MA930 Data Analysis. Class test (2020)

The exam comprises two pages with a total of five questions.
Full marks are given for correct answers to each of the five questions.
Note: Do not use calculators or computers for any numerics. Answers should be handwritten.

Q1. Generation of random numbers
The continuous random variable x can take values between 0 and 1, over which range it has the probability density

$$f(x) = 4x(1 - x^2). \tag{1}$$

Provide the formula $x = g(y)$ which will generate correctly distributed random numbers x from a flat distribution for y between 0 and 1.

Total marks: 4

Q2. Unbiased sampling of the third cumulant
The third cumulant of a random variable x is defined as

$$K_3 = \langle (x - \langle x \rangle)^3 \rangle = \langle x^3 \rangle - 3\langle x^2 \rangle \langle x \rangle + 2\langle x \rangle^3. \tag{2}$$

When a finite sample of n independent identical random variables $\{x_i\}$ are drawn, care needs to be taken in the construction of the sample estimators so they are unbiased.

(a) Argue why the following expectations are equal

$$E = \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \langle x_i x_j x_k \rangle = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \langle x_i x_j x_k \rangle \tag{3}$$

and provide the answer in terms of $\langle x \rangle$, $\langle x^2 \rangle$, $\langle x^3 \rangle$ and n.

(b) Demonstrate that

$$U_3 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (x_i - \bar{x})^3 \quad \text{where} \quad \bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j \tag{4}$$

provides an unbiased estimator for the third cumulant.

Total marks: 7

Q3. Frequentist statistics

(a) A cheap but not fully reliable test for dust allergy is compared to a definitive diagnosis. One hundred people suspected of having allergies are tested: 90 people who genuinely have an allergy are correctly identified by the test; 3 people who don't have allergies are also correctly identified by the test; and the test predicts that 92 of those tested have an allergy. What are the false positive and false negative error rates?

(b) You decide to test if a coin is biased towards heads using a significance test. You flip the coin ten times and get 9 heads and 1 tail. Without using a computer, find a sufficient approximation for the p-value that allows you to state if the result is significant at the 5% or 1% levels.

Total marks: 4

Q4. Autoregression model with multiplicative noise
Consider a stochastic process y_t in discrete time that obeys the following iterative relation

$$y_t = y_{t-1}^{\phi} e^{\epsilon_t} \tag{5}$$

where $0 < \phi < 1$ and $\{\epsilon_t\}$ are independent Gaussian random numbers with zero mean and unit variance. This process has been going since long in the past. Demonstrate that the steady-state distribution of y_t can be written

$$f(y_t) = \frac{1}{y_t\sqrt{2\pi\sigma^2}} e^{-(\log(y_t))^2/2\sigma^2} \tag{6}$$

and give the form of σ^2 as a function of ϕ.

Total marks: 4

Q5. Finite-time correction for Wiener-Khinchin theorem
The finite-time power spectrum of a stochastic signal that has an autocovariance $A(t)$ is written

$$S_T(\omega) = \frac{1}{T}\int_{-T/2}^{T/2}\int_{-T/2}^{T/2} dt dt' e^{-i\omega(t-t')} A(t-t'). \tag{7}$$

In class, the transformation $t'' = t - t'$ and $r = (t + t')/2$ was used to map the integral domain onto a diamond shape in the (t'', r) plane. An approximation was then made in which the t'' integral was extended to minus and plus infinity. The resulting rectangular integral yields the Wiener-Khinchin result

$$S(\omega) = \frac{1}{T}\int_{-T/2}^{T/2} dr \int_{-\infty}^{\infty} dt'' e^{-i\omega t''} A(t'') = \tilde{A}(\omega). \tag{8}$$

In the limit $T \to \infty$ we have $S_T \to S$; however, for finite T a difference remains. For this question you will examine this difference for the particular case where the stochastic signal is an Ornstein-Uhlenbeck process with autocovariance $A(t) = e^{-|t|/\tau}$ where τ is a positive constant. By considering the regions of integration, show that the difference $D = S(\omega) - S_T(\omega)$ is

$$D = \frac{1}{T}\left(\frac{1 - e^{-T\lambda}}{\lambda^2} + \frac{1 - e^{-T\lambda_*}}{\lambda_*^2}\right), \tag{9}$$

where $\lambda = i\omega + 1/\tau$ and λ_* is its complex conjuate, and therefore the correction D decays as $1/T$ for large T.

Total marks: 6

EXAM END

A1. Generation of random numbers

The cumulative distribution is

$$C = 2x^2 - x^4 \tag{10}$$

To generate random numbers we need the inverse of this so must solve

$$x^4 - 2x^2 + y = 0 \tag{11}$$

this has the solutions

$$x^2 = 1 \pm \sqrt{1 - y} \tag{12}$$

When y is small x is also small, so it is the lower root we want, yielding

$$x = \sqrt{1 - \sqrt{1 - y}}. \tag{13}$$

Total marks 4

A2. Unbiased sampling of the third cumulant

(a) Nothing special about i in the second equation, so there are just n of these in the first equation but with an extra normalisation so they will be equal. Now, taking the second equation we have three cases adding to n^2 elements: (i) 1 where all are equal; (ii) $3(n-1)$ where $i = j \neq k$ and two other permutations and (iii) $(n-1)(n-2)$ where they are all different, so that

$$E = \frac{1}{n^2}\left(\langle x^3\rangle + 3(n-1)\langle x^2\rangle\langle x\rangle + (n-1)(n-2)\langle x\rangle^3\right). \tag{14}$$

(b) Let's call the biased estimator B_3 where

$$\langle B_3\rangle = \frac{1}{n}\sum_{i=1}^{n}\langle (x_i - \bar{x})^3\rangle = \langle (x_i - \bar{x})^3\rangle = \langle x^3\rangle - \frac{3}{n}\sum_{j=1}^{n}\langle x_i^2 x_j\rangle + 2E \tag{15}$$

where for the last term we have used the result of the previous part of the question. The sum in the second term is just $\langle x^3\rangle + (n-1)\langle x^2\rangle\langle x\rangle$ and so we can write

$$\langle B_3\rangle = \langle x^3\rangle - \frac{3}{n}\left(\langle x^3\rangle + (n-1)\langle x^2\rangle\langle x\rangle\right) + \frac{2}{n^2}\left(\langle x^3\rangle + 3(n-1)\langle x^2\rangle\langle x\rangle + (n-1)(n-2)\langle x\rangle^3\right). \tag{16}$$

Gathering terms we have

$$\langle B_3\rangle = \langle x^3\rangle\left(1 - \frac{3}{n} + \frac{2}{n^2}\right) - 3\langle x^2\rangle\langle x\rangle\left(\frac{(n-1)}{n} - \frac{2(n-1)}{n^2}\right) + 2\langle x\rangle^3\frac{1}{n^2}(n-1)(n-2) \tag{17}$$

so that

$$\langle B_3\rangle = \frac{(n-1)(n-2)}{n^2}\left(\langle x^3\rangle - 3\langle x^2\rangle\langle x\rangle + 2\langle x\rangle^3\right) = \frac{(n-1)(n-2)}{n^2}K_3. \tag{18}$$

Hence, an unbiased estimator must be $n^2 B_3/(n-1)(n-2)$ or

$$K_3 = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\langle (x_i - \bar{x})^3\rangle. \tag{19}$$

Total marks 7

A3. Frequentist statistics

(a) Construct a table for the data

	really allergic	not really allergic
predicted allergic	90	2
predicted not allergic	5	3

where we have inferred that 5 people are predicted by the test to not have an allergy, even though they do (because it must add up to 100)

False negative rate is $P(\text{predicted not allergic}|\text{really allergic}) = 5/95 = 1/19$
False positive rate is $P(\text{predicted allergic}|\text{not really allergic}) = 2/5$

(b) Events that are as extreme or more extreme are 9 heads and 10 heads. The p-value is the sum of their probabilities under the null hypothesis

$$\text{p-value} = 10(0.5)^{10} + (0.5)^{10} = 11(0.5)^{10} \tag{20}$$

Now $(1/2)^{10} = 1/1024$ and $1024 = 1100 - 76$. So

$$\text{p-value} = \frac{11}{1100-76} = \frac{11}{1100}\frac{1}{1-76/1100} = 0.01(1 + \frac{76}{1100} + \cdots) \simeq 0.0107 \tag{21}$$

where we have used $76/1100 \simeq 77/1100$. The result is therefore significant at the 5% level but just not significant at the 1% level.

Total marks 4

A4. Autoregression model with multiplicative noise

We take the log of the equation and introduce $x_t = \log(y_t)$ to yield

$$x_t = \phi x_{t-1} + \epsilon_t. \tag{22}$$

In the steady-state (which exists for the values of ϕ given) we have $\mathrm{Var}(x) = 1/(1-\phi^2) = \sigma^2$ from the lecture notes. Additionally, x_t must be Gaussian distributed as it is formed from the addition of Gaussian random numbers. Hence the distribution of x_t is a normal with zero mean and σ^2 variance. There is a logarithmic relation between x and y so that $dx/dy = 1/y$. Hence, $f_y(y) = f_x(\log(y))/y$ and the desired result is obtained.

Total marks 4

A5. Finite-time correction for the Wiener-Khinchin theorem

(a) The transformed integral is a diamond shape, whereas the approximation is a band including the diamond shape. The difference is the integral above and below, from the edge of the diamond to plus infinity, and then the same on the left hand side to minus infinity. It is clear that the integral on the right above and below are the same, and similarly those to the left are the complex conjugate. Hence the only integral that needs to be performed is

$$I = \frac{1}{T}\int_0^{T/2} dr \int_{T-2r}^{\infty} dt'' A(t'')e^{-i\omega t''} \tag{23}$$

and the answer will be $S - S_T = 2I + 2I_*$ where the $*$ denotes the conjugate. For the unit-variance O-U process, and defining $\lambda = i\omega + 1/\tau$, we have

$$I = \frac{1}{T}\int_0^{T/2} dr \int_{T-2r}^{\infty} dt'' e^{-\lambda t''} = \frac{1}{T\lambda}\int_0^{T/2} dr e^{\lambda(2r-T)} = \frac{1}{2T\lambda^2}\left(1 - e^{-\lambda T}\right) \tag{24}$$

which together with the other components $2I + 2I_*$ gives the answer D given in the question.

Total marks 6