6.2. Discrete Probability Distributions

Discrete Uniform distribution
(diskreetti tasajakauma)

A random variable X follows the dicrete uniform distribution on the interval $[a, a+1, \ldots, b]$, if it may attain each of these values with equal probability. We have then:

$$P(X=x) = \begin{cases} \frac{1}{n} & \text{for } a \le x \le b \\ 0 & \text{otherwise,} \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a+1}{n} & \text{for } a \le x \le b \\ 1 & \text{for } x > b, \end{cases}$$

$$E(X) = \frac{1}{2}(a+b) \text{ and } V(X) = \frac{n^2-1}{12},$$

where $n = b - a + 1$ denotes the number of values that X may take.

Example. Throwing a dice:

$$E(X) = \frac{1+6}{2} = 3.5 \text{ and } V(X) = \frac{6^2-1}{12} \approx 2.92,$$

the same results as we found earlier.

Bernoulli distribution (Bernoulli-jakaumaa)

Let X be a random variable with possible values 0 and 1, and let $P(X=1) = p$. That is, the probability distribution of X is

x	0	1
$P(X = x)$	$1-p$	p

or written as a mathematical formula

$$P(X=x) = \begin{cases} p^x(1-p)^{1-x}, \text{ if } x \in [0,1] \\ 0, \text{ otherwise.} \end{cases}$$

We denote $X \sim$ Ber(p), meaning “X is Bernoulli distributed with parameter p”.

An example of a Bernoulli random variable is the result of a toss of a coin with Head, say, equal to one and Tail equal to zero.

Note. p is called the probability of ”success”, and $q = 1 - p$ the probability of ”failure”.

$$E[X] = 0 \times p^0(1-p)^1 + 1 \times p(1-p)^0 = p$$

$$V[X]=E[X^2]-(E[X])^2=p-p^2=p(1-p)=pq$$

Sums of independent Bernoulli trials

Assume that the probability that your computer will fail to perform a certain calculation within a limited time is $P(F) = 0.1$, such that the probability of success is $P(S) = 0.9$. We may model this as a Bernoulli random variable with $p = 0.9$ and $q = 0.1$.

Now suppose you let your computer try solving different but equally difficult calculations (same probability of success p) three times and you're interested in the number of successes X within these 3 trials. Note that X may be thought of as the sum of three independent Bernoulli trials with success probability p.

The $2 \times 2 \times 2 = 8$ possible outcomes of the sequences of sucesses and failures may be systematically arranged as follows

$$\begin{array}{cccc}
FFF & FFS & FSS & SSS \\
 & FSF & SFS & \\
 & SFF & SSF & \\
 & & & \\
X{=}0 & X{=}1 & X{=}2 & X{=}3
\end{array}$$

We wish to find the probability distribution of X. By independence:

$P(X=0) = P(FFF) = 0.1 \times 0.1 \times 0.1 = 0.001,$
$P(X=3) = P(SSS) = 0.9 \times 0.9 \times 0.9 = 0.729.$

Similiarly,

$P(FFS) = P(FSF) = P(SFF) = 0.1 \times 0.1 \times 0.9 = 0.009,$
$P(FSS) = P(SFS) = P(SSF) = 0.1 \times 0.9 \times 0.9 = 0.081.$

In order to find $P(X=1)$ and $P(X=2)$ we note that their constituting events are mutually disjoint and that there are $\binom{3}{1} = 3$ such events for the event $X=1$, and $\binom{3}{2} = 3$ such events for the event $X=2$, such that we may write:

$$P(X=1) = \binom{3}{1} \cdot 0.1 \cdot 0.1 \cdot 0.9 = 0.027,$$
$$P(X=2) = \binom{3}{2} \cdot 0.1 \cdot 0.9 \cdot 0.9 = 0.243.$$

All these probabilities may be expressed as:

$$P(X=x) = \binom{3}{x} \cdot 0.9^x \cdot 0.1^{3-x}, \quad x = 0, 1, 2, 3.$$

Binomial distribution (Binomijakauma)

Suppose a Bernoulli trial with success probability p is repeated *independently* n times (e.g. tossing a coin n times). Then the probability of x successes, $x = 0, 1, \ldots, n$ can be calculated with the binomial distribution:

$$P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, \ x = 0, 1, \ldots, n \\ 0, \ \text{otherwise} \end{cases}$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}, \quad m! = m(m-1)(m-2)\cdots 1.$$

A binomial variable random variable is a sum of independent Bernoulli variables, that is, $X = \sum_{i=1}^{n} X_i$ where $X_i \sim$ Ber(p). So

$$\mu = \mathsf{E}[X] = \sum_{i=1}^{n} \mathsf{E}[X_i] = \sum_{i=1}^{n} p = np,$$

$$\sigma^2 = V[X] = \sum_{i=1}^{n} V[X_i] = np(1-p) = npq,$$

and

$$\sigma = \sqrt{np(1-p)} = \sqrt{npq}.$$

Example: (Football pool betting continued.) Football pool betting consists of guessing the results of 13 matches, for each of which there are 3 possible outcomes. If the guess is completely random, then the probability of guessing the result of any single match right is $p = \frac{1}{3}$ and the number of correctly guessed matches is binomially distributed with parameters $n = 13$ and $p = \frac{1}{3}$ ($X \sim \text{Bin}(13, \frac{1}{3})$). Then:

$$E(X) = 13 \cdot \frac{1}{3} \approx 4.33, \quad V(X) = 13 \cdot \frac{1}{3} \cdot \frac{2}{3} \approx 2.89$$

and, for example,

$$P(X = 10) = \binom{13}{10} \left(\frac{1}{3}\right)^{10} \left(\frac{2}{3}\right)^{13-10} \approx 0.001435$$

Example. Two percents of a product are defective. If a lot of 100 items are ordered what is the probability that there are no defective items? What is the probability that there are at least two defective items?

Let X denote the number of defective items in the lot. Then possible values of X are $X = 0, 1, \ldots, 100$. Then $X \sim \text{Bin}(100, 0.02)$, such that

$$P(X=0) = \binom{100}{0}(0.02)^0(0.98)^{100} = (0.98)^{100} \approx 0.1326$$

$$\begin{aligned} P(X \geq 2) &= 1 - P(X \leq 1) = 1 - (P(X = 0) + P(X = 1)) \\ &= 1 - \underbrace{(0.98)^{100}}_{P(X=0)} - \underbrace{100(0.02)(0.98)^{99}}_{P(X=1) \approx 0.2707} \approx 0.5967 \end{aligned}$$

Note. The expected value is $\text{E}[X] = 100 \cdot 0.02 = 2$.

Getting binomial probabilities from Excel

If the number of trials n is large, the calculation of binomial probabilities can become quite tedious. Excel can calculate both binomial probabilities

$$b(x;n,p) = P(X{=}x) = \binom{n}{x}p^x(1-p)^{n-x}$$

and cumulative binomial probabilities

$$B(x;n,p) = P(X{\le}x) = \sum_{k=0}^{x}\binom{n}{k}p^k(1-p)^{n-k}$$

using the syntax BINOMDIST(n,x,[cumulative]).

Example:

$b(5;18,0.2)$ =BINOMDIST(5,18,0.2,FALSE),

$B(2;16,0.05)$ =BINOMDIST(2,16,0.05,TRUE).

Hypergeometric Distribution
(Hypergeometrinen jakauma)

Suppose we are interested in the number of defective items in a sample of n items from a lot containing N units, of which a are defective. The probability that the first drawing will yield a defective item is $\frac{a}{N}$, but for the second drawing it is $\frac{a-1}{N-1}$ or $\frac{a}{N-1}$, depending on whether or not the first unit was defective. Thus, the trials are not independent and the binomial distribution does not apply! The binomial distribution would apply if we do sampling with replacement (takaisinpano-otanta), namely, if each unit selected for the sample is replaced before the next sample is drawn.

For sampling without replacement note that x sucesses (defectives) can be chosen in $\binom{a}{x}$ ways, the $n-x$ failures can be chosen in $\binom{N-a}{n-x}$ ways, such that x sucesses and $n-x$ failures can be chosen in $\binom{a}{x}\binom{N-a}{n-x}$ ways. Also, n objects can be chosen from a set of N objects in $\binom{N}{n}$ ways.

The probability of getting x successes in a sample of size n is therefore

$$P(X=x) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, \ldots, n;$$

where x cannot exceed a and $n - x$ cannot exceed $N - a$. This equation defines the hypergeometric distribution $HG(N, a, n)$ with sample size n, population size N, and a sucesses in the population. Hypergeometric probabilities may be calculated in Excel using the command HYPGEOMDIST.

The expected value and variance of the hypergeometric distribution are

$$E(X) = np \quad \text{and} \quad V(X) = npq\left(\frac{N-n}{N-1}\right)$$

with $p = a/N$ and $q = 1 - p$.

For $N \gg n$ the hypergeometric distribution approaches the binomial distribution with parameters n and $p = a/N$. As a rule of thumb the binomial approximation starts being useful for $N \geq 10n$.

Example. Consider a box with 20 balls, five of which are black. Taking out 6 balls without replacement, what is the probability of picking 2 black balls?

$X :=$ number of black balls $\sim HG(20, 5, 6)$:

$$P(X = 2) = \frac{\binom{5}{2}\binom{20-5}{6-2}}{\binom{20}{6}} \approx 0.352$$

If each ball was returned before picking the next ball, then $X \sim \text{Bin}(n, p)$ with $n = 6$ and $p = 5/20 = 1/4$, such that:

$$P(X = 2) = \binom{6}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^4 \approx 0.297$$

Poisson distribution (Poisson-jakauma)

The Poisson distribution is used for modelling the occurence of events during a fixed time interval or the number of rare successes in a very large number of trials, such as:

- the number of misprints on a bookpage,
- the number of goals during a football match,
- the number of telephone calls during a fixed time interval.

Poisson distribution: $X \sim \text{Poi}(\lambda)$

$$P(X = x) = \frac{\lambda^x}{x!}e^{-\lambda}, \text{ for } x = 0, 1, 2, \ldots$$

and zero otherwise, where λ denotes the average number of events per unit time interval, and $e = 2.71828\ldots$ is the base of the natural logarithm.
Excel Syntax: POISSON(x, λ,[cumulative]).

The expected value and variance are

$$E[X] = \lambda \quad \text{and} \quad V[X] = \lambda.$$

Example. Suppose that in a certain area there are on average five traffic accidents per month, that is X=number of traffic accidents ∼Poi(5), such that $E(X) = V(X) = 5$. The probability of 4 accidents in a given month is then

$$P(X = 4) = \frac{5^4}{4!} e^{-5} \approx 0.1755.$$

The Poisson distribution can be used as an approximation for the binomial distribution if p is "small" and n is large (rules of thumb: $p \leq 0.05$ and $n \geq 20$). Then $\lambda = np$ is used.

Example. Using the Poisson approximation for our earlier example of the binomial distribution we get for $X \sim \text{Bin}(100, 0.02)$ with $\lambda = np = 100 \times 0.02 = 2$:

$$P(X = 0) \approx \frac{2^0}{0!} e^{-2} \approx 0.135$$

and

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\ &\approx 1 - e^{-2} - 2e^{-2} \approx 0.594 \end{aligned}$$

This is pretty close to the exact results found earlier:

$$P(X = 0) = 0.133 \quad \text{and} \quad P(X \geq 2) = 0.597.$$

The following conditions must be fulfilled for the Poisson distribution to be applicable:

1. The number of events in nonoverlapping time intervals must be independent.
2. The probability of occurence must be the same in all time intervals of the same length.
3. The probability of more than one event occuring during a short interval must be small relative to the occurence of only one event.
4. The probability of an event occuring in a short interval must be approximately proportional to the interval's length.

Note. The parameter λ is interval-specific in the sense that it denotes the average number of occurrences during whatever is chosen as a unit time interval. It may therefore sometimes be necessary to first transform the average number of occurences to be measured in the applicable time unit.

Geometric Distribution (Geometrinen jakauma)

Cosider repeating a Bernoulli trial with success probability p, until the first success occurs, and denote with X the number of the first successful trial. Then X follows a geometric distribution with parameter p, denoted by $X \sim \text{Geo}(p)$. Its density function is

$$P(X{=}x) = \begin{cases} p(1-p)^{x-1}, \text{ for } x = 1, 2, \ldots \\ 0, \text{ otherwise.} \end{cases}$$

$$E(X) = \frac{1}{p} \quad \text{and} \quad V(X) = \frac{1-p}{p^2}.$$

Example. A gambler decides to play lotto every week until the first time he guesses 7 out of 39 numbers right. What is the expected waiting time for this to happen?

We learned earlier that there are $\binom{39}{7} = 15\,380\,937$ possibilities to draw 7 out of 39 numbers. The waiting time for the lottery winnings is therefore geometrically distributed with parameter $p = 1/15\,380\,937$, such that:

$$E(X) = \frac{1}{p} = 15\,380\,937 \text{ weeks} \approx 295\,787 \text{ years.}$$