Introduction to Data Science with Python

WMASDS04

Week 04: Statistical Inference with Python

Outlines

- The Frequentist Approach,
- Measuring the Variability in Estimates
  - Point and Interval Estimates
- Hypothesis Testing
  - Testing Hypotheses Using Confidence Intervals
  - Testing Hypotheses Using p-Values

Terminologies

- Frequentist and Bayesian approach of inference
  - Populations and Samples
  - Parameters and Estimators
  - Prior and posterior probability
- Variables
  - Quantitative and Qualitative
  - Discrete and Continuous
  - Response and Explanatory
- Descriptive Statistics
  - Central Tendency [Mean, Median, Mode]
  - Dispersion [variance, standard deviation, coefficient of variation]
  - Skewness [symmetric, asymmetric, positively skewed, negatively skewed]
  - Kurtosis [leptokurtic, mesokurtic, platykurtic]
- Point Estimation and Interval Estimation
  - Confidence Interval
  - Confidence level
- Test of Hypothesis
  - Null Hypothesis and Alternative Hypothesis
  - Type I Error and Type II Error
  - Level of Significance and P-value
  - Power of test

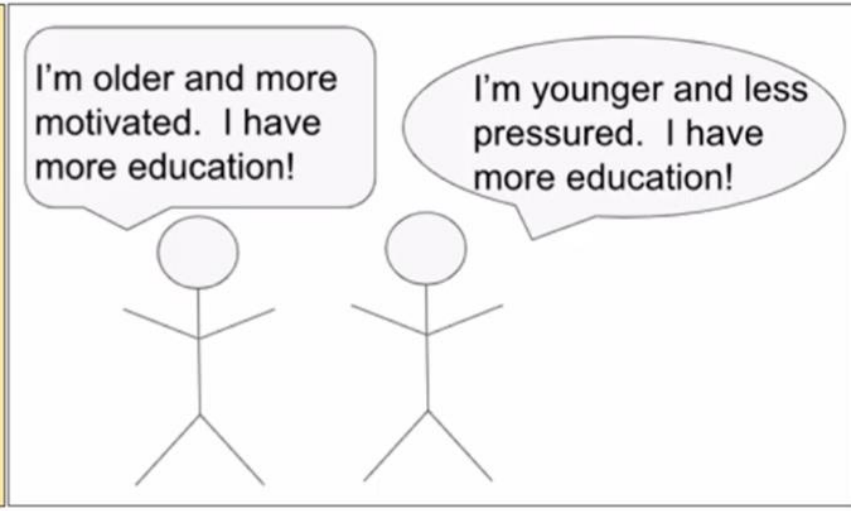
Research Question:
What proportion
of all parents with toddlers
report they use a car seat for all
travel with their toddler?

Estimate a Population Proportion with Confidence

Research Question:
What is the average cartwheel distance for all adults?

Estimate a Population Mean with Confidence

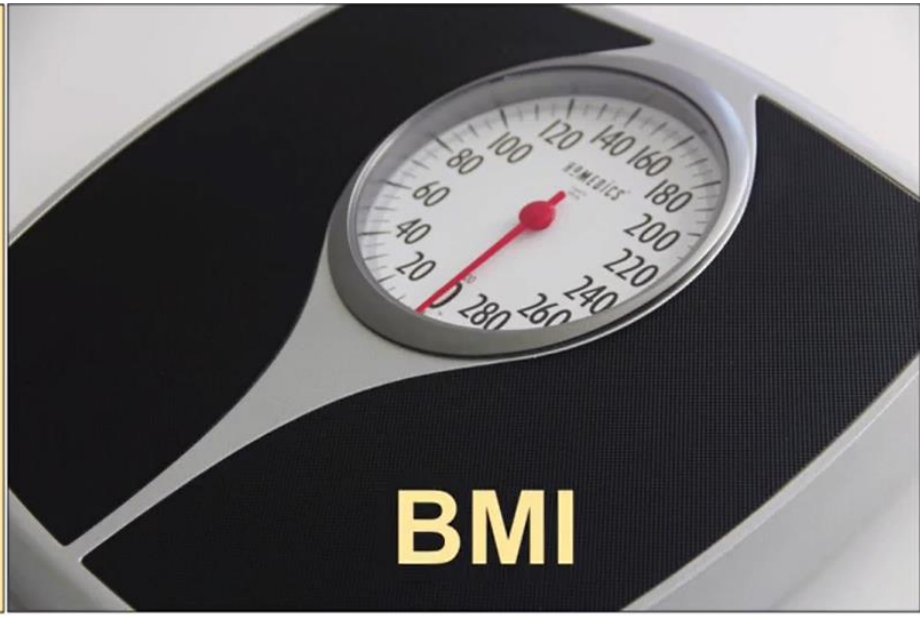
Research Question:
What is the average difference in self-reported education between older twin and younger twin?



Estimate a Population Mean Difference with Confidence

Research Question:

Considering Mexican-American adults (ages 18 - 29) living in the United States, do **males** have a **significantly higher mean** Body Mass Index than **females**?



Test a Theory about Two Population Means

Research Question:

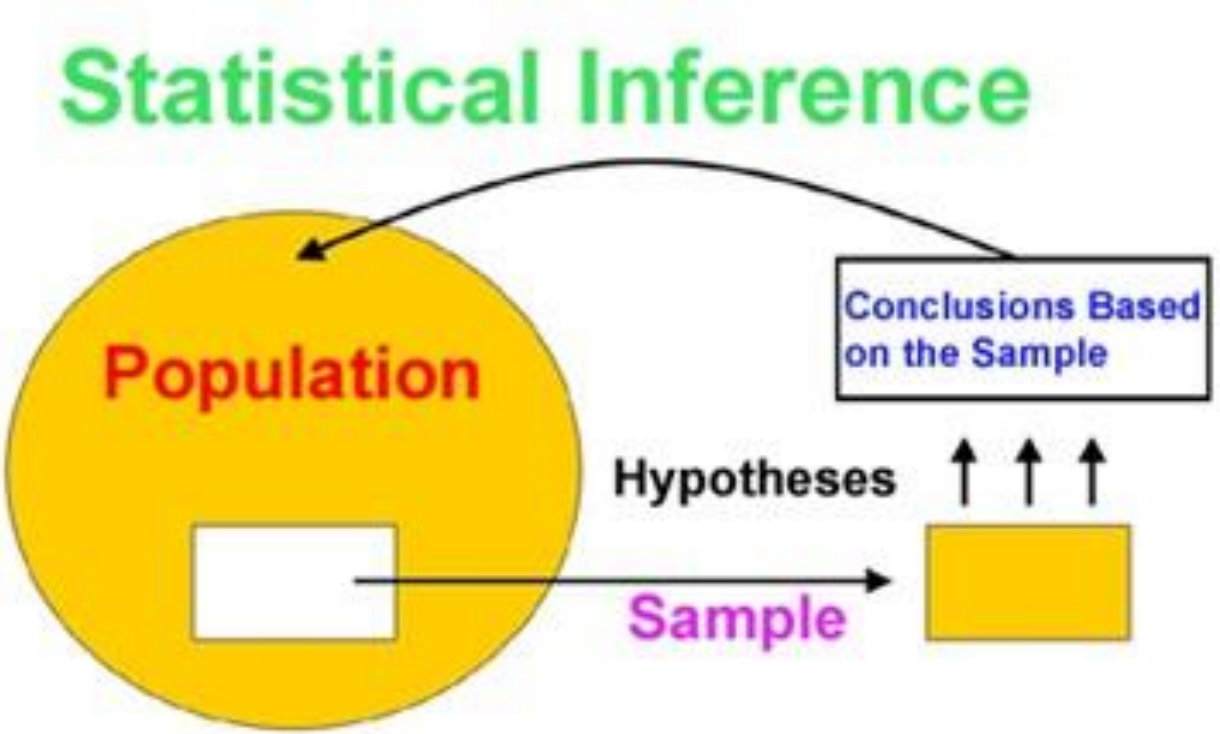
In previous years 52% of parents believed electronics and social media was cause of teenager's lack of sleep.

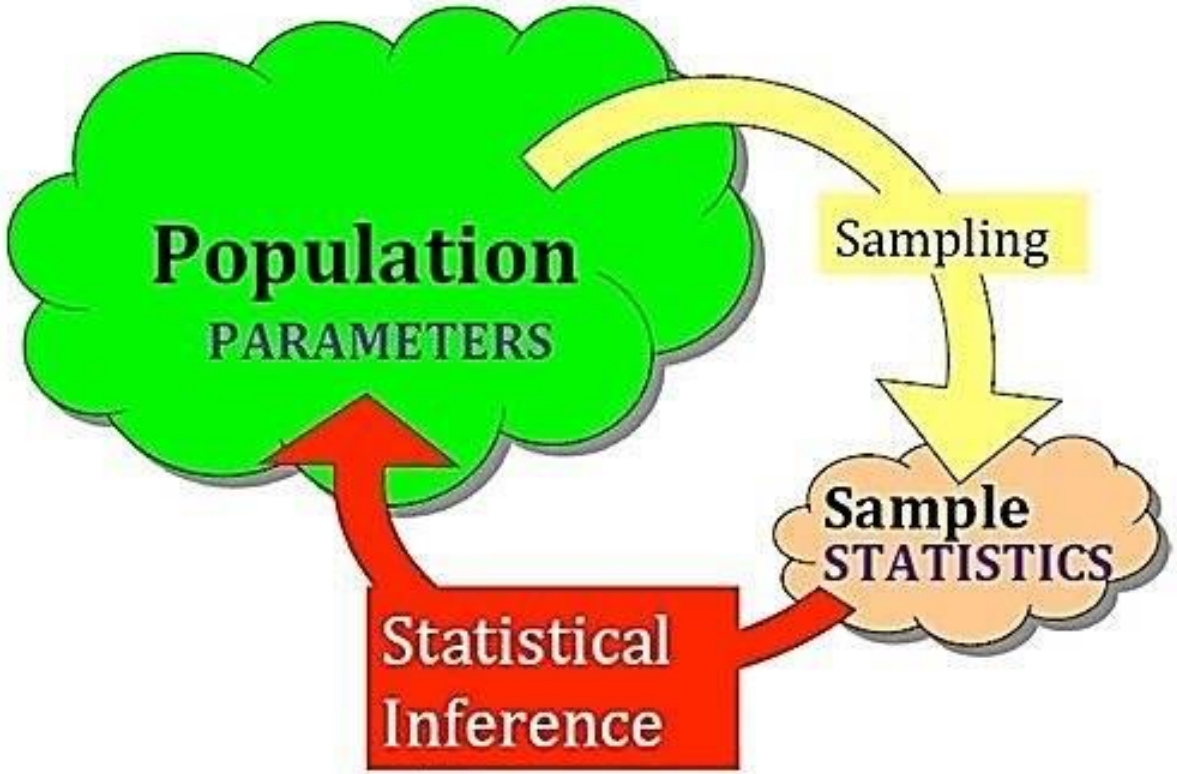
Do more parents today believe that their teenager's lack of sleep is caused due to electronics and social media?

Test a Theory about a Population Proportion

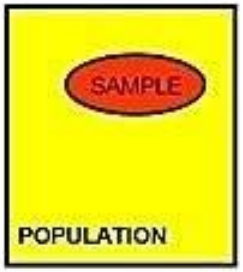
Process of Statistical Inference

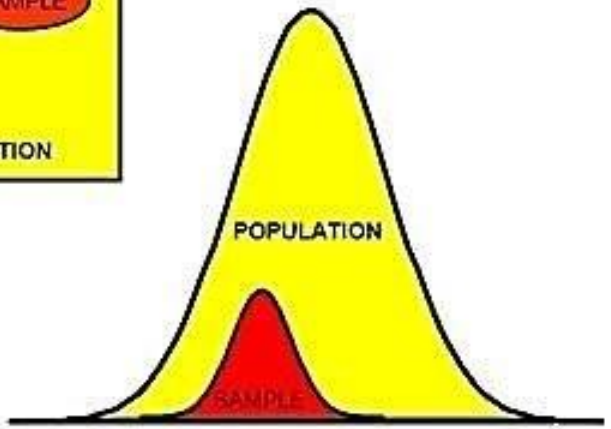
- .



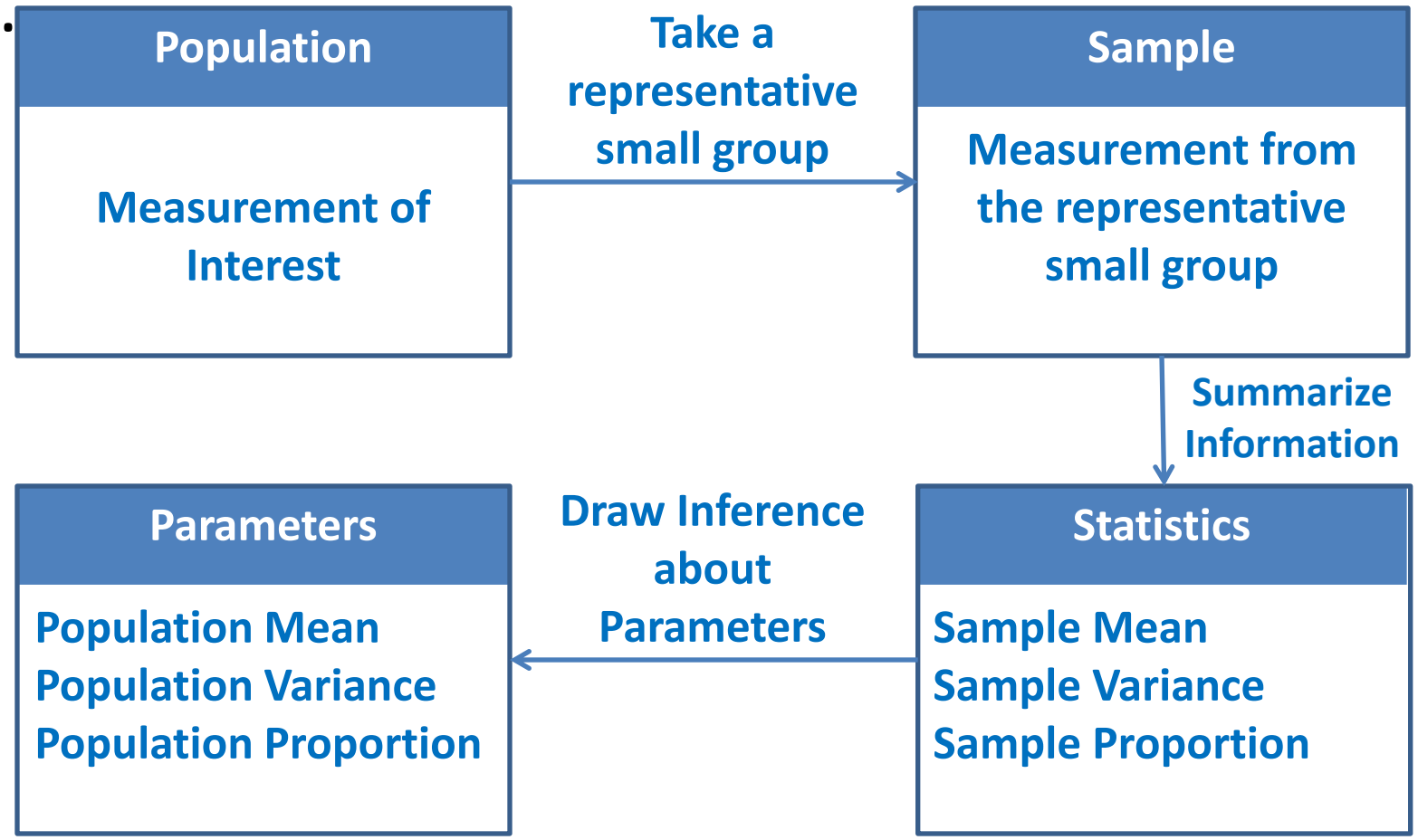


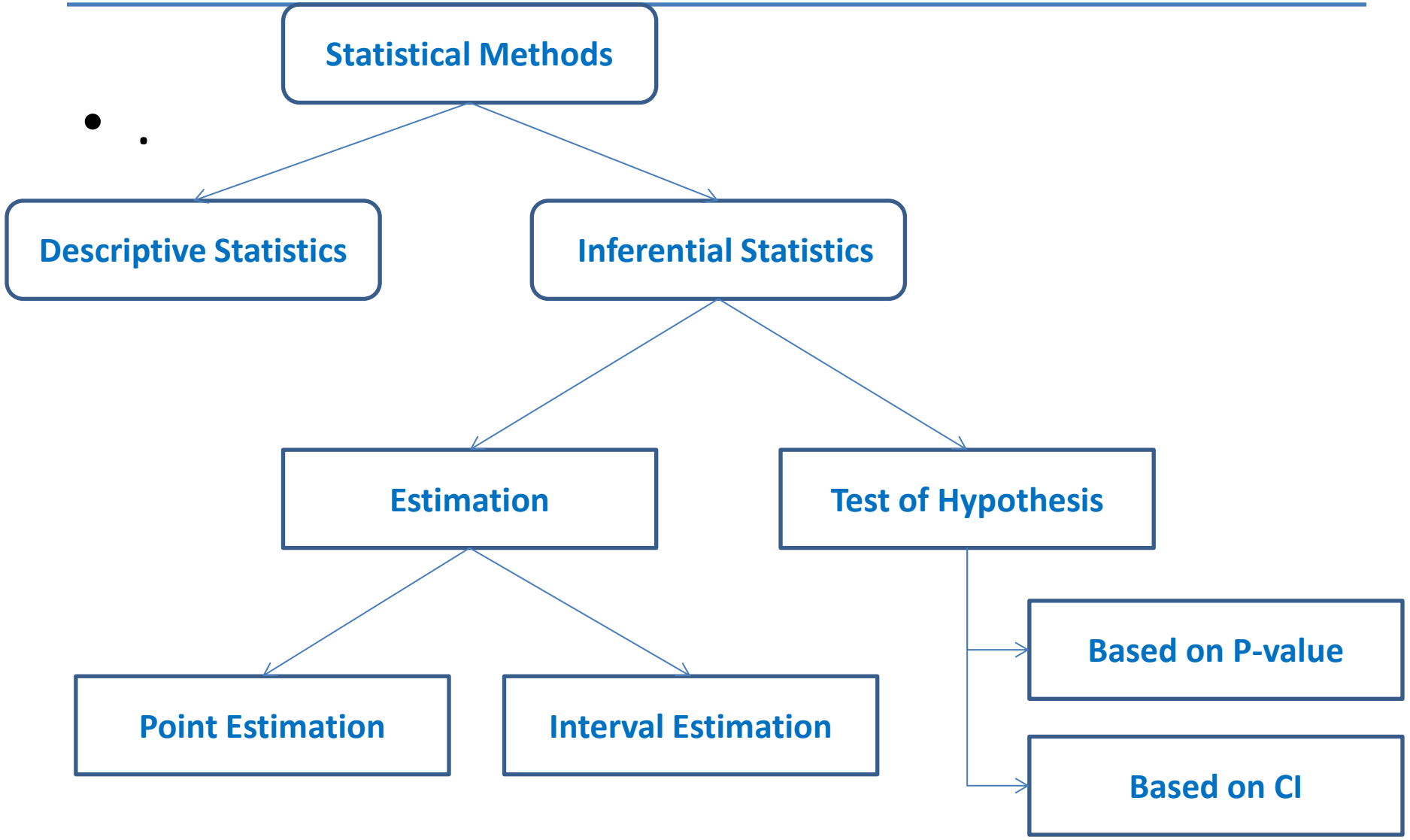
- by constructing **confidence intervals** on population parameters
- or by setting up a **hypothesis test** on a population parameter





The Process of Statistical Inference

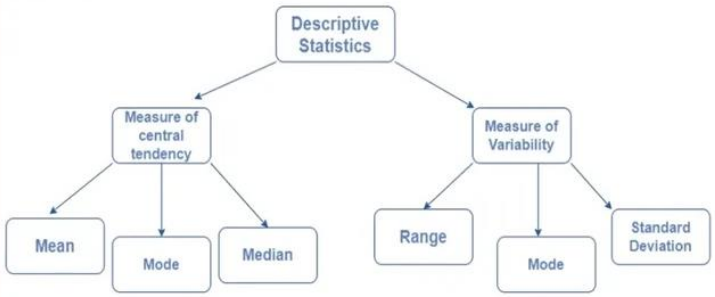


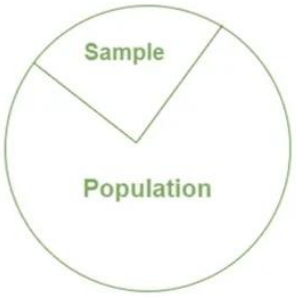

Statistical Methods
Descriptive Statistics
Inferential Statistics
Estimation
Test of Hypothesis
Point Estimation
Interval Estimation
Based on P-value
Based on CI

Descriptive and Inferential Statistics

- .

DESCRIPTIVE	INFERENTIAL
It is the analysis of data that helps to describe, show and summarize data under study	It is the analysis of random sample of data taken from a population to describe and make inference about the population
Organize, analyze and present data in a meaningful way	Compares, test and predicts data
It is used to describe a situation	It is used to explain the chance of occurrence of an event
It explain already known data and limited to a sample or population having small size	It attempts to reach the conclusion about the population
Types: Measure of central tendency & Measure of variability	Types: Estimation of parameters & Testing of hypothesis
Results are shown with help of charts, graphs, tables etc.	Results are shown with help of probability scores





Inference: Estimation and Test of Hypothesis

Point estimation

- Summarize the sample by a single value as an estimate of the population parameter.
- Ex. Average salary of junior data scientists is. 55,000 euros.

Interval estimation

- A range of values within which, we believe, the true population parameter lies with high probability.
- Ex. Average salary of junior data scientists is in the range of (52,000,55,000) With 95% confidence level.

Testing of Hypothesis

- To decide whether a statement regarding population parameter is true or false, based on sample data.
- Ex. Claim: Average salary of junior data scientists is greater than.50,000 euros annually.

Parameter vs. Statistic

- A **STATISTIC** is a descriptive measure computed from a **sample of data.**

- A **PARAMETER** is a descriptive measure computed from an **entire population** of data.

- **Inferential statistics** enables you to make an educated guess about a population parameter based on a statistic computed from a sample randomly drawn from that population.

What is Inference?

- Statistical inference is the process of generating conclusions about a population from a noisy sample.
  - In other words, statistical inference is a method of making decisions about the parameters of a population, based on random sampling.
- It helps to assess the relationship between the dependent and independent variables.
- The purpose of statistical inference to estimate the uncertainty or sample to sample variation.
- Without statistical inference we're simply living within our data.
- **With statistical inference, we're trying to generate new knowledge.**

Frequentist vs Bayesian Approach

- In the case of the *frequentist approach*,
  - the main assumption is that there is **a population**, which can be represented by several parameters, from which **we can obtain numerous random samples**. Population **parameters are fixed** but they are not accessible to the observer. The only way to derive information about these parameters is to take a sample of the population, to compute the parameters of the sample, and to use statistical inference techniques to make probable propositions regarding population parameters.
- The *Bayesian approach* is based on a consideration that
  - **data are fixed**, not the result of a repeatable sampling process, but **parameters describing data can be described probabilistically**. To this end, Bayesian inference methods focus on producing parameter distributions that represent all the knowledge we can extract from the sample and from prior information about the problem.

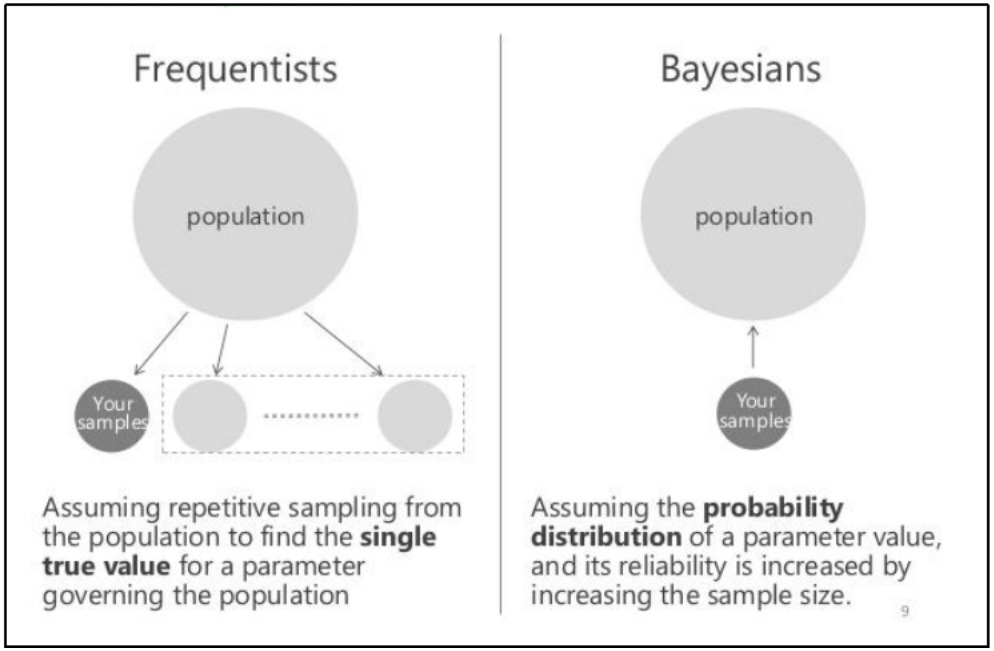
Frequentist Statistics	Bayesian Statistics
Parameters fixed	Parameters vary
Data varies	Data fixed
Probability P(D/Ø)	Likelihood P(Ø/D)
Confidence Interval	Credible Interval
No Prior	Strength of Prior

	Frequentist	Bayesian
Setup	θ is unknown and fixed	θ is a random variable
What do we want?	$\hat{\theta}$ is the best estimate of θ (but fixed)	$p(\theta \mid x, y)$ posterior distribution dictated by the data
What do we need?	Build a model using the data (x, y) *to* determine $\hat{\theta}$	Using the data determine $p(\theta)$ – prior probability
How we do it?	OLS: $\hat{\theta} = \frac{Cov(y, x_i)}{Var(x_i)}$ ML: $\mathcal{L}(x, \theta)$ likelihood function	$p(\theta \mid x, y) \propto L(x, y \mid \theta) * p(\theta)$ Find $p(\theta)$
Inference	H_0: $\theta = 0$ and H_a: $\theta \neq 0$ p-value is the probability that $\theta > \theta_c$	$p(\theta > \theta_c \mid x, y)$ read from posterior probability

Comparison of Frequentist Statistics & Bayesian

Frequentist

1. **Parameters are fixed** but unknown and **data are random**
2. Probability is a measure of **frequency of repeated events**



Bayesian

1. **Parameters are random** and **data are fixed**
2. Probability is a **degree of certainty about values**

Statistical Inference: The Frequentist Approach

- The ultimate objective of statistical inference is to produce probable propositions concerning population parameters from analysis of a sample.
  - about **point estimates**. A point estimate is a particular value that best approximates some parameter of interest. For example, the mean or the variance of the sample.
  - about **confidence intervals** or set estimates. A confidence interval is a range of values that best represents some parameter of interest.
  - about **the acceptance or rejection of a hypothesis**.
- In all these cases, the production of propositions is based on a simple assumption:
  - we can estimate the probability that the result represented by the proposition has been caused by chance.

Measuring the Variability in Estimates

- Sampling Distribution of Point Estimates
  - For example, resampling
- The Traditional Approach
  - For example, using Central Limit Theorem
- The Computationally Intensive Approach
  - For example, Bootstrapping

4.3.1.1 Sampling Distribution of Point Estimates

- The sample mean is a point estimate of the population mean.
- The problem we face is that estimates generally vary from one sample to another, and this sampling variation suggests our estimate may be close, but it will not be exactly equal to our parameter of interest. How can we measure this variability?
- Then, we can use the sampling distribution to compute a measure of the variability.
- It is very useful to think of a particular point estimate as being drawn from such a distribution.

Empirical distribution of the sample mean

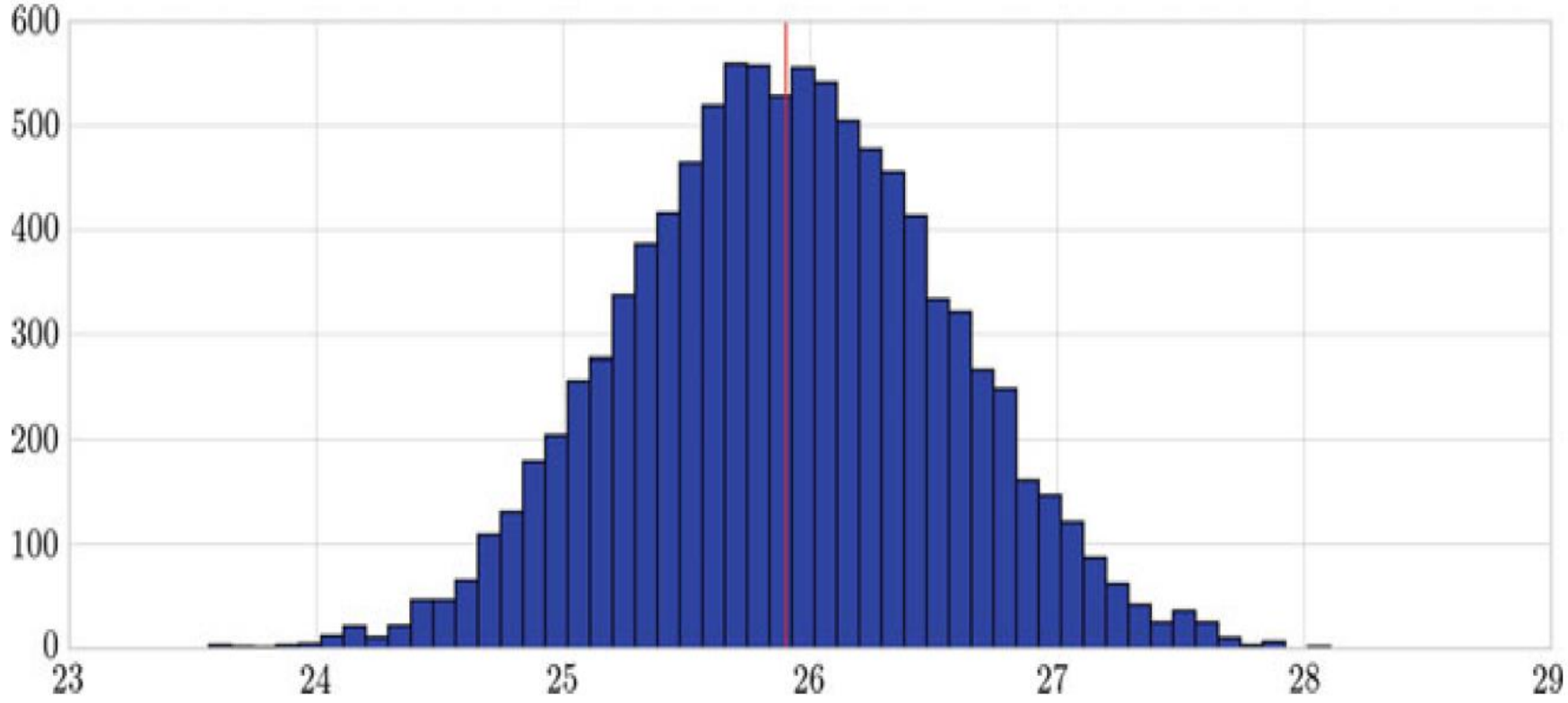


Fig. 4.1 Empirical distribution of the sample mean. In *red*, the mean value of this distribution

- In Fig. 4.1, we can see the empirical sample distribution of the mean for s = 10,000 samples with n = 200 observations from our dataset.
- This empirical distribution has been built in the following way:
  - draw all possible independent samples of a given size from a given population.
  - compute the sample mean for each sample.
  - Estimate the sampling distribution of sample mean by the empirical distribution of the sample replications.
- The probability distribution of this statistic (sample mean) is called the mean sampling distribution.

The Traditional Approach (CLT)

- The standard deviation of the sample mean σ^{-}_{x}, or standard error, can be approximated by this formula.

$$SE = \frac{\sigma_x}{\sqrt{n}}$$

- The demonstration of this result is based on the Central Limit Theorem.
- This formula uses the standard deviation of the population, which is not known, but it can be shown that if it is substituted by its empirical estimate, the estimation is sufficiently good, if n > 30 and the population distribution is not skewed.
- This allows us to estimate the standard error of the sample mean even if we do not have access to the population.
- Unlike the case of the sample mean, there is no formula for the standard error of other interesting sample estimates, such as the median.

The Computationally Intensive Approach (Bootstrapping)

- Let us consider from now that our full dataset is a sample from a hypothetical population.
- A modern alternative to the traditional approach to statistical inference is the bootstrapping method.
- In the bootstrap, we drawn observations with replacement from the original data to create a bootstrap sample or resample.
- Then, we can calculate the mean for this resample.
- By repeating this process a large number of times, we can build a good approximation of the mean sampling distribution.

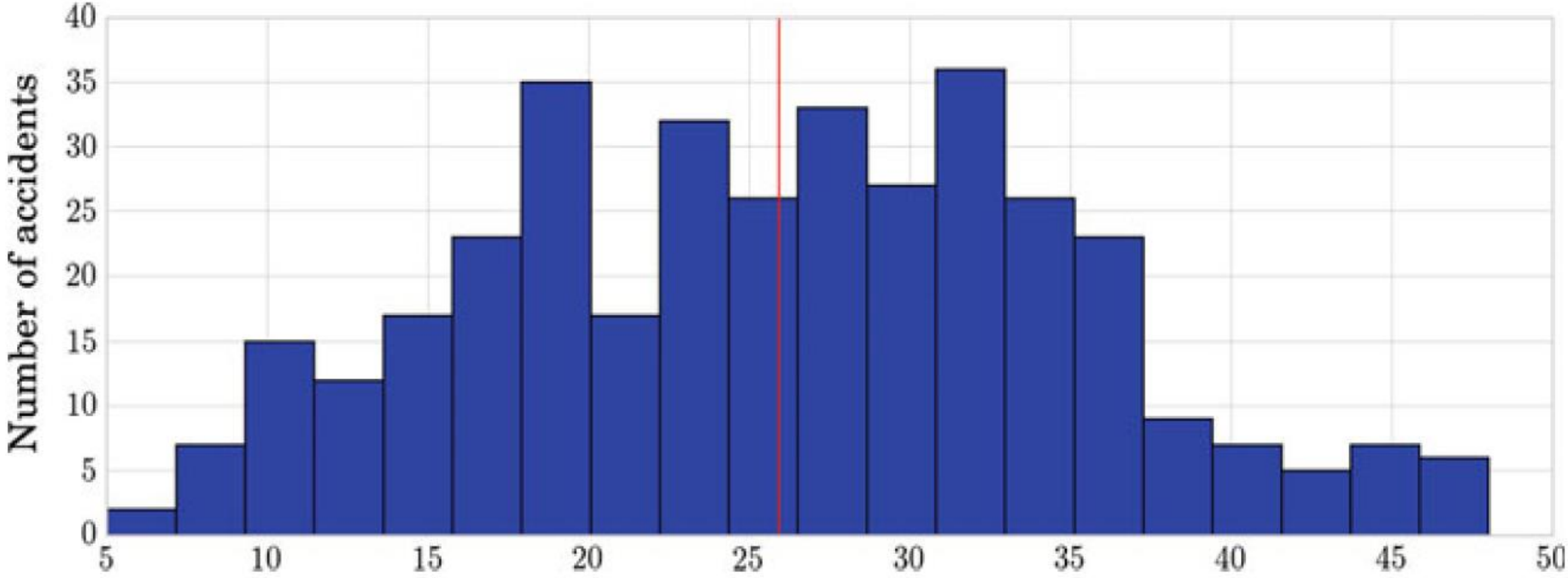
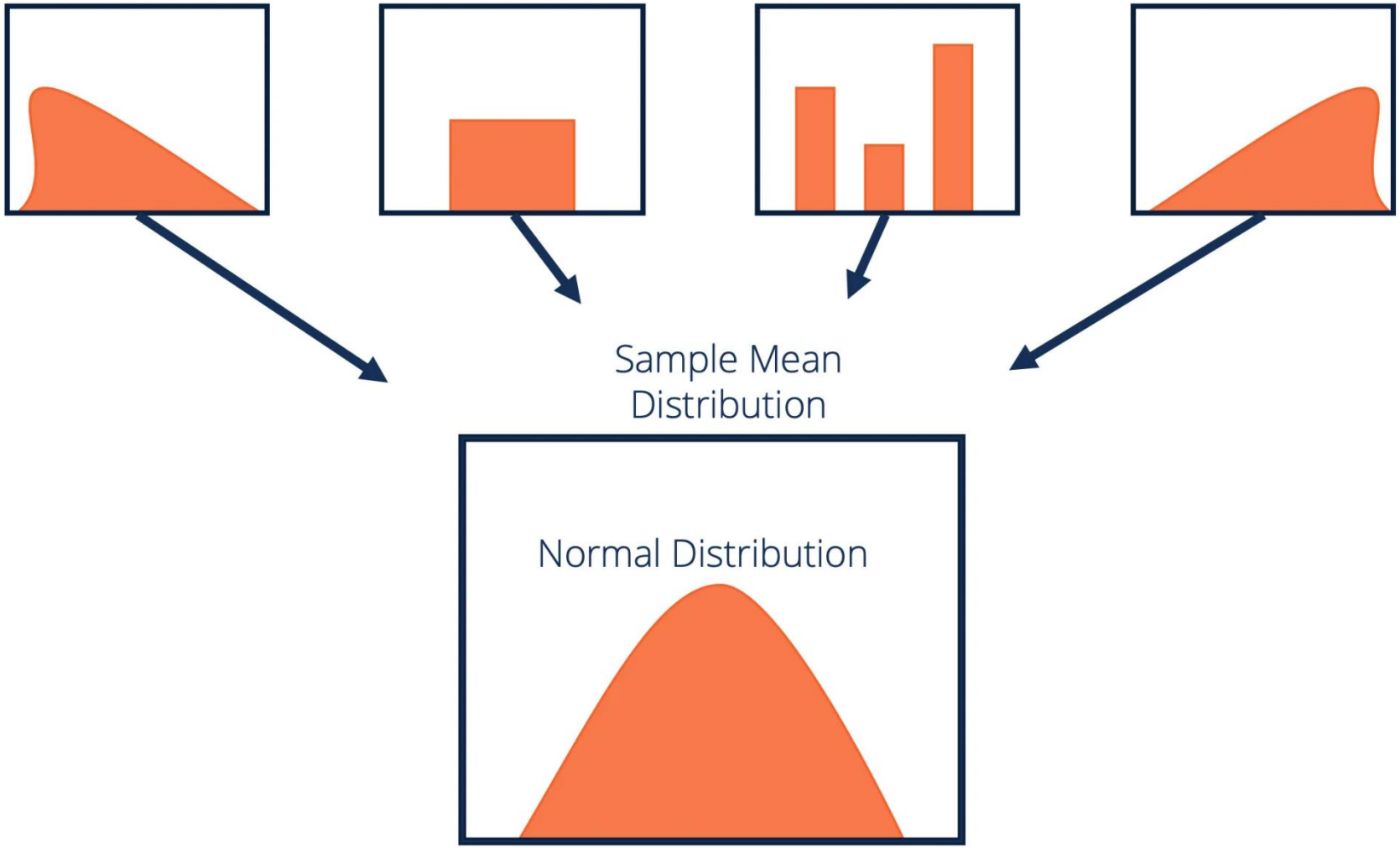


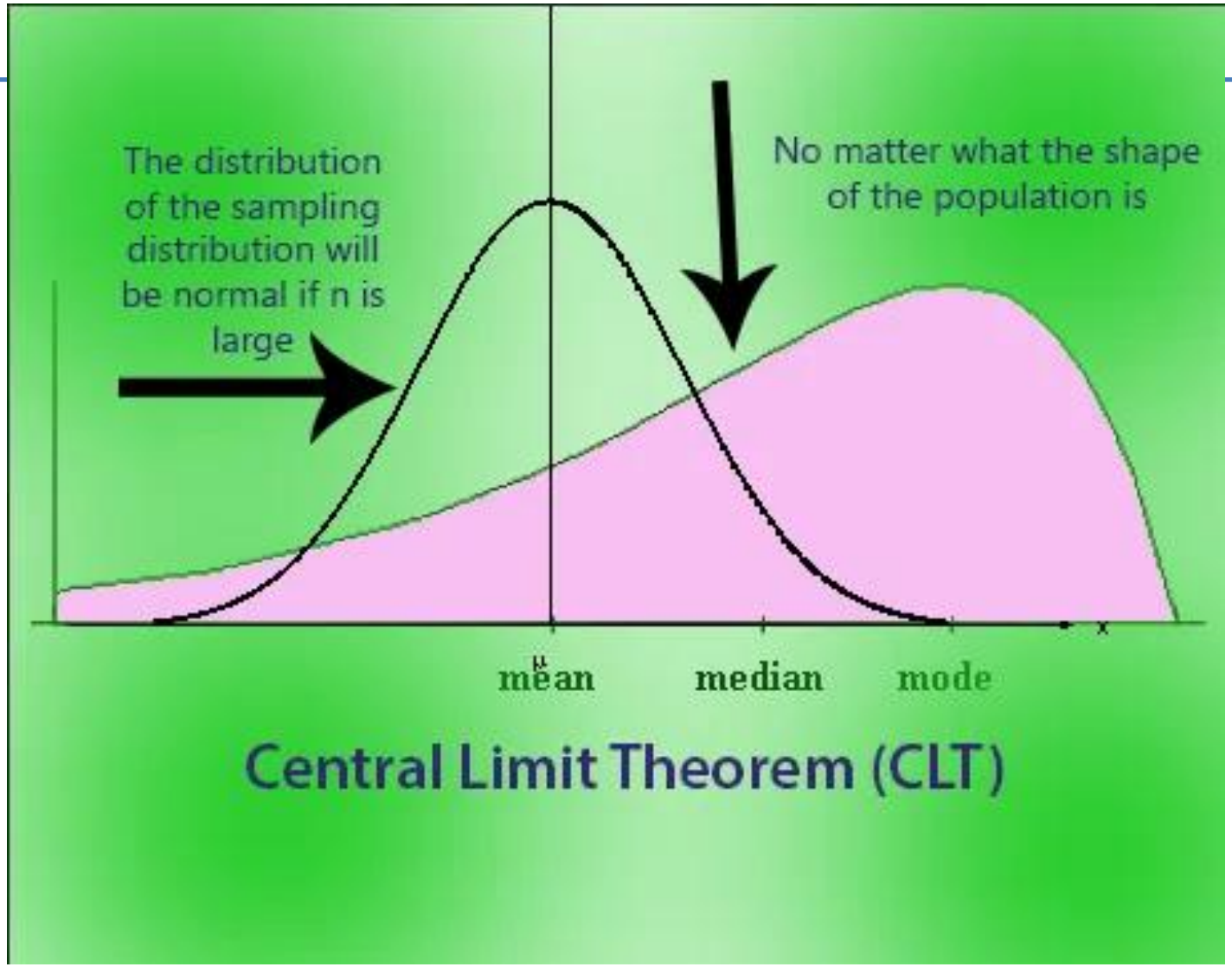
Fig. 4.2 Mean sampling distribution by bootstrapping. In *red*, the mean value of this distribution

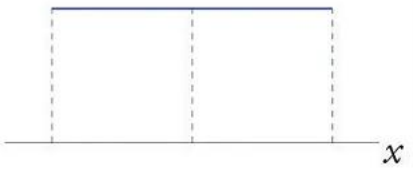
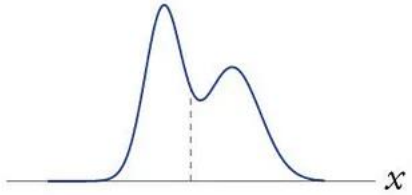
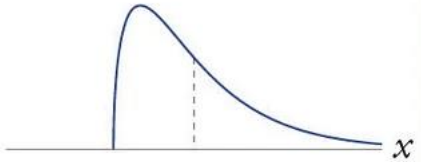
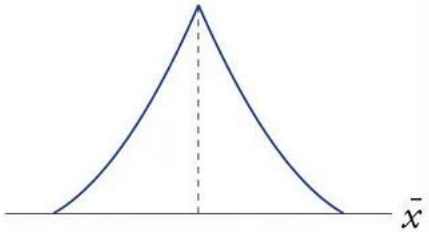
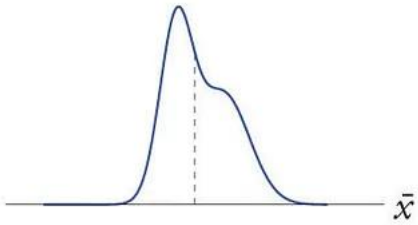
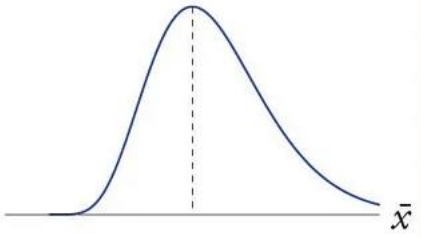
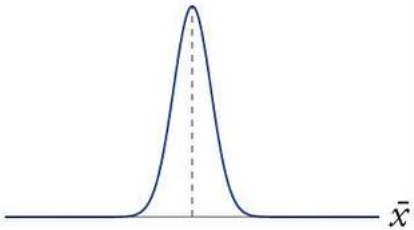
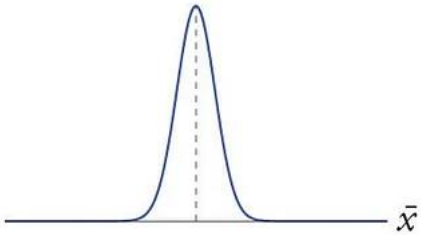
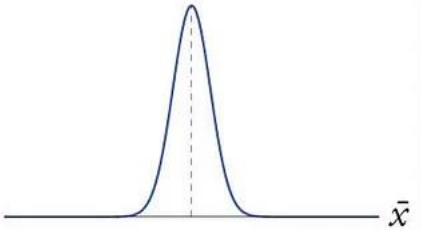
Central Limit Theorem

- **Theorem 4.1 :** Given a population with a finite mean $\boldsymbol{\mu}$ and a finite non-zero variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of σ^2/n as n, the sample size, increases.

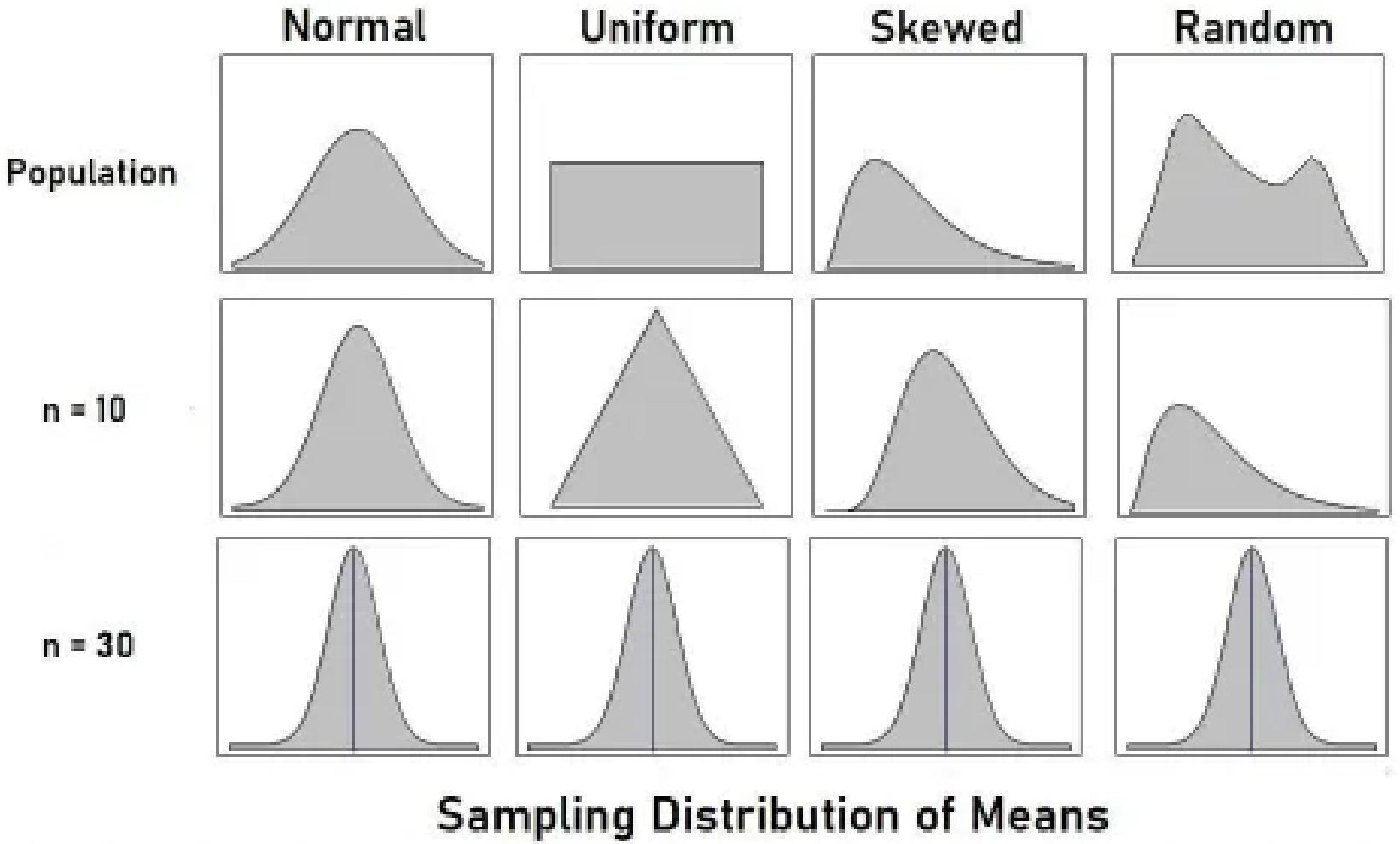

Sample Mean
Distribution
Normal Distribution

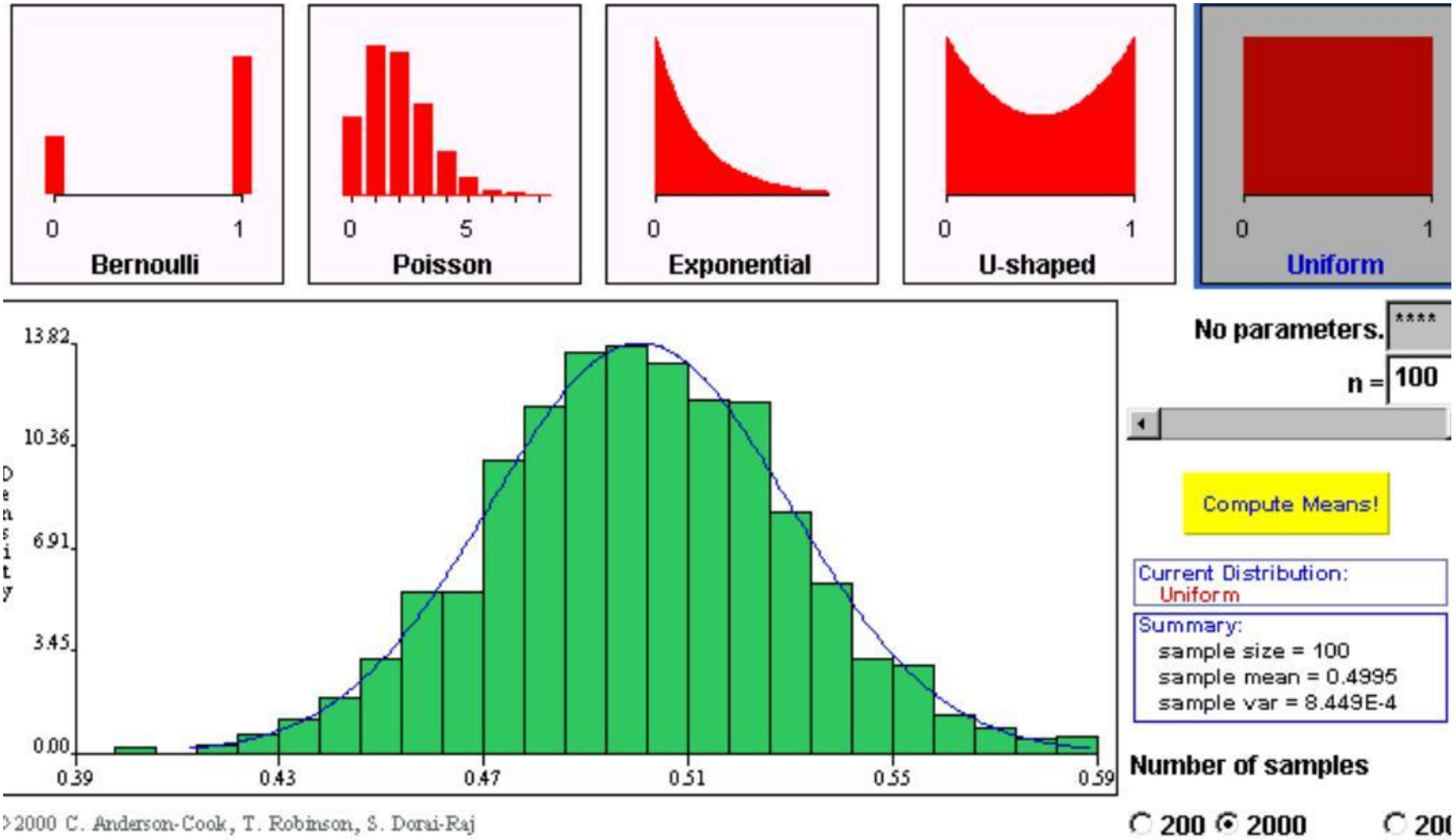
- .


The distribution of the sampling distribution will be normal if n is large
No matter what the shape of the population is
μ
mean
median
mode
x
Central Limit Theorem (CLT)

Population distribution			
Sampling distribution of $\overline{X}$ with $n = 5$			
Sampling distribution of $\overline{X}$ with $n = 30$			

CENTRAL LIMIT THEOREM




0
1
Bernoulli
0
5
Poisson
0
Exponential
0
1
U-shaped
0
1
Uniform
13.82
10.36
6.91
3.45
0.00
Density
0.39
0.43
0.47
0.51
0.55
0.59
©2000 C. Anderson-Cook, T. Robinson, S. Dorai-Raj
No parameters.
****
n =
100
Compute Means!
Current Distribution:
Uniform
Summary:
sample size = 100
sample mean = 0.4995
sample var = 8.449E-4
Number of samples
200
2000

Point Estimate and Confidence Interval

- A point estimate is the value of a statistic (estimator) that predicts the parameter value.
- An interval estimate is an interval of numbers around the point estimate, called confidence interval, within which the parameter value is believed to fall with a specified degree of confidence.

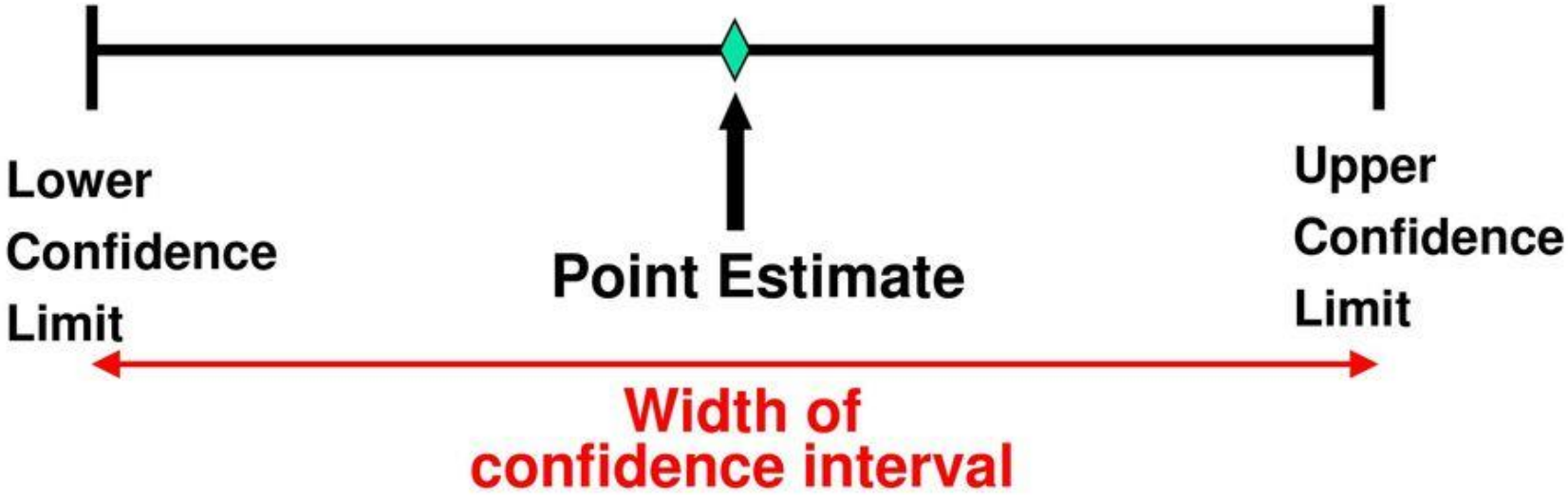
Point estimate vs confidence interval

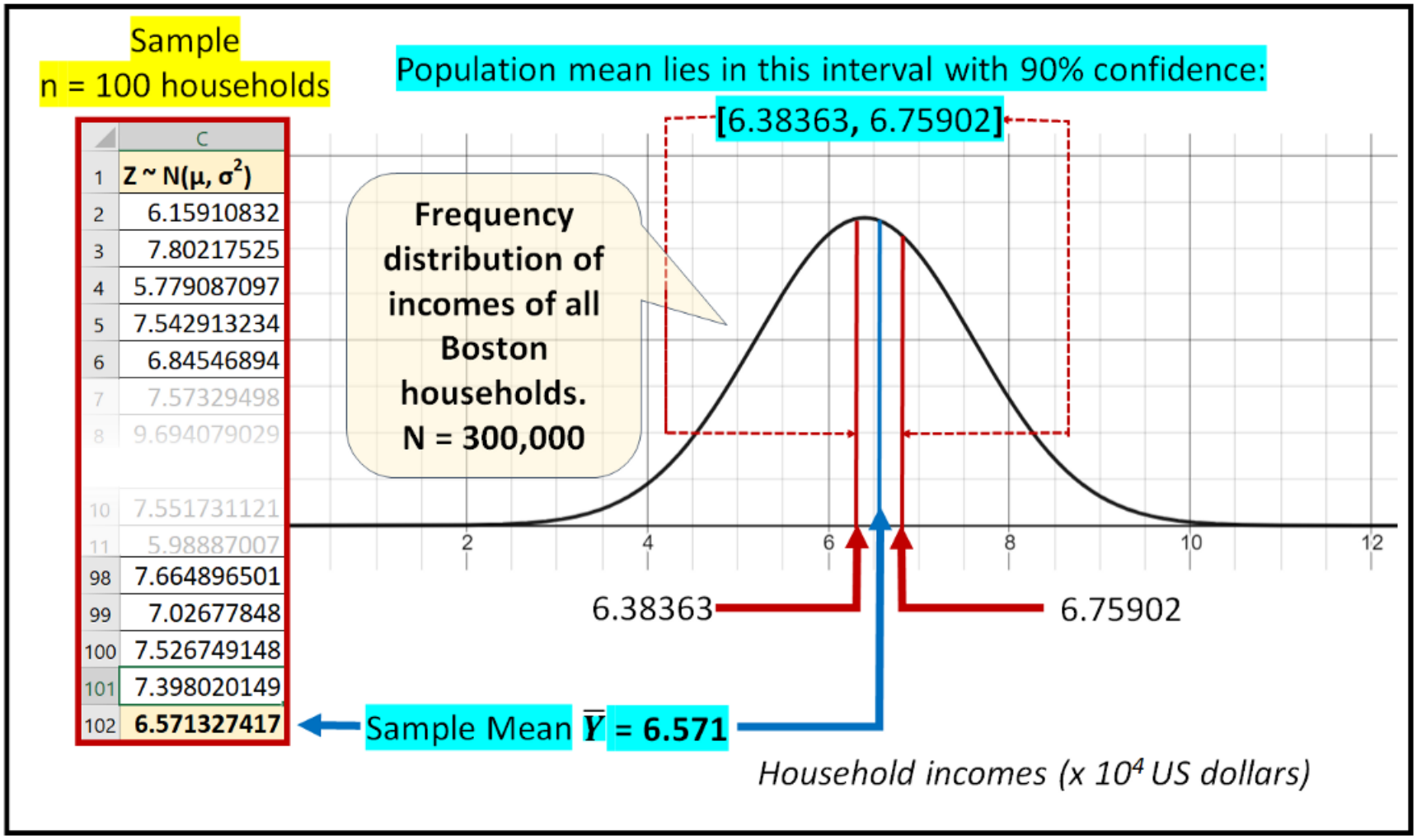
- A point estimate Θ, such as the sample mean, provides a single plausible value for a parameter. However, as we have seen, a point estimate is rarely perfect; usually there is some error in the estimate.
- That is why we have suggested using the standard error as a measure of its variability. Instead of that, a next logical step would be to provide a plausible range of values for the parameter.
- A plausible range of values for the sample parameter is called a confidence interval.

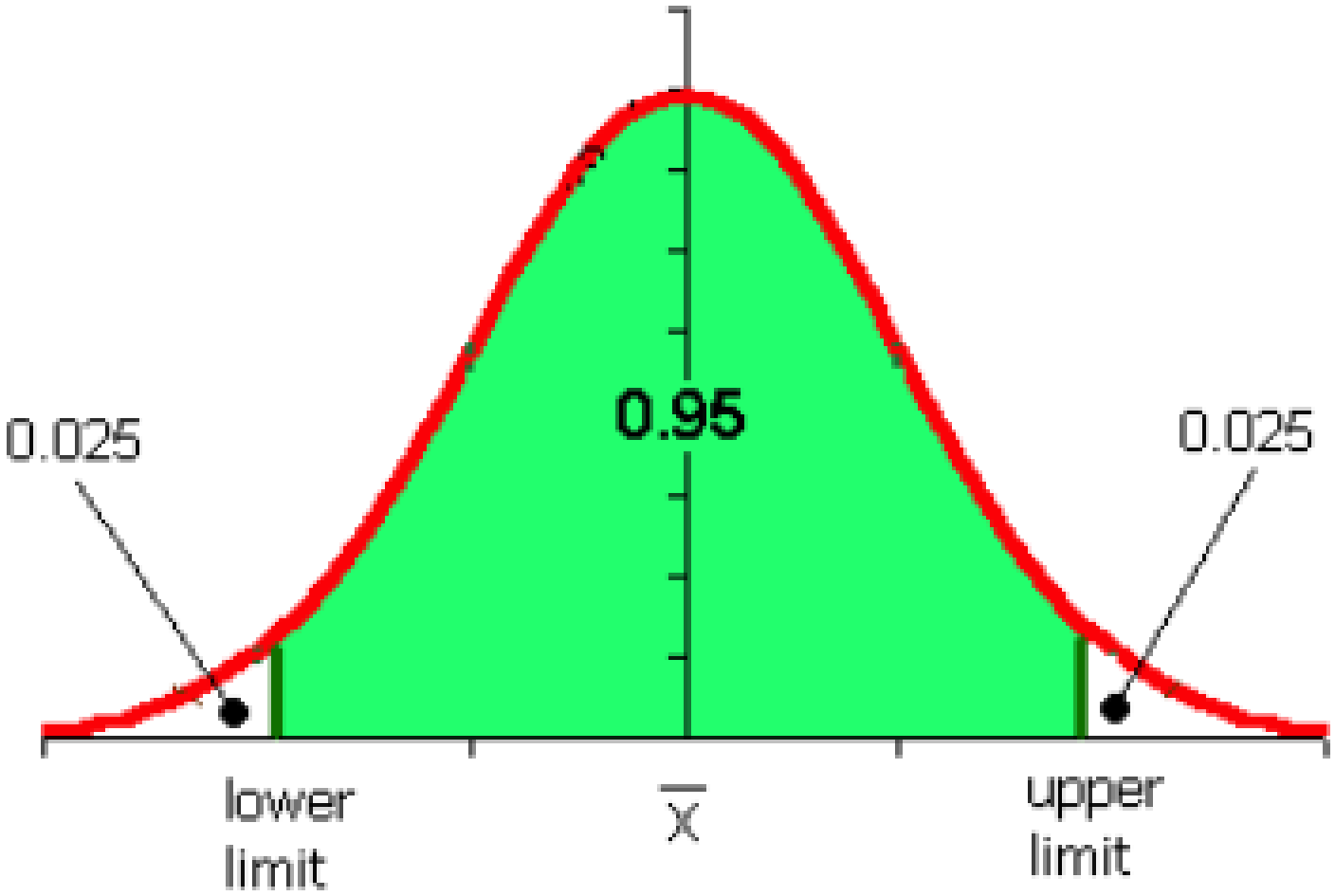
Point and Interval Estimates

DCOVA

- A point estimate is a single number,
- a confidence interval provides additional information about the variability of the estimate




Sample
n = 100 households
Population mean lies in this interval with 90% confidence:
[6.38363, 6.75902]
Frequency distribution of incomes of all Boston households. N = 300,000
C
Z ~ N(μ, σ²)
6.15910832
7.80217525
5.779087097
7.542913234
6.84546894
7.57329498
9.694079029
7.551731121
5.98887007
7.664896501
7.02677848
7.526749148
7.398020149
6.571327417
2
4
6
8
10
12
6.38363
6.75902
Sample Mean $\overline{Y}$ = 6.571
Household incomes (x 10⁴ US dollars)


0.025
0.95
0.025
lower
limit
$\overline{x}$
upper
limit

Confidence Intervals

- In general, if the point estimate follows the normal model with standard error *SE*, then a confidence interval for the population parameter is

$$\Theta \pm z \times SE$$

- where *z* corresponds to the confidence level selected:
- .

Confidence_Level	90%	95%	99%	99.9%
z_Value	1.65	1.96	2.58	3.291

95% CI of mean using bootstrapping:

- We compute a 95% confidence interval of the sample mean using bootstrapping:

1. Repeat the following steps for a large number, *s*, of times:
   a. Draw ***n*** observations with replacement from the original data to create a bootstrap sample or resample.
   b. Calculate the mean for the resample.
2. Calculate the mean of your *s* values of the sample statistic. This process gives you a "bootstrapped" estimate of the sample statistic.
3. Calculate the standard deviation of your s values of the sample statistic. This process gives you a "bootstrapped" estimate of the SE of the sample statistic.
4. Obtain the 2.5th and 97.5th percentiles of your ***s*** values of the sample statistic.

What Does“95% Confident ” Mean?

- **Suppose we took many (infinite) samples from a population and built a 95% confidence interval from each sample. Then about 95% of those intervals would contain the actual parameter.**
  - In the next figure, we show how many CI computed from 100 different samples of 100 elements from a dataset contain the real population mean.
  - If this simulation could be done with infinite different samples, 5% of those intervals would not contain the true mean.
- *We **cannot** say either that our specific sample contains the true parameter or that the interval has a 95% chance of containing the true parameter. That interpretation would not be correct under the assumptions of traditional statistics.*

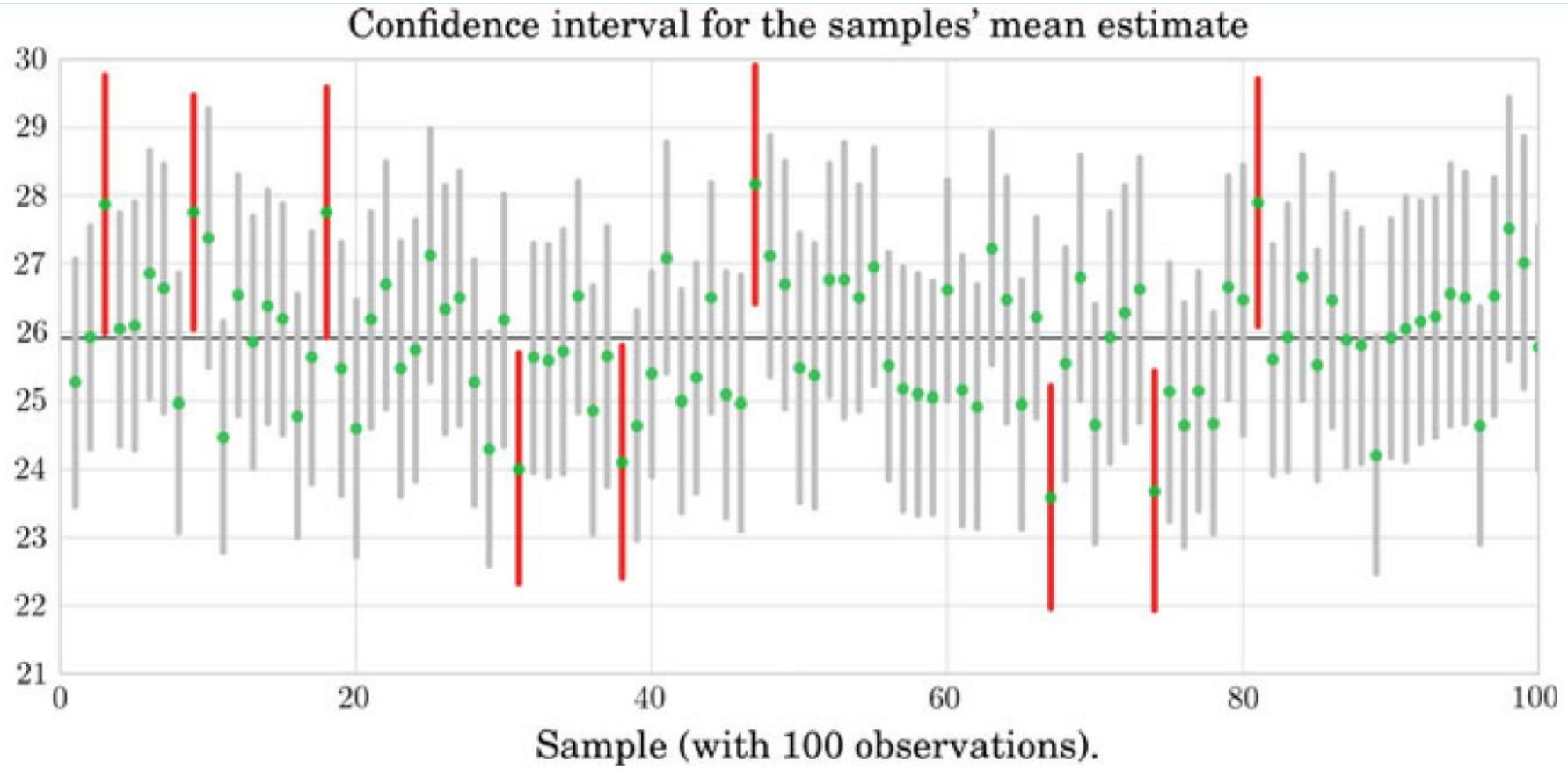
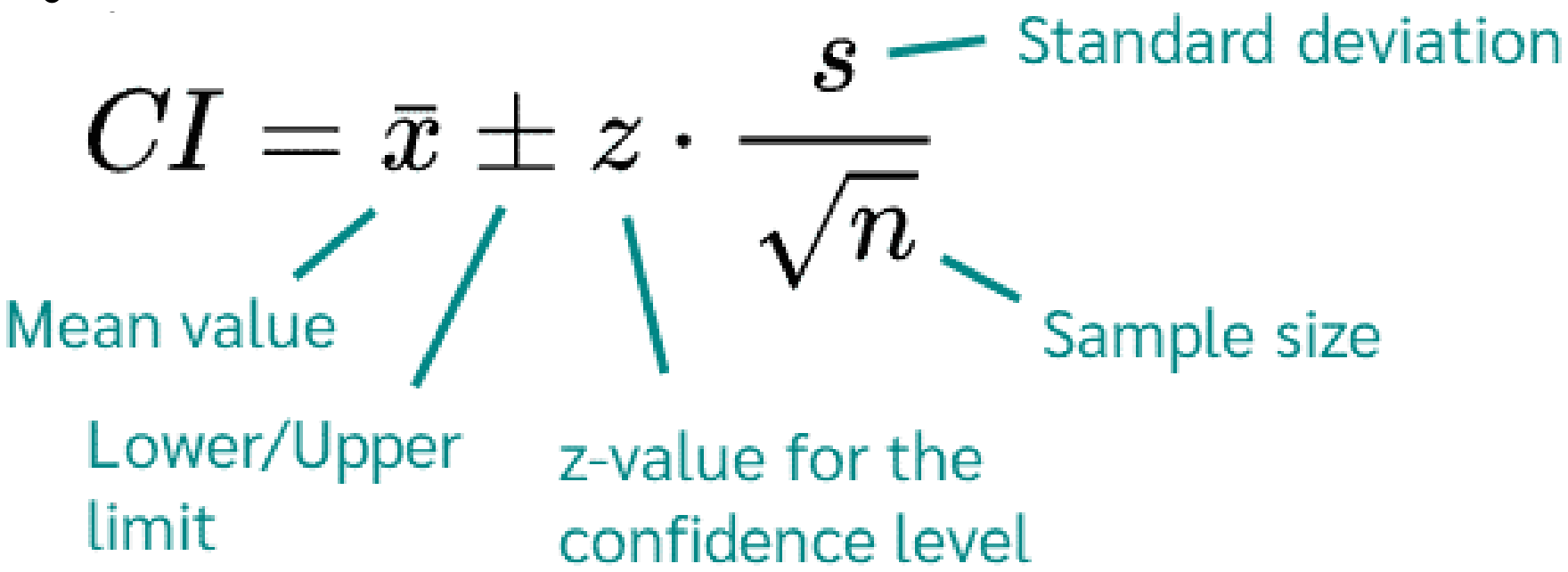


Fig. 4.3 This graph shows 100 sample means (*green points*) and its corresponding confidence intervals, computed from 100 different samples of 100 elements from our dataset. It can be observed that a few of them (those in *red*) do not contain the mean of the population (*black horizontal line*)

Interpretation of confidence interval

- The 95% confidence interval
- If the confidence interval method were used in an analyzing a large number of data sets (e.g., 100 separate samples from same population), then in the long run 95% of the confidence intervals would contain the true parameter values.

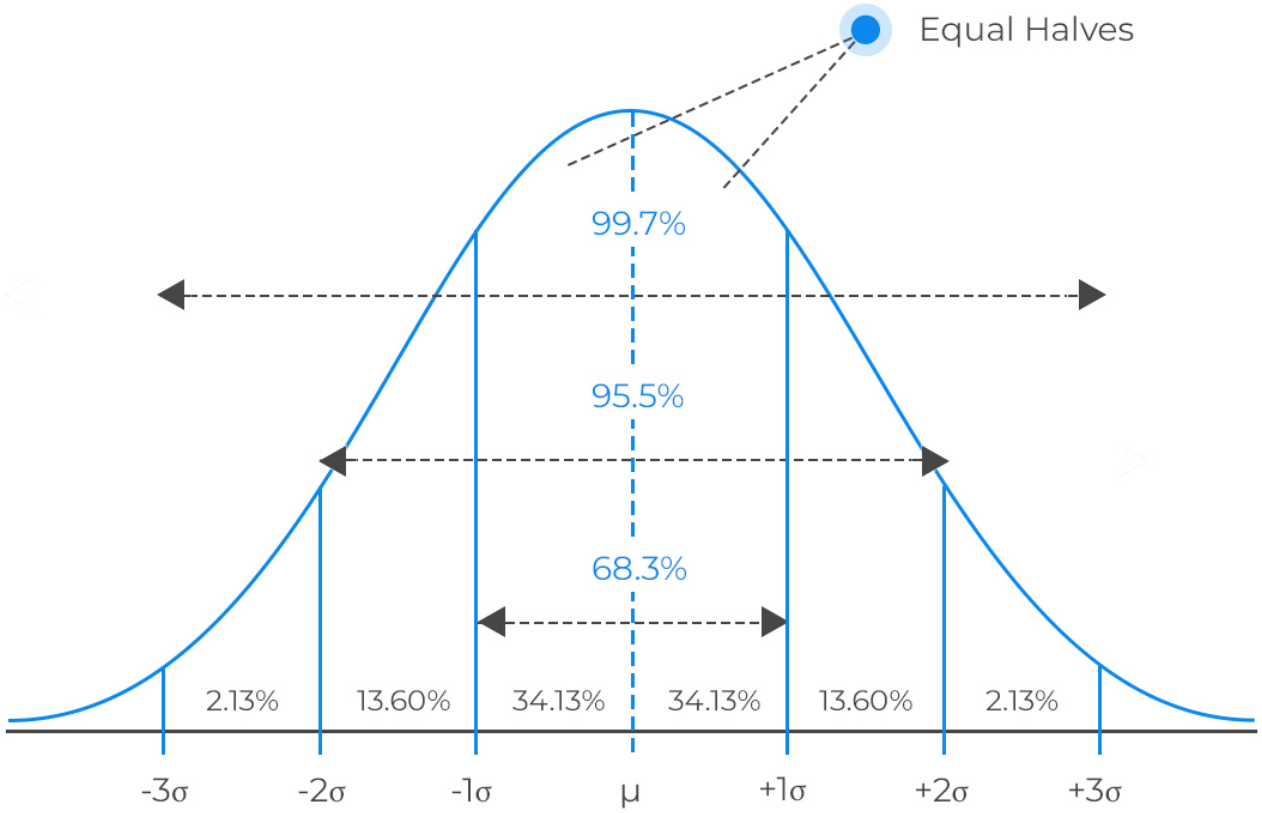
Confidence Interval for Mean

- 

Confidence Interval for Proportion

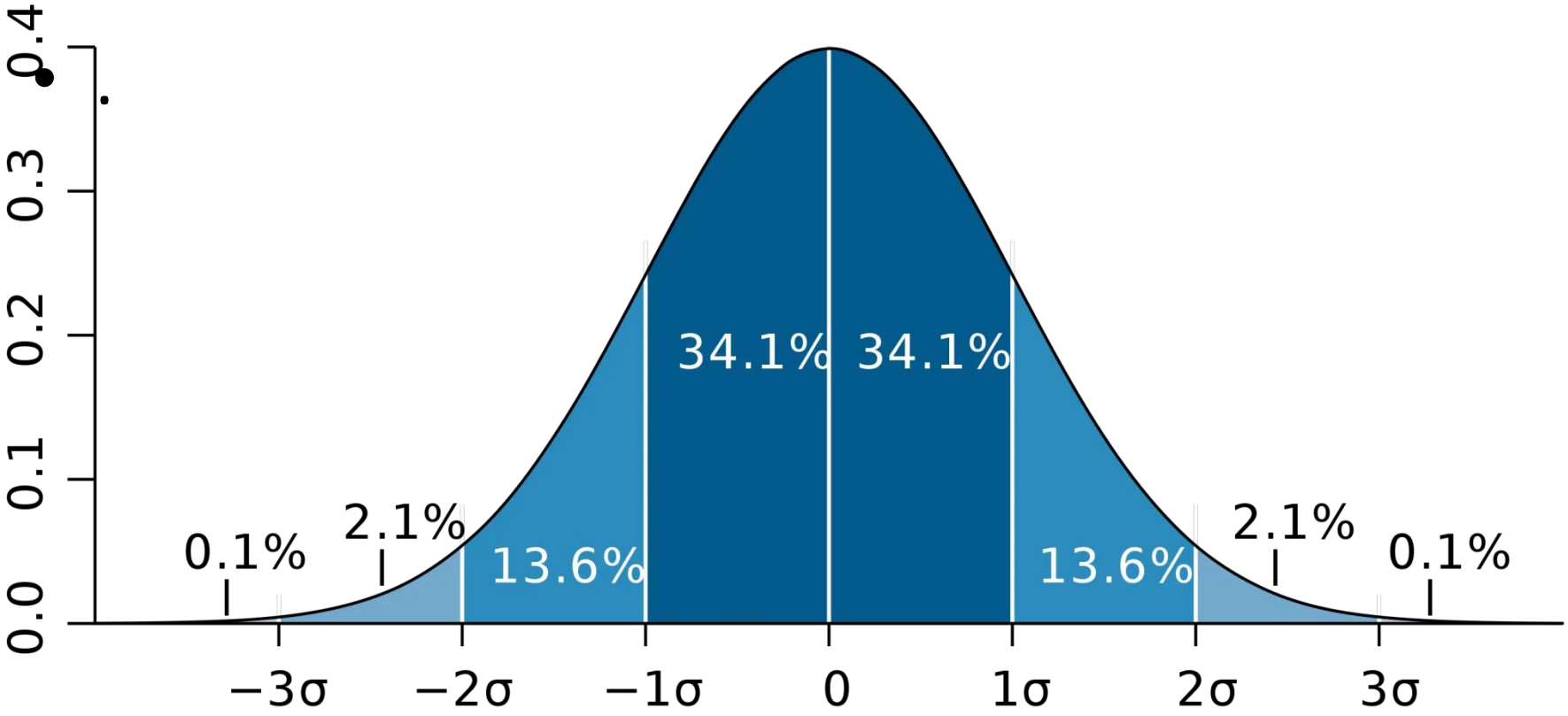
$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Shape of the normal distribution



No. of standard deviations from the mean

Shape of Normal Distribution



Hypothesis Testing

- The question about the estimate is: Are the observed effects real or not?
- Or, Were the observed effects statistically significant?
- The process of determining the statistical significance of an effect is called hypothesis testing.
- This process starts by simplifying the options into two competing hypotheses:
  - H_0: The mean number of daily traffic accidents is the same in 2010 and 2013
  - H_A: The mean number of daily traffic accidents in 2010 and 2013 is different

- H_0 the null hypothesis and it represents a skeptical point of view: the effect we have observed is due to chance (due to the specific sample bias).
- H_A is the alternative hypothesis and it represents the other point of view: the effect is real.
- The general rule of frequentist hypothesis testing: we will not discard H_0 (and hence we will not consider H_A) unless the observed effect is implausible under H_0.

Testing Hypotheses Using Confidence Intervals

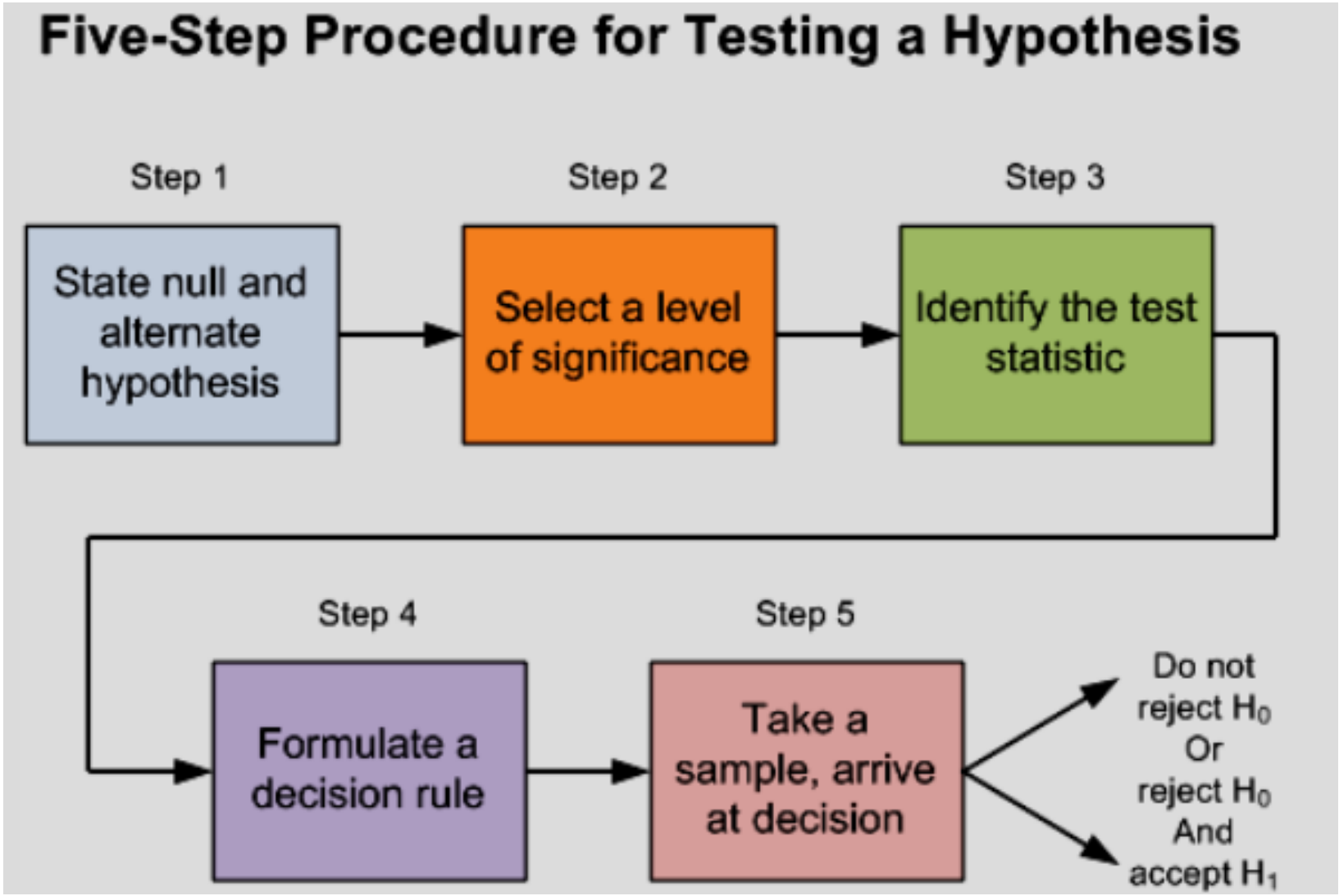
- **Hypothesis testing is built around rejecting or failing to reject the null hypothesis.**
  - **That is, we do not reject H_0 unless we have strong evidence against it.**
- But what precisely does strong evidence mean?
  - As a general rule of thumb, for those cases where the null hypothesis is actually true, we do not want to incorrectly reject H_0 more than 5% of the time.
  - This corresponds to a significance level of $\alpha = 0.05$.
- In this case, the correct interpretation of our test is as follows:
  - If we use a 95% confidence interval to test a problem where the null hypothesis is true, we will make an error whenever the point estimate is at least 1.96 standard errors away from the population parameter.
  - This happens about 5% of the time (2.5% in each tail).

Testing Hypotheses Using p-Values

- The P-value is the probability, presuming that H_0 is true, that the test statistic equals the observed value or a value even more extreme in the direction predicted by H_a.
- A small P-value, such as 0.01, means that the observed data would have been unusual, if H_0 were true.
  - the result is declared statistically significant (at 1% level of significant).
- A moderate to large *P*-value, such as 0.26 or 0.83, means the data are consistent with H_0; if H_0 were true, the observed data would not be unusual.
- Smaller P-values reflect stronger evidence against H_0.

- The goal of classical hypothesis testing is to answer the question,
  - "Given a sample and an apparent effect, what is the probability of seeing such an effect by chance?"
- Here is how we answer that question:
  - The first step is to quantify the size of the apparent effect by choosing a test statistic.
  - The second step is to define a null hypothesis, which is a model of the system based on the assumption that the apparent effect is not real.
  - The third step is to compute a p-value, which is the probability of seeing the apparent effect if the null hypothesis is true.
  - The last step is to interpret the result. **If the p-value is low, the effect is said to be statistically significant, which means that it is unlikely to have occurred by chance. In this case we infer that the effect is more likely to appear in the larger population.**

Process of Test of Hypothesis



How to Define Null and Alternative?

- .

Null Hypothesis (H_0)

-The assumption you're beginning with

-The opposite of what you're testing

Alternative Hypothesis (H_1)

-The claim you're testing

How to Define Null and Alternative?

- **Null Hypotheses**
  - H_0: Put here what is **typical** of the population, a term that characterizes "business as usual" where nothing out of the ordinary occurs.
- **Alternative Hypotheses**
  - H_1: Put here what is the **challenge**, the view of some characteristic of the population that, if it were true, would trigger some new action, some change in procedures that had previously defined "business as usual."

Type I Error(Significance Level) and Type II Error

Key:
Outcome (Probability)

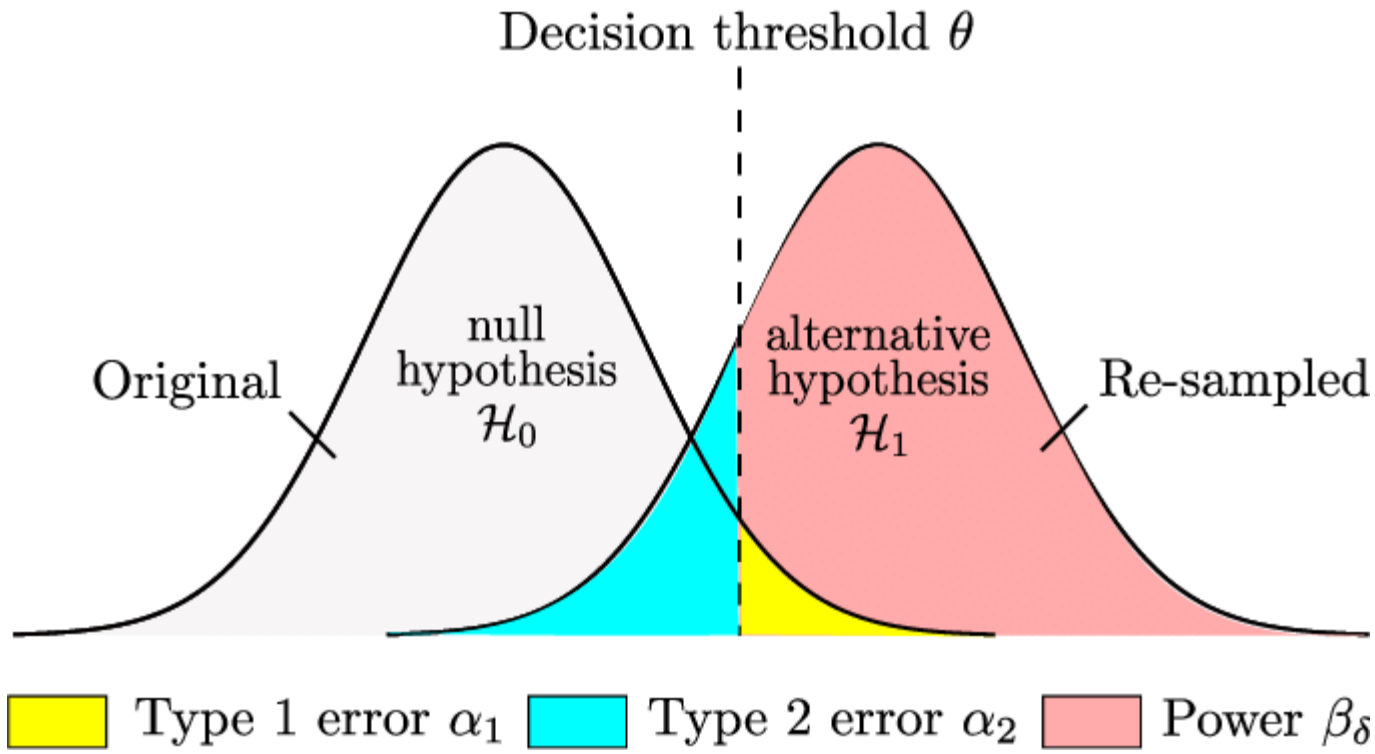
	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject H_0	**No error ($1 - \alpha$)**	**Type II Error (β)**
Reject H_0	**Type I Error (α)**	**No Error ($1 - \beta$)**

- β denotes the probability of Type II Error
- $1 - \beta$ is defined as the power of the test

Power = $1 - \beta$ = the probability that a false null hypothesis is rejected

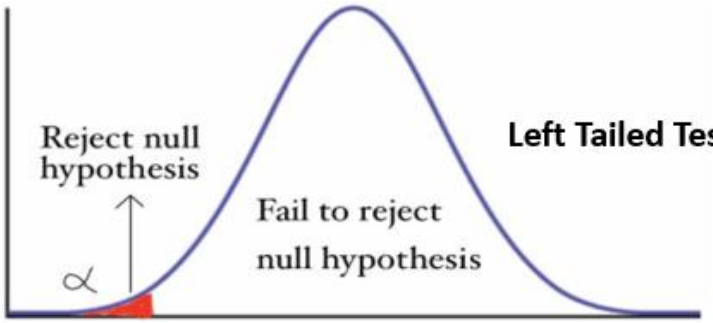
Distribution of Test Statistics under H_0 and H_1

- .



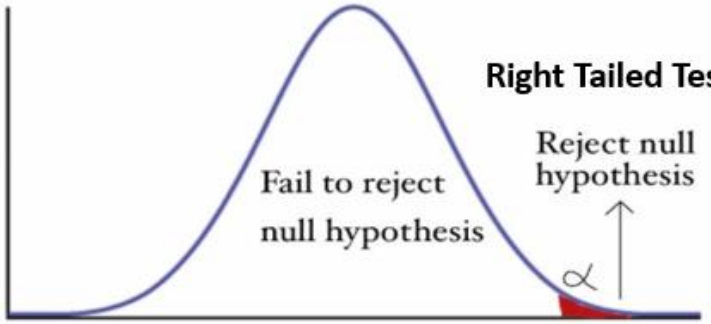
Two-sided vs one-sided tests

Null Hypothesis(H0) : $\mu \geq value$
Alternative Hypothesis(H_A) : $\mu < value$



Left Tailed Test

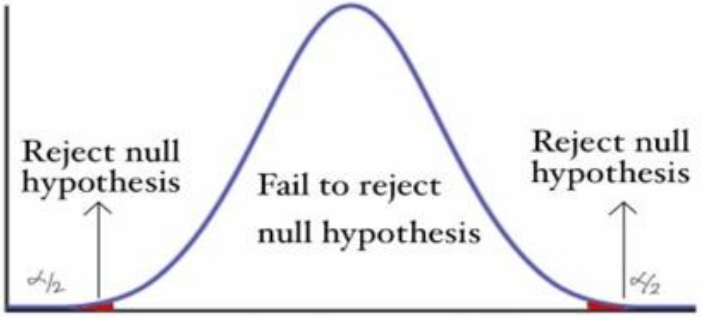
Null Hypothesis(H0) : $\mu \leq value$
Alternative Hypothesis(H_A) : $\mu > value$



Right Tailed Test

One Tailed Test

Null Hypothesis(H0) : $\mu = value$
Alternative Hypothesis(H_A) : $\mu \neq value$



Two Tailed Test

Decision Rule for Test of Hypothesis

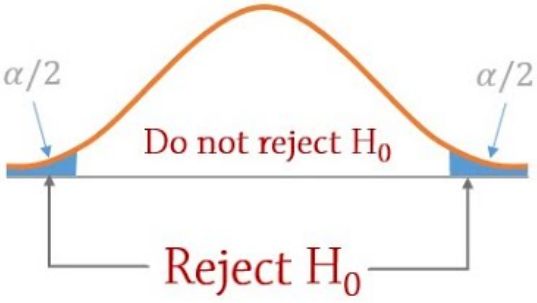
Hypothesis Testing

One-tailed

Two-tailed

$H_0: \mu = 23$

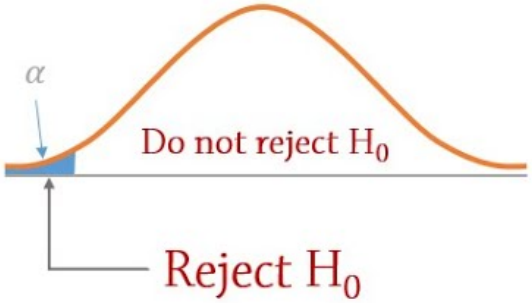
$H_1: \mu \neq 23$



Left-tailed

$H_0: \mu \geq 23$

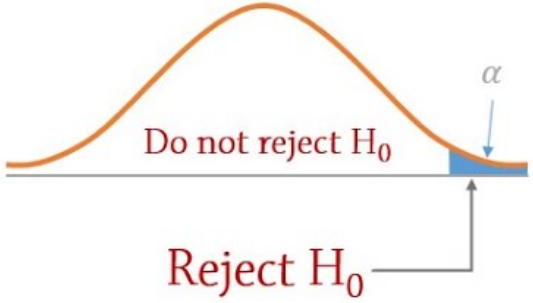
$H_1: \mu < 23$



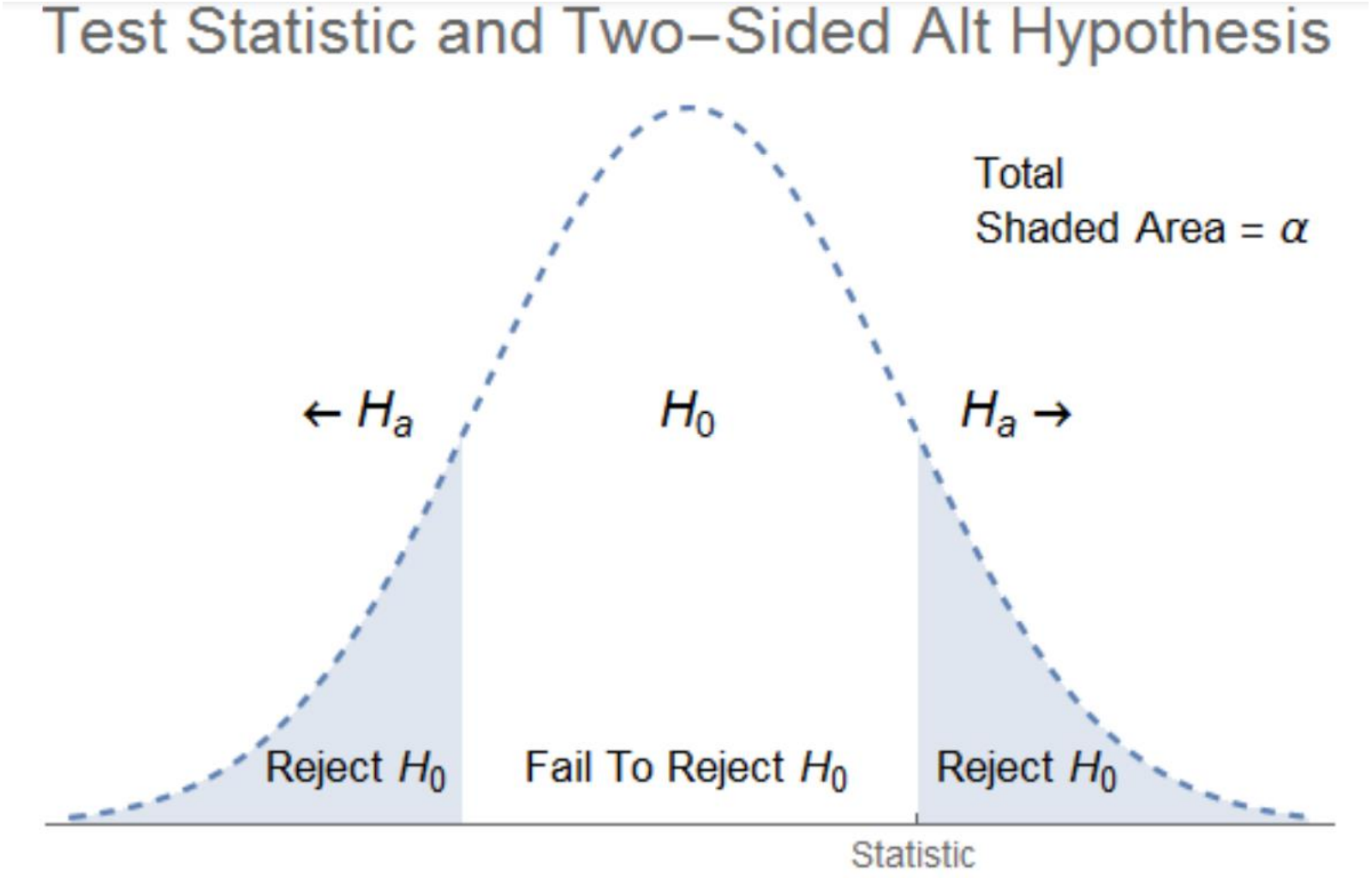
Right-tailed

$H_0: \mu \leq 23$

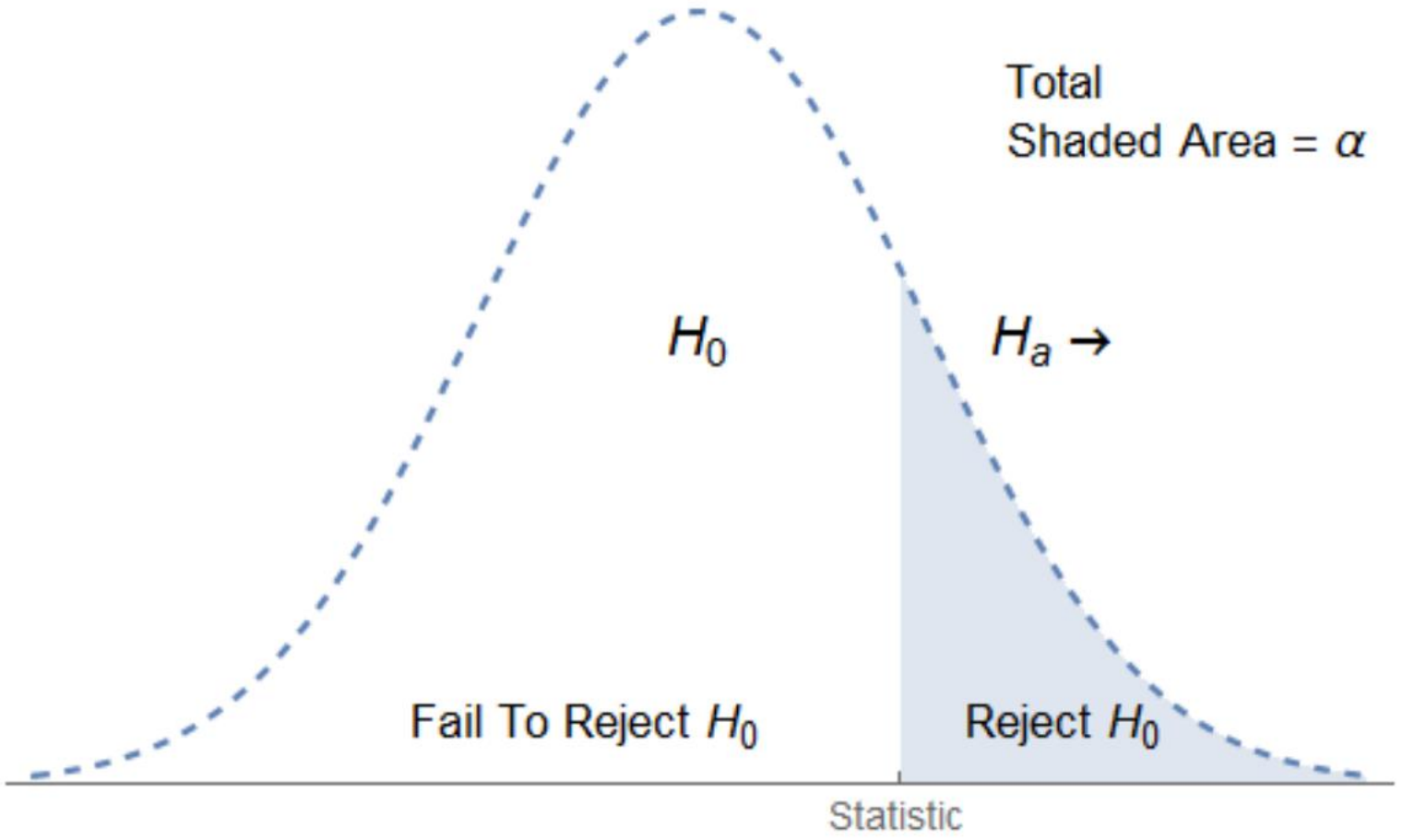
$H_1: \mu > 23$



Test Decision Based on Critical Value

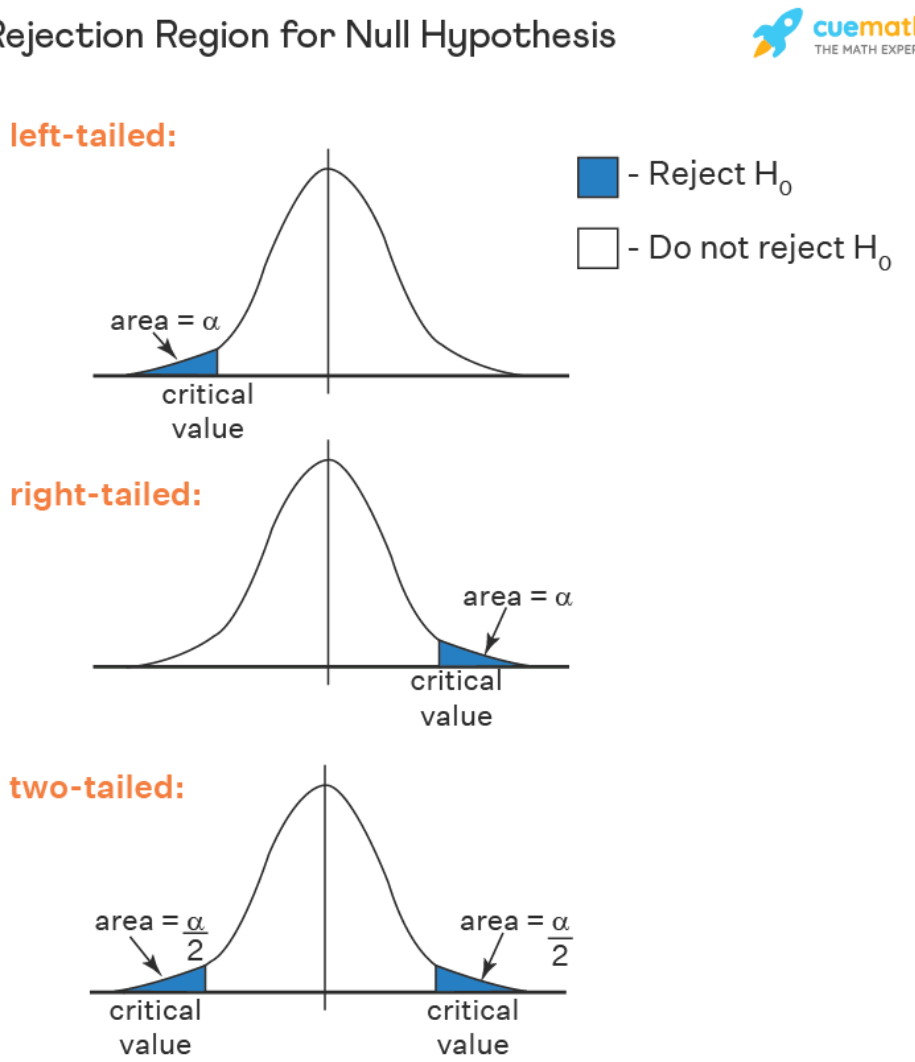
- 

Test Statistic and One–Sided Alt Hypothesis

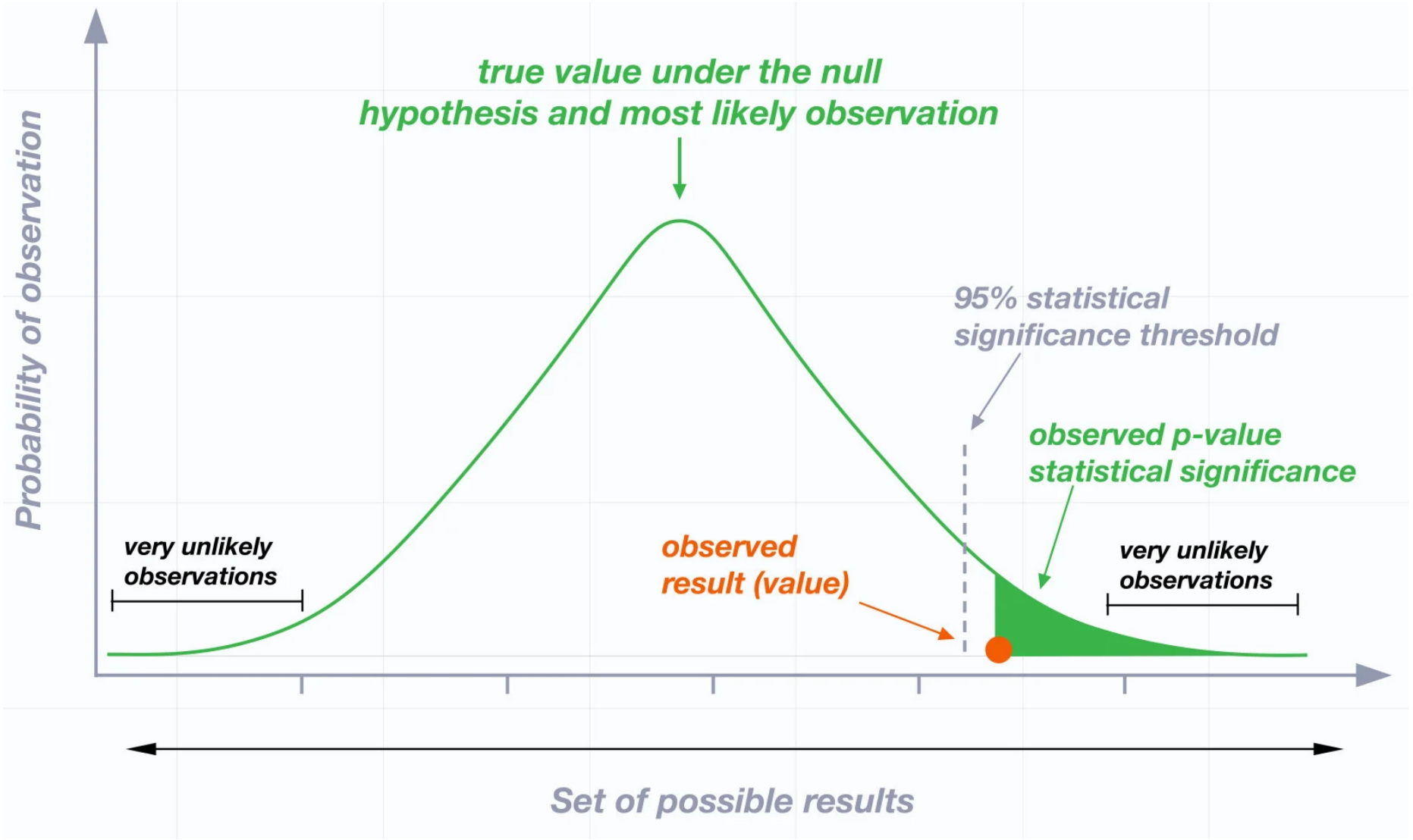
- 

Critical Value and Rejection Region

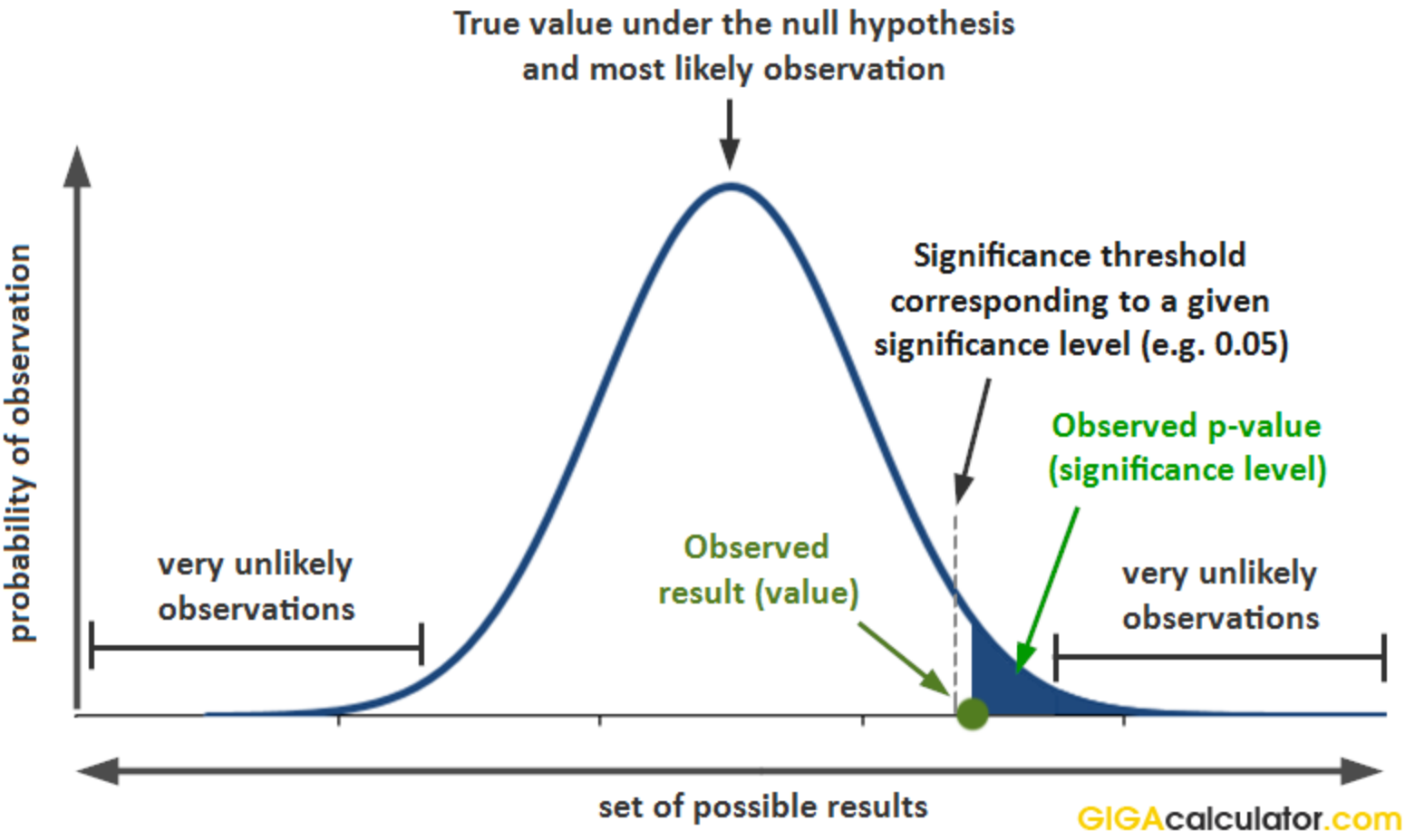
- .



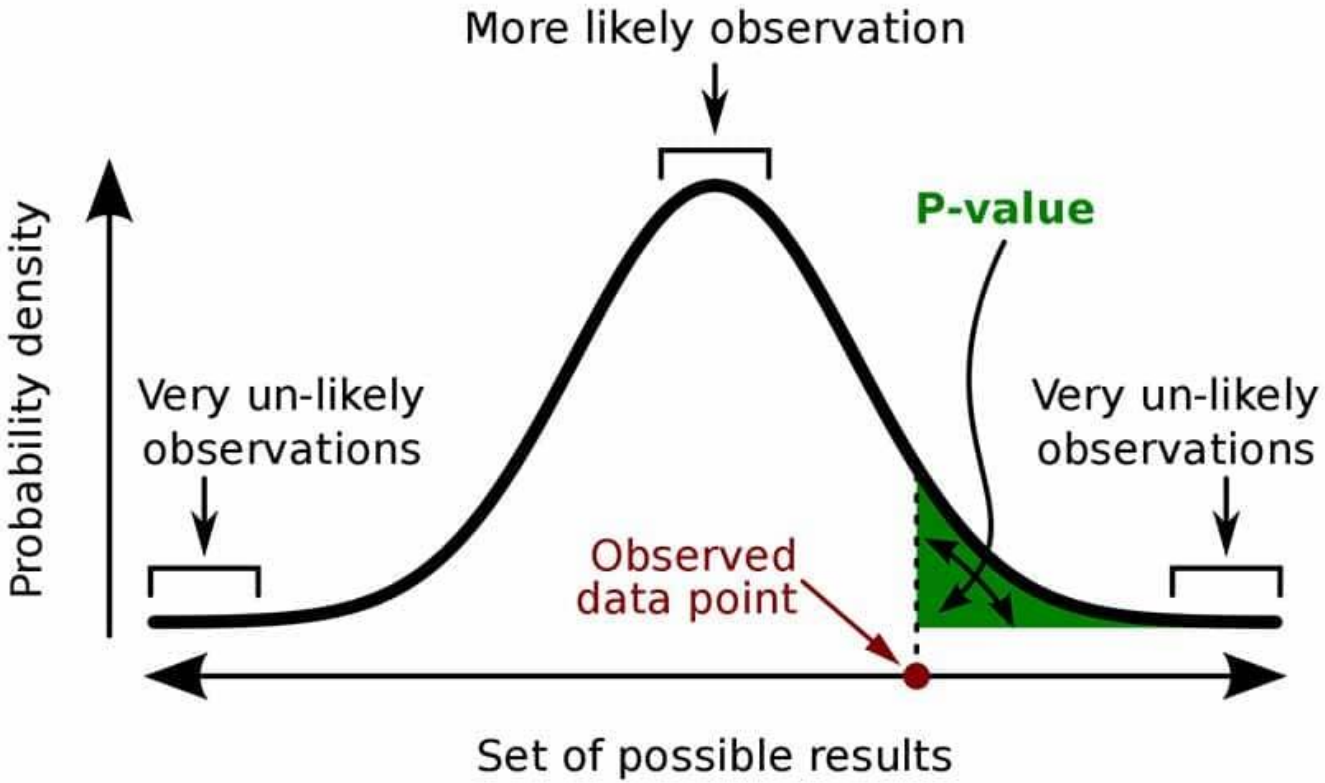
Decision Based on P-value



P-values and statistical significance explained



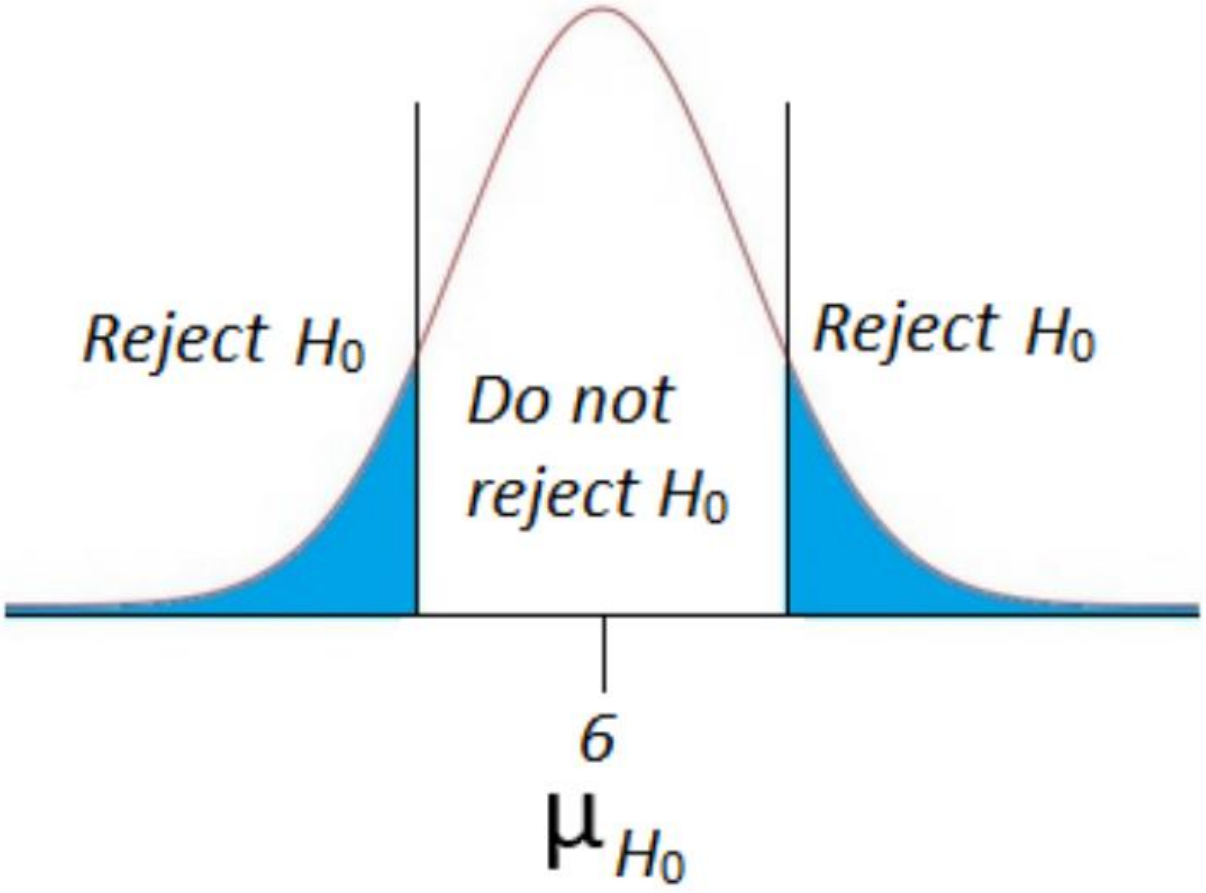
P-value



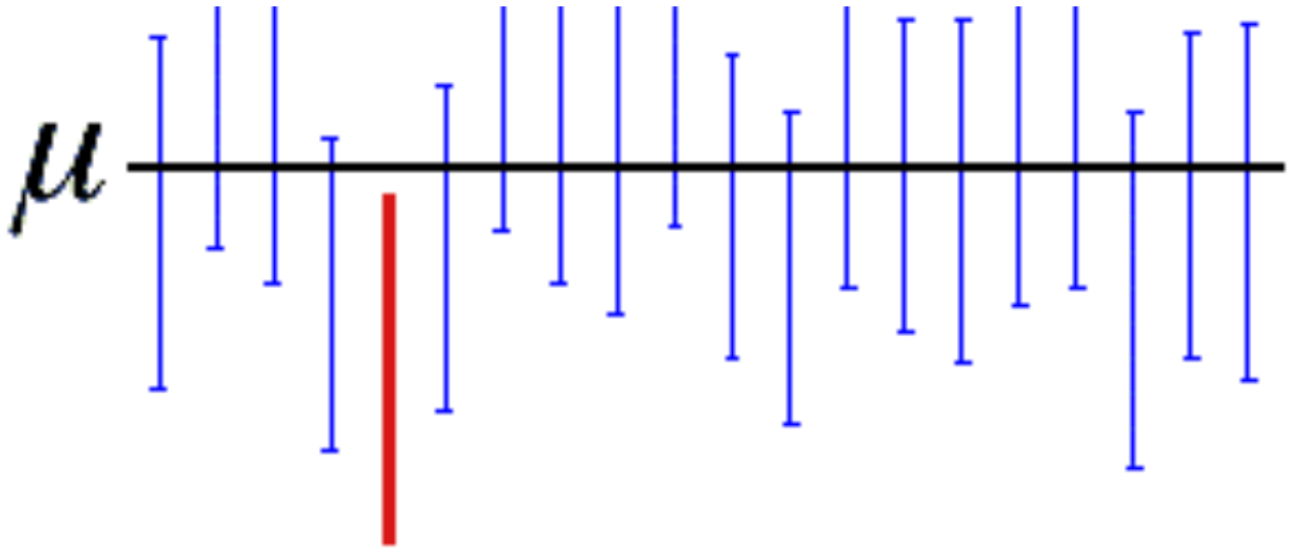
A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Test of Hypothesis based on CI

- .



Test based on CI

- 

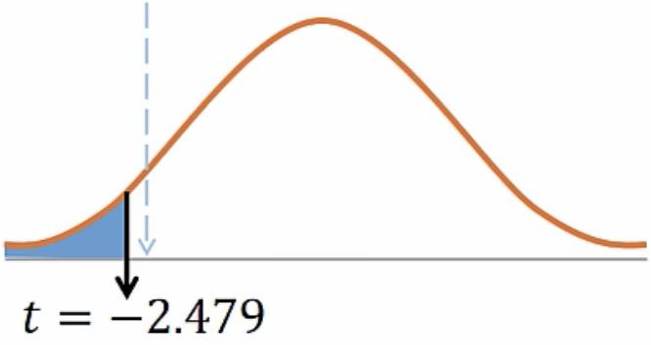
Example 01

Example: A random sample of 27 observations from a large population has a mean of 22 and a standard deviation of 4.8. Can we conclude at $\alpha = 0.01$ that the population mean is significantly below 24?

$$n = 27 \quad \bar{x} = 22 \quad s = 4.8 \quad \alpha = 0.01$$

$$H_0: \mu \geq 24$$
$$H_1: \mu < 24$$

$$\alpha = 0.01 \qquad df = 26$$



Reject H_0 if $t < -2.479$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{22 - 24}{4.6/\sqrt{27}} = -2.165$$

Since t = -2.165 is not less than -2.479,

Fail to Reject H_0

There is not enough evidence that the population mean is less than 24.

Example 01

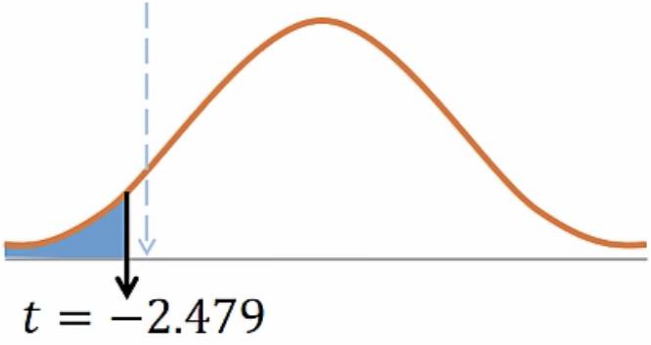
Example: A random sample of 27 observations from a large population has a mean of 22 and a standard deviation of 4.8. Can we conclude at $\alpha = 0.01$ that the population mean is significantly below 24?

$$n = 27 \quad \bar{x} = 22 \quad s = 4.8 \quad \alpha = 0.01$$

$$H_0: \mu \geq 24$$
$$H_1: \mu < 24$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{22 - 24}{4.6/\sqrt{27}} = -2.165$$

$\alpha = 0.01$ $\quad df = 26$



Reject H_0 if $t < -2.479$

Since $t = -2.165$ is not less than -2.479,

Fail to Reject H_0

There is not enough evidence that the population mean is less than 24.

Example 02

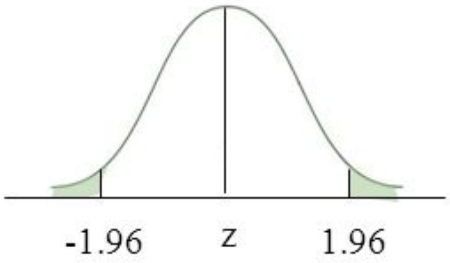
$\mu = 500$ g $\quad \sigma = 100$ g

- Step 1: H_o: $\mu_{\text{GRE after Kaplan Course}} = 500$ (There is no effect of the Kaplan training course on average GRE scores)

  H_1: $\mu_{\text{GRE after Kaplan Course}} \neq 500$ (There is an effect…)

  $\alpha = 0.05$

Step 2: Set criteria



Critical Region

$z > 1.96$

or

$z < -1.96$

Step 3: $n = 100$ $\quad \bar{X} = 525$ $\quad \sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{100}{\sqrt{100}} = \dfrac{100}{10} = 10$

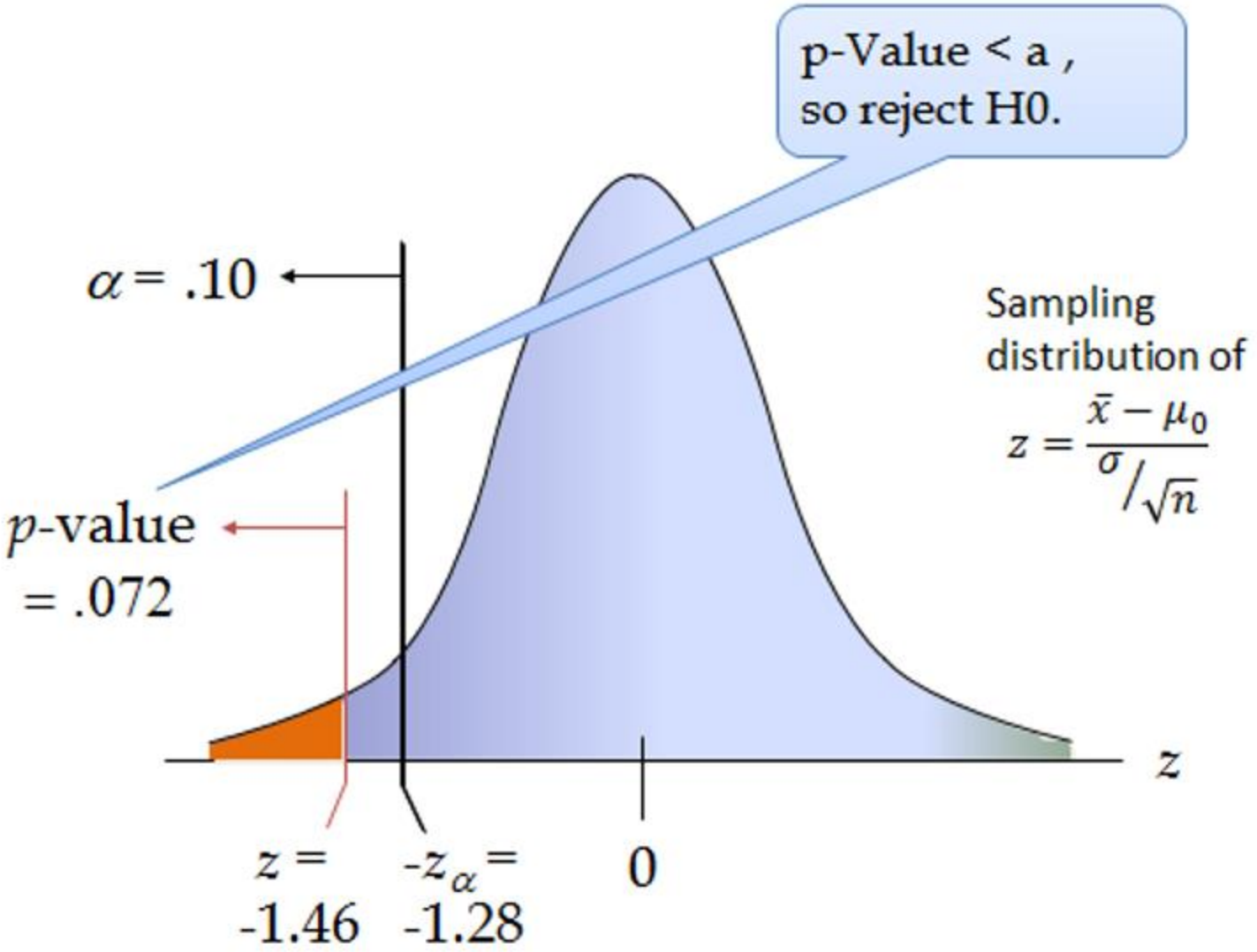
$$Z_{obt} = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{525 - 500}{10} = \frac{25}{10} = 2.5$$

Step 4: Reject H_o because Z_{obt} of 2.5 is in the critical region.

Step 5: Conclusion. The Kaplan training course significantly increased GRE scores on average, $z = 2.5$, $p < .05$.

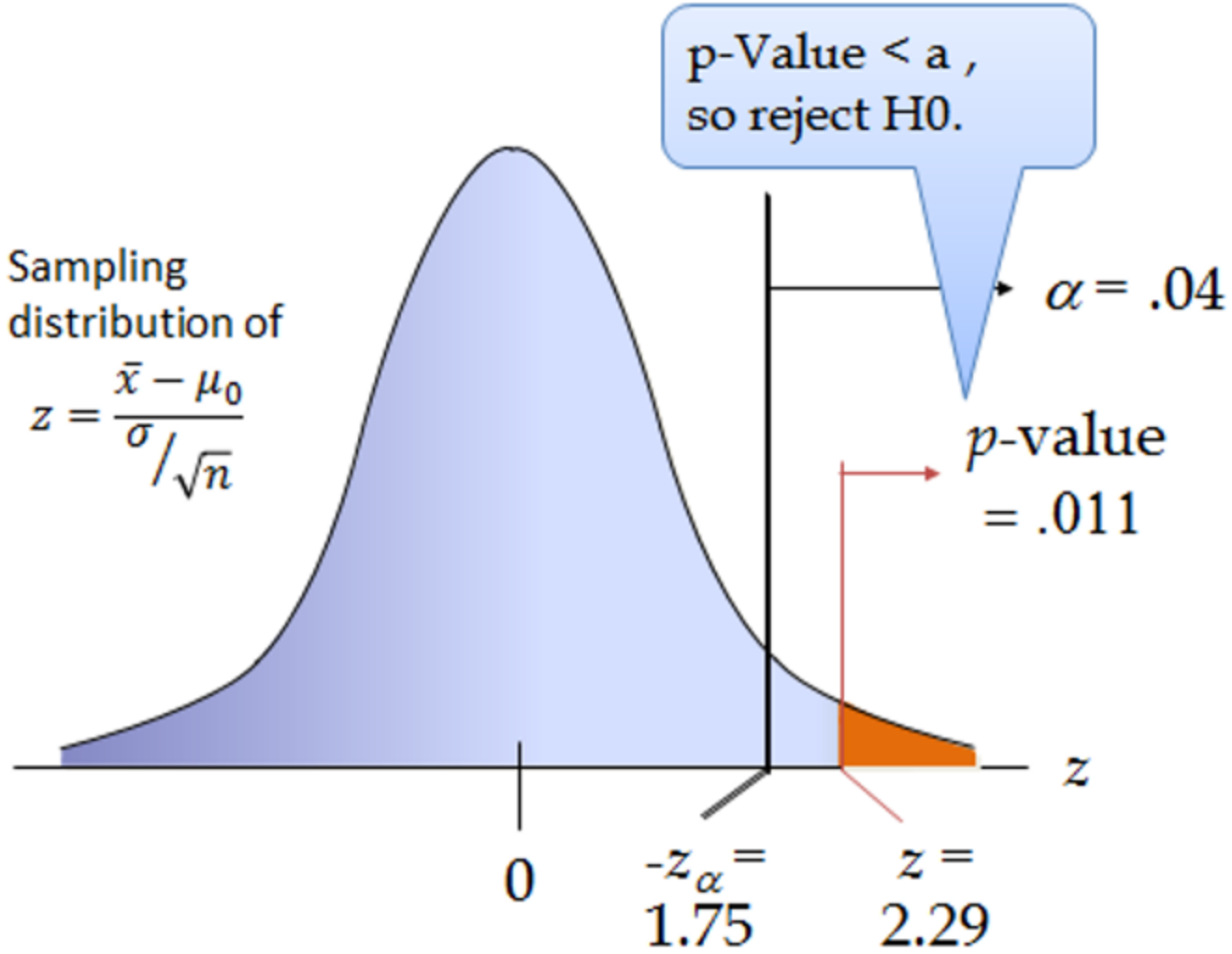
Example 03

- .

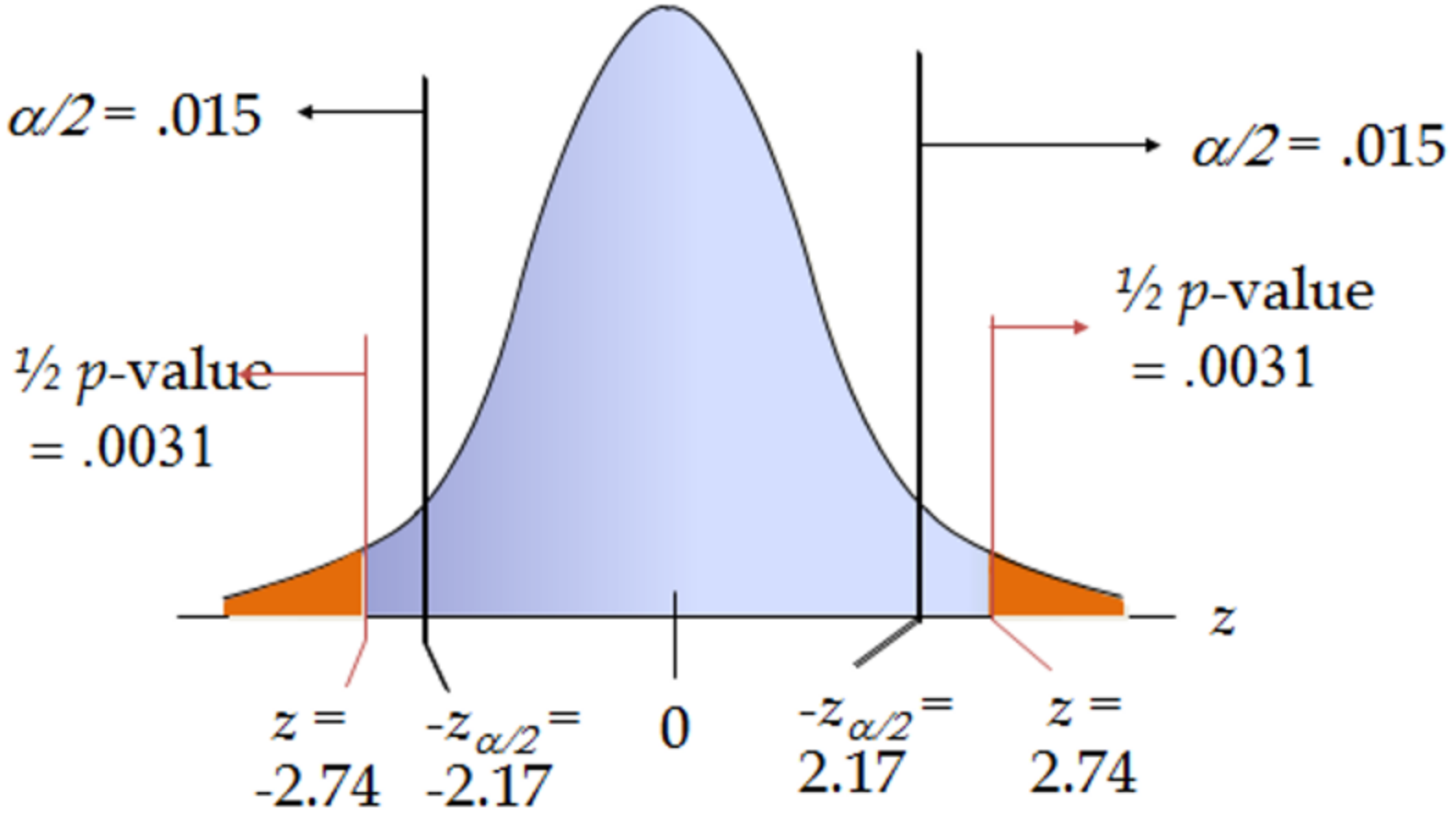

p-Value < a ,
so reject H0.
$\alpha = .10$
Sampling
distribution of
$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
p-value
= .072
z
$z =$
-1.46
$-z_\alpha =$
-1.28
0

Example 04

- .


p-Value < a ,
so reject H0.
Sampling
distribution of
$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
$\alpha = .04$
p-value
= .011
z
0
$-z_\alpha$ =
1.75
z =
2.29

Example?


α/2 = .015
α/2 = .015
½ p-value
= .0031
½ p-value
= .0031
z
z =
-2.74
-zα/2 =
-2.17
0
-zα/2 =
2.17
z =
2.74

Example 05

- **Example:** The Food and Nutrition Board of the National Academy of Sciences states that the recommended daily allowance (RDA) of iron for adult females under the age of 51 is 18 milligrams (mg). A sample of iron intake in was obtained during a 24-hour period from 45 randomly selected adult females under the age of 51. It revealed that the sample mean () was 14.68 mg. At the 1 percent significance level, does the data suggest that adult females under the age of 51 are, on average, getting less than the RDA of 18 mg of iron? Assume that the population standard deviation is 4.2 mg.

Solution: Find the values of n, $\bar{x}$ and σ:

$$n = 45,\ \bar{x} = 14.68 \text{ and } \sigma = 4.2$$

Step 1: State the null hypothesis (H_0) and the alternative hypothesis (H_a). The null and alternative hypotheses are:

$H_0: \mu = 18\ mg$
$H_a: \mu \leq 18\ mg$ $\Rightarrow$ A left/~~two/right~~-tailed test

Step 2: Decide on the siginificance level, (α) :

Degree of significance, $\alpha\%$ = 1%

Level of significance, α = 0.01

Step 3: Compute the value of the test statistics (Z) :

$$\mu_{\bar{x}} = \mu = 18$$

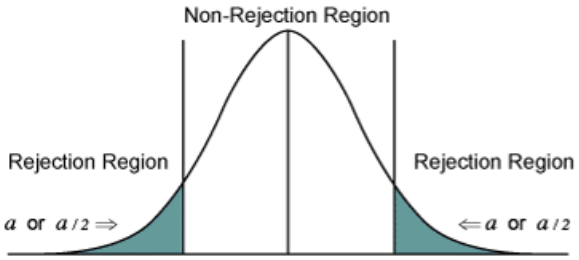
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4.2}{\sqrt{45}} \Rightarrow$$

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma/\sqrt{n}} = \frac{(14.68 - 18)}{(4.2/\sqrt{45})} = -5.30$$

Step 4: Type of the test: **Check one only**

1. Left-tailed test: [x] Critical value* $\Rightarrow$ $Z_a = -2.33$
2. Two-tailed test: [ ] Critical values* $\Rightarrow$ $\pm Z_{a/2} = \pm$ ___
3. Right-tailed test: [ ] Critical value* $\Rightarrow$ $Z_a =$ ___

*Use normal table to find the critical value(s), Z or $\pm Z_{a/2}$:



$\uparrow Z_a$ for left-tailed test $\uparrow Z_a$ for right-tailed test

$-Z_{a/2} \leftarrow$ for two-tailed test $\rightarrow Z_{a/2}$

Step 5: Compare the values of test statistics, Z, and critical value(s), Z_a or $\pm Z_{a/2}$:

		Check one only	Check one only
1.	$Z \geq Z_a \Rightarrow$	[ ]	
2.	$-Z_{a/2} \leq Z \leq Z_{a/2}$	[ ] Do not reject H_0	[ ] Otherwise reject H_0 [x]
3.	$Z \leq Z_a \Rightarrow$	[x]	

Step 6: Interpret the results of the hypothesis test:

At the 1 percent significance level, the data provides/~~does not provide~~ sufficient evidence to conclude that adult females under the age of 51 are, on average, getting less than the RDA of 18 mg of iron.

- .

Step 4: Determine the *p*-value:

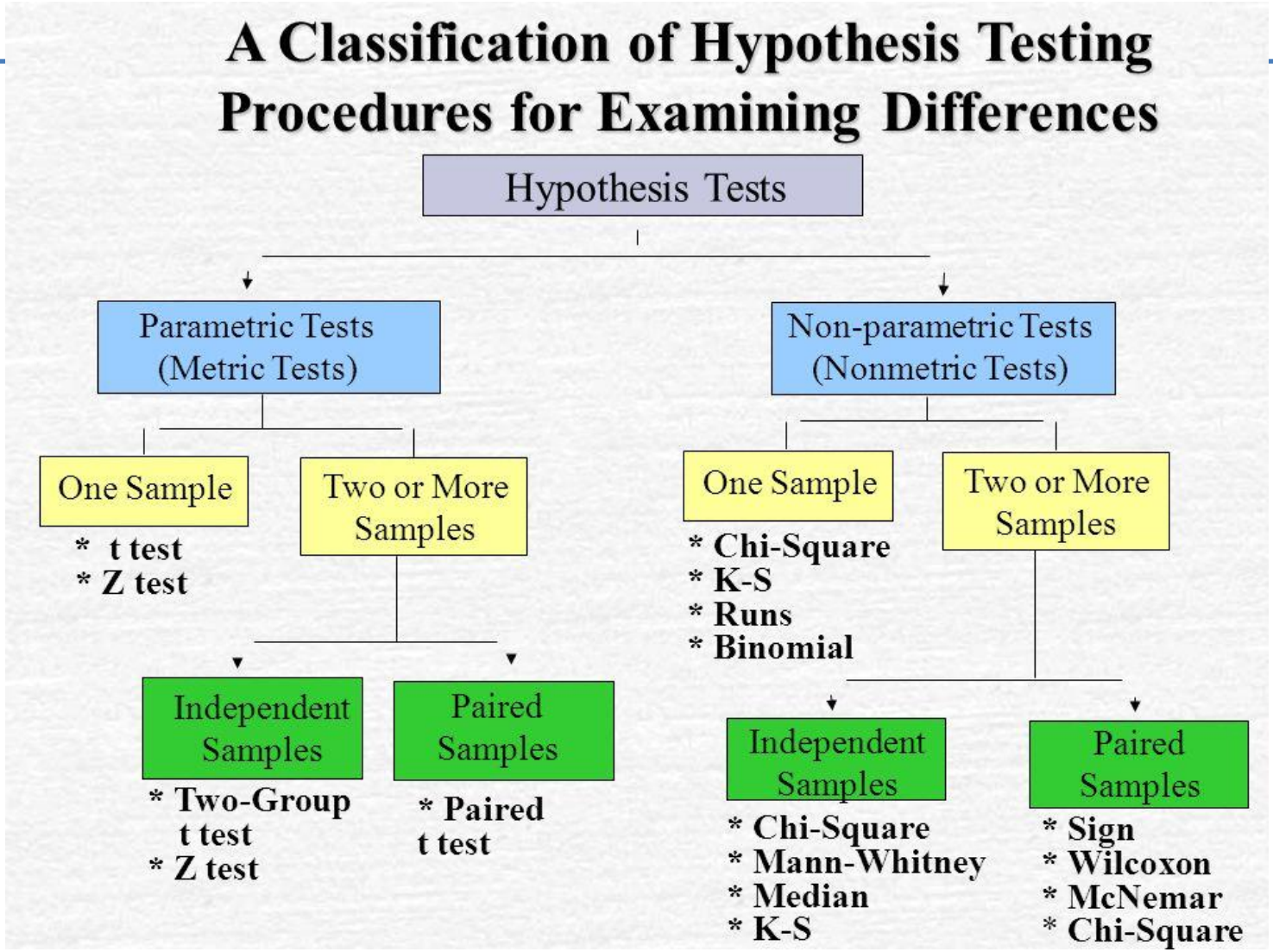
p-value $= P(z \leq Z) = P(z < -5.3) = 0.0000$ (small)

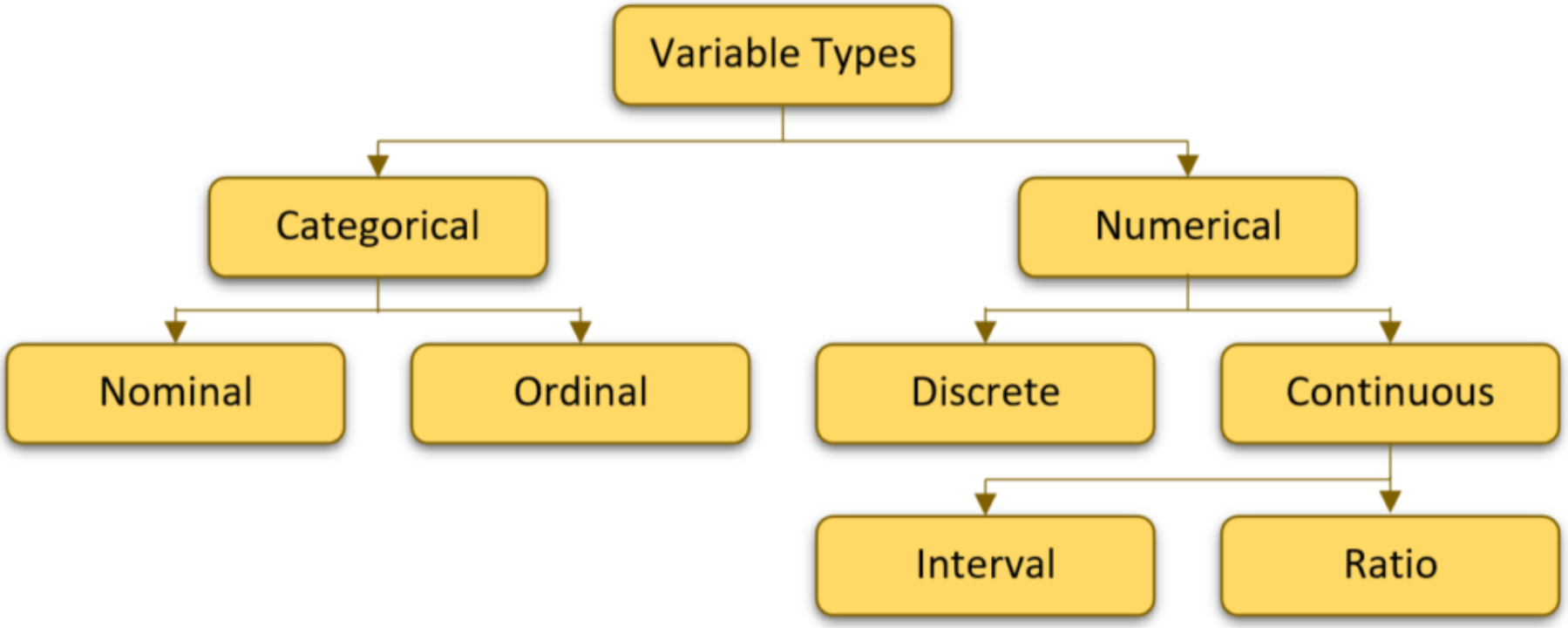
Step 5: If: Check one only Check one only

$P \leq \alpha$ [x] $\Rightarrow$ Reject H_0 [x]

$P < \alpha$ [ ] $\Rightarrow$ Do not reject H_0 [ ]

since, 0.0000 < 0.01


A Classification of Hypothesis Testing Procedures for Examining Differences
Hypothesis Tests
Parametric Tests (Metric Tests)
Non-parametric Tests (Nonmetric Tests)
One Sample
* t test
* Z test
Two or More Samples
Independent Samples
* Two-Group t test
* Z test
Paired Samples
* Paired t test
One Sample
* Chi-Square
* K-S
* Runs
* Binomial
Two or More Samples
Independent Samples
* Chi-Square
* Mann-Whitney
* Median
* K-S
Paired Samples
* Sign
* Wilcoxon
* McNemar
* Chi-Square


Variable Types
Categorical
Numerical
Nominal
Ordinal
Discrete
Continuous
Interval
Ratio

- .

Independent Variable	Dependent Variable
Feature	Target
Exogenous	Endogenous
Explanatory	Explained / Response
Predictor	Predicted
Regressor	Regressand
Manipulated	Measured / Observed
Exposure	Outcome
Input	Output

Difference between Descriptive and Inferential Statistics

Basis of Comparison	Descriptive Statistics	Inferential Statistics
Definition	Descriptive Statistics describes or summarizes the data.	Inferential statistics makes inferences or conclusions about the population based on the sample.
Purpose	The purpose of descriptive statistics is to describe a situation or an event.	On the other hand, the purpose of inferential statistics is to explain the likelihood of occurrence of an event.
Function	Descriptive Statistics helps in organizing, analyzing and presenting data in an effective and useful way.	Inferential statistics deals with comparing the data, testing i.e., making hypotheses & estimates, and predicting future results.
Conclusions	Descriptive statistics quantifies the known data i.e., it summarizes the characteristics of that data which is already known.	Whereas the Inferential Statistics tries to make inferences or learn about the population i.e., it explains beyond the data that is available.
Variables taken in consideration	Descriptive statistics measures only the provided data and does not consider any other variables. Hence, it is an objective methodology.	Inferential Statistical considers variables, sampling errors that may lead to conducting additional tests. Therefore, it is a subjective process.
Result or Output Form	Descriptive Statistics gives the result or output in the form of tables, charts, or graphs.	Inferential Statistics generates probabilities as its result.
Parameters or Sample Statistics	Descriptive statistics are applied on the entire population. The properties of the population are known as the parameters.	Inferential Statistics is applied on a subset of the population i.e., on sample. The properties of the sample are known as the sampling statistic.
Types of Statistics	The types of descriptive statistics are: • measures of central tendency • measures of variability • measures of distribution • five-point summary, and • measure of association between two variables.	The types of inferential statistics are: • hypothesis testing • confidence intervals, and • regression analysis
Measures	The measure or techniques used in the descriptive statistics are mean, median, mode, variance, standard deviation, range, IQR, frequency distribution.	The measures or techniques used for inferential data analysis are t-test, Z-test, ANOVA, Chi-Square, Linear Regression.
Examples	The example of descriptive statistics is: • 46% of employees are females • In class 8th, for 10 students, the average mark in French is 70 • Average email response rate is 7% • Range of the group is 10 • Variation or spread in the class is 20%	The example of inferential research is given. We know that in class 8th, for 10 students, the average mark in French is 70. Using this information, we can infer about the entire population of 50 students who are enrolled for French.

THANK YOU