2

Statistical definitions

Basic statistical definitions which will be used in the following chapters are explained. The following is only meant to be a quick reference on statistical notation and definitions and more elaborate textbooks can be used for a comprehensive introduction to the subject.

2.1 Probability density function

Given a continuous random variable Ψ, we can associate a *distribution function* $F(\psi)$. This is also named the cumulative density function or probability distribution function, and it describes the probability that a realization of Ψ takes a value less than or equal to ψ. We can relate it to a continuous probability density function $f(\psi)$, through

$$F(\psi) = \int_{-\infty}^{\psi} f(\psi')d\psi', \tag{2.1}$$

thus $f(\psi)$, when it exists, is just the derivative of the distribution function

$$f(\psi) = \frac{\partial F(\psi)}{\partial \psi}. \tag{2.2}$$

The probability density function (pdf) gives the probability that a random variable Ψ will take a particular value ψ. If a probability distribution has density $f(\psi)$, then the infinitesimal interval $(\psi, \psi + d\psi)$ has probability $f(\psi)d\psi$.

The pdf must satisfy the conditions

$$f(\psi) \geq 0 \quad \text{for all } \psi, \tag{2.3}$$

which states that the probability for Ψ to take a value ψ, must be positive or zero, and

G. Evensen, *Data Assimilation*, 2nd ed., DOI 10.1007/978-3-642-03711-5_2,

$$\int_{-\infty}^{\infty} f(\psi)d\psi = 1, \tag{2.4}$$

that is, the probability of finding Ψ in the space of real numbers $\Re^1$, is equal to one.

Further, given $f(\psi)$, the probability that ψ takes a value in the interval $[\psi_a, \psi_b]$ is

$$\Pr\big(\Psi \in [\psi_a, \psi_b]\big) = \int_{\psi_a}^{\psi_b} f(\psi)d\psi. \tag{2.5}$$

The most common and useful distribution is the one called the *normal or Gaussian distribution*. It is defined by its *mean* and *variance* and has a bell shaped or Gaussian form. It represents a family of distributions of the same general form, characterized by their mean μ, and the variance σ^2. The *standard normal distribution* is a normal distribution with a mean of zero and a variance of one. The normal distribution has the pdf

$$f(\psi) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\psi-\mu)^2}{2\sigma^2}\right). \tag{2.6}$$

A convenient aspect of a normal population distribution is that the following empirical "rule of thumb" can be applied to the data: $\mu \pm \sigma$ spans approximately 68% of the realizations, $\mu \pm 2\sigma$ spans approximately 95% of the realizations, and $\mu \pm 3\sigma$ spans about 99% of the realizations.

The *joint pdf* describes the probability of two events together. Given two random variables Ψ and Φ we can define the joint pdf $f(\psi, \phi)$.

The *conditional pdf* describes the probability of some event Ψ, assuming the event Φ. The conditional pdf is denoted $f(\psi|\phi)$ which is read as the pdf for Ψ given Φ. It is often called the *posterior pdf*.

The *marginal pdf* is the pdf of one event, ignoring any information about the other event. It is obtained by integrating the joint pdf over the ignored event; e.g. the marginal pdf for Ψ is $f(\psi) = \int_{-\infty}^{\infty} f(\psi, \phi)d\phi$.

We also have that

$$f(\psi|\phi) = \frac{f(\psi, \phi)}{f(\phi)}, \tag{2.7}$$

or equivalently

$$f(\psi, \phi) = f(\psi|\phi)f(\phi) = f(\phi|\psi)f(\psi). \tag{2.8}$$

The variables Ψ and Φ are said to be independent if $f(\psi, \phi) = f(\psi)f(\phi)$.

From 2.8 we can write

$$f(\psi|\phi) = \frac{f(\psi)f(\phi|\psi)}{f(\phi)}. \tag{2.9}$$

This is Bayes' theorem which is a general result in probability theory giving the conditional probability distribution of a random variable Ψ given Φ in terms of the conditional probability distribution of variable Φ given Ψ, often named

the *likelihood*, and the marginal probability distribution of Ψ alone. In the context of Bayesian probability theory, the marginal probability distribution of Ψ alone is usually called the *prior* probability distribution or simply the prior. The conditional distribution of Ψ given the "data" Φ is called the *posterior* probability distribution or just the posterior. This is a general result and will be used extensively in the following chapters.

In this book we will in several occasions refer to and use Bayesian statistics to derive and explain data assimilation methods and their properties. In particular we will use a probability density function $f(\boldsymbol{\psi})$, for the event $\boldsymbol{\psi} \in \Re^n$. This is again related to the distribution function $F(\boldsymbol{\psi})$, of the random variable $\boldsymbol{\Psi} \in \Re^n$, through the equation

$$F(\psi_1, \ldots, \psi_n) = \int_{-\infty}^{\psi_1} \cdots \int_{-\infty}^{\psi_n} f(\psi'_1, \ldots, \psi'_n) d\psi'_1 \ldots d\psi'_n, \tag{2.10}$$

and the pdf is again defined as the derivative of the distribution function.

The pdf is a positive function of dimension n and it has the property that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\psi_1, \ldots, \psi_n) d\psi_1 \ldots d\psi_n = 1. \tag{2.11}$$

Thus, the probability that $\boldsymbol{\psi}$ is located somewhere in $\Re^n$ is one. For each value of $\boldsymbol{\psi}$, $f(\boldsymbol{\psi})$ gives the probability for this particular state. The pdf $f(\boldsymbol{\psi})$ is also named the joint pdf for $(\psi_1, \ldots, \psi_n)$.

This joint pdf can be factorized into

$$f(\psi_1, \ldots, \psi_n) = f(\psi_1) f(\psi_2|\psi_1) f(\psi_3|\psi_1, \psi_2) \cdots f(\psi_n|\psi_1, \ldots, \psi_{n-1}). \tag{2.12}$$

Here $f(\psi_2|\psi_1)$ is the likelihood of ψ_2 given ψ_1, and if $n = 2$ we get just $f(\psi_1, \psi_2) = f(\psi_1) f(\psi_2|\psi_1)$, which is interpreted as the probability of ψ_1 times the likelihood of ψ_2 given ψ_1.

If the events, $(\psi_1, \ldots, \psi_n)$ are independent we can write

$$f(\psi_1, \ldots, \psi_n) = f(\psi_1) f(\psi_2) \cdots f(\psi_n). \tag{2.13}$$

We will make frequent use of the pdf of a model state $\boldsymbol{\psi}$, and the likelihood function for a vector of measurements $\boldsymbol{d}$, of the state which is written as $f(\boldsymbol{d}|\boldsymbol{\psi})$. The joint pdf of the state and the measurements can be written

$$f(\boldsymbol{\psi}, \boldsymbol{d}) = f(\boldsymbol{\psi}) f(\boldsymbol{d}|\boldsymbol{\psi}) = f(\boldsymbol{d}) f(\boldsymbol{\psi}|\boldsymbol{d}), \tag{2.14}$$

and we must have

$$f(\boldsymbol{\psi}|\boldsymbol{d}) = \frac{f(\boldsymbol{\psi}) f(\boldsymbol{d}|\boldsymbol{\psi})}{f(\boldsymbol{d})}, \tag{2.15}$$

where the denominator is just the integral of the numerator, which normalizes the numerator such that the expression integrates to one. This is Bayes' theorem, and in this context it states that the pdf of the model state given a set of measurements is proportional to the pdf of the model state times the likelihood function for the measurements.

2.2 Statistical moments

The probability density function $f(\psi)$, contains a huge amount of information, especially for high dimensional systems, and actually much more information than is normally needed. Instead of working with the full density it is often convenient to define statistical moments of the density. These are defined from the general expression of the expected value of a function $h(\Psi)$,

$$E[h(\Psi)] = \int_{-\infty}^{\infty} h(\psi) f(\psi) d\psi. \tag{2.16}$$

2.2.1 Expected value

The expected value of a random variable Ψ with distribution $f(\psi)$, is defined as

$$\mu = E[\Psi] = \int_{-\infty}^{\infty} \psi f(\psi) d\psi. \tag{2.17}$$

The expected value (or expectation) of a random variable represents the average one "expects" if an infinite number of samples are drawn from the distribution. Note that the value itself may not be expected in the general sense, it may be unlikely or even impossible, dependent on the shape of $f(\psi)$.

2.2.2 Variance

If Ψ is a random variable, the variance is given by

$$\begin{aligned} \sigma^2 = E\Big[\big(\Psi - E[\Psi]\big)^2\Big] &= \int_{-\infty}^{\infty} \big(\psi - E[\Psi]\big)^2 f(\psi) d\psi \\ &= E\big[\Psi^2\big] - E\big[\Psi\big]^2. \end{aligned} \tag{2.18}$$

That is, it is the expected value of the square of the deviation of Ψ from its own mean. In other words, it is the average of the square of the distance of each data point from the mean. It is thus the mean squared deviation. The second line in 2.18 is often used for the practical computation of the variance. It is just the second moment minus the square of the first moment.

An inconvenience is that the variance has a unit which is the square of the data unit. For this reason it is common to use the square root of the variance which is named the *standard deviation*, denoted σ. It can also easily be shown that the variance does not depend on the mean, thus the variance of $\Psi + b$ is the same as the variance of Ψ. On the other hand the variance of $a\Psi$ is $a^2\sigma^2$.

2.2.3 Covariance

Given two random variables Ψ and Φ and their respective probability density functions $f(\psi)$ and $f(\phi)$, from which we can define the joint probability $f(\psi, \phi) = f(\psi|\phi)f(\phi) = f(\phi|\psi)f(\psi)$, their covariance is defined as

$$\begin{aligned} E\Big[\big(\Psi - E[\Psi]\big)\big(\Phi - E[\Phi]\big)\Big] \\ &= \iint_{-\infty}^{\infty} \big(\psi - E[\Psi]\big)\big(\phi - E[\Phi]\big) f(\psi, \phi) d\psi d\phi \\ &= \iint_{-\infty}^{\infty} \psi\phi f(\psi, \phi) d\psi d\phi - E[\Psi]E[\Phi]. \end{aligned} \tag{2.19}$$

Note that the same conditions (2.3) and (2.4) also apply for $f(\psi, \phi)$. In the case when the random variables Ψ and Φ are independent, $f(\psi, \phi) = f(\psi)f(\phi)$ and the covariance becomes zero.

2.3 Working with samples from a distribution

Clearly when the dimension of a probability function increases to more than about 3–4 it becomes very impractical, if not impossible, to evaluate the integrals by numerical integration on a regular grid. Suppose the dimension is 10 and we need 10 grid points in each direction to have a proper representation of the density. A grid with 10^{10} nodes would then have to be stored which would require 40 Giga bytes of storage and 10^{10} additions would be needed to calculate the integral.

Fortunately there is an alternative to the direct numerical integration which often works very well even for high dimensional systems. The approach is called the Markov Chain Monte Carlo (MCMC) methods, (see e.g. *Robert and Casella*, 2004), and assumes that we have available a large number N, of realizations from the distribution $f(\psi)$.

2.3.1 Sample mean

Having a sample of independent realizations from the distribution $f(\psi)$, i.e. ψ_i, for $i = 1, N$, then the sample mean $\overline{\psi}$, is given by

$$\mu = E[\psi] \simeq \overline{\psi} = \frac{1}{N} \sum_{i=1}^{N} \psi_i. \tag{2.20}$$

The "expected value" terminology is meant to connote that $E[\Psi]$ is, in some sense, the "best guess" as to the possible outcome of Ψ, or said in another way; the expected value is the value we expect to obtain if infinitely many data are present, and the sample mean of these is computed. This is a reason why $E[\Psi]$ is often called the mean of Ψ.

2.3.2 Sample variance

The variance can be calculated from the formula

$$\begin{aligned}\sigma^2 &= E\Big[\big(\Psi - E[\Psi]\big)^2\Big] \\ &\simeq \overline{\big(\psi - \overline{\psi}\big)^2} = \frac{1}{N-1}\sum_{i=1}^{N}\big(\psi_i - \overline{\psi}\big)^2,\end{aligned} \tag{2.21}$$

where the denominator $N-1$ is used instead of N to ensure that the formula (2.21) becomes an unbiased estimator for the variance.

2.3.3 Sample covariance

The covariance can be calculated from the formula

$$\begin{aligned}\mathrm{Cov}(\psi,\phi) &= E\Big[\big(\Psi - E[\Psi]\big)\big(\Phi - E[\Phi]\big)\Big] \\ &\simeq \overline{\big(\psi - \overline{\psi}\big)\big(\phi - \overline{\phi}\big)} = \frac{1}{N-1}\sum_{i=1}^{N}\big(\psi_i - \overline{\psi}\big)\big(\phi_i - \overline{\phi}\big).\end{aligned} \tag{2.22}$$

2.4 Statistics of random fields

Of special interest for us will be the statistics of so-called random fields $\Psi(\boldsymbol{x})$ where Ψ is now a function of $\boldsymbol{x} = (x, y, z, \ldots)$.

2.4.1 Sample mean

Having an ensemble of independent samples from the distribution $f(\psi(\boldsymbol{x}))$, i.e. $\psi_i(\boldsymbol{x})$, for $i = 1, N$, then the sample mean is given by

$$\mu(\boldsymbol{x}) \simeq \overline{\psi(\boldsymbol{x})} = \frac{1}{N}\sum_{i=1}^{N}\psi_i(\boldsymbol{x}). \tag{2.23}$$

2.4.2 Sample variance

The sample variance of an ensemble of independent samples from the distribution $f(\psi(\boldsymbol{x}))$, is given as

$$\sigma^2(\boldsymbol{x}) \simeq \overline{\big(\psi(\boldsymbol{x}) - \overline{\psi(\boldsymbol{x})}\big)^2} = \frac{1}{N-1}\sum_{i=1}^{N}\big(\psi_i(\boldsymbol{x}) - \overline{\psi(\boldsymbol{x})}\big)^2. \tag{2.24}$$

2.4.3 Sample covariance

The covariance between two different locations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ for the random fields are given by

$$\begin{aligned} C_{\psi\psi}(\boldsymbol{x}_1, \boldsymbol{x}_2) &\simeq \overline{\big(\psi(\boldsymbol{x}_1) - \overline{\psi(\boldsymbol{x}_1)}\,\big)\big(\psi(\boldsymbol{x}_2) - \overline{\psi(\boldsymbol{x}_2)}\,\big)} \\ &= \frac{1}{N-1}\sum_{j=1}^{N}\big(\psi_j(\boldsymbol{x}_1) - \overline{\psi(\boldsymbol{x}_1)}\,\big)\big(\psi_j(\boldsymbol{x}_2) - \overline{\psi(\boldsymbol{x}_2)}\,\big). \end{aligned} \tag{2.25}$$

Note that if $\boldsymbol{x}_1 = \boldsymbol{x}_2$, then (2.25) reduces to the definition of variance.

The covariance of Ψ between the two locations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ defines how values of Ψ, at different locations, are "varying together" or "covarying" . For example, if the random fields Ψ are smooth we will expect that neighboring points are correlated or covarying. The covariance can therefore be a measure of smoothness.

2.4.4 Correlation

The correlation between the random variables $\Psi(\boldsymbol{x}_1)$ and $\Psi(\boldsymbol{x}_2)$ is defined by

$$\mathrm{Cor}\big(\psi(\boldsymbol{x}_1), \psi(\boldsymbol{x}_2)\big) = \frac{\mathrm{C}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\sigma(\boldsymbol{x}_1)\sigma(\boldsymbol{x}_2)}. \tag{2.26}$$

Thus, the correlation is just a normalized covariance.

2.5 Bias

One meaning is involved in what is called a biased sample; if some elements are more likely to be chosen in the sample than others, and those have a higher/lower value of the quantity being estimated, the outcome will be higher/lower than the true value.

Another kind of bias in statistics does not involve biased samples, but rather the use of a statistics whose average value differs from the value of the quantity being estimated. Suppose we are trying to estimate the true value ψ^t of a parameter ψ using an estimator $\hat{\psi}$ (that is, some function of the observed data). Then the bias of $\hat{\psi}$ is defined to be

$$E\big[\hat{\psi}\big] - \psi^t. \tag{2.27}$$

In words, this would be "the expected value of the estimator $\hat{\psi}$ minus the true value ψ^t". This may be rewritten as

$$E\big[\hat{\psi} - \psi^t\big], \tag{2.28}$$

which would read "the expected value of the difference between the estimator and the true value".

An example of a biased estimator of variance is

$$\sigma^2_{\text{biased}} = \frac{1}{N}\sum_{i=1}^{N}\left(\psi_i - \overline{\psi}\right)^2, \tag{2.29}$$

which differs from the formula (2.21) by the division by N rather than $N-1$. The proof that this is a biased estimator of the variance is left as an exercise.

2.6 Central limit theorem

The central limit theorem can be used to say something about the convergence of the moments of a sample with increasing sample size.

Assume that we draw a number of samples of the random variable Ψ, each with sample size N. We then have the following:

- The sample mean $\mu(\psi)$ from (2.23), computed from the different samples is normally distributed, independent of the distribution for Ψ.
- The standard deviation of $\mu(\psi)$ as computed from the different samples tends towards $\sigma(\Psi)/\sqrt{N}$.

Thus, if we compute the sample mean from a given sample, we can expect that the error in the computed sample mean is normally distributed and given by $\sigma(\Psi)/\sqrt{N}$. Importantly, the error decreases proportional to $1/\sqrt{N}$.

The amazing and counter-intuitive property of the central limit theorem is that no matter what the shape of the original distribution, the sampling distribution of the mean approaches a normal distribution. Furthermore, for most distributions, a normal distribution is approached very quickly as N increases.

http://www.springer.com/978-3-642-03710-8

Data Assimilation
The Ensemble Kalman Filter
Evensen, G.
2009, XXIII, 307 p., Hardcover
ISBN: 978-3-642-03710-8