ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/334726352

Statistical Modeling (CDT-13)

Preprint · July 2019

DOI: 10.13140/RG.2.2.28076.62087/1

CITATIONS
0

READS
1,719

1 author:

Luciano da F. Costa
University of São Paulo
734 PUBLICATIONS **13,156** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Statistical Modeling and Simulation View project

mathematical modeling View project

All content following this page was uploaded by Luciano da F. Costa on 20 November 2020.

The user has requested enhancement of the downloaded file.

Statistical Modeling
(CDT-13)

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

v1: 29th May. 2019
v.1.1: 20th Nov. 2020

Abstract

Probability and statistics substantially underlie scientific and technologic research. In this work we provide an introduction to of some of the most important basic statistical topics, from the concept of random experiment to parametric and non-parametric estimation, also including statistical moments as well as random variable transformations. Special attention is given to the adoption of the Dirac delta 'function' as a means of achieving a unified modeling approach to discrete and continuous random variables and probability distributions.

'La semplicità è la sofisticazione finale.'

Leonardo da Vinci.

1 Introduction

Will it rain tomorrow? Given the dry season right now in my place, it is unlikely, but not impossible. We live in a world of enduring uncertainties, and we can only be certain that nothing is absolutely certain or precisely predictable.

Along millennia, humans have had to cope with uncertainties regarding future events. Probability and statistical theory (e.g. [1, 2, 3, 4]) the most well-succeeded approaches that can help us to understand and handle uncertainties. Starting with analyses of games of chance, the field of probability and statistics found its way into virtually every scientific and technological area, from physics to biology. Currently, probability and statistics play a special role with respect to *artificial intelligence*, *pattern recognition* and *deep learning*.

Despite its vast range of applications and general importance, the basic statistical concepts and methods are relatively simple and accessible. In the current work, we present an introductory and relatively informal first look at some of the most important related topics.

We treat both discrete and continuous random variables and probability distributions in a unified manner, thanks to the adoption of the Dirac delta 'function'. Some important continuous probability distributions are presented and discussed, including the uniform, constant and normal distributions. Moment characterization of random variables is also briefly addressed, including central moments. The possibility of transforming one random variable into another, and how to obtain the probability distribution of the latter in some cases are also presented, as well as the important topics of parametric and non-parametric estimation.

2 Random Experiments, Outcomes, and Events

One important concept is statistics is that of a *random experiment* of interest, such as throwing a dice, or trying to predict the weather. More formally, a random experiment is such that its *outcome* is uncertain. For generality's sake, experiments with known outcome are also understood as random experiments, so that virtually every experiment can be treated as being random.

It is essential to specify as much precisely as possible the random of experiment of interest. For instance, in the case of the above example of throwing a dice, we need to define which dice will be used (its shape, weight, size, hardness, etc.), when and where the experiment will be performed, how the dice will be thrown (and this involves many mechanical aspects such as angle, force, height, etc.), as well

as every other aspect that can influence the experiment outcome such as air resistance, surface friction, and so on.

Given that any event in the real world is potentially influenced by an *infinite* range of effects (e.g. even the moon gravitation can have some influence, tiny as it may be, on the dice outcome [5, 6]), we conclude that random experiments *cannot be fully specified*, implying some level of error and also that unexpected or biased results can be obtained as a consequence of the incomplete description. In addition, we also have experimental errors and noise typically affecting the measurements.

This situation is completely analogue to that found in *scientific modeling* (e.g. [5]), in which the validity and predicting power of any model is limited by not being able to incorporate every effect potentially influencing the phenomenon of interest. The best one can hope for is obtaining an as much as possible complete specification of the random experiment of interest, as well as controlling its environment.

Specifying the random experiment is so important because it allows us to identify the set Ω of all possible *outcomes*, also called the *universe* set. For instance, in the case of throwing a dice and observing the result, we have $\Omega = \{1, 2, 3, 4, 5, 6\}$.

It is now possible to define *event* as any subset of Ω. Recall that the empty set $\emptyset$ is always a subset of any set, including Ω. The subset $A_0 = \emptyset$ is called the *impossible* event. Other possible events in the case of the dice example are $A_1 = \{1\}$, $A_2 = \{1, 3\}$, $A_3 = \{2, 4, 6\}$, $A_4 = \{1, 2, 3, 4, 5, 6\} = \Omega$. The latter is called the *certain* event, as it necessarily contains one of the possible outcomes of the random experiment.

It is very important to keep in mind that *event* and *outcome* are not equal. For instance, in the case of the dice, we have events that are sets of outcomes, such as $\{1, 3, 5\}$. The outcomes in the dice example are only the numbers 1, 2, 3, 4, 5, and 6. Moreover, events are necessarily sets (by being subsets), while outcomes are individual observations (e.g. the number on the facets of a dice). So, in brief:

$$\boxed{\text{Event} \neq \text{outcome.}}$$

As events are sets, given two events A and B, it is possible to consider their intersection $A \cap B$, union $A \cup B$, complementation $A^C = \Omega - A$, etc. When $A \cap B = \emptyset$, events A and B are said to be *mutually exclusive*.

3 The Concept of Probability

As defined, events provide a relatively precise specification of *any* possible result we could be interested regarding a random experiment. Observe that events include sets containing the individual experiment outcomes, but are not limited to them as they can refer to *combinations* of outcomes. For instance, in the case of dice throwing, the composite event $\{1, 2\}$ is understood as obtaining 1 *or* 2 as result.

Given that we can specify situations of particular interest as events, it is now important to associate *probabilities* to them. This can be done in more than one way, including theoretic and experimental ways, as we will discuss soon. First, let's formalize probability as a function $P()$ acting on an event A and producing a real value $P(A)$ in the interval $[0, 1]$ as a result, i.e.

$$P : A \subset \Omega \longmapsto P(A) \in [0, 1] \in \mathbb{R} \tag{1}$$

We necessarily have that

$$P(\Omega) = 1 \tag{2}$$

$$P(\emptyset) = 0 \tag{3}$$

$$P(A^C) = 1 - P(A) \tag{4}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{5}$$

We also have that when $P(A \cap B) = P(A)P(B)$, the two events A and B are said to be *independent*.

Theoretical definitions of the probability of an event typically refer to an abstract random experiment. For instance, we can consider a *hypothetical* random experiment involving a perfectly symmetric dice. In this case, it follows by *symmetry* that all outcomes have the same probability, i.e. are *equiprobable*, implying $P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = p$. Observe that, formally speaking, we cannot write $P(1)$, using $P(\{1\})$ instead, since '1' is not an event (which is necessarily a set), but an individual outcome.

Now, we need to determine p. This can be achieved by considering that the set of considered events are mutually exclusive, which allows us to apply Equation 5: $P(\{1\} \cup \{2\} \cup \{3\} \cup \{4\} \cup \{5\} \cup \{6\}) = P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) + P(\{5\}) + P(\{6\}) = P(\{1, 2, 3, 4, 5, 6\}) = 6\,p = 1$, so that $p = 1/6$.

However, when dealing with real-world situations, symmetry is hardly perfect and cannot have fuly comprehensive information about the random phenomenon of interest. Therefore, it is necessary to resource to performing the random experiment many times while taking record of the outcomes.

For instance, let's suppose that we have a specific dice to be through in specific situations, and the events corresponding to observing each of the six dice facets. Assume that the experiment was performed $N = 100000$ times, and the results were as shown in the following table.

The probability of each of these events A_i can be estimated as the ratio between the number of occurrences of

Table 1: Counts of obtaining each of the faces of a dice as an event, and respectively estimated probabilities.

event	count	*probability P*
$A_1 = \{1\}$	167938	0.167938
$A_2 = \{2\}$	170680	0.170680
$A_3 = \{3\}$	173784	0.173784
$A_4 = \{4\}$	173761	0.173761
$A_5 = \{5\}$	167582	0.167582
$A_6 = \{6\}$	146255	0.146255

A_i and the total number of random experiments N, i.e.

$$\boxed{P(A_i) = \frac{\#(A_i)}{N}} \tag{6}$$

where $\#(A_i)$ means the number of elements (or cardinality) of set A_i.

We observe that these probabilities are unlikely to be equiprobable, as they present relatively large deviations from the expected value of 1/6 after one million experiments.

The probability to be assigned to real world events is formally defined as an extension of the above equation as the number of random experiments tends to infinity, i.e.

$$\boxed{P(A_i) = \lim_{N\to\infty} \frac{\#(A_i)}{N}} \tag{7}$$

Interestingly, the *law of large numbers* (e.g. [7]) ensures that, provided a random experiment preserves its characteristics along time (a *stationary* experiment), the estimation of the respective probabilities by using Equation 7 will converge to its actual value as we approach an infinite number of samples.

However, full convergence cannot be obtained in practice, as it is impossible to perform an infinite number of experiments, and also because the experiment conditions are not completely specified and/or tend to change along time. For instance, the dice can get worn, the air density may change, etc. That is why is so important to keep the environmental conditions as stable as possible during the experiments.

4 Random Variables and Probability Distributions

Though we described a means of assigning probabilities to events corresponding directly to specific outcomes of a random experiment with discrete outcomes, it would be very convenient to incorporate *continuous measurements* often found in practice into the so far developed framework. Examples of these include the weight of an apple and the outside temperature today.

In probability and statistics, these measurements are called *random variables*. Keep in mind that we will henceforth understand random variables typically corresponding to the *outcomes* of a specific random experiment, but formally the probability of a random variable refers to the *set* containing each respective outcome.

Observe that the facets of a dice can already be understood as a numeric measurements with discrete values. It is also possible to have discrete outcomes with *categoric* nature, such as the city in which a person was born, or color names. In such cases, it is still possible to map the categorical labels into respective numeric values, though this is often imply some arbitrariness.

What is necessary now is to extend the concept of probability to continuous outcomes, that therefore can take an infinite number of values. For instance the weight of an apple is non-negative real value. All the already mentioned outcomes or measurements are called *random variables*, which are often identified by capital letters such as X. Small caps are typically reserved for expressing the *values* of the respective random variable.

In order to integrate discrete and continuous random variables into a unified approach to statistical models, we will resort to the *Dirac delta* 'function'. Formally speaking, this 'function' is not a function, but a *functional* that maps a function $g(x)$ into its value at zero, i.e. $g(0)$, as studied in a branch of mathematics called *distribution theory* (e.g. [8]).

However, for simplicity's sake, in the current text we will understand the Dirac delta 'function' more informally as the limit of some function that has area 1, such as the rectangular function $r(x)$ centered at the origin and having width a and height $1/a$. The Dirac delta function can be understood as the limit of this function as we make a smaller and smaller, i.e.

$$\delta(x) = \lim_{a\to\infty} r(x) \tag{8}$$

Though the height of this function increases continuously as its width is decreased, its unitary area is conserved, and we can write

$$\int_{-\infty}^{\infty} \delta(x)dx = 1 \tag{9}$$

Observe that the Dirac delta can also be defined as the limit of an infinite number of alternative functions, such as the gaussian normalized to have unit area.

Informally, we can also write $\delta(x)g(x) = \delta(x)g(0)$,

which expresses the 'sampling' capability of the Dirac delta function, meaning that

$$\int_{-\infty}^{\infty} \delta(x)g(x)dx = g(0) \tag{10}$$

Interestingly, we can now use the Dirac delta function to represent the six probabilities associated to the dice example as the function involving the addition of the 6 respective Dirac deltas, i.e.

$$p_\delta(x) = \frac{1}{6}\sum_{i=1}^{6} \delta(i) \tag{11}$$

which is shown in Figure 1.

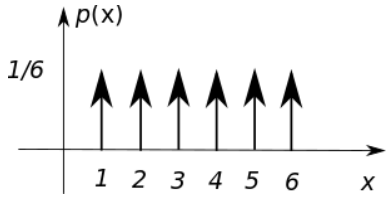

Figure 1: The probability distribution respective to the dice experiment as represented by a sum of Dirac delta 'functions' placed at the respective discrete outcome values.

The area of $p_\delta(x)$ can be simply calculated by adding the values obtained from Equation 9, yielding 1 as result. Yet, the probability value at any of the points 1, 2, 3, 4, 5, and 6 is infinite, so that the probability of any of the respective outcomes needs to be understood as the area of the respective Dirac delta distribution, in this case 1/6.

Observe that, while the set containing all positions (X) of the Dirac deltas corresponds to every possible outcome and therefore define the respectively associated universe X, each subset of the universe can be understood as specific *events* .

We can infer from the above outlined application of the Dirac delta to obtain probability functions that these representations have a *density* nature. Indeed, if we have a density function, e.g. of mass, the quantity of mass within an interval $[1, b]$ is obtained by integrating the density function along this interval. In probability, these functions are typically called *probability distributions* or densities.

Let's develop the concept of continuous probability distributions by considering two-dimensional dice with varying number of facets n, all with the same size (see Figure 2). The smallest such dice has 3 facets (an equilateral triangle). These dice are thrown within a 2D limited space (e.g. in the interstice between two sheets of glass), and the outcome is understood as corresponding to the facet resulting in touch with the floor.

The labels associated to each of the n facets are distributed uniformly between two boundary values a and b

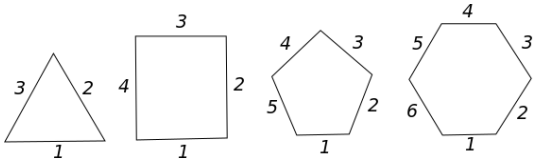

Figure 2: Two-dimensional dice with 3 to 6 facets.

(e.g. within the interval $[0, 1)$), so that the facet label associated to the $i-$th facet is $(i-1)/n$, with $i = 1, 2, \ldots, n$.

That these outcomes are discrete random variables that can be represented by Dirac delta functions of the type

$$p_\delta(x) = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{i-1}{n}\right) \tag{12}$$

If we take the limit when $n \mapsto \infty$, we get the *continuous* probability function $p_\delta(x) = 1$, which can be verified to be a probability distribution.

Formally speaking, any function $p(x)$ obeying the following conditions can be a candidate for being the *probability distribution* modeling the outcome of some random experiment:

$$p(x) \geq 0, \forall x; \tag{13}$$

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{14}$$

and we have that $p_\delta(x)$ above is indeed a probability distribution.

Given a probability distribution $p(x)$, we cannot understand the value of this function at a specific point, e.g. $p(2)$, as a probability. Instead, this value corresponds to the *density of probability* of the outcome at that specific point. So, the probability that the outcome lies within an interval $[a, b]$ can be immediately calculated as the area of $p(x)$ along this interval, i.e.

$$p(a \leq x < b) = \int_a^b p(x)dx \tag{15}$$

Fortunately, this result also holds for discrete probability distributions represented in terms of sums of Dirac delta 'functions'. For instance, we have

$$P(2 \leq X \leq 4) = \int_{1.5}^{4.5} p_\delta(x)dx = \frac{1}{2}. \tag{16}$$

All in all, we have that the adoption of the Dirac delta 'function' to represent the probability of numeric outcomes turns out to allow an unified and convenient approach capable of handling both *discrete* and *continuous* random variables in terms of 'continuous' probability distribution functions potentially involving Dirac deltas.

As with the dice example, continuous random variables X will also have respectively associated universe set and

possible events. The former corresponds to the set of X values where $p(x)$ is non-null, typically corresponding to an interval or the union of intervals, while the latter corresponds to any respective subset.

It is interesting to observe at this point that, though most real world events are probabilistic, it turns out that it is particularly challenging to produce perfectly uniform random variables. Indeed, much research has been done in devising effective methods for random number generation (e.g. [9]). So, it is somehow surprising to experience such difficulties in a probability permeated universe.

Probability distributions correspond to a key concept in probability and statistics, especially because they provide effective *models* of aspects (i.e. those quantified by the considered random variables) of a respective random experiment. Indeed, the probability function associated to a specific random variable X can be understood as provide as much information about the behavior of X as it would be possible to obtain.

It is also interesting to consider *cumulative distributions*, which are defined from the respective probability distribution functions $p(x)$ as follows

$$P(X \leq x) = \int_{-\infty}^{x} p(X)dX \tag{17}$$

So, cumulative distributions accumulate the probability along the X values, providing an indication of the overall probability of a specific observation x being comprised in the interval extending from $-\infty$ to x.

It follows immediately from Equation 17 that

$$p(x) = \frac{dP(x)}{dx}. \tag{18}$$

5 Statistical Moments

We have already seen that, given a random variable X, it is possible to model it in terms of its respective probability distribution $p(X)$. Because this function is so important for the characterization of the variable X, it is often interesting to derive some properties from it. A particularly interesting way to do so is through *functionals*, i.e. mappings from the function $p(X)$ to real values, such as the area (which yields 1). The *expectance* $E[X]$ of a random variable X is such a functional, being defined as

$$E[X] = \int_{-\infty}^{\infty} X\, p(X)dX \tag{19}$$

For instance, in the case of the uniform distribution $p(X) = 1$ in the interval $[0, 1)$, we have

$$E[X] = \int_{0}^{1} X\, p(X)dX = \frac{1}{2}X^2\Big|_0^1 = \frac{1}{2} \tag{20}$$

The expectance of X coincides with the concept of the *mean* or *average* of that random variable. $E[X]$ is also known as the *first moment* of $p(X)$. Observe that *expectance* refers to a variable, while *moment* refer to the probability distribution function.

It is interesting to observe that the interpretation of Equation 19 as the average (or mean) of the respective random variable X can be easily verified as follows. Let's go back to the dice experiment, assume that we obtained the following number of observations of each possible outcome:

Observe that this application of the expectance Equation 19 corresponds precisely to what one would typically do when calculating the average of the set of obtained random variable values: adding all obtained values (i.e. 69) and dividing the result by the total number of experiments (i.e. 20).

It is possible to extend the concept of moments to higher orders, also known as the $k-$th moments of $p(X)$:

$$E^k[X] = \mu = \int_{-\infty}^{\infty} x^k\, p(x)dx \tag{21}$$

By considering the random variable X as being transformed into another random variable $\tilde{X} = X - \mu$, where $\mu = E[x]$ is the average of X, we can define the $k-$th *central moments* of X as

$$E^k[X - \mu] = \int_{-\infty}^{\infty} (X - \mu)^k\, p(X)dX \tag{22}$$

The second central moment corresponds to the frequently used concept of *variance* $\sigma^2 = E^2[X - \mu]$, while its positive square root corresponds to the *standard deviation* $\sigma = +\sqrt{\sigma^2}$ of the random variable of interest X. Both these statistical measurements quantify the *dispersion* of X around its average, though with different units.

It is also possible to define non-dimensional quantifications of dispersion, such as the *coefficient of variation* of a random variable X defined as $c_v = \sigma/\mu$.

Interestingly, it can be shown that under certain circumstances (e.g. by using the Carleman theorem [10]), if we take all the possible moments of a probability distribution $p(x)$, this will allow us to recover $p(x)$, meaning that the set of infinite moments can provide as much information as the distribution probability about the random variable of interest X. That is so because each successive moment provides some new information about the properties of $p(x)$, until the mapping between moments and $p(x)$ becomes bijective and, therefore, invertible.

6 Some Probability Distributions

The function p_δ derived above with respect to the limiting case of a $2-$dimensional dice with an infinite number

Table 2: The mean interpretation of the the expectance $E[X]$ of a random variable X corresponding to a hypothetical dice experiment.

event	*count*	$(X)(count)$	$p(X) = \frac{count}{N}$	$(X)(p(X))$
$A_1 = \{1\}$	3	$(1)(3) = 3$	$\frac{3}{20}$	$(1)(\frac{3}{20}) = \mathbf{\frac{3}{20}}$
$A_2 = \{2\}$	4	$(2)(4) = 8$	$\frac{4}{20}$	$(2)(\frac{4}{20}) = \mathbf{\frac{16}{20}}$
$A_3 = \{3\}$	3	$(3)(3) = 9$	$\frac{3}{20}$	$(3)(\frac{3}{20}) = \mathbf{\frac{9}{20}}$
$A_4 = \{4\}$	3	$(4)(3) = 12$	$\frac{3}{20}$	$(4)(\frac{3}{20}) = \mathbf{\frac{12}{20}}$
$A_5 = \{5\}$	5	$(5)(5) = 25$	$\frac{5}{20}$	$(5)(\frac{5}{20}) = \mathbf{\frac{25}{20}}$
$A_6 = \{6\}$	2	$(6)(2) = 12$	$\frac{2}{20}$	$(6)(\frac{2}{20}) = \mathbf{\frac{12}{20}}$
Total:	$N = 20$	69	1	Mean $= 3.45 = \mathbf{\frac{69}{20}}$

of facets represents one of the most important probability distributions, called *uniform*. More specifically, the uniform distribution can be defined as

$$p(x) = \begin{cases} \mathbf{c} = \frac{\mathbf{1}}{\mathbf{b-a}}, & \text{for } x \in [a, b] \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

where $\mathbf{c}$ is a real constant value so that the total area of $p(x)$ is 1.

This distribution is particularly important because it describes a situation in which all outcomes within the interval $[a.b]$ have the same density probability. Though this distribution can be often used to model abstract, theoretical situations such as the above considered $2-$dimensional dices, there are not many natural random variables that can be precisely modeled by this distribution. Indeed, it is difficult to think of natural experiments yielding outcomes perfectly uniform in the statistical sense.

So, it is not surprising that many other probability distribution functions have been derived and proposed. As a matter of fact, it should be remembered that there is an *infinite* number of potential probability distributions, as we can always renormalize a large class (e.g. having finite area) of functions so that they are always non-negative and have unit area. In this section, we review some of the most frequently adopted probability distribution functions, in addition to the already seen uniform distribution.

Another particularly simple and singular example is the *constant* probability function, defined as

$$p(x) = \delta(x_0) \tag{24}$$

This probability distribution is associated to random experiments having *certain* event, indicating that the outcome will always be equal to x_0.

Other probability distributions are presented in Table 3, together with their respective parameters, mean and standard deviation. This table also provides the respectively involved parameters, as well as the respective mean and standard deviation of X. By *parameters* it is meant the variables that can vary from one type of random experiment to another, allowing the adaptation of the model to the specific circumstances, but which remain fixed along different realizations of the same experiment.

Of particular importance is the *normal distribution*, characterized by its symmetric bell shape. This distribution is particularly important because of the central limit theorem, which states under certain circumstances that sums of independent random variables tend to present this type of distribution.

Observe that the normal distribution involves two parameters: the average μ and standard deviation σ of the respective random variable X. This is one example of probability distribution having statistical moments as parameters, but this is not always verified for other distributions.

Table 4 provides the percentage of probability comprised in the interval with length 2 to 6 standard deviations centered at the average of any normal distribution.

Observe that the interval of X having length equal to 4 standard deviation will comprise a very significant (i.e. 95.45%) percentage of the possible outcomes of X, while 6 standard deviations will incorporate almost every possible result. Yet, it is important to keep in mind that the normal distribution tends *asymptotically* to zero at both its left and right extremities, reaching null value only at $-\infty$ and ∞, respectively.

The *log-normal* is another probability distribution, which is associated to the normal but considers only non-negative values of the random variable X, being a potential choice for modeling some random experiments producing this type of outcomes.

Table 3: Some frequently used and/or interesting probability distribution functions and their respective parameters, mean and standard deviation.

Distribution	$p(x) =$	parameters	mean (μ)	st. dev. (σ)
Uniform	$\mathbf{c}$ for $x \in [a, b]$; 0 otherwise	a, b	$\frac{a+b}{2}$	$\frac{1}{12}(b-a)^2$
Constant	$\delta(x_0)$	x_0	x_0	0
Exponential	$\lambda e^{-\lambda x}$	$\lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda}$
Normal	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ, σ	μ	σ
Log-Normal	$\frac{1}{x}\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{ln(x)-\mu}{\sigma}\right)^2}$	μ, σ	$e^{\mu+\frac{\sigma^2}{2}}$	$\sqrt{(e^{\sigma^2}-1)e^{(2\mu+\sigma^2)}}$
Logit	$\frac{1}{\sigma\sqrt{2\pi}}\frac{1}{x(1-x)} e^{-\frac{1}{2}\left[\frac{-log\left(\frac{1}{x}-1\right)-\mu}{\sigma}\right]^2}$	μ, σ	–	–
Student's t	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}^{-\frac{\nu+1}{2}}\right)$	$\nu > 0$	0 for $\nu > 1$	$\sqrt{\frac{\nu}{\nu-2}}$ for $\nu > 2$

Table 4: The probability comprised in the interval having length of 2 to 6 standard deviations centered at the average of any normal distribution.

$P(X-\sigma \leq X < X+\sigma)$	68.27%
$P(X-2\sigma \leq X < X+2\sigma)$	95.45%
$P(X-3\sigma \leq X < X+3\sigma)$	99.73%

7 Random Variable Transformations

Given a random variable, or measurement X, we can be interested in characterizing the statistical properties of a respectively derived new variable, such as $Y = X^2$ or $Y = log(X)$, provided some circumstances are met (e.g. $x > 0$ in the latter example). Such modifications of random variables are called *random variable transformations*. Since every measurement from the real world is a random variables, any scientific law, equation or relationship involve random variable transformations.

Let's consider the problem of deriving the probability distribution of a new random variable Y derived from X. We shall consider the so-called *distribution function technique*, which is presented in terms of the following example.

Let X be a random variable characterized by the respective distribution function

$$p(x) = \begin{cases} \frac{1}{3}, & \text{for } 1 \leq x < 4 \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

Let $Y = f(X) = X^2$ be a new random variable of interest. We can make

$$P(Y \leq y) = P(X^2 < y) = P(X < y^{\frac{1}{2}}) = \int_{-\infty}^{\infty} p(x)dx = \int_{1}^{y^{\frac{1}{2}}} \frac{1}{3}dx = \frac{1}{3}x\Big|_{1}^{y^{\frac{1}{2}}} = \frac{1}{3}(\sqrt{y}-1) \tag{26}$$

so as to obtain the cummulative distribution function $P(Y)$ of the new random variable Y. Now, in order to derive the density distribution $p(Y)$, all we have to do is to differentiate $P(Y \leq y)$ with respect to y, i.e.

$$p(y) = \frac{dP(Y \leq y)}{dy} = \frac{1}{6}\frac{\sqrt{y}}{y},\ 1 \leq y < 16 \tag{27}$$

which can be rewritten as $p(y) = \frac{1}{6}\frac{\sqrt{y}}{y}$, $1 \leq y < 16$.

As expected, it can be verified that $\int_{-\infty}^{\infty} p(Y)dY = 1$. Observe that this particular technique requires that the random variable transformation is one-to-one (or injective) and that $f(X)$ strictly increases along the interval of interest. An analogue method can be used for one-to-one and strictly decreasing random variable transformations (e.g. [3]).

8 The Two Main Applications of Statistical Modeling

Having discussed some of the main statistical concepts, we are now in a better position to appreciate the two main applications of statistical modeling, which are: (i) to model quantities that have intrinsic and unpredictable

variability, such as electrons emanating from a cathode ray tube; and (ii) limited precision of every possible real-world measurement, as when reading a length by using a ruler. In this section, we will briefly present and discuss each of these situations.

In the first type of application, namely modeling quantities naturally incorporating unpredictable variability, let's consider the situation depicted in Figure 3.

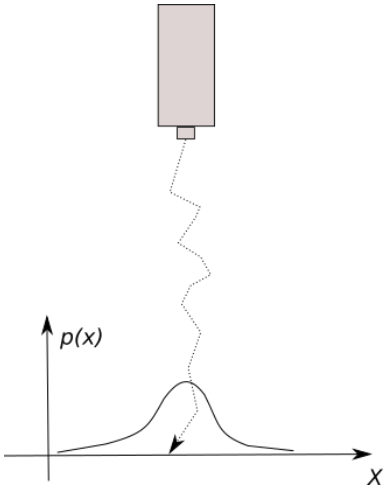

Figure 3: Electrons are produced by a cathode ray tube, traveling in interaction with air until hitting a one-dimensional surface. The position of the hit along X is a random variable that can be shown analytically to have a normal distribution. The dispersion of X is caused by interactions between the issued electrons and air molecules and ions, as well as other electrons.

We have electrons being issued from a cathode ray tube (CRT) that hit a given surface. For simplicity's sake, this hypothetical experiment is considered to be one-dimensional, in the sense that the position of the electrons when hitting the surface can be properly described by a respectively associated one-dimensional random variable X.

Because of the interactions between each electron with other particles and ions in the air, each of the issued electrons will perform a *random walk* on its way before hitting the surface. It can be shown analytically that this dynamics corresponds to a *diffusion* random process implying that the density probability of X will adhere to a gaussian function, therefore defining a respective normal distribution.

The importance of this simple example is twofold. First, it corresponds to a relatively rare situation in which the probability density function of a given random variable can be estimated analytically from physical and statistical principles. Then, we have that it provides an illustration of statistical principles being used to model variations that are intrinsic to the observed phenomenon, and not a consequence of experimental errors or limited resolution (though these will also be present in real-world experiments).

Now, let's consider the second main application of statistical modeling, namely characterizing experimental errors and/or limited resolution. Consider the situation shown in Figure 4.

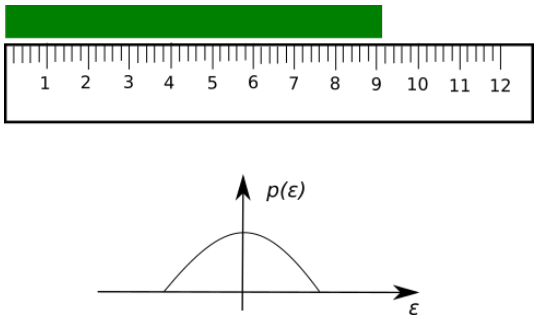

Figure 4: The application of statistics to model experimental measurement errors. An object is to be measured with a finite resolution ruler having smallest division of $\Delta X = 2cm$. The estimated length X, corresponding to a random variable, will be characterized by an inevitable error $\epsilon < \Delta X/2$ which is also a random variable presenting a respective dispersion quantifiable by using the standard deviation. A hypothetical probability density function $p(\epsilon)$ is also shown in this figure.

Now, we have a standard ruler being applied for estimating the length of a given object. Because the ruler has a limited resolution, with minimal spacing between marks corresponding to ΔX, it is impossible to read more than two significant digits in any measurement. Moreover, the estimation of the last digit will inevitably be associated to an error that will probably be smaller than $\Delta X/2$, provided the reading is properly performed. This example illustrates a practical situation in which, though the length of the object being measured is completely stable at the scale of its human observation, the resulting measurement X, itself a random variable, will inevitably contain an error $|\epsilon| < \Delta/2$. This error can also be understood as a random variable, with an associated probability density function that should not be too different from a normal distribution, though with limited support (i.e. not extending further than $|\epsilon| < \Delta/2$).

In principle, every real-world measuring device has a limited resolution and is potentially exposed to experimental errors, noise, and other perturbations. In particular, it is interesting to observe that noise can also be treated in probabilistic terms.

9 Parametric and Non-parametric Estimation

So far, we assumed that the probability distribution functions of a given random variable were somehow known, providing respective models. In some cases it is possible

to derive theoretically the distribution through mathematical developments, such as is the case with inferring that the electrons constituting the beam of a cathodic ray tube follow a normal distribution. This is achieved by taking into account that the electrons perform a random walk as they move along the beam, and this type of stochastic dynamics can be shown to lead to a normal distribution (e.g. [11]). However, in practice the distribution functions are often unknown at first, and have to be estimated.

In this section we will briefly discuss two of the main approaches for estimating the probability distribution of a random variable: (i) *parametric*; and (ii) *non-parametric*.

In the former approach, the random variable of interest is known, or assumed, to have a specific type of probability distribution, such as uniform, normal, etc. So, the estimation of the statistical model involves inferring only the involved parameters.

This can be performed according to estimating equations derived for each of the involved parameters. These equations are obtained by imposing optimal statistical adherence to the data, in approaches such as involving the *maximum likelihood* method (e.g. [12]). These estimators can be obtained from the literature with respect to a large range of distribution functions.

Let's illustrate parametric estimation with respect to statistically modeling of a random variable sampled from a normal distribution with average $\mu = 3$ and standard deviation $\sigma = \sqrt{2}$. For the purpose of our example and discussion, we assume that all that is know is the set of 8 respective observations $S = \{1.88; 2.54; 6.12; 3.14; 3.26; 6.43; 3.92; 0.47\}$. These values are illustrated in Figure 5(a).

The average and standard deviation of the values in S can be estimated as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{28}$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)^2 \tag{29}$$

where N is the number of samples, yielding $\mu_e = 3.4697$ and $\sigma = 2.0187$, which can be verified to be similar to the respectively original values.

The so obtained normal distribution is shown in solid lines in Figure 5(b) together with the normal distribution using the original parameters (dashed lines). It can be verified that this parametric estimation procedure yielded a reasonable statistical model of X, especially considering that only 8 samples were available.

The other main possibility for obtaining the probability distribution of a given random variable, namely *non-parametric estimation*, does not require the assumption of any particular type of *a priori* distribution function. Instead, the sought function is obtained through some kind of 'interpolation' of the relative frequency histogram describing the random variable being modeled.

There are several possibilities for deriving a non-parametric distribution (e.g. [13]), and here we consider only the smoothing of the relative frequency histogram by *convolving* (e.g. [14]) it with some smooth kernel (e.g. a normal function).

First, we need to define what is the *relative frequency histogram* of a given random variable X. We start with a set of N respective observations $S = \{X_1, X_2, \ldots, X_N\}$. Having identified the minimum X_m and maximum X_M respective values, we consider an interval $X \in [a, b]$ so that $a \leq X_m$ and $b \geq X_M$. This interval is divided into n subintervals, or *bins*, henceforth represented as $b_1, b_2, \ldots, b_n$.

Often, these bins are assumed to have the same size ΔX. All bins are initially set with zero count. Then, each of the elements in S is taken and the value of the bin in which it is contained is incremented. In the end, we have the *histogram* $h(i)$ of X for resolution ΔX. If the values in each of the bins are divided by N, we obtain the respective *relative frequency histogram* $f_R(i)$.

Now, we can assign a Dirac delta function at the very middle x_i of each bin, having area equal to the the respective accumulated value. More formally, we can define the function

$$F_R(x) = \sum_{i=1}^{n} f_R(i)\delta(x_i) \tag{30}$$

It can be immediately verified that this function is non-negative and has area equal to 1, and therefore already provides a valid candidate distribution function for X. Indeed, it is interesting to observe that the probability distribution function $p(x)$ of the considered random variable can be defined in terms of the following limit:

$$p(x) = \lim_{\substack{\Delta X \mapsto 0 \\ N \mapsto \infty}} F_R(X) \tag{31}$$

However, this would require an infinite number of samples, which is not available in our hypothetical example. Indeed, the distribution probability obtained from the 8 sample values has many gaps between the Dirac deltas, as well as abrupt variations. So, we are interested in interpolating this function in order to obtain a smooth, continuous interpolation.

This can be achieved by convolving $F_R(x)$ with a normal function with average 0 and standard deviation σ, yielding the approximate probability distribution

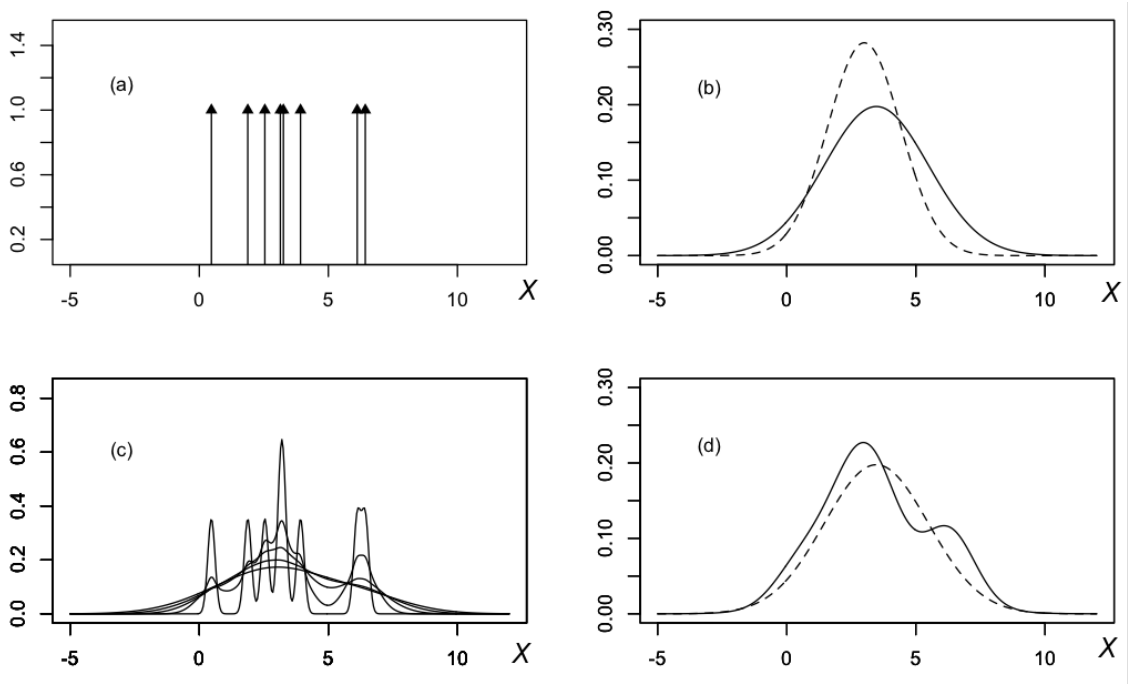

Figure 5: (a): The considered sample values represented as a sum of Dirac deltas. (b): The respective parametric estimation (solid line) compared to the original normal distribution having average $\mu = 3$ and standard deviation $\sigma = \sqrt{2}$. (c): The non-parametric estimations with a normal kernel with $mu = 0$ and $\sigma = 0.1; 0.4; 0.7; \ldots, 1.6$. (d): The chosen probability distribution (for $\sigma = 0.7$) shown in solid lines, while the previously parametrically estimated distribution is shown in dashed lines.

$$\tilde{p}(x) = [F_R * g_\sigma](\xi) = \int_{-\infty}^{\infty} g_\sigma(\xi - x) F_R(x) dx \qquad (32)$$

The convolution of a given function $f(x)$ with a sum of Dirac delta 'functions' can be verified (e.g. [14]) to correspond to the operation of adding $f(x)$ at the position of each of the Dirac deltas. Thus, given that the normal function $g_\sigma(x)$ is intrinsically smooth, it will fill the gaps between the Dirac deltas, providing an interpolation of the original density.

It is also possible to perform kernel-based non-parametric estimation of probability distributions by convolving the kernel (a normal distribution in the previous example) directly with the Dirac deltas associated to each sample in S, therefore avoiding the construction of relative frequency histograms. This procedure is simpler, but potentially more difficult to be performed quickly such as by the numerical calculation of the convolution in terms of the fast Fourier transform (e.g. [14]).

Figure 5(c) illustrates the result of the convolution of the Dirac deltas defined by the original samples – shown in Figure 5(a) – with a normal distribution with zero average and successive standard deviations $\sigma = 0.1; 0.4; 0.7; \ldots, 1.6$. This was obtained by adding this normal distribution at each of the Dirac deltas corresponding to each of the original samples X_i.

Observe that the smaller values of σ yield respective probability distributions follow more closely the original sum of Dirac deltas in Figure 5(a), implying respective gaps along the X- axis. Larger values of σ will fill these gaps, yielding a smoother probability distribution that is closer to the original normal shown by dashed lines in Figure 5(b). However, too much smoothing will imply in substantial loss of detail about the original data.

So, non-parametric estimation often requires some criterion to be applied in order to choose between the respective parametric configurations (the choice of σ values in our example). Though it is possible to obtain analytical expressions providing the best parameter choice with respect to some imposed criterion (e.g. quadratic error), for simplicity's sake here we select the smallest value of σ capable of reasonably filling the gaps and yet not smoothing too much the resulting distribution, which would imply in missing too much information (details).

Figure 5(d) illustrates in solid lines the chosen distribution (for $\sigma = 0.7$), while the normal distribution defined by the above discussed parametric estimation is shown in dashed lines. The smaller bump obtained at the righthand side of this distribution is a consequence of the relatively isolated original samples at $X = 6.12$ and $X = 6.43$. Such effects are sometimes called *random fluctuations*, which can impose some level of structure or artifacts on an otherwise uniform or smoother distribution as a consequence of limited number of samples.

10 Concluding Remarks

This text has presented in an introductory manner some of the principal concepts from the probability and statistics field. Having defined what random experiments are, it was possible to infer these correspond to a kind of scientific models taking into account uncertainties in the respectively obtained results. As such, in similar fashion to deterministic scientific counterparts, statistical models are also never guaranteed to be fully precise or complete, because incompleteness in the representation of real world quantities as well as noise and experimental errors.

Yet, statistical models are very important in science and technology because of this very same reason, i.e. they provide a principled means for coping with the uncompletness and some level of experimental error that are unavoidable in practice.

After introducing the concepts of random variables and probability distribution functions in terms of discrete outcomes (such as in the dice experiment), we resorted to the Dirac delta 'function' as a means of integrating the statistical modeling of both discrete and continuous random variables. Some types of probability distributions were presented and discussed, and the important concept of statistical moments was also described and illustrated. We also briefly discussed the useful concept of random variable transformations, as well as presented an overview of parametric and non-parametric estimation.

It is hoped that the current text can motivate and help the reader to probe further not only complementing the presented concepts, but also delving into other important topics such as multivariate statistics (e.g. [15]), decision theory, principal component analysis, and stochastic processes, among many other interesting possibilities.

Acknowledgments.

Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) for sponsorship. This work has benefited from FAPESP grant 15/22308-2.

Costa's Didactic Texts – CDTs

CDTs intend to be a halfway point between a formal scientific article and a dissemination text in the sense that they: (i) explain and illustrate concepts in a more informal, graphical and accessible way than the typical scientific article; and, at the same time, (ii) provide more in-depth mathematical developments than a more traditional dissemination work.

It is hoped that CDTs can also provide integration and new insights and analogies concerning the reported concepts and methods. We hope these characteristics will contribute to making CDTs interesting both to beginners as well as to more senior researchers.

Though CDTs are intended primarily for those who have some preliminary experience in the covered concepts, they can also be useful as summary of main topics and concepts to be learnt by other readers interested in the respective CDT theme. Observe that CDTs come with absolutely no warranty.

Each CDT focuses on a few interrelated concepts. Though attempting to be relatively self-contained, CDTs also aim at being relatively short. Links to related material are provide some complementation of the covered subjects.

The currently available set of CDTs can be found at: `https://www.researchgate.net/project/Costas-Didactic-Texts-CDTs`.

References

[1] E. Kreyszig. *Advanced Engineering Mathematics*. Wiley and Sons, 2015.

[2] M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Pearson, 2011.

[3] D. P. Bertsekas and J. N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, 2008.

[4] SOCR. Probability and statistics ebook. `http://wiki.stat.ucla.edu/socr/index.php/EBook`. [Online; accessed 27-July-2019].

[5] L. da F. Costa. Modeling: The human approach to science. Researchgate, 2019. `https://www.researchgate.net/publication/`

333389500_Modeling_The_Human_Approach_to_Science_CDT-8. Online; accessed 03-June-2019.

[6] L. da F. Costa. Quantifying complexity. Researchgate, 2019. https://www.researchgate.net/publication/332877069_Quantifying_Complexity_CDT-6. Online; accessed 15-July-2019.

[7] Wikipedia. Law of large numbers. https://en.wikipedia.org/wiki/Law_of_large_numbers. [Online; accessed 27-July-2019].

[8] G. van Dijk. *Distribution Theory*. De Gruyter Graduate Lectures, 2013.

[9] A. Kennedy. Monte Carlo methods. https://www.math.arizona.edu/~tgk/mc/notes.html. Course notes, Chapter 3. [Online; accessed 27-July-2019].

[10] S. W. M. Au-Yeung. Finding probability distributions from moments. https://www.researchgate.net/publication/240926497_Finding_Probability_Distributions_From_Moments. [Online; accessed 27-July-2019].

[11] H. C. Berg. *Random Walks in Biology*. Princeton University Press, 1993.

[12] R. B. Millar. *Maximum Likelihood Estimation and Inference*. Wiley and Sons, 2011.

[13] Wikipedia. Kernel density estimation. https://en.wikipedia.org/wiki/Kernel_density_estimation. [Online; accessed 27-July-2019].

[14] E. O. Brigham. *Fast Fourier Transform and its Applications*. Pearson, 1988.

[15] L. da F. Costa. Multivariate statistical modeling. Researchgate, 2020. https://www.researchgate.net/publication/340442989_Multivariate_Statistical_Modeling_CDT-26. Online; accessed 20-Nov-2020.

View publication stats