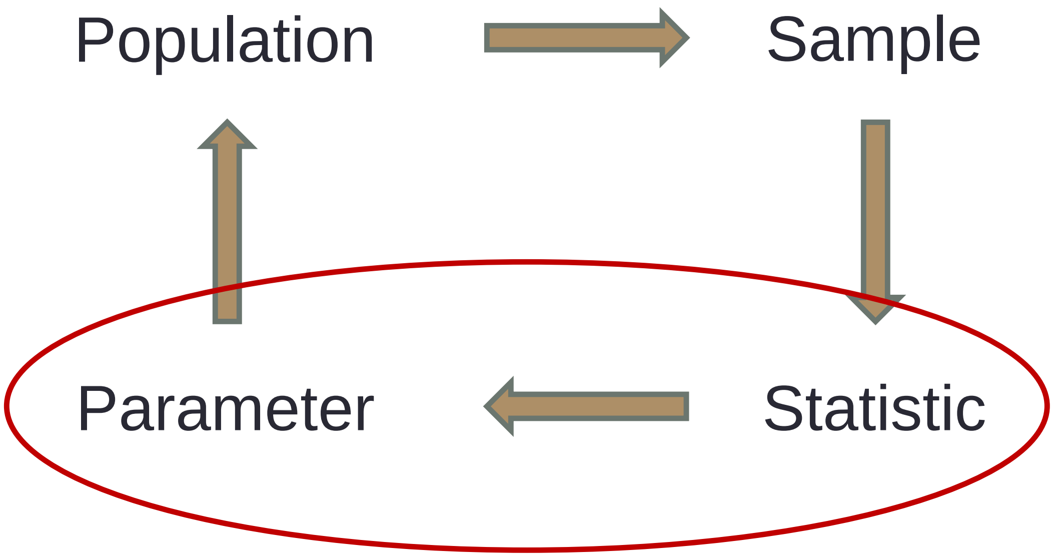
SAMPLING DISTRIBUTIONS

Analytics Primer

Parameters vs. Statistics

$$\begin{tikzcd}
\text{Population} \arrow[r] & \text{Sample} \arrow[d] \\
\text{Parameter} \arrow[u] & \text{Statistic} \arrow[l]
\end{tikzcd}$$

Parameters vs. Statistics



Parameters vs. Statistics

- **Parameter**
  - Measures computed from a population.

- **Statistic**
  - Measures computed from a sample.
  - Sample statistics is the **point estimate** of the population parameter.

Point Estimators

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

Point Estimators

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

Point Estimators

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

Point Estimators

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

Point Estimators

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

Point Estimators

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

Point Estimators

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

SAMPLING ERROR

Sampling Error

- Because statistics depend upon the sample obtained, calculated statistics will be different for different samples. However, statistics have a predictable pattern which is called the **sampling distribution**.
- A sampling distribution is a distribution of the **possible** values of a statistic for a given sample size selected from a population.

Sampling Error

- When the expected value of the sampling distribution is equal to the population parameter, the statistic (point estimator) is said to be **unbiased**.
- The absolute value of the difference between an unbiased point estimate and the corresponding population parameter is called the **sampling error**.
- Sampling error is the result of the sample being only a subset of the population.

Point Estimators

Point Estimator	Population Parameter	Sampling Error
$\bar{x}$	μ	$\lvert\bar{x} - \mu\rvert$
s^2	σ^2	$\lvert s^2 - \sigma^2\rvert$
$\hat{p}$	p	$\lvert\hat{p} - p\rvert$

Point Estimator Example

- You work for a major university admissions office and you want to know the distribution of the incoming student SAT scores as well as the proportion of students that want to live on campus.
- You sample 50 incoming students and find the following information:

Desired Information	Point Estimator
Avg. SAT Score	997
SAT Score SD	75.2
Prop. Living on Campus	0.68

Point Estimator Example

- You work for a major university admissions office and you want to know the distribution of the incoming student SAT scores as well as the proportion of students that want to live on campus.
- Later a census was taken of all students and the following population parameter information was calculated:

Desired Information	Point Estimator	Population Parameter
Avg. SAT Score	997	990
SAT Score SD	75.2	80
Prop. Living on Campus	0.68	0.72

Point Estimator Example

- You work for a major university admissions office and you want to know the distribution of the incoming student SAT scores as well as the proportion of students that want to live on campus.
- From this information we can calculate sampling error:

Point Estimator	Population Parameter	Sampling Error
997	990	7
75.2	80	4.8
0.68	0.72	0.04

Point Estimator Example

- You work for a major university admissions office and you want to know the distribution of the incoming student SAT scores as well as the proportion of students that want to live on campus.
- From this information we can calculate sampling error:

Point Estimator	Population Parameter	Sampling Error
997	990	7
75.2	80	4.8
0.68	0.72	0.04

Point Estimator Example

- You work for a major university admissions office and you want to know the distribution of the incoming student SAT scores as well as the proportion of students that want to live on campus.
- From this information we can calculate sampling error:

Point Estimator	Population Parameter	Sampling Error
997	990	7
75.2	80	4.8
0.68	0.72	0.04

Point Estimator Example

- You work for a major university admissions office and you want to know the distribution of the incoming student SAT scores as well as the proportion of students that want to live on campus.
- From this information we can calculate sampling error:

Point Estimator	Population Parameter	Sampling Error
997	990	7
75.2	80	4.8
0.68	0.72	0.04

Sampling Error

- Most of the time we do not have the ability to collect information from the whole sample to see what kind of sampling error we actually have.
- For that reason, it would be nice to know if there is a common pattern/distribution for the point estimates and therefore the sampling errors of these point estimates.

SAMPLING DISTRIBUTION FOR $\bar{x}$

Point Estimators

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

Sampling Distribution

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

- See: http://onlinestatbook.com/stat_sim/sampling_dist/

Sampling Distribution

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

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

Sampling Distribution

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

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

Sampling Distribution

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

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

Sampling Distribution

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

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

μ is the true mean of the population!

Sampling Distribution

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

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

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

Sampling Distribution

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

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

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

Notice that the ***Standard Deviation of x-bar*** is the true population standard deviation divided by square root of n (depends on sample size!)

Central Limit Theorem

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

- Sampling distribution of the sample mean:

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

Central Limit Theorem

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

- Sampling distribution of the sample mean:

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

Central Limit Theorem

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

- Sampling distribution of the sample mean:

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

Central Limit Theorem

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

- Sampling distribution of the sample mean:

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

Central Limit Theorem

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

- Sampling distribution of the sample mean:

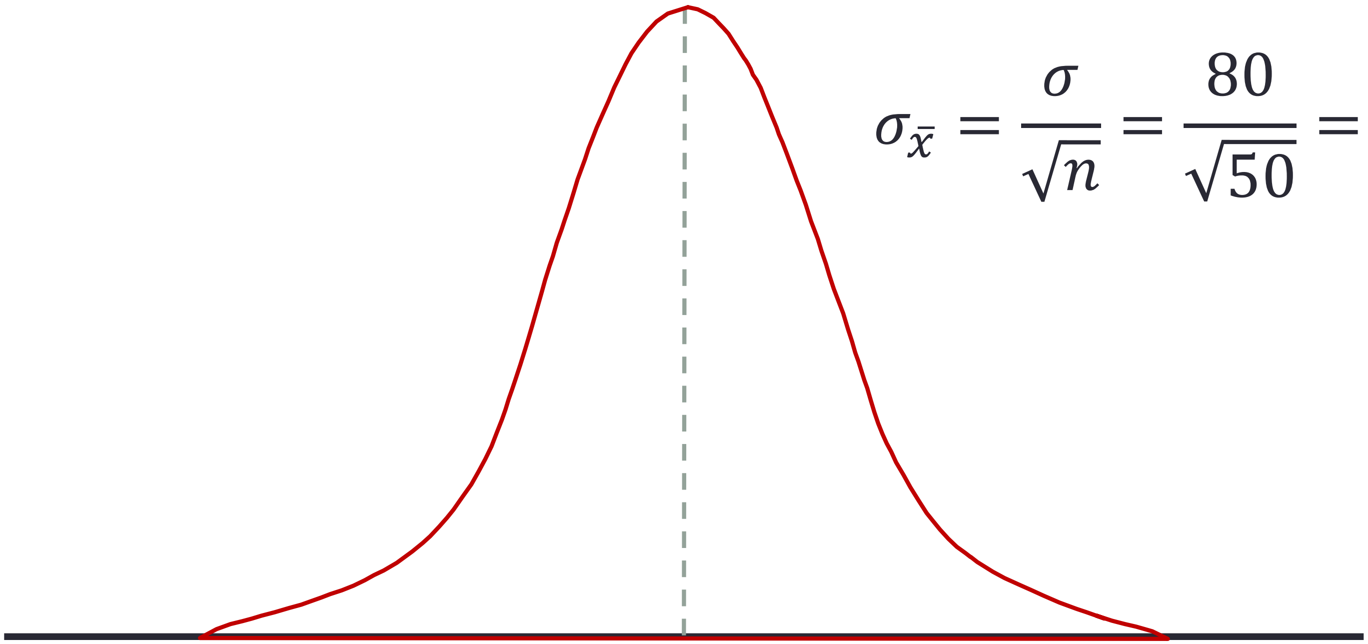
$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

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 Example

- Based on our example of SAT scores at a major university, all of the possible sample means (from samples of size 50) would have the following distribution:



$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{80}{\sqrt{50}} = 11.3$$

$$E(\bar{x}) = \mu = 990$$

Sampling Distribution Example

- Based on our example of SAT scores at a major university, all of the possible sample means (from samples of size 50) would have the following distribution:

$$\bar{x} \sim N(990, 11.3)$$

Sampling Distribution Example

- Based on our example of SAT scores at a major university(from samples of size 50)
- What is the probability that a **sample of 50 applicants** would have an **average** SAT score between 980 and 1000?

Z-Score for $\bar{x}$

- Based on our example of SAT scores at a major university (from samples of size 50)
- What is the probability that a **sample of 50 applicants** would have an **average** SAT score between 980 and 1000?

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

Z-Score for $\bar{x}$

- Based on our example of SAT scores at a major university (from samples of size 50)
- What is the probability that a **sample of 50 applicants** would have an **average** SAT score between 980 and 1000?

$$\bar{x} \sim N(990, 11.3)$$

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

Z-Score for $\bar{x}$ Example

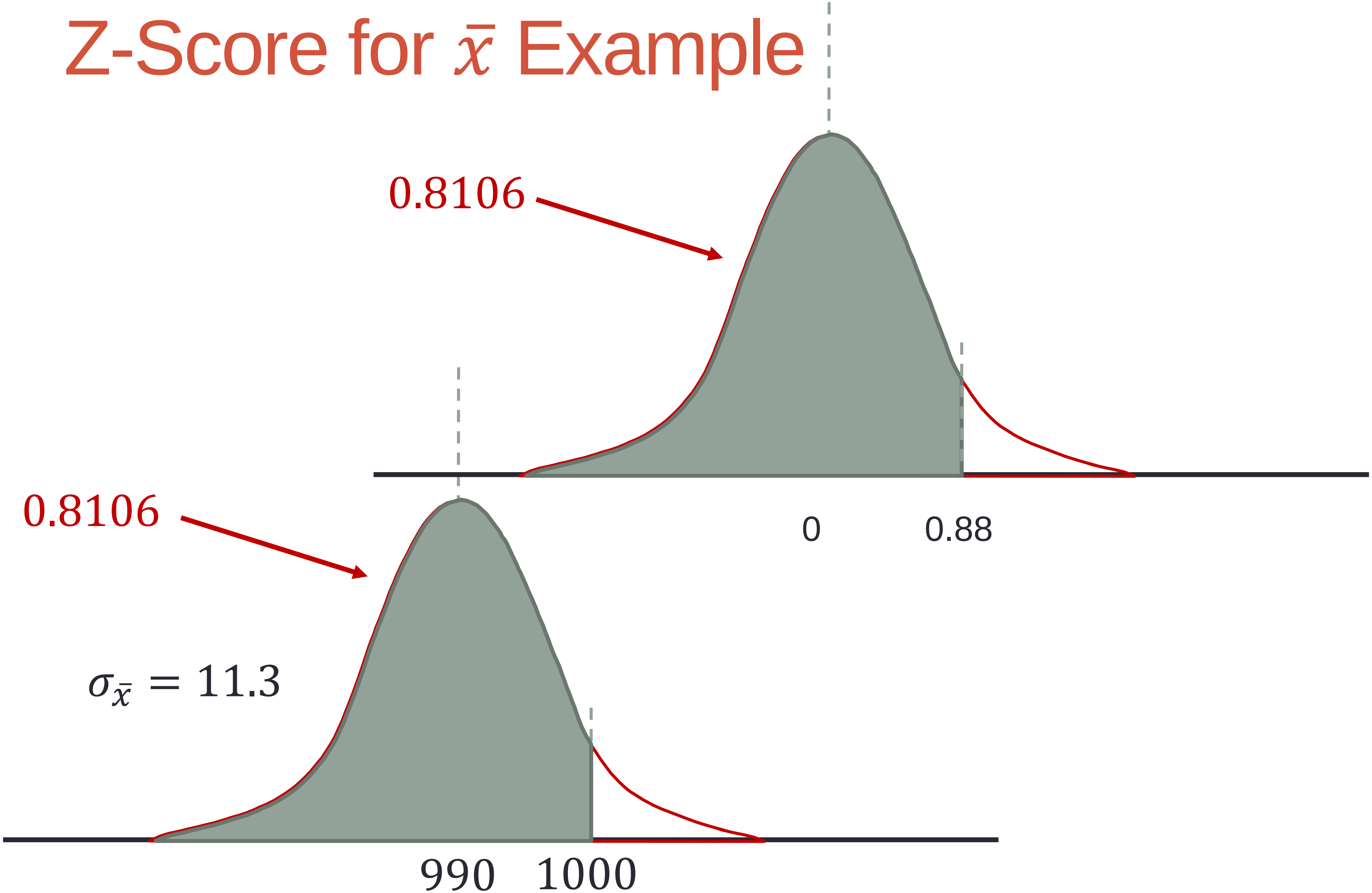
- What is the probability that a **sample of 50 applicants** would have an **average** SAT score between 980 and 1000?

$$z_{\bar{x}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{1000 - 990}{\frac{80}{\sqrt{50}}} = \frac{10}{11.3} = 0.88$$

$$P(z_{\bar{x}} \leq 0.88) = 0.8106$$

pnorm(0.88)

Z-Score for $\bar{x}$ Example



Z-Score for $\bar{x}$ Example

- What is the probability that a **sample of 50 applicants** would have an **average** SAT score between 980 and 1000?

$$z_{\bar{x}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{1000 - 990}{\frac{80}{\sqrt{50}}} = \frac{10}{11.3} = 0.88$$

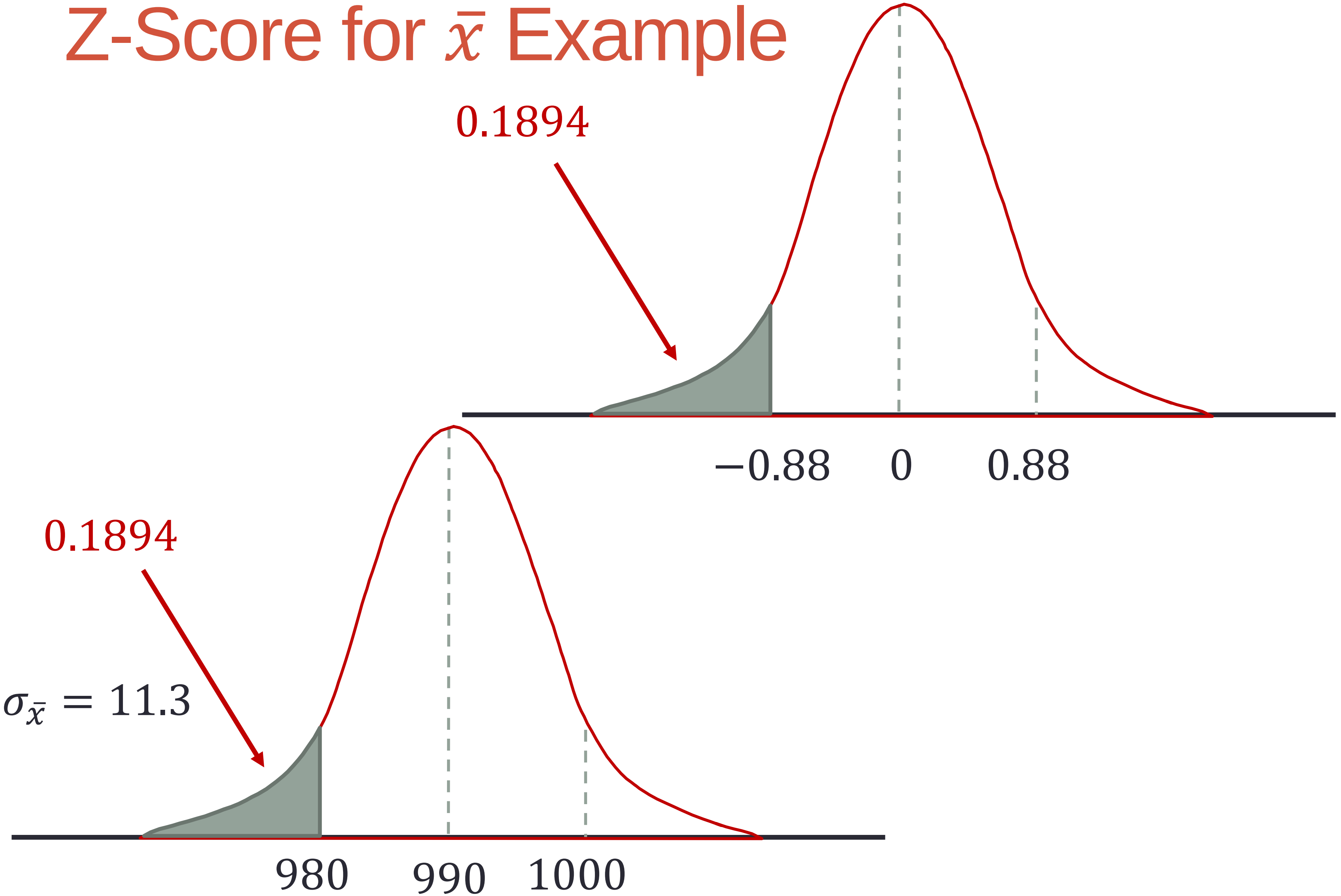
$$P(z_{\bar{x}} \leq 0.88) = 0.8106$$

$$z_{\bar{x}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{980 - 990}{\frac{80}{\sqrt{50}}} = \frac{-10}{11.3} = -0.88$$

$$P(z_{\bar{x}} \leq -0.88) = 0.1894$$

pnorm(-0.88)

Z-Score for $\bar{x}$ Example



Z-Score for $\bar{x}$ Example

- What is the probability that a **sample of 50 applicants** would have an **average** SAT score between 980 and 1000?

$$P(-0.88 \le z_{\bar{x}} \le 0.88) = 0.8106 - 0.1894$$

$$= 0.6212$$

Z-Score for $\bar{x}$ Example

- What is the probability that a **sample of 50 applicants** would have an **average** SAT score between 980 and 1000?

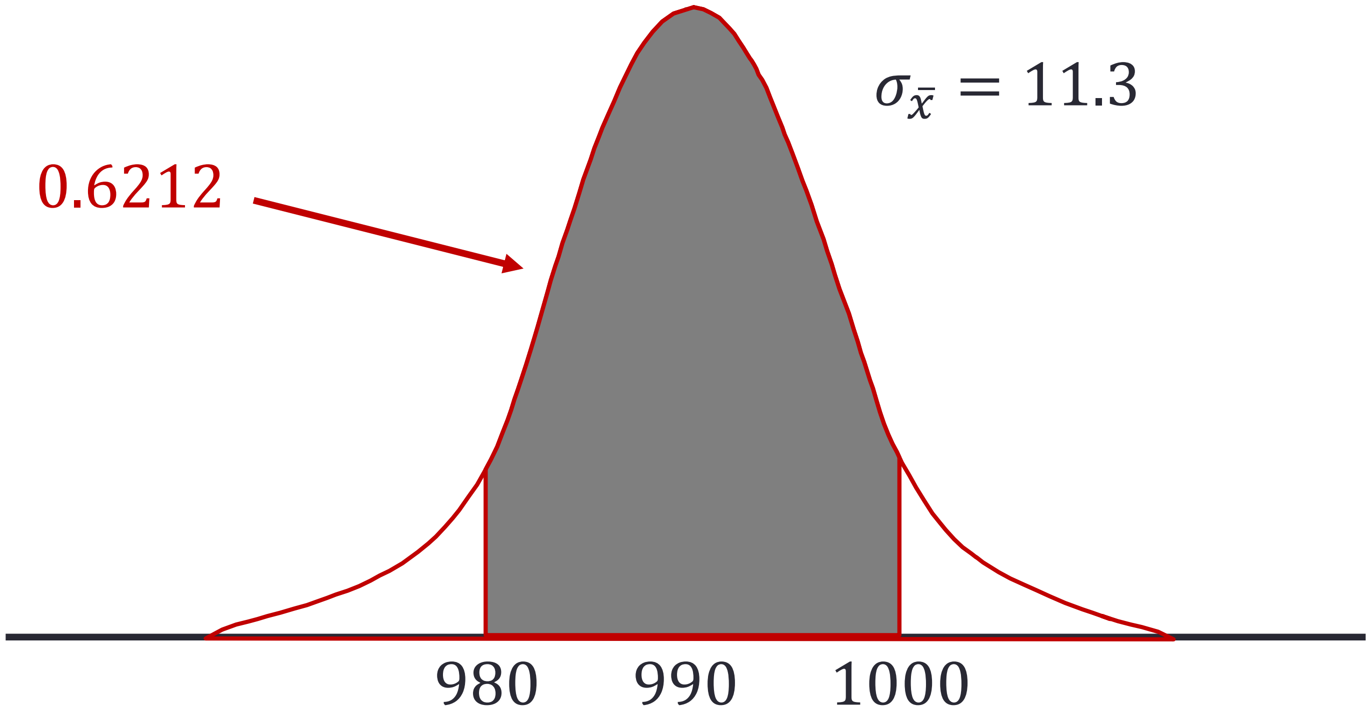
$$P(-0.88 \le z_{\bar{x}} \le 0.88) = 0.8106 - 0.1894$$

$$= 0.6212$$

$$P(980 \le \bar{x} \le 1000) = 0.6212$$

Z-Score for $\bar{x}$ Example

- What is the probability that a **sample of 50 applicants** would have an **average** SAT score between 980 and 1000?

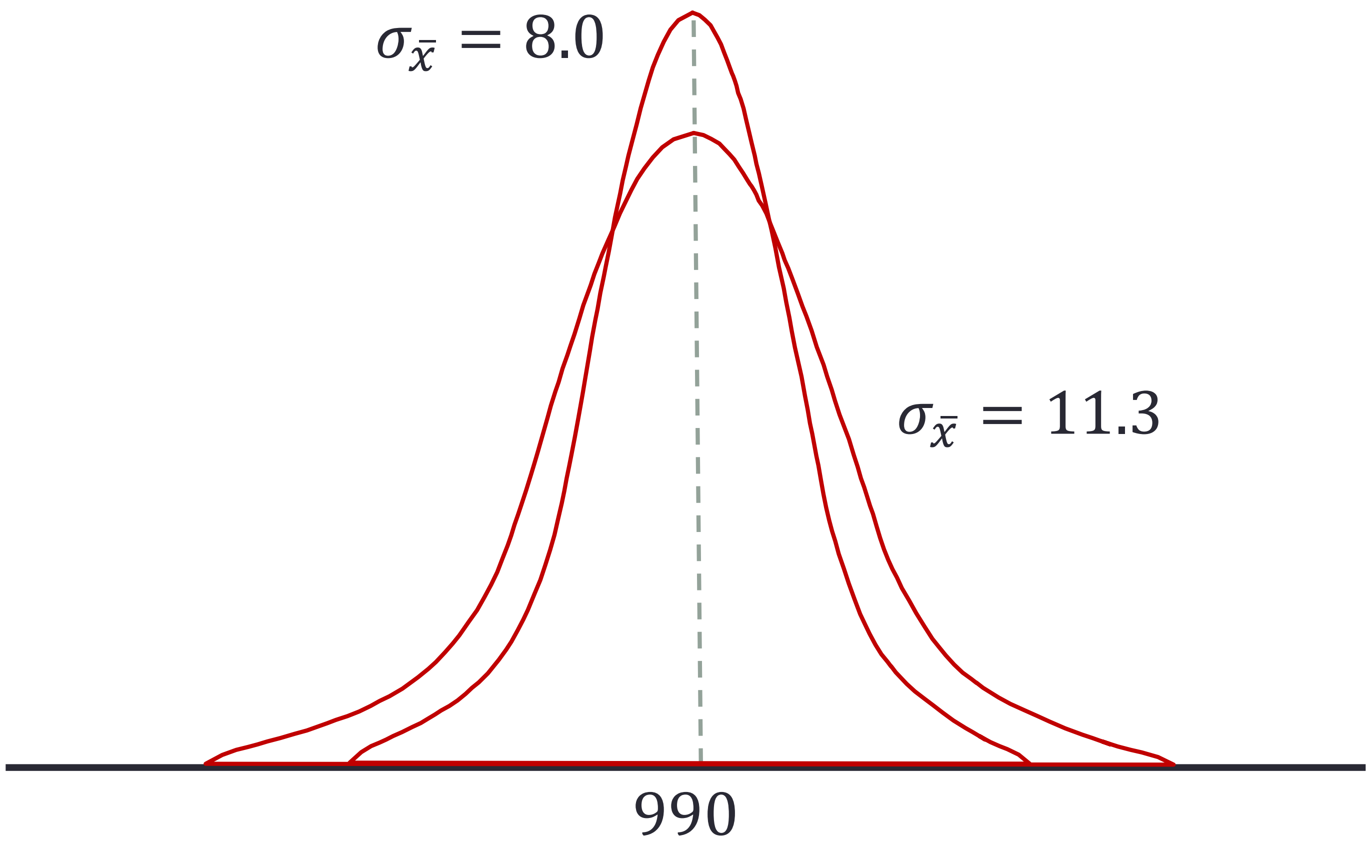


Sample Size and Sampling Distribution

- Suppose we select a sample of size 100 applicants instead of 50.
- The expected value of $\bar{x}$ remains the same: $E(\bar{x}) = \mu = 990$.
- However, the standard error of $\bar{x}$ decreases:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{80}{\sqrt{100}} = 8.0$$

Sample Size and Sampling Distribution



Example

- Assume the average daily number of web page hits a company gets follows a normal distribution with a mean of 2341.36 and s.d. of 516.79. What is the probability that a sample of 49 days over the past year has an average web page hit above 2500?
- How about a sample of 121 days instead?
- What if the distribution wasn’t normal?

Example

- Assume that I own a chain of retail stores located at major cities across the country. The daily sales in thousands of dollars at each store has a mean of 17.06 and a s.d. of 5.12. What is the probability that a sample of 64 of my stores averages sales of more than $19K?
- I am worried about one of my managers performance in retail sales. He manages 100 of my stores and they only average $14.35K in sales per day. What is the probability I randomly select 100 of my stores and get sales numbers that low?

Example

- Assume the average daily number of web page hits a company gets follows a normal distribution with a mean of 2341.36 and s.d. of 516.79. What is the probability that a sample of 49 days over the past year has an average web page hit above 2500?

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{2500 - 2341.36}{\frac{516.79}{\sqrt{49}}} = 2.15 \rightarrow 0.0158$$

Example

- Assume the average daily number of web page hits a company gets follows a normal distribution with a mean of 2341.36 and s.d. of 516.79. What is the probability that a sample of 49 days over the past year has an average web page hit above 2500?
- How about a sample of 121 days instead?

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{2500 - 2341.36}{\frac{516.79}{\sqrt{121}}} = 3.38 \rightarrow 0.0004$$

Example

- Assume the average daily number of web page hits a company gets follows a normal distribution with a mean of 2341.36 and s.d. of 516.79. What is the probability that a sample of 49 days over the past year has an average web page hit above 2500?
- How about a sample of 121 days instead?
- What if the distribution wasn’t normal? Worried for sample of 49 that our results weren’t valid. Not worried for sample of 121.

Example

- Assume that I own a chain of retail stores located at major cities across the country. The daily sales in thousands of dollars at each store has a mean of 17.06 and a s.d. of 5.12. What is the probability that a sample of 64 of my stores averages sales of more than $19K?

Example

- Assume that I own a chain of retail stores located at major cities across the country. The daily sales in thousands of dollars at each store has a mean of 17.06 and a s.d. of 5.12. What is the probability that a sample of 64 of my stores averages sales of more than $19K?

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{19 - 17.06}{\frac{5.12}{\sqrt{64}}} = 3.03 \rightarrow 0.0012$$

Example

- Assume that I own a chain of retail stores located at major cities across the country. The daily sales in thousands of dollars at each store has a mean of 17.06 and a s.d. of 5.12. What is the probability that a sample of 64 of my stores averages sales of more than $19K?
- I am worried about one of my managers performance in retail sales. He manages 100 of my stores and they only average $14.35K in sales per day. What is the probability I randomly select 100 of my stores and get sales numbers that low?

Example

- Assume that I own a chain of retail stores located at major cities across the country. The daily sales in thousands of dollars at each store has a mean of 17.06 and a s.d. of 5.12. What is the probability that a sample of 64 of my stores averages sales of more than $19K?
- I am worried about one of my managers performance in retail sales. He manages 100 of my stores and they only average $14.35K in sales per day. What is the probability I randomly select 100 of my stores and get sales numbers that low?

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{14.35 - 17.06}{\frac{5.12}{\sqrt{100}}} = -5.29 \rightarrow \approx 0$$

SAMPLING DISTRIBUTION FOR $\hat{p}$

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 (or sometimes referred to as π), 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 (or sometimes referred to as π), 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	1
003	F	1
004	M	0
005	M	0

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

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

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

$$\hat{p}_F = \frac{2}{5} = 0.4 \qquad \bar{x} = \frac{0+1+1+0+0}{5}$$

$$= 0.4$$

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?

$$np \geq 5$$

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

Number of successes and failures both at least 5.

Sampling Distribution of $\hat{p}$

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

$$np \geq 5$$

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

- 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 (≈50).

Sampling Distribution

- The **sampling distribution of $\boldsymbol{\hat{p}}$** is the probability distribution of all the possible values of the sample proportion $\hat{p}$.

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

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

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

Sampling Distribution Example

- You work for a major university admissions office and you want to know the proportion of students that want to live on campus.
- You sample **50 incoming students** and want to know the probability that the proportion of students who want to live on campus is **between 0.65 and 0.75**.

Sampling Distribution Example

- You work for a major university admissions office and you want to know the proportion of students that want to live on campus.
- You sample **50 incoming students** and want to know the probability that the proportion of students who want to live on campus is **between 0.65 and 0.75**.

$$50 \times 0.72 = 36 \geq 5$$

$$50(1 - 0.72) = 14 \geq 5$$

Z-Score $\hat{p}$

- You work for a major university admissions office and you want to know the proportion of students that want to live on campus.
- You sample **50 incoming students** and want to know the probability that the proportion of students who want to live on campus is **between 0.65 and 0.75**.

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

Z-Score $\hat{p}$

- You sample **50 incoming students** and want to know the probability that the proportion of students who want to live on campus is **between 0.65 and 0.75**.

$$z_{\hat{p}} = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.65 - 0.72}{\sqrt{\frac{0.72(1-0.72)}{50}}} = \frac{-0.07}{0.064} = -1.09$$

$$P\left(z_{\hat{p}} \leq -1.09\right) = 0.1379$$

pnorm(-1.09)

Z-Score $\hat{p}$

- You sample **50 incoming students** and want to know the probability that the proportion of students who want to live on campus is **between 0.65 and 0.75**.

$$z_{\hat{p}} = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.65 - 0.72}{\sqrt{\frac{0.72(1-0.72)}{50}}} = \frac{-0.07}{0.064} = -1.09$$

$$P\left(z_{\hat{p}} \leq -1.09\right) = 0.1379$$

$$z_{\hat{p}} = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.75 - 0.72}{\sqrt{\frac{0.72(1-0.72)}{50}}} = \frac{0.03}{0.064} = 0.47$$

$$P\left(z_{\hat{p}} \leq 0.47\right) = 0.6808$$

pnorm(0.47)

Z-Score $\hat{p}$

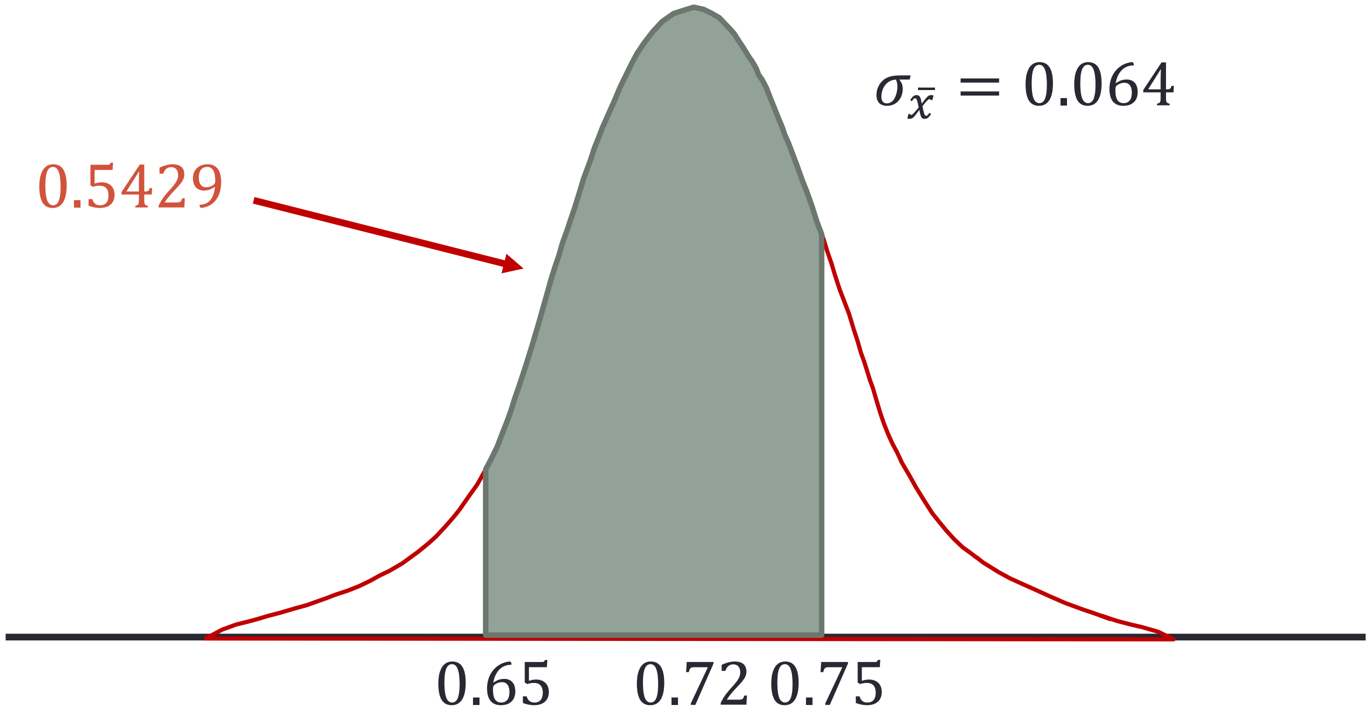
- You sample **50 incoming students** and want to know the probability that the proportion of students who want to live on campus is **between 0.65 and 0.75**.

$$P\left(-1.09 \le z_{\hat{p}} \le 0.47\right) = 0.6808 - 0.1379$$
$$= 0.5429$$

$$P(0.65 \le \hat{p} \le 0.75) = 0.5429$$

Z-Score $\hat{p}$

- You sample **50 incoming students** and want to know the probability that the proportion of students who want to live on campus is **between 0.65 and 0.75**.



Example

- The NC Board of Education is interested in gathering information about the drop out rate of high school students across the state of North Carolina. The proportion of high school students that drop out is 5.24% across the state. What is the probability that less than 8 out 169 random students across the state drop out of high school?

Example

- The NC Board of Education is interested in gathering information about the drop out rate of high school students across the state of North Carolina. The proportion of high school students that drop out is 5.24% across the state. What is the probability that less than 8 out 169 random students across the state drop out of high school?

$$\hat{p} = \frac{8}{169} = 0.0473$$

$$p = 0.0524$$

$$z_{\hat{p}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.0473 - 0.0524}{\sqrt{\frac{0.0524(1-0.0524)}{169}}} = -0.3 \rightarrow 0.3821$$