Simulering og Statistikk - Module 1, January 2014

Dr S.B. van Albada and Prof Dr H.G. Schaathun

I. Stochastic variables

In this section we introduce the concept of a *stochastic variable* by studying coin flips.

1. What possible *outcomes* do we have for a single coin flip experiment?

2. What are the *probabilities* for each of these outcomes?

3. Take a coin and flip it 15 times while keeping track of the results. Draw a *histogram* of the results.

In the following, assign the outcome 1 to heads, and the outcome 0 to tails.

The outcome of the coin flip experiment is not fixed; instead, there are two possible outcomes, both with equal probabilities P(heads) = 50% and P(tails) = 50%, or, as we use to write in probability theory, P(heads) = 0.5 and P(tails) = 0.5. We call the outcome of the coin flip experiment a *stochastic variable* (or *random variable*).

We will now *simulate* a single coin flip.

4. Input the following code into MATLAB:

```
x=rand<0.5
```

This code returns either 0 or 1, both with equal probabilities P(tails) = P(heads) = 0.5. More precisely, it generates a *(pseudo) random number* in the interval ⟨0, 1⟩. If the resulting number is smaller than 0.5, x is assigned the value 0, otherwise x gets the value 1.

Explanation: The function *rand* returns a (uniformly distributed) (pseudo) random number in the interval ⟨0, 1⟩. The statement *rand<0.5* is either true (1) or false (0). This result is assigned to the variable x.

Study: §3.1 Deterministic and stochastic models
§3.2 Stochastic experiments
Exercises: E3.1, E3.2

5. Repeat the above simulation a number of times, simulating a series of coin flips in order to verify that the outcome takes on random values from the *sample space* (norsk: *utfallsrommet*) {0, 1}.

We can also call the set of all possible outcomes the *population*.

The following code generates a series of 15 independent coin flips, and generates a histogram of the results.

```
clear
n=15
x=rand(1,n)<0.5
hist(x,0:1)
```

Explanation: The function *rand(m, n)* returns an m-by-n matrix of (uniformly distributed) pseudo random numbers, all in the interval ⟨0, 1⟩. The statement rand(1, n)<0.5 returns true (1) or false (0) for all matrix indices.

6. Input the code in MATLAB, and run it a couple of times.

A *stochastic variable* is a variable that can take on multiple values, each with its own probability. These probabilities have to add up to one: We are 100% sure that the experiment has *some* outcome. In the case of a coin flip, we have: P(heads) + P(tails) = 0.5 + 0.5 = 1.

We indicate a stochastic variable with a capital letter, like X or Y. The same letter in lower case indicates the possible *outcomes* (norsk: utfall) that the stochastic variable can take on. In the case of a coin flip, we can write $P(X=x) = 0.5$. This means that the probability that our stochastic variable X takes on the value x, equals 50%. This holds for both possible outcomes: $P(X=0) = 0.5$ (i.e., 50% probability that the outcome equals "tails"), and $P(X=1) = 0.5$ (i.e., 50% probability that the outcome equals "heads").

We can define an *event* (norsk: hendelse) as a set of outcomes. For example, the outcomes of a single roll of dice are {1, 2, ..., 6}. We might be interested in the case that we roll either 5 or 6, in other words, the event {5, 6}.

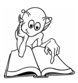
7. What is the probability for the event {5, 6}?

In case all outcomes have equal probabilities, the probability for an event can be calculated as:

$$P(\text{event}) = \frac{\text{Number of favourable outcomes}}{\text{Number of possible outcomes}} \qquad (1)$$

where an outcome is called "favourable" if it lies in the desired event.

8. What is the probability for throwing 7 eyes in two rolls of dice?



Study: §3.3 Probability
§7.1 until §7.1.1 Discrete uniform distribution
Chapter 2 Statistics

II. Discrete probability distributions

9. If we flip a coin twice, what is the probability to get the summed outcomes 0, 1 and 2, respectively?

10. Draw a bar chart with the possible outcomes on the horizontal axis and the corresponding probabilities on the vertical axis. The resulting plot shows the *probability distribution* for the coin flip experiment with two coins.

11. Take a coin and flip it twice. Repeat this experiment 15 times, and draw a histogram of the results.

12. Input the following code into MATLAB:

```
clear
n=2
trials=15
for i=1:trials
t=rand(1,n)<0.5;
x(i)=sum(t);
end
hist(x,0:n)
```

Explanation: The array *t* has length *n*. For each trial, all its elements are assigned zeroes and ones with equal probabilities, as in a virtual coin flip experiment with *n* coins. The summed outcome is then written to the i^{th} element of array *x*. Finally, a histogram of the outcomes is generated, showing the counts for the possible outcomes 0, 1, ... *n*.

13. If we flip a coin three times, what is the probability to get the outcomes 0, 1, 2 and 3, respectively?

14. Draw a bar chart with the possible outcomes on the horizontal axis and the corresponding probabilities on the vertical axis. The resulting plot shows the *probability distribution* for the coin flip experiment with three coins.

15. Take a coin and flip it three times. Repeat this experiment 15 times, and make a histogram of the results.

16. Modify the above MATLAB code and simulate a series of 15 experiments, each with 3 coin flips.

17. If we flip a coin a number of times *n*, what possible outcomes do we have if we add together the outcomes of these *n* coin flips?

18. Input the following code into MATLAB:

```
clear
n = 2; % Number of coin flips for each trial
trials = 30; % Number of trials
for i = 1:trials
t = rand(1, n) < 0.5; % For trial #i, t is an array of n coin flip
outcomes.
x(i) = sum(t); % For trial #i, x(i) is the total number of heads.
end

probabilities(:,1) = hist(x, (0:n), 1) / trials;
% The experimental probabilities for the total numbers of heads 0,
% 1, ..., n are assigned to the first column of the matrix named
% "probabilities".

for i = 1:(n+1)
probabilities(i,2) = nchoosek(n, i-1) / 2^n;
% The theoretical probabilities for the total numbers of heads 0,
```

```
% 1, ..., n are assigned to the second column of the matrix
% named "probabilities".
end

bar((0:n), probabilities, 'grouped')
% The experimental and theoretical probabilities for the total
% numbers of heads are plotted together in a histogram.
```

19. Run the script in MATLAB.
    a. In the MATLAB workspace containing all used variables, double-click on the variable *t* to show its value. Explain what you find. If necessary, rerun the code.
    b. Next, inspect the variable *x*. Explain your findings.
    c. Inspect the value of the variable *probabilities*. Again, explain your findings.
20. Adapt the above code and sample the probability distribution for the sum of 3, 5 and 10 coin flips. What do you observe?

The fascinating effect that the sum of multiple random variables very quickly approaches a *normal distribution* (or *bell curve* or *Gaussian distribution*) is called the *central limit theorem* and will be covered later.

21. What is the probability for throwing 1, 2, ...,6 eyes with one dice? Make a table and plot the results.
22. What is the probability to throw exactly or less than 0, 1, 2, ...,6 eyes with a single dice? Make a table and plot the probability distribution.

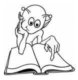
The (cumulative) distribution function (norsk: (kumulativ) fordelingsfunksjon) $F_X(x)$ is defined as the probability to get a result less than or equal to x:

$$F_X(x) = P(X \leq x) \qquad (2)$$

23. Plot the cumulative distribution for a single dice roll.
24. What is the probability for throwing 2, 3, 4, ... ,12 eyes in two rolls of dice? Make a table and plot the probability distribution.
25. Plot the cumulative distribution function for a roll with two two dice.
26. Use the following MATLAB code to retrieve the empirical cumulative distribution function for a sample consisting of 15 rolls with two dice:

```
y=[1,3,6,10,15,21,26,30,33,35,36]/36%The c.d.f. (verify the values!)
stairs(2:12, y, 'r')                % Plot the c.d.f. in red.
hold on                             % Superimpose the next plot
n=15
x=sum(ceil(6*rand(2,n)))            % Roll a pair of dice n times.
%                                     x is an array of length n.
stairs([0 sort(x)],0:1/length(x):1) % Plot the empirical c.d.f. of x.
hold off                            % End of superposition
```

27. Explain the code.
28. Increase the sample size. What do you observe?



Study: §4.0 Introduction to stochastic variables
§4.1Discrete stochastic variables
Exercises: E4.1, E4.3, E4.4, E4.5

III. Mean

Take a coin and flip it 5 times, while keeping track of the numbers of heads (1) and tails (0). Such a set of observations is called a *sample*.

29. Plot the results in a histogram.

30. Calculate your *sample mean* (i.e., the average outcome).

31. Verify that the sample mean $\bar{x}$ can be written as

$$\bar{x} = \frac{\sum_i x_i}{n} \qquad (3)$$

What does the index *i* mean?

32. The sample mean should be a number between the outcomes zero and one. What possible outcomes for the sample mean are there after 5 single coin tosses?

The *expected value* E[*X*] of a discrete stochastic variable is a weighted average over all possible outcomes:

$$\mu_X = E[X] = \sum_i x_i P(x_i) \qquad (4)$$

In the special case that all possible outcomes are equally likely, the expected value can be calculated as the sum over all possible outcomes, divided by the number of possible outcomes (see exercise 23).

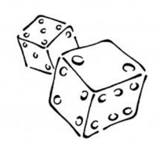
33. What is the expected value for the outcome of a single coin flip?

34. Prove that the expected value can be written as

$$\mu_X = E[X] = \frac{\sum_i x_i}{n} \qquad (5)$$

if all possible outcomes in an experiment are equally likely.

What does the index *i* mean? Compare to exercise 20!



35. Take a single dice and roll it 15 times. Write down the results.

36. Plot the results in a histogram.

37. Calculate the sample mean for your experiment.

38. What possible outcomes for the sample mean are there after two rolls of dice?

39. What is the expected value for a single roll of dice?

The following MATLAB code simulates a series of 15 dice rolls, and generates a histogram of the results.

```
n=15
x=ceil(6*rand(1,n))
hist(x,1:6)
```

Explanation: The function *rand(m, n)* returns an m-by-n matrix of uniformly distributed pseudo random numbers, all in the interval ⟨0, 1⟩. Multiplication by 6 gives random numbers in the interval ⟨0, 6⟩. Finally, the function *ceil* returns the closest larger integer.

40. Input the code into MATLAB and run it a number of times.

If we sample a population very often, and take the average over the sample means (Eq. 3), the average value will approach the population mean (Eq. 4 / 5).

Example: We repeatedly throw three dice.

- The outcomes could be: {1, 1, 4}, {2, 3, 5}, {6, 1, 4}, ...
- The sample means would then be (verify!) 2, 10/3, 11/3, ...
- The average of these sample means would then be (verify!) 2, 8/3, 3, ...
- This series will in the end converge to the expected value (verify!) 3.5.

We call the sample mean an *estimator* for the expected value.

Exercises: E5.1

IV. Standard deviation and variance

41. In this assignment we will study the *dispersion* (norsk: spredning) of statistical data. Input the following code into MATLAB and run it:

```
clear
n=10
x=ceil(2*rand(1,n))-1          % simulate n coin tosses
y=ceil(6*rand(1,n))            % simulate n dice rolls
t=1:n
plot(t,x,t,y,'LineStyle','none', 'Marker','diamond')
```

a. For which of the two experiments is the dispersion largest?
b. Collect the results of the 10 coin tosses and the 10 rolls of dice in a table.
c. Compute the sample mean for both experiments.

In order to quantify the dispersion, we need a measure for the variation around the mean. We could try to take the average difference with the sample mean:

$$\frac{\sum_i (x_i - \bar{x})}{n} \qquad (6)$$

d. Compute the sample average of the difference with the sample mean for the ten coin tosses.

e. Prove that the answer to the previous exercise should be zero. Hint: Split the summation in two parts.

As the outcome is zero independent of the spread in the data, expression (6) is *not* a good measure for the dispersion.

f. Propose a better measure for the dispersion of a sample.

The most obvious way to solve the above problem is to average the *distance* from the mean:

$$\frac{\sum_i |x_i - \bar{x}|}{n} \qquad (7)$$

g. Compute expression (7) for the ten coin tosses as well as for the ten rolls of dice. Explain your findings. Does expression (7) give a good measure for the dispersion?

The main problem with expression (7) is that the absolute value is a bit unpleasant to work with as the function is not continuously differentiable. Therefore, the commonly-used measure for dispersion is slightly different:

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}} \qquad (8)$$

s is called the *standard deviation of the sample*.

The "average" squared distance from the sample mean is called the *sample variance*:

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1} \qquad (9)$$

Note: The reason why we divide by (n-1) instead of n is subtle and will be explained later. It is therefore not entirely correct to use the word "average".

h. Compute the variance and standard deviation for the ten coin tosses as well as for the ten rolls of dice. Does the standard deviation give a good measure for the dispersion? Compare your results to your answers to the previous exercise. Explain the difference.

i. The following MATLAB code can be used to compute the variance of the coin flip sample:

```
x_diff_squared = (x-mean(x)).^2
x_variance = sum(x_diff_squared)/(n-1)
```

Explanation: The dot operator is used to apply an operation (here taking the square) to every element of a matrix.

Use and modify the code in order to verify your results in exercise c, d, e, g, and h.

We have learnt how to *measure* the spread of data in a sample. But is it also possible to *predict* the spread of a given experiment? The answer is *yes*!

In the case of a coin toss, we know that we get the outcome 0 with a probability P(0)=0.5 and the outcome 1 with a probability P(1)=0.5, and the expected value of the outcome is μ = 0.5.

j. What is the *expected value* for the squared distance from μ?
k. Explain that the formula

$$\sigma^2 = \frac{\sum_i (x_i - \bar{x})^2}{n} \qquad (10)$$

gives the desired result. σ^2 is called the *population variance*. The square root of the population variance is called the *standard deviation of the population*.

$$\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}} \qquad (11)$$

What does the index *i* stand for? Compare with Equation 8!

l. Compute the population variance and standard deviation for a roll of dice.
m. Verify your results with MATLAB.

If we sample a population very often, and take the average over the sample variances (Eq. 9), the averaged value will approach the population variance (Eq. 10).

Example: We repeatedly throw three dice.

- The outcomes could be: {1, 1, 4}, {2, 3, 5}, {6, 1, 4}, ...
- The sample variances would then be (verify!) 3, 7/3, 19/3, ...
- The average of the sample means would then be (verify!) 3, 8/3, 35/9, ...
- This series will in the end converge to the population mean (verify!) 35/12.

We call the sample variance en *estimator* for the population variance.

42. The MATLAB command `mean(x)` gives the mean of an array of values x. Use the command to find the mean of 1, 100, 10 000 and 1 000 000 simulated single coin tosses. Repeat the experiment 5 times and collect the results in a table. (You should end up with 20 mean values). What do you observe?

43. Compute the mean of 1, 100, 10 000 and 1 000 000 simulated single dice rolls. Repeat the experiment 5 times and collect the sample means in a table. (In all, you should have collected 20 sample means). What do you observe?

As you probably have found in assignment 38 and 39, the sample mean tends to approach the expected value as sample size increases. the *dispersion* (norsk: spredning) becomes smaller for larger sample sizes.

Exercises: E5.3, E5.4

V. Independent events

Repeat: §1.5 Set theory
Exercises: E1.1, E1.2

44. If we throw two dice, what is the probability P(first 3, second 2) that the first one shows 3 eyes and the second one 2 eyes? Express your answer in terms of the probabilities P(first 3) (i.e., the probability for 3 eyes on the first roll) and P(second 2) (i.e., the probability for 2 eyes on the second roll).

In the previous assignment, the events P(first 3) and P(second 2) are *independent*: The first dice has no influence whatsoever on the second dice, and vice versa. In the case of independent events A and B, the probability that *both* events will take place is simply the *product* of the probabilities that events A and B take place:

$$P(A \cap B) = P(A)P(B) \tag{12}$$

45. If we throw three dice, what is the probability P(first 3, second 2, third 2) that the first one shows 3 eyes, the second one 2 eyes and the third one 2 eyes? Express your answer in terms of the probabilities P(first 3), P(second 2) and P(third 2).

46. If we throw three dice, what is the probability P(first 3, second plus third 4) that the first one shows 3 eyes, and the *sum* of the second and third dice equals 4? Express your answer in terms of the probabilities P(first 3) and P(second plus third 4).

Study: §5.3 Independent stochastic variables

VI. Dependent events

47. In this exercise, we will see how to deal with multiple events that are *not* independent.

a. If we throw two dice, what is the probability P(first 3, first plus second 4) that the first dice shows 3 eyes, and the *sum* of the first and second dice equals 4?
b. Is it possible to express your answer in terms of the probabilities P(first 3) and P(first plus second 4)?
c. If we *know* that the first dice shows 3 eyes, what is then the probability P(first plus second 4 | first 3) that the sum of the first and second dice equals 4 eyes?
d. Is it possible to express the probability P(first 3, first plus second 4) in terms of the probabilities P(first 3) and P(first plus second 4 | first 3)?
e. If we *know* that the sum of the eyes of the two dice equals 4, what is then the probability P(first 3 | first plus second 4) that the first dice shows one eye?
f. Is it possible to express the probability P(first 3, first plus second 4) in terms of the probabilities P(first plus second 4) and P(first 3 | first plus second 4)?

In the previous assignment, the events P(first 3) and P(first plus second 4) are *dependent*. If the first dice shows 3 eyes, the second dice *has to* show one eye in order to make the sum equal to 4. In the case that events A and B are dependent, the probability that *both* events will take place *cannot* be calculated as the product of the probabilities that events A and B take place. Instead, we have to deal with the *conditional probability* P(A | B) that A happens, in the case that we *know* that B happens:

$$P(A \cap B) = P(A)P(A \mid B) \qquad (13)$$

P(A | B) is pronounced as “P(A given B)”.

Study: §3.5 Dependent events
§3.6 General product laws
§3.7 Independent events
Exercises: E3.11, E3.14, E3.15, E7.1

VII. Mutually exclusive events

48. We throw three dice. What is the probability that:

a. the sum of the first two is less than 5 (event A)?
b. the sum of the last two is larger than 9 (event B)?
c. the sum of the first two is less than 5, while the sum of the last two is larger than 9 (event A ∩ B)?

Two events A and B are *mutually exclusive* if they impossibly can happen both. In other words: If A happens, B cannot happen (and vice versa, of course). If two events are mutually exclusive, the probability that both happen is zero:

$$P(A \cap B) = 0 \quad (14)$$

VIII. Continuous stochastic variables

49. What possible *outcomes* do we have for the following measurements, assuming that we can measure infinitely accurately:
   a. the height of an arbitrary person?
   b. the weight of an arbitrary person?
   c. the age of an arbitrary person?
   d. time modulo one second?

In the previous exercise we have seen examples of random variables. However, the difference with the random variables we have seen before is that these are *continuous random variables* that can take on values over a continuous range.

50. Input the following code into MATLAB:

```
x=rand
```

This code returns a uniformly distributed continuous random *(pseudo) random number* in the interval $\langle 0, 1 \rangle$.

51. Repeat the above simulation a number of times, simulating a series of observations in order to verify that the outcome takes on random values from the sample space $\langle 0, 1 \rangle$.

52. Take a stopwatch that is able to measure at least thousands' of seconds. Stop it at an arbitrary moment (wait for some time and do not watch your clock in the mean time) and write down the measured random time modulo one second. (This should give you a number between 0 and 1). Repeat this ten times and collect your data in a table.

53. What are the *probabilities* for each of the outcomes (still assuming infinite measurement precision)?

54. Draw the empirical cumulative distribution function based on your data.

55. Draw the theoretical cumulative distribution function in the same plot.

56. Input and run the following code in MATLAB:

```
n=10;
x = rand(1,n)
stairs([0 sort(x)], 0:1/n:1, 'r') % Plot the empirical c.d.f.
hold on
y = 0:.001:1
```

```
stairs(y, y, 'b') % Plot the c.d.f.
hold off
```

57. Explain the code and compare to exercise 51 and 52.

58. Increase the sample size. What do you observe?

59. What would the probability distribution look like? (Remember that the sum of all probabilities should be one)!

As we have seen, we can define the *cumulative distribution* function for a continuous stochastic variable in the same we as we have done for a discrete stochastic variable: The cumulative distribution function $F_X(x)$ is as always defined as the probability to get a result less than or equal to x:

$$F_X(x) = P(X \leq x) \qquad (15)$$

However, the *probability distribution* has become useless, as the probability for every single value equals zero, whereas the integrated (total) probability should equal one. We therefore define a new function, the *probability density* $f_X(x)$ as the *relative* likelihood for the random variable to take on a given value. The area under the density function between two values gives the probability that the random variable falls between these values. The integral (total area) of the probability density function over the entire space is equal to one.

60. Draw the probability density for a uniformly distributed continuous random variable representing a point in time between zero and one seconds.

In order to achieve the probability density function empirically, we cannot use the exact same method as we used for the probability distribution of a discrete random variable. The problem is that we never will get the same result twice, since the probability for each event is zero. We can solve this problem by gathering together data that lie close together, a method that is called *binning*. In this way, we can achieve a histogram of the data representing the relative probabilities for the different areas. Finally, we *normalize* such that the total area under the histogram equals one.

Study: §4.2 Continuous stochastic variables
Exercises: E4.7

61. What will the probability density look like if we instead of one, each time draw *two* uniformly distributed random numbers between zero and one (for example by the stopwatch method) and add these together?

62. Take a sample of 15 measurements and draw an empirical probability density function for the previous exercise. Use a sample size of 0.4.

63. Input and run the following code in MATLAB:

```
n=15;                          % Sample size
nBins=5;                       % Number of bins in the histogram
x = rand(1,n)                  % Uniform random number in <0,1>
y = rand(1,n)                  % And another
z=(x+y)                        % Sum of two uniform random variables
```

```
bins=1/nBins:2/nBins:2-1/nBins% The array of bin centres
% (indicated by centre of first bin,
% step size, and centre of last bin)
% We need to make a bar chart as a histogram is not nomalizable:
[nelements,centers] = hist(z,bins)
bar(centers,nelements*nBins/2/sum(z),'b') % Note the normalization of
% the number of elements in order to get the probability density!
hold on
y = 0:.001:2
plot(y, 1-abs(y-1), 'r') % Theoretical probability density for
% the sum of two uniform random variables
hold off
```

64. Increase the sample size and the number of bins in order to sample the probability density function.

65. Adapt the above code and sample the probability density for the sum of 3, 4 and 5 uniformly distributed continuous random variables. What do you observe?

As in exercise 20, we are witness of the *central limit theorem*: The sum of multiple random variables very quickly approaches a *normal distribution*. The central limit theorem will be covered later.

IX. Mean and variance for a continuous random variable

The sample mean and variance for a continuous random variable are calculated in the exact same way as we have done for a discrete random variable see formulas given at exercises 32 and 41.g.

However, in order to calculate the expected value, variance and standard deviation of the population of all possible outcomes we now need to *integrate* instead of sum over all possible outcomes:

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f(x) dx \qquad (16)$$

$$\sigma_X{}^2 = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx \qquad (17)$$

66. Take your sample of 10 uniformly distributed continuous random variables from exercise 52. Compute the sample mean, standard deviation and variance.

67. Calculate the expected value, variance and standard deviation for a uniformly distributed continuous random variable in the interval <0, 1>. Compare your results to the previous exercise.

68. Take your sample of 15 random numbers from exercise 62 generated as the sum of two uniform random variables. Compute the sample mean, standard deviation and variance.

69. Calculate the expected value, variance and standard deviation for the sum of two independent uniformly distributed continuous random variables in the interval <0, 1>. Compare your results to the previous exercise.

Study: §5.1 Expected value
§5.2 Dispersion, variance, standard deviation
Exercises: E5.7, E5.12