SOLUTIONS TO
STATISTICAL METHODS EXAM

Solutions to Question 1 This question explores the moment generating function of the normal distribution.

(i) Let $X \sim N(0,1)$. The moment generating function of X is:

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

(ii) The first four derivatives of $M_X(t)$ are

$$\begin{aligned}
M_X^{'}(t) &= t \exp\left(\frac{t^2}{2}\right), \\
M_X^{''}(t) &= t^2 \exp\left(\frac{t^2}{2}\right) + \exp\left(\frac{t^2}{2}\right), \\
M_X^{'''}(t) &= t^3 \exp\left(\frac{t^2}{2}\right) + 3t \exp\left(\frac{t^2}{2}\right), \\
M_X^{''''}(t) &= t^4 \exp\left(\frac{t^2}{2}\right) + 6t^2 \exp\left(\frac{t^2}{2}\right) + 3 \exp\left(\frac{t^2}{2}\right).
\end{aligned}$$

So, $E(X) = 0$, $E(X^2) = 1$, $E(X^3) = 0$ and $E(X^4) = 3$.

(iii) The moment generating function of $Y = \mu + \sigma Z$ is:

$$M_Y(t) = E\left[\exp(t\mu + t\sigma Z)\right] = \exp(t\mu) E\left[\exp(t\sigma Z)\right] = \exp(t\mu) \exp\left(\frac{t^2\sigma^2}{2}\right).$$

(iv) The moment generating function of $S = X_1 + X_2$ is:

$$\begin{aligned}
M_S(t) &= E\left[\exp(tX_1 + tX_2)\right] \\
&= E\left[\exp(tX_1)\right] E\left[\exp(tX_2)\right] \\
&= \exp\left(t\mu_1 + \frac{t^2\sigma_1^2}{2}\right) \exp\left(t\mu_2 + \frac{t^2\sigma_2^2}{2}\right).
\end{aligned}$$

(v) We have $E(S) = E(X_1 + X_2) = E(X_1) + E(X_2) = \mu_1 + \mu_2$ and $Var(S) = Var(X_1 + X_2) = Var(X_1) + Var(X_2) = \sigma_1^2 + \sigma_2^2$.

(vi) Since

$$M_S(t) = \exp\left(t\mu_1 + \frac{t^2\sigma_1^2}{2}\right)\exp\left(t\mu_2 + \frac{t^2\sigma_2^2}{2}\right) = \exp\left[t(\mu_1+\mu_2) + \frac{t^2(\sigma_1^2+\sigma_2^2)}{2}\right],$$

it follows that S has the normal distribution with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.

Solutions to Question 2 Suppose $\widehat{\theta}$ is an estimator of θ.

(i) $\widehat{\theta}$ is an unbiased estimator of θ if $E(\widehat{\theta}) = \theta$.

(ii) $\widehat{\theta}$ is an asymptotically unbiased estimator of θ if $\lim_{n\to\infty} E(\widehat{\theta}) = \theta$.

(iii) the bias of $\widehat{\theta}$ is $E(\widehat{\theta}) - \theta$.

(iv) the mean squared error of $\widehat{\theta}$ is $E(\widehat{\theta} - \theta)^2$.

(v) $\widehat{\theta}$ is a consistent estimator of θ if $\lim_{n\to\infty} E(\widehat{\theta} - \theta)^2 = 0$.

Let X_i denote the time that it takes student i to complete a take-home exam, and suppose that $X_1, X_2, \ldots, X_n$ constitute a random sample from an exponential distribution with parameter β. Consider the following estimators for $\theta = 1/\beta$: $\widehat{\theta_1} = c \min(X_1, X_2, \ldots, X_n)$ and $\widehat{\theta_2} = 1/n \sum_{i=1}^n X_i$.

(i) Let $Z = \min(X_1, X_2, \ldots, X_n)$. Then the cdf of Z is

$$\begin{aligned}
\Pr(Z < z) &= \Pr\left(\min(X_1, X_2, \ldots, X_n) < z\right) \\
&= 1 - \Pr\left(\min(X_1, X_2, \ldots, X_n) > z\right) \\
&= 1 - \Pr\left(X_1 > z, X_2 > z, \ldots, X_n > z\right) \\
&= 1 - \Pr^n\left(X > z\right) \\
&= 1 - \exp(-n\beta z).
\end{aligned}$$

It follows that Z has an exponential distribution with parameter $n\beta$. So, $E(cZ) = cE(Z) = c/(n\beta) = c\theta/n = \theta$ if and only if $c = n$.

(ii) The variance of $\widehat{\theta_1}$ is

$$\begin{aligned}
Var\left(\widehat{\theta_1}\right) &= n^2 Var\left(Z\right) \\
&= \frac{n^2}{n^2\beta^2} \\
&= \frac{1}{\beta^2} \\
&= \theta^2
\end{aligned}$$

The MSE is the same as the variance since $\widehat{\theta_1}$ is unbiased.

(iii) The bias of $\widehat{\theta_2}$ is

$$\begin{aligned}
E\left(\widehat{\theta_2}\right) - \theta &= E\left(\frac{1}{n}\sum_{i=1}^n X_i\right) - \theta \\
&= \frac{1}{n}\sum_{i=1}^n E\left(X_i\right) - \theta
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\theta - \theta \\
&= \theta - \theta \\
&= 0.
\end{aligned}
$$

The variance of $\widehat{\theta_2}$ is

$$
\begin{aligned}
Var\left(\widehat{\theta_2}\right) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \theta^2 \\
&= \frac{\theta^2}{n}.
\end{aligned}
$$

The MSE is the same as the variance since $\widehat{\theta_2}$ is unbiased.

(iv) Clearly, $\widehat{\theta_2}$ has the smaller MSE and so it should be preferred.

Solutions to Question 3 Let the random variable Y_i be the number of typographical errors on a page of a 400-page book (for $i = 1, 2, \ldots, 400$), and suppose that the Y_i's are independent and identically distributed according to a Poisson distribution with parameter λ. Let the random variable X be the number of pages of this book that contain at least one typographical error. Suppose that you are told the value of X but are not told anything about the values of Y_i.

(i) Clearly, X has the binomial distribution with parameters $n = 400$ and $p = \Pr(Y > 0) = 1 - \Pr(Y = 0) = 1 - \exp(-\lambda)$. So, the pmf of X is

$$\begin{aligned} p(x) &= \binom{n}{x} (1-p)^{400-x} p^x \\ &= \binom{n}{x} \exp\left\{-(400-x)\lambda\right\} \left\{1 - \exp(-\lambda)\right\}^x \end{aligned}$$

for $x = 0, 1, \ldots, 400$.

(ii) The likelihood function of λ is

$$L(\lambda) = \binom{n}{x} \exp\left\{-(400-x)\lambda\right\} \left\{1 - \exp(-\lambda)\right\}^x$$

for $\lambda > 0$.

(iii) The log likelihood function of λ is

$$\log L(\lambda) = \log \binom{n}{x} - (400-x)\lambda + x \log\left\{1 - \exp(-\lambda)\right\}.$$

The first derivative of $\log L$ with respect to λ is

$$\frac{d \log L(\lambda)}{d\lambda} = x - 400 + \frac{x \exp(-\lambda)}{1 - \exp(-\lambda)}.$$

Setting this to zero and solving, we obtain $\widehat{\lambda} = \log\{400/(400-x)\}$. The second derivative of $\log L$ with respect to λ

$$\frac{d^2 \log L(\lambda)}{d\lambda^2} = -\frac{x \exp(\lambda)}{\left\{\exp(\lambda) - 1\right\}^2} < 0,$$

so $\widehat{\lambda} = \log\{400/(400-x)\}$ is indeed a maximum likelihood estimator of λ.

(iv) If $x = 25$ then $\widehat{\lambda} = \log\{400/375\} = 06453852$.

(v) If X has the binomial distribution with parameters $n = 400$ and p then the mle of p is $\widehat{p} = x/400$. So, by the invariance property the mle of λ can be obtained by setting $1 - \exp(-\lambda) = x/400$.

Solutions to Question 4 Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with the common probability density function (pdf):

$$f(x) = \theta_2 x^{\theta_2 - 1} \theta_1^{-\theta_2}$$

for $0 < x < \theta_1$, $\theta_1 > 0$ and $\theta_2 > 0$. Both θ_1 and θ_2 are unknown.

(i) The cumulative distribution function corresponding to the given pdf is

$$F(x) = \theta_2 \theta_1^{-\theta_2} \int_0^x y^{\theta_2 - 1} dy = \theta_1^{-\theta_2} x^{\theta_2}.$$

The mean corresponding to the given pdf is

$$E(X) = \theta_2 \theta_1^{-\theta_2} \int_0^{\theta_1} y^{\theta_2} dy = \frac{\theta_1 \theta_2}{\theta_2 + 1}.$$

The variance corresponding to the given pdf is

$$Var(X) = \theta_2 \theta_1^{-\theta_2} \int_0^{\theta_1} y^{\theta_2 + 1} dy - \frac{\theta_1^2 \theta_2^2}{(\theta_2 + 1)^2} = \frac{\theta_1^2 \theta_2}{\theta_2 + 2} - \frac{\theta_1^2 \theta_2^2}{(\theta_2 + 1)^2}.$$

(ii) The joint likelihood function of θ_1 and θ_2 is

$$L\left(\theta_1, \theta_2\right) = \theta_2^n \theta_1^{-n\theta_2} \left(\prod_{i=1}^n x_i \right)^{\theta_2 - 1}$$

for $\theta_1 > 0$ and $\theta_2 > 0$.

(iii) The likelihood function monotonically decreases with respect to θ_1. The lowest possible value for θ_1 is $\max(X_1, X_2, \ldots, X_n)$. So, the mle of θ_1 is $\max(X_1, X_2, \ldots, X_n)$.

(iv) The log of the joint likelihood function is

$$\log L\left(\theta_1, \theta_2\right) = n \log \theta_2 - n\theta_2 \log \theta_1 + (\theta_2 - 1) \sum_{i=1}^n \log x_i.$$

The first derivative of the log likelihood with respect to θ_2 is

$$\frac{d \log L\left(\theta_1, \theta_2\right)}{d\theta_2} = \frac{n}{\theta_2} - n \log \theta_1 + \sum_{i=1}^n \log x_i.$$

Setting this to zero and solving, we obtain $\widehat{\theta_2} = n / \{ n \log \widehat{\theta_1} - \sum_{i=1}^n \log x_i \}$, where $\widehat{\theta_1} = \max(X_1, X_2, \ldots, X_n)$. The second derivative of the log likelihood with respect to θ_2

$$\frac{d^2 \log L\left(\theta_1, \theta_2\right)}{d\theta_2^2} = -\frac{n}{\theta_2^2} < 0.$$

Hence, $\widehat{\theta_2} = n / \{ n \log \widehat{\theta_1} - \sum_{i=1}^n \log x_i \}$ is indeed the mle of θ_2.

(v) Let $Z = \max(X_1, X_2, \ldots, X_n)$. Then the cdf of Z is

$$\begin{aligned} \Pr(Z < z) &= \Pr\left(\max(X_1, X_2, \ldots, X_n) < z\right) \\ &= \Pr\left(X_1 < z, X_2 < z, \ldots, X_n < z\right) \\ &= \Pr^n\left(X < z\right) \\ &= \theta_1^{-n\theta_2} x^{n\theta_2}. \end{aligned}$$

It follows that Z has the same distribution as the given pdf with θ_2 replaced by $n\theta_2$. So, the bias $\widehat{\theta_1}$ is

$$E\left(\widehat{\theta_1}\right) - \theta_1 = \frac{n\theta_1\theta_2}{n\theta_2 + 1} - \theta_1 = -\frac{\theta_1}{n\theta_2 + 1} < 0.$$

The variance $\widehat{\theta_1}$ is

$$Var\left(\widehat{\theta_1}\right) = \frac{n\theta_1^2\theta_2}{n\theta_2 + 2} - \frac{n^2\theta_1^2\theta_2^2}{(n\theta_2 + 1)^2}$$

and the MSE $\widehat{\theta_1}$ is

$$MSE\left(\widehat{\theta_1}\right) = \frac{\theta_1^2}{(n\theta_2 + 1)^2} + \frac{n\theta_1^2\theta_2}{n\theta_2 + 2} - \frac{n^2\theta_1^2\theta_2^2}{(n\theta_2 + 1)^2}.$$

The MSE approaches zero as $n \to \infty$. Hence, $\widehat{\theta_1}$ is a biased and consistent estimator for θ_1.

Solutions to Question 5 Suppose we wish to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

(i) the Type I error occurs if H_0 is rejected when in fact $\theta = \theta_0$.

(ii) the Type II error occurs if H_0 is accepted when in fact $\theta \neq \theta_0$.

(iii) the significance level is the probability of type I error.

(iv) the power function: $\Pi(\theta) = \Pr(\text{ Reject } H_0 \mid \theta)$.

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Bernoulli distribution with parameter p. Assume $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$ has a normal distribution with mean p and variance $p(1-p)/n$.

(i) The rejection region for $H_0 : p = p_0$ versus $H_1 : p \neq p_0$ is

$$\sqrt{\frac{n}{\bar{x}(1-\bar{x})}}\,|\bar{x} - p_0| > z_{\alpha/2}.$$

(ii) The rejection region for $H_0 : p = p_0$ versus $H_1 : p < p_0$ is

$$\sqrt{\frac{n}{\bar{x}(1-\bar{x})}}\,(\bar{x} - p_0) < -z_{\alpha}.$$

(iii) The rejection region for $H_0 : p = p_0$ versus $H_1 : p > p_0$ is

$$\sqrt{\frac{n}{\bar{x}(1-\bar{x})}}\,(\bar{x} - p_0) > z_{\alpha}.$$

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Bernoulli distribution with parameter p. Assume $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$ has a normal distribution with mean p and variance $p(1-p)/n$.

(i) The power function, $\Pi(p)$, for $H_0 : p = p_0$ versus $H_1 : p \neq p_0$ is

$$\begin{aligned}
\Pi(p) &= \Pr\left(\sqrt{\frac{n}{\bar{x}(1-\bar{x})}}\,|\bar{x} - p_0| > z_{\alpha/2} \,\middle|\, p\right) \\
&= \Pr\left(|\bar{x} - p_0| > \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} z_{\alpha/2} \,\middle|\, p\right) \\
&= \Pr\left(\bar{x} > p_0 + \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} z_{\alpha/2} \text{ or } \bar{x} < p_0 - \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} z_{\alpha/2} \,\middle|\, p\right)
\end{aligned}$$

$$
\begin{aligned}
&= \Pr\left(\sqrt{n}\frac{\bar{x}-p}{\sqrt{p(1-p)}} > \sqrt{n}\frac{p_0-p}{\sqrt{p(1-p)}} + \sqrt{\frac{\bar{x}(1-\bar{x})}{p(1-p)}}z_{\alpha/2}\right. \\
&\qquad \left.\text{or } \sqrt{n}\frac{\bar{x}-p}{\sqrt{p(1-p)}} < \sqrt{n}\frac{p_0-p}{\sqrt{p(1-p)}} - \sqrt{\frac{\bar{x}(1-\bar{x})}{p(1-p)}}z_{\alpha/2}\right| p\Bigg) \\
&= \Pr\left(Z > \sqrt{n}\frac{p_0-p}{\sqrt{p(1-p)}} + \sqrt{\frac{\bar{x}(1-\bar{x})}{p(1-p)}}z_{\alpha/2}\right. \\
&\qquad \left.\text{or } Z < \sqrt{n}\frac{p_0-p}{\sqrt{p(1-p)}} - \sqrt{\frac{\bar{x}(1-\bar{x})}{p(1-p)}}z_{\alpha/2}\right| p\Bigg) \\
&= 1-\Phi\left(\sqrt{n}\frac{p_0-p}{\sqrt{p(1-p)}} + \sqrt{\frac{\bar{x}(1-\bar{x})}{p(1-p)}}z_{\alpha/2}\right) \\
&\qquad +\Phi\left(\sqrt{n}\frac{p_0-p}{\sqrt{p(1-p)}} - \sqrt{\frac{\bar{x}(1-\bar{x})}{p(1-p)}}z_{\alpha/2}\right),
\end{aligned}
$$

where $\Phi(\cdot)$ denotes the standard normal distribution function.

(ii) The power function, $\Pi(p)$, for $H_0 : p = p_0$ versus $H_1 : p < p_0$ is

$$
\begin{aligned}
\Pi(p) &= \Pr\left(\sqrt{\frac{n}{\bar{x}(1-\bar{x})}}(\bar{x}-p_0) < -z_\alpha \right| p\Bigg) \\
&= \Pr\left(\bar{x} < p_0 - \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}z_\alpha \right| p\Bigg) \\
&= \Pr\left(\sqrt{n}\frac{\bar{x}-p}{\sqrt{p(1-p)}} < \sqrt{n}\frac{p_0-p}{\sqrt{p(1-p)}} - \sqrt{\frac{\bar{x}(1-\bar{x})}{p(1-p)}}z_\alpha \right| p\Bigg) \\
&= \Pr\left(Z < \sqrt{n}\frac{p_0-p}{\sqrt{p(1-p)}} - \sqrt{\frac{\bar{x}(1-\bar{x})}{p(1-p)}}z_\alpha \right| p\Bigg) \\
&= \Phi\left(\sqrt{n}\frac{p_0-p}{\sqrt{p(1-p)}} - \sqrt{\frac{\bar{x}(1-\bar{x})}{p(1-p)}}z_\alpha\right),
\end{aligned}
$$

where $\Phi(\cdot)$ denotes the standard normal distribution function.

(iii) The power function, $\Pi(p)$, for $H_0 : p = p_0$ versus $H_1 : p > p_0$ is

$$
\begin{aligned}
\Pi(p) &= \Pr\left(\sqrt{\frac{n}{\bar{x}(1-\bar{x})}}(\bar{x}-p_0) > z_\alpha \right| p\Bigg) \\
&= \Pr\left(\bar{x} > p_0 + \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}z_\alpha \right| p\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&= \Pr\left(\sqrt{n}\frac{\bar{x}-p}{\sqrt{p(1-p)}} < \sqrt{n}\frac{p_0-p}{\sqrt{p(1-p)}} + \sqrt{\frac{\bar{x}(1-\bar{x})}{p(1-p)}}z_\alpha \middle| p\right) \\
&= \Pr\left(Z > \sqrt{n}\frac{p_0-p}{\sqrt{p(1-p)}} + \sqrt{\frac{\bar{x}(1-\bar{x})}{p(1-p)}}z_\alpha \middle| p\right) \\
&= 1-\Phi\left(\sqrt{n}\frac{p_0-p}{\sqrt{p(1-p)}} + \sqrt{\frac{\bar{x}(1-\bar{x})}{p(1-p)}}z_\alpha\right),
\end{aligned}
$$

where $\Phi(\cdot)$ denotes the standard normal distribution function.

Note that we have used the fact $\sqrt{n}\{\bar{x}-p\}/\sqrt{p(1-p)}$ has the standard normal distribution.

Solutions to Question 6 The Neyman–Pearson test rejects $H_0 : \theta = \theta_1$ in favor of $H_1 : \theta = \theta_2$ if

$$\frac{L(\theta_1)}{L(\theta_2)} = \frac{\prod_{i=1}^{n} f(X_i; \theta_1)}{\prod_{i=1}^{n} f(X_i; \theta_2)} < k$$

for some k.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter p.

(i) The most powerful test is to reject $H_0 : p = p_0$ if

$$\begin{aligned}
\frac{L(p_0)}{L(p_1)} &= \frac{p_0^{\sum_{i=1}^n X_i} (1-p_0)^{n-\sum_{i=1}^n X_i}}{p_1^{\sum_{i=1}^n X_i} (1-p_1)^{n-\sum_{i=1}^n X_i}} \\
&= \left(\frac{1-p_0}{1-p_1}\right)^n \left[\frac{p_0(1-p_1)}{p_1(1-p_0)}\right]^{\sum_{i=1}^n X_i} \\
&< k_0,
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
& \left[\frac{p_0(1-p_1)}{p_1(1-p_0)}\right]^{\sum_{i=1}^n X_i} < \left(\frac{1-p_0}{1-p_1}\right)^{-n} k_0 \\
\Longleftrightarrow\ & \sum_{i=1}^n X_i \log\left[\frac{p_0(1-p_1)}{p_1(1-p_0)}\right] < \log\left[\left(\frac{1-p_0}{1-p_1}\right)^{-n} k_0\right] \\
\Longleftrightarrow\ & \sum_{i=1}^n X_i > \left\{\log\left[\frac{p_0(1-p_1)}{p_1(1-p_0)}\right]\right\}^{-1} \log\left[\left(\frac{1-p_0}{1-p_1}\right)^{-n} k_0\right] \\
\Longleftrightarrow\ & \sum_{i=1}^n X_i > k
\end{aligned}$$

as required. Note that $\{p_0(1-p_1)\}/\{p_1(1-p_0)\} < 1$ and so $\log\{p_0(1-p_1)\}/\{p_1(1-p_0)\} < 0$.

(ii) Note that $\sum_{i=1}^n X_i$ has the binomial distribution with parameters n and p. So,

$$\Pi(p) = \Pr\left(\sum_{i=1}^n X_i > k \,\middle|\, p\right) = \Pr\left(Bin(n,p) > k\right).$$

(iii) Note that

$$\begin{aligned}
\Pr\left(Bin(5, 0.5) > 4\right) &= 0.03125, \\
\Pr\left(Bin(5, 0.5) > 3\right) &= 0.1875.
\end{aligned}$$

So, $k = 3$.

(iv) Note that

$$
\begin{aligned}
\beta &= \Pr(\text{Type II error}) \\
&= \Pr\left(\sum_{i=1}^{5} X_i \leq 3 \,\middle|\, p = 0.6\right) \\
&= \Pr\left(Bin(5, 0.6) \leq 3\right) \\
&= 0.66304.
\end{aligned}
$$