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}

$$
\begin{aligned}
E_{MMSE} &= \mathrm{E}[\sum_{i \in K_{miss}} (X_i - a_i)^2] && (1)\\
\frac{E_{MMSE}}{da_i} &= \mathrm{E}[\sum_{i \in K_{miss}} -2(X_i - a_i)] && (2)\\
&= \mathrm{E}[\sum_{i \in K_{miss}} (2a_i - 2X_i)] && (3)\\
&= \sum_{i \in K_{miss}} (2\mathrm{E}[a_i] - 2\mathrm{E}[X_i]) && \text{by the linearity of expectation} \quad (4)\\
&= \sum_{i \in K_{miss}} (2\mathrm{E}[\mu] - 2\mathrm{E}[X_i]) && \text{by substituting } \mu \text{ for } a_i \quad (5)\\
&= \sum_{i \in K_{miss}} (2\mu - 2\mathrm{E}[X_i]) && \text{by the expected value of a constant} \quad (6)\\
&= \sum_{i \in K_{miss}} (2\mu - 2\mu) && \text{since the expectation of each i.i.d. RV is } \mu \quad (7)\\
&= 0 && (8)
\end{aligned}
$$

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

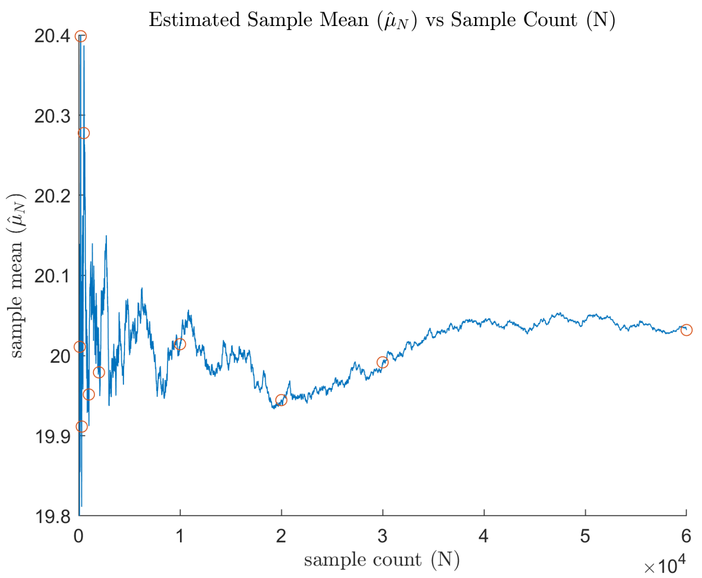

Figure 1: Sample means estimated from 10 to 60000 samples.

$\hat{\mu}_N$ behaves erratically until the sample count reaches 100 samples, then fluctuates around a value of 20 $\pm$ 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

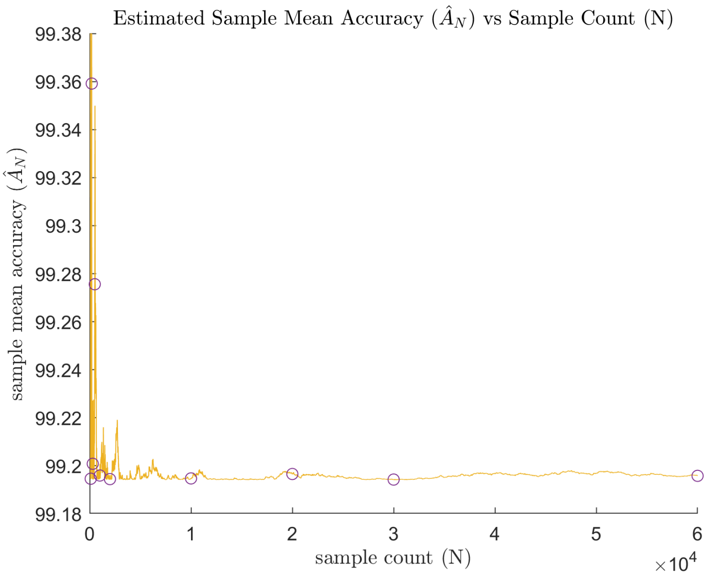

Figure 2: Sample mean accuracy estimated from 10 to 60000 samples.

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

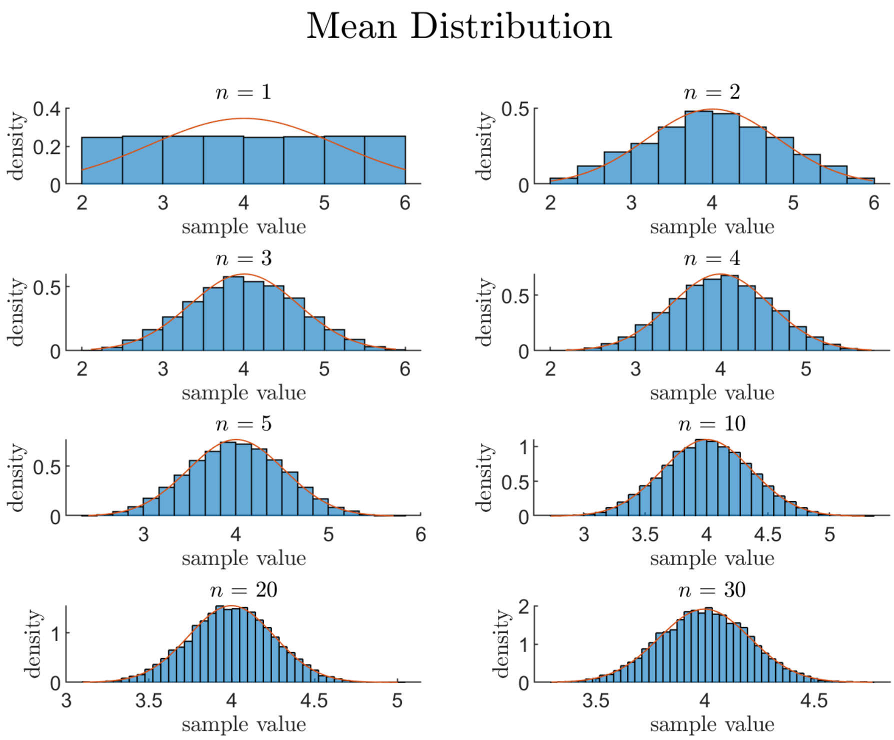

Figure 3: PDF of M_n for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

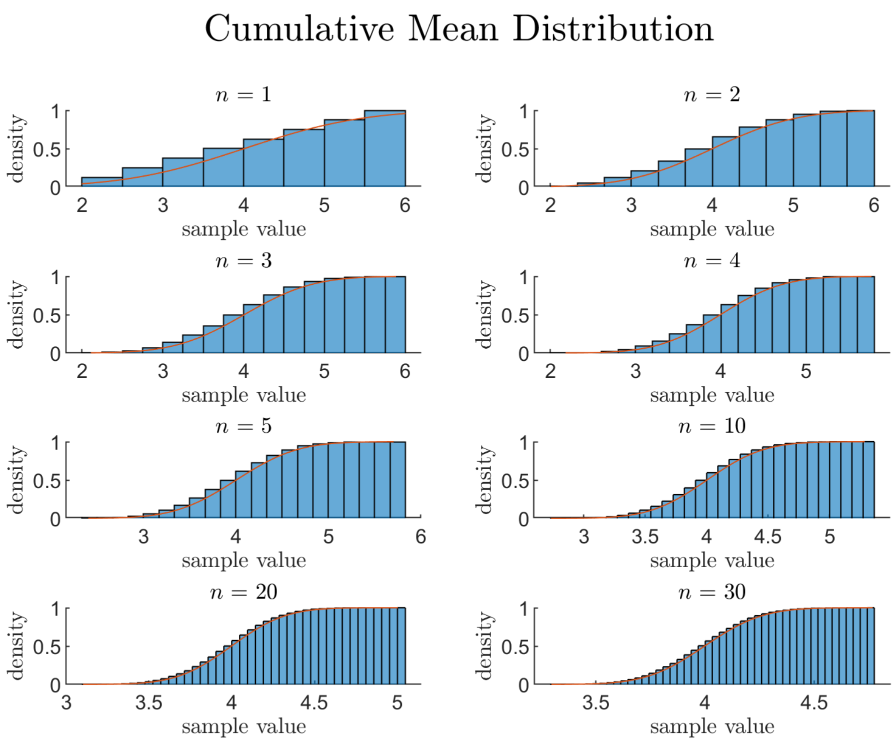

Figure 4: CDF of M_n for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

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 σ^2 $\forall\ i \in \{1, 2, 3, \dots, n\}$.

2.2.2 Mean of M_n

$$\begin{aligned}
\mathrm{E}[M_n] &= \mathrm{E}[\frac{X_1 + X_2 + \cdots + X_n}{n}] && \text{by the definition of the expectation} \quad (1)\\
&= \frac{\mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_n]}{\mathrm{E}[n]} && \text{by the linearity of expectation} \quad (2)\\
&= \frac{\mu + \mu + \cdots + \mu}{\mathrm{E}[n]} && \text{since } \mathrm{E}[X_i] = \mu\ \forall\ i \quad (3)\\
&= \frac{\mu + \mu + \cdots + \mu}{n} && \text{by the expected value of a constant} \quad (4)\\
&= \frac{n\mu}{n} && (5)\\
&= \mu && (6)
\end{aligned}$$

2.2.3 Variance of M_n

$$\begin{aligned}
Var(M_n) &= Var(\frac{X_1 + X_2 + \cdots + X_n}{n}) && \text{by the definition of } M_n \quad (1)\\
&= \frac{1}{n^2} Var(X_1 + X_2 + \cdots + X_n) && \text{since } Var(aX) = a^2 Var(X) \quad (2)\\
&= \frac{1}{n^2}[Var(X_1) + Var(X_2) + \cdots + Var(X_n)] && \text{since } X_j \perp\!\!\!\perp X_k\ \forall\ j \neq k \quad (3)\\
&= \frac{1}{n^2}[\sigma^2 + \sigma^2 + \cdots + \sigma^2] && \text{since } Var[X_i] = \sigma^2\ \forall\ i \quad (4)\\
&= \frac{n\sigma^2}{n^2} && (5)\\
&= \frac{\sigma^2}{n} && (6)
\end{aligned}$$

2.3 Multivariate Gaussian RV

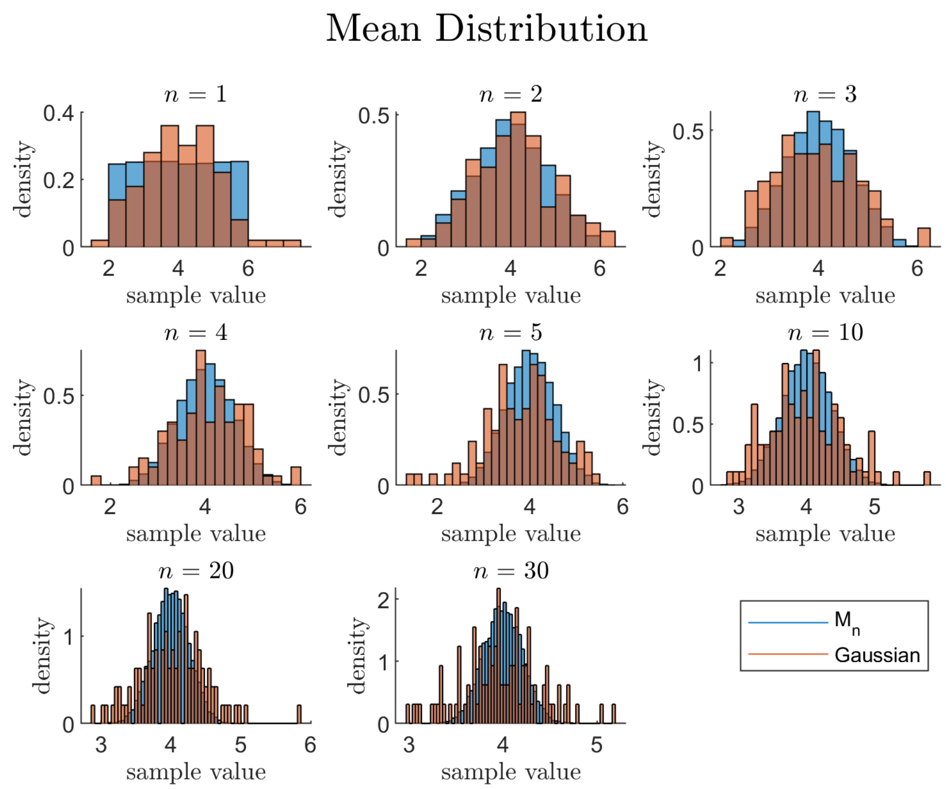

Figure 5: PDF of M_n and a Multivariate Gaussian for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

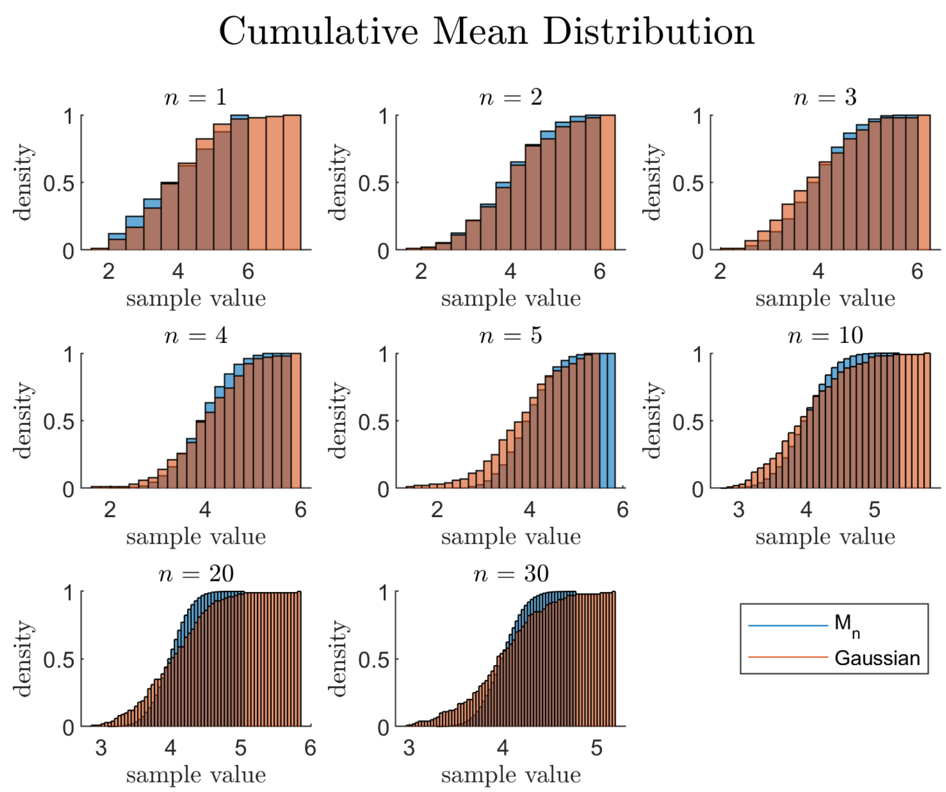

Figure 6: CDF of M_n and a Multivariate Gaussian for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

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

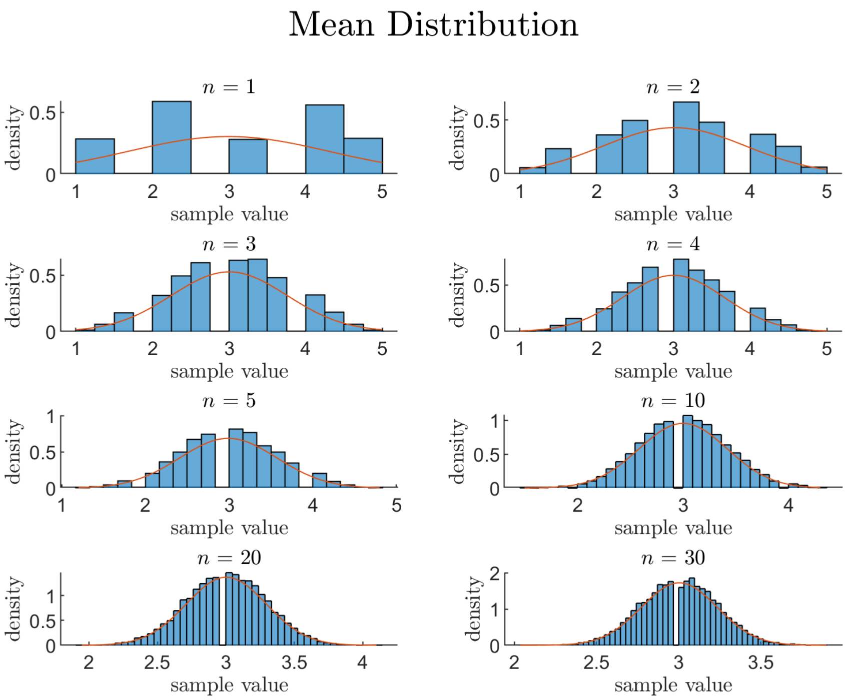

Figure 7: PDF of M_n for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

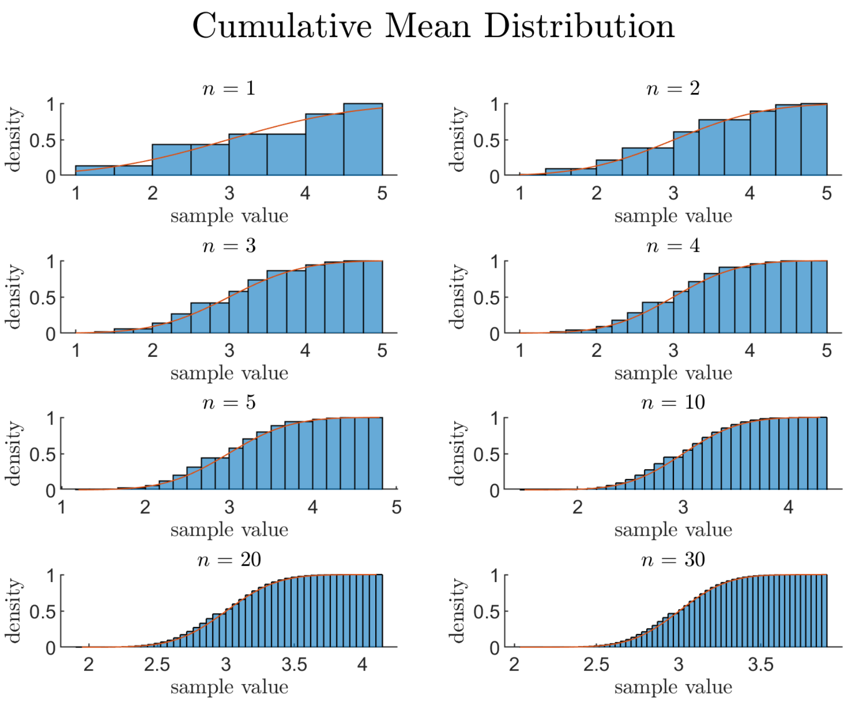

Figure 8: CDF of M_n for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

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

$$
\begin{aligned}
\mathrm{E}[X_i] &= \sum_{i=1}^{5} x_i P(x = x_i) && \text{by the definition of expectation} && (1)\\
&= 1 \cdot p + 2 \cdot 2p + 3 \cdot p + 4 \cdot 2p + 5 \cdot p && \text{by the pmf of an unfair 5-sided dice} && (2)\\
&= 21 \cdot p && && (3)\\
&= 21 \cdot \frac{1}{7} && \text{since } p = \frac{1}{7} \text{ for } \sum_{i=1}^{5} \mathrm{P}(x = x_i) = 1 && (4)\\
&= 3 && && (5)
\end{aligned}
$$

2.4.2.2 Variance of X_i

$$
\begin{aligned}
\mathrm{E}[X_i^2] &= \sum_{i=1}^{5} x_i^2 P(x = x_i) && (1)\\
&= 1^2 \cdot p + 2^2 \cdot 2p + 3^2 \cdot p + 4^2 \cdot 2p + 5^2 \cdot p && (2)\\
&= 75 \cdot p && (3)\\
&= \frac{75}{7} && (4)
\end{aligned}
$$

$$
\begin{aligned}
Var(X_i) &= \mathrm{E}[X_i^2] - (\mathrm{E}[X_i])^2 && \text{by the definition of variance} && (1)\\
&= \frac{75}{7} - 3^2 && && (2)\\
&= \frac{12}{7} && && (3)
\end{aligned}
$$

2.4.2.3 Mean of M_n

$$
\begin{aligned}
\mathrm{E}[M_n] &= \mathrm{E}[\frac{X_1 + X_2 + \cdots + X_n}{n}] && \text{by the definition of the expectation} && (1)\\
&= \frac{\mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_n]}{\mathrm{E}[n]} && \text{by the linearity of expectation} && (2)\\
&= \frac{3 + 3 + \cdots + 3}{\mathrm{E}[n]} && \text{since } \mathrm{E}[X_i] = 3 \ \forall\ i \in \{1, 2, 3, \ldots, n\} && (3)\\
&= \frac{3 + 3 + \cdots + 3}{n} && \text{by the expected value of a constant} && (4)\\
&= \frac{n \cdot 3}{n} && && (5)\\
&= 3 && && (6)
\end{aligned}
$$

2.4.2.4 Variance of M_n

$$
\begin{aligned}
Var(M_n) &= Var(\frac{X_1 + X_2 + \cdots + X_n}{n}) && \text{by the definition of } M_n \quad (1)\\
&= \frac{1}{n^2} Var(X_1 + X_2 + \cdots + X_n) && \text{since } Var(aX) = a^2 Var(X) \quad (2)\\
&= \frac{1}{n^2}[Var(X_1) + Var(X_2) + \cdots + Var(X_n)] && \text{since } X_j \perp\!\!\!\perp X_k \text{ for } j \neq k \quad (3)\\
&= \frac{1}{n^2}[\frac{12}{7} + \frac{12}{7} + \cdots + \frac{12}{7}] && \text{since } Var[X_i] = \frac{12}{7}\ \forall\ i \quad (4)\\
&= \frac{12n}{7n^2} && (5)\\
&= \frac{12}{7n} && (6)
\end{aligned}
$$

2.4.3 Multivariate Gaussian RV

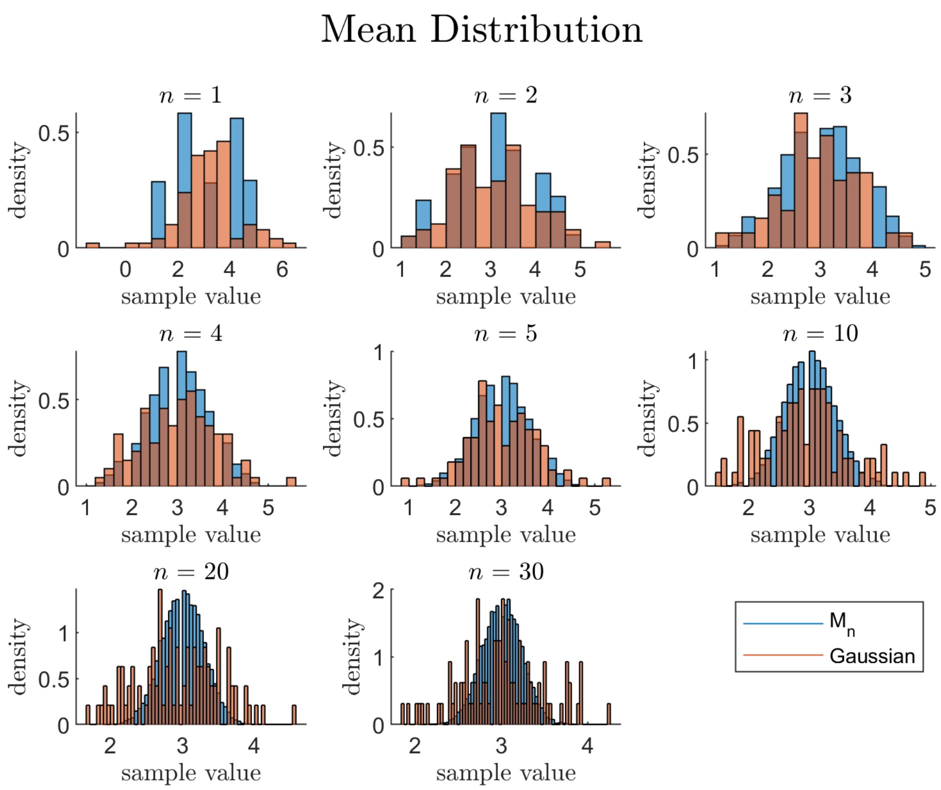

Figure 9: PDF of M_n and a Multivariate Gaussian for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

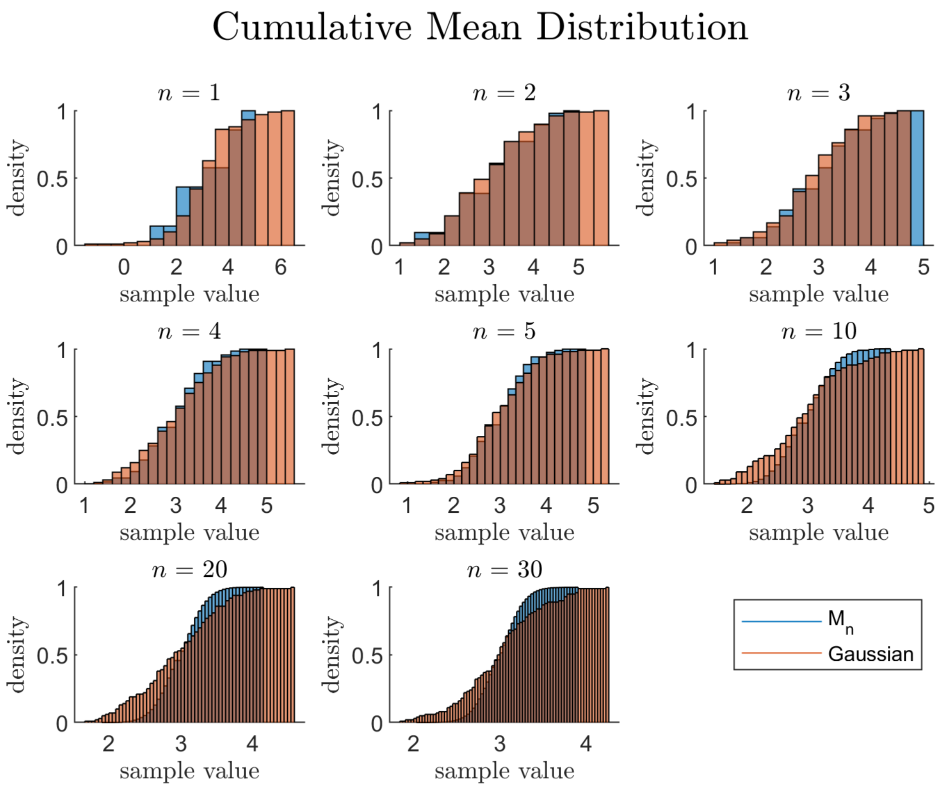

Figure 10: CDF of M_n and a Multivariate Gaussian for $n \in \{1, 2, 3, 4, 5, 10, 20, 30\}$.

3 Gaussian Discriminant Analysis

3.1 Classification Rule for $\sum = \sum_0 = \sum_1$ and $p = \frac{1}{2}$

$$
\begin{aligned}
P(y=0|\vec{x}) &\geq P(y=1|\vec{x}) && (1)\\
\frac{P(\vec{x}|y=0)P(y=0)}{P(\vec{x})} &\geq \frac{P(\vec{x}|y=1)P(y=1)}{P(\vec{x})} && \text{by Bayes' Rule} \quad (2)\\
P(\vec{x}|y=0)P(y=0) &\geq P(\vec{x}|y=1)P(y=1) && (3)\\
\frac{1}{2}P(\vec{x}|y=0) &\geq \frac{1}{2}P(\vec{x}|y=1) && \text{since } y = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases} \quad (4)\\
P(\vec{x}|y=0) &\geq P(\vec{x}|y=1) && (5)\\
f_{X,y=0}(\vec{x}) &\geq f_{X,y=1}(\vec{x}) && (6)
\end{aligned}
$$

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

$$
\begin{aligned}
K &= \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & Var(X_2) & \dots & Cov(X_2,X_n) \\ \vdots & \vdots & & \vdots \\ Cov(X_n,X_1) & \dots & & Var(X_n) \end{bmatrix} && (1)\\
&= \begin{bmatrix} Var(X_1) & 0 & \dots & 0 \\ 0 & Var(X_2) & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \dots & & Var(X_n) \end{bmatrix} && \text{since } \rho = 0 \text{ for } X_j \perp\!\!\!\perp X_k \text{ for } j \neq k \quad (2)\\
&= \begin{bmatrix} \sum & 0 & \dots & 0 \\ 0 & \sum & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \dots & & \sum \end{bmatrix} && (3)
\end{aligned}
$$

$$
\begin{aligned}
K^{-1} &= \frac{1}{\sum} \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \dots & & 1 \end{bmatrix} && (1)\\
&= \frac{I}{\sum} && (2)
\end{aligned}
$$

$$
\begin{aligned}
f_X(\vec{x}) &= \frac{\exp\{-\frac{1}{2}(\vec{x}-\mu)^T \frac{I}{\sum}(\vec{x}-\mu)\}}{(2\pi)^{\frac{n}{2}}|\sum I|^{\frac{1}{2}}} && \tag{1}\\
&= \frac{\exp\{-\frac{1}{2}(\vec{x}^T\Sigma^{-1}\vec{x}-\mu^T\Sigma^{-1}\vec{x}-x^T\Sigma^{-1}\mu+\mu^T\Sigma^{-1}\mu)\}}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}} && \text{cross multiply} \tag{2}
\end{aligned}
$$

$$
\begin{aligned}
ln(\exp) &= ln(\exp\{-\frac{1}{2}(\vec{x}^T\Sigma^{-1}\vec{x}-\mu^T\Sigma^{-1}\vec{x}-x^T\Sigma^{-1}\mu+\mu^T\Sigma^{-1}\mu)\}) && \tag{1}\\
&= -\frac{1}{2}(\vec{x}^T\Sigma^{-1}\vec{x}-\mu^T\Sigma^{-1}\vec{x}-x^T\Sigma^{-1}\mu+\mu^T\Sigma^{-1}\mu) && \tag{2}\\
&= -\frac{1}{2}(\vec{x}^T\Sigma^{-1}\vec{x}-2\mu^T\Sigma^{-1}\vec{x}+\mu^T\Sigma^{-1}\mu) && \text{since } \mu^T\Sigma^{-1}\vec{x}=x^T\Sigma^{-1}\mu \tag{3}
\end{aligned}
$$

$$
\begin{gathered}
f_{X,y=0}(\vec{x}) \geq f_{X,y=1}(\vec{x}) \tag{1}\\
-\frac{1}{2}(\vec{x}^T\Sigma^{-1}\vec{x}-2\mu_0^T\Sigma^{-1}\vec{x}+\mu_0^T\Sigma^{-1}\mu_0) \geq -\frac{1}{2}(\vec{x}^T\Sigma^{-1}\vec{x}-2\mu_1^T\Sigma^{-1}\vec{x}+\mu_1^T\Sigma^{-1}\mu_1) \tag{2}\\
\vec{x}^T\Sigma^{-1}\vec{x}-2\mu_0^T\Sigma^{-1}\vec{x}+\mu_0^T\Sigma^{-1}\mu_0 \geq \vec{x}^T\Sigma^{-1}\vec{x}-2\mu_1^T\Sigma^{-1}\vec{x}+\mu_1^T\Sigma^{-1}\mu_1 \tag{3}\\
-2\mu_0^T\Sigma^{-1}\vec{x}+2\mu_1^T\Sigma^{-1}\vec{x} \geq \mu_1^T\Sigma^{-1}\mu_1-\mu_0^T\Sigma^{-1}\mu_0 \tag{4}\\
[2(\mu_1-\mu_0)^T\Sigma^{-1}]\vec{x}+[\mu_0^T\Sigma^{-1}\mu_0-\mu_1^T\Sigma^{-1}\mu_1] \geq 0 \tag{5}
\end{gathered}
$$

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

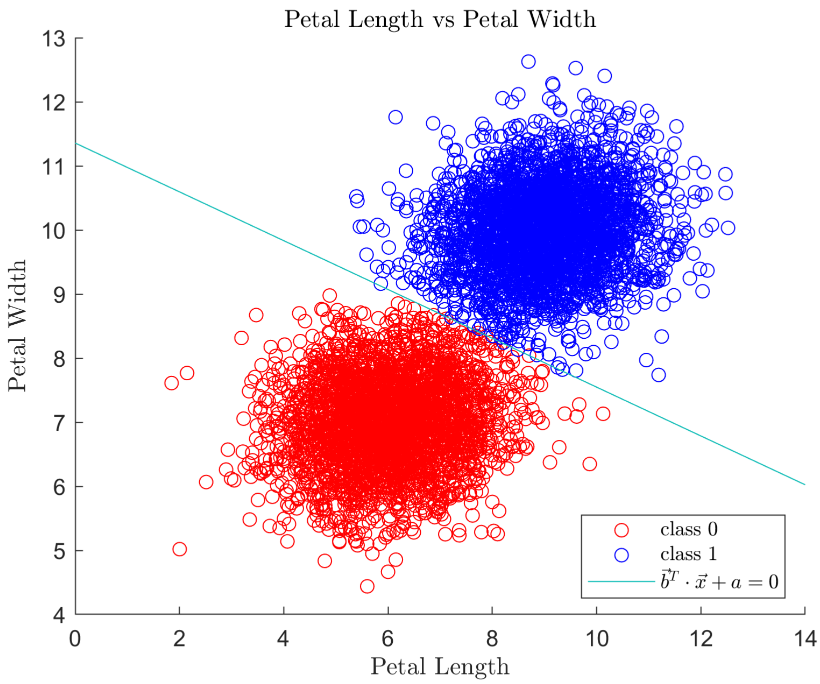

Figure 11: Scatter plot and contour line of two classes from a sample with the same covariance and based on the linear inequality with $p = \frac{1}{2}$. Based on the samples, 50.17% are categorized as class 0.

3.3 Classification Rule for $\sum = \sum_0 = \sum_1$ and a general p

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

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

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

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

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

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

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

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

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

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

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

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

3.4 Linear Inequality Contour for $\sum = \sum_0 = \sum_1$ and $p = 0.05$

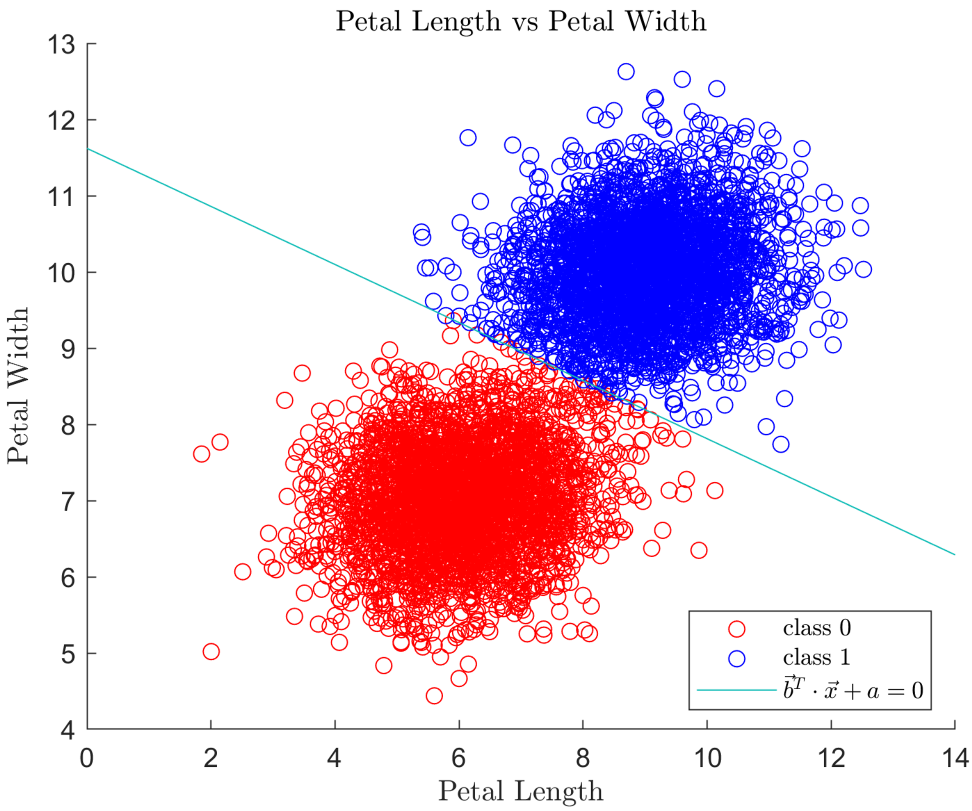

Figure 12: Scatter plot and contour line of two classes from a sample with the same covariance and based on the linear inequality with $p = 0.05$. Based on the samples, 50.85% are classified as class 0. Also, the change in p shifted the contour line towards the samples categorized as class 1. The result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

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

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

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

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

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

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

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

$$\frac{\frac{1-p}{p}\exp(\vec{x}^T\Sigma_0^{-1}\vec{x} - 2\mu_0^T\Sigma_0^{-1}\vec{x} + \mu_0^T\Sigma_0^{-1}\mu_0)}{|\sum_0|^{\frac{1}{2}}} \geq \frac{\exp(\vec{x}^T\Sigma_1^{-1}\vec{x} - 2\mu_1^T\Sigma_1^{-1}\vec{x} + \mu_1^T\Sigma_1^{-1}\mu_1)}{|\sum_1|^{\frac{1}{2}}} \tag{3}$$

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

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

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

3.6 Quadratic Inequality Contour

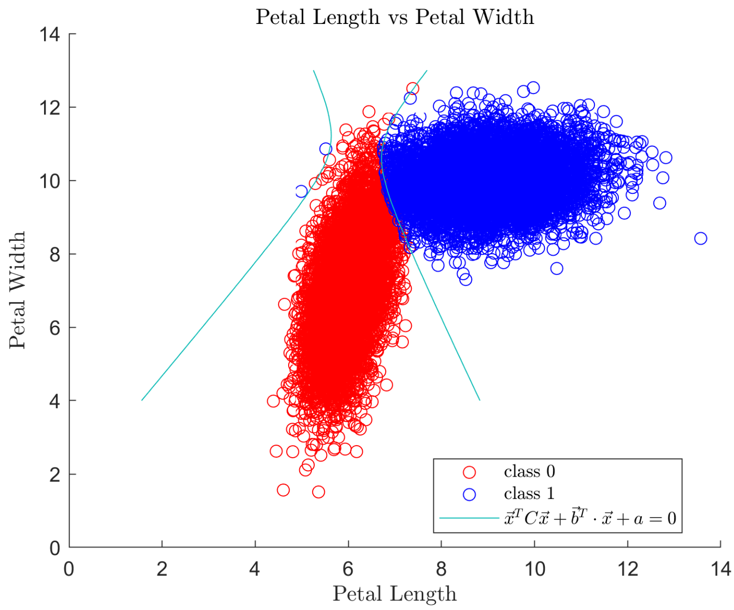

Figure 13: Scatter plot and contour line of two classes from a sample with distributions that do not necessarily have the same covariances and based on the quadratic inequality with $p = 0.5$. Based on the samples, 50.51% are classified as class 0.

4 Appendix

```
%% Read/store files
load data.txt;
load data_2.txt;
load data_3.txt;
f_path = 'D:\UCLA\Courses\EE 131A\Project\Plots';

%% Section 1

% Indices of missing and available data
K_miss      = isnan(data);
K_avail     = ~isnan(data);

% Filtered data
data_miss   = data(K_miss);
data_avail = data(K_avail);

%% Section 1.b

% Estimate sample mean
N_cont = 10:10:60000;
N_disc =
    [10,20,50,100,200,300,500,1000,2000,10000,20000,30000,60000];
mu_N_c = sample_mean(data_avail, N_cont);
mu_N_d = sample_mean(data_avail, N_disc);

% mu_N vs N plot
hold on;
plot(N_cont, mu_N_c);
scatter(N_disc, mu_N_d);
title('Estimated Sample Mean ($\hat{\mu}_N$) vs Sample Count (N)
    ', 'Interpreter', 'Latex');
xlabel('sample count (N)', 'Interpreter', 'Latex');
ylabel('sample mean ($\hat{\mu}_N$)', 'Interpreter', 'Latex');
ylim([19.8 20.4])
print(gcf, fullfile(f_path, '01_b'), '-dpng', '-r300');

%% Section 1.c

% Estimate sample mean accuracy
square_d    = (data_avail(:) - mu_N_d).^2;
square_c    = (data_avail(:) - mu_N_c).^2;
acc_d       = zeros(1, length(mu_N_d));
acc_c       = zeros(1, length(mu_N_c));

for mu_N_index = 1:length(mu_N_d)
    acc_d(mu_N_index) = sum(square_d(:,mu_N_index))/length(
```

```
        data_avail);
end

for mu_N_index = 1:length(mu_N_c)
    acc_c(mu_N_index) = sum(square_c(:,mu_N_index))/length(data_avail);
end

% A_N vs N plot
hold on;
plot(N_cont, acc_c);
scatter(N_disc, acc_d);
title('Estimated Sample Mean Accuracy ($\hat{A}_N$) vs Sample Count (N)', 'Interpreter', 'Latex');
xlabel('sample count (N)', 'Interpreter', 'Latex');
ylabel('sample mean accuracy ($\hat{A}_N$)', 'Interpreter', 'Latex');
xlim([0 60000]);
ylim([99.18 99.38]);
print(gcf, fullfile(f_path, '01_c'), '-dpng', '-r300');

%% Section 2.a

% Initializations
n          = [1,2,3,4,5,10,20,30];          % samples
samples    = 10000;
M_n        = zeros(samples, length(n));    % RVs
mean_sum   = zeros([1 samples]);
min_val    = zeros([1 length(n)]);          % subplots
max_val    = zeros([1 length(n)]);
sample_x   = zeros(100, length(n));
mean_n     = zeros([1 length(n)]);
sd_n       = zeros([1 length(n)]);

% Generate mean for n RVs
for mean_ind = 1:length(n)
    for sum_ind = 1:n(mean_ind)
        mean_sum = mean_sum + 4*rand([1 samples])+2;
    end
    M_n(:, mean_ind) = mean_sum./n(mean_ind);
    % Reset for next iteration
    mean_sum = zeros([1 samples]);
end

% Close all open figures
close all;

% PDF & CDF subplots
```

```
for plot_ind = 1:length(n)
    % Parameters
    min_val(plot_ind)       = min(M_n(:, plot_ind));
    max_val(plot_ind)       = max(M_n(:, plot_ind));
    sample_x(:, plot_ind) = linspace(min_val(plot_ind), max_val(
        plot_ind));
    mean_n(plot_ind)        = mean(M_n(:, plot_ind));
    sd_n(plot_ind)          = std(M_n(:, plot_ind));

    % Formatting (PDF)
    figure(1);
    subplot(4, 2, plot_ind);
    title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
    title(title_string, 'Interpreter', 'Latex');
    xlabel('sample value', 'Interpreter', 'Latex');
    ylabel('density', 'Interpreter', 'Latex');

    % Plot (PDF)
    hold on;
    histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'pdf');
    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);
    plot(sample_x(:, plot_ind), y_pdf);

    % Formatting (CDF)
    figure(2);
    subplot(4, 2, plot_ind);
    title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
    title(title_string, 'Interpreter', 'Latex');
    xlabel('sample value', 'Interpreter', 'Latex');
    ylabel('density', 'Interpreter', 'Latex');

    % Plot (CDF)
    hold on;
    histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'cdf');
    y_cdf = cdf('Normal', sample_x(:, plot_ind), mean_n(plot_ind
        ), sd_n(plot_ind));
    plot(sample_x(:, plot_ind), y_cdf);
end

% Save subplots
figure(1);
sgtitle('Mean Distribution', 'Interpreter', 'Latex');
print(gcf, fullfile(f_path, '02_a_pdf'), '-dpng', '-r300');
```

```
figure(2);
sgtitle('Cumulative Mean Distribution', 'Interpreter', 'Latex');
print(gcf, fullfile(f_path, '02_a_cdf'), '-dpng', '-r300');

%% Section 2.c

% PDF & CDF subplots
for plot_ind = 1:length(n)
    % Multivariate Gaussian
    mv_GRV = mvnrnd(mean_n(plot_ind), sd_n(plot_ind), length(
        sample_x(:, plot_ind)));

    % Formatting (PDF)
    figure(1);
    subplot(3, 3, plot_ind);
    title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
    title(title_string, 'Interpreter', 'Latex');
    xlabel('sample value', 'Interpreter', 'Latex');
    ylabel('density', 'Interpreter', 'Latex');

    % Plot (PDF)
    hold on;
    histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'pdf');
    histogram(mv_GRV, 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'pdf');

    % Formatting (CDF)
    figure(2);
    subplot(3, 3, plot_ind);
    title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
    title(title_string, 'Interpreter', 'Latex');
    xlabel('sample value', 'Interpreter', 'Latex');
    ylabel('density', 'Interpreter', 'Latex');

    % Plot (CDF)
    hold on;
    histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'cdf');
    histogram(mv_GRV, 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'cdf');
end

% Save subplots
figure(1);
sgtitle('Mean Distribution', 'Interpreter', 'Latex');
subplot(3,3,9);
plot(0,0,  0,0);
```

```
axis off;
legend('M_n', 'Gaussian');
print(gcf, fullfile(f_path, '02_c_pdf'), '-dpng', '-r300');

figure(2);
sgtitle('Cumulative Mean Distribution', 'Interpreter', 'Latex');
subplot(3,3,9);
plot(0,0,  0,0);
axis off;
legend('M_n', 'Gaussian');
print(gcf, fullfile(f_path, '02_c_cdf'), '-dpng', '-r300');

%% Section 2.d.a

% Initializations
n          = [1,2,3,4,5,10,20,30];        % samples
tosses     = [2 2 4 4 1 3 5];
samples    = 10000;
M_n        = zeros(samples, length(n));   % RVs
mean_sum   = zeros([1 samples]);
min_val    = zeros([1 length(n)]);        % subplots
max_val    = zeros([1 length(n)]);
sample_x   = zeros(100, length(n));
mean_n     = zeros([1 length(n)]);
sd_n       = zeros([1 length(n)]);

% Generate mean for n biased RVs
for mean_ind = 1:length(n)
    for sum_ind = 1:n(mean_ind)
        toss = tosses(randi(length(tosses), [1 samples]));
        mean_sum = mean_sum + toss;
    end
    M_n(:, mean_ind) = mean_sum./n(mean_ind);
    % Reset for next iteration
    mean_sum = zeros([1 samples]);
end

% Close all open figures
close all;

% PDF & CDF subplots
for plot_ind = 1:length(n)
    % Parameters
    min_val(plot_ind)       = min(M_n(:, plot_ind));
    max_val(plot_ind)       = max(M_n(:, plot_ind));
    sample_x(:, plot_ind) = linspace(min_val(plot_ind), max_val(
        plot_ind));
    mean_n(plot_ind)        = mean(M_n(:, plot_ind));
```

```
    sd_n(plot_ind)          = std(M_n(:, plot_ind));

    % Formatting (PDF)
    figure(1);
    subplot(4, 2, plot_ind);
    title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
    title(title_string, 'Interpreter', 'Latex');
    xlabel('sample value', 'Interpreter', 'Latex');
    ylabel('density', 'Interpreter', 'Latex');

    % Plot (PDF)
    hold on;
    histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'pdf');
    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);
    plot(sample_x(:, plot_ind), y_pdf);

    % Formatting (CDF)
    figure(2);
    subplot(4, 2, plot_ind);
    title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
    title(title_string, 'Interpreter', 'Latex');
    xlabel('sample value', 'Interpreter', 'Latex');
    ylabel('density', 'Interpreter', 'Latex');

    % Plot (CDF)
    hold on;
    histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'cdf');
    y_cdf = cdf('Normal', sample_x(:, plot_ind), mean_n(plot_ind
        ), sd_n(plot_ind));
    plot(sample_x(:, plot_ind), y_cdf);
end

% Save subplots
figure(1);
sgtitle('Mean Distribution', 'Interpreter', 'Latex');
print(gcf, fullfile(f_path, '02_d_a_pdf'), '-dpng', '-r300');

figure(2);
sgtitle('Cumulative Mean Distribution', 'Interpreter', 'Latex');
print(gcf, fullfile(f_path, '02_d_a_cdf'), '-dpng', '-r300');

%% Section 2.d.c

% PDF & CDF subplots
```

```
for plot_ind = 1:length(n)
    % Multivariate Gaussian
    mv_GRV = mvnrnd(mean_n(plot_ind), sd_n(plot_ind), length(
        sample_x(:, plot_ind)));

    % Formatting (PDF)
    figure(1);
    subplot(3, 3, plot_ind);
    title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
    title(title_string, 'Interpreter', 'Latex');
    xlabel('sample value', 'Interpreter', 'Latex');
    ylabel('density', 'Interpreter', 'Latex');

    % Plot (PDF)
    hold on;
    histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'pdf');
    histogram(mv_GRV, 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'pdf');

    % Formatting (CDF)
    figure(2);
    subplot(3, 3, plot_ind);
    title_string = strcat({'$n$ = '}, num2str(n(plot_ind)));
    title(title_string, 'Interpreter', 'Latex');
    xlabel('sample value', 'Interpreter', 'Latex');
    ylabel('density', 'Interpreter', 'Latex');

    % Plot (CDF)
    hold on;
    histogram(M_n(:, plot_ind), 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'cdf');
    histogram(mv_GRV, 'BinWidth', 1/(n(plot_ind)+1), '
        Normalization', 'cdf');
end

% Save subplots
figure(1);
sgtitle('Mean Distribution', 'Interpreter', 'Latex');
subplot(3,3,9);
plot(0,0,  0,0);
axis off;
legend('M_n', 'Gaussian');
print(gcf, fullfile(f_path, '02_d_c_pdf'), '-dpng', '-r300');

figure(2);
sgtitle('Cumulative Mean Distribution', 'Interpreter', 'Latex');
subplot(3,3,9);
```

```
plot(0,0,  0,0);
axis off;
legend('M_n', 'Gaussian');
print(gcf, fullfile(f_path, '02_d_c_cdf'), '-dpng', '-r300');

%% Section 3.b

% Parameters
bd_mu_0      = [9;10];
bd_mu_1      = [6;7];
bd_sigma     = [1.15,0.1;0.1,0.5];

% Classification inequality
b_class_vec   = 2*(bd_mu_1-bd_mu_0).'*bd_sigma^(-1)*data_2.';
b_class_const = bd_mu_0.'*bd_sigma^(-1)*bd_mu_0-bd_mu_1.'*bd_sigma^(-1)*bd_mu_1;
b_class_ans   = b_class_vec+b_class_const;

% Proportion of samples from class 0
mean(b_class_ans >= 0); % 0.5017

% Scatter plot data
b_class_0_ind = b_class_ans >= 0;
b_class_1_ind = b_class_ans <  0;
b_class_0_vec = data_2(b_class_0_ind, :);
b_class_1_vec = data_2(b_class_1_ind, :);

% Scatter plots
hold on;
scatter(b_class_0_vec(:, 1), b_class_0_vec(:, 2), 'r');
scatter(b_class_1_vec(:, 1), b_class_1_vec(:, 2), 'b');

% Contour
ineq_f_b = @(x,y) 2*(bd_mu_1-bd_mu_0).'*bd_sigma^(-1)*[x;y]+b_class_const;
fcontour(ineq_f_b, [0 14 4 13], 'LevelList', 0);

% Formatting
title('Petal Length vs Petal Width', 'Interpreter', 'Latex');
xlabel('Petal Length', 'Interpreter', 'Latex');
ylabel('Petal Width', 'Interpreter', 'Latex');
legend('class 0', 'class 1', '$\vec{b}^T\cdot\vec{x}+a=0$', 'Interpreter', 'Latex', 'Location', 'Best');
print(gcf, fullfile(f_path, '03_b_scatter'), '-dpng', '-r300');

%% Section 3.d

% Parameter
```

```
d_prob = 0.05;

% Classification inequality
d_class_vec   = 2*(bd_mu_1-bd_mu_0).'*bd_sigma^(-1)*data_2.';
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);
d_class_ans   = d_class_vec+d_class_const;

% Proportion of samples from class 0
mean(d_class_ans >= 0); % 0.5085

% Scatter plot data
d_class_0_ind = d_class_ans >= 0;
d_class_1_ind = d_class_ans <  0;
d_class_0_vec = data_2(d_class_0_ind, :);
d_class_1_vec = data_2(d_class_1_ind, :);

% Scatter plots
hold on;
scatter(d_class_0_vec(:, 1), d_class_0_vec(:, 2), 'r');
scatter(d_class_1_vec(:, 1), d_class_1_vec(:, 2), 'b');

% Contour
ineq_f_d = @(x,y) 2*(bd_mu_1-bd_mu_0).'*bd_sigma^(-1)*[x;y]+
    d_class_const;
fcontour(ineq_f_d, [0 14 4 13], 'LevelList', 0);

% Formatting
title('Petal Length vs Petal Width', 'Interpreter', 'Latex');
xlabel('Petal Length', 'Interpreter', 'Latex');
ylabel('Petal Width', 'Interpreter', 'Latex');
legend('class 0', 'class 1', '$\vec{b}^T\cdot\vec{x}+a=0$', '
    Interpreter', 'Latex', 'Location', 'Best');
print(gcf, fullfile(f_path, '03_d_scatter'), '-dpng', '-r300');

%% Section 3.f

% Parameter
f_mu_0     = [9;10];
f_mu_1     = [6;7];
f_sigma_0 = [1.15,0.1;0.1,0.5];
f_sigma_1 = [0.2,0.3;0.3,2];
f_prob     = 0.5;

% Initializations
f_class_ans = zeros(length(data_3), 1);

% Classification inequality
```

```
f_class_lin_vec   = 2*(f_mu_1.'*f_sigma_1^(-1)-f_mu_0.'*f_sigma_0^(-1))*data_3.';
f_class_det       = det(f_sigma_0)^(-1/2)*det(f_sigma_1)^(-1/2);
f_class_const     = (f_mu_0.'*f_sigma_0^(-1)*f_mu_0-f_mu_1.'*f_sigma_1^(-1)*f_mu_1)...
                        +log(((1-f_prob)/f_prob)*f_class_det);

% Inequality vector
for sample_ind = 1:length(data_3)
    f_class_quad_vec = data_3(sample_ind,:)*(f_sigma_0^(-1)-f_sigma_1^(-1))*data_3(sample_ind,:).';
    f_class_ans(sample_ind) = f_class_quad_vec+f_class_lin_vec(sample_ind)+f_class_const;
end

% Proportion of samples from class 0
mean(f_class_ans >= 0); % 0.5051

% Scatter plot data
f_class_0_ind = f_class_ans >= 0;
f_class_1_ind = f_class_ans <  0;
f_class_0_vec = data_3(f_class_0_ind, :);
f_class_1_vec = data_3(f_class_1_ind, :);

% Scatter plots
hold on;
scatter(f_class_0_vec(:, 1), f_class_0_vec(:, 2), 'r');
scatter(f_class_1_vec(:, 1), f_class_1_vec(:, 2), 'b');

% Contour
ineq_f_d = @(x,y) [x,y]*(f_sigma_0^(-1)-f_sigma_1^(-1))*[x;y]...
    +2*(f_mu_1.'*f_sigma_1^(-1)-f_mu_0.'*f_sigma_0^(-1))*[x;y]+f_class_const;
fcontour(ineq_f_d, [0 14 4 13], 'LevelList', 0);

% Formatting
title('Petal Length vs Petal Width', 'Interpreter', 'Latex');
xlabel('Petal Length', 'Interpreter', 'Latex');
ylabel('Petal Width', 'Interpreter', 'Latex');
legend('class 0', 'class 1', '$\vec{x}^TC\vec{x}+\vec{b}^T\cdot\vec{x}+a=0$', 'Interpreter', 'Latex', 'Location', 'Best');
print(gcf, fullfile(f_path, '03_f_scatter'), '-dpng', '-r300');

%% Functions

% sample mean estimator
function mu_N = sample_mean(data, samples)
    mu_N = zeros(1, length(samples));
```

```
        for i = 1:length(samples)
            mu_N(i) = mean(data(1:samples(i)));
        end
end
```