Discrete Distributions

Probability Examples c-5

Leif Mejlbro

bookboon
The eBook company

Leif Mejlbro

Probability Examples c-5 Discrete Distributions

Probability Examples c-5 – Discrete Distributions

ISBN 978-87-7681-521-9

Contents

	Introduction	**5**
1	**Some theoretical background**	**6**
1.1	The binomial distribution	6
1.2	The Poisson distribution	8
1.3	The geometric distribution	8
1.4	The Pascal distribution	9
1.5	The hypergeometric distribution	11
2	**The binomial distribution**	**12**
3	**The Poisson distribution**	**26**
4	**The geometric distribution**	**33**
5	**The Pascal distribution**	**51**
6	**The negative binomial distribution**	**54**
7	**The hypergeometric distribution**	**56**
	Index	**72**

Introduction

This is the fifth book of examples from the *Theory of Probability*. This topic is not my favourite, however, thanks to my former colleague, Ole Jørsboe, I somehow managed to get an idea of what it is all about. The way I have treated the topic will often diverge from the more professional treatment. On the other hand, it will probably also be closer to the way of thinking which is more common among many readers, because I also had to start from scratch.

The prerequisites for the topics can e.g. be found in the *Ventus: Calculus 2* series, so I shall refer the reader to these books, concerning e.g. plane integrals.

Unfortunately errors cannot be avoided in a first edition of a work of this type. However, the author has tried to put them on a minimum, hoping that the reader will meet with sympathy the errors which do occur in the text.

Leif Mejlbro
26th October 2014

1 Some theoretical background

1.1 The binomial distribution

It is well-known that the *binomial coefficient*, where $\alpha \in \mathbb{R}$ and $k \in \mathbb{N}_0$, is defined by

$$\left(\begin{array}{c} \alpha \\ k \end{array} \right) := \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k(k-1)\cdots 1} = \frac{\alpha^{(k)}}{k!},$$

where

$$\alpha^{(k)} := \alpha(\alpha-1)\cdots(\alpha-k+1)$$

denotes the k-th decreasing factorial. If in particular, $\alpha = n \in \mathbb{N}$, and $n \geq k$, then

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

is the number of combinations of k elements chosen from a set of n elements without replacement.

If $n \in \mathbb{N}_0$ and $n < k$, then

$$\left(\begin{array}{c} n \\ k \end{array} \right) = 0.$$

Also the following formulæ should be well-known to the reader

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right) + \left(\begin{array}{c} n-1 \\ k \end{array} \right), \qquad k,\, n \in \mathbb{N}.$$

$$k \left(\begin{array}{c} n \\ k \end{array} \right) = n \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right), \qquad k,\, n \in \mathbb{N}.$$

Chu-Vandermonde's formula:

$$\left(\begin{array}{c} n+m \\ k \end{array} \right) = \sum_{i=0}^{k} \left(\begin{array}{c} n \\ i \end{array} \right) \left(\begin{array}{c} m \\ k-i \end{array} \right), \qquad k,\, m,\, n \in \mathbb{N}, \quad k \leq n+m.$$

$$(-1)^k \left(\begin{array}{c} -(\alpha+1) \\ k \end{array} \right) = \left(\begin{array}{c} \alpha+k \\ k \end{array} \right), \qquad \alpha \in \mathbb{R}, \quad k \in \mathbb{N}_0.$$

The binomial series:

$$(1+x)^{\alpha} = \sum_{k=0}^{+\infty} \left(\begin{array}{c} \alpha \\ k \end{array} \right) x^k, \qquad x \in \left]-1, 1\right[, \quad \alpha \in \mathbb{R}.$$

$$(1-x)^{-(\beta)} = \sum_{k=0}^{+\infty} \left(\begin{array}{c} -(\beta+1) \\ k \end{array} \right) (-1)^k x^k = \sum_{k=0}^{+\infty} \left(\begin{array}{c} \beta+k \\ k \end{array} \right) x^k.$$

If in particular $\alpha = n \in \mathbb{N}$, then

$$(a+b)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^k\, b^{n-k}, \qquad a,\, b \in \mathbb{R}.$$

$$\left(\begin{array}{c} -\frac{1}{2} \\ k \end{array} \right) (-4)^k = \left(\begin{array}{c} 2k \\ k \end{array} \right), \qquad \text{and} \qquad \left(\begin{array}{c} \frac{1}{2} \\ k \end{array} \right) (-4)^k = \frac{1}{2k-1} \left(\begin{array}{c} 2k \\ k \end{array} \right).$$

Stirling's formula

$$n! \sim \sqrt{2\pi n} \left\{ \frac{n}{e} \right\}^n,$$

where $\sim$ denotes that the proportion between the two sides tends to 1 for $n \to +\infty$.

A *Bernoulli event* is an event of two possible outcomes, S and F, of the probabilities p and $q = 1-p$, resp.. Here S is a shorthand for *Success*, and F for *Failure*. The probability of success is of course p.

Assume given a sequence of n Bernoulli events of the same probability of success p. Then the probability of getting precisely k successes, $0 \leq k \leq n$, is given by

$$\left(\begin{array}{c} n \\ k \end{array} \right) p^k\, (1-p)^{n-k},$$

where the binomial coefficient indicates the number of ways the order of the k successes can be chosen.

A random variable X is following a *binomial distributions with parameters n (the number of events) and* $p \in \left]0, 1\right[$ *(the probability)*, if the possible values of X are 0, 1, 2, ..., n, of the probabilities

$$P\{X = k\} = \left(\begin{array}{c} n \\ k \end{array} \right) p^k\, (1-p)^{n-k}, \qquad k = 0,\, 1,\, 2,\, \ldots,\, n.$$

In this case we write $X \in B(n,p)$. It is of course possible here to include the two causal distributions, where $p = 0$ or $p = 1$ give degenerated binomial distributions.

If $X \in B(n,p)$ is following a Bernoulli distribution, then

$$E\{X\} = np \qquad \text{and} \qquad V\{X\} = np(1-p) = npq.$$

If $X \in B(n,p)$ and $Y \in B(m,p)$ are independent and binomially distributed random variables, then the sum $X + Y \in B(n+m,p)$ is again binomially distributed. We say that the *binomial distribution is reproductive in the parameter of numbers for fixed probability parameter.*

1.2 The Poisson distribution

A random variable X is following a *Poisson distribution* of parameter $a \in \mathbb{R}_+$, and we write $X \in P(a)$, if the possible values of X lie in $\mathbb{N}_0$ of the probabilities

$$P\{X = k\} = \frac{a^k}{k!}\, e^{-a}, \qquad k \in \mathbb{N}_0.$$

In this case,

$$E\{X\} = a \qquad \text{and} \qquad V\{X\} = a.$$

When we compute the probabilities of a Poisson distribution it is often convenient to apply the recursion formula

$$P\{X = k\} = \frac{a}{k}\, P\{X = k-1\}, \qquad \text{for } k \in \mathbb{N}.$$

Assume that $\{X_n\}$ is a sequence of Bernoulli distributed random variables,

$$X_n \in B\left(n\,,\, \frac{a}{n}\right), \qquad a > 0 \text{ and } n \in \mathbb{N} \text{ med } n > a.$$

Then $\{X_n\}$ converges in distribution towards a random variable $X \in P(a)$, which is Poisson distributed.

1.3 The geometric distribution

A random variable X is following a *geometric distribution of parameter* $p \in]0,1[$, and we write $\mathrm{Pas}(1,p)$, if the possible values of X lie in $\mathbb{N}$ of the probabilities

$$P\{X = k\} = pq^{k-1} = p(1-p)^{k-1}, \qquad k \in \mathbb{N}.$$

We have the following results

$$P\{X \le k\} = 1 - q^k \qquad \text{and} \qquad P\{X > k\} = q^k, \qquad k \in \mathbb{N},$$

and

$$E\{X\} = \frac{1}{p} \qquad \text{and} \qquad V\{X\} = \frac{q}{p^2}.$$

Consider a number of tests under identical conditions, independent of each other. We assume in each test that an event A occurs with the probability p.
Then we define a random variable X by

$$X = k, \qquad \text{if } A \text{ occurs for the first time in test number } k.$$

Then X is geometrical distributed, $X \in \mathrm{Pas}(1, p)$.
For this reason we also say that the geometric distribution is a *waiting time distribution*, and p is called the *probability of success*.

If $X \in \mathrm{Pas}(1, p)$ is geometrically distributed, then

$$P\{X > m + n \mid X > n\} = P\{X > m\}, \qquad m,\, n \in \mathbb{N},$$

which is equivalent to

$$P\{X > m + n\} = P\{X > m\} \cdot P\{X > n\}, \qquad m,\, n \in \mathbb{N}.$$

For this reason we also say that the geometric distribution is *forgetful*: If we *know* in advance that the event A has not occurred in the first n tests, then the probability that A does not occur in the next m tests is equal to the probability that A does not occur in a series of m tests (without the previous n tests).

1.4 The Pascal distribution

Assume that $Y_1, Y_2, Y_3, \ldots$, are independent random variables, all geometrically distributed with the probability of success p. We define a random variable X_r by

$$X_r = Y_1 + Y_2 + \cdots + Y_r.$$

This random variable X_r has the values $r,\, r+1,\, r+2, \ldots$, where the event $\{X_r = k\}$ corresponds to the event that the r-th success occurs in test number k. Then the probabilities are given by

$$P\{X_r = k\} = \begin{pmatrix} k-1 \\ r-1 \end{pmatrix} p^r q^{k-r}, \qquad k = r,\, r+1,\, r+2, \ldots.$$

We say that $X_r \in \mathrm{Pas}(r, p)$ is following a *Pascal distribution of the parametres* $r \in \mathbb{N}$ and $p \in]0, 1[$, where r is called the *parameter of numbers*, and p is called the *parameter of probability*. In this case

$$E\{X\} = \frac{r}{p} \qquad \text{and} \qquad V\{X\} = \frac{rq}{p^2}.$$

If $X \in \mathrm{Pas}(r, p)$ and $Y \in \mathrm{Pas}(s, p)$ are independent Pascal distributed random variables of the same parameter of probability, then the sum $X + Y \in \mathrm{Pas}(r + s, p)$ is again Pascal distributed. We say that the *Pascal distribution is reproductive in the parameter of numbers for fixed parameter of probability*.

Clearly, the geometric distribution is that particular case of the Pascal distribution, for which the parameter of numbers is $r = 1$. We also call the Pascal distributions *waiting time distributions*.

It happens quite often that we together with $X_r \in \mathrm{Pas}(r, p)$ consider the *reduced waiting time*, $Z_r = X_r - r$. This is following the distribution

$$P\{Z_r = k\} = \left(\begin{array}{c} k+r-1 \\ k \end{array} \right) p^r q^k = (-1)^k \left(\begin{array}{c} -r \\ k \end{array} \right) p^r q^k, \qquad k \in \mathbb{N}_0,$$

with

$$E\{Z_r\} = \frac{rq}{p} \qquad \text{and} \qquad V\{Z_r\} = \frac{rq}{p^2}.$$

We interpret the random variable Z_r by saying that it represents the number of failures before the r-th success.

If $r \in \mathbb{N}$ above is replaced by $\kappa \in \mathbb{R}_+$, we get the *negative binomially distributed* random variable $X \in \mathrm{NB}(\kappa, p)$ with the parameters $\kappa \in \mathbb{R}_+$ and $p \in]0, 1[$, where X has $\mathbb{N}_0$ as its image, of the probabilities

$$P\{X = k\} = (-1)^k \left(\begin{array}{c} -\kappa \\ k \end{array} \right) p^\kappa q^k = \left(\begin{array}{c} k+\kappa-1 \\ k \end{array} \right) p^\kappa q^k, \qquad k \in \mathbb{N}_0.$$

In particular,

$$E\{X\} = \frac{\kappa q}{p} \qquad \text{and} \qquad V\{X\} = \frac{\kappa q}{p^2}.$$

If $\kappa = r \in \mathbb{N}$, then $X \in \mathrm{NB}(r,p)$ is the reduced waiting time $X = X_r - r$, where $X_r \in \mathrm{Pas}(r,p)$ is Pascal distributed.

1.5 The hypergeometric distribution

Given a box with the total number of N balls, of which a are white, while the other ones, $b = N - a$, are black. We select *without replacement* n balls, where $n \leq a + b = N$. Let X denote the random variable, which indicates the number of white balls among the chosen n ones. Then the values of X are

$$\max\{0\,,\, n-b\},\, \ldots,\, k,\, \ldots \min\{a, n\},$$

each one of the probability

$$P\{X = k\} = \frac{\left(\begin{array}{c} a \\ k \end{array}\right)\left(\begin{array}{c} b \\ n-k \end{array}\right)}{\left(\begin{array}{c} a+b \\ n \end{array}\right)}.$$

The distribution of X is called a *hypergeometric distribution*. We have for such a hypergeometric distribution,

$$E\{X\} = \frac{na}{a+b} \qquad \text{and} \qquad V\{X\} = \frac{nab(a+b-n)}{(a+b)^2(a+b-1)}.$$

There are some similarities between an hypergeometric distribution and a binomial distribution. If we change the model above by replacing the chosen ball (i.e. selection *with replacement*), and Y denotes the random variable, which gives us the number of white balls, then

$$Y \in B\left(n\,,\, \frac{a}{a+b}\right) \qquad \text{is binomially distributed,}$$

with

$$P\{Y = k\} = \left(\begin{array}{c} n \\ k \end{array}\right) \frac{a^k \cdot b^{n-k}}{(a+b)^n}, \qquad k \in \mathbb{N}_0,$$

and

$$E\{Y\} = \frac{na}{a+b} \qquad \text{and} \qquad V\{Y\} = \frac{nab}{(a+b)^2}.$$

2 The binomial distribution

Example 2.1 *Prove the formulæ*

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right) + \left(\begin{array}{c} n-1 \\ k \end{array} \right), \qquad k,\, n \in \mathbb{N},$$

$$k \left(\begin{array}{c} n \\ k \end{array} \right) = n \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right), \qquad k,\, n \in \mathbb{N},$$

either by a direct computation, or by a combinatorial argument.

By the definition

$$\begin{aligned} \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right) + \left(\begin{array}{c} n-1 \\ k \end{array} \right) &= \frac{(n-1)!}{(k-1)!(n-k)!} + \frac{(n-1)!}{k!(n-k-1)!} = \frac{(n-1)!}{k!(n-k)!}\{k + (n-k)\} \\ &= \frac{n!}{k!(n-k)!} = \left(\begin{array}{c} n \\ k \end{array} \right). \end{aligned}$$

It is well-known that we can select k elements out of n in $\left(\begin{array}{c} n \\ k \end{array} \right)$ ways.
This can also be described by

1) we either choose the first element, and then we are left with $k-1$ elements for $n-1$ places; this will give in total $\left(\begin{array}{c} n-1 \\ k-1 \end{array} \right)$, or
2) we do not select the first element, so k elements should be distributed among $n-1$ places. This gives in total $\left(\begin{array}{c} n-1 \\ k \end{array} \right)$ ways.

The result follows by adding these two possible numbers.

By the definition,

$$k \left(\begin{array}{c} n \\ k \end{array} \right) = k \cdot \frac{n!}{k!(n-k)!} = \frac{n(n-1)!}{(k-1)!(n-k)!} = n \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right).$$

Example 2.2 *Prove the formula*

$$\sum_{k=r}^{n} \left(\begin{array}{c} k \\ r \end{array} \right) = \left(\begin{array}{c} n+1 \\ r+1 \end{array} \right), \qquad r,\, n \in \mathbb{N}_0, \quad r \leq n.$$

HINT: *Find by two different methods the number of ways of choosing a subset of* $r+1$ *different numbers from the set of numbers 1, 2, dots,* $n+1$.

Clearly, the formula holds for $n = r$,

$$\sum_{k=r}^{r} \left(\begin{array}{c} k \\ r \end{array} \right) = \left(\begin{array}{c} r \\ r \end{array} \right) = 1 = \left(\begin{array}{c} r+1 \\ r+1 \end{array} \right) = \left(\begin{array}{c} n+1 \\ r+1 \end{array} \right).$$

Assume that the formula holds for some given $n \geq r$. Then we get for the successor $n+1$, according to the first question of Example 2.1,

$$\begin{aligned}\sum_{k=r}^{n+1}\binom{k}{r} &= \sum_{k=r}^{n}\binom{k}{r}+\binom{n+1}{r} = \binom{n+1}{r+1}+\binom{n+1}{r}\\ &= \binom{n+2}{r+1} = \binom{(n+1)+1}{r+1},\end{aligned}$$

and the claim follows by induction after n for every fixed r.
Since $r \in \mathbb{N}_0$ can be any number, the claim follows in general.

ALTERNATIVELY the formula can be shown by a combinatorial argument. Given the numbers 1, 2, 3, ... , $n+1$. From this we can choose a subset of $r+1$ different elements in $\binom{n+1}{r+1}$ different ways.

Then we split according to whether the *largest selected number* $k+1$ is $r+1$, $r+2$, ..., $n+1$, giving $(n+1)-k$ disjoint subclasses.
Consider one of these classes with $k+1$, $k = r, \ldots, n$, as its largest number. Then the other r numbers must be selected from 1, 2, 3, ... , k, which can be done in $\binom{k}{r}$ ways. Hence by summation and identification,

$$\sum_{k=r}^{n}\binom{k}{r} = \binom{n+1}{r+1}.$$

Example 2.3 *Prove by means of Stirling's formula,*

$$\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}\,2^{2n}.$$

REMARK: *One can prove that*

$$\sqrt{\frac{2n}{2n+1}}\cdot\frac{2^{2n}}{\sqrt{\pi n}} < \binom{2n}{n} < \frac{2^{2n}}{\sqrt{\pi n}}.$$

Stirling's formula gives for large n,

$$n! \sim \sqrt{2\pi n}\cdot n^n\cdot e^{-n},$$

hence

$$\binom{2n}{n} = \frac{(2n)!}{n!n!} \sim \frac{\sqrt{2\pi\cdot 2n}\cdot(2n)^{2n}e^{-2n}}{\left(\sqrt{2\pi n}\cdot n^n\cdot e^{-n}\right)^2} = \frac{2\sqrt{\pi n}\cdot 2^{2n}\cdot n^{2n}\cdot e^{-2n}}{2\pi n\cdot n^{2n}\cdot e^{-2n}} = \frac{1}{\sqrt{\pi n}}\cdot 2^{2n}.$$

Example 2.4 *Prove by means of the identity*

$$(1+x)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) x^k, \qquad x \in \mathbb{R}, \quad n \in \mathbb{N},$$

the formulæ

$$n(1+x)^{n-1} = \sum_{k=1}^{n} k \left(\begin{array}{c} n \\ k \end{array} \right) x^{k-1} = n \sum_{k=1}^{n} \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right) x^{k-1}, \qquad x \in \mathbb{R}, \quad n \in \mathbb{N},$$

and

$$\frac{(1+x)^{n-1}-1}{n+1} = \sum_{k=0}^{n} \frac{1}{k+1} \left(\begin{array}{c} n \\ k \end{array} \right) x^{k+1}, \qquad x \in \mathbb{R}, \quad n \in \mathbb{N}.$$

The former result follows by a partial differentiation and the second result of Example 2.1.

The latter result follows by an integration from 0.

Note that

$$\frac{1}{k+1}\left(\begin{array}{c} n \\ k \end{array}\right) = \frac{1}{k+1}\cdot\frac{n!}{k!(n-k)!} = \frac{1}{n+1}\cdot\frac{(n+1)!}{(k+1)!((n+1)-(k+1))} = \frac{1}{n+1}\left(\begin{array}{c} n+1 \\ k+1 \end{array}\right),$$

so we also have

$$\frac{(1+x)^{n+1}-1}{n+1} = \frac{1}{n+1}\sum_{k=0}^{n}\left(\begin{array}{c} n+1 \\ k+1 \end{array}\right)x^{k+1} = \frac{1}{n+1}\sum_{k=1}^{n+1}\left(\begin{array}{c} n+1 \\ k \end{array}\right)x^{k}.$$

Example 2.5 *A random variable X is binomially distributed, $X \in B(n,p)$. Prove that*

$$E\{X(X-1)\} = n(n-1)p^2,$$

and then find the variance of X.

By a direct computation,

$$\begin{aligned} E\{X(X-1)\} &= \sum_{k=1}^{n} k(k-1)\,P\{X=k\} = \sum_{k=2}^{n} k(k-1)\left(\begin{array}{c} n \\ k \end{array}\right)p^k(1-p)^{n-k} \\ &= n\sum_{k=2}^{n}(k-1)\left(\begin{array}{c} n-1 \\ k-1 \end{array}\right)p^k(1-p)^{n-k} = n(n-1)\sum_{k=2}^{n}\left(\begin{array}{c} n-2 \\ k-2 \end{array}\right)p^k(1-p)^{n-k} \\ &= n(n-1)p^2\sum_{\ell=0}^{n-2}\left(\begin{array}{c} n-2 \\ \ell \end{array}\right)p^{\ell}(1-p)^{(n-2)-\ell} = n(n-1)p^2. \end{aligned}$$

Since $E\{X\} = np$, or if one prefers to compute,

$$\begin{aligned} E\{X\} &= \sum_{k=1}^{n} k\,P\{X_k\} = \sum_{k=1}^{n} k\left(\begin{array}{c} n \\ k \end{array}\right)p^k(1-p)^{n-k} = n\sum_{k=1}^{n}\left(\begin{array}{c} n-1 \\ k-1 \end{array}\right)p^p(1-p)^k \\ &= np\sum_{\ell=0}^{n-1}\left(\begin{array}{c} n-1 \\ \ell \end{array}\right)p^{\ell}(1-p)^{(n-1)-\ell} = np, \end{aligned}$$

we get the variance

$$\begin{aligned} V\{X\} &= E\left\{X^2\right\} - (E\{X\})^2 = E\left\{X^2\right\} - E\{X\} + E\{X\} - (E\{X\})^2 \\ &= E\{X(X-1)\} + E\{X\} - (E\{X\})^2 = n(n-1)p^2 + np - n^2p^2 = -np^2 + np \\ &= np(1-p) = npq, \end{aligned}$$

where we as usual have put $q = 1-p$.

Example 2.6 *Consider a sequence of Bernoulli experiments, where there in each experiment is the probability p of success (S) and the probability q of failure (F). A sequence of experiments may look like*

$$SFSSSFFSSFFFFSFSFF\cdots .$$

1) Find the probability that FF occurs before FS.

2) Find the probability that FF occurs before før SF.

1) Both FF and FS assume that we first have an F, thus

$$P\{FF \text{ before } FS\} = P\{F\} = q.$$

2) If just one S occurs, then FF cannot occur before SF, thus

$$P\{FF \text{ before } SF\} = P\{F \text{ in the first experiment}\} \cdot P\{F \text{ in the second experiment}\} = q^2.$$

Example 2.7 *We perform a sequence of independent experiments, and in each of them there is the probability p, where $p \in]0,1[$, that an event A occurs.*
Denote by X_n the random variable, which indicates the number of times that the event A has occurred in the first n experiments, and let X_{n+k} denote the number of times the event A has occurred in the first $n+k$ experiments.
Compute the correlation coefficient between X_n and X_{n+k}.

First note that we can write

$$X_{n+k} = X_n + Y_k,$$

where Y_k is the number of times that A has occurred in the *latest* k experiments.

The three random variables are all Bernoulli distributed,

$$X_n \in B(n,p), \qquad Y_k \in B(k,p) \quad \text{and} \quad X_{n+k} \in B(n+k,p).$$

Since X_n and Y_k are independent, we have

$$\mathrm{Cov}\,(X_n, X_n + Y_k) = V\{X_n\} = np(1-p).$$

Furthermore,

$$V\{X_{n+k}\} = (n+k)p(1-p).$$

Then

$$\varrho\,(X_n, X_{n+k}) = \frac{\mathrm{Cov}\,(X_n, X_{n+k})}{\sqrt{V\{X_n\}\; V\{X_{n+k}\}}} = \sqrt{\frac{V\{X_n\}}{V\{X_{n+k}\}}} = \sqrt{\frac{n}{n+k}}.$$

Example 2.8 1. *Let the random variable X_1 be binomially distributed, $X_1 \in B(n,p)$, (where $p \in]0,1[$). Prove that the random variable $X_2 = n - X_1$ is binomially distributed, $X_2 \in B(n,1-p)$.*

Let F denote a experiment of three different possible events A, B and C. The probabilities of these events are a, b and c, resp., where

$$a > 0, \qquad b > 0, \qquad c > 0 \quad \text{and} \quad a+b+c = 1.$$

Consider a sequence consisting of n independent repetitions of F.
Let X, Y and Z denote the number of times in the sequence that A, B and C occur.

2. *Find the distribution of the random variables X, Y and Z and find*

$$V\{X\}, \qquad V\{Y\} \quad \text{and} \quad V\{X+Y\}.$$

3. *Compute* $\mathrm{Cov}(X,Y)$ *and the correlation coefficient* $\varrho(X,Y)$.

4. *Compute $P\{X = i \wedge Y = j\}$ for $i \geq 0$, $j \geq 0$, $i + j \leq n$.*

1) This is almost trivial, because

$$P\{X_k = k\} = P\{X_1 = n-k\} = \left(\begin{array}{c} n \\ n-k \end{array} \right) p^{n-k}(1-p)^k = \left(\begin{array}{c} n \\ k \end{array} \right) (1-p)^k p^{n-k}$$

for $k = 0,\ 1,\ 2,\ \ldots,\ n$, thus $X_2 \in B(n,1-p)$.
We express this by saying that X_1 counts the successes and X_2 counts the failures.

2) Since $X \in B(n,a)$ and $Y \in B(n,b)$ and $Z \in B(n,c)$, it follows immediately that

$$V\{X\} = na(1-a) \qquad \text{and} \qquad V\{Y\} = nb(1-b).$$

We get from $X + Y + Z = n$ that $X + Y = n - Z \in B(n,1-c)$, so

$$V\{X+Y\} = V\{n-Z\} = V\{Z\} = nc(1-c).$$

3) From $a+b+c = 1$ follows that $c = 1-(a+b)$ and $1-c = a+b$. Then it follows from (2) that

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \frac{1}{2}\left(V\{X+Y\} - V\{X\} - V\{Y\}\right) \\ &= \frac{n}{2}\left\{[1-(a+b)](a+b) - a(1-a) - b(1-b)\right\} \\ &= \frac{n}{2}\left\{(a+b) - (a+b)^2 - (a+b) + \left(a^2+b^2\right)\right\} \\ &= \frac{n}{2}\left(-2ab\right) = -nab. \end{aligned}$$

Hence,

$$\varrho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{V\{X\} \cdot V\{Y\}}} = \frac{-nab}{\sqrt{na(b+c)nb(a+c)}} = -\sqrt{\frac{ab}{(a+c)(b+c)}} = -\sqrt{\frac{ab}{(1-a)(1-b)}}.$$

4) We can select i events A in $\left(\begin{array}{c} n \\ i \end{array} \right)$ ways, and then j events B in $\left(\begin{array}{c} n-i \\ j \end{array} \right)$ ways, so

$$P\{X = i, Y = j\} = \left(\begin{array}{c} n \\ i \end{array} \right) \left(\begin{array}{c} n-i \\ j \end{array} \right) a^i b^j c^{n-i-j}.$$

Remark 2.1 We get by a reduction,

$$\left(\begin{array}{c} n \\ i \end{array} \right) \left(\begin{array}{c} n-i \\ j \end{array} \right) = \frac{n!}{i!(n-i)!} \cdot \frac{(n-i)!}{j!(n-i-j)!} = \frac{n!}{i!j!(n-i-j)!} := \left(\begin{array}{c} n \\ i\,,\, j \end{array} \right)$$

analogously to the binomial distribution. Note that $i + j + (n - i - j) = n$, cf. the denominator. Therefore, we also write

$$P\{X = i,\, Y = j\} = \left(\begin{array}{c} n \\ i\,,\, j \end{array} \right) a^i b^j (1 - ab)^{n-i-j},$$

and the distribution of (X, Y) is a *trinomial distribution* (multinomial distribution or polynomial distribution). $\Diamond$

Example 2.9 *An event H has the probability p, where $0 < p < 1$, to occur in one single experiment. An experiment consists of $10n$ independent experiments.*
Let X_1 denote the random variable which gives us the number of times which the event H occurs in the first n experiments.
Then let X_2 denote the random variable, which indicates the number of times that the event H occurs in the following $2n$ experiments.
Let X_3 denote the random variable, which indicates the number of times the event H occurs in the following $3n$ experiments.
Finally, X_4 denotes the random variable, which gives us the number of times the event H occurs in the remaining $4n$ experiments.

1. *Find the distributions of X_1, X_2, X_3 and X_4.*

Using the random variables X_1, X_2, X_3 and X_4 we define the new random variables

$$Y_1 = X_2 + X_3, \quad Y_2 = X_3 + X_4, \quad Y_3 = X_1 + X_2 + X_3 \quad \textit{and} \quad Y_4 = X_2 + X_3 + X_4.$$

2. *Compute the correlation coefficients $\varrho(Y_1, Y_2)$ and $\varrho(Y_3, Y_4)$.*

1) Clearly,

$$\begin{aligned}
P\{X_1 = k\} &= \binom{n}{k} p^k (1-p)^{n-k}, && k = 0, 1, \ldots, n,\\
P\{X_2 = k\} &= \binom{2n}{k} p^k (1-p)^{2n-k}, && k = 0, 1, \ldots, 2n,\\
P\{X_3 = k\} &= \binom{3n}{k} p^k (1-p)^{3n-k}, && k = 0, 1, \ldots, 3n,\\
P\{X_4 = k\} &= \binom{4n}{k} p^k (1-p)^{4n-k}, && k = 0, 1, \ldots, 4n.
\end{aligned}$$

2) Since $Y_1 = X_2 + X_3$ is the number of times which H occurs in a sequence of experiments consisting of $5n$ experiments, then

$$P\{Y_1 = k\} = \binom{5n}{k} p^k (1-p)^{5n-k}, \qquad k = 0, 1, \ldots, 5n.$$

Analogously,

$$\begin{aligned}
P\{Y_2 = k\} &= \binom{7n}{k} p^k (1-p)^{7n-k}, && k = 0, 1, \ldots, 7n,\\
P\{Y_3 = k\} &= \binom{6n}{k} p^k (1-p)^{6n-k}, && k = 0, 1, \ldots, 6n,\\
P\{Y_4 = k\} &= \binom{9n}{k} p^k (1-p)^{9n-k}, && k = 0, 1, \ldots, 9n.
\end{aligned}$$

Since in general for a binomial distribution $X \in B(m, p)$,

$$E\{X\} = mp \qquad \text{and} \qquad V\{X\} = mp(1-p),$$

we get here

$$\begin{aligned}
V\{Y_1\} &= 5np(1-p), & V\{Y_2\} &= 7np(1-p),\\
V\{Y_3\} &= 6np(1-p), & V\{Y_4\} &= 9np(1-p).
\end{aligned}$$

Since the X_i-s are mutually independent, we get

$$\begin{aligned}
\operatorname{Cov}(Y_1, Y_2) &= E\{Y_1 Y_2\} - E\{Y_1\}\, E\{Y_2\} \\
&= E\{(X_2+X_3)(X_3+X_4)\} - (E\{X_2\}+E\{X_3\})\,(E\{X_3\}+E\{X_4\}) \\
&= E\{X_2\,(X_3+X_4)\} - E\{X_2\}\cdot(E\{X_3\}+E\{X_4\}) \\
&\qquad +E\left\{X_3^2\right\} - (E\{X_3\})^2 + E\{X_3X_4\} - E\{X_3\}\,E\{X_4\} \\
&= 0 + V\{X_3\} + 0 = 3np(1-p),
\end{aligned}$$

hence

$$\varrho(Y_1 m Y_2) = \frac{\operatorname{Cov}(Y_1, Y_2)}{\sqrt{V\{Y_1\}\cdot V\{Y_2\}}} = \frac{3np(1-p)}{\sqrt{5np(1-p)\cdot 7np(1-p)}} = \frac{3}{\sqrt{35}} = \frac{3\sqrt{35}}{35}.$$

Analogously,

$$\begin{aligned}\mathrm{Cov}\,(Y_3,Y_4) &= E\{Y_3Y_4\} - E\{Y_3\}\,E\{Y_4\}\\ &= E\{(X_1+X_2+X_3)\,(X_2+X_3+X_4)\}\\ &\qquad -E\{X_1+X_2+X_3\}\,E\{X_2+X_3+X_4\}\\ &= E\{X_1\,(X_2+X_3+X_4)\} - E\{X_1\}\,E\{X_2+X_3+X_4\}\\ &\qquad +E\{(X_2+X_3)\,X_4\} - E\{X_2+X_3\}\,E\{X_4\}\\ &\qquad +E\left\{(X_2+X_3)^2\right\} - (E\{X_2+X_3\})^2\\ &= 0+0+V\{X_2+X_3\} = V\{Y_1\} = 5np(1-p),\end{aligned}$$

hence

$$\varrho\,(Y_3,Y_4) = \frac{\mathrm{Cov}\,(Y_3,Y_4)}{\sqrt{V\{Y_3\}\cdot V\{Y_4\}}} = \frac{5np(1-p)}{\sqrt{6np(1-p)\cdot 9np(1-p)}} = \frac{5}{3\sqrt{6}} = \frac{5\sqrt{6}}{18}.$$

Example 2.10 *An assembly of 20 politicians consists of 10 from party A, 8 from party B and 2 from parti C. A committee is going to be set up.*

1) The two politicians from party C want that the committee has such a size that the probability of C being represented in the committee, when its members are chosen randomly, is at least 25 %. How big should the committee be to fulfil this demand?

2) One agrees on the size of the committee of 3. If the members are chosen randomly among the 20 politicians, one shall find the distribution of the 2-dimensional random variable (X_A, X_B). (Here, X_A indicates the number of members in the committee from A, and X_B the number of members from B, and X_C the number of members from C).

3) Find the distribution for each of the random variables X_A, X_B and X_C.

4) Find the mean and variance for each of the random variables X_A, X_B and X_C.

1) Let n, $1 \le n \le 20$, denote the size of the committee. We shall compute the probability that C is not represented in the committee. Clearly, $n \le 18$. Then the probability that C is not represented is given by

$$P\{n \text{ chosen among the 18 members of A and B}\} = \frac{18}{20}\cdot\frac{17}{19}\cdots\frac{19-n}{21-n}.$$

The requirement is that this probability is ≤ 75 %. By trial-and-error we get

$$\begin{aligned} n=1: &\quad \frac{18}{20} = 0,90,\\ n=2: &\quad \frac{18}{20}\cdot\frac{17}{19} = 0,8053,\\ n=3: &\quad \frac{18}{20}\cdot\frac{17}{19}\cdot\frac{16}{18} = 0,7158.\end{aligned}$$

We conclude that the committee should at least have 3 members.

2) We can choose 3 from 20 members in a total of

$$\binom{20}{3} = \frac{20 \cdot 19 \cdot 18}{1 \cdot 2 \cdot 3} = 20 \cdot 19 \cdot 3 = 1140 \text{ ways.}$$

This gives us the following probabilities of the various possibilities:

$$P\{3 \text{ from A}\} = \frac{1}{1140}\binom{10}{3} = \frac{120}{1140} = \frac{2}{19},$$

$$P\{2 \text{ from A and } 1 \text{ from B}\} = \frac{1}{1140}\binom{10}{2}\binom{8}{1} = \frac{360}{1140} = \frac{6}{19},$$

$$P\{2 \text{ from A and } 1 \text{ from C}\} = \frac{1}{1140}\binom{10}{2}\binom{2}{1} = \frac{90}{1140} = \frac{3}{38},$$

$$P\{1 \text{ from A and } 2 \text{ from B}\} = \frac{1}{1140}\binom{10}{1}\binom{8}{2} = \frac{280}{1140} = \frac{14}{57},$$

$$P\{1 \text{ from each of A and B and C}\} = \frac{1}{1140}\binom{10}{1}\binom{8}{1}\binom{2}{1} = \frac{160}{1140} = \frac{8}{57},$$

$$P\{1 \text{ from A and } 2 \text{ from C}\} = \frac{1}{1140}\binom{10}{1}\binom{2}{2} = \frac{10}{1140} = \frac{1}{114},$$

$$P\{3 \text{ from B}\} = \frac{1}{1140}\binom{8}{3} = \frac{56}{1140} = \frac{14}{285},$$

$$P\{2 \text{ from B and } 1 \text{ from C}\} = \frac{1}{1140}\binom{8}{2}\binom{2}{1} = \frac{56}{1140} = \frac{14}{285},$$

$$P\{1 \text{ from B and } 2 \text{ from C}\} = \frac{1}{1140}\binom{8}{1}\binom{2}{2} = \frac{8}{1140} = \frac{2}{285}.$$

By suitable interpretations, e.g.

$$P\{X_A = 2,\, X_B = 0\} = P\{X_A = 2,\, X_C = 1\},$$

etc., we obtain the distribution of (X_A, X_B),

$X_B \setminus X_A$	0	1	2	3	$X_B^\star$
0	0	$\frac{1}{114}$	$\frac{3}{38}$	$\frac{2}{19}$	$\frac{11}{57}$
1	$\frac{2}{285}$	$\frac{8}{57}$	$\frac{6}{19}$	$\star$	$\frac{44}{95}$
2	$\frac{14}{285}$	$\frac{14}{57}$	$\star$	$\star$	$\frac{28}{95}$
3	$\frac{14}{285}$	$\star$	$\star$	$\star$	$\frac{14}{285}$
$X_A^\star$	$\frac{2}{19}$	$\frac{15}{38}$	$\frac{15}{38}$	$\frac{2}{19}$	1

Table 1: The distribution of (X_A, X_B).

$$P\{X_A = 3,\, X_B = 0\} = \frac{2}{19},$$

$$P\{X_A = 2,\, X_B = 1\} = \frac{6}{19},$$

$$P\{X_A = 2,\, X_B = 0\} = \frac{3}{38},$$

$$P\{X_A = 1,\, X_B = 2\} = \frac{14}{57},$$

$$P\{X_A = 1,\, X_B = 1\} = \frac{8}{57},$$

$$P\{X_A = 1,\, X_B = 0\} = \frac{1}{114},$$

$$P\{X_A = 0,\, X_B = 3\} = \frac{14}{285},$$

$$P\{X_A = 0,\, X_B = 2\} = \frac{14}{285},$$

$$P\{X_A = 0,\, X_B = 1\} = \frac{2}{285},$$

$$P\{X_A = 0,\, X_B = 0\} = 0.$$

This distribution is best described by the table on page 20.

3) The distributions of X_A and X_B are marked in the table by $X_A^\star$ and $X_B^\star$.
We compute the distribution of X_C by

$$\begin{aligned}
P\{X_C = 2\} &= P\{X_A = 1,\, X_C = 2\} + P\{X_B = 1,\, X_C = 2\} = \frac{1}{114} + \frac{2}{285} = \frac{3}{190} \\
P\{X_C = 1\} &= P\{X_A = 2,\, X_C = 1\} + P\{X_A = 1,\, X_B = 1,\, X_C = 1\} \\
&\qquad + P\{X_B = 2\, X_C = 1\} \\
&= \frac{3}{38} + \frac{8}{57} + \frac{14}{285} = \frac{51}{190}, \\
P\{X_C = 0\} &= P\{X_A = 3\} + P\{X_A = 2,\, X_B = 1\} \\
&\qquad + P\{X_A = 1,\, X_B = 2\} + P\{X_B = 3\} \\
&= \frac{2}{19} + \frac{6}{19} + \frac{14}{57} + \frac{14}{285} = \frac{68}{95}.
\end{aligned}$$

Summing up we obtain the distributions given on page 22.

	0	1	2	3
X_A	$\frac{2}{19}$	$\frac{15}{38}$	$\frac{15}{38}$	$\frac{2}{19}$
X_B	$\frac{11}{57}$	$\frac{44}{95}$	$\frac{28}{95}$	$\frac{14}{285}$
X_C	$\frac{68}{95}$	$\frac{51}{190}$	$\frac{3}{190}$	0

Table 2: The distributions of X_A, X_B and X_C.

4) The means are

$$E\{X_A\} = 1 \cdot \frac{15}{38} + 2 \cdot \frac{15}{38} + 3 \cdot \frac{2}{19} = \frac{1}{38}(15 + 30 + 12) = \frac{57}{38} = \frac{3}{2},$$

$$E\{X_B\} = 1 \cdot \frac{44}{95} + 2 \cdot \frac{28}{95} + 3 \cdot \frac{14}{285} = \frac{1}{95}(44 + 56 + 14) = \frac{114}{95} = \frac{6}{5},$$

$$E\{X_C\} = 1 \cdot \frac{51}{190} + 2 \cdot \frac{3}{190} = \frac{57}{190} = \frac{3}{10},$$

which can also be found in other ways. We have e.g.

$$E\{X_A\} + E\{X_B\} + E\{X_C\} = 3 \qquad \text{and} \qquad E\{X_A\} = \frac{3}{2},$$

and then we only have to compute $E\{X_C\}$.

Furthermore,

$$E\left\{X_A^2\right\} = 1 \cdot \frac{15}{38} + 4 \cdot \frac{15}{38} + 9 \cdot \frac{4}{38} = \frac{1}{38}(15 + 60 + 36) = \frac{111}{38},$$

$$E\left\{X_B^2\right\} = 1 \cdot \frac{44}{95} + 4 \cdot \frac{28}{95} + 9 \cdot \frac{14}{285} = \frac{1}{95}(44 + 112 + 42) = \frac{198}{95},$$

$$E\left\{X_C^2\right\} = 1 \cdot \frac{51}{190} + 4 \cdot \frac{3}{190} = \frac{1}{190}(51 + 12) = \frac{63}{190},$$

so

$$V\{X_A\} = \frac{111}{38} - \frac{9}{4} = \frac{51}{76},$$

$$V\{X_B\} = \frac{198}{95} - \frac{36}{25} = \frac{306}{475},$$

$$V\{X_C\} = \frac{63}{190} - \frac{9}{100} = \frac{459}{1900}.$$

3 The Poisson distribution

Example 3.1 *Let X and Y be independent Poisson distributed random variables of the parameters a and b, resp.. Let $Z = X + Y$. Find*

$$P\{X = k \mid Z = n\}, \qquad k = 0,\, 1,\, 2,\, \ldots,\, n.$$

Since X and Y are independent, we get for $k = 0,\, 1{,}2,\, \ldots,\, n$, that

$$\begin{aligned} P\{X = k \mid Z = n\} &= \frac{P\{X = k \wedge X + Y = n\}}{P\{Z = n\}} = \frac{P\{X = k \wedge Y = n - k\}}{P\{Z = n\}} \\ &= \frac{P\{X = k\} \cdot P\{Y = n - k\}}{P\{Z = n\}}. \end{aligned}$$

Since $Z = X + Y$ is Poisson distributed with parameter $a + b$, we get

$$\begin{aligned} P\{X = k \mid Z = n\} &= \frac{\frac{a^k}{k!}\, e^{-a} \cdot \frac{b^{n-k}}{(n-k)!}\, e^{-b}}{\frac{(a+b)^n}{n!}\, e^{-(a+b)}} = \frac{n!}{k!(n-k)!}\, \frac{a^k b^{n-k}}{(a+b)^n} \\ &= \left(\begin{array}{c} n \\ k \end{array} \right) \left\{ \frac{a}{a+b} \right\}^k \cdot \left\{ \frac{b}{a+b} \right\}^{n-k}, \end{aligned}$$

which describes a binomial distribution $B\left(n, \frac{a}{a+b}\right)$.

Example 3.2 *Let X and Y be random variables for which*

$$X \text{ is Poisson distributed}, \qquad X \in P(a),$$

and

$$P\{Y = k \mid X = n\} = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}, \qquad k = 0,\, 1,\, \ldots,\, n,$$

where $n \in \mathbb{N}_0$ and $p \in]0, 1[$.
Prove that Y is Poisson distributed, $Y \in P(ap)$.

We get for $k \in \mathbb{N}_0$,

$$\begin{aligned} P\{Y = k\} &= \sum_{n=k}^{\infty} P\{Y = k \mid X = n\} \cdot P\{X = n\} = \sum_{n=k}^{\infty} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \cdot \frac{a^n}{n!}\, e^{-a} \\ &= \sum_{n=k}^{\infty} \frac{n!}{k!(n-k)!}\, \frac{e^{-a}}{n!}\, p^k (1-p)^{n-k} \cdot a^n = \frac{e^{-a}}{k!}\, a^k p^k \sum_{n=k}^{\infty} \frac{1}{(n+k)!}\, (1-p)^{n-k} a^{n-k} \\ &= \frac{e^{-a}}{k!}\, a^k p^k \sum_{n=0}^{\infty} \frac{1}{n!}\, \{(1-p)\}^n = \frac{e^{-a}}{k!}\, a^k p^k \sum_{n=0}^{\infty} \frac{1}{n!}\, \{(1-p)a\}^n \\ &= \frac{e^{-a}}{k!}\, a^k p^k \cdot e^{(1-p)a} = \frac{1}{k!}\, (ap)^k e^{-ap}, \end{aligned}$$

proving that $Y \in P(ap)$.

Example 3.3 *We have in* Ventus: Probability c-2 *introduced skewness of a distribution. Compute the skewness $\gamma(X)$ of a random variable X, which is Poisson distributed with parameter a. What happens to $\gamma(X)$, when $a \to \infty$?*

The skewness is defined as

$$\gamma\left(X_a\right) = \frac{1}{\sigma^3}\, E\left\{\left(X_a - \mu\right)^3\right\},$$

where

$$E\left\{\left(X_a - \mu\right)^3\right\} = E\left\{X_a^3\right\} - \mu\left(3\sigma^2 + \mu^2\right).$$

We have for a Poisson distributed random variable $X_a \in P(a)$,

$$P\left\{X_a = k\right\} = \frac{a^k}{k!}\, e^{-a}, \qquad k \in \mathbb{N}_0, \quad a > 0,$$

and

$$E\left\{X_a\right\} = \mu = a \qquad \text{and} \qquad V\left\{X_a\right\} = \sigma^2 = a.$$

It follows from

$$E\left\{X_a\left(X_a - 1\right)\left(X_a - 2\right)\right\} = \sum_{k=3}^{\infty} k(k-1)(k-2)\,\frac{a^k}{k!}\, e^{-a} = a^3 \sum_{k=3}^{\infty} \frac{a^{k-3}}{(k-3)!}\, e^{-a} = a^3,$$

and

$$\begin{aligned} X_a^3 &= X_a\left(X_a - 1\right)\left(X_a - 2\right) + 3X_a^2 - 2X_a \\ &= X_a\left(X_a - 1\right)\left(X_a - 2\right) + 3X_a\left(X_a - 1\right) + X_a, \end{aligned}$$

that

$$E\left\{X_a^3\right\} = a^3 + 3a^2 + a,$$

hence

$$E\left\{\left(X_a - \mu\right)\right\} = E\left\{X_a^3\right\} - \mu\left(3\sigma^2 + \mu^2\right) = \left(a^3 + 3a^2 + a\right) - a\left(3a - a^3\right) = a.$$

We get the skewness by insertion,

$$\gamma\left(X_a\right) = \frac{1}{\sigma^3}\, E\left\{\left(X_a - \mu\right)^3\right\} = \frac{a}{a\sqrt{a}} = \frac{1}{\sqrt{a}}.$$

Finally, it follows that

$$\lim_{a\to\infty} \gamma\left(X_a\right) = \lim_{a\to\infty} \frac{1}{\sqrt{a}} = 0.$$

X_i	n_i
0	29
1	42
2	21
3	16
4	7
5	2
6	3
≥ 7	0
	120

Example 3.4 *A carpark where one may park at most 1 hour, has 192 parking places. On the first five days (Monday – Friday) in one particular week someone has counted the number of vacant places at the time intervals of 5 minutes between 2:00 PM and 4:00 PM, i.e. performed 24 observations per day, thus in total 120 observations. The results of this investigation (which comes from Los Angeles) is shown in the table. Here X_i denotes the number of vacant places and n_i the number of times there were observed X_i vacant places.*

1) We assume that the number X of vacant places at any time between 2 PM and 4 PM is Poisson distributed of the same mean as in the given results of the observations. Compute the expected number of observations. corresponding to

$$X = 0, \quad X = 1, \quad X = 2, \quad X = 3, \quad X = 4, \quad X = 5, \quad X = 6, \quad X \geq 7.$$

2) A cardriver tries once a day on each of the first five days of the week to park in the carpark between 2 PM and 4 PM. Find the probability that he finds a vacant place on every of these days, and find the probability that he finds a vacant place just one of the five days.

1) We first determine λ by

$$\lambda = \frac{1}{120}\left(1 \cdot 42 + 2 \cdot 21 + 3 \cdot 16 + 4 \cdot 7 + 5 \cdot 2 + 6 \cdot 3\right) = \frac{47}{30} \approx 1.5667.$$

Using this λ we get

$$P\{X = k\} = \frac{\lambda^k}{k!}\, e^{-\lambda}, \qquad k \in \mathbb{N}_0.$$

When these probabilities are multiplied by the total number 120, we obtain the following table:

n	0	1	2	3	4	5	6	≥ 7
X_{observed}	29	42	21	16	7	2	3	0
X_{expected}	25.0	39.2	30.7	16.1	6.3	2.0	0.5	0.2

2) The probability that he finds a vacant place on one day is

$$P\{X \geq 1\} = 1 - P\{X = 0\} = 1 - e^{-\lambda} = 1 - \exp\left(-\frac{47}{30}\right) \approx 0.79.$$

The probability that he finds a vacant place on all 5 days is

$$(P\{X \geq 1\})^5 = \left\{1 - \exp\left(-\frac{47}{30}\right)\right\}^5 \approx 0.31.$$

The probability that he finds a vacant place on one particular day (Monday – Friday), but not on any other of the days is

$$P\{X \geq 1\} \cdot (P\{X = 0\})^4 = \left(1 - \exp\left(-\frac{47}{30}\right)\right) \cdot \exp\left(-4 \cdot \frac{47}{30}\right).$$

The probability that he finds a vacant place on precisely one of the 5 days is

$$5 \cdot P\{X \geq 1\} \cdot (P\{X = 0\})^4 = 5\left(1 - \exp\left(-\frac{47}{30}\right)\right) \cdot \exp\left(-4 \cdot \frac{47}{30}\right) \approx 0.0075.$$

Example 3.5 *A schoolboy tries modestly to become a newspaper delivery boy. He buys every morning k newspapers for 8 DKK for each, and he tries to sell them on the same day for 12 DKK each. The newspapers which are not sold on the same day are discarded, so the delivery boy may suffer a loss by buying too many newspapers.*
Experience shows that the number of newspapers which can be sold each day, X, can be assumed approximately to follow a Poisson distribution of mean 10, thus

$$P\{X = k\} = \frac{10^k}{k!}\, e^{-10}, \qquad k \in \mathbb{N}_0.$$

k	$\sum_{m=0}^{k} \frac{1}{m!}\, 10^m e^{-10}$
0	0.00005
1	0.0005
2	0.0028
3	0.0104
4	0.0293
5	0.0671
6	0.1302
7	0.2203
6	0.3329
9	0.4580
10	0.5831

Let E_k denote his expected profit per day (which possibly may be negative), when he buys k newspapers in the morning.

1) Prove that

$$E_k = \sum_{m=0}^{k} (12m - 8k)\, \frac{10^m}{m!}\, e^{-10} + \sum_{m=k+1}^{\infty} 4k\, \frac{10^m}{m!}\, e^{-10}.$$

2) Prove that

$$E_{k+1} - E_k = 4 - 12 \sum_{m=0}^{k} \frac{10^m}{m!}\, e^{-10}.$$

3) Compute by using the table, $E_{k+1} - E_k$ and E_k for $k = 0, 1, \ldots, 10$ (2 decimals).

4) Find the value of k, for which his expected profit E_k is largest.

1) He spends in total $8k$ DKK, when he buys k newspapers and earns 12 DKK for each newspaper he sells. Thus

$$\begin{aligned} E_k &= \sum_{m=0}^{k} (12m - 8k)\, \frac{10^m}{m!}\, e^{-10} + \sum_{m=k+1}^{\infty} (12k - 8k)\, \frac{10^m}{m!}\, e^{-10} \\ &= \sum_{m=0}^{k} (12m - 8k)\, \frac{10^m}{m!}\, e^{-10} + 4k \sum_{k+1}^{\infty} \frac{10^m}{m!}\, e^{-10}. \end{aligned}$$

2) Then, using the result of (1),

$$\begin{aligned}
E_{k+1} - E_k &= \sum_{m=0}^{k+1} (12m - 8k - 8)\, \frac{10^m}{m!}\, e^{-10} + 4(k+1) \sum_{m=k+2}^{\infty} \frac{10^m}{m!}\, e^{-10} \\
&\quad - \sum_{m=0}^{k} (12m - 8k)\, \frac{10^m}{m!}\, e^{-10} - 4k \sum_{m=k+1}^{\infty} \frac{10^m}{m!}\, e^{-10} \\
&= \sum_{m=0}^{k} (12m - 8k)\, \frac{10^m}{m!}\, e^{-10} - 8 \sum_{m=0}^{k} \frac{10^m}{m!}\, e^{-10} \\
&\quad + 4(k+1) \cdot \frac{10^{k+1}}{(k+1)!}\, e^{-10} + 4k \sum_{m=k+2}^{\infty} \frac{10^m}{m!}\, e^{-10} \\
&\quad + 4 \sum_{m=k+2}^{\infty} \frac{10^m}{m!}\, e^{-10} - \sum_{m=0}^{k} (1.5m - k)\, \frac{10^m}{m!}\, e^{-10} \\
&\quad - 4k \sum_{m=k+2}^{\infty} \frac{10^m}{m!}\, e^{-10} - 4k \cdot \frac{10^{k+1}}{(k+1)!}\, e^{-10} \\
&= - \sum_{m=0}^{k} \frac{10^m}{m!}\, e^{-10} + 4 \sum_{m=k+1}^{\infty} \frac{10^m}{m!}\, e^{-10} \\
&= 4 \sum_{m=0}^{\infty} \frac{10^m}{m!}\, e^{-10} - 1.5 \sum_{m=0}^{k} \frac{10^m}{m!}\, e^{-10} \\
&= 4 - 12 \sum_{m=0}^{k} \frac{10^m}{m!}\, e^{-10}.
\end{aligned}$$

k	$E_{k+1} - E_k$	E_k
0	3.9994	0.00
1	3.9940	4.00
2	3.9664	7.99
3	3.8752	11.96
4	3.6484	15.84
5	3.1948	19.48
6	2.4376	22.68
7	1.3564	25.12
8	0.0052	26.47
9	-1.4960	26.48
10	-2.9972	24.98

3) Using the formulæ

$$E_{k+1} - E_k = 4 - 12 \sum_{m=0}^{k} \sum_{m=0}^{m} \frac{10^m}{m!}\, e^{-10}$$

and

$$E_{k+1} = (E_{k+1} - E_k) + E_k$$

we get the table on the previous page.

4) We obtain the largest expected profit, 26.48 DKK, for $k = 9$, because

$$E_9 - E_8 > 0 \qquad \text{and} \qquad E_{10} - E_9 < 0.$$

4 The geometric distribution

Example 4.1 *Given a random variable X of the values 1, 2, ... of positive probabilities. Assume furthermore,*

$$P\{X > m+n\} = P\{X > m\} \cdot P\{X > n\}, \qquad m,\, n \in \mathbb{N}.$$

Prove that the distribution of X is a geometric distribution.

It follows from the assumptions that

$$P\{X > n\} = P\{X > 1 + \cdots + 1\} = (P\{X > 1\})^n, \qquad n \in \mathbb{N}.$$

Putting $a = P\{X > 1\} > 0$ we get

$$P\{X > n\} = a^n,$$

hence

$$P\{X = n\} = P\{X > n-1\} - P\{X > n\} = a^{n-1} - a^n = (1-a)a^{n-1}.$$

This shows that X is geometrically distributed with $p = 1-a$ and $q = a$, hence $X \in \mathrm{Pas}(1, 1-a)$.

Example 4.2 *Let X_1, X_2, ... be independent random variables, which are following a geometric distribution,*

$$P\{X_i = k\} = p\, q^{k-1}, \qquad k \in \mathbb{N}; \quad i \in \mathbb{N},$$

where $p > 0$, $q > 0$ and $p + q = 1$.
Prove by induktion that $Y_r = X_1 + X_2 + \cdots + X_r$ has the distribution

$$P\{Y_r = k\} = \binom{k-1}{r-1} p^r q^{k-r}, \qquad k = r,\, r+1,\, \ldots.$$

Find the mean and variance of Y_r.

The claim is obvious for $r = 1$.
If $r = 2$ and $k \geq 2$, then

$$\begin{aligned} P\{Y_2 = k\} &= \sum_{\ell=1}^{k-1} P\{X_1 = \ell\} \cdot P\{X_2 = k-\ell\} = \sum_{\ell=1}^{k-1} p\, q^{\ell-1} p\, q^{k-\ell-1} \\ &= (k-1)p^2 q^{k-2} = \binom{k-1}{1} p^2 q^{k-2} = \binom{k-1}{2-1} p^2 q^{k-2}, \end{aligned}$$

proving that the formula holds for $r = 2$.

Assume that the formula holds for some r, and consider the successor $r+1$ with $Y_{r+1} = X_{r+1} + Y_r$ and $k \geq r+1$. Then

$$\begin{aligned} P\{Y_{r+1} = k\} &= \sum_{\ell=1}^{k-r} P\{X_{r+1} = \ell\} \cdot P\{Y_r = k-\ell\} = \sum_{\ell=1}^{k-r} p\, q^{\ell-1} \binom{k-\ell-1}{r-1} p^r q^{k-\ell-r} \\ &= p^{r+1} q^{k-(r+1)} \sum_{\ell=1}^{k-r} \binom{k-\ell-1}{r-1}. \end{aligned}$$

It follows from Example 2.2 using some convenient substitutions that

$$\sum_{\ell=1}^{k-r} \binom{k-\ell-1}{r-1} = \sum_{j=r-1}^{k-2} \binom{j}{r-1} = \binom{k-1}{r},$$

hence by insertion,

$$P\{Y_{r+1} = k\} = \binom{k-1}{r} p^{r+1} q^{k-(r+1)}.$$

This is exactly the same formula, only with r replaced by $r+1$.
The claim now follows by induction.

Finally,

$$E\{Y_r\} = r\,E\{X_1\} = \frac{r}{p} \quad \text{and} \quad V\{Y_r\} = r\,V\{X_1\} = \frac{rq}{p^2}.$$

Example 4.3 *An (honest) dice is tossed, until we for the first time get a six.*
Find the probability p_n that the first six occurs in toss number n.
Let the random variable X denote the sum of the pips in all the tosses until and included the first toss in which we obtain a six. Find the mean $E\{X\}$.

Clearly,

$$p_n = P\{Y = n\} = \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{n-1}, \qquad n \in \mathbb{N}.$$

If the first six occurs in toss number n, there are only in the first $n-1$ tosses possible to obtain 1, 2, 3, 4, 5 pips, each of the probability $\frac{1}{5}$. Hence the mean of the first $n-1$ tosses is

$$\frac{1}{5}(1+2+3+4+5) = \frac{1}{5} \cdot 15 = 3.$$

In toss number n we get 6 pips, so the expected number of pips in n tosses under the condition that the firs six occurs in toss number n is

$$(n-1) \cdot 3 + 6 = 3(n+1).$$

Then the mean of X is given by

$$\begin{aligned} E\{X\} &= \sum_{n=1}^{\infty} 3(n+1)p_n = \sum_{n=1}^{\infty} 3(n+1) \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{n-1} = \frac{1}{2}\left\{\sum_{n=1}^{\infty} n\left(\frac{5}{6}\right)^{n-1} + \sum_{n=1}^{\infty}\left(\frac{5}{6}\right)^{n-1}\right\} \\ &= \frac{1}{2} \cdot \left\{\frac{1}{\left(1-\frac{5}{6}\right)^2} + \frac{1}{1-\frac{5}{6}}\right\} = \frac{1}{2}\{36+6\} = 21. \end{aligned}$$

Example 4.4 *A box contains N different slips of paper. We then draw randomly one slip from the box and then replace it in the box. Hence all experiments are identical and independent.*
Let the random variable $X_{r,N}$ denote the number of draws until we have got r different slips of paper, $(r = 1, 2, \cdots, N)$. Prove that

$$E\{X_{r,N}\} = N\left\{\frac{1}{N} + \frac{1}{N-1} + \cdots + \frac{1}{N-r+1}\right\}.$$

HINT*: Use that*

$$X_{r,N} = X_{1,N} + (X_{2,N} - X_{1,N}) + \cdots + (X_{r,N} - X_{r-1,N}),$$

and that each of these random variables is geometrically distributed.

Find $E\{X_{r,N}\}$ in the two cases $N = 10$, $r = 5$ and $N = 10$, $r = 10$.
Prove that

$$N \ln\frac{N+1}{N-r+1} < E\{X_{r,N}\} < N \ln\frac{N}{N-r},$$

and show that for r fixed and large N,

$$N \ln\frac{N}{N-r}$$

is a good approximation of $E\{X_{r,N}\}$.

$$\left(\text{*An even better approximation is* } N \ln\frac{N+\frac{1}{2}}{N-r+\frac{1}{2}}\right).$$

Since $X_{k-1,N}$ denotes the number of draws for obtaining $k-1$ different slips, $X_{k,N} - X_{k-1,N}$ indicates the number of draws which should be further applied in order to obtain a new slip. When we have obtained $k-1$ different slips of paper, then we have in the following experiments the probability

$$\frac{N-(k-1)}{N} = \frac{N-k+1}{N}$$

of getting an different slip. This proves that $X_{k,N} - X_{k-1,N}$ is geometrically distributed with

$$p = \frac{N-k+1}{N} \quad \text{and mean} \quad \frac{1}{p} = \frac{N}{N-k+1}, \qquad k = 1, 2, \ldots, N,$$

where we use the convention that $X_{0,N} = 0$, thus $X_{1,N} - X_{0,N} = X_{1,N}$.

The mean is

$$\begin{aligned}\{X_{r,N}\} &= E\{X_{1,N}\} + E\{X_{2,N} - X_{1,N}\} + \cdots + E\{X_{r,N} - X_{r-1,N}\}\\ &= N\left\{\frac{1}{N} + \frac{1}{N-1} + \frac{1}{N-2} + \cdots + \ dfrac1N-r+1\right\}.\end{aligned}$$

Putting for convenience

$$a_{r,N} = \sum_{k=N-r+1}^{N} \frac{1}{k},$$

this is written in the following short version

$$E\{X_{r,N}\} = N \cdot a_{r,N}.$$

If $N = 10$ and $r = 5$, then

$$E\{X_{5,10}\} = 10\left\{\frac{1}{10} + \frac{1}{9} + \frac{1}{8} + \frac{1}{7} + \frac{1}{7}\right\} = 6,46.$$

If $N = 10$ and $r = 10$, then

$$E\{X_{10,10}\} = 10\left\{\frac{1}{10} + \frac{1}{9} + \frac{1}{8} + \cdots + 1\right\} = 29,29.$$

We shall finally estimate

$$a_{r,N} = \sum_{k=N-r+1}^{N} \frac{1}{k}.$$

The sequence $\left(\frac{1}{k}\right)$ is decreasing, hence by the criterion of integral,

$$a_{r,N} > \int_{N-r+1}^{N+1} \frac{1}{x}\,dx = \ln\frac{N+1}{N-r+1}$$

and

$$a_{r,N} < \int_{N-r}^{N} \frac{1}{x}\,dx = \ln\frac{N}{N-r},$$

hence

$$N \cdot \ln\frac{N+1}{N-r+1} < N \cdot a_{r,N} = E\{X_{r,N}\} < N \cdot \frac{N}{N-r}.$$

Remark 4.1 Note that the difference between the upper and the lower bound can be estimated by

$$\begin{aligned} N\left\{\ln\frac{N}{N-r} - \ln\frac{N+1}{N-r+1}\right\} &= N\,\ln\left(\frac{N}{N+1}\cdot\frac{N-r+1}{N-r}\right) \\ &= N\left\{\ln\left(1+\frac{1}{N-r}\right) - \ln\left(1+\frac{1}{N}\right)\right\} \sim N\left\{\frac{1}{N-r} - \frac{1}{N}\right\} \\ &= N\cdot\frac{r}{(N-r)\cdot N} = \frac{r}{N-r} < \frac{r}{N}, \end{aligned}$$

which shows that if r is kept fixed and N is large, then

$$N\,\ln\frac{N+1}{N-r+1} \qquad \text{and} \qquad N\,\ln\frac{N}{N-r}$$

are both reasonable approximations of $E\{X_{r,N}\}$. $\Diamond$

Numerical examples

	$r=5$ $N=10$	$r=5$ $N=20$	$r=20$ $N=150$
$E\{X_{r,N}\}$	6.4563	5.5902	21.3884
$N\,\ln\frac{N+1}{N-r+1}$	6.0614	5.4387	21.3124
$N\,\ln\frac{N+\frac{1}{2}}{N-r+\frac{1}{2}}$	6.4663	5.5917	21.3885

Example 4.5 *A box contains h white balls, r red balls and s black balls ($h > 0$, $r > 0$, $s > 0$). We draw at random a ball from the box and then return it to the box. This experiment if repeated, so the experiments are identical and independent.*
Denote by X the random variable which gives the number of draws, until a white ball occurs for the first time, and let Y denote the number of draws which are needed before a red ball occurs for the first time.

1) Find the distributions of the random variables X and Y.

2) Find the means $E\{X\}$ and $E\{Y\}$.

3) Find for $n \in \mathbb{N}$ the probability that the first white ball is drawn in experiment number n, and that no red ball has occurred in the previous $n-1$ experiment, thus $P\{X = n \wedge Y > n\}$.

4) Find $P\{X < Y\}$.

1) Since X and Y are following geometric distributions, we have

$$P\{X = n\} = \frac{h}{h+r+s}\left(1 - \frac{h}{h+r+s}\right)^{n-1}, \qquad n \in \mathbb{N},$$

$$P\{Y = n\} = \frac{r}{h+r+s}\left(1 - \frac{r}{h+r+s}\right)^{n-1}, \qquad n \in \mathbb{N}.$$

2) Then,

$$E\{X\} = \frac{h+r+s}{h} \qquad \text{and} \qquad E\{Y\} = \frac{h+r+s}{r}.$$

3) When the first white ball is drawn in experiment number n, and no red ball has occurred in the previous $n-1$ experiment, then all drawn balls in the first $n-1$ experiments must be black. Hence we get the probability

$$P\{X = n \wedge Y > n\} = \left(\frac{s}{h+r+s}\right)^{n-1} \cdot \frac{h}{h+r+s}, \qquad n \in \mathbb{N}.$$

4) First solution. It follows by using (3),

$$\begin{aligned} P\{X < Y\} &= \sum_{n=1}^{\infty} P\{X = n \wedge Y > n\} = \sum_{n=1}^{\infty}\left(\frac{s}{h+r+s}\right)^{n-1} \cdot \frac{h}{h+r+s} \\ &= \frac{h+r+s}{h+r} \cdot \frac{h}{h+r+s} = \frac{h}{h+r}. \end{aligned}$$

Second solution. The number of black balls does not enter the problem when we shall find $P\{X < Y\}$, so we may without loss of generality put $s = 0$. Then by the definition of X and Y,

$$P\{X < Y\} = P\{X = 1\} = \frac{h}{h+r}.$$

In fact, if $Y > X$, then X must have been drawn in the first experiment.

Example 4.6 *Given a roulette, where the event of each game is either "red" or "black". We assume that the games are independent. The probability of the event "red" in one game is p, while the probability of the event "black" is q, where $p > 0$, $q > 0$, $p + q = 1$.*
Let the random variable X denote the number of games in the first sequence of games, where the roulette shows the same colour as in the first game.

1) Find for $n \in \mathbb{N}$ the probability that $X = n$.

2) Find the mean of the random variable X.

3) Find the values of p, for which the mean is bigger than 8.

4) Find for $p = \frac{1}{3}$ the variance of the random variable X.

1) We first note that

$$\begin{aligned} P\{X = n\} &= P\{\text{the first } n \text{ games give red and number } (n+1) \text{ gives black}\} \\ &\quad + P\{\text{the first } n \text{ games give black and number } (n+1) \text{ gives red}\} \\ &= p^n q + q^n p = pq\left(p^{n-1} + q^{n-1}\right). \end{aligned}$$

Notice that

$$\sum_{n=1}^{\infty} P\{X = n\} = \sum_{n=1}^{\infty} (P^n q + q^n p) = pq\left(\frac{1}{1-p} + \frac{1}{1-q}\right) = pq\left(\frac{1}{q} + \frac{1}{p}\right) = p + q = 1.$$

2) We infer from

$$\sum_{n=1}^{\infty} n\, z^{n-1} = \frac{1}{(1-z)^2} \qquad \text{for } |z| < 1,$$

that the mean is

$$\begin{aligned} E\{X\} &= pq\left(\sum_{n=1}^{\infty} np^{n-1} + \sum_{n=1}^{\infty} nq^{n-1}\right) = pq\left\{\frac{1}{(1-p)^2} + \frac{1}{(1-q)^2}\right\} = pq\left\{\frac{1}{q^2} + \frac{1}{p^2}\right\} \\ &= \frac{p}{q} + \frac{q}{p} = \frac{p^2+q^2}{pq} = \frac{(p+q)^2 - 2pq}{pq} = \frac{1}{pq} - 2. \end{aligned}$$

3) Now $q = 1 - p$, so

$$E\{X\} = \frac{1}{pq} - 2 = \frac{1}{p(1-p)} - 2$$

is bigger than 8 for

$$\frac{1}{p(1-p)} > 10, \qquad \text{hence for} \quad p^2 - p + \frac{1}{10} > 0.$$

Since $p^2 - p + \frac{1}{10} = 0$ for

$$p = \frac{1}{2} \pm \sqrt{\frac{1}{4} - \frac{1}{10}} = \frac{1}{2} \pm \sqrt{\frac{25-10}{100}} = \frac{1}{2} \pm \frac{\sqrt{15}}{10},$$

we get the two possibilities

$$p \in \left]0, \frac{1}{2} - \frac{\sqrt{15}}{10}\right[\qquad \text{eller} \qquad p \in \left]\frac{1}{2} + \frac{\sqrt{15}}{10}, 1\right[,$$

or approximately,

$$0 < p < 0.1127 \qquad \text{or} \qquad 0.8873 < p < 1.$$

4) In general,

$$\begin{aligned}
V\{X\} &= E\left\{X^2\right\} - (E\{X\})^2 = E\{X(X-1)\} + E\{X\} - (E\{X\})^2 \\
&= \sum_{n=2}^{\infty} n(n-1)\left\{p^n 1 + q^n p\right\} + \frac{p}{q} + \frac{q}{p} - \left(\frac{p}{q} + \frac{q}{p}\right)^2 \\
&= p^2 q \sum_{n=2}^{\infty} n(n-1)p^{n-2} + pq^2 \sum_{n=2}^{\infty} n(n-1)q^{n-2} + \frac{p}{q} + \frac{q}{p} - \frac{p^2}{q^2} - \frac{q^2}{p^2} - 2 \\
&= p^2 q \cdot \frac{2}{(1-p)^3} + pq^2 \cdot \frac{2}{(1-q)^3} + \frac{p}{q} + \frac{q}{p} - \frac{p^2}{q^2} - \frac{q^2}{p^2} - 2 \\
&= 2\,\frac{p^2 q}{q^2} + 2\,\frac{pq^2}{+}\,\frac{p}{q} + \frac{q}{p} - \frac{p^2}{q^2} - \frac{q^2}{p^2} - 2 = \frac{p^2}{q^2} + \frac{q^2}{p^2} + \frac{p}{q} + \frac{q}{p} - 2 \\
&= \frac{p}{q}\left(\frac{p}{q} + 1\right) + \frac{q}{p}\left(\frac{q}{p} + 1\right) - 2 = \frac{p}{q^2} + \frac{q}{p^2} - 2.
\end{aligned}$$

Inserting $p = \frac{1}{3}$ and $q = \frac{2}{3}$ we get

$$V\{X\} = \frac{1}{3} \cdot \frac{9}{4} + \frac{2}{3} \cdot 9 - 2 = \frac{3}{4} + 6 - 2 = \frac{19}{4}.$$

Example 4.7 *We perform a sequence of independent experiments. In each of these there is the probability p of the event A, and the probability $q = 1 - p$ of the event B $(0 < p < 1)$. Such a sequence of experiments may start in the following way,*

$$AAABBABBBBBBAAB \cdots .$$

Every maximal subsequence of the same type is called a run. *In the example above the first run AAA has the length 3, the second run BB has length 2, the third run A has length 1, the fourth run $BBBBBB$ has length 6, etc.*
Denote by $\widehat{X_n}$ the length of the n-th run, $n \in \mathbb{N}$.

1) Explain why the random variables X_n, $n \in \mathbb{N}$, are independent.

2) Find $P\{X_1 = k\}$, $k \in \mathbb{N}$.

3) Find $E\{X_1\}$ and $V\{X_1\}$.

4) Find $P\{X_2 = k\}$, $k \in \mathbb{N}$.

5) Find $E\{X_2\}$ and $V\{X_2\}$.

6) Prove that

$$E\{X_2\} \leq E\{X_1\} \quad \text{and} \quad V\{X_2\} \leq V\{X_1\}.$$

1) The probability of the event $\{X_n = k\}$ depends only of the first element of the n-th run, proving that X_n, $n \in \mathbb{N}$, are independent.

2) It follows like in Example 4.6 that (k rersults of A, resp. B)

$$E\{X_1\} = P\{A\cdots AB\} + P\{B\cdots BA\} = p^k q + q^k p = pq\left(p^{k-1} + q^{k-1}\right).$$

3) The mean is, cf. Example 4.6,

$$E\{X_1\} = \sum_{k=1}^{\infty} pq\cdot k\left(p^{k-1} + q^{k-1}\right) = pq\left\{\frac{1}{(1-p)^2} + \frac{1}{(1-q)^2}\right\} = \frac{p}{q} + \frac{q}{p} = \frac{p^2+q^2}{pq} = \frac{1}{pq} - 2.$$

Furthermore,

$$\begin{aligned} E\{X_1(X_1-1)\} &= \sum_{k=2}^{\infty} k(k-1)p^k q + \sum_{k=2}^{\infty} k(k-1)q^k p = p^2 q\sum_{k=2}^{\infty} k(k-1)p^{k-2} + pq^2\sum_{k=2}^{\infty} k(k-1)q^{k-2} \\ &= p^2 q\cdot\frac{2}{q^3} + pq^2\cdot\frac{2}{p^3} = 2\left\{\frac{p^2}{q^2} + \frac{q^2}{p^2}\right\}, \end{aligned}$$

hence

$$\begin{aligned} V\{X_1\} &= E\{X_1(X_1-1)\} + E\{X_1\} - (E\{X_1\})^2 = 2\left\{\frac{p^2}{q^2} + \frac{q^2}{p^2}\right\} + \frac{p}{q} + \frac{q}{p} - \left(\frac{p}{q} + \frac{q}{p}\right)^2 \\ &= \frac{p^2}{q^2} + \frac{q^2}{p^2} + \frac{p}{q} + \frac{q}{p} - 1 = \frac{p}{q^2} + \frac{q}{p^2} - 2. \end{aligned}$$

4) If we have ℓ copies of A and k copies of B in the first sum, and ℓ copies of B and k copies of A ib the second sum, then

$$\begin{aligned} P\{X_2 = k\} &= \sum_{\ell=1}^{\infty} P\{A\cdots AB\cdots BA\} + \sum_{\ell=1}^{\infty} P\{B\cdots BA\cdots AB\} = \sum_{\ell=1}^{\infty} p^\ell q^k\cdot p + \sum_{\ell}^{\infty} q^\ell p^k q \\ &= p^2 q^k\sum_{\ell=1}^{\infty} p^{\ell-1} + q^2 p^k\sum_{\ell=1}^{\infty} q^{\ell-1} = \frac{p^2 q^k}{1-p} + \frac{q^2 p^k}{1-q} = p^2 q^{k-1} + q^2 p^{k-1}. \end{aligned}$$

5) The mean is

$$E\{X_2\} = \sum_{k=1}^{\infty} kp^2 q^{k-1} + \sum_{k=1}^{\infty} kq^2 q^{k-1} = \frac{p^2}{(1-q)^2} + \frac{q^2}{(1-p)^2} = \frac{p^2}{p^2} + \frac{q^2}{q^2} = 2.$$

Furthermore,

$$\begin{aligned} E\{X_2(X_2-1)\} &= \sum_{k=2}^{\infty} k(k-1)q^2 q^{k-1} + \sum_{k=2}^{\infty} k(k-1)q^2 p^{k-1} \\ &= p^2 q\sum_{k=2}^{\infty} k(k-1)q^{k-2} + pq^2\sum_{k=2}^{\infty} k(k-1)p^{k-2} \\ &= p^2 q\cdot\frac{2}{(1-q)^3} + pq^2\cdot\frac{2}{(1-p)^3} = \frac{2p^2 q}{p^3} + \frac{2pq^2}{q^3} = 2\left(\frac{p}{q} + \frac{q}{p}\right), \end{aligned}$$

hence

$$\begin{aligned} V\{X_2\} &= E\{X_2(X_2-1)\} + E\{X_2\} - (E\{X_2\})^2 \\ &= 2\left\{\frac{p}{q} + \frac{q}{p}\right\} + 2 - 4 = 2\left\{\frac{p}{q} + \frac{q}{p} - 1\right\}. \end{aligned}$$

6) Now $\varphi(x) = x + \frac{1}{x}$ has a minimum for $x = 1$, so if we put $x = \frac{p}{q}$, then

$$E\{X_1\} = \frac{p}{q} + \frac{q}{p} = \frac{p}{q} + \frac{1}{\frac{p}{q}} \geq 1 + 1 = 2 = E\{X_2\}.$$

It follows from

$$V\{X_1\} = x^2 + \frac{1}{x^2} + x + \frac{1}{x} - 2$$

and

$$V\{X_2\} = 2x + \frac{2}{x} - 2,$$

that

$$\begin{aligned} V\{X_1\} - V\{X_2\} &= x^2 + \frac{1}{x^2} - x - \frac{1}{x} = x(x-1) + \frac{1-x}{x^2} = \frac{(x-1)\left(x^3-1\right)}{x^2} \\ &= \left(\frac{x-1}{x}\right)^2 \left(x^2+x+1\right) \geq 0, \end{aligned}$$

hence $V\{X_1\} \geq V\{X_2\}$. Equality is only obtained for $x = \frac{p}{q} = 1$, i.e. for

$$p = q = \frac{1}{2}.$$

Example 4.8 *We toss a sequence of tosses with an (honest) dice. The tosses are identical and independent. We define a random variable X of values in $\mathbb{N}_0$ by*

$X = n$, *if the first six is in toss number* $n+1$.

1) Find the distribution and mean of X.

2) Find for $k = 0, 1, 2, \ldots, n$, the probability that there in n tosses are precisely k fives and no six.

3) Find the probability $p_{n,k}$ that the first six occurs in toss number $n+1$, and that we have had precisely k fives, $n = k, k+1, \ldots$ in the first n tosses.

4) Find for every $k \in \mathbb{N}_0$ the probability p_k that we have got exactly k fives before we get the first six.

5) Find the expected number of fives before we get the first six.

1) It is immediately seen that

$$P\{X = n\} = \left(\frac{5}{6}\right)^n \cdot \frac{1}{6} = \frac{5^n}{6^{n+1}}, \qquad n \in \mathbb{N}_0.$$

Hereby we get the mean

$$E\{X\} = \sum_{n=0}^{\infty} n\, P\{X = n\} = \frac{5}{6^2} \sum_{n=1}^{\infty} n \left(\frac{5}{6}\right)^{n-1} = \frac{5}{6^2} \cdot \frac{1}{\left(1 - \frac{5}{6}\right)^2} = 5.$$

2) If in n tosses there are k fives and no six, this is obtained by choosing k places out of n, which can be done in $\left(\begin{array}{c} n \\ k \end{array} \right)$ sways. The remaining $n-k$ tosses are then chosen among the numbers $\{1,2,3,4\}$, so

$$P\{k \text{ fives and no six in } n \text{ tosses}\} = \left(\begin{array}{c} n \\ k \end{array} \right) \left\{\frac{1}{6}\right\}^k \left\{\frac{4}{6}\right\}^{n-k},$$

for $k = 0, 1, \ldots, n$.

3) If $n \geq k$, then

$$\begin{aligned} p_{n,k} &= P\{k \text{ fives and no six in } n \text{ tosses}\} \times P\{\text{one six in toss number } n+1\} \\ &= \binom{n}{k} \cdot \left(\frac{1}{6}\right)^k \cdot \left(\frac{4}{6}\right)^{n-k} \cdot \frac{1}{6} = \binom{n}{k} \left(\frac{1}{6}\right)^{k+1} \left(\frac{2}{3}\right)^{n-k}. \end{aligned}$$

4) This probability is

$$\begin{aligned} p_k &= \sum_{n=k}^{\infty} p_{n,k} = \frac{1}{4^k \cdot 6} \sum_{n=k}^{\infty} \binom{n}{k} \left\{\frac{2}{3}\right\}^n = \frac{1}{4^k \cdot 6} \sum_{n=k+1}^{\infty} \binom{n-1}{k} \left\{\frac{2}{3}\right\}^{n-1} \\ &= \frac{1}{4^k \cdot 6} \sum_{n=k+1}^{\infty} \binom{n-1}{(k+1)-1} \left\{\frac{2}{3}\right\}^{n-(k+1)} \left\{\frac{1}{3}\right\}^{k+1} \left\{\frac{2}{3}\right\}^{k} \left\{\frac{1}{3}\right\}^{-k-1} \\ &= \frac{1}{4^k \cdot 6} \cdot 3 \cdot 2^k \sum_{n=k+1}^{\infty} \binom{n-1}{(k+1)-1} \left\{\frac{1}{3}\right\}^{k+1} \left\{\frac{2}{3}\right\}^{n-(k+1)} \\ &= \frac{1}{2^{k+1}}, \qquad k \in \mathbb{N}_0, \end{aligned}$$

because the sum is the total probability that $Y \in \mathrm{Pa}\left(k+1, \frac{1}{3}\right)$ occurs, hence $= 1$.

An ALTERNATIVE computation is the following

$$\begin{aligned} p_k &= \sum_{n=k}^{\infty} p_{n.k} = \left(\frac{1}{6}\right)^{k+1} \sum_{n=k}^{\infty} \binom{n}{k} \left(\frac{2}{3}\right)^{n-k} = \left(\frac{1}{6}\right)^{k+1} \sum_{\ell=0}^{\infty} \binom{\ell+k}{k} \left(\frac{2}{3}\right)^{\ell} \\ &= \left(\frac{1}{6}\right)^{k+1} \sum_{\ell=0}^{\infty} \binom{\ell+k}{\ell} \left(\frac{2}{3}\right)^{\ell} = \left(\frac{1}{6}\right)^{k+1} \left(1-\frac{2}{3}\right)^{-(k+1)} = \left(\frac{1}{2}\right)^{k+1}, \qquad k \in \mathbb{N}_0. \end{aligned}$$

ALTERNATIVELY we see that every other toss than just fives or sixes are not relevent for the question. If we neglect these tosses then a five and a six occur with each the probability $\frac{1}{2}$. Then the probability of getting exactly k fives before the first six is

$$\left(\frac{1}{2}\right)^k \cdot \frac{1}{2} = \left(\frac{1}{2}\right)^{k+1}, \qquad k \in \mathbb{N}_0.$$

5) The expected number of fives coming before the first six is then

$$\sum_{k=0}^{\infty} k\, p_k = \sum_{k=1}^{\infty} \frac{k}{2^{k+1}} = \frac{1}{4} \sum_{k=1}^{\infty} k \left(\frac{1}{2}\right)^{k-1} = \frac{1}{4} \cdot \frac{1}{\left(1-\frac{1}{2}\right)^2} = 1.$$

Here ones usual intuition fails, because one would guess that the expected number is < 1.

Example 4.9 *Let X_1 and X_2 be independent random variables, both following a geometric distribution*

$$P\{X_i = k\} = pq^{k-1}, \qquad k \in \mathbb{N}, \qquad i = 1, 2.$$

Let $Y = \min\{X_1, X_2\}$.

1) Find the distribution of Y.

2) Find $E\{Y\}$ and $V\{Y\}$.

1) If $k \in \mathbb{N}$, then

$$\begin{aligned} P\{Y = k\} &= P\{X_1 = k,\, X_2 > k\} + P\{X_1 > k,\, X_2 = k\} + P\{X_1 = k,\, X_2 = k\} \\ &= 2\,P\{X_1 = k\} \cdot P\{X_2 > k\} + (P\{X_1 = k\})^2 \\ &= 2pq^{k-1} \cdot q^k + p^2 q^{2k-2} = q^{2k-2}\left\{2pq + p^2\right\} \\ &= \left(q^2\right)^{k-1} \cdot \left\{(p+q)^2 - q^2\right\} = \left(1 - q^2\right)\left(q^2\right)^{k-1}, \end{aligned}$$

proving that Y is geometrically distributed, $Y \in \mathrm{Pas}\left(1, q^2\right)$.

Alternatively we first compute

$$P\{Y > k\} = P\{X_1 > k,\, X_2 > k\} = P\{X_1 > k\} \cdot P\{X_2 > k\} = q^{2k},$$

and use that

$$P\{Y > k-1\} = P\{Y = k\} + P\{Y > k\},$$

hence by a rearrangement,

$$\begin{aligned} P\{Y = k\} &= P\{Y > k-1\} - P\{Y > k\} = q^{2k-2} - q^{2k} \\ &= \left(1 - q^2\right)\left(q^2\right)^{k-1}, \qquad \text{for } k \in \mathbb{N}, \end{aligned}$$

and it follows that Y is geometrically distributed, $Y \in \mathrm{Pa}\left(1, q^2\right)$.

2) Using a formula and replacing p by $1 - q^2$, and q by q^2, we get

$$E\{Y\} = \frac{1}{1-q^2} \qquad \text{and} \qquad V\{Y\} = \frac{q^2}{\left(1-q^2\right)^2}.$$

Example 4.10 *Let X_1 and X_2 be independent random variables, both following a geometric distribution,*

$$P\{X_i = k\} = p\,q^{k-1}, \qquad k \in \mathbb{N}, \qquad i = 1, 2.$$

Put $Z = \max\{X_1, X_2\}$.

1) Find $P\{Z = k\}$, $k \in \mathbb{N}$.

2) Find $E\{Z\}$.

1) A small consideration gives

$$P\{Z \leq n\} = P\{\max(X_1, X_2) \leq n\} = P\{X_1 \leq n\} \cdot P\{X_2 \leq n\},$$

which is also written

$$\sum_{k=1}^{n} P\{Z = k\} = \sum_{\ell=1}^{n} P\{X_1 = \ell\} \cdot \sum_{m=1}^{n} P\{X_2 = m\}.$$

Then

$$\begin{aligned}
P\{Z=b\} &= \sum_{k=1}^{n} P\{Z=k\} - \sum_{k=1}^{n-1} P\{Z=k\} \\
&= \left\{\sum_{\ell=1}^{n-1} P\{X_1=\ell\} + P\{X_1=n\}\right\} \cdot \left\{\sum_{m=1}^{n-1} P\{X_2=m\} + P\{X_2=n\}\right\} \\
&\qquad - \sum_{\ell=1}^{n-1} P\{X_1=\ell\} \cdot \sum_{m=1}^{n-1} P\{X_2=m\} \\
&= P\{X_1=n\} \cdot \sum_{m_1}^{n} P\{X_2=m\} + P\{X_2=n\} \cdot \sum_{\ell=1}^{n-1} P\{X_1=\ell\} \\
&= pq^{n-1} \sum_{k=1}^{n} pq^{k-1} + pq^{n-1} \sum_{k=1}^{n-1} pq^{k-1} \\
&= p^2 q^{n-1} \sum_{k=1}^{n} q^{k-1} + p^2 q^{n-1} \sum_{k=1}^{n-1} q^{k-1} \\
&= p^2 q^{n-1} \cdot \frac{1-q^n}{1-q} + p^2 q^{n-1} \cdot \frac{1-q^{n-1}}{1-q} \\
&= pq^{n-1}\left\{1-q^n+1-q^{n-1}\right\} = p\left\{2q^{n-1} - q^{2n-2} - q^{2n-1}\right\}.
\end{aligned}$$

Check:

$$\begin{aligned}
\sum_{n=1}^{\infty} P\{Z=n\} &= 2p \sum_{n=1}^{\infty} q^{n-1} - p \sum_{n=1}^{\infty} \left(q^2\right)^{n-1} - pq \sum_{n_1}^{\infty} \left(q^2\right)^{n-1} \\
&= 2p \cdot \frac{1}{1-q} - p(1+q) \cdot \frac{1}{1-q^2} \\
&= \frac{2p}{p} - \frac{p(1+q)}{(1-q)(1+q)} = 2-1 = 1,
\end{aligned}$$

proving that the probabilities found here actually are summed up to 1. ◊

2) From

$$\sum n = 1^{\infty} n\, z^{n-1} = \frac{1}{(1-z)^2} \qquad \text{for } |z|<1,$$

follows that

$$\begin{aligned}
E\{Z\} &= \sum_{n=1}^{\infty} n\, P\{Z=n\} = 2p \sum_{n=1}^{\infty} n\, q^{n-1} - (1+q)p \sum_{n=1}^{\infty} n \left(q^2\right)^{n-1} \\
&= 2p \cdot \frac{1}{(1-q)^2} - (1+q)p \cdot \frac{1}{\left(1-q^2\right)^2} = \frac{2p}{p^2} - \frac{(1+q)p}{\{(1+q)p\}^2} \\
&= \frac{2}{p} - \frac{1}{(1+q)p} = \frac{2+2q-1}{p(1+q)} = \frac{1+2q}{1-q^2}.
\end{aligned}$$

Example 4.11 *Let X_1 and X_2 be independent random variables, both following a geometric distribution,*

$$P\{X_i = k\} = pq^{k-1}, \qquad k \in \mathbb{N}; \qquad i = 1, 2.$$

Lad $Y = X_1 - X_2$.

1) Find $P\{Y = k\}$, $k \in \mathbb{Z}$.

2) Find $E\{Y\}$ and $V\{Y\}$.

1) Since X_1 and X_2 are independent, we get

$$P\{Y = k\} = \sum_i P\{X_1 = k + i \wedge X_2 = i\} = \sum_i P\{X_1 = k + i\} \cdot P\{X_2 = i\}.$$

Then we must split the investigations into two cases:

a) If $k \geq 0$, then

$$\begin{aligned} P\{Y = k\} &= \sum_{i=1}^{\infty} pq^{k+i-1} \cdot pq^{i-1} = \sum_{i=1}^{\infty} p^2 q^k q^{2i-2} = p^2 q^k \sum_{i=1}^{\infty} \left(q^2\right)^{i-1} = \frac{p^2}{1-q^2}\, q^k \\ &= \frac{p}{1+q}\, q^k = \frac{1-q}{1+q}\, q^k. \end{aligned}$$

b) If instead $k < 0$, then

$$\begin{aligned} P\{Y = k\} &= \sum_{i=-k+1}^{\infty} pq^{k+i-1} pq^{i-1} = \sum_{i=-k+1}^{\infty} p^2 q^k \left(q^2\right)^{i-1} = p^2 q^k \cdot \frac{\left(q^2\right)^{-k}}{1-q^2} \\ &= \frac{p^2}{1-q^2} \cdot q^{-k} = \frac{p}{1+q}\, q^{-k} = \frac{1-q}{1+q}\, q^{-k}. \end{aligned}$$

Alternatively if follows for $k < 0$ by the symmetry that

$$\begin{aligned} P\{Y = k\} &= P\{X_1 - X_2 = -|k|\} = P\{X_2 - X_1 = |k|\} \\ &= \frac{p}{1+q}\, q^{|k|} = \frac{p}{1+q}\, q^k \qquad \text{for } k \in \mathbb{N}_0. \end{aligned}$$

Summing up,

$$P\{Y = -k\} = P\{Y = k\} = \frac{p}{1+q}\, q^k \qquad \text{for } k \in \mathbb{N}_0.$$

2) The mean exists and is given by

$$E\{Y\} = E\{X_1 - X_2\} = E\{X_1\} - E\{X_2\} = 0.$$

Since X_1 and X_2 are independent, the variance is

$$V\{Y\} = V\{X_1 - X_2\} = V\{X_1\} + V\{-X_2\} = V\{X_1\} + V\{X_2\} = \frac{2q}{p^2}.$$

Example 4.12 *A man has N keys, of which only one fits into one particular door. He tries the keys one by one:*
Let X be the random variable, which indicates the number of experiments until he is able to unlock the door. Find $E\{X\}$.
If he instead of each time taking a key at random (without bothering with, if the key already has been tested) put the tested keys aside , how many tries should he use at the average?

1) In the first case we have a uniform distribution,

$$P\{X = k\} = \frac{1}{N} \qquad \text{for } k = 1,\, 2,\, \ldots,\, N,$$

hence

$$E\{X\} = \frac{1}{N}\sum_{k=1}^{N} k = \frac{N+1}{2}.$$

2) In this case the corresponding random variable is geometrically distributed with $p = \frac{1}{N}$, hence

$$P\{Y = k\} = \frac{1}{N}\left(1 - \frac{1}{N}\right)^{k-1}, \quad k \in \mathbb{N}, \qquad Y \in \text{Pas}\left(1, 1 - \frac{1}{N}\right).$$

Then finally, by some formula,

$$E\{Y\} = N.$$

5 The Pascal distribution

Example 5.1 *Let X and Y be random variables, both having values in $\mathbb{N}_0$. We say that the random variable X is* stochastically larger than Y, *if*

$$P\{X > k\} \geq P\{Y > k\} \qquad \text{for every } k \in \mathbb{N}_0.$$

1) Prove that if X is stochastically larger than Y, and if $E\{X\}$ exists, then

$$E\{Y\} \leq E\{X\}.$$

2) Let $X \in \text{Pas}\,(r, p_1)$ and $Y \in \text{Pas}\,(r, p_2)$, where $r \in \mathbb{N}$ and $0 < p_1 < p_2 < 1$. Prove that X is stochastically larger than Y.

1) We use that

$$E\{X\} = \sum_{k=0}^{\infty} P\{X > k\},$$

and analogously for Y. Then

$$E\{X\} = \sum_{k=0}^{\infty} P\{X > k\} \geq \sum_{k=0}^{n} P\{Y > k\} = E\{Y\}.$$

2) Intuitively the result is obvious, when one thinks of the waiting time distribution. This is, however, not so easy to prove, what clearly follows from the following computations.
Obviously, $P\{X > k\} \geq P\{Y > k\}$ is equivalent to

$$P\{X \leq k\} \leq P\{Y \leq k\}.$$

Hence, we shall prove or the Pascal distribution that the function

$$\varphi_{r,k}(p) = \sum_{k=r}^{k} \binom{j-1}{r-1} p^r (1-p)^{j-r}$$

is increasing in $p \in]0,1[$ for every $r \in \mathbb{N}$ and $k \geq r$.
Writing more traditionally x instead of p, we get

$$\begin{aligned} \varphi_{r,k}(x) &= \sum_{j=r}^{k} \binom{j-1}{r-1} x^r (1-x)^{j-r} = x^r \sum_{j=0}^{k-r} \binom{j+r-1}{r-1} (1-x)^j \\ &= x^r \sum_{j=0}^{k-r} \binom{j+r-1}{j} (1-x)^j. \end{aligned}$$

We put a convenient index on X, such that we get the notation

$$P\{X_r \leq k\} = \varphi_{r,k}(x), \qquad x = p \in]0,1[.$$

Then by a differentiation,

$$\begin{aligned}
\varphi'_{r,k}(x) &= \frac{r}{x}\,\varphi_{r,k}(x) - x^r \sum_{j=0}^{k-r} j \binom{j+r-1}{j} (1-x)^{j-1} \\
&= \frac{r}{x}\,\varphi_{r,k}(x) - r\,x^r \sum_{j=1}^{k-r} \binom{j+r-1}{j-1} (1-x)^{j-1} \\
&= \frac{r}{x}\,\varphi_{r,k}(x) - \frac{r}{x}\,x^{r+1} \sum_{j=0}^{k-r-1} \binom{j+r}{j} (1-x)^{j} \\
&= \frac{r}{x}\left\{ \varphi_{r,k}(x) - x^{r+1} \sum_{j=0}^{k-(r+1)} \binom{j+(r+1)-1}{j} (1-x)^{j} \right\} \\
&= \frac{r}{x}\left\{ \varphi_{r,k}(x) - \varphi_{r+1,k}(x) \right\} \\
&= \frac{r}{x}\left(P\left\{ X_r \le k \right\} - P\left\{ X_{r+1} \le k \right\} \right).
\end{aligned}$$

The event $\{X_{r+1} \le k\}$ corresponds to that success number $(r+1)$ comes at the latest in experiment number k. The probability is $\le$ the probability that success number r occurs at the latest in experiment number k, thus

$$\varphi'_{r,k}(x) = \frac{r}{x}\left(P\left\{ X_r \le k \right\} - P\left\{ X_{r+1} \le k \right\} \right) \ge 0, \qquad x \in]0,1[,$$

so $\varphi_{r,k}(x)$ is increasing in x. Therefore, if $X \in \mathrm{Pas}(r, p_1)$ and $Y \in \mathrm{Pas}(r, p_2)$, where $r \in \mathbb{N}$ and $0 < p_1 < p_2 < 1$, then

$$P\{X \le k\} \le P\{Y \le k\} \qquad \text{for } k \ge r.$$

This is equivalent to the fact that X is stochastically larger than Y.

Alternatively, the result can be proved by *induktion after* r. If $r = 1$, then

$$P\{X > k\} = (1-p_1)^k > (1-p_2)^k = P\{Y > k\},$$

and the result is proved for $r = 1$.

When $r > 1$, then we write

$$X = \sum_{i=1}^{r} X_i \qquad \text{and} \qquad Y = \sum_{i=1}^{r} Y_i,$$

where

$$X_i \in \mathrm{Pas}\,(1, p_2) \qquad \text{and} \qquad Y_i \in \mathrm{Pas}\,(1, p_2)\,,$$

and where the X_i-s (resp. the Y_i-s) are mutually independent.

The condition

$$P\{X > k\} \ge P\{Y > k\}, \qquad \text{for every } k \in \mathbb{N}_0,$$

does only concern the distributions of X, resp. Y. Therefore, we can assume that X and Y are independent, and that all the X_i-s and Y_j-s above also are independenty.

We shall prove the following lemma:

If X_1, X_2, Y_1, Y_2 are independent random variables with values in $\mathbb{N}_0$, where X_1 is stochastically larger than Y_1, and X_2 is stochastically larger than Y_2, then $X_1 + X_2$ is stochastically larger that $Y_1 + Y_2$.

PROOF. If $k \in \mathbb{N}_0$, then

$$\begin{aligned}
P\{X_1 + X_2 > k\} &= \sum_{i=0}^{\infty} P\{X_2 = i \wedge X_1 > k - i\} \\
&= \sum_{i=0}^{\infty} P\{X_2 = i\} \cdot P\{X_1 > k - i\} \\
&\geq \sum_{i=0}^{\infty} P\{X_2 = i\} \cdot P\{Y_1 > k - i\} \\
&= P\{Y_1 + X_2 > k\} = \sum_{i=0}^{\infty} P\{Y_1 = i\} \cdot P\{X_2 > k - i\} \\
&\geq \sum_{i=0}^{\infty} P\{Y_1 = i\} \cdot P\{Y_2 > k - i\} = P\{Y_1 + Y_2 > k\},
\end{aligned}$$

and the claim is proved. □

Then write

$$X = \left(\sum_{i=1}^{r-1} X_i\right) + X_r \qquad \text{and} \qquad Y = \left(\sum_{i=1}^{r-1} X_i\right) + Y_r.$$

It follows form the assumption of induction that $\sum_{i=1}^{r-1} X_i$ is stochastically larger than $\sum_{i=1}^{r-1} Y_i$. Since X_r also is stochastically larger than Y_r, it follows from the result above that X is stochastically larger than Y.

6 The negative binomial distribution

Example 6.1 *A random variable X has its distribution given by*

$$P\{X=k\} = \binom{-\kappa}{k} p^{\kappa}(-q)^k, \qquad k \in \mathbb{N}_0,$$

where $p>0$, $q>0$, $p+q=1$ and $\kappa>0$, thus $X \in \mathrm{NB}(\kappa,p)$.
Find the mean and variance of X.

First note that

$$(-1)^n \binom{-\kappa}{n} = \binom{n+\kappa-1}{n},$$

hence the mean is given by

$$\begin{aligned} E\{X\} &= \sum_{n=1}^{\infty} n \binom{n+\kappa-1}{n} p^{\kappa} q^n = \sum_{n=1}^{\infty} \kappa \binom{n+\kappa-1}{n-1} p^{\kappa} q^n = n \sum_{n=0}^{\infty} \binom{n+\kappa}{n} q^{\kappa} q^{n+1} \\ &= \frac{\kappa q}{p} \sum_{n=0}^{\infty} \binom{n+\{\kappa+1\}-1}{n} p^{\kappa+1} q^n = \frac{\kappa q}{p}. \end{aligned}$$

Furthermore,

$$\begin{aligned} E\{X(X-1)\} &= \sum_{n=2}^{\infty} n(n-1) \left(\begin{array}{c} n+\kappa-1 \\ n \end{array} \right) p^{\kappa} q^n = \sum_{n=2}^{\infty} \kappa(\kappa+1) \left(\begin{array}{c} n+\kappa-1 \\ n-2 \end{array} \right) p^{\kappa} q^n \\ &= \kappa(\kappa+1) \sum_{n=0}^{\infty} \left(\begin{array}{c} n+\{\kappa+2\}-1 \\ n \end{array} \right) p^{\kappa} q^{n+2} \\ &= \kappa(\kappa+1) \sum_{n=0}^{\infty} \left(\begin{array}{c} n+\{\kappa+2\}-1 \\ n \end{array} \right) p^{\kappa+2} q^n = \kappa(\kappa+1) \cdot \frac{q^2}{p^2}, \end{aligned}$$

whence

$$\begin{aligned} V\{X\} &= E\{X(X-1)\} + E\{X\} - (E\{X\})^2 = \kappa(\kappa+1)\,\frac{q^2}{p^2} + \kappa\,\frac{q}{p} - \kappa^2\,\frac{q^2}{p^2} \\ &= \kappa \left(\frac{q^2}{p^2} + \frac{q}{p} \right) = \frac{\kappa q}{p^2}\,(q+p) = \frac{\kappa q}{p^2}. \end{aligned}$$

7 The hypergeometric distribution

Example 7.1 BANACH'S MATCH STICK PROBLEM. *A person B has the habit of using two boxes of matches at the same time. We assume that a matchbox contains 50 matches. When B shall apply a match, he chooses at random one of the two matchboxes without noticing afterwards if it is empty. Let X denote the number of matches in one of the boxes, when we* discover *that the other one is empty, and let Y denote the number of matches in the first box, when the second one is* emptied.
It can be proved that

$$a_r = P\{X = r\} = \binom{100-r}{50}\left(\frac{1}{2}\right)^{100-r}, \qquad r = 0,\, 1,\, \ldots,\, 50.$$

Find analogously

$$b_r = P\{Y = r\}, \qquad r = 1,\, 2,\, \ldots,\, 50.$$

For completeness we give the proof of the result on a_r.

We compute $P\{X = 50-k\}$, thus we shall *use* 50 matches form one of the boxes and k matches from the other one, and then in choice number $50+k+1 = 51+k$ choose the empty box.

In the first $50+k$ choices we shall choose the box which is emptied in total 50 times, corresponding to the probability

$$\binom{50+k}{50}\left(\frac{1}{2}\right)^{50}\left(\frac{1}{2}\right)^{k} = \binom{50+k}{k}\left(\frac{1}{2}\right)^{50+k}.$$

In the next choice we shall choose the empty box, which can be done with the probability $\frac{1}{2}$. Finally, we must choose between 2 boxes, so we must multiply by 2. Summing up,

$$P\{X = 50-k\} = \binom{50+k}{k}\left(\frac{1}{2}\right)^{50+k}, \qquad k = 0,\, 1,\, \ldots,\, 50.$$

Then by the substitution $r = 50-k$,

$$a_r = P\{X = r\} = \binom{100-r}{50}\left(\frac{1}{2}\right)^{100-r}, \qquad r = 0,\, 1,\, \ldots,\, 50.$$

In order to find b_r we shall compute $P\{Y = 50-k\}$, i.e. we shall *use* 49 matches from one of the boxes and k matches from the other one, and then *choose* the box, in which there is only 1 match left. Analogously to the above we get

$$P\{Y = 50-k\} = \binom{49+k}{k}\left(\frac{1}{2}\right)^{49+k}, \qquad k = 0,\, 1,\, \ldots,\, 50.$$

Then by the substitution $r = 50-k$,

$$b_r = P\{Y = r\} = \binom{99-r}{49}\left(\frac{1}{2}\right)^{99-r}, \qquad r = 0,\, 1,\, \ldots,\, 50.$$

Example 7.2 . *(Continuation of Example 7.1).*

1) Find an explicit expression for the mean μ of the random variable X of Example 7.1.

2) Find, by using the result of Example 2.3, an approximate expression of the mean μ.

HINT TO QUESTION 1: *Start by reducing the expression*

$$50 - \mu = \sum_{r=0}^{50} (50 - r) a_r.$$

It follows from Example 7.1 that

$$a_r = P\{X = r\} = \left(\begin{array}{c} 100 - r \\ 50 \end{array} \right) \left(\frac{1}{2} \right)^{100-r}, \qquad r = 0, 1, \ldots, 50.$$

The text of the example is confusing for several reasons:

(a) On the pocket calculator TI-92 the mean is easily computed by using the command

$$\sum(r \star nCr(100 - r, 50) \star 2\hat{}(r - 100), r, 0, 50),$$

corresponding to

$$\sum_{r=0}^{50} r \left(\begin{array}{c} 100 - r \\ 50 \end{array} \right) \left(\frac{1}{2} \right)^{100-r} \approx 7.03851.$$

The pocket calculator is in fact very fast in this case.

(b) The hint looks very natural. The result, however, does not look like any expression containing $\left(\begin{array}{c} 2n \\ n \end{array} \right)$, which one uses in the approximation in Example 2.3. I have tried several variants, of which the following is the best one:

1) By using the hint we get

$$50 - \mu = \sum_{r=0}^{50} (50 - r) a_r = \sum_{r=0}^{49} (50 - r) \left(\begin{array}{c} 100 - r \\ 50 \end{array} \right) \left(\frac{1}{2} \right)^{100-r}.$$

Then by Example 2.2,

$$\left(\begin{array}{c} n + 1 \\ r + 1 \end{array} \right) = \sum_{k=r}^{n} \left(\begin{array}{c} k \\ r \end{array} \right), \qquad r, n \in \mathbb{N}_0, \quad r \leq n.$$

When we insert this result, we get

$$\begin{aligned} (50 - r) \left(\begin{array}{c} 100 - r \\ 50 \end{array} \right) &= (50 - r) \frac{(100 - r)!}{50!(50 - r)!} = 51 \cdot \frac{(100 - r)!}{51!(49 - r)!} = 51 \left(\begin{array}{c} 100 - r \\ 51 \end{array} \right) \\ &= 51 \sum_{k=50}^{99-r} \left(\begin{array}{c} k \\ 50 \end{array} \right) = 51 \sum_{j=0}^{49-r} \left(\begin{array}{c} 50 + j \\ 50 \end{array} \right) = 51 \sum_{j=0}^{49-r} \left(\begin{array}{c} 50 + j \\ j \end{array} \right). \end{aligned}$$

Then continue with the following

$$\begin{aligned}
50-\mu &= \sum_{r=0}^{49}(50-r)\binom{100-r}{50}\left(\frac{1}{2}\right)^{100-r} = 51\sum_{r=0}^{49}\binom{100-r}{51}\left(\frac{1}{2}\right)^{100-r}\\
&= 51\sum_{r=0}^{49}\sum_{j=0}^{49-r}\binom{50+j}{50}\left(\frac{1}{2}\right)^{100-r} = 51\sum_{j=0}^{49}\binom{50+j}{50}\sum_{r=0}^{49-j}\left(\frac{1}{2}\right)^{100-r}\\
&= 51\sum_{j=0}^{49}\binom{50+j}{50}\left\{\left(\frac{1}{2}\right)^{51+j}+\cdots+\left(\frac{1}{2}\right)^{100}\right\}\\
&= 51\sum_{j=0}^{49}\binom{50+j}{50}\left(\frac{1}{2}\right)^{50+j} - 51\cdot\left(\frac{1}{2}\right)^{100}\sum_{j=0}^{49}\binom{50+j}{50}\\
&= 51\sum_{j=0}^{49}\binom{50+j}{50}\left(\frac{1}{2}\right)^{50+j} - 51\cdot\left(\frac{1}{2}\right)^{100}\sum_{k=50}^{99}\binom{k}{50}\\
&= 51\sum_{r=1}^{50}\binom{100-r}{50}\left(\frac{1}{2}\right)^{100-r} - 51\left(\frac{1}{2}\right)^{100}\binom{99+1}{50+1} \qquad \text{(according to Example 2.2)}\\
&= 51\left\{\sum_{r=0}^{50}\binom{100-r}{50}\left(\frac{1}{2}\right)^{100-r} - \binom{100}{50}\left(\frac{1}{2}\right)^{100}\right\} - 51\binom{100}{51}\left(\frac{1}{2}\right)^{100}\\
&= 51-51\left\{\binom{100}{50}+\binom{100}{51}\right\}\left(\frac{1}{2}\right)^{100} = 51-51\left\{\binom{100}{50}+\frac{50}{51}\binom{100}{50}\right\}\left(\frac{1}{2}\right)^{100},
\end{aligned}$$

hence by a rearrangement,

$$\mu = 101\binom{100}{50}\left(\frac{1}{2}\right)^{100} - 1.$$

2) It follows from Example 2.3 that

$$\binom{2n}{n} \sim \frac{2^{2n}}{\sqrt{\pi n}}, \quad \text{i.e.} \quad \binom{100}{50} \sim \frac{2^{100}}{\sqrt{50\pi}} \quad \text{for } n=50,$$

hence

$$\mu = 101\binom{100}{50}\left(\frac{1}{2}\right)^{100} - 1 \sim \frac{101}{\sqrt{50\pi}} - 1 \approx 7.05863.$$

Additional remark. If we apply the more precise estimate from Example 2.3,

$$\sqrt{\frac{100}{101}}\cdot\frac{2^{100}}{\sqrt{50\pi}} < \binom{100}{50} < \frac{2^{100}}{\sqrt{50\pi}},$$

then it follows by this method that

$$7.01864 < \mu < 7.05863.$$

It was mentioned in the beginning of this example that a direct application of the pocket calculator gives

$$\mu \approx 7.03851,$$

which is very close to the mean value of the upper and lower bound above.

Example 7.3 *A box contains* $2N$ *balls, of which* $2h$ *are white,* $0 < h < N$.
Another box contains $3N$ *balls, of which* $3h$ *are white.*
If we select two balls from each of the boxes, for which box do we have the largest probability of getting 2 white balls?
Which box has the largest probability of obtaining at least one white ball?

1) Traditionally the other balls are black, so $N = h + s$.
We are dealing with hypergeometric distributions, so

$$p_2 = P\{2 \text{ white balls from } U_2\} = \frac{\binom{2h}{2}\binom{2s}{0}}{\binom{2N}{2}} = \frac{\binom{2h}{2}}{\binom{2N}{2}} = \frac{2h(2h-1)}{2N(2N-1)} = \frac{h}{N} \cdot \frac{2h-1}{2N-1},$$

and analogously,

$$p_3 = P\{2 \text{ white balls from } U_3\} = \frac{\binom{3h}{2}}{\binom{3N}{2}} = \frac{3h(3h-1)}{3N(3N-1)} = \frac{h}{N} \cdot \frac{3h-1}{3N-1}.$$

It follows from

$$\begin{aligned}\frac{3h-1}{3N-1}-\frac{2h-1}{2N-1} &= \frac{1}{(3N-1)(2N-1)}\{(6hN-3h-2N+1)-(6hN-2h-3N+1)\}\\ &= \frac{N-h}{(3N-1)(2N-1)}>0,\end{aligned}$$

that $p_2 < p_3$.

2) By interchanging h and s we get analogously that we have the largest probability of obtaining 2 black balls from U_3.
Since ({at least one white ball} is the complementary event of {two black balls}, we have the largest probability of obtaining at least one white ball from U_2.

Example 7.4 *A box contains 10 white and 5 black balls. We select without replacement 4 balls. Let the random variable X denote the number of white balls among the 4 selected balls, and let Y denote the number the number of black balls among the 4 selected balls.*

1) Compute $P\{X=i\}$, $i=0,\ 1,\ 2,\ 3,\ 4$, and $P\{Y=i\}$, $i=0,\ 1,\ 2,\ 3,\ 4$.

2) Find the means $E\{X\}$ and $E\{Y\}$.

3) Compute the variances $V\{X\}$ and $V\{Y\}$.

4) Compute the correlation coefficient between X and Y.

1) This is an hypergeometric distribution with $a=10$, $b=5$ and $n=4$, thus

$$P\{X=i\}=\frac{\binom{10}{i}\binom{5}{4-i}}{\binom{15}{4}} \quad \text{and} \quad P\{Y=i\}=\frac{\binom{5}{i}\binom{10}{4-i}}{\binom{15}{4}}.$$

By some computations,

$$P\{X=0\}=P\{Y=4\}=\frac{\binom{10}{0}\binom{5}{4}}{\binom{15}{4}}=\frac{1\cdot 5}{1365}=\frac{1}{273}\approx 0,0037,$$

$$P\{X=1\}=P\{Y=3\}=\frac{\binom{10}{1}\binom{5}{3}}{1365}=\frac{10\cdot\frac{5\cdot 4}{1\cdot 2}}{1365}=\frac{20}{273}\approx 0,0733,$$

$$P\{X=2\}=P\{Y=2\}=\frac{\binom{10}{2}\binom{5}{2}}{1365}=\frac{\frac{10\cdot 9}{1\cdot 2}\cdot\frac{5\cdot 4}{1\cdot 2}}{1365}=\frac{90}{273}\approx 0,3297,$$

$$P\{X=3\}=P\{Y=1\}=\frac{\binom{10}{3}\binom{5}{1}}{1365}=\frac{\frac{10\cdot 9\cdot 8}{1\cdot 2\cdot 3}\cdot 5}{1365}=\frac{120}{273}\approx 0,4396,$$

$$P\{X=4\} = P\{Y=0\} = \frac{\binom{10}{4}\binom{5}{0}}{1365} = \frac{\frac{10\cdot 9\cdot 8\cdot 7}{1\cdot 2\cdot 3\cdot 4}}{1365} = \frac{42}{273} \approx 0,1538,$$

where of course

$$\frac{90}{273} = \frac{30}{91}, \quad \frac{120}{273} = \frac{40}{91} \quad \text{and} \quad \frac{42}{273} = \frac{2}{13}.$$

2) The means are

$$E\{X\} = n \cdot \frac{a}{a+b} = 4 \cdot \frac{10}{10+5} = \frac{8}{3},$$

and

$$E\{Y\} = n \cdot \frac{b}{a+b} = 4 \cdot \frac{5}{10+5} = \frac{4}{3} \qquad (= 4 - E\{X\}).$$

3) The variance is

$$V\{X\} = V\{Y\} = \frac{nab(a+b-n)}{(a+b)^2(a+b-1)} = \frac{4\cdot 10\cdot 5\cdot (10+5-4)}{(10+5)^2(10+5-1)} = \frac{200\cdot 11}{225\cdot 14} = \frac{44}{63}.$$

4) It follows from $Y = 4 - X$ that

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E\{XY\} - E\{X\}E\{Y\} = E\{X(4-X)\} - E\{X\}E\{4-X\} \\ &= 4\,E\{X\} - E\left\{X^2\right\} - 4\,E\{X\} + (E\{X\})^2 = -V\{X\}, \end{aligned}$$

hence

$$\varrho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{V\{X\}\,V\{Y\}}} = \frac{-V\{X\}}{V\{X\}} = -1.$$

Example 7.5 *A collection of 100 rockets contains 10 duds. A customer buys 20 of the rockets.*

1) Find the probability that exactly 2 of the 20 rockets are duds.

2) Find the probability that none of the 20 rockets is a dud.

3) What is the expected number of duds among the 20 rockets?

This is an hypergeometric distribution with $a = 10$ (the duds) and $b = 90$ and $n = 20$. Let X denote the number of duds. Then

$$(1)\quad P\{X=i\} = \frac{\binom{10}{i}\binom{90}{20-i}}{\binom{100}{20}}, \qquad i = 0,\, 1,\, 2,\, \ldots,\, 10.$$

1) When $i = 2$, it follows from (1) that

$$P\{X=2\} = \frac{\binom{10}{2}\binom{90}{18}}{\binom{100}{20}} = \frac{10!}{2!8!}\cdot\frac{90!}{18!72!}\cdot\frac{20!80!}{100!} = \frac{90!}{100!}\cdot\frac{80!}{72!}\cdot\frac{20!}{18!}\cdot\frac{10!}{8!}\cdot\frac{1}{2!}$$
$$= \frac{80\cdot 79\cdot 78\cdot 77\cdot 76\cdot 75\cdot 74\cdot 73}{100\cdot 99\cdot 98\cdot 97\cdot 96\cdot 95\cdot 94\cdot 93\cdot 92\cdot 91}\cdot\frac{20\cdot 19\cdot 10\cdot 9}{2} = \frac{101355025}{318555566} \approx 0,3182.$$

2) When $i = 0$, it follows from (1) that

$$P\{X=0\} = \frac{\binom{10}{0}\binom{90}{20}}{\binom{100}{20}} = \frac{90!}{70!20!}\cdot\frac{80!20!}{100!} = \frac{90!}{100!}\cdot\frac{80!}{70!}$$
$$= \frac{80\cdot 79\cdot 78\cdot 77\cdot 76\cdot 75\cdot 74\cdot 73\cdot 72\cdot 71}{100\cdot 99\cdot 98\cdot 97\cdot 96\cdot 95\cdot 94\cdot 93\cdot 92\cdot 91} = \frac{15149909}{159277783} \approx 0,0951.$$

3) The expected number of duds is the mean

$$E\{X\} = n\cdot\frac{a}{a+b} = 20\cdot\frac{10}{100} = 2.$$

Example 7.6 *A box U_1 contains 2 white and 4 black balls. Another box U_2 contains 3 white and 3 black balls, and a third box U_3 contains 4 white and 2 b black balls.*
An experiment is described by selecting randomly a sample consisting of three balls from each of the three boxes. The numbers of the white balls in each of these samples are random variables, which we denote by X_1, X_2 and X_3.

1. *Compute $E\{X_1\}$, $E\{X_2\}$, $E\{X_3\}$ and $E\{X_1+X_2+X_3\}$ and $V\{X_1+X_2+X_3\}$.*

2. *Find the probability $P\{X_1 = X_2 = X_3\}$.*

Then collect all 18 balls in one box, from which we take a sample consisting of 9 balls. Then the number of white balls in the sample is a random variable, which is denoted by Y.

Find $E\{Y\}$ and $V\{Y\}$.

The boxes U_1, U_2 and U_3 are represented by

$$U_1 = \{h,\, h,\, s,\, s,\, s,\, s\}, \qquad U_2 = \{h,\, h,\, h,\, s,\, s,\, s\}, \qquad U_3 = \{h,\, h,\, h,\, h,\, s,\, s\}.$$

1) We have in all three cases hypergeometric distributions. Hence,

$$P\{X_1 = i\} = \frac{\binom{2}{i}\binom{4}{3-i}}{\binom{6}{3}}, \qquad i = 1,\, 0,\, 1,\, 2,$$

$$P\{X_2 = i\} = \frac{\binom{3}{i}\binom{3}{3-i}}{\binom{6}{3}}, \qquad i = 1,\, 0,\, 1,\, 2,\, 3,$$

$$P\{X_3 = i\} = \frac{\binom{4}{i}\binom{2}{3-i}}{\binom{6}{3}}, \qquad i = 1,\, 0,\, 1,\, 2,\, 3.$$

The means and the variances can now be found by formulæ in any textbook. However, we shall here compute all probabilities in the "hard way", because we shall need them later on:

a) When we draw from U_1 we get

$$\begin{aligned} P\{X_1=0\} &= \binom{3}{0}\cdot\frac{4}{6}\cdot\frac{3}{5}\cdot\frac{2}{4}=\frac{1}{5},\\ P\{X_1=1\} &= \binom{3}{1}\cdot\frac{2}{6}\cdot\frac{4}{5}\cdot\frac{3}{4}=\frac{3}{5},\\ P\{X_1=2\} &= \binom{3}{2}\cdot\frac{2}{6}\cdot\frac{1}{5}\cdot\frac{4}{4}=\frac{1}{5},\\ P\{X_1=3\} &= 0, \end{aligned}$$

where

$$E\{X_1\}=1\cdot\frac{3}{5}+2\cdot\frac{1}{5}=1 \quad \text{and} \quad E\left\{X_1^2\right\}=1\cdot\frac{3}{5}+4\cdot\frac{1}{5}=\frac{7}{5},$$

hence

$$V\{X_1\}=\frac{7}{5}-1^2=\frac{2}{5}.$$

b) When we draw from U_2 we get analogously

$$\begin{aligned} P\{X_2=0\} &= \binom{3}{0}\cdot\frac{3}{6}\cdot\frac{2}{5}\cdot\frac{1}{4}=\frac{1}{20},\\ P\{X_2=1\} &= \binom{3}{1}\cdot\frac{3}{6}\cdot\frac{3}{5}\cdot\frac{2}{4}=\frac{9}{20},\\ P\{X_2=2\} &= \binom{3}{2}\cdot\frac{3}{6}\cdot\frac{2}{5}\cdot\frac{3}{4}=\frac{9}{20},\\ P\{X_2=3\} &= \binom{3}{3}\cdot\frac{3}{6}\cdot\frac{2}{5}\cdot\frac{1}{4}=\frac{1}{20}, \end{aligned}$$

hence

$$\begin{aligned} E\{X_2\} &= 1\cdot\frac{9}{20}+2\cdot\frac{9}{20}+3\cdot\frac{1}{20}=\frac{9+18+3}{20}=\frac{3}{2},\\ E\left\{X_2^2\right\} &= 1\cdot\frac{9}{20}+4\cdot\frac{9}{20}+9\cdot\frac{1}{20}=\frac{9+36+9}{20}=\frac{54}{20}=\frac{27}{10},\\ V\{X_2\} &= \frac{54}{20}-\frac{9}{4}=\frac{54-45}{20}=\frac{9}{20}. \end{aligned}$$

c) When we draw from U_3 we get complementary to the draw from U_1 that

$$\begin{aligned} P\{X_3=0\} &= 0,\\ P\{X_3=1\} &= \frac{1}{5},\\ P\{X_3=2\} &= \frac{3}{5},\\ P\{X_3=3\} &= \frac{1}{5}. \end{aligned}$$

Hence we get (there are here several variants)

$$E\{X_3\}=1\cdot\frac{1}{5}+2\cdot\frac{3}{5}+3\cdot\frac{1}{5}=\frac{1+6+3}{5}=2=3-E\{X_1\},$$

and

$$V\{X_3\} = V\{X_1\} = \frac{2}{5}.$$

Then

$$E\{X_1 + X_2 + X_3\} = E\{X_1\} + E\{X_2\} + E\{X_3\} = 1 + \frac{3}{2} + 2 = \frac{9}{2}.$$

Since X_1, X_2 and X_3 are stochastically independent, we get

$$V\{X_1 + X_2 + X_3\} = V\{X_1\} + V\{X_2\} + V\{X_3\} = \frac{2}{5} + \frac{9}{20} + \frac{2}{5} = \frac{8+9+8}{20} = \frac{25}{20} = \frac{5}{4}.$$

2) It follows that

$$\begin{aligned} P\{X_1 = X_2 = X_3\} &= \sum_{k=0}^{3} P\{X_1 = k\} \cdot P\{X_2 = k\} \cdot P\{X_3 = k\} \\ &= \frac{1}{5} \cdot \frac{1}{20} \cdot 0 + \frac{3}{5} \cdot \frac{9}{20} \cdot \frac{1}{5} + \frac{1}{5} \cdot \frac{9}{20} \cdot \frac{3}{5} + 0 \cdot \frac{1}{20} \cdot \frac{1}{5} = 2 \cdot \frac{3}{5} \cdot \frac{9}{20} \cdot \frac{1}{5} = \frac{27}{250}. \end{aligned}$$

3) We have in this case the hypergeometric distribution

$$P\{Y=k\} = \frac{\binom{9}{k}\binom{9}{9-k}}{\binom{18}{9}} = \frac{\binom{9}{k}^2}{\binom{18}{9}}, \qquad i = 0,\, 1,\, \ldots,\, 9$$

with $a = b = n = 9$, hence

$$E\{Y\} = \frac{na}{a+b} = \frac{9\cdot 9}{9+9} = \frac{9}{2},$$

and

$$V\{Y\} = \frac{nab(a+b-n)}{(a+b)^2(a+b-1)} = \frac{9\cdot 9\cdot 9\cdot 9}{18^2\cdot 17} = \frac{81}{4\cdot 17} = \frac{81}{68}.$$

Example 7.7 *Given a sequence of random variables (X_n), for which*

$$P\{X_n = k\} = \frac{\binom{a_n}{k}\binom{b_n}{m-k}}{\binom{a_n+b_n}{m}}, \qquad \max(0, m-b_n) \le k \le \min(a_n, m),$$

where m, a_n, $b_n \in \mathbb{N}$, and where $a_n \to \infty$ and $b_n \to \infty$ in such a way that

$$\frac{a_n}{a_n+b_n} \to p, \qquad p \in]0,1[.$$

Prove that the sequence (X_n) converges in distribution towards a random variable X, which is binomially distributed, $X \in B(m,p)$.
Give an intuitiv interpretation of this result.

When n is sufficiently large, then $a_n \ge m$ and $b_n \ge m$. We choose n so big that this is the case. Then for $0 \le k \le m$,

$$\begin{aligned} P\{X_n=k\} &= \frac{a_n!}{k!\,(a_n-k)!}\cdot\frac{b_n!}{(m-k)!}\cdot\frac{m!}{(b_n-m+k)!}\cdot\frac{(a_n+b_n-m)!}{(a_n+b_n)!} \\ &= \binom{m}{k}\cdot\frac{a_n\,(a_n-1)\cdots(a_n-k+1)\cdot b_n\,(b_n-1)\cdots(b_n-m+k+1)}{(a_n+b_n)\,(a_n+b_n-1)\cdots(a_n+b_n-m+1)} \\ &= \binom{m}{k}\cdot\frac{a_n}{a_n+b_n}\cdot\frac{a_n-1}{a_n+b_n-1}\cdots\frac{a_n-k+1}{a_n+b_n-k+1}\cdot\frac{b_n}{a_n+b_n-k}\cdots\frac{b_n-m+k+1}{a_n+b_n-m+1} \\ &\to \binom{m}{k}p^k(1-p)^{m-k} \quad \text{for } k = 0,\, 1,\, \ldots,\, m \quad \text{når } n\to\infty. \end{aligned}$$

Remark 7.1 If we have a large number of white balls a_n, and a large number of black balls, then it is almost unimportant if we draw with replacement (binomial) or without replacement *(hypergeometric).*

An ALTERNATIVE proof, in which we apply Stirling's formula, is the following. Let

$$P\{X=k\} = \left(\begin{array}{c} m \\ k \end{array}\right) p^k (1-p)^{m-k}, \qquad k = 0,\, 1,\, \ldots,\, m, \qquad X \in B(m,p).$$

Choose N, such that $b_n \geq m$ and $a_n \geq m$ for all $n \geq N$. Then it follows from Stirling's formula for $n \geq N$,

$$\begin{aligned} P\{X=k\} - P\{X_n = k\} &= \left(\begin{array}{c} m \\ k \end{array}\right) p^k (1-p)^{m-k} - \frac{\left(\begin{array}{c} a_n \\ k \end{array}\right)\left(\begin{array}{c} b_n \\ m-k \end{array}\right)}{\left(\begin{array}{c} a_n+b_n \\ m \end{array}\right)} \\ &= \left(\begin{array}{c} m \\ k \end{array}\right) p^k (1-p)^{m-k} - \frac{a_n!}{k!\,(a_n-k)!} \cdot \frac{b_n!}{(m-k)!\,(b_n-m+k)!} \cdot \frac{m!\,(a_n+b_n-m)!}{(a_n+b_n)!} \\ &= \left(\begin{array}{c} m \\ k \end{array}\right) p^k (1-p)^{m-k} - \frac{m!}{k!(m-k)!} \cdot \frac{(a_n+b_n-m)!a_n!b_n!}{(b_n-m+k)!\,(a_n-k)!\,(a_n+b_n)!} \\ &= \left(\begin{array}{c} m \\ k \end{array}\right) p^k (1-p)^{m-k} - \left(\begin{array}{c} m \\ k \end{array}\right) \cdot \frac{(a_n+b_n-m)!}{(b_n-m+k)!\,(a_n-k)!} \cdot \frac{a_n!b_n!}{(a_n+b_n)!}, \end{aligned}$$

where

$$\begin{aligned} &(a_n+b_n-m)! \sim \sqrt{2\pi(a_n+b_n-m)} \cdot (a_n+b_n-m)^{a_n+b_n-m} \exp(-(a_n+b_n-m)), \\ &(b_n-m+k)! \sim \sqrt{2\pi(b_n-m!+k)} \cdot (b_n-m+k)^{b_n-m+k} \exp(-(b_n-m+k)), \\ &(a_n-k)! \sim \sqrt{2\pi(a_n-k)} \cdot (a_n-k)^{a_n-k} \exp(-(a_n-k)), \\ &a_n! \sim \sqrt{2\pi} \cdot a_n^{a_n} \cdot \exp(-a_n), \\ &b_n! \sim \sqrt{2\pi} \cdot b_n^{b_n} \cdot \exp(-b_n), \\ &(a_n+b_n)! \sim \sqrt{2\pi(a_n+b_n)} \cdot (a_n+b_n)^{a_n+b_n} \cdot \exp(-(a_n+b_n)). \end{aligned}$$

Since the exponentials disappear by insertion, we get the following

$$\begin{aligned} &\frac{(a_n+b_n-m)!a_n!b_n!}{(b_n-m+k)!(a_n-k)!(a_n+b_n)!} \\ &\quad\sim \sqrt{\frac{2\pi(a_n+b_n-m)\cdot 2\pi a_n \cdot 2\pi b_n}{2\pi(b_n-m+k)\cdot 2\pi(a_n-k)\cdot 2\pi(a_n+b_n)}} \times \\ &\qquad \times \frac{(a_n+b_n-m)^{a_n+b_n-m} a_n^{a_n} b_n^{b_n}}{(b_n-m+k)^{b_n-m+k}(a_n-k)^{a_n-k}(a_n+b_n)^{a_n+b_n}} \\ &\quad= \sqrt{\frac{a_n+b_n-m}{a_n+b_n} \cdot \frac{a_n}{a_n-k} \cdot \frac{b_n}{n_n-m+k}} \times \\ &\qquad \times \left(\frac{a_n+b_n-m}{a_n+b_n}\right)^{a_n+b_n} \frac{a_n^k b_n^{m-k}}{(a_n+b_n-m)^m} \left(\frac{a_n}{a_n-k}\right)^{a_n-k} \left(\frac{b_n}{b_n-m+k}\right)^{b_n-m+k} \\ &\quad\to 1 \cdot \lim_{n\to\infty}\left(1-\frac{m}{a_n+b_n}\right)^{a_n+b_n} \cdot \lim_{n\to\infty}\left(1+\frac{k}{a_n-k}\right)^{a_n-k} \times \\ &\qquad \times \lim_{n\to\infty}\left(1+\frac{m-k}{b_n-m+k}\right)^{b_n-m+k} \cdot \lim_{n\to\infty}\left(\frac{a_n}{a_n+b_n-m}\right)^{k} \left(\frac{b_n}{a_n+b_n-m}\right)^{m-k} \\ &\quad= 1 \cdot e^{-m} \cdot e^{k} \cdot e^{m-k} \cdot p^k (1-p)^m - k = p^k (1-p)^{m-k}. \end{aligned}$$

Finally, we get by insertion

$$\lim_{n\to\infty} \left(P\{X = k\} - P\{X_n = k\}\right) = 0,$$

and we have proved that $X_n \to X \in B(m, p)$ in distribution.

Example 7.8 *A deck of cards consists of the 13 diamonds. The court cards are the 3 cards jack, queen and king. We let in this example ace have the value 1.*
We take a sample of 4 cards.
Let X denote the random variable, which indicates the number of cards which are not court cards among the 4 selected cards.

1) Compute the probabilities

$$P\{X = i\}, \qquad i = 1,\, 2,\, 3,\, 4.$$

2) Find the mean $E\{X\}$.

3) We now assume that among the chosen 4 cards are precisely 2 cards which are not court cards, thus $X = 2$.
What is the expected sum (the mean) of these 2 cards which are not court cards?

1) Here X is hypergeometrically distributed with $a = 10$, $b = 3$, $n = 4$, so we get

$$P\{X = i\} = \frac{\binom{10}{i}\binom{3}{4-i}}{\binom{13}{4}} = \frac{1}{715}\binom{10}{i}\binom{3}{4-i}, \qquad i = 1,\, 2,\, 3,\, 4.$$

Hence,

$$\begin{aligned} P\{X = 1\} &= \frac{10}{715} = \frac{2}{143}, \\ P\{X = 2\} &= \frac{135}{715} = \frac{27}{143}, \\ P\{X = 3\} &= \frac{360}{715} = \frac{72}{143}, \\ P\{X = 4\} &= \frac{210}{715} = \frac{42}{143}. \end{aligned}$$

2) Then by a convenient formula the mean is

$$E\{X\} = \frac{na}{a+b} = \frac{4 \cdot 10}{13} = \frac{40}{13}.$$

Alternatively we get by a direct computation,

$$E\{X\} = \frac{1}{143}\left(1 \cdot 2 + 2 \cdot 27 + 3 \cdot 72 + 4 \cdot 42\right) = \frac{440}{143} = \frac{40}{13}.$$

3) Given that we have two cards which are not court cards. The first one can have the values 1, 2, 3, ..., 10, and the expected value is

$$\frac{1}{10}\{1+2+3+\cdot+10\} = \frac{11}{2}.$$

The second card has also the expected value $\frac{11}{2}$, hence the expected sum is

$$\frac{11}{2}+\frac{11}{2} = 11.$$

ALTERNATIVELY and more difficult we write down all 90 possibilities of the same probability, where we are aware of the order of the two cards:

$$\begin{array}{ccccc} 1+2, & 2+1, & 3+1, & \cdots, & 10+1, \\ 1+3, & 2+3, & 3+2, & \cdots, & 10+2, \\ 1+4, & 2+4, & 3+4, & \cdots, & 10+3, \\ \vdots & \vdots & \vdots & & \vdots \\ 1+10, & 2+10, & 3+10, & \cdots, & 10+9. \end{array}$$

The total sum is

$$18\cdot(1+2+3+\cdots+10) = 18\cdot 55 = 990.$$

Then the average is

$$\frac{990}{90} = 11.$$

Example 7.9 *We shall from a group G_1 consisting of 2 girls and 5 boys choose a committee with 3 members, and from another group G_2 of 3 girls and 4 boys we shall also select a committee of 3 members.Let X_1 denote the number of girls in the first committee, and X_2 the number of girls in the second committee.*

1. *Find the means $E\{X_1\}$ and $E\{X_2\}$.*

2. *Find $P\{X_1 = X_2\}$.*

The two groups G_1 and G_2 are then joined into one group G of 14 members. We shall from this group choose a committee of 6 members. Let Y denote the number of girls in this committee.

3. *Find the mean $E\{Y\}$ and the variance $V\{Y\}$.*

We are again considering hypergeometric distributions.

1) Since G_1 consists of 2 girls and 5 boys, of which we shall choose 3 members, we get

$$a = 2, \qquad b = 5, \qquad a + b = 7 \qquad \text{and} \qquad n = 3,$$

hence

$$E\{X_1\} = \frac{na}{a+b} = \frac{3 \cdot 1}{7} = \frac{6}{7}.$$

Since G_2 consists of 3 girls and 4 boys, of which we shall choose 3 members, we get

$$a = 3, \qquad b = 4, \qquad a + b = 7 \qquad \text{and} \qquad n = 3,$$

hence

$$E\{X_2\} = \frac{na}{a+b} = \frac{3 \cdot 3}{y} = \frac{9}{7}.$$

2) It follows from

$$P\{X_1 = k\} = \frac{\binom{2}{k}\binom{5}{3-k}}{\binom{7}{3}} = \frac{1}{35}\binom{2}{k}\binom{5}{3-k} \qquad \text{for } k = 0,\, 1,\, 2,$$

and

$$P\{X_2 = k\} = \frac{\binom{3}{k}\binom{4}{3-k}}{\binom{7}{3}} = \frac{1}{35}\binom{3}{k}\binom{4}{3-k} \qquad \text{for } k = 0,\, 1,\, 2,\, 3,$$

and $P\{X_1 = 3\} = 0$ that

$$P\{X_1 = X_2\} = P\{X_1 = 0\} \cdot P\{X_2 = 0\} + P\{X_1 = 1\} \cdot P\{X_2 = 1\} + P\{X_1 = 2\} \cdot P\{X_2 = 2\}.$$

Here

$$P\{X_1 = 0\} \cdot P\{X_2 = 0\} = \frac{\binom{2}{0}\binom{5}{3}}{35} \cdot \frac{\binom{3}{0}\binom{4}{3}}{35} = \frac{10 \cdot 4}{35^2} = \frac{40}{35^2},$$

$$P\{X_1 = 1\} \cdot P\{X_2 = 1\} = \frac{\binom{2}{1}\binom{5}{2}}{35} \cdot \frac{\binom{3}{1}\binom{4}{2}}{35} = \frac{2 \cdot 10 \cdot 3 \cdot 6}{35^2} = \frac{360}{35^2},$$

$$P\{X_1 = 2\} \cdot P\{X_2 = 2\} = \frac{\binom{2}{2}\binom{5}{1}}{35} \cdot \frac{\binom{3}{2}\binom{4}{1}}{35} = \frac{5 \cdot 3 \cdot 4}{35^2} = \frac{60}{35^2},$$

which gives by insertion

$$P\{X_1 = X_2\} = \frac{40 + 360 + 60}{35^2} = \frac{460}{35^2} = \frac{92}{245}.$$

3) Since G consists of 5 girls and 9 boys, of which we shall choose 6 members, we get

$$a = 5, \qquad b = 9, \qquad a + b = 14 \qquad \text{and} \qquad n = 6,$$

hence

$$E\{Y\} = \frac{na}{a+b} = \frac{6 \cdot 5}{14} = \frac{15}{7},$$

and

$$V\{Y\} = \frac{nab(a+b-n)}{(a+b)^2(a+b-1)} = \frac{6 \cdot 5 \cdot 9 \cdot (14-6)}{14^2 \cdot (14-1)} = \frac{5 \cdot 6 \cdot 8 \cdot 9}{4 \cdot 13 \cdot 49} = \frac{540}{637}.$$

Index

Banach's match stick problem, 54
Bernoulli event, 5
binomial coefficient, 4
binomial distribution, 4, 10, 64
binomial series, 5

Chu-Vandermonde's formula, 5
convergence in distribution, 64
criterion of integral, 35

decreasing factorial, 4

geometric distribution, 6, 31

hypergeometric distribution, 9, 54

negative binomial distribution, 8, 52

Pascal distribution, 7, 49
Poisson distribution, 6, 24

reduced waiting time, 8

skewness, 25
Stirling's formula, 5, 11, 56, 65
stochastically larger random variable, 49

waiting time distribution, 7
waiting time distributions, 8