Hi, we are Helena Yaben and Pablo Laso and we are going to talk about our approach during inferential analysis for malignancy detection in the breast cancer wisconsin dataset.

CHANGE

Breast cancer is the most common cancer diagnosed in women across the world. It accounts for 1 in 4 cancer cases for female, being the leading cause of death from cancer in women.

Currently, biopsy of breast constitutes the most effective technique to distinguish between malignant and benignant breast masses. Analysis of the nuclei characteristics can determine the nature of the tissue since malignant nuclei typically present distinctive variations in size, shape, or texture when compared to non-malignant nuclei.

CHANGE

Through the analysis of the Breast Cancer Wisconsin diagnostic dataset, the aim of this project is to determine if there exist differences between the malignant and benignant populations for each of the nuclei characteristics recorded. This way, we could identify those features that should be paid a special attention during the interpretation of histopathological analysis,

CHANGE

This dataset collects information about samples of tissue obtained from breast tumors by means of a technique known as Fine-Needle-Aspiration. The nuclei present in each tissue sample are further analyzed and a set of features is obtained. Finally, mean, standard error, and worst values for these features configure a dataset of 30 features.

CHANGE

Diagnosis for each sample is also annotated, constituting the target variable. The 30 features constitute the predictors for the target variable.

CHANGE

All predictors are numerical and continuous, whereas diagnosis is categorical and nominal. To further analyze our dataset, we converted feature diagnosis into numerical and discrete in a way that 0 represented non-malignancy and 1 represented malignancy. Finally, feature “ID” was dropped for further analysis.

CHANGE

Classes are imbalanced in this dataset, being the class benignant the majority class.

CHANGE

By plotting malignant versus benignant histograms for each feature, we noticed that malignant values were, in general, higher than those for benignant samples.

CHANGE LASO

By analyzing correlation matrix, we saw that, in general, mean and worst features had a great correlation with target variable. We can also saw that features were quite correlated with each other, which means that some of these features may be redundant when determining the diagnosis of a sample.

CHANGE
By scatter plotting pairs of features, we can see that just two features allow to perfectly distinguish two groups, and that only one feature can be used to predict diagnosis by means of models such as Logistic Regression.

CHANGE
In order to do parametric hypothesis testing, it is necessary to ensure that data is normally distributed. In this sense, we applied a Kolmogorov-Smirnov (KS) test, and it was seen that not all features were considered statistically normally distributed for both the malignant and benignant populations. Therefore, nonparametric methods should be preferably used.

CHANGE
During pre-processing, no missing values were found, as well as no significant reason to eliminate neither the proportion of rows with outliers nor the features with outliers.

CHANGE
During hypothesis testing, we considered two independent samples for each feature: malignant and benignant. The null hypothesis stated that the means for both populations was the same, weather the alternative hypothesis stated that means were different for each population. Therefore, this was a two-sample two-sided test, for which we followed, just for learning purposes and for comparison, both parametric and non-parametric methods.

Results were almost the same for both, finally determining that means were different for all features except for fractal dimension standard error, texture standard error, and smoothness standard error.

CHANGE

We concluded that this dataset has an excellent quality for its purpose since simple pre-processing was needed to obtain clear conclusions. Features show significant mean differences between both classes in a way that, as seen in the histograms, malignant values use to be higher than benignant ones, which mean that nuclei are in fact different between malignant and benignant tissue, and that some features can be more crucial than others when determining the final diagnosis.