

Linear combination of random variable and Joint distribution

- 单位统一

$$Y = 0.001X_4 + X_5$$

Y是一个 新的变量 (可以写成线性组合)

问题: 线性组合 合在一起 是什么分布

- 线性组合 有系数

$$Y = c_1X_1 + c_2X_2 + \cdots + c_pX_p$$

is a linear combination of X_1, \dots, X_p

- Properties

- Mean

$$E(Y) = c_1E(X_1) + c_2E(X_2) + \cdots + c_pE(X_p)$$

就是线性的

- Variance

$$V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \cdots + c_p^2V(X_p) + 2 \sum_{i < j} c_i c_j \text{cov}(X_i, X_j)$$

比较复杂 系数是平方 内含协方差 注意系数两倍 (规定了 $i < j$ 才有两倍) (两两之间相关 如果独立则

为零)

The image shows a blue chalkboard with handwritten mathematical formulas. At the top, it says $Y = X_1 + X_2$. Below that, it shows $\text{Var } Y = 4 \text{Var } X_1$ with a red checkmark. Then, it shows $\neq \text{Var } X_1 + \text{Var } X_2$ with a red 'X' and a red wavy line underneath. Below this, it shows $+ 2 \text{cov}(X_1, X_1)$ with a red checkmark and a red wavy line underneath.

平均数

- 任何分布，对变量 $X_1 \dots X_p$ 做平均，假定 $E(X_i) = \mu$ 都相等
那么 均值的均值 $E(\bar{x}) = \mu$ || \bar{x} 也是一个随机变量

每一个变量服从同一种分布，这个变量和其他变量都一样拥有同样的均值和方差

平均值的均值=任何一个变量的均值

平均值的方差=1/p任何一个变量的方差 离散度降低 $V(\bar{x}) = \sigma^2/p$

eg 比较全班女生的个人平均体温

正态分布

- 如果 X_i 服从**正态分布**，均值和方差不相等，（互相独立，没有cov）

$X_1 \sim N(1, 2)$

$X_2 \sim N(4, 5)$

$X_p \sim N(715, 1129)$

正态分布的叠加依然是正态分布，可以通过 μ 和 σ 确定一个n dist

都是线性的

If $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$ for $i = 1, 2, \dots, p$.

$$Y = c_1 X_1 + c_2 X_2 + \dots + c_p X_p$$

is a normal random variable with

$$\mu_Y = E(Y) = c_1 \mu_1 + c_2 \mu_2 + \dots + c_p \mu_p$$

and

$$\sigma_Y^2 = V(Y) = c_1^2 \sigma_1^2 + c_2^2 \sigma_2^2 + \dots + c_p^2 \sigma_p^2$$

i.e. $Y \sim N(\mu_Y, \sigma_Y^2)$ as described above.

解题示例

X1~()

X2~()

X1 ⊥ X2

Y=c1X1+c2X2

Y~()

计算ry

卡方分布

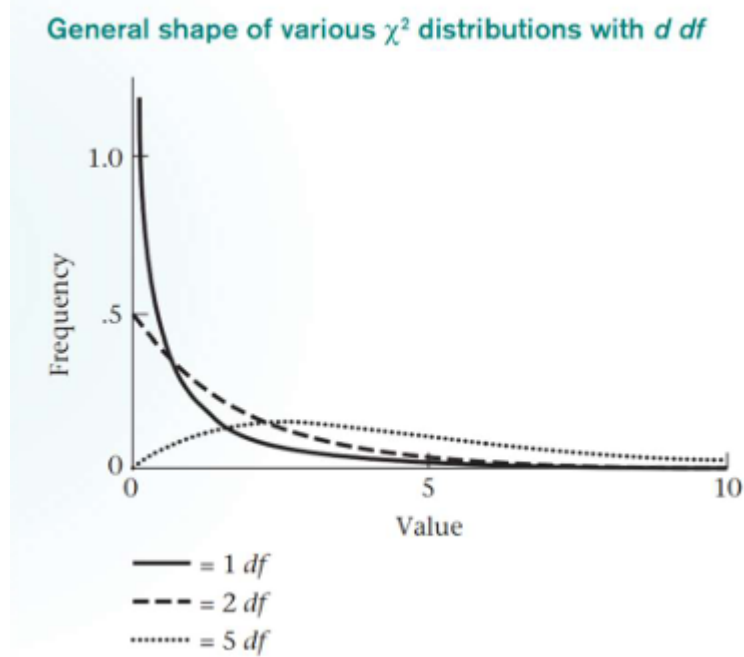
- 卡方分布Chi-square distribution, X_i 服从独立标准正态分布 (iid, 独立同分布), $G = \sum X_i^2$

$$\text{If } G = \sum_{i=1}^n X_i^2$$

where $X_1, \dots, X_n \sim N(0,1)$

只有一个参数 n ($n \geq 1$), 称为degrees of freedom(df), 记为 χ_n^2

曲线下面积为1



basic concepts of **statistics**

Population > Sample (collection of **all** elements *possessing common characteristics*)

Sample < Population (a subset of the population)

eg 总体：糖尿病患者

未来可能出现的糖尿病患者也属于这个总体

eg2 总体：2022年的糖尿病患者

100人 = 1个样本 (不等于组织样本的样本量)

实施了评价拿到黄金的人 (总体) 的调查, 有180w人参与, 约有1/4参与者 (1份随机样本) 得到随机化, 这180w人是1份样本

Sampling Methods

不随机取的也是样本，是子集就行

2.2 Sampling Methods

Sampling: the method or technique to take a sample from the population

随机抽样 (random sampling)

从有限总体中简单随机抽样或从无限总体中随机抽样。

整群抽样 (cluster sampling)

将总体划分成若干个群组，抽样时直接随机抽取群组。抽中群组中的所有抽样单位即为样本。

方便抽样 (convenience sampling)

用总体中便于取得的一些抽样单位作为样本

分层抽样 (stratified sampling)

将总体单位按某种特征或某种规则划分为不同的层(Strata)，然后从每一层中随机抽取一定量的抽样单位，组成样本。

系统抽样 (systematic sampling)

将总体中的个体按一定的顺序排列，等分成n个部分。在第一个部分内随机抽取1个个体，然后等距离在其他部分分别抽取1个个体，组成样本。

判断抽样 (judgement sampling)

由对研究总体非常了解的人主观确定总体中他认为最具代表性的个体组成样本。

概率抽样方法

非概率抽样方法

在男女比例很悬殊的地方随机抽样可能是抽不到的

怎么听着像抽卡

这种情况一定(?)要分层

整群抽样 eg 年级里抽一个班

系统抽样 eg 抽一个宿舍楼(楼号) 床号 天然顺序

方便抽样(主观)

低质量抽样-不能代表全体 虽然发了问卷，不是方便抽样，但抽到的全是经常刷主页的人

电话很贵，所以抽样到的都是有钱人，

Random sample

representative(has equal probability to be selected), independent, implementation

已知总体(出货率0.002%←这是一种分布X)，可以推断其中一个元素(出货情况未知)的出货率是0.002%

只要在总体里就符合该情况

另一个手游x 出货率0.001%←另一种分布 \tilde{x}

sampling bias

parameter in population

未知但确定

mean μ

variance σ^2

standard deviation σ

Statistics

每个变量都是随机的所以 **统计量T**也是随机的

Examples of statistics $T = T(X_1, X_2, \dots, X_n)$:

- $T = \bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ sample mean
- $T = S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$ sample variance
- $T = S = \text{sqr}t(\frac{1}{n-1} \sum (X_i - \bar{X})^2)$ sample standard deviation
- $T = M_n$ sample median
- $T \equiv 7$
- Order statistics: $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$

$T = I(X_1 < X_3)$

indicator function $T=1, F=0$

这也是一个统计量

- $T = S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$ sample variance

☆在这个情况下，如果 μ 是未知量，那么T不是统计量；如果 μ 是已知量，则T是统计量

如果能计算 \bar{X} 并用 \bar{x} 估计 μ ，如果算式中为 \bar{x} ，也是统计量，但原始是 μ 进入式子后就不是统计量

Sampling distributions 抽样分布

aka**统计量** 的分布 首先需要是个统计量

抽样分布 \neq 样本分布

抽样分布往往是一个联合分布

抽样分布是统计量的分布

统计量 T_n ，第一组样本称为 $X_1^{(1)} \sim X_n^{(1)}$...第s组 $X_1^{(s)}$

$$\begin{array}{lcl} X_1^{(1)}, \dots, X_n^{(1)} & \rightarrow & T_n^{(1)} = T(X_1^{(1)}, \dots, X_n^{(1)}) \\ X_1^{(2)}, \dots, X_n^{(2)} & \rightarrow & T_n^{(2)} = T(X_1^{(2)}, \dots, X_n^{(2)}) \\ & \vdots & \\ X_1^{(s)}, \dots, X_n^{(s)} & \rightarrow & T_n^{(s)} = T(X_1^{(s)}, \dots, X_n^{(s)}) \\ & \vdots & \end{array}$$

在重复抽样的情况下，可以估计 **统计量** 的分布

sample size n

standard error 标准误 (估计量) = 对于方差的估计 / \sqrt{n} 样本量

中位数不好算 均值使用比较广泛

一个同学测了365次体温 样本量很大 但不能代表所有人 数据可用性

抽样对象来自正态分布

iid 正态分布叠加

不正态+样本量小→根据实际情况

不正态+样本量大→CLT

central limit theorem

中心极限定理

连续型 任意分布 $\mu \sigma^2$

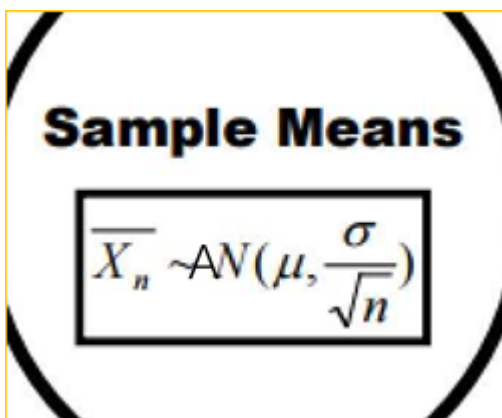
3.1 Central Limit Theorem (CLT)

23

Given that the distribution of a continuous variable in the underlying population has mean μ and standard deviation σ , the distribution of sample means computed for **samples of size n** has three important properties:

- 1. The mean of the sampling distribution is **identical** to the population mean μ .
- 2. The standard deviation of the distribution of sample **means is equal to σ/\sqrt{n}** .
- 3. Provided that **n is large enough**, the shape of the sampling distribution **is approximately normal**.

http://www.lock5stat.com/StatKey/sampling_1_quant/sampling_1_quant.html



The diagram shows the Central Limit Theorem formula for sample means. It features a large circle containing the text "Sample Means" in bold. Below this, a rectangular box contains the mathematical expression $\overline{X}_n \sim AN(\mu, \frac{\sigma}{\sqrt{n}})$.

AN: approximately

conditions

- **independence**

- sample size >30 (if not normal)

离散型

First, recall that a Binomial variable is just the sum of n Bernoulli variable:

$$S_n = \sum_{i=1}^n X_i$$

Notation:

$$S_n \sim \text{Binomial}(n, p)$$

$$X_i \sim \text{Bernoulli}(p) = \text{Binomial}(1, p) \text{ for } i = 1, \dots, n$$

X_i 是bernoulli分布那么 S_n 是二项分布

In this case,

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

\hat{p} is a sample mean!

We can use the CLT when n is large.

????

在二项分布中，既然 \hat{p} 是sample mean那么就可以用CLT

3.5 Binomial CLT

For a Bernoulli variable,

- $\mu = \text{mean} = p$
- $\sigma^2 = \text{variance} = p(1-p)$

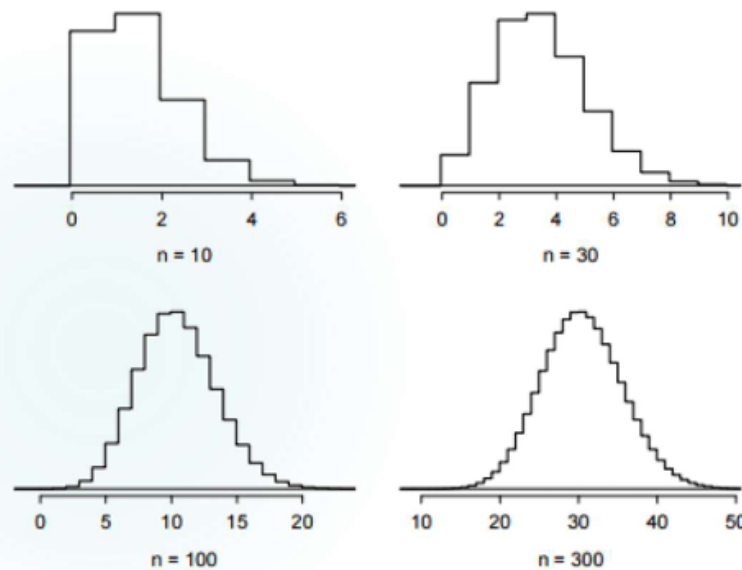
$$\bullet \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

随着 n 增加一开始是偏的其实

虽然CLT的使用条件是 $n > 30$ ，但因为此处 p 太小所以 $n=30$ 的时候偏差也很大

3.5 Histograms of number of successes

Hollow histograms of samples from the binomial model where $p = 0.10$ and $n = 10, 30, 100$, and 300 . What happens as n increases?



所以

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

$np > 15$
 $n(1-p) > 15$

$X \sim \text{Bin}(n=245, p=0.25)$