S.-T. Yau College Student Mathematics Contests 2022

Probability and Statistics

Solve every problem.

Problem 1. Let $\{X_n\}$ be a sequence of Gaussian random variables. Suppose that X is a random variable such that X_n converges to X in distribution as $n \to \infty$. Show that X is also a (possibly degenerate, *i.e.*, variance zero) Gaussian random variable.

Solution: Let $f_n(t) = \mathbb{E}\, e^{itX_n}$ be the characteristic function of X_n and $f(t) = \mathbb{E}\, e^{itX}$ be that of X. There are real numbers μ_n and σ_n such that $f_n(t) = e^{i\mu_n t - \sigma_n^2 t^2/2}$. We have $|f_n(t)|^2 \to |f(t)|^2$, hence $e^{-\sigma_n^2 t^2} \to |f(t)|^2$ for all $t \in \mathbf{R}$. Since $f(t) \neq 0$ if t is close to 0, we must have $\sigma_n^2 \to \sigma^2$ for some $\sigma \in [0, \infty)$. Now we have $e^{i\mu_n t} \to f(t)e^{\sigma^2 t^2}$ for all $t \in \mathbf{R}$ and by the dominated convergence theorem,

$$\lim_{n\to\infty} \int_0^t e^{i\mu_n s}\, ds = \int_0^t f(s) e^{\sigma^2 s^2/2}\, ds.$$

The integral on the right side does not vanish if t is close, but not equal to, 0 because the integrand is countinuous and equal to 1 at $s = 0$. On the other hand,

$$i\mu_n \int_0^t e^{i\mu_n s}\, ds = e^{i\mu_n t} - 1.$$

This gives

$$\mu_n = -i\left(f_n(t)e^{\sigma_n^2 t^2/2} - 1\right)\left(\int_0^t e^{i\mu_n s}\, ds\right)^{-1},$$

from which we see that that μ_n must converges to a finite number μ. Finally,

$$f_n(t) \to e^{i\mu t - \sigma^2 t^2/2} = f(t)$$

and X must be a (possibly denegerate) Gaussian random variable.

Problem 2. For two probability measures μ and ν on the real line $\mathbf{R}$, the total variation distance $\|\mu - \nu\|_{TV}$ is defined as

$$\|\mu - \nu\|_{TV} = \sup\{\mu(C) - \nu(C) :\ C \in \mathcal{B}(\mathbf{R})\},$$

where $\mathcal{B}(\mathbf{R})$ is the σ-algebra of Borel sets on $\mathbf{R}$. Let $\mathcal{C}(\mu, \nu)$ be the space of couplings of the probability measures μ and ν, *i.e.*, the space of $\mathbf{R}^2$ valued random variables (X, Y) defined on some (not necessarily same) probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that the marginal distributions of X and Y are μ and ν, respectively. Show that

$$\|\mu - \nu\|_{TV} = \inf\{\mathbb{P}(X \neq Y) :\ (X, Y) \in \mathcal{C}(\mu, \nu)\}.$$

For simplicity you may assume that μ and ν are absolutely continuous with respect to the Lebesgue measure on $\mathbf{R}$.

Solution: (1) Let $C \in \mathcal{B}(\mathbf{R})$ and $(X, Y) \in \mathcal{C}(\mu, \nu)$. Then

$$\mu(C) - \nu(C) = \mathbb{P}\{X \in C\} - \mathbb{P}\{Y \in C\} \leq \mathbb{P}\{X \in C,\ Y \notin C\} \leq \mathbb{P}\{X \neq Y\}.$$

Taking the supremum over $C \in \mathcal{B}(\mathbf{R})$ and then the infimum over $(X, Y) \in \mathcal{C}(\mu, \nu)$ we obtain

$$\|\mu - \nu\|_{TV} \leq \inf\{\mathbb{P}\{(X \neq Y\} :\ (X, Y) \in \mathcal{C}(\mu, \nu)\}.$$

(2) It is sufficient to a probability measure $\mathbb{P} \in \mathcal{C}(\mu, \nu)$ and a set $C \in \mathcal{B}(\mathbf{R})$ such that for $(X, Y) \in \mathbf{R}^2$ under this probability,

$$\mu(C) - \nu(C) = \mathbb{P}\{X \neq Y\}.$$

The idea is to construct $\mathbb{P}$ such that the probability $\mathbb{P}\{X = Y\}$ is the largest possible under the condition that $(X, Y) \in \mathcal{C}(\mu, \nu)$. Let $m = \mu + \nu$, or just take m to be the Lebesgue measure if μ and ν are absolutely continuous with respect to m. We have $\mu = f_1 \cdot m$ and $\nu = f_2 \cdot m$ by the Radon-Nikodym theorem. Let $f = \min\{f_1, f_2\} = f_1 \wedge f_2$. Define a probability measure $\mathbb{P}$ on $\mathbf{R}^2$ by

$$\mathbb{P}\{(X, Y) \in A \times B\} = \frac{1}{1-a} \int_{A \times B} (f_1(x) - f(x))(f_2(y) - f(y)) m(dx) m(dy) + \int_{A \cap B} f(z) m(dz).$$

Here $a = \int_{\mathbf{R}} f(z) m(dz)$ and we assume that $a < 1$; otherwise $a = 1$ and $f_1 = f_2$, and the case is trivial. Note that the first part is the product measure of $(f_1 - f) \cdot m$ and $(f_2 - f) \cdot m)$ (up to a constant) and the second part is the probability measure $f \cdot m$ on the diagonal (identified with $\mathbf{R}$) of $\mathbf{R}^2$. We have

$$\mathbb{P}\{X \in A\} = \int_A (f_1(x) - f(x)) m(dx) + \int_A f(z) m(dz) = \int_A f_1(x) m(dx) = \mu(A).$$

Similarly $\mathbb{P}\{Y \in B\} = \nu(B)$, hence $(X, Y) \in \mathcal{C}(\mu, \nu)$. On the other hand,

$$\mathbb{P}\{X \neq Y\} = \int_{\mathbf{R}} (f_1(x) - f(x)) m(dx) = 1 - a.$$

If we choose $C = \{f_1 > f_2\}$, then

$$\mu(C) - \nu(C) = \int_C (f_1(x) - f_2(x)) m(dx) = \int_{\mathbf{R}} (f_1(x) - f(x)) m(dx) = 1 - a.$$

This shows that $\mu(C) - \nu(C) = \mathbb{P}\{X \neq Y\}$.

Problem 3. We throw a fair die repeatedly and independently. Let τ_{11} be the first time the pattern 11 (two consecutive 1's) appears and τ_{12} the first time the pattern 12 (1 followed by 2) appears.

(a) Calculate the expected value $\mathbb{E}\tau_{11}$.

(b) Which is larger, $\mathbb{E}\tau_{11}$ or $\mathbb{E}\tau_{12}$? It is sufficient to give an intuitive argument to justify your answer. You can also calculate $\mathbb{E}\tau_{12}$ if you wish.

Solution:

(a) Let τ_1 be the first time the digit 1 appears. At this time, if the next result is 1, then $\tau_{11} = \tau_1 + 1$; if the next result is not 1, then the time is $\tau_1 + 1$ and we have to start all over again. This means

$$\mathbb{E}\tau_{11} = \frac{1}{6} \cdot \{\mathbb{E}\tau_1 + 1\} + \frac{5}{6} \cdot \{\mathbb{E}\tau_1 + 1 + \mathbb{E}\tau_{11}\}.$$

Solving for $\mathbb{E}\tau_{11}$ we have $\mathbb{E}\tau_{11} = 6(\mathbb{E}\tau_1 + 1)$. We need to calculate $\mathbb{E}\tau_1$. The set $\{\tau_1 \geq n\}$ is the event that that none of the first $n - 1$ results is 1, hence $\mp\{\tau_1 \geq n\} = (5/6)^{n-1}$ and

$$\mathbb{E}\tau_1 = \sum_{n=1}^{\infty} \mp\{\tau_1 \geq n\} = \sum_{n=1}^{\infty} \left(\frac{5}{6}\right)^{n-1} = 6.$$

It follows that $\mathbb{E}\tau_{11} = 6(6 + 1) = 42$.

(b) For either 11 or 12 to occur, we have to wait until the first 1 occurs. After that, if we want 11, the next digit needs to be 1; otherwise we have to start all over again (*i.e.*, waiting for the next 1). But if we want 12, the next digit needs to be 2; otherwise, we have to start all over again only if the next digit is 3 to 6 because if the next digit is 1, we have already have a start on the pattern 12. It follows that the pattern 12 has a slight advantage to occur earlier than 11. Thus we have $\mathbb{E}\tau_{12} \leq \mathbb{E}\tau_{11}$.

We can also calculate $\mathbb{E}\tau_{12}$ directly. Let τ_1 be as before and let σ be the first time a digit not equal to 1 appears. After τ_1 we wait until the first time a digit not equal to 1 appears. With probability 1/5 this digit is 2; with probability 4/5 this probability is not 2, then we have to start over again. This means that

$$\mathbb{E}\tau_{12} = \frac{1}{5} \cdot \{\mathbb{E}(\tau_1 + \sigma)\} + \frac{4}{5} \cdot \{\mathbb{E}(\tau_1 + \sigma) + \mathbb{E}\tau_{12}\}.$$

Hence $\mathbb{E}\tau_{12} = 5\mathbb{E}(\tau_1 + \sigma)$. We have seen $\mathbb{E}\tau_1 = 6$. On the other hand, $\{\sigma \geq n\}$ is the event that the first $n-1$ digits are 1, hence $\mp\{\sigma \geq n\} = (1/6)^{n-1}$ and $\mathbb{E}\,\sigma = 6/5$. It follows that

$$\mathbb{E}\tau_{12} = 5\left(6 + \frac{6}{5}\right) = 36.$$

Problem 4. Let $\{X_n\}$ be a Markov chain on a discrete state space S with transition function $p(x,y)$, $x,y \in S$. Suppose that there is a state $y_0 \in S$ and a positive number θ such that $p(x,y_0) \geq \theta$ for all $x \in S$.

(a) Show that is a positive constant $\lambda < 1$ such that for any two initial distribution μ and ν,

$$\sum_{y\in S} \left|\mathbb{P}_\mu\{X_1 = y\} - \mathbb{P}_\nu\{X_1 = y\}\right| \leq \lambda \sum_{y\in S} |\mu(y) - \nu(y)|\,.$$

(b) Show that the Markov chain has a unique stationary distribution π and

$$\sum_{y\in S} \left|\mathbb{P}_\mu\{X_n = y\} - \pi(y)\right| \leq 2\lambda^n.$$

Solution:

(a) Let $\theta = \min\{p(x,y_0)\,:\, x \in S\}$. Then $0 < \theta \leq 1$. For any two probability meausres μ and ν on the state space S, we have

$$\sum_{y\in S} |\mathbb{P}_\mu\{X_1 = y\} - \mathbb{P}_\nu\{X_1 = y\}| = \sum_{y\in S}\left|\sum_{x\in S}\{\mu(x) - \nu(x)\}\,p(x,y)\right|.$$

For the term $y = y_0$ we can replace $p(x,y_0)$ by $p(x,y_0) - \theta$ because $\sum_{x\in S}\{\mu(x) - \nu(x)\} = 1 - 1 = 0$. After this replacement, we take the absolute value of every term and exchange the order of summation. Using the fact that $p(x,y_0) - \theta \geq 0$ we have

$$\sum_{y\in S} |\mathbb{P}_\mu\{X_1 = y\} - \mathbb{P}_\nu\{X_1 = y\}| \leq \left[\sum_{y\in S} p(x,y) - \theta\right] \cdot \sum_{x\in S} |\mu(x) - \nu(x)|.$$

The first sum on the right side is $1 - \theta = \lambda < 1$. It follows that

$$\sum_{y\in S} |\mathbb{P}_\mu\{X_1 = y\} - \mathbb{P}_\nu\{X_1 = y\}| \leq \lambda \sum_{x\in S} |\mu(x) - \nu(x)|.$$

(b) Let $\mu_n(x) = \mathbb{P}_\mu\{X_n = x\}$. Then $\mu_{n+1} = \mathbb{P}_{\mu_n}\{X_1 = x\}$ and $\mu_n = \mathbb{P}_{\mu_{n-1}}\{X_1 = x\}$. By (a),

$$\sum_{x\in S} |\mu_{n+1}(x) - \mu_n(x)| \leq \lambda \sum_{x\in S} |\mu_n(x) - \mu_{n-1}(x)|.$$

It follows that

$$\sum_{x\in S} |\mu_{n+1}(x) - \mu_n(x)| \leq \lambda^n \sum_{x\in S} |\mu_1(x) - \mu(x)| \leq 2\lambda^n.$$

Since $0 \leq \lambda < 1$, the distributions μ_n converges to a distribution π, which is obviously stationary. We have by the same argument,

$$\sum_{y\in S} |\mathbb{P}_\mu\{X_n = y\} - \pi(y)| = \sum_{y\in S} |\mathbb{P}_\mu\{X_n = y\} - \mathbb{P}_\pi\{X_n = y\}| \leq 2\lambda^n.$$

If σ is another stationary distribution, then

$$\sum_{y\in S} |\sigma(y) - \pi(y)| = \sum_{y\in S} |\mathbb{P}_\sigma\{X_n = y\} - \mathbb{P}_\pi\{X_n = y\}| \leq 2\lambda^n \longrightarrow 0.$$

Hence a stationary distribtuion of the Markov chain must be unique.

Problem 5. Consider a linear regression model with p predictors and n observations:

$$\mathbf{Y} = X\beta + \mathbf{e},$$

where $X_{n\times p}$ is the design matrix, β is the unknown coefficient vector, and the random error vector $\mathbf{e}$ has a multivariate normal distribution with mean zero and $\text{Var}(\mathbf{e}) = \sigma^2 I_n$ ($\sigma^2 > 0$ unknown and I_n is the identity matrix). Here $\text{rank}(X) = k \le p$, p may or may not be greater than n, but we assume $n - k > 1$. Let $\mathbf{x}_1 = (x_{1,1}, \dots, x_{1,p})$ be the first row of X and define

$$\gamma = \frac{\mathbf{x}_1\beta}{\sigma}.$$

Find the uniformly minimum variance unbiased estimator (UMVUE) of γ or prove it does not exist.

Solution: The key points in the solution are the following.

- **(i)** Any least squares estimator, say $\hat\beta$, of β is independent of $\hat\sigma^2 = \|\mathbf{Y} - X\hat\beta\|^2/(n-k)$.
- **(ii)** $\mathbf{x}_1\beta$ is clearly estimable.
- **(iii)** Based on (i) and (ii), we can constructor an unbiased estimator, say $\hat\gamma$, of γ in terms of $\hat\beta$ and $\hat\sigma^2$, and consequently we know the estimator is a function of $X^T\mathbf{Y}$ and $\|\mathbf{Y} - X\hat\beta\|^2$.
- **(iv)** In fact, $(X^T\mathbf{Y}, \|\mathbf{Y} - X\hat\beta\|^2)$ is a complete and sufficient statistic and we conclude $\hat\gamma$ is the UMVUE of γ. More details are given below.

Let $\hat\beta = (X^TX)^- X^T Y$ be a least squares estimator of β, where $(X^TX)^-$ denotes any generalized inverse of X^TX. Let $\theta = \mathbf{x}_1\beta$, which is clearly estimable. By Gauss-Markov Theorem, we know $\hat\theta =: \mathbf{x}_1\hat\beta$ is the best linear unbiased estimator of θ. For the unbiased estimator $\hat\sigma^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2/(n-k)$, we know $(n-k)\hat\sigma^2/\sigma^2$ has χ^2_{n-k} distribution, which belongs to the Gamma family. Thus, it is readily seen that $E(1/\hat\sigma) = C/\sigma$, where C is a known constant ($C = \sqrt{n-k}\,\Gamma(\frac{n-k-1}{2})/(\sqrt{2}\,\Gamma(\frac{n-k}{2}))$).

Let $\hat\gamma = \hat\theta/(C\hat\sigma)$. Let $H = X(X^TX)^-X^T$ denote the projection matrix. Clearly, $(I_n - H)X = 0$, which implies $\text{Cov}((X^TX)^-X^T\mathbf{Y}, (I_n - H)\mathbf{Y}) = 0$. Together with the Gaussian error assumption, we know $(X^TX)^-X^T\mathbf{Y}$ and $(I_n - H)\mathbf{Y}$ are independent. It follows that $\hat\beta$ (any choice) and $\hat\sigma^2$ are independent. This leads to the unbiasedness of $\hat\gamma$.

With elementary simplifications, based on basic exponential family properties, we see that $T = (X^TY, \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2)$ is a complete and sufficient statistic. We conclude that $\hat\gamma$ is indeed unbiased and a function of a complete and sufficient statistic, and hence it must be the UMVUE of γ.

Problem 6. Let $X_1, \dots, X_{2022}$ be independent random variables with $X_i \sim N(\theta_i, i^2)$, $1 \le i \le 2022$. For estimating the unknown mean vector $\boldsymbol{\theta} \in R^{2022}$, consider the loss function $L(\boldsymbol{\theta}, \mathbf{d}) = \sum_{i=1}^{2022}(d_i - \theta_i)^2/i^2$. Prove that $\mathbf{X} = (X_1, \dots, X_{2022})$ is a minimax estimator of $\boldsymbol{\theta}$.

Recall: *If* $Y|\mu \sim N(\mu, \sigma^2)$ *and* $\mu \sim N(\mu_0, \sigma_0^2)$ *then* $\mu|Y = y \sim N\left(\frac{\mu_0/\sigma_0^2 + y/\sigma^2}{1/\sigma_0^2 + 1/\sigma^2}, \frac{1}{1/\sigma_0^2 + 1/\sigma^2}\right)$.

Solution: We show $\mathbf{X}$, as an equalizer (constant risk), achieves the limit of Bayes risks under certain priors. First, consider independent priors $\theta_i \sim N(0, \tau^2)$, $1 \le i \le 2022$. Then, the Bayes estimator δ_τ has the i-th component (estimator of θ_i) being the posterior mean $E_\tau(\theta_i|\mathbf{X}) = \frac{X_i/i^2}{1/\tau^2 + 1/i^2}$. The associated Bayes risk is $R_\tau(\delta_\tau) = \sum_{i=1}^{2022} i^{-2}\frac{1}{1/\tau^2 + 1/i^2}$. Clearly, as $\tau \to \infty$, $R_\tau(\delta_\tau) \to \sum_{i=1}^{2022} 1 = 2022$, which is identical to the Bayes risk of $\mathbf{X}$. This implies that $N(0, \tau^2)$ with $\tau \to \infty$ gives a least favorable sequence of priors and $\mathbf{X}$ is minimax.