DA_HW1

Howard

2025-02-24

Problem 1

When we play “Poker” with a deck (each player gets 5 cards out 52 regardless of the jokers), why the “Full-house” is more valuable than the “Flush”?

Ans:

- The probability of 5 cards in Full-house is : $\frac{C_1^{13} \times C_3^4 \times C_1^{12} \times C_2^4}{C_5^{52}} \approx 0.001441$.
- The probability of 5 cards in Flash is : $\frac{C_1^4 \times C_5^{13}}{C_5^{52}} \approx 0.001981$. So the probability of “Full-house” is lower than “Flash”,which means the “Full-house” is more valuable than the “Flash”.

Problem 2

Suppose two teams, E and W, are playing the NBA finals (a series of 7 games), where the series is done when E or W wins four matches firstly. If each match is independently won by team E with probability p and by team W with probability $1 - p$. Find the expected number of matches that are played, and evaluate this expected number when $p = 1/2$.

Ans:

- The probability of team E win is p,lose is $1 - p$.
- The probability of team W win is $1 - p$,lose is p.

So the expected number of matches is

$$E(N) = 4\left[p^4 + (1-p)^4\right] + 5\left[4p^4(1-p) + 4(1-p)^4 p\right] + 6\left[10p^4(1-p)^2 + 10(1-p)^4 p^2\right] + 7\left[20p^4(1-p)^3 + 20(1-p)^4 p^3\right.$$

And when $p = \frac{1}{2}$, the expected number of matches = 5.81.

Problem 3

A fountain show starts every 80 minutes, you arrive at the place at random and decide to wait for 20 minutes, what’s the probability you will witness the show?

Ans: The probability of witness the show is

$$P(T) \leq 20 = \frac{20}{80} = \frac{1}{4} = 0.25$$

So, the probability that will witness the show is 25%.

Problem 4

At the NTU post office, a pair of clerks work at distinct levels of efficiency: clerk 1’s service time follows an exponential distribution characterized by the rate μ_1, whereas clerk 2’s service time conforms to an alternative exponential distribution with rate μ_2. On a particular day, John arrived at the postal office and started receiving service from clerk 1 at precisely 8:00.

a. Mary enters at 8:10, what is the probability she sees John is still being served by clerk 1?

Ans: John starts at 8:00. Mary arrives at 8:10, so she will see John still being served if his service time exceeds 10 minutes. That probability is

$$P(T > 10) = e^{-\mu_1 * 10}$$

b. Since John is still in service, Mary goes to clerk 2 to be served. What is the probability that Mary finishes her service before John does?

Ans:

- Let X $\sim$ Exp(μ_1) be John’s remaining service time.
- Let Y $\sim$ Exp(μ_2) be Mary’s service time at clerk 2. *Since X & Y are independent, the probability that Mary finishes before John is given by

$$P(Y < X) = \frac{\mu_2}{\mu_1 + \mu_2}$$

Problem 5

John lives in city A and goes to work every morning by taking one train and then connecting to a local bus in city B. To avoid being late for work, he must arrive no later than 8h30. John always takes the train at 8h00. The trajectory between A and B takes exactly 10 minutes. According to the long-term observation, the train is of a delay probability distribution as the table below:

Ans:

- Let John’s arrival time at city B is 8:00 + T, where

$$T = 10 + d$$

  ,where d has the distribution as the table.

Bus Schedule:

- The bus departure time X follows a normal distribution: $X \sim N(20, 2^2)$.
- To be on time (arriving by 8:30 after a fixed 10-minute bus ride), John must catch a bus that departs no later than 8:20.

On-Time Condition: John is on time if:

He catches the bus: $T X $,and The bus leaves by 8:20: $X \le 20$.

Case-by-Case Analysis for Viable Delays

$d = 4 \ (T = 14)$

$$P(14 \le X \le 20) \ = \ 0.5 \ - \ \Phi\left(\tfrac{14-20}{2}\right) \ = \ 0.5 \ - \ \Phi(-3) \ \approx \ 0.49865.$$

Weighted contribution:

$$\frac{1}{16} \times 0.49865.$$

$d = 6 \ (T = 16)$

$$P(16 \le X \le 20) \ = \ 0.5 \ - \ \Phi\left(\tfrac{16-20}{2}\right) \ = \ 0.5 \ - \ \Phi(-2) \ \approx \ 0.4772.$$

Weighted contribution:

$$\frac{1}{8} \times 0.4772.$$

$d = 8 \ (T = 18)$

$$P(18 \le X \le 20) \ = \ 0.5 \ - \ \Phi\left(\tfrac{18-20}{2}\right) \ = \ 0.5 \ - \ \Phi(-1) \ \approx \ 0.3413.$$

Weighted contribution:

$$\frac{1}{2} \times 0.3413.$$

$d = 10 \ (T = 20)$

$$P(20 \le X \le 20) \ = \ 0 \quad (\text{since } X \text{ is continuous}).$$

Weighted contribution:

$$\frac{1}{4} \times 0 = 0.$$

- $d = 12 \ (T = 22)$
  Not viable, since $T > 20$.

Overall Probability

$$P(\text{on time}) \ = \ \frac{1}{16} \times 0.49865 \ + \ \frac{1}{8} \times 0.4772 \ + \ \frac{1}{2} \times 0.3413 \ \approx \ 0.26147.$$

Thus, the probability that John is late is:

$$P(\text{late}) \ = \ 1 - 0.26147 \ \approx \ 0.73853 \quad (\text{about } 73.85\%).$$

Final Result

$$\boxed{P(\text{late}) \approx 73.85\%.}$$

Problem 6

(a) After a Positive X-ray Test

- **Prior Probability:**
  - $P(C) = 0.01$ (John has cancer)
  - $P(\neg C) = 0.99$ (John does not have cancer)
- **X-ray Test Characteristics:**
  - Sensitivity: $P(\text{Positive} \mid C) = 0.99$
  - Specificity: $P(\text{Negative} \mid \neg C) = 0.99$
  - False Positive Rate: $P(\text{Positive} \mid \neg C) = 0.01$
- **Bayes' Theorem:**

$$P(C \mid \text{Positive}) = \frac{P(\text{Positive} \mid C)\, P(C)}{P(\text{Positive} \mid C)\, P(C) + P(\text{Positive} \mid \neg C)\, P(\neg C)}$$

- **Calculation:**

$$P(C \mid \text{Positive}) = \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.01 \times 0.99} = \frac{0.0099}{0.0198} = 0.5.$$

- **Conclusion:**
  - The probability that John really has cancer after a positive X-ray is **50%**.

(b) After a Second Positive MRI Test

- **MRI Test Characteristics:**
  - Sensitivity: $P(\text{Positive} \mid C) = 0.999$
  - Specificity: $P(\text{Negative} \mid \neg C) = 0.999$
  - False Positive Rate: $P(\text{Positive} \mid \neg C) = 0.001$
- **Assumption:**
  - The two tests are independent given the disease status.
- **Joint Probabilities:**
  - If John has cancer:

$$P(\text{Both Positive} \mid C) = 0.99 \times 0.999 \approx 0.98901.$$

  - If John does not have cancer:

$$P(\text{Both Positive} \mid \neg C) = 0.01 \times 0.001 = 0.00001.$$

- **Total Probability of Both Tests Positive:**

$$P(\text{Both Positive}) = 0.01 \times 0.98901 + 0.99 \times 0.00001 \approx 0.0098901 + 0.0000099 \approx 0.0099.$$

- **Bayes' Theorem for Both Tests:**

$$P(C \mid \text{Both Positive}) = \frac{0.01 \times 0.98901}{0.0099} \approx \frac{0.0098901}{0.0099} \approx 0.999.$$

- **Conclusion:**
  - After the second positive MRI test, the probability that John has cancer is approximately **99.9%**.

Final Answers

- **(a)** $P(\text{Cancer} \mid \text{Positive X-ray}) \approx 50\%.$
- **(b)** $P(\text{Cancer} \mid \text{Positive X-ray and Positive MRI}) \approx 99.9\%.$

Problem 7

Ans:

- **(a)** There are 365 choices for the first students, 364 for the second,..., 341 for the 25th.

$$\frac{365 \times 364 \times \cdots 341}{365^{25}} \approx 43.21\%$$

- **(b)**

$$P(atleastonematch) = 1 - P(alldifferent) \approx 1 - 0.431 = 0.569$$

Problem 8

Consider the hypothesis test where the null hypothesis is

$$H_0 : p = 0.4,$$

and a random sample of $n = 600$ people is surveyed. The decision rule is to accept H_0 if the number of iPhone users X is between 216 and 264 (inclusive), and to reject H_0 otherwise.

Under H_0, the number of iPhone users follows a binomial distribution:

$$X \sim \text{Binomial}(600,\ 0.4).$$

The mean and standard deviation are

$$\mu = np = 600 \times 0.4 = 240, \quad \sigma = \sqrt{np(1-p)} = \sqrt{600 \times 0.4 \times 0.6} = 12.$$

Using the normal approximation with a continuity correction:

- **For the lower tail ($X < 216$):**

  Apply the continuity correction by considering $X \leq 215.5$. Then

$$Z = \frac{215.5 - 240}{12} \approx \frac{-24.5}{12} \approx -2.0417.$$

  So,

$$P(X < 216) \approx P\big(Z < -2.0417\big) \approx 0.0206.$$

- **For the upper tail ($X > 264$):**

  Apply the continuity correction by considering $X \geq 264.5$. Then

$$Z = \frac{264.5 - 240}{12} \approx \frac{24.5}{12} \approx 2.0417.$$

  So,

$$P(X > 264) \approx P\big(Z > 2.0417\big) \approx 0.0206.$$

- **Total Type I Error Probability:**

  The Type I error probability is the probability of rejecting H_0 when it is true:

$$\alpha = P(X < 216 \text{ or } X > 264) \approx 0.0206 + 0.0206 = 0.0412.$$

Thus, the Type I error probability for this test is approximately 4.12%.

Problem 9

- **Exponential Distribution Case:**
  - Let T be the lifetime of the lightbulb, with

$$T \sim \text{Exp}(\lambda),$$

    where the mean lifetime is 10 hours. Hence, $\lambda = \frac{1}{10}$.
  - The survival function for an exponential distribution is given by

$$P(T > t) = e^{-\lambda t}.$$

- For $t = 5$ hours, the probability that the bulb lasts at least 5 more hours is

$$P(T > 5) = e^{-\frac{5}{10}} = e^{-0.5} \approx 0.6065.$$

- Thus, if the lightbulb's lifetime is exponential, there is approximately a **60.65%** chance that Jack can work for 5 hours without the bulb burning out.

- **Non-Exponential Distribution Case:**
  - When the lifetime distribution is not exponential, the **memoryless property does not hold**.
  - This means that the conditional probability

$$P(T > t + s \mid T > t)$$

is not simply equal to $P(T > s)$ but depends on the elapsed time t and the specific form of the distribution.

Problem 10

Hypotheses

- **Null Hypothesis (H_0)**: The grades in Probability & Statistics (Prob) and Operations Research (OR) are independent.
- **Alternative Hypothesis (H_a)**: The grades in Prob and OR are related (i.e., not independent).

Expected Frequencies

The expected count for each cell is computed using:

$$E_{ij} = \frac{(\text{Row Total}_i)(\text{Column Total}_j)}{100}.$$

- For cell (OR A, Prob A):

$$E_{11} = \frac{45 \times 35}{100} = 15.75.$$

- For cell (OR A, Prob B):

$$E_{12} = \frac{45 \times 30}{100} = 13.5.$$

- For cell (OR A, Prob C):

$$E_{13} = \frac{45 \times 35}{100} = 15.75.$$

- For cell (OR B, Prob A):

$$E_{21} = \frac{25 \times 35}{100} = 8.75.$$

- For cell (OR B, Prob B):

$$E_{22} = \frac{25 \times 30}{100} = 7.5.$$

- For cell (OR B, Prob C):

$$E_{23} = \frac{25 \times 35}{100} = 8.75.$$

- For cell (OR C, Prob A):

$$E_{31} = \frac{30 \times 35}{100} = 10.5.$$

- For cell (OR C, Prob B):

$$E_{32} = \frac{30 \times 30}{100} = 9.$$

- For cell (OR C, Prob C):

$$E_{33} = \frac{30 \times 35}{100} = 10.5.$$

Chi-Square Test Statistic

The chi-square statistic is:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_{ij} are the observed counts.

Calculations for each cell:

- **Cell (OR A, Prob A)**:

$$\frac{(24 - 15.75)^2}{15.75} = \frac{(8.25)^2}{15.75} \approx 4.322.$$

- **Cell (OR A, Prob B)**:

$$\frac{(11 - 13.5)^2}{13.5} = \frac{(-2.5)^2}{13.5} \approx 0.463.$$

- **Cell (OR A, Prob C)**:

$$\frac{(10 - 15.75)^2}{15.75} = \frac{(-5.75)^2}{15.75} \approx 2.100.$$

- **Cell (OR B, Prob A)**:

$$\frac{(7 - 8.75)^2}{8.75} = \frac{(-1.75)^2}{8.75} \approx 0.350.$$

- **Cell (OR B, Prob B)**:

$$\frac{(13 - 7.5)^2}{7.5} = \frac{(5.5)^2}{7.5} \approx 4.033.$$

- **Cell (OR B, Prob C)**:

$$\frac{(5 - 8.75)^2}{8.75} = \frac{(-3.75)^2}{8.75} \approx 1.607.$$

- **Cell (OR C, Prob A)**:

$$\frac{(4 - 10.5)^2}{10.5} = \frac{(-6.5)^2}{10.5} \approx 4.024.$$

- **Cell (OR C, Prob B)**:

$$\frac{(6 - 9)^2}{9} = \frac{(-3)^2}{9} = 1.$$

- **Cell (OR C, Prob C)**:

$$\frac{(20 - 10.5)^2}{10.5} = \frac{(9.5)^2}{10.5} \approx 8.595.$$

Summing these contributions:

$$\chi^2 \approx 4.322 + 0.463 + 2.100 + 0.350 + 4.033 + 1.607 + 4.024 + 1 + 8.595 \approx 26.49.$$

Degrees of Freedom

$$\text{df} = (3 - 1) \times (3 - 1) = 2 \times 2 = 4.$$

Decision Rule

At $\alpha = 0.01$, the critical value of χ^2 with 4 degrees of freedom is approximately 13.277. Since

$$26.49 > 13.277,$$

we reject the null hypothesis H_0.

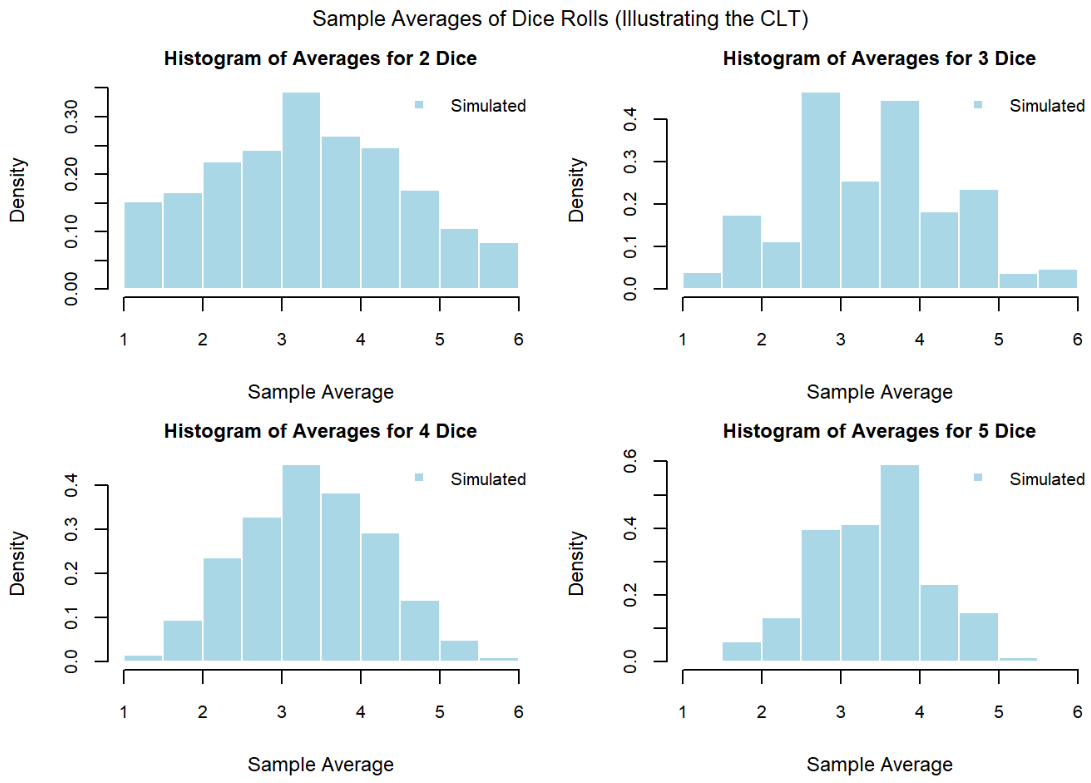
Conclusion

There is strong evidence at the 0.01 significance level to conclude that the grades in Probability & Statistics and Operations Research are related.

Problem 11

Program and simulate the averages of [2, 3, 4, 5] dice for 1000 times. Draw the four histograms for the sample averages of [2, 3, 4, 5] dice, respectively. *Reproduce the CLT results on p. 28 of slides DA01.

```
# ------------------------------------
# Simulation of sample averages of dice
# ------------------------------------

# Number of trials
num_trials <- 1000

# Dice sides
sides <- 6

# Theoretical mean and variance for one fair die
die_mean <- (sides + 1) / 2       # = 3.5
die_var  <- (sides^2 - 1) / 12    # = 35/12

# We'll simulate for n in {2, 3, 4, 5}
dice_counts <- c(2, 3, 4, 5)

# Set up a 2x2 plotting layout
par(mfrow = c(2, 2),     # 2 rows, 2 columns of plots
    oma = c(0, 0, 2, 0),
    mar = c(4, 4, 2, 1),
    cex.axis = 0.8,
    cex.lab = 0.9,
    cex.main = 0.9) # margins around each plot

for (n in dice_counts) {
  # 1. Generate the sample averages
  #    For each of the 1000 trials, roll n dice and compute the mean
  sample_averages <- replicate(num_trials, {
    mean(sample(1:sides, n, replace = TRUE))
  })

  # 2. Plot the histogram
  hist(sample_averages,
       breaks = 15,
       freq = FALSE,  # Use density (not counts)
       xlim = c(1, 6),
       main = paste("Histogram of Averages for", n, "Dice"),
       xlab = "Sample Average",
       col  = "lightblue",
       border = "white")

  # 3. Add a legend
  legend("topright",
         legend = c("Simulated"),
         col    = c("lightblue"),
         lty    = c(NA, 1),
         lwd    = c(NA, 2),
         pch    = c(15, NA),
         bty    = "n",
         cex    = 0.8)
}

# Add an overall title
mtext("Sample Averages of Dice Rolls (Illustrating the CLT)", outer = TRUE, cex = 0.8)
```



Problem 12

To validate the Kruskal’s count on p. 13 of slides DA01, we program and simulate the game with one deck of cards, i.e., 52 cards, for 10000 times. Each time, the 52 cards are randomly shuffled. We then start from the first 10 cards, and the face cards (J/Q/K) are counted as 5 steps

(a)

What is the probability that all the first 10 cards reach the same end?

```r
# 1. Create a standard 52-card deck
ranks <- c("A", "2", "3", "4", "5", "6", "7", "8", "9", "10", "J", "Q", "K")
deck  <- rep(ranks, each = 4)

# 2. Function to get step size
step_value <- function(card) {
  if (card %in% c("J", "Q", "K")) {
    return(5)          # face cards
  } else if (card == "A") {
    return(1)
  } else {
    return(as.numeric(card))  # 2..10
  }
}

# 3. Find the final card starting from position 'start'
final_card_from_position <- function(shuffled_deck, start) {
  pos <- start
  while (pos <= length(shuffled_deck)) {
    prev_pos <- pos
    pos <- pos + step_value(shuffled_deck[pos])
  }
  # The loop ends when we jump beyond the deck;
  # the "final card" is the last one we saw.
  return(shuffled_deck[prev_pos])
}

# 4. Single simulation: shuffle deck, check if first 10 ends are identical
one_sim <- function() {
  shuffled <- sample(deck)
  ends <- sapply(1:10, function(s) final_card_from_position(shuffled, s))
  length(unique(ends)) == 1  # TRUE if all ends match
}

# 5. Repeat 10,000 times and compute probability
n_sims <- 10000
results <- replicate(n_sims, one_sim())
prob_all_same <- mean(results)

prob_all_same
```

```
## [1] 0.6093
```

. ### (b)

Vary the simulation settings:

- the # of cards: [52,104];
- the # of steps for face cards = [1,3,5,7,9].

What are the 2 × 5 = 10 probabilities? Discuss your observation?

```r
# Function to assign a step value to a card.
#  (J, Q, K) return 5,
# Ace is counted as 1, and numbered cards return their numeric value.
card_value <- function(card, face_val) {
  if(card %in% c("J", "Q", "K")){
    return(face_val)
  } else if(card == "A"){
    return(1)
  } else {
    return(as.numeric(card))
  }
}

# Function to simulate the Kruskal count starting from a given position.
# The process stops when the next computed position exceeds the deck size.
# The "end" is defined as the last card reached (at position 'prev').
kruskal_end <- function(deck, start, face_val) {
  pos <- start
  repeat {
    if(pos > length(deck)) {
      return(deck[prev])
    }
    prev <- pos
    value <- card_value(deck[pos], face_val)
    pos <- pos + value
  }
}

# For one simulation iteration, compute the end for each of the first 10 starting positions.
# Return TRUE if all 10 ends are identical.
simulate_once <- function(deck, face_val) {
  ends <- sapply(1:10, function(i) kruskal_end(deck, i, face_val))
  all_same <- (length(unique(ends)) == 1)
  return(all_same)
}

# Run the simulation for a given deck size and face card step value.
simulate_probability <- function(deck_size, face_val, nrep=10000) {
  # Standard deck: 52 cards (4 copies each of 13 ranks)
  ranks <- c("A", "2", "3", "4", "5", "6", "7", "8", "9", "10", "J", "Q", "K")
  num_decks <- deck_size / 52
  deck_standard <- rep(ranks, 4)
  deck_full <- rep(deck_standard, num_decks)

  count_success <- 0
  for(i in 1:nrep) {
    deck_shuffled <- sample(deck_full)
    if(simulate_once(deck_shuffled, face_val)) {
      count_success <- count_success + 1
    }
  }
  return(count_success / nrep)
}

# Set up the combinations for deck sizes and face card values.
deck_sizes <- c(52, 104)
face_vals <- c(1, 3, 5, 7, 9)
results <- expand.grid(DeckSize = deck_sizes, FaceVal = face_vals)

# Run simulation for each combination
results$Probability <- mapply(simulate_probability, results$DeckSize, results$FaceVal)
print(results)
```

```
##    DeckSize FaceVal Probability
## 1        52       1      0.8356
## 2       104       1      0.9912
## 3        52       3      0.7306
## 4       104       3      0.9763
## 5        52       5      0.6078
## 6       104       5      0.9425
## 7        52       7      0.5034
## 8       104       7      0.8938
## 9        52       9      0.3768
## 10      104       9      0.8324
```