MATH2089
Numerical Methods and Statistics

Statistics – Chapter 2 – Descriptive Statistics

SOLUTIONS

Exercise 1

a) (i) No, the mean need not be an observed value. Consider the sample $\{0, 1\}$, its mean is 0.5.

(ii) No, that is the definition of the median.

(iii) No, as the mean need not even be an observed value. The most frequently occurring data value in the sample is called the sample **mode**.

(iv) Yes, when all the observations are equal, there is no dispersion in the sample.

(v) Yes, but this only happens when the distribution of the observed values is exactly symmetric.

b) (i) Adding 10 to all observations is like shifting the whole sample by a distance of 10. The new mean is also shifted by 10, however the dispersion in the sample is not affected by this shift so that the standard deviation is unchanged.

(ii) Multiplying all the observations by 2 is like shifting and stretching the sample. The new mean is affected by the shifting and is the initial mean multiplied by 2, while the dispersion is affected by the stretching, so that the new standard deviation is also the initial standard deviation multiplied by 2.

(iii) We have $\bar{x} = 446$ and $s_x = 5.8$ in °C. Denote y the temperature in °F. Similarly to above, we have

$$\bar{y} = \frac{9}{5} \times \bar{x} + 32 = 834.8$$

and

$$s_y = \frac{9}{5} s_x = 10.44,$$

both expressed in °F.

Exercise 2

Yes, the median is a meaningful location measure, as it is unaffected by extreme values. So the unobserved value beyond 100 hours could be anything, the median would remain unchanged. Specifically, no matter the exact value of 100^+, the median is

$$m = \frac{1}{2}(63 + 75) = 69 \text{ hours},$$

as the sample size is even and 63 and 75 are the central values of the sample.

Exercise 3

Differentiate the function $f(a) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - a)^2$ with respect to a :

$$\frac{d}{da} f(a) = \frac{-2}{n-1} \sum_{i=1}^{n} (x_i - a)$$

Set this to 0 :

$$\sum_{i=1}^{n} (x_i - a) = 0$$

it follows

$$\sum_{i=1}^{n} x_i = na$$

or

$$a = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

(you can check that the second derivative of $f(a)$ is always positive)
$\rightsquigarrow$ the sample mean is the value that minimises the sum of the squared deviations.
Note that $f(\bar{x})$ is the sample variance.

Exercise 4

a) First order the observations :

$$4.4 < 16.4 < 22.4 < 30.0 < 33.1 < 36.6 < 40.4 < 66.7 < 73.7 < 81.5 < 109.9$$

The minimal value is 4.4, the maximal value is 109.9. There are $n = 11$ observations (odd number of observations), so the median is the $(n+1)/2 = 6$th largest observation, that is,

$$m = 36.6 \text{ (MPa)}$$

This splits the sample in two equal parts :

$$4.4 < 16.4 < 22.4 < 30.0 < 33.1 < 36.6$$

and

$$36.6 < 40.4 < 66.7 < 73.7 < 81.5 < 109.9$$

(include the median in each half). The first quartile is the median of the first half and the third quartile is the median of the second half. As these subsamples have even numbers of observations, their medians are

$$q_1 = \frac{22.4 + 30}{2} = 26.2 \text{ (MPa)}$$

and

$$q_3 = \frac{66.7 + 73.7}{2} = 70.2 \text{ (MPa)}$$

$\rightsquigarrow$ the 5-number summary of the sample is thus

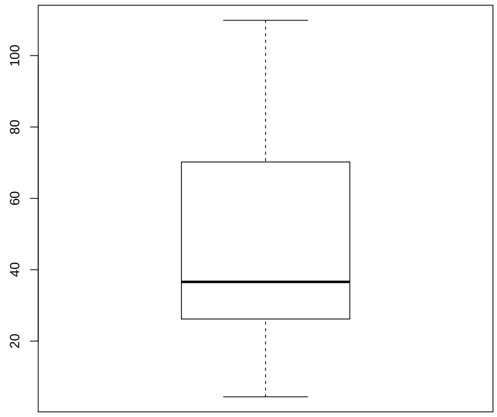
$$\{4.4, 26.2, 36.6, 70.2, 109.9\}$$

b) The interquartile range is

$$\text{iqr} = q_3 - q_1 = 70.2 - 26.2 = 44 \text{ (MPa)}$$

Thus, outliers would be the observations smaller than $q_1 - 1.5 \times \text{iqr} = 26.2 - 1.5 \times 44 = -39.8$ or larger than $q_3 + 1.5 \times \text{iqr} = 70.2 + 1.5 \times 44 = 136.2$.
$\rightsquigarrow$ there are no outliers.

c) Boxplot :



There is a slight positive skew to the data. There are no outliers.

d) The mean is

$$\bar{x} = \frac{1}{11}(4.4 + 16.4 + \ldots + 109.9) = 46.83 \text{ (MPa)},$$

the standard deviation is

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$$

$$= \sqrt{\frac{1}{10}((4.4^2 + 16.4^2 + \ldots + 109.9^2) - 11 \times 46.83^2)} = 31.996 \text{ (MPa)}$$

e) If the data set remained the same except for the largest observation being decreased, the only way that the iqr would change would be for the value to become small enough to make q_3 smaller. To do this the value of 109.9 would have to be replaced by one less than 73.7.

Exercise 5

a) Direct calculations give $\bar{x}_1 = 65.81$ (inches) and $s_1 = 2.106$ (inches). There are 37 observations in that sample, so the median is the 'middlemost' 19th smallest observation and can be found to be $m_1 = 66$ (inches).
b) Direct calculations give $\bar{x}_2 = 69.26$ (inches) and $s_2 = 2.028$ (inches). There are 50 observations in that sample, so the median is the middle between the 25th and the 26th smallest observations, both equal to 69, so the median is $m_2 = 69$ (inches).
c) A suitable back-to back stem-and-leaf plot might be (remark the particular division between stem and leafs!) :

0	61	
00	62	
00	63	
0000	64	
00000000	65	00
0000	66	0
00000000	67	0000
00000	68	0000000000
00	69	000000000000000
0	70	0000000
	71	00000
	72	00
	73	00
	74	0
	75	0

(left-hand side display : females, right-hand side display : males)

The right side appears to be shifted down several rows, which renders the fact that males are taller than females. The distribution for males appears fairly symmetric, unimodal and bell-shape, which is not really the case for the female distribution : it is slightly skewed and bimodal (two peaks).

Exercise 6

a) Direct calculations give $\bar{x} = 5.42$ (g/cm^3) and $s = 0.339$ (g/cm^3). There are 29 observations, so the median is the 15th smallest one (the 'middlemost' observation), which can be found to be

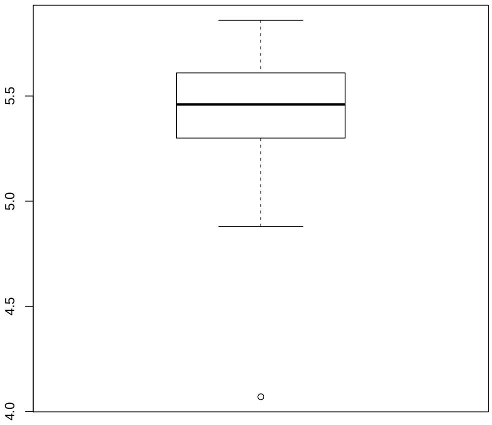
$$m = 5.46 \text{ (g/cm}^3\text{)}$$

b) This median value splits the sample in two subsamples of 15 observations each (include the median in each half). The 'middlemost' observation in the lower half (the 8th smallest) is the first quartile, found to be $q_1 = 5.30$ g/cm^3, while the 'middlemost' observation in the upper half (the 8th largest) is the third quartile, found to be $q_3 = 5.61$ g/cm^3. The interquartile range is thus

$$\text{iqr} = q_3 - q_1 = 0.31(\text{g/cm}^3),$$

and by the suggested rule, any observation outside $[q_1 - 1.5 \times \text{iqr}, q_3 + 1.5 \times \text{iqr}] = [4.835, 6.075]$ is an outlier. The value 4.07 is thus an outlier (too different from the other observations).

c) Boxplot :

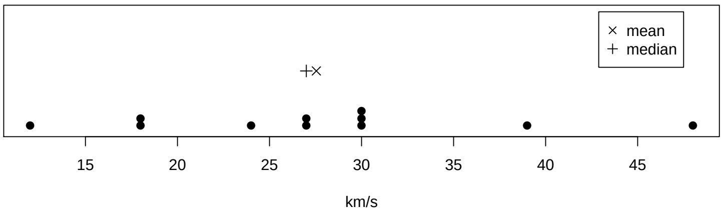


The boxplot itself is fairly symmetric, however the outlying value clearly appears. That observation was certainly corrupted by important measurement and other experimental errors.

d) The sample median should certainly be a more accurate estimate of the true earth density, as the sample mean would be affected by the outlying value.

Exercise 7

a) and b) Dotplot :



Direct calculations give $\bar{x} = 27.55$ km/s. With 11 observations, the median is the 6th smallest, that is $m = 27$ km/s.

c) Direct calculations give $s^2 = 100.47$ km^2/s^2 and $s = 10.02$ km/s.

d) The median value splits the sample in two subsamples of 6 observations each (include the median in each half). The first quartile is the middle between the 3rd and the 4th smallest observations, that is

$$q_1 = \frac{1}{2}(18 + 24) = 21 \text{ (km/s)}.$$

The third quartile is the middle between the 3rd and the 4th largest observations, that is

$$q_3 = \frac{1}{2}(30 + 30) = 30 \text{ (km/s)}.$$

e) The minimum value is $x_{(1)} = 12$ km/s, the maximum value is $x_{(11)} = 48$ km/s, the range is $x_{(11)} - x_{(1)} = 36$ km/s, and the interquartile range is $q_3 - q_1 = 9$ km/s.

Exercise 8

a) Consider the classes $[2, 3)$, $[3, 4)$, $[4, 5)$, $[5, 7)$ and $[7, 9)$. Then, the frequency distribution is

Class	[2,3)	[3,4)	[4,5)	[5,7)	[7,9)
Frequency	13	14	4	6	4

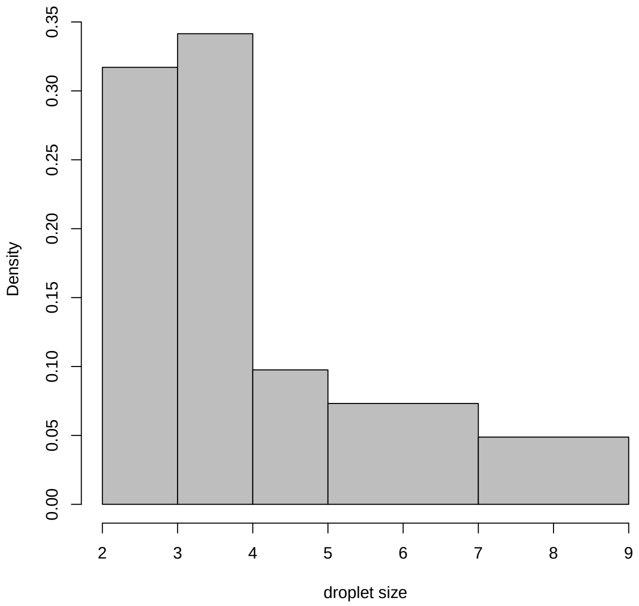
b) The relative frequency in each class is given by the (absolute) frequency divided by the total number of observations (that is to say, the relative frequency in a class is the proportion of observations in that class). With $n = 41$ observations, we find :

Class	[2,3)	[3,4)	[4,5)	[5,7)	[7,9)
Relative frequency	0.317	0.341	0.098	0.146	0.098

Note that these proportions must sum to 1 (as it is the case). Now, the density in each class is given by the relative frequency divided by the class width. Here we find :

Class	[2,3)	[3,4)	[4,5)	[5,7)	[7,9)
Density	0.317	0.341	0.098	0.073	0.049

The density histogram is :



c) Direct calculations give $\bar{x} = 3.97$ μm and $s^2 = 2.91$ μm^2.

Exercise 9

N/A

Exercise 10

a) construct the frequency histogram

```
>> histogram(Inflow)
>> xlabel('Inflow (cubic metres per second)')
>> ylabel('Frequency')
>> title('Annual maximum inflow')
```

b) transform data :

```
>> logInflow=log(Inflow);
```

construct histogram in a new window :

```
>> figure
>> histogram(logInflow)
>> xlabel('Inflow (cubic metres per second) [log scale]')
>> ylabel('Frequency')
```

c) calculate skewness :

```
>> skewness(Inflow)
  ans =
  2.9253
```

```
>> skewness(logInflow)
  ans =
  -0.0105
```

The histogram in b) is highly positively skewed (or right skewed) : the skewness is largely positive. After log-transformation, the histogram in c) is much more symmetric : the skewness is close to zero. The histogram in c) is bell-shaped, too.

d) for stem and leaf plot construction :

```
>> round(sort(logInflow)*10)/10
ans =
Columns 1 through 7
3.4000 3.8000 4.4000 4.8000 4.9000 5.0000 5.3000
Columns 8 through 14
5.4000 5.5000 5.8000 5.9000 6.0000 6.0000 6.1000
Columns 15 through 21
6.2000 6.3000 6.5000 6.6000 6.7000 6.8000 7.3000
Columns 22 through 25
7.3000 7.5000 8.1000 8.7000
```

Stemplot (Stem-and-leaf plot) :

```
3 | 48
4 | 489
5 | 034589
6 | 001235678
7 | 335
8 | 17
```

e) Common to both : Shape of distribution, spread, outliers. Frequency histogram only : scaling, class width. Stem and leaf plot only : spread of data within each “bar”.

f) sample mean, sample variance and sample standard deviation for the inflows:

```
>> mean(Inflow)
ans =
862.8400
>> var(Inflow)
ans =
1.7170e+06
>> std(Inflow)
ans =
1.3104e+03
```

$\rightsquigarrow$ $\bar{x} = 862.84$ m^3/s, $s^2 = 1717000$ (m^3/s)^2 and $s = 1310.4$ m^3/s

In Matlab, the percentiles are computed using the function `quantile` (note that quantile is another name for percentile). As the minimum and the maximum values of a sample can be regarded as the 0th percentile and 100th percentile, an elegant way to compute the five-number summary at once is :

```
>> quantile(Inflow,[0 0.25 0.5 0.75 1])
ans =
1.0e+03 *
0.0300 0.1842 0.4120 0.8143 6.1000
```

$\rightsquigarrow$ the five-number summary is $\{x_{(1)}, q_1, m, q_3, x_{(n)}\} = \{30, 184.2, 412.0, 814.3, 6100\}$ (all in m^3/s)

Note that MATLAB uses a slightly different definition of the sample percentiles than the one given in the course (MATLAB uses linear interpolation), so that the above values of q_1 and q_3 are not exactly the ones you would have found 'manually' (they are however very close). Note also that in online quizzes and the mid-semester test, students will be assessed on their ability to *use Matlab* to get a five-number summary – in any computer test they will be expected to **use** MATLAB to get the five-number summary.

g) For a density histogram with, say, 5 classes, and appropriate labelling and title :

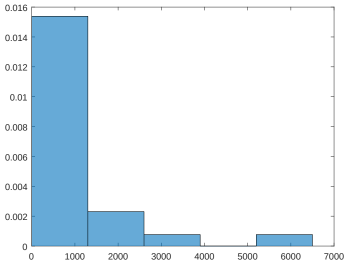
```
>> histogram(Inflow,5,'Normalization','pdf')
>> xlabel('Inflows')
>> ylabel('Density')
>> title('Density histogram for the annual maximum inflow')
```

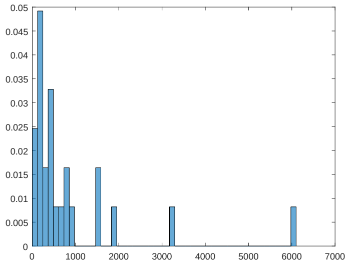
You can compare with what `histogram(Inflow,5)` produces. Here, as we have equal-width classes, only the scale differs from the frequency histogram to the density histogram : the density histogram is scaled so that its total area is 1. The general appearance of the histogram is however not affected.

h) It is straight forward to change the number of classes.

```
>> histogram(Inflow,5,'Normalization','pdf')
>> histogram(Inflow,50,'Normalization','pdf')
```

Both histograms are technically an accurate representation of the data, but they differ in how informative they are. The histogram with 5 classes is quite smooth and gives us a good idea of the shape of the data, while the histogram with 50 classes is noisy and less informative.





i) For a density histogram with classes $[0, 500)$, $[500, 1500)$, $[1500, 3000)$ and $[3000, 6500)$ and appropriate labelling and title :

```
>> edges=[0 500 1500 3000 6500];
>> histogram(Inflow,edges,'Normalization','pdf')
>> xlabel('Inflows')
>> ylabel('Density')
>> title('Density histogram for the annual maximum inflow')
```

j) The histogram shows a highly skewed distribution. As the distribution is right-skewed, the mean must be larger than the median. We know that, unlike the sample median, the sample mean is affected by outlying values: here, that means that the sample mean will be attracted to the right tail of the distribution, but not the sample median. Indeed, we previously found $\bar{x} = 862.84$ which is larger than $m = 412$.

k) frequency histogram with the mentioned classes :

```
>> histogram(Inflow,edges)
```

The frequency histogram gives a distorted representation of reality, as we get the feeling that many observations are larger than 1500 - which is not true. That is because we are using wider classes in that area, but a frequency histogram does not take the class width into account, unlike a density histogram. From the density histogram, it is clear that there are few large observations: the **areas** represent the proportions of observations in each class.

l) Write an m.file including the above commands.

Exercise 11

a) Sample means, sample variance and sample standard deviation for US presidents:

```
>> AgeUS = Age(strcmp(Country,'US'));
>> mean(AgeUS)
ans =
54.9778
>> var(AgeUS)
ans =
43.2040
>> std(AgeUS)
ans =
6.5730
```

$\leadsto$ $\bar{x}_{US} = 55.0$ years, $s^2_{US} = 43.2$ years^2 and $s_{US} = 6.6$ years

Sample means, sample variance and sample standard deviation for Australian PMs :

```
>> AgeOz = Age(strcmp(Country,'Aust'));
>> mean(AgeOz)
ans =
52.4138
>> var(AgeOz)
ans =
46.4655
>> std(AgeOz)
ans =
6.8166
```

$\leadsto$ $\bar{x}_{Aust} = 52.4$ years, $s^2_{Aust} = 46.5$ years^2 and $s_{Aust} = 6.8$ years

b) Side-by-side horizontal boxplots :

```
>> boxplot(Age,Country,'orientation','horizontal')
```

c) Sample median, sample quartiles and sample interquartile range for US presidents :

```
>> quantile(AgeUS,0:0.25:1)
ans =
42.0000 51.0000 55.0000 58.5000 70.0000
>> iqr(AgeUS)
ans =
7.5000
```

⇝ So for example $m_{US} = 55$ years, $q_{1,US} = 51$ years, $q_{3,US} = 58$ years, $\text{iqr}_{US} = 7.5$ years

Sample median, sample quartiles and sample interquartile range for Australian PMs :

```
>> quantile(AgeOz,0:0.25:1)
  ans =
  37.0000 47.0000 53.0000 56.2500 67.0000
  >> iqr(AgeOz)
  ans =
  9.2500
```

⇝ So for example $m_{Aust} = 53$ years, $q_{1,Aust} = 47$ years, $q_{3,Aust} = 56.25$ years, $\text{iqr}_{Aust} = 9.25$ years

On each country's boxplot, the quartiles are the limits of the central box, the iqr is the width of the box, and the median is the red line within the box.

d) This US president was older than 69 when he took office, so we can ask their name with :

```
Name(Age(strcmp(Country,'US'))>69)
  ans =
  cell
  'Trump'
```

e) Both the US box and the US 'whiskers' are shorter than their Australian analogs, indicating a smaller dispersion in the age of the US presidents, materialised by a smaller iqr and variance for US than for Australia. It also appears that US leaders tend to be slightly older at inauguration, which is materialised by a larger mean and larger quartiles (including a larger median), however the boxes have considerable overlap, so that the difference of ages may be not significant.

f) Write an m.file including the above commands.

Exercise 12

a) i) sample mean, sample variance and sample standard deviation for the chest decelerations:

```
>> mean(Chest)
ans =
51.6446
>> var(Chest)
ans =
90.9820
>> std(Chest)
ans =
9.5384
```

⇝ $\bar{x} = 51.64$ g, $s^2 = 90.98$ g^2 and $s = 9.54$ g

As in Exercise 1, compute the five-number summary as:

```
>> quantile(Chest,[0 0.25 0.5 0.75 1])
ans =
35 44 51 58 97
```

$\rightsquigarrow$ the five-number summary is $\{x_{(1)}, q_1, m, q_3, x_{(n)}\} = \{35, 44, 51, 58, 97\}$ (all in g)

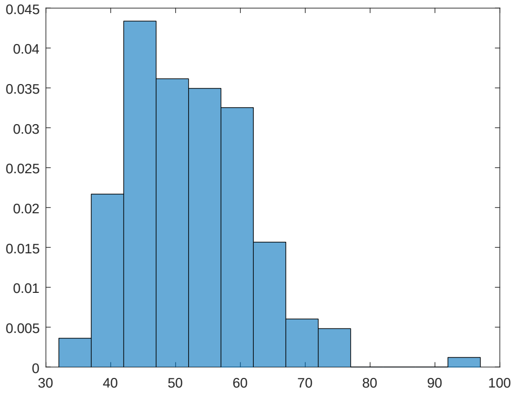
ii) A command like `Chest>=60` will return a vector of 0/1's depending on whether the condition is true or not. Thus, an easy way to compute the proportion of observations for which `Chest` is larger than or equal to 60 is to compute the mean of this vector:

```
>> mean(Chest>=60)
ans =
0.1807
```

The `mean` command will indeed sum the 0 and 1 values, which will give the total number of observations ≥ 60, and then divide by the total number of observations $\rightsquigarrow$ proportion

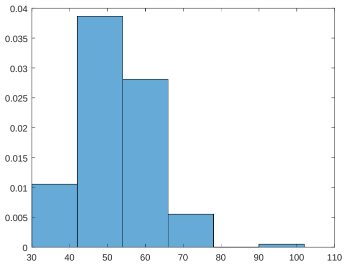
iii) The number of observations in the `Chest` vector is 166, hence an appropriate number of classes should be close to $\sqrt{166} = 12.88$. The density histogram with 13 classes is:

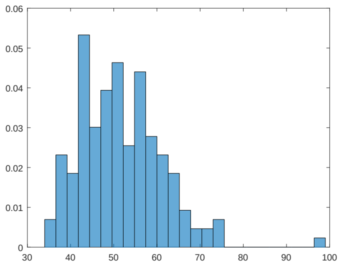
```
>> histogram(Chest,'Normalization','pdf')
```



The distribution is unimodal and right-skewed. There is an outlier at around 100. The typical value is around 50, and the range is $\sim [35, 80]$.
With 6 or 25 classes, we would get:

```
>> histogram(Chest,6,'Normalization','pdf')
>> histogram(Chest,25,'Normalization','pdf')
```

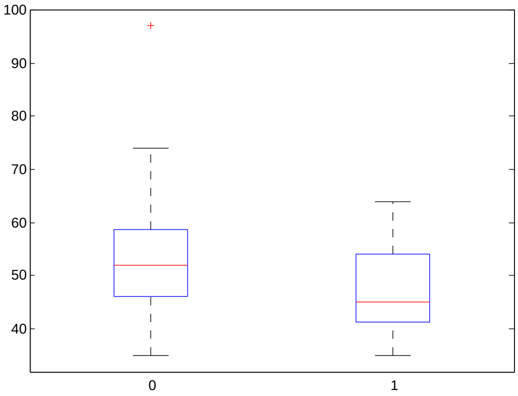




Too many classes leads to a picture with too much (irrelevant) information. Not enough classes gives us an unclear picture. The apparent shape of the distribution is affected by the number of classes chosen. A bad choice can obscure important characteristics of the data: having too few classes is as bad as having too many. It is therefore important to carefully select the number of classes: the histogram with 13 classes above is way nicer and easier to interpret than the other two.

b) i) Compare the distribution of chest deceleration for airbag (`Airbag = 1`) and non-airbag (`Airbag = 0`) cars:

```
>> boxplot(Chest,Airbag)
```



The presence of an airbag does indeed appear to prevent injuries (to some extent). The boxplot for airbag cars is centred at a lower position than the one for non-airbag cars. The median (red line) chest deceleration is around 44 g for airbag cars, while it is around 52 g for non-airbag cars. Having said that, the boxes have considerable overlap, and some inferential investigation should be carried out to conclude further.

ii) The interquartile ranges are reasonably similar (as shown by the lengths of the boxes), though the overall range of the data set is greater for the non-airbag cars (as shown by the distances between the ends of the two whiskers for each boxplot).

iii) The non-airbag data set shows a suspiciously far out value (around 100) which might require a closer look.

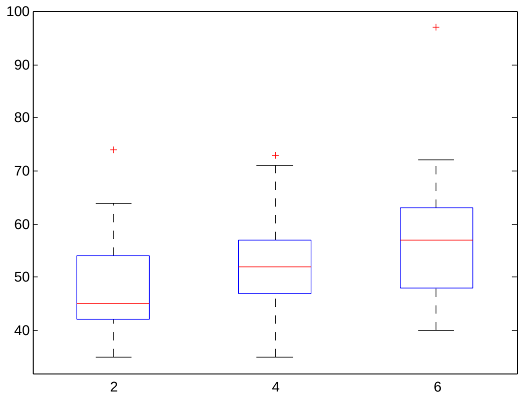
iv) The distribution for the non-airbag cars appears fairly symmetric. On the other hand, the one for the airbag cars appears to be right-skewed (the median is lower than the centre of the box). Yet, if we compute the skewness of both data sets:

```
>> skewness(Chest(Airbag==1))
ans =
0.4543
>> skewness(Chest(Airbag==0))
ans =
0.8855
```

$\leadsto$ the skewness coefficient for the non-airbag cars is strongly inflated by the presence of an outlier. This is why it is always important to spot any potential outlying value: most of the statistical procedures will be adversely affected by their presence and an appropriate treatment is required.

c) i) Compare the chest injuries of vehicles with different numbers of doors:

```
>> boxplot(Chest,Doors)
```



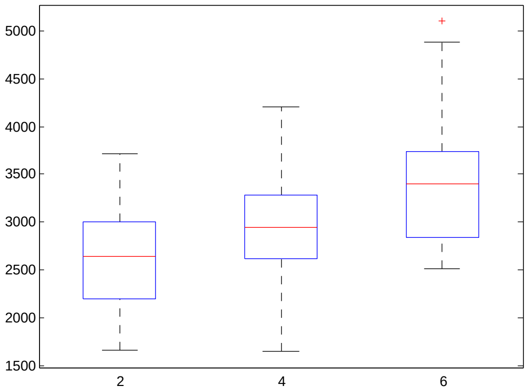
There appears to be a relationship between the number of doors and chest deceleration. The median (and the central location of the boxes) increases with the number of doors.

ii) Vehicles with two doors tend to have the least severe injuries.

iii) Many explanations could be put forward. One is related to the weight of the vehicle. Vehicles with two doors may be lighter than the others, so admit a lower inertia when colliding with the wall, so that the shock is less violent (even at equal speed).

d) Make a boxplot of `Weight` by `Doors`:

```
>> boxplot(Weight,Doors)
```



⇝ as expected, vehicle weight increases with the number of doors. This somewhat confirms the explanation given at the previous subquestion.

e) Write an m.file including the above commands.

MATH2089
Numerical Methods and Statistics

Statistics – Chapter 3 – Probability

SOLUTIONS

Exercise 1

Probability of failure is given in the series case by one or more of the components failing. Hence let A_i represent the event that the i-th component has failed, where $\mathbb{P}(A_i) = 0.01$. Hence we can consider the probability of system failure given by the probability of the union of each of these events i.e. either the first or the second or the third etc. element fails, given by $\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5)$. We can find an upper bound for this probability using the inequality in the question and summing the probability of failure for each element :

$$\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5) \leq 5 \times 0.01 = 0.05$$

Probability of the system working is given in the series case by all the components working. Hence let A_i represent the event that the i-th component is functioning without failure, where $\mathbb{P}(A_i) = 0.999$. So we have directly that $\mathbb{P}(A_i^c) = 1 - \mathbb{P}(A_i) = 0.001$. Hence we can consider the probability of the system working given by the probability of the intersection of each of these events i.e. all the first and the second and the third etc. elements function, given by $\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_{10})$. We can find a lower bound for this probability using the inequality in the question :

$$\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_{10}) \geq 1 - (\mathbb{P}(A_1^c) + \mathbb{P}(A_2^c) + \ldots + \mathbb{P}(A_{10}^c)) = 1 - 10 \times 0.001 = 0.99$$

Exercise 2

Let $A =$ "the selected student is in the civil engineering program" and $B =$ "the selected student is a male", and let x be the unknown number of females enrolled in the chemical engineering program. Then the total number of students is $4 + 6 + 6 + x = 16 + x$ students.

So, as the student is selected at random, we have

$$\mathbb{P}(A \cap B) = \frac{4}{16 + x}$$

(4 males in the civil engineering program out of $16 + x$ students),

$$\mathbb{P}(A) = \frac{10}{16 + x}$$

(10 students in the civil engineering program out of $16 + x$ students) and

$$\mathbb{P}(B) = \frac{10}{16 + x}$$

(10 males out of $16 + x$ students). The events A and B will be independent if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$, that is

$$\frac{4}{16 + x} = \frac{10}{16 + x} \times \frac{10}{16 + x}$$

Solving for x yields $x = 9$ females enrolled in the chemical engineering program. Incidentally, it is easy to show (do it!) that if two events A and B are independent, then A^c and B^c are also independent events, as well as A is independent of B^c, and B is independent of A^c. This shows that gender and program are independent variables if A and B are independent events.

Exercise 3

(1) There are three prisoners: you, A and B. Define the events $Y/A/B$ as "you/A/B is chosen to be hanged". See that $\mathbb{P}(Y) = \mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{3}$ (prior probability of being hanged).

Now, denote $J_Y/J_A/J_B$ the events "the jailer says that you/A/B will be set free". By Bayes' first rule,

$$\mathbb{P}(Y|J_A) = \mathbb{P}(J_A|Y)\frac{\mathbb{P}(Y)}{\mathbb{P}(J_A)}$$

Now, if you are chosen to be hanged, it is reasonable to suppose that the answer of the jailer is equally likely to be A or B, that is,

$$(\star) \qquad \mathbb{P}(J_A|Y) = \mathbb{P}(J_B|Y) = \frac{1}{2}$$

By the law of total probability, we also have

$$\begin{aligned}\mathbb{P}(J_A) &= \mathbb{P}(J_A|A)\mathbb{P}(A) + \mathbb{P}(J_A|B)\mathbb{P}(B) + \mathbb{P}(J_A|Y)\mathbb{P}(Y)\\ &= 0 + 1 \times 1/3 + 1/2 \times 1/3 = \frac{1}{2}\end{aligned}$$

Hence,

$$\mathbb{P}(Y|J_A) = 1/2 \times \frac{1/3}{1/2} = \frac{1}{3},$$

so that $\mathbb{P}(Y|J_A) = \mathbb{P}(Y)$. Similarly, we could find $\mathbb{P}(Y|J_B) = \mathbb{P}(Y)$, so that you'll be hanged or not independently of the answer of the jailer.

(2) $\mathbb{P}(B|J_A) = \mathbb{P}(J_A|B)\frac{\mathbb{P}(B)}{\mathbb{P}(J_A)} = 1 \times \frac{1/3}{1/2} = \frac{2}{3}$, to be compared with 'your' $\mathbb{P}(Y|J_A) = 1/3$ $\rightsquigarrow$ you'd better not take his identity.

(In fact, the previous answers (a) and (b) depend critically on the assumption ($\star$). Suppose that, if you are the one chosen to be hanged, the jailer always tells you that A will be set free, so that $\mathbb{P}(J_A|Y) = 1$ and $\mathbb{P}(J_B|Y) = 0$. How does this affect the previous answer? This problem has become known as the *Three Prisoners paradox*, or sometimes also the *Monty Hall problem*.)

Exercise 4

a) Because the events are independent, the fact that the Asian project is not successful does not affect the probability of success of the European project. Hence, $\mathbb{P}(B|A^c) = \mathbb{P}(B) = 0.7$ and the probability of the European project not being successful remains 0.3.

b) The event "at least one of the two projects is successful" is $(A \cup B)$. The additive law of probability states that

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

and because A and B are independent, $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B) = 0.4 \times 0.7 = 0.28$. Hence

$$\mathbb{P}(A \cup B) = 0.4 + 0.7 - 0.28 = 0.82.$$

c) The desired probability is

$$\mathbb{P}((A \cap B^c)|(A \cup B)) = \frac{\mathbb{P}((A \cap B^c) \cap (A \cup B))}{\mathbb{P}(A \cup B)}$$

by the definition of conditional probability. Now, $(A \cap B^c) \subset (A \cup B)$ ($(A \cap B^c)$ implies $(A \cup B)$) so that $(A \cap B^c) \cap (A \cup B) = (A \cap B^c)$, and

$$\mathbb{P}((A \cap B^c) \cap (A \cup B)) = \mathbb{P}(A \cap B^c) = \mathbb{P}(A) \times \mathbb{P}(B^c) = \mathbb{P}(A)(1 - \mathbb{P}(B))$$

again because A and B are independent. Hence,

$$\mathbb{P}((A \cap B^c)|(A \cup B)) = \frac{0.4 \times 0.3}{0.82} = 0.1463$$

Exercise 5

a) Since A and B are mutually exclusive,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) = 0.2 + 0.5 = 0.7$$

b)

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F) = 0.2 + 0.3 - 0.1 = 0.4$$

c) Since E and F are independent, $\mathbb{P}(E \cap F) = \mathbb{P}(E) \times \mathbb{P}(F) = 0.4 \times 0.25 = 0.1$. So

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F) = 0.4 + 0.25 - 0.1 = 0.55$$

d) $\mathbb{P}(A^c \cap B^c) = \mathbb{P}[(A \cup B)^c] = 1 - \mathbb{P}(A \cup B)$ and so

$$\mathbb{P}(A^c \cap B^c) = 1 - [\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)] = 1 - (0.2 + 0.4 - 0.1) = 1 - 0.5 = 0.5$$

e) Let T be the event that they test positive, and A be the event that they have the antibodies. We know from the question that:

$$\mathbb{P}(T|A) = 0.95 \qquad \mathbb{P}(T|A^c) = 0.02 \qquad \mathbb{P}(A) = 0.05$$

i) Using the law of total probability,

$$\mathbb{P}(T) = \mathbb{P}(A) \times \mathbb{P}(T|A) + \mathbb{P}(A^c) \times \mathbb{P}(T|A^c) = 0.05 \times 0.95 + 0.95 \times 0.02 = 0.0665$$

ii) Using Bayes' first law,

$$\mathbb{P}(A|T) = \frac{\mathbb{P}(T|A) \times \mathbb{P}(A)}{\mathbb{P}(T)} = \frac{0.95 \times 0.05}{0.0665} \simeq 0.714$$

Exercise 6

The command `randi` simulates values drawn at random from a set of integers. To simulate a sequence of 1,000 values (technically, a matrix of 1000 rows and 1 column) drawn at random from $\{1, 2, 3, 4, 5, 6\}$, we write

```
>> randi([1,6],1000,1)
ans =
5
2
6
2
5
5
(...)
```

Call such a sequence D:

```
>> D=randi([1,6],1000,1)
```

Then you can compute the proportions of time the events A, B and $A \cap B$ have occurred out of your 1,000 'tosses'. As these events are made up of several simple events, an easy way to do that in MATLAB is to use the command `ismember` (use `help` if you are unsure how to use it). Whether or not the event $A = \{2, 4, 6\}$ has occurred is found from

```
>> ismember(D,[2,4,6]);
```

Then the proportion of occurrences of A is

```
>> mean(ismember(D,[2,4,6]))
  ans =
  0.4970
```

Hence, in our simulation (yours should be slightly different!), $p_A = 0.4970$, indeed close to $\mathbb{P}(A) = 1/2$. Similarly, as $B = \{1, 2, 3, 4\}$, the estimation of $\mathbb{P}(B)$ is found from

```
>> mean(ismember(D,1:4))
  ans =
  0.6840
```

Hence, in our simulation (yours should be slightly different!), $p_B = 0.6840$, indeed close to $\mathbb{P}(B) = 2/3$. Finally, as $A \cap B = \{2, 4\}$, the estimation of $\mathbb{P}(A \cap B)$ is found from

```
>> mean(ismember(D,[2,4]))
  ans =
  0.3380
```

Hence, in our simulation (yours should be slightly different!), $p_{AB} = 0.3380$, indeed close to $\mathbb{P}(A \cap B) = 1/3$. It can now easily be checked that

$$p_A \times p_B = 0.4970 \times 0.6840 = 0.3399 \simeq 0.3380 = p_{AB},$$

as expected.

Now consider for instance the events $A = \{2, 4, 6\}$ =‘getting an even number’ and $B = \{1, 3, 5\}$ =‘getting an odd number’. We get:

```
>> mean(ismember(D,[2,4,6]))
  ans =
  0.4970
  >> mean(ismember(D,[1,3,5]))
  ans =
  0.5030
```

So, $p_A \times p_B = 0.4970 \times 0.5030 = 0.25$ and this is clearly not equal to $p_{AB} = 0$ (A and B cannot occur together!) These events A and B are not independent (which was quite obvious, by the way). This example also makes it clear that **mutually exclusive events cannot be independent**.

MATH2089
Numerical Methods and Statistics

Statistics – Chapter 4 – Random variables

SOLUTIONS

Exercise 1

We have $X_1, X_2, \dots, X_n \sim F$, independent of one another.

a) Denote $X_{(n)}$ the maximum value. Then, the cdf of $X_{(n)}$ is

$$\begin{aligned} F_{X_{(n)}}(x) &= \mathbb{P}(X_{(n)} \leq x) \\ &= \mathbb{P}\left(X_1 \leq x, X_2 \leq x, \dots, X_n \leq x\right) \\ &\overset{\text{ind.}}{=} \mathbb{P}(X_1 \leq x) \times \mathbb{P}(X_2 \leq x) \times \dots \times \mathbb{P}(X_n \leq x) \\ &= (F(x))^n \end{aligned}$$

Hence,

$$f_{X_{(n)}}(x) = \frac{d}{dx} F_{X_{(n)}}(x) = n\,(F(x))^{n-1} f(x)$$

b) Denote $X_{(1)}$ the minimum value. Then, the cdf of $X_{(1)}$ is

$$\begin{aligned} F_{X_{(1)}}(x) &= \mathbb{P}(X_{(1)} \leq x) \\ &= 1 - \mathbb{P}(X_{(1)} > x) \\ &= 1 - \mathbb{P}\left(X_1 > x, X_2 > x, \dots, X_n > x\right) \\ &\overset{\text{ind.}}{=} 1 - \mathbb{P}(X_1 > x) \times \mathbb{P}(X_2 > x) \times \dots \times \mathbb{P}(X_n > x) \\ &= 1 - (1 - F(x))^n \end{aligned}$$

Hence,

$$f_{X_{(1)}}(x) = \frac{d}{dx} F_{X_{(1)}}(x) = -n\,(1 - F(x))^{n-1}\,(-f(x)) = n\,(1 - F(x))^{n-1} f(x)$$

Exercise 2

Define the random variables

$$X_1 = \begin{cases} 1 & \text{if the surgery on your } \textbf{left} \text{ knee is successful} \\ 0 & \text{if not} \end{cases}$$

and

$$X_2 = \begin{cases} 1 & \text{if the surgery on your } \textbf{right} \text{ knee is successful} \\ 0 & \text{if not} \end{cases}$$

The total number of successful surgeries that you will undergo is obviously

$$X = X_1 + X_2.$$

As both X_1 and X_2 can only take the values 0 and 1, X can only take the values 0 or 1 or 2, so $S_X = \{0, 1, 2\}$. See that

- $X = 0 \iff X_1 = 0$ and $X_2 = 0$. **Because X_1 and X_2 are independent**, $\mathbb{P}(X = 0) = \mathbb{P}(X_1 = 0, X_2 = 0) = \mathbb{P}(X_1 = 0) \times \mathbb{P}(X_2 = 0) = (1 - 0.9)(1 - 0.67) = 0.033$
- $X = 2 \iff X_1 = 1$ and $X_2 = 1$. **Because X_1 and X_2 are independent**, $\mathbb{P}(X = 2) = \mathbb{P}(X_1 = 1, X_2 = 1) = \mathbb{P}(X_1 = 1) \times \mathbb{P}(X_2 = 1) = 0.9 \times 0.67 = 0.603$
- $X = 1$ in any other situations, so $\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 2) = 1 - 0.033 - 0.603 = 0.364$ (note that you can also find this value writing $(X = 1) = (X_1 = 1 \cap X_2 = 0) \cup (X_1 = 0 \cap X_2 = 1)$ and the corresponding probability)

So the pmf of X is

$$p(0) = 0.033, \qquad p(1) = 0.364, \qquad p(2) = 0.603$$

(understood that $p(x) = 0$ for any other value of x). Now, from the properties of expectations and variances, we have

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \mathbb{E}(X_2),$$

and, **because X_1 and X_2 are independent**,

$$\mathbb{V}\text{ar}(X) = \mathbb{V}\text{ar}(X_1) + \mathbb{V}\text{ar}(X_2).$$

Now, both X_1 and X_2 are Bernoulli random variables

$$X_1 \sim \text{Bern}(\pi_1) \qquad \text{and} \qquad X_2 \sim \text{Bern}(\pi_2)$$

with $\pi_1 = 0.9$ and $\pi_2 = 0.67$, so that, from the properties of the Bernoulli distribution, we have directly

$$\mathbb{E}(X_1) = \pi_1 = 0.9 \text{ and } \mathbb{V}\text{ar}(X_1) = \pi_1 \times (1 - \pi_1) = 0.9 \times 0.1 = 0.09$$

and

$$\mathbb{E}(X_2) = \pi_2 = 0.67 \text{ and } \mathbb{V}\text{ar}(X_2) = \pi_2 \times (1 - \pi_2) = 0.67 \times 0.33 = 0.2211.$$

Finally,

$$\mathbb{E}(X) = 0.9 + 0.67 = 1.57 \text{ (surgeries)}$$

(on average, 1.57 surgeries out of 2 will be successful) and

$$\mathbb{V}\text{ar}(X) = 0.09 + 0.221 = 0.3111 \text{ (surgeries}^2\text{)}$$

(Note $\mathbb{E}(X)$ and $\mathbb{V}\text{ar}(X)$ can also be derived directly from the pmf of X)

Exercise 3

a) The two conditions for f to be a legitimate density function are $f(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f(x)\,dx = 1$. As the exponential is alway positive,

$$f(x) \geq 0 \quad \forall x \iff c \geq 0.$$

The second condition gives

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} c \times e^{-0.2\,|x|}\,dx = \int_{-\infty}^{0} c \times e^{0.2\,x}\,dx + \int_{0}^{\infty} c \times e^{-0.2\,x\,dx}$$

$$= c \times \left(\left[\frac{e^{0.2\,x}}{0.2} \right]_{-\infty}^{0} + \left[\frac{e^{-0.2\,x}}{(-0.2)} \right]_{0}^{\infty} \right) = c \times \left(\frac{1}{0.2} + \frac{1}{0.2} \right) = \frac{c}{0.1}$$

As this integral must equal 1, we find

$$c = 0.1.$$

The resulting density is thus

$$f(x) = 0.1 \times e^{-0.2\,|x|} \qquad \text{for } -\infty < x < \infty,$$

sketched in the figure below.

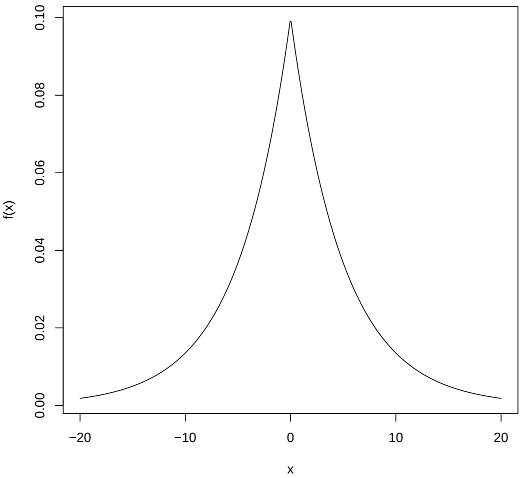
b) We computed c so as the total area under the curve, that is, $\int_{-\infty}^{\infty} f(x)\,dx$, is equal to 1. So, by symmetry, the area under the curve for $x \in (-\infty, 0]$, equal to the probability of X being in $(-\infty, 0]$, is equal to $1/2$ $\rightsquigarrow$ we expect half of the observed errors to be negative.

We have also

$$\mathbb{P}(X \leq 2) = \int_{-\infty}^{2} f(x)\,dx = 1/2 + 0.1 \int_{0}^{2} e^{-0.2\,x}\,dx = 1/2 + 0.1 \left[\frac{e^{-0.2\,x}}{(-0.2)} \right]_{0}^{2} = 0.6648$$

$\rightsquigarrow$ we expect 66.48% of the observed errors to be smaller than 2. Note that it could have been easier to write

$$\mathbb{P}(X \leq 2) = 1 - \mathbb{P}(X > 2) = 1 - \int_{2}^{\infty} f(x)\,dx,$$



and we would have found the same answer.
Finally,

$$\mathbb{P}(-1 \leq X \leq 2) = \int_{-1}^{2} f(x)\, dx = 0.1 \left(\int_{-1}^{0} e^{0.2\,x}\, dx + \int_{0}^{2} e^{-0.2\,x}\, dx \right) = \ldots = 0.2555$$

$\rightsquigarrow$ we expect 25.55% of the observed errors to be between -1 and 2.

Exercise 4

Let X be the random variable "weight of a package".

a) By definition, we have

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \frac{70}{69} \int_{1}^{70} x\, \frac{1}{x^2}\, dx = \frac{70}{69} \int_{1}^{70} \frac{1}{x}\, dx = \frac{70}{69} \left[\log(x)\right]_1^{70} = \frac{70}{69} \log(70) = 4.31 \text{ (kg)}$$

$\rightsquigarrow$ on average, the weight of a package delivered by that post office is 4.31 kg.
Similarly,

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \frac{70}{69} \int_{1}^{70} x^2\, \frac{1}{x^2}\, dx = \frac{70}{69} \int_{1}^{70} dx = \frac{70}{69}(70-1) = 70 \text{ (kg}^2\text{)},$$

so

$$\mathbb{V}\text{ar}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 70 - 4.31^2 = 51.42 \text{ (kg}^2\text{)}$$

b) Denote C the shipping cost for a package. We have that

$$C = 2.5 \times X.$$

From the properties of the expectation,

$$\mathbb{E}(C) = 2.5 \times \mathbb{E}(X) = 2.5 \times 4.31 = 10.78 \text{ (\$)}$$

For the variance, we have

$$\mathbb{V}\text{ar}(C) = 2.5^2 \times \mathbb{V}\text{ar}(X) = 6.25 \times 51.42 = 321.4 \text{ (\$}^2\text{)}$$

c) The long-term proportion of packages with weight exceeding 50 kg is

$$\mathbb{P}(X > 50) = \int_{50}^{\infty} f(x)\, dx = \frac{70}{69} \int_{50}^{70} \frac{1}{x^2}\, dx = \frac{70}{69} \left[-\frac{1}{x} \right]_{50}^{70} = \frac{70}{69} \left(\frac{1}{50} - \frac{1}{70} \right) = 0.0058$$

Exercise 5

a)
   i) a) and d)
   ii) a)
   iii) b)
   iv) c)
   v) a) large positive correlation, b) moderate negative correlation, c) no correlation, d) moderate positive correlation

b)
   i) One would expect a positive correlation since the hotter the maximum daily temperature the greater the amount of cooling utilised.
   ii) Interest rates directly impact on monthly loan repayment amounts, therefore one would expect a negative correlation between increasing interest rates and decreasing loan applications.
   iii) There would likely be little relationship between the marks of a student and their distance lived from campus.

MATH2089
Numerical Methods and Statistics

Statistics – Chapter 5 – Special random variables

SOLUTIONS

Exercise 1

a) i) $X \sim \text{Bin}(6, 0.5)$

ii) neither Binomial nor Poisson. This distribution (first occurrence of a success is a series of Bernoulli trials, see Slide 176) is called the **Geometric** distribution.

iii) $X \sim \mathcal{P}(3)$

iv) $X \sim \text{Bin}(500, 1/12)$, note that here because n is 'large' and π is 'small', the distribution of X can also be approximated by $\mathcal{P}(41.67)$ $(500 \times 1/12 = 41.67)$

v) $X \sim \mathcal{P}(250)$

vi) $X \sim \text{Bin}(2, 1/6)$

vii) neither Binomial nor Poisson

viii) neither Binomial nor Poisson

b) take the random variable in i) for instance, $X \sim \text{Bin}(6, 0.5)$. Then, from the Binomial pmf,

$$\mathbb{P}(X \leq 1) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) = \binom{6}{0} 0.5^0 0.5^6 + \binom{6}{1} 0.5^1 0.5^5 = 0.109$$

c) take the random variable in iii) for instance, $X \sim \mathcal{P}(3)$. Then, from the Poisson pmf,

$$\mathbb{P}(X \leq 1) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) = e^{-3}\frac{3^0}{0!} + e^{-3}\frac{3^1}{1!} = 0.199$$

d) Basically, the number X of errors out of 1 billion divisions would be binomially distributed:

$$X \sim \text{Bin}\left(1{,}000{,}000{,}000; \frac{1}{9{,}000{,}000{,}000}\right)$$

but these parameters are not tractable. However, we are obviously in the situation 'large n, small π' (Slide 191), so that we can use the Poisson approximation to the Binomial distribution:

$$X \sim \mathcal{P}\left(\frac{1{,}000{,}000{,}000}{9{,}000{,}000{,}000}\right) = \mathcal{P}(1/9)$$

Hence,

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - e^{-1/9} = 0.1051$$

$\rightsquigarrow$ not that unlikely

Exercise 2

Denote X the number of correct predictions that the individual gets. It is clear that

$$X \sim \text{Bin}(10, \pi),$$

where π is the probability that he predicts a right outcome for a toss. Without ESP, π would be equal to 1/2 (random guess between two alternatives), while if he has ESP, then $\pi > 1/2$. Suppose $\pi = 1/2$ (no ESP), then the probability that he gets at least seven out of ten correct is

$$\mathbb{P}(X \geq 7) = \sum_{x=7}^{10} \mathbb{P}(X = x) = \sum_{x=7}^{10} \binom{10}{x} (1/2)^x \simeq 0.17$$

$\rightsquigarrow$ not that unlikely to have guessed 7 out of 10 outcomes without ESP
$\rightsquigarrow$ there is no clear evidence that he has ESP
$\rightsquigarrow$ we do not really believe in his powers
Note that this does not mean that we are sure that the individual has no ESP! We are just not convinced enough. It is important to understand this way of thinking as it will be the basis of the **hypothesis tests** introduced in a subsequent chapter.

Exercise 3

Let X be grain size. We know that $X \sim \mathcal{N}(96, 14)$.

i) Standardising the random variable, we get

$$Z = \frac{X-96}{14} \sim \mathcal{N}(0,1).$$

Hence, from the MATLAB output,

$$\mathbb{P}(X > 100) = \mathbb{P}\left(Z > \frac{100-96}{14}\right) = \mathbb{P}(Z > 0.29) = 1 - \mathbb{P}(Z \leq 0.29) = 1 - 0.6141 = 0.3859$$

ii)

$$\begin{aligned}\mathbb{P}(50 < X < 80) &= \mathbb{P}\left(\frac{50-96}{14} < Z < \frac{80-96}{14}\right)\\ &= \mathbb{P}(-3.29 < Z < -1.14)\\ &= \mathbb{P}(Z < -1.14) - \mathbb{P}(Z < -3.29)\\ &= 0.1271 - 0.0005 = 0.1266\end{aligned}$$

(from the MATLAB output).

iii) Using the MATLAB output, by symnmetry, we have

$$\mathbb{P}(-1.645 < Z < 1.645) = 0.90$$

Thus,

$$\begin{aligned}0.90 &= \mathbb{P}\left(-1.645 < \frac{X-96}{14} < 1.645\right)\\ &= \mathbb{P}(96 - 1.645 \times 14 < X < 96 + 1.645 \times 14)\\ &= \mathbb{P}(72.97 < X < 119.03)\end{aligned}$$

$\rightsquigarrow$ the central 90% of all grain sizes are in $(72.97, 119.03)$ (μm).

Exercise 4

a) i) $\mathbb{P}(Y = 5)$

```
>> binopdf(5,20,0.35)
ans =
0.1272
```

ii) $\mathbb{P}(Y = 0)$

```
>> binopdf(0,20,0.35)
ans =
1.8125e-04
```

iii) $\mathbb{P}(Y \leq 0)$

```
>> binocdf(0,20,0.35)
ans =
1.8125e-04
```

(same as the previous one, as Y can only take value in $S_Y = \{0, 1, \dots, 20\}$)

iv) $\mathbb{P}(5 \leq Y < 15) = \mathbb{P}(Y \leq 14) - \mathbb{P}(Y \leq 4)$

```
>> binocdf(14,20,0.35)-binocdf(4,20,0.35)
ans =
0.8815
```

v) $\mathbb{P}(5 < Y \leq 15) = \mathbb{P}(Y \leq 15) - \mathbb{P}(Y \leq 5)$

```
>> binocdf(15,20,0.35)-binocdf(5,20,0.35)
ans =
0.7546
```

(not the same as the previous probability, as Binomial is a discrete distribution)

vi) $\mathbb{P}(Y \in \{1,5,9,17\}) = \mathbb{P}(Y=1) + \mathbb{P}(Y=5) + \mathbb{P}(Y=9) + \mathbb{P}(Y=17)$ (mutually exclusive events, Y cannot take simultaneously two different values in this set)

```
>> sum(binopdf([1 5 9 17],20,0.35))
ans =
0.2450
```

vii) $\mathbb{P}(Y \notin \{5,13,16\})$

```
>> 1-sum(binopdf([5 13 16],20,0.35))
ans =
0.8683
```

b) Denote X the number of correct answers out of 10 questions. Then, $X \sim \text{Bin}(10, 1/5)$.

i) $\mathbb{P}(X = 10)$

```
>> binopdf(10,10,1/5)
ans =
1.0240e-07
```

($\simeq$ 1 chance in 10 millions, don't dream)

ii) $\mathbb{P}(X \geq 5) = 1 - \mathbb{P}(X \leq 4)$

```
>> 1-binocdf(4,10,1/5)
ans =
0.0328
```

(quite unlikely as well)

iii) Denote Y the number of quizzes passed by the student. Then, $Y \sim \text{Bin}(3, 0.0328)$.

i) $\mathbb{P}(Y \geq 1) = 1 - \mathbb{P}(Y = 0)$

```
>> 1-binopdf(0,3,0.0328)
ans =
0.0952
```

Exercise 5

a) i) $\mathbb{P}(Q \leq 3)$

```
>> poisscdf(3,7)
ans =
0.0818
```

ii) $\mathbb{P}(Q < 3) = \mathbb{P}(Q \leq 2)$

```
>> poisscdf(2,7)
ans =
0.0296
```

(not the same as the previous one, as Poisson is a discrete distribution)

iii) $\mathbb{P}((Q = 19) \cup (Q = 20)) = \mathbb{P}(Q = 19) + \mathbb{P}(Q = 20)$ (mutually exclusive events, Q cannot simultaneously be equal to 19 and 20)

```
>> sum(poisspdf([19 20],7))
ans =
1.1536e-04
```

iv) $\mathbb{P}(Q > 0) = 1 - \mathbb{P}(Q = 0)$

```
>> 1-poisspdf(0,7)
ans =
0.9991
```

b) Denote X the number of stars in 16 cubic light-years. Then, $X \sim \mathcal{P}(1)$.

i) $\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X \leq 1)$

```
>> 1-poisscdf(1,1)
ans =
0.2642
```

ii) If X the number of stars in 16 cubic light-years is $\mathcal{P}(1)$, then X_n, the number of stars in $16 \times n$ cubic light-years is $\mathcal{P}(n \times 1) = \mathcal{P}(n)$. Thus we desire the values of n such that $\mathbb{P}(X_n \geq 1) = 1 - \mathbb{P}(X_n = 0) \geq 0.95$, with $X_n \sim \mathcal{P}(n)$. Trying several values of n, let say every integer between 1 an 10 :

```
>> 1-poisspdf(0,1:10)
ans =
0.6321 0.8647 0.9502 0.9817 0.9933 0.9975 0.9991 0.9997 0.9999 1.0000
```

We see that the probability will exceed 0.95 for $n \geq 3$, so that $3 \times 16 = 48$ cubic light-years of space must be studied to be sure at more than 95% to find at least one star.

Obviously, you can also solve this explicitly : as $X_n \sim \mathcal{P}(n)$, its pmf is $\mathbb{P}(X_n = x) = e^{-n}\frac{n^x}{x!}$ for $x = 0, 1, \ldots$. Hence,

$$\mathbb{P}(X_n > 0) = 1 - \mathbb{P}(X_n = 0) = 1 - e^{-n}$$

If you want this probability to be 0.95, you have

$$\begin{aligned} 1 - e^{-n} &= 0.95 \\ \iff e^{-n} &= 0.05 \\ \iff -n &= \log 0.05 \\ \iff n &= -\log 0.05 \end{aligned}$$

Now,

```
>> -log(0.05)
ans =
2.9957
```

so that $n = 2.9957$, that is, studying $16 \times 2.9957 = 47.9317$ cubic light-years of space, would be enough.

Exercise 6

a) i) $\mathbb{P}(X < 0) = 1/2$ by symmetry. Obviously we find :

```
>> unifcdf(0,-1,1)
ans =
0.5000
```

ii) $\mathbb{P}(X \leq 0) = \mathbb{P}(X < 0)$, as Uniform is a continuous distribution. For that matter, the MATLAB command is the same to compute $\mathbb{P}(X \leq x)$ and $\mathbb{P}(X < x)$: `unifcdf`

iii) $\mathbb{P}(-0.9 \leq X \leq 0.8) = \mathbb{P}(X \leq 0.8) - \mathbb{P}(X < -0.9)$

```
>> unifcdf(0.8,-1,1)-unifcdf(-0.9,-1,1)
ans =
0.8500
```

iv) $\mathbb{P}(-x \leq X \leq x) = \mathbb{P}(X \leq x) - \mathbb{P}(X < -x)$. Now, $\mathbb{P}(X < -x) = 1 - \mathbb{P}(X \geq -x) = 1 - \mathbb{P}(X \leq x)$, by symmetry. Thus, $\mathbb{P}(-x \leq X \leq x) = 2\mathbb{P}(X \leq x) - 1$. If this is to be equal to 0.9, we need $\mathbb{P}(X \leq x) = 0.95$. The command `unifinv` gives you that quantile (use `help unifinv` to learn how to use it).

```
>> unifinv(0.95,-1,1)
ans =
0.9000
```

$\rightsquigarrow x = 0.9$

Note : likewise, the commands `expinv` and `norminv` will return the quantiles of the Exponential distribution and the Normal distribution, respectively.

b) Denote X the thickness of photoresist for a wafer selected at random. We know that $X \sim U_{[0.205,0.215]}$

i) $\mathbb{P}(X > 0.2125) = 1 - \mathbb{P}(X \leq 0.2125)$

```
>> 1-unifcdf(0.2125,0.205,0.215)
ans =
0.2500
```

ii) What is x to have $\mathbb{P}(X > x) = 0.1$? That means : $\mathbb{P}(X \leq x) = 0.9$

```
>> unifinv(0.9,0.205,0.215)
ans =
0.2140
```

$\rightsquigarrow$ $x = 0.214$ is the thickness exceeded by 10% of the wafers

Exercise 7

a) i) $\mathbb{P}(W \leq 2)$

```
>> expcdf(2,2)
ans =
0.6321
```

ii) $\mathbb{P}(W < 2) = \mathbb{P}(W \leq 2)$, as Exponential is a continuous distribution. For that matter, the MATLAB command is the same to compute $\mathbb{P}(W \leq w)$ and $\mathbb{P}(W < w)$: `expcdf`

iii) $\mathbb{P}(10 < W < 13) = \mathbb{P}(W < 13) - \mathbb{P}(W \leq 10)$

```
>> expcdf(13,2)-expcdf(10,2)
ans =
0.0052
```

iv) $\mathbb{P}(W > -5) = 1 - \mathbb{P}(W \leq -5)$ must be 1, as an Exponential random variable can only assume non negative values. Indeed :

```
>> 1-expcdf(-5,2)
ans =
1
```

b) Denote X the time to failure of a randomly selected fan. Then, $X \sim \text{Exp}(1/0.0003)$

i) $\mathbb{P}(X > 10000) = 1 - \mathbb{P}(X \leq 10000)$

```
>> 1-expcdf(10000,1/0.0003)
ans =
0.0498
```

ii) $\mathbb{P}(X \leq 7000)$

```
>> expcdf(7000,1/0.0003)
ans =
0.8775
```

c) Find x such that $\mathbb{P}(X > x) = 0.95$. That means that $\mathbb{P}(X \leq x) = 0.05$

```
>> expinv(0.05,1/0.0003)
 ans =
 170.9776
```

$\rightsquigarrow$ 95% of the fans will last longer than 170.98 hours

Exercise 8

a) i) $\mathbb{P}(-1 < Z < 1) = \mathbb{P}(Z < 1) - \mathbb{P}(Z \leq -1)$

```
>> normcdf(1)-normcdf(-1)
ans =
0.6827
```

Note : the command `normcdf` uses the default values 0 and 1 for μ and σ, that is, if you do not specify them, the values returned by the command are the values for the standard Normal distribution

ii) $\mathbb{P}(-2 < Z < 2) = \mathbb{P}(Z < 2) - \mathbb{P}(Z \leq -2)$

```
>> normcdf(2)-normcdf(-2)
ans =
0.9545
```

iii) $\mathbb{P}(-3 < Z < 3) = \mathbb{P}(Z < 3) - \mathbb{P}(Z \leq -3)$

```
>> normcdf(3)-normcdf(-3)
ans =
0.9973
```

iv) this is just the 68-95-99 rule!

v) $\mathbb{P}(Z < z) = 0.95$

```
>> norminv(0.95)
ans =
1.6449
```

vi) $\mathbb{P}(Z < z) = 0.975$

```
>> norminv(0.975)
ans =
1.9600
```

vii) $\mathbb{P}(Z < z) = 0.995$

```
>> norminv(0.995)
ans =
2.5758
```

viii) These are the quantiles of level 0.95, 0.975 and 0.995 of the standard Normal distribution.

b) i) $\mathbb{P}(2 < X < 4) = \mathbb{P}(X < 4) - \mathbb{P}(X \leq 2)$

```
>> normcdf(4,3,2)-normcdf(2,3,2)
ans =
0.3829
```

ii) $\mathbb{P}(2 \leq X \leq 4) = \mathbb{P}(2 < X < 4)$, as Normal is a continuous distribution. For that matter, the MATLAB command is the same to compute $\mathbb{P}(X \leq x)$ and $\mathbb{P}(X < x)$: `normcdf`

iii) $\mathbb{P}(X \geq 4) = 1 - \mathbb{P}(X < 4)$

```
>> 1-normcdf(4,3,2)
ans =
0.3085
```

iv) $\mathbb{P}(1 < X < 5) = \mathbb{P}(X < 5) - \mathbb{P}(X \leq 1)$

```
>> normcdf(5,3,2)-normcdf(1,3,2)
ans =
0.6827
```

$\rightsquigarrow$ same probability as in a) i); Here, with $X \sim \mathcal{N}(3, 2)$, $\mu - \sigma = 1$ and $\mu + \sigma = 5$. For all normal distributions, $\mathbb{P}(\mu - \sigma < X < \mu + \sigma) = 0.6827$ (again 68-95-99 rule). You would also have that $\mathbb{P}(-1 < X < 7)$ is equal to the probability in a) ii), and $\mathbb{P}(-3 < X < 9)$ is equal to that in a) iii) (Check!)

c) Denote X the monthly exposure to PM1.0 for a cattle selected at random. We know that $X \sim \mathcal{N}(7.1, 1.5)$

i) $\mathbb{P}(X > 9) = 1 - \mathbb{P}(X \leq 9)$

```
>> 1-normcdf(9,7.1,1.5)
ans =
0.1026
```

ii) $\mathbb{P}(3 < X < 8) = \mathbb{P}(X < 8) - \mathbb{P}(X \leq 3)$

```
>> normcdf(8,7.1,1.5)-normcdf(3,7.1,1.5)
ans =
0.7226
```

iii) Find x such that $\mathbb{P}(X > x) = 0.05$, so $\mathbb{P}(X \leq x) = 0.95$

```
>> norminv(0.95,7.1,1.5)
ans =
9.5673
```

$\rightsquigarrow$ there is a 5% chance that a cattle will be exposed to more than 9.5673 μg/m^3 on a month

Exercise 9

Experiment by yourself !

MATH2089
Numerical Methods and Statistics

Statistics – Chapter 7 – Inferences concerning a mean

SOLUTIONS

Exercise 1

a) i) We are told that $\mu = 2$, so that the considered distribution is the Exp(2)-distribution, whose density is

$$f(x) = \begin{cases} \frac{1}{2}\exp\left(-\frac{x}{2}\right) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

We have thus $\mu = 2$. We also know that for the Exp(μ)-distribution, $\mathbb{V}\text{ar}(X) = \mu^2$, so that here $\sigma^2 = 2^2 = 4$. It follows $\sigma = 2$.

ii) We know that $\bar{X}$ is a random variable with mean $\mathbb{E}(\bar{X}) = \mu = 2$, variance $\mathbb{V}\text{ar}(\bar{X}) = \frac{\sigma^2}{n} = \frac{4}{48} \simeq 0.1667$ and standard deviation $\text{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{12}} \simeq 0.2887$.
By the Central Limit Theorem, $\bar{X}$ is approximately normally distributed.

b) i) We are told that $\mu = 5$, so that the considered distribution is $\mathcal{P}(5)$, whose probability mass function is

$$p(x) = e^{-5}\,\frac{5^x}{x!}, \qquad \text{for } x = 0, 1, 2, \ldots$$

We know $\mu = 5$, we also know that for the $\mathcal{P}(\lambda)$-distribution $\mathbb{V}\text{ar}(X) = \lambda$, so that here $\sigma^2 = 5$, and consequently $\sigma = \sqrt{5} \simeq 2.2361$.

ii) We know that $\bar{X}$ is a random variable with mean $\mathbb{E}(\bar{X}) = \mu = 5$, variance $\mathbb{V}\text{ar}(\bar{X}) = \frac{\sigma^2}{n} = \frac{5}{40} = 0.125$ and standard deviation $\text{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{5}}{\sqrt{10}} = 0.3536$.
By the Central Limit Theorem, $\bar{X}$ is approximately normally distributed. (n is small, so we might wonder if the CLT will work well here – maybe check for yourself on Matlab!)

c) i) The density function for the $U_{[2,5]}$-distribution is

$$f(x) = \begin{cases} \frac{1}{3} & \text{if } x \in [2,5] \\ 0 & \text{otherwise} \end{cases}$$

For the $U_{[\alpha,\beta]}$-distribution, we know that $\mathbb{E}(X) = \mu = \frac{\alpha+\beta}{2}$ and $\mathbb{V}\text{ar}(X) = \sigma^2 = \frac{(\beta-\alpha)^2}{12}$. With $\alpha = 2$ and $\beta = 5$, we directly get

$$\mu = 3.5 \qquad \text{and} \qquad \sigma^2 = 0.75$$

Hence, $\sigma = \sqrt{0.75} = 0.866$.

ii) We know that $\bar{X}$ is a random variable with mean $\mathbb{E}(\bar{X}) = \mu = 3.5$, variance $\mathbb{V}\text{ar}(\bar{X}) = \frac{\sigma^2}{n} = \frac{0.75}{15} = 0.05$ and standard deviation $\text{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.866}{\sqrt{15}} = 0.2236$.
By the Central Limit Theorem, $\bar{X}$ is approximately normally distributed (n is small, so we might wonder if the CLT will work well here – try for yourself on the app from last week!).

Exercise 2

a) Denote X the tensile strength. Then, $X \sim \mathcal{N}(75.5, 3.5)$. With samples of size $n = 6$, the sample mean is

$$\bar{X} \sim \mathcal{N}\left(75.5, \frac{3.5}{\sqrt{6}}\right)$$

Thus,

$$\mathbb{P}(\bar{X} > 75.75) = \mathbb{P}\left(Z > \frac{75.75 - 75.5}{3.5/\sqrt{6}}\right) = \mathbb{P}(Z > 0.1750) = 1 - \Phi(0.1750) = 0.4305$$

If the selected samples have size $n = 49$, then the standard deviation of the sample mean becomes $3.5/\sqrt{49} = 0.5$, and its distribution

$$\bar{X} \sim \mathcal{N}(75.5, 0.5)$$

In this situation,

$$\mathbb{P}(\bar{X} > 75.75) = \mathbb{P}\left(Z > \frac{75.75 - 75.5}{0.5}\right) = \mathbb{P}(Z > 0.5) = 1 - \Phi(0.5) = 0.3085$$

$\rightsquigarrow$ the probability of finding $\bar{X}$ 'far' from $\mu = 75.5$ decreases as the sample size increases

b) Denote X the compressive strength of concrete. Then, $X \sim \mathcal{N}(2500, 50)$. With samples of size $n = 5$ specimens, the sample mean is

$$\bar{X} \sim \mathcal{N}\left(2500, \frac{50}{\sqrt{5}}\right)$$

Thus,

$$\begin{aligned}\mathbb{P}(2499 < \bar{X} < 2510) &= \mathbb{P}\left(\frac{2499 - 2500}{50/\sqrt{5}} < Z < \frac{2510 - 2500}{50/\sqrt{5}}\right) = \mathbb{P}(-0.047 < Z < 0.4472) \\ &\simeq \Phi(0.45) - \Phi(-0.05) = 0.6736 - 0.4801 = 0.1935\end{aligned}$$

c) Denote X the waiting time of a given customer. We know that $\mu = 8.2$ min and $\sigma = 1.5$ min, so that for samples of sizes $n = 49$, we will have $\mathbb{E}(\bar{X}) = 8.2$ min and $\text{sd}(\bar{X}) = \frac{1.5}{\sqrt{49}} = 0.2143$ min. We do not know the exact distribution of X, and so not the exact distribution of the sample mean either. However, as the considered sample is 'large' ($n = 49$), we can use the Central Limit Theorem which asserts that the standardised distribution of $\bar{X}$ should be approximately close to the standard normal distribution:

$$Z = \sqrt{49}\,\frac{\bar{X} - 8.2}{1.5} = \frac{\bar{X} - 8.2}{0.2143} \overset{a}{\sim} \mathcal{N}(0, 1)$$

We can use this result to find approximate values for the required probabilities:

i) $\mathbb{P}(\bar{X} < 10) \overset{\text{CLT}}{\simeq} \mathbb{P}\left(Z < \frac{10-8.2}{0.2143}\right) = \mathbb{P}(Z < 8.39) \simeq 1$

ii) $\mathbb{P}(7 < \bar{X} < 10) \overset{\text{CLT}}{\simeq} \mathbb{P}\left(\frac{7-8.2}{0.2143} < Z < \frac{10-8.2}{0.2143}\right) = \mathbb{P}(-5.6 < Z < 8.39) \simeq 1$

iii) $\mathbb{P}(\bar{X} < 8.5) \overset{\text{CLT}}{\simeq} \mathbb{P}\left(Z < \frac{8.5-8.2}{0.2143}\right) = \mathbb{P}(Z < 1.40) = 0.9192$

Exercise 3

a) We have:

$$\mu = \mathbb{E}(X) = 0 \times 0.8 + 1 \times 0.1 + 2 \times 0.05 + 3 \times 0.05 = 0.35 \text{ (flaws)}$$

and

$$\mathbb{E}(X^2) = 0^2 \times 0.8 + 1^2 \times 0.1 + 2^2 \times 0.05 + 3^2 \times 0.05 = 0.75 \text{ (flaws}^2\text{)},$$

which yields

$$\sigma^2 = \mathbb{V}\text{ar}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 0.75 - 0.35^2 = 0.6275 \text{ (flaws}^2\text{)}$$

The standard deviation is thus

$$\sigma = \sqrt{0.6275} \approx 0.7921 \text{ (flaws)}$$

b) From random samples of size $n = 64$, the sample mean $\bar{X}$ is a random variable with mean $\mathbb{E}(\bar{X}) = \mu = 0.35$ and standard deviation $\text{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.7921}{8} \approx 0.099$

c) Samples of size $n = 64$ are certainly large enough for the Central Limit Theorem to give reliable normal approximations for the distribution of $\bar{X}$ (even if the initial distribution of X is purely discrete). Here,

$$Z = \sqrt{64}\,\frac{\bar{X} - 0.35}{0.7921} = \frac{\bar{X} - 0.35}{0.099} \overset{a}{\sim} \mathcal{N}(0,1)$$

so that

$$\mathbb{P}(\bar{X} > 0.5) \overset{\text{CLT}}{\simeq} \mathbb{P}\left(Z > \frac{0.5 - 0.35}{0.099}\right) = \mathbb{P}(Z > 1.52) = 0.0643$$

Exercise 4

a) The required confidence interval will be a two-sided t-confidence interval given by

$$\left[\bar{x} \pm t_{1-\alpha/2,n-1}\,\frac{s}{\sqrt{n}}\right]$$

With $100 \times (1-\alpha) = 99\%$, we need $\alpha = 0.01$ and $t_{0.995,10} = 3.169$. Then, the confidence interval is

$$\left[13.77 \pm 3.169\,\frac{1.15}{\sqrt{11}}\right] = [12.67, 14.87]$$

$\rightsquigarrow$ we can be 99% confident that the true mean temperature for wheat grown at that place in June is between 12.67 °C and 14.87 °C

b) We assumed *independence* – that the observations are independent of each other (and with equal mean and variance), which would be valid if we took temperature measurements at a random sample of times/places.

We assumed *normality* – that observations are normally distributed. We could try to check using a quantile plot, looking in particular for signs of skew or long tails.

c) We need a one-sided t-confidence interval like

$$\left[\bar{x} - t_{1-\alpha,n-1}\,\frac{s}{\sqrt{n}}, +\infty\right)$$

here $\alpha = 0.05$ (95% confidence level required) and $t_{10,0.95} = 1.812$, so that the confidence interval is

$$\left[13.77 - 1.812\,\frac{1.15}{\sqrt{11}}, +\infty\right) = [13.14, +\infty)$$

$\rightsquigarrow$ we can be 95% confident that the true mean temperature for wheat grown at that place in June is not below 13.14 °C.

d) We can fix $e = 0.2$ °C and use the sample size formula on Slide 276, if we treat the sample standard deviation s in place of the true standard deviation, and use the standard normal instead of t (which is OK as an approximation unless n works out to be small):

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{e}\right)^2 = \left(\frac{1.96 \times 1.15}{0.4}\right)^2 = 31.75$$

(note that we use $z_{0.975} = 1.96$).
$\rightsquigarrow$ we need about 32 observations to reach the required level of accuracy.

Exercise 5

We have observed a sample of size $n = 25$, with $\bar{x} = 4.05$ mm and $s = 0.08$ mm. As we can assume the normal distribution for the population of thickness and σ is not known, we need a one-sided t-confidence interval given by

$$\left[\bar{x} - t_{n-1,1-\alpha}\,\frac{s}{\sqrt{n}}, +\infty\right)$$

MATLAB says that $t_{24,0.95} = 1.7109$, so the interval is

$$\left[4.05 - 1.7109\,\frac{0.08}{\sqrt{25}}, +\infty\right) = [4.0226, +\infty]$$

$\rightsquigarrow$ from what we have observed, we can be 95% confident that the true mean wall thickness is larger than 4.0226 mm.

Exercise 6

a) We have observed a sample of size $n = 6$, with $\bar{x} = 26.4$ bpm and $s = 14.28$ bpm. A 95% confidence interval for μ is

$$\left[26.4 \pm 2.57\,\frac{14.28}{\sqrt{6}}\right] = [11.41, 41.39]$$

$\rightsquigarrow$ from what we have observed, we can be approximately 95% confident that the true mean increase in the pulse rate is between 11.41 bpm and 41.39 bpm when astronaut trainees are performing that task.

b) We can be (approximately) 95% confident that the maximum error is $2.57\,\frac{14.28}{\sqrt{6}} = 14.99$ bpm.

c) We assumed *independence* – that the observations are independent of each other (and with equal mean and variance), which would be valid if we took a random sample.

We assumed *normality* – that observations are normally distributed. We could check using a quantile plot but for $n = 6$ we can't tell much, and the CLT doesn't provide us a huge amount of protection against this assumption, so if data were skewed or long-tailed we would be in a bit of strife.

Exercise 7

Elementary computations yield:

$$\bar{x} = \frac{1}{10}\,(25.2 + \ldots + 19.5) = 21.90$$

and

$$s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)} = \ldots = 4.134$$

As the population is assumed to be normal and σ is unknown, the required confidence interval will be a two-sided t-confidence interval given by

$$\left[\bar{x} \pm t_{n-1,1-\alpha/2}\,\frac{s}{\sqrt{n}}\right]$$

MATLAB says $t_{9,0.975} = 2.2622$, so that the interval is

$$\left[21.9 \pm 2.2622\,\frac{4.134}{\sqrt{10}}\right] = [18.94, 24.86]$$

$\rightsquigarrow$ we can be 95% confident that the true mean fat content for the sausages of that brand is between 18.94 and 24.86 percent.

Exercise 8

A point estimate for π is the observed sample proportion, which is

$$\hat{p} = \frac{16}{48} = \frac{1}{3} \simeq 0.3333$$

Now, according to Slide 319, we can build a large sample confidence interval for π if the condition

$$n\hat{p}(1 - \hat{p}) > 5$$

is fulfilled (empirical rule). Here we have

$$48 \times \frac{1}{3} \times \frac{2}{3} = \frac{96}{9} = 10.6667,$$

so n is large enough and the true π supposedly sufficiently different to 0 or 1 for the large sample confidence interval to be reliable. This one is

$$\left[\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

which, with our data, is

$$\left[\frac{1}{3} \pm 1.96\sqrt{\frac{\frac{1}{3}(1-\frac{1}{3})}{48}}\right] = [0.2000, 0.4667]$$

$\leadsto$ we can be 95% confident that the true probability π of such a trial resulting in ignition lies between 0.2000 and 0.4667

This interval is quite wide, because a sample size of 48 is not at all large when estimating a proportion/probability.

Exercise 9

a) A p-value of 0.008 is very very small. At any usual significance level, like $\alpha = 0.10$, $\alpha = 0.05$ and $\alpha = 0.01$, we find $p < \alpha$. So there is strong evidence to reject H_0 and favour H_a.

b) A p-value of 0.08 is not really very small. However, if we test H_0 at level $\alpha = 0.10$ (that is, tolerating a 10% chance of error when rejecting H_0), we come to the conclusion that H_0 is not supported enough by the data. On the other hand, if we test H_0 at level $\alpha = 0.05$ and $\alpha = 0.01$ (or essentially, at any level α smaller than 0.08), it becomes too risky to reject H_0, so we don't.

c) A p-value of 0.80 is not at all small. It is not meaningful to carry out a test at level $\alpha > 0.8$ (why?), so in such a situation we will never reject H_0 : there is no evidence enough to claim that H_0 is not true, what the data show is consistent with H_0 (this does not mean that H_0 is automatically true!)

Exercise 10

a) We are given the observed sample proportion $\hat{p} = 36/165 = 0.2182$. Let π be the true proportion of cylinder bores which are outside the specifications. We want to test :

$$H_0 : \ \pi = \pi_0 = 0.10, \qquad \text{vs} \qquad H_a : \ \pi > \pi_0 = 0.10.$$

(one-sided alternative, for if π is actually lower than 0.10, that's even better)

From the sampling distribution

$$\sqrt{n}\,\frac{\hat{p}-\pi}{\sqrt{\pi(1-\pi)}} \overset{a}{\sim} N(0,1)$$

we can derive the (one-sided) rejection criterion at significance level $\alpha = 0.01$:

$$\text{reject } H_0 \text{ if } \hat{p} > \pi_0 + z_{1-\alpha}\sqrt{\frac{\pi_0(1-\pi_0)}{n}} = 0.10 + 2.33 \times \sqrt{\frac{0.10 \times 0.90}{165}} = 0.154$$

$\leadsto$ we **reject** H_0, as we have observed $\hat{p} = 0.2182$

The observed value of the test statistic is

$$z_0 = \sqrt{165}\,\frac{0.2182 - 0.10}{\sqrt{0.1 \times 0.9}} = 5.06$$

so that the p-value is

$$P(Z > z_0) = 1 - \Phi(5.06) \simeq 0$$

$\leadsto$ we reject H_0 and conclude that the true proportion of cylinder bores outside specification is higher than 10% (and this, with a very low chance of being wrong)

b) A 99% two-sided confidence interval for π is

$$\left[\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = \left[0.2182 \pm 2.575\sqrt{\frac{0.2182 \times 0.7818}{165}}\right] = [0.1354, 0.3010].$$

$\leadsto$ the 'target' value 0.10 is indeed outside this interval of plausible values for π

c) We need to assume that the sample is a random sample — this is given in the question, and that the central limit theorem is applicable under H_0 — since the sample size $n = 165$ satisfies both $n\pi_0(1-\pi_0) > 5$ and $n\hat{p}(1-\hat{p})$, we have some confidence about this assumption.

Exercise 11

a) We have to test

$$H_0 : \mu = 1 \qquad \text{(no change in the true mean)}$$

against

$$H_a : \mu > 1 \qquad \text{(the true mean has increased)}$$

As we assume the normal distribution for the population and $\sigma = 0.06$ is known, we will use a one-sided z-test, whose rejection criterion is given by:

$$\text{reject } H_0 \text{ if } \bar{x} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

Here, we set the significance level to $\alpha = 0.05$ (so $z_{1-\alpha} = z_{0.95} = 1.645$), $\mu_0 = 1$ (value of μ under the null hypothesis) and $n = 4$. We get :

$$\text{reject } H_0 \text{ if } \bar{x} > 1 + 1.645 \times \frac{0.06}{\sqrt{4}} = 1.0494$$

As we have observed a sample mean $\bar{x} = 1.134$ micron, we come to the conclusion that H_0 has to be rejected (at significance level 5%). The observed value of the test statistic is

$$z_0 = \sqrt{n}\,\frac{\bar{x} - \mu_0}{\sigma} = \sqrt{4} \times \frac{1.134 - 1}{0.06} = 4.46$$

so that the p-value is

$$p = 1 - \Phi(4.46) \simeq 0$$

$\rightsquigarrow$ we are almost certain that the mean layer thickness has effectively increased

b) At level 95%, the one-sided z-confidence interval for the true layer thickness mean μ is given by

$$\left[\bar{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, +\infty\right) = \left[1.134 - 1.645 \times \frac{0.06}{\sqrt{4}}, +\infty\right) = [1.0846, +\infty)$$

$\rightsquigarrow$ we are 95% confident that the current mean thickness is larger than 1.0846 micron. The previous value $\mu = 1$ micron is smaller than this bound, so we are 95% confident that μ has increased, in agreement with the result of the hypothesis test in a)

c) The rejection criterion we set up in a) was "reject H_0 when $\bar{x} > 1.0494$". Thus, if $\mu = 1.01$, we would (rightly!) reject H_0 with probability

$$1 - \beta = \mathbb{P}(\bar{X} > 1.0494 \text{ when } \mu = 1.01) = \mathbb{P}\left(Z > \sqrt{4}\,\frac{1.0494 - 1.01}{0.06}\right) = \mathbb{P}(Z > 1.31) = 0.0951$$

$\rightsquigarrow$ low probability

$\rightsquigarrow$ with such a small sample ($n = 4$), the test really struggles to distinguish $H_a : \mu = 1.01$ from $H_0 : \mu = 1$ (which we can understand, the two values being very close)

$\rightsquigarrow$ to guarantee that the type I error probability α is at most 0.05, we must accept that the type II error probability is as high as $\beta = 0.9049$

d) With a sample of size n, the rejection criterion becomes:

$$\text{reject } H_0 \text{ if } \bar{x} > 1 + 1.645 \times \frac{0.06}{\sqrt{n}}$$

The power of such a test is

$$\begin{aligned}1-\beta &= \mathbb{P}\left(\bar{X} > 1 + 1.645 \times \frac{0.06}{\sqrt{n}} \text{ when } \mu = 1.01\right)\\ &= \mathbb{P}\left(Z > \sqrt{n}\,\frac{1 + 1.645 \times \frac{0.06}{\sqrt{n}} - 1.01}{0.06}\right)\\ &= \mathbb{P}\left(Z > 1.645 - \frac{0.01\sqrt{n}}{0.06}\right)\end{aligned}$$

If we want this probability to be equal to 0.85, we need

$$1.645 - \frac{0.01\sqrt{n}}{0.06} = z_{0.15} = -1.04$$

Solving for n, we find

$$n = 259.53$$

$\rightsquigarrow$ we need at least 260 observations (far from our initial 4 observations!) to have a power of 0.85 at detecting an increase of μ to 1.01

Exercise 12

a) i) The mean stays the same
   ii) The variance gets smaller (inversely proportionally to increases in sample size)
   iii) The distribution gets closer to normal (bell-shaped)

b) To do this you need to maximise skew on the distribution being sampled from, by having a big mass of points at one end and a small dot right at the other.

Exercise 13

a) Simulate 500 independent random samples of size $n = 100$ from the Bern(0.5) distribution:

```
>> B=binornd(1,0.5,100,500);
```

Each column of `B` is one of the 500 independent random samples of size 100.

b) Vector of 500 simulated sample means, that is, a sample of size 500 drawn from the distribution of the random variable $\bar{X}$:

```
>> meanB=mean(B);
```

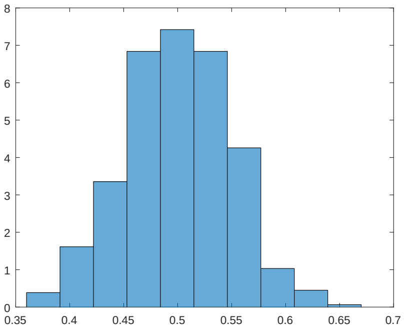
c) We know that $\mathbb{E}(\bar{X}) = \mu$ and $\mathbb{V}\text{ar}(\bar{X}) = \frac{\sigma^2}{n}$, where μ and σ^2 are the population mean and variance, respectively. Here, we know (because we *simulate*) that the population distribution is Bern($\pi = 0.5$), whose mean is $\pi = 0.5$ and variance is $\pi(1-\pi) = 0.5 \times 0.5 = 0.25$. Thus, we expect the sample mean and sample variance of the sample `meanB` to be close to $\pi = 0.5$ and $\frac{\pi(1-\pi)}{n} = \frac{0.25}{100} = 0.0025$. Indeed,

```
>> mean(meanB)
  ans =
  0.4973
  >> var(meanB)
  ans =
  0.0023
```

Note that the random samples that you simulated in your matrix `B` are not the same as those simulated here - that is why they are *random* samples - so your values `mean(meanB)` and `var(meanB)` might be different to these ones. However, they should always be close to 0.5 and 0.0025, as stated by the theory.

d) Here, we have $n = 100$, which is large enough for the Central Limit Theorem to guarantee that the sampling distribution of $\bar{X}$ should be pretty much like a normal distribution. Using our function `denhist` defined in Lab class 4, with 10 classes $(= \sqrt{100})$, we get:

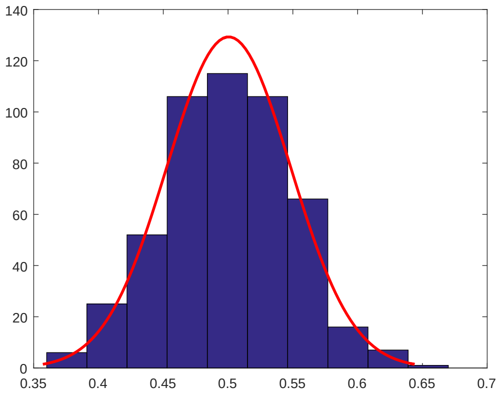
```
>> histogram(meanB,10,'Normalization','pdf')
```



Indeed, the histogram clearly shows a symmetric bell-shaped appearance.

e) We overlay the histogram with the normal density.

```
>> histfit(meanB,10,'normal')
```



The match is clear.

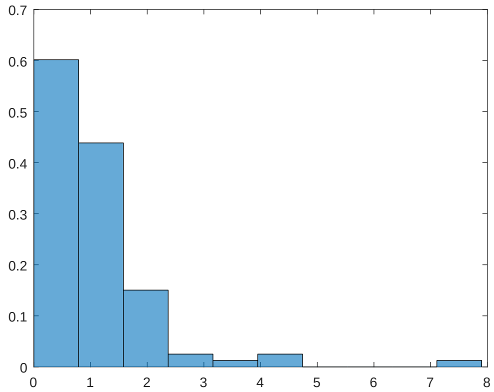
f) Write an `m-file`.

Exercise 14

a) i) Density histogram for the `kevlar90` data set:

```
>> histogram(kevlar90,10,'Normalization','pdf')
```

Note that we ask for 10 classes as $n = 101$, so that $\sqrt{n} \simeq 10$. We get:



The histogram shows a very skewed distribution, with a long right tail. The Normal distribution is therefore not at all appropriate to model this sample. On the other hand, the Exponential distribution shape should match quite well the histogram shape.

ii) Sample mean of the `kevlar90` data set:

```
>> mean(kevlar90)
ans =
1.0239
```

⇝ $\bar{x} = 1.0239$ hour.

iii) As $\bar{x} = 1.0239$, we find an estimate for μ:

$$\hat{\mu} = \bar{x} = 1.0239$$

The fitted Exponential distribution is thus Exp(1.0239), whose density is

$$f(x) = \frac{1}{1.0239} e^{-\frac{x}{1.0239}} = 0.9767 \times e^{-0.9767x}$$

for $x \geq 0$ (and 0 elsewhere). This yields an estimated standard deviation for the failure times equal to

$$\hat{\sigma} = \hat{\mu} = 1.0239.$$

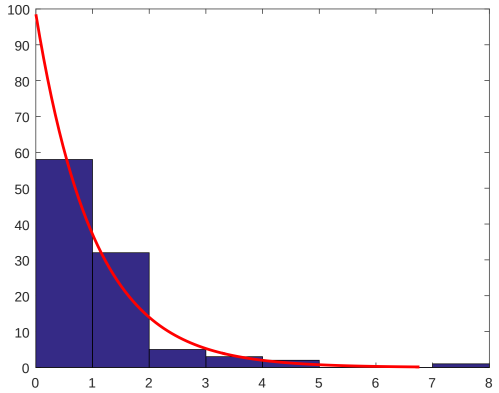
The sample standard deviation of the `kevlar90` data set set is:

```
>> std(kevlar90)
ans =
1.1173
```

⇝ $s = 1.1173$ hour, not too far away from $\hat{\sigma}$. Note that we only *assume* that the Exponential distribution is the population distribution (as it was suggested by the histogram), so there was absolutely no guarantee of finding $s \simeq \hat{\sigma}$. Precisely, finding s and $\hat{\sigma}$ considerably different would have been evidence that the Exponential distribution *is not* the population distribution. If we do believe that the Exponential distribution is the population distribution, then $\hat{\sigma} = 1.0239$ and $s = 1.1173$ are two different estimates (from two different estimators) of the same unknown quantity: the population standard deviation σ.

iv) We overlay the exponential density on the histogram of kevlar.

```
>>histfit(kevlar90,8,'exponential')
```



The agreement is reasonable. There may be too few observations in the first class and too many in the second, but this is probably not enough evidence against the Exponential model.

v) Write an `m-file`.

b) i) Simulate 1000 independent random samples of size $n = 10$ from the Exp(1.0239) distribution:

```
>> T=exprnd(mean(kevlar90),10,1000);
```

Each column of `T` is one of the 1000 independent random samples of size 10. **Careful!** In this somulation we use as the population distribution (that is the one we simulate from) the Exp($\hat{\mu}$)-distribution, so that the population mean is given by $\hat{\mu} = \bar{x}$ (the sample mean from part a)). This could be a bit confusing.

ii) Vector of 1000 simulated sample means, that is, a random sample of size 1000 drawn from the distribution of the random variable $\bar{X}$:

```
>> meanT=mean(T);
```

Note that, when applied to a matrix, the MATLAB function `mean` computes the mean of each column and returns a vector.

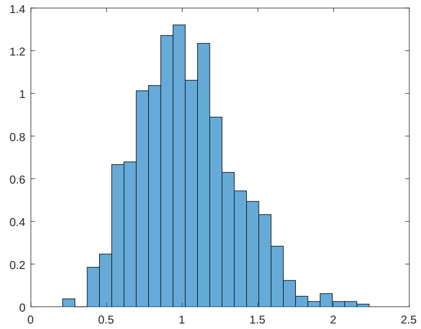
iii) We know that $\mathbb{E}(\bar{X}) = \mu$ and $\mathbb{V}\text{ar}(\bar{X}) = \frac{\sigma^2}{n}$, where μ and σ^2 are the population mean and variance, respectively. Here, we know (because we *simulate*) that the population distribution is Exp(1.0239), whose mean is 1.0239 and variance is $1.0239^2 = 1.0483$. Thus, we expect the sample mean and sample variance of the sample `meanT` to be close to 1.0239 and $\frac{1.0483}{10} = 0.1048$. Indeed,

```
>> mean(meanT)
ans =
1.0201
>> var(meanT)
ans =
0.1049
```

Note that the random samples that you simulated in your matrix `T` are not the same as those simulated here - that is why they are *random* samples - so that your results might be different to these ones. However, they should always be close to 1.0239 and 0.1048, as stated by the theory.

iv) Here, we have $n = 10$, which may not be large enough for the Central Limit Theorem to guarantee that the sampling distribution of $\bar{X}$ is approximately normal. However, we can check. Using our function `denhist` defined in Lab class 4, with 25 classes (say), we get:

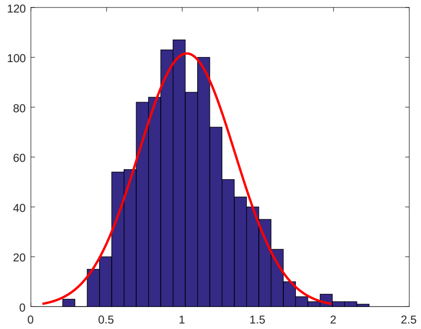
```
>> histogram(meanT,25,'Normalization','pdf')
```



The initial skewness of the Exponential distribution is still clearly visible. This said, the histogram already starts getting a bell-shape appearance.

v) The Central Limit Theorem states that, if n was larger, the sampling distribution of $\bar{X}$ would be the normal distribution with mean 1.0239 and variance 0.1048 (and so standard deviation $\sqrt{0.1048} = 0.3237$), as found in iii). Thus, we overlay the normal density using the `histfit` function:

```
>> histfit(meanT,25,'normal')
```



Despite the skewness of the histogram (due to the small value of n), the match is evident enough and the normal approximation is not too bad.

c) i) Simulate 1000 independent random samples of size $n = 500$ from the Exp(0.9767) distribution:

```
>> T=exprnd(mean(kevlar90),500,1000);
```

ii) Vector of 1000 simulated sample means, that is, a random sample of size 1000 drawn from the distribution of the random variable $\bar{X}$:

```
>> meanT=mean(T);
```

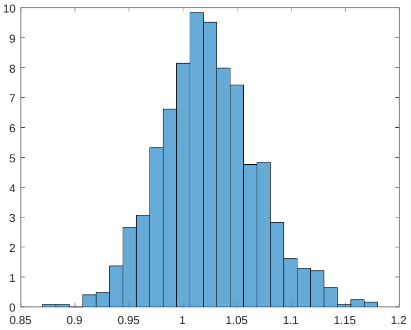
iii) We know that $\mathbb{E}(\bar{X}) = \mu$ and $\mathbb{V}\mathrm{ar}(\bar{X}) = \frac{\sigma^2}{n}$, where μ and σ^2 are the population mean and variance, respectively. Here, we know (because we *simulate*) that the population distribution is Exp(1.0239), whose mean is 1.0239 and variance is $1.0239^2 = 1.0483$. Thus, we expect the sample mean and sample variance of the sample `meanT` to be close to 1.0239 and $\frac{1.0483}{500} = 0.0021$. Indeed,

```
>> mean(meanT)
ans =
1.0229
>> var(meanT)
ans =
0.0021
```

Your results might be different to these ones, however they should always be close to 1.0239 and 0.1048, as stated by the theory. Note that the mean of $\bar{X}$ is unaffected by the sample size (and remains equal to μ, in theory), whereas the variance of $\bar{X}$ is much smaller than in b). This means that, for large n, the distribution of $\bar{X}$ is more concentrated around μ. In other words, $\bar{X}$ is more and more accurate as an estimator for μ when n gets larger and larger ('consistency' of the estimator).

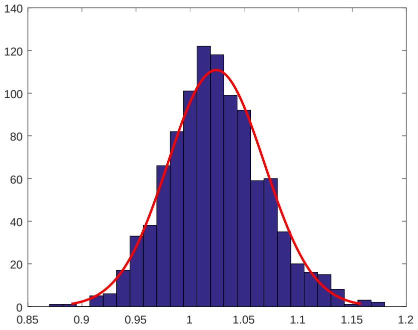
iv) With $n = 500$, the Central Limit Theorem most certainly applies and the sampling distribution of $\bar{X}$ should be approximately normal:

```
>>histogram(meanT,25,'Normalization','pdf')
```



This histogram is mostly bell-shaped (although a fair skewness is still visible). Also, compare the scale of the x-axis to the one in part b).

v) The Central Limit Theorem states that the sampling distribution of $\bar{X}$ should be the normal, so we overlay the normal density and we get:

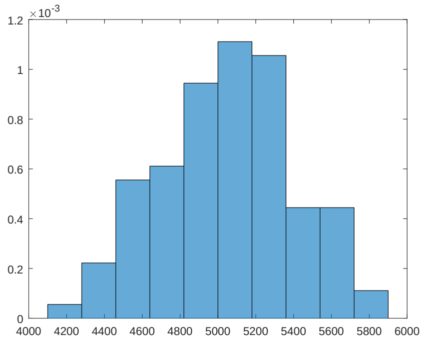


The normal approximation is now excellent.

Exercise 15

a) The `shearstrength` data set contain $n = 100$ observations. Using our function `denhist` defined in Lab class 4, with 10 classes ($= \sqrt{n}$), we get:

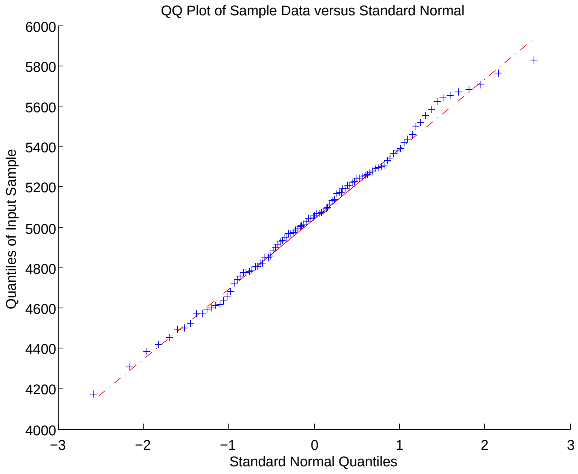
```
>>histogram(shearstrength,10,'Normalization','pdf')
```



The histogram shows an almost symmetric and bell-shaped appearance. However, there seems to be a clear cut around 5300 lb, so it is not clear whether or not the normal distribution would fit well.

b) A normal quantile plot will be more informative about the normality assumption (it is actually tailored for the job!):

```
>> qqplot(shearstrength)
```



Except a small bump starting at around 5300 lb (of course!), the dots follow very well the linear pattern. It seems clear that the normality assumption is reasonable. Note that MATLAB shows a normal quantile plot with interchanged axes compared with the plots we showed in class (here the x-axis is for the theoretical quantiles, and the y-axis is for the observed sample quantiles). The conclusion obviously remains the same: the dots must fall along a straight line to validate the normality assumption.

c) The true mean μ is well estimated by the sample mean (unbiased and efficient estimator for μ), which is:

```
>> mean(shearstrength)
   ans =
   5.0492e+03
```

$\rightsquigarrow$ $\bar{x} = 5049.2$ lb, which should be a good estimate of μ.

A 95% confidence interval is given by:

```
>> [h,p,ci]=ztest(shearstrength,mean(shearstrength),350);ci
   ci =
   1.0e+03 *
   4.9806 5.1178
```

$\rightsquigarrow$ we are 95% confident that the true value of μ lies in $[4980.6, 5117.8]$ (if the true value of σ is indeed 350 lb, obviously). Here the z-confidence interval is the one to be used as we assume the normal distribution for the population and σ is known.

d) If σ is not known, we have to estimate it by the sample standard deviation s. Here, the estimated standard deviation is:

```
>> std(shearstrength)
  ans =
  351.4525
```

This estimation brings some more variability into the procedure that needs to be taken into account. That is what a t-confidence interval achieves:

```
>> [h,p,ci]=ttest(shearstrength,mean(shearstrength));ci
  ci =
  1.0e+03 *
  4.9794 5.1189
```

$\rightsquigarrow$ in this situation (without assuming σ is known), we are 95% confident that the true value μ lies in $[4979.4, 5118.9]$. This interval is indeed slightly wider than the previous one, but not that much. That is because n is large, so that the Student distribution with many degrees of freedom (here: $n-1 = 99$ degrees of freedom) is pretty much like the standard normal distribution. Another reason for this interval to be wider than the one found in c) is that the estimated standard deviation $s = 351.4525$ is also slightly larger than the assumed $\sigma = 350$ in c).

Exercise 16

a) Simulate a matrix C with 36 rows and 500 columns containing elements that are independent values drawn from the Normal distribution with $\mu = 20$ and $\sigma = 5$:

```
>> C=normrnd(20,5,36,500);
```

Each column of C is one of the 500 independent random samples of size 36.

b) i) Compute the means of each of the 500 samples:

```
>> meanC=mean(C);
```

ii) From the given expression of the z-confidence interval, we easily compute the vector of upper bounds upz:

```
>> upz=meanC+1.96*5/6;
```

iii) From the given expression of the z-confidence interval, we easily compute the vector of lower bounds lowz:

```
>> lowz=meanC-1.96*5/6;
```

c) Compute how many of the 500 z-confidence intervals contain the true population mean $\mu = 20$:

```
>> sum((upz>=20).*(lowz<=20))
  ans =
  477
```

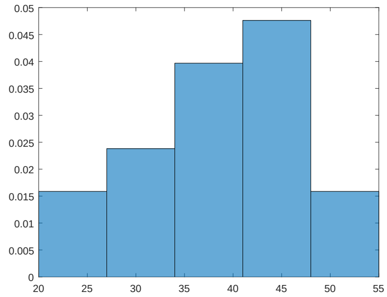
$\rightsquigarrow$ 477 out of the 500 computed 95%-confidence intervals (that is a proportion of $477/500 = 0.954$) contain the true value of μ.
Repeat with the 500 t-confidence intervals.

d) If we computed ‘infinitely many’ such confidence intervals, exactly 95% of them would contain the true value of μ (definition of the level of a confidence interval).

Exercise 17

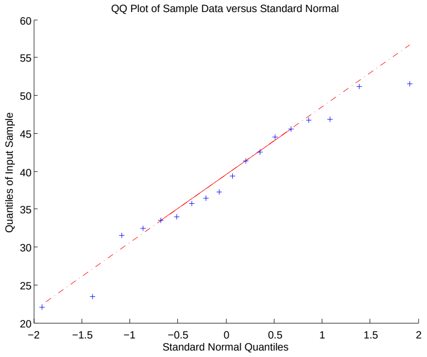
a) The data set contains $n = 18$ observations. We represent a density histogram with 5 classes:

```
>> histogram(porevolume,5,'Normalization','pdf')
```



It is not clear that the underlying distribution is well approximated by a normal distribution. However, there is no strong evidence against that assumption either! This is due to the smallish sample size (not much information to draw definite conclusion). The normal quantile plot is not really more informative:

```
>> qqplot(porevolume)
```



As there is no clear evidence against the normality assumption, we can suppose it holds (but we have to be careful!).

b) We will assume a normal population with σ known to be 8.5. At level 98% (thus with $\alpha = 0.02$), we get:

```
>> [h,p,ci]=ztest(porevolume,mean(porevolume),8.5,0.02);ci
   ci =
   34.0003 43.3219
```

$\rightsquigarrow$ we are 98% confident that the true mean value μ of residual gas saturation is in $[34.0, 43.3]$.

c) At level 95%, we get

```
>> [h,p,ci]=ztest(porevolume,mean(porevolume),8.5);ci
   ci =
   34.7344 42.5878 ;
```

$\rightsquigarrow$ we are 95% confident that the true value μ is in $[34.7, 42.6]$. This interval is shorter than the previous, as we require a smaller degree of confidence (so we gain in precision, but we lose in certainty).

MATH2089
Numerical Methods and Statistics

Statistics – Chapter 9 – Inferences concerning a difference of means

SOLUTIONS

Exercise 1

We have **paired observations**: the same 8 specimens are subjected to both methods. We have just to work with the sample of differences, whose mean is $\bar{d} = -0.4137$ (mg/l) and standard deviation $s_d = 0.3210$ (mg/l), as given.

a) As we can assume normality for the population of differences, we have to construct a 99% t-confidence interval. Matlab says $t_{7,0.995} = 3.4995$, so we have :

$$\left[\bar{d} \pm t_{n-1,1-\alpha/2}\frac{s_d}{\sqrt{n}}\right] = \left[-0.4137 \pm 3.4995 \times \frac{0.3210}{\sqrt{8}}\right] = [-0.8109, -0.0165]$$

$\rightsquigarrow$ we are 99% confident that the true average reading from the SIB method exceeds that from the MIB method by between 0.0165 and 0.8109 mg/l

b) We want to test :

$$H_0 : \mu_D = 0 \qquad \text{vs} \qquad H_a : \ \mu_D \neq 0.$$

It is a two-sided one-sample t-test. Matlab says $t_{7,0.975} = 2.3646$ so that the rejection criterion is given by :

$$\text{reject } H_0 \text{ if } \bar{d} \notin \left[0 \pm t_{n-1,1-\alpha/2}\frac{s_d}{\sqrt{n}}\right] = \left[\pm 2.3646 \times \frac{0.3210}{\sqrt{8}}\right] = [-0.2684, 0.2684]$$

$\rightsquigarrow$ as we observed $\bar{d} = -0.4137$, we reject H_0

Note that the previous confidence interval indicates that we would have rejected H_0 at level $\alpha = 0.01$ as well (the 99% confidence interval does not contain 0). The observed value of the test statistic is

$$t_0 = \sqrt{n}\,\frac{\bar{d}}{s_d} = \sqrt{8} \times \frac{-0.4137}{0.3210} = -3.75$$

so that the p-value is

$$p = 2 \times \mathbb{P}(T > |t_0|) = 2 \times \mathbb{P}(T > 3.75)$$

for $T \sim t_7$. Matlab says : $p = 0.007$. $\rightsquigarrow$ clearly reject H_0, the two methods yield significantly different results.

c) We need to assume that we have a random sample of specimens (and hence of 'differences'). There isn't much we can do to check this assumption. We used the t-distribution, so we also need to assume that the differences come from a normal distribution. Whether this is plausible could be checked by doing a quantile plot.

Exercise 2

a) We have here two independent samples (the observations are about different tyres from different brands), so we need a two-sample t-test. The given two sample standard deviations are 'of similar magnitude', so we assume the population variances may be equal (this could be properly tested, but this is beyond the scope of this course). We can calculate the pooled standard deviation, estimate of the 'true' common population variance σ :

$$s_p = \sqrt{\frac{44 \times 2200^2 + 44 \times 1500^2}{45 + 45 - 2}} = 1882.82.$$

As we have a two-sided alternative, the rejection criterion for $H_0 : \mu_1 = \mu_2$ (that is, $H_0 : \mu_1 - \mu_2 = 0$), is

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 \notin \left[-t_{n_1+n_2-2;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, t_{n_1+n_2-2;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$
$$= \left[\pm 1.99 \times 1882.82 \times \sqrt{\frac{1}{45} + \frac{1}{45}}\right] = [-789.90, 789.90]$$

Here we have observed $\bar{x}_1 - \bar{x}_2 = 42500 - 40400 = 2100$, so that we **reject** H_0.
Note that, because $\nu = 88$ is here 'large', we could approximate $t_{88,0.975}$ by $z_{0.975} = 1.96$ (why?). Note that MATLAB gives 1.9873 for $t_{88,0.975}$.
Now, we can compute the p-value associated with this test. The observed value of the test statistic is

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{2100}{1882.82\sqrt{\frac{1}{45} + \frac{1}{45}}} = 5.29,$$

so that the p-value is

$$p = 2 \times \mathbb{P}(T > 5.29) \simeq 8 \times 10^{-7}$$

for $T \sim t_{88}$ (given by MATLAB) $\rightsquigarrow$ very low risk when rejecting H_0, the mean tread lives of the two brands are significantly different.

Note that, if you don't believe in the assumption $\sigma_1 = \sigma_2$, you can use Welch-Satterthwaite's approximate result and compute

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} = 77.64$$

$\rightsquigarrow$ use the t_{77}-distribution instead of the t_{88}-distribution in the preceding development
This said, it can be found (MATLAB) that $t_{77,0.975} = 1.9912$ (for $t_{88,0.975} = 1.9873$), so that the difference we will observe in the final results will obviously be minor

b) We need a two-sided 95% t-confidence interval for $\mu_1 - \mu_2$:

$$\left[(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2;1-\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] = \left[2100 \pm 1.99 \times 1882.82 \times \sqrt{\frac{1}{45} + \frac{1}{45}}\right] = [1310, 2890]$$

$\rightsquigarrow$ we can be 95% confident that the difference between the true mean tread lives of the two brands is between 1310 and 2890. The confidence interval doesn't contain 0, as expected from part a). Since the interval is quite wide (the interval width 1580 is comparable to the sample mean difference 2100), the information about $\mu_1 - \mu_2$ might not be as precise as desirable $\rightsquigarrow$ need for more observations

c) We need to assume that the two samples are independent random samples and the population distributions have equal variance (or we have to use Welch-Satterthwaite's approximation). With the information given in the question, none of the assumptions can be checked. Note that the normality assumption is not crucial here, as the sample sizes $n_1 = n_2 = 45$ guarantee through the Central Limit Theorem that the results would remain (at least approximately) valid, even if the populations were nonnormal.

Exercise 3

Let μ_1 and μ_2 denote the true mean boredom proneness ratings for males and females respectively. We want to test:

$$H_0 : \mu_1 - \mu_2 = 0 \qquad \text{vs} \qquad H_a : \mu_1 - \mu_2 > 0.$$

Method 1. If we assume that the populations are normal and that the population variances are equal (the observed sample standard deviations are similar), we can do a two-sample t-test with pooled sample variance

$$s_p = \sqrt{\frac{96 \times 4.83^2 + 147 \times 4.68^2}{97 + 148 - 2}} = 4.74$$

As we have a one-sided test, the rejection criterion is :

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 > t_{n_1+n_2-2,1-\alpha}\, s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 1.65 \times 4.74 \times \sqrt{\frac{1}{97}+\frac{1}{148}} = 1.0217$$

Here, we have observed $\bar{x}_1 - \bar{x}_2 = 10.40 - 9.26 = 1.14 \rightsquigarrow$ we **reject** H_0 and conclude that mean boredom proneness rating for males is higher than that for females. MATLAB gives $t_{243,0.95} = 1.6511$ The observed value of the test statistic is

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}} = \frac{1.14}{4.74\sqrt{\frac{1}{97}+\frac{1}{148}}} = 1.84$$

and the p-value of the test is thus $p = P(T > 1.84)$, where $T \sim t_{243}$. MATLAB gives a p-value of 0.0335. Hence the p-value is less than the significance level given as $\alpha = 0.05$ (reject of H_0, as announced).

Method 2. Alternatively, if the populations are not assumed to be normal, we can still use a large sample test, arguing that the sample sizes are big enough for the Central Limit Theorem to make the job. Then, the rejection criterion is

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 > z_{1-\alpha}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = 1.645 \times \times \sqrt{\frac{4.83^2}{97}+\frac{4.68^2}{148}} = 1.0253$$

(note that large sample tests essentially assume that estimating the standard deviations does not affect the results, that is $s_1 \simeq \sigma_1$ and $s_2 \simeq \sigma_2 \rightsquigarrow$ no question of pooled standard deviation here). As we have observed $\bar{x}_1 - \bar{x}_2 = 1.14 \rightsquigarrow$ we **reject** H_0 and conclude that mean boredom proneness rating for males is higher than that for females. The observed value of the test statistic is

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{1.14}{\sqrt{\frac{4.83^2}{97}+\frac{4.68^2}{148}}} = 1.83$$

The (approximate) p-value is then given by

$$p = \mathbb{P}(Z > 1.83) = 0.0336$$

(from MATLAB) $\rightsquigarrow$ we reject H_0

Note that both p-values (Method 1 and Method 2) are very close, which somewhat validates the use of either method.

Exercise 4

We are asked to use a two-sample t-test to test $H_0 : \mu_1 - \mu_2 = \Delta_0 = -10$ versus $H_a : \mu_1 - \mu_2 < \Delta_0 = -10$. Since the two observed sample standard deviations are quite close, we can assume that the population variances are equal and work with the pooled standard deviation :

$$s_p = \sqrt{\frac{5 \times 5.03^2 + 5 \times 5.38^2}{6+6-2}} = 5.208$$

Now, the rejection criterion is like (why?) :

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 < \Delta_0 - t_{n_1+n_2-2,1-\alpha}\, s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = -10 - 2.764 \times 5.208 \times \sqrt{\frac{1}{12}} = -15.88$$

Here we have observed $\bar{x}_1 - \bar{x}_2 = 115.7 - 129.3 = -13.6 \rightsquigarrow$ we **do not reject** H_0, the observations do not show enough evidence that $\mu_2 - \mu_1$ is larger than 10 m and we may not believe the expert. The observed value of the test statistic is

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{s_p\sqrt{1/n_1 + 1/n_2}} = \frac{-13.6 - (-10)}{5.208\sqrt{1/6 + 1/6}} = -1.20,$$

which yield a p-value of

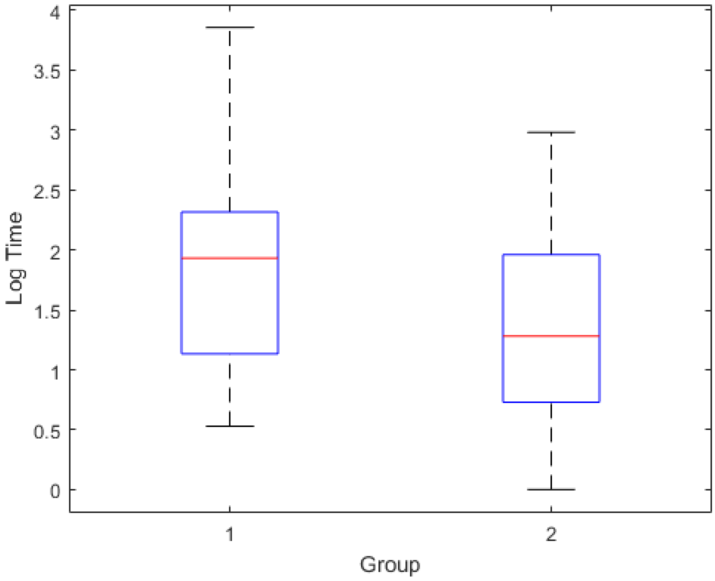
$$p = \mathbb{P}(T < -1.20) = \mathbb{P}(T > 1.20)$$

for $T \sim t_{10}$. MATLAB says : $p = 0.1289$.

$\rightsquigarrow$ too risky to reject H_0 ($p > \alpha = 0.01$).

Exercise 5

a) There appear to be some differences in the log time for the groups.

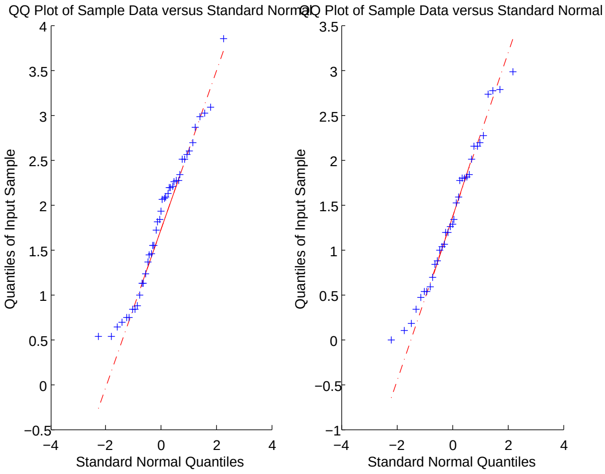


b) We have two independent samples that we would like to compare in terms of their mean. So we have to find a confidence interval for the difference in means between two independent samples. The first sample consists in $n_1 = 43$ observations, and the second sample contains $n_2 = 35$ observations. You can extract the two samples using the following code:

```
>> group=fusiondat(:,1);
>> logtime=fusiondat(:,2);
>> logtime1=logtime(group==1);
>> logtime2=logtime(group==2);
```

Now you have two vectors `logtime1` and `logtime2` for the two samples. We can check if the normal distribution is plausible in each of the subpopulations:

```
>> subplot(1,2,1)
>> qqplot(logtime1)
>> subplot(1,2,2)
>> qqplot(logtime2)
```



There is no clear evidence against the normality assumption, so we can suppose that it holds. Besides,

```
>> std(logtime1)
  ans =
  0.8137
  >> std(logtime2)
  ans =
  0.8178
```

$\rightsquigarrow$ the sample standard deviations in each sample are very similar, so it can be assumed that $\sigma_1^2 = \sigma_2^2$ (equal variances in the subpopulations).

c) The t-confidence interval for a difference in means is thus:

```
>> [h,p,ci]=ttest2(logtime1,logtime2);ci
  ci =
  0.0608 0.8003
```

$\rightsquigarrow$ we are 95% confident that the true difference in mean between the two groups is in $[0.0608, 0.8003]$. As 0 does not belong to this interval (that is, 0 is not a likely value for the difference in means), it appears that the provision of visual information does make a difference in time taken to fuse images. Note that the `ttest2` command in MATLAB by default assumes that $\sigma_1 = \sigma_2$. When this assumption is doubtful, you should specify that you want to use the formula for unequal variances using Welch-Satterthwaite's approximation for the effective degrees of freedom. Type `help ttest2` for more information.

d) We want to test :

$$H_0 : \ \mu_1 = \mu_2, \qquad \text{vs} \qquad H_a : \ \mu_1 \neq \mu_2.$$

From the sampling distribution,

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Using Matlab, we do not reject the null hypothesis since $p > 0.05$ and conclude there is no evidence log times differ between the groups.

```
>> [h, p, ci, stats] = ttest2(logTime(group == 1), logTime(group == 2))

h =

     1


p =

    0.0231


ci =

    0.0608
    0.8003


stats =

  struct with fields:

    tstat: 2.3190
       df: 76
```

```
          sd: 0.8156
```

Exercise 6

The observations are paired, so we should compute a paired-t confidence interval. After importing the data as a matrix, differences can be obtained by

```
>> abs_diff = abs(:,1) - abs(:,2)
>> [h,p,ci,stats] = ttest(abs_diff)

h =

     1


p =

   1.3146e-99


ci =

   -0.8989
   -0.8625


stats =

  struct with fields:

    tstat: -96.2166
       df: 99
       sd: 0.0915
```

The results indicate differences in the impedance between front and rear ABS sensors.

MATH2089
Numerical Methods and Statistics

STATISTICS – CHAPTER 10 – REGRESSION ANALYSIS

SOLUTIONS

Exercise 1

a) The slope of the straight line ($\beta_1 = -0.01$) is the average change in reaction time for a one degree Fahrenheit increase in the temperature of the chamber. So, with one degree Fahrenheit increase in temperature, the true average reaction time will decrease by 0.01 hour. With a 10 degree increase in temperature, the true average reaction time will decrease by $10 \times 0.01 = 0.1$ hour.

b) When $X = 200$, $\mu_{Y|X=200} = 5 - 0.01 \times 200 = 3$ hours. When $X = 250$, $\mu_{Y|X=250} = 5 - 0.01 \times 250 = 2.5$ hours.

c) We know that $Y|(X = x) \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma)$. Hence, at $X = 250$,

$$Y|(X = 250) \sim \mathcal{N}(2.5, 0.075)$$

It follows that, when $X = 250$,

$$\begin{aligned}\mathbb{P}(2.4 < Y < 2.6) &= \mathbb{P}\left(\frac{2.4 - 2.5}{0.075} < Z < \frac{2.6 - 2.5}{0.075}\right) = \mathbb{P}(-1.33 < Z < 1.33)\\ &= \Phi(1.33) - \Phi(-1.33) = 0.9082 - 0.0918 = 0.8164\end{aligned}$$

d) Out of five independent repetitions, the number of observed reaction times between 2.4 and 2.6 hours, say W, is binomially distributed, with probability of success being the one found in c) :

$$W \sim \text{Bin}(5, 0.8164)$$

Therefore,

$$\mathbb{P}(W = 5) = \binom{5}{5} 0.8164^5 (1 - 0.8164)^0 = 0.8164^5 = 0.3627$$

Exercise 2

a) For a 1 mg/cm^2 increase in dissolved material, we expect a 0.144 g/l increase in calcium content. Secondly, 86% of the observed variation in calcium content can be attributed to the linear relationship between calcium content and dissolved material (that is, what the linear regression model is able to take into account).

b) We have

$$\hat{y}(50) = 3.678 + 0.144 \times 50 = 10.878 \quad \text{(g/l)}$$

c) We know that an unbiased estimate of σ^2, the variance of the error term, is

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{e}_i^2$$

where $\hat{e}_i$'s are the residuals. See that

$$s^2 = \frac{ss_e}{n-2},$$

where ss_e is the observed error sum of squares. Now,

$$r^2 = \frac{ss_r}{ss_t} = 0.860$$

and $ss_t = 320.398$, so that

$$ss_r = 0.860 \times 320.398 = 275.542.$$

As $ss_t = ss_e + ss_r$, we also find

$$ss_e = 320.398 - 275.542 = 44.856.$$

Then,

$$s^2 = \frac{44.856}{21} = 2.136 \quad ((\text{g/l})^2)$$

and the estimated standard deviation of the error term is

$$s = \sqrt{2.136} = 1.462 \quad \text{(g/l)}$$

Exercise 3

a) Yes, the scatter plot shows that a straight line should be a good representation of the change in the response variable Y with the predictor X.

b) From the regression output the coefficient of determination is

$$r^2 = 0.931$$

indicating that 93.1% of the observed variation in mist can be attributed to the simple linear regression. The sample correlation coefficient is $\sqrt{0.931} = 0.965$ (positive sign as the slope of the regression line is positive).

c) If increasing velocity by 900 cm/sec results in an average change in the response of 0.6 mg/m^3, then the true population slope coefficient is

$$\beta_1 = \frac{0.6}{900} = 6.667 \times 10^{-4}.$$

Therefore, we would like to test the hypothesis

$$H_0 : \beta_1 = 6.667 \times 10^{-4}$$

against

$$H_a : \beta_1 < 6.667 \times 10^{-4}$$

(one-sided test). We know that

$$\sqrt{s_{xx}}\,\frac{\hat{\beta}_1 - \beta_1}{S} \sim t_{n-2}. \qquad (\star)$$

The rejection criterion for H_0 is thus

$$\text{reject } H_0 \text{ if } \hat{b}_1 < 6.667 \times 10^{-4} - t_{n-2;1-\alpha} \times \frac{s}{\sqrt{s_{xx}}}$$

The quantile of interest is given by MATLAB :

$$t_{5;0.95} = 2.015.$$

In the output, we find the estimated value of β_1

$$\hat{b}_1 = 6.2108 \times 10^{-4}$$

and its standard error

$$\frac{s}{\sqrt{s_{xx}}} = 7.579 \times 10^{-5},$$

so that the rejection criterion is

$$\text{reject } H_0 \text{ if } \hat{b}_1 < 6.667 \times 10^{-4} - 2.105 \times 7.579 \times 10^{-5} = 5.140 \times 10^{-4}.$$

Consequently, we fail to reject H_0 and conclude that if X increases by 900 units, the true average increase in the response Y is not substantially less than 0.6.

d) We will estimate β_1 using a two-sided 95% confidence interval. From $(\star)$, this interval is given by

$$\left[\hat{b}_1 \pm t_{n-2;1-\alpha/2}\frac{s}{\sqrt{s_{xx}}}\right] = [6.2108 \times 10^{-4} \pm 2.571 \times 7.579 \times 10^{-5}] = [4.26 \times 10^{-4}, 8.16 \times 10^{-4}]$$

$\rightsquigarrow$ we are 95% confident that the true average change in mist associated with a 1 cm/sec increase in velocity is between 4.26×10^{-4} and 8.16×10^{-4}.

e) Although based on only 7 residuals, the normal quantile plot does not suggest that the normal distribution is inappropriate for the model residuals.

f) The required confidence interval could be computed using the expression :

$$\left[\hat{y}(x) - t_{n-2;1-\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}, \hat{y}(x) + t_{n-2;1-\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}\right]$$

This is what has been done in the output labelled "Predicted values for new observations". The fitted value is $\hat{y}(500) = 0.7147$ and the standard error of this estimate is $s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}} = 0.0227$. The confidence interval is seen to be, at confidence level 95%,

$$[0.6562, 0.7731]$$

when $X = 500$ cm/sec. We are thus 95% confident that, when the velocity is set to 500 cm/sec, the true mean extent of mist is between 0.6562 and 0.7731 mg/mm^3. This interval can be checked visually on the graph using the red dashed lines.

g) The required prediction interval could be computed using the expression :

$$\left[\hat{y}(x) - t_{n-2;1-\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}, \hat{y}(x) + t_{n-2;1-\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}\right]$$

This is what has been done in the output labelled "Predicted values for new observations". The prediction interval is seen to be, at confidence level 95%,

$$[0.5639, 0.8654]$$

when $X = 500$ cm/sec. We are thus 95% confident that, if the velocity is set to 500 cm/sec, the next observed extent of mist will be between 0.5639 and 0.8654 mg/mm^3. This interval can be checked visually on the graph using the green dashed lines.

h) The prediction interval in part g) is wider because it contains an additional source of variability to that used to obtain the confidence interval in part g). This additional source of variability is not related to the number n of observations used to fit the regression line and cannot be reduced by increasing n. It reflects the variation in the observed response values around the regression line. These have estimated standard deviation $s = 0.054$.

i) Draw an horizontal line at $y = 1$ across to the upper green dashed lines in the figure. This corresponds to a value of x of approximately 700 cm/sec.

Exercise 4

a) Import the data as a 'Numeric Matrix'. Define the vectors:

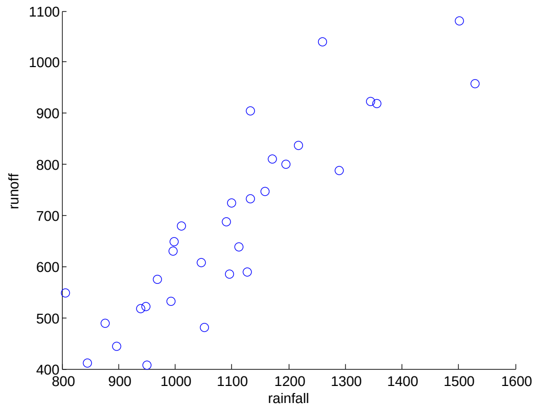
```
>> rainfall=rain(:,1);
>> runoff=rain(:,2);
```

b) Scatter-plot with labels:

```
>> scatter(rainfall,runoff)
>> xlabel('rainfall')
>> ylabel('runoff')
```

Note: alternatively, you can type `plot(rainfall,runoff,'.')` to produce the plot.

We get:



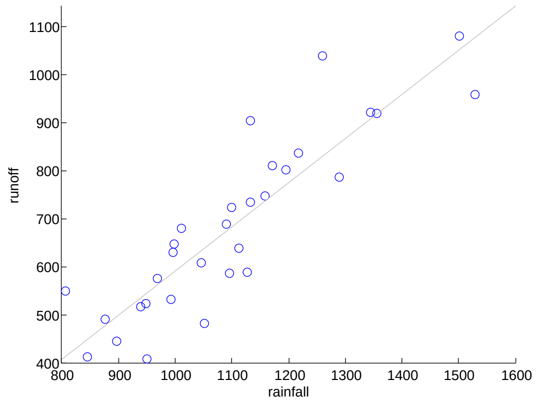
$\rightsquigarrow$ up to the natural variability inherent in the observations, there seems to be a linear increase in the runoff as the rainfall increases. Hence, a linear model like

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

with the response Y being the runoff and the predictor X being the rainfall, should be appropriate. In this expression, ε is the random error term, accounting for the natural variability present in Y even when X is fixed.

c) Add the least squares regression line:

```
>> lsline
```



d) Fit the linear regression model:

```
>> rainMod=fitlm(rainfall,runoff)
```

We get the following output:

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE         tStat       pValue  
                   ________    ________     _______    __________

    (Intercept)    -327.12       99.568     -3.2854     0.0026651
    x1             0.91946     0.089285      10.298    3.3758e-11


Number of observations: 31, Error degrees of freedom: 29
Root Mean Squared Error: 86.7
R-squared: 0.785,  Adjusted R-Squared 0.778
F-statistic vs. constant model: 106, p-value = 3.38e-11
```

(i) The estimated coefficients of the regression can be read from the output or obtained from the fitted object.

```
>> rainMod.Coefficients

                   Estimate       SE         tStat       pValue  
                   ________    ________     _______    __________

    (Intercept)    -327.12       99.568     -3.2854     0.0026651
    x1             0.91946     0.089285      10.298    3.3758e-11
```

$\rightsquigarrow$ the estimated regression coefficients are $\hat{b}_0 = -327.1249$ and $\hat{b}_1 = 0.9195$, and the estimated regression line is

$$\hat{y}(x) = -327.1249 + 0.9195 \times x$$

(ii) We get:

```
>> rainMod.MSE
ans =
7.5144e+03
```

$\rightsquigarrow$ the estimated variance of the error term is $s^2 = 7514.4$ ((mm/h)^2), that is an estimated standard deviation of $s = 86.69$ mm/h

(iii) The relevant null hypothesis to test is $H_0 : \beta_1 = 0$. We know (see Slide 448) that at significance level α, the decision rule is

$$\text{reject } H_0 \text{ if } \hat{b}_1 \notin \left[-t_{n-2,1-\alpha/2} \frac{s}{\sqrt{s_{xx}}}, t_{n-2,1-\alpha/2} \frac{s}{\sqrt{s_{xx}}} \right]$$

with a p-value given by

$$p = 2\mathbb{P}(T > |t_0|)$$

for $T \sim t_{n-2}$, with

$$t_0 = \sqrt{s_{xx}} \frac{\hat{b}_1}{s}.$$

For each item, the second value corresponds to the test for $H_0 : \beta_1 = 0$ (the first value is for $H_0 : \beta_0 = 0$). Thus, we find $\hat{b}_1 = 0.9195$, the standard error of this estimate

$$\frac{s}{\sqrt{s_{xx}}} = 0.0893,$$

the observed value of the test statistic (`t`)

$$t_0 = 10.2980$$

and the associated p-value

$$p = 0.0000$$

(very small) computed from the t_{29}-distribution (`dfe=29`).

⇝ we reject the null hypothesis $H_0 : \beta_1 = 0$ at significance level $\alpha = 0.05$ (as $p < \alpha$), which means in plain language that the rainfall does have a significant influence on the runoff.

(iv) the (estimated) expected change in the runoff for a unit change in the rainfall is the slope of the fitted straight line, that is $\hat{b}_1 = 0.9195$ ⇝ when the rainfall increases by 1 mm, the runoff increases by 0.9195 mm/h (on average).

(v) We can obtain this from the `rainMod.coefCI` object.

```
>> rainMod.coefCI
ans =
 -530.7639 -123.4858
    0.7368    1.1021
```

So the 95% confidence interval for β_1 is $[0.9195 \pm 2.0452 \times 0.0893] = [0.7369, 1.1021]$
⇝ we are 95% confident that the 'true' value of the regression slope β_1 lies between 0.7369 and 1.1021.

(vi) this is given by the r^2 coefficient (coefficient of determination). Here we get:

```
>> rainMod.Rsquared.Ordinary
ans =
    0.7853
```

⇝ $r^2 = 0.7853$: 78.53% of the variation of the runoff is explained by the variation in the rainfall. The remaining 21.47% of the variation is due to other causes, when the rainfall is fixed to a given amount.

(vii) The sample correlation between rainfall and runoff is just the square root of the r^2-coefficient. We get:

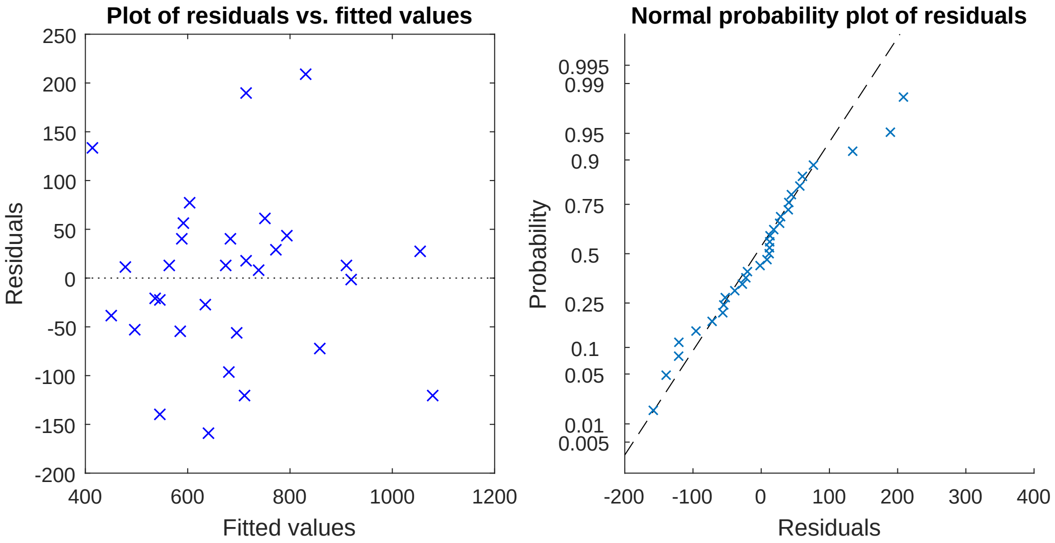
```
>> sqrt(rainMod.Rsquared.Ordinary) ans = 0.8862
```

⇝ the estimated correlation $\hat{\rho} = 0.8862$, close to 1 ⇝ quite strong linear relationship between rainfall and runoff

(viii) Plot of the residuals versus the fitted values and normal quantile plot of the residuals:

```
subplot(1,2,1)
plotResiduals(rainMod,'fitted')
subplot(1,2,2)
plotResiduals(rainMod,'probability')
```

We obtain:



⇝ no serious departure from the normality assumption (quantile plot), and no particular pattern observed in the residuals plotted against the fitted values. Besides, the variance of the residuals does not seriously vary with the fitted value
⇝ it seems reasonable to assume that the error terms were independently drawn from the distribution

$$\varepsilon \sim \mathcal{N}(0, \sigma)$$

⇝ the necessary assumptions seem to be fulfilled, which validates the model

Exercise 5

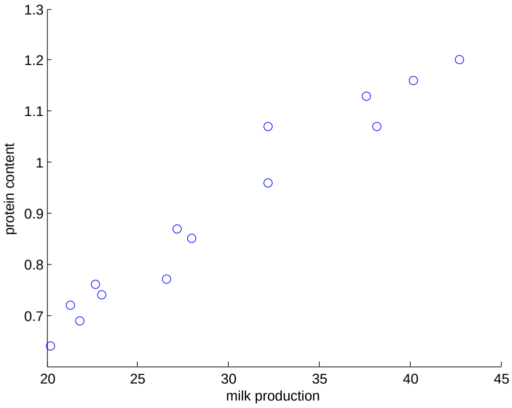
a) Import the data as a ’Numeric Matrix’. Define the vectors:

```
>> x=milk(:,1);
>> y=milk(:,2);
```

b) Scatter-plot with labels:

```
>> scatter(x,y)
>> xlabel(’milk production’)
>> ylabel(’protein content’)
```

We get:



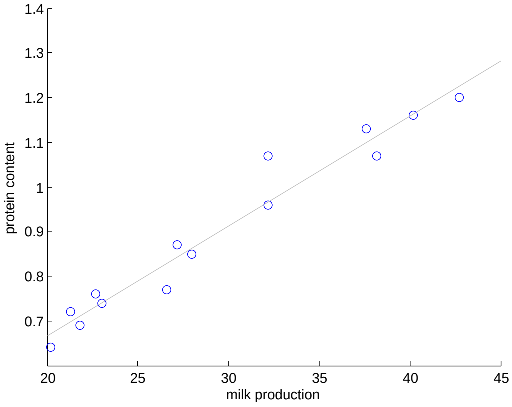
$\leadsto$ up to the natural variability inherent in the observations, there is a clear linear relationship between protein content and daily milk production. Hence, a linear model like

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

should be appropriate.

c) Add the least squares regression line:

```
>> lsline
```



d) Fit the linear regression model:

```
>> milkMod=fitlm(x,y)

milkMod =

Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   --------    ---------    ------    ---------

    (Intercept)     0.17558     0.046399    3.7841    0.0026043
    x1             0.024576    0.0015227    16.139    1.678e-09
```

```
Number of observations: 14, Error degrees of freedom: 12
Root Mean Squared Error: 0.042
R-squared: 0.956,  Adjusted R-Squared 0.952
F-statistic vs. constant model: 260, p-value = 1.68e-09
```

e) The estimated coefficients of the regression can be read from the output or obtained from the fitted object.

```
>> milkMod.Coefficients
ans =

                     Estimate       SE         tStat      pValue
                     --------    ---------    ------    ---------

    (Intercept)       0.17558     0.046399    3.7841    0.0026043
    x1               0.024576    0.0015227    16.139    1.678e-09
```

Hence, the fitted regression line is

$$\hat{y}(x) = 0.1759 + 0.0246 \times x$$

f) we have to test the hypothesis $H_0 : \beta_1 = 0$. These values are what MATLAB returns in the object `milkMod.Coefficients` above. So, $\hat{b}_1 = 0.0246$, which is indeed quite small. Does that mean that β_1 might be equal to 0? No! The absolute value of the estimated coefficient as such is not informative. It should be compared to the accuracy of that estimation ⇝ that is what we do when computing the value t_0 (ratio between the estimate and its standard error). Here the standard error for the estimate is 0.0015, and the value of t_0 (`t`) is 16.1393 ⇝ very large value! The associated p-value is 0.0000, so we can reject H_0 without too much risk ($p < \alpha$). Hence the production does have a significant influence on the protein content (the observed linear relationship is not a consequence of chance only).

g) We can obtain confidence intervals from using the `predict` function. For a 95% confidence interval, alpha is equal to 0.01.

```
>> [ypred,ypi] = predict(milkMod,30,'alpha',0.01)

ypred =
    0.9129
ypi =
    0.8785    0.9472
```

⇝ we are 99% confident that the 'true' mean protein content for cows producing 30 kg of milk per day is between 0.8785 and 0.9472 kg/day.

h) The above confidence interval refers to the mean protein content for all cows producing 30 kg/day. Here, the prediction interval is about the protein content for a single cow ⇝ much more variability in a single observation than in an average!

```
>> [ypred,ypi] = predict(milkMod,30,'alpha',0.01,'Prediction','observation')

ypred =
    0.9129
ypi =
    0.7799    1.0458
```

⇝ we are 99% confident that the protein content for that particular cow whose production is 30 kg/day will be between 0.7799 and 1.0459 kg/day.

MATH2089
Numerical Methods and Statistics

Statistics – Chapter 11 – ANOVA

SOLUTIONS

Exercise 1

Let μ_A, μ_B, μ_C, μ_D, μ_E and μ_F be the true unknown mean flow rates for nozzles of type A, B, C, D, E and F respectively. Then, the null hypothesis

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu_E = \mu_F$$

has to be tested against the alternative

$$H_a : \text{ not all the means are equal}$$

The test statistic is F and is known to follow the Fisher's F-distribution with $k-1$ and $n-k$ degrees of freedom if H_0 is true. Here, we have $k = 6$ groups and $n = 66$ observations, which leads to the Fisher's distribution $\mathcal{F}_{5,60}$. At level $\alpha = 0.05$, the rejection criterion is thus

$$\text{reject } H_0 \text{ if } f_0 > f_{5,60;0.95} = 2.37$$

Here the observed value $f_0 = 4.2$, so that we reject H_0. The associated p-value is $p = \mathbb{P}(X > 4.2)$ for $X \sim \mathcal{F}_{5,60}$. Matlab gives that

$$p = 0.0024,$$

which is indeed very small (in particular, smaller than α). We are therefore very confident when claiming that the nozzles do not have all the same mean flow rate.

This conclusion is obviously only valid if the basic assumptions for the ANOVA model are fulfilled : for each nozzle type the flow rate is normally distributed, and the standard deviation is the same for all the distributions; besides, we have independent random samples from the six types of nozzle. We can check for plausibility of the normality assumption and constant variance by plotting the residuals. In the experimental design we can plan for independent random samples.

Exercise 2

a) From the data we have immediately :

Source	degrees of freedom	sum of squares	mean square	F-statistic
Treatment	A	B	C	2.8
Error	D	E	8.2	
Total	G	H		

Since $k = 4$ groups and $n = 4 \times 12 = 48$, we have that $A = 3$, $G = 47$ and hence $D = 47 - 3 = 44$. Now, $\frac{E}{44} = 8.2$, so $E = 360.8$. Similarly, $\frac{C}{8.2} = 2.8$, so $C = 22.96$. Now, $\frac{B}{3} = 22.96$, so $B = 68.88$ and $H = B + E = 68.88 + 360.8 = 429.68$. Hence we have :

Source	degrees of freedom	sum of squares	mean square	F-statistic
Treatment	3	68.88	22.96	2.8
Error	44	360.8	8.2	
Total	47	429.68		

b) Let μ_1, μ_2, μ_3 and μ_4 be the true unknown mean yield for crops types 1, 2, 3 and 4 respectively. The null hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

has to be tested against the alternative

$$H_a : \text{ not all the means are equal}$$

The test statistic is F and is known to follow the Fisher's F-distribution with $k-1$ and $n-k$ degrees of freedom if H_0 is true. Here, we have the $\mathcal{F}_{3,44}$ distribution. At level $\alpha = 0.01$, the rejection criterion is thus

$$\text{reject } H_0 \text{ if } f_0 > f_{3,44;0.99} = 4.2606$$

Consequently, we do not reject H_0, as the observed value of the test statistic is $f_0 = 2.8$.

The associated p-value is $p = \mathbb{P}(X > 2.8)$ for $X \sim \mathcal{F}_{3,44}$. MATLAB says : $p = 0.0509$. Indeed, the p-value is larger than $\alpha = 0.01$. With this required level of certainty, it is too risky to reject H_0. We lack evidence to state that the crops do not have equal yield on average.

Exercise 3

a) Let μ_1, μ_2 and μ_3 be the true unknown mean heart rate under stress in the situations of being alone, with a friend and with a dog, respectively. The null hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

has to be tested against the alternative

$$H_a : \text{ not all the means are equal}$$

b) From the output the observed value of the test statistic $f_0 = 14.08$. The p-value, given by $p = \mathbb{P}(X > 14.08$ with $X \sim \mathcal{F}_{2,42}$, is found to be close to 0 (MATLAB says : $p = 2 \times 10^{-5}$).

c) There is thus a strong evidence to reject H_0 and to claim that the average heart rate under stress is not the same in the three groups.

d)
- Independence of errors (across and within groups), guaranteed by randomly sampling from different populations, or in this case, by randomly assigning subjects to treatment groups.
- Normality of errors, which you can check using a normal quantile plot (not available here).
- Equal variance across groups, which you can check by seeing if sample sd's are within a factor of two of each other, or by inspecting at a residual vs fits plot for similar spread at the different fitted values (also not available here).

e) The sample of women with a friend had sample mean 91.3, the sample of women with a dog had sample mean 73.5 and the sample of women being by themselves had sample mean of 82.5.

f) There are $\binom{3}{2} = 3$ comparisons : group 1 versus group 2 (i.e., control vs. friend), group 1 versus group 3 (i.e., control vs. pet) and group 2 versus group 3 (i.e., friend vs. pet). In each case we have a two-sample t-test and the test statistic will have a t_{42} distribution when H_0 is true. We take an overall significance level of 0.05. So with the Bonferonni adjustment we must have a p-value under $\frac{0.05}{3} = 0.0167$ for at least one of the pairwise test to reject H_0 at 0.05 level. In the output we find the value of $s_p = \sqrt{ms_{\text{Er}}} = 9.208$ ('pooled sample standard deviation'). For the two-sample t-test for

$$H_0^* : \mu_1 = \mu_2$$

against

$$H_a^* : \mu_1 \neq \mu_2$$

we find an observed value of the test statistic

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{82.524 - 91.325}{9.208 \times \sqrt{\frac{2}{30}}} = -2.6176$$

$\rightsquigarrow$ the associated p-value of this two-sided test is

$$p = 2 \times \mathbb{P}(T > 2.6176)$$

for $T \sim t_{42}$. MATLAB says : $p = 0.0123$. Therefore, we can conclude that $\mu_1 \neq \mu_2$, and as the p-value associated to this decision is smaller than 0.0167, this also leads to the rejection of $H_0 : \mu_1 = \mu_2 = \mu_3$ (in agreement with the ANOVA F-test).

Exercise 4

a) Here there are $k = 3$ states and $n = 3 \times 10 = 30$ observations. So, $A = 2$ and $C = 29$. Also, $D = 41081016 - 17470016 = 23611000$. Now, $B = \frac{17470016}{2} = 8735008$ and $E = \frac{B}{874482.48} = 9.9888$:

Source	df	SS	MS	F
Treatment	2	17470016	8735008	9.9888
Error	27	23611000	874482.48	
Total	29	41081016		

b) The null hypothesis is that the states are equal in terms of their average salary offered : $H_0 : \mu_1 = \mu_2 = \mu_3$, where μ_1, μ_2 and μ_3 are the average salaries offered in each state. The alternative is that this is not the case, i.e. there is at least one state which offers a salary different to the others : $H_a :$ not all μ_i are equal. The rejection criterion is

$$\text{reject } H_0 \text{ if } f_0 > f_{2,27;0.95}.$$

MATLAB says that $f_{2,27;0.95} = 3.3541$. As we have observed $f_0 = 9.9888$, we clearly reject H_0. The associated p-value is

$$p = \mathbb{P}(X > 9.9888)$$

for $X \sim \mathcal{F}_{2,27}$. MATLAB says : $p = 6 \times 10^{-4}$. We are very confident when claiming that the states differ in average starting salaries offered.

Exercise 5

b) Define μ_A, μ_B and μ_C the 'true' mean strengths for the different curing methods. Then we would like to test the null hypothesis

$$H_0 : \mu_A = \mu_B = \mu_C$$

against the alternative

$$H_a : \text{at least one is different to the others}$$

Testing such a hypothesis of equality of more than two means require an Analysis of Variance (ANOVA). However here we have a clear blocking factor: the batch. The samples for the three curing methods are not independent, as we can expect some kind of dependence between the measures relative to specimens from the same batch, even treated with different methods.

c) An ANOVA can be though off as a linear model where the predictors are categorical. We can use the `fitlm` function to fit the model, and then the `anova` function to obtain the ANOVA table. We can tell the `fitlm` function which variables are categorical, in this example it will be 1 and 2.

```
>> concreteMod=fitlm([Method,Batch],Strength,'CategoricalVars',[1,2]);
>> anova(concreteMod)

ans =

              SumSq     DF    MeanSq      F          pValue
              ------    --    ------    ------    ----------

    x1        23.229     2    11.614    8.6946     0.0022784
    x2        86.793     9    9.6437    7.2193    0.00020215
    Error     24.045    18    1.3358
```

We can see how the ANOVA procedure has split the total amount of variation in the observations into three components: the variation due to the different curing methods (here the columns, $ss_{\text{Tr}} = 23.229$), the variation due to the different batches (here the rows, $ss_{\text{Block}} = 86.793$) and the residual variation ($ss_{\text{Er}} = 24.045$). The other quantities of interest, namely the degrees of freedom for each component, the mean squares, the F-ratio and the associated p-value, have also been computed by MATLAB.

d) We reject H_0 if ms_{Tr}, here 11.6143, is much larger than $ms_{\text{Er}} = 1.3358$. We find a ratio

$$f_0 = \frac{11.6143}{1.3358} = 8.69$$

At significance level 5%, this should be compared to the quantile $f_{2,18;0.95}$ of the Fisher's F-distribution with 2 and 18 degrees of freedom, which is

```
>> finv(0.95,2,18)
  ans =
  3.5546
```

$\rightsquigarrow f_0 > f_{2,18;0.95}$, we reject H_0 at level 5%.

Check:

```
>> 1-fcdf(8.69,2,18)
  ans =
  0.0023
```

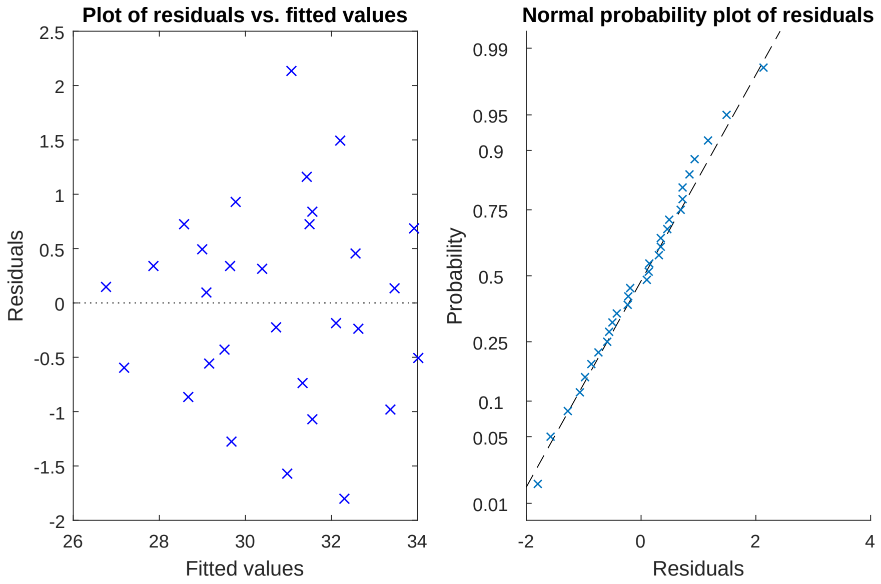
Indeed, $p < \alpha$

In plain language, the conclusion is thus that there is some significant difference in the mean strength for the different curing methods.

e) Calculate the residuals using the suggested commands. Then,

```
subplot(1,2,1)
plotResiduals(concreteMod,'fitted')
subplot(1,2,2)
plotResiduals(concreteMod,'probability')
```

We get:



No serious departure from the normality assumption (quantile plot). No particular pattern observed in the residuals plotted against the fitted values. The variance of the residuals might slightly increase with the fitted value but this is not clear, thus it seems reasonable to assume that the error terms were independently drawn from the distribution

$$\varepsilon \sim \mathcal{N}(0, \sigma)$$

⇝ the necessary assumptions seem to be fulfilled, which validates the model

f) One-way ANOVA without blocking factor:

```
>> anova1(concrete)
```

We get:

```
>> concreteMod2=fitlm(Method,Strength,'CategoricalVars',[1]);
>> anova(concreteMod2)

ans =

             SumSq     DF    MeanSq      F         pValue
             ------    --    ------    ------    --------

    x1       23.229     2    11.614    2.8292    0.076642
    Error    110.84    27    4.1051
```

⇝ at level 5%, we wouldn't have rejected H_0 ($p > \alpha$). Ignoring the blocking factor makes the test less powerful in detecting evidence against H_0