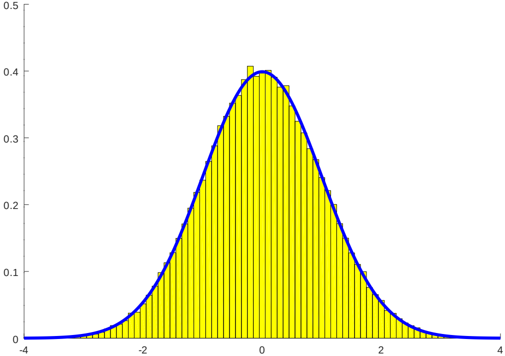
Central Limit Theorem
18.05 Spring 2022



Agenda

- Central limit theorem (CLT) (ubiquitous and important)
- Start class 7: joint distributions

Standard deviation of an average

(Board question from yesterday)

X_1, X_2, ..., X_n independent, identically distributed (i.i.d.) random variables.

All with mean μ and standard deviation σ.

Average $= \overline{X} = \dfrac{X_1 + \ldots + X_n}{n}$

$$E[\overline{X}] = \mu \;\text{ and }\; \text{standard deviation of } \overline{X} = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Reason: This is a calculation using the algebraic properties of mean and variance.

Key conclusion: The average is a better estimate of μ than any single measurement.

Standardization

Random variable X with mean μ, standard deviation σ.

Standardization: $Z = \dfrac{X - \mu}{\sigma}$.

- Z has mean 0 and standard deviation 1.
- Z is dimensionless.
- Standardizing any normal random variable produces the standard normal.
- I.e. if $X \approx$ normal then standardized $Z \approx$ standard normal.

Table question: Standardization

Suppose X is a random variable with mean μ and standard deviation σ. Let Z be the standardization of X.

(a) Give the formula for Z in terms of X, μ and σ.

(b) Use the algebraic properties of mean and variance to show Z has mean 0 and standard deviation 1.

Central Limit Theorem

Setting: X_1, X_2, ...i.i.d. with mean μ and standard dev. σ.

For each n:

$$S_n = X_1 + X_2 + \ldots + X_n \qquad \text{sum}$$

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = \frac{S_n}{n} \qquad \text{average}$$

Know:

$$E[S_n] = n\mu, \qquad \text{Var}(S_n) = n\sigma^2, \qquad \sigma_{S_n} = \sqrt{n}\,\sigma$$

$$E[\overline{X}_n] = \mu, \qquad \text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}, \qquad \sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}}.$$

Standardized sum and average: $Z_n = \dfrac{S_n - n\mu}{\sigma\sqrt{n}} = \dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$

Central Limit Theorem: For large n:

$$\overline{X}_n \approx \text{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad S_n \approx \text{N}\left(n\mu, n\sigma^2\right) \qquad Z_n \approx \text{N}(0, 1)$$

Central Limit Theorem

X_1, X_2, ...i.i.d. with mean μ and standard dev. σ.

Central Limit Theorem: For large n:

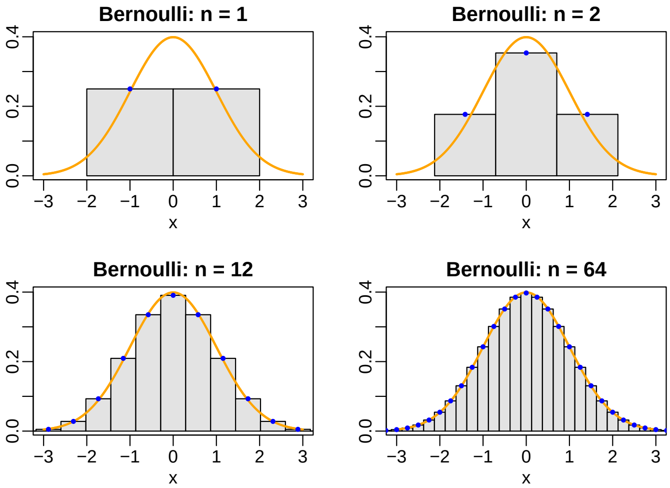
$$\overline{X}_n \approx \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad S_n \approx \mathsf{N}\left(n\mu, n\sigma^2\right) \quad Z_n \approx \mathsf{N}(0,1)$$

In words:

- $\overline{X}_n$ is approximately normal: same mean as X_i but a smaller variance.
- S_n is approximately normal.
- Standardized $\overline{X}_n$ and S_n are approximately standard normal.

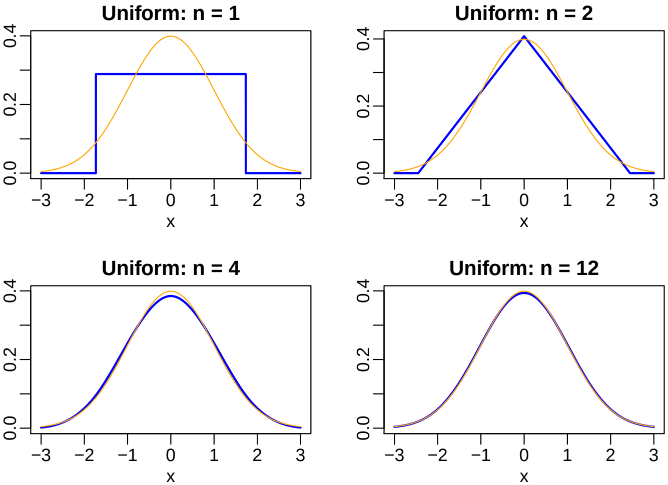
CLT: Pictures 1

The standardized average of n i.i.d. Bernoulli(0.5) random variables with $n = 1, 2, 12, 64$. (See class 6 reading for a description of how these are made.)



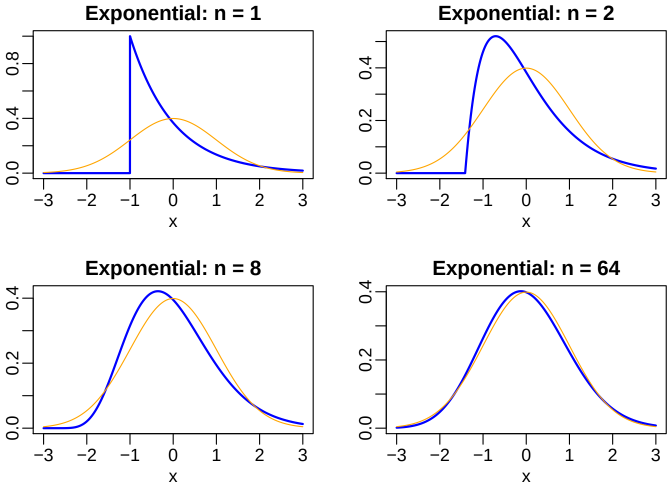
CLT: Pictures 2

Standardized average of n i.i.d. uniform random variables with $n = 1, 2, 4, 12$.



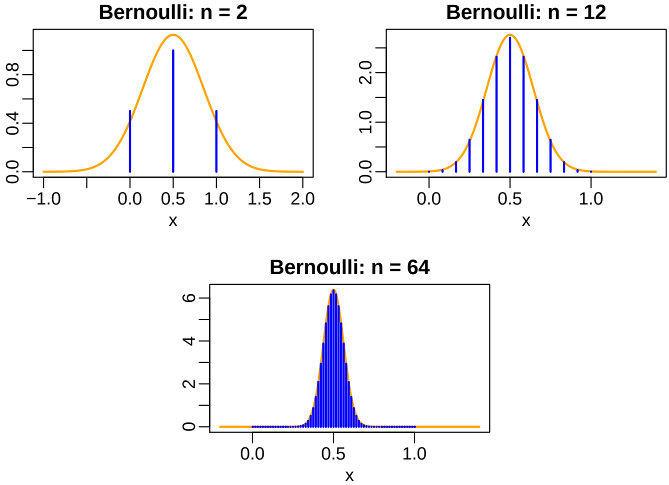
CLT: Pictures 3

The standardized average of n i.i.d. exponential random variables with $n = 1, 2, 8, 64$.



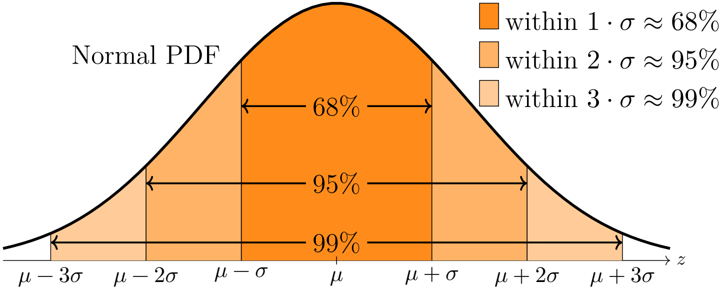
CLT: Pictures 4

The (non-standardized) average of n Bernoulli(0.5) random variables, with $n = 4, 12, 64$. (Spikier.)



Concept Question: Normal Distributions

X has normal distribution, standard deviation σ.

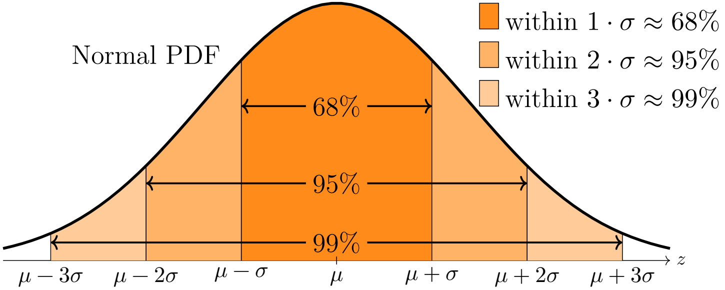


(a) $P(-\sigma < X - \mu < \sigma)$ is approximately

(i) 0.025 (ii) 0.16 (iii) 0.68 (iv) 0.84 (v) 0.95

Concept Question: Normal Distributions

X has normal distribution, standard deviation σ.



(a) $P(-\sigma < X - \mu < \sigma)$ is approximately
(i) 0.025 (ii) 0.16 (iii) 0.68 (iv) 0.84 (v) 0.95

(b) $P(X > \mu + 2\sigma)$ is approximately
(i) 0.025 (ii) 0.16 (iii) 0.68 (iv) 0.84 (v) 0.95

Board Question: CLT

(a) Carefully write the statement of the central limit theorem.

(b) To head the newly formed US Dept. of Statistics, suppose that 50% of the population supports the team of Alessandre, Gabriel, Sarah and So Hee, 25% support Jen and 25% support Jerry.

A poll asks 400 random people who they support. What is the probability that at least 55% of those polled prefer the team?

(c) What is the probability that less than 20% of those polled prefer Jen?

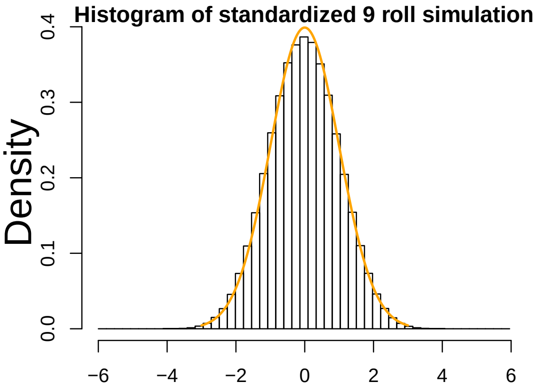
Table Question: Sampling from the standard normal distribution

How would you approximate a single random sample from a standard normal distribution using 9 rolls of a ten-sided die?

Note: $\mu = 5.5$ and $\sigma^2 = 8.25$ for a single roll of a 10-sided die.

Hint: CLT is about averages.

Histogram of 9 roll simulation



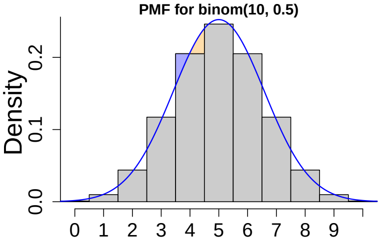
Standard normal is shown in orange.

$\overline{X}$ = average of nine rolls: $\mu = 5.5$, $\sigma = \sqrt{8.25/9}$.

Standarized statistic: $Z = \dfrac{\overline{X} - \mu}{\sigma} \approx N(0, 1)$.

Continuity correction

Approximating a discrete distribution with a continuous is ambiguous.



Here $X \sim$ binom$(10, 0.5)$. $\mu_X = 5$, $\sigma_X = \sqrt{10/4}$.

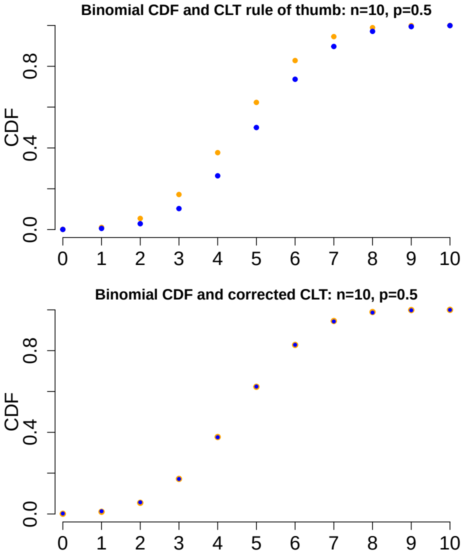
The CLT for $Y \sim \mathsf{N}(\mu_X, \sigma_X^2)$, $X \approx Y$.

We know $P(X \leq 4) = P(X \leq 4.5) = P(X \leq 4.9999)$. Should we approximate this with $F_Y(4)$, $F_Y(4.5)$, $F_Y(4.9999)$?

Rule of thumb: Use $F_Y(4)$ – convenient, often easy to compute.

Continuity correction: Use $F_Y(4.5)$ – more accurate.

Comparing rule of thumb and continuity correction



Bonus problem

Not for class. Solution will be posted with solutions for today.

An accountant rounds to the nearest dollar. We'll assume the error in rounding is uniform on [-0.5, 0.5]. Estimate the probability that the total error in 300 entries is more than \$5.