Individual Assignment

STAT5002 Introduction to Statistics

Semester 2 2025

Lecturer: Andi Han **Total Marks:** 80 **Due:** 11:59 pm Sunday 02 Nov 2025

There are **four questions** in this assignment with a total of **80** points.

Submission Format You must submit in the following format:

- You should submit **a single combined PDF** file that includes your written answers along with your code (and relevant outputs, if necessary). You may include code as screenshots but make sure they are clear and understandable.
- You should only submit **one combined file** with answers to each of the questions clearly labeled and structured.
- Markers will only mark the contents in the PDF and please do not include external links in the file.

Your submitted file must include your SID. To comply with anonymous marking policies, do not include your name anywhere in your assignment.

Instructions Below are some instructions you should follow. Failure to comply with instructions could risk mark deduction.

- You should structure your answers with proper mathematical notations, include necessary working details, assumptions, justifications for your calculations and interpretations of your results.
- You may use the output of `t.test()` and `chisq.test()` to check your results. However, you must show your work without relying on these functions unless specified in the question.
- Round your *final answers* to **two decimal places** (if necessary).

Please review your submission carefully. You may revise and resubmit your work until the due date.

Q1. (25 marks) **Unfair and Unknown Dice**

You have two six-sided dice, Die A and Die B. *Die A* is small-value biased, i.e., each of the small-value faces (1, 2, 3) has twice the probability of each of the large-value faces (4, 5, 6). The true distribution of *Die B* is unknown.

(a) You roll the *Die A* independently for 81 times and let S be the number of rolls (out of 81 rolls) with value at least 3 (i.e., 3, 4, 5, or 6). What is the expected value and standard error of S? [**5 marks**]

(b) Compute the 97% prediction interval for S in Part (a) and interpret the result. Use 5000 simulations to verify your derivation. Include the `R` code used for simulation and explain the result of simulation. [**7 marks**]

(c) You roll the *Die B* for 99 times and observe 24 of rolls with odd values (e.g., 1,3,5). Let p be the probability that the Die B outputs odd values. What is the smallest p that is consistent with your observed data, under 95% confidence level? (If you use `R`, make sure you include the `R` code and output). [**3 marks**]

(d) In Part (c), you record the observed values of the 99 rolls as follows

Value	1	2	3	4	5	6
Frequency	10	27	5	33	9	15

Based on the frequency table, you wish to know whether Die B has the same distribution as Die A. Follow the HATPC framework to perform an appropriate hypothesis test at 1% significance level (To answer, first identify the test you are going to perform). [**10 marks**]

Q2. (30 marks) **Caffeine Effect**

A sports-science group is testing whether a new 200 mg caffeine gel affects sprinters' start performance. 16 athletes perform two timed starts on the same day.

- **PRE**: baseline, no caffeine
- **POST**: 15 mins after ingesting the gel

The researchers record the reaction times in milliseconds (ms). After the experiment, each athlete self-reports whether they felt, either "more alert" or "not more alert". The recorded results are included in Table 1.

The researchers would like to know whether the Caffeine gel can *reduce* the start reaction time.

Athlete	**PRE (ms)**	**POST (ms)**	**Self-report**
1	171	160	Alert
2	162	155	Alert
3	164	158	Not alert
4	169	161	Alert
5	173	165	Alert
6	168	170	Not alert
7	158	151	Alert
8	166	157	Alert
9	176	170	Alert
10	161	155	Not alert
11	170	165	Alert
12	159	157	Not alert
13	167	160	Alert
14	163	165	Not alert
15	172	166	Alert
16	160	159	Not alert

Table 1: Synthetic reaction-time data for 16 athletes. "PRE" = baseline, "POST" = 15 mins after 200 mg caffeine gel.

a) Introduce appropriate parameters and state the null and alternative hypotheses. [**4 marks**]

b) Select and justify a suitable statistical test (including type of the test and indicating whether it is a two-sided or one-sided test). [**4 marks**]

c) What is the key assumption for the above test? Use appropriate graphical summaries to assess whether the test assumptions hold. [**4 marks**]

d) Compute the observed test statistic and associated p-value, assuming the assumption holds in part (c). Specify distribution of the test statistics and the rejection region at 5% significance level. [**6 marks**]

e) Draw your conclusion based on the calculated p-value under 5% significance level. [**4 marks**]

f) Perform a bootstrap simulation (10 000 resamples) for the test statistic and plot its histogram, and compare with the theoretical distribution. (You need to include your `R` code here) [**4 marks**]

g) Estimate the p-value from the simulated distribution and draw your conclusion under 5% significance level. [**4 marks**]

Here is the `R` code for the records that you can use.

```
pre_ms  <- c(171, 162, 164, 169, 173, 168, 158, 166,
             176, 161, 170, 159, 167, 163, 172, 160)

post_ms <- c(160, 155, 158, 161, 165, 170, 151, 157,
             170, 155, 165, 157, 160, 165, 166, 159)
```

Q3. (10 marks) **Caffeine Effect and Self-report**

Based on the same scenario and records in Q2, researchers are now interested in whether the caffeine effect differs depending on sprinters' self-reported alertness. Follow the HATPC framework to perform a **classical two-sample T-test** examining whether the average caffeine effect (computed using POST–PRE) differs between sprinters who felt alert and those who felt not alert at 5% significance level. (Carry out the classical two-sample T-test without simulation even when you think the assumptions may not hold).

Here is the `R` code that separates the records into alert group and not-alert group.

```
pre_alert <- c(171, 162, 169, 173, 158, 166, 176, 170, 167, 172)
post_alert <- c(160, 155, 161, 165, 151, 157, 170, 165, 160, 166)

pre_notalert <- c(164, 168, 161, 159, 163, 160)
post_notalert <- c(158, 170, 155, 157, 165, 159)
```

Q4. (15 marks) **Advertising and Sales**

A coffee-chain marketing team believes that increasing its weekly social-media advertising budget boosts the number of drinks sold in the same week. To quantify this effect, the team runs a 20-week pilot.

Each week i, the team measures advertising spend in thousands of dollars (x_i) and drinks sold in thousands of cups (y_i).

You may use the following `R` code.

```
# Budget in thousands of dollars
x <- c(2.0, 3.5, 4.0, 5.0, 6.5, 7.0, 8.0, 9.5, 10.0, 11.0,
       12.5, 13.0, 14.5, 15.0, 16.0, 17.5, 18.0, 19.5, 20.5, 22.0)
```

Week	**Budget** x **(k\$)**	**Sales** y **(k cups)**
1	2.0	17.0
2	3.5	23.0
3	4.0	23.2
4	5.0	28.0
5	6.5	30.8
6	7.0	33.3
7	8.0	34.9
8	9.5	41.7
9	10.0	41.6
10	11.0	46.8
11	12.5	47.7
12	13.0	50.5
13	14.5	53.1
14	15.0	52.4
15	16.0	55.0
16	17.5	56.1
17	18.0	55.5
18	19.5	52.8
19	20.5	51.9
20	22.0	50.0

Table 2: Weekly advertising spend and corresponding drink sales for a 20-week pilot campaign.

```
# Sales in thousands of cups
y <- c(17.0, 23.0, 23.2, 28.0, 30.8, 33.3, 34.9, 41.7, 41.6, 46.8,
       47.7, 50.5, 53.1, 52.4, 55.0, 56.1, 55.5, 52.8, 51.9, 50.0)
```

a) Use `R` to obtain an estimated linear regression model and interpret all the coefficients (including the intercept). [**4 marks**]

b) Use appropriate tools to examine whether the assumptions (i.e., normality of residuals, linearity, homoscedasticity) are satisfied for linear regression. [**8 marks**]

c) Suggest and briefly justify at least two other variables that could also affect weekly sales. [**3 marks**]