Questions:

1. **Stock market** The price of the ABCD stock is $256 today. Every day the stock price goes either up by 25%, or down by 20%, with equal probabilities, and all days are independent from each other. A call option allows buying the stock for $300 exactly in 4 days.
   How much is this option worth?

2. **Plane**. The Aeroflot flight to Saint Petersburg on an Airbus has 380 seats. Suppose that any customer who booked this flight will cancel with probability $1\text{-}p = 0.1$, independently of other customers. Suppose that the company sells a fixed number N of tickets, with $N \geq 380$. Let X be the number of customers present the day of the flight.

- What is the distribution of X?
- Using the central limit theorem, approximate the probability $P(X \geq n)$: plug the mean and variance from the original distribution into normal.
- What is the largest number of bookings the company can accept if it does not want to refuse customers with probability more than 0.05.
- Suppose the company chooses N equal to the value calculated in the previous question. What is the probability that at least 35 seats are left unoccupied?

3. **Batteries**. Suppose that 2 batteries are randomly (uniformly) chosen without replacement from the following collection of 12 batteries: 3 new, 4 used (still working), 5 defective. Let X denote the number of new batteries chosen. Let Y denote the number of used batteries chosen. Find the joint probability distribution (PMF), and correlation between X and Y.

4. **A normal and another**. Let X and Z be two independent real-valued random variables, $X \sim N(0, 1)$ and Z is such that $P(Z = 1) = P(Z = -1) = 0.5$. Put $Y = XZ$.

- Determine the distribution of Y.
- Compute $\mathrm{Cov}(X, Y)$.
- Is the random vector (X, Y) multivariate Gaussian? Are the variables X and Y independent?

5. **Defects**. A detail for some machine should have dimensions X and Y, which have are distributed as $N(10.5, 0.0025)$ and $N(3.2, 0.0036)$ respectively. The detail is considered good if its length is within $[10.4, 10.6]$ and its width is within $[3.15, 3.25]$, and defective otherwise.

- If the dimensions are independent of each other, what is the probability of defects?
- How do you think this probability change, if (X, Y) is joint normal with correlation 0.9?
- If you know that (X, Y) is joint normal with correlation 0.9, find the distribution of Y conditional on the event that $X = 10.55$. Find the 0.5% and 99.5% quantiles of this conditional distribution.

6. **Maximum likelihood estimates.**

   - What is the ML estimate of λ for exponential distribution? Is it biased? Is it consistent?
   - Using the second derivative of log-likelihood, express the 99% asymptotic confidence interval for λ.

7. **Model performance**. If all components of a voice assistant are healthy, its response time has exponential distribution with mean of 0.5 seconds. If there are problems with the database (they occur 0.5% of time), response time is exponential with mean 7.

   - How to build an alarm rule that will notify me in 95% of cases of broken database?
   - If the mean response time is measured once per minute (and observations are independent), what alarm will notify me with 99.9% probability if the database has been broken for 3 minutes?
   - How often will these alarms fire?

Solutions

Stock market The price of the ABCD stock is $256 today. Every day the stock price goes either up by 25%, or down by 20%, with equal probabilities, and all days are independent from each other. A call option allows buying the stock for $300 exactly in 4 days. How much is this option worth?

Solution. The possible outcomes are listed in the table below.

Scenario	Probability	Price in 4 days.	Optimal action with the option	Profit
Up 4 times	0.5^4	625	Exercise	325
3 up, 1 down	$0.5^4 \times 4$	400	Exercise	100
2 up, 2 down	$0.5^4 \times 6$	256	Ignore	0
1 up, 3 down	$0.5^4 \times 4$	256×0.8^2	Ignore	0
4 down	0.5^4	256×0.8^{24}	Ignore	0

Therefore, the expected profit from using the option is $325 \times {}^1_{16} + 100 \times {}^4_{16} + 0 \times {}^{11}_{16} = 45.3125$. If the price of the option is also $45.3125, a risk-indifferent investor would be indifferent between buying the option. Thus, this number can be considered as the fair price of the option.
The probability of 4 ups equals 0.25^4 as a product of individual probabilities of an up in a single day, because days are independent. The probability of 2 ups and 1 down equals $0.25^4 \times 4$, because there are 4 possible scenarios with this result: UUUD, UUDU, UDUU and DUUU. It can also be calculated using the formula for binomial distribution.

Plane. The Aeroflot flight to Saint Petersburg on an Airbus has 380 seats. Suppose that any customer who booked this flight will cancel with probability $1\text{-}p = 0.1$, independently of other customers. Suppose that the company sells a fixed number N of tickets, with $N \geq 380$. Let X be the number of customers present the day of the flight.

- What is the distribution of X?
- Using the central limit theorem, approximate the probability $P(X \geq n)$: plug the mean and variance from the original distribution into normal.
- What is the largest number of bookings the company can accept if it does not want to refuse customers with probability more than 0.05.
- Suppose the company chooses N equal to the value calculated in the previous question. What is the probability that at least 35 seats are left unoccupied?

Solution. We believe that all N customers come independently with probability 0.9. Then the distribution of number of successes (customers who came) is Binomial with parameters N and 0.9

But if N is large (and 380 is already large enough), then this distribution converges to normal, because it is a sum of independent and identical variables. The mean of this distribution is 0.9N, and variance is 0.9(1 − 0.9)N = 0.09N. Then $P(X \geq n) \approx 1 - F_{N(0.9N,0.09N}(n)$.

If we want the probability of refusal to be at most 0.05, it means that $F_{N(0.9N,0.09N}(n) \geq 1 - 0.05 = 0.95$.

We know, that 95% quantile of normal distribution corresponds to the mean + 1.64 standard deviations (`scipy.stats.norm.ppf(0.95)`), and this quantile should be at most 380. So we set $0.9N + \sqrt{0.09N} \times 1.64 \leq 380$, which means $N \leq 411$ (approximately).

If we set N = 411, then the mean and standard deviation would be 370 and √37 respectively. Then probability that at least 35 seats are unoccupied the is probability that total number of passengers is no larger than 380 − 35 = 345, and it can be found with normal approximation as $F(345) \approx 2 \times 10{-}5$

(`scipy.stats.norm.cdf(345, loc=370, scale=37**0.5)`). However, with original (binomial) distribution, it is much higher: 9×10^{-5} (`scipy.stats.binom.cdf(345, n=411, p=0.9)`). This discrepancy is so high because the CLT convergence is slower in areas of low density.

NOTE: $N(\mu, 2)$ denotes normal distribution with mean and variance 2.

Batteries. Suppose that 2 batteries are randomly (uniformly) chosen without replacement from the following collection of 12 batteries: 3 new, 4 used (still working), 5 defective. Let X denote the number of new batteries chosen. Let Y denote the number of used batteries chosen. Find the joint probability distribution (PMF), and correlation between X and Y.

Solution. Let's fill the table with the joint PMF. For example, first cell denotes P(X = 0, Y = 0),which corresponds to the event that both chosen batteries are defective. The first battery is defective with probability 5/12, and the probability that the second battery is defective conditional on the first being defective is 4/11.

Another example: P(X = 0, Y = 1) corresponds to the event that one battery is defective and another is used (in any order). Both orders (defective+used or used+defective) are equiprobable, so we can calculate probability of one and double it.

X \ Y	0	1	2	Total
0	5/12*4/11	2*4/12*5/11	4/12*3/11	72/132
1	2*3/12*5/11	2*4/12*3/11	0	54/132
2	3/12*2/11	0	0	6/132
Total	56/132	64/132	12/132	1

Now the mean for X can be calculated as $\mathbb{E}X = 0 \times \frac{72}{132} + 1 \times \frac{54}{132} + 2 \times \frac{6}{132} = \frac{1}{2}$, and variance as $Var(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \left(0^2 \times \frac{72}{132} + 1^2 \times \frac{54}{132} + 2^2 \times \frac{6}{132}\right) - \left(\frac{1}{2}\right)^2 = \frac{45}{132} \approx 0.341$. Similarly, $\mathbb{E}Y = \frac{2}{3}$ and $Var(Y) \approx 0.404$.

To calculate the covariance, we need to get all products of deviations from means:

X \ Y	0	1	2
0	(0-1/2)*(0-2/3)	(0-1/2)*(1-2/3)	(0-1/2)*(2-2/3)
1	(1-1/2)*(0-2/3)	(1-1/2)*(1-2/3)	(1-1/2)*(2-2/3)
2	(2-1/2)*(0-2/3)	(2-1/2)*(1-2/3)	(2-1/2)*(2-2/3)

Now we can multiply them by the corresponding probabilities (from the first table) and add together, to obtain $Cov(X,Y) = \frac{5}{12} \times \frac{4}{11} \times \left(0 - \frac{1}{2}\right) \times \left(0 - \frac{2}{3}\right) + \cdots \approx -0.1515$. Finally, the correlation $corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = -\frac{0.15}{\sqrt{0.341 \times 0.404}} \approx -0.408$. No wonder that it is large and negative: the more we chosen new batteries, the less used batteries on average we can choose after that.

A normal and another. Let X and Z be two independent real-valued random variables, $X \sim N(0, 1)$ and Z is such that $P(Z = 1) = P(Z = -1) = 0.5$. Put $Y = XZ$.

- Determine the distribution of Y.
- Compute Cov(X, Y).
- Is the random vector (X, Y) multivariate Gaussian? Are the variables X and Y independent?

Solution. To determine the distribution of Y, we may try to calculate its CDF: $P(Y \leq y) = P(Z = 1, X \leq x \cup Z = -1, X \geq -x) = \frac{1}{2}F_X(x) + \frac{1}{2}\big(1 - F_X(-x)\big)$. But the distribution of X is symmetric with respect to 0, so $F_X(x) = 1 - F_X(-x)$, and thus $P(Y \leq y) = F_X(x)$. It means that Y has a CDF of $\mathcal{N}(0, 1)$ variable, so Y itself is $\mathcal{N}(0, 1)$.

$$Cov(X, Y) = \mathbb{E}\big((X - \mathbb{E}X)(Y - \mathbb{E}Y)\big) = \mathbb{E}(XY) = \mathbb{E}(X \times XZ) = \mathbb{E}(X^2Z)$$

Because X and Z are independent, $\mathbb{E}(X^2Z) = \mathbb{E}(X^2)\mathbb{E}(Z) = 1 \times 0 = 0$.

The random vector (X, Y) is not jointly Gaussian. Its joint PDF equals $f_{XY}(x, y) = \begin{cases} \frac{1}{2}f_X(x), if\ y = \pm x \\ 0, otherwise \end{cases}$, which is clearly not the PDF of a bivariate normal vector. X and Y are not independent, because, for example, $f_X(1) > 0$, $f_Y(2) > 0$, but $f_{XY}(1,2) = 0$.

Defects. A detail for some machine should have dimensions X and Y, which have are distributed as $N(10.5, 0.0025)$ and $N(3.2, 0.0036)$ respectively. The detail is considered good if its length is within $[10.4, 10.6]$ and its width is within $[3.15, 3.25]$, and defective otherwise.

- If the dimensions are independent of each other, what is the probability of defects?
- How do you think this probability change, if (X, Y) is joint normal with correlation 0.9?
- If you know that (X, Y) is joint normal with correlation 0.9, find the distribution of Y conditional on the event that $X=10.55$. Find the 0.5% and 99.5% quantiles of this conditional distribution.

Solution. If the dimensions are independent, then probability of *no* defects is $P(X \in [10.4, 10.6], Y \in [3.15, 3.25]) = P(X \in [10.4, 10.6])P(Y \in [3.15, 3.25]) = \big(CDF_X(10.6) - CDF_X(10.4)\big)\big(CDF_Y(3.25) - CDF_Y(3.15)\big) \approx (0.977 - 0.022)(0.798 - 0.202) \approx 0.568$. Thus, probability of defects is 0.432.

If X and Y are highly correlated (but other parameters are same as above), then probability of defect should decrease, because $P(X \in [10.4, 10.6])$ will stay the same, and $P(Y \in [3.15, 3.25] \mid X \in [10.4, 10.6])$ will be higher in this case. Indeed, e.g. if $X \approx 10.5$, then with high probability $Y \approx 3.2$, and in general lack of defects in X with high confidence implies lack of defects in Y.

If $P(X, Y)$ is joint normal, then also $P(Y|X)$ is normal with parameters $\mu_{Y|X} = \mathbb{E}(Y|X) = \mu_Y + (X - \mu_X)\rho\frac{\sigma_Y}{\sigma_X}$ and $\sigma^2_{Y|X} = \sigma^2_Y(1 - \rho^2)$. If $X = 10.55$, then $\mu_{Y|X} = 3.2 + 0.05 \times 0.9 \times \frac{0.06}{0.05} = 3.254$, and $\sigma^2_{Y|X} = 0.0036 \times (1 - 0.81) \approx 0.026^2$. The 0.5% and 99.5% quantiles of this distribution are thus approximately 3.187 and 3.321. Defect probability, by the way, in this case equals $1 - P(Y \in [3.15, 3.25]|X = 10.55) \approx 1 - (0.44 - 0.00) = 0.560$.

Maximum likelihood estimates.

1. What is the ML estimate of λ for exponential distribution? Is it biased? Is it consistent?
2. Using the second derivative of log-likelihood, express the 99% asymptotic confidence interval for λ.

Solution.

1. For exponential distribution, density equals $f(x) = \lambda e^{-\lambda x}$, and likelihood for sample is therefore $\prod_{i=1}^{n} \lambda e^{-\lambda X_i}$. The log-likelihood is $\ln L = n \ln \lambda - \lambda \sum_{i=1}^{n} X_i = n(\ln \lambda - \lambda \bar{X})$. It is a concave function (a sum of two weakly concave functions), so it has a maximum where its derivative is 0. The first derivative is $\frac{\partial \ln L}{\partial \lambda} = n\left(\frac{1}{\lambda} - \bar{X}\right)$, and it equals zero if $\lambda = \frac{1}{\bar{X}}$. Therefore, maximum likelihood estimate is $\lambda = \frac{1}{\bar{X}}$ (and it coincides with the method-of-moments estimate). It is biased (because a strictly convex transformation $\frac{1}{x}$ has been applied to unbiased estimate $\bar{X}$), but it is consistent, because $\bar{X}$ estimates $\mathbb{E}X$ consistently, and $\lambda = \frac{1}{\mathbb{E}X}$.

2. The second derivative of log-likelihood is $H = \frac{\partial^2 \ln L}{\partial \lambda^2} = -\frac{n}{\lambda^2}$, and therefore the asymptotic variance of $\hat{\lambda}$ is $\sigma^2_{\hat{\lambda}} = -H^{-1} = \frac{\lambda^2}{n} \approx \frac{\hat{\lambda}^2}{n}$. We know that $\hat{\lambda}$ in large samples is distributed almost normally, as any maximum likelihood estimate. Therefore, its 99% confidence interval can be expressed as $\left[\hat{\lambda} - 2.58\frac{\hat{\lambda}}{\sqrt{n}}, \hat{\lambda} - 2.58\frac{\hat{\lambda}}{\sqrt{n}}\right]$ (here 2.58 is 0.5% quantile of standard normal distribution).

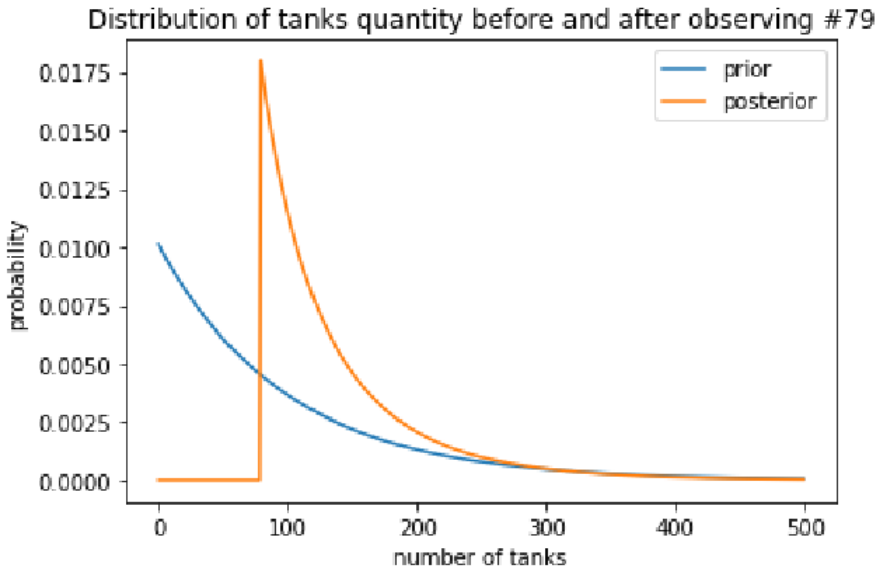
Counting tanks

1. You don't know how many tanks your enemy has, but you believe that it is on average 100. You don't know anything more about it (only that the number of tanks is non-negative), so you decide that this unknown number of tanks n has geometric distribution with mean 100. What should be the parameter p of this distribution?
2. You observed several enemy tanks, and shot one of them. After close examination, you see that it has serial number 79. Assuming that all the tanks have consecutive serial numbers, probability of seeing this particular number should be $1/n$. What is the posterior distribution for n now?
3. Calculate the mode of this posterior distribution (the most probable n).
4. Plot the prior and posterior distribution (in Python) and calculate (maybe numerically) the mean of the posterior distribution.

Solution. Geometric distribution ($p(n) = (1-p)^{n-1}p$ for natural n) has mean $\frac{1}{p}$, and therefore we can set parameter $p = \frac{1}{100}$ for the prior distribution. Thus, without any data, we believe that $p(n) = \frac{1}{100} \times \left(\frac{99}{100}\right)^{n-1}$.

After observing tank with number 79, you know for sure that $n \geq 79$, and likelihood is $p(data|n) = \frac{1}{n}$ for $n \geq 79$. The posterior distribution is thus $p(n|data) = \frac{p(n) \times p(data|n)}{p(data)} = \begin{cases} 0, if\ n < 79 \\ \frac{1}{Z} \times \frac{1}{n}\left(\frac{99}{100}\right)^n, if\ n \geq 0.79 \end{cases}$, where Z is the normalizing constant. For $n \geq 79$, it is monotonically decreasing, and thus the most probable value for n is 79.

To evaluate this distribution, we have to estimate the normalizing constant: $Z = \sum_{n=79}^{\infty} \frac{1}{n}\left(\frac{99}{100}\right)^n \approx 0.317$ (I just replaced ∞ with an arbitrary number that is large enough). The posterior mean can be evaluated as $Z = \sum_{n=79}^{\infty} n \frac{1}{Z} \frac{1}{n}\left(\frac{99}{100}\right)^n \approx 142$ tanks.



Model performance. If all components of a voice assistant are healthy, its response time has exponential distribution with mean of 0.5 seconds. If there are problems with the database (they occur 0.5% of time), response time is exponential with mean 7.

- How to build an alarm rule that will notify me in 95% of cases of broken database?
- If the mean response time is measured once per minute (and observations are independent), what alarm will notify me with 99.9% probability if the database has been broken for 3 minutes?
- How often will these alarms fire?

Solution. To make an alert that goes off in 95% cases of broken DB, we can fire it if the response time is above the 5% quantile for the broken-DB state. We can start with calculating the quantile function for exponential distribution: $\alpha = CDF(x) = 1 - e^{-\lambda x}$ corresponds to $x = -\frac{1}{\lambda}\ln(1-\alpha)$. In our case (broken DB, $\lambda = \frac{1}{7}$) the threshold is $x^* = -7\ln 0.95 = 0.359$. Unfortunately, the alarm that fires if the response time is above 0.359 seconds, will fire in $e^{-2\times 0.359} \approx 48.5\%$ cases of normal database (that is, in $0.95 \times 0.005 + 0.485 \times 0.995 = 0.49$ of all time), and thus will be practically useless.

Another alarm, that waits for 3 observations and then makes decision, can be best designed (as we know) with likelihood ratio rule. For our case, likelihood ratio is $LR = \frac{\prod_{i=1}^{3} 2e^{-2x_i}}{\prod_{i=1}^{3} \frac{1}{7}e^{-\frac{1}{7}x_i}} = 14^3 e^{-\left(2-\frac{1}{7}\right)\Sigma_{i=1}^{3} x_i}$ – and it is a function of sum (or average) of waiting times only. So we have to make decision based on average of 3 waiting times, and it has non-exponential distribution. By simulation (see the Python notebook) we find that its 0.1% quantile in case of broken DB is 0.464 seconds, and this number is a 47% quantile of our test statistic in case of no problems with DB. So our alarm again is useless.

But we can make our alarm more adequate by diminishing its recall. For example, a threshold of 1.5 seconds (average after 3 observations) corresponds to alarm probability of 97% in case of problems, and 0.44% in case of no problems, which seems a good balance (and $\frac{0.005\times 0.97}{0.005\times 0.97 + 0.995\times 0.0044} \approx 50\%$ precision).