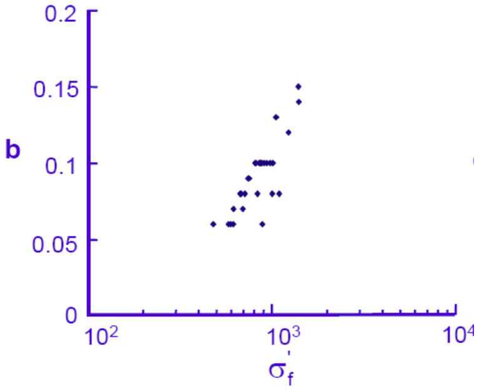
# CHAPTER 2 BASIC PROBABILITY THEORY

## 2.1 Sample Space

For any physical quantity, a set of data can be obtained through physical tests or surveys under a homogeneous condition. The set of all possible outcomes of such tests is called the sample space (or random space) Ω and each individual outcome is a sample point. For the example of fatigue tests, the sample data can be obtained about the physical quantities, as shown in Figure 2.1. The sample space can be described by a probability distribution (or mass) function or histogram.


Strain-life equation: $\frac{\Delta\varepsilon}{2} = \frac{\sigma'_f}{E}\left(2N_f\right)^b + \varepsilon'_f\left(2N_f\right)^c$ : Low-cycle fatigue

**Figure 2.1**: Fatigue Tests and Sample Data Set (Coutesy of Prof. Darrell F. Socie, UIUC, Probabilistic Fatigue, 2005)

## 2.2 Axioms and Theories of Probability

Axiom 1. For any event $E$

$$0 \le P(E) \le 1 \tag{2}$$

where $P(E)$ is the probability of the event $E$.

Axiom 2. Let the sample space be Ω. Then,

$$P(\Omega) = 1 \tag{3}$$

Axiom 3. If $E_1, E_2, \ldots, E_n$ are mutually exclusive events then

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) \tag{4}$$

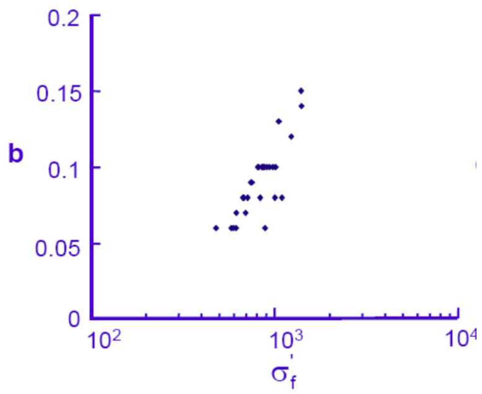
Some Useful Theorems

$$
\begin{aligned}
&P(\bar{E}) = 1 - P(E) \\
&P(\emptyset) = 0 \\
&P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) \\
&P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} : \quad \text{Conditional Probability} \\
&P(E_1 \cap E_2) = P(E_1)P(E_2) \quad \text{if } E_1 \text{ and } E_2 \text{ are independent.}
\end{aligned} \tag{5}
$$

## 2.3 Random Variables

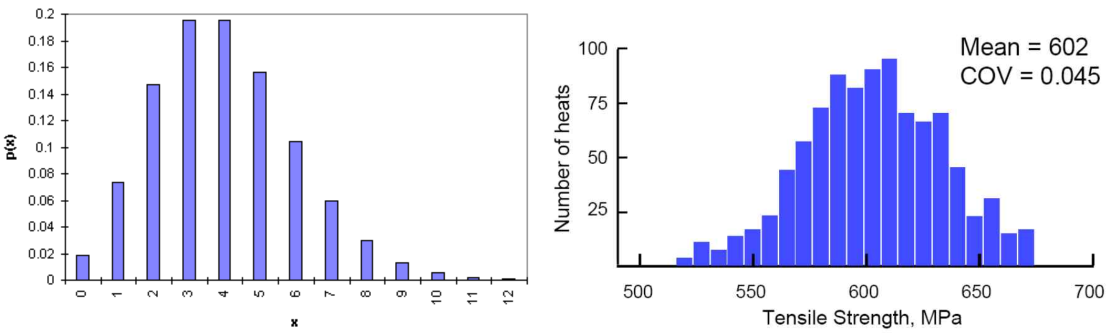
A random variable is a function which maps events in the sample space Ω into the real value $R$ where the outcomes of the event can be real or integer, continuous or discrete, success or fail, etc. The random variable is often denoted as $X: E \rightarrow R$.




The outcome of an event need not be a number, for example, the outcome of coin toss can be either "head" or "tail". However, we often want to represent outcomes as numbers.

- Discrete random variable (Fig. 2.2a):
  The outcome of an experiment is discrete. For example, specimen tensile tests with 10 kN are conducted one hundred times. Each tensile test employs 20 specimens. Let say, the random variable $X$ is the number of failed specimens in each tensile test. Then, $X$ is a discrete random variable.

- Continuous random variable (Fig. 2.2b):
  The outcome of an experiment is continuous. For example, an LED light bulb is tested until it burns out. The random variable $X$ is its lifetime in hours. $X$ can take any positive real value, so $X$ is a continuous random variable. Similar examples include the tensile strength of specimen tensile tests.




( a ) Discrete Random Variable ( b ) Continuous Random Variable

**Figure 2.2:** Random Variable

## 2.4 Univariate Distributions

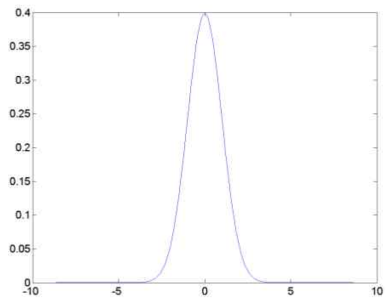
- To understand probability distributions relevant to engineering applications
- To investigate statistical properties of probability distributions
- To make use of Matlab statistical toolbox

Let $X$ be a random variable in an engineering application. The probability density function (PDF) and cumulative distribution function (CDF) of $X$ are denoted by $f_X$ and $F_X$, respectively. Their relationship is $f_X(x) = \frac{\partial}{\partial X} F_X(x)$.

*Normal Distribution (or Gaussian Distribution)*

$$y(x;\mu,\sigma) = f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

```
>> x=[-10:0.1:10];
>> y=normpdf(x,0,1);
>> plot(x,y)
```

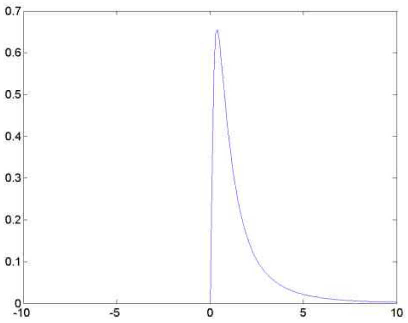


- ✓ Symmetric distribution, skewness=0, kurtosis=3
- ✓ Central limit theorem states that any distribution with finite mean and standard deviation tends to follow normal distribution
- ✓ Special case of chi-squared distribution and gamma distribution
- ✓ Dimension of fabricated part
- ✓ Uncontrolled random quantities (i.e., White Gaussian noise)

*Lognormal Distribution*

$$y(x;\mu,\sigma) = f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

```
>> x=[-10:0.1:10];
>> y=lognpdf(x,0,1);
>> plot(x,y)
```

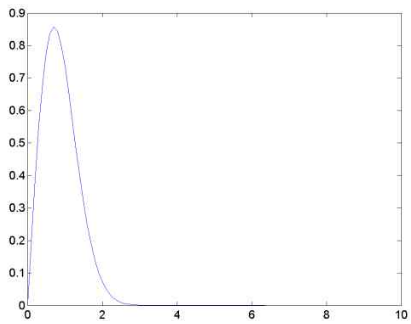



- ✓ Limited to a finite value at the lower limit
- ✓ Positively skewed
- ✓ Strengths of materials, fracture toughness

## *Weibull Distribution*

$$y(x;v,k,a) = f_X(x) = \frac{k-a}{v-a}\left(\frac{x-a}{v-a}\right)^{k-1} e^{-\left(\frac{x-a}{v-a}\right)^k}, \quad \text{2 parameter Weibull if } a=0$$

```
>> x=[0:0.1:10];
>> y=wblpdf(x,1,2);
>> plot(x,y)
```

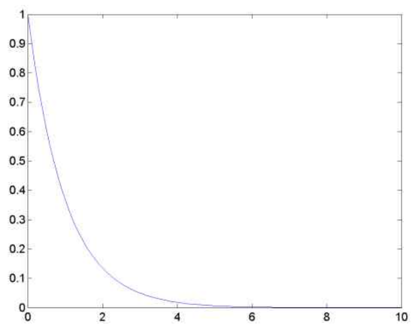



- ✓ $k$ is a shape parameter; $v$ is a scale parameter; $a$ is a location parameter
- ✓ Originally proposed for fatigue life
- ✓ Used in analysis of systems with weakest link
- ✓ Wear, fatigue, and fracture

## *Exponential Distribution*

$$y(x;\mu) = f_X(x) = \frac{1}{\mu} e^{-\frac{x-a}{\mu}}$$

```
>> x=[0:0.1:10];
>> y=exppdf(x,1);
>> plot(x,y)
```

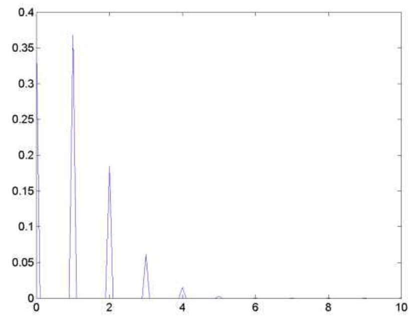



- ✓ $a$ is a location parameter; $\mu$ is a scale parameter
- ✓ Used to model data for time between failures with a constant failure rate
- ✓ Called as "memoryless random distribution"
- ✓ Continuous version of Poisson distribution to describe the number of occurrences per unit time

## *Poisson Distribution (Discrete)*

$$y(x;\lambda) = f_X(x) = \frac{\lambda^x}{x!}e^{-\lambda}$$

```
>> x=[0:0.1:10];
>> y=poisspdf(x,1);
>> plot(x,y)
```

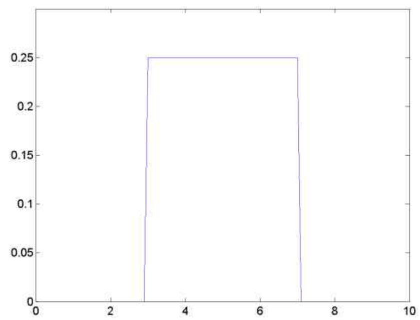



- ✓ An event occurrence in a given interval
- ✓ The occurrences are independent; called as "memoryless random distribution"
- ✓ Used to model data for the number of failed specimens (or product defects) in a given batch with a constant failure rate

## *Uniform Distribution*

$$y(x;a,b) = f_X(x) = \frac{1}{b-a}$$

```
>> x=[0:0.1:10];
>> y=unifpdf(x,3,7);
>> plot(x,y)
```

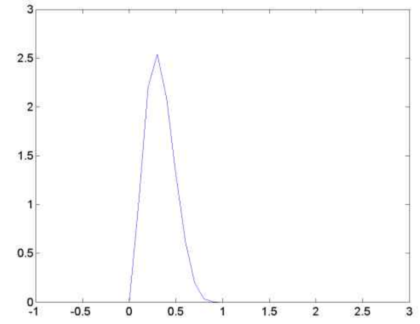



- ✓ Symmetric, skewness=0
- ✓ Equal occurrence
- ✓ Random number generator

## *Beta Distribution*

$$y(x;a,b) = f_X(x) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}, \quad B(a,b):\ \text{Beta function}$$

```
>> x=[-10:0.1:10];
>> y=betapdf(x,3,6);
>> plot(x,y)
```




- ✓ Bounded distributions
- ✓ Related to Gamma distribution
- ✓ Manufacturing tolerance
- ✓ Reliability data in a Bayesian model

*Other Distributions in Engineering*

Rayleigh distribution, Gamma distribution, Extreme Type I, II distributions, etc. Refer to http://mathworld.wolfram.com/topics/ProbabilityandStatistics.html and http://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm.

**Homework 5.1:** ***Statistical uncertainty***

Device a way to quantify the amount of statistical uncertainty. Use $n$=10, 100, 1000 for $X$~N(0,$1^2$).

**Homework 5.2**: ***Statistical modeling of material strengths***

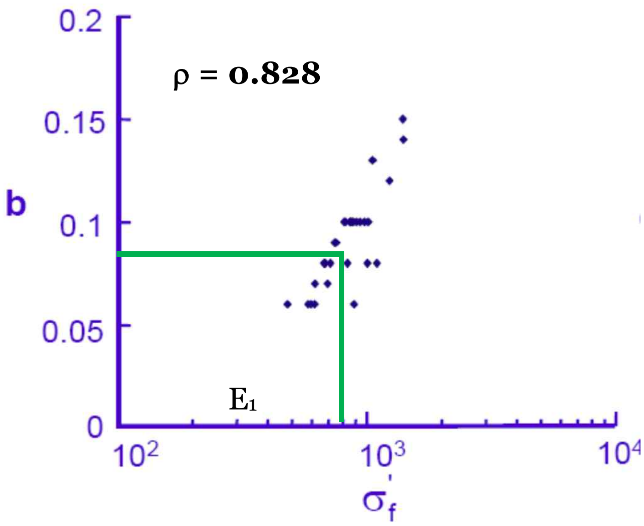
Download the excel file named 'tensile_test.xlsx' at the ETL. You can find the yield strength and tensile strength data from uniaxial tensile tests. Among the probability distribution functions listed above, you are asked to determine two best candidates to model the yield strength and tensile strength. DO NOT use any advanced techniques but rely on the basic analysis of the distribution types described above. Write your essay with at least 150 words.

### 2.5 Random Vectors (Material properties, etc.) – Statistical correlation (related to random vectors)

Suppose $X_1$ and $X_2$ are jointly distributed and joint event is defined as $X_1 \le x_1$ and $X_2 \le x_2$. The corresponding bi-variate distribution of a random vector is defined as

$$\text{Joint CDF: } F_{X_1X_2}(x_1,x_2) = P(X_1 \le x_1, X_2 \le x_2)$$

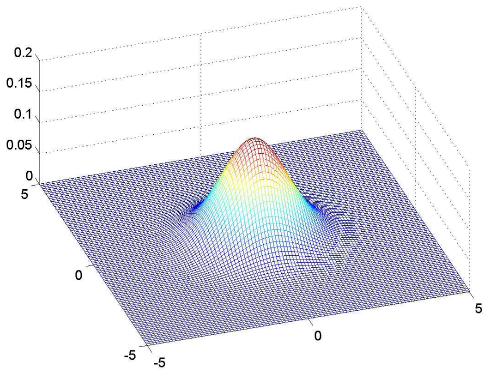
$$\text{Joint PDF: } f_{X_1X_2}(x_1,x_2) = \frac{\partial^2}{\partial x_1 \partial x_2} F_{X_1X_2}(x_1,x_2) \tag{6}$$



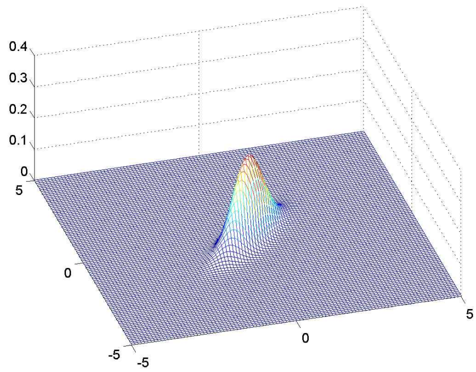
Assume that two random variables are normally distributed. To define the joint PDF of a multivariate distribution, five parameters are required, namely, the mean values of $X_1$ and $X_2$, $\mu_{X_1}$ and $\mu_{X_2}$, their standard deviations $\sigma_{X_1}$ and $\sigma_{X_2}$, and the correlation coefficient $\rho_{X_1X_2}$. The PDF of the bivariate normal distribution can be expressed as

$$f_{X_1X_2}(x_1,x_2)=\frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho_{X_1X_2}^2}}\exp\left\{-\frac{1}{2(1-\rho_{X_1X_2}^2)}\left[\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)^2 -2\rho_{X_1X_2}\frac{(x_1-\mu_{X_1})(x_2-\mu_{X_2})}{\sigma_{X_1}\sigma_{X_2}}+\left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right)^2\right]\right\} \tag{7}$$

If $X_1$ and $X_2$ are correlated, namely, $\rho_{X_1X_2}\neq 0$, $f_{X_1X_2}(x_1,x_2)$ is not symmetry.




```
>> [x1,x2]=meshgrid(-5:0.1:5);
>> f=1/(2*pi)*exp(-
(x1.^2+(x2).^2)./2);
>> mesh(x1,x2,f)
```




```
>> [x1,x2]=meshgrid(-5:0.1:5);
>> f=1/(2*pi*sqrt(1-0.8^2))*exp(-(x1.^2-
1.6*x1.*x2+x2.^2) /(2*(1-0.8^2)^2));
>> mesh(x1,x2,f)
```

Bivariate distribution of random vector can be generalized for $n$-dimensional random vector, $\mathbf{X}:\Omega\rightarrow R^n$. Joint CDF and PDF for $n$-dimensional random vector are written as

$$\text{Joint CDF:}\quad F_{\mathbf{X}}(\mathbf{x})=P\left(\bigcap_{i=1}^{n}\{X_i\le x_i\}\right)$$

$$\text{Joint PDF:}\quad f_{\mathbf{X}}(\mathbf{x})=\frac{\partial^n}{\partial x_1\cdots\partial x_n}F_{\mathbf{X}}(\mathbf{x}) \tag{8}$$

A multi-variate normal random vector is distributed as

$$f_{\mathbf{X}}(\mathbf{x})=(2\pi)^{-\frac{n}{2}}\left|\mathbf{\Sigma_X}\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{X})^T\mathbf{\Sigma_X}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{X})\right] \tag{9}$$

where $\boldsymbol{\mu}_\mathbf{X}$ and $\mathbf{\Sigma_X}$ are mean and covariance matrix of $\mathbf{X}$.

## 2.6 Conditional Probability – Statistical dependence (related to joint events)

The probability of the event $E_1$ occurrence conditional upon the event $E_2$ occurrence is defined as:

$$P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} \tag{10}$$

Let us recall the example of fatigue tests. The sample data can be obtained about the physical quantities in the damage model below.

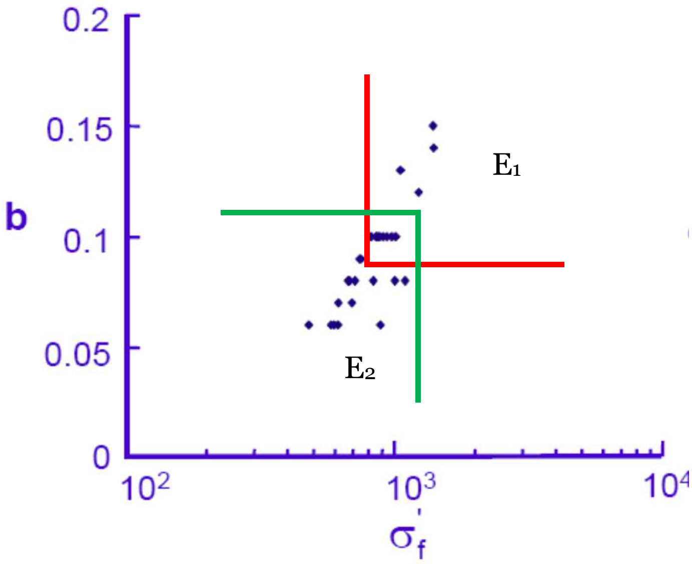
$$\frac{\Delta\varepsilon}{2} = \frac{\sigma_f'}{E}\left(2N_f\right)^b + \varepsilon_f'\left(2N_f\right)^c$$

**Exercise**: Let us consider a 20 data set for the fatigue strength coefficient ($\sigma_f'$) and exponent ($b$) used in the strain-life formula shown above. Two events are defined as

$$E_1 = \{(X_1, X_2) \mid X_1 > 8\times10^2 \text{ and } X_2 > 0.09\}$$
$$E_2 = \{(X_1, X_2) \mid X_1 < 1.02\times10^3 \text{ and } X_2 < 0.11\}$$

$P(E_1) = 8/20 = 2/5$, $P(E_2) = 16/20 = 4/5$, $P(E_1 \cap E_2) = 4/20 = 1/5$

$$P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{1/5}{4/5} = \frac{1}{4}$$
$$P(E_2 \mid E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)} = \frac{1/5}{2/5} = \frac{1}{2}$$




Bayesian statistics (or inference) is based on the conditional probability. It will be recalled in the Bayesian probability theory.

## 2.7 Statistical Moments – Quantification of randomness

- To understand the statistical moments of a random variable.
- To apply statistical moments to an uncertain response.
- To prepare uncertainty propagation analysis through a system in Sections 4 & 5.

Let $\mathbf{X} = \{X_1, \cdots, X_n\}^T$ be an $n$-dimensional random vector and $g(\mathbf{X})$ be a function of $\mathbf{X}$. In general, the $N^{\text{th}}$ statistical moment of $g(\mathbf{X})$ is defined as

$$E[g(\mathbf{X})]^N \equiv \int_\Omega g^N(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \tag{11}$$

where $f_{\mathbf{X}}(\mathbf{x})$ is the joint PDF of $\mathbf{X}$ and $\Omega$ is a sample (or random) space.

<u>2.7.1 Statistical Moments of a Random Vector</u>

First, one special case is considered to find out statistical moments of an input random variable, that is, $g(\mathbf{X}) = X_i, i = 1, \cdots, n$.

*<u>Mean of a Random Vector</u>*

Let $g(\mathbf{X}) = X_1$ and set $N$=1. The first moment of random variable $X_1$ is defined as

$$\begin{aligned} E[X_1]^1 &\equiv \int_{-\infty}^{\infty} x_1 f_{X_1}(x_1) dx_1 \\ &= \mu_{X_1} \end{aligned} \tag{12}$$

Similarly,

$$E[X_2]^1 \equiv \int_{-\infty}^{\infty} x_2 f_{X_2}(x_2) dx_2 = \mu_{X_2}$$

$$\vdots$$

$$E[X_n]^1 \equiv \int_{-\infty}^{\infty} x_n f_{X_n}(x_n) dx_n = \mu_{X_n}$$

$$\boldsymbol{\mu}_{\mathbf{X}} = \left\{ \mu_{X_1} \quad \cdots \quad \mu_{X_n} \right\}^T$$

<u>2.7.2 Covariance of a Random Vector</u>

Let $g(\mathbf{X}) = (X_i - \mu_i)(X_j - \mu_j)$. The statistical moment is defined as

$$\begin{aligned} E\left[(X_i - \mu_i)(X_j - \mu_j)\right] &\equiv \int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) f_{X_i X_j}(x_i, x_j) dx_i dx_j \\ &= \left[\Sigma_{ij}\right] = \boldsymbol{\Sigma}_{\mathbf{X}} \end{aligned} \tag{13}$$

where $f_{X_i X_j}(x_i, x_j)$ and $\Sigma_{ij}$ are the joint PDF and the covariance matrix of $X_i$ and $X_j$, respectively.

When $i = j$, the diagonal terms in the covariance matrix are obtained as

$$\begin{aligned} E\left[X_1-\mu_1\right]^2 &\equiv \int_{-\infty}^{\infty}(x_1-\mu_1)^2 f_{X_1}(x_1)dx_1 \\ &= \sigma_{X_1}^2 = \Sigma_{11} \\ &\vdots \\ E\left[X_n-\mu_n\right]^2 &\equiv \int_{-\infty}^{\infty}(x_n-\mu_n)^2 f_{X_n}(x_n)dx_n \\ &= \sigma_{X_n}^2 = \Sigma_{nn} \end{aligned} \tag{14}$$

If $i \neq j$, the off-diagonal terms in the covariance matrix are obtained as

$$\begin{aligned} E\left[(X_1-\mu_1)(X_2-\mu_2)\right] &\equiv \int_{-\infty}^{\infty}(x_1-\mu_1)(x_2-\mu_2) f_{X_1X_2}(x_1,x_2)dx_1dx_2 \\ &= \Sigma_{12} \\ &\vdots \\ E\left[(X_n-\mu_n)(X_{n-1}-\mu_{n-1})\right] &\equiv \int_{-\infty}^{\infty}(x_n-\mu_n)(x_{n-1}-\mu_{n-1}) f_{X_nX_{n-1}}(x_n,x_{n-1})dx_ndx_{n-1} \\ &= \Sigma_{nn-1} \end{aligned} \tag{15}$$

The covariance matrix is written as

$$\mathbf{\Sigma_X} = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1} & \cdots & \Sigma_{nn} \end{bmatrix}$$

### 2.7.2 Higher moments

Skewness and Kurtosis (3rd and 4th order moments)

$$\text{skewness} = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \quad \text{or} \quad \text{skewness} = \frac{\sum_{i=1}^{N}\left(x_i-\bar{X}\right)^3}{(N-1)s^3} \tag{16}$$

$$\text{kurtosis} = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] \quad \text{or} \quad \text{kurtosis} = \frac{\sum_{i=1}^{N}\left(x_i-\bar{X}\right)^4}{(N-1)s^4} \tag{17}$$

### 2.7.3 Properties of Covariance Matrix, $\mathbf{\Sigma_X}$

- $\mathbf{\Sigma_X}$ is symmetric, i.e., $\mathbf{\Sigma_X} = \mathbf{\Sigma_X}^T$
- Variance of $X_i$ is the $i$th diagonal element of $\mathbf{\Sigma_X}$, i.e., $\sigma_{X_i}^2 = \Sigma_{ii}$
- $\mathbf{\Sigma_X}$ is a positive semi-definite matrix, i.e., $\mathbf{A}^T\mathbf{\Sigma_X}\mathbf{A} \geq 0, \quad \forall \mathbf{A} \in R^n$

2.7.4 Correlation Coefficient, $\rho_{ij}$

The correlation coefficient $\rho_{ij}$ is defined as

$$\rho_{ij} \equiv \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}}\sqrt{\Sigma_{jj}}} = \frac{\Sigma_{ij}}{\sigma_i \sigma_j} \tag{18}$$

The correlation coefficient $\rho_{ij}$ is a degree of correlation between two random variables. Note that $\Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$ represents the off-diagonal elements of covariance matrix, $\mathbf{\Sigma_X}$.

- If $X_i$ and $X_j$ are independent (*i.e.*, $f_{X_1X_2}(x_1,x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$), then $X_i$ and $X_j$ are uncorreleated (*i.e.* $\rho_{ij} = 0$), but vice versa is not true.
- $-1 \le \rho_{ij} \le +1$
- If $X_j = aX_i + b$, $\rho_{ij} = \pm 1 = \mathrm{sgn}(a)$.

2.7.5 Coefficient of Variation, COV(X) = $\sigma_X / \mu_X$

**Homework 6**: *Statistical moments and joint PDF*
Use the same excel file named 'tensile_test.xlsx' at the ETL. Calculate the sampled means and standard deviations of yield strength and tensile strength. With the calculated means, standard deviations, and correlation coefficient, you can plot a joint pdf of yield strength and tensile strength. ASSUME the yield strength and tensile strength follow normal distribution.

**Homework 7**: *Read Chapter 2 of the Textbook*
Read Chapter 2 to reinforce your knowledge about the fundamentals of the engineering statistics.