Exercises Set 4 - Solution

1 Expectation values and variances

a) $\mathbb{E}(10X_1 + 2X_2) = 10 \cdot \mathbb{E}(X_1) + 2 \cdot \mathbb{E}(X_2) = 10\mu + 2\mu = 12\mu$

b) $\mathbb{E}\left(\frac{X_1+X_2+X_3+...+X_n}{n}\right) = \frac{1}{n} \cdot (\mathbb{E}(X_1) + \mathbb{E}(X_2) + ... + \mathbb{E}(X_n)) = \frac{1}{n} \cdot n\mu = \mu$
Because the X_i are independent, there is no Covariance, and the variance of a sum is just the sum of the variances. For completenexx, we do however state the covariance in the formulae below (it is then just =0).

c) $\mathrm{Var}(10X_1 + 2X_2) = 10^2 \cdot \mathrm{Var}(X_1) + 2^2 \cdot \mathrm{Var}(X_2) + 2 \cdot 20\mathrm{Cov}(X_1, X_2) = 100\sigma^2 + 4\sigma^2 = 104\sigma^2$

d) $\mathrm{Var}\left(\frac{X_1+X_2+X_3+...+X_n}{n}\right) = \frac{1}{n^2} \cdot \sum_{i=1}^{n} \left(\mathrm{Var}(X_i) + \sum_{j\neq i} \mathrm{Cov}(X_i, X_j)\right) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$

2 The influence of replicates on the precision of a measure: the average chemical composition of the interstellar medium

a) If the law follows $\mathcal{N}(74, 1)$, $P(X \geq 77) = 1 - \Phi(\frac{77-74}{1}) = 0.135\%$.
Where the value of $\Phi(z)$ can be looked up in a standard normal table ("z score") or computed via the "error function": $\Phi(z) = (\mathrm{erf}(z/\sqrt{2}) + 1)/2$

b) If it follows $\mathcal{N}(77, 1)$, $P(X \leq 74) = \Phi(\frac{74-77}{1}) = 1 - \Phi(3) = 0.135\%$.

c) The 99.99% confidence interval goes from the value below which 0.005% is of the population is situated, to the point where 99.995% is situated, such that the interval contains 99.995%-0.005% = 99.99% of the population. We find: $[x_{0.005\%}; x_{99.995\%}] = [\sigma \cdot z_{0.005\%} + \mu; \sigma \cdot z_{99.995\%} + \mu]$

$$= [-3.89\sigma + \mu; 3.89\sigma + \mu] = \begin{cases} [70.11; 77.89] & \text{for } \mathcal{N}(74, 1) \\ [73.11; 80.89] & \text{for } \mathcal{N}(77, 1) \end{cases}$$

d) If for one single measurement $\mathbb{E}(X) = 74$ and $\mathrm{Var}(X) = 1$, then for 10 independent measurements, from the central limit theorem,

$$\mu = \mathbb{E}\left(\frac{1}{10} \cdot \sum_{i=1}^{10} X_i\right) = 74 \quad \text{and} \quad \sigma^2 = \mathrm{Var}\left(\frac{1}{10} \cdot \sum_{i=1}^{10} X_i\right) = 0.1 \qquad => \quad \mathcal{N}(74, 0.1)$$

e) Given a measurement of $x = 75$, the confidence interval for the true mean μ is given as $\mu \in [x \pm \sigma \cdot z_{99.995\%}] = [71.11; 78.89]$. Unfortunately both 77% and 74% are included.
If it is the result of 10 measurements, $\mu \in [x \pm \frac{\sigma}{\sqrt{10}} \cdot z_{99.995\%}] = [73.77; 76.23]$. This time only 74% is included.
If it is the result of 1000 measurements, the confidence interval's width is $2 \cdot \frac{\sigma \cdot z_{99.995\%}}{\sqrt{1000}} = 0.246\%$

f) The bigger bound should be smaller than 77%. Then $77 - x = 2 \leq \frac{\sigma \cdot z_{99.995\%}}{\sqrt{N}}$ and $N = \frac{3.89^2}{4} = 3.78$.
Only 4 replicates would be sufficient to eliminate the model which predicts the composition of 77%.

3 Bernoulli law, Normal law and Car insurance

The probability p to have an accident during one given year, the total cost for all accidents Cost_{tot} and the average cost for an accident Cost_{av} are:

$$\begin{aligned} p &= \frac{219 + 3654 + 17759}{4'400'000} = 0.49\% \\ \text{Cost}_{tot} &= 219 \cdot 300'000 + 3654 \cdot 100'000 + 17759 \cdot 5'000 = 519'895'000 \text{ CHF} \\ \text{Cost}_{av} &= \frac{\text{Cost}_{tot}}{219 + 3654 + 17759} = 24'034 \text{ CHF} \end{aligned}$$

An average client costs $p \cdot \text{Cost}_{av} = 0.0049 \cdot 24'034 \text{ CHF} = 118 \text{ CHF}$ yearly.

To be 95% sure to be profitable, the insurance should be profitable for any accident probability in the 95% confidence interval. In the worst case the probability is:

$$p_{97.5\%} = z_{97.5\%} \cdot \sigma + \mu = 1.96\sqrt{\frac{p(1-p)}{n}} + p$$

Where n denotes the number of clients.

An insurance with n client is profitable, if it gains more money that it has to spend.

$$n \cdot \text{Client}_{Bill} \geq \frac{n}{4'400'000} 1'700'000'000 + n \cdot p_{97.5\%} \cdot \text{Cost}_{av}$$

Where the total cost of working is normalized by the number of client.

So the minimal client's bill per year is 537.5 CHF if the insurance has 10'000 clients and 507.8 CHf if it has 1 million clients.

The profit done by car insurance companies is $4'400'000 \cdot 1100 - 1'700'000'000 - \text{Cost}_{tot} = 2.62$ bilions CHF. Which makes a profit of 262'000 CHF per employees.

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

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

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i} = 1.604 \qquad \tilde{x}_1 = \frac{1}{2}(1.71 + 1.81) = 1.76 \qquad s_1 = \sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2} = 0.29$$

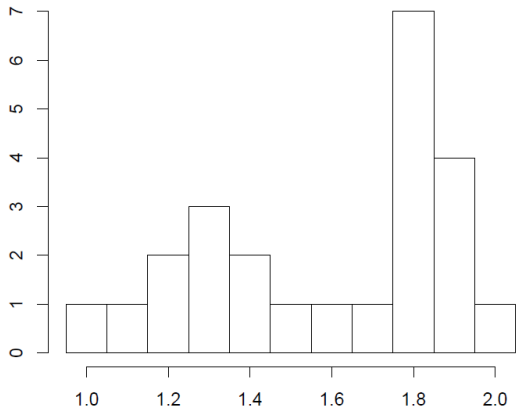
Since the median isn't close to the mean, the distribution is probably asymmetric. The histogram drawn below confirms that and clearly shows 2 populations. Because of these, the difference between $\tilde{x}_{0.25}$ and $\tilde{x}_{0.75}$ is big, so there aren't any outliers.

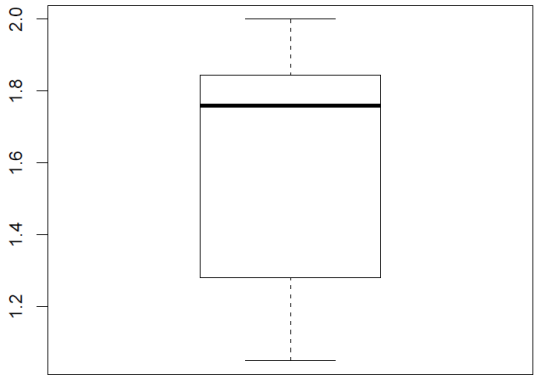
In the dataset, 14 samples over 24 (58.3%) are within the intervals $[\bar{x}_1 \pm s_1] = [1.31; 1.90]$ and every one (100%) are within $[\bar{x}_1 \pm 2 \cdot s_1] = [1.01; 2.19]$.

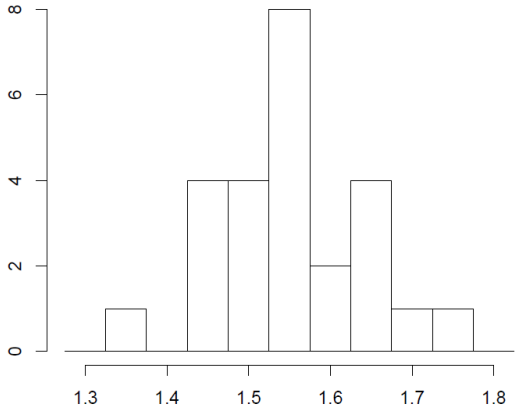
Using the normal law we get:

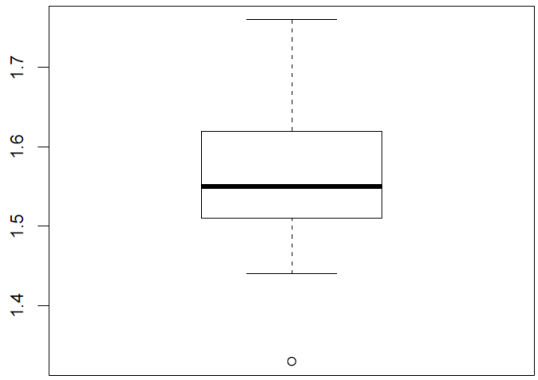
$$\begin{aligned} P(\mu - \sigma \leq X \leq \mu + \sigma) &= P(X \leq \mu + \sigma) - P(X < \mu - \sigma) \\ &= P\left(Z \leq \frac{\mu + \sigma - \mu}{\sigma}\right) - P\left(Z \leq \frac{\mu - \sigma - \mu}{\sigma}\right) \\ &= \Phi(1) - \Phi(-1) = \Phi(1) - (1 - \Phi(1)) = 0.682 \\ P(\mu - 2\sigma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

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.









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.

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.

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.

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