Group 11:

Monika Mira Szamohval

Tamas Koncz

Lilla Tagai

**Q 6:**

Q: Make summary statistics (mean, median, variance, standard deviation). Do you think these statistics are representative of the population? Why? Why no?

The mean is sensitive to outliers, while the median isn’t. The median is not representative of the population if the data is skewed. The data is skewed if the mean is not in the middle.

In this dataset, the mean and the median are very close to each other**, therefore both values are representative of the dataset.**

Both standard deviation and variance are derived from the mean of a given data set.

Variance measures the average degree to which each point differs from the mean. The greater the variance, the larger the overall data range.

Standard deviation is simply the square root of the variance. The calculation of variance uses squares because it weights outliers more heavily than data very near the mean.

**In this exercise the variance is not representative** as it weights outliers more - it’s not meant to be representative of the dataset.

**The standard deviation is representative of the dataset**, since the dataset is not skewed and there are no big outliers.

> mean(df$price\_beer)

[1] 127

> mean(df$price\_dreher)

[1] 240

> median(df$price\_beer)

[1] 129

> median(df$price\_dreher)

[1] 239

> var(df$price\_beer)

[1] 684.4444

> var(df$price\_dreher)

[1] 210

> sd(df$price\_beer)

[1] 26.16189

> sd(df$price\_dreher)

[1] 14.49138

**Q7:**

Make histograms. Try different bins. What do you see when changing bin size?

Due to the small sample size the histogram of the dataset split into 4 and 5 buckets gives the same graph:

hist(df$price\_beer, breaks= 4, xlab = "Beer prices - cheapest in store")

hist(df$price\_beer, breaks= 5, xlab = "Beer prices - cheapest in store")



The dataset split into 10 buckets:

hist(df$price\_beer, breaks= 10, xlab = "Beer prices - cheapest in store")

The picture is very similar to previous ones. The reason for this is that the dataset is not skewed and most values are in the range 120-130 and 120-140.



Since the dataset consist of 10 rows, the histogram not split into bins gives the same result:



hist(df$price\_dreher, breaks= 4, xlab = "Dreher prices")

hist(df$price\_dreher, breaks= 5, xlab = "Dreher prices")

The second set of histograms shows the price allocation of Drehers in the same stores. Here the data is evenly distributed as well, with a slight left skewness.

hist(df$price\_dreher, breaks= 4, xlab = "Dreher prices")

hist(df$price\_dreher, breaks= 5, xlab = "Dreher prices")



The overall trends remain unchanged if we change the bin sizes, the reason for this is the small sample size.

hist(df$price\_dreher, breaks= 10, xlab = "Dreher prices")

A comparision of the two price\_types show that Dreher costs almost twice as the cheapest beer in the stores, the type Dreher has no outliers and the mean = median.

The cheapest bears has a wider range with more datapoints and the mean towards the lower end.

