

ECE 131A Project
Day in the Life of a Data-Scientist

Khyle Calpe
405016683
Discussion 1A

17 March 2020

1 Data Imputation

1.1 Proof that $a_i = \mu \forall i \in K_{miss}$ minimizes E_{MMSE}

$$E_{MMSE} = E\left[\sum_{i \in K_{miss}} (X_i - a_i)^2\right] \quad (1)$$

$$\frac{E_{MMSE}}{da_i} = E\left[\sum_{i \in K_{miss}} -2(X_i - a_i)\right] \quad (2)$$

$$= E\left[\sum_{i \in K_{miss}} (2a_i - 2X_i)\right] \quad (3)$$

$$= \sum_{i \in K_{miss}} (2E[a_i] - 2E[X_i]) \quad \text{by the linearity of expectation} \quad (4)$$

$$= \sum_{i \in K_{miss}} (2E[\mu] - 2E[X_i]) \quad \text{by substituting } \mu \text{ for } a_i \quad (5)$$

$$= \sum_{i \in K_{miss}} (2\mu - 2E[X_i]) \quad \text{by the expected value of a constant} \quad (6)$$

$$= \sum_{i \in K_{miss}} (2\mu - 2\mu) \quad \text{since the expectation of each i.i.d. RV is } \mu \quad (7)$$

$$= 0 \quad (8)$$

1.2 Sample mean $\hat{\mu}_N$ over N samples

N	$\hat{\mu}_N$
10	16.3022
20	17.5534
50	18.4207
100	20.0109
200	20.3985
300	19.9114
500	20.2776
1000	19.9514
2000	19.9790
10000	20.0142
20000	19.9444
30000	19.9916
60000	20.0318

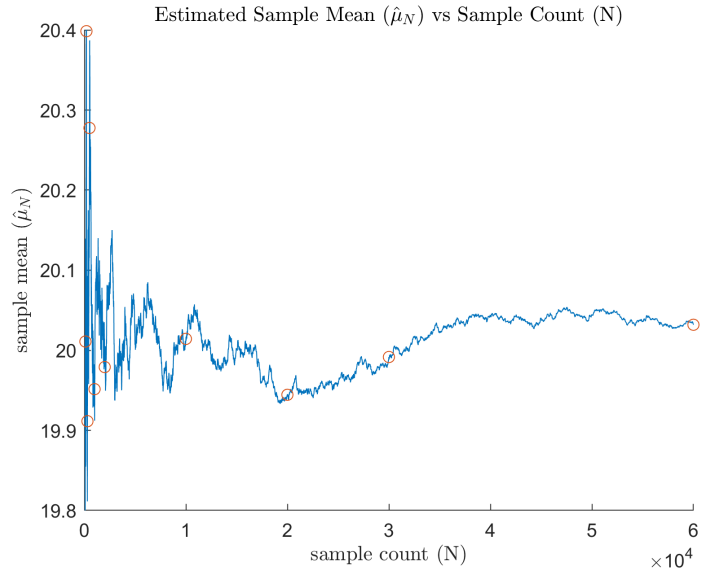


Figure 1: Sample means estimated from 10 to 60000 samples.

$\hat{\mu}_N$ behaves erratically until the sample count reaches 100 samples, then fluctuates around a value of 20 ± 0.01 . Hence, after 100 samples, $\hat{\mu}_N$ approaches the true mean.

1.3 Sample mean accuracy \hat{A}_N over N samples

N	\hat{A}_N
10	112.8127
20	105.1433
50	101.6648
100	99.1946
200	99.3591
300	99.2009
500	99.2756
1000	99.1960
2000	99.1945
10000	99.1948
20000	99.1966
30000	99.1943
60000	99.1958

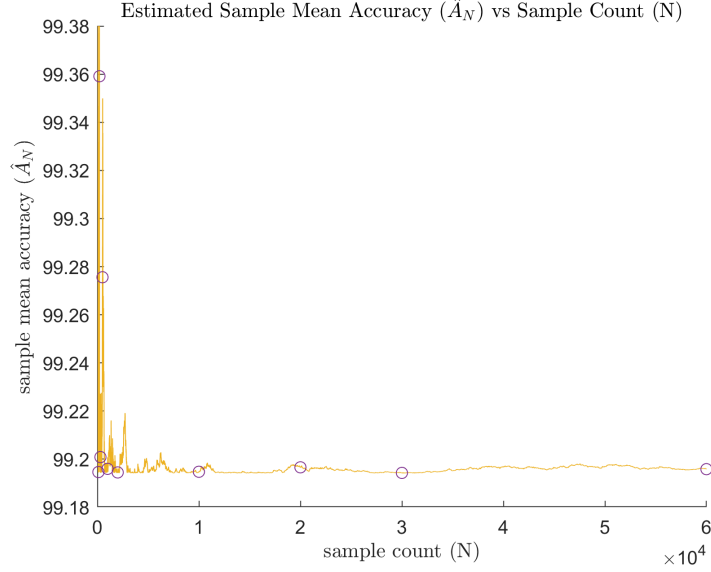


Figure 2: Sample mean accuracy estimated from 10 to 60000 samples.

As the sample count reaches 1000 samples, \hat{A}_N fluctuates approximately between 112 and 99.2. From 10000 to 60000 samples, \hat{A}_N approaches 99.2. After 10000 samples and as $\hat{\mu}_N$ approaches the true mean, \hat{A}_N reaches an approximate value.

1.4 Limiting Value of \hat{A}_N

As N approaches a large number, based on figure 2, \hat{A}_N approaches the value of 99.2. Since $\hat{\mu}_N$ approaches the true mean as N approaches a large number and the random variable X_i is positive, the difference between each sample and the true mean is non-negative. Additionally, since the data samples are not identical, the limiting value of \hat{A}_N is not zero.

1.5 Estimation of σ^2

Based on the limiting value of \hat{A}_N , the variance is approximately 99.2.

2 Central Limit Theorem

2.1 PDF & CDF of the mean M_n of a sequence of i.i.d. RVs

Mean Distribution

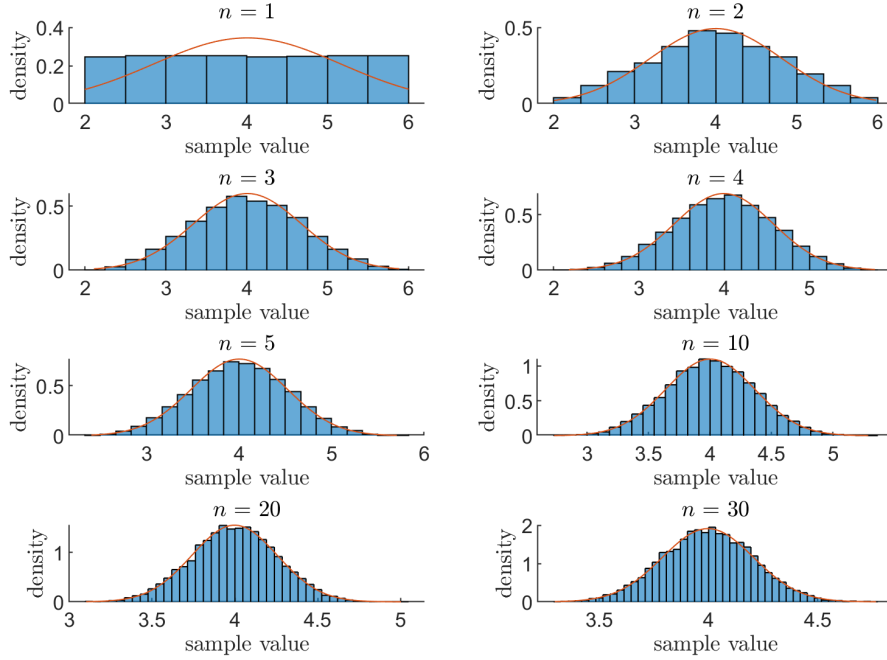


Figure 3: PDF of M_n for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

Cumulative Mean Distribution

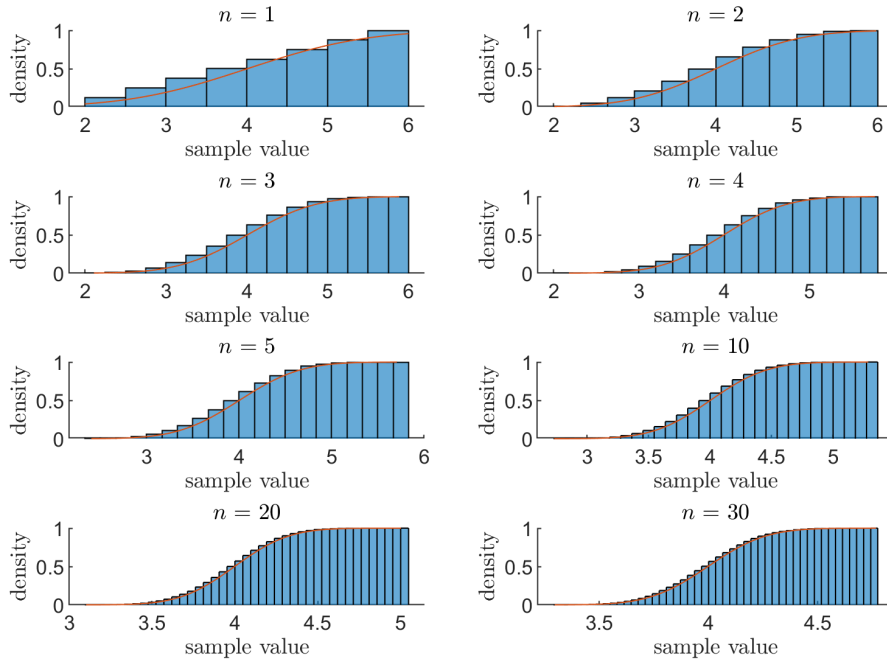


Figure 4: CDF of M_n for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

Based on the histograms, the mean and cumulative mean distributions of M_n approach the pdf and cdf of a Gaussian RV as n increases. The results show that the pdf and cdf of the sum of a sequence of i.i.d. RVs follows the pdf and cdf distributions of a Gaussian RV.

2.2 Mean and Variance of X_i and M_n

2.2.1 Mean and Variance of X_i

According to the prompt given, the mean of X_i is equal to μ and the variance of X_i is equal to $\sigma^2 \forall i \in \{1, 2, 3, \dots, n\}$.

2.2.2 Mean of M_n

$$E[M_n] = E\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] \quad \text{by the definition of the expectation} \quad (1)$$

$$= \frac{E[X_1] + E[X_2] + \dots + E[X_n]}{E[n]} \quad \text{by the linearity of expectation} \quad (2)$$

$$= \frac{\mu + \mu + \dots + \mu}{E[n]} \quad \text{since } E[X_i] = \mu \forall i \quad (3)$$

$$= \frac{\mu + \mu + \dots + \mu}{n} \quad \text{by the expected value of a constant} \quad (4)$$

$$= \frac{n\mu}{n} \quad (5)$$

$$= \mu \quad (6)$$

2.2.3 Variance of M_n

$$Var(M_n) = Var\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) \quad \text{by the definition of } M_n \quad (1)$$

$$= \frac{1}{n^2} Var(X_1 + X_2 + \dots + X_n) \quad \text{since } Var(aX) = a^2 Var(X) \quad (2)$$

$$= \frac{1}{n^2} [Var(X_1) + Var(X_2) + \dots + Var(X_n)] \quad \text{since } X_j \perp\!\!\!\perp X_k \forall j \neq k \quad (3)$$

$$= \frac{1}{n^2} [\sigma^2 + \sigma^2 + \dots + \sigma^2] \quad \text{since } Var[X_i] = \sigma^2 \forall i \quad (4)$$

$$= \frac{n\sigma^2}{n^2} \quad (5)$$

$$= \frac{\sigma^2}{n} \quad (6)$$

2.3 Multivariate Gaussian RV

Mean Distribution

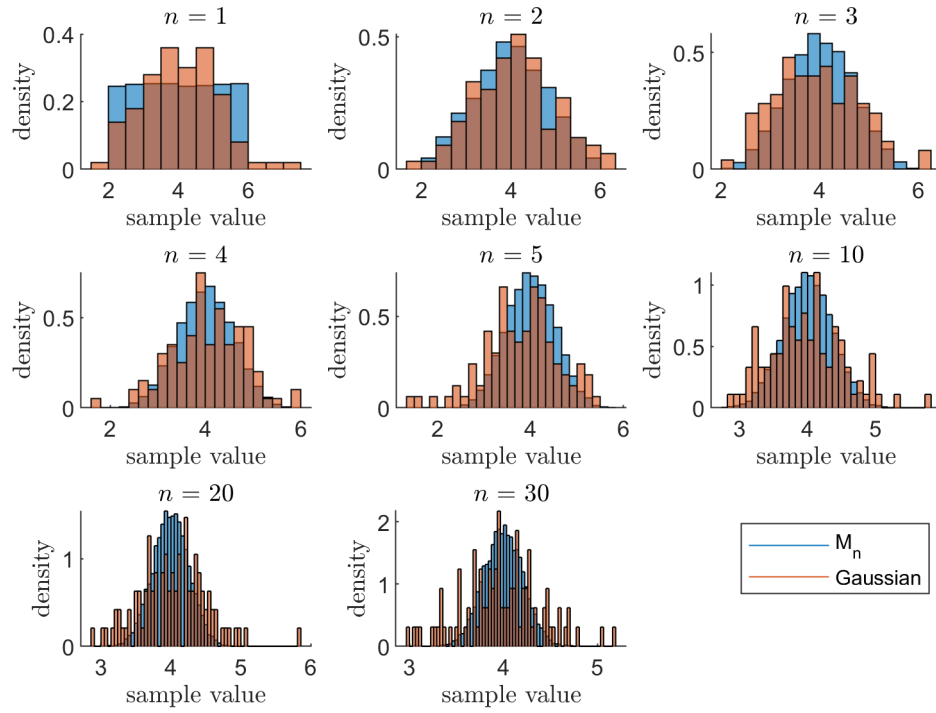


Figure 5: PDF of M_n and a Multivariate Gaussian for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

Cumulative Mean Distribution

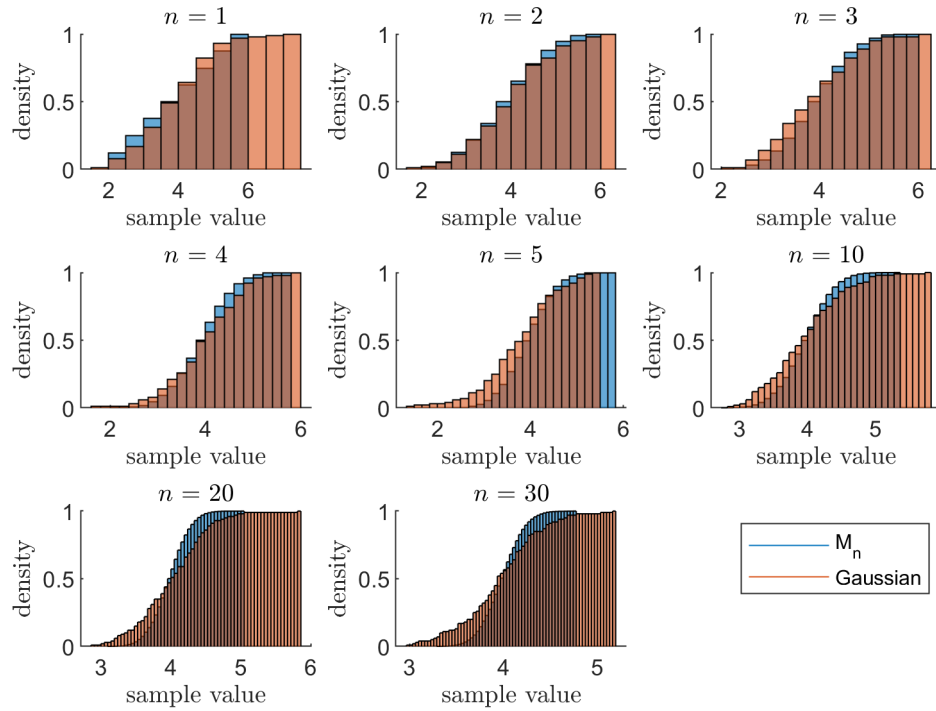


Figure 6: CDF of M_n and a Multivariate Gaussian for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

2.4 X_i representing an unfair 5-sided dice

2.4.1 PDF & CDF of the mean M_n of a sequence of biased RVs

Mean Distribution

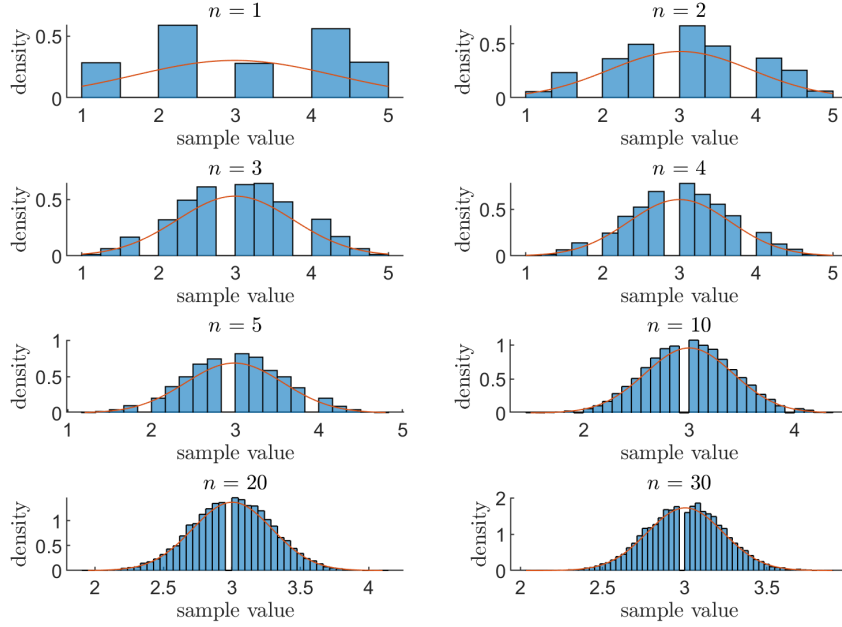


Figure 7: PDF of M_n for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

Cumulative Mean Distribution

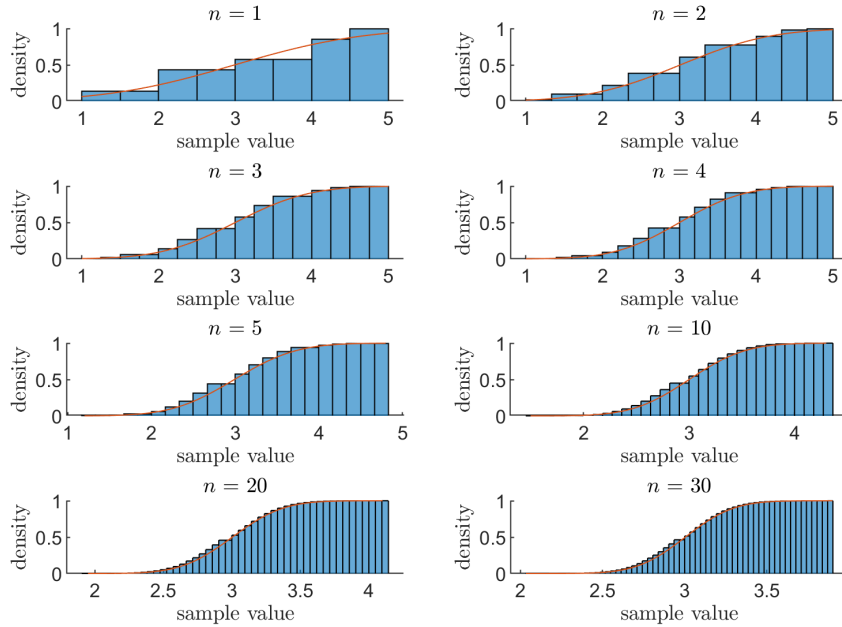


Figure 8: CDF of M_n for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

The histograms show the same Gaussian distributions as the last section. However, the pdf and cdf indicate bias for even sample values as indicated by the bimodal peaks in the pdf and the rise of the distribution for even-valued bins.

2.4.2 Mean and Variance of X_i and M_n

2.4.2.1 Mean of X_i

$$E[X_i] = \sum_{i=1}^5 x_i P(x = x_i) \quad \text{by the definition of expectation} \quad (1)$$

$$= 1 \cdot p + 2 \cdot 2p + 3 \cdot p + 4 \cdot 2p + 5 \cdot p \quad \text{by the pmf of an unfair 5-sided dice} \quad (2)$$

$$= 21 \cdot p \quad (3)$$

$$= 21 \cdot \frac{1}{7} \quad \text{since } p = \frac{1}{7} \text{ for } \sum_{i=1}^5 P(x = x_i) = 1 \quad (4)$$

$$= 3 \quad (5)$$

2.4.2.2 Variance of X_i

$$E[X_i^2] = \sum_{i=1}^5 x_i^2 P(x = x_i) \quad (1)$$

$$= 1^2 \cdot p + 2^2 \cdot 2p + 3^2 \cdot p + 4^2 \cdot 2p + 5^2 \cdot p \quad (2)$$

$$= 75 \cdot p \quad (3)$$

$$= \frac{75}{7} \quad (4)$$

$$Var(X_i) = E[X_i^2] - (E[X_i])^2 \quad \text{by the definition of variance} \quad (1)$$

$$= \frac{75}{7} - 3^2 \quad (2)$$

$$= \frac{12}{7} \quad (3)$$

2.4.2.3 Mean of M_n

$$E[M_n] = E\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] \quad \text{by the definition of the expectation} \quad (1)$$

$$= \frac{E[X_1] + E[X_2] + \dots + E[X_n]}{E[n]} \quad \text{by the linearity of expectation} \quad (2)$$

$$= \frac{3 + 3 + \dots + 3}{E[n]} \quad \text{since } E[X_i] = 3 \forall i \in \{1, 2, 3, \dots, n\} \quad (3)$$

$$= \frac{3 + 3 + \dots + 3}{n} \quad \text{by the expected value of a constant} \quad (4)$$

$$= \frac{n \cdot 3}{n} \quad (5)$$

$$= 3 \quad (6)$$

2.4.2.4 Variance of M_n

$$\text{Var}(M_n) = \text{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \quad \text{by the definition of } M_n \quad (1)$$

$$= \frac{1}{n^2} \text{Var}(X_1 + X_2 + \cdots + X_n) \quad \text{since } \text{Var}(aX) = a^2 \text{Var}(X) \quad (2)$$

$$= \frac{1}{n^2} [\text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)] \quad \text{since } X_j \perp\!\!\!\perp X_k \text{ for } j \neq k \quad (3)$$

$$= \frac{1}{n^2} \left[\frac{12}{7} + \frac{12}{7} + \cdots + \frac{12}{7} \right] \quad \text{since } \text{Var}[X_i] = \frac{12}{7} \forall i \quad (4)$$

$$= \frac{12n}{7n^2} \quad (5)$$

$$= \frac{12}{7n} \quad (6)$$

2.4.3 Multivariate Gaussian RV

Mean Distribution

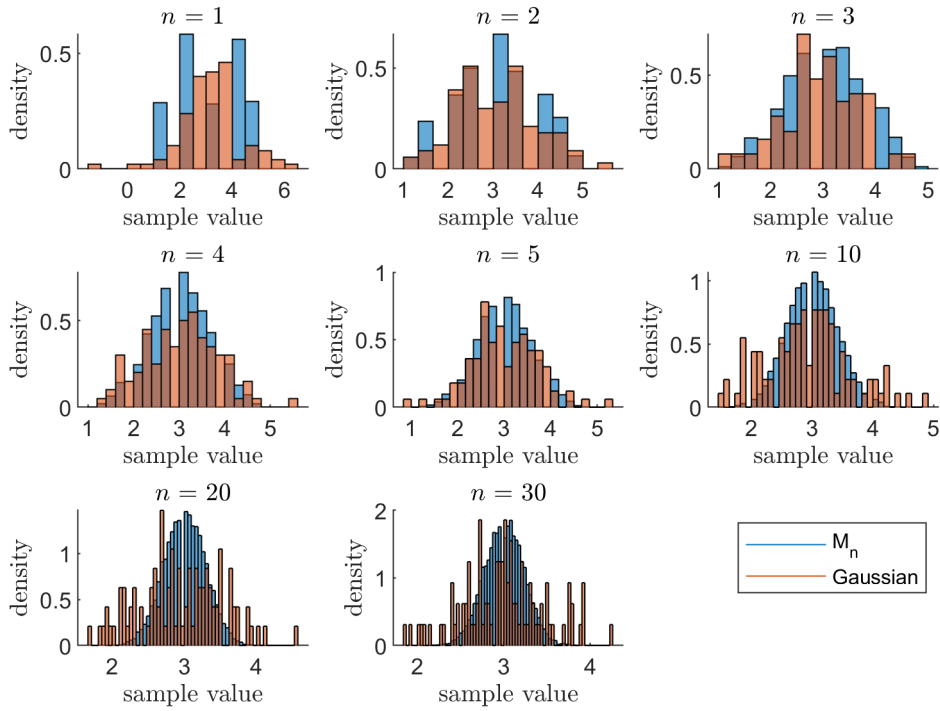


Figure 9: PDF of M_n and a Multivariate Gaussian for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

Cumulative Mean Distribution

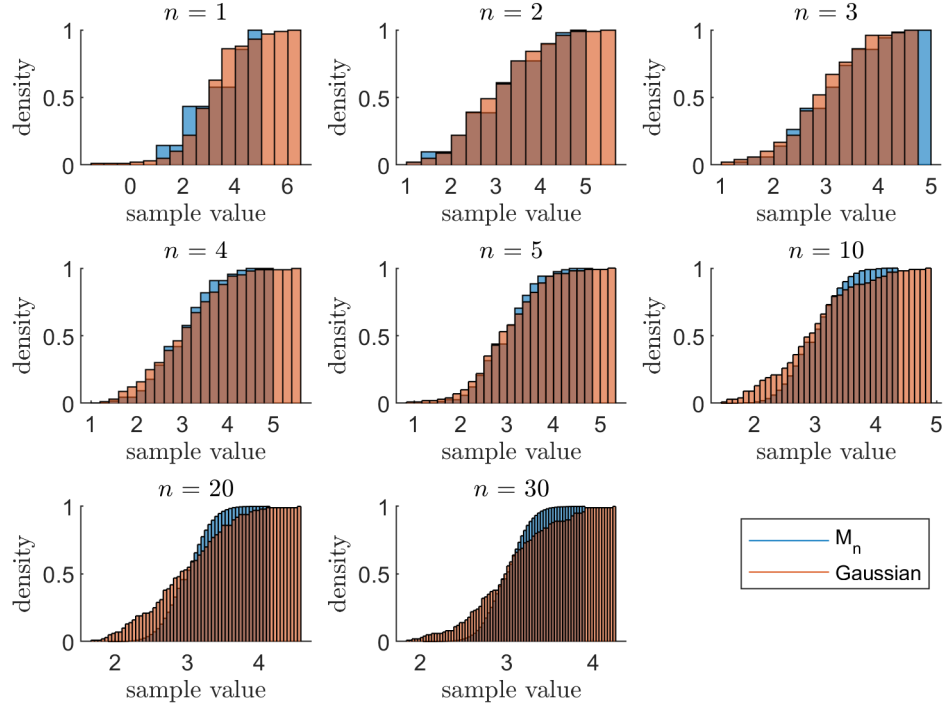


Figure 10: CDF of M_n and a Multivariate Gaussian for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

3 Gaussian Discriminant Analysis

3.1 Classification Rule for $\Sigma = \Sigma_0 = \Sigma_1$ and $p = \frac{1}{2}$

$$P(y = 0|\vec{x}) \geq P(y = 1|\vec{x}) \quad (1)$$

$$\frac{P(\vec{x}|y = 0)P(y = 0)}{P(\vec{x})} \geq \frac{P(\vec{x}|y = 1)P(y = 1)}{P(\vec{x})} \quad \text{by Bayes' Rule} \quad (2)$$

$$P(\vec{x}|y = 0)P(y = 0) \geq P(\vec{x}|y = 1)P(y = 1) \quad (3)$$

$$\frac{1}{2}P(\vec{x}|y = 0) \geq \frac{1}{2}P(\vec{x}|y = 1) \quad \text{since } y = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1 - p \end{cases} \quad (4)$$

$$P(\vec{x}|y = 0) \geq P(\vec{x}|y = 1) \quad (5)$$

$$f_{X,y=0}(\vec{x}) \geq f_{X,y=1}(\vec{x}) \quad (6)$$

$$f_X(\vec{x}) = \frac{\exp\{-\frac{1}{2}(\vec{x} - \mu)^T K^{-1}(\vec{x} - \mu)\}}{(2\pi)^{\frac{n}{2}} |K|^{\frac{1}{2}}} \quad \text{by the definition of a Gaussian pdf} \quad (1)$$

$$K = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & & \vdots \\ \text{Cov}(X_n, X_1) & \dots & & \text{Var}(X_n) \end{bmatrix} \quad (1)$$

$$= \begin{bmatrix} \text{Var}(X_1) & 0 & \dots & 0 \\ 0 & \text{Var}(X_2) & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \dots & & \text{Var}(X_n) \end{bmatrix} \quad \text{since } \rho = 0 \text{ for } X_j \perp\!\!\!\perp X_k \text{ for } j \neq k \quad (2)$$

$$= \begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & \Sigma & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \dots & & \Sigma \end{bmatrix} \quad (3)$$

$$K^{-1} = \frac{1}{\Sigma} \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \dots & & 1 \end{bmatrix} \quad (1)$$

$$= \frac{I}{\Sigma} \quad (2)$$

$$f_X(\vec{x}) = \frac{\exp\{-\frac{1}{2}(\vec{x} - \mu)^T \frac{I}{\sum} (\vec{x} - \mu)\}}{(2\pi)^{\frac{n}{2}} |\sum I|^{\frac{1}{2}}} \quad (1)$$

$$= \frac{\exp\{-\frac{1}{2}(\vec{x}^T \Sigma^{-1} \vec{x} - \mu^T \Sigma^{-1} \vec{x} - x^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu)\}}{(2\pi)^{\frac{n}{2}} |\sum|^{\frac{1}{2}}} \quad \text{cross multiply} \quad (2)$$

$$\ln(\exp) = \ln(\exp\{-\frac{1}{2}(\vec{x}^T \Sigma^{-1} \vec{x} - \mu^T \Sigma^{-1} \vec{x} - x^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu)\}) \quad (1)$$

$$= -\frac{1}{2}(\vec{x}^T \Sigma^{-1} \vec{x} - \mu^T \Sigma^{-1} \vec{x} - x^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu) \quad (2)$$

$$= -\frac{1}{2}(\vec{x}^T \Sigma^{-1} \vec{x} - 2\mu^T \Sigma^{-1} \vec{x} + \mu^T \Sigma^{-1} \mu) \quad \text{since } \mu^T \Sigma^{-1} \vec{x} = x^T \Sigma^{-1} \mu \quad (3)$$

$$f_{X,y=0}(\vec{x}) \geq f_{X,y=1}(\vec{x}) \quad (1)$$

$$-\frac{1}{2}(\vec{x}^T \Sigma^{-1} \vec{x} - 2\mu_0^T \Sigma^{-1} \vec{x} + \mu_0^T \Sigma^{-1} \mu_0) \geq -\frac{1}{2}(\vec{x}^T \Sigma^{-1} \vec{x} - 2\mu_1^T \Sigma^{-1} \vec{x} + \mu_1^T \Sigma^{-1} \mu_1) \quad (2)$$

$$\vec{x}^T \Sigma^{-1} \vec{x} - 2\mu_0^T \Sigma^{-1} \vec{x} + \mu_0^T \Sigma^{-1} \mu_0 \geq \vec{x}^T \Sigma^{-1} \vec{x} - 2\mu_1^T \Sigma^{-1} \vec{x} + \mu_1^T \Sigma^{-1} \mu_1 \quad (3)$$

$$-2\mu_0^T \Sigma^{-1} \vec{x} + 2\mu_1^T \Sigma^{-1} \vec{x} \geq \mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0 \quad (4)$$

$$[2(\mu_1 - \mu_0)^T \Sigma^{-1}] \vec{x} + [\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1] \geq 0 \quad (5)$$

3.2 Linear Inequality Contour for $\sum = \sum_0 = \sum_1$ and $p = \frac{1}{2}$

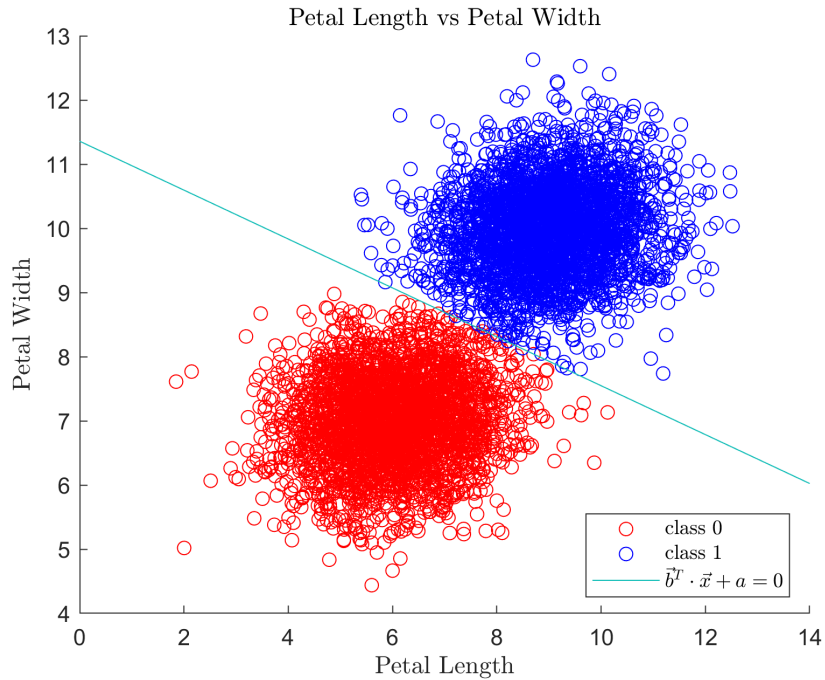


Figure 11: Scatter plot and contour line of two classes from a sample with the same covariance and based on the linear inequality with $p = \frac{1}{2}$. Based on the samples, 50.17% are categorized as class 0.

3.3 Classification Rule for $\Sigma = \Sigma_0 = \Sigma_1$ and a general p

$$P(y = 0|\vec{x}) \geq P(y = 1|\vec{x}) \quad (1)$$

$$\frac{P(\vec{x}|y = 0)P(y = 0)}{P(\vec{x})} \geq \frac{P(\vec{x}|y = 1)P(y = 1)}{P(\vec{x})} \quad \text{by Bayes' Rule} \quad (2)$$

$$P(\vec{x}|y = 0)P(y = 0) \geq P(\vec{x}|y = 1)P(y = 1) \quad (3)$$

$$(1 - p)P(\vec{x}|y = 0) \geq (p)P(\vec{x}|y = 1) \quad \text{since } y = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1 - p \end{cases} \quad (4)$$

$$(1 - p)f_{X,y=0}(\vec{x}) \geq (p)f_{X,y=1}(\vec{x}) \quad (5)$$

$$(1 - p)f_{X,y=0}(\vec{x}) \geq (p)f_{X,y=1}(\vec{x}) \quad (1)$$

$$\frac{\frac{1-p}{p} \exp(\vec{x}^T \Sigma^{-1} \vec{x} - 2\mu_0^T \Sigma^{-1} \vec{x} + \mu_0^T \Sigma^{-1} \mu_0)}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \geq \frac{\exp(\vec{x}^T \Sigma^{-1} \vec{x} - 2\mu_1^T \Sigma^{-1} \vec{x} + \mu_1^T \Sigma^{-1} \mu_1)}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \quad (2)$$

$$\frac{(1-p)}{p} \exp(\vec{x}^T \Sigma^{-1} \vec{x} - 2\mu_0^T \Sigma^{-1} \vec{x} + \mu_0^T \Sigma^{-1} \mu_0 - \vec{x}^T \Sigma^{-1} \vec{x} + 2\mu_1^T \Sigma^{-1} \vec{x} - \mu_1^T \Sigma^{-1} \mu_1) \geq 1 \quad (1)$$

$$\frac{(1-p)}{p} \exp(-2\mu_0^T \Sigma^{-1} \vec{x} + \mu_0^T \Sigma^{-1} \mu_0 + 2\mu_1^T \Sigma^{-1} \vec{x} - \mu_1^T \Sigma^{-1} \mu_1) \geq 1 \quad (2)$$

$$\frac{(1-p)}{p} \exp(2(\mu_1 - \mu_0)^T \Sigma^{-1} \vec{x} + \mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1) \geq 1 \quad (3)$$

$$\ln\left[\frac{(1-p)}{p} \exp(2(\mu_1 - \mu_0)^T \Sigma^{-1} \vec{x} + \mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1)\right] \geq 0 \quad (4)$$

$$[2(\mu_1 - \mu_0)^T \Sigma^{-1}] \vec{x} + [\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1 + \ln(\frac{1-p}{p})] \geq 0 \quad (5)$$

3.4 Linear Inequality Contour for $\Sigma = \Sigma_0 = \Sigma_1$ and $p = 0.05$

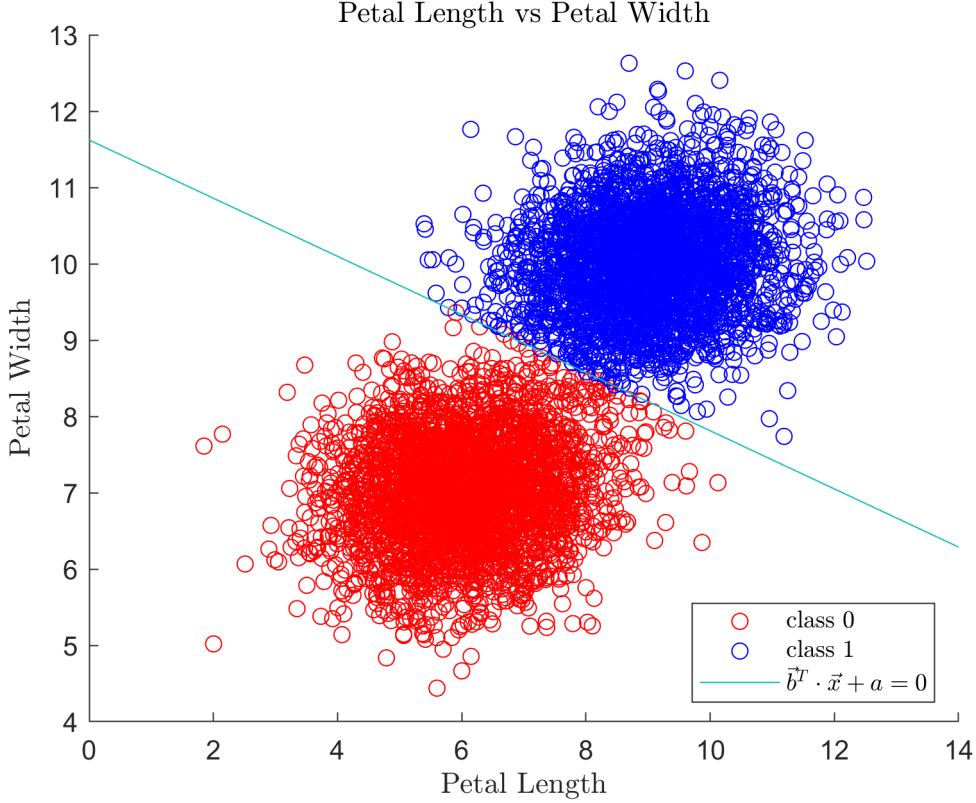


Figure 12: Scatter plot and contour line of two classes from a sample with the same covariance and based on the linear inequality with $p = 0.05$. Based on the samples, 50.85% are classified as class 0. Also, the change in p shifted the contour line towards the samples categorized as class 1. The result is expected given that the probability of a sample to be categorized as class 0, as opposed to 1, is higher.

3.5 Quadratic Inequality

$$P(y = 0|\vec{x}) \geq P(y = 1|\vec{x}) \quad (1)$$

$$\frac{P(\vec{x}|y = 0)P(y = 0)}{P(\vec{x})} \geq \frac{P(\vec{x}|y = 1)P(y = 1)}{P(\vec{x})} \quad \text{by Bayes' Rule} \quad (2)$$

$$P(\vec{x}|y = 0)P(y = 0) \geq P(\vec{x}|y = 1)P(y = 1) \quad (3)$$

$$(1 - p)P(\vec{x}|y = 0) \geq (p)P(\vec{x}|y = 1) \quad \text{since } y = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1 - p \end{cases} \quad (4)$$

$$(1 - p)f_{X,y=0}(\vec{x}) \geq (p)f_{X,y=1}(\vec{x}) \quad (5)$$

$$(1-p)f_{X,y=0}(\vec{x}) \geq (p)f_{X,y=1}(\vec{x}) \quad (1)$$

$$\frac{\frac{1-p}{p} \exp(\vec{x}^T \Sigma_0^{-1} \vec{x} - 2\mu_0^T \Sigma_0^{-1} \vec{x} + \mu_0^T \Sigma_0^{-1} \mu_0)}{(2\pi)^{\frac{n}{2}} |\Sigma|^\frac{1}{2}}} \geq \frac{\exp(\vec{x}^T \Sigma_1^{-1} \vec{x} - 2\mu_1^T \Sigma_1^{-1} \vec{x} + \mu_1^T \Sigma_1^{-1} \mu_1)}{(2\pi)^{\frac{n}{2}} |\Sigma|^\frac{1}{2}}} \quad (2)$$

$$\frac{\frac{1-p}{p} \exp(\vec{x}^T \Sigma_0^{-1} \vec{x} - 2\mu_0^T \Sigma_0^{-1} \vec{x} + \mu_0^T \Sigma_0^{-1} \mu_0)}{|\Sigma_0|^\frac{1}{2}}} \geq \frac{\exp(\vec{x}^T \Sigma_1^{-1} \vec{x} - 2\mu_1^T \Sigma_1^{-1} \vec{x} + \mu_1^T \Sigma_1^{-1} \mu_1)}{|\Sigma_1|^\frac{1}{2}}} \quad (3)$$

$$\begin{aligned} \frac{(1-p)}{p} |\Sigma_0|^{-\frac{1}{2}} |\Sigma_1|^\frac{1}{2} \exp(\vec{x}^T (\Sigma_0^{-1} - \Sigma_1^{-1}) \vec{x} \\ + 2(\mu_1^T \Sigma_1^{-1} - \mu_0^T \Sigma_0^{-1}) \vec{x} + \mu_0^T \Sigma_0^{-1} \mu_0 - \mu_1^T \Sigma_1^{-1} \mu_1) \geq 1 \end{aligned} \quad (4)$$

$$\begin{aligned} \ln\left[\frac{(1-p)}{p} |\Sigma_0|^{-\frac{1}{2}} |\Sigma_1|^\frac{1}{2} \exp(\vec{x}^T (\Sigma_0^{-1} - \Sigma_1^{-1}) \vec{x} \right. \\ \left. + 2(\mu_1^T \Sigma_1^{-1} - \mu_0^T \Sigma_0^{-1}) \vec{x} + \mu_0^T \Sigma_0^{-1} \mu_0 - \mu_1^T \Sigma_1^{-1} \mu_1)\right] \geq \ln(1) \end{aligned} \quad (5)$$

$$\begin{aligned} \vec{x}^T (\Sigma_0^{-1} - \Sigma_1^{-1}) \vec{x} + 2(\mu_1^T \Sigma_1^{-1} - \mu_0^T \Sigma_0^{-1}) \vec{x} \\ + [\mu_0^T \Sigma_0^{-1} \mu_0 - \mu_1^T \Sigma_1^{-1} \mu_1 + \ln(\frac{1-p}{p} |\Sigma_0|^{-\frac{1}{2}} |\Sigma_1|^\frac{1}{2})] \geq 0 \end{aligned} \quad (6)$$

3.6 Quadratic Inequality Contour

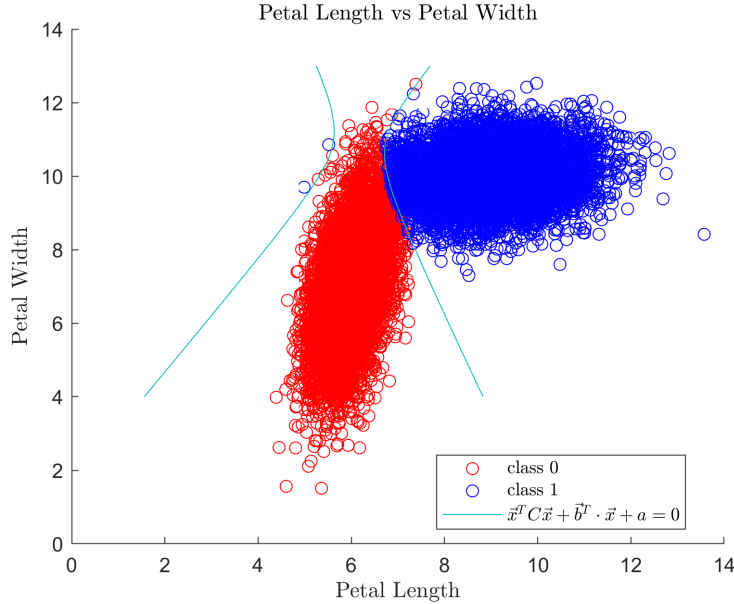


Figure 13: Scatter plot and contour line of two classes from a sample with distributions that do not necessarily have the same covariances and based on the quadratic inequality with $p = 0.5$. Based on the samples, 50.51% are classified as class 0.

4 Appendix

```
1 %% Read/store files
2 load data.txt;
3 load data_2.txt;
4 load data_3.txt;
5 f_path = 'D:\UCLA\Courses\EE 131A\Project\Plots';
6
7 %% Section 1
8
9 % Indices of missing and available data
10 K_miss = isnan(data);
11 K_avail = ~isnan(data);
12
13 % Filtered data
14 data_miss = data(K_miss);
15 data_avail = data(K_avail);
16
17 %% Section 1.b
18
19 % Estimate sample mean
20 N_cont = 10:10:60000;
21 N_disc =
    [10,20,50,100,200,300,500,1000,2000,10000,20000,30000,60000];
22 mu_N_c = sample_mean(data_avail, N_cont);
23 mu_N_d = sample_mean(data_avail, N_disc);
24
25 % mu_N vs N plot
26 hold on;
27 plot(N_cont, mu_N_c);
28 scatter(N_disc, mu_N_d);
29 title('Estimated Sample Mean ( $\hat{\mu}_N$ ) vs Sample Count (N)', 'Interpreter', 'Latex');
30 xlabel('sample count (N)', 'Interpreter', 'Latex');
31 ylabel('sample mean ( $\hat{\mu}_N$ )', 'Interpreter', 'Latex');
32 ylim([19.8 20.4])
33 print(gcf, fullfile(f_path, '01_b'), '-dpng', '-r300');
34
35 %% Section 1.c
36
37 % Estimate sample mean accuracy
38 square_d = (data_avail(:) - mu_N_d).^2;
39 square_c = (data_avail(:) - mu_N_c).^2;
40 acc_d = zeros(1, length(mu_N_d));
41 acc_c = zeros(1, length(mu_N_c));
42
43 for mu_N_index = 1:length(mu_N_d)
44     acc_d(mu_N_index) = sum(square_d(:, mu_N_index))/length(
```



```

        data_avail);
45 end
46
47 for mu_N_index = 1:length(mu_N_c)
48     acc_c(mu_N_index) = sum(square_c(:, mu_N_index))/length(
        data_avail);
49 end
50
51 % A_N vs N plot
52 hold on;
53 plot(N_cont, acc_c);
54 scatter(N_disc, acc_d);
55 title('Estimated Sample Mean Accuracy ( $\hat{A}_N$ ) vs Sample
        Count (N)', 'Interpreter', 'Latex');
56 xlabel('sample count (N)', 'Interpreter', 'Latex');
57 ylabel('sample mean accuracy ( $\hat{A}_N$ )', 'Interpreter', '
        Latex');
58 xlim([0 60000]);
59 ylim([99.18 99.38]);
60 print(gcf, fullfile(f_path, '01_c'), '-dpng', '-r300');
61
62 %% Section 2.a
63
64 % Initializations
65 n = [1,2,3,4,5,10,20,30]; % samples
66 samples = 10000;
67 M_n = zeros(samples, length(n)); % RVs
68 mean_sum = zeros([1 samples]);
69 min_val = zeros([1 length(n)]); % subplots
70 max_val = zeros([1 length(n)]);
71 sample_x = zeros(100, length(n));
72 mean_n = zeros([1 length(n)]);
73 sd_n = zeros([1 length(n)]);
74
75 % Generate mean for n RVs
76 for mean_ind = 1:length(n)
77     for sum_ind = 1:n(mean_ind)
78         mean_sum = mean_sum + 4*rand([1 samples])+2;
79     end
80     M_n(:, mean_ind) = mean_sum./n(mean_ind);
81     % Reset for next iteration
82     mean_sum = zeros([1 samples]);
83 end
84
85 % Close all open figures
86 close all;
87
88 % PDF & CDF subplots

```

```

89 for plot_ind = 1:length(n)
90     % Parameters
91     min_val(plot_ind)      = min(M_n(:, plot_ind));
92     max_val(plot_ind)      = max(M_n(:, plot_ind));
93     sample_x(:, plot_ind) = linspace(min_val(plot_ind), max_val(
        plot_ind));
94     mean_n(plot_ind)       = mean(M_n(:, plot_ind));
95     sd_n(plot_ind)        = std(M_n(:, plot_ind));
96
97     % Formatting (PDF)
98     figure(1);
99     subplot(4, 2, plot_ind);
100    title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
101    title(title_string, 'Interpreter', 'Latex');
102    xlabel('sample value', 'Interpreter', 'Latex');
103    ylabel('density', 'Interpreter', 'Latex');
104
105    % Plot (PDF)
106    hold on;
107    histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'pdf');
108    y_pdf = (1/(sd_n(plot_ind).*sqrt(2*pi))).*exp(-(1/2)*((
        sample_x(:, plot_ind)-mean_n(plot_ind))./sd_n(plot_ind))
        .^2);
109    plot(sample_x(:, plot_ind), y_pdf);
110
111    % Formatting (CDF)
112    figure(2);
113    subplot(4, 2, plot_ind);
114    title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
115    title(title_string, 'Interpreter', 'Latex');
116    xlabel('sample value', 'Interpreter', 'Latex');
117    ylabel('density', 'Interpreter', 'Latex');
118
119    % Plot (CDF)
120    hold on;
121    histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'cdf');
122    y_cdf = cdf('Normal', sample_x(:, plot_ind), mean_n(plot_ind
        ), sd_n(plot_ind));
123    plot(sample_x(:, plot_ind), y_cdf);
124 end
125
126 % Save subplots
127 figure(1);
128 sgtitle('Mean Distribution', 'Interpreter', 'Latex');
129 print(gcf, fullfile(f_path, '02_a-pdf'), '-dpng', '-r300');
130

```

```

131 figure(2);
132 sgtitle('Cumulative Mean Distribution', 'Interpreter', 'Latex');
133 print(gcf, fullfile(f_path, '02_a_cdf'), '-dpng', '-r300');
134
135 %% Section 2.c
136
137 % PDF & CDF subplots
138 for plot_ind = 1:length(n)
139     % Multivariate Gaussian
140     mv_GRV = mvnrnd(mean_n(plot_ind), sd_n(plot_ind), length(
        sample_x(:, plot_ind)));
141
142     % Formatting (PDF)
143     figure(1);
144     subplot(3, 3, plot_ind);
145     title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
146     title(title_string, 'Interpreter', 'Latex');
147     xlabel('sample value', 'Interpreter', 'Latex');
148     ylabel('density', 'Interpreter', 'Latex');
149
150     % Plot (PDF)
151     hold on;
152     histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'pdf');
153     histogram(mv_GRV, 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'pdf');
154
155     % Formatting (CDF)
156     figure(2);
157     subplot(3, 3, plot_ind);
158     title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
159     title(title_string, 'Interpreter', 'Latex');
160     xlabel('sample value', 'Interpreter', 'Latex');
161     ylabel('density', 'Interpreter', 'Latex');
162
163     % Plot (CDF)
164     hold on;
165     histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'cdf');
166     histogram(mv_GRV, 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'cdf');
167 end
168
169 % Save subplots
170 figure(1);
171 sgtitle('Mean Distribution', 'Interpreter', 'Latex');
172 subplot(3,3,9);
173 plot(0,0, 0,0);

```

```

174 axis off;
175 legend('M_n', 'Gaussian');
176 print(gcf, fullfile(f_path, '02_c-pdf'), '-dpng', '-r300');
177
178 figure(2);
179 sgtitle('Cumulative Mean Distribution', 'Interpreter', 'Latex');
180 subplot(3,3,9);
181 plot(0,0, 0,0);
182 axis off;
183 legend('M_n', 'Gaussian');
184 print(gcf, fullfile(f_path, '02_c-cdf'), '-dpng', '-r300');
185
186 %% Section 2.d.a
187
188 % Initializations
189 n = [1,2,3,4,5,10,20,30]; % samples
190 tosses = [2 2 4 4 1 3 5];
191 samples = 10000;
192 M_n = zeros(samples, length(n)); % RVs
193 mean_sum = zeros([1 samples]);
194 min_val = zeros([1 length(n)]); % subplots
195 max_val = zeros([1 length(n)]);
196 sample_x = zeros(100, length(n));
197 mean_n = zeros([1 length(n)]);
198 sd_n = zeros([1 length(n)]);
199
200 % Generate mean for n biased RVs
201 for mean_ind = 1:length(n)
202     for sum_ind = 1:n(mean_ind)
203         toss = tosses(randi(length(tosses), [1 samples]));
204         mean_sum = mean_sum + toss;
205     end
206     M_n(:, mean_ind) = mean_sum./n(mean_ind);
207     % Reset for next iteration
208     mean_sum = zeros([1 samples]);
209 end
210
211 % Close all open figures
212 close all;
213
214 % PDF & CDF subplots
215 for plot_ind = 1:length(n)
216     % Parameters
217     min_val(plot_ind) = min(M_n(:, plot_ind));
218     max_val(plot_ind) = max(M_n(:, plot_ind));
219     sample_x(:, plot_ind) = linspace(min_val(plot_ind), max_val(
        plot_ind));
220     mean_n(plot_ind) = mean(M_n(:, plot_ind));

```

```

221     sd_n(plot_ind)          = std(M_n(:, plot_ind));
222
223 % Formatting (PDF)
224 figure(1);
225 subplot(4, 2, plot_ind);
226 title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
227 title(title_string, 'Interpreter', 'Latex');
228 xlabel('sample value', 'Interpreter', 'Latex');
229 ylabel('density', 'Interpreter', 'Latex');
230
231 % Plot (PDF)
232 hold on;
233 histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
    Normalization', 'pdf');
234 y_pdf = (1/(sd_n(plot_ind).*sqrt(2*pi))).*exp(-(1/2)*((
    sample_x(:, plot_ind)-mean_n(plot_ind))./sd_n(plot_ind))
    .^2);
235 plot(sample_x(:, plot_ind), y_pdf);
236
237 % Formatting (CDF)
238 figure(2);
239 subplot(4, 2, plot_ind);
240 title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
241 title(title_string, 'Interpreter', 'Latex');
242 xlabel('sample value', 'Interpreter', 'Latex');
243 ylabel('density', 'Interpreter', 'Latex');
244
245 % Plot (CDF)
246 hold on;
247 histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
    Normalization', 'cdf');
248 y_cdf = cdf('Normal', sample_x(:, plot_ind), mean_n(plot_ind
    ), sd_n(plot_ind));
249 plot(sample_x(:, plot_ind), y_cdf);
250 end
251
252 % Save subplots
253 figure(1);
254 sgtitle('Mean Distribution', 'Interpreter', 'Latex');
255 print(gcf, fullfile(f_path, '02_d_a_pdf'), '-dpng', '-r300');
256
257 figure(2);
258 sgtitle('Cumulative Mean Distribution', 'Interpreter', 'Latex');
259 print(gcf, fullfile(f_path, '02_d_a_cdf'), '-dpng', '-r300');
260
261 %% Section 2.d.c
262
263 % PDF & CDF subplots

```

```

264 for plot_ind = 1:length(n)
265     % Multivariate Gaussian
266     mv_GRV = mvnrnd(mean_n(plot_ind), sd_n(plot_ind), length(
        sample_x(:, plot_ind)));
267
268     % Formatting (PDF)
269     figure(1);
270     subplot(3, 3, plot_ind);
271     title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
272     title(title_string, 'Interpreter', 'Latex');
273     xlabel('sample value', 'Interpreter', 'Latex');
274     ylabel('density', 'Interpreter', 'Latex');
275
276     % Plot (PDF)
277     hold on;
278     histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'pdf');
279     histogram(mv_GRV, 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'pdf');
280
281     % Formatting (CDF)
282     figure(2);
283     subplot(3, 3, plot_ind);
284     title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
285     title(title_string, 'Interpreter', 'Latex');
286     xlabel('sample value', 'Interpreter', 'Latex');
287     ylabel('density', 'Interpreter', 'Latex');
288
289     % Plot (CDF)
290     hold on;
291     histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'cdf');
292     histogram(mv_GRV, 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'cdf');
293 end
294
295 % Save subplots
296 figure(1);
297 sgtitle('Mean Distribution', 'Interpreter', 'Latex');
298 subplot(3,3,9);
299 plot(0,0, 0,0);
300 axis off;
301 legend('M_n', 'Gaussian');
302 print(gcf, fullfile(f_path, '02_d_c_pdf'), '-dpng', '-r300');
303
304 figure(2);
305 sgtitle('Cumulative Mean Distribution', 'Interpreter', 'Latex');
306 subplot(3,3,9);

```

```

307 plot(0,0, 0,0);
308 axis off;
309 legend('M_n', 'Gaussian');
310 print(gcf, fullfile(f_path, '02_d_c_cdf'), '-dpng', '-r300');
311
312 %% Section 3.b
313
314 % Parameters
315 bd_mu_0 = [9;10];
316 bd_mu_1 = [6;7];
317 bd_sigma = [1.15,0.1;0.1,0.5];
318
319 % Classification inequality
320 b_class_vec = 2*(bd_mu_1-bd_mu_0).'*bd_sigma^(-1)*data_2.';
321 b_class_const = bd_mu_0.'*bd_sigma^(-1)*bd_mu_0-bd_mu_1.'*
    bd_sigma^(-1)*bd_mu_1;
322 b_class_ans = b_class_vec+b_class_const;
323
324 % Proportion of samples from class 0
325 mean(b_class_ans >= 0); % 0.5017
326
327 % Scatter plot data
328 b_class_0_ind = b_class_ans >= 0;
329 b_class_1_ind = b_class_ans < 0;
330 b_class_0_vec = data_2(b_class_0_ind, :);
331 b_class_1_vec = data_2(b_class_1_ind, :);
332
333 % Scatter plots
334 hold on;
335 scatter(b_class_0_vec(:, 1), b_class_0_vec(:, 2), 'r');
336 scatter(b_class_1_vec(:, 1), b_class_1_vec(:, 2), 'b');
337
338 % Contour
339 ineq_f_b = @(x,y) 2*(bd_mu_1-bd_mu_0).'*bd_sigma^(-1)*[x;y]+
    b_class_const;
340 fcontour(ineq_f_b, [0 14 4 13], 'LevelList', 0);
341
342 % Formatting
343 title('Petal Length vs Petal Width', 'Interpreter', 'Latex');
344 xlabel('Petal Length', 'Interpreter', 'Latex');
345 ylabel('Petal Width', 'Interpreter', 'Latex');
346 legend('class 0', 'class 1', '$\vec{b}^T \cdot \vec{x} + a = 0$', 'Interpreter', 'Latex', 'Location', 'Best');
347 print(gcf, fullfile(f_path, '03_b_scatter'), '-dpng', '-r300');
348
349 %% Section 3.d
350
351 % Parameter

```

```

352 d_prob = 0.05;
353
354 % Classification inequality
355 d_class_vec = 2*(bd_mu_1-bd_mu_0).'*bd_sigma^(-1)*data_2.';
356 d_class_const = (bd_mu_0.'*bd_sigma^(-1)*bd_mu_0-bd_mu_1.'*
    bd_sigma^(-1)*bd_mu_1)+log((1-d_prob)/d_prob);
357 d_class_ans = d_class_vec+d_class_const;
358
359 % Proportion of samples from class 0
360 mean(d_class_ans >= 0); % 0.5085
361
362 % Scatter plot data
363 d_class_0_ind = d_class_ans >= 0;
364 d_class_1_ind = d_class_ans < 0;
365 d_class_0_vec = data_2(d_class_0_ind, :);
366 d_class_1_vec = data_2(d_class_1_ind, :);
367
368 % Scatter plots
369 hold on;
370 scatter(d_class_0_vec(:, 1), d_class_0_vec(:, 2), 'r');
371 scatter(d_class_1_vec(:, 1), d_class_1_vec(:, 2), 'b');
372
373 % Contour
374 ineq_f_d = @(x,y) 2*(bd_mu_1-bd_mu_0).'*bd_sigma^(-1)*[x;y]+
    d_class_const;
375 fcontour(ineq_f_d, [0 14 4 13], 'LevelList', 0);
376
377 % Formatting
378 title('Petal Length vs Petal Width', 'Interpreter', 'Latex');
379 xlabel('Petal Length', 'Interpreter', 'Latex');
380 ylabel('Petal Width', 'Interpreter', 'Latex');
381 legend('class 0', 'class 1', '$\vec{b}^T\cdot\vec{x}+a=0$', 'Interpreter', 'Latex', 'Location', 'Best');
382 print(gcf, fullfile(f_path, '03_d_scatter'), '-dpng', '-r300');
383
384 %% Section 3.f
385
386 % Parameter
387 f_mu_0 = [9;10];
388 f_mu_1 = [6;7];
389 f_sigma_0 = [1.15,0.1;0.1,0.5];
390 f_sigma_1 = [0.2,0.3;0.3,2];
391 f_prob = 0.5;
392
393 % Initializations
394 f_class_ans = zeros(length(data_3), 1);
395
396 % Classification inequality

```



```

397 f_class_lin_vec = 2*(f_mu_1.'*f_sigma_1^(-1)-f_mu_0.'*f_sigma_0
      ^(-1))*data_3.';
398 f_class_det      = det(f_sigma_0)^(-1/2)*det(f_sigma_1)^(-1/2);
399 f_class_const    = (f_mu_0.'*f_sigma_0^(-1)*f_mu_0-f_mu_1.'*
      f_sigma_1^(-1)*f_mu_1)...
400                  +log(((1-f_prob)/f_prob)*f_class_det);
401
402 % Inequality vector
403 for sample_ind = 1:length(data_3)
404     f_class_quad_vec = data_3(sample_ind,:)*(f_sigma_0^(-1)-
      f_sigma_1^(-1))*data_3(sample_ind,:).';
405     f_class_ans(sample_ind) = f_class_quad_vec+f_class_lin_vec(
      sample_ind)+f_class_const;
406 end
407
408 % Proportion of samples from class 0
409 mean(f_class_ans >= 0); % 0.5051
410
411 % Scatter plot data
412 f_class_0_ind = f_class_ans >= 0;
413 f_class_1_ind = f_class_ans < 0;
414 f_class_0_vec = data_3(f_class_0_ind, :);
415 f_class_1_vec = data_3(f_class_1_ind, :);
416
417 % Scatter plots
418 hold on;
419 scatter(f_class_0_vec(:, 1), f_class_0_vec(:, 2), 'r');
420 scatter(f_class_1_vec(:, 1), f_class_1_vec(:, 2), 'b');
421
422 % Contour
423 ineq_f_d = @(x,y) [x,y]*(f_sigma_0^(-1)-f_sigma_1^(-1))*[x;y]...
424               +2*(f_mu_1.'*f_sigma_1^(-1)-f_mu_0.'*f_sigma_0^(-1))*[x;y]+
      f_class_const;
425 fcontour(ineq_f_d, [0 14 4 13], 'LevelList', 0);
426
427 % Formatting
428 title('Petal Length vs Petal Width', 'Interpreter', 'Latex');
429 xlabel('Petal Length', 'Interpreter', 'Latex');
430 ylabel('Petal Width', 'Interpreter', 'Latex');
431 legend('class 0', 'class 1', '$\vec{x}^{TC}\vec{x}+\vec{b}^T\cdot\vec{x}+a=0$',
      'Interpreter', 'Latex', 'Location', 'Best');
432 print(gcf, fullfile(f_path, '03_f_scatter'), '-dpng', '-r300');
433
434 %% Functions
435
436 % sample mean estimator
437 function mu_N = sample_mean(data, samples)
438     mu_N = zeros(1, length(samples));

```

```
439     for i = 1:length(samples)
440         mu_N(i) = mean(data(1:samples(i)));
441     end
442 end
```