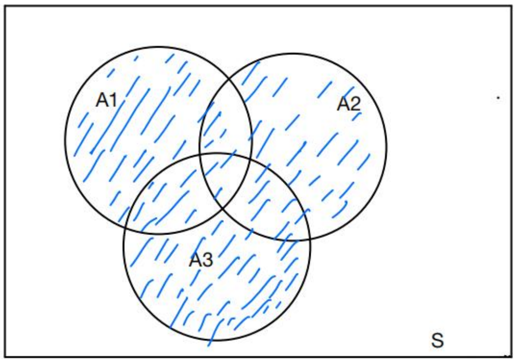
Final exam solution

1.

(a) The Venn diagram is as below



Let S be the set of all observations, and $\mathrm{area}(S) = 1$, we have

$$P(S) = \mathrm{area}(S) = 1,$$

$$P(A_i) = \mathrm{area}(A_i), \text{ for } i = 1,2,3$$

$$P(A_1 \cup A_2 \cup A_3) = \mathrm{area}(\text{shaded in blue})$$

From the diagram above, it clearly holds that

$$\mathrm{area}(\text{shaded in blue}) \leq \mathrm{area}(A_1) + \mathrm{area}(A_2) + \mathrm{area}(A_3)$$

Therefore

$$P(A_1 \cup A_2 \cup A_3) \leq P(A_1) + P(A_2) + P(A_3)$$

(b) Note that

$$\begin{aligned} c \in (A \cup B)^c &\Leftrightarrow c \notin (A \cup B) \\ &\Leftrightarrow c \notin A \text{ and } c \notin B \\ &\Leftrightarrow c \in A^c \text{ and } c \in B^c \\ &\Leftrightarrow c \in A^c \cap B^c \end{aligned}$$

Therefore, $(A \cup B)^c \subseteq A^c \cap B^c$ and $A^c \cap B^c \subseteq (A \cup B)^c$, it follows that

$$(A \cup B)^c = A^c \cap B^c$$

Substitute A, B with A^c, B^c correspondently gives

$$(A^c \cup B^c)^c = (A^c)^c \cap (B^c)^c$$

$$\Leftrightarrow A^c \cup B^c = A \cap B$$

$$\Leftrightarrow (A^c \cup B^c)^c = (A \cap B)^c$$

(c) Since A and B are independent events and do not have zero probabilities, then

$$P(A \cap B) = P(A)P(B), \text{ and } P(A), P(B) > 0$$

i. Since $P(A) = P(A \cap B^c) + P(A \cap B)$, then

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ &= P(A) + P(A)P(B) && \text{by the independent of A and B} \\ &= P(A)(1 - P(B)) \\ &= P(A)P(B^c) && \text{since } P(B^c) = 1 - P(B) \end{aligned}$$

ii. From (b) we have

$$\begin{aligned} P(A^c \cap B^c) &= P((A \cup B)^c) \\ &= 1 - P(A \cup B) \\ &= 1 - (P(A) + P(B) - P(A \cap B)) \\ &= 1 - P(A) - P(B) + P(A)P(B) \\ &= (1 - P(A)) - P(B)(1 - P(A)) \\ &= P(A^c) - P(B)P(A^c) \\ &= P(A^c)(1 - P(B)) \\ &= P(A^c)P(B^c) \end{aligned}$$

2.

(i) $X \sim \text{Binomial}(n, p)$, where $n = 100$ and $p = 0.01$

(ii) The probability of no misidentifications is

$$P(X = 0) = \binom{100}{0}(0.01)^0(1 - 0.01)^{100} = 0.3660$$

(iii) The probability of at least 2 errors is

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - \binom{100}{0}(0.01)^0(1 - 0.01)^{100} - \binom{100}{1}(0.01)^1(1 - 0.01)^{99} \\ &= 1 - \binom{100}{0}(0.01)^0(1 - 0.01)^{100} - \binom{100}{1}(0.01)^1(1 - 0.01)^{99} \end{aligned}$$

$$= 1 - 0.3660 - 0.3697$$

$$= 0.2642$$

(iv) Note that the Poisson distribution can be used to approximate the binomial distribution when the sample size, n is very large and p is very small. Here, $X \sim \text{Binomial}(n, p)$, we have $n = 100$, $p = 0.01$, thus, we use $\text{Poisson}(\lambda)$ with $\lambda = np = 1$ to approximate the distribution of X.

(v) Let

Error1: the first spam misidentifies an email.

Error2: the second spam misidentifies an email.

We have $P(\text{Error1}) = 0.01$, $P(\text{Error2}) = 0.005$. Note that the second check is in a third party app,thus Error1 and Error2 are independent,i.e.

$$P(\text{Error1 and Error2}) = P(\text{Error1})P(\text{Error2}).$$

Thus

$$P(\text{Error2}|\text{Error1}) = \frac{P(\text{Error1})P(\text{Error2})}{P(\text{Error1})} = P(\text{Error2}) = 0.005$$

From problem 1 part(c) we also have Error1 and Error2^c is independent thus

$$P(\text{Error1}|\text{Error2}^c) = \frac{P(\text{Error1})P(\text{Error2}^c)}{P(\text{Error2}^c)} = P(\text{Error1}) = 0.01$$

3.

(a) The probability that all five cards are hearts is

$$P(\text{all five cards are hearts}) = \frac{\binom{13}{5}}{\binom{52}{5}} = 0.000495$$

(b) Consider about choose 5 cards, Let X be the number of hearts, then

$$X \sim \text{Hypergeometric}(N, K, n)$$

Where $N = 52$, $K = 13$, $n = 5$, and

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

(c) Let the sample be 2D3H, i.e. a hand of five cards with 2 diamonds and 3 spades.

$$P(2D3H) = \frac{\binom{13}{2}\binom{13}{3}}{\binom{52}{5}}$$

(d) Suppose that we have 3H1D1S of different kind, discard 1D1S aside, then pickup 2 new card which are both hearts, the probability of this hand is

$$P = \frac{\binom{13}{3}\binom{13}{1}\binom{13}{1}}{\binom{52}{5}} \cdot \frac{\binom{13-3}{2}}{\binom{52-5}{2}} = \frac{\binom{13}{3}\binom{13}{1}\binom{13}{1}\binom{10}{2}}{\binom{52}{5}\binom{47}{2}}$$

4.

(a) For a Poisson Distribution with λ, if λ is large, then it is appropriate to approximate it with a Normal Distribution. In practice, if $\lambda \geq 10$, we use $N(\mu, \sigma^2)$,where $\mu = \lambda$, and $\sigma = \sqrt{\lambda}$, to approximate Poisson(λ). In our question, $X \sim \text{Poisson}(10)$, thus we use $N(10, \sqrt{10}^2)$ to approximate X.

(b) Note that $X \sim \text{Poisson}(10)$, then

$$\begin{aligned} P(X \leq 6) &= P(X \leq 6.5) \qquad \texttt{continuity correction.} \\ &= P\left(\frac{X-10}{\sqrt{10}} \leq \frac{6.5-10}{\sqrt{10}}\right) \\ &= P\left(Z \leq \frac{6.5-10}{\sqrt{10}}\right) \\ &= P(Z \leq -1.1068) \\ &= 0.1342 \end{aligned}$$

where $Z \sim N(0,1)$, and the final result is computed by the R command

```
pnorm(-1.1068)
```

(c) Let $X_i \sim \text{Poisson}(10)$, $i = 1,2,3,4,5$. Then the average of all five stores

$$Y = \frac{\sum_i^5 X_i}{5}$$

Since $X_1, X_2, \ldots, X_5$ are independent, thus

$$E(Y) = E(\frac{\sum_i^5 X_i}{5}) = \frac{1}{5}\sum_i^5 E(X_i) = \frac{1}{5}\sum_i^5 10 = 10$$

and

$$Var(Y) = Var(\frac{\sum_i^5 X_i}{5}) = \frac{1}{25}\sum_i^5 Var(X_i) = \frac{1}{25}\sum_i^5 10 = 2$$

We can also use $N(10, \sqrt{2}^2)$ to approximate Y.

(d)From part (c) , we have

$$\begin{aligned} P(Y \le 6) &= P(Y \le 6.5) \qquad \text{continuity correction} \\ &= P(\frac{Y-10}{\sqrt{2}} \le \frac{6.5-10}{\sqrt{2}}) \\ &= P(Z \le \frac{6.5-10}{\sqrt{2}}) \\ &= P(Z \le -2.4749) \\ &= 0.0067 \end{aligned}$$

The result is computed by the R command

pnorm(-2.4749**)**

5.

(a) The prior distribution of μ_A is $X \cdot N(30, 5^2)$, where $X \sim Poisson(80)$. The mean is $\mu_A = 80 \cdot 30 = 2400$, and standard deviation

$$\sigma_A = \sqrt{(80+80^2)(5^2+30^2) - 80^2 \cdot 30^2} = 483.735$$

(b) We use $N(80,80)$ to approximate $X \sim Poisson(80)$,thus we have

$$\begin{aligned} P(X \ge 101) &= P(X \ge 100.5) \\ &= P(\frac{X-80}{\sqrt{80}} \ge \frac{100.5-80}{\sqrt{80}}) \\ &= P(Z \ge \frac{100.5-80}{\sqrt{80}}) \\ &= P(Z \ge 2.292) \\ &= 1 - P(Z \le 2.292) \\ &= 0.011 \end{aligned}$$

Note that the event runs for 3 hours, therefore, the probability that the number of patrons exceeds 100 is

$$1-(1-P(X \geq 101))^3 = 0.0326$$

(c) The distribution is $N(2592.588, 426.3952^2)$

(d) Note that

$$\bar{x} = 2592.588,\ s = 426.3952, n_1 = 3$$
$$\mu_A = 2400, \sigma_A = 483.735, n = 3$$

Then

$$\widetilde{\mu_A} = \frac{\mu_A(s/\sqrt{n_1})^2 + \bar{x}(\sigma_A/\sqrt{n})^2}{(s/\sqrt{n_1})^2 + (\sigma_A/\sqrt{n})^2} = \frac{2400(426.3952/\sqrt{3})^2 + 2592.588(483.735/\sqrt{3})^2}{(426.3952/\sqrt{3})^2 + (483.735/\sqrt{3})^2} = 2508.3794$$

$$\widetilde{\sigma_A} = \sqrt{\frac{(s/\sqrt{n_1})^2 \cdot (\sigma_A/\sqrt{n})^2}{(s/\sqrt{n_1})^2 + (\sigma_A/\sqrt{n})^2}} = 184.676$$

6.

(a) The logistic regression model is the most appropriate mode to use for predicting whether a customer defaults or not.

$$P(\text{defaults} = 1) = \frac{e^{b_0 + b_1 \cdot \text{Risk\_score}}}{1 + e^{b_0 + b_1 \cdot \text{Risk\_score}}}$$

(b) Independent variable: Risk_Score

Dependent variable: The probability of default occurs: prob_def .

(c) The code:

```
# read data
Default <- read.csv('defaults.csv')

# logistic regression
m1 <- glm(pro_def ~ Risk_Score, data=Default,+ family = binomial())
summary(m1)
```

Here, we use the binomial distribution as the distribution of errors terms, since the response variable, y is binary.

(d) 95% confidence interval for the parameter (coefficient) for Risk_Score is [0.2336791,1.6566246].

(e) An increase of one unit Risk_Score increases the log-odds in favor of an `pro_def` value by an estimated $\widehat{b_1} = 1.827$ (from the R result of (f)) with 95% confidence interval between 0.234 and 1.657.

(f) The model is

$$P(\text{defaults} = 1) = \frac{e^{-6.845+1.827\cdot \text{Risk\_score}}}{1+e^{-6.845+1.827\cdot \text{Risk\_score}}}$$

If Risk_score $= 4$, then the probability to default is

$$\frac{e^{-6.845+1.827\cdot 4}}{1+e^{-6.845+1.827\cdot 4}} = 0.6137$$