Exercises Set 1 - Solution

1 Steel cable resistance

a) The mean is computed using the formula given in the course.

$$\langle x \rangle = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 11.31$$

Here, $n = 20$ is the number of samples.

To find the median and to draw the box plot, it's useful to sort the samples in ascending order.

7.1	9.2	9.3	10.1	10.1	10.5	10.8	11.1	11.2	11.3
11.5	11.6	11.8	12.2	12.2	12.4	12.6	13.3	13.7	14.2

To find the median, the 25% and 75% quartiles or any other quantiles, we first compute $n \cdot \alpha$, where α is the fraction associated to the quantile (e.g. $\alpha = 0.5$ for the median, $\alpha = 0.25$ for the first quartile).

Two things can happen: $n\alpha$ can be an integer, or not.

If it is an integer, both $x_{n\alpha}$ and $x_{n\alpha+1}$ fulfil the condition that a fraction α of the datapoints needs to smaller or equal to $\tilde{x}_\alpha$ and a fraction $1-\alpha$ of the datapoints needs to be larger or equal to $\tilde{x}_\alpha$.

$$\begin{aligned} \frac{n_{x_i \leq x_{n\alpha}}}{n} = \frac{n\alpha}{n} \geq \alpha \qquad &\text{and} \qquad \frac{n_{x_i \geq x_{n\alpha}}}{n} = \frac{n - n\alpha + 1}{n} \geq 1 - \alpha \\ \frac{n_{x_i \leq x_{n\alpha+1}}}{n} = \frac{n\alpha + 1}{n} \geq \alpha \qquad &\text{and} \qquad \frac{n_{x_i \geq x_{n\alpha+1}}}{n} = \frac{n - n\alpha}{n} \geq 1 - \alpha \\ \tilde{x}_\alpha \quad &= \quad \frac{1}{2}\left(x_{n\alpha} + x_{n\alpha+1}\right) \end{aligned}$$

If $n\alpha$ isn't an integer, only $x_{\lceil n\alpha \rceil}$ fulfils the conditions, where $\lceil a \rceil$ is the "ceiling" of a number, i.e. the first integer value above that (non-integer) number.

$$\begin{aligned} \frac{n_{x_i \leq x_{\lceil n\alpha \rceil}}}{n} = \frac{\lceil n\alpha \rceil}{n} \geq \alpha \qquad &\text{and} \qquad \frac{n_{x_i \geq \tilde{x}_{\lceil n\alpha \rfloor + 1}}}{n} = \frac{n - \lceil n\alpha \rceil + 1}{n} > \frac{n - n\alpha}{n} \geq 1 - \alpha \\ \tilde{x}_\alpha \quad &= \quad x_{\lceil n\alpha \rceil} \end{aligned}$$

In this exercise, $n\alpha$ is integer for the median, the 25% and 75% quartiles. So we find:

$$\tilde{x} \quad = \quad \frac{1}{2}\left(x_{10} + x_{11}\right) = \frac{1}{2}(11.3 + 11.5) = 11.4$$

Important: Note that there are a number of different ways to deal with the situation where a quantile does not exactly coincide with a specific point in the data set. For the median, it is the standard convention to take the middle value for an odd number of data points, and the average of the two most central values for an even number of data points. However, for other quantiles there are many

different ways to interpolate - the approach presented here is only one possibility. Importantly, the larger the data-set (i.e. the smaller the typical gaps between neighbouring data points), the smaller the difference between different conventions.

b) For the box plot, we start by computing the quartiles:

$$\begin{aligned}\tilde{x}_{0.25} &= \frac{1}{2}(x_5 + x_6) = \frac{1}{2}(10.1 + 10.5) = 10.3\\ \tilde{x}_{0.75} &= \frac{1}{2}(x_{15} + x_{16}) = \frac{1}{2}(12.2 + 12.4) = 12.3\end{aligned}$$

The mean and the median are close together, this is due to a symmetric distribution.

The maximal and minimal values for a sample not to be considered as outlier, are:

$$\tilde{x}_{0.75} + 1.5(\tilde{x}_{0.75} - \tilde{x}_{0.25}) = 15.3 \qquad \text{and} \qquad \tilde{x}_{0.25} - 1.5(\tilde{x}_{0.75} - \tilde{x}_{0.25}) = 7.3$$

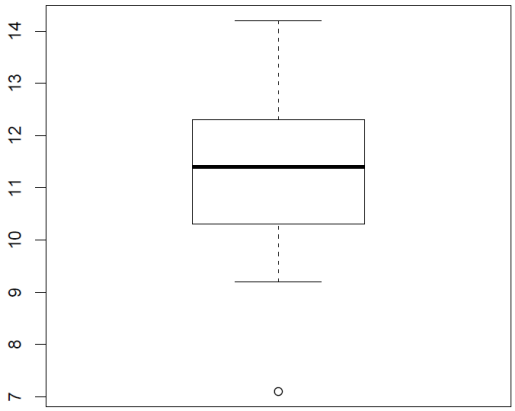

Figure 1: Box plot. The box contains all values between $\tilde{x}_{0.25}$ and $\tilde{x}_{0.75}$. The dashed line goes from the smallest to the biggest values, which aren't outliers. The small circle indicates the outliers' position.

c) After removing the only aberrant value ($x_1 = 7.1$), we get $n = 19$, and:

$$\begin{aligned}\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = 11.53\\ \tilde{x} &= x_{\lceil 0.5n \rceil} = x_{10} = 11.5\\ \tilde{x}_{0.25} &= x_{\lceil 0.25n \rceil} = x_5 = 10.5\\ \tilde{x}_{0.75} &= x_{\lceil 0.75n \rceil} = x_{15} = 12.4\end{aligned}$$

The mean and the median increase because a small value is removed. The box size decreases because we remove a value far from the mean.

d) Keeping the 19 samples (as we had an "external" reason to remove one data point), naively, we would have to find the value where at least 90% of all samples are stronger or equal than X. This means that only 10% of samples can be weaker or equal, hence we would compute $\tilde{x}_{0.1} = x_2 = 9.3$. Similarly, for 95% we obtain $\tilde{x}_{0.05} = x_1 = 9.2$.

Clearly the last case, which simply corresponds to the weakest value measured, is just based on a single data point, and $\tilde{x}_{0.1}$ only on two data points. Given that there is a significant spread in our measurements (the IQR is on the order of 20% of the median), it would be irresponsible to claim anything about a 90% or even 95% limit - especially because we are talking abou elevator cables here.

We will discuss this in more detail in future lectures, but essentially making assumptions about "all steel cables" from measurements on n steel cables requires larger and larger data sets the more we want to look at the fringes of the distribution (i.e. very high or low percentiles).

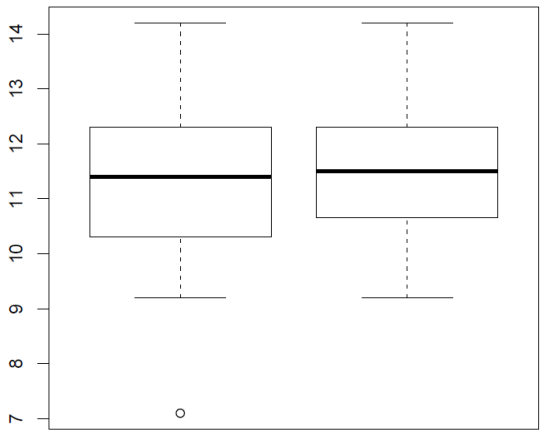

Figure 2: Box plots with all samples (on the left) and without the aberrant values (on the right).

2 Last year MX grades at the statistics exam

The computations are similar as in the first exercise with $n = 22$.

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = 4.35 \\ \tilde{x} &= \frac{1}{2}\left(x_{11} + x_{12}\right) = 4.25 \\ \tilde{x}_{0.25} &= x_{\lceil 0.25n \rceil} = x_6 = 4.00 \\ \tilde{x}_{0.75} &= x_{\lceil 0.75n \rceil} = x_{17} = 4.75 \end{aligned}$$

As before, the similarity between the mean and the median is due to a symmetric distribution.

Outliers will be greater than $\tilde{x}_{0.75} + 1.5(\tilde{x}_{0.75} - \tilde{x}_{0.25}) = 5.875$ or smaller than $\tilde{x}_{0.25} - 1.5(\tilde{x}_{0.75} - \tilde{x}_{0.25}) = 2.875$. Two values are concerned, but it doesn't make sense to remove them, because these grades are possible and can't be considered as errors like in the first exercise.

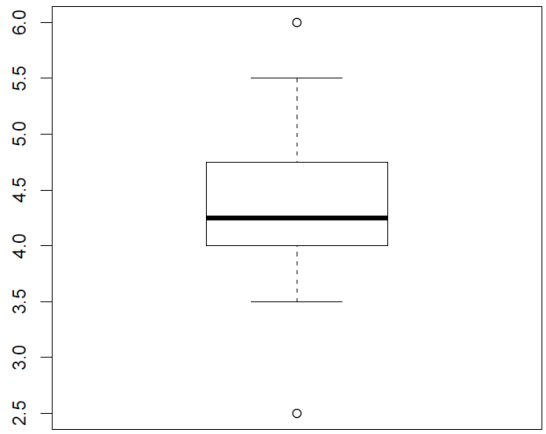

Figure 3: Box plot for the grades.

The cumulative distribution of this dataset is:

To pass the exam, a student must do at least 4:

$$\frac{n_{x_i \geq 4}}{n} = 1 - \frac{n_{x_i < 4}}{n} = 1 - \frac{n_{x_i \leq 3.75}}{n} = 1 - F(3.75) = \frac{9}{11} = 81.8\%$$

Outcome	...	2.5	...	3.5	3.75	4	4.25	4.5	4.75	5	5.25	5.5	5.75	6
Probability	0	$\frac{1}{22}$	0	$\frac{1}{11}$	$\frac{1}{22}$	$\frac{5}{22}$	$\frac{3}{22}$	$\frac{3}{22}$	$\frac{3}{22}$	$\frac{1}{22}$	0	$\frac{1}{11}$	0	$\frac{1}{22}$
CDF $F(x)$	0	$\frac{1}{22}$	$\frac{1}{22}$	$\frac{3}{22}$	$\frac{2}{11}$	$\frac{9}{22}$	$\frac{6}{11}$	$\frac{15}{22}$	$\frac{9}{11}$	$\frac{19}{22}$	$\frac{19}{22}$	$\frac{21}{22}$	$\frac{21}{22}$	1

Similarly:

$$\frac{n_{x_i \geq 5.25}}{n} = 1 - \frac{n_{x_i < 5.25}}{n} = 1 - \frac{n_{x_i \leq 5}}{n} = 1 - F(5) = \frac{3}{22} = 13.6\%$$

3 On wealth inequalities in the world

a) First, we need to inspect the data (e.g. using a text editor). The first column contains annual incomes in CHF. The second contains quantiles. To find the median, we simply need to find the annual income which corresponds to the 0.50 quantile. This can be done directly in the text editor, or using the `np.where(x==0.5)` function to find the row corresponding to the 0.5 quantile, and then finding the value of the first column at that index. We get

$$\tilde{x} = 3502$$

Finding the mean is more complex: Not all steps have the same distance (initially we have steps of 0.01, but towards the end we have steps of 0.001, and the last step is 0.002), that means not all steps correspond to the same fraction of people. Therefore we will need to use a weighted average. For the weight, we can use the difference between adjacent quantiles - so we will get 0.01 for the first 99 rows (rows 0 to 98 in Python), 0.001 for the next 8 rows, and 0.002 for the last row. As a check, we can see that these weights (which correspond to fractions), sum up to one.

Finally, the last row gives us the 100% quantile - i.e. the annual income that is small or equal to the income of all humans on earth. This means it corresponds to the highest income in the world. A staggering 11.5 billion. Clearly, this is not the income of 0.002 of all humans (that would still be about 16 million people) - and if we would make that assumption we would very strongly skew the mean (0.002 of all people earning 11.5 billion would already give a mean of 23 million, even if everyone else would earn nothing). On the other hand, if only about 10 people have an annual income of about 10 billion, this only contributes about 12.5CHF to the mean, so a handful of individuals, even with such staggering income, still does not affect the mean too much, so we can leave them out.

This means we have to make an educated guess about the highest 0.002 fraction. One possibility is to look at the increase between the 0.997 and the 0.998 quantile, which is 336217. If we expect this trend to continue, after another 0.002, we would arrive at 1976434. Hence we decide to replace the last income value by 1976434, knowing that this.

Finally, the EDCF gives us the values *below* (or exactly at) which a certain quantile lies. That is the lowest 0.01 have an income 87CHF p.a. *or less*. Hence, for the weighted income we can use the average between two limits (and taking 0 as the lower limit).

With all these operations, we arrive at a mean of 15814CHF. Clearly the precision given here is not appropriate, given the approximations we have (for example, without the last step, we would get more than 17kCHF, if we would have simply ignored the highest quantile, we would arrive at close to 13kCHF) This is about 4.5 times larger than the median - the distribution is extremely asymmetric.

b) For a box plot, we need the 0.25 and 0.75 quantiles, which can simply be read off. We restrict the plotting range to 80000, as otherwise the extreme outliers would make the plot unreadable. The mean is clearly not *representative* in any meaningful sense of the word. But the median too is of limited value - unlike in exercise 1, one cannot say that most values are somewhat close to the median, because the IQR is actually larger than the median.

c) There is no reason to remove the outliers - they are not measurement mistakes.

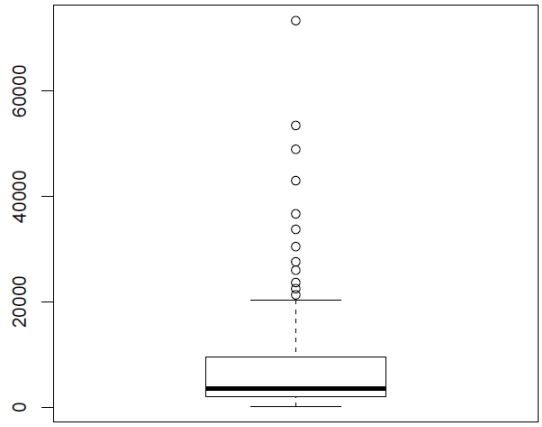

Figure 4: Box plot. Note that many outliers above 80000 exist. Also note that we are creating a box plot based on cumulated data. If we actually had the data for every human, the number of points corresponding to outliers would be many millions. Each point here summarises a large group of people (with a range of incomes) and a non-discrete representation would be more suited.

d) The total annual income is the number of humans (N), times the mean. We do not need to know N, as all our values are expressed as fractions. To find the fraction of wealthiest humans required to assemble half of the global income, we use the same weighted sum as above, but increase the endpoint of the sum until the result crosses 50%, i.e. 7906CHF. This happens somewhere between the 0.97 and 0.98 quantile, i.e. between 2 and 3% of the population accumulate 50% of the world's annual income. To find this value for 10% we would already need better data - we would have to search for the point where the weighted sum reaches 14232 CHF (and, as pointed out above, we could well be off by a few thousand here), and we do not have enough resolution in our data for this.

e) You would have 18000 CHF/year, you are between the quantiles $\tilde{x}_{0.83}$ and $\tilde{x}_{0.84}$. So between 83 and 84% of people in the world have a lower income.

f) Your average yearly salary would be 79926 CHF and your quantile a bit above 99.6%