

Data Handling Skills: Exercises for use with MS Excel

The final grades of 80 biology students are recorded in the following table:

68	84	75	82	68	90	62	88	76	93
73	79	88	73	60	93	71	59	85	75
61	65	75	87	74	62	95	78	63	72
66	78	82	75	94	77	69	74	68	60
96	78	89	61	75	95	60	79	83	71
79	62	67	97	78	85	76	65	71	75
65	80	73	57	88	78	62	76	53	74
86	67	73	81	72	63	76	75	85	77

1. Enter the data in an Excel sheet, using a single column and find:

- The sum.
- The lowest grade.
- The highest grade.
- The range.
- The sample mean or average.
- The mode.
- The median.
- The sample variance and standard deviation.

2. Using a new sheet, calculate the frequency for each value in the range of the data set above and plot the frequency distribution. Clue: introduce in a column all the value in the range, i.e., 53,54...97. Calculate in a second column the counts of each of these values that appear in the table above. Use the chart wizard to plot the range against the frequencies.

3. In Excel, construct a frequency histogram using the following bins:

50-54																			
55-59																			
60-64																			
65-69																			
70-74																			
75-79																			
80-84																			
85-89																			
90-94																			
95-99																			

Clue: check first to which row of the table above each grade belongs and fill in an empty square. Turn the page 90 degrees, leaving the left side of the table at the bottom. This is how the histogram has to look like. Now using Excel, write in a column the right limit of the bins. In the next column, calculate the count. Plot column 1 against the counts.

4. Using a new sheet, let us do the same as in the previous exercise but using relative frequencies instead.

Clue: The frequencies calculated previously have to be divided by the sample size or total number of data available.

5. Using the same sheet, we are now going to plot the relative frequency density. For a probability density histogram the total sum of the areas of the boxes has to be equal to 1.

Clue: The relative frequencies calculated in the previous exercise have to be divided by the width of each bin (class width). Note that it is constant and equal to 5 for all the classes.

6. By using your calculator check that the area of the relative frequency density is equal to 1.

Clue: the area of a rectangle is the base multiplied by the height.

Now use Excel to do this calculation.

Clue: the relative frequencies are the areas of the boxes.

7. Compare the distribution of grades with a normal distribution using the mean and standard deviation calculated in Exercise 1. First try this graphically through the histograms. Follow these steps:

- a. Open a new sheet and copy in a column the right limit of the bins used in Exercise 3.
- b. Calculate for the each of these numbers the Normal distribution with the mean and standard deviation from Exercise 1.
- c. Plot the values calculated in b. against the bin limits of a.
- d. Copy the plot of Exercise 6 and paste it on the plot calculated in c.

Do you think that there is a significant difference between both histograms?

8. Using the histogram of Exercise 6, calculate the probability of the intervals: [60,64], [75,79], [83,86], [92,99].

Clue: In a relative frequency density, the probability of an interval is the area, corresponds to the area covered by the histogram.

9. Calculate the probability of the same four intervals for the normal distribution of Exercise 7.

10. By looking at the results from Exercises 7,8, and 9, do you think that the grades are normal distributed? Check your conclusion by passing the Chi-square test.

Clue: Use the Chi-square test provided by Excel to get the p-value, then check whether it is larger or smaller than 0.05.

11. Using your calculator, determine the estimator of the standard error for the grades. In a new Excel sheet calculate the 90%, 95% and 99% confidence intervals for the mean of the grades by using the t-distribution. What is the meaning of these intervals? What can we expect and with what confidence if we chose randomly a student and ask for the grade?

12. Calculate the same confidence intervals by using a normal distribution. Compare these intervals to those obtained in the previous exercise. Explain the difference.

13. Using the one-sample t-test, check whether the sample mean is equal to an expected value of 75.

14. Using the one-sample t test check whether the sample mean is equal to an expected value of 50.

15. We want to compare the level of these biology students with a different group whose grades are given in the following table:

51	81	75	82	68	90	62	55	76	55
73	79	65	73	60	64	71	59	54	65
61	65	75	56	74	62	95	78	63	72
66	52	79	75	66	66	69	74	77	60

Plot the histogram for these data as in Exercise 5 and compare the distribution of both groups of students. Are the mean values *not* significantly different for both groups? Check your answer by applying the two-sample t-test for the means.

Clue: Calculate the t-value and the p-value as explained in the notes and compare the p-value with the t statistic for a level of 0.05.

16. In the following table salaries for 10 couples are listed:

Men	Women
20	15
30	35
30	25
20	25
20	25
30	15
40	25
30	25
40	35
40	25

Using Excel:

- Calculate the mean, variance and standard deviation for both groups.
- Calculate the mean, variance and standard deviation for both groups together.
- Calculate the mean, variance and standard deviation for the difference of salaries.
- By looking at the result so far obtained, could you say that there is a salary discrimination against women?
- Calculate the covariance and correlation between both variables. Are they independent?
- Produce a scatter plot of the variables 'Men' against 'Women' and fit a regression line. According to this line, which is the salary of a woman whose partner earns 35?

17. The mean and variance of the number of contaminants in a bacterial vaccine preserved with phenol are 63 and 25 respectively. A sample of twenty bacterial vaccines in a different preservative gave the following contamination counts

62	60	60	54	55	55	55	60	57	56
60	60	55	60	54	54	52	58	60	67

Assuming that the counts are normally distributed, has the preservative significantly changed the contamination counts?

18. The activity of an enzyme (units/gram protein) in twelve tissues infected with hepatitis and eighteen normal liver tissues was as follows:

Hepatitis liver tissue	4.15	4.48	4.22	3.94	4.52	3.70
	4.77	4.03	3.90	4.86	3.16	3.33
Normal liver tissue	3.15	4.23	3.12	3.5	3.99	4.40
	3.86	3.86	3.16	4.27	4.34	3.79
	4.28	4.63	4.98	3.52	2.77	3.18

Is there a significant difference in enzyme activity?