2024-1

Multivariate Data Analysis

Youjin Shin
(yj.shinn@catholic.ac.kr)

E(X) and Var(X) of X~B(n,p)

- When X is a binomial random variable that follows a binomial distribution, let's calculate $E(X_1)$ and $Var(X_1)$ of this X.
  - First, we calculate mean and variance of Bernoulli random variable (X_1)

$$P(X_1 = x) = p^x(1-p)^{(1-x)}, \qquad x = 0, 1$$

$$E(X_1) = \sum_x xP(X_1 = x) \quad = 0 \times (1-p) + 1 \times p = p$$

$$Var(X_1) = E(X_1 - E(X_1))^2 = \sum_x (x-p)^2 P(X_1 = x) \quad = (0-p)^2 (1-p) + (1-p)^2 p = p(1-p)$$

  - Next, we calculate mean and variance of Bernoulli random variables $(X_1, X_2, \cdots, X_n)$, where the success probability of variables is equal to P.

$$E(X) = E(X_1 + X_2 + \cdots + X_n) = nE(X_1) = np$$

$$Var\ (X) = Var(X_1 + X_2 + \cdots + X_n) = nVar(X_1) = np(1-p)$$

Binomial Distribution

◆ **As the sample size increases, the standard error of the sample proportion decreases.**

> ❖ **Example**
> Let's assume there is a constituency in Seoul with 100,000 voters. A random sample of **100 voters** from this constituency was surveyed, and the approval rating for a specific candidate in this sample was found to be **53%.** Based on this data, can we be certain that this candidate will win the election?
> (1: if you support the specific candidate /. 0 : if you do not support the candidate.)

➢ X_1 is Bernoulli Random Variable (Bernoulli Trial)

① Standard deviation

$$Var(X_1) = E(X_1 - E(X_1))^2 = \sum_x (x-p)^2 P(X_1 = x) = \cdots = p(1-p) = 0.53 \times 0.47 = 0.2491$$

$$Std(\sigma) = \sqrt{Var(X_1)} = \sqrt{0.53 \times 0.47} = 0.4991$$

Binomial Distribution

◆ **As the sample size increases, the standard error of the sample proportion decreases.**

> ❖ **Example**
> Let's assume there is a constituency in Seoul with 100,000 voters. A random sample of **100 voters** from this constituency was surveyed, and the approval rating for a specific candidate in this sample was found to be 53%. Based on this data, can we be certain that this candidate will win the election?
> (1: if you support the specific candidate /. 0 : if you do not support the candidate.)

② Standard Error

$$SEM = \frac{\sigma}{\sqrt{n}} = \frac{0.4991}{\sqrt{100}} = 0.0499$$ 약 5%

③ The range of approval ratings for the candidate :
(Estimation of approval ratings in the population)
$\mathbf{53\% \pm 5\%}$ $\mathbf{(48 \sim 58\%)}$

The approval rating range includes values below 50%, so victory could not be guaranteed!

Binomial Distribution

◆ **As the sample size increases, the standard error of the sample proportion decreases.**

❖ **(QUIZ) Example2**

Let's assume there is a constituency in Seoul with 100,000 voters. A random sample of ~~**100**~~ **2,500 voters** from this constituency was surveyed, and the approval rating for a specific candidate in this sample was found to be ~~**53**~~ **52%.** Based on this data, can we be certain that this candidate will win the election?

(1: if you support the specific candidate /. 0 : if you do not support the candidate.)

➢ X_1 is Bernoulli Random Variable (Bernoulli Trial)

① Standard deviation

② Standard Error

③ The range of approval ratings for the candidate :

Binomial Distribution

◆ **As the sample size increases, the standard error of the sample proportion decreases.**

❖ **(QUIZ) Example2**

Let's assume there is a constituency in Seoul with 100,000 voters. A random sample of ~~**100**~~ **2,500 voters** from this constituency was surveyed, and the approval rating for a specific candidate in this sample was found to be ~~**53**~~ **52%.** Based on this data, can we be certain that this candidate will win the election?

(1: if you support the specific candidate /. 0 : if you do not support the candidate.)

➢ X_1 is Bernoulli Random Variable (Bernoulli Trial)

① Standard deviation

$$Var(X_1) = E(X_1 - E(X_1))^2 = \sum_x (x-p)^2 P(X_1 = x) = \cdots = p(1-p) = 0.52 \times 0.48 = 0.2496$$

$$Std(\sigma) = \sqrt{Var(X_1)} = \sqrt{0.52 \times 0.48} = 0.4996$$

Binomial Distribution

◆ **As the sample size increases, the standard error of the sample proportion decreases.**

❖ **(QUIZ) Example2**

Let's assume there is a constituency in Seoul with 100,000 voters. A random sample of ~~**100**~~ **2,500 voters** from this constituency was surveyed, and the approval rating for a specific candidate in this sample was found to be ~~**53**~~ **52%.** Based on this data, can we be certain that this candidate will win the election?

(1: if you support the specific candidate /. 0 : if you do not support the candidate.)

② Standard Error

$$SEM = \frac{\sigma}{\sqrt{n}} = \frac{0.4996}{\sqrt{2500}} = 0.01$$

약 1%

③ The range of approval ratings for the candidate :

$$\mathbf{52\% \pm 1\%\ (51 \sim 53\%)}$$

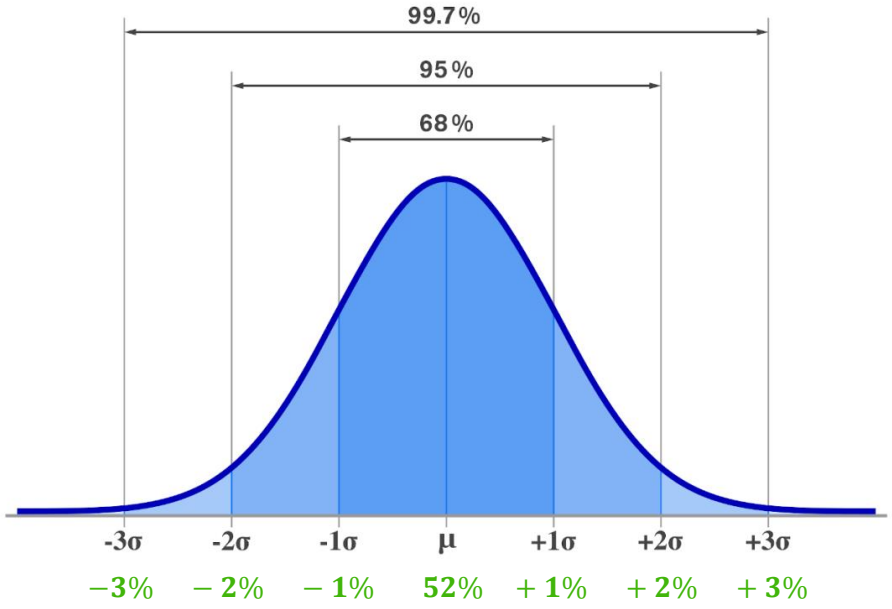
The approval rating range is entirely above 50%, making the election win almost certain.

Confidence Interval

- As a result of conducting a random survey of 2,500 people in a constituency with 100,000 voters, the approval rating for a specific candidate was 52% and the standard error was about 1%. Therefore, the estimated value of the approval rating for the parameter is as follows:

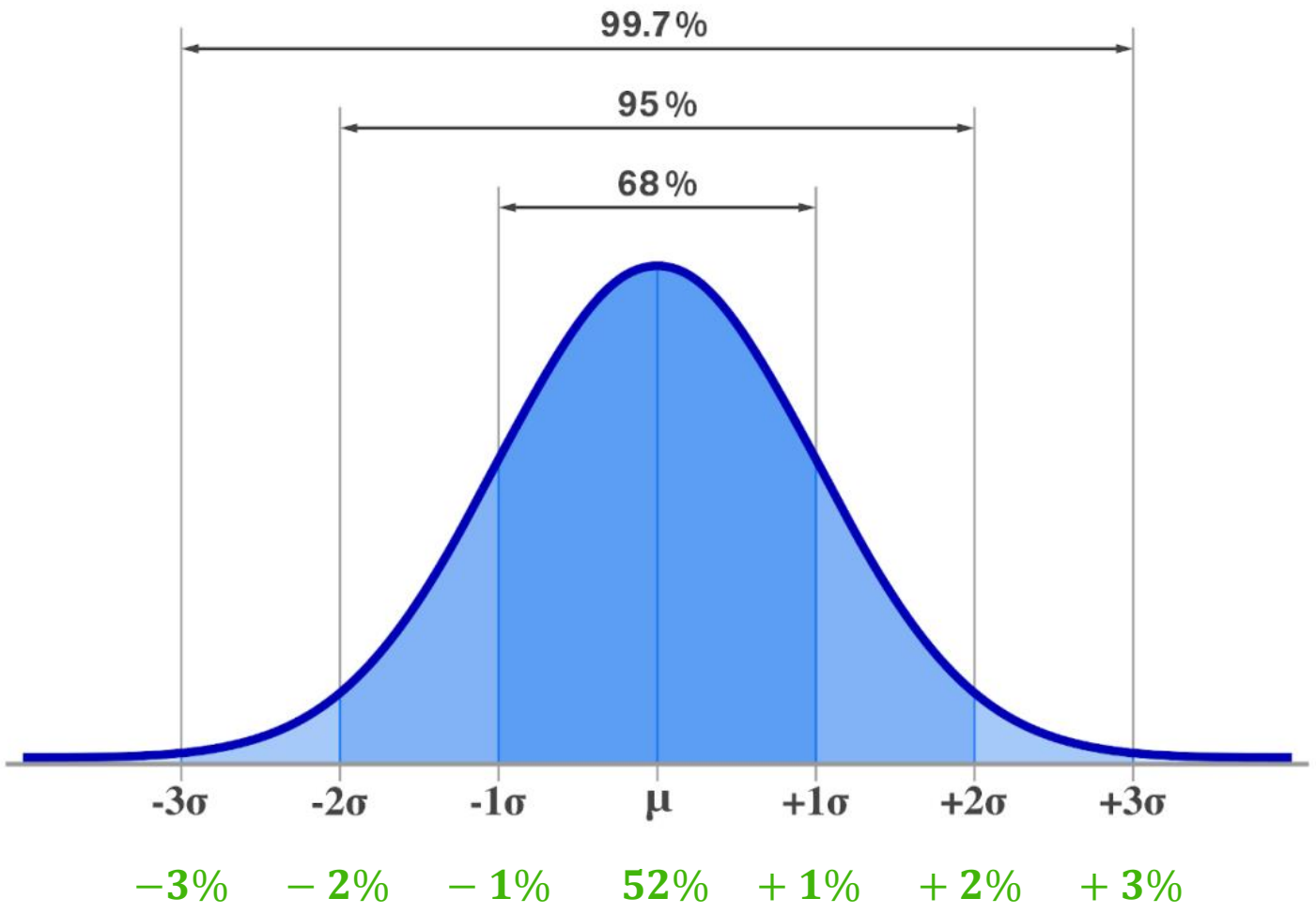
$$\mathbf{52\% \pm 1\%\ (51\sim53\%)}$$

- **Let's interpret the meaning of this value through normal distribution.**



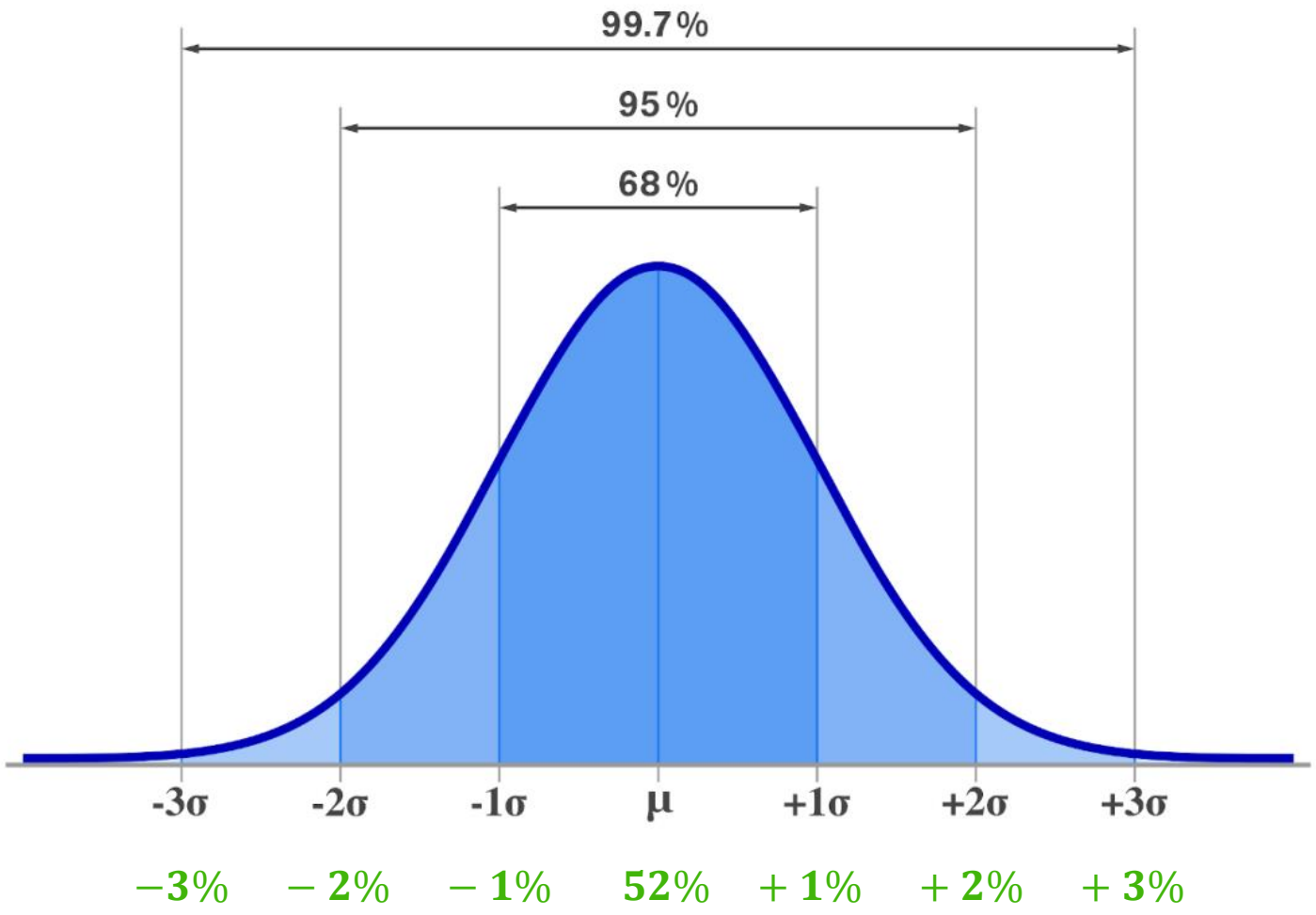
- Mean of the sample : 52
- The standard error of the sample mean (SEM) : 1

Confidence Interval



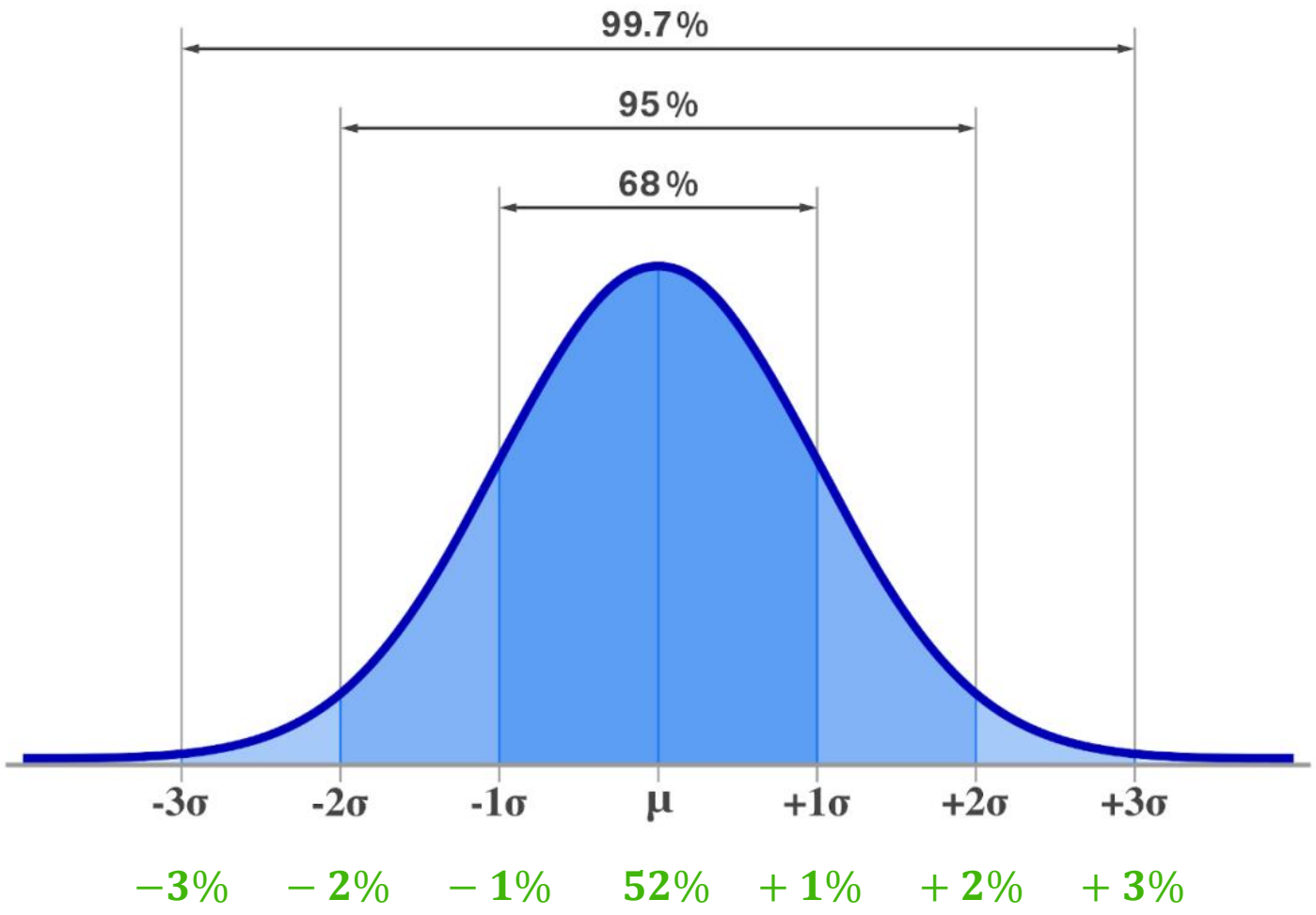
◆ Sum of the area under the normal distribution curve corresponding to $\mathbf{52\%\pm 1}\boldsymbol{SE}$ (**51%~53%**) is **68%**

Confidence Interval



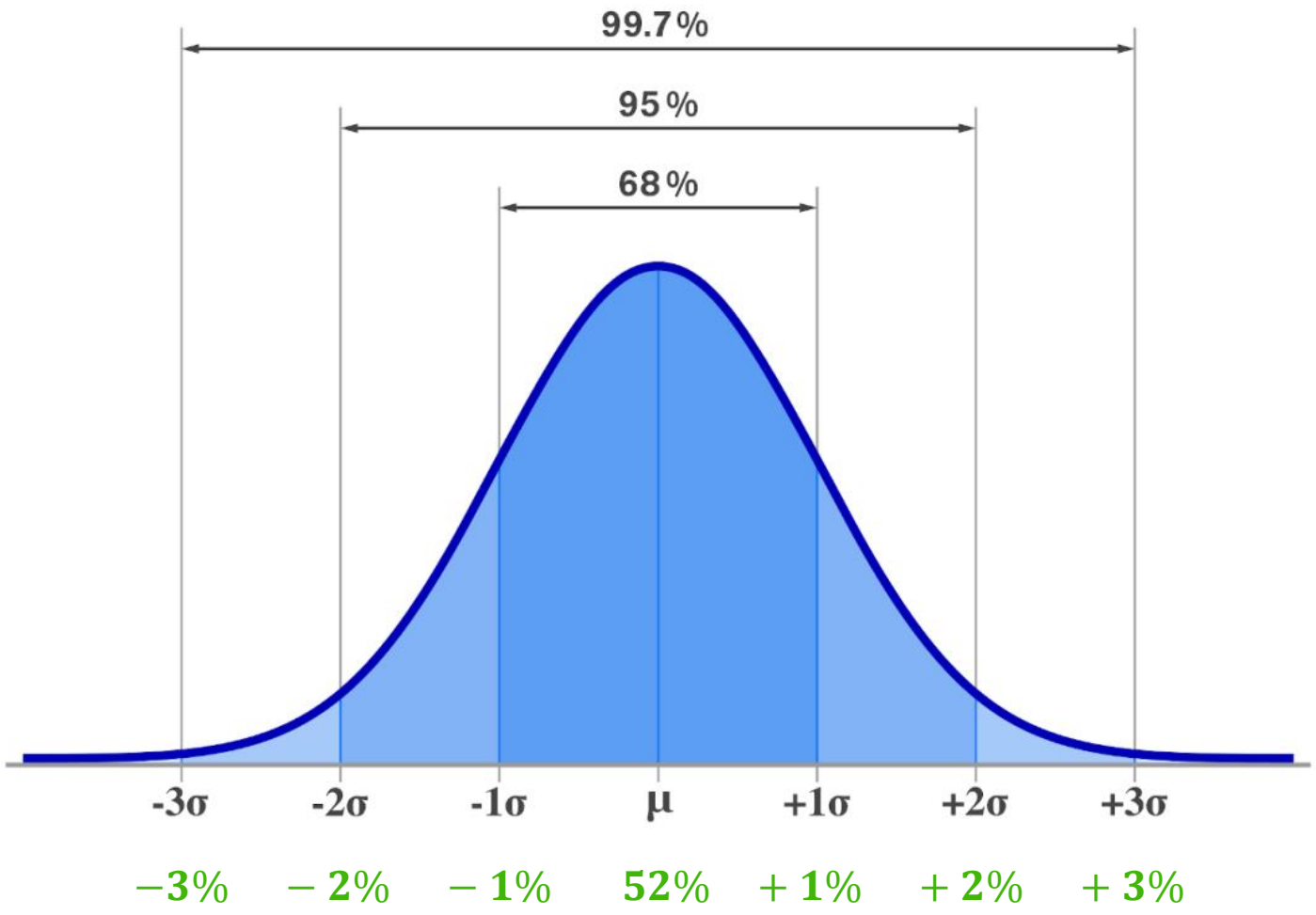
- Sum of the area under the normal distribution curve corresponding to $\mathbf{52\%\pm 1}\boldsymbol{SE}$ ($\mathbf{51\%}$~$\mathbf{53\%}$) is **68%**
- (When estimating the population proportion using the sample proportion), the **Confidence Interval (CI) with approximately 68%** confidence in the population proportion is **51-53%**.
- I have 68% confidence that the approval rating for a specific candidate will be in the range of 51-53%.

Confidence Interval



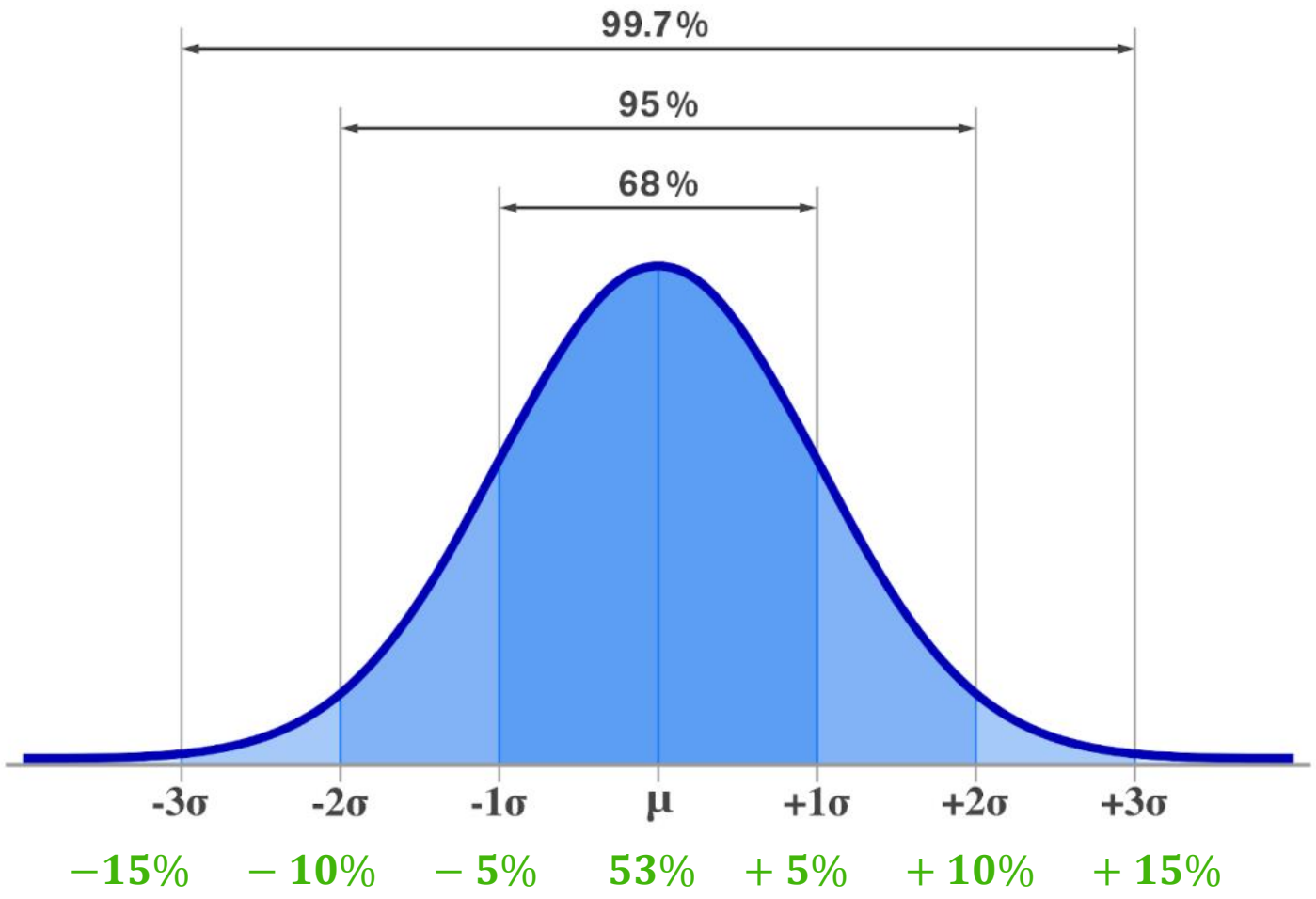
- Sum of the area under the normal distribution curve corresponding to $\mathbf{52\%} \pm \mathbf{2}\boldsymbol{SE}$ (**50**%~**54**%) is **95%**
- (When estimating the population proportion using the sample proportion), the **Confidence Interval (CI) with approximately 95%** confidence in the population proportion is **50-54%**.
- I have 95% confidence that the approval rating for a specific candidate will be in the range of 50-54%.

Confidence Interval



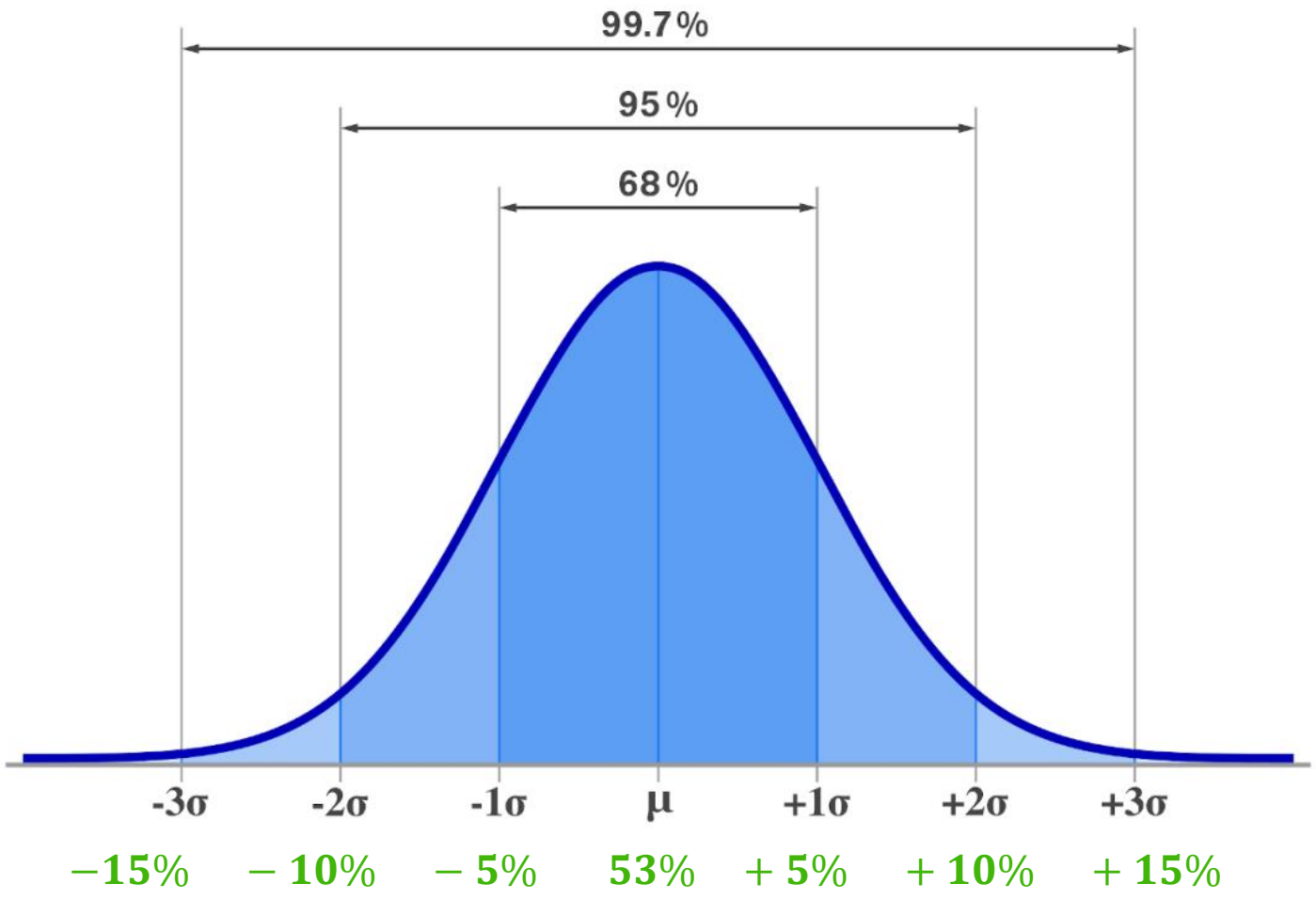
- Sum of the area under the normal distribution curve corresponding to $\mathbf{52\%\pm 3}\boldsymbol{SE}$ (**49**%~**55**%) is **99.7%**
- (When estimating the population proportion using the sample proportion), the **Confidence Interval (CI) with approximately 99.7%** confidence in the population proportion is **49-55%**.
- I have 99.7% confidence that the approval rating for a specific candidate will be in the range of 49-55%.

Confidence Interval



Q.
In the case of the example 1
(53% ± 5% (48~58%), What is the confidence interval with 95% confidence?

Confidence Interval



Q.
In the case of the example 1
(53% ± 5% (48~58%), What is the confidence interval with 95% confidence?

A. In a normal distribution, the interval where the area under the curve is 95% is **5**3%±2$\boldsymbol{SE}$ = **5**3%±10

=> **The Confidence Interval (CI) with 95% confidence is 43%~63%)**

Bootstrap

- An easy and effective way to estimate the distribution for a population is to **repeatedly sampling** from an existing dataset with replacement to create many new simulated samples.

- **Bootstrap** is a statistical method that allows for the estimation of the distribution of a statistic (like the mean, median, variance) over a population by creating many simulated samples.
- This approach is useful for assessing the variability of a statistical estimate when the underlying distribution of the population is unknown or when it's difficult to derive the sampling distribution of a statistic directly.

Bootstrap

◆ The Process of Bootstrap:

1. **Original Dataset**: Start with an original dataset of size N.
2. **Resampling**: Draw repeated samples, each of size N, with replacement from the original dataset. This means each sample may include some data points multiple times and omit others.
3. **Calculate Statistics**: For each resampled dataset, calculate the statistic of interest (mean, median, standard deviation, etc.).
4. **Distribution of the Statistic**: After many (usually thousands of) iterations of resampling and recalculating, you'll have a distribution of the statistic.

Bootstrap

◆ **Why Bootstrap Works:**

- Bootstrap works under the principle that the original sample is a good representation of the underlying population. By resampling with replacement, we create 'new' samples that could have theoretically been drawn from the population. This allows us to estimate the sampling distribution of our statistic without requiring the actual population or making strong assumptions about its distribution.

◆ **Advantages:**

- Flexibility: Can be applied to a wide range of statistics and complex estimators.
- Simplicity: Relatively easy to implement and understand, without the need for complex mathematical formulas.
- Applicability: Useful in situations where the sample size is too small to rely on traditional parametric methods, or the distribution of the population is unknown.

◆ **Limitations:**

- Accuracy: The accuracy of bootstrap estimates can suffer if the original sample is not representative of the population.
- Computationally Intensive: Requires a significant amount of computation, especially as the sample size or the number of bootstrap samples increases.

Lab: Bootstrap

- Let's use bootstrap to compare the height of coffee drinkers and non-drinkers.
- DATASET : MDA_09_coffee_dataset.csv

```
import pandas as pd
import numpy as np


np.random.seed(104)


df = pd.read_csv('data/MDA_09_coffee_dataset.csv')
print(df_sample.info())
print(df_sample.head(5))
```

	user_id	age	drinks_coffee	height
0	4509	<21	False	64.538179
1	1864	>=21	True	65.824249
2	2060	<21	False	71.319854
3	7875	>=21	True	68.569404
4	6254	<21	True	64.020226

	user_id	age	drinks_coffee	height
486	4951	<21	False	68.539632
2297	5383	>=21	True	70.594166
911	3904	<21	False	61.114035
146	4351	>=21	True	68.952444
2255	3998	<21	False	65.493818

Lab: Bootstrap

```
# Randomly sample 200 samples from the population.
df_sample = df.sample(200)
print(df_sample.head())
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 200 entries, 486 to 914
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   user_id       200 non-null    int64
 1   age           200 non-null    object
 2   drinks_coffee  200 non-null    bool
 3   height        200 non-null    float64
dtypes: bool(1), float64(1), int64(1), object(1)
memory usage: 6.4+ KB
```

Lab: Bootstrap

```
##### Confidence interval using bootstrap #####
# Let's repeat the bootstrap 10,000 times to find the 99.7% confidence interval for the height difference
# between people who do not drink coffee and people who drink coffee.

# 1. Average height difference between non-coffee drinkers and coffee drinkers

iterationNum = 1000
diffHeightList = []

for _ in range(iterationNum):
    bootSample = df_sample.sample(200, replace=True)      # sampling with replacement
    # Avg. height of people who don't drink coffee
    nonCoffeeHeightMean = bootSample[bootSample['drinks_coffee'] == False].height.mean()
    # Avg. height of people who drink coffee
    coffeeHeightMean = bootSample[bootSample['drinks_coffee'] == True].height.mean()
    diff = nonCoffeeHeightMean - coffeeHeightMean
    diffHeightList.append(diff)
```

Lab: Bootstrap

```
# When the confidence level is 99%.7, the confidence interval for the average height difference

print("Lowerbound(0.3):", np.percentile(diffHeightList, 0.3))
print("Uppperbound(99.7):", np.percentile(diffHeightList, 99.7))
```

```
Lowerbound(0.3): -3.1053882693534907
Uppperbound(99.7): -0.8117815664737708
```

* diff = nonCoffeeHeightMean - coffeeHeightMean ⇨

Diff > 0 : noncoffee group is taller

Diff < 0 : coffee group is taller

- According to the calculation results, the 99.7% confidence interval is are all negative.
- In general, we can say that people who drink coffee are taller than people who don't drink coffee.

Lab: Bootstrap

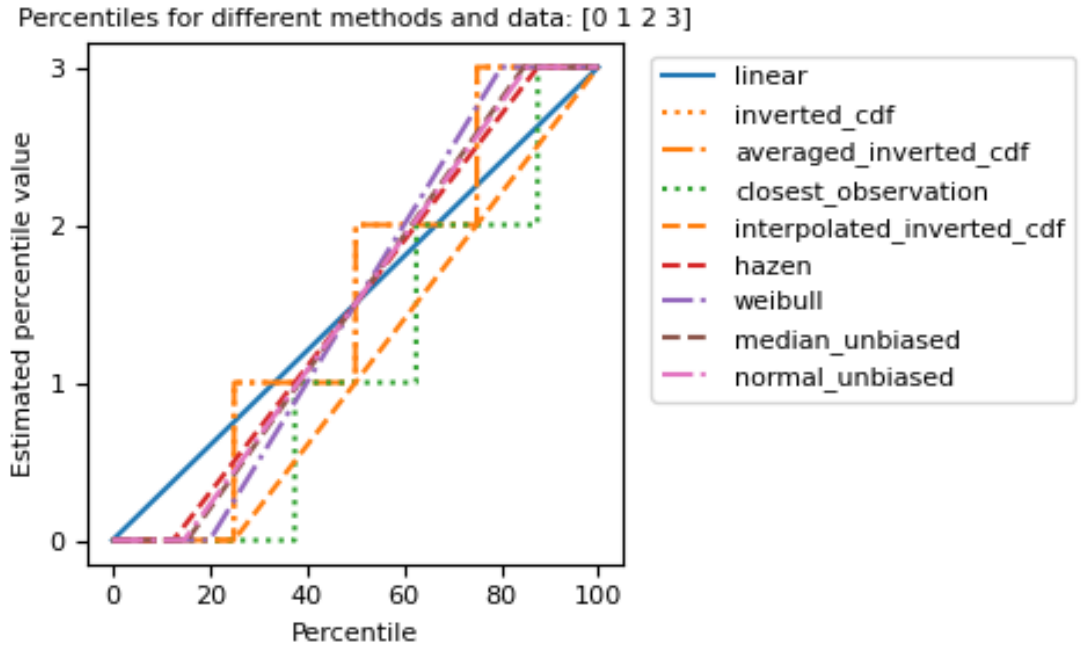
- CF (the results of np.percentile() and hand-calculation can be different each other)

  np.Percentile function has many different way
  to calculate the percentiles so values can be different

method : *str, optional*

This parameter specifies the method to use for estimating the percentile. There are many different methods, some unique to NumPy. See the notes for explanation. The options sorted by their R type as summarized in the H&F paper [1] are:

1. 'inverted_cdf'
2. 'averaged_inverted_cdf'
3. 'closest_observation'
4. 'interpolated_inverted_cdf'
5. 'hazen'
6. 'weibull'
7. 'linear' (default)
8. 'median_unbiased'
9. 'normal_unbiased'



Lab: Comparison of parameters and bootstrap CIs

- A confidence interval is determined through bootstrapping, and if the observed value is within the confidence interval, the hypothesis is determined to be true.

- In this example, since we know the original data, we will find the difference between the actual means and see if the parameter is within the confidence interval.

Lab: Comparison of parameters and bootstrap CIs

```
print("##### Height differences in the population #####")
# # 1. Average height difference between non-coffee drinkers and coffee drinkers
diffHeight = df[df['drinks_coffee'] == False].height.mean() - df[df['drinks_coffee'] == True].height.mean()
print("diffHeight : ",diffHeight)
```

```
##### Height differences in the population #####
1. diffHeight :  -1.9568024933368093
```

	Parameter (population)	Confidence Interval of Samples
Average height difference between non-coffee drinkers and coffee drinkers	-1.9568	-3.1054 ~ -0.8118

⇒ **Parameters are within the bootstrap confidence interval !!!**

Lab: Bootstrap

```
# When the confidence level is 99%.7, the confidence interval for the average height difference

print("Lowerbound(0.3):", np.percentile(diffHeightList, 0.3))
print("Uppperbound(99.7):", np.percentile(diffHeightList, 99.7))
```

```
Lowerbound(0.3): -3.1053882693534907
Uppperbound(99.7): -0.8117815664737708
```

* diff = nonCoffeeHeightMean - coffeeHeightMean ⇨

Diff > 0 : noncoffee group is taller

Diff < 0 : coffee group is taller

- According to the calculation results, the 99.7% confidence interval is are all negative.
- In general, we can say that people who drink coffee are taller than people who don't drink coffee.

- **BTW, is this really true?**

Simpson's Paradox

◆ Simpson's Paradox is a fascinating phenomenon in statistics that can lead to seemingly paradoxical results. It occurs when a trend appears in different groups of data but reverses or disappears when these groups are combined.

➢ (Example) Employment rate for men and women

: When we investigated the employment rate, we found that the employment rate for men was much higher.

** Suppose that there are only two fields (IT & Education) in this employment market.

	# of applicants	# of pass	# of fail
Male	1,000	730	73%
Female	1,000	270	27%

Simpson's Paradox

- (Example) Employment rate for men and women

	# of applicants	# of pass	# of fail
Male	1,000	730	73%
Female	1,000	270	27%

- Employment rate of IT field

	# of applicants	# of pass	# of fail
Male	900	720	80%
Female	200	180	90%

- Employment rate of Education field

	# of applicants	# of pass	# of fail
Male	100	10	10%
Female	800	90	11.25%

Simpson's Paradox

- (QUIZ) Let's figure out whether Simpson's Paradox exists in our data (the relationship between coffee drinking and height).
  - Age criteria : 21 years old
  - Required Values:
    - Average height difference between people over 21 years old and under 21 years old (Including Confidence level (99.7%) checking)
    - Average height difference between non-coffee drinkers and coffee drinkers among people under 21 years of age (Including Confidence level (99.7%) checking)
    - Average height difference between non-coffee drinkers and coffee drinkers among people over 21 years of age (Including Confidence level (99.7%) checking)

Lab: Bootstrap

```
# 2. Average height difference between people over 21 years old and under 21 years old

diffHeightListByAge = []
for _ in range(iterationNum):
    bootSample = df_sample.sample(200, replace=True)    # sampling with replacement
    over21HeightMean = bootSample[bootSample['age'] == '>=21'].height.mean()     # Avg.Height for over 21
    under21HeightMean = bootSample[bootSample['age'] == '<21'].height.mean()     # Avg.Height for under 21
    diff = over21HeightMean - under21HeightMean
    diffHeightListByAge.append(diff)

# When the confidence level is 99%.7, the confidence interval for the average height difference

print("Lowerbound(0.3):", np.percentile(diffHeightListByAge, 0.3))
print("Uppperbound(99.7):", np.percentile(diffHeightListByAge, 99.7))
```

```
Lowerbound(0.3): 3.1355516936531616
Uppperbound(99.7): 5.016793651153235
```

- diff = over21HeightMean - under21HeightMean
- Mean, CI(99.7%) are all positive numbers.

⇒ At a 99.7% confidence level,

⇒ People over 21 are taller than people under 21.

Lab: Bootstrap

```
# 3. Average height difference between non-coffee drinkers and coffee drinkers among people under 21 years of age
diffHeightListUnder21 = []
for _ in range(iterationNum):
    bootSample = df_sample.sample(200, replace=True)   # sampling with replacement
    # Average height of people under 21 years of age who do not drink coffee
    nonCoffeeHeightMeanUnder21 = bootSample.query("age == '<21' and drinks_coffee == False").height.mean()
    # Average height of people under 21 years of age who drink coffee
    coffeeHeightMeanUnder21 = bootSample.query("age == '<21' and drinks_coffee == True").height.mean()
    diff = nonCoffeeHeightMeanUnder21 - coffeeHeightMeanUnder21
    diffHeightListUnder21.append(diff)

# When the confidence level is 99%.7, the confidence interval for the average height difference

print("Lowerbound(0.3):", np.percentile(diffHeightListUnder21, 0.3))
print("Uppperbound(99.7):", np.percentile(diffHeightListUnder21, 99.7))
```

```
Lowerbound(0.3): 0.26628329321498534
Uppperbound(99.7): 2.7262103458140956
```

- diff = = nonCoffeeHeightMeanUnder21 - coffeeHeightMeanUnder21
- Mean, CI(99.7%) are all positive numbers.

⇒ At a 99.7% confidence level,

⇒ For people under 21, the non-coffee group is taller.

Lab: Bootstrap

```
# 4. # Average height difference between non-coffee drinkers and coffee drinkers among people over 21 years of age
diffHeightListOver21 = []
for _ in range(iterationNum):
    bootSample = df_sample.sample(200, replace=True)  # sampling with replacement
    # Average height of people over 21 years of age who do not drink coffee
    nonCoffeeHeightMeanOver21 = bootSample.query("age != '<21' and drinks_coffee == False").height.mean()
    # Average height of people over 21 years of age who drink coffee
    coffeeHeightMeanOver21 = bootSample.query("age != '<21' and drinks_coffee == True").height.mean()

    diff = nonCoffeeHeightMeanOver21 - coffeeHeightMeanOver21
    diffHeightListOver21.append(diff)

# When the confidence level is 99%.7, the confidence interval for the average height difference

print("Lowerbound(0.3):", np.percentile(diffHeightListOver21, 0.3))
print("Uppperbound(99.7):", np.percentile(diffHeightListOver21, 99.7))
```

```
Lowerbound(0.3): 0.4435255490158182 5
Uppperbound(99.7): 3.3338172914183075
```

- diff = = nonCoffeeHeightMeanOver21 - coffeeHeightMeanOver21
- Mean, CI(99.7%) are all positive numbers.

⇒ At a 99.7% confidence level,

⇒ For people over 21, the non-coffee group is taller.

Lab: Bootstrap

◆ Results summanry

- In the overall data, coffee drinkers were on average taller than non-coffee drinkers.
- Among people under 21, non-coffee drinkers had a taller average height.
- Among people over the age of 21, those who did not drink coffee had a taller average height.

Data	Coffee group	Non-coffer group
Overall	taller	
Under 21		taller
Over 21		taller

Lab: Comparison of parameters and bootstrap CIs

- A confidence interval is determined through bootstrapping, and if the observed value is within the confidence interval, the hypothesis is determined to be true.
- In this example, since we know the original data, we will find the difference between the actual means and see if the parameter is within the confidence interval.
- Including three more part for checking Simpson's Paradox

Lab: Comparison of parameters and bootstrap CIs

```
print("##### Height differences in the population #####")
# # 1. Average height difference between non-coffee drinkers and coffee drinkers
diffHeight = df[df['drinks_coffee'] == False].height.mean() - df[df['drinks_coffee'] == True].height.mean()
print("1. diffHeight : ",diffHeight)

# 2. Average height difference between people over 21 years old and under 21 years old
diffHeightByAge = df[df['age'] == '>=21'].height.mean() - df[df['age'] == '<21'].height.mean()
print("2. diffHeight : ",diffHeightByAge)

# 3. Average height difference between non-coffee drinkers and coffee drinkers among people under 21 years of age
diffHeightUnder21 = df.query("age == '<21' and drinks_coffee == False").height.mean() - df.query("age == '<21' and
drinks_coffee == True").height.mean()
print("3. diffHeight : ",diffHeightUnder21)

# 4. Average height difference between non-coffee drinkers and coffee drinkers among people over 21 years of age
diffHeightOver21 = df.query("age != '<21' and drinks_coffee == False").height.mean() - df.query("age != '<21' and
drinks_coffee == True").height.mean()
print("4. diffHeight : ",diffHeightOver21)
```

Lab: Bootstrap

Height differences in the population
1. diffHeight : -1.9568024933368093
2. diffHeight : 3.882291249920982
3. diffHeight : 1.69939009355123
4. diffHeight : 1.950935488978871

	Parameter (population)	Confidence Interval of Samples
1. Average height difference between non-coffee drinkers and coffee drinkers	-1.9568	-3.1054 ~ -0.8118
2. Average height difference between people over 21 years old and under 21 years old	3.8823	3.1356 ~ 5.0168
3. Average height difference between non-coffee drinkers and coffee drinkers among people under 21 years of age	1.6994	0.2663 ~ 2.7262
4. Average height difference between non-coffee drinkers and coffee drinkers among people over 21 years of age	1.9509	0.4435 ~ 3.3338

⇒ **Parameters are within the bootstrap confidence interval !!!**


Bye Bye