Student: Arthur Grossmann—Le Mauguen H00494101

F79SC: Statistics for Science

Assessment Project 2025 – Edinburgh

Academic Year: 2024-2025

Introduction

This report presents an analysis of mechanical reliability data for construction materials as part of the Statistics for Science (F78SC) assessed coursework. The aim is to investigate the behaviour of different materials (Concrete, Steel, and Wood) under stress conditions relevant to real-world structural applications.

Using a dataset of 150 materials provided in *Materials data template.xlsx*, the tasks explore several statistical concepts including random sampling, descriptive statistics, hypothesis testing (with equal and unequal variance assumptions), one-sided and two-sided t-tests, and simple linear regression analysis. All analyses are carried out using Microsoft Excel, with graphical outputs (boxplots, scatter plots, dotplots, and residual plots) used to support interpretations and conclusions.

The main body of this report includes interpretations of results, comments on the suitability of statistical methods, and the implications of findings. An Appendix at the end of the document outlines the procedures used in Excel, including formulas, commands, and relevant screenshots.

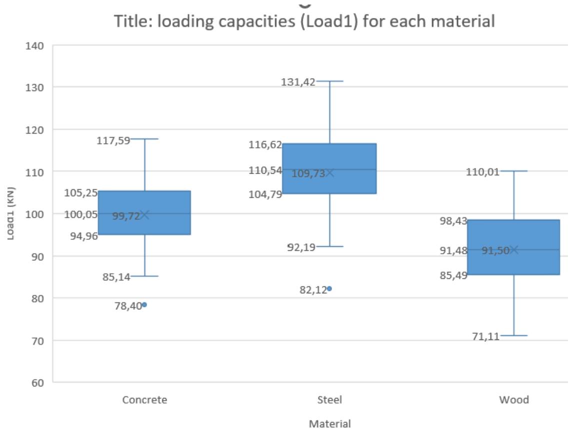
Task 1 Sampling for Reliability Testing

To randomly sample 25 concrete materials and 25 steel materials from a set of 50 each for load testing, the Analysis ToolPak Add-in in Excel was used. First, I ensured my data is organized in separate ranges for concrete and steel. In the Data Analysis dialog box, select Sampling and set the Input Range to the concrete data range. I selected Random as the sampling method and entered 25 for the number of samples. These steps were repeated for the steel data range. I used the INDEX function to create a new table that combines the sampled data from both materials. This procedure ensures that each piece has an equal likelihood of being selected for load testing.

Task 2 Investigation of Loading Capacities before Load Testing

1) Figure 1: Descriptive statistics for the loading capacities (Load1) for each material

Material	Concrete	Steel	Wood
Mean	99,71726579	109,7342	91,50045
Standard Error	1,091275081	1,32077	1,287272
Median	100,0539638	110,5378	91,48435
Mode	#N/A	#N/A	#N/A
Standard Deviation	7,716480101	9,339251	9,102385
Sample Variance	59,54406515	87,22161	82,85341
Kurtosis	0,406864768	0,675983	-0,15892
Skewness	-0,1197159	-0,50515	-0,26662
Range	39,18624273	49,30312	38,89814
Minimum	78,40106912	82,11948	71,10998
Maximum	117,5873118	131,4226	110,0081
Sum	4985,86329	5486,712	4575,022
Count	50	50	50



NB: Here the minimum value was set manually to 60

The descriptive statistics for the loading capacities (Load1) across the three materials reveal several key insights about their performance and variability.

Steel has the highest mean loading capacity at approximately 109.73, followed by Concrete with a mean of 99.72, and Wood with the lowest mean at 91.50. These mean values are closely matched by the median values in each group, suggesting relatively symmetric distributions for all three materials. The small differences between the mean and median also indicate a limited influence of extreme values or skewness.

In terms of variability, Steel shows the highest standard deviation and range, at 9.34 and 49.30 respectively. This means that Steel's loading capacity values are more widely spread compared to the other materials. Concrete and Wood exhibit slightly lower variability, with standard deviations of 7.72 and 9.10, and ranges of 39.19 and 38.90 respectively.

The skewness values are all slightly negative, with Steel at -0.51, Concrete at -0.12, and Wood at -0.27. This indicates that all distributions are mildly left-skewed, with Steel being the most asymmetric. However, none of the skewness values suggest a severe deviation from symmetry. Kurtosis values are also close to zero for each material, meaning the distributions are roughly mesokurtic and similar in peak to a normal distribution.

We can notice is the absence of a mode in the dataset. The mode represents the most frequently occurring value in a dataset, but in this case, no single loading capacity value appears more than once for any of the materials. This suggests a continuous spread of values without repetition, which is common in datasets involving measurements recorded to multiple decimal places.

To better visualize the distribution and compare the loading capacities across materials, boxplots were produced. These show the median, interquartile range, and any potential outliers. We see Steel with the highest median and the widest interquartile range, while Wood is lower on the scale with a narrower spread. We also observe very few outliers in any of the three distributions, reinforcing the impression of generally consistent data.

In conclusion, Steel offers the highest average load capacity but also the greatest variability. Concrete stands as a balanced option with a good average and moderate variability. Wood, while more consistent, has the lowest overall capacity. The boxplots reinforce these observations by illustrating differences in central tendency and spread across the three materials.

2) To test whether there is a statistically significant difference between the mean loading capacities (Load1) of concrete and steel, we conduct a two-sided independent samples t-test under the assumption of equal population variances. The significance level is set at 5%. We use the following hypotheses:

- **Null hypothesis (H_0):** The mean loading capacities of concrete and steel are equal (μ_concrete = μ_steel)
- **Alternative hypothesis (H_1):** The mean loading capacities of concrete and steel are not equal (μ_concrete ≠ μ_steel)

Assumptions for applying the t-test include the **independence of the two samples** (concrete and steel), meaning they are not paired measurements of the same objects. Each group's data appears **approximately normally distributed**, with skewness values (−0.12 for concrete, −0.51 for steel) indicating only mild asymmetry, and kurtosis values (0.41 and 0.68, respectively) suggesting no heavy tails or outliers. Boxplots support this, showing fairly symmetric distributions without extreme values. **Equal population variances** are assumed (as stated in the question), and sample standard deviations (7.72 for concrete vs. 9.34 for steel) are close enough to make this assumption reasonable.

A two-sided t-test assuming equal variances was conducted (Excel t-Test: Two-Sample Assuming Equal Variances). With 50 observations in each group, the degrees of freedom are:df=$n_{concrete}$+n_{steel}−2 =50+50−2=98.

The test statistic is calculated as: $T = \frac{\overline{X_{concrete}} - \overline{X_{steel}}}{S_p\sqrt{\frac{1}{50}+\frac{1}{50}}} \sim t_{98}$

Where ${S_p}^2$, the pooled sample variance, is:

$$ {S_p}^2 = \frac{(n_{concrete} - 1){s_{concrete}}^2 + (n_{steel} - 1){s_{steel}}^2}{(n_{concrete} - 1) + (n_{Steel} - 1)} = \frac{(50 - 1){s_{concrete}}^2 + (50 - 1){s_{steel}}^2}{98} = 73.38 $$

The computed test statistic was T= -5.85, leading to a two-tailed p-value of:

$P(T \leq -5.85 \; or \; T \geq -5.85\,) = 2 \times P(T > -5.85) = 6.6 \times 10^{-8}$

T-Test results:	Concrete	Steel
Mean	99,71727	109,7342
Variance	59,54407	87,22161
Observations	50	50
Pooled Variance	73,38284	
Hypothesized Mean Difference	0	
df	98	
t Stat	-5,84668	
P(T<=t) one-tail	3,3E-08	
t Critical one-tail	1,660551	
P(T<=t) two-tail	6,6E-08	
t Critical two-tail	1,984467	

Interpretation: Since the p-value (6.6e-8) is far below the significance threshold of 0.05 (and the absolute value of the t-statistic (|T| = 5.847) is greater than the critical value (±1.984)), we reject the null hypothesis. There is strong statistical evidence to conclude that the mean loading capacities of concrete and steel are significantly different at the 5% level. This result was expected given the values of the summary statistics and the plots presented in Part 1, which already suggested a noticeable difference in the means and distributions between the two materials.

3) We now perform a second t-test, this time without assuming equal variances. This test follows Welch's t-test approach, while maintaining the same hypotheses and underlying assumptions as in the previous analysis. The test statistic is

$$t \ = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -5.85$$

Where: $\overline{X}_1, \overline{X}_2$ are the sample means (Concrete and Steel); s_1^2, s_2^2 are the sample variances; n_1, n_2 are the sample sizes

The degrees of freedom, calculated using the Welch–Satterthwaite equation, are: $df \ = \frac{\left(\frac{s_1^2}{n^1} + \frac{s_2^2}{n^2}\right)^2}{\frac{\left(\frac{s_1^2}{n^1}\right)^2}{(n^1 - 1)} + \frac{\left(\frac{s_2^2}{n^2}\right)^2}{(n^2 - 1)}} = 95.$

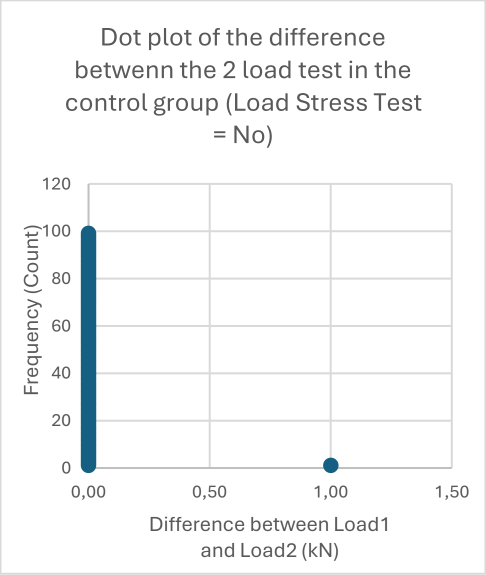
Using this value, the two tailed p-value is: $P(T \leq -5.85 \ or \ T \geq -5.85\ ) = 2 \times P(T > -5.85) = 7.05 \times 10^{-8}$.

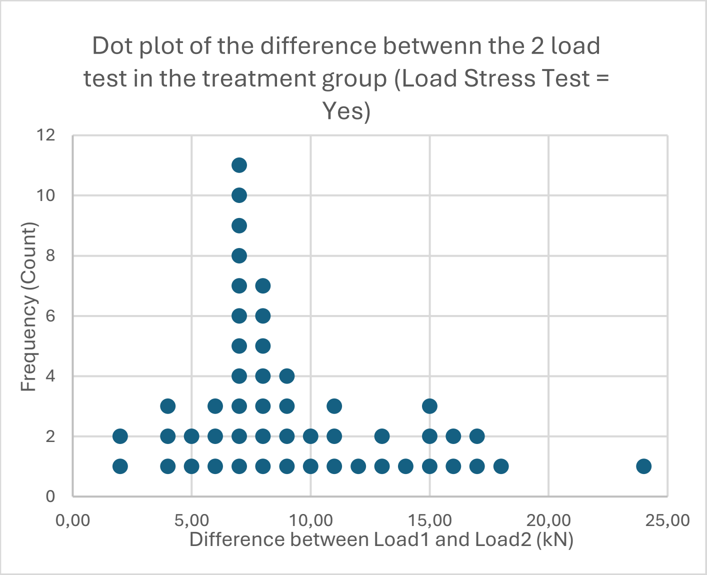
t-Test: Two-Sample Assuming Unequal Variances	*Concrete*	*Steel*
Mean	99,71727	109,7342
Variance	59,54407	87,22161
Observations	50	50
Hypothesized Mean Difference	0	
df	95	
t Stat	-5,84668	
P(T<=t) one-tail	3,53E-08	
t Critical one-tail	1,661052	
P(T<=t) two-tail	7,05E-08	
t Critical two-tail	1,985251	

Both the Toolpak and Excel formula yield nearly identical p-values, with minor differences likely due to internal rounding. In both cases, the p-value is far below the 0.05 significance level, indicating a statistically significant difference in mean loading capacities.

This result confirms the conclusion from the test in Part 2 under the equal variances assumption. Moreover, Welch's t-test, performed without assuming equal variances, also shows a statistically significant difference. Together, these results consistently support the rejection of the null hypothesis and provide strong evidence that the average loading capacities of concrete and steel are significantly different.

Task 3 Investigation of Loading Capacities after Load Testing





1) The dot plots illustrate the differences in loading capacities (Load1 – Load2) for both the control group (Load Stress Test = No) and the treatment group (Load Stress Test = Yes). Out of the 150 materials, 25 concrete and 25 steel samples were tested, these constitute the treatment group. Wood was not tested and is therefore only present in the control group. To improve clarity, the difference values were rounded prior to plotting, which reduces the number of unique values and makes the distribution patterns easier to interpret.

In the control group, where no stress test was applied, we observe very limited variation in the differences between Load1 and Load2. The differences range from 0.00 to 1.00, with the vast majority (99 out of 100) concentrated around zero. This is consistent with expectations, as no test was performed, and thus the loading capacity should remain stable. There is a single small outlier at 1.00, but it does not affect our overall interpretation. The control group also includes all wood materials, since wood samples were not subjected to the stress test, which explains the larger number of observations in this group.

In contrast, the treatment group, which includes only steel and concrete samples, displays a much wider range of differences, indicating a more pronounced effect of the stress test. The values range from 2 kN to 24 kN, with a peak around 7 kN and most differences falling between 4 kN and 11 kN. This distribution shows a clear concentration of changes in that range, suggesting that the stress test generally causes a noticeable reduction in loading capacity. Compared to the control group, the treatment group exhibits greater variability, which highlights the impact of the stress test on the structural properties of the materials tested.

2) It is claimed that performing a stress test reduces the loading capacity of steel by at most 7 kN. To investigate whether there is any evidence against this claim, we perform a one-sided t-test on the difference between Load1 (before the stress test) and Load2 (after the stress test), focusing only on the material steel. The test is conducted at a 5% significance level. We use the following hypotheses:

- **Null Hypothesis (H_0)**: The mean difference between Load1 and Load2 for steel in the treatment group is less than or equal to 7 kN.

- **Alternative Hypothesis (H_1)**: The mean difference between Load1 and Load2 for steel in the treatment group is greater than 7 kN.

The **assumptions for the paired t-test** are that the differences between Load1 and Load2 are **independent**, **approximately normally distributed**, and have roughly **equal variances** (although equal variance is less critical in this context).

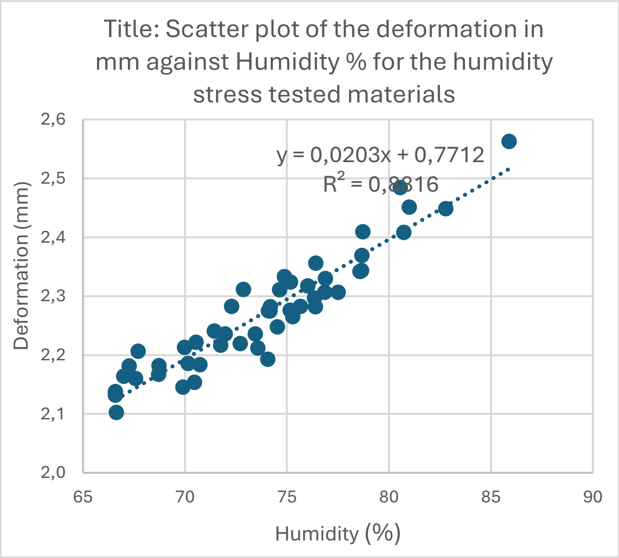
We use a one-tailed test because the hypothesis specifically tests whether the stress test reduces the loading capacity of steel by more than 7 kN, focusing on a reduction in capacity in one direction. A one-tailed test is appropriate when we are only interested in detecting a difference in one direction (in this case, a decrease in loading capacity), as opposed to a two-tailed test, which would look for differences in both directions.

t-Test: Paired Two Sample for Means	*Load1 (kN)*	*Load2 (kN)*
Mean	109,0663244	101,8251
Variance	60,03482844	62,24188
Observations	25	25
Pearson Correlation	0,985910696	
Hypothesized Mean Difference	7	
df	24	
t Stat	0,913791526	
P(T<=t) one-tail	0,184957895	
t Critical one-tail	1,71088208	
P(T<=t) two-tail	0,36991579	
t Critical two-tail	2,063898562	

The p-value for the one-tailed test is 0.185, which is greater than the significance level of 0.05. This indicates that the observed difference in loading capacity before and after the stress test is not statistically significant enough to reject the null hypothesis at the 5% level. The test statistic (t Stat) is 0.914, which is lower than the critical value for a one-tailed test at the 5% significance level (t critical = 1.711).

Based on these results, we fail to reject the null hypothesis. There is not enough evidence to conclude that the stress test reduces the loading capacity of steel by more than 7 kN. In practical terms, this suggests that the observed reduction in load-bearing capacity due to the stress test is consistent with the claim that it does not exceed 7 kN. Therefore, the original claim holds under the current data and analysis.

Task 4 Relationship between Deformation Level and Humidity

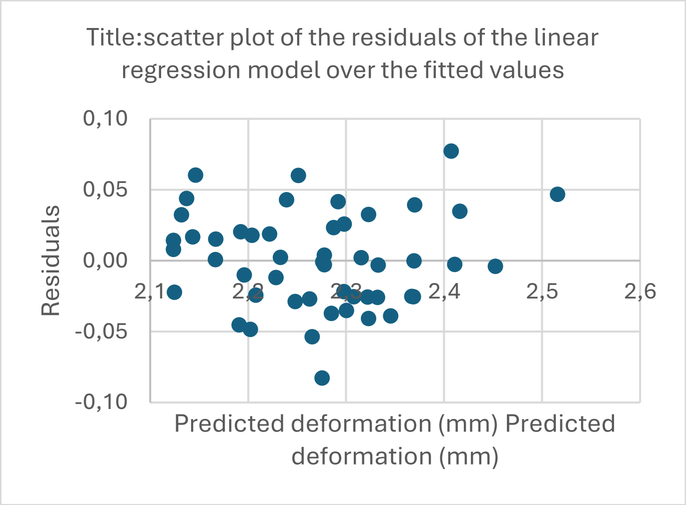


1)The scatter plot shown beside illustrates the relationship between Deformation (in mm) and Humidity (%) for materials subjected to a humidity stress test. Humidity levels in the dataset range from 66% to 86%, while deformation spans from 2.10 mm to 2.56 mm. A clear **positive correlation** emerges between Humidity and Deformation: as humidity increases, deformation tends to rise accordingly. The data points generally follow a **linear trend**, they align along a straight line, supporting the use of a linear model to describe this trend. An R² value of 0.8016 confirms a strong linear association, indicating that around 80.16% of the variation in deformation can be attributed to changes in humidity.

Based on the least squares regression, the linear equation describing this relationship is: **Deformation = α + β × Humidity.** From the scatter plot, we get: **Deformation = 0.0203 × Humidity + 0.7712**

Here, the intercept (α = 0.7712mm) corresponds to the theoretical deformation at 0% humidity, while the slope (β = 0.0203 mm) suggests that each 1% increase in humidity results in an approximate deformation increase of 0.0203 mm.

Overall, both the scatter plot and the regression analysis highlight a strong positive linear relationship between the humidity level and the deformation for materials subjected to this test. The high R² value reinforces the reliability of the model and underscores the significant role humidity plays in influencing material deformation.



2) The residual plot displays the residuals (differences between observed and predicted values) of the linear regression model against the fitted values (predicted deformation).

First, the residuals appear to be randomly scattered around the horizontal axis (residuals = 0), with no obvious pattern or trend. This randomness suggests that the model captures the linear trend in the data appropriately. Moreover, the residuals are well centered around zero and the spread remains relatively constant across the range of fitted values, indicating homoscedasticity (stable variance of the residuals). No strong outliers or influential points are visible in the plot, which further supports the validity of the model. In addition, the scatter plot of humidity versus deformation demonstrates a clear linear relationship between the two variables, supporting the assumption of linearity. The independence assumption also appears reasonable, as each material was tested separately under controlled conditions.

Taken together, these observations indicate that the key assumptions of linear regression (linearity, independence, homoscedasticity, and normality of residuals) are reasonably met. Therefore, we can conclude that the linear regression model provides a good fit for describing the relationship between humidity and deformation for the tested materials.

Appendices

Sources

- Help for Excel document available on Canvas https://canvas.hw.ac.uk/courses/28972/files/4059831?wrap=1
- Changing the language used https://support.microsoft.com/en-us/office/change-the-language-office-uses-in-its-menus-and-proofing-tools-f5c54ff9-a6fa-4348-a43c-760e7ef148f8#:~:text=Within%20any%20Office%20application%2C%20select,then%20select%20Set%20as%20Preferred.
- Excel analysis toolpak https://www.excel-easy.com/data-analysis/analysis-toolpak.html
- Course materials from the second year at ISEN Nantes
- Satterthwaite Formula for Degrees of Freedom https://www.statisticshowto.com/satterthwaite-formula/

<u>How to make sure Analysis ToolPak is enabled:</u>

- Go to File > More … > Options > Add-ins
- At the bottom, in the “Manage” box, select Excel Add-ins and click Go…
- Check the box for Analysis ToolPak and click OK

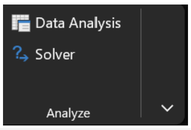
Task 1:

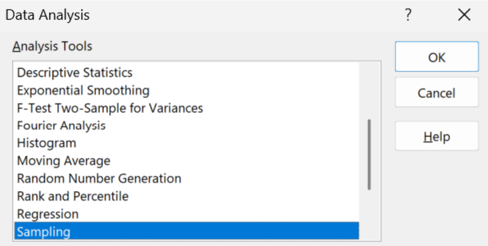
To perform the sampling procedure for reliability testing, we used the Sampling tool from the Analysis ToolPak. The goal was to randomly select 25 concrete and 25 steel materials from their respective groups of 50, ensuring that each material had an equal chance of being selected.

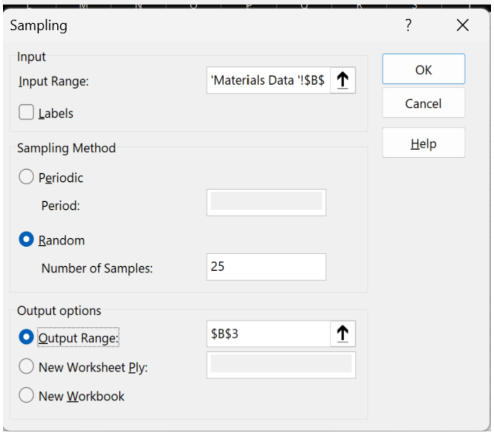
The input range for the sampling tool was set to the UIDs (unique identifiers) of the materials within each group. The number of samples was specified as 25, and the random sampling method was selected to guarantee unbiased selection.

This procedure was conducted separately for concrete and steel, resulting in two independent lists of 25 randomly selected UIDs. To retrieve the full data associated with each selected material, we used the INDEX() function. This formula allowed us to automatically extract the relevant information (e.g., material type, Load1, Load2) from the original database by matching the position of each selected UID.

This method ensures a fair and reproducible sampling process, aligning with the requirement that every individual unit has an equal probability of inclusion in the test.







Input

Input Range: 'Materials Data '!B

□ Labels

Sampling Method

○ Periodic

Period:

● Random

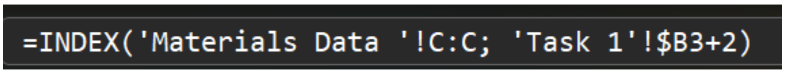
Number of Samples: 25

Output options

● Output Range: $B$3

○ New Worksheet Ply:

○ New Workbook

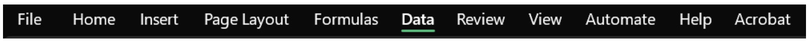


```
=INDEX('Materials Data '!C:C; 'Task 1'!$B3+2)
```

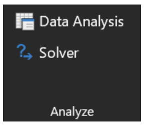
Task 2 – Q1:

How to calculate the descriptive statistics for the loading capacities (Load1) for each material:

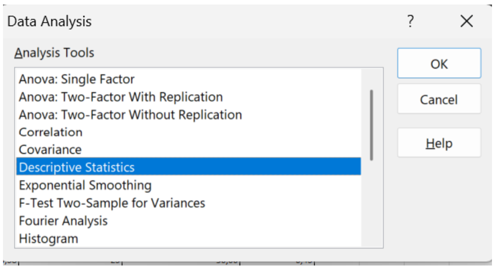
- Click the Data tab in the Excel ribbon.



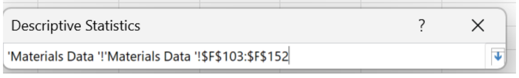
- On the right side, in the Analyze group, click Data Analysis. A dialog box will appear with various analysis tools.



- In the Data Analysis dialog box, select Descriptive Statistics from the list and click OK.

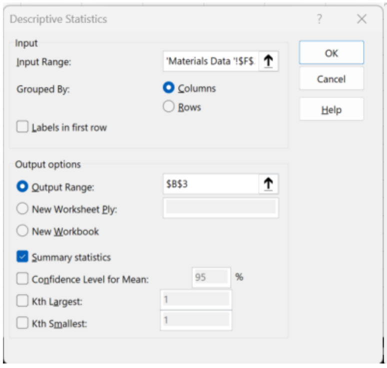


- In the Descriptive Statistics dialog box: Input Range: Enter the range of data you want to analyse, note you can choose to use data from another sheet and event another excel file (check Labels in First Row if your data range includes column headers)


Descriptive Statistics

'Materials Data '!'Materials Data '!$F$103:$F$152

- In Output Options Output Range: Select where you want the results to appear
- Choose the Statistics to Display, summary statistics here
- Click OK
- Do it 3 times and don't forget to change the output range each time:


Descriptive Statistics

Input

Input Range: 'Materials Data '!F

Grouped By: ● Columns ○ Rows

□ Labels in first row

Output options

● Output Range: $B$3

○ New Worksheet Ply:

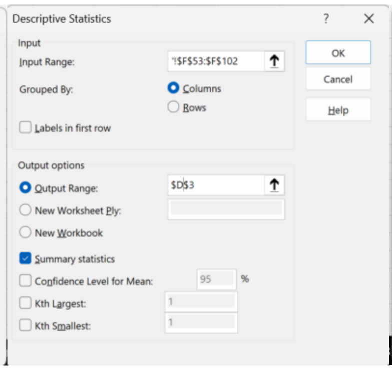
○ New Workbook

☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1


Descriptive Statistics

Input

Input Range: '!$F$53:$F$102

Grouped By: ● Columns ○ Rows

□ Labels in first row

Output options

● Output Range: $D$3

○ New Worksheet Ply:

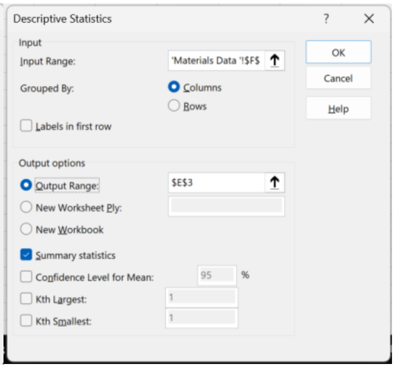
○ New Workbook

☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1


Descriptive Statistics

Input

Input Range: 'Materials Data '!F

Grouped By: ● Columns ○ Rows

□ Labels in first row

Output options

● Output Range: $E$3

○ New Worksheet Ply:

○ New Workbook

☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

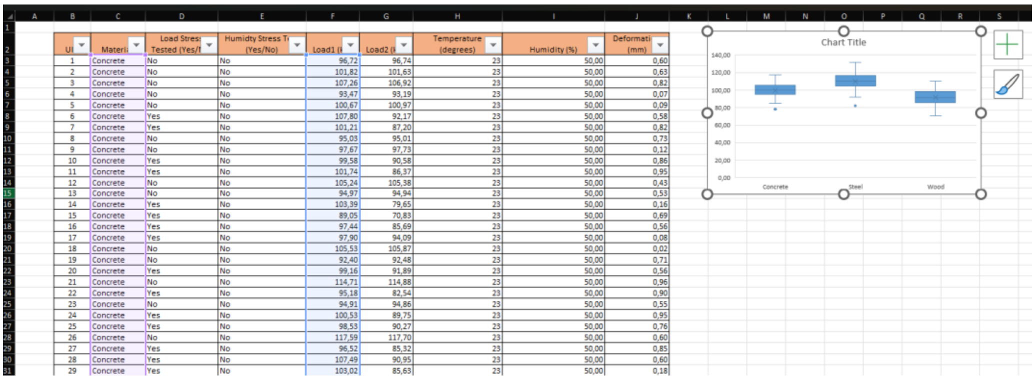
Breakdown of the terms generated by the Data Analysis Toolpak in Excel:

- Mean: The average value of the dataset. It's the sum of all values divided by the number of data points.
- Standard Error: A measure of how much the sample mean is expected to vary from the true population mean. It's calculated as the standard deviation divided by the square root of the sample size. A lower standard error means more reliable data.
- Median: The middle value when the data is ordered from smallest to largest. If there's an even number of observations, it's the average of the two middle values. It's useful for understanding the "center" of the data, especially when the data is skewed.

- Mode: The value that occurs most frequently in the dataset. Some datasets may have no mode or more than one mode (bimodal or multimodal).
- Standard Deviation: A measure of how spread out the values in the dataset are. A larger standard deviation means the values are more spread out from the mean.
- Sample Variance: The square of the standard deviation. It measures the variability of data points in the sample.
- Kurtosis: This describes the "tailedness" of the data distribution. A high kurtosis indicates a distribution with heavy tails or outliers, while low kurtosis indicates a distribution with light tails.
- Skewness: A measure of the asymmetry of the data distribution. Positive skew means the data is stretched more to the right, while negative skew means it's stretched to the left.
- Range: The difference between the maximum and minimum values in the dataset. It gives you an idea of the spread of the data.
- Minimum: The smallest value in the dataset.
- Maximum: The largest value in the dataset.
- Sum: The total sum of all the values in the dataset. (useless here)
- Count: The number of data points in the dataset. (50 for each material here)

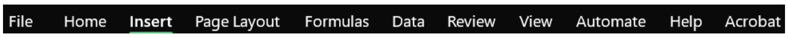
How to produce a boxplot:

- Highlight the data (using Ctrl) like in the following screenshot

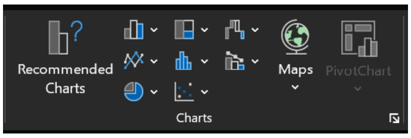


Ul	Materi	Load Stres Tested (Yes/I	Humidty Stress T (Yes/No)	Load1 (I	Load2 (I	Temperature (degrees)	Humidity (%)	Deformati (mm)
1	Concrete	No	No	96,72	96,74	23	50,00	0,60
2	Concrete	No	No	101,82	101,63	23	50,00	0,63
3	Concrete	No	No	107,26	106,92	23	50,00	0,82
4	Concrete	No	No	93,47	93,19	23	50,00	0,07
5	Concrete	No	No	100,67	100,97	23	50,00	0,09
6	Concrete	Yes	No	107,80	92,17	23	50,00	0,58
7	Concrete	Yes	No	101,21	87,20	23	50,00	0,82
8	Concrete	No	No	95,03	95,01	23	50,00	0,73
9	Concrete	No	No	97,67	97,73	23	50,00	0,12
10	Concrete	Yes	No	99,58	90,58	23	50,00	0,86
11	Concrete	Yes	No	101,74	86,37	23	50,00	0,95
12	Concrete	No	No	105,24	105,38	23	50,00	0,43
13	Concrete	No	No	94,97	94,94	23	50,00	0,53
14	Concrete	Yes	No	103,39	79,65	23	50,00	0,16
15	Concrete	Yes	No	89,05	70,83	23	50,00	0,69
16	Concrete	Yes	No	97,44	85,69	23	50,00	0,56
17	Concrete	Yes	No	97,90	94,09	23	50,00	0,08
18	Concrete	No	No	105,53	105,87	23	50,00	0,02
19	Concrete	No	No	92,40	92,48	23	50,00	0,71
20	Concrete	Yes	No	99,16	91,89	23	50,00	0,56
21	Concrete	No	No	114,71	114,88	23	50,00	0,96
22	Concrete	Yes	No	95,18	82,54	23	50,00	0,90
23	Concrete	No	No	94,91	94,86	23	50,00	0,55
24	Concrete	Yes	No	100,53	89,75	23	50,00	0,95
25	Concrete	Yes	No	98,53	90,27	23	50,00	0,76
26	Concrete	No	No	117,59	117,70	23	50,00	0,60
27	Concrete	Yes	No	96,52	85,32	23	50,00	0,85
28	Concrete	Yes	No	107,49	90,95	23	50,00	0,60
29	Concrete	Yes	No	103,02	85,63	23	50,00	0,18

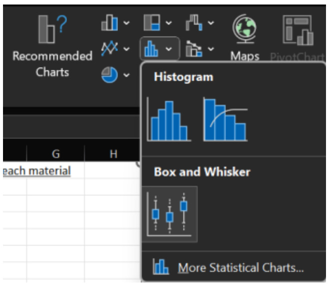
- Click the Insert tab in the Excel ribbon.



- In the Charts group (in the middle), click the Insert Statistic Chart icon



- Choose Box and Whisker.

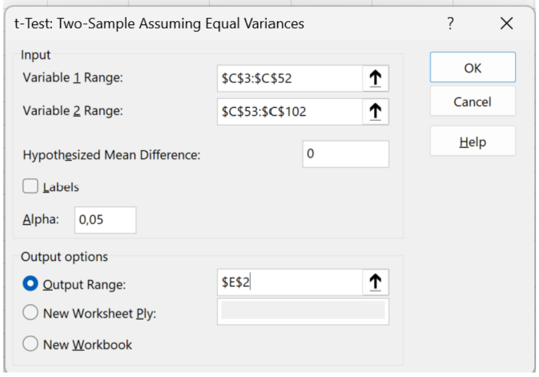


- Chart Title: Click and type to rename.
- Axis Titles: Use the plus sign (+) next to the chart to add axis labels or Chart Design > Add Chart Element > Axis Title
- To zoom in the plot: Click directly on the axis numbers in your boxplot. (here the y axis), it will highlight the axis and show a box around it. Right-click the axis and choose "Format Axis..." from the context menu, a panel will open on the right side of Excel and here set the Minimum (and maximum if you want too) value manually (here a minimum value of 60 was chosen)
- To add the value on the graph go in the Chart Design, then add Chart Element and choose where they will be located in Data Labels.

Task 2 – Q3:

How to conduct a t-Test: Two-Sample Assuming Equal Variances:

- Go to Data > Data Analysis > Choose: t-Test: Two-Sample Assuming Equal Variances
- Then input the data ranges for Load1 values of Concrete and Steel, and set: Hypothesized Mean Difference: 0 and Alpha: 0.05 (for a 5% significance level) like in the following screenshot:



t-Test: Two-Sample Assuming Equal Variances

Input

Variable 1 Range: $C$3:$C$52

Variable 2 Range: $C$53:$C$102

Hypothesized Mean Difference: 0

□ Labels

Alpha: 0,05

Output options

● Output Range: $E$2

○ New Worksheet Ply:

○ New Workbook

Task 2 – Q3:

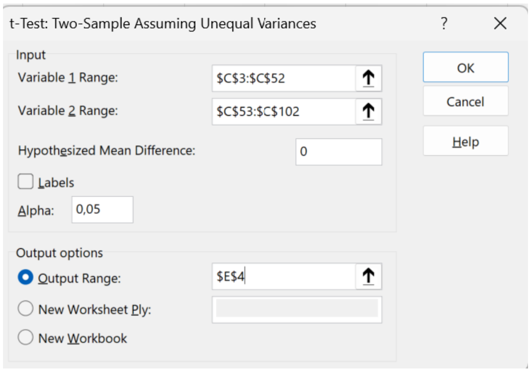
How to conduct a Welch's t-test:

= T.TEST(array1, array2, 2, 3)

Where:

- array1 is the data for Concrete
- array2 is the data for Steel
- 2 means two-tailed
- 3 indicates **unequal variances (Welch's t-test)**

Alternatively, we can also use Data > Data Analysis > t-Test: Two-Sample Assuming Unequal Variances from the Analysis ToolPak like in the screenshot:



t-Test: Two-Sample Assuming Unequal Variances

Input

Variable 1 Range: $C$3:$C$52

Variable 2 Range: $C$53:$C$102

Hypothesized Mean Difference: 0

□ Labels

Alpha: 0,05

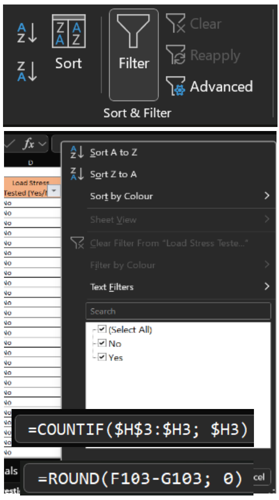
Output options

● Output Range: $E$4

○ New Worksheet Ply:

○ New Workbook

Task 3 – Q1:



```
=COUNTIF($H$3:$H3; $H3)
=ROUND(F103-G103; 0)
```

<u>How to produce a dot plot for discrete data:</u>

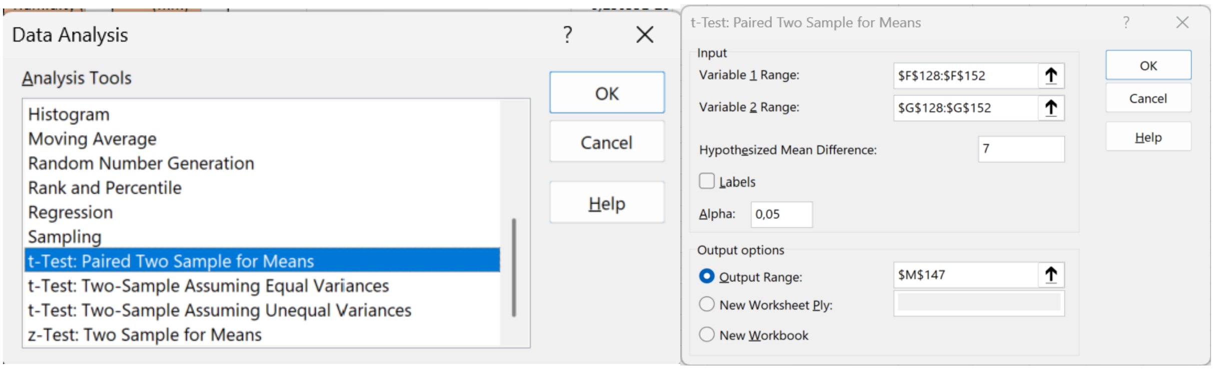
Sort the data to only keep the ones where Load Stress Test = Yes and Material = Steel. Do it using the Filter feature in Excel:

- Click on the filter arrow in the "Load Stress Test" column and uncheck all options except "Yes".
- Then, in the "Material" column, uncheck all options except "Steel".
  This will display only the rows where both conditions are met.

Create a new column and use the COUNTIF function to count how often each unique value appears. Then go to the Insert tab > choose Scatter Chart > select the one without lines (just markers).

Task 3 – Q2:

<u>How to conduct a t-Test: Paired Two Sample for Means:</u>


Data Analysis
?
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help
t-Test: Paired Two Sample for Means
Input
Output options
OK
Cancel
Help

Variable 1 Range: $F$128:$F$152
Variable 2 Range: $G$128:$G$152
Hypothesized Mean Difference: 7
□ Labels
Alpha: 0,05
● Output Range: $M$147
○ New Worksheet Ply:
○ New Workbook

- Go to Data > Data Analysis > Choose: t-Test: Paired Two Sample for Means
- Then input the data ranges for Load1 and Load 2 values and set: Hypothesized Mean Difference: 7 and Alpha: 0.05 (for a 5% significance level).
- Ensure that the order of the variables reflects the assumption: Load1 should be the higher values (before the stress test) and Load2 should be the lower values (after the stress test), so the hypothesized mean difference is positive.

Task 4 – Q1:

<u>How to to obtain a linear equation:</u>

Filter your data to select only the rows where the "Humidity Stress Test" column equals Yes. This can be done by using Excel's filter function: go to the column header for "Humidity Stress Test", click on the dropdown arrow next to the header. Select "Yes" to filter only those rows where the stress test is conducted.

Select the data corresponding to Deformation (Y-axis) and Humidity (X-axis). Highlight the two columns (Deformation and Humidity), go to the Insert tab in Excel, click on the Scatter Plot icon and choose the first option (Scatter with only Markers).

To calculate the linear equation Deformation=α+β×Humidity, we can use Excel's built-in functions for regression. Right-click on one of the data points in the scatter plot, Select Add Trendline, Choose Linear as the trendline type, Check the box for Display Equation on chart and Display R-squared value on chart.

NB: we can also use: Excel's LINEST function to obtain the coefficients α and β (=LINEST(Deformation_range, Humidity_range, TRUE, TRUE))

Task 4 – Q2:

How to to obtain the residuals of the linear regression model:

The residuals are the differences between the observed values of Deformation and the predicted values from the linear regression model.

Calculate the Fitted (Predicted) Values: Using the linear equation Deformation=α+β×Humidity obtained from the regression model, calculate the predicted Deformation values for each observation. (= (Intercept) + (Slope) * (Humidity))

In another new column, calculate the residuals by subtracting the predicted Deformation values from the observed Deformation values. (= (Observed Deformation) - (Predicted Deformation)) Then plot the values.