



Biostatistics

Microbiome: a Hands-on Experience 2025

Center for Plant Molecular Biology - ZMBP

Agenda:

- Intro to Statistics
- Getting to know the data
- Data handling and visualization in R
- Hypothesis Testing
- Predictive Models in Microbiome Analysis



Intro to Statistics

Biostatistics or medical biometry:

- a branch of statistics that applies statistical techniques and principles to scientific research to a wide range of topics in health-related fields
 - ▶ e.g. medicine, biology, and public health
- the development of new tools to study these areas.
- includes the design of biological experiments, the collection and analysis of data from those experiments, and the interpretation of the results.

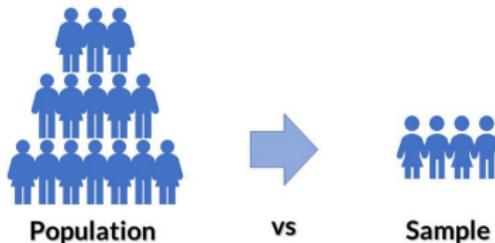
Population vs. Sample

Population: is the set of all individuals of interest

- often large and impossible to obtain measurements from all individuals
- Examples: all individuals with Migraine, all people who take a daily vitamin supplement in Germany, ...

Sample: a set of individuals selected from a population

- representative and generalizable



Parameter vs. Statistic

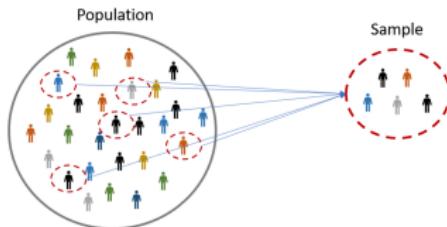
Parameter: describes a population

Statistic: describes a sample

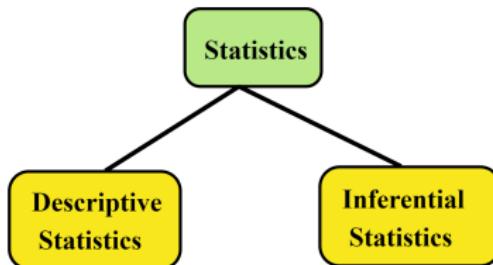
A statistic is used to estimate a parameter

For example:

- μ describes the population mean
- \bar{X} represents the sample mean (average)



Descriptive vs. Inferential



Descriptive statistics methods: are used to collect, summarize, organize, simplify, and describe data.

- **Examples:** mean, median, variance, standard deviation

Descriptive vs. Inferential

Inferential Statistics methods: concluding and making decisions concerning a population based only on a sample

- Examples: regression analysis, confidence interval, and hypothesis testing



Sampling error

- arises when a sample does not represent the whole population.
- the discrepancy between a sample statistic and the true population parameter
- Increasing the sample size and random selection can reduce the errors.
- **Example:**

true population value: $\mu = 19.5$

sample statistic: $\bar{X} = 18$

sampling error: $\mu - \bar{X} = 1.5$

there is a discrepancy of 1.5

Types of Variables

Independent (Explanatory): is the variables that can be altered or manipulated in research (e.g., caffeine dose)

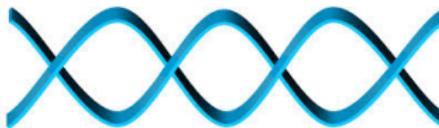
Dependent (Response): is the result of manipulation done to the variables (e.g., reaction times).

Example: X and Y in linear regression model $Y = AX + \epsilon$



Getting to know the data

GREIN:



GEO RNA-seq Experiments Interactive Navigator is an Interactive Web Platform for Re-analyzing GEO RNA-seq data and the large number (> 6,000) of already processed datasets [Mahi et al. \[2019\]](#).

Link: <http://www.ilincs.org/apps/grein/>

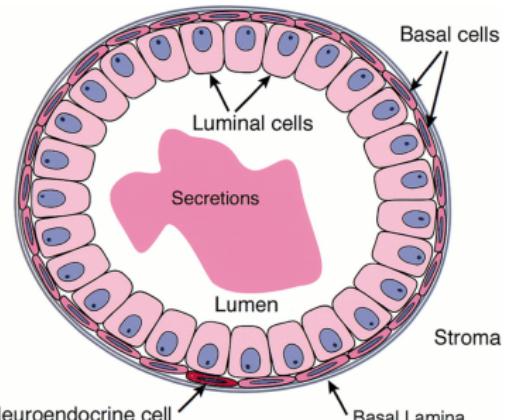
GREIN provides both raw and normalized counts (normalized for differences in sequencing depth and composition bias).

Getting started with data



Data set: GEO code GSE60450.

- RNA-seq data from the paper by [Fu et al. \[2015\]](#)
- examined expression in basal and luminal cells from mice at different stages (virgin, pregnant, and lactating)



Getting started with data



Data set: GEO code GSE60450

- There are 6 groups and 2 samples per group, 12 samples in total.

Cell Type	Group	Samples
Basel Cell	Virgin Pregnant Lactating	N=2 N=2 N=2
Luminal Cell	Virgin Pregnant Lactating	N=2 N=2 N=2

Getting to know the data



Metadata:

	characteristics	immunophenotype	developmental stage
GSM1480291	mammary gland, luminal cells, virgin	luminal cell population	virgin
GSM1480292	mammary gland, luminal cells, virgin	luminal cell population	virgin
GSM1480293	mammary gland, luminal cells, 18.5 day pregnancy	luminal cell population	18.5 day pregnancy
GSM1480294	mammary gland, luminal cells, 18.5 day pregnancy	luminal cell population	18.5 day pregnancy
GSM1480295	mammary gland, luminal cells, 2 day lactation	luminal cell population	2 day lactation
GSM1480296	mammary gland, luminal cells, 2 day lactation	luminal cell population	2 day lactation
GSM1480297	mammary gland, basal cells, virgin	basal cell population	virgin
GSM1480298	mammary gland, basal cells, virgin	basal cell population	virgin
GSM1480299	mammary gland, basal cells, 18.5 day pregnancy	basal cell population	18.5 day pregnancy
GSM1480300	mammary gland, basal cells, 18.5 day pregnancy	basal cell population	18.5 day pregnancy
GSM1480301	mammary gland, basal cells, 2 day lactation	basal cell population	2 day lactation
GSM1480302	mammary gland, basal cells, 2 day lactation	basal cell population	2 day lactation

Showing 1 to 12 of 12 samples

Getting to know the data



Counts table:

	gene_symbol	GSM1480291	GSM1480292	GSM1480293	GSM1480294
ENSMUSG00000000001	Gnai3	243.28596	255.66037	239.73819	217.10047
ENSMUSG00000000003	Pbsn	0	0	0	0
ENSMUSG00000000028	Cdc45	11.18453	13.78314	11.60091	4.2718
ENSMUSG00000000031	H19	6.30808	8.53042	7.09408	11.03901
ENSMUSG00000000037	Scml2	2.19217	4.66442	2.7959	2.49541
ENSMUSG00000000049	Apoh	0.22369	0.08404	0	0
ENSMUSG00000000056	Narf	11.27401	14.74964	26.16464	18.8213
ENSMUSG00000000058	Cav2	118.24288	112.70235	50.53489	63.40027
ENSMUSG00000000078	Klf6	2036.16657	2230.26276	1902.63241	1959.61407
ENSMUSG00000000085	Scmh1	33.68781	38.70204	9.18057	9.4318
ENSMUSG00000000088	Cox5a	126.92208	108.75231	141.96507	125.4894

Formatting the data



Converting from wide to long format:

Gene_id	Sample_1	Sample_2	Sample_3	Sample_4
Gene_1	243	255	239	205
Gene_2	11	13	10	16
Gene_3	6	8	7	4

pivot_wider()
Converting from wide to long format

Converting from long to wide format
pivot_longer()

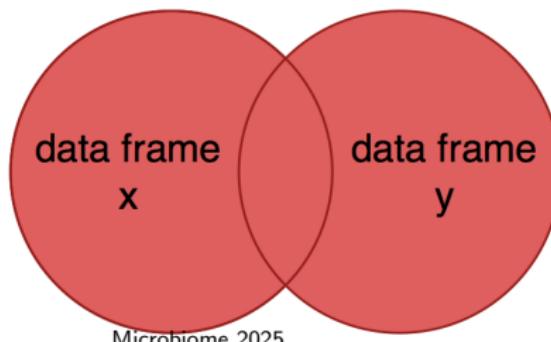
Gene_id	Sample	Counts
Gene_1	Sample_1	243
Gene_1	Sample_2	255
Gene_1	Sample_3	239
Gene_1	Sample_4	205
Gene_2	Sample_1	11
Gene_2	Sample_2	13
Gene_2	Sample_3	10
Gene_2	Sample_4	16
Gene_3	Sample_1	6
Gene_3	Sample_2	8
Gene_3	Sample_3	7
Gene_3	Sample_4	4

Formatting the data



Joining (merging) datasets:

- The process involves combining datasets that share at least some of the same observations (rows) but have different variables (columns).
- Typically, there is one variable in common, called the “key” variable, shows which rows from one dataset match the rows of the other.



Formatting the data



Joining two tables:

Long format data table

Gene_id	Sample	Counts
gene_1	Sample_1	243
gene_1	Sample_2	255
gene_1	Sample_3	239
gene_1	Sample_4	205



Metadata table

Sample	Immunophenotype	Stage
Sample_1	luminal cell	virgin
Sample_2	basal cell	virgin
Sample_3	luminal cell	pregnancy
Sample_4	basal cell	lactation



Gene_id	Sample	Counts	Immunophenotype	Stage
gene_1	Sample_1	243	luminal cell	virgin
gene_1	Sample_2	255	basal cell	virgin
gene_1	Sample_3	239	luminal cell	pregnancy
gene_1	Sample_4	205	basal cell	lactation



Data handling and visualization in R

Data handling and visualization in R

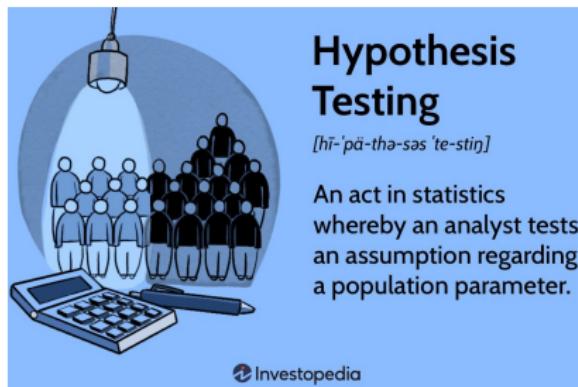


Link: <https://colab.research.google.com/drive/1mRuWrMP87i8i9TQSS3VReyyGEtRZUehC?usp=sharing>

Hypothesis Testing

Inferential statistics: is used to measure behavior in samples to learn more about the behavior in populations (often too large or inaccessible).

Hypothesis testing or significance testing is a method for testing a claim or hypothesis about a parameter in a population, using data measured in a sample.



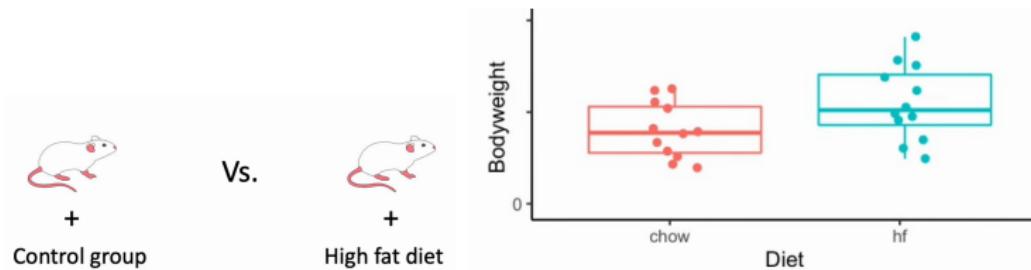
A hypothesis is

- an **educated guess** about something in the world around you.
- It should be **testable**, either by experiment or observation.
- For example a new treatment method we think might work for patients
 - ▶ Statement: If the patients receive counseling in addition to medication then their overall depression scale will decrease.
- A **statistical hypothesis test** is a method of statistical inference used to decide whether the data sufficiently supports a particular hypothesis.

Hypothesis Testing



Let's consider the mice weight example from [Winzell and Ahrén \[2004\]](#). This study characterizes the high-fat diet-fed mouse as a model for impaired glucose tolerance (IGT) and type 2 diabetes.



Question: Is there a difference in weight between mice with control vs high-fat diet?

Steps of hypothesis testing



Ideally, we undertake the following steps:

- Step 1: State the hypotheses.
- Step 2: Set the criteria for a decision.
- Step 3: Compute the test statistic.
- Step 4: Make a decision.

State the hypotheses



Null Hypothesis (H_0):

- is a statement about a population parameter, such as the population mean, that is assumed to be true.
- We will test whether the value stated in the null hypothesis is likely to be true.
- The only reason we are testing the null hypothesis is because we think it is wrong.

State the hypotheses



Alternative Hypothesis (H1):

- is a statement that directly contradicts a null hypothesis.
- states that the actual value of a population parameter is less than, greater than, or not equal to the value stated in the null hypothesis.

Null Hypothesis:

$$H_0; \mu = 85\%$$

Alternative Hypothesis:

$$H_1; \mu \neq 85\%$$

State the hypotheses



Mice example:

Null Hypothesis (H0):

there is no difference between the two diet groups.

Null model: weight = **mean** (grand mean) + residuals

Alternative Hypothesis (H1):

there is a difference between the two diet groups.

Alternative model: weight = **diet** (group mean) + residuals

A null hypothesis is **rejected** if the measured data is significantly unlikely to have occurred by chance.

Types of error



- Type I error, or α error, means rejecting the null hypothesis when the null hypothesis is true.
- Type II error, or β error, is the probability of retaining a null hypothesis that is actually false.

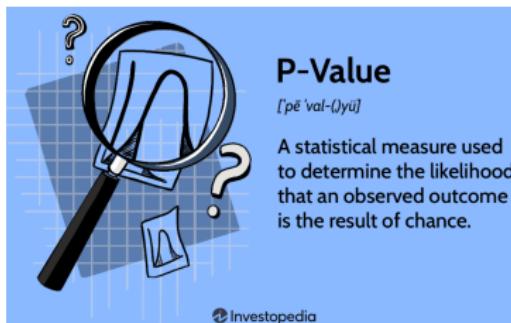
		Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	⚠ Type I Error (False positive)	✓ Correct Outcome! (True positive)	
Fail to reject null hypothesis	✓ Correct Outcome! (True negative)	⚠ Type II Error (False negative)	

The criteria for a decision: P-value



A p-value (or probability value)::

- is the probability of an α error.
- is a number describing how likely it is that your data would have occurred by random chance.
- The lower the p-value, the greater the statistical significance of the observed difference.
- a p-value doesn't tell us if the null hypothesis is true or false.
It's a piece of evidence, not a definitive proof.



The criteria for a decision: statistical significance



Statistical significance: or an α level is the level of significance or criterion for a hypothesis test.

- It is the largest probability of committing a Type I error (the highest value of p) that we will allow and still decide to reject the null hypothesis.
- In hypothesis testing, if $p \leq \alpha$, reject the null hypothesis. If $p > \alpha$, fail to reject (retain) the null hypothesis.

We directly control for the probability of a Type I error by stating an α level.

Example: Binomial test



Scenario:

- The known prevalence rate for a disease is $r = 4\%$.
- Sample: 100 test persons, $X = 9$ of them have the disease.
- H_0 : The prevalence in the test group is also $r = 4\%$.
- H_1 : The prevalence in the test group differs from $r = 4\%$.
- significance level is $\alpha = 0.05$.
- if $r = 4\%$, the probability of observing 9 or more persons with disease is $P(X \geq 9) = 0.016$, rather unlikely.

Decision: $p \leq \alpha$, we can reject the null hypothesis .

Hypothesis Types



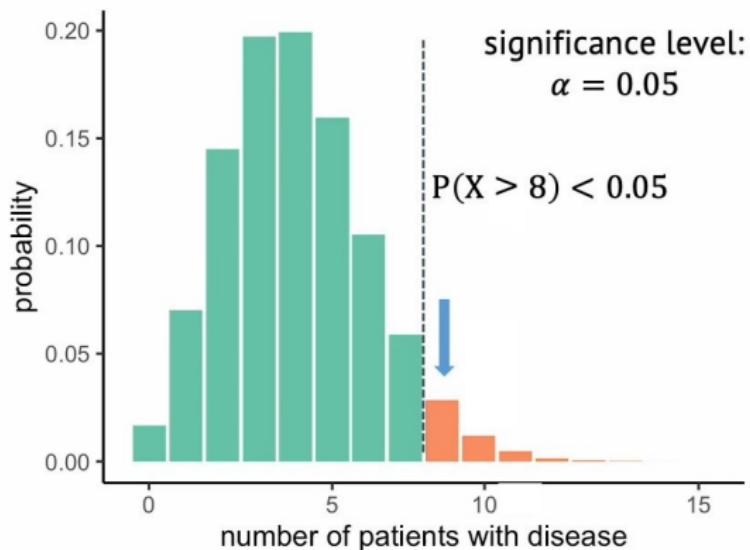
One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X < \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X \neq \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X > \mu_0$
A normal distribution curve with the left tail shaded dark grey and the right tail light grey. The region where the null hypothesis is rejected is labeled "Rejection Region". The region where the null hypothesis is accepted is labeled "Acceptance Region".	A normal distribution curve with both tails shaded dark grey. The central region where the null hypothesis is accepted is labeled "Acceptance Region". The regions where the null hypothesis is rejected in both tails are labeled "Rejection Region".	A normal distribution curve with the right tail shaded dark grey and the left tail light grey. The region where the null hypothesis is rejected is labeled "Rejection Region". The region where the null hypothesis is accepted is labeled "Acceptance Region".

Hypothesis Types: Example



One-sided Binomial test, look only in one direction:

$$H_1 : r > 0.04 \text{ or } H_1 : r < 0.04$$



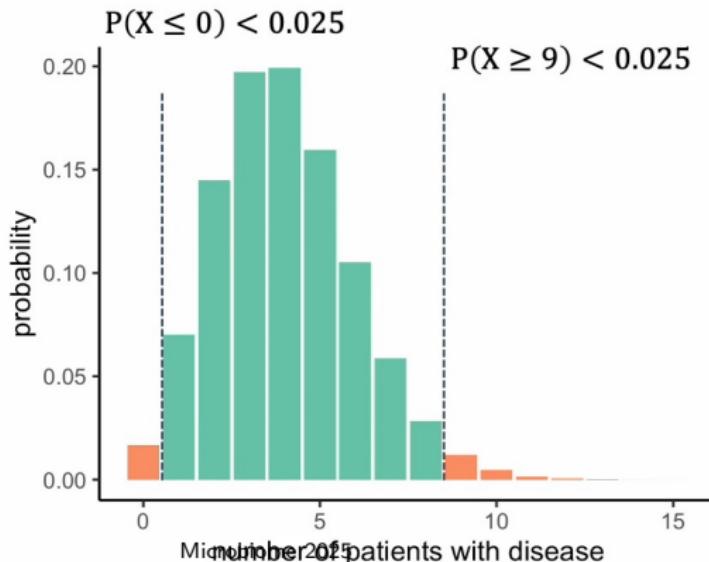
Hypothesis Types: Example



Two-sided Binomial test, look in both directions. Which number of test persons are very unlikely/extreme, assuming H0 is true?

$$H1 : r \neq 0.04$$

$$P(X = 0) = 0.017, \quad P(X > 8) = 0.019$$

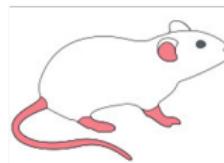
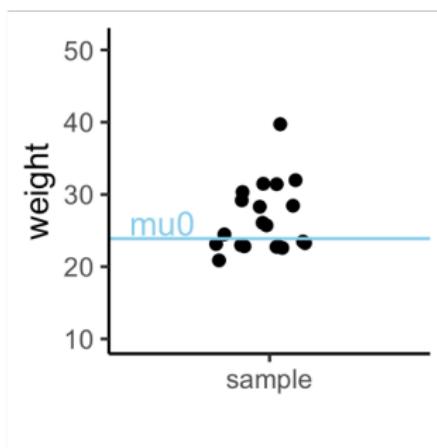


One-sample t-test



It is used for comparing one group's average to a known average.

Example: We now want to know whether the mice weights in the test group that have been fed a high-fat diet (black dots) differ significantly from a known average mouse weight (blue line).

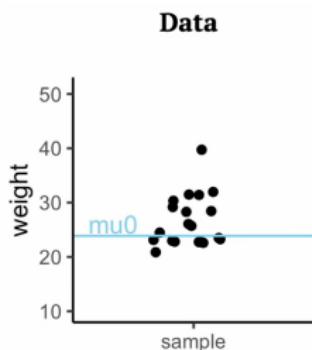


One-sample t-test



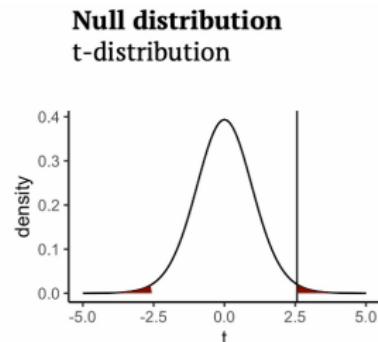
t-statistics for one-sample: $t = \frac{\bar{X} - \mu_0}{\hat{\sigma} / \sqrt{n}}$.

The larger the t-statistic, the more likely that your results will be statistically significant



Test statistic

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma} / \sqrt{n}} = 2.57$$
$$P(|t| > 2.57) = 0.019$$



Conclusion: We can reject H₀ at a 5% significance level,

the data mean is different from μ_0 .

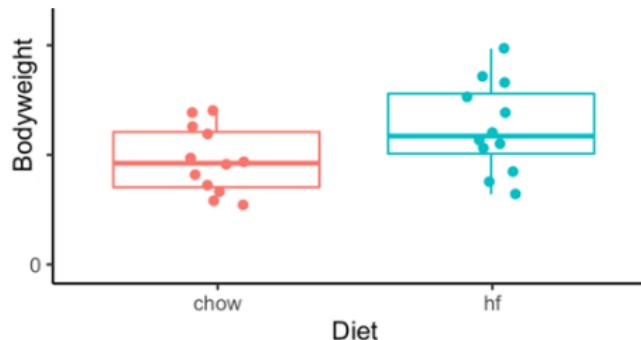
Two-sample t-test



It is used for comparing two groups with equal variances and sample sizes.

Example: a high-fat diet (Sample A) vs a control diet (Sample B).

t-statistics for two samples: $t = \frac{\bar{X}_A - \bar{X}_B}{SE}$ with $SE = \sqrt{\frac{\hat{\sigma}_A^2 + \hat{\sigma}_B^2}{n}}$

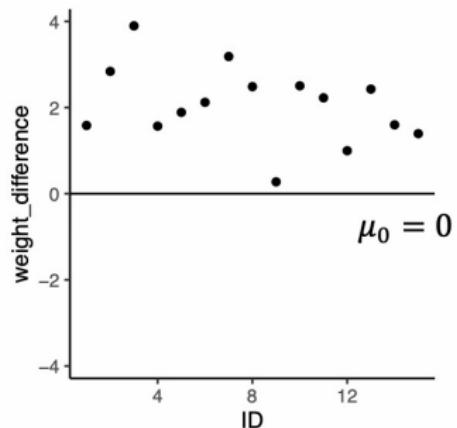


Pairing data t-test



Example: The weight of some mice is measured before and after 1 month of a high-fat diet. The issue with doing the two-sample t-test is the high variance between individual mice weights.

One-sample t-test: (H0) the mean weight gain/loss is equal to zero



Hypothesis testing in R



Link: <https://colab.research.google.com/drive/1mRuWrMP87i8i9TQSS3VReyyGEtRZUehC?usp=sharing>



Predictive Models in Microbiome Analysis

Predictive modeling is widely used in microbiome research to:

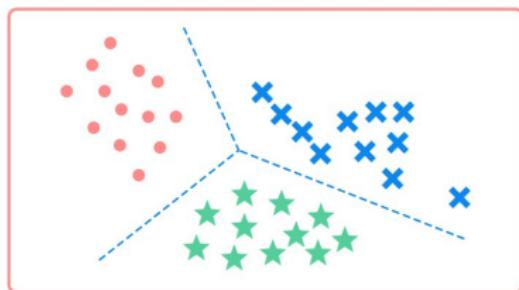
- classify microbial communities
- predict disease status
- understand environmental influences

These models leverage microbial composition and functional data to make informed predictions.

Predictive models can be categorized into and learning approaches:

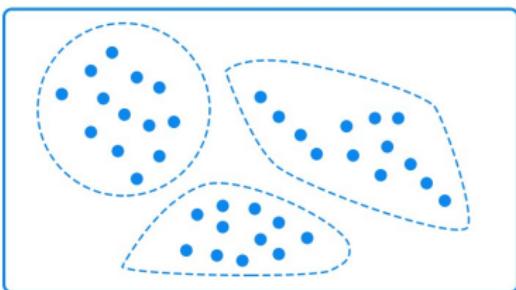
- **supervised learning approaches**
- **unsupervised learning approaches**

Classification



Supervised learning

Clustering



Unsupervised learning

Supervised Learning Models: require labeled data, where the outcome is known.

Examples in microbiome analysis:

- predicting disease status (healthy vs. diseased) based on gut microbiome profiles
- Predicting microbial abundance based on environmental factors.

Common supervised learning algorithms:

Random Forest, Support Vector Machines (SVM), Logistic Regression, Neural Networks

Unsupervised Learning Models: analyze microbiome data without predefined labels.

Examples in microbiome analysis:

- Identifying microbial community structures
- Clustering microbial profiles based on similarity

Common unsupervised learning algorithms:

K-Means Clustering, Hierarchical Clustering, Principal Component Analysis (PCA), Principal Coordinate Analysis (PCoA)

Linear Regression:

models the relationship between one or more independent variables and a continuous response (dependent, or target) variable.

For example:

- Predicting changes in microbiome composition based on environmental factors like temperature, pH, or nutrient levels.
- Disease Resistance: Analyzing how microbial communities in the soil or on plant surfaces correlate with disease resistance using regression models.
- Estimates blood sugar levels based on dietary intake and exercise patterns.

Linear Regression



Formulation:

Number of observations (Sample size): N

The number of independent variables: d

Independent variables (features): $X_1, X_2, X_3, \dots, X_d$

Response variable: Y

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_d X_d + \epsilon$$

or

$$Y = \beta X + \epsilon$$

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_d X_d + \epsilon$$

Regression coefficient β : represents the change in the dependent variable (outcome) for a one-unit change in the predictor variable while holding other predictors constant.

Example: $\beta_1 = 0.6$ means that for every one-unit increase in X_1 , the response variable Y is expected to increase by 0.6 units, assuming all other variables remain constant.

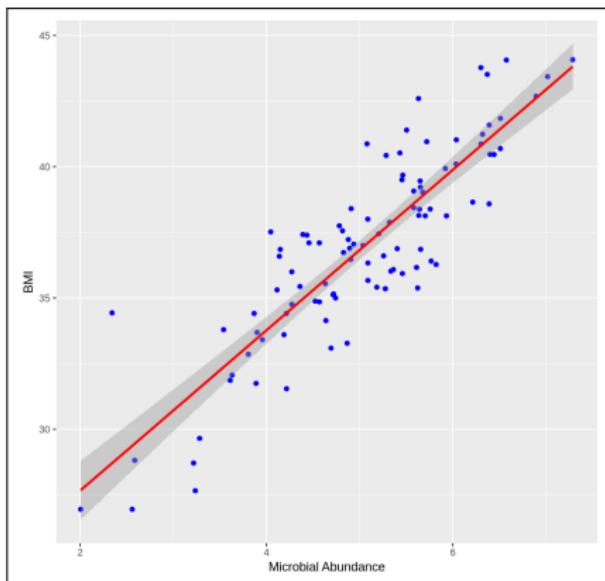
The intercept β_0 : represents the expected value of the dependent variable (Y) when all predictor variables (X) are equal to zero.

Linear Regression



Example: Predicting BMI based on the composition of the intestinal microbiome.

$$\text{BMI} = \beta_0 + \beta_1 \times \text{microbial-abundance} + \epsilon$$



Regression in R



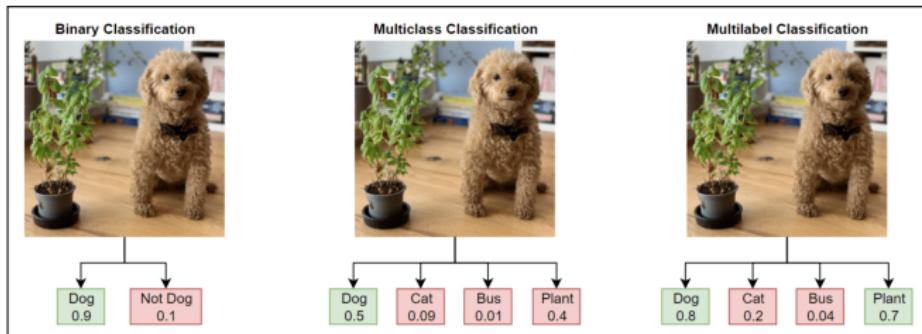
Link: <https://colab.research.google.com/drive/1mRuWrMP87i8i9TQSS3VReyyGEtRZUehC?usp=sharing>

Classification models:

are techniques for categorizing data into different classes or groups based on their features.

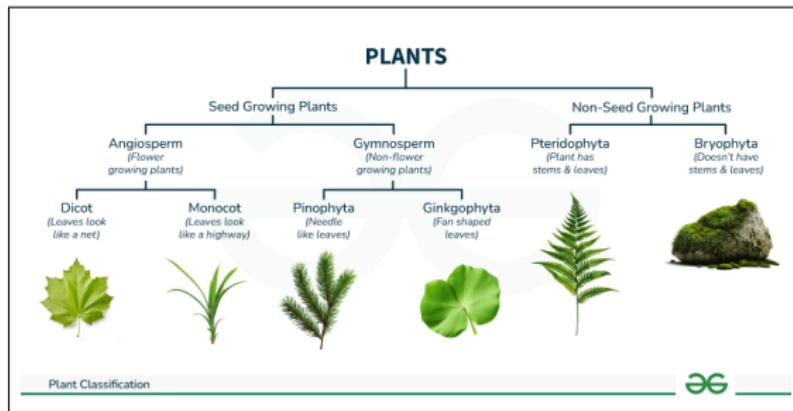
Examples:

- Binary classes: Gender (male or female), fail or pass an exam,
- Multi-class: Blood type (A, B, AB, O)



Classification models:

- Categorize samples based on microbiome data.
- Are crucial for identifying and understanding the composition of microbial communities from various samples, like soil, water, or human microbiota.



Types of Classification Methods:

Supervised Learning Methods:

- **Decision Trees:** These use a tree-like model of decisions to classify the data.
- **Random Forest:** A collection of decision trees to improve accuracy and control overfitting.
- **Logistic Regression:** Predicts the probability of a categorical dependent variable.
- **k-Nearest Neighbors (k-NN)** – Compares sequences to reference databases.

Unsupervised Learning Methods:

- **Clustering Algorithms** (e.g., k-means, hierarchical clustering): Group similar items into clusters without predefined classes.
- **Principal Component Analysis (PCA):** Reduces the dimensionality of data to find patterns.

Application in Microbiome Analysis: Taxonomic Classification

Example: 16S rRNA Gene Sequencing

- The 16S ribosomal RNA (rRNA) gene is highly conserved across bacteria and archaea, with variable regions that help differentiate species.
- Sequencing this gene allows the classification of microbes at different taxonomic levels (kingdom, phylum, genus, species).
- Classification Methods: k-Nearest Neighbors (k-NN), Random Forest,...
- Example Tools: QIIME2, Mothur, SILVA Database

Choosing Between Regression and Classification

The only difference between the two is the **type of output**.

- Regression is used to predict a continuous output (e.g., house prices).
- Classification is used for predicting a discrete output (e.g., classifying emails as spam or not).

Task	Use Regression	Use Classification
Predicting a continuous variable (e.g., metabolite levels, BMI)	✓	✗
Predicting a categorical outcome (e.g., healthy vs. diseased)	✗	✓
Microbial abundance modeling	✓	✗
Microbial community type identification	✗	✓

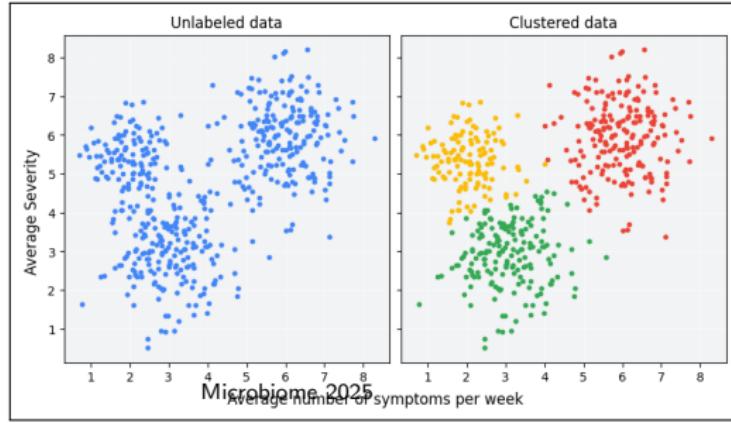
Classification in R



Link: <https://colab.research.google.com/drive/1mRuWrMP87i8i9TQSS3VReyyGEtRZUehC?usp=sharing>

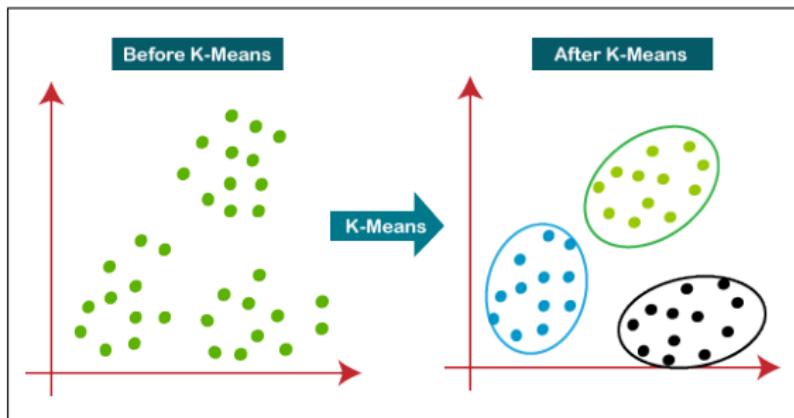
Clustering:

- is an unsupervised learning technique
- to group unlabeled samples based on their similarity to each other.
- groups similar microbiome samples based on their characteristics.
- If the samples are labeled: ——> **Classification**.



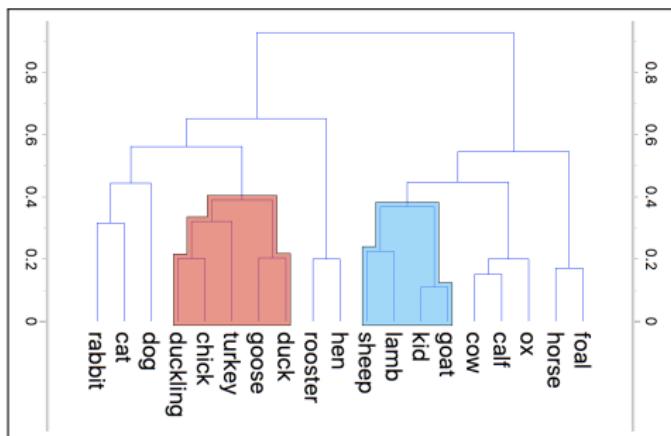
K-Means Clustering:

- Partitions data into k clusters by minimizing the variance within each cluster
- Commonly used to classify microbial community structures based on species composition.



Hierarchical Clustering:

- Builds a hierarchy of clusters using linkage methods (e.g., complete, average, single linkage).
- Suitable for identifying sub-groups in microbiome samples.

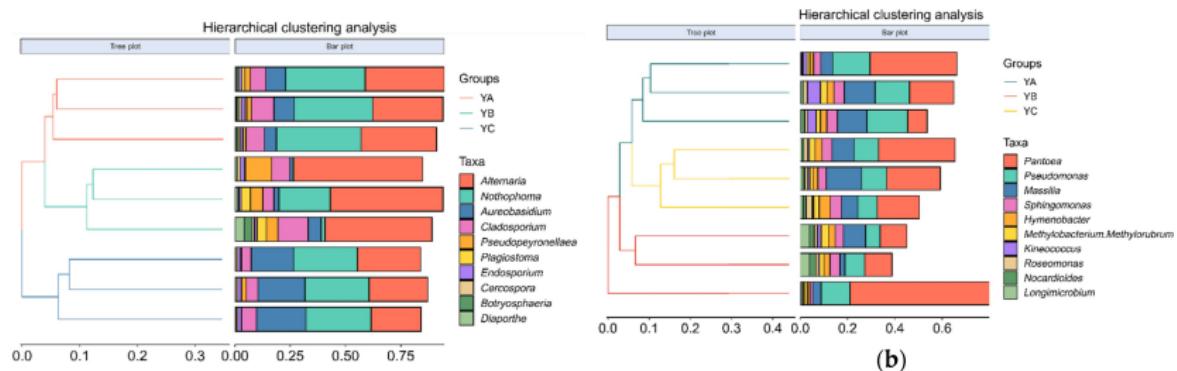


Clustering



Hierarchical clustering analysis of phyllosphere microbial communities of different Poplars at the genus level:

- Left: Hierarchical clustering analysis of phyllosphere fungal communities.
- Right: Hierarchical clustering analysis of phyllosphere bacterial communities.



Yin et al. [2024]

Clustering in R



Link: <https://colab.research.google.com/drive/1mRuWrMP87i8i9TQSS3VReyyGEtRZUehC?usp=sharing>

References

Bibliography



- N. Y. Fu, A. C. Rios, B. Pal, and et al. Egf-mediated induction of mcl-1 at the switch to lactation is essential for alveolar cell survival. *Nature Cell Biology*, 17(4):365–375, 2015. URL <https://doi.org/10.1038/ncb3117>.
- N. Mahi, M. F. Najafabadi, M. Pilarczyk, M. Kouril, and M. Medvedovic. Grein: An interactive web platform for re-analyzing geo rna-seq data. *Scientific Reports*, 9(1), 2019. URL <https://doi.org/10.1038/s41598-019-43935-8>.
- M. S. Winzell and B. Ahrén. The high-fat diet-fed mouse: a model for studying mechanisms and treatment of impaired glucose tolerance and type 2 diabetes. *Diabetes*, 35(3):215–219, 2004. URL https://doi.org/10.2337/diabetes.53.suppl_3.S215.
- Xin Yin, Weixi Zhang, Dan Li, Ran Wang, Xinyao Cong, Zhongyi Pang, Yanhui Peng, Yang Ge, Wenxu Zhu, and Changjun Ding. Factors influencing the change of phyllosphere microbial community of three *populus* spp. in the same habitat. *Forests*, 15(8:1453):215–219, 2024. URL <https://doi.org/10.3390/f15081453>.