CLT

对大量分布都有近似为正态分布， *但对二项分布有特殊的规则*

3.6 De Moivre-Laplace CLT

When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

Currently recommended
np > 15
n(1-p) > 15

应用：因为把这些全算出来太麻烦了，可以用CLT近似

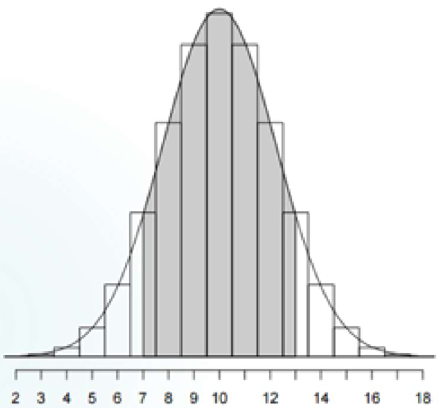
$$P(7 \leq X \leq 13) = \sum_{k=7}^{13} \binom{20}{k} 0.5^x (1-0.5)^{20-k}$$

校正偏差：
eg顶点处分别少计了一半的矩形

3.6 Improving approximation

35

Take for example a Binomial distribution where $n = 20$ and $p = 0.5$, we should be able to approximate the distribution of X using $N(10, \sqrt{5})$.



It is clear that our approximation is missing about 1/2 of P(X = 7) and P(X = 13), as $n \to \infty$ this error is very small. In this case P(X = 7) = P(X = 13) = 0.073 so our approximation is off by ≈ 7%.

当n不趋于∞时要考虑误差（如果格子数目不太大）（如果n=200+ 校正的必要性不大ry）

Binomial probability:

$$P(7 \leq X \leq 13) = \sum_{k=7}^{13} \binom{20}{k} 0.5^x (1-0.5)^{20-k}$$

？？？？

Naive approximation:

$$P(7 \leq X \leq 13) \approx P\left(Z \leq \frac{13-10}{\sqrt{5}}\right) - P\left(Z \leq \frac{7-10}{\sqrt{5}}\right)$$

Continuity corrected approximation:

$$P(7 \leq X \leq 13) \approx P\left(Z \leq \frac{13+1/2-10}{\sqrt{5}}\right) - P\left(Z \leq \frac{7-1/2-10}{\sqrt{5}}\right)$$

连续性校正近似

分别改变半格

拓展：joint distribution

联合概率 - 离散

Para	>18	<18
F	X	Y
M	Z	U

X+Y（一条边） X+Z 边际概率
相交得X（联合概率）

- 连续
  曲线+曲线→曲面
  双重积分$\int_a^b\int_c^d f_{XY}(x,y)dxdy$
  也可以通过此式只研究x 或y

CH5 Estimation

Comparison between parameters and statistics	
Statistics	*Parameters*
• Describes a sample	• Describes a population
• Always known	• Usually unknown
• Random, changes upon repeated sampling	• Fixed
• Ex: $\bar{X}, S^2, S$	• Ex: μ, σ^2, σ

选出

统计量： *不依赖未知参数，*与样本有关

Estimation估计
样本代表整体← **不能偏** 否则推断所得全是错的

统计推断inference（估计+评估合理性）

- 点估计point estimation
- 参数估计 interval estimation
  *eg*统计20个人的出货率
  统计方法：最小值
  可以得到结果（一个值）
  *eg2*统计20个人的出货率
  统计方法：平均值
  也可以得到结果（一个值）
  *eg3*统计20个人的出货率
  统计方法：男的出货率平均值
  也可以

总之方法是随便选的但准确性不一样；点估计=结果为一个值

*eg4*最大出货率&最小出货率
可以得到区间
*eg5*去掉一个最大最小ry再选剩下的最大最小
还是可以得到区间

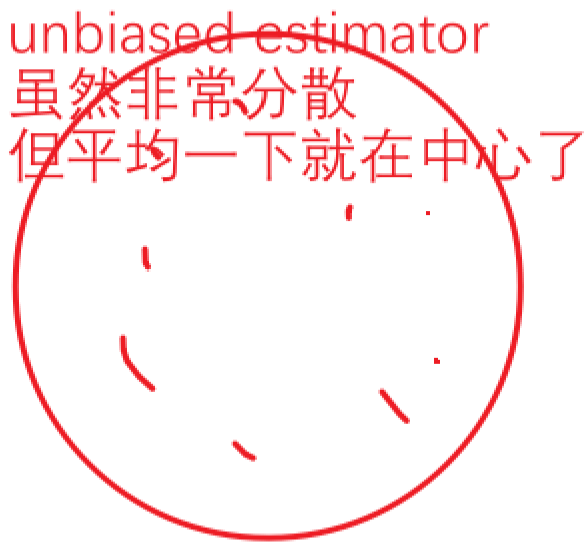
参数估计=得到范围
净含量±ry

如何评估以上方法的好坏：

点估计

- Accuracy准确性
  - 对统计量θ的估计称为$\hat{\theta}$
  - unbiased estimator E($\hat{\theta}$)=θ

  -



- Precision精确度
  - 所有估计无论是否有偏都有精度，无偏>有偏，无偏小方差>无偏大方差
  - **MSE mean square error= Bias^2+Var** 更为一般的比较 越小越好
- Consistency相合性
  - 样本量小的时候没有明显问题但随着样本量增大，不收敛的估计是一个很糟糕的估计

Standard Error

我们希望$\hat{\theta}$无偏小方差
标准误 se($\hat{\theta}$)=(var($\hat{\theta}$))$^{1/2}$
对$\hat{\theta}$的标准差、越小越好
格式：估计身高 $\hat{H}$=1.73m（se=0.03m）一定要写上se

如何评估以上方法的好坏：

区间估计

- Confidence Level置信水平是一个概率值aka置信度
  - **P(CI∋θ)** 置信区间包含总体平均值的概率（表示“**至少**有np的CI包含”，可能会更多）

抽取的100个样本，

≠样本量为100 √取了100个样本，也就是重复了100次取样，每个样本样本量n， **因此有100个CI**

- Precision精度
  - CI不是越大越好eg-∞~+∞ P=1，但没意义
  - 总得来说，窄的CI比宽的CI能提供更多的有关总体参数的信息
  - **固定**置信水平，n↑，CI越窄，但两者变化 *速度*不同

计算

（不考）极大自然估计MLE maximum likelihood estimation

想要让概率达到最大值，所对应的参数

eg抛硬币三次正面朝上（likelihood：表示该概率的参数方程，内含参数p，p取值范围0~1，方程的值就不同），当p（抛一次正面朝上，不一定=0.5）为多少时概率最大

就算p=0.000001也可能抛十次三次朝上，只不过概率非常小

总会出现最值点maximum likelihood，称此时的p为最大自然估计（最可能的θ）
$\hat{\theta}$

区间估计（基于点估计）
如果P(LB≤θ≤UP)≥95%，则CI=【LB，UB】

you need to know:

- 点估计
- critical values for the test statistic? ? ? ? ? ? ? 临界值
- se of point estimate
- sample size
  CI=point extimate ± (critical values × standard error) 这样计算也行，括号内称为margin of error

---

inference on population proportion

样本量够大

binary outcomes
bernoulli trails
对于这个二项分布而言p的大小可以被估计（样本估计总体）$\hat{p}$=Y/n
E($\hat{p}$)=E(Y)/n 证明$\hat{p}$是无偏的

Var(p̂)=Var(Y)/n^2

2.2 Property of $\hat{p}$

$\widehat{p}$ is a unbiased estimator of p. That is,

$$E(\widehat{p}) = p.$$

To quantify the precision of $\widehat{p}$,

$$\operatorname{var}(\widehat{p}) = \frac{p(1-p)}{n}$$

Question: What is the (asymptotic) distribution of $\widehat{p}$?

$$\widehat{p} \sim \mathcal{AN}\left(p, \frac{p(1-p)}{n}\right)$$

p̂的极限分布/渐进分布： **正态分布**（CLT）

如何将区间估计与分布联系在一起：（画图）

Y/n=p̂~N(p,p(1-p)/n)①

↓

标准化N(0,1)

↓

z_α是上分位数的写法，$z_{1-\alpha}$是下分位数的写法

↓

critical value 即z（分位数的值）

↓

①中p是未知的，需要用已知的p̂替换p

p̂~N(p̂,p̂(1-p̂)/n)←这是方差的 *估计值*，真实值无法计算

得出的图像就是对p̂的分布的估计

↓

确定上下界，中间的图形面积为95%（或别的）即置信区间（对称性）②

Let us define z_α be the upper α percentage point of the standard normal distribution, i.e., $P(Z > z_\alpha) = \alpha$.

An approximate $100(1-\alpha)\%$ confidence interval for p is

standard error

$$\left(\widehat{p} - \boxed{z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},}\ \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right).$$

margin of error 误差幅度

①对称：两边各α/2

②假如分布不对称

也是两边各α/2（看面积）

③只关心置信区间的上/下限

confidence limit置信界 eg -∞~上限（只关心大小不关心小多少ry

单边α

？？？

example

在观测之前都是猫箱但可以计算概率 "包含真实值的可能性"

2.4 CI - Interpretation

21

- In practice, we only perform the study once, and get CI (0.193, 0.237).
- We have no way of knowing if this interval that we calculated is one of the 95% (that covers the true parameter) or one of the 5% that does not.
- Thus **we are 95% confident that the true response rate of this new treatment in 2L NSCLC patients is between 0.193 and 0.237**
- Every time that we calculate a 95% CI, then there is a 5% chance that the CI does not cover the quantity that you are estimating.

要求：

np̂>5；n(1-p̂)>5（实际>10更好）

满足才能用CLT并近似成normal distribution

样本量小

用二项分布做 **不用正态近似**

2.7 What if the sample size is small*

Example. As part of a demographic survey of her Statistical Consultant course, Miss Wang asks students if they have any statistical consultant experience before. The following are the data from her course:

Experience	Count
Yes	7
No	14
----------	--
Total	21

Point estimate = 7/21 = 0.33

Calculate exact binomial CI satisfying $P(p_{LB} < p < p_{UB}) = 1 - \alpha,$

(e.g., Clopper-Pearson interval) ⇨

$$\sum_{k=0}^{k} \binom{n}{k} p_{UB}{}^{k} (1 - p_{UB})^{n-k} = \frac{\alpha}{2}$$

$$\sum_{k=x}^{n} \binom{n}{k} p_{LB}{}^{k} (1 - p_{LB})^{n-k} = \frac{\alpha}{2}$$

Calculation demanding…

Let R help you

清华大学统计学研究中心

point estimator of population mean

首先假定每个Y都服从正态分布N(μ,σ^2)

那么**Ȳ**一定也服从正态分布（不是估计）N(μ,σ^2**/n**)

$\overline{Y}$ is an **unbiased** estimator of μ.

$\text{se}(\overline{Y}) = \sigma/\sqrt{n}$ gives the variability of $\bar{Y}$, i.e. how “close” we expect $\bar{Y}$ to be to μ.

可以用**Ȳ**估计μ，样本s估计σ

为什么是/(n-1) *内含偏差* ***因为xi-xbar，如果xi-μ就不存在误差，分母为n*** *需要校正（如果恰好=σ才是无偏的，但估计值是有偏的，倍数为(n-1)/n，所以不/n*

☆☆☆$\hat{\sigma}^2 = s^2 = \Sigma_{i=1 \sim n}(x_i - \bar{X})^2/(n-1)$

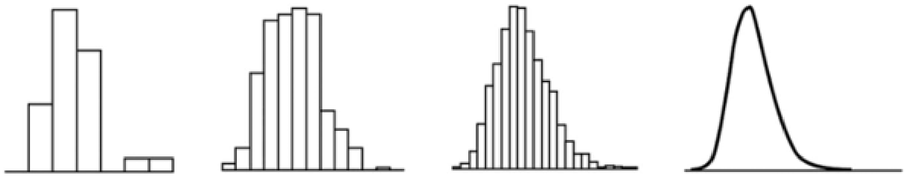
$\bar{Y}$对μ的估计到底有多好（能否确定抽样分布）

3.2 Point estimator of population mean - Example

The **standard error of the mean** is

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}.$$

For the butterfly wings, $SE_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{2.48}{\sqrt{14}} = 0.66 \text{ cm}^2$.

	$n = 28$	$n = 280$	$n = 2{,}800$	$n \to \infty$
$\bar{y}$	32.81	32.44	32.59	$\bar{y} \to \mu$
s	2.46	2.49	2.47	$s \to \sigma$
SE	0.46	0.15	0.05	$\text{SE} \to 0$
Sample distribution				

（n是样本量）s是 **样本**的标准差，s可以估计σ，σ是定值而s比较稳定；$\bar{Y}$也会趋近μ（无偏估计）

SE$\bar{Y}$=s/$n^{1/2}$

n→∞ se→0

估计值非常集中，几乎恒定

3.4 CI for μ. Assume normality and KNOWN σ

Recall if $Y_1, \ldots, Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution and σ^2 is **known**, then

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

Similar to CI derivation in population proportion, a $100(1-\alpha)\%$ CI for μ is given by

$$\left(\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

3.5 t distribution

However, population standard deviation σ is usually unknown. Replacing it with the **sample standard deviation** S, we get a new sampling distribution:

$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

a t **distribution** with degrees of freedom $\nu = n - 1$.

The t distribution was published by Gosset in 1908 & related to quality control at Guinness brewery.

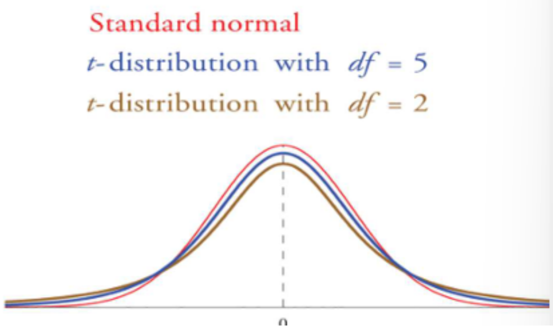
用s估计σ的时候分布不是很正态，变为t分布（但是也和标准正态分布很近）

分母是s/n$^{1/2}$

3.5 t distribution

The t distribution has the following characteristics:

1. It is continuous and symmetric about 0.
2. It is indexed by a value ν called the degrees of freedom.
3. As $\nu \longrightarrow \infty$, $t(\nu) \longrightarrow \mathcal{N}(0,1)$.
4. When compared to the standard normal distribution, the t distribution, in general, is less peaked and has more probability (area) in the tails.



永远比标准正态分布低，不会超过红色图线

3.6 CI for μ. Assume normality and UNKNOWN σ

Recall if $Y_1, \ldots, Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution and σ^2 is **unknown**, then

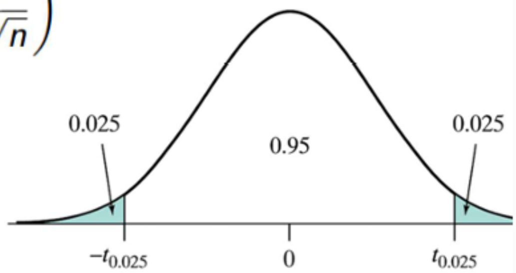
$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

A $100(1-\alpha)\%$ CI for μ is given by

$$\left(\overline{y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},\quad \overline{y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

where S is the sample standard deviation.

We replace 1.96 (from a normal) by the equivalent t distribution value, denoted $t_{0.025}$ for 95% CI.



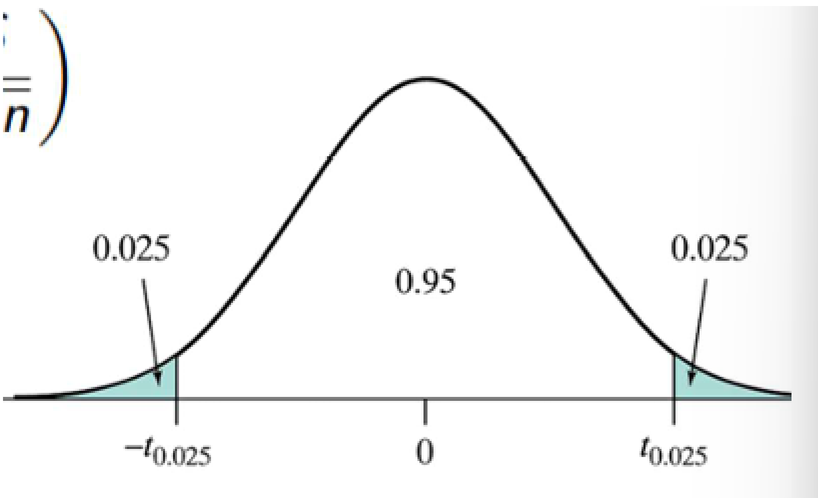
清华大学统计学研究中心

此时critical value来自t分布而不是正态分布

估计（可能?）都是随机变量
→研究它的抽样分布
→如果能找到分布的均值和方差（不知道确切的）（及其分布）
→就可以计算各种内容
（不一定所有sampling distribution都是正态分布，也可以是卡方分布 xx分布 及其变形（比如标准化操作）ry）

自由度

是一个参数 决定了整个分布的形状
上图 中心化的t分布 中心是0



对称的
找1-α对称区间时 分别找(1-α)/2（上分位数+下分位数）
qt quantile(in R 算分位数)