

# Energy Statistics

Methods for analyzing energy measurement and survey data, energy efficiency and renewable energy technologies 2024

Dr Jonathan Chambers

# Intended outcomes

- Get familiar with statistical concepts
- Learn about statistical methods including machine learning, statistics to apply for energy data
- Assignment: Application of statistical methods to a national energy survey data

# Data

- Input variables: predictors, **independent** variables, **features**, or sometimes just variables, typically denoted as  $X$
- Output variables: **response** or **dependent** variable, typically denoted as  $Y$

# Energy Data – some examples

- electricity consumption of a household
- gas consumption
- temperature
- power
- electricity generation
- ...

# Energy Data Statistics

- Statistical analysis is the basis of most Energy work
- Need to understand relations from data for many purpose
  - Energy related behaviour
  - Performance of energy systems
  - Summarising available data
  - Creating models

	<b>Statistics</b>	<b>Machine learning</b>
Subfield of...	Mathematics, economics	Computer science
Focus on..	Building models with explicitly programmed instructions	Creating systems that learn from data
Purpose	Inferences: relationships between variables	Prediction accuracy, optimisation
Prior assumptions about data	Knowledge about population	none

# Statistics

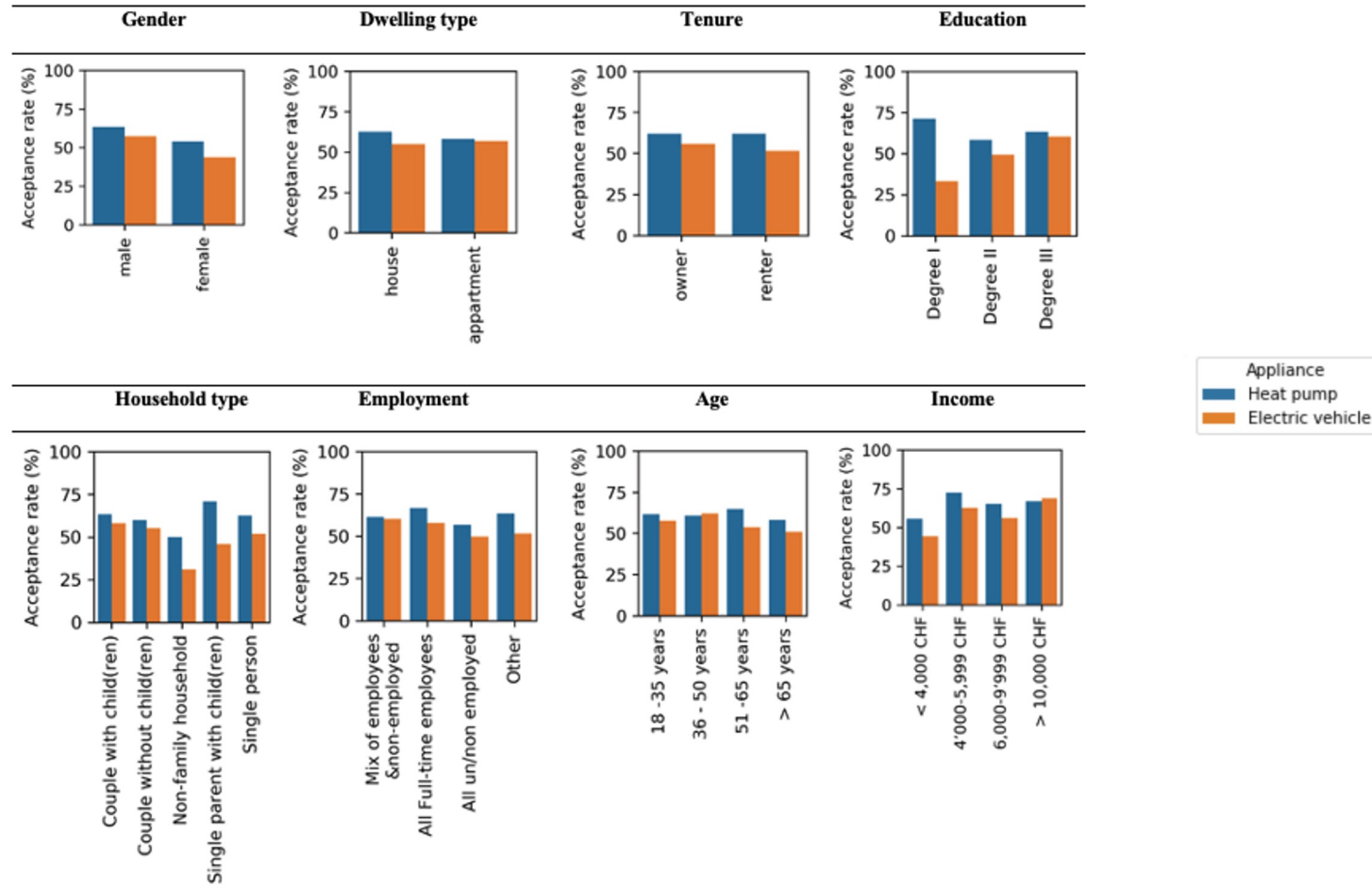
- **Descriptive statistics**
- **Inferential statistics**

# Descriptive Statistics

- A **descriptive statistic**: is a summary statistic that quantitatively describes or summarizes features from a collection of information.
- Aims
  - to summarize and describe a sample
  - to observe the data
  - often an initial analysis for **inferential statistics**



# Descriptive statistics example



# Inferential Statistics

- An **inferential statistic**: the process of used to infer properties of an underlying distribution of probability.
- Inferential statistical analysis infers properties of a population, for example by **testing hypotheses** and deriving estimates.
- Aims
  - to discover some property or general pattern about a large group by studying a smaller group of people.
  - to determine whether the findings from the sample can generalize - or be applied - to the entire population.

# Descriptive vs. Inferential statistics

- With descriptive statistics:
  - You only describe your data.
  - If you observe differences, YOU CANNOT SAY these differences is statistically significant → You need to perform tests for it.
- Perform Statistical tests to find relationship between variables,
  - Do they depend on each other?
  - Is there a significant difference between two groups?
  - Does one impact the other one?

**You find these relationships with inferential statistics!**

# Descriptive vs. Inferential statistics: examples

## Descriptive statistics:

- The mean frequency of the usage of washing machine per week of retired people is 1.2 times.
  - The mean frequency of the usage of washing machine per week of family with children is 2.3 times.
- For descriptive statistics, no statistical test is required.

## Inferential statistics:

- In Switzerland, families with children are significantly 1.9 times more likely to use washing machines than retired people.
- → To assert this (thereby claiming statistical significance), you need a statistical test!

# Concept: “Population”

- Statistics jargon often uses the word “Population”
  - because the statistics were first used to analyse actual country populations and demographics
- Means “whatever is the origin this data”, e.g. energy measurements
- Distinguish between the “sample” and “population”
  - Population: all the possible values
  - Sample: we select some values from this population
- In statistics are generally concerned with trying to understand the “real” properties or trends of the “population” from studying a “sample”

# Concept: “Null Hypothesis”

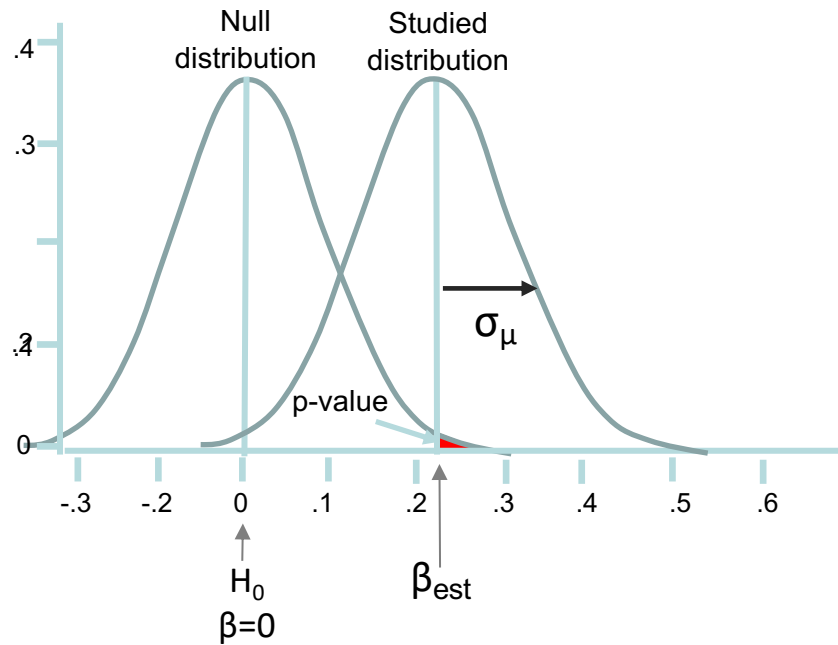
- Define a Null hypothesis as the hypothesis that there is no significant difference between specified populations, any observed difference being due to sampling or experimental error.
- For each statistical test this is given a mathematical definition
- When performing these tests we accept or reject the null hypothesis
- Rejecting a null hypothesis shows a statistically significant difference between observations **but does not automatically prove whatever theory we are testing.**

# Concept: “P-value”

- The probability that a particular statistical measure, such as the mean or standard deviation, of an assumed probability distribution will be greater than or equal to (or less than or equal to in some instances) observed results.
- Think of it as the amount of “overlap” of distributions of data.

# Concept: "P-value" - probability

Small p-value 😊



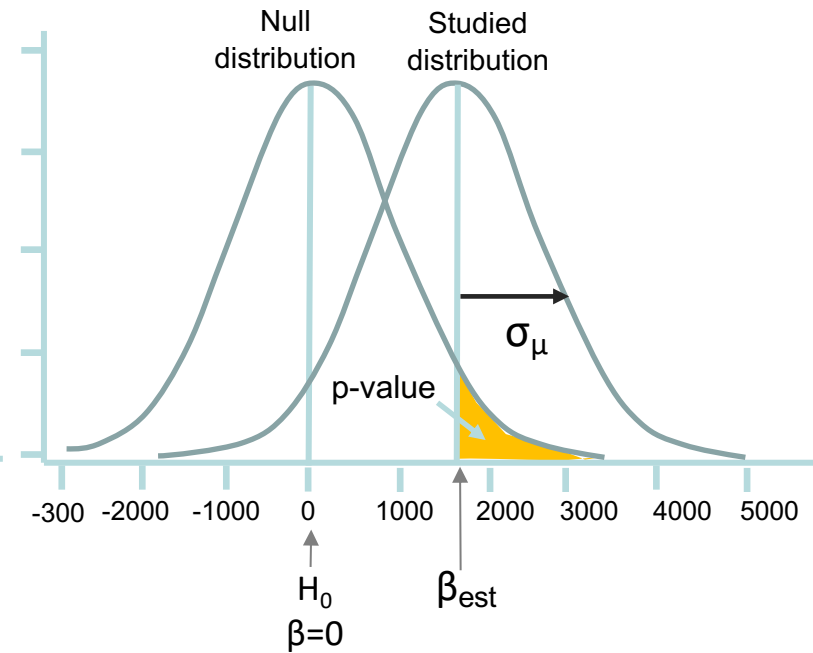
$$\beta_{est} = .22$$

is **UNLIKELY** to be observed if  $\beta = 0$

(low probability 😊)

Dependent variable is significant

High p-value 😞



$$\beta_{est} = 1600$$

is **LIKELY** to be observed if  $\beta = 0$

(high probability 😞)

dependent variable is NOT significant



# Choice of p-value

- The p-value is used to determine whether or not to reject the null hypothesis
- Choose significance level e.g. 5% (1% is also common)
  - p-value  $\leq 0.05$  -> low probability that null hypothesis is valid
  - p-value  $\geq 0.05$  -> insufficiently low probability to reject null hypothesis

# Inferential statistics methods

Chi-square

t-test

ANOVA  
(Analyse of variance)

Regression analysis

# Selecting a method using Aims, Research Questions, and Variables

Test	Chi-square	t-test	one-way ANOVA	Regression analysis
Aim	Find a relationship	Compare two groups	Compare three or more groups	Explain the dependent variable based on many categorical or continuous independent variables and one dependent continuous variable
Dependent variable	Frequency data, categorical variable e.g. ownership	Continuous data e.g. electricity consumption	Continuous data e.g. electricity consumption	Continuous data e.g. electricity consumption
Independent variable	categorical variable e.g. secondary, high school, university, male	binary categorical variable e.g. female, male	multiple categorical variable e.g. children, young, middle-age, elderly	multiple continuous or categorical variable e.g. age, sex, dwelling type, appliance ownership, floor area of house
Example Research Questions	Is there an association between gender and possession of EV for example?  Is there an association between education and possession of EV for example?	Do people living in single-family houses consume more electricity than people who live in apartments?  Do women consume more electricity than men?	Do young people significantly consume more electricity than elderly people and middle aged people?	Which dwelling characteristics or socio-economic variables have a significant impact on the consumption of electricity / thermal energy?

# Chi-square test

- Chi-square test is used when we perform **hypothesis testing on two categorical variables** from a single population.
- **Null hypothesis:** No relationship exists on the categorical variables in the population; they are independent.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where:

$\chi^2$  = Chi Square obtained

$\sum$  = the sum of

$O$  = observed score

$E$  = expected score

- **Small Chi-square:** observed and expected match well  
*(large p-value ☹)*
- **Large Chi-square:** observed and expected do not match, the null hypothesis is rejected  
*(small p-value 😊)*

# Chi-square test – example 1

RQ: Is there an association between gender / education and possession of EV?

- Ownership of an EV (electric vehicle) & Are you planning to buy it?
- Answer: No, Maybe, Yes.
- Gender: Female or Male
- Education: university, college

Null hypothesis: There is no relationship between gender and intention to buy an EV.

	no	maybe	yes
female	35% <sup>a</sup>	29% <sup>a,b</sup>	22% <sup>b</sup>
male	65% <sup>a</sup>	70% <sup>a,b</sup>	77% <sup>b</sup>

← Number of categories = n  
Degrees of freedom = df = n - 1

- We reject the null hypothesis and the relation between buying EVs and gender was significant,  $X^2(N=622) = 7.706$ , **p= 0.02**
- Men are more likely to buy the DLC of EVs whereas women are more likely to reject it.

# Chi-square test – example 2

- Education: university, high-school, college

	no	maybe	yes
university	35% <sup>a</sup>	34% <sup>a</sup>	32% <sup>a</sup>
high school	20%	16%	23%
college	45% <sup>a</sup>	40% <sup>a</sup>	45% <sup>a</sup>

- We accept (**do not reject**) the null hypothesis and the relation between buying EVs and education degree was NOT significant,  $X^2(N=622) = 2.706$ ,  $p=0.85$ .  
→ There is no relationship between education and the intention to buy an EV.

# t-test

- The t-test is an inferential statistic that is used to determine the difference or to compare the means of two groups of samples which may be related to certain features.
- It is performed on continuous variables.
- **Null hypothesis:** The difference between these group means is zero.

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \text{where} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where

$\mu$  = Proposed constant for the population mean

$\bar{x}$  = Sample mean

$n$  = Sample size (i.e., number of observations)

$s$  = Sample standard deviation

$s_{\bar{x}}$  = Estimated standard error of the mean ( $s/\text{sqrt}(n)$ )

Higher values of the t-value, also called t-score, **indicate that a large difference exists between the two sample sets.**

A large t-score indicates that the groups are different.

A small t-score indicates that the groups are similar.

# t-test - examples

**RQ1: Are people living in houses likely to consume more electricity than people who live in apartments?**

Null hypothesis: The difference between the mean of consumption of families and mean of the sample is zero.

**RQ2: Do women consume more electricity than men?**

Null hypothesis: The difference between the mean of consumption of women and mean is zero

Example results:

Women consume more electricity ( $M = 2,200\text{kWh}$ ,  $SD = 34.5$ ) compared to men ( $M = 1,200\text{kWh}$ ,  $SD = 31$ ),  $t = 4$ ,  $p = 0.04$

OR

There was no significant difference between men and women when consuming electricity,  $t(2) = 1.7$ ,  $p = 0.12$ , despite women ( $M = 2,300\text{kWh}$ ,  $SD = 34.5$ ) consumes higher than men ( $M = 2,200\text{kWh}$ ,  $SD = 34.5$ ).



# One-way ANOVA (Analysis of variance)

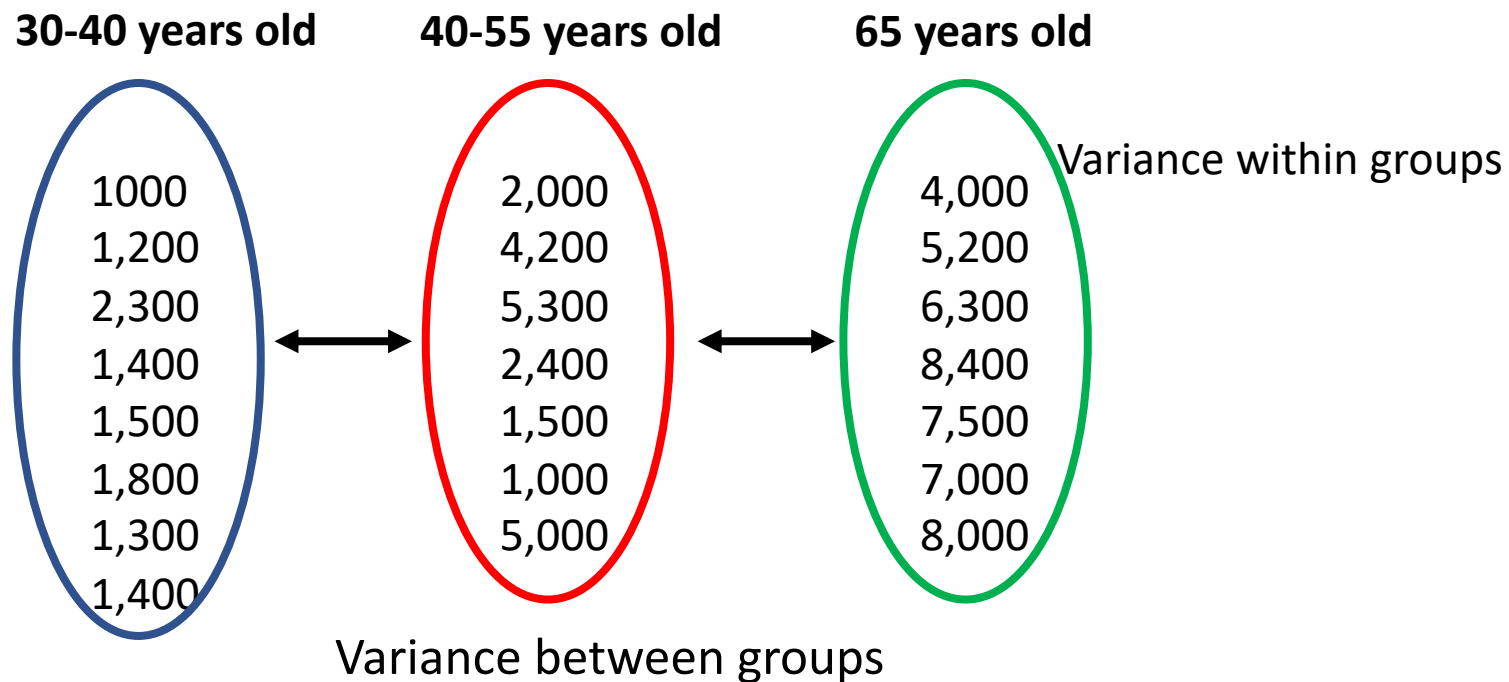
- It is also an analysis of variance and **is used to compare multiple (three or more) samples with a single test**. It is used when the categorical feature (independent variable) has more than two categories.
- The hypothesis being tested in ANOVA is (exactly same with t-test)  
**Null:** All pairs of samples are same, i.e. there is no difference in means of categorised groups.

Example research question:

Do the “40-55 years old” people consume significantly more electricity compared to the “65+ years old” and/or to the “30-40 years old”?

# ANOVA (Analysis of variance)

- The result of the ANOVA formula, the F statistic (i.e., F-ratio, F-value), allows for the analysis of multiple groups of data to determine the variability between samples and within sample.
- F-statistic = variation between sample means / variation within the samples



# ANOVA

- The bigger the  $F$ -value is, the bigger the differences in means between samples.
- However, we must calculate the  $p$ -value to see whether this difference is significant or not!
- We also don't know which group is different from the others.
- Whenever the  $F$ -value was significant (i.e.  $p$ -value is less than 0.05), indicating that there was at least one group differing from one of the others, **Tukey's HSD post-hoc tests** are further performed to identify the statistical differences in each pair of group comparison.

# Tukey's HSD test

This is a t-test!

You do multiple comparisons between groups.

- “40-55 years old” compared to “65+ years old”
- “40-55 years old” compared to “30-40 years old”
- “30-40 years old” compared to “65+ years old”


Multiple Comparisons						
Dependent Variable: elec						
Tukey HSD						
(I) seco3	(J) seco3	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	1183.16287*	206.13783	.000	620.2106	1746.1152
	3	1424.21820*	480.50808	.026	111.9742	2736.4622
	4	2917.99025*	216.00329	.000	2328.0959	3507.8846
	5	1507.77375*	341.99403	.000	573.8049	2441.7426

# ANOVA - example

- Do the “40-55 years old” people consume significantly more electricity compared to the “65+ years old” and/or to the “30-40 years old”?
- The “40-55 years old” consume significantly more electricity ( $M = 3,500\text{kWh}$ ,  $SD = 34.5$ ) compared to the “65+ years old” ( $M = 1,200\text{kWh}$ ,  $SD = 31$ ), but the difference was not significant with “30-40 years old” ( $M = 3,200\text{kWh}$ ,  $SD = 31$ ).

# Regression analysis

Is used to describe relationship between *dependent* variable  $y$  and *independent* variable(s)  $x_1, x_2, \dots$

- Generally  $y = f(x_1, x_2, \dots)$
- Linear regression:  $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + u$   
  
Error term
- Nonlinear regression,  $y = f(x_1, x_2, \dots) \dots$ 
  - e.g.:  $y = \alpha + \beta_1 x_1 + \beta_2 x_1 x_2 + u$

# Regression analysis

**Example Question:** Which type of dwelling characteristics (area, thermal insulation or year of construction) or socio-economic variable (presence of children) has a significant impact on the consumption of electricity / thermal energy?

## **Ordinary Least Square (OLS):**

- Dependent value: continuous
- Independent value: can be anything (continuous, categorical)
- Explain the impact of factors (age, sex, income) on energy consumption.

## **Multinomial logit (same formula, just 0 or 1 for input data)**

- Dependent value: categorical
- Independent value: can be anything (continuous, categorical)
- Explain the impact of factors (age, sex, income) on ownership of electric cars.

# Regression analysis - example

- Ordinary Least Square (OLS) multiple linear regression analysis is conducted to estimate the effect of all the independent variables on the electricity consumption (in kWh).
- The linear regression explaining household expenditure for electricity, heating, and fuels for private mobility can be given by:

$$y_{i,j} = \alpha + x_i' \beta_j + z_{i,j}' \gamma_j + \varepsilon_{i,j}$$

where,

$(y_{i,j})$  are the expenditure of household  $i$  in the energy domain  $j$  (electricity, heating) or the ownership of appliances.

$x_i$  is a set of respondent and household characteristics assumed to be relevant for all types of energy demand and device ownership,

$z_{i,j}$  represents a set of domain-specific controls.



# Regression analysis – Basics I

Model Summary<sup>b</sup>

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	Change Statistics			
						F Change	df1	df2	Sig. F Change
1	.492 <sup>a</sup>	.242	.240	2933.27310	.242	83.581	6	1567	.000

a. Predictors: (Constant), single\_person, age, single\_parent\_child, accom5, couple\_w\_children, couple\_wo\_children

b. Dependent Variable: elec

Variables	Beta	SE	t	Sig.
Household income	0.103	0.039	2.637	0.003
Under 30	ref.	ref.	ref.	ref.
30 to 40 years old	-0.390	0.071	-5.481	0.000
55 to 65	-0.307	0.078	-3.919	0.000
Over 65	-0.226	0.067	-3.385	0.001
Sex of the head of household (1 if female)	0.001	0.002	0.754	0.662
Floor area in m2 (log)	-0.067	0.057	-1.186	0.236
Age of the dwelling	0.007	0.014	0.468	0.640
Constant	4.556	2.2083	1.456	1.779

# Regression analysis – Basics II

Model Summary<sup>b</sup>

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	Change Statistics			
						F Change	df1	df2	Sig. F Change
1	.492 <sup>a</sup>	.242	.240	2933.27310	.242	83.581	6	1567	.000

a. Predictors: (Constant), single\_person, age, single\_parent\_child, accom5, couple\_w\_children, couple\_wo\_children

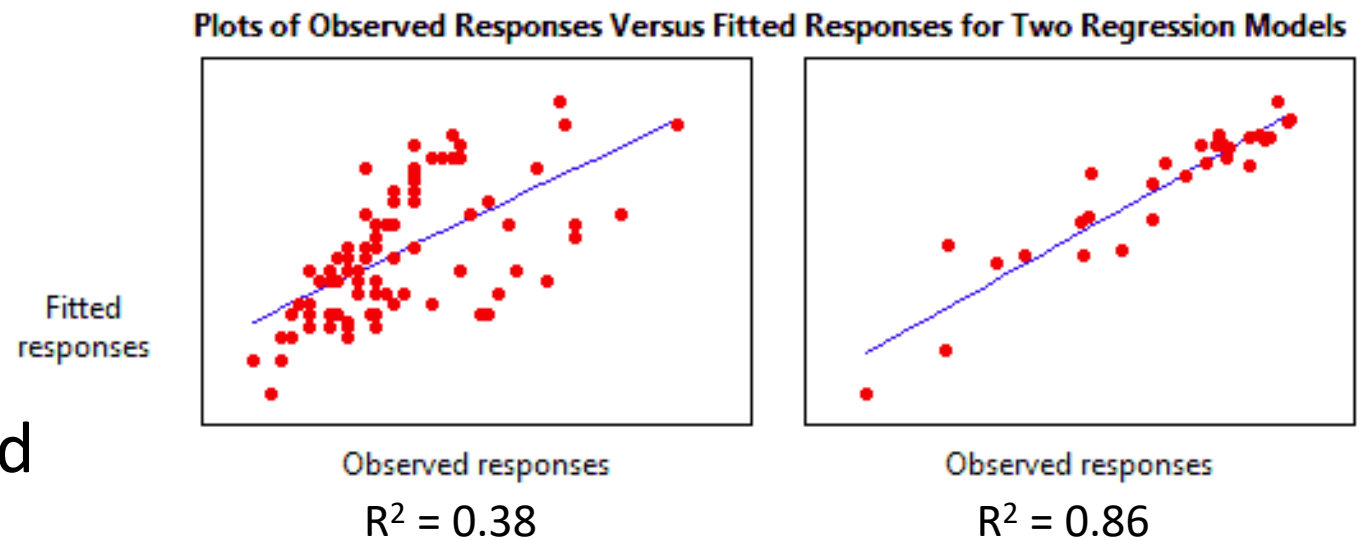
b. Dependent Variable: elec

R-square: provides an estimate of the strength of the relationship between your model and the response variable. It is not a formal hypothesis test for this relationship. i.e. how well the dependent variables is explained.

The F-test of overall significance determines whether this relationship is statistically significant. i.e. model is significant or not

# What is $R^2$ ?

- Coefficient of determination
- Is a measure of the amount of variation in the dependent variable which is explained by the independent variables
- Measures how well the dependent variable is explained



# F-test (not the same as the other one)

- Applies to **all** regression coefficients
- The null hypothesis states that the model with no independent variables fits the data as well as the model we proposed.

# Look at individual factors

Variables	Beta	SE	t	Sig.
Household income	0.103	0.039	2.637	0.003
Under 30	ref.	ref.	ref.	ref.
30 to 40 years old	-0.390	0.071	-5.481	0.000
55 to 65	-0.307	0.078	-3.919	0.000
Over 65	-0.226	0.067	-3.385	0.001
Sex of the head of household (1 if female)	0.001	0.002	0.754	0.662
Floor area in m2 (log)	-0.067	0.057	-1.186	0.236
Age of the dwelling	0.007	0.014	0.468	0.640
Constant	4.556	2.2083	1.456	1.779

- Beta: coefficient → describes the mathematical relationship between each independent and the dependent variable.
- SE = Standard error → standard deviation of the coefficient.
- t = t-value (coefficient divided by standard error)
- Significance (p-value): The p-values for the coefficients indicate whether these relationships are statistically significant.

# Regression analysis - coefficients, what do they mean, how to read them?

Table 9 Example of results of a linear regression

Variables	Gas heating	Electric heating	Electricity consumption	Ownership of dishwasher	Ownership of washing machines
Household income	0.0235	0.219	0.1975	0.220	0.120
Under 30	ref.	ref.	ref.	ref.	ref.
30 to 40 years old	0.0345**	0.23	0.2085	0.111	0.013
55 to 65	0.1862**	0.0271	0.2235	0.120	0.09
Over 65	0.1565**	0.4235	-0.0078	0.450	0.250
Sex of the head of household (1 if female)	0.0112	0.0012	0.0676	0.80	0.253
Retirement	ref.	ref.	ref.	ref.	ref.
Working household	0.0045**	0.2**	0.1785	0.2845	0.48
Household work mix	0.0155	0.211	0.1895	0.2955	0.491
Presence of children (1 if there are children)	0.4672***	0.2781***	0.2045***	0.4472***	0.2881**
Geography (1 if city)	0.1375	0.4045	-0.0268	0.4175	0.6845
Owner (1 if owner)	-0.0078	-0.0178	0.0486	0.2722	0.2622
Dwelling type (1 if house)	0.219***	0.329***	1.419***	0.261	0.261
Floor area in m2 (log)	0.845***	0.267***	0.1785*	0.2845	0.48
Age of the dwelling	0.155**	0.211**	0.0895	0.2955	0.491
Constant	4.556	2.2083	1.456	1.779	0.992

\*\* = significance  $\leq 0.05$ , \*\*\* = significance  $\leq 0.01$ .

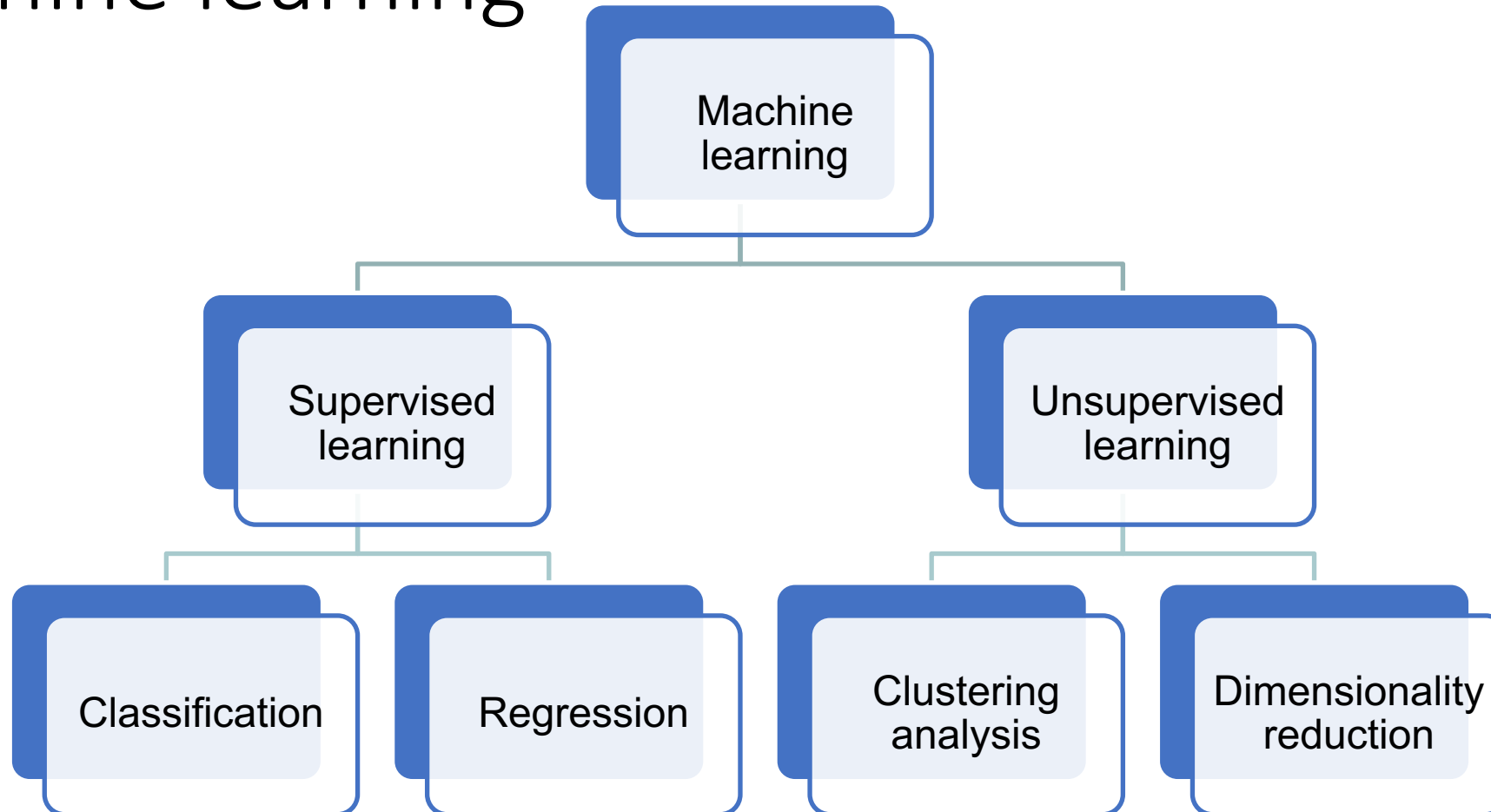
The table values represent Beta.

The stars tell you the p-value significance

# Regression analysis - example

- First, a preliminary analysis of the **multicollinearity** of the variables should be performed to assess whether there is a strong correlation between two or more variables in the regression model based on the test of **variance-inflation factors (VIF)**. The predictor VIFs shown in the regression table must all be less than 5.
- Second, F-test value should be checked, looking at the overall significance of the model.
- Third, the **R<sup>2</sup> value** must be calculated by the model. It must be greater than 0.25 for it to be a good fit model. If it's over 0.40, that's a great fit for behavioural or social analysis.
- Overall, the model should be presented as such with a Table:
  - $F(n_f, N) = 37.2, p = 0.001, R^2 = 0.295$ .
  - $n_f$  = the degree of freedom is the number of random variables that cannot be determined or fixed by an equation
  - $N$  = number of observations (for example, number of people who answered the questionnaire)

# Machine learning





# Supervised learning

- A set of methods that allow **predictions** based on **behaviours or characteristics analysed in historical data**.
- **Prediction** aspect makes it difference from **descriptive** and **inferential** statistics
- We try to make a statistics model in order to predict new values from new input data
- Example methods: Regression models, decision trees.

# Is regression a machine learning method?

- Use regression model equation to **predict** energy consumption as a form of machine learning

## Example

- Regression formula based on data:
  - Consumption =  $324 + 0.0235 * \text{income} + 0.239 * \text{dwelling type} + 0.0345 * \text{age (30-45 year)} + 0.1862 * \text{age (55-65 years)} + \dots$
- Predict the consumption of a home with an income of 1000 CHF, house (0), age 30-45 and floor space of 80m<sup>2</sup>
  - Consumption =  $324 + 0.0235 * 1000 + 0.239 * 0 + 0.0345 * 1 + 0.1862 * 0 + \dots +$
- **Consumption = 6,500 kWh**

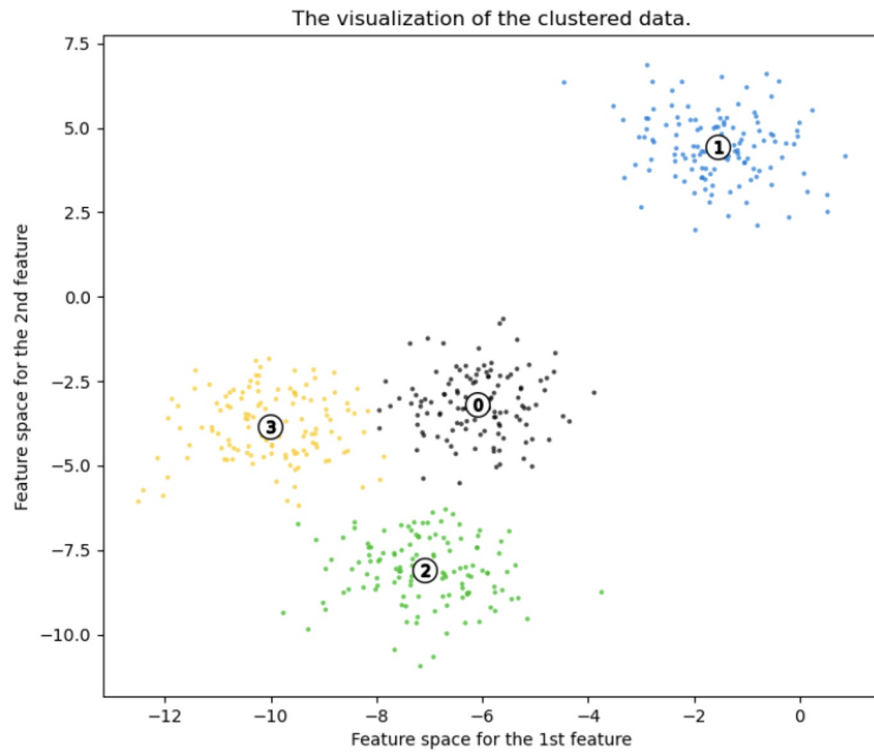
# Unsupervised learning

- A set of techniques that allow you to apply models to extract knowledge of data sets where a priori is unknown.
- Example methods: Cluster analysis, factorial analysis

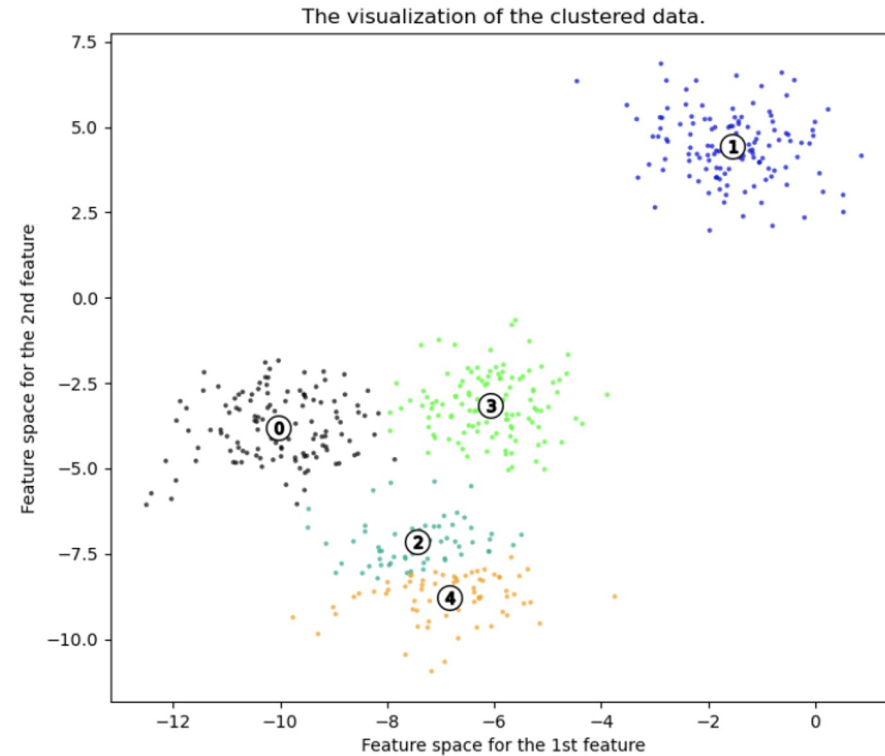
# Cluster analysis

- Primary purpose is to group features (e.g., respondents, measured values, other entities) based on the characteristics they possess.
- Minimize the heterogeneity of objects within the clusters while also maximizing the heterogeneity between clusters.
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and implementation.

# Cluster analysis



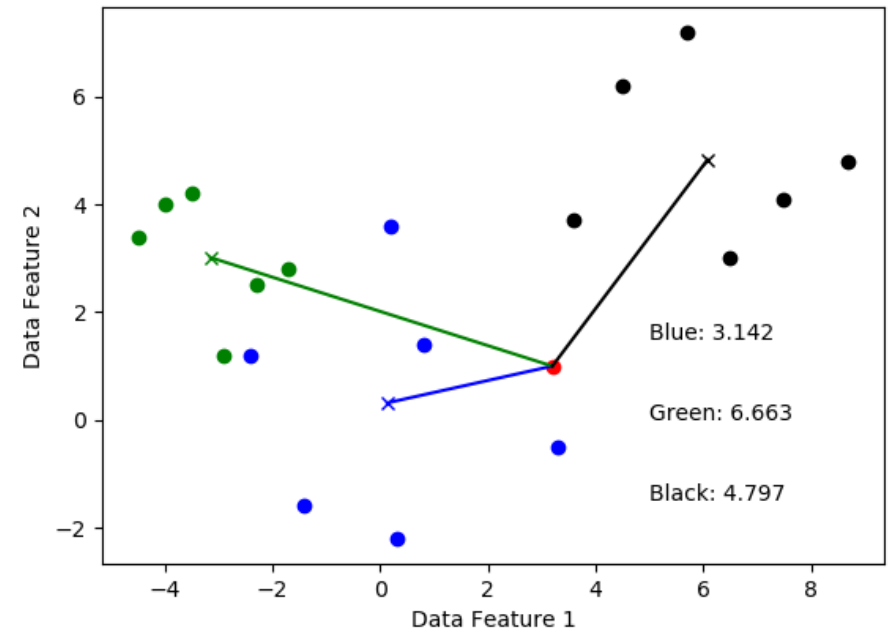
good cluster



bad cluster

# Example

	Believe in Climate change	Technophile	employed	Engage in electricity tariffs
Household 1	0	0	1	1
Household 2	0	1	0	1
Household 3	1	0	0	0
Household 4	0	1	1	1
Household 5	0	1	0	0
Household 6	1	0	1	1



# K-means

- Most used approach due to its versatility and applicability to large datasets.
- Robust and validated implementations are also readily available.
- Several indexes have been developed to optimise the chosen number of clusters “k”.

# Summary: Descriptive vs. Inferential

## Descriptive statistics

- Concerned with the describing the target population (which is your survey)
- Organise, analyse and present the data in a meaningful manner
- Final results are shown in form of charts, tables and graphs
- Tools: Measures of tendency (mean/media/mode), spread of data (standard deviation)

## Inferential statistics

- Make inferences from the sample and generalise them to the population.
- Compares, test, and predict future outcomes
- Final results are probability scores.
- Tools: Analysis of variance, t-test.



# Summary: statistical tests

- Chi-square, t-test and ANOVA are tests to find whether differences between groups are significant or not.
- You choose one of them depending on your variable.
- Define a null hypothesis.
  - Chi-square: Chi-value (frequency)
  - t-test: t-value (mean value)
  - ANOVA: F-value (mean value)
- These values always determine a p-value (significance).
  - A p-value is the probability that the results from your sample data occurred by chance.
  - Significance of your statistical test depends on the p-value ( $<0.05$ )

# Summary: regression

- Flexible way to understand relation between several input variables and an output variable
- Building block also for machine learning models

# Summary: machine learning

- Supervised and Unsupervised methods
- Especially useful for prediction