Summer Seminar: Philosophy of Statistics
Lecture Notes 3: The Concept of a Probability Model

Aris Spanos [Summer 2019]

1 Introduction

1.1 The story so far and what comes next

In chapter 2 we commenced the long journey to explore the theory of probability as a modeling framework. In an effort to motivate the various concepts needed, the discussion began with the formalization of the notion of a **random experiment** $\mathcal{E}$, defined by the conditions in table 3.1.

Table 3.1: Random Experiment ($\mathcal{E}$)

[a]	All possible distinct outcomes are known at the outset.
[b]	In any particular trial the outcome is not known in advance, but there exist discernible regularities pertaining to the frequency of occurrence associated with different outcomes.
[c]	The experiment can be repeated under identical conditions.

The mathematization of $\mathcal{E}$ took the form of a **statistical space** $[(S,\Im,\mathbb{P}(.))^n,\mathcal{G}_n^{\text{IID}}]$ where $(S,\Im,\mathbb{P}(.))$ is a *probability space* and $\mathcal{G}_n^{\text{IID}}$ is a *simple sampling space*. Unfortunately, the statistical space does not lend itself naturally or conveniently to the modeling of stochastic phenomena that give rise to numerical data.

The main purpose of this chapter is to map the abstract probability space $(S,\Im,\mathbb{P}(.))$ onto the real line where observed data live. The end result will be a reformulation of $(S,\Im,\mathbb{P}(.))$ into a **probability model**, one of the two pillars of a statistical model; the other being mapping of $\mathcal{G}_n^{\text{IID}}$ onto the real line to define a sampling model in chapter 4.

A bird's eye view of the chapter. The key to mapping the statistical space onto the real line ($\mathbb{R}$):

$$[(S,\Im,\mathbb{P}(.)),\mathcal{G}_n^{\text{IID}}]\longrightarrow\mathbb{R}, \tag{1}$$

is the concept of a **random variable**, one of the most crucial concepts of probability theory. Let us briefly unpack that mapping.

Step 1. Define the mapping we call a random variable:

$$X(.)\colon S\to\mathbb{R},$$

such that that $X(.)$ **preserves the event structure of interest** $\Im$.

Step 2. Armed with the mapping $X(.)$, the probability set-function:

$$\mathbb{P}(.)\colon \Im\to[0,1],$$

is metamorphosed into a point-to-point numerical function, the *cumulative distribution function* (cdf), defined in terms of X:

$$F_X(.)\colon \mathbb{R} \to [0,1].$$

Step 3. $F_X(.)$ is simplified by transforming it into the *density function*:

$$f_x(.)\colon \mathbb{R} \to [0,\infty).$$

The concept of a probability model is usually defined in terms of the density function.

2 The concept of a random variable

To help the reader understand the **concept of the random variable** and how it transforms the abstract statistical space $(S,\Im,\mathbb{P}(.))$ into something much easier to handle, a probability model $\Phi=\{f(x;\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta,\ x\in\mathbb{R}_X\}$, the discussion begins with the simplest case and then proceeds to consider the more complicated ones:

(i) **Finite:** the outcomes S is finite,

(ii) **Infinite/countable:** the outcomes set S is infinite but countable,

(iii) **Infinite/uncountable:** the outcomes set S is infinite and uncountable.

2.1 The case of a finite outcomes set: $S=\{s_1, s_2, ..., s_n\}$

A (discrete) random variable with respect to the event space $\Im$, is defined to be a real-valued function of the form:

$$X(.)\colon\ S \to \mathbb{R}_X, \tag{2}$$

such that all the sets defined by $\{s\colon\ X(s)=x\}$ for $x\in\mathbb{R}$ constitute events in $\Im$, denoted by:

$$A_x:=\{s\colon\ X(s)=x\}\in\Im,\ \forall x\in\mathbb{R}. \tag{3}$$

or equivalently:

$$A_x=X^{-1}(x)\in\Im,\ \forall x\in\mathbb{R},$$

where $X^{-}(.)$ denotes the pre-image of $X(.)\colon\ S \to \mathbb{R}_X$. NOTE that the pre-image does not often coincide with the inverse of the function! All functions have a pre-image, but only one-to-one functions have an inverse.

Intuitively, a random variable is a function which attaches real numbers to all the elements of S in a way which **preserves the event structure of interest** $\Im$.

Example 3.1. Consider the random experiment of "tossing a coin twice":

$$S=\{(HH),(HT),(TH),(TT)\},$$

where the event space of interest is:

$$\begin{gathered}\Im=\{S,\varnothing,A,B,C,A\cup B,A\cup C,B\cup C\},\\ A=\{(HH)\},\ B=\{(TT)\},\ C=\{(HT),(TH)\}.\end{gathered} \tag{4}$$

Let us consider the question whether the following two mappings constitute random variables relative to $\Im$ in (4):

$$X(HH)=2, \quad X(TT)=0, \quad X(HT)=X(TH)=1,$$
$$Y(HT)=1, \quad Y(TH)=2, \quad Y(HH)=Y(TT)=0.$$

For an affirmative answer the *pre-image* for each of the values in their range defines an event in $\Im$. The pre-image is found by tracing the elements of S associated with each of the values in the mappings range (figure 3.1), and determining whether the resulting subset of S belongs to $\Im$ or not.

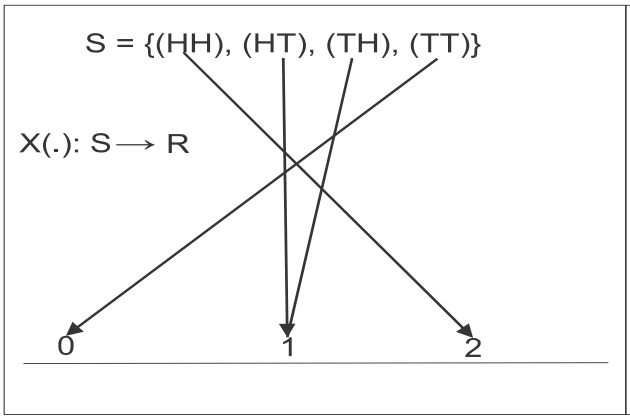

Fig. 3.1: Random variable X

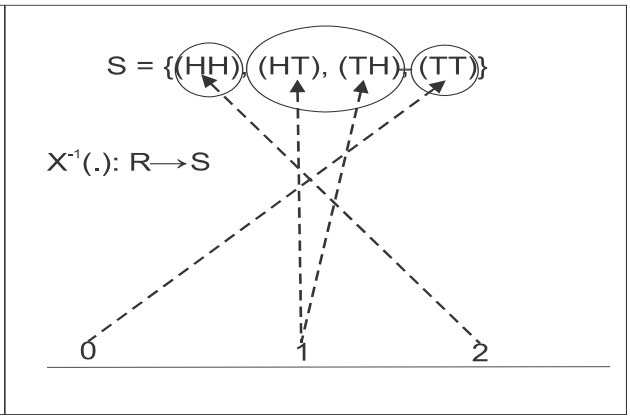

Fig. 3.2: Pre-image of random variable X

For $X(.)$ the pre-images associated with its values in $\mathbb{R}_X$:=$(0,1,2)$ yield:

$$X^{-1}(0)=\{(TT)\}=A\in\Im, \; X^{-1}(1)=\{(HT),(TH)\}=C\in\Im, \; X^{-1}(2)=\{(HH)\}=B\in\Im,$$

and since all of them belong to $\Im$, X is a random variable relative to it.

For $Y(.)$ the pre-images for the values in $\mathbb{R}_Y$:=$(0,1,2)$ yield:

$$Y^{-1}(0)=\{(HH),(TT)\}=A\cup B\in\Im, \; Y^{-1}(1)=\{(HT)\}\notin\Im, \; Y^{-1}(2)=\{(TH)\}\notin\Im. \qquad (5)$$

Hence, Y is not a random variable relative to the above $\Im$, since two of those pre-images are not events in $\Im$!

2.2 Key features of a random variable

First, the term 'random variable' is a misnomer since, in light of its definition in (3), $X(.)$ is just a real-valued function that involves no probabilities, i.e. it **is neither random nor a variable**.

Second, the concept of a random variable is always defined **relative to an event space** $\Im$, and whether or not $X(.)$ satisfies the restriction in (3) depends on $\Im$, not on $\mathbb{P}(.)$. The fact that a certain real-valued function is not a random variable with respect to a particular $\Im$ does not mean that it cannot be a random variable with

respect to some other event space. Indeed, one can define an event space, say $\Im_Y$, with respect to which Y constitutes a random variable.

Example 3.2. To see that, let us return to the pre-images of $Y^{-1}(y),\ y{=}0,1,2$ in (5) and define the events:

$$A_1{:=}Y^{-1}(0){=}\{(HH),(TT)\},\ A_2{:=}Y^{-1}(1){=}\{(HT)\},\ A_3{:=}Y^{-1}(2){=}\{(TH)\},$$

and use (A_1,A_2,A_3) to generate a field:

$$\Im_Y{:=}\sigma(Y){=}\{S,\varnothing,A_1,A_2,A_3,A_1\cup A_2,A_1\cup A_3,A_2\cup A_3\}.$$

$\Im_Y{:=}\sigma(Y)$ is known as *the minimal field* generated by the random variable Y. This concept gives rise to an alternative but equivalent definition for a random variable that generalizes directly to the case where S is unaccountable.

Random variable. The real-valued function $X(.)\colon\ S\to\mathbb{R}_X$, is said to be a *random variable* with respect to $\Im$, if the *σ-field generated by X is a subset of* $\Im$, i.e. $\sigma(X)\subseteq\Im$.

In example 3.1 $\sigma(X){=}\Im$ (verify), but in general it can be a proper subset of $\Im$.

Example 3.3. The real-valued function:

$$Z(HT){=}Z(TH){=}0,\quad Z(HH){=}Z(TT){=}1,$$

is a random variable relative to $\Im$ since $\sigma(Z){=}\{S,\varnothing,C,\overline{C}\}{\subset}\Im$ where:

$$C{:=}\{s\colon\ Z{=}0\}{:=}Z^{-1}(0){=}\{(HT),(TH)\},\ \overline{C}{:=}\{s\colon\ Z{=}1\}{:=}Z^{-1}(1){=}\{(HH),(TT)\}.$$

Equivalently, Z is a random variable relative to $\Im$ since $\sigma(Z){=}\{S,\varnothing,C,\overline{C}\}{\subset}\Im$. In light of that, when would one prefer to use Z instead of X? Z will be the random variable of choice when the event of interest is C 'one of each' and its complement $\overline{C}$ 'two of the same'. Hence, a real-valued function of the form $X(.)\colon\ S\to\mathbb{R}_X$ is a random variable relative to a particular $\Im$ when $\sigma(X){\subset}\Im$.

Third, since $\mathbb{R}_X{\subset}\mathbb{R}$ one might wonder why the restriction in (3) is in terms of $x{\in}\mathbb{R}$ and not $x{\in}\mathbb{R}_X$. It turns out that all the points $\overline{\mathbb{R}}_X{:=}(\mathbb{R}{-}\mathbb{R}_X)$ have the empty set $\varnothing$ as their pre-image, and since $\varnothing$ belongs to all event spaces (being a σ-fields):

$$X^{-1}(x){:=}\{s\colon\ X(s){=}x\}{=}\varnothing{\in}\Im,\ \text{for all } x{\in}\overline{\mathbb{R}}_X{:=}(\mathbb{R}{-}\mathbb{R}_X)\,.$$

Intuitively, a mapping $X(.)\colon\ S\to\mathbb{R}_X$ is a random variable when it preserves the event structure of a particular event space $\Im$, by ensuring that its pre-image takes the form:

$$X^{-1}(.)\colon\ \mathbb{R}\to\Im,\tag{6}$$

where for each $x{\in}\mathbb{R}_X,\ X^{-1}(x){\in}\Im$, and $X^{-1}(x){=}\varnothing{\in}\Im$ for all $x{\notin}\mathbb{R}_X$.

Example 3.4. Consider the case $\Im_0{=}\{S,\varnothing\}$ where the only $X(.)\colon\ S\to\mathbb{R}$ that is a random variable relative to $\Im_0$ is $X(s){=}c{\in}\mathbb{R}$, for all $s{\in}S$, which defines a constant is a *degenerate* random variable.

Example 3.5. For the slightly more informative case, consider $\Im=\{S,\varnothing,A,\overline{A}\}$ and the random variable can take two values, say:

$$\{s\colon\ X(s)=1\}:=A,\ \{s\colon\ X(s)=0\}:=\overline{A}.$$

The resulting random variable is the indicator function $\mathbb{I}_A(s)=\begin{cases}1, & s\in A,\\ 0, & s\in\overline{A}.\end{cases}$, that is **Bernoulli distributed**.

2.2.1 Assigning probabilities

Using the concept of a random variable we mapped S (an arbitrary set) to a subset of the real line (a set of numbers) $\mathbb{R}_X$. Because we do not want to change the original probability structure of $(S,\Im,\mathbb{P}(.))$ we imposed condition (3) to ensure that all events defined in terms of the random variable X belong to the original event space $\Im$. We also want to ensure that the same events in the original probability space (S, $\Im$, $\mathbb{P}(.)$) and the new formulation, such as $A_x=\{s\colon\ X(s)=x\}$, get assigned the same probabilities. In order to ensure that, we define the point function $f_x(.)$, which we call a **density function** as follows:

$$f_x(x):=\mathbb{P}(X=x)\text{ for all }x\in\mathbb{R}_X. \tag{7}$$

NOTE that $(X=x)$ is a shorthand notation for $\{s\colon\ X(s)=x\}$. Clearly, for $x\notin\mathbb{R}_X$, $X^{-1}(x)=\varnothing$, and thus $f_x(x)=0$, for all $x\notin\mathbb{R}_X$.

Example 3.6. In the case of the indicator function if we let $X(s):=I_A(s)$ we can define the probability density as follows:

$$f_x(1):=\mathbb{P}(X=1)=\theta\text{ and }\ f_x(0):=\mathbb{P}(X=0)=(1-\theta),$$

where $0\leq\theta\leq 1$. This is known as the **Bernoulli density**:

x	0	1
$f_x(x)$	$(1-\theta)$	θ

(8)

► **What have we gained?** In the context of the original probability space (S, $\Im$, $\mathbb{P}(.)$), where $S=\{s_1,s_2,...,s_n\}$, the probabilistic structure of the experiment was specified in terms of:

$$\mathbf{P}:=\{p(s_1),p(s_2),...,p(s_n)\},\ \text{such that}\ \textstyle\sum_{i=1}^{n}p(s_i)=1.$$

Armed with this we could assign a probability of any event $A\in\Im$ as follows. We know that all events $A\in\Im$ are just unions of certain outcomes. Given that outcomes are also mutually exclusive elementary events, we proceed to use axiom 3 (see chapter 2) to define the probability of A as equal to the sum of the probabilities assigned to each of the outcomes making up the event A, i.e. if $A=\{s_1,s_2,...,s_k\}$, then:

$$\mathbb{P}(A)=\textstyle\sum_{i=1}^{k}p(s_i).$$

Example 3.7. In the case of the random experiment of "tossing a coin twice":

$$S=\{(HH),(HT),(TH),(TT)\},\ \Im=\mathcal{P}(S),$$

where $\mathcal{P}(S)$ denotes the *power set* of S: the set of all subsets of S (see chapter 2). The random variable of interest is defined by: X- the number of "Heads". This suggest that the events of interest are:

$$A_0=\{s\colon\ X=0\}=\{(TT)\},\ A_1=\{s\colon\ X=1\}=\{(HT),(TH)\},\ A_2=\{s\colon\ X=2\}=\{(HH)\}.$$

In the case of a fair coin all four outcomes are given the same probability and thus:

$$\begin{aligned}&\mathbb{P}(A_0)=\mathbb{P}\{s\colon\ X=0\}=\mathbb{P}\{(TT)\}=\tfrac{1}{4},\\&\mathbb{P}(A_1)=\mathbb{P}\{s\colon\ X=1\}=\mathbb{P}\{(HT),(TH)\}=\mathbb{P}(HT)+\mathbb{P}(TH)=\tfrac{1}{2},\\&\mathbb{P}(A_2)=\mathbb{P}\{s\colon\ X=2\}=\mathbb{P}\{(HH)\}=\tfrac{1}{4}.\end{aligned}$$

Returning to the main focus of this chapter, we can claim that using the concept of a random variable we achieved the following mapping:

$$(S,\Im,\mathbb{P}(.))\overset{X(.)}{\rightarrow}(\mathbb{R}_X,f_x(.)),$$

where the original probabilistic structure has been transformed into:

$$\{f_x(x_1),f_x(x_2),...,f_x(x_m)\},\ \text{such that}\ \textstyle\sum_{i=1}^{m}f_x(x_i)=1,\ m\leq n;$$

the last is referred to as *the probability distribution* of a random variable X.

The question which arises at this point is to what extend the latter description of the probabilistic structure is preferable to the former. At first sight it looks as though no mileage has been gained by this transformation. However, it turns out that this is misleading and a lot of mileage has been gained for two reasons.

(a) Instead of having to specify $\{f_x(x_1),f_x(x_2),...,f_x(x_m)\}$ explicitly, we can use simple real valued functions in the form of formulae such as:

$$f_x(x;\theta)=\theta^x(1-\theta)^{1-x},\ x=0,1,\ \text{and}\ 0\leq\theta\leq1, \tag{9}$$

which specify the distribution implicitly. For each value of X the function $f_x(x)$ specifies its probability. This formula constitutes a more compact way of specifying the distribution given above.

(b) Using such density functions there is no need to know the probabilities associated with the events of interest in advance. In the case of the above formula, θ could be unknown and the set of such density functions is referred to as *a family of density functions* indexed by θ. This is particularly important for modeling purposes where such families provide the basis of probability models. In a sense, the uncertainty relating to the outcome of a particular trial (condition [c] defining a Random

Experiment) has become the uncertainty concerning the "true" value of the unknown parameter θ.

The distribution defined by (9) is known as the *Bernoulli distribution*. This distribution can be used to describe random experiments with only two outcomes.

Example 3.8. In the case of the random experiment of "tossing a coin twice":

$$S=\{(HH),(HT),(TH),(TT)\}, \Im=\{S,\varnothing,A,\overline{A}\},$$

where the event of interest is $A=\{(HH),(HT),(TH)\}$, with $\mathbb{P}(A)=\theta$, $\mathbb{P}(\overline{A})=(1-\theta)$. By defining the random variable $X(A)=1$ and $X(\overline{A})=0$, the probabilistic structure of the experiment is described by the Bernoulli density (9).

This type of random experiment can be easily extended to n repetitions of the same two-outcomes experiment, giving rise to the so-called *Binomial* distribution.

Example 3.9. Consider the random experiment of "tossing a coin n times and counting the number of Heads". The outcomes set for this experiment is defined by $S=\{H,T\}^n$ with $\mathbb{P}(H)=\theta$ and $\mathbb{P}(T)=1-\theta$. Define the random variable:

$$X\text{: the total number of } H\text{'s in } n \text{ trials.}$$

The range of values of X is $\mathbb{R}_X=\{0,1,2,3,...,n\}$, a *Binomially distributed* random variable and a density function:

$$f_x(x;\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x},\ 0\leq x\leq n,\ n=1,2,...,\ 0\leq\theta\leq 1, \tag{10}$$

where $\binom{n}{x}=\frac{n!}{(n-x)!x!}$ and $n!=n\cdot(n-1)\cdot(n-2)\cdot\cdots(3)\cdot(2)\cdot 1$.
This formula stems naturally from combinations rule discussed in chapter 2.

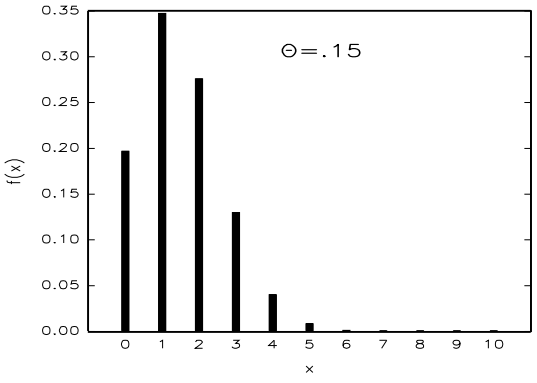

Fig. 3.3: Binomial ($n=10, \theta=.15$)

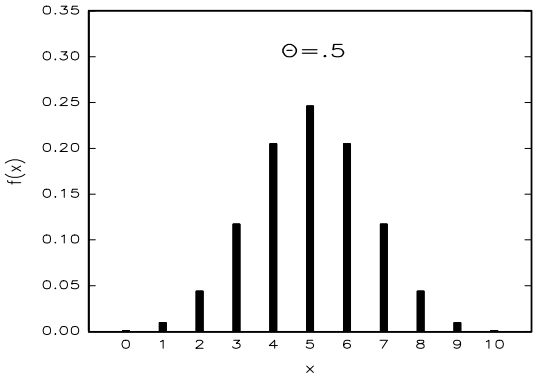

Fig. 3.4: Binomial ($n=10, \theta=.5$)

This formula can be graphed for specific values of θ. In figures 3.3 and 3.4 we can see the graph of the Binomial density function (10) with n=10 and two different values of the unknown parameter, θ=.15 and θ=.5, respectively.

The horizontal axis depicts the values of the random variable X ($\mathbb{R}_X=\{0,1,2,3,...,n\}$) and the vertical axis depicts the values of the corresponding probabilities as shown

below. The gains from this formulation are even more apparent in the case where the outcomes set S is infinite but countable. As shown next, in such a case listing the probabilities for each $s \in S$ in a table is impossible. The assignment of probabilities using a density function, however, is trivial.

2.3 The case of a countable outcomes set: $S=\{s_1, s_2, ..., s_n, ...\}$

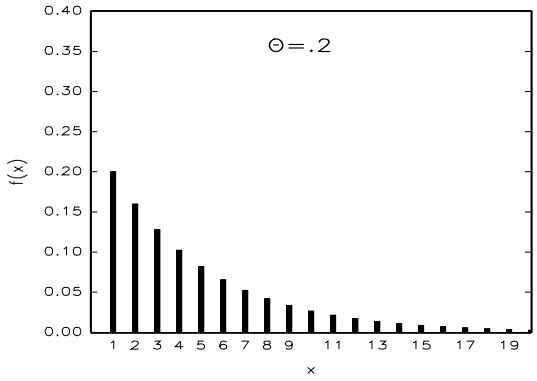

Fig. 3.5: Geometric ($n = 20, \theta = .2$)

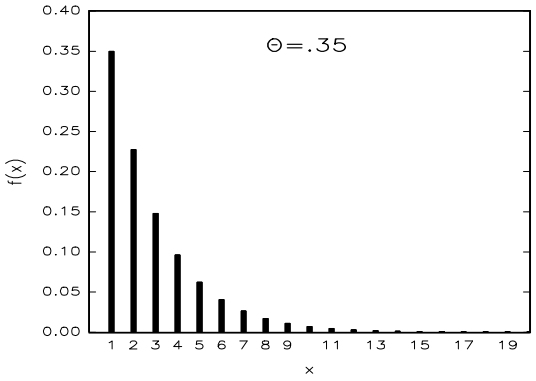

Fig. 3.6: Geometric ($n = 20, \theta = .35$)

Consider the case of the countable outcomes set $S=\{s_1, s_2, ..., s_n, ...\}$. This is a simple extension of the finite outcome set case where the probabilistic structure of the experiment is specified in terms of:

$$\{p(s_1), p(s_2), ..., p(s_n), ...\}, \text{ such that } \sum_{i=1}^{\infty} p(s_i)=1.$$

The probability of an event $A \in \Im$, is equal to the sum of the probabilities assigned to each of the outcomes making up the event A: $\mathbb{P}(A)=\sum_{\{i:s_i \in A\}} p(s_i)$.

Example 3.10. Consider the random experiment of "tossing a coin until the first H turns up". The outcomes set is:

$$S=\{(H), (TH), (TTH), (TTTH), (TTTTH), (TTTTTH), \ldots\},$$

and let the event space be the power set of S. If we define the random variable

$X(.)$- the number of trials needed to get one H, i.e.

$$X(H)=1, \; X(TH)=2, \; X(TTH)=3, \text{ etc.}$$

and $\mathbb{P}(H)=\theta$, then the density function for this experiment is:

$$f_x(x;\theta)=(1-\theta)^{x-1}\theta, \; 0 \leq \theta \leq 1, \; x \in \mathbb{R}_X=\{1, 2, 3, ...\}.$$

This is the density function of the **Geometric distribution**. This density function is graphed in figures 3.5-3.6 for n=20 and two different values of the unknown parameter θ=.20 and θ=.35, respectively. Looking at these graphs we can see why the name Geometric was given to this distribution: the probabilities decline geometrically as the values of X increase.

3 The general concept of a random variable

Having introduced the basic concepts needed for the transformation of the abstract probability space $(S, \Im, \mathbb{P}(.))$ into something more appropriate (and manageable) for modeling purposes, using the simplest case of countable outcomes set, we will now proceed to explain these concepts in their full generality.

3.1 The case of an uncountable outcomes set S

As a prelude to the discussion that follows let us see why the previous strategy of assigning probabilities to each and every outcome in the case of an uncountable set, say S=$\mathbb{R}$, will not work. The reason is very simple: the outcomes set has so many elements that it is impossible to arrange them in a sequence and thus count them. Hence, any attempt to follow the procedure used in the countable outcomes set case will lead to insurmountable difficulties. Intuitively we know that we cannot cover the real line point by point. The only way to overlay $\mathbb{R}$ or any of its uncountable subsets is to use a sequence of intervals of any one of the following forms:

$$(a,b),\ [a,b],\ [a,b),\ (-\infty,a],\ \text{where } a<b,\ a \text{ and } b \text{ real numbers.}$$

We will see in the sequel that the most convenient form for such intervals is:

$$\{(-\infty,x]\} \text{ for each } x\in\mathbb{R}. \tag{11}$$

3.1.1 The general definition of a Random Variable

A random variable relative to $\Im$ is a function $X(.)$: $S \to \mathbb{R}$, that satisfies the restriction:

$$B_x:=\{s:\ X(s)\leq x\}:=X^{-1}\left((-\infty,x]\right)\in\Im \ \text{ for all } \ x\in\mathbb{R}. \tag{12}$$

NOTICE that the only difference between this definition and that of a discrete random variable comes in the form of the events used B_x:=$\{s:\ X(s)\leq x\}$ instead of A_x:=$\{s:\ X(s)$=$x\}$.

Moreover, in view of the fact that:

$$\{s:\ X(s)=x\}\subset\{s:\ X(s)\leq x\},$$

the latter definition includes the former as a special case. From this definition we can see that the pre-image of the random variable $X(.)$ takes us from intervals $(-\infty,x],\ x\in\mathbb{R}$ back to the event space $\Im$. The set of all such intervals generates a σ-field on the real line known as the **Borel-field** and denoted by $\mathcal{B}(\mathbb{R})$:

$$\mathcal{B}(\mathbb{R})=\sigma\left((-\infty,x],\ x\in\mathbb{R}\right).$$

It is worth NOTING that we could have generated $\mathcal{B}(\mathbb{R})$ using any one of the interval forms mentioned above $(a,b),\ [a,b],\ [a,b),\ (-\infty,a]$ and all such intervals are now

elements of the Borel field $\mathcal{B}(\mathbb{R})$. Hence, in a formal sense, the pre-image of the random variable $X(.)$ constitutes a mapping from the Borel-field $\mathcal{B}(\mathbb{R})$ to the event space $\Im$ and takes the form:

$$X^{-1}(.)\colon \mathcal{B}(\mathbb{R}) \to \Im. \tag{13}$$

This ensures that the random variable $X(.)$ preserves the event structure of $\Im$ because the pre-image preserves the set-theoretic operations (see Karr, 1973):

Table 3.2: Pre-image and set theoretic operations

(i)	Union:	$X^{-1}(\bigcup_{i=1}^{\infty} B_i)=\bigcup_{i=1}^{\infty} X^{-1}(B_i)$,
(ii)	Intersection:	$X^{-1}(\bigcap_{i=1}^{\infty} B_i)=\bigcap_{i=1}^{\infty} X^{-1}(B_i)$,
(iii)	Complementation:	$X^{-1}(\overline{B})=(\overline{X^{-1}(B)})$.

3.1.2 The probability space induced by a random variable*

The next step should be to transform $\mathbb{P}(.)\colon \Im \to [0,1]$ into a mapping on the real line or more precisely on $\mathcal{B}(\mathbb{R})$. This transformation of the probability set function takes the form:

$$\mathbb{P}(X \leq x)=\mathbb{P}X^{-1}((-\infty,x])=P_X((-\infty,x]),$$

where the last compositve function has eliminated $\Im$:

$$P_X(.):=\mathbb{P}X^{-1}(.)\colon \mathcal{B}(\mathbb{R}) \to [0,1].$$

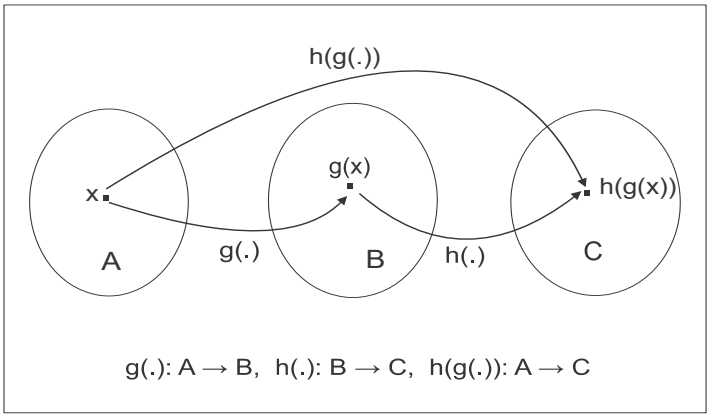

Fig. 3.7: Composite function $h(g(.))\colon A \to C$

In terms of fig. 3.7:

$$g(.)\colon A \to B \text{ corresponds to } X^{-1}(.)\colon \mathcal{B}(\mathbb{R}) \to \Im$$

$$h(.)\colon B \to C, \text{ corresponds to } \mathbb{P}(.)\colon \Im \to [0,1],$$

and A, B and C correspond to $\mathcal{B}(\mathbb{R})$, $\Im$ and $[0,1]$, respectively.

Collecting the above elements together we can see that in effect a random variable X induces a new probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X(.))$ with which we can replace the abstract probability space $(S, \Im, \mathbb{P}(.))$. The main advantage of the former over the latter is that everything takes place on the real line and not in some abstract space. In direct analogy to the countable outcomes set case, the concept of a random variable induces the following mapping:

$$\boxed{(S, \Im, \mathbb{P}(.)) \overset{X(.)}{\rightarrow} (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X(.)).}$$

That is, using the mapping $X(.)$ we traded S for $\mathbb{R}$, $\Im$ for $\mathcal{B}(\mathbb{R})$ and $\mathbb{P}(.)$ for $P_X(.)$. For reference purposes we call $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X(.))$ the probability space *induced* by a random variable X; see Galambos (1995).

Borel (measurable) functions. In probability theory we are interested not just in random variables, but also well-behaved functions of such random variables. By well-behaved functions, in calculus, we usually mean continuous or differentiable functions. In probability theory well-behaved functions refers to ones which preserve the event structure of their argument random variable A function defined by:

$$h(.)\text{: } \mathbb{R} \rightarrow \mathbb{R} \text{ such that } \{h(X) \leq x\} := h^{-1}((-\infty, x]) \in \mathcal{B}(\mathbb{R}), \text{ for all } x \in \mathbb{R},$$

is called a *Borel (measurable) function.* That is, a *Borel function* is a function which is a random variable relative to $\mathcal{B}(\mathbb{R})$. NOTE that indicator functions, monotone functions, continuous functions as well as functions with a finite number of discontinuities are Borel functions; see Khazanie (1976).

Equality of random variables. Random variables are unlike mathematical functions in so far as their probabilistic structure is of paramount importance. Hence, the concept of equality for random variables involves this probabilistic structure. Two random variables X and Y, defined on the same probability space $(S, \Im, \mathbb{P}(.))$, are said to be equal *with probability one* (or *almost surely*) if (see Karr, 1993):

$$\mathbb{P}(s\text{: } X(s) \neq Y(s)) = 0, \text{ for all } s \in S,$$

i.e., if the set $(s\text{: } X(s) \neq Y(s))$ is an event with zero probability.

4 Cumulative distribution and density functions

4.1 The concept of a cumulative distribution function

Using the concept of a random variable $X(.)$, so far we transformed the abstract probability space $(S, \Im, \mathbb{P}(.))$ into a less abstract space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X(.))$. However, we have not reached our target yet because $P_X(.) := \mathbb{P}X^{-1}(.)$ is still a set function. Admittedly it is a much easier set function because it is defined on the real line, but a set function all the same. What we prefer is a numerical point-to-point function.

The way we transform the set-function $P_X(.)$ into a numerical point-to-point function is by a clever stratagem. By viewing $P_X(.)$ as only a function of the end point of the interval $(-\infty, x]$ we define the **cumulative distribution function** (cdf):

$$F_X(.)\colon \mathbb{R} \to [0,1], \text{ defined by } F_X(x) = \mathbb{P}\{s\colon X(s) \leq x\} = P_X((-\infty, x]). \tag{14}$$

The ploy leading to this trick began a few pages ago when we argued that even though we could use any one of the following intervals (see Galambos, 1995):

$$(a,b),\ [a,b],\ [a,b),\ (-\infty, a], \text{ where } a < b,\ a \in \mathbb{R} \text{ and } b \in \mathbb{R},$$

to generate the Borel field $\mathcal{B}(\mathbb{R})$, we chose the intervals of the form: $(-\infty, x],\ x \in \mathbb{R}$.

In view of this, we can think of the cdf as being defined via:

$$\mathbb{P}\{s\colon a < X(s) \leq b\} = \mathbb{P}\{s\colon X(s) \leq b\} - \mathbb{P}\{s\colon X(s) \leq a\} = P_X((a,b]) = F_X(b) - F_X(a),$$

and then assume that $F_X(-\infty) = 0$.

The properties of the cdf $F_X(x)$ in table 3.3, where $x \to x_0^+$ reads "as x tends to x_0 through values greater than x_0", are determined by those of $(S, \Im, \mathbb{P}(.))$. In particular from axioms [**1**]-[**3**] of $\mathbb{P}(.)$ and the mathematical structure of the $\sigma-$fields $\Im$ and $\mathcal{B}(\mathbb{R})$; see Karr (1993). That is, $F_X(x)$ is a non-decreasing, right-continuous function such that $F_X(-\infty) = 0$, and $F_X(\infty) = 1$. Properties **F1** and **F3** need no further explanation but **F2** is not obvious. The right-continuity property of the cdf follows from the axiom of countable additivity [3] of the probability set function $\mathbb{P}(.)$, whose value stems with the fact that at every point of discontinuity x_0 **F2** holds.

Table 3.3: **Cumulative distribution function-properties**

F1.	$F_X(x) \leq F_X(y)$, for $x \leq y,\ x, y$ real numbers,
F2.	$\lim_{x \to x_0^+} F_X(x) = F_X(x_0)$, for any real number x_0,
F3.	$\lim_{x \to \infty} F_X(x) := F_X(\infty) = 1,\ \lim_{x \to -\infty} F_X(x) := F_X(-\infty) = 0.$

The cumulative distribution function (cdf) provides the last link in the chain of the metamorphosis of $(S, \Im, \mathbb{P}(.))$ into something more amenable to modeling. Before

we proceed to shed additional light on the cdf we need to relate it to the concept of a density function introduced in the context of discrete random variables.

The discerning reader would have noticed that in the context of discrete random variables the metamorphosis of the abstract probability space took the form:

$$(S, \Im, \mathbb{P}(.)) \overset{X(.)}{\rightarrow} (\mathbb{R}_X, f_x(.)),$$

where $\mathbb{R}_X$=$\{x_1, ..., x_n, ...\}$. $(S, \Im, \mathbb{P}(.))$ has been transformed into:

$$\{f_x(x_1), f_x(x_2), ..., f_x(x_m), ...\}, \text{ such that } \sum_{x_i \in \mathbb{R}_X} f_x(x_i)=1.$$

The last link in the metamorphosis chain was the concept of a density function:

$$f_x(.)\colon \mathbb{R}_X \rightarrow [0,1],\ f_x(x):=\mathbb{P}(X=x) \text{ for all } x \in \mathbb{R}_X.$$

On the other hand, in the context of a continuous random variable (uncountable outcomes set) the metamorphosis took the form:

$$(S, \Im, \mathbb{P}(.)) \overset{X(.)}{\rightarrow} (\mathbb{R}_X, F_X(.)),$$

with the cdf being the last link in the chain. The reason why the density function could not be defined directly in this case has been discussed extensively in the previous chapter. The gist of the argument is that in the case of an uncountable outcomes set we cannot define probability at a point but only over an interval.

4.2 The concept of a density function

At this stage two questions arise naturally. The first is whether we can define a density function in the case of a continuous random variable. The second is whether we can define a cdf in the case of a discrete random variable. Both questions will be answered in the affirmative beginning with the first.

Having defined the cumulative distribution function over intervals of the form $((-\infty, x])$ we can proceed to recover the *density function* $f_x(.)$ (when it exists). Assuming that there exists a function of the form:

$$f_x(.)\colon \mathbb{R} \rightarrow [0, \infty), \tag{15}$$

such that it is related to the cdf via:

$$F_X(x)=\int_{-\infty}^{x} f_x(u)\ du, \text{ where } f_x(u) \geq 0, \tag{16}$$

$f_x(.)$ is said to be a **density function** that corresponds to $F_X(.)$.

This recovery presupposes the existence of a non-negative function whose form one has to guess in advance. In cases where $f_x(.)$ is assumed to be *continuous*, one can recover it from $F_X(.)$ using the *fundamental theorem of calculus;* see Binmore (1993). Suppose that $f_x(x)$ is a *continuous* function of x:

(a) if $F_X(x)=\int_{-\infty}^{x} f_x(u)\,du$, then $\frac{dF_X(x)}{dx}=f_x(x)$,

(b) if $\frac{dF_X(x)}{dx}=f_x(x)$, then $\int_a^b f_x(u)\,du=F_X(b)-F_X(a)$.

Using the fundamental theorem of calculus we can recover the density function much easier using the fact that:

$$\frac{dF_X(x)}{dx}=f_x(x),\ \text{at all continuity points of } f_x(x) \text{ for } x\in\mathbb{R}.$$

Example 3.10. Consider the random experiment of measuring 'the lifetime of a light bulb' in a typical home environment. The cumulative distribution function often used to model this experiment is that of the **exponential distribution**:

$$F_X(x;\theta)=1-e^{-\theta x},\ \theta>0,\ x\in\mathbb{R}_+=[0,\infty).$$

The graph of the cdf for $\theta=3$ is shown in figure 3.8. In view of the fact that $F_X(x;\theta)$ is continuous for all $x\in\mathbb{R}_+$, we can deduce that the density function is just the derivative of this function and takes the form (see figure 3.9): $f_x(x;\theta)=\theta e^{-\theta x},\ \theta>0,\ x\in\mathbb{R}_+$.

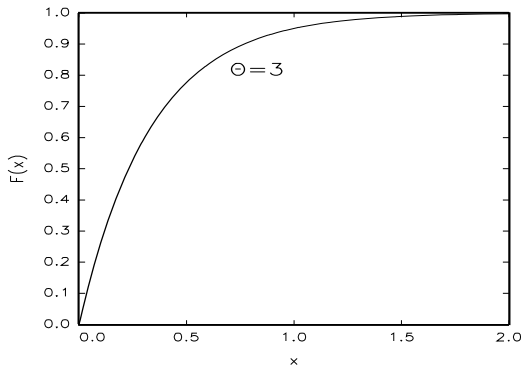

Fig. 3.8: The Exponential cdf

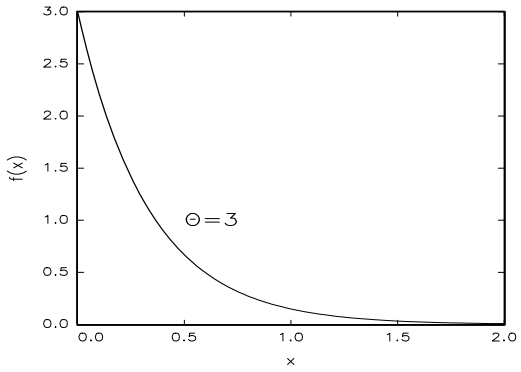

Fig. 3.9: The Exponential density

The **density function,** for *continuous random variables* (defined by (16)), satisfies the properties in table 3.4.

Table 3.4: **Density function (continuous) - Properties**

f1.	$f_x(x)\geq 0,$ for all $x\in\mathbb{R}_X$,
f2.	$\int_{-\infty}^{\infty} f_x(x)dx=1$,
f3.	$F_X(b)-F_X(a)=\int_a^b f_x(x)dx,\ a<b,\ a\in\mathbb{R}, b\in\mathbb{R}.$

Example 3.11. For what value of c is the function $f(x)=cx^2,\ 0<x<1$ a proper density? **f1** implies that $c>0$, but since $\int_0^1 cx^2dx=\left[\frac{cx^3}{3}\right]_0^1=\frac{1}{3}c$, for **f2** to hold $c=3$.

We now turn out attention to the question whether we can define a cdf in the case of discrete random variables. The definition of the cumulative distribution function given in (14) is also applicable to the case where $X(.)$ takes values in a *countable* subset of $\mathbb{R}$. For $\mathbb{R}_X$=$\{x_1, x_2, ..., x_n\}$, where $x_1 < x_2 < ... < x_n$, the cdf function of a random variable $X(.)$ is defined in terms of the density function by:

$$F_X(x_k)=\mathbb{P}(\{s\text{: } X(s) \leq x_k\})=\sum_{i=1}^{k} f_x(x_i), \text{ for } k=1, 2, ..., n. \tag{17}$$

That is, the cdf for a *discrete* random variable is a *step function* with the jumps defined by $f_x(.)$. The term *cumulative* stems from the fact that the cdf in both cases (14)-(17) accumulates the probabilities given by the density functions. This becomes apparent by ordering the values of X in ascending order $x_1 \leq x_2 \leq ... \leq x_n$ and assuming that $F_X(x_0)$=0, then $F_X(.)$ and $f_x(.)$ are related via:

$$f_x(x_i)=F_X(x_i)-F_X(x_{i-1}), \ i=1, 2, ...n.$$

The *density function*, in the case of a *discrete* random variable, has properties similar with those above with the integral replaced by a summation:

Table 3.5: **Density function (discrete) - Properties**

f1.	$f_x(x) \geq 0$, for all $x \in \mathbb{R}_X$,
f2.	$\sum_{x_i \in \mathbb{R}_X} f_x(x_i)=1$,
f3.	$F_X(b)-F_X(a)=\sum_{a<x_i \leq b} f_x(x_i), \ a<b, \ a \in \mathbb{R}, b \in \mathbb{R}$.

Example 3.12. In the case of the **Bernoulli** random variable the density function is:

$$f_x(1)=\theta \text{ and } f_x(0)=(1-\theta),$$

where $0 \leq \theta \leq 1$ (see (8)), which can be depicted as two spikes at x=0 and x=1 with heights $(1-\theta)$ and θ, respectively. The corresponding cdf takes the form $F_X(0)=\theta, F_X(1)$=1:

$$F_X(x)=\begin{cases} 0, & x<0, \\ \theta, & 0 \leq \theta < 1, \\ 1, & 1 \leq x. \end{cases}$$

The cdf is a step function with jumps at x=0 of height $(1-\theta)$ and x=1 of height θ.

Although the cdf appears to be the natural choice for assigning probabilities in cases where the random variable $X(.)$ takes values in an uncountable subset of $\mathbb{R}$, the density function offers itself more conveniently for modeling purposes. For this reason we conclude this section by mentioning some more distributions for both continuous and discrete random variables.

Continuous random variable. A random variable $X(.)$ is said to be *continuous* if its range of values is any uncountable subset of $\mathbb{R}$. A glance at the definition

(15)-(16) suggests that one should not interpret the density function of a continuous random variable as a function assigning probabilities, because the latter might take values greater than one!

Example 3.13. The most widely used distribution in probability theory and statistical inference is without a doubt the **Normal** (or **Gaussian**) distribution whose density function is:

$$f_x(x;\boldsymbol{\theta})=\tfrac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\tfrac{(x-\mu)^2}{2\sigma^2}\right\},\ \boldsymbol{\theta}:=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+,\ x\in\mathbb{R}. \tag{18}$$

The graph of this density function, shown in figure 3.10, exhibits the well-known bell shape symmetry with respect to the vertical line $x=\mu$ and $f_x(x;\boldsymbol{\theta})=f_x(2\mu-x;\boldsymbol{\theta})$, which is also the point of its maximum $f_x(\mu;\boldsymbol{\theta})=\frac{1}{\sigma\sqrt{2\pi}}$, since $\frac{df_x(x;\boldsymbol{\theta})}{dx}=-(\frac{x-\mu}{\sigma^2})f_x(x;\boldsymbol{\theta})$. The cdf for the Normal distribution is:

$$F_X(x;\boldsymbol{\theta})=\int_{-\infty}^{x}\tfrac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\tfrac{(u-\mu)^2}{2\sigma^2}\right\}du,\ \boldsymbol{\theta}:=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+,\ x\in\mathbb{R}. \tag{19}$$

The graph of this cdf, shown in figure 3.11, exhibits the distinct elongated S associated with the Normal distribution.

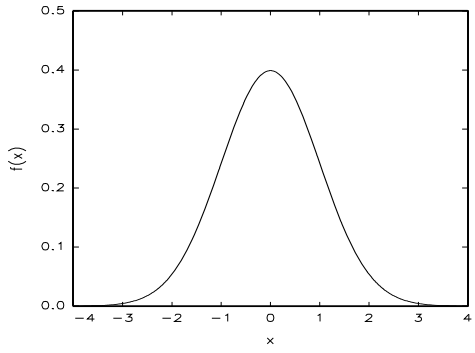

Fig. 3.10: The Normal density

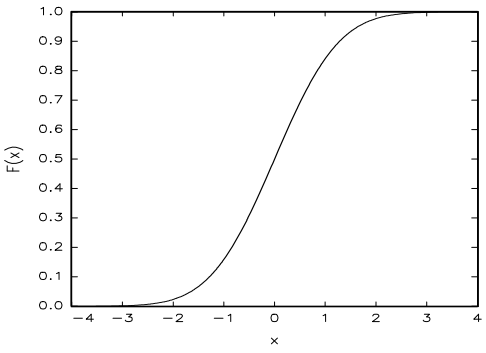

Fig. 3.11: The Normal cdf

Example 3.14. Another widely used distribution is the **Uniform** (continuous) whose density function is of the form:

$$f_x(x;\boldsymbol{\theta})=\tfrac{1}{b-a},\ \boldsymbol{\theta}:=(a,b)\in\mathbb{R}^2,\ a\leq x\leq b. \tag{20}$$

The graph of this density function for a=1 and b=3, exhibits the well-known rectangular shape. The cdf for the Uniform (continuous) distribution is:

$$F_X(x;\boldsymbol{\theta})=\tfrac{x-a}{b-a},\ \boldsymbol{\theta}:=(a,b)\in\mathbb{R}^2,\ a\leq x\leq b. \tag{21}$$

Discrete random variable. A random variable $X(.)$ is said to be *discrete* if its range $\mathbb{R}_X$ is a countable (it can be counted) subset of the real line $\mathbb{R}$; its density function is of the form:

$$f_x(.)\colon\ \mathbb{R}\to[0,1].$$

In contrast to the continuous random variable case, this definition suggests that one could interpret the density function of a discrete random variable as a function assigning probabilities.

Example 3.15. (a) The **Uniform** distribution has also a **discrete** form, with a density function:

$$f_x(x;\theta)=\frac{1}{n+1},\ n \text{ is an integer, } x=0,1,2,...,n. \tag{22}$$

The graph of this density function shows the same height spikes shape. The cdf for the Uniform (discrete) distribution is:

$$F_X(x;\theta)=\frac{x+1}{n+1},\ n \text{ is an integer, } x=0,1,2,...,n, \tag{23}$$

exhibiting jumps of the form: $p_k=\frac{1}{n+1},\ k=0,1,2,...,n.$

(b) Another widely used discrete distribution is the **Poisson** whose density function is:

$$f_x(x;\theta)=\frac{e^{-\theta}\theta^x}{x!},\ \theta>0,\ x=0,1,2,3,... \tag{24}$$

The graph of this density function, shown in figure 3.12 for $\theta=4$, where the asymmetry in the shape of the density is obvious. The cdf for the Poisson distribution is:

$$F_X(x;\theta)=\sum_{k=0}^{x}\frac{e^{-\theta}\theta^k}{k!},\ \theta>0,\ x=0,1,2,3,... \tag{25}$$

The graph of the cdf is shown in figure 3.13.

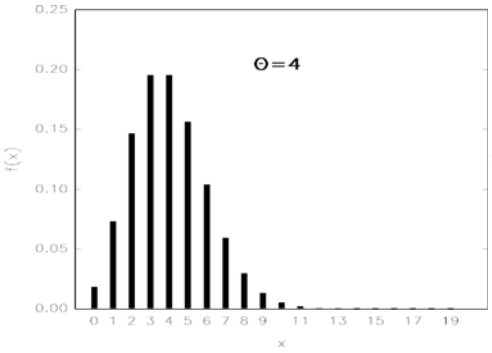

Fig. 3.12: The Poisson density

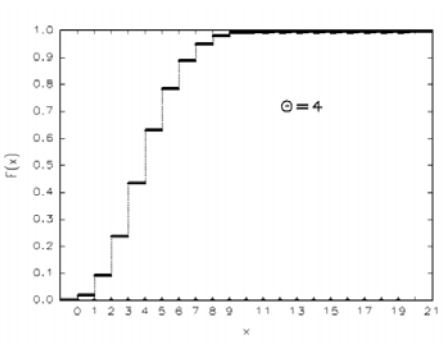

Fig. 3.13: The Poisson cdf

5 From a probability space to a probability model

Let us collect the various threads together. The primary aims has been to transform $(S, \Im, \mathbb{P}(.))$ into something more amenable to modeling with numerical data. After introducing the concept of a random variable $X(.)$: $S \to \mathbb{R}$, that preserves the event structure of $\Im$, it was used to map $(S, \Im, \mathbb{P}(.))$ into $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X(.))$. At the third stage the set function $P_X(.)$ was transformed into a numerical point-to-point function, the cdf:

$$F_X(x)=P_X((-\infty, x]), \text{ for all } x \in R.$$

This was then simplified $F_X(.)$ by introducing the density function via:

$$F_X(x)=\int_{-\infty}^{x} f_x(u)du, \ f_x(x) \geq 0, \text{ for all } x \in \mathbb{R}.$$

We then extended the formulation to allow $f(x), \ x \in \mathbb{R}_X$ to include unknown **parameter(s)** $\boldsymbol{\theta}$, $f(x;\boldsymbol{\theta}), \ \boldsymbol{\theta} \in \boldsymbol{\Theta}$. Symbolically the transformation has taken the form:

$$(S, \Im, \mathbb{P}(.)) \overset{X(.)}{\to} (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X(.)) \overset{X(.)}{\to} \{f(x;\boldsymbol{\theta}), \ \boldsymbol{\theta} \in \Theta, \ x \in \mathbb{R}_X\}.$$

We can view the mapping as: $S \Rightarrow \mathbb{R}_X$, and $[\Im, \mathbb{P}(.)] \Rightarrow \{f(x;\boldsymbol{\theta}), \ \boldsymbol{\theta} \in \Theta\}$.

The end result of this metamorphosis is that the original probability space $(S, \Im, \mathbb{P}(.))$ has become a **probability model** defined by:

$$\Phi=\{f(x;\boldsymbol{\theta}), \ \boldsymbol{\theta} \in \Theta, \ x \in \mathbb{R}_X\}. \tag{26}$$

Φ is a family of density functions indexed by a set of unknown parameters $\boldsymbol{\theta}$; one density for each possible value of $\boldsymbol{\theta}$ in the *parameter space* Θ.

Defining the probability model in terms of the density function, instead of the cdf, $\Phi_F=\{F(x;\boldsymbol{\theta}), \ \boldsymbol{\theta} \in \Theta, \ x \in \mathbb{R}_X\}$, is because it is more convenient for modeling purposes. The shape of the density function is easier to assess using graphical teachinques than that of the cdf. As shown in chapter 5, there is a helpful link between $f(x)$ and the histogram of the observed data, rendering the choice of an appropriate statistical model easier.

It is important to emphasize that the probability model has three important components: (i) the density function of a random variable X, (ii) the parameter space Θ and (iii) $\mathbb{R}_X$–the range of values of $X(.)$, which defines the **support** of the density $f_x(.)$ by $\mathbb{R}_X := \{x \in \mathbb{R}_X: \ f_x(x)>0\}$.

To illustrate the key concept of a probability model we consider two examples.

Example 3.16. (a) An interesting example of a probability model is the **Beta:**

$$\Phi=\left\{f(x;\boldsymbol{\theta})=\tfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B[\alpha,\beta]}, \ \boldsymbol{\theta}:=(\alpha,\beta) \in \mathbb{R}^2_+, \ 0<x<1\right\}.$$

In figure 3.14 several members this family of densities (one for each combination of values of $\boldsymbol{\theta}$) are shown. This probability model has two unknown parameters $\alpha > 0$

and $\beta > 0$; the parameter space is the product of the positive real line: $\Theta := \mathbb{R}^2_+$. This suggests that the set Φ has an infinity of elements, one for each combination of elements from two infinite sets. Its support is $\mathbb{R}^*_X := (0, 1)$. As can be seen, this probability model involves density functions with very different shapes depending on the values of the two unknown parameters.

(b) Another example of a probability model is the **Gamma:**

$$\Phi = \left\{ f(x;\boldsymbol{\theta}) = \frac{\beta^{-1}}{\Gamma[\alpha]} \left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left\{-\left(\frac{x}{\beta}\right)\right\},\ \boldsymbol{\theta} := (\alpha, \beta) \in \mathbb{R}^2_+, x \in \mathbb{R}_+ \right\}.$$

In figure 3.15 several members of this family of densities (one for each combination of values of $\boldsymbol{\theta}$) are shown. Again, the probability model has two unknown parameters $\alpha > 0$ and $\beta > 0$; the parameter space is the product of the positive real line: $\Theta := \mathbb{R}^2_+$. Its support is $\mathbb{R}^*_X := (0, \infty)$.

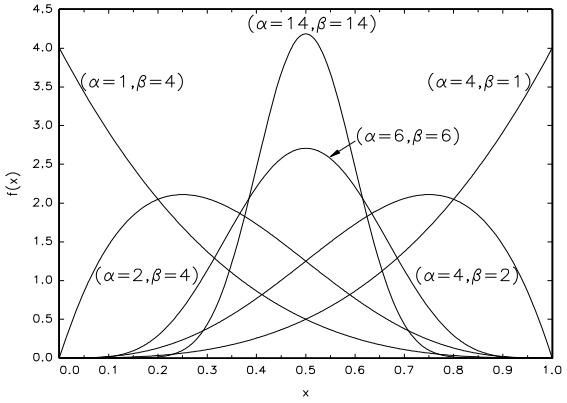

Fig. 3.14: Beta Probability Model

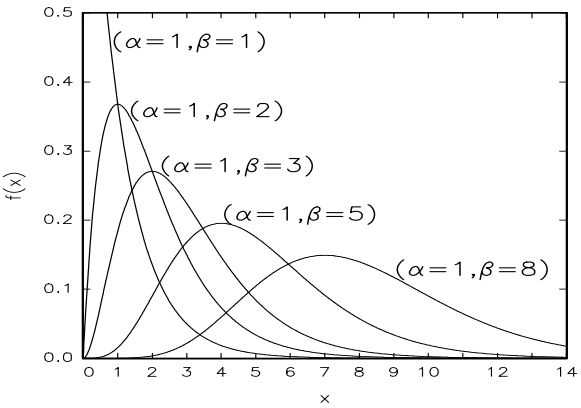

Fig. 3.15: Gamma probability model

For empirical modeling purposes the selection of a probability model is of paramount importance because it accounts for the 'distribution' chance regularities of the underlying stochastic mechanism that gave rise to the observed data in question. Our task is to chose the most appropriate family for the data in question; see Appendix 3.A for several such models. The question that naturally arises at this stage is:

▶ **How do we select an appropriate probability model?**

An oversimplified answer is that the modeler selects the probability model based on the chance regularity patterns exhibited by the particular data. Chapter 5 discusses how the histogram of the particular data can be used to make informed decisions with regard to the appropriate probability model. The best way to distinguish between similar looking distributional shapes is via measures based on moments, such as the skewness and kurtosis coefficients considered next.

5.1 Parameters and Moments

Why do we care? In the previous section we introduced the concept of a *probability model*:

$$\Phi=\{f_x(x;\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta,\ x\in\mathbb{R}_X\},$$

as a formalization of conditions [a]-[b] of a *Random Experiment*. Before we proceed to formalize condition [c] (see next chapter), we make an *important digression* to introduce a most convenient way to handle the unknown parameter(s) $\boldsymbol{\theta}$ of the probability model. In the context of statistical modeling and inference, the most efficient way to deal with the unknown parameters $\boldsymbol{\theta}$ is to relate them to the *moments* of the distribution. As mentioned in the previous section one of the important considerations in choosing a probability model is the shape of density functions. In selecting such probability models one can get ideas by looking at the histogram of the data as well as a number of numerical values, such as arithmetic averages, from descriptive statistics. These numerical values are related to what we call *moments* of the distribution and can be used to make educated guesses about the appropriateness of different probability models.

5.2 Numerical characteristics of random variables

5.3 Warning: be aware of confusing terminology

Table 3.8: **Confused and confusing Terminology -** BE WARE

	Descriptive statistics ($\mathbf{x}_0$) **Real world**	Probability Distribution **Math world**	Inferential Sample Statistics **Math world**
mean:	$\overline{x}:=\frac{1}{n}\sum_{i=1}^{n}x_i$	$\mu'_1=\int_{x\in\mathbb{R}_X} xf(x)dx$	$\widehat{\mu}'_1(\mathbf{X})=\sum_{i=1}^{n}X_i:=\overline{X}$
variance:	$\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$	$\mu_2=\int_{x\in\mathbb{R}_X}(x-\mu'_1)^2f(x)dx$	$\widehat{\mu}_2(\mathbf{X})=\sum_{i=1}^{n}(X_i-\overline{X})^2$
raw moments: $r=1,2,3,...$	$\frac{1}{n}\sum_{i=1}^{n}x_i^r$	$\mu'_r=\int_{x\in\mathbb{R}_X}x^rf(x)dx$	$\widehat{\mu}'_r(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n}X_i^r,$
central moments: $r=1,2,3,...$	$\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^r$	$\mu_r=\int_{x\in\mathbb{R}_X}(x-\mu'_1)^rf(x)dx$	$\widehat{\mu}_r(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^r$

At the outset of this section it is important to bring out a **serious problem with the terminology in statistics**, first raised by Fisher (1922).

The same terms, *mean*, *variance*, *covariance, correlation*, *skewness*, *kurtosis*, etc., are used for three related but different concepts; see table 3.8. The use of such terms in

the context of *descriptive statistics* denotes summaries of data $\mathbf{x}_0{:=}(x_1, x_2, \ldots, x_n)$, where the small letters denote actual numbers and the end result is another *number*. In the context of *probability theory*, these terms refer to mathematical concepts associated with the underlying probability model, and they denote *unknown (constant) parameters*.

In *inferential statistics* these terms refer to different functions $h(\mathbf{X})$ of the underlying sample $\mathbf{X}{:=}(X_1, X_2, \ldots, X_n)$, where the capital letters denote random variables, and they denote *random variables* of the form $Y{=}h(\mathbf{X})$ with their own sampling distributions.

Moments come in two flavors:

Raw moments: $E(X^r){=}\int_{-\infty}^{\infty} x^r{\cdot}f_x(x;\boldsymbol{\theta})dx,\ r=1,2,...,n,...$

Central moments: $E(X-\mu)^r{=}\int_{-\infty}^{\infty}(x{-}\mu)^2 f(x;\boldsymbol{\theta})dx,\ r=2,...,n,...$

5.3.1 The Mean

For $h(X){:=}X$, where X takes values in $\mathbb{R}_X$, the the *mean* of the distribution is:

$$\begin{aligned} &\textbf{Discrete:} && E(X){=}\textstyle\sum_{x_i\in\mathbb{R}_X} x_i{\cdot}f_x(x_i;\boldsymbol{\theta}),\\ &\textbf{Continuous:} && E(X){=}\textstyle\int_{-\infty}^{\infty} x{\cdot}f_x(x;\boldsymbol{\theta})dx. \end{aligned} \tag{27}$$

NOTE that the only difference in the definition between continuous and discrete random variables is the replacement of the integral by a summation. The mean is a *measure of location* in the sense that knowing what the mean of X is, we have some idea on where $f_x(x;\boldsymbol{\theta})$ is located. Intuitively, the mean represents a weighted average of the values of X, with the corresponding probabilities providing the weights. Denoting the mean by, $\mu{:=}E(X)$, the above definition suggests that μ is a function of the unknown parameters $\boldsymbol{\theta}$, i.e. $\mu(\boldsymbol{\theta})$. This provides the modeler with a direct relationship between the first moment of $f(x;\boldsymbol{\theta})$ and $\boldsymbol{\theta}$.

Example 3.21. (a) For the *Bernoulli* distribution: $\mu(\theta){:=}E(X){=}0{\cdot}(1{-}\theta){+}1{\cdot}\theta{=}\theta$, and thus, the mean coincides with the unknown parameter θ.

(b) For the *Uniform* distribution (a continuous distribution):

$$f(x;\boldsymbol{\theta}){=}\tfrac{1}{(\theta_2-\theta_1)},\ x{\in}[\theta_1,\theta_2],\ \boldsymbol{\theta}{:=}(\theta_1,\theta_2),\ -\infty{<}\theta_1{<}\theta_2{<}\infty,$$

$$\mu(\boldsymbol{\theta}){:=}E[X]{=}\textstyle\int_{\theta_1}^{\theta_2}\frac{x}{(\theta_2-\theta_1)}dx{=}\frac{1}{2}\left(\frac{1}{\theta_2-\theta_1}\right)x^2\Big|_{\theta_1}^{\theta_2}{=}\frac{\theta_1+\theta_2}{2}.$$

(c) For the *Normal* distribution:

$$f(x;\boldsymbol{\theta}){=}\tfrac{1}{\sigma\sqrt{2\pi}}\exp(-\tfrac{(x-\mu)^2}{2\sigma^2}),\ \boldsymbol{\theta}{:=}(\mu,\sigma^2){\in}\mathbb{R}\times\mathbb{R}_+,\ x{\in}\mathbb{R}, \tag{28}$$

the parameter μ is actually the *mean* of the distribution (hence the notation):

$$\begin{aligned} E(X) \ &{=}\textstyle\int_{-\infty}^{\infty} x{\cdot}(\frac{1}{\sigma\sqrt{2\pi}})\exp(-\frac{(x-\mu)^2}{2\sigma^2})dx{=}\int_{-\infty}^{\infty}\frac{\sigma z+\mu}{\sigma\sqrt{2\pi}}\exp(-\tfrac{z^2}{2})\sigma dz{=}\\ &{=}\ \tfrac{\sigma}{\sqrt{2\pi}}\textstyle\int_{-\infty}^{\infty} z\exp(-\frac{z^2}{2})dz{+}\mu\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})dz{=}0{+}\mu{\cdot}1{=}\mu. \end{aligned}$$

The second equality follows using the substitution $z=\left(\frac{x-\mu}{\sigma}\right)$ or $x=\sigma z+\mu$, with $\frac{dx}{dz}=\sigma$.

(d) **Counter-example**. Consider the case of the Cauchy distribution:

$$f(x;\theta)=\tfrac{1}{\pi(1+x^2)},\ x\in\mathbb{R}.$$

$$E(X)=\textstyle\int_{-\infty}^{\infty}\frac{x}{\pi(1+x^2)}dx=\frac{\ln(1+x^2)}{\pi}\Big|_{-\infty}^{\infty}=\infty-(-\infty)=\infty.$$

In light of this indeterminacy, $E(X)$ does not exist for the Cauchy distribution.

For random variables X_1 and X_2 and the constants a, b and c, $E(.)$ satisfies the properties in table 3.6.

Table 3.6: **Mean - Properties**

E1. $E(c)=c$,
E2. $E(aX_1+bX_2)=aE(X_1)+bE(X_2)$.

These properties designate $E(\ )$ a *linear mapping*.

Example 3.22. Let $X_1,...,X_n$ be Bernoulli distributed random variables with mean θ. Then using [**E2**] one can show that: $E(Y)=\sum_{i=1}^{n}E(X_i)=\sum_{i=1}^{n}\theta=n\theta$.

5.3.2 Variance

For $h(X):=[X-E(X)]^2$ the integral:

$$E(X-\mu)^2=\int_{-\infty}^{\infty}(x-\mu)^2 f(x;\boldsymbol{\theta})dx$$

yields *the variance*:

$$Var(X):=E[(X-E(X))^2]=\textstyle\int_{-\infty}^{\infty}[(x-\mu)^2 f_x(x;\boldsymbol{\theta})dx,$$

where in the case of discrete random variables the integral is replaced by the usual summation (see (27)). In our context the variance represents a *measure of dispersion* (variation) around the mean.

Example 3.23. (a) In the case of the *Bernoulli* model:

$$Var(X)=E(X-E(X))^2=(0-\theta)^2\cdot(1-\theta)+(1-\theta)^2\cdot\theta=\theta(1-\theta).$$

(b) Consider X with a density function $f(x)=2e^{-x},\ x>0$, $\int_0^{\infty}2e^{-2x}dx=1$:

$$\begin{aligned}E(X) &= \textstyle\int_0^{\infty}(x)2e^{-2x}dx=2\left[x\left(\frac{e^{-2x}}{-2}\right)-1\left(\frac{e^{-2x}}{4}\right)\right]_0^{\infty}=\tfrac{1}{2},\\ E(X^2) &= \textstyle\int_0^{\infty}(x^2)2e^{-2x}dx=2\left[x^2\left(\frac{e^{-2x}}{-2}\right)-2x\left(\frac{e^{-2x}}{4}\right)+2\left(\frac{e^{-2x}}{8}\right)\right]_0^{\infty}=\tfrac{1}{2}\end{aligned}$$

Hence, $Var(X)=E(X^2)-[E(X)]^2=\frac{1}{4}$.

(c) For $X \backsim \mathsf{N}(\mu, \sigma^2)$ the mean (μ) was derived in Example 3.20(c). One can show that $Var(X)=\sigma^2$ using the substitution $x=\sigma z+\mu$:

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 (\tfrac{1}{\sigma\sqrt{2\pi}}) \exp\left(-\tfrac{(x-\mu)^2}{2\sigma^2}\right) dx = \int_{-\infty}^{\infty} \tfrac{(\sigma z+\mu)^2}{\sigma\sqrt{2\pi}} \exp\left(-\tfrac{z^2}{2}\right)(\sigma)\, dz = \\ &= \sigma^2 \int_{-\infty}^{\infty} \tfrac{z^2}{\sqrt{2\pi}} \exp(-\tfrac{z^2}{2}) dz + \tfrac{2\sigma\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \tfrac{z}{\sqrt{2\pi}} \exp(-\tfrac{z^2}{2}) dz + \mu^2 \int_{-\infty}^{\infty} \tfrac{1}{\sqrt{2\pi}} \exp(-\tfrac{z^2}{2}) dz = \\ &= \sigma^2 + 0 + \mu^2 = \sigma^2 + \mu^2, \end{aligned}$$

hence the notation $X \backsim \mathsf{N}(\mu, \sigma^2)$. In figure 3.16 we can see the Normal density (with μ=0) and different values of σ^2; the greater the value of σ^2 the greater the dispersion.

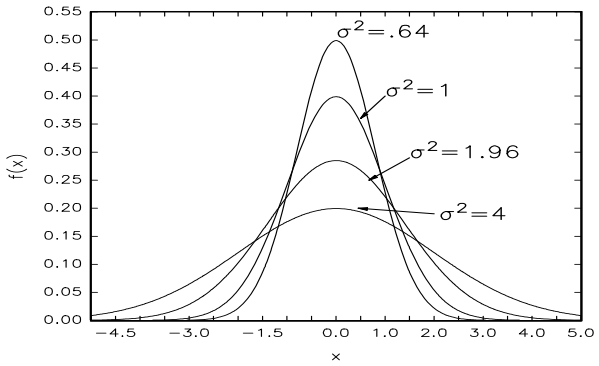

Fig. 3.16: Normal density with different σ^2

For *independent* random variables X_1 and X_2 and any constants a, b and c, $Var(.)$ satisfies the properties in table 3.7.

Table 3.7: **Variance - Properties**

V1.	$Var(c)=0$,
V2.	$Var(aX_1 + bX_2)=a^2Var(X_1)+b^2Var(X_2)$,
V3.	$Var(X)=E(X^2) - [E(X)]^2$,
V4.	$Var(X)=0$ if and only if $\mathbb{P}(X=E(X))=1$.

Bienayme's lemma. If $X_1, X_2, ..., X_n$ are Independently distributed random variables:

$$Var(\textstyle\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i^2 Var(X_i).$$

This lemma constitutes a direct extension of property **V2**.

Example 3.24. Let $X_1, X_2, ..., X_n$ be Independent Bernoulli distributed random variables with mean θ. What is the variance of $Y=a + \sum_{i=1}^n X_i$? Using [V1] and Bienayme's lemma, we can deduce that:

$$Var(Y) = \textstyle\sum_{i=1}^n Var(X_i) = \sum_{i=1}^n \theta(1-\theta) = n\theta(1-\theta).$$

A particularly useful inequality which testifies that the variance provides a measure of dispersion is that of Chebyshev.

Chebyshev **inequality**. Let Xbe a random variable with bounded variance:

$$\mathbb{P}\left(|X-E(X)|>\varepsilon\right) \leq \frac{Var(X)}{\varepsilon^2}, \text{ for any } \varepsilon > 0.$$

Appendix 9.A lists several important probabilistic inequalities.

5.3.3 Standard deviation

The square root of the variance of a random variable, named by Pearson (1894) as the *standard deviation*, is also used as a measure of dispersion: $SD(X)=[Var(X)]^{\frac{1}{2}}$.

This measure is particularly useful in statistical inference because it provides one with the best way to standardize any random variable X whose variance exists. One of the most useful practical rules in statistical inference is the following:

A random variable X is as "big" as its standard deviation $SD(X)<\infty$.

It can be used to render a random variable X free of its units of measurement by *standardizing it*: $X^*:=\frac{X}{\sqrt{[Var(X)]}}$, where $Var(X^*)=1$.

Although the mean and the variance are the most widely used moments, they do not suffice to determine the main features of a the density function $f(x)$, $x\in\mathbb{R}_X$. That is, knowing the mean (location) and the variance (variation) we know very little about the *shape* of $f(x)$. To get some idea about the shape of $f(x)$ we need to consider higher moments.

5.4 Higher moments

$E(X)$ and $Var(X)$ belong to two broader classes of numerical characteristics of random variables and their distributions, known as *higher raw* and *central moments*, respectively. The concept of *moments* in general was borrowed from classical mechanics, where the mean $E(X)$ is the abscissa of the center of gravity of the mass of the distribution, and the variance $Var(X)$ represents the moment of inertia of the mass of the distribution with respect to a perpendicular axis through the point $x=E(X)$. The first to coin the term was Karl Pearson (1893), who, at the time, was a professor of Applied Mathematics and Mechanics at University College London.

5.4.1 Higher Raw moments

For $h(X):=X^r$, $r=2,3,4,...$ the integral in (**??**) yields *the raw moments*:

$$\mu'_r(\boldsymbol{\theta}):=E\left(X^r\right)=\int_{-\infty}^{\infty} x^r f_x(x;\boldsymbol{\theta})dx, \quad r=1,2,3,...$$

Example 3.25. Exponential random variable X with a density function:

$$f_x(x;\theta)=\theta e^{-\theta x}, \; x>0, \; \theta>0,$$

the raw moments are defined by: $\mu'_r(\theta)=\int_0^\infty x^r\theta e^{-\theta x}dx$. Using the change of variables $u=\theta x,\ dx=\frac{1}{\theta}du$, we deduce that: $\mu'_r(\theta)=\int_0^\infty \frac{u^r}{\theta^r}e^{-u}du=\frac{1}{\theta^r}\int_0^\infty u^{(r+1)-1}e^{-u}du=\frac{r!}{\theta^r}$.

In practice it is often easier to derive the higher moments indirectly using the moment generating (mgf) or the characteristic function (chf) defined as the two-sided Laplace and Fourier transformations of the density function, respectively:

$$\begin{array}{ll}\text{Mgf:} & m_X(t):=E\left(e^{tX}\right)=\int_{-\infty}^{\infty}e^{tX}f(x)dx,\ \text{for } t\in(-h,h),\ h>0,\\ \text{Chf:} & \varphi_X(t):=E\left(e^{itX}\right)=\int_{-\infty}^{\infty}e^{itx}f(x)\,dx,\ \text{for } i=\sqrt{-1}.\end{array} \tag{29}$$

This enables one to derive the higher moments using differentiation instead of integration:

$$\tfrac{d^r}{dt^r}m_X(t)\big|_{t=0}:=m_X^{(r)}(0)=E(X^r),\quad \tfrac{d^r}{dt^r}\varphi_X(t)\big|_{t=0}=E(i^rX^re^{itX})\big|_{t=0}=i^rE(X^r),\ r=1,2,3,...$$

5.4.2 Higher Central moments

The concept of the variance can be extended to define the **central moments** using the sequence of functions $h(X):=(X-E(X))^r,\ r=3,4,...$ in (**??**):

$$\mu_r(\boldsymbol{\theta}):=E(X-\mu)^r=\int_{-\infty}^{\infty}(x-\mu)^r\,f(x;\boldsymbol{\theta})dx,\ r=2,3,...$$

Example 3.26. Normal distribution $[X\backsim\mathsf{N}(\mu,\sigma^2)]$:

$$E(X-\mu)^r=\begin{cases}(\frac{r!}{2^{\frac{r}{2}}(\frac{r}{2}!)})\sigma^r, & \text{for } r=2,4,6,...\\ 0, & \text{for } r=3,5,7,...\end{cases}$$

Not surprisingly, the central moments are directly related with the raw moments as well as the *cumulants,* $\kappa_r,\ r=1,2,...,$ (see Appendix 3.B) via:

$$\begin{array}{ll}\mu_2=\mu'_2-(\mu'_1)^2,\ \mu_3=\mu'_3-3\mu'_2\mu'_1+2(\mu'_1)^3, & \kappa_2=\mu_2,\ \kappa_3=\mu_3,\\ \mu_4=\mu'_4-4\mu'_3\mu'_1+6\mu'_2(\mu'_1)^2-3(\mu'_1)^4, & \kappa_4=\mu_4-3\mu_2^2,\\ \vdots & \vdots\end{array}$$

For more details see Stuart, Ord, and Arnold (1999).

Example 3.27. The first four cumulants of the Normal distribution $[X\backsim\mathsf{N}(\mu,\sigma^2)]$ are:

$$\kappa_1=\mu,\ \kappa_2=\sigma^2,\ \kappa_3=\mu_3=0,\ \kappa_4=\mu_4-3\mu_2^2=0,\ \kappa_r=0,\ r>4.$$

One of the main uses of the central moments is that they can be used to give us a more complete picture of *the distribution's shape.* By standardizing the above central moments we define a number of useful measures which enable us to get a more complete idea of the possible shape of a density function. The first important feature of a distribution's shape is that of symmetry around a given point a; often $a=E(X)$.

Symmetry. A random variable X with density $f(x)$ is said to have a symmetric distribution about a point x_0 if:

$$f(x_0-x)=f(x_0+x),\ \text{for all } x\in\mathbb{R}.$$

In terms of the cdf symmetry is defined by: $F_X(-x)=1-F_X(x)$, for all $x\in\mathbb{R}_X$.

5.4.3 The Skewness coefficient

The first index of shape, designed to gives us some idea about the possible asymmetry of a density function around the mean, is the *Skewness* coefficient defined as the standardized third central moment, introduced by Pearson (1895):

$$\textbf{Skewness}: \alpha_3(X)=\tfrac{\mu_3}{(\sqrt{\mu_2})^3}.$$

NOTE that $\sqrt{\mu_2}=[Var(X)]^{\frac{1}{2}}$ denotes the standard deviation. If the distribution is *symmetric* around the mean then, $\alpha_3=0$; the converse is not true!

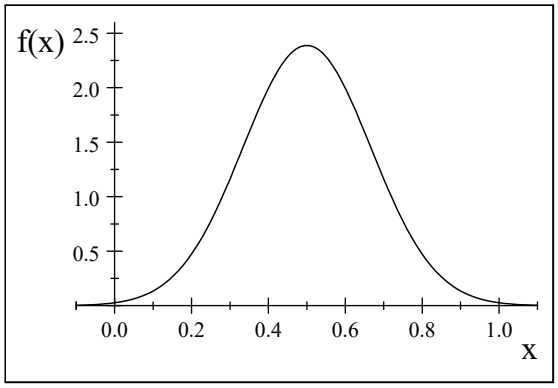

Fig. 3.17: $\mathsf{N}(.5, .0278)$ density

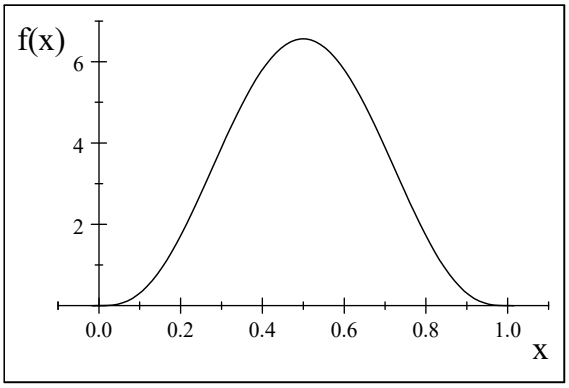

Fig. 3.18: $\mathsf{Beta}(4, 4)$ density

Example 3.28. Figure 3.17 depicts the Normal density (28), for values of the mean and variance chosen to be the same as those in fig. 3.18 as those of the Beta density (30):

$$f(x;\boldsymbol{\theta})=\tfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B[\alpha,\beta]},\ \boldsymbol{\theta}:=(\alpha,\beta)\in\mathbb{R}^2_+,\ 0<x<1, \tag{30}$$

for $\alpha=\beta=4$, in figure 3.18. Both distributions are symmetric with skewness coefficient $\alpha_3=0$; NOTE that for the Beta density $\alpha=\beta>1\rightarrow\alpha_3=\tfrac{2(\beta-\alpha)\sqrt{(\alpha+\beta+1)}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}=0$.

CAUTIONARY NOTE. $\alpha_3=0$ *does not imply* that the distribution is symmetric!

Example 3.29. For the non-symmetric distribution given below:

x	-2	1	3
$f(x)$	.4	.5	.1

(31)

$$E(X)=(-2)(.4)+1(.5)+3(.1)=0,\quad E(X^3)=(-2)^3(.4)+1(.5)+3^3(.1)=0.$$

That is, $\alpha_3=0$ despite its non-symmetry; see Romano and Siegel (1986). This example brings out the importance of looking at the graphs of the distributions and not just at some summary descriptive statistics.

Skewed distributions. The Beta distribution in (30) is non-symmetric (skewed) for $\alpha\neq\beta$. Figure 3.19 depicts two positively skewed ($\alpha_3>0$) Beta densities with values

(α=1, β=4) and (α=2, β=4), and figure 3.20 two negatively skewed density functions (α_3<0) with values (α=4, β=1) and (α=4, β=2).

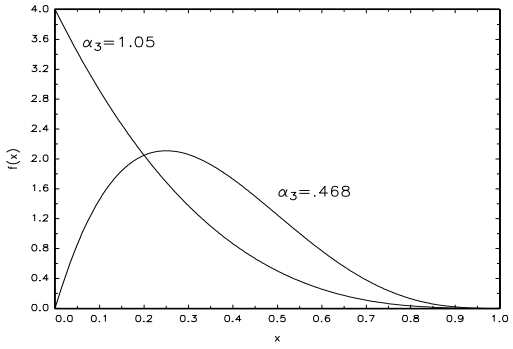

Fig. 3.19: Positively skewed densities

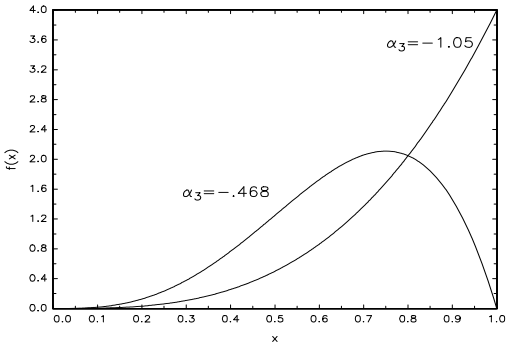

Fig. 3.20: Negatively skewed densities

5.4.4 Coefficient of Kurtosis

The skewness coefficient enables us to distinguish between a symmetric and a non-symmetric distribution but that still leaves us with the problem of distinguishing between two symmetric distributions with different shapes, such as the Normal and the bell-shape symmetric Beta densities shown in figures 3.17 and 3.18. Looking at these two graphs we can see that these two densities differ with respect to their peak and their tails. The Normal has a bell shaped peak but that of the symmetric Beta is more rounded. The Normal has longish tails extending to infinity on both sides but the Beta tails die out immediately. Intuition suggests that one way to distinguish between them is to devise a measure which measures *peakedness* in relation to tails. The *kurtosis* coefficient is such a measure, originally introduced in Pearson (1895).

The kurtosis is a standardized version of the fourth central moment:

$$\textbf{Kurtosis: } \alpha_4(X)=\frac{\mu_4}{(\mu_2)^2}.$$

Terminology. The term comes from the Greek word $\kappa\acute{\upsilon}\rho\tau\omega\sigma\eta$ which means 'curvature' and aims to measure the *peakedness* of the density function in relation to the *shape of the tails*. For the Normal distribution (28) α_4=3, and it is referred to as a *mesokurtic distribution*; *meso* comes from the Greek word $\mu\acute{\varepsilon}\sigma os$ - middle. In the case where the distribution in question has a flatter *peak* than the Normal (α_4<3), we call it *platykurtic*, and in the case where it has a more pointed peak than the Normal (α_4>3), we call it *leptokurtic*; *platy* and *lepto* come from the Greek words $\pi\lambda\alpha\tau\acute{\upsilon}s$ and $\lambda\epsilon\pi\tau\acute{o}s$ which mean *wide* and *slim,* respectively; these terms were introduced by Pearson (1906). NOTE that in some books the measure used is not α_4 but (α_4−3), the *excess kurtosis*.

Example 3.30. Returning to figures 3.17 and 3.18, the $\mathsf{N}(\mu,\sigma^2)$ is mesokurtic (α_4=3) and the $\mathsf{Beta}(4,4)$ density (Appendix 3.A) is platykurtic since:

$$\alpha_4=\frac{3(\alpha+\beta+1)\big[2(\alpha+\beta)^2+\alpha\beta(\alpha+\beta-6)\big]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)}\overset{\alpha=\beta}{=}3-\left(\frac{6}{2\alpha+3}\right)=2.455.$$

Intuitively, we can think of the kurtosis coefficient as a measure which indicates whether a symmetric distribution when compared with the Normal has thicker tails and more pointed peak or not. Viewing the Normal density as a bell-shaped pile made of plaster the sculptor shaves off part of the shoulders and adds it the tails and the peak to produce a leptokurtic distribution.

[**a**] **Leptokurtic**. Any distribution whose kurtosis coefficient α_4>3 is called leptokurtic.

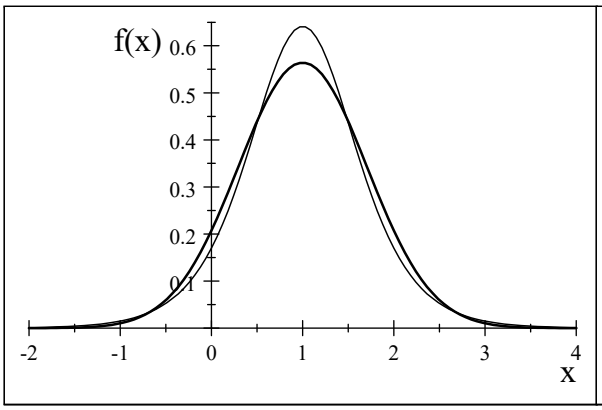

Fig. 3.21: $\mathsf{N}(1,.5)$ vs. $\mathsf{Lg}(1,.5)$ densities

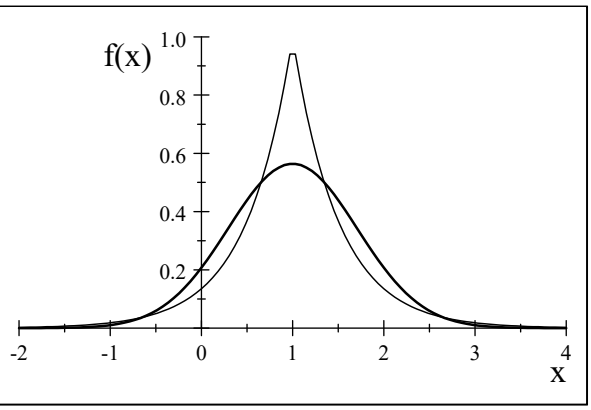

Fig. 3.22: $\mathsf{N}(1,.5)$ vs. $\mathsf{Lp}(1,.39)$ densities

Example 3.31. Figure 3.21 compares a Normal [$\mathsf{N}(1,.5)$] with a Logistic [$\mathsf{Lg}(\alpha,\beta)$] density:

$$f(x;\boldsymbol{\theta})=\tfrac{1}{2\beta}e^{-\left(\frac{|x-\alpha|}{\beta}\right)},\ \boldsymbol{\theta}:=(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}_+, x\in\mathbb{R},$$

with $\alpha=E(X)=1$, $\beta=.38985$ [$Var(X)=\frac{\beta^2\pi^2}{3}=.5$] and $\alpha_4=4.2$. Figure 3.22 compares a Normal [$\mathsf{N}(1,.5)$] with a Laplace [$\mathsf{Lp}(\alpha,\beta)$] density:

$$f(x;\boldsymbol{\theta})=\tfrac{1}{2\beta}e^{-\left(\frac{|x-\alpha|}{\beta}\right)},\ \boldsymbol{\theta}:=(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}_+, x\in\mathbb{R},$$

with $\alpha=E(X)=1$, $\beta=.5$ $\left[Var(X)=2\beta^2=.5\right]$ and $\alpha_4=6$; see Appendix 3A.

Example 3.32. Figure 3.23 compares the standard Normal [$\mathsf{N}(0,1)$] density (bold line) and the standard *Student's t* density with $\nu=5$, denoted as $\mathsf{St}(\nu=5)$:

$$f(x)=\frac{\Gamma[\frac{1}{2}(\nu+1)](\nu\pi)^{-\frac{1}{2}}}{\Gamma[\frac{1}{2}\nu]}\left(1+\frac{x^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)},\ \nu>2,\ x\in\mathbb{R}. \tag{32}$$

The Normal (meso-kurtic) and the Student's t (lepto-kurtic) differ in two respects:

(i) The tails of the $\mathsf{St}(\nu{=}5)$ are thicker, (ii) The peak of the $\mathsf{St}(\nu{=}5)$ is more pointed.

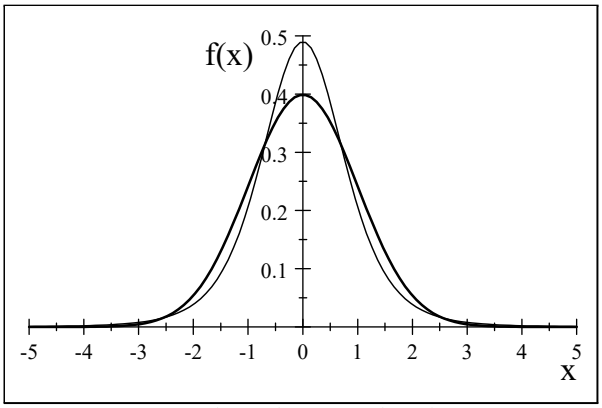

Fig. 3.23: $\mathrm{St}(\nu{=}5)$ vs. $\mathsf{N}(0,1)$ densities both scaled to have $\sqrt{Var(X)}{=}1$.

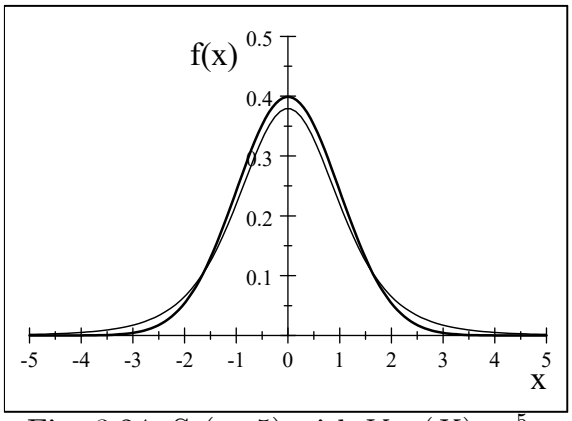

Fig. 3.24: $\mathrm{St}(\nu{=}5)$ with $Var(X){=}\frac{5}{5-2}$ vs. $\mathsf{N}(0,1)$ with $Var(X){=}1$.

WARNING: in many textbooks the graph of the Normal and Student's t distribution looks like figure 3.24 instead. The latter picture is misleading, however, because the Normal has $SD(X){=}1$ but the Student's t is $SD(X){=}\sqrt{\frac{\nu}{\nu-2}}$. Standardizing the latter to $SD(X){=}1$ yields figure 3.23, which is the relevant plot when looking at real data plots (chapter 5); see example 4.42 for the details of the transformation. In figure 3.25 we compare the Normal [N(0,1)] (bold, lowest peak) with two even more leptokurtic distributions, the $\mathsf{St}(\nu{=}3)$ and the Cauchy(.4) (highest peak) where .4 is the scale parameter.

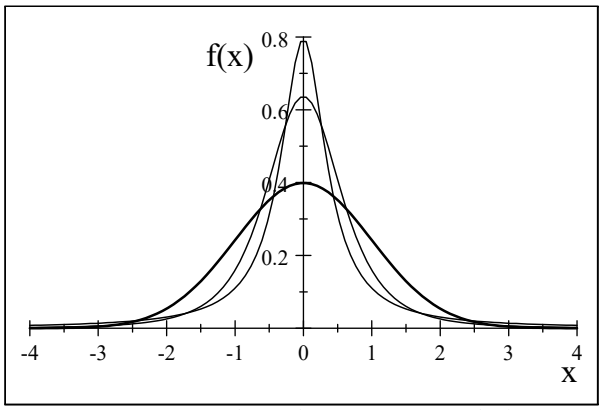

Fig. 3.25: $\mathrm{St}(\nu{=}3)$ vs. Cauchy(.4) vs. Normal [N(0,1)] densities

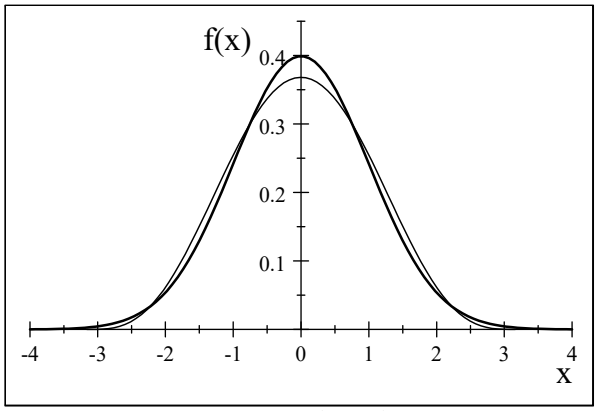

Fig. 3.26: Pearson II ($\nu{=}3$), -3.16$\leq x\leq$3.16 vs. Normal [N(0,1)] densities

[**b**] **Platykurtic**. Any distribution whose kurtosis coefficient $\alpha_4 < 3$ is called platykurtic.

Example 3.33. In figure 3.26 we compare the Normal density (bold) with a platykurtic density, the Pearson type II with $\nu{=}3$:

$$f(x){=}\left(\tfrac{\Gamma[\nu+2]}{\Gamma[.5]\cdot\Gamma[\nu+1.5](c)}\right)\left(1{-}\tfrac{x^2}{c^2}\right)^{(\nu+\frac{1}{2})},\ -c\leq x\leq c,\ c^2{=}2(\nu{+}2). \tag{33}$$

The Normal density differs from the Pearson type II in exactly the opposite way than it differs from the Student's t:

(a) The tails of the Pearson II are slimmer,
(b) The curvature of the Pearson II is less pointed.

It should be non-surprising that the Pearson II distribution is directly related to the symmetric Beta distribution (α=β); see Appendix 3.A.

The Normal, the Student's t, and the Pearson type II densities are bell shaped, but they differ in terms of their kurtosis, meso, lepto and platykurtic, respectively.

In conclusion, it must be said that the usefulness of the kurtosis coefficient is reduced in the case of non-symmetric distributions because it does not have the same interpretation as in the symmetric cases above; see Balanda and MacGillivray (1988).

Example 3.34. Consider the discrete random variable X with a density function:

x	0	1	2
$f(x)$	0.3	0.3	0.4

(34)

$E(X)$=0(.3)+1(.3) + 2(.4)=1.1, $E(X^2)$=0^2(.3)+1^2(.3)+2^2(.4)=1.9,
$E(X^3)$=0^3(.3)+1^3(.3)+2^3(.4)=3.5, $E(X^4)$=0^4(.3)+1^4(.3)+2^4(.4)=6.7.

$Var(X)$= $[0-1.1]^2(.3) + [1-1.1]^2(.3) + [2-1.1]^2(.4)$= 0.69,

$Var(X)$= $E(X^2) - [E(X)]^2$= 1.90 − 1.21=0.69,

$E\{(X-E(X))^3\}$= $[0-1.1]^3(.3)+[1-1.1]^3(.3)+[2-1.1]^3(.4)$=0.108,
$E\{(X-E(X))^4\}$= $[0-1.1]^4(.3)+[1-1.1]^4(.3)+[2-1.1]^4(.4)$=0.7017,

$\alpha_3=\left(\frac{0.108}{(0.83)^3}\right)$=0.18843, $\alpha_4=\left(\frac{0.7017}{(0.83)^4}\right)$=1.4785.

Example 3.35. Consider the continuous random variable X with: $f(x)$=$2x$, $0<x<1$.

$E(X)=\int_0^1 2x^2dx=\frac{2}{3}x^3\big|_0^1=\frac{2}{3}$, $E(X^2)=\int_0^1 2x^3dx=\frac{2}{4}x^4\big|_0^1=\frac{1}{2}$,
$Var(X)=E(X^2)-[E(X)]^2=\frac{1}{2}-\frac{4}{9}=\frac{1}{18}$, $E(X^3)=\int_0^1 2x^4dx=\frac{2}{5}x^5\big|_0^1=\frac{2}{5}$.

Invariance of skewness and kurtosis. We conclude the discussion of the skewness and kurtosis coefficients by re-iterating that their usefulness stems from the fact that they are *invariant* to *location* and *scale* changes, i.e.

$$\alpha_3(X)=\alpha_3(a+bX) \quad \text{and} \quad \alpha_4(X)=\alpha_4(a+bX).$$

5.5 The problem of moments

Despite their importance and usefulness, moments do not always exist for certain random variables of interest.

Example 3.36. When the random variable X is *Cauchy* distributed (see Appendix 3.A), none of its moments exist.

The existence of moments relates to how heavy (fat) the tail areas, [$\mathbb{P}(|X|>x)$ as x increases], of the underlying distribution are, so that when integrated (or summed) it gives rise to a finite number.

Sufficient condition. To get sense of 'how heavy is not too heavy', it can be shown that when for any $p>0$:

$$x^p\mathbb{P}(|X|>x) \underset{x\to\infty}{\to} 0,$$

then all moments lower order than p exist, i.e. $E(X^r)<\infty$ for $0\leq r<p$; see Romano and Siegel (1986).

Lower moments lemma. If $\mu'_k:=E(X^k)<\infty$ (exist) for some positive integer k, then all the raw moments of order less than k also exist, i.e.

$$E(X^i)<\infty, \text{ for all } i=1,2,...,k-1.$$

The questions that naturally arise are: given a set of moments:

$$\left\{\mu'_k:=E(X^k)<\infty,\ k=1,2,\ldots\right\},$$

▶ (i) is there a function $f(x)\geq 0$, such that: $\mu'_k=\int_{-\infty}^{\infty} x^r f(x)dx$? [existence]

▶ (ii) is the function $f(x)$ unique? [uniqueness]
i.e. does $\int_{-\infty}^{\infty} x^r f(x)dx=\int_{-\infty}^{\infty} x^r g(x)dx \Rightarrow f(x)=g(x)$?

In general, the answer to both questions is no! Under what conditions moments exist?

Lemma 1. A useful result on the *existence* of the moments is the following. A sufficient (but certainly not necessary) condition is that the support of the random variable X is a bounded interval, i.e. $\mathbb{R}_X:=[a,b]$, where $-\infty<a<b<\infty$. In this case all moments exist:

$$\mu'_k=\int_a^b x^r f(x)dx<\infty, \text{ for all } k=1,2,...$$

When the range of values of the random variable in question is unbounded the moments do not always exist. A sufficient condition for the uniqueness problem is provided by lemma 2.

Lemma 2. The moments $\{\mu'_k<\infty,\ k=1,2,...\}$ of a continuous random variable X with cdf $F(x)>0$, $x\in\mathbb{R}$, will determine it *uniquely* if the *Carleman condition* (Stoyanov, 1987) holds: $\sum_{n=1}^{\infty}(\mu'_{2n})^{-\frac{1}{2n}}=\infty$.

Moments and distributions. In the context of statistical modeling, the road from moments $\left\{\mu'_k:=E(X^k)<\infty,\ k=1,2,\ldots\right\}$ to distributions $f(x)$, $x\in\mathbb{R}_X$, is treacherous! This is because moments do not usually determine distributions uniquely even if we use an *infinite* number of them; see Simon (1998).

Example 3.37. The classic example of the non-uniqueness problem is the case of the *Log-Normal* distribution with density:

$$\varphi(x)=\tfrac{1}{x\sqrt{2\pi}}\exp\left\{-\tfrac{1}{2}(\ln x)^2\right\},\ x\in\mathbb{R}_+. \tag{35}$$

It can be shown that it cannot be determined uniquely by infinite set of its moments. Heyde (1963) shows that the density function:

$$g(x)=\tfrac{1}{c}\varphi(x)\{1+\tfrac{1}{2}\sin(2\pi k)\ln(x)\},$$

where k is any positive integer and $c>0$ (normalization constant), has the same moments $\{\mu'_k:=E(X^k)<\infty,\ k=1,2,\ldots\}$ as (35).

Moment 'matching' can be very misleading!

Example 3.38. Consider the distribution as specified below (see Romano and Siegel, 1986).

x	$\sqrt{3}$	$-\sqrt{3}$	0
$f(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{4}{6}$

(36)

The distribution in (36) has moments which match the first five moments of $Z\backsim\mathsf{N}(0,1)$:

$$E(X)=\sqrt{3}(\tfrac{1}{6})-\sqrt{3}(\tfrac{1}{6})=0, \qquad E(X^2)=3(\tfrac{1}{6})+3(\tfrac{1}{6})=1,$$
$$E(X^3)=(\sqrt{3})^3(\tfrac{1}{6})-(\sqrt{3})^3(\tfrac{1}{6})=0, \quad E(X^4)=9(\tfrac{1}{6})+9(\tfrac{1}{6})=3.$$

This example might seem a bit extreme but it should serve as a cautionary note.

On the other hand, the road from distributions to moments is smooth and extremely useful. For instance, when we are prepared to limit ourselves to a specific distribution or even class of distributions the problem becomes tractable.

Example 3.39. (a) When one assumes that $X_k\backsim\mathsf{N}(\mu,\sigma^2)$, focusing on just the first two moments is fully justified because they characterize the Normal distribution!

(b) The Pearson family of distributions is characterized by the first four moments (see ch. 12).

It is important to bring out the fact that there are implicit assumptions when practitioners declare that they will focus their statistical modeling on the first two moments of a random variable X whose distribution they leave unspecified. This is because focusing on the first two moments is an indirect distributional assumption since it assumes the existence of these moments, and thus it excludes certain other distributions such as the Cauchy. Worse, such a strategy disregards the question 'why focus only on the first two moments?' Such a strategy makes sense when it is accompanied by an explicit distributional assumptions, such as the Normal, or other distributions that are characterized by the first two moments.

5.6 Other numerical characteristics

It is sometimes the case that for certain random variables, the moments discussed above do not make sense. For example, in the case where the random variable X denotes religion of a person: 1=Christian, 2=Muslim, 3=Jewish, 4=Buddhist, the mean and variance do not make much sense. In addition, sometimes the mean and variance do not exist, as in the case of the Cauchy distribution (see next section). In such cases we need to consider other numerical characteristics.

5.6.1 Measures of location

Mode. The mode or modal value m_0 is that particular value of the random variable X which corresponds to the maximum of the density function. In the case of a discrete distribution the mode is the value x at which $f(x)$ is maximum. In the continuous case where $f(x)$ is twice differentiable the mode can be derived as the solution of:

$$\left(\tfrac{df(x)}{dx}\right)=0, \quad \text{subject to} \quad \left(\tfrac{df^2(x)}{dx^2}\right)\Big|_{x=m_0}<0. \tag{37}$$

Note that a density function can have several local maxima. When the density function has a unique maximum it is said to be *unimodal*, otherwise it is call *multimodal*.

Example 3.40. (a) Since the natural logarithm is a monotonic function, the mode of the Beta distribution in (30) can be derived by locating the maximum of the log of the density function:

$$\ln f(x;\boldsymbol{\theta})=-\ln\left(B[\alpha,\beta]\right)+(\alpha-1)\ln x+(\beta-1)\ln(1-x),$$

$$\tfrac{d\ln f(x;\boldsymbol{\theta})}{dx}=\tfrac{(\alpha-1)}{x}-\tfrac{(\beta-1)}{(1-x)}=0\rightarrow m_0=\tfrac{(\alpha-1)}{(\alpha+\beta-2)},\ \text{for } \alpha>1,\ \beta>1,$$

$$\tfrac{d^2\ln f(x;\boldsymbol{\theta})}{dx^2}\Big|_{x=m_0}=-\tfrac{(\alpha-1)}{x^2}-\tfrac{(\beta-1)}{(1-x)^2}\Big|_{x=m_0}=-\tfrac{(\alpha+\beta-2)^3}{(\alpha-1)(\beta-1)}<0,\ \text{for } \alpha>1,\ \beta>1.$$

(b) For the density function given in (34) the mode is equal to 2.

(c) The Cauchy $\mathsf{C}\,(\alpha=0,\beta=1)$ distribution has no moments, but it does have a mode, as shown in 3.24.

Median. The median of a random variable X is that particular value $m(x)$ which divides the probability into two equal halves, i.e. $m(x)$ (assuming it is unique) such that:

$$\mathbb{P}(x\leq m(x))\geq 0.5 \text{ and } \mathbb{P}(x\geq m(x))\geq 0.5.$$

In the case where the cdf is continuous and strictly increasing, $m(x)$ is defined by:

$$F(m(x))=0.5 \ \text{ and } \ m(x) \text{ is unique.}$$

Example 3.41. For $X\backsim \mathsf{N}(\mu,\sigma^2)$ the three measures of location coincide:

$$\text{mean=median=mode.} \tag{38}$$

NOTES: (a) In general, the *symmetry* of $f(x)$ does not guarantee (38). The mean might not exist, but when it does, it is unique. The median always exists but it might might not be unique. The mode might not exist (e.g. the Cantor distribution; see Karr, 1993) or it might not be unique; $f(x)$ can be multimodal.

(b) In analogy to (37), the mean and median can be viewed as the resulting from minimizing two different loss functions, $E(X-\theta)^2$ and $E(|X-\theta|)$. In particular:

$$E(X)=\arg\min E(X-\theta)^2,\ \theta\in\mathbb{R},\ \text{when } E(X^2)<\infty,$$

(Schervish, 1995) and when X is continuous and $F(x)$ is strictly increasing:

$$m(x)=\arg\min E\left|X-\theta\right|,\ \theta\in\mathbb{R}.$$

Example 3.42. For the Cauchy $\mathsf{C}(\alpha,\beta)$: median=mode=α; see fig. 3.27.

Extending the concept of a median to values x_p for p in the interval $[0,1]$, we define the quantiles.

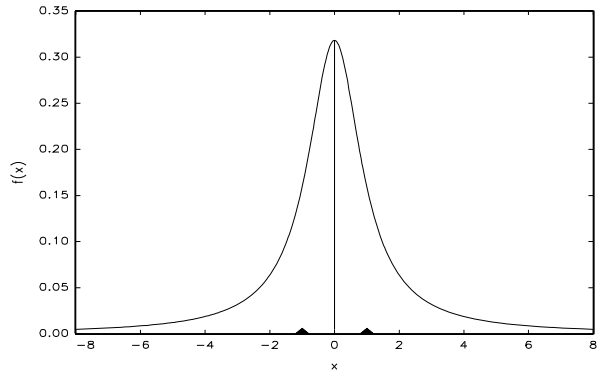

Fig. 3.27: Mode of the Cauchy density

5.6.2 Measures of dispersion

Range. The range is defined to be the difference between the largest and the smallest value taken by the random variable in question, i.e.

$$\mathrm{R}(X):=X_{\max}-X_{\min}.$$

Example 3.46. In the case of the Uniform distribution ($\mathsf{U}(a,b)$):

$$\mathrm{R}(X)=X_{\max}-X_{\min}=(b-a)$$

The Interquartile Range. It is defined to be the difference between the lower and upper quartiles:

$$\mathrm{IQR}\left(X\right):=\left(x_{.75}-x_{.25}\right).$$

The IQR(X) can be used to provide a standardization of a random variable X which is different from that based on the standard deviation ($SD(X)$) in the sense that:

$$Y=\left(\tfrac{X-m(x)}{\mathrm{IQR}(X)}\right)\rightarrow m(y)=0 \text{ and } \mathrm{IQR}\left(Y\right)=1.$$

Example 3.47. (a) In the case of the Normal distribution [$\mathsf{N}(\mu,\sigma^2)$]:

$$\mathrm{IQR}\left(X\right):=\left(x_{.75}-x_{.25}\right)=\mu+.6745\sigma-\mu+.6745\sigma=2(.6745)\sigma.$$

Figure 3.32 shows the Normal cdf for $\mathsf{N}(0,1)$ with the quantiles given in table 3.11.

Table 3.11: N(0,1) - quantiles			
q	x	$F(x)$	$f(x)$
$x_{.05}$	-1.645	.05	.103
$x_{.25}$	$-.6745$	.25	.318
$x_{.75}$	.6745	.75	.318
$x_{.95}$	1.645	.95	.103

In figure 3.33 we can see these quantiles in relation to the density function. NOTE that the maximum of the density function of $\mathsf{N}(0,1)$ is just $\left(\sqrt{2\pi}\right)^{-1} = .39894$.

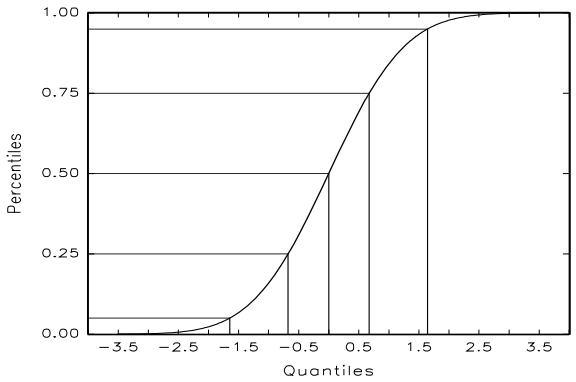

Fig. 3.32: Normal cdf: quantiles

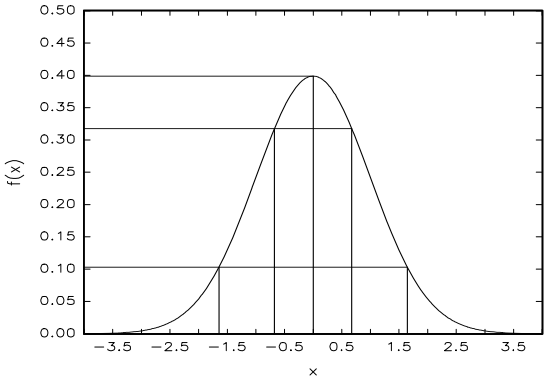

Fig. 3.33: Normal density: quantiles

(b) In the case of the Cauchy distribution considered above, we can easily see that:

$$\text{IQR}(X) = (\alpha+\beta) - (\alpha-\beta) = 2\beta.$$

This can be used as a measure of dispersion since the variance does not exist.

The quartile deviation is defined as half of the interquartile range i.e.

$$q(X) := \left(\tfrac{1}{2}\right)\left(x_{3/4} - x_{1/4}\right).$$

Example 3.48. (a) Normal $[\mathsf{N}(\mu,\sigma^2)]$: $q(X) := \left(\frac{1}{2}\right)\left(x_{3/4} - x_{1/4}\right) = (.6745)\sigma$.

(b) Cauchy $[\mathsf{C}(\alpha,\beta)]$: $q(X) := \left(\frac{1}{2}\right)\left(x_{3/4} - x_{1/4}\right) = \beta$.

The coefficient of variation, proposed by Pearson (1896), is defined to be the ratio of the standard deviation to the mean of the random variable in question, i.e.

$$\text{cv}(X) := \frac{\{Var(X)\}^{\frac{1}{2}}}{E(X)}.$$

6 Summary

In this chapter the abstract *probability space* $(S, \Im, \mathbb{P}(.))$ has been mapped onto the real line, where numerical data live. The end result is a probability model comprising a family of densities indexed by a small number of unknown parameters ($\boldsymbol{\theta}$):

$$\Phi=\{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta,\ x\in\mathbb{R}_X\}.$$

Its basic elements are: (i) a density function $f(x;\boldsymbol{\theta}),\ x\in\mathbb{R}_X$, where $\mathbb{R}_X:=\{x\in\mathbb{R}:\ f(x;\boldsymbol{\theta})>0\}$ denotes its support and (ii) a *parameter space* $\Theta\subset\mathbb{R}^p$-the set of all possible values of $\boldsymbol{\theta}$. Both elements play important roles in choosing an appropriate statistical model. As shown in chapter 5, the moments of the distribution are directly related to the distributional shapes taken by density functions. The relationship between the unknown *parameters* $\boldsymbol{\theta}$ of the probability model and the *moments* of the distribution in question is very important. The concepts introduced during this digression will prove indispensable for modeling and inference purposes. An important consideration in making a decision in relation to the appropriate model is the richness of the our choice menu. This is why in Appendix 3.A several important probability models are given for reference purposes.

In the next chapter we complete the transformation of the statistical space into a statistical model by mapping the abstract sampling space $\mathcal{G}_n^{\text{IID}}$ onto the real line as well. This gives rise to a set of random variables $\mathbf{X}:=(X_1, X_2, ..., X_n)$ with an IID probabilistic structure.

Important concepts

Random variable, density function, Bernoulli, Binomial, Poisson and Geometric (discrete) distributions, Borel (measurable) function, equality of two random variables, cumulative distribution function, Normal (Gaussian), Beta, Gamma, Exponential, Uniform, Student's t and Cauchy (continuous) distributions, parameters of a distribution, parameter space, moments of a distribution, distribution of a function of a random variable, probability integral transformation, mean and variance of a distribution, higher raw and central moments of a distribution, standard deviation, skewness and kurtosis coefficients, the problem of moments, mode and median of a distribution, quantiles, quantile function, probability model.

Crucial distinctions

Discrete vs. continuous random variables, sample vs. distribution moments, substantive (structural) vs. statistical parameters, mean vs. median vs. mode, probability space vs. probability model, Student's t vs. Normal distribution, the cumulative distribution function vs. the quantile function.

Essential ideas

- From a mathematical perspective, a random variable X is neither random nor a variable. It is a real-valued function $X(.)$: $S \to \mathbb{R}$ that preserves the events of interest and related events.

- The concept of a random variable $X(.)$ is used to transform the abstract probability space $(S, \Im, \mathbb{P}(.))$ into a probability model on the real line where numerical data live.
- In statistical modeling the moments of a distribution can provide effective ways to simplify both the modeling and the inference facets. Keep in mind, however, that the relationship from distributions to moments is the best way to proceed, because the reverse is easy to transgress.
- The meaningfulness of the numerical characteristics of a distribution $f(x)$, including its moments, depend crucially on the scale of measurement [nominal, ordinal, interval, ratio] of X.
- Just because two distributions have the same first few moments does not mean they are even similar!

7 Questions and Exercises

1. Explain why the abstract probability space is inappropriate for modeling purposes.

2. (a) "A *random variable* is neither random nor a variable". Discuss.

(b) "The concept of a random variable is a relative concept". Discuss.

(c) Explain the difference between the inverse and the pre-image of a function.

3. Consider the random experiment of casting two dice and counting the total number of dots appearing on the uppermost faces. The random variable X takes the value 0 when the total number of dots is odd and 1 when the total number of dots is even.

(a) Derive the density function of the random variable X assuming that the two dice are symmetric.

(b) Derive the density function of the random variable X assuming that the two dice are non-symmetric.

4. Discuss the difference between the following probability set functions in terms of their domain:

$$\mathbb{P}(X \leq x) = \mathbb{P}X^{-1}((-\infty, x]) = P((-\infty, x]).$$

5. In the case of the random experiment of "tossing a coin twice":

$$S = \{(HH), (HT), (TH), (TT)\}, \Im = \{S, \varnothing, A, \overline{A}\},$$

where $A = \{(HH), (HT), (TH)\}$. Consider the following functions:

(i) $X(HH)=1,\ X(HT)=2,\ X(TH)=2,\ X(TT)=1,$

(ii) $Y(HH)=1,\ Y(HT)=0,\ Y(TH)=0,\ Y(TT)=0,$

(iii) $Z(HH)=1,\ Z(HT)=1,\ Z(TH)=1,\ Z(TT)=7405926.$

(a) Which of the functions (i)-(iii) constitute random variables with respect to $\Im$?

(b) Compare the sigma-fields generated by each function to the event space of interest $\Im$, and relate the result to your answer in (a).

6. Compare and contrast the concepts of a discrete random variable and a continuous random variable.

7. Describe briefly the transformation of the probability space $(S, \Im, \mathbb{P}(.))$ into a *probability model* of the form: $\Phi=\{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta,\ x\in\mathbb{R}_X\}$.

Explain the relationship between the components of the probability space and the probability model.

8. Explain the main components of a generic *probability model* above.

9. Why do we care about the moments of a distribution? How do the moments provide a way to interpret the unknown parameters?

10. For the Exponential distribution the density function is: $f(x;\theta)=\theta e^{-\theta x},\ \theta>0,\ x>0$.

(a) Derive its mean and variance. (b) Derive its mode.

11. Consider the function: $f(x)=140\,[x^3(1-x)^3]\,,\ 0<x<1$.

(a) Show that this is indeed a proper density function for a random variable X.
(b) Derive the mean, mode, variance and kurtosis of X.

12. Consider the function: $f(x)=\frac{x}{2},\ 0<x<2$.

(a) Show that this is a proper density function for a random variable X.
(b) Derive the mean and variance of X.

13. Consider the discrete random variable X whose distribution is given below:

x	-1	0	1
$f(x)$	0.2	0.4	0.4

(a) Derive its mean, variance, skewness and kurtosis coefficients.
(b) Derive its mode and coefficient of variation.

14. (a) State the properties of a density function.
(b) Contrast the properties of the expected value and variance operators.
(c) Let X_1 and X_2 be two Independent random variables with the same mean μ and variance σ^2. Derive the mean and variance of the function: $Y=\frac{1}{3}X_1+\frac{2}{3}X_2$.

15. Explain how the properties of the variance are actually determined by those of the mean operator.

16. Explain how the moment generating function can be used to derive the moments.

17. Explain the concept of skewness and discuss why $\alpha_3=0$ does not imply that the distribution in question is symmetric.

18. Explain the concept of kurtosis and discuss why it is of limited value when the distribution is non-symmetric.

19. For a Weibull distribution with parameters (α=3.345, β=3.45) derive the kurtosis coefficient using the formulae in appendix 3.A.

20. Explain why matching moments between two distributions can lead to misleading conclusions.

21. Compare and contrast the cumulative distribution function (cdf) and the quantile function.

22. Explain the concepts of a percentile and a quantile and how they are related.

23. Why do we care about probabilistic inequalities?

24. "Moments do not characterize distributions in general and when they do we often need an infinite number of moments for the characterization". Discuss.

25. Explain the probability integral and the probability integral transformations. How useful can they be in simulating non-uniform random variables?