Exercises Set 5

1 Demonstrate an unbiased estimator for the variance

We need to prove that the bias of the estimator equals zero. We have stated that each X_j has the same mean, μ, i.e. the expectation value of every mean is μ.

$$\mathbb{E}(X_j) \quad = \quad \mathbb{E}(X_1) = ... = \mathbb{E}(X_n) = \mathbb{E}(\bar{X}) = \mu$$

The bias is the difference between the expectation value of our estimator, and the thing we want to estimate, in this case the real variance.

Note that the expectation value is a linear function (it is just a weighted sum), so we have, in general $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$, for any two random variables X and Y. Furthermore $\mathbb{E}(aX) = a\mathbb{E}(X)$ for any constant a.

$$\begin{aligned}\text{Bias}(s^2) &= \mathbb{E}(s^2) - \sigma^2 \\ &= \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) - \sigma^2\end{aligned}$$

Because of the summing rule above, we can take the sum out of the expectation value.

$$= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \quad - \sigma^2$$

And since all X_i are the same, the sum over n is just the expectation of any X_j (we could have for example just have chosen X_1) times n.

$$= \frac{n}{n-1}\left\{\mathbb{E}(X_j^2) - 2\cdot\mathbb{E}(X_j\bar{X}) + \mathbb{E}(\bar{X}^2)\right\} - \sigma^2$$

To compute the remaining terms, we write out the expression for $\overline{X}$ explicitly. Note that the expectation value of the square of a random variable is not just the expectation value squared.

$$= \frac{n}{n-1}\left\{\mathbb{E}(X_j^2) - 2\cdot\mathbb{E}\left(X_j\frac{1}{n}\sum_{i=1}^{n}X_i\right) \qquad + \mathbb{E}\left(\frac{1}{n^2}\left(\sum_{i=1}^{n}X_i\right)^2\right)\right\} - \sigma^2$$

The term in the middle will once contain X_j^2 (i.e. X_1^2 if we had chosen X_1), so there the expectation value is $\mathbb{E}(X_j^2)$. In the other $n-1$ times it will contain X_iX_j with $i \neq j$. These are two identical and independent distributions. Because of that, the joint probability is just the product of probabilities ($P(x_j \cap x_i) = P(x_j)\cdot P(x_i)$, and the expectation value of the product is just the product of the expectation value i.e. $\mathbb{E}(X_j \cdot X_i) = \mathbb{E}(X_j)\cdot\mathbb{E}(X_i)$ now. As each of those products (for any i) is equivalent, the sum is just $(n-1)$ times the product $\mathbb{E}(X_j)\cdot\mathbb{E}(X_i) = \mathbb{E}(X_j)^2$.

$$= \frac{n}{n-1}\left\{\mathbb{E}(X_j^2) - \frac{2}{n}\left((n-1)\cdot\mathbb{E}(X_j)^2 + \mathbb{E}(X_j^2)\right) + \frac{1}{n^2}\mathbb{E}\left(\left(\sum_{i=1}^{n}X_i\right)^2\right)\right\} - \sigma^2$$

For the next term, a similar rule applies, we have n times a product of some X_j with itself, and $n(n-1)$ times the product of two independent random variables.

$$
\begin{aligned}
&= \frac{n}{n-1}\left\{\frac{n-2}{n}\mathbb{E}(X_j^2) - \frac{2(n-1)}{n}\mathbb{E}(X_j)^2 + \frac{1}{n^2}\left(n\cdot\mathbb{E}(X_j^2) + n(n-1)\cdot\mathbb{E}(X_j)^2\right)\right\} - \sigma^2 \\
&= \frac{n}{n-1}\left\{\frac{n-1}{n}\mathbb{E}(X_j^2) - \frac{n-1}{n}\mathbb{E}(X_j)^2\right\} - \sigma^2 \\
&= \mathbb{E}(X_j^2) - \mathbb{E}(X_j)^2 - \sigma^2 \\
&= \mathrm{Var}(X_j) - \sigma^2 = 0
\end{aligned}
$$

2 Two different estimators of a macromolecule length measurement

a) The cumulative distribution function (i.e. the probability to find a value smaller or equal than x) is:

$$
P(X \le x) \;=\; \int_{-\infty}^{x} f(x)dx = \int_0^x c\cdot dx = \begin{cases} 0 & x \le 0 \\ c\cdot x & 0 < x \le \theta \\ c\cdot\theta & x > \theta \end{cases}
$$

Since $P(X \le \infty) = P(X \le \theta)$ should be one for the probability to be normalized, we have to have $c = 1/\theta$.

b) The expectation value of the first estimator can be found by operations on the expectation operator. To find the expectation value of any measurement X_j, we take the integral over all values weighted by the cumulative distribution function, which just gives us $\theta/2$ as expected.

$$
\mathbb{E}(\theta_1) \;=\; \mathbb{E}\left(\frac{2}{n}\sum_{i=1}^{n} X_i\right) = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = 2\cdot\mathbb{E}(X_j) = 2\int_0^\theta \frac{x}{\theta}dx = \theta
$$

For the second one, we use the density function, which is the derivative of the cumulative distribution.

$$
\begin{aligned}
F_{\theta_2}(t) &= P(\max_i X_i \le t) = P(X_1 \le t, ..., X_n \le t) = P(X_j \le t)^n = \left(\frac{t}{\theta}\right)^n \\
f_{\theta_2}(t) &= \frac{d}{dt}F_{\theta_2}(t) = n\frac{t^{n-1}}{\theta^n} \\
\mathbb{E}(\theta_2) &= \mathbb{E}(\max_i X_i) = \int_0^\theta f_{\theta_2}(x)x\cdot dx = n\int_0^\theta \frac{x^n}{\theta^n}dx = \frac{n}{n+1}\theta
\end{aligned}
$$

So the second estimator is biased. A non-biased estimator would be $\theta_2^* = \frac{n+1}{n}\theta_2$.

c) The mean square errors are just the variances because the bias is zero:

$$\begin{aligned}
MSE(\theta_1) &= \mathbb{E}\left(\left(\frac{2}{n}\sum_{i=1}^{n} X_i - \theta\right)^2\right) = \frac{4}{n^2}\mathbb{E}\left(\left(\sum_{i=1}^{n} X_i\right)^2\right) - \frac{4\theta}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) + \theta^2 \\
&= \frac{4}{n^2}\left(n\cdot\mathbb{E}(X_j^2) + n(n-1)\cdot\mathbb{E}(X_j)^2\right) - \frac{4\theta}{n}\frac{n\theta}{2} + \theta^2 \\
&= \frac{4}{n}\left(\mathbb{E}(X_j^2) + (n-1)\frac{\theta^2}{4}\right) - \theta^2 = \frac{4}{n}\int_0^\theta \frac{x^2}{\theta}dx - \frac{\theta^2}{n} \\
&= \frac{\theta^2}{3n} \\
MSE(\theta_2^*) &= \mathbb{E}\left(\left(\frac{n+1}{n}\theta_2 - \theta\right)^2\right) = \frac{(n+1)^2}{n^2}\mathbb{E}\left(\theta_2^2\right) - \frac{2\theta(n+1)}{n}\mathbb{E}(\theta_2) + \theta^2 \\
&= \frac{(n+1)^2}{n^2}\int_0^\theta f_{\theta_2}(x)x^2\cdot dx - \theta^2 = \frac{(n+1)^2}{n}\int_0^\theta \frac{x^{n+1}}{\theta^n}dx - \theta^2 \\
&= \theta^2\left(\frac{(n+1)^2}{n(n+2)} - 1\right) \\
&= \frac{\theta^2}{n(n+2)}
\end{aligned}$$

d) θ_2^* is the best estimator for θ, because its mean square error is smaller. (Technically they are the same for n=1, but clearly for 1 measurement the variance it not well defined). Importantly, the MSE decreases quadratically with the number of measurements, rather than just linearly.

3 The distribution function of a mean

a) For the mean of two random variables to give some value m, they have to sum to $2m$. This is the case if the first gives $m-t$ and the second gives $m+t$, where t can be any number between $-\infty$ to ∞. Alternatively we could also say that if the first gives t then the second has to give $2m-t$ - this is equivalent and is found in the literature more often so we will use the latter approach here, but the first approach also works.

The first probability distribution is given by

$$p_1(x_1) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{x_1^2}{2\sigma^2}\right] \tag{1}$$

where the standard deviation $\sigma = 4$ and the second is given by

$$p_2(x_2) = \frac{1}{\rho\sqrt{2\pi}}\exp\left[-\frac{x_2^2}{2\rho^2}\right] \tag{2}$$

where we have now used ρ to denote the standard deviation, so that the equations below don't become too messy by having σ with indices. Here $\rho = 3$.

The mean, $m = (x_1 + x_2)/2$, has a probability distribution $p_{\text{mean}}(m)$ which we now want to find using the argument above. That is, we multiply the two probability distributions, setting $x_1 = t$ and $x_2 = 2m - t$. To simplify things (as normalisation is a bit subtle here) we look at the un-normalized probability distributions for the moment. That means we leave out the normalization factors in p_1 and p_2 above and just calculate what p_{mean} will be proportional to (denoted by the sign $\propto$. Then, in the end, we can ensure that the integral of p_{mean} over all possible means is equal to 1 to find the normalization pre-factor.

So we get

$$p_{\text{mean}}(m) \propto \int_{-\infty}^{+\infty} p_1(t)p_1(2m-t)dt$$

which so far is still general. Putting in the probability distributions we use (and using σ and ρ to keep things general - but you can also directly put numbers in, which makes the calculation easier) we get:

$$p_{\text{mean}}(m) \propto \int_{-\infty}^{+\infty} \exp\left[-\frac{t^2}{2\sigma^2}\right] \cdot \exp\left[-\frac{(2m-t)^2}{2\rho^2}\right] dt \tag{3}$$

$$\propto \int_{-\infty}^{+\infty} \exp\left[-\left\{\frac{t^2}{2\sigma^2} + \frac{(2m)^2}{2\rho^2} + \frac{t^2}{2\rho^2} + \frac{-4mt}{2\rho^2}\right\}\right] dt \tag{4}$$

We can take the term containling only m, no t, out of the integral

$$\propto \exp\left[-\frac{(2m)^2}{2\rho^2}\right] \int_{-\infty}^{+\infty} \exp\left[-\left\{\frac{t^2}{2\sigma^2} + \frac{t^2}{2\rho^2} + \frac{-4mt}{2\rho^2}\right\}\right] dt \tag{5}$$

$$\propto \exp\left[-\frac{2m^2}{\rho^2}\right] \int_{-\infty}^{+\infty} \exp\left[-\left\{\frac{(\rho^2+\sigma^2)}{2\sigma^2\rho^2} t^2 + \frac{-2m}{\rho^2} t\right\}\right] dt \tag{6}$$

So we have an integral of the generalized Gaussian function $\exp\left[-(at^2+bt)\right]$ with $a = \frac{(\rho^2+\sigma^2)}{2\sigma^2\rho^2}$ and $b = -2m/\rho^2$. When integrated from $-\infty$ to ∞ this function gives $\sqrt{\pi/a}\exp\left[b^2/4a\right]$ - see below or online for a proof. The square root has no dependence on m, meaning it is just a constant pre-factor which we can leave out since we are only looking at proportionalities (we will determine the correct proportionality later). So we get

(7)

$$\propto \exp\left[-\frac{2m^2}{\rho^2}\right] \cdot \exp\left[\left\{\frac{\left(\frac{-2m}{\rho^2}\right)^2}{4\left(\frac{(\rho^2+\sigma^2)}{2\sigma^2\rho^2}\right)}\right\}\right] \tag{8}$$

$$\propto \exp\left[-\frac{2m^2}{\rho^2}\right] \cdot \exp\left[\left\{\frac{4m^2}{4\rho^4} \frac{2\sigma^2\rho^2}{(\rho^2+\sigma^2)}\right\}\right] \tag{9}$$

Putting together the exponentials

$$\propto \exp\left[\left\{\frac{2m^2\sigma^2}{\rho^2(\rho^2+\sigma^2)} - \frac{2m^2}{\rho^2}\right\}\right] \tag{10}$$

$$\propto \exp\left[\left\{\frac{2m^2\sigma^2}{\rho^2(\rho^2+\sigma^2)} - \frac{2m^2(\rho^2+\sigma^2)}{\rho^2(\rho^2+\sigma^2)}\right\}\right] \tag{11}$$

$$\propto \exp\left[\left\{\frac{2m^2(\sigma^2-\rho^2-\sigma^2)}{\rho^2(\rho^2+\sigma^2)}\right\}\right] \tag{12}$$

$$\propto \exp\left[-\frac{m^2}{2[(\rho^2+\sigma^2)/4]}\right] \tag{13}$$

Which we recognize as a Gaussian distribution with a mean of zero, and a variance of $(\rho^2+\sigma^2)/4$ (or equivalently, a standard deviation of $\sqrt{(\rho^2+\sigma^2)}/2$. We know how such a distribution needs to be normalized (see Eq. 1), giving us the final result:

$$p_{\text{mean}}(m) = \exp\left[-\frac{m^2}{2[(\rho^2+\sigma^2)/4]}\right] \tag{14}$$

b) As derived above, we have a mean of 0 and a standard deviation of $\sqrt{(\rho^2+\sigma^2)}/2$.

We could also have derived that without knowledge of the probability distributions themselves, only by knowing their respective means and standard deviations.

For the mean m, because the mean of each individual distribution is 0, and they are independent, the mean of m is also 0.

We could also have derived this from the rules for adding variances (which can always be derived from the definition of the variance).

The variance of m would then be

$$\begin{aligned}\mathrm{Var}(m) &= \mathrm{Var}((x_1 + x_2)/2) \\ &= \mathrm{Var}(x_1/2) + \mathrm{Var}(x_2/2)) \\ &= \mathrm{Var}(x_1)/4 + \mathrm{Var}(x_2)/4 \\ &= (\sigma^2 + \rho^2)/4\end{aligned}$$

And the standard deviation of m is the square root of that, $\sqrt{\sigma^2 + \rho^2}/2$.

Note however, that although we were able to compute the mean and standard deviation of m without knowing the distribution function, we did not *a priori* know the full probability distribution function of m. It turns out that for Gaussian distribution functions, the mean of two independent Gaussian distributions is also Gaussian. We just proved this for the case where they have same the center, but it is true even if they have different centers.

But for other distributions, this is NOT true in general. For example, if you take two uniform distributions, the mean will not just be uniform. (We proved the discrete version of this by looking at the mean of two Bernoulli distributions).

Proof of the integral of the generalized Gaussian

We assume that we know the usual Gaussian integral:

$$\int_{-\infty}^{+\infty} \exp\left[-t^2\right] dt = \sqrt{\pi} \tag{15}$$

for which a number of proofs exist.

Shift the Gaussian function so that it is not centered at 0 but at some finite value c, giving $\exp\left[-(t-c)^2\right]$. The integral of this function should obviously be unchanged, as we are going from $-\infty$ to ∞ so the center does not matter. We can do the calculation explicitly, by substituting $t = v + c$ with $dt = dv$. The integration limits are still at infinity of course ($\infty - c = \infty$), so we get:

$$\int_{-\infty}^{+\infty} \exp\left[-(t-c)^2\right] dt = \int_{-\infty}^{+\infty} \exp\left[-v^2\right] dv = \sqrt{\pi} \tag{16}$$

We can then add a pre-factor to get the scaled Gaussian function $\exp\left[-at^2\right]$, where a is a positive number. The integral of this function is found by substituting $t = u/\sqrt{a}$ with $dt = du/\sqrt{a}$. The integration limits are still at infinity of course (as $\infty\sqrt{a} = \infty$), so we get:

$$\int_{-\infty}^{+\infty} \exp\left[-at^2\right] dt = \int_{-\infty}^{+\infty} \exp\left[-u^2\right] \frac{1}{\sqrt{a}} du = \frac{\sqrt{\pi}}{\sqrt{a}} \tag{17}$$

To find the generalised form with a linear term, we have to "complete the square", i.e. we have to bring the expression below into a form like above.

$$\int_{-\infty}^{+\infty} \exp\left[-\{at^2 + bt\}\right] dt = \int_{-\infty}^{+\infty} \exp\left[-\left\{at^2 + bt + \frac{b^2}{4a} - \frac{b^2}{4a}\right\}\right] dt \tag{18}$$

$$= \int_{-\infty}^{+\infty} \exp\left[-\left\{\left(\sqrt{a}t + \frac{b}{2\sqrt{a}}\right)^2 - \frac{b^2}{4a}\right\}\right] dt \tag{19}$$

The last term can be taken out of exponential and then out of the integral, as it does not depend on t.

$$\tag{20}$$

$$\exp\left[\frac{b^2}{4a}\right]\int_{-\infty}^{+\infty}\exp\left[-\left\{\left(\sqrt{a}t+\frac{b}{2\sqrt{a}}\right)^2\right\}\right]dt \tag{21}$$

$$=\exp\left[\frac{b^2}{4a}\right]\int_{-\infty}^{+\infty}\exp\left[-\left\{a\left(t+\frac{b}{2a}\right)^2\right\}\right]dt \tag{22}$$

The integral now has the form of a shifted and scaled Gaussian. The shift does nothing to the integral, the scaling comes in as derived above, so overall we get:

$$\tag{23}$$

$$=\exp\left[\frac{b^2}{4a}\right]\frac{\sqrt{\pi}}{\sqrt{a}} \tag{24}$$

$$\tag{25}$$

Meaning

$$\int_{-\infty}^{+\infty}\exp\left[-\{at^2+bt\}\right]dt=\sqrt{\frac{\pi}{a}}\exp\left[-\frac{b^2}{4a}\right] \tag{26}$$