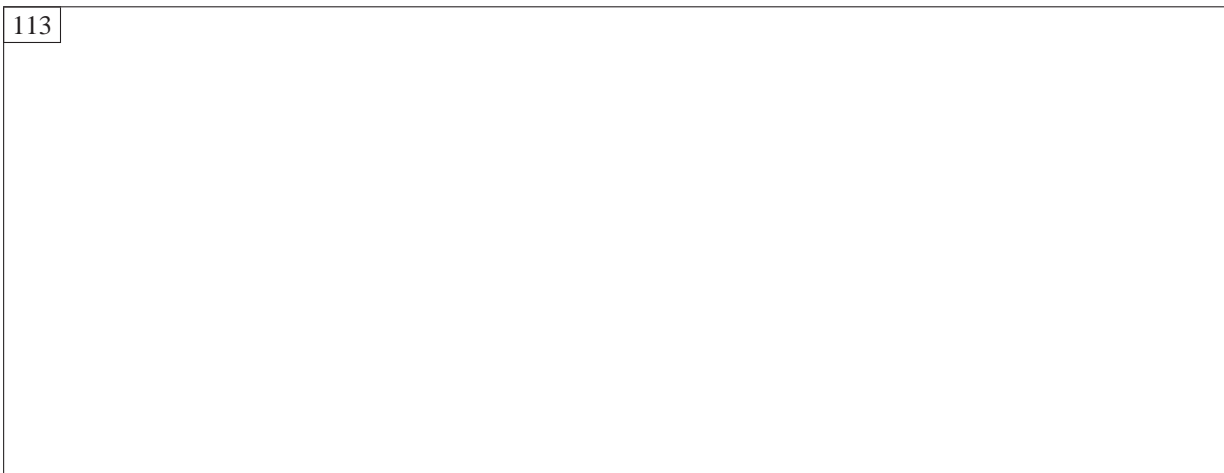


## Handout 15 *Random variables*

**104. Random variables.** From a mathematical point of view there is very little else we can do with general sample spaces over and above what we did on the last handout. The subject becomes more interesting if we associate numbers with the outcomes of the random process. Mathematically, this is nothing other than a function from the sample space to the real numbers  $\mathbb{R}$ . The combination of sample space, probability function and number assignment, is called a **random variable**. The usual letter for a random variable is  $X$ . Some examples:

1. The sample space for a die already consists of numbers, so we can associate with outcome  $n$  the real number  $n$  again. No big deal. But we can also associate with every even number the value  $-1$ , with every odd number the value  $1$ . In this way we use the die for an experiment with only two outcomes.



2. A coin toss yields either “Head” or “Tails.” These are not numbers and we can associate, for example, 0 with “Head” and 1 with “Tails.”
3. The experiment of throwing a coin 10 times has a sample space of 1024 sequences of Heads/Tails. Instead of assigning a different number to each of these, we can instead record just the total number of “Heads” in the sequence. Thus the range of this random variable is  $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ .
4. In Roulette a little ball is thrown on a spinning wheel with 37 slots, numbered from 0 to 36. Players can bet on any number and on many different subsets of numbers. Suppose we put 10 pounds on the number 14. Then we can associate with every outcome from the Roulette wheel our win or loss: With every number except 14 we associate a loss of 10 pounds, that is, the number  $-10$ , and with 14 we associate £ 350, which is what we would get (in addition to our bet) from the croupier if that number came up.

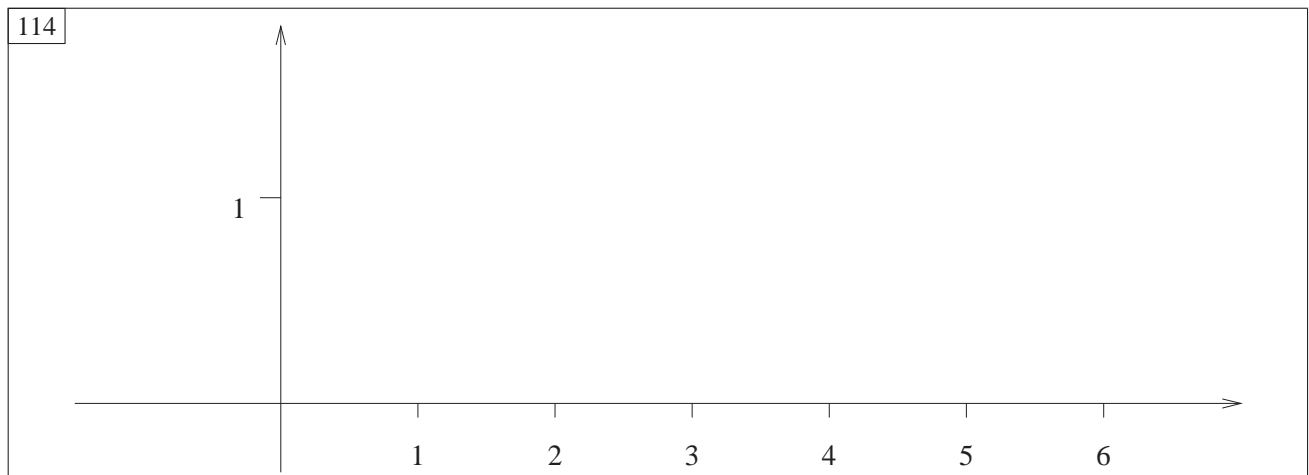
Every number  $r$  in the range of a random variable  $X$  has a probability attached to it, which is the probability of the set of outcomes of the random process which have  $r$  associated with it. One writes  $p(X = r)$  for this probability, although mathematically, what we are talking about is the probability of the event  $X^{-1}(\{r\})$  as  $X$  is just a function from the sample space to  $\mathbb{R}$ .

In fact, one can now forget about the original random process altogether and consider as sample space the real numbers and as probability function the one which is based on the values for individual numbers in the range of  $X$  as we have just defined it. How about other events on  $\mathbb{R}$ ? Obviously, the real numbers (an uncountable set!) has lots and lots of subsets, and we may not really be interested in the probability attached to each of them (well, to tell the truth, it’s not possible in general to attach a probability to all of them anyway). The events that one usually considers are the following:

- $L_r \stackrel{\text{def}}{=} \{x \in \mathbb{R} \mid x < r\}$ , that is, the “half-line” to the left of the value  $r$ . The probability of the random variable ending up in  $L_r$  is written as  $p(X < r)$ .
- $I_{r,s} \stackrel{\text{def}}{=} \{x \in \mathbb{R} \mid r < x < s\}$ , that is, the interval from  $r$  to  $s$ . The probability for this event is written as  $p(r < X < s)$ .

Considering the half-line events, there is one for every real number  $r \in \mathbb{R}$ , and we can *plot* the corresponding probabilities along the real line. Since the events get bigger and bigger as  $r$  grows, this function will grow also as we move along the real axis from left to right. Finally, this function tends to zero as we go towards  $-\infty$  and to 1 as we go towards  $+\infty$ . It is called the **cumulative distribution function** of  $X$ .

Let's construct the one that is associated with the random variable that associates the face value with the outcome of a roll of a fair die:



**105. Expected value.** Once we have associated numbers with the outcome of a random process via a random variable, we can do some calculations. The most basic one is that of the **expected value**. For  $X$  a random variable on a *finite* sample space  $S$ , we define the **expected value** of  $X$  as the *weighted average* of all the values that  $X$  can take:

$$E[X] \stackrel{\text{def}}{=} \sum_{a \in S} X(a) \times p(\{a\})$$

Let's look at the examples from before and see what their expected values are:

1. For the fair die we get  $E[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$ . Note that we can never get 3.5 as an *actual outcome* from a roll of a die, so the word “expected value” has a purely technical meaning in probability theory; it is not the value you should expect to see, but the weighted average of the values you may see.

The second variable associates  $-1$  with even and  $1$  with odd outcomes. For the expected value we get  $E[X] = 1 \times \frac{1}{6} - 1 \times \frac{1}{6} + 1 \times \frac{1}{6} - 1 \times \frac{1}{6} + 1 \times \frac{1}{6} - 1 \times \frac{1}{6} = 0$ , as (I hope) you did indeed expect.

2. For the coin toss we compute  $E[X] = 0 \times \frac{1}{2} + 1 \times \frac{1}{2} = 0.5$
3. This is a long computation but the answer is intuitive:

$$\begin{aligned} E[X] &= 0 \times \frac{\binom{10}{0}}{1024} + 1 \times \frac{\binom{10}{1}}{1024} + 2 \times \frac{\binom{10}{2}}{1024} + 3 \times \frac{\binom{10}{3}}{1024} + 4 \times \frac{\binom{10}{4}}{1024} + \\ &\quad 5 \times \frac{\binom{10}{5}}{1024} + 6 \times \frac{\binom{10}{6}}{1024} + 7 \times \frac{\binom{10}{7}}{1024} + 8 \times \frac{\binom{10}{8}}{1024} + 9 \times \frac{\binom{10}{9}}{1024} + 10 \times \frac{\binom{10}{10}}{1024} \\ &= \frac{1}{1024} \times (0 \times 1 + 1 \times 10 + 2 \times 45 + 3 \times 120 + 4 \times 210 + \\ &\quad 5 \times 252 + 6 \times 210 + 7 \times 120 + 8 \times 45 + 9 \times 10 + 10 \times 1) \\ &= \frac{10}{1024} \times (1 + 10 + 45 + 120 + 210 + 126) \\ &= \frac{5120}{1024} \\ &= 5 \end{aligned}$$

We can interpret this as saying that *on average* we should expect to see five “Heads” in 10 throws of a fair coin.

4. This is also interesting, as it computes our “expected profit” from placing the £ 10 bid on the number 14:

$$E[X] = 36 \times -10 \times \frac{1}{37} + 350 \times \frac{1}{37} = -0.27$$

This says that we can expect to *lose* 27 pennies from such a bet. In gambling lingo, this is called the “house advantage.”

If we have defined some random variable  $X$  then we can create new random variables by applying a further function to the outcome from  $X$ . For example, we can add some fixed value  $c$  to every outcome. The effect on the expected value will be:

$$E[X + c] = E[X] + c$$

that is, the expected value also gets “shifted along.” If we multiply every outcome with some fixed factor  $c$ , then this, too, has a simple effect on the expected value:

$$E[c \times X] = c \times E[X]$$

Taking these two operations together, we get the general affine transformation  $y = ax + b$ . For this we have:

$$E[aX + b] = aE[X] + b$$

**106. Variance and standard deviation.** The expected value gives us a first idea for the “behaviour” of a random variable. However, it can not distinguish between the case where the only possible value of  $X$  is 0 (with probability 1) and the case where both  $-1$  and  $1$  can occur with probability  $\frac{1}{2}$ . In both cases the expected value is 0.

This is where the idea of **variance** comes into play which tells us something about how far away the values of the random variable tend to be from the expected value. Now, the distance could be negative or positive depending whether the outcome is below or above the expected value. In order to measure it correctly we would have to take the absolute value of this, but it turns out that by taking the *square of the distance*, rather than the distance itself, we solve the problem of negative values *as well as* getting much nicer formulas.<sup>20</sup> So we set:

$$\text{Var}[X] \stackrel{\text{def}}{=} \sum_{a \in S} (X(a) - E[X])^2 \times p(\{a\})$$

The disadvantage of taking squares is that we can *not* interpret variance as the “average distance from the mean” in an intuitive sense. We recover some of the connection to actual distance by taking the square root of the variance; the result is called **standard deviation**:

$$\text{StdDev}[X] \stackrel{\text{def}}{=} \sqrt{\text{Var}[X]}$$

However, since we first square, then add up for every outcome of the random process and then take a square root, standard deviation, too, is not what you would call the “average distance from the mean.” It tends to be reasonably close, though.

Example: If the random variable can have values  $-5$  and  $5$  and if the probability for each is  $\frac{1}{2}$ , then the expected value is 0, the variance  $5^2 \times \frac{1}{2} + 5^2 \times \frac{1}{2} = 25$ , and the standard deviation  $5$  — which is what one would hope to see. However, if  $-5$  and  $5$  happen with probability  $\frac{1}{4}$  and the value  $0$  is obtained with probability  $\frac{1}{2}$ , then the expected value is still 0, the variance is  $\frac{25}{2}$  and standard deviation is  $\frac{5}{\sqrt{2}}$ . The “average distance,” on the other hand, would be  $5 \times \frac{1}{4} + 5 \times \frac{1}{4} = \frac{5}{2}$ .

If we apply the simple transformations to  $X$  that we also considered for the expected value, then by simply examining the definitions we find:

$$\begin{aligned} \text{Var}[X + c] &= \text{Var}[X] & \text{and} & & \text{StdDev}[X + c] &= \text{StdDev}[X] \\ \text{Var}[c \times X] &= c^2 \times \text{Var}[X] & \text{and} & & \text{StdDev}[c \times X] &= c \times \text{StdDev}[X] \\ \text{Var}[aX + b] &= a^2 \times \text{Var}[X] & \text{and} & & \text{StdDev}[aX + b] &= a \times \text{StdDev}[X] \end{aligned}$$

I hope these formulas “make sense” to you: adding a constant only “moves the outputs along” which does not change the average distance from the expected value; multiplying by a constant  $c$  “spreads out” the values by factor  $c$ , and this increases (or decreases) the average distance from the expected value accordingly.

**107. Some classical discrete random variables.** Let’s look at some random variables that are useful for building probabilistic models of random (or uncertain) phenomena in computing.

**Bernoulli**<sup>21</sup> This is based on a random process with two outcomes,  $a$  and  $b$ , where the first happens with probability  $p$  (which is some fixed number in  $[0, 1]$ ) and the second happens with probability  $1 - p$  (often abbreviated to  $q$ ). The Bernoulli random variable associates 1 with  $a$  and 0 with  $b$ .

In computing, this may model the probability with which a message gets from sender to receiver without distortion. It is the model that underlies the theory of **error-correcting codes**.

The expected value of this variable is  $1 \times p + 0 \times q = p$ , and the variance is  $p - p^2$ .

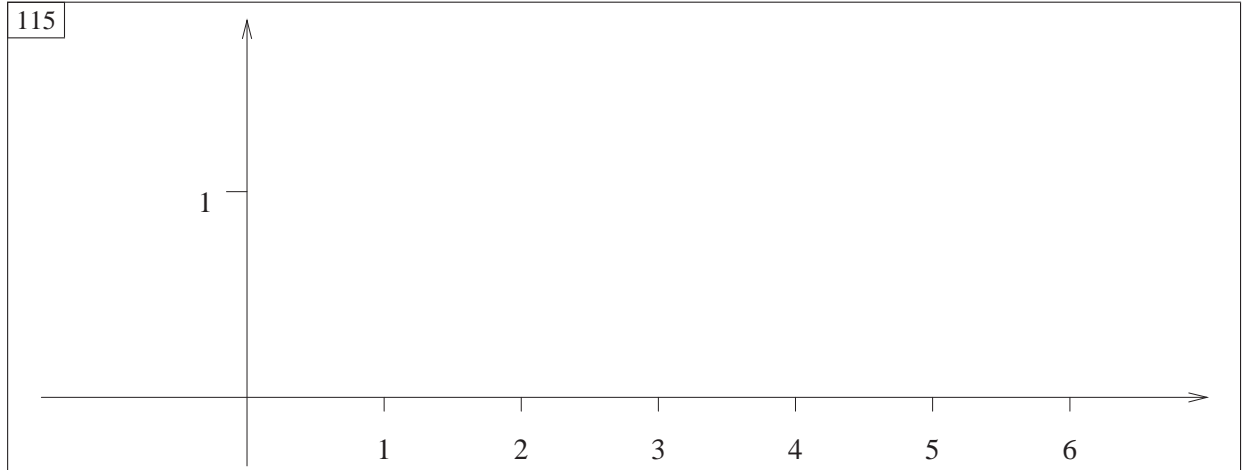
<sup>20</sup>The reason for this phenomenon is related to distance in  $n$ -dimensional space, which we discussed at the beginning of this course.

<sup>21</sup>Named after Jakob Bernoulli, 1654–1705, a member of a Swiss dynasty of mathematicians.

**Binomial** This is based on using  $n$  independent copies of the process that underlies the Bernoulli variable and associates with each of the  $2^n$  many outcomes the number of times the first outcome  $a$  occurred. The possible values of this variable, therefore, are  $0, 1, \dots, n$ . The probability of the random variable assuming value  $i$  is:

$$p(X = i) = \binom{n}{i} \times p^i q^{n-i}$$

Read this as saying that there are  $\binom{n}{i}$ -many ways to choose  $i$  positions in the vector of outcomes of length  $n$ . In each case, the probability of getting this vector is  $p^i$  (for each outcome  $a$ ) times  $q^{n-i}$  (for each outcome  $b$ ). As a picture ( $n = 6, p = \frac{1}{2}$ ):



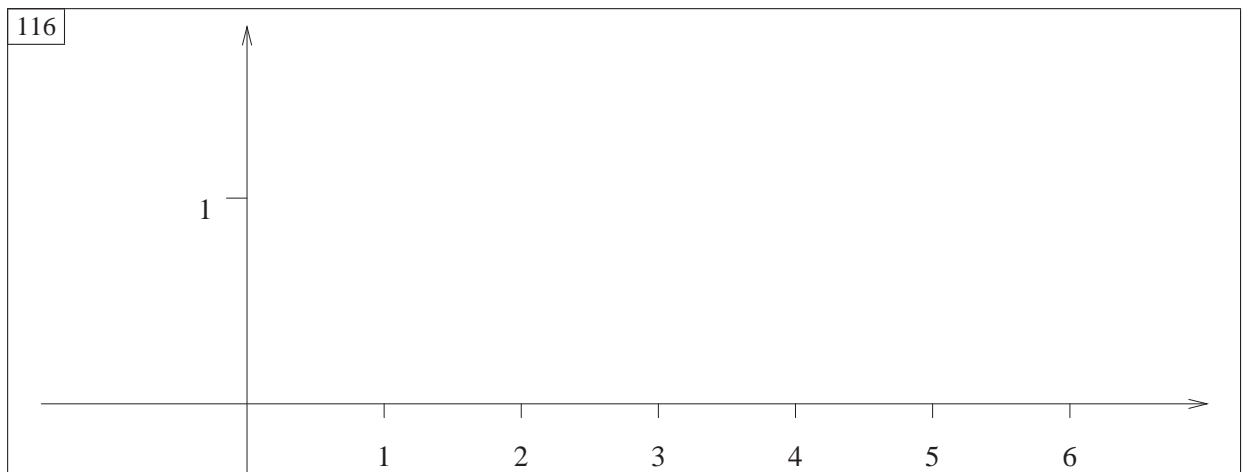
Whether the binomial variable is the right model to use depends very much on the independence of the  $n$  runs of the random process. For example, transmission errors tend to occur in “bursts” which distort several messages in a row, rather than happening randomly and independently for each individual message.

The expected value of the binomial random variable is  $np$ , and the variance is  $npq$ .

**Geometric** In this case we perform the random process with two outcomes *until we obtain outcome a* for the first time. The individual rounds are again assumed to be independent of each other. The random variable associated with this is the number of rounds until  $a$  occurred; it can take values  $1, 2, 3, \dots$ , that is, all natural numbers except zero. The probability that  $i$  rounds are necessary is:

$$p(X = i) = q^{i-1} p$$

Unless  $q = 0$ , the probability is highest for  $i = 1$  and after this falls steadily (in the fashion of a “geometric sequence”). As a picture ( $p = \frac{3}{4}$ ):



An example from computing is given by the “Ethernet” protocol in which the nodes of the network share a common communication medium (the “ether”). For transmission a node sends its message simply into the common medium. At the same time it listens to the traffic on the medium to check whether its message could be heard or whether there was a collision with a message from another node. In the latter case, it “backs off” for a while and tries again. The geometric random variable describes the number of trials required before the message gets through undisturbed (assuming that the traffic on the medium is reasonably constant).

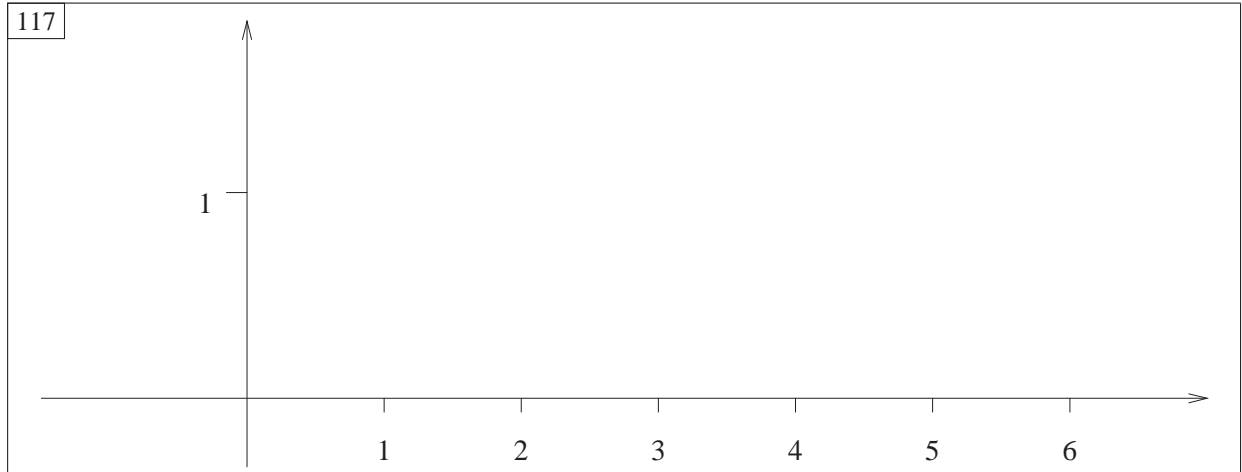
The expected value of the geometric random variable is  $\frac{1}{p}$ , and the variance  $\frac{1-p}{p^2}$ .

**108. Two idealized random variables.** The examples above are all based on a simple random process with two outcomes. If we believe that such a random process exists and can be invoked independently as often as we like, then we are justified also to believe in the binomial and the geometric random variables. In contrast, the next two examples are *not* themselves based on a finite, discrete, and realizable process. Instead, they can be seen as *approximations* to the binomial variable.

**Poisson**<sup>22</sup> The possible outcomes are the natural numbers  $\mathbb{N}$  (including zero) where the probability for each number  $i$  is

$$p(X = i) = \frac{\lambda^i}{i!} \times e^{-\lambda}$$

Here  $\lambda > 0$  is a parameter that can be chosen freely while  $e$  is the real number 2.71828... (known as **Euler's constant** or **Napier's constant**). A picture for  $\lambda = 2$ :



The Poisson variable is a very good approximation to the binomial variable in case the parameter  $n$  of the latter is big and the probability  $p$  small. More precisely, we have for the Poisson variable  $X_\lambda$  with parameter  $\lambda$  and the binomial variable  $X_{n,p}$ , where  $\lambda = np$ :

$$p(X_{n,p} = i) \approx p(X_\lambda = i)$$

or filling in the definitions:

$$\binom{n}{i} \times p^i q^{n-i} \approx \frac{\lambda^i}{i!} \times e^{-\lambda}$$

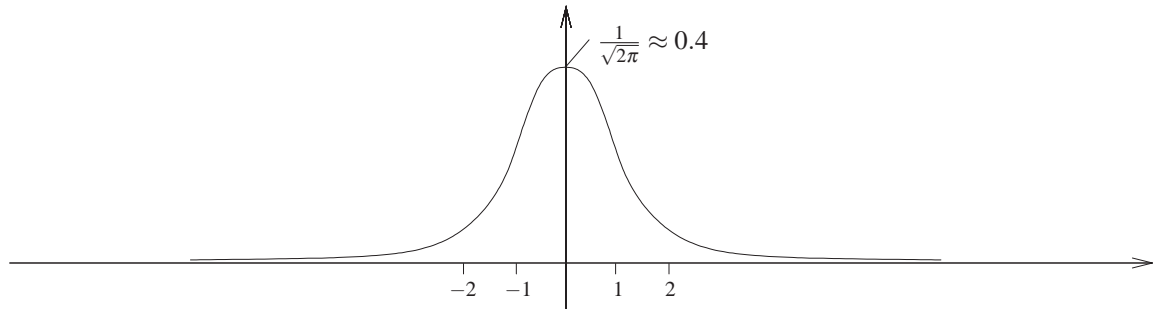
The situation that is so well described by the Poisson variable appears very frequently in the real world. For example, it describes the number of requests received by some agency *in a certain fixed time interval*. This could be the number of operating system requests on a multi-user system, or the number of emails received within a 24-hour period.

The expected value of the Poisson random variable is  $\lambda$ , and this is also the variance. When the variable is used to describe the arrival of requests in a time interval  $v$ , then  $\lambda$  is called the **rate of arrival** (as it is indeed the average number of requests that are received in an interval of length  $v$ ).

**Normal distribution** This is also an approximation to the binomial variable; it is appropriate when the parameter  $n$  of the latter is large and  $p$  *not* very small (or otherwise we should use Poisson). Another difference to Poisson is that it is not discrete but a **continuous** random variable. What is meant by this is that the range of the variable is allowed to be *any real number*. The usual methods of probability models need to be reconsidered for this because the reals form an uncountable set (as we have seen on Handout 10). It is not possible to give positive probabilities to individual real numbers, but only to *intervals* of reals. The way to compute the probability, then, is to take the *area* under a **density function**. A picture makes this clear:

<sup>22</sup>Named after Siméon Denis Poisson, 1781–1840, French mathematician and physicist.

118: Normal distribution  $N(0, 1)$



The density function for the **normal distribution** (also known as the **Gaussian distribution** or the **bell curve**) has two parameters  $\mu$  and  $\sigma$  and is usually denoted by  $N(\mu, \sigma^2)$ . Its equation is:

$$N(\mu, \sigma^2)(x) = \frac{1}{\sqrt{2\pi}\sigma} \times e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The computation of expected value and variance must be redefined for continuous random variables (which is done using integration); for the normal distribution we get exactly  $\mu$  for the expected value and  $\sigma^2$  for the variance.

Since we know what interpretation the parameters  $\mu$  and  $\sigma$  have, we know that in order to approximate the binomial random variable with parameters  $n$  and  $p$  with a normal distribution  $N(\mu, \sigma^2)$  we should choose  $\mu$  to be  $np$  and  $\sigma^2$  to be  $npq$ .

**109. The Central Limit Theorem.** We have said that the normal distribution approximates the binomial random variable but in fact it can be shown that it approximates *all* random variables in the following sense. If  $X$  is a random variable (discrete or continuous), and assuming that we can repeat the random process underlying  $X$  in an independent fashion as often as we like, we can consider  $n$  copies of  $X$ , denoted  $X_1, X_2$ , and so on, and the new random variable  $Y_n$  that is computed as

$$Y_n \stackrel{\text{def}}{=} X_1 + X_2 + \dots + X_n$$

The variable  $Y_n$  will have expected value  $n \times E[X]$  and variance  $n \times \text{Var}[X]$ .

One then further adjusts  $Y_n$  so that the expected value will be 0 and the variance 1, resulting in a variable that we call  $Z_n$ :

$$Z_n \stackrel{\text{def}}{=} \frac{Y_n - nE[X]}{\sqrt{n\text{Var}[X]}}$$

The **Central Limit Theorem** now states that the probabilities associated with the variable  $Z_n$  can be approximated by the probabilities that are computed from the standard normal distribution  $N(1,0)$ . As a formula, using the cumulative distribution function associated with random variables:

119: Central Limit Theorem

$$n \text{ large} \implies p(Z_n \leq r) \approx p(N(0, 1) \leq r)$$

The theorem explains why the bell curve is so often found when analyzing a random natural phenomenon: Very often, the randomness is the result of a large number of individual and independent choices and we can only observe their cumulative effect. The measured variable then is very much like our  $Y_n$  above and as the theorem shows, it is close to the normal distribution even if the individual processes were not.

**110. Practical advice.** In the exam I expect you to be able to

- compute expected value, variance, and standard variation for simple random variables (with finite range);
- use the Poisson random variable to model an appropriate situation.

We did not have enough time to explore how one can use the normal distribution to estimate probabilities, so that will not be part of the exam. Also, I don't expect you to memorize the formulas for expected value and variance of the various random variables listed in this handout.