Homework 3

Aiden Kenny
STAT GR5204: Statistical Inference
Columbia University

December 14, 2020

Question 1

Suppose that $\boldsymbol{X} \overset{\text{iid}}{\sim} \chi^2(\theta)$, 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 $\boldsymbol{X}$ is given by

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

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

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

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 $\sum_{i=1}^{8} \log X_i \geq \log c_* := k$.

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 ith 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 $H_0 : p_1 = \cdots = p_k = 1/k$ 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 4

With significance α, we are testing whether or not each of the r groups have the same proportion for each of the c bloodtypes, that is, we are testing $H_0 : p_{1_j} = \cdots , = p_{r_j}$ for all $j = \{1, \ldots, c\}$ against $H_A : H_0$ is false (in this example, we have $2 \leq c \leq 4$). There are n total observations in the data broken down as follows: N_{ij} is the observation in the ith group and jth blood type, N_{i*} is the number of observations in the ith group across all c bloodtypes, and N_{*j} is the number of observations in the jth bloodtype across all r groups. Using the χ^2 test for homogenity, our test statistic is

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}, \quad \text{where} \quad \hat{E}_{ij} = \frac{N_{i*} \cdot N_{*j}}{n}.$$

We would the reject H_0 when $Q \geq q_\alpha$, where q_α is the $1-\alpha$ quantile of the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom. For this example, the contingency table is given to us, where we have $r = 3$, $c = 4$, and we want to use a significance level of $\alpha = 0.1$. Calculating the rest of the necessary values to obtain Q by hand is tediuos and will not be written here. Using `R`, I found that $Q = 6.829$ and $q_{0.1} = 10.645$, so we would not reject H_0. That is, the proportion of the four blood types seem to be the same for each of the three groups.

Question 5

When we manipulate a conjugacy table such that the row totals and column totals are the same, we are keeping N_{i*} and N_{*j} the same for all i, j, and so $\hat{E}_{ij}$ will remain the same as well. Therefore, when looking at the formula for Q, we can increase its value by making each N_{ij} farther away from $\hat{E}_{ij}$. Considering two groups i, k and two blood types j, ℓ, we can alter the value of Q as follows: $N_{ij} \leftarrow N_{ij} - \epsilon$, $N_{i\ell} \leftarrow N_{i\ell} + \epsilon$, $N_{kj} \leftarrow N_{kj} + \epsilon$, and $N_{k\ell} \leftarrow N_{k\ell} - \epsilon$, where ϵ is some number. As a more concrete example, from the contingency table given to us we have $N_{12} = 6$, $N_{13} = 5$, $N_{22} = 24$, and $N_{23} = 7$, and by having $\epsilon = 1$, the new values become $N_{12} = 5$, $N_{13} = 6$, $N_{22} = 25$, and $N_{23} = 6$. Using these alternate values give $Q = 8.983$.

Question 6

For this question, for all $i, j = \{1, 2\}$ we have $N_{i*} = N_{*j} = 2n$ and $\hat{E}_{ij} = 2n \cdot 2n/4n = n$, and so the test statistic for the χ^2 test of independence is

$$Q = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(N_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{4a^2}{n}.$$

We reject H_0 if $Q \geq q_\alpha$, where q_α is the $1 - \alpha$ quantile of the χ^2 distribution with 1 degree of freedom, and so we would reject H_0 when $a \geq \sqrt{nq_\alpha}/2$ or $a \leq -\sqrt{nq_\alpha}/2$. When $\alpha = 0.01$, we have $q_{0.01} = 6.635$.

Question 7

From this data set, it is possible to see how the proportions of depression levels changed for the patients that took the drug. However, becuase there is no control group to compare the results to another type of treatment (like another drug or a placebo), we cannot tell if the changed in proportions are any more significant relative to another treatment.

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 $\epsilon \sim \mathrm{N}(0, \sigma^2)$.

(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

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

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

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

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

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

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

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

Question 13