1. Practice – 2014

Average, Sample Standard Deviation, Histogram, Box-Plot

1. Exercise

The power of the cars trespassing the border station was noted.

- Calculated the average power and the sample standard deviation of the power.
- Create a histogram and compare it with the preceding values.
- Create a boxplot for the data.

Solution:

1. Calculate the *average*: **F3=AVERAGE(**
   - We need to give the array where the measurement are:
     - Left click on the first power, that is cell **B4**
     - With **CTRL+SHIFT+↓** we can select the whole array
     - Close the bracket **)** and press **ENTER**
     - With **CTRL+↑** we can jump up to the top of the page.
2. The *standard deviation* (denoted by s) is the square root of the average of the squared differences of the values from their average value: $s = \sqrt{\frac{1}{N}\sum(\bar{x} - x_i)^2}$.

   The *sample standard deviation* (denoted by s^*): $s^* = \sqrt{\frac{1}{N-1}\sum(\bar{x} - x_i)^2}$.

   First calculete the sample standard deviation using the formula:
   - We can only sum up the elements of an array, hence first we need to calculate the squared differences of the values from their average value in column C: **C4=(B4-F$3)^2**
     - Double click ont he right bottom corner of C4.
     - Remark: putting the $ sign before the column or row number we can fix it, so it won't change when we pull down the formula. Try the formula with and without the $ sign and see the difference!
   - Calculate N, how many elements are in column B?: **F4=COUNT(B4:B4197)**
     - Select the array the same way we did at the average!
   - Now we can calculate the sample standard deviation: **F5=SQRT(SUM(C4:C4197)/(F4-1))**

   Now use the Excel function to calculate it: **F6=STDEV.S(B4:B4197)**

   Remark 1: F5 must equal to F6!
   Remark 2: To calculate the standard deviation use STDEV.P(Array).

3. Make a *Histogram*!
   - Determine how many subintervals we want: **M5=ROUND(LOG(F4;2)+1;0)** = 13
     - If N<100 use SQRT(N) instead of LOG(N;2).
   - In L8:L20 make a list of the serial numbers: 1,2, 3..13.
   - Determine the boundaries of the intervals, so that each interval would contain about the same amount of elements. In our case there are 13 subintervals, so about 1/13th of the elements fall in each interval.
     PERCENTILE.INC(Array;k) is a number where the k-th of the Array is smaller than it.
     - Lower boundaries: **M8=PERCENTILE.INC(B$4:B$4197;(L-1)/13)**, and double click

    - Upper boundaries: M8=PERCENTILE.INC(B$4:B$4197;L/13), and double click
- Count how many elemnts are in each interval exactly, this is the frequencies of the intervals: O8=COUNIF(B$4:B$4197;"<"&N8)- COUNIF(B$4:B$4197;"<"&M8), and double click
    - With the formula the elements on the interval boundaries fall into the right hand side from the boundary. To consider the value on the last interval's upper boundary we need to allow equality there: O20=COUNIF(B$4:B$4197;"<="&N20)- COUNIF(B$4:B$4197;"<"&M20)
    - The sum of the frequencies must be N: O21=SUM(O8:O20) must give the value in F4!
- The area of the rectalgles should equal to the relative frequencies of the intervals. The heights of each rectangle can be calculated from the previous coulumns (Area/Width = (Frequency/N)/(Upper Boundary – Lower Boundary): P8=(O8/F$4)/(N8-M8)
- Now we have everything to draw the Histogram. We just need to create the x-y coordinates of the points that neded to be connected. For that a macro was written. To run the macro:
    - Click on the lower boundary of the first interval (M8)
    - Go to: Developer/Macros/HisztogramAdatElokeszito
- To make the diagram select the generated data in columns Q and R. Go to: Insert/Chart/Scatter/Scatter with Straight Lines
- To place the average in the diagram we need the coordinates of the points that needed to be connected:
    - For the X coordinates: L23=F3 and L24=F3
    - For the Y coordinates we need the height of the tallest rectangle and the minimum Y coordinate as well: M23=MIN(R8:R47) and M24=MAX(R8:R47)
- Right click on the diagram, then Select data/Add
    - Series name: Average
    - Series X value: Select L23:L24
    - Series Y value: Select M23:M24

4. Make a *Box-Plot*!
    - We need to find the median and the lower and upper quartiles. For this let's define the quartiles. The k-th quartile is the number where the k/4-th of the data is lower and the rest (4-k)/k-th is larger than it.
        - The 0^{th} quartile is the minimum,
        - the 1^{st} quartile is the lower quartile, Q1
        - the 2^{nd} quartile is the median, Q2
        - the 3^{rd} quartile is the upper quartile, Q3
        - the 4^{th} quartile is the maximum, Q4
    - Calculate them.
        - In cells E11:E15 write the number of the quartiles, 1, 2…4
        - Than F11=QUARTILE.INC(B$4:B$4197;E11), and double click
    - Excel can't make a Box-plot, we need to do it by ourselves. We will
        - make three rectangles with the heights Q1, Q2-Q1 and Q3-Q2
        - put them on top of each other
        - put a negative error bar on Q1 with the value Q1-Q0
        - put a positive error bar on Q3 with the value Q4-Q3

- Calculate the values in I11: I15
- Select I12:I14, then go to Insert/Bar/2-D bar/Stacked bar
- Right click on the diagram, then Select data, then click on Switch Row/Column, then OK
- Click on the rectangle on the bottom. To put a negative error bar:
    - Layout/Analysis/Error bars/More Error bar options
    - Display: minus
    - Error amount: custom
    - Leave the positive value as it is, in the negative value select Q1-Q0 in I11
- To put a positive error bar on the upper rectangle.
    - Display: plus
    - Error amount: custom
    - In the positive value select Q4-Q3 in I15, the negative value does not change
- Change the format of the bottom rectangle:
    - Left click on the bottom rectangle
    - Format data series/Fill/No fill
    - Format data series /Border color/No line
- We can see that 50% of the car powers fall between 71 and 100 hp. This region is called the “inter quartile region” (IQR)
- The scope of the upper 25% is much larger than of the lower 25%.