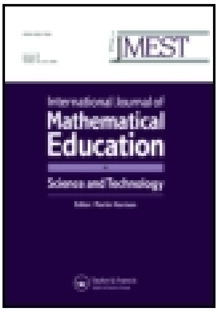

International Journal of Mathematical Education in Science and Technology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tmes20

Students' misconceptions about random variables

Farida Kachapova [a] & Ilias Kachapov [b]

[a] School of Computing and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand

[b] Examination Academic Services, University of Auckland, Auckland, New Zealand

Available online: 16 Dec 2011

International Journal of Mathematical Education in Science and Technology, 2011, 1–9, iFirst

CLASSROOM NOTE

Students' misconceptions about random variables

Farida Kachapova[a*] and Ilias Kachapov[b]

[a]*School of Computing and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand;* [b]*Examination Academic Services, University of Auckland, Auckland, New Zealand*

(*Received 23 June 2011*)

This article describes some misconceptions about random variables and related counter-examples, and makes suggestions about teaching initial topics on random variables in general form instead of doing it separately for discrete and continuous cases. The focus is on post-calculus probability courses.

Keywords: misconception; counter-example; mixed random variable; singular random variable; Riemann–Stieltjes integral; expected value

1. Introduction

Much research has been done on students' thinking about probability and related misconceptions (see, e.g. [1–6]). Many papers are devoted to the concept of probability but not to the concept of random variable (see, e.g. [7–9]). Random variable is '… a fundamental stochastic idea' and '… is the support of many probability and statistics subjects' [7]. So, it is important that the students in probability courses gain a deep understanding of this concept. In this article, we focus on post-calculus probability courses, which are part of university courses in quantitative areas, such as mathematical studies, physics and engineering. Students in such courses have differentiation and integration skills, so they can master probability concepts at a more advanced level than the introductory level of basic statistics courses.

In Section 2, we look at some common students' misconceptions about events and random variables. In Section 3, we make suggestions about teaching random variables. In most probability courses, random variables are taught in two separate chapters containing the material on discrete variables and continuous ones, respectively. So many topics are taught twice. Besides, some theorems that are true for arbitrary random variables are stated only for the discrete and continuous ones. So, instead of losing generality we suggest to use the Riemann–Stieltjes integral to introduce the expectation of a random variable in general form and to develop the theory for arbitrary random variables as long as possible before teaching particular properties of the discrete and continuous variables. Our suggestions do not target

*Corresponding author. Email: farida.kachapova@aut.ac.nz

ISSN 0020–739X print/ISSN 1464–5211 online

http://dx.doi.org/10.1080/0020739X.2011.644332
http://www.tandfonline.com

introductory service courses in statistics, where students have a limited mathematical background and common teaching methods with minimum mathematics are appropriate.

2. Common misconceptions

2.1. *Event*

One of the common misconceptions in probability is the belief that an event is an arbitrary subset of a sample space, which is usually correct in the case of a finite sample space but not in the general case; sometimes Venn diagrams can strengthen this misconception. The misconception can be avoided if we emphasize that a subset A is an event if its probability $\mathrm{P}(A)$ exists; no theory can be developed for sets with undefined probabilities.

2.2. *Random variable*

Random variable is a fundamental concept in probability theory that is not intuitive, so some students develop misconceptions about it. One common misconception is that a random variable X is any real-valued function on the sample space. Students can overcome this misconception by learning the most important characteristic of the random variable – its distribution function F_X: $F_X(x) = \mathrm{P}(X \leq x)$. For the distribution function to exist, the probability $\mathrm{P}(X \leq x)$ should be defined for any real number x; therefore X is a valid random variable only if for any $x \in \mathbf{R}$ the probability $\mathrm{P}(X \leq x)$ is defined, that is $\{X \leq x\}$ is an event.

2.3. *Continuous random variable*

Some students' misconceptions in probability relate to continuous random variables, which make a harder topic than discrete random variables. One of the misconceptions is defining a continuous random variable as a variable with a non-countable set of values (sometimes this set is expected to be an interval or a combination of intervals). This leads to the misconception that any random variable is either discrete or continuous. To help students to overcome these misconceptions, we can emphasize the standard definition: a random variable X is *continuous* if its distribution function F can be expressed as $F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$ for some integrable non-negative function f, which is called the *density function* of X. A simple example of a mixed random variable can clarify the matter further.

Example 1: *Mixed random variable*. Consider the following function (Figure 1):

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ 0.2x & \text{if } 0 \leq x < 1, \\ 0.2x + 0.6 & \text{if } 1 \leq x < 2, \\ 1 & \text{if } x \geq 2. \end{cases}$$

Clearly F is non-decreasing, right-continuous and has limits 0 and 1 at $-\infty$ and ∞, respectively. So F is the distribution function of some random variable X.

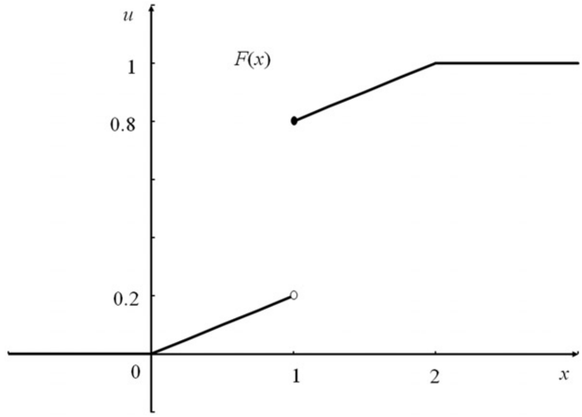

Figure 1. The distribution function of the mixed random variable.

Here, $P(X=1)=0.6$. If we assume that X has a density function $f(x)$, then $f(x)=F'(x)$ for any real x except points 0, 1, and 2; so

$$0.6 = P(X=1) \leq F(1) = \int_{-\infty}^{1} f(x)\mathrm{d}x = \int_{0}^{1} 0.2\mathrm{d}x = 0.2,$$

which is a contradiction. Therefore X is not continuous though the set of its values is the interval $(0, 2)$ and it is not countable.

This random variable is an example of a mixed variable: it is discrete at point 1 (since the probability of 1 is 0.6) and it is continuous on intervals $(0, 1)$ and $(1, 2)$.

Another common students' misconception is thinking of a continuous random variable as the variable with a continuous distribution function; the condition of continuity of the distribution function is necessary but not sufficient. Any singular random variable can be used as a counter-example here, since the distribution function of such a variable is continuous and its derivative equals 0 almost surely; therefore the singular random variable does not have a density function and is not continuous. The following is a well-known example of a singular variable (see, e.g. [10]); the students need only basic knowledge of calculus to understand the example.

Example 2: *Singular random variable*. We will define a distribution function F as the Cantor function. For any $x \in [0, 1]$, denote by $(x_1, x_2, x_3, \ldots)$ the expression of x in base 3, that is $x = \frac{x_1}{3} + \frac{x_2}{3^2} + \frac{x_3}{3^3} + \cdots$. Denote

$$k = \min\{r : x_r = 1\}$$

if there is no r with $x_r = 1$, then $k = \infty$. Define

$$F(x) = \sum\nolimits_{r=1}^{k} \frac{\mathrm{sign}(x_r)}{2^r}, \quad \text{where} \quad \mathrm{sign}\,(a) = \begin{cases} 1 & \text{if } a > 0, \\ 0 & \text{if } a = 0. \end{cases}$$

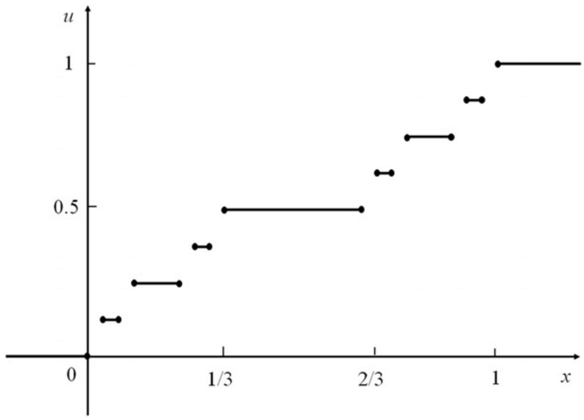

Figure 2. The Cantor function as the distribution function of the singular random variable.

This is the analytical definition of the Cantor function (Figure 2). Geometrically, it can be described as follows. Divide [0, 1] into three equal parts and define F on the middle part by $F(x)=\frac{1}{2}$ for $x\in\left[\frac{1}{3},\frac{2}{3}\right]$. Next, divide each of the remaining two intervals into three equal parts and define F in the middle parts by:

$$F(x)=\frac{1}{2^2}\text{ for }x\in\left[\frac{1}{9},\frac{2}{9}\right]$$

and

$$F(x)=\frac{3}{2^2}\text{ for }x\in\left[\frac{7}{9},\frac{8}{9}\right].$$

This process is repeated infinite number of times. For other definitions of the Cantor function, see [11]. Define $F(x)=0$ for $x<0$, and $F(x)=1$ for $x>1$.

Clearly $F(0)<F(1)$, the function F is continuous, non-decreasing and its derivative equals zero almost surely. Hence F is a singular function and is the distribution function of some singular random variable Y.

The following theorem also helps the students to understand that discrete and continuous are not the only types of random variables (for details, see [12]).

Lebesgue decomposition: any random variable X can be uniquely represented in the form: $X=\alpha_1X_d+\alpha_2X_c+\alpha_3X_s$, where numbers $\alpha_i\geq 0$ $(i=1,2,3)$, $\alpha_1+\alpha_2+\alpha_3=1$, and X_d, X_c, and X_s are discrete, continuous, and singular random variables, respectively.

3. Teaching suggestions

Many textbooks define a random variable and its distribution function in general form, then give most definitions and statements about random variables twice: first in discrete case, then in continuous case. This is good for a basic probability course. For a university course, we suggest a more holistic teaching approach with more emphasis on the general theory for all random variables and with less attention to particular details of the discrete and continuous cases.

After defining a random variable X, we recommend to introduce its distribution function F_X, survival function S_X by $S_X(x) = \mathrm{P}(X > x)$, and quantile function F_X^{-1} by

$$F_X^{-1}(u) = \inf\{x \in \mathbb{R} : F_X(x) \geq u\}, u \in [0,\ 1],\ \text{with}\ \inf \varnothing = +\infty\ \text{by convention.}$$

The quantile function is used for simulating continuous distributions and is usually introduced later with other simulation methods but it is more logical to introduce this function and its properties at the beginning of the course.

Next, we suggest to introduce the Riemann–Stieltjes integral with respect to a distribution function F. In a post-calculus course, the students can master this concept without difficulty, because there is a natural analogy with the Riemann integral, where increments of x are replaced with increments of F. For a random variable X with the distribution function F, the expectation is defined by the following Riemann–Stieltjes integral:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x \mathrm{d}F(x).$$

Thus, the students see that a random variable does not have to be discrete or continuous to have an expectation. This is illustrated by calculating the expectations for the aforementioned mixed and singular random variables.

Example 3: Find the expectation $\mathrm{E}(X)$ for the mixed random variable X from Example 1.

Solution: For the continuous part of the variable the density equals $f(x) = F'(x) = 0.2$, $x \in (0, 1) \cup (1, 2)$. So the expectation equals

$$\begin{aligned}\mathrm{E}(X) &= 1 \times P(X = 1) + \int_0^1 xf(x)\mathrm{d}x + \int_1^2 xf(x)\mathrm{d}x \\ &= 1 \times (0.8 - 0.2) + \int_0^2 0.2x\mathrm{d}x = 1.\end{aligned}$$

The following is a more rigorous proof that demonstrates how the answer can be derived from the general definition of expectation; this can be used in more advanced probability courses.

Since the random variable X is bounded, the expectation $\mathrm{E}(X)$ exists and $\mathrm{E}(X) = \int_0^2 x\mathrm{d}F(x)$. This integral is a limit of integral sums.

For an odd integer $n = 2m + 1$ ($m > 0$), divide $[0, 2]$ into n equal sub-intervals:

$$0 = x_0 < x_1 < \cdots < x_m < 1 < x_{m+1} < \cdots < x_n = 2.$$

For $i = 0, 1, \ldots, n-1$, denote $\xi_i = x_{i+1}$ and $\Delta F_i = F(x_{i+1}) - F(x_i)$. Then, for $i \neq m$:

$$\Delta F_i = 0.2 \Delta x_i. \tag{1}$$

The integral sum equals

$$S_n = \sum_{i=0}^{n-1} \xi_i \Delta F_i = S_n^{(1)} + S_n^{(2)} + S_n^{(3)},$$

where

$$S_n^{(1)} = \xi_m \Delta F_m, \quad S_n^{(2)} = \sum_{i=0}^{m-1} \xi_i \Delta F_i \text{ and } S_n^{(3)} = \sum_{i=m+1}^{n-1} \xi_i \Delta F_i.$$

Here

$$S_n^{(1)} = x_{m+1}[F(x_{m+1}) - F(x_m)] \underset{n\to\infty}{\to} 1 \cdot \left[\lim_{x\to 1+} F(x) - \lim_{x\to 1-} F(x)\right] = 0.8 - 0.2 = 0.6.$$

By (1),

$$S_n^{(2)} = \sum_{i=0}^{m-1} \xi_i \cdot 0.2\Delta x_i = 0.2 \sum_{i=0}^{m-1} \xi_i \Delta x_i \underset{n\to\infty}{\to} 0.2 \int_0^1 x\mathrm{d}x,$$

the Riemann integral.

Similarly,

$$S_n^{(3)} = 0.2 \sum_{i=m+1}^{n-1} \xi_i \Delta x_i \underset{n\to\infty}{\to} 0.2 \int_1^2 x\mathrm{d}x.$$

So

$$E(X) = \lim_{n\to\infty} S_n = 0.6 + 0.2 \int_0^1 x\mathrm{d}x + 0.2 \int_1^2 x\mathrm{d}x = 0.6 + 0.2 \int_0^2 x\,\mathrm{d}x = 1.$$

Similarly, it can be shown that for the singular random variable Y from Example 2, the expectation $\mathrm{E}(Y) = \frac{1}{2}$. This also follows from the symmetry of the distribution function with respect to point $\left(\frac{1}{2}, \frac{1}{2}\right)$.

The singular random variable seems to be an artificial mathematical construction. Mixed random variables, on the other hand, have natural applications, in actuarial science in particular. 'A mixed type rv frequently encountered in actuarial science is an insurance risk for which there is a probability mass in zero (the probability of non-occurrence of claims), while the claim amount given that a claim occurs is a continuous rv.' [13, p. 17]. We illustrate this with the following two examples (for more examples, see [14]).

Example 4: Consider an insurance payment X against accidental loss with excess of \$200 and maximum payment of \$2000. Assume that the loss is below \$200 with probability 0.08 and above \$2200 with probability 0.12, with the uniform distribution in between. Find the distribution and expectation of X.

Solution: Clearly, $\mathrm{P}(X=0)=0.08$, $\mathrm{P}(X=2000)=0.12$, $\mathrm{P}(0 < X < 2000)=0.8$, and X is uniformly distributed between 0 and 2000. The distribution function of X is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ 0.0004x + 0.08 & \text{if } 0 \le x < 2000, \\ 1 & \text{if } x \ge 2000. \end{cases}$$

For the continuous part of the variable, the density equals $f(x) = F'(x) = 0.0004$, $x \in (0, 2000)$. So the expectation (expected payment) equals

$$\begin{aligned}\mathrm{E}(X) &= 0 \times P(X=0) + 2000 \times P(X=2000) + \int_0^{2000} x f(x)\mathrm{d}x \\ &= 2000 \times 0.12 + \int_0^{2000} 0.0004x\,\mathrm{d}x = \ \$1040.\end{aligned}$$

Example 5: Consider an insurance claim of size X, where X equals 0 with probability 0.2 and the distribution of its positive values is proportional to exponential distribution with parameter 0.005. Find the distribution and expectation of X.

Solution: Clearly $\mathrm{P}(X=0) = 0.2$ and $\mathrm{P}(X > 0) = 0.8$, so the distribution function of X is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 - 0.8e^{-0.005x} & \text{if } x \geq 0. \end{cases}$$

For the continuous part of the variable, the density equals $f(x) = F'(x) = 0.004\,e^{-0.005x}$, $x > 0$. So the expectation (expected claim size) equals

$$\mathrm{E}(X) = 0 \times P(X=0) + \int_0^{\infty} x f(x)\mathrm{d}x = \int_0^{\infty} 0.004x\,e^{-0.005x}\,\mathrm{d}x = \ \$160.$$

The properties of expectation can be stated and proved in general form using the properties of the Riemann–Stieltjes integral. Next, it is logical to introduce other numerical characteristics of random variables in general form through expectation, such as moments, variance, standard deviation, covariance and Pearson correlation. Besides general properties of these characteristics, it is useful to consider Chebyshev's inequality and Hoeffding's equality [15]:

$$\mathrm{Cov}(X, Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[F_{X,Y}(x, y) - F_X(x)F_Y(y)\right]\mathrm{d}x\,\mathrm{d}y,$$

where $F_{X,Y}$ is the joint distribution function of X and Y, and $\mathrm{Cov}(X, Y)$ is their covariance. The Hoeffding equality is not sufficiently known, though it has useful applications, and it is easily stated and proved [16, p. 1139].

Finally, formulas can be derived for the expectation of a discrete random variable and a continuous one. In addition to the standard formulas, it is suggested to include these interesting alternative formulas:

$$\mathrm{E}(X) = \int_0^{\infty} S_X(x)\mathrm{d}x - \int_{-\infty}^{0} F_X(x)\mathrm{d}x \quad \text{for a continuous } X \text{ and}$$

$$\mathrm{E}(X) = \sum_{k=0}^{\infty} P(X > k) - \sum_{k=-\infty}^{0} P(X < k) \quad \text{for a discrete, integer-valued } X.$$

They have quite simple proofs.

4. Discussion

The described approach is not suggested for introductory statistics courses, where students' mathematical background is limited. Our suggestions are intended for post-calculus probability courses, which are part of university courses in quantitative areas. The students in these courses enjoy the mathematical side of probability and are not entirely focused on statistical applications.

A standard one-semester course in calculus provides the students with necessary differentiation and integration skills; these students are familiar with Riemann integral and therefore they can understand the Riemann–Stieltjes integral, which is defined similarly, only increments of x are replaced with increments of y. So the expectation of a random variable can be defined in general form through a Riemann–Stieltjes integral.

The described general approach to teaching random variables covers all types of variables, not only discrete and continuous ones. This is the way random variables are taught in actuarial science where mixed variables have important applications and the expected value has to be introduced through the Riemann–Stieltjes integral.

This approach was used by the authors for several years in post-calculus probability courses at the Auckland University of Technology (New Zealand) and the Moscow Technological University (Russia). Our case studies show that this approach helps:

- to avoid the misconception that discrete and continuous variables are the only possible ones;
- to avoid tedious repetition of definitions and proofs;
- to produce more interesting proofs in the general case.

We noticed that the described examples of unusual random variables stimulated students' interest in the subject and their critical thinking. The students got interested what other types of random variables exist, and Lebesgue decomposition theorem helped to answer such questions.

Many university students have got a basic knowledge of discrete and continuous variables from high school. In university courses we can build on this knowledge, generalize and develop it to a more advanced level. The universal approach to random variables can help students to distinguish between general and type-specific properties of random variables. Together with the aforementioned counter-examples, the described teaching approach can also help students to avoid misconceptions and gain a deeper understanding of the basic probability concepts. More advanced counter-examples in probability can be found in [17]. The general teaching approach provides students with an abstract knowledge of random variables, which is important in a variety of applications. In particular, this knowledge can be applied in actuarial science (as shown in Examples 4 and 5), so there is no need to study mixed random variables separately as a third type of variables.

Acknowledgements

The authors wish to thank the referees and the editor for their valuable comments and suggestions that helped to improve this article.

References

[1] B.A. Warner, D. Pendergraft, and T. Webb, *That was Venn, this is now*, J. Stat. Ed. 6(1) (1998). Available at http://www.amstat.org/publications/jse/v6n1/warner.html
[2] J. Ancker, *The language of conditional probability*, J. Stat. Ed. 14(2) (2006). Available at http://www.amstat.org/publications/jse/v14n2/ancker.html
[3] A. Rossman and T. Short, *Conditional probability and education reform: Are they compatible?* J. Stat. Ed. 3(2) (1995). Available at http://www.amstat.org/publications/jse/v3n2/rossman.html
[4] M.D. Larsen, *Advice for new and student lecturers on probability and statistics*, J. Stat. Ed. 14(1) (2006). Available at http://www.amstat.org/publications/jse/v14n1/larsen.html
[5] T. Leviatan, *Conceptual, computational and didactic aspects of teachers' training in probability and statistics*, 54th Session of ISI, Berlin, 2003, pp. 586–589.
[6] E. San Martin, *Piaget's viewpoint on the teaching of probability: A breaking-off with the traditional notion of chance?* ICOTS-7, Salvador, Bahia, Brazil, 2006.
[7] B.R. Hernandez, J. Albert Huerta, and C. Batanero, *An exploratory study of students' difficulties with random variables*, ICOTS-7, Salvador, Bahia, Brazil, 2006.
[8] T.K. Miller, *The random variable concept in introductory statistics*, ICOTS-5, Singapore, June 1998, pp. 1222–1223.
[9] K. Ballman, *Greater emphasis on variation in an introductory statistics course*, J. Stat. Ed. 5(2) (1997). Available at http://www.amstat.org/publications/jse/v5n2/ballman.html
[10] P. Billingsley, *Probability and Measure*, 3rd ed., Wiley, New York, 1995.
[11] H.L. Royden, *Real Analysis*, 3rd ed., Prentice-Hall, Upper Saddle River, NJ, 1988.
[12] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 2, Wiley, New York, 1971.
[13] M. Denuit, J. Dhaene, M.J. Goovaerts, and R. Kaas, *Actuarial Theory for Dependent Risks*, Wiley, New York, 2005.
[14] R. Kaas, M.J. Goovaerts, J. Dhaene, and M. Denuit, *Modern Actuarial Risk Theory: Using R*, 2nd ed., Wiley, New York, 2009.
[15] W. Hoeffding, *Masstabinvariante korrelationstheorie*, Schriften Math. Inst. Inst. Angew. Math. Univ. 5 (1940), pp. 179–233.
[16] E.L. Lehmann, *Some concepts of dependence*, Ann. Math. Statist. 34 (1966), pp. 1315–1321.
[17] J. Stoyanov, *Counterexamples in Probability*, 2nd ed., Wiley, New York, 1997.