50.034 – Introduction to Probability and Statistics

January–May Term, 2019

Homework Set 8

Due by: Week 12 Cohort Class (17 Apr 2019 or 18 Apr 2019)

NOTE: **Make-up for Friday's cohort class** (originally 19th April) will be on **17th April** (Wednesday), 2–4pm, CC14 (2.507).

Reminder: There is **Mini-quiz 4** in Week 12 during your cohort class.

Note: The tables of values for the standard normal distribution, the χ^2 distribution, and the t-distribution, can be found at the back of the course textbook.

Question 1. Let $\{X_1, \dots, X_{10}\}$ be a random sample of observable normal random variables with unknown mean μ and known variance 16. Let $\overline{X}_{10}$ denote the sample mean of $\{X_1, \dots, X_{10}\}$. Let $\mathcal{H} = \{\mathcal{H}_c\}_{c\in\mathbb{R}}$ be a collection of hypothesis tests, where each $\mathcal{H}_c$ represents a hypothesis test with null hypothesis $H_0 : \mu = 2$, test statistic $T = |\overline{X}_{10} - 2|$, and rejection region $R = [c, \infty)$.

(i) Find the value of $c \in \mathbb{R}$ such that $\mathcal{H}_c$ is a level 0.05 test that has the highest power among all level 0.05 tests in $\mathcal{H}$.

(ii) Determine the p-value of $\mathcal{H}$, given the following observed values:

$$X_1 = 1, X_2 = 2, X_3 = 1, X_4 = 2, X_5 = 1, X_6 = 2, X_7 = 1, X_8 = 2, X_9 = 1, X_{10} = 2.$$

Solution. (i) First of all, since $\text{var}(X_i) = 16$ for each i, it follows that $\text{var}(\overline{X}_{10}) = \frac{16}{10}$. The given null hypothesis is $H_0 : \mu = 2$, so if indeed H_0 is true, then $Z = \frac{\sqrt{10}(\overline{X}_{10}-2)}{4}$ has the standard normal distribution. We are given that H_0 is rejected if $T = |\overline{X}_{10} - 2| \geq c$. Note that

$$\begin{aligned}\Pr(T \geq c) &= 1 - \Pr(|\overline{X}_{10} - 2| \leq c) = 1 - \Pr(-\tfrac{\sqrt{10}c}{4} \leq \tfrac{\sqrt{10}(\overline{X}_{10}-2)}{4} \leq \tfrac{\sqrt{10}c}{4})\\ &= 1 - [\Phi(\tfrac{\sqrt{10}c}{4}) - \Phi(-\tfrac{\sqrt{10}c}{4})] = 1 - [\Phi(\tfrac{\sqrt{10}c}{4}) - (1 - \Phi(\tfrac{\sqrt{10}c}{4}))] = 2 - 2 \cdot \Phi(\tfrac{\sqrt{10}c}{4}),\end{aligned}$$

where $\Phi(z)$ denotes the standard normal cumulative distribution function.

We want a significance level of 0.05, which means we want $2 - 2 \cdot \Phi(\frac{\sqrt{10}c}{4}) \leq 0.05$, or equivalently, $\Phi(\frac{\sqrt{10}c}{4}) \geq 0.975$. From the table of values for the standard normal distribution, $\Phi(1.96) = 0.9750$, hence $\Phi(\frac{\sqrt{10}c}{4}) \geq 0.975$ implies that $\frac{\sqrt{10}c}{4} \geq 1.96$, or equivalently, that $c \geq \frac{7.84}{\sqrt{10}} \approx 2.4792$. To maximize the power of $\mathcal{H}$, we need to find the smallest possible c satisfying $c \geq 2.4792$, thus $c = 2.4792$.

(ii) Given the observed values $X_1 = 1, X_2 = 2, X_3 = 1, X_4 = 2, X_5 = 1, X_6 = 2, X_7 = 1, X_8 = 2, X_9 = 1, X_{10} = 2$, the corresponding observed value for $\overline{X}_{10}$ is $\frac{1+2+1+2+1+2+1+2+1+2}{10} = 1.5$. Thus, the null hypothesis H_0 will be rejected whenever $c \leq |1.5 - 2| = 0.5$. Note that $c \leq 0.5$ corresponds to the significance level $2 - 2 \cdot \Phi(\frac{\sqrt{10}c}{4}) \geq 2 - 2 \cdot \Phi(0.3953) \approx 2 - 2(0.6554) \approx 0.6892$. Therefore, the p-value of $\mathcal{H}$ is approximately 0.6892.

Question 2. Let $\{X_1, \dots, X_{20}\}$ be a random sample of observable normal random variables with unknown mean μ and unknown variance σ^2. Let $\overline{X}_{20}$ and s_{20}^2 denote the sample mean and unbiased sample variance respectively of $\{X_1, \dots, X_{20}\}$. Let $\mathcal{H} = \{\mathcal{H}_c\}_{c\in\mathbb{R}}$ be a collection of hypothesis tests, where each $\mathcal{H}_c$ represents a hypothesis test with null hypothesis $H_0 : \mu = 0$, test statistic $T = \left|\frac{\overline{X}_{20}}{s_{20}}\right|$, and rejection region $R = [c, \infty)$.

(i) Find the value of $c \in \mathbb{R}$ such that $\mathcal{H}_c$ is a level 0.05 test that has the highest power among all level 0.05 tests in $\mathcal{H}$.

(ii) Determine the p-value of $\mathcal{H}$, given the following observed values:

$X_1 = -1,\quad X_2 = 0,\quad X_3 = 1,\quad X_4 = 0,\quad X_5 = -1,\quad X_6 = 0,\quad X_7 = 1,\quad X_8 = 0,\quad X_9 = -1,\quad X_{10} = 0,$
$X_{11} = 1,\quad X_{12} = 0,\quad X_{13} = -1,\quad X_{14} = 0,\quad X_{15} = 1,\quad X_{16} = 0,\quad X_{17} = -1,\quad X_{18} = 0,\quad X_{19} = 1,\quad X_{20} = 0.$

Solution. (i) The given null hypothesis is $H_0 : \mu = 0$, so if indeed H_0 is true, then $Z = \frac{\sqrt{20}(\overline{X}_{20})}{s_{20}(X_1,\ldots,X_{20})}$ has the t-distribution with 19 degrees of freedom. We are given that H_0 is rejected if $T = |\frac{\overline{X}_{20}}{s_{20}}| \geq c$. Note that

$$\begin{aligned}\Pr(T \geq c) &= 1 - \Pr(|\tfrac{\overline{X}_{20}}{s_{20}}| \leq c) = 1 - \Pr(-\sqrt{20}c \leq \tfrac{\overline{X}_{20}}{s_{20}} \leq \sqrt{20}c) \\ &= 1 - [F(\sqrt{20}c) - F(-\sqrt{20}c)] = 1 - [F(\sqrt{20}c) - (1 - F(\sqrt{20}c))] \\ &= 2 - 2 \cdot F(\sqrt{20}c),\end{aligned}$$

where $F(z)$ denotes the cumulative distribution function of Z.
We want a significance level of 0.05, which means we want $2 - 2 \cdot F(\sqrt{20}c) \leq 0.05$, or equivalently, $F(\sqrt{20}c) \geq 0.975$. From the table of values for the t-distribution, $F(2.093) = 0.975$, hence $F(\sqrt{20}c) \geq 0.975$ implies that $\sqrt{20}c \geq 2.093$, or equivalently, that $c \geq \frac{2.093}{\sqrt{20}} \approx 0.4680$. To maximize the power of $\mathcal{H}$, we need to find the smallest possible c satisfying $c \geq 0.4680$, thus $c = 0.4680$.

(ii) Given the observed values for $X_1, \ldots, X_{20}$, we can calculate the corresponding observed value for $\overline{X}_{20}$ to equal 0. Thus, the null hypothesis H_0 will be rejected whenever $c \leq |0| = 0$. Note that $c \leq 0$ corresponds to the significance level $2 - 2 \cdot F(0) \geq 2 - 2(0.5) = 1$. Therefore, the p-value of $\mathcal{H}$ is exactly 1.

Question 3. Let $\{X_1, \ldots, X_{20}\}$ be a random sample of 20 observable normal random variables with unknown mean μ and unknown variance σ^2. Let $\overline{X}_{20}$ and $\hat{\sigma}^2_{20}$ denote the sample mean and biased sample variance respectively of $\{X_1, \ldots, X_{20}\}$. Consider a hypothesis test $\mathcal{H}$ with null hypothesis $H_0 : \sigma^2 = 25$. Let $T = \hat{\sigma}^2_{20}$ be the test statistic, and let $R = [c, \infty)$ be the rejection region of $\mathcal{H}$, where c is some constant to be determined. Find the value of c that maximizes the power of $\mathcal{H}$ at significance level 0.1.

Solution. The given null hypothesis is $H_0 : \sigma^2 = 25$, so if indeed H_0 is true, then $Z = \frac{20\hat{\sigma}^2_{20}}{\sigma^2} = 0.8\hat{\sigma}^2_{20}$ has the χ^2 distribution with 19 degrees of freedom. We are given that H_0 is rejected if $T = \hat{\sigma}^2_{20} \geq c$. Note that

$$\Pr(T \geq c) = 1 - \Pr(\hat{\sigma}^2_{20} \leq c) = 1 - \Pr(0.8\hat{\sigma}^2_{20} \leq 0.8c) = 1 - F(0.8c)$$

where $F(z)$ denotes the cumulative distribution function of Z.

We want a significance level of 0.1, which means we want $1 - F(0.8c) \leq 0.1$, or equivalently, $F(0.8c) \geq 0.9$. From the table of values for the χ^2 distribution, $F(27.20) = 0.9$, hence $F(0.8c) \geq 0.9$ implies that $0.8c \geq 27.20$, or equivalently, that $c \geq 34$. To maximize the power of $\mathcal{H}$, we need to find the smallest possible c satisfying $c \geq 34$, thus $c = 34$.

Question 4. Let $\{X_1, \ldots, X_{20}\}$ be a random sample of observable Poisson random variables with unknown mean θ. Suppose we are given that the parameter space of θ contains only two possible values 1 and 2. Find a most powerful hypothesis test with significance level 0.05, such that its null hypothesis is $H_0 : \theta = 1$. Please give a complete description of this most powerful hypothesis test $\mathcal{H}$, including the test statistic and rejection region of $\mathcal{H}$.

Solution. We want a hypothesis test with null hypothesis $H_0 : \theta = 1$ and alternative hypothesis $H_1 : \theta = 2$. Since both the null hypothesis and the alternative hypothesis are simple hypotheses, the Neyman–Pearson lemma tells us that a most powerful hypothesis test with the given hypotheses H_0 and H_1 at significance level 0.05 is a hypothesis test $\mathcal{H}$ whose test statistic is the likelihood ratio, and whose rejection region is $[c, \infty)$, for some suitable value of c to be determined.

Given any vector $\mathbf{x} = (x_1, \ldots, x_{20})$ of possible observed values for $(X_1, \ldots, X_{20})$ and any real value $\lambda \geq 0$, let $\mathcal{L}(\lambda|\mathbf{x})$ denote the likelihood function of θ given $\mathbf{x}$, evaluated at $\theta = \lambda$. Since $X_1, \ldots, X_{20}$ are iid Poisson random variables with parameter θ, the conditional probability mass function of each X_i is

$$f_{X_i|\theta}(x_i|\lambda) = \begin{cases} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}, & \text{if } x_i = 0, 1, 2, 3, \ldots; \\ 0, & \text{otherwise.} \end{cases}$$

thus it follows from the independence of $X_1, \ldots, X_{20}$ that the likelihood function of θ evaluated at $\theta = \lambda$ is

$$\begin{aligned} \mathcal{L}(\lambda|\mathbf{x}) &= \prod_{i=1}^{20} f_{X_i|\theta}(x_i|\lambda) \\ &= \begin{cases} \dfrac{\lambda^{(x_1 + \cdots + x_{20})} e^{-20\lambda}}{x_1! \cdots x_{20}!}, & \text{if } x_i \text{ is a non-negative integer for all i;} \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

By assumption, each x_i is a possible value for X_i and so must be a non-negative integer. Consequently, the likelihood ratio of $\mathbf{x}$ is

$$\frac{\mathcal{L}(2|\mathbf{x})}{\mathcal{L}(1|\mathbf{x})} = \frac{\frac{2^{(x_1+\cdots+x_{20})} e^{-40}}{x_1! \cdots x_{20}!}}{\frac{e^{-20}}{x_1! \cdots x_{20}!}} = 2^{(x_1+\cdots+x_{20})} e^{-20}.$$

Thus, we could set the test statistic of $\mathcal{H}$ to be $T = 2^{(X_1+\cdots+X_{20})} e^{-20}$. Note that

$$\begin{aligned} \Pr(T \geq c) &= 1 - \Pr(2^{(X_1+\cdots+X_{20})} \leq c \cdot e^{20}) = 1 - \Pr((X_1 + \cdots + X_{20}) \log 2 \leq 20 + \log c) \\ &= 1 - \Pr((X_1 + \cdots + X_{20}) \leq \tfrac{20+\log c}{\log 2}) = 1 - \Pr(Y \leq \tfrac{20+\log c}{\log 2}) \end{aligned}$$

This implies that $\Pr(T \geq c) \leq 0.05$ if and only if $\Pr(Y \leq \frac{20+\log c}{\log 2}) \geq 0.95$. So, if indeed H_0 is true, then Y is Poisson with parameter 20, which implies that

$$\Pr(Y \leq \tfrac{20+\log c}{\log 2}) = \sum_{k=0}^{\lfloor \frac{20+\log c}{\log 2} \rfloor} e^{-20} \tfrac{20^k}{k!}.$$

We check that $\sum_{k=0}^{27} e^{-20} \frac{20^k}{k!} \approx 0.9475 < 0.95$ while $\sum_{k=0}^{28} e^{-20} \frac{20^k}{k!} \approx 0.9657 > 0.95$, thus

$$\Pr(Y \leq \tfrac{20+\log c}{\log 2}) \geq 0.95$$

if and only if $\frac{20+\log c}{\log 2} \geq 28$, or equivalently, if and only if

$$c \geq \exp((28 \log 2) - 20) \approx 0.5533.$$

So, if $\mathcal{H}$ is a hypothesis test with null hypothesis $H_0 : \theta = 1$, test statistic $T = 2^{(X_1+\cdots+X_{20})} e^{-20}$, and rejection region $[0.5533, \infty)$, then $\mathcal{H}$ is a most powerful test with null hypothesis H_0 and significance level 0.05.

Question 5. Let $\{X_1, \ldots, X_{15}\}$ be a random sample of observable Bernoulli random variables with unknown parameter θ. Assume that the parameter space of θ is the interval $[0, 1]$. Find a uniformly most powerful hypothesis test with significance level 0.1, such that its null hypothesis is $H_0 : \theta \geq 0.4$. Please give a complete description of this uniformly most powerful hypothesis test $\mathcal{H}$, including the test statistic and rejection region of $\mathcal{H}$.

Solution. Note: This question would be straightforward if you had done the assigned readings (Section 9.3 of course textbook in particular).

Consider the null hypothesis $H_0 : \theta \geq 0.4$ and alternative hypothesis $H_1 : \theta < 0.4$. By Example 9.3.3 of the course textbook, the joint probability mass function of $X_1, \ldots, X_{15}$ has what is called a monotone likelihood ratio in the statistic $Y = X_1 + \cdots + X_{15}$ (as defined in Definition 9.3.2). Thus, by Theorem 9.3.1 of the course textbook, a hypothesis test that rejects H_0 when $Y \leq c$ (for some real number c) will be a most powerful test for the hypotheses H_0 and H_1.

For each specific choice of c, the size of $\mathcal{H}$ will be $\Pr(Y \leq c | \theta = 0.4)$. Since Y is a sum of iid Bernoulli random variables with parameter θ, it follows that Y is a binomial random variable with parameters 15 and θ, thus

$$\Pr(Y \leq c | \theta = 0.4) = \sum_{x=0}^{c} \binom{15}{x} (0.4)^x (0.6)^{15-x}.$$

We check that $\sum_{x=0}^{3} \binom{15}{x} (0.4)^x (0.6)^{15-x} \approx 0.0905 < 0.1$ while $\sum_{x=0}^{4} \binom{15}{x} (0.4)^x (0.6)^{15-x} \approx 0.2173 > 0.1$, thus $\Pr(Y \leq c | \theta = 0.4) \leq 0.1$ if and only if $c \leq 3$. Therefore, if $\mathcal{H}$ is a hypothesis test with null hypothesis $H_0 : \theta \geq 0.4$, test statistic $T = X_1 + \cdots + X_{15}$, and rejection region $(-\infty, 3]$, then $\mathcal{H}$ is a most powerful test with null hypothesis H_0 and significance level 0.1.