MONASH UNIVERSITY DEPARTMENT OF MATHEMATICS

Approaches to Conditioning

Honours Thesis

Konstantin Gredeskoul
Supervisor: Dr Joseph G. Kupka

October, 1995

Abstract

In this thesis we shall discuss and analyze different approaches to conditioning and their applications.

The traditional Measure-Theoretic graduate-level approach is examined and compared to a relatively new approach that is based on the construction of families of regular conditional probabilities. The latter one has many advantages over the traditional, being more intuitive, simple and applicable to real-life problems, while remaining rigorous and quite general.

As an important followup to conditional probabilities, the concept of conditional density and distribution functions is attentively studied in many details, using both approaches as starting points.

Several examples and applications of the methods are studied and analyzed. The analysis part was partially accomplished using statistical packages MINITAB and S-PLUS. S-PLUS was also used to graph the results. The numerical integration was done in algebraic package MAPLE-V. The rest of the analysis and simulations were conducted using written-on-the-fly compact programs in C.

Acknowledgments

The author is immeasurably grateful[1] to his supervisor, Dr. Joseph G. Kupka. His support, advice and, more importantly, constructive criticism, were of the highest possible value throughout the year. We discussed most of the ideas presented in the thesis and his response was always competent, impeccably correct[2] and very useful.

Many thanks go to my lovely wife Masha, who completed her thesis just two days before I completed mine. Despite the fact that she was also engaged in the last Honours year, she was always very supportive and inspiring.

And finally, I wish to thank my parents, my sister and my grandparents for making my life wonderful and happy for, already, 21 years.

Thanks you all very much!

Konstantin Gredeskoul
27th October, 1995
Monash University
Melbourne, Australia.

[1] There is a doubt if one can define any measure on the space of gratefulness. Since it is certainly beyond the scope of the thesis to discuss different aspects of defining measure on such uncommon spaces, we shall always assume that gratefulness has an infinite measure.

[2] Almost everywhere.

"... Well, in our country", said Alice, "you'd generally get to somewhere else — if you ran very fast for a long time, as we've been doing."

"A slow sort of country!" said the Queen. "Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!..."

Alice and the Red Queen
"Through the Looking Glass"
Lewis Carroll

Contents

1 Introduction

Conditional probability arises naturally in elementary probability theory. For two events A and B, the probability of an event A given that event B has occured is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{1}$$

if the conditioning event B has a non-zero probability. Although this definition makes a lot of sense in many intuitive examples, it is very important to investigate whether the restriction $P(B) \neq 0$ is really necessary.

In reality, very often we would like to be able to handle conditional probabilities where the conditioning event is known to have probability zero. As an example, suppose we wish to introduce or define the *conditional distribution function* of two continuous random variables. Applying the above elementary definition (1) to the distribution function we run into various sorts of problems. Indeed, let X and Y be two random variables of continuous type with ranges R_X and R_Y respectively. Then, according to the elementary definition of the conditional probability, we can write

$$\begin{aligned} F(y|x) &= P(Y \leq y|X = x) \\ &= P(Y \leq y, X = x)/P(X = x) \\ &= 0/0 \\ &= ?, \end{aligned}$$

since $P(X = x) = 0$ for all $x \in \mathbb{R}$, as X is continuous. Thus, Definition (1) is not capable of dealing with such cases, which are in no sense degenerate!

As Kolmogorov (1933, page 51) noted, "...the concept of conditional probability with regard to an isolated hypothesis whose probability equals 0 is inadmissible." Unfortunately, the rigorous approach to this problem is far beyond boundaries of elementary probability theory. It is aimed in this work to give a summary of two different approaches to defining and using conditional probabilities with the least number of restrictions. The following approaches will be discussed:

1. The measure theoretic approach. This can be thought of as a traditional graduate-level approach which requires some broad knowledge of measure theory.

2. Approach via Regular Conditional Probabilities. The approach was developed by J. Chang and D. Pollard (Yale University, 1993) and reconstructed by J. Kupka (Monash University, 1994) so that it contains a rigorous treatment of conditional probabilities with almost no measure theory and is introduced via the concept of expectation; hence it should be suitable for teaching at the undergraduate level.

In general, it is a deep theoretical problem to provide a universal rigorous definition in place of (1). It requires a lot of abstract mathematics. We can, however, approach the problem from a more particular point of view — we aim to define $P(A|X=x)$ for some random variable X and *any* event A from our event space Ω. *Thus we will narrow B to an event* $\{X=x\}$. And yet the task is very important because, for example, the conditional distribution function of two continuous random variables mentioned above can be *defined* as

$$F(y|x) = P(Y \leq y|X=x),$$

which does not make sense unless $P(A|X=x)$ is defined for a continuous random variable X.

The basic idea behind approaches 1 and 2 is the same: if the traditional conditional probability of an event A given $X=x$ is considered by fixing X at x and, perhaps, varying A, a more general approach is obtained by a change of viewpoint, namely, by considering $P(A|X=x)$ for *fixed* A as a function of x. However, different treatments of this idea lead to different constructions which will be studied here.

2 Notation and Basic Theorems

Before proceeding to conditional probabilities, we should introduce the notation used throughout the thesis.

2.1 Measure-Theoretic Background

Let $\mathcal{F}$ be a collection of subsets of a set Ω. Then $\mathcal{F}$ is called a *σ-field* if $\Omega \in \mathcal{F}$ and $\mathcal{F}$ is closed under complementation and countable union, that is,

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
3. $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

The smallest σ-field containing all n-dimensional intervals (rectangles) $(a, b]$ with $a, b \in \mathbb{R}^n$, is called *the class of Borel sets of* $\mathbb{R}^n$, written $\mathcal{B}(\mathbb{R}^n)$ or just $\mathcal{B}$. A *measure* μ on a σ-field $\mathcal{F}$ is a non-negative extended real-valued countably additive function μ on $\mathcal{F}$. Explicitly, $\mu : \mathcal{F} \to [0, +\infty]$ and whenever $A_1, A_2, \ldots$ form a finite or countably infinite collection of disjoint sets in $\mathcal{F}$ we have $\mu\left(\bigcup_n A_n\right) = \sum_n \mu(A_n)$. μ is called a *probability measure* if $\mu(\Omega) = 1$. The triple $(\Omega, \mathcal{F}, \mu)$ is then called a *probability measure space.* If $(\Omega, \mathcal{F})$ is a measurable space and $f : \Omega \to \mathbb{R}^n$, then f is said to be *Borel measurable* if f is measurable relative to the σ-fields $\mathcal{F}$ and $\mathcal{B}$, ie. $f^{-1}(B) \in \mathcal{F}$ for each Borel set $B \in \mathcal{B}$. When $n = 1$, a sufficient condition for Borel measurability is that $\{\omega : f(\omega) > c\} \in \mathcal{F}$ for any real c. This is sometimes taken as a definition in the one-dimensional case. A certain condition is said to hold *almost everywhere* with respect to the measure μ (written a.e. $[\mu]$) if there is at most a set $B \in \mathcal{F}$ of μ-measure 0 such that condition holds outside of B. In other words, *almost everywhere* means "everywhere, with a possible exception of a set of measure 0".

A nonnegative, finitely additive set function μ on the field $\mathcal{F}$ is said to be *σ-finite* on $\mathcal{F}$ if and only if Ω can be written as $\bigcup_{n=1}^{\infty} A_n$ where the A_n belong to $\mathcal{F}$ and $\mu(A_n) < \infty$ for all n. If μ is a measure on the σ-field $\mathcal{F}$ and λ is a signed measure on $\mathcal{F}$ we say that λ is *absolutely continuous* with respect to μ if and only if $\mu(A) = 0$ implies $\lambda(A) = 0$ $(A \in \mathcal{F})$.

The following two theorems will be used in the measure theoretic context of the thesis:

Theorem 1 (Radon-Nikodym) *Let μ be a σ-finite measure and λ a signed measure on the σ-field $\mathcal{F}$ of subsets of Ω. Assume that λ is absolutely continuous with respect to μ. Then there is a Borel measurable function $g : \Omega \to \bar{\mathbb{R}}$ such that*

$$\lambda(A) = \int_A g d\mu, \quad \textit{for all } A \in \mathcal{F}.$$

If h is another such function, then $g = h$ a.e. $[\mu]$.

Theorem 2 (Fubini) *Let $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$, $\mu = \mu_1 \times \mu_2$, where μ_j is a σ-finite measure on $\mathcal{F}_j$, $j = 1, 2$. If f is a Borel measurable function on $(\Omega, \mathcal{F})$ such that $\int_\Omega f d\mu$ exists, then*

$$\begin{aligned}\int_\Omega f d\mu &= \int_{\Omega_1}\int_{\Omega_2} f d\mu_2 d\mu_1 \\ &= \int_{\Omega_2}\int_{\Omega_1} f d\mu_1 d\mu_2 \quad .\end{aligned}$$

2.2 Probability Theory Framework

A *sample space* Ω is a set whose points are in one-to-one correspondence with the possible outcomes of a random experiment. Let $\mathcal{F}$ be a σ-algebra on Ω. An *event* A is defined to be an $\mathcal{F}$-measurable subset of Ω, i.e. $A \in \mathcal{F}$. A *probability* P on Ω is a probability measure on $(\Omega, \mathcal{F})$, as defined earlier.

To simplify the meaning of an event we can say that among all subsets of Ω, most will be measurable but a few won't. We then call all the measurable subsets of Ω *events* and ignore non-measurable subsets.

A *random variable* X on a probability space $(\Omega, \mathcal{F}, P)$ is a Borel measurable function from Ω onto $R_X \subset \bar{\mathbb{R}}$[3]. The set R_X is called the *range* of X. A random variable X is said to be *discrete* if there exists a set E with $P(X \in E) = 0$ such that the range of X on the compliment E^c is finite or countably infinite. The *distribution function* of a random variable X is the function $F = F_X$ from $\mathbb{R}$ to the interval $[0, 1]$ given by $F(x) = P\{\omega : X(\omega) \leq x\}$, $x \in \mathbb{R}$. The *probability function* p_X of a discrete random variable is defined to be $p_X(x) = P\{X = x\}$, $x \in \mathbb{R}$. The *probability measure on* $\mathbb{R}$ *induced by a random variable* X is defined by $P_X(A) = P\{X \in A\}$ for all $A \in \mathcal{B}(\mathbb{R})$. A random variable X is said to be *absolutely continuous* if there exists a nonnegative real-valued Borel measurable function f on $\mathbb{R}$ such that $F(x) = \int_{-\infty}^{x} f(t)dt$, $x \in \mathbb{R}$, where the dt denotes integration with respect to Lebesgue measure. Such a function f is then called a *density function* of X. A random variable X is said to be *continuous* if its distribution function F_X is continuous on all of $\mathbb{R}$.

And finally, here are several propositions which will be used later. A random variable X is continuous if and only if for all $x \in \mathbb{R}$ we have $P(X = x) = 0$. If a random variable X is absolutely continuous, it is also continuous. The converse of this is not true, since continuity of the distribution function F_X does not automatically imply the existence of a density function.

[3] $\bar{\mathbb{R}} \overset{def}{=} \mathbb{R} \cup \{\infty\}$

3 Measure-Theoretic Approach

3.1 Definitions

Theorem 3 *Let X be a random variable and let A be a fixed event in Ω. Then there is a nonnegative real-valued Borel measurable function g_A such that for each set $B \in \mathcal{B}$,[4].*

$$P(\{X \in B\} \cap A) = \int_B g_A(x) dP_X(x)$$

Further, if h is another such function then $g = h$ almost everywhere $[P_X]$.

Proof:

Let $\mu = P_X$ and $\lambda(B) = P(\{X \in B\} \cap A)$ for $B \in \mathcal{B}$. Then λ is a finite measure on $\mathcal{B}$ and absolutely continuous with respect to μ, since $P_X(B) = 0$ implies $P(\{X \in B\} \cap A) = 0$, hence $\lambda(B) = 0$. The result then follows from the Radon-Nikodym theorem. □

The above theorem is a key to the measure-theoretic definition we need. Intuitively, conditional probability $P(A|X = x)$ should satisfy

$$P(\{X \in B\} \cap A) = \int_B P(A|X = x) dP_X(x)$$

and since the theorem tells us that such a function $g_A(x)$ exists and is essentially unique, we may *define* the conditional probability $P(A|X = x)$ to be the function $g_A(x)$. Thus we have,

Definition 3.1 *The **conditional probability** $P(A|X = x)$ is defined to be $g_A(x)$. It is essentially unique for a given A.*

This definition is indirect: to show that some probability is conditional we should first *guess* the function $g_A(x)$ and then *verify* the equation of Theorem 3. The first step may require a good deal of intuition, the second — some messy integration. Thus, while the above definition is rigorous and fills the gap, it is not very intuitive and not easy to use without some knowledge of measure theory.

Anyway, since we still have an abstract definition of conditional probability, we can ask ourselves whether the old-fashioned $P(A|B) = P(A \cap B)/P(B)$ coincides with the above definition for "normal" cases, i.e. those for which $P(B) > 0$. We can illustrate that this is so by considering a discrete random variable X over it's *essential range* defined by $R_X = \{x \in \mathbb{R} : P(X = x) > 0\}$.

[4] By $dP_X(x)$ we mean integration with respect to the probability measure induced by X

3.2 Discrete Case

Let X be a discrete random variable with probability function p_X defined by $p_X(x) = P(X = x)$. Note that $p_X(x) \neq 0$ for all $x \in R_X$, so the "traditional" definition

$$P(A|X = x) = \frac{P(\{X = x\} \cap A)}{P(X = x)}$$

makes sense over R_X. We wish to show that it satisfies the equation of Theorem 3 and thus coincides with the newly introduced $g_A(x)$.

To show that, we start with $P(\{X \in B\} \cap A)$ for an event $A \subset \Omega$ and B being a Borel subset of $\mathbb{R}$.

$$\begin{aligned} P(\{X \in B\} \cap A) &= \sum_{x \in B} P(\{X = x\} \cap A) \\ &= \sum_{x \in B} P(A|X = x)P(X = x) \\ &= \sum_{x \in B} P(A|X = x)p_X(x) \\ &= \int_B P(A|X = x)dP_X(x) \end{aligned}$$

from which it follows that

$$g_A(x) = P(A|X = x)$$

since $g_A(x)$ is uniquely determined on R_X by the Theorem 3. Indeed, if $P(X = x) > 0$, then $g = h$ a.e.$[P_X]$ means simply that $g(x) = h(x)$ for all $x \in R_X$. Hence, when X is discrete and the probability that $X = x$ is not zero, the conditional probability defined above coincides with the traditional definition. □

It is important to mention that the integral $\int_B f(x)dP_X(x)$ is taken with respect to the measure P_X induced by X, so that in this discrete case the integral becomes $\sum_{x \in B} f(x)p_X(x)$.

3.3 Bivariate Case

3.3.1 Conditional density function

This section is devoted to another very important application of Definition 3.1. Here we will be dealing with two jointly continuous random variables X and Y with ranges $R_X \subset \mathbb{R}$ and $R_Y \subset \mathbb{R}$. Let also the distribution function of X be F and let the distribution of Y depend on a sample value x of X. And finally, let X and Y have joint density function $f_{X,Y}$ and marginal densities f_X and f_Y.

Our aim is to rigorously define the conditional density function $h(y|x)$. Although the task may look simple, many books at the undergraduate level introduce the conditional density function as we'll do below, but very rarely give a justification in terms of probabilities. See, for example, Hogg and Craig (1970, page 64).

Definition 3.2 *The* ***conditional density*** *of Y given $X = x$ is defined to be*

$$h(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \tag{2}$$

where $f_X(x) \neq 0$ and 0 otherwise.

There is nothing wrong with this definition in the sense that we have introduced a new symbol $h(y|x)$ and called it a "conditional density". However, we would also like to know whether this definition is related to the meaning of conditional probability as defined earlier.

To answer this question we must understand what the desired properties of the density $h(y|x)$ are. So let $B, C \in \mathcal{B}(\mathbb{R})$ and let $A = \{Y \in C\} \in \mathcal{F}$. First of all, since $h(y|x)$ is a density function *of Y given that $X = x$* we should expect that $P(Y \in C|X = x)$ can be obtained by integrating $h(y|x)$ with respect to y over C, i.e.

$$P(A|X = x) = P(Y \in C|X = x) = \int_C h(y|x)dy.$$

This equation is, therefore, the required connection with conditional probability which is defined in terms of Definition 3.1. So if we can prove the above equation we can *justify* the existence and meaningfulness of the symbol $h(y|x)$ and thus, Definition 3.2.

Keeping in mind that A was defined as the event $\{Y \in C\}$, we need to show that

$$P(\{X \in B\} \cap \{Y \in C\}) = \int_B \left[\int_C h(y|x)dy\right] f_X(x)dx, \tag{3}$$

thus proving that $g_A(x) = P(Y \in C|X = x) = \int_C h(y|x)dy$ as stated above.

Starting with the left-hand-side of (3) and noting that (2) can be rewritten as $f_{XY}(x,y) = h(y|x)f_X(x)$ where $f_X(x) \neq 0$ we get

$$\begin{aligned} P(\{X \in B\} \cap \{Y \in C\}) &= P((X,Y) \in B \times C) \\ &= \int_{B\times C} f_{XY}(x,y)dydx \\ &= \int_B \int_C f_{XY}(x,y)dydx \quad \text{by the Fubini Theorem} \\ &= \int_B \int_C h(y|x)f_X(x)dydx \\ &= \int_B \left[\int_C h(y|x)dy\right] f_X(x)dx \end{aligned}$$

as required to establish $P(Y \in C|X = x) = \int_C h(y|x)dy$. □

Note, that C and B above are one-dimensional subsets of $\mathbb{R}$. A slightly more general result may be obtained by considering C as a subset of $\mathbb{R}^2$, so that event A becomes $A = \{(X, Y) \in C\}$. So let $C(x) = \{y : (x, y) \in C\}$ be the projection onto the y-axis of the intersection of the two-dimensional region C and the line $\{X = x\}$. The regions are shown on the Figure 1.

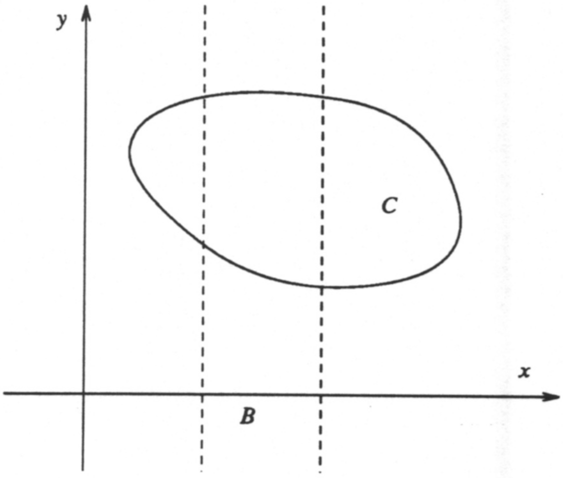

Figure 1: Regions $B \times \mathbb{R}$ and C in the xy-plane.

To prove that

$$P(A|X = x) = P((X, Y) \in C|X = x) = \int_{C(x)} h(y|x)dy,$$

we proceed as before,

$$\begin{aligned}
P(\{X \in B\} \cap \{(X, Y) \in C\}) &= P(\{(X, Y) \in B \times \mathbb{R}\} \cap \{(X, Y) \in C\}) \\
&= P((X, Y) \in (B \times \mathbb{R}) \cap C) \\
&= \int_{(B\times\mathbb{R})\cap C} f_{X,Y}(x, y)dydx \\
&= \int_B \int_{C(x)} f_{X,Y}(x, y)dydx \quad \text{(Fubini Theorem)} \\
&= \int_B \int_{C(x)} h(y|x)f_X(x)dydx \\
&= \int_B \left[\int_{C(x)} h(y|x)dy\right] f_X(x)dx \\
&= \int_B P(A|X = x)f_X(x)dx
\end{aligned}$$

as claimed. □

The above approach to the definition of the conditional density function of two continuous random variables is used in Ash (1972), however the author only considers the $A = \{(X, Y) \in C\}$ case where he makes very few or no justifications. For example, Ash assumes and then uses the formula

$$P(Y \in C(x)|X = x) = \int_{C(x)} h(y|x) f_X(x) dx.$$

However, according to Definition 3.1 we only know how to handle a *fixed* (with respect to x) event A. The event $\{Y \in C(x)\}$ is no longer fixed — it varies with x, so the conclusions of this assertion can lead to unjustified or wrong results.

We did not use this assertion and yet, justified every single step above. So the results should be clear and less confusing.

To summarize, the aim of this example was to produce a meaningful definition of conditional density of two continuous random variables and establish a relation between the density and conditional probability in the sense of Definition 3.1.

3.3.2 Reverse conditioning

In the light of the above discussion, we can solve the following interesting problem.

Let X be an absolutely continuous random variable. If $X = x$, let Y be another absolutely continuous random variable with conditional density $f_{Y|X}(y|x)$. We already know how to find $P(Y \in C|X = x)$; the expression $P(Y \in C|X = x)$ is then interpreted as a probability of an event determined by Y given that some information about X is known.

Conversely, it would be interesting to know whether the knowledge of Y can tell us something about X. So we aim to find $P(X \in B|Y = y)$ in this section.

Since X and Y are jointly absolutely continuous, their joint density function exists and satisfies

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Since the conditional density is defined by the above formula independently of any experiment, we can simply interchange x and y to get

$$f_{X|Y}(x|y) = \frac{f_{Y,X}(y,x)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

since, clearly, $f_{Y,X}(y,x) = f_{X,Y}(x,y)$ for all $x, y \in \mathbb{R}^2$. Because also

$$f_Y(y) = \int f_{X,Y}(t,y) dt,$$

we can write

$$f_{X|Y}(x|y) = \frac{f_X(x) f_{Y|X}(y|x)}{\int f_X(t) f_{Y|X}(y|t) dt} \tag{4}$$

and hence, finally,

$$P(X \in B|Y = y) = \int_B \frac{f_X(x) f_{Y|X}(y|x)}{\int f_X(t) f_{Y|X}(y|t) dt} dx.$$

In fact, the above formula (4) is the continuous prototype of the well-known Bayes Rule, which, in the discrete case, states

$$P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X = x)}{\sum_t P(Y = y|X = t)P(X = t)}.$$

This formula gives a probability for a "cause given effect" — the reverse of the usual "effect given cause" model.

3.4 Example

This section is devoted to a simple and intuitive example to demonstrate another application of Definition 3.1. Later, we shall approach this example from another starting point — with a different definition of conditional probability.

Let X be a uniform random variable between 0 and 1. Once the value x of X is known, let Y be a binomial random variable with probability of success $p = x$ and some fixed number of trials n. Let's denote by y a sample realization of Y. Suppose that we are interested in the unconditional distribution of Y, i.e. $P(Y = i)$ for some fixed $i = 0, 1, \ldots, n$.

To solve the problem, we first note that $f_X(x) = 1$ if $x \in [0, 1]$ and 0 otherwise. Also, we intuitively assume that for $A = \{Y = y\}$,

$$g_A(x) = P(Y = y|X = x) = \binom{n}{y} x^y (1 - x)^{n-y}$$

is the conditional probability function of Y given that $X = x$.

Then, for $i \in R_Y$, for $A = \{Y = i\}$ and $B = \mathbb{R}$, the required probability distribution can be calculated as follows:

$$\begin{aligned} P(A) &= P(\{X \in B\} \cap A) \\ &= \int_B g_A(x) dP_X(x) \\ &= \int_{[0,1]} P(Y = i|X = x) \times 1dx, \\ \text{hence,} \quad P(Y = i) &= \int_0^1 \binom{n}{i} x^i (1 - x)^{n-i} \times 1dx \\ &= \frac{n!}{i!(n-i)!} \times \frac{\Gamma(i+1)\Gamma(n-i+1)}{\Gamma(i+1+n-i+1)} \times 1 \\ &= \frac{n!}{i!(n-i)!} \times \frac{i!(n-i)!}{(n+1)!} \\ &= \frac{1}{n+1}, \end{aligned}$$

by recognizing the Beta$(i + 1, n - i + 1)$ density function under the integral sign. Hence Y is *rectangular* (or uniform) on $\{0, 1, \ldots, n\}$.

3.5 Summary

The measure theoretic approach provides a very large set of tools for probability theory. It is, essentially, the only way of making a rigorous approach to probability and in particular to conditional probabilities. It was shown that the approach described above leads to expected and known results in simple cases (e.g. discrete), while filling the gap in previously undefined cases (e.g. continuous).

From another point of view, even if it is necessary to carry out research or experiment in probability with understanding and rigor, learning the measure theoretic background may be too time consuming. For that reason, the introduction of some other approach may be useful.

As we shall see, the approach via Regular Conditional Probabilities does not require knowledge of measure theory, while being both intuitive and rigorous. We shall now proceed to the next section, which is devoted to this.

4 Regular Conditional Probabilities

4.1 Definitions

Definition 4.1 *Let X be a random variable on $(\Omega, \mathcal{F}, P)$ with range R_X. A* ***regular conditional probability*** *for X is a family $\{P_x\}_{x \in R_X}$ such that*

1. *$\forall x \in R_X,\ P_x$ is a probability on $(\Omega, \mathcal{F})$*
2. *$\forall x \in R_X,\ P_x(X \neq x) = 0$*
3. *$\forall A \in \mathcal{F}, P(A) = \mathbf{E}[g_A(X)]$, where $g_A(x) = P_x(A)$.*

The symbol $P(A|X = x)$ is then identified with $P_x(A)$.

Also, to avoid the use of superfluous g_A notation, we shall often write

$$\begin{aligned} P(A) &= \mathbf{E}[g_A(X)] \\ &= \mathbf{E}_{P_X}[g_A(x)] \\ &= \mathbf{E}_{P_X}[P(A|X = x)] \\ &= \mathbf{E}_{P_X}[P_x(A)] \\ &= \int P(A|X = x) dP_X(x) \end{aligned}$$

This approach does not require knowledge of measure theory; one must only be familiar with the concept of expectation to understand the definition. This approach is also more intuitive and provides better tools for calculating and finding conditional probabilities, as will be illustrated.

Terminology: any model that makes assumptions about regular conditional probabilities will be referred to as an **rcp** model.

Given the above definition, it is important to verify whether P_x satisfies the condition of Measure-Theoretic Definition 3.1. This result is established by the following lemma, which will also be used later.

Lemma 4 *For a fixed $A \in \mathcal{F}$ and for all $B \in \mathcal{B}(\mathbb{R})$,*

$$P(X \in B, A) = \int_B P_x(A) dP_X(x).$$

Proof:

We have,

$$\begin{aligned} P_x(X \in B, A) &= P_x(X \in B, A, X = x) + P_x(X \in B, A, X \neq x) \\ &= P_x(X \in B, A, X = x) + 0 \\ &= P_x(A) I_B(x), \end{aligned}$$

where $I_B(x)$ is the indicator function[5] of B. This is true because

$$(X \in B, X = x) = \begin{cases} (X = x) & \text{if } x \in B \\ \emptyset & \text{if } x \notin B \end{cases}$$

It now follows that

$$\begin{aligned} P(X \in B, A) &= \int P_x(X \in B, A) dP_X(x) \\ &= \int P_x(A) I_B(x) dP_X(x) \\ &= \int_B P_x(A) dP_X(x), \end{aligned}$$

as required. □

Therefore, $P_x(A)$ is the same as $g_A(x)$ introduced in the Definition 3.1.

4.2 Discrete Case

As before, the natural thing to do now is to ask whether the **rcp** model coincides with elementary conditional probabilities for a discrete variable X with essential range $R_X = \{x : P(X = x) > 0\}$.

Theorem 5 *For a discrete random variable X with range $R_X = \{x : P(X = x) > 0\}$ and an event A from Ω the following holds for every $x_0 \in R_X$*

$$P_{x_0}(A) = \frac{P(A, X = x_0)}{P(X = x_0)}.$$

Proof: We first note that if $x \neq x_0$ then the following relations hold:

$$A \cap \{X = x_0\} \subseteq A \cap \{X \neq x\} \subseteq \{X \neq x\},$$

hence, by properties (1) and (2) of the definition,

$$\begin{aligned} 0 \leq P_x(A, X = x_0) &\leq P_x(A, X \neq x) \\ &\leq P_x(X \neq x) \\ &= 0. \end{aligned}$$

Now, combining property (3) and the above, we get

$$\begin{aligned} P(A, X = x_0) &= \sum_x P_x(A, X = x_0) p_X(x) \\ &= P_{x_0}(A, X = x_0) P(X = x_0) \end{aligned}$$

[5] The indicator function of a set B is defined by

$$I_B(x) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{if } x \notin B \end{cases}$$

Hence,

$$P_{x_0}(A, X = x_0) = \frac{P(A, X = x_0)}{P(X = x_0)}.$$

Finally, since by (1) P_{x_0} is a probability, we must also have

$$\begin{aligned} P_{x_0}(A) &= P_{x_0}(A, X = x_0) + P_{x_0}(A, X \neq x_0) \\ &= P_{x_0}(A, X = x_0) + 0 \\ &= P(A, X = x_0)/P(X = x_0). \end{aligned}$$

as claimed. □

Corollary 6 *For any discrete variable X, the family of rcps $\{P_x\}_{x \in R_X}$ exists and is uniquely determined by properties (1), (2) and (3). It coincides with the conditional probabilities defined in the elementary sense.*

4.3 Conditional Distributions and Densities

In this section we will show how conditional distributions and densities of random variables can be derived from the rcps. We will also demonstrate an application of the approach to derivation of the formula for t distribution — the fastest known way to get to the formula.

4.3.1 Single event conditioning

Let us consider probabilities of events *given* that something of non-zero probability has already occurred. We shall call it *single event conditioning.* So let H be an event with $P(H) > 0^6$. If A is another event, then we know that the elementary definition holds, so we may write

$$P_H(A) = P(A|H) = \frac{P(A \cap H)}{P(H)}$$

Having H fixed, we regard P_H as another probability function on $(\Omega, \mathcal{F})$. Hence, if Y is a discrete random variable with respect to P, it will be discrete with respect to P_H. Similarly, if Y is an absolutely continuous random variable (P), then Y is absolutely continuous (P_H). If, however, Y is a mixed random variable with respect to P, it can be anything in P_H.

For example, let us define a mixed random variable Y as uniform between 0 and 1 with probability $\frac{1}{2}$, 2 with probability $\frac{1}{4}$ and 3 with probability $\frac{1}{4}$. Let $H = (\{Y = 2\} \cup \{Y = 3\})$ be fixed. Observe that now Y is discrete with respect to P_H, since the P_H-probability that $Y = 2$ or $Y = 3$ is 1, and the P_H-probability that Y is neither 2 nor 3 is zero. Hence Y may be regarded as a discrete random

[6] It is important to remember that $P(H) > 0$ throughout this section.

variable with essential range $R_Y = \{2,3\}$ and with P_H-probability distribution given by $P_H(Y=2) = \frac{1}{2}$ and $P_H(Y=3) = \frac{1}{2}$.

We may then define a distribution function given H as

$$\begin{aligned} F_Y(y|H) &= P_H(Y \leq y) \\ &= \frac{P(Y \leq y, H)}{P(H)}, \end{aligned}$$

and if this distribution function has a density, we can call it $f_Y(y|H)$ — the conditional density of Y given H.

It is interesting to note that generally there is no relationship between the P-density $f_Y(y)$ and the new conditional density $f_Y(y|H)$. The only exception is the case when $H = \{Y \in C\}$[7]. Here, the relationship is clearly

$$f_Y(y|Y \in C) = \begin{cases} f_Y(y)/P(Y \in C) & \text{if } y \in C \\ 0 & \text{otherwise.} \end{cases}$$

4.3.2 Arbitrary random variable conditioning

We can transfer the ideas of the previous section to the case when $H = (X = x)$ for an arbitrary random variable X, in which case the probability of H may be zero.

So let us consider the family of rcps $\{P_x\}_{x \in R_X}$. In general, the family $\{P_x\}$ is an abstract mathematical object which only satisfies the three defining properties. If, however, we have a particular experiment in which x is found at the Stage 1 and treated as a parameter at the Stage 2, then P_x can be interpreted as a "revised probability P" — with knowledge that something (namely, $X = x$) has already happened. This idea was already demonstrated in the uniform-binomial case of Example 3.4 and will be discussed in more details later.

Let now Y be another random variable. As above, we can *define* a conditional distribution function of Y given x as

$$F_{Y|X}(y|x) = P_x(Y \leq y) = P(Y \leq y | X = x)$$

and then $f_{Y|X}(y|x)$ is defined as the density for $F_{Y|X}(y|x)$, if such density exists. As we shall show, when X is absolutely continuous, the density will exist if and only if X and Y are jointly absolutely continuous. This is precisely the context in which we may introduce the concept of *conditional density function.*

4.3.3 Conditional density function

In this section we shall construct an approach in which the formula (2) for the conditional density will *appear* as the only reasonable choice for definition of conditional density.

[7] This was the case with the example above, since H was defined as $(Y \in \{2,3\})$. There is no density with respect to P_H, however, in this case.

Before we turn to our construction, let us state two Conventions which are vital for understanding the further development.

Convention 1 *In a probability model in which $f(y)$ is thought of as a density for Y, we are specifying the formula*

$$P(Y \in C) = \int_C f(y)dy$$

for calculating $P(A)$ for $A = (Y \in C)$, but for no other A.

Convention 2 *Similarly, in a probability model in which a function $h(x,y)$ ($x \in R_X$, $y \in \mathbb{R}$) is thought of as a conditional density for Y given X, we are specifying the formula*

$$P_x(Y \in C) = \int_C h(x,y)dy$$

for calculating the $P_x(A)$ values for $A = (Y \in C)$ and for no other A.

These two Conventions basically state the properties that we want the density function to have. It's interesting that it is sufficient to understand the meaning of the density function in the light of the two Conventions in order to derive the usual defining formula, as will be shown.

So let $f_{Y|X}(y|x)$ be the function satisfying the formula of Convention 2 (existence of such a function will be proven later in this section) and let X have a density function $f_X(x)$. Then, according to property (3) of the definition of regular conditional probabilities, we must have,

$$\begin{aligned}
P(Y \in C) &= \mathbf{E}_{P_X}[P(Y \in C|X = x)] \\
&= \int_{\mathbb{R}} P_x(Y \in C) f_X(x)dx \\
&= \int_{\mathbb{R}} \left(\int_C f_{Y|X}(y|x)dy \right) f_X(x)dx \\
&= \int_{\mathbb{R}} \int_C f_{Y|X}(y|x) f_X(x)dydx \\
&= \int_C \left(\int_{\mathbb{R}} f_{Y|X}(y|x) f_X(x)dx \right) dy, \quad \text{(Fubini Theorem)}
\end{aligned}$$

which by Convention 1 implies that

$$f_Y(y) = \int_{\mathbb{R}} f_{Y|X}(y|x) f_X(x)dx. \tag{5}$$

Since earlier in the chapter we assumed that X and Y are jointly absolutely continuous, their joint density function $f_{X,Y}(x,y)$ exists and so we also have

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)dx. \tag{6}$$

Since both (5) and (6) are true for *all* y, we may *think* that the integrands are equal as well. In other words, we conjecture the formula

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \tag{7}$$

which is precisely equation (2) which defined the conditional density function in the Measure-Theoretic part earlier.

We shall now use rcps to *verify* this relationship, by showing that

$$f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x)$$

is indeed a joint density function for X and Y. So we must check that

$$P(X \leq a, Y \leq b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f_{Y|X}(y|x)f_X(x)dydx$$

Let A be the event $(X \leq a, Y \leq b)$. We then note, that

$$\begin{aligned} P_x(X \leq a, Y \leq b) &= P_x(X = x, X \leq a, Y \leq b) \\ &= \begin{cases} P_x(X = x, Y \leq b) & \text{if } x \leq a \\ 0 & \text{if } x > a \end{cases} \\ &= \begin{cases} P_x(Y \leq b) & \text{if } x \leq a \\ 0 & \text{if } x > a \end{cases} \end{aligned}$$

We can then use the properties of rcps $\{P_x\}$ to calculate $P(A)$, as follows:

$$\begin{aligned} P(A) &= \int_{\mathbb{R}} P_x(A)f_X(x)dx \\ &= \int_{\mathbb{R}} P_x(X \leq a, Y \leq b)f_X(x)dx \\ &= \int_{-\infty}^{a} P_x(Y \leq b)f_X(x)dx \\ &= \int_{-\infty}^{a}\left(\int_{-\infty}^{b} f_{Y|X}(y|x)dy\right)f_X(x)dx \\ &= \int_{-\infty}^{a}\int_{-\infty}^{b} f_{Y|X}(y|x)f_X(x)dydx \end{aligned}$$

which shows that the integrand is in fact the joint density of X and Y, as required.

So we have shown that *if* an $f_{Y|X}(y|x)$ function exists and satisfies the condition of Convention 2, *then* $f_{Y|X}(y|x)f_X(x)$ is a joint density function of X and Y. To prove the converse, i.e. that *if* $f_{X,Y}(x,y)$ is a joint density *then* $f_{X,Y}(x,y)/f_X(x)$ is a conditional density satisfying Convention 2, we need to verify that

$$P_x(Y \in C) = \int_C \frac{f_{X,Y}(x,y)}{f_X(x)}dy,$$

at least a.e. $[P_X]$. The LHS and the RHS of the above equation are functions of x, say $f(x)$ and $g(x)$. To establish the equality $f(x) = g(x)$ a.e. $[P_X]$ it suffices to show that for all $B \in \mathcal{B}(\mathbb{R})$,

$$\int_B f(x)dP_X(x) = \int_B g(x)dP_X(x).$$

So, integrating the RHS for an arbitrary $B \in \mathcal{B}(\mathbb{R})$ we get

$$\begin{aligned}\int_B \left(\int_C \frac{f_{X,Y}(x,y)}{f_X(x)}dy\right) f_X(x)dx &= \int_B\int_C f_{X,Y}(x,y)dydx \\ &= P((X,Y) \in (B \times C)) \\ &= P(X \in B, Y \in C).\end{aligned}$$

On the other hand, by Lemma 4,

$$\int_B P_x(Y \in C)dP_X(x) = P(X \in B, Y \in C),$$

whence the LHS is the same as the RHS, as claimed. □

Thus, we proved,

Theorem 7 *In the light of Convention 2, when X is absolutely continuous the conditional density function $f_{Y|X}$ exists if and only if the joint density function $f_{X,Y}$ exists. In both cases, they can be connected by the equation*

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

when $f_X(x)$ is not zero. The conditional density can be defined to be equal to zero when $f_X(x) = 0$.

To summarize, we defined the conditional density of Y given X by the means of Convention 2, which lead us to the formula introduced earlier as a defining property of conditional density. In this case we *derived* the formula (2) from our "intuitive" Convention and showed the equivalence of the existence of joint density and conditional density.

4.3.4 Limit approach

In the previous sections we tried to relate abstract P_x-probabilities to ordinary conditional probabilities. This was partially done using the two Conventions.

In *nice* cases, which will be specified later, it is possible to provide even more intuitive connection between conditional densities and ordinary conditional probabilities via the concept of a limit. An attempt to do that will be made in this section.

We start with the simplest relation,

$$P_x(A) = P(A|X = x) = \lim_{\substack{a\to x^- \\ b\to x^+}} P(A|a \le X \le b) \tag{8}$$

which simply says that $P(A|X = x)$ is essentially the probability of A given that X is about x, i.e..

$$P(A|X = x) = P(A|X \approx x).$$

Although this approach is very simple, there do not seem to be any general results which either (i) state conditions on A and X which guarantee (8), or (ii) ensure that there exists an rcp for which (8) is true. So whenever this approach is used, great care should be taken to prevent possible errors.

If the family $\{P_x\}$ is partially determined by a given conditional density $f_{Y|X}$ via the equation

$$P_x(Y \in C) = \int_C f_{Y|X}(y|x)dy,$$

we may try to relate P_x's and conditional probabilities using the theorem:

Theorem 8 *If f is integrable on some interval $[a, b]$ containing x_0 and continuous at x_0, then*

$$f(x_0) = \lim_{\substack{a\to x_0^- \\ b\to x_0^+}} \frac{1}{b-a}\int_a^b f(t)dt$$

This is a basic result of Real Analysis, which we are not going to prove here. Instead, we can apply the theorem to our density functions, which, in cases that we shall call *nice cases*, will have properties stated in the Theorem 8.

So let f_X be a density function of a random variable X satisfying the conditions of the theorem, i.e. continuous at x_0. We then have

$$f_X(x_0) = \lim_{\substack{a\to x_0^- \\ b\to x_0^+}} \frac{P(a \le X \le b)}{b-a}$$

since $P(a \le X \le b) = \int_a^b f_X(x)dx$. This is the required relationship between f_X and genuine probabilities.

Further, if Y is another random variable, $C \in \mathcal{B}(\mathbb{R})$, then

$$P(a \le X \le b, Y \in C) = \int_a^b \int_C f_{X,Y}(x,y)dydx$$

for jointly absolutely continuous X and Y. Hence,

$$\begin{aligned} P(Y \in C|a \le X \le b) &= \frac{P(a \le X \le b, Y \in C)}{P(a \le X \le b)} \\ &= \frac{\int_a^b \int_C f_{X,Y}(x,y)dydx}{\int_a^b f_X(x)dx} \end{aligned}$$

Now, if $f_X(x)$ and $g(x) = \int_C f_{X,Y}(x,y)dy$ are both continuous at x_0 as functions of x only and $f_X(x_0) \neq 0$, then, dividing numerator and denominator by $(b-a)$ and applying Theorem 8 we get

$$\begin{aligned}
P_{x_0}(Y \in C) &= \int_C f_{Y|X}(y|x_0)dy \\
&= \int_C \frac{f_{X,Y}(x_0,y)}{f_X(x_0)}dy \\
&= \frac{\int_C f_{X,Y}(x_0,y)dy}{f_X(x_0)} \\
&= \frac{g(x_0)}{f_X(x_0)} \\
&= \lim_{\substack{a\to x_0^- \\ b\to x_0^+}} \left(\frac{\int_a^b \int_C f_{X,Y}(x,y)dydx}{\int_a^b f_X(x)dx} \right) \qquad \text{(by Theorem 8)} \\
&= \lim_{\substack{a\to x_0^- \\ b\to x_0^+}} P(Y \in C | a \leq X \leq b).
\end{aligned}$$

Again, this is the required connection between the "mysterious" $P_{x_0}(Y \in C)$ and probabilities defined in the traditional sense.

4.3.5 Problems with the limit approach

We can now ask ourselves why don't we use the formula

$$P(A|X=x) = \lim_{\substack{a\to x^- \\ b\to x^+}} P(A|a \leq X \leq b)$$

to *define* the conditional probability given X, if X is continuous?

The answer is that the above formula is applicable as long as X is a very "nice" random variable, with a very "nice" density function. If X or it's density have some nasty properties, then the formula would not be true. In fact, even Theorem 8, which we used to derive the result, will not be true anymore.

As an example, suppose that X is a random variable with the range $R_X = [0,1]$ and density f_X defined by the equation

$$f_X(x) = \begin{cases} 1 & \text{if } x \text{ is irrational,} \\ 0 & \text{if } x \text{ is rational.} \end{cases}$$

Thus, $f_X(x)$ is just $1 - \delta(x)$ restricted to $[0,1]$, where $\delta(x)$ is the well-known Dirichlet function.

This density function is not even integrable in the Riemann sense. One could argue, though, that since rational numbers form a set of Borel measure 0, the

Lebesgue integral of $f_X(x)$ is finite. However, even Lebesgue integration does not help here, because if x is a rational number between 0 and 1, then $f_X(x) = 0$, but

$$\frac{1}{b-a}\int_a^b f_X(t)dt = 1$$

for all $0 \leq a \leq b \leq 1$, so the limit is also 1! This clearly contradicts the statement of Theorem 8, which is not surprising, since the conditions of the theorem are not satisfied.

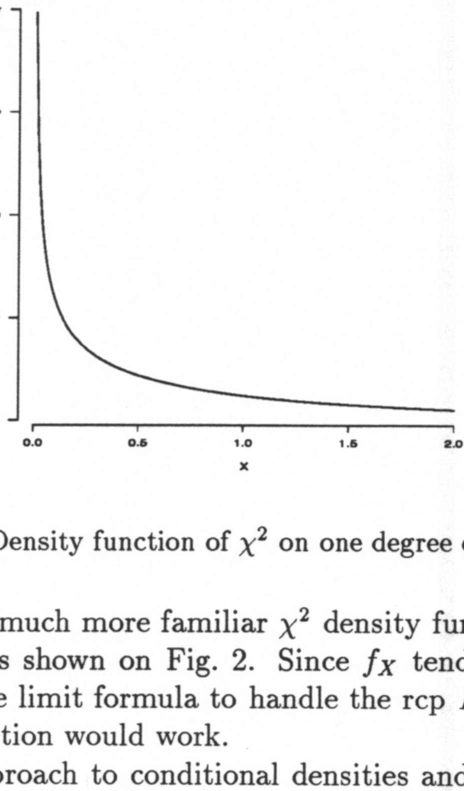

Figure 2: Density function of χ^2 on one degree of freedom.

Another example is much more familiar χ^2 density function on one degree of freedom. The density is shown on Fig. 2. Since f_X tends to infinity at 0, it is not clear how to use the limit formula to handle the rcp P_0. At all other points, however, the limit definition would work.

Thus, the Limit Approach to conditional densities and probabilities works in many cases when both densities and random variables are continuous and "well-behaved", so to speak. In these cases the limit formulae show clearly the relationship between abstract rcps and real-world probabilities. The approach, however, is limited to these cases and seems not general enough to provide a definition in the more difficult cases. It is not even clear whether

$$P_x(A) = \lim_{\substack{a \to x^- \\ b \to x^+}} P(A|a \leq X \leq b)$$

satisfies the axioms of regular conditional probabilities, even if the limit did exist.

4.4 Application to Derivation of the t-density

In this section we shall demonstrate how conditional probabilities can serve as a powerful computational tool. We shall derive the density function of Student's t random variable defined by

$$T = \frac{X}{\sqrt{Y/k}}$$

where X is a standard normal random variable and Y is a χ^2 random variable on k degrees of freedom, independent of X.

4.4.1 Standard method

Traditionally, the density function of t is derived using the transformation technique, as shown below.

The joint density function of X and Y is just a product of marginal densities, since X and Y are independent. So,

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}\Gamma(k/2)2^{k/2}} y^{k/2-1} e^{(-\frac{1}{2}y)} e^{(-\frac{1}{2}x^2)}, \quad y > 0.$$

Now we define a transformation

$$\begin{cases} T = X/\sqrt{Y/k} \\ V = Y \end{cases} \iff \begin{cases} X = T \times \sqrt{V/k} \\ Y = V \end{cases},$$

the Jacobian of which is

$$\begin{aligned} \mathrm{J} &= \det \begin{bmatrix} \partial x/\partial t & \partial x/\partial v \\ \partial y/\partial t & \partial y/\partial v \end{bmatrix} \\ &= \det \begin{bmatrix} \sqrt{\frac{v}{k}} & \frac{t}{2\sqrt{vk}} \\ 0 & 1 \end{bmatrix} \\ &= \sqrt{\frac{v}{k}} \end{aligned}$$

So, the joint density of T and V is

$$f_{T,V}(t,v) = \sqrt{\frac{v}{k}} \frac{1}{\sqrt{2\pi}\Gamma(k/2)2^{k/2}} v^{k/2-1} e^{(-\frac{1}{2}v)} e^{(-\frac{1}{2k}t^2 v)}, \quad v > 0$$

and hence,

$$\begin{aligned} f_T(t) &= \int_{-\infty}^{\infty} f_{T,V}(t,v) dv \\ &= \frac{1}{\sqrt{2k\pi}\Gamma(k/2)2^{k/2}} \int_0^{\infty} v^{\frac{k}{2}-1+\frac{1}{2}} e^{(-\frac{1}{2}(1+t^2/k)v)} dv. \end{aligned} \tag{9}$$

This integral can be evaluated by recognizing the Gamma distribution under the integral sign. After multiplying and dividing by required coefficients, we get the final formula for the t-density function:

$$f_T(t) = \frac{\Gamma[(k+1)/2]}{\Gamma[k/2]} \frac{1}{\sqrt{k\pi}} \frac{1}{(1+t^2/k)^{(k+1)/2}}.$$

□

4.4.2 Using conditional densities

Using conditional densities we can get to the integral (9) in about one line, as follows.

As above, let X be a standard normal random variable and Y is a χ^2 random variable on k degrees of freedom, independent of X. Then we can define, for a fixed $Y = y$,

$$T_y = \frac{X}{\sqrt{y/k}} \quad \sim \quad N\left(0, \frac{1}{\sqrt{y/k}}\right).$$

By the theorem, apparently due to Kupka, the Normal density of T_y is the conditional density $f_{T|Y}$, so we immediately have

$$\begin{aligned} f_T(t) &= \int_{-\infty}^{\infty} f_{T|Y}(t|y) f_Y(y) dy \\ &= \int_0^{\infty} \sqrt{\frac{y}{k}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2k}t^2} \frac{1}{2^{k/2}\Gamma(k/2)} y^{k/2-1} e^{-\frac{y}{2}} dy \end{aligned}$$

which is, after rearrangement, precisely the same as the integral (9). □

Apparently, this is the fastest route to the integral, taking which we get the final formula for the t-density function.

4.5 Example

The rcps are conditional probabilities, which means that except for the discrete case it is not possible to make statistical tests to estimate these probabilities. This idea is illustrated in this section.

Here we return to the Example 3.4. If we wish to estimate in an experiment the conditional probability $P(Y = i|X = x)$ for a fixed x and i, we must run the experiment until the value of X is x — an event of probability zero since X is continuous. So rcps are the probabilities which do exist but are hard to "touch". This is in analogy with electrons — we all know they exist but nobody has ever seen or touched one. It is possible, however, to conduct an experiment which involves theoretical *assumed* properties of electrons and check observable consequences of the model. If they agree with theoretical results derived using assumptions about electrons (which cannot be verified directly) we may think that our assumptions are valid.

Similarly with rcps — if we can not check directly our assumptions we can often use our assumptions to derive results that *can* be checked in a real-life experiment, as we shall show later.

4.5.1 Unconditional distribution of Y

Now, we shall use rcps to answer another previously discussed question: *what is the unconditional distribution of* Y? Let $A = \{Y = i\}$ be an event for a fixed value $i = 0, \ldots, n$. To find $P(A)$ we start with our *assumption* about $P_x(A)$:

$$P_x(Y = i) = P(Y = i|X = x) = \binom{n}{i} x^i (1 - x)^{n-i}$$

By property (3) of the definition we then have

$$\begin{aligned} P(Y = i) &= \mathbf{E}_{P_X}[P_x(Y = i)] \\ &= \int P_x(Y = i) f_X(x) dx \\ &= \int_0^1 \binom{n}{i} x^i (1 - x)^{n-i} \times 1 \, dx \\ &= \frac{1}{n + 1} \end{aligned}$$

as was shown before — by recognizing a Beta distribution under the integral sign. Hence, Y is uniform on 0,1,...,n.

4.5.2 Simulation

Since the result was derived using the rcps, it should be tested in a real-life experiment. Fortunately, this experiment is easy to conduct: a simple program in C was written to simulate 10,000 uniform variables over the interval [0,1] and for each simulate one binomial variable with $n = 49$. The program is shown below:

```
/*
 * Program to simulate MAX_X pairs of random variables:
 *      X - uniform(0,1)
 *      Y - binomial(49,x)
 * where x is a sample value of X.
 *
 * (C) 1995 Konstantin Gredeskoul, Monash University.
 */

#include <stdio.h>
#include <stdlib.h>
```

```
#define MAX_X 10000
#define MAX_Y 49

int binomial(float p, int n);

void main()
{
        FILE *output;
        int i, y;
        float x;

        output = fopen("simulate.out", "w");     /* Open file for writing    */

        for (i=1; i<=MAX_X; ++i) {
            x = (float)(rand()+1)/RAND_MAX;      /* Get uniform x            */
            y = binomial(x, MAX_Y);              /* Get binomial Y given p=x */
            fprintf(output,"%10.8f\t%d\n",x,y);  /* Write results to a file  */
        }
}

int binomial(float p, int n)                     /* The function returns bin-
                                                    omial r.v. with parameters
                                                    n and p.                 */
{
        int i, y=0;
        float z;

        if (p<0 || p>1)
           return(-1);
        else
           for (i=1;i<=n;++i) {                  /* Make 50 runs of:         */
               z = (float)(rand()+1)/RAND_MAX;   /* Get z - uniform (0,1)    */
               if (z<=p) ++y;                    /* If z < p, increment y    */
           }
        return(y);                               /* Return y                 */
}
```

Then the S-Plus package was used to analyze and graph the data; the simulated Y-values are shown on Fig. 3 — they form an acceptable uniform distribution.

4.5.3 χ^2 test of goodness-of-fit

To make sure that they form an acceptable uniform distribution, another program on C was written to test 10,000 simulated y-values for goodness-of-fit using the χ^2 test. It was not possible to load and test 10,000 y-values into Minitab due to lack of memory. The program, however, worked quickly by reading each y-value, incrementing the counter and disposing any y-value, no longer needed.

The program is shown below:

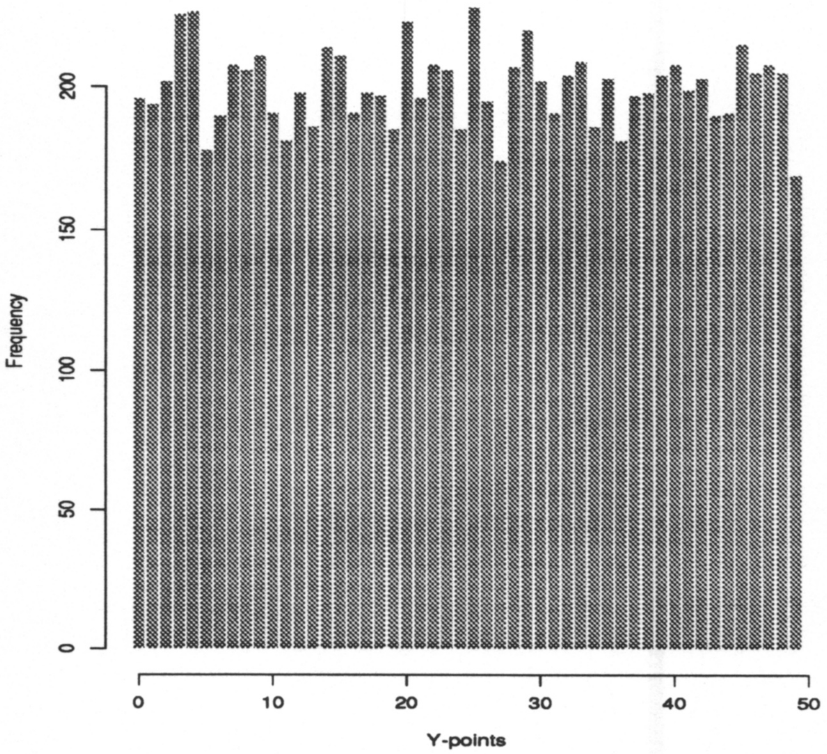

Figure 3: Histogram of 10,000 empirical Y values.

```
/*
 * Program to read n values from a file and conduct a
 * goodness-of-fit Chi-square test on the data.
 *
 * (C) Copyright 1995, Konstantin Gredeskoul
 */

#include<stdio.h>
#include<math.h>

#define N 50

void main(int argc, char *argv[])
{
        int     i, c=0, count[N]={0};
        float   chisq=0, expected;
        FILE    *inp;

        if (argc == 2) {
           printf("Opening file %s for reading...\n", argv[1]);
           inp = fopen(argv[1], "r");
           if (inp!=NULL)
           printf("Reading data...\n");
           while (fscanf(inp, "%d\n", &i)==1){
                 ++count[i];
                 ++c;
           }
           fclose(inp);

           printf("Closing file...\nCalculating Chi-square statistic...\n");

           expected=(float)c/N;
           for (i=0; i<=N-1; ++i)
               chisq+=pow(count[i]-expected,2)/expected;

           printf("Chi-square = %5.3f, on %d d.f.\nStop.", chisq, N-1);
        }
}
```

And here is the output of the above program:

```
Opening file chisq.txt for reading...
Reading data...
Closing file...
Calculating Chi-square statistic...
Chi-square = 43.400, on 49 d.f.
Stop.
```

The probability corresponding to the value 43.400 was found using MINITAB. The extract of the (edited) MINITAB session is shown below:

```
MTB > CDF C1;
SUBC>   Chisquare 49.

         x     P( X <= x)
   43.4000        0.3013

MTB > let k1 = 0.3013
MTB > let k2 = 1 - k1
MTB > print k2

K2        0.698700
```

Since $P(\chi^2_{49} > 43.400) = 0.6987$ is certainly not significant, we accept the null hypothesis that the distribution of Y is uniform over $\{0, \ldots, 49\}$.

4.5.4 Distribution of Y given $X \in [a, b]$

Although it is impossible to simulate the Y values given a particular value of X, it is quite possible to obtain an empirical Y-distribution given that X lies in some interval $[a, b]$ since the probability that $X \in [a, b]$ is not zero for all $0 \leq a < b \leq 1$. We shall consider, for example, the interval: $[0.8, 0.9]$.

Theoretically, assuming the rcp model we have the following probabilities,

$$\begin{aligned}
P(Y = i | a \leq X \leq b) &= \frac{P(Y = i, a \leq X \leq b)}{P(a \leq X \leq b)} \\
&= \frac{P(Y = i, a \leq X \leq b)}{b - a} \\
&= \frac{1}{b-a} \int_0^1 P_x(Y = i, a \leq X \leq b) f_X(x) dx \\
&= \frac{1}{b-a} \int_a^b P_x(Y = i) f_X(x) dx \quad \text{by property (2)} \\
&= \frac{1}{b-a} \int_a^b \binom{n}{i} x^i (1-x)^{n-i} dx
\end{aligned}$$

For each value of $i = 0, 1, \ldots, n$ these integrals have to be evaluated numerically. This was accomplished using the algebraic package **Maple-V** and integrating inside of a loop, as shown below:

```
> with(combinat, numbcomb);
> for i from 30 to 49 do
      evalf(10*int(numbcomb(49,i)*x^i*(1-x)^(49-i),x=0.8..0.9),20)
  od;
>
```

The results of Maple's calculations are given in the Table 1 and graphed on Fig. 4. The theoretical probabilities are shown as connected circles, while simulated values are shown as bars. Since simulated data fit theoretical values, we conclude that the assumptions of the rcp model were correct.

4.5.5 What if X is discrete?

Recall, that we showed that the unconditional distribution of Y is discrete uniform when X is a continuous uniform random variable.

What happens if X is uniform but *discrete*? Would the distribution of Y still be uniform? To answer this question, we can use rcps to calculate the exact probabilities for each value of Y and check whether they are all the same. So let X be a discrete uniform random variable, with essential range $R_X = \{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\}$ for some integer m. What is the unconditional probability that $Y = i$ for some fixed $i = 0, \ldots, n$?

Y	Probability	Y	Probability
30	.0005620518	40	.1063443302
31	.0003956246	41	.1269602333
32	.0012600785	42	.1374890209
33	.0028890942	43	.1333657228
34	.0061563624	44	.1136191576
35	.0121294347	45	.0825404783
36	.0220604103	46	.0489275089
37	.0370104687	47	.0220886682
38	.0572224960	48	.0067186362
39	.0814171961	49	.0010279005

Table 1: Y-values and exact theoretical probabilities.

Since, as earlier, we have

$$P_x(Y=i) = \binom{n}{i} x^i (1-x)^{n-i}$$

we can write

$$\begin{aligned} P(Y=i) &= \mathbf{E}_{P_X}[P_x(Y=i)] \\ &= \sum_{j=0}^{m} P(Y=i|X=j/m) \times p_X\left(\frac{j}{m}\right) \\ &= \sum_{j=0}^{m} \binom{n}{i} \left[\frac{j}{m}\right]^i \left[1-\frac{j}{m}\right]^{n-i} \times \frac{1}{m+1} \\ &= \frac{1}{m+1} \sum_{j=0}^{m} \binom{n}{i} \left[\frac{j}{m}\right]^i \left[1-\frac{j}{m}\right]^{n-i} \end{aligned}$$

For each i the above sum has to be evaluated numerically. A program in C was written to calculate the above probabilities for $m = 10$ and $n = 20$. The program and the output are shown below:

```
/*
 * Calculation of probabilities P(Y=i) when Y is a (conditional) binomial
 * random variable with probability of success p=x, where
 * X is another discrete random variable with range {j/m} for j=0,...,m.
 *
 * (C) Copyright 1995, Konstantin Gredeskoul
 */
```

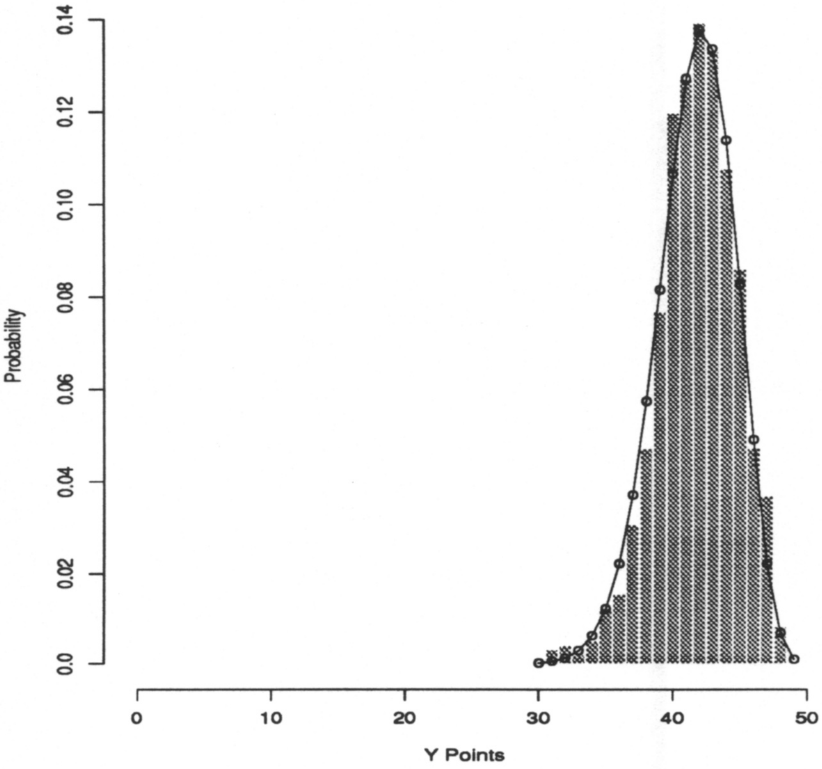

Figure 4: Sample Y-histogram and theoretical values for $x \in [0.8, 0.9]$.

```
#include <stdio.h>
#include <stdlib.h>
#include <math.h>
#define N 20

double factorial(int);
double choose(int, int);

void main()
{
        int i, y;
        double x;
        float sum, total=0;

        x=factorial(20);

        for (y=0; y<=N; ++y){
                sum=0;

                for (x=0.0; x<=1.0; x+=0.1)
```

```
                        sum += choose(N,y) * pow(x,y) * pow(1-x,N-y);

                sum/=11;
                printf("Y = %2d     Probability = %f\n", (int)y, sum);
                total+=sum;
        }
        printf("-----------------------------------\n");
        printf("           Total prob. = %f\n", total);
}

double factorial(int n)                             /* Factorial function  */
{
        int i;
        double tmp=1;

        for (i=1;i<=n;++i) tmp*=i;
        return(tmp);
}

double choose(int n, int m)                         /* n choose m function */
{
        double tmp;

        if (m==0 || m==n || n==1)
                return(1);
        if (m==1 || m==n-1)
                return(n);
        if (m>n)
                return(-1);

        tmp=factorial(n)/(factorial(n-m)*factorial(m));
        return(tmp);
}
```

And the output:

```
Y =  0     Probability = 0.103086
Y =  1     Probability = 0.030469
Y =  2     Probability = 0.041201
Y =  3     Probability = 0.043688
Y =  4     Probability = 0.043480
Y =  5     Probability = 0.043283
Y =  6     Probability = 0.043278
Y =  7     Probability = 0.043290
Y =  8     Probability = 0.043291
Y =  9     Probability = 0.043290
Y = 10     Probability = 0.043290
Y = 11     Probability = 0.043290
Y = 12     Probability = 0.043291
Y = 13     Probability = 0.043290
Y = 14     Probability = 0.043278
```

```
Y = 15     Probability = 0.043283
Y = 16     Probability = 0.043480
Y = 17     Probability = 0.043688
Y = 18     Probability = 0.041201
Y = 19     Probability = 0.030469
Y = 20     Probability = 0.103086
-----------------------------------
           Total prob. = 1.000000
```

It can be seen that probabilities are unequal. Hence we can conclude that the unconditional distribution of Y is not uniform in this case.

In fact, the probabilities form a rather interesting distribution shown in Fig. 5. The two peaks at 0 and 20 can be explained by the fact that if $X = 0$ with probability $\frac{1}{11} = 0.09$ then $Y = 0$ with probability 1. Similarly, if $X = 1$ with probability 0.09, then $Y = 20$ with probability 1. For all other values of X, the values of Y may vary, hence the pattern on the histogram.

If X was distributed over $\{\frac{1}{m}, \frac{2}{m}, \ldots, \frac{m-1}{m}\}$, then the unconditional distribution of Y would not have the two peaks at 0 and 20, but it would still not quite be uniform, as shown on the Fig. 6.

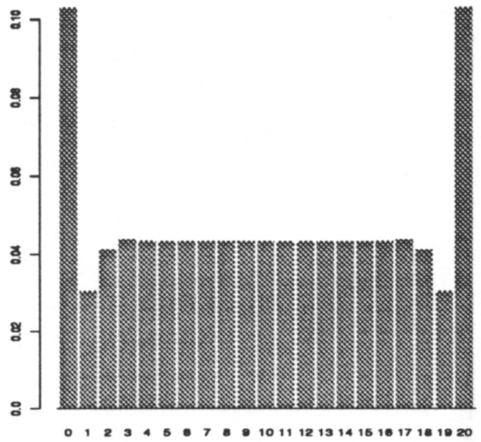

Figure 5: Exact probability distribution of Y when X is discrete.

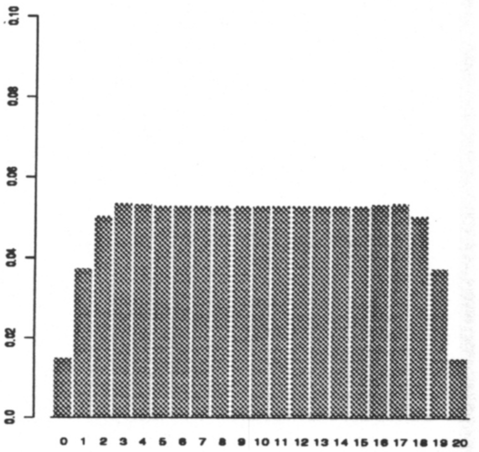

Figure 6: Exact probability distribution of Y when X is discrete, but not 0 or 1.

5 Bibliography

1. **Ash, Robert B.** "Real Analysis and Probability", *1972, Academic Press, NY.*

2. **Ash, Robert B.** "Basic Probability Theory", *1970, John Wiley & Sons, Inc.*

3. **Chang, J. T. and Pollard, D.** "Conditioning as disintegration", *1993, Submitted to Statistical Science.*

4. **Cohn, Donald L.** "Measure Theory", *1980, Birkha-user, Boston.*

5. **Hogg, R. V. and Craig, A. T.** "Introduction to Mathematical Statistics", 3rd edition, *1970, The Macmillan Company.*

6. **Kolmogorov, A. N.** "Foundations of Probability", 2nd English edition, *1933, Chelsea, New-York.*

7. **Kupka, J. G.** "Regular Conditional Probabilities".

8. **Lehmann, E. L.** "Testing Statistical Hypotheses", 2nd edition, *1986, John Wiley & Sons.*

9. **Lehmann, E. L.** "Theory of Point Estimation", *1983, John Wiley & Sons.*

10. **Mendenhall, W., Wackerly, D. D. and Scheaffer, R. L** "Mathematical Statistics with Applications", 4th edition, *1990, PWS-KENT.*

11. **Mood, A. M., Graybill, F. A. and Boes, D. C** "Introduction to the Theory of Statistics", 3d edition, *1974, McGraw-Hill Series in Probability and Statistics.*