Chapter 8: Use Case Notebook for Instructors

Ram Gopal, Dan Philps, and Tillman Weyde

2022

Contents

Load functions to compute p-value

```
p_rtail = function(sampdist,tstat)
  {
  temp = density(sampdist)
  df = data.frame(temp$x, temp$y)
  formula1 = df$temp.x<tstat
  df1 = df[formula1,]
  plot(df, col = "red", type = "h")
  points(df1, col = "green", type = "h")
  pvalue = length(sampdist[sampdist>tstat])/(length(sampdist))
  return(pvalue)
  }

p_ltail = function(sampdist,tstat)
  {
  temp = density(sampdist)
  df = data.frame(temp$x, temp$y)
  formula1 = df$temp.x>tstat
  df1 = df[formula1,]
  plot(df, col = "red", type = "h")
  points(df1, col = "green", type = "h")
  pvalue = length(sampdist[sampdist<tstat])/(length(sampdist))
  return(pvalue)
  }

p_2tail = function(sampdist,tstat)
  {
hyp = mean(sampdist)
cutoff1 = hyp - abs(tstat-hyp)
  cutoff2 = hyp + abs(tstat-hyp)
  temp = density(sampdist)
```

```
  df = data.frame(temp$x, temp$y)
  formula1 = df$temp.x<cutoff1 | df$temp.x>cutoff2
  df1 = df[formula1,]
  plot(df, col = "green", type = "h")
  points(df1, col = "red", type = "h")
  pvalue = length(sampdist[sampdist<cutoff1 |  sampdist>cutoff2])/(length(sampdist))
  return(pvalue)
  }
```

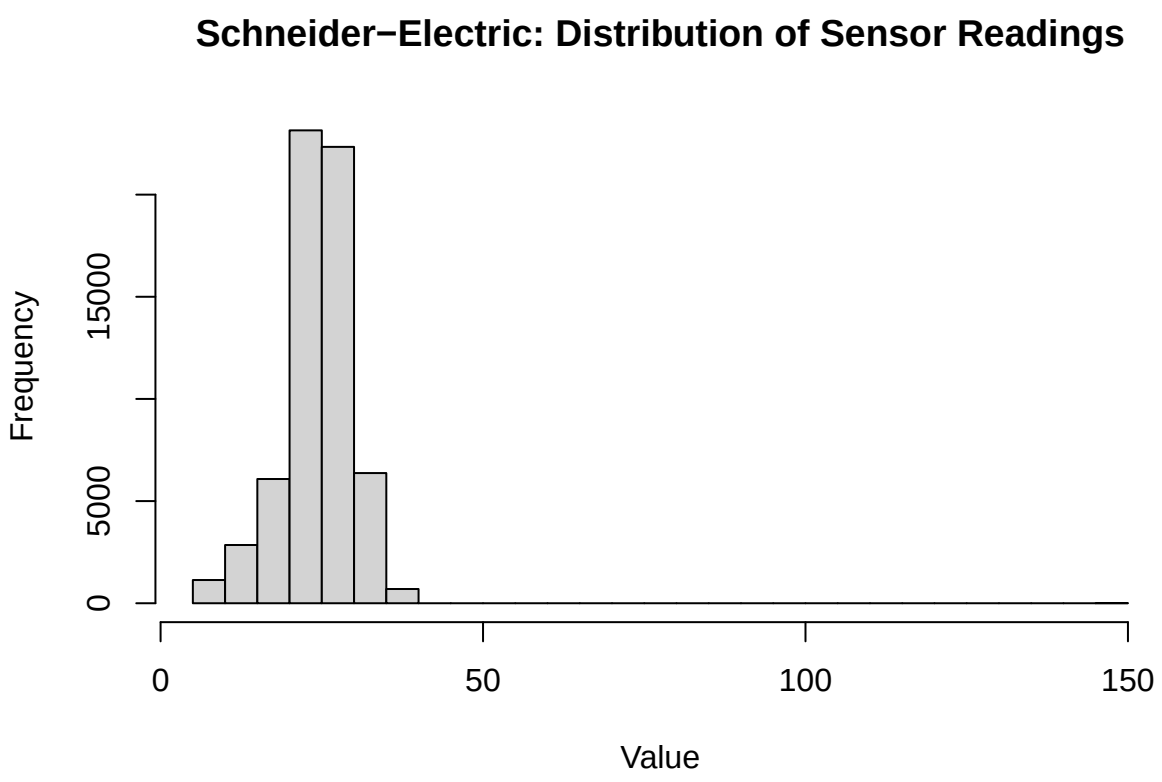
Use Case: Statistical Testing to Detect Cyber Attacks

The Schneider-Electric dataset shows records of temperature sensor readings, relevant to IoT devices. Readings outside of the usual indicate a problem. Exactly the same principles could apply to monitoring employee activities in the WFH (Work from home) era, or monitoring team members for excellent or weak productivity: we are looking for exceptions.

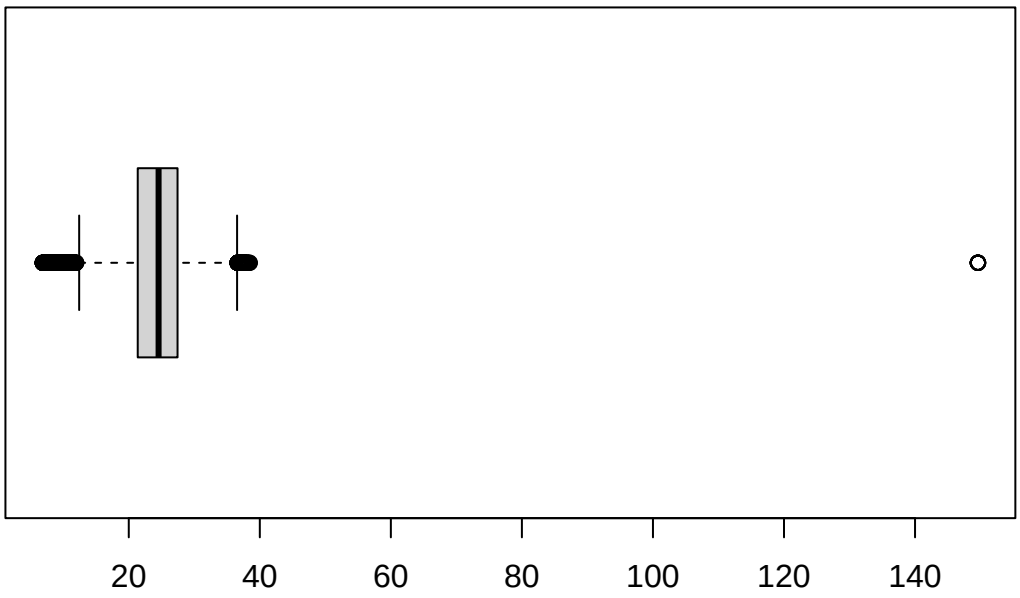
We will explore two questions: what are the bounds of normal operation of the Schneider sensors (i.e., confidence intervals)? What is the probability that a recent high sensor reading represents failure or cyber-attack (i.e., statistical testing)?

To spot exceptions, outliers, we can plot a frequency distribution of sensor readings taking over a number of months, using 50 bins and a column chart. By eye we can see that the distribution is approximately normal, but with a number of outliers:

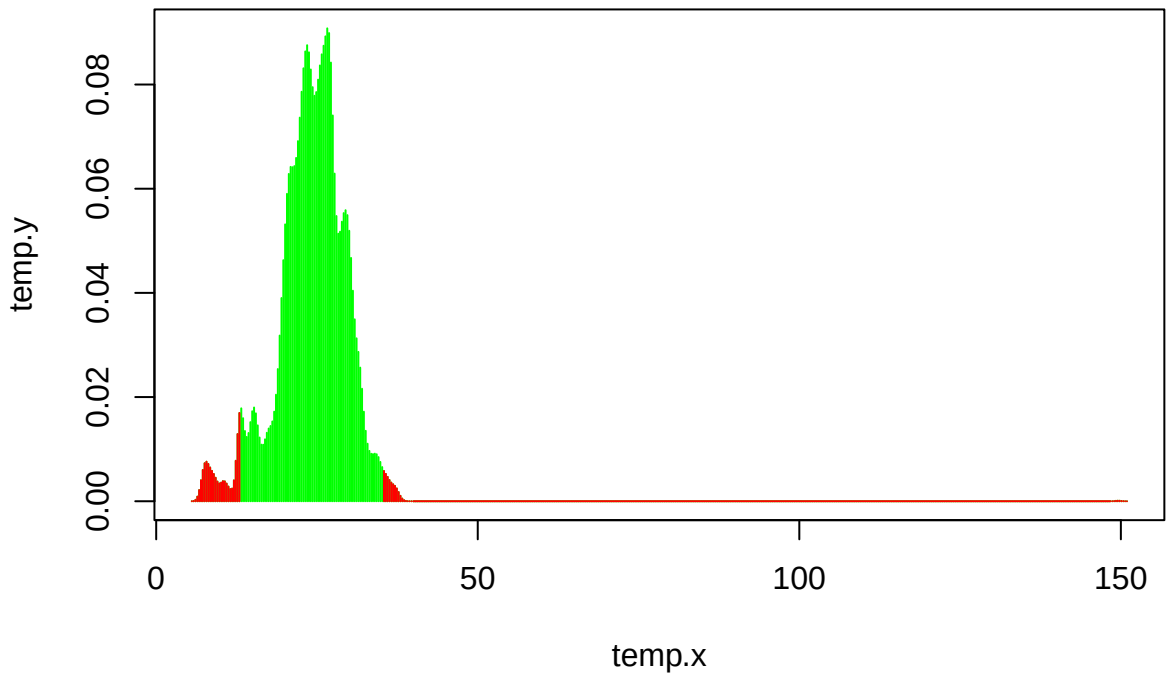
```
df <- read.csv("../../data/sensor-fault-detection.csv")
hist(df$Value,breaks=50,
     main = "Schneider-Electric: Distribution of Sensor Readings", xlab="Value")
```




```
boxplot(df$Value,horizontal = T)
```




```
p_2tail(df$Value,13)
```




```
## [1] 0.04193
```

In practice the sensor reading of 13 has a p-value of 0.0419 tells us that based on the assumptions of our test, there is only a 4.19% probability the null hypothesis should be not rejected. If we decided to have a 5% significance level, we would reject the null hypothesis, meaning that we would conclude that the sensor reading was not a normal operating reading and may represent a failure or cyber-attack.