EPFL
Autumn Semester 2025

Probability and statistics for materials science
Prof. Gregor Jotzu

Exercise Set 4 - Solution

1 Nanoparticles with a normal distribution

The z-value is $z = (x - \mu)/\sigma$. To get $\Phi(z)$ you can look it up in a z-table ("standard normal table") e.g. on Wikipedia, z-table.net, ztable.io (has nice illustrations)... Or compute it using

$$\Phi(z) = \frac{1}{2}(1 + \text{erf}[z/\sqrt(2)]) = \frac{1}{2}(\text{erfc}[-z/\sqrt(2)])$$

where erf is the "error function" and erfc the "complementary error function" (which some calculators feature). Also keep in mind that $\Phi(-z) = 1 - \Phi(z)$

a) 90% interval means between a cumulative probability ($\Phi(z)$) of 5% and 95%. For $\Phi(z) = 95\%$ the z-value is 1.645, for 5% it is consequently -1.645.

In method A this corresponds to 359nm to 441nm.

In method B this corresponds to 374nm to 506nm.

b) $z_A = -2$ and $z_B = -2.25$. In principle this is enough to say that method B is better (lower z-value means lower probability).

We can still caclulate the probabilities, which are $P_A(x \leq 350\text{nm}) = 0.02275$ and $P_B(x \leq 350\text{nm}) = 0.01222$. So method B is almost twice as good.

c) $z_A = -4$ and $z_B = -3.5$, so now method A is better. $P_A(x \leq 300\text{nm}) = 0.00003$ and $P_B(x \leq 300\text{nm}) = 0.00023$, so it is indeed much better. Note that although the mean of method B is further away from 300nm when measured in nm, it is less far away when measured in standard deviations.

d) The z-value we look for is still the same (1.645). But now the standard deviation *of the mean* for $N = 1'000'000$ nanoparticles is given by $\sigma_{\text{mean}} = \sigma/\sqrt{1'000'000} = \sigma/1'000$ (this you get from the central limit theorem or just from the fact that they are i.i.d.). So for method A we have the range 399.96nm to 400.04nm, and for method B we have 439.93nm to 440.07

e) So in this case we know σ_{mean} and can use it to estimate σ of the population/process. This gives $\sigma = \sigma_{\text{mean}} * \sqrt{N}$, where N is now $10'000$. So $\sigma = 10$nm for method C, lower than for any of the other methods.

The z-value is now -4 for the requirement in b, and -9 for the requirement in c (giving a probability of only about 1 in 10^{19} for particles below 300nm!). So in both cases your method C is the best. Well done!

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": $\Phi(z) = (\text{erf}(z/\sqrt{2}) + 1)/2$

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 Poisson's law and supernova explosions

a) The random variable X can be any integer equal or greater than zero. The cumulative probability over all the possible value should be equal to 1 (so it is a well-defined probability distribution)..

$$\sum_{k=0}^{\infty} P(X = k) = C \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = C \cdot e^{\lambda} = 1 \qquad => \qquad C = e^{-\lambda}$$

b) To obtain the mean we have to take a sum over all possible values, weighted by their probabilities.

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} P(X = k) \cdot k = e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \cdot k$$

We can drop the $k = 0$ element of the sum as this one is now 0 (we multiply with 0 there)

$$= e^{-\lambda} \cdot \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \cdot k$$

Then we use the fact that $k/k! = 1/(k-1)!$ and also factor out one λ before the sum.

$$= \lambda e^{-\lambda} \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

We substitute $(k-1)$ for some dummy variable (here called l), so we recognize again the form of the Maclaurin series of the exponential function.

$$= \lambda e^{-\lambda} \cdot \sum_{l=0}^{\infty} \frac{\lambda^l}{l!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

c) The value of k is the number of events per century, hence $\lambda = \frac{1}{4}$.

d) Now we can simply evaluate the Poissonian distribution function:

$$
\begin{aligned}
P(X=2) &= e^{-0.25}\cdot\frac{0.25^2}{2} = 0.0243 \\
P(X\geq 2) &= 1-P(X=1)-P(X=0) = 1-e^{-0.25}(0.25+1) = 0.0265 \\
P(X\geq 1) &= 1-P(X=0) = 1-e^{-0.25} = 0.221 \\
P(X=0) &= e^{-0.25} = 0.779
\end{aligned}
$$

4 Aircraft wing probability of failure

a) First we compute the probability of unsafe fly for both kinds of plane. All engine's failures are independent from each other and a Binomial distribution is considered.

$$
\begin{aligned}
P_2 = P(\text{2-engines plane fails}) &= P(\text{2 engines fail}) = p^2 = 10^{-12} \\
P_4 = P(\text{4-engines plane fails}) &= P(\text{2 engines fail on the same side}) + P(\text{3 or 4 engines fail}) \\
&= 2\cdot p^2(1-p)^2 + \binom{4}{3}p^3(1-p) + p^4 = 2\cdot p^2(1-p)\cdot(1+p) + p^4 \\
&= 2\cdot p^2(1-p^2) + p^4 = 2\cdot p^2 - p^4 = 2\cdot 10^{-12} - 10^{-24}
\end{aligned}
$$

So, the 4-reactors planes is slightly less safe.

b) Let Y be the random variable which indicates the number of failing planes due to reactor issues per year. Its average is:

$$
\mathbb{E}(Y) = 365\cdot 24'000\cdot\frac{P_2+P_4}{2} = 13.14\cdot 10^{-6}
$$

c) We can model the event a plane fail with a Binomial distribution of parameter P_2 for 2-reactors planes and P_4 for the other ones. The number of trials $n = 0.5\cdot 365\cdot 24'000$ is the same for both kinds of planes.

$$
\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}(\mathcal{B}(P_2,n)+\mathcal{B}(P_4,n)) = \mathrm{Var}(\mathcal{B}(P_2,n)) + \mathrm{Var}(\mathcal{B}(P_4,n)) \\
&= nP2(1-P_2) + nP4(1-P_4) = 13.14\cdot 10^{-6} \\
\sigma_Y &= \sqrt{\mathrm{Var}(Y)} = 3.62\cdot 10^{-3}
\end{aligned}
$$

Both variances simply add because the event a 2-reactors plane fails is independent to the event a 4-reactors plane fails.

Alternative approach

Since both $n\cdot P_2$ and $n\cdot P_4$ are really small, it is also possible to use Poisson's distribution with parameter $\lambda_2 = n\cdot P_2$ resp. $\lambda_4 = n\cdot P_4$. In fact this law is an approximation of the Binomial distribution for highly improbable events.

The average and standard deviation are:

$$
\begin{aligned}
\mathbb{E}(Y) &= \mathbb{E}(\mathcal{P}(\lambda_2)+\mathcal{P}(\lambda_4)) = \lambda_2+\lambda_4 = 13.14\cdot 10^{-6} \\
\mathrm{Var}(Y) &= \mathrm{Var}(\mathcal{P}(\lambda_2)+\mathcal{P}(\lambda_4)) = \lambda_2+\lambda_4 = 13.14\cdot 10^{-6} \\
\sigma_Y &= \sqrt{\mathrm{Var}(Y)} = 3.62\cdot 10^{-3}
\end{aligned}
$$

Which give the same result as before. We use $\mathbb{E}(\mathcal{P}(\lambda)) = \lambda$ (see exercise 2 above) and $\mathrm{Var}(\mathcal{P}(\lambda)) = \lambda$.

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\
\mathbb{E}(X^2) &= \sum_{k=0}^{\infty} P(X = k) \cdot k^2 = e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \cdot k^2 \\
&= e^{-\lambda} \cdot \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \cdot k(k-1) + e^{-\lambda} \cdot \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \cdot k \\
&= \lambda e^{-\lambda} \cdot \left(\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \cdot (k-1) + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \right) \\
&= \lambda e^{-\lambda} \cdot \left(\lambda \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \sum_{l=0}^{\infty} \frac{\lambda^l}{l!} \right) \\
&= \lambda e^{-\lambda} \cdot \left(\lambda e^{\lambda} + e^{\lambda} \right) = \lambda^2 + \lambda \\
\mathrm{Var}(X) &= \lambda^2 + \lambda - \lambda^2 = \lambda
\end{aligned}$$