Problem solving with statistics

GROUP 11

GROUP MEMBERS

XIAO, LEAH /5615963

(Problem Analysis and Data Explore)

LI, JIAYI /5630918

(Data Analysis and Python Programming)

TSAI, YUN-LIEN /5672048

(Decision Making)

LEE, TING-YI /5610529

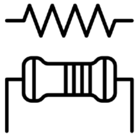
(Limitation of Data and Excel Analysis)

Contents

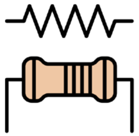
Problem Analysis

Background: The resistance manufacturer claims to offer a resistance identical to the one currently used by the kitchen manufacturer, with the same price but a longer lifespan.

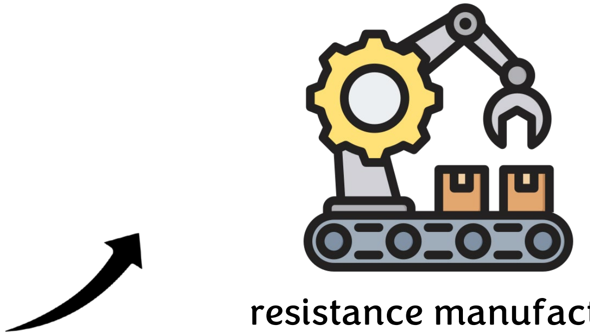
kitchen manufacturer



Old



New



resistance manufacturer

Objective: Determine whether to accept the resistance manufacturer's offer for a basic resistance material.

Accelerated Life Test	
Old (h)	New (h)
11	16
15	21
19	25
22	27
24	32
27	34
28	35
30	36
31	40
33	41
35	43
36	44
37	47
38	48
39	49
42	50
43	52
46	54
49	55
64	61

Explore the Data

Resistance Lifetime Analysis (Hours)	Old	New
Mean	33.45	40.5
Standard Error	2.778	2.711
Median	34	42
Standard Deviation	12.424	12.124
Sample Variance	154.366	147.000
Kurtosis	0.738	-0.550
Skewness	0.385	-0.380
Range	53	45
Mininum	11	16
Maxinum	64	61
25%	26.25	33.50
75%	39.75	49.25
Sum	669	810
Count	20	20

The mean of the new resistance is higher than the old resistance, but it needs to be confirmed whether this difference is statistically significant.

The degree of dispersion is very similar, and the distribution range is comparable.

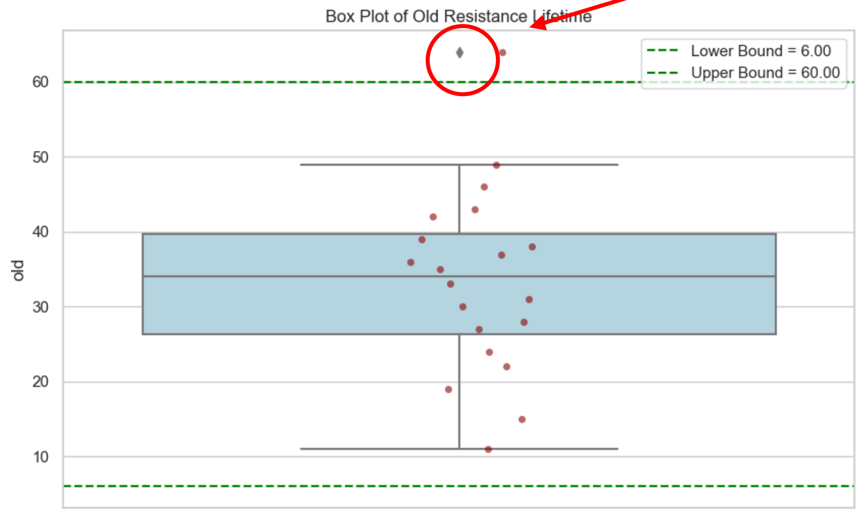
Both skewness and kurtosis are close to the theoretical values of a normal distribution (0 and 0) and can be further validated using hypothesis test.

Explore the Data

Boxplot value (Old)
Quartile 1 (Q1) = 26.25
Quartile 2 (Median) = 34
Quartile 3 (Q3) = 39.75
IQR = 13.5

Boundaries for outliers (Old)
Extreme min value = 6
Extreme max value = 60

Outlier



Boxplot value (New)
Quartile 1 (Q1) = 33.5
Quartile 2 (Median) = 42
Quartile 3 (Q3) = 49.25
IQR = 15.75

Boundaries for outliers (New)
Extreme min value = 9.875
Extreme max value = 72.875

Explore the Data

Discussion

- Investigate the source of the data
- Data collection process problem
- Experimental environment problem

Solution

- **Delete outliers:** may reduce sample size, especially when the sample size is small (n=20), which may affect the stability of statistical tests
- **Replace with median:** compresses the distribution range of data causes the data to lose variability
- **Replace with upper bound:** while limiting the effect of extreme values on the mean and variance, it partially retains the information of outliers

Statistical Comparison of Resistance Lifetimes

Analysis Approach

- Problem Definition
- Mean Difference Testing
- Test Method Selection

Normality Check

- Verification of Distribution
- Chi-square Test

Variance Analysis

- F-test for two variances

Mean Comparison

- 2 sample t-test
- Results Analysis

Outlier Analysis

- Outlier Treatment
- Reanalysis with Adjusted Data

Test Results

- t-test results on original data
- t-test results after replacing outliers

Problem Definition

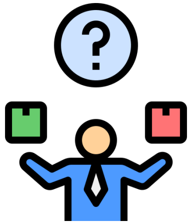
Mean Difference Testing

Test Method Selection

Do New Resistance Last Longer ?

New

μ_1



Old

μ_2

Problem Definition

Mean Difference Testing

Test Method Selection

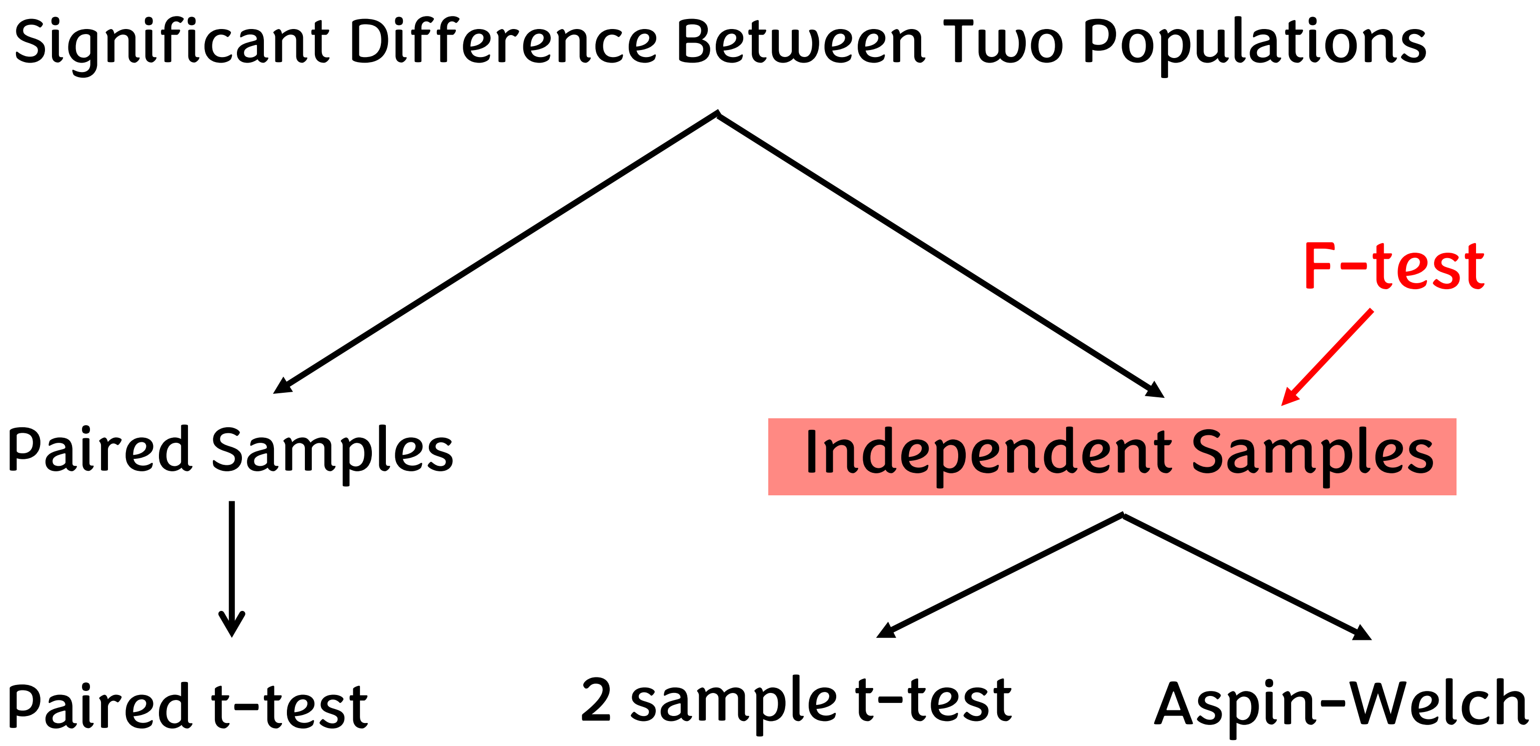
Significant Difference Between Two Populations

- Paired Samples
  - Paired t-test
- Independent Samples
  - 2 sample t-test
  - Aspin-Welch

Problem Definition

Mean Difference Testing

Test Method Selection



Null hypothesis	Histogram	CDF	Chi-Square & P - value	Decision

Null hypothesis | Histogram | CDF | Chi-Square & P - value | Decision

H_0: Old data come from a normal distribution

H_1: Old data does not come from a normal distribution

Old Resistance

H_0: New data come from a normal distribution

H_1: New data does not come from a normal distribution

New Resistance

Develop the classes	
No.of classes	5.3220
Round No.of classes	5
Class range	10.6

Old Resistance

Develop the classes	
No.of classes	5.3220
Round No.of classes	5
Class range	9.0

New Resistance

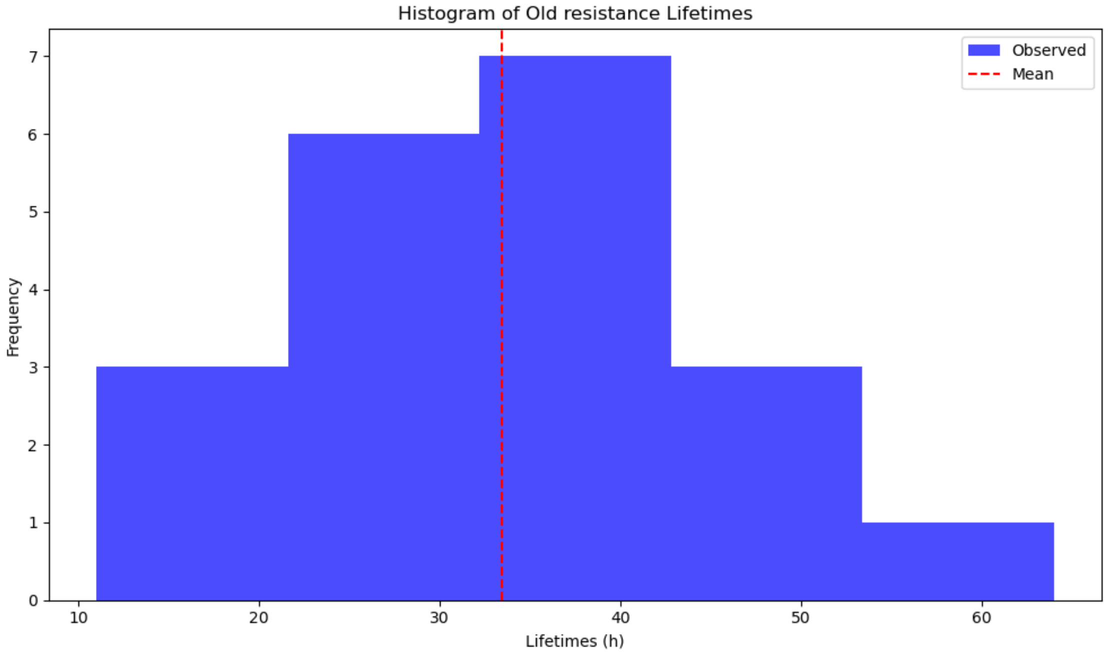
	Length	Frequency
0	21.6	3
1	32.2	6
2	42.8	7
3	53.4	3
4	64.0	1

Old Resistance

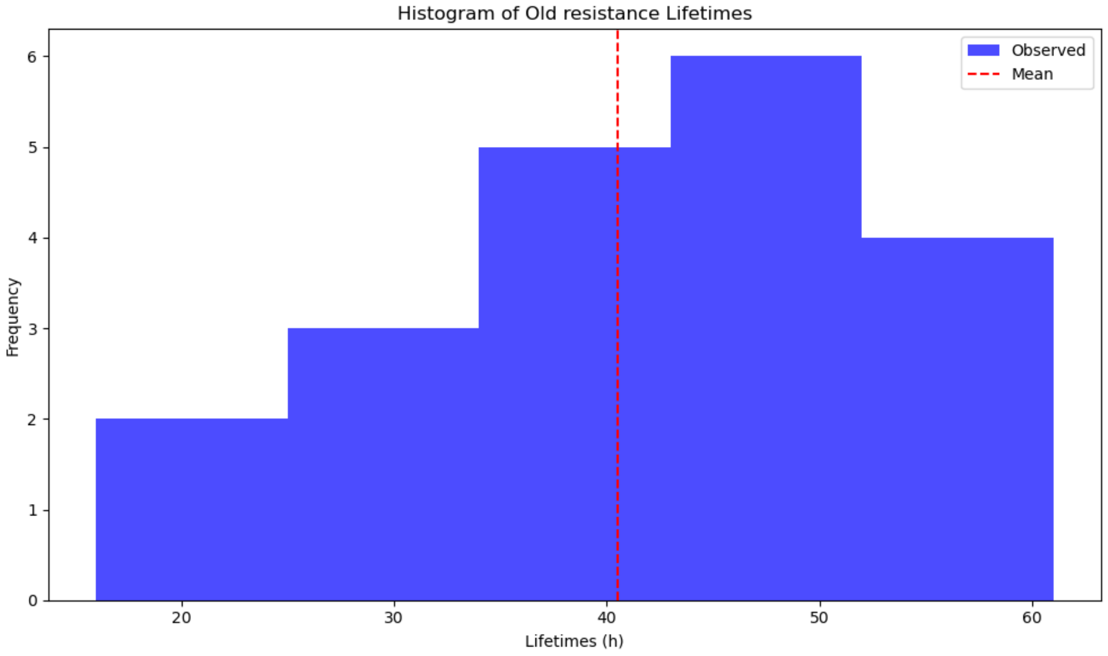
	Length	Frequency
0	25.0	2
1	34.0	3
2	43.0	5
3	52.0	6
4	61.0	4

New Resistance

Null hypothesis | **Histogram** | CDF | Chi-Square & P - value | Decision



Old Resistance



New Resistance

Null hypothesis | Histogram | **CDF** | Chi-Square & P - value | Decision

	CDF	Bin Frequency	Expected Value
0	0.170101	0.170101	3.402010
1	0.459931	0.289830	5.796603
2	0.77414	0.314210	6.284191
3	0.945831	0.171691	3.433822
4	0.993031	0.047200	0.944000
5	Sum =	0.993031	20.000000

Old Resistance

	CDF	Bin Frequency	Expected Value
0	0.100551	0.100551	2.011019
1	0.295941	0.195390	3.907799
2	0.581681	0.285740	5.714807
3	0.828564	0.246882	4.937645
4	0.954564	0.126000	2.520001
5	Sum =	0.954564	20.000000

New Resistance

Null hypothesis Histogram CDF **Chi-Square & P - value** Decision

	χ^2
0	0.047505
1	0.007137
2	0.081535
3	0.054808
4	0.003322
Sum	0.194307

P = 0.90847

Old Resistance

	χ^2
0	0.000060
1	0.210886
2	0.089408
3	0.228570
4	0.869205
Sum	1.398129

P = 0.52379

New Resistance

The old data derives from a normal distribution

- The p-value(0.90847) is larger than the α=5%
- There is no sufficient evidence to reject the null hypothesis

Old Resistance

The new data derives from a normal distribution

- The p-value(0.52379) is larger than the α=5%
- There is no sufficient evidence to reject the null hypothesis

New Resistance

Assumption Check

Method Selection

Accurate Calculation

Result Presentation

Decision

- ✓ The samples come from a normal distribution
- ✓ Samples are independent

Assumption Check

Method Selection

Accurate Calculation

Result Presentation

Decision

F test for two variances

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Where:
σ_1^2: the variance of the new resistance lifetime
σ_2^2: the variance of the current resistance lifetime

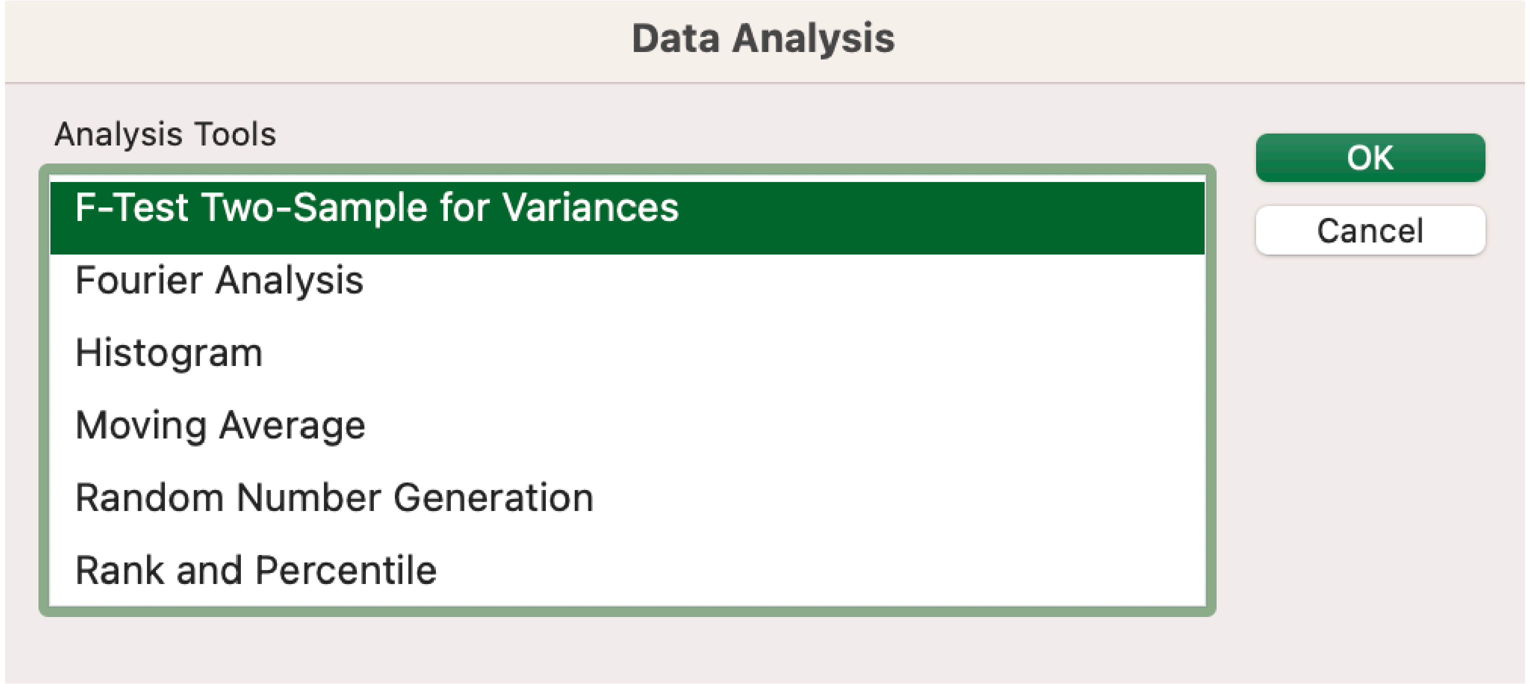
Assumption Check

Method Selection

Accurate Calculation

Result Presentation

Decision



Assumption Check

Method Selection

Accurate Calculation

Result Presentation

Decision

	old(h)	new(h)
Mean	33.45	40.5
Variance	154.3657895	147
Observations	20	20
df	19	19
F	1.050107411	
P(F<=f) one-tail	0.458123948	
F Critical one-tail	2.168251601	

$F_{crit} = F_{19,19,0.025}$ 2.526450934 (from python)

P(F<=f)*2 0.916247895 > 0.05

Assumption Check

Method Selection

Accurate Calculation

Result Presentation

Decision

No significant difference in variances between the two types of resistance

- The p-value(0.91625) is larger than the α=5%
- There is no sufficient evidence to reject the null hypothesis

Assumption Check

Method Selection

✓ **The samples come from a normal distribution**

Accurate Calculation

✓ $\sigma_1^2 = \sigma_2^2$

Result Presentation

Decision

Assumption Check

Method Selection

Accurate Calculation

Result Presentation

Decision

2 sample t-test

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Where:

μ_1: the population mean lifetime of new resistance

μ_2: the population mean lifetime of the current resistance

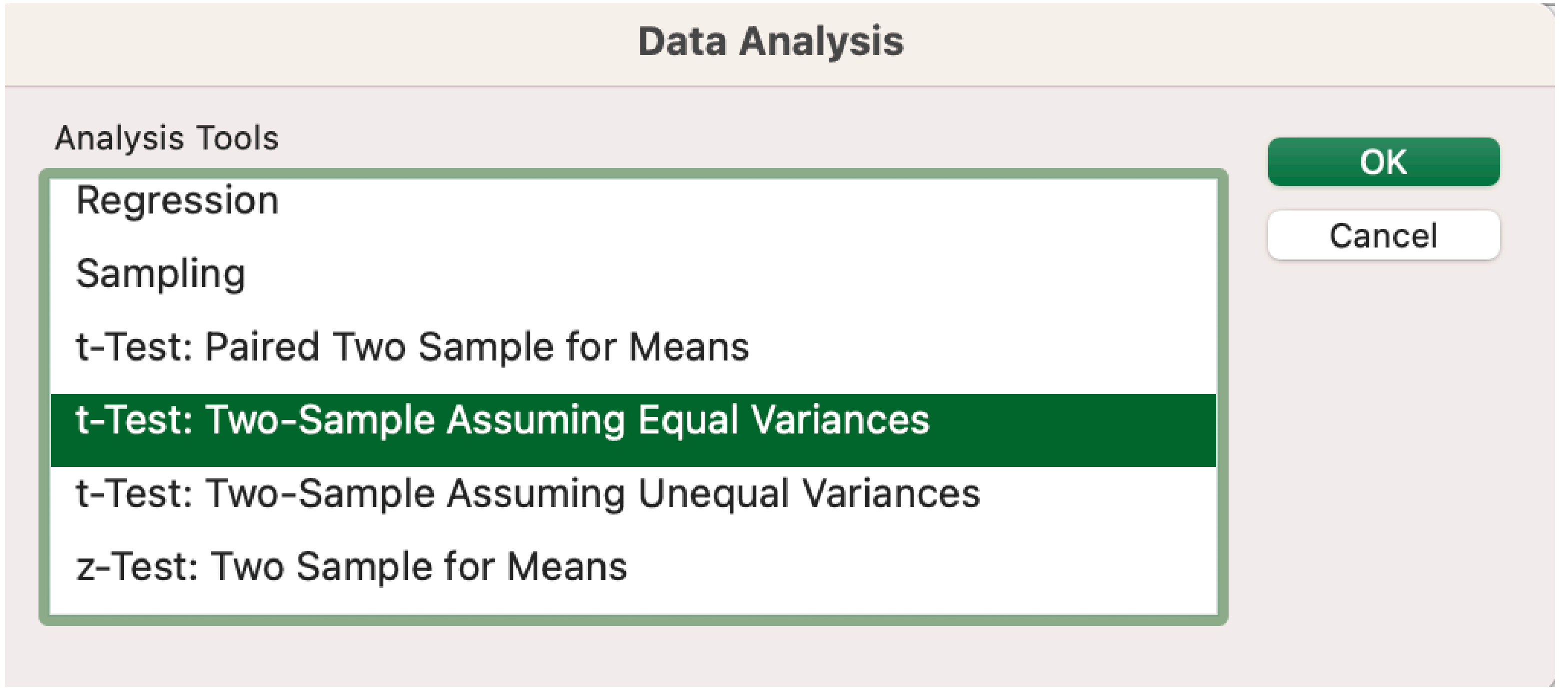
Assumption Check

Method Selection

Accurate Calculation

Result Presentation

Decision



Assumption Check

Method Selection

Accurate Calculation

Result Presentation

Decision

	new(h)	old(h)
Mean	40.5	33.45
Variance	147	154.365789
Observations	20	20
Pooled Variance	150.6828947	
Hypothesized Mean Difference	0	
df	38	
t Stat	1.816172685	
P(T<=t) one-tail	0.038619744	
t Critical one-tail	1.68595446	
P(T<=t) two-tail	0.077239488	
t Critical two-tail	2.024394164	

Assumption Check

Method Selection

Accurate Calculation

Result Presentation

Decision

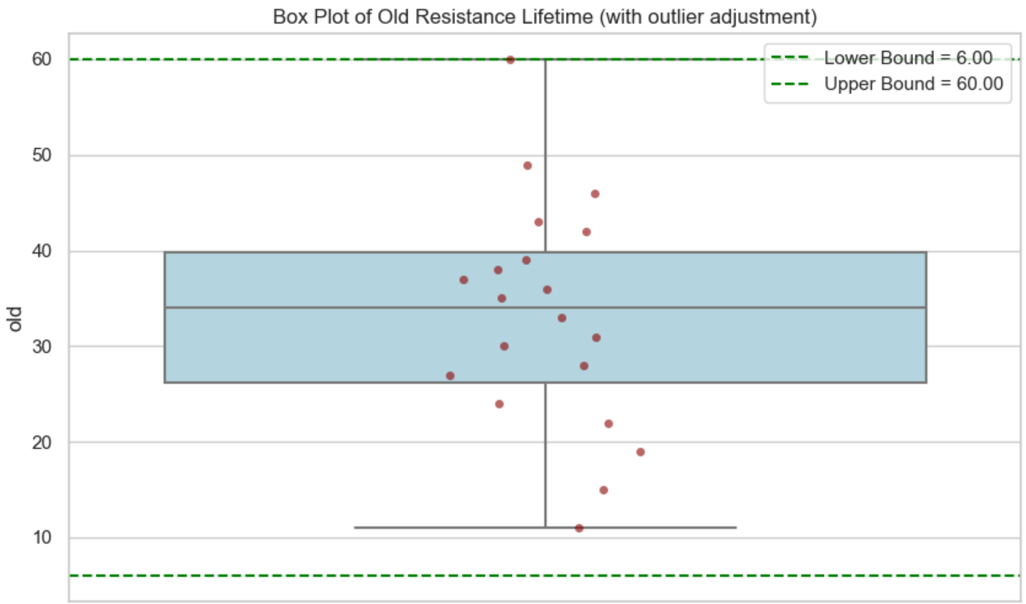
Reject H_0, the evidence suggests that the lifetimes of new resistance is greater than that of the old resistance

- The p-value(0.03862) is less than the α=5%
- There is sufficient evidence to reject the null hypothesis

Resistance Lifetime Analysis with Outlier Adjustment (Hours)	
Mean	33.2
Standard_Error	2.667
Median	34
Standard Deviation	11.929
Sample Variance	142.303
Kurtosis	0.170
Skewness	0.154
Range	49
Minimum	11
Maximum	60
25%	26.25
75%	39.75
Sum	665
Count	20

Boxplot Value and Boundaries for outliers	
Quartile 1(Q1)	26.25
Quartile 2(Q2)	34
Quartile 3(Q3)	39.75
IQR	13.5
Extreme min value	6
Extreme max value	60.0

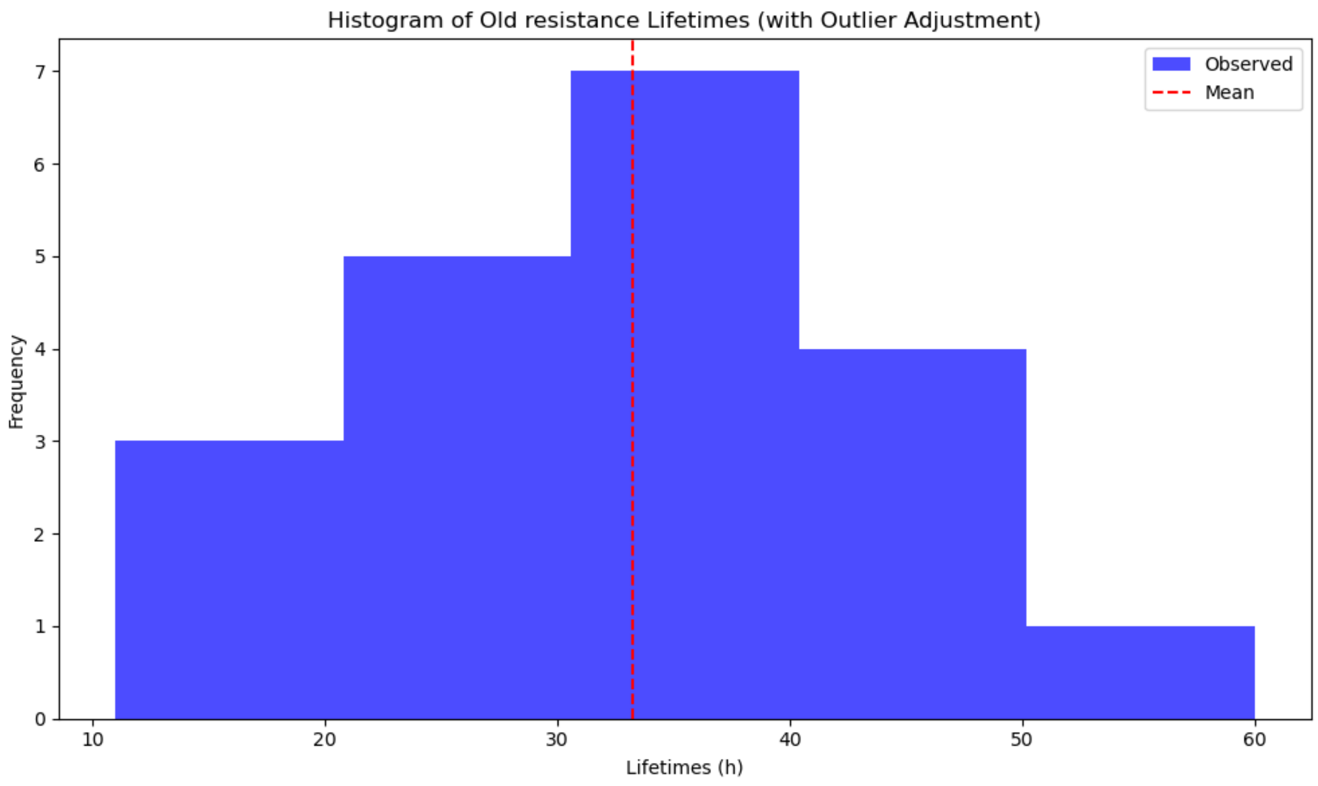
Resistance Lifetime Analysis with Outlier Adjustment (Hours)	
Mean	33.2
Standard_Error	2.667
Median	34
Standard Deviation	11.929
Sample Variance	142.303
Kurtosis	0.170
Skewness	0.154
Range	49
Minimum	11
Maximum	60
25%	26.25
75%	39.75
Sum	665
Count	20



Understand the data | **Histogram** | CDF | Chi-Square & P - value | Decision

Develop the classes	
No.of classes	5.3220
Round No.of classes	5
Class range	9.8

	Length	Frequency
0	20.8	3
1	30.6	5
2	40.4	7
3	50.2	4
4	60.0	1



Understand the data | Histogram | **CDF** | Chi-Square & P - value | Decision

	CDF	Bin Frequency	Expected Value
0	0.148319	0.148319	2.966384
1	0.4121	0.263781	5.275614
2	0.725539	0.313439	6.268782
3	0.922327	0.196788	3.935759
4	0.987533	0.065206	1.304120
5	Sum =	0.987533	20.000000

	χ^2
0	0.000381
1	0.014399
2	0.085293
3	0.001049
4	0.070921
Sum	0.172041

P = 0.91999

	χ^2
0	0.000381
1	0.014399
2	0.085293
3	0.001049
4	0.070921
Sum	0.172041

P = 0.91999

The adjusted old data derives from a normal distribution

- The p-value(0.91999) is larger than the α=5%
- There is no sufficient evidence to reject the null hypothesis

	old_adjusted(h)	new(h)
Mean	33.25	40.5
Variance	142.3026316	147
Observations	20	20
df	19	19
F	0.968045113	
P(F<=f) one-tail	0.472154629	
F Critical one-tail	0.461201089	

P(F<=f)*2	0.944309259> 0.05

No significant difference in variances between the two types of resistance

- The p-value(0.94431) is larger than the α=5%
- There is no sufficient evidence to reject the null hypothesis

	new(h)	old_adjusted(h)
Mean	40.5	33.25
Variance	147	142.302632
Observations	20	20
Pooled Variance	144.651316	
Hypothesized Mean Difference	0	
df	38	
t Stat	1.90623664	
P(T<=t) one-tail	0.03210175	
t Critical one-tail	1.68595446	
P(T<=t) two-tail	0.06420351	
t Critical two-tail	2.02439416	

Reject H_0, the evidence suggests that the lifetimes of new resistance is greater than that of the adjusted old resistance

- The p-value(0.0321O) is less than the α=5%
- There is sufficient evidence to reject the null hypothesis

Old data **New data**

P =0.03861

Adjusted old data **New data**

P = 0.03210

Both with and without outlier replacement, the test results consistently indicate that the new resistances have a significantly longer lifetime than the old ones.

Data-Driven Decision Making

What should the kitchen manufacturer do?

Based on the statistical results at a 95% confidence level, the new resistance demonstrates a significantly longer lifespan compared to the current resistance.

Given these findings, it is recommended that the kitchen manufacturer transition to the new resistance to take advantage of its superior performance.

Switch to the new resistance

- **Potential Performance Improvement:** The new resistance shows a higher average lifetime from the dataset. It could reduce replacement frequency and maintenance costs, providing opportunities for long-term product improvements.

Considerations

- **Request a Larger Dataset:** The sample size of 20 may be insufficient to detect lifetime differences.
- **Field Testing:** Conduct small-scale real-world testing to validate the new resistance's performance and compatibility with existing products.
- **Supplier Evaluation:** Assess the new supplier's reputation, quality assurance processes, and ability to provide consistent delivery.
- **Cost-Benefit Analysis:** Weigh the potential cost savings from reduced maintenance against the upfront costs of testing and implementation.

If decide to keep using the old resistance

- **Proven Performance:** The current resistances have a stable performance history and are already integrated into the production process.
- **Avoidance of Transition Risks:** Keeping the old resistance eliminates the risks associated with supplier changes, such as production delays or quality issues.
- **Cost Stability:** There are no additional costs for testing or adjusting new things.

Considerations

- **Performance Monitoring:** Continuously monitor the performance of the old resistance to ensure it remains satisfactory over time.
- **Reassessment Opportunities:** Periodically reassess the potential benefits of switching to new technology as the market evolves.

Limitation of data analysis

In carrying out our analysis, what difficulties does this data present?

DATA QUALITY ISSUES

- Check the process of data collection
- Dealing with the outlier
- Retest by RD in the trial phase

VOLUME OF DATA

- More sample data leads to more statistical power
- Apply industrial sampling standard
- Conduct the power analysis of two sample size. Ex. Set the effect size to further calculate the statistical power

BIAS

- Require more data and ensure the sample is randomly selected.
- Evaluate the effect of bias on kitchen products when use new material

DOMAIN KNOWLEDGE TO SUPPORT THE DATA ANALYSIS

- Need more domain knowledge to support the data to overall assess the reliability of analysis. Ex. Testing environmental parameter
- The current analysis cannot support the claim that except for lifespan, new and old resistance are exact the same.

Appendix

Boxplot value and figure

```
Q1_old = old_df['old'].quantile(0.25)
Q3_old = old_df['old'].quantile(0.75)
IQR_old = Q3_old - Q1_old
lower_bound_old = Q1_old - (1.5 * IQR_old)
upper_bound_old = Q3_old + (1.5 * IQR_old)

Q1_new = new_df['new'].quantile(0.25)
Q3_new = new_df['new'].quantile(0.75)
IQR_new = Q3_new - Q1_new
lower_bound_new = Q1_new - (1.5 * IQR_new)
upper_bound_new = Q3_new + (1.5 * IQR_new)

sns.set(style="whitegrid", context="notebook", palette="pastel")
plt.figure(figsize=(10, 6))
sns.boxplot(data=old_df, y='old', color='lightblue')
sns.stripplot(data=old_df, y='old', color='darkred', alpha=0.6)
#border line
plt.axhline(lower_bound_old, color='green', linestyle='--', label=f'Lower Bound = {lower_bound_old:.2f}')
plt.axhline(upper_bound_old, color='green', linestyle='--', label=f'Upper Bound = {upper_bound_old:.2f}')
plt.title('Box Plot of Old Resistance Lifetime')
plt.legend()
plt.show()

plt.figure(figsize=(10, 6))
sns.boxplot(data=new_df, y='new', color='lightblue')
sns.stripplot(data=new_df, y='new', color='darkred', alpha=0.6)
plt.axhline(lower_bound_new, color='green', linestyle='--', label=f'Lower Bound = {lower_bound_new:.2f}')
plt.axhline(upper_bound_new, color='green', linestyle='--', label=f'Upper Bound = {upper_bound_new:.2f}')
plt.title('Box Plot of New Resistance Lifetime')
plt.legend()
plt.show()
```

Appendix

Cumulative Distribution Function

```
mean_old = np.mean(old_df)
std_dev_old = np.std(old_df, ddof=1)
n_old = len(old_df)
k_old = int(1 + 3.322 * np.log10(n_old))
classrange_old=np.ptp(old_df)/k_old
hist_old, bin_edges_old = np.histogram(old_df, bins=k_old)
cdf_values_old = stats.norm.cdf(bin_edges_old+classrange_old, mean_old,
std_dev_old)
mean_new = np.mean(new_df)
std_dev_new = np.std(new_df, ddof=1)
n_new = len(new_df)
k_new = int(1 + 3.322 * np.log10(n_new))
classrange_new=np.ptp(new_df)/k_new
print(f"classrange:{classrange_new}")
hist_new, bin_edges_new = np.histogram(new_df, bins=k_new)
cdf_values_new = stats.norm.cdf(bin_edges_new+classrange_new, mean_new,
std_dev_new)
```

Histogram

```
plt.hist(old_data, bins=bin_edges_old, color='blue', alpha=0.7, label='Observed')
# average values
plt.axvline(mean_old, color='red', linestyle='--', label='Mean')
plt.title('Histogram of Old resistance Lifetimes (with Outlier Adjustment)')
plt.xlabel('Lifetimes (h)')
plt.ylabel('Frequency')
plt.legend()
plt.tight_layout()
plt.show()
plt.figure(figsize=(10, 6))
plt.hist(new_data, bins=bin_edges_new, color='blue', alpha=0.7, label='Observed')
plt.axvline(mean_new, color='red', linestyle='--', label='Mean')
plt.title('Histogram of New resistance Lifetimes')
plt.xlabel('Lifetimes (h)')
plt.ylabel('Frequency')
plt.legend()
plt.show()
```

Appendix

Chi-Square & P - value

```
bin_frequencies_old = []
for i in range(0, len(cdf_values_old)-1):
    if i==0:
      frequency = cdf_values_old[0]
    else:
      frequency= cdf_values_old[i] - cdf_values_old[i - 1]
    bin_frequencies_old.append(frequency)
bin_frequencies_old= np.array(bin_frequencies_old)
expected_values_old = bin_frequencies_old * n_old
expected_frequencies_normalized_old = expected_values_old * (hist_old.sum() / expected_values_old.sum())
chi_square_values_old = (hist_old - expected_values_old) ** 2 / expected_values_old
chi_square_df_old = pd.DataFrame({'χ^2': chi_square_values_old})
chi_square_stat_old, p_value_old = stats.chisquare(hist_old, f_exp=expected_frequencies_normalized_old,ddof=2)

bin_frequencies_new = []
for i in range(0, len(cdf_values_new)-1):
    if i==0:
      frequency = cdf_values_new[0]
    else:
      frequency= cdf_values_new[i] - cdf_values_new[i - 1]
    bin_frequencies_new.append(frequency)
bin_frequencies_new= np.array(bin_frequencies_new)
expected_values_new = bin_frequencies_new * n_new
expected_frequencies_normalized_new = expected_values_new * (hist_new.sum() / expected_values_new.sum())
chi_square_values_new = (hist_new - expected_values_new) ** 2 / expected_values_new
chi_square_df_new = pd.DataFrame({'χ^2': chi_square_values_new})
chi_square_stat_new, p_value_new = stats.chisquare(hist_new, f_exp=expected_frequencies_normalized_new,ddof=2)
```

Thank you for your attention!

Group 11