Statistics for Data Science

Marina Anca Cidota
Faculty of Mathematics and Computer Science
University of Bucharest

cidota@fmi.unibuc.ro

References

1. Robert Shumway, David Stoffer. Time Series Analysis and its Applications with R Examples, 4th edition, Springer, 2017
2. Peter Brockwell, Richard Davis. Time Series: Theory and Methods, Springer-Verlag, 1987
3. James Hamilton. Time Series Analysis, Princeton University Press, 1994
4. Norman Matloff. Probability and Statistics for Data Science, CRC Press, 2019.
5. Monica Dumitrecu, Anton Batatorescu. Applied Statistics using the R system, Ed. Universitatii din Bucuresti, 2006.
6. Ion Vaduva. Analiza dispersionala, Ed. Tehnica, 1970
7. Peter Bruce, Andrew Bruce. Practical Statistics for Data Scientists, O'Reilly, 2017
8. Hadley Wickham, Garrett Grolemund. R for Data Science, O'Reilly, 2016.

Introduction

Def. (Ω, K, P) is called a probability space, where

1) Ω - sample space = the set of all possible outcomes
2) $K \subset \mathcal{P}(\Omega)$ the power set of Ω (the set of all subsets of Ω)

   K is a σ-algebra $\begin{cases} \emptyset \in K \\ A \in K \Rightarrow \overline{A} \in K \\ A_n \in K \Rightarrow \bigcup_{n=1}^{\infty} A_n \in K \end{cases}$

3) $P : K \to [0,1]$ is a probability measure function

   $P(\Omega) = 1$

   - a measure function: $\begin{cases} P : K \to [0, \infty) \\ P(\emptyset) = 0 \\ P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) \text{ if } A_n \in K, \\ \qquad A_i \cap A_j = \emptyset, \ \forall\, i \neq j. \end{cases}$

The elements of K are called random events.

Introduction

<u>Def</u>. if S is a set, the σ-algebra generated by S, written $\mathcal{B}(S)$, is:

$$\mathcal{B}(S) \overset{def}{=} \bigcap_{\substack{\mathcal{F} \supset S \\ \mathcal{F} \text{ is } \sigma\text{-algebra}}} \mathcal{F}$$

If $S = \mathbb{R}$ then $\mathcal{S} \overset{not}{=} \mathcal{B}(S) = \mathcal{B}\big((a,b) \mid a, b \in \mathbb{R} \big)$

If S is finite or countable then $\mathcal{S} = \mathcal{P}(S)$.

$(S, \mathcal{S})$ is a measurable space.

<u>Def</u>. $X : (\Omega, \mathcal{K}) \to (S, \mathcal{S})$ is a random variable if

$$\forall B \in \mathcal{S} \Rightarrow X^{-1}(B) \in \mathcal{K}.$$

$$P \circ X^{-1} : \mathcal{S} \to [0,1], \quad (P \circ X^{-1})(B) = P(X^{-1}(B))$$

Introduction

The cumulative distribution function (CDF) of X is

$$F(x) \stackrel{\text{def}}{=} (P \circ X^{-1})(-\infty, x) = P(X < x)$$

$$F : \mathbb{R} \to [0,1]$$

$$F(-\infty) = 0; \quad F(\infty) = 1; \quad F \text{ is monotonically increasing}$$

- for X a continuous random variable, $P \circ X^{-1}$ is specified by the density function $f(x)$,

$$f(x) \geq 0 \quad \forall x$$

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

$$f(x) = F'(x)$$

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

Introduction

- for X a discrete random variable, $P \circ X^{-1}$ is specified by $\{p(x), x \in S\}$, $p(x) \geq 0, \forall x$

$$\sum_{x \in S} p(x) = 1$$

$$X : \begin{pmatrix} x_1 & \dots & x_n \\ p_1 & \dots & p_n \end{pmatrix}$$

$$F(x) = \sum_{x_i < x} P(x = x_i)$$

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx \quad \left(\text{or } \sum_{i=1}^{n} x_i p_i \text{ for the discrete case}\right)$$

$$Var(X) = E\left(\left(X - E(X)\right)^2\right) = E(X^2) - E(X)^2$$

$\sigma = \sqrt{Var(X)}$ – standard deviation

Stochastic models of different phenomena of interest

- Categorical random variables/vectors.
- Quantitative discrete/continuous random variables/vectors.

Def. A stochastic process is a family of random variables $\{X_t, t \in T\}$ defined on a probability space $(\Omega, \mathcal{K}, P)$.

time T, states S

time $\mathbb{N} / \mathbb{Z} / \mathbb{R} / [0, \infty)$

state discrete$/ \mathbb{R} / \mathbb{R}^d$

Stochastic models of different phenomena of interest

• Observations in "transversal" studies (at a fixed moment of time)

$\{X_1, \dots, X_n\}$ random variables, independent and identically distributed (iid) like the stochastic model $X : \Omega \to S$.

statistical data are the observed values $(x_1, \dots, x_n) \in S^n$.

• Observations in "longitudinal" studies (on a fixed time interval)

$\{X_t, t \leq t_n\} \subset \{X_t, t \in T\}$

statistical data are the observed part of the process trajectory $(x_t, t \leq t_n)$

Stochastic models of different phenomena of interest

Parametric statistics

Hypothesis: statistical data come from a stochastic model whose distribution has a known functional form, but it depends on an unknown parameter $\theta \in \mathbb{R}^k$, $k \geq 1$.

- statistical data are available;
- Looking for estimators of θ or tests on hypotheses over θ.

Example of stochastic models

1. categorical 1-dim random variable

X = satisfaction degree

$S = \{ vun, un, sa, vsa \}$

$$P \circ X^{-1} : \begin{pmatrix} vun & un & sa & vsa \\ p_1 & p_2 & p_3 & p_4 \end{pmatrix} \qquad p_i \geq 0, \quad \sum_{i=1}^{4} p_i = 1$$

2. categorical 2-dim random vector

$X = (X_1, X_2)' =$ (intention to vote, satisfaction degree about the current situation)

$S = \{ (yes, vun), (yes, un), (yes, sa), (yes, vsa),$
$(no, vun), (no, un), (no, sa), (no, vsa) \}$

Example of stochastic models

Quantitative discrete random variables

3. discrete uniform distribution – e.g. rolling a dice

$$X \sim U\{1, \ldots, n\}, \quad n \in \mathbb{N}, \quad n \geq 2$$

$$S = \{1, 2, \ldots, n\}$$

$$P(X = x) = \frac{1}{n}, \quad x = 1, 2, \ldots, n$$

$$E(X) = \frac{n+1}{2}$$

$$Var(X) = \frac{n^2 - 1}{12}$$

Example of stochastic models

4. Bernoulli distribution – e.g. having "success" in a trial with two possible outcomes: success/failure.

$$X \sim B(1, \theta), \quad \theta \in (0,1) \qquad S = \{0,1\}$$

$$P_\theta(X = x) = \theta^x (1-\theta)^{1-x}, \quad x = 0,1$$

$$X : \begin{pmatrix} 0 & 1 \\ 1-\theta & \theta \end{pmatrix}$$

$$E(X) = \theta$$

$$Var(X) = \theta(1-\theta)$$

Example of stochastic models

5. Binomial distribution – e.g. number of "successes" in n independent trials with two possible outcomes: success/failure.

$$X \sim Bi(n, \theta), \quad \theta \in (0,1), \quad S = \{0, 1, \dots, n\}$$

$$P_\theta(X = x) = C_n^x \cdot \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, \dots, n$$

$$E(X) = n \cdot \theta$$

$$Var(X) = n \cdot \theta (1-\theta)$$

Example of stochastic models

6. Poisson distribution – e.g. number of defective devices in a very large volume lot.

$X \sim Po(\theta), \quad \theta \in (0, \infty), \quad S = \mathbb{N}$

$P_\theta (X = x) = \frac{\theta^x}{x!} e^{-\theta}, \quad x = 0, 1, \ldots$

$E(X) = \theta$

$Var(X) = \theta$

Example of stochastic models

7. Geometric distribution – e.g. the moment of the first "success" in a sequence of independent trials with two possible outcomes: success/failure.

$$X \sim Ge(\theta), \quad \theta \in (0,1), \quad S = \mathbb{N}^*$$

$$P_\theta(X = x) = (1-\theta)^{x-1} \cdot \theta \qquad x = 1, 2, \ldots$$

$$E(X) = \frac{1}{\theta}$$

$$Var(X) = \frac{1-\theta}{\theta^2}$$

Example of stochastic models

Quantitative discrete random vectors

8. Multinomial distribution - e.g. sampling with replacement

   n trials - extractions from a ballot box containing balls of d colors

$$X = (X_1, \ldots, X_d)' \sim M(n; p_1, \ldots, p_d)$$

$$P_\theta\left(X = (x_1, \ldots, x_d)'\right) = \frac{n!}{x_1! \ldots x_d!} p_1^{x_1} \ldots p_d^{x_d}, \text{ where}$$

$$S = \left\{ (x_1, \ldots, x_d)' \mid x_i \in \{0, 1, \ldots, n\} \; \forall i = \overline{1,d}, \; \sum_{i=1}^{d} x_i = n \right\}$$

$$\theta = (p_1, \ldots, p_d), \quad p_i \in [0,1] \; \forall i = \overline{1,d}, \; \sum_{i=1}^{d} p_i = 1$$

$$E(X_i) = n p_i$$

$$i = \overline{1,d}$$

$$\mathrm{Var}(X_i) = n p_i (1 - p_i) \qquad \mathrm{Cov}(X_i, X_j) = -n p_i p_j, \; i \neq j$$

Example of stochastic models

Quantitative continuous random variables

9. continuous uniform distribution - e.g throw a dart at random at the interval $[0, \theta]$ - all the points are equally like to be hit.

$$X \sim U(0, \theta), \quad \theta \in (0, \infty)$$

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & x \in [0, \theta] \\ 0, & x \notin [0, \theta] \end{cases}$$

$$E(X) = \frac{\theta}{2}$$

$$Var(X) = \frac{\theta^2}{12}$$

Example of stochastic models

10. Exponential distribution - e.g. the simplest model for lifetime

$X \sim Expo(\theta), \quad \theta \in (0, \infty)$

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right), & x \in [0, \infty) \\ 0, & x \in (-\infty, 0) \end{cases}$$

$E(X) = \theta$

$Var(X) = \theta^2$

Example of stochastic models

11. Gamma distribution – e.g. general model for lifetime

$X \sim Gamma(\alpha, \theta), \quad \alpha \in (0, \infty), \quad \theta \in (0, \infty)$

$$f(x; \alpha, \theta) = \begin{cases} \frac{1}{\Gamma(\alpha) \cdot \theta^{\alpha}} \cdot x^{\alpha - 1} \cdot \exp\left(-\frac{x}{\theta}\right), & x \in [0, \infty) \\ 0, & x \in (-\infty, 0) \end{cases}, \text{ where}$$

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} \, dx$$

$$E(X) = \alpha \cdot \theta$$

$$Var(X) = \alpha \cdot \theta^2$$

Example of stochastic models

12. Normal (Gaussian) distribution

$$X \sim N(\mu, \sigma^2), \quad \theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$$

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \quad x \in \mathbb{R}$$

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

Example of stochastic models

Quantitative continuous random vectors

13. Normal distribution $N(d; 0, I)$

- is the product of d normal distributions $N(0,1)$

$X = (X_1, \dots, X_d)' \sim N(d; 0, I)$

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \cdot \exp\left\{-\frac{1}{2}\mathbf{x}'\mathbf{x}\right\}, \quad \mathbf{x} = (x_1, \dots, x_d)' \in \mathbb{R}^d$$

Example of stochastic models

Quantitative continuous random vectors

13. Normal distribution $N(d; 0, I)$

- is the product of d normal distributions $N(0,1)$

$X = (X_1, \dots, X_d)' \sim N(d; 0, I)$

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \cdot \exp\left\{-\frac{1}{2}\mathbf{x}'\mathbf{x}\right\}, \quad \mathbf{x} = (x_1, \dots, x_d)' \in \mathbb{R}^d$$

14. Normal distribution $N(d; \mu, \Sigma)$

$X = (X_1, \dots, X_d)' \sim N(d; \mu, \Sigma)$

$$f(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} (\det \Sigma)^{\frac{1}{2}}} \cdot \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right\},$$

$\mathbf{x} = (x_1, \dots, x_d)' \in \mathbb{R}^d$

$E(X) = \mu$

$Cov(X, X') = E\left((X - E(X))(X - E(X))'\right) = \Sigma$

Σ is symmetric and positive definite matrix

To do:

Write algorithms to generate the following random variables/vectors:

1) $X : \begin{pmatrix} a_1 & & a_m \\ p_1 & & p_m \end{pmatrix} \qquad \sum_{i=1}^{m} p_i = 1 \ , \ p_i \geq 0$

2) $X \sim Ge(\theta) \qquad \theta \in (0,1)$

3) $X = (X_1, \dots, X_d)' \sim M(n; p_1, \dots, p_d) \quad \sum_{i=1}^{d} p_i = 1 \ , \ p_i \geq 0$