**Introductory Statistics**

This week’s sheet is about descriptive statistics and outliers.

**Problem 1: Calculating sample mean, sample variance and sample standard deviation**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 7 | 8 | 7 | 8 | 8 | 11 | 7 | 9 | 10 | 12 | 9 | 10 | 12 |
| 10 | 10 | 9 | 8 | 11 | 11 | 11 | 9 | 9 | 12 | 10 | 11 | 10 |

The table above gives the shoes sizes of 26 members of a football club.

**Tasks:**

* Enter the data into column A in an Excel sheet. Use the first rows for titles/descriptions of each column.
* In column B calculate the squares of each value.
* Calculate the sum of all x and all x2 and from this calculate **the sample mean, sample variance and sample standard deviation** once according to SCHAUM and once according to the lecture notes.
* Create a frequency table on the side.
* From the table, following the method above, calculate the **sample mean, sample variance and sample standard deviation** using the corresponding formula for the sums. The function SUMPRODUCT might be useful here. Make sure they agree with your first value each time.
* Finally use the original table to obtain **sample mean, sample variance and sample standard deviation** using the **Data - Data analysis** option selecting **Descriptive Statistics** and ticking **Summary statistics**. Use it also to find the 7th smallest and the 5th largest value for this data set.
* Compare your calculated values with those given by Excel.

**Questions**

* Which formula does Excel use for calculating the sample variance and sample standard deviation?
* What are the **median** and the **mode** of this sample?

**Problem 2: Identifying potential outliers**

* Open the data on hurricanes.
* Usinga **column chart** plot the data
* Using **Data – Data Analysis – Rank and Percentile** identify the **lower and upper quartile** and the **median** by interpolating between the corresponding kth statistics. Using the box and whisker method identify **potential outliers**. (Careful data is given in descending order!)
* Using **Data … Descriptive Statistics – Summary statistics** find mean and standard deviation.
* Use these values to calculate the z-values for all data point. Identify **potential outliers** again.
Recall the z-value is the number of standard deviations that a value is away from the mean. (Lecture 4)
* Copy the table twice, once deleting the top hurricane and once the top two hurricanes.
* Repeat the summary statistics and the analysis for outliers using z-values.

**Questions**

* How do **sample mean, sample standard deviation and quartiles** for the hurricane data change when the extreme value(s) are excluded?
* What do you observe for the recalculated **z-values** when excluding the extreme case(s)?
* Can you give reason for including/excluding the top two hurricanes?