Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#223 Approximations (LLN & CLT)

été 2020

Central Limit Theorem

Suppose $X_1, X_2, \cdots$ independent, with mean μ and variance σ^2.
Let $S_n = X_1 + X_2 + \cdots + X_n$, then $\mathbb{E}(S_n) = n\mu$ and $\text{Var}(S_n) = n\sigma^2$.
Heuristic CLT: for reasonably large n,

$$S_n \approx \mathcal{N}(n\mu, n\sigma^2), \text{ or } \sqrt{n}(\overline{X} - \mu) \approx \mathcal{N}(0, \sigma^2)$$

Application: a coin with bias p is tossed n times. Let S_n denote the number of *heads*,

$$S_n \approx \mathcal{N}(np, np(1-p))$$

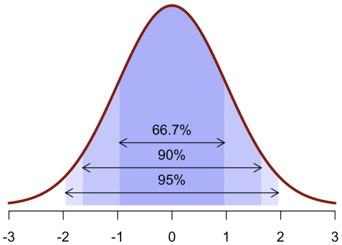
let p_n denote the proportion of *heads*, then

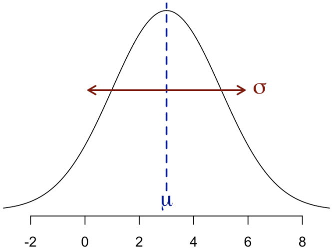
$$p_n \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \text{ or } \sqrt{n}(p_n - p) \approx \mathcal{N}\left(0, p(1-p)\right)$$

Gaussian Distribution

$$Y \text{ has a } \mathcal{N}(0,1) \text{ distribution, } f(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right),$$

$$Y \text{ has a } \mathcal{N}(\mu,\sigma^2) \text{ distribution, } f(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

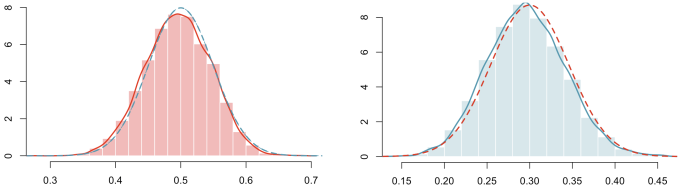




Proportion

Algorithm 1: Gaussian approximation

initialization : $p \leftarrow 1/2$;
for *i=1,2,...,m* **do**
    $\boldsymbol{x} \leftarrow$ sample $\{0,1\}$ with probability $\{1-p,p\}$ n times;
    $f[i] \leftarrow \overline{x}$



Henre with $n = 100$ (and $m = 10,000$)

Proportion

Consider a university with $25,000$ registered student. 400 students are chosen at random, 217 are living with their parents. Estimate the fraction of students living with their parents

$$\widehat{p} = \frac{217}{400} = 54.25\%$$

Give a confidence interval

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

since we don't know the standard deviation $\sqrt{p(1-p)}$,
... use $\sqrt{\widehat{p}(1-\widehat{p})}$

Proportion

Over five years, no student got caught cheating in a course.
Estimate the yearly probability to have a student cheating

$$\widehat{p} = \frac{0}{5} = 0.00\%$$

Give a 95% confidence interval

(...) use a Bayesian approach (#231)