technische universität
dortmund

Mathematical Statistics and Application in Industry

Seminar Empirical Processes

A little bit of mathematics to obtain statistical methods with less assumptions

Marléne Baumeister, Markus Pauly

Empirical distribution function I

- Starting point in 1930-1940: *empirical distribution function*.
- Let $X_1, \ldots, X_n$ be i.i.d. with distribution function F.
- **Theorem I** (Glivenko-Cantelli, 1933):

$$||\widehat{F}_n - F||_\infty = \sup_{-\infty<x<\infty} |\widehat{F}_n(x) - F(x)| \to 0 \text{ almost surely.}$$

Empirical distribution function II

- For some $t_1, ..., t_k \in \mathbb{R}$ holds

$$\sqrt{n}\left(\widehat{F}_n(t_i) - F(t_i)\right)_{i=1}^{k} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}\right),$$

  where the $k \times k$-matrix $\mathbf{\Sigma}$ has the $(i,j) - th$ element $F\left(\min((t_i, t_j)\right) - F(t_i)F(t_j)$. This result was extended to the corresponding empirical process by

- **Theorem II** (Donsker, 1952):

$$\sqrt{n}(\widehat{F}_n - F) \xrightarrow{d} B_0 \circ F \text{ in } D(\mathbb{R}),$$

  where B_0 is a Brownian bridge on $[0,1]$ and $D(\mathbb{R})$ denotes the Skorokhod space.

Applications

- Many estimators for $\hat{\theta}_n$ and test statistics T_n can be written in terms of empirical distribution functions:
  - Sample median: $m_n = \hat{F}_n^{-1}(0.5)$ (in the same way: general quantiles).
  - Sample mean: $\overline{x}_n = \int x \, \mathrm{d}\hat{F}_n(x)$
  - Mann-Whitney-Test or the estimator for the concordance measure:

$$\frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{1}\{X_i > Y_j\} = \int \Big(1 - \hat{F}_n(y)\Big) \, \mathrm{d}\hat{G}_n(y),$$

    where $\hat{F}_n$ and $\hat{G}_m$ are the empirical distribution functions of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, respectively.
  - And a lot more, e.g. Kaplan-Meier estimator, Kolmogorov-Smirnov statistic etc.

Application II

- In all previous situations we have

$$\widehat{\theta}_n = \Phi(\hat{F}_n) \quad \text{(one-sample)}$$
$$T_n = \Phi(\hat{F}_n, \hat{G}_n) \quad \text{(two-sample)}$$

- To obtain asymptotic **confidence intervals** or **asymtotically valid tests**, we can combine the Donsker Theorem and a functional delta-Method, e.g.

$$\sqrt{n}\Big(\Phi(\hat{F}_n) - \Phi(F)\Big) \xrightarrow{d} \Phi'_F(B_0 \circ F).$$

- More recently, empirical process theory is used to analyse theoretical properties of AI and ML algorithms.

My Application: quantile-based ANOVA and MANOVA

- Asymptotic level-α-tests for hypotheses like

$$\mathcal{H}_0 : \mathbf{m}_1 = \cdots = \mathbf{m}_k,$$

  with group-specific (vectors of) medians $\mathbf{m}_i$, $i \in \{1, \ldots, k\}$.
- Constructed by the Donsker Theorem and the Functional Delta Method.

Advantages:

- Less assumption: no distribution, no homoscedasty.
- Works well for heavy-tailed data.
- Allows incorporation of complex factorial structures.

Possible topics

Source: Vaart and Wellner (2000)

- Introduction (Stochastic convergence for general spaces)
- General Donsker and Glivenko-Cantelli results for

$$\mathcal{F} \ni f \mapsto \frac{1}{n}\sum_{i=1}^{n} f(X_i) \qquad \left(\text{Recall } \mathbb{R} \ni t \mapsto \hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbf{1}\{X_i \leq t\}}_{f_t(X_i)}\right)$$

- Functional Delta-Method
- Bootstrap for one-sample settings
- Bootstrap and Permutation for k-sample settings.
- Application of the theory in different fields (e.g. biostatistics) of recent research.

Modus/dates

- Required Qualification: Knowledge about measure-theoretic probability theory (e.g. stochastic convergence), finished Bachelor is recommended
- 2-3 blocks during the lecture time
- Currently planned as a seminar in presence
- First meeting in the first lecture week (to find dates etc., maybe a first introduction about basic concepts by myself)
- talk (40-60 minutes) and a short seminar paper.

Registration

- Binding registration via Email untill **27.03.2023**:

  marlene.baumeister@tu-dortmund.de

  Please add the following information:
  - Study program,
  - Prior knowledge about probability theory (which lecture, course),
  - If possible: desired topic.
- Further Information will soon be given on our homepage.

References

Vaart, A. W. van der and Jon A Wellner (2000). *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer.