Sampling distribution

Sample mean $\bar{x} \sim N$

Sample proportion $\hat{p} \sim N$

$$\bar{x} = \frac{\sum x_i}{n} \xrightarrow[\text{theorem}]{\text{Central limit}} N(n\mu_x, \sqrt{n} S_x)$$

$$\bar{x} \sim N\left(\mu_x, \frac{S_x}{\sqrt{n}}\right)$$

$$\hat{p} = \frac{\sum x_i}{N}$$

$x_i \sim$ Benuoli distribution

$\sum x_i \sim$ Binomial (N, p)

$\uparrow$ population proportion

$Np \geq 10, \ N(1-p) \geq 10$

$\downarrow$

$$\sum x_i \sim N\left(np, \sqrt{np(1-p)}\right)$$

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

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right) \Leftarrow$$

assumption. 1) SRS. $\Rightarrow$ unbias $\Rightarrow \mu_{\bar{x}} = E(\bar{x}) = \mu_x$

2) $n \geq 30$ or $X \sim$ Normal

2.1) how to check is $X \sim$ Normal

2.1.1). sample has no skewness or no outlier. $\Rightarrow$ population is normal

2.1.2). Q-Q plot of sample is all on a straight line.

3) $n < 0.1N \Rightarrow$ independent sampling

$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$

$$\Rightarrow \quad Var(\bar{x}) = Var\left(\frac{\sum x_i}{n}\right)$$

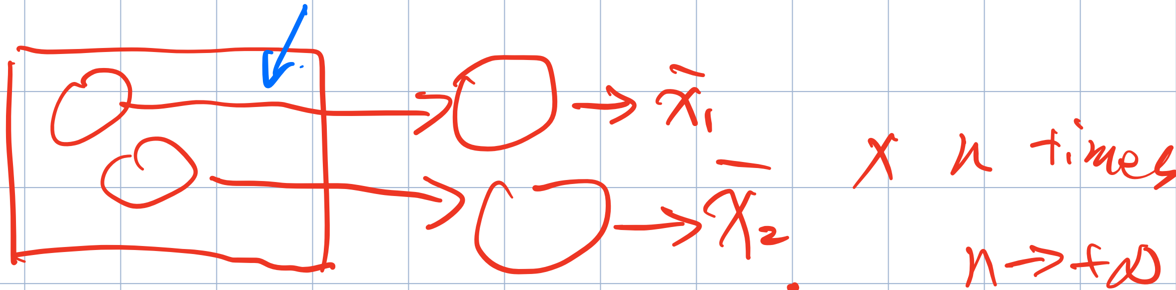
$$= \frac{1}{n^2} Var(x_1 + x_2 + \cdots + x_n)$$

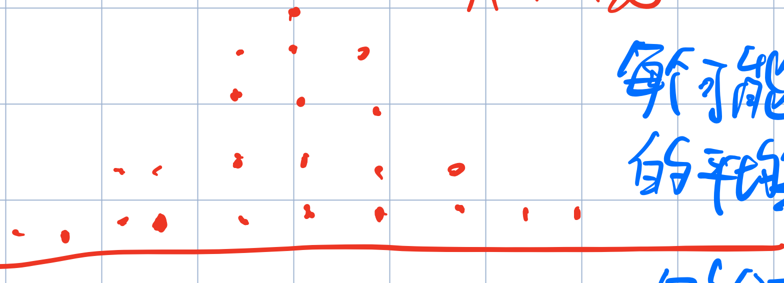
independent $\Rightarrow$ $$= \frac{1}{n^2}[Var(x_1) + Var(x_2) + \cdots Var(x_n)]$$

$$= \frac{1}{n^2} \times n \, Var(X)$$

$$= \frac{Var(X)}{n}$$

$$\sigma_{\bar{x}} = \sqrt{\frac{Var(X)}{n}} = \frac{\sigma_x}{\sqrt{n}}$$

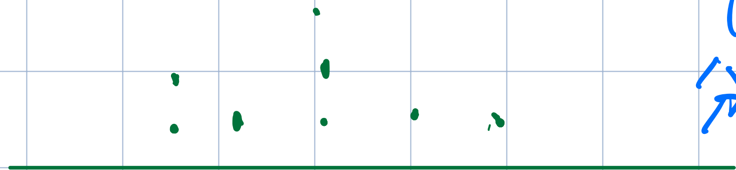




每个可能的样本
的平均身高
的分布

多个样本 $\bar{x}$ 的分布

sampling distribution

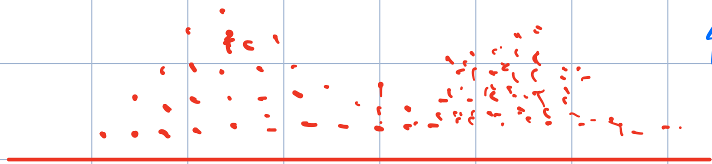


一个班级身高
分布

一个样本内数
据的分布

sample distribution

一个学校的身高分布



population distribution

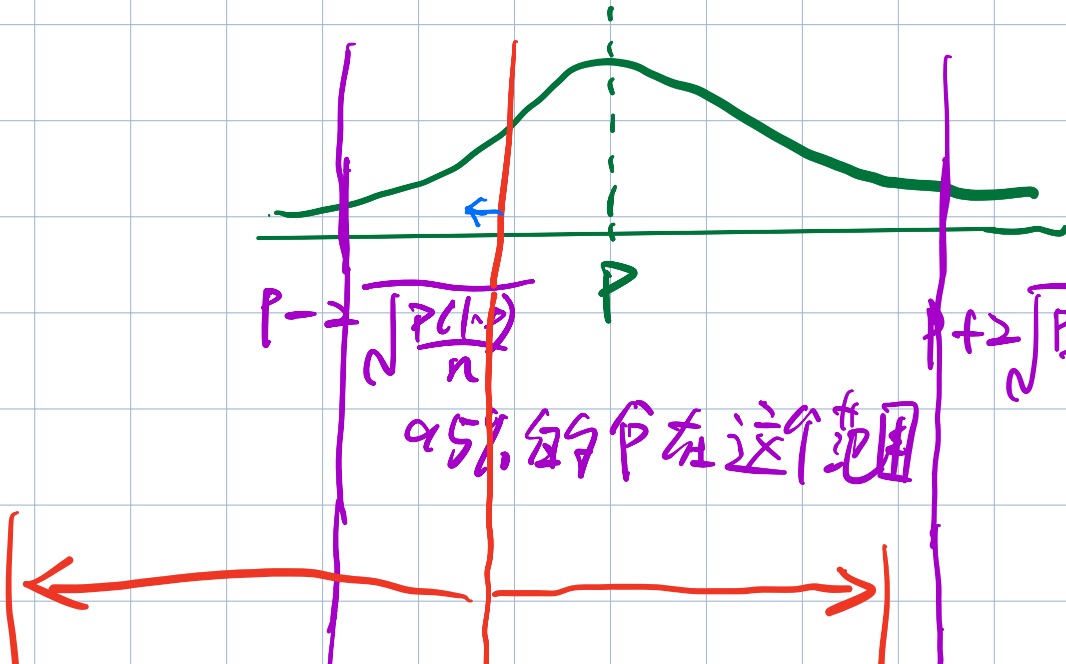
3 conditions $\rightarrow \bar{x} \sim N(\mu_x, \frac{\sigma_x}{\sqrt{n}})$

another 3 conditions $\Rightarrow \hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$

已知 $n=50$. $\hat{p}=0.47$. 求 p 大致范围.

population proportion ⇒ 未知.

当符合 3 条件时. $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$



95%的情况下 P在这个范围

$\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ $\hat{p}$ $\hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

0.27 0.67

$0.67 + 0.64 = 1.31 \Rightarrow 31\%$

$$2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < 3\%$$

95% → 1.96

$$\frac{1.96\sqrt{\hat{p}(1-\hat{p})}}{0.03} < \sqrt{n}$$

$$\sqrt{n} > \frac{2\sqrt{\hat{p}(1-\hat{p})}}{0.03}$$

$$n > \frac{2^2}{0.03^2}\,\hat{p}(1-\hat{p})$$

$$n > \frac{2^2}{0.05^2} \cdot 0.5^2.$$

$$n > 1111$$

置信区间 confidence interval.

C.I.

置信度 confidence level.

C-level

误差幅度. Margin of Error

M.O.E.

临界值 Critical Value.

C.V.

标准误. Standard Error

S.E.

CI: sample statistics ± MOE.

CI: statistics ± C.V. × S.E.

!!! 置信度不是概率.

4-step process

State：
1) 声明要做什么.
2) 声明细节.
3) 声明使用符号指代变量.

Plan
1) 声明使用方法.
2) 该方法所用条件.

Do
1) 用计算器计算.
2) 写清结果.

Conclude
做结论，汇报结果

State "We are constructing a C.I. (1)

for p at C-level = 95%. (2)

where p is the true proportion of female students in all students in our school (3)

Plan 1-sample z-interval for p.

只抽了一个样.
只给一个 pop. param. 做估计.

sampling distri 的形态、
z 代表正态分布.

要估计的 population parameter.

如果是 for $p_1 - p_2$ 则是 2-sample
两个 proportion 之差.

Do: 手算 C.I. 公式:

Statistics ± Critical Value × Standard Error

用 C-level 算出

$$\hat{p} \pm \text{invNorm}\left(\text{area}: \frac{1+C\text{-level}}{2},\ \mu: 0,\ \sigma: 1\right) \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.47 \pm 1.96 \cdot \sqrt{\frac{0.47 \times 0.53}{50}}$$

D0. 用计算器算

By using calculator 1-PropZ int $\left(\begin{array}{l} X: 23 \\ n: 50 \\ C\text{-level}: 0.95 \end{array} \right)$

We get (0.345, 0.618).

Conclude 汇报结果.

We are 95% confident that the true proportion of female student among all students in our school. is between 0.345 and 0.618.

Confidence interval done.

Hypothesis testing.

假设 检验

当别人说一个 claim，你想证明他对 or 错。

1) 当作 claim 已经成立。

2) 取样，算概率。
p-value

3) 若 p-value 较小，
说明不该抽到此样

4) 说明原假设错误

① 定假设

1) 你看到别人的 claim

eg. $p>0.7$ or $p=0.7$

2) 你同意? 不同意?

↓ 你想证 $p>0.7$ → $H_A: p>0.7$ 备则假设

↓ 你想证 $p<0.7$ → $H_A: p<0.7$

>, < 或 ≠ 改为 =,

$H_0: P=0.7$ 原假设. $H_0: P=0.7$

通过推翻原假设

来证明备则假设.

通过[原假设成立时,

抽到了极小概率样本. 但不应该

抽到极小概率样本.

故原假设不成立]

来推翻原假设.

1) 原假设已成立.

$p=0.7$

$$\Rightarrow \hat{p} \sim N\left(0.7, \sqrt{\frac{0.7\times 0.3}{50}}\right)$$

$\hat{p}=0.47$ $\hat{p}<0.47$

$$\text{p-value} = P(\hat{p} \le 0.47) = \text{normalcdf}\left(\begin{array}{l} \text{lower}: 0 \\ \text{upper}: 0.47 \\ \mu: 0.7 \\ \sigma: \sqrt{\frac{0.7 \times 0.3}{50}} \end{array}\right)$$

$$\approx 0$$

$\text{p-value} < 0.05 \Rightarrow$ 拒绝原假设.

$\Rightarrow H_A$ 成立. $p < 0.7$

①定假设. ②验条件 ③算p-值 ④出结论

State: ①假设是什么.

② 使用的数学符号代表什么变量.

③ 用的α-level

Plan: ① 写明使用的方法.

1-sample z-test for p

② 验条件.

1) SRS

2) 10% $n < 10\% N$

★ 3) $np \geq 10$ $n(1-p) \geq 10$

这里是p 不是p̂.

Do. 1) 用计算器

by using calculator

1-PropZ test

P_0: 0.7

x: 23

n: 50

prop: $< P_0$

We get p-value = 0.0001 ≈ 0.

Z-statistics ≈ −3.70

Conclude: Because our p-value (≈ 0) is less than α-level (0.05). We do have convincing evidence that the true proportion of female student among all students in our school is less than 0.7

α-level: 自己设定的标准，小于α则拒绝H_0