Quantitative Methods – I (Statistics)

A. Y. 2022-23

Prof. Lorenzo Cavallo

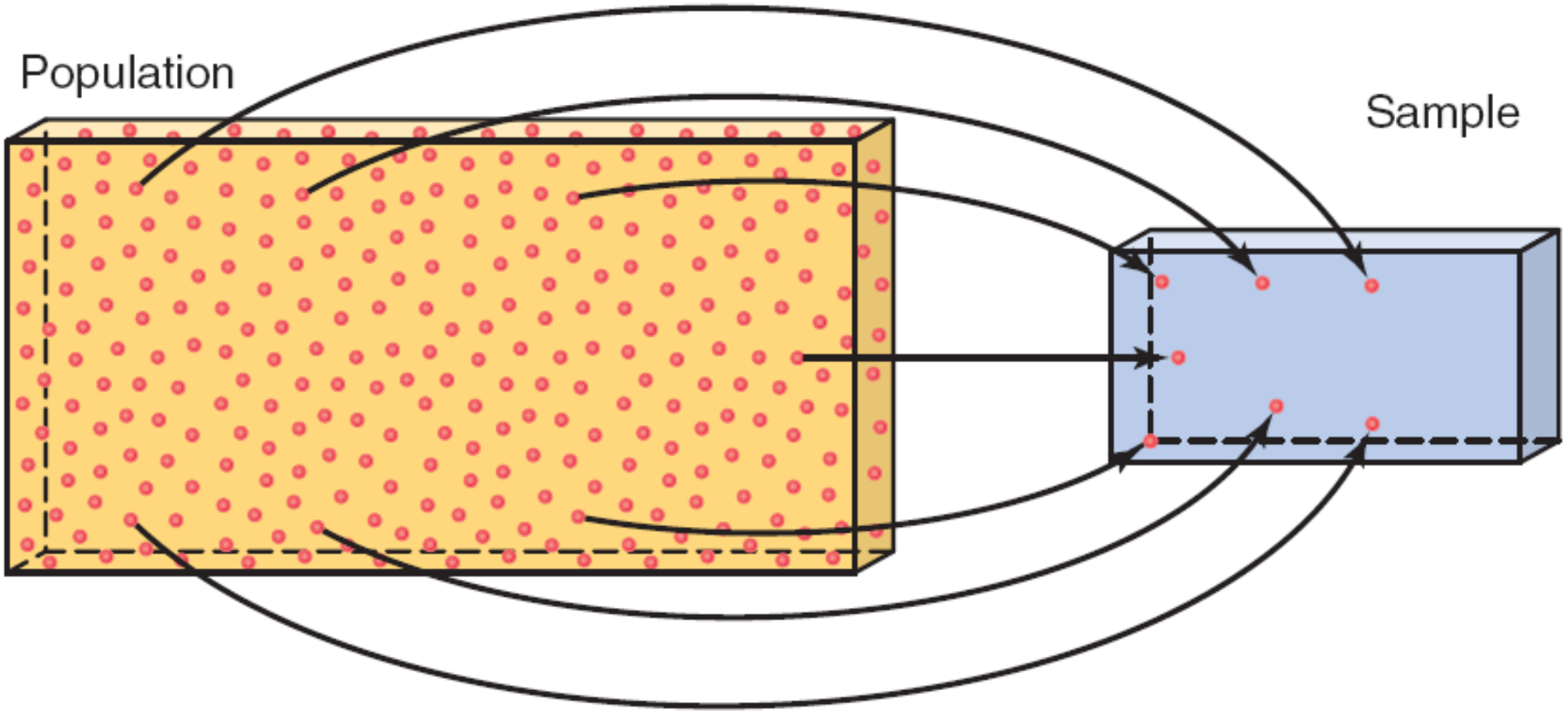
Chapter 7

Sampling Distributions

Sampling Distributions: Outline

1. Definitions: sampling distributions, sampling errors, non-sampling errors
2. Sample mean: $\bar{X}$
   - expected value, variance, and shape of the sampling distribution
3. Sample proportion: $\hat{p}$
   - expected value, variance, and shape of the sampling distribution

Recall: Population vs Samples



Recall: Population vs Sample

Population vs ***Sample***

The set of all the units/subjects of interest

EXAMPLE: **All** residents in NY

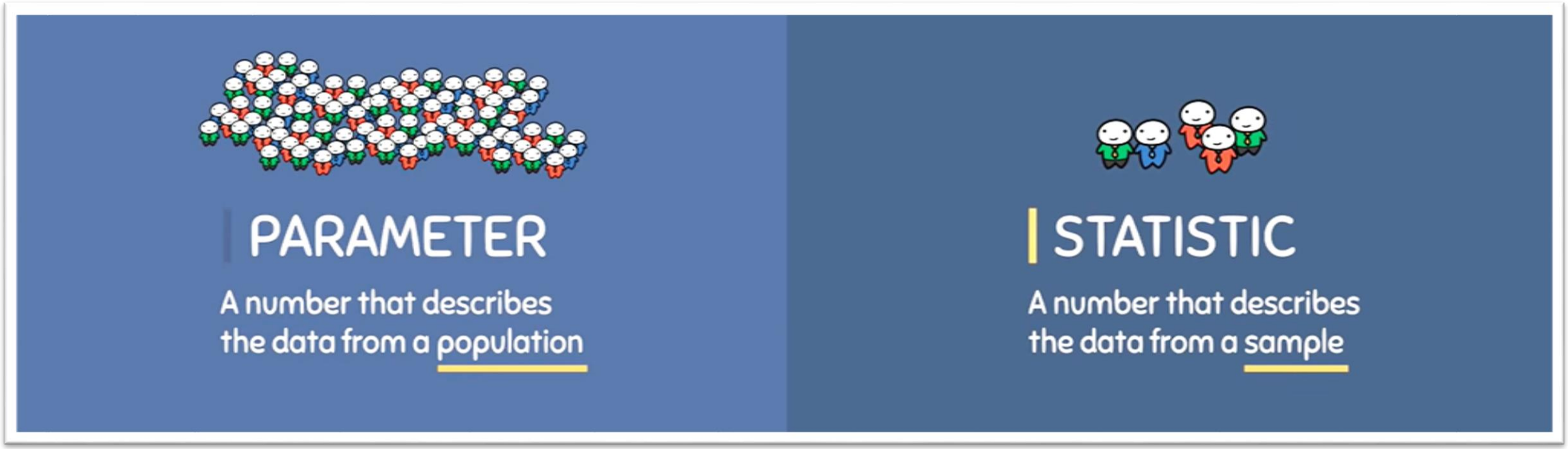
- The population is a compete set.
- It contains all members of a specified group.
- The measurable quality is called a parameter.
- Reports are a true rapresentation of options.

A subset of the units/subjects of interest

EXAMPLE: **A subset of** residents in NY

- The sample is a subset of the population.
- It is a subset that represents the entire population.
- The measurable quality is called a statistic.
- Reports have margin of error and confidence interval.

Recall: parameter vs statistic



Recall: parameter vs statistic

Population vs ***Sample***

The set of all the units/subjects of interest

EXAMPLE: **All** residents in NY

A subset of the units/subjects of interest

EXAMPLE: **A subset of** residents in NY

Parameter vs ***Statistic***

Numerical summary **of the population**

EXAMPLE: Average number of books read last year **by ALL residents in NY**

Numerical summary **of the sample**

EXAMPLE: Average number of books read last year **by 1000 selected residents in NY**

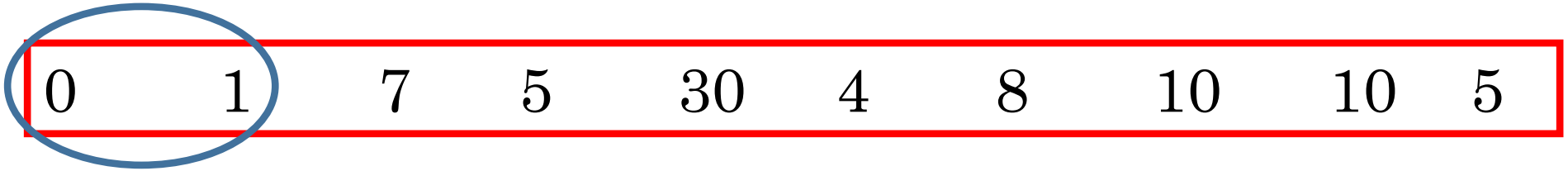
Recall: parameter vs statistic

	SAMPLE	POPULATION	
MEAN	$\bar{x}$	μ	
STANDARD DEVIATION	s	σ	
	STATISTIC	PARAMETER	

Recall: parameter vs sample statistics

Many samples might generate from the same population ➔ many sample statistics might result.

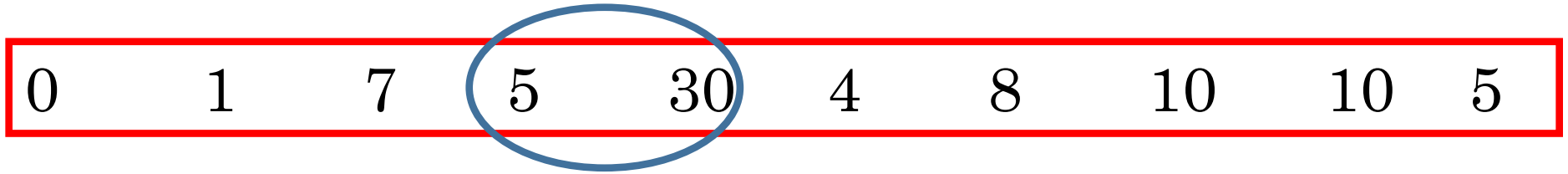
Ex: Suppose red square contains the population. Parameter of interest is average nr of books read last year. ➔ parameter is: 8



..and this possible sample ➔ sample statistics is: 0.5

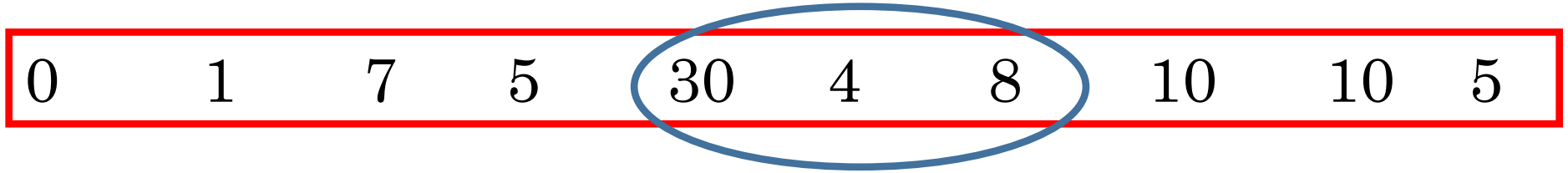
Recall: parameter vs sample statistics

Same population ➔ parameter is: 8



..but different sample ➔ sample statistics is: 17.5

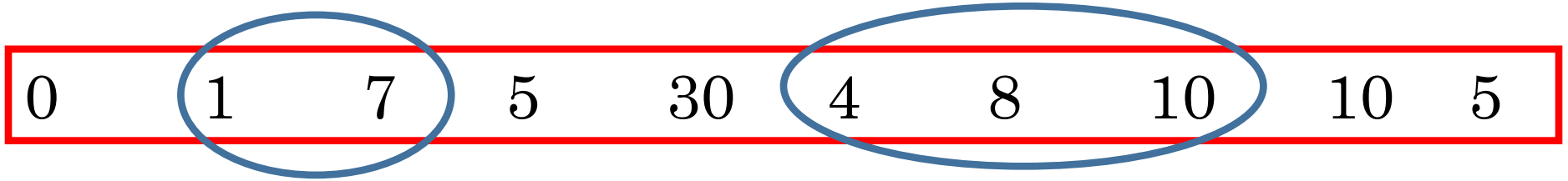
Same population ➔ parameter is: 8



..but different sample ➔ sample statistics is: 14

Sampling distribution: definition

Same population ➔ parameter is: 8



..but different sample ➔ sample statistics is: 6

➔Population parameter ***constant*** (at a given point in time), while sample statistic ***changes***

➔Sample statistics is a random variable with its own probability distribution, named ***sampling distribution***.

Sampling Distribution: example

Take a class of 5 students, and let X= final evaluation (over 100)

Student	A	B	C	D	E
X	70	78	80	80	95

In this population, average evaluation is 80.6, with st.dev. 8.09

Consider samples of 3 randomly selected students.

- How many?
- Which ones?:

Sampling Distribution: example

Take a class of 5 students, and let X= final evaluation (over 100)

Student	A	B	C	D	E
X	70	78	80	80	95

In this population, average evaluation is 80.6, with st.dev. 8.09

Consider samples of 3 randomly selected students.

- How many? $\binom{5}{3} = 10$
- Which ones?: ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE

Sampling Distribution: example

Sample	Scores in the Sample	$\bar{x}$
ABC	70, 78, 80	76.00
ABD	70, 78, 80	76.00
ABE	70, 78, 95	81.00
ACD	70, 80, 80	76.67
ACE	70, 80, 95	81.67
ADE	70, 80, 95	81.67
BCD	78, 80, 80	79.33
BCE	78, 80, 95	84.33
BDE	78, 80, 95	84.33
CDE	80, 80, 95	85.00

➔

Sampling distribution of $\bar{x}$, with $n = 3$

$\bar{x}$	$P(\bar{x})$
76.00	.20
76.67	.10
79.33	.10
81.00	.10
81.67	.20
84.33	.20
85.00	.10
	$\Sigma P(\bar{x}) = 1.00$

Sampling and non-sampling errors

Non-sampling errors: done during collection, recording, and tabulation of data

Sampling error: difference between the value of sample statistics and the value of the population parameter.

For the mean:

$$\text{Sampling error} = \bar{x} - \mu$$

(assuming that no non-sampling errors are made)

Non-sampling errors: possible causes

1. Non-random (and hence non-representative) sample. E.g.

   Landon vs Roosvelt in 1936 US presidential elections (Youtube)

2. Not fully understood questions (and hence wrong answers)

3. Respondents intentionally give false information to sensitive questions (e.g. on income, wealth, bad or risky habits). See e.g.:

   The under-reporting of financial wealth in the Survey on Household Income and Wealth (paper)

4. Mistakes in entering the answers while taking the interview or on the computer afterwards

Sampling errors: possible causes

Just chance!

They are due to the use of samples rather than the population. And only chance decides which sample will finally be drawn.

Sample	Scores in the Sample	$\bar{x}$
ABC	70, 78, 80	76.00
ABD	70, 78, 80	76.00
ABE	70, 78, 95	81.00
ACD	70, 80, 80	76.67
ACE	70, 80, 95	81.67
ADE	70, 80, 95	81.67
BCD	78, 80, 80	79.33
BCE	78, 80, 95	84.33
BDE	78, 80, 95	84.33
CDE	80, 80, 95	85.00

Suppose that the selected sample is drawn among those with $n = 3$:

➔ sampling error is $\bar{x} - \mu = 81.67 - 80.6 = 1.07$

Estimator

Parameter

Sampling and non-sampling errors

$$\text{Sampling error} = \bar{x} - \mu$$

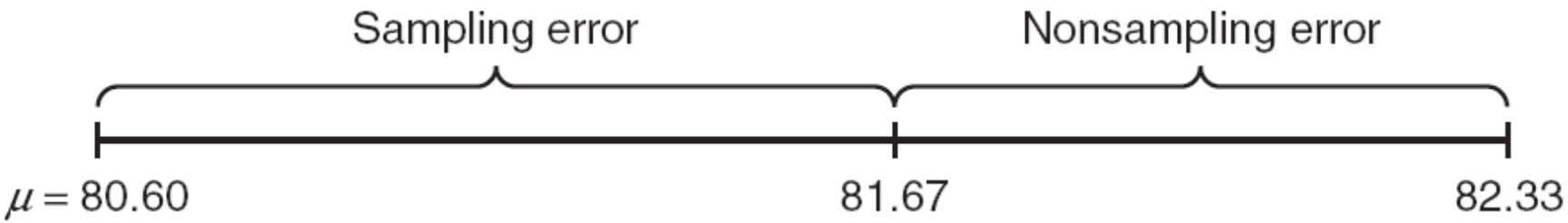
In our example, $\mu = 80.6$ ➔sampling errors are:

$\bar{x}$	$\boldsymbol{sampling\ error = \bar{x} - \mu}$
76.00	-4.60
76.67	-3.93
79.33	-1.27
81.00	0.40
81.67	1.07
84.33	3.73
85.00	4.40

Non-sampling and sampling errors

Suppose that, with $n = 3$, we draw the selected sample, but value 80 is wrongly recoded as 82 ➔ $\bar{x} = 82.33$. Only one portion of this is error is sampling error…

Sample	Scores in the Sample	$\bar{x}$
ABC	70, 78, 80	76.00
ABD	70, 78, 80	76.00
ABE	70, 78, 95	81.00
ACD	70, 80, 80	76.67
ACE	70, 80, 95	81.67
ADE	70, 80, 95	81.67
BCD	78, 80, 80	79.33
BCE	78, 80, 95	84.33
BDE	78, 80, 95	84.33
CDE	80, 80, 95	85.00



Note: since we do not know μ, we actually cannot quantify sampling and non-sampling errors

Mean and Standard Deviation of $\overline{x}$

Expected value of the sampling distribution of $\overline{x}$, denoted by

$$\mu_{\overline{x}} \; or \; E(x)$$

Standard deviation of the sampling distribution of $\overline{x}$, denoted by

$$\sigma_{\overline{x}}$$

It is also called **standard error**

Mean and Standard Deviation of $\overline{x}$

Expected value of the sampling distribution of $\overline{x}$, denoted by $\mu_{\overline{x}}$ or E(x)

In our example:

$$\mu_{\overline{x}} = \sum \bar{x} P(\bar{x}) = \mu = 80.6$$

The expected value of $\overline{x}$ coincides with the parameter μ

$\overline{x}$	$P(\overline{x})$	$\overline{x}P(\overline{x})$	$\overline{x}^2 P(\overline{x})$
76.00	0.20	15.20	1155.20
76.67	0.10	7.67	587.83
79.33	0.10	7.93	629.32
81.00	0.10	8.10	656.10
81.67	0.20	16.33	1334.00
84.33	0.20	16.87	1422.31
85.00	0.10	8.50	722.50
Total	1.00	80.60	6507.26

$\bar{x}$ is an Unbiased Estimator for μ

If the expected value (or mean) of an estimator is equal to the value of the parameter of interest ➔ the estimator is said to be **unbiased**.

In our example, this condition is met since $E(\bar{x}) = \mu_{\bar{x}} = \mu = 80.6$.

More generally

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

➔$\bar{x}$ is an **unbiased estimator** for μ

Standard dev. of $\bar{x}$ and Standard Error

In our example,

$$\mathrm{Var}(\bar{x}) = \sum x^2 P(\bar{x}) - (E(x))^2 = 6527.06 - 80.6^2 = 10.9$$

So,

$$St.dev(\bar{x}) = \sqrt{10.9} = 3.03 \neq \sigma = 8.09$$

$St.dev(\bar{x})$ is a **biased estimator** for σ

Standard dev. of $\bar{x}$ and Standard Error

$$St.dev(\bar{x}) = \sqrt{10.9} = 3.03 \neq \sigma = 8.09$$

$St.dev(\bar{x})$ is a **biased estimator** for σ.

More generally, for sufficiently large samples (30 obs +), we have that:

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

"the standard error"

The **unbiased estimator** for σ

$\bar{x}$ is a Consistent Estimator for μ

Look at the formula:

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

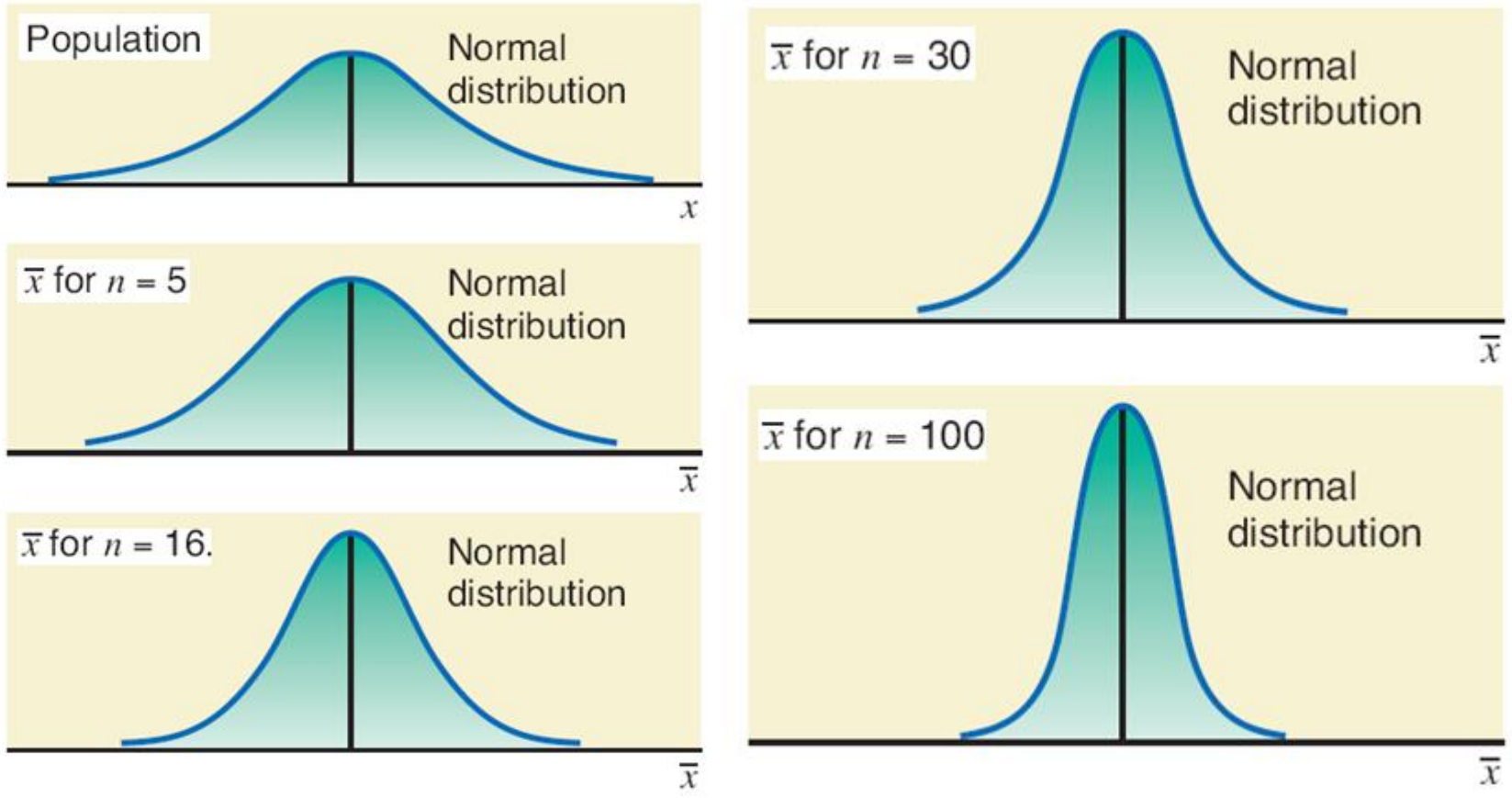
1) For $n > 1$, $\sigma_{\bar{x}} < \sigma$ ➔ the dispersion of the sampling distribution is smaller than the dispersion of the population distribution

2) $\sigma_{\bar{x}}$ decreases as n increases ➔ If the standard error of an estimator decreases as the sample size increases, the estimator is said to be consistent.

$\overline{x}$: shape of the sampling distribution

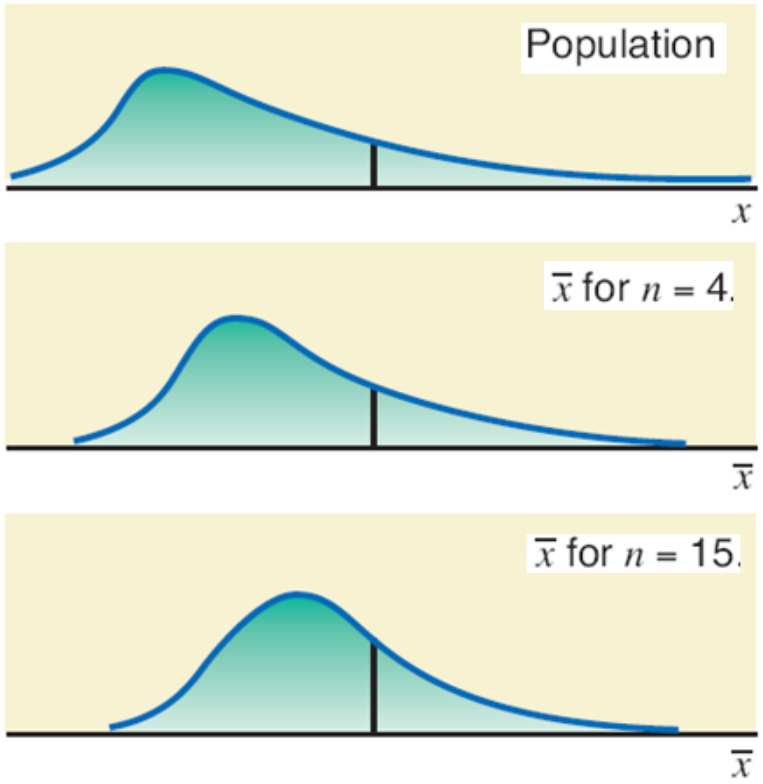
1) If the population from which samples are drawn is **normally distributed**, i.e. if $X \sim N(\mu; \sigma^2)$ ➔ <u>regardless for sample size n</u>,

$$\overline{x} \sim N(\mu; \frac{\sigma^2}{n})$$

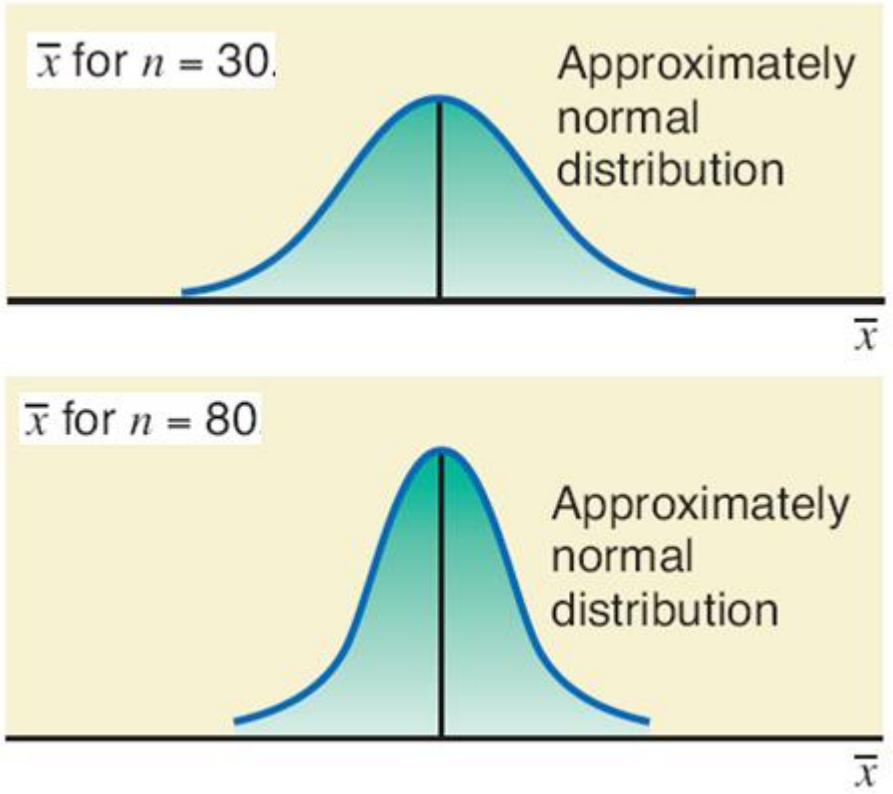


$\overline{x}$: shape of the sampling distribution

2) If the population is **NOT normally distributed** (or unknown) ➔ in large samples, i.e. $n > 30$, still $\overline{x} \sim N(\mu; \frac{\sigma^2}{n})$



CENTRAL LIMIT THEOREM



Sampling Distribution of the Mean

When the population is normally distributed

Shape: Regardless of sample size, the distribution of sample means will be normally distributed.

Center: The mean of the distribution of sample means is the mean of the population. Sample size does not affect the center of the distribution.

Spread: The standard deviation of the distribution of sample means, or the standard error, is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Standardizing a Sample Mean on a Normal Curve

The **standardized z-score** is how far above or below the sample mean is compared to the population mean in units of standard error.

"How far above or below" = sample mean minus μ

"In units of standard error" = divide by $\frac{\sigma}{\sqrt{n}}$

Standardized sample mean

$$z = \frac{sample\ mean\ -\ \mu}{standard\ error} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Central Limit Theorem

According to the **Central Limit Theorem (CLT),** the larger the sample size, the more normal the distribution of sample means becomes.

The CLT is central to the concept of statistical inference because it permits us to draw conclusions about the population based strictly on sample data without having knowledge about the distribution of the underlying population.

https://www.youtube.com/watch?v=b5xQmk9veZ4

Sampling Distribution of the Mean

When the population is **not normally distributed**

Shape: When the sample size taken from such a population is sufficiently large, the distribution of its sample means will be **approximately normally distributed** regardless of the shape of the underlying population those samples are taken from.

According to the **Central Limit Theorem**, the larger the sample size, the more normal the distribution of sample means becomes.

Sampling Distribution of the Mean

When the population is **not normally distributed**

Center: The mean of the distribution of sample means is the mean of the population, μ. Sample size does not affect the center of the distribution.

Spread: The standard deviation of the distribution of sample means, or the standard error, is

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

Example: Standardizing a Mean

Ex: When a production machine is properly calibrated, it requires an average of 25 seconds per unit produced, with a standard deviation of 3 seconds. For a simple random sample of $n = 36$ units, the sample mean is found to be 26.2 seconds per unit. When the machine is properly calibrated, what is the probability that the mean for a simple random sample of this size will be at least 26.2 seconds?

Example: Standardizing a Mean

Ex: When a production machine is properly calibrated, it requires an average of 25 seconds per unit produced, with a standard deviation of 3 seconds. For a simple random sample of $n = 36$ units, the sample mean is found to be 26.2 seconds per unit. When the machine is properly calibrated, what is the probability that the mean for a simple random sample of this size will be at least 26.2 seconds?

Standardized sample mean:

$$\bar{x} = 26.2,\ \mu = 25,\ \sigma = 3$$

$$z = \frac{26.2 - 25}{\frac{3}{\sqrt{36}}} = 2.40$$

$$P(\bar{x} \geq 26.2) = P(z \geq 2.40) = 0.0082$$

$\bar{x}$: summing up

1) $\bar{x}$ is an **unbiased estimator** for μ, since $E(\bar{x}) = \mu$

2) $\bar{x}$ is a **consistent estimator** for μ, since $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ decreases with n

3) $\bar{x}$ often has a **nice distribution**, since $\bar{x} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$. This happens:

- <u>regardless of sample size n</u>, if population is normally distributed, i.e. $X \sim N(\mu; \sigma^2)$
- <u>with sufficiently large sample size n</u>, if population is not normally distributed (or unknown)

Sampling distribution of $\overline{x}$: example I

According to the 2015 Physician Compensation Report, the average income of American physicians in 2014 was \$196,000. Assuming that that the earnings are normally distributed with standard deviation \$20,000, calculate the mean and standard deviation of $\overline{x}$ and describe the shape of its sampling distribution when:

(a) $n = 16$ (b) $n = 50$ (c) $n = 1000$

Sampling distribution of $\overline{x}$: example I

According to the 2015 Physician Compensation Report, the average income of American physicians in 2014 was \$196,000. Assuming that that the earnings are normally distributed with standard deviation \$20,000, calculate the mean and standard deviation of $\overline{x}$ and describe the shape of its sampling distribution when:

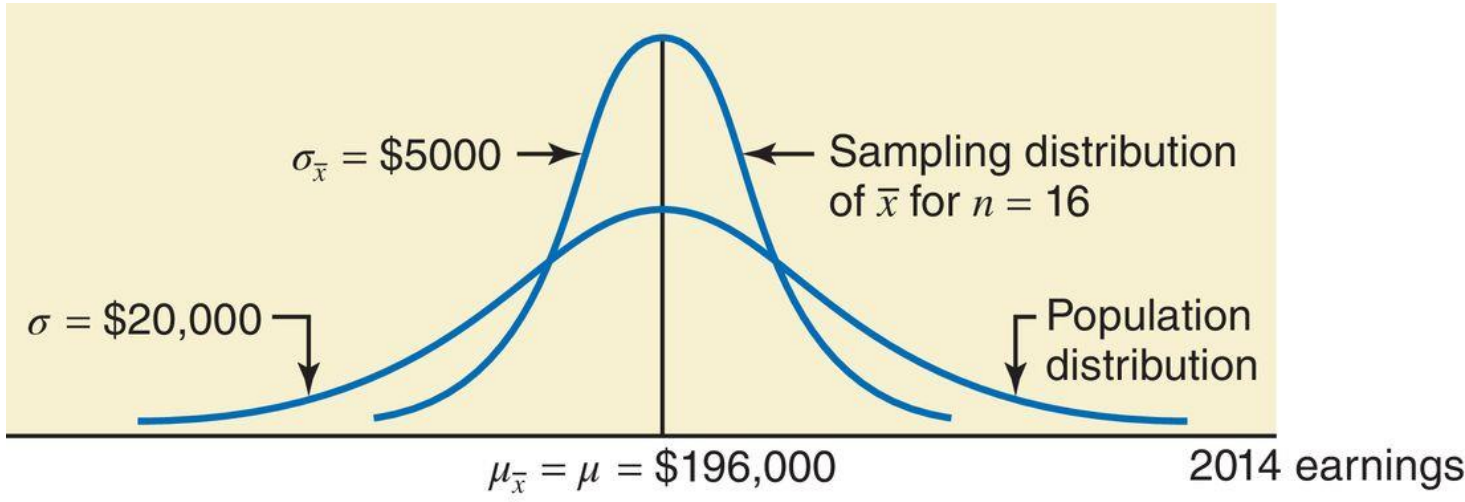
(a) $n = 16$ (b) $n = 50$ (c) $n = 1000$

Solution (a)

The ***average*** income is normally distributed with

$\mu_{\bar{x}} = \mu = \$196{,}000$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$20{,}000}{\sqrt{16}} = \$5{,}000$



Sampling distribution of $\overline{x}$: example I

According to the 2015 Physician Compensation Report, the average income of American physicians in 2014 was \$196,000. Assuming that that the earnings are normally distributed with standard deviation \$20,000, calculate the mean and standard deviation of $\overline{x}$ and describe the shape of its sampling distribution when:

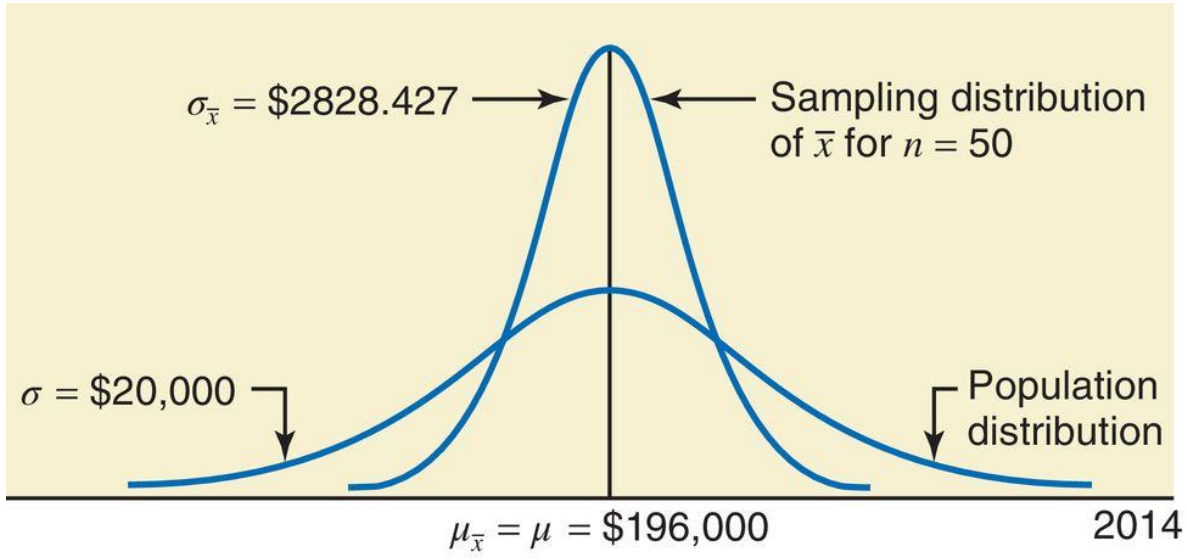
(a) $n = 16$ (b) $n = 50$ (c) $n = 1000$

Solution (b)

The ***average*** income is normally distributed with

$\mu_{\bar{x}} = \mu = \$196{,}000$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$20{,}000}{\sqrt{50}} = \$2{,}828.427$



Sampling distribution of $\overline{x}$: example I

According to the 2015 Physician Compensation Report, the average income of American physicians in 2014 was \$196,000. Assuming that that the earnings are normally distributed with standard deviation \$20,000, calculate the mean and standard deviation of $\overline{x}$ and describe the shape of its sampling distribution when:

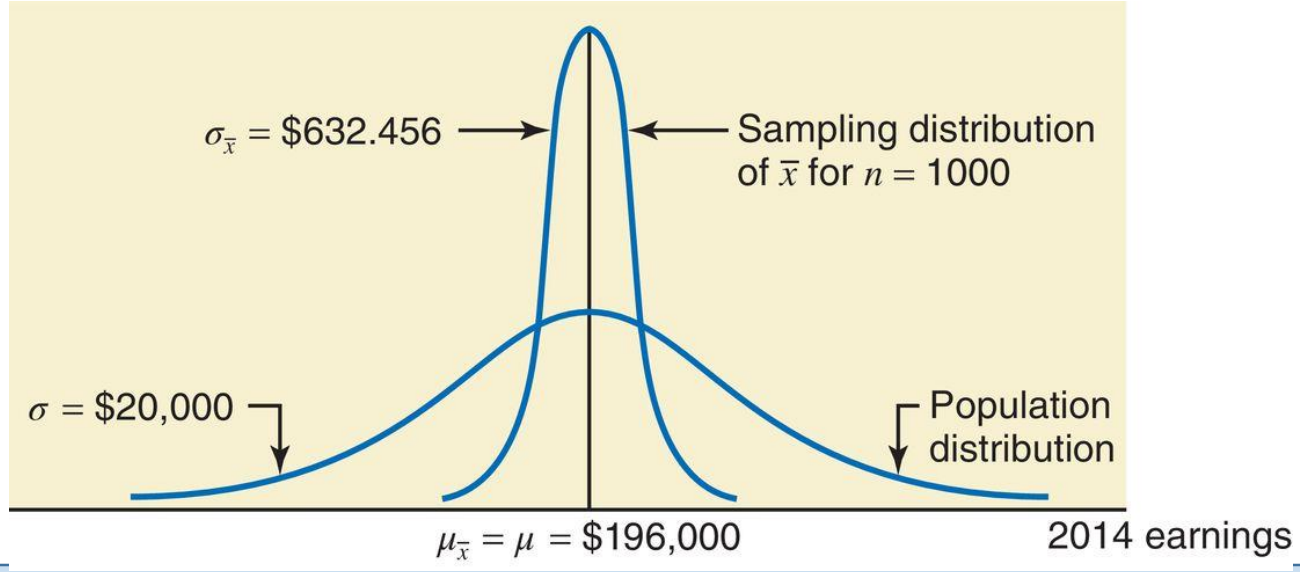
(a) $n = 16$ (b) $n = 50$ (c) $n = 1000$

Solution (c)

The ***average*** income is normally distributed with

$\mu_{\bar{x}} = \mu = \$196{,}000$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$20{,}000}{\sqrt{1000}} = \$632.456$



Sampling distribution of $\overline{x}$: example II (a)

The distribution of the rents paid in a small city, x, is skewed to the right. Knowing that the average is \$1550 and that the standard deviation is \$225, describe the sampling distribution of $\overline{x}$ when $n = 10$.

Sampling distribution of $\overline{x}$: example II (a)

The distribution of the rents paid in a small city, x, is skewed to the right. Knowing that the average is \$1550 and that the standard deviation is \$225, describe the sampling distribution of $\overline{x}$ when $n = 10$.

Solution

The ***average*** rent paid has an unknown distribution, since the distribution of the population is certainly not normal (because it's skewed) and n is not high enough to apply the Central Limit Theorem.

Sampling distribution of $\overline{x}$: example II (b)

The distribution of the rents paid in a small city, x, is skewed to the right. Knowing that the average is \$1550 and that the standard deviation is \$225, calculate the mean and standard deviation of $\overline{x}$ and describe the shape of its sampling distribution when $n = 100$.

Sampling distribution of $\overline{x}$: example II (b)

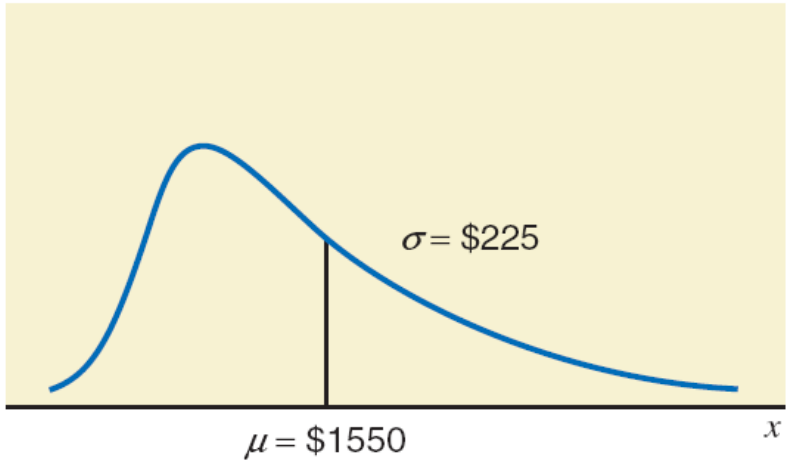
The distribution of the rents paid in a small city, x, is skewed to the right. Knowing that the average is \$1550 and that the standard deviation is \$225, calculate the mean and standard deviation of $\overline{x}$ and describe the shape of its sampling distribution when $n = 100$.

Solution

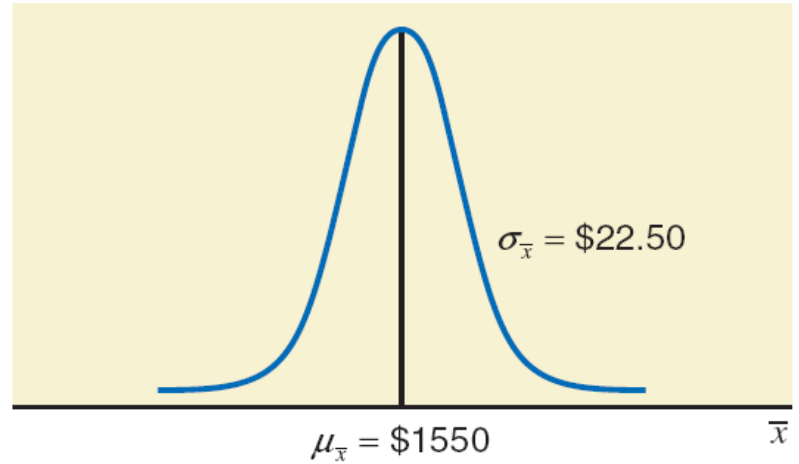
The ***average*** rent paid is normally distributed with

$\mu_{\bar{x}} = \mu = \$1550$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$225}{\sqrt{100}} = \$22.5$



(a) Population distribution.



(b) Sampling distribution of $\overline{x}$ for $n = 100$.

Sampling distribution of $\overline{x}$: example III

Assume that the exam scores of all examinees is normally distributed with a mean of 1020 and a standard deviation of 153. Let $\overline{x}$ be the mean score of a random sample of a set of examinees. Describe the shape of its sampling distribution when the sample size is

(a) $n = 16$ (b) $n = 50$ (c) $n = 1000$

Sampling distribution of $\bar{x}$: example III

Assume that the exam scores of all examinees is normally distributed with a mean of 1020 and a standard deviation of 153. Let $\bar{x}$ be the ***average*** score of a random sample of a set of examinees. Describe the shape of its sampling distribution when the sample size is:

(a) $n = 16$ (b) $n = 50$ (c) $n = 1000$

Solution

average score is normally distributed with:

(a)	(b)	(c)
$\mu_{\bar{x}} = \mu = 1020$ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{153}{\sqrt{16}} = 38.25$	$\mu_{\bar{x}} = \mu = 1020$ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{153}{\sqrt{50}} = 21.637$	$\mu_{\bar{x}} = \mu = 1020$ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{153}{\sqrt{1000}} = 4.838$

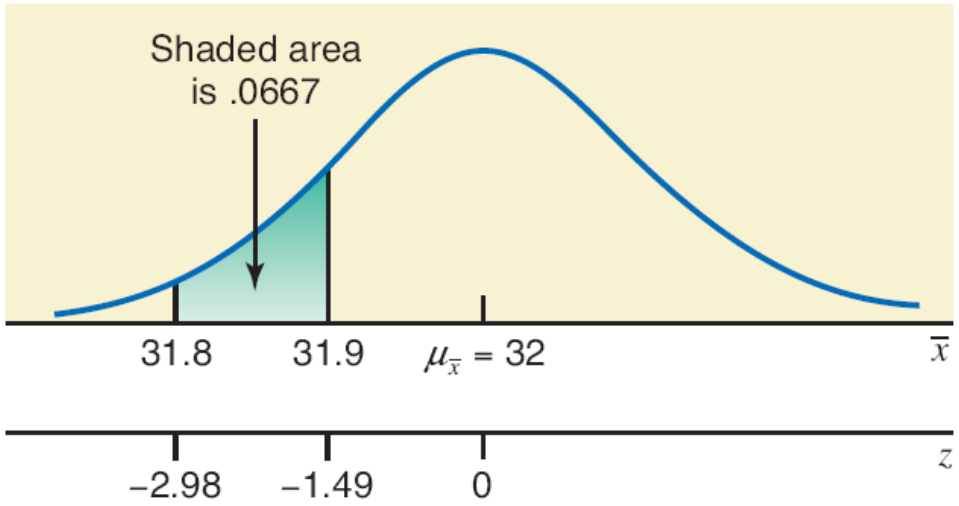
Sampling distribution of $\overline{x}$: application I

The weight of the packages of a certain brand of cookies, x, is normally distributed with $\mu = 32$ and $\sigma = 0.3$ ounces. Find the probability that the ***average*** weight of a random sample of 20 packages is between 31.8 and 31.9 ounces.

Sampling distribution of $\overline{x}$: application I

The weight of the packages of a certain brand of cookies, x, is normally distributed with $\mu = 32$ and $\sigma = 0.3$ ounces. Find the probability that the ***average*** weight of a random sample of 20 packages is between 31.8 and 31.9 ounces.

Solution. If $X \sim N(\mu = 32; \sigma^2 = 0.09)$ ➔ $\overline{x} \sim N(\mu = 32; \frac{\sigma^2}{n} = 0.0045)$



➔ $P(31.8 < \overline{x} < 31.9) = 0.0667$

Sampling distribution of $\overline{x}$: application II

The annual cost of a checking account at major U.S. banks has a mean of \$400 and a standard deviation of \$30. Consider a random sample of 225 US checking accounts, what is the probability that the ***average*** annual cost of the checking accounts in this sample is:

(a) within \$4 of the population mean?
(b) less than the population mean by \$2.70 or more?

Sampling distribution of $\overline{x}$: application II

The annual cost of a checking account at major U.S. banks has a mean of \$400 and a standard deviation of \$30. Consider a random sample of 225 US checking accounts, what is the probability that the ***average*** annual cost of the checking accounts in this sample is:

(a) within \$4 of the population mean?

Population distribution is unknown, but $n > 30$ ➔ CLT

$$\overline{x} \sim N(\mu_{\bar{x}} = \$400; \frac{\sigma^2}{n} = \frac{900}{225} = 4)$$

$$P(\$396 < \overline{x} < \$404) = P(-2 < z < 2) = 0.9544$$

Sampling distribution of $\overline{x}$: application II

The annual cost of a checking account at major U.S. banks has a mean of \$400 and a standard deviation of \$30. Consider a random sample of 225 US checking accounts, what is the probability that the ***average*** annual cost of the checking accounts in this sample is:

(b) less than the population mean by \$2.70 or more?

$$\overline{x} \sim N(\mu_{\bar{x}} = \$400; \frac{\sigma^2}{n} = \frac{900}{225} = 4)$$

$$P(\overline{x} < 400 - 2.7 = 397.3) = P(z < -1.35) = 0.0885$$

$\hat{p}$: sample proportion

Recall the CENTRAL LIMIT THEOREM: if the population is **NOT normally** distributed (or unknown) ➔ if $n > 30$, still $\overline{x} \sim N(\mu; \frac{\sigma^2}{n})$

Particular case, when X$\sim Ber(p)$.
Remember, $\mu = E(X) = p$ and $\sigma^2 = V(X) = p(1-p)$

In this case, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{p}$, i.e. proportion of 1 ("successes") over n

Sample proportion

$$\hat{p} \sim N(\mu_{\hat{p}} = p; \sigma^2_{\hat{p}} = \frac{p(1-p)}{n})$$

Sample proportion: example

According to a New York Times poll, 55% of US citizens consider owning a home part of the American Dream. Consider a random sample of 2000 US citizens, find mean and standard deviation of $\hat{p}$ and describe its sampling distribution.

Sample proportion: example

According to a New York Times poll, 55% of US citizens consider owning a home part of the American Dream. Consider a random sample of 2000 US citizens, find mean and standard deviation of $\hat{p}$ and describe its sampling distribution.

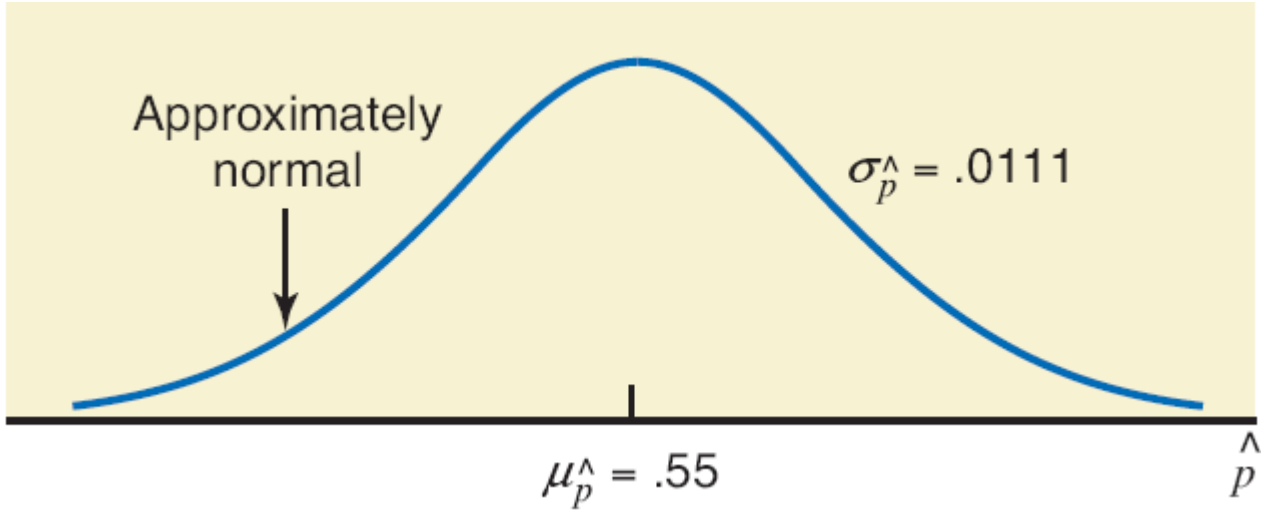
Solution

average proportion is normally distributed with

$\mu_{\hat{p}} = p = 0.55$

$\sigma^2_{\hat{p}} = \frac{p(1-p)}{n} = \frac{0.55(0.45)}{2000} = 0.000124$

➔$\sigma_{\hat{p}} = 0.0111$



Sample proportion: application

75% of American adults think that college education is too expensive for most people. Find the probability that 76.5% to 78% of adults in a random sample of 1400 Americans will hold this opinion.

Sample proportion: application

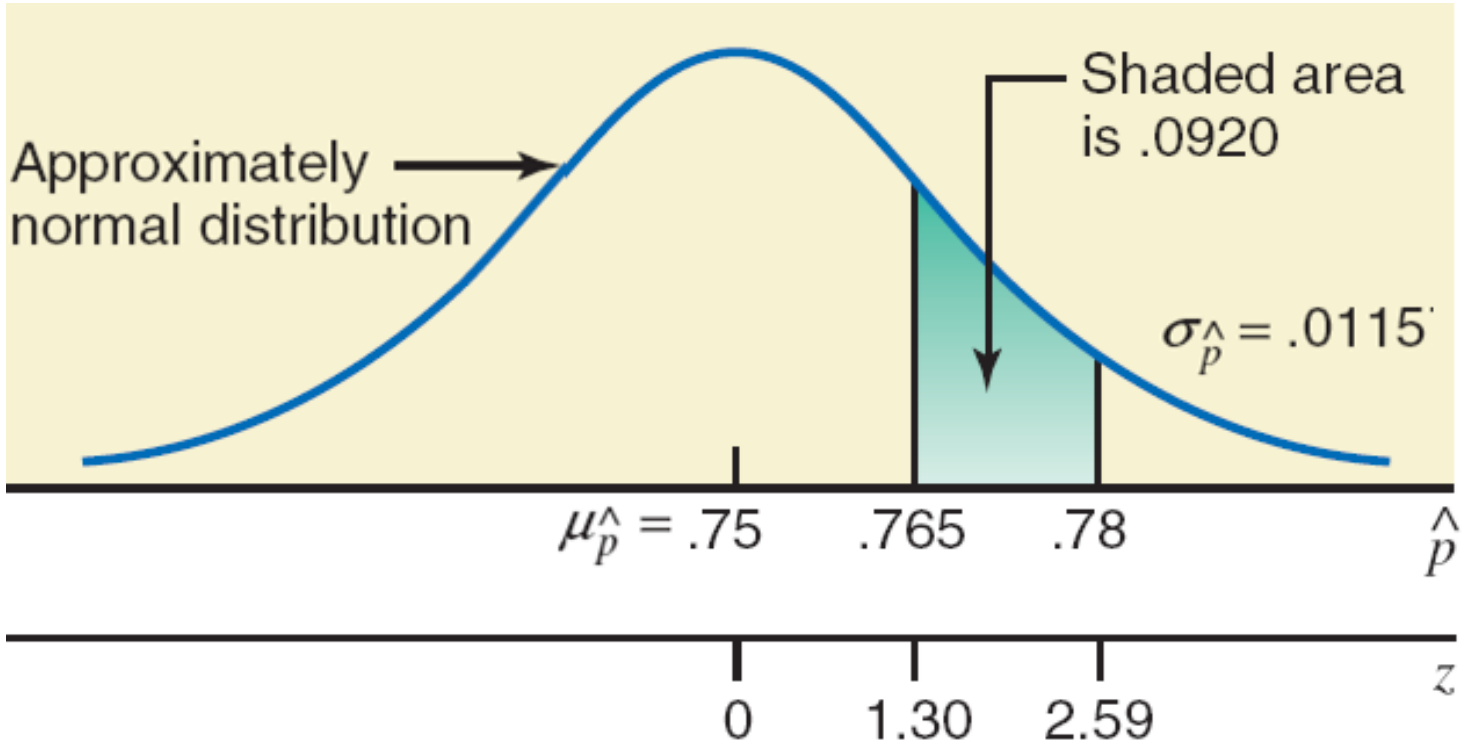
75% of American adults think that college education is too expensive for most people. Find the probability that 76.5% to 78% of adults in a random sample of 1400 Americans will hold this opinion.

Solution

average proportion is normally distributed with

$\mu_{\hat{p}} = p = 0.75$

$\sigma_{\hat{p}} = \sqrt{\frac{0.75(0.25)}{1400}} = 0.0115$



➔ $P(0.765 < \hat{p} < 0.78) =$

$= P(1.30 < z < 2.59) = 0.092$

Sample proportion: application II

Maureen Webster, who is running for mayor in a large city, claims that 53% of eligible voters will vote for her. Find the probability that less than 49% of a random sample of 400 registered voters will favour Maureen.

Sample proportion: application II

Maureen Webster, who is running for mayor in a large city, claims that 53% of eligible voters will vote for her. Find the probability that less than 49% of a random sample of 400 registered voters will favour Maureen.

Solution

$\mu_{\hat{p}} = p = 0.53$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.53(0.47)}{400}} = 0.025$$

➔ $\mathrm{P}(\hat{p} < 0.49) =$
$= P(z < -1.60) = 0.0548$