STAT 111: Mathematical Statistics II

David Yang

Spring 2024

Abstract

These notes arise from my studies in STAT 111: Mathematical Statistics II, taught by Professor Phil Everson, at Swarthmore College. I am responsible for all faults in this document, mathematical or otherwise. Feel free to message me with any suggestions or corrections at dyang5@swarthmore.edu.

Contents

1 Discrete Probability Distributions

1.2 Indicator Variables

a) **Define $I(A)$ to be an indicator variable for the event A, meaning $I_{(A)} = 1$ if A occurs and $I_{(A)} = 1$ if A^c occurs. Relate this to a Bernoulli random variable. Show how an indicator variable is the *fundamental bridge* between probability and expected value, in that $P(A) = E(I_{(A)})$. Use this to prove Boole's inequality: $P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n)$. Try to think of a non-trivial example.**

We can write $I(A)$ as

$$I(A) = \begin{cases} 1 & \text{if } A \\ 0 & \text{if } A^c \end{cases}$$

where A occurs with probability of $P(A)$, and A^c occurs with probability $1 - P(A)$. This is equivalent to a $\text{Bern}(P(A))$ random variable.

For the fundamental bridge, note that $E[I(A)] = P(A) \cdot 1 + (1 - P(A)) \cdot 0 = P(A)$.

To prove Boole's Inequality, we will show that $P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n)$. Rewriting the probability values on the left and right hand sides as expected values of indicator variables, we know that

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = E[I_{A_1 \cup A_2 \cup \cdots \cup A_n}].$$

On the other hand,

$$P(A_1) + P(A_2) + \cdots + P(A_n) = E[I_{A_1}] + E[I_{A_2}] + \cdots + E[I_{A_n}] = E[I_{A_1} + I_{A_2} + \cdots + I_{A_n}]$$

Thus, to prove Boole's inequality, it suffices to show that

$$E[I_{A_1 \cup A_2 \cup \cdots \cup A_n}] \leq E[I_{A_1} + I_{A_2} + \cdots + I_{A_n}].$$

Note that $I_{A_1 \cup A_2 \cup \cdots \cup A_n}$ can only be either 0 or 1, since it is an indicator. In the former case, none of A_1 to A_n have occurred, so $I_{A_1} + I_{A_2} + \cdots + I_{A_n} = 0$. In the latter case, if $I_{A_1 \cup A_2 \cup \cdots \cup A_n} = 1$, then at least one of A_1 to A_n has occurred, so $I_{A_1} + I_{A_2} + \cdots + I_{A_n} \geq 1$.

It follows that

$$I_{A_1 \cup A_2 \cup \cdots \cup A_n} \leq I_{A_1} + I_{A_2} + \cdots + I_{A_n}.$$

Taking the expected value of both sides, we arrive at Boole's Inequality.

Example of Boole's Inequality: Let A represent the event of a fair coin flip; $A_i = 1$ if flip i is heads, and $A_i = 0$ if it is tails. Boole's Inequality tells us that if we flip the coin 5 times, the probability we flip at least one heads is less than or equal to 5 times the probability we flip a heads on any single coin flip. Thus, the probability we flip at least one heads is at most $5 \cdot \frac{1}{2} = \frac{5}{2}$; this is trivial as this is greater than 1.

b) **A special case of Boole's inequality occurs when the events all have the same probability. Suppose n graduates all throw their caps in the air and then retrieve a cap at random. Find an expression for the probability that none of the students retrieve their own cap (a derangement). Find the limit of this probability if the number of caps $n \to \infty$. *Hint: For $i = 1, ..., n$, let A_i represent the event that person i retrieves their own cap. Then $(A_1 \cup A_2 \cup ... \cup A_n)^c$ is the event that nobody ends up with their own cap.***

The probability of a derangement is

$$P(A_1 \cup A_2 \cup ... \cup A_n)^c = 1 - P(A_1 \cup A_2 \cup ... \cup A_n).$$

Taking the limit of this expression as $n \to \infty$ and using the Principle of Inclusion Exclusion, we know that this

$$\begin{aligned}\lim_{n\to\infty} 1 - P(A_1 \cup A_2 \cup ... \cup A_n) &= 1 - \left[\sum_{k=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \dots\right] \\ &= \lim_{n\to\infty} 1 - \left[n \cdot \frac{1}{n} - \binom{n}{2}\left(\frac{1}{n} \cdot \frac{1}{n-1}\right) + \binom{n}{3}\left(\frac{1}{n} \cdot \frac{1}{n-1} \cdot \frac{1}{n-2}\right)\right]\end{aligned}$$

Simplifying, this becomes

$$\lim_{n\to\infty} 1 - \left[1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \dots\right] = \lim_{n\to\infty} \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \dots$$

Recall the Taylor Expansion of e^x: $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$. We recognize the above expression as e^{-1}:

$$\begin{aligned}e^{-1} = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} &= \lim_{n\to\infty} \left(1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \dots\right) \\ &= \lim_{n\to\infty} \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \dots\end{aligned}$$

Thus,

$$\lim_{n\to\infty} P(A_1 \cup A_2 \cup ... \cup A_n)^c = \boxed{\frac{1}{e}}.$$

c) **In the caps example, let X represent the number of graduates who retrieve their own cap. Explain why the distribution of X is approximately Poisson(1) when n is large (see 1(c)). Compare the exact probabilities of $X = 0$ and $X = 1$ to the corresponding Poisson probabilities when $n = 5$.**

Note that the probability that a given graduate receives their own cap (assuming nothing about all other graduate cap arrangements) is $p = \frac{1}{n}$. Though we are essentially sampling without replacement when throwing all the graduate caps in the air and retrieving them at random, we know from problem 1(b) that the distribution X for the number of graduates who receive their own cap converges to Binom(n, p) when n is large.

Furthermore, note that $\lambda = np = n\frac{1}{n} = 1$ is fixed. Consequently, the limit of $P(X = x)$ as n is large, by problem 1(c), is simply Poisson(λ) = Poisson(1). We conclude that the distribution of X is approximately Poisson(1) when n is large.

Using the approximation $X \sim$ Poisson(1) for large n, we can approximate the probability of a derangement

$$f_x(x = 0, \lambda = 1) = \frac{1^0 e^{-1}}{0!} = \frac{1}{e}$$

and the probability that exactly one graduate gets their own cap:

$$f_x(x = 1, \lambda = 1) = \frac{1^1 e^{-1}}{1!} = \frac{1}{e}.$$

We can also compare these approximations with the exact probabilities of $X = 0$ and $X = 1$ for $n = 5$. Note that the probability of a derangement for $n = 5$ can be calculated using the formula from 2(b):

$$\begin{aligned} P(X = 0) &= 1 - \left[5 \cdot \frac{1}{5} - \binom{5}{2}\frac{1}{5} \cdot \frac{1}{4} + \binom{5}{3}\frac{1}{5} \cdot \frac{1}{4} \cdot \frac{1}{3} - \binom{5}{4}\frac{1}{5} \cdot \frac{1}{4} \cdot \frac{1}{3} \cdot \frac{1}{2} + \binom{5}{5}\frac{1}{5} \cdot \frac{1}{4} \cdot \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{1}\right] \\ &= \frac{11}{30} \approx 0.3\overline{6} \end{aligned}$$

Note that the probability that $X = 1$ can be calculated by "picking" the graduate to get their own cap (of which there are 5 possibilites), multiplying this by the probability that the chosen graduate gets their own hat ($\frac{1}{5}$), and then multiplying this by the probability of a derangement with $n = 4$ (each of the remaining four graduates does not get their own cap):

$$\begin{aligned} P(X = 1) &= 5 \cdot \frac{1}{5} \cdot P(\text{derangement for } n = 4 \text{ students}) \\ &= 1 - \left[4 \cdot \frac{1}{4} - \binom{4}{2}\frac{1}{4} \cdot \frac{1}{3} + \binom{4}{3}\frac{1}{4} \cdot \frac{1}{3} \cdot \frac{1}{2} - \binom{4}{4}\frac{1}{4} \cdot \frac{1}{3} \cdot \frac{1}{2}\frac{1}{1}\right] \\ &= \frac{3}{8} = 0.375 \end{aligned}$$

We can see that even for small n (in our case, $n = 5$), the expected probabilities approach $\frac{1}{e} \approx 0.36788$.

2 Continuous Random Variables

2.1 Intro to Continuous Random Variables

Assumptions for a Poisson process in time include

- Events are independent
- Rate of events through time is constant
- Events are not simultaneous

Suppose that $T_1, \ldots, T_n$ are i.i.d. Expo(λ) random variables.

Definition 2.1
The **first order statistic**, defined to be the smallest of these, is $T_{(1)}$ follows the Expo($n\lambda$) distribution.

Proof. Consider the CDF of $T_{(1)}, P(T_{(1)} \leq t)$. Note that

$$\begin{aligned} P(T_{(1)} \leq t) &= 1 - P(T_{(1)} > t) \\ &= 1 - \left(e^{-\lambda t}\right)^n \\ &= 1 - e^{-n\lambda t} \end{aligned}$$

which is the CDF for an Expo($n\lambda$) distribution. $\square$

2.2 The Differential Argument and the Gamma Distribution

a) By definition,

$$P(X \in [x, x+dx)) = \int_x^{x+dx} f_x(y)\, dy.$$

For small deviation dx, we can approximate the integral as $f_x(x)\,[(x+dx) - x] = f_x(x)\, dx$.

Note that this expression becomes exact when dx approaches 0: formally, we have

$$\lim_{dx \to 0} \frac{P(X \in [x, x+dx))}{dx} = \lim_{dx \to 0} \frac{F(x+dx) - F(x)}{dx} = f_x(x)$$

because the second expression is the limit form of the derivative of the CDF at x, which is precisely the value of the PDF at x: $f_x(x)$.

We can use the Differential argument to derive the Exponential density function. Since the CDF of an Exponential variable at time t measures the probability that an event has occured by time t, we can measure the probability that the first Poisson event occurs in the time interval $[t, t+dt)$ by subtracting the CDF at times $t+dt$ and time t:

$$
\begin{aligned}
P(\text{first occurence in } [t, t+dt)) &= \left(1 - e^{-\lambda(t+dt)}\right) - \left(1 - e^{-\lambda t}\right) \\
&= e^{-\lambda t} - e^{-\lambda(t+dt)}
\end{aligned}
$$

Dividing this expression by dt and taking the limit as dt approaches 0, we derive the Exponential variable density at time t:

$$
f_t(t) = \lim_{dt \to 0} \frac{e^{-\lambda t} - e^{-\lambda(t+dt)}}{dt}.
$$

Using L'Hopital's Rule, this becomes

$$
\begin{aligned}
f_t(t) &= \lim_{dt \to 0} \frac{e^{-\lambda t} - e^{-\lambda(t+dt)}}{dt} \\
&= \lim_{dt \to 0} \frac{\lambda e^{-\lambda(t+dt)}}{1} \\
&= \lambda e^{-\lambda t}.
\end{aligned}
$$

b) We can similarly use the differential argument to derive the Gamma(k, λ) density; let X be the time of the kth event for a Poisson process with rate λ. We can model the probability of $k-1$ events occurring before time x using Pois(λx), where λx represents the expected number of events to occur in x units of time. Similarly, we can model the probability of 1 event occurring in the interval $[x, x+dx)$ using Pois(λdx).

Thus, the probability of $k-1$ events occurring before time x and a kth event occurring in $[x, x+dx)$ is

$$
\begin{aligned}
&P(k-1 \text{ events before time } x)P(1 \text{ event before time } t) \\
&= \frac{(\lambda x)^{k-1} e^{-\lambda x}}{(k-1)!} \frac{(\lambda dx)^1 e^{-\lambda dx}}{1!} \\
&= \frac{\lambda^k e^{-\lambda(x+dx)} dx}{(k-1)!} x^{k-1}
\end{aligned}
$$

Dividing this probability by dx and taking the limit as dx approaches 0, we find that

$$
\begin{aligned}
f_x(k, \lambda) &= \lim_{dx \to 0} \frac{\left(\frac{\lambda^k e^{-\lambda(x+dx)} dx}{(k-1)!} x^{k-1}\right)}{dx} \\
&= \lim_{dx \to 0} \frac{\lambda^k e^{-\lambda(x+dx)}}{(k-1)!} x^{k-1} \\
&= \frac{\lambda^k e^{-\lambda x}}{(k-1)!} x^{k-1}
\end{aligned}
$$

which matches the Gamma(k, λ) density we expect.

Note that if $X_1 \sim$ Gamma(k_1, λ) and $X_2 \sim$ Gamma(k_2, λ) are independent, intuitively, $X_1 + X_2 \sim$ Gamma$(k_1 + k_2, \lambda)$, as we can think of X_1 and X_2 representing the sum for waiting times in two non-overlapping time intervals. We will reserve discussion for why the sum of two independent Gamma variables with different λ values is not Gamma for part (d).

c) Recall the form of the Gamma PDF derived in part (b):

$$f_x(k, \lambda) = \frac{\lambda^k e^{-\lambda x}}{(k-1)!} x^{k-1}.$$

Note that our derivation measures the average wait time for k events to occur, where k is an integer. However, the Gamma Distribution also extends to values of k which are not necessarily integers – all positive real k, in fact. To understand the extension and subsequent definition of the Gamma Distribution, we introduce the notion of a Gamma Function, an extension of the factorial function for real numbers. For integer k, $\Gamma(k) = (k-1)!$, which matches with the normalizing constant in our present PDF.

In general, if α is a positive real value replacing our k parameter, we define the PDF of $X \sim$ Gamma(α, λ) to be

$$f_x(\alpha, \lambda) = \frac{\lambda^\alpha e^{-\lambda x}}{\Gamma(\alpha)} x^{\alpha-1}.$$

We can derive the definition of Gamma Function by forcing this PDF to integrate to 1 as x ranges from 0 to ∞. Note that

$$\int_0^\infty \frac{\lambda^\alpha e^{-\lambda x}}{\Gamma(\alpha)} x^{\alpha-1}\, dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{-\lambda x} x^{\alpha-1}\, dx.$$

Apply the change of variables $y = \lambda x$. It follows that $x = \frac{y}{\lambda}$ so $dx = \frac{1}{\lambda} dy$. Our integral now becomes

$$\frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{-\lambda x} x^{\alpha-1}\, dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{-y} \left(\frac{y}{\lambda}\right)^{\alpha-1} \left(\frac{1}{\lambda}\, dy\right).$$

Simplifying and setting the resulting expression to 1 (since our PDF should integrate to 1), we find that

$$\frac{\int_0^\infty e^{-y} y^{\alpha-1}\, dy}{\Gamma(\alpha)} = 1.$$

This inspires our definition of the **Gamma Function**:

$$\boxed{\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1}\, dy}$$

which is defined for any positive real values of the parameter α.

With the above definition of the Gamma Function, we can derive a nice recursive definition of the Gamma Function, one which will illuminate its relationship with the factorial function.

Note that by definition, $\Gamma(\alpha+1) = \int_0^\infty e^{-y} y^\alpha \, dy$.

Using Integration by Parts, we find that

$$\begin{aligned}\Gamma(\alpha+1) &= \left[y^\alpha(-e^{-y})\right]_0^\infty - \int_0^\infty (-e^{-y})\alpha y^\alpha \, dy \\ &= 0 + \alpha \int_0^\infty (e^{-y}) y^\alpha \, dy \\ &= \alpha\Gamma(\alpha).\end{aligned}$$

Since $\Gamma(1) = \int_0^\infty e^{-y} y^{1-1} \, dy = 1$ by definition, it follows recursively that $\Gamma(\alpha) = (\alpha-1)!$.

d) A final powerful property of the Gamma distribution is that for any constant $c > 0$, if $X \sim \text{Gamma}(\alpha, \lambda)$, then $Y = cX \sim \text{Gamma}\left(\alpha, \frac{\lambda}{c}\right)$. This property can be thought of as a changing in the units for the λ parameter: intuitively, if X measures the waiting time for α events to occur given a rate of λ in some unit of time (e.g seconds), $Y = cX$ may measure the waiting time for α events to occur given a rate of λ in a new unit of time (e.g. minutes, where $c = \frac{1}{60}$).

Note: This intuition, which relates the λ parameter to a given unit of time, also gives us an idea of why the sum of two Gamma variables with different λ values is not itself Gamma: we cannot sum the waiting times for events of processes measured in different units of time.

Suppose that $Y = cX$. Comparing the CDFs of X and Y, we find that

$$F_Y(y) = P(cX < y) = P\left(X < \frac{y}{c}\right) = F_X\left(\frac{y}{c}\right).$$

Taking the derivative of both sides, we find a relationship between the PDFs. Since $y = cx$, we have

$$f_Y(y) = \frac{1}{c} f_X\left(\frac{y}{c}\right) = \frac{1}{c} f_X(x).$$

The scaling property itself can be proved by comparing the PDFs for X and Y. Note that

$$f_x(\alpha, \lambda) = \frac{\lambda^\alpha e^{-\lambda x}}{\Gamma(\alpha)} x^{\alpha-1}.$$

Consider $\frac{1}{c}f_x(\alpha,\lambda) = \frac{1}{c}\frac{\lambda^\alpha e^{-\lambda x}}{\Gamma(\alpha)}x^{\alpha-1}$. We can rewrite this expression in a few equivalent forms:

$$\begin{aligned}
\frac{1}{c}f_x(\alpha,\lambda) &= \frac{1}{c}\frac{\lambda^\alpha e^{-\lambda x}}{\Gamma(\alpha)}x^{\alpha-1} \\
&= \left(\frac{1}{c}\right)^\alpha c^{\alpha-1}\frac{\lambda^\alpha e^{-\lambda x}}{\Gamma(\alpha)}x^{\alpha-1} \\
&= \left[\left(\frac{1}{c}\right)^\alpha \lambda^\alpha\right]\left(c^{\alpha-1}x^{\alpha-1}\right)\frac{\lambda^\alpha e^{-\frac{\lambda}{x}(cx)}}{\Gamma(\alpha)} \\
&= \left(\frac{\lambda}{c}\right)^\alpha (cx)^{\alpha-1}\frac{e^{-\frac{\lambda}{c}(cx)}}{\Gamma(\alpha)}
\end{aligned}$$

Note that the resulting expression is precisely the PDF for a Gamma $\left(\alpha, \frac{\lambda}{c}\right)$ RV, so

$$Y = cX \sim \text{Gamma}\left(\alpha, \frac{\lambda}{c}\right)$$

as desired.

2.3 The Uniform Distribution

Definition 2.2

Let F be the CDF for a continuous random variable. The **inverse function** $G(u)$ is defined to satisfy $G(u) = x$ if $F(x) = u$.

Suppose that $U \sim \text{Unif}(0,1)$. Then $X = G(U)$ has CDF F.

Remark. Let $U \sim \text{Unif}(0,1)$. Then $X = G(U) \sim \text{Expo}(\lambda)$.

2.4 The Normal and Chi-Square Distributions

Definition 2.3 (Chi-Square)

The sum of k independent squared standard normal variables follows a Chi-square distribution with k degrees of freedom, denoted $\chi^2_(k)$.

Remark. $\chi^2_(1) \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$.

In general, $\chi^2_(\nu) \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{1}{2}\right)$ due to the properties of a sum Gamma RVs.

Furthermore, due to the scaling property of Gamma RVs, if $X \sim \text{Gamma}(\alpha,\lambda)$, then $Y = 2\lambda X \sim \text{Gamma}\left(\alpha, \frac{1}{2}\right)$ which is equivalent to a $\chi^2_{2\alpha}$ random variable.

2.5 The Beta Distribution

Definition 2.4

The **Beta Distribution** is the conjugate prior of the binomial distribution; if we assume a coin flip has a probability of success following a Beta distribution, the posterior distribution, after accounting for new data, is also a Beta distribution.

Remark. Let $V_1 \sim \text{Gamma}(a, \lambda)$ and $V_2 \sim \text{Gamma}(b, \lambda)$. Then

$$X = \frac{V_1}{V_1 + V_2} \sim \text{Beta}(a, b).$$

Remark. Let $U_1, \ldots, U_n$ be i.i.d. Unif$(0, 1)$ random variables. Then the kth order statistic U_k follows a Beta$(k, n - k + 1)$ distribution for $k = 1, \ldots, n$.

2.6 The Beta, F, and t distributions

Definition 2.5 (Relationships between Gamma, Beta, and F^* Distributions)

Let $V_1 \sim \text{Gamma}(a, \lambda)$, $V_2 \sim \text{Gamma}(b, \lambda)$.

Then $X = \frac{V_1}{V_1+V_2} \sim \text{Beta}(a, b)$.

Then $R = \frac{V_1}{V_2} = \frac{X}{1-X}$ has PDF

$$f_R(r) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} r^{a-1} \left(\frac{1}{1+r}\right)^{a+b}.$$

Furthermore, $Y = c\frac{V_1}{V_2} = cR \sim F^*(a, b, c)$ has PDF

$$\begin{aligned} f_Y(y) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{y^{a-1}c^b}{(c+y)^{a+b}} I(y > 0) \\ &\propto \frac{y^{a-1}c^b}{(c+y)^{a+b}} I(y > 0) \end{aligned}$$

for $a > 0$, $b > 0$, $c > 0$. This is known as the $F^*(a, b, c)$ **density.**

Definition 2.6 (F-distribution)

The F**-distribution** is defined in terms of Chi-Square random variables: if $V_1 \sim \chi_(m_1)^2$ is independent of $V_2 \sim \chi_(m_2)^2$, then

$$F_{(m_1, m_2)} \sim \frac{\left(\frac{V_1}{m_1}\right)}{\left(\frac{V_2}{m_2}\right)}.$$

Remark. $F_{(n,m)} = F^* \left(\frac{n}{2}, \frac{m}{2}, \frac{m}{n}\right)$.

Definition 2.7

If $Z \sim N(0,1)$ is independent of $V \sim \chi^2_{(m)}$, then

$$T = \frac{Z}{\sqrt{\frac{V}{m}}} \sim t_(m),$$

the **t-distribution** with m degrees of freedom. Furthermore, $T^2 \sim F_{(1,m)}$.

3 Expected Value

3.1 Definition of Expected Value

Definition 3.1
For a discrete RV X, the **expected value of** X is $\mathbb{E}[X] = \sum_i x_i P(x_i)$.

For a continuous RV X, $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx$.

Remark. The expected value of a random variable may not always be defined – the integral may be unbounded, the integrand may be nonintegrable, or the random variable could be both discrete and continuous.

Example. $X \sim \text{Cauchy}(1)$ has pdf $f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}$ has undefined expected value, as the absolute integral diverges.

3.2 Linearity of Expectation

Definition 3.2 (Linearity of Expectation)
If $X_1, X_2, \ldots, X_n$ are random variables, then

$$\mathbb{E}[X_1 + X_2 + \cdots + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]$$

regardless of whether $X_1, X_2, \ldots, X_n$ *are independent or not.*

Example. Suppose that n missiles are targeted independent by n intercepts, each choosing a target at random. Find the expected number of missiles targets.

Answer. Define an indicator variable I_k for each $k \in \{1, \ldots, n\}$ such that

$$I_k = \begin{cases} 1 & \text{if missile } k \text{ is targeted} \\ 0 & \text{otherwise} \end{cases}$$

Define X to be the number of missiles targeted. Note that

$$\mathbb{E}[X] = \mathbb{E}[I_1) + \cdots + \mathbb{E}[I_n).$$

Note that $\mathbb{E}[I_k)$ for any k is $1 - \left(\frac{n-1}{n}\right)^n$, so $\mathbb{E}[X] = n(1 - \left(\frac{n-1}{n}\right)^n)$.
As $n \to \infty$, it follows that

$$\mathbb{E}[X] = \lim_{n\to\infty} n\left[1 - \left(\frac{n-1}{n}\right)^n\right] = n\left(\frac{e-1}{e}\right).$$

⊛

3.3 LOTUS and Variance

Theorem 3.1 (LOTUS (Law of the Unconscious Statistician))**.** For a random variable X, and a fixed function g then $Y = g(x)$ has

$$\mathbb{E}[Y] = \begin{cases} \sum_x g(x)p(x) & \text{if X is discrete} \\ \int_{-\infty}^{\infty} g(x)f(x)\,dx & \text{if X is continuous} \end{cases}$$

where $p(x)$ and $f(x)$ are the pmf and pdf of X, respectively.

This holds only if $\sum_x |g(x)|p(x)$ or $\int_{-\infty}^{\infty} |g(x)|f(x)\,dx$ converge.

Theorem 3.2 (Markov's Inequality)**.** For a random variable X,

$$P(X \geq t) \leq \frac{\mathbb{E}\left[X\right]}{t}.$$

Theorem 3.3 (Chebyshev's Inequality)**.** For a random variable X with mean μ and variance σ^2 and for $t > 0$,

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

3.4 Moment Generating Functions

Definition 3.3 (Moment Generating Function)

The **Moment Generating Function** of a random variable X is

$$M_X(t) = \mathbb{E}\left[e^{tX}\right].$$

The kth derivative of $M_x(t)$ at $t = 0$ is $\mathbb{E}\left[X^k\right]$, the kth moment of the distribution of X.

Theorem 3.4 (Uniqueness Property of MGFs)**.** If two random variables X and Y have the same MGFs for all t in an interval containing 0, then $F_X(x) = F_Y(x)$ i.e. they must have the same distribution for all x.

Remark. $M_{a+bX}(t) = e^{at}M_X(bt)$.

Furthermore, for independent random variables $X_1, \ldots, X_n$, each with MGF M_{X_i}, the MGF for $Y = \sum X_i$ is

$$M_Y(t) = \prod M_{X_i}(t).$$

3.5 Inequalities and Approximate Methods

Theorem 3.5 (Jensen's Inequality)**.** If g is a convex function, then

$$\mathbb{E}\left[g(X)\right] \geq g(\mathbb{E}\left[X\right]).$$

If g is concave, then

$$\mathbb{E}\left[g(X)\right] \leq g(\mathbb{E}\left[X\right]).$$

Using Taylor approximations, we can derive approximations for the mean and variance of a random variable under a transformation g.

Lemma 3.1. If X is a random variable with mean μ_X and variance σ_X^2 and the transformed variable $Y = g(X)$,

$$\mathbb{E}\left[y\right] \approx g(\mu_x) + \frac{1}{2}\sigma_X^2 g''(\mu_X)$$

and

$$\mathrm{Var}\left[Y\right] \approx g'(\mu_X)^2 \mathrm{Var}\left[X\right].$$

3.6 Conditional Expectation

Remark. Intuitively (without working with the definitions), conditional expectations given events and random variables are different: conditional expectation given an event is a value whereas conditional expectation given a random variable yields a random variable.

Theorem 3.6 (Law of Total Expectation: Adam's Law)**.** $\mathbb{E}\left[Y\right] = \mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right]$

Theorem 3.7 (Law of Total Variance: Eve's Law)**.**

$$\mathrm{Var}\left[Y\right] = \mathbb{E}\left[\mathrm{Var}\left[Y \mid X\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[Y \mid X\right]\right]$$

4 Joint and Conditional Distributions

4.1 Joint, Marginal, and Conditional Distributions

Definition 4.1 (Joint, Marginal, and Conditional)
The **joint pdf** describes the probability density of observing both $X = x$ and $Y = y$ simultaneously.

The **marginal pdf of** Y represents the probability density of observing $Y = y$, irrespective of the value of X.

The **conditional pdf of** Y given $X = x$ represents the probability density of observing $X = x$, given that $Y = y$.

4.2 Covariance

Definition 4.2 (Covariance and Correlation)
The **covariance of** X **and** Y is defined as

$$\begin{aligned}\operatorname{Cov}[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\end{aligned}$$

The **correlation of** X **and** Y is

$$\rho = \frac{\operatorname{Cov}[X, Y]}{\sigma_X \sigma_Y}.$$

where $-1 \leq \rho \leq 1$.

4.3 Multivariate Normal

a) A random vector $\mathbf{X} = (X_1, \ldots, X_k)$ is said to have a Multivariate Normal (MVN) distribution if every linear combination of the X_j follows a Normal distribution. That is,

$$t_1 X_1 + \cdots + t_k X_k$$

follows a Normal distribution for any choice of constants $t_1, \ldots, t_k$.

This definition may allow for vectors that do not have a proper joint density function: this is more clear with the definition of the joint density function below (the covariance matrix for a vector with correlated entries will not be invertible).

b) If $\mathbf{Y} \sim N_n(\mu, \mathbf{V})$, then the joint PDF for $\mathbf{y} = (y_1, \ldots, y_n)$ is

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{\exp\left[-\frac{1}{2}(\mathbf{y} - \mu)^T \mathbf{V}^{-1} (\mathbf{y} - \mu)\right]}{\sqrt{(2\pi)^n |\det \mathbf{V}|}}$$

It follows that this is an extension of the bivariate Normal density to larger dimensions. Note that the bivariate case is a specific case of the Multivariate Normal where $n = 2$. If $\mathbf{Y} \sim N_2(\mu, \mathbf{V})$, then $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and $\mathbf{V} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$.

One can check that indeed, the expression

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \frac{\exp\left[-\frac{1}{2}(\mathbf{y}-\mu)^T\mathbf{V}^{-1}(\mathbf{y}-\mu)\right]}{\sqrt{(2\pi)^n|\det \mathbf{V}|}} \\ &= \frac{\exp\left[-\frac{1}{2}\begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix}^T \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}^{-1} \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix}\right]}{\sqrt{(2\pi)^n|\sigma_1^2\sigma_2^2(1-p^2)|}}, \end{aligned}$$

obtained by plugging in the respective expressions for the bivariate case, matches the bivariate Normal density. The calculation/algebra is left as an exercise to the reader.

c) Suppose that V is block diagonal; without loss of generality, we can assume that

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & 0 \\ 0 & \mathbf{V}_2 \end{bmatrix}.$$

By properties of block matrices, we know that $\det \mathbf{V} = (\det \mathbf{V}_1)(\det \mathbf{V}_2)$. Furthermore,

$$\mathbf{V}^{-1} = \begin{bmatrix} \mathbf{V}_1 & 0 \\ 0 & \mathbf{V}_2 \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{V}_1^{-1} & 0 \\ 0 & \mathbf{V}_2^{-1} \end{bmatrix}$$

Suppose that $Y_1 \sim N_{n_1}(\mu_1, \mathbf{V}_1)$ and $Y_2 \sim N_{n_2}(\mu_2, \mathbf{V}_2)$, where $n_1 + n_2 = n$, μ_1 and μ_2 form μ of Y, and similarly, $\mathbf{V}_1$ and $\mathbf{V}_2$ form $\mathbf{V}$. Note that

$$\exp\left[-\frac{1}{2}(\mathbf{y}-\mu)^T\mathbf{V}^{-1}(\mathbf{y}-\mu)\right] = \exp\left[-\frac{1}{2}(\mathbf{y}_1-\mu_1)^T\mathbf{V_1}^{-1}(\mathbf{y}_1-\mu_1)\right]\exp\left[-\frac{1}{2}(\mathbf{y}_2-\mu_2)^T\mathbf{V_2}^{-1}(\mathbf{y}_2-\mu_2)\right].$$

Furthermore,

$$\sqrt{(2\pi)^n|\det \mathbf{V}|} = \sqrt{(2\pi)^{n_1+n_2}|\det \mathbf{V_1}\det \mathbf{V}_2|}.$$

It follows that

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \frac{\exp\left[-\frac{1}{2}(\mathbf{y}-\mu)^T\mathbf{V}^{-1}(\mathbf{y}-\mu)\right]}{\sqrt{(2\pi)^n|\det \mathbf{V}|}} \\ &= \frac{\exp\left[-\frac{1}{2}(\mathbf{y}_1-\mu_1)^T\mathbf{V_1}^{-1}(\mathbf{y}_1-\mu_1)\right]}{\sqrt{(2\pi)^{n_1}|\det \mathbf{V_1}|}}\frac{\exp\left[-\frac{1}{2}(\mathbf{y}_2-\mu_2)^T\mathbf{V_2}^{-1}(\mathbf{y}_2-\mu_2)\right]}{\sqrt{(2\pi)^{n_2}|\det \mathbf{V_2}|}} \\ &= f_{\mathbf{Y}_1}(\mathbf{y}_1)f_{\mathbf{Y}_2}(\mathbf{y}_2) \end{aligned}$$

where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are two independent sub-vectors of $\mathbf{Y}$.

d) Consider the covariance between $\overline{Y}$ and $Y_i - \overline{Y}$ for any i from 1 to n. Note that

$$\text{Cov}\left[\overline{Y}, Y_i - \overline{Y}\right] = \text{Cov}\left[\overline{Y}, Y_i\right] - \text{Cov}\left[\overline{Y}, \overline{Y}\right]$$

by the linearity property of covariance. Since $\overline{Y} = \sum_{i=1}^{n} Y_i$, it follows that

$$\begin{aligned}
\operatorname{Cov}\left[\overline{Y}, Y_i - \overline{Y}\right] &= \operatorname{Cov}\left[\overline{Y}, Y_i\right] - \operatorname{Cov}\left[\overline{Y}, \overline{Y}\right] \\
&= \operatorname{Cov}\left[\frac{1}{n} Y_i + \frac{1}{n} \sum_{j \mid i \neq j \,\&\, j \geq 1}^{n} Y_j, Y_i\right] - \frac{\sigma^2}{n} \\
&= \frac{1}{n} \operatorname{Cov}\left[Y_i, Y_i\right] - \frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0.
\end{aligned}$$

Thus, we find that $\operatorname{Cov}\left[\overline{Y}, Y_i - \overline{Y}\right] = 0$ for $i = 1, \ldots, n$, so $\overline{Y}$ is uncorrelated with each $Y_i - \overline{Y}$.

Since

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

by definition, and $\overline{Y}$ is independent from each $Y_i - \overline{Y}$, it follows that $\overline{Y}$ is independent of s^2.