PhD notes II: Statistics with SPSS

Contents

PhD notes II: Statistics with SPSS

1 Normality tests

Population = ALL samples on Earth — Sample = OUR reduced group of measurements

Does my data follow a normal (parametric) or non-parametric distribution?

1.1 Kurtosis and Skewness

This option does not sort by ”type of tissue”, considers ALL tissues.

1. Analyze > Descriptive Statistics > Descriptives...

2. Options > Distribution: Check Kurtosis and Skewness

1.2 Shapiro-Wilk

Contrasts the normality of a data set.

1. Analyze > Descriptive Statistics > Explore...

2. Factor list: “tejido” (IF wanna sort by tissue)

3. Statistics > Descriptives: 95% > Plots > Check Normality plots with tests

4. At output: “Tests of Normality” chart

1) Null hypothesis: a given sample fit a normal distribution.

2) Output: 0<w<1; for small w, we reject the null hypothesis.

1.3 Kolmogorov-Smirnov

Quantifies the distance between the empiric distribution function and the cumulative (evaluated at ’x’): the probability of X to acquire a value $\leq$ than ’x’. It tests for normality as it compares the distance between both functions.

1. Analyze > Descriptive Statistics > Explore...

2. Factor list: “tejido” (IF wanna sort by tissue)

3. Statistics > Descriptives: 95% > Plots > Check Normality plots with tests

4. At output: “Tests of Normality” chart

Particular case of testing for normality → **Lilliefors**
1) $\bar{x}$ and σ
2) Find maximum discrepancy between empirical and cumulative function.

1.4 Q-Q plot

Graphical method for comparing two probability distributions by plotting their quantiles; ”¿Hasta qué valor está el 25 por ciento de mis datos?” Q_1(25%), Q_2(50%), Q_3(75%) and Q_4(100%).

1. Analyze > Descriptive Statistics > Explore...

2. Factor list: “tejido” (IF wanna sort by tissue)

3. Statistics > Descriptives: 95% > Plots > Check Normality plots with tests

4. At output: “Normal Q-Q plot” and ”Detrended Normal Q-Q plot”

+) q-quantiles → values that partition a finite set of values into ’q’ subsets of equal sizes.

+) point (x,y) → corresponds to 1 of the quantiles of the 2nd distribution (y) against the same quantile of the first distribution (x).

+) When 2 distributions:
* similar points → y=x line (normal distribution)
* different but linearly related → points are on a line but not necessary y=x.
* Atypical values: when one of the distributions is more skewed (distorted) than the other (heavier tails).

A Q-Q plot is a more powerful approach to do a comparison rather than using common histograms.

2 Statistical tools

2.1 Boxplot

To build a data filter so that the outlying and extreme values (points that can origin distortion on results) are identified and can be removed.

1. Analyze > Descriptive Statistics > Explore...

2. Factor list: "tejido" (IF wanna sort by tissue)

3. Statistics > Descriptives: 95% and Outliers > Plots > Normality plots with tests

Boxplot charts allow a visual interpretation of population distribution by splitting input data into five pieces of information: the minimum value, the first quartile, the median value, the third quartile and the maximum value, in such a way that 50% of the information is arranged into a box and the remaining data is plotted in whiskers. Particularly, the length of the box and its whiskers gives information about the data scattering meanwhile its symmetry characterizes data distribution tendency.

When the median of one box (i.e., a given tissue) does not fit within the boxes of the other tissues, the analyzed M-metric has the potential to differentiate such tissue by respect to the others.

Boxplot chart also represents the outlier values as individual spots. The mild (extreme) outlier values are placed 1.5 times (3 times) the length of the box far from the median and are represented by circles (stars). The values that differ a lot from the distribution tendency are typically removed as they are associated with experimental errors in the measurements.

2.2 Kruskal-Wallis

Non-parametric method for testing whether samples originate from the same distribution.

1. Analyze > Non parametric tests > Independent samples

2. Objective > Customize analysis

3. Field > Test fields (all 27 metrics); Groups > "tejido"

4. Settings > Test options > Significance lvl 0.05; Confidence interval 95%

It is used to determine if there are significant statistically differences between two or more data distributions, i.e., if two or more data groups have the same distribution origin. Particularly, it defines a null hypothesis that all the studied data comes from the same probability distribution and compares the mean value of each M-metrics for a given tissue data with the mean of the three remaining tissues. In other words, this test permits us to know if the M-metrics distribution for all tissues can be separated into four well-differentiated groups: bone, tendon, muscle and myotendinous junction tissue. This test outputs two indicators: the significance and the chi-square approximation (redundancy, choose only one). On one hand, if the significance value is smaller than 0.05, the null hypothesis is rejected so we can assume with a statistical confidence of 95% that our input data derives from different probability distribution groups, which would be the ideal case in order to have discriminatory potential. On the other hand, large chi-square values result into which M-metrics provide a better tissue differentiation.

2.3 ROC curve

Statistical technique used to describe the performance of a classifier (an algorithm or a particular variable) when classifying measures into two categories, by plotting the true positive rate (TPR), or sensitivity, against the false positive rate (FPR), or 1-specificity, for multiple threshold value.

1. Analyze > Classify > ROC Analysis
2. Test variable: prob_tendo
3. State variable: Tendo; Value of state variable: 1
4. Options... > Include cutoff value for positive classification
5. Test direction > Larger test result indicates more positive test
6. Distribution assumption > Non parametric
7. Display... > Plot: ROC curve, with diagonal reference line
8. Print: Standard error and confidence level and Classifier evaluation metrics
9. Output: "Classifier evaluation metrics": K-S Statistics → Cutoff (Youden's Index).

For each classifier and measure, there are four possible outcomes depending on what the sample actually is (real value) and how is classified by the model (prediction). Such classification is represented in the so-called confusion matrix. In a ROC curve plot, several points are of interest, for instance, the lower left point (0, 0) and the upper right point (1, 1). Those points represent a classifier never issuing a positive classification, which means that will never commit false positive errors but neither true positives, and the opposite one, always issuing positive classifications. Consequently, both thresholds lack of predictive information, just as any point in the diagonal line y=x connecting them (see diagonal red line in Fig. 4). Intuitively, the perfect point corresponds to (0, 1), the upper left point, which represent the perfect classification with the maximum sensitivity and specificity. Usually, to compare the performance of classifiers (in our case, to estimate which component provides better sensitivity-specificity values for a particular tissue), the area under the ROC curve (AUC) is calculated, with values ranging from 0, when the variable has no predictive capability, to 1, with 100% both sensitivity and specificity.

3 Principal Components Analysis

The basic assumption of factor analysis is that for a collection of observed variables there are a set of **underlying or latent variables called factors** (smaller than the number of observed variables), that can explain the interrelationships among those variables. The goal is to reduce the number of variables to explain and to interpret the results.

1. Analyze > Dimension reduction > Factor... > Factor analysis box

2. Transfer the variables you want to include (Vermell PA,..., Blau Phi) into Variables box by using the arrow button

3. Click on Descriptives button > Factor analysis: Descriptives box

4. -Statistics- already selected Initial solution; check Univariate descriptives; and at -Correlation matrix-: Coefficients, Inverse, Significance levels, Determinant, Reproduced, Anti-image and KMO and Bartlett's test of sphericity > Continue > Factor analysis box

5. Click on Extraction > Factor Analysis: Extraction box; Max iterations: 25; method: Principal Components

6. Keep defaults but also select: -Analyze- Correlation matrix; -Display- Unrotated factor solution and Scree plot; -Extract- Fixed number of factors to extract (10) > Continue > Factor Analysis box

7. Click on Rotation > Factor Analysis: Rotation box; Max iterations: 25

8. In -Method-, select None > -Display- Rotated solution and Loading plot(s) > Continue > Factor Analysis box

9. Click on Scores > Factor Analysis: Factor Scores

10. Check Save as variables > Bartlett > Display factor score coefficient matrix > Continue > Factor Analysis box

11. Options... > Factor Analysis: Options box > Check -Missing values- Exclude cases pairwise and -Coefficient display format- Sorted by size > Continue > OK

+A) Eigenvalues represent the total amount of variance that can be explained by a given principal component. They can be positive or negative in theory, but in practice they explain variance which is always positive. **If eigenvalues are greater than zero, then it's a good sign.** Since variance cannot be negative, negative eigenvalues imply the model is ill-conditioned. Eigenvalues close to zero imply there is item multicollinearity, since all the variance can be taken up by the first component. Eigenvalues are also **the sum of squared component loadings across all items for each component, which represent the amount of variance in each item that can be explained by the principal component**.

+B) Eigenvectors represent a weight for each eigenvalue. The eigenvector times the square root of the eigenvalue gives the component loadings which can be interpreted as the correlation of each item with the principal component.

+C) Component Matrix: The components can be interpreted as the correlation of each item with the component. Each item has a loading corresponding to each of the other components. The square of each loading represents the proportion of variance (think of it as an r^2 statistic) explained by a particular component. If you keep going on adding the squared loadings cumulatively down the components, you find that it sums to 1 or 100%. This is also known as the communality, **and in a PCA the communality for each item is equal to the total variance.**

+D) Total variance explained: Recall that the **eigenvalue represents the total amount of variance that can be explained by a given principal component**. Starting from the first component, each subsequent component is obtained from partialling out the previous component. Therefore **the first component explains the most variance, and the last component explains the least**. Looking at the Total Variance Explained table, you will get the total variance explained by each component. **If we extract the same**

number of components as the number of items, the Initial Eigenvalues column will be the same as the Extraction Sums of Squared Loadings column.

3.1 Kaiser-Meyer-Olkin

Measure of Sample Adequacy:
1) KMO values between 0.8 and 1 indicate the sampling is adequate.

2) KMO values less than 0.6 indicate the sampling is not adequate and that remedial action should be taken.

3) KMO Values close to zero means that there are large partial correlations compared to the sum of correlations.

3.2 Bartlett's test of Sphericity

Testing the difference between the correlation matrix with the identity matrix.
Sig (p-value)<0.001

4 Binary Logistic Regression model

+ The regression does not necessarily converge if a non-orthogonal basis is being used. We previously perform a Principal Components Analysis to build and orthogonal PC basis in a 10 dimension-space. +

1. Analyze > Regression... > Binary Logistic...

2. Dependent: ”Muscul”

3. Save... > Predicted values > Probabilities

4. Options... > Statistics and Plots > Classification plots, Hosmer-Lemeshow goodness-of-fit, Iteration history

5. Probability for Stepwise > Entry: 0,05; Removal: 0,10

6. Classification cutoff: change the value depending on the corresponding Youden's Index; Maximum iterations: 20

7. Include constant in model

8. Covariates: C1, C2, C3, C4, C5, C6, C7, C8, C9, C10

9. Method: Backward Wald

10. At output NEW VARIABLE ARISES: ”PRE 1” → corresponds to ”prob muscul”.

11. At output: ”Model Summary” table → Cox and Snell + Nagelkerke R-squared values.

We have chosen a logistic function because it is valid for non-parametric data and because it does not require the relation between the predictors and the probability of a target outcome to be linear, neither the residuals to be normally distributed and exhibit constant variance, as it is the case, for instance, of other fitting methods as those based on the ordinary least square regression (OLS). Particularly, we have previously verified the non-dependency of the model predictors (principal components) and their connection with the dichotomic dependent variable. Moreover, the multicollinearity and interaction between predictors is removed so stable predictive models and convergent iteration processes are achieved.

By applying a **stepwise regression approach** (backward elimination) and using **Wald estimator** throughout multiple steps, the principal components are removed until only the most significant ones remain. What is more, **Hosmer-Lemeshow** test is computed at each step.

To illustrate the goodness-of-fit or explanatory capacity of the four predictive models, the **R-squared of Nagelkerke and R-squared of Cox and Snell** indicators are computed. In this way, because of the predictive model efficacy is analyzed as of its particular classificatory table (sensitivity and specificity), a threshold is required to establish whether a given tissue is or not of a particular category. As previously done with the principal component values, such threshold is determined by the **Youden’s Index** of the ROC curve associated with the probability sampling distribution of each model (type of tissue): the optimal cut-off corresponds to the farthest point from the ROC diagonal (the maximum value of the ratio sensitivity + specificity – 1). In particular, the ROC curve was computed for each of the four logistic regression functions with the aim of ensuring the good performance of the designed algorithm.

4.1 R-squared values: pseudo or not?

As a starting point, recall that a non-pseudo R-squared is a statistic generated in ordinary least squares (OLS) regression that is often used as a goodness-of-fit measure.

When analyzing data with a logistic regression, an equivalent statistic to R-squared does not exist. The model estimates from a logistic regression are maximum likelihood estimates arrived at through an iterative process. They are not calculated to minimize variance, so the OLS approach to goodness-of-fit does not apply. However, to evaluate the goodness-of-fit of logistic models, several pseudo R-squareds have been developed. These are "pseudo" R-squareds because they look like R-squared in the sense that they are on a similar scale, ranging from 0 to 1 (though some pseudo R-squareds never achieve 0 or 1) with higher values indicating better model fit, but they cannot be interpreted as one would interpret an OLS R-squared and different pseudo R-squareds can arrive at very different values. Note that most software packages report the natural logarithm of the likelihood due to floating point precision problems that more commonly arise with raw likelihoods.

References:
1) Cox, D. R., and E. J. Snell. 1989. The Analysis of Binary Data, 2nd ed. London: Chapman and Hall.
2) Nagelkerke, N. J. D. 1991. A note on the general definition of the coefficient of determination. Biometrika, 78:3, 691-692.

- R-squared Nagelkerke:It adjusts Cox and Snell's so that the range of possible values extends to 1. **Then, if the full model perfectly predicts the outcome and has a likelihood of 1, Nagelkerke R-squared = 1**.

- R-squared Cox and Snell: the ratio of the likelihoods reflects the improvement of the full model over the intercept model (**the smaller the ratio, the greater the improvement**). Note that Cox and Snell's pseudo R-squared has a maximum value that is not 1: if the full model predicts the outcome perfectly and has a likelihood of 1, Cox and Snell's is then less than one.