

Homework 3

Aiden Kenny

STAT GR5204: Statistical Inference

Columbia University

December 14, 2020

Question 1

Suppose that $\mathbf{X} \stackrel{\text{iid}}{\sim} \chi^2(\theta)$, where $\theta \in \mathbb{N}$ is unknown. We would like to test $H_0 : \theta \leq 8$ against $H_A : \theta > 8$, using a UMP test δ^* with a specified significance $\alpha_* \in (0, 1)$. The joint density of \mathbf{X} is given by

$$f(\mathbf{x} | \theta) = \prod_{i=1}^n \frac{x_i^{\theta/2-1} e^{-x_i/2}}{2^{\theta/2} \Gamma(\theta/2)} = 2^{-n\theta/2} \cdot \Gamma^{-n}(\theta/2) \cdot \left(\prod_{i=1}^n x_i \right)^{n(\theta/2-1)} \cdot \exp \left(-\frac{1}{2} \sum_{i=1}^n x_i \right)$$

To determine δ^* , we will look at the likelihood ratio. If we have two values θ_1, θ_2 such that $\theta_1 < \theta_2$, then then likelihood ratio is

$$\begin{aligned} \frac{f(\mathbf{x} | \theta_2)}{f(\mathbf{x} | \theta_1)} &= \frac{2^{-n\theta_2/2} \cdot \Gamma^{-n}(\theta_2/2) \cdot (\prod_i x_i)^{n(\theta_2/2-1)} \cdot \exp(-\frac{1}{2} \sum_i x_i)}{2^{-n\theta_1/2} \cdot \Gamma^{-n}(\theta_1/2) \cdot (\prod_i x_i)^{n(\theta_1/2-1)} \cdot \exp(-\frac{1}{2} \sum_i x_i)} \\ &= 2^{n(\theta_1-\theta_2)/2} \left(\frac{\Gamma(\theta_1/2)}{\Gamma(\theta_2/2)} \right)^n \left(\prod_{i=1}^n x_i \right)^{n(\theta_2-\theta_1)/2}, \end{aligned}$$

which is a monotone increasing function of the test statistic $\prod_{i=1}^n x_i$. Therefore, the UMP test is δ^* : reject H_0 if $\prod_{i=1}^n X_i \geq c_*$, where c_* is chosen such that the test has a significance level α_* . Taking the log of both sides gives us $\sum_{i=1}^n \log X_i \geq \log c_* := k$.

Question 2

Question 3

From a random sample of n people, we observe which of the k different cereal brands that each person prefers. Let N_i be the number of people who prefer the i th cereal (so $\sum_i N_i = n$), and let p_i be the proportion of people who prefer cereal i . With a significance of α , we would like to test $H_0 : p_1 = \dots = p_k = 1/k$ against $H_A : H_0$ is false. To conduct this test we will use the χ^2 goodness-of-fit test, from which our test statistic is

$$Q = \sum_{i=1}^k \frac{(N_i - n/k)^2}{n/k} = \frac{k}{n} \sum_{i=1}^k \left(N_i^2 - \frac{2nN_i}{k} + \frac{n^2}{k^2} \right) = \frac{k}{n} \sum_{i=1}^k N_i^2 - n.$$

For this test, we reject H_0 when $Q \geq q_\alpha$, where q_α is the $1 - \alpha$ quantile of the χ^2 distribution with $k - 1$ degrees of freedom. As a result, we will reject H_0 when $\sum_i N_i^2 \geq n(q_\alpha + n)/k$. From the specific example we have $n = 400$, $k = 5$, and $q_{0.01} = 13.273$, so in order to reject H_0 , we would need $\sum_i N_i^2 \geq 33062.136$.

Question 8

By definition, the least-squares estimate for the intercept is $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Rearranging gives us $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, and so the least-squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ will always pass through the point (\bar{x}, \bar{y}) .

Question 9

- (a) The least-squares coefficients for the model are given by $\hat{\beta}_0 = 40.893$ and $\hat{\beta}_1 = 0.548$.

Question 10

Let (\mathbf{x}, \mathbf{y}) be the vectors of observations for the predictor x and the response Y . From the data, we have $n = 10$, $\bar{x} = 2.33$, $\bar{y} = 0.81$, $\|\mathbf{x} - \bar{x}\mathbf{1}\|^2 = 36.081$, and $(\mathbf{y} - \bar{y}\mathbf{1})^T(\mathbf{x} - \bar{x}\mathbf{1}) = 24.717$. Here we are assuming that $Y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.

- (a) The MLEs for β_0 , β_1 , and σ^2 are given by

$$\hat{\beta}_1 = \frac{(\mathbf{y} - \bar{y}\mathbf{1})^T(\mathbf{x} - \bar{x}\mathbf{1})}{\|\mathbf{x} - \bar{x}\mathbf{1}\|^2} = 0.685, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -0.786, \quad \hat{\sigma}^2 = \frac{\|\mathbf{y} - \hat{\beta}_0\mathbf{1} - \hat{\beta}_1\mathbf{x}\|^2}{n} = 0.938.$$

- (b) The variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ is given by

$$\text{Var}[\hat{\beta}_1] = \frac{\sigma^2}{\|\mathbf{x} - \bar{x}\mathbf{1}\|^2} = 0.0277\sigma^2, \quad \text{Var}[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\|\mathbf{x} - \bar{x}\mathbf{1}\|^2} \right) = 0.25\sigma^2.$$

- (c) The covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$, and therefore the correlation, is

$$\text{Cov}[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x}\sigma^2}{\|\mathbf{x} - \bar{x}\mathbf{1}\|^2} = -0.0646\sigma^2, \quad \text{Cor}[\hat{\beta}_0, \hat{\beta}_1] = \frac{\text{Cov}[\hat{\beta}_0, \hat{\beta}_1]}{\sqrt{\text{Var}[\hat{\beta}_0] \cdot \text{Var}[\hat{\beta}_1]}} = -0.775.$$

Question 11

Suppose β_0, β_1 are the coefficients from the linear model in question 10, and we want to estimate $\theta = 3\beta_0 - 2\beta_1 + 5$. Because $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators for the coefficients, we can estimate θ with $\hat{\theta} = 3\hat{\beta}_0 - 2\hat{\beta}_1 + 5$. The MSE of $\hat{\theta}$, which is just its variance, is given by

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Var}[3\hat{\beta}_0 - 2\hat{\beta}_1 + 5] = 9\text{Var}[\hat{\beta}_0] + 4\text{Var}[\hat{\beta}_1] - 12\text{Cov}[\hat{\beta}_0, \hat{\beta}_1] = 2.0245\sigma^2.$$

Question 12

We know that the MLE and least-squares estimates of β_0 and β_1 are the same, so $\hat{\beta}_0 = -0.786$ and $\hat{\beta}_1 = 0.685$. Using this linear model, when $x = 2$, we predict $\hat{Y} = -0.786 + 0.685 \cdot 2 = 0.584$. The variance of \hat{Y} is given by

$$\mathbb{E}[(\hat{Y} - Y)^2] = \sigma^2 \left(1 + \frac{1}{n} + \frac{(2 - \bar{x})^2}{\|\mathbf{x} - \bar{x}\mathbf{1}\|^2} \right) = 1.103\sigma^2.$$

Question 13