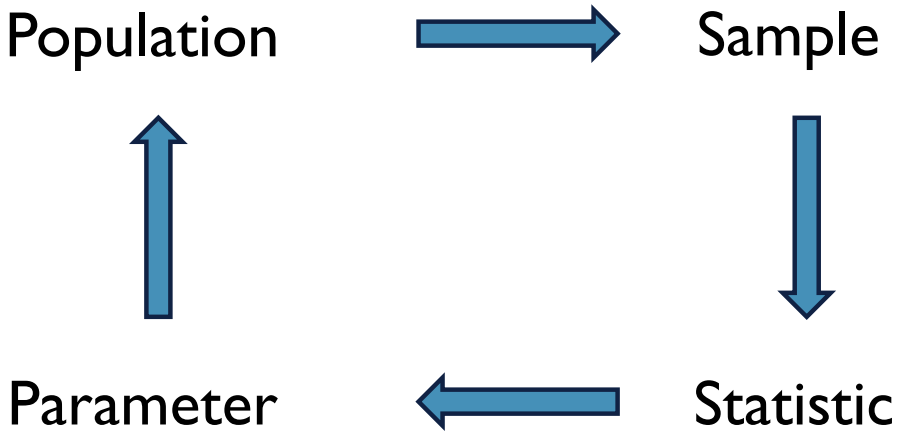
DISTRIBUTIONS OF STATISTICS FROM DATA

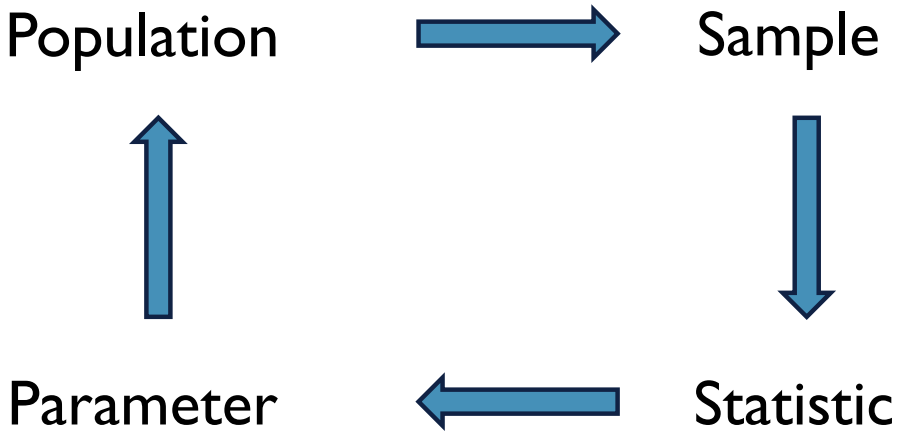
ST101 – DR. ARIC LABARR

REVIEW



- Population – set of all objects/individuals of interest.
- Sample – subset of the population that information is actually obtained.
- Statistic – measures computed from a sample.
- Parameter – measures computed from a population.

PARAMETERS VS. STATISTICS



- Population – set of all objects/individuals of interest.
- Sample – subset of the population that information is actually obtained.
- **Statistic** – measures computed from a sample.
- **Parameter** – measures computed from a population.

POINT ESTIMATORS

Point Estimator (Statistic)	Population Parameter
$\bar{x}$	μ
s^2	σ^2
$\hat{p}$	p

- Sample statistics are **point estimates** (single number estimates) of a population parameter.
- Different population parameters have different corresponding sample statistics.

SAMPLES ARE ESTIMATES

- Samples are *estimates* of the population.
- Statistics are *estimates* of the parameters.
- With any estimation, comes a chance of making errors.

SAMPLES VS. POPULATIONS

Population: 1, 3, 5, 5, 7, 9, 4, 6, 10, 2 $\mu = 5.2$

SAMPLES VS. POPULATIONS

Population: 1, 3, 5, 5, 7, 9, 4, 6, 10, 2 $\mu = 5.2$

Sample 1: 1, 10, 6, 9 $\bar{x}_1 = 6.5$

SAMPLES VS. POPULATIONS

Population: 1, 3, 5, 5, 7, 9, 4, 6, 10, 2 $\mu = 5.2$

Sample 1: 1, 10, 6, 9 $\bar{x}_1 = 6.5$

Sample 2: 1, 3, 2, 5 $\bar{x}_2 = 2.75$

SAMPLING ERROR

Population: 1, 3, 5, 5, 7, 9, 4, 6, 10, 2 $\mu = 5.2$

Sample 1: 1, 10, 6, 9 $\bar{x}_1 = 6.5$

Sample 2: 1, 3, 2, 5 $\bar{x}_2 = 2.75$

Both estimates are wrong!

SAMPLING ERROR

- Samples are *estimates* of the population.
- Statistics are *estimates* of the parameters.
- With any estimation, comes a chance of making errors.
- **Sampling error** occurs when there is a difference between a sample point estimate and the corresponding population parameter.

SAMPLING ERROR

Population: 1, 3, 5, 5, 7, 9, 4, 6, 10, 2 $\mu = 5.2$

Sample 1: 1, 10, 6, 9 $\bar{x}_1 = 6.5$ $\bar{x}_1 - \mu = 6.5 - 5.2 = 1.3$

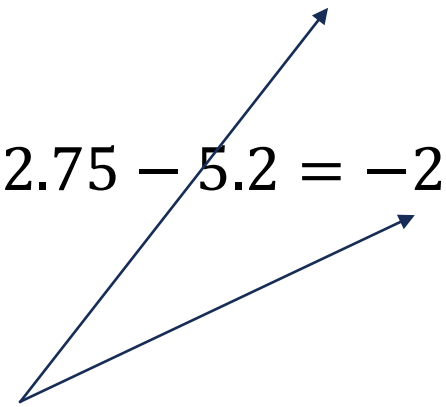
Sample 2: 1, 3, 2, 5 $\bar{x}_2 = 2.75$ $\bar{x}_2 - \mu = 2.75 - 5.2 = -2.45$

SAMPLING ERROR

Population: 1, 3, 5, 5, 7, 9, 4, 6, 10, 2 $\mu = 5.2$

Sample 1: 1, 10, 6, 9 $\bar{x}_1 = 6.5$ $\bar{x}_1 - \mu = 6.5 - 5.2 = 1.3$

Sample 2: 1, 3, 2, 5 $\bar{x}_2 = 2.75$ $\bar{x}_2 - \mu = 2.75 - 5.2 = -2.45$



Sampling Error!

SAMPLING ERROR

Population: 1, 3, 5, 5, 7, 9, 4, 6, 10, 2 $\quad \mu = 5.2$

Sample 1: 1, 10, 6, 9 $\quad \bar{x}_1 = 6.5$ $\quad \bar{x}_1 - \mu = 6.5 - 5.2 = 1.3$

Sample 2: 1, 3, 2, 5 $\quad \bar{x}_2 = 2.75$ $\quad \bar{x}_2 - \mu = 2.75 - 5.2 = -2.45$

Typically all we know! Rarely have the parameter to measure sampling error.

SAMPLING ERROR

Population: 1, 3, 5, 5, 7, 9, 4, 6, 10, 2 $\quad \mu = 5.2$

Sample 1: 1, 10, 6, 9 $\quad \bar{x}_1 = 6.5 \quad \bar{x}_1 - \mu = 6.5 - 5.2 = 1.3$

Sample 2: 1, 3, 2, 5 $\quad \bar{x}_2 = 2.75 \quad \bar{x}_2 - \mu = 2.75 - 5.2 = -2.45$

If sample statistics (like the sample mean) had a predictable pattern, then the errors would have a typical pattern as well!

SUMMARY

- Sample statistics are point estimates (single number estimates) of a population parameter.
- Sampling error occurs when there is a difference between a sample point estimate and the corresponding population parameter.
- If sample statistics (like the sample mean) had a predictable pattern, then the errors would have a typical pattern as well.

SAMPLING DISTRIBUTION FOR $\bar{x}$

DISTRIBUTIONS OF STATISTICS FROM DATA

POINT ESTIMATORS

Point Estimator (Statistic)	Population Parameter
$\bar{x}$	μ
s^2	σ^2
$\hat{p}$	p

- Sample statistics are **point estimates** (single number estimates) of a population parameter.
- Different population parameters have different corresponding sample statistics.

SAMPLING DISTRIBUTION

- The **sampling distribution of $\bar{x}$** is the probability distribution of all the possible values of the sample mean $\bar{x}$.

SAMPLING DISTRIBUTION

- The **sampling distribution of $\bar{x}$** is the probability distribution of all the possible values of the sample mean $\bar{x}$.
- The **sampling distribution of $\bar{x}$** has a mean (expected value) and variance as well.

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

$$SD(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

SAMPLING DISTRIBUTION

- The **sampling distribution of $\bar{x}$** is the probability distribution of all the possible values of the sample mean $\bar{x}$.
- The **sampling distribution of $\bar{x}$** has a mean (expected value) and variance as well.

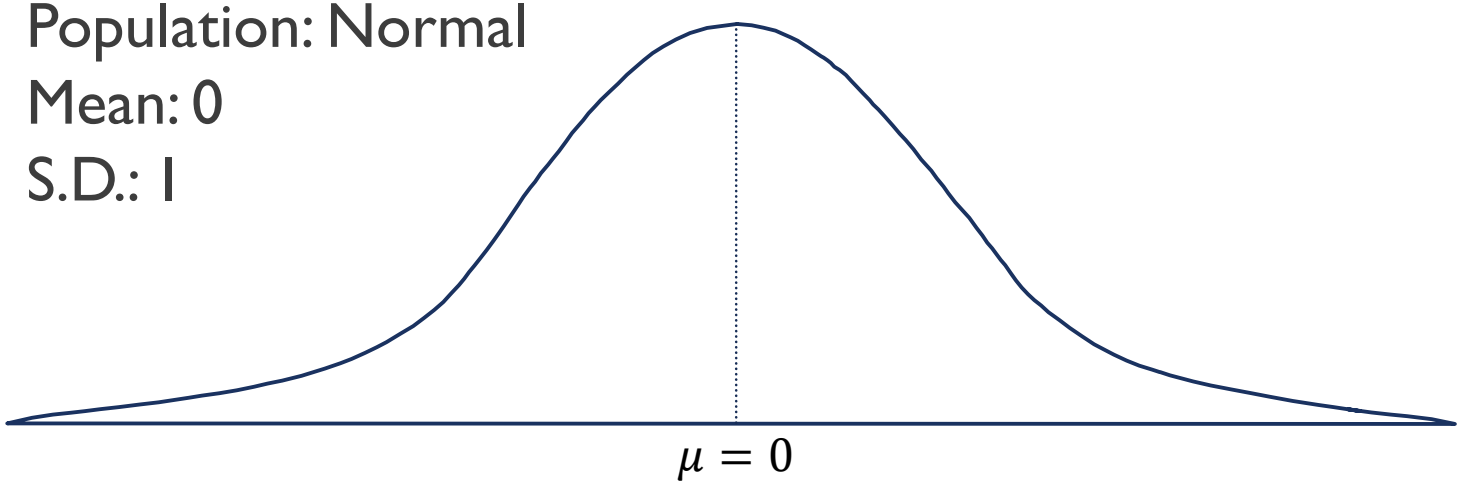
$$E(\bar{x}) = \mu_{\bar{x}} = \mu$$

$$SD(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- If sample larger than 5% of population: $SD(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$

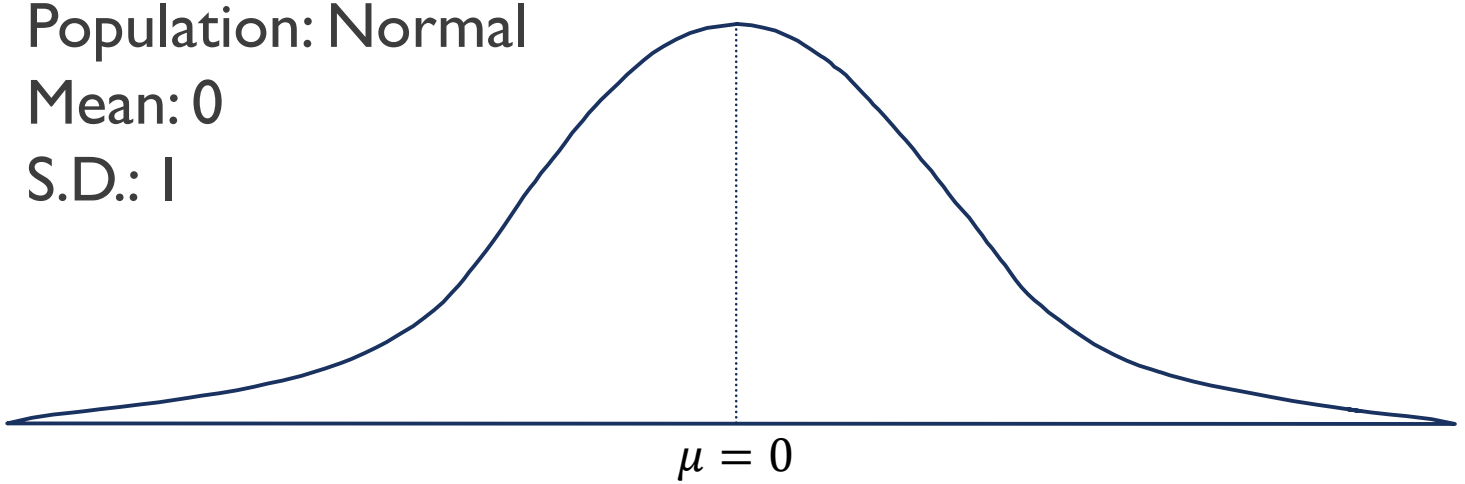
MANY, MANY SAMPLES

Population: Normal
Mean: 0
S.D.: 1



MANY, MANY SAMPLES

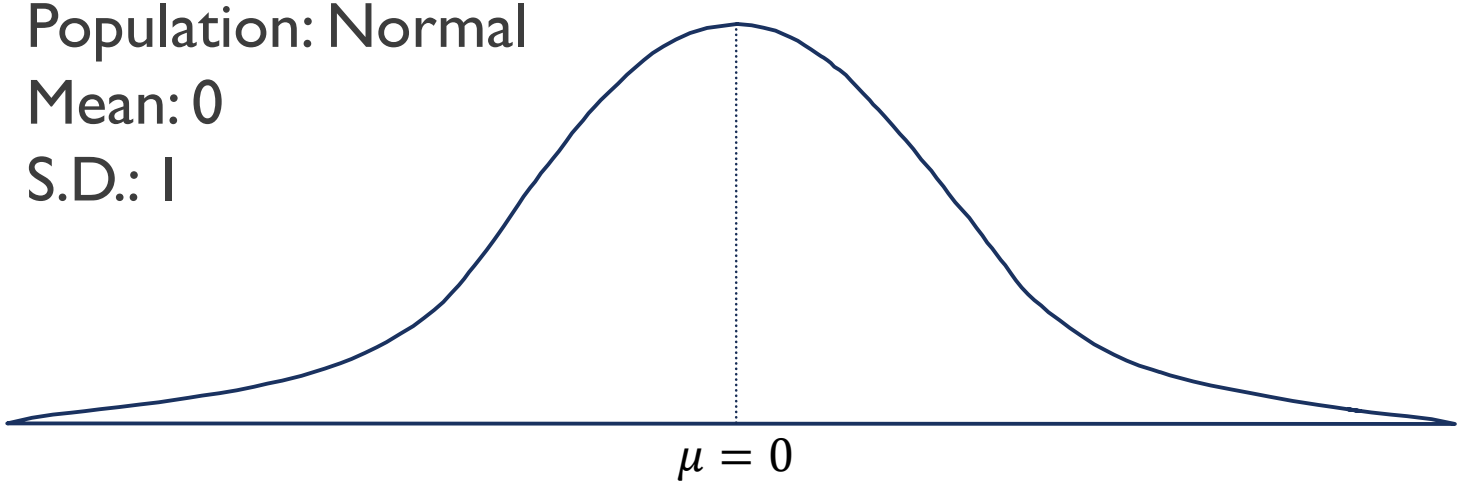
Population: Normal
Mean: 0
S.D.: 1



Sample 1: -1.4, 0.2, -1.7, 2.1, -2.0, 0.5, 1.6, -1.2, 0.6, 0.2 $\longrightarrow$ $\bar{x}_1 = -0.1$

MANY, MANY SAMPLES

Population: Normal
Mean: 0
S.D.: 1

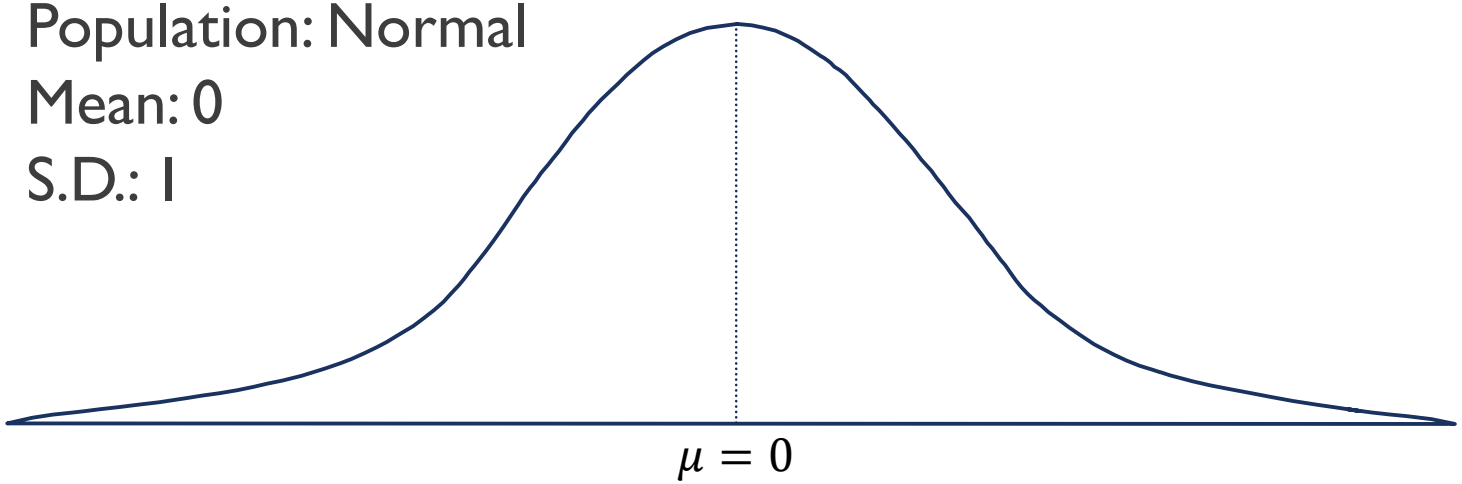


Sample 1: -1.4, 0.2, -1.7, 2.1, -2.0, 0.5, 1.6, -1.2, 0.6, 0.2 → $\bar{x}_1 = -0.1$

Sample 2: 0.8, -0.3, -0.6, -1.1, -1.3, 0.4, -0.9, -0.4, -1.0, -1.2 → $\bar{x}_2 = -0.6$

MANY, MANY SAMPLES

Population: Normal
Mean: 0
S.D.: 1



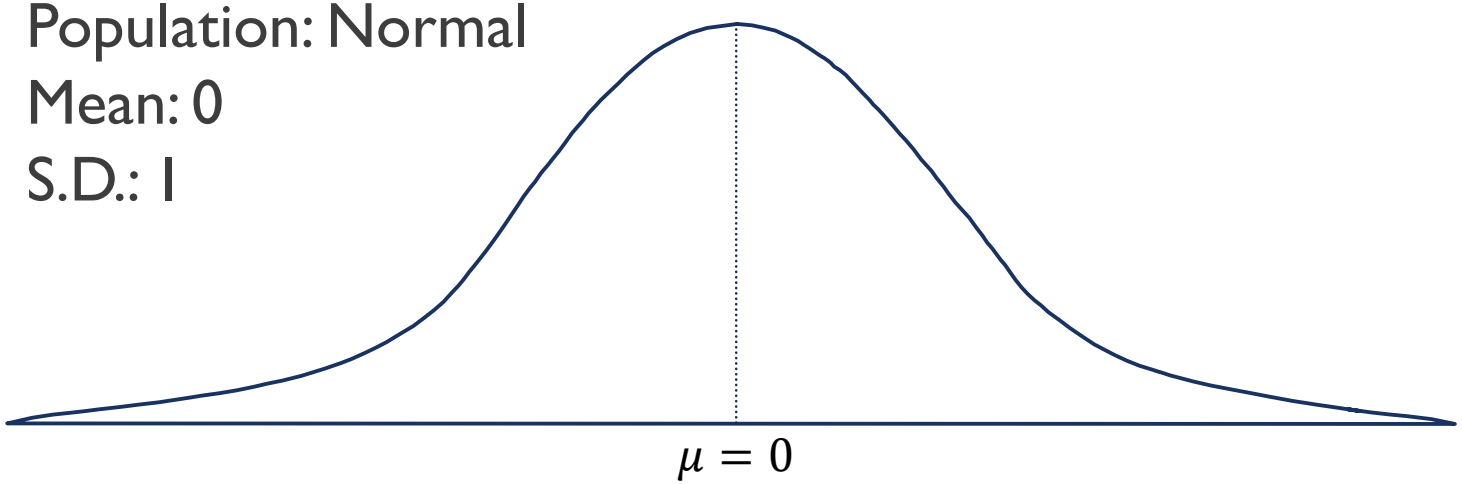
Sample 1: -1.4, 0.2, -1.7, 2.1, -2.0, 0.5, 1.6, -1.2, 0.6, 0.2 → $\bar{x}_1 = -0.1$

Sample 2: 0.8, -0.3, -0.6, -1.1, -1.3, 0.4, -0.9, -0.4, -1.0, -1.2 → $\bar{x}_2 = -0.6$

Sample 3: -0.2, 2.2, 0.7, 0.5, 1.2, -0.1, -0.6, -0.6, 0.7, -0.6 → $\bar{x}_3 = 0.3$

MANY, MANY SAMPLES

Population: Normal
Mean: 0
S.D.: 1



Sample 1: -1.4, 0.2, -1.7, 2.1, -2.0, 0.5, 1.6, -1.2, 0.6, 0.2 → $\bar{x}_1 = -0.1$

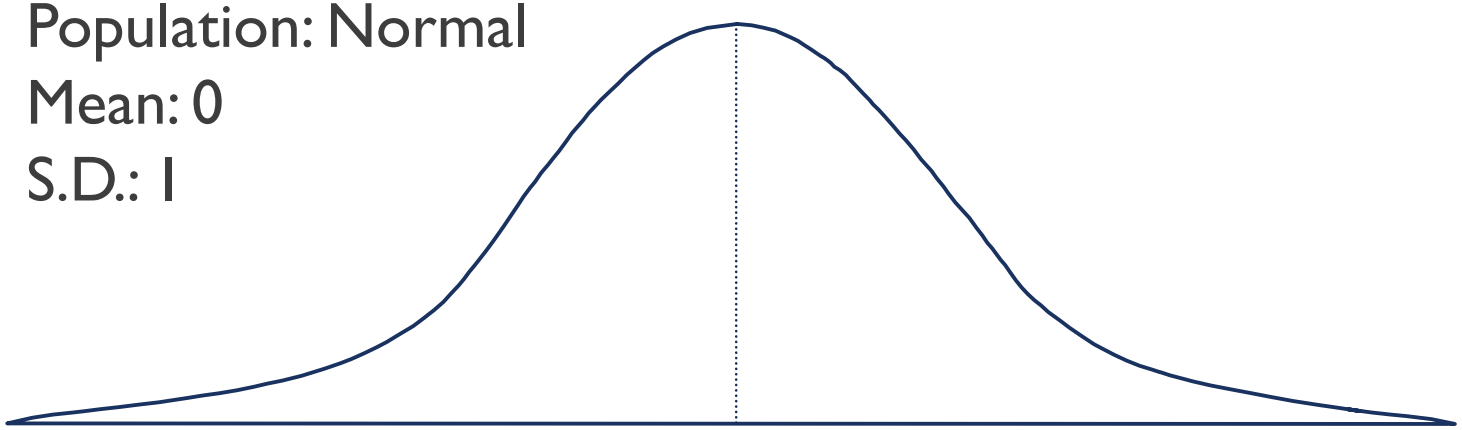
Sample 2: 0.8, -0.3, -0.6, -1.1, -1.3, 0.4, -0.9, -0.4, -1.0, -1.2 → $\bar{x}_2 = -0.6$

Sample 3: -0.2, 2.2, 0.7, 0.5, 1.2, -0.1, -0.6, -0.6, 0.7, -0.6 → $\bar{x}_3 = 0.3$

Sample 4: 2.0, -1.2, 1.6, 0.6, -0.8, 1.2, 0.8, 0.9, 0.5, -1.2 → $\bar{x}_4 = 0.4$

MANY, MANY SAMPLES

Population: Normal
Mean: 0
S.D.: 1



Sample 1: -1.4, 0.2, -1.7, 2.1, -2.0, 0.5, 1.6, -1.2, 0.6, 0.2 → $\bar{x}_1 = -0.1$

Sample 2: 0.8, -0.3, -0.6, -1.1, -1.3, 0.4, -0.9, -0.4, -1.0, -1.2 → $\bar{x}_2 = -0.6$

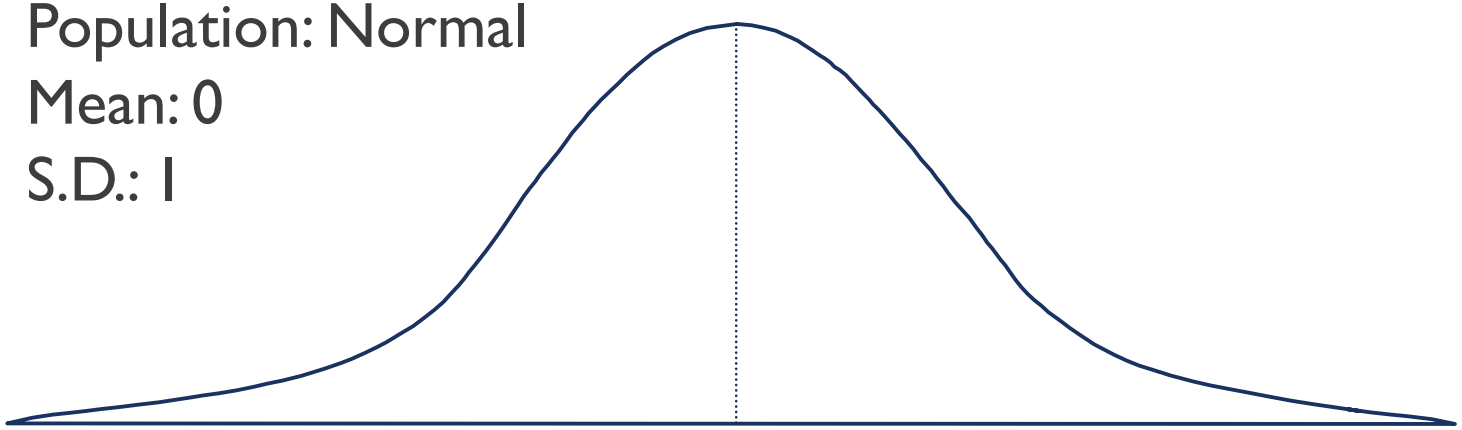
Sample 3: -0.2, 2.2, 0.7, 0.5, 1.2, -0.1, -0.6, -0.6, 0.7, -0.6 → $\bar{x}_3 = 0.3$

Sample 4: 2.0, -1.2, 1.6, 0.6, -0.8, 1.2, 0.8, 0.9, 0.5, -1.2 → $\bar{x}_4 = 0.4$

⋮ ⋮

DISTRIBUTION OF SAMPLE MEANS

Population: Normal
Mean: 0
S.D.: 1



Sample 1: -1.4, 0.2, -1.7, 2.1, -2.0, 0.5, 1.6, -1.2, 0.6, 0.2 → $\bar{x}_1 = -0.1$

Sample 2: 0.8, -0.3, -0.6, -1.1, -1.3, 0.4, -0.9, -0.4, -1.0, -1.2 → $\bar{x}_2 = -0.6$

Sample 3: -0.2, 2.2, 0.7, 0.5, 1.2, -0.1, -0.6, -0.6, 0.7, -0.6 → $\bar{x}_3 = 0.3$

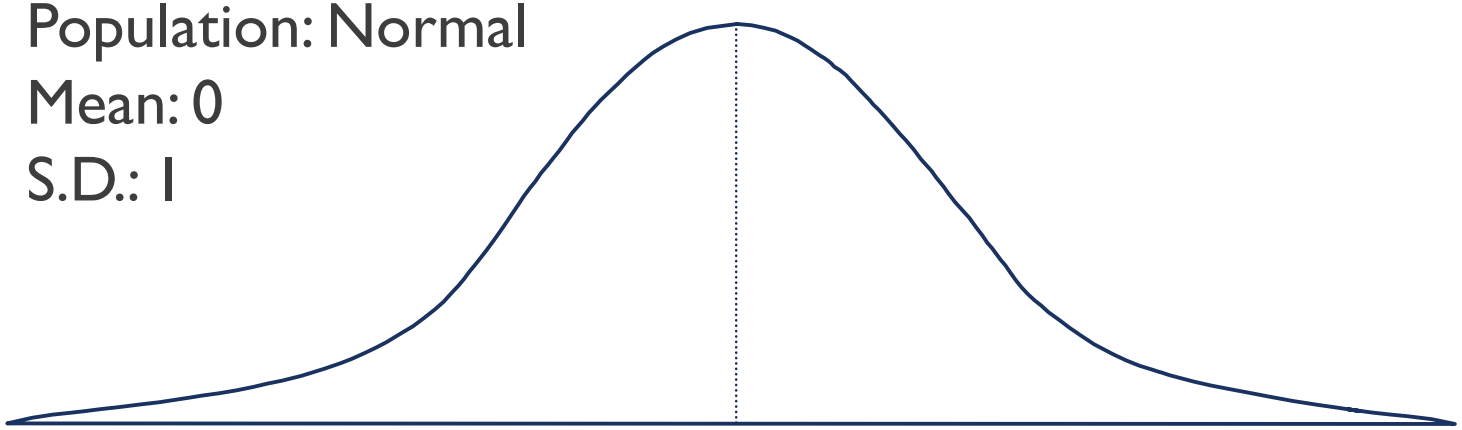
Sample 4: 2.0, -1.2, 1.6, 0.6, -0.8, 1.2, 0.8, 0.9, 0.5, -1.2 → $\bar{x}_4 = 0.4$

⋮ ⋮

What is the distribution of the sample means?

DISTRIBUTION OF SAMPLE MEANS

Population: Normal
Mean: 0
S.D.: 1



Sample 1: -1.4, 0.2, -1.7, 2.1, -2.0, 0.5, 1.6, -1.2, 0.6, 0.2 → $\bar{x}_1 = -0.1$

Sample 2: 0.8, -0.3, -0.6, -1.1, -1.3, 0.4, -0.9, -0.4, -1.0, -1.2 → $\bar{x}_2 = -0.6$

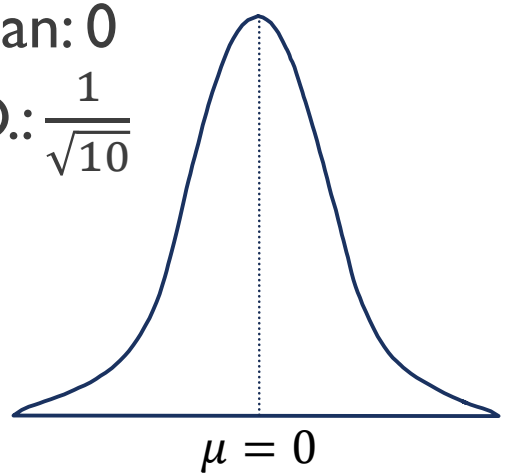
Sample 3: -0.2, 2.2, 0.7, 0.5, 1.2, -0.1, -0.6, -0.6, 0.7, -0.6 → $\bar{x}_3 = 0.3$

Sample 4: 2.0, -1.2, 1.6, 0.6, -0.8, 1.2, 0.8, 0.9, 0.5, -1.2 → $\bar{x}_4 = 0.4$

⋮ ⋮

What is the distribution of the sample means?

Dist of $\bar{x}$: ~Normal
Mean: 0
S.D.: $\frac{1}{\sqrt{10}}$

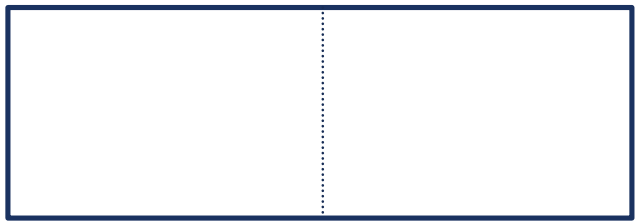


MANY, MANY SAMPLES

Population: Uniform

Mean: 0

S.D.: 1



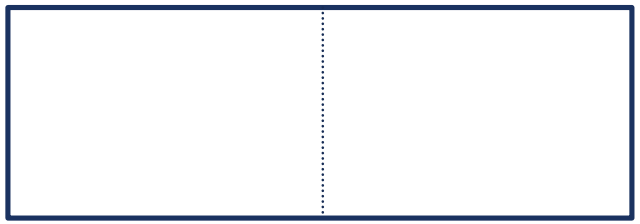
$\min = -1.73$ $\mu = 0$ $\max = 1.73$

MANY, MANY SAMPLES

Population: Uniform

Mean: 0

S.D.: 1



$\min = -1.73$ $\mu = 0$ $\max = 1.73$

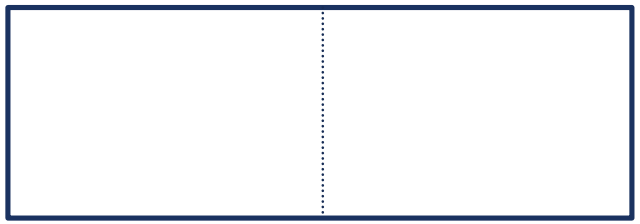
Sample 1: 0.7, -0.8, -0.2, 0.1, -0.6, 1.5, 1.6, -0.7, 0.7, 0.4 $\longrightarrow$ $\bar{x}_1 = 0.3$

MANY, MANY SAMPLES

Population: Uniform

Mean: 0

S.D.: 1



$\min = -1.73$ $\mu = 0$ $\max = 1.73$

Sample 1: 0.7, -0.8, -0.2, 0.1, -0.6, 1.5, 1.6, -0.7, 0.7, 0.4 → $\bar{x}_1 = 0.3$

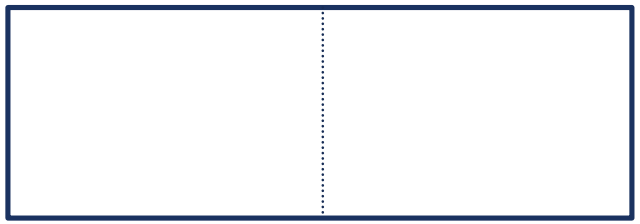
Sample 2: -1.0, -0.5, 0.1, -1.2, 0.1, 1.7, 1.5, 1.1, -1.7, -0.8 → $\bar{x}_2 = -0.1$

MANY, MANY SAMPLES

Population: Uniform

Mean: 0

S.D.: 1



$\min = -1.73$ $\mu = 0$ $\max = 1.73$

Sample 1: 0.7, -0.8, -0.2, 0.1, -0.6, 1.5, 1.6, -0.7, 0.7, 0.4 → $\bar{x}_1 = 0.3$

Sample 2: -1.0, -0.5, 0.1, -1.2, 0.1, 1.7, 1.5, 1.1, -1.7, -0.8 → $\bar{x}_2 = -0.1$

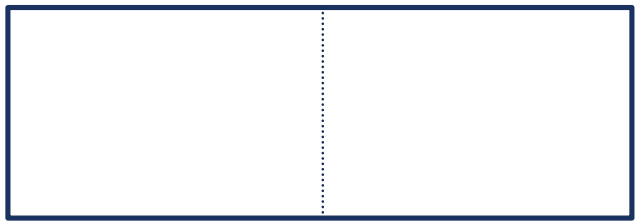
Sample 3: -0.9, -1.7, 0.2, 0.1, 1.3, -1.4, -1.2, 0.3, -0.1, 1.5 → $\bar{x}_3 = -0.2$

MANY, MANY SAMPLES

Population: Uniform

Mean: 0

S.D.: 1



$\min = -1.73$ $\mu = 0$ $\max = 1.73$

Sample 1: 0.7, -0.8, -0.2, 0.1, -0.6, 1.5, 1.6, -0.7, 0.7, 0.4 → $\bar{x}_1 = 0.3$

Sample 2: -1.0, -0.5, 0.1, -1.2, 0.1, 1.7, 1.5, 1.1, -1.7, -0.8 → $\bar{x}_2 = -0.1$

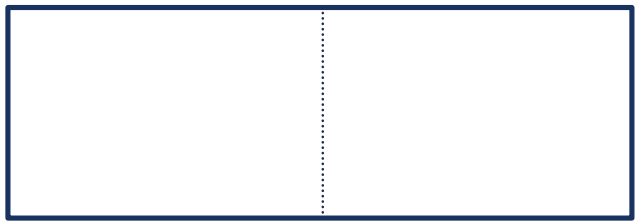
Sample 3: -0.9, -1.7, 0.2, 0.1, 1.3, -1.4, -1.2, 0.3, -0.1, 1.5 → $\bar{x}_3 = -0.2$

Sample 4: -0.6, -0.2, 0.8, 0.8, -0.7, -0.6, 1.6, -0.6, 0.6, -0.1 → $\bar{x}_4 = 0.1$

⋮ ⋮

DISTRIBUTION OF SAMPLE MEANS

Population: Uniform
Mean: 0
S.D.: 1



$\min = -1.73$ $\mu = 0$ $\max = 1.73$

Sample 1: 0.7, -0.8, -0.2, 0.1, -0.6, 1.5, 1.6, -0.7, 0.7, 0.4 → $\bar{x}_1 = 0.3$

Sample 2: -1.0, -0.5, 0.1, -1.2, 0.1, 1.7, 1.5, 1.1, -1.7, -0.8 → $\bar{x}_2 = -0.1$

Sample 3: -0.9, -1.7, 0.2, 0.1, 1.3, -1.4, -1.2, 0.3, -0.1, 1.5 → $\bar{x}_3 = -0.2$

Sample 4: -0.6, -0.2, 0.8, 0.8, -0.7, -0.6, 1.6, -0.6, 0.6, -0.1 → $\bar{x}_4 = 0.1$

⋮ ⋮

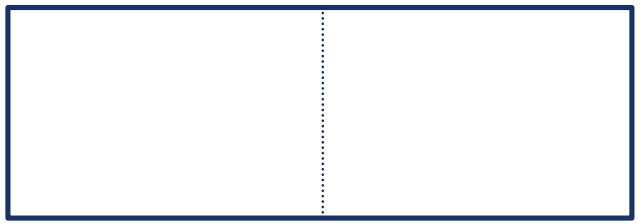
What is the distribution of the sample means?

DISTRIBUTION OF SAMPLE MEANS

Population: Uniform

Mean: 0

S.D.: 1



$\min = -1.73$ $\mu = 0$ $\max = 1.73$

Sample 1: 0.7, -0.8, -0.2, 0.1, -0.6, 1.5, 1.6, -0.7, 0.7, 0.4 → $\bar{x}_1 = 0.3$

Sample 2: -1.0, -0.5, 0.1, -1.2, 0.1, 1.7, 1.5, 1.1, -1.7, -0.8 → $\bar{x}_2 = -0.1$

Sample 3: -0.9, -1.7, 0.2, 0.1, 1.3, -1.4, -1.2, 0.3, -0.1, 1.5 → $\bar{x}_3 = -0.2$

Sample 4: -0.6, -0.2, 0.8, 0.8, -0.7, -0.6, 1.6, -0.6, 0.6, -0.1 → $\bar{x}_4 = 0.1$

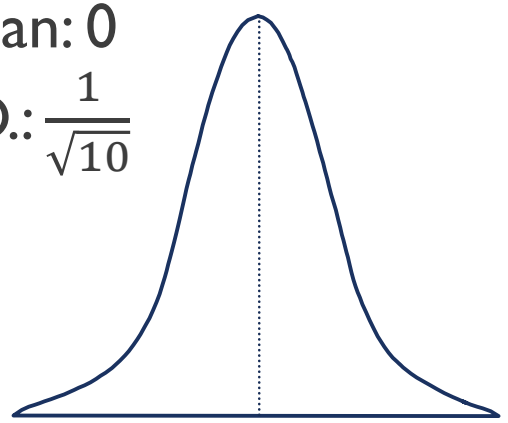
⋮ ⋮

What is the distribution of the sample means?

Dist of $\bar{x}$: ~Normal

Mean: 0

S.D.: $\frac{1}{\sqrt{10}}$



$\mu = 0$

CENTRAL LIMIT THEOREM

- If we use a large sample ($n \geq 50$), the **Central Limit Theorem (CLT)** states that the sampling distribution of $\bar{x}$ is approximately Normally distributed, **regardless of the population distribution**.

CENTRAL LIMIT THEOREM

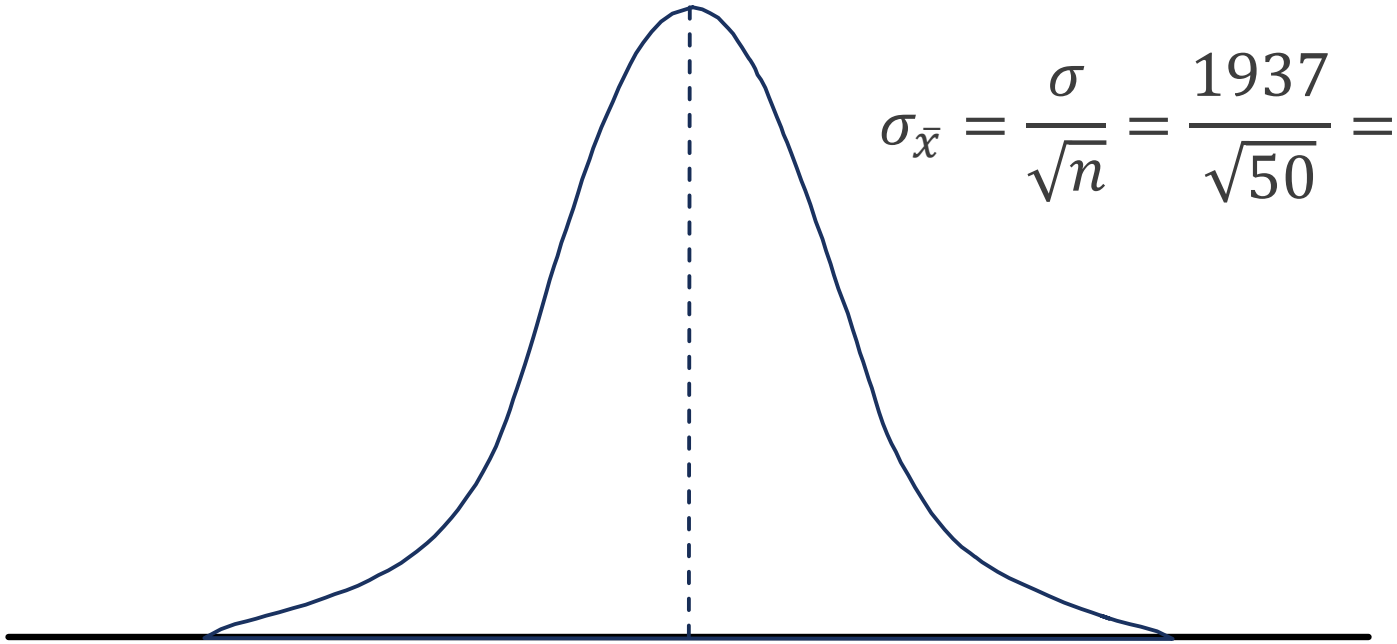
- If we use a large sample ($n \geq 50$), the **Central Limit Theorem (CLT)** states that the sampling distribution of $\bar{x}$ is approximately Normally distributed, **regardless of the population distribution**.
- If we use a small sample ($n < 50$), the sampling distribution of $\bar{x}$ is approximately Normally distributed **only if the population distribution is Normal**.

SAMPLING DISTRIBUTION BIKE DATA EXAMPLE

- The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the probability that **a sample of 50 days** has **an average** between 4,000 and 5,000 total users?

Z-SCORE FOR $\bar{x}$

- Based on our previous example, all of the possible sample means (from samples of size 50) would have the following distribution:



$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1937}{\sqrt{50}} = 273.93$$

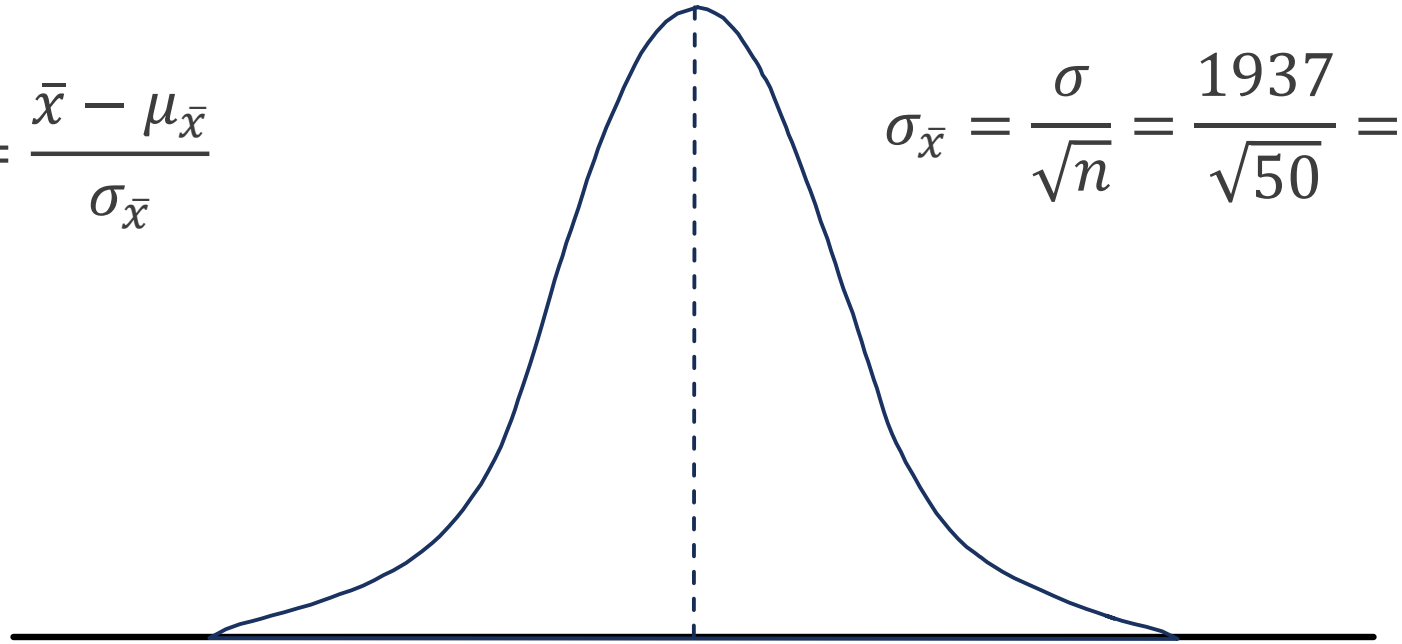
$$E(\bar{x}) = \mu = 4{,}504$$

Z-SCORE FOR $\bar{x}$

- Based on our previous example, all of the possible sample means (from samples of size 50) would have the following distribution:

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1937}{\sqrt{50}} = 273.93$$



$$E(\bar{x}) = \mu = 4{,}504$$

SAMPLING DISTRIBUTION BIKE DATA EXAMPLE

- The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the probability that **a sample of 50 days** has **an average** between 4,000 and 5,000 total users?

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

SAMPLING DISTRIBUTION BIKE DATA EXAMPLE

- The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the probability that **a sample of 50 days** has **an average** between 4,000 and 5,000 total users?

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

SAMPLING DISTRIBUTION BIKE DATA EXAMPLE

- The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the probability that **a sample of 50 days** has **an average** between 4,000 and 5,000 total users?

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{5000 - 4504}{\frac{1937}{\sqrt{50}}} = \frac{496}{273.93} = 1.81$$

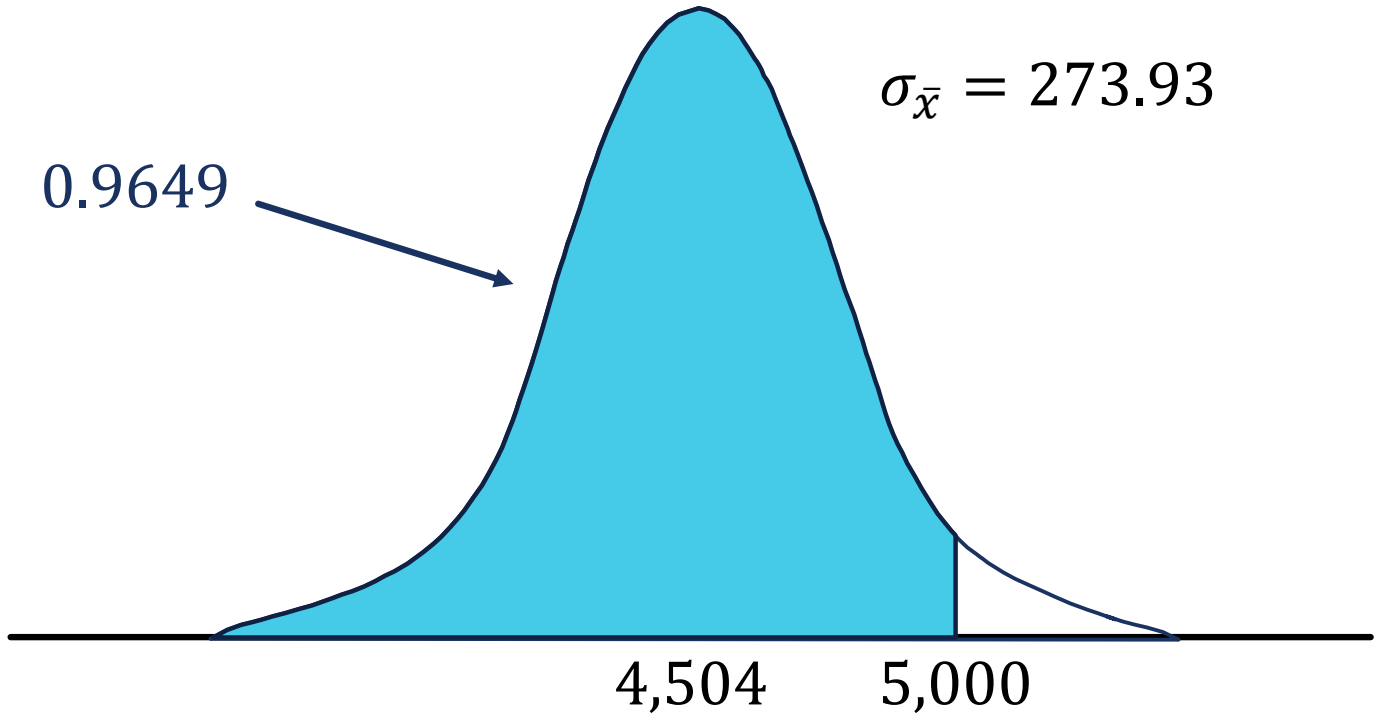
SAMPLING DISTRIBUTION BIKE DATA EXAMPLE

- The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the probability that **a sample of 50 days** has **an average** between 4,000 and 5,000 total users?

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{5000 - 4504}{\frac{1937}{\sqrt{50}}} = \frac{496}{273.93} = 1.81 \qquad P(z_{\bar{x}} \leq 1.81) = 0.9649$$

SAMPLING DISTRIBUTION BIKE DATA EXAMPLE

- The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the probability that **a sample of 50 days** has **an average** between 4,000 and 5,000 total users?



SAMPLING DISTRIBUTION BIKE DATA EXAMPLE

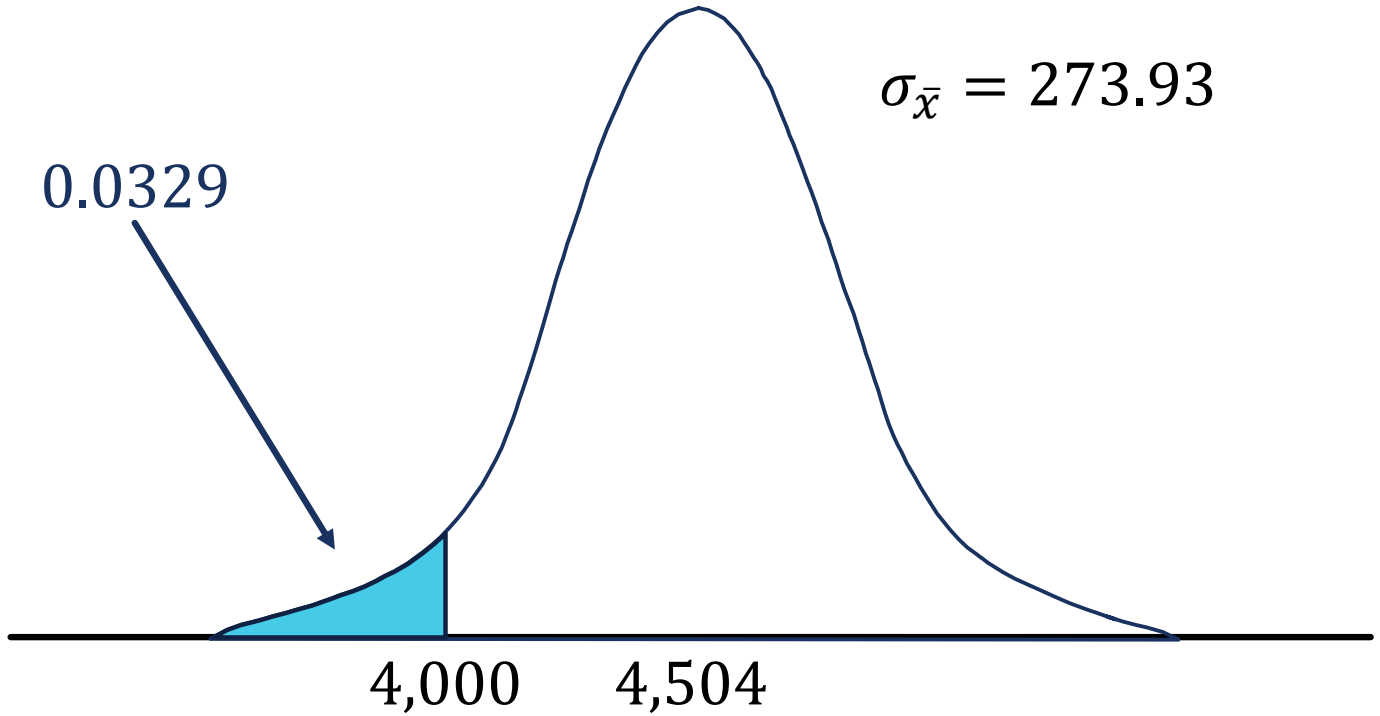
- The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the probability that **a sample of 50 days** has **an average** between 4,000 and 5,000 total users?

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{5000 - 4504}{\frac{1937}{\sqrt{50}}} = \frac{496}{273.93} = 1.81 \qquad P(z_{\bar{x}} \leq 1.81) = 0.9649$$

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{4000 - 4504}{\frac{1937}{\sqrt{50}}} = \frac{-504}{273.93} = -1.84 \qquad P(z_{\bar{x}} \leq -1.84) = 0.0329$$

SAMPLING DISTRIBUTION BIKE DATA EXAMPLE

- The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the probability that **a sample of 50 days** has **an average** between 4,000 and 5,000 total users?



SAMPLING DISTRIBUTION BIKE DATA EXAMPLE

- The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the probability that **a sample of 50 days** has **an average** between 4,000 and 5,000 total users?

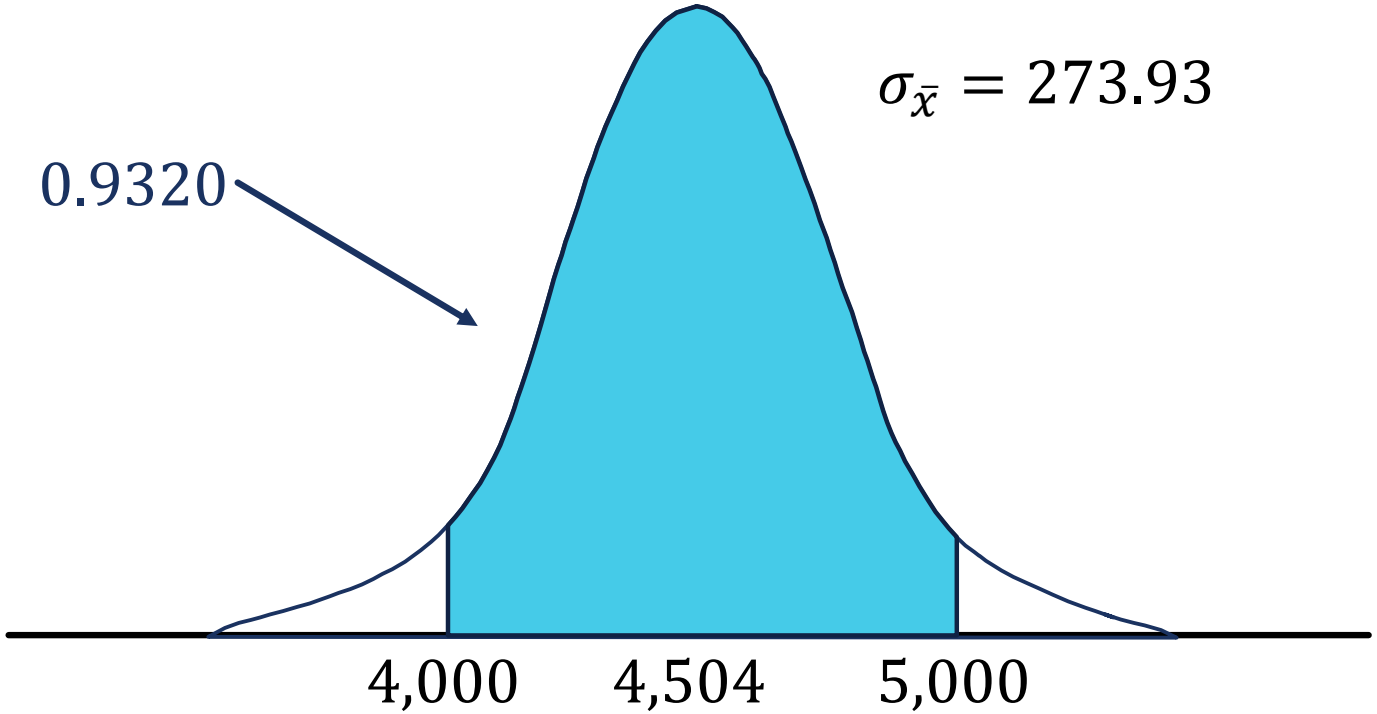
$$P(-1.84 \leq z_{\bar{x}} \leq 1.81) = 0.9649 - 0.0329$$

$$= 0.9320$$

$$P(4{,}000 \leq \bar{x} \leq 5{,}000) = 0.9320$$

SAMPLING DISTRIBUTION BIKE DATA EXAMPLE

- The average daily number of total users is 4,504 with a standard deviation of 1,937. What is the probability that **a sample of 50 days** has **an average** between 4,000 and 5,000 total users?

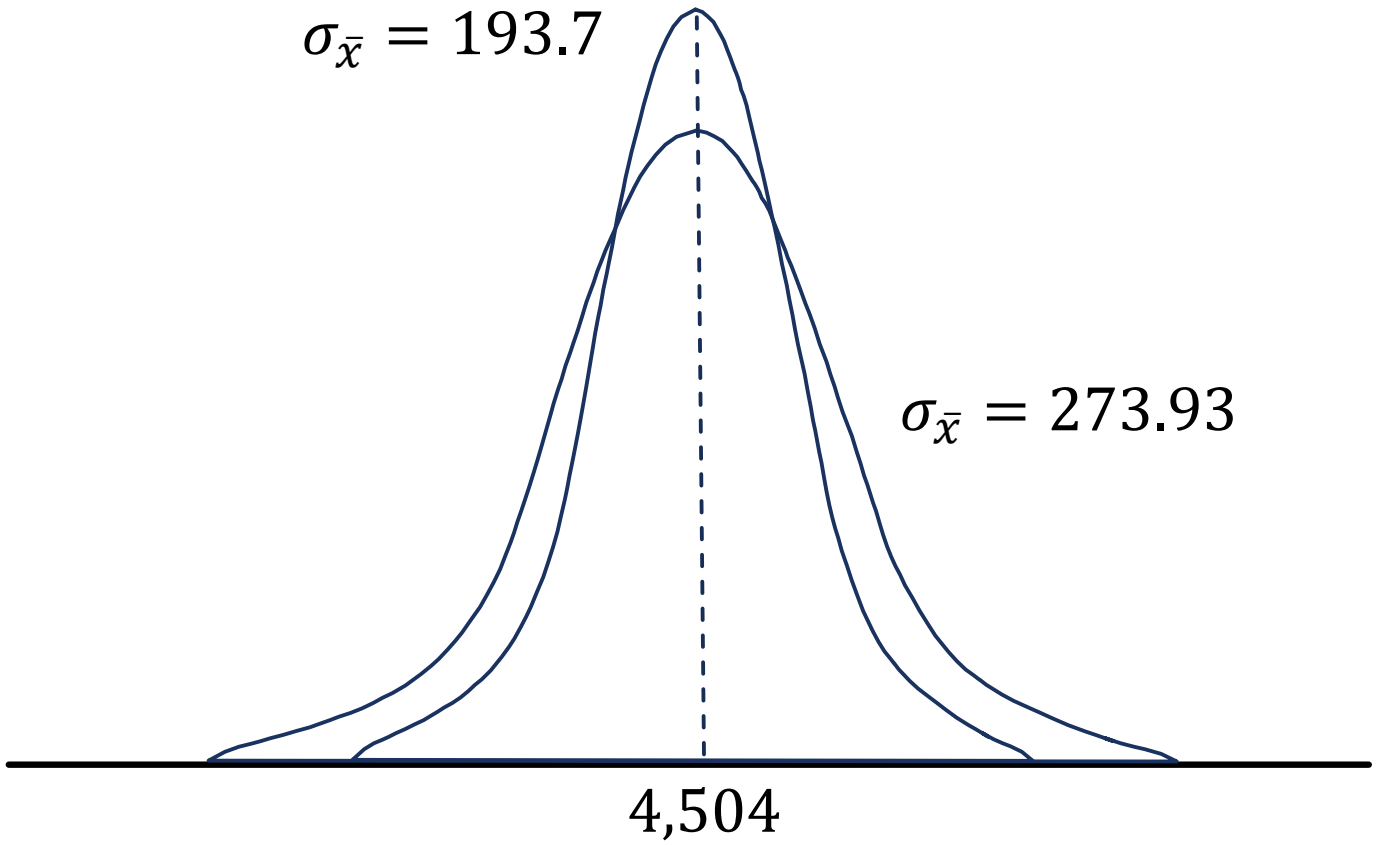


SAMPLE SIZE AND SAMPLING DISTRIBUTION

- Suppose we select a sample of size 100 days instead of 50.
- The expected value of $\bar{x}$ remains the same: $E(\bar{x}) = \mu = 4{,}504$.
- However, the standard error of $\bar{x}$ decreases:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1{,}937}{\sqrt{100}} = 193.7$$

SAMPLE SIZE AND SAMPLING DISTRIBUTION



SUMMARY

- The sampling distribution of $\bar{x}$ is the probability distribution of all the possible values of the sample mean $\bar{x}$.
- The sampling distribution of $\bar{x}$ has a mean (expected value) of μ and variance of $\frac{\sigma}{\sqrt{n}}$.
- If we use a large sample ($n \geq 50$), the Central Limit Theorem (CLT) states that the sampling distribution of $\bar{x}$ is approximately Normally distributed, regardless of the population distribution.

SAMPLING DISTRIBUTION FOR $\hat{p}$

DISTRIBUTIONS OF STATISTICS FROM DATA

PROPORTIONS

- Means are not the only thing of interest in a population.
- Another typical problem would be to estimate the proportion of the population, p, that has a certain attribute.
- Since we cannot view the whole population, we have to use the **sample proportion**, $\hat{p}$, for this estimate.

PROPORTIONS

- Means are not the only thing of interest in a population.
- Another typical problem would be to estimate the proportion of the population, p, that has a certain attribute.
- Since we cannot view the whole population, we have to use the **sample proportion**, $\hat{p}$, for this estimate.
- What is the sampling distribution of the sample proportion?

SAMPLING DISTRIBUTION OF $\hat{p}$

- Sample proportions are similar to sample means.

Customer ID	Gender	Gender Numeric
001	M	0
002	F	I
003	F	I
004	M	0
005	M	0

$$\hat{p}_F = \frac{2}{5} = 0.4$$

SAMPLING DISTRIBUTION OF $\hat{p}$

- Sample proportions are similar to sample means.

Customer ID	Gender	Gender Numeric
001	M	0
002	F	1
003	F	1
004	M	0
005	M	0

$$\hat{p}_F = \frac{2}{5} = 0.4$$

$$\bar{x} = \frac{0+1+1+0+0}{5}$$

$$= 0.4$$

SAMPLING DISTRIBUTION OF $\hat{p}$

- Sample proportions are similar to sample means.
- The **sampling distribution of $\hat{p}$** is approximately the **Normal distribution** whenever the sample size is large.
- How large is large enough?

$$np \geq 5$$

$$n(1-p) \geq 5$$

SAMPLING DISTRIBUTION OF $\hat{p}$

- Sample proportions are similar to sample means.
- The **sampling distribution of $\hat{p}$** is approximately the **Normal distribution** whenever the sample size is large.
- How large is large enough?

$$np \geq 5$$
$$n(1-p) \geq 5$$

← At least 5 in each of the two categories!

SAMPLING DISTRIBUTION OF $\hat{p}$

- How large is large enough?

$$np \geq 5$$

$$n(1-p) \geq 5$$

- For values of p near 0.5, sample sizes as small as 10 can afford a Normal approximation.
- With very small (approaching 0) or large (approaching 1) values of p, much larger samples are needed.

SAMPLING DISTRIBUTION

- The **sampling distribution of $\hat{p}$** is the probability distribution of all the possible values of the sample proportion $\hat{p}$.
- The **sampling distribution of $\hat{p}$** has a mean (expected value) and variance as well.

$$E(\hat{p}) = \mu_{\hat{p}} = p$$

$$SD(\hat{p}) = \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

SAMPLING DISTRIBUTION BIKE DATA EXAMPLE

- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. In your data, 63% of the days are clear or cloudy. What is the probability that you sample 50 days and less then half of them are clear or cloudy?

SAMPLING DISTRIBUTION BIKE DATA EXAMPLE

- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. In your data, 63% of the days are clear or cloudy. What is the probability that you sample 50 days and less then half of them are clear or cloudy?

$$50 \times 0.63 = 31.5 \geq 5$$

$$50(1 - 0.63) = 18.5 \geq 5$$

Z-SCORE $\hat{p}$

- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. In your data, 63% of the days are clear or cloudy. What is the probability that you sample 50 days and less then half of them are clear or cloudy?

$$z_{\hat{p}} = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$$

Z-SCORE $\hat{p}$

- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. In your data, 63% of the days are clear or cloudy. What is the probability that you sample 50 days and less then half of them are clear or cloudy?

$$z_{\hat{p}} = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Z-SCORE $\hat{p}$

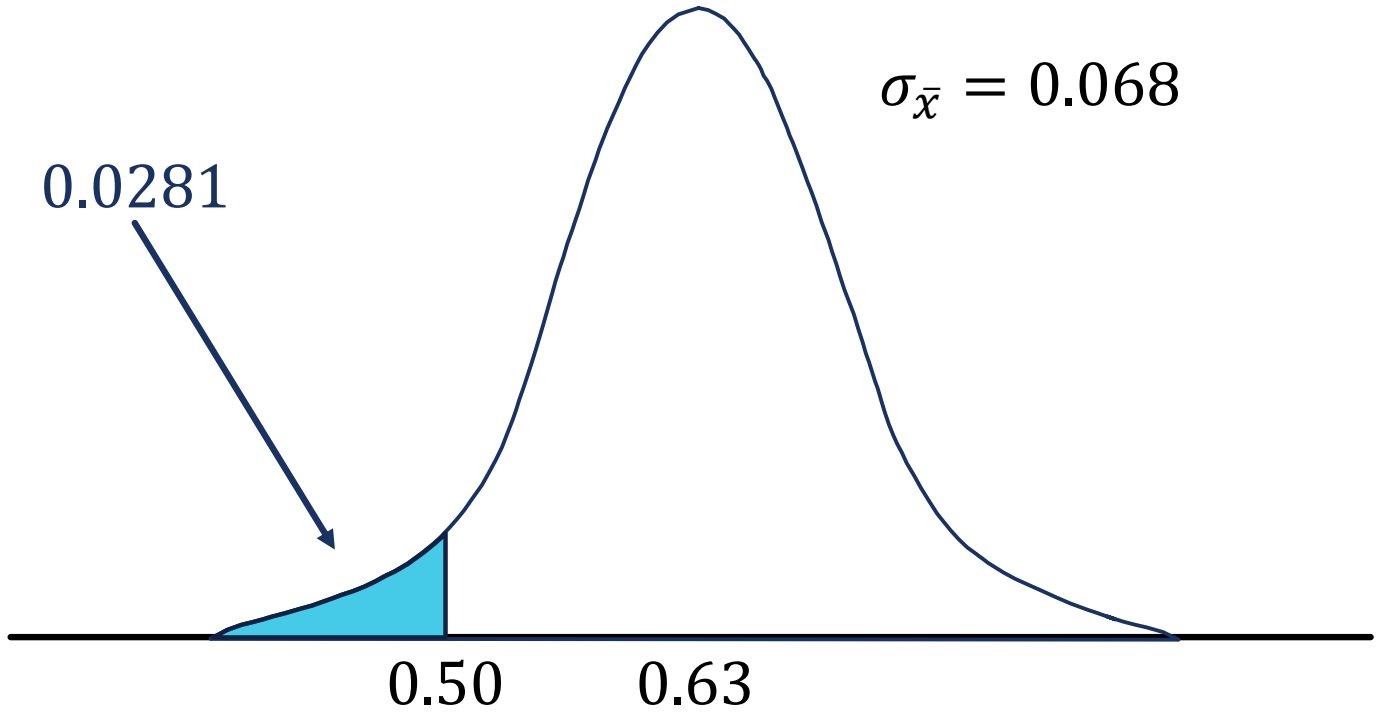
- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. In your data, 63% of the days are clear or cloudy. What is the probability that you sample 50 days and less then half of them are clear or cloudy?

$$z_{\hat{p}} = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.50 - 0.63}{\sqrt{\frac{0.63(1-0.63)}{50}}} = \frac{-0.13}{0.068} = -1.91$$

$$P\left(z_{\hat{p}} \leq -1.91\right) = 0.0281$$

Z-SCORE $\hat{p}$

- You think that people are more likely to rent a bike on a clear or cloudy day compared to misty / rain / snow. In your data, 63% of the days are clear or cloudy. What is the probability that you sample 50 days and less then half of them are clear or cloudy?



SUMMARY

- Another typical problem would be to estimate the proportion of the population that has a certain attribute, p, with the sample proportion, $\hat{p}$.
- The sampling distribution of $\hat{p}$ is approximately the Normal distribution whenever the sample size is large (both $np \geq 5$ and $n(1-p) \geq 5$).