STAT 23400 Lecture 10: Central Limit Theorem

Review: Expected value, Covariance, and Correlations

- Expected Value for $g(X, Y)$:

$$E[g(X,Y)] = \begin{cases} \sum_{x,y} g(x,y) f(x,y) & \text{in discrete case ,} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y) dxdy & \text{in continuous case.} \end{cases}$$

- Expected value is linear: $E(aX + bY) = aE(X) + bE(Y)$
- If X and Y are independent, then $E(g(X)h(Y)) = E(g(X))E(h(Y))$.
- Covariance of X and Y (with means μ_X and μ_Y),

$$Cov(X; Y) = E((X - \mu_X)(Y - \mu_Y)) = E(XY) - \mu_X \mu_Y$$

- Correlation coefficient

$$\rho_{XY} = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Review: Linear Combinations of Random Variables

- The expected value and variance for linear combinations of random variables are

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = a_1 E(X_1) + a_2 E(X_2) + \cdots + a_n E(X_n)\,.$$

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i) + \sum_{i \neq j} a_i a_j Cov(X_i, X_j)$$

- When $X_1, \ldots, X_n$ are independent, we have

$$Var(X_1 + X_2 + \ldots + X_n) = Var(X_1) + Var(X_2) + \ldots + Var(X_n).$$

- For $X \sim \text{Binom}(n, p)$, we proved that

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

Outline for Today

- Topics:
  - Central Limit Theorem (Section 6.1-6.2 in MMSA).
  - Normal Approximation to Binomial Distribution (Section 6.2 in MMSA).
  - CLT in Exponential and bimodal distributions.

Linear Combinations of Normal Random Variables

Suppose $X_1, \ldots, X_n$ are joint normal r.v.'s with

$$X_i \sim N(\mu_i, \sigma_i^2) \qquad \text{for } i = 1, 2, \ldots, n$$
$$Cov(X_i, X_j) = \sigma_{ij} \qquad \text{for } 1 \leq i \neq j \leq n$$

Let

$$X = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n \,.$$

X has a **normal** distribution with mean

$$E(X) = \sum_{i=1}^{n} a_i E(X_i) = a_1 \mu_1 + a_2 \mu_2 + \ldots + a_n \mu_n$$

and variance

$$Var(X) = \text{Var}(\sum_{i=1}^{n} a_i X_i) = \sum\nolimits_{i=1}^{n} a_i^2 \sigma_i^2 + \sum\nolimits_{i \neq j} a_i a_j \sigma_{ij}$$

Example: Linear Combination of Normal R.V.

Example. Let X_1, X_2, X_3 be i.i.d normal r.v.'s with mean μ and variances σ^2. What is the distribution of $Y = X_1 - X_2 + X_3$?

Central Limit Theorem

Properties of the Mean of i.i.d. RV's: Mean and Variance

Suppose $X_1, \ldots, X_n$ are i.i.d. rv's with mean μ and variance σ^2.

- Consider the sample mean $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$, we have

$$\boxed{E(\overline{X}) = \mu, \qquad \text{Var}\,(\overline{X}) = \frac{\sigma^2}{n}.}$$

Sampling Distribution of the Sample Mean

Note that $\overline{X}$ itself is a r.v. What is the distribution of $\overline{X}$?

- called the **sampling distribution of the sample mean.**
- depends on the population distribution. Here are some example.
  - If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, then $\overline{X} \sim N(\mu, \sigma^2/n)$.
  - If $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, then $n\overline{X} \sim \text{Binom}(n, p)$
  - If $X_1, \ldots, X_n \sim \text{Exponential}(\lambda)$, then $\overline{X} \sim \text{Gamma}(n, n\lambda)$ (See Section 4.4 in MMSA)
  - $\vdots$

Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots$ be a sequence of **i.i.d.** random variables with **mean** μ and **variance** σ^2. Then, when n is large,

- the distribution of the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately
$$N\left(\mu, \frac{\sigma^2}{n}\right).$$
- the distribution of the sum $\sum_{i=1}^{n} X_i$ is approximately
$$N(n\mu, n\sigma^2).$$

Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d. $\sim F$ with $\mu = E[X_i]$ and $\sigma^2 = Var[X_i]$, for $i = 1, \ldots, n$, and

$$S_n = \sum_{i=1} X_i, \qquad \bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}.$$

As $n \to \infty$,

$$T_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\sum_{i=1}^n (X_i - \mu)}{\sqrt{n}\sigma} \to N(0; 1),$$

and

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to N(0; 1).$$

Proof idea

Since mgf characterize the corresponding distribution completely, it suffices to show that, as $n \to \infty$, the mgf of T_n converges to the mgf of $Z \sim N(0, 1)$. In other words, we need

$$m_{T_n}(t) \to m_Z(t)$$

for all t.

This can be done using Taylor expansion.

Example 1: Card Game

- Recall the card game: draw ONE card from a well-shuffled deck of cards and may get a reward based on the card drawn as follows.

Event	reward X	$p(x)$
Heart (not ace)	\$1	12/52
Ace	\$5	4/52
King of spades	\$10	1/52
All else	\$0	35/52
Total		1

- The card drawn is placed back to the deck before he draws the card for the next game.
- Let X_i be the reward he get in the ith game, then X_i's are i.i.d. and his total reward from the 300 games is

$$X_1 + X_2 + \cdots + X_n$$

Example 1: Card Game

- Recall that in the card game, we can calculate the mean and variance as

$$\mu = 0.81 \text{ and } \sigma^2 = \text{Var}(X) = \frac{9260}{52^2}\,.$$

- So if a gambler played the game 300 times, his expected value, variance of his total reward is

$$\begin{aligned}
\text{E}(X_1 + \cdots + X_{300}) &= 300\mu = 300 \times 0.81 \approx 243.308 \\
\text{Var}(X_1 + \cdots + X_{300}) &= 300\sigma^2 = 300 \times \frac{9260}{52^2} \\
\text{SD}(X_1 + \cdots + X_{300}) &= \sqrt{300 \times \frac{9260}{52^2}} = 32.052
\end{aligned}$$

- The gambler is expected to get \$243.308 from the 300 games, with a standard deviation \$32.052.

Example 1: Card Game

Q: What is the probability that the gambler can earn \$250 or more from the 300 games?

Normal Approximation to Binomial Distribution

Normal Approximation to Binomial Distribution

Normal approximation to a binomial random variable is a special case of CLT:

$$Y = \sum_{i=1}^{n} X_i \sim \text{Binom}(n, p),$$

where $X_1, X_2, ..., X_n$ are n **independent Bernoulli** random variables with parameter p.
Therefore,

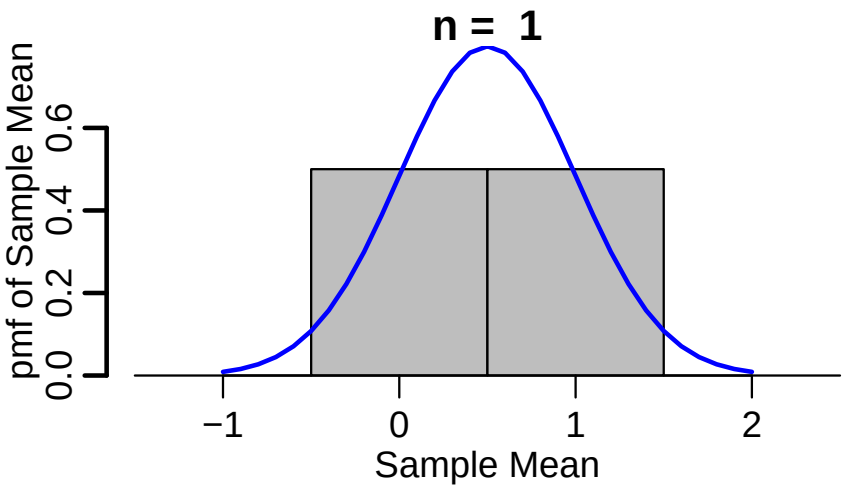
$$E(X_i) = p, \qquad Var(X_i) = p(1-p).$$

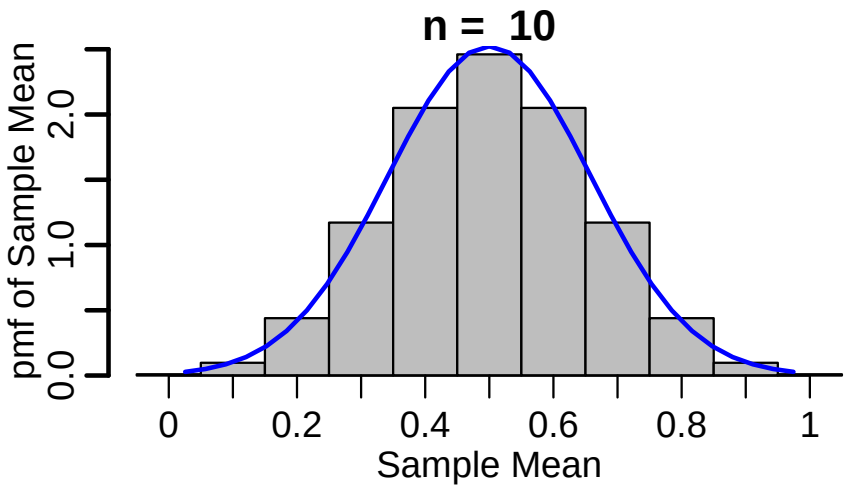
By CLT, for large n,

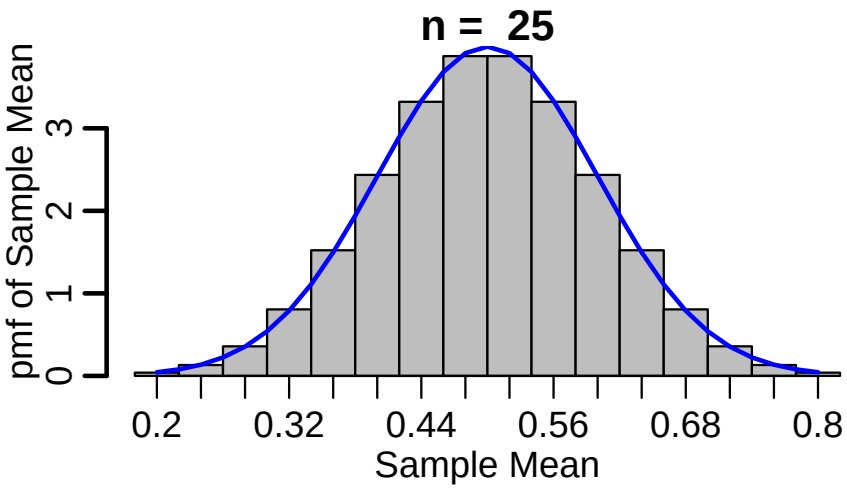
$Y \sim \text{Binom}(n, p)$ is approximately distributed as $N(np, np(1-p))$.

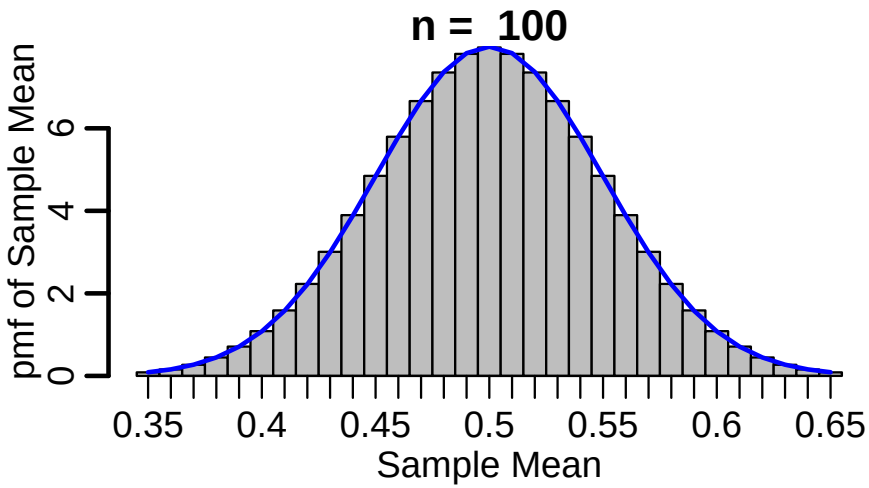
Normal Approximation to $Binom(n, p = 0.5)$

$X_1, \ldots, X_n \sim$ Bernoulli$(p = 0.5)$, the sampling distribution of $\overline{X}$ is



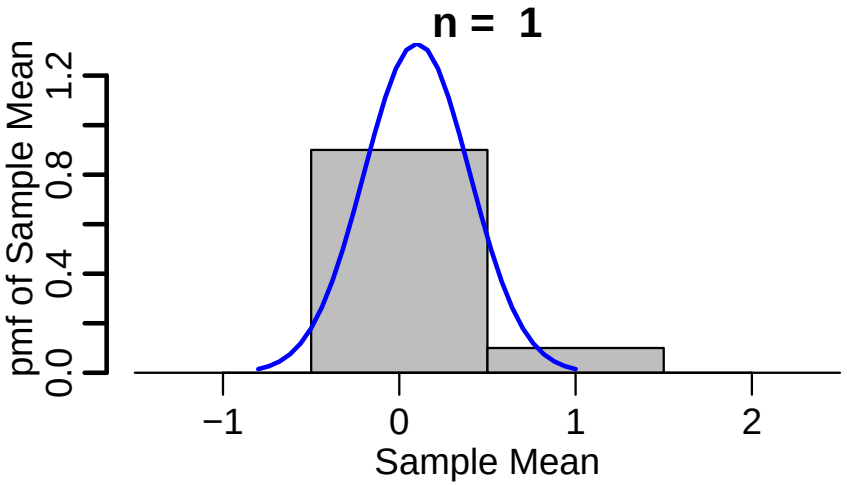


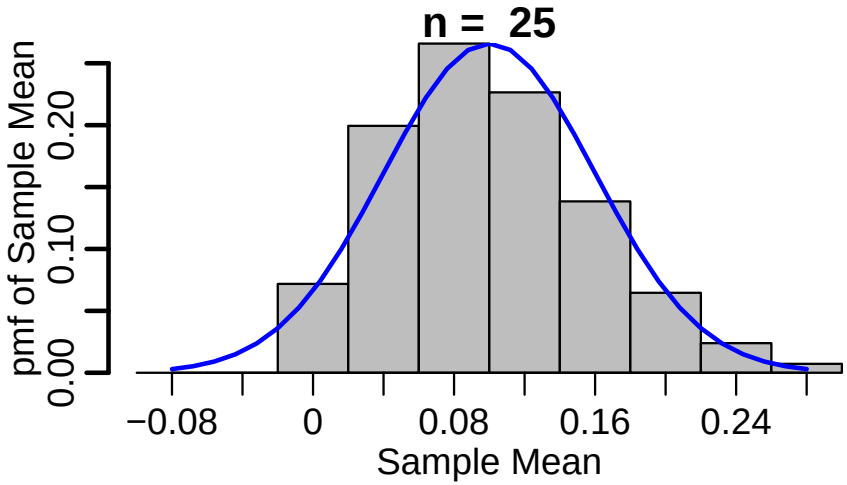


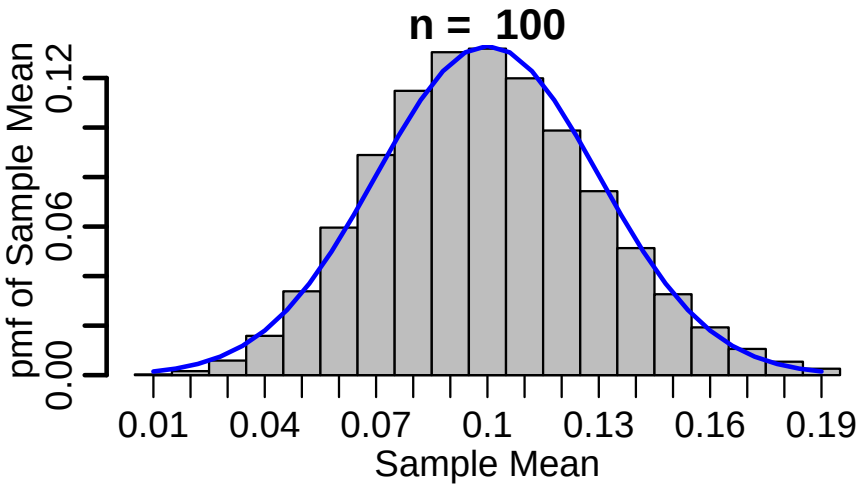


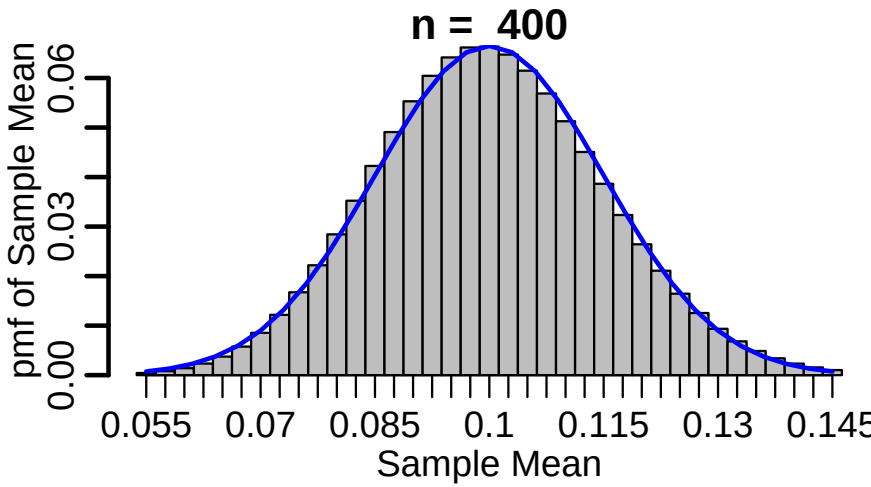
Normal Approximation to $Binom(n, p = 0.1)$

$X_1, \ldots, X_n \sim$ Bernoulli$(p = 0.1)$, the sampling distribution of $\overline{X}$ is









Note that the skewness of the distribution of the sample mean diminishes as n increases.

Example 3: Roulette

With a perfectly balanced roulette wheel, red numbers should turn up 18 in 38 of the time. To test its wheel, one casino records the results of 3800 plays. Let X be the number of reds the casino got.

Q1: If the roulette wheel is perfectly balanced, what is the chance that $X \geq 1890$?

Q2 If the casino gets 1890 reds, do you think the roulette wheel should be calibrated?

Example 3: Roulette

Q1: If the roulette wheel is perfectly balanced, what is the chance that $X \geq 1890$?

Q2 If the casino gets 1890 reds, do you think the roulette wheel should be calibrated?

How Large n Has to Be to Use CLT?

- If the population is normal, then any n will do.
- If the population distribution is symmetric, then n should be at least 30 or so.
- The more skew or irregular the population, the larger n has to be
- For the Binomial distribution, a rule of thumb is that n should be such that

$$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10.$$

CLT for Exponential Distribution

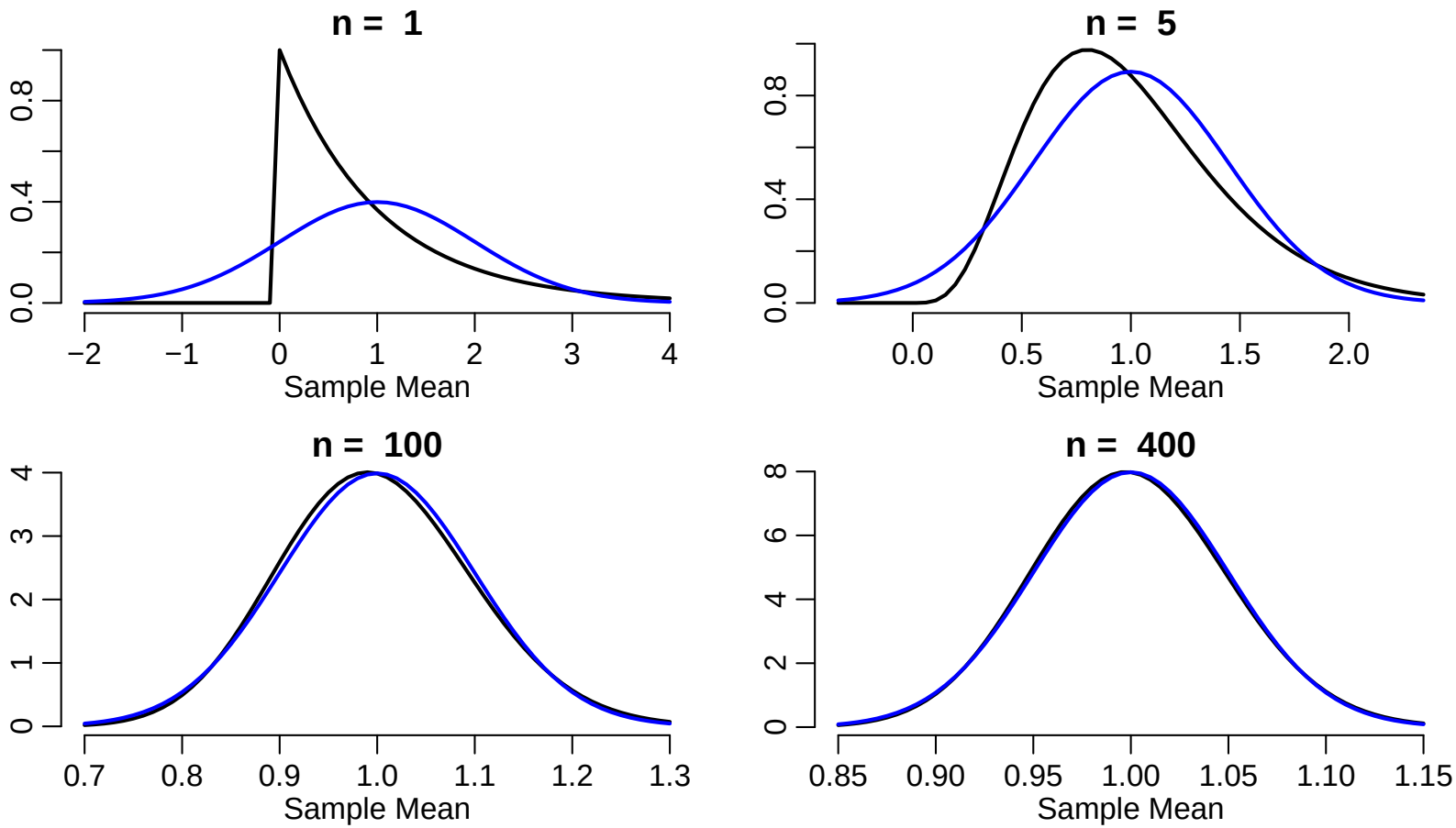
If the population distribution is exponential with density

$$f(x) = e^{-x}, \quad \text{for } x > 0, \quad \mu = 1,\ \sigma^2 = 1$$

CLT for Exponential Distribution

black curve: the exact sampling distribution of $\overline{X}$,

blue curve: the normal approximation



CLT for Bimodal Distribution

If the population distribution is Bimodal with density

$$f(x) = \frac{0.5}{\sqrt{2\pi}(0.1)} \exp\left(-\frac{(x-1)^2}{2(0.1)^2}\right) + \frac{0.5}{\sqrt{2\pi}(0.1)} \exp\left(-\frac{(x+1)^2}{2(0.1)^2}\right)$$

CLT for Bimodal Distribution

Black curve: the exact sampling distribution of $\overline{X}$,

Blue curve: the normal approximation