TD 3

Adrian Valente

November 18, 2020

Goals: open and explore a dataset, visualize all kinds of relationships between variables, compute a summary statistic and a confidence interval for it, choose statistical test for categorical data independence.

1 Exloratory analysis tools (theory)

- Print a summary of the Titanic dataset. Which variables are categorical and which are continuous?
- How to visualize the relationship between two categorical variables? Between two continuous ones? Between a continuous and a categorical variable?
- How to quantify the existence of a relationship between two categorical variables? Between two continuous ones? Between a continuous and a categorical variable?

2 Linear regression: the Anscombe dataset

- Print a summary of the Anscombe dataset. It is actually a set of four 2D dataset exhibiting the same correlations but very different profiles between variables
- For each pair of variables (x_i, y_i) visualize the relationship with a scatter plot. Add a `geom_smooth` object to visualize the result of a linear regression.
- For each pair of variables (x_i, y_i) compute the Pearson correlation coefficient, and fit a linear regression.

3 Statistical inference with the Titanic dataset

- At the core of statistics stands the notion of *model*. We consider that the data that we are given (under the form of *samples* ($\mathbf{x}_1, \ldots, \mathbf{x}_N$)) has been generated by a certain process which: - can be described mathematically, - incorporates randomness, - and depends on certain parameters

  (formally we say that each $\mathbf{x}_i$ is sampled from a parametrized probability distribution $f(\mathbf{x}; \theta)$). The objective of the statistician can be 3-fold: - to estimate the values of the parameters θ, and estimate the uncertainty of his knowledge (which gives the domain of *estimators* and *confidence intervals*) - to confront alternative models between each other (which gives the domain of *statistical tests* to measure *p-values*) - to characterize the relevance of each component of his model (with the *effect sizes* associated to tests).

- Think of models that could give rise to the Titanic survival rate. Try to formalize them mathematically.

- We will consider the simplest one, that for each passenger a Bernoulli sample is generated to determine its survival. What is the parameters associated to this model? How to estimate it? Can you generate a confidence interval for it? (if you are mathematically oriented, I encourage you to think about how a CI can be computed with the central limit theorem).

- Let us consider a more complex model. We will still consider that for each passenger survival is determined by a Bernoulli sample, but from a sex-dependent distribution. What are the parameters, and can you estimate them? Can you generate confidence intervals? Do the CIs for the 2 parameters overlap?

- We would like to know which model to choose, between those two. What statistical test would be appropriate? What are its null and alternative hypotheses? Now run it. What effect size could you associate to it?

- Is there any influence of age on the survival rate? Which statistical tests can answer to that question?

- Unrelated to survival now: we want to model the continuous random variables age and fare. What would be a good probability distribution

that could fit them?

- How can you simultaneously test for an effect of all variables on survival?

4 A cautionary tale: Simpson's paradox

- In R, load the UCBAdmissions dataset `data(UCBAdmissions)`. Notice it is already under the form of a contingency table.
- Compute the acceptance rate for men and women overall, and then by department. How is this possible?
- The original article is Bickel et al., "Sex bias in graduate admissions: data from Berkeley", 1975
- "All models are wrong, some are useful."