Summer Seminar: Philosophy of Statistics
Lecture Notes 4: A Simple Statistical Model

Aris Spanos [Summer 2019]

1 Introduction

1.1 The story so far, a summary

Chapter 2 initiated the formalization of a *simple chance mechanism* known as a *random experiment* $\mathcal{E}$ into a *simple statistical space*:

$$\left[(S,\Im,\mathbb{P}(.))^n,\ \mathcal{G}_n^{\text{IID}}=\{\mathcal{A}_1,\mathcal{A}_2,\mathcal{A}_3,...,\mathcal{A}_n\}\right]$$

In light of the fact that numerical data live on the real line, the concept of a random variable $X(.)$ was used to transform $(S,\Im,\mathbb{P}(.))$ into a probability model:

Probability Space $(S,\Im,\mathbb{P}(.))$	$\rightsquigarrow$	Probability model $\Phi=\{f(x;\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta,\ x\in\mathbb{R}_X\}$

Φ denotes a family of density functions $f(x;\boldsymbol{\theta})$, indexed by $\boldsymbol{\theta}$ in Θ.

The **primary objective** of this chapter is to map the statistical space onto the real line by defining the sampling model:

Sampling space $\mathcal{G}_n^{\text{IID}}=\{\mathcal{A}_1,\mathcal{A}_2,\mathcal{A}_3,...,\mathcal{A}_n\}$	$\rightarrow$	Sampling model $\mathbf{X}:=(X_1,X_2,X_3,...,X_n).$

The transformation involves two important concepts in probability theory: Independence and Identical Distribution (IID). The resulting sampling model will, when combined with the probability model gives rise to a *simple statistical model*.

1.2 From random trials to a random sample: a first view

As argued in chapter 2 a simple sampling space $\mathcal{G}_n^{\text{IID}}:=\{\mathcal{A}_1,\mathcal{A}_2,...,\mathcal{A}_n\}$ is a set of *random trials,* which are both:

$$\textbf{Independent (I):}\quad \mathbb{P}_{(n)}(\mathcal{A}_1\cap\mathcal{A}_2\cap\cdots\cap\mathcal{A}_k)=\textstyle\prod_{i=1}^k\mathbb{P}_i(\mathcal{A}_i),\ \text{for } k=2,3,...,n, \tag{1}$$

$$\textbf{Identically Distributed (ID):}\quad \mathbb{P}_1(.)=\mathbb{P}_2(.)=\cdots=\mathbb{P}_n(.)=\mathbb{P}(.). \tag{2}$$

Independence is related to the condition that 'the outcome of any one trial *does not change* the probability of any other trial', or formally:

$$\mathbb{P}_{(n)}(\mathcal{A}_k|\mathcal{A}_1,\mathcal{A}_2,...\mathcal{A}_{k-1},\mathcal{A}_{k+1},...,\mathcal{A}_n)=\mathbb{P}_k(\mathcal{A}_k),\ \text{for } k=1,2,...,n. \tag{3}$$

The second pertains to "keeping the same probabilistic setup from one trial to the next", ensuring that the events and probabilities associated with the different outcomes remain the same for all trials.

Having introduced the concept of a *random variable* in chapter 3, it is natural to map $\mathcal{G}_n^{\text{IID}}$ onto the real line to transform the trials $\{\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_n\}$ into a set of random variables $\mathbf{X}:=(X_1, X_2, ..., X_n)$. The set function $\mathbb{P}_{(n)}(.)$ will be transformed into the *joint distribution function* $f(x_1, x_2, ..., x_n)$. Using these two concepts we can define the concept of a *random sample* $\mathbf{X}$ to be a set of Independent and Identically Distributed (IID) random variables. The basic new concept needed for the formalization is that of a *joint distribution function*.

A bird's eye view of the chapter. In section 2 we introduce the concept of a joint distribution using the simple bivariate case for expositional purposes. In section 3 we relate the concept of the joint distribution to that of the marginal (univariate) distribution. Section 4 introduces the concept of conditioning and conditional distributions as it relates to both the joint and marginal distributions. In section 5 we define the concept of *independence* using the relationship between the joint, marginal and conditional distributions. In section 6 we define the concept of *identically distributed* in terms of the joint and marginal distributions and proceed to define the concept of a random sample in section. In section 7 we introduce the concept of a *function of random variables* and its distribution with the emphasis placed on applications to the concept of an *ordered* random sample. Section 8 completes the transformation of a simple *statistical space* into a simple *statistical model*.

2 Joint distributions of random variables

The concept of a joint distribution is undoubtedly one of the most important concepts in both probability theory and statistical inference. As in the case of a single random variable the discussion will proceed to introduce the concept from the simple to the more general case. In this context simple refers to the case of *countable* outcomes sets which give rise to *discrete* random variables. After we introduce the basic ideas in this simplified context we proceed to discuss them in their full generality.

2.1 Joint distributions of discrete random variables

In order to understand the concept of a set of random variables (a *random vector*) we consider first the two random variable case since the extension of the ideas to n random variables is simple in principle, but complicated in terms of notation.

Random vector. Consider the two *simple random variables* (random variables) $X(.)$ and $Y(.)$ defined on the same probability space $(S, \Im, \mathbb{P}(.))$, i.e.

$$X(.)\colon S \to \mathbb{R}, \text{ such that } X^{-1}(x) \in \Im, \text{ for all } x \in \mathbb{R},$$

$$Y(.)\colon S \to \mathbb{R}, \text{ such that } Y^{-1}(y) \in \Im, \text{ for all } y \in \mathbb{R}.$$

REMARK: recall that $Y^{-1}(y)=\{s\colon Y(s)=y,\ s \in S\}$ denotes the *pre-image* of the function $Y(.)$ and not its inverse. Viewing them separately we can define their individual density functions, as explained in the previous chapter, as follows:

$$\mathbb{P}(s\colon X(s)=x)=f_x(x)>0,\ x \in \mathbb{R}_X,\ \mathbb{P}(s\colon Y(s)=y)=f_y(y)>0,\ y \in \mathbb{R}_Y,$$

where $\mathbb{R}_X$ and $\mathbb{R}_Y$ denote the *support* of the density functions of X and Y. Viewing them together we can think of each pair $(x,y)\in\mathbb{R}_X\times\mathbb{R}_Y$ as events of the form:

$$\{s\colon X(s)=x,\ Y(s)=y\}:=\{s\colon X(s)=x\}\cap\{s\colon Y(s)=y\},\ (x,y)\in\mathbb{R}_X\times\mathbb{R}_Y.$$

In view of the fact the event space $\Im$ is a σ-field, and thus close under intersections, the mapping:

$$\mathbf{Z}(.,.):=(X(.),Y(.))\colon S\to\mathbb{R}^2,$$

is a *random vector* since the pre-image of $\mathbf{Z}(.)$ belongs to the event space $\Im$:

$$\mathbf{Z}^{-1}(x,y)=[(X^{-1}(x))\cap(Y^{-1}(y))]\in\Im,$$

since, by definition, $X^{-1}(x)\in\Im$ and $Y^{-1}(y)\in\Im$ (being a σ-field; see chapter 3).

Joint density. The joint density function is defined by:

$$f(.,.)\colon \mathbb{R}_X\times\mathbb{R}_Y\to[0,1],$$
$$f(x,y)=\mathbb{P}\{s\colon X(s)=x,\ Y(s)=y\},\ (x,y)\in\mathbb{R}_X\times\mathbb{R}_Y.$$

Example 4.1. Consider the case of the random experiment of tossing a fair coin twice, giving rise to the set of outcomes: $S=\{(HH),(HT),(TH),(TT)\}$.

Let us define the random variables $X(.)$ and $Y(.)$ on S as follows:

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

We can construct the individual density functions as follows:

x	0	1
$f(x)$	.25	.75

y	0	1
$f(y)$	.25	.75

(4)

To define the joint density function we need to specify all the events:

$$(X=x,Y=y),\ x\in\mathbb{R}_X,\ y\in\mathbb{R}_Y,$$

and then attach probabilities to these events. In view of the fact that:

$$\begin{aligned}(X=0,Y=0)=\{\ \}=\emptyset &\to f(x=0,y=0)=.00,\\ (X=0,Y=1)=\{(TT)\} &\to f(x=0,y=1)=.25,\\ (X=1,Y=0)=\{(HH)\} &\to f(x=1,x=0)=.25,\\ (X=1,Y=1)=\{(HT),(TH)\} &\to f(x=1,y=1)=.50.\end{aligned}$$

That is, the joint density takes the form:

$y\diagdown x$	0	1
0	.00	.25
1	.25	.50

(5)

If we compare this joint density (5) with the univariate densities (4), there is no obvious relationship. As argued in the next chapter, however, the difference between the joint probabilities $f(x,y), x{\in}\mathbb{R}_X,\ y{\in}\mathbb{R}_Y$, and the product of the individual probabilities $(f(x){\cdot}f(y))$ for $x{\in}\mathbb{R}_X,\ y{\in}\mathbb{R}_Y$, reflects the dependence between the random variables X and Y. At this stage it is crucial to note that a most important feature of the joint density function $f(x,y)$, is that it provides a general description of the dependence between X and Y.

Before we proceed to consider the continuous random variables case it is instructive to consider a particularly simple case of a bivariate discrete density function.

Example 4.2. The previous example is a particular case of a well known discrete joint distribution, the *Bernoulli* distribution given below:

Table 4.1: Bernoulli density

$y \diagdown x$	0	1
0	$p(0,0)$	$p(1,0)$
1	$p(0,1)$	$p(1,1)$

(6)

where $p(i,j)$ denotes the joint probability for $X{=}i$ and $Y{=}j$, $i,j{=}0,1$. The Bernoulli joint density takes the form:

$$f(x,y){=}p(0,0)^{(1-y)(1-x)}p(0,1)^{(1-y)x}p(1,0)^{y(1-x)}p(1,1)^{xy},\ x{=}0,1,\ \ y{=}0,1.$$

2.2 Joint distributions of continuous random variables

In the case where the outcomes set S is *uncountable,* the random variables defined on it are said to be **continuous** because their range of values is a piece of the real line $\mathbb{R}$.

Random vector. Consider the two continuous random variables (random variables) $X(.)$ and $Y(.)$ defined on the same probability space $(S,\Im,\mathbb{P}(.))$, i.e.

$$X(.)\colon\ S\to\mathbb{R},\ \text{such that}\ X^{-1}((-\infty,x])\in\Im,\ \text{for all}\ x{\in}\mathbb{R},$$

$$Y(.)\colon\ S\to\mathbb{R},\ \text{such that}\ Y^{-1}((-\infty,y])\in\Im,\ \text{for all}\ y{\in}\mathbb{R}.$$

Viewing them separately we can define their individual cumulative distribution functions (cdf) (see chapter 3), as follows:

$$\mathbb{P}(s\colon\ X(s){\leq}x){=}\mathbb{P}\left(X^{-1}(-\infty,x]\right){=}P_X\left((-\infty,x]\right){=}F_X(x),\ x{\in}\mathbb{R},$$

$$\mathbb{P}(s\colon\ Y(s){\leq}y){=}\mathbb{P}\left(Y^{-1}(-\infty,y]\right){=}P_Y\left((-\infty,y]\right)\ {=}F_Y(y),\ y{\in}\mathbb{R}.$$

Viewing them together we can associate with each pair $(x,y){\in}\mathbb{R}\times\mathbb{R}$ events of the form:

$$\{s\colon\ X(s){\leq}x,\ Y(s){\leq}y\}{:=}\{s\colon\ X(s){\leq}x\}{\cap}\{s\colon\ Y(s){\leq}y\},\ (x,y){\in}\mathbb{R}\times\mathbb{R}.$$

As in the discrete random variable case, since $\Im$ is a $\sigma-$field (close under intersections) the mapping:

$$\mathbf{Z}(.,.){:=}\left(X(.),Y(.)\right)\colon\ S\to\mathbb{R}^2,$$

constitutes a **random vector**; the pre-image of $\mathbf{Z}(.)$:

$$\mathbf{Z}^{-1}\left((-\infty,x]\times(-\infty,y]\right)=\left[(X^{-1}((-\infty,x]))\cap(Y^{-1}((-\infty,y]))\right]\in\Im,$$

since $X^{-1}((-\infty,x])\in\Im$ and $Y^{-1}((-\infty,y])\in\Im$ by definition.

The **joint cumulative distribution function** (cdf) is defined by:

$$F_{XY}(.,.)\colon \mathbb{R}^2\rightarrow[0,1],$$

$$F_{XY}(x,y)=\mathbb{P}\{s\colon X(s)\leq x,\ Y(s)\leq y\}=P_{XY}((-\infty,x]\times(-\infty,y]),\ (x,y)\in\mathbb{R}^2.$$

The joint cdf can also be defined on intervals of the form $(a,b]$ taking the form:

$$\mathbb{P}\{s\colon x_1<X(s)\leq x_2,\ y_1<Y(s)\leq y_2\}=F(x_2,y_2)-F(x_1,y_2)-F(x_2,y_1)+F(x_1,y_1).$$

The **joint density function**, assuming that $f(x,y)\geq 0$ exists, is defined via:

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(u,v)\,du\,dv.$$

NOTE that the subscripts will often be omitted when there is no possibility of confusion. In the case where $F(x,y)$ is differentiable at (x,y) we can derive the joint density by partial differentiation:

$$f(x,y)=\left(\frac{\partial F^2(x,y)}{\partial x\partial y}\right)\ \text{at all continuity points of } f(x,y).$$

Example 4.3. Let the joint cdf be that of the bivariate Exponential distribution:

$$F(x,y)=1-e^{-x}-e^{-y}+e^{-x-y}\rightarrow\quad f(x,y)=\left(\frac{\partial F^2(x,y)}{\partial x\partial y}\right)=e^{-x-y},\ x\geq 0,\ y\geq 0.$$

In the case of *continuous* random variables we can think of the joint density as being defined over an interval of the form $(x<X\leq x+dx, y<Y\leq y+dy)$ as follows:

$$\mathbb{P}(x<X\leq x+dx,\ y<Y\leq y+dy)=f(x,y)dx\,dy.$$

Hence, as in the univariate case (see chapter 3), the joint density function takes values greater than one, i.e.

$$f(.,.)\colon \mathbb{R}\times\mathbb{R}\rightarrow[0,\infty).$$

In direct analogy to the univariate case, $f(x,y)$ satisfies similar properties (table 4.2).

Table 4.2: Joint density function - Properties

[bf1	$f(x,y)\geq 0,$ for all $(x,y)\in\mathbb{R}_X\times\mathbb{R}_Y,$
[bf2]	$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)dxdy=1,$
[bf3]	$F_{XY}(a,b)=\int_{-\infty}^{a}\int_{-\infty}^{b}f(x,y)\,dx\,dy,$
[bf4]	$f(x,y)=\left(\frac{\partial F^2(x,y)}{\partial x\partial y}\right),$ at all continuity points of $f(x,y).$

NOTE: in the discrete case the above integrals become summations over all values of X and Y, i.e., for $x_1<x_2<\cdots<x_n<\cdots$ and $y_1<y_2<\cdots<y_n<\cdots$:

$$\textbf{[bf2]}' \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} f(x_i,y_j)=1, \quad \textbf{[bf3]}' \; F(x_k,y_m)=\sum_{i=1}^{k}\sum_{i=1}^{m} f(x_i,y_j).$$

Example 4.4. An important *discrete* bivariate distribution, is the *Binomial*, (or trinomial as often called) whose density takes the form:

$$f(x,y;\boldsymbol{\theta})=\left(\tfrac{n!}{x!y!(n-x-y)!}\right)\theta_1^x\theta_2^y(1-\theta_1-\theta_2)^{n-x-y},\ \theta_i\in[0,1],\ i=1,2,$$

here $\boldsymbol{\theta}:=(\theta_1,\theta_2)$, n is an integer such that $x+y\leq n,\ x,y=0,1,2,...n$.

Example 4.5. The most important *continuous* bivariate distribution is the *Normal*, whose density is:

$$f(x,y;\boldsymbol{\theta})= \tfrac{(1-\rho^2)^{-\frac{1}{2}}}{2\pi\sqrt{\sigma_{11}\sigma_{22}}}\exp\left\{-\tfrac{1}{2(1-\rho^2)}\left(\left(\tfrac{y-\mu_1}{\sqrt{\sigma_{11}}}\right)^2-2\rho\left(\tfrac{y-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\tfrac{x-\mu_2}{\sqrt{\sigma_{22}}}\right)+\left(\tfrac{x-\mu_2}{\sqrt{\sigma_{22}}}\right)^2\right)\right\}, \quad (7)$$

where $\boldsymbol{\theta}:=(\mu_1,\mu_2,\sigma_{11},\sigma_{22},\rho)\in\mathbb{R}^2\times\mathbb{R}^2_+\times[-1,1]$, $x\in\mathbb{R}$, $y\in\mathbb{R}$. In view of its apparent complexity, the bivariate density given in (7), is often denoted by:

$$\begin{pmatrix} Y \\ X \end{pmatrix} \backsim \mathsf{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right), \quad (8)$$

where $\sigma_{12}:=\rho\sqrt{\sigma_{11}\sigma_{22}}$. A special case of this distribution, known as the **standard bivariate Normal**, is defined by the special values of the parameters: $\mu_1=\mu_2=0,\ \sigma_{11}=\sigma_{22}=1$. Its density function takes the form:

$$f(x,y;\boldsymbol{\theta})=\tfrac{1}{2\pi\sqrt{(1-\rho^2)}}\exp\left\{-\tfrac{1}{2(1-\rho^2)}\left[x^2-2\rho xy+y^2\right]\right\}; \quad (9)$$

$f(x,y;\boldsymbol{\theta})$ with $\boldsymbol{\theta}:=(0,0,1,1,0.2)$ is shown in figure 4.1. The details of the bell shape of the surface can be seen from the inserted contours which can be viewed intuitively as lines we get by slicing the surface at different heights.

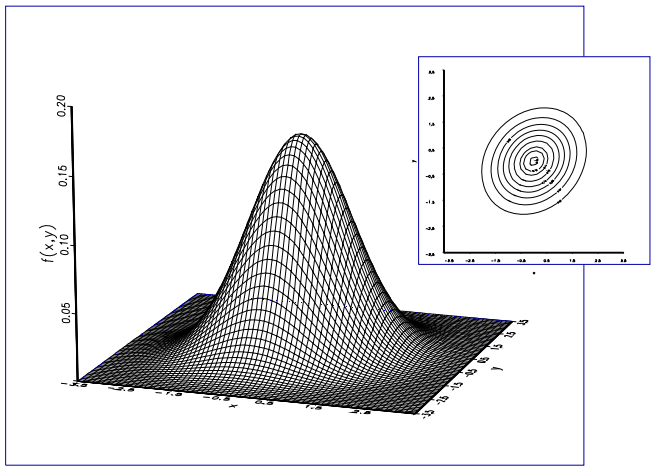


Fig. 4.1: Bivariate Normal density surface with contours inserted

Several additional bivariate distributions are listed in appendix 4.A.

2.3 Joint moments of random variables

As in the case of univariate distributions the best way to interpret the unknown parameters is via the moments. In direct analogy to the univariate case, we define **the joint product moments** of order (k, m) by:

$$\mu'_{km}=E\{X^k Y^m\},\ k, m=0, 1, 2, ..,$$

and the **joint central moments of order** (k, m) is defined by:

$$\mu_{km}=E\{(X - E(X))^k (Y - E(Y))^m\},\ k, m=0, 1, 2, ...$$

The first two joint product and central moments are:

$$\begin{array}{ll} \mu'_{10}=E(X),\ \mu'_{01}=E(Y), & \mu_{10}=0,\ \mu_{01}=0, \\ \mu'_{20}=E(X)^2 + Var(X), & \mu_{20}=Var(X), \\ \mu'_{02}=E(Y)^2 + Var(Y), & \mu_{02}=Var(Y), \\ \mu'_{11}=E(X\cdot Y). & \mu_{11}=E\left[(X-E(X))\,(Y-E(Y))\right]. \end{array}$$

The most important and widely used joint moment is the **covariance**, defined by:

$$\mu_{11}:=Cov(X, Y)=E\{[X - E(X)][Y - E(Y)]\}. \tag{10}$$

Example 4.6. Consider the joint Normal distribution whose density is given in (7). We know from chapter 3 that the parameters $(\mu_1, \mu_2, \sigma_{11}, \sigma_{22})$ correspond to the moments:

$$\mu_1=E(Y),\ \ \mu_2=E(X),\ \ \sigma_{11}=Var(Y),\ \ \sigma_{22}=Var(X).$$

The additional parameter σ_{12} turns out to be the covariance between the two random variables, i.e. $\sigma_{12}:=Cov(X, Y)$.

Example 4.7. Let us derive the covariance of X and Y, using the joint density given below:

$y \diagdown x$	0	1	2	$f_y(y)$
0	0.2	0.2	0.2	0.6
2	0.1	0.1	0.2	0.4
$f_x(x)$	0.3	0.3	0.4	1

(11)

First, we need to derive the *moments* of the univariate distributions:

$$\begin{array}{rl} E(X) & =(0)(.3)+(1)(.3)+(2)(.4)=1.1,\ E(Y)=(0)(.6)+(2)(.4)=.8, \\ Var(X) & =[0 - 1.1]^2(.3) + [1 - 1.1]^2(.3) + [2 - 1.1]^2(.4)=.69, \\ Var(Y) & =[0 - .8]^2(.6) + [2 - .8]^2(.4)=.96. \end{array}$$

Using these moments we proceed to derive the covariance:

$$\begin{array}{rl} Cov(X, Y) & =E\{[X-E(X)][Y-E(Y)]=[0-1.1][0 - .8](.2)+ \\ & +[0-1.1][2-.8](.1)+[1-1.1][0-.8](.2)+[1-1.1][2-.8](.1)+ \\ & +[2-1.1][0-.8](.2)+[2-1.1][2-.8](.2)=.12. \end{array}$$

Let us verify (10). In view of the fact that:

$E(X\cdot Y)=\ \ (0)(0)(.2)+(0)(2)(.1)+(1)(0)(.2)+(1)(2)(.1)+(2)(0)(.2)+(2)(2)(.2)=1.0,$

we can conclude that: $Cov(X,Y)=\ \ 1.0-(1.1)(.8)=.12,$ confirming the above value.

Table 4.3: Covariance - Properties

C1.	$Cov(X,Y)=E(XY)-E(X)\cdot E(Y)$,
C2.	$Cov(X,Y)=Cov(Y,X)$,
C3.	$Cov(aX+bY,Z)=aCov(X,Z)+bCov(Y,Z)$, for $(a,b)\in\mathbb{R}^2$.
C4.	$Cov(X,X)=Var(X)$,
C5.	$Cov(X,Y)=0$, when X and Y are independent.

The **C1** property of the coveraince (table 4.3) shows the relationship between the raw and central joint moments for $k=m=1$. The covariance is equal to the first **joint product moment** $E(XY)$ minus the product of the two means. The second property refers to the symmetry of the covariance with respect to the two random variables involved. The third property follows directly from the linearity of the expectation operator $E(.)$.

Correlation coefficient. For any two random variables X and Y such that $Var(X)<\infty,\ Var(Y)<\infty$, defined on the same probability space $(S,\Im,\mathbb{P}(.))$, the *correlation coefficient* is defined by:

$$Corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}}.$$

Example 6.6. For the joint distribution in example 4.16, let us derive the correlation coefficient between X and Y, given:

$$E(X)=1.1,\ E(Y)=.8,\ \ Var(X)=.69,\ Var(Y)=.96,\ Cov(X,Y)=.12.$$

The correlation coefficient is: $Corr(X,Y)=\frac{0.12}{\sqrt{(.69)\cdot(.96)}}=.147.$

Table 6.3: Correlation coefficient - Properties

$\boldsymbol{\rho}$**1**.	$-1\leq Corr(X,Y)\leq 1$,
$\boldsymbol{\rho}$**2**.	$Corr(aX+b,cY+d)=Corr(X,Y)$, for $(a,b,c,d)\in\mathbb{R}^4,\ (a\cdot c)>0$,
$\boldsymbol{\rho}$**3**.	$Corr(X,Y)=\pm 1$, if and only if $Y=a_0+a_1X,\ (a_0,a_1)\in\mathbb{R}^2$.

The first property relating to the range of values for the correlation coefficient follows from the *Schwarz inequality* (Appendix 9.A):

$$|Cov(X,Y)|\leq [Var(X)]^{\frac{1}{2}}\,[Var(Y)]^{\frac{1}{2}}\,.$$

The second property follows from the definition of the correlation coefficient which renders it invariant to linear transformations. Proving the third property is rather involved, and thus omitted. It does, however, bring out the fact that correlation is a measure of linear dependence as the following example attests.

Example 6.7. Let X be uniformly distributed between minus one and plus one, denoted by

$$X \backsim \mathsf{U}(-1,1), \text{ and } Y{:=}X^2.$$

That is, X and Y are perfectly dependent (but non-linearly); knowledge of one determines the other completely. We can show, however, that the two are *uncorrelated*. In view of the fact that $f_x(x){=}\frac{1}{2}$ and $E(X){=}0$:

$$Cov(X,Y){=}E(XY){-}E(X){\cdot}E(Y){=}E(X^3){-}E(X){\cdot}E(X^2),$$

Hence, X and Y are uncorrelated if $E(X^3){=}0$:

$$E(X^3){=}\textstyle\int_{-1}^{1} x^3\left(\frac{1}{2}\right)dx{=}\frac{1}{2}\left[\left(\frac{1}{4}\right)x^4\Big|_{-1}^{1}\right]{=}\frac{1}{2}\left[\left(\frac{1}{4}\right)-\left(\frac{1}{4}\right)\right]{=}0.$$

Hence, the general conclusion we can draw about the discussion is that:

$$\textbf{independence} \;\Rightarrow\; \textbf{non-correlation}$$

but the converse is not true: **non-correlation** $\not\Rightarrow$ **independence**

2.4 The n-random variables joint distribution

Extending the concept of a random variable from a 2-dimensional to an n-dimensional random vector $\mathbf{X}(.){:=}(X_1(.),X_2(.),...,X_n(.))$ is straight forward:

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

where $\mathbb{R}^n{:=}\mathbb{R}{\times}\mathbb{R}{\times}\cdots{\times}\mathbb{R}$ denotes the Cartesian product of the real line (chapter 2).

The n-variable function $\mathbf{X}(.)$ is said to be a *random vector* relative to $\Im$ if:

$$\mathbf{X}(.)\colon S \to \mathbb{R}^n, \text{ such that } \mathbf{X}^{-1}((-\infty,\mathbf{x}]){\in}\Im, \text{ for all } \mathbf{x}{\in}\mathbb{R}^n_X,$$

where $\mathbf{x}{:=}(x_1,x_2,...,x_n)$ and $(-\infty,\mathbf{x}]{:=}(-\infty,x_1]{\times}(-\infty,x_2]{\times}\cdots{\times}(-\infty,x_n]$. NOTE that all the random variables $(X_1(.),X_2(.),...,X_n(.))$ are defined on the same outcomes set S and relative to the same event space $\Im$.

In view of the fact that $\Im$ is a σ-field we know that $\mathbf{X}{:=}(X_1,X_2,...,X_n)$ is a random vector relative to $\Im$ if and only if the random variables $(X_1,X_2,...,X_n)$ are random variables relative to $\Im$. This is because $X_k^{-1}(-\infty,x_k]{\in}\Im$ for all $k{=}1,2,...,n$, and so does their intersection:

$$\left(\textstyle\bigcap_{k=1}^{n} X_k^{-1}(-\infty,x_k]\right){\in}\Im.$$

The above concepts introduced above for the two random variable case can be easily extended to the n-random variable case, that satisfies similar properties as shown in table 4.4.

Table 4.4: Multivariate density - properties

[mf1]	$f(x_1, x_2, ..., x_n) \geq 0$, for all $(x_1, x_2, ..., x_n) \in \mathbb{R}^n_X$,
[mf2]	$\int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f(x_1, x_2, ..., x_n) dx_1 dx_2 \cdots dx_n = 1$,
[mf3]	$F(x_1, x_2, ..., x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f(u_1, u_2, ..., u_n)\, du_1 du_2 \cdots du_n$.

3 Marginal distributions

The second component of condition [c], relating to the independence of the trials is defined in terms of a simple relationship between the joint density function $f(x_1, x_2, ..., x_n; \boldsymbol{\phi})$ and the density functions of the individual random variables $X_1, X_2, ..., X_n$, referred to as the *marginal distributions*. Let us see how the marginal is related to the joint distribution.

It should come as no surprise to learn that from the joint distribution one can always recover the **marginal** (univariate) **distributions** of the individual random variables involved. In terms of the joint cdf, the marginal distribution is derived via a limiting process:

$$F_X(x) = \lim_{y \to \infty} F(x, y) \quad \text{and} \quad F_Y(y) = \lim_{x \to \infty} F(x, y).$$

Example 4.11. Let us consider the case of the bivariate exponential cdf:

$$F(x, y) = F(x, y) = (1 - e^{-\alpha x}) \left(1 - e^{-\beta y}\right), \ \alpha > 0, \ \beta > 0, \ x > 0, \ y > 0.$$

Given that $\lim_{n \to \infty} (e^{-n}) = e^{-\infty} = 0$, we can deduce that:

$$F_X(x) = \lim_{y \to \infty} F(x, y) = 1 - e^{-\alpha x}, \ x > 0, \quad F_Y(y) = \lim_{x \to \infty} F(x, y) = 1 - e^{-\beta y}, \ y > 0.$$

Let us see how the marginalization is defined in terms of the density functions. In view of the fact that:

$$F_X(x) = \lim_{y \to \infty} F(x, y) = \lim_{y \to \infty} \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y) dy dx = \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} f(x, y) dy \right] dx,$$

and the relationship between $F_X(x)$ and $f_x(x)$, we can deduce that:

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \ x \in \mathbb{R}_X. \tag{12}$$

Similarly, in terms of the joint density function, the marginal density function of Y is derived via:

$$f_y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx, \ y \in \mathbb{R}_Y. \tag{13}$$

That is, *marginalization* amounts to integrating out the other random variable.

Example 4.12. Let us consider the case of the bivariate exponential density:

$$f(x,y)=e^{-x-y},\ x>0,\ y>0,$$

where the random variables X and Y are continuous. The formula in (12) suggests that to derive the marginal distribution of X, one needs to integrate out the random variable Y from $f(x,y)$:

$$f_x(x)=\textstyle\int_0^\infty e^{-x-y}dy=e^{-x}.$$

Example 4.13. Consider the bivariate standard Normal density (9). In order to derive the marginal density of X, we need to integrate out Y, and vice versa. The manipulations for such a derivation are rather involved (and thus omitted) but the result is particularly useful. It turns out that:

$$f_x(x)=\int\limits_{-\infty}^{\infty} f(x,y)\,dy=\tfrac{1}{\sqrt{2\pi}}\exp\{-\tfrac{1}{2}x^2\},\quad f_y(y)=\int\limits_{-\infty}^{\infty} f(x,y)\,dx=\tfrac{1}{\sqrt{2\pi}}\exp\{-\tfrac{1}{2}y^2\}.$$

That is, both marginal distributions are (standard) Normal: $X\backsim\mathsf{N}(0,1)$ and $Y\backsim\mathsf{N}(0,1)$.

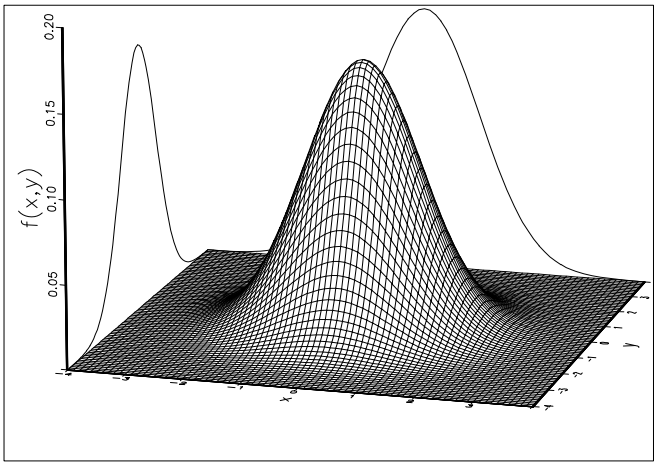


Fig. 4.2: Bivariate Normal density with projected marginal densities

Marginalization. We can visualize the derivation of the marginal distribution of X, from the bivariate distribution $f(x,y)$, as *projecting* the bivariate surface into the $[x,f(x,y)]$ plane. As shown in figure 4.2, projecting a bell-shaped surface onto a plane opposite yields a bell-shape for both marginal distributions. Intuitively, going from the joint to the marginal density amounts to ignoring the information relating to the particular dimension represented by the random variable integrated out.

In the **discrete random variable case,** we can derive the marginal distribution of one random variable, from the joint density $f(x,y)$, by *summing out* the other random variable For example, the derivation of the marginal density of X takes the form of summing over all the values of Y, say $y_1<y_2<y_3<\cdots<y_n<\cdots$, as follows:

$$f_x(x)=\textstyle\sum_{i=1}^{\infty} f(x,y_i),\quad x\in\mathbb{R}_X. \tag{14}$$

Similarly, the marginal density of Y takes the form of summing over all the values of X, say $x_1<x_2<x_3<\cdots<x_n<\cdots$:

$$f_y(y)=\sum_{i=1}^{\infty} f(x_i,y), \quad y\in\mathbb{R}_Y. \tag{15}$$

Example 4.14. The joint density of the Bernoulli distribution is well defined, if the probabilities $p(i,j)$ for $i,j=0,1$, in addition to being non-negative also satisfy certain additional restrictions as required by the marginal distributions. The marginal distributions of X and Y are given below:

x	0	1
$f_x(x)$	$p_{\cdot 1}$	$p_{\cdot 2}$

y	0	1
$f_y(y)$	$p_{1\cdot}$	$p_{2\cdot}$

(16)

$$p_{\cdot 1}=p(0,0)+p(0,1), \qquad p_{1\cdot}=p(0,0)+p(1,0),$$
$$p_{\cdot 2}=p(1,0)+p(1,1). \qquad p_{2\cdot}=p(0,1)+p(1,1).$$

For these marginal distributions to make sense they need to satisfy the properties of univariate density functions [**f1**]-[**f3**] (see chapter 3). This suggests that their probabilities must add up to one, i.e. $p_{\cdot 1}+p_{\cdot 2}=1$ and $p_{1\cdot}+p_{2\cdot}=1$.

Example 4.16. Let us derive the marginal distribution of X from the joint density given below:

$y\diagdown x$	0	1	2
0	0.2	0.2	0.2
2	0.1	0.1	0.2

(17)

The formula in (12) suggests that by summing down the columns we derive the marginal density of X and summing over rows we derive the marginal density of Y:

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

y	0	2
$f_y(y)$	0.6	0.4

(18)

These are clearly proper density functions, given that:

$$f_x(x)\geq 0,\ f_x(0)+f_x(1)+f_x(2)=1, \text{ and } \quad f_y(x)\geq 0,\ f_y(0)+f_y(2)=1.$$

The two marginal densities are shown with the joint density below.

$y\diagdown x$	0	1	2	$f_y(y)$
0	0.2	0.2	0.2	0.6
2	0.1	0.1	0.2	0.4
$f_x(x)$	0.3	0.3	0.4	1

(19)

Looking at the last column we can see that the probabilities associated with the values of Y contain no information relating to X.

4 Conditional Distributions

4.1 Conditional probability

Let us return to chapter 2 and remind ourselves of the concept of conditional probability using our favorite example.

Example 4.17. Consider the random experiment of "tossing a fair coin twice", whose outcomes set is:

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

Assuming that $A=\{(HH),(HT),(TH)\}$ is an event of interest, without any additional information common sense suggests that $\mathbb{P}(A)=\frac{3}{4}$. However, in the case where there exists some additional information, say somebody announces that in a particular trial "the first coin is a T", the situation changes. The available information defines the event $B=\{(TH),(TT)\}$ and knowing that B has occurred invalidates the probability $\mathbb{P}(A)=\frac{3}{4}$. This is because the information implies that, in this particular trial, the outcomes (HH) and (HT) cannot occur. That is, instead of S the set of all possible distinct outcomes, given that B has occurred, is just B. This suggests that the new probability of A, given that B has occurred, denoted by $\mathbb{P}(A|B)$, is different. Common sense suggests that $\mathbb{P}(A|B)=\frac{1}{2}$, because A includes one of the two possible distinct outcomes. ▶ How can we formalize this argument?

The formula for the conditional probability of event A given event B, takes the form:

$$\mathbb{P}(A|B)=\tfrac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)},\ \text{for}\ \mathbb{P}(B)>0. \tag{20}$$

In the above Example, $\mathbb{P}(A\cap B)=\mathbb{P}(TH)=\frac{1}{4}$, $\mathbb{P}(B)=\frac{1}{2}$, and thus $\mathbb{P}(A\cap B)=\frac{\frac{1}{4}}{\frac{1}{2}}=\frac{1}{2}$, which confirms the common sense answer.

4.2 Conditional density functions

As in the case of the joint and marginal distributions we will consider the simple discrete random variables case first and then proceed to discuss the general continuous random variables case.

4.2.1 Discrete random variables

In the case of two discrete random variables X and Y, if we define the events:

$$A=\{Y=y\}\ \text{and}\ B=\{X=x\},$$

then the translation of the above formulae in terms of density functions takes the form:

$$\begin{aligned}&\mathbb{P}(X=x)=f(x),\ x\in\mathbb{R}_X,\ \mathbb{P}(Y=y)=f(y),\ y\in\mathbb{R}_Y,\\&\mathbb{P}(Y=y,X=x)=f(x,y),\ (x,y)\in\mathbb{R}_X\times\mathbb{R}_Y,\\&\mathbb{P}(Y=y|X=x)=f(y|x),\ y\in\mathbb{R}_Y,\end{aligned}$$

giving rise to the conditional density formula:

$$f(y|x)=\tfrac{f(x,y)}{f_x(x)},\ \text{for}\ f(x)>0,\ y\in\mathbb{R}_Y, \tag{21}$$

where $f(y|x)$ denotes the conditional density of Y given that $X{=}x$.

Example 4.18. Consider the joint density function for the discrete random variables X and Y given in (19). From the above formula we can see that the conditional density of Y given $X{=}0$ takes the form:

$$f(y|x{=}0){=}\tfrac{f(x=0,y)}{f_x(x=0)},\ y{\in}\mathbb{R}_Y{:=}\{0,2\}.$$

This suggests that the conditional probabilities $f(y|x{=}0)$, for $y{\in}\mathbb{R}_Y$, are weighted averages of the joint probabilities $f(x{=}0,y)$, for $y{\in}\mathbb{R}_Y$, with the marginal probability $f_x(x{=}0)$ providing the weight. Hence:

$$f(y|x{=}0){=}\begin{cases}\frac{f(x=0,y=0)}{f_x(x=0)}{=}\ \ \frac{0.2}{0.3}{=}\frac{2}{3},\ y{=}0,\\ \frac{f(x=0,y=2)}{f_x(x=0)}{=}\ \ \frac{0.1}{0.3}{=}\frac{1}{3},\ y{=}2.\end{cases}$$

The conditional density is shown below:

y	0	2
$f(y\|x{=}0)$	$\frac{2}{3}$	$\frac{1}{3}$

(22)

4.2.2 Continuous random variables

In the case of two continuous random variables X and Y we cannot use the events $A{=}\{Y{=}y\}$ and $B{=}\{X{=}x\}$ in order to transform (20) in terms of density functions, because as we know in such a case $P(X{=}x){=}0$ and $P(Y{=}y){=}0$ for all $x{\in}\mathbb{R}$, $y{\in}\mathbb{R}$. As in the case of the definition of the joint and marginal density functions we need to consider events of the form:

$$A{=}\{X\leq x\}\text{ and }B{=}\{Y\leq y\}.$$

However, even in the case of continuous random variables we would like to be able to refer to the conditional distribution of Y given $X{=}x$. The way we get around the mathematical difficulties is by way of the conditional cumulative distribution function defined as follows:

$$F_{Y|X}(y|X{=}x){=}\lim_{h\to 0^+}\tfrac{\mathbb{P}(Y\leq y,\ x\leq X\leq x+h)}{\mathbb{P}(x\leq X\leq x+h)},$$

where $h\to 0^+$ reads "*as h tends to 0 through values greater than 0*". After some mathematical manipulations we can show that:

$$F_{Y|X}(y|X{=}x){=}\lim_{h\to 0^+}\tfrac{\mathbb{P}(Y\leq y,\ x\leq X\leq x+h)}{\mathbb{P}(x\leq X\leq x+h)}{=}\int_{-\infty}^{y}\tfrac{f(x,u)}{f_x(x)}du.$$

This suggests that in the case of two continuous random variables X and Y we could indeed define the conditional density function as in (21) but $f(y|x)$ we should not interpret it as assigning probabilities because:

$$f(.|x)\colon\ \mathbb{R}_Y\to[0,\infty).$$

As we can see, the *conditional density* is a proper density function, in so far as, in the case of continuous random variables, it satisfies the properties in table 4.5.

Table 4.5: Conditional density - Properties

[cf1]	$f(y\|x) \geq 0$, for all $y \in \mathbb{R}_Y$,
[cf2]	$\int_{-\infty}^{\infty} f(y\|x)\, dy = 1$,
[cf3]	$F(y\|x) = \int_{-\infty}^{y} f(u\|x)\, du$.

NOTE: in the case of discrete random variables the integrals are replaced with summations.

Example 4.19. Consider the case where the joint density function takes the form:

$$f(x,y) = 8xy,\ 0 < x < y,\ 0 < y < 1.$$

The marginal densities of X and Y can be derived from the joint density by integrating out Y and X, respectively:

$$f_x(x) = \int_x^1 (8xy)dy = 4xy^2|_{y=x}^{y=1} = 4x(1-x^2),\ 0 < x < 1,$$

$$f_y(y) = \int_0^y (8xy)dx = 4x^2y|_{x=0}^{x=y} = 4y^3,\ 0 < y < 1.$$

REMARK: The only difficulty in the above derivations is to notice that the range of X is constrained by Y and vice versa. Using these results we can deduce that:

$$f(y|x) = \frac{8xy}{4x(1-x^2)} = \frac{2y}{(1-x^2)},\quad x < y < 1,\ 0 < x < 1,$$

$$f(x|y) = \frac{8xy}{4y^3} = \frac{2x}{y^2},\qquad 0 < x < y,\ 0 < y < 1.$$

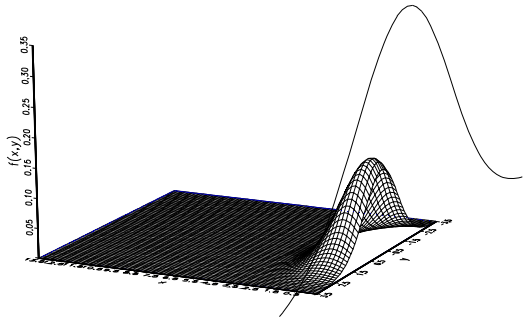


Fig. 4.3: Bivariate Normal density with conditional density at $x = -.5$

Example 4.20. Consider the bivariate standard *Normal distribution*. As seen in the previous section, in the case where $f(x,y)$ is Normal, the marginal distributions $f_x(x)$ and $f_y(y)$ are also Normal. Hence, ignoring the details of the mathematical derivations, the conditional density of Y given $X=x$ can be derived as follows:

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{(2\pi\sqrt{(1-\rho^2)})^{-1}\exp\{-[2(1-\rho^2)]^{-1}(x^2-2\rho xy+y^2)\}}{(\sqrt{2\pi})^{-1}\exp\{-\frac{1}{2}x^2\}} =$$

$$= [2\pi(1-\rho^2)]^{-\frac{1}{2}}\exp\left\{-[2(1-\rho^2)]^{-1}(x^2-2\rho xy+y^2)+\frac{1}{2}x^2\right\}.$$

Using the equality: $[2(1-\rho^2)]^{-1}(x^2-2\rho xy+y^2)+\frac{1}{2}x^2=\ [2(1-\rho^2)]^{-1}(y-\rho x)^2$, the conditional density takes the form:

$$f(y|x)=\frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp\left\{-\frac{1}{2(1-\rho^2)}[y-\rho x]^2\right\}. \tag{23}$$

This shows that $f(y|x)$ is also Normal with mean ρx and variance $(1-\rho^2)$:

$$(Y|X=x) \backsim \mathsf{N}\left(\rho x, (1-\rho^2)\right).$$

The conditional density $f(y|x=\text{-}.5)$ can be visualized as the one dimensional density we get by *slicing* the joint density using a perpendicular plane, parallel to the y-axis and passing through the point $x=\text{-}.5$. In figure 4.3 we can see how the slicing of the bivariate surface at $x=\text{-}.5$ scaled by $[1/f_x(\text{-}.5)]$ yields (23).

4.3 Conditional moments

The conditional density, being a proper density function, also enjoys numerical characteristics analogous to marginal density functions. In particular, for continuous random variables we can define the **conditional moments:**

$$\text{Raw:}\qquad E(Y^r|X=x)=\int_{-\infty}^{\infty} y^r f(y|x)dy,\quad r=1,2,...,$$

$$\text{Central:}\quad E\left\{(Y-E[Y|X=x])^r|X=x\right\}=\int_{-\infty}^{\infty}[y-E(y|x)]^r f(y|x)dy,\ r=2,3,...$$

NOTE that the only difference between the marginal and conditional moments is that the relevant distribution with respect to which $E(.)$ is defined is now the conditional.

In the case of *discrete* random variables we replace the integrals with summations as exemplified in the case of the first of these conditional moments:

$$\text{Conditional mean:}\qquad E(Y|X=x)=\textstyle\sum_{y\in\mathbb{R}_Y} y\cdot f(y|x),$$

$$\text{Conditional variance:}\quad Var(Y|X=x)=\textstyle\sum_{y\in\mathbb{R}_Y}[y-E(y|x)]^2\cdot f(y|x).$$

Example 4.22. *Discrete distribution*, no unknown parameters. For the conditional density (22):

$$E(Y|X=0)=0\cdot(\tfrac{2}{3})+2\cdot(\tfrac{1}{3})=(\tfrac{2}{3}),\ Var(Y|X=0)=[0-(\tfrac{2}{3})]^2(\tfrac{2}{3})+[2-(\tfrac{2}{3})]^2(\tfrac{1}{3})=(\tfrac{24}{27}).$$

Example 4.23. *Continuous distribution, no unknown parameters.* Consider the case where the joint density function takes the form:

$$f(x,y)=8xy,\ 0<x<y,\ 0<y<1.$$

As shown above, the marginal densities of x and y are:

$$f(x)=4x(1-x^2),\ 0<x<1 \text{ and } \ f(y)=4y^3,\ 0<y<1.$$

$$f(y|x)=\frac{8xy}{4x(1-x^2)}=\frac{2y}{(1-x^2)},\quad x<y<1,\ 0<x<1,\quad f(x|y)=\frac{8xy}{4y^3}=\frac{2x}{y^2},\quad 0<x<y,\ 0<y<1.$$

$$E(Y|X{=}x) = \int_x^1 y\left(\tfrac{2y}{(1-x^2)}\right)dy = \tfrac{2}{(1-x^2)}\int_x^1 y^2 dy = \tfrac{2}{(1-x^2)}\left(\tfrac{1}{3}y^3\big|_{y=x}^{y=1}\right) = \tfrac{2}{3}\left(\tfrac{1-x^3}{1-x^2}\right),$$

$$E(X|Y{=}y) = \int_0^y x\left(\tfrac{2x}{y^2}\right)dx = \tfrac{2}{y^2}\left(\tfrac{1}{3}x^3\big|_{x=0}^{x=y}\right) = \tfrac{2}{y^2}\left(\tfrac{1}{3}y^3\right) = \tfrac{2}{3}y,$$

$$Var(X|Y{=}y) = \int_0^y [x-\tfrac{2}{3}y]^2\left(\tfrac{2x}{y^2}\right)dx = \int_0^y [x^2+\tfrac{4}{9}y^2-\tfrac{4}{3}xy]\left(\tfrac{2x}{y^2}\right)dx =$$

$$= \int_0^y [(\tfrac{2x^3}{y^2})+\tfrac{8}{9}x-\tfrac{8}{3}(\tfrac{x^2}{y})]dx = (\tfrac{x^4}{2y^2})+\tfrac{4}{9}x^2-\tfrac{8}{9}(\tfrac{x^3}{y})\Big|_{x=0}^{x=y} = \tfrac{1}{18}y^2.$$

Example 4.24. *Continuous distribution, with unknown parameters.* Consider the case of the bivariate (standard) Normal distribution discussed in the previous sub-section. It was shown that the conditional distribution of Y given $X{=}x$ takes the form:

$$(Y|X{=}x) \backsim \mathsf{N}\left(\rho x,\ (1-\rho^2)\right).$$

This suggests that $E(Y|X{=}x){=}\rho x$, and $Var(Y|X{=}x){=}(1-\rho^2)$.

The conditional moments are of interest in modeling dependence, because they often provide the most flexible way to capture the important aspects of probabilistic dependence (see chapter 6).

4.4 Marginalization vs. conditioning

Marginal and conditional densities, viewed in relation to the joint density function:

$$\begin{aligned} \text{Joint} \quad & f(.,.)\text{: } (\mathbb{R}\times\mathbb{R}) \to [0,\infty), \\ \text{Marginal} \quad & f_y(.)\text{: } \mathbb{R} \to [0,\infty), \\ \text{Conditional} \quad & f(.|x)\text{: } \mathbb{R} \to [0,\infty), \end{aligned}$$

have one thing in common: they are both univariate densities. In the case of the marginal density $f_y(.)$ the information relating to the other random variable X is suppressed (integrated out). On the other hand, in the case of the conditional density $f(.|x)$ retains part of the information relating to X; the information $X{=}x$.

The formula (21), defining the conditional density can be re-arranged to yield:

$$f(x,y){=}f(y|x){\cdot}f_x(x), \text{ for all } (x,y){\in}\mathbb{R}_X{\times}\mathbb{R}_Y. \tag{24}$$

This reduces the bivariate density $f(x,y)$, into a product of two univariate densities, $f(y|x)$ and $f_x(x)$. This reduction is important in relation to the concept of independence. Before we consider that, however, let us elaborate on the intuition underlying marginalization and conditioning.

Example 4.30. Contemplate the following scenario. You wake up in a hospital covered in plaster from head to toe with only the eyes, ears and mouth showing and suffering from complete amnesia. A nurse, who just came on duty, walks in and informs you that based on the report he had just read: you have been involved in a car accident, you are in bad shape (but out of danger) and you are likely to remain in hospital for a while. The first questions that come to mind are: ▶ who am I? and ▶ can I afford the bills? The nurse seems to be reading your mind, but the only thing

he could offer is the joint distribution, shown below, pertaining to the broader local community you come from, where X denotes *age* bracket and Y *income* bracket:

X=1: (18-35), X=2: (36-55), X=3: (56-70),
Y=0: poor, Y=1: middle income, Y=2: rich.

$x \diagdown y$	0	1	2	$f_x(x)$
1	.20	.10	.01	.31
2	.10	.25	.06	.41
3	.15	.05	.08	.28
$f_y(y)$	.45	.40	.15	1

(25)

A glance at the joint probabilities brings some more confusion because the highest probability is attached to the event $(X{=}2, Y{=}1)$ (middle aged and middle income) and the lowest probability is attached to the event $(X{=}1, Y{=}2)$ (young but rich!). In an attempt to re-assure yourself you ignore income (as of secondary importance) for a moment and look at the marginal density of X. The probability of being in the age bracket X=3 (56-70) (irrespective of income) is $f_x(x{=}3){=}.28$ is lower than the probabilities of being either young $f_x(x{=}1){=}.31$ or middle-aged $f_x(x{=}2){=}.41$; a sigh of relief but not much comfort because $f_x(x{=}1){=}.31$ is not very much higher than $f_x(x{=}3){=}.28$! During this syllogism the nurse remembers that according to the report you were driving a Porsche! This additional piece of information suddenly changes the situation. Unless you were a thief speeding away when the accident happened (an unlikely event in a crime-free small community), you can assume that $Y{=}2$ has happened. ▶ How does this change the joint probabilities?

The relevant probabilities now are the conditional probability of X given $Y{=}2$:

$$f(x|y{=}2){=}\begin{cases} \frac{f(x=1,y=2)}{f_y(y=2)}{=}\ \ \frac{.01}{.15}{=}.067, & x{=}1, \\ \frac{f(x=2,y=2)}{f_y(y=2)}{=}\ \ \frac{.06}{.15}{=}.400, & x{=}2, \\ \frac{f(x=3,y=2)}{f_y(y=2)}{=}\ \ \frac{.08}{.15}{=}.533, & x{=}3. \end{cases}$$

A glance at these conditional probabilities and you are begging the nurse to take the plaster off to check how old you are; there is more than 50% chance your are a senior!

4.5 Conditioning on events vs. random variables

In most textbooks the formulae for assigning probabilities to events are translated directly into analogous formulae using density functions that seem very similar; see table 4.7 but ignore the quantifiers in boxes. Despite the mnemonic value of the similarity, it turns out to be highly misleading because random variables do not

represent a single event since they take more than one value.

Table 4.7: **Joint and conditional probability formulae**

Events: $\mathbb{P}(.)$	**Random variables:** $f(.)$
[i] $\mathbb{P}(A\|B)=\frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)},\ \mathbb{P}(B)>0,$	[i]* $f(y\|x)=\frac{f(x,y)}{f(x)},$ for $f(x)>0,$ $\boxed{\forall y\in\mathbb{R}_Y}$
[ii] $\mathbb{P}(A\|B)=\mathbb{P}(A\|B)\cdot\mathbb{P}(B)=\mathbb{P}(B\|A)\cdot\mathbb{P}(A),$ for $\mathbb{P}(A)>0,\ \mathbb{P}(B)>0,$	[ii]* $f(x,y)=f(x\|y)\cdot f(y)=f(y\|x)\cdot f(x),$ for $f(x)>0,\ f(y)>0,$ $\boxed{\forall(x,y)\in\mathbb{R}_X\times\mathbb{R}_Y,}$
[iii] $\mathbb{P}(A\|B)=\frac{\mathbb{P}(B\|A)\cdot\mathbb{P}(A)}{\mathbb{P}(B)},\ \ \mathbb{P}(B)>0,$	[iii]* $f(y\|x)=\frac{f(x\|y)f(y)}{f(x)},$ for $f(x)>0,$ $\boxed{?}$

That is, the formulae [i]*-[iii]* *without* the quntifiers [in boxes] are both *technically incorrect* and *highly misleading* for several reasons.

(a) One cannot simply replace A and B with Y and X because a random variable is *not an event* in itself. Random variables can be associated with more than one event whose totality defines the σ-field $\sigma(X)$ generated by X; see chapter 3.

(b) The joint, conditional and marginal density functions are defined at particular values of X and Y which need to be specified explicitly using the relevant quantifier. In that sense, for the formulae [i]*-[iii]* to be formally correct, the *missing quantifiers* (in boxes) for the relevant values of X and Y need to be added.

(c) Although the conditioning in [i]-[iii] is symmetric, for conditional densities it is *non-symmetric* with respect to the random variables X and Y.

In light of (a)-(c), the proper way to define [i]* includes a quantifier $\forall y\in\mathbb{R}_Y$:

$$\text{[i]}^{**}\ f(y|X=x)=\frac{f(X=x,y)}{f(X=x)},\ \text{for}\ f(x)>0,\ \forall y\in\mathbb{R}_Y. \tag{26}$$

That is, $f(y|X=x)$ is defined at a single value of the conditioning random variable X, say $X=x$, and all values of y in $\mathbb{R}_Y$.Moreover, for each value $X=x$, $f(y|X=x)$, $\forall y\in\mathbb{R}_Y$, defines a different conditional distribution, each of which constitutes a proper density function since: $f(y|X=x)\geq 0,\ \forall y\in\mathbb{R}$, and $\sum_{y\in\mathbb{R}_Y} f(y|X=x)dy=1$.

Example 4.31. To illustrate what the conditional density formula in (26) represents, let us evaluate $f(y|X=x),\ \forall y\in\mathbb{R}_Y$ and $f(x|Y=y),\ \forall x\in\mathbb{R}_X$ in the case of the joint distribution in table 4.8.

$x\diagdown y$	0	1	2	$f_x(x)$
1	.20	.10	.01	.31
2	.10	.25	.06	.41
3	.15	.05	.08	.28
$f_y(y)$	.45	.40	.15	1

Table 4.8

$$f(y|x{=}1){=}\begin{cases}\frac{f(x=1,y=0)}{f_x(x=1)}{=}\frac{.20}{.31}, & y{=}0\\ \frac{f(x=1,y=1)}{f_x(x=1)}{=}\frac{.10}{.31}, & y{=}1\\ \frac{f(x=1,y=2)}{f_x(x=1)}{=}\frac{.01}{.31}, & y{=}2\end{cases}$$

y	0	1	2
$f(y\|x{=}1)$	.645	.323	.032

$$f(y|x{=}2){=}\begin{cases}\frac{f(x=2,y=0)}{f_x(x=2)}{=}\frac{.10}{.41}, & y{=}0\\ \frac{f(x=2,y=1)}{f_x(x=2)}{=}\frac{.25}{.41}, & y{=}1\\ \frac{f(x=2,y=2)}{f_x(x=2)}{=}\frac{.06}{.41}, & y{=}2\end{cases}$$

y	0	1	2
$f(y\|x{=}2)$	.244	.610	.146

$$f(y|x{=}3){=}\begin{cases}\frac{f(x=3,y=0)}{f_x(x=3)}{=}\frac{.15}{.28}, & y{=}0\\ \frac{f(x=2,y=1)}{f_x(x=3)}{=}\frac{.05}{.28}, & y{=}1\\ \frac{f(x=2,y=2)}{f_x(x=3)}{=}\frac{.08}{.28}, & y{=}2\end{cases}$$

y	0	1	2
$f(y\|x{=}3)$	.536	.179	.285

(27)

Similarly, the conditional density of X given $Y{=}y$ take the form:

$$f(X|Y{=}y){=}\tfrac{f(x,Y=y)}{f(Y=y)},\ \text{for } f(y){>}0,\ \forall x{\in}\mathbb{R}_X. \tag{28}$$

When evaluated using the joint density in table 4.8 yields:

$$f(x|y{=}0){=}\begin{cases}\frac{f(x=1,y=0)}{f_y(y=0)}{=}\frac{.20}{.45}, & x{=}1\\ \frac{f(x=2,y=0)}{f_y(y=0)}{=}\frac{.10}{.45}, & x{=}2\\ \frac{f(x=3,y=0)}{f_y(y=0)}{=}\frac{.15}{.45}, & x{=}3\end{cases}$$

x	1	**2**	3
$f(x\|y{=}0)$	.445	.222	.333

$$f(x|y{=}1){=}\begin{cases}\frac{f(x=1,y=1)}{f_y(y=1)}{=}\frac{.10}{.40}, & x{=}1\\ \frac{f(x=2,y=1)}{f_y(y=1)}{=}\frac{.25}{.40}, & x{=}2\\ \frac{f(x=3,y=1)}{f_y(y=1)}{=}\frac{.05}{.40}, & x{=}3\end{cases}$$

x	1	**2**	3
$f(x\|y{=}1)$	.250	.625	.125

$$f(x|y{=}2){=}\begin{cases}\frac{f(x=1,y=2)}{f_y(y=2)}{=}\frac{.01}{.15}, & x{=}1\\ \frac{f(x=2,y=2)}{f_y(y=2)}{=}\frac{.06}{.15}, & x{=}2\\ \frac{f(x=3,y=2)}{f_y(y=2)}{=}\frac{.08}{.15}, & x{=}3\end{cases}$$

x	1	**2**	3
$f(x\|y{=}2)$	.067	.400	.533

(29)

The features of the conditional density brought out above suggest that the multiplication rule in terms of events in [ii] does not extend to random variables as in [ii]* without the quantifier (in the box). When the relevant quantifier for the two conditional densities is added, it becomes obvious from (27) and (29) that the two sides are not equal (one is equating apples and eggs):

$$\{f(y|X{=}x){\cdot}f(X{=}x),\forall y{\in}\mathbb{R}_Y\}\neq\{f(x|Y{=}y){\cdot}f(Y{=}y),\forall x{\in}\mathbb{R}_X\} \tag{30}$$

The multiplication rule only holds when X and Y take all their values:

$$f(x,y){=}f(y|X{=}x){\cdot}f(X{=}x){=}f(x|Y{=}y){\cdot}f(Y{=}y),\ \forall(x,y){\in}\mathbb{R}_X{\times}\mathbb{R}_Y, \tag{31}$$

but such a case what one has in (31) is essentially the joint distribution *reparametrized* in terms of *all* the conditional and marginal distributions.

The most misleading conversion from events to random variables is from formula [iii] to [iii]*, known as Bayes rule; see chapter 3. For events A and B, $\mathbb{P}(A\cap B)=\mathbb{P}(B|A)\cdot\mathbb{P}(A)$ but for random variables one cannot replace $f(X{=}x,y)$ with $f(X{=}x|y)f(y)$. Once the missing quantifier $\forall y{\in}\mathbb{R}_Y$, is added to the proper definition of $f(y|X{=}x)$ in [iii]*, the resulting formula is absurd:

$$f(y|X{=}x){=}\tfrac{f(X=x|y)f(y)}{f(X=x)},\ \text{for } f(x){>}0,\ \forall y{\in}\mathbb{R}_Y. \tag{32}$$

To shed some light on what (32) represents, let us begin with the key component $\{f(X{=}x|y),\ \forall y{\in}\mathbb{R}_Y\}$. On closer examination, it turns out that it is neither the conditional density of X given $Y{=}y$, $\{f(y|X{=}x),\forall y{\in}\mathbb{R}_Y\}$, nor the conditional density of Y given $X{=}x$, $\{f(X|Y{=}y),\forall x{\in}\mathbb{R}_X\}$. Instead, $\{f(X{=}x|y),\ \forall y{\in}\mathbb{R}_Y\}$ in (32) represents a 'cannibalized' probability function that is *not* even a proper density in the sense that it sums to one.

Example 4.31 (continued). The evaluation of $f(X{=}x|y),\ y{\in}\mathbb{R}_Y$, for the joint distribution in table 4.8 gives rise to table 4.9.

y	$f(X{=}2,y)$	$f(X{=}2\|y)$	$f(y)$
0	.10	.222	.45
1	.25	.625	.40
2	.06	.400	.15
	.41	1.247	$\times$

Table 4.9 (33)

In light of the fact that for the summation of its probabilities in table 4.9 yields:

$$\textstyle\sum_{y\in\mathbb{R}_Y} f(X{=}2|y){=}1.247,\quad \sum_{y\in\mathbb{R}_Y} f(X{=}2,y){=}f(X{=}2){=}.41, \tag{34}$$

shows that $f(X{=}2,y){=}f(X{=}2|y){\cdot}f(y)$ is not a proper density. Indeed, when summed over all values of Y, yields $f(X{=}2)$, *not* the marginal distribution of X, but just one value. Hence, $f(X{=}x|y),\ y{\in}\mathbb{R}_Y$ in (32) represents one piece from the three conditional densities in (29), despite claims by Bayesian textbooks that:

(a) $f(X{=}x|y)f(y)$ represents a product of a conditional and a marginal density; see Lindley (1965), p. 118, and Robert (2007), pp. 8-9.

That is not a true claim. To render [iii]* meaningful one needs to add a quantifier, and the only one that can make claim (a) true is:

$$\text{[iii]}^{**}\ f(y|x){=}\tfrac{f(x|y)f(y)}{f(x)},\ \text{for } f(x){>}0,\ \forall(x,y){\in}\mathbb{R}_X{\times}\mathbb{R}_Y.$$

This, however, turns [iii]** into the joint distribution $f(y,x)$ *reparameterized* in terms of the conditional $f(x|y)$ and marginal distributions $f(x)$ and $f(y)$.

It is important emphasize that the above results do not depend on the fact that the example used is in terms of discrete random variables. The only difference when (X,Y) are continuous is that the summation will be replaced by integration.

Armed with the concepts of joint, marginal and conditional distributions for random variables, we can proceed to formalize the concepts of independence and identical distributions.

5 Independence

5.1 Independence in the two random variable case

As seen in chapter 2, two events A and B which belong to the same event space $\Im$, are said to be **independent** if:

$$\mathbb{P}(A \cap B)=\mathbb{P}(A)\cdot\mathbb{P}(B).$$

By translating the arbitrary events A and B into events of the form: A:=(s: $X(s) \leq x$) and B:=(s: $Y(s) \leq y$), $s \in S$, the above condition becomes:

$$\begin{aligned} &\mathbb{P}(X \leq x, Y \leq y)=\mathbb{P}(X \leq x)\cdot\mathbb{P}(Y \leq y), \text{ for all } (x,y)\in\mathbb{R}^2. \\ &F_{XY}(x,y)=F_X(x)\cdot F_Y(y), \text{ for all } (x,y)\in\mathbb{R}^2, \end{aligned} \tag{35}$$

where $F_{XY}(.,.)$ denotes the joint cumulative distribution function (cdf). In terms of the density functions, X and Y are said to be *independent* if:

$$f(x,y)=f_x(x)\cdot f_y(y), \text{ for all } (x,y)\in\mathbb{R}^2. \tag{36}$$

That is, the joint density is equal to the product of the two marginal density functions. In other words the only case where the joint density contains no additional information from that contained in the marginal density functions is the case where the random variables are independent.

It is important to NOTE that in view of (37), when X and Y are *independent:*

$$f(y|x)=\frac{f_x(x)\cdot f_y(y)}{f_x(x)}=f_y(y) \text{ for all } y\in\mathbb{R}_Y. \tag{37}$$

Similarly, $f(x|y)=f_x(x)$, for all $x\in\mathbb{R}_X$. That is, when Y and X are independent, conditioning on X does not affect the marginal density of Y and vice versa. This provides a more intuitive way to understand the concept of independence.

Example 4.32. Consider the bivariate density (25). The random variables X and Y are *not independent* since for the first value $(X,Y)=(1,0)$:

$$f(1,0)=(.20) \neq f_x(1)\cdot f_y(0)=(.31)(.45)=(.1395).$$

Example 4.33. Consider the bivariate density given below.

$x \diagdown y$	0	1	$f_x(x)$
0	0.3	0.2	0.5
2	0.3	0.2	0.5
$f_y(y)$	0.6	0.4	1

(38)

To check whether X and Y are independent, we need to verify that the equality in (36) holds, for *all* values of X and Y:

$(X,Y)=(0,0)$: $f(0,0)=f_x(0)\cdot f_y(0)=(.3)=(.5)(.6)$,

$(X,Y)=(1,0)$: $f(1,0)=f_x(1)\cdot f_y(0)=(.3)=(.5)(.6)$,

$(X,Y)=(0,2)$: $f(0,2)=f_x(0)\cdot f_y(2)=(.2)=(.5)(.4)$,

$(X,Y)=(1,2)$: $f(1,2)=f_x(1)\cdot f_y(2)=(.2)=(.5)(.4)$.

These results suggest that (36) holds, and thus X and Y are independent.

Example 4.34. In the case where (X, Y) are jointly Normally distributed, with density as defined in (9), we can deduce that when ρ=0, X and Y are independent. This follows by a simple substitution of the restriction ρ=0 in the joint density:

$$\begin{aligned} f(x,y) &= (\tfrac{(1-\rho^2)^{-\frac{1}{2}}}{2\pi}) \exp\{-\tfrac{1}{2(1-\rho^2)}[x^2-2\rho xy+y^2]\}\Big|_{\rho=0} = \tfrac{1}{2\pi}\exp\left\{-\tfrac{1}{2}[x^2+y^2]\right\} = \\ &= \left((\tfrac{1}{\sqrt{2\pi}})\exp\{-\tfrac{1}{2}x^2\}\right)\left((\tfrac{1}{\sqrt{2\pi}})\exp\{-\tfrac{1}{2}y^2\}\right) = f_x(x)\cdot f_y(y), \end{aligned}$$

where $f_x(x)$ and $f_y(y)$ are standard Normal densities.

REMARK: The last example provides an important clue to the concept of independence by suggesting that when the joint density $f(x, y)$ can be factored into a product of two non-negative functions $u(x)$ and $v(y)$ i.e.

$$f(x, y)=u(x)\cdot v(y),$$

where $u(.) \geq 0$ depends only on x and $v(.) \geq 0$ depends only on y, then X and Y are independent.

Example 4.35. In the case where (X, Y) are jointly exponentially distributed, with density:

$$f(x, y; \theta)= [(1+\theta x)(1+\theta y)-\theta] \exp\{-x-y-\theta xy\},\ x>0,\ y>0,\ \theta>0.$$

It's obvious that X and Y are independent only when θ=0, since the above factorization can be achieved only in that case:

$$f(x, y; 0)= [(1+\theta x)(1+\theta y)-\theta] \exp\{-x-y-\theta xy\}|_{\theta=0} =(e^{-x})(e^{-y}).$$

5.2 Independence in the n random variable case

The extension of the above definitions of independence from the two to the n-variable case is not just a simple matter of notation. As argued in the previous chapter, the events $A_1, A_2, ..., A_n$ are *independent* if the following condition holds:

$$\mathbb{P}(A_1 \cap A_2 \cap ... \cap A_k)=\mathbb{P}(A_1)\cdot\mathbb{P}(A_2)\cdots\mathbb{P}(A_k),\ \text{for } k=2, 3, ..., n. \tag{39}$$

Independence. The random variables $X_1, X_2, ..., X_n$ are said to be *independent* if the following condition holds:

$$F(x_1, x_2, ..., x_n)=F_1(x_1)\cdot F_2(x_2)\cdots F_n(x_n),\ \text{for all } \mathbf{x}\in\mathbb{R}^n. \tag{40}$$

where for $\mathbf{x}$:=$(x_1, ..., x_n)$. In terms of the density functions, *independence* can be written in the form:

$$f(x_1, x_2, ..., x_n)=f_1(x_1)\cdot f_2(x_2)\cdot\cdots\cdot f_n(x_n),\ \text{for all } \mathbf{x}\in\mathbb{R}^n. \tag{41}$$

From (41) we can see that the qualification *for all subsets of* $\{A_1, A_2, ..., A_n\}$ in the case of events has been replaced with the qualification *for all* $\mathbf{x}\in\mathbb{R}^n$. In other words, in the case of random variables we do not need to check (41) for any subsets of the set $X_1, X_2, ..., X_n$, but we need to check it for all values $\mathbf{x}\in\mathbb{R}^n$.

The above definition completes the first stage of our quest for transforming the concept of random trials. The independence given in the introduction in terms of trials (1) has now been recast in terms of random variables as given in (41). We consider the second scale of our quest in the next section.

6 Identical Distributions and random samples

As mentioned in the introduction, the concept of random trials has two components: independence and identical distributions. Let us consider the recasting of the identically distributed component in terms of random variables.

6.1 Identically Distributed random variables

Example 4.37. Consider the Bernoulli density function:

$$f(x;\theta)=\theta^x(1-\theta)^{1-x},\ x=0,1,$$

where $\theta=\mathbb{P}(X=1)$. Having a sample of n independent trials, say $(X_1, X_2, ..., X_n)$, amounts to assuming that the random variables $X_1, X_2, ..., X_n$ are *independent*, with each X_i having a density function of the form:

$$f(x_i;\theta_i)=\theta_i^{x_i}(1-\theta_i)^{1-x_i},\ x_i=0,1,\ i=1,2,...,n,$$

where $\theta_i=\mathbb{P}(X_i=1),\ i=1,2,...,n$. Independence in this case ensures that:

$$f(x_1,...,x_n;\boldsymbol{\phi})=\prod_{i=1}^n f_i(x_i;\theta_i)=\prod_{i=1}^n \theta_i^{x_i}(1-\theta_i)^{1-x_i},\ x_i=0,1,$$

where $\boldsymbol{\phi}:=(\theta_1,\theta_2,...,\theta_n)$. Obviously, this does not satisfy the identically distributed component. For that to be the case we need to impose the restriction that for all trials the probabilistic structure remains the same, i.e. the random variables $X_1, X_2, ..., X_n$ are also *identically distributed* in the sense that:

$$f(x_i;\theta_i)=\theta^{x_i}(1-\theta)^{1-x_i},\ x_i=0,1,\ i=1,2,...,n.$$

Let us formalize the concept of identically distributed random variables in the case of arbitrary but independent random variables, beginning with the 2-variable case. In general, the joint density involves the unknown parameters $\boldsymbol{\phi}$, and the equality in (36) takes the form:

$$f(x,y;\boldsymbol{\theta})=f_x(x;\boldsymbol{\theta}_1)\cdot f_y(y;\boldsymbol{\theta}_2),\ \text{for all } (x,y)\in\mathbb{R}_X\times\mathbb{R}_Y,$$

where the marginal distributions $f_x(x;\boldsymbol{\theta}_1)$ and $f_y(y;\boldsymbol{\theta}_2)$ can be very different.

Two independent random variables are said to be *identically distributed* if $f_x(x;\boldsymbol{\theta}_1)$ and $f_y(y;\boldsymbol{\theta}_2)$ are the same density functions, denoted by:

$$f_x(x;\boldsymbol{\theta}_1)\equiv f_y(y;\boldsymbol{\theta}_2),\ \text{for all } (x,y)\in\mathbb{R}_X\times\mathbb{R}_Y,$$

where the equality sign $\equiv$ is used to indicate that all the marginal distributions have the same functional form and the same unknown parameters, i.e. $f_x(.)=f_y(.)$ and $\boldsymbol{\theta}_1=\boldsymbol{\theta}_2$.

Example 4.38. Consider the case where: $f(x,y;\boldsymbol{\theta})=\left(\frac{\theta_1}{\theta_2}\right)\frac{e^{-\frac{y}{\theta_2}}}{x^2},\ x\geq 1,\ y>0.$

It is clear that X and Y are independent with marginal densities:

$$f_x(x;\theta_1)=\tfrac{\theta_1}{x^2},\ x\geq 1,\quad f_y(y;\theta_2)=\tfrac{1}{\theta_2}e^{-\frac{y}{\theta_2}},\ y>0.$$

However, the random variables X and Y are not identically distributed because neither of the above conditions for ID are satisfied. In particular, the two marginal densities belong to different families of densities ($f_x(x;\theta_1)$ belongs to the Pareto and $f_y(y;\theta_2)$ belongs to the Exponential families), they depend on different parameters $\theta_1\neq\theta_2$ and the two random variables X and Y have different ranges of values.

Example 4.39. Consider the three bivariate distributions (a)-(c) given below.

$x\diagdown y$	0	2	$f_x(x)$
1	0.18	0.12	0.3
2	0.42	0.28	0.7
$f_y(y)$	0.6	0.4	1

(a)

$x\diagdown y$	0	1	$f_x(x)$
0	0.18	0.12	0.3
1	0.42	0.28	0.7
$f_y(y)$	0.6	0.4	1

(b)

$x\diagdown y$	0	1	$f_x(x)$
0	0.36	0.24	0.6
1	0.24	0.16	0.4
$f_y(y)$	0.6	0.4	1

(c)

The random variables (X,Y) are independent in all three cases (verify!). The random variables in (a) are not Identically Distributed because $\mathbb{R}_X\neq\mathbb{R}_Y$, and $f_x(x)\neq f_y(y)$ for some $(x,y)\in\mathbb{R}_X\times\mathbb{R}_Y$. The random variables in (b) are not Identically Distributed because even though $\mathbb{R}_X=\mathbb{R}_Y$, $f_x(x)\neq f_y(y)$ for some $(x,y)\in\mathbb{R}_X\times\mathbb{R}_Y$. Finally, the random variables in (c) are Identically Distributed because $\mathbb{R}_X=\mathbb{R}_Y$, and $f_x(x)=f_y(y)$ for all $(x,y)\in\mathbb{R}_X\times\mathbb{R}_Y$.

Example 4.40. When $f(x,y;\boldsymbol{\theta})$ is *bivariate Normal*, as specified in (7), the two marginal density functions have the same functional form but $\boldsymbol{\theta}:=(\mu_1,\mu_2,\sigma_{11},\sigma_{22})$, $\boldsymbol{\theta}_1:=(\mu_1,\sigma_{11})$ and $\boldsymbol{\theta}_2:=(\mu_2,\sigma_{22})$, are usually different. Hence, for the random variables X and Y to be identically distributed, the two means and two variances should coincide: $\mu_1=\mu_2$ and $\sigma_{11}=\sigma_{22}$, i.e.

$$f(x;\boldsymbol{\theta}_1)=\tfrac{1}{\sqrt{2\pi\sigma_{11}}}e^{-\frac{1}{2\sigma_{11}}[x-\mu_1]^2},\quad f(y;\boldsymbol{\theta}_2)=\tfrac{1}{\sqrt{2\pi\sigma_{11}}}e^{-\frac{1}{2\sigma_{11}}[y-\mu_1]^2}.$$

The concept of Identically Distributed random variables can be easily extended to the n-variable case in a straight forward manner.

Identical Distributions. The random variables $(X_1,X_2,...X_n)$ are said to be *identically distributed* if:

$$f_k(x_k;\boldsymbol{\theta}_k)\equiv f(x_k;\boldsymbol{\theta}),\ \text{for all } k=1,2,...,n.$$

This has two dimensions:

(i) $f_1(.)=f_2(.)=f_3(.)=\cdots\equiv f_n(.)=f(.)$, and (ii) $\boldsymbol{\theta}_1=\boldsymbol{\theta}_2=\boldsymbol{\theta}_3=...=\boldsymbol{\theta}_n=\boldsymbol{\theta}$.

Example 4.41. Let X_1 and X_2 be independent Normal random variables with densities:

$$f(x_i)=\tfrac{1}{\sqrt{2\pi}}\exp(-\tfrac{(x_i-\mu_i)^2}{2}),\ x_i\in\mathbb{R},\quad i=1,2.$$

X_1 and X_2 are *not* identically distributed because they have different means: $\mu_1\neq\mu_2$.

6.2 A random sample of random variables

Our first formalization of condition [c] of a *random experiment* $\mathcal{E}$, where:

[c] it can be repeated under identical conditions,

took the form of a set of *random trials* $\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, ..., \mathcal{A}_n\}$ which are both *independent* and *identically distributed* (IID):

$$\mathbb{P}_{(n)}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap ... \cap \mathcal{A}_k) = \mathbb{P}(\mathcal{A}_1) \cdot \mathbb{P}(\mathcal{A}_2) \cdots \mathbb{P}(\mathcal{A}_k), \text{ for } k=2, 3, ..., n. \tag{42}$$

Using the concept of a *sample* $\mathbf{X}:=(X_1, X_2, ..., X_n)$, where X_i denotes the i-th trial, we can proceed to formalize condition [c] in the form of a set of random variables, $X_1, X_2, ..., X_n$, that are both *Independent* (I) and *Identically Distributed* (ID).

Random sample. The sample $\mathbf{X}^{\text{IID}}_{(n)}:=(X_1, X_2, ..., X_n)$ is called a *random sample* if the random variables involved are:

(a) **I**: $\quad f(x_1, x_2, ..., x_n; \boldsymbol{\phi}) \overset{\text{I}}{=} \prod_{k=1}^{n} f_k(x_k; \boldsymbol{\theta}_k)$ for all $\mathbf{x} \in \mathbb{R}^n_X$,

(b) **ID**: $\quad f_k(x_k; \boldsymbol{\theta}_k) = f(x_k; \boldsymbol{\theta})$, for all $k=1, 2, ..., n$,

where $\mathbf{x}:=(x_1, ..., x_n)$. That is, the joint density for $\mathbf{X}^{\text{IID}}_{(n)}:=(X_1, X_2, ..., X_n)$ is:

$$f(x_1, x_2, ..., x_n; \boldsymbol{\phi}) \overset{\text{I}}{=} \prod_{k=1}^{n} f_k(x_k; \boldsymbol{\theta}_k) \overset{\text{ID}}{=} \prod_{k=1}^{n} f(x_k; \boldsymbol{\theta}), \text{ for all } \mathbf{x} \in \mathbb{R}^n_X. \tag{43}$$

The first equality follows from the independence condition and the second from the Identical Distribution condition. NOTE that $f_k(x_k; \boldsymbol{\theta}_k)$ denotes the marginal distribution of $X_k(.)$, derived via:

$$f_k(x_k; \boldsymbol{\theta}_k) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, ..., x_{k-1}, x_k, x_{k+1}, ..., x_n; \boldsymbol{\phi}) dx_1 ... dx_{k-1} dx_{k+1} ... dx_n.$$

As argued in chapter 2, the formalization of a random experiment was chosen to motivate several concepts because it was simple enough to avoid unnecessary complications. It was also stated, however, that simple stochastic phenomena within the intended scope of a simple statistical model are rarely encountered in economics. One of our first tasks, once the transformation is completed, is to extend it. In preparation for that extension we note at this stage that the concept of a random sample is a very special form of what we call a sampling model.

Sampling model. A *sampling model* is a set of random variables $(X_1, X_2, ..., X_n)$ (a *sample*) with a certain *probabilistic structure*, which relates the observed data to the probability model.

In so far as the sampling model is concerned we note that from the modeling viewpoint the basic components of a random sample: $\mathbf{X}:=(X_1, X_2, ..., X_n)$, are the assumptions of (i) Independence, and (ii) Identical Distribution. The validity of these assumptions can often be assessed using a battery of graphical techniques, discussed in chapters 5-6, as well as formal misspecification tests discussed in chapter 15.

In an attempt to show how easy it is to end up with a non-random sample, it is shown in the next subsection that a simple re-arrangement of the sample gives rise to a non-random sample.

7 A simple statistical model

7.1 From a random experiment to a simple statistical model

At this stage, the mapping of the a simple statistical space $\left[(S,\Im,\mathbb{P}(.))^n,\ \mathcal{G}_n^{\text{IID}}\right]$ framed as a formalization of the concept of a random experiment $\mathcal{E}$, defined by the conditions [a]-[c] onto the real line has been completed. In chapter 3 the mapping took the form:

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

In this chapter we transformed the simple sampling space into a random sample:

$$\boxed{\mathcal{G}_n^{\text{IID}}=\{\mathcal{A}_1,\mathcal{A}_2,\mathcal{A}_3,...,\mathcal{A}_n\} \overset{X(.)}{\rightarrow} \mathbf{X}:=(X_1,X_2,...,X_n)\text{-IID.}}$$

Collecting the results of chapters 2-4 together we define a generic simple model in table 4.8.

Table 4.8: Simple generic statistical model

[i] Probability model: $\Phi=\{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta,\ x\in\mathbb{R}_X\}$,
[ii] Sampling model: $\mathbf{X}:=(X_1,X_2,...,X_n)$ is a random sample.

The concept of the statistical model constitutes a key contribution of probability theory to the theory of statistical modeling and inference. All forms of parametric statistical inference begin with a prespecified statistical model that constitutes a parsimonious description of the stochastic mechanism that could have given rise to the data in question. Hence, a sound understanding of the form and structure of a simple statistical model of the form given above is imperative.

7.2 Examples of simple Statistical models

7.2.1 Simple Normal model

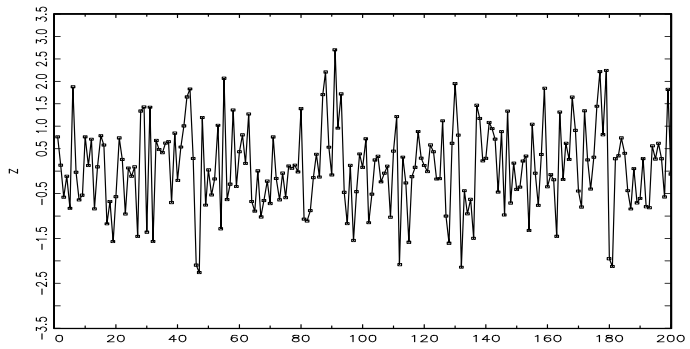
Mathematical world

Simple Normal model

[i] Probability model: $\Phi=\left\{f(x;\boldsymbol{\theta})=\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(x-\mu)^2}{2\sigma^2}),\ \boldsymbol{\theta}:=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+,\ x\in\mathbb{R}\right\}$,
$\mu=E(X),\ \sigma^2=Var(X)$

[ii] Sampling model: $\mathbf{X}:=(X_1,X_2,...,X_n)$ is a random sample.

Real world of data



A typical realization of a NIID process

7.2.2 Simple Log-Normal

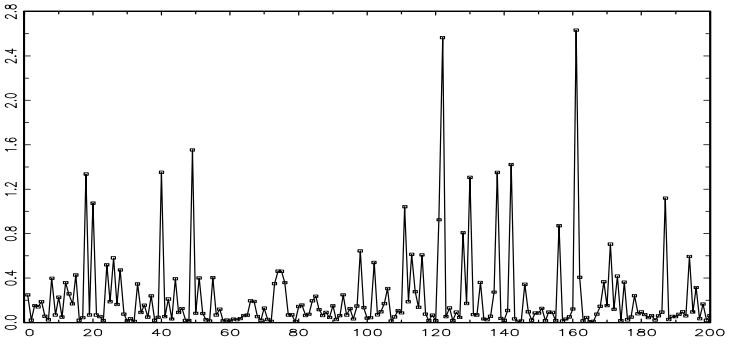
Mathematical world

Simple log-Normal model

[i] Probability model: $\Phi=\left\{f(x;\boldsymbol{\theta})=\frac{1}{x}\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\},\ \boldsymbol{\theta}:=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+,\ x\in\mathbb{R}_+\right\}$, $\mu=E(X),\ \sigma^2=Var(X)$

[ii] Sampling model: $\mathbf{X}:=(X_1,X_2,...,X_n)$ is a random sample.

Real world of data



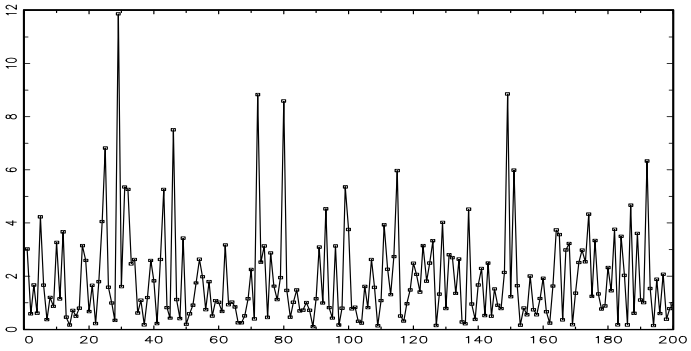
Typical realization of Log-Normal IID data

7.2.3 Simple Exponential model

Mathematical world

Simple Exponential model
[i] Probability model: $\Phi=\left\{f(x;\theta)=\frac{1}{\theta}e^{-\left(\frac{x}{\theta}\right)},\ \theta\in\mathbb{R}_+,\ x\in\mathbb{R}_+\right\}$ $\theta=E(X)$
[ii] Sampling model: $\mathbf{X}:=(X_1,X_2,...,X_n)$ is a random sample.

Real world of data



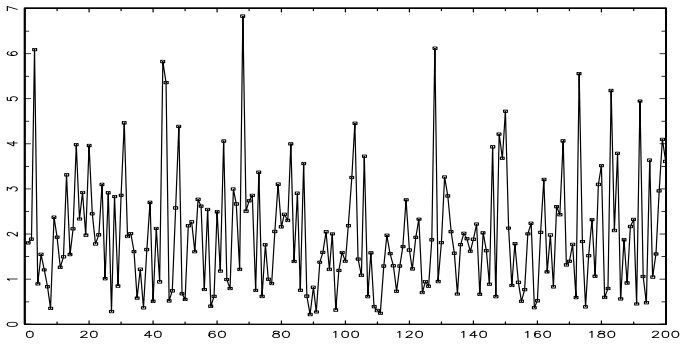
Typical realization of Exponential IID data

7.2.4 Simple Weibull model

Mathematical world

Simple Weibull model
[i] Probability model: $\Phi = \left\{ f(x;\boldsymbol{\theta}) = \frac{\beta x^{\beta-1}}{\alpha^{\beta}} \exp\left\{ -\left(\frac{x}{\alpha}\right)^{\beta} \right\},\ \boldsymbol{\theta} := (\alpha,\beta) \in \mathbb{R}^2_+,\ x>0 \right\}$
[ii] Sampling model: $\mathbf{X} := (X_1, X_2, ..., X_n)$ is a random sample.

Real world of data



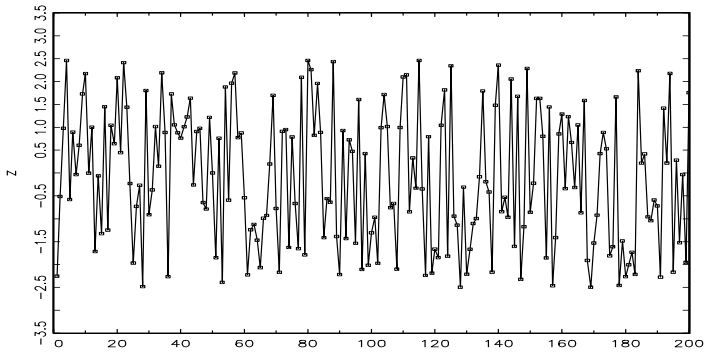
Typical realizationo of Weibull IID data

7.2.5 Simple Uniform model

Mathematical world

Simple Uniform model
[i] Probability model: $\Phi = \left\{ f(x;\boldsymbol{\theta}) = \frac{1}{(b-a)},\ \boldsymbol{\theta} := (a,b),\ a \leq x \leq b \right\}$
[ii] Sampling model: $\mathbf{X} := (X_1, X_2, ..., X_n)$ is a random sample.

Real world of data



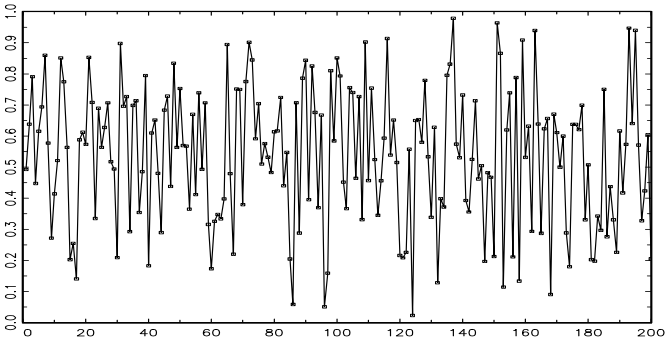
Typical realization of Uniform IID data

7.2.6 Simple Beta model

Mathematical world

Simple Beta model
[i] Probability model: $\Phi=\left\{f(x;\boldsymbol{\theta})=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B[\alpha,\beta]},\boldsymbol{\theta}:=(\alpha,\beta)\in\mathbb{R}^2_+,\ 0\leq x\leq 1\right\}$
[ii] Sampling model: $\mathbf{X}:=(X_1,X_2,...,X_n)$ is a random sample.

Real world of data



Typical realization of Beta (α,β) IID data

7.2.7 Simple Student's t model

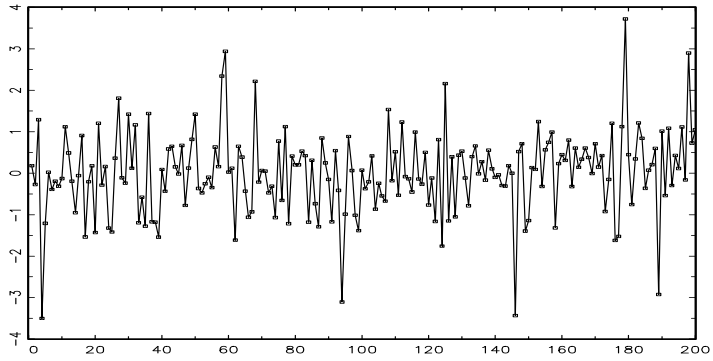
Mathematical world

Simple Student's t model

[i] Probability model: $\Phi=\left\{ f(x;\boldsymbol{\theta})=\frac{\Gamma[\frac{1}{2}(\nu+1)](\sigma^2\nu\pi)^{-\frac{1}{2}}}{\Gamma[\frac{1}{2}\nu]}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{1}{2}(\nu+1)},\ \boldsymbol{\theta}:=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+,\ x\in\mathbb{R}\right\}$

[ii] Sampling model: $\mathbf{X}:=(X_1,X_2,...,X_n)$ is a random sample.

Real world of data



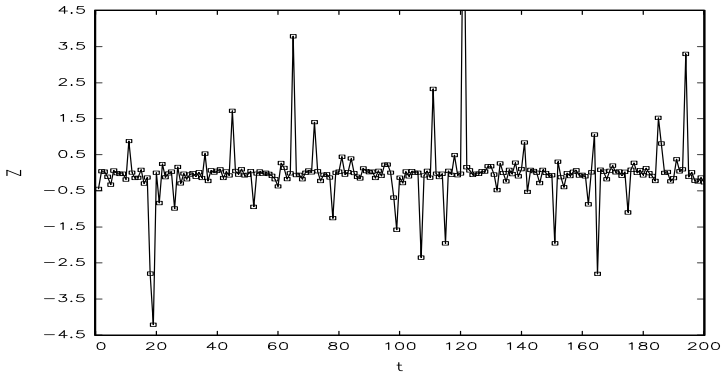
Typical realization of Student's t IID data

7.2.8 Simple Cauchy model

Mathematical world

Simple Cauchy model
[i] Probability model: $\Phi=\left\{f(x;\boldsymbol{\theta})=\frac{1}{\pi\beta[1+\{(x-\alpha)^2/\beta\}},\ \boldsymbol{\theta}:=(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}_+,\ x\in\mathbb{R}\right\}$
[ii] Sampling model: $\mathbf{X}:=(X_1,X_2,...,X_n)$ is a random sample.

Real world of data



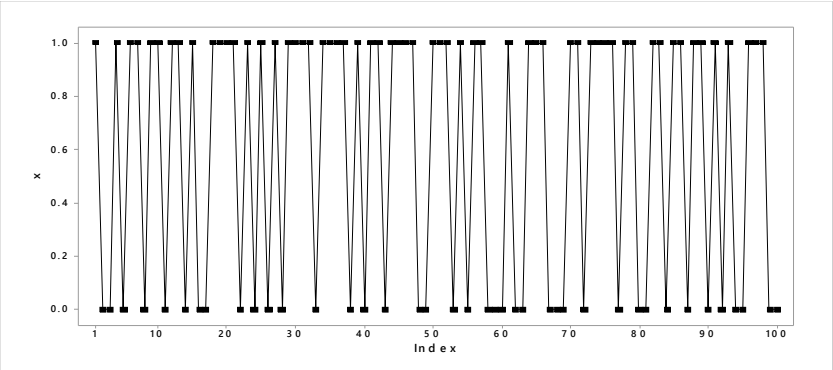
Typical realization of Cauchy IID data

7.2.9 Simple Bernoulli model

Mathematical world

Simple Bernoulli model
[i] Probability model: $\Phi=\{f(x;\theta)=\theta^x(1-\theta)^{1-x}, 0\leq\theta\leq 1, x=0,1\},\ \theta=E(X)$
[ii] Sampling model: $\mathbf{X}:=(X_1, X_2, ..., X_n)$ is a random sample.

Real world of data



Typical realization of Bernoulli IID data

As mentioned above, every form of statistical inference begins with a prespecified statistical model. This specification amounts to choosing a set of probabilistic assumptions which the modeler deems appropriate for describing the stochastic mechanism that gave rise to the data set in question. The choice of an appropriate statistical model constitutes perhaps the most difficult, and at the same time the most crucial, decision a modeler has to make; in comparison, the decision of choosing a good estimator for θ is trivial.

What renders the above statistical model *simple* is the assumption of a *random sample*, that is $(X_1, X_2, ..., X_n)$ are IID random variables and thus the distribution of the sample takes the simple form:

$$f(x_1, x_2, ..., x_n; \boldsymbol{\theta})=\prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}_X^n.$$

Making an appropriate choice of a statistical model will require the modeler to develop both a formal and an intuitive understanding of these probabilistic assumptions. Due to the simplicity of the random experiment, its formalization gives rise to a statistical model which is not adequate for modeling most stochastic phenomena in many fields, including econometrics.

Important concepts

Random vector, bivariate and multivariate distributions, joint density function, bivariate Exponential and Normal distributions, joint moments, covariance, skewness and kurtosis coefficients for bivariate distributions, marginal distributions, conditional distributions, conditional moments (raw and central), truncation, hazard function, independence among random variables, identical distributions for random variables, functions of random variables, distributions of functions of random variables, ordered sample, distributions of ordered statistics, simple (generic) statistical model, simple Bernoulli model, simple Normal model, statistical identification of parameters, parameterization, reparameterization.

Crucial distinctions

Discrete vs. continuous random vectors, conditional probability vs. conditional distributions, marginalization vs. conditioning, conditioning on events vs. conditioning on random variables, marginal vs. conditional moments, sampling space vs. sampling model, statistical vs. substantive parameterizations, statistical vs. structural identification.

Essential ideas

- From a mathematical perspective, a random variable X maps a simple sampling space $\mathcal{G}_n^{\text{IID}}=\{\mathcal{A}_1,\mathcal{A}_2,\mathcal{A}_3,...,\mathcal{A}_n\}$ into a simple sampling model $\mathbf{X}:=(X_1,X_2,...,X_n)$, where the elements of $\mathbf{X}$ are Independent and Identically Distributed (IID), preserving the event structure $\Im$ of the original probability space $(S,\Im,\mathbb{P}(.))$.
- The concepts of a random variable $X(.)$ in conjunction with the concept of a σ-field $\Im$ render going from one random variable to a random vector $\mathbf{X}:=(X_1,X_2,...,X_n)$ easy, because joint distributions $f(x_1,x_2,...,x_n)$, involve the joint occurrence of $X_1,X_2,...,X_n$ which is well-defined because $\Im$ is close under intersections.
- Conditioning an event A on a event B is mathematically and conceptually very different from conditioning a random variable Y on a related random variable X, since random variables define more than one event; often an infinite number.
- Conditional distributions provide the key to modeling dependence among many random variables; see chapters 6-7.
- The concepts of Independence and Identically Distributed (IID) for a random vector $\mathbf{X}:=(X_1,X_2,...,X_n)$ presuppose the designation of an ordering $(1,2,...,n)$ relative to which they are defined.
- It is imperative to distinguish between the statistical and the substantive parameterizations of interest, and ensure that both sets of parameters are identified.

8 Questions and Exercises

1. Consider the discrete Uniform distribution with density $f_x(x;\theta)=\frac{1}{n+1}$, n is an integer, $x=0,1,2,...,n$. Derive $E(X)$ and $Var(X)$; note that: $\sum_{k=0}^{n}k=\frac{1}{2}(n+1)$, $\sum_{k=0}^{n}k^2=\frac{1}{6}n(2n+1)(n+1)$

2. "Marginalizing amounts to throwing away all the information relating to the random variable we are summing (integrating) out". Comment.

3. Consider the random experiment of tossing a coin twice and define the random variables: X-number of H's, and Y=|number of H's - number of T's|.

Derive the joint distribution of (X, Y), assuming a fair coin, and check whether the two random variables are independent.

4. Let the joint density function of two random variables X and Y be:

$x \diagdown y$	-1	1
-1	.2	.1
0	.2	.1
1	.2	.2

(a) Derive the marginal distributions of X and Y. (b) Determine whether X and Y are independent.

(c) Verify your answer in (b) using the conditional distribution(s).

5. Define the concept of independence for two random variables X and Y in terms of the joint, marginal and conditional density functions.

6. Explain the concept of a *random sample* and explain why it is often restrictive for most economic data series.

7. Describe briefly the formalization of the condition: [c] we can repeat the experiment under identical conditions, in the form of the concept of a random sample.

8. Explain intuitively why it makes sense that when the joint distribution $f(x, y)$ is Normal the marginal distributions $f_x(x)$ and $f_y(y)$ are also Normal.

9. Define the raw and central joint moments and show that:
$Cov(X, Y)$=$E(XY)$−$E(X)\cdot E(Y)$. Why do we care about these moments?

10. (a) Explain the concept of an ordered sample.

(b) Explain intuitively why an ordered random sample is neither independent nor identically distributed.

11. Explain the concepts of identifiability and parameterization.

12. "In relating statistical models to (economic) theoretical models we often need to reparameterize/restrict the former in order to render the estimated parameters theoretically meaningful." Explain.