Elements of statistical analysis

Ensemble average

The concept of an *ensemble average* is based upon the existence of independent statistical events. For example, consider a number of individuals who are simultaneously flipping unbiased coins. If a value of one is assigned to a head and the value of zero to a tail, then the *arithmetic average* of the numbers generated is defined as:

$$X_N = \frac{1}{N} \Sigma x_n \qquad (2.1)$$

where our nth flip is denoted as x_n and N is the total number of flips.

The key is the events must be independent

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Ensemble average

Now imagine that we are trying to establish the nature of a random variable, x. The nth *realization* of x is denoted as x_n. The *ensemble average* of x is denoted as X (or $\langle x \rangle$), and *is defined as*

$$X = \langle x \rangle \equiv \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} x_n \qquad (2.2)$$

Obviously it is impossible to obtain the ensemble average experimentally, since we can never have an infinite number of independent realizations. The most we can ever obtain is the arithmetic mean for the number of realizations we have. For this reason the arithmetic mean can also referred to as the *estimator* for the true mean or ensemble average.

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Ensemble average

Unless stated otherwise, all analyses will utilise the concept of ensemble average.

This means that we need to be aware of or take in to account the "statistical differences" between the true mean and the estimates of mean

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Ensemble average

In general, the x_n could be realizations of any random variable. The X defined by equation 2.2 represents the ensemble average of it. The quantity X is sometimes referred to as the *expected value* of the random variable x, or even simply its *mean*.

For example, the velocity vector at a given point in space and time, $\vec{x}, t$, in a given turbulent flow can be considered to be a random variable, say $u_i(\vec{x}, t)$. If there were a large number of identical experiments so that the $u_i^{(n)}(\vec{x}, t)$ in each of them were identically distributed, then the ensemble average of $u_i^{(n)}(\vec{x}, t)$ would be given by

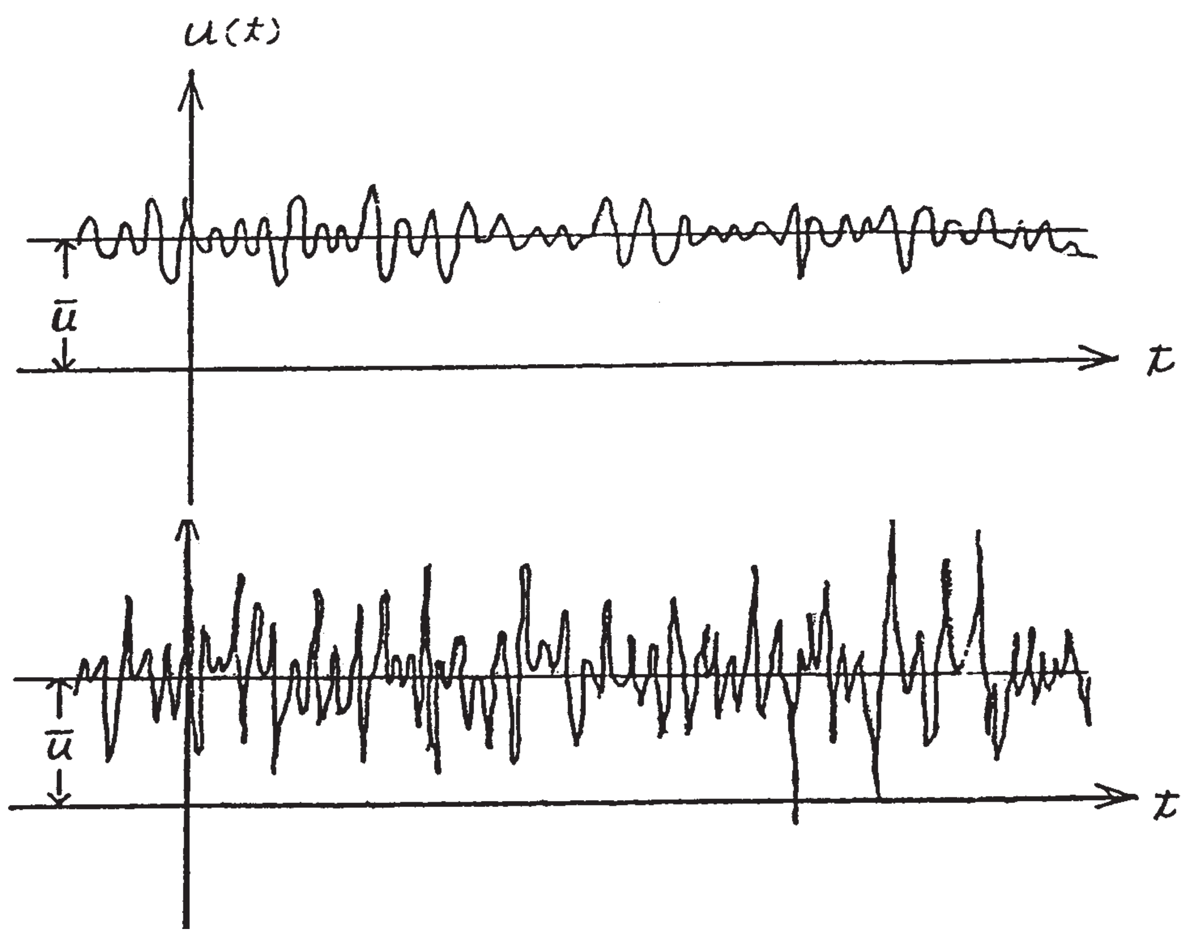
$$\langle u_i(\vec{x}, t) \rangle = U_i(\vec{x}, t) \equiv \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} u_i^{(n)}(\vec{x}, t) \tag{2.3}$$

source: Turbulence for the 21st century, W. K. George

Turbulence: lecture

Elements of statistical analysis

Fluctuations about the mean



Possible to distinguish between these two signals by looking at the fluctuations

$$x' = x - X$$

$$\langle x' \rangle = 0$$

The signals can be distinguished by calculating the variance

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Fluctuations about the mean

variance is defined as:

$$var[x] \equiv \langle (x')^2 \rangle = \langle [x - X]^2 \rangle \tag{2.6}$$

$$= \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} [x_n - X]^2 \tag{2.7}$$

Note that the variance, like the ensemble average itself, can never really be measured, since it would require an infinite number of members of the ensemble.

Elements of statistical analysis

Fluctuations about the mean

The variance can also referred to as the *second central moment of* x. The word central implies that the mean has been subtracted off before squaring and averaging. The reasons for this will be clear below. If two random variables are identically distributed, then they must have the same mean and variance.

The variance is closely related to another statistical quantity called the *standard deviation* or root mean square (rms) value of the random variable x,which is denoted by the symbol, σ_x. Thus,

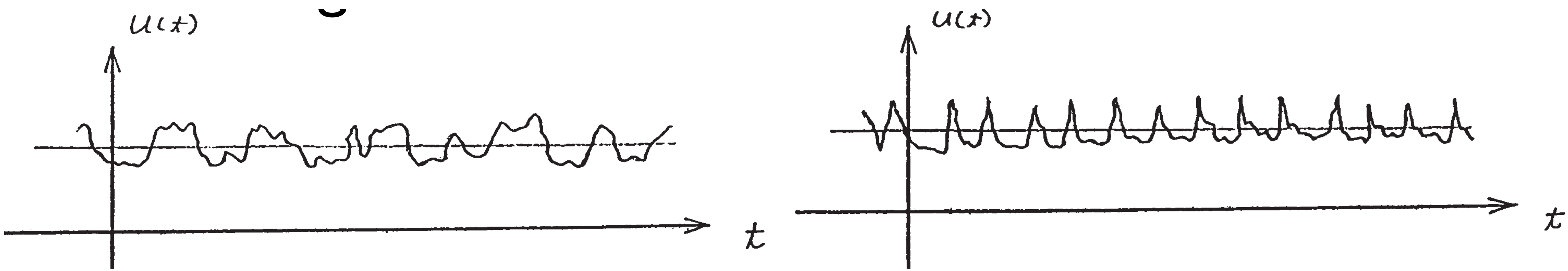
$$\sigma_x \equiv (var[x])^{1/2} \tag{2.9}$$

or $\sigma_x^2 = var[x]$.

Elements of statistical analysis

Higher moments

Two signals with the same mean and variance



The m-th moment of the random variable is defined as:

$$\langle x^m \rangle = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} x_n^m \tag{2.10}$$

It is usually more convenient to work with the *central moments* defined by:

$$\langle (x')^m \rangle = \langle (x - X)^m \rangle = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} [x_n - X]^m \tag{2.11}$$

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Probability

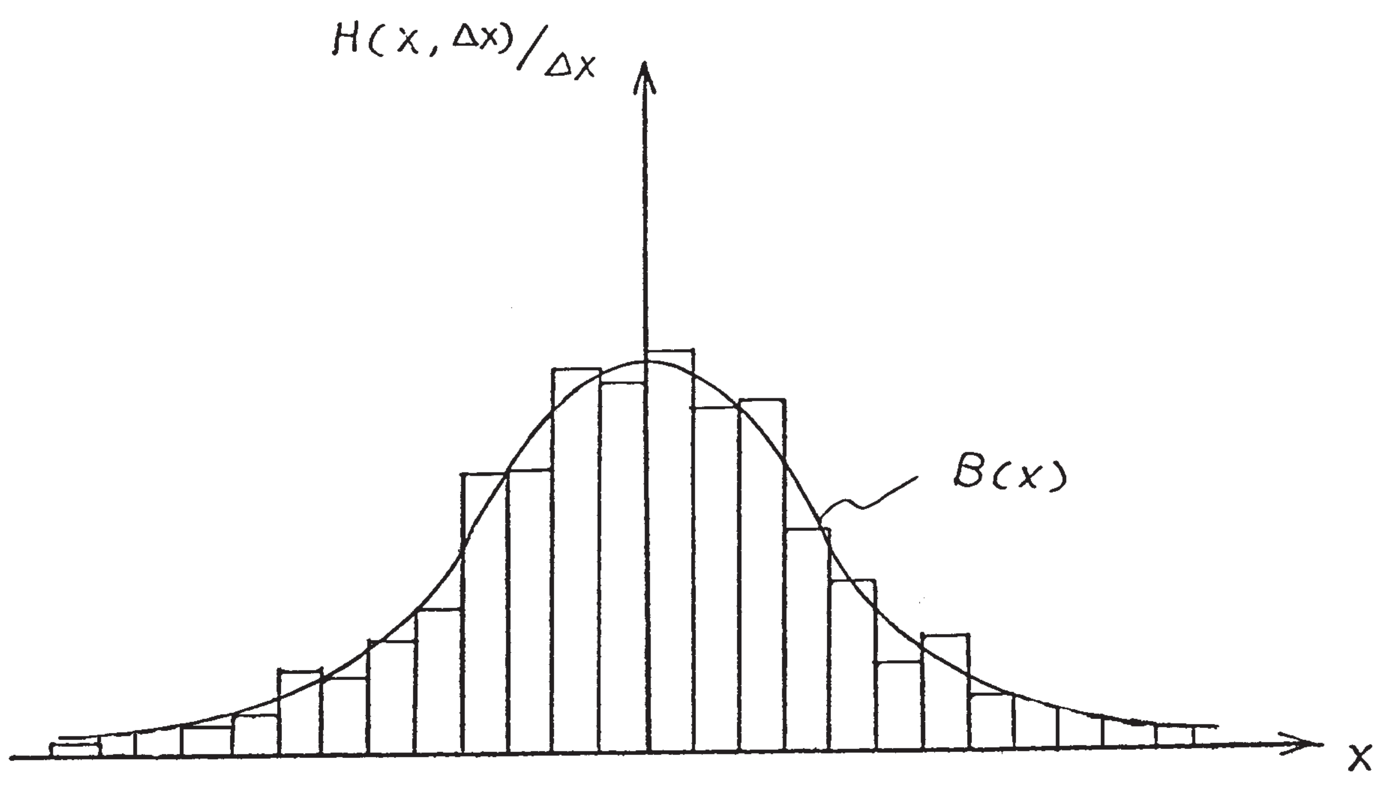
The frequency of occurrence of a given *amplitude* (or value) from a finite number of realizations of a random variable can be displayed by dividing the range of possible values of the random variables into a number of slots (or windows). Since all possible values are covered, each realization fits into only one window. For every realization a count is entered into the appropriate window. When all the realizations have been considered, the number of counts in each window is divided by the total number of realizations. The result is called the **histogram** (or *frequency of occurrence* diagram). From the definition it follows immediately that the sum of the values of all the windows is exactly one.

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Probability

The shape of a histogram depends on the *statistical distribution of the random variable*, but it also depends on the total number of realizations, N, and the size of the slots, Δc. The histogram can be represented symbolically by the function $H_x(c, \Delta c, N)$ where $c \leq x < c + \Delta c$, Δc is the slot width, and N is the number of realizations of the random variable. Thus the histogram shows the relative frequency of occurrence of a given value range in a given ensemble.



source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Probability

If the number of realizations, N, increases without bound as the window size, Δc, goes to zero, the histogram divided by the window size goes to a limiting curve called the *probability density function*, $B_x(c)$. That is,

$$B_x(c) \equiv \lim_{\substack{N \to \infty \\ \Delta c \to 0}} H(c, \Delta c, N)/\Delta c \tag{2.12}$$

Note that as the window width goes to zero, so does the number of realizations which fall into it, NH. Thus it is only when this number (or relative number) is divided by the slot width that a meaningful limit is achieved.

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Probability

The **probability density function** (or **pdf**) has the following properties:

- Property 1:
$$B_x(c) > 0 \tag{2.13}$$
always.

- Property 2:
$$Prob\{c < x < c + dc\} = B_x(c)dc \tag{2.14}$$
where $Prob\{\ \ \}$ is read "the probability that".

- Property 3:
$$Prob\{c < x\} = \int_{-\infty}^{x} B_x(c)dc \tag{2.15}$$

- Property 4:
$$\int_{-\infty}^{\infty} B_x(x)dx = 1 \tag{2.16}$$

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Probability

Since $B_x(c)dc$ gives the probability of the random variable x assuming a value between c and $c + dc$, any moment of the distribution can be computed by integrating the appropriate power of x over all possible values. Thus the n-th moment is given by:

$$\langle x^n \rangle = \int_{-\infty}^{\infty} c^n B_x(c) dc \tag{2.17}$$

If the probability density is given, the moments of all orders can be determined. For example, the variance can be determined by:

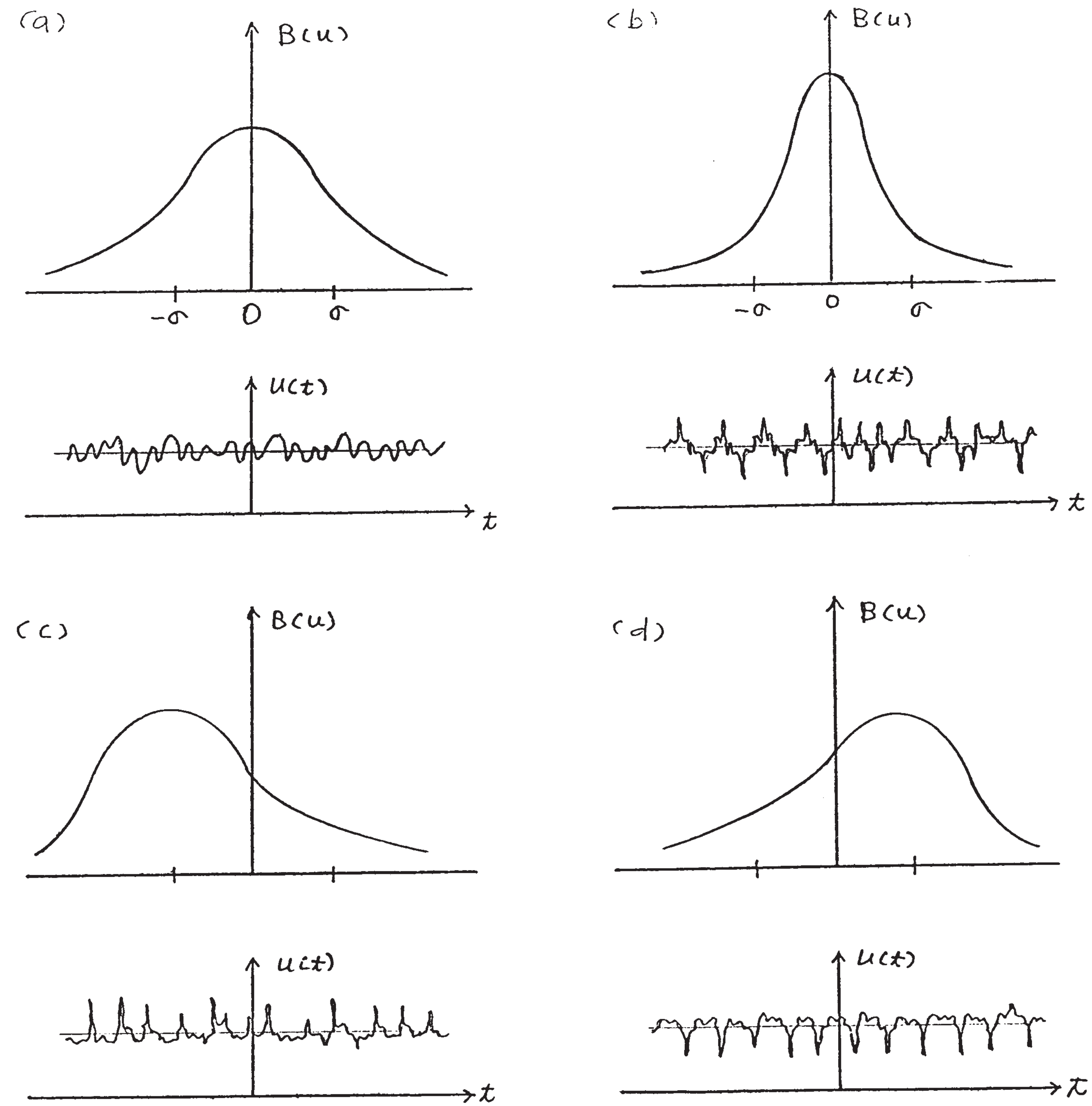
$$var\{x\} = \langle (x - X)^2 \rangle = \int_{-\infty}^{\infty} (c - X)^2 B_x(c) dc \tag{2.18}$$

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Probability

What are the distinguishing moments for these distributions?



source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Skewness and Kurtosis

Because of their importance in characterizing the shape of the pdf, it is useful to define scaled (or normalized) versions of third and fourth central moments: the *skewness* and *kurtosis* respectively. The *skewness* is defined as third central moment divided by the three-halves power of the second; i.e.,

$$S = \frac{\langle (x - X)^3 \rangle}{\langle (x - X)^2 \rangle^{3/2}} \tag{2.26}$$

The *kurtosis* is defined as the fourth central moment divided by the square of the second; i.e.,

$$K = \frac{\langle (x - X)^4 \rangle}{\langle (x - X)^2 \rangle^2} \tag{2.27}$$

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Probability distribution

Sometimes it is convenient to work with the **probability distribution** instead of with the probability density function. The probability distribution is defined as the probability that the random variable has a value less than or equal to a given value. Thus from equation 2.15, the probability distribution is given by

$$F_x(c) = Prob\{x < c\} = \int_{-\infty}^{c} B_x(c')dc' \tag{2.21}$$

Note that we had to introduce the integration variable, c', since c occurred in the limits.

Equation 2.21 can be inverted by differentiating by c to obtain

$$B_x(c) = \frac{dF_x}{dc} \tag{2.22}$$

Elements of statistical analysis

Normal distribution

One of the most important pdf's in turbulence is the Gaussian or Normal distribution defined by

$$B_{xG}(c) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-(c-X)^2/2\sigma_x^2} \tag{2.23}$$

where X is the mean and σ_x is the standard derivation. The factor of $1/\sqrt{2\pi}\sigma_x$ insures that the integral of the pdf over all values is unity as required. It is easy to prove that this is the case by completing the squares in the integration of the exponential (see problem 2.2).

Distribution completely determined by first two moments

It is straightforward to show by integrating by parts that all the even central moments above the second are given by the following recursive relationship,

$$\langle (x - X)^n \rangle = (n-1)(n-3)...3.1\sigma_x^n \tag{2.24}$$

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Joint pdfs and joint moments

It is usually important in turbulence to consider joint statistics of flow variables

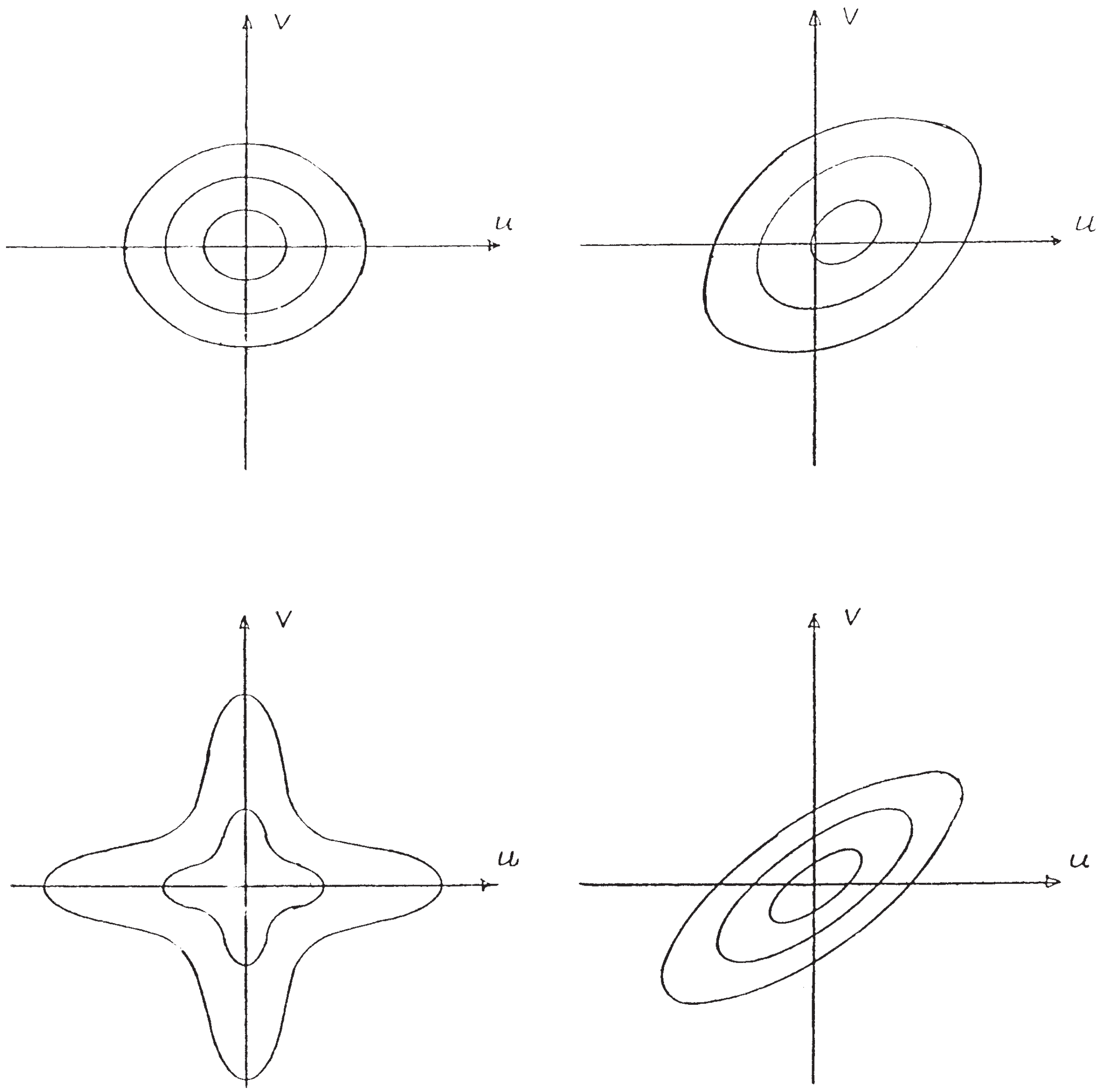
For example if u and v are two random variables, there are three second-order moments which can be defined $\langle u^2 \rangle$, $\langle v^2 \rangle$, and $\langle uv \rangle$. The product moment $\langle uv \rangle$ is called the *cross-correlation* or *cross-covariance*. The moments $\langle u^2 \rangle$ and $\langle v^2 \rangle$ are referred to as the *covariances*, or just simply the *variances*. Sometimes $\langle uv \rangle$ is also referred to as the *correlation*.

The joint probability density function can be constructed using the joint histogram following a procedure described before, but, now, we need to do this with two variables

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Joint pdfs and joint moments



$$u' = u - U$$
$$v' = v - V$$

u' and v' are random variables about the mean

$\langle u'v' \rangle > 0$ (positive correlation)

$\langle u'v' \rangle < 0$ (negative correlation)

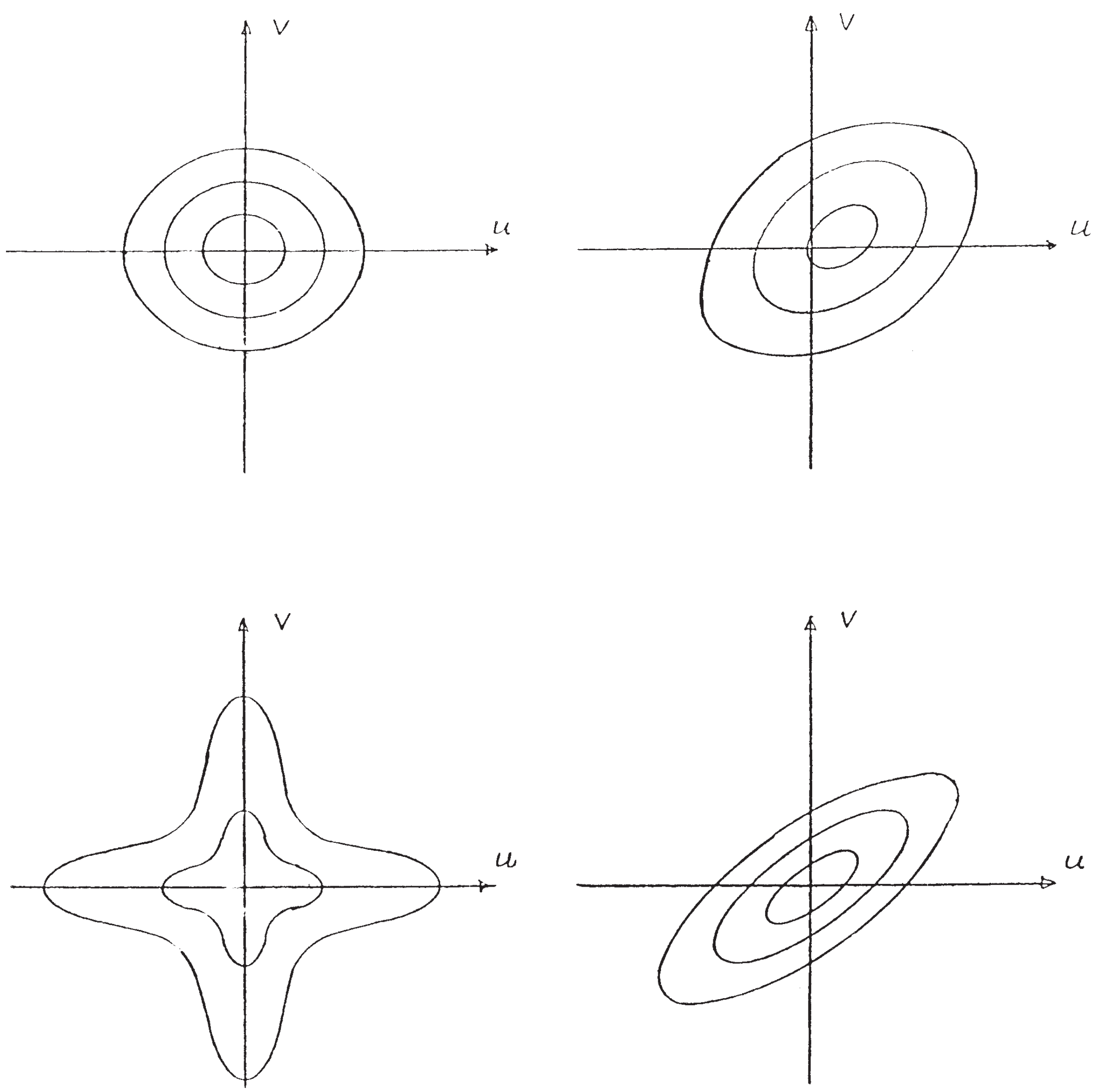
$\langle u'v' \rangle = 0$ (uncorrelated)

source: Turbulence for the 21st century, W. K. George

Turbulence: lecture

Elements of statistical analysis

Joint pdfs and joint moments



$$\rho_{uv} \equiv \frac{\langle u'v' \rangle}{\left[\langle u'^2 \rangle \langle v'^2 \rangle\right]^{1/2}}$$

correlation coefficient (-1 to 1)

1 : perfect correlation
-1: perfect anti-correlation

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Joint pdfs and joint moments

As with the single-variable pdf, there are certain conditions the joint probability density function must satisfy. If $B_{uv}(c_1, c_2)$ indicates the jpdf of the random variables u and v, then:

- Property 1:

$$B_{uv}(c_1, c_2) > 0 \tag{2.31}$$

  always.

- Property 2:

$$Prob\{c_1 < u < c_1 + dc_1, c_2 < v < c_2 + dc_2\} = B_{uv}(c_1, c_2)dc_1, dc_2 \tag{2.32}$$

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Joint pdfs and joint moments

- Property 3:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} B_{uv}(c_1, c_2)dc_1dc_2 = 1 \tag{2.33}$$

- Property 4:

$$\int_{-\infty}^{\infty} B_{uv}(c_1, c_2)dc_2 = B_u(c_1) \tag{2.34}$$

where B_u is a function of c_1 only.

- Property 5:

$$\int_{-\infty}^{\infty} B_{uv}(c_1, c_2)dc_1 = B_v(c_2) \tag{2.35}$$

where B_v is a function of c_2 only.

Elements of statistical analysis

Joint pdfs and joint moments

If the joint probability density function is known, the *joint moments* of all orders can be determined. Thus the m,n-th joint moment is

$$\langle u^m v^n \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c_1^m c_2^n B_{uv}(c_1, c_2) dc_1 dc_2 \tag{2.36}$$

where m and n can take any value. The corresponding central-moment is:

$$\langle (u-U)^m (v-V)^n \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (c_1 - U)^m (c_2 - V)^n B_{uv}(c_1, c_2) dc_1 dc_2 \tag{2.37}$$

In the preceding discussions, only two random variables have been considered. The definitions, however, can easily be generalized to accommodate any number of random variables. In addition, the joint statistics of a single random variable at different times or at different points in space could be considered. This will be discussed later when stationary and homogeneous random processes are considered.

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Statistical independence and lack of correlation

Definition: Statistical Independence Two random variables are said to be *statistically independent* if their joint probability density is equal to the product of their marginal probability density functions. That is,

$$B_{uv}(c_1, c_2) = B_u(c_1) B_v(c_2) \tag{2.39}$$

It is easy to see that statistical independence implies a complete lack of correlation; i.e., $\rho_{uv} \equiv 0$. From the definition of the cross-correlation,

$$\begin{aligned}
\langle (u-U)(v-V) \rangle &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (c_1 - U)(c_2 - V) B_{uv}(c_1, c_2) dc_1 dc_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (c_1 - U)(c_2 - V) B_u(c_1) B_v(c_2) dc_1 dc_2 \\
&= \int_{-\infty}^{\infty} (c_1 - U) B_u(c_1) dc_1 \int_{-\infty}^{\infty} (c_2 - V) B_v(c_2) dc_2 \\
&= 0 \qquad (2.40)
\end{aligned}$$

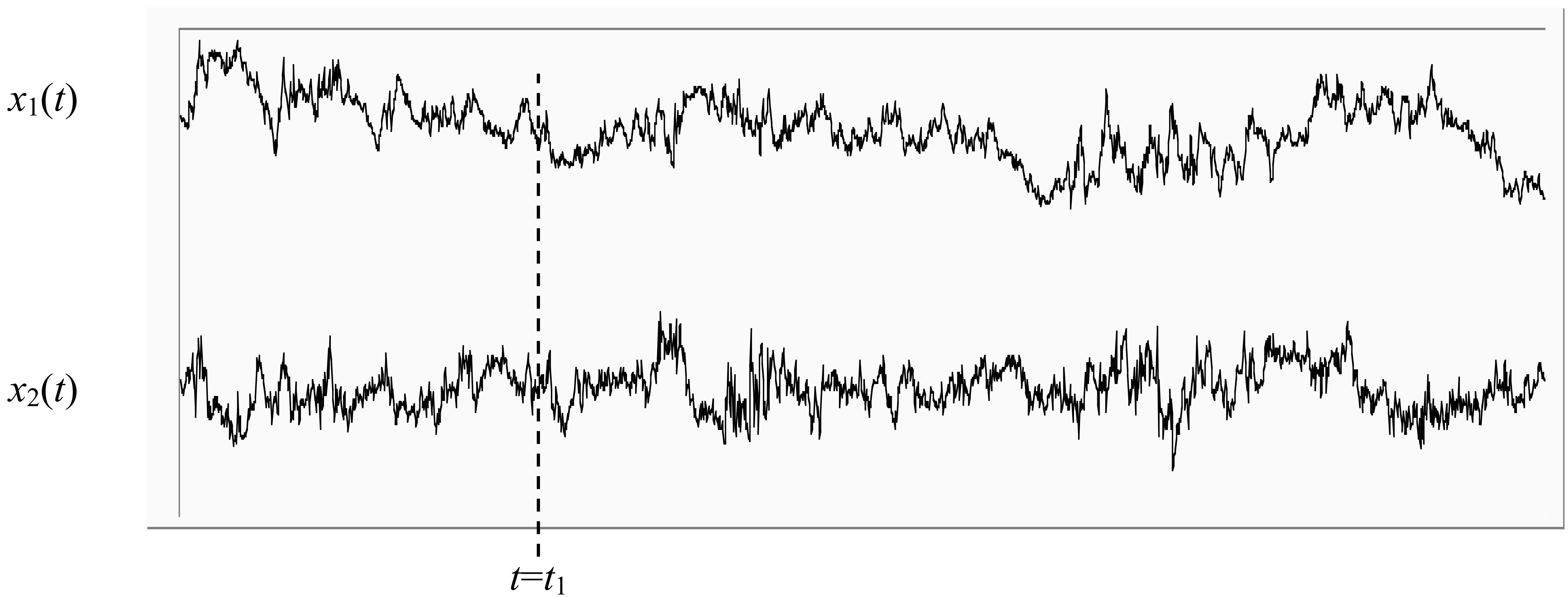
source: Turbulence for the 21st century, W. K. George

Turbulence: lecture

Elements of statistical analysis

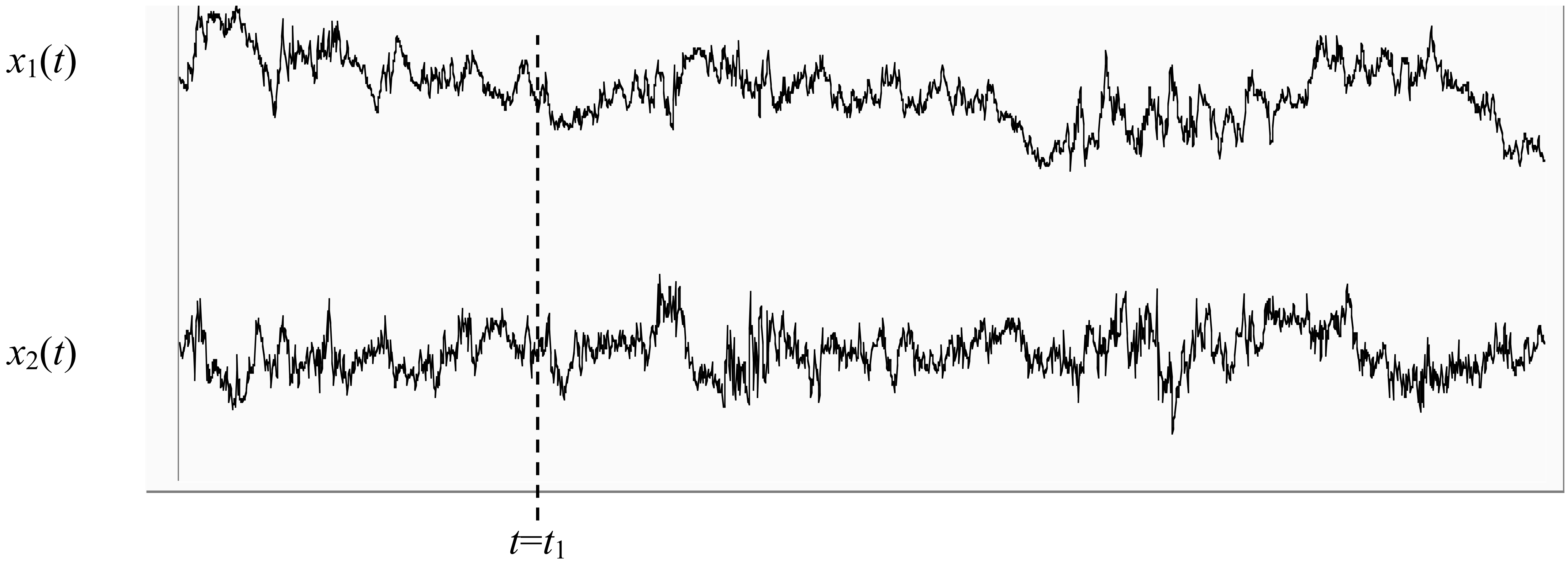
Stationarity and Ergodicity

Take two realisations of all possible time histories of some property of a random (or turbulent) flow:



Elements of statistical analysis

Stationarity and Ergodicity



These samples could be produced by doing the experiment twice. The collection of all possible realisations is the 'random process'. The mean value at time t_1 of <u>all</u> the samples is:

$$\overline{x(t_1)} = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{\infty} x_n(t_1),$$

and similarly for powers of $x(t)$. This is called an <u>ensemble</u> average.

Turbulence: lecture

Elements of statistical analysis

Stationarity and Ergodicity

If all these ensemble averages do not vary with t_1 the process is stationary. In other words, a stationary process is one in which all possible moments and joint-moments are time-invariant.

But it is also possible to describe the properties of $x(t)$ by computing time averages over specific samples (realisations). Consider the n'th sample, $x_n(t)$. It has a mean value given by:

$$\overline{x_n} = \lim_{T \to \infty} \frac{1}{T} \int_0^T x_n(t) \mathrm{d}t.$$

If $x(t)$ is stationary AND $\overline{x_n}$ does not depend on n, the process is ergodic.

Elements of statistical analysis

Stationarity and Ergodicity

For ergodic processes, the time averages are equal to the corresponding ensemble averages

$$\overline{x(t)} = \overline{x_n}$$

Only stationary processes are ergodic

In practice, random data representing physical phenomena are ergodic. Therefore, we can analyse them based on single observed time history, provided that the data record is long enough for the quantity of interest

Elements of statistical analysis

Stationarity and Ergodicity

For ergodic processes, the time averages are equal to the corresponding ensemble averages

$$\overline{x(t)} = \overline{x_n}$$

Only stationary processes are ergodic

In practice, random data representing physical phenomena are ergodic. Therefore, we can analyse them based on single observed time history, provided that the data record is long enough for the quantity of interest

Elements of statistical analysis

Estimation from finite number of realisations

The question of *convergence* of the estimator can be addressed by defining the square of **variability of the estimator**, say $\epsilon^2_{X_N}$, to be:

$$\epsilon^2_{X_N} \equiv \frac{var\{X_N\}}{X^2} = \frac{\langle (X_N - X)^2 \rangle}{X^2} \tag{2.49}$$

Now we want to examine what happens to ϵ_{X_N} as the number of realizations increases. For the estimator to converge it is clear that ϵ_x should decrease as the number of samples increases.

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Estimation from finite number of realisations

We examine the variance of X_N,

$$
\begin{aligned}
var\{X_N\} &= \langle (X_N - X)^2 \rangle \\
&= \langle \left[\frac{1}{N} \sum_{n=1}^{N} x_n \; - \; X) \right]^2 \rangle && (2.50) \\
&= \langle \left[\frac{1}{N} \sum_{n=1}^{N} x_n - \frac{1}{N} \sum_{n=1}^{N} X) \right]^2 \rangle && (2.51) \\
&= \langle \left[\frac{1}{N} \sum_{n=1}^{N} (x_n - X) \right]^2 \rangle && (2.52)
\end{aligned}
$$

since $\langle X_N \rangle = X$ from equation 2.46. Using the fact that the operations of averaging and summation commute, the squared summation can be expanded as follows:

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Estimation from finite number of realisations

Using the fact that summation and averaging commute,

$$\begin{aligned}
\left\langle\left[\sum_{n=1}^{N}(x_n - X)\right]^2\right\rangle &= \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\langle(x_n - X)(x_m - X)\rangle \\
&= \frac{1}{N^2}\sum_{n=1}^{N}\langle(x_n - X)^2\rangle \\
&= \frac{1}{N}var\{x\}, \qquad (2.53)
\end{aligned}$$

where the next to last step follows from the fact that the x_n are assumed to be statistically independent samples (and hence uncorrelated), and the last step from the definition of the variance.

source: Turbulence for the 21st century, W. K. George

Elements of statistical analysis

Estimation from finite number of realisations

The square of the variability of X_N is given by

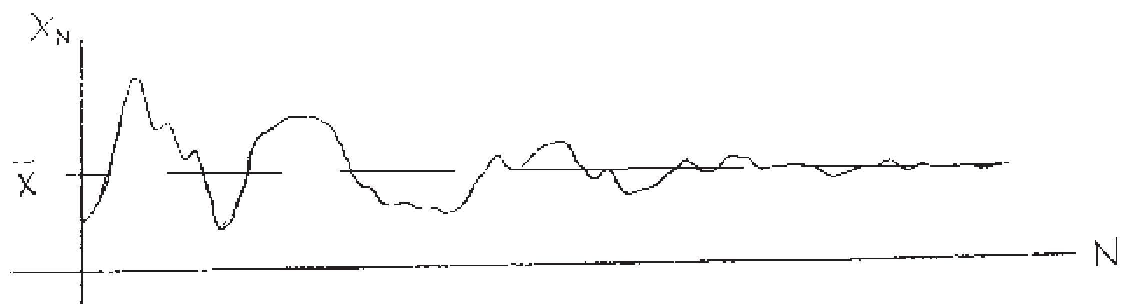
$$\begin{aligned} \epsilon^2_{X_N} &= \frac{1}{N}\frac{var\{x\}}{X^2} \\ &= \frac{1}{N}\left[\frac{\sigma_x}{X}\right]^2 \end{aligned} \qquad (2.54)$$

Thus *the variability of the estimator depends inversely on the number of independent realizations, N, and linearly on the relative fluctuation level of the random variable itself, σ_x/X.* Obviously if the relative fluctuation level is zero (either because there the quantity being measured is constant and there are no measurement errors), then a single measurement will suffice. On the other hand, as soon as there is any fluctuation in the x itself, the greater the fluctuation (relative to the mean of x, $\langle x \rangle = X$), then the more independent samples it will take to achieve a specified accuracy.

source: Turbulence for the 21st century, W. K. George

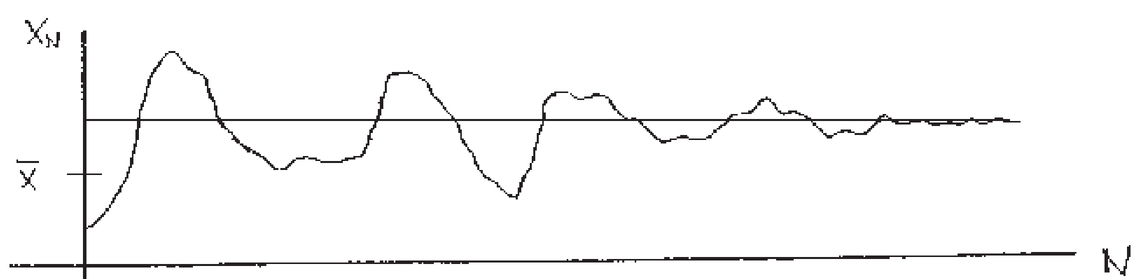
Elements of statistical analysis

Estimation from finite number of realisations



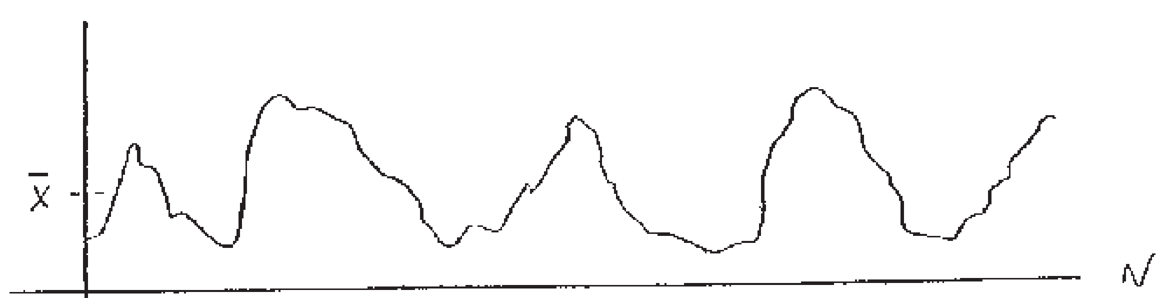
2.8 a: Convergence to true mean

Convergence to true mean



2.8 b: Convergence to incorrect mean (bias)

Convergence to wrong mean



2.8 c: No convergence at all

No convergence

source: Turbulence for the 21st century, W. K. George