Probability and Statistics Interview Guide

Given two independent sequences of random variables $(X_k^1)_{k=1}^{\infty}, (X_k^2)_{k=1}^{\infty}$, respectively converging to X^1, X^2 in distribution, we know that $(X_k^1 + X_k^2)_{k=1}^{\infty}$ converges to $X^1 + X^2$ in distribution. If we have now countably infinite number of independent sequences of random variables $\{(X_k^j)_{k=1}^{\infty}\}_{j=1}^{\infty}$, where each sequence converges in distribution to X^j, can we claim the sequence $S_k = \sum_{j=1}^{\infty} X_k^j$ converges to $X^1 + X^2 + \cdots$ in distribution?

Solution -

The answer is no.

Intuitively, you can see that there is no claim regrading the rate of convergence of any of the sequences, thus they can converge "as slow as we want", which will make the sum of all sequences (in the same index) ambiguous. A concrete example can be (of constant random variables, for simplicity)

$$X_k^j = 1_{k \leq j}$$

Each of the sequences of course converges to 0 since it is constant for k large enough, and their sum is $S_k = \sum_{j=1}^{\infty} X_k^j = k$, which diverge.

Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space and $\{U_n\}_{n \in \mathbb{N}}$ a collection of random variables such that $U_n \sim \text{Unif}(0, n)$ (uniform distribution) for all $n \in \mathbb{N}$. Prove that if $g : \mathbb{R} \to \mathbb{R}$ is a Borel measurable function satisfying $\lim_{x \to \infty} |g(x)| = 0$, then $g(U_n)$ converges in probability to 0.

Solution -

Given $\varepsilon > 0$, there is $a > 0$ such that

$$|g(x)| < \varepsilon, \qquad \text{whenever} \quad x > a$$

.

It follows that

$$\{|g(U_n)| > \varepsilon\} \subset \{U_n \leq a\}$$

Hence

$$P[|g(U_n)| > \varepsilon] \leq P[U_n \leq a] = \frac{\min(a, n)}{n} \xrightarrow{n \to \infty} 0$$

Two gamblers are playing coin toss game: Gambler A has (n+1) coins and B has n coins. What is the probability that A will have more heads than B if both flip all their coins.

Solution -

Firstly, we observe that the probability of A getting more heads than B is symmetrical to the probability of A getting more tails than B. This symmetry arises from the fact that the coin flips are independent events, and there is no inherent bias towards heads or tails.

Now, when A has one more coin than B (n+1 coins for A and n coins for B), A can either have more heads than B or more tails than B, but not both. This is because the total number of coin flips is fixed, and each coin can only result in either a head or a tail.

Combining these two observations, we conclude that the probability of A obtaining more heads than B is equal to the probability of A obtaining more tails than B, and since these are the only two possible outcomes, each with an equal chance, the overall probability is 1/2 or 50%. This is a result of the symmetry in the setup, highlighting that, under these conditions, A is equally likely to get more heads than B as it is to get more tails than B.

Jason throws n darts at a dartboard, aiming for the center. The second dart lands farther from the center than the first, and so does the third, the fourth, ..., the n-1 th dart. If Jason throws the nth dart aiming for the center, what is the probability that the nth throw is further from the center than the first? Assume Jason's skillfulness is constant.

Solution -

$D_1, \ldots, D_{n+1}$ are iid random variables which measure the distance of each throw to the center. We naturally assume that $P(\{D_1 = D_2\}) = 0$ (this is the case e.g., when the underlying distribution of D_1 is continuous). As a consequence $P\big(\bigcap_{i\neq j}\{D_i \neq D_j\}\big) = 1$.

The question is to compute $P\big(\{D_{n+1} > D_1\} \big| \{D_n > D_1\} \cap \ldots \cap \{D_2 > D_1\}\big)$, i.e., the ratio

$$\frac{P\big(\{D_{n+1} > D_1\} \cap \{D_n > D_1\} \cap \ldots \cap \{D_2 > D_1\}\big)}{P\big(\{D_n > D_1\} \cap \ldots \cap \{D_2 > D_1\}\big)}.$$

It suffices to compute the numerator. The denominator will then follow by replacing $n+1$ with n. Note that

$$\begin{aligned} P\big(\{D_{n+1} > D_1\} \cap \ldots \cap \{D_2 > D_1\}\big) &= P\big(\{D_1 < \min(D_2, \ldots, D_{n+1})\}\big) \\ &= P\big(\{D_1 < \min(D_2, \ldots, D_{n+1})\} \cap \bigcap_{i \neq j} \{D_i \neq D_j\}\big) \\ &= P\Big(\bigsqcup_{\sigma \in Sym(\{2,\ldots,n+1\})} \{D_1 < D_{\sigma(2)} < \ldots < D_{\sigma(n+1)}\}\Big) \\ &= \sum_{\sigma \in Sym(\{2,\ldots,n+1\})} P\big(\{D_1 < D_{\sigma(2)} < \ldots < D_{\sigma(n+1)}\}\big), \end{aligned}$$

where $Sym(\{2, \ldots, n+1\}$ is the set of permutations on $\{2, \ldots, n+1\}$.

For each fixed σ, by the iid assumption, the vector $(D_1, D_{\sigma(2)}, \ldots, D_{\sigma(n+1)})$ has the same distribution as $(D_1, D_2, \ldots, D_{n+1})$. Consequently,

$$P\big(\{D_1 < D_{\sigma(2)} < \ldots < D_{\sigma(n+1)}\}\big) = P\big(\{D_1 < D_2 < \ldots < D_{n+1}\}\big),$$

and $P\big(\{D_{n+1} > D_1\} \cap \ldots \cap \{D_2 > D_1\}\big) = n!\, P\big(\{D_1 < D_2 < \ldots < D_{n+1}\}\big)$.

Similarly,

$$\begin{aligned} 1 &= P\Big(\bigsqcup_{\sigma \in Sym(\{1,\ldots,n+1\})} \{D_{\sigma(1)} < D_{\sigma(2)} < \ldots < D_{\sigma(n+1)}\}\Big) \\ &= (n+1)!\, P\big(\{D_1 < D_2 < \ldots < D_{n+1}\}\big), \end{aligned}$$

thus $P\big(\{D_1 < D_2 < \ldots < D_{n+1}\}\big) = \frac{1}{(n+1)!}$ and finally

$$P\big(\{D_{n+1} > D_1\} \cap \ldots \cap \{D_2 > D_1\}\big) = \frac{n!}{(n+1)!} = \frac{1}{n+1}.$$

The ratio we are interested in is therefore

$$\frac{1/(n+1)}{1/n} = \frac{n}{n+1}.$$

- Let X be a random variable with probability function $\mathrm{f}(x) = \lambda \mathrm{e}^{-\lambda x}$ if $x > 0$ and $\lambda > 1$ Is a *constant*. 0 otherwise.
- Calculate the expected value of the variable $Y = e^X$ by first finding the density function of Y and applying the elementary definition of expectation.
- As a second method use the unconscious statistician's theorem.

Solution -

$$f_Y(y) = \frac{\lambda}{y}\exp\{-\lambda \log y\} = \frac{\lambda}{y}\exp\{\log y^{-\lambda}\} = \lambda y^{-\lambda-1}$$

$$f_Y(y) = \lambda y^{-(\lambda+1)}$$

$y, \lambda > 1$

thus

$$\mathbb{E}[Y] = \int_1^\infty \lambda y^{-\lambda} dy = \frac{\lambda}{\lambda - 1}$$

LOTUS

$$\mathbb{E}[Y] = \int_0^\infty \lambda e^{-\lambda x} e^x dx = \int_0^\infty \lambda e^{-(\lambda-1)x} dx = \frac{\lambda}{\lambda-1}\int_0^\infty (\lambda-1)e^{-(\lambda-1)x} dx =$$

$$\frac{\lambda}{\lambda-1}\underbrace{\int_0^\infty \theta e^{-\theta x} dx}_{=1} = \frac{\lambda}{\lambda-1}$$

A stick is broken into 3 pieces, by randomly choosing two points along its unit length, and cutting it. What is the expected length of the middle part?

Solution -
Integrate from 0 to 1, x * x/2 + (1-x) * (1-x)/2 = 1/3. Logic: if one cut is at distance x from left, with probability x, the second cut comes before it, and the expected length of the middle piece is x/2.. Similarly with prob (1-x) it, the middle piece is expected to have length (1-x)/2. Thus adding and integrating from 0 to 1.

Let X, Y be independent random variables exponentially distributed with parameter 1. Find the $E(|X-Y|)$.

Solution -

In general $|x-y| = \max(x,y) - \min(x,y)$ so that:

$$\begin{aligned}
\mathbb{E}\,|X-Y| &= \mathbb{E}\max(X,Y) - \mathbb{E}\min(X,Y) \\
&= \int_0^\infty P(\max(X,Y) > z)\,dz - \int_0^\infty P(\min(X,Y) > z)\,dz \\
&= \int_0^\infty P(X > z \vee Y > z)\,dz - \int_0^\infty P(X > z \wedge Y > z)\,dz \\
&= \int_0^\infty P(X > z) + P(Y > z) - 2P(X > z)\,P(Y > z)\,dz \\
&= \int_0^\infty 2e^{-z} - 2e^{-2z}dz \\
&= \left[-2e^{-z} + e^{-2z}\right]_0^\infty \\
&= 1
\end{aligned}$$

What is $\Pr(X+Y<0)$ where $X \sim U(0,1)$ and $Y \sim N(0,1)$? X and Y are independent

Solution -

Assuming X, Y are independent:

We want to Y-average $Pr(X < -Y)$, which at fixed Y is 0 if $Y \geq 0$, 1 if $Y < -1$ and $-Y$ otherwise. The average is

$$\int_{-\infty}^{-1} f_Y(y)dy - \int_{-1}^{0} yf(y)dy = \Phi(-1) + \tfrac{1-e^{-1/2}}{\sqrt{2\pi}} \approx 0.315.$$

Let $X_l, X_2, \ldots$ be i.i.d. with $E(X_1) = u$, $\mathrm{Var}(X_1) = \sigma^2$ and $u_2 = E[(X_1 - u)^4]$. Let $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ where $\bar{X}_n = \frac{\sum_{n=1}^{n} X_i}{n}$. Prove that $\frac{\sqrt{n}(S_n^2 - \sigma^2)}{\sqrt{u_2 - \sigma^4}} \to N(0,1)$ in distribution sense.

Solution -

We can replace X_n by $X_n - u$, hence we can assume that $u = 0$. We haved

$$\begin{aligned}\sqrt{n}\left(S_n^2 - \sigma^2\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(X_i^2 - \sigma^2\right) - 2\sqrt{n}\overline{X_n}^2 + \sqrt{n}\overline{X_n}^2 \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(X_i^2 - \sigma^2\right) - \sqrt{n}\overline{X_n}^2.\end{aligned}$$

Now, notice that

$$\sqrt{n}\overline{X_n}^2 = \frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\right)^2,$$

hence by Slutsky's lemma and the central limit theorem, $\sqrt{n}\overline{X_n}^2 \to 0$ in probability.

In order to reach the wanted conclusion, use the central limit theorem for the i.i.d. centered sequence $\left(\left(X_i^2 - \sigma^2\right)/\sqrt{u_2 - \sigma^2}\right)_{i \geqslant 1}$ (because the variance of $X_1^2 - \sigma^2$ is $\mathbb{E}\left[X_1^4\right] - \sigma^4 = u_2 - \sigma^4$).

There are 100 ropes in a bag. In each step, two **rope ends** are picked at random, tied together and put back into a bag. The process is repeated until there are no free ends.
What is the expected number of loops at the end of the process?

Solution -

Suppose, at a given moment, there are n loose ends. Once you choose the first loose end, you have $n - 1$ other loose ends to tie the rope to, each with probability $\frac{1}{n-1}$. Note that only one of the loose ends is on the same rope as the first loose end.

When $n = 200$ (remember n is the number of loose ends and there are 2 loose ends per rope), the answer is simply $\frac{1}{200-1} + \frac{1}{200-3} + \cdots + \frac{1}{200-199}$. We add, since n is arbitrary for any given moment and specifies a different time of you looping the rope.

A fair coin is tossed repeatedly until 5 consecutive heads occurs.

What is the expected number of coin tosses?

Solution -

Lets calculate it for n consecutive tosses the expected number of tosses needed.

Lets denote E_n for n consecutive heads. Now if we get one more head after E_{n-1}, then we have n consecutive heads or if it is a tail then again we have to repeat the procedure.

So for the two scenarios:

1. $E_{n-1} + 1$
2. $E_n + 1$ (1 for a tail)

So, $E_n = \frac{1}{2}(E_{n-1} + 1) + \frac{1}{2}(E_{n-1} + E_n + 1)$, so $E_n = 2E_{n-1} + 2$.

We have the general recurrence relation. Define $f(n) = E_n + 2$ with $f(0) = 2$. So,

$$\begin{aligned} f(n) &= 2f(n-1) \\ \implies f(n) &= 2^{n+1} \end{aligned}$$

Therefore, $E_n = 2^{n+1} - 2 = 2(2^n - 1)$

For $n = 5$, it will give us $2(2^5 - 1) = 62$.

A British coin has a portrait of Queen Elizabeth II on the 'Heads' side and 'ONE POUND' written on the 'tails' side, while an Indian coin has a portrait of Mahatma Gandhi on the 'Heads' side and ' 10 RUPEES' written on the 'tails' side.

These two coins are tossed simultaneously twice in sucession. The result of the first toss was 'heads' for both the coins. What is the probability that the result of second toss had a '10 RUPEES' side.

Solution -

The events are independent so the answer is simply $\frac{1}{2}$ if the coin is fair

Let X be a right continuous Markov process with left limits and generator L. Why is $f(X_t) - f(X_0) - \int Lf(X_s)ds$ a martingale for every $f \in D(L)$?

Solution -

Let $\{T_t\}$ be the semigroup of X, i.e. $T_t f(x) = E_x f(X_t)$. Then by Markov property

$$E_x[f(X_t) \mid F_s] = T_t f(X_s), \qquad s < t.$$

As $Lf = \frac{d}{du} T_u f|_{u=0}$, by the semigroup property of $\{T_t\}_t$ we have $T_r Lf = \frac{d}{du} T_u f|_{u=r}$.

So

$$E[\int_s^t Lf(X_r)dr \mid F_s] = \int_s^t T_r Lf(X_s)dr = T_t f(X_s) - T_s f(X_s).$$

Suppose S_n is a simple random walk started from $S_0 = 0$. Denote M_n to be the maximum of the walk in the first n steps, i.e. $M_n = \max_{k \leq n} S_k$. Show that M_n is not a Markov chain, but that $Y_n = M_n - S_n$ is a Markov chain.

Solution -

To show that the process M is not a Markov chain, one can consider two different paths of the process M between the times 0 and 4:

- First assume that $(M_n)_{0 \leqslant n \leqslant 4} = (0, 1, 1, 1, 2)$. Then $S_2 = 0$ hence S_3 is conditionally uniformly distributed on $\{-1, 1\}$ and the last step $1 \to 2$ has conditional probability $\frac{1}{2} \cdot P(X_4 = 1) = \frac{1}{4}$.
- Now assume that $(M_n)_{0 \leqslant n \leqslant 4} = (0, 0, 0, 1, 2)$. Then $S_3 = 1$ with full conditional probability hence the last step $1 \to 2$ has conditional probability $P(X_4 = 1) = \frac{1}{2}$.

To summarize, what this specific example shows is that the conditional probability of the step $M_3 = 1 \to M_4 = 2$ depends not only on the fact that $M_3 = 1$ but on $(M_k)_{0 \leqslant k \leqslant 2}$ as well.

To show that the process $Y = M - S$ is a Markov chain, one can note the following:

- If $Y_n = 0$, then $S_n = M_n$ hence:
  - Either $X_{n+1} = 1$ and then $M_{n+1} = M_n + 1$ and $S_{n+1} = S_n + 1$, thus $Y_{n+1} = 0$.
  - Or $X_{n+1} = -1$ and then $M_{n+1} = M_n$, $S_{n+1} = S_n - 1$ and $Y_{n+1} = 1$.
- If $Y_n \geqslant 1$, then $S_n \leqslant M_n - 1$ hence $S_{n+1} \leqslant M_n$ thus $M_{n+1} = M_n$ and $Y_{n+1} = Y_n - X_{n+1}$.

To summarize, the conclusion follows from the identity

$$Y_{n+1} = \max\{Y_n - X_{n+1}, 0\}$$

How many solutions are there to the equation

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 29$$

where $x_i, i = 1, 2, 3, 4, 5, 6$ are nonnegative integers such that

a) $x_i > 1$ for $i = 1, 2, 3, 4, 5, 6$?

c) $x_1 \leq 5$?

Solution -

Throw 2 balls in each box, to make sure i) is satisfied, and then solve $x_1 + x_2 + .. + x_6 = 17$ for each of $x_1 = 1, 2, 3$ separately, using the formula $(n + k - 1)C(k - 1)$, where n is the total sum, and k is the number of terms in the sum.

If $\hat{\theta} = \sqrt{\frac{3}{n}\sum_{i=1}^{n} Y_i^2}$ and $\{Y_i\}_{i=1}^{n} \sim U[0,\theta]$ and they are iid, is $\hat{\theta}$ in unbiased estimator of θ?

Solution -

It is biased. For $n = 1$,

$$\mathsf{E}\sqrt{3Y_1^2} = \sqrt{3}\,\mathsf{E}Y_1 = \frac{\sqrt{3}}{2}\theta < \theta.$$

However, it is consistent because $n^{-1}\sum_{i=1}^{n} Y_i^2 \to \theta^2/3$ a.s.

Say I have $X \sim \mathcal{N}(a,b)$ and $Y \sim \mathcal{N}(c,d)$. Is XY also normally distributed?

Is the answer any different if we know that X and Y are independent?

Solution -

The product of two Gaussian random variables is distributed, in general, as a linear combination of two Chi-square random variables:

$$XY = \frac{1}{4}(X+Y)^2 - \frac{1}{4}(X-Y)^2$$

Now, $X+Y$ and $X-Y$ are Gaussian random variables, so that $(X+Y)^2$ and $(X-Y)^2$ are Chi-square distributed with 1 degree of freedom.

If X and Y are both zero-mean, then

$$XY \sim c_1 Q - c_2 R$$

where $c_1 = \frac{Var(X+Y)}{4}$, $c_2 = \frac{Var(X-Y)}{4}$ and $Q, R \sim \chi_1^2$ are central.

The variables Q and R are independent if and only if $Var(X) = Var(Y)$.

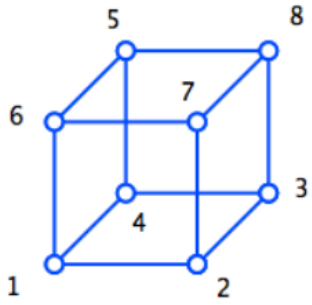
In general, Q and R are noncentral and dependent.

There is a cube and an ant is performing a random walk on the edges where it can select any of the 3 adjoining vertices with equal probability. What is the expected number of steps it needs till it reaches the diagonally opposite vertex?

Solution -

Problems such as these fall in the category of Markov chains and one way to solve this is through first step analysis.

We shall denote the vertices of the cube by numbers from 1 to 8 with 1 and 8 being the opposite ends of the body diagonal.



Let v_i denote the expected number of steps to reach the vertex numbered 8 starting at vertex numbered i.

$$v_1 = 1 + \tfrac{1}{3}(v_2 + v_4 + v_6);\ v_2 = 1 + \tfrac{1}{3}(v_1 + v_3 + v_7);\ v_3 = 1 + \tfrac{1}{3}(v_2 + v_4 + v_8);$$
$$v_4 = 1 + \tfrac{1}{3}(v_1 + v_3 + v_5);\ v_5 = 1 + \tfrac{1}{3}(v_4 + v_6 + v_8);\ v_6 = 1 + \tfrac{1}{3}(v_1 + v_5 + v_7);$$
$$v_7 = 1 + \tfrac{1}{3}(v_6 + v_2 + v_8);\ v_8 = 0;$$

Note that by symmetry you have $v_2 = v_4 = v_6$ and $v_3 = v_5 = v_7$.

Hence, $v_1 = 1 + v_2$ and $v_2 = 1 + \frac{1}{3}(v_1 + 2v_3)$ and $v_3 = 1 + \frac{2}{3}v_2$.

Solving we get

$$v_1 = 10$$
$$v_2 = v_4 = v_6 = 9$$
$$v_3 = v_5 = v_7 = 7$$

Hence, the expected number of steps to reach the diagonally opposite vertex is 10.

Suppose that an urn contains 8 red balls and 4 white balls. We draw 2 balls from the urn without replacement. If we assume that at each draw each ball in the urn is equally likely to be chosen, what is the probability that both balls are red?

Solution -

- R_1 is the event that first draw is red
- R_2 is the event that second draw is red

$$\begin{aligned} P(R_1 \cap R_2) &= P(R_1) \times P(R_2|R_1) \\ &= \frac{8}{12} \times \frac{7}{11} \\ &= \frac{56}{132} \\ &= \frac{14}{33} \end{aligned}$$

Choose a random number between 0 and 1 and record its value. Do this again and add the second number to the first number. Keep doing this until the sum of the numbers exceeds 1. What's the expected value of the number of random numbers needed to accomplish this?

Solution -

Here is a (perhaps) more elementary method. Let X be the amount of numbers you need to add until the sum exceeds 1. Then (by linearity of expectation):

$$\mathbb{E}[X] = 1 + \sum_{k \geq 1} \Pr[X > k]$$

Now $X > k$ if the sum of the first k numbers $x_1, \ldots, x_k$ is smaller than 1. This is exactly equal to the volume of the k-dimensional set:

$$\left\{ (x_1, \ldots, x_k) : \sum_{i=1}^{k} x_i \leq 1,\ x_1, \ldots, x_k \geq 0 \right\}$$

This is known as the k-dimensional *simplex*. When $k = 1$, we get a line segment of length 1. When $k = 2$, we get a right isosceles triangle with sides of length 1, so the area is $1/2$. When $k = 3$, we get a triangular pyramid (tetrahedron) with unit sides, so the volume is $1/6$. In general, the volume is $1/k!$, and so

$$\mathbb{E}[X] = 1 + \sum_{k \geq 1} \frac{1}{k!} = e.$$