POLITECNICO DI MILANO





Recap on probability and statistics

Marco Lovera

Dipartimento di Scienze e Tecnologie Aerospaziali, Politecnico di Milano

Regardless of the specific approach to estimation, some background on random variables and probability is needed to formulate and solve problems.

Random experiment

We start by introducing the concept of a random experiment, which has the following three components

- Sample space: Ω
- Events of interest: **C**
- Probability function: P.

Sample space

The sample space Ω is defined as the set of outcomes of an experiment.

Example: tossing a coin twice (H=Heads, T=Tails).

Ω= {HH; HT; TT; TH}

An *event* is a subset of Ω.

Examples:

- event "at least one head" is {HH; HT; TH}
- event "no more than one head" is {HT; TH; TT}.

Events of interest

We need to enumerate the set of events of interest that can occur when carrying out an experiment.

In probability theory, this set is defined based on the following properties:

- the "empty set" belongs to **C**
- If event A $\in$ **C**, then its complement $A_c = (\Omega - A) \in$ **C**
- If for $N < \infty$ events $A_1, A_2, \ldots, A_N \in$ **C**, then

$$\bigcup_i A_i \in \mathbf{C}$$

Finally, a probability function P assigns a number (probability) to each event in **C**.

P is a function mapping **C** to the [0, 1] interval, satisfying:

- $P(\Omega) = 1$

- If for $N < \infty$ events $A_1, A_2, \ldots, A_N \in$ **C**, and

$$A_i \bigcap A_j = 0, \quad \forall i, j$$

then

$$P(\bigcup_i A_i) = \sum_i P(A_i)$$

Random variable

A variable *v* is called a *random variable* if its value depends on the outcome of a random experiment.

Formally *v* is defined as the output of a function $\phi(\cdot)$ mapping the sample space Ω into the range space V of *v*:

$$\phi(\cdot) : \Omega \to V$$

For a subset D of V, how can we compute $P(v \in D)$?

The image of D through $\phi^{-1}(\cdot)$ is needed, so that we can define

$$P(v \in D) = P(\phi^{-1}(D)).$$

This calls for some attention, as $P(\cdot)$ is defined only on the elements of **C**, so the above makes sense provided that

$$\phi^{-1}(D) \in \mathbf{C}.$$

Example

Real random variables

A random variable *v* is called *real* if

$$V = \bar{\mathbb{R}} = \{-\infty, \mathbb{R}, +\infty\}$$

Therefore in view of the previous definitions

$$P(v \in [a, b]) = P(\phi^{-1}([a, b]))$$

and we must ensure that all the intervals [*a*,*b*] belong to **C**.

For this to hold, we only need to define

$$P(v \in [-\infty, q]) = P(\phi^{-1}([-\infty, q])), \quad \forall q \in \mathbb{R}$$

as the probability for an arbitrary [a,b] interval follows by intersection.

So, we need to ensure that

$$\phi^{-1}([-\infty, q]) \in \mathbf{C}, \quad \forall q \in \mathbb{R}$$

As a conclusion, for a given random experiment, we say that *v* is a well defined real random variable if

$$v = \phi(s), \qquad \phi(\cdot) : \Omega \to \bar{\mathbb{R}}$$

$$\phi^{-1}([-\infty, q]) \in \mathbf{C}, \qquad \forall q \in \bar{\mathbb{R}}$$

$$P(\phi^{-1}(-\infty)) = P(\phi^{-1}(+\infty)) = 0$$

By definition, the probability distribution of a random variable v is a function

$$F(\cdot) : \bar{\mathbb{R}} \to [0, 1]$$

given by

$$F(q) = P(v \leq q) = P(\phi^{-1}([-\infty, q]))$$

Main properties of probability distribution functions

- $F(-\infty) = 0$
- $F(+\infty) = 1$
- $F(\cdot)$ monotonically increasing.
- $F(\cdot)$ right-continuous.
- $\lim_{q \to \infty} F(q) = 1$
- $\mathrm{F}(\cdot)$ piece-wise continuous.

Main use of probability distribution functions: probabilities can be easily expressed in terms of their values, *i.e.*,

$$P(v \in (a,b]) = F(b) - F(a)$$

and

$$P(v \in [a,b]) = F(b) - F(a^-)$$

Continuous vs discrete random variables

A real random variable is called

- continuous if $F(q)$ is a continuous function
- discrete if $F(q)$ is a step-wise function.

In view of the above properties we have that $F(\cdot)$ is differentiable almost everywhere (a.e.), *i.e.*, for all q except for discontinuities.

Therefore $dF(q)/dq$ is well defined a.e. and we can let

$$f(q) = \frac{dF(q)}{dq}$$

almost everywhere.

In the sense of generalised derivatives, the above holds also for discontinuities, leading to impulses in the derivative.

Probability density functions

Recalling that

$$P(v \in [a,b]) = F(b) - F(a^-)$$

and the definition

$$f(q) = \frac{dF(q)}{dq}$$

we have in turn that

$$P(v \in [a,b]) = \int_a^b f(q)dq$$

Expected value

The expected value of a random variable is defined as

$$E[v] = \int_{-\infty}^{+\infty} q f(q) dq$$

(and does not necessarily exist for all *f*).

If it exists it denotes the “center of mass” of the density function.

If *f*(*q*) is symmetric around $\bar{q}$, then $\bar{q} = E[v]$.

Variance and standard deviation

The variance of a random variable is defined as

$$\text{Var}[v] = \sigma^2(v) = \int_{-\infty}^{+\infty} (q - E[v])^2 f(q) dq$$

(and does not necessarily exist for all *f*).

As $f(q) \geq 0$, then also $\text{Var}[v] \geq 0$.

The standard deviation (root mean square) of a random variable is given by

$$\sigma[v] = \sqrt{\text{Var}[v]}.$$

Interpretation of σ

The Chebyshev inequality states that

$$P(|v - E[v]| > \epsilon) \leq \frac{\sigma^2[v]}{\epsilon^2}, \forall \epsilon > 0.$$

Therefore, letting $\epsilon = 2\sigma[v]$ we get

$$P(|v - E[v]| > 2\sigma[v]) \leq \frac{\sigma^2[v]}{4\sigma^2[v]} = 0.25$$

so *regardless of the distribution* the interval centered in E[*v*] with half-width $2\sigma[v]$ covers at least 0.75 probability.

The order *k* moment of a random variable is defined as

$$m_k[v] = \int_{-\infty}^{+\infty} q^k f(q) dq$$

and clearly

$$m_0[v] = \int_{-\infty}^{+\infty} f(q) dq = 1$$

$$m_1[v] = \int_{-\infty}^{+\infty} q f(q) dq = E[q].$$

The second order moment

$$m_2[v] = \int_{-\infty}^{+\infty} q^2 f(q)dq$$

is related to the variance and the expected value as follows:

$$\begin{aligned}\mathsf{Var}[v] &= \int_{-\infty}^{+\infty} (q - E[v])^2 f(q)dq = \\ &= \int_{-\infty}^{+\infty} (q^2 + E[v]^2 - 2qE[v]) f(q)dq = \\ &= m_2[v] + E[v]^2 - 2E[v]E[v] = m_2[v] - E[v]^2.\end{aligned}$$

Consider a random variable *v* and let

$$w = g(v)$$

where

$$g(\cdot) : \bar{\mathbb{R}} \to \bar{\mathbb{R}}$$

It can be shown that if *v* is a well defined random variable then so is *w.*

In terms of expected value we have

$$E[w] = \int_{-\infty}^{+\infty} q f_w(q) dq = \int_{-\infty}^{+\infty} g(q) f_v(q) dq$$

Letting

$$w = g(v) = (v - E[v])^2$$

we have that

$$E[w] = E[(v - E[v])^2] = \int_{-\infty}^{\infty} (q - E[v])^2 f_v(q)dq = \text{Var}[v].$$

and similarly

$$w = v^k \qquad \Rightarrow \qquad E[w] = E[v^k] = m_k[v].$$

In the same way we have that (as the expectation operator is *linear*):

$$w = \alpha v \qquad \Rightarrow \qquad E[w] = E[\alpha v] = \alpha E[v]$$

A Gaussian (normal) random variable has a density function of the form

$$f(q) = \alpha e^{-\beta q^2}, \qquad \alpha, \beta > 0$$

and α, β such that the density function has unit area.

Gaussian densities can be more effectively expressed in terms of expected value and variance.

Scalar Gaussian random variables

Indeed letting

$$\mu = E[v], \qquad \sigma^2 = \mathrm{Var}[v]$$

it can be shown that

$$f(q) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(q-\mu)^2}{2\sigma^2}}.$$

Shorthand notation:

$$v \sim G(\mu, \sigma^2) \qquad v \sim N(\mu, \sigma^2)$$

Linear propagation and standard Gaussian

Linear propagation: given $v \sim G(\mu, \sigma^2)$ and *w=a+bv* then

$$w \sim G(a + b\mu, b^2\sigma^2)$$

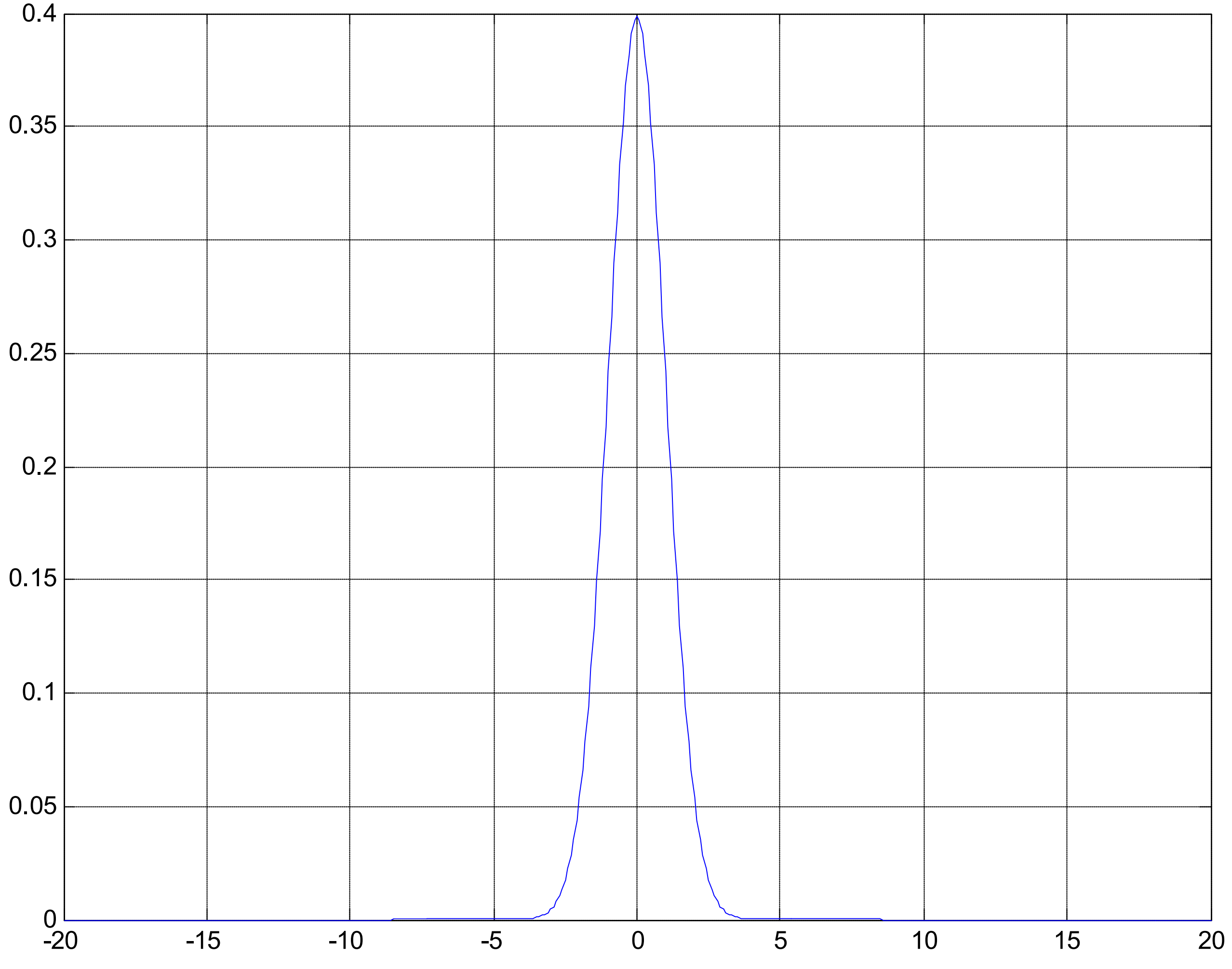
Based on this, for a generic $v \sim G(\mu, \sigma^2)$ we can define

$$w = \frac{v - \mu}{\sigma}$$

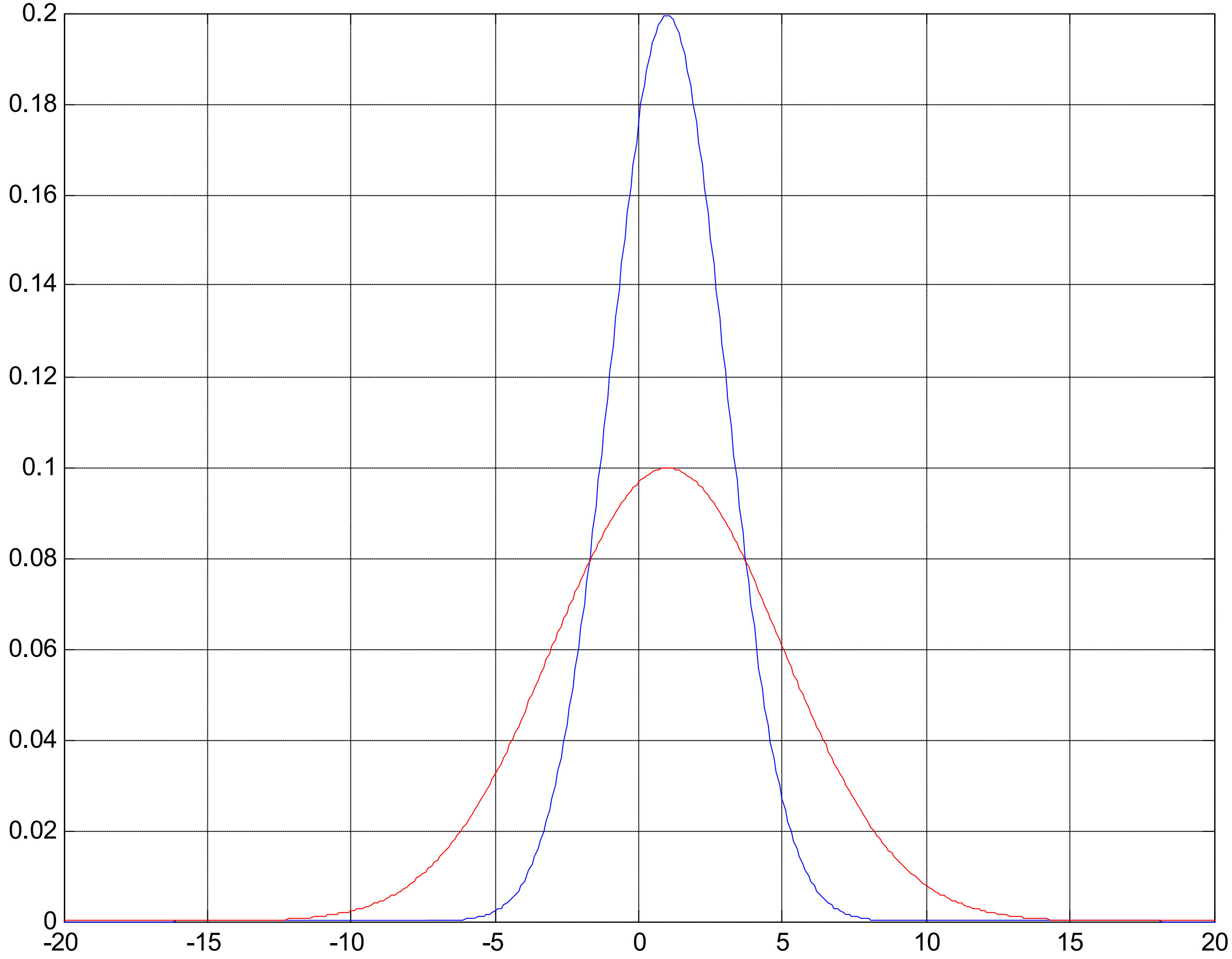
for which clearly $w \sim G(0, 1)$

The corresponding density function is $f(q) = \frac{1}{\sqrt{2\pi}} e^{\frac{-q^2}{2}}$ which is known as the standard Gaussian.

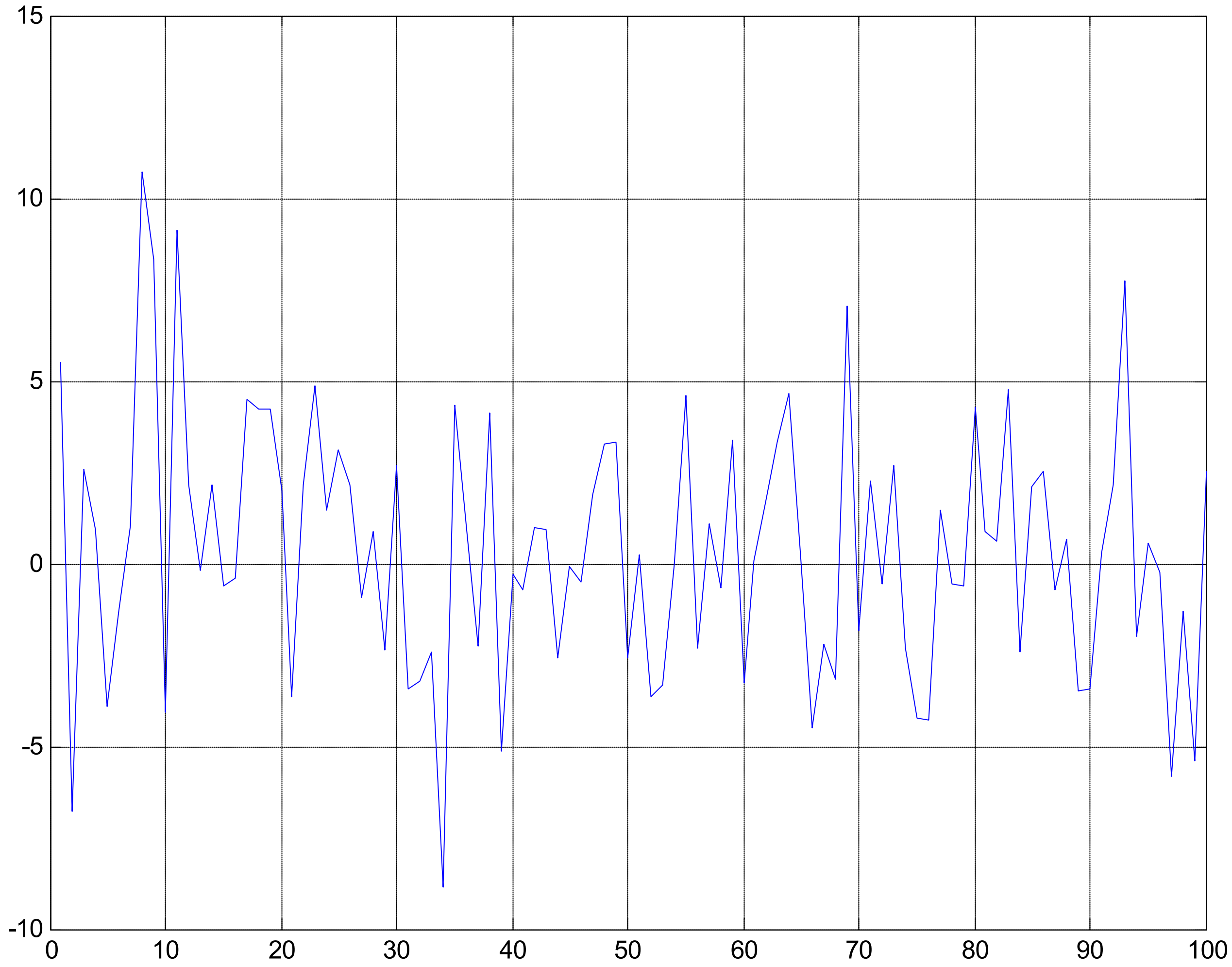
Example: the standard Gaussian



Example: μ=1, σ=2 and σ=4



Example: random data



Problem: density estimation

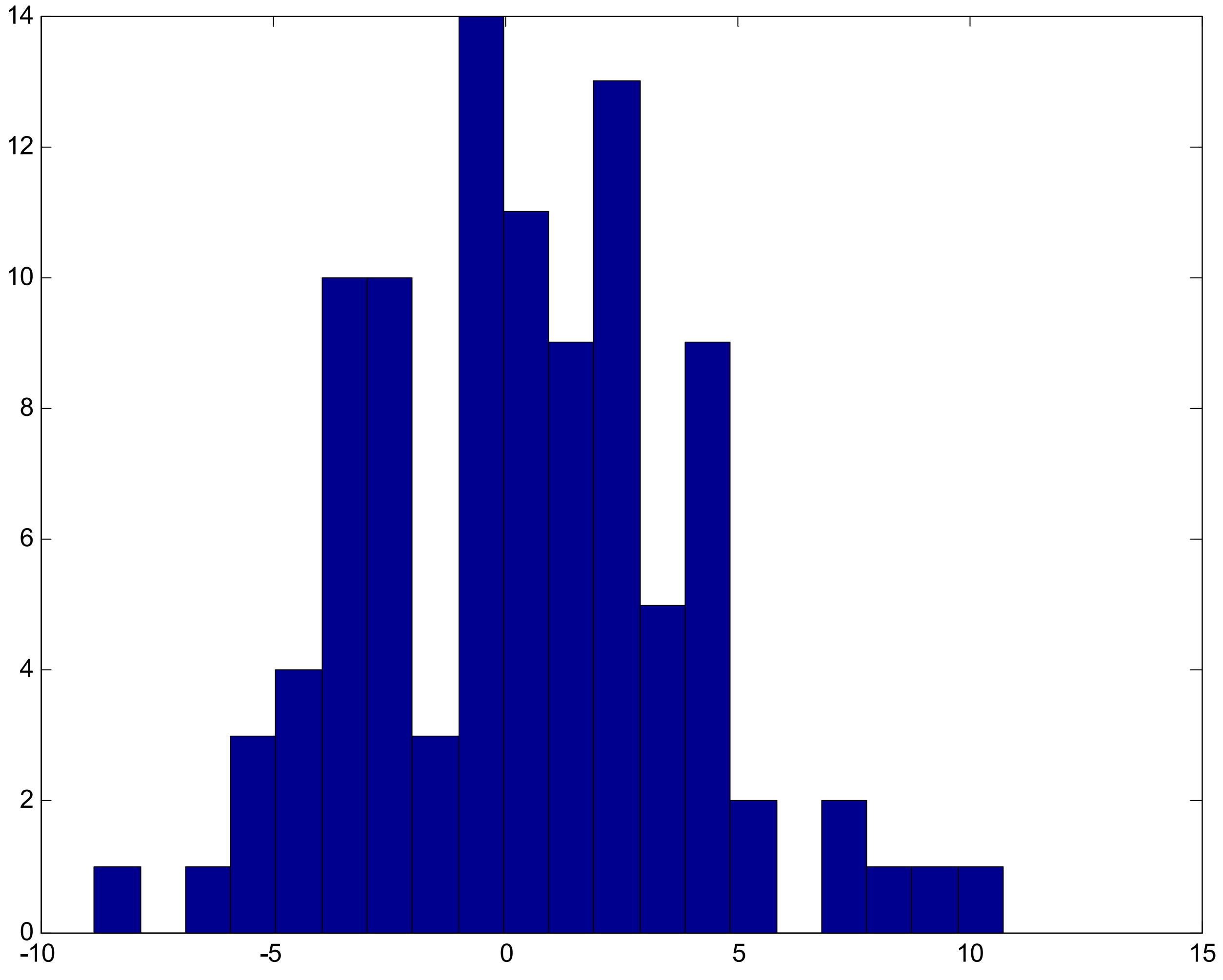
Assumption: data are extractions from a random variable with Gaussian density with unknown μ and σ.

Problem: estimating the density from data.

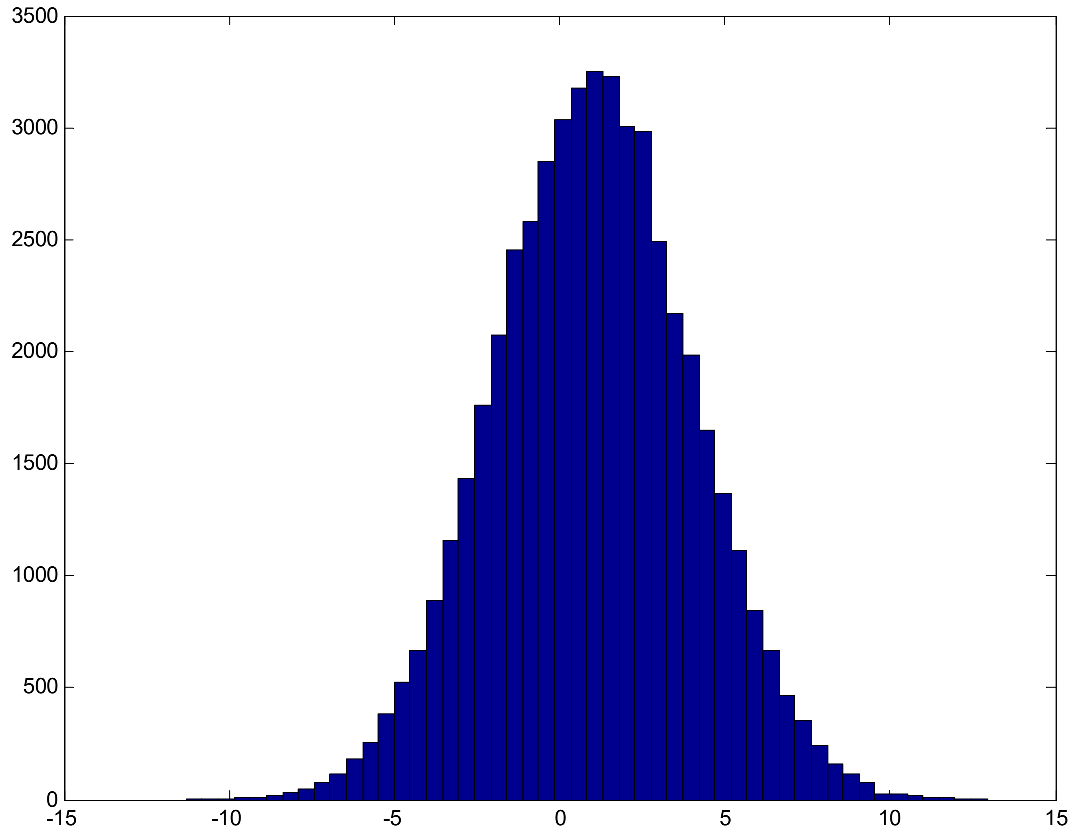
Possible approaches:

- Nonparametric: try to reconstruct the value of $f(q)\ \forall\ q$
- Parametric: try to estimate μ and σ and then "plug" the estimates in place of the true values.

Data histogram



If the dataset is very long then accurate nonparametric estimates can be built



Suitable *estimators* for μ and σ must be devised.

Let's pick the intuitive ones:

$$\widehat{\mu} = \frac{1}{N}\sum_i v_i$$

$$\widehat{\sigma}^2 = \frac{1}{N}\sum_i (v_i - \widehat{\mu})^2$$

and see what happens.

Parametric approach

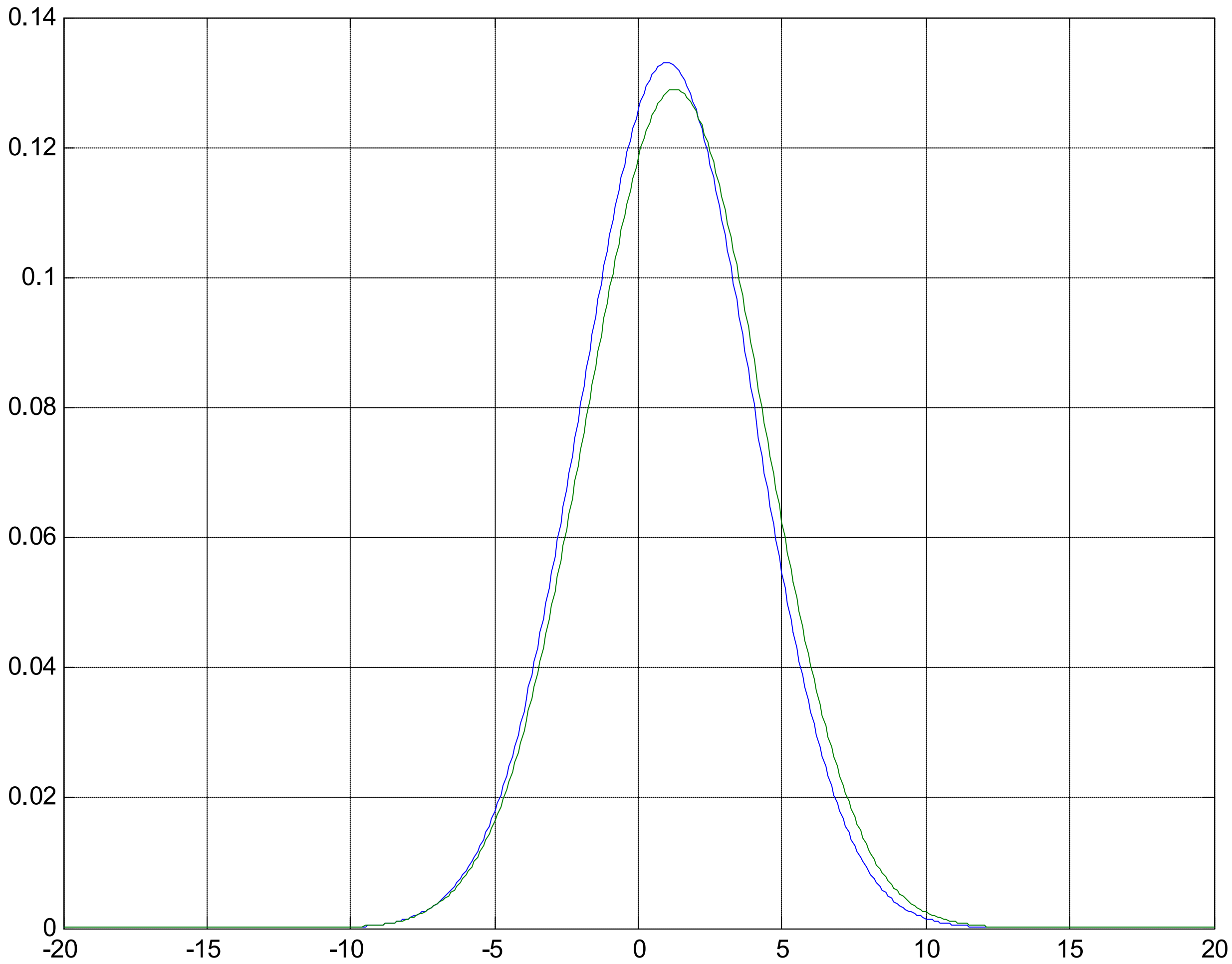
For the original dataset of 100 samples:

$$\hat{\mu} = 1.2781 \qquad \hat{\sigma}^2 = 9.578$$

True values:

$$\mu = 1 \qquad \sigma^2 = 9$$

Estimated density



Open questions

We can think about this result in many different ways (keeping in mind that in real problems the true values are unknown!):

- What happens if the esperiment is repeated?
- How accurate are these estimates?
- What happens if the length of the dataset increases?

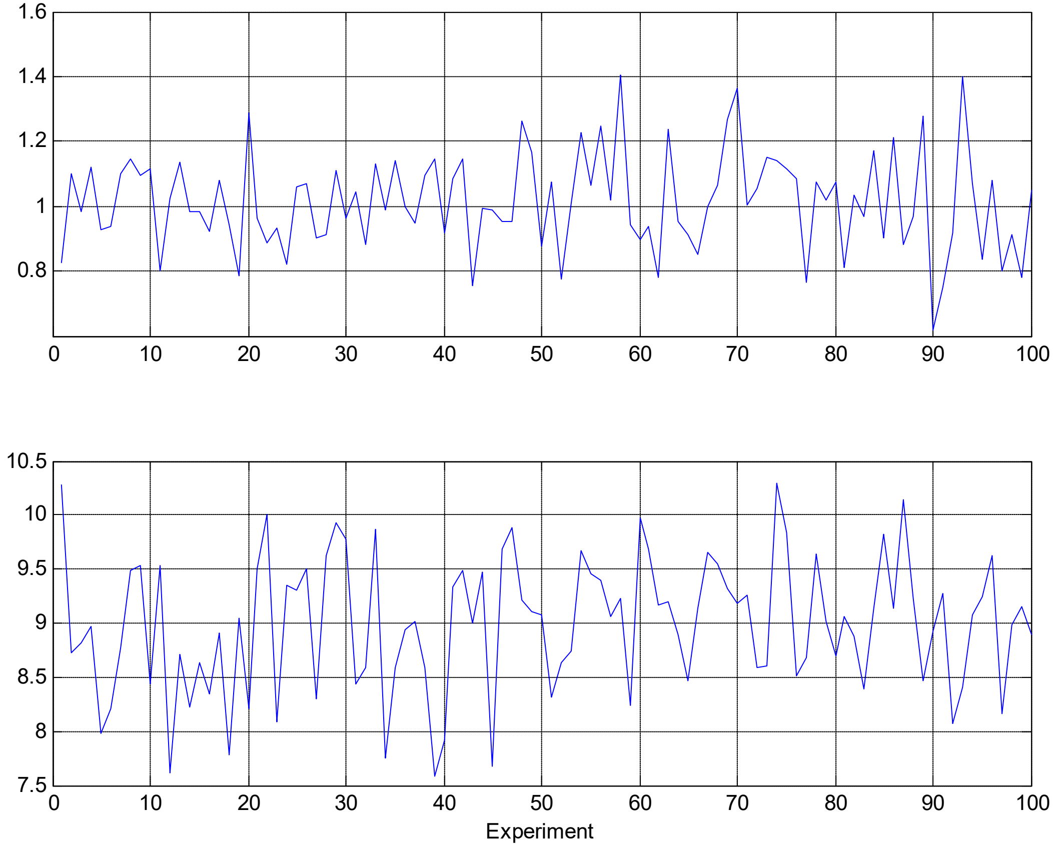
Repeated experiment

$$\hat{\mu} = 1.2461 \quad \hat{\sigma}^2 = 8.4484$$

$$\hat{\mu} = 0.7302 \quad \hat{\sigma}^2 = 9.4722$$

$$\hat{\mu} = 1.1630 \quad \hat{\sigma}^2 = 9.8497$$

We now repeat the experiment *many* times (M=100) and look at the outcomes in terms of estimates of μ and σ^2



Comments on the results:

- As expected the estimates of μ and σ^2 are also random variables
- We can then study the repeated estimates as *data*, looking at their properties.
- Mean of the estimates:

$$\frac{1}{M}\sum_i \widehat{\mu}_i = 1.0136 \quad \frac{1}{M}\sum_i \widehat{\sigma}_i^2 = 8.9788$$

Comments on the results:

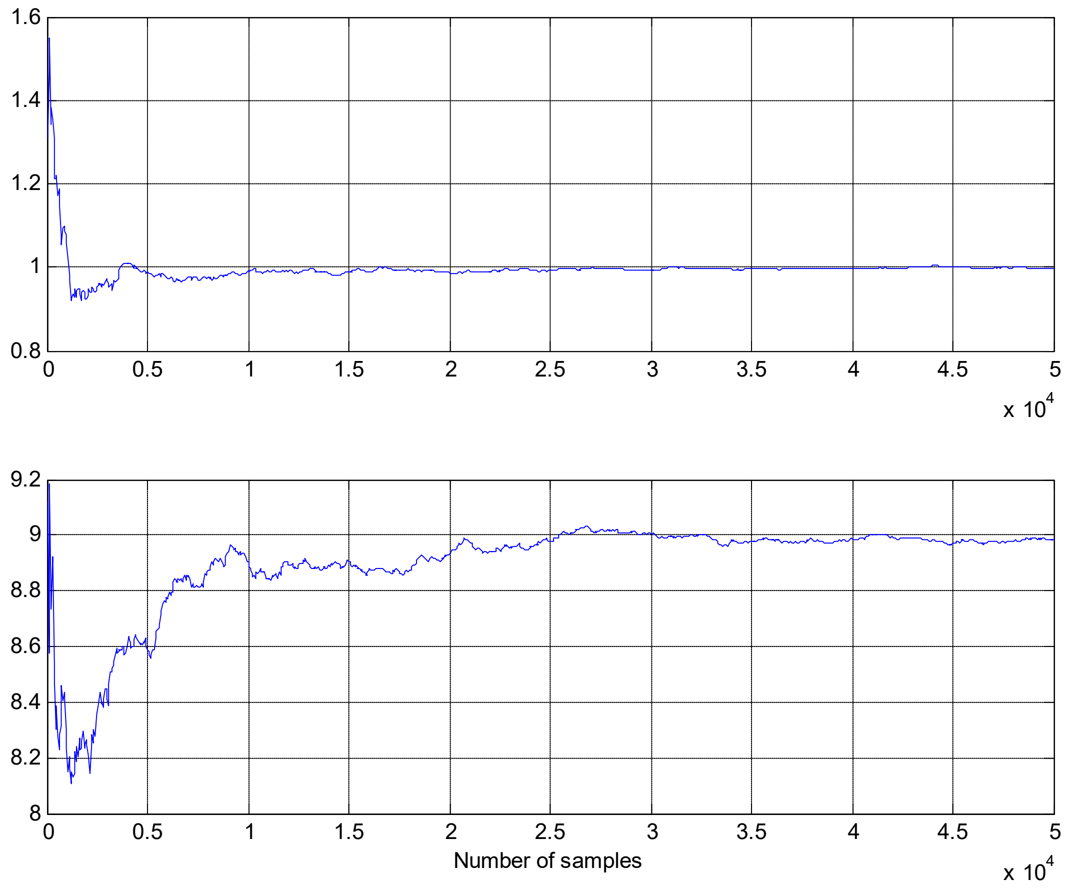
- Standard deviation of the estimates:

  0.1509 0.6232

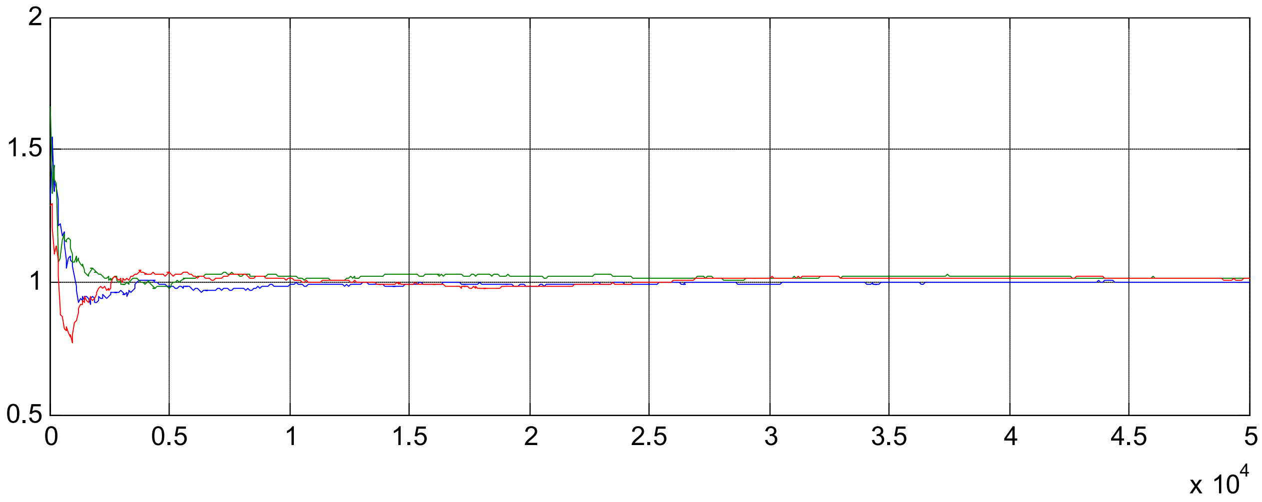
- So in conclusion
  - *On average* the estimators seem to provide correct results
  - However the estimates are random, so the standard deviation provides information about the probability of errors (remember Chebyshev inequality).

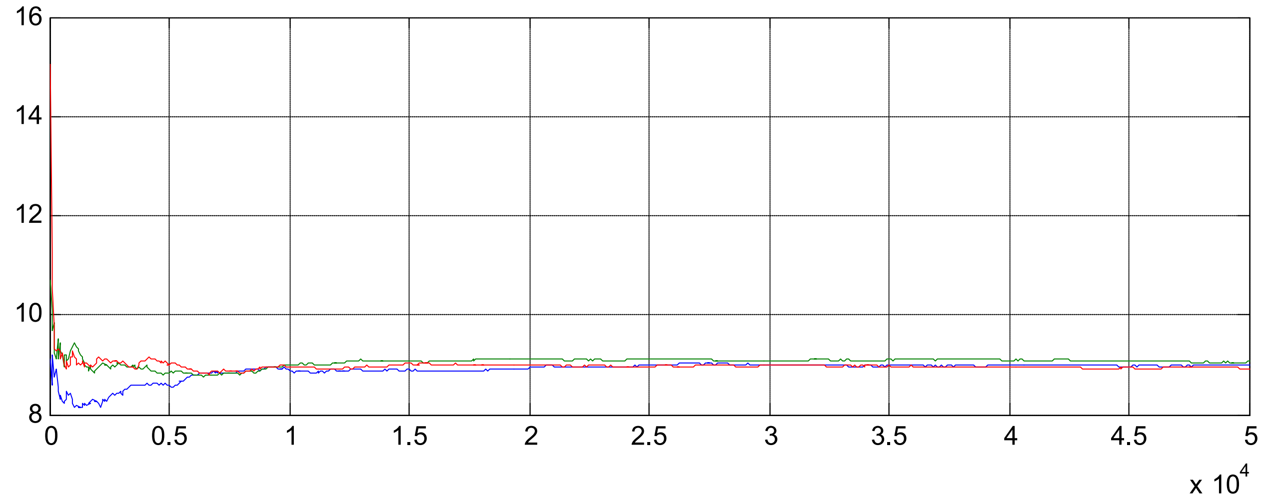
Increasing data length

What happens to the estimates if the length of the data set increases?



Increasing data length: different experiments





A vector

$$v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

is a well-defined random variable with respect to a random experiment (Ω, **C**, P) subject to suitable extensions of the conditions defined in the scalar case.

Let first

$$q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix}$$

Then *v* is a well defined random variable if it depends on the outcomes of the experiment via a function

$$v = \phi(s), \qquad \phi(\cdot) : \Omega \to \bar{\mathbb{R}}^n$$

such that

$$\phi^{-1}(v_1 \leq q_1, v_2 \leq q_2, \ldots, v_n \leq q_n) \in \mathbf{C}, \qquad \forall q \in \bar{\mathbb{R}}^n$$

and if

$$P(v_i = \pm\infty) = 0 \qquad i = 1, \ldots, n.$$

The (joint) probability distribution for the vector random variable *v* is defined as

$$F(q_1, q_2, \ldots, q_n) = P(v_1 \leq q_1, v_2 \leq q_2, \ldots, v_n \leq q_n)$$

If one is interested in the (marginal) distribution of a single component q_i, then it can be obtained as

$$F_i(q_i) = F(\infty, \ldots, \infty, q_i, \infty, \ldots, \infty)$$

Note that in general the joint distribution cannot be reconstructed from the sole knowledge of the marginals.

By generalising the scalar definition we have that

$$f(q_1, q_2, \ldots, q_n) = \frac{\partial^n F(q_1, q_2, \ldots, q_n)}{\partial q_1 \partial q_2 \ldots \partial q_n}$$

and the individual marginal densities can be obtained as

$$f_i(q_i) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} f(q_1, q_2, \ldots, q_n) dq_1 \ldots dq_n$$

where integration is carried out over all components except the i_{th} one.

Expected value: vector case

By extending the scalar definition we have

$$E[v] = \int_{\mathbb{R}^n} q f(q) dq = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix} f(q_1, q_2, \ldots, q_n) dq_1 \ldots dq_n$$

which can be equivalently written as

$$E[v] = \begin{bmatrix} \int_{-\infty}^{+\infty} q_1 f_1(q_1) dq_1 \\ \int_{-\infty}^{+\infty} q_2 f_2(q_2) dq_2 \\ \vdots \\ \int_{-\infty}^{+\infty} q_n f_2(q_n) dq_n \end{bmatrix} = \begin{bmatrix} E[v_1] \\ E[v_2] \\ \vdots \\ E[v_n] \end{bmatrix}$$

Consider a vector random variable *v* and let

$$w = g(v)$$

where

$$g(\cdot) : \bar{\mathbb{R}}^n \to \bar{\mathbb{R}}^n$$

Then in terms of expected value we have

$$E[w] = \int_{\bar{\mathbb{R}}^n} q f_w(q) dq = \int_{\bar{\mathbb{R}}^n} g(q) f_v(q) dq$$

Linearity of expected value: vector case

Given a vector random variable *v* let

$$w = Av = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$$

then

$$E[w] = E[Av] = AE[v]$$

Similarly, column-wise we have

$$w = Av, A = \begin{bmatrix} \alpha_1 & \dots & \alpha_n \end{bmatrix}$$

$$E[w] = E[Av] = AE[v] = \begin{bmatrix} \alpha_1 & \dots & \alpha_n \end{bmatrix} E[v] = \sum_{i=1}^{n} \alpha_i E[v_i].$$

The variance of a vector random variable is defined as

$$\text{Var}[v] = \int_{\mathbb{R}^n} (q - E[v])(q - E[v])^T f(q) dq$$

Clearly Var[v] is a square $n \times n$ matrix, which can be equivalently defined as

$$\text{Var}[v] = E[(v - E[v])(v - E[v])^T]$$

It appears from both expressions that Var[v] is a symmetric positive semi-definite matrix.

Covariance matrix

In scalar form we have

$$\mathrm{Var}[v] = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & & \\ c_{n1} & \cdots & c_{nn} \end{bmatrix}$$

where

- $c_{ii} = \mathrm{Var}[v_i]$ is the variance of v_i
- $c_{ij} = E[(v_i - E[v_i])(v_j - E[v_j])]$ is the covariance index between v_i and v_j.

Covariance and second order moment

As in the scalar case, defining the second order moment as

$$m_2[v] = \int_{\bar{\mathbb{R}}^n} qq^T f(q)dq$$

we have

$$\mathsf{Var}[v] = E[vv^T] - E[v]E[v]^T = m_2[v] - E[v]E[v]^T.$$

The correlation matrix is defined as

$$\rho[v] = \begin{bmatrix} \bar{c}_{11} & \cdots & \bar{c}_{1n} \\ \vdots & & \\ \bar{c}_{n1} & \cdots & \bar{c}_{nn} \end{bmatrix}$$

where $\bar{c}_{ij} = \dfrac{c_{ij}}{\sqrt{c_{ii}c_{jj}}}.$

It follows from the definition that $\bar{c}_{ii} = 1$ and $|\bar{c}_{ij}| \leq 1$.

Two random variables v_1 and v_2 are said to be incorrelated if for the vector random variable

$$v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

we have that

$$\bar{c}_{12}[v] = 0.$$

Theorem: random variables v_1 and v_2 are incorrelated if and only if

$$E[v_1 v_2] = E[v_1]E[v_2].$$

To prove it we compute c_{12}:

$$\begin{aligned} c_{12} &= E[(v_1 - E[v_1])(v_2 - E[v_2])] \\ &= E[v_1 v_2] + E[E[v_1]E[v_2]] - E[v_1 E[v_2]] - E[v_2 E[v_1]] = \\ &= E[v_1 v_2] - E[v_1]E[v_2] \end{aligned}$$

Therefore $c_{12} = 0 \quad \Leftrightarrow \quad E[v_1 v_2] = E[v_1]E[v_2].$

Incorrelation and independence

Two random variables v_1 and v_2 are said to be independent if

$$f(q_1, q_2) = f_1(q_1) f_2(q_2).$$

Theorem: two independent random variables are also incorrelated.

To prove it we compute $E[v_1 v_2]$:

$$\begin{aligned} E[v_1 v_2] &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} q_1 q_2 f(q_1, q_2) dq_1 dq_2 = \\ &= \int_{-\infty}^{+\infty} q_1 f_1(q_1) dq_1 \int_{-\infty}^{+\infty} q_2 f_2(q_2) dq_2 = \\ &= E[v_1] E[v_2]. \end{aligned}$$

(the converse is not true in general)

Consider two random variables v_1 and v_2 and their sum w:

$$w = v_1 + v_2$$

Clearly for the expected value we have

$$E[w] = E[v_1] + E[v_2]$$

but for the variance:

$$\begin{aligned}
\mathrm{Var}[w] = E[(w - E[w])^2] &= E[(v_1 + v_2 - E[v_1] - E[v_2])^2] = \\
&= E[(v_1 - E[v_1])^2 + (v_2 - E[v_2])^2] + \\
&+ 2E[(v_1 - E[v_1])(v_2 - E[v_2])] = \\
&= \mathrm{Var}[v_1] + \mathrm{Var}[v_2] + 2c_{12}
\end{aligned}$$

Linear propagation: expected value and variance

For arbitrary linear combinations of v_1 and v_2

$$z = \alpha_1 v_1 + \alpha_2 v_2$$

For the expected value we have

$$E[z] = \alpha_1 E[v_1] + \alpha_2 E[v_2]$$

but for the variance:

$$\mathrm{Var}[z] = \alpha_1^2 \mathrm{Var}[v_1] + \alpha_2^2 \mathrm{Var}[v_2] + 2\alpha_1\alpha_2 c_{12}$$

Vector Gaussian random variables

A vector of random variables

$$v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

is said to be Gaussian (or, equivalently, its components are said to be jointly Gaussian) if it has a density function of the form

$$f(q) = \alpha e^{-q^T B q}, \quad \alpha > 0, \quad B = B^T > 0$$

As in the scalar case, letting

$$\mu = E[v], \qquad C = \mathrm{Var}[v]$$

it can be shown that

$$f(q) = \frac{1}{\sqrt{\det[C]}(2\pi)^{n/2}} e^{\frac{1}{2}(q-\mu)^T C^{-1}(q-\mu)}.$$

Shorthand notation:

$$v \sim G(\mu, C) \qquad v \sim N(\mu, C)$$

Consider a vector Gaussian random variable such that

$$v \sim G(\mu, C)$$

then

$$v_i \sim G(\mu_i, C_{ii})$$

i.e., the components of a vector Gaussian random variable are in turn Gaussian random variables.

The converse is not true in general.

Consider a set of *independent* Gaussian random variables such that

$$v_i \sim G(\mu_i, C_{ii})$$

then

$$v \sim G(\mu, C)$$

If v_1 and v_2 are Gaussian and incorrelated then they are also independent.

Proof: follows from properties of the exponential.

Consider a *n*-dimensional vector Gaussian random variable

$$v \sim G(\mu_v, C_v)$$

and apply the linear transformation

$$w = Av + b$$

where

- A $m \times n$, $m \leq n$ and rank(A)=m
- b $m \times 1$

Then *w* is Gaussian and

$$w \sim G(\mu_w, C_w)$$

where

$$\mu_w = A\mu_v + b$$

$$C_w = AC_vA^T$$

If v_1 and v_2 are jointly Gaussian such that

$$v_1 \sim G(\mu_1, \sigma_1^2)$$
$$v_2 \sim G(\mu_2, \sigma_2^2)$$

then their linear combination

$$w = \alpha_1 v_1 + \alpha_2 v_2$$

is also Gaussian and such that

$$\mu_w = \alpha_1 \mu_1 + \alpha_2 \mu_2$$

$$\sigma_w^2 = \alpha_1^2 \sigma_1^2 + \alpha_2^2 \sigma_2^2 + 2\alpha_1 \alpha_2 c_{12}$$

Sequences of random variables

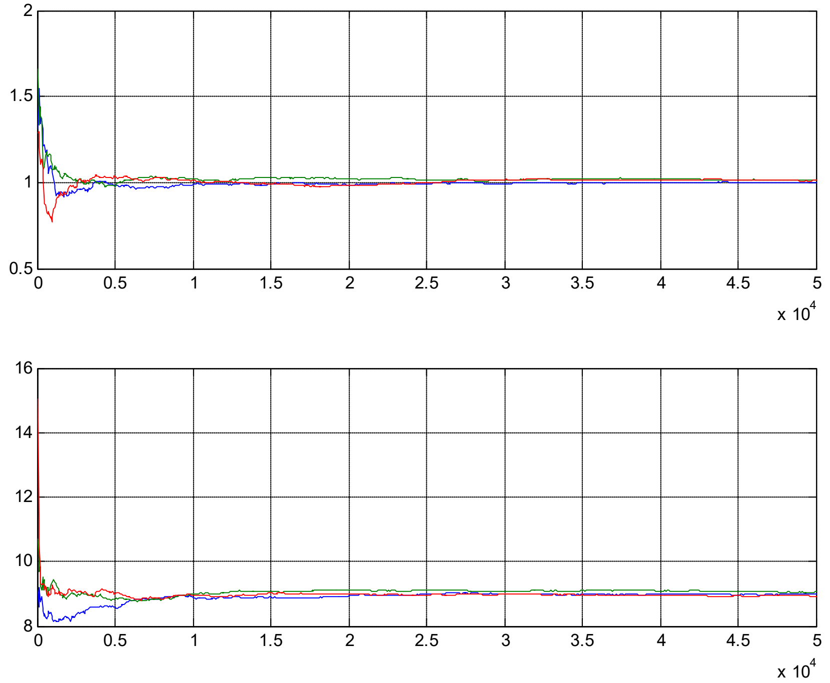
For a given random experiment we have a *sequence* o random variables defined as

v_1, v_2, ..., v_t

and we want to define notions of asymptotic limit for the sequence, i.e.,

$$\lim_{t \to \infty} v_t$$

Example: the study of the properties of an estimator for increasing data lenghts.



Critical issue:

the sequence depends on the outcome of the experiment!

Sequences of random variables: sure convergence

Various notions of convergence can be defined.

Main distinction between *sure* and *almost sure* convergence.

Sure (or strong) convergence:

$$\lim_{t \to \infty} v_t = a$$

is equivalent to

$$\forall \epsilon > 0 \quad \exists t_\epsilon : |v_t - a| < \epsilon \quad \forall t > t_\epsilon \quad \forall s \in \Omega$$

Sequences of random variables: almost sure convergence

A*lmost sure* convergence is defined in terms of the set *A* of outcomes of the experiment for which the sequence converges:

$$A = \left\{ s \in \Omega : \lim_{t \to \infty} v_t = a \right\}$$

If P(*A*)=1 then $\lim_{t \to \infty} v_t = a$ with probability 1 (almost surely).

Consider the limit value *a*, define an interval [a-ϵ, a+ϵ] and consider the set of events

$$B_1(\epsilon) = \{s \in \Omega : |v_1(s) - a| < \epsilon\}$$

B_1 is in turn an event so we can compute its probability:

$$P(B_1(\epsilon)) = g_1(\epsilon)$$

Repeating the process for increasing *t* we have the numerical sequence $g_1, g_2, ..., g_t$

Then, we say that v_t converges in probability to *a*

$$\mathrm{plim}_{t\to\infty} v_t = a$$

if

$$\lim_{t\to\infty} g_t(\epsilon) = 1 \quad \forall \epsilon > 0.$$

Sequences of random variables: mean and mean square convergence

Let

$$\mu_t = E[v_t]$$

then we say that the sequence convergence in mean if

$$\lim_{t \to \infty} \mu_t = a$$

Similarly, let

$$h_t = E[(v_t - a)^2]$$

then we say that the sequence has mean square convergence, denoted as

$$\text{l.i.m.}_{t \to \infty} v_t = a$$

if $\lim_{t \to \infty} h_t = 0$

Properties of mean and mean square convergence

The following implications hold:

If $\quad \text{l.i.m.}_{t\to\infty} v_t = a \Rightarrow \lim_{t\to\infty} E[v_t] = a$

If $\quad \lim_{t\to\infty} E[v_t] = a \quad \text{and} \quad \lim_{t\to\infty} Var[v_t] = a$

$\Rightarrow \text{l.i.m.}_{t\to\infty} v_t = a$

If $\quad \lim_{t\to\infty} E[v_t] = a \quad \text{then}$

$\lim_{t\to\infty} Var[v_t] = a \Leftrightarrow \text{l.i.m.}_{t\to\infty} v_t = a$

Sequences of random variables: convergence in distribution

Up to now convergence to a *constant value* has been considered.

$$\text{plim}_{t\to\infty} v_t = a$$

What if *a* is a random variable, with distribution F(*a*)?

Denoting with F(*q*,*t*) the distribution of v_t, if

$$\lim_{t\to\infty} F(q,t) = F_a(q), \quad \forall q$$

then we say that

$$\lim_{t\to\infty} v_t = a$$

in distribution.

In the Gaussian case:

if $a \sim G(\mu, \sigma^2)$

then we say that v_t is asymptotically Gaussian:

$$v_t \sim AsG(\mu, \sigma^2)$$

Sequences of random variables: summary

Summing up, the following implications hold.

Sure convergence $\Rightarrow$ A.s. convergence

A. s. convergence $\Rightarrow$ Convergence in prob.

Convergence in prob. $\Rightarrow$ Convergence in distr.

But also...

Mean square convergence $\Rightarrow$ Convergence in prob.

Mean square convergence $\Rightarrow$ Convergence in mean

Consider N independent real random variables v_i such that

$$E[v_i] = \mu, \quad \forall i$$

and their sum

$$x_N = \sum_i v_i$$

Then the following results hold.

Theorem 1:

if the v_i are identically distributed then

$$\lim_{N \to \infty} \frac{x_N}{N} = \mu, \quad \text{a.s. and m.s.}$$

Theorem 2:

if the v_i are such that

$$Var[v_i] \leq C, \quad \forall i$$

then

$$\lim_{N\to\infty} \frac{1}{N}(x_N - E[x_N]) = 0, \quad \text{a.s. and m.s.}$$

Central limit theorem

Consider N independent and identically distributed real random variables v_i such that

$$E[v_i] = \mu, \quad \forall i$$

$$Var[v_i] = \sigma^2, \quad \forall i$$

then their sum

$$x_N = \sum_i v_i$$

is such that $E[x_N] = N\mu, \quad Var[x_N] = N\sigma^2$

and $y_N = \dfrac{x_N - N\mu}{\sqrt{N\sigma^2}} \sim AsG(0,1).$

Conditional probability

Consider a random experiment defined by {Ω, **C**, P} and study the probabilities of two events *A* and *C*.

The conditional probability of *A* given *C* is defined as

$$P(A\backslash C) = \frac{P(A \cap C)}{P(C)}$$

Example: rolling a dice.

$$P(A \backslash C) = \frac{P(A \cap C)}{P(C)}$$

$\Omega=\{1,2,3,4,5,6\}$

C= all subsets of Ω

Consider

A={1,2,3,5} and C={2,4,6}.

Clearly P(A)=4/6=2/3 and P(C)=3/6=1/2.

$$P(A \setminus C) = \frac{P(A \cap C)}{P(C)}$$

A∩C= {2}, so P(A∩C)=1/6.

Therefore

P(A\C)=1/3.

Conditional probability

If now we *fix* C and consider the function

$$P(\cdot \backslash C)$$

defined in **C** and taking values in [0,1], we have defined the probability of any event in **C** given event C.

It has to be checked that this function is a well-defined probability function, *i.e.*, it satisfies the properties defined earlier on.

P is a function mapping C to the [0, 1] interval, satisfying:

- P(Ω) = 1: $P(\Omega \backslash C) = \dfrac{P(\Omega \cap C)}{P(C)} = \dfrac{P(C)}{P(C)} = 1$
- If for N < ∞ events $A_1, A_2, \ldots, A_N \in$ **C**, and

$$A_i \bigcap A_j = 0, \quad \forall i, j$$

then

$$P(\bigcup_i A_i) = \sum_i P(A_i)$$

The second property holds as we have the following.

Conditional probability

$$P(\bigcup_i A_i \backslash C) = \frac{P(\bigcup_i A_i \cap C)}{P(C)} = \frac{P(\bigcup_i (A_i \cap C))}{P(C)} =$$

$$= \frac{\sum_i P((A_i \cap C))}{P(C)} = \sum_i P(A_i \backslash C)$$

Constrained random experiment

We can now consider a *constrained* random experiment defined by

{Ω, **C**, P(·\C)}

as a random experiment constrained to the event C.

Partition of Ω

A partition of Ω is defined as a set

$$\Pi = \{C_1, C_2, \ldots, C_n\}, \quad C_i \subseteq \Omega$$

with the following properties:

- The sets C_i are all disjoint
- $\cup_i C_i = \Omega$.

Total probability theorem

Given a random experiment and a partition Π such that

$$\Pi \subseteq \mathbf{C}$$

and

$$P(C_i) \neq 0$$

then we have

$$P(A) = \sum_i P(A \backslash C_i) P(C_i) \quad \forall A \in \mathbf{C}$$

Proof: A can be written as

$$A = A \cap \Omega = A \cap (\cup_i C_i) = \cup_i (A \cap C_i)$$

so in terms of probabilities

$$P(A) = P(\cup_i (A \cap C_i)) = \sum_i P(A \cap C_i) = \sum_i P(A \backslash C_i) P(C_i)$$

For two events A and $B \in \mathbf{C}$ with P(A), P(B) $\neq$ 0 it holds that

$$P(A \backslash B) = \frac{P(B \backslash A) P(A)}{P(B)}$$

Proof: multiply both sides by P(B) to get P(A $\cap$ B) on both sides of the equation.

Extension of Bayes Theorem

Let

$$\Pi = \{C_1, C_2, \ldots, C_n\}, \quad C_i \subseteq \mathbf{C}$$

a partition of Ω and consider an event B $\in$ **C**.

Then

$$P(A_i \backslash B) = \frac{P(B \backslash A_i) P(A_i)}{\sum_i P(B \backslash A_i) P(A_i)}.$$

Usual nomenclature:

- $P(A_i)$: *a priori* probability
- $P(A_i \backslash B)$: *a posteriori* probability

with respect to the conditioning to *B*.

Independent events

Two events A and B $\in$ **C** are called *independent* if and only if

$$P(A \cap B) = P(A)P(B)$$

Clearly for independent events we have, in terms of conditional probabilities

$$P(A \backslash B) = P(A)$$

$$P(B \backslash A) = P(B)$$

The above ideas can lead to the definition of conditional distributions and conditional densities, as follows.

Consider a random experiment and a random variable v defined on it.

Then pick an event $C \in \mathbf{C}$: $P(C) \neq 0$.

Then the distribution function for v conditional to C is defined as the distribution function for the constrained experiment.

Conditional distribution function

Consider the random experiment {Ω, **C**, P(·\C)} and random variable *v*, then the conditional distribution is

$$F(q\backslash C) = \frac{P(v \leq q, s \in C)}{P(C)}, \qquad \forall q \in \bar{\mathbb{R}}$$

where we can write equivalently

$$P(v \leq q, s \in C) = P(\phi^{-1}([-\infty, q]) \cap C)$$

Conditional probability density function

A conditional probability density function for a given conditional distribution can be defined as

$$f(q \backslash C) = \frac{dF(q \backslash C)}{dq}$$

Total probability theorem (for distributions)

Consider a partition

$$\Pi = \{C_1, C_2, \ldots, C_n\}, \qquad C_i \subseteq \mathbf{C}$$

such that $P(C_i) \neq 0 \ \forall\, i$.

Then

$$F(q) = \sum_i F(q \backslash C_i) P(C_i), \qquad \forall q \in \bar{\mathbb{R}}$$

If the conditioning event is given by

$$C = \phi^{-1}([-\infty, r]), \quad r \in \bar{\mathbb{R}}$$

then by definition

$$F(q \backslash C) = \frac{P(v \leq q, v \leq r)}{P(v \leq r)} = \frac{P(v \leq q, v \leq r)}{F(r)}$$

But clearly $P(v \leq q, v \leq r) = P(v \leq \min(q, r))$ so

$$F(q \backslash C) = \begin{cases} \frac{F(q)}{F(r)} & q \leq r \\ 1 & q > r \end{cases}$$

More on conditional densities

As a consequence, if

$$F(q \backslash C) = \begin{cases} \frac{F(q)}{F(r)} & q \leq r \\ 1 & q > r \end{cases}$$

then in terms of densities we have

$$f(q \backslash C) = \frac{dF(q \backslash C)}{dq} = \begin{cases} \frac{f(q)}{F(r)} & q \leq r \\ 0 & q > r \end{cases}$$

or equivalently

$$f(q \backslash C) = \frac{dF(q \backslash C)}{dq} = \begin{cases} \frac{f(q)}{\int_{-\infty}^{r} f(w)dw} & q \leq r \\ 0 & q > r \end{cases}$$

More on conditional distributions and densities

For a generic conditioning event E we have the conditional density

$$f(q \backslash E) = \begin{cases} \dfrac{f(q)}{\int_E f(w)dw} & q \notin E \\ 0 & q \in E \end{cases}$$

and the corresponding distribution

$$F(q \backslash v \in E) = \int_{-\infty}^{q} f(r \backslash v \in E) dr.$$

Given a real random variable *v* and the conditional density function f(q\C) the conditional expectation of *v* given C is defined as

$$E[v\backslash C] = \int_{-\infty}^{+\infty} q f(q\backslash C) dq.$$

Furthermore, if C is defined on *v*, we have

$$E[v\backslash v \in E] = \int_{-\infty}^{+\infty} q f(q\backslash v \in E) dq = \int_{E} q f(q\backslash v \in E) dq =$$
$$= \frac{\int_E q f(q) dq}{\int_E f(q) dq}.$$

Vector conditional distribution

Consider the random experiment {Ω, **C**, P(·\C)} and a vector random variable *v*, then the conditional distribution is

$$F(q\backslash C) = \frac{P(v_1 \leq q_1, \ldots, v_n \leq q_n, s \in C)}{P(C)}, \qquad \forall q \in \bar{\mathbb{R}}^n$$

where we can write equivalently

$$\begin{aligned} &P(v_1 \leq q_1, \ldots, v_n \leq q_n, s \in C) = \\ =&P(\phi^{-1}(v_1 \leq q_1, \ldots, v_n \leq q_n) \cap C) \end{aligned}$$

Vector conditional density

Similarly, for the conditional density function we get

$$f(q_1, \dots, q_n \backslash C) = \frac{\partial F(q_1, \dots, q_n \backslash C)}{\partial q_1 \dots \partial q_n}.$$

and if the event C is defined on v as v $\in$ E we get

$$f(q_1, \dots, q_n \backslash C) = \begin{cases} \frac{f(q_1, \dots, q_n)}{\int_E f(q_1, \dots, q_n) dq_1, \dots, dq_n} & q \notin E \\ 0 \quad q \in E \end{cases}$$

Vector conditional density: a particular case

What if the conditioning event corresponds to a line?

We get a conditional density given by (*n*=2 case)

$$f_1(q_1 \backslash v_2 = q_2) = \frac{f(q_1, q_2)}{f_2(q_2)}$$

At the level of vector conditional densities they can be stated as

$$f_1(q_1) = \int_{-\infty}^{+\infty} f_1(q_1 \backslash q_2) f_2(q_2) dq_2$$

$$f_1(q_1 \backslash q_2) = \frac{f_2(q_2 \backslash q_1) f_1(q_1)}{f_2(q_2)}$$