NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

MSc DATA SCIENCE AND INFORMATION TECHNOLOGIES

BIOINFORMATICS – BIOMEDICAL DATA SCIENCE

BIOSTATISTICS

Midterm Exercises

Eleni Panagiotopoulou

Professor: Dr. Dionisis Linardatos

Athens

2023-2024

1

A survey shows the following selection of snacks purchased according to gender.

Snack	Male	Female
Hotdog	8	12
Toasts	13	9
Peanuts	9	6
Popcorn	8	10

The claim that snack preference is related to the gender of the consumer will be tested.

Answer

Observations of qualitative characteristics are generally approached using chi-square (X^2) hypothesis tests. This test is useful when the presence (or absence) of a relationship between qualitative quantities is investigated. In this case, these qualities correspond to the gender and snack preference and the claim the relationship between these two qualities is examined. However, this approach is valid given that certain conditions are met regarding the available data:

1. The minimum theoretically expected frequency should be bigger than 1.
2. There are more than 20 observations.
3. Less than 20% of the theoretically expected frequencies are smaller than 5.
4. In the case of 2x2 contingency tables with 20-40 observations, none of the theoretically expected values is less than 5 (not our case).

Funding the hypothesis

H_0: Snack preferences do not differ between men and women.

H_1: Snack preference is different between men and women.

Significance level: a = 0.05

In order to continue with our analysis, the matrix with expected values is calculated. This will also play part in defining whether or not the chi-square hypothesis test is going to deliver reliable results.

Calculation of expected frequencies

The contingency matrix for our data will be calculated in order to check if the aforementioned requirements are satisfied. The expected frequencies are calculated using Excel:

Snack	Male	Female	Row sums
Hotdog	10.13	9.87	20.00
Toasts	11.15	10.85	22.00
Peanuts	7.60	7.40	15.00
Popcorn	9.12	8.88	18.00
Column sums	38.00	37.00	75.00

Figure 1. Matrix of the expected frequencies of the data.

1. Minimum theoretically expected frequency: 7.4> 1: CONDITION MET
2. Observations=75 > 20: CONDITION MET
3. None of the theoretically expected frequencies is smaller than 5: CONDITION MET

Calculation of x^2 table

The equation of chi-squared is shown below:

$$x^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i}$$

	Observed frequencies (OF)		Expected frequencies (E)		O - E		(O - E)^2/E	
	Male	Female	Male	Female	Male	Female	Male	Female
Hotdog	8	12	10.13	9.87	-2.13	2.13	0.45	0.46
Toasts	13	9	11.15	10.85	1.85	-1.85	0.31	0.32
Peanuts	9	6	7.60	7.40	1.40	-1.40	0.26	0.26
Popcorns	8	10	9.12	8.88	-1.12	1.12	0.14	0.14
Column sums	38	37	38	37	0.00	0.00	1.15	1.18

Figure 2. Calculation of X^2 using Excel. The observed and the expected frequencies are gathered in a table. Then, their difference is calculated as well as their difference squared divided by the expected value in each case. The latter, after being summed up, corresponds to x^2.

We can conclude that x^2 = 1,15 + 1,18 = 2,33.

The degrees of freedom are calculated through the formula:

$$df = (c - 1) * (r - 1)$$

where:

c: number of columns, r: number of rows

The degrees of freedom in our case equal df = (2 – 1) * (4 – 1) = 3.

Using disstool from MATLAB, we are able to find the probability of rejecting the null hypothesis.

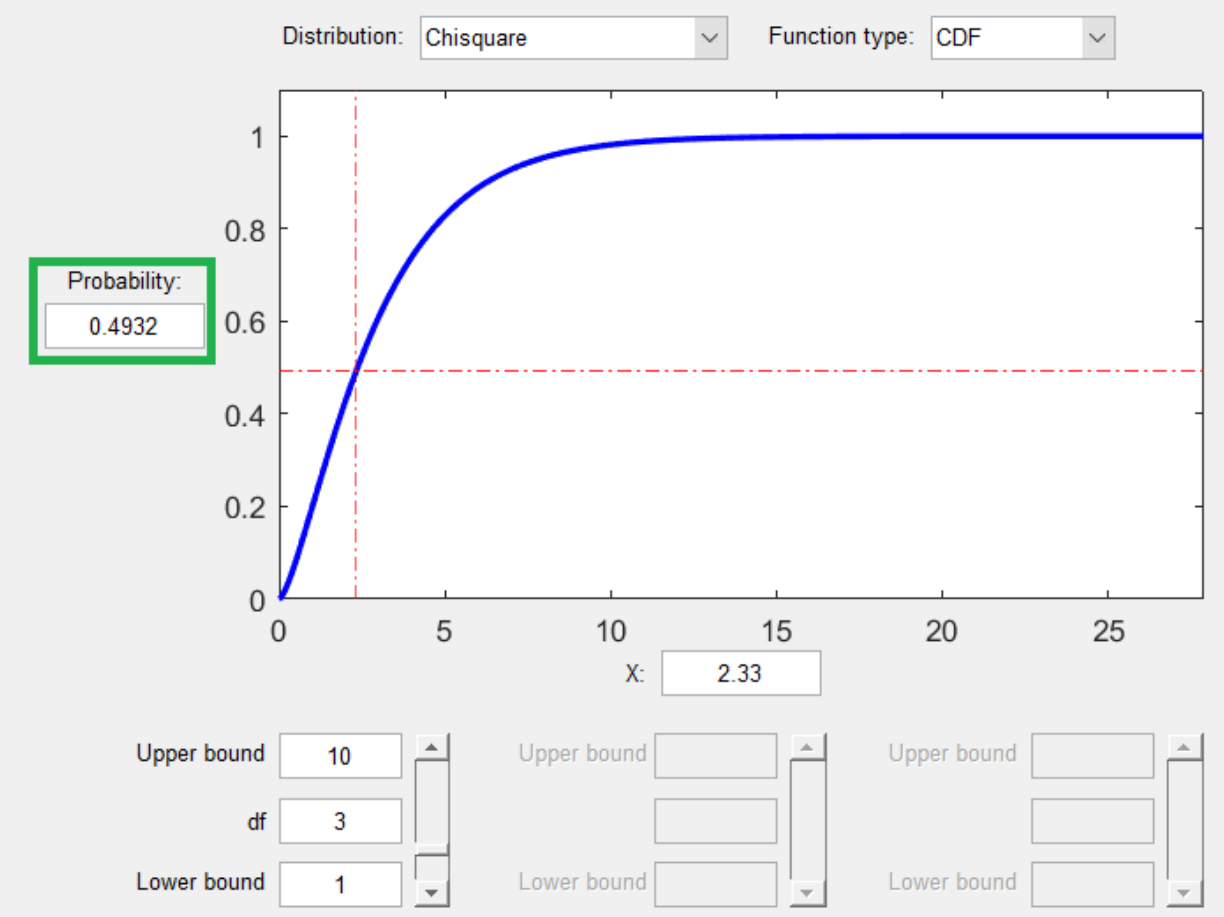


Figure 3. The probability given that the distribution is chi square and X is 2,33 with 3 degrees of freedom. The probability equals 0.49.

For χ^2 = 2.33 and 3 degrees of freedom, the probability equals 0.49. Since P = 0.49 < 0.95 (given that we opted for significance level a = 0.05), the null hypothesis CANNOT be rejected. This experiment suggests that there is no significant difference when it comes to snack preference between men and women.

2

Scientists investigated the level of caffeine in cups of a specific coffee served by a single specialty coffee shop. Cups were purchased on eight consecutive days and the amount of caffeine in each of the eight cups (measured in milligrams) is provided in the following table:

317	305	465	498	517	353	357	410

Determine whether the median amount of caffeine in the cups of the specific coffee differs from 400 milligrams.

Answer

We start by setting the level of significance ***a* = 0.05**. We need to determine whether or not the mean value of a sample differs from a specific value of 400 mg. To approach this, we need to use two-tailed hypothesis test for the mean value. The test is trustworthy in the case that the sample comes from a normal distribution, thus before continuing, a test for normality should be carried out. In our case, the observations (n) of the sample are significantly less than 30, so the Lilliefors test is used. The latter tests the null hypothesis that:

- H_0: The sample under study comes from a normal distribution with undefined mean and variance.

Against the alternative hypothesis:

- H_1: The sample does not come from a normal distribution.

The command in MATLAB: [h, p, lstat, cv] = lillietest(sample);

The output ***h* = 0** and ***p* = 0.3619 > 0.05** suggests that the null hypothesis cannot be rejected. Thus, the two-tailed hypothesis test will provide us with trustworthy results. In order to investigate if the mean value of the sample is equal to 400 mg, we proceed with *ttest()* in MATLAB, since this is the ideal test for samples that come from normal distributions with unknown variance. We set the null (H_0) and the alternative hypothesis (H_1):

- H_0: $\mu = 400$ mg
- H_1: $\mu \neq 400$ mg

The command in MATLAB: [h_mean, p_mean, ci, s] = ttest(sample, 400);

Since ***h (h_mean) = 0*** and ***p-value (p_mean) = 0.9274 > 0.05 = a***, the null hypothesis cannot be rejected. Overall, the caffeine amount of each cup follows a normal distribution and there is not enough data to reject the hypothesis that the mean value of caffeine in each cup is 400 mg.

3

Scientists investigated the way wild monkeys learn to retrieve a coconut from the opposite side of a river. More specifically, 24 wild monkeys were assigned to one of three experimental conditions:

a. *Observation of another monkey: Monkeys watched another monkey retrieving the coconut by building a bridge across the river stream.*
b. *Observation of a human: Monkeys watched a human retrieving the coconut.*
c. *Banana reward: Monkeys were allowed to do what they wanted but each time they managed to retrieve the coconut they were rewarded with a banana.*

After learning, the monkeys were required to perform the task of retrieving a coconut from the opposite side of a river. The time taken to retrieve the coconut (in minutes) was measured and recorded to the following table:

Observation of monkey	*Observation of human*	*Banana reward*
7	*15*	*6*
13	*13*	*9*
3	*7*	*3*
8	*8*	*9*
9	*6*	*5*
2	*13*	*1*
8	*7*	*3*
7	*9*	*10*

Given that the time taken to retrieve the coconut follows a normal distribution, carry out an appropriate analysis to test the hypothesis that the time is equal for all learning conditions. If the hypothesis doesn't hold, which condition leads to the fastest learning?

Answer

We begin by setting the significance level to ***a = 0.05***. We need to determine the effectiveness of each learning method (monkey observation, human observation, or banana reward) by comparing the time it takes to the subjects to retrieve the coconut from the other side of the river. The learning method is deemed more efficient if the mean value of retrieval time is less than the one from the other methods. To start off the analysis, the necessity to compare the mean values from each sample leads us to opt for the one-way analysis of variance (one-way ANOVA). In order for this method to bring trustworthy results, certain requirements must be met.

- **Independence of the samples**: This is fulfilled since the monkeys that were assigned to each learning method are not the same (24 monkeys overall, 8 for each learning method).

- **Normality**: This is also fulfilled since the information that the time to retrieve a coconut follows a normal distribution is provided.
- **Equality of variances**
  In order for equality of variances to be satisfied, we check the validity of that assumption. The *vartestn()* MATLAB command uses as the null hypothesis that the equality of the sample's variances stands, while the alternative hypothesis is the opposite. A two-tailed test is carried out with the following command:

  ```
  [p_var, stats_var] = vartestn(X);
  ```

  Since ***p_var (p-value) = 0.996 > 0.05***, the null hypothesis cannot be rejected, satisfying the third assumption for one-way ANOVA.

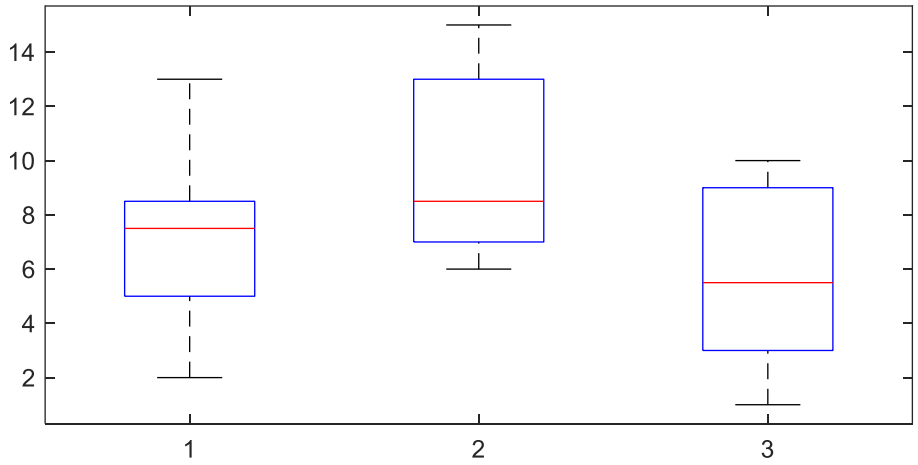


Figure 4. Output boxplots of vartestn(). These boxplots do not strongly suggest that the samples come from a normal distribution given that the median and the mean of the samples do not coincide. However, the analysis will be carried out given that normality is provided. The x-axis contains the learning methods, 1 corresponds to monkey observation, 2 to human observation and 3 to banana reward. Y-axis includes the time values.

Thus, we can proceed with one-way ANOVA. The command:

```
[p, table, stats] = anova1(X);
```

The outcome of the analysis ***p (p-value) = 0.0792 > 0.05 = a*** suggests that the null hypothesis cannot be rejected. This leads us to believe that the mean value of each method does not differ between the samples, indicating that all 3 learning methods have the same efficacy for the monkeys.

Source	SS	df	MS	F	Prob>F
Columns	66.083	2	33.0417	2.87	0.0792
Error	241.875	21	11.5179		
Total	307.958	23			

Figure 5. ANOVA table

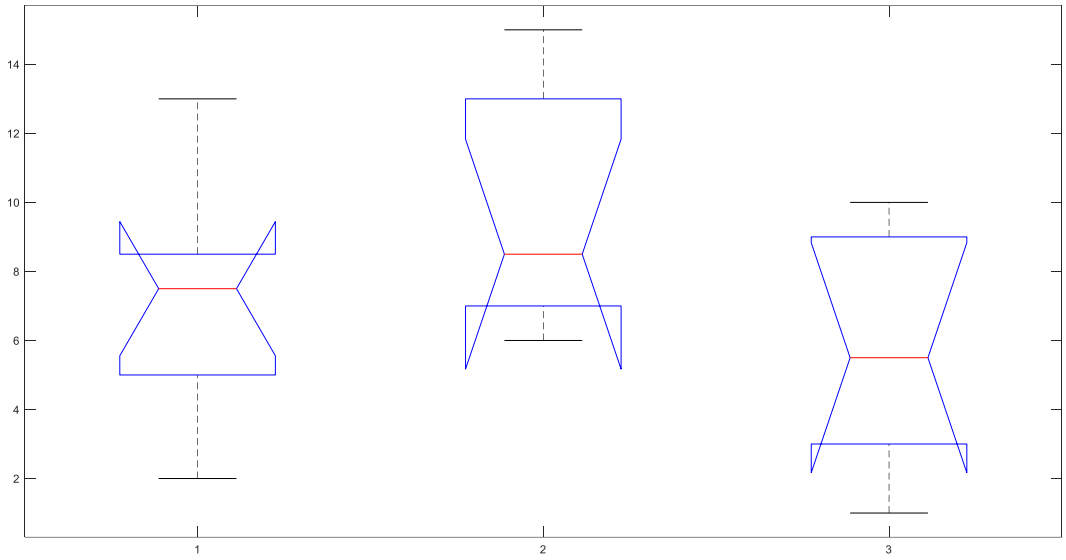


Figure 6. Violin plots of the ANOVA. The x-axis contains the learning methods, 1 corresponds to monkey observation, 2 to human observation and 3 to banana reward. Y-axis includes the time values.

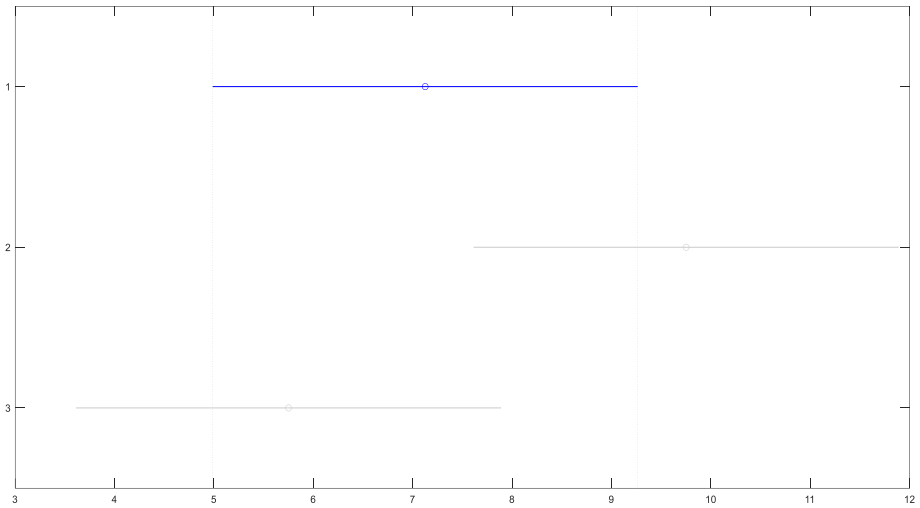
In a significance level of ***a = 0.05***, the mean values of time for each method are equal. However, the p-value of the one-way ANOVA equals 0.0792, a value very close to our level of significance. With the provided data, we cannot conclude a significant difference. For a deeper inspection of the results, the MATLAB command *multcompare()* is used in order to visualize the means of the samples with their respective variance all together in a plot. The command:

```
compare_means = multcompare(stats);
```

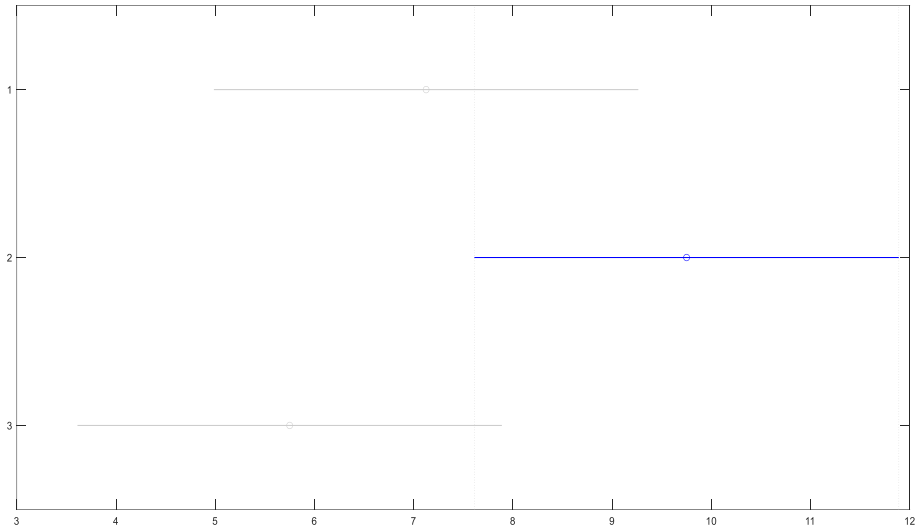
The comparison now takes place in pairs between all the samples. The output of the analysis:

Sample a	Sample b	Lower limit	a-b	Upper limit	P-value
1	2	-6.9022	-2.6250	1.6522	0.2902
1	3	-2.9022	1.3750	5.6522	0.7009
2	3	-0.2772	4.0000	8.2772	0.0695

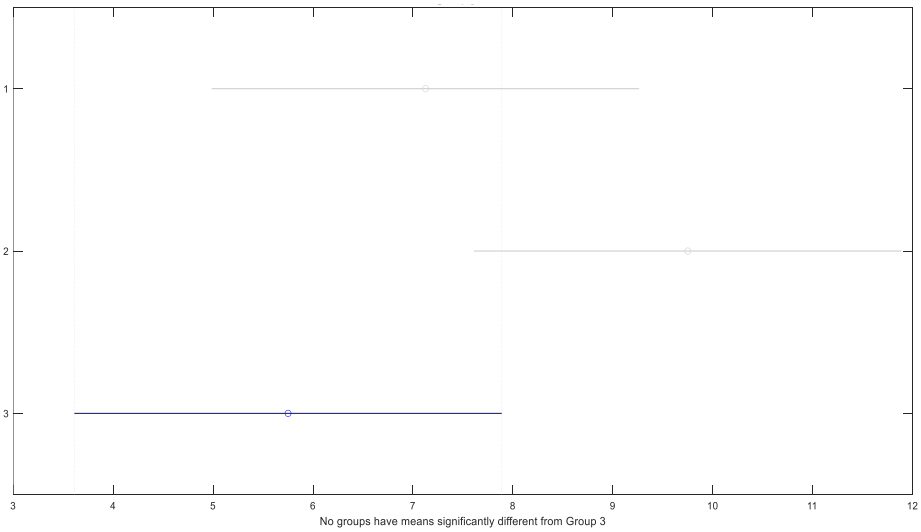
The upper and lower limit are the confidence intervals of the difference between the means of the compared samples. In the case of 2 – 3 comparison we can see that 0 is contained in the interval [-0.2772, 8.2772] but also that the p-value is 0.0695, a little larger than the significance level. Probably, with some more observations for each monkey learning method our results could change. Current data may not suggest statistical difference between the means, but future analysis with more observations could possibly bring new results. In the pictures that follow, the means and variances are visualized. The output comes from the same command.



No groups have means significantly different from Group 1



No groups have means significantly different from Group 2



No groups have means significantly different from Group 3

4

Assuming that the sodium level in human blood follows a normal distribution with a mean value of 160 and a standard deviation of 20, that is the probability that the sodium level in a person's blood is:

a. *less than 140,*
b. *between 125 and 140,*
c. *greater than 180.*

Note: You may use the disttool graphical interface.

Answer

The disttool graphical interface is used for visualization of the normal distribution with Mu=160 and Sigma=20. In order to find the probabilities in question, the CDF (Cumulative Distribution Function) will be used as function type. In probability theory and statistics, the cumulative distribution function (CDF) of a real-valued random variable X, or just distribution function of X, evaluated at x, is the probability that X will take a value less than or equal to x.

a. x < 140

When x = 140, by definition, disttool provides the probability that X will take a value equal or less than 140.

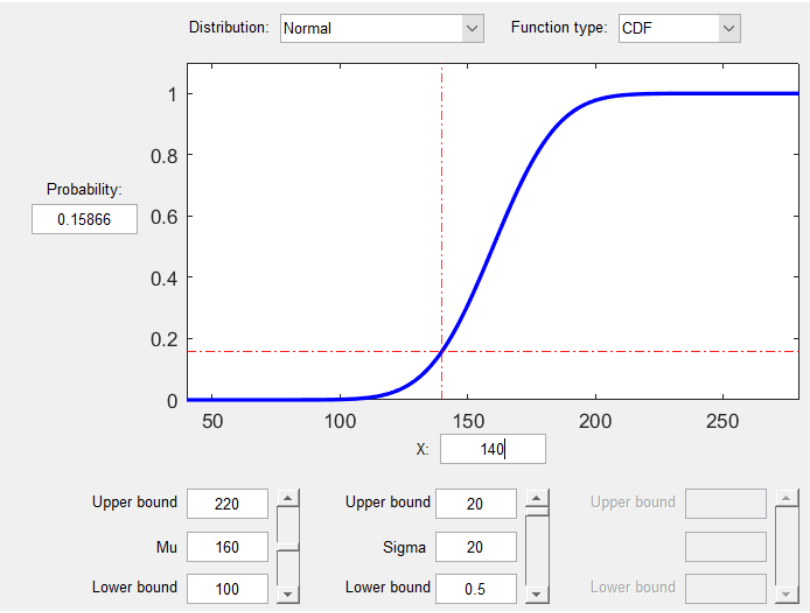


Figure 7. For given X=140, the probability is 0.15866.

We can conclude that ***P(x<140) = 0.159***.

b. 125 < x < 140

The probability in question can be calculated as P(125<x<140) = P(x>125) – P(x>140). In order to calculate the probability of x to be found between 125 and 140, the cumulative probability function will be used to provide the probability if x = 125.

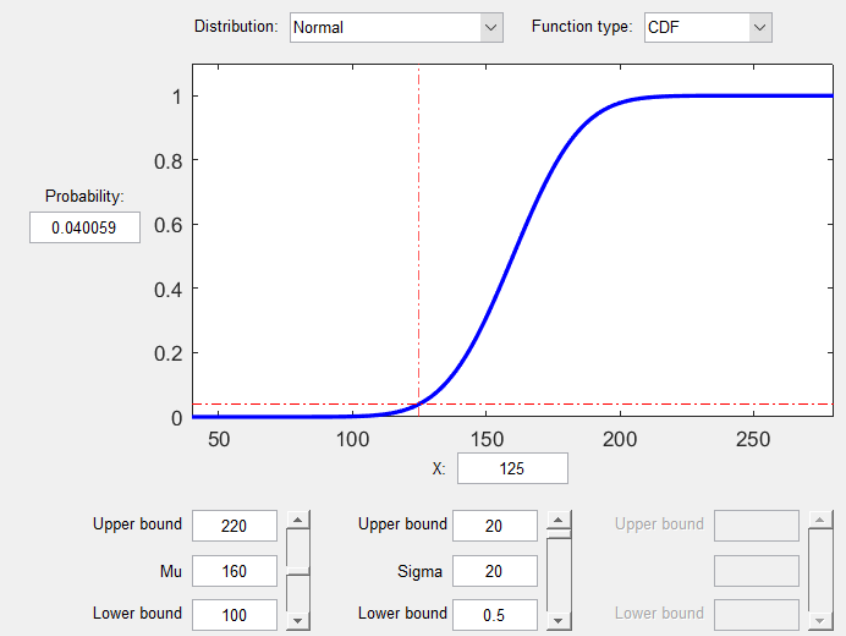


Figure 8. For X = 125, Probability = 0.04

So, P(x<125) = 0.04. Given the nature of the distribution we are able to calculate the probability that x > 125.

P(x>125) = 1 - P(x<125) => ***P(x>125) = 0.96***

However, x needs to also be constrained to being lower than 140. Knowing that P(x<140) = 0.159, it can be calculated:

P(x>140) = 1 - P(x<140) => ***P(x>140) = 0.84***

Now, when it comes to the probability in question, P(125<x<140) = P(x>125) – P(x>140) = 0.96 – 0.84

P(125<x<140) = 0.12

c. x > 180

Due to the nature of a normal distribution, this relationship P(x < $\mu - \sigma$) = P(x > $\mu + \sigma$) holds. In our case $\mu - \sigma = 160 - 20 = 140$, while $\mu + \sigma = 160 + 20 = 180$. The equation above, given our data for the normal distribution, translates to the following:

P(x<140) = P(x>180) = 0.159

We will corroborate that P(x>180) = 0.159, using the CDF of our normal distribution.

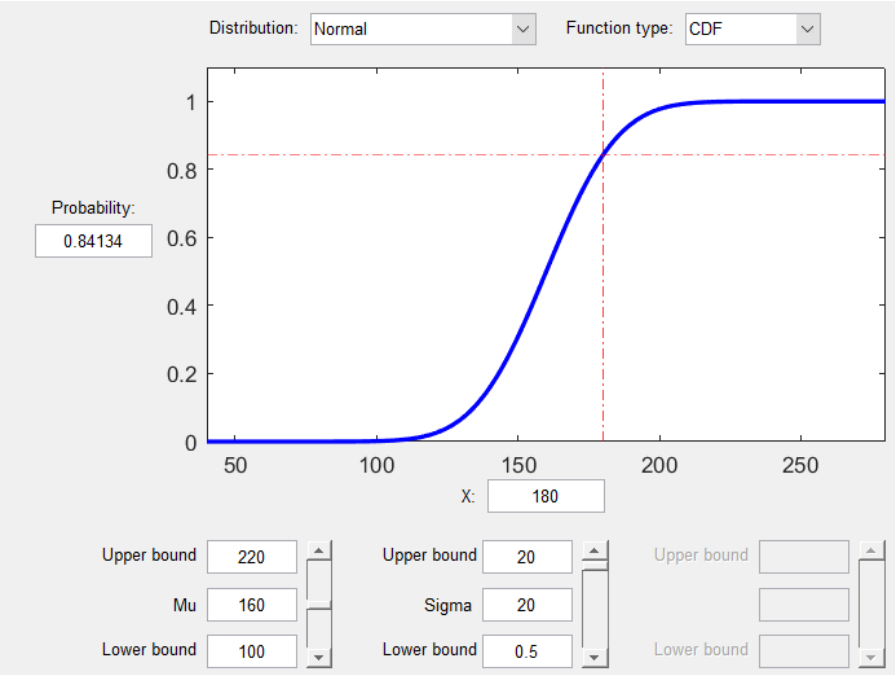


Figure 9. For x = 180, Probability = 0.841.

From the graph, P(x<180) = 0.841. So, P(x>180) = 1 - P(x<180) = 1 – 0.841.

P(x>180) = 0.159

5

A microbiologist wished to determine whether there is any difference in the time it takes to make a product from two different starters: lactobacillus acidophilus (A) and bulgaricus (B). Seven batches of the product were made with each of the starters. The table below shows the time taken, in hours, to make each batch:

Starter A	*7.3*	*6.9*	*8.2*	*6.1*	*6.3*	*7.4*	*6.8*
Starter B	*6.4*	*6.1*	*6.7*	*6.9*	*6.3*	*5.7*	*5.5*

Assuming that both sets of times may be considered to be random samples from normal populations, test the hypothesis that the mean time taken to make the product is the same for both starters. Present all steps of the hypothesis testing.

Answer

To begin, the significance level is set to ***a = 0.05***. We need to compare the time taken in hours for each species to make a product. This problem will be approached with two-tailed hypothesis test for the mean value. The assumptions that must be met are:

- **Sample independence**: This is fulfilled since our two samples are two different species of *Lactobacillus sp.*
- **Normality**: The normality of the distribution is provided.

However, this test has different approaches depending on whether the variances of the samples, when unknown, are equal or unequal. In our case, we check which case holds using the MATLAB command *vartestn()*. The null hypothesis is that the two sample variances are equal while the alternative hypothesis states the opposite. The command used:

```
[p_var,stats_var] = vartestn(x);
```

The **p_var (p-value) = 0.4255 > 0.05 = a**, indicating that the null hypothesis cannot be rejected. Thus, we proceed with the case where the samples have equal variances. We set the null and the alternative hypothesis:

- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 \neq \mu_2$

Now we perform a two-tailed hypothesis test for the mean value testing if the means of the samples are equal. The command used:

```
[h, p, ci, stats] = ttest2(sample_a, sample_b, 0.05, 'both');
```

The output ***h = 1*** and ***p (p-value) = 0.0376 < 0.05 = a***, means that the null hypothesis is rejected. The mean time in hours that it takes both species to make the same product is significantly different. The 95% confidence interval of the difference of the population mean values is ***ci = [0.0524, 1.4905]*** which includes the theoretical difference of "1". This shows that the difference $\mu_a - \mu_b > 0$, or that $\mu_a > \mu_b$. This means that the microorganism *Lactobacillus bulgaricus* is more efficient in producing the same batch of the product as it takes less time in comparison to the microorganism *Lactobacillus acidophilus*. We calculate the means using the following functions in MATLAB:

```
mean_a = mean(sample_a);
mean_b = mean(sample_b);
```

The output shows that the *mean_a = 7 h*, while *mean_b = 6.2* h. The variable mean_a corresponds to the time in hours it takes *Lactobacillus acidophilus* to make a batch of the product, while mean_b to *Lactobacillus bulgaricus*.