统计

2021.3 动物园的猪@piginzoo.com

由来

- 盆友的问题：

A、B两个系统，分别有1000万号码，要判断是否不一致号码大于10000了，需要至少采样多少可以判断号码一致性出问题了？

- 抽样多少个合适啊？
- 抽样之后，计算出一个频率来，可以用这个频率可以代表整体么？
- 这个抽样频率值有多可信？可量化么？

统计

- 概率：大家都知道，直接跳过
- 统计：
  - 统计要干啥？
    - “从总体中抽取样本构造统计量，然后通过样本性质去推断总体性质”
    - 讲人话：就是要用抽样来估算总体
  - 估算总体的啥？
    - 总体的分布如果知道，那就是估计他的参数了
    - 总体的分布不知道，那只能估点泛泛指标（均值、方差）
  - 统计就3问题：
    - 抽样分布
    - 参数估计
    - 假设检验

那就开始抽吧

- 抽出来的那个叫样本
- 1个,2个,…一堆样本，然后用它们各种组合就得到统计量
  - 统计量=T(X1,X2,…,Xn)
- 不同抽样，统计量就不同，他是个“随机变量”呀
  - 来来，看一些统计量
    - 样本均值：$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$
    - 样本方差：$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

概念

- 总体分布
- 总体均值
- 总体方差
- 总体比例 $\boldsymbol{\pi}$：这个特殊说一下，比如人群中男的比例

- 统计量：可视为一种随机变量
- 抽样分布：有统计量形成的分布
- 抽出的样本的均值、方差等，都是统计量
- 样本比例的抽样分布: **p**

先说说“样本均值”这货 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$

- 默念：这货是统计量；这货是随机变量；这货也会有分布
- 中心极限定理

独立同分布的中心极限定理

设随机变量X_1，X_2，……X_n，……独立同分布，并且具有有限的数学期望和方差：$E(X_i)=\mu$，$D(X_i)=\sigma^2 \neq 0(k=1,2,\ldots)$，则对任意x，分布函数

$$F_n(x)=P\left\{\frac{\sum_{i=1}^{n}X_i-n\mu}{\sigma\sqrt{n}}\le x\right\}$$

满足

$\chi^2(n-1)$ 称为自由度为 $n-1$ 的卡方分布

$$\lim_{n\to\infty}F_n(x)=\lim_{n\to\infty}\left\{\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}\le x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt=\Phi(x)$$

该定理说明，当n很大时，随机变量$Y_n=\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}$近似地服从标准正态分布N(0，1)。因此，当n很大时，$\sum_{i=1}^{n}X_i=\sqrt{n}\sigma Y_n+n\mu$近似地服从正态分布$N(n\mu, n\sigma^2)$。该定理是中心极限定理最简单又最常用的一种形式，在实际工作

讲人话：样本均值，符合正态分布： $\bar{x}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$

当 n 比较大时，$\bar{X}$ 近似服从 $N\left(\mu,\frac{\sigma^2}{n}\right)$，等价地有 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$

再说说“样本比例”这个统计量

- 样本比例 $\hat{p}=\frac{X}{n}$
- 当n足够大，$\hat{p}$服从均值为π、方差为$\frac{\pi(1-\pi)}{n}$的正态分布

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

再说说“样本方差”这个统计量

- 样本方差和总体分布有关（之前均值是无关的）
- 就拿正态分布来说

设总体分布为 $N(\mu,\sigma^2)$ 的正态分布，则样本方差 S^2 的分布为

$$(n-1)\,S^2/\sigma^2 \sim \chi^2(n-1)$$

$\chi^2(n-1)$ 称为自由度为 $n-1$ 的卡方分布

- 那么，问题来了，啥是卡方分布？

- 统计就3问题：
  - 抽样分布 ✔
  - 参数估计 🡐
  - 假设检验

参数估计

- 前面谈的估计量都咋分布
- 我们的“正事”是用这些估计量来估总体
- 都估啥：“总体的均值、方差、比例”
- 两种估法：
  - 点估计
  - 区间估计

参数估计 之 点估计

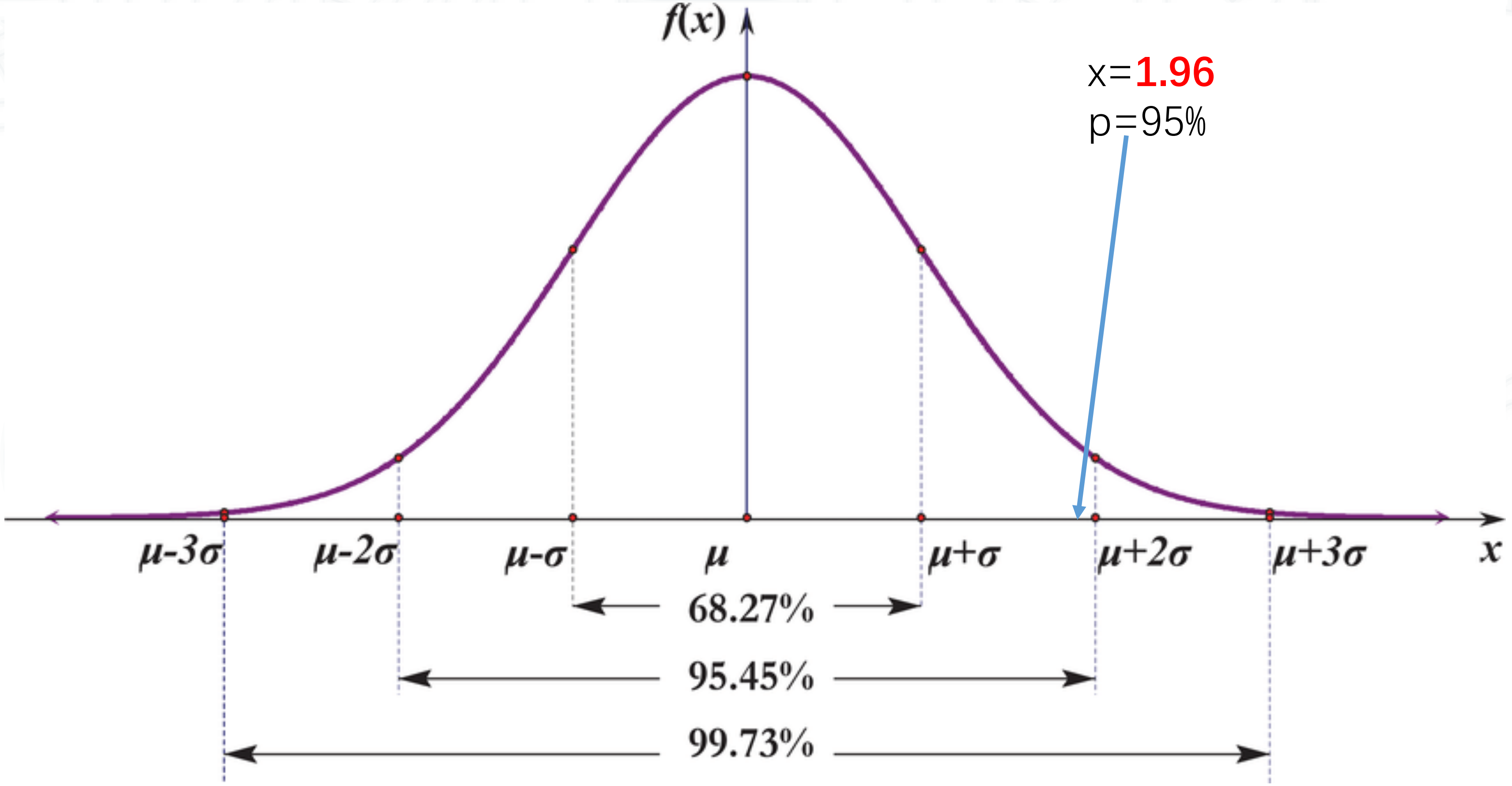
- 用 $\bar{x}$ 来估计总体均值μ
- 用样本比例p来估计总体比例π
- 用样本方差s来估计总体方差δ

- 点估计的问题：
  - 无法给出估计的可靠性度量

基础知识

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

- 正态分布和3**δ**

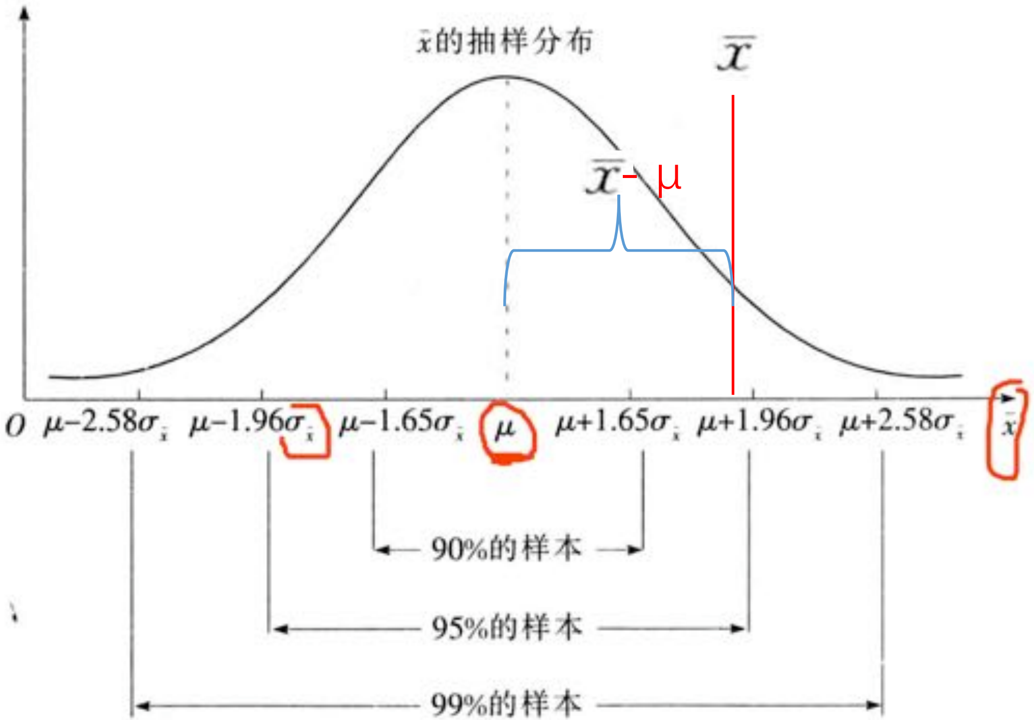


参数估计 之 区间估计

- 给出总体参数的一个范围
- 范围是由统计量+-后得到
- 对估计值和真实值差异给出概率度量
- 3δ原则

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

总体均值、方差



参数估计 区间估计

- 置信水平 ***α***
- z值： $z_{\alpha/2}$

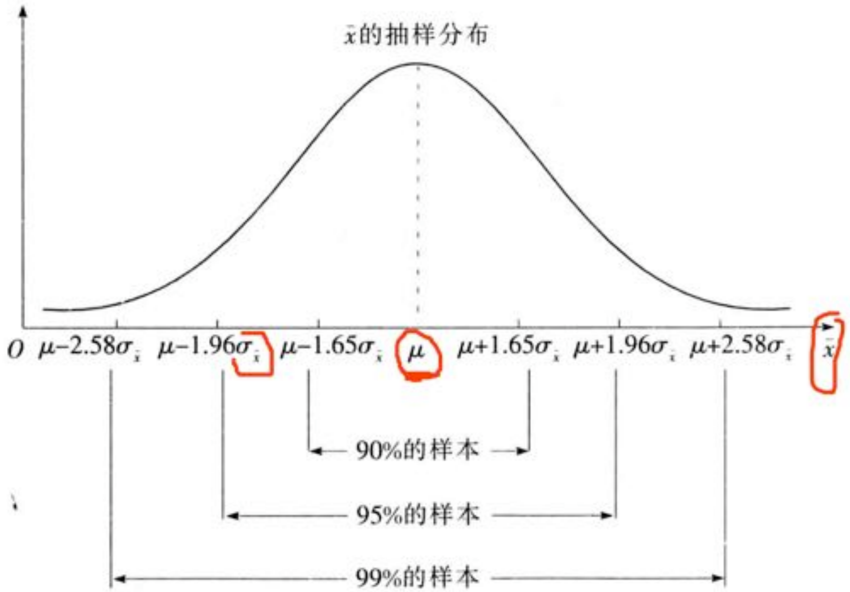
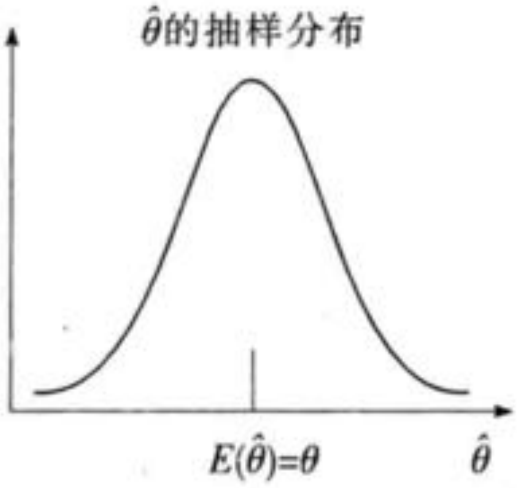


表 7—1 常用置信水平的 $z_{\alpha/2}$ 值

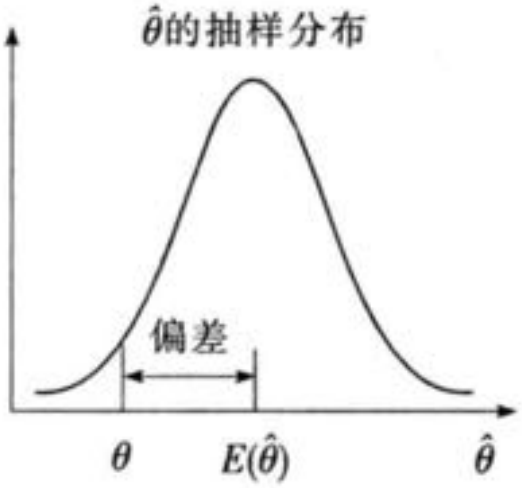
置信水平	α	$\alpha/2$	$z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.96
99%	0.01	0.005	2.58

参数估计的评价

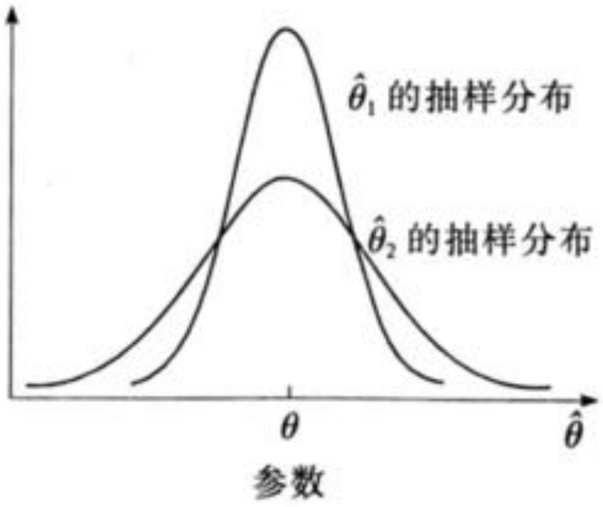
- 无偏性：估计量的期望=总体参数
  - $E(\bar{x})=\mu, E(p)=\pi \quad E(s^2)=\sigma^2$
- 有效性： 两个估计量，和总体参数方差更小的更有效
- 一致性：样本加大，估计量会更接近总体的参数值

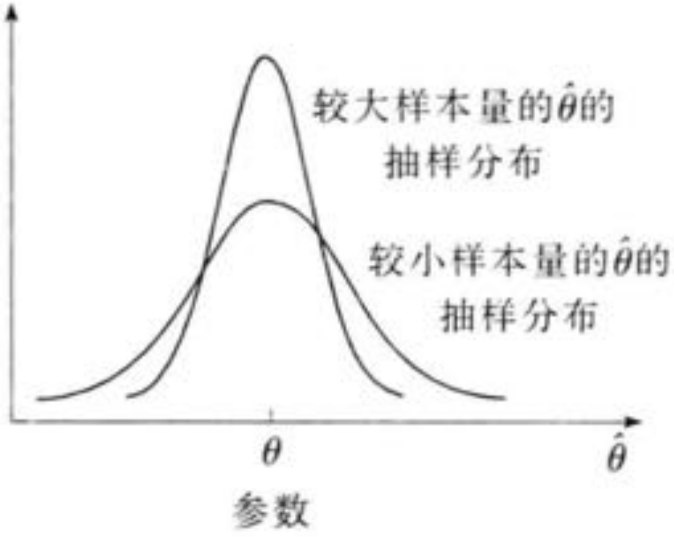


(a)无偏估计量



(b)有偏估计量





参数估计：均值区间估计

- Case1 方差**δ**已知（不用知道均值，当然啦，你要估嘛）
  - Case1.1：如果小样本+总体符合正态
  - Case1.2：大样本

$$z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1) \longrightarrow \bar{x}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ 称为置信下限，$\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ 称为置信上限

α 是事先所确定的一个概率

值，也称为风险值，它是总体均值不包括在置信区间的概率；$1-\alpha$ 称为置信水平；

如果总体服从正态分布但 σ^2 未知，或总体并不服从正态分布，只要是在大样本条件下，式（7.1）中的总体方差 σ^2 就可以用样本方差 s^2 代替，这时总体均值 μ 在 $1-\alpha$ 置信水平下的置信区间可以写为：

$$\bar{x}\pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

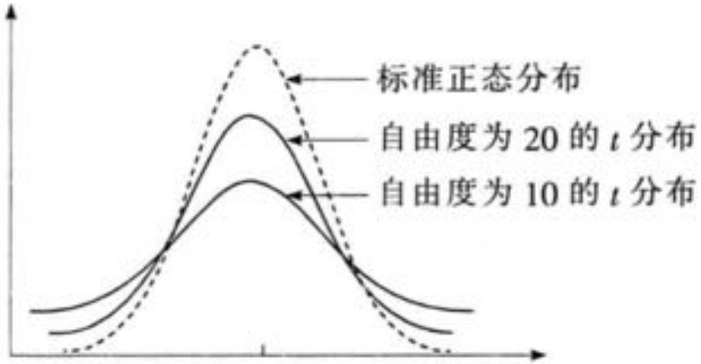
参数估计：均值区间估计

- Case2：如果小样本+总体符合正态+方差未知
- *注：大样本前面讨论过了*

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

- 样本均值，符合自由度为n-1的T分布

什么是T分布？



根据 t 分布建立的总体均值 μ 在 $1-\alpha$ 置信水平下的置信区间为：

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

卡方分布 χ^2

定义 6.3 设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立，且 $X_i(i=1, 2, \cdots, n)$ 服从标准正态分布 $N(0, 1)$，则它们的平方和 $\sum_{i=1}^{n} X_i^2$ 服从自由度为 n 的 χ^2 分布。

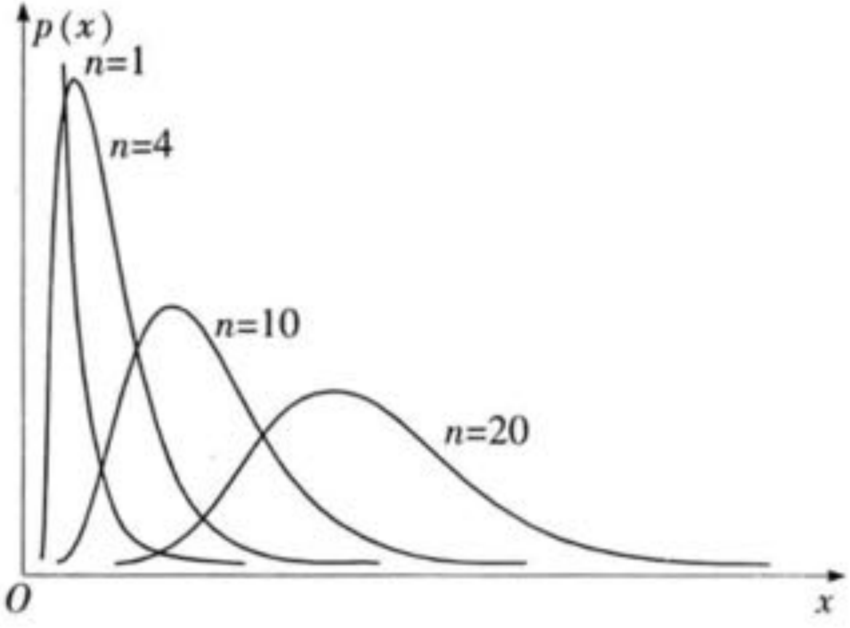
- n叫做自由度
- χ^2、T、F分布密度函数很复杂，都不给出了

χ^2 分布的数学期望为：

$$E(\chi^2) = n$$

χ^2 分布的方差为：

$$D(\chi^2) = 2n$$



t分布

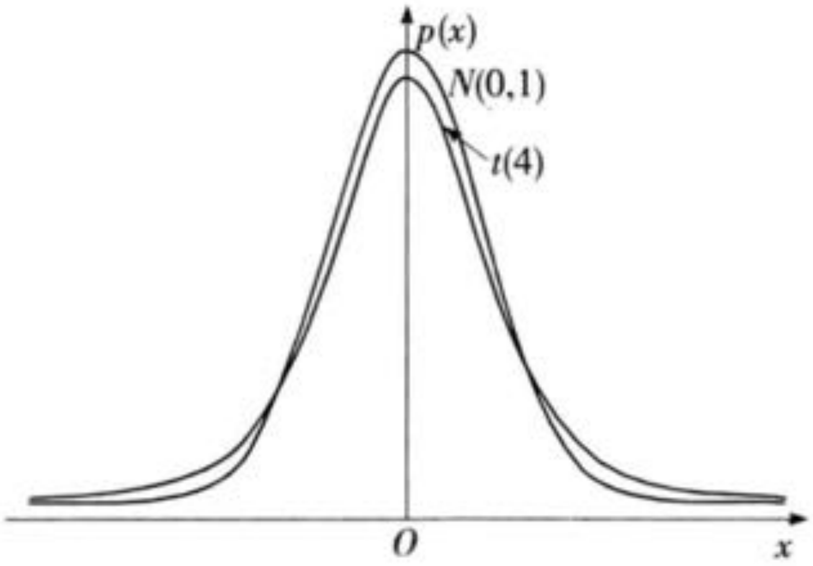
定义 6.4 设随机变量 $X \sim N(0, 1)$，$Y \sim \chi^2(n)$，且 X 与 Y 独立，则

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

称为 t 分布，记为 $t(n)$，其中，n 为自由度

当 $n \geqslant 2$ 时，t 分布的数学期望 $E(t) = 0$。

当 $n \geqslant 3$ 时，t 分布的方差 $D(t) = \frac{n}{n-2}$。



设 X_1，X_2，…，X_n 是来自正态分布 $N(\mu, \sigma^2)$ 的一个样本，$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$，则 $\frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t(n-1)$

称为服从自由度为（$n-1$）的 t 分布

参数估计：总体比例的区间估计

- 先回忆啥是“比例”
- 只讨论大样本情况

当样本量足够大时，比例 p 的抽样分布可用正态分布近似

p 的数学期望为 $E(p)=\pi$；p 的方差为 $\sigma_p^2=\dfrac{\pi(1-\pi)}{n}$

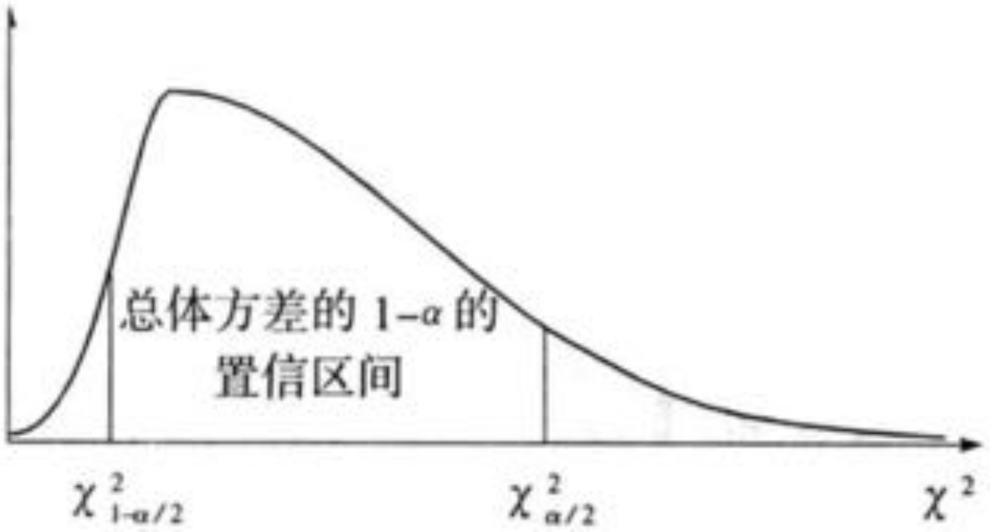
- 标准化后：$$z=\frac{p-\pi}{\sqrt{\pi(1-\pi)/n}}\sim N(0,1)$$

- 总体比例$\boldsymbol{\pi}$在1-**α**的置信水平下的置信区间为：

$$p\pm z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$$

参数估计：总体方差的区间估计

- 只讨论正态总体
- 样本方差符合自由度n-1的χ^2分布



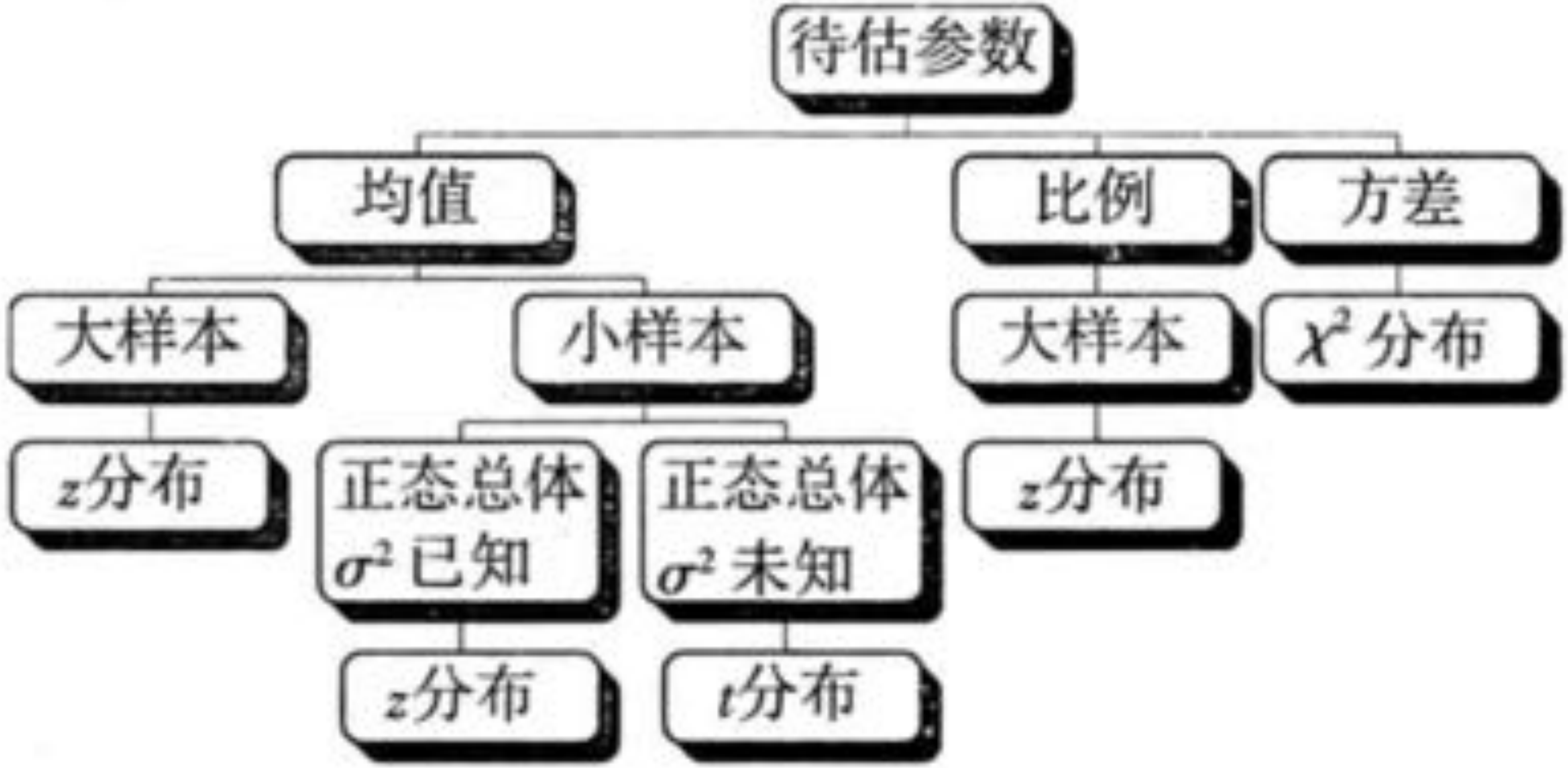
$$\chi^2_{1-\alpha/2} \leqslant \chi^2 \leqslant \chi^2_{\alpha/2}$$

由于 $\dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$，可用它来代替 χ^2，于是有

$$\chi^2_{1-\alpha/2} \leqslant \frac{(n-1)s^2}{\sigma^2} \leqslant \chi^2_{\alpha/2}$$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

总体参数估计总结

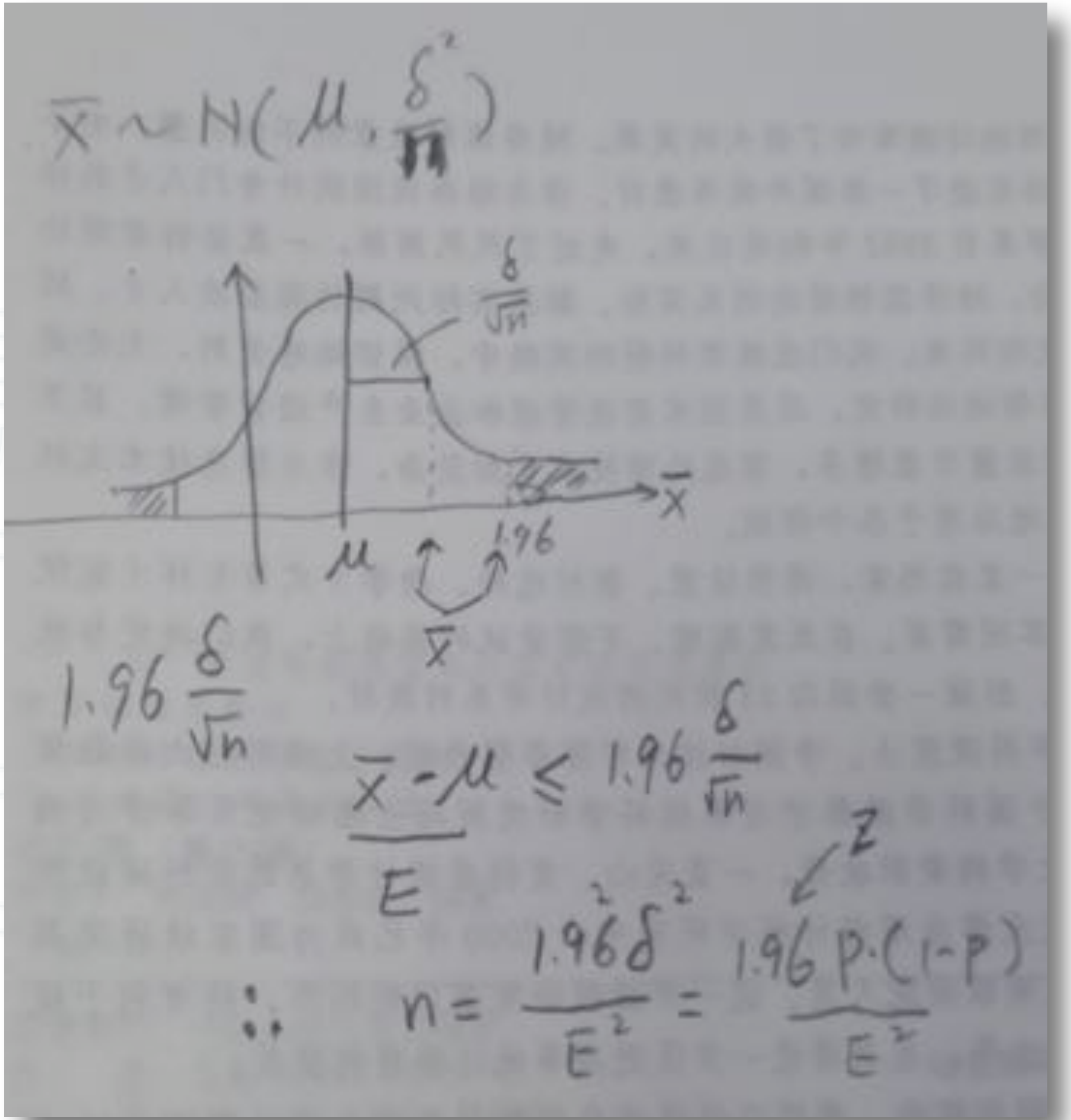


待估参数
均值
比例
方差
大样本
小样本
大样本
χ^2 分布
z分布
正态总体 σ^2 已知
正态总体 σ^2 未知
z分布
z分布
t分布

样本量的确定

- 抽样，你抽多少合适？
- 怎么定义合适？
- 估均值的时候需要多少样本？
- 估比例的时候需要多少样本？

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



（二）简单随机抽样，给定估计比例$\hat{P}$的精度（不考虑回答率）

1. 有限总体或不重复抽样情形

必要样本量计算公式

$$n=\frac{z^2\hat{P}\left(1-\hat{P}\right)}{e^2+\frac{z^2\hat{P}\left(1-\hat{P}\right)}{N}}$$

2. 无限总体或重复抽样情形

必要样本量计算公式

$$n=\frac{z^2\hat{P}(1-\hat{P})}{e^2}$$

为确定n，需要知道

- 期望的误差界限e
- 与给定置信水平相对应的Z
- 总体大小N
- 总体方差估计$\hat{P}(1-\hat{P})$

其中总体方差估计通常需要根据历史数据获取或取最大方差0.25。

这个式子怎么来的？

- 可以使用概率方式来解决这个问题，设随机变量X为不一致，0为不一致，1位一致，所以p(0)=0.0001，p(1)=0.9999，这样，1000万号码，就是1000个号码出现不一致。
- 然后，我需要去随机采样，我要保证，采样到某个数量后，我计算不一致的比率，用这些采样样本，去计算我的区间估计，假设我需要置信度为95%的话，也就是2∂，然后我反向计算，这个时候需要的样本容量
- p=0.0001（1000万里有1000个就要触发业务，所以概率是0.0001是一个边界值）
- z=2，也就是2个sigma，也可以是1.96，大概是95%的置信度。
- e=0.000005，是误差，1000人，误差超过500个人都是可以接受的，也就是误差是0.00005，
- 然后套入公式2

- 统计就3问题：
  - 抽样分布 ✔
  - 参数估计 ✔
  - 假设检验🡸

假设检验

由统计资料得知，1989 年某地新生儿的平均体重为 3 190 克，现从 1990 年的新生儿中随机抽取 100 个，测得其平均体重为 3 210 克，问 1990 年的新生儿与 1989 年相比，体重有无显著差异？

- 参数估计，是用样本去估总体参数
- 而假设检验，是先假设一个参数情况，然后用样本去验证它
- 原假设： $H_0: \mu = 3\,190$(克)
- 备择假设： $H_1: \mu \neq 3\,190$(克)
- 两者互斥，只能接受一个
- 否定一个，意味着接受另外一个

假设检验

标准正态分布

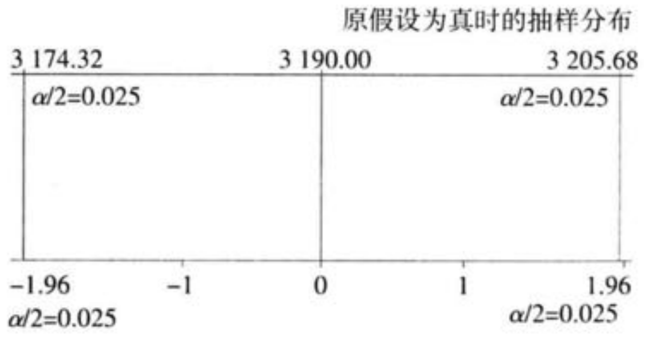
$H_0: \mu = 3\ 190$（克）

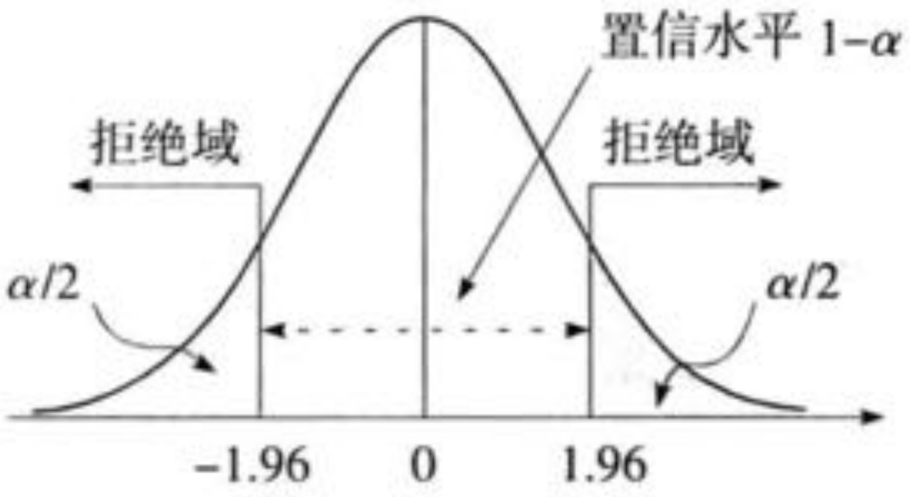
$H_1: \mu \neq 3\ 190$（克）

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \rightarrow N(0,\ 1)$$

$\mu = 3\ 190$，$\sigma = 80$，$n = 100$，$\alpha = 0.05$ 时得到的置信区间

（3174.32 ~ 3205.68）





假设检验

- 第1类错误：**α**假设，弃真错误：原假设H0为真，但是我们拒绝了
- 第2类错误：**β**假设，取伪错误：原假设H0为假，但是我们接受了

项目	没有拒绝 H_0	拒绝 H_0
H_0 为真	$1-\alpha$（正确决策）	α（弃真错误）
H_0 为伪	β（取伪错误）	$1-\beta$（正确决策）