mid_term

March 2, 2025

1 Mid-Term Exam

1.1 Cheatsheet

You may or may not use the formulas below. Feel free to use them if feel needed.

$$P(A \cap B) = P(A)P(B) \text{ if A and B are independent}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
$$E(X) = \sum xp(x)$$
$$E[\sum X_i] = \sum E[X_i]$$
$$Var(X) = E[X - E[X]] = E[X^2] - E[X]^2$$
$$Var(\sum X_i) = \sum Var(X_i) \quad \text{if } X_i \text{ are independent with each other}$$
$${}_nC_k = \frac{n!}{k!(k-r)!}$$
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

1.2 Q1

Consider a random simple experiment of drawing cards from a deck of cards. This deck of cards has 4 suits: 13 hearts, 13 diamonds, 13 spades, and 13 clubs. Each card has a number ranging from 1 to 13.

1. Drawing one card from this deck of cards, what is the probability of drawing a card with a number that is larger or equal to 10, and has a red color (heart or diamond).
2. Drawing one card from this deck of cards, given that it has a black color (spade or club), what is the probability of drawing a card with a number less or equal to 3.
3. Drawing 10 cards with replacements from this deck of cards, what is the probability of getting at least 8 (larger or equal to 8) club cards?
4. Drawing 3 cards with replacement from this deck of cards, what is the expected value and variance of the sum of the numbers of the drawn 3 cards?

5. Drawing 5 cards without replacement from this deck of cards, what is the probability of getting a flush (all 5 cards having the same suit: heart, diamond, spade, or club)? Hint: Calculate the total number of possible combinations of drawing 5 cards from 52 cards, then calculate total number of possible combinations of drawing 5 cards of the same suit.

1.3 Q2

Suppose 70% of people like cats, and 60% of people like dogs, and 50% of people likes both dogs and cats.

1. Now, a friend told you that he does not like dogs, what is the probability that he likes cats?
2. You conducted a survey to randomly sample 10 people for their pet preferences. What is the expected value and the variance of the number of participants who like both cats and dogs?
3. Draw the probability mass function of the number of people who like both cats and dogs with a bar plot.

1.4 Q3

For a random variable X, where $X \sim N(10, 100)$.

1. What is the population mean and population standard deviation for variable X?
2. Given that $P(X < 29.6) = 0.975$, find k, where $P(k < X < 29.6) = 0.95$.
3. Write down the Python code to find $P(-10 < X < 50)$.

Hint: Recall that the Python function of calculating the cumulative probability function for $P(X < x)$ where $X \sim N(\mu, \sigma^2)$ is

```
from scipy import stats
stats.norm.cdf(x, loc=mean, scale=standard_deviation)
```

1.5 Q4

You conducted a study investigating people's attitudes toward using AI tools in their health care. People's attitude is measured as a continuous variable X, where a positive value indicates a positive attitude and a negative value indicates a negative attitude.

You hypothesized that people may have a negative attitude for using AI tools in their health care. Thus you randomly sampled 10 participants and measured their attitudes. The data is the following:

$-1, -2, 0, 1, -1, 1, -2, -1, -1, 1$

1. Calculate the sample mean and sample variance.
2. You decided to conduct a **two-tailed** t-test to test your hypothesis. State your null hypothesis and alternative hypothesis.
3. Construct the t statistic and state the sampling distribution for your t statistic.
4. Given your calculated t statistic and its sampling distribution, write down the Python code to calculate the p value.

Hint: Recall that the Python code for calculating the cumulative probability function for $P(T < x)$ where $T \sim t_k$ is

```
from scipy import stats
```

```
stats.t.cdf(x, df=k)
```

5. Assuming you want to do a **one-tailed** t test instead. What is the rejection region for your sample mean $\bar{X}$ at significance level $\alpha = 0.05$? Given that $P(T < -1.833) = 0.05$ where $T \sim t_9$.
6. For this **one-tailed** t test, you assume your null hypothesis $H_1 : \mu = -0.5$, calculate the power of your study at significance level $\alpha = 0.05$. Just write down the code to calculate the power.

1.6 Q5

A researcher wants to examine whether there is a relationship between gender (male, female) and preference for a type of movie genre (action, comedy, drama). The researcher surveys 200 people, and the data is summarized in the following table:

Gender	**Action**	**Comedy**	**Drama**	**Total**
Male	40	30	20	90
Female	20	50	40	110
Total	60	80	60	200

Using a significance level of 0.05, determine if there is a significant association between gender and movie preference.

Given that the critical value for $\chi^2 * = 5.991$, $P(X < \chi^2 *) = 0.95$, where $X \sim \chi^2_2$

1.7 Q6 (Extra credit 5 points)

One of the most important methods to estimate parameters in statistical models is the **Maximum Likelihood Estimation** (MLE) method.

This method takes several steps to estimate parameters:

1. Construct the likelihood function of parameters $\hat{\theta}$ for one observed data X_i using the probability density function. $L(\hat{\theta}|X_i) = pdf(X_i|\hat{\theta})$.
2. Construct the likelihood function of parameters for all observed data, assuming samples are *i.i.d.* $L(\hat{\theta}) = \prod pdf(X_i|\hat{\theta})$
3. Construct the log-likelihood function of parameters for all observed data. $LogL(\hat{\theta}) = log(\prod pdf(X_i|\hat{\theta})) = \sum log(pdf(X_i|\hat{\theta}))$
4. Find the parameters $\hat{\theta}$ that minimize the negative log-likelihood function of parameters for all observed data. $\hat{\theta} = argmin_{\hat{\theta}}(-LogL(\hat{\theta}))$

The found parameters $\hat{\theta}$ is the MLE estimator for the parameters in the model because these parameters maximize the likelihood (probability of observing the data).

In Python, we usually use `scipy.optimize.minimize` function from `scipy` package to minimize a given function.

Now, complete the following block of codes to estimate a population mean μ for a normal distributed random variable X, given the population variance $\sigma^2 = 1$ and following observed samples X_i.

$1, 2, 3, 4, 5, 6, 7, 8, 9$

```
from scipy.optimize import minimize
from scipy import stats
import numpy as np

observed_samples = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9])
population_variance = 100
population_standard_deviation = [PLEASE COMPLETE HERE]

def likelihood_function_for_single_data(theta, sample):
    likelihood = stats.norm.pdf([PLEASE COMPLETE HERE],
                                loc=[PLEASE COMPLETE HERE],
                                scale=population_standard_deviation)
    return likelihood

def log_likelihood_function_for_single_data(theta, sample):
    likelihood = likelihood_function_for_single_data(theta, sample)
    log_likelihood = np.log(likelihood)
    return log_likelihood

def log_likelihood_function_for_all_data(theta, data):
    log_likelihood_all = 0
    for sample in data:
        log_likelihood = log_likelihood_function_for_single_data(theta, sample)
        log_likelihood_all = log_likelihood_all + [PLEASE COMPLETE HERE]
    return log_likelihood_all

def function_to_minimize(theta, data=observed_samples):
    log_likelihood_all = log_likelihood_function_for_all_data(theta, data)
    negative_log_likelihood_all = [PLEASE COMPLETE HERE]
    return negative_log_likelihood_all

x0 = 0 #x0 is an initial guess for parameter for the optimization algorithm to start with
res = minimize(function_to_minimize, x0)
mu_estimate = res.x
print(mu_estimate)
```

You will find the MLE estimator for the population mean is roughly the same as our sample mean, which is an unbiased estimator for the population mean.