Analysis of Trends in Temperatures

Authors:

Thomas Beurksens - i6296221

Timothy Hanson - i6286194

Alexandra Holikova - i6309418

Mathematical Statistics

Faculty of Business and Economics

March 2025

Contents

1 Introduction

In recent decades, climate change has become one of the most pressing global challenges. A key indicator of this phenomenon is the rise in average temperatures across the world. Understanding local temperature trends over time is important for identifying how global climate patterns affect specific regions.

This project analyzes long-term temperature data from the Netherlands to investigate whether there is statistical evidence of a significant upward trend. We focus primarily on annual average temperatures from De Bilt, a centrally located and well-documented weather station Visser [1], and compare the findings with data from Eelde and Maastricht to assess the consistency of patterns across regions.

Our approach starts with visual analysis to identify potential trends. We then apply statistical tools including confidence intervals, hypothesis testing, and linear regression to evaluate whether the observed changes are statistically significant. We use bootstrap methods to improve the reliability of our results, especially in cases where traditional assumptions may not hold.

By combining these techniques, we aim to draw meaningful conclusions about temperature trends in the Netherlands and contribute to the broader understanding of how climate change may be reflected in long-term local data.

2 Problem Statement

The aims of this paper is to determine whether there is statistical evidence of a long-term increase in average temperatures in the Netherlands. Using historical annual temperature data from De Bilt, we investigate whether observed changes over time can be confirmed through, hypothesis testing, linear regression, and bootstrap methods.

2.1 Research Question

Is there statistically significant evidence of an upward trend in average annual temperatures in the Netherlands over the past century, based on data from De Bilt ?

3 Annual Temperature Data Analysis

In this section we analyze the annual temperatures, and split our timespan into decades. The goal is to identify trends across decades and estimate the reliability of these averages using confidence intervals.

To better understand the reliability of each 10-year average temperature, we calculated a 95% confidence interval for every period. This interval gives a range within which we expect the true average temperature to fall, based on the variation in the data and the size of the sample. Since we only have a limited number of years per decade (usually 10), we used the t-distribution:

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

which is more appropriate for smaller sample sizes. The confidence interval takes into account both how much the temperatures vary within a decade and how many data points are available. In general, more variation or fewer data points will lead to wider intervals, indicating more uncertainty.

These intervals were then visualized as error bars in the final graph to show not just the estimated averages, but also the level of confidence we have in each one. Our confidence interval is given by:

$$C(X) = \left[\bar{x} - t_{0.025,n-1} \times \frac{S}{\sqrt{n}},\ \bar{x} + t_{0.025,n-1} \times \frac{S}{\sqrt{n}}\right]$$

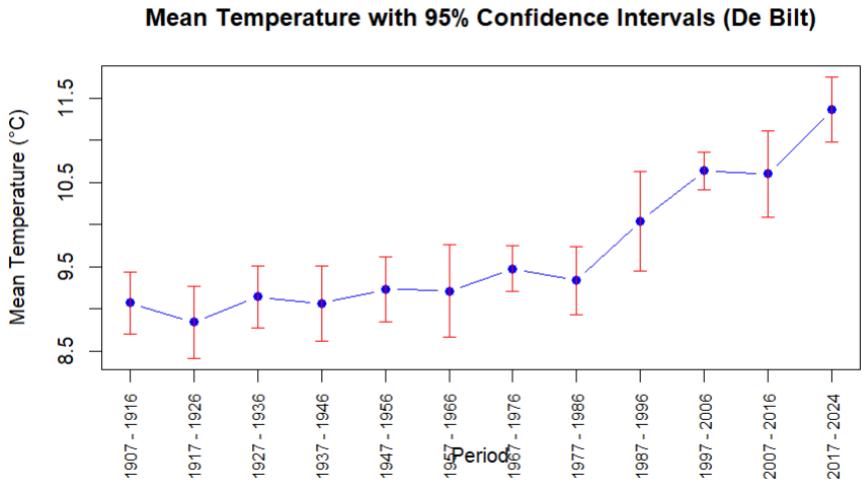

Figure 1: Mean temperature per decade in De Bilt with 95% confidence intervals.

Figure 1 displays the average temperature in De Bilt for each 10-year period from 1907 to 2024. The points represent the mean temperature for each decade, with the connecting line showing overall trends.

The plot illustrates a general upward trend in temperatures over the past century, with some variability across decades. The use of 10-year periods and confidence intervals provides a clearer picture of long-term warming, while also highlighting the degree of statistical uncertainty. Based on this, we proceed with formal inference methods to validate our initial observations.

4 Subsample Comparison Using t-Tests

In this part, we investigate whether there is a statistically significant increase in average annual temperatures in De Bilt after 1970. 1970 was chosen as a cutoff point since it is often used as a climate benchmarkIPCC [2] . To assess this, we perform two types of t-tests: a one-sample t-test using the pre-1970 mean as a benchmark, and a two-sample t-test assuming equal variances. We then compare the results and discuss which test is more appropriate for our context.

4.1 Hypotheses

To assess changes in mean temperature over time, we divided the dataset into two groups:

- Pre-1970: years 1907–1970
- Post-1970: years 1971–2022

We test the following hypotheses for both tests:

$$H_0 : \mu_{\text{post}} \leq \mu_{\text{pre}}$$
$$H_A : \mu_{\text{post}} > \mu_{\text{pre}}$$

4.2 Assumptions

Both tests rely on the following assumptions:

- **Independence:** Temperature observations from different years are independent.
- **Identically distributed:** Each group is assumed to come from a distribution with constant variance.
- **Normality:** The data within each group are approximately normally distributed. Due to large sample sizes, the Central Limit Theorem supports this assumption.

4.3 One-Sample t-Test

We begin by performing a one-sample t-test using the pre-1970 mean as the benchmark to test whether the mean temperature after 1970 is significantly higher.

The test statistic is calculated as:

$$t = \frac{\bar{X}\text{post} - \mu_0}{s/\sqrt{n}}, \quad \text{where } \mu_0 = \bar{X}\text{pre} \tag{1}$$

Results

- $t = 9.618, \quad df = 53, \quad$ critical value $t_{0.05} = 1.674$
- $p \approx 1.61 \times 10^{-13}$

Since $t > t_{0.05}$ and $p < 0.05$, we reject H_0 and conclude that post-1970 temperatures are significantly higher.

4.4 Two-Sample t-Test

We also perform a two-sample t-test, assuming equal variances between the two groups. The test statistic is defined as:

$$t = \frac{\bar{X}\text{post} - \bar{X}\text{pre}}{s_p \sqrt{\frac{1}{n_{\text{post}}} + \frac{1}{n_{\text{pre}}}}} \tag{2}$$

where s_p is the pooled standard deviation:

$$s_p = \sqrt{\frac{(n_{\text{post}} - 1)s_{\text{post}}^2 + (n_{\text{pre}} - 1)s_{\text{pre}}^2}{n_{\text{post}} + n_{\text{pre}} - 2}} \tag{3}$$

The degrees of freedom are computed as:

$$df = n_{\text{post}} + n_{\text{pre}} - 2 \tag{4}$$

Results

- $t = 8.519, \quad df = 116, \quad$ critical value $t_{0.05} = 1.658$
- $p \approx 3.386 \times 10^{-14}$

Since $t > t_{0.05}$ and $p < 0.05$, we reject H_0 and conclude that there is a significant increase in temperature after 1970.

4.5 Confidence Intervals

To account for uncertainty in our estimates, we compute 95% confidence intervals for both group means:

$$\bar{X} \pm t_{1-\alpha/2,\ df} \cdot \frac{s}{\sqrt{n}} \tag{5}$$

The resulting intervals are:

- Pre-1970: [8.967, 9.253]
- Post-1970: [10.026, 10.501]

As these intervals do not overlap, they further support the conclusion of a significant temperature increase.

4.6 Comparison between one and two-sample t-Tests

Both the one-sample and two-sample t-tests provide strong evidence that temperatures in De Bilt have increased significantly since 1970. However, the two-sample t-test is more appropriate for our context because it accounts for estimation uncertainty in both groups. The one-sample test treats the pre-1970 mean as fixed, which may underestimate variability and lead to overconfident conclusions.

Therefore, we base our main conclusion on the two-sample t-test, which shows strong statistical evidence of a significant and meaningful increase in average annual temperatures in De Bilt after 1970.

5 Linear Regression Model

In this section, we investigate the presence of a linear upward trend in average annual temperatures using regression analysis. Compared to simple averages, regression offers a more precise measure of the trend over time. We estimate the following linear model:

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad \forall i = 1, \ldots, n$$

In this formulation:

- Y_i is the dependent variable representing the average annual temperature in year i,
- x_i is the explanatory variable denoting the year (e.g., 1907, 1908, ...),
- α is the intercept, representing the estimated average temperature when $x_i = 0$,
- β is the slope coefficient, indicating the average change in temperature per year,
- ε_i is the error term, capturing unobserved random variation and measurement noise.

Since a linear regression model is being used the following assumptions must be made. We assume that the linear model is correct $\forall i = 1, \ldots, n$ and we assume that x_i is a constant and not random.

Furthermore, it is also assumed that the mean of ε_i is equal to zero and has some variance σ^2 . Furthermore, $\varepsilon_1, \ldots, \varepsilon_n$ are all uncorrelated and its distribution is left unspecified. The model is estimated as an ordinary least squared estimation which aims to minimize the sum of the squared distances between Y_i and $\alpha + \beta x_i$, by doing so we get least squares estimators for α and β which are $\alpha_L s$ and $\beta_L s$. This was implemented and the visual representation of the regression function can be found in Figure 2.

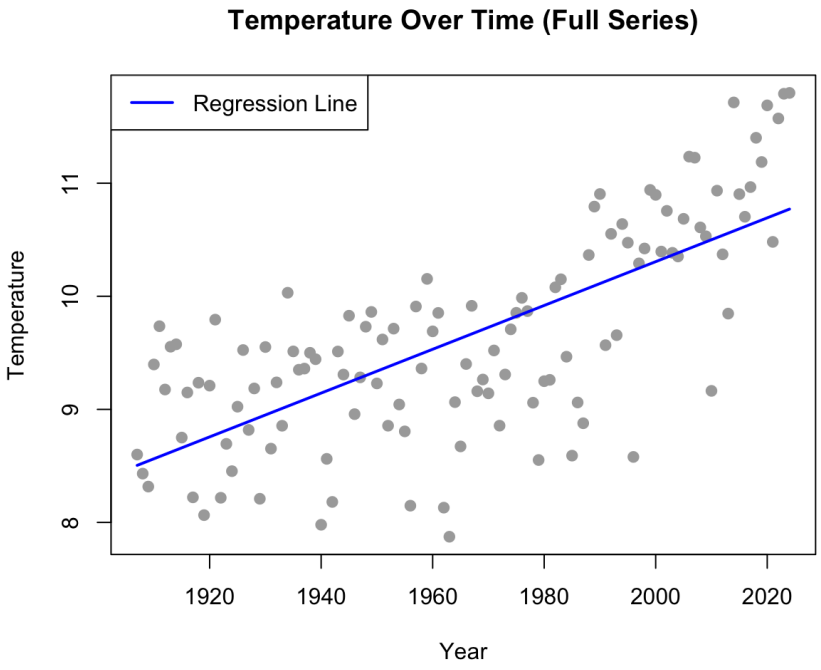

Figure 2: Temperature observations in De Bilt from 1907 to 2022 with a fitted linear regression line.

Once estimated the following poluation model was found

$$Y_i = -28.4346 + 0.0193 x_i \tag{6}$$

Although the plot suggests an upward-sloping trend, this visual observation alone is not sufficient to draw definitive conclusions. Estimation uncertainty remains, especially given that the sample size, while moderate (117 observations), may not be large enough to guarantee precise inference. Additionally, time series data can violate classical assumptions such as independence and homoscedasticity. To address this, we rely on asymptotic analysis, which assumes that the sample size is sufficiently large for statistical approximations to hold.

The asymptotic distribution of the OLS estimator $\hat{\beta}$ is derived using the central limit theorem (CLM), which ensures that under suitable conditions, the standardized estimator converges in

distribution to the normal distribution as the sample size increases.

$$\frac{\hat{\beta}_{n,LS} - \beta}{\sqrt{S_n^2/S_{xx}}} \xrightarrow{d} \mathcal{N}(0,1)$$

Since the true parameter β is unobserved, we estimate it using a 95% confidence interval based on the asymptotic normality of the OLS estimator. The confidence interval is given by:

$$C(X) = \left\{ \hat{\beta}_{n,\text{LS}} \pm z_{0.025} \sqrt{\frac{S_n^2}{S_{xx}}} \right\}$$

where $\hat{\beta}_{n,\text{LS}}$ is the least squares estimator, S_n^2 is the estimator of the error variance, $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, and $z_{0.025} \approx 1.96$ is the 97.5th percentile of the standard normal distribution.

By substituting the values obtained from the regression estimation, the following confidence interval is found:

$$C(X) = \{0.01589; 0.02284\}$$

It can be said that the true value of β lies between these two values with 95% certainty. The estimator $\hat{\beta}_{n,LS}$ lies in this interval as well. Furthermore, from this confidence interval it can be concluded that β is greater that 0 providing the evidence needed for an upward trend. In order to solidify this claim a hypothesis test is conducted.

The null and alternative hypotheses are defined as:

$$H_0 : \beta \leq 0 \qquad \text{vs.} \qquad H_1 : \beta > 0$$

We reject the null hypothesis if the test statistic exceeds the critical value from the standard normal distribution:

$$\frac{\hat{\beta}_{n,\text{LS}} - \beta_0}{\sqrt{S_n^2/S_{xx}}} > z_{0.025}$$

Using our computed values, we find that:

$$\frac{\hat{\beta}_{n,\text{LS}} - 0}{\sqrt{S_n^2/S_{xx}}} = 10.93183 > z_{0.025} = 1.959$$

Since the test statistic exceeds the critical value, we reject the null hypothesis at the 5% significance level. This provides strong statistical evidence in favor of the alternative hypothesis, supporting the conclusion that there is a significant upward trend in temperatures over time.

So far, the complete time series has been used it spans from 1907 to 2024. However, there is reason to believe that this temperature change is more pronounce after the 1970s. Hartmann et al. [3], says that there was a brief cooling period between 1970 and 1979 where certain parts of the northern hemisphere including the Netherlands had inconsistent temperature. So while there wasn't a dramatic cooling plateau between 1970-1979, this period is seen as a transition period. Given this the analysis will be for the years of 1979 until the 2024.

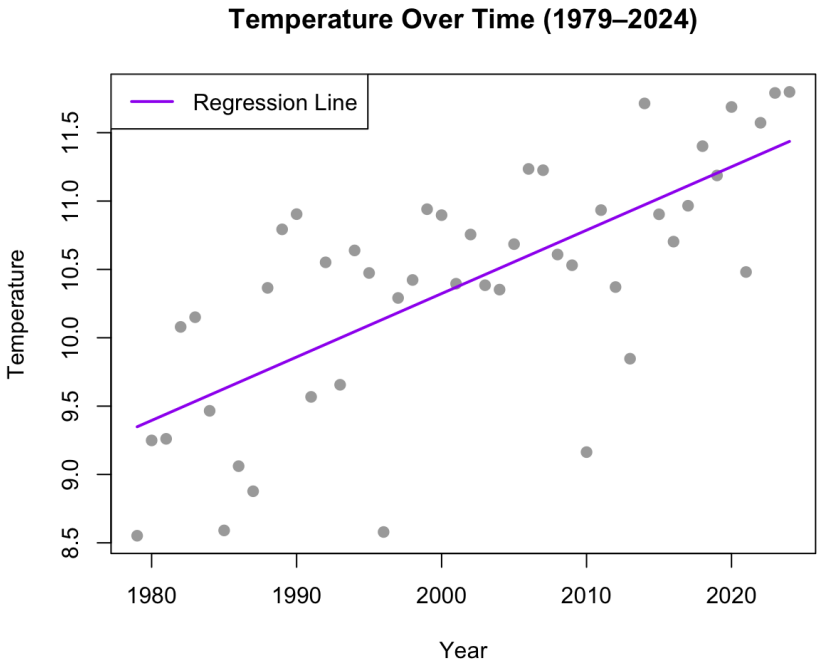

Figure 3: Average annual temperatures in De Bilt from 1979 to 2024. The plot is restricted to this period to investigate whether the upward trend becomes more prominent in recent decades.

In Figure 3, a linear regression line is fitted using the same procedure described earlier.

This new regression line has a slope of 0.04638959 and an intercept of -82.4561 .Furthermore, in order to prove the validity of the results, an asymptotic analysis is conducted in the same manner as before,where we assume n to be large enough. We find the following confidence interval:

$$C(X) = \{0.0327257, 0.06005349\}$$

Performing a one-sided z-test, we obtain a test statistic of 6.654, which is greater than the critical value $z_{0.025} = 1.959$. Therefore, we reject the null hypothesis H_0. This result provides strong evidence suggesting an upward trend in average annual temperatures.

When comparing the two plots, the regression line in Figure 3 is notably steeper than that in Figure 2. Specifically, the estimated slope for the restricted post-1979 sample is 0.0463, compared to

0.0194 for the full time series. Given that the estimated slope in the restricted period is substantially higher, there is compelling evidence that the rate of temperature increase has accelerated since the 1970s.

That said, this does not imply that there was no increase in temperature before 1970—rather, the data suggest that the warming trend has intensified in more recent decades.

6 Bootstrapping Implementation

6.1 Relevance to Temperature Trend Analysis

The bootstrap method (Efron et al. [4]) provides a robust approach to statistical inference that is particularly valuable for our analysis of temperature trends for several reasons:

- **Assumption robustness**: Traditional regression inference assumes normally distributed, homoscedastic errors. Climate data may violate these assumptions due to potential autocorrelation and changing variability over time.
- **Small sample concerns**: While our full dataset spans 1907–2022, subsamples (e.g., post-1979 data) have limited observations where asymptotic approximations may be less reliable.
- **Complex inference**: The bootstrap provides straightforward confidence intervals for our trend estimates without relying on parametric assumptions.

6.2 Bootstrap Methods: Theory and Application

The bootstrap is a non-parametric resampling technique that allows us to estimate the sampling distribution of a statistic by drawing repeated samples from the data. Unlike classical inference, bootstrap methods do not require assumptions about normality or known variances.

General Bootstrap Algorithm

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a sample and let $\hat{\theta} = s(X)$ be a statistic of interest (e.g., the sample mean, a regression coefficient, etc.).

The bootstrap proceeds as follows:

1. Generate B bootstrap samples:

$$X^{*(b)} = \{x_1^{*(b)}, x_2^{*(b)}, \ldots, x_n^{*(b)}\}, \quad \text{sampled with replacement from } X, \quad b = 1, \ldots, B$$

2. For each bootstrap sample, compute the statistic:

$$\hat{\theta}^{*(b)} = s(X^{*(b)})$$

3. Use the empirical distribution $\{\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}\}$ to:

   - Approximate the sampling distribution of $\hat{\theta}$
   - Estimate the standard error:

$$\widehat{\mathrm{SE}}_{\mathrm{boot}}(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{*(b)} - \bar{\theta}^{*}\right)^2}$$

   - Construct confidence intervals using the percentile method:

$$\mathrm{CI}_{1-\alpha}^{\mathrm{boot}} = \left[\hat{\theta}^{*(\alpha/2)}, \hat{\theta}^{*(1-\alpha/2)}\right]$$

Bootstrap Hypothesis Testing

To test a null hypothesis $H_0 : \theta = \theta_0$, we compute the test statistic under the observed data:

$$T = \hat{\theta} - \theta_0$$

Then we generate bootstrap replicates under H_0 (e.g., by resampling from a centered distribution or pooled data) and calculate:

$$T^{*(b)} = \hat{\theta}^{*(b)} - \hat{\theta}, \quad b = 1, \ldots, B$$

The p-value is then estimated as:

$$p_{\mathrm{boot}} = \frac{1}{B}\sum_{b=1}^{B} 1(T^{*(b)} \geq T)$$

Application to Temperature Means

Using $B = 1000$ replications, we applied the bootstrap to estimate confidence intervals for average temperatures before and after 1970 in De Bilt:

$$\begin{aligned} \text{Pre-1979 (bootstrap):} &\quad [8.959, 9.249] \\ \text{Post-1979 (bootstrap):} &\quad [10.007, 10.476] \end{aligned}$$

These closely match the classical confidence intervals and show no overlap, supporting a statistically significant increase.

Bootstrap Test for Difference in Means

We bootstrapped the difference in means $\Delta = \bar{x}_{\text{post}} - \bar{x}_{\text{pre}}$, obtaining:

$$\hat{\Delta} = 1.153^{\circ}\text{C}$$

The one-sided bootstrap p-value (from 1000 resamples) was:

$$p_{\text{boot}} = 0.000$$

indicating a highly significant increase in post-1970 temperatures.

Bootstrap for Linear Trend Estimation

We applied the bootstrap to estimate the slope $\hat{\beta}_1$ from the linear regression model:

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

Bootstrap confidence intervals for the slope were:

- **Full sample (1907–2024):** $[0.01602, 0.02299]$
- **Post-1979:** $[0.03028, 0.0566]$

This suggests that not only is there a warming trend, but that the rate of warming has increased in more recent decades.

Conclusion

The bootstrap results confirm and reinforce our findings from classical inference. The consistent patterns across both methodologies enhance the robustness of our conclusions. The bootstrap's flexibility and minimal assumptions make it particularly valuable in climatological studies, where standard assumptions (e.g., homoscedasticity, normality) may not hold.

7 Extension to Monthly Data

In this section, we extend our analysis from annual to monthly temperature data. This allows us to examine seasonal variations more closely.

7.1 Data Grouping and Seasonal Averages

To study seasonal effects, we split the monthly data into two key seasons:

- **Winter:** January–March
- **Summer:** June–August

We then grouped observations by decade and calculated the mean temperature for each season and decade. For each average, we construct **95% confidence intervals**.

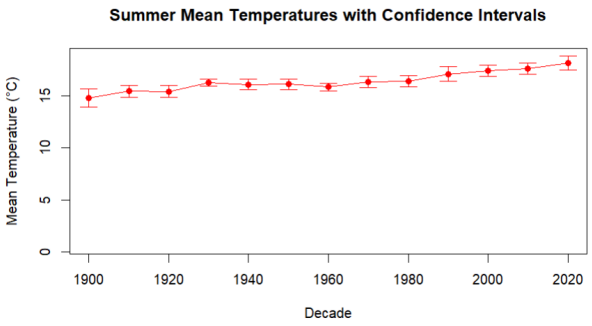

Figure 4: Summer mean temperature with 95% confidence intervals by decade

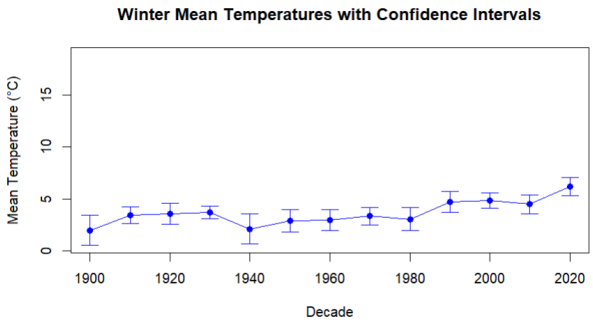

Figure 5: Winter mean temperature with 95% confidence intervals by decade

Figures 4 and 5 display the seasonal means and their associated confidence intervals by decade.

Observed Trends

By comparing the first and last observations for each season, we find:

$$\text{Winter 1900: } 1.99^\circ\text{C} \quad \text{vs} \quad \text{Winter 2020: } 6.21^\circ\text{C}$$
$$\text{Summer 1900: } 14.76^\circ\text{C} \quad \text{vs} \quad \text{Summer 2020: } 18.13^\circ\text{C}$$

While the overall increase in temperature is similar for both seasons, winter exhibits greater decade-to-decade variability, resulting in narrower confidence intervals.

7.2 Linear Regression Analysis

We perform linear regressions using the monthly temperature Y_i as the dependent variable, and a dummy variable x_i for the year:

$$\text{Winter:} \quad Y_i = -33.62 + 0.02 \cdot x_i$$
$$\text{Summer:} \quad Y_i = -27.03 + 0.02 \cdot x_i$$

In both cases, the positive slope confirms a long-term upward trend in seasonal temperatures.

7.3 Hypothesis Testing

We formally tested whether average temperatures after 1970 were significantly higher than those before 1970, using the following hypotheses:

$$H_0 : \mu_{\text{post}} \leq \mu_{\text{pre}}$$
$$H_a : \mu_{\text{post}} > \mu_{\text{pre}}$$

Given our large sample size, we apply an asymptotic z-test. By the CLT, the sampling distribution of the sample mean approximates a normal distribution. The test statistic is given by:

$$Z = \frac{\bar{X}_{\text{post}} - \mu_0}{s/\sqrt{n}} \quad \text{where } \mu_0 = \bar{X}_{\text{pre}}$$

Calculated z-statistics:

$$\text{Winter:} \quad Z = 5.197$$
$$\text{Summer:} \quad Z = 8.95$$

In both cases, the null hypothesis is rejected at typical significance levels, indicating a statistically significant increase in mean temperature after 1970.

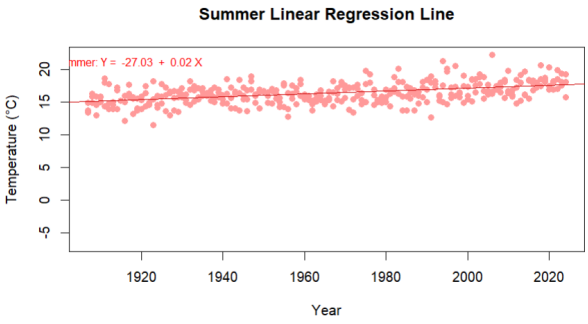

Figure 6: Summer temperature regression line

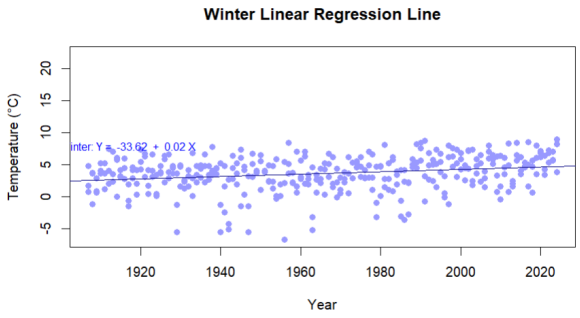

Figure 7: Winter temperature regression line

Both winter and summer months exhibit an upward trend in temperature. Summer temperatures are more consistent across decades, as evidenced by tighter confidence intervals, while winter temperatures show more variability.

8 Conclusion

Our analysis provides strong statistical evidence of a long-term increase in average temperatures in the Netherlands. Both one-sample and two-sample t-tests revealed significantly higher temperatures after 1970, with non-overlapping confidence intervals confirming this trend.

Linear regression further supported these findings, showing a positive and statistically significant slope over the full sample. When restricting the sample to post-1979, the slope nearly doubled, indicating an acceleration in warming in recent decades.

Bootstrap methods confirmed the robustness of these results, yielding similar confidence intervals and significance levels without relying on strict distributional assumptions. The extension to monthly data showed consistent increases in both summer and winter temperatures, with clear upward trends across decades. Together, the evidence points to a clear and statistically significant warming trend in Dutch temperatures over the past century, with a marked increase since the 1970s. These results have further implications within policy making, in regard to greenhouse gas emissions. And provide an eerie warning for the future, but with the implementation of the correct type of policy, things could slow down or even revert.

References

[1] H. Visser. *The significance of climate change in the Netherlands: An analysis of historical and future trends (1901–2020) in weather conditions, weather extremes and temperature-related impacts*. Tech. rep. 550002007/2005. Project S/550002/01/TO, Uncertainties, Transparency and Communication: Tools for Uncertainty Analysis. Bilthoven, Netherlands: RIVM - National Institute for Public Health and the Environment, 2005.

[2] IPCC. *Climate Change 2013: The Physical Science Basis*. Working Group I Contribution to the Fifth Assessment Report. 2013. URL: `https://www.ipcc.ch/report/ar5/wg1/`.

[3] D. L. Hartmann et al. *Climate Change 2013: The Physical Science Basis*. Technical Report. Contribution of Working Group I to the Fifth Assessment Report of the IPCC. Intergovernmental Panel on Climate Change, 2013.

[4] Bradley Efron et al. *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC, 1993. ISBN: 9780412042317.

9 Appendix

````
---
title: "Final_assignment␣Mathematical␣statistics"
format: html
Authors:Timothy,Alexandra,Thomas:(TAT)
---

## Code final project Mathematical statistics

Load data

```{r}
# --- Load Annual Data ---
data <- read.csv2("/Users/timothyhanson/Documents/Temperature␣Data/AnnualTemp.csv"
    )
colnames(data)[1:2] <- c("Year", "DeBilt")
data$Year <- as.numeric(data$Year)
data$DeBilt <- as.numeric(gsub(",", ".", gsub("␣", "", data$DeBilt)))

```

Sampling + averages

```{r}
# Extract years
years <- data[, 1]  # First column assumed to be years

````
````

```r
# Define time range
start_year <- 1907
end_year <- 2024

# Initialize result table
results_df <- data.frame(Period = character(), De_Bilt = numeric(),
    stringsAsFactors = FALSE)

# Compute 10-year means
for (start in seq(start_year, end_year, by = 10)) {
  end <- min(start + 9, end_year)  # Cap at 2024
  period_indices <- which(years >= start & years <= end)  # Indices for this range
  subset_DB <- data[period_indices, 2]  # De Bilt temps

  # Store mean
  results_df <- rbind(results_df, data.frame(
    Period = paste(start, "-", end),
    De_Bilt = mean(subset_DB, na.rm = TRUE)
  ))
}

# Show means
print(results_df)

# CI function for De Bilt temps
calc_CI <- function(data, confidence_level = 0.95) {
  data <- na.omit(data)
  n <- length(data)
  if (n == 0) return(c(NA, NA))  # No data case

  # Compute CI
  mean_val <- mean(data)
  sd_val <- sd(data)
  t_value <- qt(1 - (1 - confidence_level) / 2, df = n - 1)
  margin_error <- t_value * (sd_val / sqrt(n))

  return(c(mean_val - margin_error, mean_val + margin_error))
}

# Initialize CI table
ci_df <- data.frame(Period = character(), De_Bilt_CI = character(),
    stringsAsFactors = FALSE)

# Compute 10-year CIs
for (start in seq(start_year, end_year, by = 10)) {
  end <- min(start + 9, end_year)
  period_indices <- which(years >= start & years <= end)
  subset_DB <- data[period_indices, 2]

```

```
  # Store CI
  ci_DB <- calc_CI(subset_DB)
  ci_df <- rbind(ci_df, data.frame(
    Period = paste(start, "-", end),
    De_Bilt_CI = paste("[", round(ci_DB[1], 2), ",", round(ci_DB[2], 2), "]")
  ))
}

# Show CIs
print(ci_df)

# Merge mean and CI data
final_df <- merge(results_df, ci_df, by = "Period")

# Extract CI bounds
ci_bounds <- do.call(rbind, strsplit(gsub("\\[|\\]", "", final_df$De_Bilt_CI), ","
    ))
final_df$CI_Lower <- as.numeric(ci_bounds[, 1])
final_df$CI_Upper <- as.numeric(ci_bounds[, 2])

# Define x-axis positions
x_pos <- 1:nrow(final_df)

# Plot mean temps with CI
plot(
  x_pos, final_df$De_Bilt,
  type = "b",
  ylim = range(final_df$CI_Lower, final_df$CI_Upper),
  xaxt = "n",
  xlab = "Period",
  ylab = "Mean Temperature ( C )",
  main = "Mean Temperature with 95% Confidence Intervals (De Bilt)",
  col = "blue",
  pch = 16
)

axis(1, at = x_pos, labels = final_df$Period, las = 2, cex.axis = 0.8)

arrows(
  x0 = x_pos, y0 = final_df$CI_Lower,
  x1 = x_pos, y1 = final_df$CI_Upper,
  angle = 90, code = 3, length = 0.05, col = "red"
)

```

Hypothesis Testing

```{r}
```

```
# Split groups
pre <- data$DeBilt[data$Year <= 1970]
post <- data$DeBilt[data$Year > 1970]

# Sample sizes & stats
n_pre <- length(pre)
n_post <- length(post)
mean_pre <- mean(pre)
mean_post <- mean(post)
var_pre <- var(pre)
var_post <- var(post)

# One-sample t-test (post vs. pre mean)
t1 <- (mean_post - mean_pre) / (sqrt(var_post / n_post))
df1 <- n_post - 1
p1 <- 1 - pt(t1, df1)

cat("One-sample t-test:\n")
cat("t = ", round(t1, 4), " | df = ", df1, " | p = ", format(p1, scientific = TRUE), "\
    n\n")

# Two-sample t-test (equal variance)
df2 <- n_pre + n_post - 2
pooled_sd <- sqrt(((n_pre - 1)*var_pre + (n_post - 1)*var_post) / df2)
t2 <- (mean_post - mean_pre) / (pooled_sd * sqrt(1/n_pre + 1/n_post))
p2 <- 1 - pt(t2, df2)

cat("Two-sample t-test:\n")
cat("t = ", round(t2, 4), " | df = ", df2, " | p = ", format(p2, scientific = TRUE), "\
    n\n")

# CIs
alpha <- 0.05
t_crit <- qt(1 - alpha / 2, df2)
ci_pre <- mean_pre + c(-1, 1) * t_crit * sqrt(var_pre / n_pre)
ci_post <- mean_post + c(-1, 1) * t_crit * sqrt(var_post / n_post)

cat("95% CI for pre-1970 mean:", round(ci_pre, 2), "\n")
cat("95% CI for post-1970 mean:", round(ci_post, 2), "\n")

```

Linear Regression

```{r}


# --- Full dataset regression ---
x <- data$year
```

```r
y <- data$temp
x_bar <- mean(x)
y_bar <- mean(y)

b1 <- sum((x - x_bar) * (y - y_bar)) / sum((x - x_bar)^2)
b0 <- y_bar - b1 * x_bar
y_pred <- b0 + b1 * x
residuals <- y - y_pred

# --- Compute confidence interval and stats ---
SST <- sum((y - y_bar)^2)
SSE <- sum(residuals^2)
R_squared <- 1 - (SSE / SST)

n <- length(x)
s_squared <- SSE / (n - 2)
SE_b1 <- sqrt(s_squared / sum((x - x_bar)^2))
SE_b0 <- sqrt(s_squared * (1/n + x_bar^2 / sum((x - x_bar)^2)))

t_b1 <- b1 / SE_b1
t_b0 <- b0 / SE_b0

# 95% Z-confidence interval (asymptotic)
alpha <- 0.05
z_alpha <- qnorm(1 - alpha / 2)
CI_lower <- b1 - z_alpha * SE_b1
CI_upper <- b1 + z_alpha * SE_b1

# --- Print full regression output ---
cat("FULL DATASET REGRESSION (all years)\n")
cat("Intercept (b0):", b0, "\n")
cat("Slope (b1):", b1, "\n")
cat("R-squared:", R_squared, "\n")
cat("Standard Error (slope):", SE_b1, "\n")
cat("Standard Error (intercept):", SE_b0, "\n")
cat("Z-test statistic (slope  0):", (1 - 0) / SE_k1, "\n")
cat("Z-based 95% CI for slope: [", CI_lower, ",", CI_upper, "]\n\n")

# --- Plot full data and regression line ---
plot(x, y, main = "Temperature Over Time (Full Series)", xlab = "Year", ylab = "
    Temperature", pch = 19, col = "darkgray")
lines(x, y_pred, col = "blue", lwd = 2)
legend("topleft", legend = "Regression Line", col = "blue", lwd = 2)

# --- Subset data: 1979 to 2024 ---
filtered_data <- subset(data, year >= 1979 & year <= 2024)
c <- filtered_data$year
z <- filtered_data$temp
c_bar <- mean(c)
```

```r
z_bar <- mean(z)

# --- Regression on subset ---
k1 <- sum((c - c_bar) * (z - z_bar)) / sum((c - c_bar)^2)
k0 <- z_bar - k1 * c_bar
z_pred <- k0 + k1 * c
residuals_subset <- z - z_pred

SST_subset <- sum((z - z_bar)^2)
SSE_subset <- sum(residuals_subset^2)
R_squared_subset <- 1 - (SSE_subset / SST_subset)

n_subset <- length(c)
s_squared_subset <- SSE_subset / (n_subset - 2)
SE_k1 <- sqrt(s_squared_subset / sum((c - c_bar)^2))
SE_k0 <- sqrt(s_squared_subset * (1/n_subset + c_bar^2 / sum((c - c_bar)^2)))
t_k1 <- k1 / SE_k1
t_k0 <- k0 / SE_k0


# --- Print subset regression output ---
cat("SUBSET␣REGRESSION␣(1979  2024  )\n")
cat("Intercept␣(k0):", k0, "\n")
cat("Slope␣(k1):", k1, "\n")
cat("R-squared:", R_squared_subset, "\n")
cat("Standard␣Error␣(slope):", SE_k1, "\n")
cat("Standard␣Error␣(intercept):", SE_k0, "\n")
cat("Z-value␣(slope):", t_k1, "\n\n")

# --- Plot subset regression ---
plot(c, z, main = "Temperature␣Over␣Time␣(1979  2024  )", xlab = "Year", ylab = "
    Temperature", pch = 19, col = "darkgray")
lines(c, z_pred, col = "purple", lwd = 2)
legend("topleft", legend = "Regression␣Line", col = "purple", lwd = 2)
# --- 95% Z-confidence interval (asymptotic) for slope from 1979  2024  ---
CI_lower_subset <- k1 - z_alpha * SE_k1
CI_upper_subset <- k1 + z_alpha * SE_k1

# Print confidence interval and test statistic
cat("Z-based␣95%␣CI␣for␣slope␣(1979  2024  ):␣[", CI_lower_subset, ",", CI_upper_
    subset, "]\n")
cat("Z-test␣statistic␣(slope␣  ␣0):", (k1 - 0) / SE_k1, "\n")
```

```

**Bootstrap**

```{r}
```
```

```r
# Split data into pre-1970 and post-1970
pre_data <- subset(data, Year < 1970)
post_data <- subset(data, Year >= 1970)

# Define temperature vectors
pre <- pre_data$DeBilt
post <- post_data$DeBilt

# Define sample sizes
n_pre <- length(pre)
n_post <- length(post)

# Means
mean_pre <- mean(pre)
mean_post <- mean(post)

# --- BOOTSTRAP SECTION ---
set.seed(123)  # for reproducibility
B <- 1000      # number of bootstrap replications

# --- Bootstrap Confidence Intervals for the Mean ---
boot_mean <- function(data, B) {
  n <- length(data)
  replicate(B, mean(sample(data, n, replace = TRUE)))
}

boot_pre <- boot_mean(pre, B)
boot_post <- boot_mean(post, B)

# Confidence intervals
ci_boot_pre <- quantile(boot_pre, c(0.025, 0.975))
ci_boot_post <- quantile(boot_post, c(0.025, 0.975))

cat("Bootstrap 95% CI for pre-1970 mean:", round(ci_boot_pre, 3), "\n")
cat("Bootstrap 95% CI for post-1970 mean:", round(ci_boot_post, 3), "\n\n")

# --- Bootstrap Hypothesis Test (mean difference post - pre) ---
obs_diff <- mean_post - mean_pre
combined <- c(pre, post)

boot_diff <- replicate(B, {
  sample_combined <- sample(combined, length(combined), replace = TRUE)
  sample_post <- sample_combined[1:n_post]
  sample_pre <- sample_combined[(n_post+1):(n_post + n_pre)]
  mean(sample_post) - mean(sample_pre)
})

p_boot <- mean(boot_diff >= obs_diff)

```

```r
cat("Bootstrap hypothesis test:\n")
cat("Observed mean difference (post - pre):", round(obs_diff, 3), "\n")
cat("Bootstrap p-value (one-sided):", format(p_boot, scientific = TRUE), "\n\n")

# --- Bootstrap for Regression Slope ---
x <- data$Year
y <- data$DeBilt
n_total <- length(y)

boot_regression <- function(x, y, B) {
  replicate(B, {
    idx <- sample(1:n_total, n_total, replace = TRUE)
    coef(lm(y[idx] ~ x[idx]))[2]
  })
}

boot_slope_all <- boot_regression(x, y, B)
ci_slope_all <- quantile(boot_slope_all, c(0.025, 0.975))

# Post-1979 regression
post1979_data <- subset(data, Year > 1979)
x_post <- post1979_data$Year
y_post <- post1979_data$DeBilt

boot_reg_post <- function(x, y, B) {
  replicate(B, {
    idx <- sample(1:length(y), length(y), replace = TRUE)
    coef(lm(y[idx] ~ x[idx]))[2]
  })
}

boot_slope_post <- boot_reg_post(x_post, y_post, B)
ci_slope_post <- quantile(boot_slope_post, c(0.025, 0.975))

cat("Bootstrap 95% CI for slope (1907  2024  ):", round(ci_slope_all, 5), "\n")
cat("Bootstrap 95% CI for slope (post-1979):", round(ci_slope_post, 5), "\n")

```

Monthly Data

```{r}
# Load and clean data
data <- read.csv2("/Users/User/OneDrive/E&OR year 2/Mathematical Statistics/
    MonthlyTemp.csv")
colnames(data) <- trimws(colnames(data))  # Trim column names
data$Date <- as.character(data$Date)      # Convert Date to character
data$Year <- as.integer(substr(data$Date, 1, 4))  # Get year
data$Month <- as.integer(substr(data$Date, 5, 6)) # Get month
```

```r

# Label seasons and filter for Winter and Summer
data$Season <- ifelse(data$Month %in% c(1, 2, 3), "Winter", ifelse(data$Month %in%
    c(6, 7, 8), "Summer", "Other"))
winter_summer_data <- data[data$Season %in% c("Winter", "Summer"), ]
winter_summer_data$Decade <- floor(winter_summer_data$Year / 10) * 10

# Prepare results table
results <- data.frame(Decade = integer(), Season = character(), Mean_Temperature =
    numeric(),
                      CI_Lower = numeric(), CI_Upper = numeric(), Slope = numeric
                          (), Intercept = numeric(), RSS = numeric())

# Loop over seasons
for (season in c("Winter", "Summer")) {
  season_data <- winter_summer_data[winter_summer_data$Season == season, ]  #
      Filter by season

  # Loop over decades
  for (decade in unique(season_data$Decade)) {
    subset_data <- season_data[season_data$Decade == decade, ]
    if (nrow(subset_data) < 2) next  # Skip if too few observations

    # Mean temp and 95% CI
    mean_temp <- mean(subset_data$De.Bilt, na.rm = TRUE)
    sd_temp <- sd(subset_data$De.Bilt, na.rm = TRUE)
    n <- length(subset_data$De.Bilt)
    se_temp <- sd_temp / sqrt(n)
    t_value <- qt(0.975, df = n - 1)
    ci_lower <- mean_temp - t_value * se_temp
    ci_upper <- mean_temp + t_value * se_temp

    results <- rbind(results, data.frame(Decade = decade, Season = season,
                                         Mean_Temperature = mean_temp, CI_Lower =
                                             ci_lower, CI_Upper = ci_upper,
                                         Slope = NA, Intercept = NA, RSS = NA))

    # Linear regression for current decade
    season_data$Decade_Dummy <- ifelse(season_data$Decade == decade, 1, 0)
    X <- season_data$Decade_Dummy
    Y <- season_data$De.Bilt
    X_bar <- mean(X, na.rm = TRUE)
    Y_bar <- mean(Y, na.rm = TRUE)
    numerator <- sum((X - X_bar) * (Y - Y_bar), na.rm = TRUE)
    denominator <- sum((X - X_bar)^2, na.rm = TRUE)
    beta1 <- numerator / denominator
    beta0 <- Y_bar - beta1 * X_bar

    # Save regression results
```

```
    results$Slope[results$Decade == decade & results$Season == season] <- beta1
    results$Intercept[results$Decade == decade & results$Season == season] <-
        beta0

    # Residual sum of squares
    Y_hat <- beta0 + beta1 * X
    residuals <- Y - Y_hat
    rss <- sum(residuals^2, na.rm = TRUE)
    results$RSS[results$Decade == decade & results$Season == season] <- rss

    # Output RSS
    cat("Season:", season, "Decade:", decade, "RSS:", rss, "\n")
  }
}

# Display results
print(results)

# Plot Winter mean temps with confidence intervals
plot(results$Decade[results$Season == "Winter"], results$Mean_Temperature[results$
    Season == "Winter"],
     type = "o", col = "blue", xlab = "Decade", ylab = "Mean Temperature ( C )",
     main = "Winter Mean Temperatures with Confidence Intervals",
     ylim = range(results$CI_Lower, results$CI_Upper, na.rm = TRUE), pch = 16)
arrows(results$Decade[results$Season == "Winter"], results$CI_Lower[results$Season
    == "Winter"],
       results$Decade[results$Season == "Winter"], results$CI_Upper[results$Season
           == "Winter"],
       angle = 90, code = 3, length = 0.1, col = "blue")

# Plot Summer mean temps with confidence intervals
plot(results$Decade[results$Season == "Summer"], results$Mean_Temperature[results$
    Season == "Summer"],
     type = "o", col = "red", xlab = "Decade", ylab = "Mean Temperature ( C )",
     main = "Summer Mean Temperatures with Confidence Intervals",
     ylim = range(results$CI_Lower, results$CI_Upper, na.rm = TRUE), pch = 16)
arrows(results$Decade[results$Season == "Summer"], results$CI_Lower[results$Season
    == "Summer"],
       results$Decade[results$Season == "Summer"], results$CI_Upper[results$Season
           == "Summer"],
       angle = 90, code = 3, length = 0.1, col = "red")

# Plot Winter data and regression line
plot(winter_summer_data$Year[winter_summer_data$Season == "Winter"],
     winter_summer_data$De.Bilt[winter_summer_data$Season == "Winter"],
     type = "p", col = rgb(0.6, 0.6, 1), pch = 16, xlab = "Year", ylab = "
         Temperature ( C )",
     main = "Winter Linear Regression Line",
     ylim = range(winter_summer_data$De.Bilt, na.rm = TRUE))
```

```r

# Add regression line and formula for Winter
winter_model <- lm(De.Bilt ~ Year, data = winter_summer_data[winter_summer_data$
    Season == "Winter", ])
abline(winter_model, col = rgb(0, 0, 0.5), lty = 1)
text(x = min(winter_summer_data$Year[winter_summer_data$Season == "Winter"]) + 10,
     y = max(winter_summer_data$De.Bilt[winter_summer_data$Season == "Winter"]) -
         1,
     labels = paste("Winter: Y = ", round(coef(winter_model)[1], 2), " + ", round(
         coef(winter_model)[2], 2), "X"),
     col = "blue", cex = 0.8)

# Plot Summer data and regression line
plot(winter_summer_data$Year[winter_summer_data$Season == "Summer"],
     winter_summer_data$De.Bilt[winter_summer_data$Season == "Summer"],
     type = "p", col = rgb(1, 0.6, 0.6), pch = 16, xlab = "Year", ylab = "
         Temperature ( C )",
     main = "Summer Linear Regression Line",
     ylim = range(winter_summer_data$De.Bilt, na.rm = TRUE))

# Add regression line and formula for Summer
summer_model <- lm(De.Bilt ~ Year, data = winter_summer_data[winter_summer_data$
    Season == "Summer", ])
abline(summer_model, col = rgb(0.8, 0, 0), lty = 1)
text(x = min(winter_summer_data$Year[winter_summer_data$Season == "Summer"]) + 10,
     y = max(winter_summer_data$De.Bilt[winter_summer_data$Season == "Summer"]) -
         1,
     labels = paste("Summer: Y = ", round(coef(summer_model)[1], 2), " + ", round(
         coef(summer_model)[2], 2), "X"),
     col = "red", cex = 0.8)

# Split data before and after 1970
pre_1970_winter <- winter_summer_data[winter_summer_data$Year >= 1900 & winter_
    summer_data$Year < 1970 & winter_summer_data$Season == "Winter", ]
post_1970_winter <- winter_summer_data[winter_summer_data$Year >= 1970 & winter_
    summer_data$Year <= 2020 & winter_summer_data$Season == "Winter", ]
pre_1970_summer <- winter_summer_data[winter_summer_data$Year >= 1900 & winter_
    summer_data$Year < 1970 & winter_summer_data$Season == "Summer", ]
post_1970_summer <- winter_summer_data[winter_summer_data$Year >= 1970 & winter_
    summer_data$Year <= 2020 & winter_summer_data$Season == "Summer", ]

# Z-test function
z_test <- function(pre_temp, post_temp) {
  mean_post <- mean(post_temp)
  sd_post <- sd(post_temp)
  n_post <- length(post_temp)
  mean_pre <- mean(pre_temp)
  se_post <- sd_post / sqrt(n_post)
  z_stat <- (mean_post - mean_pre) / se_post
```

```
  p_value <- 1 - pnorm(z_stat)
  return(list(z_stat = z_stat, p_value = p_value))
}

# Run z-tests and print results
winter_test <- z_test(pre_1970_winter$De.Bilt, post_1970_winter$De.Bilt)
cat("Winter - Z-statistic:", round(winter_test$z_stat, 3), "\n")
cat("Winter - P-value:", round(winter_test$p_value, 3), "\n")

summer_test <- z_test(pre_1970_summer$De.Bilt, post_1970_summer$De.Bilt)
cat("Summer - Z-statistic:", round(summer_test$z_stat, 3), "\n")
cat("Summer - P-value:", round(summer_test$p_value, 3), "\n")

# Interpret Winter result
if (winter_test$p_value < 0.05) {
  cat("Winter - Significant increase after 1970.\n")
} else {
  cat("Winter - No significant change after 1970.\n")
}

# Interpret Summer result
if (summer_test$p_value < 0.05) {
  cat("Summer - Significant increase after 1970.\n")
} else {
  cat("Summer - No significant change after 1970.\n")
}

```
```
```