Exercise Set 5 - Solution

1 Steel cable resistance

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

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

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

To find the median (and other quantiles) 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 $\tilde{x}_\alpha$). So we take $\tilde{x}_\alpha$ to be the average of the two values.

$$\frac{n_{x_i \le x_{n\alpha}}}{n} = \frac{n\alpha}{n} \ge \alpha \quad \text{and} \quad \frac{n_{x_i \ge x_{n\alpha}}}{n} = \frac{n - n\alpha + 1}{n} \ge 1 - \alpha$$
$$\frac{n_{x_i \le x_{n\alpha+1}}}{n} = \frac{n\alpha + 1}{n} \ge \alpha \quad \text{and} \quad \frac{n_{x_i \ge x_{n\alpha+1}}}{n} = \frac{n - n\alpha}{n} \ge 1 - \alpha$$
$$\tilde{x}_\alpha = \frac{1}{2}\left(x_{n\alpha} + x_{n\alpha+1}\right)$$

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.

$$\frac{n_{x_i \le x_{\lceil n\alpha \rceil}}}{n} = \frac{\lceil n\alpha \rceil}{n} \ge \alpha \quad \text{and} \quad \frac{n_{x_i \ge \tilde{x}_{\lceil n\alpha \rfloor + 1}}}{n} = \frac{n - \lceil n\alpha \rceil + 1}{n} > \frac{n - n\alpha}{n} \ge 1 - \alpha$$
$$\tilde{x}_\alpha = x_{\lceil n\alpha \rceil}$$

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

$$\tilde{x} = \tilde{x}_{0.5} = \frac{1}{2}\left(x_{10} + x_{11}\right) = \frac{1}{2}(11.3 + 11.5) = 11.4$$
$$\tilde{x}_{0.25} = \frac{1}{2}\left(x_5 + x_6\right) = \frac{1}{2}(10.1 + 10.5) = 10.3$$
$$\tilde{x}_{0.75} = \frac{1}{2}\left(x_{15} + x_{16}\right) = \frac{1}{2}(12.2 + 12.4) = 12.3$$

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

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)

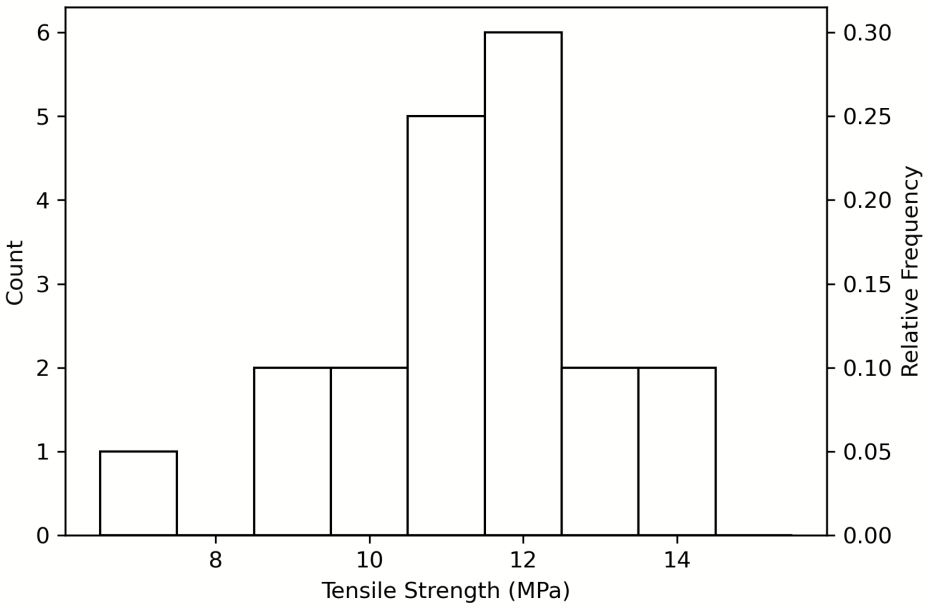


Figure 1: Tensile strength histogram

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.

d) Keeping only 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 about 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).

e) The variance is 2.79MPa^2, and the standard deviation is 1.67MPa. The mean is 11.31MPa, thus the interval in the exercise is [9.64,12.98]MPa. Counting gives a total of 14 out of 20 points in this interval, a fraction $p = \frac{14}{20} = 70\%$.

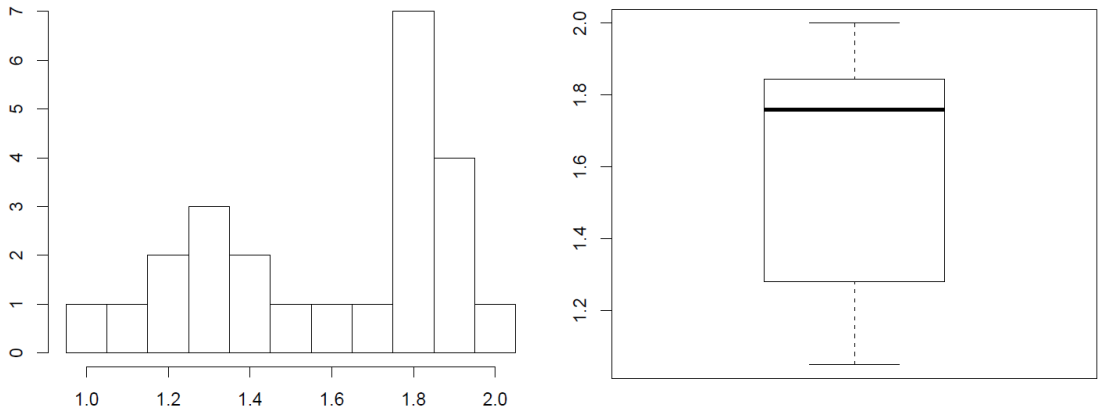
2 Molecular Beam Epitaxy (MBE), Normal law and precision

a) For the first dataset, the mean, the median and the standard deviation are:

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i} = 1.604 \qquad \tilde{x}_1 = \frac{1}{2}(1.71 + 1.81) = 1.76 \qquad \tilde{s}_1 = \sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2} = 0.29$$

Since the median isn't close to the mean, the distribution is probably asymmetric. The histogram drawn below confirms that and clearly shows 2 populations. Because of these, the difference between $\tilde{x}_{0.25}$ and $\tilde{x}_{0.75}$ is big, so there aren't any outliers.

b)



c) In the dataset, 14 samples over 24 (58.3%) are within the intervals $[\bar{x}_1 \pm \tilde{s}_1] = [1.31; 1.90]$ and every one (100%) are within $[\bar{x}_1 \pm 2 \cdot \tilde{s}_1] = [1.01; 2.19]$.

d) Using the normal law we get:

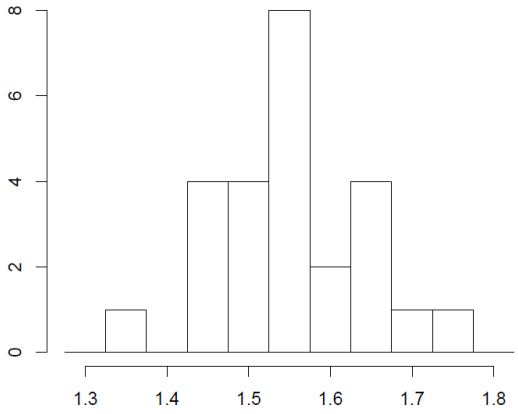
$$\begin{aligned} P(\mu - \sigma \leq X \leq \mu + \sigma) &= P(X \leq \mu + \sigma) - P(X < \mu - \sigma) \\ &= P\left(Z \leq \frac{\mu + \sigma - \mu}{\sigma}\right) - P\left(Z \leq \frac{\mu - \sigma - \mu}{\sigma}\right) \\ &= \Phi(1) - \Phi(-1) = \Phi(1) - (1 - \Phi(1)) = 0.682 \\ P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &= \Phi(2) - \Phi(-2) = 2 \cdot \Phi(2) - 1 = 0.954 \end{aligned}$$

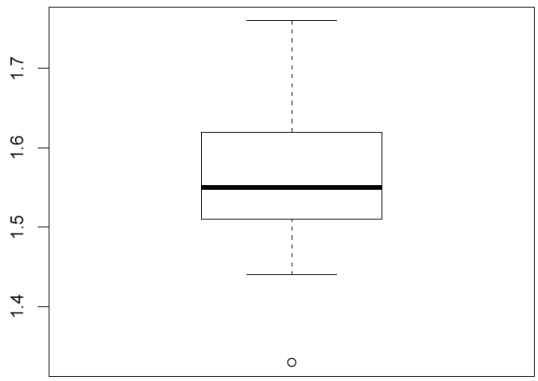
The distribution is too asymmetric and too widespread to be well represented by a Gaussian. This is also clear when looking at the first histogram or the corresponding box plot.

f) The mean, the median and the standard deviation are:

$$\bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i} = 1.556 \qquad \tilde{x}_2 = 1.55 \qquad s_2 = \sqrt{\frac{1}{n_2}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2} = 0.092$$

For the second dataset $\bar{x}_2 \simeq \tilde{x}_2$, so the distribution can be symmetric. The box plot shows that 1.33 is an outlier, and the histogram shows there is only one population.





72% of the samples are within the interval $[\bar{x}_2 \pm s_2] = [1.46; 1.65]$ and 92% are in $[\bar{x}_2 \pm 2 \cdot s_2] = [1.37; 1.74]$. These proportions look more like the Gaussian ones than those form the first dataset do. This time the distribution can be represented by a Gaussian.

The quality increases because the data now follow the Gaussian law and the standard deviation is reduced.

f) To use the Gaussian distribution, we have to estimate

$$\mu = \bar{x}_2 = 1.556 \qquad \text{and} \qquad \sigma = \sqrt{\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (x_{2i} - \mu)^2} = 0.094$$

The probability a chips lie within 1.45 and 1.55 is

$$\begin{aligned} P(1.45 \leq X \leq 1.55) &= P(X \leq 1.55) - P(X < 1.45) \\ &= P\left(Z \leq \frac{1.55 - \mu}{\sigma}\right) - P\left(Z < \frac{1.45 - \mu}{\sigma}\right) \\ &= \Phi(-0.06) - \Phi(-1.12) \\ &= 1 - \Phi(0.06) - (1 - \Phi(1.12)) \\ &= \Phi(1.12) - \Phi(0.06) = 0.345 \end{aligned}$$

Since 0.345 is smaller than 0.95, Intel won't start a partnership with the lab. Of course, even if the criterion had been fulfilled, but only when one specific technician operates the device, it would mean that the method is not very reproducible - one should really take the distribution that arises from a multitude of operators.

h) Next we search a confidence interval. We look for $x_{0.05}, x_{0.95}$ such that $P(X \leq x_{0.05}) = \Phi(z_1) = 5\%$, and $P(X \leq x_{0.95}) = \Phi(z_2) = 95\%$. z_2 can be found on a distribution table. To find z_1, we have to remember the centred and reduced Normal law is symmetric around 0.

$$\begin{aligned} z_1 = -1.65 = \frac{x_{0.05} - \mu}{\sigma} \qquad &=> \qquad x_{0.05} = z_1 \cdot \sigma + \mu = 1.40 \\ z_2 = 1.65 = \frac{x_{0.95} - \mu}{\sigma} \qquad &=> \qquad x_{0.95} = z_2 \cdot \sigma + \mu = 1.71 \end{aligned}$$

3 Maxwell-Boltzmann law and the particle speeds in an ideal gas

a) The speed of the particle v can be any positive real number. As before, the sum of the probability for all possible values should give 1.

$$\int_0^\infty P(V = v) dv = \int_0^\infty C \cdot v^2 e^{-\frac{m \cdot v^2}{2kT}} dv = -C \cdot \left[v \cdot \frac{kT}{m} e^{-\frac{m \cdot v^2}{2kT}} \right]_0^\infty + \int_0^\infty \frac{CkT}{m} \cdot e^{-\frac{m \cdot v^2}{2kT}} dv$$

Here we have used "integration by parts" (i.e. $\int fg' = fg - \int f'g$). The first term vanishes since the exponential function decreases much faster than x increases (try typing $10 \cdot \exp[-10]$ and $100 \cdot \exp[-100]$ into your calculator!).

The second term can be found from the scaled Gaussian integral (see last week's exercise) and taking into account the fact that the Gaussian function is symmetric around 0 - so the integral from 0 to ∞ is just half of the total integral.

It follows:

$$\int_0^\infty P(V=v)dv = \int_0^\infty \frac{CkT}{m} \cdot e^{-\frac{m \cdot v^2}{2kT}} dv = \frac{CkT}{m} \cdot \sqrt{\pi \frac{2kT}{4m}} = 1 \qquad => \qquad C = \frac{\pi}{2} \left(\frac{2m}{\pi kT}\right)^{3/2}$$

b) The average speed is given by. Again we start by integrating by parts, and for the last integral we look at the derivative of $\exp[-ax^2]$ and work backwards.:

$$\begin{aligned}
\mathbb{E}(V) &= \int_0^\infty C \cdot v^3 \cdot e^{-\frac{m \cdot v^2}{2kT}} dv \\
&= -C \cdot \left[v^2 \cdot \frac{kT}{m} e^{-\frac{m \cdot v^2}{2kT}} \right]_0^\infty + \frac{CkT}{m} \cdot \int_0^\infty 2v \cdot e^{-\frac{m \cdot v^2}{2kT}} \\
&= \frac{CkT}{m} \cdot \left[\frac{-2kT}{m} e^{-\frac{m \cdot v^2}{2kT}} \right]_0^\infty \\
&= 2C \cdot \frac{(kT)^2}{m^2} = \sqrt{\frac{8kT}{\pi m}} = 473 \text{ m/s}
\end{aligned}$$