Chapter 2

Probability and simulation

Contents

2 Probability and simulation
2.1 Random variable . . . 1
2.2 Discrete random variables . . . 4
2.2.1 Introduction to simulation . . . 7
2.2.2 Mean and variance . . . 10
2.3 Discrete distributions . . . 18
2.3.1 Binomial distribution . . . 18
2.3.2 Hypergeometric distribution . . . 21
2.3.3 Poisson distribution . . . 23
2.4 Continuous random variables . . . 27
2.4.1 Mean and Variance . . . 29
2.5 Continuous distributions . . . 30
2.5.1 Uniform distribution . . . 30
2.5.2 Normal distribution . . . 31
2.5.3 Log-Normal distribution . . . 38
2.5.4 Exponential distribution . . . 39
2.6 Simulation of random variables . . . 42
2.7 Identities for the mean and variance . . . 45
2.8 Covariance and correlation . . . 48
2.9 Independence of random variables . . . 51
2.10 Functions of normal random variables . . . 56
2.10.1 The χ^2-distribution . . . 57
2.10.2 The t-distribution . . . 62
2.10.3 The F-distribution . . . 69

Glossaries **73**

Acronyms **74**

In this chapter elements from probability theory are introduced. These are needed to form the basic mathematical description of randomness. For example for calculating the probabilities of outcomes in various types of experimental or observational study setups. Small illustrative examples, such as e.g. dice rolls and lottery draws, and natural phenomena such as the waiting time between radioactive decays are used as throughout. But the scope of probability theory and it's use in society, science and business, not least engineering endeavour, goes way beyond these small examples. The theory is introduced together with illustrative code examples, which the reader is encouraged to try and interact with in parallel to reading the text. Many of these are of the learning type, cf. the discussion of the way Python is used in the course in Section 1.5.

2.1 Random variable

The basic building blocks to describe random outcomes of an experiment are introduced in this section. The definition of an *experiment* is quite broad. It can be an experiment, which is carried out under controlled conditions e.g. in a laboratory or flipping a coin, as well as an experiment in conditions which are not controlled, where for example a process is observed e.g. observations of the GNP or measurements taken with a space telescope. Hence, an experiment can be thought of as any setting in which the outcome cannot be fully known. This for example also includes measurement noise, which are random "errors" related to the system used to observe with, maybe originating from noise in electrical circuits or small turbulence around the sensor. Measurements will always contain some noise.

First the *sample space* is defined:

Definition 2.1

The *sample space* S is the set of all possible outcomes of an experiment.

Example 2.2

Consider an experiment in which a person will throw two paper balls with the purpose of hitting a wastebasket. All the possible outcomes forms the sample space of this experiment as

$$S = \{(\text{miss},\text{miss}), (\text{hit},\text{miss}), (\text{miss},\text{hit}), (\text{hit},\text{hit})\}. \tag{2-1}$$

Now a *random variable* can be defined:

Definition 2.3

A *random variable* is a function which assigns a numerical value to each outcome in the sample space. In this book random variables are denoted with capital letters, e.g.

$$X, Y, \ldots . \tag{2-2}$$

Example 2.4

Continuing the paper ball example above, a random variable can be defined as the number of hits, thus

$$X\big((\text{miss,miss})\big) = 0, \tag{2-3}$$
$$X\big((\text{hit,miss})\big) = 1, \tag{2-4}$$
$$X\big((\text{miss,hit})\big) = 1, \tag{2-5}$$
$$X\big((\text{hit,hit})\big) = 2. \tag{2-6}$$

In this case the random variable is a function which maps the sample space S to positive integers, i.e. $X : S \to \mathbb{N}_0$.

Remark 2.5

The random variable represents a value of the outcome *before* the experiment is carried out. Usually the experiment is carried out n times and there are random variables for each of them

$$\{X_i : 1, 2, \ldots, n\}. \tag{2-7}$$

After the experiment has been carried out n times a set of values of the random variable is available as

$$\{x_i : 1, 2, \ldots, n\}. \tag{2-8}$$

Each value is called a *realization* or *observation* of the random variable and is denoted with a small letter sub-scripted with an index i, as introduced in Chapter 1.

Finally, in order to quantify probability, a random variable is associated with a *probability distribution*. The distribution can either be discrete or continuous depending on the nature of the outcomes:

- Discrete outcomes can for example be: the outcome of a dice roll, the number of children per family, or the number of failures of a machine per year. Hence some countable phenomena which can be represented by an integer.
- Continuous outcomes can for example by: the weight of the yearly harvest, the time spend on homework each week, or the electricity generation per hour. Hence a phenomena which can be represented by a continuous value.

Furthermore, the outcome can either be unlimited or limited. This is most obvious in the discrete case, e.g. a dice roll is limited to the values between 1 and 6. However it is also often the case for continuous random variables, for example many are non-negative (weights, distances, etc.) and proportions are limited to a range between 0 and 1.

Conceptually there is no difference between the discrete and the continuous case, however it is easier to distinguish since the formulas, which in the discrete case are with sums, in the continuous case are with integrals. In the remaining of this chapter, first the discrete case is presented and then the continuous.

2.2 Discrete random variables

In this section discrete distributions and their properties are introduced. A discrete random variable has discrete outcomes and follows a discrete distribution.

To exemplify, consider the outcome of one roll of a fair six-sided dice as the random variable X^{fair}. It has six possible outcomes, each with equal probability. This is specified with the *probability density function*.

Definition 2.6 The *pdf* of a discrete random variable

For a discrete random variable X the *probability density function* (*pdf*) is

$$f(x) = P(X = x). \tag{2-9}$$

It assigns a probability to every possible outcome value x.
A discrete *pdf* fulfils two properties: there are no negative probabilities for any outcome value

$$f(x) \geq 0 \text{ for all } x, \tag{2-10}$$

and the probabilities for all outcome values sum to one

$$\sum_{\text{all } x} f(x) = 1. \tag{2-11}$$

Example 2.7

For the fair dice the *pdf* is

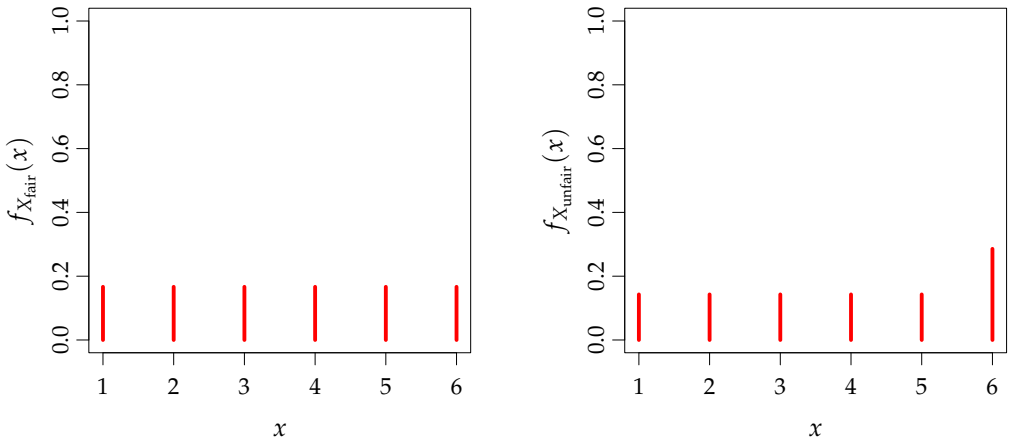
x	1	2	3	4	5	6
$f_{X^{\text{fair}}}(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

If the dice is not fair, maybe it has been modified to increase the probability of rolling a six, the *pdf* could for example be

x	1	2	3	4	5	6
$f_{X^{\text{unfair}}}(x)$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{2}{7}$

where X^{unfair} is a random variable representing the value of a roll with the unfair dice.

The pdfs are plotted: the left plot shows the pdf of a fair dice and the right plot the pdf of an unfair dice:



Remark 2.8

Note that the *pdf*s has subscript with the symbol of the random variable to which they belong. This is done when there is a need to distinguish between *pdf*s e.g. for several random variables. For example if two random variables X and Y are used in same context, then: $f_X(x)$ is the *pdf* for X and $f_Y(x)$ for Y, similarly the sample standard deviation s_X is for X and s_Y is for Y, and so forth.

The *cumulated distribution function (cdf)*, or simply the *distribution function*, is often used.

Definition 2.9 The *cdf* of a discrete random variable

The *cumulated distribution function* (*cdf*) for the discrete case is the probability of realizing an outcome below or equal to the value x

$$F(x) = P(X \leq x) = \sum_{j \text{ where } x_j \leq x} f(x_j) = \sum_{j \text{ where } x_j \leq x} P(X = x_j). \qquad (2\text{-}12)$$

The probability that the outcome of X is in a range is

$$P(a < X \leq b) = F(b) - F(a). \qquad (2\text{-}13)$$

For the fair dice the probability of an outcome below or equal to 4 can be calculated

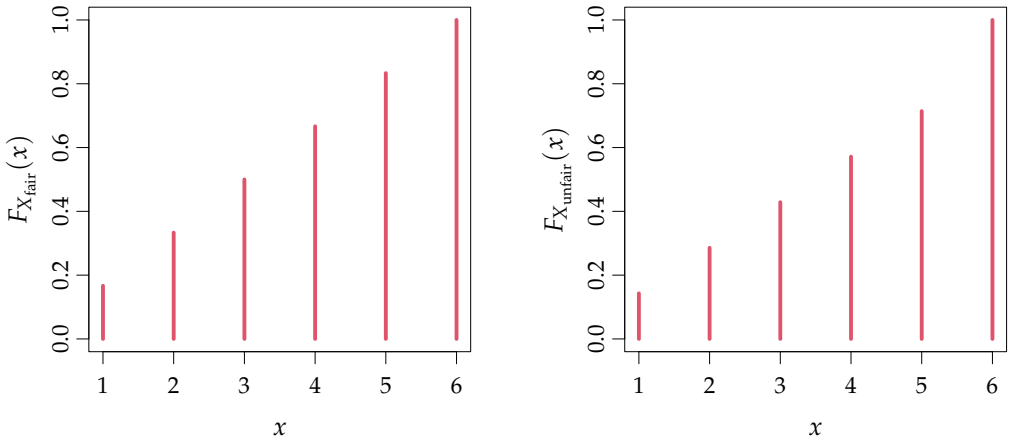
$$F_{X^{\text{fair}}}(4) = \sum_{j=1}^{4} f_{X^{\text{fair}}}(x_j) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{2}{3}. \tag{2-14}$$

|||| Example 2.10

For the fair dice the *cdf* is

x	1	2	3	4	5	6
$F_{X^{\text{fair}}}(x)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1

The cdf for a fair dice is plotted in the left plot and the cdf for an unfair dice is plotted in the right plot:



2.2.1 Introduction to simulation

One nice thing about having computers available is that we try things in virtual reality - this we can here use here to play around while learning how probability and statistics work. With the *pdf* defined an experiment can easily be *simulated*, i.e. instead of carrying out the experiment in reality it is carried out using a model on the computer. When the simulation includes generating random numbers it is called a *stochastic simulation*. Such simulation tools are readily available within Python, and it can be used for as well learning purposes as a way to do large scale complex probabilistic and statistical computations. For now it will be used in the first way.

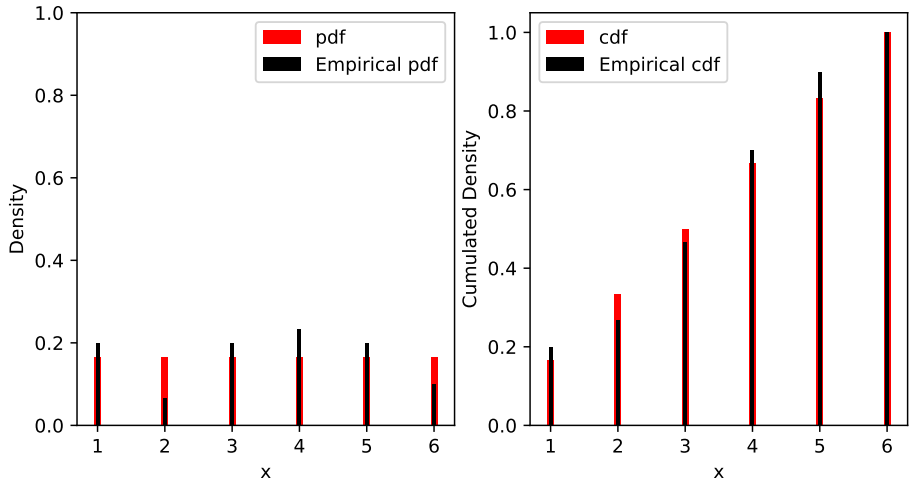
Example 2.11 Simulation of rolling a dice

Let's simulate the experiment of rolling a dice using the following

```
# Make a random draw from (1,2,3,4,5,6) with equal probability for each outcome
np.random.choice(range(1, 7), size=1)
```

The simulation becomes more interesting when the experiment is repeated many times, then we have a sample and can calculate the *empirical density function* (or *empirical pdf or density histogram*, see Section 1.6.1) as a discrete histogram and actually "see" the shape of the *pdf*

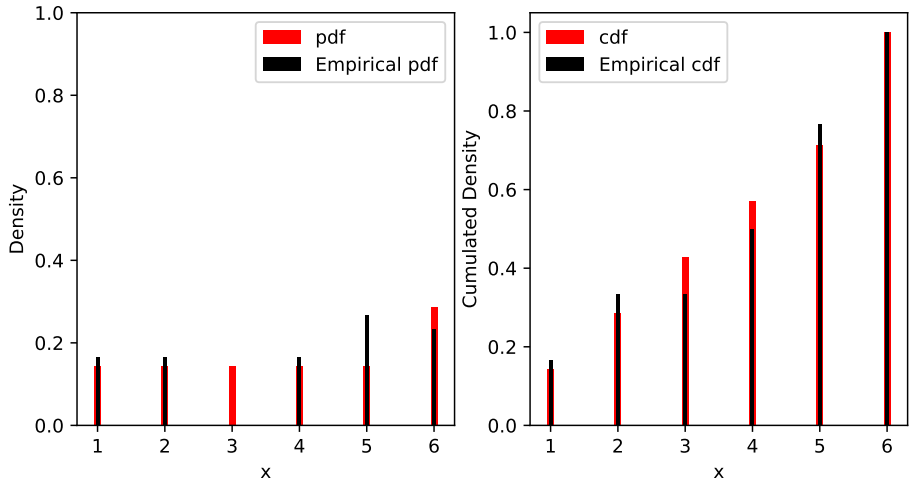
```
# Simulate a fair dice
# Number of simulated realizations
n = 30
# Draw independently from the set (1,2,3,4,5,6) with equal probability
xFair = np.random.choice(range(1, 7), size=n, replace=True)
# Count the number of each outcome using the bincount function
counts = np.bincount(xFair)
# Plot the pdf
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(8,4))
ax1.bar(range(1,7), [1/6]*6, color='red')
# Plot the empirical pdf
ax1.bar(range(1,7), counts[1:7]/n)
# Plot the cdf
ax2.bar(range(1,7), np.cumsum([1/6]*6), color='red')
# Add the empirical cdf
ax2.bar(range(1,7), np.cumsum(counts[1:7]/n))
```



Try simulating with different number of rolls n and describe how this affects the accuracy of the empirical *pdf* compared to the *pdf*?

Now repeat this with the unfair dice

```
# Simulate an unfair dice
# Number of simulated realizations
n = 30
# Draw independently from the set (1,2,3,4,5,6) with higher
# probability for a six
probs = [1/7, 1/7, 1/7, 1/7, 1/7, 2/7]
xUnfair = np.random.choice(range(1, 7), size=n, replace=True, p=probs)
counts = np.bincount(xUnfair)
# Plot the pdf
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(8,4))
ax1.bar(range(1,7), probs, color='red')
# Plot the empirical pdf
ax1.bar(range(1,7), counts[1:7]/n)
# Plot the cdf
ax2.bar(range(1,7), np.cumsum(probs), color='red')
# Add the empirical cdf
ax2.bar(range(1,7), np.cumsum(counts[1:7]/n))
```



Compare the fair and the unfair dice simulations:

How did the empirical *pdf* change?

By simply observing the empirical *pdf* can we be sure to distinguish between the fair and the unfair dice?

How does the number of rolls n affect how well we can distinguish the two dices?

One reason to simulate becomes quite clear here: it would take considerably more time to actually carry out these experiments. Furthermore, sometimes calculating the theoretical properties of random variables (e.g. products of several random variables etc.) are impossible and simulations can be a useful way to obtain such results.

Random number sequences generated with software algorithms have the properties of real random numbers, e.g. they are independent, but are in fact deterministic sequences depending on a *seed*, which sets an initial value of the sequence. Therefore they are named *pseudo random numbers*, since they behave like and are used as random numbers in simulations, but are in fact deterministic sequences.

|||| Remark 2.12 Random numbers and seed in Python

In Python the initial values can be set with a single number called the *seed* as demonstrated with the following Python code. As default the seed is created from the time of start-up of a new instance of Python. A way to generate truly (i.e. non-pseudo) random numbers can be to sample some physical phenomena, for example atmospheric noise as done at www.random.org.

```
# The random numbers generated depends on the seed

# Set the seed
np.random.seed(127)
# Generate a (pseudo) random sequence
print(np.random.rand(10))


[0.524 0.040 0.186 0.773 0.552 0.086 0.441 0.716 0.671 0.473]

# Generate again and see that new numbers are generated
print(np.random.rand(10))


[0.906 0.105 0.175 0.089 0.650 0.071 0.460 0.907 0.094 0.633]

# Set the seed and the same numbers as before just after the
# seed was set are generated
np.random.seed(127)
print(np.random.rand(10))


[0.524 0.040 0.186 0.773 0.552 0.086 0.441 0.716 0.671 0.473]
```

2.2.2 Mean and variance

In Chapter 1 the *sample mean* and the *sample variance* were introduced. They indicate respectively the centring and the spread of the observations in a sample. In this section the *mean* and *variance* are introduced. They are properties of the distribution of a random variable, they are called *population parameters*. The mean indicates where the distribution is centred. The variance indicates

the spread of the distribution.

Mean and expected value

The *mean* (μ) of a random variable is the population parameter which most statistical analysis focus on. It is formally defined as a function $\mathrm{E}(X)$: the *expected value* of the random variable X.

|||| Definition 2.13 Mean value

The mean of a discrete random variable X is

$$\mu = \mathrm{E}(X) = \sum_{j=1}^{\infty} x_j f(x_j), \tag{2-15}$$

where x_j is the value and $f(x_j)$ is the probability that X takes the outcome value x_j.

The mean is simply the weighted average over all possible outcome values, weighted with the corresponding probability. As indicated in the definition there might be infinitely many possible outcome values, hence, even if the total sum of probabilities is one, then the probabilities must go sufficiently fast to zero for increasing values of X in order for the sum to be defined.

|||| Example 2.14

For the fair dice the mean is calculated by

$$\mu_{x^{\text{fair}}} = \mathrm{E}(X^{\text{fair}}) = 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = 3.5,$$

for the unfair dice the mean is

$$\mu_{x^{\text{unfair}}} = \mathrm{E}(X^{\text{unfair}}) = 1\frac{1}{7} + 2\frac{1}{7} + 3\frac{1}{7} + 4\frac{1}{7} + 5\frac{1}{7} + 6\frac{2}{7} \approx 3.86.$$

The mean of a random variable express the limiting value of an average of many outcomes. If a fair dice is rolled a really high number of times the sample mean of these will be very close to 3.5. For the statistical reasoning related to the use of a sample mean as an estimate for μ, the same property ensures that envisioning

many sample means (with the same n), a meta like thinking, then the mean of such many repeated sample means will be close to μ.

After an experiment has been carried out n times then the *sample mean* or *average* can be calculated as previously defined in Chapter 1

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i}^{n} x_i. \tag{2-16}$$

It is called a *statistic*, which means that it is calculated from a sample. Note the use of a hat in the notation over μ: this indicates that it is an *estimate* of the real underlying mean.

Our intuition tells us that the estimate ($\hat{\mu}$) will be close to true underlying expectation (μ) when n is large. This is indeed the case, to be more specific $\mathrm{E}\left[\frac{1}{n}\sum X_i\right] = \mu$ (when $\mathrm{E}[X_i] = \mu$), and we say that the average is a central estimator for the expectation. The exact quantification of these qualitative statements will be covered in Chapter 3.

Now play a little around with the mean and the sample mean with some simulations.

|||| Example 2.15 Simulate and estimate the mean

Carrying out the experiment more than one time an estimate of the mean, i.e. the sample mean, can be calculated. Simulate rolling the fair dice

```
# Number of realizations
n = 30
# Simulate rolls with a fair dice
xFair = np.random.choice(range(1, 7), size=n, replace=True)
# Calculate the sample mean
xFair.sum()/n
```

```
np.float64(3.3333333333333335)
```

```
# or
xFair.mean()
```

```
np.float64(3.3333333333333335)
```

Let us see what happens with the sample mean of the unfair dice by simulating the same number of rolls

```
# Simulate an unfair dice

# n realizations
probs = [1/7, 1/7, 1/7, 1/7, 1/7, 2/7]  # Higher probability for a six
xUnfair = np.random.choice(range(1, 7), size=n, replace=True, p=probs)
# Calculate the sample mean
xUnfair.mean()


np.float64(4.166666666666667)
```

Consider the mean of the unfair dice and compare it to the mean of the fair dice (see Example 2.14). Is this in accordance with your simulation results?

Let us again turn to how much we can "see" from the simulations and the impact of the number of realizations n on the estimation. In statistics the term *information* is used to refer to how much information is embedded in the data, and therefore how accurate different properties (parameters) can be estimated from the data.

Repeat the simulations several times with $n = 30$. By simply comparing the sample means from a single simulation can it then be determined if the two means really are different?

Repeat the simulations several times and increase n. What happens with to the 'accuracy' of the sample mean compared to the real mean? and thereby how well it can be inferred if the sample means are different?

Does the information embedded in the data increase or decrease when n is increased?

Variance and standard deviation

The second most used population parameter is the *variance* (or standard deviation). It is a measure describing the spread of the distribution, more specifically the spread away from the mean.

Definition 2.16 Variance

The variance of a discrete random variable X is

$$\sigma^2 = \mathrm{V}(X) = \mathrm{E}[(X-\mu)^2] = \sum_{i=1}^{\infty}(x_i-\mu)^2 f(x_i), \tag{2-17}$$

where x_i is the outcome value and $f(x_i)$ is the *pdf* of the ith outcome value. The *standard deviation* σ is the square root of the variance.

The variance is the expected value (i.e. average (weighted by probabilities)) of the squared distance between the outcome and the mean value.

Remark 2.17

Notice that the variance cannot be negative.

The standard deviation is measured on the same scale (same units) as the random variable, which is not case for the variance. Therefore the standard deviation can much easier be interpreted, when communicating the spread of a distribution.

Consider how the expected value is calculated in Equation (2-15). One can think of the squared distance as a new random variable that has an expected value which is the variance of X.

Example 2.18

The variance of rolls with the fair dice is

$$\begin{aligned}\sigma^2_{x^{\text{fair}}} &= \mathrm{E}[(X^{\text{fair}} - \mu_{X^{\text{fair}}})^2] \\ &= (1-3.5)^2\frac{1}{6} + (2-3.5)^2\frac{1}{6} + (3-3.5)^2\frac{1}{6} + (4-3.5)^2\frac{1}{6} + (5-3.5)^2\frac{1}{6} + (6-3.5)^2\frac{1}{6} \\ &= \frac{70}{24} \\ &\approx 2.92.\end{aligned}$$

It was seen in Chapter 1, that after an experiment has been carried out n times

the *sample variance* can be calculated as defined previously by

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{2-18}$$

and hence thereby also *sample standard deviation* s.

Again our intuition tells us that the statistic (e.g. *sample variance*), should in some sense converge to the true variance - this is indeed the case and the we call the sample variance a central estimator for the true underlying variance. This convergence will be quantified for a special case in Chapter 3.

The sample variance is calculated by:

- Take the sample mean: $\bar{x}$
- Take the distance for each sample: $x_i - \bar{x}$
- Finally, take the average of the squared distances (using $n-1$ in the denominator, see Chapter 1)

|||| Example 2.19 Simulate and estimate the variance

Return to the simulations. First calculate the sample variance from n rolls of a fair dice

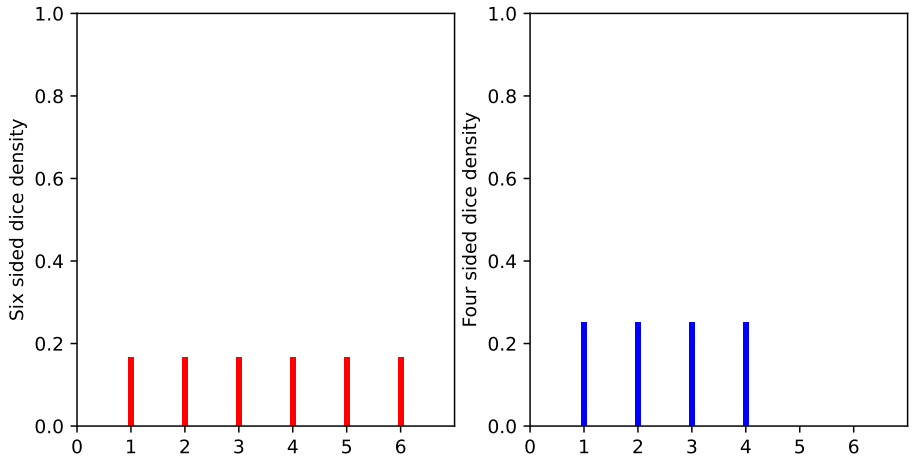
```
# Simulate a fair dice and calculate the sample variance

# Number of realizations
n = 30
# Simulate
xFair = np.random.choice(range(1,7), size=n, replace=True)
# Calculate the distance for each sample to the sample mean
distances = xFair - xFair.mean()
# Calculate the average of the squared distances
sum(distances**2)/(n-1)


np.float64(2.791954022988505)


# Or use the built in function
xFair.var(ddof=1)


np.float64(2.7919540229885054)
```

Let us then try to play with variance in the dice example. Let us now consider a four-sided dice. The *pdf* is

x	1	2	3	4
$F_{X^{\text{fairFour}}}(x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

Plot the *pdf* for both the six-sided dice and the four-sided dice

```
# Plot the pdf of the six-sided dice and the four-sided dice
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(8,4))
ax1.bar(range(1, 7), [1/6] * 6, color='red')
ax2.bar(range(1, 5), [1/4] * 4, color='blue')
```




```
# Calculate the means and variances of the dices

# The means
muXSixsided = np.sum(np.array([1,2,3,4,5,6])*1/6)
muXFoursided = np.sum(np.array([1,2,3,4])*1/4)
# The variances
print(np.sum((np.array([1,2,3,4,5,6]) - muXSixsided)**2 * 1/6))


2.916666666666667


print(np.sum((np.array([1,2,3,4]) - muXFoursided)**2 * 1/4))


1.25
```

Which dice outcome has the highest variance? is that as you had anticipated?

2.3 Discrete distributions

In this section the discrete distributions included in the material are presented. See the overview of all distributions in the collection of formulas Section A.2.1.

In the Python library Scipy, implementations of many different distributions are available. For each distribution at least the following is available

- The *pdf* is available by using `'.pmf()'`, e.g. for the binomial distribution `scipy.stats.binom.pmf()` (use `'.pmf()'` for discrete cases and `'.pdf()'` for continuous cases)
- The *cdf* is available by using `'.cdf()'`, e.g. `scipy.stats.binom.cdf()`
- The quantiles by using `'.ppf()'`, e.g. `scipy.stats.binom.ppf()`
- Random number generation by using `'.rvs()'`, e.g. `scipy.stats.binom.rvs()`

Examples of these functions are demonstrated below in this section for the discrete and later for the continuous distributions, see them demonstrated for the normal distribution in Example 2.45.

2.3.1 Binomial distribution

The binomial distribution is a very important discrete distribution and appears in many applications, it is presented in this section. In statistics it is typically used for proportions as explained in Chapter 7.

If an experiment has two possible outcomes (e.g. failure or success, no or yes, 0 or 1) and is repeated more than one time, then the number of successes may be binomial distributed. For example the number of heads obtained after a certain number of flips with a coin. Each repetition must be independent. In relation to random sampling this corresponds to successive draws with replacement (think of drawing notes from a hat, where after each draw the note is put back again, i.e. the drawn number is replaced again).

Definition 2.20 Binomial distribution

Let the random variable X be binomial distributed

$$X \sim B(n, p), \tag{2-19}$$

where n is number of independent draws and p is the probability of a success in each draw.
The binomial *pdf* describes probability of obtaining x successes

$$f(x; n, p) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{2-20}$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}, \tag{2-21}$$

is the number of distinct sets of x elements which can be chosen from a set of n elements. Remember that $n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1$.

Theorem 2.21 Mean and variance

The mean of a binomial distributed random variable is

$$\mu = np, \tag{2-22}$$

and the variance is

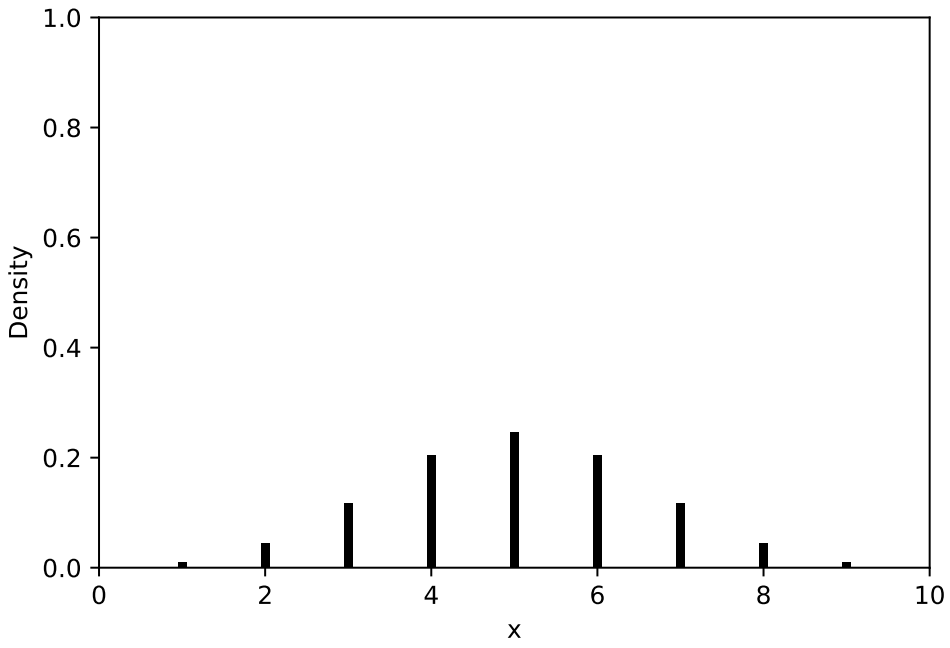
$$\sigma^2 = np(1-p). \tag{2-23}$$

Actually this can be proved by calculating the mean using Definition 2.13 and the variance using Definition 2.16.

|||| Example 2.22 Simulation with a binomial distribution

The binomial distribution for 10 flips with a coin describe probabilities of getting x heads (or equivalently tails)

```
# Simulate a binomial distributed experiment

# Number of flips
nFlips = 10
# The possible outcomes are (0,1,...,nFlips)
xSeq = list(range(0,nFlips))
# Use the binom.pmf() function which returns the pdf
pdfSeq = stats.binom.pmf(xSeq, nFlips, 0.5)
# Plot the density
plt.bar(xSeq, pdfSeq, color='black',  width=0.1)
```



Example 2.23 Simulate 30 successive dice rolls

In the previous examples successive rolls of a dice was simulated. If a random variable which counts the number of sixes obtained X^{six} is defined, it follows a binomial distribution

```
# Simulate 30 successive dice rolls
Xfair = np.random.choice(range(1,7), size=30, replace=True)
# Count the number sixes obtained
sum(Xfair==6)


np.int64(5)


# This is equivalent to
stats.binom.rvs(n=30, p=1/6)


4
```

2.3.2 Hypergeometric distribution

The hypergeometric distribution describes number of successes from successive draws without replacement.

|||| Definition 2.24 Hypergeometric distribution

Let the random variable X be the number of successes in n draws without replacement. Then X follows the hypergeometric distribution

$$X \sim H(n, a, N), \tag{2-24}$$

where a is the number of successes in the N elements large population. The probability of obtaining x successes is described by the hypergeometric *pdf*

$$f(x; n, a, N) = P(X = x) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}. \tag{2-25}$$

The notation

$$\binom{a}{b} = \frac{a!}{b!(a-b)!}, \tag{2-26}$$

represents the number of distinct sets of b elements which can be chosen from a set of a elements.

|||| Theorem 2.25 Mean and variance

The mean of a hypergeometric distributed random variable is

$$\mu = n\frac{a}{N}, \tag{2-27}$$

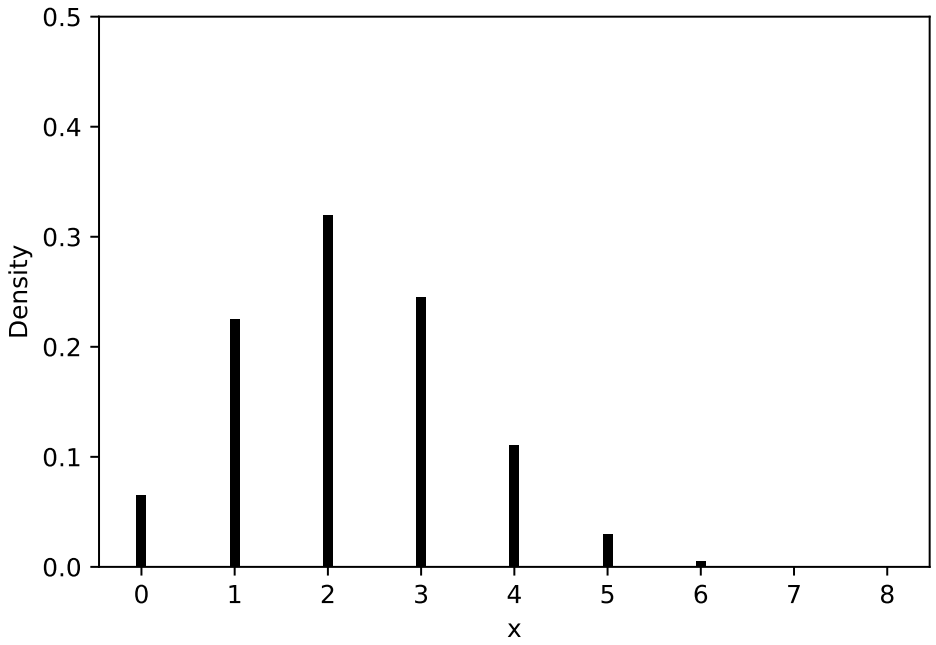
and the variance is

$$\sigma^2 = n\frac{a(N-a)}{N^2}\frac{N-n}{N-1}. \tag{2-28}$$

|||| Example 2.26 Lottery probabilities using the hypergeometric distribution

A lottery drawing is a good example where the hypergeometric distribution can be applied. The numbers from 1 to 90 are put in a bowl and randomly drawn without replacement (i.e. without putting back the number when it has been drawn). Say that you have the sheet with 8 numbers and want to calculate the probability of getting all 8 numbers in 25 draws.

```
# The probability of getting x numbers of the sheet in 25 drawings

# Number of successes in the population
a = 8
# Size of the population
N = 90
# Number of draws
n = 25
# Plot the pdf, note: parameters names are different in the python-function (here using
plt.bar(np.arange(0,9), stats.hypergeom.pmf(np.arange(0,9), N, a, n), color='black',  w
```



2.3.3 Poisson distribution

The Poisson distribution describes the probability of a given number of events occurring in a fixed interval if these events occur with a known average rate and independently of the distance to the last event. Often it is events in a time interval, but can as well be counts in other intervals, e.g. of distance, area or volume. In statistics the Poisson distribution is usually applied for analyzing for example counts of: arrivals, traffic, failures and breakdowns.

Definition 2.27 Poisson distribution

Let the random variable X be Poisson distributed

$$X \sim Po(\lambda), \tag{2-29}$$

where λ is the rate (or intensity): the average number of events per interval. The Poisson *pdf* describes the probability of x events in an interval

$$f(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}. \tag{2-30}$$

Theorem 2.28 Mean and variance

A Poisson distributed random variable X has exactly the rate λ as the mean

$$\mu = \lambda, \tag{2-31}$$

and variance

$$\sigma^2 = \lambda. \tag{2-32}$$

Example 2.29

The Poisson distribution is typically used to describe phenomena such as:

- the number radioactive particle decays per time interval, i.e. the number of clicks per time interval of a Geiger counter
- calls to a call center per time interval (λ does vary over the day)
- number of mutations in a given stretch of DNA after a certain amount of radiation
- goals scored in a soccer match

One important feature is that the rate can be scaled, such that probabilities of occurrences in other interval lengths can be calculated. Usually the rate is denoted with the interval length, for example the hourly rate is denoted as λ^{hour}

and can be scaled to the minutely rate by

$$\lambda^{\text{minute}} = \frac{\lambda^{\text{hour}}}{60}, \tag{2-33}$$

such the probabilities of x events per minute can be calculated with the Poisson *pdf* with rate λ^{minute}.

|||| Example 2.30 Rate scaling

You are enjoying a soccer match. Assuming that the scoring of goals per match in the league is Poisson distributed and on average 3.4 goals are scored per match. Calculate the probability that no goals will be scored while you leave the match for 10 minutes.

Let $\lambda^{90\text{minutes}} = 3.4$ be goals per match and scale this to the 10 minute rate by

$$\lambda^{10\text{minutes}} = \frac{\lambda^{90\text{minutes}}}{9} = \frac{3.4}{9}. \tag{2-34}$$

Let X be the number of goals in 10 minute intervals and use this to calculate the probability of no events a 10 minute interval by

$$P(X = 0) = f(0, \lambda^{10\text{minutes}}) \approx 0.685, \tag{2-35}$$

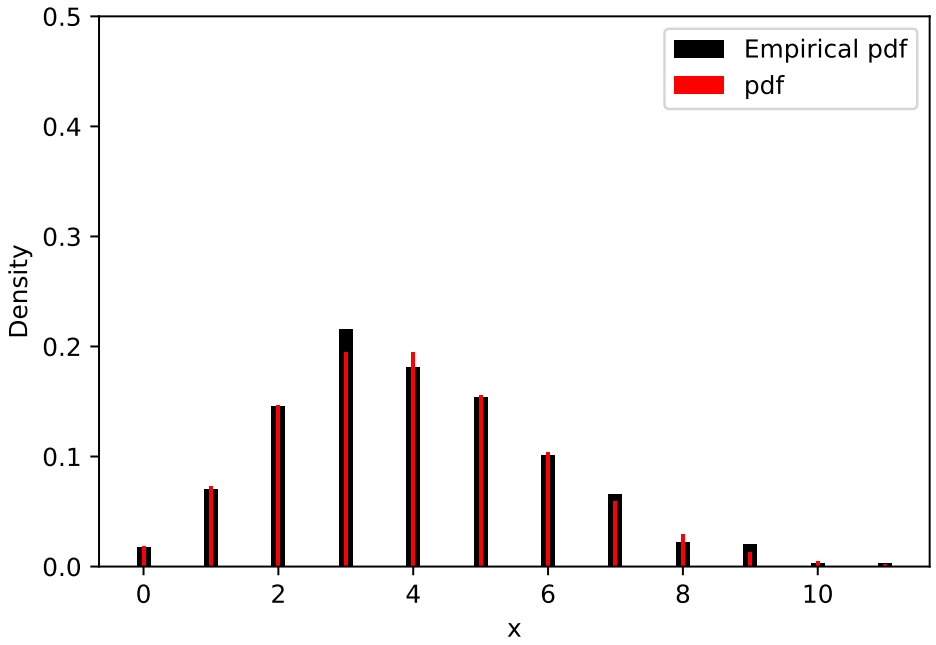
which was found with the following code

```
# Probability of no goals in 10 minutes

# The Poisson pdf (using poisson.pmf() function)
stats.poisson.pmf(0, 3.4/9)


np.float64(0.6853827910309876)
```

Example 2.31 Poisson distributed random variable

Simulate a Poisson distributed random variable to see the Poisson distribution

```
# Simulate a Poisson random variable

# The mean rate of events per interval
lamb = 4
# Number of realizations
n = 1000
# Simulate
x = stats.poisson.rvs(lamb, size=n)
# Plot the empirical pdf
values, counts = np.unique(x, return_counts=True)
plt.bar(values, counts/n, color='black',  width=0.2, label='Empirical pdf')
# Add the pdf to the plot
plt.bar(values, stats.poisson.pmf(values, lamb), color='red',  width=0.05, label='pdf')
plt.show()
```



2.4 Continuous random variables

If an outcome of an experiment takes a continuous value, for example: a distance, a temperature, a weight, etc., then it is represented by a continuous random variable.

|||| Definition 2.32 Density and probabilities

The *pdf* of a continuous random variable X is a non-negative function for all possible outcomes

$$f(x) \geq 0 \text{ for all } x, \tag{2-36}$$

and has an area below the function of one

$$\int_{-\infty}^{\infty} f(x)dx = 1. \tag{2-37}$$

It defines the probability of observing an outcome in the range from a to b by

$$P(a < X \leq b) = \int_{a}^{b} f(x)dx. \tag{2-38}$$

For the discrete case the probability of observing an outcome x is equal to the *pdf* of x, but this is not the case for a continuous random variable, where

$$P(X = x) = P(x < X \leq x) = \int_{x}^{x} f(u)du = 0, \tag{2-39}$$

i.e. the probability for a continuous random variable to be realized at a single number $P(X = x)$ is zero.

The plot in Figure 2.1 shows how the area below the *pdf* represents the probability of observing an outcome in a range. Note that the normal distribution is used here for the examples, it is introduced in Section 2.5.2.

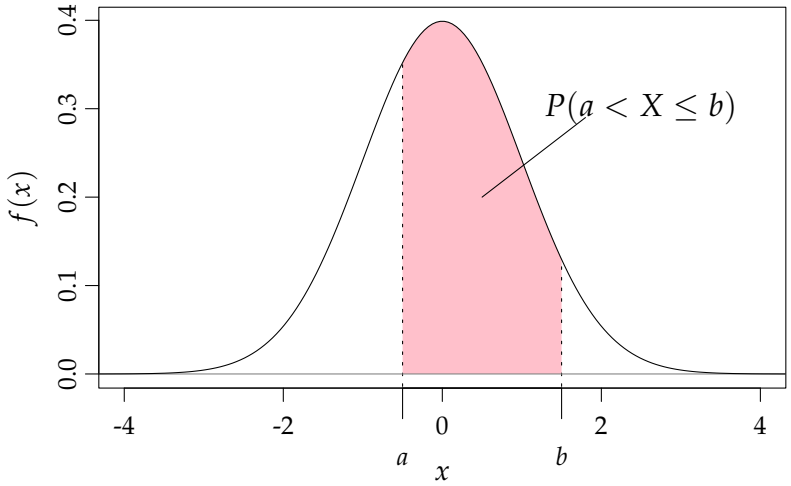


Figure 2.1: The probability of observing the outcome of X in the range between a and b is the area below the *pdf* spanning the range, as illustrated with the coloured area.

|||| Definition 2.33 Distribution

The *cdf* of a continuous variable is defined by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du, \tag{2-40}$$

and has the properties (in both the discrete and continuous case): the *cdf* is non-decreasing and

$$\lim_{x \to -\infty} F(x) = 0 \text{ and } \lim_{x \to \infty} F(x) = 1. \tag{2-41}$$

The relation between the *cdf* and the *pdf* is

$$P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)dx, \tag{2-42}$$

as illustrated in Figures 2.1 and 2.2.

Also as the *cdf* is defined as the integral of the *pdf*, the *pdf* becomes the derivative of the *cdf*

$$f(x) = \frac{d}{dx}F(x) \tag{2-43}$$

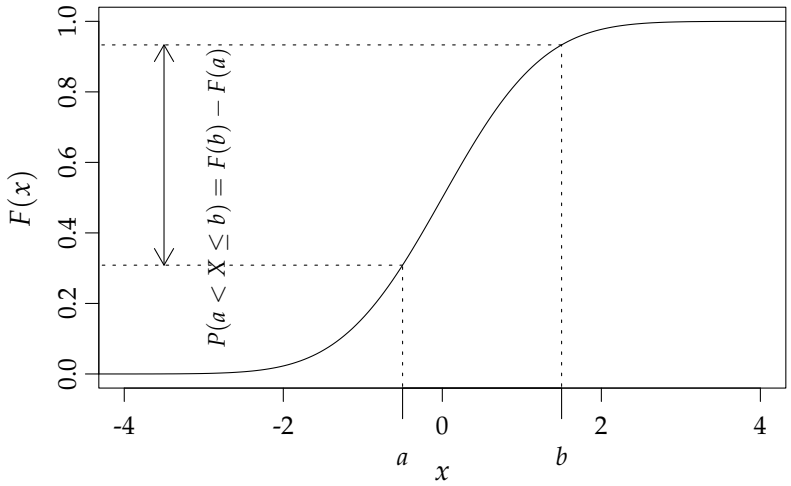


Figure 2.2: The probability of observing the outcome of X in the range between a and b is the distance between $F(a)$ and $F(b)$.

2.4.1 Mean and Variance

Definition 2.34 Mean and variance

For a continuous random variable the mean or expected value is

$$\mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) dx, \tag{2-44}$$

hence similar as for the discrete case the outcome is weighted with the *pdf*. The variance is

$$\sigma^2 = \mathrm{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx, \tag{2-45}$$

The differences between the discrete and the continuous case can be summed up in two points:

- In the continuous case integrals are used, in the discrete case sums are used.
- In the continuous case the probability of observing a single value is always zero. In the discrete case it can be positive or zero.

2.5 Continuous distributions

2.5.1 Uniform distribution

A random variable following the *uniform distribution* has equal density at any value within a defined range.

Definition 2.35 Uniform distribution

Let X be a uniform distributed random variable

$$X \sim U(\alpha, \beta), \tag{2-46}$$

where α and β defines the range of possible outcomes. It has the *pdf*

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{for } x \in [\alpha, \beta] \\ 0 & \text{otherwise} \end{cases}. \tag{2-47}$$

The uniform *cdf* is

$$F(x) = \begin{cases} 0 & \text{for } x < \alpha \\ \frac{x-\alpha}{\beta-\alpha} & \text{for } x \in [\alpha, \beta) \\ 1 & \text{for } x \geq \beta \end{cases}. \tag{2-48}$$

In Figure 2.3 the uniform *pdf* and *cdf* are plotted.

Theorem 2.36 Mean and variance of the uniform distribution

The mean of a uniform distributed random variable X is

$$\mu = \frac{1}{2}(\alpha + \beta), \tag{2-49}$$

and the variance is

$$\sigma^2 = \frac{1}{12}(\beta - \alpha)^2. \tag{2-50}$$

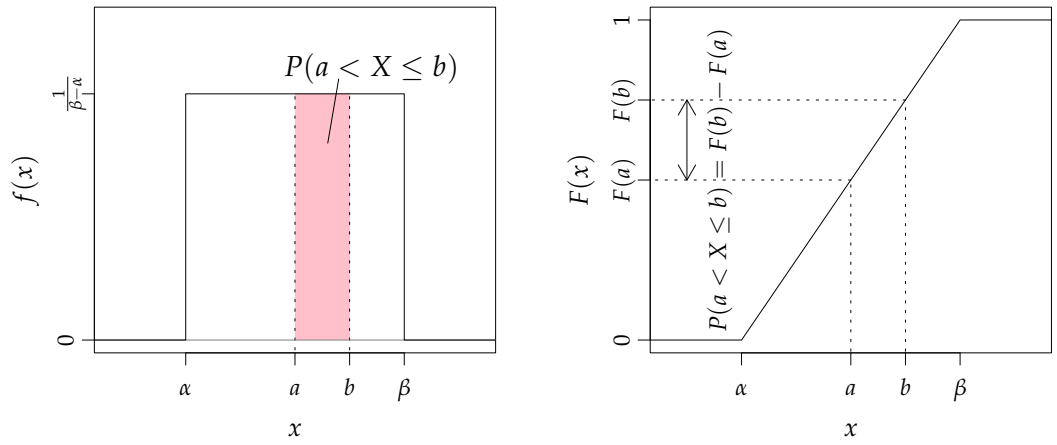


Figure 2.3: The uniform distribution *pdf* and cdf.

2.5.2 Normal distribution

The most famous continuous distribution is the normal distribution for many reasons. Often it is also called the Gaussian distribution. The normal distribution appears naturally for many phenomena and is therefore used in extremely many applications, which will be apparent in later chapters of the book.

|||| Definition 2.37 Normal distribution

Let X be a normal distributed random variable

$$X \sim N(\mu, \sigma^2), \tag{2-51}$$

where μ is the mean and σ^2 is the variance (remember that the standard deviation is σ). Note that the two parameters are actually the mean and variance of X.
It follows the normal *pdf*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{2-52}$$

and the normal *cdf*

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du. \tag{2-53}$$

Theorem 2.38 Mean and variance

The mean of a Normal distributed random variable is

$$\mu, \tag{2-54}$$

and the variance is

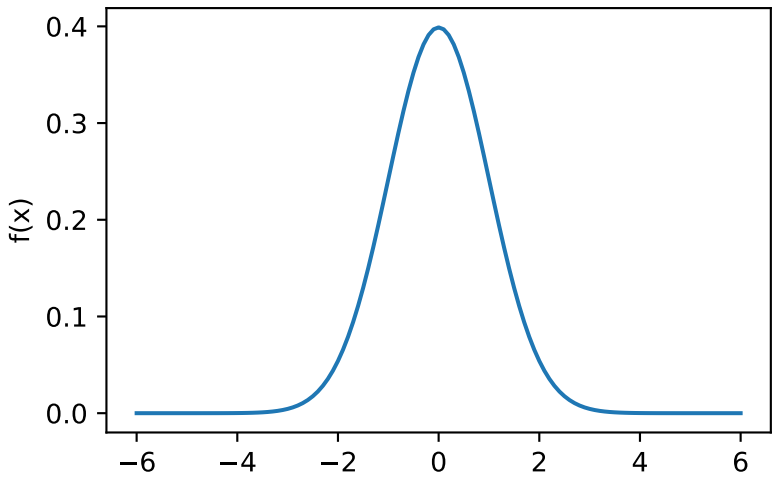
$$\sigma^2. \tag{2-55}$$

Hence simply the two parameters defining the distribution.

Example 2.39 The normal pdf

Example: Let us play with the normal *pdf*

```
# Play with the normal distribution

# The mean and standard deviation
muX = 0
sigmaX = 1
# A sequence of x values
xSeq = np.arange(-6, 6.1, 0.1)
##
pdfX = 1/(sigmaX*np.sqrt(2*np.pi)) * np.exp(-(xSeq-muX)**2/(2*sigmaX**2))
# Plot the pdf
plt.plot(xSeq, pdfX)
plt.ylabel('f(x)')
```



Try with different values of the mean and standard deviation. Describe how this change the position and spread of the *pdf*?

Theorem 2.40 Linear combinations of normal random variables

Let $X_1, \ldots, X_n$ be independent normal random variables, then any linear combination of $X_1, \ldots, X_n$ will follow a normal distribution, with mean and variance given in Theorem 2.56.

Use the mean and variance identities introduced in Section 2.7 to find the mean and variance of the linear combination as exemplified here:

Example 2.41

Consider two normal distributed random variables

$$X_1 \sim N(\mu_{X_1}, \sigma^2_{X_1}) \quad \text{and} \quad X_2 \sim N(\mu_{X_2}, \sigma^2_{X_2}). \tag{2-56}$$

The difference

$$Y = X_1 - X_2, \tag{2-57}$$

is normal distributed

$$Y \sim N(\mu_Y, \sigma^2_Y), \tag{2-58}$$

where the mean is

$$\mu_Y = \mu_{X_1} - \mu_{X_2}, \tag{2-59}$$

and

$$\sigma^2_Y = \sigma^2_{X_1} + \sigma^2_{X_2}, \tag{2-60}$$

where the mean and variance identities introduced in Section 2.7 have been used.

Standard normal distribution

Definition 2.42 Standard normal distribution

The standard normal distribution is the normal distribution with zero mean and unit variance

$$Z \sim N(0,1), \tag{2-61}$$

where Z is the standardized normal random variable.

Historically before the widespread use of computers the standardized random variables were used a lot, since it was not possible to easily evaluate the *pdf* and *cdf*, instead they were looked up in tables for the standardized distributions. This was smart since transformation into standardized distributions requires only a few simple operations.

Theorem 2.43 Transformation to the standardized normal random variable

A normal distributed random variable X can be transformed into a standardized normal random variable by

$$Z = \frac{X - \mu}{\sigma}. \tag{2-62}$$

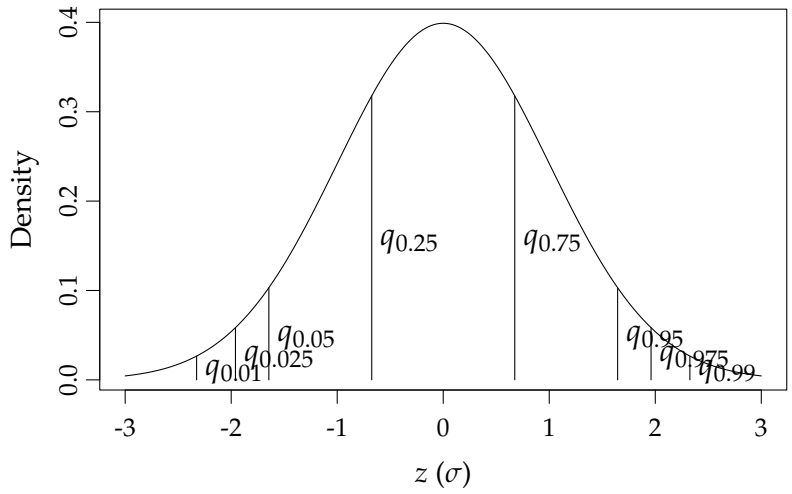
Example 2.44 Quantiles in the standard normal distribution

The most used quantiles (or percentiles) in the standard normal distribution are

Percentile	1%	2.5%	5%	25%	75%	95%	97.5%	99%
Quantile	0.01	0.025	0.05	0.25	0.75	0.95	0.975	0.99
Value	-2.33	-1.96	-1.64	-0.67	0.67	1.64	1.96	2.33

Note that the values can be considered as standard deviations (i.e. for Z the standardized normal then $\sigma_Z = 1$), which holds for any normal distribution.

The most used quantiles are marked on the plot



Note that the units on the x-axis is in standard deviations.

Normal *pdf* details

In order to get insight into how the normal distribution is formed consider the following steps. In Figure 2.4 the result of each step is plotted:

1. Take the distance to the mean: $x - \mu$
2. Square the distance: $(x - \mu)^2$
3. Make it negative and scale it: $\frac{-(x-\mu)^2}{(2\sigma^2)}$
4. Take the exponential: $e^{\frac{-(x-\mu)^2}{(2\sigma^2)}}$
5. Finally, scale it to have an area of one: $\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{(2\sigma^2)}}$

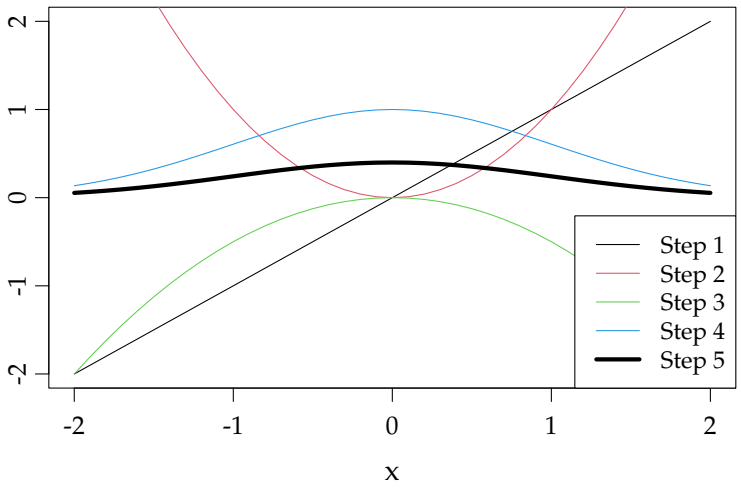


Figure 2.4: The steps involved in calculating the normal distribution *pdf*.

|||| Example 2.45 Python functions for the normal distribution

In Python functions to generate values from many distributions are implemented. For the normal distribution the following functions are available:

```
# Do it for a sequence of x values
xSeq = np.arange(-3, 4)
# The pdf
stats.norm.pdf(xSeq, 0, 1)


array([0.004, 0.054, 0.242, 0.399, 0.242, 0.054, 0.004])


# The cdf
stats.norm.cdf(xSeq, 0, 1)


array([0.001, 0.023, 0.159, 0.500, 0.841, 0.977, 0.999])


# The quantiles
stats.norm.ppf([0.01,0.025,0.05,0.5,0.95,0.975,0.99], 0, 1)


array([-2.326, -1.960, -1.645,  0.000,  1.645,  1.960,  2.326])


# Generate random normal distributed realizations
stats.norm.rvs(0, 1, size=10)


array([-1.043,  0.050, -0.592, -0.840,  0.460,  0.150,  0.021, -1.221,
       -0.638, -1.024])


# Calculate the probability that the outcome of X is between a and b
a = 0.2
b = 0.8
stats.norm.cdf(b, 0, 1) - stats.norm.cdf(a, 0, 1)


np.float64(0.20888489197750038)


# See more details in online documentation for scipy.stats.norm
```

Use the functions to make a plot of the normal *pdf* with marks of the 2.5%, 5%, 95%, 97.5% quantiles.

Make a plot of the normal *pdf* and a histogram (empirical *pdf*) of 100 simulated realizations.

2.5.3 Log-Normal distribution

If a random variable is log-normal distributed then its logarithm is normally distributed.

Definition 2.46 Log-Normal distribution

A log-normal distributed random variable

$$X \sim LN(\alpha, \beta^2), \tag{2-63}$$

where α is the mean and β^2 is the variance of the normal distribution obtained when taking the natural logarithm to X.
The log-normal *pdf* is

$$f(x) = \frac{1}{x\sqrt{2\pi}\beta} e^{-\frac{(\ln x - \alpha)^2}{2\beta^2}}. \tag{2-64}$$

Theorem 2.47 Mean and variance of log-normal distribution

Mean of the log-normal distribution

$$\mu = e^{\alpha + \beta^2/2}, \tag{2-65}$$

and variance

$$\sigma^2 = e^{2\alpha+\beta^2}(e^{\beta^2} - 1). \tag{2-66}$$

The log-normal distribution occurs in many fields, in particular: biology, finance and many technical applications.

2.5.4 Exponential distribution

The usual application of the exponential distribution is for describing the length (usually time) between events which, when counted, follows a Poisson distribution, see Section 2.3.3. Hence the length between events which occur continuously and independently at a constant average rate.

Definition 2.48 Exponential distribution

Let X be an exponential distributed random variable

$$X \sim Exp(\lambda), \tag{2-67}$$

where λ is the average rate of events.

It follows the exponential *pdf*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}. \tag{2-68}$$

Theorem 2.49 Mean and variance of exponential distribution

Mean of an exponential distribution is

$$\mu = \frac{1}{\lambda}, \tag{2-69}$$

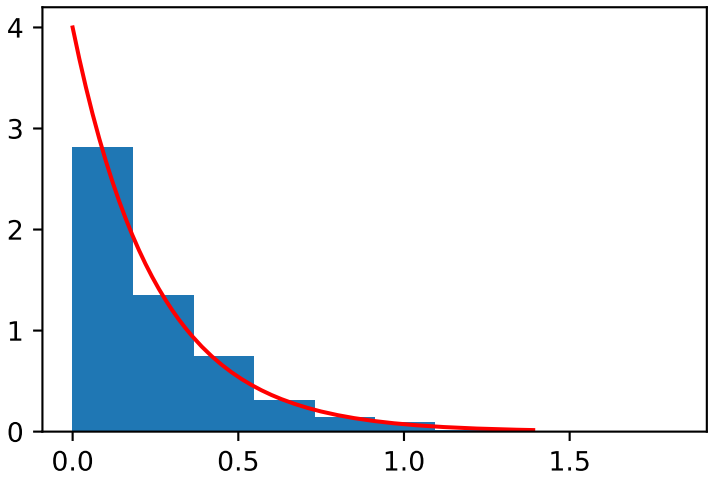
and the variance is

$$\sigma^2 = \frac{1}{\lambda^2}. \tag{2-70}$$

Example 2.50 Exponential distributed time intervals

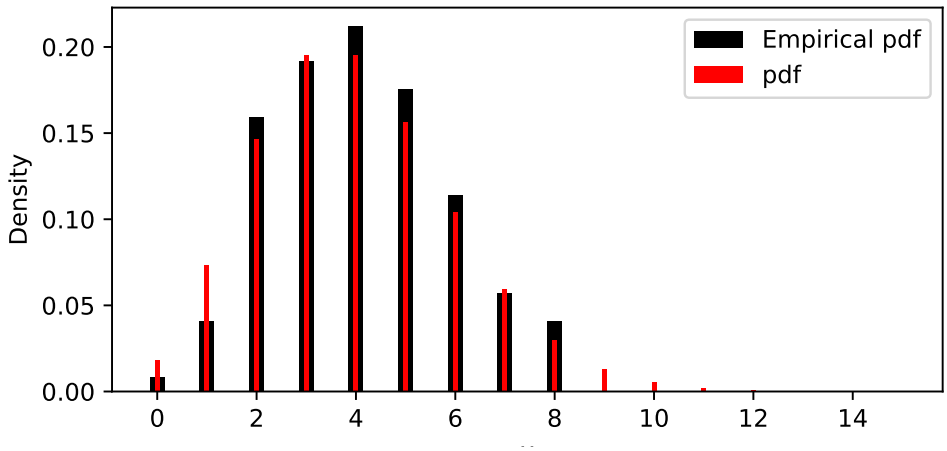
Simulate a so-called Poisson process, which has exponential distributed time interval between events

```
# Simulate exponential waiting times

# The rate parameter: events per time
lamb = 4
# Number of realizations
n = 1000
# Simulate
wait_times = stats.expon.rvs(loc=0, scale=1/lamb, size=n)
# The empirical pdf
plt.hist(wait_times, density=True)
# Add the pdf to the plot
x = np.arange(0,1.4,0.01)
plt.plot(x, stats.expon.pdf(x, loc=0, scale=1/lamb), color='red')
plt.show()
```



Furthermore check that by counting the events in fixed length intervals that they follow a Poisson distribution.

```
# Check the relation to the Poisson distribution
# by counting the events in each interval

# Sum up to get the running time
running_times = np.cumsum(wait_times)
# Use the hist function to count in intervals between the breaks,
# here 0,1,2,...
counts, bin_edges = np.histogram(running_times, bins=np.arange(np.ceil(running_times.ma
plt.bar(np.arange(len(np.bincount(counts))), np.bincount(counts)/len(counts), color='bl

# Add the Poisson pdf to the plot
poisson_pmf = stats.poisson.pmf(np.arange(0, 16), lamb)
plt.bar(np.arange(0, 16), poisson_pmf, color='red', width=0.1, label='pdf')
```



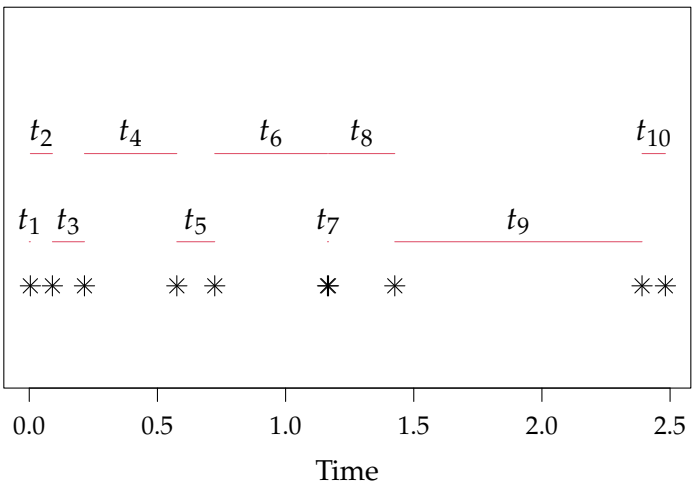


Figure 2.5: Exponential distributed time intervals between events forms a so-called Poisson process.

2.6 Simulation of random variables

The basic concept of simulation was introduced in Section 2.2.1 and we have already applied the in-built functions in Python for generating random numbers from any implemented distribution, see how in Section 2.3.1. In this section it is explained how realizations of a random variable can be generated from any probability distribution – it is the same technique for both discrete and continuous distributions.

Basically, a computer obviously cannot create a result/number, which is random. A computer can give an output as a function of an input. (Pseudo) random numbers from a computer are generated from a specially designed algorithm - called a random number generator, which once started can make the number x_{i+1} from the number x_i. The algorithm is designed in such a way that when looking at a sequence of these values, in practice one cannot tell the difference between them and a sequence of real random numbers. The algorithm needs a start input, called the "seed", as explained above Remark 2.12. Usually, you can manage just fine without having to worry about the seed issue since the program itself finds out how to handle it appropriately. Only if you want to be able to recreate exactly the same results you need to set seed value.

Actually, a basic random number generator typically generates (pseudo) random numbers between 0 and 1 in the sense that numbers in practice follow the uniform distribution on the interval 0 to 1, see Section 2.35. Actually, there is a simple way how to come from the uniform distribution to any kind of distribution:

Theorem 2.51

If $U \sim \text{Uniform}(0,1)$ and F is a distribution function for any probability distribution, then $F^{-1}(U)$ follow the distribution given by F

Recall, that the distribution function F in Python is given by the `'.cdf()'` versions of the distributions, while F^{-1} is given by the `'.ppf()'` versions.

Example 2.52 Random numbers in Python

We can generate 100 normally distributed $N(2, 3^2)$ numbers similarly the following two ways:

```
# Generate 100 normal distributed values
random_numbers = stats.norm.rvs(loc=2, scale=3, size=100)
# Similarly, generate 100 uniform distributed values from 0 to 1 and
# # put them through the inverse normal cdf
uniform_random_numbers = stats.uniform.rvs(loc=0, scale=1, size=100)
stats.norm.ppf(uniform_random_numbers, loc=2, scale=3)
```

Example 2.53 Simulating the exponential distribution

Consider the exponential distribution with $\lambda = 1/\beta = 1/2$, that is, with density function

$$f(x) = \lambda e^{-\lambda x},$$

for $x > 0$ and 0 otherwise. The distribution function is

$$F(x) = \int_0^x f(t)dt = 1 - e^{-0.5x}.$$

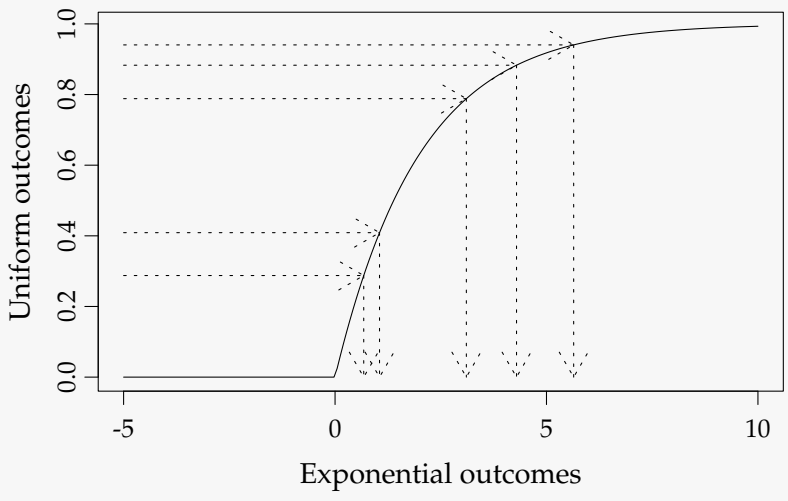
The inverse of this distribution function can be found by solving

$$u = 1 - e^{-0.5x} \Leftrightarrow x = -2\log(1-u).$$

So if random numbers $U \sim \text{Uniform}(0,1)$ then $-2\log(1-U)$ follows the exponential distribution with $\lambda = 1/2$ (and $\beta = 2$). We confirm this in the code given below:

```
# Three equivalent ways of simulating the exponential distribution
# with lambda=1/2
re1 = -2*np.log(1-stats.uniform.rvs(loc=0, scale=1, size=10000))
re2 = stats.expon.ppf(stats.uniform.rvs(loc=0, scale=1, size=10000), loc=0, scale=2)
re3 = stats.expon.rvs(loc=0, scale=2, size=10000)

# Check the means and variances of each
print(re1.mean(), re2.mean(), re2.mean())


2.0039553521283056 1.9948804494574823 1.9948804494574823


print(re1.var(), re2.var(), re2.var())


3.89520722053301 3.7967058951210935 3.7967058951210935
```

This can be illustrated by plotting the distribution function (cdf) for the exponential distribution with $\lambda = 1/2$ and 5 random outcomes



But since Python has already done all this for us, we do not really need this as long as we only use distributions that have already been implemented in Python.

2.7 Identities for the mean and variance

Rules for calculation of the mean and variance of linear combinations of independent random variables are introduced here. They are valid for both the discrete and continuous case.

Theorem 2.54 Mean and variance of linear functions

Let $Y = aX + b$ then

$$\mathrm{E}(Y) = \mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b, \tag{2-71}$$

and

$$\mathrm{V}(Y) = \mathrm{V}(aX + b) = a^2\,\mathrm{V}(X). \tag{2-72}$$

Random variables are often scaled (i.e. aX) for example when shifting units:

Example 2.55

The mean of a bike shops sale is 100 bikes per month and varies with a standard deviation of 15. They earn 200 Euros per bike. What is the mean and standard deviation of their earnings per month?

Let X be the number of bikes sold per month. On average they sell $\mu_X = 100$ bikes per month and it varies with a variance of $\sigma_X^2 = 225$. The shops monthly earnings

$$Y = 200X,$$

has then a mean and standard deviation of

$$\mu_Y = \mathrm{E}(Y) = \mathrm{E}(200X) = 200\,\mathrm{E}(X) = 200 \cdot 100 = 20000 \text{ Euro/month},$$
$$\sigma_Y = \sqrt{\mathrm{V}(Y)} = \sqrt{\mathrm{V}(200X)} = \sqrt{200^2\,\mathrm{V}(X)} = \sqrt{40000 \cdot 225} = 3000 \text{ Euro/month}.$$

Theorem 2.56 Mean and variance of linear combinations

The mean of a linear combination of independent random variables is

$$\mathrm{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1\,\mathrm{E}(X_1) + a_2\,\mathrm{E}(X_2) + \cdots + a_n\,\mathrm{E}(X_n), \tag{2-73}$$

and the variance

$$\mathrm{V}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1^2\,\mathrm{V}(X_1) + a_2^2\,\mathrm{V}(X_2) + \cdots + a_n^2\,\mathrm{V}(X_n). \tag{2-74}$$

Example 2.57

Lets take a dice example to emphasize an important point. Let X_i represent the outcome of a roll with a dice with mean μ_X and standard deviation σ_X.

Now, consider a scaling of a single roll with a dice, say five times

$$Y^{\text{scale}} = 5X_1,$$

then the mean will scale linearly

$$\mathrm{E}(Y^{\text{scale}}) = \mathrm{E}(5X_1) = 5\,\mathrm{E}(X_1) = 5\,\mu_X,$$

and the standard deviation also scales linearly

$$\sigma^2_{Y^{\text{scale}}} = \mathrm{V}(5X_1) = 5^2\,\mathrm{V}(X_1) = 5^2\,\sigma_X^2 \;\Leftrightarrow\; \sigma_{Y^{\text{scale}}} = 5\,\sigma_X.$$

Whereas *for a sum* of five rolls

$$Y^{\text{sum}} = X_1 + X_2 + X_3 + X_4 + X_5,$$

the mean will similarly scale linearly

$$\begin{aligned}\mathrm{E}(Y^{\text{sum}}) &= \mathrm{E}(X_1 + X_2 + X_3 + X_4 + X_5)\\ &= \mathrm{E}(X_1) + \mathrm{E}(X_2) + \mathrm{E}(X_3) + \mathrm{E}(X_4) + \mathrm{E}(X_5)\\ &= 5\,\mu_X,\end{aligned}$$

however the standard deviation will increase only with the square root

$$\begin{aligned}\sigma^2_{Y^{\text{sum}}} &= \mathrm{V}(X_1 + X_2 + X_3 + X_4 + X_5)\\ &= \mathrm{V}(X_1) + \mathrm{V}(X_2) + \mathrm{V}(X_3) + \mathrm{V}(X_4) + \mathrm{V}(X_5)\\ &= 5\,\sigma_X^2 \;\Leftrightarrow\\ \sigma_{Y^{\text{sum}}} &= \sqrt{5}\,\sigma_X.\end{aligned}$$

This is simply because when applying the sum to many random outcomes, then the high and low outcomes will even out each other, such that the variance will be smaller for a sum than for a scaling.

2.8 Covariance and correlation

In this chapter we have discussed mean and variance (or standard deviation), and the relation to the sample mean and sample variance, see Section 2.2.2. In Chapter 1 Section 1.4.3 we discussed the sample covariance and sample correlation, these two measures also have theoretical justification, namely covariance and correlation, which we will discuss in this section. We start by the definition of covariance.

Definition 2.58 Covariance

Let X and Y be two random variables, then the covariance between X and Y, is

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]\,. \tag{2-75}$$

Remark 2.59

It follows immediately from the definition that $\mathrm{Cov}(X,X) = \mathrm{V}(X)$ and $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

An important concept in statistics is independence (see Section 2.9 for a formal definition). We often assume that realizations (random variables) are independent. If two random variables are independent then their covariance will be zero, the reverse is however not necessarily true (see also the discussion on sample correlation in Section 1.4.3).

The following calculation rule apply to covariance between two random variables X and Y:

Theorem 2.60 Covariance between linear combinations

Let X and Y be two random variables, then

$$\mathrm{Cov}(a_0 + a_1X + a_2Y, b_0 + b_1X + b_2Y) = a_1b_1\,\mathrm{V}(X) + a_2b_2\,\mathrm{V}(Y) + (a_1b_2 + a_2b_1)\,\mathrm{Cov}(X,Y). \tag{2-76}$$

Proof

Let $Z_1 = a_0 + a_1X + a_2Y$ and $Z_2 = b_0 + b_1X + b_2Y$ then

$$\begin{aligned}\mathrm{Cov}(Z_1, Z_2) &= \mathrm{E}[(a_1(X - \mathrm{E}[X]) + a_2(Y - \mathrm{E}[Y]))(b_1(X - \mathrm{E}[X]) + b_2(Y - \mathrm{E}[Y]))] \\ &= \mathrm{E}[a_1(X - \mathrm{E}[X])b_1(X - \mathrm{E}[X])] + \mathrm{E}[a_1(X - \mathrm{E}[X])b_2(Y - \mathrm{E}[Y])] + \\ &\quad \mathrm{E}[a_2(Y - \mathrm{E}[Y])b_1(X - \mathrm{E}[X])] + \mathrm{E}[a_2(Y - \mathrm{E}[Y])b_2(Y - \mathrm{E}[Y])] \\ &= a_1b_1\,\mathrm{V}(X) + a_2b_2\,\mathrm{V}(Y) + (a_1b_2 + a_2b_2)\,\mathrm{Cov}(X, Y). \qquad (2\text{-}77)\end{aligned}$$

■

Example 2.61

Let $X \sim N(3, 2^2)$ and $Y \sim N(2, 1)$ and the covariance between X and Y given by $\mathrm{Cov}(X, Y) = 1$. What is the variance of the random variable $Z = 2X - Y$?

$$\begin{aligned}\mathrm{V}(Z) &= \mathrm{Cov}[2X - Y, 2X - Y] = 2^2\,\mathrm{V}(X) + \mathrm{V}(Y) - 4\,\mathrm{Cov}(X, Y) \\ &= 2^2 2^2 + 1 - 4 = 13.\end{aligned}$$

We have already seen in Section 1.4.3 that the sample correlation measures the observed degree of linear dependence between two random variables – calculated from samples observed on the same observational unit e.g. height and weight of people. The theoretical counterpart is the correlation between two random variables – the true linear dependence between the two variables:

Definition 2.62 Correlation

Let X and Y be two random variables with $\mathrm{V}(X) = \sigma_x^2$, $\mathrm{V}(Y) = \sigma_y^2$, and $\mathrm{Cov}(X, Y) = \sigma_{xy}$, then the correlation between X and Y is

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}. \qquad (2\text{-}78)$$

Remark 2.63

The correlation is a number between -1 and 1.

Example 2.64

Let $X \sim N(1, 2^2)$ and $\epsilon \sim N(0, 0.5^2)$ be independent random variables, find the correlation between X and $Z = X + \epsilon$.

The variance of Z is

$$\mathrm{V}(Z) = \mathrm{V}(X + \epsilon) = \mathrm{V}(X) + \mathrm{V}(\epsilon) = 4 + 0.25 = 4.25.$$

The covariance between X and Z is

$$\mathrm{Cov}(X, Z) = \mathrm{Cov}(X, X + \epsilon) = \mathrm{V}(X) = 4,$$

and hence

$$\rho_{xz} = \frac{4}{\sqrt{4.25 \cdot 4}} = 0.97.$$

2.9 Independence of random variables

In statistics the concept of independence is very important, and in order to give a formal definition of independence we will need the definition of two-dimensional random variables. The probability density function of a two-dimensional discrete random variable, called the joint probability density function, is,

|||| Definition 2.65 Joint *pdf* of two-dimensional discrete random variables

The *pdf* of a two-dimensional discrete random variable $[X, Y]$ is

$$f(x,y) = P(X = x, Y = y), \tag{2-79}$$

with the properties

$$f(x,y) \geq 0 \text{ for all } (x,y), \tag{2-80}$$

$$\sum_{\text{all } x} \sum_{\text{all } y} f(x,y) = 1. \tag{2-81}$$

|||| Remark 2.66

$P(X = x, Y = y)$ should be read: the probability of $X = x$ *and* $Y = y$.

|||| Example 2.67

Imagine two throws with an fair coin: the possible outcome of each throw is either head or tail, which will be given the values 0 and 1 respectively. The complete set of outcomes is (0,0), (0,1), (1,0), and (1,1) each with probability $1/4$. And hence the *pdf* is

$$f(x,y) = \frac{1}{4}; \quad x = \{0,1\}, y = \{0,1\},$$

further we see that

$$\begin{aligned}\sum_{x=0}^{1} \sum_{y=0}^{1} f(x,y) &= \sum_{x=0}^{1} (f(x,0) + f(x,1)) = f(0,0) + f(0,1) + f(1,0) + f(1,1) \\ &= 1.\end{aligned}$$

The formal definition of independence for a two dimensional discrete random variable is:

Definition 2.68 Independence of discrete random variables

Two discrete random variables X and Y are said to be independent if and only if

$$P(X = x, Y = y) = P(X = x)P(Y = y). \tag{2-82}$$

Example 2.69

Example 2.67 is an example of two independent random variables, to see this write the probabilities

$$P(X = 0) = \sum_{y=0}^{1} f(0, y) = \frac{1}{2},$$
$$P(X = 1) = \sum_{y=0}^{1} f(1, y) = \frac{1}{2}.$$

similarly $P(Y = 0) = \frac{1}{2}$ and $P(Y = 1) = \frac{1}{2}$, now we see that $P(X = x)P(Y = y) = \frac{1}{4}$ for all possible x and y, and hence

$$P(X = x)P(Y = y) = P(X = x, Y = y) = \frac{1}{4}.$$

Example 2.70

Now imagine that for the second throw we don't see the outcome of Y, but only observe the sum of X and Y, denote it by

$$Z = X + Y.$$

Lets find out if X and Z are independent. In this case the for all outcomes $(0, 0)$, $(0, 1)$, $(1, 1)$, $(1, 2)$ the joint *pdf* is

$$P(X = 0, Z = 0) = P(X = 0, Z = 1) = P(X = 1, Z = 1) = P(X = 1, Z = 2) = \frac{1}{4}.$$

The *pdf* for each variable is: for X

$$P(X = 0) = P(X = 1) = \frac{1}{2},$$

and for Z

$$P(Z=0) = P(Z=2) = \frac{1}{4} \text{ and } P(Z=1) = \frac{1}{2},$$

thus for example for the particular outcome $(0,0)$

$$P(X=0)P(Z=0) = \frac{1}{2}\cdot\frac{1}{4} = \frac{1}{8} \neq \frac{1}{4} = P(X=0,Z=0),$$

the *pdf*s are not equal and hence we see that X and Z are not independent.

|||| Remark 2.71

In the example above it is quite clear that X and Z cannot be independent. In real applications we do not know exactly how the outcomes are realized and therefore we will need to assume independence (or test it).

To be able to define independence of continuous random variables, we will need the *pdf* of a two-dimensional random variable:

|||| Definition 2.72 *Pdf* of two dimensional continous random variables

The *pdf* of a two-dimensional continous random variable $[X,Y]$ is a function $f(x,y)$ from $\mathbb{R}^2$ into $\mathbb{R}_+$ with the properties

$$f(x,y) \geq 0 \text{ for all } (x,y), \tag{2-83}$$

$$\int\int f(x,y)dxdy = 1. \tag{2-84}$$

Just as for one-dimensional random variables the probability interpretation is in form of integrals

$$P\big((X,Y)\in A\big) = \int_A f(x,y)dxdy, \tag{2-85}$$

where A is an area.

Example 2.73 Bivariate normal distribution

The most important two-dimensional distribution is the bivariate normal distribution

$$f(x_1,x_2) = \frac{1}{2\pi\sqrt{|\Sigma|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$
$$= \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}-\sigma_{12}^2}} e^{-\frac{\sigma_{22}(x_1-\mu_1)^2+\sigma_{11}(x_2-\mu_2)^2-2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2)}{2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}},$$

where $\boldsymbol{x} = (x_1,x_2)$, and $\boldsymbol{\mu} = [\mathrm{E}(X_1),\mathrm{E}(X_2)]$, and Σ is the so-called variance-covariance matrix with elements $(\Sigma)_{ij} = \sigma_{ij} = \mathrm{Cov}(X_i,X_j)$, note that $\sigma_{12} = \sigma_{21}$, $|\cdot|$ is the determinant, and Σ^{-1} is the inverse of Σ.

Definition 2.74 Independence of continous random variables

Two continous random variables X and Y are said to be independent if

$$f(x,y) = f(x)f(y). \tag{2-86}$$

We list here some properties of independent random variables.

Theorem 2.75 Properties of independent random variables

If X and Y are independent then

$$\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y), \tag{2-87}$$

and

$$\mathrm{Cov}(X,Y) = 0. \tag{2-88}$$

Let $X_1,\ldots,X_n$ be independent and identically distributed random variables then

$$\mathrm{Cov}(\overline{X}, X_i - \overline{X}) = 0. \tag{2-89}$$

Proof

$$\begin{aligned} \mathrm{E}(XY) &= \int\int xyf(x,y)dxdy = \int\int xyf(x)f(y)dxdy \\ &= \int xf(x)dx \int yf(y)dy = \mathrm{E}(X)\,\mathrm{E}(Y) \end{aligned} \tag{2-90}$$

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{E}[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))] \\ &= \mathrm{E}[XY] - \mathrm{E}[\mathrm{E}(X)Y] - \mathrm{E}[X\,\mathrm{E}(Y)] + \mathrm{E}(X)\,\mathrm{E}(Y) \\ &= 0. \end{aligned} \tag{2-91}$$

$$\begin{aligned} \mathrm{Cov}(\overline{X}, X_i - \overline{X}) &= \mathrm{Cov}(\overline{X}, X_i) - \mathrm{Cov}(\overline{X}, \overline{X}) \\ &= \frac{1}{n}\sigma^2 - \frac{1}{n^2}\,\mathrm{Cov}\Big(\sum X_i, \sum X_i\Big) \\ &= \frac{1}{n}\sigma^2 - \frac{1}{n^2}n\sigma^2 = 0. \end{aligned} \tag{2-92}$$

■

Remark 2.76

Note that $\mathrm{Cov}(X,Y) = 0$ does not imply that X and Y are independent. However, if X and Y follow a bivariate normal distribution, then if X and Y are uncorrelated then they are also independent.

2.10 Functions of normal random variables

This section will cover some important functions of a normal random variable. In general the question of how an arbitrary function of a random variable is distributed cannot be answered on closed form (i.e. directly and exactly calculated) – for answering such questions we must use simulation as a tool, as covered details in Chapter 4. We have already discussed simulation as a learning tool, which will also be used in this section.

The simplest function we can think of is a *linear combination of normal random variables*, which we from Theorem 2.40 know *will follow a normal distribution*. The mean and variance of this normal distribution can be calculated using the identities given in Theorem 2.56.

|||| Remark 2.77

Note that combining Theorems 2.40 and 2.75, and Remark 2.76 imply that $\overline{X}$ and $X_i - \overline{X}$ are independent.

In addition to the result given above we will cover three additional distributions: χ^2-distribution, t-distribution and the F-distribution, which are all very important for the statistical inference covered in the following chapters.

2.10.1 The χ^2-distribution

The χ^2-distribution (chi-square) is defined by:

Definition 2.78

Let X be χ^2 distributed, then its *pdf* is

$$f(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1}e^{-\frac{x}{2}}; \quad x \geq 0, \tag{2-93}$$

where $\Gamma\left(\frac{\nu}{2}\right)$ is the Γ-function and ν is the degrees of freedom.

An alternative definition (here formulated as a theorem) of the χ^2-distribution is:

Theorem 2.79

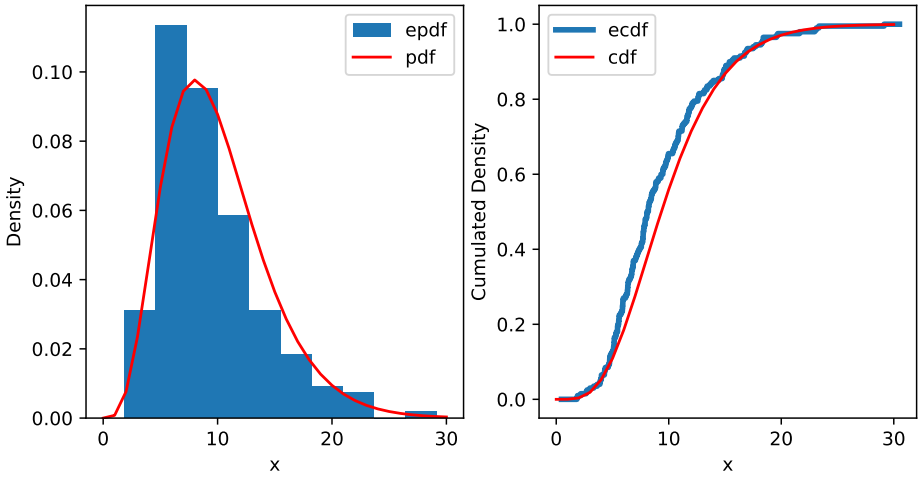
Let $Z_1, \ldots, Z_\nu$ be independent random variables following the standard normal distribution, then

$$\sum_{i=1}^{\nu} Z_i^2 \sim \chi^2(\nu). \tag{2-94}$$

We will omit the proof of the theorem as it requires more probabilty calculus than covered here. Rather a small example that illustrates how the theorem can be checked by simulation:

|||| Example 2.80 simulation of χ^2-distribution

```
# Simulate 10 realizations from a standard normal distributed variable
n = 10
stats.norm.rvs(loc=0, scale=1, size=n)
# Now repeat this 200 times and calculate the sum of squares each time
# Note: the use of the function replicate: it repeats the
#       expression in the 2nd argument k times, see ?replicate
x = [np.sum(stats.norm.rvs(loc=0, scale=1, size=n)**2) for _ in range(200)]
x = np.array(x)
# Plot the epdf of the sums and compare to the theoretical chisquare pdf
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(8,4))
ax1.hist(x, density=True, label='epdf')
ax1.plot(range(0,31), stats.chi2.pdf(range(0,31), df=n), color="red", label='pdf')
# and the ecdf compared to the cdf
stats.ecdf(x).cdf.plot(ax2, label='ecdf')
ax2.plot(range(0,31), stats.chi2.cdf(range(0,31), df=n), color="red", label='cdf')
```



In the left plot the empirical *pdf* is compared to the theoretical *pdf* and in the right plot the empirical *cdf* is compared to the theoretical *cdf*.

Theorem 2.81

Given a sample of size n from the normal distributed random variables X_i with variance σ^2, then the sample variance S^2 (viewed as random variable) can be transformed into

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}, \tag{2-95}$$

which follows the χ^2-distribution with degrees of freedom $\nu = n - 1$.

Proof

Start by rewriting the expression

$$\begin{aligned}
\frac{(n-1)S^2}{\sigma^2} &= \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 \\
&= \sum_{i=1}^{n} \left(\frac{X_i - \mu + \mu - \overline{X}}{\sigma} \right)^2 \\
&= \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 + \sum_{i=1}^{n} \left(\frac{\overline{X} - \mu}{\sigma} \right)^2 - 2 \sum_{i=1}^{n} \frac{(\overline{X} - \mu)(X_i - \mu)}{\sigma^2} \\
&= \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 + n \left(\frac{\overline{X} - \mu}{\sigma} \right)^2 - 2n \left(\frac{\overline{X} - \mu}{\sigma} \right)^2 \\
&= \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 - \left(\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \right)^2,
\end{aligned} \tag{2-96}$$

we know that $\frac{X_i - \mu}{\sigma} \sim N(0,1)$ and $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, and hence the left hand side is a $\chi^2(n)$ distributed random variable minus a $\chi^2(1)$ distributed random variable (also $\overline{X}$ and S^2 are independent, see Theorems 2.75, and 2.40, and Remark 2.76). Hence the left hand side must be $\chi^2(n-1)$.

■

If someone claims that a sample comes from a specific normal distribution (i.e. $X_i \sim N(\mu, \sigma^2)$), then we can examine probabilities of specific outcomes of the sample variance. Such calculation will be termed hypethesis test in later chapters.

|||| Example 2.82 Milk dose machines

A manufacture of machines for dosing milk claims that their machines can dose with a precision defined by the normal distribution with a standard deviation less than 2% of the dose volume in the operation range. A sample of $n = 20$ observations was taken to check if the precision was as claimed. The sample standard deviation was calculated to $s = 0.03$.

Hence the claim is that $\sigma \leq 0.02$, thus we want to answer the question: if $\sigma = 0.02$ (i.e. the upper limit of the claim), what is then the probability of getting the sampling deviation $s \geq 0.03$?

```
# Chi-square milk dosing precision
# The sample size
n = 20
# The claimed deviation
sigma = 0.02
# The observed sample standard deviation
s = 0.03
# Calculate the chi-square statistic
chiSq = (n-1)*s**2 / sigma**2
# Use the cdf to calculate the probability of getting the observed
# sample standard deviation or higher
1 - stats.chi2.cdf(chiSq, df=n-1)


np.float64(0.0014022691601097703)
```

It seems very unlikely that the standard deviation is below 0.02 since the probability of obtaining the observed sample standard deviation under this condition is very small. The probability we just found will be termed a p-value in later chapters - the p-value a very fundamental in testing of hypothesis.

The probability calculated in the above example will be called the p-value in later chapters and it is a very fundamental concept in statistics.

|||| Theorem 2.83 Mean and variance

Let $X \sim \chi^2(\nu)$ then the mean and variance of X is

$$\mathrm{E}(X) = \nu; \quad \mathrm{V}(X) = 2\nu. \tag{2-97}$$

We will omit the proof of this theorem, but it is easily checked by a symbolic calculation software (like e.g. Maple).

Example 2.84

We want to calculate the expected value of the sample variance (S^2) based on n observations with $X_i \sim N(\mu, \sigma^2)$. We have already seen that $\frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$ and we can therefore write

$$\begin{aligned} \mathrm{E}(S^2) &= \frac{\sigma^2}{n-1}\frac{n-1}{\sigma^2}\mathrm{E}(S^2) \\ &= \frac{\sigma^2}{n-1}\mathrm{E}\left(\frac{n-1}{\sigma^2}S^2\right) \\ &= \frac{\sigma^2}{n-1}(n-1) = \sigma^2, \end{aligned}$$

and we say that S^2 is a central estimator for σ^2 (the term *estimator* is introduced in Section 3.1.3). We can also find the variance of the estimator

$$\begin{aligned} \mathrm{V}(S^2) &= \left(\frac{\sigma^2}{n-1}\right)^2 \mathrm{V}\left(\frac{n-1}{\sigma^2}S^2\right) \\ &= \frac{\sigma^4}{(n-1)^2}2(n-1) = 2\frac{\sigma^4}{n-1}. \end{aligned}$$

Example 2.85 Pooled variance

Suppose now that we have two different samples (not yet realized) $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ with $X_i \sim N(\mu_1, \sigma^2)$ and $Y_i \sim N(\mu_2, \sigma^2)$ (both *i.i.d.*). Let S_1^2 be the sample variance based on the X's and S_2^2 be the sample variance based on the Y's. Now both S_1^2 and S_2^2 will be central estimators for σ^2, and so will any weighted average of the type

$$S^2 = aS_1^2 + (1-a)S_2^2; \quad a \in [0,1].$$

Now we would like to choose a such that the variance of S^2 is as small as possible, and hence we calculate the variance of S^2

$$\begin{aligned} \mathrm{V}(S^2) &= a^2 2\frac{\sigma^4}{n_1-1} + (1-a)^2 2\frac{\sigma^4}{n_2-1} \\ &= 2\sigma^4\left(a^2\frac{1}{n_1-1} + (1-a)^2\frac{1}{n_2-1}\right). \end{aligned}$$

In order to find the minimum we differentiate with respect to a

$$\begin{aligned}\frac{\partial \mathrm{V}(S^2)}{\partial a} &= 2\sigma^4\left(2a\frac{1}{n_1-1}-2(1-a)\frac{1}{n_2-1}\right)\\ &= 4\sigma^4\left(a\left(\frac{1}{n_1-1}+\frac{1}{n_2-1}\right)-\frac{1}{n_2-1}\right)\\ &= 4\sigma^4\left(a\frac{n_1+n_2-2}{(n_1-1)(n_2-1)}-\frac{1}{n_2-1}\right),\end{aligned}$$

which is zero for

$$a=\frac{n_1-1}{n_1+n_2-2}.$$

In later chapters we will refer to this choice of a as the pooled variance (S_p^2), inserting in (2-98) gives

$$S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}.$$

Note that S_p^2 is a weighted (proportional to the number of observations) average of the sample variances. It can also be shown (you are invited to do this) that $\frac{n_1+n_2-2}{\sigma^2}S_p^2 \sim \chi^2(n_1+n_2-2)$. Further, note that the assumption of equal variance in the two samples is crucial in the calculations above.

2.10.2 The t-distribution

The *t-distribution* is the *sampling distribution* of the sample mean standardized with the sample variation. It is valid for all sample sizes, however for larger sample sizes ($n > 30$) the difference between the t-distribution and the normal distribution is very small. Hence for larger sample sizes the normal distribution is often applied.

|||| Definition 2.86

The t-distribution *pdf* is

$$f_T(t)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{2-98}$$

where ν is the degrees of freedom and $\Gamma()$ is the Gamma function.

The relation between normal random variables and χ^2-distributed random variables are given in the following theo:rem

Theorem 2.87

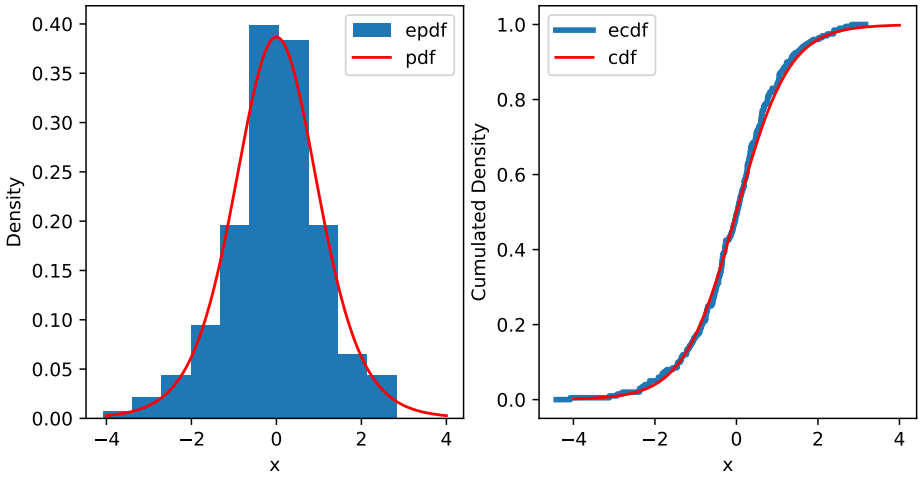
Let $Z \sim N(0,1)$ and $Y \sim \chi^2(\nu)$, then

$$X = \frac{Z}{\sqrt{Y/\nu}} \sim t(\nu). \tag{2-99}$$

We will not prove this theorem, but show by an example how this can be illustrated by simulation:

Example 2.88 Relation between normal and χ^2

```
# Set simulate parameters
nu = 8; k = 200
# Generate the simulated realizations
z = stats.norm.rvs(size=k)
y = stats.chi2.rvs(size=k, df=nu)
x = z/np.sqrt(y/nu)
# Plot
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(8,4))
ax1.hist(x, density=True, label='epdf')
ax1.plot(range(-4,5), stats.t.pdf(range(-4,5), df=nu), color="red", label='pdf')
stats.ecdf(x).cdf.plot(ax2, label='ecdf', linewidth=3)
ax2.plot(range(-4,5), stats.t.cdf(range(-4,5), df=nu), color="red", label='cdf')
```



In the left plot the empirical *pdf* is compared to the theoretical *pdf* and in the right plot the empirical *cdf* is compared to the theoretical *cdf*.

The t-distribution arises when a sample is taken of a normal distributed random variable, then the sample mean standardized with the sample variance follows the t-distribution.

Theorem 2.89

Given a sample of normal distributed random variables $X_1, \ldots, X_n$, then the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1), \tag{2-100}$$

follows the t-distribution, where $\overline{X}$ is the sample mean, μ is the mean of X, n is the sample size and S is the sample standard deviation.

Proof

Note that $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ which inserted in Equation (2.87) gives

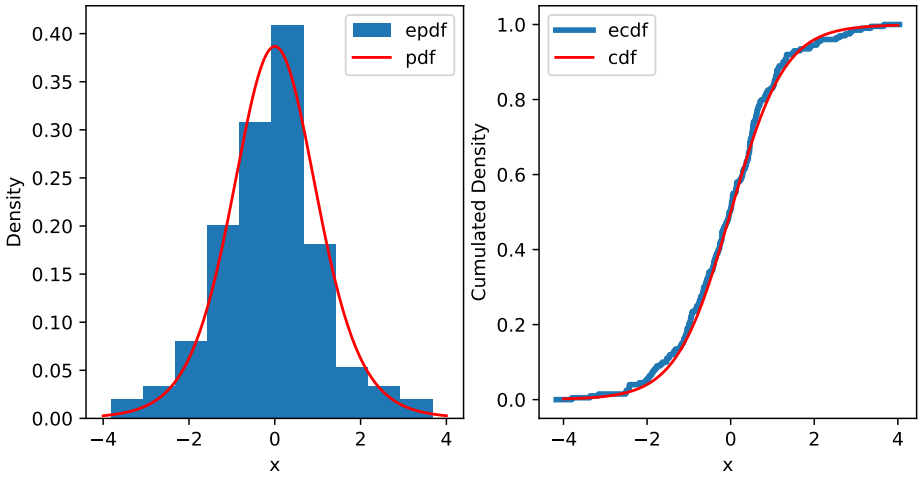
$$\begin{aligned} T &= \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}} = \frac{\frac{\overline{X}-\mu}{1/\sqrt{n}}}{\sqrt{S^2}} \\ &= \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1). \end{aligned} \tag{2-101}$$

■

We could also verify this by simulation:

Example 2.90 Simulation of t-distribution

```
# Simulate
n = 8; k = 200; mu = 1.0; sigma = 2.0
# Repeat k times the simulation of a normal dist. sample:
# return the values in a (n x k) matrix
x = [stats.norm.rvs(loc=mu, scale=sigma, size=n) for _ in range(k)]
x = np.array(x)
xbar = np.array([np.mean(i) for i in x])
s    = np.array([np.std(i)  for i in x])
tobs = (xbar - mu)/(s/np.sqrt(n))
# Plot
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(8,4))
ax1.hist(tobs, density=True, label='epdf')
ax1.plot(np.arange(-4,4.01,0.01), stats.t.pdf(np.arange(-4,4.01,0.01), df=nu), color="r
stats.ecdf(tobs).cdf.plot(ax2, label='ecdf', linewidth=3)
```



In the left plot the empirical *pdf* is compared to the theoretical *pdf* and in the right plot the empirical *cdf* is compared to the theoretical *cdf*.

Note that $\overline{X}$ and S are random variables, since they are the sample mean and standard deviation of a sample consisting of realizations of X, but the sample is not taken yet.

Very often samples with only few observations are available. In this case by assuming normality of the population (i.e. the X_i's are normal distributed) and for a some mean μ, the t-distribution can be used to calculate the probability of obtaining the sample mean in a given range.

Example 2.91 Electric car driving distance

An electric car manufacture claims that their cars can drive on average 400 km on a full charge at a specified speed. From experience it is known that this full charge distance, denote it by X, is normal distributed. A test of $n = 10$ cars was carried out, which resulted in a sample mean of $\bar{x} = 382$ km and a sample deviation of $s = 14$.

Now we can use the t-distribution to calculate the probability of obtaining this value of the sample mean or lower, if their claim about the mean is actually true:

```
# Calculate the probability of getting the sample mean under the
# conditions that the claim is actually the real mean

# A test of 10 cars was carried out
n = 10
# The claim is that the real mean is 400 km
muX = 400
# From the sample the sample mean was calculated to
xMean = 393
# And the sample deviation was
xSD = 14
# Use the cdf to calculate the probability of obtaining this
# sample mean or a lower value
stats.t.cdf((xMean-muX)/(xSD/np.sqrt(n)), df=n-1, loc=0, scale=1)


np.float64(0.0741523536832797)
```

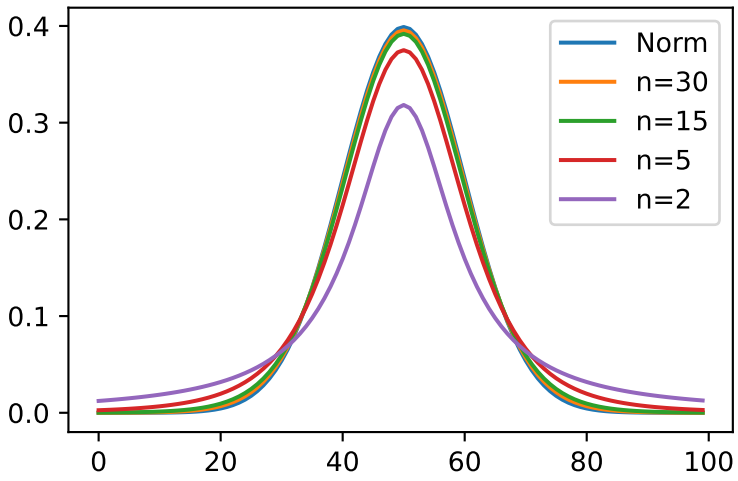
If we had the same sample mean and sample deviation, how do you think changing the number of observations will affect the calculated probability? Try it out.

The t-distribution converges to the normal distribution as the simple size increases. For small sample sizes it has a higher spread than the normal distribution. For larger sample sizes with $n > 30$ observations the difference between the normal and the t-distribution is very small.

Example 2.92 *t*-distribution

Generate plots to see how the *t*-distribution is shaped compared to the normal distribution.

```
# Plot the t-distribution for different sample sizes

# First plot the standard normal distribution
x = np.arange(-5,5,0.1)
plt.plot(stats.norm.pdf(x, loc=0, scale=1), label='Norm')
# Add the t-distribution for 30 observations
plt.plot(stats.t.pdf(x, df=30-1, loc=0, scale=1), label='n=30')
# Add the t-distribution for 15, 5 and 2 observations
plt.plot(stats.t.pdf(x, df=15-1, loc=0, scale=1), label='n=15')
plt.plot(stats.t.pdf(x, df=5-1, loc=0, scale=1), label='n=5')
plt.plot(stats.t.pdf(x, df=2-1, loc=0, scale=1), label='n=2')
# Add a legend
plt.legend()
plt.show()
```



How does the number of observations affect the shape of the *t*-distribution *pdf* compared to the normal *pdf*?

Theorem 2.93 Mean and variance

Let $X \sim t(\nu)$ then the mean and variance of X is

$$\mathrm{E}(X) = 0; \quad \nu > 1, \tag{2-102}$$

$$\mathrm{V}(X) = \frac{\nu}{\nu - 2}; \quad \nu > 2. \tag{2-103}$$

We will omit the proof of this theorem, but it is easily checked with a symbolic calculation software (like e.g. Maple).

Remark 2.94

For $\nu \leq 1$ the expectation (and hence the variance) is not defined (the integral is not absolutely convergent), and for $\nu \in (1, 2]$ $(1 < \nu \leq 2)$ the variance is equal ∞. Note that this does not violate the general definition of probability density functions.

2.10.3 The F-distribution

The F-distribution is defined by:

Definition 2.95

The F-distribution *pdf* is

$$f_F(x) = \frac{1}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1} \left(1 + \frac{\nu_1}{\nu_2} x\right)^{-\frac{\nu_1+\nu_2}{2}}, \tag{2-104}$$

where ν_1 an ν_2 are the degrees of freedom and $B(\cdot, \cdot)$ is the Beta function.

The F-distribution appears as the ratio between two independent χ^2-distributed random variables:

Theorem 2.96

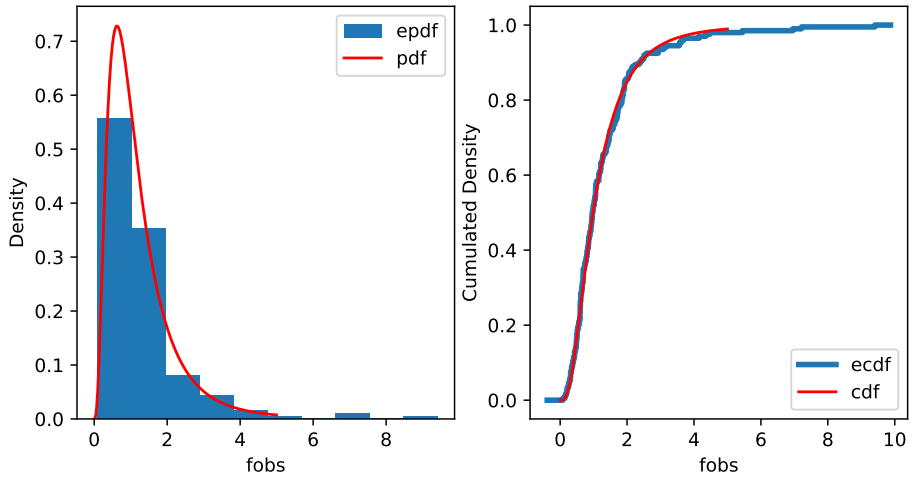
Let $U \sim \chi^2(\nu_1)$ and $V \sim \chi^2(\nu_2)$, be independent then

$$F = \frac{U/\nu_1}{V/\nu_2} \sim F(\nu_1, \nu_2). \tag{2-105}$$

Again we will omit the proof of the theorem and rather show how it can be visualized by simulation:

Example 2.97 *F*-distribution

```
# Simulate
nu1 = 8; nu2 = 10; k = 200
u = stats.chi2.rvs(size=k, df=nu1)
v = stats.chi2.rvs(size=k, df=nu2)
fobs = (u/nu1) / (v/nu2)
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(8,4))
ax1.hist(fobs, density=True, label='epdf')
ax1.plot(np.arange(0,5.01,0.01), stats.f.pdf(np.arange(0,5.01,0.01), dfn=nu1, dfd=nu2),
stats.ecdf(fobs).cdf.plot(ax2, label='ecdf', linewidth=3)
ax2.plot(np.arange(0,5.01,0.01), stats.f.cdf(np.arange(0,5.01,0.01), dfn=nu1, dfd=nu2),
```



In the left plot the empirical *pdf* is compared to the theoretical *pdf* and in the right plot the empirical *cdf* is compared to the theoretical *cdf*.

Theorem 2.98

Let $X_1, \ldots, X_{n_1}$ be independent and sampled from a normal distribution with mean μ_1 and variance σ_1^2, further let $Y_1, \ldots, Y_{n_2}$ be independent and sampled from a normal distribution with mean μ_2 and variance σ_2^2. Then the statistic

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1), \tag{2-106}$$

follows an F-distribution.

Proof

Note that $\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1)$ and $\frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$ and hence

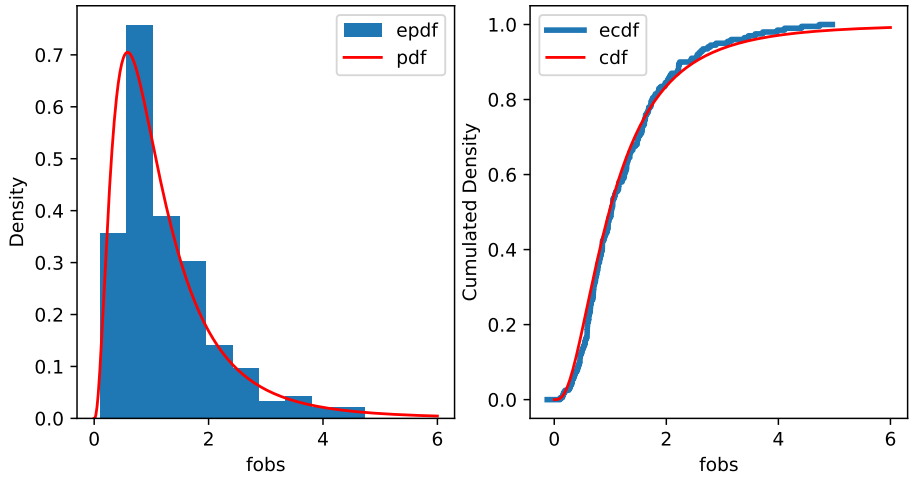
$$\frac{\frac{(n_1-1)S_1^2}{\sigma_1^2(n_1-1)}}{\frac{(n_2-1)S_2^2}{\sigma_2^2(n_2-1)}} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F(n_1 - 1, n_2 - 1). \tag{2-107}$$

■

We can also illustrate this sample version by simulation:

Example 2.99 Relation between normal and F-distribution

```
# Simulate
n1 = 8; n2 = 10; k = 200
mu1 = 2; mu2 = -1
sigma1 = 2; sigma2 = 4
s1 =  np.array([np.std(stats.norm.rvs(size=n1, loc=mu1, scale=sigma1)) for _ in range(k
s2 =  np.array([np.std(stats.norm.rvs(size=n2, loc=mu2, scale=sigma2)) for _ in range(k
fobs = (s1**2 / sigma1**2) / (s2**2 / sigma2**2)
# Plot
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(8,4))
ax1.hist(fobs, density=True, label='epdf')
ax1.plot(np.arange(0,6.01,0.01), stats.f.pdf(np.arange(0,6.01,0.01), dfn=nu1-1, dfd=nu2
stats.ecdf(fobs).cdf.plot(ax2, label='ecdf', linewidth=3)
ax2.plot(np.arange(0,6.01,0.01), stats.f.cdf(np.arange(0,6.01,0.01), dfn=nu1-1, dfd=nu2
```



In the left plot the empirical *pdf* is compared to the theoretical *pdf* and in the right plot the empirical *cdf* is compared to the theoretical *cdf*.

Remark 2.100

Of particular importance in statistics is the case when $\sigma_1 = \sigma_2$, in this case

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1). \tag{2-108}$$

Theorem 2.101 Mean and variance

Let $F \sim F(\nu_1, \nu_2)$ then the mean and variance of F is

$$\mathrm{E}(F) = \frac{\nu_2}{\nu_2 - 2}; \quad \nu_2 > 2, \tag{2-109}$$

$$\mathrm{V}(F) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}; \quad \nu_2 > 4. \tag{2-110}$$

Glossaries

Binomial distribution [Binomial fordeling] If an experiment has two possible outcomes (e.g. failure or success, no or yes, 0 or 1) and is repeated more than one time, then the number of successes is binomial distributed 18, 20, 21

χ^2-distribution [χ^2-fordeling (udtales: chi-i-anden fordeling)] 57, 59

Continuous random variable [Kontinuert stokastisk variabel] If an outcome of an experiment takes a continuous value, for example: a distance, a temperature, a weight, etc., then it is represented by a continuous random variable 3, 27, 29, 53

Correlation [Korrelation] The sample correlation coefficient are a summary statistic that can be calculated for two (related) sets of observations. It quantifies the (linear) strength of the relation between the two. See also: Covariance 48, 49

Covariance [Kovarians] The sample covariance coefficient are a summary statistic that can be calculated for two (related) sets of observations. It quantifies the (linear) strength of the relation between the two. See also: Correlation 48–50

Degrees of freedom [Frihedsgrader] The number of "observations" in the data that are free to vary when estimating statistical parameters often defined as $n - 1$ 57, 59, 62, 69

Discrete random variable [Diskret stokastisk variabel] A discrete random variable has discrete outcomes and follows a discrete distribution 4, 11, 14, 51, 52

Distribution [Fordeling] Defines how the data is distributed such as, normal distribution, cumulated distribution function, probability density function exponential distribution, log-normal distribution, Poisson distribution, uniform distribution, hypergeometric distribution, binomial distribution, t-distribution, F-distribution 3

Expectation [Forventningsværdi] A function for calculating the mean. The value we expect for a random variable (or function of random variables), hence of the population 12, 69

Exponential distribution [Eksponential fordelingen] The usual application of the exponential distribution is for describing the length (usually time) between events which, when counted, follows a Poisson distribution 39, 43

***F*-distribution** [*F*-fordelingen] The *F*-distribution appears as the ratio between two independent χ^2-distributed random variables 69

Histogram [Histogram] The default histogram uses the same width for all classes and depicts the raw frequencies/counts in each class. By dividing the raw counts by n times the class width the density histogram is found where the area of all bars sum to 1 7, 38

Hypergeometric distribution [Hypergeometrisk fordeling] 21, 22

Independence [Uafhængighed] 48, 51–53

(Statistical) Inference [Statistisk inferens (følgeslutninger baseret på data)] 56

Interval [Interval] Data in a specified range 23–25, 40, 42

Log-normal distribution [Lognormal fordeling] 38

Normal distribution [Normal fordeling] 27, 31, 34, 35, 38, 54, 57, 60, 62, 67, 68, 71

***P*-value** [*p*-værdi (for faktisk udfald af en teststørrelse)] 60

Sample mean [Stikprøvegennemsnit] The average of a sample 10, 12, 13, 15, 48, 62, 64, 66, 67

t-distribution [*t*-fordeling] 62

Acronyms

ANOVA Analysis of Variance *Glossary:* Analysis of Variance

cdf cumulated distribution function 5, *Glossary:* cumulated distribution function

CI confidence interval *Glossary:* confidence interval

CLT Central Limit Theorem *Glossary:* Central Limit Theorem

IQR Inter Quartile Range *Glossary:* Inter Quartile Range

LSD Least Significant Difference *Glossary:* Least Significant Difference

pdf probability density function *Glossary:* probability density function