Part IB: Statistics
Examples Sheet 1 Solutions

Please send all comments and corrections to jmm232@cam.ac.uk.

2. If $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$ are independent, derive the distribution of $\min(X, Y)$. If $X \sim \Gamma(\alpha, \lambda)$ and $Y \sim \Gamma(\beta, \lambda)$ are independent, derive the distributions of $X + Y$ and $X/(X + Y)$.

Solution: In the first case, we have:

$$\text{Prob}(\min(X, Y) > z) = \text{Prob}\left(X > z \text{ and } Y > z\right) = \text{Prob}(X > z)\text{Prob}(Y > z) = e^{-(\lambda+\mu)z}.$$

Hence $\min(X, Y) \sim \text{Exp}(\lambda + \mu)$.

In the second case, we can use moment-generating functions. We have $X \sim \Gamma(\alpha, \lambda)$ and $Y \sim \Gamma(\beta, \lambda)$, and hence:

$$M_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha+\beta}.$$

Hence we have $X + Y \sim \Gamma(\alpha + \beta, \lambda)$.

In the third case, we consider a transformation $t = x/(x + y)$ and $u = x + y$. Then $x = tu, y = u(1 - t)$. It follows that the Jacobian of the transformation is:

$$J = \det \begin{pmatrix} u & t \\ -u & 1 - t \end{pmatrix} = u.$$

Since $f_{X,Y}(x, y) \propto x^{\alpha-1}y^{\beta-1}e^{-\lambda(x+y)}$, it follows that the join density of (T, U) is given by:

$$f_{(T,U)}(t, u) \propto u(tu)^{\alpha-1}(u(1 - t))^{\beta-1}e^{-\lambda u}.$$

In particular, integrating out u, we obtain:

$$T = \frac{X}{X + Y} \sim \text{Beta}(\alpha, \beta).$$

This proof also shows that $U = X + Y$ is gamma-distributed as $\Gamma(\alpha + \beta, \lambda)$.

A much more general argument for approaching this question is to use delta functions. The density of $Z = \min(X, Y)$ is given by:

$$f_Z(z) \propto \lambda\mu \int_0^\infty \int_0^\infty dx dy\, \delta(z - \min(x, y)) e^{-\lambda x} e^{-\mu y}.$$

Observe that this integral can be split into two regions, $0 \leq x \leq y$ (in which $\min(x, y) = x$) and $0 \leq y \leq x$ (in which $\min(x, y) = y$). Hence we can write:

$$\begin{aligned}
f_Z(z) &\propto \lambda\mu \int_0^\infty dy \int_0^y dx\, \delta(z - x) e^{-\lambda x - \mu y} + \lambda\mu \int_0^\infty dx \int_0^x dy\, \delta(z - y) e^{-\lambda x - \mu y} \\
&= \lambda\mu \int_0^\infty dy\, 1_{\{y:0<z<y\}}(y) e^{-\lambda z - \mu y} + \lambda\mu \int_0^\infty dx\, 1_{\{x:0<z<x\}}(x) e^{-\lambda x - \mu z} \\
&= 1_{[0,\infty)}(z) \left(\lambda\mu e^{-\lambda z} \int_z^\infty dy\, e^{-\mu y} + \lambda\mu e^{-\mu z} \int_z^\infty dx\, e^{-\lambda x} \right) \\
&= 1_{[0,\infty)}(z) \left(\lambda e^{-\lambda z - \mu z} + \mu e^{-\lambda z - \mu z} \right) \\
&= \begin{cases} (\lambda + \mu) e^{-(\lambda+\mu)z} & \text{for } z > 0, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

where $1_S(x)$ is the indicator function defined by:

$$1_S(x) = \begin{cases} 1 & \text{if } x \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Hence $\min(X, Y) \sim \text{Exp}(\lambda + \mu)$.

3. (a) Let $X_1, ..., X_n$ be independent Poisson random variables with X_i having parameter $i\theta$ for some $\theta > 0$. Find a real-valued sufficient statistic T, and compute its distribution. Show that the maximum likelihood estimator $\hat{\theta}$ of θ is unbiased.

   (b) For some $n \geq 2$, let $X_1, ..., X_n \sim \mathrm{Exp}(\theta)$. Find a minimal sufficient statistic T, and compute its distribution. Show that the maximum likelihood estimator $\hat{\theta}$ of θ is biased but asymptotically unbiased. Find an injective function h on $(0, \infty)$ such that, writing $\psi = h(\theta)$, the maximum likelihood estimator $\hat{\psi}$ of the new parameter ψ is unbiased.

Solution: This is our first problem when we start doing some actual statistics. Throughout this sheet, the statistical problem we are interested in is the problem of *estimation*, which can be described as follows:

The statistical estimation problem: Suppose that $\mathbf{X} = (X_1, ..., X_n)$ is a vector of independent identically distributed random variables, drawn from some distribution with pmf/pdf $f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta})$ dependent on a vector of parameters $\boldsymbol{\theta}$. Using the sample $\mathbf{X}$, we wish to:

(S1) *Estimate* the value of $\boldsymbol{\theta}$. To do so, we define an *estimator*:

$$\hat{\boldsymbol{\theta}} = T(\mathbf{X}),$$

where T is a function of the sample $\mathbf{X}$. Note any function of the sample is called a *statistic*, but a statistic that is specifically designed to estimate the value of underlying parameters of a distribution is called an *estimator*.

(S2) Provide a region which represents the *uncertainty* in our measurement; such a region, in frequentist statistics, is called a *confidence set*. A $100\alpha\%$ *confidence set* is a (random) set $A(\mathbf{X})$ such that:

$$\mathrm{Prob}\left(\boldsymbol{\theta} \in A(\mathbf{X})\right) = \alpha.$$

That is, confidence sets are random constructions dependent on the sample that have a probability α of covering the true value of $\boldsymbol{\theta}$.

In this first problem, we are interested solely in point-estimates of the parameters, and we don't provide any sort of uncertainty estimates.

There are lots of properties we might hope that estimators obey, which we explore in this problem:

Properties of estimators: We define the following properties of estimators $\hat{\boldsymbol{\theta}} = T(\mathbf{X})$:

- The *bias* of an estimator is defined to be:

$$\mathrm{bias}(\hat{\boldsymbol{\theta}}) = \mathbb{E}[\boldsymbol{\theta}] - \boldsymbol{\theta}.$$

We would like the bias to be small (in magnitude) so that estimators are likely to return the true values of the parameters. We say that a statistic is *unbiased* if $\mathbb{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$.

- The *mean squared error* of an estimator is defined to be:

$$\mathrm{mse}(\hat{\boldsymbol{\theta}}) = \mathbb{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2].$$

  This is an alternative measure of bias. We would again like this to be small so that estimators are closer to the true values of the parameters. Observe that:

$$\mathrm{mse}(\hat{\boldsymbol{\theta}}) = \mathbb{E}[(\hat{\boldsymbol{\theta}} - \mathbb{E}[\hat{\boldsymbol{\theta}}])^2 + (\mathbb{E}[\boldsymbol{\theta}] - \theta)^2] = \mathrm{var}(\hat{\boldsymbol{\theta}}) + \mathrm{bias}(\hat{\boldsymbol{\theta}})^2.$$

  This relation is called the *bias-variance tradeoff*. It shows that mean square error can sometimes be reduced by having a biased estimator with a lower variance.

- An estimator $\hat{\boldsymbol{\theta}} = T(\mathbf{X})$ is called *sufficient* for $\boldsymbol{\theta}$ if for all $\mathbf{t}$, $f_{\mathbf{X}|T(\mathbf{X})=\mathbf{t}}(\mathbf{x})$ is independent of $\boldsymbol{\theta}$. How should we think about this? It encodes the idea that if we have already computed $T(\mathbf{X})$ from the sample, then any further knowledge that could be extracted from the sample will not give further information on $\boldsymbol{\theta}$. It is desirable for an estimator to be sufficient, else we are not exploiting all the information that we could from the sample $\mathbf{X}$!

- A sufficient estimator $T(\mathbf{X})$ is called *minimal* if it is a function of every other sufficient statistic. That is, for any other sufficient statistic T', we have $T'(\mathbf{x}) = T'(\mathbf{y})$ implies $T(\mathbf{x}) = T(\mathbf{y})$.

We can spot sufficient statistics and minimal sufficient statistics using two standard criterion:

Proposition: Let $T = T(\mathbf{X})$ be an estimator. We have:

(a) THE FACTORISATION THEOREM. $T(\mathbf{X})$ is *sufficient* if and only if we have:

$$f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = g(T(\mathbf{x}), \boldsymbol{\theta})h(\mathbf{x}),$$

for some functions g, h.

(b) $T(\mathbf{X})$ is minimal sufficient if:

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})}$$

is independent of $\boldsymbol{\theta}$ if and only if $T(\mathbf{x}) = T(\mathbf{y})$.

These criteria will be consistently useful throughout the sheet.

(a) Now we have reviewed the theory from the lectures, we can start the problem properly. Observe that the distribution of $\mathbf{X} = (X_1, ..., X_n)$ is given by:

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \frac{\theta^{x_1}(2\theta)^{x_2}...(n\theta)^{x_n}}{x_1!...x_n!}e^{-\theta}e^{-2\theta}...e^{-n\theta} = \left(\theta^{x_1+\cdots+x_n}e^{-\frac{1}{2}n(n+1)\theta}\right)\cdot\left(\frac{1^{x_1}...n^{x_n}}{x_1!...x_n!}\right),$$

which by the factorisation theorem shows that:

$$T(\mathbf{X}) = \sum_{k=1}^{n} X_k$$

is a sufficient statistic. It is a sum of Poisson variables, which is itself Poisson, hence $T(\mathbf{X}) \sim \text{Po}\left(\frac{1}{2}n(n+1)\theta\right)$.

The maximum likelihood estimator $\hat{\theta}$ of θ can be obtained by maximising $f_{\mathbf{X}}(\mathbf{x}|\theta)$, or equivalently, maximising $\log(f_{\mathbf{X}}(\mathbf{x}|\theta))$. We have:

$$\frac{\partial}{\partial\theta}\log(f_{\mathbf{X}}(\mathbf{x}|\theta)) = \frac{x_1 + ... + x_n}{\theta} - \frac{1}{2}n(n+1),$$

which reveals the maximum likelihood estimator is indeed:

$$\hat{\theta} = \frac{2}{n(n+1)}\sum_{k=1}^{n} X_k.$$

To show it is unbiased, we compute $\mathbb{E}[\hat{\theta}] - \theta$. We have:

$$\mathbb{E}[\hat{\theta}] = \frac{2}{n(n+1)}\sum_{k=1}^{n} k\theta = \frac{2\theta}{n(n+1)}\cdot\frac{n(n+1)}{2} = \theta,$$

and hence the MLE is indeed unbiased.

(b) The distribution of $\mathbf{X} = (X_1, ..., X_n)$ is given by:

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \theta^n e^{-\theta(x_1+...+x_n)}.$$

By the factorisation theorem, a sufficient statistic is:

$$T(\mathbf{X}) = \sum_{k=1}^{n} X_k.$$

We can prove this is minimal as follows. Observe that:

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_{\mathbf{X}}(\mathbf{y}|\theta)} = e^{-\theta(T(\mathbf{x})-T(\mathbf{y}))},$$

which is independent of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Hence T is also minimal. This is enough for both minimal and sufficient.

On the other hand, the maximum likelihood estimator can be computed by taking the logarithmic derivative:

$$\frac{\partial}{\partial\theta}f_{\mathbf{X}}(\mathbf{x}|\theta) = \frac{n}{\theta} - (x_1 + \cdots + x_n),$$

hence the maximum likelihood estimator is:

$$\hat{\theta} = \frac{n}{\sum_{k=1}^{n} X_k}.$$

To show this is biased, we need the distribution of the maximum likelihood estimator. Observe that if $X_i \sim \text{Exp}(\theta)$, then $X_i \sim \Gamma(1, \theta)$, because the exponential distribution is a special case of the gamma distribution. In Question 2 we showed that the sum of two gamma distributions is gamma-distributed, and hence we have that:

$$\sum_{k=1}^{n} X_k \sim \Gamma(n, \theta).$$

It follows by the law of the unconscious statistician that:

$$\mathbb{E}\left[\frac{n}{\sum_{k=1}^{n} X_k}\right] = \int_0^\infty \frac{n}{x} \cdot \frac{\theta^n x^{n-1} e^{-\theta x}}{\Gamma(n)}\, dx = \frac{n\theta}{(n-1)} \int_0^\infty \frac{\theta^{n-1} x^{n-2} e^{-\theta x}}{\Gamma(n-1)}\, dx = \frac{n\theta}{n-1},$$

where the integral must be one because it is integral over the distribution function for $\Gamma(n-1, \theta)$. We have used the property of the gamma function $\Gamma(z+1) = z\Gamma(z)$. It is asymptotically unbiased, since as $n \to \infty$, we have:

$$\mathbb{E}\left[\frac{n}{\sum_{k=1}^{n} X_k}\right] \to \theta,$$

as required.

Finally we are asked to find an injective function $h : (0, \infty) \to \mathbb{R}$ such that by writing $\psi = h(\theta)$, the maximum likelihood estimator of $\hat{\psi}$. We use the invariance property of the MLE:

> **Invariance of the MLE:** If $\hat{\theta}$ is the maximum likelihood estimator for the parameter θ, and $\alpha = g(\theta)$ is a function of θ, then the maximum likelihood estimator for the parameter α is $\hat{\alpha} = g(\hat{\theta})$.

In our case, we choose $\psi = h(\theta) = 1/\theta$. Then:

$$\mathbb{E}[\hat{\psi}] = \mathbb{E}\left[\frac{1}{\hat{\theta}}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right].$$

We already determined the distribution of this sum though, given by $\Gamma(n, \theta)$. Hence the expectation is:

$$\mathbb{E}[\hat{\psi}] = n\frac{n}{\theta} = \frac{1}{\theta} = \psi,$$

so this estimator is unbiased.

4. For some $n \geq 2$ let $X_1, ..., X_n \sim \text{Unif}(\theta, 2\theta)$ for some $\theta > 0$. Show that $\tilde{\theta} = 2X_1/3$ is an unbiased estimator of θ. Use the Rao-Blackwell theorem to find an unbiased estimator $\hat{\theta}$ which is a function of a minimal sufficient statistic and which satisfies $\text{Var}(\hat{\theta}) < \text{Var}(\tilde{\theta})$ for all $\theta > 0$.

Solution: The distribution of $\mathbf{X} = (X_1, ..., X_n)$ is given by:

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \begin{cases} 1/\theta^n, & \text{if } \theta \leq x_1, ..., x_n \leq 2\theta, \\ 0, & \text{otherwise.} \end{cases} = \frac{1}{\theta^n} 1_{[\theta, 2\theta]^2}(\min\{x_1, ..., x_n\}, \max\{x_1, ..., x_n\}).$$

We note that:

$$\mathbb{E}[X_1] = \int_{\theta}^{2\theta} \frac{x}{\theta}\, dx = \frac{1}{\theta}\left[\frac{x^2}{2}\right]_{\theta}^{2\theta} = \frac{(2\theta)^2}{2\theta} - \frac{\theta^2}{2\theta} = \frac{3\theta}{2}.$$

Hence $\mathbb{E}[\tilde{\theta}] = \frac{2}{3}\mathbb{E}[X_1] = \theta$, so indeed $\tilde{\theta}$ is an unbiased estimator of θ.

Next, we recall the Rao-Blackwell theorem:

The Rao-Blackwell theorem: Let T be a sufficient statistic for θ, and let $\tilde{\theta}$ be an estimator for θ with $\mathbb{E}[\tilde{\theta}^2] < \infty$ for all θ. Define:

$$\hat{\theta}(\mathbf{x}) := \mathbb{E}\left[\tilde{\theta}(\mathbf{X}) \middle| T(\mathbf{X}) = T(\mathbf{x})\right]$$

for all $\mathbf{x}$. Then $\hat{\theta}$ has the same bias as $\tilde{\theta}$, and for all θ, we have:

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \leq \mathbb{E}[(\tilde{\theta} - \theta)^2].$$

The inequality is strict unless $\tilde{\theta}$ is a function of T.

Proof: First, observe that $\hat{\theta}(\mathbf{x})$ is well-defined, since $\mathbb{E}[\tilde{\theta}(\mathbf{X})]$ could depend on θ, but since T is sufficient, we have that the condition distribution of $\mathbf{X}$ given $T(\mathbf{X})$ is independent of θ. Hence $\mathbb{E}[\tilde{\theta}(\mathbf{X})|T(\mathbf{X}) = T(\mathbf{x})]$ is independent of θ, and is a function of the data solely.

Note that by the law of total expectation, we have:

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[\mathbb{E}[\tilde{\theta}|T(\mathbf{X}) = T(\mathbf{x})]] = \mathbb{E}[\tilde{\theta}],$$

so indeed they have the same bias.

Similarly, by the law of total variance, we have:

$$\text{var}(\tilde{\theta}) = \mathbb{E}[\text{var}(\tilde{\theta}|T)] + \text{var}(\mathbb{E}[\tilde{\theta}|T]) = \mathbb{E}[\text{var}(\tilde{\theta}|T)] + \text{var}(\hat{\theta}).$$

Hence $\text{var}(\tilde{\theta}) \geq \text{var}(\hat{\theta})$, and it follows that:

$$\text{mse}(\tilde{\theta}) \geq \text{mse}(\hat{\theta}),$$

since the estimators have the same bias. Equality holds if and only if $\text{var}(\tilde{\theta}|T) = 0$. □

In the case of our problem, we must therefore start by finding a minimal sufficient statistic. First, observe from the way of writing the pdf above, we have:

$$T(\mathbf{x}) = (\min(x_1, ..., x_n), \max(x_1, ..., x_n))$$

is a sufficient statistic, by the factorisation theorem. It is minimal since:

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_{\mathbf{Y}}(\mathbf{y}|\theta)} = \frac{1_{[\theta,2\theta]^2}(T(\mathbf{x}))}{1_{[\theta,2\theta]^2}(T(\mathbf{y}))}.$$

This is independent of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. As a result, we can use the Rao-Blackwell theorem to find an unbiased estimator which is a function of this minimal sufficient statistic. We have:

$$\begin{aligned}
\hat{\theta} &= \mathbb{E}\left[\frac{2}{3}X_1|T\right] \\
&= \frac{2}{3}\mathbb{E}\left[X_1|T \text{ and } X_1 = \min(\mathbf{x})\right]\mathbb{P}(X_1 = \min(\mathbf{x})) + \frac{2}{3}\mathbb{E}\left[X_1|T \text{ and } X_1 = \max(\mathbf{x})\right]\mathbb{P}(X_1 = \max(\mathbf{x})) \\
&\qquad + \frac{2}{3}\mathbb{E}\left[X_1|T \text{ and } \min(\mathbf{x}) < X_1 < \max(\mathbf{x})\right]\mathbb{P}(\min(\mathbf{x}) < X_1 < \max(\mathbf{x})) \\
&= \frac{2}{3}\left(\frac{\min(\mathbf{x})}{n} + \frac{\max(\mathbf{x})}{n} + \frac{n-2}{2n}\left(\max(\mathbf{x}) + \min(\mathbf{x})\right)\right) \\
&= \frac{1}{3}\left(\min(\mathbf{x}) + \max(\mathbf{x})\right).
\end{aligned}$$

5. Let $X_1, ..., X_n \sim \text{Unif}(0, \theta)$. Find the maximum likelihood estimator $\hat{\theta}$ of θ. By considering the distribution of $\hat{\theta}/\theta$ and for $\alpha \in (0, 1)$, find an appropriate, one-sided $100(1-\alpha)$% confidence interval for θ based on $\hat{\theta}$.

Solution: If $\mathbf{X} = (X_1, ..., X_n)$, the density function of $\mathbf{X}$ is given by:

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \begin{cases} 1/\theta^n, & \text{if } 0 \leq x_1, ..., x_n \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, if $x_1, ..., x_n$ are fixed, we have that $1/\theta^n$ is maximised over the range $x_1, ..., x_n \leq \theta$ by choosing θ to be as small as possible, i.e. the maximum likelihood estimator should be taken to be $\hat{\theta} = \max(X_1, ..., X_n)$.

The distribution of $\hat{\theta}/\theta$ is most easily obtained by considering the CDFs. We have:

$$\begin{aligned} \mathbb{P}(\hat{\theta}/\theta \leq z) &= \mathbb{P}\left(\max(X_1, ..., X_n)/\theta \leq z\right) \\ &= \mathbb{P}(X_1 \leq \theta z)\mathbb{P}(X_2 \leq \theta z)...\mathbb{P}(X_n \leq \theta z) \\ &= \mathbb{P}(X_1 \leq \theta z)^n \\ &= \begin{cases} 1, & \text{if } z > 1, \\ \left(\int_0^{\theta z} \frac{1}{\theta}\, dx\right)^n, & \text{if } 0 < z < 1, \\ 0, & \text{if } z < 0. \end{cases} \\ &= \begin{cases} 1, & \text{if } z > 1, \\ z^n, & \text{if } 0 < z < 1, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

We have just computed the probability $\mathbb{P}(\hat{\theta} \leq \theta \leq \frac{\hat{\theta}}{z}) = \mathbb{P}(z \leq \hat{\theta}/\theta \leq 1) = 1 - z^n$. Choosing $z^n = \alpha$, i.e. $z = \alpha^{1/n}$, we obtain the confidence interval, given by:

$$\left(\max(X_1, ..., X_n), \frac{\max(X_1, ..., X_n)}{\alpha^{1/n}}\right).$$

This is one-sided in the sense that $\mathbb{P}(\hat{\theta} \leq \theta) = 1$, because the maximum of the sample will never exceed θ.

6. Suppose that $X_1 \sim N(\theta_1, 1)$ and $X_2 \sim N(\theta_2, 1)$ independently, where θ_1 and θ_2 are unknown. Show that both the square S and the circle C in $\mathbb{R}^2$ given by:

$$S = \{(\theta_1, \theta_2) : |\theta_1 - X_1| < 2.236, |\theta_2 - X_2| < 2.236\},$$

$$C = \{(\theta_1, \theta_2) : (\theta_1 - X_1)^2 + (\theta_2 - X_2)^2 \leq 5.991\}$$

are 95% confidence sets for (θ_1, θ_2). *Hint:* $\Phi(2.236) = (1+\sqrt{0.95})/2$ *where* Φ *is the distribution function of a* $N(0, 1)$ *random variable.* What might be a sensible criterion for choosing between S and C?

Solution: For the square, we have:

$$\mathbb{P}((\theta_1, \theta_2) \in S(X_1, X_2)) = \mathbb{P}(|\theta_1 - X_1| < 2.236)^2,$$

by independence. This probability is given by:

$$\mathbb{P}(|\theta_1 - X_1| < 2.236) = \mathbb{P}(-2.236 < X_1 - \theta_1 < 2.236) = \Phi(2.236) - \Phi(-2.236) = 2\Phi(2.236) - 1,$$

for Φ the cumulative distribution function of the unit normal distribution (since $X_1 - \theta_1 \sim N(0, 1)$). In particular, we have:

$$\mathbb{P}((\theta_1, \theta_2) \in S(X_1, X_2)) = (2\Phi(2.236) - 1)^2 = \left(\sqrt{0.95}\right)^2 = 0.95,$$

so this is indeed a 95% confidence set.

In the case of the circle, we have that the likelihood is given by:

$$f_{(X_1,X_2)}((x_1, x_2)|(\theta_1, \theta_2)) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1-\theta_1)^2 - \frac{1}{2}(x_2-\theta_2)^2}.$$

In particular, we have:

$$\text{Prob}((X_1, X_2) \in C) = \frac{1}{2\pi} \iint\limits_C e^{-\frac{1}{2}(x_1-\theta_1)^2 - \frac{1}{2}(x_2-\theta_2)^2} \, dx_1 dx_2.$$

Introducing polar coordinates around the point (θ_1, θ_1) in the integral, we have:

$$\begin{aligned}
\frac{1}{2\pi} \iint\limits_C e^{-\frac{1}{2}(x_1-\theta_1)^2 - \frac{1}{2}(x_2-\theta_2)^2} \, dx_1 dx_2 &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\sqrt{5.991}} r e^{-\frac{1}{2}r^2} \, dr d\theta \\
&= \left[-e^{-\frac{1}{2}r^2}\right]_0^{\sqrt{5.991}} \\
&= 1 - e^{-\frac{1}{2} \cdot 5.991} \\
&\approx 0.9499884...,
\end{aligned}$$

as required (we see that the exact value for the radius R of the circle in this case should be such that $0.95 = 1 - e^{-\frac{1}{2}R}$, i.e. $R = -2\log(0.05)$).

Finally, we are asked how to choose between the confidence sets. Clearly, it is better to have smaller confidence sets because then we are 'more certain' of where our parameters lie! The area of the first is $(2 \cdot 2.236)^2 \approx 19.998784$, whilst the area of the second is $\pi \cdot 5.991 \approx 18.8213...$. As a result, the circle should be preferred.

7. Suppose the number of defects in a silicon wafer can be modelled with a Poisson distribution for which the parameter λ is known to be either 1 or 1.5. Suppose the prior mass function for λ is:

$$\pi_\lambda(1) = 0.4, \qquad \pi_\lambda(1.5) = 0.6.$$

A random sample of five wafers finds $x = (3, 1, 4, 6, 2)$ defects respectively. Show that the posterior distribution for λ given x is:

$$\pi_{\lambda|Z}(1|x) = 0.012, \qquad \pi_{\lambda|X}(1.5|x) = 0.988.$$

Solution: The posterior distribution is:

$$\pi_{\lambda|Z}(\lambda|x) \propto \pi_{Z|\lambda}(x|\lambda)\pi_\lambda(\lambda).$$

Observe that:

$$\pi_{Z|\lambda}(x|\lambda) = \prod_{n\in x} \frac{\lambda^n e^{-\lambda}}{n!} = \frac{\lambda^{3+1+4+6+2} e^{-5\lambda}}{3!4!6!2!}.$$

Hence we have:

$$\pi_{\lambda|Z}(1|x) = \frac{(e^{-5}/(3!4!6!2!) \cdot 0.4}{(e^{-5}/(3!4!6!2!) \cdot 0.4 + ((1.5)^{16}e^{-7.5}/(3!4!6!2!) \cdot 0.6} \approx 0.0122137...,$$

$$\pi_{\lambda|Z}(1.5|x) = \frac{((1.5)^{16}e^{-7.5}/(3!4!6!2!) \cdot 0.6}{(e^{-5}/(3!4!6!2!) \cdot 0.4 + ((1.5)^{16}e^{-7.5}/(3!4!6!2!) \cdot 0.6} \approx 0.987786...$$

8.

(a) Suppose $X = (X_1, ..., X_n)$ has probability density function $f_X(\cdot; \theta)$, and suppose T is a sufficient statistic for θ. Let $\hat{\theta}_{\text{MLE}}$ be the unique maximum likelihood estimator of θ. Show that $\hat{\theta}_{\text{MLE}}$ is a function of T.

(b) Now adopt a Bayesian perspective, and suppose that the parameter θ has a prior density function π_θ. Let the estimator $\hat{\theta}_{\text{Bayes}}$ be the unique minimiser of the expected value of the loss function L under the posterior distribution. Show that $\hat{\theta}_{\text{Bayes}}$ is also a function of T.

Solution: (a) Since T is a sufficient statistic, by the factorisation theorem we may write:

$$f_{\mathbf{X}}(\mathbf{x}; \theta) = g(\theta; T(\mathbf{x}))h(\mathbf{x}).$$

The maximum likelihood estimator is then:

$$\hat{\theta}_{\text{MLE}} = \text{argmax}_\theta \left(g(\theta; T(\mathbf{x}))h(\mathbf{x}) \right) = \text{argmax}_\theta \left(g(\theta; T(\mathbf{x})) \right),$$

so indeed $\hat{\theta}_{\text{MLE}}$ is a function of $T = T(\mathbf{x})$.

(b) In a Bayesian approach, the posterior density is given by:

$$\pi_{\theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{\mathbf{X}|\theta}(\mathbf{x}; \theta)\pi_\theta(\theta)}{f_{\mathbf{X}}(\mathbf{x})}.$$

The Bayes estimator is the minimiser of:

$$h(\delta) = \int_\Theta L(\theta, \delta)\pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta.$$

In particular, we have:

$$\begin{aligned}
\hat{\theta}_{\text{Bayes}} &= \text{argmin}_\delta \int_\Theta L(\theta, \delta)\pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \\
&= \text{argmin}_\delta \int_\Theta L(\theta, \delta) \frac{f_{\mathbf{X}|\theta}(\mathbf{x}; \theta)\pi_\theta(\theta)}{f_{\mathbf{X}}(\mathbf{x})}\, d\theta \\
&= \text{argmin}_\delta \int_\Theta L(\theta, \delta) \frac{g(\theta; T(\mathbf{x}))h(\mathbf{x})\pi_\theta(\theta)}{f_{\mathbf{X}}(\mathbf{x})}\, d\theta \\
&= \text{argmin}_\delta \int_\Theta L(\theta, \delta) g(\theta; T(\mathbf{x}))\pi_\theta(\theta)\, d\theta.
\end{aligned}$$

So again, $\hat{\theta}_{\text{Bayes}}$ only depends on the data through $T(\mathbf{x})$.

9. Let $X_1, ..., X_n$ be independent and identically distributed with conditional probability density function $f(x|\theta) = \theta x^{\theta-1} 1_{\{0 \leq x \leq 1\}}$ for some $\theta > 0$. Suppose the prior distribution for θ is $\Gamma(\alpha, \lambda)$. Find the posterior distribution of θ given $X = (X_1, ..., X_n)$ and the Bayesian point estimator of θ under the quadratic loss function.

Solution: The likelihood is:

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \theta^n (x_1 ... x_n)^{\theta-1} 1_{[0,1]^2}(\min\{x_1, ..., x_n\}, \max\{x_1, ..., x_n\}).$$

The posterior distribution is therefore:

$$\begin{aligned}
\pi_{\theta|\mathbf{X}}(\theta|\mathbf{x}) &\propto f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi_\theta(\theta) \\
&= \theta^n (x_1 ... x_n)^{\theta-1} 1_{[0,1]^2}(\min\{x_1, ..., x_n\}, \max\{x_1, ..., x_n\}) \cdot \theta^{\alpha-1} e^{-\lambda\theta} \\
&\propto \theta^{n+\alpha-1} \exp\left((\theta - 1)\log(x_1 ... x_n)\right) e^{-\lambda\theta} \\
&\propto \theta^{n+\alpha-1} \exp\left(-\left(\lambda - \sum_{k=1}^{n} \log(x_k)\right)\theta\right).
\end{aligned}$$

Thus we see that:

$$\theta|\mathbf{X} \sim \text{Gamma}\left(n + \alpha, \lambda - \sum_{k=1}^{n} \log(X_k)\right).$$

The Bayesian point estimator under the quadratic loss function is:

$$\begin{aligned}
\hat{\theta}_{\text{Bayes}} &= \text{argmin}_\delta \int_\Theta L(\theta, \delta)\pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \\
&= \arg\min_\delta \int_0^\infty (\theta - \delta)^2 \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta.
\end{aligned}$$

Observe that if $h(\delta)$ is the integral in the argument of the argmin_δ on the right hand side, we have at the Bayes estimator:

$$0 = h'(\delta) = 2\int_0^\infty (\theta - \delta)\pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta,$$

so that:

$$\delta \int_0^\infty \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta = \int_0^\infty \theta \pi_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta.$$

This implies that:

$$\hat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta|\mathbf{X}] = \frac{n + \alpha}{\lambda - \sum_{k=1}^{n} \log(X_k)}.$$

10. (**Law of small numbers**) For each $n \in \mathbb{N}$, let $X_{n1}, ..., X_{nn} \sim \text{Bernoulli}(p_n)$ and let $S_n = \sum_{i=1}^{n} X_{ni}$. Prove that if $np_n \to \lambda \in (0, \infty)$ as $n \to \infty$, then for each $x \in \{0, 1, 2, ...\}$,

$$\mathbb{P}(S_n = x) \to \mathbb{P}(Y = x),$$

as $n \to \infty$ where $Y \sim \text{Poisson}(\lambda)$.

Solution: Observe that immediately we have $S_n \sim \text{Bin}(n, p_n)$. Hence we have:

$$\begin{aligned}
\mathbb{P}(S_n = x) &= \binom{n}{x} p_n^x (1 - p_n)^{n-x} \\
&= \frac{n!}{x!(n-x)!} p_n^x (1 - p_n)^{n-x} \\
&= \frac{1}{x!} \underbrace{\frac{n(n-1)...(n-x+1)}{n^x}}_{\to 1} \underbrace{(np_n)^x}_{\to \lambda^x} \underbrace{\left(1 - \frac{np_n}{n}\right)^{n-x}}_{\to e^{-\lambda}} \\
&= \frac{\lambda^x e^{-\lambda}}{x!},
\end{aligned}$$

which is a Poisson distribution with parameter λ, as required.

11. For some $n \geq 3$, let $\epsilon_1, ..., \epsilon_n \sim N(0,1)$, set $X_1 = \epsilon_1$ and $X_i = \theta X_{i-1} + (1-\theta^2)^{1/2}\epsilon_i$ for $i = 2, ..., n$ and some $\theta \in (-1, 1)$. Find a sufficient statistic for θ that takes values in a subset of $\mathbb{R}^3$.

Solution: We have, by nesting conditional probabilities, that:

$$f_{\mathbf{X}}(x_1, ..., x_n) = f_{X_n|X_1,...,X_{n-1}}(x_n) f_{X_{n-1}|X_1,...,X_{n-2}}(x_{n-1}) ... f_{X_1}(x_1).$$

Now observe that:

$$\mathbb{E}[X_n | X_{n-1} = x_{n-1}] = \theta x_{n-1}, \qquad \mathrm{Var}[X_n | X_{n-1} = x_{n-1}] = 1 - \theta^2.$$

Hence the final distribution is:

$$\begin{aligned} f_{\mathbf{X}}(x_1, ..., x_n) &= \frac{1}{\sqrt{2\pi(1-\theta^2)}} \exp\left(-\frac{1}{2(1-\theta^2)}(x_n - \theta x_{n-1})^2\right) ... \frac{1}{\sqrt{2\pi(1-\theta^2)}} \exp\left(-\frac{1}{2(1-\theta^2)}(x_2 - \theta x_1)^2\right) \\ &\quad \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_1^2\right) \\ &= \frac{1}{\sqrt{(2\pi)^n (1-\theta^2)^{(n-1)}}} \exp\left(-\frac{1}{2(1-\theta^2)}\left((x_n^2 + ... + x_2^2) - 2\theta(x_n x_{n-1} + ... x_2 x_1) + \theta^2(x_{n-1}^2 + ... + x_1^2) - \frac{1}{2}x_1^2\right)\right) \\ &= \frac{1}{\sqrt{(2\pi)^n (1-\theta^2)^{(n-1)}}} \exp\left(-\frac{1}{2(1-\theta^2)}\left((x_n^2 + ... + x_1^2) - 2\theta(x_n x_{n-1} + ... x_2 x_1) + \theta^2(x_{n-1}^2 + ... + x_2^2)\right)\right). \end{aligned}$$

Hence, by the factorisation theorem, we immediately see that a sufficient statistic is:

$$T(\mathbf{x}) = \begin{pmatrix} x_1^2 + ... + x_n^2 \\ x_1 x_2 + ... + x_{n-1} x_n \\ x_2^2 + ... + x_{n-1}^2 \end{pmatrix}.$$

12. (**Harder**) Let $\hat{\theta}$ be an unbiased estimator of $\theta \in \Theta = \mathbb{R}$ satisfying $\mathbb{E}_\theta(\hat{\theta}^2) < \infty$ for all $\theta \in \Theta$. We say that $\hat{\theta}$ is a uniform minimum variance unbiased (UMVU) estimator if $\mathrm{Var}_\theta(\hat{\theta}) \leq \mathrm{Var}_\theta(\tilde{\theta})$ for all $\theta \in \Theta$ and any other unbiased estimator $\tilde{\theta}$. Prove that a necessary and sufficient condition for $\hat{\theta}$ to be a UMVU estimator is that $\mathbb{E}_\theta(\hat{\theta}U) = 0$ for all $\theta \in \Theta$ and all estimators U with $\mathbb{E}_\theta(U) = 0$ and $\mathbb{E}_\theta(U^2) < \infty$ (i.e. '$\hat{\theta}$ is uncorrelated with every unbiased estimator of 0'). Is the estimator $\hat{\theta}$ in Question 4 a UMVU estimator?

Solution: Suppose that $\hat{\theta}$ is a UMVU estimator for θ. Given any unbiased zero estimator U, we have that, for any λ, $\hat{\theta} + \lambda U$ is also an unbiased estimator of θ. Hence by assumption

$$\mathrm{Var}_\theta(\hat{\theta} + \lambda U) \geq \mathrm{Var}_\theta(\hat{\theta}).$$

Expanding the left hand side, this becomes:

$$\lambda^2 \mathrm{Var}_\theta(U) + \lambda \mathrm{Cov}_\theta(\hat{\theta}, U) \geq 0.$$

However, this is a quadratic which takes on negative values unless $0 = \mathrm{Cov}_\theta(\hat{\theta}, U) = \mathbb{E}[\hat{\theta}U] - \mathbb{E}[\hat{\theta}]\mathbb{E}[U] = \mathbb{E}[\hat{\theta}U]$ as required. This proves necessity.

On the other hand, we can prove sufficiency as follows. Suppose that $\hat{\theta}$ is an unbiased estimator satisfying $\mathbb{E}[\hat{\theta}U] = 0$ for all unbiased zero estimators U. Suppose that $\tilde{\theta}$ is any other unbiased estimator. Then $\hat{\theta} - \tilde{\theta}$ is an unbiased zero estimator, since $\mathbb{E}[\hat{\theta} - \tilde{\theta}] = \theta - \theta = 0$. It follows that:

$$0 = \mathbb{E}[\hat{\theta}(\hat{\theta} - \tilde{\theta})] = \mathbb{E}[\hat{\theta}^2] - \mathbb{E}[\hat{\theta}\tilde{\theta}].$$

Since $\mathbb{E}[\hat{\theta}] = \mathbb{E}[\tilde{\theta}]$, this implies that $\mathrm{Var}_\theta(\hat{\theta}) = \mathrm{Cov}_\theta(\hat{\theta}, \tilde{\theta}) \leq \sqrt{\mathrm{Var}_\theta(\hat{\theta})\mathrm{Var}_\theta(\tilde{\theta})}$. It follows that:

$$\mathrm{Var}_\theta(\hat{\theta}) \leq \mathrm{Var}_\theta(\tilde{\theta}),$$

as required. □

The estimator in Question 4 is:

$$\hat{\theta}(\mathbf{X}) = \frac{1}{3}\left(\max(\mathbf{X}) + \min(\mathbf{X})\right).$$

This is unbiased, as we showed in Question 4. It is a UMVU estimator if and only if $\mathbb{E}[\hat{\theta}U] = 0$ for any unbiased zero estimator U. Idea is to produce a U where this is non-zero. Really difficult, see lecturer's solutions.

Part IB: Statistics
Examples Sheet 2 Solutions

Please send all comments and corrections to jmm232@cam.ac.uk.

1. Let X have density function:

$$f(x|\theta) = \frac{\theta}{(x+\theta)^2}, \qquad x > 0,$$

where $\theta \in (0, \infty)$ is an unknown parameter. Find the likelihood ratio test of size 0.05 of $H_0 : \theta = 1$ against $H_1 : \theta = 2$ and show that the probability of Type II error is $19/21$.

Solution: The likelihood ratio test has the critical region:

$$\Lambda_x(H_0; H_1) = \frac{2/(x+2)^2}{1/(x+1)^2} = \frac{2(x+1)^2}{(x+2)^2}.$$

Observe that since:

$$\frac{d}{dx}\log(\Lambda_x(H_0; H_1)) = \frac{2}{x+1} - \frac{2}{x+2} > 0,$$

we have that the likelihood is monotonically increasing. Hence the likelihood ratio test rejects H_0 at 5% in the region $X > c$ where $\mathbb{P}_{\theta=1}(X > c) = 0.05$. In other words, we need:

$$\int_c^\infty \frac{1}{(1+x)^2} = \left[-\frac{1}{1+x}\right]_c^\infty = \frac{1}{1+c} = 0.05.$$

Rearranging, we have $c = 19$. The probability of a Type II error is:

$$\mathbb{P}_{\theta=2}(X \leq c) = \int_0^{19} \frac{2}{(x+2)^2} = \left[-\frac{2}{x+2}\right]_0^{19} = 1 - \frac{2}{21} = \frac{19}{21},$$

as required.

2. Let $X \sim N(\mu, 1)$ where μ is unknown. Find the most powerful test of sizes 0.05 and 0.01 for the following hypotheses:

(a) $H_0 : \mu = 0$ vs $H_1 : \mu = 4$.

(b) $H_0 : \mu = 4$ vs $H_1 : \mu = 0$.

Explain how to interpret your results when the realised value is $X(\omega) = 2.1$.

Solution: (a) The Neyman-Pearson lemma tells us that likelihood ratio is the uniformly most powerful test, for any size. In the first case, the likelihood ratio is:

$$\Lambda_x(H_0; H_1) = \frac{e^{-\frac{1}{2}(x-4)^2}}{e^{-\frac{1}{2}x^2}} = e^{-\frac{1}{2}(x-4)^2 + \frac{1}{2}x^2} = e^{4x-8}.$$

This is a monotonically increasing function, so the critical region is of the form $\{X > c\}$. The probability of a Type I error is therefore:

$$\mathbb{P}_{\mu=0}(X > c) = 1 - \mathbb{P}_{\mu=0}(X \leq c) = 1 - \Phi(c).$$

Hence, for the tests of size 0.05 and 0.01, we have tests:

- $C_1 = (\Phi^{-1}(0.95), \infty) = (1.64, \infty)$;
- $C_2 = (\Phi^{-1}(0.99), \infty) = (2.33, \infty)$.

Hence if $X = 2.1$, we reject H_0 at 95% significance, and do not reject H_0 at 99% significance.

(b) In the second case, the likelihood ratio is:

$$\Lambda_x(H_0; H_1) = \frac{e^{-\frac{1}{2}x^2}}{e^{-\frac{1}{2}(x-4)^2}} = e^{8-4x}.$$

This is a monotonically decreasing function, so the critical region is of the form $\{X < c\}$. The probability of a Type I error is therefore:

$$\mathbb{P}_{\mu=4}(X < c) = \mathbb{P}_{\mu=4}(X - 4 < c - 4) = \Phi(c - 4).$$

Hence, for the tests of size 0.05 and 0.01, we have tests:

- $C_1 = (-\infty, 4 + \Phi^{-1}(0.05)) = (-\infty, 4 - 1.64) = (-\infty, 2.36)$;
- $C_2 = (-\infty, 4 + \Phi^{-1}(0.01)) = (-\infty, 4 - 2.33) = (-\infty, 1.67)$.

Hence if $X = 2.1$, we reject H_0 at 95% significance, but we do not reject H_0 at 99% significance.

3. Let $X_1, ..., X_n, Y_1, ..., Y_n$ be independent, with $X_1, ..., X_n \sim \text{Exp}(\theta_1)$ and $Y_1, ..., Y_n \sim \text{Exp}(\theta_2)$. Recalling the forms of the relevant MLEs from Sheet 1, show that the likelihood ratio of $H_0 : \theta_1 = \theta_2$ and $H_1 : \theta_1 \neq \theta_2$ is a monotone function of $|T - 1/2|$, where:

$$T = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i}.$$

By writing down the distribution of T under H_0, express the likelihood ratio test of size α in terms of $|T - 1/2|$ and the quantiles of a beta distribution.

Solution: The likelihood ratio is now for a composite hypothesis, so we take the ratio of the suprema of the likelihoods:

$$\Lambda_{\mathbf{x},\mathbf{y}}(H_0; H_1) = \frac{\sup_{\theta_1,\theta_2} \theta_1^n e^{-\theta_1(x_1+...+x_n)} \theta_2^n e^{-\theta_2(y_1+...+y_n)}}{\sup_\theta \theta^{2n} e^{-\theta(x_1+...+x_n+y_1+...+y_n)}}$$

Let:

$$\bar{x} = \sum_{k=1}^{n} x_k, \qquad \bar{y} = \sum_{k=1}^{n} y_k.$$

Then, using the expression for the maximum likelihood estimator of an exponential distribution, we have:

$$\Lambda_{\mathbf{x},\mathbf{y}}(H_0; H_1) = \frac{(n/\bar{x})^n e^{-n} (n/\bar{y})^n e^{-n}}{(2n/(\bar{x}+\bar{y}))^{2n} e^{-2n}} = \frac{1}{2^{2n}} \frac{(\bar{x}+\bar{y})^{2n}}{\bar{x}^n \bar{y}^n}$$

Now, observe that:

$$T(1-T) = \frac{\bar{x}\bar{y}}{(\bar{x}+\bar{y})^2},$$

hence we can write:

$$\Lambda_{\mathbf{x},\mathbf{y}}(H_0; H_1) = \frac{1}{2^{2n}} \left(\frac{1}{T(1-T)} \right)^n = \frac{1}{2^{2n}} \left(\frac{1}{1/4 - (T-1/2)^2} \right)^n.$$

This is evidently a monotone decreasing function of $|T - 1/2|$. Thus the critical region takes the form $\{|T - 1/2| > c\}$. Next, observe that under the null hypothesis we have that:

$$\bar{X} \sim \Gamma(n, \theta), \qquad \bar{Y} \sim \Gamma(n, \theta),$$

by the first examples sheet. Hence by Sheet 1, Question 2, we have:

$$T \sim \text{Beta}(n, n).$$

In particular, we see that to obtain a test of size α, we require c to satisfy:

$$\alpha = \mathbb{P}_{H_0}(|T - 1/2| > c) = \mathbb{P}(T - 1/2 < -c) + \mathbb{P}(T - 1/2 > c).$$

Evidently, under the null hypothesis the distribution of T is symmetric about $1/2$, so that:

$$\mathbb{P}_{H_0}(T - 1/2 > c) = \mathbb{P}_{H_0}(T - 1/2 < -c).$$

It follows that:

$$\alpha = 2\mathbb{P}_{H_0}(T > c+1/2) \quad \Rightarrow \quad 1 - \frac{\alpha}{2} = \mathbb{P}(T < c+1/2) \quad \Rightarrow \quad \text{B}_{\alpha/2} = c+1/2 \quad \Rightarrow \quad c = \text{B}_{\alpha/2} - \frac{1}{2},$$

where $\text{B}_{\alpha/2}$ is the upper $\alpha/2$ point of the $\text{Beta}(n, n)$ distribution (i.e. the point above which there is $\alpha/2$ probability remaining in the tail. Thus the critical region is:

$$|T - 1/2| > \text{B}_{\alpha/2} - \frac{1}{2}.$$

4. A machine produces biodegradable plastic articles (many of which are defective) in bunches of three articles at a time. Under the null hypothesis that each article has a constant (but unknown) probability θ of being defective, write down the probabilities $p_i(\theta)$ of a bunch having i defective articles, for $i = 0, 1, 2, 3$. In a trial run in which 512 bunches were produced, the numbers of bunches with i defective articles were 213 ($i = 0$), 228 ($i = 1$), 57 ($i = 2$) and 14 ($i = 3$). Carry out Pearson's χ^2 test at the 5% level of the null hypothesis, explaining carefully why the test statistic should be referred to the χ^2_2 distribution.

Solution: The null hypothesis is:

$$p_0(\theta) = (1-\theta)^3, \qquad p_1(\theta) = 3\theta(1-\theta)^2, \qquad p_2(\theta) = 3\theta^2(1-\theta), \qquad p_3(\theta) = \theta^3.$$

Under the null hypothesis, the number of defective articles in a trial run with 512 bunches (so $512 \cdot 3 = 1536$ articles) is distributed as $X \sim \mathrm{Bin}(1536, \theta)$. The maximum likelihood estimator is just the mean, $X/1536$, which is given by:

$$\frac{228 + 57 \cdot 2 + 14 \cdot 3}{1536} = \frac{1}{4}.$$

Hence the expected data is:

$$(e_0, e_1, e_2, e_3) = 512 \cdot (27/64, 27/64, 9/64, 1/64) = (216, 216, 72, 8).$$

It follows that the χ^2-statistic is given by:

$$\chi^2 = \frac{(216-213)^2}{216} + \frac{(216-228)^2}{216} + \frac{(72-57)^2}{72} + \frac{(8-14)^2}{8} = \frac{25}{3} \approx 8.33.$$

The unconstrained model has independent probabilities which must only sum to one. Thus it has 3 free parameters. The constrained model has one free parameter. Hence by Wilk's theorem, we expect the asymptotic distribution of the χ^2-statistic to be χ^2_2. Hence we reject above $\chi^2_2(0.05) \approx 5.991$. Thus we reject the null hypothesis.

5. Let f_0 and f_1 be probability mass functions on a countable set $\mathcal{X}$. State and prove a version of the Neyman-Pearson lemma for a size α test of $H_0 : f = f_0$ against $H_1 : f = f_1$ assuming that α is such that there exists a likelihood ratio test of exact size α.

Solution: We claim that out of all tests of size at most α, the likelihood ratio test with critical region:

$$C = \left\{ x : \frac{f_1(x)}{f_0(x)} \geq k \right\}$$

has the most power (i.e. minimises the probability of a Type II error), where k is chosen such that:

$$\mathbb{P}(\text{Type I error}) = \mathbb{P}(X \in C | H_0) = \sum_{x \in C} f_0(x) = \alpha.$$

We follow the proof in lectures that we used for continuous random variables.

Let the probability of a Type II error under the likelihood ratio test be:

$$\beta := \mathbb{P}(\text{Type II error}) = \mathbb{P}(X \notin C | H_1) = \sum_{x \in C^c} f_1(x)$$

Let C_* be the critical region of some other test with size at most α, and let α_*, $1 - \beta_*$ be its size and power respectively. Then, we have:

$$\begin{aligned}
\beta - \beta_* &= \sum_{x \in C^c} f_1(x) - \sum_{x \in C_*^c} f_1(x) \\
&= \sum_{x \in C^c \cap C_*} f_1(x) + \sum_{x \in C_\cap^c C_*^c} f_1(x) - \sum_{x \in C_*^c \cap C^c} f_1(x) - \sum_{x \in C_*^c \cap C} f_1(x) \\
&= \sum_{x \in C^c \cap C_*} f_1(x) - \sum_{x \in C_*^c \cap C} f_1(x) \\
&= \sum_{x \in C^c \cap C_*} \frac{f_1(x)}{f_0(x)} f_0(x) - \sum_{x \in C_*^c \cap C} \frac{f_1(x)}{f_0(x)} f_0(x) \\
&\leq k \left(\sum_{x \in C^c \cap C_*} f_0(x) - \sum_{x \in C_*^c \cap C} f_0(x) \right)
\end{aligned}$$

because in the first term, $x \in C^c$, and in the second term, $x \in C$ (so that $f_1(x)/f_0(x) \geq k$, and thus $-f_1(x)/f_0(x) \leq -k$). Next, add and subtract terms to give:

$$\begin{aligned}
\beta - \beta_* &\leq k \left(\sum_{x \in C^c \cap C_*} f_0(x) + \sum_{x \in C^c \cap C_*} f_0(x) + \sum_{x \in C \cap C_*} f_0(x) - \sum_{x \in C_*^c \cap C} f_0(x) - \sum_{x \in C \cap C_*} f_0(x) \right) \\
&= k \left(\sum_{x \in C_*} f_0(x) - \sum_{x \in C} f_0(x) \right) \\
&= k(\alpha_* - \alpha) \leq 0,
\end{aligned}$$

since $\alpha_* \leq \alpha$ by assumption. Thus $\beta \leq \beta_*$, and it follows that $1 - \beta \geq 1 - \beta_*$. Thus the likelihood ratio test is the most powerful test.

6. A random sample of 59 people from the planet Krypton yielded the results below.

		Eye-colour	
		Blue	Brown
Sex	Male	19	10
	Female	9	21

Carry out a Pearson's χ^2 test at the 5% level of the null hypothesis that sex and eye-colour are independent factors on Krypton. Now carry out the corresponding test at the 5% level of the null hypothesis that each of the cell probabilities is equal to $1/4$. Comment on your results.

Solution: First, we carry out the test of independence. Under H_0, the expected entries are:

$$e_{ij} = \frac{N_{i+}N_{+j}}{n^2} \qquad \Rightarrow \qquad e_{11} = \frac{(19+10)(19+9)}{59} = \frac{812}{59}, \qquad e_{12} = \frac{899}{59}, \qquad e_{21} = \frac{840}{59}, \qquad e_{22} = \frac{930}{59}.$$

whilst the observed are $o_{11} = 19$, $o_{12} = 10$, $o_{21} = 9$ and $o_{22} = 21$. Hence we have:

$$\chi^2 = \frac{(812/59 - 19)^2}{812/59} + \frac{(899/59 - 10)^2}{899/59} + \frac{(840/59 - 9)^2}{840/59} + \frac{(930/59 - 21)^2}{930/59} \approx 7.46.$$

Under H_0, the column probabilities must sum to 1, and the row probabilities must sum to 1. Hence there are $(2-1)+(2-1) = 2$ degrees of freedom. Under H_1, the probabilities must simply sum to 1, hence there are $4 - 1 = 3$ degrees of freedom. Hence we are performing a χ^2-test with 1 degree of freedom. We have $\chi^2_1(0.05) \approx 3.84$, so we reject H_0 (i.e. there is evidence to suggest the factors are not independent).

On the other hand, performing a test under H_0 where the cell probabilities equal $1/4$, we have:

$$e_{ij} = \frac{59}{4}.$$

Hence, we have:

$$\chi^2 = \frac{(59/4 - 19)^2}{59/4} + \frac{(59/4 - 10)^2}{59/4} + \frac{(59/4 - 9)^2}{59/4} + \frac{(59/4 - 21)^2}{59/4} \approx 7.64.$$

This time, under H_0, there are no free degrees. Under H_1, the probabilities simply sum to 1, hence there are $4 - 1 = 3$ degrees of freedom. Hence we are performing a χ^2-test with 3 degrees of freedom. We have $\chi^2_3(0.05) \approx 7.81$, so we do not reject H_0 (i.e. there is insufficient evidence to suggest that the factors are different from $1/4$).

7. Write down from lectures the model and hypotheses for a test of homogeneity in a two-way contingency table. By first deriving the MLEs under each hypothesis, show that the likelihood ratio and Pearson's χ^2 tests are identical to those for the independence test. Apply the homogeneity test to the data below from a clinical trial for a drug, obtained by randomly allocating 150 patients to three equal groups (so the row totals are fixed).

	Improved	No difference	Worse
Placebo	18	17	15
Half dose	20	10	20
Full dose	25	13	12

Solution: Since patients are allocated to three equal groups, the model is:

$$(N_{i1}, N_{i2}, N_{i3}) \sim \text{multinomial}(50, p_{i1}, p_{i2}, p_{i3})$$

for each row $i = 1, 2, 3$. Under the two hypotheses, we have:

- H_0 is such that for each i, we have $p_{1i} = p_{2i} = p_{3i} = p_i$ for some p_i, i.e. the probabilities depend only on the column. We require $p_1 + p_2 + p_3 = 1$.
- H_1 is such that the probabilities p_{ij} are independent. We then require $p_{i1} + p_{i2} + p_{i3} = 1$ for each row $i = 1, 2, 3$.

Under H_1, we have the likelihood:

$$L = \prod_{i,j=1}^{3} p_{ij}^{N_{ij}}$$

Taking the logarithm, we have:

$$\log(L) = \sum_{i,j=1}^{3} N_{ij} \log(p_{ij}).$$

Defining a Lagrangian where we impose the condition that the row sum condition, we have:

$$\mathcal{L} = \sum_{i,j=1}^{3} N_{ij} \log(p_{ij}) - \sum_{i=1}^{3} \lambda_i \left(\sum_{j=1}^{3} p_{ij} - 1 \right).$$

Differentiating, we have:

$$\frac{\partial \mathcal{L}}{\partial p_{ij}} = \frac{N_{ij}}{p_{ij}} - \lambda_i, \qquad \frac{\partial \mathcal{L}}{\partial \lambda_i} = 1 - \sum_{j=1}^{3} p_{ij} = 0.$$

Hence, we have:

$$p_{ij} = \frac{N_{ij}}{\lambda_i} \qquad \Rightarrow \qquad 1 = \sum_{j=1}^{3} p_{ij} = \frac{1}{\lambda_i} \sum_{j=1}^{3} N_{ij} = \frac{1}{\lambda_i} N_{i+},$$

which yields:

$$p_{ij} = \frac{N_{ij}}{N_{i+}}$$

for the maximum likelihood estimators under the alternative hypothesis.

Under H_0, we have the likelihood:

$$L = \prod_{i,j=1}^{3} p_j^{N_{ij}}$$

Taking the logarithm, we have:

$$\log(L) = \sum_{i,j=1}^{3} N_{ij} \log(p_j).$$

Defining a Lagrangian where we impose the column sum condition, we have:

$$\mathcal{L} = \sum_{i,j=1}^{3} N_{ij} \log(p_j) - \lambda \left(\sum_{j=1}^{3} p_j - 1 \right),$$

which yields the minimum through:

$$0 = \sum_{i=1}^{3} \frac{N_{ij}}{p_i} - \lambda, \qquad \sum_{j=1}^{3} p_j = 1.$$

Hence we have:

$$p_j = \sum_{i=1}^{3} \frac{N_{ij}}{\lambda} = \frac{N_{+j}}{\lambda} \qquad \Rightarrow \qquad 1 = \frac{1}{\lambda} \sum_{j=1}^{3} N_{+j} = \frac{1}{\lambda} N.$$

It follows that:

$$p_j = \frac{N_{+j}}{N}.$$

Using this data, the generalised likelihood ratio statistic is given by:

$$2 \log \Lambda(H_0; H_1) = 2 \sum_{i,j=1}^{3} N_{ij} \log \left(\frac{N_{ij} N}{N_{i+} N_{+j}} \right) = 2 \sum_{i,j=1}^{3} o_{ij} \log \left(\frac{o_{ij}}{e_{ij}} \right),$$

where $e_{ij} = N_{i+} N_{+j} / N$, which is exactly the same as the test for independence. The null hypothesis has dimension 2, whilst the alternative hypothesis has dimension 6. Hence there are $6 - 2 = 4$ degrees of freedom. This is what we would have had for an independence test too.

Applying the test, we have $N_{i+} = 50$, $N_{+1} = 63$, $N_{+2} = 40$, $N_{+3} = 47$, and $N = 150$.

$$(e_{11}, e_{12}, e_{13}, e_{21}, e_{22}, e_{23}, e_{31}, e_{32}, e_{33}) = (63/3, 40/3, 47/3, 63/3, 40/3, 47/3, 63/3, 40/3, 47/3).$$

Hence we have the χ^2-statistic:

$$\begin{aligned} \chi^2 = {} & \frac{(63/3 - 18)^2}{63/3} + \frac{(40/3 - 17)^2}{40/3} + \frac{(47/3 - 15)^2}{47/3} \\ & + \frac{(63/3 - 20)^2}{63/3} + \frac{(40/3 - 10)^2}{40/3} + \frac{(47/3 - 20)^2}{47/3} \\ & + \frac{(63/3 - 25)^2}{63/3} + \frac{(40/3 - 13)^2}{40/3} + \frac{(47/3 - 12)^2}{47/3} \\ \approx {} & 5.173 \end{aligned}$$

Observe that $\chi^2_4(0.05) = 9.488$. Hence we do not reject H_0 (i.e. we do not reject that the probabilities only depend on the column).

8. Let $X_1, ..., X_n \sim \text{Exp}(\theta)$ be identically distributed and independent. Find the likelihood ratio test of size α of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ where $\theta_1 > \theta_0$ and derive an expression for the power function. Is the test uniformly most powerful for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$? Is it uniformly most powerful for testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$?

Solution: The likelihood ratio is given by:

$$\Lambda(H_0; H_1) = \left(\frac{\theta_1}{\theta_0}\right)^n \frac{e^{-\theta_1(X_1+...+X_n)}}{e^{-\theta_0(X_1+...+X_n)}} = \left(\frac{\theta_1}{\theta_0}\right)^n \exp\left(-(\theta_1 - \theta_0)\sum_{i=1}^n X_i\right).$$

Hence, the likelihood ratio is a monotone decreasing function of:

$$T = \sum_{i=1}^n X_i.$$

It follows that the critical region is of the form $\{T < c\}$, where $\mathbb{P}(T < c|H_0) = \alpha$. Under the null hypothesis, recall that the sum of exponentials is a gamma distribution, so:

$$T = \sum_{i=1}^n X_i \sim \Gamma(n, \theta_0).$$

Recall also from the distributions sheet, that this implies $2\theta_0 T \sim \chi^2_{2n}$. As a result, have:

$$\alpha = \mathbb{P}(T < c|H_0) = \mathbb{P}(2\theta_0 T < 2\theta_0 c|H_0) = F^{(2n)}(2\theta_0 c),$$

where $F^{(2n)}$ is the cumulative distribution function of the χ^2-distribution with $2n$ degrees of freedom. Rearranging the above, we see that:

$$1 - \alpha = \mathbb{P}(2\theta_0 T > 2\theta_0 c|H_0),$$

hence we should take:

$$c = \frac{\chi^2_{2n}(1-\alpha)}{2\theta_0},$$

where $\chi^2_{2n}(1-\alpha)$ is the upper $1-\alpha$ point of the χ^2_{2n} distribution.

The *power function* is $w(\theta) = \mathbb{P}(\text{reject } H_0|\text{true value of parameter is } \theta)$, i.e. the probability of rejecting H_0 on an observation of the data, given that the true value of the parameter is θ. We would like the power function to be large on H_1 (i.e. we want the probability of rejecting H_0 to be large if the true value of θ is different from H_0), and we would like the power function to be small on H_0.

By the above calculation, the power function in this problem is:

$$w(\theta) = F^{(2n)}(2\theta c) = F^{(2n)}\left(\frac{\theta}{\theta_0}\chi^2_{2n}(1-\alpha)\right).$$

We say that the test is *uniformly most powerful* if:

- The worst possible size is at most α. That is,
$$\sup_{\theta \in \Theta_0} w(\theta) \leq \alpha,$$
where Θ_0 is the set of null hypothesis values for the parameter. Fortunately, we have $\Theta_0 = \{\theta_0\}$, hence the supremum is just $w(\theta_0) = F^{(2n)}(2\theta_0 c) = \alpha$, by construction.

- For any other test C_* with size less than or equal to α, and with power function w_*, we want $w(\theta) \geq w_*(\theta)$ for all $\theta \in \Theta_1$ (that is, our test is more likely to detect the alternative hypothesis versus any other test).

  Now, the Neyman-Pearson lemma for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ can be rephrased as saying: amongst all power functions $w_*(\theta)$ satisfying $w_*(\theta_0) \leq \alpha$ (i.e. size at most α), the one with the greatest value of $w_*(\theta_1)$ is the power function of the likelihood ratio test. This applies for each $\theta_1 > \theta_0$, and importantly for our likelihood ratio is independent of θ_1; hence $w(\theta) \geq w_*(\theta)$ for all $\theta > \theta_0$, as required. Thus the test is uniformly most powerful.

In the other case, we have $\Theta_0 = \{\theta : \theta \leq \theta_0\}$ and $\Theta_1 = \{\theta : \theta > \theta_0\}$.

- On Θ_0, we have that $w(\theta)$ is now *increasing*, and hence:

$$\sup_{\theta \in \Theta_0} w(\theta) = w(\theta_0) = \alpha,$$

  again by construction.

- For any other test with power function $w_*(\theta)$ and size at most α for these hypotheses, we have:

$$w_*(\theta_0) \leq \sup_{\theta \leq \theta_0} w_*(\theta) \leq \alpha,$$

  by definition. Thus by the Neyman-Pearson argument above, we are once again done - for all values of $\theta_1 > \theta_0$, we have that $w(\theta) \geq w_*(\theta)$, as required. The test remains the uniformly most powerful test.

9. If $X \sim N(0,1)$ and $Y \sim \chi^2_n$ are independent, we say that $T = X/\sqrt{Y/n}$ has a t-distribution with n degrees of freedom and write $T \sim T_n$. Show that the probability density function of T is:

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{1}{(n\pi)^{1/2}} \frac{1}{(1+t^2/n)^{(n+1)/2}}, \qquad t \in \mathbb{R}.$$

Solution: Let $t = x/\sqrt{y/n}$ and $w = y$. Then the inverse transformation is:

$$x = t\sqrt{w/n}, \qquad y = w.$$

The Jacobian of the inverse transformation is:

$$J(t,w) = \det \begin{pmatrix} \sqrt{w/n} & \frac{1}{2}t(nw)^{-1/2} \\ 0 & 1 \end{pmatrix} = \sqrt{w/n}.$$

Hence, the joint distribution is given by:

$$f_{(T,W)}(t,w) = \sqrt{\frac{w}{n}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(t\sqrt{w/n}^2} \cdot \frac{w^{n/2-1}e^{-w/2}}{2^{n/2}\Gamma(n/2)},$$

where recall that if $Y \sim \chi^2_n$, then $Y \sim \Gamma(n/2, 1/2)$. Simplifying, we note that:

$$f_{(T,W)}(t,w) = \frac{1}{\sqrt{2\pi n}2^{n/2}\Gamma(n/2)} e^{-\frac{w}{2}\left(\frac{t^2}{n}+1\right)} w^{(n-1)/2}.$$

The marginal density of T is therefore:

$$f_T(t) = \frac{1}{\sqrt{2\pi n}2^{n/2}\Gamma(n/2)} \int_0^\infty dw\, w^{(n-1)/2} e^{-\frac{1}{2}(t^2/n+1)w}.$$

Making the substitution $u = \frac{1}{2}\left(\frac{t^2}{n}+1\right)w$, we have:

$$du = \frac{1}{2}\left(\frac{t^2}{n}+1\right) dw,$$

and hence we have:

$$\begin{aligned} f_T(t) &= \frac{1}{\sqrt{2\pi n}2^{n/2}\Gamma(n/2)} \frac{2^{(n+1)/2}}{(t^2/n+1)^{(n+1)/2}} \int_0^\infty du\, u^{(n+1)/2-1} e^{-u}\, du \\ &= \frac{1}{\sqrt{2\pi n}2^{n/2}\Gamma(n/2)} \frac{2^{(n+1)/2}}{(t^2/n+1)^{(n+1)/2}} \Gamma\left(\frac{n+1}{2}\right) \\ &= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{1}{(\pi n)^{1/2}} \frac{1}{(t^2/n+1)^{(n+1)/2}}, \end{aligned}$$

as required.

10. Let $X_1, ..., X_n \sim N(\mu, \sigma^2)$ be independent identically distributed, where σ^2 is unknown, and suppose we are interested in testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Letting $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ and $S_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2$, show that the likelihood ratio can be expressed as:

$$L_X(H_0, H_1) = \left(1 + \frac{T^2}{n-1}\right)^{n/2},$$

where $T = \frac{n^{1/2}(\bar{X}-\mu_0)}{(S_{XX}/(n-1))^{1/2}}$. Determine the distribution of T under H_0, and hence determine the size α likelihood ratio test.

Solution: The likelihood is:

$$L(\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2\right),$$

and hence the log-likelihood is given by:

$$\log(L(\mu, \sigma^2)) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2.$$

Under the alternative hypothesis, we then require the MLEs to satisfy:

$$\frac{\partial}{\partial\sigma^2}\log(L) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (X_i - \mu)^2 = 0, \qquad \frac{\partial}{\partial\mu}\log(L) = -\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu) = 0.$$

This reveals that:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2.$$

Under the null hypothesis, we instead have that the MLE of σ^2 must satisfy:

$$\hat{\hat{\sigma}}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2.$$

It follows that the required likelihood ratio is given by:

$$\begin{aligned}
\Lambda(H_0; H_1) &= \frac{(2\pi\hat{\hat{\sigma}}^2)^{n/2} \exp\left(-\frac{1}{2\hat{\hat{\sigma}}^2}\sum_{i=1}^n (X_i - \bar{X})^2\right)}{(2\pi\hat{\sigma}^2)^{n/2} \exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^n (X_i - \mu_0)^2\right)} \\
&= \left(\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)^{n/2}
\end{aligned}$$

Now observe that:

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu_0)^2 &= \sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-\mu_0)^2 \\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\mu_0) + n(\bar{X}-\mu_0)^2 \\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2(\bar{X}-\mu_0)(n\bar{X}-n\bar{X}) + n(\bar{X}-\mu_0)^2 \\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu_0)^2.\end{aligned}$$

Hence the likelihood ratio can be rewritten as:

$$\Lambda(H_0;H_1) = \left(1+\frac{n(\bar{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)^{n/2} = \left(1+\frac{T^2}{n-1}\right)^{n/2},$$

where:

$$T = \frac{\sqrt{n}(\bar{X}-\mu_0)}{\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)^{1/2}} = \frac{\sqrt{n}(\bar{X}-\mu_0)}{(S_{XX}/(n-1))^{1/2}},$$

as required. Therefore, the likelihood ratio test rejects H_0 for large values of $|T|$.

Under H_0, we have $\bar{X} \sim N(\mu, \sigma^2/n)$, and hence $\sqrt{n}(\bar{X}-\mu)/\sigma \sim N(0,1)$. On the other hand, in the lectures we saw in the the multivariate normal theory section that $\bar{X}, S_{XX}$ are independent and:

$$\frac{1}{\sigma^2}S_{XX} \sim \chi^2_{n-1}$$

and hence it follows that:

$$T = \frac{\sqrt{n}(\bar{X}-\mu)/\sigma}{\left(\frac{1}{n-1}\frac{1}{\sigma^2}S_{XX}\right)^{n/2}}$$

is t-distributed, with $T \sim t_{n-1}$. It follows that the likelihood ratio test rejects H_0 in the region $\{|T| > c\}$, where the probability of a Type I error is:

$$\alpha = \mathbb{P}(|T|>c|H_0) = \mathbb{P}(T<-c|H_0) + \mathbb{P}(T>c|H_0) = 2\mathbb{P}(T>c|H_0),$$

using the fact that the t-distribution is symmetric. Hence we have:

$$1-\frac{\alpha}{2} = \mathbb{P}(T<c|H_0),$$

so we should choose c to be the upper $\alpha/2$ point of the Student t-distribution with $n-1$ degrees of freedom.

11. Statisticians A and B obtain independent samples $X_1, ..., X_{10}$ and $Y_1, ..., Y_{17}$ respectively, both from a $N(\mu, \sigma^2)$ distribution with both μ and σ unknown. They estimate (μ, σ^2) by $(\bar{X}, S_{XX}/9)$ and $(\bar{Y}, S_{YY}/16)$ respectively, where for example, $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$ and $S_{XX} = \sum_{i=1}^{10}(X_i - \bar{X})^2$. Given that the values $\bar{X} = 5.5$ and $\bar{Y} = 5.8$, which statistician's estimate of σ^2 is more probable to have exceeded the true value by more than 50%? Find this probability (approximately) in each case.

Solution: The key insight we need in this question is that the sample mean and sample variance, $\bar{X}$, S_{XX}, are independent of one another, as we proved in the lectures. We also have:

$$\frac{S_{XX}}{\sigma^2} \sim \chi^2_{n-1},$$

where n is the number of elements of a sample. It follows that:

$$\mathbb{P}\left(\frac{S_{XX}}{9} \geq 1.5\sigma^2\right) = \mathbb{P}\left(\frac{S_{XX}}{\sigma^2} \geq 13.5\right) = 1 - F^{(9)}(13.5) \approx 0.14,$$

where $F^{(n)}$ is the cumulative distribution function of a n degree of freedom χ^2-distribution. Similarly, we have:

$$\mathbb{P}\left(\frac{S_{YY}}{16} \geq 1.5\sigma^2\right) = \mathbb{P}\left(\frac{S_{YY}}{\sigma^2} \geq 24\right) = 1 - F^{(16)}(24) \approx 0.09.$$

Consequently, the first statistician's estimate is more likely to have exceeded the true value by more than 50%.

12. (**Harder**) In Question 5, does there exist a version of the Neyman-Pearson lemma when a likelihood ratio test of exact size α does not exist?

Solution: Yes, there does - if we allow for the use of *randomised* tests. Not too difficult - see lecturer's solutions.

Part IB: Statistics
Examples Sheet 3 Solutions

Please send all comments and corrections to jmm232@cam.ac.uk.

1. Let $X \sim N_n(\mu, \Sigma)$, and let A be an arbitrary $m \times n$ matrix. Prove directly from the definition that AX has an m-variate normal distribution $AX \sim N_m(A\mu, A\Sigma A^T)$.

Solution: Recall the definition of a multivariate normal distribution: we say that X has a multivariate normal distribution if for every $t \in \mathbb{R}^n$, we have $t^T X$ has a normal distribution.

In this case, we have $t^T(AX) = (A^T t)^T X$, which has a normal distribution since X is multivariate normal distributed. The ith component of its mean is given by:

$$\mathbb{E}[AX]_i = \sum_{j=1}^{n} \mathbb{E}[A_{ij}X_j] = \sum_{j=1}^{n} A_{ij}\mathbb{E}[X_j] = \sum_{j=1}^{n} A_{ij}\mu_j = (A\mu)_i,$$

so the mean is $A\mu$ as required. Similarly, the (i,j)th component of its covariance is given by:

$$\operatorname{cov}\left((AX)_i, (AX)_j\right) = \sum_{k,l=1}^{n} \operatorname{cov}\left(A_{ik}X_k, A_{jl}X_l\right) = \sum_{k,l=1}^{n} A_{ik}A_{jl}\operatorname{cov}\left(X_k, X_l\right) = \sum_{k,l=1}^{n} A_{ik}A_{jl}\Sigma_{kl} = (A\Sigma A^T)_{ij},$$

so the covariance is $A\Sigma A^T$ as required.

---

2. Let X and Y be $N(0,1)$ random variables such that $\operatorname{cov}(X,Y) = 0$. Show by example that X and Y need not be independent. (Hint: Take $Y = XZ$ where X and Z are independent and the distribution of Z is chosen appropriately.)

Solution: Assume we have some Z such that $Y = XZ$, and X, Z are independent. Then the covariance between X and Y is given by:

$$\operatorname{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X^2 Z] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdz\, x^2 z f_X(x) f_Z(z),$$

where $f_X(x)$ is the probability distribution function of X and $f_Z(z)$ is the probability distribution function of Z. We see that if $f_Z(z)$ is chosen to be symmetric about $z = 0$, then this integral vanishes since it is odd in z.

On the other hand, the probability distribution function of Y is given by the convolution:

$$f_Y(y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdz\, \delta(y - xz) f_X(x) f_Z(z) = \int_{-\infty}^{\infty} \frac{dz}{|z|} f_X\left(\frac{y}{z}\right) f_Z(z).$$

If we choose $f_Z(z) \propto \delta(z + 1/2) + \delta(z - 1/2)$, then this gives:

$$f_Y(y) \propto 2f_X(2y) + 2f_X(-2y) = 4f_X(2y),$$

since f_X is normally distributed with mean 0, so it is an even function. Thus f_Y is consistently normally distributed with this choice of Z.

Obviously, for $Y = XZ$ chosen in this way, Y is dependent on X up to the overall sign, so is not an independent normal distribution. This occurs because the relationship between X, Y is non-linear, whereas covariance is a measure of linear dependence between variables. However, if two variables are jointly normally distributed, the $\operatorname{cov}(X,Y) = 0$ *does* imply that they are independent.

3. Let $X \sim N_n(\mu, \sigma^2 I)$ where I is the $n \times n$ identity matrix, and let P be an $n \times n$ orthogonal projection matrix; i.e. $P^2 = P = P^T$. Show that the random vectors PX and $(I - P)X$ are independent.

Solution: Consider the vector:

$$Z = \begin{pmatrix} P \\ I - P \end{pmatrix} X,$$

which must be multivariate normal distributed by Question 1. Its covariance is given by:

$$\begin{pmatrix} P \\ I - P \end{pmatrix} \sigma^2 I \begin{pmatrix} P \\ I - P \end{pmatrix}^T = \sigma^2 \begin{pmatrix} P \\ I - P \end{pmatrix} \begin{pmatrix} P & I - P \end{pmatrix} = \sigma^2 \begin{pmatrix} P^2 & P - P^2 \\ P - P^2 & (I - P)^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} P & 0 \\ 0 & I - P \end{pmatrix},$$

since $(I - P)^2 = I - 2P + P^2 = I - 2P + P = I - P$, using the properties of a projection operator. It follows that the covariance of the variable Z is given by a block diagonal matrix; in particular, the first n components must be independent of the second n components, as required.

---

4. Let $X \sim N_n(\mu, \Sigma)$, and let X_1 denote the first n_1 components of X. Let μ_1 denote the first n_1 components of μ, and let Σ_{11} denote the upper left $n_1 \times n_1$ block of Σ. Show that $X_1 \sim N_{n_1}(\mu_1, \Sigma_{11})$.

Solution: Let:

$$A = \begin{pmatrix} I_{n_1} & 0 \end{pmatrix},$$

be the matrix which projects out the first n_1 components of X (so an $n_1 \times n_1$ identity matrix, with zeroes attached to pad to the size of X). Then $X_1 = AX$ is normally distributed, with mean $A\mu = \mu_1$ (the first n_1 components of μ), and covariance:

$$A\Sigma A^T = \begin{pmatrix} I_{n_1} & 0 \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I_{n_1} \\ 0 \end{pmatrix} = \begin{pmatrix} I_{n_1} & 0 \end{pmatrix} \begin{pmatrix} \Sigma_{11} \\ \Sigma_{12}^T \end{pmatrix} = \Sigma_{11},$$

as required.

5. Consider the simple linear regression model:

$$Y_i = a + bx_i + \epsilon_i, \qquad i = 1, ..., n,$$

where $\epsilon_1, ..., \epsilon_n \sim N(0, \sigma^2)$ are independent and identically distributed, and $\sum_{i=1}^{n} x_i = 0$. Derive from first principles explicit expressions for the MLEs $\hat{a}, \hat{b}$ and $\hat{\sigma}^2$. Show that we can obtain the same expressions if we regard the simple linear regression model as a special case of the general linear regression model $Y = X\beta + \epsilon$ and specialise the formulae $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $\hat{\sigma}^2 = n^{-1}||Y - X\hat{\beta}||^2$.

Solution: We have $Y_i \sim \mathcal{N}(a + bx_i, \sigma^2 I)$, so the likelihood of observing $(Y_1, ..., Y_n)$ to take the value $\mathbf{y} = (y_1, ..., y_n)$ is given by:

$$f(\mathbf{y}|a, b, \mathbf{x}) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2\right)$$

Taking the logarithm, we have:

$$\ell(a, b, \mathbf{x}) = \log(f(\mathbf{y}|a, b, \mathbf{x})) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2 .$$

Taking the derivative with respect to a, we have:

$$\frac{\partial \ell}{\partial a} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \qquad \Rightarrow \qquad a = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{Y},$$

is the MLE for a, i.e. $\hat{a} = \bar{Y}$. Taking the derivative with respect to b, we have:

$$\frac{\partial \ell}{\partial b} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0 \qquad \Rightarrow \qquad b = \frac{\mathbf{x}\cdot\mathbf{y}}{\mathbf{x}\cdot\mathbf{x}}$$

is the MLE for b, i.e. $\hat{b} = (\mathbf{x}\cdot\mathbf{Y})/\mathbf{x}\cdot\mathbf{x}$. Finally, taking the derivative with respect to σ^2 we have:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = 0 \qquad \Rightarrow \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{a} - \hat{b}x_i\right)^2 .$$

We can obtain the same expressions from the general model if we take $\beta = (a, b)^T$ and:

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

In this case, we have:

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \sum_{i=1}^{n} x_i^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{x}\cdot\mathbf{x} \end{pmatrix}.$$

We therefore have:

$$(X^TX)^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & (\mathbf{x}\cdot\mathbf{x})^{-1} \end{pmatrix}.$$

We also have:

$$X^TY = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix} = \begin{pmatrix} \bar{Y} \\ \mathbf{x}\cdot\mathbf{Y} \end{pmatrix}.$$

Hence we have:

$$\hat{\beta} = (X^TX)^{-1}X^TY = \begin{pmatrix} \bar{Y} \\ (\mathbf{x}\cdot\mathbf{Y})/(\mathbf{x}\cdot\mathbf{x}) \end{pmatrix},$$

as anticipated. The formula for $\hat{\sigma}^2$ is immediate.

---

6. The relationship between the range in metres, Y, of a howitzer with muzzle velocity v metres per second fired at angle of elevation α degrees is assumed to be:

$$Y = \frac{v^2}{g}\sin(2\alpha) + \epsilon,$$

where $g = 9.81$ and $\epsilon \sim N(0, \sigma^2)$. Estimate v from the following independent observations made on 9 shells.

α (deg)	5	10	15	20	25	30	35	40	45
$\sin(2\alpha)$	0.1736	0.3420	0.5	0.6428	0.7660	0.8660	0.9397	0.9848	1
range (m)	4860	9580	14080	18100	21550	24350	26400	27700	28300

Solution: This question can be done directly using the previous exercise. First, we need to normalise. We note that the model takes the form:

$$Y_i = \frac{v^2}{g}\sin(2\alpha_i) + \epsilon_i = \frac{v^2}{2ng}\sum_{i=1}^{n}\sin(2\alpha_i) + \frac{v^2}{g}\left(\sin(2\alpha_i) - \frac{1}{n}\sum_{i=1}^{n}\sin(2\alpha_i)\right) + \epsilon_i,$$

so we should take:

$$a = \frac{v^2}{2ng}\sum_{i=1}^{n}\sin(2\alpha_i), \qquad b = \frac{v^2}{g}, \qquad x_i = \sin(2\alpha_i) - \frac{1}{n}\sum_{i=1}^{n}\sin(2\alpha_i).$$

The MLE for a is given by:

$$\hat{a} = \frac{1}{9}\sum_{i=1}^{9} y_i = \frac{4860 + 9580 + 14080 + 18100 + 21550 + 24350 + 26400 + 27700 + 28300}{9} \approx 19435.56...$$

and the mean of the values $\sin(2\alpha_i)$ is given by:

$$\frac{1}{9}\sum_{i=1}^{9}\sin(2\alpha_i) = \frac{0.1738 + 0.3420 + 0.5 + 0.6428 + 0.7660 + 0.8660 + 0.9397 + 0.9849 + 1}{9} \approx 0.690578...$$

Hence, an estimate for v can be obtained as:

$$v \approx \sqrt{\frac{g\hat{a}}{\frac{1}{9}\sum_{i=1}^{9}\sin(2\alpha_i)}} \approx \sqrt{\frac{9.81 \cdot 19435.56}{0.690578}} \approx 525.4\,\text{ms}^{-1}.$$

7. Consider the one-way analysis of variance (ANOVA) model:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \qquad i = 1, ..., I, \qquad j = 1, ..., n_i,$$

where $(\epsilon_{ij}) \sim N(0, \sigma^2)$ are independent and identically distributed. Derive from first principles explicit expressions for the MLEs $\hat{\mu}_1, ..., \hat{\mu}_I$ and $\hat{\sigma}^2$. Show that we can obtain the same expressions if we regard the ANOVA model as a special case of the general linear regression model $Y = X\beta + \epsilon$ and specialise the formulae $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $\hat{\sigma}^2 = n^{-1}||Y - X\hat{\beta}||^2$.

Solution: The likelihood is given by:

$$f\left(Y_{ij}|\boldsymbol{\mu}, \sigma^2\right) = \frac{1}{\sqrt{(2\pi\sigma^2)^{n_1+...+n_I}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \mu_i\right)^2\right).$$

Taking the logarithm, we have:

$$\ell\left(\boldsymbol{\mu}, \sigma^2\right) = -\frac{n_1 + ... + n_I}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \mu_i\right)^2.$$

Taking the derivative, we have:

$$\frac{\partial \ell}{\partial \mu_k} = -\frac{1}{\sigma^2}\sum_{j=1}^{n_k}\left(Y_{kj} - \mu_k\right) = 0 \qquad \Rightarrow \qquad \hat{\mu}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{kj},$$

i.e. the MLEs are the column averages of the Y_k. The estimator for $\hat{\sigma}^2$ can be obtained from:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n_1 + ... + n_I}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{I}\sum_{k=1}^{n_i}\left(Y_{ij} - \mu_i\right) r \qquad \Rightarrow \qquad \hat{\sigma}^2 = \frac{1}{n_1 + ... + n_I}\sum_{i=1}^{I}\sum_{k=1}^{n_i}\left(Y_{ij} - \hat{\mu}_i\right)^2.$$

We already saw in lectures that this can be written in matrix form with:

$$Y = \begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ \vdots \\ Y_{1,n_1} \\ Y_{2,1} \\ Y_{2,2} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{I,1} \\ Y_{I,2} \\ \vdots \\ Y_{I,n_I} \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdot & 0 \\ \vdots & \vdots & \ddots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{pmatrix}.$$

where the X matrix is an $n \times I$ matrix, with the first 'block' being an $n_1 \times I$ matrix with ones on its first column, the second 'block' being an $n_2 \times I$ matrix with ones on its second column, etc.

Observe that in this form, we have:

$$X\beta = \begin{pmatrix} \mu_1 \\ \mu_1 \\ \vdots \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \vdots \\ \mu_2 \\ \vdots \\ \mu_I \\ \mu_I \\ \vdots \\ \mu_I \end{pmatrix},$$

so that the $\hat{\sigma}^2$ formula follows immediately. On the other hand, we have:

$$X^T X = \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_I \end{pmatrix}, \qquad X^T Y = \begin{pmatrix} \sum_{j=1}^{n_1} Y_{1j} \\ \sum_{j=1}^{n_2} Y_{2j} \\ \vdots \\ \sum_{j=1}^{n_I} Y_{Ij} \end{pmatrix}.$$

Hence computing $(X^T X)^{-1} X^T Y$, we get the formulae for $\hat{\mu}$ that we obtained directly.

8. Consider the linear model $Y = X\beta + \epsilon$, where $\mathbb{E}\epsilon = 0$ and $\mathrm{Cov}(\epsilon) = \sigma^2\Sigma$, for some unknown parameter σ^2 and known positive definite matrix Σ. Derive the form of the Generalised Least Squares estimator $\tilde{\beta}^{\mathsf{GLS}}$, defined by:

$$\tilde{\beta}^{\mathsf{GLS}} = \arg\min_{\beta}(Y - X\beta)^T\Sigma^{-1}(Y - X\beta).$$

State and prove a version of the Gauss-Markov theorem for $\tilde{\beta}^{\mathsf{GLS}}$.

Solution: In index notation, we have:

$$(Y - X\beta)^T\Sigma^{-1}(Y - X\beta) = (Y_i - X_{ik}\beta_k)\Sigma^{-1}_{ij}(Y_j - X_{jl}\beta_l).$$

Taking a derivative with respect to β_a, we have at stationary values:

$$0 = -X_{ia}\Sigma^{-1}_{ij}(Y_j - X_{jl}\beta_l) - (Y_i - X_{ik}\beta_k)\Sigma^{-1}_{ij}X_{ja} = -2X_{ia}\Sigma^{-1}(Y_j - X_{jl}\beta_l),$$

by symmetry of the covariance matrix Σ. Restoring the vector notation, we have:

$$0 = X^T\Sigma^{-1}(Y - X\beta) \qquad \Rightarrow \qquad X^T\Sigma^{-1}X\beta = X^T\Sigma^{-1}Y.$$

Since Σ is positive definite, we have that $X^T\Sigma^{-1}X$ is positive definite (assuming X has full rank), which implies it is invertible. Thus we have:

$$\beta = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y.$$

The version of the Gauss-Markov theorem that we shall state and prove is the following 'correlated' version: if β^* is any other linear unbiased estimator of β (where by linear, we mean linear in the data Y), we have:

$$t^T\mathrm{Cov}\left(\tilde{\beta}^{\mathsf{GLS}}\right)t \leq t^T\mathrm{Cov}(\beta^*)t$$

for all vectors t.

Proof: We first observe that the covariance of the generalised least squares estimator is given by:

$$\begin{aligned}\mathrm{Cov}\left(\tilde{\beta}^{\mathsf{GLS}}\right) &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\left(\sigma^2\Sigma\right)\left[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\right]^T \\ &= \sigma^2(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\Sigma\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} \\ &= \sigma^2(X^T\Sigma^{-1}X)^{-1},\end{aligned}$$

where we have used the result on multiplying a multivariate normal vector Y by a matrix.
Now let $\beta^* = AY$ be any linear unbiased estimator of β. Since it is unbiased, $AX\beta = \mathbb{E}[\beta^*] = \beta$. This holds for any β, so we must have $AX = I$. Now,

$$\begin{aligned}\mathrm{Cov}(\beta^*) &= \mathbb{E}[(\beta^* - \beta)(\beta^* - \beta)^T] \\ &= \mathbb{E}[(AY - \beta)(AY - \beta)^T] \\ &= \mathbb{E}[(AX\beta + A\epsilon - \beta)(AX\beta + A\epsilon - \beta)^T] \\ &= \mathbb{E}[A\epsilon\epsilon^TA^T] \\ &= A\mathbb{E}[\epsilon\epsilon^T]A^T \\ &= \sigma^2A\Sigma A^T.\end{aligned}$$

Now, we let $B = A - (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$, the difference between A and the least-squares linear mapping. Then:

$$BX = AX - (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X = AX - I = 0.$$

Therefore, we have in terms of B:

$$\begin{aligned}\mathrm{Cov}(\beta*) &= \sigma^2\left(B + (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\right)\Sigma\left(B + (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\right)^T \\ &= \sigma^2\left(B\Sigma B^T + (X^T\Sigma X)^{-1}\right) \\ &= \sigma^2 B\Sigma B^T + \mathrm{Cov}\left(\tilde{\beta}^{\mathsf{GLS}}\right)\end{aligned}$$

Now, since $t^T B\Sigma B^T t \geq 0$ (since Σ is positive definite), we have:

$$t^T\mathrm{Cov}(\beta^*)t \geq t^T\mathrm{Cov}\left(\tilde{\beta}^{\mathsf{GLS}}\right)t,$$

as required.

Compare with the usual Gauss-Markov theorem: in a full rank linear model $Y = X\beta + \epsilon$, with $\mathrm{Cov}(\epsilon) = \sigma^2 I$, then for any unbiased linear estimator β^* of β (where linear means linear in the data, Y), we have:

$$\mathrm{Var}(t^T\hat{\beta}) \leq \mathrm{Var}(t^T\beta^*)$$

where $\hat{\beta}$ is the least-squares estimator, and t is any vector.