STAT 201

Week 7

Lecture goals:

By the end of this lecture, the students are expected to be able to:

- Describe the Law of Large Numbers.
- Describe a Normal distribution.
- Explain the Central Limit Theorem and its role in constructing confidence intervals.
- Write a computer script to calculate confidence intervals based on the assumption of normality / the Central Limit Theorem.
- Discuss the potential limitations of these methods.
- Decide whether to use asymptotic theory or bootstrapping to compute estimator uncertainty.

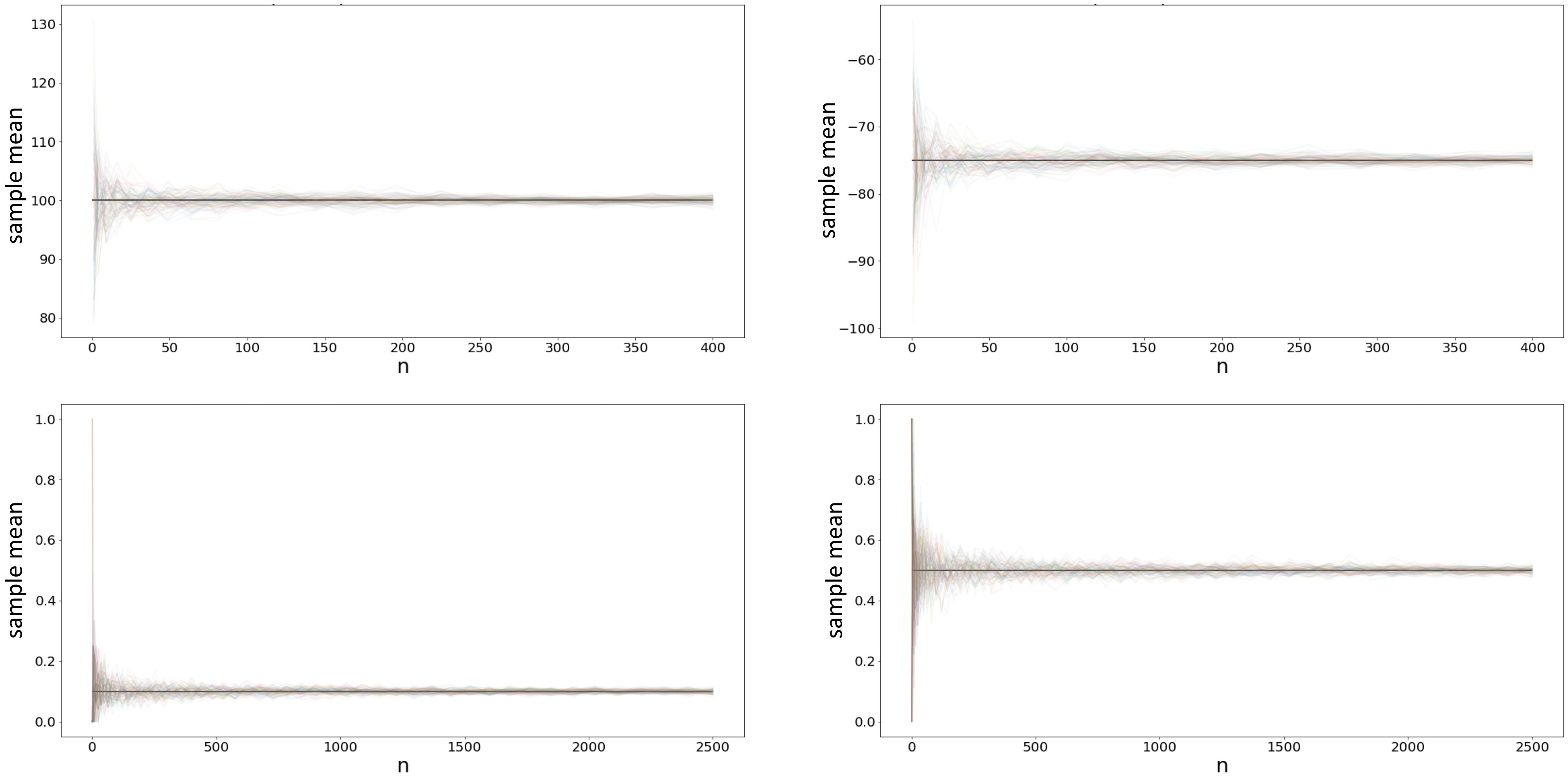
Law of Large Numbers

(Strong) Law of Large Numbers

- The Law of Large Numbers (LLN) states that the sample average converges to the population mean.

- In other words, as the sample size increases, the sample average gets closer and closer to the population mean with higher and higher probability.

(Strong) Law of Large Numbers



Normal Distribution

Gaussian distribution

- The Gaussian (or Normal) distribution is one of the most (if not the most) important distribution in statistics.

- Many of the methods in statistics and data analysis assume Normality. Besides, the Central Limit Theorem assigns a central role for the Gaussian distribution.

- Today, we are going to explore the Normal distribution in more detail.

Gaussian distribution

- Let $X \sim N(\mu, \sigma)$ be a random variable. The density of X is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \qquad x \in \mathbb{R}$$

- Note that the term

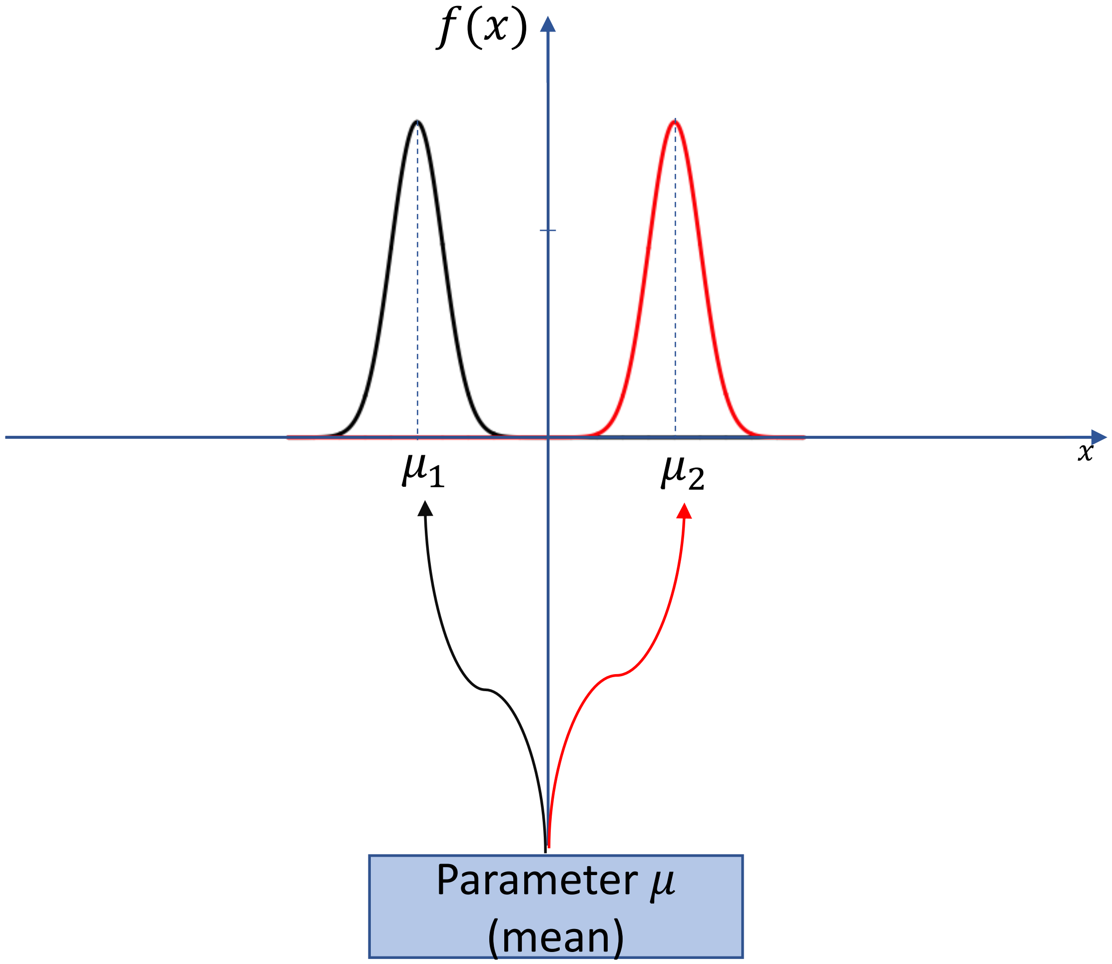
$$\frac{(x-\mu)^2}{\sigma^2}$$

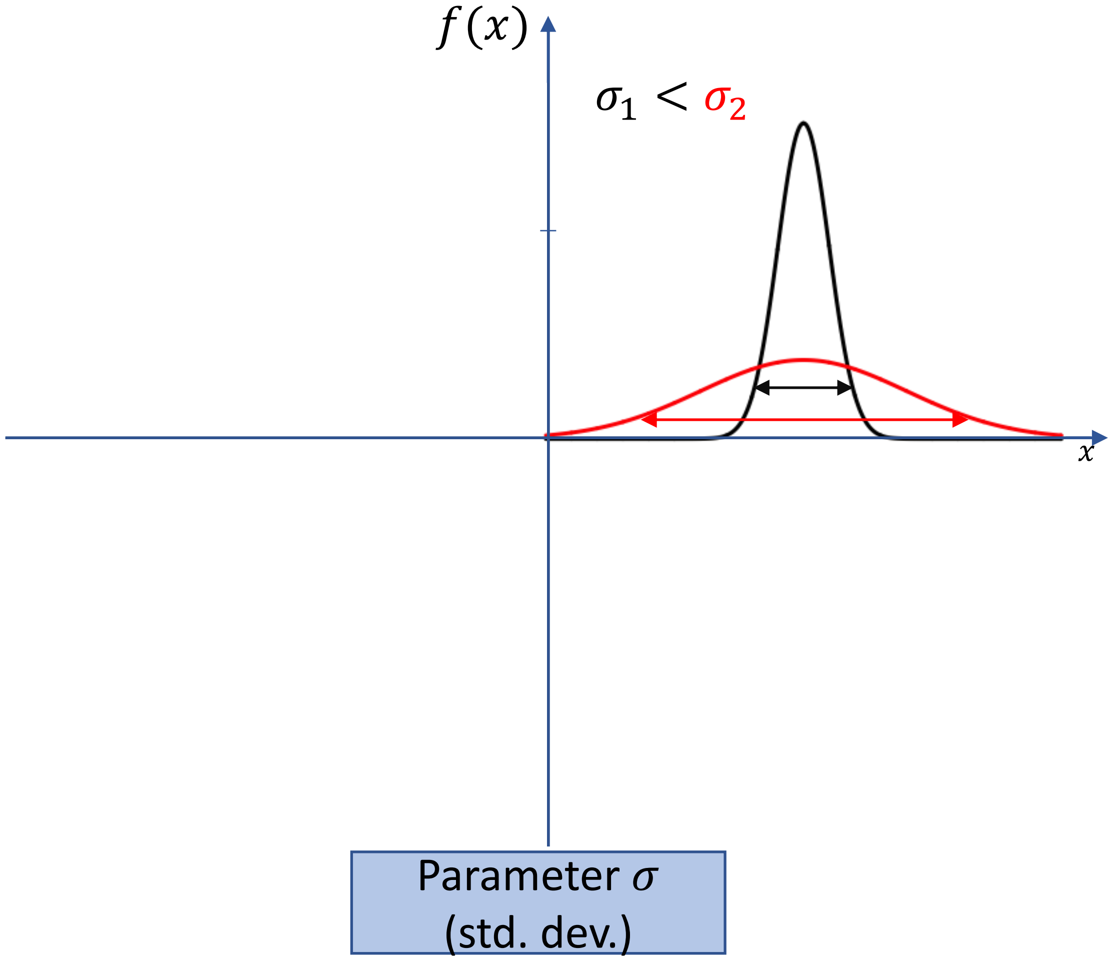
measures the square of the distance between x and μ in terms of standard deviations.

Gaussian distribution

- Unimodal and bell-shaped;
- Symmetric around the mean, μ;
- The standard deviation σ, controls the spread of the curve (wider or narrower);

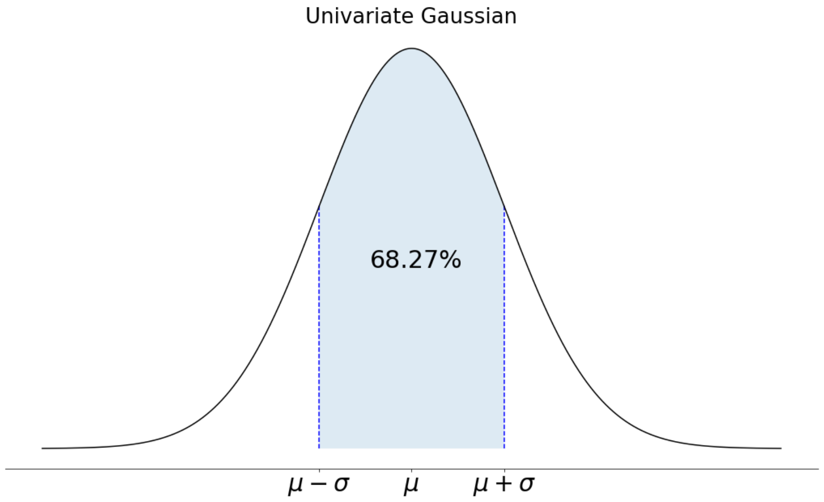
Gaussian distribution

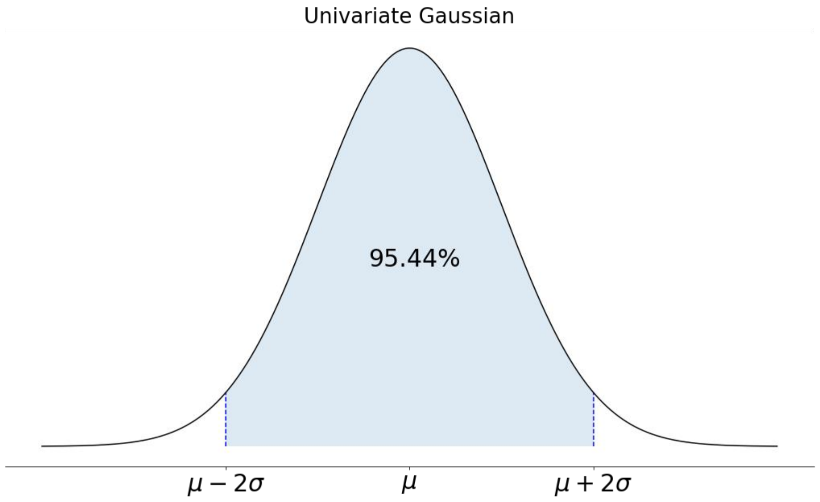


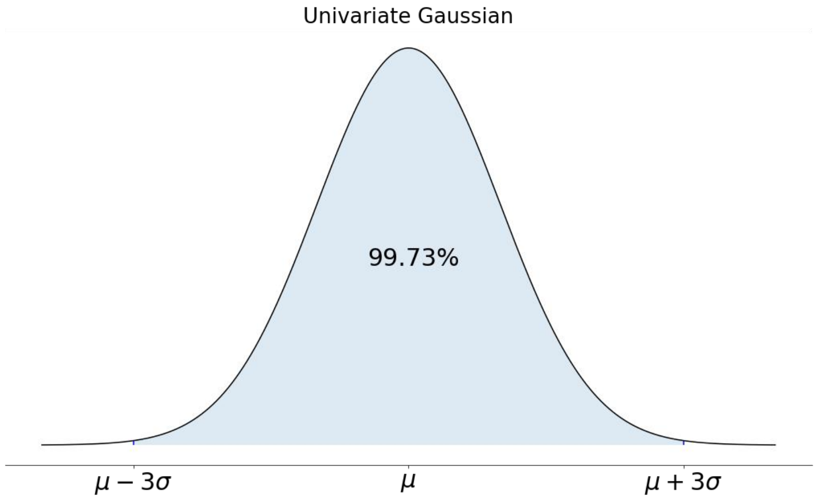


Gaussian distribution

- Regardless of the values of μ and σ we have that:

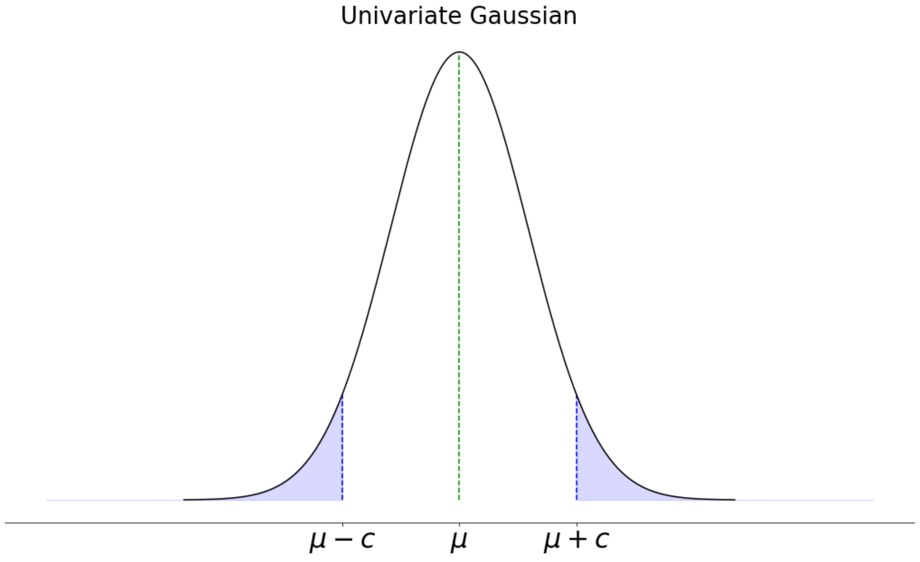






Gaussian distribution

- Let $X \sim N(\mu, \sigma)$ be a random variable. Then,
  - $E[X] = \mu$
  - $Var(X) = \sigma^2$
  - X is symmetric around the mean, which means:
  $$P(X \geq \mu + c) = P(X \leq \mu - c)$$
  for any constant c.



- Also, $\frac{X-\mu}{\sigma} \sim N(0,1)$. The $N(0,1)$ is known as *Standard Normal*.

Gaussian distribution

- Unfortunately, the CDF of the Gaussian distribution does not have a closed-form.

- We need to use software packages to get the desired probability or quantiles.

R:

Probabilities: e.g., $P(X \leq 3)$
```
pnorm(3, μ, σ)
```

Quantiles: e.g., $P(X \leq x) = 0.95$
```
qnorm(0.95, μ, σ)
```

Central Limit Theorem

Central Limit Theorem

- Let $x_1, x_2, \dots, x_n$ be a random sample from a population.
- The CLT states that for large sample sizes (large n) the sampling distribution of the sample mean (or sample proportion) will converge to the Normal distribution.
- $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

Central Limit Theorem: Assumptions

- The central limit theorem makes the following assumptions:
  - The sample is drawn in an independent fashion.

  - In general, if your sample size is greater than 10% of the population size, there will be a severe violation of independence.

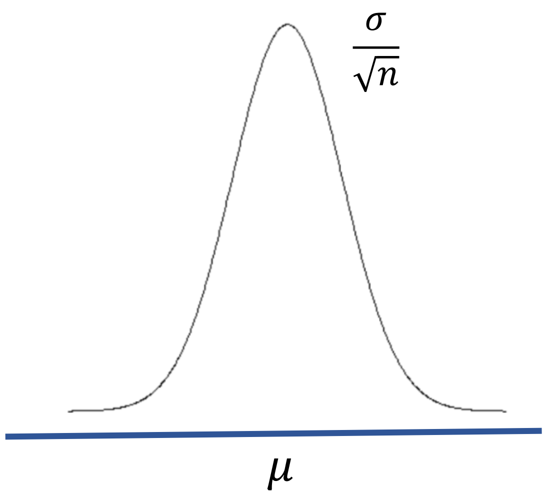
  - The sample size must be large enough.
    - For the proportion, you can check if $n \times p \geq 10$ and $n \times (1-p) \geq 10$.
    - For the sample mean, there is no universal guideline, and we might need a large sample size. Usually, however, sample sizes between 30 and 50 are enough to get a reasonable approximation (but it is not guaranteed).

Confidence Intervals based on CLT

Confidence intervals based on CLT: Mean

- Let $x_1, x_2, \dots, x_n$ be a random sample from a population with mean μ and standard deviation σ.

- Assuming the CLT conditions are satisfied, we have that:

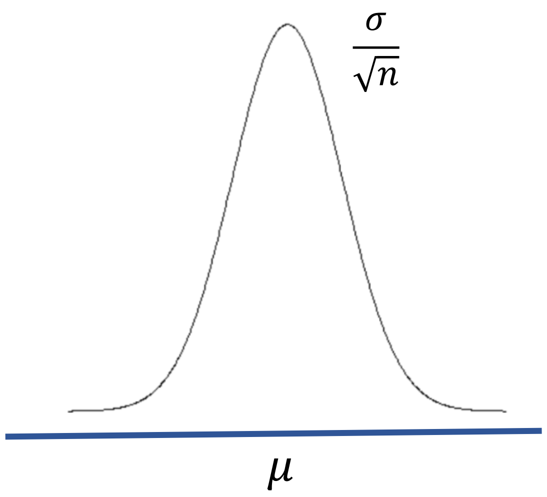


$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Confidence intervals based on CLT: Mean

- Let $x_1, x_2, \dots, x_n$ be a random sample from a population with mean μ and standard deviation σ.

- Assuming the CLT conditions are satisfied, we have that:



$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Therefore,

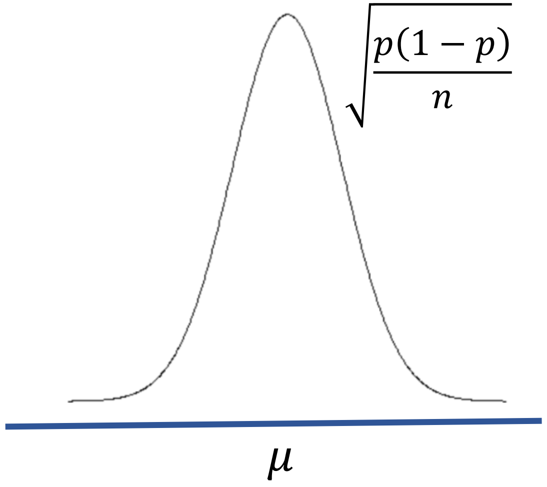
$$IC(\mu, \alpha) = \bar{x} \pm z^*_{1-\alpha} \times \frac{s}{\sqrt{n}}$$

<u>**Note:**</u> we could actually get a better approximation using t-distribution that you are going learn next week.
However, for large n the Normal and t-distributions are quite close. In fact, for $n \geq 50$, both distributions are essentially the same.

Confidence intervals based on CLT: Proportion

- Let $x_1, x_2, \dots, x_n$ be a random sample from a population with proportion p.

- Assuming the CLT conditions are satisfied, we have that:

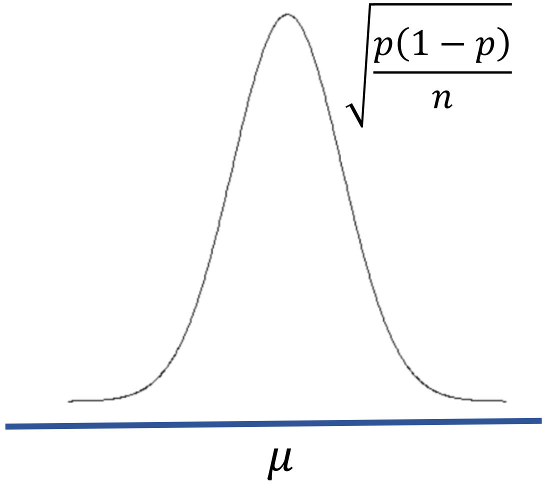


$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Confidence intervals based on CLT: Proportion

- Let $x_1, x_2, \ldots, x_n$ be a random sample from a population with proportion p.

- Assuming the CLT conditions are satisfied, we have that:



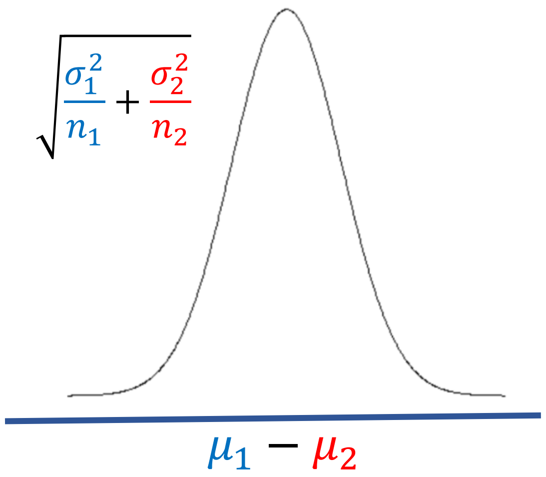
$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Therefore,

$$IC(p, \alpha) = \hat{p} \pm z^*_{1-\alpha} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Confidence intervals based on CLT: Difference in means

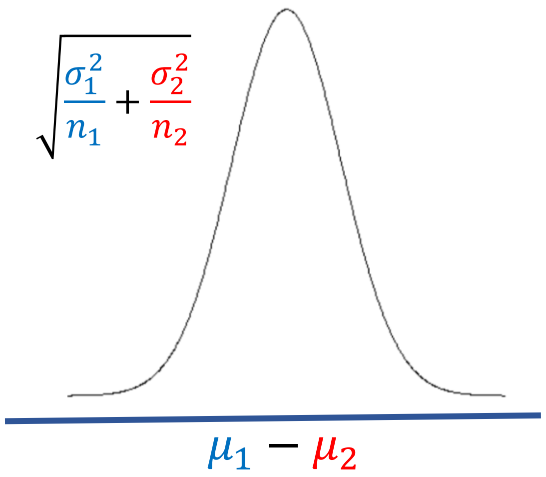
- Let $x_1, x_2, \dots, x_n$ be a random sample from a population with mean μ_1 and standard deviation σ_1.
- Let $y_1, y_2, \dots, y_n$ be a random sample from a population with mean μ_2 and standard deviation σ_2.
- Assuming the CLT conditions are satisfied, we have that:



$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Confidence intervals based on CLT: Difference in means

- Let $x_1, x_2, \dots, x_n$ be a random sample from a population with mean μ_1 and standard deviation σ_1.
- Let $y_1, y_2, \dots, y_n$ be a random sample from a population with mean μ_2 and standard deviation σ_2.
- Assuming the CLT conditions are satisfied, we have that:



$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Therefore,

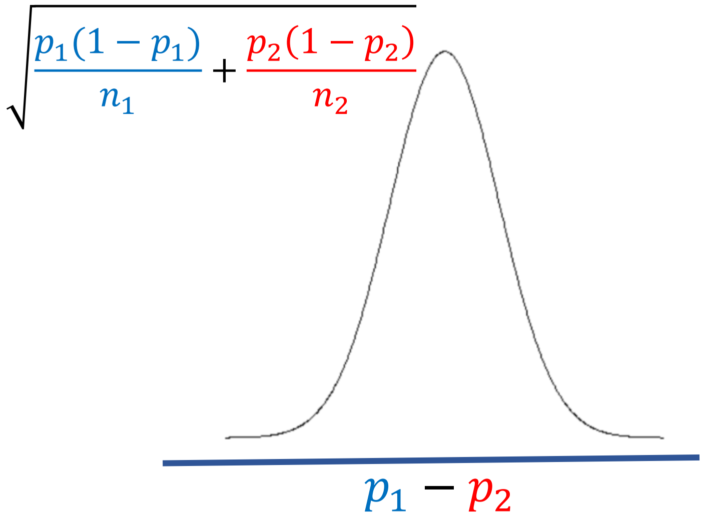
$$IC(\mu, \alpha) = \bar{x} - \bar{y} \pm z^*_{1-\alpha} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

<u>Note:</u> we could actually get a better approximation using t-distribution that you are going learn next week.
However, for large n the Normal and t-distributions are quite close. In fact, for $n \geq 50$, both distributions are essentially the same.

Confidence intervals based on CLT: Difference in proportions

- Let $x_1, x_2, \dots, x_n$ be a random sample from a population with proportion p_1.
- Let $y_1, y_2, \dots, y_n$ be a random sample from a population with proportion p_2.
- Assuming the CLT conditions are satisfied, we have that:

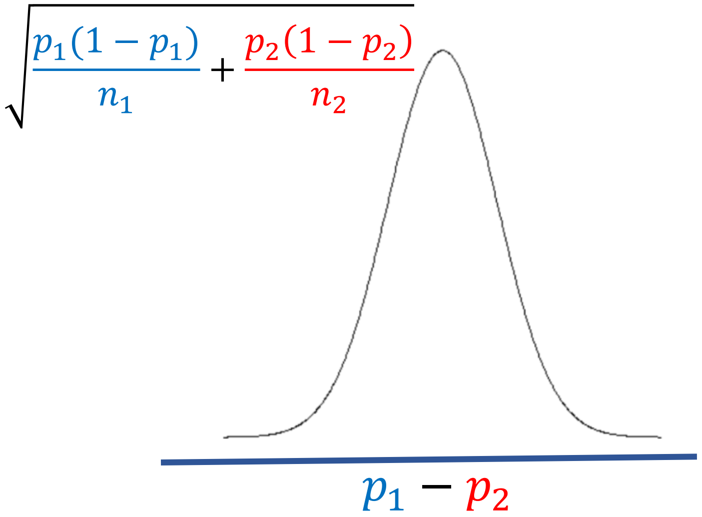
$$\bar{X} - \hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$



Confidence intervals based on CLT: Difference in proportions

- Let $x_1, x_2, \dots, x_n$ be a random sample from a population with proportion p_1.
- Let $y_1, y_2, \dots, y_n$ be a random sample from a population with proportion p_2.
- Assuming the CLT conditions are satisfied, we have that:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$



Therefore,

$$IC(\mu, \alpha) = \hat{p}_1 - \hat{p}_2 \pm z^*_{1-\alpha} \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$