Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

# TRINITY BUSINESS SCHOOL

## COVER SHEET

**Module: Foundations of Business Analytics**

**Module Code: BU7142-202223**

**MSc Programme: Business Analytics**

This sheet **must** be attached to your assessment. **The onus is on the student to keep a copy of all submissions.**

I have read and understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at: http://www.tcd.ie/calendar

I have completed the Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write

| Student Number | Student Surname (BLOCK CAPS) | Student name |
|---|---|---|
| 22326199 | COBOS | José Ramón |

**BU7142 Foundations of Business Analytics**
**Individual Assignment 2022-23**

**1. [15]** You have an online shop on Amazon selling wireless keyboard and mouse. A data set containing 100 online visitors to your shop shows the following:

| **Buy a Mouse** | **Buy a Keyboard** | |
|---|---|---|
| | Yes | No |
| Yes | 25 | 20 |
| No | 10 | 45 |

**(a) [3]** What is the probability that a customer buys a keyboard and a mouse?
25%, because you have 25 customers that have bought the buy and the keyboard and 100 customers in total so the probability is 25/100.

**(b) [3]** What is the probability that a customer makes a purchase in your shop?

$P(M \cup K) = P(M) + P(K) - P(M \cap K)$

The probability that a customer buys in your store is the probability that he buys either a mouse or a keyboard, so it is calculated by the union of two events.

$P(M) = 45\% \;||\; P(K) = 35\% \;||\; P(K \cap M) = 25\%$
$P(M \cup K) = 45\% + 35\% - 25\% = 55\%$

**(c) [3]** Given that a customer buys a mouse, what is the probability that the customer also buys a keyboard?

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

That is because we want to know of all the customers who have bought a mouse (45), how many of them have also bought a keyboard, i.e. the intersection (25).

$P(K/M) = P(K \cap M) / P(M) = 25 / 45 = 56\%$

**(d) [6]** Are the events, buying mouse and buying keyboard, independent? Why or why not?

We can provide the independence of two events by comparing $P(M \cap K) = P(M) * P(K)$

$P(M \cap K) = 25\%$

P(M) * P(K) = 45% * 35% = 16%.

We can say that these two events aren't independent.

**2. [16]** Suppose that you have €1,000,000 and you are contemplating the purchase of two investments, Tesla and Zoom. One year from now, Tesla can be sold at €X per €1 invested, and Zoom can be sold for €Y per €1 invested. You regard X and Y to be statistically independent random variables. X and Y both are normally distributed with mean of 1.1 and standard deviation of 0.1.

**(a) [4]** If you put all your money in Tesla, what is the probability that you will be able to sell it one year from now at a positive profit?

The probability to sell it one year from now at a positive profit is P(x > 1) and as X ~ N(1.1, 0.1) We can standardize the distribution to calculate that probability as a normal distribution (0, 1).

P(X > 1) = P(Z > J) || J = (1 – 1.1) / 0.1 = -1

P( Z > -1) = P(Z < 1) = 84%.

**(b) [4]** If you split your money evenly between the two investments, what is the probability you will be able to sell your portfolio a year from now at a positive profit?

As X ~ N(1.1, 0.1) and Y ~ N(1.1, 0.1) and are independent, you can sum them and get a new random variable. We can calculate the μ and Std of the new variable like this:

$$E(L)=a_1E(X_1)+a_2E(X_2)+\ldots+a_nE(X_n) \text{ and}$$
$$V(L)= a_1^2V(X_1)+a_2^2V(X_2)+\ldots+a_n^2V(X_n).$$

In our example a1 and a2 are 0.5, so the new random variable L has E(L) = 1.1 and Std(L) = 0.1

And the result is the same because the new random variable L is distributed the same as X and Y.

P(L > 1) = 84%

**(c) [8]** What is the optimal allocation strategy between the two investments? And Why? Please provide calculation and/or explanation.

As we saw in the previous exercise it doesn't matter how to combine as they are independent and the same distributed random variables.

**3. [16]** You are interested in the average age of the employees in a large company. The relevant data are not easily obtainable for the entire company, so a sample must be taken. It is assumed that the standard deviation of the age is 10 years.

**(a) [8]** How large must the sample be, in order for an 85% confidence interval to be no more than two years wide?

to make the CI at 85% 2 years, we clear the n from the CI equation by equaling 1 to make it 1 year above and 1 year below to make it a 2-year interval.

CI85% = $\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

$z_{\alpha/2} = 1.44$ $\sigma = 10$

$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1$

n = 207.36

We need a sample of 208 people or more to be able to create a 2-years CI at 85%

**(b) [8]** If a 95% confidence interval for the population mean of the age that is calculated from a particular sample is [42.1,42.7], what is the sample mean and sample size for this particular sample?

The mean of the sample is 42.4, because the CI95% is [42.1, 42.7], so we have a 0.6 interval leaving 0.3 below and 0.3 above.

CI95% = $\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

$z_{\alpha/2} = 1.96$ $\sigma = 10$

$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 0.3$

n = 4268.44

And the sample size is 4268.44 because we have to do the same as in the previous exercise but with a 0.6 CI95%

**4. [10]** There are three doors. Behind one of them is a prize. Monty Hall, the host, asks you to pick a door. Suppose you pick Door #1. Monty opens one of the other two doors (say Door #3) and show, there is nothing behind Door #3. He gives you the choice of either sticking with your original choice, #1, or switching to Door #2. Use probability theory to explain and prove which strategy, switch or not switch, is better.

The probability of choosing at the first opportunity the door that hides the car is 1/3, so the probability that the car is in other doors is 2/3.

If I keep my original choice I win if I originally chose the car (with probability 1/3), while if I change, I win if I originally chose one of the two goats (with probability 2/3). Therefore, I must change my choice if I want to maximize the probability of winning the car.

**5.[22]** A souvenir store manager in Dublin would like to know how the price can affect the sales of souvenirs. In his data, he obtained 1000 pairs of prices (unit: €) and volume of sales (unit: hundreds) for the souvenirs.

A simple linear regression was run with Sales as the dependent variable and Price as the independent variable with the following results.

Regression Statistics

| R | R Square | Adj. RSqr | StErr of EST |
|---|---|---|---|
| 0.676 | 0.457 | 0.441 | 5.184 |

Summary Table

| Variable | Coeff. | Std. Err |
|---|---|---|
| Intercept | 72.168 | 7.332 |
| Price | -2.147 | 0.599 |

**(a) [3]** Write down the estimated regression equation. Explain this regression equation.

Sales (in hundred) = 72.17 -2.15 * price (in €)

Whit this regression we can deduce that he is going to sell a mean of 7217 units of a product independently to the price with a standard deviation of 733 units, and every increasing euro is going to make the sales decrease by 215 units on average, with a standard deviation of 0.59

**(b) [3]** What's the value for R Square? Explain the meaning of this value.

R2 = 0.457

It means that our linear regression explains 45.7% of the variation of the real data.

**(c) [6]** Test the significance of the model at α=0.1 level. Write down the hypothesis you are testing, the statistic you use to evaluate the test, and your conclusion.

The null hypothesis h0 is that the parameter is not significant, and h1 is the opposite.
The statistic is -3.58, calculated as the estimate b1 / std b1.

T statistic = estimate b1 / std b1 || -2.147 / 0.599 = -3.58

Comparing the statistic with the Z α/2 we can conclude that the parameter is significant at α = 0.1 level.

**(d) [4]** Use the regression output to construct a 99% confidence interval for the predicted sales when the price is set to be 15.

CI99% = $\hat{y} \pm 1.96 \times \text{standard error of the estimate}$

Sales = 72.17 – 2.15 * 15

Sales = 39.97 (in hundred)

39.97 +- 2.58 * 5.184 = [26.59, 53.34]

**(e) [6]** A colleague in the toy store claims that £1 increase in the price will lead to less than 2.5 (hundred) decrease in sales. Is there sufficient evidence to reject that claim at α=0.05 level?

No, because we must use the left-tailed test since we must know if -2.5 is outside the probability set by the 95% area. And for that, we must know either the t statistic to compare it with the value in the tables or the p-value to compare it with the alpha, but we cannot calculate either.

**6.[21]** In this course, we have studied the impact of GDP and Population on a country's performance in summer Olympic Games. In addition to GDP and Population, there are other important factors that may affect a country's performance. Please choose **THREE factors** and briefly discuss your research plan to study the impact of these factors on a country's performance in summer Olympic Games, i.e., which statistical tools, models, data and variables, etc. you will need and why. Please note that you should focus on the explanation and justification of your research plan and do NOT need to actually collect the data or carry out the test.

To carry out this study I have selected the variables of well-being index, degree of urbanization and geographical and climatological conditions.

To begin, I would compile the information and data that I need to carry out the study, considering the veracity of the sources, as well as the importance of the data found in the different databases.

Once the data has been collected, an initial model would be proposed to try to explain

sports performance through the previously mentioned variables. This first model would consist of a linear regression to try to find out if there is a linear relationship between the variables. I would use excel if I don't have a very large database.

Once the model has been built, I would pay special attention to the signs of the coefficients of the parameters of this first model, to see if the relationship I am looking for is the one it really has. After verifying that the relationship is correct, I would proceed to evaluate the significance of each of the parameters, proposing significance hypotheses for each of them and thus ensuring that all the variables that I am using are useful when explaining sports performance. Finally, once the initial linear regression model has been established and verified, I would try to improve this model by looking for non-linear relationships in order to improve the significance of the parameters and to increase the R squared, which is the indicator that summarizes how well my model explains the dependent variable, in this last step, I would also try to locate possible errors that my model may have, such as multicollinearity.