Nour Moghazi (900225966)
Omar Moustafa (900222400)
Abdelrahman Baioumi (900223218)
Nour Kahky (900221042)

Statistical Inference Project Report

1. Introduction:

In this project, the main goal is to thoroughly analyze and interpret a real dataset that was obtained from a sample of clients of a wholesale distributor. The objectives include understanding the spending patterns of clients across different product categories, drawing conclusions about the population, and making informed data analysis decisions.

The research objectives are the following:

- Utilize descriptive statistics to observe and understand the main characteristics of the dataset
- Utilize inferential statistics to estimate population parameters as well as hypothesis testing
- Explore spending habits across multiple different product categories, channels, and regions
- Find gender differences in spending behavior

To achieve the research objectives listed above, the following procedures will be taken:

- Utilize descriptive statistics to summarize and analyze the main characteristics of the dataset using measures, tables, and graphs
- Utilize inferential statistics such as point estimation, interval estimation, and hypothesis testing
- Estimate parameters such as the mean, proportion, and variance, and calculate the 95% confidence intervals for each of these estimates
- Explore the difference between means and proportions of multiple different groups and test certain hypotheses about population characteristics

By using these particular techniques, gaining a deeper understanding of how people choose to spend their money and make informed decisions will be achieved.

2. Descriptive Statistics Section:

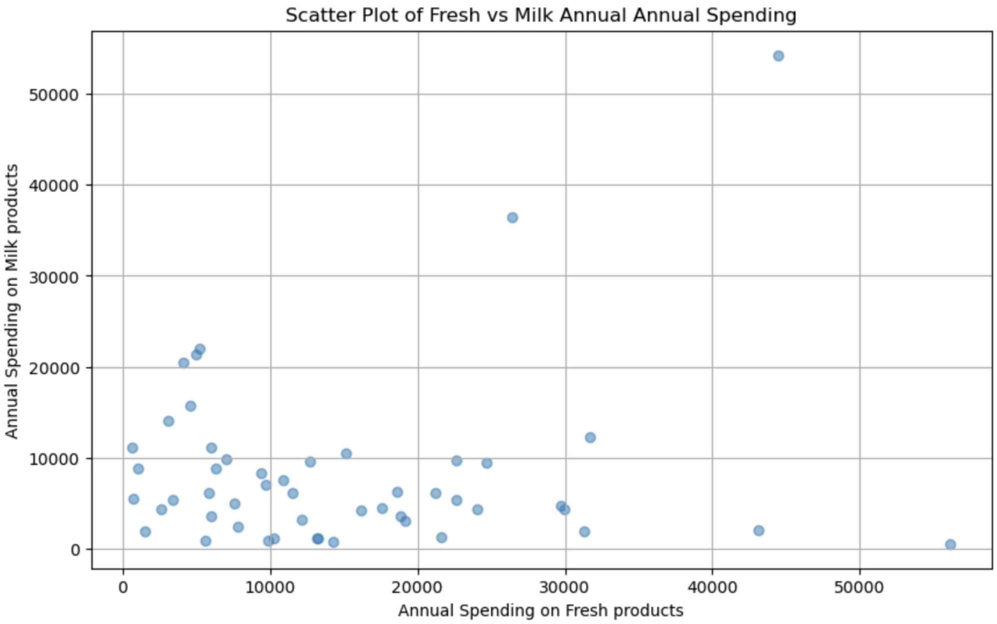
The dataset contains the following variables:

1. Channel: A categorical attribute indicating the type of sales channel (e.g., retail or hotel/restaurant/cafes).
2. Region: A categorical attribute indicating the region where customers are located.
3. Fresh: Annual spending on fresh products (continuous).
4. Milk: Annual spending on milk products (continuous).
5. Grocery: Annual spending on grocery products (continuous).
6. Frozen: Annual spending on frozen products (continuous).
7. Detergents Paper: Annual spending on detergents and paper products (continuous).
8. Delicatessen: Annual spending on delicatessen products (continuous).
9. Gender: A categorical attribute (seems incorrectly included given the context; might represent a binary attribute with values 1 or 0).

Descriptive Table of All Variables with Measures:

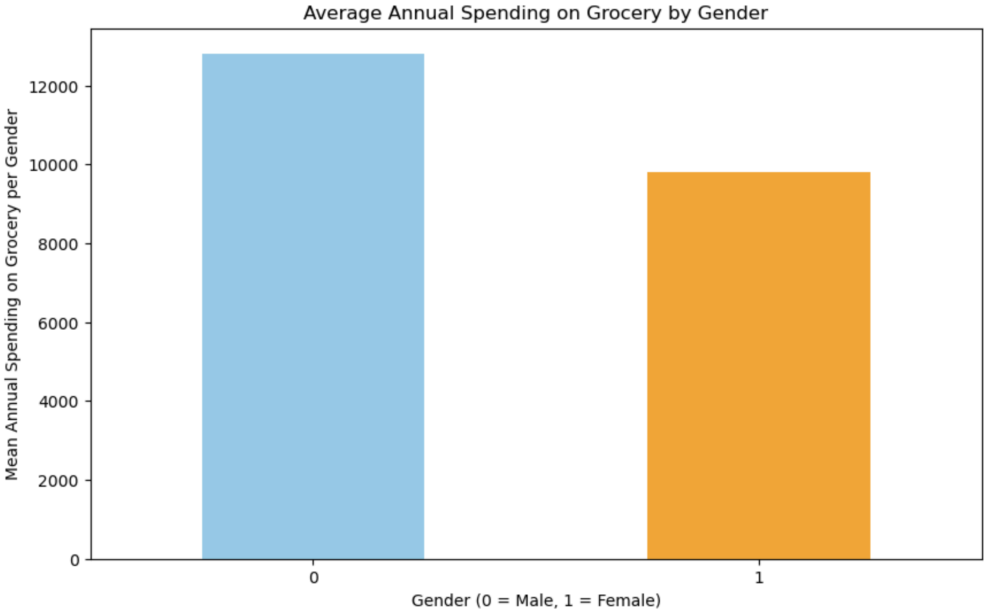
	Mean	**Median**	**Range**	**Standard Deviation**	**Variance**
Channel	1.62	2	1	0.49031	0.24041
Region	3	3	0	0	0
Fresh	14921.18	11822.5	55529	12183.13	148428656.59
Milk	8225.64	5450.5	53704	9442.35	89157973.52
Grocery	10710.01	7561	55016	9376.45	87917868.99
Frozen	2243.3	1310	9969	2561.36	6560576.9
Detergents Paper	3912.6	2953.5	24071	4109.37	16886921.77
Delicatessen	2380.92	1451.5	16477	2825.15	798146.4
Gender	0.68	1	1	0.4712	0.222

Graphs (Histograms, Bar Charts, & Scatter Plots):



Interpretation of the above Scatter Plot:

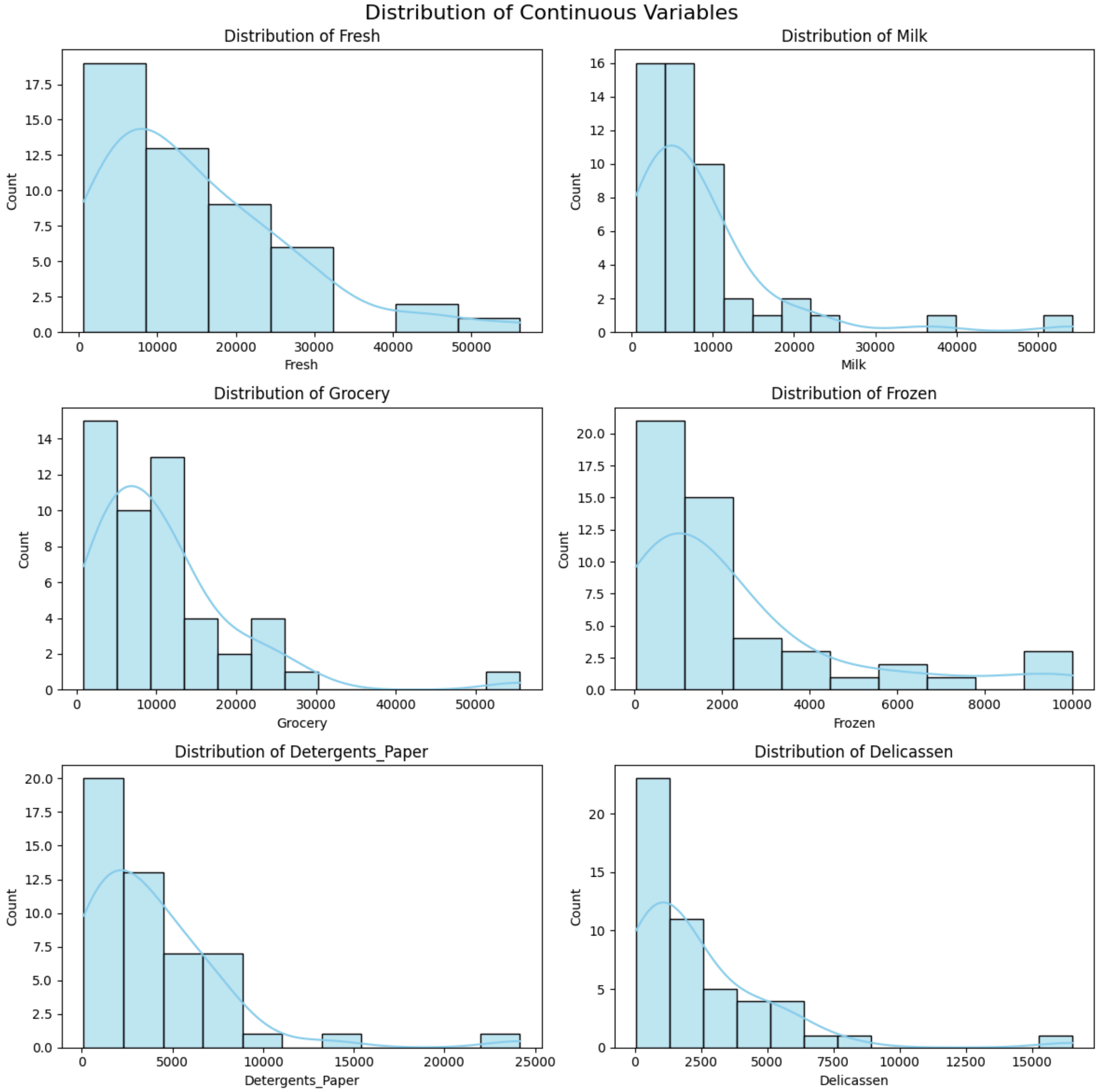
The above visualization indicates that there might be a strong negative correlation between the annual spending on fresh products and the annual spending on milk products. In other words, customers who spend more on fresh products tend to spend less on milk products and those who spend more on milk products tend to spend less on fresh products.



Interpretation of the above Bar Chart:

From the above visualization, it is very clear that the average annual spending on groceries for males is significantly higher than for females as it shows that the bar for the annual spending on groceries for males is much higher than for females, indicating that, on average, males tend to spend much more annually on groceries on an in comparison to females.

Distributed Histogram Analysis



Visual Analysis:

- **Fresh**, **Milk**, and **Grocery** items have a right-skewed distribution, which means most customers spend less, with a few spending a lot.
- **Frozen** and **Detergent Paper** also show a right-skew but have a slightly flatter spread, suggesting more uniform spending in these categories compared to the others.
- **Delicatessen** spending is highly skewed, with most customers spending small amounts and a few outliers spending significantly more.

3. Inferential Statistics

1. FRESH

    a. Choose one parameter of interest suggest a good point estimator and calculate its value

Parameter of Interest: Mean (μ)

Point Estimator: Sample Mean ($\bar{x}$)

Calculation: Sample Mean = $\bar{x}$ = Sum of all data points / Number of data points =
(12669 + 7057 + … + 11519 + 4967) / 50 = 746059 / 50 **= 14921.18**

Explanation: Due to its efficiency and unbiasedness, the sample mean is regarded as a reliable estimator of the population mean. The sample mean tends to be around the population mean when the sample size is large, which in this case it is indeed since the sample size is equal to 50, making the sample mean an accurate estimator.

   b. Get an interval estimate at a 95% confidence level for each parameter proposed in (a) and interpret your results.

Calculation: x bar ± Z * (s/(n^0.5))
—> x bar = Sample Mean = 14921.18
—> Z = z-score that corresponds to the 95% confidence interval = 1.96
—> s = sample standard deviation = 12183.13
—> n = sample size = 50
—> CI = 14921.18 ± 1.96 * (12183.13 / (50^0.5)) **= [11544.19, 18298.17]**
—> We are 95% confident that the population mean lies in the interval **[11544.19, 18298.17]**

2. MILK

    a. Choose one parameter of interest suggest a good point estimator and calculate its value

Parameter of Interest: Population variance (σ^2)

Point Estimator: Sample variance (s^2)

Calculation: Sample Variance = s^2 = (1/(n-1)) * ∑ (xi - $\bar{x}$)^2 = 89157973.52

Explanation: The sample variance is a good choice for a point estimator for the population variance as it is an unbiased estimator and it works to measure the average squared deviation of the data points from their mean which gives an estimate of the spread of the data.

b. Get an interval estimate at a 95% confidence level for each parameter proposed in (a) and interpret your results.

CI = $(((n-1)s^2) / X^2,_{\alpha/2}\ , ((n-1)s^2) / X^2,_{1-\alpha/2}\ )$

—> For a 95% CI: $\alpha/2 = 0.025$, $1 - \alpha/2 = 0.975$

—> n = sample size = 50

—> $((n-1)s^2) / X^2,\ _{\alpha/2 =}$ (50-1) * 89157973.52 / X^2, 0.025 = 4368740702.48 / 71.42
= 61169710.2

—> $((n-1)s^2) / X^2,\ _{\alpha/2 =}$ (50-1) * 89157973.52 / X^2, 0.975 = 4368740702.48 / 30.96
= 141109195.82

—> CI = **[61169710.2, 141109195.82]**

—> We are 95% confident that the population variance lies in the interval **[61169710.2, 141109195.82]**

3. Average Annual Grocery Spending Per Gender

a. Choose one parameter of interest suggest a good point estimator and calculate its value

Parameter of Interest: Population mean difference ($\mu 1 - \mu 2$) between the mean annual grocery spending per gender

Point Estimator: Sample mean difference ($\bar{x}1$ - $\bar{x}2$) between the mean annual grocery spending per gender

Calculation:

$\bar{x}1$ = Mean annual grocery spending for males = (1/n1) * ∑x1i —> n1 = 16

—> $\bar{x}1$ = (18881 + 12974 + … + 21955 + 55571)/16 = 12814.94

$\bar{x}2$ = Mean annual grocery spending for females = (1/n2) * ∑x2i —> n2 = 34

—> $\bar{x}2$ = (7561 + 9568 + … + 10868 + 28921)/34 = 9818.44

—> $\bar{x}1$ - $\bar{x}2$ = 12814.94 - 9818.44 **= 2996.5**

Explanation: This is a good point estimator as it will lead to assessing the difference in average spending between genders which would lead to getting an estimate of how much higher or lower the spending is for one gender compared to the other.

b. Get an interval estimate at a 95% confidence level for each parameter proposed in (a) and interpret your results.

CI = ($\bar{x}1$ - $\bar{x}2$) ± MOE = ($\bar{x}1$ - $\bar{x}2$) ± t * ((s1^2/n1)+(s2^2/n2))^0.5

—> ($\bar{x}1$ - $\bar{x}2$) = 2996.5, n1 = 16, s1^2 = 182597860, n2 = 34, s2^2 = 44585344

—> df = 16 + 34 - 2 = 48 degrees of freedom —> $t_{\alpha/2}$ for a 95% CI at 48 df = 2.013

—> MOE = 2.013 * ((182597860/16)+(44585344/34))^0.5 = 2.013 * 3567.03 = 7180.43

—> CI = 2996.5 ± 7180.43 = **[-4183.93, 10176.93]**

—> We are 95% confident that the difference between population means lies in the interval **[-0.52, 0.036]**

4. Proportion of Channel 1 with respect to Proportion of Channel 2

a. Choose one parameter of interest suggest a good point estimator and calculate its value

Parameter of Interest: Population proportion difference (p1 - p2) between Channel 1 and 2

Point Estimator: Sample proportion difference ($\hat{p}1$ - $\hat{p}2$) between Channel 1 and 2

Calculation: n1 = x, n2 = y, N = 50

$\hat{p}1$ = n1/N = 19/50 = 0.38

$\hat{p}2$ = n2/N = 31/50 = 0.62

—> $\hat{p}1$ - $\hat{p}2$ = 0.38 - 0.62 **= -0.24**

Explanation: This parameter of interest and point estimator were chosen to directly assess the difference in proportions between Channel 1 and 2 which give an estimate of how much higher or lower the proportion of Channel 1 is in comparison to Channel 2 showing the potential differences in customer distribution across the two channels.

b. Get an interval estimate at a 95% confidence level for each parameter proposed in (a) and interpret your results.

CI = ($\hat{p}1$ - $\hat{p}2$) ± MOE = ($\hat{p}1$ - $\hat{p}2$) ± Z * (($\hat{p}1$(1-$\hat{p}1$)/n1)+($\hat{p}2$(1-$\hat{p}2$)/n2))^0.5

—> ($\hat{p}1$ - $\hat{p}2$) = 0.38 - 0.62 = -0.24

—> Z = z-score that corresponds to the 95% confidence interval = 1.96

—> (($\hat{p}1$(1-$\hat{p}1$)/n1)+($\hat{p}2$(1-$\hat{p}2$)/n2))^0.5 = ((0.38(1-0.38)/19)+(0.62(1-0.62)/31))^0.5 = 0.141

—> MOE = 1.96 * 0.141 = 0.27636

—> CI = -0.24 ± 0.27636 = **[-0.51636, 0.03636]**

—> We are 95% confident that the difference between population means lies in the interval **[-0.51636, 0.03636]**

3) c. Choose two main variables in your dataset and get a point and interval estimate (with a 0.05 significance level) for the difference between the means/proportions of 2 different groups.

1. Average Fresh Spending for each Gender

n_1(males) = 16

n_2 (females) = 34

σ_1 = 11840.5 = 11841

σ_2 = 12505.34 = 12505

$\bar{x}_1$= 15552.5 =15553

$\bar{x}_2$ = 14624.09 = 14624

- Point Estimate:

$\bar{x}_1 - \bar{x}_2$= 15553 - 14624 = 929

- Interval Estimate:

$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \times \sqrt{\sigma_1/n_1 + \sigma_2/n_2}$

929 ± 1.96*((11841/16)+(12505/34))^0.5

929 ± 1.96*(740+368)^0.5

929 ± 1.96*(33.3)

We are 95% confident that the difference between genders' annual spending on fresh products means lies in the interval **[863.758, 994.242]**

2. Milk Variable in Both Channels

- Point Estimate:

  The point estimate is simply going to be the difference between the mean annual amount spent on milk in channel 1($\bar{x}_1$) and channel 2 ($\bar{x}_2$).

  $\bar{x}_1$= 2611

  $\bar{x}_2$ = 11667

A sufficient point estimator would be ($\bar{x}_2 - \bar{x}_1$) which in this case equals 11667 - 2611 = 9056.

- Interval Estimate:

  The function for the 95% confidence interval is defined as:

  $(\bar{x}_2 - \bar{x}_1) \pm z_{\alpha/2} \times \sqrt{\sigma_1/n_1 + \sigma_2/n_2}$

To get the interval estimate, we must calculate the standard deviation for milk in channel 1 (σ_1) and channel 2 (σ_2), the *z* value, and the number of entries for each channel (n_1, n_2).

$n_1 = 19$

$n_2 = 31$

$\sigma_1 = 1648$

$\sigma_2 = 10573$

The *z* value for a 95% confidence interval is 1.96

Plugging the values into the equation, we get the 95% confidence interval: **[9015, 9096]**

3) d. Formulate 3 claims that your research team would like to test about the population characteristics. State the claim in the form of null and alternative hypotheses, test your hypothesis, and state if you should reject or fail to reject the null hypothesis at a 0.05 significance level.

Claim 1:

The mean annual spending on fresh products of the population is no more than $20000.
H0: µ <= 20000
H1: µ > 20000
—> µ0 = 20000, $\bar{x}$ = 14921.18, s = 12183.13, n = 50
—> significance level = α = 0.05 —> Z = z-score that corresponds to (α = 0.05) = 1.645
—> ZC = ($\bar{x}$ - µ0) / (s/(n^0.5)) = (20000 - 14921.18) / (12183.13 / (50^0.5)) = 5078.82/1722.95
= 2.95 —> ZC = 2.95 lies in the rejection region
—> We **reject the null hypothesis**, H0, under α = 0.05

Claim 2:

The mean annual spending on milk products of the population is no more than $4000.
H0: µ <= 4000
H1: µ > 4000
—> µ0 = 4000, $\bar{x}$ = 8225.64, s = 9442.35, n = 50
—> significance level = α = 0.05 —> Z = z-score that corresponds to (α = 0.05) = 1.645
—> ZC = ($\bar{x}$ - µ0) / (s/(n^0.5)) = (8225.64 - 4000) / (9442.35 / (50^0.5)) = 4225.64/1335.35
= 3.16 —> ZC = 3.16 lies in the rejection region
—> We **reject the null hypothesis**, H0, under α = 0.05

Claim 3:

The mean spending of females on groceries is no more than $15000.

H_0: μ <= 15000

H_1: μ > 15000

—> μ_0 = 15000, $\bar{x}$ = 9818.44, s = 6677.23, n = 34

—> significance level = α = 0.05 —> Z = z-score that corresponds to (α = 0.05) = 1.645

—> Z_C = ($\bar{x}$ - μ_0) / (s/(n^0.5)) = (15000 - 9818.44) / (6677.23 / (34^0.5)) = 5181.56/1145.14 **= 4.52** —> Z_C = 4.52 lies in the rejection region

—> We **reject the null hypothesis**, H_0, under α = 0.05

4. Conclusion with Main Findings and Decisions

From the Descriptive Statistics section, we can conclude that males spend more money annually on groceries in general than females. We can also see that people spend the most money on milk, fresh products, and groceries with milk being the highest with an average of $9000 annually. The type of product that people spend the least amount of money on is frozen products, with an annual average of $1000. From the graphs conducted, we can confirm that milk is the most popular product and the least popular are frozen foods.

From the Inferential Statistics section, we can see that we are 95% confident that the average annual spending on fresh products for one gender is higher than the other. This calculation verifies that males do indeed spend more on groceries in general than females, which we already suspected from the graphs. The width of the confidence interval [863.758, 994.242] indicates some uncertainty about the true difference since this is a relatively large interval. The milk variable in both channels' confidence interval was [9015, 9096] and from that, we can conclude that Channel 2 is used more than Channel 1.

We suggest that the distributors should focus more on the sales of frozen products and delicatessen since they have the least amount of money spent on them. They should market them better or even offer sales and other discounts on these products to help popularize them. Furthermore, expanding on milk and fresh products can also be beneficial. Since they are already popular, they should offer more options on these products to utilize these sales using a wider variety. Making use of Channel 2 is also highly recommended as it is the more popular channel of the two.

An area we would wish to improve in the data itself would be the population size. We recommend gathering more data as this dataset only contains 50 rows. Even though this number is considered to be a large value for N, it is still relatively small. Our suggestion is to increase the population size for more valid and reliable calculations to be made in the future.

Overall, our analysis provides distributors with valuable insights, including their most and least popular products, the gender representing a more loyal customer base, as well as the preference for the more commonly used channel. By leveraging these insights, distributors can improve their product offers and marketing tactics to better suit customer preferences and increase their revenue to higher levels of success and customer satisfaction.