STATISTICAL METHODS FOR FINANCE: EXAMINATION SOLUTIONS 2017-18

Q1 (*Fisher score function; Fisher information*).
(i) *The Fisher score function.* This is the derivative of the log-likelihood:

$$s(\theta) := \partial \log L(\theta)/\partial\theta = \ell'(\theta). \qquad (s) \quad [2]$$

(ii) *The Fisher information (per reading). Under the regularity conditions below*, this is defined as (for the equalities, see (b) below)

$$I(\theta) := E[\{\ell'(\theta)\}^2] = -E[\ell''(\theta)] : \qquad I(\theta) = E[s^2(\theta)] = -E[s'(\theta)], \quad (I) \ [3]$$

(iii) The joint density $f = f(x_1, \dots, x_n; \theta) = f(x; \theta)$. This integrates to 1: $\int f(x;\theta)dx = 1$ (with dx n-dimensional Lebesgue measure): $\int f = 1$. We assume $f(x;\theta)$ smooth enough for use to differentiate under the integral sign (w.r.t. dx) w.r.t. θ, *twice* (once to get the mean $E[s(\theta)] = 0$, twice to get the variance $var(s(\theta)) = I(\theta)$). Doing this once gives the mean:

$$\int \frac{\partial f}{\partial\theta} = \frac{\partial}{\partial\theta}\int f = \frac{\partial}{\partial\theta}1 = 0 : \quad \int\Big(\frac{1}{f}\frac{\partial f}{\partial\theta}\Big).f = 0 : \quad \int\Big(\frac{\partial}{\partial\theta}\log f\Big).f = 0.$$

Now $E[g(X)] = \int g(x)f(x;\theta)dx = \int gf$, so

$$E\Big[\frac{\partial \log L}{\partial\theta}\Big] = 0 : \ E\Big[\frac{\partial \ell}{\partial\theta}\Big] = 0 : \ E[\ell'(\theta)] = 0 : \ E[s(\theta)] = 0. \qquad (a) \ [5]$$

Differentiating under the integral sign wrt θ again gives the variance:

$$\frac{\partial}{\partial\theta}\int\Big(\frac{1}{f}\frac{\partial f}{\partial\theta}\Big).f = 0 : \int \frac{\partial}{\partial\theta}\Big[\Big(\frac{1}{f}\frac{\partial f}{\partial\theta}\Big).f\Big] = 0 : \int\Big[\Big(\frac{1}{f}\frac{\partial f}{\partial\theta}\Big)\frac{\partial f}{\partial\theta} + f\frac{\partial}{\partial\theta}\Big(\frac{1}{f}\frac{\partial f}{\partial\theta}\Big)\Big] = 0.$$

As the bracket in the second term is $\partial \log f/\partial\theta$, this says

$$\int\Big[\Big(\frac{1}{f}\frac{\partial f}{\partial\theta}\Big)^2 + \frac{\partial}{\partial\theta}\Big(\frac{\partial \log f}{\partial\theta}\Big)\Big]f = 0, \qquad \int\Big[\Big(\frac{\partial \log f}{\partial\theta}\Big)^2 + \frac{\partial^2}{\partial\theta^2}(\log f)\Big]f = 0 :$$

$$E\Big[\Big(\frac{\partial}{\partial\theta}\log L\Big)^2 + \frac{\partial^2}{\partial\theta^2}\log L\Big] = 0 : \ E[\{\ell'(\theta)\}^2 + \ell''(\theta)] = 0 : \ E[s(\theta)^2 + s'(\theta)] = 0.$$

(b) [5]

By (a) and (I): the score fn $s(\theta) := \ell'(\theta)$ has mean 0 and variance $I(\theta)$. [2]
(iv) These results are useful in establishing the concept of *information* in Statistics, and in particular in proving the *Cramér-Rao inequality*, giving the minimum-variance lower bound for the variance of unbiased estimators. [3]
[Seen, lectures]

Q2 (*Efficiency; asymptotic efficiency*).
(i)

$$\ell = -n \log 2 - \sum |x_i - \theta|.$$

To maximise this – i.e. minimise $\sum |x_i - \theta|$ – draw a graph. From this, the sum is minimised by $\theta = Med$, and increases linearly on either side of the sample median (a.s. unique if the sample size is odd). So the MLE is $\hat{\mu} = Med$ (and with sample size even: anything between the 'central two' – irrelevant for large samples).
(ii) With one reading, as above, ℓ decreases with slope -1 to the right of Med, slope +1 to the left of Med. So $(\ell')^2 = 1$ (except at $\lambda = Med$, where the derivative is not defined – but we are going to integrate, and so can neglect null sets, e.g. single points). So $I = \int (\partial \log f/\partial\theta)^2 f = \int f = 1$, as f is a density. So the CR bound is $1/n$.

We are given that Med is asymptotically normal, and that its mean is $med = \theta$, so Med is asymptotically unbiased. By symmetry, the population median is $med = \theta$, where the density is $\frac{1}{2}$. So $4f(med)^2 = 1$, and the asymptotic variance of the sample median is $1/n$, the CR bound, so Med is also asymptotically efficient. **[10]**
. (ii)

$$f(x;\mu) = \frac{1}{\pi(1 + (x-\mu)^2)}, \qquad \ell = \log f = c - \log[1 + (x-\mu)^2],$$

$$\ell' = \frac{2(x-\mu)}{1 + (x-\mu)^2}, \qquad \ell'(\mathbf{x};\theta) = 2\sum_1^n \frac{(x_i - \mu)}{1 + (x_i - \mu)^2}.$$

(Note: efficiency iff ℓ' factorises in the form $\ell'(\mathbf{x};\theta) = A(\theta)(u(\mathbf{x}) - \theta)$. The likelihood here does not factorise, so there is no efficient estimator.)

The information per reading is

$$E[(\ell')^2] = \int (\partial f/\partial\mu)^2 f = \frac{4}{\pi}\int \frac{(x-\mu)^2}{[1+(x-\mu)^2]^3}dx = \frac{4}{\pi}\int \frac{x^2}{[1+x^2]^3}dx = \frac{4}{\pi}I,$$

say. Given $I = \pi/8$, the information per reading as $\frac{1}{2}$. So the information in a sample of size n is $n/2$, and the MLE has asymptotic variance $2/n$. As in (i), the sample median has asymptotic variance $\pi^2/4n$. So the asymptotic efficiency is their ratio, $8/\pi^2 \sim 81\%$ **[10]**
. [Seen – Problems]

Q3 (*Independence of linear and quadratic forms*).
(i) The joint CF of $A\mathbf{x}$ and $B\mathbf{x}$ is

$$\phi(\mathbf{u},\mathbf{v}) := E\exp\{i\mathbf{u}^T A\mathbf{x} + i\mathbf{v}^T B\mathbf{x}\} = E\exp\{i(A^T\mathbf{u}+B^T\mathbf{v})^T\mathbf{x}\}.$$

This is the CF of $\mathbf{x}$ at argument $\mathbf{t} = A^T\mathbf{u}+B^T\mathbf{v}$, so

$$\phi(\mathbf{u},\mathbf{v}) = \exp\{i(\mathbf{u}^T A + \mathbf{v}^T B)\mu - \frac{1}{2}(A^T\mathbf{u}+B^T\mathbf{v})^T\Sigma(A^T\mathbf{u}+B^T\mathbf{v})\}$$

$$= \exp\{i(\mathbf{u}^T A+\mathbf{v}^T B)\mu - \frac{1}{2}[\mathbf{u}^T A\Sigma A^T\mathbf{u}+\mathbf{u}^T A\Sigma B^T\mathbf{v}+\mathbf{v}^T B\Sigma A^T\mathbf{u}+\mathbf{v}^T B\Sigma B^T\mathbf{v}]\}.$$

This factorises into a product of functions of $\mathbf{u}$ and $\mathbf{v}$ iff the two cross-terms in $\mathbf{u}$ and $\mathbf{v}$ vanish, that is, iff $A\Sigma B^T = 0$ and $B\Sigma A^T = 0$; by symmetry of Σ, these are equivalent. If $\Sigma = \sigma^2 I$, the condition is $AB^T = 0$. **[5]**
(ii) For P a symmetric projection,

$$x^T P x = x^T P^T P x = (Px)^T (Px),$$

which simplifies from *quadratic forms* to *linear* forms. So: if $x^T P_1 x$, $x^T P_2 x$ are quadratic forms with P_i projections, they are independent iff

$$P_1 P_2 = 0 :$$

P_1, P_2 are *orthogonal projections* – i.e. their ranges are orthogonal subspaces,

$$(P_1 x).(P_2 x) = 0 \quad \forall\ x: \quad x^T P_1^T P_2 x = 0 \quad \forall x; \quad P_1^T P_2 = 0 \quad \forall x; \quad P_1 P_2 = 0$$

for P_i symmetric. **[3]**
(iii) P and $I - P$ are (orthogonal, symmetric) projections:

$$(I-P)^2 = I-2P+P^2 = 1-2P+P = I-P; \quad P(I-P) = P-P^2 = P-P = 0.$$

[2]
(iv) If λ is an eigenvalue of A, λ^2 is an eigenvalue of A^2. So if a projection P has eigenvalue λ, $\lambda^2 = \lambda$: $\lambda = 0$ or 1. For a projection, the trace – the sum of the eigenvalues – is the number of non-zero e-values; this is the rank. So: *For a projection, the eigenvalues are 0 or 1, and the trace is the rank.*

By Spectral Decomposition, a symmetric projection matrix P can be diagonalised by an orthogonal transformation O to a diagonal matrix D:

$$O^T P O = D, \qquad P = ODO^T;$$

the diagonal entries d_{ii} are 0 or 1 (above); re-order so that the 1s come first. So with $y := O^T x$,

$$x^T P x = x^T O D O^T x = y^T D y = y_1^2 + \ldots + y_r^2.$$

Normality is preserved under orthogonal transformations (the density depends on $x_1^2 + \cdots + x_n^2 = \|x\|^2 = \|y\|^2$), so also $y \sim N(0, \sigma^2 I)$. So $y_1^2 + \ldots + y_r^2$ is σ^2 times the sum of r independent squares of standard normal variates, and this sum is $\chi^2(r)$ (by definition of chi-square):

$$x^T P x \sim \sigma^2 \chi^2(r).$$

If P has rank r, $I - P$ has rank $n - r$ (with n the sample size):

$$x^T (I - P) x \sim \sigma^2 \chi^2(n - r),$$

and the two quadratic forms are independent. **[5]**
(v) The result above gives independence of SSE and SSR, the sums of squares for error and for regression. **[2]**
(vi) The result extends to projections summing to the identity: $P_1 + \cdots + P_k = I$. These are orthogonal; their quadratic forms are independent chi-squares (*Cochran's theorem).* So we can form F-*statistics*, for use in *testing hypotheses* in regression. **[3]**
[Seen – lectures]

Q4 (*Wold decomposition*).
(i) Consider a stationary process (X_t), with variance σ^2. If we are *given* the values of X_s up to X_{t-q}, this knowledge makes X_t *less variable*, so

$$\sigma_q^2 := var(X_t | \cdots, X_{t-q-2}, X_{t-q-1}, X_{t-q}) \leq \sigma^2.$$

As we increase q, the information given decreases, so X_t given this information becomes more variable: σ_q^2 increases with q. So

$$0 \leq \sigma_q^2 \uparrow \sigma_\infty^2 \leq \sigma^2 \qquad (q \to \infty).$$

One possibility is that $\sigma_q = 0$ for all q, and then $\sigma_\infty = 0$ also. So then X_t is non-random (deterministic), "given the remote past" (example: $X_t = a\cos(\omega t + b)$ with a, b, ω random: X_t is random, but the time-dependence is trivial). Such a process is called *singular* or *purely deterministic.* **[4]**

At the other extreme, we may have

$$\sigma_q \uparrow \sigma_\infty = \sigma \qquad (q \to \infty).$$

Then as information given recedes into the past, its influence dies away to nothing (as "it should"). Such a process is called *purely nondeterministic.* **[4]**

(ii) **Theorem (Wold Decomposition Theorem: Wold (1938))**. A (strictly) stationary stochastic process (X_t) possesses a unique decomposition

$$X_t = Y_t + Z_t,$$

where (i) Y_t is purely deterministic, (ii) Z_t is purely nondeterministic, (iii) Y_t, Z_t are uncorrelated, (iv) Z_t is a general linear process,

$$Z_t = \sum \phi_i \epsilon_{t-i}, \text{with the } \epsilon_t \text{ uncorrelated.} \qquad [4]$$

(iii) **Corollary.** If (X_t) has no purely deterministic component – so

$$X_t = \sum\nolimits_{i=0}^\infty \psi_i \epsilon_{t-i}, \qquad \sum \psi_i^2 < \infty, \qquad (\epsilon_t) \quad WN(\sigma^2) \ --$$

then
(i) $\gamma_k := cov(X_t, X_{t+k}) = \sigma^2 \sum_{i=0}^\infty \psi_i \psi_{i+k}$,
(ii) $\gamma_k \to 0$, $\rho_k := corr(X_t, X_{t+k}) \to 0$ $(k \to \infty)$. **[4]**

Proof.

$$\gamma_k = cov(X_t, X_{t+k}) = E[X_t, X_{t+k}] = E[(\sum\nolimits_{i=0}^{\infty} \psi_i \epsilon_{t-i})(\sum\nolimits_{j=0}^{\infty} \psi_j \epsilon_{t-k-j})]$$

$$= \sum \sum\nolimits_{i,j} \psi_i \psi_j E[\epsilon_{t-i} \epsilon_{t-k-j}].$$

Here $E[.] = 0$ unless $i = j + k$, when it is σ^2, so

$$\gamma_k = \sigma^2 \sum\nolimits_{j=0} \psi_j \psi_{j+k},$$

proving (i). For (ii), use the Cauchy-Schwarz inequality:

$$|\gamma_k| = \sigma^2 |\sum\nolimits_{i=0}^{\infty} \psi_i \psi_{i+k}| \leq \sigma^2 (\sum\nolimits_{i=0}^{\infty} \psi_i^2)^{1/2} \sum\nolimits_{i=0}^{\infty} \psi_{i+k}^2)^{1/2} \to 0 \quad (k \to \infty),$$

as $\sum \psi_i^2 < \infty$, so $\sum_{i=k}^{\infty} \psi_i^2$ is the tail of a convergent series. // **[4]**
[Seen – lectures]

Q5 (*Poisson distribution with Gamma prior*).

Data: $x = (x_1, \cdots, x_n)$, x_i independent, Poisson $P(\theta)$ param. θ:

$$f(x|\theta) = \Pi_1^n f(x_i|\theta) = \theta^{x_1+\cdots+x_n} e^{-n\theta}/x_1! \cdots x_n! = \theta^{n\bar{x}} e^{-n\theta}/\Pi x_i!,$$

where $\bar{x} := \frac{1}{n}\Sigma x_i$ is the sample mean.
Prior: the Gamma density $\Gamma(a, b) \quad (a, b > 0)$:

$$f(\theta) = \frac{a^b \theta^{b-1}}{\Gamma(b)} e^{-a\theta} \qquad (\theta > 0):$$

$$f(x|\theta)f(\theta) = \frac{a^b}{\Gamma(b)\Pi x_i!} \theta^{n\bar{x}+b-1} e^{-(n+a)\theta},$$

$$f(\theta|x) \propto f(x|\theta)f(\theta) = const.\theta^{n\bar{x}+b-1} e^{-(n+a)\theta}.$$

This has the form of a Gamma density. So, it *is* a Gamma density, $\Gamma(n + a, n\bar{x} + b)$:

$$f(\theta|x) = \frac{(n+a)^{n\bar{x}+b}}{\Gamma(n\bar{x}+b)}.\theta^{n\bar{x}+b-1} e^{-(n+a)\theta} \qquad (\theta > 0). \qquad [10]$$

Means. For $\Gamma(a, b)$, the mean is

$$\begin{aligned} E[\theta] &= \int_0^\infty \theta f(\theta) d\theta = \frac{a^b}{\Gamma(b)}. \int_0^\infty \theta^b e^{-a\theta} d\theta \\ &= \frac{a^b}{\Gamma(b)}.\Gamma(b+1)/a^{b+1} \quad \text{(substituting } t := a\theta) \\ &= b/a \qquad (\text{as } \Gamma(x+1) = x\Gamma(x)): \end{aligned}$$

the prior mean is b/a. [4]
The posterior mean is $(n\bar{x} + b)/(n + a)$; the data mean is $\bar{x}$. Write

$$\lambda := a/(n+a), \qquad \text{so } 1 - \lambda = n/(n+a): \qquad \text{since}$$

$$\frac{n\bar{x}+b}{n+a} = \frac{a}{n+a}.\frac{b}{a} + \frac{n}{n+a}.\bar{x},$$

posterior mean $(n\bar{x}+b)/(n+a) = \lambda$. prior mean $b/a + (1-\lambda)$. sample mean $\bar{x}$.

This is a weighted average, or *compromise*, of the sample and prior means, with weights proportional to their *precisions*: n (the sample size, a measure of the data precision) and a (the rate of decay of $\Gamma(a, b)$, a measure of the prior precision). [6]
[Seen, lectures]

Q6 (*Mixed models*).
(i) **Theorem (Edgeworth)**. If μ is an n-vector, $\mathbf{\Sigma} > \mathbf{0}$ a symmetric positive definite $n \times n$ matrix, then

$$f(x) := \frac{1}{(2\pi)^{\frac{1}{2}n}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)\}$$

is an n-dimensional probability density function (of a random n-vector X);
X has CF $\phi(\mathbf{t}) = \exp\{i\mathbf{t}^T\mu - \frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\}$;
X is multinormal $N(\mu, \mathbf{\Sigma})$. [3]
(ii) Bayes's Theorem:

$$f(\theta|x) = f(\theta, x)/f(x) = f(\theta)f(x|\theta)/f(x). \qquad (*) \quad [3]$$

(iii) If

$$y|u \sim N(A\beta + Bu, R), \qquad u \sim N(0, D),$$

$$\begin{aligned} f(y,u) &= f(y|u)f(u) \\ &= const. \exp\{-\frac{1}{2}(y - A\beta - Bu)^T R^{-1}(y - A\beta - Bu)\}. \exp\{-\frac{1}{2}u^T D^{-1}u\} \\ &= const. \exp\{-\frac{1}{2}[u^T(B^T R^{-1}B + D^{-1})u - 2u^T B^T R^{-1}(y - A\beta) + \text{function of } y\}. \end{aligned}$$

So also

$$f(u|y) = f(u,y)/f(y)$$

$$= const. \exp\{-\frac{1}{2}[u^T(B^TR^{-1}B+D^{-1})u-2u^TB^TR^{-1}(y-A\beta)+\text{fn of } y]\}. \quad [6]$$

From Edgeworth's theorem, if $X \sim N(\mu, \Sigma)$,

$$f(x) = const. \exp\{-\frac{1}{2}[x^T\Sigma^{-1}x - 2x\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu]\}.$$

Comparing, $u|y \sim N(\mu, \Sigma)$ with $\Sigma^{-1} = B^TR^{-1}B + D^{-1}$, from the quadratic term in u. The linear term in u then gives

$$B^TR^{-1}(y-A\beta) = \Sigma^{-1}\mu, \qquad \mu = \Sigma B^TR^{-1}(y-A\beta) = (B^TR^{-1}B+D^{-1})^{-1}B^TR^{-1}(y-A\beta).$$
[5]

So

$$u|y \sim N(\mu, \Sigma), \quad \mu = (B^TR^{-1}B+D^{-1})^{-1}B^TR^{-1}(y-A\beta)), \quad \Sigma = (B^TR^{-1}B+D^{-1})^{-1}.$$
[3]

[Seen – lectures and Problems]

N. H. Bingham