Basic Statistical Methods

Patricia Sebastián León
psebastian@cipf.es

Bioinformatics and Genomics Department

Centro de Investigación Príncipe Felipe, Valencia, Spain

VII International Course of **M**assive **D**ata **A**nalysis

Index

Outline

What is Statistics?

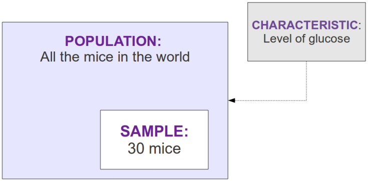
Statistics

Is the science of the collection, organization and interpretation of data.

What is Statistics?

Statistics

Is the science of the collection, organization and interpretation of data.



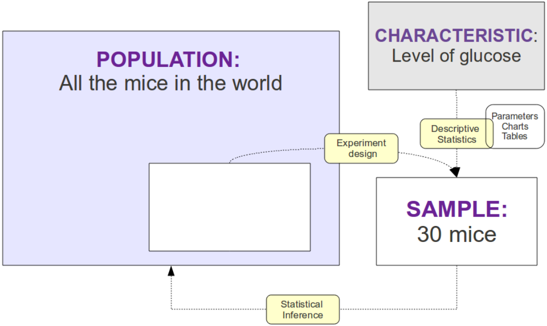
- **Variable**: Is the measure of the characteristic of interest for our study.
- **Population**: The universal set of all objects or individuals under study.
- **Sample**: A subset of the population.

Goal of Statistics

The goal of the statistical inference is to extend the sample information to the population and to provide a measurement of the probability of the error you are making

Goal of Statistics

The goal of the statistical inference is to extend the sample information to the population and to provide a measurement of the probability of the error you are making



Random experiment

Definition

A random experiment is an experiment, trial or observation whose outcome cannot be predicted with certainty, before the experiment is run. It is usually assumed that the experiment can be repeated indefinitely under essentially the same conditions

Random experiment

Definition

A random experiment is an experiment, trial or observation whose outcome cannot be predicted with certainty, before the experiment is run. It is usually assumed that the experiment can be repeated indefinitely under essentially the same conditions

Examples:

- Tossing a coin
- Measuring the expression of a gene
- Giving a drug to a mouse

Random variable

Definition

A random variable X associates a numerical value to each of the possible results of a random experiment

Random variable

Definition

A random variable X associates a numerical value to each of the possible results of a random experiment

Examples:

- $X = \begin{cases} 0, & heads \\ 1, & tails \end{cases}$
- X ={Light intensity of a probe in a microarray experiment}={0.112, 3.2, -2.73, ...}
- $X = \begin{cases} 0, & The\ drug\ is\ effective \\ 1, & The\ drug\ is\ NOT\ effective \end{cases}$

Types of random variables

There are two possible types of random variables:

1. Discrete variables.
2. Continuous variables.

Types of random variables

Discrete variable

A discrete variable is one variable that cannot take on all values within its limits.

Types of random variables

Discrete variable

A discrete variable is one variable that cannot take on all values within its limits.

Examples:

- X= Giving a drug to a mouse
  $$=\begin{cases} 0, & \textit{The drug is effective} \\ 1, & \textit{The drug is NOT effective} \end{cases}$$
- X = Number of tails obtained when tossing a coin 10 times = {0, 1, 2, 3, 4, ..., 10}
- X = Number of *counts* in a NGS experiment = {0, 1, 2, ..., n}

Types of random variables

Continuous variable

A continuous variable is one variable that can take on all values within its limits

Types of random variables

Continuous variable

A continuous variable is one variable that can take on all values within its limits

Examples:

- X = Gene expression level in a microarray experiment ={0.112, 3.2, 2.73, ...}
- X = Height of a person in a given population = {1.67m, 1.50m, 2.01m, 1.90m, ...}
- X = Time (in minutes) to get home from work every day = {15.5minutes, 20.12minutes, 17.6 minutes, ...}

Probability Distributions

Probability Distribution

A probability distribution identifies:

- When the random variable is discrete it identifies the probability of each value of the variable.
- When the random variable is continuous it identifies the probability of the value falling within a particular interval.

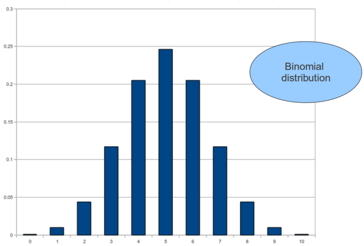
Discrete distributions

Discrete distribution

A discrete probability distribution is a probability function that only can take values in a discrete set of values.

Example of discrete probability distribution

X = {Number of tails obtained when tossing a coin 10 times} = {0, 1, 2, ..., 10}



f(2) = P(X = 2) = P(2 tails in 10 throwings) = 0.044
f(5) = P(X = 5) = P(5 tails in 10 throwings) = 0.25

Discrete distributions

Some examples of discrete distributions:

- Binomial
- Poisson
- Negative binomial
- Hypergeometric

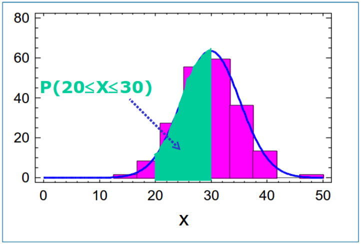
Continuous distributions

Continuous distribution

A continuous probability distribution is a probability distribution that can take values within all values of an interval

Example of continuous probability distribution

X = Time (in minutes) to get home from work every day = {15.5minutes, 20.12minutes, 17.6 minutes, ...}



$P(20 \leq X \leq 30)$ = P(taking a time between 20 and 30 minutes to go home from work) = 0.45

Continuous distributions

Some examples of continuous distributions:

- Uniform
- Exponential
- Normal
- Student's t

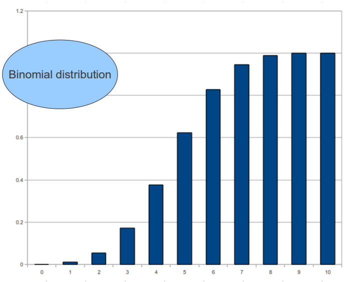
Cumulative distribution

Cumulative distribution

The cumulative distribution function (CDF), or just distribution function, describes the probability that a random variable X with a given probability distribution will be found at a value less than or equal to x.

Example of discrete cumulative distribution

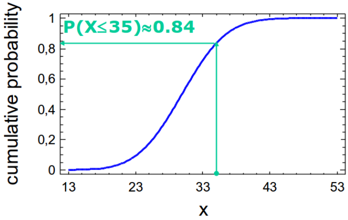
X = {Number of tails obtained when tossing a coin 10 times} = {0, 1, 2, ..., 10}



$F(2) = P(X \leq 2) = P(\text{2 tails or less in 10 throwings}) = P(X = 0) + P(X = 1) + P(X = 2) = 0.001 + 0.01 + 0.044 = 0.55$

Example of continuous cumulative distribution

X = Time (in minutes) to get home from work every day = {15.5minutes, 20.12minutes, 17.6 minutes, ...}



F(35) = P(X ≤ 35) = P(Taking less than or equal to 35 minutes to go home from work) = 0.84

Outline

Hypothesis testing

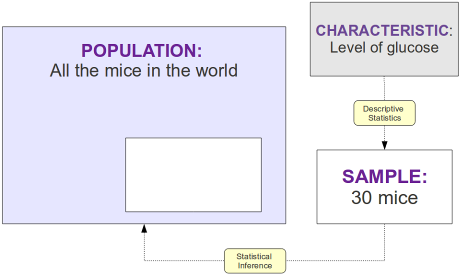
Hypothesis Testing

A statistical hypothesis test is a method to make decisions using data, whether from a controlled experiment or an observational study (not controlled)

Hypothesis testing

Hypothesis Testing

A statistical hypothesis test is a method to make decisions using data, whether from a controlled experiment or an observational study (not controlled)



Hypothesis Testing

Steps

1. Hypothesis about the population
2. Random sample
3. Summarizing the information (statistic)
4. Does the information given by the sample support the hypothesis? Are we making any error?

Hypothesis Testing

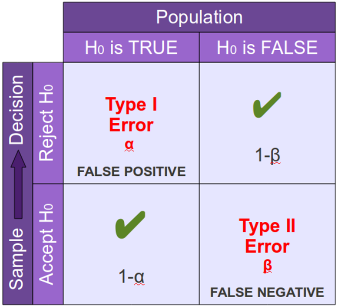
Steps

1. Hypothesis about the population
2. Random sample
3. Summarizing the information (statistic)
4. Does the information given by the sample support the hypothesis? Are we making any error?

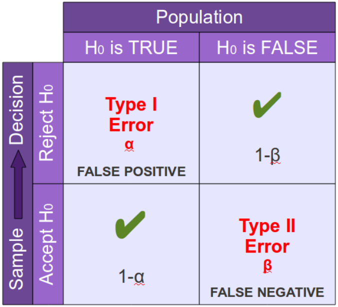
Decision rule

H_0: Null hypothesis
H_1: Alternative hypothesis

Errors on Hypothesis Testing



Errors on Hypothesis Testing



Significance level $= \alpha =$ P(Type I error) $=$ P(Rejecting H_0 when H_0 is TRUE)

Confidence level $= 1 - \alpha$

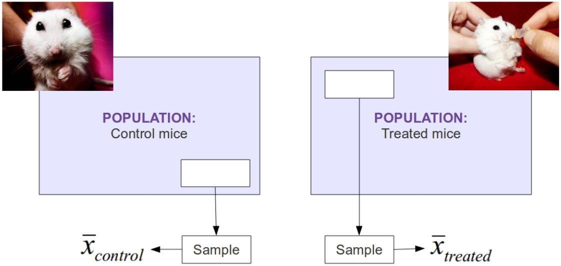
$\beta =$ P(Type II error) $=$ P(Failing to reject H_0 when H_0 is FALSE)

Power $= 1 - \beta =$ P(Rejecting H_0 when H_0 is FALSE)

Example of a Hyphothesis Test

We are comparing the average level of glucose of two groups of mice. The first one is the control and the second one has been treated with a drug.

$$¿\mu_{control} = \mu_{treated}?$$

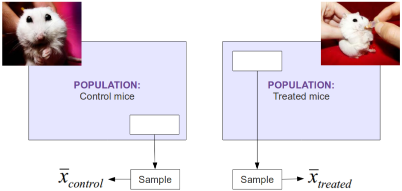


t-test (two class comparation)

t-test

A t-test is a parametric testing for comparing means between two groups. In this type of test, the statistic has a Student's t distribution if the null hypothesis is true.

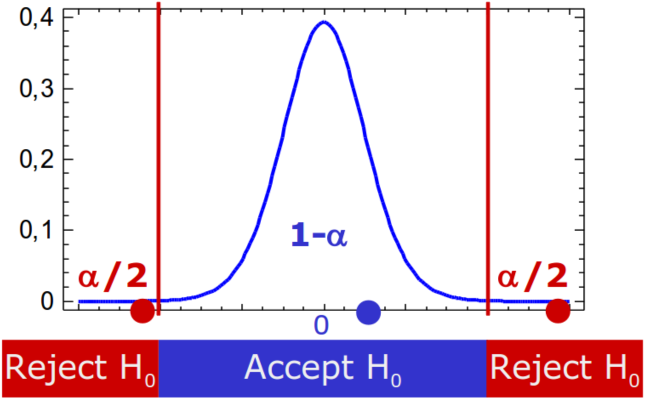
t-test example



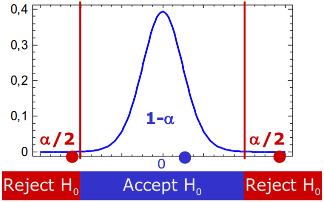
$$H_0 = \mu_{control} = \mu_{treated}$$
$$H_1 = \mu_{control} \neq \mu_{treated}$$

t-statistic: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \sim$ Student's t (if H_0 is TRUE)

t-test



p-value



- A result is called significant if it is unlikely to have ocurred by chance
- The p-value mesures the significance of a result
- The smaller the p-value is, the more significant the result is said to be

Outline

1. Introduction to Statistics
2. Hypothesis Testing
3. Parametric and non-parametric tests
   - Definitions
   - Some non-parametric tests
4. Multiple testing

Parametric and non-parametric tests

Parametric tests

It is assumed that the data are sampled from a population that follows a known probability distribution.

Parametric and non-parametric tests

Parametric tests

It is assumed that the data are sampled from a population that follows a known probability distribution.

- t-test
- ANOVA

Parametric and non-parametric tests

Parametric tests

It is assumed that the data are sampled from a population that follows a known probability distribution.

- t-test
- ANOVA

Non-Parametric tests

No assumptions about the population probability distribution

Parametric and non-parametric tests

Parametric tests

It is assumed that the data are sampled from a population that follows a known probability distribution.

- t-test
- ANOVA

Non-Parametric tests

No assumptions about the population probability distribution

- Wilcoxon
- Kruskal-Wallis
- Kolmorov-Smirnov
- Fisher's exact test

Wilcoxon test

Wilcoxon test

The Wilcoxon test involves comparisons of differences between measurements, so it requires that the data are measured at an interval level of measurement.

Kolmogorov-Smirnov test

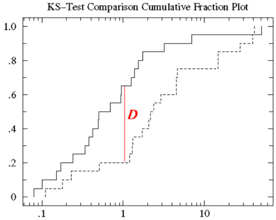
K-S test

Kolmogorov-Smirnov test is a nonparametric test for determining whether two underlying distributions differ, or whether an underlying probability distribution differs from a hypothesized distribution.

Example of a Kolmogorov-Smirnov test

We are testing two sets of data have the same probability distribution.

H_0: Same probability distribution
H_1: Different probability distribution



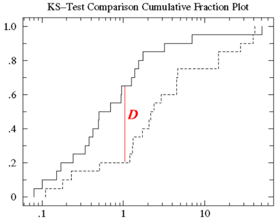
For each group:

1. Ranking data from the smallest to the largest
2. Calculating cumulative probability
3. Comparing them

Example of a Kolmogorov-Smirnov test

We are testing two sets of data have the same probability distribution.

H_0: Same probability distribution
H_1: Different probability distribution



- If H_0 is TRUE, the D (K-S distance) will be small
- The smaller the p-value is, the larger the distance between the two distributions will be.

Fisher's exact test

Fisher's exact test

Fisher's exact test is used to determine whether there is any relationship between two categorical variables (with two levels).

Example of a Fisher's exact test

Does a GO term appear with more frequency in list1 or in list2 or the frequency is more or less the same for both lists?

H_0: GO and gene lists are independent variables

		GO:0006950		
		Yes	No	Total
Genes list	List1	20 (67%)	10(33%)	30
	List2	20(29%)	50(71%)	70
	Total	40	60	100

Example of a Fisher's exact test

		GO:0006950		
		Yes	No	Total
Genes list	List1	20	10	30
	List2	20	50	70
	Total	40	60	100

		GO:0006950		
		Yes	No	Total
Genes list	List1	21	9	30
	List2	19	51	70
	Total	40	60	100

... and so on ...

Hypergeometric distribution

$$p = \frac{30!70!40!60!}{100!20!10!20!50!}$$

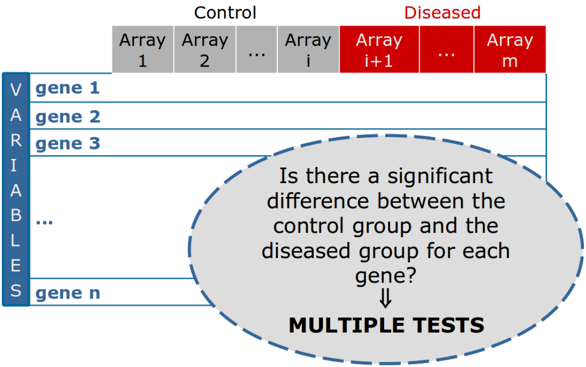
$$\sum$$

$$p = \frac{30!70!40!60!}{100!21!9!19!51!}$$

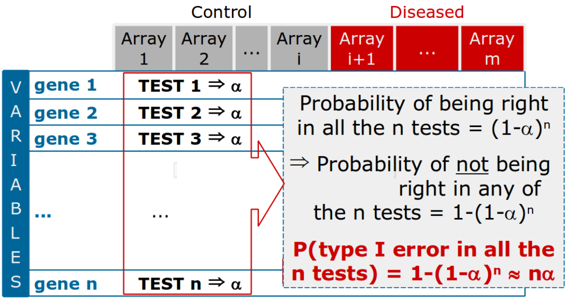
p-value = 0.00068 → Reject H_0

Outline

Multiple test



Example

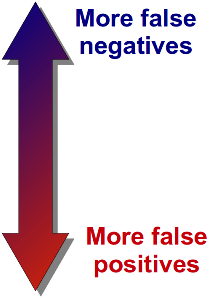


Multiple test corrections adjust p-values derived from statistical tests to correct for ocurrence of false positives. The number of this false positives increases as the number of test incresases

Type of error control

	H_0 not rejected	H_0 rejected	Total
H_0 true	**U**	**V** Type I Error	**n_0**
H_0 false	**T** Type II Error	**S**	**$n-n_0$**
Total	**n-R**	**R**	**n**

- **FWER** (Family-wise error rate) $= P(V > 0)$: Probability to reject one hypothesis by mistake is not more than α
- **FDR** (False discovery rate) $= E(V/R)$: Expected proportion of type I errors among the rejected hypothesis.



Bonferroni method

Bonferroni method

Is a solution for the problem of multiple testing. It reduces the allowable error α (p-value cutoff) for each test, dividing α by the number of tests n ($\frac{\alpha}{n}$). The resulting overall error does not exceed the desired limit.

Bonferroni method

Bonferroni method

Is a solution for the problem of multiple testing. It reduces the allowable error α (p-value cutoff) for each test, dividing α by the number of tests n ($\frac{\alpha}{n}$). The resulting overall error does not exceed the desired limit.

EXAMPLE:

To obtain $\alpha = 0{,}05$ with $n = 10$ test, take the p-value cutoff $\alpha = \frac{0{,}05}{10} = 0{,}005$ and the overall α will not exceed 0.05

Adjusted p-value

Adjusted p-value

Multiple testing correction adjusts the individual p-value of each test to keep the overall error rate (or false positive rate) to less than or equal to the user-specified p-value cutoff.

Adjusted p-value

Adjusted p-value

Multiple testing correction adjusts the individual p-value of each test to keep the overall error rate (or false positive rate) to less than or equal to the user-specified p-value cutoff.

EXAMPLE:

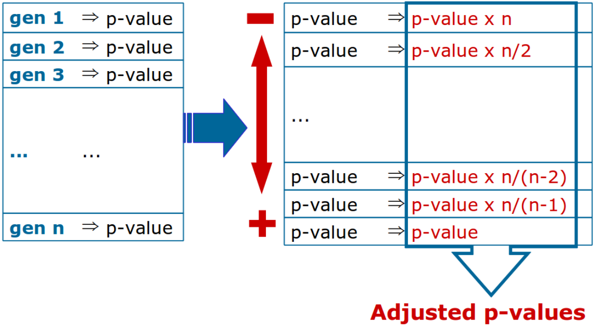
Bonferroni method:

For each test:

Adjusted p-value = p-value x n

BUT... Bonferroni method raises the number of false negatives and fails to identify significant differences in the data.

Benjamini & Hochberg



Some interesting links

- http://stattreck.com
- http://en.frestatistics.info
- http://www.statsoft.com/textbook
- htpp://udel.edu/∼mcdonald/statintro.html
- http://www.aiaccess.net/e_gm.html
- htpp://www.onlinestatbook.com/rvls.html