Exercise Set 3 - 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 completeness, 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) $\mathrm{Var}\left(\frac{X_1+X_2+X_3+...+X_n}{n}\right) = \frac{1}{n^2} \cdot \sum_{i=1}^{n} \left(\mathrm{Var}(X_i) + \sum_{j\neq i} \mathrm{Cov}(X_i, X_j)\right) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$

e) We have

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

and

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

Based on that we find

$$\mathbb{E} = \sum_{i=1}^{n} p_i \cdot x_i = \frac{1}{n} \sum_{i=1}^{n} i = \frac{n+1}{2}$$

and

$$\begin{aligned}
\mathbb{V} = \mathbb{E}(X^2) & - (\mathbb{E}(\mathbb{X}))^2 && (1)\\
= \sum_{i=1}^{n} p_i (x_i)^2 & - \left(\frac{n+1}{2}\right)^2 && (2)\\
= \frac{1}{n} \sum_{i=1}^{n} i^2 & - \left(\frac{n+1}{2}\right)^2 && (3)\\
= \frac{1}{n} \frac{n(n+1)(2n+1)}{6} & - \frac{(n+1)(n+1)}{4} && (4)\\
= (n+1) \left(\frac{2n+1}{6}\right. & \left. - \frac{n+1}{4}\right) && (5)\\
= (n+1) \left(\frac{4n+2}{12}\right. & \left. - \frac{3n+3}{12}\right) && (6)\\
= \frac{(n+1)(n-1)}{12} & && (7)
\end{aligned}$$

2 Bernoulli law, Normal law and Car insurance

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

$$\begin{aligned} p &= \frac{219+3654+17759}{4'400'000} = 0.004916... \approx 0.49\% \\ C_{\text{total}} &= 219 \cdot 300'000 + 3654 \cdot 100'000 + 17759 \cdot 5'000 = 519'895'000 \text{ CHF} \\ C_{\text{av}} &= \frac{C_{\text{total}}}{219+3654+17759} = 24'033.6... \text{ CHF} \approx 24\text{kCHF} \end{aligned}$$

The total cost for accident payouts, averaged per client is $p \cdot C_{\text{av}} = 0.0049 \cdot 24'033.6...$ CHF $=$ 118.15795...CHF yearly, which could alternatively be calculated directly as $C_{\text{total}}/(4.4 \times 10^6)$

In addition there are the fixed costs, and per client those amount to 1.7×10^9CHF$/(4.4 \times 10^6) = 386.\overline{36}$CHF.

Summing those two up we get about 504.52CHF yearly (which you can round to about 500CHF yearly for a final result with two significant digits - but keep the more precise value for the following).

Note that for the company to hence make zero profit *on average*, i.e. make a profit 50% of the time and a loss 50% of the time (which could for example be the policy of a state-based insurance) it would have to charge clients exactly this amount. But in the following we want to find out how much it has to charge to make a profit in 95% of the years.

b) To be 95% sure to be profitable, the probability to make a profit (i.e. lose less than 0 CHF) has to integrate to 0.95. That is, any result *smaller than* a certain limit is acceptable.

For an individual client, the probability of accident (the Bernoulli probability) is 0.0049. For 10′000 clients, we assume that each client's probability is independent and get to a Binomial distribution with $n = 10'000$ and $p = 0.004916....$

This has an expectation value of $np = 10'000 \cdot 0.004916.. = 49.16...$ accidents, corresponding to $49.16... \cdot 24033.6...CHF= 1.18 \times 10^6$CHF accident-related total cost, or 118.1...CHF per customer (as above). **To this we add the fixed costs per customer, giving again an expectation value of $\mu = 504.52...$CHF.**

The standard deviation is $\sqrt{(np(1-p))} = 6.99...$ accidents corresponding to $0.1681... \times 10^6$CHF on the total cost **or** $\sigma = \frac{\sqrt{(np(1-p))}}{n} = 16.81...$**CHF on the per-customer cost.**

We now approximate the possible end-of-year costs using a Gaussian/Normal distribution with μ the and σ determined above.

Capturing 95% of the total outcomes, given that we have a one-sided distribution (we are always fine with fewer accidents !) corresponds to a **z-value of about 1.65**, since $\Phi(1.65) \approx 0.95$.

So with 95% probability, a customer will cost (in CHF) less than

$$z \cdot \frac{\sqrt{(np(1-p)}}{n} + \mu = z \cdot 16.81... + 504.52... = 532.258... \approx 532$$

So the minimal client's bill per year is about 532 CHF if the insurance has 10′000 clients.

b) For one million clients, we see that the standard deviation of the per-client cost reduces (it will now be 10 times smaller as it scales with $1/\sqrt{(n)}$. So we get $z \cdot 1.681... + 504.52... \approx$ 507CHF.

d) Here we can just go back to the values in the table. The made by car insurance companies is $4'400'000 \cdot 1100 - 1'700'000'000 - C_{\text{total}} = 2.62$billion CHF. Which gives a profit of 262′000CHF per employee.

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

$$p_1(x_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\,[-\frac{x_1^2}{2\sigma^2}] \tag{8}$$

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

$$p_2(x_2) = \frac{1}{\rho\sqrt{2\pi}} \exp\,[-\frac{x_2^2}{2\rho^2}] \tag{9}$$

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:

$$p_{\text{mean}}(m) \propto \int_{-\infty}^{+\infty} \exp\,[-\frac{t^2}{2\sigma^2}] \cdot \exp\,[-\frac{(2m-t)^2}{2\rho^2}]dt \tag{10}$$

$$\propto \int_{-\infty}^{+\infty} \exp\,[-\left\{\frac{t^2}{2\sigma^2} + \frac{(2m)^2}{2\rho^2} + \frac{t^2}{2\rho^2} + \frac{-4mt}{2\rho^2}\right\}]dt \tag{11}$$

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

$$\propto \exp\,[-\frac{(2m)^2}{2\rho^2}] \int_{-\infty}^{+\infty} \exp\,[-\left\{\frac{t^2}{2\sigma^2} + \frac{t^2}{2\rho^2} + \frac{-4mt}{2\rho^2}\right\}]dt \tag{12}$$

$$\propto \exp\,[-\frac{2m^2}{\rho^2}] \int_{-\infty}^{+\infty} \exp\,[-\left\{\frac{(\rho^2+\sigma^2)}{2\sigma^2\rho^2}t^2 + \frac{-2m}{\rho^2}t\right\}]dt \tag{13}$$

So we have an integral of the generalized Gaussian function $\exp[-(at^2+bt)]$ 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 $\sqrt{\pi/a}\exp[b^2/4a]$ - 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

$$\tag{14}$$

$$\propto \exp[-\frac{2m^2}{\rho^2}]\cdot\exp[\left\{\frac{\left(\frac{-2m}{\rho^2}\right)^2}{4\left(\frac{(\rho^2+\sigma^2)}{2\sigma^2\rho^2}\right)}\right\}] \tag{15}$$

$$\propto \exp[-\frac{2m^2}{\rho^2}]\cdot\exp[\left\{\frac{4m^2}{4\rho^4}\frac{2\sigma^2\rho^2}{(\rho^2+\sigma^2)}\right\}] \tag{16}$$

Putting together the exponentials

$$\propto \exp[\left\{\frac{2m^2\sigma^2}{\rho^2(\rho^2+\sigma^2)} - \frac{2m^2}{\rho^2}\right\}] \tag{17}$$

$$\propto \exp[\left\{\frac{2m^2\sigma^2}{\rho^2(\rho^2+\sigma^2)} - \frac{2m^2(\rho^2+\sigma^2)}{\rho^2(\rho^2+\sigma^2)}\right\}] \tag{18}$$

$$\propto \exp[\left\{\frac{2m^2(\sigma^2-\rho^2-\sigma^2)}{\rho^2(\rho^2+\sigma^2)}\right\}] \tag{19}$$

$$\propto \exp[-\frac{m^2}{2[(\rho^2+\sigma^2)/4]}] \tag{20}$$

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. 8), giving us the final result:

$$p_{\text{mean}}(m) = \exp\left[-\frac{m^2}{2[(\rho^2+\sigma^2)/4]}\right] \tag{21}$$

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

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

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:

$$\int_{-\infty}^{+\infty} \exp[-t^2]dt = \sqrt{\pi} \tag{22}$$

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 $\exp[-(t-c)^2]$. 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:

$$\int_{-\infty}^{+\infty} \exp[-(t-c)^2]dt = \int_{-\infty}^{+\infty} \exp[-v^2]dv = \sqrt{\pi} \tag{23}$$

We can then add a pre-factor to get the scaled Gaussian function $\exp[-at^2]$, 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:

$$\int_{-\infty}^{+\infty} \exp[-at^2]dt = \int_{-\infty}^{+\infty} \exp[-u^2]\frac{1}{\sqrt{a}}du = \frac{\sqrt{\pi}}{\sqrt{a}} \tag{24}$$

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.

$$\int_{-\infty}^{+\infty} \exp[-\{at^2 + bt\}]dt = \int_{-\infty}^{+\infty} \exp[-\left\{at^2 + bt + \frac{b^2}{4a} - \frac{b^2}{4a}\right\}]dt \tag{25}$$

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

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

$$\tag{27}$$

$$\exp\left[\frac{b^2}{4a}\right] \int_{-\infty}^{+\infty} \exp[-\left\{\left(\sqrt{a}t + \frac{b}{2\sqrt{a}}\right)^2\right\}]dt \tag{28}$$

$$= \exp\left[\frac{b^2}{4a}\right] \int_{-\infty}^{+\infty} \exp[-\left\{a\left(t + \frac{b}{2a}\right)^2\right\}]dt \tag{29}$$

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:

$$\tag{30}$$

$$= \exp\left[\frac{b^2}{4a}\right]\frac{\sqrt{\pi}}{\sqrt{a}} \tag{31}$$

$$\tag{32}$$

Meaning

$$\int_{-\infty}^{+\infty} \exp\left[-\{at^2+bt\}\right]dt = \sqrt{\frac{\pi}{a}}\exp\left[-\frac{b^2}{4a}\right] \tag{33}$$