CSC196: Analyzing Data

Sampling Distributions

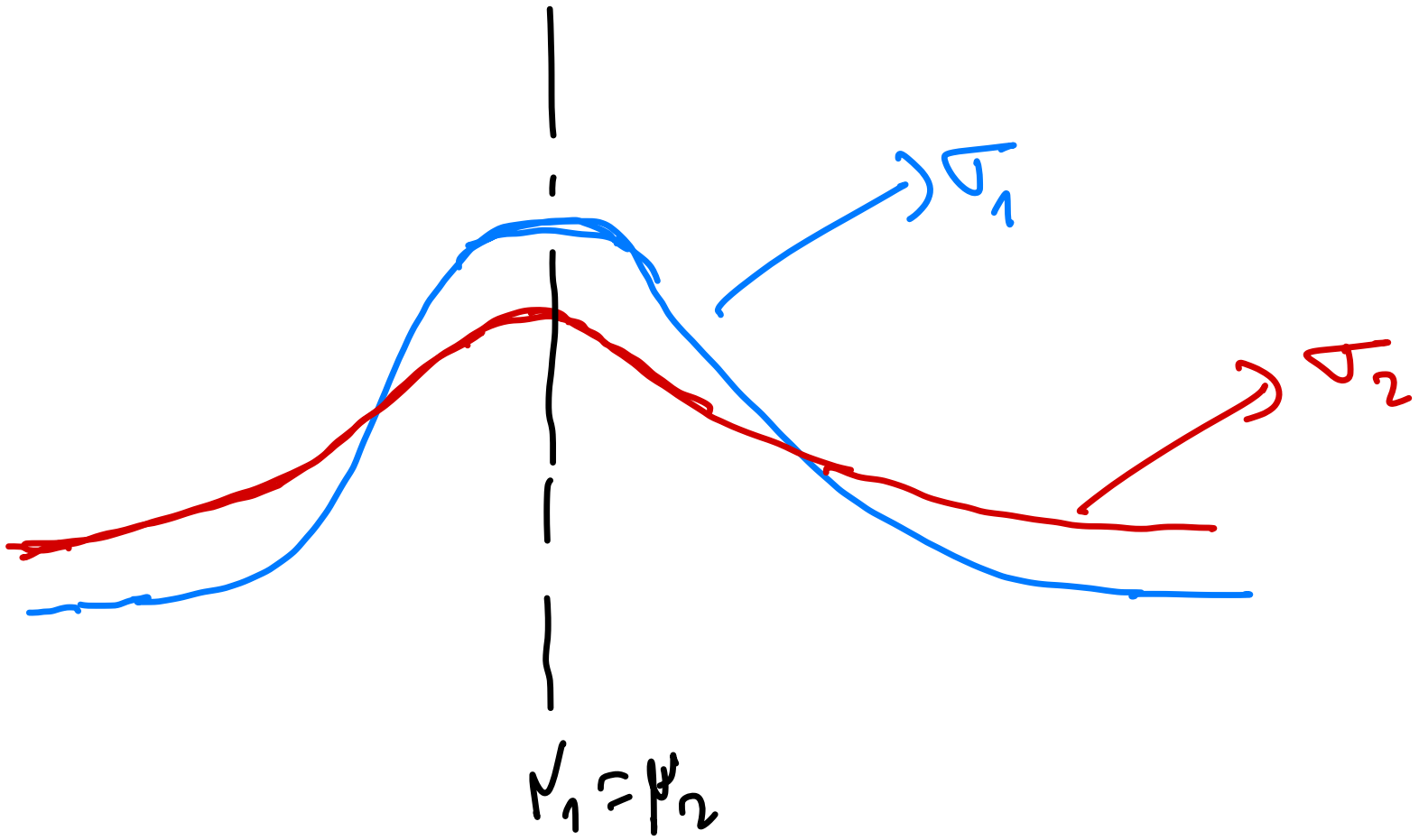
Jason Pacheco and Cesim Erten

QUIZ

Draw two normal distributions with the same mean $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$ on the same plot.

QUIZ

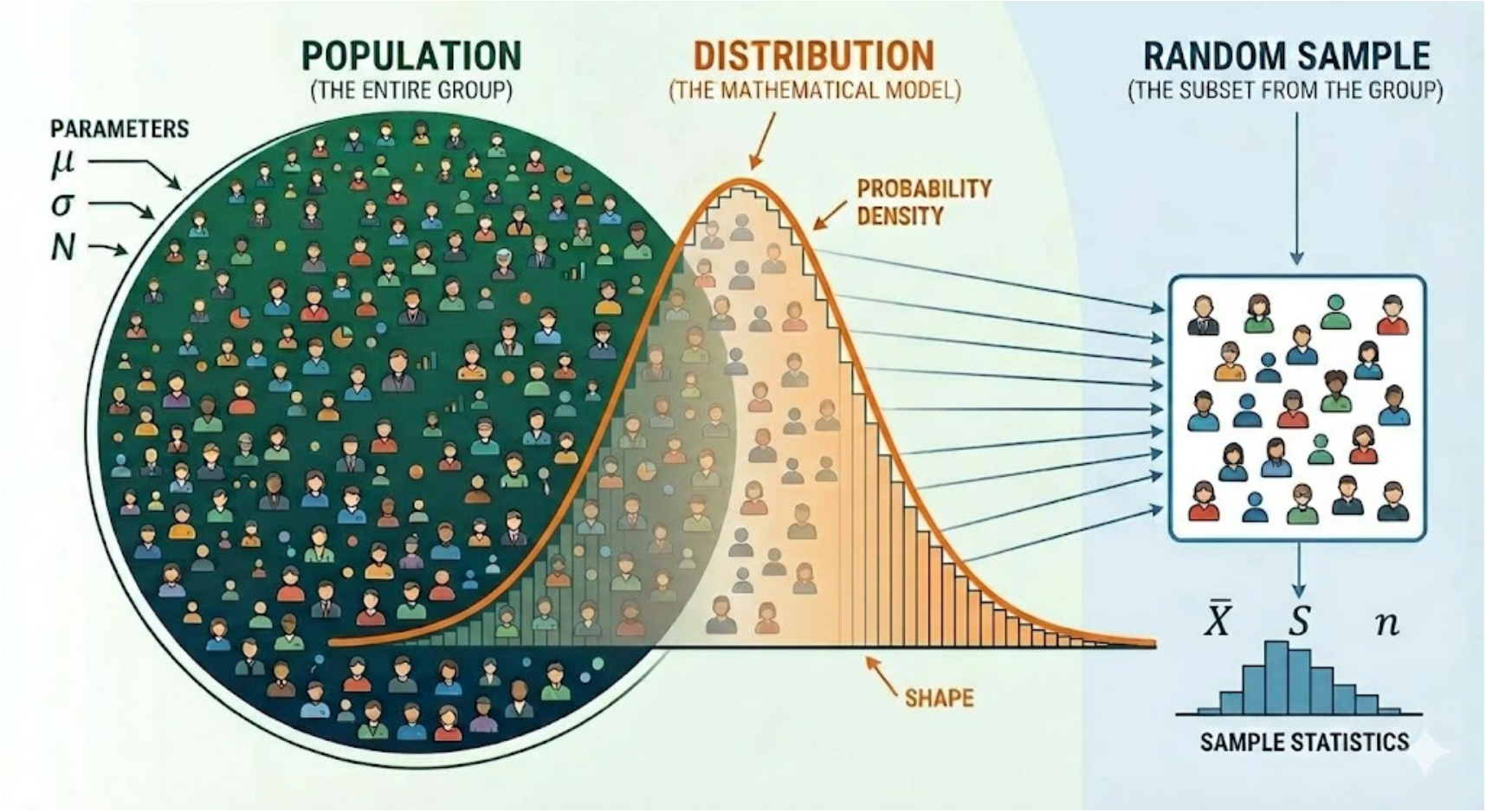
Draw two normal distributions with the same mean $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$ on the same plot.



Outline

- Random Sampling
- Sampling Distributions
- Central Limit Theorem and Sampling Distribution of Means

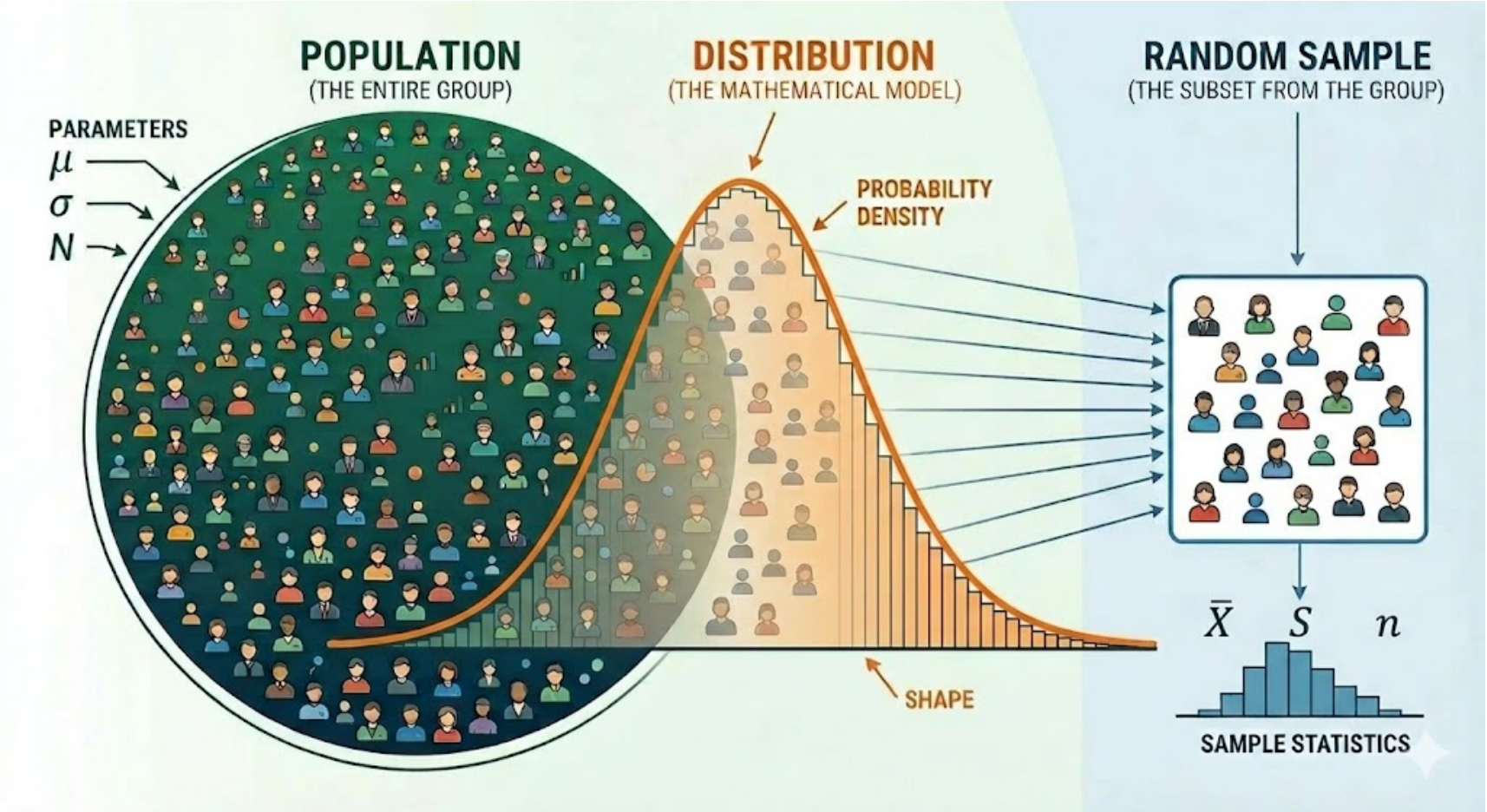
Random Sampling



A population is the totality of the observations with which we are concerned.

"Normal population" $\equiv$ A population whose observations are values of a random variable having a normal distribution

Random Sampling



A sample is a subset of a population.

In a random sample observations are made independently and at random:

Independent random variables $X_1, X_2, \cdots, X_n$ each with the same probability distribution $f(x)$.

Random Sampling

Any function of the random variables constituting a random sample is called a statistic. Let's review some statistics from initial lectures:

Let $X_1, \cdots, X_n$ iid (independently, identically distributed) random variables:

Sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Sample median: $\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if n is odd} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{otherwise} \end{cases}$

Sample mode: Value in the sample that occurs the most.

Random Sampling

Any function of the random variables constituting a random sample is called a statistic. Let's review some statistics from initial lectures:

Let $X_1, \cdots, X_n$ iid (independently, identically distributed) random variables:

Sample variance: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$

Sample standard deviation: $S = \sqrt{S^2}$

Sample range: $X_{max} - X_{min}$

Sampling Distributions and Sampling Distribution of Means

Probability distribution of a statistic is called a sampling distribution.

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

Note that $\overline{X}$ is a random variable:

$E(\overline{X}) =$

$Var(\overline{X}) =$

Sampling Distributions and Sampling Distribution of Means

Probability distribution of a statistic is called a sampling distribution.

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

Note that $\overline{X}$ is a random variable:

$E(\overline{X}) = E\left(\frac{1}{n}\sum x_i\right)$

$Var(\overline{X}) =$

Sampling Distributions and Sampling Distribution of Means

Probability distribution of a statistic is called a sampling distribution.

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

Note that $\overline{X}$ is a random variable:

$E(\overline{X}) = E\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n} E\left(\sum x_i\right)$

$Var(\overline{X}) =$

Sampling Distributions and Sampling Distribution of Means

Probability distribution of a statistic is called a sampling distribution.

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

Note that $\overline{X}$ is a random variable:

$E(\overline{X}) = E\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n}E\left(\sum x_i\right) = \frac{1}{n}\sum E(x_i) = \frac{1}{n} * n * \mu$

$Var(\overline{X}) =$

Sampling Distributions and Sampling Distribution of Means

Probability distribution of a statistic is called a sampling distribution.

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

Note that $\overline{X}$ is a random variable:

$E(\overline{X}) = E\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n}E\left(\sum x_i\right) = \frac{1}{n}\sum E(x_i) = \frac{1}{n} * n * \mu$

$= \mu$

$Var(\overline{X}) =$

Sampling Distributions and Sampling Distribution of Means

Probability distribution of a statistic is called a sampling distribution.

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

Note that $\overline{X}$ is a random variable:

$E(\overline{X}) = E\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n}E\left(\sum x_i\right) = \frac{1}{n}\sum E(x_i) = \frac{1}{n} * n * \mu = \mu$

$Var(\overline{X}) =$

(same as distribution mean μ)

Sampling Distributions and Sampling Distribution of Means

Probability distribution of a statistic is called a sampling distribution.

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

Note that $\overline{X}$ is a random variable:

$E(\overline{X}) = E\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n}E\left(\sum x_i\right) = \frac{1}{n}\sum E(x_i) = \frac{1}{n} * n * \mu = \mu$

$Var(\overline{X}) = Var\left(\frac{1}{n}\sum x_i\right)$

Sampling Distributions and Sampling Distribution of Means

Probability distribution of a statistic is called a sampling distribution.

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

Remember:
$Var(aX) = a^2 Var(X)$

Note that $\overline{X}$ is a random variable:

$E(\overline{X}) = E\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n}E\left(\sum x_i\right) = \frac{1}{n}\sum E(x_i) = \frac{1}{n} * n * \mu = \mu$

$Var(\overline{X}) = Var\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n^2} Var\left(\sum X_i\right)$

Sampling Distributions and Sampling Distribution of Means

Probability distribution of a statistic is called a sampling distribution.

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

Remember: X_i are independent

Note that $\overline{X}$ is a random variable:

$E(\overline{X}) = E\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n} E\left(\sum x_i\right) = \frac{1}{n}\sum E(x_i) = \frac{1}{n} * n * \mu = \mu$

$Var(\overline{X}) = Var\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n^2} Var\left(\sum X_i\right) = \frac{1}{n^2}\sum Var(X_i)$

Sampling Distributions and Sampling Distribution of Means

Probability distribution of a statistic is called a sampling distribution.

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

Note that $\overline{X}$ is a random variable:

$E(\overline{X}) = E\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n} E\left(\sum x_i\right) = \frac{1}{n}\sum E(x_i) = \frac{1}{n} * n * \mu = \mu$

$Var(\overline{X}) = Var\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n^2} Var\left(\sum X_i\right) = \frac{1}{n^2}\sum Var(X_i) = \frac{1}{n^2} * n * \sigma^2$

Sampling Distributions and Sampling Distribution of Means

Probability distribution of a statistic is called a sampling distribution.

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

Note that $\overline{X}$ is a random variable:

$E(\overline{X}) = E\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n}E\left(\sum x_i\right) = \frac{1}{n}\sum E(x_i) = \frac{1}{n} * n * \mu = \mu$

$Var(\overline{X}) = Var\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n^2}Var\left(\sum X_i\right) = \frac{1}{n^2}\sum Var(X_i) = \frac{1}{n^2} * n * \sigma^2 = \frac{\sigma^2}{n}$

Sampling Distributions and Sampling Distribution of Means

Probability distribution of a statistic is called a sampling distribution.

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

Note that $\overline{X}$ is a random variable:

$E(\overline{X}) =$

$Var(\overline{X}) =$

In summary =

Expected value of sample mean is the same as the mean of the underlying distribution (and is not related to the size of sample, n)

Sampling Distributions and Sampling Distribution of Means

Probability distribution of a statistic is called a sampling distribution.

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

Note that $\overline{X}$ is a random variable:

$E(\overline{X}) =$

$Var(\overline{X}) =$

In summary =

Variance of sample mean is inversely proportional to the size of sample, n.

Examples Combining Sample Mean with Chebyshev's

Review: X a random variable $\Rightarrow \forall\, t > 0,\ P(\,|X - \mu| \geq t) \leq \dfrac{\sigma^2}{t^2}$

Examples Combining Sample Mean with Chebyshev's

Review: X a random variable $\Rightarrow \forall\, t > 0,\ P(|X - \mu| \geq t) \leq \dfrac{\sigma^2}{t^2}$

(Earlier we covered a special form of Chebyshev's Theorem:

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

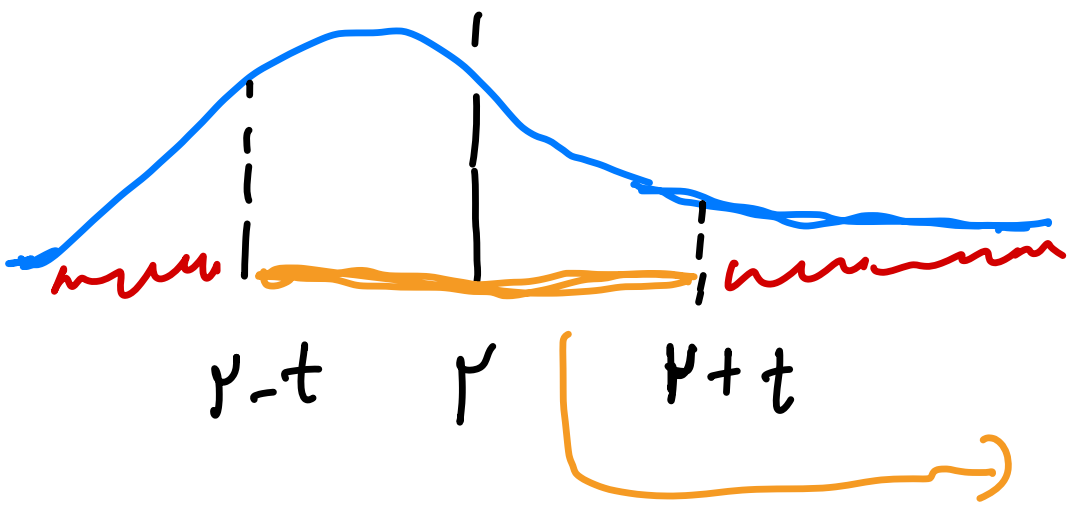
Examples Combining Sample Mean with Chebyshev's

Review: X a random variable $\Rightarrow \forall\, t > 0,\ P(|X - \mu| \geq t) \leq \dfrac{\sigma^2}{t^2}$

(Earlier we covered a special form of Chebyshev's Theorem:

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

we can deduce this from the general form stated above:



$$P(\mu - t < X < \mu + t) \geq 1 - \frac{\sigma^2}{t^2}$$

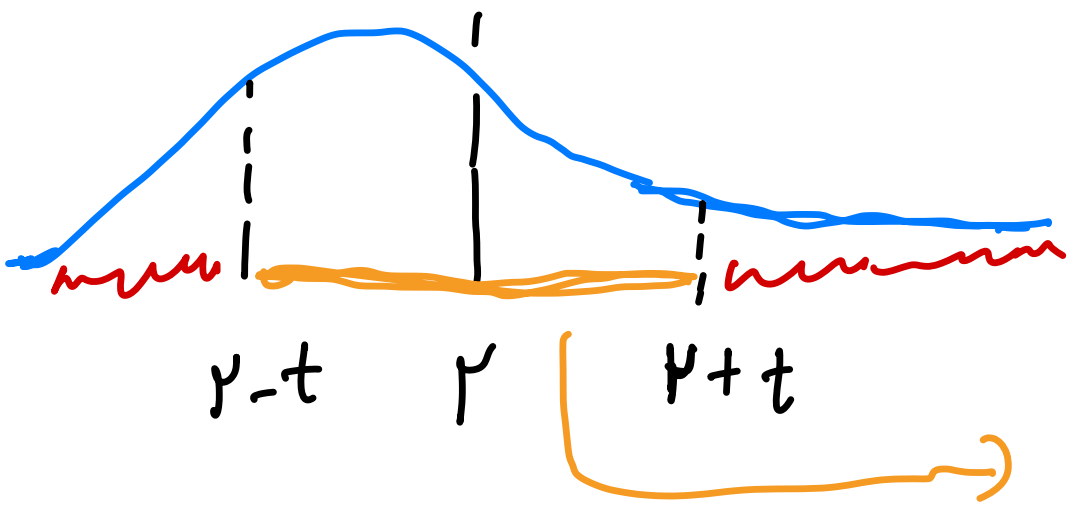
Examples Combining Sample Mean with Chebyshev's

Review: X a random variable $\Rightarrow \forall\, t > 0,\ P(|X - \mu| \geq t) \leq \dfrac{\sigma^2}{t^2}$

(Earlier we covered a special form of Chebyshev's Theorem:

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

we can deduce this from the general form stated above:



Replacing t with kσ gives us the exact special form.

$$P(\mu - t < X < \mu + t) \geq 1 - \frac{\sigma^2}{t^2}$$

Examples Combining Sample Mean with Chebyshev's

Example: Poll $100{,}000$ people. Assume each person in the population votes for candidate C independently with probability p.

Bound probability that poll is off by $\geq 1\,\%$?

Use Chebyshev's Theorem.

First find maximum possible variance of a Bernoulli RV.

Solution:

Examples Combining Sample Mean with Chebyshev's

Example: Poll $100{,}000$ people. Assume each person in the population votes for candidate C independently with probability p.

Bound probability that poll is off by $\geq 1\,\%$?

Use Chebyshev's Theorem.

First find maximum possible variance of a Bernoulli RV.

Solution:

$X_i = 1$ if i votes C and $X_i = 0$ otherwise

Examples Combining Sample Mean with Chebyshev's

Example: Poll $100{,}000$ people. Assume each person in the population votes for candidate C independently with probability p.

Bound probability that poll is off by $\geq 1\,\%$?

Use Chebyshev's Theorem.

First find maximum possible variance of a Bernoulli RV.

Solution:

$X_i = 1$ if i votes C and $X_i = 0$ otherwise

$E(X_i) = p \Rightarrow E(\bar{x}) = p$

Examples Combining Sample Mean with Chebyshev's

Example: Poll $100{,}000$ people. Assume each person in the population votes for candidate C independently with probability p.

Bound probability that poll is off by $\geq 1\,\%$?

Use Chebyshev's Theorem.

First find maximum possible variance of a Bernoulli RV.

Solution:

$X_i = 1$ if i votes C and $X_i = 0$ otherwise

$E(X_i) = p \Rightarrow E(\bar{x}) = p$

$Var(X_i) = p(1-p) \leq 0.25$

Rectangle with sides p and $(1-p)$ has maximum area when $p = (1-p)$

Examples Combining Sample Mean with Chebyshev's

Example: Poll $100{,}000$ people. Assume each person in the population votes for candidate C independently with probability p.

Bound probability that poll is off by $\geq 1\,\%$?

Use Chebyshev's Theorem.

First find maximum possible variance of a Bernoulli RV.

Solution:

$X_i = 1$ if i votes C and $X_i = 0$ otherwise

$E(X_i) = p \Rightarrow E(\bar{x}) = p$ $\quad Var(X_i) = p(1-p) \leq 0.25$

OR equivalently: find p where derivative of $p - p^2$ is 0.

Examples Combining Sample Mean with Chebyshev's

Example: Poll $100{,}000$ people. Assume each person in the population votes for candidate C independently with probability p.

Bound probability that poll is off by $\geq 1\,\%$?

Use Chebyshev's Theorem.

First find maximum possible variance of a Bernoulli RV.

Solution:

$X_i = 1$ if i votes C and $X_i = 0$ otherwise

$E(X_i) = p \Rightarrow E(\bar{x}) = p$

$Var(X_i) = p(1-p) \leq 0.25$

$\Rightarrow Var(\bar{X}) \leq \frac{0.25}{100{,}000}$

Examples Combining Sample Mean with Chebyshev's

Example: Poll $100{,}000$ people. Assume each person in the population votes for candidate C independently with probability p.

Bound probability that poll is off by $\geq 1\,\%$?

Use Chebyshev's Theorem.

First find maximum possible variance of a Bernoulli RV.

Solution:

$X_i = 1$ if i votes C and $X_i = 0$ otherwise

$E(X_i) = p \Rightarrow E(\bar{x}) = p$

$Var(X_i) = p(1-p) \leq 0.25$

$\Rightarrow Var(\bar{X}) \leq \frac{0.25}{100.000}$

Applying Chebyshev's =

$P((\bar{X} - p) \geq 0.01) \leq \frac{0.25}{100.000} \times \frac{1}{(0.01)^2} = 2.5\%$