**TABLE OF CONTENTS**

[**QUESTION 1** 2](#_Toc99527728)

[**QUESTION 2** 6](#_Toc99527729)

[**QUESTION 3** 9](#_Toc99527730)

# **QUESTION 1**

(A). First step is to find the margin of error for the population mean life of compact fluorescent light bulbs in this shipment at the 95% level of confidence.

Since sample size (n) = 81 ≥ 30, the formula for the margin of error is

where, e = margin of error

Z = Z-value

α = level of significance

𝜎 = population standard deviation

n = sample size

Knowing that the level of confidence is 95% = 0.95

1 − α = 0.95, you also know that α = 0.05. Therefore, using the Z−table, you will find the z−value;

= 1 – α

= 0.95

= 0.95

= 0.975

Therefore, = 1.96

and the margin of error is

The 95% confidence interval estimate for the population mean life of compact fluorescent light bulbs in this shipment is;

= 0.95

Where, = sample mean

e = margin error

= population mean

= 0.95

= 0.95

Therefore, you can say that the population mean life of compact fluorescent light bulbs in this shipment will be between 7382.22 hours and 7817.78 hours with 95% level of confidence (or with a 5% chance of being wrong).

(B). Based on the 95% confidence interval in (a),

The minimum mean life estimate (lower confidence limit) should be 7382.22 and

The maximum mean life estimate (upper confidence limit) should be 7817.78.

From the calculations, a good estimate for the mean life of the bulbs should only be between this confidence interval and therefore, the manufacturer has **NO** right to state that the compact fluorescent light bulbs have a mean life of 8,000 hours.

(C). Yes, it can be assumed that the population compact fluorescent light bulb life is normally distributed. This is because the sample size (n) is large enough to apply the Central Limit Theorem (n >= 30)

Where n = 81.

In probability theory, the central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution as the sample size becomes larger. As the sample size gets bigger and bigger, the mean of the sample will get closer to the actual population mean.

Standard Normal Distribution is a type of probability distribution that is symmetric about the average or the mean, depicting that the data near the average or the mean are occurring more frequently when compared to the data which is far from the average or the mean. A score on the standard normal distribution can be termed as the “Z-value”.

(D). First step is to find the margin of error for the population mean life of compact fluorescent light bulbs in this shipment at the 95% level of confidence.

Since sample size (n) = 81 ≥ 30, the formula for the margin of error is

where, e = margin of error

Z = Z-value

α = level of significance

𝜎 = population standard deviation

n = sample size

Knowing that the level of confidence is 95% = 0.95

1 − α = 0.95, you also know that α = 0.05. Therefore, using the Z−table, you will find the z−value;

= 1 – α

= 0.95

= 0.95

= 0.975

Therefore, = 1.96

and the margin of error is

The 95% confidence interval estimate for the population mean life of compact fluorescent light bulbs in this shipment is;

= 0.95

Where, = sample mean

e = margin error

= population mean

= 0.95

= 0.95

Therefore, you can say that the population mean life of compact fluorescent light bulbs in this shipment will be between 7425.78 hours and 7774.22 hours with 95% level of confidence (or with a 5% chance of being wrong).

Based on the 95% confidence interval,

The minimum mean life estimate (lower confidence limit) should be 7425.78 and

The maximum mean life estimate (upper confidence limit) should be 7774.22.

From the calculations, a good estimate for the mean life of the bulbs should only be between this confidence interval and therefore, the manufacturer has **NO** right to state that the compact fluorescent light bulbs have a mean life of 8,000 hours.

# **QUESTION 2**

(A).

H0 : difference = 0

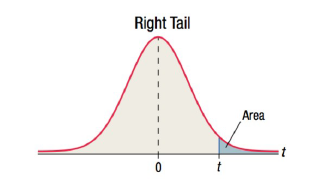
H1 : difference ≠ 0

This means that;

The null hypothesis (H0) states that there is no difference between the average hourly wage of men and women.

While the alternative hypothesis (H1) states that there is a difference (pay gap) between the average hourly wage of men and women.

(B).



The t-value of the regression test is **-4.95**, this is the test statistic.

The system of hypothesis is a right-tailed test.

H0 : average hourly pay for men = average hourly pay for women

H1 : average hourly pay for men > average hourly pay for women

To calculate the critical value (CV) = Zα

Knowing that the level of significance (α) = 5% = 0.05

So, CV = Z0.05

The t-distribution table helps to find the critical value of this right-tail distribution.

CV = 1.64

Knowing that the t-value of -4.95 is further away from the critical value of 1.64 or rejection region. We have to fail to reject the Null hypothesis. This means that at the 5% level of significance, we cannot affirm that the average hourly wage of men is significantly higher than the average of women.

(C). All coefficients are jointly estimated, so during a multiple regression, it is expected that every new variable will change the other coefficients of the variables already in the model. So, by including the “degree” variable, it can be predicted that the coefficient for the male (gender) variable will surely be affected. Multiple regression model is more realistic because the “male” variable alone cannot possibly be the only variable/factor affecting hourly wage in a real-life situation.

At a significant level of 0.05, it can be seen in the second regression result that the gender and qualification of an individual significantly affects their hourly pay – this is obvious since the p-values of these two variables (which is 0.000) are less than 0.05.

Adding new variables to a model generally increases R2, which helps to improve the overall goodness of fit of the model. It is important to note that when the “degree” variable was added to the second regression, the R2 slightly increased from 0.0260 to 0.0749. The R2 increased by 0.0489 (4.89%) – this is a good sign that the model has improved.

In the first model, R2 is 0.0260 which means 2.6% of variation in hourly wage is accounted for by the “male” (gender) variable of whether the individual is a male or female. The remainder of the variation in hourly wage (97.4%) is affected by other factors (u).

In the second model, R2 is 0.0749 which means 7.49% of variation in hourly wage is accounted for by the “male” variable (of whether the individual is a male or female) and the “degree” (whether the individual has a degree or not). The remainder of the variation in hourly wage (92.51%) is affected by other factors (u).

(D).

Based on the results in Table 3, below is the regression model,

rwage = 12.84 + 3.22male + 5.77degree + u

Interpreting the estimated parameters of this model

̂1. If male=0 and degree=0, the predicted wage rate is £12.84 per hour. So, this means that regardless of an individual’s gender or whether or not they have a qualification, their wage is constant at £12.84 per hour.

2. Holding the other variables including “degree” constant, being a male individual can increase the wages of an individual by £3.22 per hour (i.e. on average).

3. Holding the other variables including “male” constant, having a degree can increase the wages of an individual by £5.77 per hour (i.e. on average).

By looking at the regression model, the p-values of 0.000 are less than 0.05, and so we can say that the two dependent variables (“male” and “degree”) are significantly affecting the hourly rate and so, we can reject the null hypothesis and indeed claim that the average hourly wage of the men are significantly higher than that of the women – so there is a gender pay gap.

By including the “degree” variable however, we can see that this also affects the pay rate and it has to also be considered.

The R2 is low at 0.0749 and this calls our attention to the need to look into other variables that may actually be affecting or causing a pay gap.

To further examine the gender pay, a number of other factors need to be considered like the years of experience, whether permanent/part-time employee, age, and so on. There is still a number of variables that can be added that can even strengthen the R2 and improve the goodness of fit of the model.

# **QUESTION 3**

(A).

Based on the results in Table 5, below is the regression model,

rlnwage = 2.24 + 0.16Degree + 0.058Experience - 0.001SquaredExperience - 0.024Worktraining + u

(B). The R2 explains the overall best fit of the model.

R2 is 0.1206 which means 12.06% of variation in log of hourly wage is accounted for by the “Degree”, “Experience”, “SquaredExperience”, and “Worktraining” variable. So factors such as if the person has a degree, years of experience or on-the-job training are all represented.

However, the R2 is low and for the regression model to improve or have a better fit, other explanatory factors/variables that can potentially affect the log of the hourly wage needs to be added. These factors will increase R2 and account for the remaining 87.94% that will explain variations in the dependent variable.

(C).

Interpreting the estimated parameters of this model

̂1. Holding all dependent variables constant, the predicted log of hourly wage for each individual is £2.24 per hour. So, this means that regardless of the other factors/variables, their wage is constant at this amount. Because the p-value (0.000) for the constant is lower than 0.05 (at 95% confidence), the constant is statistically significant in the model.

2. Holding the other variables constant, being a Degree can increase the hourly wages by 0.16%, on an average (i.e. on average). Because the p-value (0.001) for the “Degree” variable is lower than 0.05 (at 95% confidence), the variable is statistically significant in the model.

3. Holding the other variables constant, an additional year of experience increases the hourly pay by 0.058% on average (i.e. on average). Because the p-value (0.000) for the “Experience” variable is lower than 0.05 (at 95% confidence), the variable is statistically significant in the model.

4. Holding the other variables constant, an additional unit of squared years of experience decreases the hourly wages by 0.001%, on an average (i.e. on average). Because the p-value (0.001) for the “SquaredExperience” variable is lower than 0.05 (at 95% confidence), the variable is statistically significant in the model.

5. Holding the other variables constant, receiving the on-the-job training decreases the hourly pay by 0.024% on average (i.e. on average). However, because the p-value (0.602) for the “Worktraining” variable is higher than 0.05 (at 95% confidence), the variable is not statistically significant in the model.

Overall, the p-value (0.000) of the whole model is lower than 0.05 (at 95% confidence) and is statistically significant. However, the R2 of 0.1206 can be considered as low and new explanatory variables that could be affecting the outcome of rlnwage need to be included in the multiple regression model.

To calculate the average after how many years students can expect the highest earning;

rlnwage = ……………………………….equation 1 if all other variables are constant

rlnwage = 2.24

rlnwage = ………………….equation 2 if all other variables are constant

2.24 = 0.058 x

(D).

Holding the other variables constant, receiving the on-the-job training decreases the hourly pay by 0.024% on average (i.e. on average). However, because the p-value (0.602) for the “Worktraining” variable is higher than 0.05 (at 95% confidence), the variable is not statistically significant in the model.

As the coefficient of the “Worktraining” variable is negative and also insignificant, it is clear that it does not have significant effects on the hourly pay of individuals. This can also be because receiving an on-the-job training is regarded as voluntary so this means whether or not, an individual does the training, he/she does not need to be concerned that this would affect their wage.

Because of this, it is safe to agree that this variable (“Worktraining”) can be removed to improve the validity of the model. Removing the variable might even help to improve or increase the R2, thereby improving the overall goodness of fit of the multiple regression model.

(E).

Run all the mean values from Table 4 through the regression model;

rlnwage = 2.24 + 0.16Degree + 0.058Experience - 0.001SquaredExperience - 0.024Worktraining

rlnwage = 2.24 + 0.16(0.41) + 0.058(15.99) - 0.001(362.08) - 0.024(1.59)

rlnwage = 2.24 + 0.0656 + 0.92742 - 0.36208 – 0.03816

**rlnwage = 2.83**.

To check if the result is also within a 95% confidence interval;

Since sample size (n) = 588 ≥ 30, the formula for the margin of error is

where, e = margin of error

Z = Z-value

α = level of significance

𝜎 = population standard deviation

n = sample size

Knowing that the level of confidence is 95% = 0.95

1 − α = 0.95, you also know that α = 0.05. Therefore, using the Z−table, you will find the z−value;

= 1 – α

= 0.95

= 0.95

= 0.975

Therefore, = 1.96

and the margin of error is

The 95% confidence interval estimate for the mean of log of hourly wages is;

= 0.95

Where, = sample mean

e = margin error

= population mean

= 0.95

= 0.95

Therefore, you can say that the log of hourly wage will be between 2.703 and 2.797 with 95% level of confidence (or with a 5% chance of being wrong).

**This means that the estimated mean log of hourly wage (rlnwage = 2.83) for when all the coefficients of the explanatory variables are considered, does not fall within the confidence interval.**

**This is to conclude that the estimators under-estimate of the return to degree.**

A regression model is underestimated if one or more important predictor variables is missing, thereby producing biased regression coefficients and, overestimating the mean of square error which can be seen on the regression result table. A biased variable is a variable that has significant variables in the error term and omitted from the regression model. In Table 5, the multiple regression analysis is underestimated because the r2 is only explaining 12.06% of the variables which means it is closer to 0 than 1. This implies that important variables are missing to estimate.

Usually, if the estimators do not give an estimate, it is important to include other factors or variables that might actually help to give a better estimate of the independent variable. A number of other factors need to be considered like the position of the individual in the company, whether permanent/part-time employee, age, and so on. Including more variables to the regression model will help to increase R2 as well and improve the fitness of the regression line.