Fall 2020 STAT 88 Final Solutions

Wooseok Ha

Dec 2020

1. A random variable X has the distribution shown in the table below. Let μ_X and σ_X denote the mean and standard deviation of X.

x	-1	0	1
$P(X = x)$	$1/18$	$16/18$	$1/18$

(a) Show that $\mu_X = 0$ and $\sigma_X = \frac{1}{3}$.

(b) Use the probability distribution of X to calculate $P(|X - \mu_X| \geq 3\sigma_X)$. Compare this exact probability with the upper bound provided by Chebyshev's inequality and show that the bound provided by Chebyshev's inequality is attained.

(c) Now we take n independent random samples, $X_1, X_2, \ldots, X_n$, drawn from the same probability distribution as X. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean. Apply Chebyshev's inequality to the sample mean to determine the smallest sample size n so that $P(|\bar{X} - \mu_X| \geq 3\sigma_X)$ is less than 0.01. Here μ_X and σ_X are still the mean and standard deviation of X in previous parts.

(d) Do you think the same statement of part(b) holds for the sample mean $\bar{X}$, i.e. if we use the probability distribution of $\bar{X}$ to calculate $P(|\bar{X} - E(\bar{X})| \geq 3 \cdot \mathrm{SD}(\bar{X}))$, it attains the bound provided by Chebyshev's inequality? Show your work. *(Hint: consider the case when n is large and use the fact that $\Phi(-3) = 1 - \Phi(3) \approx 0.0013$.)*

Solution

(a) From the distribution table, we can calculate $E(X) = 0$ and $E(X^2) = \frac{1}{9}$. Then $\sigma_X = \sqrt{E(X^2) - (EX)^2} = \frac{1}{3}$.

(b) Using the probability distribution of X, we obtain

$$P(|X - \mu_X| \geq 3\sigma_X) = P(|X| \geq 1) = P(X = 1) + P(X = -1) = \frac{1}{9}.$$

Meanwhile, Chebyshev's inequality gives the bound

$$P(|X - \mu_X| \geq 3\sigma_X) \leq \frac{\mathrm{Var}(X)}{9\sigma_X^2} = \frac{1}{9}.$$

Hence the bound provided by Chebyshev's inequality is attained.

(c) We know $E(\bar{X}) = E(X) = 0$ and $\mathrm{SD}(\bar{X}) = \frac{\sigma_X}{\sqrt{n}}$. Applying Chebyshev's inequality, we get

$$P(|\bar{X} - \mu_X| \geq 3\sigma_X) \leq \frac{\mathrm{Var}(\bar{X})}{9\sigma_X^2} = \frac{1}{9n}.$$

To find $\frac{1}{9n} \leq 0.01$, we must have $n \geq \frac{100}{9}$. So $n = 12$.

(d) Chebyshev's inequality gives

$$P(|\bar{X} - \mu_{\bar{X}}| \geq 3 \cdot \mathrm{SD}(\bar{X})) \leq \frac{1}{9}.$$

On the other hand, when n is large, we know from CLT that $\bar{X}$ is approximately normal. Then

$$P(|\bar{X} - \mu_{\bar{X}}| \geq 3 \cdot \mathrm{SD}(\bar{X})) \approx P(|Z| \geq 3) = 2 \cdot \Phi(-3) = 0.0026,$$

where $Z \sim \mathcal{N}(0, 1)$. So the Chebyshev's bound is not attained.

2. An experiment was conducted to observe the effect of a new antibiotic to treat an infection. Suppose that 100 randomly selected patients are given a standard antibiotic, and 85 of them recovered from an infection. Of the 100 randomly selected patients given a new antibiotic, 90 patients recovered from an infection. The patients in the two groups are independent of each other.

(a) Let p_1 denote the probability of recovery under the standard treatment and let p_2 denote the probability of recovery under the new treatment. Construct an approximate 95% confidence interval for $p_1 - p_2$, the difference of percentages of recovery under the standard treatment and the new treatment.

(b) Perform a test at a 5% level whether the percentage of recovery under the standard antibiotic differs from the percentage of recovery under the new antibiotic. You should follow the 5 steps for hypothesis testing discussed in class. *(Hint: use the fact that $\Phi(-1) \approx 0.1587$.)*

(c) Now suppose that 200 patients participated in the experiment and out of 200, a simple random sample of 100 participants received the standard treatment and the remaining participants received the new treatment. Again, 85 patients recovered in the first group, and 90 patients recovered in the second group. Perform a test whether the standard and the new antibiotic have different effects on the infection. State your null hypothesis, alternative hypothesis, test statistics, and p-value. You can write the p-value as an expression of finite sum and don't have to make a conclusion about rejecting or accepting the null.

Solution

(a) The confidence interval for $p_1 - p_2$ is given by

$$\bar{X} - \bar{Y} \pm 2\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}.$$

Plugging in $\bar{X} = 0.85$ and $\bar{Y} = 0.9$,

$$\begin{aligned}
&0.85 - 0.9 \pm 2\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}} \\
&\approx 0.85 - 0.9 \pm 2\sqrt{\frac{0.85(1-0.85)}{100} + \frac{0.9(1-0.9)}{100}} \\
&\approx (-0.1433, 0.0433).
\end{aligned}$$

(b) The null and alternative hyothesis are $H_0 : p_1 = p_2 = p$ and $H_A : p_1 \neq p_2$. Our test statistic is $\bar{X} - \bar{Y}$. Under H_0, we approximate p by the weighted average of $\bar{X}$ and $\bar{Y}$, i.e.

$$p = 0.5 \cdot 0.85 + 0.5 \cdot 0.9 = 0.875.$$

Then the null distribution of $\bar{X} - \bar{Y}$ is approximately normal with mean 0 and SD

$$\sqrt{p(1-p)/100 + p(1-p)/100} \approx 0.0505,$$

and so

$$p\text{-value} = P(|\bar{X} - \bar{Y}| \geq 0.05) = 2 \cdot P(Z \geq \frac{0.05}{0.0505}) \geq 2 \cdot P(Z \geq 1) = 0.3173.$$

Since p-value is bigger than 5%, we do not reject H_0.

(c) The null hypothesis is that the standard and the new antibiotic have the same effect on the infection. The alternative hypothesis is that the two antibiotics have different effects. The test statistic X is the number of patients in the second group who recovered from the infection. Under H_0, X follows $\mathrm{HG}(200, 175, 100)$ and $E(X) = 87.5$. The p-value is

$$p\text{-value} = P(X \geq 90) + P(X \leq 85) = \sum_{g=90}^{100} \frac{\binom{175}{g}\binom{25}{100-g}}{\binom{200}{100}} + \sum_{g=0}^{85} \frac{\binom{175}{g}\binom{25}{100-g}}{\binom{200}{100}}.$$

3. Let $X_1, X_2, \ldots, X_n$ be independent exponentially distributed random variables with mean $1/\lambda$.

(a) Show that $S = \min\{X_1, X_2, \ldots, X_n\}$ has an exponential distribution with parameter λn.

(b) Find an unbiased estimator of $1/\lambda$ that can be written in terms of S.

(c) Let T be the unbiased estimator found in part(b). Find the mean squared error (MSE) of T and compare it with the MSE of the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.*(Hint: use the result of part(a).)*

Solution

(a) We have

$$P(S \le x) = P(\min\{X_1, X_2, \ldots, X_n\} \le x) = 1 - P(\min\{X_1, X_2, \ldots, X_n\} > x) = 1 - P(X_1 > x)^n.$$

The survival function of exponential distribution is $e^{-\lambda x}$, so we get

$$P(S \le x) = 1 - e^{-n\lambda x}.$$

This is the CDF of the exponential distribution with mean $1/\lambda n$.

(b) Since $E(S) = \frac{1}{\lambda n}$, nS is an unbiased estimator of $1/\lambda$.

(c) Let $T = nS$. Since both T and $\bar{X}$ are unbiased, the mean squared error of T and $\bar{X}$ are given by their respective variances. $S \sim \text{Exp}(n\lambda)$, so $\text{Var}(T) = n^2\text{Var}(S) = \frac{n^2}{n^2\lambda^2} = \frac{1}{\lambda^2}$, while $\text{Var}(\bar{X}) = \frac{\text{Var}(X_1)}{n} = \frac{1}{n\lambda^2}$. So

$$\text{MSE}(T) > \text{MSE}(\bar{X}),$$

as long as $n > 1$.

4. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. random draws from Uniform(a, b) and our goal is to estimate the unknown parameters a and b based on n samples. Let $T_1 = \min\{X_1, X_2, \ldots, X_n\}$ and $T_2 = \max\{X_1, X_2, \ldots, X_n\}$.

(a) Show that $E(T_1 - a) = \frac{b-a}{n+1}$ and $E(b - T_2) = \frac{b-a}{n+1}$.

(b) Find the expectations of $T_1 + T_2$ and $T_1 - T_2$. *(Hint: combine the two equations given in part(a).)*

(c) Find unbiased estimators of a and b in terms of both T_1 and T_2. *(Hint: combine the expressions for $E(T_1 + T_2)$ and $\frac{n+1}{n-1}E(T_1 - T_2)$ using the result of part(b).)*

Solution

(a) Note that both $b - T_2$ and $T_1 - a$ represent length of a single Gap whose expectation is $\frac{b-a}{n+1}$, the total gap divided by the total number of Gaps. So we obtain

$$E(T_1 - a) = E(b - T_2) = \frac{b-a}{n+1}.$$

(b) If we add the two equations, we get

$$E(T_1 - T_2 - a + b) = 2 \cdot \frac{b-a}{n+1}.$$

Rearranging terms,

$$E(T_1 - T_2) = \frac{n-1}{n+1}(a - b).$$

If we subtract the two equations, we get

$$E(T_1 + T_2 - a - b) = 0.$$

Rearranging terms,

$$E(T_1 + T_2) = a + b.$$

(c) From part(b), we know

$$E(T_1 + T_2) = a + b \quad \text{and} \quad \frac{n+1}{n-1}E(T_1 - T_2) = a - b.$$

Adding the two equations and dividing by 2,

$$E(\frac{n}{n-1}T_1 - \frac{1}{n-1}T_2) = a.$$

Subtracting the two equations and dividng by 2,

$$E(-\frac{1}{n-1}T_1 + \frac{n}{n-1}T_2) = b.$$

5. Consider rolling a die for which the probability of rolling a six is $p \in (0, 1)$. David wants to test if the die is fair, i.e. there is no bias towards a six. David conducted 10 trials and obtained 1 six which was from the last trial. Then David left the lab.

(a) Danny continues David's work and performs a hypothesis testing for whether the die is biased towards a six or not. Formally, the null hypothesis is $H_0 : p = \frac{1}{6}$ and the alternative hypothesis is $H_A : p < \frac{1}{6}$. Danny uses the number of sixes out of David's 10 trials as the test statistic. Find the p-value.

(b) In fact, David actually stopped his experiment immediately after 1 six, because his boss had instructed him to do so. David looked at the p-value that Danny obtained in part(a) and explains to him that the p-value should be changed. Explain why Danny's p-value should be changed. Which test statistic should be used and what is the distribution of the test statistic under the null distribution? Find the correct p-value. *(Hint: the fact that the experiment was stopped immediately after the first six should remind you of the distribution familiar to you.)*

Solution

(a) Let X be the number of sixes out of 10 trials. Under H_0, we have $X \sim \text{Binom}(10, 1/6)$. Then

$$p\text{-value} = P(X \le 1) = P(X = 0) + P(X = 1) = \left(\frac{5}{6}\right)^{10} + 10\frac{1}{6}\left(\frac{5}{6}\right)^{9} \approx 0.4845.$$

(b) Let Y be the number of trials until first six. Then Y follows a Geometric distribution. Under H_0, $Y \sim \text{Geom}(1/6)$. Then

$$p\text{-value} = P(Y \ge 10) = 1 - P(Y \le 9) = \left(\frac{5}{6}\right)^{9} \approx 0.1938.$$

6. Let (X, Y) be a random pair and let $\widehat{Y} = \widehat{a}X + \widehat{b}$ be the linear regression of Y based on X. For notation, we use $r = r(X, Y)$ to denote the correlation between X and Y and use $D = Y - \widehat{Y}$ to denote the residual. We write $E(X) = \mu_X$, $\text{Var}(X) = \sigma_X$, $E(Y) = \mu_Y$, and $\text{Var}(Y) = \sigma_Y$.

(a) Show that $D = Y - \widehat{Y}$ and $\widehat{Y}$ are uncorrelated, i.e. $r(D, \widehat{Y}) = 0$.

(b) Show that $\text{Var}(Y)$ can be decomposed into $\text{Var}(Y) = \text{Var}(\widehat{Y}) + \text{Var}(D)$.

Solution

(a) We know from Ch11.5 of the textbook that $r(D, X) = 0$. Since $\widehat{Y}$ is a linear function of X, $r(D, \widehat{Y}) = 0$.

(b) We know that

$$\text{Var}(D) = (1 - r^2)\sigma_Y^2,$$

and

$$\text{Var}(\widehat{Y}) = \text{Var}(\widehat{a}X + \widehat{b}) = \widehat{a}^2\sigma_X^2 = r^2\sigma_Y^2.$$

Adding the two, we get $\text{Var}(\widehat{Y}) + \text{Var}(D) = \text{Var}(Y)$.

7. For each part (a) - (e) of this question, write down the numbers of ALL statements that are true. No credit will be given for partially correct answers (i.e. if a false statement is selected, or if a true statement is omitted.)

(a) Let $(5, 10)$ be an instance of 95% confidence interval for a parameter θ. Which of the following statement is true:

  i The chance that the interval $(5, 10)$ contains θ is 95%.
  ii The chance that the interval $(5, 10)$ contains θ is 0% or 100%.
  iii None of the above.

(b) Let T_1 and T_2 be two different estimators for a parameter θ. Which of the following statement is true:

  i If T_1 and T_2 are both unbiased estimators of θ, then $\frac{T_1+T_2}{2}$ is also an unbiased estimator.
  ii If T_1 is an unbiased estimator but T_2 is biased, then $\text{MSE}(T_1)$ is smaller than $\text{MSE}(T_2)$.
  iii If T_1 and T_2 are both unbiased estimators of θ and $\text{Var}(T_1) < \text{Var}(T_2)$, then $\text{MSE}(T_1)$ is smaller than $\text{MSE}(T_2)$.
  iv None of the above.

(c) Let X be a continuous random variable with density f and CDF F. Which of the following statement is true:

  i f can be derived from F.
  ii F can be derived from f.
  iii For any x, $P(X = x) = 0$.
  iv $\int_{-\infty}^{\infty} F(x)dx = 1$.

(d) Let $X_1, X_2, \ldots, X_n$ be i.i.d. with mean μ and SD σ. Which of the following statement is true:

  i If n is large enough, $S = X_1 + X_2 + \cdots + X_n$ approximately follows a normal distribution.
  ii If $X_1, X_2, \ldots, X_n$ follow normal distribution, $S = X_1 + X_2 + \cdots + X_n$ exactly follows a normal distribution.
  iii As n increases, $S = X_1 + X_2 + \cdots + X_n$ is more concentrated around its mean $n\mu$, i.e. for any $\epsilon > 0$, $P(|S - n\mu| < \epsilon) \to 1$ as $n \to \infty$.
  iv None of the above.

(e) Let X have a Uniform$(0, 1)$. What is the distribution of $Y = -2 \log X$:

  i Exponential distribution.
  ii Normal distribution.
  iii Uniform distribution.
  iv None of the above.

Solution

(a) Answer: ii.

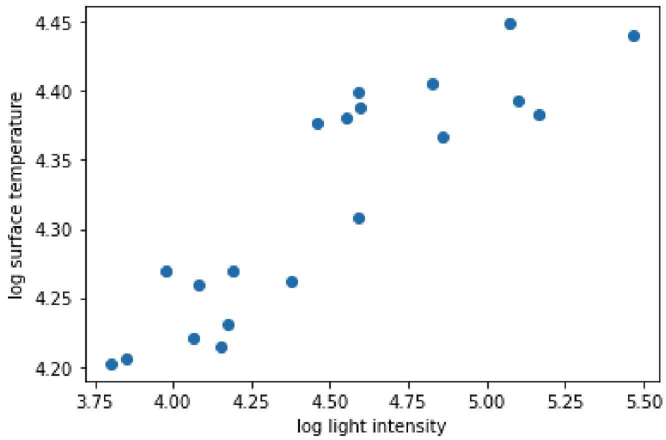
(b) Answer: i, iii.

(c) Answer: i, ii, iii.

(d) Answer: i, ii.

(e) Answer: 1.

8. A study was conducted to determine whether a linear relationship exists between log surface temperature (response) and log light intensity (predictor variable) for some nearby stars. Below is a scatter plot based on a random sample of 20 stars.



(a) From the scatter plot, discuss if the assumptions of the Simple Linear Regression Model are satisfied.

Now we run the linear regression on this data and the summary of the Python regression output is given below.

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.6032	0.087	41.427	0.000	3.420	3.786
x	0.1597			0.000		

We additionally know that the sample mean and SD of log light intensity are 4.50 and 0.46, the sample mean and SD of log surface temperature are 4.32 and 0.082, and the SD of the residuals is 0.0039.

(b) What is the predicted log surface temperature if the log light intensity is 5?

(c) Find the standard error of $\widehat{\beta}_1$, SE($\widehat{\beta}_1$).

(d) Find the 95% confidence interval for β_1. You may use one of the following percentile outputs from python:

```
stats.t.ppf(0.975, df=18) = 2.1
stats.t.ppf(0.950, df=18) = 1.73
stats.t.ppf(0.975, df=20) = 2.09
stats.t.ppf(0.950, df=20) = 1.72
```

(e) Find the sample correlation between log light intensity and log surface temperature. The sample correlation between $(x_1, x_2, \ldots, x_n)$ and $(Y_1, Y_2, \ldots, Y_n)$ is defined as

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

Solution

(a) The linear relationship between the two variables holds. The constant variance assumption also holds.

(b) We have $\widehat{\beta}_0 + \widehat{\beta}_1 \cdot 5 = 4.4017$.

(c) We have $\widehat{\sigma} = 0.0039$. So

$$\mathrm{SE}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{0.0039}{\sqrt{20} \cdot 0.046} \approx 0.01896.$$

(d) The 95% CI is $\widehat{\beta}_1 \pm 2.1 \cdot SE = (0.119, 0.2)$.

(e) The correlation between log light intensity and log surface temperature is given by

$$r = \widehat{\beta}_1 \cdot \frac{\widehat{\sigma}_X}{\widehat{\sigma}_Y} = 0.89$$