Homework 2

Aiden Kenny
STAT GR5204: Statistical Inference
Columbia University

Novermber 26, 2020

Question 1

Suppose that $Y \sim \text{Bin}(100, p)$, and we want to make an inference about the value of p. We test $H_0 : p = 0.08$ against $H_A : p < 0.08$, and our test δ will reject H_0 if and only if $Y = 6$.

(a) The significance α is the probability of making a Type I error,

$$\alpha(\delta) = \Pr(Y = 6 \,|\, p = 0.08) = \binom{100}{6}(0.08)^6(0.92)^{94} = 0.123.$$

(b) Suppose that $p = 0.04$. The probability of a Type II error, β, is

$$\beta(\delta) = \Pr(Y \neq 6 \,|\, p = 0.04) = 1 - \Pr(Y = 6 \,|\, p = 0.04) = 1 - \binom{100}{6}(0.04)^6(0.96)^{94} = 0.895.$$

Question 2

For a random variable $Y \sim \text{Binom}(n, p)$, if n is large enough we can approximate it using a normal distribution with the same mean and variance, i.e. $Y \sim \text{N}(np, np(1-p))$. Then the sample proportion, $\hat{p} = Y/n$, is also normally distributed as $\hat{p} \sim \text{N}(p, p(1-p)/n)$, and standardizing $\hat{p}$ gives us $(\hat{p} - p)/\sqrt{p(1-p)/n} \sim \text{N}(0, 1)$. It can then be shown that a $100\gamma\%$ confidence interval for $\hat{p}$ is given by

$$\mathcal{I} = \left(\frac{\hat{p} + z_\gamma^2/2n}{1 + z_\gamma^2/2n} - z_\gamma \cdot \frac{\sqrt{\hat{p}(1-\hat{p})/n + z_\gamma^2/4n^2}}{1 + z_\gamma^2/n} \;,\; \frac{\hat{p} + z_\gamma^2/2n}{1 + z_\gamma^2/2n} + z_\gamma \cdot \frac{\sqrt{\hat{p}(1-\hat{p})/n + z_\gamma^2/4n^2}}{1 + z_\gamma^2/n} \right),$$

where $z_\gamma = \Phi^{-1}((1+\gamma)/2)$. For this example, we have $n = 300$ and $\hat{p} = 75/300 = 1/4$, and to get a 90% confidence interval, we have $z_\gamma = 1.645$. Therefore, a 90% confidence interval for p is $(0.212, 0.294)$.

Question 3

Suppose we have a random sample $\boldsymbol{X} \overset{\text{iid}}{\sim} \text{Gamma}(4, \beta)$, so $\mathbb{E}[X_i] = 4\beta$ and $\text{Var}[X_i] = 4\beta^2$. The expected value and variance of the sample mean $\bar{X}$ is then given by $\mathbb{E}[\bar{X}] = 4\beta$ and $\text{Var}[\bar{X}] = 4\beta^2/n$, and from the CLT we have $\sqrt{n}(\bar{X} - 4\beta)/2\beta = \sqrt{n}(\bar{X}/2\beta - 2) \sim \text{N}(0, 1)$. If $z_\gamma = \Phi^{-1}((1+\gamma)/2)$, then we have $\gamma = \Pr(-z_\gamma \leq \sqrt{n}(\bar{X}/2\beta - 2) \leq z_\gamma)$. Rearranging to get β in the middle gives us

$$\mathcal{I} = \left(\frac{2\bar{X}}{2 + z_\gamma/\sqrt{n}} \;,\; \frac{2\bar{X}}{2 - z_\gamma/\sqrt{n}} \right).$$

For this random sample, $n = 25$, and because we want a 95.4% confidence interval (oddly specific), we have $z_\gamma = 2$, so the confidence interval is given by $\mathcal{I} = \left(\, 5\bar{X}/6 \;,\; 5\bar{X}/4 \, \right)$.

Question 4

Suppose that $X \sim \text{Binom}(100, p)$, where $p \in (1/4, 1/2)$ is unknown. We test $H_0 : 1/2$ against $H_A : p = 1/4$ using δ : reject H_0 if $X \leq 3$. That is, our rejection region is $\mathcal{S}_X = \{0, 1, 2, 3\}$, and so the power function for this test is

$$\pi(p \,|\, \delta) = \Pr(X \in \mathcal{S}_X \,|\, p) = \sum_{k=0}^{3} \binom{10}{k} p^k (1-p)^{n-k}.$$

Question 5

Suppose $X_1, X_2 \overset{\text{iid}}{\sim} \text{Exp}(1/\theta)$, where $\theta > 0$, which means $\mathbb{E}[X_i] = \theta$ and $\text{Var}[X_i] = \theta^2$. The joint density is

$$f_{\boldsymbol{X}}(\mathbf{x}) = \frac{\mathrm{e}^{-x_1/\theta}}{\theta} \cdot \frac{\mathrm{e}^{-x_2/\theta}}{\theta} = \frac{\mathrm{e}^{-(x_1+x_2)/\theta}}{\theta^2}$$

Consider the transformation $Y_1 = X_1 + X_2$ and $Y_2 = X_2$, which implies $Y_1 \sim \text{Gamma}(2, 1/\theta)$ and $Y_2 \sim \text{Exp}(1/\theta)$. We note that because $f_{\boldsymbol{X}}(\mathbf{x})$, which is also the likelihood function of the random sample, is a function of y_1 and θ, so it is a sufficinet statistic. In vector notation, the transformation is given by $\boldsymbol{Y} = \mathbf{A}\boldsymbol{X}$, where $\mathbf{A} = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$, and so $\boldsymbol{X} = \mathbf{A}^{-1}\boldsymbol{Y}$, where $\mathbf{A}^{-1} = \left[\begin{smallmatrix} 1 & -1 \\ 0 & 1 \end{smallmatrix}\right]$. Therefore, the joint density of Y_1 and Y_2 is

$$g_{\boldsymbol{Y}}(\mathbf{y}) = f_{\boldsymbol{X}}(\mathbf{x}) \cdot \left| \det\left(\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}}\right) \right| = f_{\boldsymbol{X}}\left(\mathbf{A}^{-1}\mathbf{y}\right) \cdot |\det(\mathbf{A})| = \frac{\mathrm{e}^{-(y_1-y_2+y_2)/\theta}}{\theta^2} \cdot 1 = \frac{\mathrm{e}^{-y_1/\theta}}{\theta^2},$$

for $0 < y_2 < y_1$. Because $Y_2 \sim \text{Exp}(1/\theta)$, we have $\mathbb{E}[Y_2] = \theta$, and so Y_2 is an unbiased estimator for θ. The conditional density of Y_2 given Y_1 is

$$g_{Y_2|Y_1}(y_2 \,|\, y_1) = \frac{g_{\boldsymbol{Y}(\mathbf{y})}}{g_{Y_1}(y_1)} = \frac{\mathrm{e}^{-y_1/\theta}/\theta^2}{y_1 \mathrm{e}^{-y_1/\theta}} = \frac{1}{y_1},$$

and so the conditional expectation of Y_2 given Y_1 is

$$\mathbb{E}[Y_2 \,|\, y_1] = \int y_2 \cdot g_{Y_2|Y_1}(y_2 \,|\, y_1) \, \partial y_2 = \int_0^{y_1} \frac{y_2}{y_1} \, \partial y_2 = \left. \frac{y_2^2}{2y_1} \right|_0^{y_1} = \frac{y_1}{2} := \varphi(y_1).$$

As is standard, we can define the random variable $\varphi(Y_1)$, and immediately see that $\mathbb{E}[\varphi(Y_1)] = \mathbb{E}[Y_2] = \theta$. In addition, we have

$$\mathbb{E}[\varphi(Y_1)^2] = \int \varphi(y_1)^2 \cdot g_{Y_1}(y_1) \, \partial y_1 = \int_0^\infty \frac{y_1^2}{4} \cdot \frac{y_1 \mathrm{e}^{-y_1/\theta}}{\theta^2} = \frac{3\theta^2}{2} \int_0^\infty \frac{y_1^{4-1} \cdot \mathrm{e}^{-y_1/\theta}}{\theta^4 \Gamma(4)} \, \partial y_1 = \frac{3\theta^2}{2},$$

and so $\text{Var}[\varphi(Y_1)] = \mathbb{E}[\varphi(Y_1)^2] - \mathbb{E}[\varphi(Y_1)]^2 = \theta^2/2$.

Question 6

Suppose that $\boldsymbol{X} \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$. We test $H_0 : \theta = 1/2$ against $H_A : \theta < 1/2$ using our procedure δ : reject H_0 if the test statistic $Y = \mathbf{1}^T \boldsymbol{X} = \sum X_i \leq 2$. Since it is a sum of independent Poisson random variables, our test statistic is also a Poisson random variable, namely $Y \sim \text{Poisson}(n\theta)$, and the rejection region for Y is $\mathcal{S}_Y = \{0, 1, 2\}$. Therefore, for $\theta \in (0, 1/2]$, the power function of δ is given by

$$\pi(\theta \,|\, \delta) = \Pr(Y \in \mathcal{S}_Y \,|\, \theta) = \sum_{k=0}^{2} \frac{(n\theta)^k \mathrm{e}^{-n\theta}}{k!}.$$

A graph of $\pi(\theta \,|\, \delta)$ can be found in Figure 1. The values of $\pi(\theta \,|\, \delta)$ when $\theta = \{1/2, 1/3, 1/4, 1/6, 1/12\}$ are given by $0.062, 0.238, 0.423, 0.677, 0.920$, respectively, and have also been marked on Figure 1. The significance of this test is given by $\alpha(\delta) = \pi(1/2 \,|\, \delta) = 0.062$.

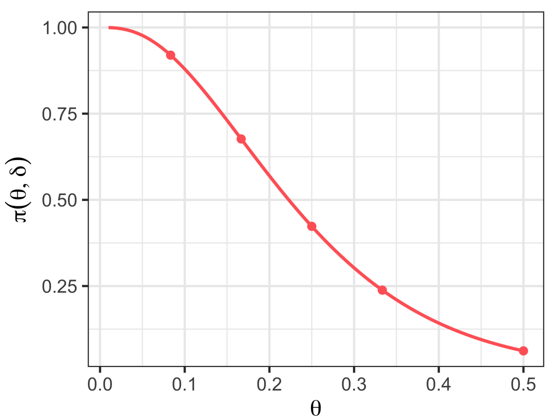

Figure 1: The power function $\pi(\theta \mid \delta)$.

Question 7

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$, where θ is unknown, and let $Y_1, \ldots, Y_n$ be the n order statistics of the random sample. We are testing $H_0 : \theta = 1$ against $H_A : \theta > 1$ with procedure δ : reject H_0 if $Y_n \geq c$, where c is an unknown constant that will be determined by the significance of δ. If Y_n is the nth order statistic of $\boldsymbol{X}$, its cdf and density are respectively given by $F_{Y_n}(y_n) = (y_n/\theta)^n$ and $f_{Y_n}(y_n) = n y_n^{n-1}/\theta^n$ for $y_n \in [0, \theta]$.

(b) The rejection region of Y_n is $\mathcal{S}_{Y_n} = [c, \theta]$, and so the power function is given by

$$\pi(\theta \mid \delta) = \Pr(Y_n \in \mathcal{S}_{Y_n} \mid \theta) = F_{Y_n}(\theta) - F_{Y_n}(c) = 1 - \frac{c^n}{\theta^n}.$$

(a) To have a significance level of $\alpha = 0.05$, we must have $\alpha(\delta) = \pi(1 \mid \delta) = 1 - c^n \overset{\text{set}}{=} 0.05$, and solving for c gives us $c = \sqrt[n]{0.95}$, and the power function with this level of significance is $\pi(\theta \mid \delta) = 1 - 0.95/\theta^n$. The question specifically asks for $n = 4$, in which case $c = \sqrt[4]{0.95}$ and $\pi(\theta \mid \delta) = 1 - 0.95/\theta^4$. More generally, for any specified significance level α_*, we will have $c = \sqrt[n]{1 - \alpha_*}$ and $\pi(\theta \mid \delta) = 1 - (1 - \alpha_*)/\theta^n$.

Question 8

Let $\boldsymbol{X} \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. We want to test $H_0 : \sigma^2 = \sigma_0^2$ against $H_A : \sigma^2 > \sigma_0^2$ using procedure δ : reject H_0 if $nS^2/\sigma_0^2 \geq c$, where c will be determined by the significance of the test. Our test statistic is based off of $S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$, and we know that $T(\boldsymbol{X}) \coloneqq nS^2/\sigma_0^2 \sim \chi^2(n-1)$. The rejection region for our test statistic is $\mathcal{S}_{T(\boldsymbol{X})} = [c, \infty)$, and we would like to find a value of c such that the significance is $\alpha = 0.025$.

$$\alpha(\delta) = \pi(\sigma_0^2 \mid \delta) = \Pr\big(T(\boldsymbol{X}) \in \mathcal{S}_{T(\boldsymbol{X})} \mid \sigma_0^2\big) = \int_c^\infty \frac{t^{(n-3)/2} \cdot \mathrm{e}^{-t/2}}{2^{(n-1)/2}\Gamma\left((n-1)/2\right)} \, \partial t \overset{\text{set}}{=} 0.025.$$

There is no way to derive a closed form relationship between c and α like we did in question 7, but using `R` we can write a function to find the value of c for a given significance value α. When $n = 13$ and $\alpha = 0.025$, we have $c = 23.337$. To find a general formula of $\pi(\sigma^2 \mid \delta)$, we notice that if $T(\boldsymbol{X}) \geq c$, then $S^2 \geq c\sigma_0^2/n$, which serives as a rejection region for S^2. Since $nS^2/\sigma^2 \sim \chi^2(n-1)$ for any value of σ^2, it can be shown that $S^2 \sim \text{Gamma}\big((n-1)/2, n/2\sigma^2\big)$. Therefore, the power function is given by

$$\pi(\sigma^2 \mid \delta) = \Pr(S^2 \geq c\sigma_0^2/n \mid \sigma^2) = \int_{c\sigma_0^2/n}^\infty \frac{x^{(n-3)/2} \cdot \mathrm{e}^{-nx/2\sigma^2}}{\big(n/2\sigma^2\big)^{(n-1)/2}\Gamma\big((n-1)/2\big)} \, \partial x,$$

where c is chosen to make $\alpha(\delta) = 0.025$ (or any specified value, for that matter). Again, there is no closed form solution to this function, but using `R` we can find its values, and given a specified value of σ_0^2, we will always have $\pi(\sigma_0^2 \mid \delta) = \alpha$ as desired.

Question 9

If $\boldsymbol{X} \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ with unknown μ and σ^2, then a γ% confidence interval for μ is given by

$$\mathcal{I} = \Big(\ \bar{X} - t_\gamma(n) \cdot S/\sqrt{n}\ ,\ \bar{X} + t_\gamma(n) \cdot S/\sqrt{n}\ \Big),$$

where $t_\gamma(n) = T_{n-1}^{-1}((1+\gamma)/2)$ is the $(1+\gamma)/2$th quantile of the t distribution with df $= n-1$ and S is the sample standard deviation. The length of this confidence interval is given by

$$\Delta = \max(\mathcal{I}) - \min(\mathcal{I}) = \Big(\bar{X} + t_\gamma(n) \cdot S/\sqrt{n}\Big) - \Big(\bar{X} - t_\gamma(n) \cdot S/\sqrt{n}\Big) = 2t_\gamma(n) \cdot S/\sqrt{n}.$$

The squared length is then given by $\Delta^2 = 4t_\gamma^2(n) \cdot S^2/n$. Because the sample variance is an unbiased estimator for σ^2, we have $\mathbb{E}[\Delta^2] = \mathbb{E}\big[4t_\gamma^2(n) \cdot S^2/n\big] = 4t_\gamma^2(n) \cdot \sigma^2/n$. We now set $\mathbb{E}[\Delta^2] < \sigma^2/2$, and after some cancellations, we see that we need $t_\gamma^2(n)/n < 1/8$. There is no way to find a closed-form expression for this, so we will have to check the value of $t_\gamma^2(n)/n$ for increasing values of n. I set up a `while` loop in `R` to solve for it, and when $\gamma = 0.9$, we find that $n = 24$ is the smallest value of n such that $\mathbb{E}[\Delta^2] < \sigma^2/2$.

Question 10

Let $\boldsymbol{X} \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \sigma^2)$, where θ is unknown and σ^2 is known, and we assume prior that $\theta \sim \mathrm{N}(\mu, \nu^2)$, where both μ and ν^2 are known.

(a) Since normal distributions are are conjugate to normal sampling, it follows that $\theta \mid \mathbf{x} \sim \mathrm{N}(\tilde{\mu}, \tilde{\sigma}^2)$, where

$$\tilde{\mu} = \frac{\sigma^2\mu + n\nu^2\bar{x}}{\sigma^2 n\nu^2} \quad \text{and} \quad \tilde{\sigma}^2 = \frac{\sigma^2\nu^2}{\sigma^2 + n\mu^2}.$$

We also know that $(\theta \mid \mathbf{x} - \tilde{\mu})/\tilde{\sigma} \sim \mathrm{N}(0,1)$, and so a 95% confidence interval for $\theta \mid \mathbf{x}$ is given by

$$\mathcal{I} = \Big(\ \tilde{\mu} - \Phi^{-1}(0.975) \cdot \tilde{\sigma}\ ,\ \tilde{\mu} + \Phi^{-1}(0.975) \cdot \tilde{\sigma}\ \Big).$$

(b) We can think of our interval $\mathcal{I}$ as a function of ν^2. To examine what happens to $\mathcal{I}(\nu^2)$ as $\nu^2 \to \infty$, we will first look at $\tilde{\mu}$ and $\tilde{\sigma}$. Using L'Hopital's rule, we have

$$\lim_{\nu^2 \to \infty} \tilde{\mu} = \lim_{\nu^2 \to \infty} \frac{\sigma^2\mu + n\nu^2\bar{x}}{\sigma^2 n\nu^2} = \lim_{\nu^2 \to \infty} \frac{n\bar{x}}{n} = \bar{x},$$
$$\lim_{\nu^2 \to \infty} \tilde{\sigma} = \lim_{\nu^2 \to \infty} \sqrt{\frac{\sigma^2\nu^2}{\sigma^2 + n\mu^2}} = \sqrt{\lim_{\nu^2 \to \infty} \frac{\sigma^2\nu^2}{\sigma^2 + n\mu^2}} = \sqrt{\lim_{\nu^2 \to \infty} \frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}},$$

and so $\mathcal{I}(\nu^2) \to \big(\ \bar{x} - \Phi^{-1}(0.975) \cdot \sigma/\sqrt{n}\ ,\ \bar{x} + \Phi^{-1}(0.975) \cdot \sigma/\sqrt{n}\ \big)$, which is a 95% confidence interval for θ.

Question 11

Let $\boldsymbol{X} \overset{\text{iid}}{\sim} \mathrm{Unif}(0, \theta)$, where θ is unknown, and let $Y = X_{(n)}$ be the nth order statistic.

(a) Let $F_i(x) = \Pr(X_i \le x) = x/\theta$ for all $i \in \{1, \dots, n\}$. Then the cdf of Y is $G(y) = \prod_{i=1}^n F_i(y) = (y/\theta)^n$, since all of the X_i's are independent. The density of Y is then given by $g(y) = ny^{n-1}/\theta^n$ for $y \in [0, \theta]$. By letting $W = Y/\theta$, we have $Y = \theta W$ and $\partial Y/\partial W = \theta$, so the density of Y/θ is given by $h(w) = g\big(y(w)\big) \cdot \theta = nw^{n-1}$ for $w \in [0,1]$. The cdf is then seen to be $H(w) = w^n$, and so the quantile is given by $H^{-1}(w) = \sqrt[n]{w}$.

(b) Since we must have $y \leq \theta$, it is natural that Y will underestimate θ, and is therefore biased. We have

$$\mathbb{E}[Y] = \int_0^\theta y \cdot \frac{ny^{n-1}}{\theta^n}\, \partial y = \frac{n}{\theta^n}\int_0^\theta y^n\, \partial y = \frac{n}{\theta^n} \cdot \frac{y^{n+1}}{n+1}\bigg|_0^\theta = \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \frac{n\theta}{n+1},$$

and so the bias is $\text{bias}(Y) = \mathbb{E}[Y] - \theta = -\theta/(n+1)$.

(d) Using the cdf of Y/θ, for any interval involving Y/θ, we have $\Pr(a \leq Y/\theta \leq b) = b^n - a^n$, where $a, b \in (0, 1]$. Rearranging the terms inside the interval gives us $\Pr(Y/b \leq \theta \leq Y/a) = b^2 - a^2 \overset{\text{set}}{=} \gamma$. That is, as long as we impose the constraint that $b^2 - a^2 = \gamma$, any interval $(Y/b, Y/a)$ is a $\gamma\%$ confidence interval for θ.

Question 12

Suppose that $X \sim \mathrm{P}$, where P is an unknown distribution, and we want to test $H_0 : \mathrm{P} = \text{Unif}(0, 1)$ against $H_1 : \mathrm{P} = \mathrm{N}(0, 1)$. The densities of both distributions (within their support) are given by $f_0(x) = 1$ for $x \in [0, 1]$ and $f_2(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, respectively, and so our test statistics is