50.034 – Introduction to Probability and Statistics

January–May Term, 2019

Homework Set 7

Due by: Week 11 Cohort Class (11 Apr 2019 or 12 Apr 2019)

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

Question 1. Let $\{X_1, \dots, X_n\}$ be a random sample of n observable normal random variables with mean μ and variance σ^2. Assume that $n \geq 3$, and that $\sigma^2 \neq 0$. Let $\hat{\sigma}^2$ denote the biased sample variance of this random sample $\{X_1, \dots, X_n\}$. For each of the two inequalities given below, determine the smallest possible value for the sample size n such that the inequality is satisfied:

(i) $\Pr\big(\frac{\hat{\sigma}^2}{\sigma^2} \leq 1.2\big) \geq 0.8$.
(ii) $\Pr\big(|\hat{\sigma}^2 - 0.77\sigma^2| \leq 0.37\sigma^2\big) \geq 0.725$.

Solution. (i) The random variable $\frac{n\hat{\sigma}^2}{\sigma^2}$ has a χ^2 distribution with $(n-1)$ degrees of freedom. Note that $\Pr\big(\frac{\hat{\sigma}^2}{\sigma^2} \leq 1.2\big) \geq 0.8$ is equivalent to $\Pr\big(\frac{n\hat{\sigma}^2}{\sigma^2} \leq 1.2n\big) \geq 0.8$. To find the smallest value for n satisfying this inequality, it suffices to find an integer value for n such that

- $\Pr\big(\frac{n\hat{\sigma}^2}{\sigma^2} \leq 1.2n\big) \geq 0.8$; and
- $\Pr\big(\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \leq 1.2(n-1)\big) < 0.8$.

From the table of values for the χ^2 distribution (at the back of the course textbook), we find that

- $\Pr\big(\frac{19\hat{\sigma}^2}{\sigma^2} \leq 1.2 \times (19) = 22.8\big) > \Pr\big(\frac{19\hat{\sigma}^2}{\sigma^2} \leq 22.76\big) = 0.8$.
- $\Pr\big(\frac{18\hat{\sigma}^2}{\sigma^2} \leq 1.2 \times (18) = 21.6\big) < \Pr\big(\frac{18\hat{\sigma}^2}{\sigma^2} \leq 21.61\big) = 0.8$.

Therefore, the smallest sample size n satisfying $\Pr\big(\frac{\hat{\sigma}^2}{\sigma^2} \leq 1.2\big) \geq 0.8$ is $n = 19$.

(ii) Recall that the random variable $\frac{n\hat{\sigma}^2}{\sigma^2}$ has a χ^2 distribution with $(n-1)$ degrees of freedom. Note that

$$\begin{aligned}
\Pr\big(|\hat{\sigma}^2 - 0.77\sigma^2| \leq 0.37\sigma^2\big) \geq 0.725 &\iff \Pr\big(|\tfrac{\hat{\sigma}^2}{\sigma^2} - 0.77| \leq 0.37\big) \geq 0.725 \\
&\iff \Pr\big(-0.37 \leq \tfrac{\hat{\sigma}^2}{\sigma^2} - 0.77 \leq 0.37\big) \geq 0.725 \\
&\iff \Pr\big(0.4 \leq \tfrac{\hat{\sigma}^2}{\sigma^2} \leq 1.14\big) \geq 0.725 \\
&\iff \Pr\big(0.4n \leq \tfrac{n\hat{\sigma}^2}{\sigma^2} \leq 1.14n\big) \geq 0.725.
\end{aligned}$$

To find the smallest value for n satisfying this inequality, it suffices to find an integer value for n such that

- $\Pr\big(0.4n \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq 1.14n\big) \geq 0.725$; and
- $\Pr\big(0.4(n-1) \leq \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \leq 1.14(n-1)\big) < 0.725$.

From the table of values for the χ^2 distribution, we find that

$$\begin{aligned}
\Pr\big(0.4(17) \leq \tfrac{17\hat{\sigma}^2}{\sigma^2} \leq 1.14(17)\big) &= \Pr\big(6.8 \leq \tfrac{17\hat{\sigma}^2}{\sigma^2} \leq 19.38\big) \\
&= \Pr\big(\tfrac{17\hat{\sigma}^2}{\sigma^2} \leq 19.38\big) - \Pr\big(\tfrac{17\hat{\sigma}^2}{\sigma^2} \leq 6.8\big) \\
&> \Pr\big(\tfrac{17\hat{\sigma}^2}{\sigma^2} \leq 19.37\big) - \Pr\big(\tfrac{17\hat{\sigma}^2}{\sigma^2} \leq 6.908\big) \\
&= 0.75 - 0.025 = 0.725.
\end{aligned}$$

$$\begin{aligned}\Pr\big(0.4(16) \le \tfrac{16\hat{\sigma}^2}{\sigma^2} \le 1.14(16)\big) &= \Pr\big(6.4 \le \tfrac{16\hat{\sigma}^2}{\sigma^2} \le 18.24\big)\\ &= \Pr\big(\tfrac{16\hat{\sigma}^2}{\sigma^2} \le 18.24\big) - \Pr\big(\tfrac{16\hat{\sigma}^2}{\sigma^2} \le 6.4\big)\\ &< \Pr\big(\tfrac{16\hat{\sigma}^2}{\sigma^2} \le 18.25\big) - \Pr\big(\tfrac{16\hat{\sigma}^2}{\sigma^2} \le 6.262\big)\\ &= 0.75 - 0.025 = 0.725.\end{aligned}$$

Therefore, the smallest sample size n satisfying $\Pr\big(|\hat{\sigma}^2 - 0.77\sigma^2| \le 0.37\sigma^2\big) \ge 0.725$ is $n = 17$.

Question 2. Consider a statistical model consisting of observable exponential random variables $X_1, \ldots, X_n$ that are conditionally iid given the parameter θ. Let $\hat{\theta}$ be the maximum likelihood estimator of θ. What is the conditional probability density function of the random variable $n(\hat{\theta})^{-1}$ given $\theta = 5$? (Hint: Theorem 5.7.7 of the course textbook may be useful.)

Solution. As covered in Lecture 16 (slide 21), the maximum likelihood estimator of θ is $\hat{\theta} = \frac{1}{\overline{X}_n} = \frac{n}{X_1+\cdots+X_n}$, where $\overline{X}_n$ denotes the sample mean of $\{X_1, \ldots, X_n\}$. We are given that each X_i has the exponential distribution with parameter θ, or equivalently, each X_i has the gamma distribution with parameters 1 and θ. Since $X_1, \ldots, X_n$ are independent, it follows from Theorem 5.7.7 of the course textbook that $X_1 + \cdots + X_n$ has the gamma distribution with parameters n and θ. Since $n(\hat{\theta})^{-1} = X_1 + \cdots + X_n$, it follows that the conditional probability density function of $n(\hat{\theta})^{-1}$ given $\theta = 5$, which we denote by $f_{n(\hat{\theta})^{-1}|\theta}(x|5)$, is the probability density function of a gamma random variable with parameters n and 5. Therefore, using $\Gamma(n) = (n-1)!$ (for any positive intege n), we get

$$f_{n(\hat{\theta})^{-1}|\theta}(x|5) = \begin{cases} \frac{5^n}{(n-1)!}x^{n-1}e^{-5x}, & \text{if } x \ge 0.\\ 0, & \text{if } x < 0;\end{cases}$$

Question 3. An automated juice vending machine dispenses orange juice in cups. A total of 25 dispensed cups of orange juice are collected, and the amount of juice in each cup is measured. Consider a statistical model consisting of a random sample $\{X_1, \ldots, X_{25}\}$ of normal random variables with unknown mean μ and unknown variance σ^2, where each X_i represents the amount of juice (in ml) in the i-th dispensed cup. Let $\overline{X}_{25}$ and s_{25}^2 be the sample mean and the unbiased sample variance respectively of $\{X_1, \ldots, X_{25}\}$.

(i) Find an exact 90% confidence interval for μ in terms of $\overline{X}_{25}$ and s_{25}^2.

(ii) After measuring the amount of juice in these 25 dispensed cups, it was noticed that the 25 cups have a mean amount of 251 ml and a standard deviation of 4 ml. Using this information, find an observed value of the 90% confidence interval that you found in the previous part.

Solution. (i) Note that $Z = \frac{\sqrt{25}(\overline{X}_{25}-\mu)}{s_{25}} = \frac{5(\overline{X}_{25}-\mu)}{s_{25}}$ has the t-distribution with 24 degrees of freedom. Let $F(z)$ be the cumulative distribution function of Z. Then for any real number $r > 0$,

$$\begin{aligned}\Pr(\overline{X}_{25} - r < \mu < \overline{X}_{25} + r) &= \Pr(-r < \overline{X}_{25} - \mu < r) = \Pr\Big(-\tfrac{5r}{s_{25}} < \tfrac{5(\overline{X}_{25}-\mu)}{s_{25}} < \tfrac{5r}{s_{25}}\Big)\\ &= F\big(\tfrac{5r}{s_{25}}\big) - F\big(-\tfrac{5r}{s_{25}}\big) = F\big(\tfrac{5r}{s_{25}}\big) - \Big(1 - F\big(\tfrac{5r}{s_{25}}\big)\Big)\\ &= 2\cdot F\big(\tfrac{5r}{s_{25}}\big) - 1.\end{aligned}$$

Thus $\Pr(\overline{X}_{25} - r < \mu < \overline{X}_{25} + r) = 0.9$ if and only if $2\cdot F\big(\frac{5r}{s_{25}}\big) - 1 = 0.9$, or equivalently, if and only if $F\big(\frac{5r}{s_{25}}\big) = 0.95$.

From the table of values for the t-distribution (at the back of the course textbook), the closest value of z (that we can find from the table) satisfying $F(z) = 0.95$ for 24 degrees of freedom is $z = 1.711$, hence $\frac{5r}{s_{25}} \approx 1.711$, which implies $r \approx 0.3422s_{25}$. Therefore an exact 90% confidence interval for μ is $(\overline{X}_{25} - 0.3422s_{25}, \overline{X}_{25} + 0.3422s_{25})$.

(ii) We are given that some specific observed values $X_1 = x_1, \dots, X_{25} = x_{25}$ satisfy $\overline{x}_{25} = \frac{x_1+\cdots+x_{25}}{25} = 251$, as well as

$$\frac{1}{25}\sum_{i=1}^{25}(x_i - \overline{x}_{25})^2 = 4^2 = 16,$$

which is the observed value of the biased sample variance $\hat{\sigma}^2_{25}(X_1, \dots, X_{25})$. Note that

$$s^2_{25}(X_1, \dots, X_{25}) = \frac{25}{24}\hat{\sigma}^2_{25}(X_1, \dots, X_{25}),$$

hence $s_{25}(x_1, \dots, x_{25}) = \sqrt{\frac{25}{24}}\hat{\sigma}_{25}(x_1, \dots, x_{25}) = \frac{5}{\sqrt{24}}(4) = \frac{20}{\sqrt{24}}$.
Therefore, an observed value of the 90% confidence interval found in the previous part is the open interval

$$\Big(251 - 0.3422\tfrac{20}{\sqrt{24}}, 251 + 0.3422\tfrac{20}{\sqrt{24}}\Big) \approx (249.603, 252.397).$$

Question 4. At a Fuji apple farm in Fujisaki, Japan, 10 Fuji apples are randomly selected and weighed. Consider a statistical model consisting of a random sample $\{X_1, \dots, X_{10}\}$ of normal random variables with unknown mean μ and unknown variance σ^2, where each X_i represents the weight (in grams) of the i-th selected apple. Let $\overline{X}_{10}$ and s^2_{10} be the sample mean and the unbiased sample variance respectively of $\{X_1, \dots, X_{10}\}$.

(i) Find an exact 95% confidence interval for μ in terms of $\overline{X}_{10}$ and s^2_{10}.

(ii) The measurements of all 10 weights (in grams) are indicated below:

$$148, 150, 155, 154, 152, 148, 155, 160, 152, 145.$$

Using these values, find an observed value of the 95% confidence interval for μ that you found in the previous part.

Solution. (i) Note that $Z = \frac{\sqrt{10}(\overline{X}_{10}-\mu)}{s_{10}}$ has the t-distribution with 9 degrees of freedom. Let $F(z)$ be the cumulative distribution function of Z. Then for any real number $r > 0$,

$$\begin{aligned}\Pr(\overline{X}_{10} - r < \mu < \overline{X}_{10} + r) &= \Pr(-r < \overline{X}_{10} - \mu < r) = \Pr\Big(-\tfrac{\sqrt{10}r}{s_{10}} < \tfrac{\sqrt{10}(\overline{X}_{10}-\mu)}{s_{10}} < \tfrac{\sqrt{10}r}{s_{10}}\Big)\\ &= F\big(\tfrac{\sqrt{10}r}{s_{10}}\big) - F\big(-\tfrac{\sqrt{10}r}{s_{10}}\big) = F\big(\tfrac{\sqrt{10}r}{s_{10}}\big) - \Big(1 - F\big(\tfrac{\sqrt{10}r}{s_{10}}\big)\Big)\\ &= 2 \cdot F\big(\tfrac{\sqrt{10}r}{s_{10}}\big) - 1.\end{aligned}$$

Thus $\Pr(\overline{X}_{10}-r < \mu < \overline{X}_{10}+r) = 0.95$ if and only if $2\cdot F\big(\frac{\sqrt{10}r}{s_{10}}\big)-1 = 0.95$, or equivalently, if and only if $F\big(\frac{\sqrt{10}r}{s_{10}}\big) = 0.975$.
From the table of values for the t-distribution (at the back of the course textbook), the closest value of z (that we can find from the table) satisfying $F(z) = 0.975$ for 9 degrees of freedom is $z = 2.262$, hence $\frac{\sqrt{10}r}{s_{10}} \approx 2.262$, which implies $r \approx 0.7153s_{10}$. Therefore an exact 95% confidence interval for μ is $(\overline{X}_{10} - 0.7153s_{10}, \overline{X}_{10} + 0.7153s_{10})$.

(ii) We are given the vector of observed values

$$(x_1, \dots, x_{10}) = (148, 150, 155, 154, 152, 148, 155, 160, 152, 145).$$

We calculate that $\overline{x}_{10} = \frac{x_1+\cdots+x_{10}}{10} = 151.9$. We also calculate that the observed value of the unbiased sample variance is

$$\begin{aligned}\frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x}_{10})^2 &= \frac{1}{9}\Big[(-3.9)^2 + (-1.9)^2 + (3.1)^2 + (2.1)^2 + (0.1)^2 \\ &\qquad + (-3.9)^2 + (3.1)^2 + (8.1)^2 + (0.1)^2 + (-6.9)^2\Big] \\ &= \frac{170.9}{9} \approx 18.989.\end{aligned}$$

Therefore, an observed value of the 95% confidence interval found in the previous part is the open interval

$$\left(151.9 - 0.7153\sqrt{\tfrac{170.9}{9}}, 151.9 + 0.7153\sqrt{\tfrac{170.9}{9}}\right) \approx \big(148.783, 155.017\big).$$

Question 5. At a steel factory, the tensile strength of 20 randomly cut steel sample pieces are measured. Consider a statistical model consisting of a random sample $\{X_1, \ldots, X_{20}\}$ of normal random variables with unknown mean μ and unknown variance σ^2, where each X_i represents the tensile strength (in MPa) of the i-th steel sample piece. Let $\overline{X}_{20}$ and s_{20}^2 be the sample mean and the unbiased sample variance respectively of $\{X_1, \ldots, X_{20}\}$.

(i) Find an exact 95% confidence interval for the variance σ^2 in terms of $\overline{X}_{20}$ and s_{20}^2.

(ii) After the 20 measurements have been collected, it was noticed that the 20 steel sample pieces have a mean tensile strength of 355 MPa and a standard deviation of 25 MPa. Using this information, find an observed value of the 95% confidence interval that you found in the previous part.

Solution. (i) Let $\hat{\sigma}_{20}^2$ be the unbiased sample variance of $\{X_1, \ldots, X_{20}\}$. Note that $Z = \frac{20\hat{\sigma_{20}}^2}{\sigma^2}$ has the χ^2 distribution with 19 degrees of freedom. Let $F(z)$ be the cumulative distribution function of Z. Then for any real number $c_1, c_2 > 0$,

$$\begin{aligned}\Pr(c_1\hat{\sigma}_{20}^2 < \sigma^2 < c_2\hat{\sigma}_{20}^2) &= \Pr\left(\tfrac{1}{c_2\hat{\sigma}_{20}^2} < \tfrac{1}{\sigma^2} < \tfrac{1}{c_1\hat{\sigma}_{20}^2}\right) = \Pr\left(\tfrac{20}{c_2} < \tfrac{20\hat{\sigma}_{20}^2}{\sigma^2} < \tfrac{20}{c_1}\right) \\ &= F\big(\tfrac{20}{c_1}\big) - F\big(\tfrac{20}{c_2}\big).\end{aligned}$$

We want to find values for c_1, c_2 such that $\Pr(c_1\hat{\sigma}_{20}^2 < \sigma^2 < c_2\hat{\sigma}_{20}^2) = F\big(\frac{20}{c_1}\big) - F\big(\frac{20}{c_2}\big) = 0.95$. There are infinitely many such possible pairs of values for c_1, c_2. We shall find one specific pair.

Note that $0.975 - 0.025 = 0.95$. From the table of values for the χ^2-distribution (at the back of the course textbook), we have $F(32.85) = 0.975$ and $F(8.907) = 0.025$ (for 19 degrees of freedom), so we could set $\frac{20}{c_1} = 32.85$, $\frac{20}{c_2} = 8.907$, or equivalently, set $c_1 \approx 0.6088$, $c_2 \approx 2.2454$. Thus, an exact 95% confidence interval for the variance σ^2 is $(0.6088\hat{\sigma}_{20}^2 < \sigma^2 < 2.2454\hat{\sigma}_{20}^2)$.

Finally, since we are asked to express the exact 95% confidence interval in terms of $\overline{X}_{20}$ and s_{20}^2, we shall use the fact that $\hat{\sigma}_{20}^2 = \frac{19}{20}s_{20}^2$ to get the following exact 95% confidence interval for σ^2:

$$(0.57836s_{20}^2 < \sigma^2 < 2.1331s_{20}^2).$$

(ii) We are given that some specific observed values $X_1 = x_1, \ldots, X_{20} = x_{20}$ satisfy $\overline{x}_{20} = \frac{x_1+\cdots+x_{20}}{20} = 355$, as well as $\frac{1}{20}\sum_{i=1}^{20}(x_i - \overline{x}_{20})^2 = 25^2 = 625$, which is the observed value of the biased sample variance $\hat{\sigma}_{20}^2(X_1, \ldots, X_{20})$.

Therefore, an observed value of the 95% confidence interval found in the previous part is the open interval

$$\left(0.6088\hat{\sigma}_{20}^2 < \sigma^2 < 2.2454\hat{\sigma}_{20}^2\right) \approx (380.5, 1403.4).$$