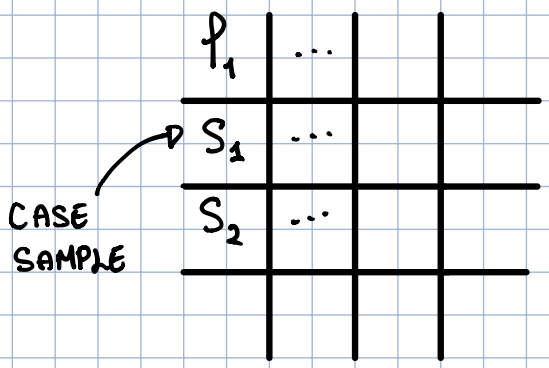
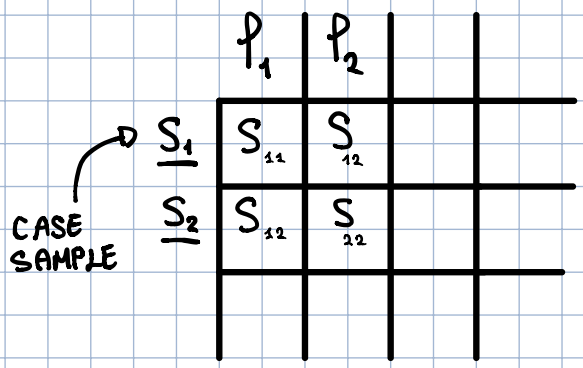
MULTI-VARIATE ANALYSIS

We are interested in extending the analysis of a dataset to the MULTI-VARIATE case. In particular, if previously we studied the dataset one feature (e.k.a. one column vector) at a time, now we want to consider several features jointly.



(UNI-VARIATE VIEW)



(MULTI-VARIATE VIEW)

Here the s_i are scalar values and we count the # of scalars to estimate the feature's distribution.

Here the CASE SAMPLE are not scalars but raw-vectors.

PMF ESTIMATION

Given a dataset with discrete features f_1, f_2, if we want to estimate their joint pmf we have to count the raw vectors and compute for each unique raw vector its frequency.

Thus, we have,

$$\text{pmf}_{f_1+f_2}(\underline{s}_i) := \frac{\#\text{ of rows equals to } \underline{s}_i}{\text{total \# of rows}}$$

Q: What if a particular row never appears in the dataset? Don't we want to apply some SMOOTHING?

SCATTER PLOTS

Scatter plots allow us to visualize if there are any spheres in which all the samples of a particular class are situated.

NOISY DATA is data in which all the samples are uniformly distributed with respect to the class.

A scatter plot allows us to understand if the data is noisy or not.

TODO: Add image of scatter plots.

PMF VS PDF

Consider a dataset that contains numerical data such as the width or the height of something.

When we estimate the distribution of said feature, should we estimate a pmf or a pdf?

If we want to be rigorous, we have to treat the feature as a continuous r.v., and thus we have to estimate a pdf. However, if the initial accuracy of the dataset is low (i.e. few decimal digits), we can DISCRETIZE the dataset and treat the feature as a discrete r.v.

INT NUMBERS → PMF (DISCRETE R.V.)
FLOAT NUMBERS → PDF (CONTINUOUS R.V.)
↖ DISCRETIZATION

A possible method to discretize a dataset is to apply the FLOOR() function.