Chapter 2

Probability and simulation (solutions to exercises)

Contents

```python
import numpy as np
import pandas as pd
import scipy.stats as stats
import matplotlib.pyplot as plt
import statsmodels as sm
import statsmodels.stats.proportion as smprop
```

2.1 Discrete random variable

Exercise 2.1 Discrete random variable

a) Let X be a stochastic variable. When running the Python-command `stats.binom.pmf(4, 10, 0.6)` Python returns 0.1115. written as:

```python
print(stats.binom.pmf(4, 10, 0.6))
```

```
0.11147673600000009
```

What distribution is applied and what does 0.1115 represent?

b) Let X be the same stochastic variable as above. The following are results from Python:

```python
print(stats.binom.cdf(4,10,0.6))
```

```
0.16623861760000005
```

```python
print(stats.binom.cdf(5,10,0.6))
```

```
0.3668967424000001
```

Calculate the following probabilities: $P(X \leq 5)$, $P(X < 5)$, $P(X > 4)$ and $P(X = 5)$.

c) Let X be a stochastic variable. From Python we get:

```
print(stats.poisson.pmf(k=4,mu=3))

0.16803135574154085
```

What distribution is applied and what does 0.16803 represent?

d) Let X be the same stochastic variable as above. The following are results from Python:

```
print(stats.poisson.cdf(4,3))

0.8152632445237722

print(stats.poisson.cdf(5,3))

0.9160820579686966
```

Calculate the following probabilities: $P(X \leq 5)$, $P(X < 5)$, $P(X > 4)$ and $P(X = 5)$.

2.2 Course passing proportions

|||| Exercise 2.2 Course passing proportions

a) If a passing proportion for a course given repeatedly is assumed to be 0.80 on average, and there are 250 students who are taking the exam each time, what is the expected value, μ and standard deviation, σ, for the number of students who do not pass the exam for a randomly selected course?

2.3 Notes in a box

Exercise 2.3 Notes in a box

A box contains 6 notes:

On 1 of the notes there is the number 1
On 2 of the notes there is the number 2
On 2 of the notes there is the number 3
On 1 of the notes there is the number 4

Two notes are drawn at random from the box, and the following random variable is introduced: X, which describes the number of notes with the number 4 among the 2 drawn. The two notes are drawn without replacement.

a) The mean and variance for X, and $P(X = 0)$ are?

b) The 2 notes are now drawn with replacement. What is the probability that none of the 2 notes has the number 1 on it?

2.4 Consumer survey

Exercise 2.4 Consumer survey

In a consumer survey performed by a newspaper, 20 different groceries (products) were purchased in a grocery store. Discrepancies between the price appearing on the sales slip and the shelf price were found in 6 of these purchased products.

a) At the same time a customer buys 3 random (different) products within the group consisting of the 20 goods in the store. The probability that no discrepancies occurs for this customer is?

2.5 Hay delivery quality

|||| Exercise 2.5 Hay delivery quality

A horse owner receives 20 bales of hay in a sealed plastic packaging. To control the hay, 3 bales of hay are randomly selected, and each checked whether it contains harmful fungal spores.

It is believed that among the 20 bales of hay 2 bales are infected with fungal spores. A random variable X describes the number of infected bales of hay among the three selected.

a) The mean of X, (μ_X), the variance of X, (σ_X^2) and $P(X \geq 1)$ are?

b) Another supplier advertises that no more than 1% of his bales of hay are infected. The horse owner buys 10 bales of hay from this supplier, and decides to buy hay for the rest of the season from this supplier if the 10 bales are error-free.
What is the probability that the 10 purchased bales of hay are error-free, if 1% of the bales from a supplier are infected (p_1) and the probability that the 10 purchased bales of hay are error-free, if 10% of the bales from a supplier are infected (p_{10})?

2.6 Newspaper consumer survey

|||| Exercise 2.6 Newspaper consumer survey

In a consumer survey performed by a newspaper, 20 different groceries (products) were purchased in a grocery store. Discrepancies between the price appearing on the sales slip and the shelf price were found in 6 of these purchased products.

a) Let X denote the number of discrepancies when purchasing 3 random (different) products within the group of the 20 products in the store. What is the mean and variance of X?

2.7 A fully automated production

Exercise 2.7 A fully automated production

On a large fully automated production plant items are pushed to a side band at random time points, from which they are automatically fed to a control unit. The production plant is set up in such a way that the number of items sent to the control unit on average is 1.6 item pr. minute. Let the random variable X denote the number of items pushed to the side band in 1 minute. It is assumed that X follows a Poisson distribution.

a) What is the probability that there will arrive more than 5 items at the control unit in a given minute is?

b) What is the probability that no more than 8 items arrive to the control unit within a 5-minute period?

2.8 Call center staff

|||| Exercise 2.8 Call center staff

The staffing for answering calls in a company is based on that there will be 180 phone calls per hour randomly distributed. If there are 20 calls or more in a period of 5 minutes the capacity is exceeded, and there will be an unwanted waiting time, hence there is a capacity of 19 calls per 5 minutes.

a) What is the probability that the capacity is exceeded in a random period of 5 minutes?

b) If the probability should be at least 99% that all calls will be handled without waiting time for a randomly selected period of 5 minutes, how large should the capacity per 5 minutes then at least be?

2.9 Continuous random variable

|||| Exercise 2.9 Continuous random variable

a) The following Python commands and results are given:

```
print(stats.norm.cdf(2))

0.9772498680518208

print(stats.norm.cdf(2,1,1))

0.8413447460685429

print(stats.norm.cdf(2,1,2))

0.6914624612740131
```

Specify which distributions are used and explain the resulting probabilities (preferably by a sketch).

b) What is the result of the following command: `stats.norm.ppf(stats.norm.cdf(2))`?

c) The following Python commands and results are given:

```
print(stats.norm.ppf(0.975))

1.959963984540054

print(stats.norm.ppf(0.975,1,1))

2.959963984540054

print(stats.norm.ppf(0.975,1,2))

4.919927969080108
```

State what the numbers represent in the three cases (preferably by a sketch).

2.10 The normal pdf

|||| Exercise 2.10 The normal pdf

a) Which of the following statements regarding the probability density function of the normal distribution $N(1, 2^2)$ is false?

1. The total area under the curve is equal to 1.0
2. The mean is equal to 1^2
3. The variance is equal to 2
4. The curve is symmetric about the mean
5. The two tails of the curve extend indefinitely
6. Don't know

Let X be normally distributed with mean 24 and variance 16

b) Calculate the following probabilities:
   - $P(X \leq 20)$
   - $P(X > 29.5)$
   - $P(X = 23.8)$

2.11 Computer chip control

|||| Exercise 2.11 Computer chip control

A machine for checking computer chips uses on average 65 milliseconds per check with a standard deviation of 4 milliseconds. A newer machine, potentially to be bought, uses on average 54 milliseconds per check with a standard deviation of 3 milliseconds. It can be used that check times can be assumed normally distributed and independent.

a) What is the probability that the time savings per check using the new machine is less than 10 milliseconds is?

b) What is the mean (μ) and standard deviation (σ) for the total time use for checking 100 chips on the new machine is?

2.12 Concrete items

Exercise 2.12 Concrete items

A manufacturer of concrete items knows that the length (L) of his items are reasonably normally distributed with $\mu_L = 3000$ mm and $\sigma_L = 3$ mm. The requirement for these elements is that the length should be not more than 3007 mm and the length must be at least 2993 mm.

a) The expected error rate in the manufacturing will be?

b) The concrete items are supported by beams, where the distance between the beams is called L_{beam} and can be assumed normal distributed. The concrete items length is still called L. For the items to be supported correctly, the following requirements for these lengths must be fulfilled: $90 \text{ mm} < L - L_{\text{beam}} < 110$ mm. It is assumed that the mean of the distance between the beams is $\mu_{\text{beam}} = 2900$ mm. How large may the standard deviation σ_{beam} of the distance between the beams be if you want the requirement fulfilled in 99% of the cases?

2.13 Online statistic video views

Exercise 2.13 Online statistic video views

In 2013, there were 110,000 views of the DTU statistics videos that are available online. Assume first that the occurrence of views through 2014 follows a Poisson process with a 2013 average: $\lambda_{365days} = 110000$.

a) What is the probability that in a randomly chosen half an hour there is no occurrence of views?

b) There has just been a view, what is the probability that you have to wait more than fifteen minutes for the next view?

2.14 Body mass index distribution

|||| Exercise 2.14 Body mass index distribution

The so-called BMI (Body Mass Index) is a measure of the weight-height-relation, and is defined as the weight (W) in kg divided by the squared height (H) in meters:

$$BMI = \frac{W}{H^2}.$$

Assume that the population distribution of BMI is a log-normal distribution with $\alpha = 3.1$ and $\beta = 0.15$ (hence that $\log(BMI)$ is normal distributed with mean 3.1 and standard deviation 0.15).

a) A definition of "being obese" is a BMI-value of at least 30. How large a proportion of the population would then be obese?

2.15 Bivariate normal

|||| Exercise 2.15 Bivariate normal

a) In the bivariate normal distribution (see Example 2.73), show that if Σ is a diagonal matrix then (X_1, X_2) are also independent and follow univariate normal distributions.

b) Assume that Z_1 and Z_2 are independent standard normal random variables. Now let X and Y be defined by

$$X = a_{11}Z_1 + c_1,$$
$$Y = a_{12}Z_1 + a_{22}Z_2 + c_2.$$

Show that an appropriate choice of $a_{11}, a_{12}, a_{22}, c_1, c_2$ can give any bivariate normal distribution for the random vector (X, Y), i.e. find $a_{11}, a_{12}, a_{22}, c_1, c_2$ as a function of μ_X, μ_Y and the elements of Σ.

Note that $\Sigma_{ij} = \text{Cov}(X_i, X_j)$ (i.e. here $\Sigma_{12} = \Sigma_{21} = \text{Cov}(X, Y)$), and that any linear combination of random normal variables will result in a random normal variable.

c) Use the result to simulate 1000 realization of a bivariate normal random variable with $\boldsymbol{\mu} = (1, 2)$ and

$$\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

and make a scatter plot of the bivariate random variable.

2.16 Sample distributions

Exercise 2.16 Sample distributions

a) Verify by simulation that $\frac{n_1+n_2-2}{\sigma^2}S_p^2 \sim \chi^2(n_1+n_2-2)$ (See Example 2.85). You may use $n_1 = 5$, $n_2 = 8$, $\mu_1 = 2$, $\mu_2 = 4$, and $\sigma^2 = 2$.

b) Show that if $X \sim N(\mu_1, \sigma^2)$ and $Y \sim N(\mu_2, \sigma^2)$, then

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

Verify the result by simulation. You may use $n_1 = 5$, $n_2 = 8$, $\mu_1 = 2$, $\mu_2 = 4$, and $\sigma^2 = 2$.

2.17 Sample distributions 2

Exercise 2.17 Sample distributions 2

Let $X_1, ..., X_n$ and $Y_1, ..., Y_n$, with $X_i \sim N(\mu_1, \sigma^2)$ and $Y_i \sim N(\mu_2, \sigma^2)$ be independent random variables. Hence, two samples before they are taken. S_1^2 and S_2^2 are the sample variances based on the X's and the Y's respectively. Now define a new random variable

$$Q = \frac{S_1^2}{S_2^2} \tag{2-1}$$

a) For n equal $2, 4, 8, 16$ and 32 find:

1. $P(Q < 1)$
2. $P(Q > 2)$
3. $P\left(Q < \frac{1}{2}\right)$
4. $P\left(\frac{1}{2} < Q < 2\right)$

b) For at least one value of n illustrate the results above by direct simulation from independent normal distributions. You may use any values of μ_1, μ_2 and σ^2.