Biostatistics

Th4

Experimental Design

Carina Silva
(*carina.silva@estesl.ipl.pt*)
Mestrado em Farmácia

Contents

Contents

Contents

Contents

Introduction

- Statistics starts with a problem, continues with the collection of data, proceeds with the data analysis and finishes with conclusions.

Introduction

- Statistics starts with a problem, continues with the collection of data, proceeds with the data analysis and finishes with conclusions.
- To formulate the problem correctly you must understand the biological problem.

Introduction

- Statistics starts with a problem, continues with the collection of data, proceeds with the data analysis and finishes with conclusions.
- To formulate the problem correctly you must understand the biological problem.
- Understand the objective. You'll almost always find something, but that something may just be a coincidence.

Introduction

- Statistics starts with a problem, continues with the collection of data, proceeds with the data analysis and finishes with conclusions.
- To formulate the problem correctly you must understand the biological problem.
- Understand the objective. You'll almost always find something, but that something may just be a coincidence.
- Put the problem into statistical terms. This is a challenging step and where irreparable errors are sometimes made. Once the problem is translated into the language of statistics, the solution is often routine.

Statistical models

- Regression analysis is used for explaining or modeling the relationship between a single variable Y, called the *response*, *output* or *dependent variable*, and one or more *predictor*, *input*, *independent* or *explanatory variables*, $X_1, ..., X_p$.

Statistical models

- Regression analysis is used for explaining or modeling the relationship between a single variable Y, called the *response*, *output* or *dependent variable*, and one or more *predictor*, *input*, *independent* or *explanatory variables*, $X_1, ..., X_p$.
- The response must be a continuous variable but the explanatory variables can be continuous, discrete or categorical.

Statistical models

- Regression analysis is used for explaining or modeling the relationship between a single variable Y, called the *response*, *output* or *dependent variable*, and one or more *predictor*, *input*, *independent* or *explanatory variables*, $X_1, ..., X_p$.
- The response must be a continuous variable but the explanatory variables can be continuous, discrete or categorical.
- When we have continuous explanatory variables, we have a *regression analysis*.

Statistical models

- Regression analysis is used for explaining or modeling the relationship between a single variable Y, called the *response*, *output* or *dependent variable*, and one or more *predictor*, *input*, *independent* or *explanatory variables*, $X_1, ..., X_p$.
- The response must be a continuous variable but the explanatory variables can be continuous, discrete or categorical.
- When we have continuous explanatory variables, we have a *regression analysis*.
- When we have continuous and categorical explanatory variables we have *Analysis of Covariance*.

Statistical models

- Regression analysis is used for explaining or modeling the relationship between a single variable Y, called the *response*, *output* or *dependent variable*, and one or more *predictor*, *input*, *independent* or *explanatory variables*, $X_1, ..., X_p$.
- The response must be a continuous variable but the explanatory variables can be continuous, discrete or categorical.
- When we have continuous explanatory variables, we have a *regression analysis*.
- When we have continuous and categorical explanatory variables we have *Analysis of Covariance*.
- When we have just qualitative predictors (factors), we have to perform *Analysis of Variance* (ANOVA).

Statistical models

- Regression analysis is used for explaining or modeling the relationship between a single variable Y, called the *response*, *output* or *dependent variable*, and one or more *predictor*, *input*, *independent* or *explanatory variables*, $X_1, ..., X_p$.
- The response must be a continuous variable but the explanatory variables can be continuous, discrete or categorical.
- When we have continuous explanatory variables, we have a *regression analysis*.
- When we have continuous and categorical explanatory variables we have *Analysis of Covariance*.
- When we have just qualitative predictors (factors), we have to perform *Analysis of Variance* (ANOVA).
- When the responde variable is qualitative we can use *Logistic Regression*.

Introduction

- The name ANOVA stands for Analysis of Variance and is used because the original thinking was to try to partition the overall variance in the response to each of the factors and the error.

Introduction

- The name ANOVA stands for Analysis of Variance and is used because the original thinking was to try to partition the overall variance in the response to each of the factors and the error.
- Can be considered a generalization of the *t*-test for two independent samples.

Introduction

- The name ANOVA stands for Analysis of Variance and is used because the original thinking was to try to partition the overall variance in the response to each of the factors and the error.
- Can be considered a generalization of the t-test for two independent samples.
- Predictors (qualitative variables) are now typically called factors which have some number of levels (treatments).

Introduction

- The name ANOVA stands for Analysis of Variance and is used because the original thinking was to try to partition the overall variance in the response to each of the factors and the error.
- Can be considered a generalization of the t-test for two independent samples.
- Predictors (qualitative variables) are now typically called factors which have some number of levels (treatments).
- The parameters are now often called effects.

The model

- ANOVA tests the null hypothesis that the population means are all equal. The alternative is that they are not all equal. This alternative could be true because all of the means are different or simply because one of them differs from the rest.

$$H_0 : \mu_1 = ... = \mu_p \quad vs. \quad H_1 : \exists'(i,j) : \mu_i \neq \mu_j \text{(some i,j=1,...,p; i}\neq\text{j)}$$

The model

- ANOVA tests the null hypothesis that the population means are all equal. The alternative is that they are not all equal. This alternative could be true because all of the means are different or simply because one of them differs from the rest.

$$H_0 : \mu_1 = ... = \mu_p \quad vs. \quad H_1 : \exists'(i,j) : \mu_i \neq \mu_j \text{(some i,j=1,...,p; i}\neq\text{j)}$$

- If we reject the null hypothesis, we need to perform some further analysis to draw conclusions about which population means differ from which others.

The model

- ANOVA tests the null hypothesis that the population means are all equal. The alternative is that they are not all equal. This alternative could be true because all of the means are different or simply because one of them differs from the rest.

$$H_0 : \mu_1 = ... = \mu_p \quad vs. \quad H_1 : \exists'(i,j) : \mu_i \neq \mu_j \text{(some i,j=1,...,p; i}\neq\text{j)}$$

- If we reject the null hypothesis, we need to perform some further analysis to draw conclusions about which population means differ from which others.
- The one-way ANOVA model is
  $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, for $i = 1, ..., p$ and $j = 1, ..., n_i$.

The model

- ANOVA tests the null hypothesis that the population means are all equal. The alternative is that they are not all equal. This alternative could be true because all of the means are different or simply because one of them differs from the rest.

$$H_0 : \mu_1 = ... = \mu_p \quad vs. \quad H_1 : \exists'(i,j) : \mu_i \neq \mu_j \text{(some i,j=1,...,p; i}\neq\text{j)}$$

- If we reject the null hypothesis, we need to perform some further analysis to draw conclusions about which population means differ from which others.
- The one-way ANOVA model is $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, for $i = 1, ..., p$ and $j = 1, ..., n_i$.
- The ε_{ij} are assumed to be $N(0, \sigma)$ distributed. The parameters of the model are: p population means (μ_i) and the common standard deviation σ.

Assumptions

- The sample sizes n_i my differ, but the standard deviation is assumed to be the same in all of the populations.

Assumptions

- The sample sizes n_i my differ, but the standard deviation is assumed to be the same in all of the populations.
- Rule: If the largest sample standard deviation is less than twice the smallest sample standard deviation, we can use methods based on the condition that the population standard deviations are equal and our results will still be approximately correct.

Assumptions

- The sample sizes n_i my differ, but the standard deviation is assumed to be the same in all of the populations.
- Rule: If the largest sample standard deviation is less than twice the smallest sample standard deviation, we can use methods based on the condition that the population standard deviations are equal and our results will still be approximately correct.
- The response variable must be Normally distributed in each level of the factor (treatments).

Assumptions

- The sample sizes n_i my differ, but the standard deviation is assumed to be the same in all of the populations.
- Rule: If the largest sample standard deviation is less than twice the smallest sample standard deviation, we can use methods based on the condition that the population standard deviations are equal and our results will still be approximately correct.
- The response variable must be Normally distributed in each level of the factor (treatments).
- Independence of the responde variable in each factor level.

Example 1

Many studies have suggested that there is a link between exercise and healthy bones. Exercise stresses the bones and this causes them to get stronger. One study examined the effect of jumping on the bone density of growing rats. There were three treatments: a control with no jumping, a low-jump condition (the jump height was 30 centimeters), and a high-jump condition (60 centimeters). After 8 weeks of 10 jumps per day, 5 days per week, the bone density of the rats (expressed in mg/cm) was measured. Here are the data:

Group	Bone density (mg/cm)									
Control	611	621	614	593	593	653	600	554	603	569
Low jump	635	605	638	594	599	632	631	588	607	596
High jump	650	622	626	626	631	622	643	674	643	650

The ANOVA table

Source of	df	SS	MS	*F*-ratio
Between groups	$p-1$	$\sum_{i=1}^{p} n_i(\bar{y}_i-\bar{y})^2$	$\frac{\sum_{i=1}^{p} n_i(\bar{y}_i-\bar{y})^2}{p-1}$	$\frac{MS_{Groups}}{MS_{Residual}}$
Residual	$\sum_{i=1}^{p} n_i - p$	$\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2$	$\frac{\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2}{\sum_{i=1}^{n} n_i - p}$	
Total	$\sum_{i=1}^{p} n_i - 1$	$\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij}-\bar{y})^2$		

The ANOVA table

- Sum of squares between groups: is the sum of all the squared differences **between** the individual means and the overall mean.
- Sum of squares ´within groups or residuals: is the sum of all the squared differences between individual ´data and the group mean within each group.
- Mean squares: is the result by dividing the sum of squares by its d.f. This is then a measure of the average deviation of individual values from their respective mean, since the df is about the same as the number of terms in the sum.
- F-ratio: Ratio of the two mean squares. When you publish the F statistic, should be written as $F_{2,27}$ or F(2,27).

Nonparametric ANOVA Analysis

When the ANOVA assumptions fails:

- Kruskall-Wallis test (or ANOVA by ranks)

Exercise 2

An entomologist is studying the vertical distribution of a fly species in a deciduous forest and obtains five collections of the flies in three different vegetation layers: herb, shrub and tree. The entomologist wants to test is the abundance of the flies is the same in all three vegetation layers.

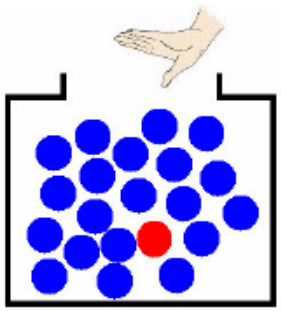
Herbs	14	12.1	9.6	8.2	10.2
Shrubs	8.4	5.1	5.5	6.6	6.3
Trees	6.9	7.3	5.8	4.1	5.4

Multiple Comparisons

If the null hypothesis is rejected, there are differences between the treatment means, but exactly which means differ is not specified. Two situations could occur:

- Post-hoc comparisons: for further exploration of the data after a significant effect has been found.
- Planned comparisons: hypothesis specified before the analysis commences (orthogonal comparisons).

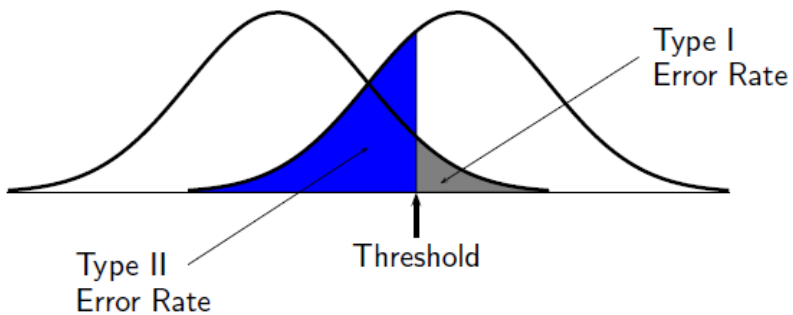
The multiple comparison problem



- Imagine a solution with 20 spheres: 19 are blue and 1 is red. What are de odds of randomly sampling the red sphere by chance? It is 1 out of 20.
- Now let's say that you get to sample a single sphere (and put it back into the solution) 20 times. Have a much higher chance to sample the red sphere. This is exactly what happens when testing several thousand tests at the same time.

One Test, One Threshold

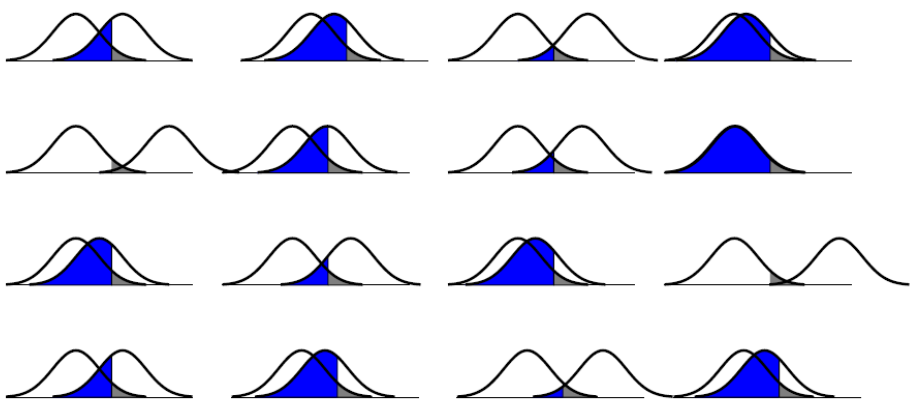
With a single hypothesis test, we choose a rejection threshold to control the Type I error rate,



while achieving a desirable Type II error rate for relevant alternatives.

The multiple comparison problem

Many Tests, One Threshold With Multiple tests, the problem is more complicated.



Each test has possible Type I and Type II errors, and there are many possible ways to combine them. The probability of a Type I error grows with the number of tests.

The Multiple Testing Problem

- Perform m simultaneous hypothesis tests with a common procedure.
- For any given procedure, classify the results as follows:

	Rejected H_0	Not Rejected H_0	
True H_0	V	U	m_0
False H_0	S	T	m_1
	R	$m - R$	m

All quantities except m, m-R, and R are unobserved.

- The problem is to choose a procedure that balances the competing demands of sensitivity and specificity.

- V: r.v. which represents the number of false positive features, $P(V \geq 1) = 1 - P(V = 0) = 1 - (1 - \alpha)^n$, where α is the probability of rejecting the null hypothesis when it is true (*Type I Error*), $P(\text{Rej } H_0 | H_0 \text{ True})$.

Number of hypothesis tested (n)	False positives incidence ($n \times \alpha$)	Probability of 1 or more false positives by chance ($1 - (1 - 0.05)^n$)
1	$1/20 = 0.05$	0.050
2	$2 \times (1/20) = 0.1$	0.098
20	$20 \times (1/20) = 1$	0.642
100	$100 \times (1/20) = 5$	0.994

Problem: When many hypotheses are tested, the probability of a type I error (false positive) increases sharply with the number of hypotheses.

FWER method: Single-step approach: Bonferroni

- FWER: Family-wise error rate: the probability of at least one type I error, FWER=P($V \geq 1$) and guarantees P(V $>$ 0)$\leq \alpha$.

- The Bonferroni's Method is the simplest way to achieve control of the FWER at any desired level α.
- Simply choose $c = \alpha/m$.
- With this value of c, the FWER will be no larger than α for any family of m tests.

Example 1:

- Suppose we conduct 5 tests and obtain the following p-values for tests 1 through 5.

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

- Which tests' null hypotheses will you reject if you wish to control the FWER at level 0.05?
- Use the Bonferroni's Method to answer this question.

Solution

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

$p_1 = 0.042 > 0.01$
$p_2 = 0.001 \leq 0.01$
$p_3 = 0.031 > 0.01$
$p_4 = 0.014 > 0.01$
$p_5 = 0.007 \leq 0.01$

The cutoff for significance is $c = 0.05/5 = 0.01$ using the Bonferroni's Method. Thus we would reject the null hypothesis for tests 2 and 5 with the Bonferroni's Method.

Some considerations

- FWER is appropriate when you want to guard against ANY false positives.
- However, in many cases (particularly in genomics) we can live with a certain number of false positives.
- FWER is to conservative because it depends on the overall number of tests (m).
- Holm's method is less conservative than the Bonferroni's Method.
- The methods will provide the same results for many data sets, but sometimes Holm's method will result in more rejected null hypotheses.

FDR – False Discovery Rate

In practice, however, many biologists seem willing to accept that some errors will occur, as long as this allows findings to be made. For example a researcher might consider acceptable a small proportion of errors (say 10%, 20%) between her findings. In this case, the researcher is expressing interest in controlling the **false discovery rate** (FDR).

- Unlike a significance level which is determined before looking at the data, FDR is a post data measure of confidence.
- FDR uses information available in the data to estimate the proportion of false positive results that have occurred.
- If one obtains a list of differentially expressed genes where the FDR is controlled at, say, 20%, one will expect that a 20% of these genes will represent false positive results.

The Benjamini and Hochberg Procedure for Strongly Controlling FDR at Level α

- Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ denote the m p-values ordered from smallest to largest.
- Find the **largest integer** k so that $p_{(k)} \leq \frac{k \times \alpha}{m}$.
- If no such k exists, set $c = 0$ (declare nothing significant).
- Otherwise set $c = p_{(k)}$ (reject the null hypotheses corresponding to the smallest k p-values).

Example 1 (cont.):

Test	1	2	3	4	5
p-value	0.042	0.001	0.031	0.014	0.007

$$p_{(1)} = 0.001 \leq 1 \times 0.05/5 = 0.01$$
$$p_{(2)} = 0.007 \leq 2 \times 0.05/5 = 0.02$$
$$p_{(3)} = 0.014 \leq 3 \times 0.05/5 = 0.03$$
$$p_{(4)} = 0.031 \leq 4 \times 0.05/5 = 0.04$$
$$p_{(5)} = 0.042 \leq 5 \times 0.05/5 = 0.05$$

The **Benjamini and Hochberg's Method** would reject the null hypotheses for all 5 tests. Here, $k = 5$ and $c = p_{(5)} = 0.042$.

Some final considerations

FWER vs FDR

- The decision of controlling FDR or FWER depends on the goals of the experiment.
- If the objective is *gene fishing*, allowing a certain number of false positives to be reasonable, then FDR is preferable.
- If instead one is working with a shorter number of hypotheses, in which we want to verify if some specific ones are significant, then FWER is the appropriate criteria.
- FDRs are more appropriate in large sets of hypotheses.
- There are several methods to perform multiple comparisons. Some of them too conservative (Bonferroni, Scheffé) and others too liberal like LSD method. Somewhere in the middle are the HSD (or Tukey test) or Holm Sidak test.

Remarks

- Which multiple tests correction should be used? As long as the conditions you have for the data meet with the assumptions in particular multiple tests corrections, use the one that gives the highest power. **Using an FDR method is common these days.**
- 5% (or 95% confidence) is a convention, not a magic number (same to 1% or 0.1%). If you do not have any particular reason to favour a particular threshold, use a convention.

Exercise 3

Consider the dataset of the exercise 1. Since the null hypothesis of the ANOVA was rejected, perform a multiple comparison analysis.

Two-Way ANOVA

Many experiments involve the study of the effects of two or more factors. By a factorial design, we mean that in each complete trial or replication of the experiment, all possible combinations of the levels of the factors are investigated.

- Factorial ANOVA with fixed effects: A fixed factor contains all levels of the factor of interest in the design.

Two-Way ANOVA

Many experiments involve the study of the effects of two or more factors. By a factorial design, we mean that in each complete trial or replication of the experiment, all possible combinations of the levels of the factors are investigated.

- Factorial ANOVA with fixed effects: A fixed factor contains all levels of the factor of interest in the design.
- Factorial ANOVA with random effects: A random factor contains only a sample of the possible levels of the factor, and the intent is to generalize to all other levels.

Two-Way ANOVA

Many experiments involve the study of the effects of two or more factors. By a factorial design, we mean that in each complete trial or replication of the experiment, all possible combinations of the levels of the factors are investigated.

- Factorial ANOVA with fixed effects: A fixed factor contains all levels of the factor of interest in the design.
- Factorial ANOVA with random effects: A random factor contains only a sample of the possible levels of the factor, and the intent is to generalize to all other levels.
- Factorial ANOVA with mixed effects: The model contains fixed and random factors.

Two-way ANOVA with fixed factors

The model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

The interaction effect $(\alpha\beta)_{ij}$ is interpreted as that part of the mean response not attributable to the additive effect of α_i and β_j. For example, you may enjoy strawberries and cream individually, but the combination is superior. In contrast, you may like fish and ice cream but not together.

Hypothesis

- Factors A and B are additive: $H0 : (\alpha\beta)_{ij} = 0$ vs $H1 : (\alpha\beta)_{ij} \neq 0$
- Factor A: $H0 : (\alpha)_i = 0$ vs $H1 : (\alpha)_i \neq 0$
- Factor B: $H0 : (\beta)_j = 0$ vs $H1 : (\beta)_j \neq 0$

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
FACTOR A	SSA	$a-1$	$MSA = \frac{SSA}{a-1}$	$F = \frac{MSA}{MSE}$
FACTOR B	SSB	$b-1$	$MSB = \frac{SSB}{b-1}$	$F = \frac{MSB}{MSE}$
INTERACTION AB	$SSAB$	$(a-1)(b-1)$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F = \frac{MSAB}{MSE}$
ERROR	SSE	$ab(r-1)$	$MSE = \frac{SSE}{ab(r-1)}$	
TOTAL	$TotalSS$	$abr-1$		

The Analysis of Variance Table for a 2-Factor Factorial Design

Assumptions

- Normality of the residuals.
- Homogeneity of the variances.
- Independency of observations.

Example

Various studies have shown that interethnic differences in drugmetabolizing enzymes exist. A study was conducted to determine whether differences exist in pharmacokinetics of the tricyclic antidepressant nortriptyline between Hispanics and Anglos (Gaviria *et al.*,1986). The study consisted of five males and five females from each ethnic group. We would like to determine whether ethnicity or sex differences exist, or whether there is an interaction between the two variables on the outcome variable total clearance (ml/min/kg).

- The model is $Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$
  - μ is the overall mean
  - α_i is the effect of i^{th} level of factor A (ethnicity: 1=Hispanic, 2=Anglo)
  - β_j is the effect of j^{th} level of factor B (sex: 1=female; 2=male)
  - $\alpha\beta_{ij}$ effect of interaction of the i^{th} level of ethnicity and the j^{th} level of sex.

Hispanics		Anglos	
Females	Males	Females	Males
10.5	5.4	7.1	5.7
8.3	7.1	10.8	3.8
8.5	6.1	12.3	7.8
6.4	10.8	7.0	4.4
6.5	4.1	7.9	9.9

Exercise

An investigator intends to study the final result of a certain chemical process. The two factors of interest are temperature and pressure. He chose three levels for each factor and randomly assigned 2 tests for each temperature and pressure level. The data are as follows:

Temp. \ Pressure	Type 1	Type 2	Type 3
Low	86.3 86.1	84 85.2	85.8 87.3
Medium	88.5 89.4	87.3 89.9	898 90.3
High	89.1 91.7	90.2 93.2	91.3 93.7