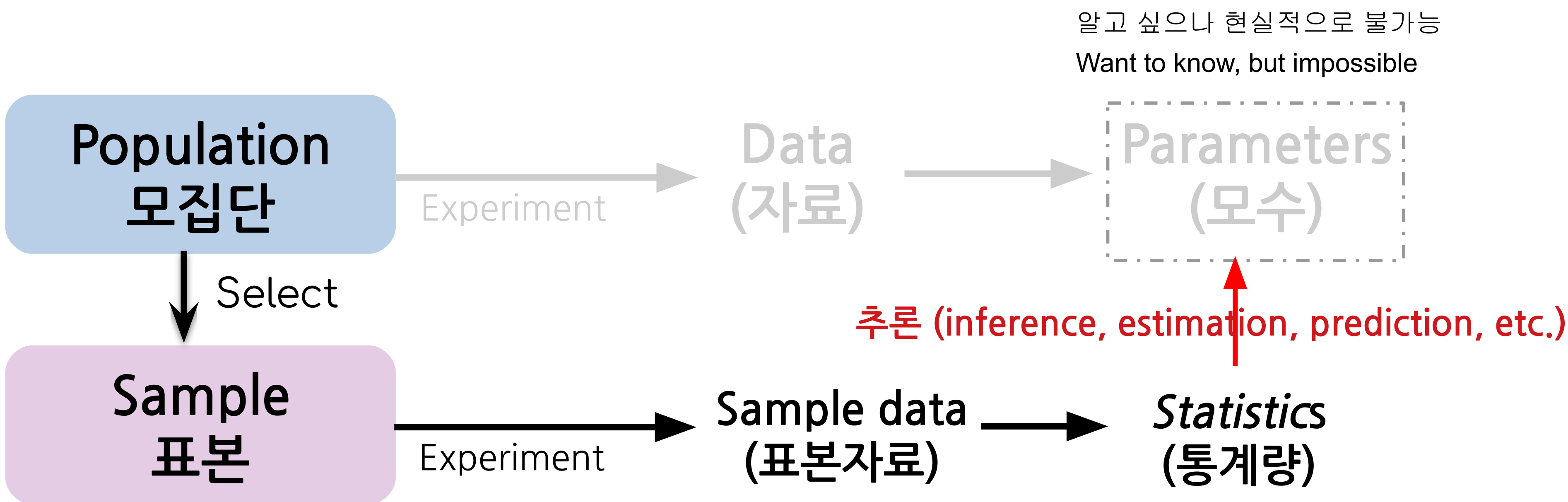
통계분석
Statistical Analysis

Inferential Statistics : 추론통계



Sampling

- Let us consider a population.
  X = Random variable of the population distribution

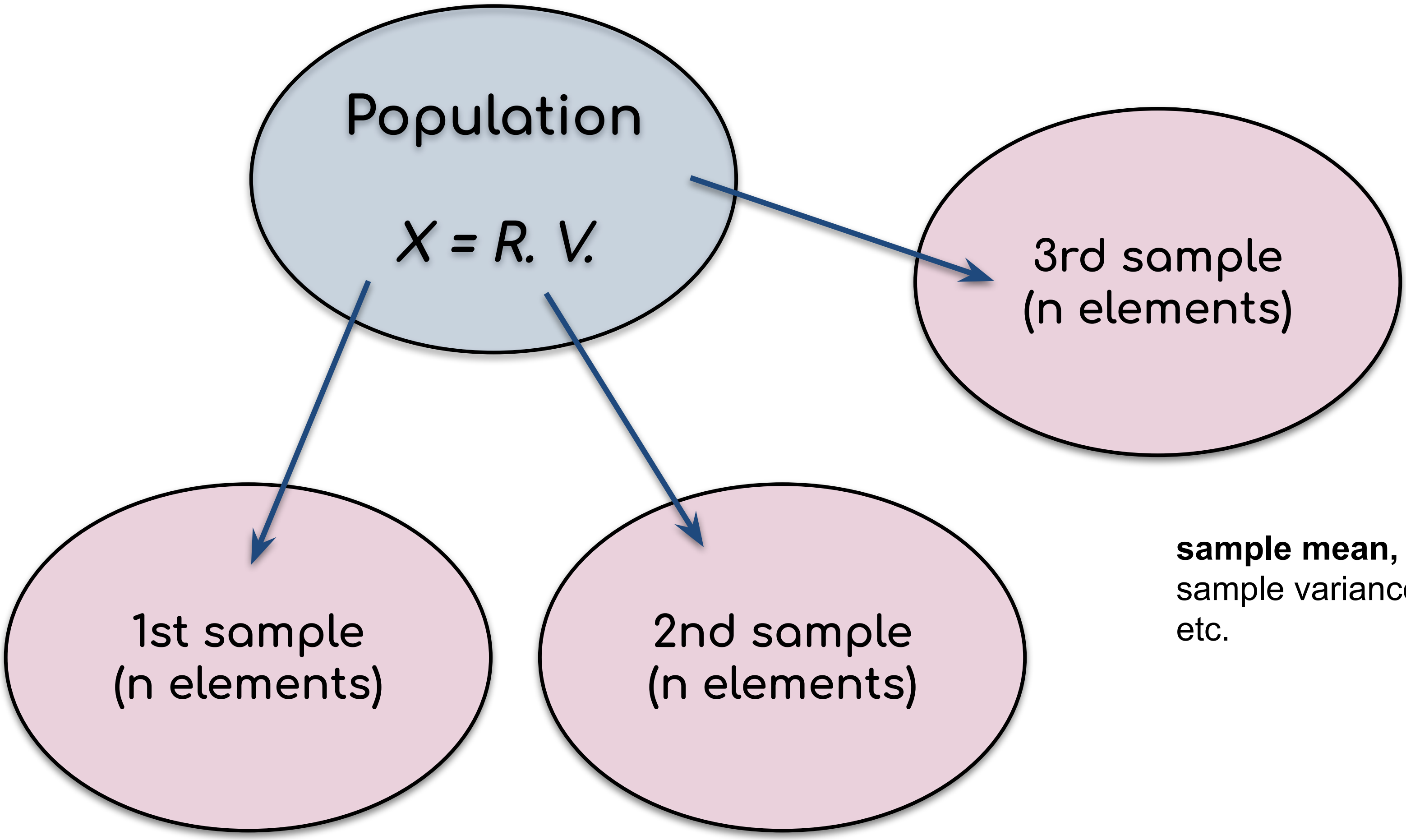
- Let us do n observations.
  We have n data, $\{x_1, x_2, \cdots, x_n\}$. **[The first sample]**

- When we do another n observations,
  We have a new set of n data, $\{x'_1, x'_2, \cdots, x'_n\}$. **[The 2nd sample]**

- Two set of sample data are different.
  There is uncertainty about sampling points $\{x_i\}$.

- Each of x_i can be regarded as a random variable.
  from the same population distribution.

Sampling



sample mean,
sample variance,
etc.

sample mean,
sample variance,
etc.

sample mean,
sample variance,
etc.

What is the distribution of sample statistics?

Random Sample and Statistic

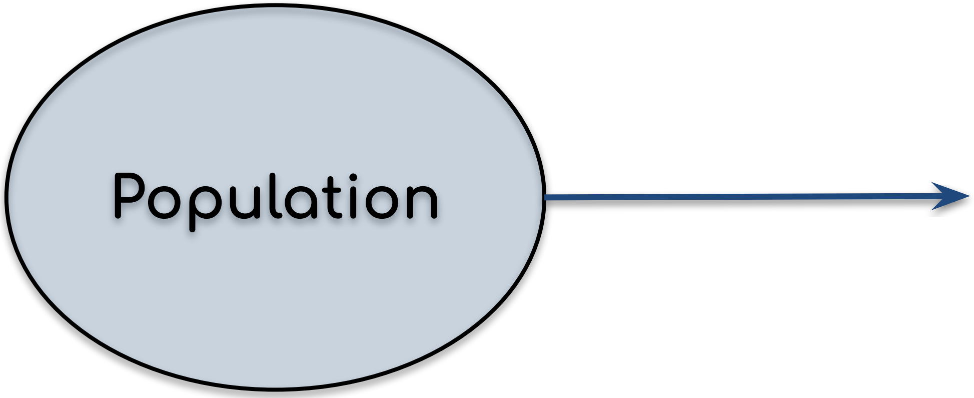
- If (1) X_is are independent of one another and (2) every X_i has the same probability distribution,

  Random Variables (RVs) $\boldsymbol{X_1, X_2, \ldots, X_n}$ = a ***Random Sample*** of size $\boldsymbol{n}$

- A ***statistic*** is any quantity calculated from sampling data.

- Due to uncertainty of sample data, a statistic is also a ***random variable***.

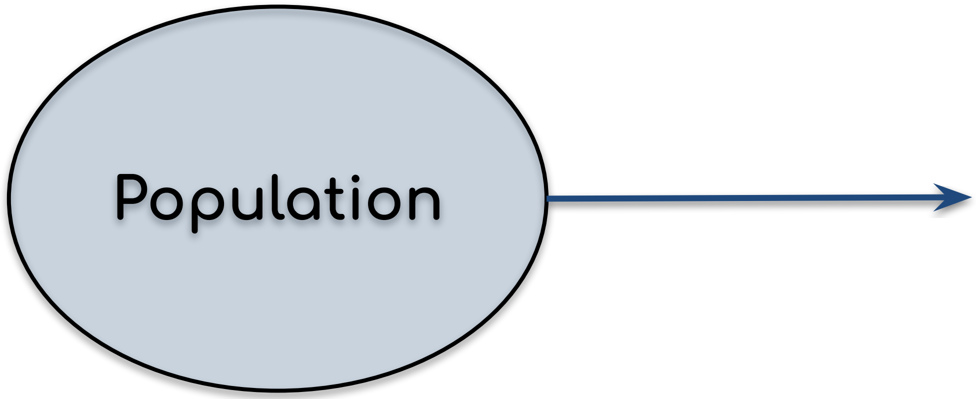
Statistics



n observations
$\{x_1, x_2, \cdots, x_n\}$

$$\text{statistic} = f(x_1, x_2, \cdot, x_n)$$

Sample Mean

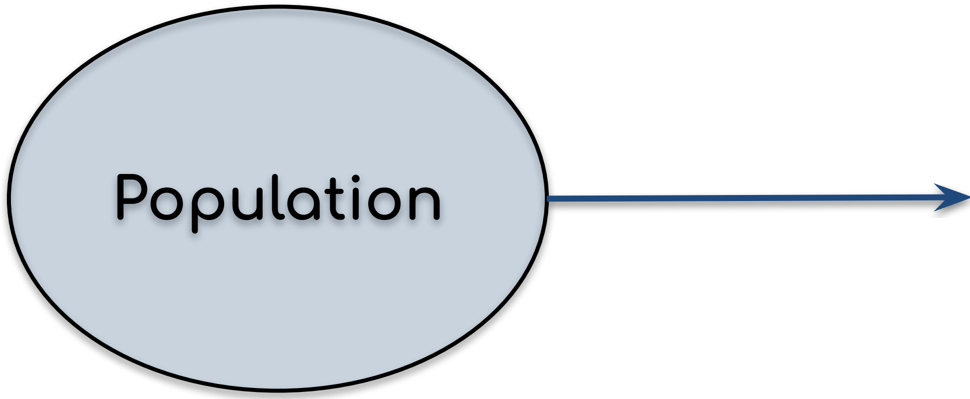


n observations
$\{x_1, x_2, \cdots, x_n\}$

$$\text{Sample Mean} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$\text{Population Mean} = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Sample Median



n observations

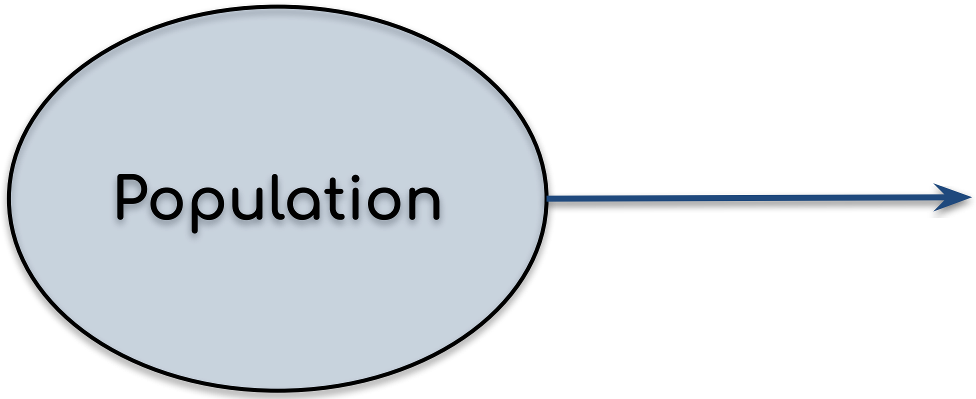
$$\{x_1, x_2, \cdots, x_n\}$$

$$x_1 \leq x_2 \leq \cdots \leq x_n$$

$$\begin{aligned} \text{sample median} = \tilde{x} &= x_{(n+1)/2} \text{ [n is odd]} \\ &= \frac{1}{2}\left[x_{n/2} + x_{n/2+1}\right] \text{ [n is even]} \end{aligned}$$

$$\text{population median} = \tilde{\mu} = \text{the middle of the ordered population values}$$

Sample Variance / Sample Standard Deviation



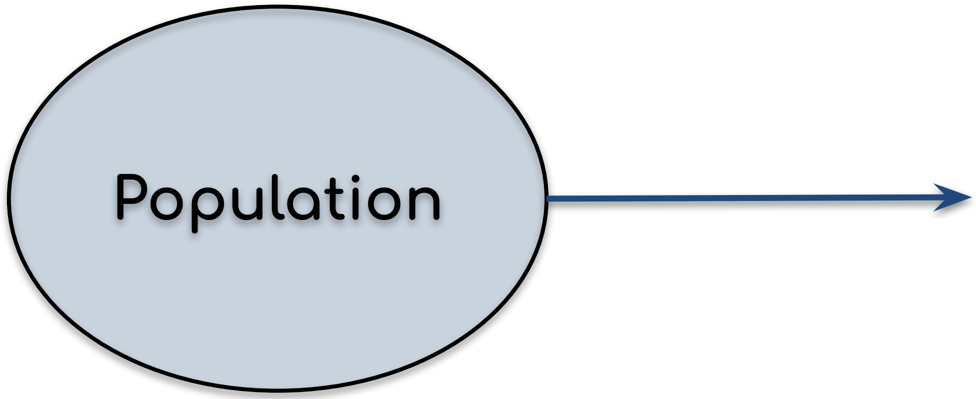
n observations
$\{x_1, x_2, \cdots, x_n\}$

$$\text{deviation from the mean} = x_i - \bar{x}$$

$$\text{sum of squared deviations} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\text{sample variance} = ?? \quad \frac{1}{n}?$$

Sample Variance / Sample Standard Deviation



n observations

$\{x_1, x_2, \cdots, x_n\}$

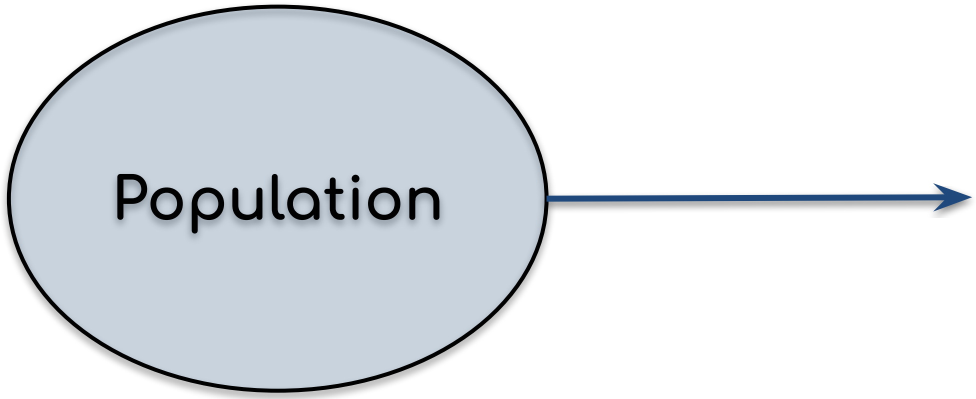
deviation from the mean $= x_i - \bar{x}$

sum of squared deviations $= \sum_{i=1}^{n}(x_i - \bar{x})^2$

sample variance $= ??$

You might want to write $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Sample Variance / Sample Standard Deviation



n observations
$\{x_1, x_2, \cdots, x_n\}$

$$\text{deviation from the mean} = x_i - \bar{x}$$

$$\text{sum of squared deviations} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\text{sample variance} = ??$$

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Sample Variance / Sample Standard Deviation

$$\text{sample variance} = ?? \quad \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Unbiased sample variance

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Biased sample variance

Variance of empirical distribution

Sample Variance / Sample Standard Deviation

$$\text{sample variance} =?? \qquad \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Unbiased sample variance

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Biased sample variance

Variance of empirical distribution

$\sigma^2 = \text{population variance}$

We would like to estimate the population variance.

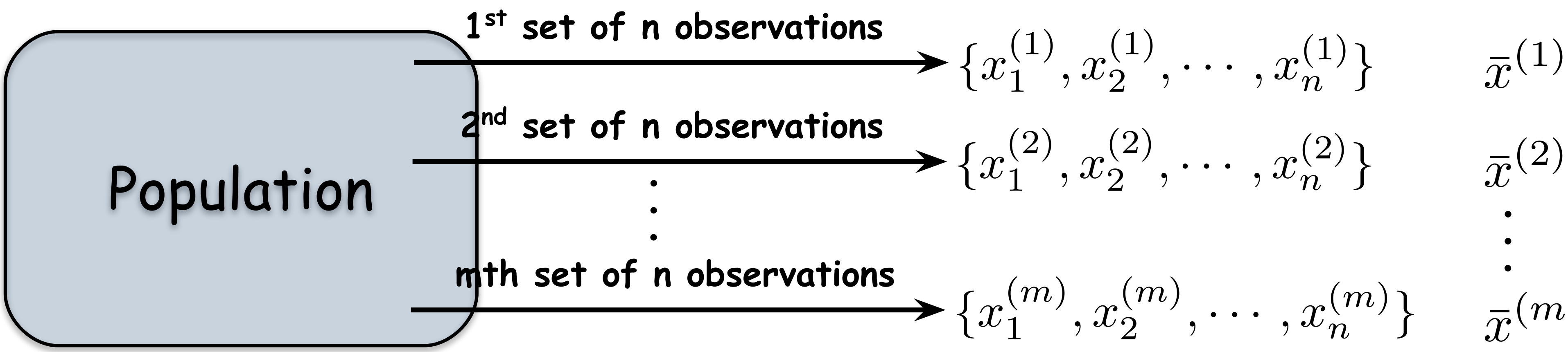
Biased sample variance tends to underestimate the population variance.

Sample Variance / Sample Standard Deviation

- $\text{sample variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2$
- $\text{sample standard deviation} = s = \sqrt{s^2}$

- $\text{population standard deviation} = \sigma = \sqrt{\sigma^2}$
- $\sigma^2 = \text{population variance}$

Distribution of Statistics



We would like to know ***distributions of statistics*** (sample mean, sample variance, etc.), which are random variables.

Distribution of Sample Mean

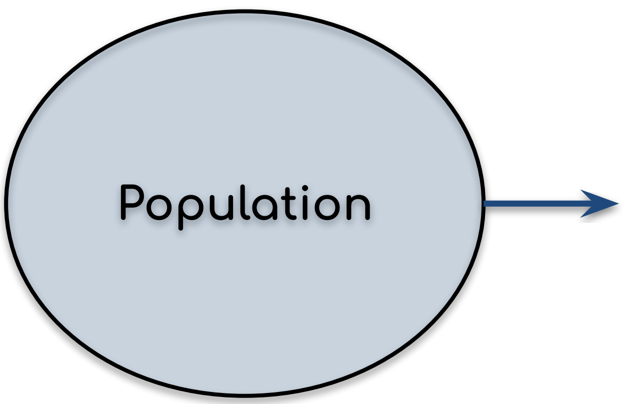
$X_1, X_2, \cdots, X_n$: Random sample from $\underline{\text{a distribution with mean } \mu \text{ and variance } \sigma^2}$

population distribution

$$\text{Sample Mean} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow$$

- Expectation Value of Sample Mean

$$E(\bar{X}) = \mu_{\bar{X}} = \mu$$



n observations
$\{x_1, x_2, \cdots, x_n\}$

- Variance of Sample Mean

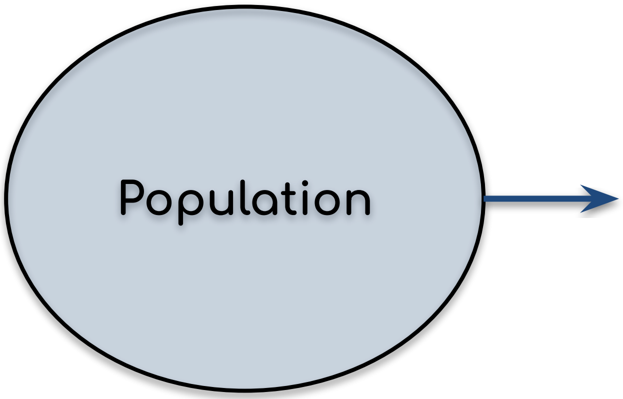
$$V(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

- Standard Deviation of Sample Mean

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- Here we are discussing properties of sample means, NOT sample variances.

Distribution of Sample Mean



n observations

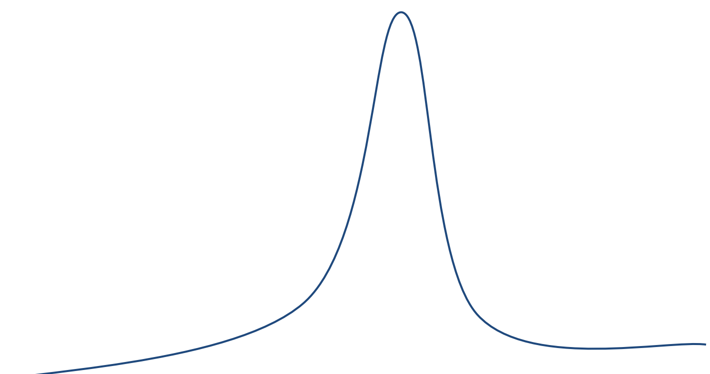
$$\{x_1, x_2, \cdots, x_n\}$$

- Variance of Sample Mean

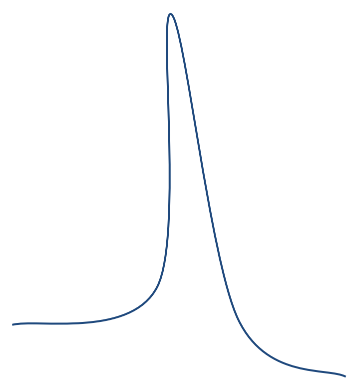
$$V(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

- Standard Deviation of Sample Mean

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$



sample mean of n=10



sample mean of n=40

- Here we are discussing properties of sample means, NOT sample variances.

Distribution of Sample Mean

- Expectation Value of Sample Mean

$$\text{Sample Mean} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow E(\bar{X}) = \mu_{\bar{X}} = \mu$$

$$u(X_1, \cdots X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E[u(X_1, \cdots X_n)] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

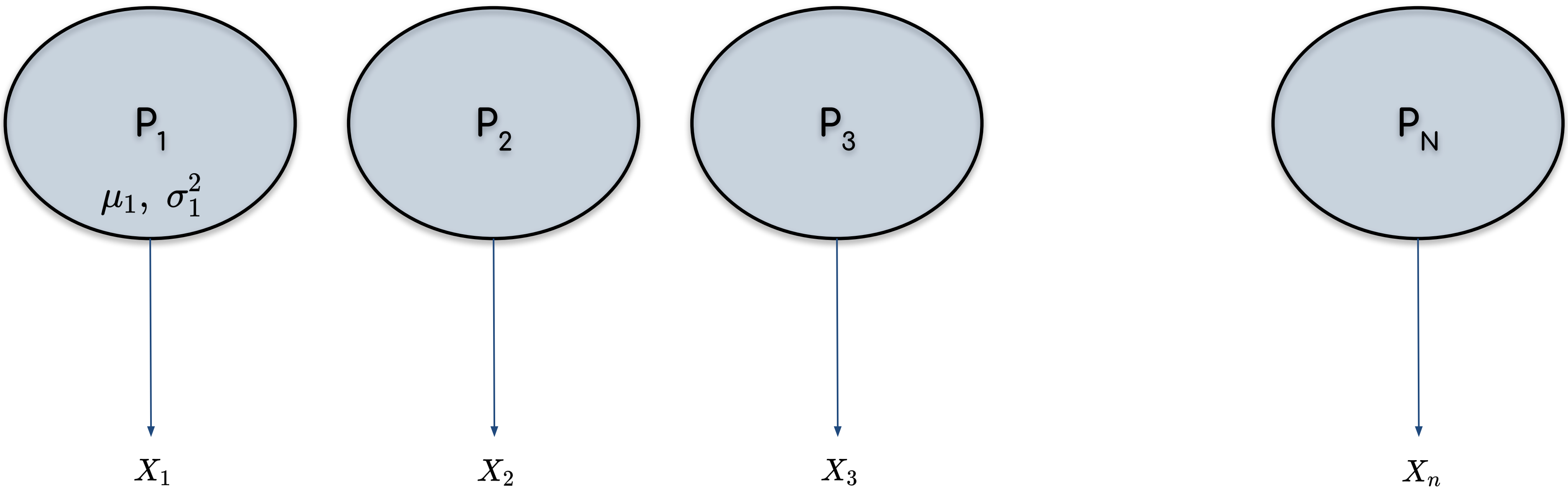
$$E[u(X_1, \cdots X_n)] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} n \times \mu = \mu$$

Distribution of Linear Combinations

- $X_1, X_2, \cdots, X_n$: Random variables

  $X_1, X_2, \cdots, X_n$ have means $\mu_1, \mu_2, \cdots, \mu_n,$ respectively.

  $X_1, X_2, \cdots, X_n$ have variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2,$ respectively.



Distribution of Linear Combinations
Generalization of Sample Mean

- $X_1, X_2, \cdots, X_n$: Random variables

  $X_1, X_2, \cdots, X_n$ have means $\mu_1, \mu_2, \cdots, \mu_n,$ respectively.

  $X_1, X_2, \cdots, X_n$ have variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2,$ respectively.

- Let us consider a linear combination,

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

for some constants $a_1, a_2, \cdots, a_n$

- Mean

$$E(Y) = E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1E(X_1) + \cdots + a_nE(X_n)$$
$$= a_1\mu_1 + \cdots + a_n\mu_n$$

It does not matter whether $X_1, X_2, \cdots, X_n$ are independent or not

Distribution of Linear Combinations

- $X_1, X_2, \cdots, X_n$: Random variables

  $X_1, X_2, \cdots, X_n$ have means $\mu_1, \mu_2, \cdots, \mu_n,$ respectively.

  $X_1, X_2, \cdots, X_n$ have variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2,$ respectively.

- Let us consider a linear combination,

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

for some constants $a_1, a_2, \cdots, a_n$

- Variance of Y

$$V(Y) = V(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j)a_ia_j$$

If $X_1, X_2, \cdots, X_n$ are independent,

$$V(Y) = V(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1^2V(X_1) + \cdots + a_n^2V(X_n)$$

$$= a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2$$

Variance of Sample Mean

$X_1, X_2, \cdots, X_n$: Random sample from a distribution with mean μ and variance σ^2

population distribution

These variables are **independent of one another**.
(Imagine the example of tossing coin many times. Each trial is independent of the other trials.

$$V[u(X_1, \cdots X_n)] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}\left(\frac{1}{n}X_i, \frac{1}{n}X_j\right)$$

$$= \sum_{i=1}^{n} \mathrm{Cov}\left(\frac{1}{n}X_i, \frac{1}{n}X_i\right) + \cancelto{0}{\sum_{i \neq j} \mathrm{Cov}\left(\frac{1}{n}X_i, \frac{1}{n}X_j\right)}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Cov}(X_i, X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{V}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

Distribution of Linear Combinations
Generalization of Sample Mean

- $X_1, X_2, \cdots, X_n$: Random variables

  $X_1, X_2, \cdots, X_n$ have means $\mu_1, \mu_2, \cdots, \mu_n,$ respectively.

  $X_1, X_2, \cdots, X_n$ have variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2,$ respectively.

- Let us consider a linear combination,

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$$

for some constants $a_1, a_2, \cdots, a_n$

- We can calculate expectation value and variance of Y, but we do not know the probability distribution of a random variable Y, which is what we want to know.
- In general, we do not know the exact expression of the distribution for Y.
- (Special case I) For sample mean of very large n, the central limit theorem gives us an approximate expression of the probability distribution.
- (Special case II) When the population distribution is normal, then we can know the distribution of Y.

Large Number of Samples

$X_1, X_2, \cdots, X_n$: Random sample from a distribution with mean μ and variance σ^2

Independent and identically-distributed (i.i.d.)

When n becomes large,

1) The Law of Large Numbers
2) The Central Limit Theorem**

The Law of Large Numbers

$X_1, X_2, \cdots, X_n$: Random sample from a distribution with mean μ and variance σ^2
Independent and identically-distributed (i.i.d.)

- **<u>As *n* increases</u>, the probability that the sample mean is close to the population mean goes to *1*.**

- **The sample mean based on a large *n* tends to be closer to the population mean than does the sample mean based on a small *n*.**

In general, $\mu \neq \bar{x}$

When n $\rightarrow \infty$, $P(\bar{x} = \mu) \approx 1$

Formal discussion: https://en.wikipedia.org/wiki/Law_of_large_numbers

The Central Limit Theorem

$X_1, X_2, \cdots, X_n$: Random sample from a distribution with mean μ and variance σ^2
Independent and identically-distributed (i.i.d.)

- For a sufficiently large n,

  $\bar{X}$ is approximately a normal distribution with $\mu_{\bar{X}} = \mu, \sigma^2_{\bar{X}} = \sigma^2/n$

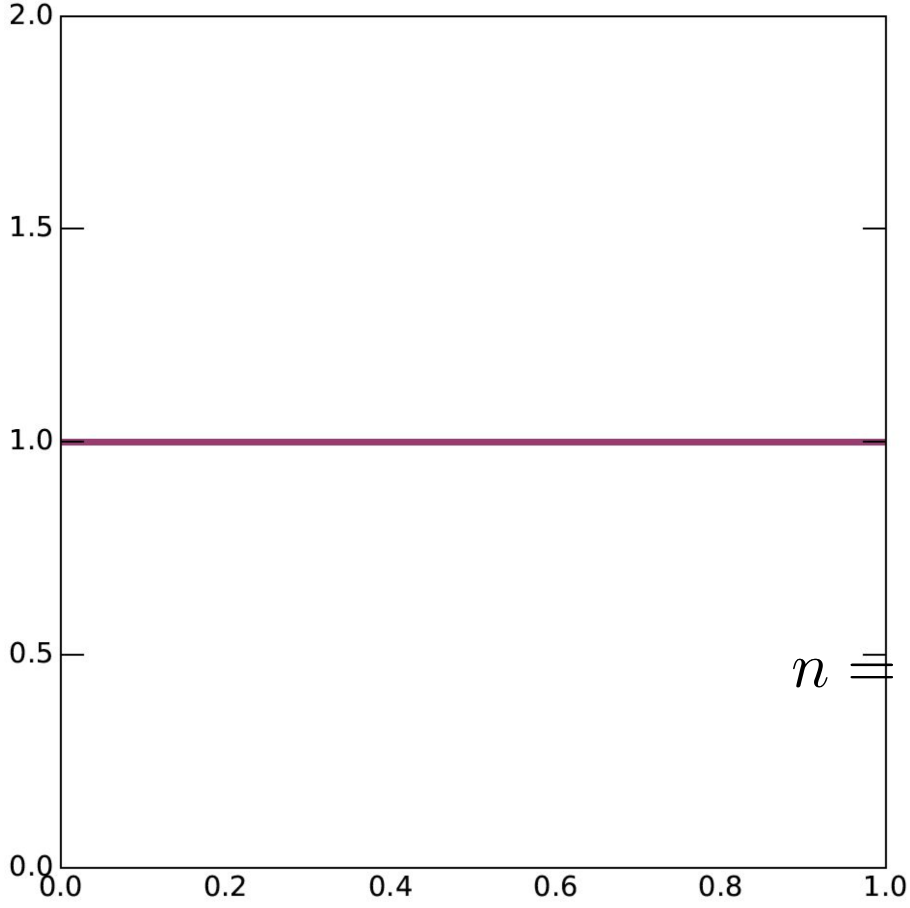
- When n is large, $\underline{\bar{X} \sim N(\mu, \sigma^2/n)}$

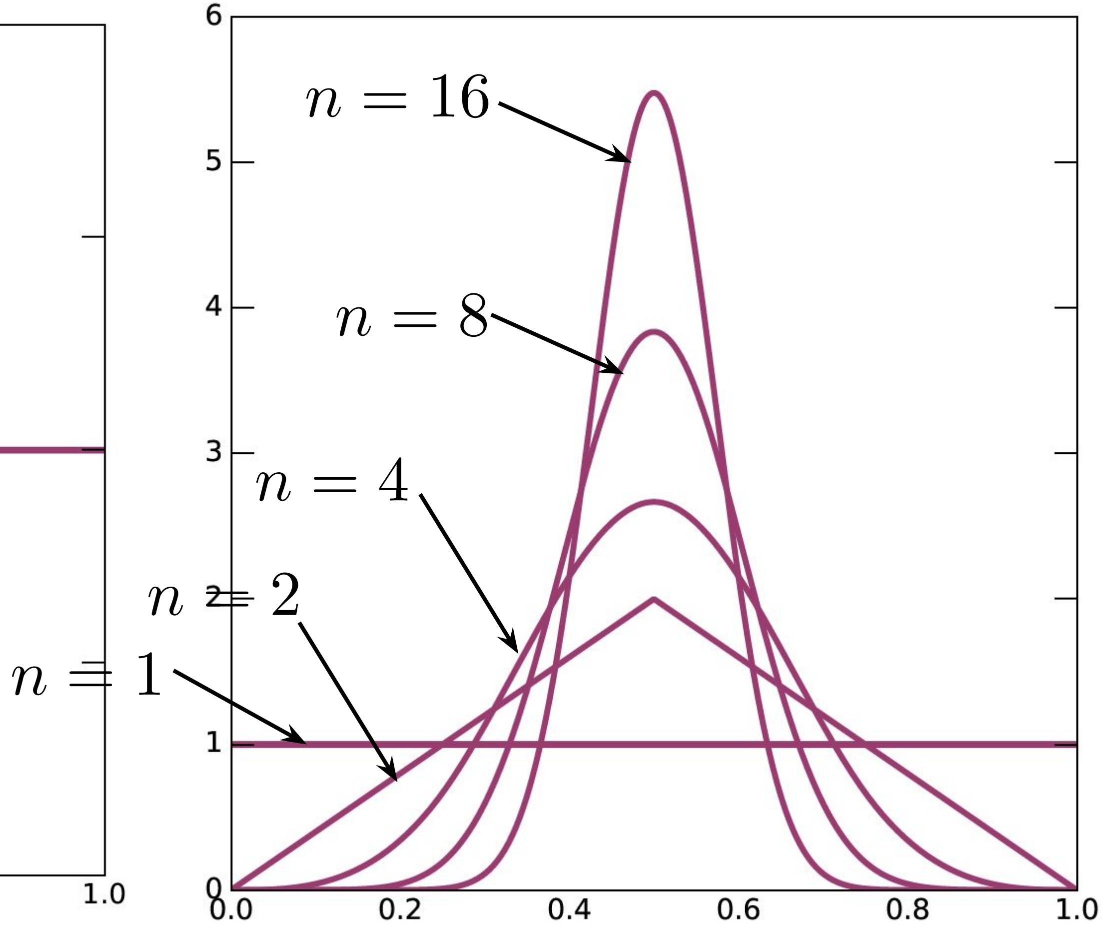
- How large is n?

Example: Uniform Distributions

$X_1, X_2, \cdots, X_n$: Random sample from a distribution with mean μ and variance σ^2

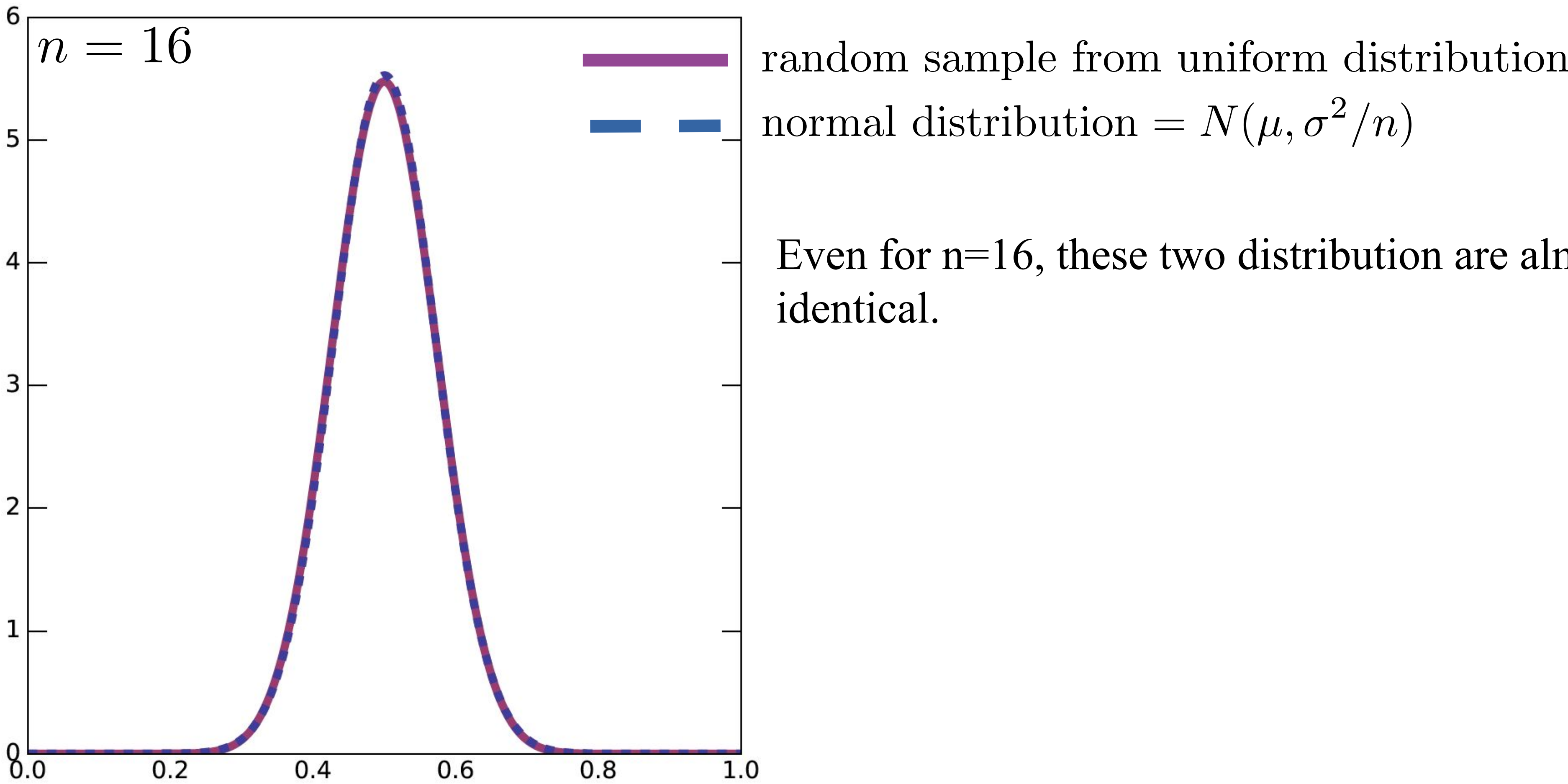
Independent and identically-distributed (i.i.d.)





Example: Uniform Distributions

$X_1, X_2, \cdots, X_n$: Random sample from a distribution with mean μ and variance σ^2
Independent and identically-distributed (i.i.d.)



Even for n=16, these two distribution are almost identical.

The Central Limit Theorem

$X_1, X_2, \cdots, X_n$: Random sample from a distribution with mean μ and variance σ^2
Independent and identically-distributed (i.i.d.)

- For a sufficiently large n,

  $\bar{X}$ is approximately a normal distribution with $\mu_{\bar{X}} = \mu, \sigma^2_{\bar{X}} = \sigma^2/n$

- When n is large, $\bar{X} \sim N(\mu, \sigma^2/n)$

- How large is n?

  Roughly speaking, ***<u>n > 30</u>*** might be large enough to use the central limit theorem.

  Here we do not have to know the population distribution.

Random Sample from Normal Distributions

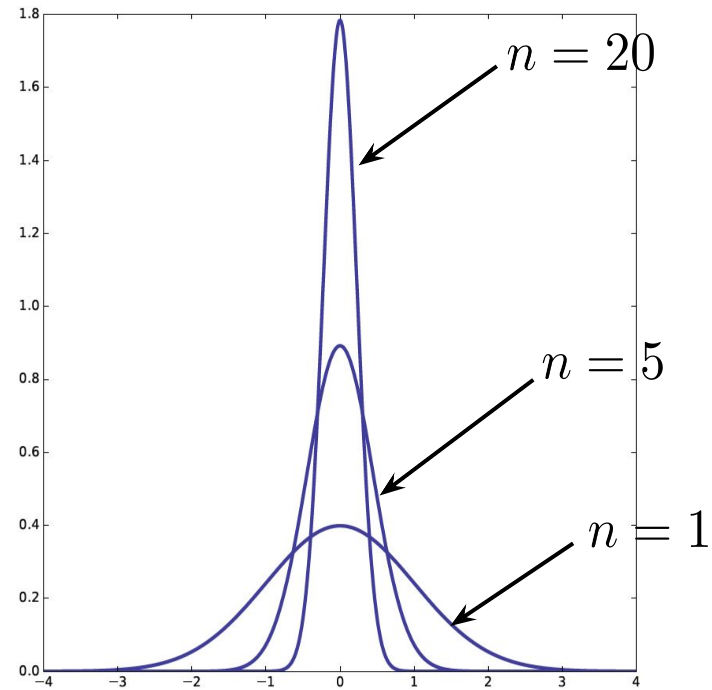
$$X_1, X_2, \cdots, X_n : \text{Random sample from } N(\mu, \sigma^2)$$

The population follows the normal distribution.

- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is normally distributed $\sim N(\mu, \sigma^2/n)$

NOT APPROXIMATION, BUT **EXACT**
Central Limit Theorem NOT APPLIED HERE
NOT PROVED; RESULTS JUST GIVEN

- As n increases, the normal distribution of $\bar{X}$ becomes sharper.



Random Sample from Normal Distributions

$X_1, X_2, \cdots, X_n$: Random sample from $N(\mu, \sigma^2)$

What is the distribution of any linear combination?: $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$
(Generalization of the previous slide)

- Any linear combination of $X_1, \cdots, X_n$ is normally distributed.

$$a_1X_1 + a_2X_2 + \cdots + a_nX_n \sim N(\mu', \sigma'^2)$$

$$\mu'?, \quad \sigma'?$$

NOT APPROXIMATION, BUT **EXACT**
Central Limit Theorem NOT APPLIED HERE
NOT PROVED; RESULTS JUST GIVEN