Foundations of Probability – Calculations Document

University of Strathclyde

Table of Contents

List of Figures

List of Tables

Data Analysis

Question 1 – Subset Worksheet

A specific subset of data was created using the version 21.1.0 of the Minitab software. A base was set for random numbers, the random number generator was initialised with the number "202476449". Random data was then generated using the Bernoulli distribution with a probability of 0.6, specifically 6150 rows in column C11. After generating the random Bernoulli data, the subset was filtered to only include rows where the value in column C11 was 1. The subset of filtered data was then saved to an Excel file (.xlsx) with the name "202476449-myprojectdata.xlsx", the first 5 rows of the subset data is shown below in Table 1.

Table 1 - First 5 value for Subset of Filtered Data

Mother Age	Feeding Type	Tried Breast Feeding	Smoked	Type of Birth	Maternal BMI	Pulse	Bump Size	Iron Level	Time	Column_11
20-24	Breast	1	0	Forceps	Healthy	61	34.9	9.9	2	1
20-24	Breast	1	1	Caesarean - Elective	Healthy	71	34.5	10.3	7	1
20-24	Breast	1	0	Vaginal	Healthy	73	35.4	11	2.5	1
20-24	Breast	1	0	Vaginal	Healthy	75	34.1	12	0.5	1
20-24	Breast	1	0	Vaginal	Healthy	76	35.2	11.2	1	1

Question 2 – Variables

Iron Level

Figure 1 shows the output of the boxplot of Iron Level. The boxplot was selected for its central tendency and spread where you can visually spot the median and the interquartile ranges. It can also reveal outliers and can give insights into the data's symmetry and skewness. We can see, the boxplot shows the median is around 11 with the interquartile ranging from approximately 10.5 to 11.5 indicating the middle 50% of iron levels. The whiskers show the extent of values that are not considered outliers. Notice that there are several outliers, particularly below the bottom whisker suggesting a slight skewness.

A histogram was chosen for its ability to represent the frequency distribution of the continuous variable, which is particularly useful for identifying the shape of the distribution, again detecting any skewness and spotting unusual patterns. The distribution of iron levels appears roughly symmetric and bell-shaped, which is a characteristic of normal distribution.

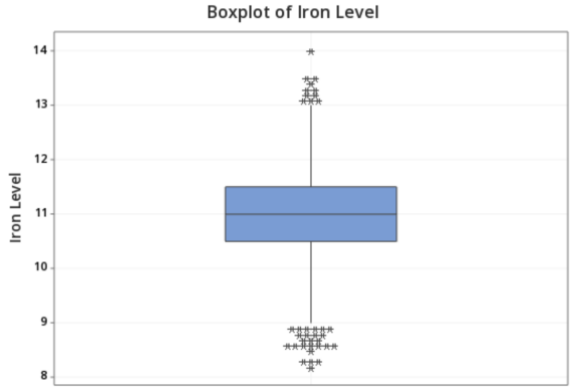

Figure 1 - Boxplot of Iron Levels

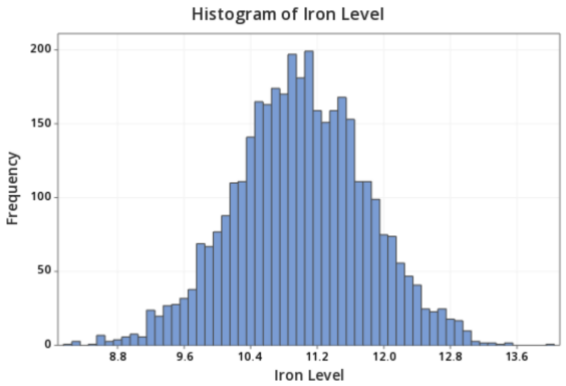

Figure 2 - Histogram of Iron Levels

A descriptive statistic table summary provides a comprehensive summary allowing for an assessment of its central tendency, variability and distribution which can inform subsequent analyses and interpretations. The headers are summarised below:

- N: Sample Size – indicated data robustness.
- N*: missing values – indicates data completeness.
- Mean: Average value – central data point.
- SE Mean: Precision of mean – estimate reliability.
- StDev: Data spread – variability measure.
- Minimum: Lowest Value – data range state.
- Q1: Lower quartile.
- Median: Middle value – central tendency.
- Q3: Upper quartile.
- Maximum: Highest value – data range end.

Table 2 - Minitab Descriptive Statistics Output for Iron Level

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Iron Level	3653	0	11.001	0.0133	0.801	8.200	10.500	11.000	11.500	14.000

Type of Birth

For “Type of Birth”, which is a categorical variable a bar chart will be used as it can show the frequency of each category, providing a clear visual comparison of the different types of birth. It allowed a quick visual comparison and easily understandable representation of the distribution of categorical data.

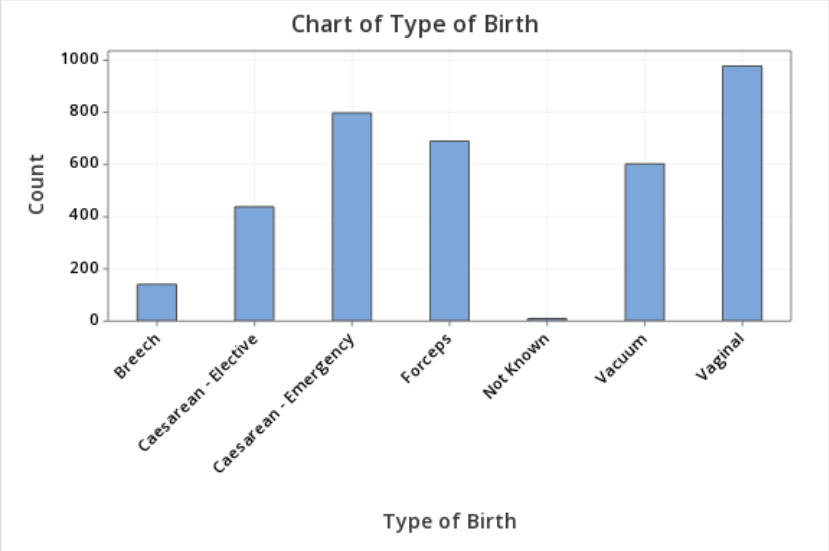

Figure 3 - Chart showing type of birth

A frequency table was chosen for its straightforward representation of categorical data, making it a valuable complement to a visual bar chart. It enumerates each category within 'Type of Birth,' providing a clear count that captures the sample's distribution. By including percentages, the table facilitates proportional analysis, enabling a quick assessment of each category's relative frequency. Both the chart and the table reveal that vaginal births were the most frequent, followed by Caesarean – Emergency. The ‘Not Known’ category, with a count of 9, accounts for merely 0.25% of the dataset. Although its incidence is low, making it plausible to exclude, it was decided to retain this category to maintain the integrity of the dataset's entirety.

Table 3 - Tally for Type of birth

Tally

Type of Birth	Count	Percent
Breech	140	3.83
Caesarean - Elective	437	11.96
Caesarean - Emergency	798	21.85
Forceps	690	18.89
Not Known	9	0.25
Vacuum	602	16.48
Vaginal	977	26.75
N=	3653	

Question 3 – Bump Size Analysis

Within Minitab, a subset of the data was generated to concentrate exclusively on the "Vaginal" and "Caesarean – Emergency" categories within the "Type of Birth" variable with 1775 rows. Figure 4 illustrates a boxplot of bump size stratified by type of birth. Upon initial inspection, the medians and interquartile ranges appeared remarkably similar.

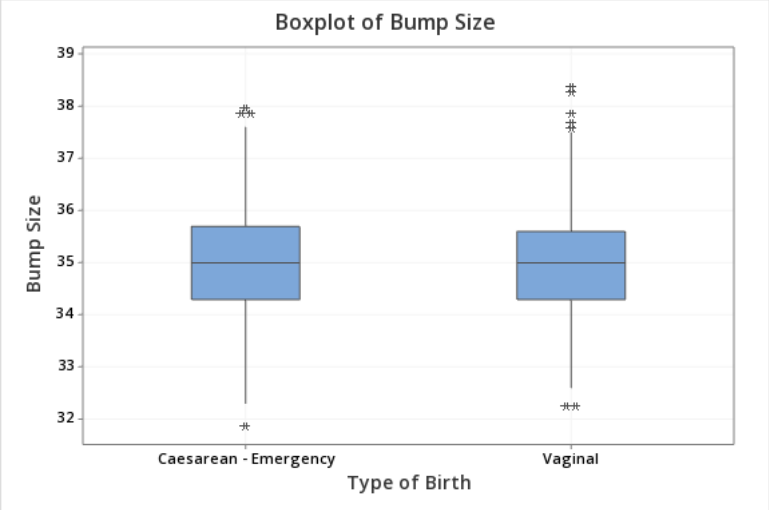

Figure 4 - Boxplots of Bump Size by Type of Birth

Given the nature of the data, the employment of a t-test was contingent upon its conformity to the relevant statistical assumptions. Although Figure 5 indicated preliminary sings of normal distribution for the variables, this observation alone needed further validation. Should the data have not followed a normal distribution, the consideration of a non-parametric alternative, such as the Mann-Whitney U test would have been warranted.

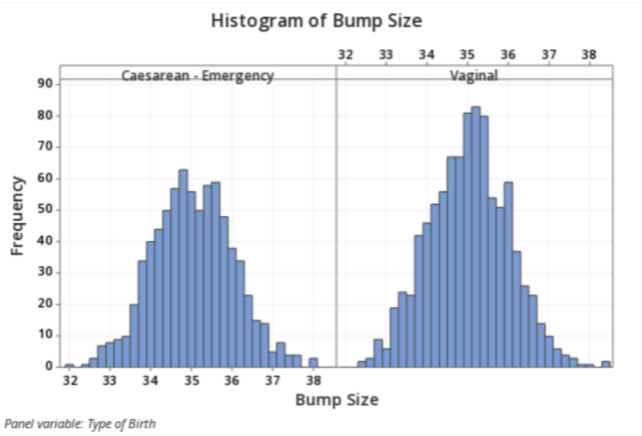

Figure 5 - Histograms of Bump Size by Type of Birth

Figure 6 shows the output from Minitab for the probability plot of the bump size for the “Caesarean – Emergency” and “Vaginal” variables. Notice the P-Values of 0.132 and 0.089 for the “Caesarean – Emergency” and “Vaginal” variables respectively. Given these P-values, both groups’ bump size does not significantly deviate from normality suggesting the assumption for conducting a t-test are met.

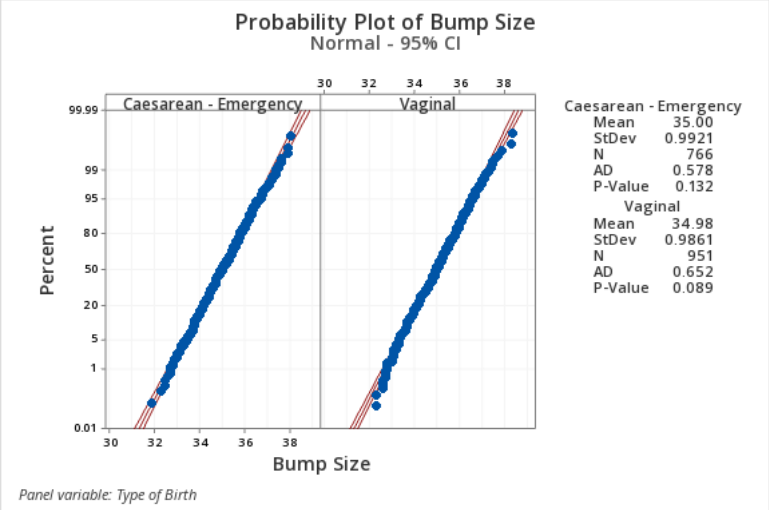

Figure 6 - Probability Plot of Bump Size

A Two-Sample T-Test was performed in Minitab with the output shown in Figure 7 below. A T-value of 0.51 is relatively low, suggesting that the mean bump size between the 2 group is not very different. The Degrees of Freedom (DF) which is 1633 is the function of the sample sizes of the two groups suggesting a large sample size for this example. The P-Value of 0.612 is much higher than the common alpha level of 0.05, indicating that there is no statistically significant difference in bump size between the 2 groups.

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
0.51	1633	0.612

Figure 7 - T-Test Output

Question 4 – Proportion of Patients who Smoked

The values for those that smoked is a binary choice, '0' representing the participant did not smoke during the pregnancy, or '1' representing they did smoke during pregnancy. Figure 8 below shows that 3232 did not smoke during pregnancy, while 421 did.

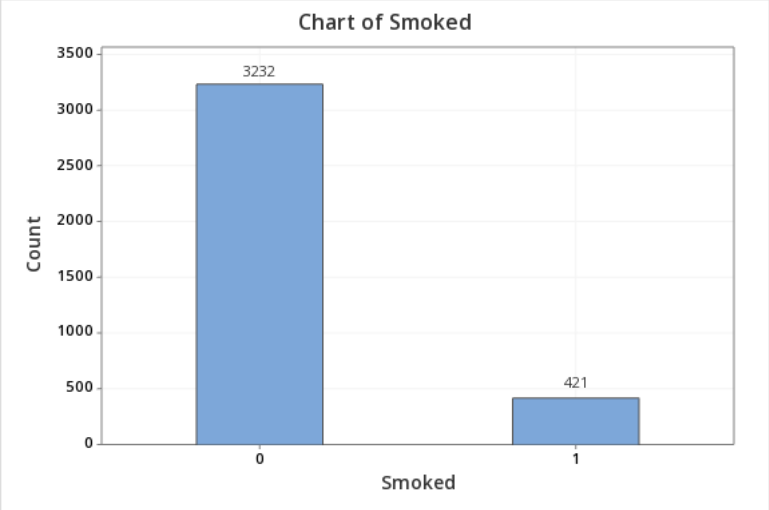

Figure 8 - Bar chart of Smoked

Table 4 shows below that 11.52% of patients smoked while pregnant while 88.48% did not. In total there are 3653 values, and none are missing.

Table 4 - Descriptive Statistics: Smoked

Statistics

Variable	Smoked	Total Count	N	N*	CumN	Percent
Smoked	0	3232	3232	0	3232	88.4752
	1	421	421	0	3653	11.5248

Figure 9 shows the results for calculating the 95% confidence interval for the proportion of patients who smoked while pregnant. It used the approximation method, which is appropriate for large sample sizes when the sample proportion is not extremely close to 0 or 1. The confidence interval gives us a range of plausible values for the true proportion of the population. With 95% confidence, we can say the proportion who smoked is between 10.49% and 12.56% which has a width of 2.07%. Given the sample size of 3653, the width of the interval is relatively narrow, indicating a decent level of precision.

Method

p: event proportion

Normal approximation method is used for this analysis.

Descriptive Statistics

N	Event	Sample p	95% CI for p
3653	421	0.115248	(0.104893, 0.125603)

Figure 9 - Minitab output proportion of patients who smoked

Question 5 – Feeding Type and Maternal BMI

To investigate the potential association between Feeding Type and Maternal BMI, a preliminary data visualisation was conducted by plotting the distributions of both variables using bar charts. This exploratory step is crucial as it provides a visual assessment of the data, allowing for the identification of any patterns or inconsistencies that could influence the outcome of a Chi-square test.

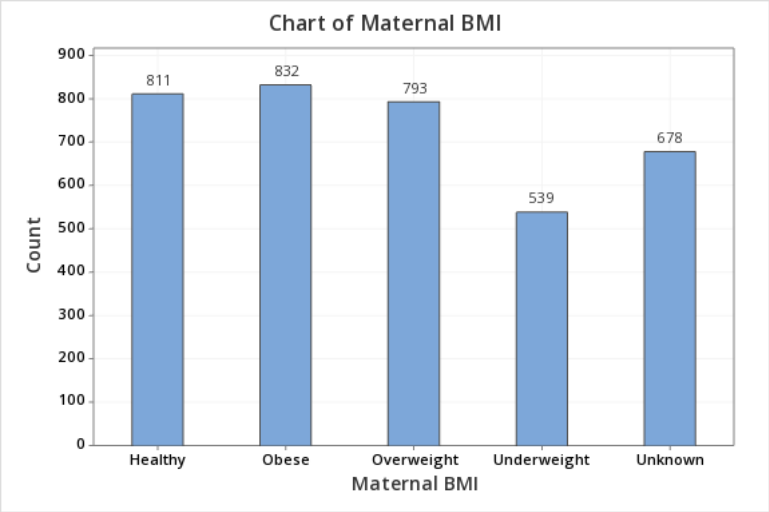

Figure 10 - Bar Chart of Maternal BMI

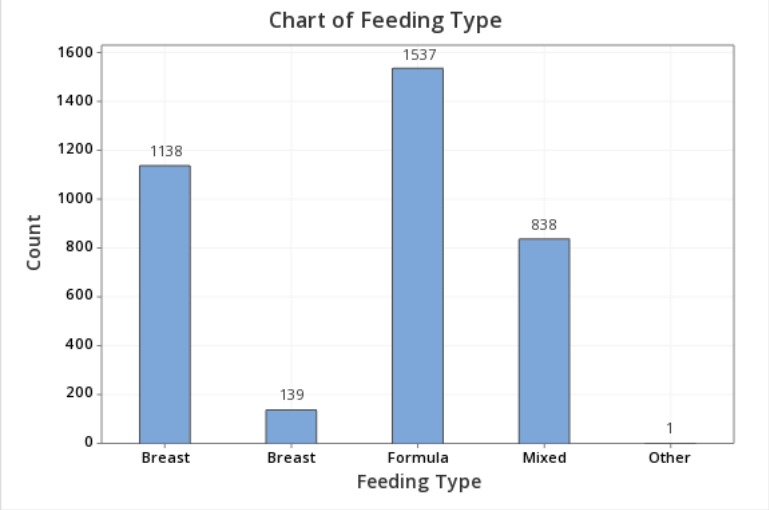

Figure 11 - Bar chart of Feeding Type

Before executing the Chi-square test, a meticulous data cleansing process was essential to address several issues within the dataset that could potentially skew the results. Initial scrutiny revealed the presence of an "Unknown" category within the Maternal BMI

variable. Additionally, the Feeding Type variable contained a nominal "Other" category. Such instances of ambiguous or sparse data can undermine the robustness of the Chi-square test, as the test's accuracy is predicated on enough observations in each cell of the contingency table. All expected cell frequencies must be greater than 5 to warrant the application of the test ensuring a reliable p-value.

Upon further examination, a discrepancy was detected within the Feeding Type variable—two distinct categories labelled as "Breast" were identified. This duplication was posited to be an error and thus, the less prevalent category with 139 occurrences was removed.

Adjacent to the observed values, the expected frequencies are displayed. These are derived from the marginal totals for each category, assuming the feeding types would be uniformly distributed.

Table 5 shows the observed frequencies of the Feeding Type categories across four Maternal BMI classification. Adjacent to the observed values, the expected frequencies are displayed. These are derived from the marginal totals for each category, assuming the feeding types would be uniformly distributed.

Table 5 - Observed Frequencies Table

Rows: Feeding Type Columns: Maternal BMI

	Healthy	Obese	Overweight	Underweight	All
Breast	257	254	243	172	926
	254.0	257.2	246.8	168.0	
Formula	336	354	338	230	1258
	345.0	349.4	335.3	228.2	
Mixed	190	185	180	116	671
	184.0	186.4	178.9	121.7	
All	783	793	761	518	2855

Table 6 shows below the Chi-Square test results, the Chi-square statistic measures how much the observed frequencies deviate from the expected frequencies, a low value like 1.047 indicates a small deviation. The P-value provides the probability of observing a Chi-square statistic at least as extreme as the one calculated, assuming the null hypotheses is true. The high P-value of 0.984 is much greater than the typical alpha value of 0.05, leading us to fail to reject the null hypotheses. From this we can concluded that there was no evidence of an association between Feeding Type and Maternal BMI.

Table 6 - Chi-square Test results

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	1.043	6	0.984
Likelihood Ratio	1.047	6	0.984

Probability

Question 1 – Predicting Outcomes for future mothers

Observed counts of mothers who smoked

Table 7 presents a cross-tabulation of the number of women categorised by their smoking status during pregnancy and whether they tried breastfeeding from the dataset. These values were obtained from Minitab.

Table 7 - Counts of mothers who smoked

	Did Not Smoke	Smoked	Total
Did Not Try Breastfeeding	1137	141	1278
Tried Breastfeeding	2095	280	2375
Total	3232	421	3653

Probability tree for breastfeeding

Figure 12 below visualises the likelihood of the mothers' smoking habits and whether they tried breastfeeding during pregnancy. Starting from a single root, the tree branches into two primary paths based on whether the mothers smoked or did not smoke during pregnancy. Each of these paths then further divides into whether the mothers tried or did not try breastfeeding. The final probability at the ends of each path represents the combined likelihood of each scenario occurring within the population studied. Python 3.12.2 and matplotlib 3.8.3 were used to create the tree graphic only.

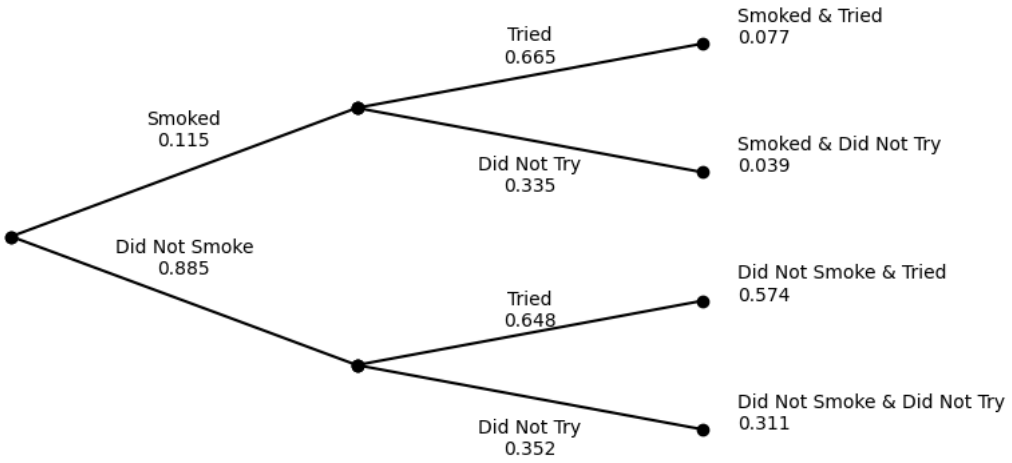

Figure 12 - Probability Tree of Smoking & Breastfeeding during Pregnancy

Elementary elements and associated probabilities

1. Mothers Who Smoked and Tried Breastfeeding:
    - Elementary Event: E_1 = Smoked, Tried Breastfeeding
    - Probability: $P(E_1)$ = Probability of Smoking x Probability of Trying Breastfeeding | Smoked = 0.115 x 0.335 = 0.077
2. Mothers Who Smoked and Did Not Try Breastfeeding:
    - Elementary Event: E_2 = Smoked, Did Not Try Breastfeeding
    - Probability: $P(E_2)$ = Probability of Smoking x Probability of Not Trying Breastfeeding | Smoked = 0.115 x 0.335 = 0.039

3. Mothers Who Did Not Smoke and Tried Breastfeeding:
    - Elementary Event: E_3 = Did Not Smoke, Tried Breastfeeding
    - Probability: $P(E_3)$ = Probability of Not Smoking x Probability of Trying Breastfeeding | Did Not Smoked = 0.885 x 0.648 = 0.574
4. Mothers Who Did Not Smoke and Did Not Try Breastfeeding:
    - Elementary Event: E_4 = Did Not Smoke, Did Not Try Breastfeeding
    - Probability: $P(E_4)$ = Probability of Not Smoking x Probability of Not Trying Breastfeeding | Did Not Smoke = 0.885 x 0.352 = 0.311

Probability that mother smoked given she tried breastfeeding

Bayes' Theorem is given below:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

A is the event that a mother smoked during pregnancy.

B is the event that a mother tried breastfeeding.

When we input the values:

$$(A|B) = \frac{0.655 \times 0.115}{0.650}$$

$$(A|B) = 0.118$$

Probability that in a sample of 40 new mothers at least one smoked

To calculate the probability that at least one of the 40 new mothers who tried breastfeeding smoked, we can use the complement rule.

$$P(At\ least\ one\ smoked) = 1 - P(None\ smoked)$$

Given:

$n = 40$

$p = P(Smoked) = 0.115$

First, $P(None\ smoked)$was calculated using binomial probability for $k = 0$ (where k is the number of successes, i.e. instances of smoking):

$$P(k = 0) = \binom{n}{0} p^0 (1-p)^n$$

$$P(k = 0) = \binom{40}{0} 0.115^0 (1-0.115)^{40}$$

$$P(k = 0) = 7.55x10^{-3}$$

$$P(At\ least\ one\ smoked) = 1 - 7.55x10^{-3}$$

$$P(At\ least\ one\ smoked) = 0.992$$

Question 2 - Time Variable

Initial Analysis of Time Spent

Figure 13 shows a histogram of the time in days spent in hospital after giving birth. It indicates that the distribution of time spent in hospital after giving birth is right-skewed, with most mothers staying for a shorter duration.

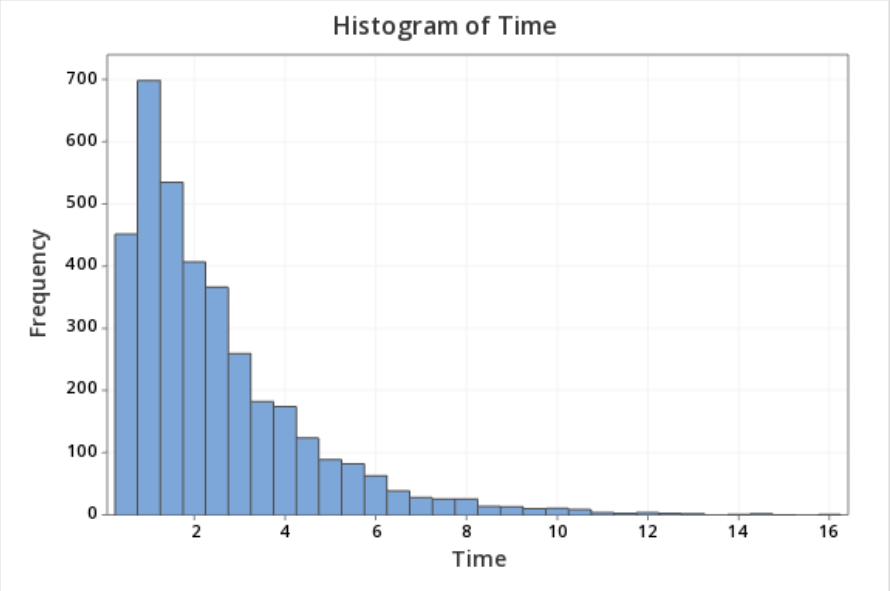

Figure 13 - Histogram of Time Spent in Hospital

Table 8 presents the descriptive statistics for the variable time. The skewness coefficient of 1.99 corroborates the presence of a right skewed distribution as seen in Figure 13. It is important to note the mean of 2.53 days, which is slightly higher than the median value of 2 days agreeing with previous observations suggesting most stays are relatively short.

Table 8 - Minitab Descriptive Statistics Output

Statistics

Variable	Total Count	N*	Mean	SE Mean	StDev	Variance	Q1	Median	Q3	Range
Time	3653	0	2.5361	0.0347	2.0954	4.3907	1.0000	2.0000	3.5000	15.5000

Variable	Mode	N for Mode	Skewness
Time	1	698	1.99

Probability Distribution

Given the empirical distribution's characteristics a Gamma distribution could describe the time spent in hospital after giving birth. The Gamma distribution is described by two parameters, shape (α) and scale (θ) and its probability density function is given by:

$$\int(x;\alpha;\theta) = \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)}, x > 0, \alpha > 0, \theta > 0$$

To relate the Gamma distribution parameters to the dataset, we can use the mean and variance as follows:

Mean $= \alpha\theta$

Variance = $\alpha\theta^2$

From the descriptive statistics:

Mean = 2.54 $days$

Variance = 4.39 $days^2$

Solving The equations for $\alpha$and θ:

$\alpha = \frac{\mu^2}{\sigma^2}$

$\theta = \frac{\sigma^2}{\mu}$

Calculating the values of α and θ:

Shape parameter $(\alpha) = 1.46$

Scale parameter $(\theta) = 1.73$

Therefore, the PDF of Gamma Distribution that might describe the time spent in hospital after giving birth, using these parameter values, is:

$$\int(x;1.46;1.73) = \frac{x^{1.46-1}e^{-\frac{x}{1.73}}}{1.73^{1.46}\Gamma(1.46)}, x > 0$$

Fitting a selection of probability distributions

Figure 14 shows the probability plots for the time spent in hospital after giving birth, each compared to four different theoretical distributions: Normal, Exponential, Lognormal, and Gamma. A probability plot is a graphical technique to assess whether a data set follows a given distribution.

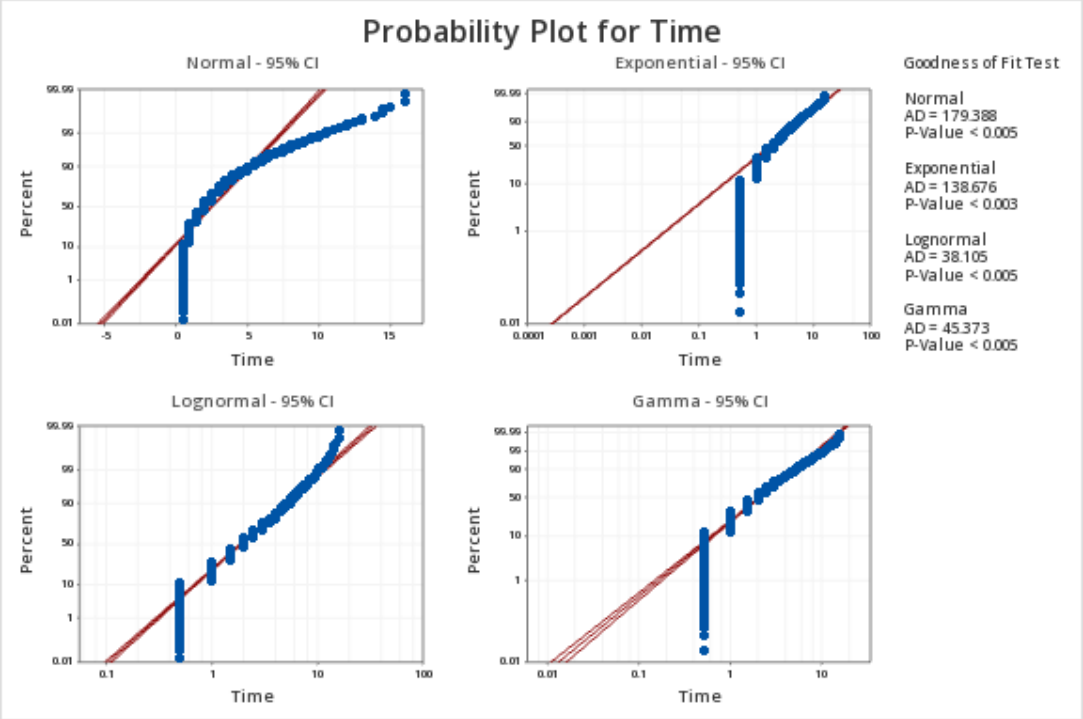

Figure 14 - Probability plots

- The red line represents the fitted theoretical distribution.
- The blue dots represent the empirical data.
- The confidence bands around the line show the 95% confidence intervals.

Table 9 shows the results of a goodness-of-fit test for different theoretical distributions compared to the dataset. The Anderson-Darling (AD) statistic is used here to quantify how well the data follows a given distribution. Lower values of the AD statistic suggest a better fit, a lower P-value indicates there is stronger statistical evidence against the null hypothesis.

Table 9 - Minitab output Goodness of Fit Test

Goodness of Fit Test

Distribution	AD	P
Normal	179.388	<0.005
Exponential	138.676	<0.003
Lognormal	38.105	<0.005
Gamma	45.373	<0.005

Despite none of the p-values being above the conventional 0.05 threshold, which would suggest a good fit, in practice, the Lognormal distribution's lower AD value compared to the Gamma makes it a more compelling choice for modelling this variable.

References

Matplotlib. (2024, March 26). *MatplotLib 3.8.3*. Retrieved from PyPi: https://pypi.org/project/matplotlib/

Minitab. (2024, March 26). *Minitab*. Retrieved from Minitab: https://www.minitab.com/en-us/

Public Health Scotland. (2024, March 26). *Births in Scotland*. Retrieved from Public Heatlh Scotland: https://www.opendata.nhs.scot/dataset/births-in-scottish-hospitals

python. (2024, March 26). *Downloads*. Retrieved from Python: https://www.python.org/downloads/