Danmarks Tekniske Universitet

02323 Introduction to Statistics Fall 22

Project 1: Heating in Sønderborg

October 14, 2022

Written by:
Marcus Wahlers Sand (s215827)

1 Introduction

The following report is handed in for the first mandatory project in the Introduction to Statistics course at DTU (02323).

> The assignment consists of two parts. The first part focuses on descriptive analysis of the data. The second part is primarily about confidence intervals and hypothesis tests. [3]

Below you will find the introductory text describing the project:

> This project focuses on the daily heat consumption for four houses in Sønderborg during the period from October 2008 to June 2011. The relation between the surrounding climate and the heat consumption can give insight into the actual level of insulation of the houses. E.g. the relation between wind speed and heat consumption is an indicator of how airtight/exposed the houses are.
>
> Ideally, this kind of data can be used to make an empirical energy signature for houses. As long as the residents' behaviour doesn't change too much during the period of observation, the analysis will be independent of the indoor temperature. The houses are detached single-family houses. [3]

2 Descriptive Analysis

2.1 Question a

Write a short description of the data.

In the dataset, we have daily observations of the heat consumption in four houses together with the daily averages of centrally observed climatic variables [3]. There are three climatic variables (TA, G & Ws) and one heat consumption variable for each house ($Q1$, $Q2$, $Q3$ & $Q4$). These variables are all quantitative. We also get a date variable, t, in order to tell at what time the observation took place.

The dataset consists of 973 observations, with a period ranging from October 2nd, 2008, to June 1st, 2011. With 973 observations, we have one "observation" per day - or actually, I would rather that we call it one row per day since some days are missing the heat consumption variables. Some days are missing both heat consumption and climatic variables (for example, March 28th, 2011), making the

date variable the only variable filled out for the given row. It is worth mentioning that some days only have data on the heat consumption of *some* houses, meaning there is missing data from the houses. This could be caused by forgetting to submit the data or having the data retrieval fail. For House 3, 511 days are missing data - a little over 50% of the days. Below you can find a table showing the variables (Table 1).

Variable	**Explanation**
t	Date
Ta	Ambient air temperature ($°C$)
G	Global radiation (W/m^2)
Ws	Wind speed (m/s)
$Q1$	Heat consumption in House 1 (kW/day)
$Q2$	Heat consumption in House 2 (kW/day)
$Q3$	Heat consumption in House 3 (kW/day)
$Q4$	Heat consumption in House 4 (kW/day)

Table 1: Table showing the variables in the dataset. [3]

2.2 Question b

Make a density histogram of the daily heat consumption of House 1.

Below (Figure 1) you will find the density histogram of the daily heat consumption of House 1.

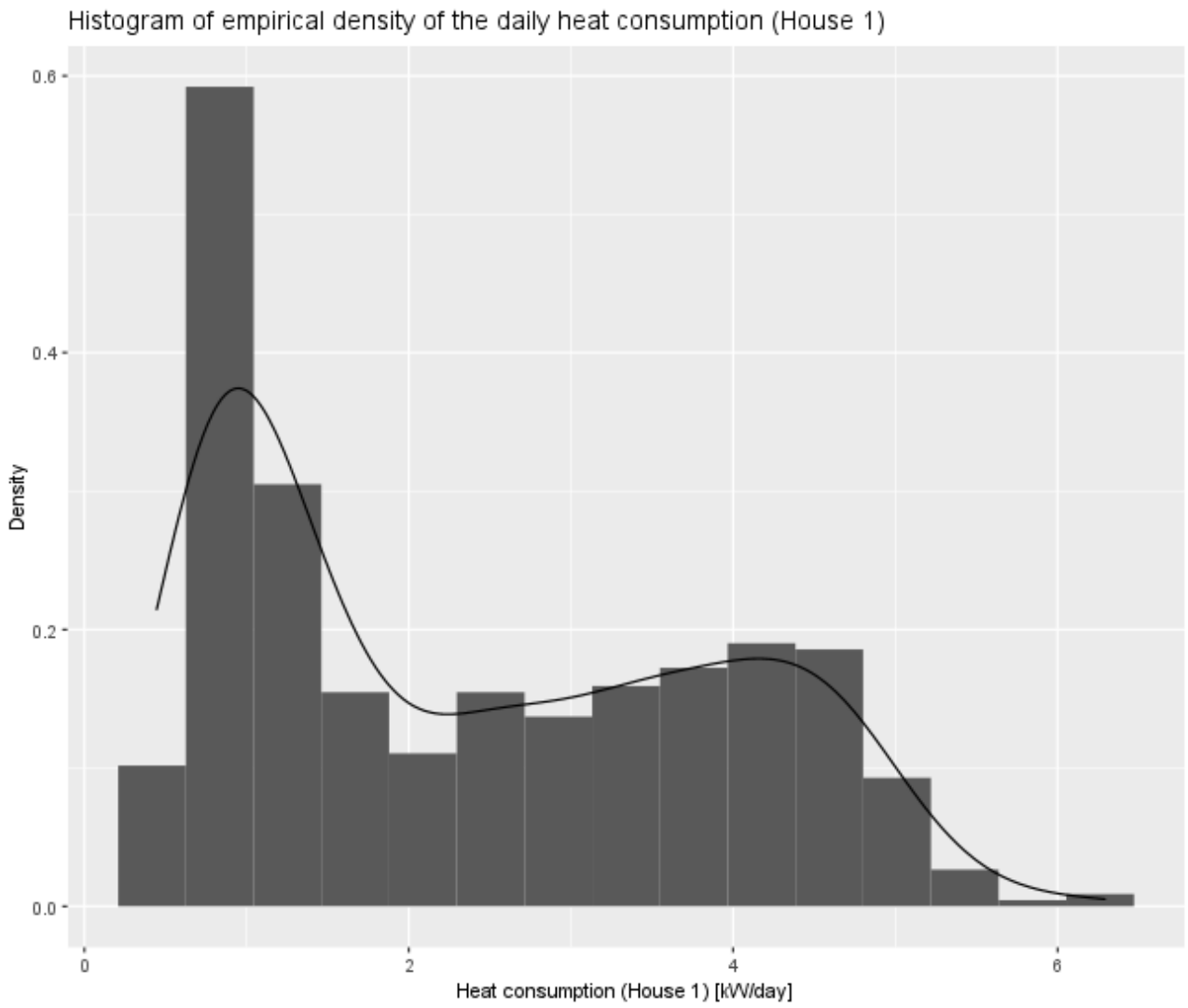


Figure 1: Histogram of the empirical density of the daily heat consumption (House 1)

The empirical distribution of the daily heat consumption is right-skewed (the mean and median are not in the center, and the tail is on the right side) and ranges from around 0.4 kW/day (having used next to no heat) to around 6.3 kW/day - and unless the house is producing and providing heat to others, the heat "consumption" cannot be negative (and even then, you would probably separate the produced heat from the consumed heat, so in that case, the heat consumption cannot be negative). This is quite a big variance in min and max values, but looking at the data, the variance makes sense, given the seasons bring different needs for heat (you don't need that much heat in the summer, but in the winter, you'd want to get some heat).

2.3 Question c

Make a plot illustrating the daily heat consumption over time for the period 2 October 2008 to 1 October 2010.

Below you can see the plot showing the daily heat consumption over time (Figure 2). You can see how during the winter, the heat consumption of all houses rises to their local maximums (when looking at each calendar year alone).

The heat consumption, from looking at the plot, seems very similar across all the houses. You can notice how House 4 seems to use a lot less heat than the other houses during the summer, dropping down to around 0 around May-June, whilst the other houses seem to wait till July (maybe House 4 turns off their radiators earlier than the other houses?).

From looking at the graph, everything looks pretty normal, seeing as the coldest months in Denmark usually are January and February.

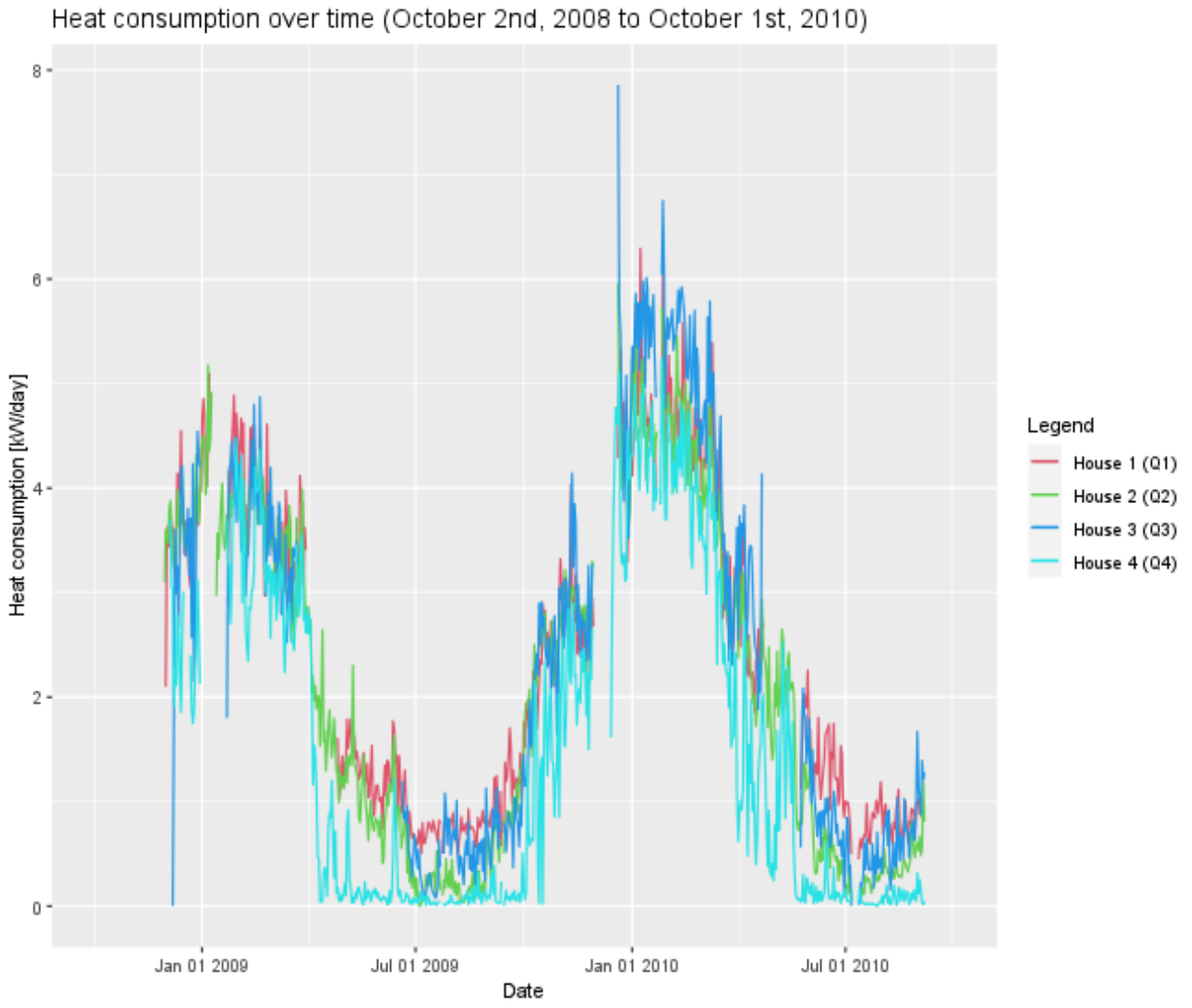


Figure 2: Plot showing the daily heat consumption of all houses for the period 2 October 2008 to 1 October 2010.

2.4 Question d

Make a box plot of the daily heat consumption in January-February 2010 by house.

Below (Figure 3) you can see the box plot showing the daily heat consumption in January-February 2010 by house.

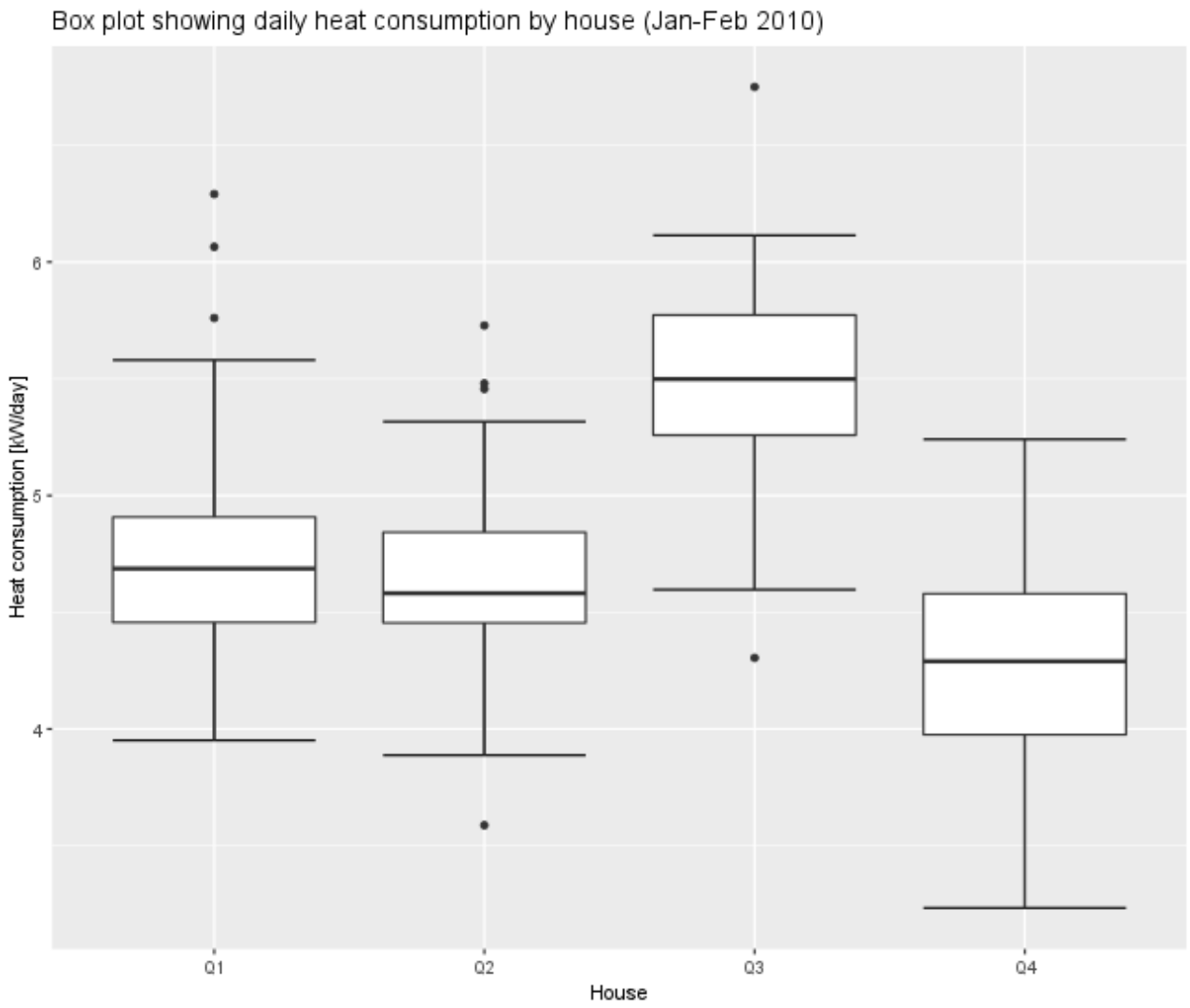


Figure 3: Box Plot showing the daily heat consumption of all houses for the period January 2010 to February 2010.

Through the box plot, we can read the minimum value (within a distance of $1.5 \times IQR$ from Q_1), lower quartile (Q_1), median (Q_2), upper quartile (Q_3) and the maximum value (within a distance of $1.5 \times IQR$ from Q_3) of the daily heat consumption for each house in January-February 2010. If we "ignore" the outliers, and focus on the whiskers and quartiles, then House 1 (Q1; not to be confused with the lower quartile), House 3 (Q3), and House 4 (Q4) look pretty symmetrical - though House 3 has an average maximum value which is not that far away from the upper quartile, making it kind of skewed, but not as much as House 2 (Q2).

We can observe three extreme observations above the whiskers on House 1 and House 2, 1 extreme observation above House 3, and one extreme observation below House 2 and House 3.

House 1 and House 2 seem very similar in their daily heat consumption (with House 1 having some higher extreme observations). House 3 generally had a way higher daily consumption of heat in January 2010 than all the other houses, and House 4 generally had lower daily consumption of heat than the other houses, with the lowest value being way below the others' lowest values, but with the highest value being around the same as House 1 and House 2's whiskers.

2.5 Question e

Compute the relevant summary statistics for the daily heat consumption of each of the four houses during the first two months of 2010 and present the data in a table.

House	Number of obs.	Sample mean	Sample variance	Sample Std. dev.	Lower quartile	Median	Upper quartile
	n	$(\bar{x})$	(s^2)	(s)	(Q_1)	(Q_2)	(Q_3)
House 1	55	4.76	0.21	0.46	4.46	4.69	4.91
House 2	56	4.61	0.19	0.43	4.45	4.58	4.84
House 3	55	5.47	0.19	0.44	5.26	5.50	5.77
House 4	57	4.28	0.17	0.42	3.98	4.29	4.58

Table 2: Relevant summary statistics for the daily heat consumption of each one of the four houses during the first two months of 2010.

Through the table above (Table 2) we can get way more "precise" information than from the box plot. We can read precise values, and as something new, we have access to the sample standard deviance (and hence also the sample variance), which is not readable on a box plot. We also know how many observations have been made for each house.

3 Statistical Analysis

3.1 Question f

Specify separate statistical models describing the daily heat consumption of each of the four houses.

We estimate the parameters of the four models by finding the sample mean and the sample standard deviation for the daily heat consumption of each of the four houses - and then we can assume the statistical models. We will assume the statistical models to express that the samples are normally distributed - and then we will perform model validation afterward.

We will be estimating the parameters using the observations during the first two months of 2010 since that is also the data which is used in the provided Q-Q plot example code, which is said can be used to perform model validation (so it would not make much sense to use the parameters from the whole period - and the whole period is probably not normally distributed either).

House 1:

$$\bar{x} = E(x) = \sum_{\text{all } x} x f(x) = 4.76$$

$$s = \sqrt{s^2} = \sqrt{Var(x)} = \sqrt{\sum_{\text{all } x} (x - \bar{x})^2 f(x)} = 0.46$$

$$X \sim N(\bar{x}, s^2) \implies x \sim N(4.76, 0.46^2)$$

House 2:

$$\bar{x} = E(x) = \sum_{\text{all } x} x f(x) = 4.61$$

$$s = \sqrt{s^2} = \sqrt{Var(x)} = \sqrt{\sum_{\text{all } x}(x - \bar{x})^2 f(x)} = 0.43$$

$$X \sim N(\bar{x}, s^2) \implies x \sim N(4.61, 0.43^2)$$

House 3:

$$\bar{x} = E(x) = \sum_{\text{all } x} x f(x) = 5.47$$

$$s = \sqrt{s^2} = \sqrt{Var(x)} = \sqrt{\sum_{\text{all } x}(x - \bar{x})^2 f(x)} = 0.44$$

$$X \sim N(\bar{x}, s^2) \implies x \sim N(5.47, 0.44^2)$$

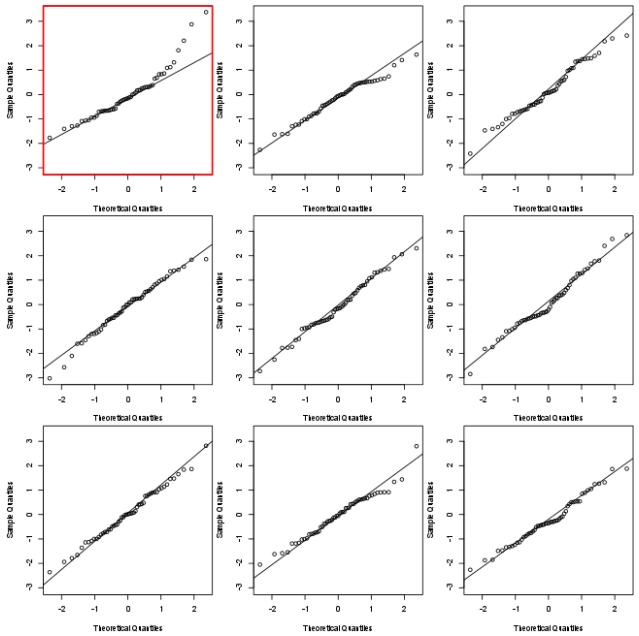
House 4:

$$\bar{x} = E(x) = \sum_{\text{all } x} x f(x) = 4.28$$

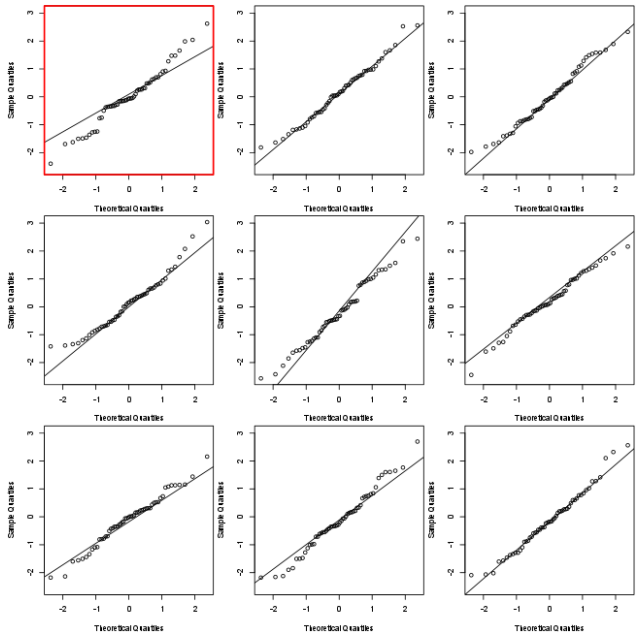
$$s = \sqrt{s^2} = \sqrt{Var(x)} = \sqrt{\sum_{\text{all } x}(x - \bar{x})^2 f(x)} = 0.42$$

$$X \sim N(\bar{x}, s^2) \implies x \sim N(4.28, 0.42^2)$$

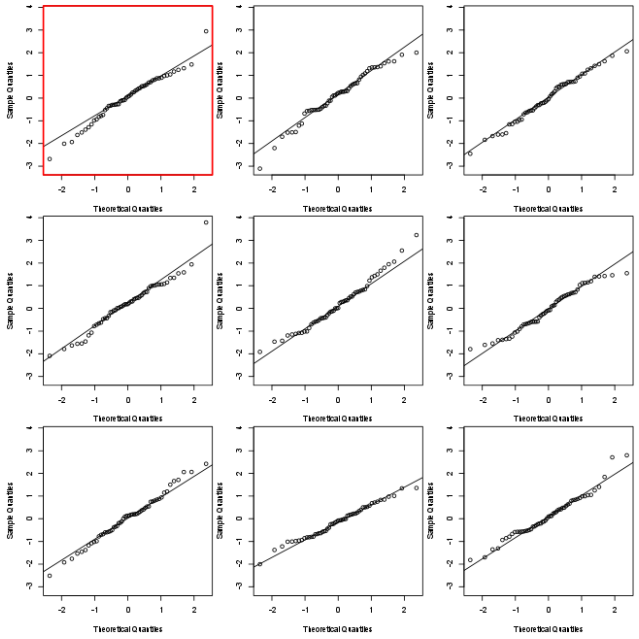
Now, we will try to fit the data into Q-Q plots to perform model validation and check if the statistical models are invalid. The plots are shown in the figure below (Figure 4).



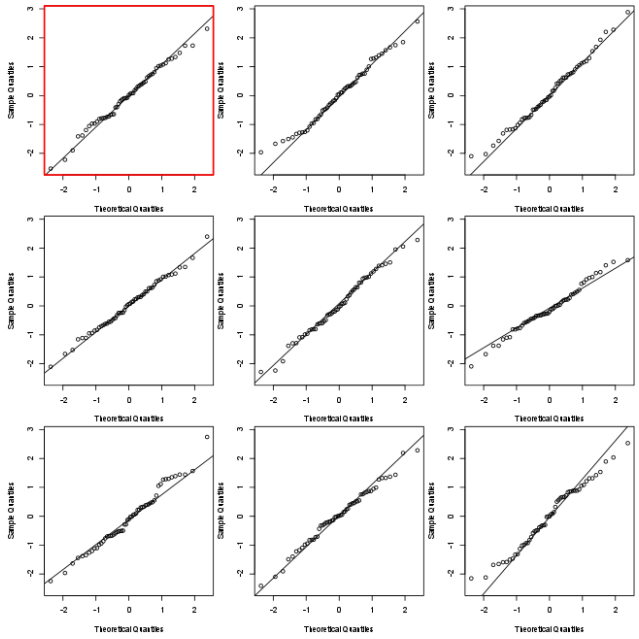
(a) Multiple (simulated) Q-Q Plots for House 1.



(b) Multiple (simulated) Q-Q Plots for House 2.



(c) Multiple (simulated) Q-Q Plots for House 3.



(d) Multiple (simulated) Q-Q Plots for House 4.

Figure 4: Multiple (simulated) Q-Q Plots for all houses during January - February 2010.

As can be observed from the figures, the non-simulated plots (marked with a red border) don't stick out from the crowd of simulated Q-Q plots by an awful lot. We will therefore assume that the model is valid. If they had stood out from the crowd, then we would have had to perform model validation through other means, such as Normality Tests.

We can do this because we are dealing with a distribution of a mean of i.i.d. random variables, and because we have a sample size that is "large enough" (we have at least 55 samples for each house) - which fulfills the CLI (Central Limit Theorem, Theorem 3.14 [1]). This also means that we can now perform confidence intervals and hypothesis tests.

3.2 Question g

State the formula for a 95% confidence interval (CI) for the mean daily heat consumption of House 1 during January - February 2010.

Section 3.1.2 of the book uses Method 3.9 [1] to calculate the confidence interval. Because we want the 95% confidence interval, we pick an a-value of 0.05, which means that we have to find $1 - a/2$, then we get 0.975. This means we need to calculate $t_{0.975}$ to calculate our confidence interval. One can use R to find the value or look the value up in a t-value table since to calculate the t-distribution, there is the need for an unobservable value, μ. The t-distribution function is from Theorem 3.5 [1].

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

$$t_{0.975} = qt(0.975, n-1) = qt(0.975, 54) = 2.00$$

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \implies \bar{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}} \implies 4.76 \pm 2.00 \cdot \frac{0.46}{\sqrt{55}} = [4.63, 4.88]$$

In the table below (Table 3), you can find the CI for all of the houses from January to February 2010.

	Lower bound of CI	**Upper bound of CI**
House 1	4.63	4.88
House 2	4.49	4.72
House 3	5.35	5.59
House 4	4.17	4.39

Table 3: 95% confidence interval (CI) for the mean daily heat consumption of all four houses during January - February 2010.

3.3 Question h

Carry out a hypothesis test in order to assess whether the mean daily heat consumption of House 1 during January-February 2010 is significantly different from 2.38 kW/day.

This can be done by testing the following hypothesis:

$$H_0 : \mu_{House1} = 2.38,$$
$$H_1 : \mu_{House1} \neq 2.38.$$

We will use $\alpha = 0.05$ (5%) as the significance level. This is our "risk level".

Below is the formula for calculating the value of the test statistic (Method 3.23 (3-21) [1]):

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Below you can see the distribution of the test statistic, with 54 (n-1) degrees of freedom (Figure 5).

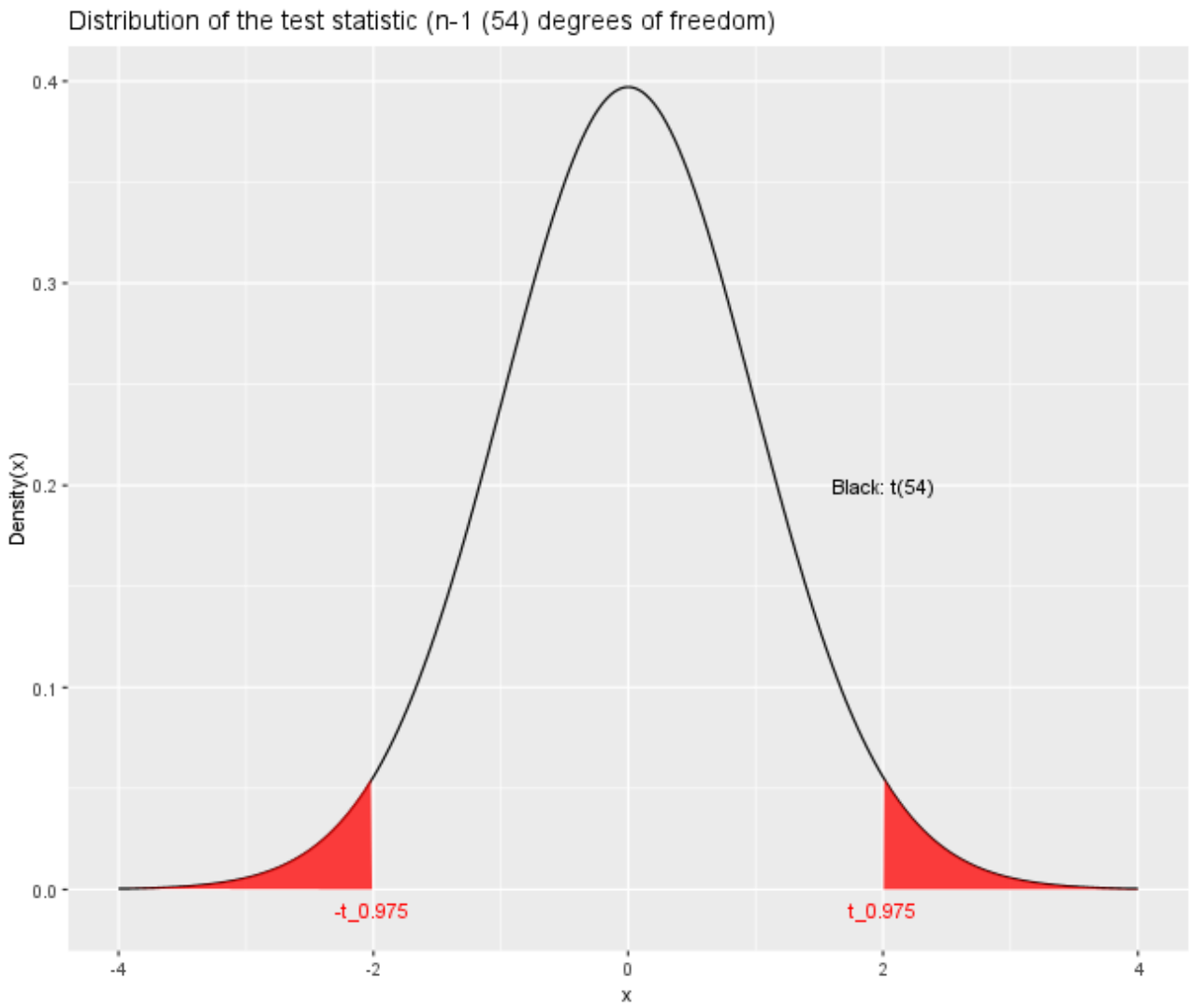


Figure 5: Distribution of the test statistic for House 1, which had 55 observations in January-February 2010, making the degrees of freedom for the distribution 54.

Now we will be calculating the value of the test statistic:

$$\begin{aligned} t_{obs} &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{4.76 - 2.38}{0.46/\sqrt{55}} \\ &= \frac{2.38}{0.46/7.42} \\ &= \frac{2.38}{0.06} \\ &= 38.42 \end{aligned}$$

And now the p-value:

$$\begin{aligned} p\text{-value} &= 2 \cdot P(T > |t_{obs}|) \\ &= 2 \cdot (1 - pt(abs(t_{obs}), df = n - 1)) \end{aligned}$$

$$\begin{aligned} &= 2 \cdot (1 - pt(abs(38.42), df = 55 - 1)) \\ &= 2 \cdot (1 - 1) \\ &= 2 \cdot 0 \\ &= 0 \end{aligned}$$

We don't even have to consult the p-value table (Table 3.1 [1]) to know that this is very strong evidence against H_0, and it is way below our significance level, $\alpha = 0.05$ (5%) - and as such we discard H_0 and go with the alternative hypothesis, that $\mu_{House1} \neq 2.38$.

This outcome was expected, as 2.38 was well outside the bounds of House 1's CI (Confidence Interval), with the mean daily heat consumption in January-February 2010 being twice as high as H_0. As such, this hypothesis test was unnecessary.

3.4 Question i

Carry out a hypothesis test in order to investigate whether the mean daily heat consumption during the first two months of 2010 differs between House 1 and 2.

This can be done by testing the following hypothesis (Method 3.49 (3-47) [1]):

$$\delta = \mu_{House2} - \mu_{House1},$$

$$H_0 : \delta = \delta_0 = 0,$$

$$H_1 : \delta \neq \delta_0.$$

We will use $\alpha = 0.05$ (5%) as the significance level. This is our "risk level".

Below is the formula (Welch two-sample) for calculating the value of the test statistic (Method 3.49 (3-48) [1]):

$$t_{obs} = \frac{(\bar{x_1} - \bar{x_2}) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

We will calculate the degrees of freedom using Theorem 3.50 (3-50) [1]:

$$v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{(\frac{0.46^2}{55} + \frac{0.43^2}{56})^2}{\frac{(0.46^2/55)^2}{55-1} + \frac{(0.43^2/56)^2}{56-1}} = 108.26$$

Below you can see the distribution of the test statistic, with 108.26 (v) degrees of freedom (Figure 6).

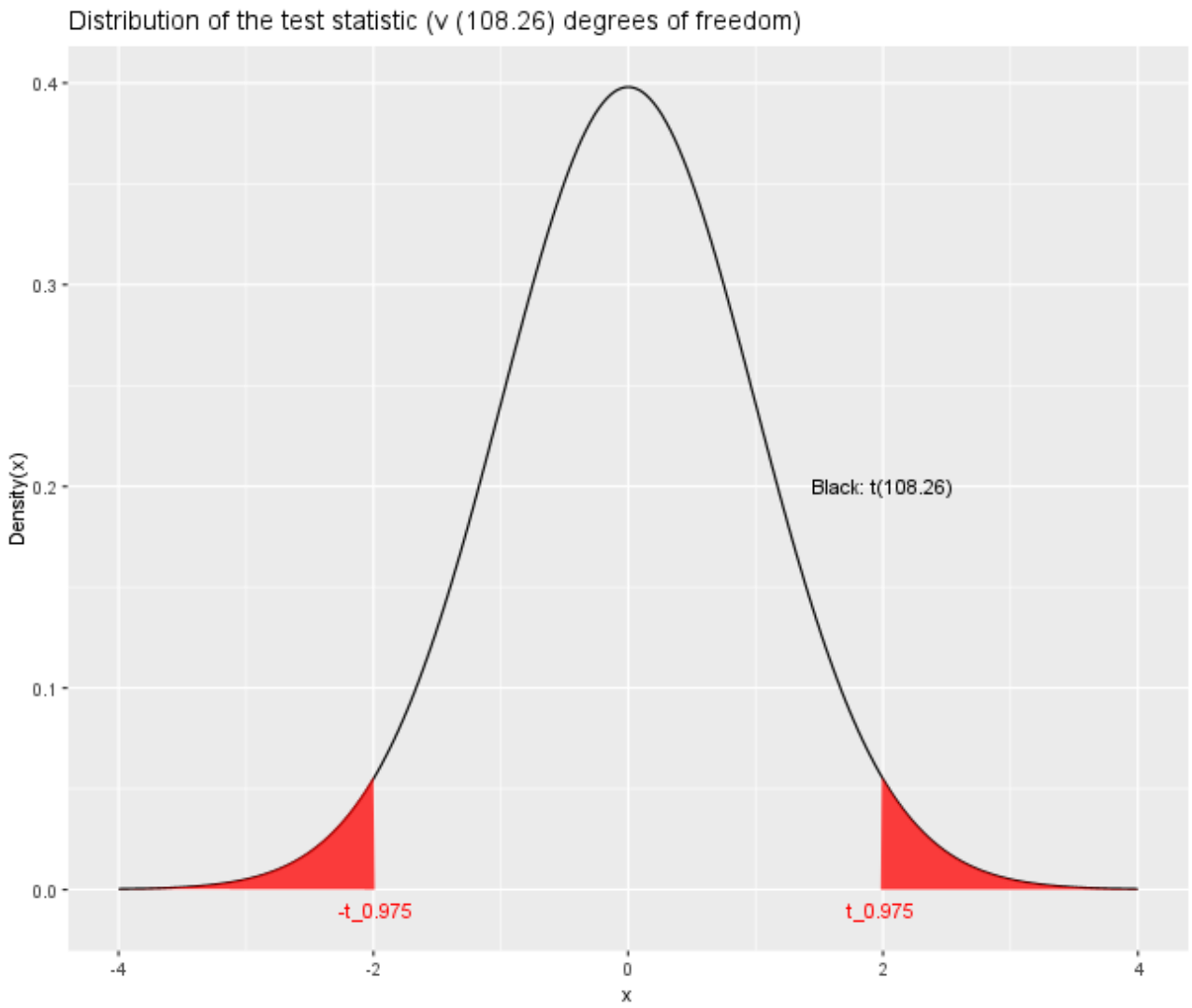


Figure 6: Distribution of the test statistic for House 1 and House 2, which had 55 and 56 observations in January-February 2010, making the degrees of freedom for the distribution 108.26.

Now we will be calculating the value of the test statistic:

$$t_{obs} = \frac{(\bar{x_1} - \bar{x_2}) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{(4.76 - 4.61) - 0}{\sqrt{0.46^2/55 + 0.43^2/56}} = 1.75$$

And now the p-value:

$$\begin{aligned} p\text{-value} &= 2 \cdot P(T > |t_{obs}|) \\ &= 2 \cdot (1 - pt(abs(t_{obs}), df = v)) \\ &= 2 \cdot (1 - pt(abs(1.75), df = 108.26)) \\ &= 2 \cdot (1 - 0.96) \\ &= 2 \cdot 0.04 \\ &= 0.08 \end{aligned}$$

Now, if we consult the p-value table (Table 3.1 [1]) we can interpret that there is weak evidence against H_0, and the p-value is also above our significance level, $\alpha = 0.05$ - and as such we **do not discard** H_0.

Given that the sample means of House 1 and House 2 were quite similar, this outcome was also expected, seeing as the difference between the two sample means was $0.15kW/day$, with the daily heat consumption of House 1 being slightly higher than the daily heat consumption of House 2 during January-February 2010.

3.5 Question j

Comment on whether it was necessary to carry out the statistical test in the previous question or if the same conclusion could have been drawn from the confidence intervals alone.

The lower bound of House 1's confidence interval is inside the bounds of House 2's confidence interval. Still, the upper bound of House 1's confidence interval also goes out of the bounds of House 2's confidence interval.

In other words, we are 95% confident that values ranging from $[4.63, 4.72]$ could have been possible shared population means. The two sample means are 4.76 and 4.61, both being outside of this "shared confidence interval" - hence it would probably still be a good idea to carry out the hypothesis test, even if we were pretty sure that the conclusion could have been drawn from confidence intervals alone (and even then, you should still carry out the hypothesis test).

3.6 Question k

Using data from the whole period of observation, state the formula for computing the correlation between the daily heat consumption of House 1 (Q1) and the global radiation (G).

Using Definition 1.18 (eq. 1-11), 1.19 (eq. 1-12), and 2.62 (eq. 2-78) [1], we can compute the correlation by first computing the sample covariance. Afterward, we can use the sample covariance and the sample deviations to calculate the sample correlation using the formulas below.

Sample covariance (def. 1.18) [1]:

$$s_{xy} = \frac{1}{n-1}\sum_{i=n}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Sample correlation (def. 1.19) [1]:

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

We will start by calculating the needed values using the data from the whole period of observation, October 2008 through June 2011. To do this, we only look at the values of the dataset where both Q1 and G are present (i.e., we don't want to count observations where we have Q1 and not G, and vice-versa). In R, we can do this with a subset, and after that, we can count how many observations are in the subset and perform the sample mean and sample standard deviation functions on that subset:

$$n = 509$$

$$\bar{x} = mean(Q1) = 2.37$$

$$\bar{y} = mean(G) = 143.85$$

$$\bar{s_x} = sd(Q1) = 1.48$$

$$\bar{s_y} = sd(G) = 105.93$$

We will now be inserting values and calculating to find the sample correlation:

$$\begin{aligned} s_{xy} &= \frac{1}{n-1}\sum_{i=n}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{509-1}\sum_{i=n}^{n}(x_i - 2.37)(y_i - 143.85) \\ &= \frac{1}{508} - 60144.82 \\ &= -118.40 \end{aligned}$$

$$\begin{aligned} r &= \frac{s_{xy}}{s_x \cdot s_y} \\ &= \frac{-118.40}{1.48 \cdot 105.93} \\ &= \frac{-118.40}{157.12} \\ &= -0.75 \end{aligned}$$

Now we make a scatter plot of one variable against the other (see Figure 7):

Figure 7: Scatter plot between the daily heat consumption of House 1 (Q1) and the global radiation (G) using the data from the whole period of observation.

As we can see on the scatter plot, the general trend is negative, corresponding to our calculated value of r being negative. From some quick research online on the properties of sample correlation coefficients, a value of -0.75 "indicates variables which can be considered highly correlated" [2]. We can see that there is some correlation in the scatter plot. Still, from the correlation coefficient alone, I had expected a "closer spread" of values - but this is probably also due to how "unstable", "random", and "fluctuating" global radiation is.

A Appendix

A.1 R Code

Please refer to the R-code, which was handed in with this submission.

B Bibliography

[1] Per B. Brockhoff, Jan K. Møller, Elisabeth W. Andersen, Peder Bacher, and Lasse E. Christiansen. *Introduction to Statistics at DTU*. Polyteknisk Kompendie, May 2022.

[2] Keith G. Calkins. Applied statistics - lesson 5: Correlation coefficients. https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm, Jul 2005.

[3] DTU. Project 1: Heating in sønderborg. https://www2.compute.dtu.dk/courses/introstat/projects/soenderborg1.zip, 2022.