Probability Notes

June 4, 2020

Contents

1 Probability Theorems

1.1 Set Theorems

For any three sets, the following hold true

$$A = (A \cap B) \cup (A \cap B^c) \text{ where } B \text{ and } B^c \text{ are disjoint}$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

1.2 Basic Probability Rules

If $A \cap B = \phi$, then $P(A \cup B) = P(A) + P(B)$

$P(A|B)P(B) = P(B|A)P(A) = P(A \cap B)$ Bayes' Theorem

$P(A) = P(A \cap B) + P(A \cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c)$

$P(A \cap B \cap C) = P(A)P(B|A)P(C|B, A)$ Chain Rule

1.2.1 Total Probability Theorem

Let A_1, A_2, .. ,A_n be n disjoint events that completely cover the event space, and B be another event, then

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n)$$
$$\text{or, } P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

1.3 Independence

Two events A and B are independent iff

$$P(A \cap B) = P(A)P(B)$$

Note that *independence* is not the same as *disjoint*

$$A \cap B = \phi \Rightarrow P(A \cap B) = 0 \text{ but } P(A) \neq P(B) \neq 0$$

Multiple events $A_1, A_2, \ldots, A_n$ are independent iff

$$P(A_i \cap A_j \cap \ldots \cap A_k) = P(A_i)P(A_j) \,..\, P(A_k) \;\; \forall\, i, j, \ldots, k \mid i, j, \ldots, k \in 1, 2, \ldots, n$$

Conditional Independence is similar to the above equation. For an event C,

$$P(A_i \cap A_j \cap \ldots \cap A_k|C) = P(A_i|C)P(A_j|C) \,..\, P(A_k|C) \;\; \forall\, i, j, \ldots, k \mid i, j, \ldots, k \in 1, 2, \ldots, n$$

1.4 Joint Probability Distributions

Joint Probability Distributions are defined for two or more than two variables. In this section, we only consider two variables. The formal definition is

$$P_{XY}(x, y) = P(X = x \text{ and } Y = y)$$

Based on this definition, the following theorems follow

$$\sum_x \sum_y P_{XY}(x,y) = 1$$

$$P_X(x) = \sum_y P_{XY}(x,y) \qquad \text{Marginal Probability}$$

$$P_{X|Y}(x|y) = P_{X|Y}(X = x|Y = y) = \frac{P_{XY}(x,y)}{P_Y y}$$

$$\sum_x P_{X|Y}(x|y) = 1 \qquad \text{Since Y is fixed and we sum over all X's}$$

$$P_{XYZ}(x,y,z) = P_X(x)P_{Y|X}(y|x)P_{Z|X,Y}(z|x,y) \qquad \text{Chain Rule}$$

1.5 Expected Value

Before going to expected value, let's define a Random Variable

Random Variable X is a linear map : $\mathbb{R} \to \mathbb{R}$. The value taken by the variable is denoted by x

X will have an associated probability distribution, i.e., $P_X(X = x)$. Using these quantities, we have

$$E[X] = \sum_x x P_X(X = x) \qquad \text{Expected Value}$$

Based on this definition, the following theorems for expected value follow

$$E[\alpha] = \alpha E[\alpha X] = \alpha E[X]$$

$$E[\alpha X + \beta] = \alpha E[X] + \beta$$

$$E[g(X)] = \sum_x g(x) P_X(X = x)$$

$$E[X^2] = \sum_x x^2 P_X(X = x) \qquad \text{Also called } \textit{Second Moment}$$

$$E[X|A] = \sum_x x P_{X|A}(X|A)$$

$$E[g(X)|A] = \sum_x g(x) P_{X|A}(X|A)$$

$$E[X + Y + Z] = E[X] + E[Y] + E[Z] \qquad \text{Linearity of Expectation}$$

$$E[XY] = \sum_X \sum_Y xy P_{XY}(x,y)$$

$$E[g(X,Y)] = \sum_X \sum_Y g(xy) P_{XY}(x,y)$$

$$E[XY] = E[X]E[Y] \qquad \text{if X and Y are } \textit{independent}$$

where $\alpha, \beta \in \mathbb{R}$, $g(X) : \mathbb{R} \to \mathbb{R}$, and A is an event, X, Y, Z are Random Variables

1.5.1 Total Expectation Theorem

The *Total Expectation Theorem* is the natural extension of the *Total Probability Theorem* Let $A_1, A_2, \ldots, A_n$ be n disjoint events that completely cover the event space, and X be random variable, then

$$E[X] = E[X|A_1]P(A_1) + E[X|A_2]P(A_2) + \cdots + E[X|A_n]P(A_n)$$

$$\text{or, } E[X)] = \sum_{i=1}^{n} E[X|A_i]P(A_i)$$

1.6 Variance

The formal definition of variance is

$$Var(X) = E[(X - \bar{X})^2] = E[X^2] - E[X]^2$$

Using this definition, the following theorems follow

$$\begin{aligned} E[X^2] &= E[X]^2 + Var(X) \\ Var(\alpha) &= 0 \\ Var(\alpha X + \beta) &= \alpha^2 Var(X) \\ Var(X + Y) &= Var(X) + Var(Y) \text{ if } X \text{ and } Y \text{ are } \textit{independent} \text{ random variables} \end{aligned}$$

1.7 Cumulative Probability Distribution

Cumulative probability distribution is defined for both discrete and continuous variables

$$F_x(X) = P(X \leq x) = \begin{cases} \int_{-\inf}^{x} p_X(t)dt & X \text{ is a discrete random variable} \\ \sum_{k<=x} P_X(k) & X \text{ is a continuous random variable} \end{cases}$$

2 Covariance and Correlation

For any two random variables X and Y,

$$\begin{aligned} Cov(X, Y) &= E[(X - \overline{X})(Y - \overline{Y})] = E[XY] - E[X]E[Y] \\ Cov(X, X) &= Var(X) \\ Corr(X, Y) &= E[(\frac{X - \overline{X}}{\sigma_X})(\frac{Y - \overline{Y}}{\sigma_Y})] \\ &= \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \end{aligned}$$

Key points to note

- *Inpdependence* $\Rightarrow Cov(X, Y) = Corr(X, Y) = 0$, but the converse is ***not*** true
- Correlation is dimensionless and $-1 \leq Corr(X, Y) \leq 1$ with value close to 0 implying minimal relation and values close to $-1, 1$ implying perfect relation

3 Iterated Expectation and Variance

The law of iterated expectation tells the following about expectation and variance

$$\begin{aligned} E[E[X|Y]] &= E[X] \\ Var(X) &= E[Var(X|Y)] + Var(E[X|Y]) \end{aligned}$$

Proof for Iterated Expectation

$$\begin{aligned} P(X) &= \sum_y P(X|Y)P(Y) \\ \Rightarrow E[X] &= \sum_x xP(X) = \sum_x \sum yxP(X|Y)P(Y) \\ &= \sum_y P(Y) \sum_x xP(X|Y) = \sum_y P(Y)E[X|Y] \\ \text{or, } E[X] &= E[E[X|Y]] \qquad\qquad E[X|Y] \text{ is a function of } X \text{ and not } Y \end{aligned}$$

Proof for Variance

$$
\begin{aligned}
Var(X) &= E[X^2] - E[X]^2 \\
Var(X|Y) &= E[(X - \overline{X})^2|Y] = E[X^2|Y] - E[X|Y]^2 && 1 \\
Var[E(X|Y)] &= E[E(X|Y)^2] - E[E[X|Y]]^2 \\
&= E[E[(X|Y)]^2] - E[X]^2 && 2 \\
E[Var(X|Y)] &= E[E[X^2|Y]] - E[E[X|Y]^2] && \text{from 1} \\
&= E[X^2] - E[E[X|Y]^2] && 3 \\
E[Var(X|Y)] + Var(E[X|Y]) &= E[X^2] - E[X]^2 && \text{adding 2 and 3} \\
&= Var(X)
\end{aligned}
$$

4 Random number of Random Variables

Let X_i be independent identically distributed Random Variables and let $Y = \sum_{i=1}^{N} X_i$ be the sum of N such random variables where N itself is a random variable. Then,

$$
\begin{aligned}
Y &= X_1 + X_2 + \cdots + X_N \\
E[Y|N = n] &= \sum_{i=1}^{n} E[X_i] \\
&= NE[X] \\
E[Y] &= E[E[Y|N]] = E[NE[X]] \\
&= E[N]E[X] && \text{since } E[X] \text{ will be a number} \\
Var(Y) &= E[Var(Y|N)] + Var(E[Y|N]) \\
&= E[NVar(X)] + Var(NE[X]) \\
&= E[N]Var(X) + E[X]^2 Var(N)
\end{aligned}
$$

5 Convolutions

Convolution operations are defined for both CDF and PDF/PMFs. Let X and Y be random independent variables, then

$$
\begin{aligned}
F_{X+Y}(x) &= F_X * F_Y = \int_{\mathbb{R}} F_X(x - y) dF_Y(y) \\
p_{X+Y}(x) &= p_X * p_Y = \int_{\mathbb{R}} p_X(x - y) p_Y(y) dy
\end{aligned}
$$

We can extend the idea to n independent variables as

$$F_X^{n*} = F_X * \cdots * F_X \ n \text{ times}$$

It has the following properties for positive random variable X_is

1.

$$F_X^{n*}(x) \leq F_X^n(x)$$

This can be proven as

$$P(X_1 + \cdots + X_n \leq x) \leq P(X_1 \leq x, \ldots, X_n \leq x)$$
$$P(X_1 + \cdots + X_n \leq x) \leq \prod_{i=1}^{n} P(X \leq x) \text{ by independence}$$
$$\text{or, } F_X^{n*}(x) \leq F_X^n(x)$$

2.

$$F_X^{n*}(x) \geq F_X^{n+1}(x)$$

which follows immediately from the fact that

$$P(X_1 + \cdots + X_n \leq x) \geq P(X_1 \leq x, \ldots, X_{n+1} \leq x)$$

since the volume of the regions denoting the sums will be lower in the higer dimensions. This can be quickly verified by considering $X_1 \leq 1$ and $X_1 + X_2 <= 1$.

6 Moment Generating Function

Moment generating function is defined as the following for all values of t

$$\phi(t) = E[e^{tX}] = \begin{cases} \sum_x e^{tx} p_X(x) & \text{for discrete case} \\ \int_{-\inf}^{\inf} e^{tx} f_X(x) & \text{for continuous case} \end{cases}$$

This function is called the moment generating function because all the moments of the random variable X can be obtained by successively differentiating the function $\phi(t)$.

$$\begin{aligned} \phi'(t) &= \frac{d}{dt} E[e^{tX}] \\ &= E[\frac{d}{dt} e^{tX}] \\ &= E[Xe^{tX}] \\ \text{mean} &= E[X] \\ &= \phi'(0) \end{aligned}$$

Continuing in a similar fashion,

$$\begin{aligned} \phi''(t) &= \frac{d}{dt} E[Xe^{tX}] \\ &= E[\frac{d}{dt} Xe^{tX}] \\ &= E[X^2 e^{tX}] \\ \text{variance} &= \phi''(0) \\ &= E[X^2] \end{aligned}$$

In general, for any $n > 0$, the n^{th} derivative will give the n^{th} moment

$$\phi^n(0) = E[X^n]$$

There exists a one to one correspondence between the moment generating function and the distribution function of a random variable, similar to Lagrangian multipliers.

6.1 Moment Generating Function for Sum of Independent RV

An important property is in the context of sum of two or more random variables. The **moment generating of sum of two independent random variables is simply the product of the moment generating functions of the two individual random variables**

$$\begin{aligned}\phi_{X+Y}(t) &= E[e^{t(X+Y)}] \\ &= E[e^{tX}e^{tY}] \\ &= E[e^{tX}]E[e^{tY}] \\ \boxed{\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)} &\text{ for independent random variables}\end{aligned}$$

7 Binomial Random Variable

Binomial Random Variable X is defined as the number of successes in an experiment with n independent trials, where each trial can only have two outcomes, *success* or *failure*.
Let X_i denote the Random Variable corresponding to the individual trials, with probability of success p. Then we have the following

$$X_i = \begin{cases} 1 & \text{if success in trial i} \\ 0 & \text{otherwise} \end{cases} \qquad \text{indicator variable}$$

$$X = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

$$P(X = k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

7.1 Mean and Variance

First let's calculate the mean and variance for a single trial X_i

$$\begin{aligned} E[X_i] &= 1 * p + 0 * (1-p) &&= p \\ Var(X_i) &= (1-p)^2 p + (0-p)^2(1-p) &&= p(1-p) \end{aligned}$$

We know that all $X_i's$ are independent. Hence, the mean and variance for X become

$$\begin{aligned} E[X] = E[\sum_{i=1}^{n} X_i] &= \sum_{i=1}^{n} E[X_i] &&= np \\ Var(X) = Var(\sum_{i=1}^{n} X_i) &= \sum_{i=1}^{n} Var(X_i) &&= np(1-p) \end{aligned}$$

8 Continuous Uniform Random Variable

A uniform random variable is defined as follows

$$f_X(x) = \begin{cases} \frac{1}{b-a} & if a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

8.1 Mean and Variance

$$\begin{aligned} E[X] &= \int_a^b x\frac{1}{b-a}dx = [\frac{x^2}{2(b-a)}]_a^b \\ &= \frac{a+b}{2} \\ Var(X) &= \int_a^b (x - \frac{a+b}{2})^2 \frac{1}{b-a}dx \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

9 Gaussian Distribution

The gaussian distribution (or normal distribution) is defined between $-\inf$ and inf. It is parametrized by mean μ and variance σ, $X \sim \mathcal{N}(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

As already described,

$$\begin{aligned} E[X] &= \mu \\ Var(X) &= \sigma^2 \end{aligned}$$

A *Standard Normal* is defined as a normal distribution with $\mu = 0$ and $\sigma^2 = 1$
Any normal distribution can be converted to a standard normal as $X = \frac{X-\mu}{\sigma}$
If $Y = aX + b$, then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

10 Counting Process

Counting process is used in scenarios when we want to count the occurrence of a certain event. N_t denotes the number of events till time t starting from 0. It is assumed that $N_0 = 0$. Formal definition is

A random process $\{N_t, t \in [0, \inf)\}$ is said to be a counting process if N_t is the number of events from time $t = 0$ upto time t. For a counting process, we assume

1. $N_0 = 0$
2. $N_t \in \{0, 1, 2, \cdots\}$ for all $t \in [0, \inf)$
3. for $0 \leq s < t, N_t - N_s$ shows the number of events that occur in the interval $(s, t]$

10.1 Independent Increments

We say that a continuous time counting process N_t has independent increments if for all $0 \leq t_1 < t_2 < \cdots < t_n$, the random variables

$$N_{t_2} - N_{t_1},\ N_{t_3} - N_{t_2}, \ldots,\ N_{t_n} - N_{t_{n-1}}$$

are independent.

Note that these differences are nothing but the number of arrivals in a given time interval. Thus, we are equivalently saying that **the number of arrivals in any two disjoint intervals are**

independent.

A very simple consequence of this property is:
Suppose we wise to find the probability of 2 arrivals in the interval $(1, 2]$ and 3 arrivals in the interval $(3, 5]$. Then,

$$P(\text{2 arrivals in } (1,2] \text{ and 3 arrivals in } (3,5]) = P(\text{2 arrivals in } (1,2])P(\text{3 arrivals in } (3,5]))$$

since the arrivals in disjoint intervals are independent.

10.2 Stationary Increments

We say that a continuous time counting process N_t has stationary increments if for all $t_2 > t_1 \geq 0$ and for all $r > 0$, $N_{t_2} - N_{t_1}$ and N_{t_2+r} and N_{t_1+r} are independent.

In other words, **the number of arrivals in a given time interval is invariant to it's location**. Note that the number of arrivals in the time interval between t_1 and t_2 is nothing but $N_{t_2} - N_{t_1}$. By the above statement, if the process has stationary increments, then this quantity is same as $N_{t_2-t_1}$, which is the distribution of the counting process itself.

11 Renewal Process

This is a fundamental stochastic process useful in modelling arrivals and interarrival times. Some definitions will make the usage clear.

Let S_i denote the ith renewal time or the time when the ith arrival takes place. By definition, $S_0 = 0$. We can also define

$$\begin{aligned} S_n &= S_{n-1} + \xi_n \\ S_n &= \xi_1 + \xi_2 + \cdots + \xi_{n-1} \end{aligned}$$

where ξ_i are positive ($P(\xi > 0) = 1$) independent identically distributed variables representing the interarrival times. We also define

$$\begin{aligned} N_t &= \underset{k}{\text{argmax}}\{S_k \leq t\} \\ \{S_n > t\} &= \{N_t < n\} \end{aligned}$$

or, N_t is simply the number of arrivals till the time t.

Define the following quantity

$$\begin{aligned} F^{n*} &= F_\xi * \ldots * F_\xi \ n \text{ times} \\ u(t) &= \sum_{i=1}^{\inf} F^{n*}(t) \end{aligned}$$

It can be shown that the function $u(t)$ converges. The expectation of N_t then becomes

$$\begin{aligned}
E[N_t] &= E[\text{number of } n \text{ such that } S_n \le t] \\
&= E[\sum_{n=1}^{\inf} I(S_n \le t)] && \text{sum of Indicators will equal } n \\
&= \sum_{n=1}^{\inf} P(S_n \le t) && \text{since } E[\text{Indicator}] \text{ is just the function inside indicator} \\
&= \sum_{n=1}^{\inf} F^{n*}(t) && \text{by defining cumulative as sum of } \xi\text{s} \\
&= u(t)
\end{aligned}$$

11.1 Laplace Transform

For a density function f defined from $\mathbb{R}^{\ge 0} \to \mathbb{R}$, Laplace transform is

$$L_f(s) = \int_{\mathbb{R}^{\ge 0}} e^{-sx} f(x) dx$$

The following properties hold for this transform

1. If f is a probability density function, then

$$E[e^{-sx}] = L_f(s)$$

2. if f_1 and f_2 are two probability density functions, then

$$L_{f_1 * f_2}(s) = L_{f_1}(s) L_{f_2}(s)$$

3. If F is the cumulative probability distribution for X and p is the probability density function, then

$$L_{F_X}(s) = \frac{L_{p_X}(s)}{s}$$

which can be proven using integration by parts as follows

$$L_{F_X}(s) = \int_{\mathbb{R}^{\ge 0}} F_X(x) \frac{d(e(-sx))}{s} = 0 + \frac{1}{s} \int_{\mathbb{R}^{\ge 0}} p_X(x) e^{-sx} dx$$

11.2 Calculating the Expectation

Armed with the concept of a Laplace transform, we make the following observation first

$$\begin{aligned}
u(t) &= \sum_{i=1}^{\inf} F^{n*}(t) = F(t) + \sum_{i=2}^{\inf} F^{n*}(t) \\
&= F(t) + \big(\sum_{i=1}^{\inf} F^{n*}(t) \big) * F(t) \\
&= F(t) + u(t) * F(t) \\
u(t) &= F(t) + u(t) * p(t)
\end{aligned}$$

where p is the probability density function and the last line stems from the fact that $\int u * F = \int u(x-y) dF(y) = \int u(x-y) p(y) dy$. Taking Laplace transform on both sides,

$$L_u(s) = L_F(s) + L_u(s)L_p(s)$$
$$L_u(s) = \frac{L_p(s)}{s} + L_u(s)L_p(s) \text{ from 3}$$
$$L_{u(s)} = \frac{L_p(s)}{s(1 - L_p(s))}$$

The last equation can be used to calculate the laplace transform of $u(t)$ and consecutively guess the functional form of $u(t)$.

11.3 Limit Theorems for Renewal Processes

The following two theorems hold true for Renewal processes

1. If $E[\xi] = \mu < \text{inf}$, then

$$\lim_{t\to\text{inf}} \frac{N_t}{t} = \frac{1}{\mu}$$

which is analogous to the strong law of large numbers. This can be proven as follows

$$S_{N_t} \leq t \leq S_{N_t+1} \text{ from the definition of } N_t$$
$$\text{or, } \frac{N_t}{S_{N_t+1}} \leq \frac{N_t}{t} \leq \frac{N_t}{S_{N_t}}$$

we can calculate the limits on the two bounds as

$$\lim_{t\to\text{inf}} \frac{N_t}{S_{N_t}} = \lim_{n\to\text{inf}} \frac{n}{S_n} = \frac{1}{\mu}$$

from the strong law of large numbers applied to $\lim_{n\to\text{inf}} \frac{S_n}{n}$. Similarly, one can show

$$\lim_{t\to\text{inf}} \frac{N_t}{S_{N_t+1}} = \lim_{t\to\text{inf}} \frac{N_t}{N_t+1} \lim_{t\to\text{inf}} \frac{N_t+1}{S_{N_t+1}} = 1 * \frac{1}{\mu}$$

2. If $Var(\xi) = \sigma^2 < \text{inf}$, then

$$\lim_{t\to\text{inf}} \frac{N_t - t/\mu}{\sigma\sqrt{t}/\mu^{3/2}} = \mathcal{N}(0,1)$$

which is analogous to the central limit theorem. It can be proven by considering the CLT on ξs

$$\lim_{n\to\text{inf}} P(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x) = \text{CDF of } \mathcal{N}(0,1)$$
$$\text{or, } \lim_{n\to\text{inf}} P(S_n \leq n\mu + \sigma\sqrt{n}x) = \text{CDF of } \mathcal{N}(0,1)$$
$$\text{or, } \lim_{n\to\text{inf}} P(N_t \geq n) = \text{CDF of } \mathcal{N}(0,1) \text{ from definition of } N_t, \text{ where } t = n\mu + \sigma\sqrt{n}x$$

We substitute $n\mu = t$ for very large value of n, since the total time will become total variables into the expected time for one ξ when n is large. Hence,

$$n = \frac{t}{\mu} - \frac{\sigma\sqrt{t}}{\mu^{3/2}}x$$
$$\lim_{n\to\text{inf}} P(N_t \geq n) = \lim_{n\to\text{inf}} P(\frac{N_t - t/\mu}{\sigma\sqrt{t}/\mu^{3/2}} \leq x) = \text{CDF of } \mathcal{N}(0,1)$$

12 Bernoulli Process

Bernoulli process falls under the family of random processes, which are random variables continuously evolving over time. Bernoulli process can be described as a sequence of independent Bernoulli trials, where each trial has only two outcomes : success with $P(success) = p$ and failure.

$$P_{X_t}(x_t) = \begin{cases} p & \text{if } X_t = 1 \\ 1-p & \text{if } X_t = 0 \end{cases}$$
$$E[X_t] = p$$
$$Var(X_t) = p(1-p)$$

12.1 Mean and Variance

Number of successes S in n time slots

$$P(S = k) = \binom{n}{k} p^k (1-p)^{n-k}$$
$$E[S] = np$$
$$Var(S) = np(1-p)$$

12.2 Interarrival Times (Geometric Random Variable)

Let T_1 denote the number of trials till the first success

$$P(T_1 = t) = (1-p)^{t-1}p \qquad t \in 1, 2, \ldots$$
$$E[T_1] = \frac{1}{p}$$
$$Var(T_1) = \frac{1-p}{p^2}$$

This process is memoryless as all future coin flips are independent of whatever has happened till now. Also, the distribution is a **Geometric Random Variable**.

12.3 Sum of Interarrival times

We are interested in the total time till k arrivals. Let this random variable be Y_k

$$Y_k = T_1 + T_2 + \cdots + T_k \qquad \text{where } T_i\text{'s are i.i.d geometric with parameter } p$$
$$P(Y_k = t) = P(k-1 \text{ arrivals between } t = 1 \text{ to } t = t \text{ and last arrival at time } t)$$
$$= \binom{t-1}{k-1} p^k (1-p)^{t-k} \qquad \forall\, t \geq k$$
$$E[Y_k] = \sum_{i=1} kE[T_i]$$
$$= \frac{k}{p}$$
$$Var(Y_k) = \sum_{i=1}^{k} Var(T_i)$$
$$= \frac{k(1-p)}{p^2}$$

13 Exponential Distribution

Exponential distribution is characterized by the parameter λ and has the following probability distribution

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \lambda e^{-\lambda x} & \text{otherwise} \end{cases}$$

Exponential distribution is used to represent the interarrival time probability distribution in the context of Poisson Process. The cumulative distribution is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{otherwise} \end{cases}$$

$$\begin{aligned} P(X > x) &= \int_x^{\inf} \lambda e^{-\lambda x} dx \\ &= e^{-\lambda x} \end{aligned}$$

13.1 Mean and Variance

The mean of the distribution is given by

$$\begin{aligned} E[x] &= \int_0^{\inf} \lambda x e^{-\lambda x} dx \\ &= [-xe^{-\lambda x}]_0^{\inf} + \int_0^{\inf} e^{-\lambda x} dx = \frac{1}{\lambda} \end{aligned}$$

$$\boxed{E[X] = \frac{1}{\lambda}}$$

where we used integration by parts, $\int uv' = uv - \int u'v$ and substituted $u = x$ and $v = -e^{-\lambda x}/\lambda$.

For variance, we first calculate the value of $E[x^2]$

$$\begin{aligned} E[x^2] &= \int_0^{\inf} \lambda x^2 e^{-\lambda x} dx \\ &= [-x^2 e^{-\lambda x}]_0^{\inf} + \int_0^{\inf} 2x e^{-\lambda x} dx \\ &= [\frac{-2xe^{-\lambda x}}{\lambda}]_0^{\inf} - [\frac{2e^{-\lambda x}}{\lambda^2}]_0^{\inf} \\ &= \frac{2}{\lambda^2} \\ Var(X) &= E[X^2] - E[X]^2 \end{aligned}$$

$$\boxed{Var(X) = \frac{1}{\lambda^2}}$$

The above property can be generalized for the nth power as well

$$E[X^n] = \frac{n!}{\lambda^n}$$

13.2 Memoryless Property

A fundamental mathematical property of the exponential distribution is the memoryless property. In summary, this means that whatever has transpired till now will not affect the future distribution. Mathematically $P(T > t+s) is independent of t$

$$\begin{aligned} P(T > t+s | T > t) &= \frac{P(T > t+s \text{ and } T > t)}{P(T > t)} \\ &= \frac{P(T > t+s)}{P(T > t)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} \\ &= e^{-\lambda s} \end{aligned}$$

$$\boxed{P(T > t+s | T > t) = P(T > s)}$$

14 Poisson Process

14.1 Poisson Random Variable

A random variable X is said to be $Poisson(\mu)$ if it has the following probability distribution

$$p_X(x = k) = \begin{cases} e^{-\mu}\frac{\mu^k}{k!} & \text{for all } x = \{0,\ 1,\ 2, \cdots\} \\ 0 & \text{otherwise} \end{cases}$$

The sum of n independent Poisson variables is also Poisson

$$X_1 + X_2 + \cdots + X_n \sim Poisson(\mu_1 + \mu_2 + \cdots + \mu_n)$$

14.2 Mean and Variance

Expected value is calculated as follows

$$\begin{aligned} E[X] &= \sum_{k=0}^{\inf} k e^{-\mu}\frac{\mu^k}{k!} = \mu e^{-\mu}\sum_{k=1}^{\inf}\frac{\mu^{k-1}}{(k-1)!} \\ &= \mu e^{-\mu}\sum_{k=0}^{\inf}\frac{\mu^k}{k!} \end{aligned}$$

$$\boxed{E[X] = \mu}$$

Variance can be calculated using $Var(X) = E[X^2] - E[X]^2$

$$\begin{aligned} E[X^2] &= \sum_{k=0}^{\inf} k^2 e^{-\mu}\frac{\mu^k}{k!} = \mu e^{-\mu}\sum_{k=1}^{\inf} k\frac{\mu^{k-1}}{(k-1)!} \\ &= \mu e^{-\mu}\sum_{k=0}^{\inf}(k+1)\frac{\mu^k}{k!} \\ &= \mu e^{-\mu}\Big(\mu\sum_{k=1}^{\inf}\frac{\mu^{k-1}}{(k-1)!} + \sum_{k=0}^{\inf}\frac{\mu^k}{k!}\Big) \\ &= \mu e^{-\mu}(\mu e^{\mu} + e^{\mu}) \\ Var(X) &= E[X^2] - E[X]^2 \end{aligned}$$

$$\boxed{Var(X) = \mu}$$

Thus, mean and variance is the same for a Poisson variable.

14.3 Poisson Process

Poisson process also falls in the realm of random processes but is different from Bernoulli process as it is a continuous time process. This process is very commonly used to model arrival times and number of arrivals in a given time interval.

$$P(k,\tau) = \text{Probability of } k \text{ arrivals in interval of duration } \tau$$

$$\sum_k P(k,\tau) = 1 \qquad \text{for a given } \tau$$

Assumptions

- The Probability is dependent only on τ and not the *location* of the interval
- Number of arrivals in disjoint time intervals are *independent*

14.4 A Special Counting Process

A counting process $N_t : t \in [0, \inf)$ is a Poisson process with rate λ if

1. $N_0 = 0$
2. N_t is composed of independent and stationary increments
3. The number of arrivals in any time interval $\tau > 0$ has $Possion(\lambda\tau)$ distribution

Hence, for a Poisson process, the number of arrivals in any interval is dependent only on the length of that interval and not the location. Further, the number of arrivals in the interval will follow a Poisson distribution.

14.5 Derivation from Bernoulli Process

For a very small interval δ,

$$P(k,\delta) = \begin{cases} 1-\lambda\delta & k = 0 \\ \lambda\delta & k = 1 + O(\delta^2) \\ 0 & k > 2 \end{cases}$$

$$\lambda = \lim_{\delta\to 0} \frac{P(1,\delta)}{\delta} \qquad \text{arrival rate per unit time}$$

$$\begin{aligned} E[k] &= (\lambda\delta) * 1 + (1-\lambda\delta) * 0 \\ &= \lambda\delta \\ \tau &= n\delta \end{aligned}$$

The last equation clearly implies that we can approximate the whole process as a bernoulli process where we have n miniscule time intervals with at most one arrival per interval.

$$\begin{aligned} P(k\ arrivals) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \binom{n}{k} (\frac{\lambda\delta}{n})^k (1-\frac{\lambda\delta}{n})^{n-k} \\ \lambda\tau &= np \qquad \text{or, arrival rate * time = E[arrivals]} \\ Poisson &= \lim_{\delta\to 0, n\to \text{inf}} Bernoulli \\ or,\ P(k,\tau) &= \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!} \qquad k = 0, 1, \cdots, \text{ for a given } \tau \\ where,\ \sum_k P(k,\tau) &= 1 \qquad \text{for a given } \tau \end{aligned}$$

Let N_t denote the no of arrivals till time t, then

$$E[N_t] = \lambda t$$
$$Var(N_t) = \lambda t$$

14.6 Time till kth arrival

Suppose the k^{th} arrival happens at a time t. Then we are saying that there have been $k-1$ arrivals till time t and the k^{th} arrival happens at time t (precisely in an interval of $[t, t+\delta]$). Let Y_k be the required time,

$$\begin{aligned} f_{Y_k}(t)\delta &= P(t \leq Y_k \leq t+\delta) \\ &= P(k-1 \text{ arrivals in } [0,t])(\lambda\delta) \\ &= \frac{(\lambda t)^{k-1}}{(k-1)!}e^{-\lambda t}(\lambda\delta) \\ f_{Y_k}(t) &= \frac{\lambda^k t^{k-1}}{(k-1)!}e^{-\lambda t} \end{aligned}$$ Erlang Distribution

14.7 Time of 1st Arrival

Using the Erlang Distribution described above, we have

$$f_{Y_1}(t) = \lambda e^{-\lambda t}$$

$Y_k = T_1 + T_2 + \cdots + T_k$ where all T_i are independent and exponential distributions.

14.8 Renewal Process

Poisson process can be seen as a special case of a renewal process, when the interarrival times are all exponentially distributed.

$$\begin{aligned} \text{Interarrival time } \xi_i &= \lambda e^{-\lambda t} \\ \text{Number of arrivals } P(N_t = n) &= \frac{(\lambda t)^n}{n!}e^{-\lambda t} \\ \text{Time till } n\text{th arrival } P(S_n = t) &= \lambda\frac{(\lambda t)^{n-1}}{(n-1)!}e^{-\lambda x} \text{ for } t > 0 \\ \text{Cumulative distribution } P(S_n \leq t) &= \begin{cases} 1 - e^{-\lambda t}\sum_{k=1}^{n-1}\frac{(\lambda t)^k}{k!} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

14.9 Merging of Poisson Processes

Merging of two Poisson processes is also a Poisson process. Consider two flasbulbs of Red and Green colours, flashing as Possion processes with rates λ_1 and λ_2. Then the process denoting the combined flashing of the two bulbs is also Poisson.
Consider a very small interval of time δ. In this small interval, any of the individual bulbs can have at most one flashes (since we ignore higher order terms). Thus, the following four possibilities arise

Thus, the combined process is Poisson with parameter $\lambda_1 + \lambda_2$

$$P(\text{arrival happened from first process}) = \frac{\lambda_1\delta}{\lambda_1\delta + \lambda_2\delta} = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

0	Red	$\overline{Red}$
$Green$	$\lambda_1\delta\lambda_2\delta$	$(1-\lambda_1\delta)\lambda_2\delta$
$\overline{Green}$	$\lambda_1\delta(1-\lambda_2\delta)$	$(1-\lambda_1\delta)(1-\lambda_2\delta)$

Table 1: Base Probabilities for flashes

0	Red	$\overline{Red}$
$Green$	0	$\lambda_2\delta$
$\overline{Green}$	$\lambda_1\delta$	$(1-(\lambda_1+\lambda_2)\delta)$

Table 2: Probabilities after ignoring δ^2 terms

14.10 Splitting of Poisson Process

Suppose we have a Poisson process with parameter λ which we split into two processes up and down, with probabilities p and $1-p$. The two resulting processes are also Poisson with different parameters.
Consider a small time slot of length δ. Then,

$$\begin{aligned} P(\text{arrival in this time slot}) &= \lambda\delta \\ P(\text{arrival in up slot}) &= \lambda\delta p \\ P(\text{arrival in down slot}) &= \lambda\delta(1-p) \end{aligned}$$

Thus, up and down are themselves Poisson with parameters λp and $\lambda(1-p)$ respectively.

14.11 Random Indcidence for Poisson

Suppose we have a Poisson process with parameter λ running forever. We show up at a random time instant. What is the length of the chosen interarrival time (the total of the time from the last arrival to the next arrival).
Let $T_1^{'}$ denote the time that has elapsed since the last arrival and T_1 be the time till the next arrival. Note that the reverse process is also Poisson with the same parameter. Thus,

$$E[\text{interarrival time}] = E[T_1^{'} + T_1] = \frac{1}{\lambda} + \frac{1}{\lambda} = \frac{2}{\lambda}$$

This may seem paradoxical since the time difference between any two arrivals in a Poisson process is same and it's expected length is $\frac{1}{\lambda}$, whereas we got an interval twice this length. The paradox is resolved by considering the fact that when we choose a random point in time, it is more likely to fall in an interval of larger size than the smaller ones (since probability will be proportional to the length of the interval).

Consider a separate example where we want to compare two values $E[\text{size of a family}]$ and $E[\text{size of a family of a given person}]$.
The two value will be different due to the underlying nature of the way experiment is conducted. For the first, we randomly choose families and average their sizes. Here, family of any size is equally likely to be picked. In the second case, we first pick a person from the population, get their family size, and then average the sizes of the families. Note that, this experiment is biased since the we are more likely to select people from larger families (or equivalently, it is more likely that we pick a person from a large family since the probability of picking is proportional to the family size). Hence, the second value will likely be larger and the two quantities are not equal.

14.12 Non Homogenous Poisson Process

Sometimes, it may not be accurate to use a simple Poisson process to model arrival. For example, a restaurant will not have the same rate of influx throughout the day. This rate itself is a function of time. In such cases, we model the arrivals as Non Homogenous Poisson Process.

For such a process, we have $\lambda(t) : [0, \inf) \to [0, \inf)$ and the counting process N_t is non homogenous if the following hold

1. $N_0 = 0$
2. The increments to N_t are **independent but not stationary**
3. For any small time interval δ, the probability of more than 1 arrival in the interval is zero

The distribution of arrivals in a time interval is still Poisson, but the Poisson parameter is now dependent on the location of the interval itself (since the process does not have stationary increments)

$$N_{t+s} - N_t \sim Poisson(\int_t^{t+s} \lambda(\alpha) d\alpha)$$

15 Gamma Distribution

A random variable is said to have a Gamma distribution if for parameters $\lambda > 0, \alpha > 0$ it has the following probability distribution

$$p_X(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The denominator in the above fomula acts as nothing but a normalization constant and is defined as

$$\begin{aligned} \Gamma(\alpha) &= \int_0^{\inf} \lambda e^{-\lambda x}(\lambda x)^{\alpha-1} \\ &= \int_0^{\inf} e^{-y} y^{\alpha-1} dy \qquad \text{by letting } \lambda x = y \\ &= (\alpha - 1) \int_0^{\inf} e^{-y} y^{\alpha-2} dy \text{ using integration by parts} \\ &= (\alpha - 1)\Gamma(\alpha - 1) \end{aligned}$$

Note that at $\alpha = 1$, $\Gamma(1) = \int_0^{\inf} \lambda e^{-\lambda x} = 1$. Hence, if α is an integer, $\Gamma(\alpha) = \alpha!$ using the recursion relation derived above.

For a fixed λ, as the value of α becomes large, the distribution takes the form of a normal distribution.

15.1 Mean and Variance

Mean and variance are easily obtainable for this using the moment generating function. Recall

$$\phi(t) = E[e^{tX}]$$
$$\phi^n(t) = E[X^n]$$

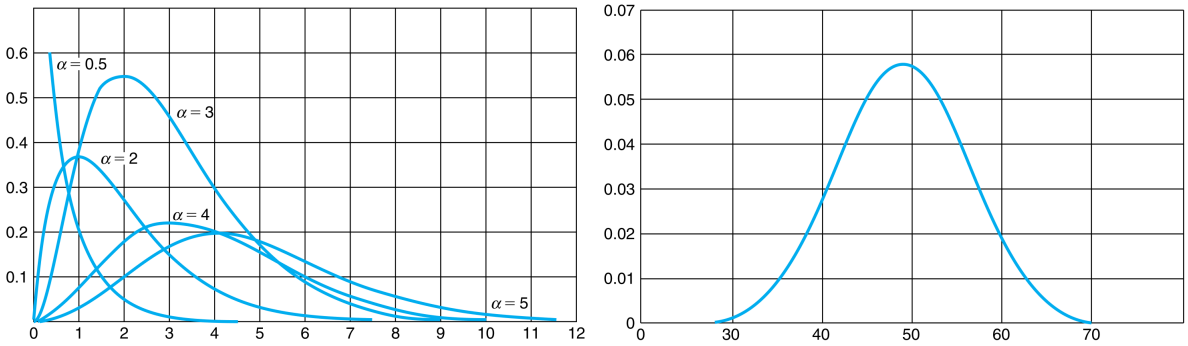


Figure 1: Gamma distribution for $\lambda = 1$ and different values of α. The bottom figure shows the distribution for $\alpha = 50$.

For the current distribution,

$$\begin{aligned}\phi(t) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^{\inf} e^{tx} e^{-\lambda x} x^{\alpha-1} \\ &= (\frac{\lambda}{\lambda - t})^\alpha\end{aligned}$$

Differentiating,

$$\begin{aligned}\phi'(t) &= \frac{\alpha\lambda^\alpha}{(\lambda-t)^{\alpha+1}} \\ \phi''(t) &= \frac{\alpha(\alpha+1)\lambda^\alpha}{(\lambda-t)^{\alpha+2}} E[X] \qquad\qquad = \phi'(0) \\ \boxed{E[X] = \frac{\alpha}{\lambda}} \\ Var(X) &= \phi''(0) \\ \boxed{Var(X) = \frac{\alpha}{\lambda^2}}\end{aligned}$$

15.2 Sum of Gamma Distributions

Let $X_1, X_2, \ldots, X_n$ be n random variables that are gamma distributed with parameters $(\alpha_1, \lambda), (\alpha_2, \lambda), \ldots, (\alpha_n, \lambda)$. Then the distribution of the sum of these random variables is itself a gamma distribution with the parameters $\alpha' = \sum_{i=1}^n \alpha_i$ and $\lambda' = \lambda$

16 Chi-Square Distribution

If $Z_1, Z_2, \ldots, Z_n$ are n independent standard normal variables, then the random variable X

$$\begin{aligned}X &= Z_1^2 + Z_2^2 + \cdots + Z_n^2 \\ \text{then, } X &\sim \chi_n^2\end{aligned}$$

i.e., X follows the chi-square distribution with n degrees of freedom.

If we add two chi-square distributed variables with degrees of freedom n_1 and n_2, then the resultant variable itself is chi-square distributed with $n_1 + n_2$ degrees of freedom. This simply follows from the fact that the sum of the two random variables is nothing but sum of $n_1 + n_2$

standard normal squared variables which is nothing but a chi-square variable with $n_1 + n_2$ degrees of freedom.

If $X \sim \chi^2_n$, then $\chi^2_{\alpha,n}$ is

$$P(X \geq \chi^2_{\alpha,n}) = \alpha$$

This quantity is usually listed in mathematical tables since they are heavily used in hypothesis testing.

16.1 Relation between Chi-Square and Gamma Distribution

Consider the moment generating function for a chi-square random variable with $n = 1$ degrees of freedom

$$\begin{aligned}
E[e^{tX}] &= E[e^{tZ^2}] \;\; Z \sim \mathcal{N}(0,1) \\
&= \int_{-\inf}^{\inf} e^{tx^2} f_Z(x)dx \;\; \text{since } E[g(x)] = \sum_x g(x)p(x) \\
&= \frac{1}{\sqrt{2\pi}} \sum_{-\inf}^{\inf} e^{tx^2} e^{-x^2/2} \\
&= \frac{1}{\sqrt{2\pi}} \sum_{-\inf}^{\inf} e^{-x^2(1/2-t)}
\end{aligned}$$

Using $\int_{-\inf}^{\inf} e^{-a(x+b)^2} = \sqrt{\frac{\pi}{a}}$

$$\begin{aligned}
E[e^{tX}] &= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\pi}{1/2-t}} \\
&= \frac{1}{\sqrt{1-2t}}
\end{aligned}$$

Extending this idea to the case of n degrees of freedom,

$$\begin{aligned}
E[e^{tX}] &= E[e^{t(Z_1^2+Z_2^2+\cdots+Z_n^2)}] \\
&= E[\prod_{i=1}^{n} e^{tZ_i^2}] \\
&= \prod_{i=1}^{n} E[e^{tZ_i^2}] \;\; \text{since } Z_i \text{ are independent} \\
&= (1-2t)^{-n/2} \;\; \text{from the derivation above}
\end{aligned}$$

But, the quantity just derived is nothing but the moment generating function of the Gamma distribution with parameters $(n/2, 1/2)$. Hence, by the uniqueness of the moment generating function, we are forced to conclude that the **probability density function of a chi-square variable with n degrees is same as that of a Gamma distribution with parameters (n/2, 1/2)**.
Thus,

$$f_X(x) = \frac{\frac{1}{2}e^{-x/2}(\frac{x}{2})^{(n/2)-1}}{\Gamma(\frac{n}{2})} \;\; x > 0$$

16.2 Mean and Variance

Since the distribution of a chi-square variable is identical to a Gamma distribution,

$$\boxed{E[X] = n}$$
$$\boxed{Var(x) = 2n}$$

17 t-Distribution

Let Z be a standard normal random variable and let χ^2_n be a chi-square random variable. Assuming these two random variables are independent, the radom variable T_n is

$$T_n = \frac{Z}{\sqrt{\chi^2_n/n}}$$

is said to have a t-distribution with n degrees of freedom.
This distribution is symmetric around the normal, and as n increases, the distribution becomes more and more like the standard normal distribution.

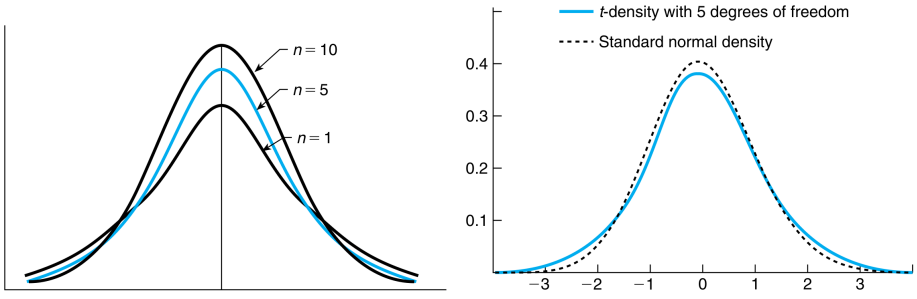


Figure 2: t-distribution for different degrees of freedom, and comparison with standard normal

From figure 2, we see that t-distribution is heavier tailed than a standard normal. Translation, this means that a larger value is more likely to occur under a t-distribution than a standard normal. Furthermore, the heavy tails imply more variance than the standard normal.

For $\alpha between 0$ and 1, let $t_{\alpha,n}$ be such that

$$P(T_n \geq t_{\alpha,n}) = \alpha$$

By symmetry around the origin,

$$\begin{aligned} P(T_n \leq -t_{\alpha,n}) &= \alpha \\ \text{or, } P(T_n \geq -t_{\alpha,n}) &= 1 - \alpha \\ \text{and, } -t_{\alpha,n} &= t_{1-\alpha,n} \end{aligned}$$

These standard values are available in math charts since they form the basis of the t test.

17.1 Mean and Variance

The following are stated without proof

$$\boxed{E[T_n] = 0}$$
$$\boxed{Var(T_n) = \frac{n}{n-2}}$$

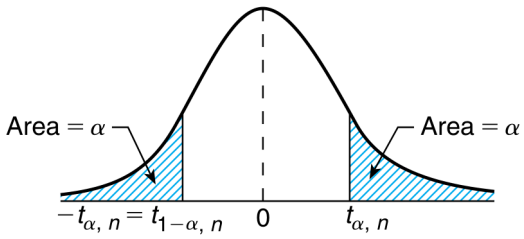


Figure 3: visual representation of $t_{\alpha,n}$

In the limit of large n, the variance is close to 1, which is consistent with the fact that the distribution resembles a standard normal in that limit.

18 F-Distribution

If χ^2_n and χ^2_m are two independent chi-squared distributions with n and m degrees of freedom respectively, then the variable $F_{n,m}$ defined as

$$F_{n,m} = \frac{\chi^2_n/n}{\chi^2_m/m}$$

is said to have an F-distribution with n and m degrees of freedom.
For any α between 0 and 1, we define $F_{\alpha,n,m}$ as

$$P(F_{n,m} > F_{\alpha,n,m}) = \alpha$$

These values are available in standard tables for different combinations of α, n and m.

Consider

$$\begin{aligned}
\alpha &= P(\frac{\chi^2_n/n}{\chi^2_m/m} > F_{\alpha,n,m}) \\
&= P(\frac{\chi^2_m/m}{\chi^2_n/n} < \frac{1}{F_{\alpha,n,m}}) \\
&= 1 - P(\frac{\chi^2_m/m}{\chi^2_n/n} \geq \frac{1}{F_{\alpha,n,m}}) \\
\text{or, } P(\frac{\chi^2_m/m}{\chi^2_n/n} \geq \frac{1}{F_{\alpha,n,m}}) &= 1 - \alpha \\
\text{or, } P(F_{m,n} \geq F_{1-\alpha,m,n}) &= 1 - \alpha \\
\text{Hence, } \frac{1}{F_{\alpha,n,m}} &= F_{1-\alpha,m,n}
\end{aligned}$$

19 Logistics Distribution

A random variable X is said to have a logistics distribution with parameters μ and v if its cumulative density function is of the form

$$F_X(x) = \frac{e^{(x-\mu)/v}}{1 + e^{(x-\mu)/v}}, \quad \forall -\inf < x < \inf$$

Differentiating to get the density function

$$f_x(x) = \frac{e^{(x-\mu)/v}}{v(1 + e^{(x-\mu)/v})^2}, \quad \forall -\inf < x < \inf$$

19.1 Mean

$$\boxed{E[X] = \mu}$$

$v =$ dispersion parameter

20 Markov Process

Markov Process is a discrete time process that is not memoryless. Here the random variable takes several possible states, and the probability distribution is defined in such a way that P(transition from state 1 to state 2) is dependent on state 1.

Let X_n be the random variable denoting the state after n transitions and X_0 will represent the starting state (which can be given or random). Markov assumption states that *Given the current state, past does not matter*. Armed with these,

$$\begin{aligned}
p_{ij} &= P(\text{next state } j \mid \text{current state } i) \\
p_{ij} &= P(X_{n+1} = j | X_n = i) = P(X_{n+1} = j | X_n = i, X_{n-1}, \ldots, X_0) \\
r_{ij}(n) &= P(X_n = j | X_0 = i) \qquad \text{or, in state } j \text{ after } n \text{ steps} \\
r_{ij}(n) &= \sum_{k=1}^{m} r_{ik}(n-1) p_{kj}
\end{aligned}$$

20.1 Recurring and Transient States

A state i is called *recurrent* if, starting from i, and travelling anywhere, it is always possible to return to i. If a state is not recurrent, it is *transient*. States in a recurrent class are periodic if they can be grouped into $d > 1$ groups so that all transitions from one group lead to the next group.

20.2 Steady State Probabilities

Do $r_{ij}(n)$ converge to some π_j (independent of i) ?
Yes if,

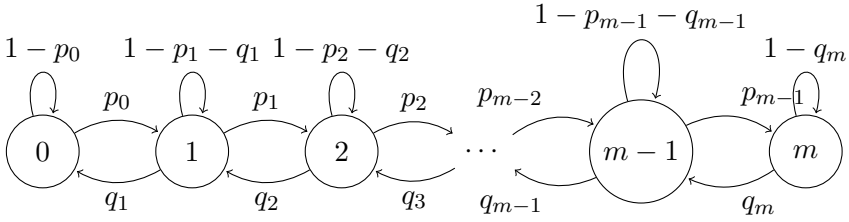
- recurrent states are all in a single class
- single recurrent class is not periodic (otherwise oscillations are possible)

Assuming yes,

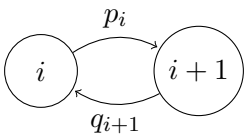
$$\begin{aligned}
r_{ij}(n) &= \sum_k r_{ik}(n-1) p_{kj} \\
\lim_{n \to \text{inf}} r_{ij}(n) &= \sum_k r_{ik}(n-1) p_{kj} \\
\pi_{ij} &= \sum_k \pi_{ik} p_{kj} \qquad \text{balance equations} \\
\sum_i \pi_i &= 1 \\
\text{frequency of transitions } k \to j &= \pi_k p_{kj} \qquad \text{in one step} \\
\text{frequency of transitions into } j &= \sum_k \pi_k p_{kj} \qquad \text{influx from all connected states}
\end{aligned}$$

20.3 Birth Death Process

Consider the checkout counter example. The states are represented by the number of people currently being processed, and we always move n to $[n-1, n, n+1]$, i.e., either the people in the queue decrease by one, remain same or increase by one. Let the probability for moving up be p and moving down be q.



Let's estimate the steady state probabilities. Consider the following diagram splitting the chain into two parts through the two adjacent states



In this case, to maintain steady state, long term frequency of left-right transition should be same as right left transition, i.e., $\pi_i p_i = \pi_{i+1} q_i$
In the special case of $p_i = p$ and $q_i = q \; \forall \; i$,

$$\rho = \frac{p}{q} \qquad \text{load factor}$$

$$\pi_{i+1} = \pi_i \frac{p}{q} = \pi_i \rho$$

$$\pi_i = \pi_0 \rho^i \qquad i = 0, \ldots, m$$

$$\text{Using } \sum_{i=0}^{m} \pi_0 \rho^i = 1,$$

$$\pi_0 = \frac{1}{\sum_{i=0}^{m} \rho^i}$$

$$\text{if } p < q \text{ and } m \to \inf,$$

$$\pi_0 = 1 - \rho$$

$$\pi_i = (1-\rho)\rho^i$$

$$E[X_n] = \frac{\rho}{1-\rho} \qquad \text{Exponential Distribution}$$

When $\rho = 1$ or $p = q$, then all states are equally likely - symmetric random walk.

20.4 Absorption Probabilities

let a_i denote the probability of absorption and μ_i denote the expected no of steps until absorption starting from state i. Then,

$$a_i = \sum_j a_j p_{ij} \qquad \text{outflux to the possible states}$$

$$\mu_i = 1 + \sum_j \mu_j p_{ij}$$

For multipe absorption states, we can possibly consider them together as a group and calculate the relevant quantities.
For a given state s,

$$\begin{aligned} E[\text{steps to first time reach } s \text{ from } i] &= t_i \\ t_i &= E[min\{n \geq 0 \text{ such that } X_n = s\}] \\ t_s &= 0 \\ t_i &= 1 + \sum_j t_j p_{ij} \qquad \text{outflux to all possible states} \end{aligned}$$

Mean recurrence time (mean time to reach back a state) for s

$$\begin{aligned} t_s^* &= E[min\{n \geq 1 \text{ such that } X_n = s\} | X_0 = s] \\ t_s^* &= 1 + \sum_j t_j p_{ij} \end{aligned}$$

21 Central Limit Theorem

21.1 Weak Law of Large Numbers

Suppose we want to know the mean height of penguins in the world. The absolutely correct answer can be obtained by taking the average of the entire population. But ths is not practical, and often we will have to resort to estimating the quantity through a sample. Let there be n penguins in the sample and $X_1, X_2, \ldots, X_n$ be the random variables denoting their heights. Then,

$$\begin{aligned} M_n &= \frac{X_1 + X_2 + \cdots + X_n}{n} \\ \lim_{n \to \text{inf}} E[M_n] &= E[X] = \text{The true mean} \end{aligned}$$

21.2 Markov Inequality/Chebychev Inequality

For nonnegative random variable X,

$$\begin{aligned} E[X] &= \sum_x x p_X(x) \geq \sum_{x \geq a} x p_X(x) && \text{discrete case} \\ &= \int_x x p_X(x) \ \geq \int_{x \geq a} x p_X(x) && \text{continuous case} \end{aligned}$$

Applying the above set of inequalities to the variable $X - \mu$

$$E[(X-\mu)^2] \geq a^2 P((X-\mu)^2 \geq a^2)$$
$$\text{or, } Var(X) \geq a^2 P(\mid X-\mu \mid \geq a)$$

For continuous case,

$$\begin{aligned}\sigma^2 &= \int_{-\inf}^{\inf} (x-\mu)^2 f_X(x)dx \\ &\geq \int_{-\inf}^{\mu-c} (x-\mu)^2 f_X(x)dx + \int_{\mu+c}^{\inf} (x-\mu)^2 f_X(x)dx \\ &\geq c^2 P(\mid X-\mu \mid \geq c)\end{aligned}$$

Hence,

$$P(\mid X-\mu \mid \geq c^2) \leq \frac{\sigma^2}{c^2}$$

or,

$$\boxed{P(\mid X-\mu \mid \geq k\sigma) \leq \frac{1}{k^2}} \quad \text{where } c = k\sigma$$

Going back to the problem of estimating the mean,

$$M_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$
$$E[M_n] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu \quad \text{expectation of expectation}$$
$$Var(M_n) = \sum_{i=1}^{n} Var(\frac{X_i}{n}) = \frac{\sigma^2}{n} \quad \text{since } X_i \text{ are independent}$$

$$\boxed{P(\mid M_n - \mu \mid \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}}$$

or, as $n \to \inf,\ M_n - \mu \to 0$, ϵ is the error bound/confidence.

21.3 Central Limit Theorem

Chebychev's inequality gives a loose bound. We can do better with CLT. Let X be a random variable with mean μ and variance σ^2, and let X_i be independent identically distributed random variables with the same distribution as X. Then,

$$S_n = X_1 + X_2 + \cdots + X_n$$
$$Z_n = \frac{S_n - E[S_n]}{\sigma_n} \quad \text{random variable with mean 0 and variance 1}$$
$$= \frac{S_n - nE[X]}{\sqrt{n}\sigma}$$
$$\text{or, } S_n = \sqrt{n}\sigma Z_n + nE[X]$$
$$\text{In } \lim_{n\to\inf} Z_n \to Z \text{ (standard normal)}$$
$$\text{or, } \boxed{Z = \frac{S_n - nE[X]}{\sqrt{n}\sigma}} \text{ only for CDF (no comment on PDF/PMF)}$$
$$\text{Thus, } \boxed{P(Z > c) = P(\frac{S_n - nE[X]}{\sqrt{n}\sigma} > c)}$$

By defining the confidence on how close we desire S_n to the actual mean, we can calculate the required value of the n using standard normal CDF tables. However, we need to have an estimate of variance of the distribution in order to do the estimate of n.

22 Distribution of Sample Mean and Variance

Let $X_1, X_2, \ldots, X_n$ be independent random variables from a distribution having mean μ and variance σ^2. From the central limit theorem,

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \sim \mathcal{N}(0, 1)$$

or, the sum of the random variables follows the distribution of a standard normal as the value of n becomes large. Typically, the property starts to manifest as soon as n becomes around 30.

22.1 Sample Mean

Let the random variable $\overline{X}$ denote the sample mean and is defined as

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Note that any scaled version of a normal distribution is also normal. Thus, **The mean of n independent random variables coming from the same distribution also follows a normal distribution for sufficiently large n**.

Note that sample mean is itself a random variable and thus has a distribution. This happens because the quantity itself is the average of several random variables, which are instances of the same probability distribution.

$$E[\overline{X}] = E[\frac{X_1 + X_2 + \cdots + X_n}{n}] \tag{1}$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \tag{2}$$

$$\boxed{E[\overline{X}] = \mu} \tag{3}$$

Similiarly, the variance of the sample mean can be computed as follows

$$\begin{aligned} Var(\overline{X}) &= Var(\frac{X_1 + X_2 + \cdots + X_n}{n}) \\ &= Var(\frac{X_1}{n}) + Var(\frac{X_2}{n}) + \cdots + Var(\frac{X_n}{n}) \text{ using independence} \\ &= n\frac{\sigma^2}{n^2} \text{ using } Var(aX) = a^2 Var(X) \end{aligned}$$

$$\boxed{Var(\overline{X}) = \frac{\sigma^2}{n}}$$

Hence, for a population of mean μ and variance σ^2, the E[sample mean] is still μ but the variance of the sample mean shrinks by a factor of n. Stated in a different manner, this means that the spread of the sample mean reduces as we take the mean from more and more observations. This directly translates into the fact that our confidence on the estimate of the sample mean increases with more observations.

22.2 Sample Variance

The sample variance is defined as

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

where $\overline{X}$ is the sample mean. Similar to the sample mean, this is also a random variable. We divide by $n-1$ so that the estimator is unbiased, as shown below by calculating the mean of the estimator

$$\begin{aligned}
E[S^2] &= E[\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}] \\
&= \frac{1}{n-1}\sum_{i=1}^{n} E[(X_i - \overline{X})^2] \\
&= \frac{1}{n-1}\sum_{i=1}^{n} E[X_i^2 - 2X_i\overline{X} + \overline{X}^2] \\
&= \frac{1}{n-1}\big(\sum_{i=1}^{n} E[X_i^2] - E[2\overline{X}\sum_{i=1}^{n} X_i] + \sum_{i=1}^{n} E[\overline{X}^2]\big) \\
&= \frac{1}{n-1}\big(\sum_{i=1}^{n} E[X_i^2] - E[2n\overline{X}^2] + nE[\overline{X}^2]\big) \\
\text{Using } E[X^2] &= Var(X) + E[X]^2, \\
E[S^2] &= \frac{1}{n-1}\big(\sum_{i=1}^{n} E[Var(X_i) + E[X_i]^2] - nE[\overline{X}^2]\big) \\
&= \frac{1}{n-1}(n\sigma^2 + n\mu^2 - n(\frac{\sigma^2}{n} + \mu^2)) \\
&\boxed{E[S^2] = \sigma^2}
\end{aligned}$$

i.e., the mean of the sample variance is same as the variance of the distribution (population variance). Further, the division by $n-1$ to calculate sample variance comes from the fact that we are already using $\overline{X}$ as an estimate for sample mean which takes away one degree of freedom.

22.3 Distributions for a Normal Population

Consider $X_1, X_2, \ldots, X_n$ be independently derived from a normal population with mean μ and variance σ^2
i.e., $X_i \sim \mathcal{N}(\mu, \sigma^2) \forall i = 1, 2, \ldots, n$

Based on the derivations above,

$$E[\overline{X}] = \mu$$
$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

And since the sum of normal random variables is also normal,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

which is similar to the central limit theorem.

From the derivation above for the sample variance,

$$E[S^2] = \sigma^2$$

Now let's calcluate the distribution of S^2

$$\begin{aligned}
S^2 &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \\
(n-1)S^2 &= \sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&= \sum_{i=1}^{n}((X_i - \mu) - (\overline{X} - \mu))^2 \\
&= \sum_{i=1}^{n}((X_i - \mu)^2 + (\overline{X} - \mu)^2 - 2(X_i - \mu)(\overline{X} - \mu)) \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 + n(\overline{X} - \mu)^2 - 2(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 + n(\overline{X} - \mu)^2 - 2n(\overline{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2 \\
\frac{(n-1)S^2}{\sigma^2} &= \sum_{i=1}^{n}(\frac{X_i - \mu}{\sigma})^2 - (\frac{\overline{X} - \mu}{\sigma/\sqrt{n}})^2 \quad \text{to make standard normals} \\
\text{or,} \quad \frac{(n-1)S^2}{\sigma^2} + (\frac{\overline{X} - \mu}{\sigma/\sqrt{n}})^2 &= \sum_{i=1}^{n}(\frac{X_i - \mu}{\sigma})^2
\end{aligned}$$

The right hand side is a chi-square variable with n degrees of freedom and the second part of the right hand side is a chi-square variable with 1 degree of freedom. We know that sum of independent chi-square variables is also a chi-square variable with degrees of freedom equal to the sum of individual degrees of freedom. Hence, it follows that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

and also the fact that **for a normal population, the sample mean and sample variance are independent variables with normal and chi-square distributions respectively**. This independence is a unique property for a normal distribution and is useful in parameter estimation and hypothesis testing.

Another interesting observation from the above derivations is

$$\boxed{\sqrt{n}\frac{\overline{X} - \mu}{S} \sim t_{n-1}}$$

$$\text{whereas } \sqrt{n}\frac{\overline{X} - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

Note that the denominator is in the first equation is sample variance. The derivation is

$$\frac{Z}{\sqrt{\chi^2_n/n}} \sim t_n \text{ definition}$$

$$\text{or, } \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}\frac{1}{n-1}}} \sim t_{n-1}$$

$$\text{or, } \sqrt{n}\frac{\overline{X}-\mu}{S} \sim t_{n-1}$$

23 Parameter Estimation

In probability theory, we usually have the distribution known to us and we try to use this information to obtain theoretical results applicable on population of this distribution. However, in statistics, we are given the samples drawn from the population, and we are interested in estimating the parameters of this population. For instance, it can be known that the samples are drawn from a normal distribution and the objective is to use the samples to obtain the mean and variance of the normal distribution. It is possible to partially know the parameters in some cases, for example mean is unknown but the variance is known. Following are some methods to obtain the *estimates* of these parameters.

23.1 Maximum Likelihood Estimator

Maximum Likelihood Estimator or MLE is based on the idea to **find that value of θ that maximizes the probability of observing the given set of samples of the population**. Alternately, let x_i for $i = 1, 2, \ldots, n$ be n samples drawn from a population whose distribution is parametrized by θ (can be a vector as well). Then we define the likelihood function as

$$\text{likelihood } = f(x_1, x_2, \ldots, x_n|\theta)$$

i.e., the joint probability (or density) of occurrence of all the samples under the given distribution for some value of θ. We aim to maximize this likelihood to get the estimate of θ. It is often the case that taking logarithm of both sides allows for an easy way of estimation. Note that both likelihood an log likelihood are maximized by the same value of estimator.

23.1.1 MLE for Bernoulli Variable

Suppose we observe n independent samples from a Bernoulli process and the aim is to find the MLE for the probability of success.
Let p denote the probability of success. Then,

$$P(X_i = x) = p^x(1-p)^{1-x} \text{ where } x \text{ is 0 or 1}$$

$$\text{or, } P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$$

Since all the samples are independent, the joint probability or likelihood is simply the product of all the probabilities

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|p) &= p^{x_1}(1-p)^{1-x_1} \ldots p^{x_n}(1-p)^{1-x_n} \\ &= p^{\sum_{i=1}^{x_i}}(1-p)^{n-\sum_{i=1}^{x_i}} \\ log(f(x_1, x_2, \ldots, x_n|p)) &= \sum_{i=1}^{x_i} log(p) + (n - \sum_{i=1}^{x_i}) log(1-p) \end{aligned}$$

Taking the derivative with respect to p to maximize,

$$\frac{d}{dp}log(f(x_1, x_2, \ldots, x_n|p)) = \sum_{i=1}^{x_i}\frac{1}{p} - (n - \sum_{i=1}^{x_i})\frac{1}{1-p} = 0$$

$$\text{or, } \boxed{\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

which is the proportion of successful trials in the sample.

23.1.2 MLE for Poisson Variable

Suppose that we observe n random samples obtained from a poisson process. Recall

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Hence, the joint distribution can be written as

$$\begin{aligned}
f(x_1, x_2, \ldots, x_n|p) &= \prod_{i=1}^{n}\frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
f(x_1, x_2, \ldots, x_n|p) &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \\
log(f(x_1, x_2, \ldots, x_n|p)) &= -n\lambda + (\sum_{i=1}^{n} x_i)log(\lambda) - \sum_{i=1}^{n} log(x_i!) \\
\frac{d}{dp}(f(x_1, x_2, \ldots, x_n|p)) &= -n + (\sum_{i=1}^{n} x_i)\frac{1}{\lambda} \\
&= 0 \\
\text{or, } \boxed{\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}}
\end{aligned}$$

i.e., the average rate is simply the average of all the observed arrivals.

23.1.3 MLE for Normal Variable

Suppose we observe n samples from a normal population whose mean is μ and variance is σ^2. We will aim to obtain MLE estimates for both the mean and variance. Recall

$$P(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}}exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

Hence, the likelihood will be

$$f(x_1, x_2, \ldots, x_n|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

$$log(f(x_1, x_2, \ldots, x_n|\mu, \sigma^2)) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}(\sum_{i=1}^{n}(x_i-\mu)^2)$$

$$\frac{d}{d\mu}(log(f(x_1, x_2, \ldots, x_n|\mu, \sigma^2))) = -\frac{1}{\sigma^2}(\sum_{i=1}^{n}(x_i-\mu))$$

$$= 0$$

$$\text{or, } \boxed{\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

$$\frac{d}{d\sigma}(log(f(x_1, x_2, \ldots, x_n|\mu, \sigma^2))) = -\frac{n}{\sigma} - \frac{1}{2\sigma^2}(\sum_{i=1}^{n}(x_i-\mu)^2)$$

$$= 0$$

$$\text{or, } \boxed{\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}$$

$$\text{where } \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Note that the estimator for variance is different from the sample variance

$$\text{MLE } \sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}$$

$$S^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}$$

23.1.4 MLE for Uniform Random Variable

Consider oberving n samples from a uniform distribution with the parameter θ. Then,

$$f(x_1, x_2, \ldots, x_n|\theta) = \frac{1}{\theta^n}$$

which is clearly maximized when θ is minimum. But since θ has to be at least as large as the maximum observed value,

$$\boxed{\hat{\theta} = max(x_1, x_2, \ldots, x_n)}$$

23.2 Interval Estimates

The MLE estimates calculated above are estimates and do not reflect the true value. We expect the true value of the parameter to be close to the estimate, but not exactly equal to it. Hence, it makes sense to give an interval instead of a single estimate to reflect our confidence in the estimated value of the parameter.

Note that the below confidence intervals imply that α percent of times, the constructed interval will contain the true mean μ, when the same calculation is repeated with multiple samples. The calculations of intervals do not imply that μ is contained in the interval with α confidence. We calculate an interval that falls on μ rather than telling the interval that μ falls in.

23.2.1 Confidence interval for Mean of Normal Distribution when Variance is Known

Consider the problem of estimation of the mean of a normal distribution with known variance σ^2. Since we know that the MLE for mean is just the sample mean, and the sample mean follows a normal distribution,

$$P(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96) = 0.95 \text{ using standard normal tables}$$
$$\text{or, } P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

Thus, we are 95% confident that the true value of the mean lies in the range

$$(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

This trick to calculate the confidence interval can be generalized for any value of confidence. Recall that

$$P(Z > z_\alpha) = \alpha$$
$$P(Z < -z_\alpha) = \alpha$$

Hence, for a given confidence level α, the two sided confidence interval is

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$
$$P(-z_{\alpha/2} < \sqrt{n}\frac{\overline{X} - \mu}{\sigma} < z_{\alpha/2}) = 1 - \alpha$$
$$P(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$$\boxed{\text{two sided } 1-\alpha \text{ confidence interval} = (\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})}$$

where $\overline{x}$ is the observed value of $\overline{X}$.

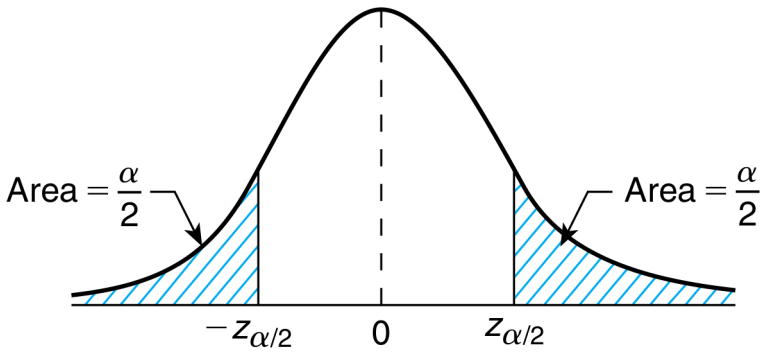


Figure 4: Visualization of the double sided confidence interval on standard normal.

In a very similar manner, we can calculate the one sided confidence interval. Here, we are only

interested in the lower or upper bound of the said interval. The other side is inf or $-\inf$.

$$P(Z > z_\alpha) = \alpha$$
$$P(\sqrt{n}\frac{\overline{X} - \mu}{\sigma} > z_\alpha) = \alpha$$
$$P(\mu < \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}}) = \alpha$$
$$\boxed{\text{Lower } 1-\alpha \text{ confidence interval} = (-\inf, \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}})}$$

$$P(Z < -z_\alpha) = \alpha$$
$$P(\sqrt{n}\frac{\overline{X} - \mu}{\sigma} < -z_\alpha) = \alpha$$
$$P(\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}} < \mu) = \alpha$$
$$\boxed{\text{Upper } 1-\alpha \text{ confidence interval} = (\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \inf)}$$

The interpretation of the right sided confidence interval is that we are $1-\alpha$ confident that the value of the mean is more than the lower end of the interval. In a similar way, the left side interval gives the upper bound on the value of mean with the desired confidence.

23.2.2 Confidence interval for Mean of Normal Distribution when Variance is Unknown

The derivation of confidence intervals in this case is similar to the above, with the only difference of using a t-distribution. Recall

$$\sqrt{n}\frac{\overline{X} - \mu}{S} \sim t_{n-1}$$

where S is the sample variance. Following similar derivation to the known variance case,

$$\text{two sided } 1-\alpha \text{ confidence interval} = (\overline{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}})$$
$$\text{Lower } 1-\alpha \text{ confidence interval} = (-\inf, \overline{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}})$$
$$\text{Upper } 1-\alpha \text{ confidence interval} = (\overline{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}}, \inf)$$

where s is the observed value of the sample variance S. However, notice that the intervals calculated will usually be larger than those when the variance is known because t-distribution is heavier tailed than a standard normal and thus has higher variance.

23.2.3 Confidence interval for Variance of Normal Distribution when Mean is Unknown

Recall that

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

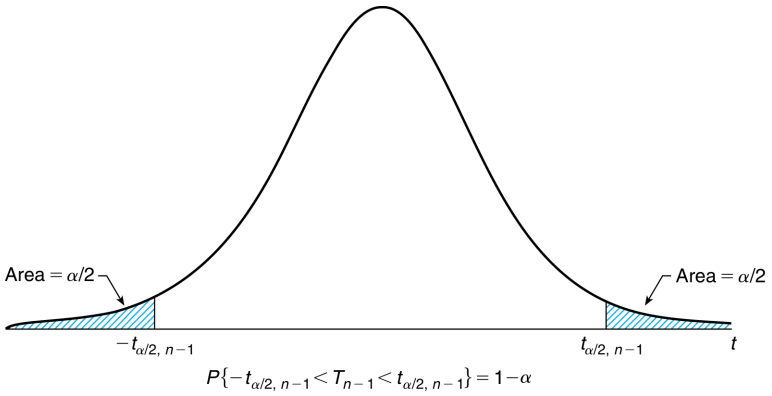


Figure 5: Visualization of the double sided confidence interval on standard normal.

By noting that σ^2 is always positive, we have the following

$$\text{two sided } 1-\alpha \text{ confidence interval} = (\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}})$$
$$\text{Lower } 1-\alpha \text{ confidence interval} = (0, \frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}})$$
$$\text{Upper } 1-\alpha \text{ confidence interval} = (\frac{(n-1)s^2}{\chi^2_{\alpha,n-1}}, \text{inf})$$

23.2.4 Estimating Difference in Means of Two Normal Populations

Suppose the following two independent sets of random variables

$$X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$Y_1, Y_2, \ldots, Y_m \sim \mathcal{N}(\mu_1, \sigma_1^2)$$

Then, we are interseted in the distribution of $\mu_1 - \mu_2$. It is intuitive to see that the MLE estimator of this quantity is nothing but $\overline{X} - \overline{Y}$. Also, since $\overline{X}$ and $\overline{Y}$ are both normally distributed,

$$\overline{X} - \overline{Y} \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m})$$

Consequently, using the confidence intervals derived for the case of a mean of a single normal distribution, we have the following intervals when the standard deviations are known

$$\text{two sided } 1-\alpha \text{ confidence interval} = (\overline{x} - \overline{y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \overline{x} - \overline{y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}})$$
$$\text{Lower } 1-\alpha \text{ confidence interval} = (-\text{inf}, \overline{x} - \overline{y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}})$$
$$\text{Upper } 1-\alpha \text{ confidence interval} = (\overline{x} - \overline{y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \text{inf})$$

where $\overline{x}$ and $\overline{y}$ are estimates of $\overline{X}$ and $\overline{Y}$ respectively.
A more challenging scenario arises when the variances are not known. In that case, it is only

logical to try to estimate the intervals using sample variances (themselves random variables)

$$S_1^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$
$$S_2^2 = \frac{\sum_{i=1}^{m}(Y_i - \overline{Y})^2}{m-1}$$

However, the distribution useful for deriving the confidence intervals

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$$

is complicated and depends on the unknown variances. The variable is friendly if we assume that the two unknown variances are both same.

Assuming $\sigma_1 = \sigma_2 = \sigma$, it follows

$$(n-1)\frac{S_1^2}{\sigma^2} \sim \chi_{n-1}^2$$
$$(m-1)\frac{S_2^2}{\sigma^2} \sim \chi_{m-1}^2$$
$$(n-1)\frac{S_1^2}{\sigma^2} + (m-1)\frac{S_2^2}{\sigma^2} \sim \chi_{n+m-2}^2$$

since S_1^2 and S_2^2 are independent chi-square random variables and from section 16 that the sum of such variables is also chi-square.
We already know

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim \mathcal{N}(0,1)$$

and we also know that the ratio of a standard normal and the square root of a chi-square divided by it's degrees of freedom is a t-distribution

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \div \sqrt{\frac{(n-1)\frac{S_1^2}{\sigma^2} + (m-1)\frac{S_2^2}{\sigma^2}}{n+m-2}} \sim t_{n+m-2}$$

Let

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

Then,

$$P(-t_{\alpha/2,n+m-2} \leq \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \leq t_{\alpha/2,n+m-2}) = 1 - \alpha$$

Two sided $1-\alpha$ confidence interval $= (\overline{x} - \overline{y} - t_{\alpha/2,n+m-2}s_p\sqrt{\frac{1}{n} + \frac{1}{m}}, \overline{x} - \overline{y} + t_{\alpha/2,n+m-2}s_p\sqrt{\frac{1}{n} + \frac{1}{m}})$

where s_p is the sample estimate for S_p. Lower confidence interval is derived in a similar fashion to the previous derivations, but the upper confidence interval is the lower confidence interval of $\mu_2 - \mu_1$.

23.2.5 Confidence Interval for Mean of Bernoulli Random Variable

Suppose we obtain a sample of n independent Bernoulli random variables, where the probability of success is p. Let X denote the no of successes. Using the CLT for large n,

$$\frac{X - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0,1)$$

It is not tractable to calculate the confidence intervals from this formulation. Let $\hat{p} = X/n$ denote the MLE of the mean p. Substituiting in the denominator of above,

$$P(-z_{\alpha/2} \leq \frac{X - np}{\sqrt{n\hat{p}(1-\hat{p})}} \leq z_{\alpha/2}) \approx 1 - \alpha$$

$$\text{Two sided } 1-\alpha \text{ confidence interval} \approx (\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$$

and one sided confidence intervals can be obtained in similar manner to previous derivations.

23.3 Evaluating an Estimator

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be the set of random variables sampled from a population whose parameters are defined by θ. Let $d(\boldsymbol{X})$ denote an estimator of θ. Then

$$r(d, \theta) = E[(d(\boldsymbol{X}) - \theta)^2]$$

denotes the mean squared estimator of $\boldsymbol{X}$. Although it is rare to find an estimator that minimizes this error, we can certainly find the minima under the set of estimators satisfying a certain criteria.
The bias of an estimator is defined as

$$b_\theta(d) = E[d(\boldsymbol{X})] - \theta$$

If the bias is zero, then the estimator is called an **unbiased estimator**. That is, the expected value of the estimator is same as the parameter being estimated.

For an unbiased estimator, the mean square error is

$$\begin{aligned} r(d, \theta) &= E[(d(\boldsymbol{X}) - \theta)^2] \\ &= E[(d(\boldsymbol{X}) - E[d(\boldsymbol{X})])^2] \\ &= Var(d(\boldsymbol{X})) \end{aligned}$$

i.e., the mean squared error of an unbiased estimator is equal to its variance.

Let $X_1, X_2, \ldots, X_n$ be sampled from a distribution whose mean is θ. Then,

$$d(\boldsymbol{X}) = \sum_{i=1}^{n} \lambda_i X_i$$

$$\text{and } \sum_{i=1}^{n} \lambda_i = 1$$

is also an unbiased estimator because

$$\begin{aligned} E[\sum_{i=1}^{n} \lambda_i X_i] &= \sum_{i=1}^{n} \lambda_i E[X_i] \\ &= \theta \sum_{i=1}^{n} \lambda_i \\ &= \theta \end{aligned}$$

23.3.1 Combining Unbiased Estimators

Suppose we have n unbiased estimators $d_1, \ldots, d_n$ for a parameter θ with different independent variances

$$E[d_i] = \theta \quad Var(d_i) = \sigma_i^2$$

Then, a weighted combination of these estimators is also an unbiased estimator of θ (assuming that the weights sum up to 1). Suppose we wish to find a set of weights that minimize the mean squared error to get the best estimator, then

$$d = \sum_{i=1}^{n} w_i d_i \text{ where } \sum_{i=1}^{n} w_i = 1$$

$$r(d, \theta) = Var(d) \text{ (derived above)}$$

$$Var(d) = Var(\sum_{i=1}^{n} w_i d_i)$$

$$= \sum_{i=1}^{n} w_i Var(d_i) \text{ by independence}$$

$$= \sum_{i=1}^{n} w_i \sigma_i^2$$

$$\text{So we minimize } \sum_{i=1}^{n} w_i \sigma_i^2 - \lambda(\sum_{i=1}^{n} w_i - 1)$$

$$\text{Taking the derivative for any } i,\ 0 = 2\sigma_i w_i - \lambda$$

$$\text{Using } \sum_{i=1}^{n} w_i = 1$$

$$w_i = \frac{1/\sigma_i^2}{1/\sigma_1^2 + 1/\sigma_2^2 + \cdots + 1/\sigma_n^2}$$

or, the weights for the estimators are inversely proportional to their individual variances. This is useful in situations when we have n independent results for evaluation of a parameter, and we want to increase our confidence in the estimator by combining all these independent results.

23.3.2 Relation between Bias and Variance

The result obtained above that the mean squared error of an unbiased estimator is it's variance can be generalized for the case of any estimator as follows

$$\begin{aligned} r(d, \theta) &= E[(d - \theta)^2] \\ &= E[(d - E[d] + E[d] - \theta)^2] \\ &= E[(d - E[d])^2 + (E[d] - \theta)^2 + 2(d - E[d])(E[d] - \theta)] \\ &= E[(d - E[d])^2] + E[(E[d] - \theta)^2] + 2(E[d] - \theta)E[d - E[d]] \\ &= E[(d - E[d])^2] + (E[d] - \theta)^2 \end{aligned}$$

$$\boxed{r(d, \theta) = Var(d) + b_\theta^2(d)}$$

where we have noted that $E[d - E[d]] = E[d] - E[d] = 0$ and $E[d] - \theta$ is a constant since $E[d]$ itself is a constant.

23.4 Bayes Estimator

The Bayes estimator is the expected value of the parameter given the data. It utilizes the Bayes theorem in order to arrive at the estimator value

$$E[\theta|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = \int \theta f(\theta|x_1, x_2, \ldots, x_n)$$
$$f(\theta|x_1, x_2, \ldots, x_n) = \frac{p(\theta)f(x_1, x_2, \ldots, x_n|\theta)}{\int p(\theta)f(x_1, x_2, \ldots, x_n|\theta)d\theta}$$

where $p(\theta)$ is the assumed prior distribution on the parameter θ. Based on our knowledge of the process, this can be uniform, normal etc.

24 Hypothesis Testing

Hypothesis testing is a set of procedures used to test a hypothesis about statistical parameters. Based on the results of the procedure, we decide whether to accept or reject the hypothesis. This can be as simple as deciding whether the mean of a population is more than 1 or not.

Whenever a hypothesis is accepted, it does not mean that the hypothesis is true, but rather that the data is consistent with it.

Suppose we have a population that is characterized by the distribution F_θ and we are interested in making statistical comments about the value of the parameters θ. **The hypothesis that specifies the statement that we are going to test about the parameter is called the null hypothesis** and is denoted by H_0. Note that the statement of the null hypothesis can either comment on the exact value of the parameter, or comment on an inequality satisfied by the parameter. When the hypothesis completely specifies the distribution, it is called a *simple hypothesis* and in the other case, it is called *composite hypothesis*.

To test the hypothesis, suppose we have available with us n independent samples from the population. Based on these samples, we must define a n dimensional region C such that if the sample falls in this region, we reject the null hypothesis and vice versa. This region C is called the critical region.

$$\text{Reject } H_0 \text{ if } (X_1, X_2, \ldots, X_n) \in C$$
$$\text{Acecpt } H_0 \text{ if } (X_1, X_2, \ldots, X_n) \notin C$$

Two types of errors can result when checking a hypothesis

- *type I error*: Reject H_0 when it is correct
- *type II error*: Accept H_0 when it is incorrect

Since hypothesis testing is about checking if the given data is consistent with the hypothesis, the error we make on rejecting the hypothesis when it is correct should be low. This is consistent with the confidence intervals discussed earlier. We denote *type I error* by α meaning that there is only α chance that the hypothesis will be incorrectly rejected by the test, and is called the level of significance of the test.

A lower significance level or lower α implies that we require stronger evidence against the null hypothesis to reject it.
A classical approach while testing the parameters of a population will be to first determine a point estimator of the parameter and then determine the distribution of the said estimator.

Hypothesis test will usually involve checking whether the estimator lies in a selcted region, for which we can determine the relevant confidence intervals through the distribution of the estimator.

24.1 Test around Mean of Normal Population

24.1.1 Known Variance

Consider n samples $X_1, X_2, \ldots, X_n$ drawn from a normal distribution with unknown mean μ and known variance σ^2. We have the following hypothesis

$$\begin{aligned} \text{null hypothesis} \quad & H_0 : \mu = \mu_0 \\ \text{alternate hypothesis} \quad & H_1 : \mu \neq \mu_0 \end{aligned}$$

The sample mean $\overline{X}$ is clearly a natural choice for the estimator of the mean. It is intuitive to define the critical region in such a manner that we reject H_0 if the estimator is far off from μ_0 and vice versa

$$C = \{X_1, X_2, \ldots X_n : |\overline{X} - \mu_0| > c\}$$

for some suitably chosen c. We also know that the mean of a normal population has a normal distribution. Hence for some significance level α (*type I error*),

$$P_{\mu_0}(|\overline{X} - \mu_0| > c) = \alpha$$

where the subscript denotes the fact that the probability is being calculated under the assumption of $\mu = \mu_0$. Under this assumption, $\overline{X}$ is normally distributed with mean mu_0.

$$\begin{aligned} P_{\mu_0}(|\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}| > \frac{c\sqrt{n}}{\sigma}) = \alpha \\ 2P_{\mu_0}(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{c\sqrt{n}}{\sigma}) = \alpha \end{aligned}$$

but we know that these are tabulated values

$$\begin{aligned} P(Z > z_{\alpha/2}) &= \alpha/2 \\ \text{or,} \quad \frac{c\sqrt{n}}{\sigma} &= z_{\alpha/2} \\ \text{or,} \quad c &= \frac{\sigma z_{\alpha/2}}{\sqrt{n}} \end{aligned}$$

or simply put in terms of the hypothesis test,

$$\begin{aligned} \text{Reject} \quad H_0 \quad \text{if} \quad & \left|\frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0)\right| > z_{\alpha/2} \\ \text{Accept} \quad H_0 \quad \text{if} \quad & \left|\frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0)\right| \leq z_{\alpha/2} \end{aligned}$$

where α is the *type I error* and should ideally be low.

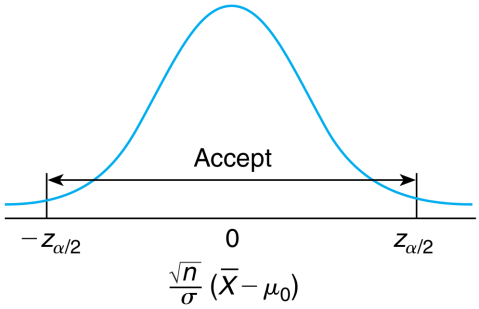


Figure 6: Acceptance Region for Hypothesis $\mu = \mu_0$

24.1.2 p-value

We can also determine an inequality on the significance leve, meaning that α above a certain threshold, we will always reject the null hypothesis and vice versa.
This threshold is calculated via the probability that the estimator distribution is as large as the test statistic

$$P(|Z| > \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0)) = p\text{-}value$$

This *p-value* is interpreted as the threshold for the significance level, i.e., at any significance level less than this value we will accept the null hypothesis and vice versa. This is because the closer the statistic is to zero, the higher the chance that the mean is actually equal to μ_0. Thus, a very small *p-value* implies that the statistic is far away from the mean and we can reject H_0.

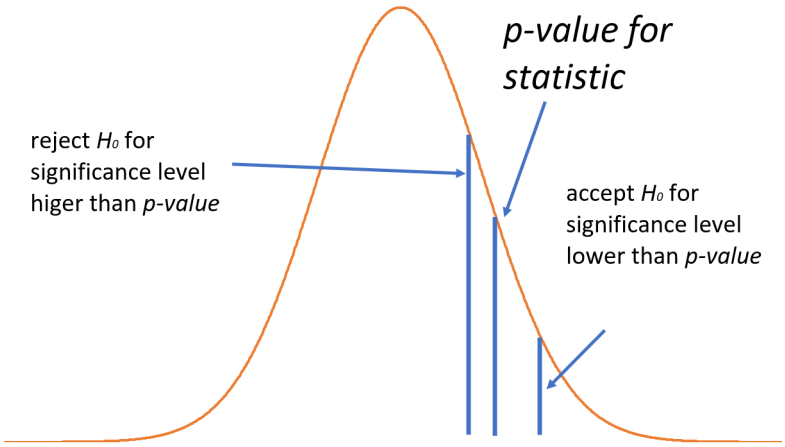


Figure 7: *p-value* and significance levels on a standard normal distribution

Note that *p-value* is not particular to the normal distribution. We can define it in the context of a t-distribution as well. *p-value* represents the probability value associated with the test statistic under the distribution the test statistic follows.

Furthermore, if we have a predefined value of the significance level, any *p-value* lower than this level implies it is very likely for the mean to be different, calling for rejecting H_0. This is visually represented in figure 7.

We have not yet commented on the *type II error*. Consider $\beta(\mu)$ as probability of accepting H_0

when the mean is μ

$$\begin{aligned}\beta(\mu) &= P_\mu(\text{accepting } H_0 \text{ when the mean is } \mu)\\ &= P(\text{statistic is } \ \leq z_{\alpha/2})\\ &= P(|\frac{\sqrt{n}}{\sigma}(\overline{X}-\mu_0)| \leq z_{\alpha/2})\\ &= P(-z_{\alpha/2} \leq \frac{\sqrt{n}}{\sigma}(\overline{X}-\mu_0) \leq z_{\alpha/2})\end{aligned}$$

But, under the premise that μ is the correct mean,

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

Thus,

$$\begin{aligned}\beta(\mu) &= P(-z_{\alpha/2} - \frac{\sqrt{n}}{\sigma}\mu \leq \frac{\sqrt{n}}{\sigma}(\overline{X}-\mu_0) - \frac{\sqrt{n}}{\sigma}\mu \leq z_{\alpha/2} - \frac{\sqrt{n}}{\sigma}\mu)\\ &= P(-z_{\alpha/2} - \frac{\sqrt{n}}{\sigma}(\mu+\mu_0) \leq \frac{\sqrt{n}}{\sigma}(\overline{X}-\mu) \leq z_{\alpha/2} - \frac{\sqrt{n}}{\sigma}(\mu+\mu_0))\\ &= \Phi(z_{\alpha/2} - \frac{\sqrt{n}}{\sigma}(\mu-\mu_0)) - \Phi(-z_{\alpha/2} - \frac{\sqrt{n}}{\sigma}(\mu-\mu_0))\\ &= \Phi(\frac{\sqrt{n}}{\sigma}(\mu_0-\mu)+z_{\alpha/2}) - \Phi(\frac{\sqrt{n}}{\sigma}(\mu_0-\mu)-z_{\alpha/2})\end{aligned}$$

where Φ is the standard normal cumulative distribution function.
$\beta(\mu)$ is called the Operating Charectistic. The value of this function is only dependent on the gap between μ_0 and μ. For a fix α, as the gap grows, we move away from the centre of the standard normal. As such, the difference in the two terms of $\beta(\mu)$ keeps decreasing. It is maximum when $\mu = \mu_0$.

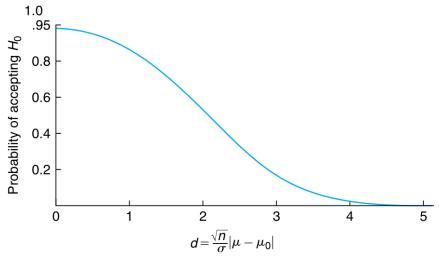


Figure 8: Curve of $\beta(\mu)$ for a fixed α

The function $1-\beta(\mu)$ is called the *power function* and is the probability of rejection of H_0 when the true mean is μ. This function is useful in calculating the value of the sample size so that the probability of accepting $H_0 : \mu = \mu_0$ when the true mean is μ_1 is β. We solve the equation $\beta(\mu_1) = \beta$ and try to guess the value of n because analytical solution is not possible.

$$n \approx \frac{(z_{\alpha/2}+z_\beta)^2\sigma^2}{(\mu_1-\mu_0)^2}$$

is the approximate solution assuming α is negligible.

24.1.3 One Sided Test

In one sided test, we test the equality of the mean with a fixed value vs a single sided inequality of the mean being larger than, or smaller than that fixed value.

$$\begin{aligned} \text{null hypothesis} \quad & H_0 : \mu = \mu_0 \\ \text{alternate hypothesis} \quad & H_1 : \mu > \mu_0 \end{aligned}$$

Note that the variance of the distribution is known in this case. Clearly, the critical region (rejection region) is one where the large values of μ are unlikely

$$C = X_1, \ldots, X_n : \overline{X} - \mu_0 > c$$

for some constant c chosen based on the significance level α. Equivalently,

$$\begin{aligned} P(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha) &= \alpha \\ \text{or, } \overline{X} &> z_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0 \end{aligned}$$

is the rejection region based on the sample mean.

$$\begin{aligned} \text{Reject} \quad H_0 \quad \text{if} \quad & \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0) > z_\alpha \\ \text{Accept} \quad H_0 \quad \text{if} \quad & \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0) \leq z_\alpha \end{aligned}$$

The *p-value* is similarly calculated as the probability that the standard normal is at least as large as this test statistic. Similar to the two sided test, operating characteristic curve can be defined

$$\begin{aligned} \beta(\mu) &= P(\text{Accepting } H_0) \\ &= P(\overline{X} \leq z_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0) \\ &= P(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}) \\ &= P(Z \leq z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}) \end{aligned}$$

where Z is the standard normal.

Special Note The tests discussed above have been derived under the assumption that the sample mean has a normal distribution. But, by central limit theorem, the sample mean of any large population will tend towards a normal distribution. Hence, the hypothesis tests will remain valid provided the population has known variance σ.

24.1.4 Unknown Variance

We proceed in a manner similar to the known variance case but use sample variance instead. Recall

$$\sqrt{n}\frac{\overline{X} - \mu_0}{S} \sim T_{n-1}$$

which is a t-distributed random variable with $n-1$ degrees of freedom. Since t-distribution also has specially defined values $t_{\alpha,n-1}$ similar to z_α, we can simply use the following 2-sided tests at significance level α

$$\begin{aligned} \text{Reject} \quad H_0 \quad \text{if} \quad & \left|\frac{\sqrt{n}}{S}(\overline{X}-\mu_0)\right| > t_{\alpha/2,n-1} \\ \text{Accept} \quad H_0 \quad \text{if} \quad & \left|\frac{\sqrt{n}}{S}(\overline{X}-\mu_0)\right| \leq t_{\alpha/2,n-1} \end{aligned}$$

Further, *p-values* are defined using the same statistic $\sqrt{n}(\overline{X}-\mu_0)/S$ and for any significance level which is less than the *p-value* probability that the t-statistic is greater than this statistic $\sqrt{n}(\overline{X}-\mu_0)/S$, we will reject $H_0 : \mu = \mu_0$. We accept the null hypothesis when the significance level is larger than the *p-value*.

Similar to the known variance case, we have the one sided tests defined as below

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0$$

$$\begin{aligned} \text{Reject} \quad H_0 \quad \text{if} \quad & \frac{\sqrt{n}}{S}(\overline{X}-\mu_0) > t_{\alpha,n-1} \\ \text{Accept} \quad H_0 \quad \text{if} \quad & \frac{\sqrt{n}}{S}(\overline{X}-\mu_0) \leq t_{\alpha,n-1} \end{aligned}$$

and the other side

$$H_0 : \mu \geq \mu_0 \quad \text{versus} \quad H_1 : \mu < \mu_0$$

$$\begin{aligned} \text{Reject} \quad H_0 \quad \text{if} \quad & \frac{\sqrt{n}}{S}(\overline{X}-\mu_0) < -t_{\alpha,n-1} \\ \text{Accept} \quad H_0 \quad \text{if} \quad & \frac{\sqrt{n}}{S}(\overline{X}-\mu_0) \geq -t_{\alpha,n-1} \end{aligned}$$

and we can calculate the *p-value* as well in the above cases using the test statistic.

24.2 Testing Equality of Means of Two Normal Populations

24.2.1 Known Variances

Consider the two populations as

$$\begin{aligned} X_1, X_2, \ldots, X_n &\sim \mathcal{N}(\mu_x, \sigma_x^2) \\ Y_1, Y_2, \ldots, Y_m &\sim \mathcal{N}(\mu_y, \sigma_y^2) \end{aligned}$$

Then, the difference in the sample means will itself be a normal distribution (since it is the difference of two normals), and the test statistic can be defined as below

$$\begin{aligned} \overline{X} - \overline{Y} &\sim \mathcal{N}(\mu_x - \mu_y, \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}) \\ T = \frac{(\overline{X}-\overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} &\sim \mathcal{N}(0,1) \end{aligned}$$

when H_0 is true,

$$T = \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim \mathcal{N}(0, 1)$$

and we compare the following hypotheses

$$\begin{aligned} & H_0 : \mu_x = \mu_y && \text{versus} \quad H_1 : \mu_x \neq \mu_y \\ \text{or,} \quad & H_0 : \mu_x - \mu_y = 0 && \text{versus} \quad H_1 : \mu_x - \mu_y \neq 0 \end{aligned}$$

we reject H_0 when the difference between the means is large, i.e., H_0 is testing whether the test statistic is close to zero or not

$$\begin{aligned} \text{Reject} \quad H_0 \quad \text{if} \quad T > & z_{\alpha/2} \\ \text{Accept} \quad H_0 \quad \text{if} \quad T \leq & z_{\alpha/2} \end{aligned}$$

where the variances of both the populations are known.

In a very similar fashion, the hypthesis testing rules for one sided test can be derived.
For

$$H_0 : \mu_x \leq \mu_y \quad \text{versus} \quad H_1 : \mu_x > \mu_y$$

We reject H_0 when the difference $\mu_x - \mu_y$ is highly positive

$$\begin{aligned} \text{Reject} \quad H_0 \quad \text{if} \quad & \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} > z_\alpha \\ \text{Accept} \quad H_0 \quad \text{if} \quad & \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \leq z_\alpha \end{aligned}$$

For the other side of the test, we use the same criteria as above, but switching the sets X and Y.

24.2.2 Unknown but Equal Variances

We consider the same two populations of Xs and Ys, but this time the variances are unknown. for simplicity of analysis, we assume that the unknown variances are same

$$\sigma_x = \sigma_y = \sigma$$

From section 23.2.4, we know

$$\begin{aligned} S_p^2 &= \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2} \\ \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} &\div \frac{S_p}{\sigma} \sim t_{n+m-2} \end{aligned}$$

If H_0 is true, $\mu_x = \mu_y$ and we have

$$T = \frac{(\overline{X} - \overline{Y})}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{n+m-2}$$

and we have the critical region defined as

$$\begin{aligned} \text{Reject} \quad H_0 \quad \text{if} \quad T &> t_{\alpha/2,n+m-2} \\ \text{Accept} \quad H_0 \quad \text{if} \quad T &\leq t_{\alpha/2,n+m-2} \end{aligned}$$

and for the one sided hypothesis

$$H_0 : \mu_x \leq \mu_y \quad \text{versus} \quad H_1 : \mu_x > \mu_y$$

We reject H_0 when the difference $\mu_x - \mu_y$ is highly positive

$$\begin{aligned} \text{Reject} \quad H_0 \quad \text{if} \quad T &> t_{\alpha,n+m-2} \\ \text{Accept} \quad H_0 \quad \text{if} \quad T &\leq z_{\alpha,n+m-2} \end{aligned}$$

24.2.3 Unknown and Unequal Variances

We consider the natural test statistic as follows

$$\frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

Even when H_0 is true, the above is not a simple distribution to solve for. If we make the additional assumption that n and m are large, then

$$\frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \sim \mathcal{N}(0, 1)$$

and the same criteria for accepting and rejecting H_0 discussed in section 24.2.1 are applicable, but after replacing population variances with sample variances.

24.2.4 Unknown and Unequal Variances

Suppose we want to observe the change in a quantity in a sample, after some kind of intervention. A simple example can be change in the mileage of a car after installation of a catalytic converter. Suppose we have n samples with us, and for each of the sample, we associate X_i with the measurement of the quantity before intervention and Y_i with the quantity post intervention. Note that X_i is not independent of Y_i because they come from the same i^{th} sample. Hence, the test discussed in section 24.2.1 is not applicable.

Instead, we consider the quantity $W = X - Y$ and assume that W_i come from a normal population. We can then consider the hypothesis

$$H_0 : \mu_w = 0 \quad \text{versus} \quad H_1 : \mu_w \neq 0$$

Using the results derived in section 24.1.4, we have

$$\begin{aligned} \text{Reject} \quad H_0 \quad \text{if} \quad \sqrt{n}\frac{\overline{W}}{S_w} &> t_{\alpha/2,n-1} \\ \text{Accept} \quad H_0 \quad \text{if} \quad \sqrt{n}\frac{\overline{W}}{S_w} &\leq t_{\alpha/2,n-1} \end{aligned}$$

One sided tests can be derived in exactly the same manner as section 24.1.4 and the concepts discussed in 24.1.2 still hold true.

24.3 Tests around Variance of Normal Population

For a n sized sample of independent observations from a normal population, we are interested in checking

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{versus} \quad \sigma^2 \neq \sigma_0^2$$

Recall from section 22.3

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Then if H_0 is true, our test statistic

$$TS = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

and from the test simply becomes

$$\begin{aligned} \text{Accept} \quad H_0 \quad \text{if} \quad \chi^2_{1-\alpha/2,n-1} \leq TS \leq \chi^2_{\alpha/2,n-1} \\ \text{Reject} \quad H_0 \qquad\qquad\qquad\qquad \text{otherwise} \end{aligned}$$

One sided test can be done in a similar manner, comparing with $\chi^2_{1-\alpha,n-1}$ or $\chi^2_{\alpha,n-1}$ based on which side we want to reject H_0.

24.3.1 Comparing Variance of Two Normal Populations

We are interested in comparing

$$H_0 : \sigma_x^2 = \sigma_y^2 \quad \text{versus} \quad \sigma_x^2 \neq \sigma_y^2$$

Recall that the ratio of sample variance with population variance is t-distributed, and the ratio of two t-distributed variables has an F-distribution. Hence, when H_0 is true,

$$TS = \frac{S_x^2}{S_y^2} \sim F_{n-1,m-1}$$

Noting that F-distribution is always positive, the region for accepting H_0 simply become

$$\begin{aligned} \text{Accept} \quad H_0 \quad \text{if} \quad F_{1-\alpha/2,n-1,m-1} \leq TS \leq F_{\alpha/2,n-1,m-1} \\ \text{Reject} \quad H_0 \qquad\qquad\qquad\qquad \text{otherwise} \end{aligned}$$

24.4 Tests around Bernoulli Population

Suppose we have a set of n samples and we want to test how many of them satisfy a property (or equivalently, success). Let p be the fraction of population satisfying he property and we want to check if this equals p_0

$$H_0 : p \leq p_0 \quad \text{versus} \quad p > p_0$$

i.e., we reject this batch if the size of sample not satisfying the property (defective) is more than some predefined quantity/significance p_0.

We reject when the defectives in the sample (X) are more than a threshold k

$$P(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} = \sum_{i=k}^{n} p^i (1-p)^{n-i}$$

which is an increasing function in p. Hence, when H_0 is true,

$$P(X \geq k) \leq \sum_{i=k}^{n} p_0^i (1-p_0)^{n-i}$$

and we reject when $X \geq k^*$ depending on the significance level α

$$k^* = \text{minimum} \quad k \quad \text{where} \quad \sum_{i=k}^{n} p_0^i (1-p_0)^{n-i} \leq \alpha$$

because there can be multiple k which satisfy the above equation, and we want to reject H_0 as soon as the number of defectives in sample X is more than the minimum k.

The test can also be done using *p-value*

$$\begin{aligned} \textit{p-value} &= P(Bin(n, p_0) \geq x) \\ &= \sum_{i=x}^{n} p_0^i (1-p_0)^{n-i} \end{aligned}$$

where x is the count of defects in the sample. We reject H_0 at any $\alpha > \textit{p-value}$ since in that situation the number of defects required will be much less than x.
For large n, X will behave like a normal distribution and when H_0 is true,

$$\frac{X - np_0}{\sqrt{np_0(1-p_0)}} \sim \mathcal{N}(0,1)$$

and criteria discussed in section 24.1.1 hold.

25 Linear Regression

We are given pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ (all independent) where we assume that x and y are governed by the linear relation

$$y \approx \theta_0 + \theta_1 x$$

The aim is to determine the model which is parametric consisting of two parameters θ_0 and θ_1. We find it using the least squares estimate, i.e., minimizing

$$\underset{\theta_0, \theta_1}{\text{minimize}} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x)^2$$

The true model also includes noise and is given by

$$Y_i = \theta_0 + \theta_1 X_i + W_i$$

where we assume the noise $W_i \sim \mathcal{N}(0, \sigma^2)$ and is independently and identically distributed. Observing some X and Y is same as observing the noise.

$$P(X = x, Y = y) = P(W = y - \theta_0 - \theta_1 x) = \frac{1}{\sqrt{2\pi\sigma_2}} \exp\left(- \frac{(y - \theta_0 - \theta_1 x)^2}{2\sigma^2} \right)$$

$$\begin{aligned} P(X_1 = x_1, Y_1 = y_1, \ldots, X_n = x_n, Y_n = y_n) &= \prod_{i=1}^{n} P(X_1 = x_i, Y_i = y_i) \\ &= \prod_{i=1}^{n} W_i = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_2}} \exp\left(- \frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2} \right) \end{aligned}$$

Maximizing the above product is maximizing the likelihood of the occurrence of the data under the model parameters θ_0 and θ_1. Since taking log will not change the maxima, we usually maximize the log likelihood

$$\underset{\theta_0, \theta_1}{\text{maximize}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2}) = \underset{\theta_0, \theta_1}{\text{minimize}} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

We can take derivatives with respect to the parameters of the above function to get the estimate for the parameters as

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \\ \hat{\theta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{E[(X - \overline{X})(Y - \overline{Y})]}{E[(X - \overline{X})^2]} = \frac{Cov(X, Y)}{Var(X)} \\ \hat{\theta}_0 &= \bar{y} - \hat{\theta}_1 \bar{x} \end{aligned}$$

The above formulae can also be derived if the additives are a function of X. Since the linear relationship will still be respected and the loglikelihood can be maximized to get the estimates of the parameters.

Some useful notation

$$\begin{aligned} S_{xY} &= \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \overline{Y}) &&= (\sum_{i=1}^{n} x_i Y_i) - n\bar{x}\overline{Y} \\ S_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 &&= (\sum_{i=1}^{n} x_i^2) - n\bar{x}^2 \\ S_{YY} &= \sum_{i=1}^{n} (Y_i - \overline{Y})^2 &&= (\sum_{i=1}^{n} Y_i^2) - n\overline{Y}^2 \end{aligned}$$

25.1 Mean and Variance of Coefficients

First note that

$$\begin{aligned} E[Y_i] &= E[\theta_0 + \theta_1 X_i + W_i] = \theta_0 + \theta_1 X_i \\ E[\overline{Y}] &= (\sum_{i=1}^{n} E[Y_i])/n = \theta_0 + \theta_1 \overline{X} \\ Var(Y_i) &= \sigma_2 \end{aligned}$$

Thus,

$$\begin{aligned}
E[\hat{\theta}_1] &= E\left[\frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] \\
&= E\left[\frac{\sum_{i=1}^n (x_i - \overline{x})(E[Y_i] - E[\overline{Y}])}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] \\
&= \theta_1 \\
E[\hat{\theta}_0] &= E[\overline{Y} - \hat{\theta_1}\bar{x}] = \theta_0
\end{aligned}$$

meaning that our estimates of the parameters are unbiased and their error will equal the variance

$$\begin{aligned}
Var(\hat{\theta}_1) &= Var\left(\frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2}\right) = Var\left(\frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}\right) \\
&= \frac{1}{(\sum_{i=1}^n (x_i - \overline{x})^2)^2} \sum_{i=1}^n (x_i - \overline{x})^2 Var(Y_i) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
Var(\hat{\theta}_0) &= Var(\overline{Y} - \hat{\theta_1}\bar{x}) = Var\left(\sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right) \\
&= \frac{\sigma^2}{n^2}\left(\sum_{i=1}^n \left(\frac{\sum_{i=1}^n x_i^2 - n\bar{x}x_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right)^2\right) = \frac{\sigma^2}{n^2(\sum_{i=1}^n x_i^2 - n\bar{x}^2)^2}(n(\sum_{i=1}^n)^2 - n^2\bar{x}^2(\sum_{i=1}^n)^2) \\
&= \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n\big((\sum_{i=1}^n x_i^2) - n\bar{x}^2\big)}
\end{aligned}$$

because both the estimators are linear combinations of independent identically distributed normal random variables $Y_i s$, and the variance of linear combination of independent random variables is simply the sum of variances multiplied by squares of coefficients.

Thus, $\hat{\theta_0}$ **and** $\hat{\theta}_1$ **are both normally distributed random variables.** with the following distributions

$$\begin{aligned}
\hat{\theta}_1 &\sim \mathcal{N}\left(\theta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right) \\
\hat{\theta}_0 &\sim \mathcal{N}\left(\theta_1, \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n\big((\sum_{i=1}^n x_i^2) - n\bar{x}^2\big)}\right)
\end{aligned}$$

25.2 Distribution of Residual

Residuals and the SS_R are defined as

$$\begin{aligned}
R &= Y - (\theta_0 + \theta_1 X) \\
SS_R &= \sum_{i=1}^n R_i^2 = \sum_{i=1}^n (Y - \theta_0 - \theta_1 X)^2 \\
&= \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}
\end{aligned}$$

SS_R is itself a random variable and it can be shown that

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$
$$E[\frac{SS_R}{\sigma^2}] = n-2$$
$$E[\frac{SS_R}{n-2}] = \sigma^2$$

since SS_R/σ^2 is the sum of squares of normally distributed variables ($E[Y] = \theta_0 + \theta_1 X$) and two degrees of freedoms are already taken up by the coefficients. Further, SS_R is an unbiased estimator of the variance of the error terms σ^2, and is also independent of the coefficients.

25.3 Inferences Concerning Coefficients

We are most interseted in checking whether a coefficient has an effect or not

$$H_0 : \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_1 \neq 0$$

We know from above derivations that

$$\frac{\hat{\theta}_1 - \theta_1}{\sigma^2/S_{xx}} \sim \mathcal{N}(0,1)$$
$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

and both the random variables are independent of each other. Hence their division is t-distributed random variable and when H_0 is true, $\theta_1 = 0$

$$\frac{\sqrt{S_{xx}}\hat{\theta}_1/\sigma}{\sqrt{\frac{SS_R}{\sigma^2(n-2)}}} = \hat{\theta}_1\sqrt{\frac{(n-2)S_{xx}}{SS_R}} = TS \sim t_{n-2}$$

We do this since we do not know the exact value of σ^2 and need to eliminate it with a sample derived version. The hypothesis test at significance level α simply becomes

$$\text{Reject} \quad H_0 \quad \text{if} \quad |TS| > t_{\alpha/2,n-2}$$
$$\text{Accept} \quad H_0 \quad \text{if} \quad |TS| \leq t_{\alpha/2,n-2}$$

which can be converted to a *p-value* using the TS and t-distribution. A small *p-value* will lead to rejection of H_0 meaning that the data provides evidence of a relationship between dependent and independent variables.

A confidence interval for θ_1 at $1-\alpha$ confidence can be obtained as follows

$$P(-t_{\alpha/2,n-2} < (\hat{\theta}_1 - \theta_1)\sqrt{\frac{(n-2)S_{xx}}{SS_R}} < t_{\alpha/2,n-2}) = 1-\alpha$$

$$\text{Confidence Interval is} \quad \left(\hat{\theta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{SS_R}{(n-2)S_{xx}}} < \theta_1 < \hat{\theta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{SS_R}{(n-2)S_{xx}}}\right)$$

The hypothesis test for θ_0 can be done in the exact same manner as θ_1 by considering the following test statistic

$$TS = (\hat{\theta}_1 - \theta_1)\sqrt{\frac{n(n-2)S_{xx}}{(\sum_{i=1}^n x_i^2)SS_R}} \sim t_{n-2}$$

25.4 Inferences Concerning Mean Response

For any new point x_0, the unbiased estimator for the response is

$$\begin{aligned} y_0 &= \hat{\theta}_0 + \hat{\theta}_1 x_0 \\ E[y_0] &= E[\hat{\theta}_0] + E[\hat{\theta}_1]E[x_0] = \theta_0 + \theta_1 x_0 \end{aligned}$$

To get the distribution of this mean response, note that

$$\begin{aligned} Y_0 &= \hat{\theta}_0 + \hat{\theta}_1 x_0 = \overline{Y} - \hat{\theta}_1 \bar{x} + \hat{\theta}_1 x_0 \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i + (x_0 - \bar{x})\frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x - \bar{x})^2} \\ &= \sum_{i=1}^{n}\left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}}\right)Y_i \end{aligned}$$

which is a linear combination of independent normally distributed random variables $Y_i s$. Thus, the mean response is also a normally distributed random variable and we can get the confidence intervals by considering the mean and variance of this random variable

$$\begin{aligned} Var(\hat{\theta}_0 + \hat{\theta}_1 x_0) &= \sum_{i=1}^{n}\left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}}\right)^2 Var(Y_i) \\ \hat{\theta}_0 + \hat{\theta}_1 x_0 &\sim \mathcal{N}\left(\theta_0 + \theta_1 x_0, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right) \end{aligned}$$

To eliminate σ^2,

$$\begin{aligned} SS_R/\sigma^2 &\sim \chi^2_{n-2} \\ \frac{(\hat{\theta}_0 + \hat{\theta}_1 x_0) - (\theta_0 + \theta_1 x_0)}{\sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]} &\div \sqrt{\frac{SS_R}{(n-2)\sigma^2}} \sim t_{n-2} \end{aligned}$$

and the confidence intervals for confidence $1 - \alpha$ become

$$(\hat{\theta}_0 + \hat{\theta}_1 x_0) \pm t_{\alpha/2, n-2}\sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\left(\frac{SS_R}{n-2}\right)}$$

25.5 Inferences Concerning Future Response

The above section 25.4 discussed the distribution of the mean response. In many scenarios, we are interested in the distribution of the actual response Y at input x_0, which takes the noise into account as well. We note

$$\begin{aligned} Y_0 &\sim \mathcal{N}(\theta_0 + \theta_1 x_0, \sigma^2) \\ \hat{\theta}_0 + \hat{\theta}_1 x_0 &\sim \mathcal{N}\left(\theta_0 + \theta_1 x_0, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right) \\ Y_0 - \hat{\theta}_0 - \hat{\theta}_1 x_0 &\sim \mathcal{N}\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right) \end{aligned}$$

Now we utilise the distribution of SS_R to eliminate σ^2 and get to the t-distribution

$$\frac{Y_0 - \hat{\theta}_0 - \hat{\theta}_1 x_0}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \div \sqrt{\frac{SS_R}{(n-2)\sigma^2}} \sim t_{n-2}$$

and the **prediction** interval for the response (not mean response is) at $1-\alpha$ confidence

$$(\hat{\theta}_0 + \hat{\theta}_1 x_0) \pm t_{\alpha/2,n-2}\sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\left(\frac{SS_R}{n-2}\right)}$$

Note that **prediction interval is the interval where we expect the value of a random variable to lie, whereas the confidence interval is the one where the value of a parameter estimate to lie.**

25.6 Coefficient of Determination

Let's consider the variation in respone Y

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

and the variation in the response after removing the effect of inputs

$$SS_R = \sum_{i=1}^{n}(Y_i - \theta_0 - \theta_1 x_0)^2$$

and thus,

$$S_{YY} - SS_R$$

is the variation explained by the inputs. We define R^2 as

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}} = 1 - \frac{SS_R}{S_{YY}}$$

R^2 is the proportion of total variance explained by the inputs. A value close to 1 implies most of the variance is explained by the inputs whereas 0 means little variance is explained by inputs.

It can also be shown that the absolute value of correlation coefficient between x and Y equals the coefficient of determination. Thus, we know the value of R^2 for simple linear regression directly by r.

25.7 Weighted Least Squares

Suppose we know that the variance of Y is dependent on Y itself in the form $Var(Y_i) \propto \sigma^2/w_i$, i.e., the weights are known only upto a constant. In this case, we minimize the weighted least squares to obtain the coefficients

$$\underset{\theta_0,\theta_1}{\text{minimize}} \sum_{i=1}^{n} w_i(Y_i - \theta_0 - \theta_1 x_i)^2$$

26 Life Testing

This section develops statistical methods around estimtating distribution of variables indicating the lifetime of a particular object. For instance, if the lifetime has an exponential distribution, we utilise the sample to obtain the parameters of the exponential distribution.

Let X be a continuous random variable denoting lifetime of an item, having cumulative distribution F and density function f, then

$$\text{hazard function or failure rate,} \quad \lambda(t) = \frac{f(t)}{1-F(t)}$$

$$P(X \in (t, t+dt)|X > t) = \frac{P(X \in (t, t+dt), X > t)}{P(X > t)} = \frac{P(X \in (t, t+dt))}{P(X > t)} \approx \frac{f(t)}{1-F(t)}dt$$

$\lambda(t)$ **denotes the conditional probability that an item of age t will fail in the next moment.**

For exponential distribution, $\lambda(t) = (\lambda e^{-\lambda x})/e^{-\lambda x} = \lambda$ because of the memoryless propoerty.

Hazard function uniquely determines the cumulative distribution F

$$\lambda(s) = \frac{\frac{d}{ds}F(s)}{1-F(s)} = \frac{d}{ds}(-log(1-F(s)))$$

$$F(t) = 1 - exp\big(-\int_0^s \lambda(s)ds\big)$$

26.1 Exponential Distribution: Stopping at rth failure

Suppose we have n items with exponentially distributed lifetime with unknown parameter and we wish to estimate the mean $\theta(note \lambda = 1/\theta)$. We observe the items until r failures and try to estimate θ. Let X_i denote the lifetime of the i^{th} item with the following notation

$$x_1 \leq x_2 \leq \ldots \leq x_r \quad \text{for} \quad i_1, i_2, \ldots, i_n$$

i.e., $X_{i_j} = x_j$. Then the joint likelihood becomes

$$\begin{aligned}
L &= \big(\prod_{i=1}^{r} \frac{1}{\theta} e^{-x_i/\theta}\big)\big(\prod_{j=r+1}^{n} e^{-x_r/\theta}\big) \\
&= \frac{1}{\theta^r} exp\big\{-\frac{1}{\theta}\big(\sum_{i=1}^{r} x_i + (n-r)x_r\big)\big\} \\
\log(L) &= -r\log(\theta) - \frac{\sum_{i=1}^{r} x_i}{\theta} - \frac{(n-r)x_r}{\theta} \\
\frac{d}{d\theta}\log(L) &= -\frac{r}{\theta} + \frac{\sum_{i=1}^{r} x_i}{\theta^2} + \frac{(n-r)x_r}{\theta^2} \\
\hat{\theta} &= \frac{\sum_{i=1}^{r} x_i + (n-r)x_r}{r} = \frac{\tau}{r}
\end{aligned}$$

where we note that for $n-r$ items, the lifetime is known only to be more than x_r and thus we use the cumulative probability of lifetime $> x_r$ in the likelihood equation. We can replace the values with random variables in above equations.

τ is the total time on test, i.e. the total time of survival of each item for the duration the test ran (X_r). Now, we can rewrite τ using the differences between consecutive times of failures. Note than all items survive for X_1 time, $n-1$ items survive for at least $X_2 - X_1$ time, and so on till $n-r+1$ items survive for additional $X_r - X_{r-1}$ time. Thus,

$$\tau = nX_1 + (n-1)(X_2 - X_1) + \cdots + (n-r+1)(X_r - X_{r-1})$$

and from answer, we know that X_1 is exponential with mean θ/n and thus, nX_1 has mean θ. By memoryless property, $X_2 - X_1$ is also exponential with mean $\theta/(n-1)$ and so $(n-1)(X_2 - X_1)$ has mean θ. Thus, τ is the sum of independent exponential variables and is a Gamma distribution with parameters $(r, 1/\theta)$. Since Gamma distribution is related to a χ^2 distribution (from section 16.1)

$$\frac{2\tau}{\theta} \sim \chi^2_{2r}$$
$$P(\chi^2_{1-\alpha/2,2r} < \frac{2\tau}{\theta} < \chi^2_{\alpha/2,2r}) = 1 - \alpha$$
$$\theta \in \left(\frac{2\tau}{\chi^2_{\alpha/2,2r}}, \frac{2\tau}{\chi^2_{1-\alpha/2,2r}} \right) \quad \text{with confidence } 1 - \alpha$$

27 Simulation, Random Numbers, Permutation Tests

27.1 Random Numbers

We can generate random numbers using the following equation

$$x_{n+1} = (ax_n + c) mod(m)$$

x_n takes the values $1, 2, \ldots, m-1$ and we take x_n/m as the pseudo random number, which is uniformly distributed between $(0, 1)$ for suitable choice of a, c, m.

27.1.1 Permutation of Integers

Suppose we want to generate a permutation of integers from $1, 2, \ldots, n$ such that each of the permutations is equally likely. Assuming we have a uniform random generator U with us,

$$P(Int(kU) + 1 = i) = P(Int(kU) = i - 1) = P(i - 1 \leq kU < i)$$
$$= P(\frac{i-1}{k} \leq U < \frac{i}{k}) = \frac{1}{k}$$

which gives us randomly generated random integers between 1 and k with equal probability. An easy way to generate permutation is

1. Choose a permutation $r_1, r_2, \ldots, r_n$ which can just be $r_j = j$
2. Let $k = n$
3. Choose a random number U and let $I = Int(kU) + 1$
4. Interchange numbers at position k and I
5. $k = k - 1$
6. if $k > 1$ goto step 3 else return permutation

The above algorithm can also be used to get a random subset of size r from a set $1, \ldots, k$ by simply running the algorithm till $k = r$ since the elements in the last r positions can be selected. For $r > n/2$, we find the $k = n - r$ elements not in the subset.

27.2 Bootstrap Method

27.3 Generating Discrete Random Variables

Suppose we want to generate the random variable X with probability mass function

$$P(X = x_i) = p_i, i = 1, 2, \ldots, n \ \sum_{i=1}^{n} p_i = 1$$

Then using a uniform random generator U, we can generate the discrete random variable using

$$X = x_i \quad \text{if} \quad p_1 + p_2 + \cdots + p_{i-1} \leq U < p_1 + p_2 + \cdots + p_i$$

i.e., we divide the number line at points $p_1, p_1 + p_2, \ldots, 1$ and choose the i^{th} interval such that U falls in that interval. This algorithm is valid since

$$P(a \leq U < b) = b - a$$
$$P(\sum_{j=1}^{i-1} p_j \leq U < \sum_{j=1}^{i} p_j) = p_i$$

This method is known as *discrete inverse transform method.*

27.3.1 Binomial Random Variable

To generate a Bernoulli random variable, we simply select $X = 1$ if $U < p$ otherwise $X = 0$. Similarly a binomial random variable can be generated using individual Bernoulli variables as described. A more efficient method is to use the inverse transform method. For number of successes $0, 1, 2, \ldots, n$, we must calculate the probability mass function. This can be done efficiently using recursion

$$p_i = P(X = i) = \binom{n, i}{p}^i (1 - p)^{n-i}$$
$$\frac{p_{i+1}}{p_i} = \frac{n - i}{i + 1} \frac{p}{1 - p}$$

The algorithm is then simply

1. Assign $i = 0, P = p_0 = (1 - p)^n, F = P, b = p/(1 - p)$
2. Generate random number $U \in (0, 1)$
3. if $U \leq F, X = i$, stop else continue
4. Update P to get p_{i+1}, $P = Pb\frac{n-i}{i+1}$
5. Update the cumulative probability $F = F + P$
6. increase $i = i + 1$, goto 3

The average number of iterations taken by the algorithm $= E[X + 1] = np + 1$ since total values checked are $n + 1$.

28 Exercises

28.1 Problems

1. **Independence in Complements**
   Given $A \perp B$, show $A \perp B^c$ and $A^c \perp B^c$. Solution

2. **Conditional Independence**
   A, B, and C are independent with $P(C) > 0$. Show that $A \perp B|C$. Solution

3. **Geometry of Meeting**
   R and J have to meet at a given place and each will arrive at the given place independent of each other with a delay of 0 to 1hr uniformly distributed. The pairs of delays are all equally likely. The first to arrive waits for 15 minutes and leaves. What is the probability of meeting ? Solution

4. **Expectation of Function**
   Let X and Y be random variables with $Y = g(X)$. Show $E[Y] = \sum_x g(x)p_X(x)$. Solution

5. **Cumulative Distribution Function**
   A random variable X is a combination of a continuous and discrete distribution as follows

$$f_X(x) = \begin{cases} 0.5 & a \leq x \leq b \\ 0.5 & \text{x} = 0.5 \\ 0 & \text{otherwise} \end{cases}$$

   Find the Cumulative Distribution of X. Solution

6. **Number of tosses till first head**
   When tossing a fair coin, what is the $E[\#$ tosses till the first H$]$. Solution

7. **Iterated Expectation Proof**
   For discrete variables, show $E[X] = E[E[X|Y]]$. Solution

8. **Iterated Expectation for three variables**
   For three random variables X, Y and Z, show $E[Z|X] = E[E[Z|X, Y]|X]$. Solution

9. **Iterated Expectation practice**
   A class has two sections denoted by the random variable Y. Let X denote the quiz score of a student. Given that section 1 has 10 students, section 2 has 20 students, $E[X|Y = 1] = 90, E[X|Y = 2] = 60, Var(X|Y = 1) = 10, Var(X|Y = 2) = 20$, find $E[X]$ and $Var(X)$. Solution

10. **Hat Problem**
    n people throw their hats in a box and then pick a hat at random. What is the expected number of people who pick their own hat ? Solution

11. **Breaking a stick**
    A stick of length l is broken first at X uniformly chosen between $[0, l]$, and then at Y, uniformly chosen between $[0, X]$. Find the expected length of the shorter part. Solution

12. **Convolution of Exponentials**
    Suppose $X \sim exp(\lambda)$ and $Y \sim exp(\mu)$, find the probability distribution $p_{X+Y}(x)$.

13. **Triangles from a Stick**
    We have a stick of length 1. We randomly choose two points on the stick and break the stick at those points. Calculate the probability that the three pieces form a triangle. Solution

14. **PMF of g(X)**
Let X be uniform in $[0, 2]$, then find the PMF of $Y = X^3$. Solution

15. **Waiting for Taxi**
A taxi stand and bus stop near Al's home are at the same location. Al goes there and if a taxi is waiting $P = \frac{2}{3}$, he boards it. Otherwise, he waits for a taxi or bus to come, whichever is first. Taxi takes anywhere between 0 to 10 mins (uniform) while a bus arrives in exactly 5 mins. He boards whichever is first. Find CDF and E[wait time]. Solution

16. **Bayes Theorem**
Let Q be a continuous random variable with PDF

$$f_Q(q) = \begin{cases} 6q(1-q) & 0 \leq q \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

where Q represents $P(success)$ for a Bernoulli X, i.e., $P(X = 1|Q = q) = q$. Find $f_{Q|X}(q|x) \forall x \in [0, 1] and q$. Solution

17. **A Normal Transformation**
Let $X \sim \mathcal{N}(0, 1)$ and $Y = g(X)$. Find $p_Y(y)$.

$$g(t) = \begin{cases} -t & t \leq 0 \\ \sqrt{t} & t > 0 \end{cases}$$

Solution

18. **Binomial Shooter**
A shooter takes 10 hits in a shooting range and each shot has $p = 0.2$ of hitting target independent of each other. Let X = number of hits. Find

(a) PMF of X
(b) $P(no\ hits)$
(c) $P(scoring\ more\ than\ misses)$
(d) $E[X]$ and $Var(X)$
(e) Suppose the entry is \$3 and each shot fetches \$2. Let Y = profit. Find $E[Y] and Var(Y)$.
(f) Suppose entry is free and total reward is square of number of hits. Let Z be profit. Find $E[Z]$.

Solution

19. **Mosquito and Tick**
Every second, a mosquito lands with $P = 0.5$. Once it lands, it bites with $P = 0.2$. Let X be the time between successive mosquito bites. Find $E[X]$ and $Var(X)$.
Now suppose a tick comes into play independent of mosquito. It lands with $P = 0.1$ and once landed, bites with) $= 0.7$. Let Y be the time between successive bug bites. Find $E[Y]$ and $Var(Y)$. Solution

20. **HH or TT**
Given a coin with $P(H) = p$, find the E[number of tosses till HH or TT]. Solution

21. **A Three Coin Game**
Let 3 fair coins be tossed at every turn. Given all coins and turns are independent, calculate the following (assuming success is defined as all three coins landing the same side up))

(a) PMF of K, no of trials upto but not including the 2^{nd} success

(b) E and Var of M, the E[number of tails] before first success.

Solution

22. **Linear Expectations**
Bob conducts trials in a similar manner to Problem 21, but with four coins. He repeatedly removes a coin at success until just a single coin remains. Calculate the Expected number of tosses till the finish of experiment. Solution

23. **Papers Drawn with Replacement**
Suppose there are n papers in a drawer. We take one paper, sign it, and then put it back into the drawer. We take one more paper out and if it is not signed, we sign it and put it back in the drawer. If the paper is already signed, we simply put it back in the drawer. We repeat this process until all the papers are signed. Find the E[papers drawn till all papers are signed]. What is the value of this quantity as $n \to large$. Solution

24. **A Three Variable Inequality**
Let X, Y, Z be three exponentially distributed random variables with parameters λ, μ, and ν respectively. Find $P(X < Y < Z)$. Solution

25. **Poisson Emails**
You get emails according to a Poisson process at the rate of 5 messages/hour. You check email every 30 minutes. Find

- P(no new message)
- P(one new message)

Solution

26. **Poisson Fishing**
We go fishing where we catch fishes at the rate of $0.6/hour$. We fish for two hours. If we do not catch a fish in the first two hours, we fist until the first catch. Find the following

- P(fish for > 2 hours)
- P(fish for > 2 but < 5 hours)
- P(catch at least two fish)
- E[fish]
- E[Total fishing time]
- E[future fishing time—fished for two hours]

Solution

27. **Poisson Lightbulbs**
We have three identical but independent lightbulbs whose lifetimes are modelled by a Poisson process with parameter λ. Given that we start all the three bulbs together, find the E[time until last bulb dies out]. Solution

28. **Two Poisson Lightbulbs**
Beginning at $t = 0$, we begin using bulbs one at a time until failure. Any broken bulb is immediately replaced. Each new bulb is selected independently and equally likely from type A(exponential life with $\lambda = 1$) or type B(exponential life with $\lambda = 3$). Lifetimes of all bulbs are independent.

(a) Find E[time until first failure].

(b) P(no bulb failure before time t).

(c) Given that there are no failures until time t, determine the conditional probability that the first bulb used is of type A.

(d) Find the probability that the total illumination by two type B bulbs > one type A.

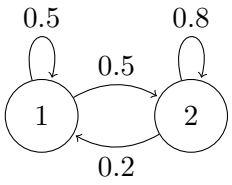
(e) Suppose the process terminates after 12 bulbs fail. Determine the expected value and variance of the total illumination provided by type B bulbs while the process is in operation.

(f) Given there are no failures until time t, find the expected value of time until first failure.

Solution

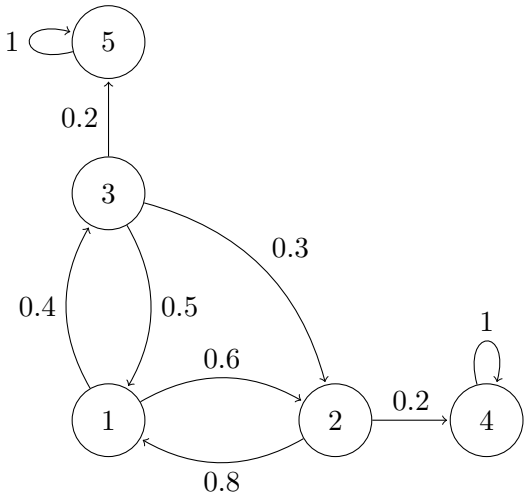
29. **Minimum of Exponentials**
Given n independent exponential random variables with different parameters, find the distribution for their minimum. Solution

30. **Steady State Markov Process**
Find the steady state probabilites of the following Markov Process



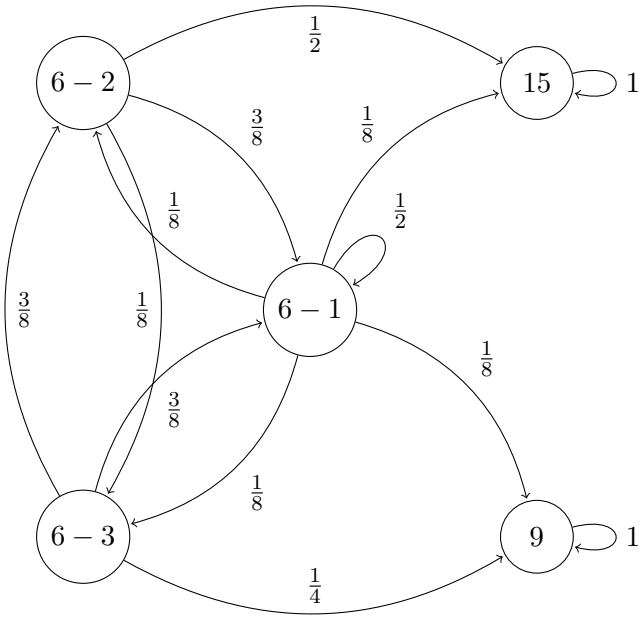
Solution

31. **Absorption Probabilities**



Calculate the absorption probabilites for state 4 and expected time to absortion from all states. (for absorption time, assume $p_{35} = 0$ and $p_{32} = 0.5$) Solution

32. **Selecting Courses with Markov Process**



Consider the above markov process for changing courses. The probability being in some course tomorrow given a course today is mentioned along the edges. Suppose we start with course 6-1 (Note that course 6 is the combination of courses 6-1, 6-2 and 6-3). Calculate the following

(a) P(eventually leaving course 6).

(b) P(eventually landing in course 15).

(c) E[number of days till leaving course 6].

(d) At every switch for 6-2 to 6-1 or 6-3 to 6-1, we buy an ice cream (but a maximum of two). Calculate the E[number of ice creams before leaving course 6].

(e) Suppose we end up in 15. What is the E[number of steps to reach 15].

(f) Suppose we don't want to take course 15. Accordingly, when in 6-1, we stay there with probability 1/2 while other three options have equal probabilities. If we are in 6-2, probability of going to 6-1 and 6-3 are in the same ratio as before. Calculate the E[number of days until we enter course 9].

(g) Assuming $P(X_{n+1} = 15|X_n = 9) = P(X_{n+1} = 9|X_n = 15) = P(X_{n+1} = 15|X_n = 15) = P(X_{n+1} = 9|X_n = 9) = 1/2$, what is $P(X_n = 15)$ and $P(X_n = 9)$ far into the future.

(h) Suppose $P(X_{n+1} = 6-1|X_n = 9) = 1/8, P(X_{n+1} = 9|X_n = 9) = P(X_{n+1} = 15|X_n = 15) = 7/8$. What is the E[number of days till return to 6-1].

Solution

33. **Estimating Binomial with CLT, 1/2 correction**
Given a Bernoulli Process with $n = 36$ and $p = 0.5$, find $P(S_n \leq 21)$. Solution

34. **Sample Variance for Normal Distribution**
The time it takes a central processing unit to process a certain type of job is normally distributed with mean 20 seconds and standard deviation 3 seconds. If a sample of 15 such jobs is observed, what is the probability that the sample variance will exceed 12 ? Solution

35. **MLE Estimate**
Suppose we observe n independent and identically distributed samples $x_1, x_2, \ldots, x_n$ from an exponential distribution. Estimate the parameter of the exponential. Solution

36. **Bayes Estimator for Normal Distribution**
Suppose $X_1, X_2, \ldots, X_n$ are from a normal distribution with unknown mean θ and known variance σ_0^2, and suppose the mean has a prior normal ditribution with mean μ and variance σ^2. Calculate the Bayes estimator for the mean θ. Solution

37. **LMS Estimate**
Given the prior $f_{\Theta|(\theta)}$, uniform in $[4, 10]$, and $f_{X|\Theta}(x|\theta)$ is uniform in $[\theta-1, \theta+1]$, estimate the posterior of θ. Solution

38. **Probability Convergence**
Let X be uniformly distributed between $[-1, 1]$. Let $X_1, X_2, \ldots, X_n$ be independently and identically distributed with the same distribution as X. Find whether the following sequences are convergent in probability and also find the limit.

    (a) X_i
    (b) $Y_i = X_i/i$
    (c) $Z_i = (X_i)^i$

    Solution

39. **Age of Smokers vs Non Smokers**
One often hears that the death rate of a person who smokes is, at each age, twice that of a nonsmoker. What does this mean? Does it mean that a nonsmoker has twice the probability of surviving a given number of years as does a smoker of the same age? Solution

28.2 Solutions

1. Question

    (a)

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

    (b)

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

2. Question

$$P(A \cap B|C) = \frac{P(A \cap B \cap C)}{P(C)} = P(A)P(B) = P(A|C)P(B|C) \quad \text{Due to independence}$$

3. Question
Suppose R arrives at x hours. J has to arrive between x hrs to x hrs + 15 mins. Similarly if J arrives at y hours, R has to arrive between y hours to y hours + 15 mins. These are regions enclosed by the regions $x \leq 1, y \leq 1, y \leq x + \frac{1}{4}$ *and* $y \geq x - \frac{1}{4}$. The probability is then $1 - P(\textit{not meeting}) = 1 - 2(\frac{1}{2}\frac{3}{4}\frac{3}{4}) = \frac{7}{16}$.

4. Question

$$\begin{aligned} E[Y] = \sum_y y p_Y(y) &= \sum_y \sum_{x:g(x)=y} p_X(x) = \sum_y \sum_{x:g(x)=y} y p_X(x) \\ &= \sum_y \sum_{x:g(x)=y} g(x) p_X(x) = \sum_x g(x) p_X(x) \end{aligned}$$

5. Question
Cumulative Distribution of X can be found by integration and is as follows

$$f_X(x) = \begin{cases} 0 & x < 0 \\ 0.5x & 0 \leq x < 0.5 \\ 0.75 & \text{x} = 0.5 \\ 0.75 + 0.5(x - 0.5) & 0.5 < x \leq 1 \\ 1 & 1 < x \end{cases}$$

6. Question
Let X be the # of tosses till first H. Then, $(X = 1) \cap (X > 1) = \phi$. Using *Total Expectation Theorem*

$$\begin{aligned} E[X] &= P(X = 1)E[X|X = 1] + P(X > 1)E[X|X > 1] \\ &= 0.5 * 1 + 0.5E[X] \\ \Rightarrow E[X] &= 2 \end{aligned}$$

$P(X = 1) = 0.5$ because then we get the head in the first toss itself. Since $P(X = 1) + P(X > 1) = 1$, we have $P(X > 1) = 0.5$. $E[X] = E[X|X > 1]$ because the tosses are *independent* and thus memoryless.

7. Question
Note that $E[X|Y]$ is a function of y.

$$\begin{aligned} E[E[X|Y]] &= \sum_y E[X|Y] p_Y(y) \\ &= \sum_y \sum_x x p_{X|Y} p_Y \\ &= \sum_y \sum_x x p_{X,Y}(x, y) \\ &= \sum_x x \sum_y p_{X,Y}(x, y) \\ &= \sum_x x p_X(x) \\ &= E[X] \end{aligned}$$

8. Question
Note that $E[Z|X,Y]$ will be a function of both X and Y.

$$\begin{aligned}
E[Z|X,Y] &= \sum_z z p_{Z|X,Y}(z|x,y) \\
E[E[Z|X,Y]|X] &= \sum_y E[Z|X,Y] p_{X,Y|X}(x,y|x) \\
&= \sum_y \sum_z z p_{Z|X,Y}(z|x,y) p_{Y|X}(y|x) \\
&= \sum_y \sum_z z \frac{p_{X,Y,Z}(x,y,z)}{p_X(x)} \\
&= \sum_z z \sum_y \frac{p_{X,Y,Z}(x,y,z)}{p_X(x)} \\
&= \sum_z z \frac{p_{X,Z}(x,z)}{p_X(x)} \\
&= \sum_z z p_{Z|X}(z|x) \\
&= E[Z|X]
\end{aligned}$$

9. Question
We use the formulae from iterated expectation to calculate these.

$$\begin{aligned}
P_Y(y) &= \begin{cases} \frac{1}{3} & y = 1 \\ \frac{2}{3} & y = 2 \end{cases} \\
E[X] &= E[E[X|Y]] = \sum_y E[X|Y] P(Y) \\
&= 90 * \frac{1}{3} + 60 * \frac{2}{3} \\
Var(X) &= E[Var(X|Y)] + Var(E[X|Y]) \\
&= \sum_y Var(X|Y) P(Y) + ((90 - E[E[X|Y]])^2 \frac{1}{3} + (60 - E[E[X|Y]])^2 \frac{2}{3}) \\
&= \frac{650}{3}
\end{aligned}$$

10. Question
Let X denote the number of people who pick their own hat. We have been asked $E[X]$. Let X_i be a binary random variable denoting whether the i^{th} person picked their own hat, i.e.,

$$\begin{aligned}
X_i &= \begin{cases} 1 & \text{if } i^{th} \text{ person picks their own hat} \\ 0 & \text{otherwise} \end{cases} \\
P(X_i = 1) &= \frac{1}{n} \\
E[X_i] &= 1 * \frac{1}{n} + 0 * (1 - \frac{1}{n}) = \frac{1}{n}
\end{aligned}$$

Consequently

$$E[X] = E[\sum_{i=1}^{n} X_i] = \sum_{n=1}^{n} E[X_i] = 1$$

It is interesting to see the variance of X. Note that the formula for variance is $E[X^2] - E[X]^2$. Thus,

$$X^2 = (\sum_{i=1}^{n} X_i)^2 = \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} X_i X_j$$

$$E[X^2] = \sum_{i=1}^{n} E[X_i^2] + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} E[X_i X_j]$$

Note that X_i and X_j are not independent since after the first person has picked the hat, only $n-1$ hats remain

$$X_i X_j = \begin{cases} 1 & \text{if } X_i = X_j = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$P(X_i X_j = 1) = P(X_i = 1) P(X_j = 1 | X_i = 1) = \frac{1}{n} * \frac{1}{n-1}$$

$$E[X_i X_j] = 1 * (\frac{1}{n} * \frac{1}{n-1}) + 0 * (1 - \frac{1}{n} * \frac{1}{n-1}) = \frac{1}{n(n-1)}$$

$$E[X_i^2] = 1^2 \frac{1}{n} + 0^2 (1 - \frac{1}{n}) = \frac{1}{n}$$

Putting these values in the original equation for variance

$$E[X_2] = n\frac{1}{n} + \frac{1}{n}\frac{1}{n-1}(\frac{n(n-1)}{2} * 2) = 2$$

$$Var(X) = 2 - 1^2 = 1$$

11. Question

The following is the joint probability distribution of X and Y

$$f_{XY}(x, y) = f_X(x) f_{Y|X}(y|x) = \frac{1}{l}\frac{1}{x} = \frac{1}{xl} \; \forall \, 0 \leq y \leq x \leq 1$$

Using marginal probabilities, we can calculate $f_Y(y) and E[Y] as$

$$f_Y(y) = \int f_{XY}(x, y) dx = \int_y^l \frac{1}{xl} dx = \frac{1}{l} \log \frac{l}{y}$$

Note that for any y, $y \leq x \leq l$

$$E[Y] = \int y f_Y(y) = \int_0^l l y \frac{1}{l} \log \frac{l}{y} = \frac{l}{4}$$

This problem can also be approched using iterated expectation

$$E[Y] = E[E[Y|X]] = E[\text{uniform random variable between 0 and } x]$$
$$= E[\frac{X}{2}] = \frac{1}{2} E[X]$$
$$= \frac{l}{4}$$

12. Question

Note that the required probability distribution is given by the following formula

$$p_{X+Y}(x) = \int_{-\inf}^{\inf} p_X(x-y) * p_Y(y)dy$$

However, note that the exponential distribution is not positive everywhere. For values < 0, the probability density is 0. Hence, we break the integral into three parts as follows

$$p_{X+Y}(x) = \int_{-\inf}^{0} p_X(x-y) * p_Y(y)dy + \int_{0}^{x} p_X(x-y) * p_Y(y)dy + \int_{x}^{\inf} p_X(x-y) * p_Y(y)dy$$

Carefully note that for y in range $(-\inf, 0]$, $p_Y(y) = 0$, and in the range $[x, \inf)$, $x - y < 0$, which implies $p_X(x) = 0$. Hence,

$$\begin{aligned} p_{X+Y}(x) &= \int_{0}^{x} p_X(x-y) * p_Y(y)dy \\ &= \lambda\mu \exp(-\lambda x) \int_{0}^{x} \exp((\lambda-\mu)y)dy \\ &= \frac{\lambda\mu}{\lambda-\mu} \exp(-\lambda x)(\exp((\lambda-\mu)x) - 1) \\ &= \frac{\lambda\mu}{\lambda-\mu}(\exp(\mu x) - \exp(-\lambda x)) \end{aligned}$$

13. Question

Assume that we break the stick at points X and Y. Assume $X < Y$. Then for the stick to form a triangle, the three lengths $X, Y - X$ and $1 - Y$ should satisfy the following three inequalities

$$\begin{aligned} X + (Y - X) &> 1 - Y \\ (Y - X) + (1 - Y) &> X \\ X + (1 - Y) &> Y - X \end{aligned}$$

which is nothing but the triangluar region between the points $(0, 0.5), (0.5, 0.5)$ and $(0.5, 1)$ and has the area of $1/8$. We should also consider the case $Y < X$ and by symmetry, the area is same. Now, X and Y comprise of the entire square region $X \leq 1$ and $Y \leq 1$. Hence the required probability is $2 * 1/8 = 1/4$.

14. Question

Always solve such questions using the cumulative distribution approach.

$$P(X \leq x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2}x & 0 \leq x \leq 2 \\ 1 & 2 < x \end{cases}$$

$$\begin{aligned} P(Y \leq y) = P(X^3 \leq y) &= P(X \leq y^{\frac{1}{3}}) \\ &= \begin{cases} 0 & y < 0 \\ \frac{1}{2}y^{\frac{1}{3}} & 0 \leq y^{\frac{1}{3}} \leq 2 \\ 1 & 2 < y^{\frac{1}{3}} \end{cases} \end{aligned}$$

$$\begin{aligned} f_Y(y) &= \frac{dP(Y <= y)}{dy}(y) \\ &= \begin{cases} 0 & y < 0 \\ \frac{1}{6}y^{\frac{-2}{3}} & 0 \leq y \leq 8 \\ 0 & 8 < y \end{cases} \end{aligned}$$

15. Question
Let X be the waiting time and $F_X(x)$ be the CDF. Then,

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{2}{3} & x = 0 \\ \frac{2}{3} + \frac{1}{30}x & 0 < x < 5 \\ 1 & 5 \leq x \end{cases}$$

The PDF is simply the derivate of the CDF. Thus, expectation is

$$E[X] = \frac{2}{3}(0) + \int_0^5 \frac{1}{30} x dx + \frac{1}{6}(5) = \frac{5}{4} mins$$

16. Question
From Bayes' theorem

$$\begin{aligned} f_{Q|X}(q|x) &= \frac{f_{X|Q}(x|q) f_Q(q)}{f_X(x)} \\ &= \frac{f_{X|Q}(x|q) f_Q(q)}{\int_0^1 f_{X|Q}(x|q) f_Q(q) dq} \end{aligned}$$

We will need to solve separately for $x = 0$ and $x = 1$ as x is discrete.

$$f_{Q|X=0}(q|x=0) = \frac{(1-q) * 6q(1-q)}{\int_0^1 (1-q) * 6q(1-q) dq} = 12q(1-q)^2$$
$$f_{Q|X=1}(q|x=1) = \frac{q * 6q(1-q)}{\int_0^1 q * 6q(1-q) dq} = 12q^2(1-q)$$

17. Question
Questions of this type must only be approached through CDF. First find the CDF of Y and then it's PDF.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g(X) <= y) \\ &= P(X \in [-y, 0] \cup X \in [0, y^2]) \\ &= (F_X(0) - F_X(-y)) + (F_X(y^2) - F_X(0)) \\ &= F_X(y^2) - F_X(-y) \\ p_Y(y) &= \frac{F_Y(y)}{dy} \\ &= \frac{dF_X(y^2)}{dx} \frac{d(y^2)}{dy} - \frac{dF_X(-y)}{dx} \frac{d(-y)}{dy} \\ &= 2y p_X(y^2) + p_X(-y) \\ &= 2y \frac{1}{\sqrt{2\pi}} e^{-\frac{y^4}{2}} + \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \end{aligned}$$

18. Question

(a) $P(X = k) = \binom{10}{k} 0.2^k 0.8^{10-k}$

(b) $P(no\ hits) = 0.8^{10}$

(c) $P(X >= 6)\ \sum_{k=6}^{10} \binom{10}{k} 0.2^k 0.8^{10-k}$

(d) $E[X] = np = 2, Var(X) = np(1-p) = 1.6$ for Bernoulli distribution

(e) $Y = 2X - 3, E[Y] = 2E[X] - 3 = 1, Var(Y) = 4Var(X) = 6.4$

(f) $Z = X^2, E[Z] = E[X^2] = Var(X) + E[X]^2 = 5.6$

19. Question

For the mosquito, $P(bite) = P(land)P(bite|land) = 0.1$. X is a geometric random variable. $E[X] = 1/p = 10$ and $Var(X) = \frac{1-p}{p^2} = 90$.
For the mosquito and tick combined, $P(\text{mosquito and tick}) = 0.1 + 0.1 * 0.7 - 0.1 * 0.1 * 0.7 = 0.163$. This is again a geometric random variable with $E[Y] = 1/0.163$ and $Var(Y) = (1 - 0.163)/(0.163^2)$.s

20. Question

This quantity can be calculated using the law of total expectation

$$E[X] = E[X|A_1]P(A_1) + E[X|A_2]P(A_2) + \cdots + E[X|A_n]P(A_n) \quad \text{where } A_i \text{ are disjoint}$$

Let H_1 denote heads at first toss, H_2 denote heads at the second toss, T_1 denote tails at first toss and T_2 denote tails at the second toss. Then,

$$\begin{aligned} E[X] &= E[X|H_1]P(H_1) + E[X|T_1]P(T_1) \\ E[X|H_1] &= E[X|H_1H_2]P(H_2|H_1) + E[X|H_1T_2]P(T_2|H_1) \\ &= 2p + (1 + E[X|T_1])(1-p) \\ E[X|T_1] &= E[X|T_1T_2]P(T_2|T_1) + E[X|T_1H_2]P(H_2|T_1) \\ &= 2(1-p) + (1 + E[X|H_1])p \end{aligned}$$

$E[X|H_1T_2] = 1 + E[X|T_1]$ because the tails after the first heads implies the first heads is now irrelevant and we have wasted one toss on the heads. The remaining process is same as starting from the first coin toss as tails.
Solving for the conditional expectations,

$$\begin{aligned} E[X|H_1] &= \frac{3 - 2p + p^2}{1 - p + p^2} \\ E[X|T_1] &= \frac{2 + p^2}{1 - p + p^2} \\ E[X] &= \frac{2 + p - p^2}{1 - p + p^2} \end{aligned}$$

21. Question

Define X as the following random variable

$$X = \begin{cases} 1, p = \frac{1}{4} & HHH \text{ or } TTT \\ 0, p = \frac{3}{4} & \text{otherwise} \end{cases}$$

(a) K is simply a binomial distribution, where we want the 2^{nd} success to happen at the $K+1$th trial.

$$p_K(k) = \binom{k}{1}\frac{1}{4}^2\frac{3}{4}^{k-1} \qquad \text{since the last trial is success}$$

(b) M = number of tails before first success. Let the success be at $N+1$. Defin Y as

$$Y = \begin{cases} 1 & p=\frac{1}{2} \quad HHT,\ HTH, \text{ or } THH \\ 2 & p=\frac{1}{2} \quad HTT,\ THT, \text{ or } TTH \end{cases}$$

$$E[Y] = 1*\frac{1}{2} + 2*\frac{1}{2}$$

$$Var(Y) = (1-\frac{3}{2})^2*\frac{1}{2} + (2-\frac{3}{2})^2*\frac{1}{2}$$

$$E[N+1] = \frac{1}{p} = 4$$

$$Var(N+1) = Var(N) = \frac{1-p}{p^2} = \frac{1-\frac{1}{4}}{\frac{1}{4}^2}$$

$$M = Y_1 + Y_2 + \cdots Y_N$$

$$E[M] = E[Y_1 + Y_2 + \cdots Y_N]$$

$$Var(M) = Var(Y_1 + Y_2 + \cdots Y_N)$$

Note that both Y and N are random variables here. Using the formulae for random number of random variables,

$$E[M] = E[E[M|N]] = E[NE[Y]] = E[N]E[Y] = (4-1)*\frac{3}{2} = \frac{9}{2}$$

$$Var(M) = Var(E[M|N]) + E[Var(M|N)] = Var(NE[Y]) + E[NVar(Y)]$$

$$= E[Y]^2 Var(N) + E[N]Var(Y) = \frac{9}{4}*12 + 3*\frac{1}{4} = \frac{111}{4}$$

22. Question

Let X be the number of tosses till the first coin is removed. This is a geometric random variable with $P(\text{success}) = \frac{1}{8}$. then $E[X] = 1/p = 8$. Now Y be the number of tosses till the second coin is removed (counting tosses after removal of first coin). Note that geometric random variables are memory less and what happened before the start of the "experiment" will not matter. Thus, $E[Y] = 1/(1/4) = 4$. Similarly, Z is the tosses till the last coin is removed and $E[Z] = 1/(1/2) = 2$. Note that the number of tosses till the end of experiment is simply $X+Y+Z$. $E[X+Y+Z] = E[X] + E[Y] + E[Z] = 14$.

23. Question

Note that the process till the end is a combination of multiple binomial process, such that any process lasts till the first success. Suppose we sign a paper and keep this in the drawer. Now the total signed papers in the drawer is k out of n and the $P(\text{success}) = \frac{n-k}{n}$ and $E[\text{draws till next unsigned paper}] = \frac{1}{p} = \frac{n}{n-k}$. Total draws

$$E = \frac{n}{1} + \frac{n}{2} + \cdots + \frac{n}{n}$$

$$= n(1 + \frac{1}{2} + \cdots + \frac{1}{n})$$

$$\lim_{n \to large} E = n\log(n)$$

24. Question
A very straightforward way is to use a triple integral

$$P(X < Y < Z) = \int_0^{\inf} \int_0^z \int_0^y \lambda e^{-\lambda x} \mu e^{-\mu y} \nu e^{-\nu z} dx dy dz = \frac{\lambda\mu}{(\lambda+\mu+\nu)(\mu+\nu)}$$

$P(X < Y < Z)$ can be broken down as $P(X < min(Y,Z))P(Y < Z)$.
Consider just $P(Y < Z)$

$$P(Y < Z) = \int_0^{\inf} \int_0^z \mu e^{-\mu y} \nu e^{-\nu z} dy dz = \frac{\mu}{\mu+\nu}$$

Thus, when two exponential processes are considered, probaility of arrival of 1st before 2nd is simply the percentage ratio of parameters. Thus,

$$\begin{aligned} P(X < min(Y,Z)) &= \frac{\lambda}{\lambda + (\mu+\nu)} \qquad Y \text{ and } Z \text{ can be combined as a single process} \\ P(Y < Z) &= \frac{\mu}{\mu+\nu} \\ P(X < Y < Z) &= P(X < min(Y,Z))P(Y < Z) \\ &= \frac{\lambda\mu}{(\lambda+\mu+\nu)(\mu+\nu)} \end{aligned}$$

25. Question We can model the arrival process like a Poisson process. $\lambda = 5$ and $\tau = \frac{1}{2}$

$$\begin{aligned} P(\lambda, \tau, k) &= \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!} \\ P(5, \frac{1}{2}, 0) &= \frac{(5 * \frac{1}{2})^0 e^{-5*\frac{1}{2}}}{0!} \\ P(5, \frac{1}{2}, 1) &= \frac{(5 * \frac{1}{2})^1 e^{-5*\frac{1}{2}}}{1!} \end{aligned}$$

26. Question

- P(fish for > 2 hours) $= P(k = 0, \tau = 2) = e^{-0.6*2}$
- P(fish for > 2 but < 5 hours) = P(first catch in $[2, 5]$ hours) $= P(k = 0, \tau = 2)(1 - P(k = 0, \tau = 3)$ which is no fish in $[0, 2]$ but at least 1 fish in the next 3 hours (which will be independent of first 2 hours)
- P(catch at least two fish) = P(at least 2 catches before 2 hours) $= 1 - P(k = 0, \tau = 2) - P(k = 1, \tau = 2)$
- E[fish] has two possibilities, either single fish after 2 hours, or many fist before 2 hours. $E[fish] = E[fish|\tau \leq 2](1 - P(\tau > 2)) + E[fish|\tau > 2]P(\tau > 2) = (0.6 * 2) * (1 - P(k = 0, \tau = 2)) + 1 * P(k = 0, \tau = 2)$
- E[Total fishing time] $= 2 + P(k = 0, \tau = 2)\frac{1}{\lambda}$, since we fish for atlest 2 hours
- E[future fishing time—fished for two hours] can be obtained using the memoryless property of Poisson process. The expected time till first arrival is independent of what has happened till now. Thus, $E[T_1] = \frac{1}{\lambda}$

27. Question

Start with the merged Poisson process which will denote the time till the first bulb will fail. For this process, $\lambda^{'} = 3\lambda$. Hence, $E[\text{first bulb fails}] = \frac{1}{3\lambda}$. After the first bulb dies out, we are left with a process with $\lambda^{'} = 3\lambda$. Due to memoryless property, $E[\text{second bulb fails}] = \frac{1}{2\lambda}$ and consequently $E[\text{last bulb fails}] = \frac{1}{\lambda}$.

Note the above two times denote the time difference, i.e. the time taken for the bulb to die out after the last bulb died out. Thus, $E[\text{time until last bulb dies out}] = \frac{1}{3\lambda} + \frac{1}{2\lambda} + \frac{1}{\lambda}$

28. Question

(a)

$$\begin{aligned} E[\text{time till failure}] &= E[\text{time till failure}|A]P(A) + E[\text{time till failure}|B]P(B) \\ &= \frac{1}{\lambda_A}\frac{1}{2} + \frac{1}{\lambda_B}\frac{1}{2} \\ &= \frac{1}{2}(1 + \frac{1}{3}) = \frac{2}{3} \end{aligned}$$

(b) Let C denote the event of no failure till time t. $P(C)$ for a given λ will be $\int_t^{\text{inf}} \lambda e^{-\lambda t}$. Then,

$$\begin{aligned} P(C) &= P(C|A)P(A) + P(C|B)P(B) \qquad \text{Using total probability theorem} \\ &= e^{-t}(\frac{1}{2}) + e^{-3t}(\frac{1}{2}) \\ &= \frac{1}{2}(e^{-t} + e^{-3t}) \end{aligned}$$

(c) Let C denote the event of no failure till time t. Then,

$$\begin{aligned} P(A|C) &= \frac{P(C|A)P(A)}{P(C)} \\ &= \frac{P(C|A)P(A)}{P(C|A)P(A) + P(C|B)P(B)} \\ &= \frac{\frac{1}{2}e^{-t}}{\frac{1}{2}(e^{-t} + e^{-3t})} \\ &= \frac{1}{1 + e^{-2t}} \end{aligned}$$

(d) Let T_{B1}, T_{B2} and T_A denote the life times of the first B bulb, second B bulb and the A bulb respectively. First consider the solution to $P(T_{B1} + T_{B2} = t)$

$$\begin{aligned} P(T_{B1} + T_{B2} = t) &= \int_0^t P(T_{B1} = t_1)P(T_{B2} = t - t_1)dt_1 \qquad \text{Using independence} \\ &= \int_0^t 3e^{-3t_1}3e^{-3(t-t_1)}dt_1 \\ &= \int_0^t 9e^{-3t}dt_1 \\ &= 9te^{-3t} \end{aligned}$$

Now, we can rewrite the requred probability in a slightly different format

$$\begin{aligned}
P(T_{B1} + T_{B2} > T_A) &= P(T_{B1} + T_{B2} = t)P(T_A \leq t) \\
&= \int_0^{\inf} 9te^{-3t}(\int_0^t e^{-t_1}dt_1)dt \\
&= \int_0^{\inf} 9te^{-3t}(1 - e^{-t})dt \\
&= \int_0^{\inf} 9te^{-3t} - 9te^{-4t}dt
\end{aligned}$$

Using integration by parts, $\int uv' = uv - \int u'v$ and choosing $u = t, v = e^{-3t}/3$,

$$\begin{aligned}
P(T_{B1} + T_{B2} > T_A) &= \left[9[te^{-3t}]_0^{\inf} - 3\int_0^{\inf} e^{-3t}dt - 9[te^{-4t}]_0^{\inf} + \frac{9}{4}\int_0^{\inf} e^{-4t}dt\right] \\
&= 0 + 1 - 0 - \frac{9}{16} = \frac{7}{16}
\end{aligned}$$

(e) Let there be N bulbs of type B out of the 12 bulbs. Clearly N is a random variable and can be seen as the "successes" of choosing a given bulb as B. and the probability of choosing any ith bulb as B is $1/2$.
Let the life time of any bulb of type B be T. Then the total lifetime of all the type B bulbs will be NT, which is nothing but the sum of a random number of random variables.

$$\begin{aligned}
E[NT] &= E[N]E[T] = np * \frac{1}{\lambda} = 12 * \frac{1}{2} * \frac{1}{3} = 2 \\
Var(NT) &= E[Var(NT|N)] + Var(E[NT|N]) = E[N]Var(T) + E[T]^2Var(N) \\
&= np * \frac{1}{\lambda^2} + (\frac{1}{\lambda})^2 np(1-p) = 1
\end{aligned}$$

(f) Let D be the event that the lifetime is greater thatn t or $T > t$. Then,

$$\begin{aligned}
E[T|D] &= E[T|D, A]P(A|D) + E[T|D, B]P(B|D) \\
&= t + (E[T - t|D, A]P(A|D) + E[T - t|D, B]P(B|D)) \\
&= t + (\frac{1}{1}P(A|D) + \frac{1}{3}P(B|D)) && \text{Using memoryless property} \\
&= t + (\frac{1}{1 + e^{-2t}} + \frac{1}{3}(1 - \frac{1}{1 + e^{-2t}})) && \text{Using part 28c} \\
&= t + \frac{1}{3} + \frac{2}{3}\frac{1}{1 + e^{-2t}}
\end{aligned}$$

29. Question
Let X_i denote the random variables with respective parameters λ_i.
Note that

$$\begin{aligned}
P(min(X_1, X_2, \ldots, X_n) > x) &= P(X_1 > x, X_2 > x, \ldots, X_n > x) \\
&= \prod_{i=1}^{n}(1 - P(X \leq x)) \\
&= \prod_{i=1}^{n} e^{-\lambda_i x} = exp(-\sum_{i=1}^{n}\lambda_i x)
\end{aligned}$$

which is an exponential random variable with the rates added together.

30. Question

Using balance equations, we have

$$\begin{aligned}\pi_1 &= \pi_1 p_{11} + \pi_2 p_{21}\\ \pi_2 &= \pi_1 p_{12} + \pi_2 p_{22}\\ \pi_1 + \pi_2 &= 1\end{aligned}$$

Solving, $\pi_1 = \frac{2}{7}$ and $\pi_2 = \frac{5}{7}$

31. Question

Let a_i denote the abosorption probabilites into state 4 starting from i

$$\begin{aligned}a_5 &= 0, a4 = 1\\ a_i &= \sum_j a_j p_{ij}\\ a_2 &= a_1 p_{21} + a_4 p_{24}\\ a_3 &= a_1 p_{31} + a_2 p_{32} + a_5 p_{35}\\ a_1 &= a_2 p_{12} + a_3 p_{13}\end{aligned}$$

Solving, $a_1 = \frac{9}{14}, a_2 = \frac{5}{7}$ and $a_3 = \frac{15}{28}$

Let μ_i denote the expected time till absorption starting from i, then

$$\begin{aligned}\mu_4 &= 0\\ \mu_1 &= 1 + \mu_2 p_{12} + \mu_3 p_{13}\\ \mu_2 &= 1 + \mu_1 p_{21} + \mu_4 p_{24}\\ \mu_3 &= 1 + \mu_1 p_{31} + \mu_2 p_{32}\end{aligned}$$

Solving, $\mu_1 = \frac{55}{4}, \mu_2 = 12$ and $\mu_3 = \frac{111}{8}$

32. Question

(a) The probability of eventually leaving course 6 is 1 as states 15 and 9 are absorbing states.

(b) Here we have to calculate the probability of absortion into state 15. Let a_i denote the probability of absorption into state 15 from state i. Then, $a_{15} = 1$ and $a_9 = 0$. Using equations from 20.4,

$$\begin{aligned}a_{6-1} &= \frac{1}{2}a_{6-1} + \frac{1}{8}a_{6-2} + \frac{1}{8}a_{6-3} + \frac{1}{8}a_9 + \frac{1}{8}a_{15}\\ a_{6-2} &= \frac{1}{2}a_{15} + \frac{3}{8}a_{6-1} + \frac{1}{8}a_{6-3}\\ a_{6-3} &= \frac{1}{4}a_9 + \frac{3}{8}a_{6-1} + \frac{3}{8}a_{6-2}\end{aligned}$$

Solving the 3 equations, 3 variable system, $a_{6-1} = 105/184, a_{6-2} = 143/184$ and $a_{6-3} = 93/184$.

(c) Let μ_i denote the expected number of steps to get absorbed starting from state i. Then, $\mu_{15} = \mu_9 = 0$. Using equations from 20.4,

$$\begin{aligned}\mu_{6-1} &= 1 + \frac{1}{2}\mu_{6-1} + \frac{1}{8}\mu_{6-2} + \frac{1}{8}\mu_{6-3} + \frac{1}{8}\mu_9 + \frac{1}{8}\mu_{15}\\ \mu_{6-2} &= 1 + \frac{1}{2}\mu_{15} + \frac{3}{8}\mu_{6-1} + \frac{1}{8}\mu_{6-3}\\ \mu_{6-3} &= 1 + \frac{1}{4}\mu_9 + \frac{3}{8}\mu_{6-1} + \frac{3}{8}\mu_{6-2}\end{aligned}$$

Solving, $\mu_{6-1} = 81/23, \mu_{6-2} = 63/23$ and $\mu_{6-3} = 77/23$.

(d) This question can be done in a manner similar to the equations described above but with a small adjustment. Note that, we can either have 0, 1, or 2 ice creams. Consider $v_i(j)$ as the probability of making j additional ice creams from 6-2 to 6-1 or 6-3 to 6-1 transitions, given the current state is i. Note $v_{15}(0) = v_9(0) = 1$. Then,

$$\begin{aligned}
v_{6-1}(0) &= \frac{1}{2}v_{6-1}(0) + \frac{1}{8}v_{6-2}(0) + \frac{1}{8}v_{6-3}(0) + \frac{1}{8}v_9(0) + \frac{1}{8}v_{15}(0) \\
v_{6-2}(0) &= \frac{1}{2}v_{15}(0) + \frac{3}{8}(0) + \frac{1}{8}v_{6-3}(0) \\
v_{6-3}(0) &= \frac{1}{4}v_9(0) + \frac{3}{8}(0) + \frac{3}{8}v_{6-2}(0)
\end{aligned}$$

Some of the transitions have been directly replaced with 0 as we are considering 0 ice creams and thus those transitions are not possible (6-2 to 6-1 for instance). Solving, $v_{6-1}(0) = 46/61, v_{6-2}(0) = 34/61$ and $v_{6-3}(0) = 28/61$.

The same way, we can construct equations for 1 additional steps where $v_{15}(1) = v_9 = 0$.

$$\begin{aligned}
v_{6-1}(1) &= \frac{1}{2}v_{6-1}(1) + \frac{1}{8}v_{6-2}(1) + \frac{1}{8}v_{6-3}(1) + \frac{1}{8}v_9(1) + \frac{1}{8}v_{15}(1) \\
v_{6-2}(1) &= \frac{1}{2}v_{15}(1) + \frac{3}{8}v_{6-1}(0) + \frac{1}{8}v_{6-3}(1) \\
v_{6-3}(1) &= \frac{1}{4}v_9(1) + \frac{3}{8}v_{6-1}(0) + \frac{3}{8}v_{6-2}(1)
\end{aligned}$$

In the second equation, after going from 6-2 to 6-1, we can only get 0 more ice creams. Hence, some of the values have been replaced with the $v_i(0)$ calculated above. Solving, $v_{6-1}(1) = 690/3721, v_{6-2}(1) = 1242/3721$ and $v_{6-3}(1) = 1518/3721$.

Note that since the total ice creams are 0, 1, or 2, we have $v_{6-1}(0) + v_{6-1}(1) + v_{6-1}(2) = 1$. $E[\text{ice creams}] = 0 * v_{6-1}(0) + 1 * v_{6-1}(1) + 2 * v_{6-1}(2) = 1140/3721$

(e) We need to recalculate the the transition probabilities since we are conditioning on the event A that we land up in state 15.

$$\begin{aligned}
P_{ij|A} &= P(X_{n+1} = j | X_i = i, A) \\
&= \frac{P(X_{n+1} = j, X_n = i, A)}{P(X_n = i, A)} \\
&= \frac{P(A|X_{n+1} = j, X_n = i)P(X_{n+1} = j|X_n = i)P(X_n = i)}{P(A|X_n = i)P(X_n = i)} \\
&= \frac{P(A|X_{n+1} = j)P(X_{n+1} = j|X_n = i)}{P(A|X_n = i)} \\
&= \frac{a_j}{a_i}P_{ij}
\end{aligned}$$

where a_i is the probability of absorption into state 15 starting from state i. Since markov process is only dependent on the last state, absoprtion probabilities are not dependent on n.

We can write equations similar to 20.4 for calculating the expected number of steps

with the adjusted transition probabilities

$$\mu_{6-1} = 1 + \frac{a_{6-1}}{a_{6-1}}\frac{1}{2}\mu_{6-1} + \frac{a_{6-2}}{a_{6-1}}\frac{1}{8}\mu_{6-2} + \frac{a_{6-3}}{a_{6-1}}\frac{1}{8}\mu_{6-3} + \frac{a_{15}}{a_{6-1}}\frac{1}{8}\mu_{15} + \frac{a_9}{a_{6-1}}\frac{1}{8}\mu_9$$
$$\mu_{6-2} = 1 + \frac{a_{6-1}}{a_{6-2}}\frac{3}{8}\mu_{6-1} + \frac{a_{6-3}}{a_{6-2}}\frac{1}{8}\mu_{6-3} + \frac{a_{15}}{a_{6-2}}\frac{1}{2}\mu_{15}$$
$$\mu_{6-3} = 1 + \frac{a_{6-1}}{a_{6-3}}\frac{3}{8}\mu_{6-1} + \frac{a_{6-2}}{a_{6-3}}\frac{3}{8}\mu_{6-2} + \frac{a_9}{a_{6-3}}\frac{1}{4}\mu_9$$

where $\mu_{15} = \mu_9 = 0, a_{15} = 1$, and $a_9 = 0$. The absorption probabilities can be taken from the part 32b. Solving, $\mu_{6-1} = 1763/483$.

(f) The changed probabilites become $P(X_{n+1} = 15|X_n = 6-1) = P(X_{n+1} = 6-2|X_n = 6-1) = P(X_{n+1} = 6-3|X_n = 6-1) = 1/6, P(X_{n+1} = 6-1|X_n = 6-2) = 3/4$ and $P(X_{n+1} = 6-3|X_n = 6-2) = 1/4$. We then use equations from 20.4 to calculate the expected values

$$\mu_{6-1} = 1 + \frac{1}{2}\mu_{6-1} + \frac{1}{6}\mu_{6-2} + \frac{1}{6}\mu_{6-3} + \frac{1}{6}\mu_9$$
$$\mu_{6-2} = 1 + \frac{3}{4}\mu_{6-1} + \frac{1}{4}\mu_{6-3}$$
$$\mu_{6-3} = 1 + \frac{1}{4}\mu_9 + \frac{3}{8}\mu_{6-1} + \frac{3}{8}\mu_{6-2}$$

where $\mu_{15} = 0$. Solving, $\mu_{6-1} = 86/13, \mu_{6-2} = 98/13$ and $\mu_{6-3} = 82/13$.

(g) If we look carefully at the new probabilities, states 15 and 9 become recurrent. Far into the future, we are sure to land up in those states, and will be in either one of those. By symmetry, the two should be same. $\pi_{15} = \pi_9 = 1/2$.

(h) We assume that 6-1 is an absorbing state, and accordingly calculate the probabilities. Note that there will not be an equation for 6-1 since we are then already in the final state.

$$\mu_{6-2} = 1 + \frac{1}{8}\mu_{6-3} + \frac{1}{2}\mu_{15}$$
$$\mu_{6-3} = 1 + \frac{3}{8}\mu_{6-2} + \frac{1}{4}\mu_9$$
$$\mu_9 = 1 + \frac{7}{8}\mu_9$$
$$\mu_{15} = 1 + \frac{7}{8}\mu_{15}$$

Solving, $\mu_{6-2} = 344/61, \mu_{6-3} = 312/61$ and $\mu_9 = \mu_{15} = 8$. Plugging these into the following equation (which corresponds to taking one step out of 6-1),

$$\mu_{6-1} = 1 + \frac{1}{2}\mu_{6-1} + \frac{1}{8}\mu_{15} + \frac{1}{8}\mu_{6-2} + \frac{1}{8}\mu_{6-3} + \frac{1}{8}\mu_{15} = \frac{265}{61}$$

33. Question The exact answer will be

$$\sum_{k=0}^{21} \binom{36}{k} (\frac{1}{2})^{36} = 0.8785$$

But the same can be estimated using the CLT as follows

$$\mu = np = 18$$
$$\sigma^2 = np(1-p) = 9$$
$$P(S_n \leq 21) \approx P(\frac{S_n - 18}{3} \leq \frac{21-18}{3}) \approx 0.843$$

Our estimate is in the rough range of the answer but not quite close. We can do better using the $\frac{1}{2}$ correction

$$P(S_n \leq 21) = P(S_n < 22) \text{ since } S_n \text{ is an integer}$$
$$\text{Consider } P(S_n <= 21.5) \text{ as a compromise between the two}$$
$$P(S_n <= 21.5) = P(\frac{S_n - 18}{3} \leq \frac{21.5-18}{3}) \approx 0.879$$

In a similar manner, $P(S_n = 19) = P(18.5 \leq S_n \leq 19.5)$ using $\frac{1}{2}$ correction.

34. Question We know that for a normal population,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

In the given problem, $n = 15$ and $\sigma^2 = 9$. Hence,

$$P(S^2 > 12) = P(\frac{14S^2}{9} > \frac{14}{9}12)$$
$$= P(\chi^2_{14} > \frac{56}{3}) \qquad = 0.178 \text{ from standard tables or chi-square calculators}$$

35. Question
Since the observations are independent, the likelihood of all the observations under some θ is given by

$$p_{X|\Theta}(x|\theta) = \prod_{i=1}^{n} \theta \exp(-\theta x_i)$$
$$log(p_{X|\Theta}(x|\theta)) = nlog(\theta) - \theta(\sum_{i=1}^{n} x_i)$$

Taking the derivatie and maximizing with respect to θ, $\hat{\theta}_{MLE} = \frac{n}{\sum_{i=1}^{n} x_i}$

36. Question

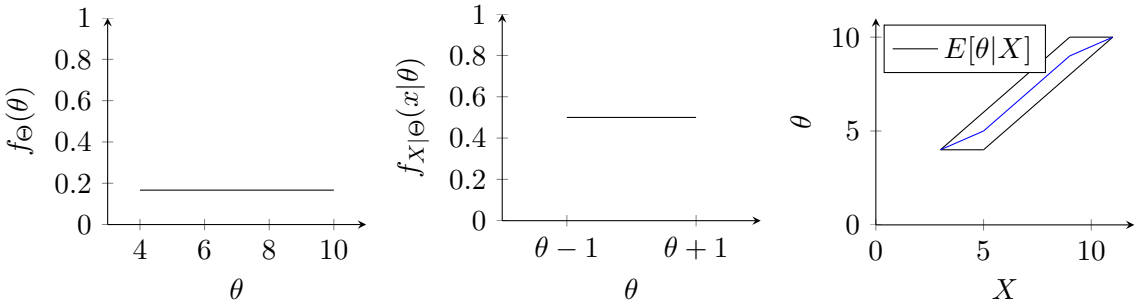
We first begin by calculating the posterior of θ given observations

$$
\begin{aligned}
f(\theta|X_1, X_2, \ldots, X_n) &= \frac{f(X_1, X_2, \ldots, X_n|\theta)p(\theta)}{\int f(X_1, X_2, \ldots, X_n|\theta)p(\theta)d\theta} \\
f(X_1, X_2, \ldots, X_n|\theta) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} exp(-\frac{\sum_{i=1}^n (X_i - \theta)^2}{2\sigma_0^2}) \\
p(\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(\theta - \mu)^2}{2\sigma^2}) \\
f(X_1, X_2, \ldots, X_n|\theta)p(\theta) &= \frac{1}{2\pi\sigma_0\sigma} exp(-\frac{\sum_{i=1}^n (X_i - \theta)^2}{2\sigma_0^2} - \frac{(\theta - \mu)^2}{2\sigma^2}) \\
\frac{\sum_{i=1}^n (X_i - \theta)^2}{2\sigma_0^2} + \frac{(\theta - \mu)^2}{2\sigma^2} &= \frac{\theta^2(\sigma_0^2 + n\sigma^2) - 2\theta(\sigma_0^2\mu + \sigma^2 \sum_{i=1}^n X_i) + (\sigma_0^2\mu^2 + \sigma^2 \sum_{i=1}^n X_i^2)}{2\sigma_0^2\sigma^2} \\
&= \frac{\left(\theta - \frac{\sigma_0^2\mu + \sigma^2 \sum_{i=1}^n X_i}{\sigma_0^2 + n\sigma^2}\right)^2 + \left(\frac{\sigma_0^2\mu^2 + \sigma^2 \sum_{i=1}^n X_i^2}{\sigma_0^2 + n\sigma^2} - \left(\frac{\sigma_0^2\mu + \sigma^2 \sum_{i=1}^n X_i}{\sigma_0^2 + n\sigma^2}\right)^2\right)}{(2\sigma_0^2\sigma^2)/(\sigma_0^2 + n\sigma^2)} \\
f(X_1, X_2, \ldots, X_n|\theta)p(\theta) &= \left(\frac{1}{2\sigma_0\sigma} exp\left(\frac{\left(\frac{\sigma_0^2\mu^2 + \sigma^2 \sum_{i=1}^n X_i^2}{\sigma_0^2 + n\sigma^2} - \left(\frac{\sigma_0^2\mu + \sigma^2 \sum_{i=1}^n X_i}{\sigma_0^2 + n\sigma^2}\right)^2\right)}{2\sigma_0^2\sigma^2)/(\sigma_0^2 + n\sigma^2)}\right) \times \right. \\
&\quad \left. \left(\frac{1}{\sqrt{2\pi(2\sigma_0^2\sigma^2)/(\sigma_0^2 + n\sigma^2)}} exp\left(-\frac{\left(\theta - \frac{\sigma_0^2\mu + \sigma^2 \sum_{i=1}^n X_i}{\sigma_0^2 + n\sigma^2}\right)^2}{2\sigma_0^2\sigma^2/(\sigma_0^2 + n\sigma^2)}\right)\right)\right. \\
&= C \times \mathcal{N}\left(\frac{\sigma_0^2\mu + \sigma^2 \sum_{i=1}^n X_i}{\sigma_0^2 + n\sigma^2}, \frac{2\sigma_0^2\sigma^2}{\sigma_0^2 + n\sigma^2}\right) \\
f(\theta|X_1, X_2, \ldots, X_n) &= \frac{C \times \mathcal{N}\left(\frac{\sigma_0^2\mu + \sigma^2 \sum_{i=1}^n X_i}{\sigma_0^2 + n\sigma^2}, \frac{2\sigma_0^2\sigma^2}{\sigma_0^2 + n\sigma^2}\right)}{C \times \int_{-\inf}^{\inf} \mathcal{N}\left(\frac{\sigma_0^2\mu + \sigma^2 \sum_{i=1}^n X_i}{\sigma_0^2 + n\sigma^2}, \frac{2\sigma_0^2\sigma^2}{\sigma_0^2 + n\sigma^2}\right) d\theta} \\
&= \mathcal{N}\left(\frac{\sigma_0^2\mu + \sigma^2 \sum_{i=1}^n X_i}{\sigma_0^2 + n\sigma^2}, \frac{2\sigma_0^2\sigma^2}{\sigma_0^2 + n\sigma^2}\right) \\
&= \mathcal{N}\left(\frac{\sigma_0^2\mu + n\sigma^2\overline{X}}{\sigma_0^2 + n\sigma^2}, \frac{2\sigma_0^2\sigma^2}{\sigma_0^2 + n\sigma^2}\right)
\end{aligned}
$$

since the integral of a normal over the entire domain is equal to the total probability which is 1. Thus, the new estimate of the mean is the weighted sum of the prior mean and the estimated mean from the observations, with the weights being proportional to the inverse of the standard deviations. The new standard deviation is also the geometric mean of the original standard deviations.

37. Question

We need to evaluate $f_{\Theta|X}(\theta|x)$ in order to get $E[\Theta|X]$.
$f_{X,\Theta}(x, \theta) = f_X(x)f_{\Theta|X}(\theta|x)$ which is a parallelogram on the $\theta - x$ plane at the points $(3, 4)$, $(5, 4)$, $(9, 10)$ and $(11, 10)$. Then $E[\Theta|X]$ can be obtained by drawing vertical lines on the planes and calculating the $E[\theta]$ over that line. It is a line which bends at two points.



38. Question

(a) No, since X_i is also uniform in $[-1, 1]$

(b) Yes, $E[Y_i] = 0$ by symmetry. For $\epsilon > 0$,

$$\begin{aligned}\lim_{i\to\text{inf}}(P|Y_i - \mu_i| > \epsilon) &= \lim_{i\to\text{inf}} P(|\frac{X_i}{i} - 0| > \epsilon) \\ &= \lim_{i\to\text{inf}} P(\frac{X_i}{i} > \epsilon \text{ and } \frac{X_i}{i} < -\epsilon) \\ &= \lim_{i\to\text{inf}}[P(X_i > i\epsilon) + P(X_i < -i\epsilon)] = 0\end{aligned}$$

(c) Yes, $E[Y_i] = 0$ by symmetry. For $\epsilon > 0$,

$$\begin{aligned}\lim_{i\to\text{inf}} P(|Z_i - 0| > \epsilon) &= \lim_{i\to\text{inf}} P((X_i)^i > \epsilon \text{ or } (X_i)^i < -\epsilon) \\ &= \lim_{i\to\text{inf}}[\frac{1}{2}(1 - \epsilon^{1/i}) + \frac{1}{2}(1 - \epsilon^{1/i})] \\ &= \lim_{i\to\text{inf}}(1 - \epsilon^{1/i}) = 0\end{aligned}$$

39. Question

Based on the definitions in section 26, death rate is simply the hazard function, i.e., $\lambda_s = 2\lambda_n$ or death rate in smokers is twice that in non smokers. Now,

$$\begin{aligned}P(t > B|t > A) &= \frac{P(t > B, t > A)}{P(t > A)} &&= \frac{P(t > B)}{P(t > A)} \\ &= \frac{1 - F(B)}{1 - F(A)} &&= \frac{exp(-\int_0^B \lambda(t)dt)}{exp(-\int_0^A \lambda(t)dt)} \\ &= exp(-\int_A^B \lambda(t)dt) \\ P_s(t > B|t > A) &= exp(-\int_A^B \lambda_s(t)dt) &&= exp(-\int_A^B 2\lambda_n(t)dt) \\ &= exp(-\int_A^B \lambda_n(t)dt)^2 &&= P_n(t > B|t > A)^2\end{aligned}$$

or, the conditional probability of survival till an age for a smoker is sqaure that of a non smoker (note that probability < 1).