

Homework 2

Aiden Kenny

STAT GR5204: Statistical Inference

Columbia University

November 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,

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

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

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

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. $Y \sim N(np, np(1-p))$. Then the sample proportion, $\hat{p} = Y/n$, is also normally distributed as $\hat{p} \sim N(p, p(1-p)/n)$, and standardizing \hat{p} gives us $(\hat{p} - p)/\sqrt{p(1-p)/n} \sim N(0, 1)$. It can then be shown that a 100% 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 $z_{0.90} = 1.645$. Therefore, a 90% confidence interval for p is $(0.212, 0.294)$.

Question 9

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

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

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

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

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 $\mathbb{E}[\Delta^2] = \mathbb{E}[4t_\gamma^2(n) \cdot S^2/n] = 4t_\gamma^2(n) \cdot \sigma^2/n$. 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 $\mathbf{X} \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)$, where θ is unknown and σ^2 is known, and we assume prior that $\theta \sim N(\mu, \nu^2)$, where both μ and ν^2 are known.

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

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

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

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

- (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 $\nu^2 \rightarrow \infty$, we will first look at $\tilde{\mu}$ and $\tilde{\sigma}$. Using L'Hopital's rule, we have

$$\begin{aligned} \lim_{\nu^2 \rightarrow \infty} \tilde{\mu} &= \lim_{\nu^2 \rightarrow \infty} \frac{\sigma^2 \mu + n \nu^2 \bar{x}}{\sigma^2 + n \nu^2} = \lim_{\nu^2 \rightarrow \infty} \frac{n \bar{x}}{n} = \bar{x}, \\ \lim_{\nu^2 \rightarrow \infty} \tilde{\sigma} &= \lim_{\nu^2 \rightarrow \infty} \sqrt{\frac{\sigma^2 \nu^2}{\sigma^2 + n \nu^2}} = \sqrt{\lim_{\nu^2 \rightarrow \infty} \frac{\sigma^2 \nu^2}{\sigma^2 + n \nu^2}} = \sqrt{\lim_{\nu^2 \rightarrow \infty} \frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}, \end{aligned}$$

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

Question 11

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

- (a) Let $F_i(x) = \Pr(X_i \leq x) = x/\theta$ 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 $h(w) = g(y(w)) \cdot \theta = nw^{n-1}$ 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

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

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 $\Pr(Y/b \leq \theta \leq Y/a) = b^2 - a^2 \stackrel{\text{set}}{=} \gamma$. 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 $X \sim P$, where P is an unknown distribution, and we want to test $H_0 : P = \text{Unif}(0, 1)$ against $H_1 : P = N(0, 1)$. 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