**Day 3 Individual Assignment**

|  |
| --- |
| Name: |

**Use the format “YourName\_Assignment3.docx” for assignment submission.**

**Day 3 Part I. Hypothesis Testing**

Q1. For the following questions, please suggest suitable null and alternative hypothesis.

|  |
| --- |
| **Business Context 1: Financial Value of Global Arms Trade** |
| Stockholm International Peace Research Institute (SIPRI) manages the SIPRI Arms Transfers Database, which records all international transfers of major conventional arms since 1950.  The Trend Indicator Value (TIV) data produced by the Database is designed to serve as an indicator of the volume of military equipment transferred in the world. However, there are significant limitations in using the data to estimate the TIVs. First, the data sets used are based on different definitions and methodologies and are thus not directly comparable. Second, several states (e.g. the United Kingdom and the USA) do not release data on arms exports but only on arms export agreements and licenses, while other states (e.g. China) do not release any financial data on total arms exports, export licenses, agreements or orders.  Nonetheless, by adding together the data that states have made available on the financial value of their arms exports, as well as estimates for those that only provide data on arms export licenses, agreements or orders, it is possible to estimate the rough total value of the global arms trade. For example, the estimate of the total value of the global arms trade for 2019 was $118 billion. However, you believe that the true figure is much higher.  Ref: <https://www.sipri.org/databases/financial-value-global-arms-trade> |
| Null Hypothesis: The total value of the global arms trade for 2019 was $118 billion  Alternative Hypothesis: The total value of the global arms trade for 2019 was higher than $118 billion |

|  |
| --- |
| **Business Context 2: Weight Loss Program in Local Gym** |
| The COVID-19 pandemic had naturally brought about changes in lifestyle and related behaviors. According to a 2022 Ipsos survey conducted online among a nationally representative sample of 500 adult Singaporeans aged 18 years and above, the average weight gain of adult Singaporeans is 5.7kg, up from 4.8kg reported in the previous 2020 survey. The most significant weight gain was reported among 18 – 24-year-olds with 59% of them saying they have gained weight.  With the loosening of COVID-19 restrictions, a local gym came up with a weight loss program to help its members get back in shape. Based on past data, it is believed that the average weight of the members should be around 90kg based on the above survey conducted. However, after conducting the first fitness session, the fitness trainer believes that the average weight to be much higher. If the average weight of the members is indeed much higher, the gym may have to consider having more frequent or long-duration workouts for its members. A random sample of 15 participants yields an average weight of 96kg and a standard deviation of 8kg.  Ref: <https://www.ipsos.com/en-sg/nearly-4-10-singaporeans-report-gaining-weight-pandemic-more-so-among-young-adults> |
| Null Hypothesis: The average weight of the members is equal to 90kg  Alternative Hypothesis: the average weight of the members is much higher than 90kg |

|  |
| --- |
| **Business Context 3: Gender Pay Gap** |
| The gender gap in pay has remained relatively stable in the United States over the past 15 years or so. In 2020, women earned 84% of what men earned, according to a Pew Research Center analysis of median hourly earnings of both full- and part-time workers. Based on this estimate, it would take an extra 42 days of work for women to earn what men did in 2020.  Much of this gap has been explained by measurable factors such as educational attainment, occupational segregation and work experience. The narrowing of the gap is attributable in large part to gains women have made in each of these dimensions.  Even though women have increased their presence in higher-paying jobs traditionally dominated by men, such as professional and managerial positions, women as a whole continue to be overrepresented in lower-paying occupations relative to their share of the workforce. This may contribute to gender differences in pay.  A survey company randomly sampled 1,500 United States residents and the result indicates that the female workers only earned 81% of their male counterparts. The company argues that the information on gender pay gap provided by Pew Research Center is inaccurate.  Ref: <https://www.pewresearch.org/fact-tank/2021/05/25/gender-pay-gap-facts/> |
| Null Hypothesis: Women earned 84% of what men earned  Alternative Hypothesis: Women not earned 84% of what men earned |

Q2. Today, Data Science as a field of research and practice is quite well-established. Banks, insurance companies, retailers, health care providers, and government agencies have substantial data science groups; large financial services firms may have hundreds of data scientists. Data scientists have been recruited to address societal crises, counting and predicting Covid-19 cases and deaths, and address climate-related disasters.

The dataset provided to you (“ds\_salaries.csv”) consists of 120 randomly sampled data science jobs, aggregated from the online job portal ai-jobs.net. Descriptions of the variables in this dataset are extracted from the codebook into the **Table 1** below:

**Table 1**

|  |  |
| --- | --- |
| **Variable(s)** | **Description** |
| work\_year | The year the salary was paid |
| employment\_type | The type of employment for the role:  PT: Part-time  FT: Full-time  CT: Contract  FL: Freelance |
| job\_title | The role worked in during the year |
| salary\_in\_usd | The salary in USD (FX rate divided by avg. USD rate for the respective year via fxdata.foorilla.com) |
| company\_size | The average number of people that worked for the company during the year:  S: Less than 50 employees (small)  M: Between 50 to 250 employees (medium)  L: More than 250 employees (large) |

Ref: <https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century>   
Ref: <https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries>

It is claimed that the typical full-time Data Scientist’s salary in a large company is USD135,000. Is there enough evidence to support this claim at the 0.05 level of significance?

**Please submit your R code for Question 2 and 3 together using this format “YourName\_Assignment3.R”**

|  |
| --- |
| Null Hypothesis: The typical full-time Data Scientist’s salary in a large company is USD135,000  Alternative Hypothesis: The typical full-time Data Scientist’s salary in a large company is not equal to USD135,000  data\_q2=read.csv('ds\_salaries.csv')  ##########################  # Hypothesis:  # Null Hypothesis: The typical full-time Data Scientist’s salary in a large company is USD 135,000  # Alternative Hypothesis: The typical full-time Data Scientist’s salary in a large company is not equal to USD135,000  ##########################  ## extract data of full time data scientist salary in large company  data\_q2\_2=data\_q2[which((data\_q2$employment\_type == 'FT') & (data\_q2$job\_title=='Data Scientist') & (data\_q2$company\_size == 'L') ),]  ## t-test  t.test(data\_q2\_2$salary\_in\_usd,mu=135000)  #######################  # result  #######################  # One Sample t-test  # data: data\_q2\_2$salary\_in\_usd  # t = -0.39764, df = 13, p-value = 0.6973  # alternative hypothesis: true mean is not equal to 135000  # 95 percent confidence interval:  # 89899.29 166078.85  # sample estimates:  # mean of x  # 127989.1  # p-value = 0.6973 > 0.025  # Accept Null Hypothesis  #######################  # Conclusion  #######################  # The typical full-time Data Scientist’s salary in a large company is equal to USD135,000 |

**Day 3 Part II. One-way and Two-way ANOVA**

Q3. Using the same dataset (“ds\_salaries.csv”) in Q3, answer the two questions below by testing the hypothesis at 95% confidence level:

1. Does the mean salary (in USD) for data science jobs, regardless of employment type and job title, differs by company size? If yes, how does the company size matter?
2. Is there an interaction effect between company size and employment type?

|  |
| --- |
| # Q3.1  data\_q3=read.csv('ds\_salaries.csv')  tapply(data\_q3$salary\_in\_usd,data\_q3$company\_size,summary)  anova\_res\_q3=aov(salary\_in\_usd~company\_size,data = data\_q3)  summary(anova\_res\_q3)  ##########################  # Hypothesis:  # Null Hypothesis: The mean salary (in USD) for data science jobs, regardless of employment type and job title, NOT differs by company size  # Alternative Hypothesis:The mean salary (in USD) for data science jobs, regardless of employment type and job title, differs by company size  ##########################  #######################  # result  #######################  # Df Sum Sq Mean Sq F value Pr(>F)  # company\_size 2 1.627e+10 8.135e+09 2.644 0.0753 .  # Residuals 117 3.600e+11 3.077e+09  # ---  # Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  # P=0.0753 > 0.025  # Accept NULL Hypothesis  #######################  # Conclusion  #######################  # The mean salary (in USD) for data science jobs, regardless of employment type and job title, NOT differs by company size  # Q3.2  anova\_res\_q3\_2=aov(salary\_in\_usd~company\_size\*employment\_type,data = data\_q3)  summary(anova\_res\_q3\_2)  #######################  # result  #######################  # Df Sum Sq Mean Sq F value Pr(>F)  # company\_size 2 1.627e+10 8.135e+09 2.683 0.0727 .  # employment\_type 2 1.270e+10 6.349e+09 2.094 0.1279  # company\_size:employment\_type 1 1.635e+09 1.635e+09 0.539 0.4643  # Residuals 114 3.456e+11 3.032e+09  # ---  # Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  # P=0.4643 > 0.025  # Accept NULL Hypothesis  #######################  # Conclusion  #######################  # There is not an interaction effect between company size and employment type |

**Day 3 Part I & II. Conceptual Understanding**

Q4. Answer all three questions below with clear explanations.

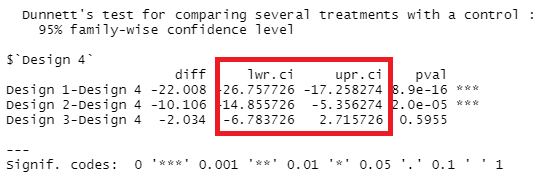
1. Is it possible to have more than two outcomes from hypothesis testing?

|  |
| --- |
|  |

2. When will a big value for α (significance level) make sense in hypothesis testing?

|  |
| --- |
|  |

3. When we perform Multiple Comparison Tests for one-way ANOVA, the outputs came with p-values and also confidence intervals. For example, using Dunnett method, we have obtained the lower and upper confidence limits (“lwr.ci” and “upr.ci”):



Explain the meanings of these confidence limits.

|  |
| --- |
|  |