Assignment: Data Analysis Groupwork

By

Anushka Jain – 22200246
Mingbo Zhang – 22204725
Pratik Pore – 22200170

MIS41130: Statistical Methods

Professor: Dr. Annunziata Esposito Amideo

UCD Michael Smurfit Graduate Business School

25th November 2022

Contents:

1: Abstract……...3

2: Introduction……....3

3: Statistical Analysis……....3

3.1 Descriptive Analysis……..3

3.2 Inferential Analysis……6

3.2.1 Confidence Intervals……6

3.2.2 Hypothesis Testing……..7

4: Conclusion……10

5: Recommendations……...11

6: References……11

7: Contributions……12

1. Abstract/Executive Summary:

In recent years, Ireland's minimum wage and per capita income have risen. In the context of rising overall per capita income, we wanted to investigate average earnings for various types of employees in various economic sectors and identify differences in salary levels across sectors. Each economic sector in the data set is divided into All NACE economic sectors excluding activities A, T, and U (B to S), Business and services (B to N, R, S), Industry (B to E), and Public administration, education, and health, while employees are divided into three types of employees: Production, transport, craft, and other manual workers, Clerical, sales, and service employees, and Managers, professionals, and associated professionals. Descriptive statistics and inferential statistics are our main statistical methods. We use descriptive statistics to describe the estimates of average of earnings of employees in the sample data and use inferential analysis to determine whether the earning of employees in different Sectors are equal. The final research results show that the earning of employees in different departments are not equal. The average earning level of Industry employees is higher. The average earning of employee in Public Administration, education and health sector are generally lower.

2. Introduction:

Ireland has a basic wage of 10.5€ per hour, which is mostly implied for manual laborers because they are the lowest wage earners. To support themselves financially, most students work part-time. Pratik from our group works at a grocery store and earns 10.50€ per hour, but my university peers who work in the public sector, such as the James Joyce Library, earn nearly 14.7€ per hour, or another colleague who works at UPS warehouse earns almost 12.9€ per hour. As a result, we were perplexed that, even though we work in the same industry, but in different sectors, our earnings differed by a great margin. That's why we looked for data on working people's earnings in Ireland from various industries, which led us to the data set.

The 3 sectors which we are targeting are Manual Labors, Service Employees, Managers & Business Professionals. We have downloaded the dataset (O'Riordan, 2022) from the Central Statistics Office's website in Ireland. The data was gathered from an authentic government website and has already been sampled, implying our dataset is a sample dataset and not that of a population.

Our variable is estimates of average earnings in € of 3 different types of employees in 4 different working sectors from 2010 Quarter 2 to 2022 Quarter 1 which is a random continuous variable.

3. Statistical analysis

3.1 Descriptive

During the descriptive analysis, we calculated the Mean, Median, Standard Deviation, Range, Minimum value, Maximum value, and variance. Because our dataset is a time series, arranging the data in ascending or descending order would lose the time significance, and thus are not calculating the interquartile ranges.

i. Production, transport, craft, and other manual workers:

Descriptive Measures of samples		All NACE economic sectors excluding activities A,T and U (B to S)	Business and services (B to N,R,S)	Industry (B to E)	Public administration, education and health (O to Q)
Measures of Centrality	Mean	534.7041667	554.8377083	696.6964583	472.8541667
	Median	522.29	543.03	709.115	467.18
Measures of Variation	Std Deviation	45.17931078	54.08373875	66.75489905	35.24728575
	Range	170.89	208	260.6	135.3
	Min	477.18	483.98	564.37	417
	Max	648.07	691.98	824.97	552.3
	Variance	2041.170123	2925.050797	4456.216547	1242.371152

The average earnings in the industry sector are the highest, while the average earnings in the public sector are the lowest of the four sectors. The median is nearly identical to the sector's mean earnings. The earning range for the industry sector is the highest, while the earning range for the public sector is the lowest.

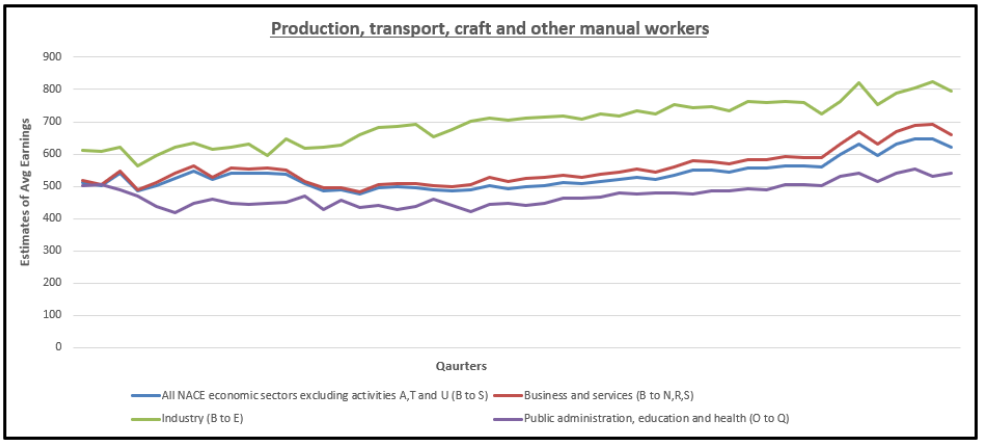

Fig 1

The graph depicts the overall rise in average earnings of serviced employees across all four sectors from 2010 Q2 to 2022 Q1. We can conclude that the Industry (B to E) sector has a high average wage for serviced employees.

ii. Clerical, Sales, and Serviced Employees:

Descriptive Measures of samples		All NACE economic sectors excluding activities A,T and U (B to S)	Business and services (B to N,R,S)	Industry (B to E)	Public administration, education and health (O to Q)
Measures of Centrality	Mean	503.3495833	489.0135417	785.159375	559.9272917
	Median	485.41	464.675	771.595	553.715
Measures of Variation	Std Deviation	45.98369748	56.25762691	57.68618506	32.50666621
	Range	168.87	201.26	264.01	135.21

	Min	452.73	432.61	680.33	508.44
	Max	621.6	633.87	944.34	643.65
	Variance	2114.500434	3164.920585	3327.695946	1056.683348

According to the descriptive analysis for serviced employees, the maximum and minimum average earnings for the Industry sector (B to E) are 785.15 and 489.01, respectively. The lowest average earnings range is in public administration, education, and health care.

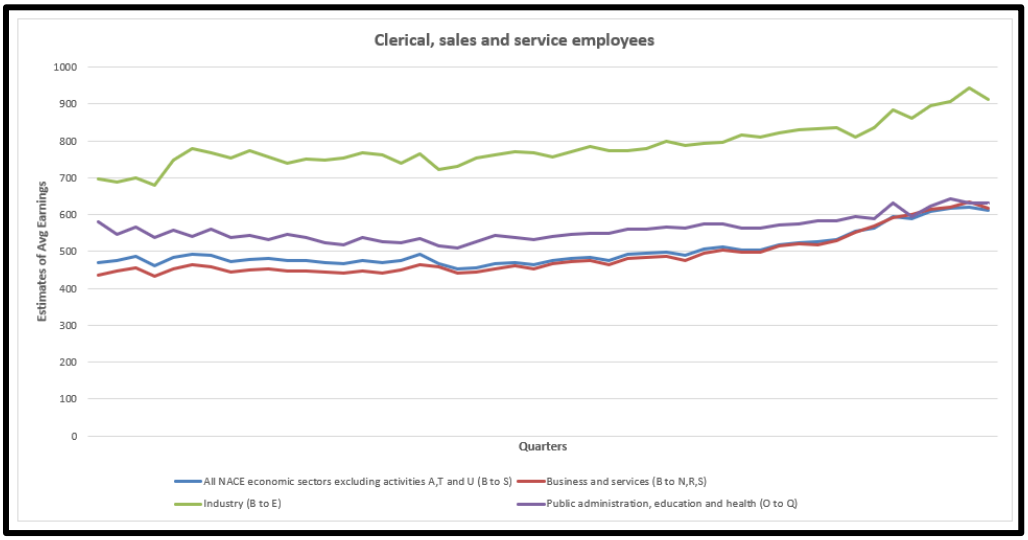

Fig 2

The graph depicts the overall rise in average earnings of serviced employees across all four sectors from 2010 Q2 to 2022 Q1. We can conclude that the Industry (B to E) sector has a high average wage for service employees.

iii. Managers, Professionals, and assorted professionals:

Descriptive Measures of samples		**All NACE economic sectors excluding activities A,T and U (B to S)**	**Business and services (B to N,R,S)**	**Industry (B to E)**	**Public administration, education and health (O to Q)**
Measures of Centrality	**Mean**	1212.761458	1303.046667	1486.011042	1146.758125
	Median	1203.715	1293.435	1496.23	1142.11
Measures of Variation	**Std Deviation**	95.45309783	102.9343809	91.60424317	101.1298486
	Range	339.75	381.65	336.22	363.68
	Min	1041.41	1113.74	1312.61	967.88
	Max	1381.16	1495.39	1648.83	1331.56
	Variance	9111.293885	10595.48676	8391.337367	10227.24628

Managers in the industry sector have the highest average earnings, while those in public administration, education, and health have the lowest. Median of Industry sector is highest 1486.23 and for the other 3 sectors it is 1303.046, 1212.76 and 1146.75. Among all sectors, the

public, education, and health sectors have the lowest minimum and maximum values. Thus, based on the data, the public, education, and health sectors have the lowest average earnings for Managers, while the industry sector has the highest earnings.

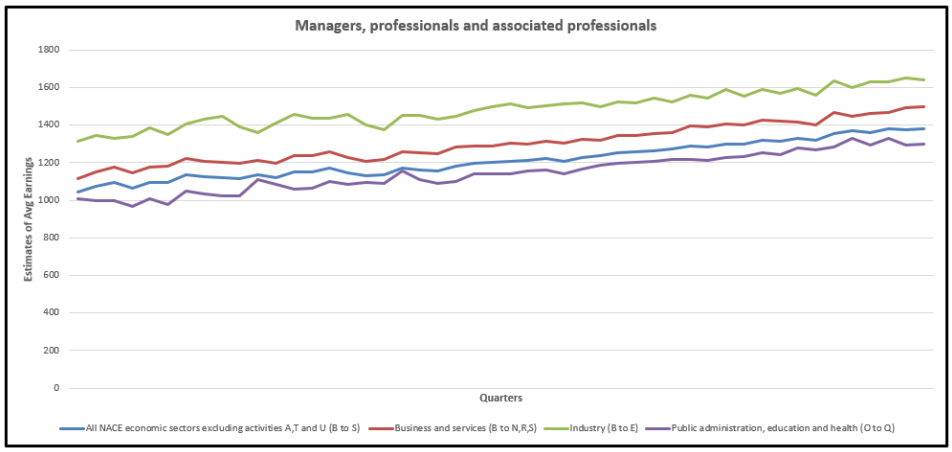

Fig 3

The line plot depicts the quarterly average earnings of Managers for all four sectors from 2010 to 2022. The average earnings of managers in the industry sector are the highest. The average earnings of the business and services sector exceed those of the NACE and public administration sectors combined.

3.2 Inferential Analysis

In inferential studies, we analyze sample data with objective to describe the entire population. Moreover, as per Central Limit Theorem (Weiss, 2010); If sample is large enough; then $\bar{x}$ is normally distributed regardless of the distribution of population mean and thus, we assume that the sample is randomly distributed.

Our data revolves around the earnings of the working sector, so we take the mean earnings of all working class irrespective of their sectors and compare it with the estimates of average earnings. Time series data is itself a realization of a sequence of unknown random variables over time. Assuming that some sequence of random variables is defined only at discrete points on the time axis, then it is a discrete stochastic process. It will obviously be a random sample.

3.2.1 Confidence Intervals

We want to know the mean estimate of average earnings of each working sector's employees, so we find a point estimate with 95% confidence that the population mean will fall between the specified intervals.

We use t-distribution with *n-1* degrees of freedom to find confidence intervals as our population standard deviation (σ) is not known. Using formula:

$$Confidence\ Interval = \bar{x} \pm t_{\frac{\alpha}{2}}.\frac{s}{\sqrt{n}} \qquad where\ degrees\ of\ freedom = n - 1;\ \alpha = 1 - 0.95$$

Confidence Intervals for Manual Workers

Parameters	Population 1	Population 2	Population 3	Population 4
$\bar{x}$ (sample mean)	534.704	554.838	696.696	472.854
Median	522.29	543.03	709.115	467.18
s (sample st dev)	45.179	54.084	66.755	35.247
n (sample size)	48	48	48	48
1-α (confidence)	0.95	0.95	0.95	0.95
$t_{\alpha/2}$	2.012	2.012	2.012	2.012
LL (confidence interval Lower Limit)	521.585	539.133	677.313	462.619
UL (confidence interval Upper Limit)	547.823	570.542	716.080	483.089

Interpretation:

We are 95% confident that the mean of the estimates of the average earnings of manual workers in various working sector lie between the lower and upper limits as specified in the above table.

Similarly, the confidence interval for Service Employees and Managers is as: -

Confidence Intervals for Service Employees				
Parameters	**Population 1**	**Population 2**	**Population 3**	**Population 4**
$\bar{x}$ (sample mean)	503.350	489.014	785.159	559.927
Median	485.41	464.675	771.595	553.715
s (sample st dev)	45.984	56.258	57.686	32.507
n (sample size)	48	48	48	48
1-α (confidence)	0.95	0.95	0.95	0.95
$t_{\alpha/2}$	2.012	2.012	2.012	2.012
LL (confidence interval Lower Limit)	489.997	472.678	768.409	550.488
UL (confidence interval Upper Limit)	516.702	505.349	801.910	569.366

Confidence Intervals for Managers				
Parameters	**Population 1**	**Population 2**	**Population 3**	**Population 4**
$\bar{x}$ (sample mean)	1212.761	1303.047	1486.011	1146.758
Median	1203.715	1293.435	1496.23	1142.11
s (sample st dev)	95.453	102.934	91.604	101.130
n (sample size)	48	48	48	48
1-α (confidence)	0.95	0.95	0.95	0.95
$t_{\alpha/2}$	2.012	2.012	2.012	2.012
LL (confidence interval Lower Limit)	1185.045	1273.158	1459.412	1117.393
UL (confidence interval Upper Limit)	1240.478	1332.936	1512.610	1176.123

Now we use the confidence intervals obtained above to make decisions about hypothesized values of a population mean.

3.2.2 Hypothesis Testing

We want to see if the mean estimates of average earnings of different ranking positions is different or not in all four working sectors. Since we wanted to compare our variable in 4 populations, we

have used one-way ANOVA. Further, before using ANOVA, we have checked and verified if we satisfy all the assumptions or not. The ratio of largest and smallest standard deviation comes out to be less than 2.

i. Problem as a hypothesis test:

For Production, transport, craft and other manual workers, we see that we have one variable (average earnings) and 4 populations (4 different working sectors)

$Employees: Production, transport, craft\ and\ other\ manual\ workers$
$Population\ 1: Employees\ of\ All\ NACE\ economic\ sectors\ excluding\ activities\ A, T and\ U$
$Population\ 2: Employees\ of\ Business\ and\ services\ (B\ to\ N, R, S)$
$Population\ 3: Employees\ of\ Industry\ (B\ to\ E)$
$Population\ 4: Employees\ of\ Public\ administration, education\ and\ health\ (O\ to\ Q)$

Next, we denote the means of the variable "estimates of average earnings" for the 4 populations μ_1, μ_2, μ_3 and μ_4 respectively.

$\mu_1 = mean\ estimate\ of\ average\ earnings\ of\ population\ 1$
$\mu_2 = mean\ estimate\ of\ average\ earnings\ of\ Population\ 2$
$\mu_3 = mean\ estimate\ of\ average\ earnings\ of\ Population\ 3$
$\mu_4 = mean\ estimate\ of\ average\ earnings\ of\ Population\ 4$

ii. Finally, we state our hypothesis that we want to test:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4\ (mean\ estimates\ of\ average\ earnings\ is\ same)$
$H_a: Not\ all\ the\ means\ are\ equal$

iii. Performing the test at 5% significance level: $\alpha = 0.05$

iv. We then compute the value of test statistic using excel:

$$F = \frac{MSTR}{MSE}$$

v.

Computations	**Values**			
			p-value	0.000000
f-statistic	**161.1624**		p-value can only be calculated via technology, i.e. no stats table	
f-critical	**2.6526**			
df_n	3			
df_d	188		is p value less than or equal to alpha?	Yes
Reject H_0?	**Yes**		**Reject H_0?**	**Yes**

Through critical-value approach, we see that f-statistic falls in the rejection region of f-critical and thus we reject H_0. Similarly, through *p*-value approach, we see that $\boldsymbol{p \leq \alpha}$; we reject H_0.

vi. **Interpretation**: At the 5% significance level, the data provide sufficient evidence to conclude that a difference exists in mean estimates of average earnings by Production, transport, craft, and other manual workers among the four working sectors. Evidently, at least two of the sectors have different mean estimates of average earnings.

vii. Now that we have incurred that our Null hypothesis has been rejected, i.e., the mean estimates of average earnings for manual workers for 4 working sectors is not equal, we now want to analyze if for any 2 sectors, the mean estimates is same. We then proceed to perform a pooled *t*-test assuming equal variances at 5% significance level as ratio of sample standard deviations comes out to be less than 2 (**54.084/45.179 = 1.197**). Here we have considered only the first 2 working sectors.

Hypothesis 2: pooled t-test assuming equal variances (at 5% significance level)	
Null Hypothesis (H_0): $\mu_1 = \mu_2$	H_0: Mean estimate of average earnings of manual workers is **equal** for Working Sector (1) and Working Sector (2)
Alternative Hypothesis (H_a): $\mu_1 < \mu_2$ (left tailed)	H_a: Mean estimate of average earnings of manual workers in working sector (1) is **less than** that for Working Sector (2)

Assumptions:	
Simple Random Sample	Since it is a time series data; therefore, Assumption 1 is satisfied.
Independent Sample	The samples are given as independent samples; therefore, Assumption 2 is satisfied.
Normal Populations	As per Central Limit Theorem (Weiss, 2010): If sample is large enough (>=30); then x bar is normally distributed regardless of the distribution of population mean.
Equal Standard Deviations / Rule of thumb the rule of 2	As per this rule, ratio of largest and smallest standard deviations is: 54.084/45.179 = **1.197** which is <2. Thus, by the rule of 2, we can consider Assumption 4 satisfied.

	Population 1	**Population 2**		
Population mean (μ)	**μ_1**	**μ_2**		
Sample size (n)	48	48		
Sample mean ($\bar{x}$)	534.70	554.84		
Sample St dev (s)	45.179	54.084		
Degrees of freedom	47	47		
Level of significance	0.05			
S(pooled)	49.831			
t-statistic	-1.979			
t-critical	-1.661		**p-value**	0.025
degrees of freedom	94			
Reject H_0?	**Yes**		**Reject H_0?**	**Yes**

viii. **Interpretation:** At the 5% significance level, the data provides sufficient evidence to reject the null hypothesis that is the mean estimates of average earnings by Production, transport, craft, and other manual workers in working sector (1) is less than that in working sector (2).

Similarly, for us to compare the mean estimates of average earnings of each pair of sectors, we would need to perform the test ${}^{4}_{2}\mathrm{C}$ times i.e., 6 times (Kim, 2014). We see that as we implement multiple *t*-tests with pre-set significance level of α = 0.05, we are adding multiple chances of error. Since α level is set for each comparison, we are increasing the total error rate to $1 - (0.95)^6$. Thus, *t*-test will not give us the comparison between all the sectors.

ix. Further, we again implement ANOVA for 2 more categories of employees which are:
- Clerical, sales and service employees (*service employees*)
- Managers, professionals and associated professionals (*managers*)

x. We do this to analyze if the population mean estimates of average earnings of service employees and managers is equal in 4 working sectors. We make sure all the assumptions are satisfied and the ratio of largest to the smallest standard deviation is less than 2.

After computing our test statistic and results on excel, for Clerical, sales and service employees we see:

Computations	**Values**				
			p-value	0.000000	
f-statistic	**374.6500**		p-value can only be calculated via technology, i.e. no stats table		
f-critical	**2.6526**				
df_n	3				
df_d	188		is p value less than or equal to alpha?	Yes	
Reject H_0?	**Yes**		**Reject H_0?**	**Yes**	

xi. At the 5% significance level, the data provides sufficient evidence to conclude that a difference exists in mean estimates of average earnings of Clerical, sales and service employees among the four working sectors.
Evidently, at least two of the sectors have different mean estimates of average earnings.

xii. Similarly for Managers, professionals and associated professionals:

Computations	**Values**				
			p-value	0.000000	
f-statistic	**108.6144**		p-value can only be calculated via technology, i.e., no stats table		
f-critical	**2.6526**				
df_n	3				
df_d	188		is p value less than or equal to alpha?	Yes	
Reject H_0?	**Yes**		**Reject H_0?**	**Yes**	

At the 5% significance level, the data provide sufficient evidence to conclude that a difference exists in mean estimates of average earnings by Managers, professionals and associated professionals among the four working sectors. Evidently, at least two of the sectors have different mean estimates of average earnings.

4. Conclusion:

Based on the above analysis, we have drawn the following conclusions:

i. Industry (B to E) employees generally get higher average earnings (compared to other sector employees). All types of employees include Clerical, sales and service employees, Production, transport, craft and other manual workers, Managers, professionals and associated professionals can get higher earnings in the industry sector.
ii. Public administration, education and health employees generally get lower average earnings (compared to other sector employees). Production, transport, craft and other manual workers, Managers, professionals and associated professionals can get lower earnings in the Public administration, education and health sector.
iii. Based on hypothesis testing conclude that a difference exists in mean estimates of average earnings for both the three types of employees among the four working sectors.

5. Recommendations

The average earnings for working positions vary depending on the sector, according to statistical analysis data. As a result, it is suggested that if one wishes to earn a higher average salary as a Managers, professionals, and associated professionals, or as Production, transport, craft, and other manual workers, they should avoid public administration, health, and education, and instead work in the Industry (B to E) sector. If a clerical, sales, or service employee wants to earn a higher average salary, they should avoid working in the business and service sector, which has the lowest average earnings.

6. References

i. O'Riordan, J (2022). Estimates of Average Earnings. Available at https://data.cso.ie/table/EHQ13
ii. Kim, H (2014). Analysis of variance (ANOVA) comparing means of more than two groups. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3916511/
iii. Weiss, N. A. (2010). Introductory Statistics. Addison-Wesley.

CONTRIBUTION

Team Member Names	% to Reading and Topic Appreciation	% to Analysis and Discussion	% to Report Preparation
Anushka Jain (22200246)	33.33%	33.33%	33.33%
Mingbo Zhang (22204725)	33.33%	33.33%	33.33%
Pratik Pore (22200170)	33.33%	33.33%	33.33%

We all contributed equally to the completion of this task. We each attempted to brainstorm what data to analyze that we could relate to in our daily lives. Then we attempted to research earnings and employment type by clicking on the various links provided in the assignment questionnaire. Following the selection of the dataset, we developed our own descriptive analysis and hypothesis, which we discussed and individually analyzed before collectively validating each approach. Finally, we compiled the excel file and the report and verified it all together.

Thank you!