Is Germany getting hotter?

Karthik

In this study, I show that the average temperature of Germany since the 90's has increased significantly than the years before.

The data set used contains 138 years of monthly temperature averages in Germany back to 1881. **We can see that the 5 of the 10 warmest years since 1881 occur from year 2011**.

The average temperature rose by **1.06 C** since the 90's.

By making a **hypothesis test**, based on the historical data and using **bootstrap**, one can show that the weather in Germany has really changed.

Loading the data

```r
library(readxl)
germany_old= read_excel("dw_temperature in Germany.xlsx")
```

To meet our objective, we modify our question of interest as:

Is the average Germany temperature since the 90's, statistically significant and higher than the years before dating back to 1881?

In order to answer it, we will make a transformation on the data to have categorical variables indicating the century and whether the average temperature is from a year after or before the 90's.

Processing the data by creating new columns of our choice

```r
germany_old$Time_Era = ifelse(germany_old$year>1990,
                              "After 90s", "Before 90s")

germany_old$Century = with(germany_old, ifelse(year<=1899,19,
                      ifelse(year>=1900&year<=1999,20,
                             ifelse(year>=2000,21,""))))

germany_old$Avg_temp_year = rowMeans(germany_old[,2:13])
```

Sorting data and reshuffling the columns

```r
germany_old = germany_old[with(germany_old, order(-Avg_temp_year)),]

germany_old = germany_old[, c(16,14,15, 1:13)]
germany_old = germany_old[, c(4,1,2,3, 5:16)]
```

Let's make a ranking of the top 10 temperatures?

```r
df = head(germany_old,10)
df
```

```
## # A tibble: 10 x 16
##     year Avg_temp_year Time_Era Century January February March April   May
##    <dbl>         <dbl> <chr>    <chr>     <dbl>    <dbl> <dbl> <dbl> <dbl>
## 1   2014          10.3  After 9~ 21          2.1      4.3   6.9  10.8  12.4
## 2   2015           9.94 After 9~ 21          2.2      0.7   5.2   8.4  12.3
## 3   2000           9.88 After 9~ 21          1.1      3.9   5.2  10.2  14.6
## 4   2007           9.87 After 9~ 21          4.8      3.9   6.2  11.5  14.2
## 5   1994           9.68 After 9~ 20          3       -0.2   6.3   7.9  12.6
## 6   2011           9.63 After 9~ 21          1        0.9   4.9  11.6  13.9
## 7   2002           9.57 After 9~ 21          1.2      5.1   5.4   8.1  13.8
## 8   2017           9.57 After 9~ 21         -2.2      2.9   7.2   7.4  14.1
## 9   2016           9.55 After 9~ 21          1        3.3   4     7.9  13.7
## 10  1934           9.55 Before ~ 20          0.5      1.2   4.1  10.3  13.3
## # ... with 7 more variables: June <dbl>, July <dbl>, August <dbl>,
## #   September <dbl>, October <dbl>, November <dbl>, December <dbl>
```

9 from 1990 on. 8 if you start from 2000. 6 if you expand to 2007. 5 occurred from 2011 year on.

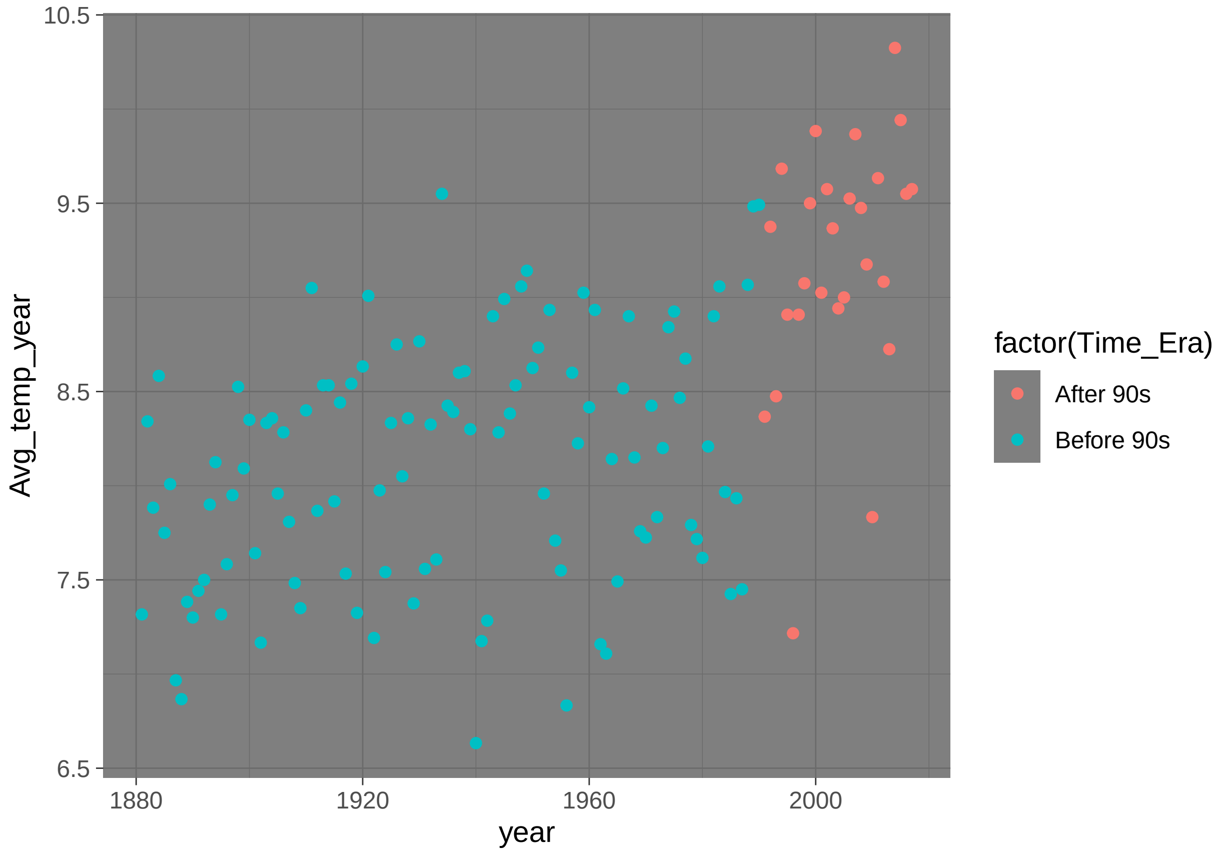
Removing the row with missing data

```
germany_old_1 = germany_old[-138,]
```

Let's take a deep investigation on these numbers

Scatter Plot of the data

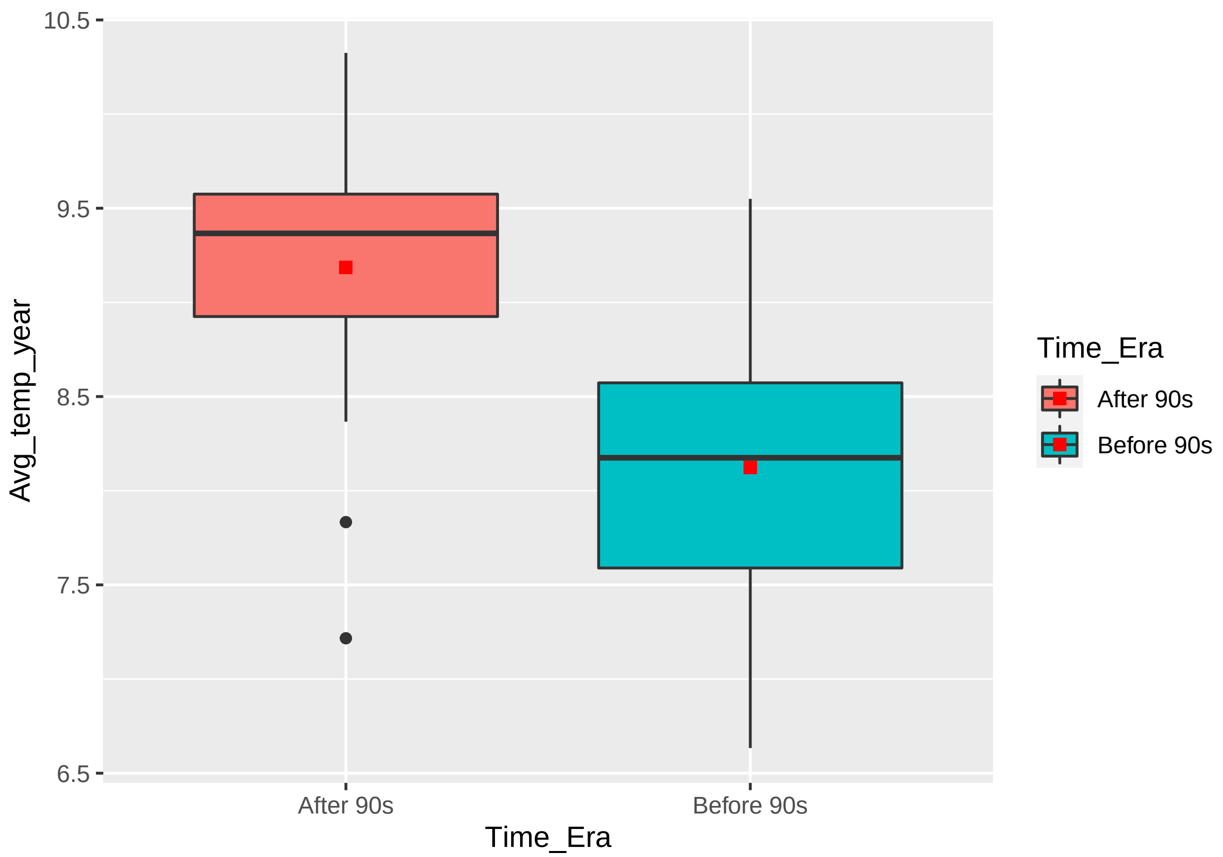
```
library(ggplot2)
sp = ggplot(germany_old_1, aes(y=Avg_temp_year, x=year)) +
        geom_point(aes(color = factor(Time_Era))) +
       theme_dark()
sp
```



The plot above really suggests that the average temperature in the years after the 90's shifted towards higher values. A boxplot will help us to have a better intuition on these numbers.

Box Plot of the data with means located

```
bp = ggplot(germany_old_1, aes(y=Avg_temp_year,
      x = Time_Era, fill = Time_Era)) + geom_boxplot() +
  stat_summary(fun.y = "mean", geom = "point",
  colour = "red", shape = 15, size = 2)

bp
```



The boxplot shows that the value at 25% of after 90's is greater than the one at 75% of the years from before 90's.

Value of Means

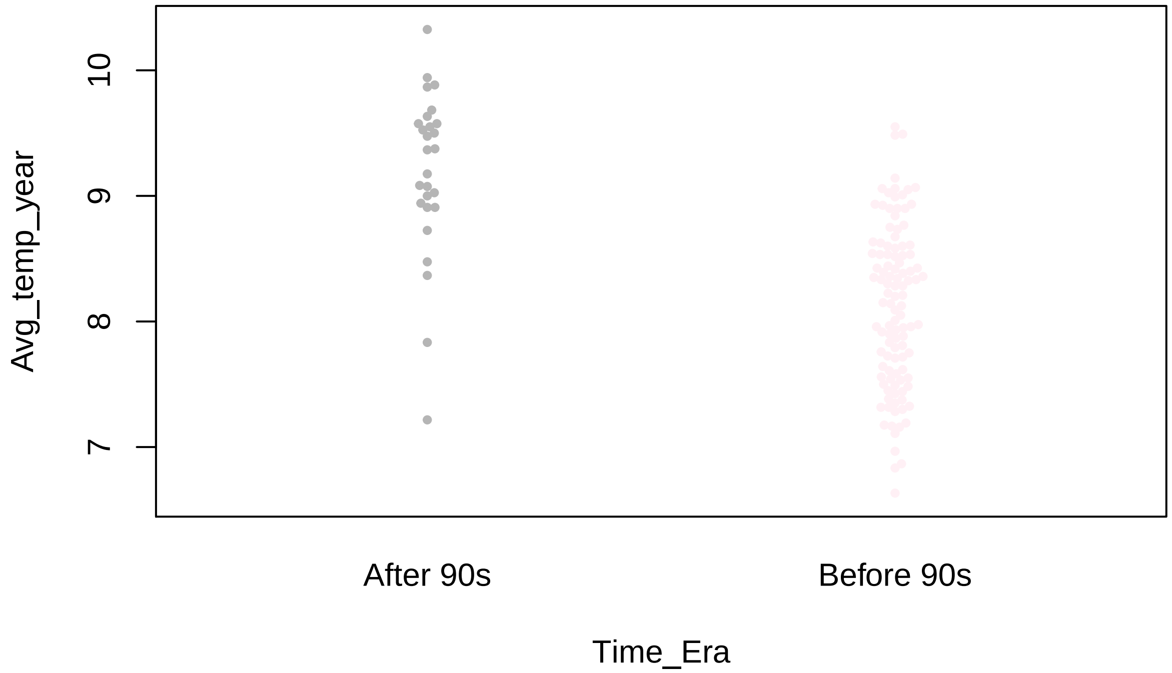
```
means=with(germany_old_1, by(Avg_temp_year, Time_Era, mean))
means
```

```
## Time_Era: After 90s
## [1] 9.185494
## ------------------------------------------------------------
## Time_Era: Before 90s
## [1] 8.121894
```

The average temperature rose by 1.06 C since the 90's. A rise of about 1 C is linked to huge consequences.

Beeswarm Plot - Visulazing the data points according to groups

```
library("beeswarm")
bs=beeswarm(Avg_temp_year~Time_Era, data=germany_old_1,
  col=sample(colors(),10), pch=19, cex=0.5, method="swarm")
```



What do you think about the distributions? Are they coming from the same source?

I think the problem now is well stated.

Hypothesis testing

Let us make the following Null hypothesis:

The difference of almost 1.06 C in the average temperatures between the two periods are due to random fluctuations in the temperature distribution, i.e., nothing has changed in the Germany's weather

Our work is to challenge this hypothesis, by calculating the probability of observing such a difference or higher, if nothing has changed in the weather of Germany.

If it is small enough, then we have evidence that Germany is really getting hotter and we reject the Null Hypothesis. If such a probability is high we can't claim any change based on this data.

We simulate a **null world** and to calculate the probabilities

First, the null world. In this world, nothing has changed since the 90's, so any of the temperatures of the two groups are possible during the whole period anytime.

To do that, We shuffle the temperatures between the two groups and relabel them because we assume the weather process was always the same for the last 138 years.

Testing difference in means using a permutation test

```
library(mosaic)
germany = germany_old_1[,c(3,2)]
summary(germany)
```

```
##    Time_Era         Avg_temp_year
##  Length:137         Min.   : 6.633
##  Class :character   1st Qu.: 7.725
##  Mode  :character   Median : 8.358
##                     Mean   : 8.332
##                     3rd Qu.: 8.908
##                     Max.   :10.325
```

```
germany$Time_Era = as.vector(germany$Time_Era)
```

Calculate the observed difference in means

```
observed = mean(Avg_temp_year~Time_Era, data = germany) %>%
  diff()
observed
```

```
## Before 90s
##    -1.0636
```

To simulate a single trial, we need to shuffle Avg_temp_year labels

```
mean(shuffle(Avg_temp_year)~Time_Era, data = germany) %>%
  diff()
```

```
## Before 90s
## 0.03001684
```

We will make permutations and relabeling our data 10000 times to plot the probability distribution of average temperature differences among our two groups (before and after 90's) in the so called null world.

Create randomization distribution

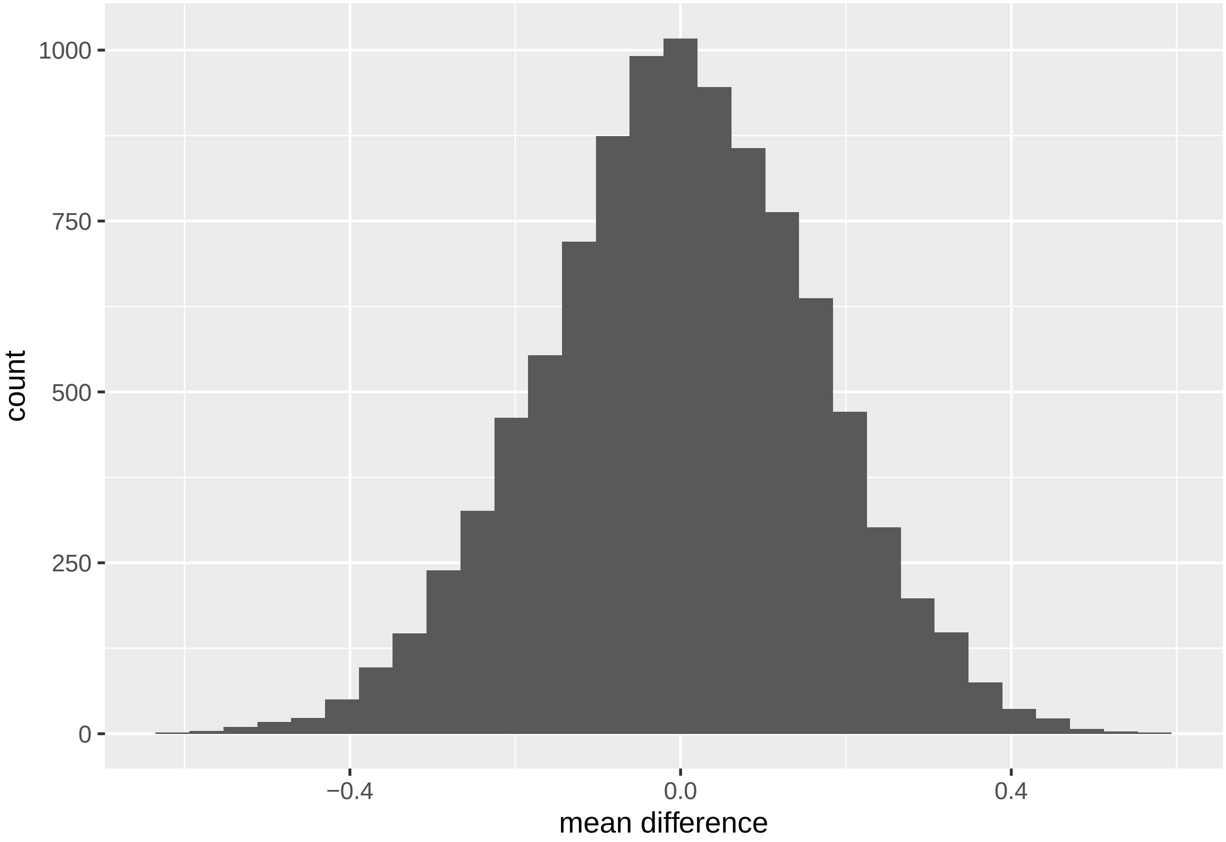
```
germany_null = do(10000)*mean(shuffle(Avg_temp_year)~Time_Era,data = germany) %>%
  diff()

head(germany_null)
```

```
##     Before.90s
## 1 -0.132584175
## 2  0.043470819
## 3 -0.043403479
## 4  0.173782267
## 5 -0.044172278
## 6 -0.006885522
```

Plotting the randomization distribution

```
ggplot(data = germany_null) +
      geom_histogram(mapping = aes(x=Before.90s)) +
  xlab("mean difference")
```

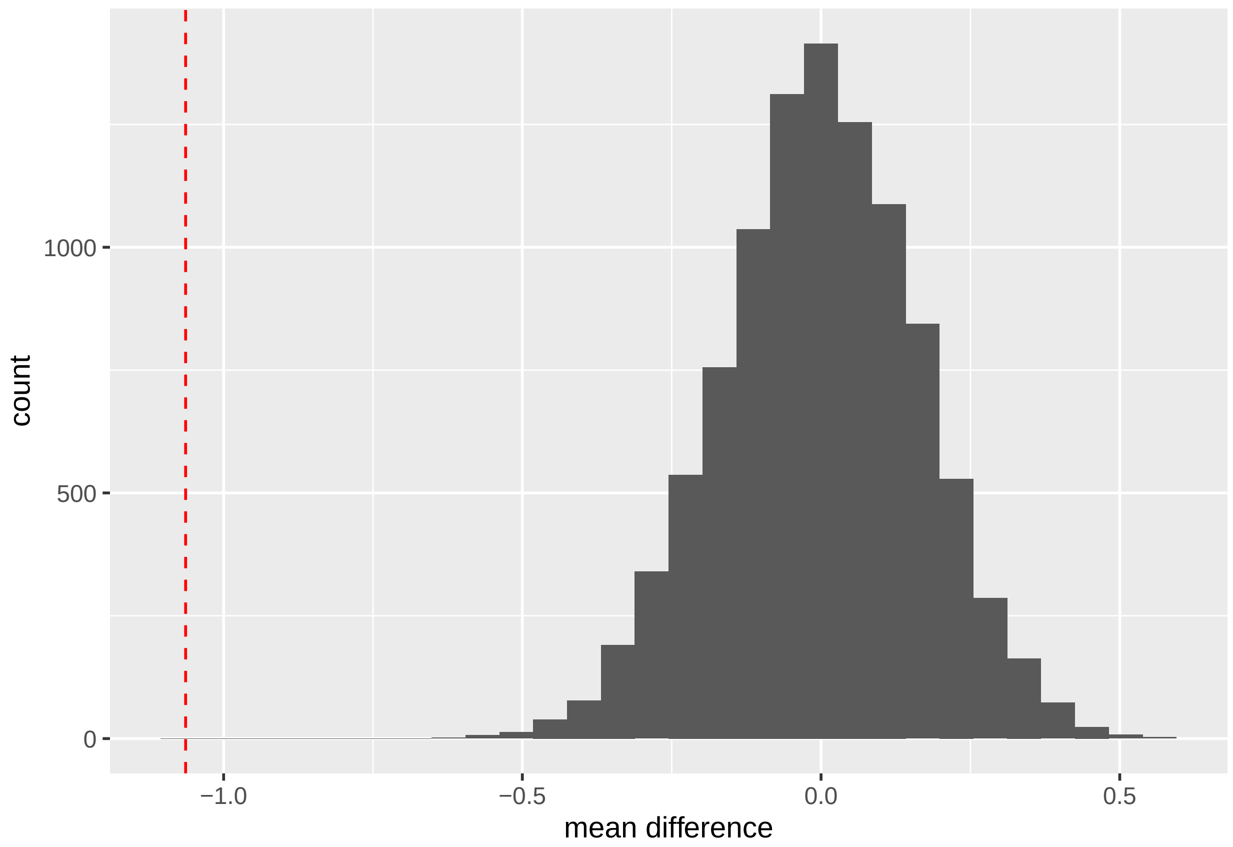
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



Superimpose a line indicating the observation

```r
ggplot(data = germany_null) +
  geom_histogram(mapping = aes(x=Before.90s)) +
  xlab("mean difference") +
  geom_vline(xintercept = observed, linetype = 2, color = "Red")
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



As we can see above, in the null world it would be **EXTREMELY RARE** to observe a difference like **1.06 C**

How rare?

Calculate the probability

```
prop(~Before.90s <= observed, data = germany_null)
```

```
## prop_TRUE
##         0
```

As expected by inspecting the plot, the p-value has turned out to be zero

With such a low p-value, we can confidently reject the null hypothesis.

Therefore, we conlcude that

Germany is really getting hotter. The observed temperature average difference (1.06 C) would not happen if the temperature before and after 90's were coming from the same distribution