STA261 S19: Final Exam Solutions

No aids. 180 minutes. Write all answers directly beneath where the question is asked. Use pages 11-14 for rough work. **I strongly recommend you do all your rough work on the back pages, then copy your final answer into the space provided for each question. You will NOT be allowed to write your answers elsewhere if you waste space.**

1. (4) Let $X_1, \ldots, X_n$ be an IID sample from a standard Normal distribution having density $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$. Let $S_n = X_1 + \cdots + X_n$.

a) (1) Show that the <u>moment generating function</u> of X_1 is $M_X(t) = \exp(t^2/2)$.

Solution.

$$\begin{aligned}
M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} dx && \text{(0.25 marks)} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{tx - \frac{1}{2}x^2\right\} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-t)^2 + \frac{1}{2}t^2\right\} dx \\
&= e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-t)^2\right\} dx \\
&= e^{t^2/2}. && \text{(0.75 marks)}
\end{aligned}$$

b) (.5) Show that $M_X(t)$ is a fixed point of the transformation $\phi(x) \to \phi(x/\sqrt{n})^n$, that is, $M_X(t/\sqrt{n})^n = M_X(t)\ \forall\ n \in \mathbb{N}$.

Solution.

$$M_X(t/\sqrt{n})^n = \left(e^{t^2/2n}\right)^n = e^{t^2/2} = M_X(t). \qquad \text{(0.5 marks)}$$

c) (1.5) Show that $S_n/n \xrightarrow{p} 0$. For full marks show all work including computing all necessary quantities and stating any theorems you use and what conditions must be true for them to hold.

Solution. First,

$$EX = M'_X(0) = te^{t^2/2}\big|_{t=0} = 0. \qquad \text{(0.25 marks)}$$

Then,

$$\text{Var}(X) = EX^2 - (EX)^2 = EX^2 = M''_X(0) = \left(e^{t^2/2} + t^2e^{t^2/2}\right)\big|_{t=0} = 1. \quad \text{(0.25 marks)}$$

So, since each sample is independent with mean 0 and finite variance, the weak LLN implies that $S_n/n \xrightarrow{p} 0$ (1 mark).

d) (1) Show that $S_n/\sqrt{n} \stackrel{d}{=} X_1$ directly, i.e. without just invoking the Central Limit Theorem.

Solution.

$$M_{S_n/\sqrt{n}}(t) = M_X(t/\sqrt{n})^n = M_{X_1}(t).$$

Since the moment generating function uniquely defines the distribution, $S_n/\sqrt{n} \stackrel{d}{=} X_1$ (1 mark).

2. (4) Suppose $X_1, \dots, X_n$ is an IID random sample from a distribution with density f_θ and parameter $\theta \in \Theta$. We put a prior distribution $\pi(\theta)$ on θ.

a) (1) Define (mathematically) what it means for a statistic $T = T(X)$ to be sufficient for θ.

Solution. For the likelihood function L, whenever $T(s_1) = T(s_2)$,

$$L(\theta \mid T(s_1)) = c(s_1, s_2) L(\theta \mid T(s_2)),$$

where $c(s_1, s_2)$ does not depend on $X_1, \dots, X_n$ (1 mark).

b) (1) Give a mathematical expression for the posterior distribution of $\theta | X_1, \dots, X_n$.

$$p(\theta \mid X_1, \dots, X_n) = \frac{L(\theta \mid X_1, \dots, X_n)\pi(\theta)}{\int_{\nu \in \Theta} L(\nu \mid X_1, \dots, X_n)\pi(\nu) d\nu}. \qquad \text{(1 mark)}$$

Note you already asked this on test 3

c) (2) Show that T is sufficient if and only if the posterior distribution of $\theta | X_1, \dots, X_n$ is equal to the posterior distribution of $\theta | T(X)$ for every $\theta \in \Theta$ and for every choice of prior.

Solution.

$$\begin{aligned}
p(\theta \mid X_1, \dots, X_n) &= \frac{f(X_1, \dots, X_n \mid \theta)\pi(\theta)}{f(X_1, \dots, X_n)} \\
&= \frac{f(X_1, \dots, X_n \mid \theta, T(X)) g(T(X) \mid \theta)\pi(\theta)}{f(X_1, \dots, X_n)} \\
&= \frac{f(X_1, \dots, X_n \mid T(X)) g(T(X) \mid \theta)\pi(\theta)}{f(X_1, \dots, X_n)} \quad \text{(iff sufficiency)} \\
&= \frac{f(X_1, \dots, X_n, T(X)) g(T(X) \mid \theta)\pi(\theta)}{f(X_1, \dots, X_n) g(T(X))} \\
&= \frac{f(X_1, \dots, X_n) g(T(X) \mid \theta)\pi(\theta)}{f(X_1, \dots, X_n) g(T(X))} \quad (T \text{ is entirely determined by } X) \\
&= p(\theta \mid T(X)). \qquad \text{(1 mark)}
\end{aligned}$$

However, since these are equalities they work both ways and the equality of lines 2 and 3 holds if and only if T is sufficient, so the conditions are equivalent (1 mark).

ALTERNATE SOLUTION: for the "only if" part, assume the posterior distribution of $\theta | X_1, \dots, X_n$ is equal to the posterior distribution of $\theta | T(X)$ for every $\theta \in \Theta$ and for every choice of prior. Then

$$\begin{aligned}
& \pi(\theta | T(X)) = \pi(\theta | X) \\
\Longrightarrow\ & \frac{\pi(T(X)|\theta)}{m(T(x))} = \frac{\pi(X|\theta)}{m(X)} \\
\Longrightarrow\ & \frac{\pi(T(X)|\theta)}{\pi(X|\theta)} = \frac{m(T(X))}{m(X)}
\end{aligned} \tag{0.1}$$

Since this ratio is free of θ, by the factorization theorem $T(X)$ is sufficient for θ.

3. (4) Suppose we observe an IID random sample $X_1, \ldots, X_n$ from a Binomial(n, θ) distribution, with $P(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$. We wish to perform Bayesian inference on θ.

a) (.5) The beta distribution with parameters $\alpha > 0, \beta > 0$ has density $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ for $0 < x < 1$. Use this fact to evaluate the integral $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$.

Solution.

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \quad \text{(0.5 marks)}$$

b) (1) Suppose we put a Beta(α, β) prior on θ. Evaluate the normalizing constant, i.e. the denominator of the posterior distribution.

Solution. Denote the normalizing constant by C.

$$\begin{aligned}
p(\theta \mid X_1, \ldots, X_n) &= \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{C} \\
\int_0^1 p(\theta \mid X_1, \ldots, X_n)d\theta &= \frac{1}{C}\int_0^1 \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\
C &= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}d\theta \\
C &= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+x+\beta+n-x)} \\
C &= \frac{\Gamma(n)\Gamma(\alpha+\beta)\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(n-x)\Gamma(x)\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)}. \qquad \text{(1 mark)}
\end{aligned}$$

NOTE: they may not use the gamma function notation for the $\binom{n}{x}$ term; be careful not to take marks off here.

FURTHER NOTE: the solution here is actually presented as for a single X, which isn't quite right. The question states we have a random sample from a binomial, so the x should be a $\sum_{i=1}^n X_i$ and the $\binom{n}{x}$ should be a $\prod_{i=1}^n \binom{n}{x_i}$. These don't change the calculations so don't be too picky here.

c) (.5) Define what is meant by a conjugate prior distribution.

Solution. The posterior is in the same family as the prior (just different parameters) (0.5 marks).

d) (1) Show that the posterior distribution of $\theta | X_1, \dots, X_n$ under the prior from b) is

$$\text{Beta}\left(\alpha + \sum_{i=1}^{n} X_i, \beta + n - \sum_{i=1}^{n} X_i\right)$$

Solution. Using the normalizing constant C,

$$p(\theta \mid X_1, \dots, X_n) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}.$$

However, x is just from $P(X = x)$ where $X = \sum_{i=1}^{n} X_i$, so the claim is shown (1 mark).

e) (1) Write down an expression for a normal approximation to the posterior for this example. You don't have to compute any likelihood-related quantities.

Solution.

$$\theta | X_1, \dots, X_n \overset{\approx}{\sim} \text{Normal}\left(\widehat{\theta}_p, \frac{1}{j_p(\widehat{\theta}_p)}\right)$$

where $\widehat{\theta}_p = \text{argmax}_\theta \; g(\theta)$ and $j_p(\theta) = -\frac{\partial^2 g(\theta)}{\partial \theta^2}$ where $g(\theta) = \ell(\theta) + \log \pi(\theta)$.

4. (8) Let Y_i be an independent but **not identically distributed** random sample from the family of normal distributions, $Y_i \overset{IND}{\sim} \text{Normal}(\mu_i, \sigma^2)$. Suppose we also observe fixed covariates x_i, one for each Y_i, and we wish to estimate μ_i according to the linear model $\mu_i = \beta x_i$ for some fixed, unknown parameter $\beta \in \mathbb{R}$. To repeat: x_i are fixed and known; β is fixed and unknown. We are estimating β only.

a) (2) With σ^2 a fixed, known constant, write down the likelihood and log-likelihood for β.

Solution. The likelihood is

$$\begin{aligned} L(\beta \mid Y_1, \ldots, Y_n) &= \prod_{i=1}^{n} p(Y_i; \mu_i = \beta x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(Y_i - \beta x_i)^2}{2\sigma^2} \right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{\sum_{i=1}^{n} (Y_i - \beta x_i)^2}{2\sigma^2} \right\}. \qquad \text{(1 mark)} \end{aligned}$$

The log-likelihood is

$$\ell(\beta \mid Y_1, \ldots, Y_n) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} (Y_i - \beta x_i)^2}{2\sigma^2}. \qquad \text{(1 mark)}$$

b) (2) Write down the score function for β and find the maximum likelihood estimator for β.

Solution. The score function is

$$S(\beta \mid Y_1, \ldots, Y_n) = \frac{\partial \ell(\beta \mid Y_1, \ldots, Y_n)}{\partial \beta} = \frac{\sum_{i=1}^{n} x_i (Y_i - \beta x_i)}{\sigma^2}. \qquad \text{(1 mark)}$$

Setting equal to zero gives

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}. \qquad \text{(1 mark)}$$

c) (2) Find the observed information and expected information for β and the variance of the maximum likelihood estimator for β.

Solution. First,

$$\frac{\partial^2 \ell(\beta \mid Y_1, \ldots, Y_n)}{\partial \beta^2} = -\frac{\sum_{i=1}^n x_i^2}{\sigma^2}. \qquad \text{(0.5 marks)}$$

The observed and expected information are both thus thus

$$I(Y_1, \ldots, Y_n) = \frac{\sum_{i=1}^n x_i^2}{\sigma^2}. \qquad \text{(1 mark)}$$

The variance of the MLE is

$$\mathrm{Var}(\widehat{\beta}) = I^{-1}(Y_1, \ldots, Y_n) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}. \qquad \text{(1 mark)}$$

d) (2) If we didn't know σ^2 but wished to estimate it together with β, would our estimates for β change? What about its estimated variance? You may (actually, you should) use facts about normal families to answer this question; you should not do all the calculations again.

Solution.

$\widehat{\beta}$ would be unchanged. It's estimated variance would be the same formula but with σ^2 replaced with $\widehat{\sigma^2}$. In normal families the location and spread are independent. They may say *anything* reasonable here to get full marks. Other potential explanations are that the MLE for $\widehat{\beta}$ doesn't depend on σ, the score function for $\widehat{\beta}$ doesn't depend on σ, etc.

5. (5) Suppose we have data $X_1, \ldots, X_n$ independently and identically distributed according to a parametric statistical model $\{F_\theta : \theta \in \Theta\}$. Let $U = U(X)$ be an unbiased estimator for θ.

a) (1) What is $\mathbb{E}(U)$?

Solution.

$$E(U) = \theta. \qquad \text{(1 mark)}$$

b) (1) Let $T = T(X)$ be a sufficient statistic. Define a new estimator $S(X) = \mathbb{E}(U(X)|T(X))$. Show that $S(X)$ is unbiased. *Hint: recall that for any random variables X, Y, $\mathbb{E}(X) = \mathbb{E}\mathbb{E}(X|Y)$.*

Solution. *The hint kind of gives it away...*

$$\mathbb{E}S(X) = \mathbb{E}\mathbb{E}(U(X) \mid T(X)) = \mathbb{E}U(X) = \theta. \qquad \text{(1 mark)}$$

c) (1) Show that $S(X)$ has variance which is at least as small as the variance of ***$U(X)$, not $T(X)$. *Hint: for any random variables X, Y, $\mathrm{Var}(X) = \mathrm{Var}(\mathbb{E}(X|Y)) + \mathbb{E}\,\mathrm{Var}(X|Y)$. Variance is non-negative.*

Solution.

$$\begin{aligned}
\mathrm{Var}(U(X)) &= \mathrm{Var}(\mathbb{E}[U(X) \mid T(X)]) + \mathbb{E}[\mathrm{Var}(U(X) \mid T(X)] \\
\mathrm{Var}(U(X)) &= \mathrm{Var}(S(X)) + \mathrm{Var}(U(X) \mid T(X)) && \text{(0.5 marks)} \\
\mathrm{Var}(S(X)) &= \mathrm{Var}(U(X)) - \mathrm{Var}(U(X) \mid T(X)) \\
\mathrm{Var}(S(X)) &\leq \mathrm{Var}(U(X)). && \text{(0.5 marks)}
\end{aligned}$$

d) (2) Suppose, in addition, ***$T(X)$, not $S(X)$ has the weird-looking property that for any function h, $\mathbb{E}h(T(X)) = 0 \implies h \equiv 0$, called "completeness". Prove that $S(X)$ has minimum variance out of all unbiased estimators. This question is very hard, but give it a shot anyways— I believe in you. *Hint: define two unbiased estimators $U_1(X)$ and $U_2(X)$. Condition them on $T(X)$, and subtract the result from each other. Use completeness. How does what you get imply the result?*

Solution. As per the hint, define $\phi_1(T) = \mathbb{E}(U_1 \mid T)$ and $\phi_2(T) = \mathbb{E}(U_2 \mid T)$. Then, $\mathbb{E}\phi_1 = \mathbb{E}\phi_2 = \theta$. Set $h(T) = \phi_1(T) - \phi_2(T)$ so that $\mathbb{E}h = \mathbb{E}\phi_1 - \mathbb{E}\phi_2 = \theta - \theta = 0$ (0.5 marks). By completeness, this implies $h(T) \equiv 0$, so $\phi_1 \equiv \phi_2$ (0.5 marks). That is, regardless of which unbiased statistic is used, conditioning on $T(X)$ will give the same estimator, and thus they will have the same variance (0.5 marks). This combined with the result from part c) implies $S(X)$ is the UMVUE (0.5 marks).

THIS PAGE IS FOR ROUGH WORK. NOTHING ON THIS PAGE WILL BE MARKED.

THIS PAGE IS FOR ROUGH WORK. NOTHING ON THIS PAGE WILL BE MARKED.

THIS PAGE IS FOR ROUGH WORK. NOTHING ON THIS PAGE WILL BE MARKED.

THIS PAGE IS FOR ROUGH WORK. NOTHING ON THIS PAGE WILL BE MARKED.