

Chapter 2

A Quick Review of Basic Probability and Statistics

This course presumes knowledge of Chapters 1 to 3 of “Introduction to Probability Models” by Sheldon M. Ross. This material is also largely covered in the course text by P. Bremaud.

2.1 Probability: The Basics

Ω : sample space
 $\omega \in \Omega$: sample outcome
 $A \subseteq \Omega$: event
 $X : \Omega \rightarrow S$: “S-valued random variable”
 P : a probability (distribution / measure) on Ω

A probability has the following properties:

1. $0 \leq P\{A\} \leq 1$ for each event A .
2. $P\{\Omega\} = 1$
3. for each sequence A_1, A_2, \dots of mutually disjoint events

$$P\left\{\bigcup_{i=1}^{\infty} A_i\right\} = \sum_{i=1}^{\infty} P\{A_i\}$$

2.2 Conditional Probability

The conditional probability of A, given B, written as $P\{A|B\}$, is defined to be

$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}}.$$

It is a probability on the new sample space $\Omega_B \subset \Omega$; $P\{A|B\}$ is interpreted as the likelihood / probability that A occurs given knowledge that B has occurred.

Conditional probability is fundamental to stochastic modeling. In particular in modeling “causality” in a stochastic setting, a causal connection between B and A means:

$$P\{A|B\} \geq P\{A\}.$$

2.3 Independence

Two events A and B are independent of one another if

$$P\{A|B\} = P\{A\}$$

i.e. $P\{A \cap B\} = P\{A\}P\{B\}$. Knowledge of B 's occurrence has no effect on the likelihood that A will occur.

2.4 Discrete Random Variables

Given a discrete random variable (rv) X which takes on values in $S = \{x_1, x_2, \dots\}$, its probability mass function is defined by:

$$P_X(x_i) = P\{X = x_i\}, \quad i \geq 1.$$

Given a collection X_1, X_2, \dots, X_n of S -valued rvs, its joint probability mass function (pmf) is defined as

$$P_{(X_1, X_2, \dots, X_n)}(x_1, x_2, \dots, x_n) = P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}.$$

The conditional pmf of X given $Y = y$ is then given by

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

The collection of rvs X_1, X_2, \dots, X_n are *independent* if

$$P_{(X_1, X_2, \dots, X_n)}(x_1, x_2, \dots, x_n) = P_{X_1}(x_1) \cdot P_{X_2}(x_2) \cdots P_{X_n}(x_n)$$

for all $(x_1, \dots, x_n) \in S^n$.

2.5 Continuous Random Variables

Given a continuous rv X taking values in \mathbb{R} , its probability density function $f_X(\cdot)$ is the function satisfying:

$$P\{X \leq x\} = \int_{-\infty}^x f_X(t) dt.$$

We interpret $f_X(x)$ as the “likelihood” that X takes on a value x . However, we need to exercise care in that interpretation. Note that

$$P\{X = x\} = \int_x^x f_X(t) dt = 0,$$

so the probability that X takes on precisely the value x (to infinite precision) is zero. The “likelihood interpretation” comes from the fact that

$$\frac{P\{X \in [a - \epsilon, a + \epsilon]\}}{P\{X \in [b - \epsilon, b + \epsilon]\}} = \frac{\int_{a-\epsilon}^{a+\epsilon} f_X(t) dt}{\int_{b-\epsilon}^{b+\epsilon} f_X(t) dt} \xrightarrow{\epsilon \rightarrow 0} \frac{f_X(a)}{f_X(b)}$$

so that $f_X(a)$ does indeed measure the relative likelihood that X takes on a value a (as opposed, say, to b).

Given a collection X_1, X_2, \dots, X_n of real-valued continuous rvs its joint probability density function (pdf) is defined as the function $f_{(X_1, X_2, \dots, X_n)}(\cdot)$ satisfying

$$P\{X_1 \leq x_1, \dots, X_n \leq x_n\} = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{(X_1, X_2, \dots, X_n)}(t_1, t_2, \dots, t_n) dt_1 \cdots dt_n.$$

Again, $f_{(X_1, \dots, X_n)}(x_1, \dots, x_n)$ can be given a likelihood interpretation. The collection X_1, X_2, \dots is independent if

$$f_{(X_1, X_2, \dots, X_n)}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

for all $(x_1, \dots, x_n) \in \mathbb{R}^n$.

Finally, the conditional pdf of X given $Y = y$ is given by

$$f_{X|Y}(x|y) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)}.$$

2.6 Sums of Random Variables

Many applications will require we compute the distribution of a sum $S_n = X_1 + X_2 + \dots + X_n$ where the X_i 's are jointly distributed real-valued rvs. If the X_i 's are continuous then

$$f_{S_n}(z) = \int_{-\infty}^{\infty} f_{X_n|S_{n-1}}(z-y|y) f_{S_{n-1}}(y) dy.$$

If the X_i 's are independent rvs,

$$f_{S_n}(z) = \int_{-\infty}^{\infty} f_{X_n}(z-y) f_{S_{n-1}}(y) dy.$$

This type of integral is known, in applied mathematics, as a *convolution integral*. So, $f_{S_n}(\cdot)$ can be computed recursively (in the independent setting) via $n-1$ convolution integrals.

A corresponding result holds in the discrete setting (with integrals replaced by sums).

2.7 Expectations

If X is a discrete real-valued rv, its expectation is defined as

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

(assuming the sum exists); if X is a continuous rv, its expectation is just

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

(assuming the integral exists).

Suppose that we wish to compute the expectation of $Y = g(X_1, \dots, X_n)$, where (X_1, \dots, X_n) is a jointly distributed collection of continuous rvs. The above definition requires that we first compute the pdf of Y and then calculate $E[Y]$ via the integral

$$E[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy.$$

Fortunately, there is an alternative approach to computing $E[Y]$ that is often easier to implement.

Result 2.1: In the above setting, $E[Y]$ can be compute as:

$$E[Y] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \dots, x_n) f_{(X_1, \dots, X_n)}(x_1, \dots, x_n) dx_1 \cdots dx_n.$$

Similarly in the discrete setting, if $Y = g(X_1, \dots, X_n)$, $E[Y]$ can be computed as

$$E[Y] = \sum_{x_1 \in S} \cdots \sum_{x_n \in S} g(x_1, \dots, x_n) P_{(X_1, \dots, X_n)}(x_1, \dots, x_n).$$

Remark 2.1: In older editions of his book, Sheldon Ross referred Result 2.1 as the “Law of the Unconscious Statistician”!

Example 2.1: Suppose X is a uniformly distributed rv on $[0, 1]$, so that

$$f_X(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{o.w.} \end{cases}$$

Let $Y = X^2$.

Approach 1 to computing $E[Y]$: Note that $P\{Y \leq y\} = P\{X^2 \leq y\} = P\{X \leq \sqrt{y}\} = \sqrt{y}$. So,

$$f_Y(y) = \frac{d}{dy} y^{\frac{1}{2}} = \frac{1}{2} y^{-\frac{1}{2}}.$$

Hence,

$$E[Y] = \int_0^1 y f_Y(y) dy = \frac{1}{2} \int_0^1 y^{\frac{1}{2}} dy = \frac{1}{2} \left[\frac{2}{3} y^{\frac{3}{2}} \right]_0^1 = \frac{1}{3}$$

Approach 2 to computing $E[Y]$:

$$E[Y] = \int_0^1 g(x) f_X(x) dy = \int_0^1 x^2 dx = \frac{1}{3}.$$

The expectation of a rv is interpreted as a measure of a rv’s “central tendency.” It is one of several summary statistics that are widely used in communicating the essential features of a probability distribution.

2.8 Commonly Used Summary Statistics

Given a rv X , the following are the most commonly used “summary statistics.”

1. *Mean of X :* The mean of X is just its expectation $E[X]$. We will see later, in our discussion of the law of large numbers, why $E[X]$ is a key characteristic of X ’s distribution.

2. *Variance of X :*

$$\text{var}(X) = E[(X - E[X])^2]$$

This is a measure of X ’s variability.

3. *Standard Deviation of X :*

$$\sigma(X) = \sqrt{\text{var}(X)}$$

This is a measure of variability that scales appropriately under a change in the units used to measure X (e.g. if X is a length, changing units from feet to inches multiplies the variance by 144, but the standard deviation by 12).

4. *Squared Coefficient of Variation:*

$$c^2(X) = \frac{\text{var}(X)}{\mathbb{E}[X]^2}$$

This is a dimensionless measure of variability that is widely used when characterizing the variation that is present in a non-negative rv X (e.g. task durations, component lifetimes, etc).

5. *Median of X :* this is the value n having the property that

$$P\{X \leq m\} = \frac{1}{2} = P\{X \geq m\}$$

(and is uniquely defined when $P\{X \leq \cdot\}$ is continuous and strictly increasing). It is a measure of the “central tendency” of X that complements the mean. Its advantage, relative to the mean, is that it is less sensitive to “outliers” (i.e. observations that are in the “tails” of X that have a big influence on the mean, but very little influence on the median).

6. *p^{th} quantile of X :* The p^{th} quantile of X is that value q having the property that

$$P\{X \leq q\} \triangleq F_X(q) = p$$

i.e. $q = F_X^{-1}(p)$.

7. *Inter-quartile range:* This is the quantity:

$$F_X^{-1}\left(\frac{3}{4}\right) - F_X^{-1}\left(\frac{1}{4}\right);$$

it is a measure of variability that, like the median, is (much) less sensitive to outliers than is the standard deviation.

2.9 Conditional Expectation

The conditional expectation of X , given $Y = y$ is just the quantity

$$\mathbb{E}[X|Y = y] = \sum_x x p_{X|Y}(x|y)$$

where X is discrete and

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx$$

when X is continuous. We can similarly define $\mathbb{E}[X|Y_1 = y_1, \dots, Y_n = y_n] = \mathbb{E}[X|\vec{Y} = \vec{y}]$ (where $\vec{Y} = (Y_1, \dots, Y_n)^T$ and $\vec{y} = (y_1, \dots, y_n)^T$). We sometimes denote $\mathbb{E}[X|Y = y]$ as $\mathbb{E}_y[X]$.

Note that expectations can be computed by “conditioning”:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \mathbb{E}[X|Y = y] f_X(y) dy$$

(if Y is continuous), and

$$\mathbb{E}[X] = \sum_y \mathbb{E}[X|Y = y] p_X(y)$$

(if Y is discrete). These equations can be rewritten more compactly as

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]].$$

This identity can be generalized. If $\vec{Y}_m = (Y_1, \dots, Y_m)^T$ and $\vec{Y}_n = (Y_1, \dots, Y_n)^T$, then

$$\mathbb{E}[X|\vec{Y}_m] = \mathbb{E}\left[\mathbb{E}[X|\vec{Y}_n]|\vec{Y}_m\right]$$

if $n \geq m$. This is often referred to as the “tower property” of conditional expectation.

2.10 Important Discrete Random Variables

1. *Bernoulli*(p) rv: $X \sim \text{Ber}(p)$ if $X \in \{0, 1\}$, and

$$\text{P}\{X = 1\} = p = 1 - \text{P}\{X = 0\}.$$

Application: Coin tosses, defective / non-defective items, etc.

Statistics:

$$\text{E}[X] = p \quad \text{var}(X) = p(1 - p)$$

2. *Binomial*(n, p) rv: $X \sim \text{Bin}(n, p)$ if $X \in \{0, 1, \dots, n\}$ and

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

Applications: Number of heads in n coin tosses; number of defectives in a product shipment of size n .

Statistics:

$$\text{E}[X] = np \quad \text{var}(X) = np(1 - p)$$

3. *Geometric*(p) rv: $X \sim \text{Geom}(p)$ if $X \in \{0, 1, \dots\}$ and

$$\text{P}\{X = k\} = p(1 - p)^k, \quad k \geq 0.$$

Applications: Number of coin tosses before the first head, etc.

Statistics:

$$\text{E}[X] = \frac{1 - p}{p} \quad \text{var}(X) = \frac{1 - p}{p^2}$$

A closely related variant, also called a geometric rv, arises when $X \in \{1, 2, \dots\}$, and

$$\text{P}\{X = k\} = p(1 - p)^{k-1}, \quad k \geq 1.$$

Here the statistics are:

$$\text{E}[X] = \frac{1}{p} \quad \text{var}(X) = \frac{1 - p}{p^2}$$

This time, it is the number of tosses required to observe the first head.

4. *Poisson*(λ) rv: $X \sim \text{Poisson}(\lambda)$ if $X \in \{0, 1, 2, \dots\}$ and

$$\text{P}\{X = k\} = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \geq 0.$$

Applications: Number of defective pixels on a high-definition TV screen, etc.

Statistics:

$$\text{E}[X] = \lambda \quad \text{var}(X) = \lambda$$

The Poisson rv arises as an approximation to a binomial rv when n is large and p is small. For example, if there are n pixels on a screen and the probability a given pixel is defective is p , then the total number of defectives on the screen is $\text{Bin}(n, p)$. In this setting, n is large and p is small. The binomial probabilities are cumbersome to work with when n is large because of the binomial coefficient that appear. As a result, we seek a suitable approximation. We propose the approximation

$$\text{Bin}(n, p) \stackrel{\mathcal{D}}{\approx} \text{Poisson}(np)$$

when n is large and p is small (where $\stackrel{\mathcal{D}}{\approx}$ denotes “has approximately the same distribution as”). This approximation is supported by the following theorem.

Theorem 2.1. $P\{\text{Bin}(n, p) = k\} \rightarrow P\{\text{Poisson}(\lambda) = k\}$ as $n \rightarrow \infty$, provided $np \rightarrow \lambda$ as $n \rightarrow \infty$.

Outline of proof: We will prove this for $k = 0$; the general case is similar. Note that

$$P\{\text{Bin}(n, p) = 0\} = (1 - p)^n = \left(1 - \frac{\lambda}{n} + o(n^{-1})\right)^n \rightarrow e^{-\lambda}$$

as $n \rightarrow \infty$ (where $o(a_n)$ represents a sequence having the property that $o(a_n)/a_n \rightarrow 0$ as $n \rightarrow \infty$).

2.11 Important Continuous Random Variables

1. *Uniform(a, b) rv:* $X \sim \text{Unif}(a, b)$, $a < b$ if

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{o.w.} \end{cases}$$

Applications: Arises in random number generation, etc.

Statistics:

$$E[X] = \frac{a+b}{2} \quad \text{var}(X) = \frac{(b-a)^2}{12}$$

2. *Beta(α, β) rv:* $X \sim \text{Beta}(\alpha, \beta)$, $\alpha, \beta > 0$, if

$$f_X(x) = \begin{cases} \frac{x^\alpha(1-x)^\beta}{B(\alpha, \beta)} & 0 \leq x \leq 1 \\ 0 & \text{o.w.} \end{cases}$$

where $B(\alpha, \beta)$ is the “normalization factor” chosen to ensure that $f_X(\cdot)$ integrates to one, i.e.

$$B(\alpha, \beta) = \int_0^1 y^\alpha(1-y)^\beta dy.$$

Applications: The Beta distribution is a commonly used “prior” on the Bernoulli parameter p .

Exercise 2.1: Compute the mean and variance of a Beta(α, β) rv in terms of the function $B(\alpha, \beta)$.

3. *Exponential(λ) rv:* $X \sim \text{Exp}(\lambda)$, $\lambda > 0$ if

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

Applications: Component lifetime, task duration, etc.

Statistics:

$$E[X] = \frac{1}{\lambda} \quad \text{var}(X) = \frac{1}{\lambda^2}$$

4. *Gamma(λ, α) rv:* $X \sim \text{Gamma}(\lambda, \alpha)$, $\lambda, \alpha > 0$, if

$$f_X(x) = \begin{cases} \frac{\lambda(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} & x \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

where

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

is the “gamma function.”

Applications: Component lifetime, task duration, etc.

Statistics:

$$E[X] = \frac{\alpha}{\lambda} \quad \text{var}(X) = \frac{\alpha}{\lambda^2}$$

5. *Gaussian / Normal rv:* $X \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}, \sigma^2 > 0$, if

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

Applications: Arises all over probability and statistics (as a result of the “central limit theorem”).

Statistics:

$$E[X] = \mu \quad \text{var}(X) = \sigma^2$$

Note that $N(\mu, \sigma^2) \stackrel{D}{=} \mu + \sigma N(0, 1)$, where $\stackrel{D}{=}$ denotes equality in distribution. (In other words, if one takes a $N(0, 1)$ rv, scales it by σ and adds μ on to it, we end up with a $N(\mu, \sigma^2)$ rv.)

6. *Weibull(λ, α) rv:* $X \sim \text{Weibull}(\lambda, \alpha)$, $\lambda, \alpha > 0$, if

$$P\{X > x\} = e^{-(\lambda x)^\alpha}$$

for $x \geq 0$. Hence:

$$f_X(x) = \begin{cases} \alpha \lambda^\alpha x^{\alpha-1} e^{-(\lambda x)^\alpha} & x \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

Applications: Component lifetime, task duration, etc.

Statistics:

$$E[X] = \frac{\Gamma(1 + \frac{1}{\alpha})}{\lambda} = \mu \quad \text{var}(X) = \frac{\Gamma(1 + \frac{2}{\alpha})}{\lambda^2} - \mu^2$$

7. *Pareto(λ, α) rv:* $X \sim \text{Pareto}(\lambda, \alpha)$, $\lambda, \alpha > 0$, if

$$f_X(x) = \begin{cases} \frac{\lambda \alpha}{(1 + \lambda x)^{\alpha+1}} & x \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

The Pareto distribution has a “tail” that decays to zero as a power of x (rather than exponentially rapidly (or faster) in x). As a result, a Pareto rv is said to be a “heavy tailed” rv.

Applications: Component lifetime, task duration, etc.

2.12 Some Quick Illustrations of Basic Probability

Example 2.2: (Capture / Recapture sampling)

We wish to estimate the number N of fish that are present in a lake. We start by visiting the lake, catching k fish, tagging each of the k fish and releasing them. (This is the “capture” phases.) A month later, we revisit the lake, catch n fish (the “recapture” phase), and count the number X of tagged fish that are present in the sample. How do we estimate the number N of fish in the lake?

Solution: Note that $X \sim \text{Bin}(n, p)$, where p is the probability that a fish is tagged. After a month, we assume that the tagged fish are “well mixed” with the total population of size N , so $p = k/N$. Given that $E[X] = np$, this suggests equating X to nk/N . In other words,

$$N \approx \frac{nk}{X}.$$

Example 2.3: (Poker)

What is the probability that a five card hand contains k hearts? ($k = 0, 1, \dots, 5$)

Solution: Note

- There are $\binom{13}{k}$ ways to choose k hearts from the 13 hearts present in the deck.
- There are $\binom{39}{5-k}$ ways to choose the remaining $5-k$ cards from the 39 “non-hearts” present in the deck.
- There are $\binom{52}{5}$ ways to choose 5 cards from a deck of 52 cards.

So,

$$P \{k \text{ hearts in a hand of 5}\} = \frac{\binom{13}{k} \binom{39}{5-k}}{\binom{52}{5}}$$

Example 2.4: (“Let’s make a Deal”)

A prize is behind one of three doors. Goats are behind the other 2 doors. We choose Door 1. If we choose correctly, we get the prize. The host opens one of Door 2 or Door 3 to expose a goat. Should we change our choice of door from Door 1 to the remaining unopened door (selected from either Door 2 or Door 3)?

Solution: Let Y be the door that the host exposes. Assume that the host knows what is behind each door, and never exposes the door behind which is the prize. Let P be the rv corresponding to the door the prize is behind so

$$P \{\text{the prize is behind door } k\} = P \{P = k\}$$

Then

$$P \{P = 1 | Y = 2\} = \frac{P \{P = 1 \cap Y = 2\}}{P \{Y = 2\}}$$

But,

$$P \{P = 1 \cap Y = 2\} = P \{P = 1\} P \{Y = 2 | P = 1\} = \frac{1}{3} \cdot \frac{1}{2},$$

where the $\frac{1}{2}$ presumes that if the price is behind the door we initially select, then the host randomly chooses to open one of the two doors with goats behind them at random. On the other hand,

$$\begin{aligned} P\{Y = 2\} &= P\{P = 1\}P\{Y = 2|P = 1\} + P\{P = 2\}P\{Y = 2|P = 2\} + P\{P = 3\}P\{Y = 2|P = 3\} \\ &= \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1 \\ &= \frac{1}{2}. \end{aligned}$$

So,

$$P\{P = 1|Y = 2\} = \frac{1}{3}.$$

Similarly,

$$P\{P = 2|Y = 2\} = 0, \quad P\{P = 3|Y = 2\} = \frac{2}{3},$$

so we should indeed change our choice of door in response to the information the host reveals.

Remark 2.2: For a more detailed discussion see Appendix B for more on the Monty Hall, Let's Make a Deal Problem..

Example 2.5: How should we model the idea that as a component ages, it becomes less reliable?

Solution: Let T be a continuous rv corresponding to the component lifetime. For $h > 0$ and fixed, consider

$$P\{T \in [t, t+h]|T > t\} = \frac{P\{T \in [t, t+h] \cap T > t\}}{P\{T > t\}} = \frac{P\{T \in [t, t+h]\}}{P\{T > t\}}.$$

This conditional probability is the likelihood the component fails in the next h time units give that it has survived to time t . Reduction in reliability as the component ages amounts to asserting that $P\{T \in [t, t+h]|T > t\}$ should be a decreasing function of t , i.e.

$$P\{T \in [t, t+h]|T > t\} \nearrow 0$$

in t . Note that when h is small,

$$P\{T \in [t, t+h]|T > t\} \approx \frac{f(t)}{\bar{F}(t)}h$$

where f is the density of T and $\bar{F}(t) \triangleq 1 - F(t) = P\{T > t\}$. Accordingly, $r(t) \triangleq f(t)/\bar{F}(t)$ is called the *failure rate* (at time t) of T . Modeling reduction in reliability as the component ages amounts to requiring that $r(t)$ should be increasing in t . In other words, T has an *increasing failure rate function*.

New components often exhibit a “burn-in” phases where they are subject to immediate (or rapid) failure because of the presence of manufacturing defects. Once a component survives through the burn-in phases, its reliability improves. Such components have *decreasing failure rate function* (at least through the end of the burn-in phase).

Most manufactured components have a failure rate that is “bathtub shaped”, see Figure 2.1

Over the operational interval $[t_1, t_2]$, the failure rate is essentially constant. This makes identifying the “constant failure rate” distribution interesting (since the failure rate of a component is often constant over the great majority of its design lifetime).

Suppose

$$r(t) = \lambda.$$

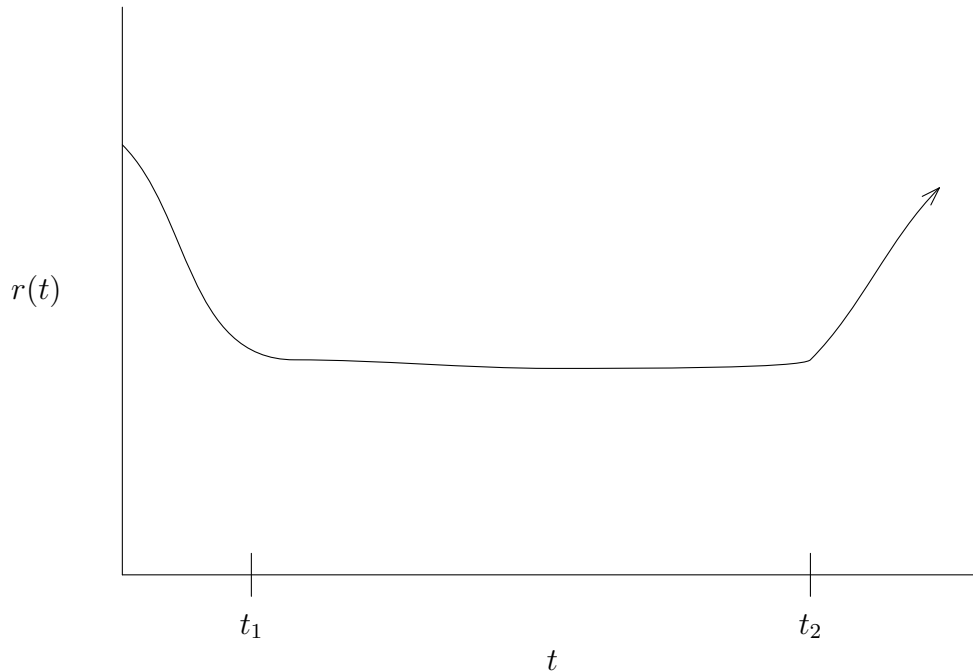


Figure 2.1: Sample Failure Rate Function

Then,

$$\frac{f(t)}{\bar{F}(t)} = \lambda,$$

so that

$$-\frac{d}{dt} \bar{F}(t) = \lambda \bar{F}(t).$$

We conclude that

$$-\frac{d}{dt} \log \bar{F}(t) = \lambda$$

so that

$$\log \bar{F}(t) = \log \bar{F}(0) - \lambda t.$$

Since T is positive, $\bar{F}(0) = 1$ and hence $\bar{F}(t) = e^{-\lambda t}$. In other words, $T \sim \text{Exp}(\lambda)$. So, exponential rvs are the unique rvs having a constant failure rate.

If $T \sim \text{Weibull}(\lambda, \alpha)$, then $\log \bar{F}(t) = -(\lambda t)^\alpha$. So,

$$r(t) = \frac{d}{dt} (\lambda t)^\alpha = \lambda \alpha t^{\alpha-1};$$

so if $\alpha < 1$, T has a decreasing failure rate, which if $\alpha > 1$, T has an increasing failure rate. When $\alpha = 1$, T has a constant failure rate and is exponentially distributed.

2.13 Statistical Parameter Estimation: The Method of Maximum Likelihood

In building stochastic models, it is often the case that observational data exists that can be used to help guide the construction of an appropriate model. In particular, the existing data can be used to help estimate the

model parameters. Statisticians call the process of fitting the parameters of a model to data the “parameter estimation problem” (“estimation” for short).

To provide a concrete example, consider the problem of building a stochastic model to represent the number of defective pixels on a high-definition television screen. We argued earlier, in Section 10 of this chapter, that a good model for the number X of defective pixels on such a screen is to assume that it follows a Poisson distribution with parameter λ^* . We now wish to estimate λ^* .

We select five such screens and count the number of defective pixels on each of the five screens, leading to counts of 0, 3, 4, 2 and 5, respectively. We view the five observations as a *random sample* from the distribution of X , by which we mean that the five observations are the realized values of five iid (independent and identically distributed) rvs X_1, X_2, \dots, X_5 having a common Poisson (λ^*) distribution.

Maximum likelihood is generally viewed as the “gold standard” method for estimating statistical parameters. We will discuss later the theoretical basis for why maximum likelihood is a preferred approach to estimating parameters. The method of maximum likelihood asserts that one should:

Estimate the parameter λ^* as that value $\hat{\lambda}$ that maximizes the likelihood of observing the given sample.

In this case, the likelihood of observing 0,3,4,2 and 5 under a Poisson (λ) model is:

$$L(\lambda) = \frac{e^{-\lambda}\lambda^0}{0!} \cdot \frac{e^{-\lambda}\lambda^3}{3!} \cdot \frac{e^{-\lambda}\lambda^4}{4!} \cdot \frac{e^{-\lambda}\lambda^2}{2!} \cdot \frac{e^{-\lambda}\lambda^5}{5!} = \frac{e^{-5\lambda}\lambda^{14}}{0!3!4!2!5!}$$

The maximizer $\hat{\lambda}$ of $L(\cdot)$ is equal to the maximizer of the log-likelihood $\mathcal{L}(\cdot)$, namely

$$\mathcal{L}(\lambda) = -5\lambda + 14 \log \lambda - \log(0!3!4!2!5!).$$

The maximizer $\hat{\lambda}$ satisfies:

$$\left. \frac{d}{d\lambda} \mathcal{L}(\lambda) \right|_{\lambda=\hat{\lambda}} = -5 + \frac{14}{\hat{\lambda}} = 0.$$

i.e.

$$\hat{\lambda} = \frac{14}{5}.$$

2.13.1 MLE for an Exp (λ) Random Variable

More generally, if X_1, X_2, \dots, X_n is a random sample from a Poisson (λ^*) distribution, then the likelihood is:

$$L_n(\lambda) = \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!},$$

having maximizer

$$\hat{\lambda}_n = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

In other words, $\hat{\lambda}_n$ is just the (arithmetic) mean of the sample. This so-called *sample mean* is usually denoted as \bar{X}_n . We next work out the maximum likelihood estimator (MLE) for normally distributed and gamma distributed rvs.

2.13.2 MLE for a $N(\mu, \sigma^2)$ Random Variable

Here we work out the MLE for normally distributed (Gaussian) rvs. Suppose that we observe a random sample X_1, X_2, \dots, X_n of iid observations from a $N(\mu^*, \sigma^{*2})$ distribution. The corresponding likelihood is:

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}}$$

and the log-likelihood is

$$\mathcal{L}_n(\mu, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log(2\pi).$$

The MLE for (μ^*, σ^{*2}) is the value $(\hat{\mu}_n, \hat{\sigma}_n^2)$ satisfying

$$\frac{\partial}{\partial \mu} \mathcal{L}_n(\hat{\mu}_n, \hat{\sigma}_n^2) = \sum_{i=1}^n \frac{(x_i - \hat{\mu}_n)}{\hat{\sigma}_n^2} = 0$$

and

$$\frac{\partial}{\partial \sigma^2} \mathcal{L}_n(\hat{\mu}_n, \hat{\sigma}_n^2) = -\frac{n}{2\hat{\sigma}_n^2} + \sum_{i=1}^n \frac{(x_i - \hat{\mu}_n)^2}{2\hat{\sigma}_n^4} = 0.$$

This yields:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i = \bar{X}_n \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_n)^2.$$

Remark: It turns out that the estimators that are most frequently used by statisticians to estimate the parameters (μ^*, σ^{*2}) for Gaussian models are the following. Estimate μ^* via \bar{X}_n and estimate σ^{*2} via

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu}_n)^2 = \frac{n}{n-1} \hat{\sigma}_n^2.$$

The estimator s_n^2 is what statisticians call the *sample variance*. Note that when n is reasonably large, s_n^2 and $\hat{\sigma}_n^2$ are almost identical. But for small samples, s_n^2 and $\hat{\sigma}_n^2$ differ. Statisticians generally prefer s_n^2 to $\hat{\sigma}_n^2$ because s_n^2 is undefined when $n = 1$ (as it should be) and s_n^2 is *unbiased* as an estimator of σ^{*2} , by which we mean that

$$E[s_n^2] = \sigma^{*2}$$

for $n \geq 2$.

Exercise 2.2: Prove that s_n^2 is an unbiased estimator for σ^{*2} when $n \geq 2$.

2.13.3 MLE for a Gamma (λ, α) Random Variable

Suppose that we observe a random sample X_1, X_2, \dots, X_n of iid observations from a Gamma (λ^*, α^*) population. The corresponding likelihood is

$$L_n(\lambda, \alpha) = \prod_{i=1}^n \frac{\lambda(\lambda x_i)^{\alpha-1} e^{-\lambda x_i}}{\Gamma(\alpha)}$$

and the log-likelihood is

$$\mathcal{L}_n(\lambda, \alpha) = n\alpha \log \lambda + (\alpha - 1) \sum_{i=1}^n \log x_i - \lambda \sum_{i=1}^n x_i - n \log \Gamma(\alpha).$$

For this example, there is no “closed form” for the maximizer $(\hat{\lambda}_n, \hat{\alpha}_n)$ of $\mathcal{L}_n(\cdot)$; the MLE $(\hat{\lambda}_n, \hat{\alpha}_n)$ must be computed numerically. This example illustrates a key point about MLE’s. While they are the statistical “gold standard”, they are often notoriously difficult to compute (even in the presence of powerful computers).

Exercise 2.3: Compute the MLE for a random sample from a Weibull (λ^*, α^*) population.

Exercise 2.4: Compute the MLE for a random sample from a Unif (a^*, b^*) population.

Exercise 2.5: Compute the MLE for a random sample from a Bin (n, p^*) population (where n is known).

Exercise 2.6: Compute the MLE for a random sample of Beta (α^*, β^*) population.

2.13.4 MLE as a “Gold Standard”

Let us now return the question of why maximum likelihood is the “gold standard” estimation method. Suppose that we have a random sample from a normally distributed population in which μ^* is unknown, but the variance σ^{*2} is known to equal one. Recall that for a normal distribution, μ^* characterizes both the mean and the median. This suggests estimating μ^* via either the estimator \bar{X}_n (the sample mean) or m_n , the sample median. (The sample median is the $(k+1)$ th largest observation when $n = 2k+1$, and the median is defined as the arithmetic average of the k th and $(k+1)$ th largest observations when $n = 2k$.) Since the sample is random, the estimators \bar{X}_n and m_n are themselves random variables. The hope is that when the sample size n is large, \bar{X}_n and m_n will be close to μ^* . The preferred estimator is clearly the one that has tendency to be closer to μ^* .

One way to mathematically characterize this preference is to study the rate of convergence of the estimator to μ^* . We will see in Chapter 2 that both \bar{X}_n and m_n obey “central limit theorems” that assert that \bar{X}_n and m_n are, for large n , asymptotically normally distributed with common mean μ^* and variances $\eta_1^{2/n}$ and $\eta_2^{2/n}$, respectively. Our preference should obviously be for the estimator with the smaller value $\eta_i^{2/n}$.

The estimator \bar{X}_n is the MLE in this Gaussian setting. As the “gold standard” estimator, it will come as no surprise that η_1^2 is always less than or equal to η_2^2 . So \bar{X}_n is to be preferred to m_n as an estimator of the parameter μ^* in a $N(\mu^*, 1)$ population. It is the fact that the MLE has the fastest possible rate of convergence among all possible estimators of an unknown statistical parameter that has led to its adoption as the “gold standard” estimator. (For those of you familiar with statistics, the MLE achieves (in great generality) the Cramér-Rao lower bound that describes a theoretical lower bound on the variance of an (unbiased) estimator of an unknown statistical parameter.) As a consequence, it is typical that in approaching parameter estimation problems, the first order of business is to study the associated MLE. If computation of the MLE is analytically or numerically tractable, then one would generally adopt the MLE as one’s preferred estimator.

2.14 The Method of Moments

An alternative approach to estimating model parameters is the “method of moments.” Let us illustrate this idea in the setting of a gamma distributed random sample.

Given a random sample X_1, X_2, \dots, X_n of iid observations for a Gamma (λ^*, α^*) population, recall that:

$$E[X] = \frac{\alpha^*}{\lambda^*} \quad \text{and} \quad \text{var}(X) = \frac{\alpha^*}{\lambda^{*2}}.$$

One expects that that if the sample size n is large, then the sample mean \bar{X}_n and sample variance s_n^2 will be close to $E[X]$ and $\text{var}(X)$, respectively. (This will follow from the Law of Large Number, to be discussed in Chapter 2.) This suggests that

$$\bar{X}_n \approx \frac{\alpha^*}{\lambda^*} \quad \text{and} \quad s_n^2 \approx \frac{\alpha^*}{\lambda^{*2}}.$$

The “methods of moments” estimators $\hat{\alpha}$ and $\hat{\lambda}$ for α^* and λ^* are obtained by replacing the approximations with the equalities:

$$\bar{X}_n = \frac{\hat{\alpha}}{\hat{\lambda}} \quad \text{and} \quad s_n^2 = \frac{\hat{\alpha}}{\hat{\lambda}^2}.$$

This leads to the estimators

$$\hat{\alpha} = \frac{\bar{X}_n^2}{s_n^2} \quad \text{and} \quad \hat{\lambda} = \frac{\bar{X}_n}{s_n^2}.$$

Note that the method of moments estimators $\hat{\alpha}$ and $\hat{\lambda}$ can be computed in analytical closed form, unlike the MLE (which in this gamma setting must be computed numerically). Hence, the method of moments estimators are (at least in this example) more tractable. On the other hand, they are *inefficient* statistically, because they do not achieve the Cramér-Rao lower bound. So, method of moments estimators do not typically fully exploit all the statistical information that is present in a sample (unlike MLE’s). The advantage of the method of moments is that they can offer a computationally tractable alternative to parameter estimation in settings where maximum likelihood is too difficult to implement numerically.

In general, if the rv X has a distribution that depends on d unknown statistical parameters $\theta_1^*, \theta_2^*, \dots, \theta_d^*$ one writes down expressions for the first d moments of the rv X (i.e. $E[X^k]$ for $k = 1, \dots, d$) in terms of the d parameters, leading to the equations

$$E[X^k] = f_k(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_d), \quad k = 1, \dots, d.$$

The method of moments estimators $\hat{\theta}_1, \dots, \hat{\theta}_d$ are obtained by equating the population moments to the sample moments, namely as solution to the simultaneous equations

$$\frac{1}{n} \sum_{i=1}^n X_i^k = f_k(\hat{\theta}_1, \dots, \hat{\theta}_d), \quad k = 1, \dots, d.$$

Exercise 2.7: Compute the method of moment estimators for a $N(\mu^*, \sigma^{*2})$ populations.

Exercise 2.8: Compute the method of moments estimators for a $\text{Unif}(a^*, b^*)$ population.

Exercise 2.9: Compute the method of moments estimators for Beta (α^*, β^*) population.

Exercise 2.10: Compute the method of moments estimators for a Weibull (λ^*, α^*) population.

2.15 Bayesian Statistics

Consider a case where we are attempting to estimate a Bernoulli parameter p^* corresponding to the probability that a given manufactured item is defective. With a good manufacturing process in place p^* should be small.

In this cases, if we test n items, it is likely that all n items are non-defective. In other words, the random sample X_1, X_2, \dots, X_n from such a Bernoulli population is likely to be one in which $X_i = 0$ for $1 \leq i \leq n$. The maximum likelihood estimator (and method of moments estimator) \hat{p}_n for p^* is given by

$$\hat{p}_n = \bar{X}_n = 0.$$

Given the experimental data observed, this is perhaps a reasonable estimate for p^* .

But it is unlikely that a company would base any of its operational decisions on such an estimate of p^* . Nobody truly believes that they have a flawless manufacturing process. One has a prior belief that p^* is

positive. Bayesian statistical methods offer a means of taking advantage of such prior information.

In a Bayesian approach to statistical analysis, one would view p^* as itself being a random variable. The distribution of p^* (the so-called “prior distribution” on p^*) reflects the statistician’s beliefs about the likely values of p^* in the absence of any experimental data. In our Bernoulli example, one possible prior would be the uniform distribution on $[0, 1]$. Having postulated a prior, we now observe a random sample X_1, X_2, \dots, X_n . Conditional on $\hat{p} = p$, the likelihood of the sample is just

$$L_n(p) = \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i}.$$

We now wish to compute a new “prior distribution” on \hat{p} that reflects the influence of the observed sample on the prior. We do this by taking advantage of the basic ideas of conditional probability. In this statistical setting, this application of conditional probability is often called “Bayes’ rule.” In particular, the posterior distribution is just the distribution of \hat{p} , given X_1, \dots, X_n . This translates into

$$f(p|X_1, \dots, X_n) = \frac{p^{S_n} (1-p)^{n-S_n}}{\int_0^1 r^{S_n} (1-r)^{n-S_n} dr}$$

where $S_n = X_1 + \dots + X_n$. In particular, if $X_1 = \dots = X_n = 0$, we find that

$$f(p|X_1 = 0, \dots, X_n = 0) = (n+1)(1-p)^n.$$

The mean of the posterior distribution is

$$\int_0^1 p f(p|X_1 = 0, \dots, X_n = 0) dp = \frac{1}{n+2}.$$

(Note the when $n = 0$, the mean is $\frac{1}{2}$, which coincides with the mean of the uniform prior.) Thus, the Bayesian approach here leads to an analysis that seems more consistent with usage of statistics in an operational decision-making environment.

Such a Bayesian approach to statistical analysis can be applied in any setting in which the underlying data is assumed to follow a parametric distribution. In particular, suppose that the random sample X_1, \dots, X_n is a collection of observations from a population having a density function $f(\cdot; \theta^*)$, where θ^* is the “true” value of the unknown parameter θ . Suppose $p(\cdot)$ is a density corresponding to a prior distribution on θ^* . Bayes’ rule dictates that the posterior distribution on θ^* equals

$$f(\theta|X_1, \dots, X_n) \triangleq \frac{p(\theta) \prod_{i=1}^n f(X_i; \theta)}{\int_{\Lambda} p(\theta') \prod_{i=1}^n f(X_i; \theta') d\theta'}$$

where Λ is the set of all possible values for the parameter θ .

Exercise 2.11: Suppose that we observe a random sample from a $N(\mu^*, \sigma^{*2})$ population.

1. Compute the posterior distribution on μ^* when μ^* has the prior that is $N(\tau, 1)$ distributed.
2. Repeat 1 for a general prior.

Exercise 2.12: Suppose that we observe a random sample from a $\text{Ber}(p^*)$ population.

1. Compute the posterior distribution on p^* when p^* has a prior that is $\text{Beta}(\alpha, \beta)$ distributed.
2. Repeat 1 for a general prior.

The priors that are postulated in part 1 of the above problems are called “conjugate priors” for the normal and Bernoulli distributions, respectively. Note that use of a conjugate prior simplifies computation of the posterior.

2.16 The Law of Large Numbers

One of the two most important results in probability is the law of large numbers (LLN).

Theorem 2.2. *Suppose that $(X_n : n \geq 1)$ is a sequence of iid rv's. Then,*

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{P} E(X_1)$$

as $n \rightarrow \infty$ ¹.

This result is easily to prove when X_i 's have finite variance. The key is the following inequality, called *Markov's inequality*.

Proposition 2.1: Suppose that W is a non-negative rv. Then,

$$P(W > w) \leq \frac{1}{w}E(W).$$

Proof. Note that if W is a continuous rv,

$$\begin{aligned} P(W > w) &= \int_w^\infty f(x) dx \\ &\leq \int_w^\infty \left(\frac{x}{w}\right) f(x) dx \quad (\text{since } \frac{x}{w} \geq 1 \text{ when } x \geq w) \\ &\leq \int_0^\infty \left(\frac{x}{w}\right) f(x) dx = \frac{1}{w}E(W). \end{aligned}$$

The proof is similar for discrete rv's. □

An important special case is called *Chebyshev's inequality*.

Proposition 2.2: Suppose that X_i 's are iid with common (finite) variance σ^2 . If $S_n = X_1 + \cdots + X_n$, then

$$P\left\{\left|\frac{S_n}{n} - E X_1\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

Proof. Put $W = (S_n - nE(X_1))^2$ and $w = n^2\epsilon^2$. Note that $E(W) = \text{var}(S_n) = n\sigma^2$, so

$$P\left(\left|\frac{S_n}{n} - E(X_1)\right| > \epsilon\right) = P(W > w) \leq \frac{\sigma^2}{n\epsilon^2}.$$

□

Theorem 2.2 is an immediate consequence of Proposition 2.2. Let's now apply the LLN.

The LLN guarantees that even though the “sample average” $n^{-1}(X_1 + \cdots + X_n)$ is a rv, it “settles down” to something deterministic and predictable when n is large, namely $E(X_1)$. Hence, even though the individual X_i 's are unpredictable, their average (or mean) is predictable. The fact that the average $n^{-1}(X_1 + \cdots + X_n)$ settles down to the expectation $E(X_1)$ is a principal reason for why the expectation of a rv is the most widely used “measure of central tendency” (as opposed, for example, to the median of the distribution).

¹Here \xrightarrow{P} denotes convergence in probability. For a definition see Appendix A

2.17 Central Limit Theorem

The second key limit result in probability is the *central limit theorem* (CLT). (It is so important that it is the “central” theorem of probability!)

Note that the LLN approximation (2.2) is rather crude:

$$P(X_1 + \cdots + X_n \leq x) \approx \begin{cases} 0, & x < np \\ 1, & x \geq np \end{cases}$$

Typically, we’d prefer an approximation that tells us how close $P(X_1 + \cdots + X_n \leq x)$ is to 0 when $x < np$ and how close to 1 when $x \geq np$. The CLT provides exactly this additional information.

Theorem 2.3. *Suppose that the X_i ’s are iid rv’s with common (finite) variance σ^2 . Then, if $S_n = X_1 + \cdots + X_n$*

$$\frac{S_n - nE(X_1)}{\sqrt{n}} \Rightarrow \sigma N(0, 1) \quad (2.1)$$

as $n \rightarrow \infty$.

The CLT (2.1) supports the use of the approximation

$$S_n \stackrel{\mathfrak{D}}{\approx} nE(X_1) + \sqrt{n}\sigma N(0, 1) \quad (2.2)$$

when n is large. The approximation (2.2) is valuable in many different problem settings. We now illustrate its use with an example.

An outline of the proof of the CLT is given later in the notes.

2.18 Moment Generating Functions

A key idea of in applied mathematics is that of the Laplace transform. The Laplace transform also is a useful tool in probability. In the probability context, the Laplace transform is usually called the *moment generating function* (of the rv).

Definition 2.1: The moment generating function of a rv X is the function $\varphi_X(\theta)$ defined by

$$\varphi_X(\theta) = E(\exp(\theta x)).$$

This function can be computed in “closed form” for many of the distributions encountered most frequently in practice:

- Bernoulli rv: $\varphi_X(\theta) = (1 - p) + pe^\theta$
- Binomial(n, p) rv: $\varphi_X(\theta) = ((1 - p) + pe^\theta)^n$
- Geometric(p) rv: $\varphi_X(\theta) = p/(1 - (1 - p)e^\theta)$
- Poisson(λ) rv: $\varphi_X(\theta) = \exp(\lambda(e^\theta - 1))$
- Uniform(a, b) rv: $\varphi_X(\theta) = (e^{\theta b} - e^{\theta a})/\theta(b - a)$
- Exponential(λ) rv: $\varphi_X(\theta) = \lambda(\lambda - \theta)^{-1}$
- Gamma(λ, α) rv: $\varphi_X(\theta) = \left(\frac{\lambda}{\lambda - \theta}\right)^\alpha$
- Normal(μ, σ^2) rv: $\varphi_X(\theta) = \exp(\theta\mu + \frac{\sigma^2\theta^2}{2})$

The moment generating function (mgf) of a rv X gets its name from the fact that the moments (i.e. $E(X^k)$ for $k = 1, 2, \dots$) of the rv X can easily be computed from knowledge of $\varphi_X(\theta)$. To see this, note that if X is continuous, then

$$\begin{aligned} \frac{d^k}{d\theta^k} \varphi_X(\theta) &= \frac{d^k}{d\theta^k} E(\exp(\theta X)) \\ &= \frac{d^k}{d\theta^k} \int_{-\infty}^{\infty} e^{\theta x} f(x) dx \\ &= \int_{-\infty}^{\infty} \frac{d^k}{d\theta^k} e^{\theta x} f(x) dx \\ &= \int_{-\infty}^{\infty} x^k e^{\theta x} f(x) dx \\ &= E(X^k \exp(\theta X)). \end{aligned}$$

In particular,

$$\frac{d^k}{d\theta^k} \varphi_X(0) = E(X^k).$$

Example 2.6: Suppose that X is exponentially distributed with parameter λ . Note that

$$\varphi_X(\theta) = \lambda/(\lambda - \theta)^{-1} = 1/(1 - \frac{\theta}{\lambda})^{-1} = \sum_{k=0}^{\infty} \frac{1}{\lambda^k} \theta^k \quad (2.3)$$

On the other hand, $\varphi_X(\theta)$ has the power series representation,

$$\varphi_X(\theta) = \sum_{k=0}^{\infty} \frac{1}{k!} \frac{d^k}{d\theta^k} \varphi_X(0) \theta^k \quad (2.4)$$

Equating coefficients in (2.3) and (2.4), we find that

$$\frac{d^k}{d\theta^k} \varphi_X(0) = \frac{k!}{\lambda^k},$$

so that

$$E(X^k) = \frac{k!}{\lambda^k}.$$

Note that we were able to compute all the moments of an exponential rv without having to repeatedly compute integrals.

Another key property of mgf's is the fact that uniquely characterizes the distribution of the rv. In particular, if X and Y are such that $\varphi_X(\theta) = \varphi_Y(\theta)$ for all values of θ , then

$$P(X \leq x) = P(Y \leq x)$$

for all x .

This property turns out to be very useful when combined with the following proposition.

Proposition 2.3: Let the X_i 's be independent rv's, and put $S_n = X_1 + \dots + X_n$. Then,

$$\varphi_{S_n}(\theta) = \prod_{i=1}^n \varphi_{X_i}(\theta).$$

Proof. Note that

$$\begin{aligned}
 \varphi_{S_n}(\theta) &= \mathbb{E}(\exp(\theta(X_1 + \cdots + X_n))) \\
 &= \mathbb{E}\left(\prod_{i=1}^n \exp(\theta X_i)\right) \\
 &= \prod_{i=1}^n \mathbb{E}(\exp(\theta X_i)) \quad (\text{due to independence}) \\
 &= \prod_{i=1}^n \varphi_{X_i}(\theta).
 \end{aligned}$$

□

In other words, the mgf of a sum of independence rv's is trivial to compute in terms of the mgf's of the summands. So, one way to compute the exact distribution of a sum of n independent rv's X_1, \dots, X_n is:

1. Compute $\varphi_{X_i}(\theta)$ for $1 \leq i \leq n$.
2. Compute

$$\varphi_{S_n}(\theta) = \prod_{i=1}^n \varphi_{X_i}(\theta).$$

3. Find a distribution/rv Y such that

$$\varphi_{S_n}(\theta) = \varphi_Y(\theta)$$

for all θ .

Then,

$$P(S_n \leq x) = P(Y \leq x).$$

Example 2.7: Suppose the X_i 's are iid Bernoulli rv's with parameter p . Then,

$$\varphi_{S_n}(\theta) = (1 - p + pe^\theta)^n.$$

But $(1 - p + pe^\theta)^n$ is the mgf of a Binomial rv with parameter n and p . So,

$$P(S_n = k) = P(\text{Bernoulli}(n, p) = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Theorem 2.4. Let $(X_n : 1 \leq n \leq \infty)$ be a sequence of rv's with mgf's $(\varphi_{X_n}(\theta) : 1 \leq n \leq \infty)$. If, for each θ ,

$$\varphi_{X_n}(\theta) \rightarrow \varphi_{X_\infty}(\theta)$$

as $n \rightarrow \infty$, then

$$X_n \Rightarrow X_\infty$$

as $n \rightarrow \infty$.