

# Hypothesis Testing and Analysis of Variance

# Descriptive and Inferential Statistics

**Descriptive Statistics:** The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

**Inferential statistics** takes data from a sample and makes inferences about the larger population from which the sample was drawn.

# Descriptive Statistics

- With **descriptive statistics** we condense a set of known numbers into a few simple values (either **numerically or graphically**) to simplify an understanding of those data that are available to us.
- This is analogous to writing up a **summary of a lengthy book**.
- The book summary is a tool for conveying the gist of a story to others.
- The **mean** and **standard deviation** of a set of numbers is a tool for conveying the gist of the individual numbers (without having to specify each and every one).

# Inferential Statistics

- The branch of statistics that analyzes sample data to draw conclusions about a population.

# Inferential Statistics

- **Inferential statistics** is used to make claims about the populations that give rise to the data we collect.
- This requires that we go beyond the data available to us.
- Consequently, the claims we make about populations are always subject to error; hence the term "**inferential statistics**" and not deductive statistics.
- Inferential statistics encompasses a variety of procedures to ensure that the **inferences are sound and rational**, even though they may not always be correct.
- Hence in short, inferential statistics enables us to make confident decisions in the face of uncertainty.
- At best, we can only be confident in our statistical assertions, but never certain of their accuracy.

# Relation between Descriptive and Inferential Statistics

Statistics  
(=“state  
arithmetic”)

## **Descriptive: describe data**

- How rich are our citizens on average? → **Central Tendency**
- Are there many differences between rich and poor? → **Variability**
- Are more intelligent people richer? → **Association**
- How many people earn this money? → **Probability distribution**
- Tools: tables (all kinds of summaries), graphs (all kind of plots), distributions (joint, conditional, marginal, ...), statistics (mean, variance, correlation coefficient, histogram, ...)

## **Inferential: derive conclusions and make predictions**

- Is my country so rich as my neighbors? → **Inference**
- To measure richness, do I have to consider EVERYONE? → **Sampling**
- If I don't consider everyone, how reliable is my estimate? → **Confidence**
- Is our economy in recession? → **Prediction**
- What will be the impact of an expensive oil? → **Modelling**
- Tools: Hypothesis testing, Confidence intervals, Parameter estimation, Experiment design, Sampling, Time models, Statistical models (ANOVA, Generalized Linear Models, ...)

# Kernel density estimate

- Kernel density estimation (KDE), also known as the Parzen's window, is one of most well-known approaches to estimate the underlying probability density function of a dataset.

# Probability Distribution Function (PDF)

- A **probability distribution is a function** that describes the likelihood of obtaining the possible values that a random variable can assume.
- In other words, the values of the **variable vary based** on the underlying probability distribution.



# Probability Density Function (PDF)

- ❑ A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence.
- ❑ **Probability Distribution Prerequisites**
  - To understand probability distributions, it is important to understand variables, random variables, and some notation.
  - A **variable** is a symbol ( $A$ ,  $B$ ,  $x$ ,  $y$ , etc.) that can take on any of a specified set of values.
  - When the value of a variable is the outcome of a [statistical experiment](#), that variable is a **random variable**.
  - Generally, statisticians use a capital letter to represent a random variable and a lower-case letter, to represent one of its values. For example,
    - $X$  represents the random variable  $X$ .
    - $P(X)$  represents the probability of  $X$ .
    - $P(X = x)$  refers to the probability that the random variable  $X$  is equal to a particular value, denoted by  $x$ . As an example,  $P(X = 1)$  refers to the probability that the random variable  $X$  is equal to 1.

# Probability Density Function (PDF)

## Probability Distributions

- Example will make clear relationship between random variables and probability distributions.
- Suppose you flip a coin two times.
- Have four possible outcomes: **HH, HT, TH, and TT.**
- Now, let variable  $X$  represent number of Heads in this experiment.
- The variable  $X$  can take on the values 0, 1, or 2.
- $X$  is a random variable; because its value is determined by outcome of a **statistical experiment**.
- A **probability distribution** is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence.
- The table below, which associates each outcome with its probability, is an example of a probability distribution.

Number of heads	Probability
0	0.25
1	0.50
2	0.25

**Above table represents the probability distribution of the random variable  $X$ .**

# Cumulative Probability Distributions

- A **cumulative probability** refers to the probability that value of a random variable falls within a specified range.
- Let us return to the coin flip experiment.
- If we flip a coin two times, we might ask: What is the probability that the coin flips would result in one or fewer heads?
- The answer would be a cumulative probability.
- It would be the probability that the coin flip experiment results in zero heads plus the probability that the experiment results in one head.
$$P(X \leq 1) = P(X = 0) + P(X = 1) = 0.25 + 0.50 = 0.75$$
- Like a probability distribution, a cumulative probability distribution can be represented by a table or an equation.
- In the table below, the cumulative probability refers to the probability than the random variable  $X$  is less than or equal to  $x$ .

Number of heads: $x$	Probability: $P(X = x)$	Cumulative Probability: $P(X \leq x)$
0	0.25	0.25
1	0.50	0.75
2	0.25	1.00

# Uniform Probability Distribution

- The simplest probability distribution occurs when all of the values of a random variable occur with equal probability. This probability distribution is called the **uniform distribution**.
- Suppose the random variable  $X$  can assume  $k$  different values. Suppose also that the  $P(X = x_k)$  is constant. Then,

$$P(X = x_k) = 1/k$$

- Suppose a die is tossed. What is the probability that the die will land on 5?
  - ❑ *Solution:* When a die is tossed, there are 6 possible outcomes represented by:  $S = \{ 1, 2, 3, 4, 5, 6 \}$ .
  - ❑ Each possible outcome is a random variable ( $X$ ), and each outcome is equally likely to occur. Thus, we have a uniform distribution. Therefore, the  $P(X = 5) = 1/6$ .

## Discrete Probability Distributions

If a random variable is a discrete variable, its probability distribution is called a **discrete probability distribution**.

## Continuous probability distribution

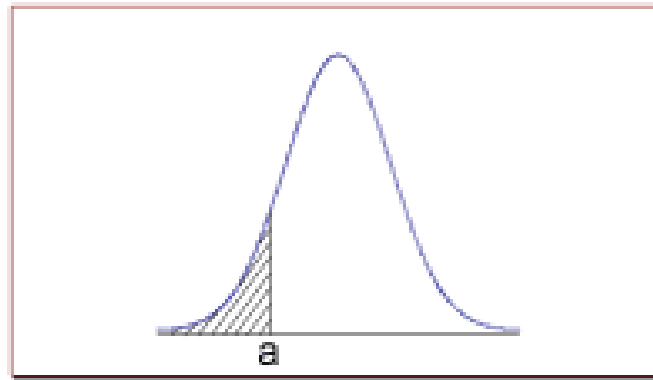
- If a random variable is a continuous variable, its probability distribution is called a **continuous probability distribution**.

# Probability Density Function

- Most often, the equation used to describe a continuous probability distribution is called a **probability density function**.
- A **probability distribution is a function** that describes the likelihood of obtaining the possible values that a random variable can assume.
- Sometimes, it is referred to as a **density function**, a **PDF**, or a **pdf**.
- **PDF Properties:**
  1. Since the continuous random variable is defined over a continuous range of values (**called the domain of the variable**), the graph of the density function will also be continuous over that range.
  2. The area bounded by the curve of the **density function and the x-axis is equal to 1**, when computed over the domain of the variable.
  3. The probability that a random variable assumes a value between  **$a$  and  $b$**  is equal to the area under the density function bounded by  **$a$  and  $b$** .

# Probability Density Function

- For example, consider the probability density function shown in the graph below.
- Suppose we wanted to know the probability that random variable  $X$  was less than or equal to  $a$ .
- The probability that  $X$  is less than or equal to  $a$  is equal to the **area under the curve** bounded by  $a$  and minus infinity - as indicated by the shaded area



## Note:

- The shaded area in the graph represents the probability that the random variable  $X$  is less than or equal to  $a$ . (cumulative probability).
- However, the probability that  $X$  is *exactly* equal to  $a$  would be zero.
- A continuous random variable can take on an infinite number of values.
- The probability that it will equal a specific value (such as  $a$ ) is always zero.

# Kernel Density Estimation (KDE)

- [Kernel density estimation \(KDE\)](#), also known as Parzen's window, is one of most well-known approaches to estimate the underlying probability density function of a dataset.
- A random variable  $x$  has a probability distribution  $p(x)$ .
- In probability theory and statistics, a **probability distribution** is the mathematical function that gives the probabilities of occurrence of different possible **outcomes** for an experiment.
- The relationship between the outcomes of a random variable and its probability is referred to as the probability density, or simply the "*density*."
- If a random variable is continuous, then the probability can be calculated via probability density function, or PDF for short.
- [Kernel density estimation](#) is a **non-parametric** method of estimating the probability density function (PDF) of a continuous random variable. It is non-parametric because it does not assume any underlying distribution for the variable.



# Kernel Density Estimation (KDE)

- Discuss fundamentals of kernel function and its use to estimate kernel density
- Gaussian kernel is used for density estimation and bandwidth optimization.
- Maximum likelihood cross-validation method is used for bandwidth optimization.

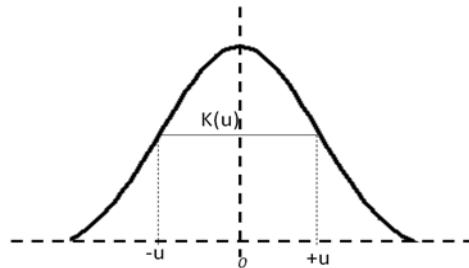
# Kernel Density Estimation (KDE)

## What Is a Kernel?

- Kernel is simply a function which satisfies following three properties as mentioned below.
- Kernel functions are used to estimate density of random variables and as weighing function in [non-parametric regression](#).

## Kernel Properties

1. The first property of a kernel function is that **it must be symmetrical**



2. The **area under the curve of the function must be equal to one**.  
Mathematically, this property is expressed as

$$\int_{-\infty}^{+\infty} K(u) du = 1$$

3. The value of kernel function, which is the density, **can not be negative**,  $K(u) \geq 0$  for all  $-\infty < u < \infty$ .

# Kernel Density Estimation (KDE)

## Construct Kernels

- Can use Gaussian kernel function to estimate kernel density and to optimize bandwidth using example data sets.
- The equation for Gaussian kernel is:

$$K(x) = \frac{1}{h\sqrt{2\pi}} e^{-0.5\left(\frac{x-x_i}{h}\right)^2}$$

Where  ***$x_i$***  is the ***observed data*** point.

***$x$***  is the ***value where kernel function is***  
computed and

***$h$***  is called the ***bandwidth***.

# Kernel Density Estimation (KDE)

## Example

- Let's say, we have marks obtained by six students in a subject.
- Will construct kernel at each data point using Gaussian kernel function.

$$x_i = \{65, 75, 67, 79, 81, 91\}$$

$$x_1 = 65, x_2 = 75 \dots x_6 = 91.$$

- Three inputs are required to develop a kernel curve around a data point.
  - I. The observation data point which is  $x_i$
  - II. The value of  $h$
  - III. A linearly spaced series of data points which houses the observed data points where  $K$  values are estimated.  $X_j = \{50, 51, 52 \dots 99\}$

# Kernel Density Estimation (KDE)

## Example

Calculation of  $K$  values for all values of  $X_j$  for a given values of  $x_i$  and  $h$  is shown in the table below; where  $x_i = 65$  and  $h = 5.5$ .

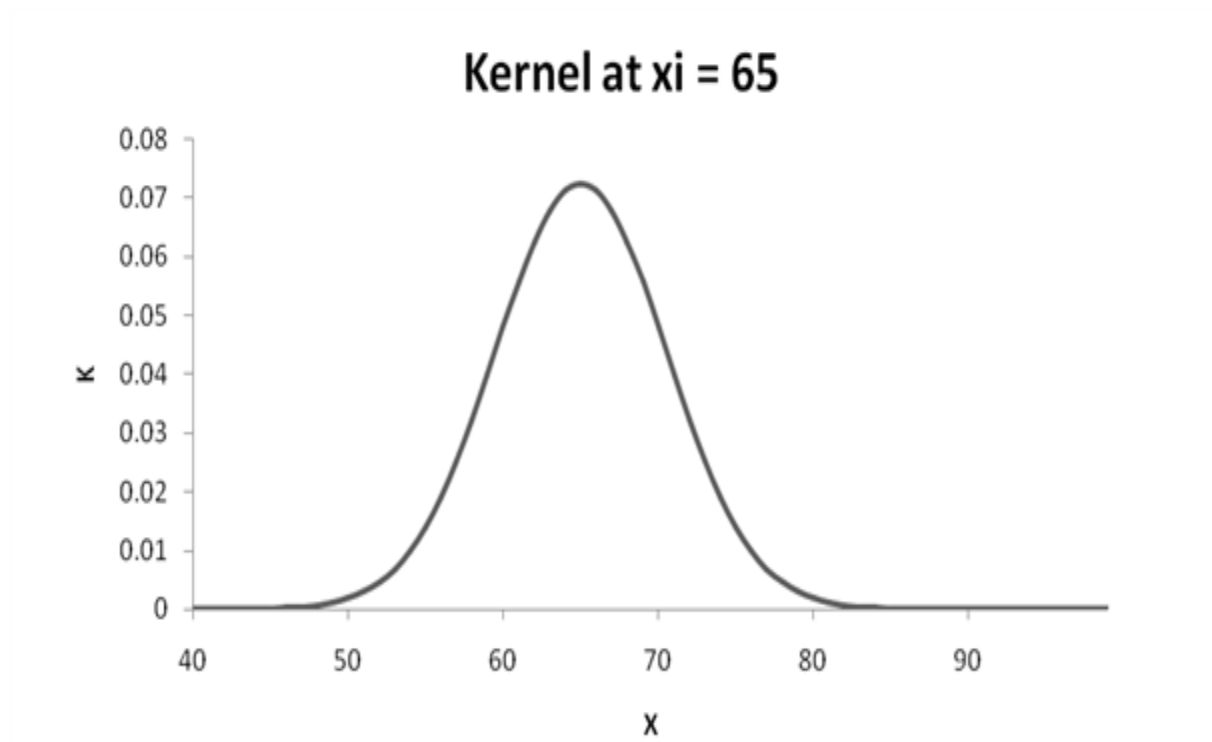
$X_j$	$x_i$	$h$	$A = \frac{1}{h\sqrt{2\pi}}$	$B = -0.5\left(\frac{X_j - x_i}{h}\right)^2$	$K = Ae^B$
50	65	5.5	0.072536	-3.71901	0.00175958
51	65	5.5	<b>0.072536</b>	<b>-3.23967</b>	<b>0.002841733</b>
52	65	5.5	0.072536	-2.79339	0.00444018
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
99	65	5.5	<b>0.072536</b>	<b>-19.1074</b>	<b>0.000000000365</b>
Sum					1.000

$$K(x) = \frac{1}{h\sqrt{2\pi}} e^{-0.5\left(\frac{x-x_i}{h}\right)^2}$$

# Kernel Density Estimation (KDE)

## Example

$X_j$  and  $K$  are plotted below to visualize the kernel.



# Kernel Density Estimation (KDE)

## Example

Similarly, at all six observed data points, kernel values are estimated as shown in the table and plotted below.

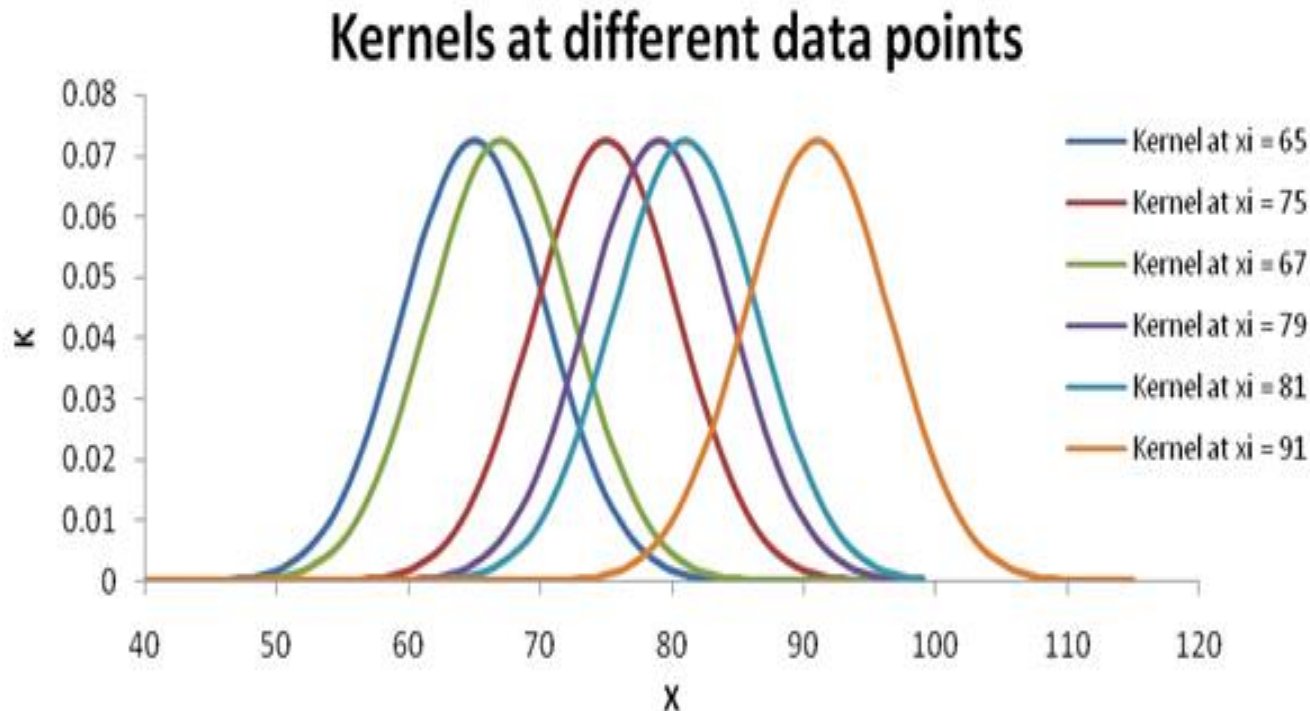
$x$	$K(x)$					
	$x_i = 65$	$x_i = 75$	$x_i = 67$	$x_i = 79$	$x_i = 81$	$x_i = 91$
50	0.00175958	0.00000237	0.00061093	0.00000007	0.00000001	0.00000000
51	0.00284173	0.00000532	0.00105409	0.00000017	0.00000003	0.00000000
52	0.00444018	0.00001157	0.00175958	0.00000042	0.00000007	0.00000000
53	0.00671214	0.00002433	0.00284173	0.00000102	0.00000017	0.00000000
-	-	-	-	-	-	-
.	.	.	.	.	.	.
-	-	-	-	-	-	-
78	0.00444	0.06251	0.009817	0.071347	0.06251	0.00444
79	0.002842	0.05568	0.006712	0.072536	0.067895	0.006712
80	0.00176	0.047984	0.00444	0.071347	0.071347	0.009817
81	0.001054	0.040007	0.002842	0.067895	0.072536	0.01389
.	.	.	.	.	.	.
99	0.00000	0.0000000	0.00000	0.000000	0.000342567	0.025184586

- Value of kernel function is nearly 0 for  $x_j$  values those are quite far from  $x_i$ .
- For instance kernel density value at  $x_j = 99$  is zero when  $x_i = 65$ .

# Kernel Density Estimation (KDE)

## Example

Similarly, at all six observed data points, kernel values are estimated as shown in the table and plotted below.





# Kernel Density Estimation (KDE)

## Kernel Density Estimation (KDE)

- Computed individual kernels over data points
- Composite density values are calculated for whole data set
- It is estimated simply by adding the kernel values ( $K$ ) from all  $\mathbf{X}_j$
- With reference to the above table,  $KDE$  for whole data set is obtained by adding **all row values**.
- The sum is then normalized by dividing the **number of data points**, which is six in this example
- Normalization is done to bring the **area under  $KDE$  curve to one**.
- Therefore, the equation to calculate  $KDE$  for every  $\mathbf{X}_j$  is expressed as:

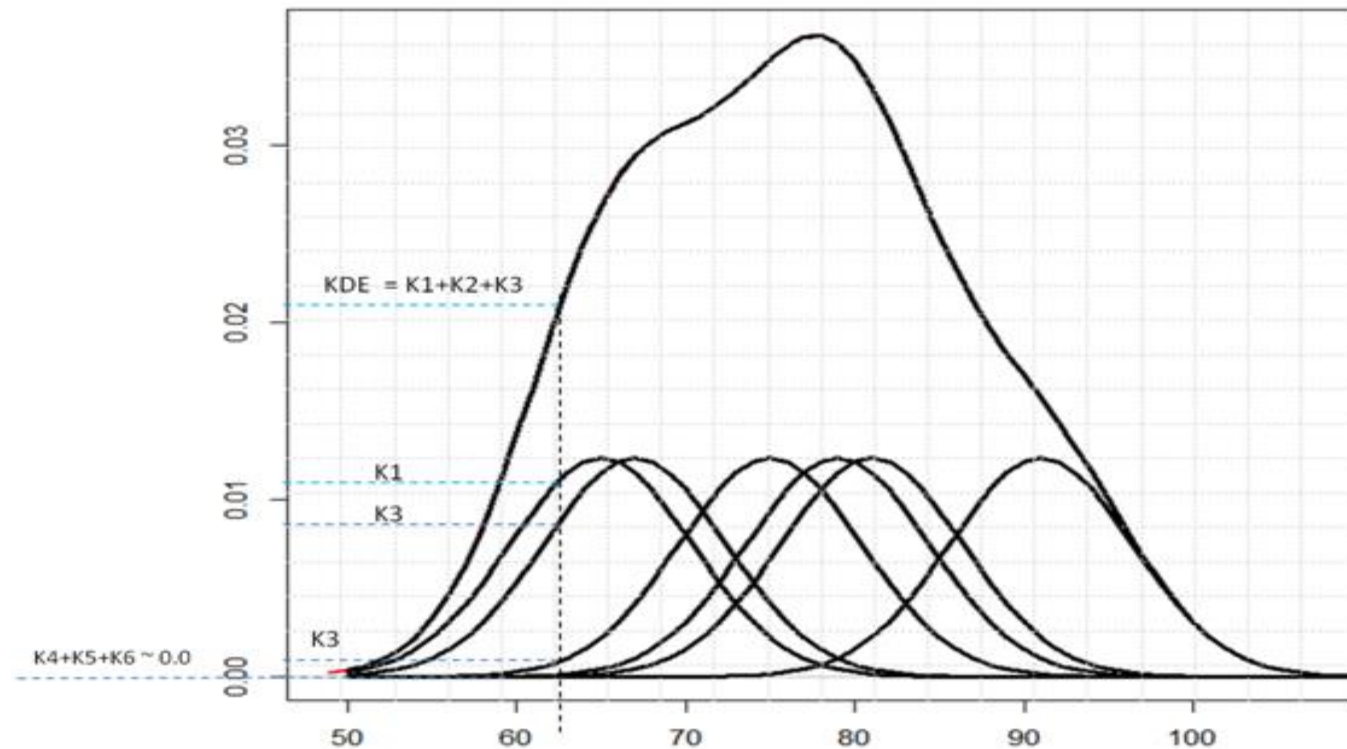
$$KDE_j = \frac{1}{n} \sum_{i=1}^{i=n} \frac{1}{h\sqrt{2\pi}} e^{-0.5\left(\frac{x_j - x_i}{h}\right)^2}$$

Where  $n$  is the number of data points.

# Kernel Density Estimation (KDE)

## Kernel Density Estimation (KDE)

The **KDE** after adding all six normalized kernels is shown below in the plot.



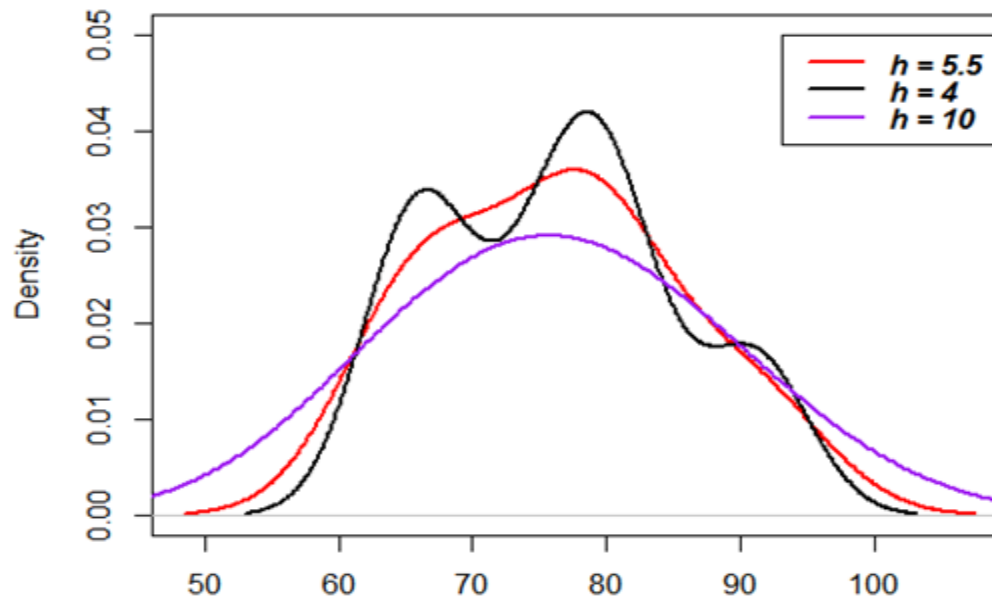
# Kernel Density Estimation (KDE)

## Bandwidth Optimization

- Bandwidth ( $h$ ) of a kernel function plays an important role to fit the data appropriately.
- A low value of bandwidth estimates density with lot of variance whereas high value of  $h$  produces large bias.
- Therefore estimation of an optimal value of  $h$  is very important to build most meaningful and accurate density.

# Kernel Density Estimation (KDE)

- As shown in the plot below, **three different** values of bandwidths produce **three different curves**.
- The **black one** gives lot of **variation in the density** values which **doesn't look realistic** whereas the **purple one fails** to explain the **actual density** by hiding information.
- There are several methods proposed by researchers to optimize the value of bandwidth in kernel density estimation.
- One among those is **maximum likelihood cross validation** method.



# Kernel Density Estimation (KDE)

- There are several methods proposed by researchers to optimize the value of bandwidth in kernel density estimation.
- One among those is **maximum likelihood cross validation** method.

# Cumulative Distribution Function (CDF)

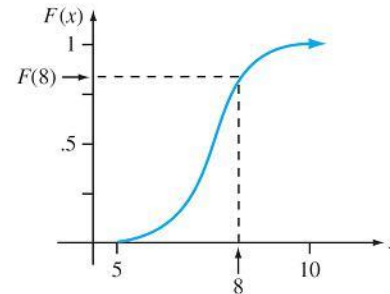
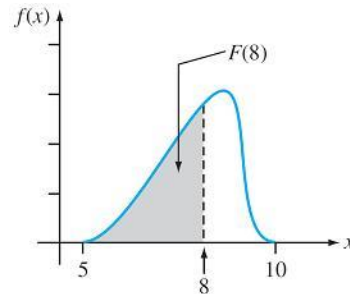
- The cumulative distribution function (cdf)  $F(x)$  for a discrete random variable  $X$  gives, for any specified number  $x$ , the probability  $P(X \leq x)$ .
- It is obtained by summing the probability mass function (pmf)  $p(y)$  over all possible values  $y$  satisfying  $y \leq x$ .
- In probability and statistics, a **probability mass function (pmf)** is a function that gives the probability that a discrete random variable is exactly equal to some value. Sometimes it is also known as the **discrete density function**.
- The cdf of a continuous random variable gives the same probabilities  $P(X \leq x)$  and is obtained by integrating the pdf  $f(y)$  between the limits.

# Cumulative Distribution Function (CDF)

- The **cumulative distribution function**  $F(x)$  for a continuous random variable  $X$  is defined for every number  $x$  by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(y) dy$$

- For each  $x$ ,  $F(x)$  is the area under the density curve to the left of  $x$ . This is illustrated in Figure, where  $F(x)$  increases smoothly as  $x$  increases.



# Cumulative Distribution Function (CDF)

- The **cumulative distribution function (CDF)** of a random variable is another method to describe the distribution of random variables.
- The advantage of the CDF is that it can be defined for any kind of random variable (**discrete, continuous, and mixed**).
- The cumulative distribution function (CDF) of random variable  $X$  is defined as

$$F_X(x) = P(X \leq x), \text{ for all } x \in \mathbb{R}.$$

- Note that CDF is defined for  $x \in \mathbb{R}$



# Cumulative Distribution Function (CDF)

- Example: **Toss a coin twice**. Let  $X$  be number of observed heads. Find CDF of  $X$ .
- Have four possible outcomes: HH, HT, TH, and TT.

The range of  $X$  is  $R_X = \{0, 1, 2\}$  and its PMF is given by

$$P_X(0) = P(X = 0) = \frac{1}{4},$$

$$P_X(1) = P(X = 1) = \frac{1}{2},$$

$$P_X(2) = P(X = 2) = \frac{1}{4}.$$

To find the CDF, we argue as follows. First, note that if  $x < 0$ , then

$$F_X(x) = P(X \leq x) = 0, \text{ for } x < 0.$$

Next, if  $x \geq 2$ ,

$$F_X(x) = P(X \leq x) = 1, \text{ for } x \geq 2.$$

Next, if  $0 \leq x < 1$ ,

$$F_X(x) = P(X \leq x) = P(X = 0) = \frac{1}{4}, \text{ for } 0 \leq x < 1.$$

Finally, if  $1 \leq x < 2$ ,

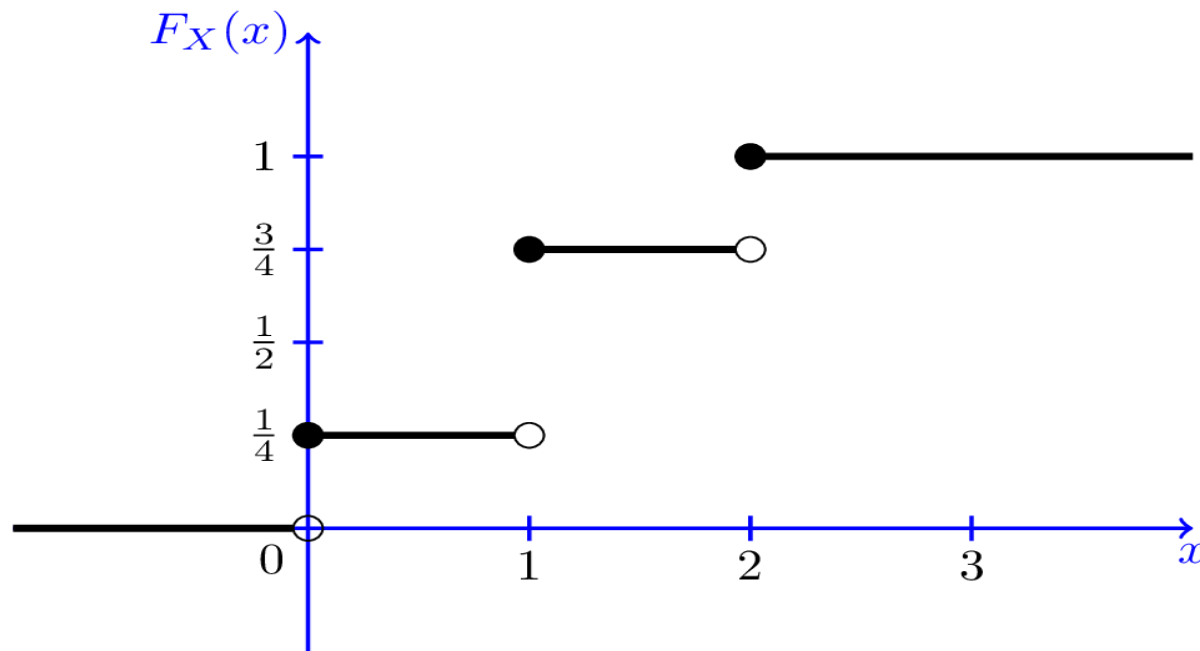
$$F_X(x) = P(X \leq x) = P(X = 0) + P(X = 1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}, \text{ for } 1 \leq x < 2.$$

Thus, to summarize, we have

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{4} & \text{for } 0 \leq x < 1 \\ \frac{3}{4} & \text{for } 1 \leq x < 2 \\ 1 & \text{for } x \geq 2 \end{cases}$$

# Cumulative Distribution Function (CDF)

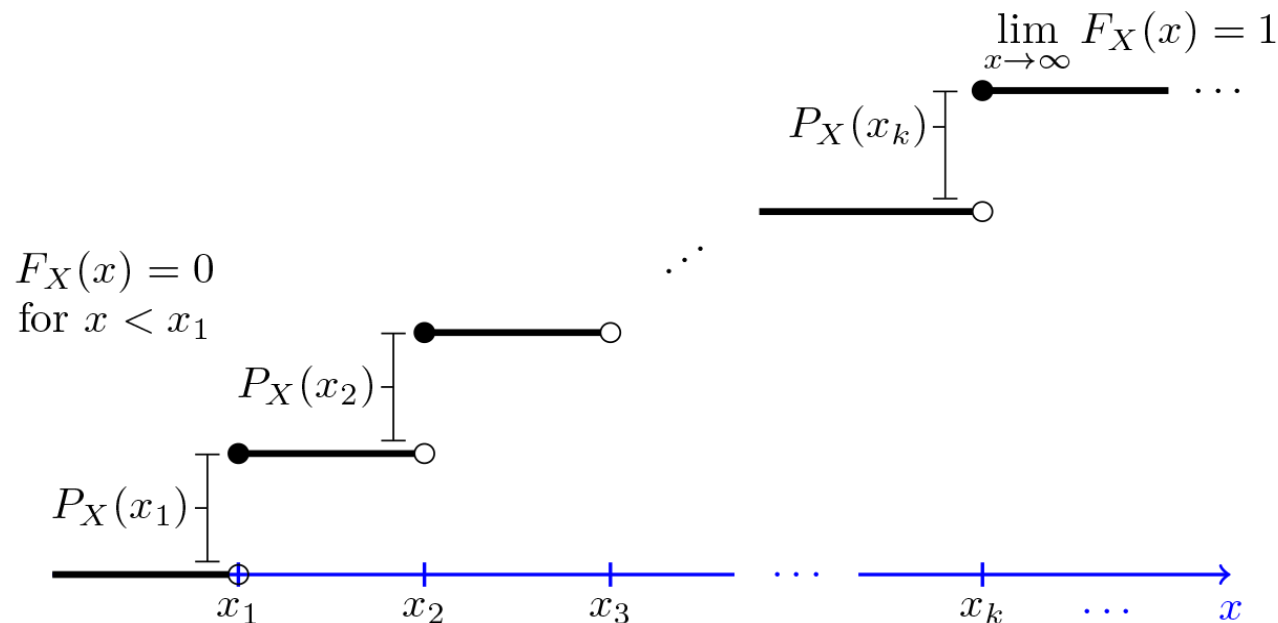
- To find CDF of a random variable, need to find the function for entire real line.
- For discrete random variables, we must be careful when to use "<" or "≤".
- Note that CDF is flat between points in  $R_X$  and jumps at each value in range.
- The size of the jump at each point is equal to the probability at that point.
- For, example, at point  $x=1$ , the CDF jumps from  $1/4$  to  $3/4$ . The size of the jump here is  $3/4 - 1/4 = 1/2$  which is equal to  $PX(1)$ .
- Also, note that the open and closed circles at point  $x=1$  indicate that  $FX(1)=3/4$  and not  $1/4$ .



# Cumulative Distribution Function (CDF)

- Let  $X$  be a discrete random variable with range  $R_X = \{x_1, x_2, x_3, \dots\}$ , such that  $x_1 < x_2 < x_3 < \dots$
- Assume that range  $R_X$  is bounded from below, i.e.,  $x_1$  is the smallest value in  $R_X$ .
- If this is not case then  $F_X(x)$  approaches zero as  $x \rightarrow -\infty$  rather than hitting zero.
- CDF is in the form of a **staircase**.
- CDF starts at 0; i.e.,  $F_X(-\infty) = 0$  and jumps at each point in the range.
- CDF stays flat between  $x_k$  and  $x_{k+1}$ , so we can write  

$$F_X(x) = F_X(x_k), \text{ for } x_k \leq x < x_{k+1}.$$
- CDF is always a non-decreasing function



# Cumulative Distribution Function (CDF)

- CDF completely describes distribution of a discrete random variable.
- Can find the PMF values by looking at values of jumps in the CDF function.
- If we have the PMF, we can find the CDF from it.
- If  $R_X = \{x_1, x_2, x_3, \dots\}$ , we can write

$$F_X(x) = \sum_{x_k \leq x} P_X(x_k).$$

# Cumulative Distribution Function (CDF)

## Example :

Let  $X$  be a discrete random variable with range  $R_X = \{1, 2, 3, \dots\}$ . Suppose the PMF of  $X$  is given by

$$P_X(k) = \frac{1}{2^k} \text{ for } k = 1, 2, 3, \dots$$

- a. Find and plot the CDF of  $X$ ,  $F_X(x)$ .
- b. Find  $P(2 < X \leq 5)$ .
- c. Find  $P(X > 4)$ .

## Solution

First, note that this is a valid PMF. In particular,

$$\sum_{k=1}^{\infty} P_X(k) = \sum_{k=1}^{\infty} \frac{1}{2^k} = 1 \text{ (geometric sum)}$$

- a. To find the CDF, note that

$$\text{For } x < 1, \quad F_X(x) = 0.$$

$$\text{For } 1 \leq x < 2, \quad F_X(x) = P_X(1) = \frac{1}{2}.$$

$$\text{For } 2 \leq x < 3, \quad F_X(x) = P_X(1) + P_X(2) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

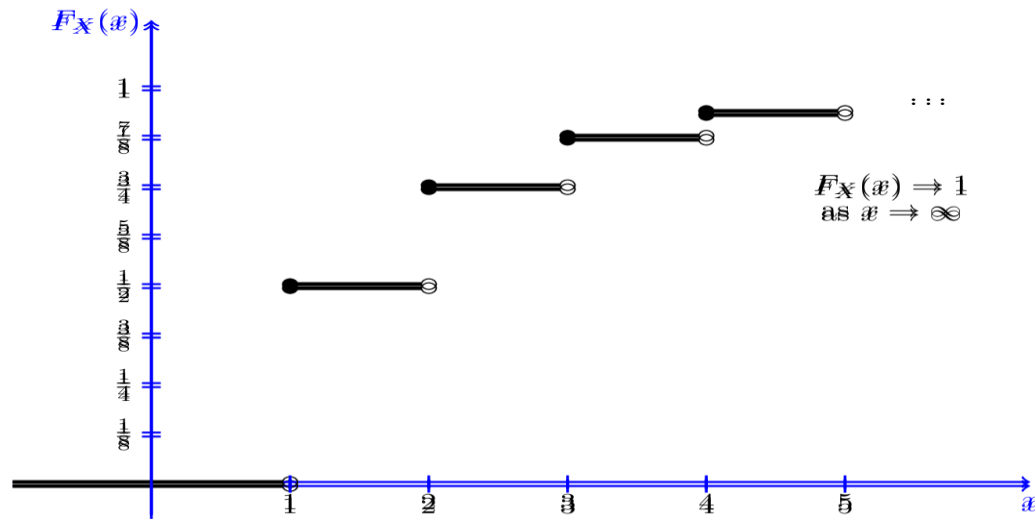
In general we have

$$\text{For } 0 < k \leq x < k+1,$$

$$F_X(x) = P_X(1) + P_X(2) + \dots + P_X(k)$$

$$= \frac{1}{2} + \frac{1}{4} + \dots + \frac{1}{2^k} = \frac{2^k - 1}{2^k}.$$

# Cumulative Distribution Function (CDF)



CDF for random variable

# Cumulative Distribution Function (CDF)

b. To find  $P(2 < X \leq 5)$ , we can write

$$P(2 < X \leq 5) = F_X(5) - F_X(2) = \frac{31}{32} - \frac{3}{4} = \frac{7}{32}.$$

Or equivalently, we can write

$$P(2 < X \leq 5) = P_X(3) + P_X(4) + P_X(5) = \frac{1}{8} + \frac{1}{16} + \frac{1}{32} = \frac{7}{32},$$

which gives the same answer.

c. To find  $P(X > 4)$ , we can write

$$P(X > 4) = 1 - P(X \leq 4) = 1 - F_X(4) = 1 - \frac{15}{16} = \frac{1}{16}.$$

# Hypothesis testing

- A statistical hypothesis is an expectation about a population.
- Usually it is formulated as a claim that a population parameter takes a particular value or falls within a specific range of values.
- On the basis of information from a sample we assess if a hypothesis makes sense or not.
- The significance test is, just like the confidence interval, a method of **inferential statistics**.
- Each significance test is based on two hypotheses: the **null hypothesis** and the **alternative hypothesis**.
- If you do a significance test, you assume that the null hypothesis is true unless your data provide strong evidence against it.



# Hypothesis testing

- The **Hypothesis Testing** is a statistical **test** used to determine whether the **hypothesis** assumed for the sample of data stands true for the **entire population** or not.
- Simply, the **hypothesis** is an assumption which is tested to determine the relationship between two data sets.
- A test of hypotheses is a statistical process for deciding between two competing assertions about a population parameter.

# Hypothesis testing

- Testing of hypothesis is one of the two important ways of studying unknown population parameters.
- In testing of hypothesis, generally some assumption is made regarding unknown population parameter; this assumption is called the hypothesis.
- Null hypothesis and alternate hypothesis are the two different types of hypothesis.
- The hypothesis of no difference is always called the null hypothesis and any alternate to the null hypothesis is called the alternate hypothesis.
- The notation “ $H_0$ ” denotes the null hypothesis, and the notation “ $H_1$ ” denotes the alternate hypothesis.

# Hypothesis testing

Hypothesis testing is formulated in terms of two hypotheses:

$H_0$ : the null hypothesis;

$H_1$ : the alternate hypothesis.

- The **null hypothesis** is the initial statistical claim that the population mean is equivalent to the claimed.
- The **null hypothesis** is a general statement or default position that there is no relationship between two variables.
- The **null hypothesis**  $H_0$  is the status quo hypothesis, representing what has been assumed about the value of the parameter.
- An **alternative hypothesis** states that there is statistical significance between two variables.
- The **alternative hypothesis** or research hypothesis  $H_a$  represents an alternative claim about the value of the parameter.

# Hypothesis testing

The hypothesis we want to test is if  $H_1$  is “likely” true. So, there are two possible outcomes:

- Reject  $H_0$  and accept  $H_1$  because of sufficient evidence in the sample in favour of  $H_1$ ;
- Do not reject  $H_0$  because of insufficient evidence to support  $H_1$ .

## Very important!!

Note that failure to reject  $H_0$  does not mean the null hypothesis is true. There is no formal outcome that says accept  $H_0$ . It only means that we do not have sufficient evidence to support  $H_1$ .

# Hypothesis testing

## Example

In a jury trial the hypotheses are:

$H_0$ : defendant is innocent;

$H_1$ : defendant is guilty.

- $H_0$  (innocent) is rejected if  $H_1$  (guilty) is supported by evidence beyond reasonable doubt.
- Failure to reject  $H_0$  (prove guilty) does not imply innocence, only that the evidence is **insufficient** to reject it.

# Hypothesis testing

A company manufacturing RAM chips claims the defective rate of the population is 5%. Let  $p$  denote the true defective probability. We want to test if:

$$H_0: p = 0.05$$

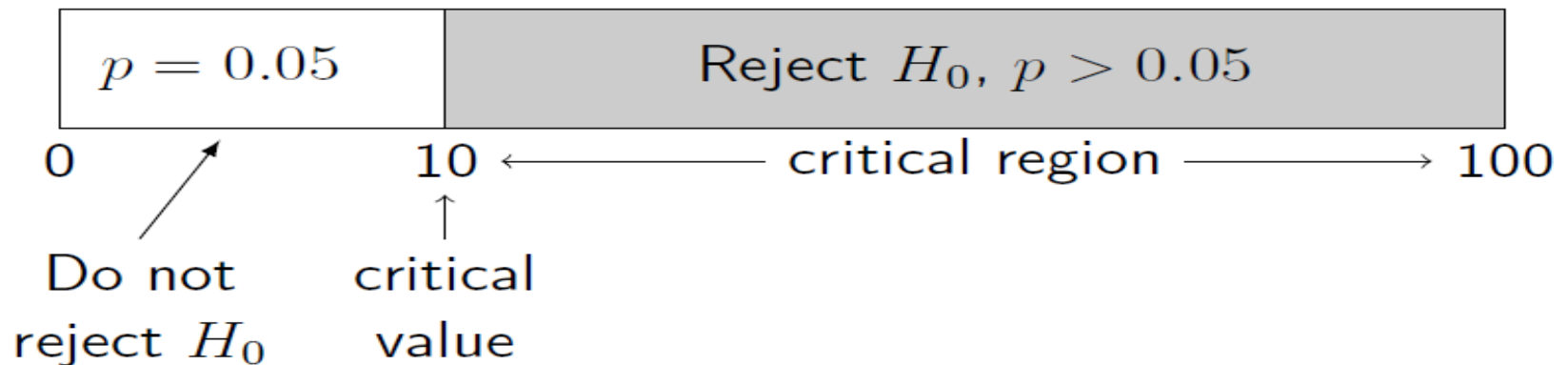
$$H_1: p > 0.05$$

We are going to use a sample of 100 chips from the production to test.

# Hypothesis testing

Let  $X$  denote the number of defective in the sample of 100. Reject  $H_0$  if  $X \geq 10$  (chosen “arbitrarily” in this case).

$X$  is called the *test statistic*.



- Why did we choose a critical value of 10 for this example? Because this is a **Bernoulli process**, the expected number of defectives in a sample is  **$np$** . So, if  $p = 0.05$  we should expect  $100 \times 0.05 = 5$  defectives in a sample of 100 chips. Therefore, 10 defectives would be **strong evidence** that  $p > 0.05$ .

# Hypothesis testing

In a jury trial the hypotheses are:

$H_0$ : defendant is innocent;

$H_1$ : defendant is guilty.

Four possible outcomes of the criminal trial hypothesis test

		Reality	
		$H_0$ true: Defendant did not commit crime	$H_0$ false: Defendant did commit crime
Jury's Decision	Reject $H_0$ : Find defendant guilty	Type I error	Correct decision
	Do not reject $H_0$ : Find defendant not guilty	Correct decision	Type II error

- The acceptance of  $H_1$  when  $H_0$  is true is called a **Type I error**. The probability of committing a type I error is called the **level of significance** and is denoted by  $\alpha$ .
- Failure to reject  $H_0$  when  $H_1$  is true is called a Type II error. The probability of committing a **type II error** is denoted by  $\beta$ .



# Hypothesis testing

Because we are making a decision based on a finite sample, there is a possibility that we will make mistakes.

The possible outcomes are:

	$H_0$ is true	$H_1$ is true
Do not reject $H_0$	Correct decision	Type II error
Reject $H_0$	Type I error	Correct decision

Rejecting the null hypothesis  $H_0$  when it is true is defined as a **type I error**.

Failing to reject the null hypothesis when it is false is defined as a **type II error**.

# Hypothesis testing

- The probability of a Type I error is denoted  $\alpha$
- The probability of a Type II error is denoted  $\beta$ .
- For a constant sample size, a decrease in  $\alpha$  is associated with an increase in  $\beta$ , and vice versa.
- In statistical analysis,  $\alpha$  is usually fixed at some small value, such as 0.05, and called the *level of significance (los)*.

# Hypothesis testing

- For example, suppose that we are interested in the **burning rate** of a **solid propellant** used to power aircrew escape systems.
- Burning rate is a **random variable**.
- Can be described by a **probability distribution**.
- Suppose that our interest focuses on the **mean burning rate** (a parameter of this distribution).
- We are interested in deciding whether or not the mean burning rate is 50 centimeters per second.
- We may express this formally as
  - $H_0$ : 50 centimeters per second – **null hypothesis**
  - $H_1$ :  $\neq 50$  centimeters per second – **alternate hypothesis**
- Hypotheses are always statements about the population or distribution under study, not statements about the sample.
- A procedure leading to a decision about a particular hypothesis is called a **test of a hypothesis**.

# Hypothesis testing

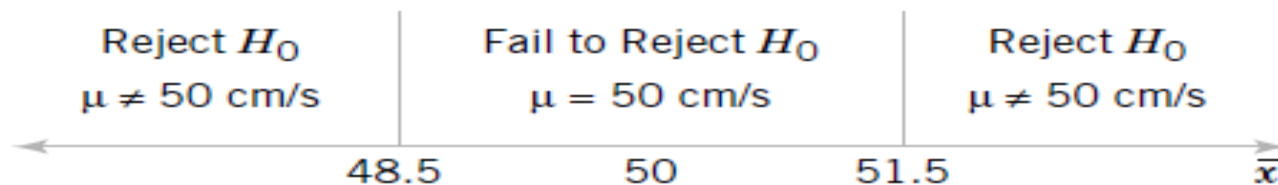
- Hypothesis-testing procedures rely on using the information in a random sample from the population of interest
- If this information is consistent with the hypothesis -hypothesis is true;
- if this information is inconsistent with the hypothesis - hypothesis is false.
- truth or falsity of a particular hypothesis can never be known with certainty, unless we can examine the entire population
- This is usually impossible in most practical situations
- Hypothesis-testing procedure should be developed with the probability of reaching a wrong conclusion in mind

# Hypothesis testing

- Structure of hypothesis-testing problems is identical in all applications.
- The **null hypothesis** is hypothesis we wish to test.
- **Rejection of the null hypothesis** always leads to **accepting the alternative hypothesis**.
- In our treatment of hypothesis testing, the null hypothesis will always be stated so that it specifies an exact value of the parameter.
- The alternate hypothesis will allow parameter to take on several values.
- Testing the hypothesis involves taking a random sample, computing a **test statistic** from the sample data.
- Then using test statistic to make a decision about null hypothesis.

# Tests of Statistical Hypotheses

- To illustrate the general concepts, consider propellant burning rate
- The null hypothesis is that the mean burning rate is 50 centimeters per second, and the alternate is that it is not equal to 50 centimeters per second.
- That is, we wish to test
  - $H_0$ : 50 centimeters per second
  - $H_1$ :  $\neq$  50 centimeters per second
- Suppose that a sample of  $n=10$
- Sample mean is an estimate of true population mean  $\mu$  and sample mean  $\bar{x}$ .
- Suppose that if  $48.5 \leq \bar{x} \leq 51.5$  - will not reject the null hypothesis
- If  $\bar{x} < 48.5$  or  $\bar{x} > 51.5$  - reject the null hypothesis



# Tests of Statistical Hypotheses

- Because our decision is based on random variables, probabilities can be associated with the type I and type II errors
- Probability of making a type I error is denoted by the Greek letter  $\alpha$

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

- Probability of making a type II error

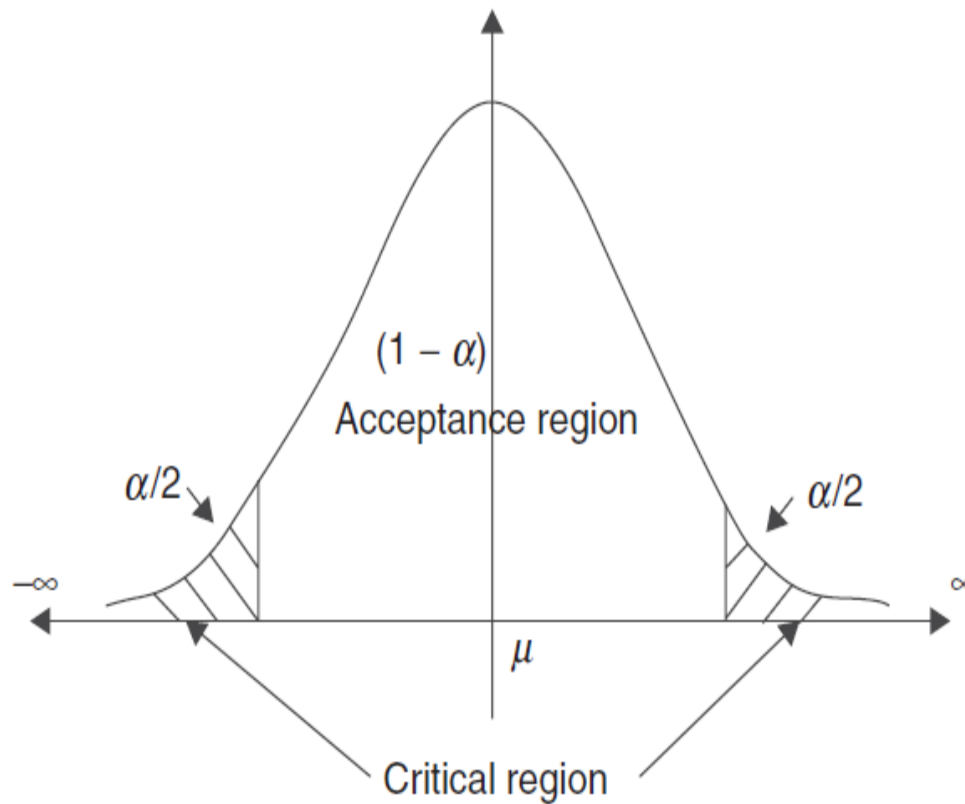
$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false})$$

# Hypothesis testing

- The decision regarding acceptance or rejection of null hypothesis is made with the help of certain region under the probability density curve of sampling distribution.
- This region is called the **critical region (CR)**
- If value of statistic, calculated, falls in this region, then null hypothesis  **$H_0$  is rejected**
- Types of critical regions
  - ✓ Two-tailed critical region and
  - ✓ one-tailed critical region



# Two-tailed critical region



□ A test of any hypothesis such as

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

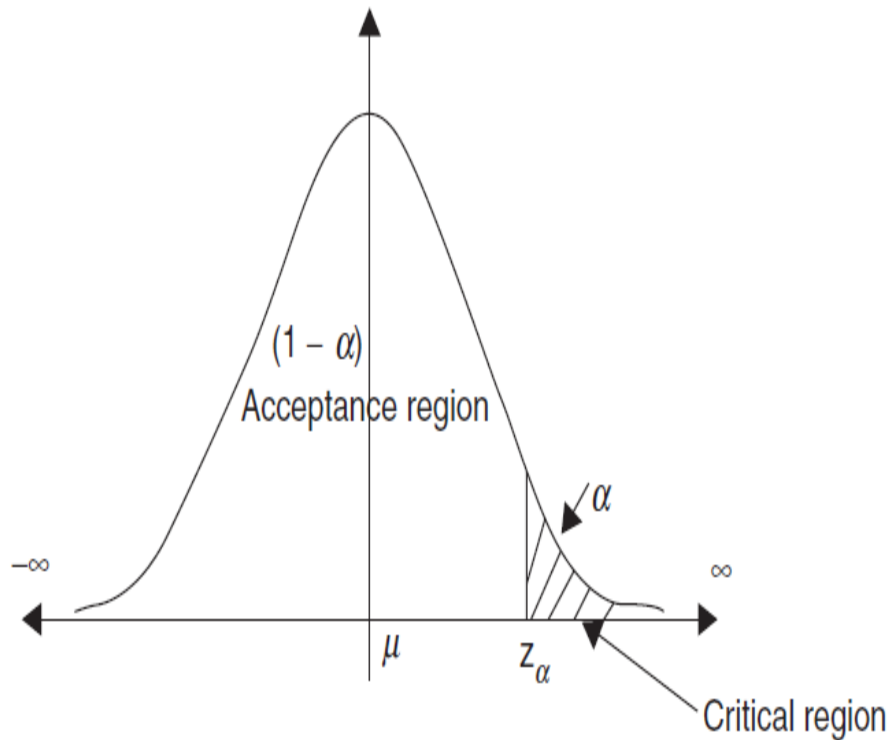
is called a **two-sided test**.

□ Need to detect differences from hypothesized value of mean that lie on either side of  $\mu_0$ .

□ Critical region is split into **two parts**

□ **Equal probability** placed in each tail of the distribution

# Right-tailed critical region

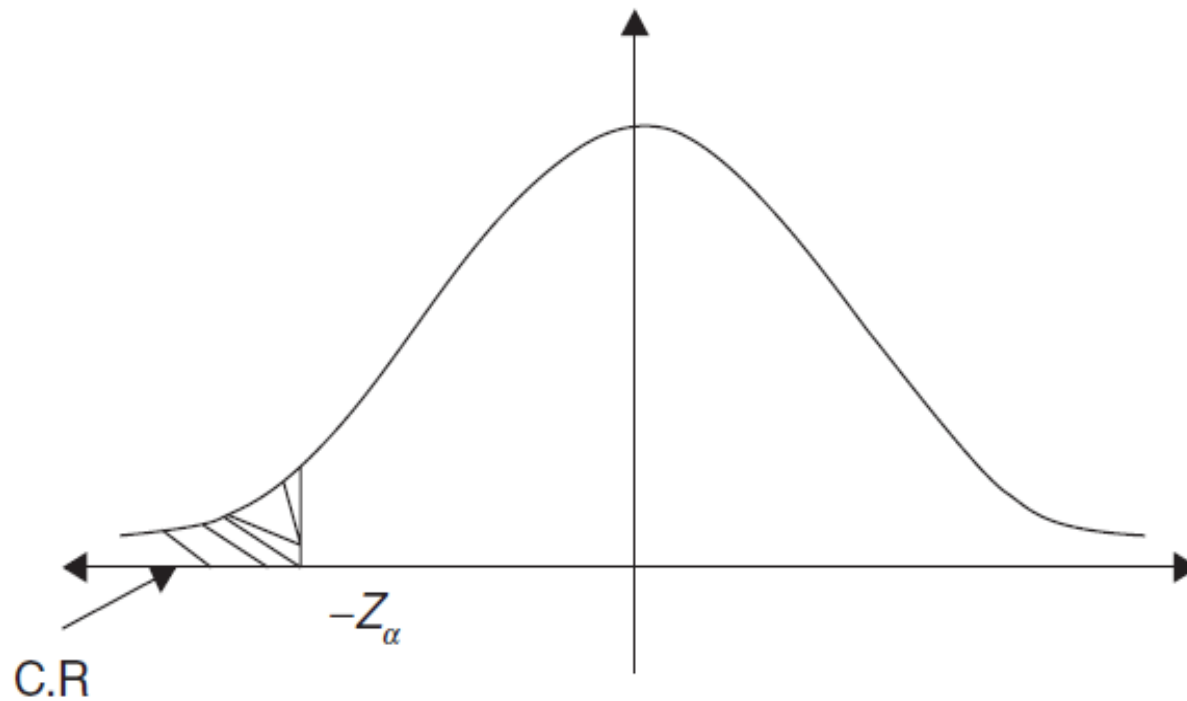


- Many hypothesis-testing problems naturally involve a **one-sided** alternative hypothesis, such as

$$\begin{array}{ll} H_0: \mu = \mu_0 & \text{or} & H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 & & H_1: \mu < \mu_0 \end{array}$$

- These tests are called **one-tailed** tests

# Left-tailed critical region



14.3

# Hypothesis testing

When the test statistic has the standard normal distribution:

Symbol in $H_a$	Terminology	Rejection Region
$<$	Left-tailed test	$(-\infty, -z_\alpha]$
$>$	Right-tailed test	$[z_\alpha, \infty)$
$\neq$	Two-tailed test	$(-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty)$

When the test statistic has Student's  $t$ -distribution:

Symbol in $H_a$	Terminology	Rejection Region
$<$	Left-tailed test	$(-\infty, -t_\alpha]$
$>$	Right-tailed test	$[t_\alpha, \infty)$
$\neq$	Two-tailed test	$(-\infty, -t_{\alpha/2}] \cup [t_{\alpha/2}, \infty)$

# General Procedure for Hypothesis Tests

1. From the problem context, identify the parameter of interest.
2. State the null hypothesis,  $H_0$ .
3. Specify an appropriate alternative hypothesis, .
4. Choose a significance level (level of significance ( $\alpha$ ))
5. Determine an appropriate test statistic.
6. State the rejection region for the statistic.
7. Compute any necessary sample quantities, substitute these into the equation for the test statistic, and compute that value.
8. Decide whether or not  $H_0$  should be rejected and report that in the problem context.

# One Sample T-Test

## One Sample $t$ Test

The One Sample  $t$  Test determines whether the sample mean is statistically different from a known or hypothesized population mean. The One Sample  $t$  Test is a parametric test.

This test is also known as:

- Single Sample  $t$  Test

The variable used in this test is known as:

- Test variable

In a One Sample  $t$  Test, the test variable is compared against a "test value", which is a known or hypothesized value of the mean in the population.

# One Sample T-Test

## Data Requirements

Your data must meet the following requirements:

1. Test variable that is continuous (i.e., interval or ratio level)
2. Scores on the test variable are independent (i.e., independence of observations)
  - There is no relationship between scores on the test variable
  - Violation of this assumption will yield an inaccurate  $p$  value
3. Random sample of data from the population
4. Normal distribution (approximately) of the sample and population on the test variable
  - Non-normal population distributions, especially those that are thick-tailed or heavily skewed, considerably reduce the power of the test
  - Among moderate or large samples, a violation of normality may still yield accurate  $p$  values
5. Homogeneity of variances (i.e., variances approximately equal in both the sample and population)
6. No outliers

# One Sample T-Test

## Hypotheses

The null hypothesis ( $H_0$ ) and (two-tailed) alternative hypothesis ( $H_1$ ) of the one sample  $T$  test can be expressed as:

$H_0: \mu = \bar{x}$  ("the sample mean is equal to the [proposed] population mean")

$H_1: \mu \neq \bar{x}$  ("the sample mean is not equal to the [proposed] population mean")

where  $\mu$  is a constant proposed for the population mean and  $\bar{x}$  is the sample mean.



# One Sample T-Test

## Test Statistic

The test statistic for a One Sample  $t$  Test is denoted  $t$ , which is calculated using the following formula:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

where

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where

$\mu$  = Proposed constant for the population mean

$\bar{x}$  = Sample mean

$n$  = Sample size (i.e., number of observations)

$s$  = Sample standard deviation

$s_{\bar{x}}$  = Estimated standard error of the mean ( $s/\text{sqrt}(n)$ )

The calculated  $t$  value is then compared to the critical  $t$  value from the  $t$  distribution table with degrees of freedom  $df = n - 1$  and chosen confidence level.

If the calculated  $t$  value > critical  $t$  value, then we reject the null hypothesis.

# One Sample T-Test

- ❑ The one sample t test compares the mean of your sample data to a known value.
- ❑ For example, you might want to know how your sample mean compares to the population mean.
- ❑ You should run a one sample t-test when
  - ✓ you don't know the population standard deviation or
  - ✓ you have a small sample size.

# One Sample T-Test Example

**Sample question:** your company wants to improve sales. Past sales data indicate that the average sale was \$100 per transaction. After training your sales force, recent sales data (taken from a sample of 25 salesmen) indicates an **average sale of \$130**, with a standard deviation of \$15. Did the training work? Test your hypothesis at a **5% alpha level**.

# One Sample T-Test Example

**Step 1:** Write your null hypothesis statement. The accepted hypothesis is that there is no difference in sales, so:

$$H_0: \mu = \$100.$$

**Step 2:** Write your alternate hypothesis. This is the one you're testing. You think that there *is* a difference (that the mean sales increased), so:

$$H_1: \mu > \$100.$$

# One Sample T-Test Example

**Step 3:** Identify the following pieces of information you'll need to calculate the test statistic. The question should give you these items:

1. **The sample mean( $\bar{x}$ ).** This is given in the question as \$130.
2. **The population mean( $\mu$ ).** Given as \$100 (from past data).
3. **The sample standard deviation( $s$ ) = \$15.**
4. **Number of observations( $n$ ) = 25.**

# One Sample T-Test Example

**Step 4:** Insert the items from above into the t score formula.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = (130 - 100) / ((15 / \sqrt{25}))$$

$$t = (30 / 3) = 10$$

**Calculated t-value : 10**

# One Sample T-Test Example

**Step 5:** Find the t-table value. You need two values to find this:

1. The alpha level: given as 5% in the question.
2. The degrees of freedom, which is the number of items in the sample (n) minus 1:  $25 - 1 = 24$ .

Look up 24 degrees of freedom in the left column and 0.05 in the top row. The intersection is 1.711.

One-tailed critical t-value : **1.711**

**In statistics, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary.**

# One Sample T-Test Example

## Step 6:

- Compare Step 4 to Step 5.
  - The value from Step 4 **does not** fall into the range calculated in Step 5, so we can reject the null hypothesis.
  - The value of 10 falls into the rejection region (the left tail).
  - **In other words, it's highly likely that the mean sale is greater.**
- The sales training was probably a success**



# One Sample T-Test Example


T-Distribution Table (One Tail)

13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.689
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.660
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
1000	1.282	1.646	1.962	2.330	2.581	3.098	3.300

# One Sample T-Test Example

Many doctors recommend having a total cholesterol level below 200 mg/dl. We will test to see if the 1952 population from which the Dixon and Massey sample was gathered is statistically different, on average, from this recommended level.

```
data dixonmassey;  
    input Obs chol52 chol62 age cor dchol age1t50 $;  
datalines;
```



1	240	209	35	0	-31	y
2	243	209	64	1	-34	n
3	250	173	61	0	-77	n
4	254	165	44	0	-89	y
5	264	239	30	0	-25	y
6	279	270	41	0	-9	y
7	284	274	31	0	-10	y
8	285	254	48	1	-31	y
9	290	223	35	0	-67	y
10	298	209	44	0	-89	y
11	302	219	51	1	-83	n
12	310	281	52	0	-29	n
13	312	251	37	1	-61	y
14	315	208	61	1	-107	n
15	322	227	44	1	-95	y
16	337	269	52	0	-68	n
17	348	299	31	0	-49	y
18	384	238	58	0	-146	n
19	386	285	33	0	-101	y
20	520	325	40	1	-195	y

# One Sample T-Test Example

**Step 1:** Write your null hypothesis statement.

$$H_0: \mu = 200.$$

**Step 2:**

$$H_1: \mu \neq 200$$

**Step 3:** Identify the following req. information

$$\alpha=0.05$$

Our sample of  $n=20$  has mean = 311.15 and  $s = 64.3929$ .

$df = 19$ , so reject  $H_0$  if  $|t| > 2.093$

**Step 4:** Insert the items from above into the t score formula

$$t = \frac{311.15 - 200}{64.3939 / \sqrt{20}} = 7.723$$

**Step 5:** Find the t-table value for  $df=19$  and  $\alpha=0.05$

**Two-tailed critical t-value : 2.093**

**Step 6:**  $|t| > 2.093$  so **We reject  $H_0$**

# One Sample T-Test Example

A professor wants to know if introductory statistics class has a good grasp of basic math. Six students are chosen at random from the class and given a math proficiency test. The professor wants the class to be able to score above 70 on the test. The six students get scores of 62, 92, 75, 68, 83, and 95. Can the professor have 90 percent confidence that the mean score for the class on the test would be above 70?

# One Sample T-Test Example

**Step 1:** Write your null hypothesis statement.

$$H_0: \mu = 70.$$

**Step 2:**

$$H_1: \mu > 70$$

**Step 3:** Identify the following req. information

$$\alpha = 0.10$$

Our sample of  $n=6$  has mean = 79.17 and  $s = 13.17$ .

$df = 5$ , so reject  $H_0$  if  $|t| > 1.4759$

**Step 4:** Insert the items from above into the t score formula

$$t = \frac{79.17 - 70}{13.17 / \sqrt{6}} = 1.71$$

**Step 5:** Find the t-table value for  $df=5$  and  $\alpha=0.10$

**One-tailed critical t-value :** 1.4759

**Step 6:**  $|t| > 1.4759$  so **We reject  $H_0$**

**Professor has evidence that the class mean on the math test would be at least 70.**

# One Sample T-Test Example

A Little League baseball coach wants to know if his team is representative of other teams in scoring runs. Nationally, the average number of runs scored by a Little League team in a game is 5.7. He chooses five games at random in which his team scored 5, 9, 4, 11, and 8 runs. Is it likely that his team's scores could have come from the national distribution? Assume an alpha level of 0.05.

# One Sample T-Test Example

**Step 1:** Write your null hypothesis statement.

$$H_0: \mu = 5.7.$$

**Step 2:**

$$H_1: \mu \neq 5.7$$

**Step 3:** Identify the following req. information

$$\alpha=0.05$$

Our sample of  $n=5$  has mean = 7.4 and  $s = 2.88$ .

$df = 4$ , so reject  $H_0$  if  $|t| > 2.1318$ .

**Step 4:** Insert the items from above into the t score formula

$$t = \frac{7.4 - 5.7}{2.88 / \sqrt{5}} = 1.32$$

**Step 5:** Find the t-table value for  $df=4$  and  $\alpha=0.05$

**Two-tailed critical t-value : 2.1318**

**Step 6:**  $|t| < 2.1318$  so **We Can't reject  $H_0$**

- Mean of this team is equal to the population mean.
- The coach cannot conclude that his team is different from the national distribution on runs scored.

# Independent Samples $t$ -Test

## Independent Samples $t$ Test

The Independent Samples  $t$  Test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. The Independent Samples  $t$  Test is a parametric test.

This test is also known as:

- Independent  $t$  Test
- Independent Measures  $t$  Test
- Independent Two-sample  $t$  Test
- Student  $t$  Test
- Two-Sample  $t$  Test
- Uncorrelated Scores  $t$  Test
- Unpaired  $t$  Test
- Unrelated  $t$  Test

The variables used in this test are known as:

- Dependent variable, or test variable
- Independent variable, or grouping variable



# Independent Samples t-Test

## Common Uses

The Independent Samples  $t$  Test is commonly used to test the following:

- Statistical differences between the means of two groups
- Statistical differences between the means of two interventions
- Statistical differences between the means of two change scores

**Note:** The Independent Samples  $t$  Test can only compare the means for two (and only two) groups. It cannot make comparisons among more than two groups. If you wish to compare the means across more than two groups, you will likely want to run an ANOVA.

# Independent Samples t-Test

## Hypotheses

The null hypothesis ( $H_0$ ) and alternative hypothesis ( $H_1$ ) of the Independent Samples  $t$  Test can be expressed in two different but equivalent ways:

$H_0: \mu_1 = \mu_2$  ("the two population means are equal")

$H_1: \mu_1 \neq \mu_2$  ("the two population means are not equal")

OR

$H_0: \mu_1 - \mu_2 = 0$  ("the difference between the two population means is equal to 0")

$H_1: \mu_1 - \mu_2 \neq 0$  ("the difference between the two population means is not 0")

where  $\mu_1$  and  $\mu_2$  are the population means for group 1 and group 2, respectively. Notice that the second set of hypotheses can be derived from the first set by simply subtracting  $\mu_2$  from both sides of the equation.

# Independent Samples t-Test

## Levene's Test for Equality of Variances

Recall that the Independent Samples  $t$  Test requires the assumption of *homogeneity of variance* -- i.e., both groups have the same variance. SPSS conveniently includes a test for the homogeneity of variance, called **Levene's Test**, whenever you run an independent samples  $t$  test.

The hypotheses for Levene's test are:

$H_0: \sigma_1^2 - \sigma_2^2 = 0$  ("the population variances of group 1 and 2 are equal")

$H_1: \sigma_1^2 - \sigma_2^2 \neq 0$  ("the population variances of group 1 and 2 are not equal")

This implies that if we reject the null hypothesis of Levene's Test, it suggests that the variances of the two groups are not equal; i.e., that the homogeneity of variances assumption is violated.

# Independent Samples t-Test

## Test Statistic

The test statistic for an Independent Samples  $t$  Test is denoted  $t$ . There are actually two forms of the test statistic for this test, depending on whether or not equal variances are assumed. SPSS produces both forms of the test, so both forms of the test are described here. **Note that the null and alternative hypotheses are identical for both forms of the test statistic.**

## EQUAL VARIANCES ASSUMED

When the two independent samples are assumed to be drawn from populations with identical population variances (i.e.,  $\sigma_1^2 = \sigma_2^2$ ), the test statistic  $t$  is computed as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{with} \quad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Where

$\bar{x}_1$  = Mean of first sample

$\bar{x}_2$  = Mean of second sample

$n_1$  = Sample size (i.e., number of observations) of first sample

$n_2$  = Sample size (i.e., number of observations) of second sample

$s_1$  = Standard deviation of first sample

$s_2$  = Standard deviation of second sample

$s_p$  = Pooled standard deviation

The calculated  $t$  value is then compared to the critical  $t$  value from the  $t$  distribution table with degrees of freedom  $df = n_1 + n_2 - 2$  and chosen confidence level. If the calculated  $t$  value is greater than the critical  $t$  value, then we reject the null hypothesis.

# Independent Samples t-Test

## EQUAL VARIANCES NOT ASSUMED

When the two independent samples are assumed to be drawn from populations with unequal variances (i.e.,  $\sigma_1^2 \neq \sigma_2^2$ ), the test statistic  $t$  is computed as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where

$\bar{x}_1$  = Mean of first sample

$\bar{x}_2$  = Mean of second sample

$n_1$  = Sample size (i.e., number of observations) of first sample

$n_2$  = Sample size (i.e., number of observations) of second sample

$s_1$  = Standard deviation of first sample

$s_2$  = Standard deviation of second sample

The calculated  $t$  value is then compared to the critical  $t$  value from the  $t$  distribution table with degrees of freedom

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2}\right)^2}$$

and chosen confidence level. If the calculated  $t$  value > critical  $t$  value, then we reject the null hypothesis.

# Independent Samples t-Test Example

A study of the effect of caffeine on muscle metabolism used eighteen male volunteers who each underwent arm exercise tests. Nine of the men were randomly selected to take a capsule containing pure caffeine one hour before the test. The other men received a placebo capsule. During each exercise the subject's respiratory exchange ratio (RER) was measured. (RER is the ratio of  $\text{CO}_2$  produced to  $\text{O}_2$  consumed and is an indicator of whether energy is being obtained from carbohydrates or fats).

**The question of interest to the experimenter was whether, on average, caffeine changes RER.**

The two populations being compared are “**men who have not taken caffeine**” and “**men who have taken caffeine**”.

# Independent Samples t-Test Example

The results were as follows:

	RER(%)	
	Placebo	Caffeine
	105	96
	119	99
	100	94
	97	89
	96	96
	101	93
	94	88
	95	105
	98	88
Mean	100.56	94.22
SD	7.70	5.61

# Independent Samples t-Test Example

## ### How to Perform Levene's Test in Python

```
In [ ]: import scipy.stats as stats
```

```
In [4]: group1 = [105,119,100,97,96,101,94,95,98]
group2 = [96,99,94,89,96,93,88,105,88]
```

```
In [5]: #Levene's test centered at the mean
stats.levene(group1, group2, center='mean')
```

```
Out[5]: LeveneResult(statistic=0.19717011725750563, pvalue=0.66296156344230539)
```

```
In [6]: #Levene's test centered at the median
stats.levene(group1, group2, center='median')
```

```
Out[6]: LeveneResult(statistic=0.05302226935312828, pvalue=0.82080358870458536)
```

```
# In both methods, the p-value is not less than .05. This means in both cases we would fail to reject
the null hypothesis.
```

```
In other words, the two groups have equal variances.
```

## EQUAL VARIANCES ASSUMED



# Independent Samples t-Test Example

**Step 1:** State the null hypothesis,  $H_0$  and an appropriate alternative hypothesis

Null hypothesis is

$H_0$ : On average, caffeine has no effect on RER,  
with an alternative (or experimental) hypothesis,

$H_1$ : On average, caffeine changes RER (2-tail test), or

$H_1$ : On average, caffeine reduces RER (1-tail case).

# Independent Samples t Test Example

**Step 2:** Calculate the test statistic  $t$  is computed as ( $n_1=n_2=9$ )

Sample 1 mean = 100.56, Sample 2 mean = 94.22

Sample 1 SD = 7.70, Sample 2 SD = 5.61

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{with} \quad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$s_p = \sqrt{\frac{(9-1) * (7.70)^2 + (9-1) * (5.61)^2}{9+9-2}} = 6.74$$

$$t = \frac{100.56 - 94.22}{6.74 * \sqrt{\frac{1}{9} + \frac{1}{9}}} = \frac{6.34}{3.17} = 2.00$$

**Step 3:** Find the t-table value for  $df = n_1 + n_2 - 2 = 16$  and  $\alpha = 0.05$

**Two-tailed critical t-value :** 2.1199

**Step 4:** If the calculated t value is greater than the critical t value, then we reject the null hypothesis.

**So  $2.00 < 2.11$ , null hypothesis not rejected**

# Paired Sample t-Test

## Paired Samples $t$ Test

The Paired Samples  $t$  Test compares two means that are from the same individual, object, or related units. The two means can represent things like:

- A measurement taken at two different times (e.g., pre-test and post-test with an intervention administered between the two time points)
- A measurement taken under two different conditions (e.g., completing a test under a "control" condition and an "experimental" condition)
- Measurements taken from two halves or sides of a subject or experimental unit (e.g., measuring hearing loss in a subject's left and right ears).

This test is also known as:

- Dependent  $t$  Test
- Paired  $t$  Test
- Repeated Measures  $t$  Test

The variable used in this test is known as:

- Dependent variable, or test variable (continuous), measured at two different times or for two related conditions or units

# Paired Sample t-Test

## ➤ Common Uses

The Paired Samples  $t$  Test is commonly used to test the following:

- ☐ Statistical difference between two time points
- ☐ Statistical difference between two conditions
- ☐ Statistical difference between two measurements
- ☐ Statistical difference between a matched pair

- **Note:** Paired Samples  $t$  Test can only compare the means for two (and only two) related (paired) units on a continuous outcome that is normally distributed.
- The Paired **Samples  $t$  Test is not appropriate** for analyses involving the following:
1. unpaired data;
  2. comparisons between more than two units/groups;
  3. a continuous outcome that is not normally distributed; and
  4. an ordinal/ranked outcome.

# Paired Sample t-Test

## Note

- To compare unpaired means between two groups on a continuous outcome that is normally distributed, choose the Independent Samples  $t$  Test.
- To compare unpaired means between more than two groups on a continuous outcome that is normally distributed, choose ANOVA.
- To compare paired means for continuous data that are not normally distributed, choose the nonparametric Wilcoxon Signed-Ranks Test.
- To compare paired means for ranked data, choose the nonparametric Wilcoxon Signed-Ranks Test.

# Paired Sample t-Test

## Data Requirements

1. Dependent variable that is **continuous** (i.e., interval or ratio level)
  - The paired measurements must be recorded in two separate variables.
2. Related samples/groups (i.e., dependent observations)
  - The subjects in each sample, or group, are the same.
  - This means that the subjects in the first group are also in the second group.
3. Random sample of data from the population
4. Normal distribution (approximately) of the difference between the paired values
5. No outliers in the difference between the two related groups.

**Note:** When testing assumptions related to normality and outliers, you must use a variable that represents the difference between the paired values - not the original variables themselves.

**Note:** When one or more of the assumptions for Paired Samples  $t$  Test are not met, you may want to run the nonparametric Wilcoxon Signed-Ranks Test instead.

# Paired Sample t-Test

## Hypotheses

The hypotheses can be expressed in two different ways that express the same idea and are mathematically equivalent:

$H_0: \mu_1 = \mu_2$  ("the paired population means are equal")

$H_1: \mu_1 \neq \mu_2$  ("the paired population means are not equal")

OR

$H_0: \mu_1 - \mu_2 = 0$  ("the difference between the paired population means is equal to 0")

$H_1: \mu_1 - \mu_2 \neq 0$  ("the difference between the paired population means is not 0")

where

- $\mu_1$  is the population mean of variable 1, and
- $\mu_2$  is the population mean of variable 2.

# Paired Sample t-Test

## Test Statistic

The test statistic for the Paired Samples  $t$  Test, denoted  $t$ , follows the same formula as the one sample  $t$  test.

$$t = \frac{\bar{x}_{\text{diff}}}{s_{\bar{x}}}$$

where

$$s_{\bar{x}} = \frac{s_{\text{diff}}}{\sqrt{n}}$$

where

$\bar{x}_{\text{diff}}$  = Sample mean of the differences

$n$  = Sample size (i.e., number of observations)

$s_{\text{diff}}$  = Sample standard deviation of the differences

$s_{\bar{x}}$  = Estimated standard error of the mean ( $s/\text{sqrt}(n)$ )

The calculated  $t$  value is then compared to the critical  $t$  value with  $df = n - 1$  from the  $t$  distribution table for a chosen confidence level.

If the calculated  $t$  value is greater than the critical  $t$  value, then we reject the null hypothesis (and conclude that the means are significantly different)



# Paired Sample t-Test - Example

- ❑ Suppose a sample of  $n$  students were given a diagnostic test before studying a particular module and then again after completing the module.
- ❑ We want to find out if, in general, our teaching leads to improvements in students' knowledge/skills (i.e. test scores).
- ❑ We can use results from our sample of students to draw conclusions about the impact of this module in general.
- ❑ Let  $x$  = test score before the module,  
     $y$  = test score after the module
- ❑ Assume  $\alpha = 0.05$

# Paired Sample t-Test - Example

Using the above example with  $n = 20$  students, the following results were obtained:

Student	Pre-module score	Post-module score	Difference
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1

# Paired Sample t-Test - Example

**Step 1:** State the null hypothesis,  $H_0$  and an appropriate alternative hypothesis

## Null hypothesis

$$H_0: \mu_1 = \mu_2$$

No impact of this module in general i.e. mean difference is zero

## Alternative hypothesis

$$H_1: \mu_1 \neq \mu_2$$

On average, the module does lead to improvements.

# Paired Sample t-Test - Example

**Step 2:** Calculating mean and standard deviation of differences.

Calculating the mean and standard deviation of the differences gives:

$$\bar{d} = 2.05 \text{ and } s_d = 2.837. \text{ Therefore, } SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{2.837}{\sqrt{20}} = 0.634$$

So, we have:

$$t = \frac{2.05}{0.634} = 3.231 \quad \text{on 19 df}$$

**Step 3 :** If the calculated t value is greater than the critical t value, then we reject the null hypothesis.

**So  $3.231 > 2.093$ , null hypothesis is rejected**

**Therefore, there is strong evidence that, on average, module does lead to improvements.**

# CHI-squared Test

- The Chi-square test is intended to test how likely it is that an observed distribution is due to chance.
- It is also called a "**goodness of fit**" statistic, because it measures how well the observed distribution of data fits with the distribution that is expected if the **variables are independent**.

# Chi-squared Test

- A Chi-square test is designed to analyze **categorical data**.
- It will not work with **parametric or continuous** data (such as height in inches).
- For example, if you want to test whether attending class influences how students perform on an exam, using test scores (from 0-100) as data **would not be appropriate for a Chi-square test**.
- However, arranging students into the **categories "Pass" and "Fail"** would.
- Additionally, data in a Chi-square grid should not be in the **form of percentages, or anything other than frequency (count) data**.
- Thus, by dividing a class of 54 into groups according to whether they attended class and whether they passed the exam, you might construct a data set like this:

	Pass	Fail
Attended	25	6
Skipped	8	15

# CHI-squared Test

- Another way to describe Chi-square test is that it tests the null hypothesis that the **variables are independent**.
- The test compares the observed data to a model that distributes the data according to the **expectation** that the **variables are independent**.
- Wherever the observed data **doesn't fit the model**, the likelihood that the variables are **dependent becomes stronger**, thus proving the null hypothesis incorrect!

# CHI-squared Test

- ❑ The **Chi-Square Test of Independence** determines whether there is an association between categorical variables (i.e., whether the variables are independent or related).
- ❑ It is a **nonparametric** test.
- ❑ This test is also known as:
  - **Chi-Square Test of Association.**
- ❑ This test utilizes a **contingency table to analyze the data.**
- ❑ A contingency table (also known as a *cross-tabulation*, *crosstab*, or *two-way table*) is an arrangement in which **data is classified according to two categorical variables.**
- ❑ The **categories for one variable** appear in the **rows**, and the **categories for the other variable** appear in **columns.**
- ❑ Each variable must have **two or more** categories.
- ❑ Each cell reflects the total **count of cases for a specific pair of categories.**



# CHI-squared Test

## Common Uses

- The Chi-Square Test of Independence is commonly used to test the following:
  - Statistical independence or association between two or more categorical variables.
- The Chi-Square Test of Independence can only compare categorical variables.
- It cannot make comparisons between continuous variables or between categorical and continuous variables.
- Additionally, the Chi-Square Test of Independence only assesses associations between categorical variables, and can not provide any inferences about causation.

# CHI-squared Test

## Data Requirements

Your data must meet the following requirements:

1. Two categorical variables.
2. Two or more categories (groups) for each variable.
3. Independence of observations.
  - ☐ There is no relationship between the subjects in each group.
  - ☐ The categorical variables are not "paired" in any way (e.g. pre-test/post-test observations).
4. Relatively large sample size.
  - ☐ Expected frequencies for each cell are at least 1.
  - ☐ Expected frequencies should be at least 5 for the majority (80%) of the cells.

# CHI-squared Test

## Hypotheses

The null hypothesis ( $H_0$ ) and alternative hypothesis ( $H_1$ ) of the Chi-Square Test of Independence can be expressed in **two different but equivalent ways**:

$H_0$ : "[Variable 1] is independent of [Variable 2]"

$H_1$ : "[Variable 1] is not independent of [Variable 2]"

**OR**

$H_0$ : "[Variable 1] is not associated with [Variable 2]"

$H_1$ : "[Variable 1] is associated with [Variable 2]"

# CHI-squared Test

## Test Statistic

The test statistic for the Chi-Square Test of Independence is denoted  $X^2$ , and is computed as:

$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where

$o_{ij}$  is observed cell count in  $i$ th row and  $j$ th column of table

$e_{ij}$  is expected cell count in  $i$ th row and  $j$ th column of the table, computed as

$$e_{ij} = \frac{\text{row } i \text{ total} * \text{col } j \text{ total}}{\text{grand total}}$$

- The quantity  $(o_{ij} - e_{ij})$  is sometimes referred to as the **residual of cell  $(i, j)$ , denoted  $r_{ij}$** .
- The calculated  $X^2$  value is then compared to the critical value from the  **$X^2$  distribution table** with **degrees of freedom  $df = (R - 1)(C - 1)$**  and chosen confidence level.
- **If the calculated  $X^2$  value > critical  $X^2$  value, then we reject the null hypothesis.**

# CHI-squared Test Example

**Example: "Which pet do you prefer?"**

	Cat	Dog
Men	207	282
Women	231	242

By doing the calculations (shown later), we come up with:

- **calculated  $X^2$  value = 4.102**
- **Calculated  $X^2$  value = 4.102 > critical  $X^2$  value = 3.841, then we reject the null hypothesis**
- This result is thought of as being "significant" meaning we think the variables are **not** independent.
- In other words, Gender is linked to Pet Preference.  
**(Men and Women have different preferences for Cats and Dogs)**

# CHI-squared Test Example

## State hypotheses

$H_0$ : "[Variable 1] is independent of [Variable 2]"

Gender and preference for cats or dogs are independent.

$H_1$ : "[Variable 1] is not independent of [Variable 2]"

Gender and preference for cats or dogs are not independent

OR

$H_0$ : "[Variable 1] is not associated with [Variable 2]"

Gender and preference for cats or dogs are independent.

$H_1$ : "[Variable 1] is associated with [Variable 2]"

Gender and preference for cats or dogs are not independent.

# CHI-squared Test Example

Lay the data out in a table

	Cat	Dog
Men	207	282
Women	231	242

Add up rows and columns:

	Cat	Dog	Total
Men	207	282	489
Women	231	242	473
	438	524	962

# CHI-squared Test Example

	Cat	Dog	Total
Men	207	282	489
Women	231	242	473
	438	524	962

Calculate "Expected Value" for each entry:

Multiply each row total by each column total and divide by the overall total:

	Cat	Dog	
Men	$\frac{489 \times 438}{962}$	$\frac{489 \times 524}{962}$	489
Women	$\frac{473 \times 438}{962}$	$\frac{473 \times 524}{962}$	473
	438	524	962

Which gives us:

	Cat	Dog	
Men	222.64	266.36	489
Women	215.36	257.64	473
	438	524	962



# CHI-squared Test Example

Subtract expected from observed, square it, then divide by expected:

	Cat	Dog	
Men	$\frac{(207-222.64)^2}{222.64}$	$\frac{(282-266.36)^2}{266.36}$	489
Women	$\frac{(231-215.36)^2}{215.36}$	$\frac{(242-257.64)^2}{257.64}$	473
	438	524	962

Which gets us:

	Cat	Dog	
Men	1.099	0.918	489
Women	1.136	0.949	473
	438	524	962

Now add up those calculated values:

$$1.099 + 0.918 + 1.136 + 0.949 = 4.102$$

**Calculated X2 Value = 4.102**

# CHI-squared Test Example

## Degrees of Freedom

First we need a "Degree of Freedom"

$$\text{Degree of Freedom} = (\text{rows} - 1) \times (\text{columns} - 1)$$

For our example we have 2 rows and 2 columns:

$$\mathbf{DF} = (2 - 1)(2 - 1) = 1 \times 1 = \mathbf{1}$$

# CHI-squared Test Example

Table of the chi square distribution – Appendix J, p. 915

df	Level of Significance $\alpha$								
	0.200	0.100	0.075	0.050	0.025	0.010	0.005	0.001	0.0005
1	1.642	2.706	3.170	3.841	5.024	6.635	7.879	10.828	12.116
2	3.219	4.605	5.181	5.991	7.378	9.210	10.597	13.816	15.202
3	4.642	6.251	6.905	7.815	9.348	11.345	12.838	16.266	17.731
4	5.989	7.779	8.496	9.488	11.143	13.277	14.860	18.467	19.998
5	7.289	9.236	10.008	11.070	12.833	15.086	16.750	20.516	22.106
6	8.558	10.645	11.466	12.592	14.449	16.812	18.548	22.458	24.104
7	9.803	12.017	12.883	14.067	16.013	18.475	20.278	24.322	26.019
8	11.030	13.362	14.270	15.507	17.535	20.090	21.955	26.125	27.869
9	12.242	14.684	15.631	16.919	19.023	21.666	23.589	27.878	29.667
10	13.442	15.987	16.971	18.307	20.483	23.209	25.188	29.589	31.421
11	14.631	17.275	18.294	19.675	21.920	24.725	26.757	31.265	33.138
12	15.812	18.549	19.602	21.026	23.337	26.217	28.300	32.910	34.822
13	16.985	19.812	20.897	22.362	24.736	27.688	29.820	34.529	36.479

**critical  $X^2$  value for  $df=1$  and  $\alpha =0.05$  is 3.841**

# CHI-squared Test Example

**If the calculated  $X^2$  value  $>$  critical  $X^2$  value, then we reject the null hypothesis.**

- **Calculated  $X^2$  value = 4.102  $>$  critical  $X^2$  value = 3.841, then we reject the null hypothesis**
- This result is thought of as being "significant" meaning we think the variables are not independent
- In other words, Gender is linked to Pet Preference.  
**(Men and Women have different preferences for Cats and Dogs)**

# Fisher's test

- ❑ Fisher's exact test is a statistical significance test used in the analysis of **contingency tables**.
- ❑ Although in practice it is employed when sample sizes are small, it is valid for all **sample sizes**.
- ❑ It is named after its inventor, **Ronald Fisher**, and is one of a class of exact tests.
- ❑ Because significance of the deviation from a null hypothesis (e.g., P-value) can be calculated exactly rather than relying on an approximation that becomes exact in the limit as the sample size grows to infinity.
- ❑ With large samples, a **chi-squared test** (or better yet, a **G-test**) can be used.

# Fisher's test

For example, a sample of teenagers might be divided into male and female on the one hand, and those that are and are not currently studying for a statistics exam on the other.

	Men	Women	Row total
Studying	1	9	10
Not-studying	11	3	14
Column total	12	12	24

**Notations:** represent cells by letters  $a$ ,  $b$ ,  $c$  and  $d$ , call totals across rows and columns *marginal totals*, and represent the grand total by  $n$

	Men	Women	Row Total
Studying	$a$	$b$	$a + b$
Non-studying	$c$	$d$	$c + d$
Column Total	$a + c$	$b + d$	$a + b + c + d (=n)$

# Fisher's test

- Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution:

$$p = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b} \binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! n!}$$

where  $\binom{n}{k}$  is the binomial coefficient and the symbol

! indicates the factorial operator.

# Fisher's test

With the following data above, this gives

	Men	Women	Row total
Studying	1	9	10
Not-studying	11	3	14
Column total	12	12	24

$$p = \binom{10}{1} \binom{14}{11} / \binom{24}{12} = \frac{10! 14! 12! 12!}{1! 9! 11! 3! 24!} \approx 0.001346076$$

- ❑ This gives exact hypergeometric probability of observing this particular arrangement of the data, assuming the given marginal totals, on the null hypothesis that **men and women are equally likely to be studiers**.
- ❑ The smaller the value of  $p$ , greater the evidence for rejecting the null hypothesis;
- ❑ Here the evidence is strong that men and women are not equally likely to be studiers.



# Fisher's test

- We can calculate the **exact probability** of any arrangement of the 24 teenagers into four cells of the table.
- Fisher showed that to generate a significance level, we need consider only the cases where the **marginal totals are the same** as in the observed table,
- Among those, only the cases where the arrangement is as extreme as the observed arrangement.
- In the example, there are 11 such cases.
- Of these only one is more extreme in same direction as our data; it looks like this:

	Men	Women	Row Total
Studying	0	10	10
Non-studying	12	2	14
Column Total	12	12	24

For this table (with extremely unequal studying proportions) the probability is  $p = \frac{\binom{10}{0} \binom{14}{12}}{\binom{24}{12}} \approx 0.000033652$ .

# Fisher's test

- To calculate the significance of the observed data, i.e. the total probability of observing data as extreme or more extreme if the **null hypothesis** is true, we have to calculate the values of  $p$  for both these tables, and add them together.
- This gives a **one-tailed test**, with  $p$  approximately  $0.001346076 + 0.000033652 = 0.001379728$ .
- This value can be interpreted as sum of evidence provided by observed data—or any more extreme table—for the null hypothesis.
- **Smaller the value of  $p$ , the greater the evidence for rejecting the null hypothesis.**

# Fisher's test

- For a **two-tailed test** we must also consider tables that are equally **extreme**, but in the opposite direction.
- Unfortunately, classification of the tables according to whether or not they are '**as extreme**' is **problematic**.
- An approach used by **fisher.test function in R** is to compute the p-value by summing the probabilities for all tables with probabilities less than or equal to that of the **observed table**.
- In example here, the 2-sided p-value is twice the 1-sided value.
- In general these can differ substantially for tables with small counts.

**Discussion applicable to 2x2 contingency tables, Need technique applicable to table of any size.**

# Fisher's test

Let there exist two such variables  $X$  and  $Y$ , with  $m$  and  $n$  observed states, respectively. Now form an  $m \times n$  matrix in which the entries  $a_{ij}$  represent the number of observations in which  $x = i$  and  $y = j$ .

Calculate the row and column sums  $R_i$  and  $C_j$ , respectively, and the total sum of the matrix.

$$N = \sum_i R_i = \sum_j C_j$$

Then calculate the conditional probability of getting the actual matrix given the particular row and column sums, given by

$$P_{\text{cutoff}} = \frac{(R_1! R_2! \cdots R_m!)(C_1! C_2! \cdots C_n!)}{N! \prod_{i,j} a_{ij}!},$$

# Fisher's test

- Now find all possible matrices of nonnegative integers consistent with the row and column sums  $R_i$  and  $C_j$ .
- For each one, calculate the associated conditional probability
- The sum of these probabilities must be 1.
- The test is most commonly applied to  $2 \times 2$  matrices, and is computationally unmanageable for large  $m$  or  $n$ .
- In case of  $2 \times 2$  matrix, P-value of test can be simply computed by sum of all P-values which are  $\leq P_{\text{cutoff}}$

# Fisher's test

- For an example application of the 2x2 test
- Let X be a journal, say either Mathematics Magazine or Science,
- Let Y be the number of articles on the topics of mathematics and biology appearing in a given issue of one of these journals.
- If Mathematics Magazine has five articles on math and one on biology, and Science has none on math and four on biology, then the relevant matrix would be

	Math. Mag.	Science	
math	5	0	$R_1 = 5$
biology	1	4	$R_2 = 5$
	$C_1 = 6$	$C_2 = 4$	$N = 10.$

# Fisher's test

	Math. Mag.	Science	
math	5	0	$R_1 = 5$
biology	1	4	$R_2 = 5$
	$C_1 = 6$	$C_2 = 4$	$N = 10.$

Computing  $P_{\text{cutoff}}$  gives

$$P_{\text{cutoff}} = \frac{5!^2 6! 4!}{10! (5! 0! 1! 4!)} = 0.0238,$$

and the other possible matrices and their  $P$ s are

$$\begin{array}{ll} \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} P = 0.2381 & \begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix} P = 0.2381 \\ \begin{bmatrix} 3 & 2 \\ 3 & 2 \end{bmatrix} P = 0.4762 & \begin{bmatrix} 1 & 4 \\ 5 & 0 \end{bmatrix} P = 0.0238, \end{array}$$

The sum of  $P$ -values less than or equal to  $P_{\text{cutoff}} = 0.0238$  is then 0.0476 is less than 0.05

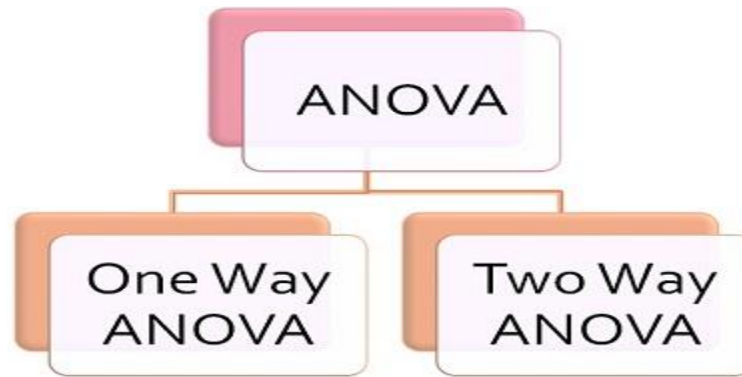
Therefore, in this case, there would be a statistically significant association between the journal and type of article appearing.

# Analysis of Variance (ANOVA)

- Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample.
- ANOVA was developed by the statistician Ronald Fisher.
- In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means.
- ANOVA is a form of statistical hypothesis testing heavily used in the analysis of experimental data.
- Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.



# ANOVA Types



The key differences between one-way and two-way ANOVA are summarized clearly below.

1. A one-way ANOVA is primarily designed to enable the equality testing between three or more means. A two-way ANOVA is designed to assess the interrelationship of two independent variables on a dependent variable.
2. A one-way ANOVA only involves one factor or independent variable, whereas there are two independent variables in a two-way ANOVA.
3. In a one-way ANOVA, the one factor or independent variable analyzed has three or more categorical groups. A two-way ANOVA instead compares multiple groups of two factors.

# One-Way vs Two-Way ANOVA Differences Chart

	One-Way ANOVA	Two-Way ANOVA
<b>Definition</b>	A test that allows one to make comparisons between the means of three or more groups of data.	A test that allows one to make comparisons between the means of three or more groups of data, where two independent variables are considered.
<b>Number of Independent Variables</b>	One.	Two.
<b>What is Being Compared?</b>	The means of three or more groups of an independent variable on a dependent variable.	The effect of multiple groups of two independent variables on a dependent variable and on each other.
<b>Number of Groups of Samples</b>	Three or more.	Each variable should have multiple samples.

# One Way ANOVA

- One-Way ANOVA ("analysis of variance") compares the means of two or more independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different.
- One-Way ANOVA is a parametric test.
- This test is also known as:
  - ✓ One-Factor ANOVA
  - ✓ One-Way Analysis of Variance
  - ✓ Between Subjects ANOVA
- The variables used in this test are known as:
  - ❑ Dependent variable
  - ❑ Independent variable (also known as grouping variable, or *factor*)
    - ✓ This variable divides cases into two or more mutually exclusive *levels*, or groups

# One Way ANOVA

## Data Requirements

1. Dependent variable that is continuous (i.e., interval or ratio level)
2. Independent variable that is categorical (i.e., two or more groups)
3. Cases that have values on both the dependent and independent variables
4. Independent samples/groups (i.e., independence of observations)
5. Random sample of data from the population
6. Normal distribution (approximately) of the dependent variable
7. Homogeneity of variances (i.e., variances approximately equal across groups)
8. No outliers

# One Way ANOVA

## Hypotheses

The null and alternative hypotheses of one-way ANOVA can be expressed as:

$H_0$ :  $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$  ("all  $k$  population means are equal")

$H_1$ : At least one  $\mu_i$  different ("at least one of  $k$  population means is not equal to others")

*where*

- $\mu_i$  is the population *mean* of the  $i^{\text{th}}$  group ( $i = 1, 2, \dots, k$ )

# One Way ANOVA

## KEY TAKEAWAYS

### Key Points

- The F-test is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled.
- Perhaps the most common F-test is that which tests the hypothesis that the means and standard deviations of several populations are equal. (Note that all populations involved must be assumed to be normally distributed.)
- The F-test is sensitive to non-normality.
- The F-distribution is skewed to the right, but as the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.

### Key Terms

- **ANOVA:** Analysis of variance—a collection of statistical models used to analyze the differences between group means and their associated procedures (such as “variation” among and between groups).
- **Type I error:** Rejecting the null hypothesis when the null hypothesis is true.
- **F-Test:** A statistical test using the F-distribution, most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled.

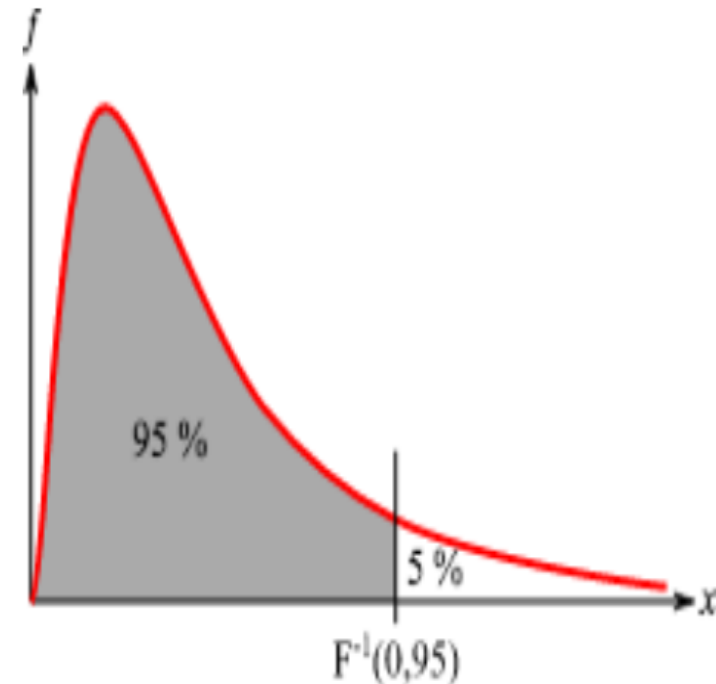
# One Way ANOVA

## The F-Distribution

The F-distribution exhibits the following properties, as illustrated in the above graph:

1. The curve is not symmetrical but is skewed to the right.
2. There is a different curve for each set of degrees of freedom.
3. The F-statistic is greater than or equal to zero.
4. As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.

The F-statistic also has a common table of values, as do z-scores and t-scores.



**F-distribution:** The F-distribution is skewed to the right and begins at the x-axis, meaning that F-values are always positive.

# One Way ANOVA

- The F test as a one-way analysis of variance is used to assess whether the expected values of a quantitative variable within several **pre-defined groups differ from each other**.
- For example, suppose that a medical trial compares four treatments. ANOVA F-test can be used to assess whether
  - ❑ any of treatments is on average superior, or inferior, to the others versus
  - ❑ the null hypothesis that all four treatments yield same mean response.



# One Way ANOVA

- Can carry out pairwise tests among the groups/treatments
- **Advantage of the ANOVA F-test is**
  - do not need to pre-specify which treatments are to be compared, and
  - do not need to adjust for making multiple comparisons.
- The **disadvantage of the ANOVA FF-test is**
  - that if we reject the null hypothesis, we do not know **which treatments** can be said to be **significantly different** from the others.
  - If the F-test is performed at level  $\alpha$  we cannot state that **treatment pair with the greatest mean difference** is significantly different at level  $\alpha$ .

# One Way ANOVA

The formula for the one-way ANOVA F-test statistic is:

$$F = \frac{\text{explained variance}}{\text{unexplained variance}} \quad \text{or} \quad F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

The “explained variance,” or “between-group variability” is:

$$\sum_i \frac{n_i (\bar{Y}_i - \bar{Y})^2}{(K - 1)} \quad \text{where } \bar{Y}_i \text{ denotes sample mean in the } i^{\text{th}} \text{ group,}$$

$n_i$  is number of observations in the  $i^{\text{th}}$  group  
 $\bar{Y}$  denote overall mean of the data  
 $K$  denotes the number of groups

The “unexplained variance”, or “within-group variability” is:

$$\sum_{ij} \frac{(\bar{Y}_{ij} - \bar{Y}_i)^2}{(N - K)} \quad \text{where } \bar{Y}_{ij} \text{ is the } j^{\text{th}} \text{ observation in } i^{\text{th}} \text{ out of } K \text{ groups}$$

$N$  is the overall sample size

This F-statistic follows the F– distribution with  $K - 1, N - K$  degrees of freedom under the null hypothesis.

# One Way ANOVA

- The statistic will be large if the between-group variability is large relative to the within-group variability
- This is unlikely to happen if the population means of the groups all have the same value.
- Note that when there are only two groups for the one-way ANOVA F-test,  $F=t^2$  where  $t$  is Student's t-statistic.

# One Way ANOVA Example

Four sororities took a random sample of sisters regarding their grade means for the past term.

The data were distributed as follows:

- Sorority 1: 2.17, 1.85, 2.83, 1.69, 3.33
- Sorority 2: 2.63, 1.77, 3.25, 1.86, 2.21
- Sorority 3: 2.63, 3.78, 4.00, 2.55, 2.45
- Sorority 4: 3.79, 3.45, 3.08, 2.26, 3.18

Using a significance level of 1%, **is there a difference in mean grades among the sororities?**

# One Way ANOVA Example

## Solution

- Let  $\mu_1, \mu_2, \mu_3, \mu_4$  be population means of the sororities.
- Null hypothesis claims that sorority groups are from same normal distribution.
- Alternate hypothesis says that at least two of the sorority groups come from populations with different normal distributions.
- Sample size for each group is 5 - example of a balanced design, since each factor (i.e., sorority) has same number of observations.

# One Way ANOVA Example

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$H_a$ : Not all of means  $\mu_1, \mu_2, \mu_3, \mu_4$  are equal

Distribution for test:  $F_{3,16}$  where  $k=4$  groups and  $n=20$  samples in total

$$df_{\text{numerator}} = k - 1 = 4 - 1 = 3$$

$$df_{\text{denominator}} = n - k = 20 - 4 = 16$$

Calculate the test statistic:  **$F=2.23$**

**F statistic critical value** for  $F_{3,16}$  is 9.01 (F Distribution Table)

Calculated F-statistic < **F statistic critical value**

**Reject null hypothesis**

# One Way ANOVA Example

## F critical values

		Degrees of freedom in the numerator								
<i>p</i>		1	2	3	4	5	6	7	8	9
16	.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
	.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05
	.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
	.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98
17	.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
	.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
	.010	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68
	.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75

# One Way ANOVA Example

- **P-Value:** P-Value is a statistical test that determines the probability of extreme results of the statistical hypothesis test, taking the Null Hypothesis to be correct.
- A p value is used in hypothesis testing to help you support or reject the null hypothesis.
- The p value is the evidence against a **null hypothesis**.
- The smaller the p-value, the stronger the evidence that you **should reject the null hypothesis**.

**P-Value from F-Ratio Calculator (ANOVA)**

F-ratio value:	<input type="text" value="2.23"/>
DF- numerator:	<input type="text" value="3"/>
DF- denominator:	<input type="text" value="16"/>

Significance Level:

☒ .01   ☐ .05   ☐ .10

The p-value is .124181. The result is *not* significant at  $p < .01$ .

<https://www.socscistatistics.com/pvalues/fdistribution.aspx>



# One Way ANOVA Example

Probability statement:  $p\text{-value} = P(F > 2.23) = 0.1241$

Compare  $\alpha$  and the  $p\text{-value}$ :  $\alpha = 0.01$ ,  $p\text{-value} = 0.1241$

**Make a decision:**

Since  $\alpha < p\text{-value}$ , you **cannot reject  $H_0$**

**Conclusion:**

**There is not sufficient evidence to conclude that there is a difference among the mean grades for the sororities.**

# Two Way ANOVA

- A biologist wants to compare mean growth for **three different levels of fertilizer**.
- Suppose the biologist wants to ask this same question but with **two different species of plants while still testing the three different levels of fertilizer**.
- The biologist needs to investigate not only the average growth between the two species (main effect A) and the average growth for the three levels of fertilizer (main effect B), but also the **interaction or relationship between the two factors of species and fertilizer**.
- Two-way analysis of variance allows the biologist to answer the question about **growth affected by species and levels of fertilizer**, and to account for the variation due to both factors simultaneously.

# Two Way ANOVA

- We now consider analysis in which two factors can explain variability in the response variable.
- Remember that we can deal with factors by controlling them, by fixing them at specific levels, and randomly applying the treatments so the effect of uncontrolled variables on the response variable is minimized.
- With two factors, we need a **factorial experiment**.

		Factor B (fertilizer)		
		Level 1	Level 2	Level 3
Factor A (species)	Species 1	1.2, 2.4, 2.6, 2.2	2.4, 2.7, 2.7, 2.9	3.1, 3.0, 3.2, 3.4
	Species 2	0.6, 0.9, 1.0, 0.9	2.1, 2.3, 2.0, 1.9	0.7, 0.5, 0.6, 0.5

- These six combinations are referred to as **treatments**.
- The experiment is called a 2 x 3 factorial experiment.

# Two Way ANOVA

## Notation

$k$  = number of levels of factor A

$l$  = number of levels of factor B

$kl$  = number of treatments (each one a combination of a factor A level and a factor B level)

$m$  = number of observations on each treatment

# Two Way ANOVA

We now partition the variation even more to reflect the main effects (Factor A and Factor B) and the interaction term:

$$SSTo = SSA + SSB + SSAB + SSE$$

Where

**SSTo** is total sums of squares, with associated degrees of freedom  $klm - 1$

**SSA** is factor A main effect sums of squares, with associated degrees of freedom  $k - 1$

**SSB** is factor B main effect sums of squares, with associated degrees of freedom  $l - 1$

**SSAB** is interaction sum of squares, with associated degrees of freedom  $(k - 1)(l - 1)$

**SSE** is error sum of squares, with associated degrees of freedom  $kl(m - 1)$

# Two Way ANOVA

null and alternative hypotheses for a two-way ANOVA.

1. H0: There is no interaction between factors  
H1: There is a significant interaction between factors
2. H0: There is no effect of Factor A on the response variable  
H1: There is an effect of Factor A on the response variable
3. H0: There is no effect of Factor B on the response variable  
H1: There is an effect of Factor B on the response variable

# Two Way ANOVA

## Two-way ANOVA table:

Source of variation	df	Sums of squares	Mean square	F
Factor A	$k - 1$	SSA	$MSA = \frac{SSA}{k - 1}$	$F_A = \frac{MSA}{MSE}$
Factor B	$l - 1$	SSB	$MSB = \frac{SSB}{l - 1}$	$F_B = \frac{MSB}{MSE}$
Interaction AB	$(k - 1)(l - 1)$	SSAB	$MSAB = \frac{SSAB}{(k - 1)(l - 1)}$	$F_{AB} = \frac{MSAB}{MSE}$
Error	$kl(m - 1)$	SSE	$MSE = \frac{SSE}{kl(m - 1)}$	
Total	$klm - 1$	SSTo		

If p-value is smaller than  $\alpha$  (level of significance), you will reject the null hypothesis.

# Two Way ANOVA

- If the p-value is smaller than  $\alpha$  (level of significance), you will reject the null hypothesis.
- When we conduct a two-way ANOVA, we always first test the hypothesis regarding the interaction effect.
- If the null hypothesis of no interaction is rejected, we do NOT interpret the results of the hypotheses involving the main effects.
- If the interaction term is NOT significant, then we examine the two main effects separately.



***Thank You !!!***

# Two Way ANOVA

<https://courses.lumenlearning.com/suny-natural-resources-biometrics/chapter/chapter-6-two-way-analysis-of-variance/>

# Two Way ANOVA

<https://courses.lumenlearning.com/suny-natural-resources-biometrics/chapter/chapter-6-two-way-analysis-of-variance/>

# One way – two way annova

<https://www.technologynetworks.com/informatics/articles/one-way-vs-two-way-anova-definition-differences-assumptions-and-hypotheses-306553#:~:text=A%20one%2Dway%20ANOVA%20only,in%20a%20two%2Dway%20ANOVA.&text=In%20a%20one%2Dway%20ANOVA%2C%20the%20one%20factor%20or%20independent,multiple%20groups%20of%20two%20factors.>

<https://courses.lumenlearning.com/boundless-statistics/chapter/one-way-anova/>

<https://libguides.library.kent.edu/SPSS/OneWayANOVA>

# Fisher's exact test

<https://mathworld.wolfram.com/FishersExactTest.html#:~:text=Fisher's%20exact%20test%20is%20a,associations%20between%20two%20categorical%20variables.&text=column%20sums%20and-,,these%20probabilities%20must%20be%201.>

[https://en.wikipedia.org/wiki/Fisher%27s\\_exact\\_test](https://en.wikipedia.org/wiki/Fisher%27s_exact_test)

<https://online.stat.psu.edu/stat504/node/89/>

# CHI-squared Test Example

<https://www.mathsisfun.com/data/chi-square-test.html>

# CHI-squared Test

<https://libguides.library.kent.edu/SPSS/ChiSquare>

<https://www.ling.upenn.edu/~clight/chisquared.htm#:~:text=The%20Chi%2Dsquare%20test%20is,if%20the%20variables%20are%20independent.>

# T-Distribution Table & t-Value Calculator Online

T-Distribution Table

<https://www.statisticshowto.com/tables/t-distribution-table/>

Student t-Value Calculator Online

<http://www.ttable.org/student-t-value-calculator.html>



IMP

<https://libguides.library.kent.edu/SPSS/OneSampletTest>

<https://libguides.library.kent.edu/SPSS/IndependentTTest>

Good Example

<https://www.statisticshowto.com/one-sample-t-test/>

## T-test

A t-test is used to compare the mean of two given samples. Like a z-test, a t-test also assumes a normal distribution of the sample. A t-test is used when the population parameters (mean and standard deviation) are not known.

There are three versions of t-test

- 1. Independent samples t-test which compares mean for two groups*
- 2. Paired sample t-test which compares means from the same group at different times*
- 3. One sample t-test which tests the mean of a single group against a known mean.*

<https://towardsdatascience.com/statistical-tests-when-to-use-which-704557554740>

[https://www.sagepub.com/sites/default/files/upm-binaries/40007\\_Chapter8.pdf](https://www.sagepub.com/sites/default/files/upm-binaries/40007_Chapter8.pdf)

[https://us.sagepub.com/sites/default/files/upm-ssets/98047\\_book\\_item\\_98047.pdf](https://us.sagepub.com/sites/default/files/upm-ssets/98047_book_item_98047.pdf)

[https://2012books.lardbucket.org/pdfs/beginning-statistics/s12-testing\\_hypotheses.pdf](https://2012books.lardbucket.org/pdfs/beginning-statistics/s12-testing_hypotheses.pdf)

<https://crumplab.github.io/statistics/t-tests.html>

<https://www.statisticssolutions.com/manova-analysis-one-sample-t-test/>

<https://towardsdatascience.com/statistical-tests-when-to-use-which-704557554740>

A statistical hypothesis is an expectation about a population. Usually it is formulated as a claim that a population parameter takes a particular value or falls within a specific range of values. On the basis of information from a sample we assess if a hypothesis makes sense or not. The significance test is, just like the confidence interval, a method of **inferential statistics**. Each significance test is based on two hypotheses: the **null hypothesis** and the **alternative hypothesis**. If you do a significance test, you assume that the null hypothesis is true unless your data provide strong evidence against it.

<https://stattrek.com/probability-distributions/probability-distribution.aspx>

<https://statisticsbyjim.com/basics/probability-distributions/#:~:text=A%20probability%20distribution%20is%20a,on%20the%20underlying%20probability%20distribution.>

## Supervised Learning

- Humans learn from past experiences, machines learn from past instances!

5:02



4:54 / 12:21



## Questions to ask in Supervised Learning

- **Training phase:**
  - What are the features? How do you represent them?
  - What is the classification model / algorithm?
  - What are the model parameters?
- **Inference phase:**
  - What is the expected performance? What is a good measure?