The central limit theorem (part 1)

Lecture 10b (STAT 24400 F24)

Sample sums & sample means

$X_1, X_2, \ldots$ are i.i.d. from some distrib. with mean μ and variance σ^2.

We will study the distributions of:

- The sample sum $S_n = X_1 + \cdots + X_n$
- The sample mean $\bar{X} = \frac{S_n}{n}$

$$\mathbb{E}(S_n) = \sum_{i=1}^{n} \mathbb{E}(X_i) = n\mu, \quad \text{Var}(S_n) = \sum_{i=1}^{n} \text{Var}(X_i) = n\sigma^2$$

$$\mathbb{E}(\bar{X}) = \frac{\mathbb{E}(S_n)}{n} = \mu, \quad \text{Var}(\bar{X}) = \frac{\text{Var}(S_n)}{n^2} = \frac{\sigma^2}{n}$$

Central limit theorem

e.g., the distrib. has finite 3rd moment $\mathbb{E}(|X_i|^3)$

If $X_1, X_2, \ldots$ are i.i.d. from a (reasonable) distribution,
then for sufficiently large n, by the Central Limit Theorem,

e.g. $n > 30$

$$\left(\text{Distribution of } \frac{S_n - n\mu}{\sqrt{n} \cdot \sigma}\right) \approx N(0,1)$$

which implies

$$(\text{Distribution of } S_n) \approx N\left(n\mu, n\sigma^2\right)$$

and

$$(\text{Distribution of } \bar{X}) \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Central limit theorem

More formally: for any fixed $x \in \mathbb{R}$, writing Φ as the CDF of $N(0,1)$, the CLT implies

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{n} \cdot \sigma} \le x\right) = \Phi(x)$$

Notations:

$$Z_n \xrightarrow{n\to\infty} Z \qquad \text{in distribution}$$

or, in CDF's,

$$F_{Z_n}(x) \xrightarrow{n\to\infty} F_Z(x), \qquad \forall x \in \mathbb{R}$$

where

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}}, \qquad Z \sim N(0,1)$$

Standardization & calculating normal probabilities

For $X \sim \mathrm{N}(\mu, \sigma^2)$, any linear transformation of X is normal:

$$aX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2)$$

Standardization means choosing the transformation that will yield $\mathrm{N}(0,1)$:

$$Z = \frac{X - \mu}{\sigma} \sim \mathrm{N}(0,1)$$

Calculate CDF of X:

$$F(x) = \mathbb{P}(X \le x) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \mathbb{P}\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

can look up values of Φ
in textbook/software

CLT for binomial

Let $X \sim \mathrm{Binomial}(n, p)$. We have calculated

$$\mathbb{E}(X) = np, \quad \mathrm{Var}(X) = np(1-p)$$

In fact, X is approximately normal. Why?

- Let $X_i = \mathbb{1}_{\text{success on } i\text{th trial}}$

  $\to$ X_i's are *i.i.d.* with mean $\mu = p$, variance $\sigma^2 = p(1-p)$

- $X = S_n = X_1 + \cdots + X_n$, n large,

$$CLT \quad \Rightarrow \quad X \;\approx\; \mathrm{N}(np, np(1-p))$$

Example: Binomial

Suppose coins are manufactured with a 25% chance of Heads.
What is the distribution of the proportion of Heads after n tosses?

The outcome of a single toss is Bernoulli(0.25)
$\rightsquigarrow$ $\mu = 0.25$, $\sigma^2 = 0.25(1-0.25) = 0.1875$

$$\mathbb{E}(\bar{X}) = \mu = 0.25, \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{0.1875}{n}.$$

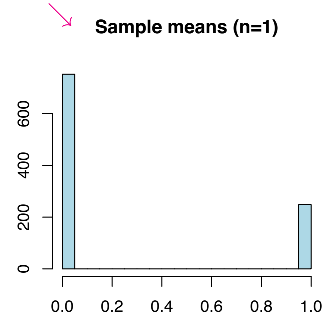
To study the distribution of $\bar{X}$ empirically,
suppose we ask 1000 people to *each* toss the coin n times,
and record the outcomes.

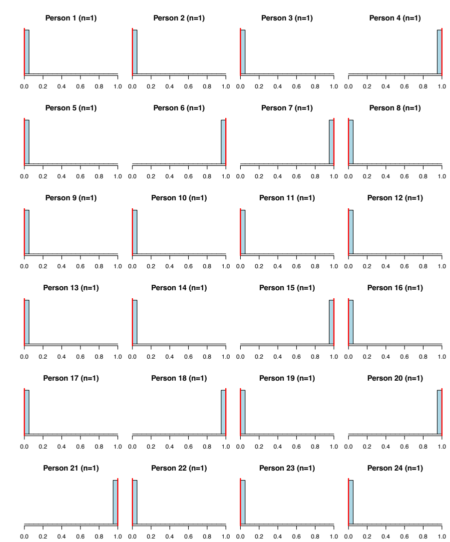
Example: binomial (very small sample size $n = 1$)

Run experiment with $n = 1$:

the individual results (1000 draws of $\bar{X}$)

the empirical distribution of all 1000 results

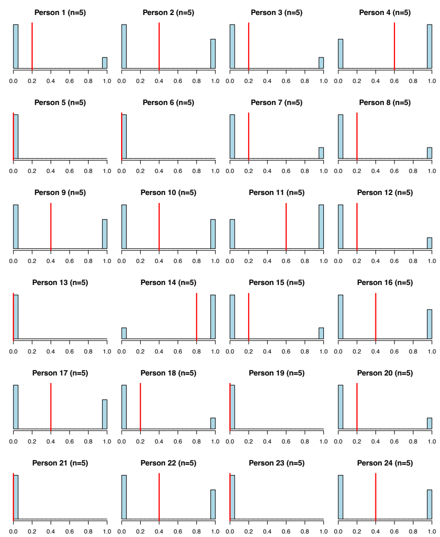




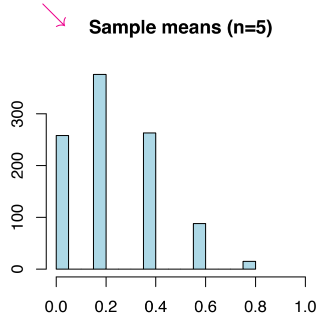
Example: binomial (moderately small sample size $n = 5$)

Run experiment with $n = 5$:

the individual results (1000 draws of $\bar{X}$)



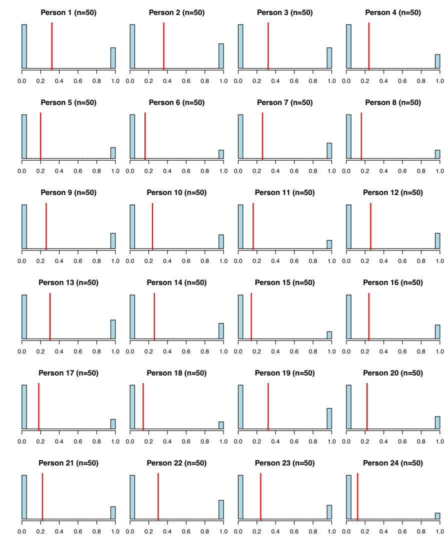
the empirical distribution of all 1000 results



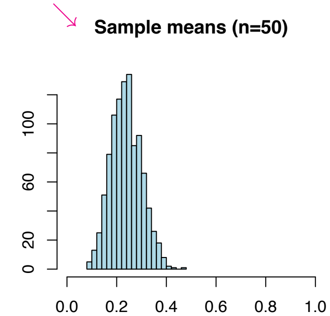
Example: binomial (moderately large sample size $n = 50$)

Run experiment with $n = 50$:

the individual results (1000 draws of $\bar{X}$)



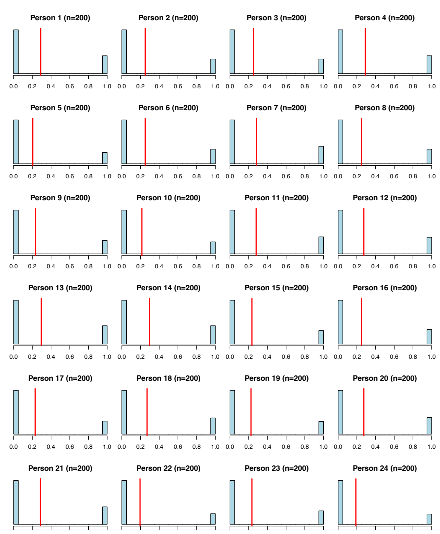
the empirical distribution of all 1000 results



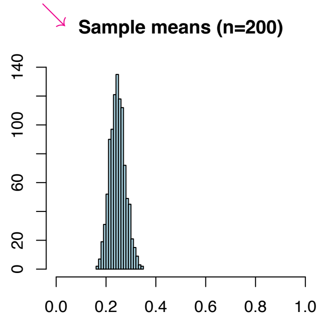
Example: binomial (very large sample size $n = 200$)

Run experiment with $n = 200$:

the individual results (1000 draws of $\bar{X}$)



the empirical distribution of all 1000 results



Example: binomial (CLT for calculating probability)

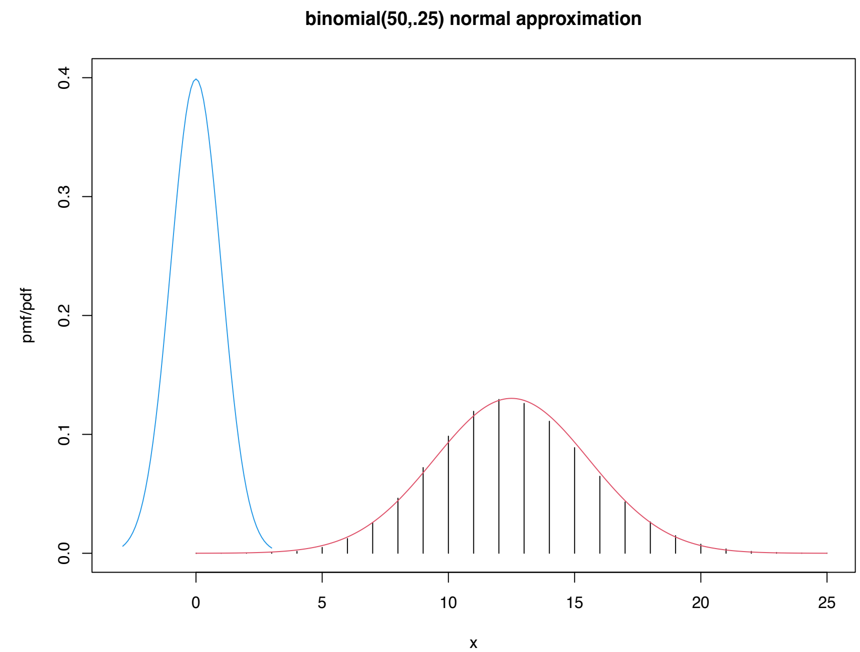
Using the coin that gives 25% chance Heads, if we toss the coin 50 times, what is the probability that we get no more than 10 Heads?

Let X = total # of Heads. $X \sim \text{binomial}(n, p)$, $n = 50, p = 0.25$.

$$\mathbb{E}(X) = np = 12.5, \quad \text{Var}(X) = np(1-p) = 9.375$$

$$\rightsquigarrow \quad X \approx \text{N}(12.5, 9.375) \quad \text{by CLT}$$

$$\mathbb{P}(X \le 10) = \mathbb{P}\left(\frac{X - 12.5}{\sqrt{9.375}} \le \underbrace{\frac{10 - 12.5}{\sqrt{9.375}}}_{=-0.8165}\right) \approx \Phi(-0.8165) = 0.2071$$



Issues in using continuous distribution to approximate discrete distribution

Instead of

$$\mathbb{P}(X \leq 10) = \mathbb{P}\left(\frac{X-12.5}{\sqrt{9.375}} \leq \underbrace{\frac{10-12.5}{\sqrt{9.375}}}_{=-0.8165}\right) \approx \Phi(-0.8165) = 0.2071$$

Alternatively we could compute:

$$\mathbb{P}(X < 11) = \mathbb{P}\left(\frac{X-12.5}{\sqrt{9.375}} < \underbrace{\frac{11-12.5}{\sqrt{9.375}}}_{=-0.4899}\right) \approx \Phi(-0.4899) = 0.3121$$

- We are approximating a discrete distrib. with a continuous distribution. The approximation is coarse when n is not large.
- More accurate if using the middle point: 10.5 instead of 10 or 11 (this is called "continuity correction", not required in this course).

CLT for negative binomial

Let $X \sim \text{NegativeBinomial}(k, p)$

$= $ how many trials to get k successes, if trials are $i.i.d.$ with prob. p

- Let $X_i = \#$ of trials to obtain the ith success, after the $(i-1)$th success was attained.
- The X_i's are also independent, and each had Geometric(p) distribution.

$$\rightsquigarrow \quad \text{mean} \quad \mu = \frac{1}{p}, \qquad \text{variance} \quad \sigma^2 = \frac{1-p}{p^2}$$

- $X = S_k = X_1 + \cdots + X_k$. If k is moderately large,

$$\rightsquigarrow \qquad X \approx \mathrm{N}\left(\frac{k}{p}, \frac{k(1-p)}{p^2}\right)$$

Example: gambling

A gambler plays a game where at each round:
Win \$8 with probability 0.1; otherwise, lose \$1.
What is the probability of being ahead after 20 rounds?

Let $X_i =$ winnings on round i, and let $n = 20$.

$$\mathbb{E}(X_i) = 0.1 \cdot 8 + 0.9 \cdot (-1) = -0.1$$

$$\mathbb{E}(X_i^2) = 0.1 \cdot 8^2 + 0.9 \cdot (-1)^2 = 7.3 \quad \Rightarrow \quad \mathrm{Var}(X_i) = 7.3 - (-0.1)^2 = 7.29$$

$$\text{Winnings from 20 rounds} = S_n \underset{\substack{\uparrow \\ \text{by CLT}}}{\approx} \mathrm{N}(20 \cdot (-0.1), 20 \cdot 7.29) = \mathrm{N}(-2, 145.8)$$

$$\mathbb{P}\begin{pmatrix}\text{gambler is ahead} \\ \text{after 20 rounds}\end{pmatrix} = \mathbb{P}(S_n > 0) = \mathbb{P}\left(\underbrace{\frac{S_n-(-2)}{\sqrt{145.8}}}_{\approx \mathrm{N}(0,1)} > \underbrace{\frac{0-(-2)}{\sqrt{145.8}}}_{=0.1656}\right)$$

$$\approx 1 - \Phi(0.1656) = 0.4342$$