

DATA, INFERENCE & APPLIED MACHINE LEARNING (COURSE 18-785)

Names : Lucie Niyomutoni

AndrewID: lniyomut

**USED LIBRARY**

* **Pandas**
* **Matplotlib**
* **Xlrd**
* **scipy**

Cisco@220004209

# QUESTION 1

## STEPS

1. **Formulating Hypotheses**

The first step is to establish both the **null hypothesis and** the **alternative hypothesis**:

* The **null hypothesis (H₀)** assumes that there is no significant difference between the women’s energy intake and the recommended value of **7725 kJ**. In other words, the average energy intake is equal to 7725 kJ.

**H₀: mu = 7725 kJ**

* The **alternative hypothesis (H₁)** assumes that the women’s energy intake does differ from the recommended value, meaning it could be either higher or lower. Therefore, we have a two-tailed test.

**H₁: mu # 7725 kJ**

1. **Select the Significance Level**

We use a standard significance level of **alpha = 0.05**. This means that we are willing to accept a 5% chance of rejecting the null hypothesis when it is actually true. If the p-value from the test is below 0.05, we reject the null hypothesis; otherwise, we do not.

1. **Determine the Type of Test**

Since the alternative hypothesis suggests the energy intake could be either higher or lower than the recommended 7725 kJ, we use a two-tailed test. This test checks for any significant difference in either direction from the hypothesized mean.

1. **Calculate the Sample Mean**

We calculate the sample mean to understand the average daily energy intake for the 11 women. The sample mean represents the central tendency of the given data.

1. **Identify the Population Mean**

The population mean is the recommended energy intake of 7725 kJ. This is the value we are comparing the sample mean to, to test if the sample mean is statistically different.

1. **Calculate the Sample Standard Deviation and Standard Error of the Mean (SEM)**

The **sample standard deviation** measures how much the individual values deviate from the sample mean, reflecting the variability in the data. We then calculate the **standard error of the mean (SEM)**, which shows how much the sample mean is expected to vary from the true population mean. It is calculated by dividing the sample standard deviation by the square root of the sample size.

1. **Compute the t-Statistic**

Using the formula for the t-statistic, we determine how many standard errors the sample mean is away from the population mean. The formula is:

This gives us the test statistic, which we will compare against the critical value for a two-tailed test.

1. **Determining Degrees of Freedom**

We calculate the degrees of freedom (df) for this test, which is the sample size minus 1. This value is used to determine the critical value of t from the t-distribution table.

df = n – 1

1. **Calculate the p-Value**

The p-value indicates the probability of obtaining a test statistic as extreme as, or more extreme than, the one observed, assuming that the null hypothesis is true. If the p-value is less than the significance level (alpha = 0.05), we reject the null hypothesis.

## RESULT

1. **Sample Mean** :The average energy intake for the 11 women is **6753.64 Kj**
2. **Sample Standard Deviation:** The variation in the energy intake values is **1142.12 kJ.**
3. **Standard Error of the Mean (SEM):** The expected variability of the sample mean is **344.36 kJ**.
4. **t-Statistic** : the calculated t-statistic is **-2.82**, showing how far the sample mean is from the population mean in terms of standard errors.
5. **Degrees of Freedom:** The degrees of freedom for this test is **10**.
6. **p-Value**: The p-value obtained is **0.0181**, which is less than the significance level **(alpha = 0.05).**

Since the p-value is lower than 0.05, we reject the **null hypothesis.** This means the data provides sufficient evidence to suggest that the average energy intake of the women differs significantly from the recommended 7725 kJ.

## INSIGHT

The results indicate that the **average energy intake of the 11 women is significantly lower than the recommended value of 7725 kJ**. This has important real-world implications:

1. **Health Considerations**: A lower-than-recommended energy intake can lead to undernutrition, which may result in fatigue, weakened immunity, and reduced muscle mass. If this trend persists, it could contribute to long-term health problems.

2. Dietary Recommendations: The results suggest that the women may not be consuming enough energy to meet their daily requirements. Dietary adjustments might be necessary to increase caloric intake, which could involve more nutrient-dense foods such as whole grains, protein-rich foods, and healthy fats.

3. Public Health Implications: On a larger scale, if this pattern of insufficient energy intake is common among a wider population, it may indicate a need for nutritional education programs. Public health initiatives could focus on improving awareness about balanced diets and caloric needs, especially among women.

4. Further Investigation: It would be beneficial to explore the reasons behind this lower energy intake. Is it due to economic factors (inability to afford sufficient food), cultural reasons (dietary restrictions), or personal choice (dieting or lack of appetite)? Understanding the root cause can help design more targeted interventions.

In summary, the evidence suggests that women in this sample are not meeting their recommended daily energy intake, highlighting the need for attention to dietary habits and public health efforts aimed at promoting proper nutrition.

# QUESTION 2

## STEPS

1. **Hypothesis Formulation**

* Null Hypothesis( ): There is no significant difference between the mean GOES scores for Guinness served in Ireland and elsewhere.

: =

* **Alternative Hypothesis ():** The mean GOES score for Guinness served in Ireland is significantly higher than that for Guinness served elsewhere. :>

1. **Significance Level**

I used a significance level of , which means we will reject the null hypothesis if the p-value is less than 0.05

1. **Appropriate test selection**

* Since I am comparing the means of two independent samples (Guinness served in Ireland vs. Elsewhere) , a two-sampled t-test is the correct test
* The samples are independent , as the pints consumed in Ireland and elsewhere are separate

1. **Calculating degree of freedom**

The degree of freedom (df) for a two-sample t-test are calculated as follow:

Where

So ,

1. Calculating t-statistic

The formula for the t-statistic in a two-sample t-test is :

,

Where : is the mean GOES score for Ireland

is the mean GOES score for Elsewhere

is the standard deviation for Ireland

is the standard deviation for elsewhere

is the sample size for Ireland

is the sample size for elsewhere

Then I calculated the standard error (SE) and statistics by using and

1. **Determining the P-value**

**To find the p-value for the obtained t-statistic , I used python library**

## RESULT

After going through all of the above steps and running the program the results were



The calculated t-statistic is approximately 11.64. This high value indicates a strong difference between the means of the two groups being compared (Guinness served in Ireland vs. elsewhere).

The degrees of freedom for the test is 101, which is based on the total sample sizes of both groups.

The resulting p-value is approximately 0.00000. This p-value is significantly lower than the chosen significance level of 𝛼=0.05

Based on these results, the null hypothesis is rejected , concluding that there is a statistically significant difference in the mean GOES scores between Guinness served in Irish pubs and those served elsewhere. Specifically, the mean score for Guinness served in Ireland (74) is considerably higher than the average score for Guinness served elsewhere (57).

Since the p-value is significantly less than the alpha level of 0.050.050.05, we reject the null hypothesis. This indicates that there is a statistically significant difference in the GOES scores, supporting the claim that Guinness served in an Irish pub is perceived to taste significantly better than Guinness served elsewhere.

## INSIGHT

The significant difference in enjoyment scores suggests that the cultural context and ambiance of Irish pubs play a crucial role in the experience of drinking Guinness. Factors such as traditional music, the presence of knowledgeable staff, and a sense of community may enhance the overall enjoyment of the beverage

The study also implies that businesses serving Guinness outside of Ireland might consider enhancing their service quality and ambiance to match the enjoyment levels experienced in Irish pubs. This could involve training staff, improving the overall setting, or hosting themed events to replicate the Irish pub atmosphere

Irish pubs can leverage these findings in their marketing strategies. By highlighting the unique aspects of the Irish pub experience, they can attract tourists and locals alike, emphasizing that drinking Guinness in Ireland is not just about the beer but also about the experience

The results may influence consumer preferences, suggesting that people may be willing to travel to Ireland for a more enjoyable Guinness experience. This insight can encourage travel and tourism sectors to promote trips centered around authentic culinary experiences.

# QUESTION 3

## STEPS

1. I first downloaded data from the World Bank Indicators for the year 2013. The data included two key indicators:

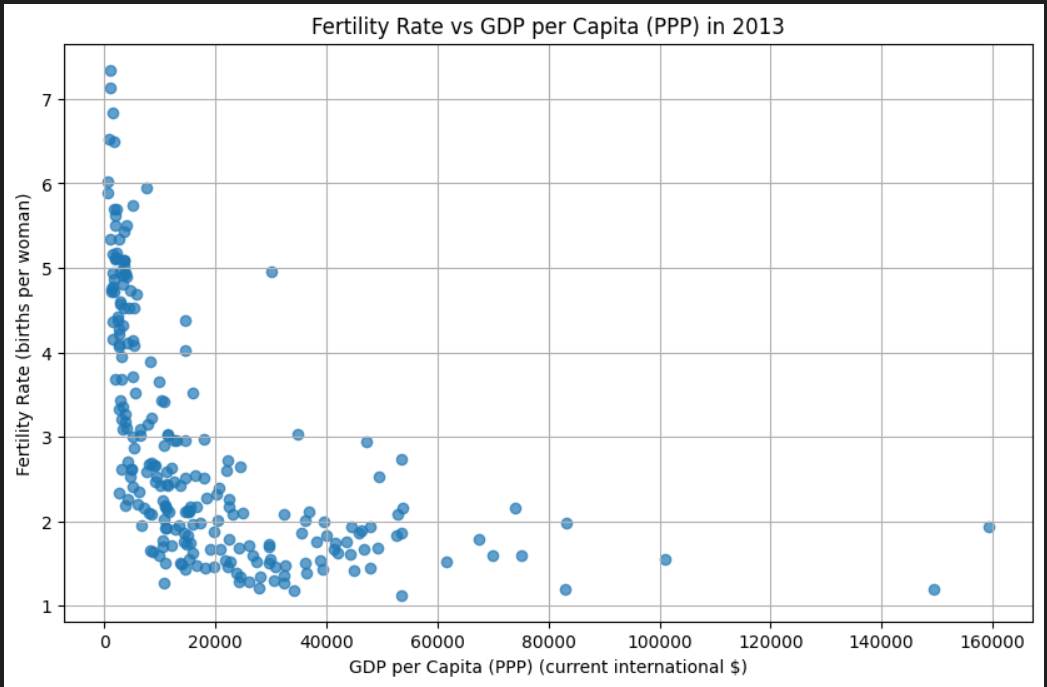
**Fertility rate, total (births per woman**) and **GDP per capita (PPP) (current international $)**

1. I ensured that the necessary Python libraries were installed. Specifically, I used **pandas** for data manipulation and **matplotlib** for data visualization. Additionally, I installed **xlrd** to facilitate the reading of Excel files
2. I used the pandas library to read the Excel files containing the fertility and GDP data
3. I extracted the relevant data for the year 2013 from both datasets
4. I computed the correlation coefficient to assess the strength and direction of the relationship between fertility rates and GDP per capita.
5. I generated a scatter plot to visually represent the relationship between GDP per capita and fertility rates

## RESULT

After going through all the above steps the results were a scatter plot illustrating the relationship between GDP per capita and fertility rates for various countries in 2013, and The correlation coefficient .

**Correlation Coefficient**: After performing the calculations, the correlation coefficient was found to be approximately **-0.5171011715833227**. This indicates a moderate to strong negative correlation between GDP per capita and fertility rates.



The graph above illustrates the relationship between fertility rate (total births per woman) and GDP per capita (current international $), with fertility rate represented on the vertical (Y) axis and GDP per capita on the horizontal (X) axis.

Notably, within the GDP per capita range of $0 to $60,000, many countries exhibit fertility rates between 1 and 3 births per woman. In contrast, countries with GDP per capita between $0 and $20,000 display a wider range of fertility rates, spanning from 1 to 7 births per woman.

This trend clearly indicates that as GDP per capita increases, fertility rates tend to decrease. In other words, wealthier countries generally have fewer children per woman. This observation suggests a negative correlation between GDP per capita and fertility rate: as nations become richer, the average number of children born to women decreases.

## INSIGHT

Countries with higher GDP per capita tend to have better access to education, healthcare, and family planning services. As economies grow, people are more likely to have access to contraception and reproductive health care, enabling them to make informed decisions about family size. For instance, nations like Germany and Japan, which have high GDP per capita, also have low fertility rates, as families often prioritize education, careers, and a higher quality of life over having many children.

One of the key factors contributing to lower fertility rates in wealthier countries is increased access to education, especially for women. Educated women often marry later, have fewer children, and invest more in each child's development and education.

Higher GDP per capita also leads to improved healthcare systems, which reduce infant mortality rates. As a result, families no longer feel the need to have as many children to ensure that some survive. This phenomenon is more common in lower-income countries, such as Sub-Saharan Africa.

In high-income countries, people often prioritize their careers and economic opportunities over having large families. The desire to maintain a certain lifestyle and higher living standards can lead couples to delay having children or limit family size

QUESTION 4