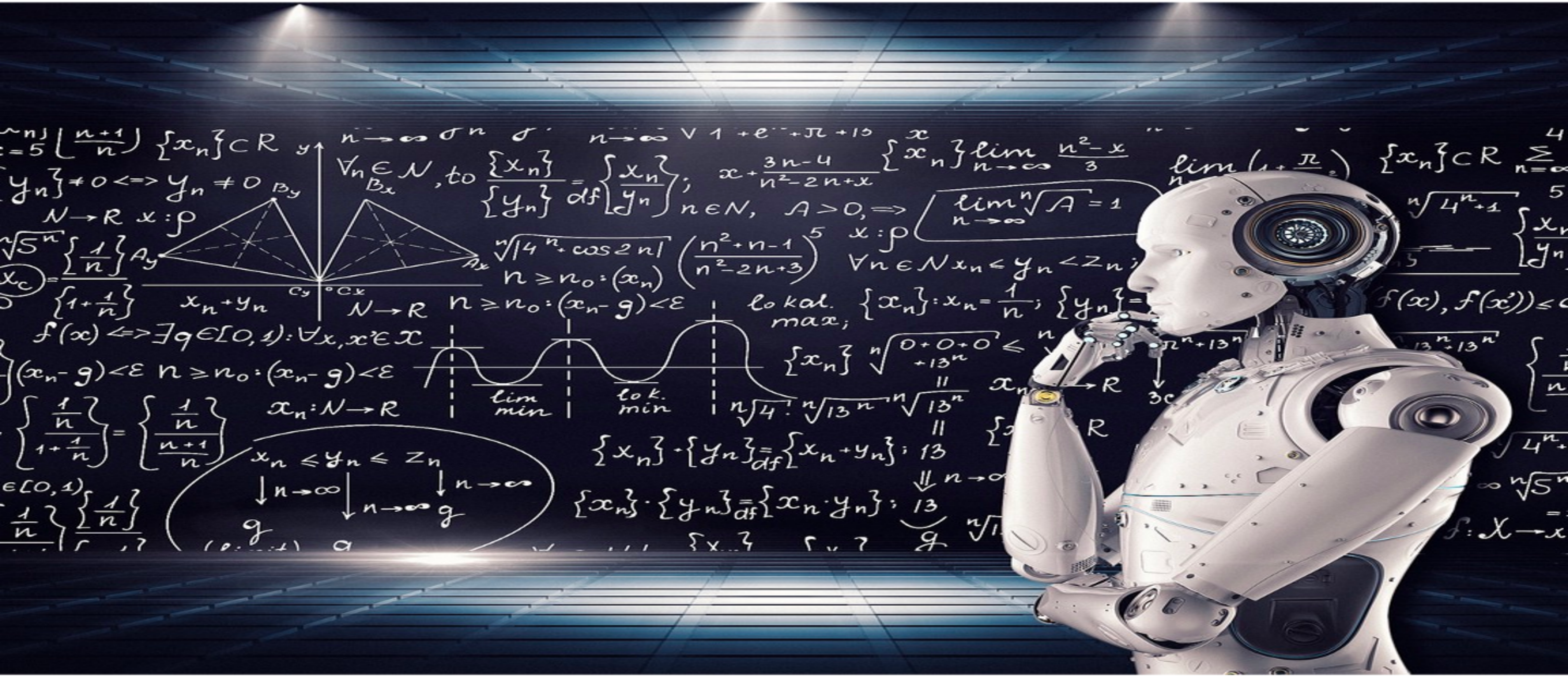
Exam Questions



Question 1

Let $\boldsymbol{X} = (X_1, \dots, X_k) \sim Multinomial\left(n, \boldsymbol{p} = (p_1, \dots, p_k)\right)$

- Derive the formula for the correlation of $\left(X_i + X_j\right)$ with X_m:

$$\rho\left(X_i + X_j, X_m\right), \qquad i \neq j \neq m$$

Recall that:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

- Derive the formula for:

$$Cov\left(e^{X_i}, e^{X_j}\right), \qquad i \neq j$$

Question 1

Let

$$\boldsymbol{X} = (X_1, X_2, X_3, X_4) \sim Multinomial\left(20, \boldsymbol{p} = \left(\frac{1}{11}, \frac{2}{11}, \frac{3}{11}, \frac{5}{11}\right)\right)$$

- Calculate $\rho(X_1 + X_3, X_4)$

Now let:

$$\boldsymbol{Y} = (X_1, X_2, X_3, X_4) \sim Multinomial\left(2, \boldsymbol{p} = \left(\frac{1}{11}, \frac{2}{11}, \frac{3}{11}, \frac{5}{11}\right)\right)$$

- Calculate the entropy of $\boldsymbol{Y}$

Question 2

Let T be the Coupon Collector random variable with n types. Namely, T represents the number of coupons you need to collect from a uniform distribution over n types before having collected at least one representative of each type.

Define $T(n,\alpha), \quad \alpha \leq 1$ as the number of coupons you need to collect from a uniform distribution over n types before having collected at least one representative from a fraction α of the types. For example, if $\alpha = 0.25$ we stop collecting when we have representatives from $\left\lceil \frac{n}{4} \right\rceil$ types.

- TRUE or FALSE:

$$E[T(98,0.1)] > 0.1E[T(98,1)]$$

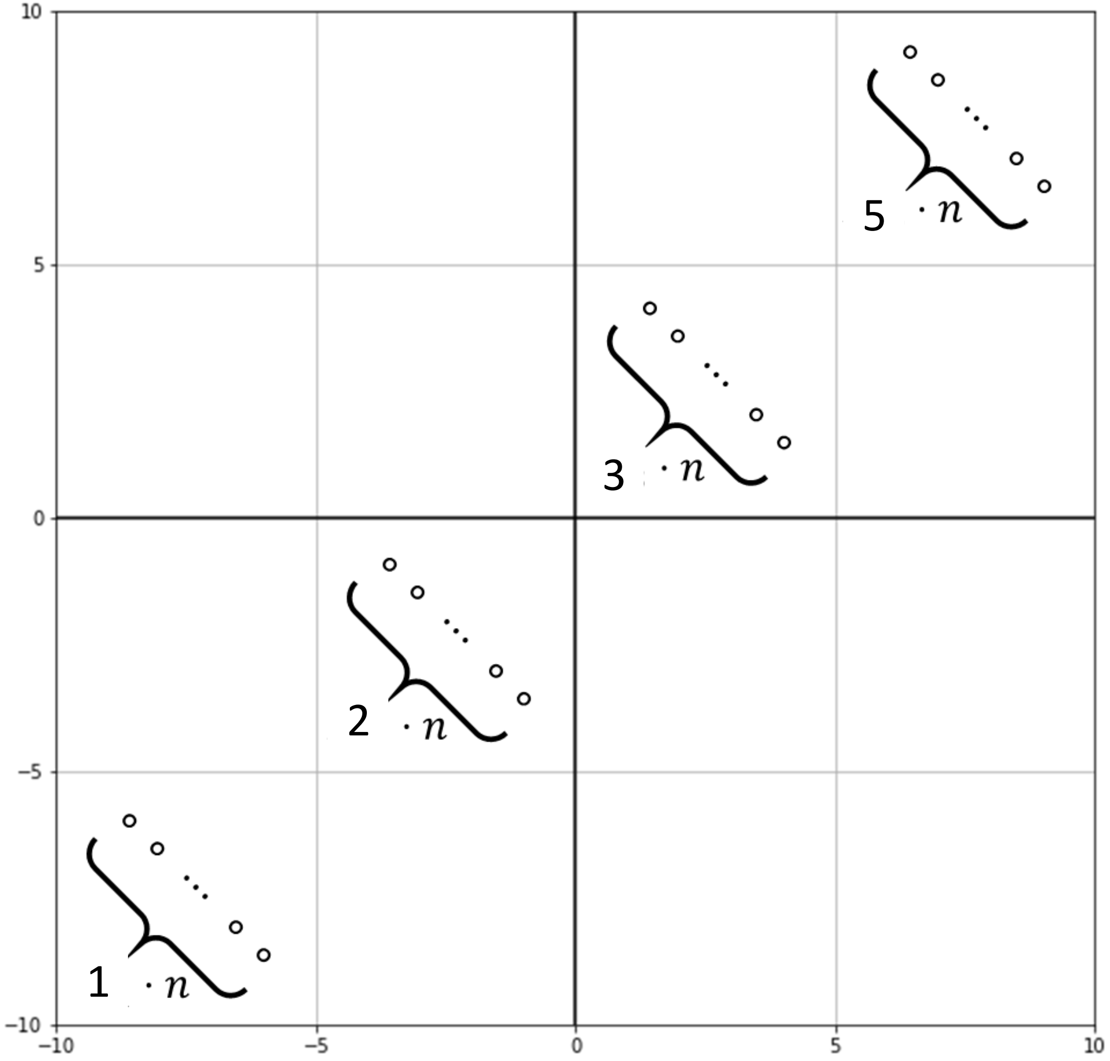
Question 3

Let $\tau(n)$ be the Kendall correlation of the dataset $\Delta(n)$.

Find:

$$\lim_{n\to\infty} \tau(n)$$

Prove your answer.



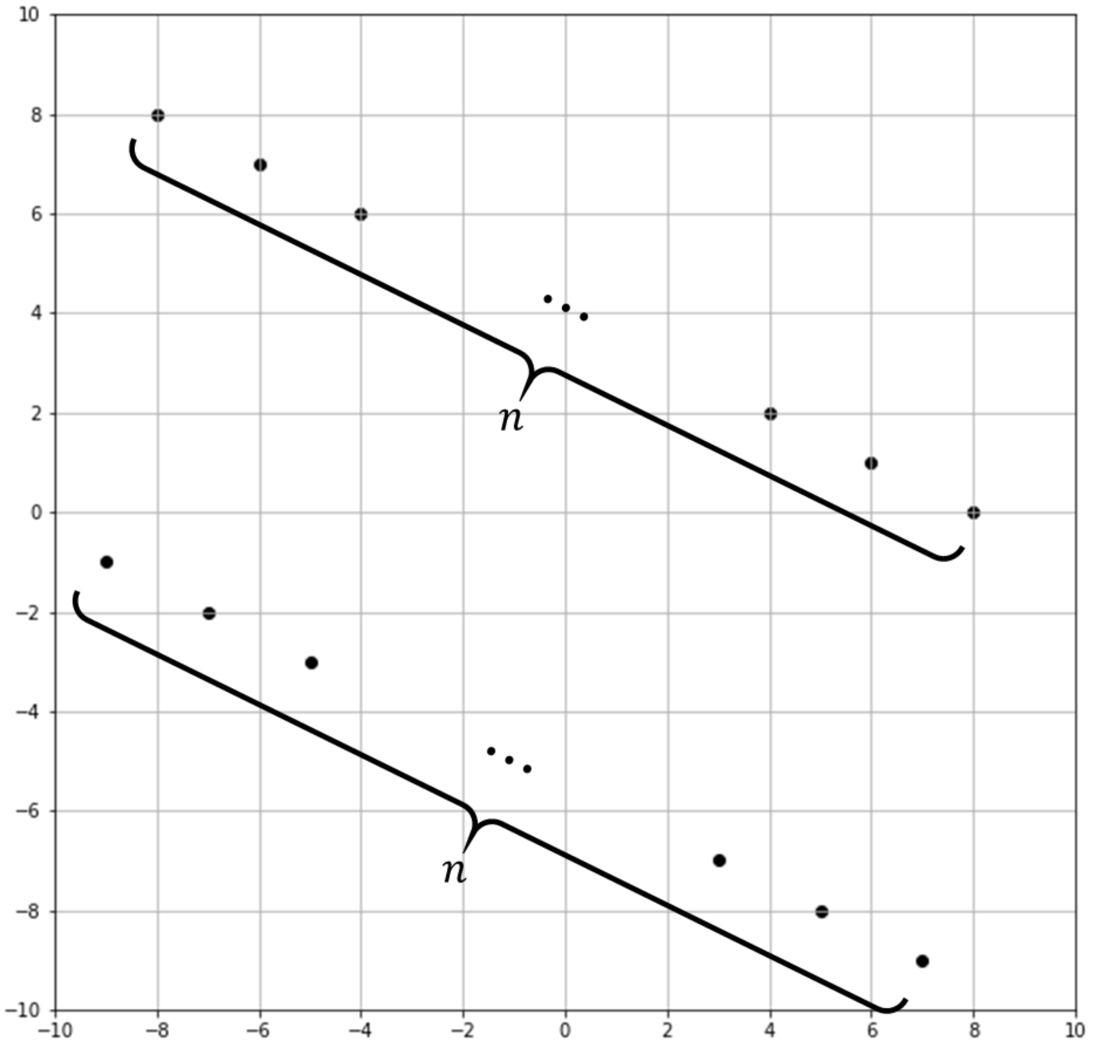
Question 3

Let $\tau(n)$ be the Kendall correlation of the dataset $\Delta(n)$.

Find:

$$\lim_{n\to\infty} \tau(n)$$

Prove your answer.



Question 3

- Let $L, U \in \mathbb{N}$, where $L < U$.
- Can there be two datasets Δ_1, Δ_2, each consisting of five integers in the range $\{100 \cdot L, 100 \cdot U\}$ that satisfy $\tau_{Kendall}(\Delta_1, \Delta_2) \leq 0$ and $\rho_{Spearman}(\Delta_1, \Delta_2) > 0$? If your answer is yes, show the data. If your answer is no, prove it.

Question 4

Let X be a random variable with a median value $Med(X) = m$.

Recall that this means that $P(X \leq m) = 0.5$.

Use Φ, as the CDF of the standard normal distribution.

- If $X \sim Normal(\mu, \sigma^2)$, what is m? Show your derivation.
- If $X \sim LogNormal(\mu, \sigma^2)$ distribution, what is m? Show your derivation.
- Let $Y = X^2, Z = X^3$.
  For each of the following, state if it's TRUE or FALSE.
  - $Med(Y) = m^2$
  - $Med(Z) = m^3$

  Prove your answer or provide a counter example.

Question 5

The Ministry Of Random Environmental Protection (MOREP) in Randomistan is performing a survey on the air quality in two cities – Random Heights (RH) and Stochastic Grove (SG).

The survey includes measurements of 100 different particle types in n_1 neighborhoods in RH and n_2 neighborhoods in SG. Following the measurements, MOREP scientists perform a one-sided WRS test on each particle type to support their assertion that **the level of particles in RH is higher than the level in SG**. The p-values of the tests are given as $p_1, p_2, \dots, p_{100}$. Assume that the p-values are ordered so that $p_1 \le p_2 \le \cdots \le p_{100}$.

Throughout this question, show your calculations or explain what information is missing to facilitate an answer to the question

Question 5

- What does p_{15} need to be so that MOREP can report at least 15 particle types to support their argument at $FDR = 0.05$?
- What does p_{15} need to be so that MOREP can report at least 15 particle types to support their argument at $FDR = 0.01$?
- What does p_{45} need to be so that MOREP can report at least 30 particle types to support their argument at $FDR = 0.05$?
- What does p_{45} need to be so that MOREP can report at least 60 particle types to support their argument at $FDR = 0.05$?