

KANSREKENING MET VOORTGEZETTE INTEGRAALREKENING

prof.dr. H.C. Tijms, Afdeling Econometrie, Vrije Universiteit Amsterdam

Versie januari 2007

(basismateriaal bij de lerarencursus “Continue Kansrekening met Integraalrekening” voor wiskunde-D op het VWO, Vrije Universiteit, 19 januari 2007)¹

VOORWOORD. Deze bijdrage beoogt om voor het wiskunde-D onderwijs een aanzet te geven tot een module waarin kansrekening voor continue stochastische variabelen op een natuurlijke en motiverende wijze samengaat met voortgezette analyse en integraalrekening. Continue kansrekening biedt binnen het nieuwe vak wiskunde-D geweldige mogelijkheden om op functionele wijze integraalrekening te verdiepen en te verlevendigen en aan de hand van concrete, niet-gekunstelde problemen het wezenlijke belang van integraalrekening te tonen. In veel van deze problemen komen vraagstellingen van geometrische aard naar voren. Kansrekening met analyse, geometrie en simulatie is een gouden combinatie voor het nieuwe wiskunde-D vak.

Het materiaal in deze bijdrage is een bewerking van gedeeltes uit de hoofdstukken 5, 7 en 10 van het boek H.C. Tijms, *Understanding Probability*, Cambridge University Press, 2003.

CONTENTS

1. The normal curve
2. Random variables
3. A first introduction to probability density function
 - 3.1 Normal density function
 - 3.2 Percentiles
4. Concept of probability density function
 - 4.1 Verification of a probability density
5. Expected value
 - 5.1 Variance
 - 5.2 Drunkard’s walk

Appendix: Geometric Probability

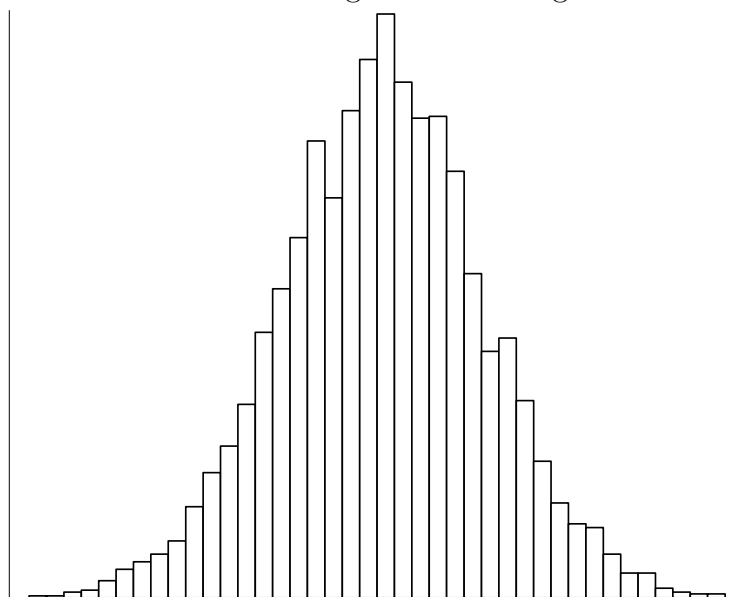
Uitwerkingen van de opgaven

¹De wiskundeleraren Alex van den Brandhof (Vossius Gymnasium Amsterdam) en Fred Pach (Montessori Lyceum Amsterdam) bewerken het materiaal voor een Nederlandse editie en zijn voornemens er een (Zebra)boekje van te maken dat direct geschikt is voor gebruik in de klas binnen een module voortgezette kansrekening voor het nieuwe vak Wiskunde-D.

1 The normal curve

In many practical situations, histograms of measurements approximately follow a bell-shaped curve. A histogram is a bar chart that divides the range of values covered by the measurements into intervals of the same width, and shows the proportion of the measurements in each interval. For example, let's say you have the height measurements of a very large number of Dutch men between 20 and 30 years of age. To make a histogram, you break up the range of values covered by the measurements into a number of disjoint adjacent intervals each having the same width, say width Δ . The height of the bar on each interval $[j\Delta, (j+1)\Delta)$ is taken such that the area of the bar is equal to the proportion of the measurements falling in that interval (the proportion of measurements within the interval is divided by the width of the interval to obtain the height of the bar). The total area under the histogram in Figure 5.1 is thus standardized to one.

Figure 1: A histogram of data

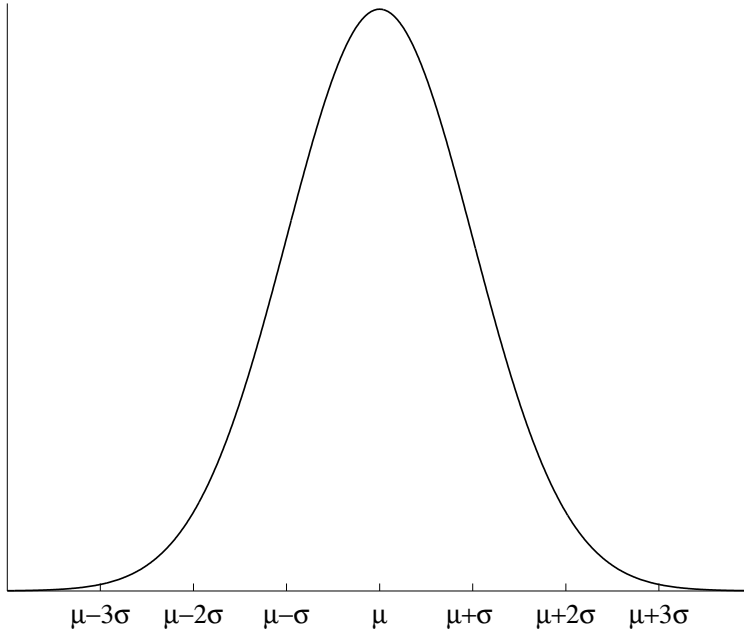


Making the width Δ of the base intervals of the histogram smaller and smaller, the graph of the histogram will begin to look more and more like the bell-shaped curve shown in Figure 2.

The bell-shaped curve in Figure 2 can be described by a function $f(x)$ of the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}.$$

Figure 2: The normal curve



This function is defined on the real line and has two parameters μ and σ , where μ (the location parameter) is a real number and σ (the shape parameter) is a positive real number. The characteristic bell-shaped curve in Figure 2 is called the *normal curve*. It is also known as the Gaussian curve (of errors), after the famous mathematician/astronomer Carl Friedrich Gauss (1777-1855), who showed in a paper from 1809 that this bell curve is applicable with regard to the accidental errors that occur in the taking of astronomical measurements. It is usual to attribute the discovery of the normal curve to Gauss. However, the normal curve was discovered by the mathematician Abraham de Moivre (1667-1754) around 1730 when solving problems connected with games of chance. The pamphlet ‘Approximato ad Summani Terminorum Binomi $(a + b)^n$ in Seriem Expansis’ containing this discovery was first made public in 1738 in the second edition of De Moivre’s masterwork *Doctrine of Chance*. Also a publication of Pierre Simon Laplace (1749-1829) from 1778 contains the normal curve function and emphasizes its importance. De Moivre anticipated Laplace and the latter anticipated Gauss. One could say that the normal curve is a natural law of sorts, and it is worth noting that each of the three famous mathematical constants $\sqrt{2}$, $\pi = 3.141\dots$ and $e = 2.718\dots$ play roles in its makeup. Many natural phenomena, such as the height of men, harvest yields, errors in physical measurements, luminosity of stars, returns on stocks, etc., can be described by

a normal curve. The Belgian astronomer and statistician Adolphe Quetelet (1796-1894) was the first to recognize the universality of the normal curve and he fitted it to a large collection of data taken from all corners of science, including economics and the social sciences. Many in the eighteenth and nineteenth centuries considered the normal curve a God-given law. The universality of the bell-shaped Gaussian curve explains the popular use of the name normal curve for it. Later on in the text we shall present a mathematical explanation of the frequent occurrence of the normal curve with the help of the central limit theorem. But first we will give a few notable facts about the normal curve. It has a peak at the point $x = \mu$ and is symmetric around this point. Second, the total area under the curve is 1. Of the total area under the curve, approximately 68% is concentrated between points $\mu - \sigma$ and $\mu + \sigma$ and approximately 95% is concentrated between $\mu - 2\sigma$ and $\mu + 2\sigma$. Nearly the entire area is concentrated between points $\mu - 3\sigma$ and $\mu + 3\sigma$. For example, if the height of a certain person belonging to a particular group is normally distributed with parameters μ and σ , then it would be exceptional for another person from that same group to measure in at a height outside of the interval $(\mu - 3\sigma, \mu + 3\sigma)$.

After an intermezzo over the concept of random variable and the concept of probability density function, we return to the normal distribution in paragraph 3.1.

2 Random variables

In many chance experiments we are more interested in some function of the outcome of the experiment values than in the actual outcomes. A *random variable* is simply a function that is defined on the sample space of the experiment and assigns a numerical value to each possible outcome of the experiment. For example, in the experiment that consists of tossing a fair coin three times, the random variable X could be defined as the number of times the coin turns up heads. Or: in the experiment consisting of the simultaneous rolling of a pair of dice, the random variable X could be defined as the sum of the values rolled, or as the greater of the two values rolled. Intuitively, a random variable is a variable that takes on its values by chance. A random variable gets its value only after the underlying chance experiment has been performed. It is common to use upper case letters such as X, Y and Z to denote random variables, and lower case letters x, y , and z to denote their possible numerical values. A random variable that can take on only a finite (or countable) number of values is called a *discrete* random variable. A *continuous* random variable is a random variable that can take

on a continuum of values (a finite or an infinite interval). A nice example of a continuous random variable is the decay time of a radioactive particle. But the number of particles emitted by a radioactive source in a fixed time interval is a discrete random variable.

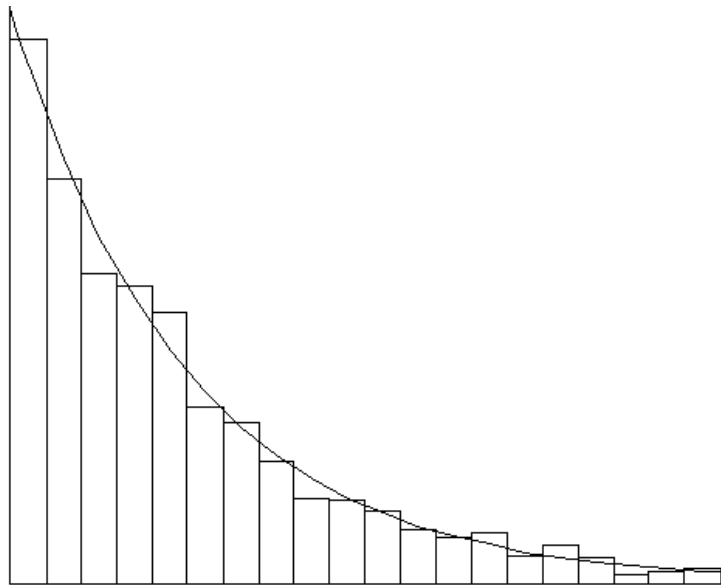
The concept of random variable is always a difficult concept for beginners. For an intuitive understanding of this concept, the best way is to see a random variable as a variable that takes on its values by chance. This view makes clear that any function of a random variable X , such as X^2 and $\sin(X)$, is also a random variable.

3 A first introduction to probability density function

Before giving further properties of the normal curve, it is helpful, informally, to discuss the concept of a probability density function. The function $f(x)$ describing the normal curve is an example of a probability density function. Any non-negative function for which the total area under the graph of the function equals 1 is called a *probability density function*. Any probability density function underlies a so-called *continuous random variable*. Such a variable can take on a continuum of values. The random variable describing the height of a randomly chosen person is an example of a continuous random variable if it is assumed that the height can be measured in infinite precision. Another example of a continuous random variable is the annual rainfall in a certain area. Any probability density function can be seen as a ‘smoothed out’ version of a probability histogram: if you take sufficiently many independent samples from a continuous random variable and the width of the base intervals of the histogram depicting the relative frequencies of the sampled values within each base interval is sufficiently narrow, then the histogram will resemble the probability density function of the continuous random variable. The probability histogram is made up of rectangles such that the area of each rectangle equals the proportion of the sampled values within the range of the base of the rectangle. For this normalization, the total area (or integral) under the histogram is equal to one. The area of any portion of the histogram is the proportion of the sampled values in the designated region. It is also the probability that a random observation from the continuous random variable will have a value in the designated region. As an illustration, take the decay time of a radioactive particle. The decay time is a continuous random variable. Figure 3 displays the probability histogram of a large number of observations for the waiting times between counts from radioactive decay.

Where the probability histogram in Figure 1 resembles a probability density function of the form $(\sigma\sqrt{2\pi})^{-1}e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$ for some values of the parameters μ and $\sigma > 0$, the probability histogram in Figure 3 resembles a probability density of the form $\lambda e^{-\lambda x}$ for some value of the parameter $\lambda > 0$. The area of the histogram between the base points t_1 and t_2 approximates the probability that the waiting time between counts will fall between t_1 and t_2 time units. It is remarked that a probability density function of the form $\lambda e^{-\lambda x}$ for some value of the parameter $\lambda > 0$ is called an *exponential density function*. The exponential density function has also many applications in practice. As another example, the times between serious earthquakes in a certain region can often be described by continuous random variables with an exponential density function.

Figure 3: A histogram of data



Taking the foregoing in mind, you may accept the fact that a continuous random variable X cannot be defined by assigning probabilities to individual values. For any number a , the probability that X takes on the value a is 0. Instead, a continuous random variable is described by assigning probabilities to intervals via a probability density function. In paragraph 4 it will be proved that the probability $P(a \leq X \leq b)$, being the probability that the

continuous random variable X takes on a value between a and b , satisfies

$$P(a \leq X \leq b) = \text{the area under the graph of the density function } f(x) \text{ between points } a \text{ and } b$$

for any real numbers a and b with $a < b$ when $f(x)$ is the probability density function of X . Readers who are familiar with integral calculus will recognize the area under the graph of $f(x)$ between a and b as the integral of $f(x)$ from a to b . Mathematically,

$$P(a \leq X \leq b) = \int_a^b f(x) dx.$$

Any introductory course in integral calculus shows that the area under the graph of $f(x)$ between a and b can be approximated through the sum of the areas of small rectangles by dividing the interval $[a, b]$ into narrow subintervals of equal width. In particular, taking $a = v$ and $b = v + \Delta$ for Δ small, the area under the graph of $f(x)$ between v and $v + \Delta$ is approximately equal to $f(v)\Delta$ when Δ is small enough. In other words, $f(v)\Delta$ is approximately equal to the probability that the random variable X takes on a value in a small interval around v of width Δ . In view of this interpretation, it is reasonable to define the *expected value* of a continuous random variable X by

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx.$$

This definition is the continuous analogon of the definition $E(X) = x_1P(X = x_1) + x_2P(X = x_2) + \cdots + x_nP(X = x_n)$ for a discrete random variable X that can take on only the finite number of values x_1, x_2, \dots, x_n .¹

3.1 Normal density function

A continuous random variable X is said to have a *normal distribution* with parameters μ and σ if

$$P(a \leq X \leq b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} dx$$

for any real numbers a and b with $a \leq b$. The corresponding normal density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} \quad \text{for } -\infty < x < \infty.$$

¹A shorthand notation for $a_1 + a_2 + \cdots + a_n$ is $\sum_{i=1}^n a_i$. This shorthand notation is often used in mathematics. The \int -sign is the continuous analogon of the \sum -sign.

The notation X is $N(\mu, \sigma^2)$ is often used as a shorthand for X is normally distributed with parameters μ and σ . Theoretically, a normally distributed random variable has the whole real line as its range of possible values. However, a normal distribution can also be used for a non-negative random variable provided that the normal distribution assigns a negligible probability to the negative axis. Advanced integral calculus is required to prove for an $N(\mu, \sigma^2)$ random variable X that

$$E(X) = \mu \quad \text{and} \quad E[(X - \mu)^2] = \sigma^2.$$

Thus, the parameter μ gives the expected value of X and the parameter σ gives an indication of the spread of the random variable X around its expected value. The parameter σ is the standard deviation of the random variable X . The concept of standard deviation will be discussed in more detail in Section 5.1.

An important result is:

if a random variable X is normally distributed with parameters μ and σ , then for each two constants $a \neq 0$ and b the random variable $U = aX + b$ is normally distributed with parameters $a\mu + b$ and $|a|\sigma$.

This result states that any linear combination of a normally distributed random variable X is again normally distributed. In particular, the random variable

$$Z = \frac{X - \mu}{\sigma}$$

is normally distributed with parameters 0 and 1. A normally distributed random variable Z with parameters 0 and 1 is said to have a *standard normal* distribution. The shorthand notation Z is $N(0, 1)$ is often used. The special notation

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{1}{2}x^2} dx$$

is used for the cumulative probability distribution function $P(Z \leq z)$ of Z . The derivative of $\Phi(z)$ is the standard normal density function which is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad \text{for } -\infty < z < \infty.$$

The quantity $\Phi(z)$ gives the area under the graph of the standard normal density function left from the point $x = z$. No closed form of the cumulative distribution function $\Phi(z)$ exists. In terms of calculations, the integral for

$\Phi(z)$ looks terrifying, but mathematicians have shown that the integral can be approximated with extreme precision by the quotient of two suitably chosen polynomials. This means that in practice the calculation of $\Phi(x)$ for a given value of x presents no difficulties at all and can be accomplished very quickly.

All calculations for an $N(\mu, \sigma^2)$ distributed random variable X can be reduced to calculations for the $N(0, 1)$ distributed random variable Z by using the linear transformation $Z = (X - \mu)/\sigma$. Writing $P(X \leq a) = P((X - \mu)/\sigma \leq (a - \mu)/\sigma)$ and noting that $\Phi(z) = P(Z \leq z)$, it follows that

$$P(X \leq a) = \Phi\left(\frac{a - \mu}{\sigma}\right).$$

An extremely useful result is the following:

the probability that a normally distributed random variable will take on a value that lies $z > 0$ or more standard deviations above the expected value is equal to $1 - \Phi(z)$, as is the probability of a value that lies z or more standard deviations below the expected value.

This important result is the basis for a rule of thumb that is much used in statistics when testing hypotheses. The proof of the result is easy. Letting Z denote the standard normal random variable, it holds that

$$\begin{aligned} P(X \geq \mu + z\sigma) &= P\left(\frac{X - \mu}{\sigma} \geq z\right) = P(Z \geq z) = 1 - P(Z < z) \\ &= 1 - \Phi(z). \end{aligned}$$

The reader should note that $P(Z < z) = P(Z \leq z)$, because Z is a continuous random variable and so $P(Z = z) = 0$ for any value of z . Since the graph of the normal density function of X is symmetric around $x = \mu$, the area under this graph left from the point $\mu - z\sigma$ is equal to the area under the graph right from the point $\mu + z\sigma$. In other words, $P(X \leq \mu - z\sigma) = P(X \geq \mu + z\sigma)$. This completes the proof of the above result.

3.2 Percentiles

In applications of the normal distribution, percentiles are often used. For a fixed number p with $0 < p < 1$, the $100p\%$ percentile of a normally distributed random variable X is defined as the number x_p for which

$$P(X \leq x_p) = p.$$

In other words, the area under the graph of the normal density function of X left from the percentile point x_p is equal to p . The percentiles of the $N(\mu, \sigma^2)$ distribution can be expressed in terms of the percentiles of the $N(0, 1)$ distribution. The $100p\%$ percentile of the standard normal distribution is denoted as z_p and is thus the solution of the equation

$$\Phi(z_p) = p.$$

It is enough to tabulate the percentiles of the standard normal distribution. If the random variable X has an $N(\mu, \sigma^2)$ distribution, then it follows from

$$P(X \leq x_p) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x_p - \mu}{\sigma}\right) = \Phi\left(\frac{x_p - \mu}{\sigma}\right)$$

that its $100p\%$ percentile x_p satisfies $(x_p - \mu)/\sigma = z_p$. Hence

$$x_p = \mu + \sigma z_p.$$

A much used percentile of the standard normal distribution is the 95% percentile

$$z_{0.95} = 1.6449.$$

Let's illustrate the use of percentiles by means of the following example: of the people calling in for travel information, how long do 95% of them spend on the line with an agent when the length of a telephone call is normally distributed with an expected value of four minutes and a standard deviation of half a minute? The 95% percentile of the call-conclusion time is $4 + 0.5 \times z_{0.95} = 4.82$ minutes. In other words, 95% of the calls are concluded within 4.82 minutes.

In inventory control, the normal distribution is often used to model the demand distribution. Occasionally, one finds oneself asking experts in the field for educated guesses with regard to the expected value and standard deviation of the normal demand distribution. But even such experts often have difficulty with the concept of standard deviation. They can, however, provide an estimate (educated guess) for the average demand, and they can usually even estimate the threshold level of demand that will only be exceeded with a 5% chance, say. Let's say you receive an estimated value of 75 for this threshold, against an estimated value of 50 for the average level of demand. From this, you can immediately derive what the expected value μ and the standard deviation σ of the normal demand distribution are. Obviously, the expected value μ is 50. The standard deviation σ follows from the relationship $x_p = \mu + \sigma z_p$ with $x_p = 75$ and $z_p = 1.6449$. This gives $\sigma = 15.2$. The same idea of estimating μ and σ through an indirect approach may be useful in

financial analysis. Let the random variable X represent the price of a stock next year. Suppose that an investor expresses his/her belief in the future stock price by assessing that there is a 25% probability of a stock price being below \$80 and a 25% probability of a stock price being above \$120. Estimates for the expected value μ and standard deviation σ of the stock price next year are then obtained from the equations $(80 - \mu)/\sigma = z_{0.25}$ and $(120 - \mu)/\sigma = z_{0.75}$, where $z_{0.25} = -0.67449$ and $z_{0.75} = 0.67449$. This leads to $\mu = 100$ and $\sigma = 5.45$.

Problem 1 The weight of a pack of soap powder is normally distributed with a mean of $\mu = 1000$ grams and a standard deviation of $\sigma = 5$ grams. What is the percentage of packs whose weight falls 10 grams below the average weight of 1000 grams?

Problem 2 Radars detect flying objects by measuring the power reflected from them. The reflected power of an aircraft can be modelled as a normally distributed random variable with expected value μ and standard deviation σ . An aircraft will be correctly identified by the radar if its reflected power is larger than $\mu + \frac{1}{4}\sigma$. What is the probability that an aircraft will be correctly identified?

Problem 3 The annual rainfall in Amsterdam is normally distributed with an expected value of 799.5 mm and a standard deviation of 121.4 mm. Over many years, what is the proportion of years that the annual rainfall in Amsterdam is no more than 550 mm?

Problem 4 The cholesterol level for an adult male of a specific racial group is normally distributed with an expected value of 5.2 mmol/L and a standard deviation of 0.65 mmol/L. Which cholesterol level is exceeded by 5% of the population?

Problem 5 Gestation periods of humans are normally distributed with an expected value of 266 days and a standard deviation of 16 days. What is the percentage of births that are more than 20 days overdue?

4 Concept of probability density

The most simple example of a continuous random variable is the random choice of a number from the interval $(0,1)$. The probability that the randomly chosen number will take on a pre-specified value is zero. It makes only sense to speak of the probability of the randomly chosen number falling in a given subinterval of $(0,1)$. This probability is equal to the length of that subinterval. For example, if a dart is thrown at random to the interval

$(0,1)$, the probability of the dart hitting exactly the point 0.25 is zero, but the probability of the dart landing somewhere in the interval between 0.2 and 0.3 is 0.1 (assuming that the dart has an infinitely thin point). No matter how small Δx is, any subinterval of the length Δx has probability Δx of containing the point at which the dart will land. You might say that the probability mass associated with the landing point of the dart is smeared out over the interval $(0,1)$ in such a way that the density is the same everywhere.² For the random variable X denoting the point at which the dart will land, we have that $P(X \leq a) = a$ for $0 \leq a \leq 1$ can be represented as $P(X \leq a) = \int_0^a f(x)dx$ with the density function $f(x)$ identically equal to 1 on the interval $(0,1)$. In order to introduce the concept of probability density within a general framework, it is instructive to consider the following example.

Example 1 A stick of unit length is broken at random into two pieces. What is the probability that the ratio of the length of the shorter piece to that of the longer piece is smaller than a for any $0 < a < 1$?

Solution. The sample space of the chance experiment is the interval $(0,1)$, where the outcome $\omega = u$ means that the point at which the stick is broken is a distance u from the beginning of the stick. Let the random variable X denote the ratio of length of the shorter piece to that of the longer piece of the broken stick. Denote by $F(a)$ the probability that the random variable X takes on a value smaller than or equal to a . Fix $0 < a < 1$. The probability that the ratio of the length of the shorter piece to that of the longer piece is smaller than or equal to a is nothing else than the probability that a random number from the interval $(0,1)$ falls either in $(\frac{1}{1+a}, 1)$ or in $(0, 1 - \frac{1}{1+a})$. The latter probability is equal to $2(1 - \frac{1}{1+a}) = \frac{2a}{1+a}$. Thus,

$$F(a) = \begin{cases} 0 & \text{for } a \leq 0 \\ \frac{2a}{1+a} & \text{for } 0 < a < 1 \\ 1 & \text{for } a \geq 1. \end{cases}$$

Denoting by $f(a) = \frac{2}{(1+a)^2}$ the derivative of $F(a)$ for $0 < a < 1$ and letting $f(a) = 0$ outside the interval $(0,1)$, it follows that

$$F(a) = \int_{-\infty}^a f(x)dx \quad \text{for all } a.$$

²The dart problem is an one-dimensional geometric probability problem. The Appendix discusses several geometric probability problems for which the sample space is a bounded region in a two-dimensional space. Also, in these problems probability mass is not assigned to individual points but to subregions. A characteristic feature of geometric probability problems is that the probability mass assigned to any subregion is the ratio of the area of the subregion and the area of the whole region.

In this specific example, we have a continuous analog of the cumulative probability $F(a)$ in the discrete case: if X is a discrete random variable having possible values a_1, a_2, \dots with associated probabilities p_1, p_2, \dots , then the probability that X takes on a value smaller than or equal to a is represented by

$$F(a) = \sum_{i:a_i \leq a} p_i \quad \text{for all } a.$$

We now come to the definition of a continuous random variable. Let X be a random variable that is defined on a sample space with probability measure P . It is assumed that the set of possible values of X is uncountable and is a finite or infinite interval on the real line.

Definition 1 *The random variable X is said to be (absolutely) continuously distributed if a function $f(x)$ exists such that*

$$P(X \leq a) = \int_{-\infty}^a f(x) dx \quad \text{for each real number } a,$$

where the function $f(x)$ satisfies

$$f(x) \geq 0 \quad \text{for all } x \quad \text{and} \quad \int_{-\infty}^{\infty} f(x) dx = 1.$$

The notation $P(X \leq a)$ stands for the probability that is assigned by the probability measure P to the set of all outcomes ω for which $X(\omega) \leq a$. The function $P(X \leq x)$ is called the (*cumulative*) *probability distribution function* of the random variable X , and the function $f(x)$ is called the *probability density function* of X . Unlike the probability distribution function of a discrete random variable, the probability distribution function of a continuous random variable has no jumps and is continuous everywhere.

Beginning students often misinterpret the non-negative number $f(a)$ as a probability, namely as the probability $P(X = a)$. This interpretation is wrong. Nevertheless, it is possible to give an intuitive interpretation of the non-negative number $f(a)$ in terms of probabilities. Before doing this, we present another example of a continuous random variable with a probability density function.

Example 2 Suppose that the lifetime X of a battery has the cumulative probability distribution function

$$P(X \leq x) = \begin{cases} 0 & \text{for } x < 0, \\ \frac{1}{4}x^2 & \text{for } 0 \leq x \leq 2, \\ 1 & \text{for } x > 2. \end{cases}$$

The probability distribution function $P(X \leq x)$ is continuous and is differentiable at each point x except for the two points $x = 0$ and $x = 2$. Also, the derivative is integrable. We can now conclude from the fundamental theorem of integral calculus that the random variable X has a probability density function. This probability density function is obtained by differentiation of the probability distribution function and is given by

$$f(x) = \begin{cases} \frac{1}{2}x & \text{for } 0 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

In each of the finite number of points x at which $P(X \leq x)$ has no derivative, it does not matter what value we give $f(x)$. These values do not affect $\int_{-\infty}^a f(x) dx$. Usually, we give $f(x)$ the value 0 at any of these exceptional points.

4.1 Interpretation of the probability density

The use of the word ‘density’ originated with the analogy to the distribution of matter in space. In physics, any finite volume, no matter how small, has a positive mass, but there is no mass at a single point. A similar description applies to continuous random variables. To make this more precise, we first express $P(a < X \leq b)$ in terms of the density $f(x)$ for any constants a and b with $a < b$. Noting that the event $\{X \leq b\}$ is the union of the two disjoint events $\{a < X \leq b\}$ and $\{X \leq a\}$, it follows that $P(X \leq b) = P(a < X \leq b) + P(X \leq a)$. Hence,

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= \int_{-\infty}^b f(x) dx - \int_{-\infty}^a f(x) dx \quad \text{for } a < b \end{aligned}$$

and so

$$P(a < X \leq b) = \int_a^b f(x) dx \quad \text{for } a < b.$$

In other words, the area under the graph of $f(x)$ between the points a and b gives the probability $P(a < X \leq b)$. Next, we find that

$$\begin{aligned} P(X = a) &= \lim_{n \rightarrow \infty} P\left(a - \frac{1}{n} < X \leq a\right) \\ &= \lim_{n \rightarrow \infty} \int_{a - \frac{1}{n}}^a f(x) dx = \int_a^a f(x) dx, \end{aligned}$$

using the continuity property of the probability measure P stating that $\lim_{n \rightarrow \infty} P(A_n) = P(\lim_{n \rightarrow \infty} A_n)$ for any non-increasing sequence of events A_n . Hence, we arrive at the conclusion

$$P(X = a) = 0 \quad \text{for each real number } a.$$

This formally proves that, for a continuous random variable X , it makes no sense to speak of the probability that the random variable X will take on a *prespecified* value. This probability is always zero. It only makes sense to speak of the probability that the continuous random variable X will take on a value in some interval. Incidentally, since $P(X = c) = 0$ for any number c , the probability that X takes on a value in an interval with endpoints a and b is not influenced by whether or not the endpoints are included. In other words, for any two real numbers a and b with $a < b$, we have

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b).$$

The fact that the area under the graph of $f(x)$ can be interpreted as a probability leads to an intuitive interpretation of $f(a)$. Let a be a given continuity point of $f(x)$. Consider now a small interval of length Δa around the point a , say $[a - \frac{1}{2}\Delta a, a + \frac{1}{2}\Delta a]$. Since

$$P(a - \frac{1}{2}\Delta a \leq X \leq a + \frac{1}{2}\Delta a) = \int_{a - \frac{1}{2}\Delta a}^{a + \frac{1}{2}\Delta a} f(x) dx$$

and

$$\int_{a - \frac{1}{2}\Delta a}^{a + \frac{1}{2}\Delta a} f(x) dx \approx f(a)\Delta a \quad \text{for } \Delta a \text{ small,}$$

we obtain that

$$P(a - \frac{1}{2}\Delta a \leq X \leq a + \frac{1}{2}\Delta a) \approx f(a)\Delta a \quad \text{for } \Delta a \text{ small.}$$

In other words, the probability of random variable X taking on a value in a *small* interval around point a is approximately equal to $f(a)\Delta a$ when Δa is the length of the interval. You see that the number $f(a)$ itself is *not* a probability, but it is a relative measure for the likelihood that random variable X will take on a value in the immediate neighborhood of point a . Stated differently, the probability density function $f(x)$ expresses how densely the probability mass of random variable X is smeared out in the neighborhood of point x . Hence, the name of density function. The probability density function provides the most useful description of a continuous random variable. The graph of the density function provides a good picture of the likelihood of the possible values of the random variable.

4.2 Verification of a probability density

In general, how can we verify whether a random variable X has a probability density? In concrete situations, we first determine the cumulative distribution function $F(a) = P(X \leq a)$ and next we verify whether $F(a)$ can be written in the form of $F(a) = \int_{-\infty}^a f(x) dx$. A sufficient condition is that $F(x)$ is continuous at every point x and is differentiable except for a finite number of points x . The following two examples are given in illustration of this point.

Example 3 Let the random variable be defined by $X = U^2$, where U is a random number from the interval $(0,1)$. What is the probability density function of X ?

Solution. The approach is to derive first the cumulative probability distribution function of X . Using the fact that the probability of the random U falling in an subinterval of length u equals u for any $0 < u < 1$, we find

$$P(X \leq x) = P(U^2 \leq x) = P(U \leq \sqrt{x}) = \sqrt{x} \quad \text{for } 0 < x < 1.$$

Differentiating $P(X \leq x)$ gives that the random variable X has the density function

$$f(x) = \begin{cases} \frac{1}{2\sqrt{x}} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Example 4 A point is picked at random in the inside of a circular disk with radius r . Let the random variable X denote the distance from the center of the disk to this point. Does the random variable X have a probability density function and, if so, what is its form?

Solution. To answer the question, we first define a sample space with an appropriate probability measure P for the random experiment. The sample space is taken as the set of all points (x, y) in the two-dimensional plane with $x^2 + y^2 \leq r^2$. Since the point inside the circular disk is chosen at random, we assign to each well-defined subset A of the sample space the probability

$$P(A) = \frac{\text{area of region } A}{\pi r^2}.$$

The cumulative probability distribution function $P(X \leq x)$ is easily calculated. The event $X \leq a$ occurs if and only if the randomly picked point falls in the disk of radius a with area πa^2 . Therefore

$$P(X \leq a) = \frac{\pi a^2}{\pi r^2} = \frac{a^2}{r^2} \quad \text{for } 0 \leq a \leq r.$$

Obviously, $P(X \leq a) = 0$ for $a < 0$ and $P(X \leq a) = 1$ for $a > r$. Since the expression for $P(X \leq x)$ is continuous at every point x and is differentiable except at the points $x = 0$ and $x = a$, it follows that X has a probability density function which is given by

$$f(x) = \begin{cases} \frac{2x}{r^2} & \text{for } 0 < x < r, \\ 0 & \text{otherwise.} \end{cases}$$

All of the foregoing examples follow the same procedure in order to find the probability density function of a random variable X . The cumulative probability distribution function $P(X \leq x)$ is determined first and this distribution function is differentiated next.

Problem 6 (a) Let the random variable X be defined by $X = V^2$, where V is a number chosen at random from the interval $(-10, 10)$. What is the probability density of X ?

(b) Let U be a number chosen at random from the interval $(0, 1)$. What is the probability density function of the random variable $X = \sqrt{U}$?

(c) Let U be a number chosen at random from the interval $(0, 1)$. What is the probability density function of the random variable $X = -\ln(U)$?

Problem 7 A point Q is chosen at random inside the unit square. The unit square consists of the points (x, y) with $0 \leq x \leq 1$ and $0 \leq y \leq 1$. Let the random variable V denote the sum of the coordinates of the point Q . Use a little geometry to calculate the probabilities $P(V \leq 0.5)$ and $P(V \leq 1.5)$. In general, what is the probability $P(V \leq v)$ for $0 \leq v \leq 2$? What is the probability density function of V ?

Problem 8 A point Q is chosen at random inside the unit square. Let the random variable W denote the product of the coordinates of the point Q . Calculate first the probability $P(W \leq 0.5)$. Next calculate the probability $P(W \leq w)$ for $0 \leq w \leq 1$. What is the probability density function of W ?

Problem 9 The number X is chosen at random between 0 and 1. Determine the probability density function of each of the random variables $V = X/(1 - X)$ and $W = X(1 - X)$.

Problem 10 A stick of unit length is broken at random into two pieces. the random variable X represent the length of the shorter piece. What is the probability density of X ? Also, use the probability distribution function of X to give an alternative derivation of the probability density of the random variable $X/(1 - X)$ from Example 1.

Problem 11 Suppose you decide to take a ride on the ferris wheel at an amusement park. The ferris wheel has a diameter of 30 meters. After several

turns, the ferris wheel suddenly stops due to a power outage. What random variable determines your height above the ground when the ferris wheel stops? What is the probability that this height is not more than 22.5 meters? And the probability of no more than 7.5 meters? What is the probability density function of the random variable governing the height above the ground?

5 Expected value

The expected value of a continuous random variable X with probability density function $f(x)$ is defined by

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx$$

provided that the integral $\int_{-\infty}^{\infty} |x|f(x) dx$ is finite (the latter integral is always well-defined by the non-negativity of the integrand). It is then said that $E(X)$ exists. In the case that X is a non-negative random variable, the integral $\int_0^{\infty} xf(x) dx$ is always defined when allowing ∞ as possible value. In this case it is convenient to say that $E(X) = \int_0^{\infty} xf(x) dx$ always exists. The definition of expected value in the continuous case parallels the definition $E(X) = \sum x_i p(x_i)$ for a discrete random variable X with x_1, x_2, \dots as possible values and $p(x_i) = P(X = x_i)$. For dx small, the quantity $f(x) dx$ in a discrete approximation of the continuous case corresponds with $p(x)$ in the discrete case. The summation becomes an integral when dx approaches zero. Results for discrete random variables are typically expressed as sums. The corresponding results for continuous random variables are expressed as integrals.

As an illustration, consider the random variable X from Example 4. The expected value of the distance X equals

$$E(X) = \int_0^r x \frac{2x}{r^2} dx = \frac{2}{3} \frac{x^3}{r^2} \Big|_0^r = \frac{2}{3} r.$$

Example 1 (continued) A stick of unit length is broken at random into two pieces. What is the expected value of the ratio of the length of the shorter piece to that of the longer piece? What is the expected value of the ratio of the length of the longer piece to that of the shorter piece?

Solution. Denote by the random variable V the ratio of the length of the shorter piece to that of the longer piece and by the random variable W the ratio of the length of the longer piece to that of the shorter piece. In Example

1 we showed that V has the probability distribution function $F(v) = \frac{2v}{v+1}$ with probability density $f(v) = \frac{2}{(v+1)^2}$ for $0 < v < 1$. Hence,

$$\begin{aligned} E(V) &= \int_0^1 v \frac{2}{(v+1)^2} dv = 2 \int_0^1 \frac{1}{v+1} dv - 2 \int_0^1 \frac{1}{(v+1)^2} dv \\ &= 2\ln(v+1) \Big|_0^1 + 2 \frac{1}{v+1} \Big|_0^1 = 2\ln(2) - 1. \end{aligned}$$

In order to calculate $E(W)$, note that $W = \frac{1}{V}$. Hence, $P(W \leq w) = P(V \geq \frac{1}{w})$ for $w > 1$. This leads to $P(W \leq w) = 1 - \frac{2}{w+1}$ for $w > 1$. Thus, the random variable W has the probability density function $\frac{2}{(w+1)^2}$ for $w > 1$ and so

$$E(W) = \int_1^\infty w \frac{2}{(w+1)^2} dw = 2\ln(w+1) \Big|_1^\infty + 2 \frac{1}{w+1} \Big|_1^\infty = \infty.$$

A little calculus was enough to find a result that otherwise is difficult to obtain from a simulation study. In a simulation study the convergence of the average of the simulated values for the random variable W is too slow in order to conclude that $E(W) = \infty$ (analogously, by summing the terms of the infinite series $\sum_{n=1}^\infty \frac{1}{n}$ on your computer or hand calculator, you will not "discover" that the value of the harmonic series is infinity large.

Example 6 Suppose that the random variable X is $N(0, \sigma^2)$ distributed. What is the probability density function of the random variable $V = |X|$? what is the expected value of V ?

Solution. Using the fact that X/σ is $N(0, 1)$ distributed, we have

$$\begin{aligned} P(V \leq v) &= P(-v \leq X \leq v) = P\left(\frac{-v}{\sigma} \leq \frac{X}{\sigma} \leq \frac{v}{\sigma}\right) \\ &= \Phi\left(\frac{v}{\sigma}\right) - \Phi\left(\frac{-v}{\sigma}\right) \quad \text{for } v > 0. \end{aligned}$$

Differentiation gives that V has the probability density function

$$\frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}v^2/\sigma^2} \quad \text{for } v > 0.$$

The expected value of V is calculated as

$$\begin{aligned} E(V) &= \int_0^\infty v \frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}v^2/\sigma^2} dv = \frac{-2\sigma^2}{\sigma\sqrt{2\pi}} \int_0^\infty de^{-\frac{1}{2}v^2/\sigma^2} \\ &= \frac{-2\sigma}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2/\sigma^2} \Big|_0^\infty = \frac{\sigma\sqrt{2}}{\sqrt{\pi}}. \end{aligned}$$

An application of this example to the one-dimensional random walk will be given later.

Remark: The expected values in Example 5 can also be calculated by using the so-called substitution rule. This rule states that for any function $g(x)$, the expected value of the induced random variable $g(X)$ can be calculated from

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x) dx$$

when $f(x)$ is the probability density of the original random variable X . In Example 5 the random variable V represents the ratio of the length of the shorter piece to that of the longer piece when a stick is broken at random into two pieces. In the example we calculated $E(V)$ by determining first the density function of V and then applying the definition of $E(V)$. However, the substitution rule provides a simpler way to calculate $E(V)$ by using the fact that $V = g(U)$ when U is a random number from the interval $(0,1)$ and the function $g(u)$ is defined by $g(u) = u/(1-u)$ for $0 < u \leq \frac{1}{2}$ and $g(u) = (1-u)/u$ for $\frac{1}{2} < u < 1$. This gives

$$\begin{aligned} E(V) &= \int_0^{1/2} \frac{u}{1-u} du + \int_{1/2}^1 \frac{1-u}{u} du = 2 \int_{1/2}^1 \frac{1-u}{u} du \\ &= 2\ln(u) - 2u \Big|_{1/2}^1 = 2\ln(2) - 1. \end{aligned}$$

Problem 12 The javelin thrower Big John throws the javelin more than x meters with probability $P(x)$, where $P(x) = 1$ for $0 \leq x < 50$, $P(x) = \frac{1,200-(x-50)^2}{1,200}$ for $50 \leq x < 80$, $P(x) = \frac{(90-x)^2}{400}$ for $80 \leq x < 90$, and $P(x) = 0$ for $x \geq 90$. What is the expected value of the distance thrown in his next shot?

Problem 13 A point is chosen at random inside the unit circle. What is the probability density of the distance of the point to the center of the circle. What is the expected value of the distance?

Problem 14 A point is chosen at random inside a triangle with height h and base of length b . Let the random variable X denote the perpendicular distance from the point to the base. What is the probability density of X ? What is the expected value of X ? *Hint:* The probability $P(X > x)$ can be expressed as the ratio of the areas of two triangles.

Problem 15 A point is chosen at random inside an equilateral triangle with unit side. Let the random variable X denote the perpendicular distance to the nearest side of the triangle. Verify that X has the probability density

function $f(x) = 4\sqrt{3}(1 - 2\sqrt{3}x)$ for $0 < x < \frac{1}{2\sqrt{3}}$. What is the expected value of X ?

Problem 16 A point is chosen at random inside the unit circle. Let the random variable V denote the absolute value of the x -coordinate of the point. What is the expected value of V ?

Problem 17 You spin a game board spinner in a round box whose circumference is marked with a scale from 0 to 1. When the spinner comes to rest, it points to a random number between 0 and 1. After your first spin, you have decide whether to spin the spinner for a second time. Your payoff is \$1000 times the total score of your spins as long as this score does not exceed 1; otherwise, your payoff is zero. What strategy maximizes the expected value of your payoff?

5.1 Variance

A measure of the spread of the random variable X around its expected value $\mu = E(X)$ is the variance. The *variance* of the random variable X is defined by as the expected value of the random variable $(X - \mu)^2$ and is denoted by $\text{var}(X)$. That is,

$$\text{var}(X) = E[(X - \mu)^2].$$

Another common notation for the variance of X is $\sigma^2(X)$. Why not use $E(|X - \mu|)$ as the measuring gauge for the spread? The answer is simply that it is much easier to work with $E[(X - \mu)^2]$ than with $E(|X - \mu|)$. The variance $\sigma^2(X) = E[(X - \mu)^2]$ can also be seen in the famed Chebyshev's inequality:

$$P(|X - \mu| \geq a) \leq \text{var}(X)/a^2$$

for every constant $a > 0$. This inequality is generally applicable regardless of what form the distribution of X takes on. For specific distributions the bound can be considerably sharpened in most cases. For example, if the random variable X has a normal distribution with expected value μ and standard deviation σ , then $P(|X - \mu| \geq a) \leq 2\sigma \approx 0.05$ while Chebyshev's inequality gives the bound 0.25.

The variance of X does not have the same dimension as the values of the random variable X . For example, if the values of X are expressed in dollars, then the dimension of $\sigma^2(X)$ will be equal to (dollars)². A measure for the spread that has the same dimension as the random variable X is the *standard deviation*. It is defined as

$$\sigma(X) = \sqrt{\text{var}(X)}.$$

Let X be a continuous random variable with density $f(x)$. For any given function $g(x)$, the expected value of the random variable $g(X)$ can be calculated from

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x) dx$$

provided that the integral exists. Letting $\mu = E(X)$, the variance of the random variable X , which is defined by $\text{var}(X) = E[(X - \mu)^2]$, can be calculated from

$$\text{var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx.$$

The variance of X is usually calculated by using the formula $\text{var}(X) = E(X^2) - \mu^2$, leading to

$$\text{var}(X) = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2.$$

The variance of X does not have the same dimension as the values of X . Therefore, one often uses standard deviation of the random variable X , which is defined by

$$\sigma(X) = \sqrt{\text{var}(X)}.$$

As an illustration, we calculate the variance of the random variable X from Example 4:

$$\text{var}(X) = \int_0^r x^2 \frac{2x}{r^2} dx - \left(\frac{2}{3}r\right)^2 = \frac{2r^2}{4} - \frac{4}{9}r^2 = \frac{1}{18}r^2.$$

The standard deviation of the distance from the randomly selected point inside the circle to the origin is $\sigma(X) = \sqrt{\text{var}(X)} = 0.2357r$.

Example 5 Let the random variable X represent a number drawn at random from the interval (a, b) . What are the expected value and the variance of X ?

Solution. The probability that X will fall into a subinterval of width w is $\frac{w}{b-a}$. Hence, $P(X \leq x) = \frac{x-a}{b-a}$ for $a \leq x \leq b$ and so the density function $f(x)$ of X is given by $f(x) = \frac{1}{b-a}$ for $a < x < b$ and $f(x) = 0$ otherwise. This gives

$$E(X) = \int_a^b x \frac{1}{b-a} dx = \frac{1}{2} \frac{x^2}{b-a} \Big|_a^b = \frac{1}{2} \frac{b^2 - a^2}{b-a} = \frac{a+b}{2},$$

using the fact that $b^2 - a^2 = (b-a)(b+a)$. Similarly, we find

$$E(X^2) = \int_a^b bx^2 \frac{1}{b-a} dx = \frac{1}{3} \frac{x^3}{b-a} \Big|_a^b = \frac{1}{3} \frac{b^3 - a^3}{b-a} = \frac{a^2 + ab + b^2}{3},$$

using the fact that $b^3 - a^3 = (b^2 + ab + a^2)(b - a)$. Thus

$$\text{var}(X) = \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

Problem 18 A point Q is chosen at random inside a sphere with radius r . What are the expected value and the standard deviation of the distance from the center of the sphere to the point Q ?

Problem 19 The lifetime (in months) of a battery is a random variable X satisfying $P(X \leq x) = 0$ for $x < 5$, $P(X \leq x) = [(x - 5)^3 + 2(x - 5)]/12$ for $5 \leq x \leq 7$ and $P(X \leq x) = 1$ for $x > 7$. What are the expected value and the standard deviation of X ?

Problem 20 In a roll of textile flaws occur occasionally. Let the random variable X denote the distance between successive flaws. The random variable X satisfies $P(X \leq x) = 0$ for $x \leq 5$, $P(X \leq x) = \frac{1}{25}(x - 5)^2$ for $5 < x < 10$ and $P(X \leq x) = 1$ for $x > 10$. What are the expected value and the standard deviation of X ?

Problem 21 Consider Problem 11 again. Calculate the expected value and standard deviation of the height above the ground when the ferris wheel stops.

Problem 22 Let X be a continuous random variable with probability density $f(x)$ and finite expected value $E(X)$.

- (a) What constant c minimizes $E[(X - c)^2]$ and what is the minimal value of $E[(X - c)^2]$?
- (b) Prove that $E(|X - c|)$ is minimal if c is chosen equal to the median of X , where the *median* of X is any value m for which $P(X \leq m) = P(X \geq m) = \frac{1}{2}$.³

Problem 23 Suppose that the continuous random variable X has the probability density function $f(x) = (\alpha/\beta)(\beta/x)^{\alpha+1}$ for $x > \beta$ and $f(x) = 0$ for $x \leq \beta$ for given values of the parameters $\alpha > 0$ and $\beta > 0$. This density is called the *Pareto* density, which provides a useful probability model for income distributions among others.

- (a) Calculate the expected value, the variance and the median of X .

³The median is sometimes a better measure for a random variable than the expected value. For example, this is the case for income distributions.

- (b) Assume that the annual income of employed measured in thousands of dollars in a given country follows a Pareto distribution with $\alpha = 2.25$ and $\beta = 2.5$. What percentage of the working population has an annual income of between 25 and 40 thousand dollars?

Problem 24 A stick of unit length is broken at random into two pieces. Let the random variable X represent the length of the shorter piece. What is the median of the random variable $(1 - X)/X$?

5.2 Drunkard's walk⁴

The drunkard's walk is named for the drunkard exiting a pub who takes a step to the right with a probability of $\frac{1}{2}$ or a step to the left with a probability of $\frac{1}{2}$. Each successive step is executed independently of the others. The following question arises: what is the expected distance back to the point of origin after the drunkard has taken many steps? This question seemingly falls into the category of pure entertainment, but in actuality, nothing could be further from the truth. The drunkard's walk (often named the *random walk*) has many important applications in physics, chemistry, astronomy and biology. These applications usually consider two- or three-dimensional representations of the drunkard's walk. The biologist looks at the transporting of molecules through cell walls. The physicist looks at the electrical resistance of a fixed particle. The chemist looks for explanations for the speed of chemical reactions. The climate specialist looks for evidence of global warming, etc. The model of the drunkard's walk is extremely useful for this type of research. We first look at the model of the drunkard walking along a straight line. Plotting the path of the drunkard's walk along a straight line is much the same as tracing the random walk of the fair-coin toss. Imagine a drunkard at his point of origin. His steps are of unit length, and there is a probability of $\frac{1}{2}$ that in any given step he will go to the right and a probability of $\frac{1}{2}$ that he will go to the left. The drunkard has no memory, i.e., the directions of the man's successive steps are independent of one another. Define the random variable D_n as

D_n = the drunkard's distance from his starting point after n steps.

Then, it holds that

$$E(D_n) \approx \sqrt{\frac{2}{\pi}n},$$

where the symbol \approx stands for 'is approximately equal to'. This approximation requires that n is sufficiently large (the approximation is already

⁴This paragraph contains more advanced material.

quite accurate from $n = 10$ onwards. Using the central limit theorem and the results of Example 6, the approximation for $E(D_n)$ can be rather easily derived. The central theorem is the queen of probability theory. A first rudimentary version of the central limit theorem was already published by the English mathematician Abraham de Moivre around 1730.⁵ In its general form, the central limit theorem states that the sum of n independent random variables each having the same probability distribution with expected value μ and standard deviation σ has approximately a normal distribution with expected value $n\mu$ and standard deviation $\sigma\sqrt{n}$ provided that n large. How large n should be depend on the specific distribution of the individual random variables. Taking this famous for granted, it is not difficult to derive the approximation for $E(D_n)$ in the one-dimensional drunkard's walk. The idea is simple: the position of the drunkard after n steps can be represented as the sum of n independent random variables each having the same distribution with expected value $\mu = 0$ and standard deviation $\sigma = 1$. To prove this, let the random variable X_i be equal to 1 if the drunkard goes to the right in his i th step and be equal to -1 if the drunkard goes the left in his i th step. Then, for any $n \geq 1$,

$$\text{the position of the drunkard after } n \text{ steps} = X_1 + X_2 + \dots + X_n.$$

Moreover, the individual random variables X_1, X_2, \dots, X_n are independent of each other and have each the same probability distribution

$$P(X_i = 1) = P(X_i = -1) = \frac{1}{2}.$$

The expected value μ and the standard deviation σ of this probability distribution are easily calculated as

$$\mu = 1 \times \frac{1}{2} + (-1) \times \frac{1}{2} = 0, \quad \sigma = \sqrt{1^2 \times \frac{1}{2} + (-1)^2 \times \frac{1}{2} - 0^2} = 1.$$

Thus, by the central limit theorem, we can conclude that the position of the drunkard after n steps is approximately $N(0, \sqrt{n})$ distributed for n sufficiently large. Using the fact that the random variable D_n is equal to the absolute value of the position of the drunkard after n steps ($D_n = |X_1 + X_2 + \dots + X_n|$), an application of the results in Example 6 gives the approximative square-root formula for $E(D_n)$.

⁵The French-born Abraham de Moivre (1667-1754) lived most of his life in England. The protestant De Moivre left France in 1688 to escape religious persecution. He was a good friend of Isaac Newton, who admired his mathematical genius, and supported himself by calculating odds for gamblers and insurers and by giving private lessons to students.

In Problem 25 below you are asked to derive an approximate formula for $E(D_n)$ for the drunkard's walk in dimensions two and three, while in Problem 26 we ask you to apply the formula for dimension three to estimate the average number of years it takes a photon to travel from the sun's core to its surface. In the drunkard's walk in higher dimensions, a drunkard (particle) starts at the origin and in each step the particle travels a unit distance in a randomly chosen direction. The direction of each successive step is determined independently of the others. Denoting again by the random variable D_n the distance of the drunkard from its starting point after n steps, the following deep result about the probability distribution of D_n is stated without proof. For the drunkard's walk in dimension two,

$$P(D_n \leq u) \approx 2 \int_0^{u/\sqrt{n}} e^{-x^2} x \, dx \quad \text{for } u > 0,$$

while for the drunkard's walk in dimension three

$$P(D_n \leq u) \approx \frac{3\sqrt{6}}{\sqrt{\pi}} \int_0^{u/\sqrt{n}} e^{-\frac{3}{2}x^2} x^2 \, dx \quad \text{for } u > 0,$$

provided that n is sufficiently large.

Problem 25 Use partial integration to derive from the approximate distribution of D_n that

$$E(D_n) \approx \frac{1}{2} \sqrt{\pi n}$$

for the drunkard's walk in dimension 2 and that

$$E(D_n) \approx \sqrt{\frac{8n}{3\pi}}$$

for the drunkard's walk in dimension 3.

Problem 26 A photon starting at the center of the sun has a countless number of collisions on its way to the sun's surface. The distance traveled by a photon between two collisions can be measured as 6×10^{-6} mm. The sun's radius measures 70,000 km. A photon travels at the light speed of 300,000 km per second. Use the result in Problem 21 for the drunkard's model for dimension three to estimate the average number of years it takes a photon to reach the surface of the sun starting from the center of the sun (once a photon has reached the surface of the sun, it takes the photon only 8 minutes to travel to the earth; the distance from the sun to the earth is 149,600,000 km).

5.3 Appendix: Geometric probability

In this appendix we consider geometric probability problems. These problems are a special case of probability problems having a continuous rather than a discrete sample space. The probability mass is not assigned to individual points but to subsets of the sample space according to the uniform distribution. That is, assuming that the sample space is a bounded region in a two-dimensional space, the probability mass assigned to any subregion is the area of the subregion divided by the area of the whole region.

Example A.1 (continued). Consider the experiment in which a random point in a circle with radius R is chosen by a blindfolded person throwing a dart at a dartboard. How do we calculate the probability of the dart hitting the bull's-eye?

Solution. The sample space of this experiment consists of the set of pairs of real numbers (x, y) where $x^2 + y^2 \leq R^2$. This sample space is uncountable. The assumption of the dart hitting the dartboard at a random point is translated by assigning the probability

$$P(A) = \frac{\text{the area of the region } A}{\pi R^2}$$

to each subset A of the sample space. If the bull's-eye of the dartboard has radius b , the probability of the dart hitting the bull's-eye is $\pi b^2 / (\pi R^2) = b^2 / R^2$. The following observation is made. The probability that the dart will hit a *prespecified* point is zero. It makes only sense to speak of the probability of hitting a given region of the dartboard. This observation expresses a fundamental difference between a probability model with a finite or countably infinite sample space and a probability model with an uncountable sample space.

Example A.2 A floor is ruled with equally spaced parallel lines a distance D apart. A needle of length L is dropped at random on the floor. It is assumed that $L \leq D$. What is the probability that the needle will intersect one of the lines? This problem is known as Buffon's needle problem. ⁶

Solution. This geometric probability problem can be translated into the picking of a random point in a certain region. Let y be the distance from the center of the needle to the closest line and let x be the angle at which the needle falls, where x is measured against a line parallel to the lines on the

⁶An experimental demonstration of Buffon's needle problem can be found in the educational software package ORSTAT2000 that can be downloaded free of charge from <http://staff.feweb.vu.nl/tijms>.

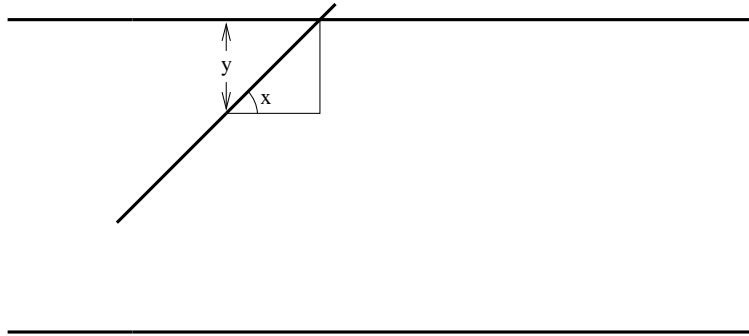


Figure 4: The landing of Buffon's needle.

floor; see Figure 7.1. The sample space of the experiment can be taken as the rectangle R consisting of the points (x, y) with $0 \leq x \leq \pi$ and $0 \leq y \leq \frac{1}{2}D$. The needle will land on a line only if the hypotenuse of the right-angled triangle in Figure 7.1 is less than half of the length L of the needle. That is, we get an intersection only if $\frac{y}{\sin(x)} < \frac{1}{2}L$. Thus, the probability that the needle will intersect one of the lines equals the probability that a point (x, y) chosen at random in the rectangle R satisfies $y < \frac{1}{2}L \sin(x)$. In other words, the area under the curve $y = \frac{1}{2}L \sin(x)$ divided by the total area of the rectangle R gives the probability of an intersection. This ratio is

$$\frac{\int_0^\pi \frac{1}{2}L \sin(x) dx}{\frac{1}{2}\pi D} = \frac{-L \cos(x)}{\pi D} \Big|_0^\pi$$

and so

$$P(\text{needle intersects one of the lines}) = \frac{2L}{\pi D}.$$

Problem A.1 The game of franc-carreau was a popular game in eighteenth-century France. In this game, a coin is tossed on a chessboard. The player wins if the coin does not fall on one of the lines of the board. Suppose now that a round coin with a diameter of d is blindly tossed on a large table. The surface of the table is divided into squares whose sides measure a in length, such that $a > d$. Define an appropriate probability space and calculate the probability of the coin falling entirely within the confines of a square. *Hint:* consider the position of the coin's middle point.

Problem A.2 Two people have agreed to meet at the train station between 12.00 and 1:00 P.M. Independently of one another, each person is to appear at a completely random moment between the hours of 12.00 and 1.00. What is the probability that the two persons will meet within 10 minutes of one another?

Problem A.3 The numbers B and C are chosen at random between 0 and 1, independently of each other. What is the probability that the quadratic equation $x^2 + Bx + C = 0$ has real roots? What is this probability if the numbers B and C are chosen at random between -1 and 1 ? In general, what is this probability if the numbers B and C are chosen at random between $-q$ and q for any $q > 0$?

Problem A.4 A dart is thrown at random on a rectangular board. The board measures 20 cm by 50 cm. A hit occurs if the dart lands within 5 cm of any of the four corner points of the board. What is the probability of a hit?

Problem A.5 A point is chosen at random inside a triangle with height h and base of length b . What is the probability that the perpendicular distance from the point to the base is larger than d ? What is the probability that the random point and the base of the triangle will form a triangle with an obtuse angle when the original triangle is equilateral?

UITWERKINGEN VAN DE OPGAVEN KANSREKENING MET VOORTGEZETTE INTEGRAALREKENING

Problem 1 Laat de stochastische variabele X het gewicht in grammen van een pak zeepoeder aangeven, de bijbehorende kansverdeling is $N(1000, 5^2)$. Uit

$$\begin{aligned} P(X \leq 990) &= P\left(\frac{X - 1000}{5} \leq \frac{990 - 1000}{5}\right) \\ &= P(Z \leq -2) = \Phi(-2) = 0.0228 \end{aligned}$$

volgt dat het gevraagde percentage gelijk is aan 2.28%.

Problem 2 Laat X de gereflecteerde kracht van een vliegtuig aangeven met bijbehorende kansverdeling $N(\mu, \sigma^2)$. De kans dat een vliegtuig correct wordt geïdentificeerd is gelijk aan

$$\begin{aligned} P\left(X > \mu + \frac{1}{4}\sigma\right) &= P\left(\frac{X - \mu}{\sigma} > \frac{\mu + \frac{1}{4}\sigma - \mu}{\sigma}\right) \\ &= P(Z > 1/4) = \Phi(-1/4) = 0.4013. \end{aligned}$$

Problem 3 De fractie van de jaren dat er minder dan 550 mm regen valt is gelijk aan

$$P\left(Z \leq \frac{550 - 799.5}{121.4}\right) = \Phi(-2.05519) = 0.0199.$$

Problem 4 Uit $z_{0.95} = 1.6449$ volgt dat het gezochte cholesterolniveau gelijk is aan

$$x_{0.95} = 5.2 + 0.65 \times 1.6449 = 6.2692 \text{ mmol/L.}$$

Problem 5 Uit

$$P\left(Z > \frac{20}{16}\right) = 1 - \Phi(5/4) = 0.1056.$$

volgt dat 10.56% van de kinderen meer dan 20 dagen te laat wordt geboren.

Problem 6 (a) De cumulatieve kansverdelingsfunctie van $X = V^2$ is

$$P(X \leq x) = P(V^2 \leq x) = P(-\sqrt{x} \leq V \leq \sqrt{x}) = \frac{2\sqrt{x}}{20} = \frac{\sqrt{x}}{10}.$$

voor $0 < x < 100$. Hierin is gebruik gemaakt van de kansdichtheid van V , die gelijk is aan $f(x) = 1/20$ voor $-10 < x < 10$ en $f(x) = 0$ anders. Differentiëren van $P(X \leq x)$ geeft dat de kansdichtheidsfunctie van X gegeven wordt door

$$g(x) = \begin{cases} \frac{1}{20\sqrt{x}} & \text{als } 0 < x < 100 \\ 0 & \text{anders.} \end{cases}$$

(b) Voor de stochast $X = \sqrt{U}$ geldt

$$P(X \leq x) = P(\sqrt{U} \leq x) = P(U \leq x^2) = x^2 \quad \text{voor } 0 \leq x \leq 1.$$

De stochastische variabele X heeft de dichtheidsfunctie $f(x) = 2x$ voor $0 < x < 1$ en $f(x) = 0$ elders.

(c) Voor de stochastische variabele $X = -\ln(U)$ geldt

$$\begin{aligned} P(X \leq x) &= P(-\ln(U) \leq x) = P(\ln(U) \geq -x) \\ &= P(U \geq e^{-x}) = 1 - P(U \leq e^{-x}) \quad \text{voor } x > 0, \end{aligned}$$

waarbij de laatste gelijkheid het feit gebruikt dat $P(U < u) = P(U \leq u)$ voor een continue stochast U . Aangezien $P(U \leq u) = u$ voor $0 \leq u \leq 1$, volgt dat

$$P(X \leq x) = 1 - e^{-x}, \quad x > 0.$$

Uiteraard, $P(X \leq x) = 0$ voor $x \leq 0$. Door differentiatie van $P(X \leq x)$ volgt dat X een kansdichtheid $f(x)$ heeft met $f(x) = e^{-x}$ voor $x > 0$ and $f(x) = 0$ voor $x \leq 0$.

Problem 7 De punten Q waarvoor de som van de coördinaten hoogstens 0.5 is, liggen onder de lijn $y = 0.5 - x$. Deze lijn vormt een driehoek in het eenheidsvierkant. Wanneer we de oppervlakte van de driehoek delen door de oppervlakte van de eenheidsvierkant krijgen we de kans

$$P(V \leq 0.5) = \frac{0.25}{1} = 0.25.$$

Vervolgens berekenen we op dezelfde wijze $P(V > 1.5) = 0.125$, waaruit volgt dat $P(V \leq 1.5) = 1 - P(V > 1.5) = 0.875$. Dit kunnen we generaliseren tot

$$P(V \leq v) = \begin{cases} \frac{1}{2}v^2 & \text{als } 0 \leq v \leq 1 \\ 1 - \frac{1}{2}(2 - v)^2 & \text{als } 1 < v \leq 2. \end{cases}$$

Differentiëren van de verdelingsfunctie geeft de kansdichtheid

$$f(v) = \begin{cases} v & \text{als } 0 < v \leq 1 \\ 2 - v & \text{als } 1 < v < 2 \\ 0 & \text{anders.} \end{cases}$$

Problem 8 De punten (x, y) die voldoen aan $W \leq 0.5$ zijn de punten in het gebied $\{(x, y) : 0 \leq x \leq 0.5, 0 \leq y \leq 1\}$ en de punten onder de curve $y = \frac{1}{2x}$ met x tussen 0.5 en 1 Dit geeft

$$P(W \leq 0.5) = 0.5 + \int_{0.5}^1 \frac{0.5}{x} dx = 0.5 - 0.5 \ln(0.5) = 0.8466$$

Algemener, voor $0 \leq w \leq 1$ vinden we

$$P(W \leq w) = w + \int_w^1 \frac{w}{x} dx = w - w \ln(w).$$

Differentiëren van de verdelingsfunctie geeft de kansdichtheid

$$f(w) = \begin{cases} -\ln(w) & \text{als } 0 < w < 1 \\ 0 & \text{anders.} \end{cases}$$

Problem 9 De kansverdelingsfuncties zijn

$$P(V \leq v) = P\left(\frac{X}{1-X} \leq v\right) = P\left(X \leq \frac{v}{1+v}\right) = \frac{v}{1+v}$$

voor $v \geq 0$ en

$$\begin{aligned} P(W \leq w) &= P(X(1-X) \leq w) = P(X^2 - X + w \geq 0) \\ &= 1 - P(X^2 - X + w < 0) \\ &= 1 - P\left(\frac{1 - \sqrt{1-4w}}{2} < X < \frac{1 + \sqrt{1-4w}}{2}\right) \\ &= 1 - \sqrt{1-4w} \end{aligned}$$

voor $0 \leq w \leq 1/4$. De kansdichtheidsfuncties van V en W zijn

$$f(v) = \begin{cases} \frac{1}{(1+v)^2} & \text{als } v \geq 0 \\ 0 & \text{anders} \end{cases},$$

respectievelijk

$$g(w) = \begin{cases} \frac{2}{\sqrt{1-4w}} & \text{als } 0 < w < 1/4 \\ 0 & \text{anders} \end{cases}.$$

Problem 10 De lengte van het kortste stuk is maximaal 0.5. Voor $0 \leq x \leq 0.5$ geldt dat de kans $P(X \leq x)$ gelijk is aan de kans dat een random gekozen

getal uit $(0,1)$ in het deelinterval $(0, x)$ of het deelinterval $(1 - x, 1)$ valt. Dit betekent dat $P(X \leq x) = 2x$ met kansdichtheid $f(x) = 2$ voor $0 < x < 0.5$ en $f(x) = 0$ elders. Laat $Y = X/(1 - X)$, dan

$$P(Y \leq y) = P\left(X \leq \frac{y}{1+y}\right) = \frac{2y}{1+y},$$

voor $0 \leq y \leq 1$. De kansdichtheid van Y is $g(y) = \frac{2}{(1+y)^2}$ voor $0 < y < 1$ en $g(y) = 0$ elders.

Problem 11 De hoek tussen de verticale as en het lijnsegment van het middelpunt van het rad naar het punt waar jij je bevindt, wordt aangegeven met de stochastische variabele X (zie figuur 5). Deze is maximaal op het laagste punt, namelijk $X = \pi$ rad ($=180^\circ$). De hoek X is beschouwen als een random getal tussen 0 en π radialen, dus $P(X \leq x) = x/\pi$ en $f(x) = 1/\pi$ voor $0 < x < \pi$. De straal van het rad en dus ook de afstand van de as tot de grond is 15 meter, de hoogte waarop je zit wordt dus gegeven door $Y = 15 + 15 \cos(X)$. Dit geeft de kansverdelingsfunctie

$$\begin{aligned} P(Y \leq y) &= P(15 + 15 \cos(X) \leq y) \\ &= P\left(X \geq \arccos\left(\frac{y}{15} - 1\right)\right) \\ &= 1 - \frac{\arccos\left(\frac{y}{15} - 1\right)}{\pi}, \end{aligned}$$

voor $0 \leq y \leq 30$. Hieruit volgen $P(Y \leq 22.5) = 2/3$ en $P(Y \leq 7.5) = 1/3$. De afgeleide van $\arccos(z)$ is $-\frac{1}{\sqrt{1-z^2}}$, dus de kansdichtheidsfunctie van Y is

$$g(y) = \begin{cases} \frac{1}{15\pi\sqrt{1-(y/15-1)^2}} & \text{als } 0 < y < 30 \\ 0 & \text{anders.} \end{cases}$$

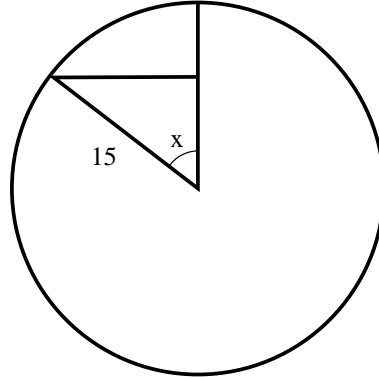
Problem 12 De kansverdelingsfunctie is

$$P(X \leq x) = \begin{cases} \frac{(x-50)^2}{1200} & \text{als } 50 \leq x < 80 \\ 1 - \frac{(90-x)^2}{400} & \text{als } 80 \leq x < 90. \end{cases}$$

Nadat $P(X \leq x)$ voor beide gevallen is gedifferentieerd kan de verwachting als volgt worden berekend.

$$\begin{aligned} E(X) &= \int_0^\infty x f(x) dx \\ &= \int_{50}^{80} x \frac{x-50}{600} + \int_{80}^{90} x \frac{90-x}{200} \\ &= \frac{x^3 - 75x^2}{1800} \Big|_{50}^{80} + \frac{135x^2 - x^3}{200} \Big|_{80}^{90} = 73\frac{1}{3} \end{aligned}$$

Figuur 5: Het rad met een straal van 15 meter.



Problem 13 De eenheids­cirkel heeft straal 1 en oppervlakte π . De gebeurtenis dat de afstand X tot het middelpunt hoogstens x is voor $0 \leq x \leq 1$ treedt alleen op als het random gekozen punt in de cirkel met middelpunt $(0, 0)$ en straal x valt. De kansverdeling van X wordt dus gegeven door

$$P(X \leq x) = \frac{\text{Oppervlakte cirkel met straal } x}{\text{Oppervlakte eenheids­cirkel}} = \frac{\pi x^2}{\pi} = x^2$$

voor $0 \leq x \leq 1$. De kansdichtheidsfunctie is $f(x) = 2x$ voor $0 < x < 1$ en $f(x) = 0$ elders. De verwachting van X wordt gegeven door

$$E(X) = \int_0^{\infty} x f(x) dx = \int_0^1 2x^2 dx = \frac{2}{3} x^3 \Big|_0^1 = \frac{2}{3}.$$

Problem 14 De punten waarvoor de afstand X tot de basis groter dan x is vormen een driehoek met hoogte $h - x$ en basis $\frac{h-x}{h}b$ (zie ook figuur 10). Hieruit volgt de kans

$$P(X \leq x) = 1 - P(X > x) = 1 - \frac{1/2 \times (h-x)^2/h \times b}{1/2 \times h \times b} = 1 - \left(\frac{h-x}{h}\right)^2$$

voor $0 \leq x \leq h$. De kansdichtheidsfunctie is $f(x) = \frac{2(h-x)}{h^2}$ voor $0 < x < h$ en $f(x) = 0$ elders. De verwachting van X wordt gegeven door

$$E(X) = \int_0^{\infty} x f(x) dx = \int_0^h x \frac{2(h-x)}{h^2} dx = \frac{2}{h^2} \left(\frac{1}{2} h x^2 - \frac{1}{3} x^3 \right) \Big|_0^h = \frac{1}{3} h.$$

Problem 15 Laat de stochastische variabele X de afstand aangeven van het random gekozen punt tot de zijde die het dichtst bij ligt. Voor elke vaste x met $0 < x < \frac{1}{2\sqrt{3}}$ geldt dat de punten die meer dan een afstand x van elk van de zijden van de gelijkzijdige driehoek liggen een ingesloten gelijkzijdige driehoek vormen waarvan de zijden lengte $1 - 2x\sqrt{3}$ hebben. Dit betekent dat $P(X > x)$ gelijk is aan de oppervlakte $\frac{1}{4}\sqrt{3}(1 - 2x\sqrt{3})^2$ van de ingesloten gelijkzijdige driehoek gedeeld door de oppervlakte $\frac{1}{4}\sqrt{3}$ van de oorspronkelijke gelijkzijdige driehoek. Dit geeft $P(X \leq x) = 1 - (1 - 2x\sqrt{3})^2$ voor $0 < x < \frac{1}{2\sqrt{3}}$. Differentieren geeft $f(x) = 4\sqrt{3}(1 - 2\sqrt{3}x)$ voor $0 < x < \frac{1}{2\sqrt{3}}$ en $E(X) = \frac{1}{6\sqrt{3}}$.

Problem 16 De eenheidscirkel bestaat uit de componenten $y = -\sqrt{1 - x^2}$ en $y = \sqrt{1 - x^2}$, voor $-1 \leq x \leq 1$, en heeft als oppervlakte π . De absolute waarde van de x -coördinaat geven we aan met de stochastische variabele Z . Het gedeelte van de eenheidscirkel dat tussen de lijnen $y = -z$ en $y = z$ ligt kan worden opgedeeld in twee driehoeken met basis $2\sqrt{1 - z^2}$ en hoogte z en twee cirkelsectoren met hoek $2 \arcsin z$ (zie figuur 6). Hieruit volgt voor Z de kansverdelingsfunctie

$$P(Z \leq z) = \frac{2z\sqrt{1 - z^2}}{\pi} + \frac{2 \arcsin z}{\pi} = \frac{2}{\pi} \left(z\sqrt{1 - z^2} + \arcsin(z) \right)$$

en de kansdichtheidsfunctie

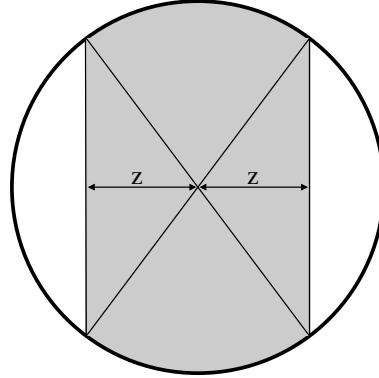
$$f(z) = \frac{2}{\pi} \left(\sqrt{1 - z^2} + \frac{-2z^2}{2\sqrt{1 - z^2}} + \frac{1}{\sqrt{1 - z^2}} \right) = \frac{4}{\pi} \sqrt{1 - z^2},$$

voor $0 < z < 1$. De verwachting is gelijk aan

$$E(Z) = \frac{4}{\pi} \int_0^1 z\sqrt{1 - z^2} dz = -\frac{4}{3\pi} (1 - z^2)^{3/2} \Big|_0^1 = \frac{4}{3\pi}.$$

Problem 17. Laat de stochastische variabele X_a de eindscore aangeven wanneer de spinner voor een tweede maal gedraaid gaat worden nadat de spinner bij de eerste draai een score van a gegeven had (elke draai van de spinner levert een random getal uit interval $(0,1)$ op). In feite heeft de stochastische variabele X_a zowel een discrete als een continue component: de kans dat de stochast X_a de waarde 0 aanneemt is gelijk aan de kans a dat de tweede draai een score tussen $1 - a$ en 1 geeft zodat het totaal de waarde 1 overschrijdt, terwijl de resterende kansmassa van de stochast X_a volgens continu en gelijkmatig over het interval $(a, 1)$ is uitgesmeerd. Intuitief zal

Figuur 6: Het gebied waarvoor $Z \leq z$ is gearceerd.



duidelijk zijn dat $E(X_a)$ berekend wordt als

$$E(X_a) = \int_0^{1-a} (a+x)dx + \int_{1-a}^1 0 \cdot dx = a(1-a) + \frac{1}{2}(1-a)^2.$$

Dit betekent dat als de eerste draai van de spinner een score a geeft, je doorgaat voor een tweede draai alleen als $a(1-a) + \frac{1}{2}(1-a)^2 > a$. De oplossing van de vergelijking $a(1-a) + \frac{1}{2}(1-a)a^2 = a$ wordt gegeven door $a^* = \sqrt{2} - 1$. De optimale strategie is dus om te stoppen na de eerste draai als deze draai een score groter dan $\sqrt{2} - 1 = 0.4142$ geeft en anders door te gaan. Zonder nadere toelichting vermelden we dat de verwachtingswaarde van de uitbetaling onder de optimale strategie gegeven wordt door

$$\$1000 \times \left[\int_0^{\sqrt{2}-1} \left(a(1-a) + \frac{1}{2}a^2 \right) da + \int_{\sqrt{2}-1}^1 a da \right] = \$609.48.$$

Opmerking. Het is interessant deze oplossing te vergelijken met die van het discrete rad met de getallen $1, 2, \dots, 1000$ erop. In dat geval is $E(X_a)$ gelijk aan

$$\frac{1}{1000} \sum_{k=1}^{1000-a} (a+k) = \frac{1}{1000} \left(a(1000-a) + \frac{1}{2}(1000-a)(1000-a+1) \right).$$

Deze uitdrukking is groter dan a zolang a kleiner dan of gelijk aan 414 is. Voor het discrete rad met de getallen $1, 2, \dots, 1000$ erop is het dus optimaal om na de eerste draai te stoppen als deze een score groter dan 414 geeft (voor het discrete rad met de getallen $1, 2, \dots, 100$ erop is het optimale omslagpunt $a^* = 41$).

Problem 18 De inhoud van een bol met straal r is $\frac{4}{3}\pi r^3$, dus de kans dat de afstand tot het middelpunt hoogstens x is wordt gegeven door

$$P(X \leq x) = \frac{4/3\pi x^3}{4/3\pi r^3} = \left(\frac{x}{r}\right)^3$$

voor $0 \leq x \leq r$. De kansdichtheidsfunctie is $f(x) = 3x^2/r^3$ en de verwachting is gelijk aan

$$E(X) = \int_0^r x \frac{3x^2}{r^3} dx = \frac{3x^4}{4r^3} \Big|_0^r = \frac{3}{4}r.$$

Verder geldt

$$E(X^2) = \int_0^r x^2 \frac{3x^2}{r^3} dx = \frac{3x^5}{5r^3} \Big|_0^r = \frac{3}{5}r.$$

De standaarddeviatie wordt gegeven door

$$\sigma(X) = \sqrt{\text{var}(X)} = \sqrt{\frac{3}{5}r^2 - \left(\frac{3}{4}r\right)^2} = \sqrt{\frac{3}{80}r^2} = 0.1936r.$$

Problem 19. De kansdichtheidsfunctie van de continue stochastische variabele X is

$$f(x) = \frac{3(x-5)^2 + 2}{12} = \frac{3x^2 - 30x + 77}{12}$$

voor $0 < x < r$. Hieruit volgt dat

$$E(X) = \int_5^7 x \frac{3x^2 - 30x + 77}{12} dx = \frac{1}{16}x^4 - \frac{5}{6}x^3 + \frac{77}{24}x^2 \Big|_5^7 = 6\frac{1}{3}$$

$$E(X^2) = \int_5^7 x^2 \frac{3x^2 - 30x + 77}{12} dx = \frac{1}{20}x^5 - \frac{5}{8}x^4 + \frac{77}{36}x^3 \Big|_5^7 = 40\frac{17}{45}.$$

De standaarddeviatie wordt dus gegeven door

$$\sigma(X) = \sqrt{40\frac{17}{45} - \left(6\frac{1}{3}\right)^2} = \sqrt{1\frac{1}{45}} = 1.0111.$$

Problem 20 De kansdichtheidsfunctie van X is $f(x) = \frac{2}{25}x - \frac{2}{5}$ voor $5 < x < 10$. Hieruit volgt dat

$$E(X) = \int_5^{10} x \left(\frac{2}{25}x - \frac{2}{5}\right) dx = \frac{2}{75}x^3 - \frac{1}{5}x^2 \Big|_5^{10} = 8\frac{1}{3}$$

$$E(X^2) = \int_5^{10} x^2 \left(\frac{2}{25}x - \frac{2}{5}\right) dx = \frac{1}{50}x^4 - \frac{2}{15}x^3 \Big|_5^{10} = 70\frac{5}{6}.$$

De standaarddeviatie wordt dus gegeven door

$$\sigma(X) = \sqrt{70\frac{5}{6} - \left(8\frac{1}{3}\right)^2} = \sqrt{1\frac{7}{18}} = 1.1785.$$

Problem 21 Met behulp van het resultaat van Problem 11 kan de verwachting worden berekend, die uiteraard gelijk is aan 15 meter :

$$\begin{aligned} E(Y) &= \int_0^{30} y \frac{1}{15\pi\sqrt{1 - (y/15 - 1)^2}} dy \\ &= \frac{-15}{\pi} \int_0^{30} \frac{-(y/15 - 1)}{15\sqrt{1 - (y/15 - 1)^2}} dy + \frac{15}{\pi} \int_0^{30} \frac{-1}{15\sqrt{1 - (y/15 - 1)^2}} dy \\ &= \frac{-15}{\pi} \sqrt{1 - \left(\frac{y}{15} - 1\right)^2} \Big|_0^{30} + \frac{15}{\pi} \arcsin\left(\frac{y}{15} - 1\right) \Big|_0^{30} \\ &= 0 + \frac{15}{\pi} \left(\frac{\pi}{2} + \frac{\pi}{2}\right) = 15 \end{aligned}$$

Het tweede moment $E(Y^2) = \int_0^{30} y^2 g(y) dy$ berekenen we met numerieke integratie. Dit geeft de waarde $E(Y^2) = 337.4$ meter, waaruit de waarde $\sigma(Y) = \sqrt{3337.4 - 15^2} = 10.6$ meter volgt voor de standaarddeviatie.

Problem 22(a) We bepalen het minimum van de functie $h(c) = E[(X - c)^2]$.

$$\begin{aligned} h(c) &= E[(X - c)^2] = \int_{-\infty}^{\infty} (x - c)^2 f(x) dx \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - 2c \int_{-\infty}^{\infty} x f(x) dx + c^2 \int_{-\infty}^{\infty} f(x) dx \\ &= E(X^2) - 2cE(X) + c^2 \end{aligned}$$

Wanneer we de afgeleide $h'(c) = -2E(X) + 2c$ gelijkstellen aan nul volgt $c = E(X)$ met $h(E(X)) = E[(X - E(X))^2] = \text{var}(X)$.

(b) De afgeleide van $E(|X - c|)$ wordt gegeven door

$$\begin{aligned} \frac{d}{dc} \int_{-\infty}^{\infty} |x - c| f(x) dx &= \frac{d}{dc} \int_{-\infty}^c (c - x) f(x) dx + \frac{d}{dc} \int_c^{\infty} (x - c) f(x) dx \\ &= cf(c) + \int_{-\infty}^c f(x) - cf(c) - \int_c^{\infty} f(x) dx \\ &= P(X \leq c) - P(X > c) = 2P(X \leq c) - 1 \end{aligned}$$

Deze afgeleide is gelijk aan nul voor $P(X \leq c) = \frac{1}{2}$.

Problem 23(a) Stel dat X Pareto-verdeeld is met parameters $\alpha > 0$ en $\beta > 0$, dan is de verwachting gelijk aan

$$\begin{aligned} E(X) &= \int_{\beta}^{\infty} x(\alpha/\beta)(\beta/x)^{\alpha+1} dx = \int_{\beta}^{\infty} \alpha\beta^{\alpha}x^{-\alpha} dx \\ &= \frac{\alpha\beta^{\alpha}}{1-\alpha}x^{-\alpha+1} \Big|_{\beta}^{\infty} = \frac{\alpha\beta}{\alpha-1}, \end{aligned}$$

wanneer $\alpha > 1$. De verwachting is oneindig ($E(X) = \infty$) voor $0 < \alpha \leq 1$. Verder geldt voor $\alpha > 2$ dat

$$\begin{aligned} E(X^2) &= \int_{\beta}^{\infty} x^2(\alpha/\beta)(\beta/x)^{\alpha+1} dx = \int_{\beta}^{\infty} \alpha\beta^{\alpha}x^{-\alpha+1} dx \\ &= \frac{\alpha\beta^{\alpha}}{2-\alpha}x^{-\alpha+2} \Big|_{\beta}^{\infty} = \frac{\alpha\beta^2}{\alpha-2}, \end{aligned}$$

en $E(X^2) = \infty$ voor $0 < \alpha \leq 2$. De variantie voor $\alpha > 2$ wordt vervolgens gegeven door

$$\text{var}(X) = \frac{\alpha\beta^2}{\alpha-2} - \left(\frac{\alpha\beta}{\alpha-1}\right)^2 = \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}.$$

De kansverdelingsfunctie is

$$P(X \leq x) = \int_{\beta}^x (\alpha/\beta)(\beta/y)^{\alpha+1} dy = 1 - \left(\frac{\beta}{x}\right)^{\alpha}$$

voor $\alpha > 0$, hieruit volgt de mediaan $x = \beta 2^{1/\alpha}$.

(b) Het percentage is 0.37% en volgt uit

$$\begin{aligned} P(25 < X \leq 40) &= P(X \leq 40) - P(X \leq 25) \\ &= (2.5/25)^{2.25} - (2.5/40)^{2.25} = 0.0037. \end{aligned}$$

Problem 24 Met kans 0.5 is het kortste stuk kleiner dan 0.25, hieruit volgt de mediaan $(1-0.25)/0.25 = 3$. Het volgt tevens uit de kansverdelingsfunctie

$$P\left(\frac{1-X}{X} \leq a\right) = P\left(X \geq \frac{1}{1+a}\right) = 1 - \frac{2}{1+a} \quad \text{voor } 0 < a < 0.5$$

Problem 25 In dimensie 2 wordt de kansdichtheid van D_n voor $u > 0$ gegeven door

$$f(u) \approx \frac{2}{\sqrt{n}} e^{-u^2/n} \frac{u}{\sqrt{n}} = \frac{2u}{n} e^{-u^2/n}.$$

De verwachting is

$$\begin{aligned} E(D_n) &\approx \int_0^\infty -u \frac{-2u}{n} e^{-u^2/n} du = -u e^{-u^2/n} \Big|_0^\infty - \int_0^\infty -e^{-u^2/n} du \\ &= 0 - 0 + \sqrt{\pi n} \int_0^\infty \frac{1}{\sqrt{n/2} \sqrt{2\pi}} e^{-\frac{1}{2} \frac{u^2}{n/2}} du \\ &= \frac{1}{2} \sqrt{\pi n} \int_{-\infty}^\infty \frac{1}{\sqrt{n/2} \sqrt{2\pi}} e^{-\frac{1}{2} \frac{u^2}{n/2}} du = \frac{1}{2} \sqrt{\pi n}, \end{aligned}$$

gebruikmakend van het feit dat de laatste integraal de integraal is van een $N(0, \frac{1}{2}n)$ dichtheid en dus gelijk aan 1 is. In dimensie 3 is de kansdichtheidsfunctie

$$f(u) \approx \frac{3\sqrt{6}}{\sqrt{n\pi}} e^{-\frac{3u^2}{2n}} \frac{u^2}{n}$$

voor $u > 0$ en de verwachting

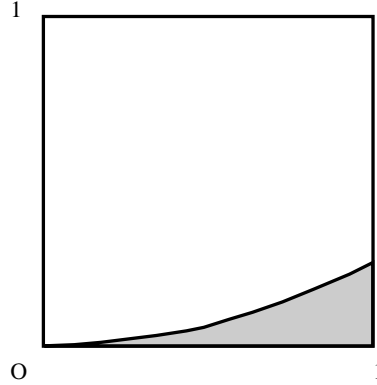
$$\begin{aligned} E(D_n) &\approx \int_0^\infty u \frac{3\sqrt{6}}{\sqrt{n\pi}} e^{-\frac{3u^2}{2n}} \frac{u^2}{n} du = \int_0^\infty \frac{-\sqrt{6}u^2 - 3u}{\sqrt{n\pi}} e^{-\frac{3u^2}{2n}} du \\ &= \frac{-\sqrt{6}u^2}{\sqrt{n\pi}} e^{-\frac{3u^2}{2n}} \Big|_0^\infty - \int_0^\infty \frac{-2\sqrt{6}u}{\sqrt{n\pi}} e^{-\frac{3u^2}{2n}} du \\ &= \frac{-2\sqrt{6}n}{3\sqrt{\pi}} \int_0^\infty \frac{-3u}{n} e^{-\frac{3u^2}{2n}} du = \frac{-2\sqrt{6}n}{3\sqrt{\pi}} e^{-\frac{3u^2}{2n}} \Big|_0^\infty = \sqrt{\frac{8n}{3\pi}}. \end{aligned}$$

Problem 25 Wanneer de eenheid van afstand 6×10^{-6} mm. is, dan is de straal van de zon $7/6 \times 10^{16}$ eenheden en de lichtsnelheid $1/2 \times 10^{17}$ eenheden per seconde. Na enig rekenwerk blijkt dat gemiddelde reistijd ongeveer 10 miljoen jaar is.

Problem A.1 De aanpak is om het middelpunt van de munt op te vatten als een random gekozen punt in het vierkant $\Omega = \{(x, y) : 0 \leq x \leq a, 0 \leq y \leq a\}$ (de uitkomstenruimte van het kansexperiment). Aan elk deelgebied A van de uitkomstenruimte kennen we de kans $P(A) = \text{opp}(A)/\text{opp}(\Omega)$ toe. De gezochte kans is gelijk aan de kans $P(A^*)$ met $A^* = \{(x, y) : \frac{d}{2} \leq x \leq a - \frac{d}{2}, \frac{d}{2} \leq y \leq a - \frac{d}{2}\}$. Dus

$$P(\text{munt valt binnen een vierkant}) = \frac{(a-d)^2}{a^2}.$$

Figuur 7: Het gebied waarvoor $C \leq \frac{1}{4}B^2$ is gearceerd.



Problem A.2 We nemen aan dat de tijd in oneindige precisie gemeten wordt. Vertaal het probleem als het kiezen van een random punt binnen het eenheidsvierkant $\{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$, waarbij de twee componenten x en y corresponderen met de aankomsttijdstippen van de twee personen (tijdseenheid is dus het uur). Het oppervlakte van de strook $\{(x, y) : |x - y| \leq \frac{1}{6}\}$ geeft de kans dat de twee personen binnen tien minuten van elkaar aankomen. De gevraagde kans is dus $1 - \frac{5}{6} \times \frac{5}{6} = 0.3056$.

Problem A.3 De vergelijking heeft reële oplossingen als $B^2 - 4C \geq 0$ en dus $C \leq \frac{1}{4}B^2$. Wanneer B en C onafhankelijk van elkaar random gekozen worden tussen 0 en 1, wordt de kans gegeven door het gearceerde oppervlakte in figuur 7. De uitkomstenruimte van het kansexperiment wordt dus gegeven door het eenheidsvierkant $\{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$. De kansmaat P kent aan elke deelverzameling A van de uitkomstenruimte de kans $P(A) = \text{opp}(A)$ toe. De kans is dan gelijk aan

$$\int_0^1 \frac{1}{4}x^2 dx = \frac{1}{12}.$$

Voor het geval dat B en C onafhankelijk van elkaar random worden gekozen tussen -1 en 1 , is de kans gelijk aan

$$\frac{1}{4} \left(2 + \int_{-1}^1 \frac{1}{4}x^2 dx \right) = 0.5417$$

In het algemene geval, wanneer B en C onafhankelijk van elkaar random worden gekozen uit $(-q, q)$, moet onderscheid worden gemaakt tussen $0 < q < 4$ (zie figuur 8) en $q \geq 4$ (zie figuur 9). De oppervlakte van het gearceerde

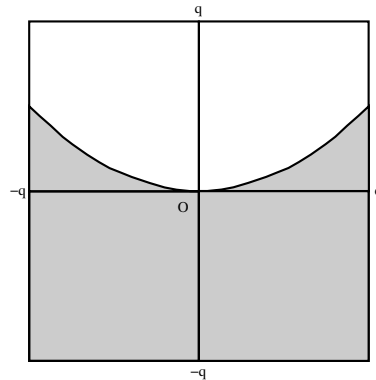
gebied moet worden gedeeld door de oppervlakte $4q^2$ van de rechthoek $R = \{(x, y) : -q \leq x \leq q, -q \leq y \leq q\}$ welke de uitkomstenruimte van het kansexperiment representeert. Voor $0 < q < 4$ geldt voor de gezochte kans de formule

$$\frac{2q^2 + 2 \int_0^q \frac{1}{4}x^2 dx}{4q^2} = \frac{1}{2} + \frac{q}{24} \quad \text{voor } 0 < q < 4$$

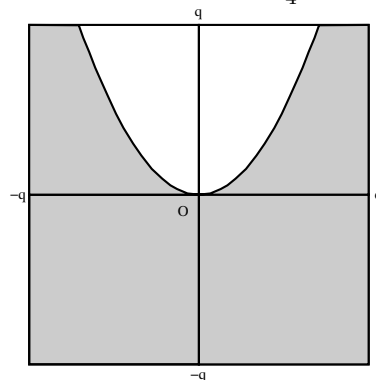
en voor $q \geq 4$ is de formule voor de gezochte kans:

$$\frac{2q^2 + 2[\int_0^{2\sqrt{q}} \frac{1}{4}x^2 dx + q(q - 2\sqrt{q})]}{4q^2} = 1 - \frac{2}{3\sqrt{q}} \quad \text{voor } q \geq 4.$$

Figuur 8: Het gebied waarvoor $C \leq \frac{1}{4}B^2$ is gearceerd ($0 < q < 4$).

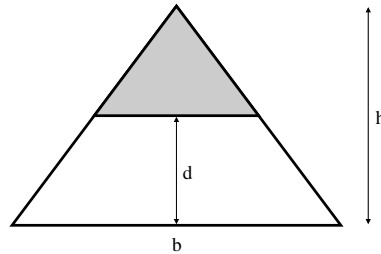


Figuur 9: Het gebied waarvoor $C \leq \frac{1}{4}B^2$ is gearceerd ($q \geq 4$).



Problem A.4 De kans is $\frac{25\pi}{20 \times 50} = \frac{\pi}{40}$.

Figuur 10: Het gebied waarvoor de afstand tot de basis groter is dan d is gearceerd.



Problem A.5 De kans dat in de driehoek met basis b en hoogte h de afstand van het random gekozen punt tot de basis groter is dan d , is gelijk aan $\left(\frac{h-x}{h}\right)^2$ (zie figuur 10 en Problem 14).

Voor de tweede vraag, noem de basis AB en noteer met C het random gekozen punt in de gelijkzijdige driehoek. Thales geeft dat hoek C in driehoek ABC stomp is als punt C binnen de cirkel ligt waarvan AB de middellijn is (zie figuur 11). De oppervlakte van de gelijkzijdige driehoek is $\frac{1}{4}\sqrt{3}b^2$. De oppervlakte van het gearceerde gebied in figuur 11 is $\frac{1}{4}\sqrt{3}(b/2)^2 + \frac{1}{6} \times \pi(b/2)^2 + \frac{1}{4}\sqrt{3}(b/2)^2$. Het quotient van deze twee oppervlaktes geeft de kans dat het random gekozen punt C in het gearceerde gebied valt en dit is ook de kans dat driehoek ABC stomp is. De gezochte kans is dus $\frac{1}{2} + \frac{\pi}{6\sqrt{3}}$.

Figuur 11: Het gebied waarvoor de driehoek stomphoekig zal zijn is gearceerd.

