Outliers? Don't Panic!

Philip He

9/7/2023

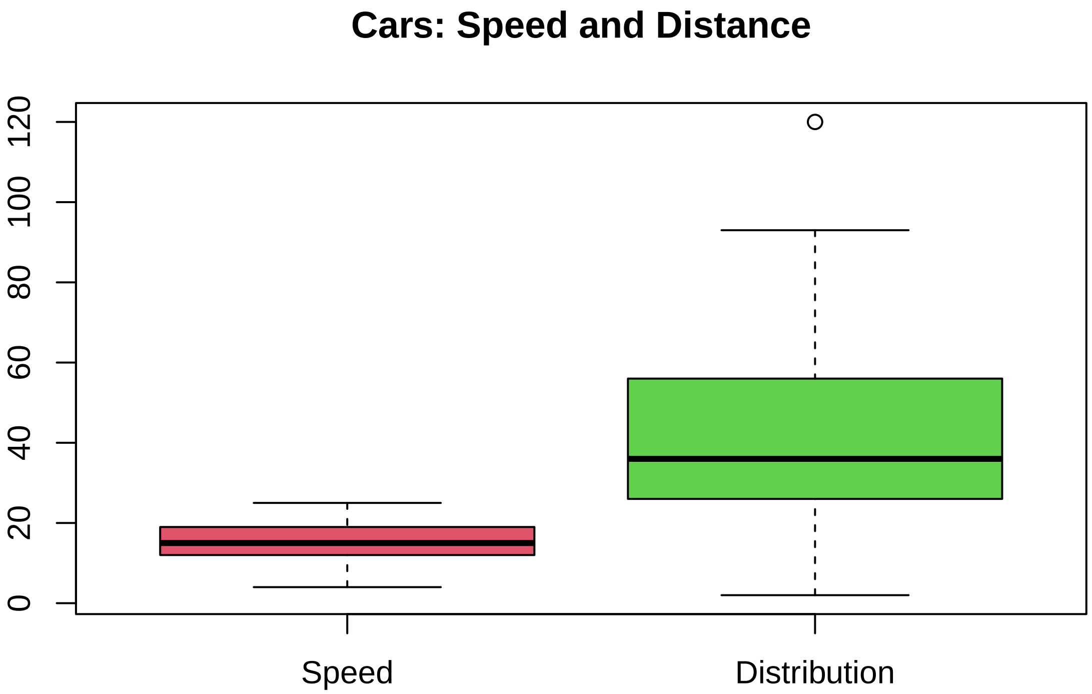
Outliers Revealed

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

The dataset `cars` has two variables `speed` and `dist` (distance). The descriptive summary is above. Let's visualize the distributions of both variables by boxplot.

```
boxplot(cars, col=2:3, main="Cars: Speed and Distance", names = c("Speed", "Distribution"))
```



It appears there is one outlier in `dist` variable whose value is greater than the upper fence, defined as $Q3 + 1.5 * IQR$. You might wonder why there is such outlier?

Outliers Neither Bad nor Good

Are outliers bad? Good? It depends on the cause if known. Albert Einstein's scientific contributions are probably outliers.

Detection of Outliers

A typical non-parametric approach to detect outliers is using boxplot where outliers are noted by the dots outside of the lower and upper fences, defined as $Q1 - 1.5 * IQR$ and $Q3 + 1.5 * IQR$, where $IQR = Q3 - Q1$.

To intuitively understand what this means, we look at the standard normal distribution $N(0, 1)$: $Q1 = -0.674$ and $Q3 = 0.674$. So $IQR = 1.35$, the upper fence is 2.698. In the unit of standard deviation σ, the upper fence is about 2.7σ. What is the odds to observe such outliers? The chance of a random draw from the standard normal distribution falling on the right of the upper fence is just 0.3%! It is 3 out of 1000.

```
Q1 <- qnorm(0.25)
Q3 <- qnorm(0.75)
IQR <- Q3 - Q1
Upper_fence <- Q3 + 1.5*IQR
Lower_fence <- Q1 - 1.5*IQR
chance_greater_than_upper_fence <- 1 - pnorm(Upper_fence)
Q1
```

```
## [1] -0.6744898
```

```
Q3
```

```
## [1] 0.6744898
```

```
IQR
```

```
## [1] 1.34898
```

```
Upper_fence
```

```
## [1] 2.697959
```

```
Lower_fence
```

```
## [1] -2.697959
```

```
chance_greater_than_upper_fence
```

```
## [1] 0.003488302
```

For multivariates, mahalanobis distance is usually used to detect outliers that incorporates the correlation between two variables. It is a useful tool to explore data when fitting a linear regression model. R has a built-in function to calculate the mahalanobis distance: `mahalanobis(x, center, cov)`.

Reasons of Outliers

Outliers represent low-likelihood data points. When outliers are observed, scientists should perform careful evaluation about the causes. Some possible causes include:

1. Measurement error

2. Incorrect unit

3. Human error, for example, wrong decimal place

4. Transmission error

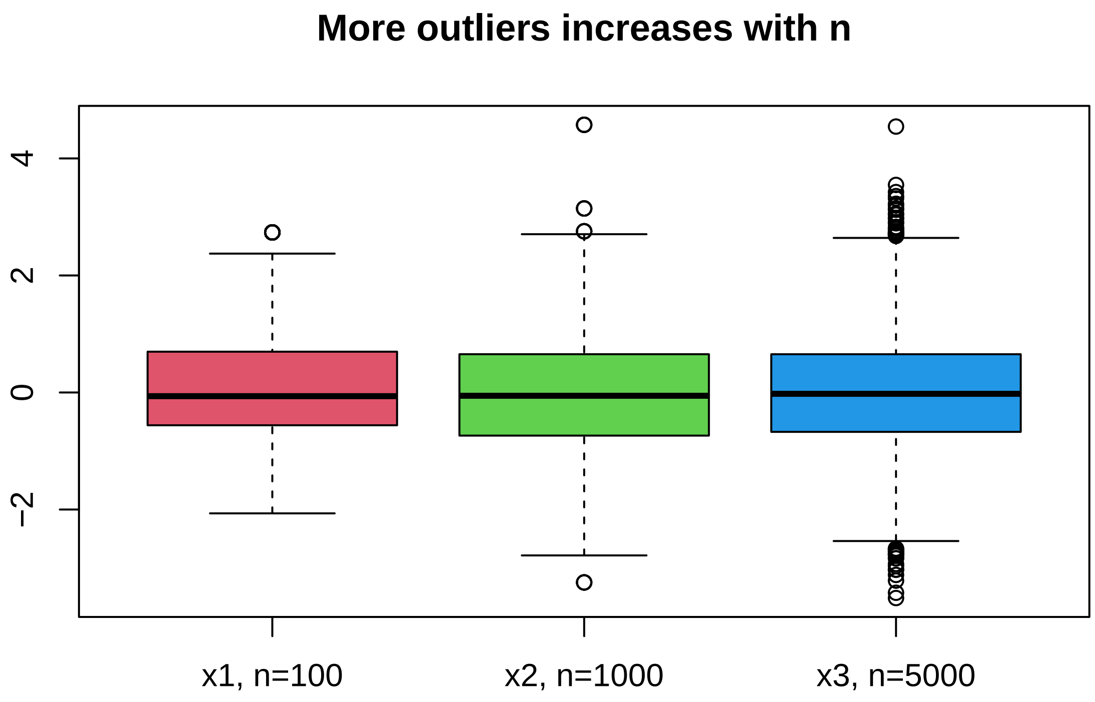
5. Missing data in data processing

6. Coding error in data processing

7. Incorrect specimen

8. When size of sample is large, outliers will be more frequent. This is because the probability of observing any outliers is increasing with n.

```r
set.seed(2023)
x1 = rnorm(100)
x2 = rnorm(1000)
x3 = rnorm(5000)
boxplot(cbind(x1, x2, x3), col=2:4,
        names=c("x1, n=100", "x2, n=1000", "x3, n=5000"),
        main="More outliers increases with n")
```


More outliers increases with n
x1, n=100
x2, n=1000
x3, n=5000