Final Exam Solutions

Introduction to Statistics(통계의 이해 영어강의) **2025 1st semester**

Section(교반): ______ Cadet Number(교번): __________ Name(성명): ______________ Score: ____________

- All solutions must include a detailed step-by-step explanation.
- If an answer has more than four decimal places, round to the **fourth decimal place**.

1. A city tested a beta version of a new public transportation app by surveying a random sample of $n = \mathbf{850}$ commuters. Among them, $x = \mathbf{137}$ reported being "very satisfied" in the survey. Answer the following questions. [13 points]

(1) Construct an approximate 95% confidence interval for the proportion of users p who reported being "very satisfied." You do not need to check the conditions for the Central Limit Theorem.

Solution:

We are given:

$$n = 850, \quad \hat{p} = \frac{137}{850} \approx 0.1612.$$

The 95% confidence interval is calculated as:

$$\hat{p} \pm z_{0.025} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.1612 \pm 1.96 \cdot \sqrt{\frac{0.1612(1-0.1612)}{850}} \approx 0.1612 \pm 0.0247.$$

Thus, the 95% confidence interval is approximately:

$$\boxed{(0.1365,\ 0.1859)}$$

(2) Choose the most appropriate word to complete each sentence:

(a) As the sample size increases, the standard error of the sample proportion $\hat{p}$ becomes (larger / smaller).

(b) As the sample size increases, the length of the confidence interval becomes (larger / smaller).

(c) The length of a 99% confidence interval is (larger / smaller) than that of a 95% confidence interval.

Solution: smaller / smaller / larger.

Reference Table

Satterthwaite's df : $\psi = \frac{(s_A^2/n_A + s_B^2/n_B)^2}{(s_A^2/n_A)^2/(n_A-1) + (s_B^2/n_B)^2/(n_B-1)}$, $S_p^2 = \frac{(n_A-1)S_A^2 + (n_B-1)S_B^2}{n_A+n_B-2}$			
$z_{0.005} = 2.5758$	$z_{0.01} = 2.3263$	$z_{0.025} = 1.9600$	$z_{0.05} = 1.6449$
$t_{0.005}(10) = 3.1693$	$t_{0.01}(10) = 2.7638$	$t_{0.025}(10) = 2.2281$	$t_{0.05}(10) = 1.8125$
pt(2.8, df = 8, lower.tail = FALSE) = 0.0116		pt(2.8, df = 9, lower.tail = FALSE) = 0.0104	
pt(1.5556, df = 8, lower.tail = FALSE) = 0.0792		pt(1.5556, df = 9, lower.tail = FALSE) = 0.0771	
pt(-0.2654, df = 50, lower.tail = TRUE) = 0.3959		pt(-1.3613, df = 50, lower.tail = TRUE) = 0.0898	
pt(4.1458, 10, lower.tail = FALSE) = 0.0010		pt(4.1458, 12, lower.tail = FALSE) = 0.0007	
pt(0.7951, 10, lower.tail = FALSE) = 0.2225		pt(0.7951, 12, lower.tail = FALSE) = 0.2210	
pt(-2.327, 13, lower.tail = TRUE) = 0.0184		pt(25.012, 13, lower.tail = FALSE) = 1.1120e-12	

Section(교반): _____ Cadet Number(교번): __________ Name(성명): ______________ Score: ____________

2. A company developed an 8-week diet program and claims that it reduces men's body fat. To evaluate this claim, the company measured the body fat of **9** male participants before and after completing the program. Let μ_{before} and μ_{after} represent the population mean body fat before and after the program, respectively, and let $\mu_{\text{diff}} = \mu_{\text{before}} - \mu_{\text{after}}$ denote the mean reduction. We conduct **a paired t-test** at the significance level $\alpha = 0.05$ to test whether the program effectively reduces body fat on average. [12 points]

(1) State the null and alternative hypotheses. (Use a **one-sided** test.)

Solution:

$$H_0 : \mu_{\text{diff}} = 0, \qquad H_A : \mu_{\text{diff}} > 0$$

(2) Find the test statistic and its null distribution.

Solution:

Under H_0, the test statistic is given by

$$T = \frac{\bar{X}_{\text{diff}} - 0}{S_{\text{diff}}/\sqrt{n_{\text{diff}}}} = \frac{\bar{X}_{\text{diff}}}{S_{\text{diff}}/3} \sim t(8),$$

where $\bar{X}_{\text{diff}}$ is the sample mean of the differences, S_{diff} is the sample standard deviation of the differences, and $n_{\text{diff}} = 9$ is the number of paired observations.

(3) Using the R code and R output provided, write down the p-value and complete the hypothesis test. **State the conclusion in the context of the data.**

```
before <- c(14, 15, 15, 17, 16, 16, 17, 15, 16)
after  <- c(12, 14, 12, 15, 14, 17, 13, 13, 17)
t.test(before, after, alternative = "greater", paired = TRUE)
```

```
        Paired t-test

data:  before and after
t = 2.8, df = [    ], p-value = [    ]
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
 0.5224733       Inf
sample estimates:
mean difference
       1.555556
```

Solution:

$$p\text{-value} = P(T > 2.8) \approx \boxed{0.0116},$$

where $T \sim t(8)$. Since the p-value is less than $\alpha = 0.05$, we reject the null hypothesis.

Conclusion: The data provide strong evidence that the 8-week diet program reduces men's body fat on average at the significance level $\alpha = 0.05$.

Final Exam Solutions

Introduction to Statistics(통계의 이해 영어강의) **2025 1st semester**

Section(교반): _____ Cadet Number(교번): __________ Name(성명): ______________ Score: ____________

3. The U.S. Environmental Protection Agency (EPA) collects fuel economy data annually. Highway fuel efficiency (measured in miles per gallon, MPG) was recorded for **22** randomly selected cars with automatic transmission and **30** randomly selected cars with manual transmission. Conduct a **two-sample t-test** to determine whether the population mean highway mileage for cars with automatic transmission (μ_A) differs from that of cars with manual transmission (μ_M). **Assume that the variances of highway fuel efficiency are equal for both transmission types.** [15 points]

(1) State the null and alternative hypotheses.

Solution:

$$H_0 : \mu_A - \mu_M = 0, \qquad H_A : \mu_A - \mu_M \neq 0$$

(2) Find the test statistic and its null distribution.

Solution:

Since we assume equal variances, we use the pooled standard deviation. Under H_0, the test statistic is:

$$T = \frac{\bar{X}_A - \bar{X}_M}{S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_M}}} = \frac{\bar{X}_A - \bar{X}_M}{S_p \cdot \sqrt{\frac{1}{22} + \frac{1}{30}}} \sim t(df)$$

where S_p is the pooled standard deviation and $df = n_A + n_M - 2 = 50$.

(3) Below are summary statistics on highway fuel efficiency (MPG) from random samples of cars with automatic and manual transmissions. Compute the observed test statistic.

Group	Mean	SD	n
Automatic	25.92	5.29	22
Manual	27.88	5.01	30

Solution:

First, calculate the pooled standard deviation:

$$s_p = \sqrt{\frac{(21)(5.29^2) + (29)(5.01^2)}{50}} \approx \sqrt{26.3114} \approx 5.1295$$

Then, compute the test statistic:

$$\frac{25.92 - 27.88}{5.1295 \cdot \sqrt{\frac{1}{22} + \frac{1}{30}}} \approx \boxed{-1.3613}$$

(4) Report the p-value and complete the hypothesis test by choosing the most appropriate word in the sentence below:

p-value = ____________________

Since the p-value is (less / greater) than $\alpha = 0.05$, we (reject / do not reject) the null hypothesis. There (is / is not) sufficient statistical evidence of a difference in average highway fuel efficiency between cars with automatic and manual transmissions at the significance level $\alpha = 0.05$.

Solution:

P-value is $2P(T < -1.3613) = 2 \times 0.0898 = 0.1796$, where $T \sim t(50)$.

greater / do not reject / is not

Final Exam Solutions

Introduction to Statistics(통계의 이해 영어강의) **2025 1st semester**

Section(교반): _____ Cadet Number(교번): __________ Name(성명): ______________ Score: ____________

4. A survey on cadet life satisfaction was conducted during a training session at a military academy. A total of **268** responses were grouped by company (**8** companies in total). We conduct **a one-way ANOVA** to test whether the average cadet life satisfaction score differs across companies at the significance level $\alpha = 0.05$. Assume that all conditions for conducting ANOVA are satisfied. [15 points]

Company	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**All**
Mean	3.80	3.45	3.68	3.22	3.91	3.34	3.70	3.51	3.58
SD	0.25	0.28	0.22	0.27	0.20	0.30	0.24	0.29	0.27
n	33	35	30	32	34	35	32	37	268

(1) Let $\mu_1, \mu_2, \ldots, \mu_8$ represent the true mean cadet life satisfaction for each of the 8 companies. State the null and alternative hypothesis.

Solution:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_8, \qquad H_A : \text{not } H_0.$$

(2) Below is part of ANOVA output. Fill in the empty cells.

Source	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Company		12.9853			< 0.001
Residuals		17.5223			
Total	267	30.5076			

Solution:

Degrees of freedom for Company = $k - 1 = 8 - 1 = 7$

Degrees of freedom for Residuals = $n - k = 268 - 8 = 260$

Mean square for Company: $12.9853/7 \approx 1.8550$

Mean square for Residuals: $17.5223/260 \approx 0.0674$

F-statistic: $1.8550/0.0674 \approx 27.5223$

(3) Do you reject the null hypothesis? **State the conclusion in the context of the data.**

Solution: Since the p-value is less than $\alpha = 0.05$, we reject the null hypothesis. There is strong statistical evidence that cadet life satisfaction differs across the eight companies at the significance level $\alpha = 0.05$.

Final Exam Solutions

Introduction to Statistics(통계의 이해 영어강의) **2025 1st semester**

Section(교반): _____ Cadet Number(교번): __________ Name(성명): ______________ Score: ____________

5. A university studied **12** pairs of twins by giving each twin a psychological test that measures aggressiveness toward their sibling. We conduct **a hypothesis test for the population correlation coefficient** to see if there is a **positive** relationship between the older twin's and younger twin's aggressiveness scores $(H_0 : \rho = 0,\, H_A : \rho > 0)$ at the significance level $\alpha = 0.05$. Assume that the scores for older twins (x) and younger twins (y) follow a bivariate normal distribution. [15 points]

Twin Pair i	1	2	3	4	5	6	7	8	9	10	11	12
Older Twin (x_i)	86	71	77	68	91	71	76	90	69	70	87	80
Younger Twin (y_i)	88	77	76	64	92	65	86	90	65	80	80	71

$$S_{xx} = \sum_{i=1}^{12}(x_i - \bar{x})^2 = 810, \quad S_{yy} = \sum_{i=1}^{12}(y_i - \bar{y})^2 = 1188, \quad S_{xy} = \sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y}) = 780$$

(1) Find the null distribution of the test statistic: $\dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}}$,

where R is the sample correlation coefficient and n is the sample size.

Solution: Under H_0, the test statistic is

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{R\sqrt{10}}{\sqrt{1-R^2}} \sim t(10)$$

(2) Compute the sample correlation coefficient r.

Solution:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{780}{\sqrt{810 \cdot 1188}} \approx \boxed{0.7951}$$

(3) Compute the observed test statistic.

Solution:

$$\frac{0.7951 \cdot \sqrt{10}}{\sqrt{1 - 0.7951^2}} \approx \boxed{4.1458}$$

(4) Report the p-value and complete the hypothesis test by choosing the most appropriate word in the sentence below:

p-value = ____________________

Since the p-value is (less / greater) than $\alpha = 0.05$, we (reject / do not reject) the null hypothesis. There (is / is not) statistically significant evidence of a positive relationship between the aggressiveness scores of older and younger twins at the significance level $\alpha = 0.05$.

Solution:

P-value is $P(T > 4.1458) = 0.0010$, where $T \sim t(50)$.

less / reject / is

Final Exam Solutions

Introduction to Statistics(통계의 이해 영어강의) **2025 1st semester**

Section(교반): _____ Cadet Number(교번): __________ Name(성명): ______________ Score: ____________

6. A research team investigates the relationship between newborns' height (x, in cm) and weight (y, in kg) in a city using the linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ independently.}$$

The team collects data from **16** randomly selected newborns. Answer the following questions. [10 points]

Infant No. i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Height (x_i, cm)	46.2	47.5	48.1	48.3	48.8	49.0	49.5	49.9	50.2	50.6	50.8	51.0	51.3	51.6	52.0	52.4
Weight (y_i, kg)	3.42	3.48	3.67	3.82	3.62	3.64	3.98	3.90	3.75	3.95	3.83	3.85	3.99	3.71	3.78	4.41

$$\bar{x} = \frac{1}{16}\sum_{i=1}^{16} x_i = 49.825, \quad \bar{y} = \frac{1}{16}\sum_{i=1}^{16} y_i = 3.8$$

$$S_{xx} = \sum_{i=1}^{16}(x_i - \bar{x})^2 = 45.45, \quad S_{yy} = \sum_{i=1}^{16}(y_i - \bar{y})^2 = 0.8096, \quad S_{xy} = \sum_{i=1}^{16}(x_i - \bar{x})(y_i - \bar{y}) = 4.414$$

(1) Compute the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

Solution:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{4.414}{45.45} = \boxed{0.0971}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3.8 - 0.0971 \cdot 49.825 = 3.8 - 4.838 = \boxed{-1.038}$$

(2) Define a new explanatory variable $x_i' = 10x_i$, converting height from centimeters to millimeters. Using x_i' as the predictor variable, we fit a new regression model:

$$y_i = \beta_0' + \beta_1' x_i' + \varepsilon_i', \quad \varepsilon_i' \sim N(0, \sigma^2) \text{ independently.}$$

Find the least square estimates for the new regression coefficients, $\hat{\beta}_0'$ and $\hat{\beta}_1'$.

Solution:

$$S_{x'x'} = 100 \cdot S_{xx} = 100 \cdot 45.45 = 4545, \quad S_{x'y} = 10 \cdot S_{xy} = 10 \cdot 4.414 = 44.14$$

$$\hat{\beta}_1' = \frac{S_{x'y}}{S_{x'x'}} = \frac{44.14}{4545} = \boxed{0.0097}$$

$$\bar{x}' = 10 \cdot \bar{x} = 498.25 \quad \Rightarrow \quad \hat{\beta}_0' = \bar{y} - \hat{\beta}_1' \cdot \bar{x}' = 3.8 - 0.0097 \cdot 498.25 = 3.8 - 4.838 = \boxed{-1.038}$$

Final Exam Solutions

Introduction to Statistics(통계의 이해 영어강의) **2025 1st semester**

Section(교반): ______ Cadet Number(교번): ___________ Name(성명): _______________ Score: _____________

7. A research team is studying the relationship between smoking status (x, where $x = 1$ if the person is a smoker and $x = 0$ otherwise) and lung capacity (y, measured in 1000ml) in residents of a city. They fit a simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ independently.}$$

A random sample of **5** smokers and **10** non-smokers was collected. Use the information below to answer the following questions. [20 points]

Subject No. i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Smoke (x_i)	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
Lung capacity ($y_i, \times 1000$ ml)	30	31	38	30	35	32	48	42	35	40	36	41	47	36	34

R Code and Output:

```
smoke <- c(1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
breath <- c(30, 31, 38, 30, 35, 32, 48, 42, 35, 40, 36, 41, 47, 36, 34)
summary(lm(breath ~ smoke))
```

```
Call:
lm(formula = breath ~ smoke)

Residuals:
   Min     1Q Median     3Q    Max
 -7.10  -3.10  -1.80   2.55   8.90

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    39.100      1.563   25.012  2.22e-12 ***
smoke          -6.300      2.708   -2.327   0.0368 *
---
Signif. codes: 0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

Residual standard error: 4.944 on 13 degrees of freedom
Multiple R-squared: 0.294, Adjusted R-squared: 0.2397
F-statistic: 5.414 on 1 and 13 DF, p-value: 0.03679
```

Final Exam Solutions

Introduction to Statistics(통계의 이해 영어강의) **2025 1st semester**

Section(교반): _____ Cadet Number(교번): __________ Name(성명): ______________ Score: ____________

(1) Estimate the average lung capacity for smokers ($x = 1$) and the average lung capacity for non-smokers ($x = 0$).

Solution:

From the R output, $\hat{\beta}_0 = 39.1$, $\hat{\beta}_1 = -6.3$.

If $x = 1$, then $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 = 39.1 - 6.3 = \boxed{32.8}$

If $x = 0$, then $\hat{y} = \hat{\beta}_0 = \boxed{39.1}$

(2) Compute the sample means $\bar{x} = \frac{1}{15}\sum_{i=1}^{15} x_i$ and $\bar{y} = \frac{1}{15}\sum_{i=1}^{15} y_i$. (Hint: use the identity $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$)

Solution:

$$\bar{x} = \frac{5 \cdot 1 + 10 \cdot 0}{15} = \frac{5}{15} = \boxed{0.3333}$$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x} = 39.1 + (-6.3)(0.3333) = \boxed{37.0002}$$

(3) Compute the observed sample correlation coefficient r between x_i and y_i for $i = 1, \cdots, 15$

Solution:

The coefficient of determination is $r^2 = 0.294$ from the R output.

Since $\hat{\beta}_1 = \frac{s_y}{s_x} r = -6.3 < 0$, $r = -\sqrt{0.294} = \boxed{-0.5422}$.

(4) We test the following hypothesis about the slope parameter β_1:

$$H_0 : \beta_1 = 0, \quad H_A : \beta_1 < 0$$

(a) Find the null distribution of the test statistic: $\frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$,

where $\hat{\beta}_1$ is the least square estimate of β_1, $\hat{\sigma}^2 = \frac{1}{15-2}\sum_{i=1}^{15}(y_i - \hat{y}_i)^2$, and $S_{xx} = \sum_{i=1}^{15}(x_i - \bar{x})^2$.

Solution:

$$T = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t(13),$$

since the degrees of freedom is $15 - 2 = 13$.

(b) Write down the p-value for this test. Using the significance level $\alpha = 0.05$, do you reject the null hypothesis?

Solution:

From the R output, the observe test statistic is -2.327 and p-value for the two-sided test is $2P(T < -2.327) = 0.0368$. Therefore, p-value for the one-sided test is $P(T < -2.327) = \boxed{0.0184}$.

Since the p-value is less than $\alpha = 0.05$, we reject the null hypothesis.