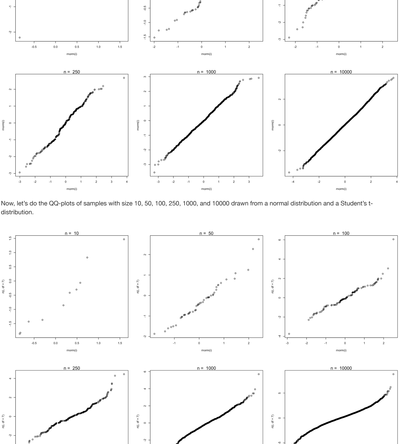
Assessment 1

Joana Levtcheva, CID 01252821

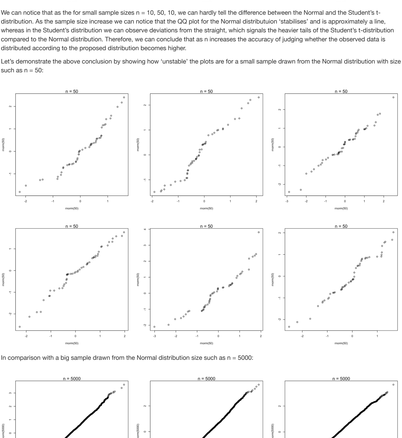
Question 1

Part a

First, let's do the QQ-plots of samples with size 10, 50, 100, 250, 1000, and 10000 both drawn from a normal distribution.

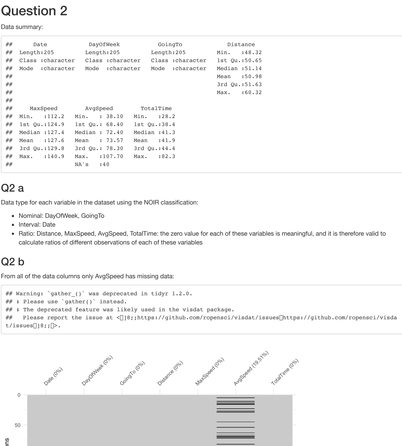


Now, let's do the QQ-plots of samples with size 10, 50, 100, 250, 1000, and 10000 drawn from a normal distribution and a Student's t-distribution.

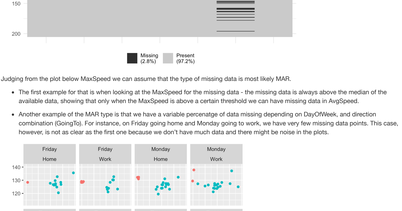


We can notice that as the for small sample sizes n = 10, 50, 10, we can hardly tell the difference between the Normal and the Student's t-distribution. As the sample size increase we can notice that the QQ plot for the Normal distribution 'stabilises' and is approximately a line, whereas in the Student's distribution we can observe deviations from the straight, which signals the heavier tails of the Student's t-distribution compared to the Normal distribution. Therefore, we can conclude that as n increases the accuracy of judging whether the observed data is distributed according to the proposed distribution becomes higher.

Let's demonstrate the above conclusion by showing how 'unstable' the plots are for a small sample drawn from the Normal distribution with size such as n = 50:



In comparison with a big sample drawn from the Normal distribution size such as n = 5000:



Part b

There is a greater empirical probability density in sample in the left tail, the corresponding quantiles for the empirical sample are higher for these quantiles below 50% than the corresponding quantiles of the theoretical distribution.On the opposite right tail, the values of the quantiles above 50% in the sample are less than the corresponding theoretical quantiles. This means the observed data has no skew and its distribution is platykurtic compared to the Gaussian distribution, meaning it has lighter tails and negative excess kurtosis.

Question 2

Data summary:

```
##       Date              DayOfWeek          GoingTo            Distance    
##  Length:205         Length:205         Length:205         Min.   :48.32  
##  Class :character   Class :character   Class :character   1st Qu.:50.65  
##  Mode  :character   Mode  :character   Mode  :character   Median :51.14  
##                                                           Mean   :50.98  
##                                                           3rd Qu.:51.63  
##                                                           Max.   :60.32  
##                                                                          
##    MaxSpeed        AvgSpeed        TotalTime     
##  Min.   :112.2   Min.   : 38.10   Min.   :28.2  
##  1st Qu.:124.9   1st Qu.: 68.40   1st Qu.:38.4  
##  Median :127.4   Median : 72.40   Median :41.3  
##  Mean   :127.6   Mean   : 73.57   Mean   :41.9  
##  3rd Qu.:129.8   3rd Qu.: 78.30   3rd Qu.:44.4  
##  Max.   :140.9   Max.   :107.70   Max.   :82.3  
##                  NA's   :40
```

Q2 a

Data type for each variable in the dataset using the NOIR classification:

- Nominal: DayOfWeek, GoingTo
- Interval: Date
- Ratio: Distance, MaxSpeed, AvgSpeed, TotalTime: the zero value for each of these variables is meaningful, and it is therefore valid to calculate ratios of different observations of each of these variables

Q2 b

From all of the data columns only AvgSpeed has missing data:

```
## Warning: `gather_()` was deprecated in tidyr 1.2.0.
## ℹ Please use `gather()` instead.
## ℹ The deprecated feature was likely used in the visdat package.
##   Please report the issue at <]8;;https://github.com/ropensci/visdat/issues]8;;https://github.com/ropensci/visdat/issues]8;;]>.
```


Judging from the plot below MaxSpeed we can assume that the type of missing data is most likely MAR.

- The first example for that is when looking at the MaxSpeed for the missing data - the missing data is always above the median of the available data, showing that only when the MaxSpeed is above a certain threshold we can have missing data in AvgSpeed.
- Another example of the MAR type is that we have a variable percenatge of data missing depending on DayOfWeek, and direction combination (GoingTo). For instance, on Friday going home and Monday going to work, we have very few missing data points. This case, however, is not as clear as the first one because we don't have much data and there might be noise in the plots.

To validate our conclusion we can compute the percentage of missing AvgSpeed values by DayOfWeek and GoingTo:

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## # A tibble: 10 × 5
## # Groups:   DayOfWeek, GoingTo [10]
##    DayOfWeek GoingTo variable n_miss pct_miss
##    <chr>     <chr>   <chr>      <int>    <dbl>
##  1 Friday    Home    AvgSpeed       1     7.69
##  2 Friday    Work    AvgSpeed       4    28.6 
##  3 Wednesday Home    AvgSpeed       3    13.0 
##  4 Wednesday Work    AvgSpeed       5    20.8 
##  5 Tuesday   Home    AvgSpeed       5    20.8 
##  6 Tuesday   Work    AvgSpeed       5    20.8 
##  7 Monday    Home    AvgSpeed       3    15   
##  8 Monday    Work    AvgSpeed       2    10.5 
##  9 Thursday  Home    AvgSpeed       6    30   
## 10 Thursday  Work    AvgSpeed       6    25
```

Part c

There could be many ways for imputing missing data. One as mentioned in the problem statement is just taking the sample mean for the entire column. A slightly more complex version of this method is grouping the data by certain qualities, for example day of week or dierction, and then fillinf the missing data with the mean of these groups.

However, in our case there is an obvious way to not just find a good way to impute the data but to fill the data with its actual value. This can be done by the following formula: Distance * 60 / TotalTime. We have that the distance is equal to velocity * time, in our case we notice that TotalTime is given in minutes and that's why we have the multiplication by 60.

This is an anylitical, exact approach, so we should not be concerned about how good and (un)biased our estimation is. We are working with not too small or too big numbers, and they are up to the first decimal point, so errors from computation are not expected to bias the calculation. It is good to check if we are not going to divide by 0:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    28.2    38.4    41.3    41.9    44.4    82.3
```

The minimum value of TotalTime is 28.2 so there should not be a problem.

Part d

Computing the modified Z-scores for the same data x and showing the data rows corresponding to the outliers:

```
## # A tibble: 4 × 7
##   Date       DayOfWeek GoingTo Distance MaxSpeed AvgSpeed TotalTime
##   <chr>      <chr>     <chr>      <dbl>    <dbl>    <dbl>     <dbl>
## 1 11/21/2011 Monday    Work        52.2     127.     38.1      82.3
## 2 8/18/2011  Thursday  Work        52.3     138.     NA        61.2
## 3 7/26/2011  Tuesday   Home        51.3     122.     43.7      70.5
## 4 7/18/2011  Monday    Work        54.5     126.     49.9      65.5
```