STAT1301 Advanced Analysis of Scientific Data

Semester 2, 2025, Assignment 2

Caleb Yates s49886350

1 Introduction

Throughout the report, the following syntactical shortcuts and notation will be used.

If X is a random variable of the sample space Ω, an abbreviation of set notation is as follows:

$$\text{Abbreviate } \{d : \forall d \in \Omega \text{ and } \mathrm{X}(d) = x\} \tag{1}$$

$$\text{As } \{\mathrm{X} = x\} \tag{2}$$

Additionally, when thinking in terms of sets becomes obselete,

$$\text{Abbreviate } \mathrm{P}(\{d : \forall d \in \Omega \text{ and } \mathrm{X}(d) = x\}) \tag{3}$$

$$\text{As } \mathrm{P}(\mathrm{X} = x) \tag{4}$$

The abbriviation will be used with inequalities as well.

Given some random variable X, there must exist a function mapping from the sample space Ω to the domain of X, which can be at most $\mathbb{R}$. This function is (intuitively) named X. This function incidentally defines the random variable, which is the motivating reason for using its letter to represent its mapping. The notation Domain[X] will be used throughout this report to indicate the domain of the function mapping X and hence the random variable X itself by definition.

Also, $N(\mu, \sigma)$ indicates that σ is $\sqrt{\text{Var}}$, aka the standard deviation. This is opposed to the syntax of $N(\mu, \sigma^2)$. For clarity, $\sigma =$ will always be explicitely written to avoid ambiguity.

Various probability (and set) theorems are used throughout this report. For clarity, the following are named:

$$\mathrm{P}(\{\mathrm{X} < x\}) = \mathrm{P}(\{x > \mathrm{X}\}) \forall x \tag{5}$$

Above (5) is true for any random variable X.

$$\mathrm{X} \sim N(\mu = 0, \sigma) \tag{6}$$

$$\implies \mathrm{P}(\{\mathrm{X} < x\}) = 1 - \mathrm{P}(\{\mathrm{X} > -x\}) \tag{7}$$

$$\iff \mathrm{P}(\{\mathrm{X} < -x\}) = 1 - \mathrm{P}(\{\mathrm{X} > x\}) \tag{8}$$

When X is a symmetrical distribution around 0, for example a standard normal distribution Z, above (6) is true.

2 Question 1

To begin, lets define the sample space

$$\Omega = \{(a,b) \in \{1,2,3,4,5,6\}\}$$
$$|\Omega| = 36$$

Notice this is uniform, and hence that a and b are independant

$$\mathrm{P}(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{36} \tag{9}$$

2.1 Part a)

Let X be a random variable representing the payout of a given dice roll $(a,b) \in \Omega$:

$$\mathrm{X}((a,b) \in \Omega) = a \cdot b$$

Let f_X be the PMF of X. Note $f_\mathrm{X}(x \in \Omega) = \mathrm{P}(\{\mathrm{X} = x\})$. By cases, the probability distribution of X can be deduced:

$$f_\mathrm{X}(1) = \frac{1}{36} \qquad f_\mathrm{X}(8) = \frac{2}{36} \qquad f_\mathrm{X}(18) = \frac{2}{36}$$
$$f_\mathrm{X}(2) = \frac{2}{36} \qquad f_\mathrm{X}(9) = \frac{1}{36} \qquad f_\mathrm{X}(20) = \frac{2}{36}$$
$$f_\mathrm{X}(3) = \frac{2}{36} \qquad f_\mathrm{X}(10) = \frac{2}{36} \qquad f_\mathrm{X}(24) = \frac{2}{36}$$
$$f_\mathrm{X}(4) = \frac{3}{36} \qquad f_\mathrm{X}(12) = \frac{4}{36} \qquad f_\mathrm{X}(25) = \frac{1}{36}$$
$$f_\mathrm{X}(5) = \frac{2}{36} \qquad f_\mathrm{X}(15) = \frac{2}{36} \qquad f_\mathrm{X}(30) = \frac{2}{36}$$
$$f_\mathrm{X}(6) = \frac{4}{36} \qquad f_\mathrm{X}(16) = \frac{1}{36} \qquad f_\mathrm{X}(36) = \frac{1}{36}$$

For all other values x, $f_\mathrm{X}(x) = 0$

2.2 Part b)

This makes determining the expected value of X trivial:

$$\mathrm{E}(\mathrm{X}) = \sum_{c \in \Omega} \mathrm{X}(c)\mathrm{P}(c) \tag{10}$$
$$= \sum_{x \in \mathrm{Domain[X]}} x\mathrm{P}(\{\mathrm{X} = x\}) \tag{11}$$
$$= 1 \cdot f_\mathrm{X}(1) + 2 \cdot f_\mathrm{X}(2) + \cdots 30 \cdot f_\mathrm{X}(30) + 36 \cdot f_\mathrm{X}(36) \tag{12}$$
$$= \frac{1}{36} + \frac{4}{36} + \cdots \frac{60}{36} + \frac{36}{36} \tag{13}$$
$$= \frac{441}{36} = \frac{49}{4} = 12.25 \tag{14}$$

2.3 Part c)

Evaluating Var(X) is similarly trivial

$$
\begin{aligned}
\mathrm{Var(X)} &= \mathrm{E}[(\mathrm{X} - \mathrm{E}(X))^2] && (15)\\
&= \sum_{c\in\Omega} (\mathrm{X}(c) - \frac{49}{4})^2 \mathrm{P}(\{c\}) && (16)\\
&= \sum_{x\in\mathrm{Domain[X]}} (x - \frac{49}{4})^2 \mathrm{P}(\{\mathrm{X} = x\}) && (17)\\
&= (1 - \frac{49}{4})^2 \cdot \frac{1}{36} + (2 - \frac{49}{4})^2 \cdot \frac{2}{36} + \cdots (30 - \frac{49}{4})^2 \cdot \frac{2}{36} + (36 - \frac{49}{4})^2 \cdot \frac{1}{36} && (18)\\
&= \frac{11515}{144} \approx 79.97 && (19)\\
\implies \sigma_{\mathrm{X}} &= \sqrt{\mathrm{Var(X)}} = \sqrt{\frac{11515}{144}} \approx 8.942 && (20)
\end{aligned}
$$

3 Question 2

Understanding this question in terms of a sample space isn't very fruitful. Ω is completely unspecified, we can only deduce that $|\Omega| \geq (0, 20)$, which implies it is continuous. $\mathrm{P}(A) : \exists A \in \Omega$ is also completely unknown.

3.1 Part a)

Let X be the continuous random variable of algae growth as measured in grams of biomass produced. Note Domain[X] $= (0, 20)$.

Since X is a random variable, its PDF f_{X} must sum to 1:

$$
\begin{aligned}
1 &= \int_{c\in\Omega} \mathrm{P}(\{c\}) && (21)\\
&= \int_{x\in\mathrm{Domain}[X]} \mathrm{P}(\{\mathrm{X} = x\}) && (22)\\
&= \int_0^{20} c(x^2 - 60x + 800)\mathrm{d}x && (23)\\
&= c[\frac{1}{3}x^3 - 30x^2 + 800x]_{x=0}^{x=20} && (24)\\
1/c &= [\frac{1}{3}(20)^3 - 30(20)^2 + 800(20)] - [0 - 0 + 0] && (25)\\
1/c &= \frac{20000}{3} && (26)\\
c &= \frac{3}{20000} && (27)
\end{aligned}
$$

3.2 Part b)

Let F_{X} be the CDF of X:

$$
\begin{aligned}
F_{\mathrm{X}} &= \int_{-\infty}^{x} f_{\mathrm{X}}(x)\mathrm{d}x && (28)\\
&= \int_0^{x} c(x^2 - 60x + 800)\mathrm{d}x && (29)\\
&= c[\frac{1}{3}x^3 - 30x^2 + 800x]_{x=0}^{x=x} && (30)\\
\frac{F_{\mathrm{X}}}{c} &= [\frac{1}{3}x^3 - 30x^2 + 800x] - [\frac{1}{3}0^3 - 30\cdot 0^2 + 800\cdot 0] && (31)\\
\implies F_{\mathrm{X}} &= c(\frac{1}{3}x^3 - 30x^2 + 800x) \text{ for } 0 \leq x \leq 20 && (32)\\
&= \frac{1}{20000}x^3 - \frac{9}{2000}x^2 + \frac{3}{25}x && (33)
\end{aligned}
$$

3.3 Part c)

$$\begin{aligned}
\mathrm{E(X)} &= \int_{x \in \mathrm{Domain[X]}} x f_{\mathrm{X}} \mathrm{d}x && (34)\\
&= \int_0^{20} x \cdot c(x^2 - 60x + 800)\mathrm{d}x && (35)\\
\frac{\mathrm{E(X)}}{c} &= \int_0^{20} x^3 - 60x^2 + 800x\mathrm{d}x && (36)\\
&= [\frac{1}{4}x^4 - 20x^3 + 400x^2]_{x=0}^{x=20} && (37)\\
&= [\frac{1}{4}(20)^4 - 20(20)^3 + 400(20)^2] - [0 - 0 + 0] && (38)\\
&= 40000 - 160000 + 160000 && (39)\\
\mathrm{E(X)} &= c \cdot 40000 && (40)\\
\mathrm{E(X)} &= 6 \text{ grams} && (41)
\end{aligned}$$

3.4 Part d)

$$\begin{aligned}
&\mathrm{P}(\{\mathrm{X} > 10\}|\{\mathrm{X} > 2\}) && (42)\\
&= \frac{\mathrm{P}(\{\mathrm{X} > 10\} \cap \{\mathrm{X} > 2\})}{\{\mathrm{X} > 2\}} && (43)\\
&= \frac{\mathrm{P}(\{\mathrm{X} > 10\})}{\mathrm{P}(\{\mathrm{X} > 2\})} && (44)
\end{aligned}$$

From the CDF definition of X, $\mathrm{P}(\{\mathrm{X} < x\}) = f_{\mathrm{X}}(x)$

$$\begin{aligned}
\implies \mathrm{P}(\{\mathrm{X} > 10\}) &= 1 - \mathrm{P}(\{\mathrm{X} < 10\}) && (45)\\
&= 1 - F_{\mathrm{X}}(10) && (46)\\
&= 1 - \frac{4}{5} && (47)\\
&= \frac{1}{5} && (48)
\end{aligned}$$

$$\begin{aligned}
\implies \mathrm{P}(\{\mathrm{X} > 2\}) &= 1 - \mathrm{P}(\{\mathrm{X} < 2\}) && (49)\\
&= 1 - F_{\mathrm{X}}(2) && (50)\\
&= 1 - \frac{139}{625} && (51)\\
&= \frac{486}{625} && (52)
\end{aligned}$$

$$\implies \frac{\mathrm{P}(\{\mathrm{X} > 10\})}{\mathrm{P}(\{\mathrm{X} > 2\})} = \frac{\frac{1}{5}}{\frac{486}{625}} = \frac{125}{486} \approx 0.2572 \qquad (53)$$

Therefore, the probability that the biomass exceeds 10 grams, given that it is detectable, is above in (53) $= \frac{125}{486}$.

4 Question 3

Assume that $p = 0.25$ for all the products, not just the 25 that were sampled.

The sample space for this is again completely unspecified, and the P probability function is practically useless for this question. For convenience, the sample space Ω is therefore defined as the domain of X, representing the number of products passing the specific inspection.

$$\Omega = \{1, 2, 3...24, 25\}$$

This makes the definition of X trivial, and its domain incidentally the entire sample space:

$$\mathrm{X}(a \in \Omega) = a$$

4.1 Part a)

Since each product has a $p = 0.25$ probability of passing inspection, and there are 25 products, and it is assumed each inspection and product is independant of each other, X is a binomial distribution:

$$\mathrm{X} \sim \mathrm{Bin}(n = 25, p = 0.25)$$

Notes the following theorems about binomial distributions and X:

$$\mathrm{P}(\{\mathrm{X} = x\}) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{25}{x} 0.25^x \cdot 0.75^{25-x}$$
$$\mathrm{E}(\mathrm{X}) = np = \frac{25}{4}$$
$$\mathrm{Var}(\mathrm{X}) = np(1-p) = \frac{75}{16}$$

4.2 Part b)

Let X_2 be the random variable representing the probability distribution of X with an n parameter such that the probability of finding a defect-free product eexceeds 99%:

$$\mathrm{X}_2 \sim \mathrm{Bin}(n, p = 0.25)$$

$$\begin{aligned}
\mathrm{P}(\{\mathrm{X}_2 \geq 1\}) &> 0.99 && (54)\\
0.99 &< \mathrm{P}(\{\mathrm{X}_2 \geq 1\}) && (55)\\
0.99 &< 1 - \mathrm{P}(\{\mathrm{X}_2 = 0\}) && (56)\\
0.99 - 1 &< -\binom{n}{0}(0.25)^0(0.75)^n && (57)\\
0.01 &> 1 \cdot 1 \cdot 0.75^n && (58)\\
\log_{0.75} 0.01 &> n && (59)\\
\implies n &< \log_{0.75} 0.01 \approx 16.008 && (60)
\end{aligned}$$

Therefore the minimum (integer) sample size is $n = 16$.

4.3 Part c)

The random variable Y is dependant on X. Given a possibility $a \in \Omega$ from the sample space, $\mathrm{Y}(a)$ explicitely depends upon $\mathrm{X}(a)$ such that it exactly equals:

$$\begin{aligned} \mathrm{Y}(a \in \Omega) &= 3\mathrm{X}(a) - (25 - \mathrm{X}(a)) && (61)\\ &= 4\mathrm{X}(a) - 25 && (62) \end{aligned}$$

This allows us to calculate E(X) and Var(X) relatively easily using probability theorems:

$$\begin{aligned} \mathrm{E}(\mathrm{Y}) &= \mathrm{E}(4\mathrm{X} - 25) && (63)\\ &= 4\mathrm{E}(\mathrm{X}) - 25 && (64)\\ &= 4 \cdot \frac{25}{4} - 25 && (65)\\ &= 0 && (66) \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(\mathrm{Y}) &= \mathrm{Var}(4\mathrm{X} - 25) && (67)\\ &= 4^2\mathrm{Var}(\mathrm{X}) && (68)\\ &= 16 \cdot \frac{75}{16} && (69)\\ &= 75 && (70) \end{aligned}$$

4.4 Part d)

Since Y is defined in terms of X, this isn't too difficult to evaluate:

$$\begin{aligned} \mathrm{P}(\{\mathrm{Y} \geq 27\}) &= \mathrm{P}(\{4\mathrm{X} - 25 \geq 27\}) && (71)\\ &= \mathrm{P}(\{4\mathrm{X} \geq 52\}) && (72)\\ &= \mathrm{P}(\{\mathrm{X} \geq 13\}) && (73)\\ &\approx 0.00337 && (74) \end{aligned}$$

This can be calculated by running 1 - pbinom(12, 25, 0.25) in R

5 Question 4

Let $\Omega = (-\infty, +\infty)$ in units ℃, representing the continuous range of possible temperatures in the storeroom. An argument could be made to limit this to $(-\infty, 8)$.

Let X be a random variable for the temperature inside the storeroom.

5.1 Part a)

$$\mathrm{X} \sim N(\mu = 7.5℃, \sigma = 0.3℃)$$

$$\begin{aligned}
\mathrm{P}(\{7.2 < \mathrm{X} < 8\}) &= \mathrm{P}(\{\frac{7.2 - 7.5}{0.3} < \frac{\mathrm{X} - \mu}{\sigma} < \frac{8 - 7.5}{0.3}\}) && (75) \\
&= \mathrm{P}(\{-\frac{2}{3} < \mathrm{Z} < \frac{5}{3}\}) && (76) \\
&= \mathrm{P}(\{\mathrm{Z} < \frac{5}{3}\}) - \mathrm{P}(\{-\frac{2}{3} < \mathrm{Z}\}) && (77) \\
&= \mathrm{P}(\{\mathrm{Z} < \frac{5}{3}\}) - \mathrm{P}(\{\mathrm{Z} > -\frac{2}{3}\}) && (78) \\
&= \mathrm{P}(\{\mathrm{Z} < \frac{5}{3}\}) - (1 - \mathrm{P}(\{\mathrm{Z} < \frac{2}{3}\})) && (79) \\
&= \mathrm{P}(\{\mathrm{Z} < \frac{5}{3}\}) + \mathrm{P}(\{\mathrm{Z} < \frac{2}{3}\}) - 1 && (80)
\end{aligned}$$

Using stats tables this equals $0.9515 + 0.7454 - 1 = 0.6969$. Using R running pnorm($\frac{5}{3}$) - pnorm($-\frac{2}{3}$) $= 0.6997$.

5.2 Part b)

$$\mathrm{X} \sim N(\mu, \sigma = 0.3℃)$$

$$\begin{aligned}
\mathrm{P}(\{\mathrm{X} > 8℃\}) &= 1\% && (81) \\
0.01 &= \mathrm{P}(\{\mathrm{X} > 8\}) && (82) \\
&= 1 - \mathrm{P}(\{\mathrm{X} < 8\}) && (83) \\
0.99 &= \mathrm{P}(\{\mathrm{X} < 8\}) && (84) \\
&= \mathrm{P}(\{\frac{\mathrm{X} - \mu}{\sigma} = \frac{8 - \mu}{\sigma}\}) && (85) \\
0.99 &= \mathrm{P}(\{\mathrm{Z} = \frac{8 - \mu}{3}\}) && (86) \\
& && (87)
\end{aligned}$$

Let z be the value which satisfies $\mathrm{P}(\{\mathrm{Z} < z\}) = 0.99$.

$$
\begin{aligned}
\implies \frac{8-\mu}{3} &= z && (88)\\
8-\mu &= 3z && (89)\\
-\mu &= 3x-8 && (90)\\
\mu &= 8-3z && (91)
\end{aligned}
$$

Using the stats table, $z \approx 2.33$ which implies $\mu \approx 8 - 3 \cdot 2.33 = 1.01$℃. Using R, $z =$ qnorm(0.99) ≈ 2.326348, which implies $\mu \approx 8 - 3 \cdot 2.326348 \approx 1.021$℃.

5.3 Part c)

We are given no information about the parameters of X

$$\mathrm{X} \sim N(\mu, \sigma)$$

$$
\begin{aligned}
\mathrm{P}(\{\mu - 1℃ < \mathrm{X} < \mu + 1℃\}) &= 95\% && (92)\\
0.95 &= \mathrm{P}(\{\frac{(\mu-1)-\mu}{\sigma} < \frac{\mathrm{X}-\mu}{\sigma} < \frac{(\mu-1)-\mu}{\sigma}\}) && (93)\\
&= \mathrm{P}(\{\frac{-1}{\sigma} < \mathrm{Z} < \frac{+1}{\sigma}\}) && (94)\\
&= 1 - (\mathrm{P}(\{\mathrm{Z} < \frac{-1}{\sigma}\}) + \mathrm{P}(\{\mathrm{Z} > \frac{+1}{\sigma}\})) && (95)\\
&= 1 - 2\mathrm{P}(\{\mathrm{Z} < \frac{-1}{\sigma}\}) && (96)\\
0.05 &= 2\mathrm{P}(\{\mathrm{Z} < \frac{-1}{\sigma}\}) && (97)\\
0.025 &= \mathrm{P}(\{\mathrm{Z} < \frac{-1}{\sigma}\}) && (98)\\
1 - 0.025 &= 1 - \mathrm{P}(\{\mathrm{Z} < \frac{-1}{\sigma}\}) && (99)\\
0.975 &= \mathrm{P}(\{\mathrm{Z} < \frac{+1}{\sigma}\}) && (100)\\
\text{Let } z \text{ be the solution to } 0.975 &= \mathrm{P}(\{\mathrm{Z} < z\}) && (101)\\
\implies z &= \frac{+1}{\sigma} && (102)\\
\implies \sigma &= \frac{1}{z} && (103)
\end{aligned}
$$

Using the stats table, $z \approx 1.96$ which implies $\sigma \approx \frac{1}{1.96} \approx 0.510204 \approx 0.51$. Using R $z =$ qnorm(0.975) ≈ 1.959964 which implies $\sigma \approx \frac{1}{1.959964} \approx 0.5102135 \approx 0.51$.