

Introduction to Inference

Jose Toledo Luna

2026-10-20

INTRODUCTION

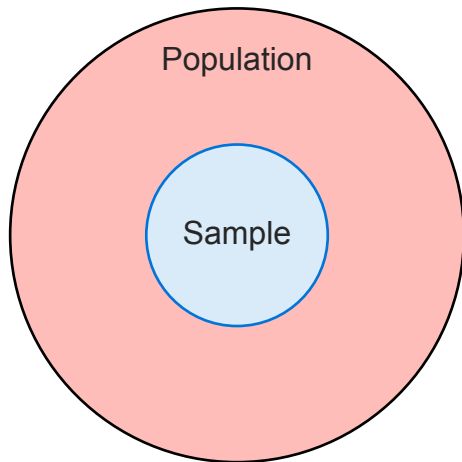
Inference uses the concepts of probability and data analysis tools to

- draw inferences about populations via samples taken from a population
- estimate the uncertainty (or level of confidence) associated with the inference

Principles of inference will be illustrated in the setting of

- estimating the average (mean) for a characteristic of a well-defined population
- drawing conclusions about that characteristic

TERMINOLOGY



- **Population:** The entire group of individuals or observations of interest
- **Parameter:** A numerical characteristic of the population
- **Sample:** A subset of the population
- **Statistic:** A numerical characteristic calculated from the sample

NOTATIONS

After identifying the target population, we are interesting in a numerical characteristic of this population (**parameter**). Here, we are interested in averages.

- μ : unkown population mean
- σ : unknown population standard deviation
- n : the number of observations in a sample drawn from the population
- $\bar{x}$: sample mean from a sample taken from the population
- s : sample standard deviation from the sample used to calculate $\bar{x}$

YOUTH RISK BEHAVIOR SURVEILLANCE SYSTEM (YRBSS)

The **Youth Risk Behavior Surveillance System** is a national survey conducted by the Centers for Disease Control and Prevention (CDC) to monitor health-related behaviors among U.S. high school students.

Roughly 2.6 million high school students participated between 1991 and 2013 in more than 1,100 separate surveys

The [openintro::yrbss](#) dataset contains:

- 13,583 observations and 13 variables
- Demographic characteristics
- Physical characteristics
- Health-related behaviors

Variable	Description
age	Age, in years.
gender	Gender.
grade	School grade.
hispanic	Hispanic or not.
race	Race / ethnicity.
height	Height, in meters (3.28 feet per meter).
weight	Weight, in kilograms (2.2 pounds per kilogram).
hours_tv_per...	How many hours of TV do you typically watch on a school night?
strength_training...	How many days did you do strength training (e.g. lift weights) in the last 7 days?
school_night...	How many hours of sleep do you typically get on a school night?

There are other variables in this dataset. To view their description, access the dataset documentation with `?openintro::yrbss`

ESTIMATING PARAMETERS

The CDC used the responses of the 13,000 students to estimate the health behaviors of the **target population**

- 21.2 million high school aged students in the U.S. in 2013

The mean weight among the 21.2 million students is an example of a **population parameter** and denoted with μ

It is infeasible to obtain the mean weight among all 21.2 million students, so instead we **estimate** it using a sample, such as the mean weight in the 13000 students in YRBSS

The sample mean, denoted as $\bar{x}$ is a **point estimate** of the population mean μ

SAMPLING FROM A POPULATION

The exact values of population parameters are unknown. Hence, the need for statistical inference

Suppose our target population consists of the 13583 individuals in [yrbss](#) and we are interested in estimating the mean weight (μ) of the students.

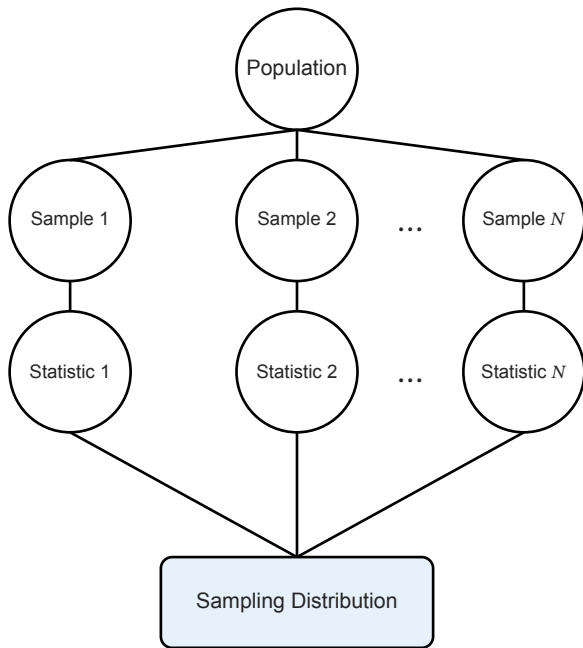
- The population mean (μ) is known for this scenario, $\mu = 67.91$

We can observe the effect of sampling variability by taking many samples from this population and calculating the sample mean ($\bar{x}$) for each sample

If we take sufficiently many samples from this population and calculate their sample means, we can construct a **sampling distribution** of $\bar{X}$

SAMPLING DISTRIBUTIONS

1. Begin with a target population, and identify the **parameter** of interest. For example, the population mean (μ)
2. Randomly draw multiple samples of size (n) from the population
3. For each sample drawn, compute the **sample statistic**
4. The resulting distribution represents the **sampling distribution** of the sample statistic
5. The sampling distribution provides information about the variability of the sample statistic. Key characteristics such as the mean, variance, and shape of the distribution can be described.



SAMPLING DISTRIBUTION OF $\bar{X}$

Given the population of all 13583 high school students in [yrbss](#), we will take many random samples of size $n = 20$

	student	weight
1	491	57.15
2	3721	61.24
3	11594	54.43
4	12656	61.24
17	930	57.61
18	9752	61.24
19	5645	63.05
20	7519	61.69

Sample 1 Mean: 63.50526

	student	weight
1	7560	62.14
2	1646	68.04
3	6919	49.90
4	10214	74.84
17	33	66.23
18	1671	70.76
19	946	55.79
20	10564	53.52

Sample 2 Mean: 66.837

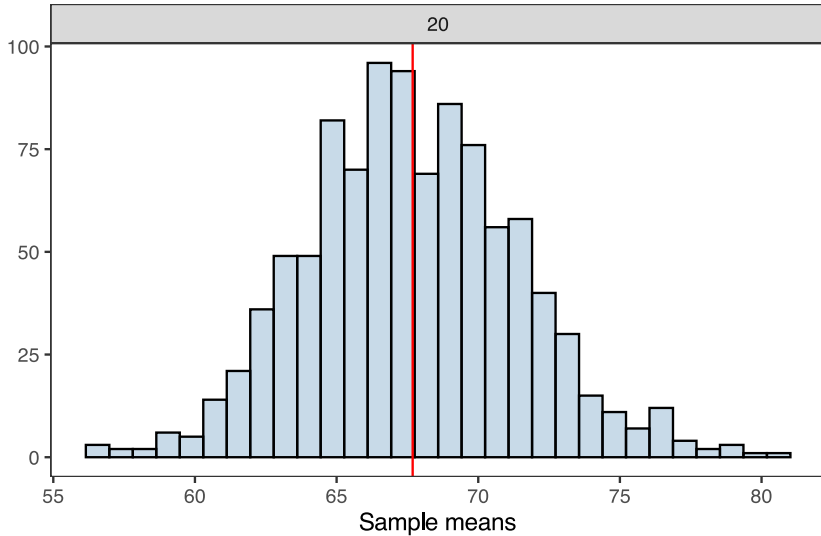
Continue the earlier process of obtaining random samples from the population and obtaining their sample mean.

We will take 1000 samples of size $n = 20$ and calculate the sample mean $\bar{x}$ for each sample. Below is the distribution of all those sample means

Recall: The population mean for the weights of the 13583 students was $\mu = 67.91$

Sampling distribution

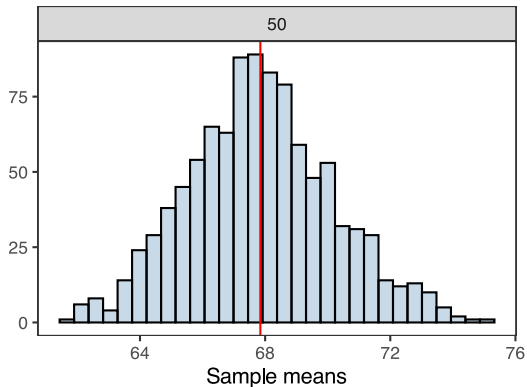
Mean = 67.685, Standard error = 3.743



Still taking 1,000 samples but with sizes $n = 50, 100$ for each sample

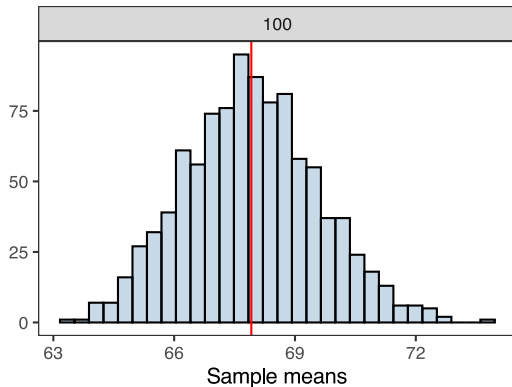
Sampling distribution

Mean = 67.847, Standard error = 2.325



Sampling distribution

Mean = 67.916, Standard error = 1.662



CENTRAL LIMIT THEOREM

The **Central Limit Theorem** states that if we repeatedly take sufficiently large independent random samples of size n from a population with mean μ and standard deviation σ , the sampling distribution of the sample mean $\bar{X}$ will be approximately normal, with mean μ and standard error $\sigma/\sqrt{n}$.

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

Note, the standard deviation of the sampling distribution of the sample mean (*or any other statistic*) is referred to as the *standard error*.

In practice, we do not know σ so we estimate it using (s) the sample standard deviation.

- **Estimation:** Allows us to estimate population parameters. The mean of the sampling distribution will approximately be the population parameter.
- **Variability:** The standard deviation of the sampling distribution (known as the **standard error**) measures how much the statistic varies from sample to sample
- **Inference:** Forms the basis for **hypothesis testing** and **confidence intervals**. We can quantify uncertainty and draw conclusions about the population parameters

CONFIDENCE INTERVALS

A **confidence interval** provides an estimate for a population parameter along with a **margin of error** that gives a plausible range of values for the population parameter

A confidence interval has the following form

$$\begin{aligned} CI &= \text{point estimate} \pm \text{margin of error} \\ &= (\text{point estimate} - \text{margin of error}, \text{point estimate} + \text{margin of error}) \\ &= (\text{Lower Bound}, \text{Upper Bound}) \end{aligned}$$

where the **point estimate** is used to estimate the population parameter of interest, and the **margin of error** is calculated based on the variability of the statistic's sampling distribution

CONFIDENCE LEVELS

Interpretation: The resulting confidence interval indicates the range within which we are confident the true population parameter lies.

For example, with a 95% confidence level, we expect that 95 out of 100 such intervals would contain the population parameter (on average).

“We are 95% confident the true population parameter lies between (lower bound) and (upper bound)” in the context of the application

CONFIDENCE INTERVAL: POPULATION MEAN

The target parameter is the population mean (μ), then the **point estimate** will be the sample mean $\bar{x}$

The **margin of error** is computed as

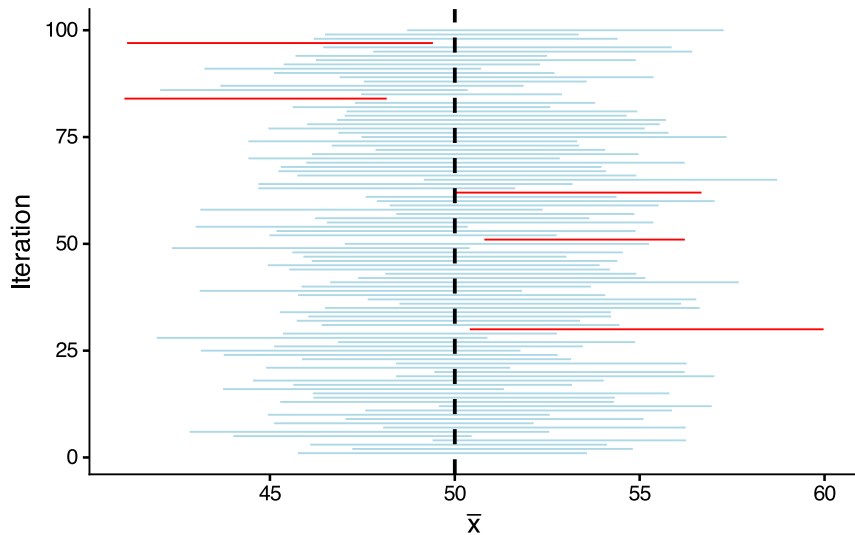
$$\text{Margin of Error} = \text{Critical Value} \times \text{Standard Error}$$

where the standard error is the standard deviation of the sampling distribution of the sample mean $\frac{\sigma}{\sqrt{n}}$

The **critical value** will be discussed later on

95% coverage

Empirical coverage: 95% (95/100)



t -DISTRIBUTION

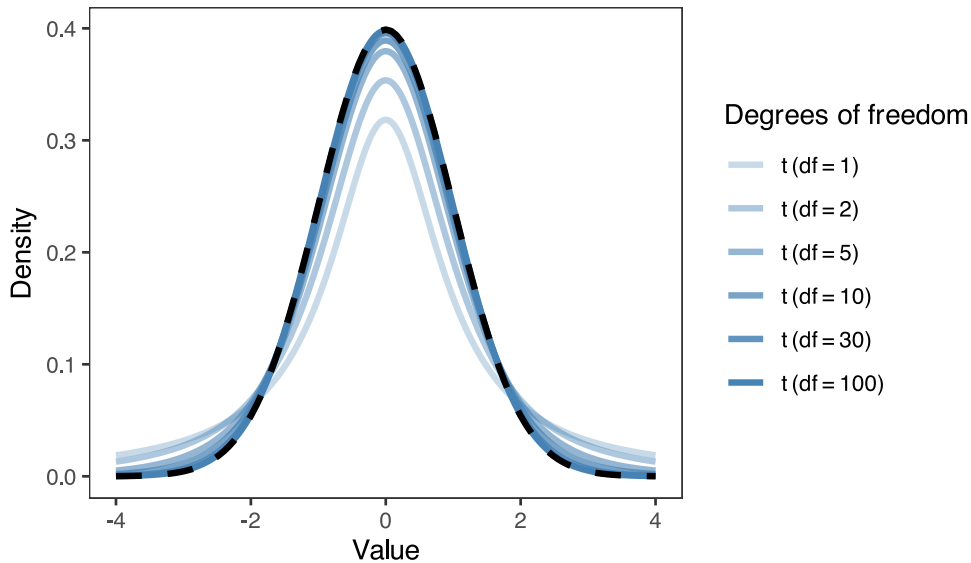
The t -distribution, also known as **Student's t -distribution**, is similar to the standard normal distribution but has heavier tails.

The heavier tails accounts for the additional uncertainty introduced by estimating the population standard deviation from the sample

The t -distribution is always centered at zero and has a single parameter, the degrees of freedom

The degrees of freedom (df) describes the precise form of the bell-shaped t -distribution. This is also a symmetrical distribution, where $t_{df}(a) = t_{df}(-a)$

Various t-distributions



The larger the degrees of freedom, the more closely the t -distribution resembles the standard normal distribution.

When modeling $\bar{X}$ using the t -distribution becomes essential whenever the following conditions

- If the population is not normally distributed, but the sample size is sufficiently large (usually $n \geq 30$)
- When the population standard deviation (σ) is unknown, it is estimated from the sample using the sample standard deviation (s).

A 95% CONFIDENCE INTERVAL

A 95% confidence interval for a population mean μ based on a single sample with mean $\bar{x}$ and sample standard deviation s is

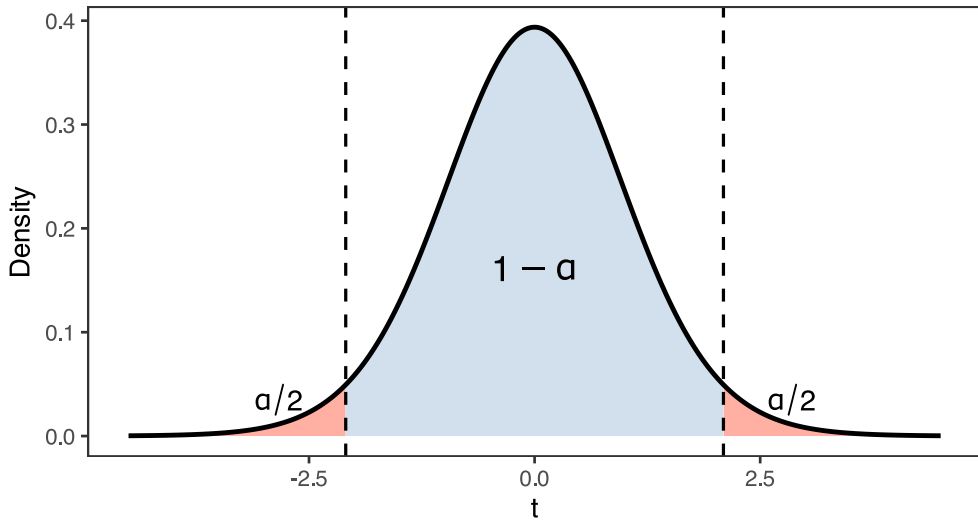
$$CI = \text{point estimate} \pm \text{margin of error}$$

$$\begin{aligned} &= \bar{x} \pm t^* \left(\frac{s}{\sqrt{n}} \right) \\ &= \left(\bar{x} - t^* \left(\frac{s}{\sqrt{n}} \right), \bar{x} + t^* \left(\frac{s}{\sqrt{n}} \right) \right) \end{aligned}$$

where t^* is the critical value on a t -distribution with $df = n - 1$, and a cumulative probability of 0.975.

t-distribution critical values

with $df = (n-1)$



A $100(1 - \alpha)$ CONFIDENCE INTERVAL

In general, a $100(1 - \alpha)\%$ confidence interval for a population mean μ based on a single sample with mean $\bar{x}$ and sample standard deviation s is

$$CI = \text{point estimate} \pm \text{margin of error}$$

$$= \bar{x} \pm t_{1-\frac{\alpha}{2}, n-1} \left(\frac{s}{\sqrt{n}} \right)$$

$$= \left(\bar{x} - t_{1-\frac{\alpha}{2}, n-1} \left(\frac{s}{\sqrt{n}} \right), \quad \bar{x} + t_{1-\frac{\alpha}{2}, n-1} \left(\frac{s}{\sqrt{n}} \right) \right)$$

Where α is the significance level desired. For example, a 95% confidence interval is obtained using $\alpha = 0.05$

YRBSS: CONFIDENCE INTERVAL FOR μ

Revisit YRBSS dataset, we are interested in estimating the mean weight of all students which is $\mu=67.91$. Take a random sample of size $n = 20$

	student	weight
1	491	57.15
2	3721	61.24
3	11594	54.43
4	12656	61.24
17	930	57.61
18	9752	61.24
19	5645	63.05
20	7519	61.69

which gives us a sample mean $\bar{x} = 63.51$ (*point-estimate*) and sample standard deviation $s = 11.43$

Let's construct a 95% confidence interval for μ . We have the following:

$$\alpha = 0.05, n = 20, df = n - 1 = 19, \bar{x} = 63.51, s = 11.43$$

We need to compute the critical value $t_{1-\alpha/2, df} = t_{0.975, 19}$.

In R, `qt(p, df)` calculates t such that $p = P(T \leq t)$

```
alpha = 0.05  
n = 20  
qt( p = 1-alpha/2, df = n-1)
```

```
[1] 2.093024
```

Interpretation the quantile which gives a cumulative probability of 0.975 from a t -distribution with 19 degrees of freedom is **2.093**

Putting all of this together, we can construct the 95% confidence interval for μ .

$$\begin{aligned} CI &= \left(\bar{x} - t_{1-\frac{\alpha}{2}, n-1} \left(\frac{s}{\sqrt{n}} \right), \quad \bar{x} + t_{1-\frac{\alpha}{2}, n-1} \left(\frac{s}{\sqrt{n}} \right) \right) \\ &= \left(63.51 - 2.093 \left(\frac{11.43}{\sqrt{20}} \right), \quad 63.51 + 2.093 \left(\frac{11.43}{\sqrt{20}} \right) \right) \\ &= (58.16, 68.86) \end{aligned}$$

Interpretation: We are 95% confident the population mean weight of all high school students is between **58.16** and **68.86** kg.

Note: This interval does contain the population mean $\mu = 67.91$

NATIONAL HEALTH AND NUTRITION SURVEY (NHANE)

The National Health and Nutrition Examination Survey (NHANES) is another survey conducted by the CDC.

Purpose: to assess the health and nutritional status of adults and children in the United States

- The `NHANES` dataset in the `NHANES` package contains responses from 10,000 participants.
- The `nhanes.samp.adult` dataset in the `openintro` package contains responses from a random sample from participants who were age 21 or older.

We will treat `nhanes.samp.adult` as our sample and think of the adult participants in the `NHANES` dataset as the population

INTRODUCTION TO HYPOTHESIS TESTING

Continuing with **NHANE** dataset, we are interesting in answering the question:

Do U.S. adults tend to be overweight ?

Body mass index (BMI) is an approximate scale used to assess weight status that adjusts for height.

- When the weight is measure in kg and height in meters (*in YRBSS data set*),

$$BMI = \frac{weight}{height^2}$$

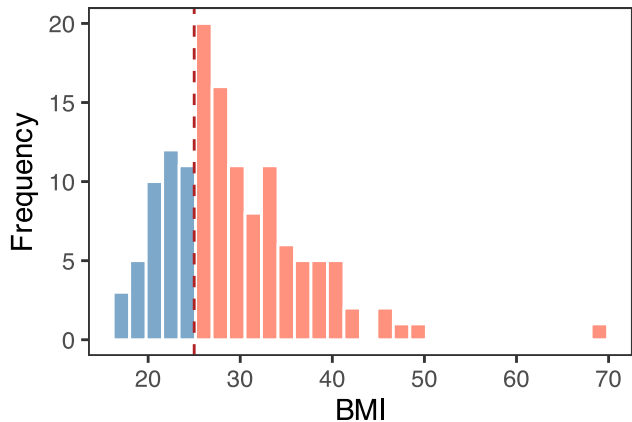
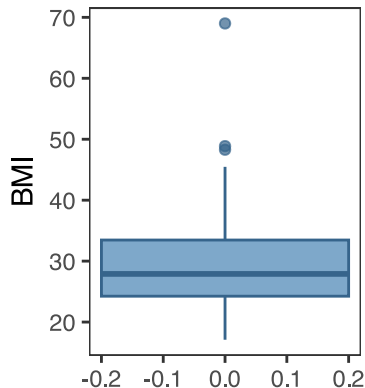
- If the weight is measured in lbs and height in inches

$$BMI = 703 \left(\frac{weight}{height^2} \right)$$

The World Health Organization (WHO) defines Body Mass Index (BMI) as

Category	BMI Range
Underweight	< 18.5
Normal	$18.5 - 24.99$
Overweight	$25 - 29.99$
Obese	≥ 30

Here is the distribution of the BMI for the sample of U.S. adults from
nhanes.samp.adult data set



HYPOTHESIS TESTING APPROACH

Observation come from either of two competing population distributions:

- **Null Distribution**:: A distribution that has been assumed to be “true” in the past
- **Alternative Distribution**: A new distribution induced by an intervention or a changing condition.

We assume that observations come from the null distribution, unless

- The value of an observed statistic (i.e., $\bar{x}$) is so **extreme** that it is considered unlikely to have occurred under the null distribution

Steps in hypothesis testing:

1. Formulate null and alternative hypotheses
2. Specify a significant level (α)
3. Calculate a test statistic
4. Calculate a p -value
5. State a conclusion in the context of the original problem

NULL AND ALTERNATIVE HYPOTHESIS

The **null hypothesis** (H_0) posits a distribution for the population that reflects no change

H_0 can be thought of as representing the status quo

The **alternative hypothesis** (H_a) claims a “real” difference between the distribution of the observed data and the null-hypothesized distribution

Lets go back to our research question:

Do U.S. adults tend to be overweight ?

There are several choices for H_0 and H_a

- $H_0 : \mu_{BMI} = 25 = \mu_0$ (*cutoff of the overweight range*)

The choice of **one-** or **two-sided alternative** is context-dependent

- $H_a : \mu_{BMI} > 25$ (*above overweight range*)
- $H_a : \mu_{BMI} < 25$ (*below overweight range*)
- $H_a : \mu_{BMI} \neq 25$ (*not overweight cutoff value*)

SPECIFYING A SIGNIFICANCE LEVEL α

The **significance level** (α) can be thought of as the value that quantifies how rare an event must be in order to represent **sufficient evidence** against H_0

Typically, α is chosen to be 0.05, 0.01, or some other small value

CALCULATE A TEST STATISTIC

The **test statistic** measures the discrepancy between the observed data and what we would be expected if the null hypothesis were true

- Specifically, how many standard deviations is the observed sample value from the population value under the null hypothesis ?

When testing hypotheses about a mean, the test statistic is

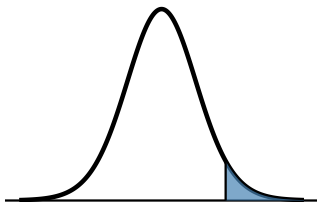
$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

where the test statistic T follows a t -distribution with $(n - 1)$ degrees of freedom

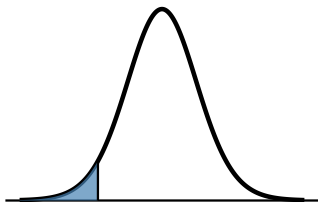
CALCULATE A p -VALUE

The p -value is the probability of observing a result **as or more extreme** than the observed sample value (**test statistic**), if the null hypothesis is true

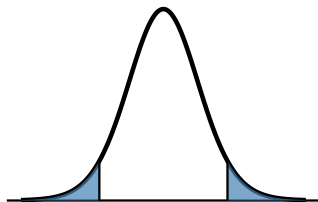
$$H_A: \mu > \mu_0$$



$$H_A: \mu < \mu_0$$



$$H_A: \mu \neq \mu_0$$



The smaller the p -value, the stronger the evidence against the null hypothesis

- If the p -value is as small or smaller than α , we **reject the null hypothesis**

The result is statistically significant at level α

- If the p -value is larger than α , we **fail to reject the null hypothesis**

The result is not statistically significant at level α . In other words, the evidence does not contradict the null hypothesis.

DRAW A CONCLUSION

! Important

Not rejecting H_0 is not the same as proving that H_0 is true. We simply do not have sufficient evidence that H_0 is not true!

State the conclusion in the context of the original problem

This is the part most often omitted by students, but it is the most important!

Simply stating: “Since the p -value was less than or equal to the significance level α we reject the null hypothesis” does not warrant full credit

HYPOTHESIS TEST: NHANES

We are going to carry out a hypothesis test to answer the following question:

Do U.S. adults tend to be overweight ?

$$H_0 : \mu_{\text{BMI}} = 25$$

$$H_a : \mu_{\text{BMI}} > 25$$

We can treat `openintro::nhanes.samp.adult` as a random sample of $n = 135$ from the population .

```
nhanes_sample = openintro::nhanes.samp.adult
```

We need obtain the test statistic T , so we must obtain the sample mean ($\bar{x}$), and sample standard deviation (s)

First we create vector containing all the students BMI measurements

```
nhanes_sample_bmi = nhanes_sample$BMI
```

Then compute $\bar{x}$ and s

```
x_bar = mean(nhanes_sample_bmi)
s = sd(nhanes_sample_bmi)
n = length(nhanes_sample_bmi)
```

Here, $\bar{x} = 29.1$ and $s = 7.55$. Now we can construct the test statistic

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

```
mu_0 = 25 #null value  
t = (x_bar - mu_0)/(s/sqrt(n))
```

```
[1] 6.306551
```

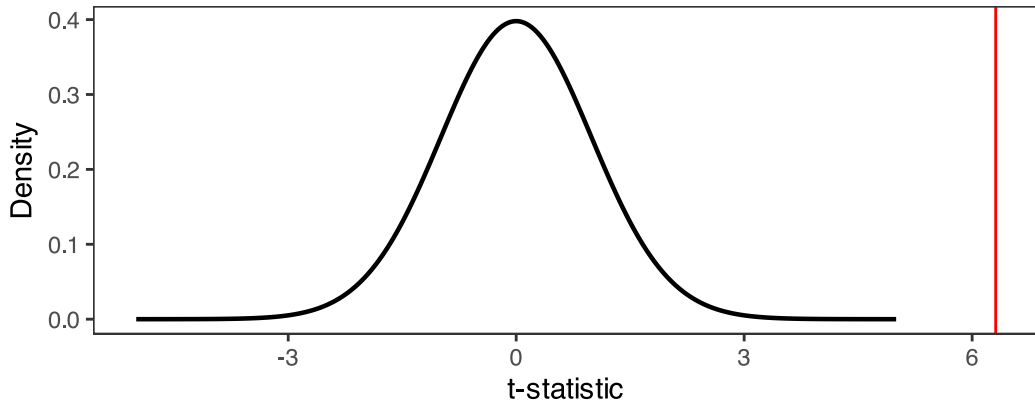
We now calculate the p -value which correspond to $P(T_{134} \geq 6.31)$

```
[1] 1.92244e-09
```

The probability we observe a test statistic as extreme as the one we observed ($t = 6.31$) is nearly 0.

We have sufficient statistical evidence to reject the null hypothesis and suggest that the average BMI of U.S adults is greater than 25, i.e they are overweight

t-distribution (df = 134)



In R, a hypothesis test for population mean μ can easily be performed using `t.test()`

```
bmi_test <- t.test(nhanes_sample_bmi, mu = 25, alternative = "greater")
```

One Sample t-test

data: nhanes_sample_bmi

t = 6.3066, df = 134, p-value = 1.922e-09

alternative hypothesis: true mean is greater than 25

95 percent confidence interval:

28.02288 Inf

sample estimates:

mean of x

29.09956

DECISION ERRORS

In hypothesis testing, decision errors occur when the conclusion drawn from the test does not reflect the true state of nature

	Fail to reject H_0	Reject H_0
H_0 is true	Correct decision $1 - \alpha$	Type I error "false positive" α
H_0 is false	Type II error "false negative" β	Correct decision $1 - \beta$ (Power)

HYPOTHESIS TESTS AND CONFIDENCE INTERVALS

The relationship between a hypothesis test and the corresponding confidence interval is defined by α

- **Hypothesis Test:** Is the observed sample statistic sufficiently far from the hypothesized population parameter to reject the hypothesis?
- **Confidence Interval:** Is a proposed population parameter sufficiently close to the observed sample statistic to be considered plausible?

In both cases, what constitutes “sufficiently far” or “sufficiently close” depends on the significance level α and the sampling variability.

In practice, if a $100(1 - \alpha)\%$ confidence interval for a population mean does not contain a hypothesized value μ_0 , then

- the data contradict the null hypothesis $H_0 : \mu = \mu_0$ at significance level α
- the implied two-sided alternative hypothesis is $H_a : \mu \neq \mu_0$