

BIOSTATISTICS WITH

SUMMER WORKSHOP
(MITGEST network)

24-27 JULY 2024

Maria Chiara Mimmi, PhD

WORKSHOP SCHEDULE

- 4 days
 - 1. Intro to R and data analysis
 - 2. Statistical inference & hypothesis testing
 - 3. Modeling correlation and regression
 - 4. Intro to Machine Learning
- Each day will include:
 - Frontal class (MORNING)
 - Practical training with R about the topics discussed in the morning. (AFTERNOON)

DAY 3 – LECTURE OUTLINE

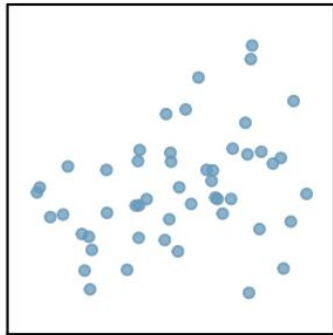
- Testing for a **correlation** hypothesis (relationship of variables)
 - Pearson rho analysis (param)
 - Spearman test (no param)
- Measures of **association**
 - Chi-Square Test of Independence
 - example
 - Fisher's Exact Test
- From correlation/association to **causation**
 - introduction to experiments
 - Example: Linear regression models
 - Example: Multiple Linear Regression
- From causation to **prediction**
 - introduction to Machine Learning
 - Supervised algorithms
 - Unsupervised algorithms

Measures of correlation

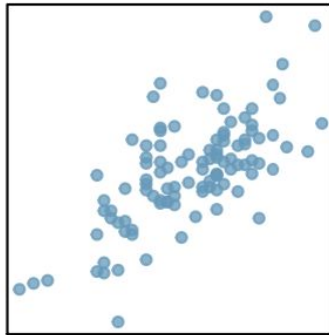
Defining correlation

- **Correlation** is a numerical summary statistic that measures the *strength of a linear relationship* between two variables
 - denoted by **r** (correlation coefficient) which takes values **between -1 and 1**

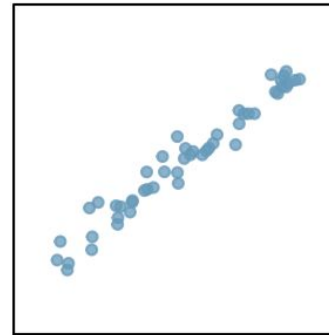
positive
correlation



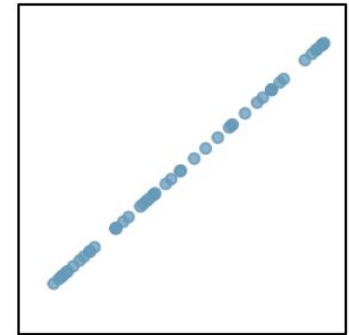
$R = 0.33$



$R = 0.69$

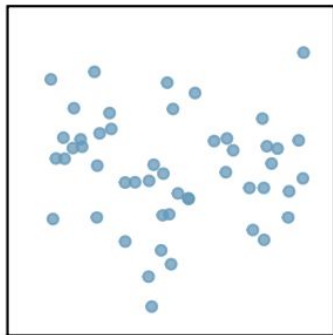


$R = 0.98$

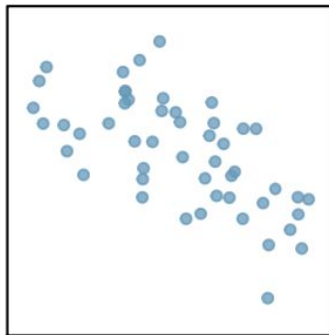


$R = 1.00$

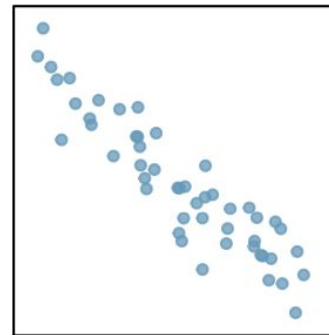
negative
correlation



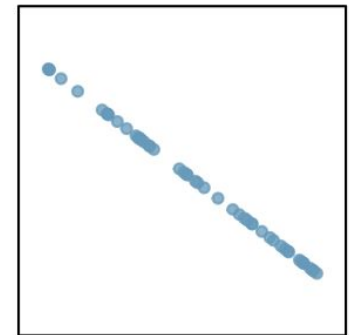
$R = -0.08$



$R = -0.64$



$R = -0.92$



$R = -1.00$

Source: Vu, J., & Harrington, D. (2021). *Introductory Statistics for the Life and Biomedical Sciences*. Retrieved from <https://www.openintro.org/book/biostat/>

Most used measures of correlation

Correlation coefficient	Type of relationship	Levels of measurement	Data distribution
Pearson's r (for population)	Linear	Two quantitative (interval or ratio) variables	Normal distribution
Spearman's (for population)	Non-linear	Two ordinal, interval or ratio variables	Any distribution
Cramér's V (Cramér's ϕ)	Non-linear	Two nominal variables	Any distribution
Kendall's (τ)	Non-linear	Two ordinal, interval or ratio variables	Any distribution

Correlation and covariance

For two quantitative variables, the basic statistics of interest are the sample covariance and/or sample correlation, which correspond to and are estimates of the corresponding population parameters.

The sample **covariance is a measure of how much two variables “co-vary”**, i.e., how much (and in what direction) should we expect one variable to change when the other changes. Positive covariance values suggest variables change in the same direction. Negative covariances suggest variables change in the opposite direction. And covariances near zero suggest that the two variables vary independently of each other.

Covariances tend to be hard to interpret, so we often use **correlation** instead. The correlation has the nice property that it is always between -1 and +1, with -1 being a “perfect” negative linear correlation, +1 being a perfect positive linear correlation and 0 indicating that X and Y are uncorrelated.

The symbol r or $r_{x,y}$ is often used for sample correlations.

The general formula for sample covariance is

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

It is worth noting that $\text{Cov}(X, X) = \text{Var}(X)$.

The formula for the sample correlation is

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{s_x s_y}$$

where s_x is the standard deviation of X and s_y is the standard deviation of Y .

$$\text{Variance} = \sigma^2 = \sum (x_i - \mu)^2 / N$$

	X	Y	Z
X	5.00	1.77	-2.24
Y	1.77	7.0	3.17
Z	-2.24	3.17	4.0

Table 4.5: A Covariance Matrix

	X	Y	Z
X	1.0	0.3	-0.5
Y	0.3	1.0	0.6
Z	-0.5	0.6	1.0

Table 4.6: A Correlation Matrix

The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

Correlation between 2 numerical variables

Pearson's correlation (parametric test)

Pearson's correlation

Pearson correlation (r) measures a linear association between 2 CONTINUOUS variables (x and y) or 2 dichotomous variables

- It's also known as a parametric correlation test because it depends to the distribution of the data.
- The Pearson correlation evaluates the linear relationship between two continuous variables.

FORMULA

WHERE:

x and y are two vectors of length n

\bar{x} and \bar{y} correspond to the means of x and y , respectively.

We can test the statistical significance of the correlation statistic as well.

The p-value (significance level) of the correlation can be determined by calculating

with

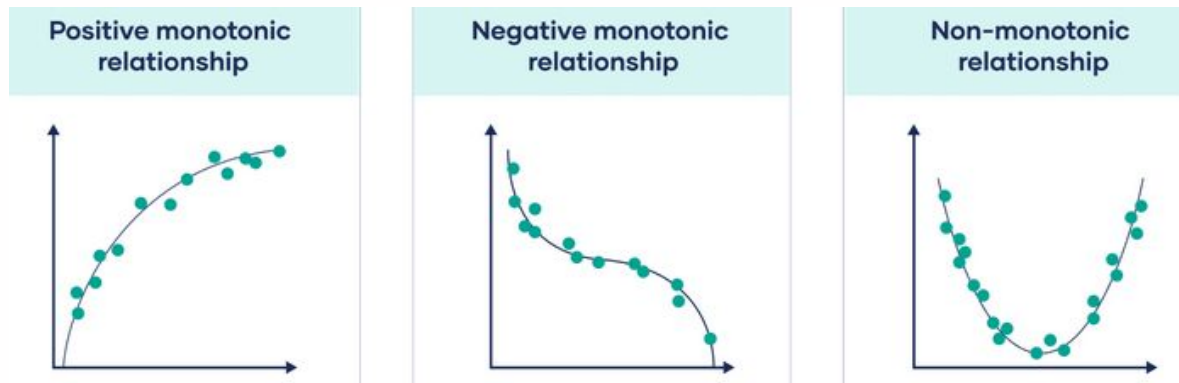
Correlation between 2 numerical variables

Spearman's correlation (non parametric test)

Spearman's rank order correlation coefficient

Spearman's correlation (or) is a **nonparametric alternative to Pearson's correlation**, used for

- continuous data with a **non linear, monotonic** relationships, or
- **ordinal** data (e.g. Likert scale survey questions: *strongly agree, agree, etc.*)



FORMULA

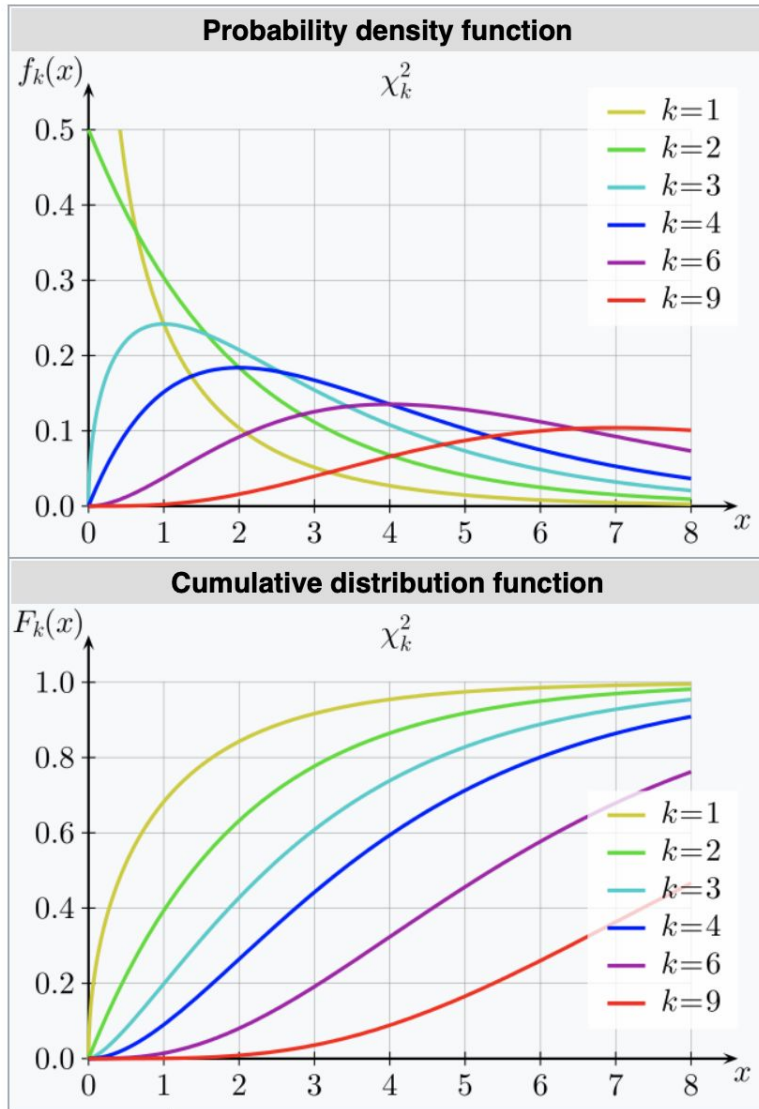
- where:
 - is Spearman's coefficient of rank correlation.
 - is the difference between the ranks for each pair.
 - is the number of paired observations.

Hypothesis Test: Rank Correlation

Chi Squared Distributions

The chi-squared distribution

chi-squared



- The **chi-squared distribution** () is a family of continuous probability distributions
- It results from the sum of squares of k normally distributed random variables, where k is the number of degrees of freedom (df)
- The **mean** is equal to the df and the **variance** is equal to $2 \times df$

Notation	$\chi^2(k)$ or χ_k^2
Parameters	$k \in \mathbb{N}^*$ (known as "degrees of freedom")
Support	$x \in (0, +\infty)$ if $k = 1$, otherwise $x \in [0, +\infty)$
PDF	$\frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} e^{-x/2}$
CDF	$\frac{1}{\Gamma(k/2)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$
Mean	k
Median	$\approx k \left(1 - \frac{2}{9k}\right)^3$
Mode	$\max(k - 2, 0)$
Variance	$2k$

Applications of the Chi-Square Test

- Unlike the **Normal distribution**, very few real-world observations follow a **chi-square distribution**, but it is used extensively in hypothesis testing (also due to its close relationship with the normal).
 - As **k** increases, the distribution looks more and more similar to a normal distribution
- The **Chi-square test** helps to answer the following questions:
 - 1. Independence test**
 - Are two categorical variables independent of each other?
 - for example, does gender have an impact on whether a person has a Netflix subscription or not?
 - 2. Distribution (or Goodness of fit) test**
 - Are the observed values of two categorical variables equal to the expected values?
 - One question could be, is one of the three video streaming services Netflix, Amazon, and Disney subscribed to above average?
 - 3. Homogeneity test**
 - Are two or more samples from the same population?
 - One question could be whether the subscription frequencies of the three video streaming services Netflix, Amazon and Disney differ in different age groups.

Correlation between 2 categorical variables

Chi Squared test of independence

A useful tool for categorical variables: contingency tables

- A **contingency table** summarizes data for 2 categorical variables (each value in the table representing the times a particular combination of outcomes occurs)
- Below we see 2 categorical variables “**gender**” (male, female) and “**has Netflix subscription**” (yes, no)

Frequency			SUM
	Male	Female	
Netflix yes	10	13	23
Netflix no	15	14	29
SUM	25	27	

- The **row totals** (counts across each row) and the **column totals** (counts across each column) are the **marginal totals**
- Frequencies can also be shown as proportions

Computing the Chi-Square Test of Independence

- E.g. suppose we are testing the independence of the two categorical variables “**gender**” (male, female) and “**has Netflix subscription**” (yes, no)
- The test performs a **comparison** of these two contingency tables:

<u>Observed</u> Frequency		
	Male	Female
Netflix yes	10	13
Netflix no	15	14

<u>Expected</u> Frequency		
	Male	Female
Netflix yes	$(23 \times 25) / 52 = 11.06$	$(23 \times 27) / 52 = 11.94$
Netflix no	$(29 \times 25) / 52 = 13.94$	$(29 \times 27) / 52 = 15.06$

IMPORTANT ASSUMPTIONS TO NOTICE:

- The assumption for the **chi-squared ()** test statistic is that the expected frequencies per cell are > 5
- The **chi-squared ()** test uses only the categories but NOT rankings

Computing the Chi-Square Test of Independence (computation)

- Let's compute the example for the two variables “gender” and “has Netflix subscription”

Observed Frequency		
	Male	Female
Netflix yes	10	13
Netflix no	15	14

Expected Frequency		
	Male	Female
Netflix yes	$(23 \times 25) / 52 = 11.06$	$(23 \times 27) / 52 = 11.94$
Netflix no	$(29 \times 25) / 52 = 13.94$	$(29 \times 27) / 52 = 15.06$

- The **chi-squared ()** test statistic is calculated via:
- where:
 - observed frequency and
 - Expected frequency**
 - calculated for each cell in the contingency table
- The test assumptions are:
 - H_0 : (null hypothesis) The two variables are independent.
 - H_1 : (alternative hypothesis) The two variables are not independent. (i.e. they are associated)
 -

Interpreting the Chi-Square Test of Independence

- The **chi-squared ()** test statistic calculated value:
- BY THE CRITICAL REGION: Looking at the distribution, for a significance level of 5% and a *df* of 1, the **critical chi-squared value = 3.841**
 - → Since the **calculated chi-squared value=0.35** is smaller, we **FAIL TO REJECT the null** (H_0 : *The two categorical variables are independent*)
- BY THE p VALUE: Also, the **p-value** associated to the and is **0.5541**.
 - → Since this p-value is not less than 0.05, we fail to reject the null hypothesis.
- This means we do not have sufficient evidence to say that there is an association between gender and political having a Netflix account!

Chi Squared test (another application)

Goodness of Fit Test for one categorical variable

Chi-Square Goodness of Fit Test

- GOAL: a **Chi-Square goodness of fit test** is used to **determine whether or not a categorical variable follows a *hypothesized distribution***.
 - With **high** goodness of fit, the values expected based on the model are **close to** the observed values
 - With **low** goodness of fit, the values expected based on the model are **far from** the observed values
- EXAMPLES OF APPLICATION:
 - *Is this sample drawn from a population with 90% right-handed and 10% left-handed people?*
 - *Do offspring have with an equal probability of inheriting all possible genotypic combinations (i.e., unlinked genes)?*
- HYPOTHESIS FORMULATION
 - Null Hypothesis (H_0): The population follows the specified distribution.
 - Alternative Hypothesis (H_a): The population does not follow the specified distribution.

Chi-Square Goodness of Fit Test (computation)

- FORMULA: The formula is essentially the same as in the independence test
- where O_i = Observed Frequencies and E_i = Expected Frequencies
- ... with k (number of groups minus 1)
- WHEN SHOULD WE USE IT? (assumptions)
 1. We are testing the distribution of **one categorical variable**
 - if you have a continuous variable, it should be converted to categorical (this is called *data binning*) or a different test can be used (like the Kolmogorov–Smirnov goodness of fit test for continuous variables)
 2. The sample was randomly selected from the population.
 3. There are a minimum of 5 observations expected in each group.

Chi-Square Goodness of Fit Test (example)

- GOAL: examine the appropriateness of hypothesized distribution for a dataset
- CASE: In the FAMuSS study (we'll see later in the lab) volunteers were observed at a university, so we test if their distribution by categorical variable **race** is the same as (i.e. *representative of*) the general US population?

Race	African.American	Asian	Caucasian	Other	Total
FAMuSS (Observed)	27	55	467	46	595
US Census (Expected)	76.16	5.95	478.38	34.51	595

- where O_i = Observed Frequencies and E_i = Expected Frequencies
- ... with k (number of groups minus 1)
- The χ^2 statistic is extremely large, and the associated p-value $< 0.001 \rightarrow$
- We **reject the null hypothesis** (= the sample proportions should equal the population proportions)... in fact, we can see for example the higher Asian representation in sample

Correlation between 2 categorical variables - Fisher's Exact Test

(alternative to the Chi-Square Test of
Independence)

Fisher's Exact Test

- Fisher's Exact Test is used to determine whether or not there is a significant association between two categorical variables.
- It is typically used as an alternative to the Chi-Square Test of Independence when one or more of the cell counts in a 2×2 table is less than 5.
- Fisher's Exact Test uses the following null and alternative hypotheses:
 - H_0 : (null hypothesis) The two variables are independent.
 - H_1 : (alternative hypothesis) The two variables are not independent.

Calculate **effect size** after a Chi-Square Test

3 alternatives to assess “strength” of the association (if any)

Three Ways to Calculate Effect Size for a Chi-Square Test

- So, we have seen 2 commonly used **Chi-Square tests**:
 - **Chi-Square Test for Independence**: Used to determine whether or not there is a significant association between two categorical variables from a single population.
 - **Chi-Square Test for Goodness of Fit**: Used to determine whether or not a categorical variable follows a hypothesized distribution
- For both of these tests, we obtain a **p-value** that tells us “if” an association is found (i.e. we should reject the null hypothesis of the test or not).
- Then, we may wonder about the **effect size** of the test (i.e. “how strong” an association is)
- There are 3 ways to measure **effect size**:
 1. **Phi (ϕ)**
 - for 2 x 2 contingency table
 2. **odds ratio (OR)**
 - for 2 x 2 contingency table
 3. **Cramer’s V (V)**
 - for larger tables
 - example in lab

Odds Ratio (OR)

Given the following 2 x2 table:

Effect Size	# Successes	# Failures
Treatment Group	A	B
Control Group	C	D

The odds ratio would be calculated as:

$$\text{Odds ratio} = (AD) / (BC)$$

When to Use

It's appropriate to calculate the odds ratio only when you're working with a 2 x 2 contingency table. Typically the odds ratio is calculated when you're interested in studying the odds of success in a treatment group relative to the odds of success in a control group.

How to Interpret

There is no specific value at which we deem an odds ratio be a small, medium, or large effect, but the further away the odds ratio is from 1, the higher the likelihood that the treatment has an actual effect.

It's best to use domain specific expertise to determine if a given odds ratio should be considered small, medium, or large.

Correlation between... 1 numerical variable and 1 categorical variables

... we have actually met before 😊

Correlation between 1 numerical variable and 1 categorical variables

- Recall that we have already encountered methods for for comparing numerical data across groups in the previous lessons
 1. Using **side-by-side boxplots** for visual comparison of how the distribution of a numerical variable differs by category
 2. Using **One-Way ANOVA** for testing relationships between Numerical and Categorical variables
 - i.e. the extension of the t-test for more than 2 groups

From association to causation

observation studies v. experimental studies

Association vs Causation → experimental studies

- So far, we have looked at “**observational studies**” which focus on measuring variables of interest
 - even if we may find **CORRELATION OR ASSOCIATION**, but it **DOES NOT IMPLY CAUSATION!**
 - in fact there can be “**hidden variables**” that affect the relationship between the **explanatory variable** and the **response variable**
- “**Experimental studies**” help us studying **causation** in that they are “*designed to provoke a response*”
 - they involve applying a **TREATMENT** to an **experimental unit** (or subject) and observing its **EFFECT**
 - they follow some design **PRINCIPLES** that provide robust evidence for causation

A conceptual framework to understand different types of statistical modeling

1. association/correlation → observational studies

- aimed at summarizing or representing the data structure, without an underlying causal theory
- may help form hypotheses for explanatory and predictive modeling

2. causal explanation → experimental studies

- aimed at testing causal connection between treatment and outcome variable
- prevalent in “causal theory-heavy” fields (like: economics, **psychology**, environmental science, etc.)

3. empirical prediction → algorithmic (ML) and data-mining modeling

- aimed at predicting new or future observations
- prevalent in fields like natural language processing, **bioinformatics**, etc.. In **epidemiology**, there is more of a mix causal explanation & empirical prediction.

• NOTE:

- The same modeling approach (e.g. fitting a regression model) can be used for all of the above goals
- While they shouldn't be confused, **Explanatory power** and **predictive accuracy** are complementary goals: e.g. in bioinformatics (which has little theory and abundance of data), predictive models are pivotal in generating avenues for causal theory.

Supervised regression

Linear /multiple linear models

Supervised classification

Discriminant analysis

Unsupervised clustering

Hierarchical clustering

Unsupervised association

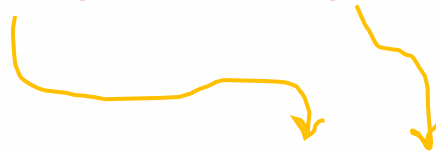
PCA

Linear relationships between X and Y and regression

Simple Linear Regression

Functional (linear) relationship and regression

- The correlation coefficient gave us information about the degree to which points (corresponding to x and y pairs) were clustered around a straight line ... but nothing about the **slope** of that line
- **regression analysis** will produce this kind of information
 - i.e., we want to know exactly how the two variables are related
 - if we hypothesize a linear relationship, a **functional form** will provide an **intercept** and a **slope**

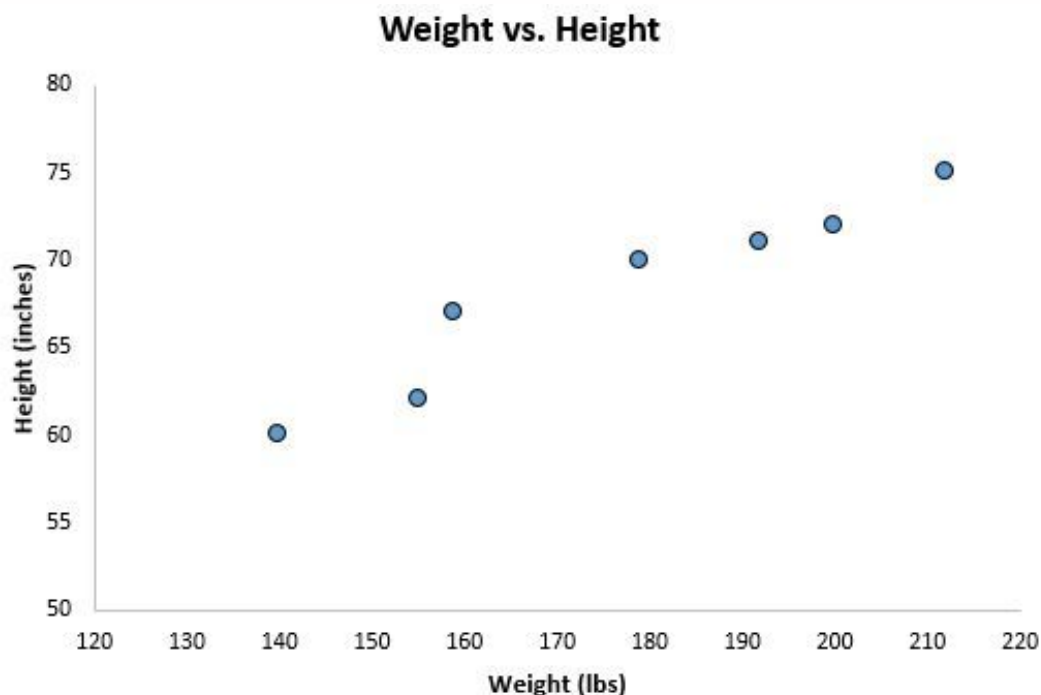


Simple linear regression: example

- **Simple linear regression** is a statistical method you can use to understand the relationship between two variables, (the predictor variable) and (the response variable).

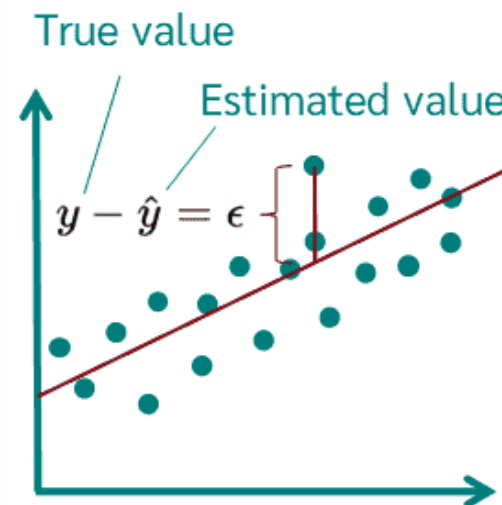
-

Weight (lbs)	Height (inches)
140	60
155	62
159	67
179	70
192	71
200	72
212	75



Linear regression (Ordinary Least Square)

- The **OLS regression line** such as to minimize the difference between estimated values and actual ones
 - in fact, OLS seeks the minimum *sum of squared distances between each point and the regression line*
 - it is the “best fitting” line given any data of points
- Note the different notations below: like with previous **inferential statistics methods**, we are making statements on the **population of interest** based on some **sample data** available



Population data	Sample data
= true Y values	= estimated (or predicted) Y values based on X values
= true intercept	= estimated intercept
= true slope/coefficient on x	= <i>estimated</i> slope/coefficient on x
= true residual or unobserved part of y	= <i>estimated</i> residual (error), or unobserved part of Y)

OLS Linear regression equation interpretation

- The formula for the line of best fit is written as:
$$\hat{y} = \beta_0 + \beta_1 x$$
 - where \hat{y} is the predicted value of the response variable (**height**), β_0 is the y-intercept, β_1 is the regression coefficient, and x is the value of the predictor variable (**weight**).
- For example, in the case of
 - $\beta_0 = 32.7830$. This means when the predictor variable weight is zero pounds, the predicted height is 32.7830 inches. Sometimes the value for β_0 can be useful to know, but not in this specific example
 - $\beta_1 = 0.2001$. This means that a one unit increase in x is associated with a 0.2001 unit increase in \hat{y} . In this case, a one pound increase in weight is associated with a 0.2001 inch increase in height.

Assumptions of linear regression

- For the results of a linear regression model to be valid and reliable, we need to check that the following four assumptions are met:
 1. **Linear relationship:** There exists a linear relationship between the independent variable, x , and the dependent variable, y
 2. **Normality:** The residuals of the model are normally distributed.
 - Check normality (OF RESIDUALS) with the known methods (QQplot, Shapiro-Wilk, Kolmogorov Smirnov)
 3. **Homoscedasticity:** The residuals have constant variance at every level of x .
 4. **Independence:** The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
 - This is mostly relevant when working with time series data. Ideally, we don't want there to be a pattern among consecutive residuals.

Diagnostic plotting: residuals

A **residual** is the vertical distance between a data point and the regression line. $y_i - \hat{y}_i$

- y_i : The **actual response** value for the i th observation
- \hat{y}_i : The **predicted response** value based on the multiple linear regression model

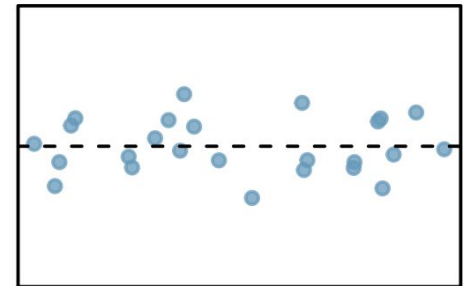
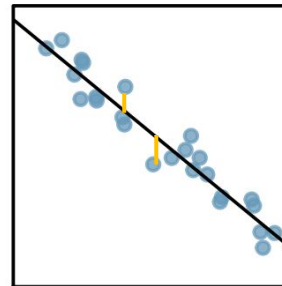
We want to see a residual plot where data shows random scatter above and below the horizontal line

In the example on the right:

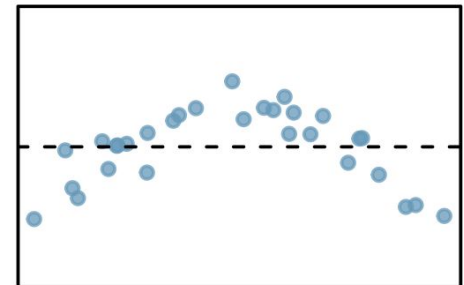
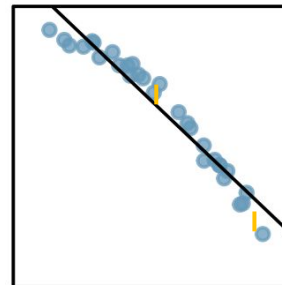
- **Case 1)** linear model is a **particularly good fit!**
- **Case 2)** the original data cycles below and above the regression line
- **Case 3)** the variability of the residuals is not constant; the residuals are slightly more variable for larger predicted values.

Best fitting
line

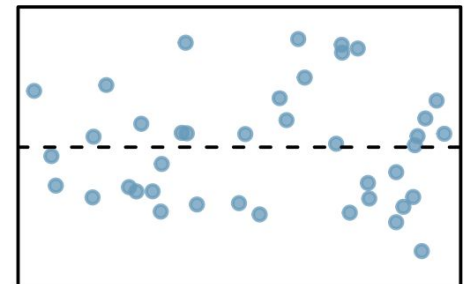
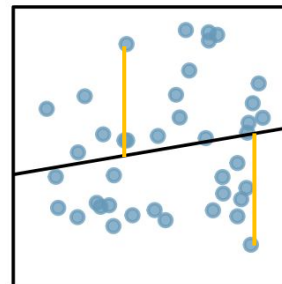
(corresponding)
Residual plots



1



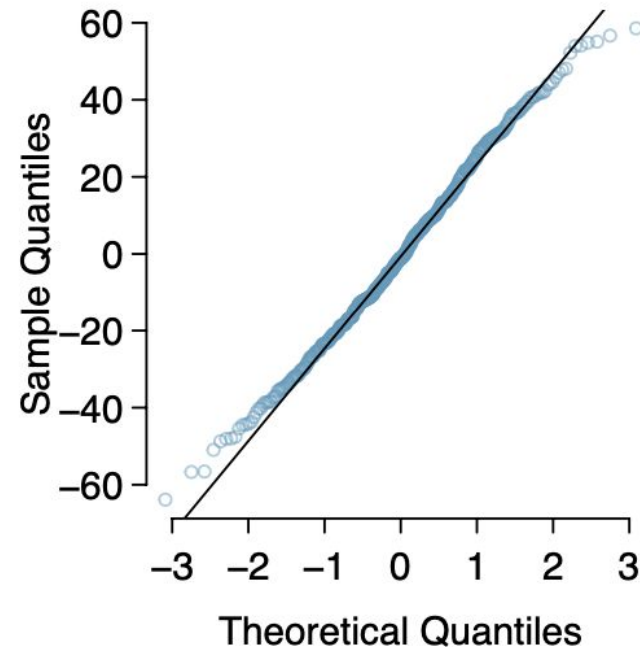
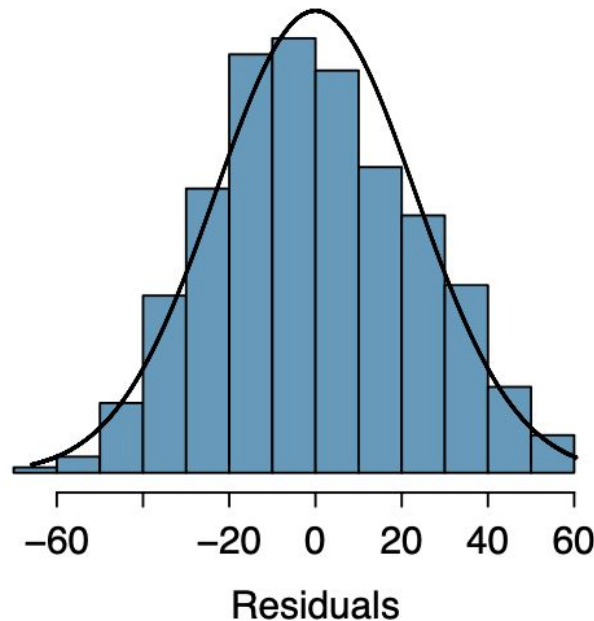
2



3

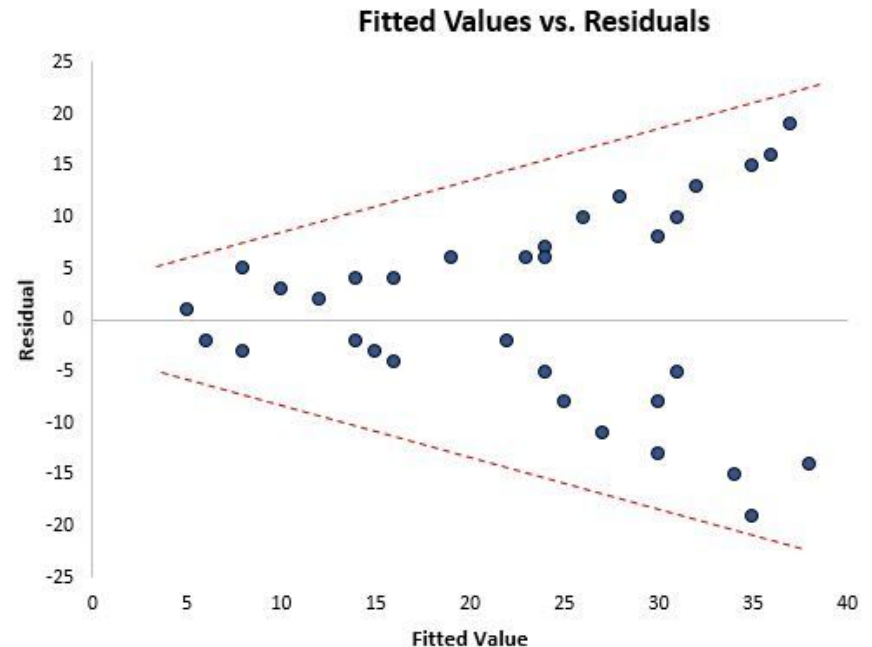
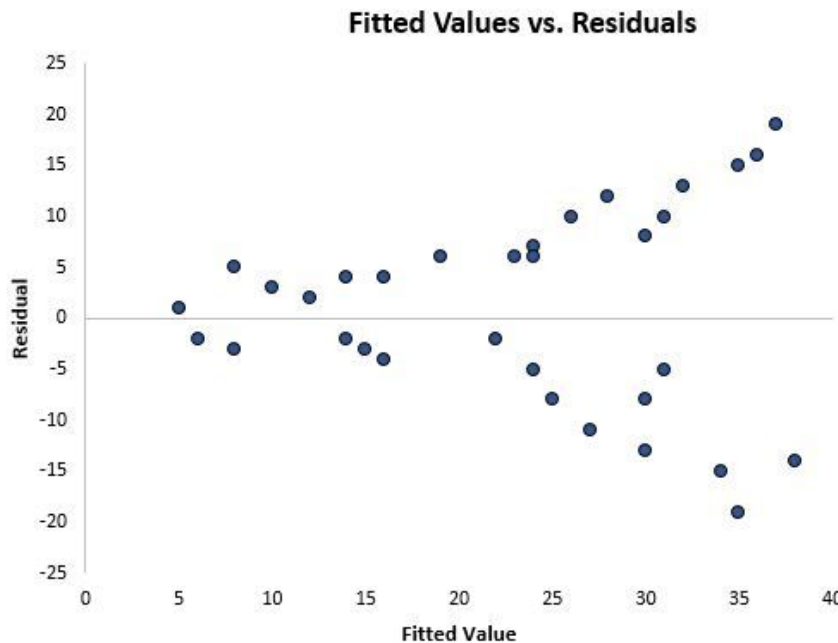
Diagnostic plotting: normality of residuals

- The residuals of the model are normally distributed.
- Check normality (OF RESIDUALS) with the known methods (QQplot, Shapiro-Wilk, Kolmogorov Smirnov)



Diagnostic plotting: Homoscedasticity

- ASSUMPTION: The residuals (i.e. the error term) have constant variance at every level of x (“homoscedasticity”)
- When this is not true, the results of the regression model might be unreliable
- This assumption can be verified by:
 - the “Residual vs. Fitted” plot
 - the Breusch-Pagan Test or the White Test



The Coefficient of Determination

One way to measure how well the least squares regression line “fits” the data is using the **coefficient of determination**, denoted as R^2 .

The coefficient of determination is the proportion of the variance in the response variable that can be explained by the predictor variable.

The coefficient of determination can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all.

A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable.

An R^2 between 0 and 1 indicates just how well the response variable can be explained by the predictor variable.

For example, an R^2 of 0.2 indicates that 20% of the variance in the response variable can be explained by the predictor variable; an R^2 of 0.77 indicates that 77% of the variance in the response variable can be explained by the predictor variable.

Regression

Multiple Linear Regression

Multiple Linear Regression

to understand the relationship between *multiple* predictor variables and a response variable

a multiple linear regression model takes the form:

$$\mathbf{Y} = \beta_0 + \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2 + \dots + \beta_p\mathbf{X}_p + \epsilon$$

where:

- \mathbf{Y} : The response variable
- \mathbf{X}_j : The j^{th} predictor variable
- β_j : The average effect on Y of a one unit increase in X_j , holding all other predictors fixed
- ϵ : The error term

The values for $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are chosen using **the least square method**, which minimizes the sum of squared residuals (RSS):

$$\text{RSS} = \sum (y_i - \hat{y}_i)^2$$

where:

Σ : A greek symbol that means *sum*

y_i : The actual response value for the i^{th} observation

\hat{y}_i : The predicted response value based on the multiple linear regression model

D	E	F	G	H	I	J	K
SUMMARY OUTPUT							
<i>Regression Statistics</i>							
Multiple R	0.857						
R Square	0.734						
Adjusted R Square	0.703						
Standard Error	5.366						
Observations	20						
ANOVA							
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>		
Regression	2	1350.76	675.38	23.46	0.00		
Residual	17	489.44	28.79				
Total	19	1840.20					

Regression SS is the total variation in the dependent variable that is explained by the regression model.

$$\sum (\hat{y} - \bar{y})^2$$

Residual SS — is the total variation in the dependent variable that is left unexplained by the regression model.

$$\sum (y - \hat{y})^2$$

Total SS — is the sum of both, regression and residual SS

D	E	F	G	H	I	J	K
SUMMARY OUTPUT							
<i>Regression Statistics</i>							
Multiple R	0.857						
R Square	0.734						
Adjusted R Square	0.703						
Standard Error	5.366						
Observations	20						
ANOVA							
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>		
Regression	2	1350.76	675.38	23.46	0.00		
Residual	17	489.44	28.79				
Total	19	1840.20					

Mean Squared Errors (MS) — are the mean of the sum of squares or the sum of squares divided by the degrees of freedom for both, regression and residuals.

$$\text{Regression MS} = \sum (\hat{y} - \bar{y})^2 / \text{Reg. df} \quad \text{Residual MS} = \sum (y - \hat{y})^2 / \text{Res. df}$$

F — is used to test the hypothesis that the slope of the independent variable is zero. ($\beta_j = 0$ means the explanatory variable is useless)

Mathematically, it can also be calculated as

$$F = \text{Regression MS} / \text{Residual MS}$$

Significance F — is nothing but the p-value for the null hypothesis that the coefficient of the independent variable is zero

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	67.67	2.82	24.03	0.00	61.73	73.61
hours	5.56	0.90	6.18	0.00	3.66	7.45
prep_exams	-0.60	0.91	-0.66	0.52	-2.53	1.33

Predictors parameters

A p-value below 0.05 indicates 95% confidence that the slope of the regression line is not zero and hence there is a significant linear relationship between the dependent and independent variables.

A p-value greater than 0.05 indicates that the slope of the regression line may be zero and that there is not sufficient evidence at the 95% confidence level that a significant linear relationship exists between the dependent and independent variables.

<i>Regression Statistics</i>	
Multiple R	0.810350635
R Square	0.656668152
Adjusted R Square	0.655978731
Standard Error	0.082783542
Observations	500

R² (R Square) — represents the power of a model. It shows the amount of variation in the dependent variable the independent variable explains and always lies between values 0 and 1. As the R² increases, more variation in the data is explained by the model and better the model gets at prediction. A low R² would indicate that the model doesn't fit the data well and that an independent variable doesn't explain the variation in the dependent variable well.

$R^2 = \text{Regression Sum of Squares} / \text{Total Sum of Squares}$

Adjusted R² — is R² multiplied by an adjustment factor. This is used while comparing different regression models with different independent variables. This number comes in handy while deciding on the right independent variables in multiple regression models.

Standard Error — This is the estimated standard deviation of the error of the regression equation and is a good measure of the accuracy of the regression line. It is the square root of the residual mean squared errors.

Std. Error = $\sqrt{(\text{Res. MS})}$

From causation to prediction

Machine learning models

Defining Machine Learning

- Machine learning is a very broad topic and a highly active research area. In the life sciences, much of what is described as “*precision medicine*” is an application of machine learning to biomedical data. The general idea is to **predict or discover outcomes from measured predictors**.
 - Can we discover new types of cancer from gene expression profiles?
 - Can we predict drug response from a series of genotypes?
 -

Supervised or Unsupervised ML algorithms

....another conceptual framework

A Quick Introduction to Supervised vs. Unsupervised Learning

The field of machine learning contains a massive set of algorithms that can be used for understanding data. These algorithms can be classified as:

1. Supervised Learning Algorithms:

building a model to estimate or predict an output based on one or more inputs.

2. Unsupervised Learning Algorithms:

finding structure and relationships among inputs. There is no “supervising” output.

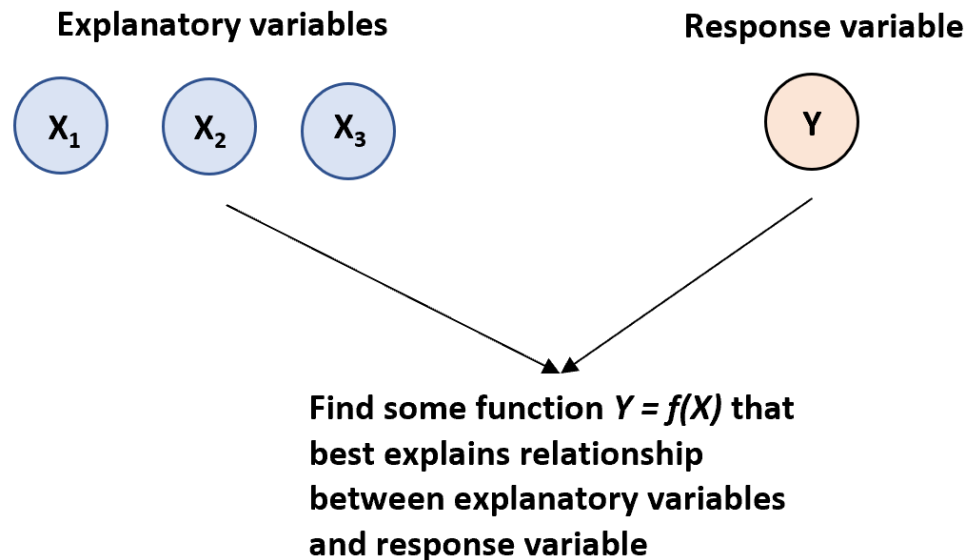
Supervised Learning Algorithms

A **supervised learning algorithm** can be used when we have one or more explanatory variables ($X_1, X_2, X_3, \dots, X_p$) and a **response variable** (Y) and we would like to find some function that describes the relationship between the explanatory variables and the response variable:

$$\mathbf{Y} = f(\mathbf{X}) + \epsilon$$

where f represents systematic information that X provides about Y and where ϵ is a random error term

Supervised Learning



There are two main types of supervised learning algorithms:

- 1. Regression:** The output variable is continuous (e.g. weight, height, time, etc.)
- 2. Classification:** The output variable is categorical (e.g. male or female, pass or fail, benign or malignant, etc.)

There are two main reasons that we use supervised learning algorithms:

1. Prediction: We often use a set of explanatory variables to predict the value of some response variable (e.g. using *square footage* and *number of bedrooms* to predict *home price*)

2. Inference: We may be interested in understanding the way that a response variable is affected as the value of the explanatory variables change (e.g. how much does home price increase, on average, when the number of bedrooms increases by one?)

Depending on whether our goal is inference or prediction (or a mix of both), we may use different methods for estimating the function f . For example, linear models offer easier interpretation but non-linear models that are difficult to interpret may offer more accurate prediction.

Here is a list of the most commonly used **supervised** learning algorithms:

- **Linear regression**
- Logistic regression
- Linear discriminant analysis
- Quadratic discriminant analysis
- Decision trees
- Naive bayes
- Support vector machines
- Neural networks

Unsupervised ML algorithms

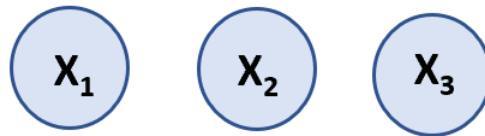
Example of PCA

Unsupervised Learning Algorithms

An **unsupervised learning algorithm** can be used when we have a list of variables ($X_1, X_2, X_3, \dots, X_p$) and we would simply like to find underlying structure or patterns within the data.

Unsupervised Learning

Explanatory variables



Find some underlying structure
or patterns within the data

There are two main types of unsupervised learning algorithms:

1. Clustering: Using these types of algorithms, we attempt to find “clusters” of [observations](#) in a dataset that are similar to each other.

2. Association: Using these types of algorithms, we attempt to find “rules” that can be used to draw associations. For example, if a patient has a high biomarker X, he will have a low biomarker Y.

Here is a list of the most commonly used unsupervised learning algorithms:

- **Principal component analysis**

- K-means clustering

- K-medoids clustering

- **Hierarchical clustering**

- Apriori algorithm

Summary: Supervised vs. Unsupervised Learning

The following table summarizes the differences between supervised and unsupervised learning algorithms:

	Supervised Learning	Unsupervised Learning
Description	Involves building a model to estimate or predict an output based on one or more inputs.	Involves finding structure and relationships from inputs. There is no “supervising” output.
Variables	Explanatory and Response variables	Explanatory variables only
End goal	Develop model to (1) predict new values or (2) understand existing relationship between explanatory and response variables	Develop model to (1) place observations from a dataset into a specific cluster or to (2) create rules to identify associations between variables.
Types of algorithms	(1) Regression and (2) Classification	(1) Clustering and (2) Association

Supervised regression
Linear /multiple linear models

Supervised classification
Discriminant analysis

Unsupervised clustering
Hierarchical clustering

Unsupervised association
PCA

old

Chi-Square Test of Independence

A **Chi-Square Test of Independence** is used to determine whether or not there is a significant association between two categorical variables.

Examples

- We want to know if education level and marital status are associated so we collect data about these two variables on a simple random sample of 50 people.

Chi-Square Test of Independence: Formula

A Chi-Square test of independence uses the following null and alternative hypotheses:

H₀: (null hypothesis) The two variables are independent.

H₁: (alternative hypothesis) The two variables are *not* independent. (i.e. they are associated)

We use the following formula to calculate the Chi-Square test statistic X^2 :

$$X^2 = \sum (O - E)^2 / E$$

where:

Σ : means "sum"

O: observed value

E: expected value

If the p-value that corresponds to the test statistic X^2 with $(\text{\#rows}-1) * (\text{\#columns}-1)$ degrees of freedom is less than your chosen significance level then you can reject the null hypothesis

Chi-Square Test of Independence: Example

Suppose we want to know whether or not gender is associated with political party preference. We take a simple random sample of 500 voters and survey them on their political party preference. The following table shows the results of the survey:

	Republican	Democrat	Independent	Total
Male	120	90	40	250
Female	110	95	45	250
Total	230	185	85	500

Step 1: Define the hypotheses.

We will perform the Chi-Square test of independence using the following hypotheses:

- H_0 : Gender and political party preference are independent.
- H_1 : Gender and political party preference are *not* independent

Step 2: Calculate the expected values.

Next, we will calculate the expected values for each cell in the contingency table using the following formula:

Expected value = (row sum * column sum) / table sum.

For example, the expected value for Male Republicans is: $(230 * 250) / 500 = \mathbf{115}$.

We can repeat this formula to obtain the expected value for each cell in the table:

	Republican	Democrat	Independent	Total
Male	115	92.5	42.5	250
Female	115	92.5	42.5	250
Total	230	185	85	500

Step 3: Calculate $(O-E)^2 / E$ for each cell in the table.

Next we will calculate $(O-E)^2 / E$ for each cell in the table where:

- O**: observed value
- E**: expected value

For example, Male Republicans would have a value of: $(120-115)^2 / 115 = \mathbf{0.2174}$.

We can repeat this formula for each cell in the table:

	Republican	Democrat	Independent
Male	0.2174	0.0676	0.1471
Female	0.2174	0.0676	0.1471

Step 4: Calculate the test statistic X^2 and the corresponding p-value.

$$X^2 = \sum (O-E)^2 / E = 0.2174 + 0.2174 + 0.0676 + 0.0676 + 0.1471 + 0.1471 = \mathbf{0.8642}$$

According to the [Chi-Square Score to P Value Calculator](#), the p-value associated with $X^2 = 0.8642$ and $(2-1) * (3-1) = 2$ degrees of freedom is **0.649198**.

Step 5: Draw a conclusion.

Since this p-value is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that there is an association between gender and political party preference.

Chi-Square Goodness of Fit Test

A **Chi-Square goodness of fit test** is used to determine whether or not a categorical variable follows a hypothesized distribution.

Example:

We want to know if the percentage of M&M's that come in a bag are as follows: 20% yellow, 30% blue, 30% red, 20% other. To test this, we open a random bag of M&M's and count how many of each color appear.

We can use a Chi-Square goodness of fit test to determine if there is a statistically significant difference in the number of expected counts for each level of a variable compared to the observed counts.

Chi-Square Goodness of Fit Test: Formula

A Chi-Square goodness of fit test uses the following null and alternative hypotheses:

- **H₀: (null hypothesis)** A variable follows a hypothesized distribution.
- **H₁: (alternative hypothesis)** A variable does not follow a hypothesized distribution.

We use the following formula to calculate the Chi-Square test statistic X^2 :

$$X^2 = \sum (O-E)^2 / E$$

where:

Σ : means “sum”

O: observed value

E: expected value

If the p-value that corresponds to the test statistic X^2 with $n-1$ degrees of freedom (where n is the number of categories) is less than your chosen significance level (common choices are 0.10, 0.05, and 0.01) then you can reject the null hypothesis.

Chi-Square Goodness of Fit Test: Example

A shop owner claims that an equal number of customers come in on each weekday. To test this hypothesis, an independent researcher records the number of customers that come into the shop on a given week and finds the following:

Monday: 50 customers

Tuesday: 60 customers

Wednesday: 40 customers

Thursday: 47 customers

Friday: 53 customers

Step 1: Define the hypotheses.

We will perform the Chi-Square goodness of fit test using the following hypotheses:

H_0 : An equal number of customers come into the shop each day.

H_1 : An equal number of customers do not come into the shop each day.

eviterei
l'esempio x
restare sulla
correlazione
tra 2 var

eviterei
l'esempio x
restare sulla
correlazione
tra 2 var

Chi-Square Goodness of Fit Test: Example

Step 2: Calculate $(O-E)^2 / E$ for each day.

There were a total of 250 customers that came into the shop during the week. Thus, if we expected an equal amount to come in each day then the expected value "E" for each day would be 50.

Step 3: Calculate the test statistic X^2 .

$$X^2 = \sum (O-E)^2 / E = 0 + 2 + 2 + 0.18 + 0.18 = 4.36$$

Step 4: Calculate the p-value of the test statistic X^2 .

According to the [Chi-Square Score to P Value Calculator](#), the p-value associated with $X^2 = 4.36$ and $n-1 = 5-1 = 4$ degrees of freedom is **0.359472**.

Step 5: Draw a conclusion.

Since this p-value is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that the true distribution of customers is different from the distribution that the shop owner claimed.