Performing Statistical Analysis with MATLAB

USING MATLAB FOR STATISTICAL ANALYSIS

Martin Burger
STATS PROGRAMMING TUTOR

MATLAB for Analysis

Focus on summary statistics as part of exploratory data analysis

Managing expectations about the course and MATLAB

- Proprietary software with trial option

Dataset import and exploration

The importance of summary statistics

- Summaries depend on the data class

The MATLAB Environment for Stats and Machine Learning

Main: Statistics and Machine Learning Toolbox™

Functions can be written in base MATLAB

Functions can be purchased in toolboxes

- Proved and tested functions with documentation
- Additional fee to the basic plan

Your company or education institute might offer access to MATLAB and toolboxes

MATLAB Applications

User friendly

Easy to use

Interactive

Graphical interface

Practices Applied in the Course

Functions from MATLAB toolboxes

MATLAB applications

Course Structure

MATLAB for Statistical Analysis

Engineering examples

Methods of varying complexity levels

Implementing MATLAB's toolboxes

Statistical methods from simple to advanced

Data visualizations, tests and modeling

Course Structure

Exploratory analysis with summary statistics and data visualizations

Probability distributions and their identification

Curve fitting and interpolation

Advanced statistical methods: Durability analysis, design of experiment, process control

Generalized Workflow in Statistical Analysis

Importing the Dataset

Download the Course Dataset

Most demos are done on the machines.csv dataset

Measurements from five agriculture machines operated by three teams

Each team works on each machine every day

Measurements:

- Output units
- Electricity consumption
- Input material weight

Summary Statistics in MATLAB

Understanding the statistics of a sample starts with summaries

The output depends on the variable class

- Five or six number summaries for numeric variables
- Counts for categorical variables

Statistics can be combined with data visualizations

- Category counts with bar chart
- Five number summary with box plot
- Frequency distribution with histogram

Summary Statistics for Numeric Variables

```
mymatrix =
    machines(:,{'Weight_Used','Electricity_Consumption'});

mymatrix = table2array(mymatrix);
```

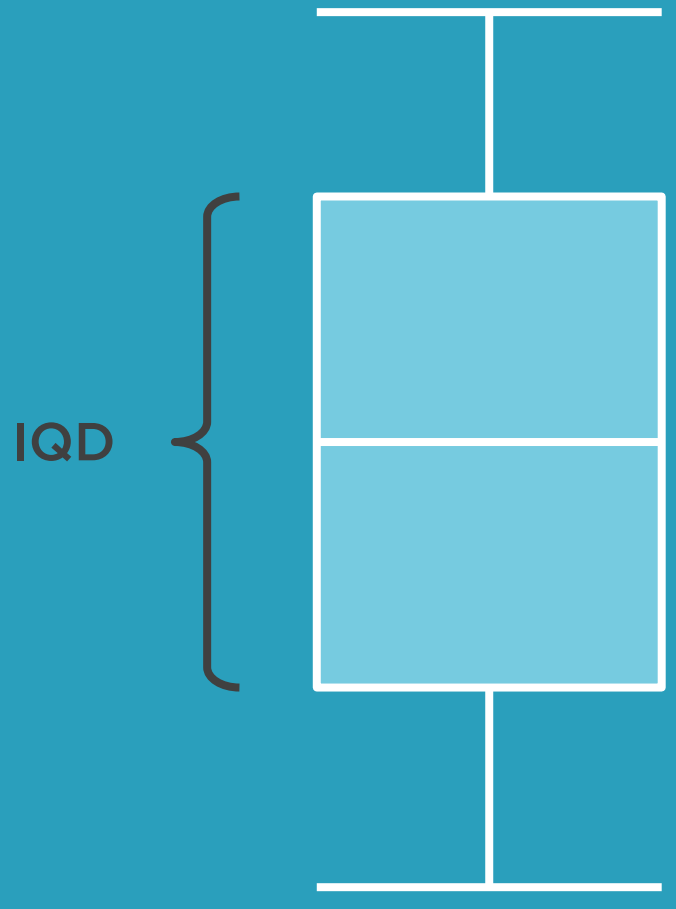
Extracting Columns from a Table

Use the indexing operator to access certain variables

- `table(rows,columns)`
- Access multiple columns by their name: `{'col1', 'col2'}`

Convert the object into a matrix with `table2array()`

Reading the Box Plot



1.5x IQD

3rd quartile

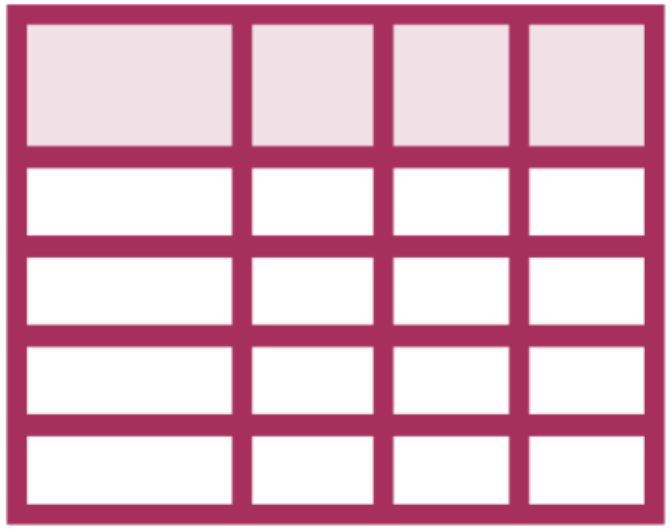
Median

1st quartile

-1.5x IQD

Contingency Tables for Categorical Variables

Count Based Summaries



Counting the number of instances per category

Nomenclature:

- Categorical variable, factor and grouping variable refer to the same class
- Categories, levels and groups refer to the values of that class

Count based summary: Crosstab or contingency table

Contingency Table Example

Counting the number of instances across two grouping variables

Teams

Machines

	S1	S2	S3
H1	**6**	**6**	**6**
H2	**6**	**6**	**6**
H3	**6**	**6**	**6**
H4	**6**	**6**	**6**
H5	**6**	**6**	**6**

Evaluating the Crosstab

All possible combinations among two factor variables show up in equal numbers

- Balanced setup

Statistical tools for grouping variables often assume a balanced setup

- Adjustments might be required

Engineering datasets are often balanced

Grouped Statistics

```
summary_table = grpstats(machines, {'Team', 'Machine'},…
                         {'mean', 'sum'},…
                         'DataVars', {'Units', 'Weight_Used'});
```

Improving the output and incorporating more data {…} and more stats {…}

- Specify the table name
- Add the grouping variable(s)
- Add the statistics
- In case of mixed classes (factors and numeric) the variables for the summary statistics must be declared after ‘DataVars’
- Add numeric variable(s)

Correlations Between Numeric Variables

Correlation

Statistical relationship between two (or more) variables.
The availability of tools to check the presence of a correlation depends on the data class.

Correlation Between Numeric Variables

Correlation coefficient

Scatterplot with correlation line

Correlation Matrix (ρ)

	Weight	Electricity
Weight	**1**	**0.6188**
Electricity	**0.6188**	**1**

Evaluating the Correlation Coefficient

Positive correlation
$(\rho_{X,Y} > 0.5\ )$

No correlation
$(\rho_{X,Y} \approx 0)$

Negative correlation
$(\rho_{X,Y} < -0.5)$

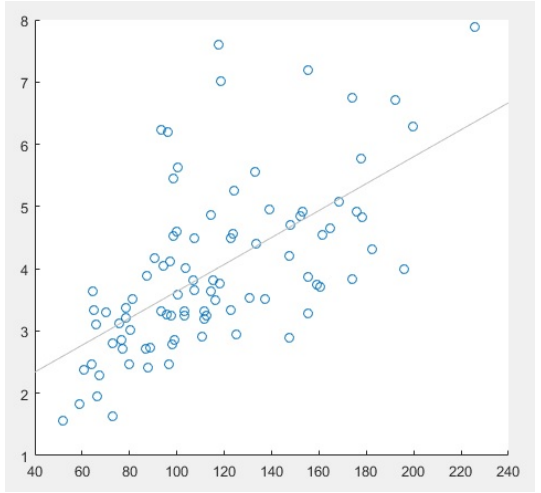
Are Weight and Electricity Correlated?

	Weight	Elect…
Weight	1	0.6188
Elect…	0.6188	1

Correlation coefficient indicates positive correlation

	Weight	Elect…
Weight	1	≈0
Elect…	≈0	1

P-value indicates that the coefficient is significant



Scatterplot with least squares line indicates a positive correlation

Summary: Using MATLAB for Statistical Analysis

Summary

MATLAB as a data science tool

Housekeeping

Data import and exploration

- Data overview with summary statistics
- Structure, variable classes, probability distributions, detecting issues

Visual representations of summary statistics

Up Next: Understanding Statistical Probability Distributions