

Research Article

Open Access

K. R. Koch*

Bayesian statistics and Monte Carlo methods

<https://doi.org/10.1515/jogs-2018-0003>

Received September 27, 2017; accepted January 29, 2018

Abstract: The Bayesian approach allows an intuitive way to derive the methods of statistics. Probability is defined as a measure of the plausibility of statements or propositions. Three rules are sufficient to obtain the laws of probability. If the statements refer to the numerical values of variables, the so-called random variables, univariate and multivariate distributions follow. They lead to the point estimation by which unknown quantities, i.e. unknown parameters, are computed from measurements. The unknown parameters are random variables, they are fixed quantities in traditional statistics which is not founded on Bayes' theorem. Bayesian statistics therefore recommends itself for Monte Carlo methods, which generate random variates from given distributions. Monte Carlo methods, of course, can also be applied in traditional statistics. The unknown parameters, are introduced as functions of the measurements, and the Monte Carlo methods give the covariance matrix and the expectation of these functions. A confidence region is derived where the unknown parameters are situated with a given probability. Following a method of traditional statistics, hypotheses are tested by determining whether a value for an unknown parameter lies inside or outside the confidence region.

The error propagation of a random vector by the Monte Carlo methods is presented as an application. If the random vector results from a nonlinearly transformed vector, its covariance matrix and its expectation follow from the Monte Carlo estimate. This saves a considerable amount of derivatives to be computed, and errors of the linearization are avoided. The Monte Carlo method is therefore efficient. If the functions of the measurements are given by a sum of two or more random vectors with different multivariate distributions, the resulting distribution is generally not known. The Monte Carlo methods are then needed to obtain the covariance matrix and the expectation of the sum.

Keywords: Confidence Region, Error Propagation, Hypothesis Test, Probability, Random Variable, Univariate and Multivariate Distributions

1 Introduction

The foundation of statistics is built on the theory of probability. Plausibility or uncertainty are expressed by probability. Bayesian statistics defines the probability for statements or propositions so that the probability is understood as a measure of the plausibility of statements. Degrees of probability are expressed by real numbers. Three rules are sufficient for the complete development of the theory of probability. Bayesian statistics therefore allows an intuitive approach to the method of statistics. This means analyzing measurements to estimate unknown phenomena, the so-called unknown parameters, to establish a confidence region where the unknown parameters are situated with a given probability and to test hypotheses introduced for the unknown parameters.

The paper is organized as follows: Section 2 defines Bayesian statistics and Section 3 covers distributions. Section 4 presents point estimation, estimation of a confidence region and hypothesis testing. Section 5 deals with Monte Carlo methods, which means generating random variates from given distributions. Section 6 gives the conclusions.

2 Bayesian statistics

Starting from a cause one wants to deduce the consequences. This type of reasoning is used in mathematics. Deductive reasoning is thus well suited for mathematical proofs.

Often after observing certain effects, one would like to obtain the underlying causes. Instead of deductive reasoning, one is therefore faced with plausible reasoning. It

*Corresponding Author: K. R. Koch: Institute for Geodesy and Geoinformation, Theoretical Geodesy Group, University of Bonn, Nussallee 17, 53115 Bonn, Germany, E-mail: koch@geod.uni-bonn.de

allows to deduce possible causes from effects. The effects are registered by observations or measurements. Analyzing these data may lead to the possible causes, cf. Zellner [1971], Box and Tiao [1973], Koch [2007a], Koch [2018].

2.1 Conditional probability

A statement or a proposition depends in general on the question, whether a further statement is true. One writes $A|B$ to denote the situation that A is true under the condition that B is true. The probability of $A|B$ is called conditional probability and is denoted by

$$P(A|B) . \quad (1)$$

It gives a measure for the plausibility of the statement $A|B$ or a measure for the uncertainty of the plausible reasoning. Conditional probabilities are well suited to express empirical knowledge. The statement B points to available knowledge and $A|B$ to the statement A in the context specified by B . The probability $P(A|B)$ expresses how available knowledge is relevant for additional knowledge. This representation allows to structure knowledge and to consider the change of knowledge. Decisions under uncertainties can therefore be reached in case of changing information.

2.2 Laws of probability

The quantitative laws, which are fulfilled by the conditional probability, may be derived solely by logical and consistent reasoning. This was shown by Cox [1946]. He introduces a certain degree of plausibility for the statement $A|B$, i.e. for the statement that A is true given that B is true. Jaynes [2003] formulates three basic requirements for the plausibility:

1. Degrees of plausibility are represented by real numbers.
2. The qualitative correspondence with common sense is asked for.
3. The reasoning has to be consistent.

A relation is derived between the plausibility of the product AB and the plausibility of the statement A and the statement B given that the proposition C is true, which contains additional information or background information about the context in which statements A and B are being made. The probability is introduced as a function of the plausibility. Using this approach, Cox [1946], see also Jaynes [2003], Loredó [1990], and Sivia [1996], obtains by extensive derivations, which need not to be given here, the

product rule of probability

$$P(AB|C) = P(A|C)P(B|AC) = P(B|C)P(A|BC) \quad (2)$$

with

$$P(S|C) = 1 \quad (3)$$

where $P(S|C)$ denotes the probability of the sure statement, i.e. the statement S is with certainty true given that C is true. From the relation between the plausibility of the statement A and the plausibility of its negation \bar{A} under the condition C , the sum rule of probability follows

$$P(A|C) + P(\bar{A}|C) = 1 . \quad (4)$$

Cox's derivations only show that the sum should be set to a constant to obtain a scale for the problem. Unity is chosen in agreement with traditional statistics which is not founded on Bayes' theorem Eq. (17).

If $S|C$ in Eq. (3) denotes the sure statement, then $\bar{S}|C$ is the impossible statement, i.e. \bar{S} is certainty false given that C is true. The probability of this impossible statement follows with $P(S|C) + P(\bar{S}|C) = 1$ from Eq. (4) and $P(S|C) = 1$ from Eq. (3) by

$$P(\bar{S}|C) = 0 . \quad (5)$$

Thus, the probability $P(A|C)$ is a real number between zero and one

$$0 \leq P(A|C) \leq 1 . \quad (6)$$

It has to be mentioned that the three rules Eq. (2), Eq. (3) and Eq. (4) are sufficient to derive the following laws of probability which are needed in Bayesian statistics. They are obtained, as explained at the beginning of this Section 2.2, by logical and consistent reasoning.

2.3 Generalized Sum Rule

By applying De Morgan's law, the generalized sum rule

$$P(A + B|C) = P(A|C) + P(B|C) - P(AB|C) \quad (7)$$

is obtained instead of the sum rule Eq. (4). Let the statements A and B in Eq. (7) now be mutually exclusive. It means that the condition C requires that A and B cannot be simultaneously true, thus

$$P(AB|C) = 0 . \quad (8)$$

We get with Eq. (8) instead of Eq. (7) the generalized sum rule for the two mutually exclusive statements A and B by

$$P(A + B|C) = P(A|C) + P(B|C) . \quad (9)$$

This rule shall be generalized to the case of n mutually exclusive statements A_1, A_2, \dots, A_n . Hence, Eq. (8) gives

$$P(A_i A_j | C) = 0 \quad \text{for } i \neq j, i, j \in \{1, \dots, n\}, \quad (10)$$

and we obtain for $n = 3$ with Eq. (7) and with Eq. (9)

$$\begin{aligned} P(A_1 + A_2 + A_3 | C) &= P(A_1 + A_2 | C) + P(A_3 | C) \\ &\quad - P((A_1 + A_2)A_3 | C) \\ &= P(A_1 | C) + P(A_2 | C) + P(A_3 | C) \end{aligned}$$

because of

$$P((A_1 + A_2)A_3 | C) = P(A_1 A_3 | C) + P(A_2 A_3 | C) = 0$$

which follows with the distributive law and with Eq. (10). Correspondingly we find

$$P(A_1 + A_2 + \dots + A_n | C) = P(A_1 | C) + P(A_2 | C) + \dots + P(A_n | C). \quad (11)$$

If the statements A_1, A_2, \dots, A_n are not only mutually exclusive but also exhaustive, which means that the background information C stipulates that one and only one statement must be true and if one is true the remaining statements must be false, then we obtain with Eq. (3) from Eq. (11)

$$P(A_1 + A_2 + \dots + A_n | C) = \sum_{i=1}^n P(A_i | C) = 1. \quad (12)$$

2.4 Bayes' Theorem

The probability of the statement AB given C and the probability of the statement $A\bar{B}$ given C , where \bar{B} denotes the negation of B , follow from the product rule Eq. (2). Thus, we obtain after adding the probabilities

$$P(AB | C) + P(A\bar{B} | C) = [P(B|AC) + P(\bar{B}|AC)]P(A|C). \quad (13)$$

The sum rule Eq. (4) leads to

$$P(B|AC) + P(\bar{B}|AC) = 1$$

and therefore

$$P(A|C) = P(AB|C) + P(A\bar{B}|C). \quad (14)$$

If instead of AB and $A\bar{B}$ the statements AB_1, AB_2, \dots, AB_n under the condition C are given, we find in analogy to Eq. (13)

$$\begin{aligned} P(AB_1 | C) + P(AB_2 | C) + \dots + P(AB_n | C) \\ = [P(B_1|AC) + P(B_2|AC) + \dots + P(B_n|AC)]P(A|C). \end{aligned}$$

If B_1, \dots, B_n given C are mutually exclusive and exhaustive statements, we find with Eq. (12) the generalization of Eq. (14)

$$P(A|C) = \sum_{i=1}^n P(AB_i | C) \quad (15)$$

or with Eq. (2)

$$P(A|C) = \sum_{i=1}^n P(B_i | C)P(A|B_i C). \quad (16)$$

If the product rule Eq. (2) is solved for $P(A|BC)$, Bayes' theorem is obtained

$$P(A|BC) = \frac{P(A|C)P(B|AC)}{P(B|C)}. \quad (17)$$

In common applications of Bayes' theorem, A denotes the statement about an unknown phenomenon. B represents the statement which contains information about the unknown phenomenon and C the statement for background information. $P(A|C)$ is called prior probability, $P(A|BC)$ posterior probability and $P(B|AC)$ likelihood. The prior probability of the statement concerning the unknown phenomenon, before information has been gathered, is modified by the likelihood, that is by the probability of the information given the statement about the phenomenon. This leads to the posterior probability of the statement about the unknown phenomenon under the condition that the information is available. The probability $P(B|C)$ in the denominator of Bayes' theorem may be interpreted as normalization constant which will be shown by Eq. (19).

The bibliography of Thomas Bayes, creator of Bayes' theorem, and references for the publications of Bayes' theorem may be found in Press [1989, p. 15 and 173].

If mutually exclusive and exhaustive statements A_1, A_2, \dots, A_n are given, we obtain with Eq. (16) for the denominator of Eq. (17)

$$P(B|C) = \sum_{i=1}^n P(A_i | C)P(B|A_i C) \quad (18)$$

and Bayes' theorem Eq. (17) takes on the form

$$P(A_i | BC) = P(A_i | C)P(B|A_i C) / c \quad \text{for } i \in \{1, \dots, n\}. \quad (19)$$

The constant c acts as a normalization constant because the probabilities of the statements A_i sum with

$$c = \sum_{i=1}^n P(A_i | C)P(B|A_i C) \quad (20)$$

to one

$$\sum_{i=1}^n P(A_i | BC) = \sum_{i=1}^n P(A_i | C)P(B|A_i C) = 1 \quad (21)$$

in agreement with Eq. (12), see also the equivalent result Eq. (62) for the generalized Bayes' theorem.

The normalization constant in Eq. (19) is frequently omitted, in which case Bayes' theorem Eq. (17) is represented by

$$P(A|BC) \propto P(A|C)P(B|AC) \quad (22)$$

where \propto denotes proportionality. Hence,

posterior probability \propto prior probability \times likelihood .

3 Distributions

So far, the statements have been kept very general. In the following, they shall refer to the numerical values of variables, i.e. to real numbers. They are called random variables in agreement with the terminology of traditional statistics, although the statements may refer to any variables not only to values resulting from random experiments. Random variables frequently applied in the sequel are the unknown parameters which describe unknown phenomena. They represent in general fixed quantities, for instance, the unknown coordinates of a point at the rigid surface of the earth. The unknown parameters are treated in detail in Section 3.5. Random variables are also given as measurements, observations or data. They follow from measuring experiments or in general from random experiments whose results are registered digitally. Another source of data are surveys with numerical results. The measurements or observations are carried out and the data are collected to gain information about the unknown parameters. The analysis of the data is explained in Section 4.

The rules obtained for the probabilities of statements and Bayes' theorem hold analogously for the probability density functions of random variables, which are derived by the statements concerning their values. To get these results, the rules of probability derived so far are sufficient. Summing the probability density functions in case of discrete random variables has only to be replaced by an integration in case of continuous random variables. We will treat in the following only continuous random variables and simplify the notation by omitting the reference to the background information C .

3.1 Continuous Distribution

Let X be a continuous random variable with the values $x \in \mathbb{R}$ in the interval $-\infty < x < \infty$. The probability $P(X < x)$ of

the statement $X < x$ is called the probability distribution function or shortly distribution function $F(x)$

$$F(x) = P(X < x) . \quad (23)$$

Let the distribution function $F(x)$ be continuous and continuously differentiable. Given these assumptions, the probability shall be determined that the random variable X takes on values in the interval $a \leq X < b$. The following three statements are considered

$$X < a, \quad X < b \quad \text{and} \quad a \leq X < b .$$

The statement $X < b$ results as the sum of the statements $X < a$ and $a \leq X < b$. Since the two latter statements are mutually exclusive, we find by the sum rule Eq. (9)

$$P(X < b) = P(X < a) + P(a \leq X < b)$$

or

$$P(a \leq X < b) = P(X < b) - P(X < a)$$

and with the distribution function Eq. (23)

$$P(a \leq X < b) = F(b) - F(a) .$$

The fundamental theorem of calculus gives, cf. Blatter [1974, II, p.15],

$$F(b) - F(a) = \int_{-\infty}^b p(x)dx - \int_{-\infty}^a p(x)dx = \int_a^b p(x)dx \quad (24)$$

with

$$dF(x)/dx = p(x) . \quad (25)$$

We call $p(x)$ the continuous probability density function or, abbreviated, continuous density function, also continuous distribution for the random variable X . The distribution for a one-dimensional continuous random variable X is called univariate distribution.

The distribution function $F(x)$ from Eq. (23) follows with the density function $p(x)$ according to Eq. (24) as an area function by

$$F(x) = \int_{-\infty}^x p(t)dt \quad (26)$$

where t denotes the variable of integration. The distribution function $F(x)$ of a continuous random variable is obtained by an integration of the continuous probability density function $p(x)$.

We obtain $P(a \leq X < b) = P(a < X < b)$ because of Eq. (24). Thus, we will only work with open intervals $a < X < b$ in the sequel. We find for the interval $x < X < x + dx$ with Eq. (24)

$$P(x < X < x + dx) = p(x)dx \quad (27)$$

because of

$$\int_{-\infty}^{x+dx} p(t)dt = \int_{-\infty}^x p(t)dt + \int_{-\infty}^{x+dx} p(t)dt = F(x) + F(x+dx)$$

and therefore

$$dF(x)/dx = p(x) \quad \text{and} \quad dF(x+dx)/dx = p(x)dx .$$

The values x of the random variable X are defined by the interval $-\infty < x < \infty$ so that $X < \infty$ represents an exhaustive statement. It therefore follows from Eq. (12)

$$F(\infty) = P(X < \infty) = 1 .$$

Because of Eq. (6), we have $F(x) \geq 0$, which according to Eq. (26) is only valid if $p(x) \geq 0$. Thus, the two conditions are obtained which the probability density function $p(x)$ for the continuous random variable X has to fulfill

$$p(x) \geq 0 \quad \text{for} \quad \int_{-\infty}^{\infty} p(x)dx = 1 . \quad (28)$$

We find for the distribution function $F(x)$ with Eq. (5), since $F(-\infty)$ represents the probability of the impossible statement,

$$F(-\infty) = 0, \quad F(\infty) = 1 \quad \text{and} \quad F(x_i) \leq F(x_j) \quad \text{for} \quad x_i < x_j . \quad (29)$$

The statement $X < x_j$ follows from the sum of the two statements $X < x_i$ and $x_i < X < x_j$. The latter statements are mutually exclusive and $P(x_i < X < x_j) \geq 0$ holds, therefore $P(X < x_j) \leq P(X < x_i)$.

The uniform distribution is introduced as an example. The random variable X has the uniform distribution with parameters a and b , if its probability density function $p(x|a, b)$ is given by

$$p(x|a, b) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for } x < a \quad \text{and} \quad x > b . \end{cases} \quad (30)$$

As the density function $p(x)$ is constant in the interval $a \leq x \leq b$, one speaks of a uniform distribution. The distribution function $F(x; a, b)$ of the uniform distribution is computed with Eq. (26) by

$$F(x; a, b) = \int_a^x \frac{dt}{b-a} \quad \text{for} \quad a \leq x \leq b .$$

We therefore obtain

$$F(x; a, b) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \leq x \leq b \\ 1 & \text{for } x > b . \end{cases} \quad (31)$$

The density function Eq. (30) satisfies both conditions of Eq. (28) because of $p(x|a, b) \geq 0$ and $F(\infty; a, b) = 1$. An example for a multivariate uniform distribution will be presented in Section 4.1.

3.2 Multidimensional Continuous Distributions

An n -dimensional continuous random variable X_1, \dots, X_n takes on the values $x_1, \dots, x_n \in \mathbb{R}$ in the intervals $-\infty < x_k < \infty$ with $k \in \{1, \dots, n\}$. The distribution function $F(x_1, \dots, x_n)$ for this random variable is defined as in Eq. (23) by

$$F(x_1, \dots, x_n) = P(X_1 < x_1, \dots, X_n < x_n) . \quad (32)$$

It represents corresponding to Eq. (24) the probability that the random variables X_k take on values in the given intervals $x_{ku} < X_k < x_{ko}$ for $k \in \{1, \dots, n\}$

$$\begin{aligned} P(x_{1u} < X_1 < x_{1o}, \dots, x_{nu} < X_n < x_{no}) \\ = \int_{x_{nu}}^{x_{no}} \dots \int_{x_{1u}}^{x_{1o}} p(x_1, \dots, x_n) dx_1 \dots dx_n \end{aligned} \quad (33)$$

with

$$\partial^n F(x_1, \dots, x_n) / \partial x_1 \dots \partial x_n = p(x_1, \dots, x_n) . \quad (34)$$

We call $p(x_1, \dots, x_n)$ the n -dimensional continuous probability density function or, abbreviated, continuous density function or multivariate distribution for the n -dimensional continuous random variable X_1, \dots, X_n .

The distribution function $F(x_1, \dots, x_n)$ is obtained with Eq. (32) by the density function $p(x_1, \dots, x_n)$

$$F(x_1, \dots, x_n) = \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_1} p(t_1, \dots, t_n) dt_1 \dots dt_n \quad (35)$$

where t_1, \dots, t_n denote the variables of integration. The conditions which a density function $p(x_1, \dots, x_n)$ has to fulfill follows in analogy to Eq. (28) with

$$\begin{aligned} p(x_1, \dots, x_n) \geq 0 \quad \text{and} \\ \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} p(x_1, \dots, x_n) dx_1 \dots dx_n = 1 . \end{aligned} \quad (36)$$

The n -dimensional continuous random variable X_1, \dots, X_n will be often denoted in the following by the $n \times 1$ continuous random vector \mathbf{x} with

$$\mathbf{x} = |X_1, \dots, X_n|' . \quad (37)$$

The values of the random vector of a continuous random variable are collected in the vector \mathbf{x} with

$$\mathbf{x} = |x_1, \dots, x_n|', \quad -\infty < x_k < \infty, \quad k \in \{1, \dots, n\}. \quad (38)$$

The reason for not distinguishing in the sequel between the vector \mathbf{x} of random variables and the vector \mathbf{x} of values of the random variables follows from the notation of vectors and matrices. It labels the vectors by small letters and the matrices by capital letters. If the distinction is necessary, it will be explained by additional comments.

The n -dimensional continuous probability density function for the continuous random vector \mathbf{x} follows with Eq. (37) or Eq. (38) instead of Eq. (34) by

$$p(\mathbf{x}). \quad (39)$$

3.3 Marginal Distribution

If an n -dimensional continuous random variable X_1, \dots, X_n is given, the marginal distribution function for X_1, \dots, X_i is obtained in analogy to Eq. (32) up to Eq. (35) by

$$\begin{aligned} F(x_1, \dots, x_i, \infty, \dots, \infty) &= P(X_1 < x_1, \dots, X_i < x_i) \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \int_{-\infty}^{x_i} \dots \int_{-\infty}^{x_1} p(t_1, \dots, t_i, t_{i+1}, \dots, t_n) dt_1 \dots dt_n \end{aligned} \quad (40)$$

with the variables t_1, \dots, t_n of integration. The marginal probability density function $p(x_1, \dots, x_i)$ for X_1, \dots, X_i results from

$$\partial^i F(x_1, \dots, x_i, \infty, \dots, \infty) / \partial x_1 \dots \partial x_i = p(x_1, \dots, x_i) \quad (41)$$

hence

$$\begin{aligned} p(x_1, \dots, x_i) \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} p(x_1, \dots, x_i, x_{i+1}, \dots, x_n) dx_{i+1} \dots dx_n. \end{aligned} \quad (42)$$

The marginal density function $p(x_1, \dots, x_i)$ therefore follows by integrating over the values of the random variables X_{i+1}, \dots, X_n . We introduce with Eq. (38)

$$\mathbf{x}_1 = |x_1, \dots, x_i|' \quad \text{and} \quad \mathbf{x}_2 = |x_{i+1}, \dots, x_n|' \quad (43)$$

and get in a more compact notation

$$p(\mathbf{x}_1) = \int_{\mathcal{X}_2} p(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_2 \quad (44)$$

where \mathcal{X}_2 denotes the domain for integrating \mathbf{x}_2 .

3.4 Conditional Distribution

To find the conditional continuous probability density function for a two-dimensional continuous random variable X_1, X_2 , one has to be aware that $P(X_2 = x_2) = 0$ holds for the continuous random variable X_2 because of Eq. (24). The statement AB in the product rule Eq. (2) must therefore be formulated such that we obtain for the continuous random variable X_1, X_2

$$\begin{aligned} &P(X_1 < x_1, x_2 < X_2 < x_2 + \Delta x_2) \\ &= P(x_2 < X_2 < x_2 + \Delta x_2) P(X_1 < x_1 | x_2 < X_2 < x_2 + \Delta x_2). \end{aligned}$$

This leads with Eq. (17) and Eq. (40), with the density function $p(x_1, x_2)$ for X_1 and X_2 and with the variables t_1 and t_2 of integration to, cf. Koch [1999, p. 91],

$$P(X_1 < x_1 | x_2 < X_2 < x_2 + \Delta x_2) = \frac{\int_{x_2}^{x_2 + \Delta x_2} \int_{-\infty}^{x_1} p(t_1, t_2) dt_1 dt_2}{\int_{x_2}^{x_2 + \Delta x_2} \int_{-\infty}^{\infty} p(t_1, t_2) dt_1 dt_2}. \quad (45)$$

The probability $P(X_1 < x_1 | X_2 = x_2)$, which corresponds to the conditional distribution function $F(x_1 | x_2)$, needs to be determined. It follows from Eq. (45) by the limit process $\Delta x_2 \rightarrow 0$. We find with the marginal density $p(x_2) = \int_{-\infty}^{\infty} p(t_1, t_2) dt_1$ from Eq. (42)

$$\begin{aligned} F(x_1 | x_2) &= \lim_{\Delta x_2 \rightarrow 0} P(X_1 < x_1 | x_2 < X_2 < x_2 + \Delta x_2) \\ &= \frac{\int_{-\infty}^{x_1} p(t_1, t_2) dt_1 \Delta x_2}{p(x_2) \Delta x_2}. \end{aligned} \quad (46)$$

By differentiating with respect to x_1 in analogy to Eq. (34), the conditional continuous probability density function $p(x_1 | x_2)$ for X_1 is obtained under the conditions that the value x_2 of X_2 is given

$$p(x_1 | x_2) = \frac{p(x_1, x_2)}{p(x_2)}. \quad (47)$$

Starting from the n -dimensional continuous random variable X_1, \dots, X_n with the density function $p(x_1, \dots, x_n)$, the conditional continuous density function for the random variables X_1, \dots, X_i given the values x_{i+1}, \dots, x_n for X_{i+1}, \dots, X_n is obtained analogously by

$$p(x_1, \dots, x_i | x_{i+1}, \dots, x_n) = \frac{p(x_1, \dots, x_i, x_{i+1}, \dots, x_n)}{p(x_{i+1}, \dots, x_n)}. \quad (48)$$

If corresponding to Eq. (37) the continuous random variables are arranged in the continuous random vectors

$$\mathbf{x}_1 = |X_1, \dots, X_i|' \quad \text{and} \quad \mathbf{x}_2 = |X_{i+1}, \dots, X_n|' \quad (49)$$

and corresponding to Eq. (38) the values of the continuous random variables in

$$\mathbf{x}_1 = |x_1, \dots, x_i|' \quad \text{and} \quad \mathbf{x}_2 = |x_{i+1}, \dots, x_n|', \quad (50)$$

we get instead of Eq. (48) the conditional continuous probability density function for the continuous random vector \mathbf{x}_1 given the values for \mathbf{x}_2 by

$$p(\mathbf{x}_1|\mathbf{x}_2) = \frac{p(\mathbf{x}_1, \mathbf{x}_2)}{p(\mathbf{x}_2)}. \quad (51)$$

The conditional continuous density function Eq. (48) or Eq. (51) satisfies in analogy to Eq. (28)

$$p(x_1, \dots, x_i|x_{i+1}, \dots, x_n) \geq 0 \quad (52)$$

and

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} p(x_1, \dots, x_i|x_{i+1}, \dots, x_n) dx_1 \dots dx_i = 1. \quad (53)$$

3.5 Generalized Bayes' Theorem

Bayes' theorem Eq. (17), which has been derived for the probability of statements, shall now be generalized such that it is valid for the probability density functions of continuous random variables. If \mathbf{x} and \mathbf{y} are continuous random variables, we obtain with Eq. (51)

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})} \quad (54)$$

under the condition that values of the random vector \mathbf{y} are given, which because of Eq. (38) are also denoted by \mathbf{y} . Furthermore, we have

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})}. \quad (55)$$

If these two equations are solved for $p(\mathbf{x}, \mathbf{y})$ and the resulting expressions are equated, the generalized Bayes' theorem is found in a form corresponding to Eq. (17)

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x})p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})}. \quad (56)$$

Since the vector \mathbf{y} contains fixed values, $p(\mathbf{y})$ is constant. Bayes' theorem is therefore often applied in the form corresponding to Eq. (22)

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{x})p(\mathbf{y}|\mathbf{x}). \quad (57)$$

The components of the continuous random vector \mathbf{x} are now identified with unknown parameters. i.e. quantities which describe unknown phenomena. The set of vectors \mathbf{x} , which collects all possible values for the parameters, is called the parameter space \mathcal{X} , thus $\mathbf{x} \in \mathcal{X}$. To estimate \mathbf{x} , measurements or observations, also called data, have to be taken which contain information about the unknown parameters. The values \mathbf{y} of the random vector \mathbf{y} represent

the given observations. The probability density function $p(\mathbf{x})$ contains information about the parameters \mathbf{x} before the data \mathbf{y} have been taken. One calls $p(\mathbf{x})$ therefore prior density function for the unknown parameters \mathbf{x} . By taking into account the measurements \mathbf{y} , the probability density function $p(\mathbf{x}|\mathbf{y})$ follows. It is named posterior density function for the parameters \mathbf{x} . Via the density function $p(\mathbf{y}|\mathbf{x})$ the information available in the data \mathbf{y} reaches the parameters \mathbf{x} . Since the data \mathbf{y} are given, this density function is not interpreted as a function of \mathbf{y} but as a function of the parameters \mathbf{x} . It is called likelihood function. Thus,

$$\text{posterior density function} \propto \text{prior density function} \times \text{likelihood function}. \quad (58)$$

The measurements modify the prior density function by the likelihood function and lead to the posterior density function for the unknown parameters.

The density function in the denominator of Eq. (56) may be expressed by the marginal density function Eq. (44) and then by Eq. (55) for the random vector \mathbf{y}

$$p(\mathbf{y}) = \int_{\mathcal{X}} p(\mathbf{x}, \mathbf{y}) d\mathbf{x} = \int_{\mathcal{X}} p(\mathbf{x})p(\mathbf{y}|\mathbf{x}) d\mathbf{x} \quad (59)$$

where the parameter space \mathcal{X} denotes the domain over which \mathbf{x} has to be integrated. Bayes' theorem Eq. (56) then follows in a form corresponding to Eq. (19) by

$$p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})/c \quad (60)$$

with the normalization constant c

$$c = \int_{\mathcal{X}} p(\mathbf{x})p(\mathbf{y}|\mathbf{x}) d\mathbf{x} \quad (61)$$

so that the posterior density function $p(\mathbf{x}|\mathbf{y})$ integrates with

$$\int_{\mathcal{X}} p(\mathbf{x}|\mathbf{y}) = \int_{\mathcal{X}} p(\mathbf{x})p(\mathbf{y}|\mathbf{x}) = 1 \quad (62)$$

to one in agreement with Eq. (53).

The vector \mathbf{x} of unknown parameters is defined in Bayes' theorem as random vector with which a prior density and a posterior density function are associated. This approach is contrary to the one of traditional statistics which defines the vector of unknown parameters as a vector of constants. But this does not mean that the vector \mathbf{x} of parameters in Bayes' theorem Eq. (56) may not represent constants like the coordinates of a point at the rigid surface of the earth. Not the constants are random, it is the knowledge about the values of the constants which are random and for which probability is introduced.

3.6 Variance and Covariance

Let $\mathbf{x} = |X_1, \dots, X_n|'$ be the $n \times 1$ continuous random vector with the probability density function $p(x_1, \dots, x_n)$, then $\mu_{x_1, \dots, x_n}^{(k)}$ with $k = \sum_{i=1}^n k_i$ and $k_i \in \mathbb{N}$ denotes the k th multivariate moment of X_1, \dots, X_n

$$\begin{aligned} \mu_{x_1, \dots, x_n}^{(k)} &= E(X_1^{k_1} X_2^{k_2} \dots X_n^{k_n}) \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} x_1^{k_1} x_2^{k_2} \dots x_n^{k_n} p(x_1, \dots, x_n) dx_1 \dots dx_n. \end{aligned} \tag{63}$$

It gives with $k_i = 1$ and $k_j = 0$ for $i \neq j$ the expected value μ_i or $E(X_i)$ of the random variable X_i .

Moments about the expected values μ_i of the random variables X_i are called central moments

$$E((X_1 - \mu_1)^{k_1} (X_2 - \mu_2)^{k_2} \dots (X_n - \mu_n)^{k_n}). \tag{64}$$

The second central moment σ_{ij} or $C(X_i, X_j)$ of the random variables X_i and X_j is called the covariance

$$\begin{aligned} \sigma_{ij} &= C(X_i, X_j) = E((X_i - \mu_i)(X_j - \mu_j)) \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) p(x_1, \dots, x_n) dx_1 \dots dx_n \end{aligned} \tag{65}$$

and the second central moment σ_i^2 or $V(X_i)$ of the random variable X_i the variance

$$\begin{aligned} \sigma_i^2 &= V(X_i) = \sigma_{ii} = C(X_i, X_i) = E((X_i - \mu_i)^2) \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} (x_i - \mu_i)^2 p(x_1, \dots, x_n) dx_1 \dots dx_n. \end{aligned} \tag{66}$$

Considering Eq. (36), it is obvious from the definition of the variance that $\sigma_i^2 \geq 0$ holds. The positive square root of σ_i^2 is called the standard deviation σ_i , thus $\sigma_i = \sqrt{\sigma_i^2}$. The variance σ_i^2 is a measure of dispersion of the random variable X_i about its expected value μ_i and therefore a measure for the accuracy of a random variable which represents a measurement.

The variances and covariances of a random vector $\mathbf{x} = |X_1, \dots, X_n|'$ are collected in the covariance matrix $D(\mathbf{x})$, also called variance-covariance matrix or dispersion matrix

$$\begin{aligned} D(\mathbf{x}) &= (\sigma_{ij}) = (C(X_i, X_j)) = E((\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))') \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \dots & \dots & \dots & \dots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{pmatrix}. \end{aligned} \tag{67}$$

It follows from Eq. (65) $\sigma_{ij} = \sigma_{ji}$ so that the covariance matrix is symmetric. In addition, it is positive definite or positive semidefinite, cf. Koch [1999, p. 100].

Let the $n \times 1$ random vector \mathbf{x} with the covariance matrix $D(\mathbf{x})$ be linearly transformed with the $m \times n$ matrix \mathbf{A} and the $m \times 1$ vector \mathbf{b} of constants into the $m \times 1$ random vector \mathbf{y} by $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$. Its $m \times m$ covariance matrix $D(\mathbf{y})$ follows with

$$D(\mathbf{y}) = D(\mathbf{A}\mathbf{x} + \mathbf{b}) = D(\mathbf{A}\mathbf{x})\mathbf{A}' \tag{68}$$

since we obtain with the definition Eq. (67) of a covariance matrix $D(\mathbf{y}) = E((\mathbf{y} - E(\mathbf{y}))(\mathbf{y} - E(\mathbf{y}))')$ and $D(\mathbf{y}) = E((\mathbf{A}\mathbf{x} + \mathbf{b} - \mathbf{A}E(\mathbf{x}) - \mathbf{b})(\mathbf{A}\mathbf{x} - \mathbf{A}E(\mathbf{x}))') = \mathbf{A}E((\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))')\mathbf{A}'$.

In case of nonlinear transformations, the matrix \mathbf{A} contains as elements the derivatives of the transformed values of the random vector \mathbf{x} with respect to the values of \mathbf{x} , cf. Koch [1999, p. 100]. Eq. (68) is then called the law of error propagation.

4 Estimation and Hypothesis testing

4.1 Point Estimation

The knowledge of the posterior density function for the unknown parameters from Bayes' theorem allows to estimate the unknown parameters. This is called point estimation. A decision is needed for the estimation, and it results in a loss. The true state of the system is characterized by the true values \mathbf{x} of the unknown parameters. The loss is a function of the estimates $\hat{\mathbf{x}}$ and of the true values \mathbf{x} of the parameters so that it will be called $L(\mathbf{x}, \hat{\mathbf{x}})$. It must express the quality of the estimate $\hat{\mathbf{x}}$. It therefore has to increase for bad estimates.

A simple loss function results from the sum of squares $(\mathbf{x} - \hat{\mathbf{x}})'(\mathbf{x} - \hat{\mathbf{x}})$ of the errors $\mathbf{x} - \hat{\mathbf{x}}$ of the estimates $\hat{\mathbf{x}}$ of the unknown parameters \mathbf{x} . This sum of squares is generalized by means of the covariance matrix $D(\mathbf{x}) = \Sigma$ for the random vector \mathbf{x} . It shall be positive definite. Its inverse Σ^{-1} is proportional to the weight matrix of \mathbf{x} , cf. [Koch, 1999, p. 104]. Thus, the quadratic loss function

$$L(\mathbf{x}, \hat{\mathbf{x}}) = (\mathbf{x} - \hat{\mathbf{x}})' \Sigma^{-1} (\mathbf{x} - \hat{\mathbf{x}}) \tag{69}$$

is chosen where the squares of the errors $\mathbf{x} - \hat{\mathbf{x}}$ are weighted by Σ^{-1} . This loss function leads to the well known method of least squares for the linear model.

To determine the posterior expected loss of the quadratic loss function Eq. (69), its expected value has to

be computed by the posterior density function $p(\mathbf{x}|\mathbf{y})$ for \mathbf{x} from Eq. (56). We obtain

$$E[L(\mathbf{x}, \hat{\mathbf{x}})] = \int_{\mathcal{X}} (\mathbf{x} - \hat{\mathbf{x}})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \hat{\mathbf{x}}) p(\mathbf{x}|\mathbf{y}) d\mathbf{x} \quad (70)$$

where \mathcal{X} denotes the parameter space over which the values of \mathbf{x} have to be integrated. The minimum of the expected loss is obtained by, cf. [Koch, 2007a, p. 66],

$$\hat{\mathbf{x}}_B = E(\mathbf{x}) . \quad (71)$$

The expected value is computed according to Eq. (70) by the posterior density function $p(\mathbf{x}|\mathbf{y})$ for \mathbf{x} . The estimate $\hat{\mathbf{x}}_B$, which is called Bayes estimate, therefore follows with

$$\hat{\mathbf{x}}_B = \int_{\mathcal{X}} \mathbf{x} p(\mathbf{x}|\mathbf{y}) d\mathbf{x} . \quad (72)$$

The vector \mathbf{x} of unknown parameters is a random vector. Its Bayes estimate $\hat{\mathbf{x}}_B$ is a fixed quantity. To express the dispersion of $\hat{\mathbf{x}}_B$, the covariance matrix $D(\mathbf{x})$ for the vector \mathbf{x} of unknown parameters is introduced according to Eq. (67) by the posterior density function $p(\mathbf{x}|\mathbf{y})$

$$\begin{aligned} D(\mathbf{x}) &= E((\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))') \\ &= \int_{\mathcal{X}} (\mathbf{x} - \hat{\mathbf{x}}_B)(\mathbf{x} - \hat{\mathbf{x}}_B)' p(\mathbf{x}|\mathbf{y}) d\mathbf{x} . \end{aligned} \quad (73)$$

We will look at the more general case that the $n \times 1$ vector \mathbf{x} of unknown parameters is defined by the $n \times 1$ vector $\mathbf{f}(\mathbf{y})$ of functions of the $n \times 1$ random vector \mathbf{y} of measurements

$$\mathbf{x} = \mathbf{f}(\mathbf{y}) . \quad (74)$$

A simple example of the functions $\mathbf{f}(\mathbf{y})$ is the sum of the vector of measurements and of the vector of systematic effects in laser scanning, the former vector with a multivariate normal distribution and the latter vector with a multivariate uniform distribution [Koch and Brockmann, 2016]. The expected value $E(\mathbf{f}(\mathbf{y}))$ and the covariance matrix $D(\mathbf{f}(\mathbf{y}))$ of the vector $\mathbf{f}(\mathbf{y})$ follow by the probability density function $p(\mathbf{x}|\mathbf{y})$ for the vector \mathbf{x} by

$$E(\mathbf{f}(\mathbf{y})) = \int_{\mathcal{X}} \mathbf{f}(\mathbf{y}) p(\mathbf{x}|\mathbf{y}) d\mathbf{x} , \quad (75)$$

$$D(\mathbf{f}(\mathbf{y})) = \int_{\mathcal{X}} (\mathbf{f}(\mathbf{y}) - E(\mathbf{f}(\mathbf{y})))(\mathbf{f}(\mathbf{y}) - E(\mathbf{f}(\mathbf{y})))' p(\mathbf{x}|\mathbf{y}) d\mathbf{x} . \quad (76)$$

Let m random variates \mathbf{x}_j be generated for the vector \mathbf{x} from the probability density function $p(\mathbf{x}|\mathbf{y})$

$$\mathbf{x}_j = (x_{ij}) \quad \text{for } i \in \{1, \dots, n\}, j \in \{1, \dots, m\} . \quad (77)$$

The Monte Carlo estimate $\hat{D}(\mathbf{f}(\mathbf{y}))$ of the $n \times n$ covariance matrix $D(\mathbf{f}(\mathbf{y}))$ of the vector $\mathbf{f}(\mathbf{y})$ is then obtained by the sum, cf. Leonard and Hsu [1999, p. 275], Koch [2007a, p. 225],

$$\hat{D}(\mathbf{f}(\mathbf{y})) = \frac{1}{m} \sum_{j=1}^m (\mathbf{x}_j - \hat{E}(\mathbf{x}))(\mathbf{x}_j - \hat{E}(\mathbf{x}))' \quad (78)$$

with the estimate $\hat{E}(\mathbf{f}(\mathbf{y}))$ of the expectations $E(\mathbf{f}(\mathbf{y}))$ and with the estimate $\hat{E}(\mathbf{x})$ of the expectation $E(\mathbf{x})$

$$\hat{E}(\mathbf{f}(\mathbf{y})) = \hat{E}(\mathbf{x}) = \frac{1}{m} \sum_{j=1}^m \mathbf{x}_j . \quad (79)$$

The example cited for the vector $\mathbf{f}(\mathbf{y})$ from Eq. (74) is normally distributed for the vector of measurements and multivariate uniformly distributed for the systematic effects. The Monte Carlo estimate is then needed, as the sum is not normally distributed anymore. This has also been pointed out by Alkhatib and Kutterer [2013], who determined by Monte Carlo methods the variances and covariances of measurements.

Random variates \mathbf{x}_j are generated for the unknown parameters \mathbf{x} by Eq. (77). The unknown parameters \mathbf{x} are random variables while in traditional statistics they are fixed quantities. Bayesian statistics therefore recommends itself for Monte Carlo methods. The Monte Carlo estimate $\hat{D}(\mathbf{f}(\mathbf{y}))$ of the vector $\mathbf{f}(\mathbf{y})$ of measurements and the estimate $\hat{E}(\mathbf{f}(\mathbf{y}))$ of the expectations of $\mathbf{f}(\mathbf{y})$ establish the Monte Carlo methods. The law of error propagation was derived by Eq. (68). It is a method of traditional statistics. Monte Carlo methods for solving the law are therefore not restricted to Bayesian statistics but can be applied by traditional statistics as well. This intimate connection between Bayesian statistics and traditional statistics is a new result, it has not been mentioned before in any paper.

Outcomes which otherwise cannot be obtained, like the results for the example mentioned in connection with Eq. (74), are supplied by the Monte Carlo methods. However, one has to be aware that 100 000 random variates need to be generated to get the covariances and expectations with at least two significant figures [Koch, 2008]. This could be considered a disadvantage of the Monte Carlo methods. However in view of the speed of computers nowadays, this is not a serious drawback. Examples for Monte Carlo methods in addition to the one mentioned for Eq. (74) are given in Section 5.

4.2 Estimation of Confidence Region

In contrast to the point estimation of Section 4.1, a region shall now be determined in which the vector \mathbf{x} of unknown

parameters is situated with a given probability. We obtain by the posterior density function $p(\mathbf{x}|\mathbf{y})$ for \mathbf{x} from Bayes' theorem Eq. (56) the probability

$$P(\mathbf{x} \in \mathcal{X}_u|\mathbf{y}) = \int_{\mathcal{X}_u} p(\mathbf{x}|\mathbf{y})d\mathbf{x} \quad (80)$$

that the vector \mathbf{x} belongs to the subspace \mathcal{X}_u of the parameter space \mathcal{X} with $\mathcal{X}_u \subset \mathcal{X}$. One is often interested to find the subspace where most of the probability, for instance 95%, is concentrated. Given a probability, there are obviously many possibilities to establish such a subspace. A region of values for \mathbf{x} within the subspace, however, should be more probable than a region outside the subspace. It will be therefore required that the probability density function of each point within the subspace is equal to or greater than the density function of a point outside the subspace. The region of highest posterior density, also called H.P.D. region, is thus obtained.

Let $p(\mathbf{x}|\mathbf{y})$ be the posterior density function for the vector \mathbf{x} of unknown parameters, the subspace \mathcal{X}_B with $\mathcal{X}_B \subset \mathcal{X}$ is then called a $1 - \alpha$ H.P.D. region or confidence region, if

$$P(\mathbf{x} \in \mathcal{X}_B|\mathbf{y}) = \int_{\mathcal{X}_B} p(\mathbf{x}|\mathbf{y})d\mathbf{x} = 1 - \alpha \quad \text{and} \\ p(\mathbf{x}_1|\mathbf{y}) \geq p(\mathbf{x}_2|\mathbf{y}) \quad \text{for } \mathbf{x}_1 \in \mathcal{X}_B, \mathbf{x}_2 \notin \mathcal{X}_B. \quad (81)$$

The value for α is generally chosen to be $\alpha = 0.05$, but also $\alpha = 0.1$ or $\alpha = 0.01$ are selected.

Let the vector \mathbf{x} of unknown parameters contain only the random variable x as component. The confidence interval for x is then defined by Eq. (81). It is simply computed by Monte Carlo methods. Random variates x_j with $j \in \{1, \dots, m\}$ for x are generated, see Eq. (77). The interval between the maximum and the minimum of the m values is divided into small intervals, let say $m - 1$ intervals. The probabilities of the random variates x_j lying within the small intervals are determined by the relative frequencies. The probabilities at both ends of the interval are then added such that Eq. (81) is fulfilled, until the probability α is reached [Koch, 2008].

To get a confidence ellipse, random variates x_{ij} with $i, j \in \{1, \dots, m\}$ for the vector \mathbf{x} are generated from $p(\mathbf{x}|\mathbf{y})$ in the rectangular region, where the confidence ellipse is situated. The region is divided into a mesh of small rectangles. They are ordered in terms of decreasing probabilities obtained by relative frequencies. The probabilities are added at both ends of the interval such that Eq. (81) is fulfilled, until the probability α is reached [JCGM, 2011].

To obtain three-dimensional confidence ellipsoids, again random variates x_{ij} with $i, j \in \{1, \dots, m\}$ for

the vector \mathbf{x} are generated from $p(\mathbf{x}|\mathbf{y})$ in the three-dimensional region, where the confidence ellipsoid is situated. The region is divided into small rectangular volume elements. Relative frequencies are computed for the volume elements. The frequencies are scaled to the interval $[0,1]$ to find the probability density functions p_{ij} for $i, j \in \{1, \dots, m\}$. The densities are sorted in decreasing order. However, even for a large number of generated random variates, there will be many volume elements having the same densities, especially at the boundary of the confidence ellipsoid so that the densities cannot be ordered.

The following method has been applied to solve this problem. The densities p_{ij} are arranged together with the coordinates of the volume elements in decreasing order. The highest density value p_{ij} together with the coordinates of the volume element determine the center of the confidence ellipsoid. Distances are computed from the center of the confidence ellipsoid to the volume elements and sorted in increasing order together with the densities p_{ij} and the coordinates of the volume elements. The densities are summed such that Eq. (81) is fulfilled, until the probability α is reached [Koch and Brockmann, 2016].

To construct a confidence hyperellipsoid, the probability density functions $p(x_j|\mathbf{y})$ for the random variates x_j from $p(\mathbf{x}|\mathbf{y})$ are sorted in decreasing order. By summing the density functions up to the index b for which $1 - \alpha$ holds true, the point \mathbf{x}_b with $p_b = p(\mathbf{x}_b|\mathbf{y})$ at the boundary of the confidence hyperellipsoid is obtained. If the vectors \mathbf{x}_j are graphically depicted as points and if neighboring points are selected with smaller density values than p_b and with larger ones, the boundary of the confidence hyperellipsoid is obtained by interpolating between these points. This method, of course, can also be applied to compute confidence ellipses or confidence ellipsoids.

4.3 Hypothesis Testing

Propositions concerning the unknown parameters may be formulated as hypotheses, and methods for deciding whether to accept or to reject the hypotheses are called hypothesis tests. Let $\mathcal{X}_0 \subset \mathcal{X}$ and $\mathcal{X}_1 \subset \mathcal{X}$ be subspaces of the parameter space \mathcal{X} and let \mathcal{X}_0 and \mathcal{X}_1 be disjoint, i.e. $\mathcal{X}_0 \cap \mathcal{X}_1 = \emptyset$. The assumption that the vector \mathbf{x} of unknown parameters is element of the subspace \mathcal{X}_0 is called the null hypothesis H_0 and the assumption that \mathbf{x} is element of \mathcal{X}_1 the alternative hypothesis H_1 .

We are interested in the point null hypothesis. It is defined by

$$H_0 : \mathbf{x} = \mathbf{x}_0 \quad \text{versus} \quad H_1 : \mathbf{x} \neq \mathbf{x}_0. \quad (82)$$

where only the vector \mathbf{x}_0 in \mathcal{X}_0 is given and \mathcal{X}_1 is the complement of \mathcal{X}_0 . According to a method of traditional statistics, the point null hypothesis is tested such that it is accepted if the point \mathbf{x}_0 lies within the $1 - \alpha$ confidence region for the parameter vector [Koch, 1999, p. 301]. This is the case, if the inequality

$$p(\mathbf{x}_0|\mathbf{y}) > p(\mathbf{x}_b|\mathbf{y}) \quad (83)$$

is fulfilled where \mathbf{x}_b denotes a point at the boundary of the confidence region \mathcal{X}_B . Its density p_b is introduced by

$$p_b = p(\mathbf{x}_b|\mathbf{y}) \quad (84)$$

and a point \mathbf{x}_0 lies within the confidence region \mathcal{X}_B , if

$$p(\mathbf{x}_0|\mathbf{y}) > p_b. \quad (85)$$

The hypothesis H_0 is then accepted, otherwise it is rejected. See the last paragraph of Section 4.2 for an application of Eq. (85) and see Koch and Brockmann [2016] for the use of the hypothesis test.

5 Error Propagation by Monte Carlo methods

The law of error propagation has been derived by Eq. (68). It is a method of traditional statistics. Monte Carlo methods for solving it are therefore not restricted to Bayesian statistics. If we know the distribution of the vector \mathbf{x} and due to the central limit theorem, it is frequently the normal distribution, random variates x_j are drawn like in Eq. (77) from this distribution for \mathbf{x}

$$x_j = (x_{ij}) \quad \text{for } i \in \{1, \dots, n\}, j \in \{1, \dots, m\}. \quad (86)$$

The covariance matrix $D(\mathbf{x})$ of the nonlinearly transformed vector \mathbf{x} , mentioned in connection with Eq. (68), and its expectation $E(\mathbf{x})$ follow from the Monte Carlo estimate Eq. (78) and Eq. (79). This will save a considerable amount of derivatives to be computed so that errors of the linearization are avoided. Thus, the Monte Carlo methods become more efficient, cf. Siebert and Sommer [2004], Acko and Godina [2005]. For geodetic applications, this has been pointed out by Hennes [2007] and demonstrated by Koch [2008].

Let the random vector $\mathbf{f}(\mathbf{y})$ from Eq. (74) be represented by a sum of two or more random vectors with multivariate normal distributions and with different multivariate distributions. The sum is then not normally distributed anymore. The Monte Carlo estimate is therefore needed, to

compute the covariance matrix and the expectation of the sum of the vectors, i.e. the variations and the mean values of the sum. This was already pointed out for Eq. (78) and Eq. (79).

More examples are presented where Monte Carlo methods were applied: Alkhatib and Schuh [2007] discussed estimating covariance matrices for large-scale least squares problems when determining the geopotential from satellite data. Alkhatib et al. [2009] compared modeling variances of random and systematic errors by Monte Carlo and fuzzy techniques. Gundlich and Kusche [2008] solved quasi-linear models. Koch [2005] determined the maximum degree of harmonic coefficients in geopotential models. Koch [2015] checked whether a false null hypothesis is rejected with a power of 0.80. Koch [2017] computed the minimal detectable biases (MDB) or minimal detectable outliers by the Expectation Maximization (EM) algorithm. Koch [2018] presented Monte Carlo methods. Koch and Kargoll [2015] controlled results for answering the question whether rectangular or polar coordinates are better suited for laser scanning to detect outliers. Kusche [2003] estimated weights of observations in satellite geodesy by applying a stochastic trace estimation which is a Monte Carlo method. Lehmann [2013] found percentage points for maximal normalized and studentized residuals. Nowel [2016] applied hypothesis tests in deformation analysis based on robust M-estimation. Pany et al. [2011] determined the impact of tropospheric, clock and measurement errors on the repeatability of VLBI positions. Rubinstein [1981] published a textbook on Monte Carlo methods. Xu [2001] computed GPS decorrelation.

6 Conclusions

It is demonstrated that the Bayesian approach allows an intuitive way to derive statistics. Three rules, which are derived by logical and consistent reasoning, are sufficient for the derivation. They lead to distributions, point estimation, estimation of confidence regions and hypothesis tests. The point estimation computes the vector of unknown parameters from measurements, while a confidence region determines the area where the unknown parameters are situated with a given probability. Hypothesis are tested by deciding whether a value for the vector of unknown parameters lies inside or outside the confidence region. This is a method which originates from traditional statistics. Error propagation has also been developed for traditional statistics. Solving it by Monte Carlo methods turns out to be more efficient than the usual way

of computing error propagations. This shows that Monte Carlo methods are not restricted to Bayesian statistics, they can be used as well in traditional statistics. Furthermore, Monte Carlo methods cover a wide spectrum of applications as demonstrated by the examples given in Section 5.

Acknowledgement: The author is indebted to Jan Martin Brockmann for his valuable comments.

References

- Acko B. and Godina A., 2005, Verification of the conventional measuring uncertainty evaluation model with Monte Carlo simulation, *Int J Simul Model*, 4, 76–84.
- Alkhatib H. and Kutterer H., 2013, Estimation of measurement uncertainty of kinematic TLS observation process by means of Monte Carlo methods, *J Applied Geodesy*, 7, 125–133.
- Alkhatib H. and Schuh W.-D., 2007, Integration of the Monte Carlo covariance estimation strategy into tailored solution procedures for large-scale least squares problems, *J Geodesy*, 81, 53–66.
- Alkhatib H., Neumann I. and Kutterer H., 2009, Uncertainty modeling of random and systematic errors by means of Monte Carlo and fuzzy techniques, *J Applied Geodesy*, 3, 67–79.
- Blatter C., 1974, *Analysis I, II, III*, Springer, Berlin.
- Box G.E.P. and Tiao G.C., 1973, *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading.
- Cox R.T., 1946, Probability, frequency and reasonable expectation, *American J of Physics*, 14, 1–13.
- Gundlich B. and Kusche J., 2008, Monte Carlo integration for quasi-linear models. In Peiliang Xu, Jingnan Liu, and Athanasios Dermanis, editors, *VI Hotine-Marussi Symposium on Theoretical and Computational Geodesy*, p. 337–344, Springer, Berlin, Heidelberg.
- Hennes H., 2007, Konkurrierende Genauigkeitsmaße – Potential und Schwächen aus der Sicht des Anwenders, *Allgemeine Vermessungs-Nachrichten*, 114, 136–146.
- Jaynes E.T., 2003, *Probability theory. The logic of science*, Cambridge University Press, Cambridge.
- JCGM, 2011, Evaluation of measurement data—Supplement 2 to the “Guide to the expression of uncertainty in measurement”—Extension to any number of output quantities, JCGM 102:2011. Joint Committee for Guides in Metrology, www.bipm.org/en/publications/guides/.
- Koch K.R., 1999, *Parameter Estimation and Hypothesis Testing in Linear Models*, 2nd Ed. Springer, Berlin.
- Koch K.R., 2005, Determining the maximum degree of harmonic coefficients in geopotential models by Monte Carlo methods, *Studia Geophysica et Geodaetica*, 49, 259–275.
- Koch K.R., 2007, *Introduction to Bayesian Statistics*, 2nd Ed. Springer, Berlin.
- Koch K.R., 2008, Evaluation of uncertainties in measurements by Monte Carlo simulations with an application for laserscanning, *J Applied Geodesy*, 2, 67–77.
- Koch K.R., 2015, Minimal detectable outliers as measures of reliability, *J Geodesy*, 89, 483–490.
- Koch K.R., 2017, Expectation Maximization algorithm and its minimal detectable outliers, *Studia Geophysica et Geodaetica*, 61, 1–18.
- Koch K.R., 2018, Monte Carlo methods. *GEM—International Journal on Geomathematics*, 9(1), DOI 10.1007/s13137-017-0101-z.
- Koch K.R. and Brockmann J.M., 2016, Systematic effects in laser scanning and visualization by confidence regions, *J Applied Geodesy*, 10, 247–257.
- Koch K.R. and Kargoll B., 2015, Outlier detection by the EM algorithm for laser scanning in rectangular and polar coordinate systems, *J Applied Geodesy*, 9, 162–173.
- Kusche J. (2003), A Monte-Carlo technique for weight estimation in satellite geodesy, *J Geodesy*, 76, 641–652.
- Lehmann R., 2013, On the formulation of the alternative hypothesis for geodetic outlier detection, *J Geodesy*, 87, 373–386.
- Leonard T. and Hsu J.S.J., 1999, *Bayesian Methods*. Cambridge University Press, Cambridge.
- Loredo T. J., 1990, From Laplace to Supernova SN 1987a: Bayesian inference in astrophysics, In P. F. Fougère, editor, *Maximum Entropy and Bayesian Methods*, p. 81–142, Dordrecht, Kluwer Academic Publ.
- Nowel K., 2016, Application of Monte Carlo method to statistical testing in deformation analysis based on robust M-estimation, *Survey Review*, 48 (348), 212–223.
- Pany A., Böhm J., MacMillan D., Schuh H., Nilsson T. and Wresnik J., 2011, Monte Carlo simulations of the impact of troposphere, clock and measurement errors on the repeatability of VLBI positions, *J Geodesy*, 85, 39–50.
- Press S.J., 1989, *Bayesian Statistics: Principles, Models, and Applications*, Wiley, New York.
- Rubinstein R.Y., 1981, *Simulation and the Monte Carlo Method*, Wiley, New York.
- Siebert B.R.L. and Sommer K.-D., 2004, Weiterentwicklung des GUM und Monte-Carlo-Techniken, *tm—Technisches Messen*, 71, 67–80.
- Sivia D.S., 1996, *Data Analysis, a Bayesian Tutorial*, Clarendon Press, Oxford.
- Xu P., 2001, Random simulation and GPS decorrelation, *J Geodesy*, 75, 408–423.
- Zellner A., 1971, *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York.