**BUSINESS STATISTICS AND DATA MODELLING (5Z7V0058\_2223\_9)**

**22557501**

**Tabitha Inyamu Ujah**

**(NB: ALL FIGURES ARE NOT REPRSENTATIVE OF SAMSUNG FIGURES BUT WERE OBTAINED FROM FREE DATA SETS FROM KAGGLE)**

2023

**Table of Contents**

[**1.0**   **Company Profile** 2](#_Toc124154411)

[**2.0**   **Summary of Statistical Methods** 2](#_Toc124154412)

[**2.1 Regression Analysis** 2](#_Toc124154413)

[**2.2 Analysis of Variance (ANOVA)** 12](#_Toc124154414)

[**2.3 Cluster Analysis** 17](#_Toc124154415)

[**3.0 Issues and Processes Involved in the Implementation and Adoption of Statistical Analysis to Aid Decision Making** 22](#_Toc124154416)

[**References** 24](#_Toc124154417)

# 

# **1.0 Company Profile**

Samsung is of the biggest manufacturers of consumer electronic products in the world established in 1969 with its headquarters in Seoul, South Korea. It operates in diverse geographies namely, Asia, the Americas, Europe, Middle East, and Africa. The company earned approximately $222 billion in revenue and approximately $31.8 billion in net income in 2021 according to their 2021 income statement. In total, they are made up of 98, 557 employees worldwide (Samsung Electronics Co., 2023).

One of the most important responsibilities for Samsung's leadership is to keep their company up to date with the latest technologies, product development, and business strategies to maintain their market position (Jang, et al., 2019). In September 2016, Samsung halted the sales of Samsung Note 7 phones because of a defect in their lithium batteries that caused the phones to explode (Thomas, et al., 2017). Nevertheless, Samsung is focused on conducting business in a sustainable manner. it has committed to using more renewable energy, attaining net zero emissions throughout the whole company by 2050, and making strategic investments in emerging sustainable technologies (Djatmiko & Pradanab, 2016).

# **2.0 Summary of Statistical Methods**

## **2.1 Regression Analysis**

Regression analysis is used for investigating relationship between variables. There is usually a dependent variable and an independent variable. It is both a descriptive and prescriptive analytical technique that is used to build models that can be used to predict a particular event to make appropriate business decisions (Anderson, et al., 2017). Based on the interactions between dependent and independent factors, regression analysis can be beneficial for forecasting the results and changes in dependent variables. The results of regression Analysis are affected by the sample size of data, missing data or nature of the data (Ali & Ahtisham, 2021).

**2.1.1 Illustration**

The illustration below shows the use of multiple linear regression analysis to determine the relationship between sales of Samsung phones (dependent variable) and its corporate success factors such as marketing, technology, and branding (independent variables). The data is a time series data of the weekly amounts attributed by these factors over the last three (3) years and the corresponding sales recorded.

The null hypothesis, H0: β1 = β2 = β3, being that none of the variables has a significant effect on the dependent variable, sales.

The alternative hypothesis, HA: One or more of the independent variables is not equal or has a significant effect on the dependent variable, sales

Firstly, to determine the relationship between sales and the independent variables the scatter plot in Figure 2.1.1 below is produced.

Sales and technology are strongly related, which implies that the more money invested in technology development, the larger the gain in sales. Marketing and branding also show positive relationship but the data points of marketing are closer to the line of best fit compared to branding. Marketing shows a great relationship with technology and less relationship with branding. Branding has the most dispersed data points for all other variables a bit further from the line of best fit.

Chart, scatter chart

Description automatically generated

**Figure 2.1.1: Scatter Plot for each variable**

From the table below, Marketing and technology show a high correlation, significant at the 1% level with a correlation co-efficient value of 0.89There exists a moderate correlation between technology& marketing and marketing &branding with a correlation coefficient value of 0.627 and 0.663 respectively at the 1% level of significance. Overall, there is a strong relationship between the dependent variable, sales and the independent variables, technology>>>Marketing>>Branding with values 1.000, .890 and .627 at the 1% level of significance.

**Table

Description automatically generated**

**Figure 2.1.2a: Correlation Table**

To further investigate the regression model, the tables below were obtained.

**Table

Description automatically generated**

**Figure 2.1.1b: Model Summary**

The adjusted R square shows that 99.9% of the variability in sales is explained by the variability in technology, branding and marketing and explained by the model.

Table

Description automatically generated

**Figure 2.1.4 Anova Table**

The sum of squares confirms the above statement that 99.9% of the variability in sales is explained by the variability in the independent variables. The model is significant with a p-value of <.001. therefore, we reject the null hypothesis that the co-efficient of the independent variable is zero and accept the alternative hypothesis at the 1% confidence level one or more of the independent variables has a significant effect on the dependent variable. Therefore, relationship exists between dependent and independent variable.

Table

Description automatically generated

**Figure 2.1.5 Coefficients**

To further investigate collinearity from the correlations table above, marketing recording a tolerance of .190 which tends closer to zero and a VIF (Variance Inflation Factor) of 5.2(above 5) shows great dependence on other independent variables and hence collinearity. Although, its p-value is shown to be highly insignificant at the 5% confidence. The technology variable shows similar behaviour as the marketing variable with tolerance of .206 (tending close to zero) and VIF of 4.8(close to 5) but a p-value of high significance at the 1% level of significance. In contrast, branding records a tolerance of .55 and VIF of 1.802 which tends towards 1 which indicates a more lack of dependence on other variables.

**Table

Description automatically generated**

**Figure 2.1.6 Collinearity Diagnostics**

In the table above, the condition index column has no record of value above 30 but there is a value of 12 in the 4th dimension. Technology and marketing have a value of .92 and .89 respectively (above 0.5) for the variance proportions, therefore it is safe to assume collinearity between technology and marketing based on all evidence shown.

The regression model is run without marketing to give the output below.

Table

Description automatically generated

**Figure 2.1.7**

The adjusted R square still shows that 99.9% of the variability in sales is explained by the variability in technology, branding and explained by the model.

Table

Description automatically generated

**Figure 2.1.8**

The model is highly significant at the 1% level of confidence, and we accept the alternate hypothesis that one or more of the independent variables has a significant effect on the dependent variable, sales.

Table

Description automatically generated

**Figure 2.1.9**

Running the model again without the marketing variable, shows acceptable tolerance value and VIF of .609 and 1.643 for both variables but branding has a p-value that is highly insignificant at the 5% confidence level. We accept the null hypothesis and reject the alternate hypothesis that branding has no significant effect on sales.

Table

Description automatically generated

**Figure 2.1.10**

Technology still has a .98 variance proportion; the model is then run again with only the technology variance.

Table

Description automatically generated

**Figure 2.1.11**

Table

Description automatically generated

**Figure 2.1.12**

Table

Description automatically generated

**Figure 2.1.13**

For the final model**, y = -1.172 + 3.569x** adjusted R square shows that 99.9% of the variability in sales is explained by the variability in technology.

The model is significant with a p-value of <.001. We reject the null hypothesis and accept the alternate hypothesis that the technology variable has a significant effect on the dependent variable, sales.

Also, the constant for the technology variable is significant with a value of .032. We reject the null hypothesis that the co-efficient of the independent variable is zero and accept the alternative hypothesis at the 1% confidence level. Therefore, relationship exists between dependent and independent variable

This shows technology is a major driver of sales for Samsung Phones.

**2.1.2 Test for Validity**

The regression model is based on the below assumptions.

1. That the error between the predicted level of the dependent variable and the actual values is constant
2. The distribution of the residuals conforms to a normal distribution (with a mean of zero)
3. The residuals are independent of one another

The residual analysis below is used to support the validity of the statistical model and its assumptions.

Table

Description automatically generated

**Figure 2.1.2a**

3% (5 cases) of the 156 cases produced errors outside 2 deviations, the deviations are still within acceptable deviations with mostly positive but negative number included as well.

Table

Description automatically generated

**Figure 2.1.2b**

Chart, histogram

Description automatically generated

**Figure 2.1.2c**

A histogram of the residuals approximate to a normal distribution as can be observed in the histogram above. The P-P plot below also confirms this.

Chart, line chart

Description automatically generated

**Figure 2.1.2d**

Chart, scatter chart

Description automatically generated

**Figure 2.1.2e**

The scatter plot is a plot of the independent variable(technology) on the x axis and the residuals on the y axis. A constant error of variance (homoscedasticity) is observed as the pattern shows all the plots to be within -3 to +3 on the y axis.

Table

Description automatically generated

**Figure 2.1.2f**

From the residual analysis above, it supports the assumption that the mean of the residuals is zero. The Shapiro-wilk’s test of normality is highly insignificant at the 5% confidence level, therefore we accept the null hypothesis and reject the alternate hypothesis, the distribution does conform to the normal distribution. The Q-Q plot and histogram also support this stateme

Chart, histogram

Description automatically generated

**Figure 2.1.2g**

Chart, line chart

Description automatically generated

**Figure 2.1.2h**

The above chart showed constant variance against the predicted level of dependent variables and based on the table below which shows insignificance at the 5% confidence level to the Breusch-Pagan Test of heteroscedasticity, we can therefore accept the residuals does indeed show constant variance.

Therefore, we can accept that the model **y = -1.172 + 3.569x is valid.**

Table

Description automatically generated

**Figure 2.1.2i**

## **2.2 Analysis of Variance (ANOVA)**

Analysis of variance is a statistical tool used to detect differences between experimental group means. It is used to determine whether the observed differences in sample means are large enough to reject the null hypothesis. It can be used to analyse data obtained through an observation of a study (Anderson, et al., 2017).

**2.2.1 Illustration**

Following the recall of the Samsung note 7 phones, Samsung has resorted to identifying 5 different manufacturers of the lithium batteries and asked that they submit safety thresholds for one (1) batch of batteries produced in the last quarter of 2022 in anticipation for production in the year 2023.

To evaluate the best supplier with the lowest safety threshold per batch anova is used to determine the difference in means to prevent a repeat of the September 2016 error.

Table

Description automatically generated

**Figure 2.2.1**

The null hypothesis, H0: µ1 = µ2 = µ3 = µ4 = µ5

The alternative hypothesis, HA: Not all population means are equal

Firstly, the safety threshold data provided by each supplier is confirmed to approximate to a normal distribution.

From the histogram we can see that the A, C, D, & E have normal distributions which is further shown with the Q-Q plots as their plots tend to the 45-degree line and B tends away from the line.

Based on the values from the Shapiro-wilk’s test of normality for each supplier, statistically, A, C, D, & E distributions are not significantly different from the normal at the 5% level and B is significantly different from the normal at the 5% level of significance. We accept the alternative hypothesis and reject the null hypothesis. Since B is approximately 18% of the distributions, a homogeneity test of the variances can be carried out.

Table

Description automatically generated

**Figure 2.2.2**

Chart, histogram

Description automatically generatedChart, line chart, scatter chart

Description automatically generated

Chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, histogram

Description automatically generatedChart, line chart, scatter chart

Description automatically generated

Chart, histogram

Description automatically generatedChart, line chart

Description automatically generated

Chart, histogram

Description automatically generatedChart, line chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

**Figure 2.2.3**

A picture containing text, receipt, screenshot

Description automatically generated

**Figure 2.2.4**

**EXPLANATION OF CHARTS**

Supplier B has the most dispersion as shown in the box plots and confirmed with the standard deviation it has the lowest median value (safety threshold) and supplier D has the highest median value (safety threshold) with 32 and 38 respectively.

Based on the mean test of the homogeneity of variances, the safety threshold of each supplier is highly significantly different at the 5% and 1% level of significance with a value 0.001 which is <<< 0.05. Therefore we reject the null hypothesis H0 : µ1 = µ2 = µ3 = µ4 = µ5 and accept the alternative hypothesis, HA that not all population means are equal.

Although we have the statement above that there is difference in the variances of the safety threshold of each supplier and the anova test is weakened, we have the F value is 3.175 and the f statistic (p-value) is statistically significant with a value of 0.014 at the 5% confidence level.

Chart, line chart

Description automatically generated

**Figure 2.2.5**

**Means Plot**

To further support the earlier statement above of difference in group variance, the Welch test and the Brown-Forsythe test was carried out and output is shown below.

**Table

Description automatically generated**

**Figure 2.2.6**

We obtained the same result that there is indeed a significant difference in the variances of the safety threshold of each supplier at the 1% level of significance in the welch test and 5% level of significance in the Brown-Forsythe test.

**Post Hoc Test**

To determine where infact the difference is within the groups of variances we carry out the Games-Howell test post hoc test.

Table

Description automatically generated

**Figure 2.2.7**

Here, there is a difference between the safety threshold of supplier A and E and a difference between Supplier D and E as well. Supplier E has the lowest safety threshold, and it is the least dispersed out of every other supplier (this means the values of safety threshold are more consistent). Supplier A on the other hand is the least dispersed of the three with closest mean averages (supplier A, B & C). Supplier D although has the highest safety threshold is not too far behind with dispersion after A & E.

Supplier D mean value is statistically bigger safety threshold than Supplier E and Supplier A is statistically bigger than Supplier E at the 5% level of significance

## **2.3 Cluster Analysis**

Cluster Analysis is a multivariate method which aims to classify a sample of subjects based on a set of measured variables into several different groups such that the similar subjects are placed in the same group. It is also known as segmentation analysis. The aim is to identify homogenous groups that share the same characteristics. This analysis unlike the two above, does not make a distinction between dependent and independent variables.

**2.3.1 Illustration**

To determine what combination of employee benefits is currently preferred by different clusters or groups of employees to position the company to tailor these benefits according to employee profile and to channel resources towards this with the emerging economic downtown and the anticipated scramble to retain talent amongst the big tech companies and organisations.

An employee benefit survey data was obtained for the total 347 employees at the headquarters in Seoul.

A screenshot of a computer

Description automatically generated with low confidence

**Figure 2.3.1**

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Table

Description automatically generated

**Figure 2.3.2**

Table

Description automatically generated

**Figure 2.3.3**

Table

Description automatically generated

**Figure 2.3.4**

Table

Description automatically generated

**Figure 2.3.5**

From the anova table it is evident Corporate ISA contributes insignificantly to the overall analysis into clusters. Data was explored to remove outliers and correlations between the variables were tested and one of highly correlated variables – Academic assistance, flexible vacation & Insurance (circled above) removed to avoid increasing the weight per benefit. Each of the clusters have been named according to the benefits the rank as most important. The clusters profiles have been profiled as below based on benefits the consider most important

Cluster 1 – Early parents (prioritise parental assistance)

Cluster 2 – Middle aged parents of teenagers (Mortgage Scheme)

Cluster 3 – Generation z (prioritise gadget insurance)

Cluster 4 – Introvert Comfort Parents (prioritise work from home)

Cluster 5 – Baby Bomers (prioritise club membership and pension scheme)

Cluster 6 – Millennial Techies Parents (prioritise club membership, Gadget insurance& parental assistance)

# **3.0 Issues and Processes Involved in the Implementation and Adoption of Statistical Analysis to Aid Decision Making**

Statistical methods are being adopted in organisations because they serve as the foundation of process improvement models, are essential to performance-based improvement initiatives and industrial problem solving creating a competitive edge for organisations (Card, et al., 2008; Abraham, 2007). Statistical models contribute to standardizing the processing of information in certain decision types thereby increasing efficiency with which organisations respond in similar situations (Søbjerg, et al., 2021). Decisions made using statistical procedures are better because they are based on facts rather than opinions, which makes the process and results more objective (Card, et al., 2008). (Lepenioti, et al., 2020)also further states that statistical methods are superior to clinical judgements.

(Abraham, 2007) stated implementation of statistical tools should be in line with the three arbitrary levels in a typical organisation; Strategic, Managerial and Operational and these levels help identify and emphasize different tools to be directed at the different levels. Alternatively, (Card, et al., 2008) groups decision making into control of individual activities and predictions of future and final process outcomes

Decision making at the strategic level should be based on facts supported by appropriate data and understanding of variation as emphasis should be placed on statistical thinking. In the managerial level, systems are devised to support the implementation direction of the upper management, an understanding of statistical thinking and tools is required at this level. Then at the operational level, statistical tools are implemented through the systems designed at the managerial level. For instance, a process improvement engineer should be proficient in experiment design and process control (Abraham, 2007; Lepenioti, et al., 2020)

The challenges involved in statistical methods for professional judgment can be group into three main categories

1. Ethical issues

This involves inherent limitations that exist in information systems when the data utilised is inaccurate or the scale of the data being set is beyond the range of common software tools. It also includes considering the integrity and completeness of the data to avoid “noise”. Statistical methods are only as good as the data included; hence accuracy poses an ethical issue. Also, most conceptual framework and analytical methods are only useful in specific contexts and when applied wrongly provide incomplete or inaccurate data for end users and decision makers (Søbjerg, et al., 2021; Li, et al., 2021; Earley, 2015)

1. Legal and organisational issues

Abiding by legislation regarding data usage means important data may not be included in statistical models. The GDPR General Data Protection Regulation effected May 2018 has a strong focus on data protection of quantitative data and the rights of individuals (Søbjerg, et al., 2021). Operational level personnel involved in day-to-day tasks tend to believe that statistical model is telling them what they already know. The challenge is getting the staff to think beyond what they already know to the additional insights they can obtain from the statistical model (Card, et al., 2008).

1. Technical Issues

This refers to the availability, validity and reliability of the calculations based on the statistical model. It also refers to the integrity and quality of data from the systems from which it is sourced as it impacts heavily upon implementation success (Søbjerg, et al., 2021; El-Adaileh & Foster, 2019). The quality of data-driven decision-making depends not only on the data itself but also closely related to the process of data acquisition and analysis.

Finally, the support of management has been cited as one of the major factors that affect implementation of statistical analysis. One of the biggest obstacles a BI implementation team confronts is garnering support from the management and the rest of the organisation.

# **References**

Abraham, B., 2007. Implementation of Statistics in Business and Industry. *Revista Colombiana de Estadística,* 30(1), pp. 1- 11.

Ali, P. & Ahtisham, Y., 2021. Understanding and Interpreting Regression Analysis. *Evidence-Based Nursing,* 24(4), pp. 116 - 118.

Anderson, D. R. et al., 2017. *Statistics for Business and Economics.* 13th ed. Boston, MA: Cengage Learning.

Card, D. N., Domzalski, K. & Glyn, D., 2008. Making Statistics Part of Decision Making in an Engineering Organization. *IEEE Software,* 25(3), pp. 37- 47.

Djatmiko, T. & Pradanab, R., 2016. Brand Image and Product Price; Its Impact for Samsung Smartphone Purchasing Decision. *Procedia - Social and Behavioral Sciences,* 219(12), pp. 221-227.

Earley, C. E., 2015. Data analytics in auditing: Opportunities. *Business Horizons,* 58(dx.doi.org/10.1016/j.bushor.2015.05.002), pp. 493 - 500.

El-Adaileh, N. A. & Foster, S., 2019. Successful Business Intelligence Implementation : A systematic literature review. *Journal of Work-Applied Management,* 11(2), pp. 121-132.

Jang, S. H., Lee, S. M., Kim, T. & Choi, D., 2019. Planting and harvesting innovation - an analysis of Samsung Electronics. *International Journal of Quality Innovation,* 5(7), pp. 1-16.

Lepenioti, K., Bousdekis, A., Apostolou, D. & Mentzas, G., 2020. Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management,* Volume 50, pp. 57-70.

Li, C., Chen, Y. & Shang, Y., 2021. A review of industrial big data for decision making in intelligent. *Engineering Science and Technology, An International Journal,* 101021(29), pp. 2-13.

Samsung Electronics Co., L., 2023. *Samsung.* [Online]   
Available at: https://www.samsung.com/global/ir/financial-information/earnings-release/  
[Accessed 02 January 2023].

Søbjerg, L. M. et al., 2021. Using risk factor statistics in decision-making: prospects andchallenges. *European Journal of Social Work,* 24(5), pp. 788-801.

Thomas, E., Zeron, A. & O’Rourke, J. S., 2017. Samsung Electronics Company, Ltd.:Galaxy Note 7 Crisis. *The Eugene D. Fanning Center for Business Communication, Mendoza College of Business, University of Notre Dame,* DOI: 10.4135/9781526489043(1), pp. 1-17.