STAT1301 Advanced Analysis of Scientific Data

Semester 2, 2025, Assignment 2

Caleb Yates s49886350

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:

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

The above abbriviation will be used with inequalities as well, e.g. $\mathrm{P}(\{\mathrm{X} < x\})$ or $\mathrm{P}(\{\mathrm{X} > x\})$.

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.

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

$$\mathrm{P}(\{\mathrm{X} < x\}) = \mathrm{P}(\{x > \mathrm{X}\})\forall x \tag{1}$$

$$\mathrm{P}(\{\mathrm{X} < x\}) = 1 - \mathrm{P}(\{\mathrm{X} > x\})\forall x \tag{2}$$

Above (1) and (2) are true for any random variable X.

$$\begin{aligned} &\mathrm{X} \sim N(\mu = 0, \sigma) \\ &\implies \mathrm{P}(\{\mathrm{X} < x\}) = \mathrm{P}(\{\mathrm{X} > -x\}) \\ &\iff \mathrm{P}(\{\mathrm{X} < -x\}) = \mathrm{P}(\{\mathrm{X} > x\}) \end{aligned} \tag{3}$$

When X is a symmetrical distribution around 0, for example the standard normal distribution Z, above (3) is true.

Also, an equivalent formula for E(X) was used:

$$\mathrm{E}(\mathrm{X}) = \sum_{c \in \Omega} \mathrm{X}(c) \cdot \mathrm{P}(\{c\}) \tag{4}$$

Recall that $\mathrm{X}(c)$ is the function mapping for the random variable X. For simple cases of one random variable, this can be simplified to (5) by noticing that the sample space can be partitioned into

$$\mathrm{E}(\mathrm{X}) = \sum_{x \in \mathrm{Domain[X]}} x \cdot \mathrm{P}(\{\mathrm{X} = x\}) \tag{5}$$

Question 1

To begin, lets define the sample space

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

Notice this follows a uniform probability distribution:

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

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:

$$f_{\mathrm{X}}(1) = \frac{1}{36} \quad f_{\mathrm{X}}(2) = \frac{2}{36} \quad f_{\mathrm{X}}(3) = \frac{2}{36} \quad f_{\mathrm{X}}(4) = \frac{3}{36} \quad f_{\mathrm{X}}(5) = \frac{2}{36} \quad f_{\mathrm{X}}(6) = \frac{4}{36}$$

$$f_{\mathrm{X}}(8) = \frac{2}{36} \quad f_{\mathrm{X}}(9) = \frac{1}{36} \quad f_{\mathrm{X}}(10) = \frac{2}{36} \quad f_{\mathrm{X}}(12) = \frac{4}{36} \quad f_{\mathrm{X}}(15) = \frac{2}{36} \quad f_{\mathrm{X}}(16) = \frac{1}{36}$$

$$f_{\mathrm{X}}(18) = \frac{2}{36} \quad f_{\mathrm{X}}(20) = \frac{2}{36} \quad f_{\mathrm{X}}(24) = \frac{2}{36} \quad f_{\mathrm{X}}(25) = \frac{1}{36} \quad f_{\mathrm{X}}(30) = \frac{2}{36} \quad f_{\mathrm{X}}(36) = \frac{1}{36}$$

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

Part b)

This makes determining the expected value of X trivial:

$$\begin{aligned}
\mathrm{E}(\mathrm{X}) &= \sum_{c \in \Omega} \mathrm{X}(c)\mathrm{P}(\{c\}) \\
&= \sum_{x \in \mathrm{Domain}[\mathrm{X}]} x\mathrm{P}(\{\mathrm{X} = x\}) \\
&= 1 \cdot f_{\mathrm{X}}(1) + 2 \cdot f_{\mathrm{X}}(2) + \cdots 30 \cdot f_{\mathrm{X}}(30) + 36 \cdot f_{\mathrm{X}}(36) \\
&= \frac{1}{36} + \frac{4}{36} + \cdots \frac{60}{36} + \frac{36}{36} \\
&= \frac{441}{36} = \frac{49}{4} = \$12.25
\end{aligned}$$

Therefore the expected of X is \$12.25

Part c)

Evaluating Var(X) is similarly trivial

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

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.

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:

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

Part b)

Let F_{X} be the CDF of X:

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

Part c)

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

Part d)

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

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

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

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

$$\implies \frac{\mathrm{P}(\{\mathrm{X} > 10\})}{\mathrm{P}(\{\mathrm{X} > 2\})} = \frac{\frac{1}{5}}{\frac{486}{625}} = \frac{125}{486} \approx 0.2572 \tag{6}$$

Therefore, the probability that the biomass exceeds 10 grams, given that it is detectable, is above in (6) $= \frac{125}{486}$.

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:

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

These definition are not necessary to solve this question, however, and are included only for completeness.

Part a)

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

$$\mathrm{P}(\{\mathrm{X} = x\}) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{25}{x} 0.25^x \cdot 0.75^{25-x}$$

$$\mathrm{E}(\mathrm{X}) = np = \frac{25}{4}$$

$$\mathrm{Var}(\mathrm{X}) = np(1-p) = \frac{75}{16}$$

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 exceeds 99%:

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

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

Therefore the minimum (integer) sample size is $n = 17$.

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:

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

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

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

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

Part d)

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

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

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

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.

Part a)

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

$$\begin{aligned}
\mathrm{P}(\{7.2 < \mathrm{X} < 8\}) &= \mathrm{P}(\{\frac{7.2 - 7.5}{0.3} < \frac{\mathrm{X} - \mu}{\sigma} < \frac{8 - 7.5}{0.3}\}) \\
&= \mathrm{P}(\{-1 < \mathrm{Z} < \frac{5}{3}\}) \\
&= \mathrm{P}(\{\mathrm{Z} < \frac{5}{3}\}) - \mathrm{P}(\{-1 < \mathrm{Z}\}) \\
&= \mathrm{P}(\{\mathrm{Z} < \frac{5}{3}\}) - \mathrm{P}(\{\mathrm{Z} > -1\}) \text{ from (1)} \\
&= \mathrm{P}(\{\mathrm{Z} < \frac{5}{3}\}) - (1 - \mathrm{P}(\{\mathrm{Z} < 1\})) \text{ from (3)} \\
&= \mathrm{P}(\{\mathrm{Z} < \frac{5}{3}\}) + \mathrm{P}(\{\mathrm{Z} < 1\}) - 1
\end{aligned}$$

Using stats tables this equals $0.9525 + 0.8413 - 1 = 0.7938$. Using R running pnorm($\frac{5}{3}$) - pnorm($-1$) $= 0.7935544 \approx 0.7936$.

Part b)

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

$$\begin{aligned}
\mathrm{P}(\{\mathrm{X} > 8℃\}) &= 1\% \\
0.01 &= \mathrm{P}(\{\mathrm{X} > 8\}) \\
&= 1 - \mathrm{P}(\{\mathrm{X} < 8\}) \text{ from (2)} \\
0.99 &= \mathrm{P}(\{\mathrm{X} < 8\}) \\
&= \mathrm{P}(\{\frac{\mathrm{X} - \mu}{\sigma} = \frac{8 - \mu}{\sigma}\}) \\
0.99 &= \mathrm{P}(\{\mathrm{Z} = \frac{8 - \mu}{0.3}\})
\end{aligned}$$

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

$$\begin{aligned}
\implies \frac{8-\mu}{0.3} &= z \\
8-\mu &= 0.3z \\
-\mu &= 0.3z - 8 \\
\mu &= 8 - 0.3z
\end{aligned}$$

Using the stats table, $z \approx 2.33$ which implies $\mu \approx 8 - 0.3 \cdot 2.33 = 7.301$℃. Using R, $z =$ qnorm(0.99) ≈ 2.326348, which implies $\mu \approx 8 - 0.3 \cdot 2.36348 = 7.302096 \approx 7.302$℃.

Part c)

We are given no information about the parameters of X

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

Note, $N(\mu, \sigma)$ indicates that σ is $\sqrt{\mathrm{Var}}$, aka the standard deviation. This is opposed to the syntax of $\mathrm{X} \sim N(\mu, \sigma^2)$.

$$\begin{aligned}
\mathrm{P}(\{\mu - 1℃ < \mathrm{X} < \mu + 1℃\}) = 95\% \\
0.95 &= \mathrm{P}(\{\mu - 1 < \mathrm{X} < \mu + 1\}) \\
0.95 &= \mathrm{P}(\{\frac{(\mu-1)-\mu}{\sigma} < \frac{\mathrm{X}-\mu}{\sigma} < \frac{(\mu+1)-\mu}{\sigma}\}) \\
&= \mathrm{P}(\{\frac{-1}{\sigma} < \mathrm{Z} < \frac{+1}{\sigma}\}) \\
&= \mathrm{P}(\{\frac{-1}{\sigma} < \mathrm{Z}\} \cap \{\mathrm{Z} < \frac{+1}{\sigma}\})
\end{aligned}$$

Note that the complement of $\{\frac{-1}{\sigma} < \mathrm{Z}\} \cap \{\mathrm{Z} < \frac{+1}{\sigma}\}$ is $\{\mathrm{Z} < \frac{-1}{\sigma}\} \cup \{\mathrm{Z} > \frac{+1}{\sigma}\}$

$$\begin{aligned}
0.95 &= 1 - (\mathrm{P}(\{\mathrm{Z} < \frac{-1}{\sigma}\} \cup \{\mathrm{Z} > \frac{+1}{\sigma}\})) \\
&= 1 - (\mathrm{P}(\{\mathrm{Z} < \frac{-1}{\sigma}\}) + \mathrm{P}(\{\mathrm{Z} > \frac{+1}{\sigma}\})) \\
&= 1 - 2\mathrm{P}(\{\mathrm{Z} < \frac{-1}{\sigma}\}) \\
0.05 &= 2\mathrm{P}(\{\mathrm{Z} < \frac{-1}{\sigma}\}) \\
0.025 &= \mathrm{P}(\{\mathrm{Z} < \frac{-1}{\sigma}\}) \\
1 - 0.025 &= 1 - \mathrm{P}(\{\mathrm{Z} < \frac{-1}{\sigma}\}) \\
0.975 &= \mathrm{P}(\{\mathrm{Z} < \frac{+1}{\sigma}\}) \text{ from (3)}
\end{aligned}$$

Let z be the solution to $0.975 = \mathrm{P}(\{\mathrm{Z} < z\})$

$$\implies z = \frac{+1}{\sigma}$$
$$\implies \sigma = \frac{1}{z}$$

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 $\sigma \approx \frac{1}{1.959964} \approx 0.5102135 \approx 0.51$℃.