Statistics for Data Science

MSc Data Science WiSe 2019/20

Prof. Dr. Dirk Ostwald

Frequentist inference

(11) Nonparametric inference

A statistical model $\mathcal{P}$ is a set of probability distributions.

- A *parametric statistical model* is a statistical model that can be parameterized by a finite number of parameters.

- A *nonparametric statistical model* is a statistical model that cannot be parameterized by a finite number of parameters. More specifically, a set of distributions $\mathcal{P}$ is said to behave nonparametrically, if it is not possible to identify a finite-dimensional space Θ, such that there exists a bijective relationship between Θ and $\mathcal{P}$, in the sense that each member $\mathbb{P} \in \mathcal{P}$ can be identified by only one member $\theta \in \Theta$, and vice versa.

Bibliographic remarks

The presented material follows Moeschlin (2001, Sections 2.1 - 2.4) in the discussion of the empirical distribution, histograms, and Kernel density estimation. The discussion of the bootstrap is based on Wasserman (2004, Section 8). For an introduction to nonparametric hypothesis testing based on permutation tests see Pesarin and Salmaso (2010).

Nonparametric inference

- The empirical distribution
- Histogram density estimation
- Kernel density estimation
- The bootstrap

Overview

- The *empirical distribution* is a method to estimate the probability of an event A by evaluating a normalized count of the number of observations satisfying the event A. As an estimator, the empirical distribution is unbiased and consistent.
- *Histogram density estimation* is a method to estimate PDFs using empirical distributions by evaluating a normalized and binwidth-adjusted count of the number of observations satisfying a given event. As estimators, histograms are consistent.
- *Kernel density estimation* is a method to estimate PDFs using smoothed empirical distributions and obtaining continuous PDF estimates. Kernel density estimators are consistent.
- The *bootstrap* is a method to estimate the variance of a statistic using the observations' empirical distribution and computer-based resampling.

Nonparametric inference

- **The empirical distribution**
- Histogram density estimation
- Kernel density estimation
- The bootstrap

Definition (Empirical distribution)

Let $X = (X_1, ..., X_n)$ denote a real-valued random sample and let $x = (x_1, ..., x_n)$ denote a realization of X. Then

$$Q_n^x(A) := \frac{1}{n}|\{i \in \mathbb{N}_n | x_i \in A\}| = \frac{1}{n}\sum_{i=1}^{n} 1_A(x_i),\ A \in \mathcal{B} \tag{1}$$

is a probability measure on $(\mathbb{R}, \mathcal{B})$. The function

$$Q_n : \mathcal{B} \times \mathbb{R}^n \to [0,1],\ (A, x) \mapsto Q(A, x) := \frac{1}{n}\sum_{i=1}^{n} 1_A(x_i) =: Q_n^x(A) \tag{2}$$

is called the *empirical distribution* of X.

Remarks

- Recall that $1_A(x) = 1$, if $x \in A$, and $1_A(x) = 0$, if $x \notin A$.
- Note that $Q_n^x(\mathbb{R}) = \frac{1}{n}\sum_{i=1}^{n} 1_{\mathbb{R}}(x_i) = \frac{1}{n}\sum_{i=1} 1 = \frac{1}{n} \cdot n = 1$.

Theorem (Unbiasedness of the empirical distribution)

Let $X = (X_1, ..., X_n)$ be a real-valued random sample with distribution $\mathbb{P}$ and let $A \in \mathcal{B}$. Then

$$Q_n(A, \cdot) : \mathbb{R}^n \to [0, 1], x \mapsto Q_n(A, x) := \frac{1}{n} \sum_{i=1}^{n} 1_A(x_i) \tag{3}$$

is an unbiased estimator of $\mathbb{P}(X_i \in A)$

Proof

$$\begin{aligned} \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} 1_A(x_i) \right) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(1_A(x_i)\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(0 \cdot \mathbb{P}(X_i \notin A) + 1 \cdot \mathbb{P}(X_i \in A)\right) \\ &= \mathbb{P}\left(X_i \in A\right). \end{aligned} \tag{4}$$

□

Definition (Empirical distribution function)

Let $X = (X_1, ..., X_n)$ denote a real-valued random sample, let $x = (x_1, ..., x_n)$ denote a realization of X, and let Q^x denote the probability measure defining the the empirical distribution of X. The cumulative distribution function of Q_n^x is given by

$$P_n^x(\xi) := \frac{1}{n}|\{i \in \mathbb{N}_n | x_i \leq \xi\}| = \frac{1}{n}\sum_{i=1}^{n} 1_{]-\infty,\xi]}(x_i), \text{ for all } \xi \in \mathbb{R}. \tag{5}$$

The function

$$P_n : \mathbb{R} \times \mathbb{R}^n \to [0,1], (\xi, x) \mapsto P_n(\xi, x) := \frac{1}{n}\sum_{i=1}^{n} 1_{]-\infty,\xi]}(x_i) := P_n^x(\xi) \tag{6}$$

is called the *empirical distribution function* of the sample X.

Remark

- Recall that $1_{]-\infty,\xi]}(x) = 1$, if $x \in]-\infty, \xi]$, and $1_{]-\infty,\xi]}(x) = 0$, if $x \notin]-\infty, \xi]$.

Theorem (Consistency of the empirical distribution (Glivenko-Cantelli))

Let $X = (X_1, ..., X_n)$ be a real-valued sample of size n with cumulative distribution function P and let P_n denote the empirical distribution function of X. Then

$$\mathbb{P}\left(x \in \mathbb{R}^n | \lim_{n\to\infty}\left(\sup_{\xi\in\mathbb{R}} |P_n(\xi, x) - P(\xi)|\right) = 0\right) = 1. \tag{7}$$

Remark

- For $n \to \infty$, P_n converges to P.
- For proofs, see Cantelli (1933); Van der Vaart (2000).

Nonparametric inference

- The empirical distribution
- **Histogram density estimation**
- Kernel density estimation
- The bootstrap

Definition (Histogram)

Let $X = (X_1, ..., X_n)$ denote a real-valued random sample of size n and let

$$Q_n^x(A) := \frac{1}{n} \sum_{i=1}^{n} 1_A(x_i),\ A \in \mathcal{B} \tag{8}$$

be a probability measure on $(\mathbb{R}, \mathcal{B})$. For $b > 0$ consider the partition of $\mathbb{R}$ given by

$$\mathcal{P}_b := \{]ib, (i+1)b] | i \in \mathbb{Z}\}, \tag{9}$$

such that

$$q_i(b) := Q_n\left(]ib, (i+1)b]\right) = \frac{1}{n} \sum_{i=1}^{n} 1_{]ib,(i+1)b]}(x_i) \tag{10}$$

is an estimate of $\mathbb{P}\left(]ib, (i+1)b]\right)$. Then

$$\hat{h}_n : \mathbb{R} \to \mathbb{R}_{\geq 0}, x, \mapsto \hat{h}_n(x) := \frac{1}{b} \sum_{i \in \mathbb{Z}} q_i(b) \cdot 1_{]ib,(i+1)b]}(x) \tag{11}$$

is known as the *histogram of X with binwidth b.*

Remark

- Note that for $x \in]ib, (i+1)b]$ the histogram evaluates to $\hat{h}_n(x) = \frac{1}{b} Q_n\left(]ib, (i+1)b]\right)$
- $\hat{h}_n(x)$ is thus the number of realizations in $]ib, (i+1)b]$ divided by nb.

Theorem (Consistency of the histogram density estimator)

Let $X = (X_1, ..., X_n)$ be a real-valued random sample with PDF p. For $c > 0, 0 < \alpha < 1$, let

$$\hat{h}_n : \mathbb{R} \to \mathbb{R}_{>0}, x, \mapsto \hat{h}_n(x) := \frac{1}{b_n} \sum_{i \in \mathbb{Z}} q_i(b_n) \cdot 1_{]ib_n, (i+1)b_n]}(x_i), \tag{12}$$

where

$$b_n := \frac{c}{n^\alpha} \tag{13}$$

denotes the binwidth of the histogram density estimator $\hat{h}_n$. Then $\hat{h}_n$ is a consistent estimator of p, i.e.,

$$\mathbb{P}\left(\left\{x \in \mathbb{R}^n | \lim_{n \to \infty} \left(\sup_{\xi \in \mathbb{R}} |\hat{h}_n(\xi) - p(\xi)|\right) = 0\right\}\right) = 1. \tag{14}$$

Remarks

- For $n \to \infty$ and with an adaptive binwidth, $\hat{h}_n$ converges to p.
- For a proof, see Révész (1968).

Nonparametric inference

- The empirical distribution
- Histogram density estimation
- **Kernel density estimation**
- The bootstrap

Kernel density estimation

- aims to estimate an unknown PDF by a continuous function.
- is a form of "smoothed" PDF histogram estimation.
- originates from the works of Rosenblatt (1956) and Parzen (1962).

In the following,

- we first review *homothetic transformations* of random variables,
- the *convolution* of random variables, and
- the convolution of a PDF with the Dirac delta measure.

We then

- define the kernel density estimator and commonly used kernels,
- and state a consistency theorem for the kernel density estimator.

Theorem (Homothetic transformation of continuous random variables)

Let X be a random variable with PDF p_X and let

$$f_b : \mathbb{R} \to \mathbb{R},\, x \mapsto f_b(x) := bx \text{ for } b > 0 \tag{15}$$

be a *homothetic transformation*. Then the random variable $Y := f_b(X)$ has PDF

$$p_Y : \mathbb{R} \to \mathbb{R}_{\geq 0}, y \mapsto p_Y(y) = \frac{1}{b} p_X\left(\frac{y}{b}\right) \tag{16}$$

Proof We first note that the inverse of f_b is given by

$$f_b^{-1} : \mathbb{R} \to \mathbb{R},\, y \mapsto f_b^{-1}(y) = \frac{y}{b}, \tag{17}$$

such that $f_b^{-1}(f_b(x)) = bx/b = x$. We next note that $f_b'(x) = b > 0$, such that

$$\frac{1}{|f'(f^{-1}(y))|} = \frac{1}{b} \tag{18}$$

The theorem then follows with the univariate probability density transform theorem.

Theorem (Convolution of random variables)

Let X and Y be two independent random variable with PDFs p_X and p_Y, respectively. Let

$$Z := X + Y \tag{19}$$

be the sum of X and Y. Then a PDF of Z is given by the *convolution* of the PDFs of X and Y,

$$p_Z : \mathbb{R} \to \mathbb{R}_{\geq 0}, z \mapsto p_Z(z) = (p_X * p_Y)(z), \tag{20}$$

where

$$(p_X * p_Y)(z) := \int_{-\infty}^{\infty} p_X(z - \xi) p_Y(\xi)\, d\xi = \int_{-\infty}^{\infty} p_X(\xi) p_Y(z - \xi)\, d\xi. \tag{21}$$

Remark

- For a proof, see DeGroot and Schervish (2012, Theorem 3.9.4).

Convolution of a continuous random variable with the Dirac measure

Consider the convolution of a continuous random variable X with PDF p_X and the Dirac measure

$$\delta_{x_j} : \mathbb{R} \to \mathbb{R}, x \mapsto \delta_{x_i}(x) := \begin{cases} 1, & x = x_i \\ 0, & x \neq x_i \end{cases} \text{ for } x_i \in \mathbb{R} \tag{22}$$

Then, with some abuse of concepts and notation, we have

$$(p_X * \delta_{x_i})(x) := \int_{-\infty}^{\infty} p_X(x - \xi)\delta_{x_i}(\xi)\, d\xi = p_X(x - x_i). \tag{23}$$

Definition (Kernel density estimator)

Let $k : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a PDF and let

$$k_b : \mathbb{R} \to \mathbb{R}_{\geq 0}, x \mapsto \frac{1}{b} k\left(\frac{x}{b}\right) \tag{24}$$

be the PDF resulting from the application of the homothetic transformation f_b. Then the function

$$\hat{k}_n : \mathbb{R} \to \mathbb{R}_{\geq 0}, x \mapsto \hat{k}_n(x) = \frac{1}{n} \sum_{i=1}^{n} k_b\left(x - x_i\right) \tag{25}$$

is referred to as *kernel density estimator (KDE) with kernel k and bandwidth b.*

Remarks

- $k_b(x - x_i)$ is a "squeezed and shifted" version of the kernel k.
- For symmetric kernels, $k_b(x - x_i)$ is centered on the observed values x_i.
- Each k_b is a PDF, hence the sum of n k_b's divided by n is a PDF.
- In contrast to histograms, KDEs yield continuous PDF estimates.
- Like histograms, KDEs depend on a bandwidth (binwidth) parameter b.

Commonly employed kernel functions include

The Gaussian kernel

$$k(x) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \tag{26}$$

The Cauchy kernel

$$k(x) := \frac{\pi}{1+x^2} \tag{27}$$

The Picard kernel

$$k(x) := \frac{1}{2} \exp(-|t|) \tag{28}$$

The triangle kernel

$$k(x) := \begin{cases} \frac{1}{\sqrt{6}} - \frac{|x|}{6}, & |t| \leq \sqrt{6} \\ 0, & |t| \leq \sqrt{6}. \end{cases} \tag{29}$$

Theorem (Consistency of the kernel density estimator)

Let $X = (X_1, ..., X_n)$ be a real-valued random sample with PDF p. For $c > 0, 0 < \alpha < 1/2$, and let

$$\hat{k}_n : \mathbb{R} \to \mathbb{R}_{\geq 0}, x \mapsto \hat{k}_n(x) = \frac{1}{n} \sum_{i=1}^{n} k_{b_n} (x - x_i) , \tag{30}$$

where

$$b_n := \frac{c}{n^\alpha} \tag{31}$$

denotes the bandwidth of the kernel density estimator $\hat{k}_n$. Then $\hat{k}_n$ is a consistent estimator of p, i.e.,

$$\mathbb{P}\left(\left\{ x \in \mathbb{R}^n | \lim_{n \to \infty} \left(\sup_{\xi \in \mathbb{R}} |\hat{k}_n(\xi) - p(\xi)| \right) = 0 \right\} \right) = 1. \tag{32}$$

Remark

- For $n \to \infty$ and with an adaptive bandwidth, $\hat{k}_n$ converges to p.
- Note the similarity to the histogram density estimator consistency theorem.
- For a mathematical rigorous form and proof, see Van der Vaart (2000).

Nonparametric inference

- The empirical distribution
- Histogram density estimation
- Kernel density estimation
- **The bootstrap**

The Bootstrap

A method for estimating the variance of a statistic.

- Let $X_1, ..., X_n \sim \mathbb{P}$ for an unknown distribution $\mathbb{P}$.
- Let $T_n = f(X_1, ..., X_n)$ be a statistic and us be interested in $\mathbb{V}_{\mathbb{P}}(T_n)$.

  $\rightarrow$ Estimate $\mathbb{V}_{\mathbb{P}}(T_n)$ by $\mathbb{V}_{Q_n}(T_n)$.

  $\rightarrow$ Approximate $\mathbb{V}_{Q_n}(T_n)$ by computer-based sampling.

For comprehensive introductions to the bootstrap approach, see Efron and Tibshirani (1994) and Davison and Hinkley (1997).

Approximating distribution variances by simulation

→ Weak law of large numbers under continuous transformations

Let $X_1, ..., X_m \sim P$ and $X := X_1$. Then

$$\frac{1}{m}\sum_{i=1}^{m} f(X_i) \xrightarrow[m\to\infty]{\mathrm{P}} \mathbb{E}_P\left(f(X)\right)$$

In particular

$$\frac{1}{m}\sum_{i=1}^{m} X_i \xrightarrow[m\to\infty]{\mathrm{P}} \mathbb{E}_P\left(X\right) \text{ and } \frac{1}{m}\sum_{i=1}^{m} X_i^2 \xrightarrow[m\to\infty]{\mathrm{P}} \mathbb{E}_P\left(X^2\right)$$

Hence

$$S^2(X) = \left(\frac{1}{m}\sum_{i=1}^{m} X_i\right)^2 + \frac{1}{m}\sum_{i=1}^{m} X_i^2 \xrightarrow[m\to\infty]{\mathrm{P}} \mathbb{E}_P\left(X\right)^2 + \mathbb{E}_P\left(X^2\right) = \mathbb{V}_P(X).$$

Approximating $\mathbb{V}_{Q_n}(T_n)$

- Q_n allocates probability mass of $\frac{1}{n}$ to each data point $x_i, i = 1, ..., n$.
- One draw $X^* \sim Q_n \Rightarrow$ A uniform random draw from $x_1, ..., x_n$.
- Creating one i.i.d. sample $X_1^*, ..., X_n^* \sim Q_n$

  $\Rightarrow n$-fold sampling with replacement from $x_1, ..., x_n$.
- Evaluation of $T_n^* := f(X_1^*, ..., X_n^*)$.
- Repeat $m \to \infty$ times.
- Use the sample variance $S^2(T^*)$ of T_n^* as approximation of $\mathbb{V}_{Q_n}(T_n)$

Note the two approximations

$$\mathbb{V}_{\mathbb{P}}(T_n) \approx \mathbb{V}_{Q_n}(T_n) \approx S^2(T_n^*). \tag{33}$$

The first approximation is usually not so good, but the second is good.

A bootstrap statistic variance estimation algorithm

Input A data set $x_1, ..., x_n$.

Output An approximation $S^2(T^*)$ of $\mathbb{V}_{Q_n}(T_n)$.

For $j = 1, ..., m$

Sample $x_1^*, ..., x_n^*$ by sampling from $x_1, ..., x_n$ n-times with replacement.

Evaluate $t_{n,j}^* = f(x_1^*, ..., x_n^*)$.

Set

$$S^2(T^*) := \frac{1}{m} \sum_{j=1}^{m} \left(t_{n,j}^* - \frac{1}{m} \sum_{j=1}^{m} t_{n,j}^* \right)^2 \tag{34}$$

Return $S^2(T^*)$.

Overview

- The *empirical distribution* is a method to estimate the probability of an event A by evaluating a normalized count of the number of observations satisfying the event A. As an estimator, the empirical distribution is unbiased and consistent.
- *Histogram density estimation* is a method to estimate PDFs using empirical distributions by evaluating a normalized and binwidth-adjusted count of the number of observations satisfying a given event. As estimators, histograms are consistent.
- *Kernel density estimation* is a method to estimate PDFs using smoothed empirical distributions and obtaining continuous PDF estimates. Kernel density estimators are consistent.
- The *bootstrap* is a method to estimate the variance of a statistic using the observations' empirical distribution and computer-based resampling.

References

Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari*, 4(421-424).

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*, volume 1. Cambridge university press.

DeGroot, M. H. and Schervish, M. J. (2012). *Probability and Statistics*. Pearson Education.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Moeschlin, O. (2001). *Angewandte Statistik*. FernUniversitaet in Hagen.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Pesarin, F. and Salmaso, L. (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons.

Révész, P. (1968). *The laws of large numbers*, volume 4. Academic Press.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer.