Note 2: Brief Review of Probability Distribution

1 Review of Probability

Random variables are denoted by X, Y, Z, etc. The *cumulative distribution function (c.d.f.)* of a random variable X is denoted by $F(x) = P(X \leq x),\ -\infty < x < \infty$, and if the random variable is continuous then its probability density function is denoted by $f(x)$ which is related to $F(x)$ via

$$\begin{aligned} f(x) &= F'(x) = \frac{d}{dx}F(x) \\ F(x) &= \int_{-\infty}^{x} f(y)dy. \end{aligned}$$

The *probability mass function (p.m.f.)* of a discrete random variable is given by

$$p(k) = P(X = k),\ -\infty < k < \infty,$$

for integers k.

$1 - F(x) = P(X > x)$ is called the *tail* of X and is denoted by $\overline{F}(x) = 1 - F(x)$. Whereas $F(x)$ increases to 1 as $x \to \infty$, and decreases to 0 as $x \to -\infty$, the tail $\overline{F}(x)$ decreases to 0 as $x \to \infty$ and increases to 1 as $x \to -\infty$.

If a r.v. X has a certain distribution with c.d.f. $F(x) = P(X \leq x)$, then we write, for simplicity of expression,

$$X \sim F. \tag{1}$$

1.1 Moments and variance

The expected value of a r.v. is denote by $E(X)$ and defined by

$$\begin{aligned} E(X) &= \sum_{k=-\infty}^{\infty} kp(k),\ \text{discrete case}, \\ E(X) &= \int_{-\infty}^{\infty} xf(x)dx,\ \text{continuous case}. \end{aligned}$$

$E(X)$ is also referred to as the *first moment* or mean of X (or of its distribution).

Higher moments $E(X^n),\ n \geq 1$ can be computed via

$$\begin{aligned} E(X^n) &= \sum_{k=-\infty}^{\infty} k^n p(k),\ \text{discrete case}, \\ E(X^n) &= \int_{-\infty}^{\infty} x^n f(x)dx,\ \text{continuous case}, \end{aligned}$$

and more generally $E(g(X))$ for a function $g = g(x)$ can be computed via

$$\begin{aligned} E(g(X)) &= \sum_{k=-\infty}^{\infty} g(k)p(k),\ \text{discrete case}, \\ E(g(X)) &= \int_{-\infty}^{\infty} g(x)f(x)dx,\ \text{continuous case}. \end{aligned}$$

(Leting $g(x) = x^n$ yields moments for example.)

Finally, the variance of X is denoted by $Var(X)$, defined by $E\{|X - E(X)|^2\}$, and can be computed via

$$Var(X) = E(X^2) - E^2(X), \tag{2}$$

the second moment minus the square of the first moment.

We usually denote the variance by $\sigma^2 = Var(X)$ and when necessary (to avoid confusion) include X as a subscript, $\sigma_X^2 = Var(X)$. $\sigma = \sqrt{Var(X)}$ is called the *standard deviation* of X.

For any r.v. X and any number a

$$E(aX) = aE(X), \text{ and } Var(aX) = a^2 Var(X). \tag{3}$$

For any two r.v.s. X and Y

$$E(X + Y) = E(X) + E(Y). \tag{4}$$

If X and Y are independent, then

$$Var(X + Y) = Var(X) + Var(Y). \tag{5}$$

The above properties generalize in the obvious fashion to to any finite number of r.v.s.

In general (independent or not)

$$Var(X + Y) = Var(X) + V(Y) + 2Cov(X, Y),$$

where

$$Cov(X, Y) \stackrel{\text{def}}{=} E(XY) - E(X)E(Y),$$

is called the *covariance* between X and Y, and is usually denoted by $\sigma_{X,Y} = Cov(X, Y)$.

When $Cov(X, Y) > 0$, X and Y are said to be *positively correlated*, whereas when $Cov(X, Y) < 0$, X and Y are said to be *negatively correlated*. When $Cov(X, Y) = 0$, X and Y are said to be *uncorrelated*, and in general this is weaker than independence of X and Y: *there are examples of uncorrelated r.v.s. that are not independent*. Note in passing that $Cov(X, X) = Var(X)$.

The *correlation coefficient* of X, Y is defined by

$$\rho = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y},$$

and it always holds that $-1 \leq \rho \leq 1$. When $\rho = 1$, X and Y are said to be perfectly (positively) correlated.

1.2 Moment generating functions

The moment generating function (mgf) of a r.v. X (or its distribution) is defined for all $s \in (-\infty, \infty)$ by

$$\begin{aligned} M(s) &\stackrel{\text{def}}{=} E(e^{sX}) \\ &= \int_{-\infty}^{\infty} e^{sx} f(x) dx \quad \left(= \sum_{-\infty}^{\infty} e^{sk} p(k) \text{ in the discrete r.v. case}\right) \end{aligned} \tag{6}$$

It is so called because it generates the moments of X by differentiation at $s = 0$:

$$M'(0) = E(X), \tag{7}$$

and more generally

$$M^{(n)}(0) = E(X^n),\ n \geq 1. \tag{8}$$

The mgf uniquely determines a distribution in that no two distributions can have the same mgf. So knowing a mgf characterizes the distribution in question.

If X and Y are independent, then $E(e^{s(X+Y)}) = E(e^{sX}e^{sY}) = E(e^{sX})E(e^{sY})$, and we conclude that *the mgf of an independent sum is the product of the individual mgf's.*

Sometimes to stress the particular r.v. X, we write $M_X(s)$. Then the above independence property can be concisely expressed as

$$M_{X+Y}(s) = M_X(s)M_Y(s),\ \text{when } X \text{ and } Y \text{ are independent.}$$

Remark 1.1 For a given distribution, $M(s) = \infty$ is possible for some values of s, but there is a large useful class of distributions for which $M(s) < \infty$ for all s in *a neighborhood of the origin*, that is, for $s \in (-\epsilon, \epsilon)$ with $\epsilon > 0$ suffiently small. Such distributions are referred to as *light-tailed* because their tails can be shown to tend to zero quickly. There also exists distributions for which no such neighborhood exists and this can be so even if the distribution has finite moments of all orders (see the lognormal distribution for example). A large class of such distributions are referred to as *heavy-tailed* because their tails tend to zero slowly.

Remark 1.2 For non-negative r.v.s. X, it is sometimes more common to use the *Laplace transform*, $\mathcal{L}(s) = E(e^{-sX}),\ s \geq 0$, which is always finite, and then $(-1)^n\mathcal{L}^{(n)}(0) = E(X^n),\ n \geq 1$.

For discrete r.v.s. X, it is sometimes more common to use

$$M(z) = E(z^X) = \sum_{k=-\infty}^{\infty} z^k p(k),\ |z| \leq 1$$

for the mgf in which case moments can be generated via $M'(1) = E(X)$, $M''(1) = E((X)(X-1))$, $M^{(n)}(1) = E(X(X-1)\cdots(X-(n-1))),\ n \geq 1$.

1.3 Examples of well-known distributions

Discrete case

1. ***Bernoulli distribution*** *with success probability p:* With $0 < p < 1$ a constant, X has p.m.f. $p(k) = P(X = k)$ given by

$$\begin{aligned} p(1) &= p, \\ p(0) &= 1-p, \\ p(k) &= 0,\ \text{otherwise.} \end{aligned}$$

Thus X only takes on the values 1 (success) or 0 (failure).

A simple computation yields

$$\begin{aligned} E(X) &= p \\ Var(X) &= p(1-p) \\ M(s) &= pe^s + 1 - p. \end{aligned}$$

Bernoulli r.v.s. arise naturally as the *indicator function*, $X = I\{A\}$, of an event A, where

$$I\{A\} \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if the event } A \text{ occurs;} \\ 0, & \text{otherwise.} \end{cases}$$

Then $p = P(X = 1) = P(A)$ is the probability that the event A occurs. For example, if you flip a coin once and let $A = \{\text{coin lands heads}\}$, then for $X = I\{A\}$, $X = 1$ if the coin lands heads, and $X = 0$ if it lands tails. Because of this elementary and intuitive coin-flipping example, a Bernoulli r.v. is sometimes referred to as a coin flip, where p is the probability of landing heads.

Observing the outcome of a Bernoulli r.v. is sometimes called *performing a Bernoulli trial*, or experiment.

Keeping in the spirit of (1) we denote a Bernoulli p r.v. by

$$X \sim Bern(p).$$

2. ***Binomial distribution*** *with success probability p and n trials:* If we consecutively perform n independent Bernoulli p trials, $X_1, \ldots, X_n$, then the total number of successes $X = X_1 + \cdots + X_n$ yields the Binomial r.v. with p.m.f.

$$p(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k}, & \text{if } 0 \le k \le n; \\ 0, & \text{otherwise.} \end{cases}$$

In our coin-flipping context, when consecutively flipping the coin exactly n times, $p(k)$ denotes the probability that exactly k of the n flips land heads (and hence exactly $n-k$ land tails).

A simple computation (utilizing $X = X_1 + \cdots + X_n$ and independence) yields

$$\begin{aligned} E(X) &= np \\ Var(X) &= np(1-p) \\ M(s) &= (pe^s + 1 - p)^n. \end{aligned}$$

Keeping in the spirit of (1) we denote a binomial n, p r.v. by

$$X \sim bin(n, p).$$

3. **G*eometric distribution*** *with success probability p:* The number of independent Bernoulli p trials required until the first success yields the geometric r.v. with p.m.f.

$$p(k) = \begin{cases} p(1-p)^{k-1}, & \text{if } k \ge 1; \\ 0, & \text{otherwise.} \end{cases}$$

In our coin-flipping context, when consecutively flipping the coin, $p(k)$ denotes the probability that the k^{th} flip is the first flip to land heads (all previous $k-1$ flips land tails). The tail of X has the nice form $\overline{F}(k) = P(X > k) = (1-p)^k,\ k \ge 0$.

It can be shown that

$$\begin{aligned} E(X) &= \frac{1}{p} \\ Var(X) &= \frac{(1-p)}{p^2} \\ M(s) &= \frac{pe^s}{1-(1-p)e^s}. \end{aligned}$$

(In fact, computing $M(s)$ is straightforward and can be used to generate the mean and variance.)

Keeping in the spirit of (1) we denote a geometric p r.v. by

$$X \sim geom(p).$$

Note in passing that $P(X > k) = (1-p)^k,\ k \geq 0$.

Remark 1.3 As a variation on the geometric, if we change X to denote the number of failures before the first success, and denote this by Y, then (since the first flip might be a success yielding no failures at all), the p.m.f. becomes

$$p(k) = \begin{cases} p(1-p)^k, & \text{if } k \geq 0; \\ 0, & \text{otherwise,} \end{cases}$$

and $p(0) = p$. Then $E(Y) = (1-p)p^{-1}$ and $Var(Y) = (1-p)p^{-2}$. *Both of the above are called the geometric distribution, and are related by* $Y = X - 1$.

4. ***Poisson distribution*** *with mean (and variance)* λ *:* With $\lambda > 0$ a constant, X has p.m.f.

$$p(k) = \begin{cases} e^{-\lambda}\frac{\lambda^k}{k!}, & \text{if } k \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

The Poisson distrubution has the interesting property that both its mean and variance are identical $E(X) = Var(X) = \lambda$. Its mgf is given by

$$M(s) = e^{\lambda(e^s - 1)}.$$

The Poisson distribution arises as an approximation to the binomial (n, p) distribution when n is large and p is small: Letting $\lambda = np$,

$$\binom{n}{k} p^k (1-p)^{n-k} \approx e^{-\lambda}\frac{\lambda^k}{k!},\ 0 \leq k \leq n.$$

Keeping in the spirit of (1) we denote a Poisson λ r.v. by

$$X \sim Poiss(\lambda).$$

Continuous case

1. **U*niform distribution*** *on* (a, b)*:* With a and b constants, X has density function

$$f(x) = \begin{cases} \frac{1}{b-a}; & \text{if } x \in (a, b) \\ 0, & \text{otherwise,} \end{cases}$$

c.d.f.

$$F(x) = \begin{cases} \frac{x-a}{b-a}, & \text{if } x \in (a, b); \\ 1, & \text{if } x \geq b; \\ 0, & \text{if } x \leq a, \end{cases}$$

and tail

$$\overline{F}(x) = \begin{cases} \frac{b-x}{b-a}, & \text{if } x \in (a,b); \\ 0, & \text{if } x \geq b; \\ 1, & \text{if } x \leq a. \end{cases}$$

A simple computation yields

$$\begin{aligned} E(X) &= \frac{a+b}{2} \\ Var(X) &= \frac{(b-a)^2}{12} \\ M(s) &= \frac{e^{sb} - e^{sa}}{s(b-a)}. \end{aligned}$$

When $a = 0$ and $b = 1$, this is known as the *uniform distribution over the unit interval*, and has density $f(x) = 1, x \in (0,1)$, $E(X) = 0.5$, $Var(X) = 1/12$, $M(s) = s^{-1}(e^s - 1)$.
Keeping in the spirit of (1) we denote a uniform (a,b) r.v. by

$$X \sim unif(a,b).$$

2. **E*xponential distribution:*** With $\lambda > 0$ a constant, X has density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0; \\ 0, & \text{if } x < 0, \end{cases}$$

c.d.f.

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \geq 0; \\ 0, & \text{if } x < 0, \end{cases}$$

and tail

$$\overline{F}(x) = \begin{cases} e^{-\lambda x}, & \text{if } x \geq 0; \\ 1, & \text{if } x < 0, \end{cases}$$

A simple computation yields

$$\begin{aligned} E(X) &= \frac{1}{\lambda} \\ Var(X) &= \frac{1}{\lambda^2} \\ M(s) &= \frac{\lambda}{\lambda - s}. \end{aligned}$$

The exponential is famous for having the unique *memoryless property*,

$$P(X - y > x | X > y) = P(X > x), \ x \geq 0, \ y \geq 0,$$

in the sense that it is the unique distribution with this property.
(The geometric distribution satisfies a discrete version of this.)
Keeping in the spirit of (1) we denote an exponential λ r.v. by

$$X \sim exp(\lambda).$$

The exponential distribution can be viewed as approximating the distribution of the *time until the first success* when performing an independent $Bern(p)$ trial every Δt units of time with $p = \lambda \Delta t$ and Δt very small; as $\Delta t \to 0$, the approximation becomes exact.

3. ***Normal distribution*** *with mean μ and variance σ^2: $N(\mu, \sigma^2)$:* The normal distribution is extremely important in applications because of the Central Limit Theorem (CLT). With $-\infty < \mu < \infty$ (the mean) and $\sigma^2 > 0$ (the variance) constants, X has density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \ -\infty < x < \infty.$$

This is also called the *Gaussian* distribution. We denote it by $N(\mu, \sigma^2)$. When $\mu = 0$ and $\sigma^2 = 1$ it is called the *standard* or unit normal, denoted by $N(0, 1)$. If Z is $N(0, 1)$, then $X = \sigma Z + \mu$ is $N(\mu, \sigma^2)$. Similarly, if X is $N(\mu, \sigma^2)$, then $Z = (x - \mu)/\sigma$ is $N(0, 1)$. Thus the c.d.f. $F(x)$ can be expressed in terms of the c.d.f. of a unit normal Z. We therefore give the $N(0, 1)$ c.d.f. the special notation $\Theta(x)$;

$$\Theta(x) = P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-y^2}{2}} dy,$$

and we see that

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= P(\sigma Z + \mu \leq x) \\ &= P(Z \leq (x - \mu)/\sigma) \\ &= \Theta((x - \mu)/\sigma). \end{aligned}$$

$\Theta(x)$ does not have a closed form (e.g., a nice formula that we can write down and plug into); hence the importance of good numerical recipes for computing it, and tables of its values.

The moment generating function of $N(\mu, \sigma^2)$ can be shown to be

$$M(s) = e^{s\mu + s^2\sigma^2/2}.$$

Keeping in the spirit of (1) we denote a $N(\mu, \sigma^2)$ r.v. by

$$X \sim N(\mu, \sigma^2).$$

4. L***ognormal distribution:*** If Y is $N(\mu, \sigma^2)$, then $X = e^Y$ is a non-negative r.v. having the *lognormal distribution*; Its natural logarithm $Y = \ln(X)$ yields a normal X has density

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(\ln(x)-\mu)^2}{2\sigma^2}}, & \text{if } x \geq 0; \\ 0, & \text{if } x < 0. \end{cases}$$

Observing that $E(X)$ and $E(X^2)$ are simply the moment generating function of $N(\mu, \sigma^2)$ evaulated at $s = 1$ and $s = 2$ respectively yields

$$\begin{aligned} E(X) &= e^{\mu + \frac{\sigma^2}{2}} \\ Var(X) &= e^{2\mu + \sigma^2}(e^{\sigma^2} - 1). \end{aligned}$$

(It can be shown that $M(s) = \infty$ for any $s > 0$.)

The lognormal distribution plays an important role in financial engineering since it is frequently used to model stock prices. As with the normal distribution, the c.d.f. does not have a closed form, but it can be computed from that of the normal via $P(X \leq x) = P(Y \leq \ln(x))$ due to the relation $X = e^Y$, and we conclude that $F(x) = \Theta((\ln(x) - \mu)/\sigma)$. Thus computations for $F(x)$ are reduced to dealing with $\Theta(x)$, the c.d.f. of $N(0, 1)$.

Keeping in the spirit of (1) we denote a lognormal μ, σ^2 r.v. by

$$X \sim lognorm(\mu, \sigma^2).$$

5. *Pareto distribution:* With constant $a > 0$, X has density

$$f(x) = \begin{cases} ax^{-(1+a)}, & \text{if } x \geq 1; \\ 0, & \text{if } x < 1, \end{cases},$$

c.d.f.

$$F(x) = \begin{cases} 1 - x^{-a}, & \text{if } x \geq 1; \\ 0, & \text{if } x \leq 1, \end{cases}$$

and tail

$$\overline{F}(x) = \begin{cases} x^{-a}, & \text{if } x \geq 1; \\ 1, & \text{if } x \leq 1. \end{cases}$$

(In many applications, a is an integer.) A simple computation yields

$$\begin{aligned} E(X) &= \frac{a}{a-1},\ a > 1;\ (= \infty \text{ otherwise}) \\ Var(X) &= \frac{a}{a-2} - \Big(\frac{a}{a-1}\Big)^2,\ a > 2;\ (= \infty \text{ otherwise}). \end{aligned}$$

(It can be shown that $M(s) = \infty$ for any $s > 0$.)

It is easily seen that $E(X^n) = \infty$ for all $n \geq a$: *The Pareto distribution has infinite moments for high enough n.* The Pareto distribution has the important feature that its tail $\overline{F}(x) = x^{-a}$ tends to 0, as $x \to \infty$, slower than does any exponential tail $e^{-\lambda x}$ or any lognormal tail. It is an example of a distribution with a very *heavy* or *fat* tail. Data suggests that the distribution of stock prices resembles the Pareto more than it does the widely used lognormal.

Keeping in the spirit of (1) we denote a Pareto a r.v. by

$$X \sim Pareto(a).$$

Remark 1.4 Variations on the Pareto distribution exist which allow the mass to start at different locations; $\overline{F}(x) = (c/(c + x))^a$, $x \geq 0$ with $c > 0$ and $a > 0$ constants for example.

1.5 Strong Law of Large Numbers and the Central limit theorem (CLT)

A *stochastic process* is a collection of r.v.s. $\{X_t : t \in T\}$ with index set T. If $T = \{0, 1, 2, \ldots\}$ is the discrete set of integers, then we obtain a sequence of random variables $X_0, X_1, X_2, \ldots$ denoted by $\{X_n : n \geq 0\}$ (or just $\{X_n\}$). In this case we refer to the value X_n as the *state* of the process at time n. For example X_n might denote the stock price of a given stock at the end of the n^{th} day. If time n starts at $n = 1$, then we write $\{X_n : n \geq 1\}$ and so on. If time is continuous (meaning that the index set $T = [0, \infty)$) then we have a continuous-time stochastic process denoted by $\{X_t : t \geq 0\}$.

A very special (but important) case of a discrete-time stochastic process is when the r.v.s. are independent and identically distributed (i.i.d.). In this case there are two classical and fundamental results, the strong law of large numbers (SLLN) and the central limit theorem (CLT):

Theorem 1.1 (SLLN) *If $\{X_n : n \geq 1\}$ are i.i.d. with finite mean $E(X) = \mu$, then w.p.1.,*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mu, \ n \to \infty.$$

One of the practical consequences of the SLLN is that we can, for n large enough, use the approximation

$$E(X) \approx \frac{1}{n}\sum_{i=1}^{n} X_i,$$

when trying to determine an apriori unknown mean $E(X) = \mu$. For this reason the SLLN is fundamental in *Monte Carlo Simulation*.

We now state the central limit theorem:

Theorem 1.2 (CLT) *If $\{X_n : n \geq 1\}$ are i.i.d. with finite mean $E(X) = \mu$ and finite non-zero variance $\sigma^2 = Var(X)$, then*

$$Z_n \stackrel{\text{def}}{=} \frac{1}{\sigma\sqrt{n}}\Big(\sum_{i=1}^{n} X_i - n\mu\Big) \Longrightarrow N(0,1), \ n \to \infty, \ \textit{in distribution};$$

(in other words, $\lim_{n\to\infty} P(Z_n \leq x) = \Theta(x), \ -\infty < x < \infty$, where $\Theta(x)$ is the cdf of $N(0,1)$.)

The CLT allows us to approximate sums of i.i.d. r.v.s. endowed with any c.d.f. F (even if unknown) by the c.d.f. of a normal, as long as the variance of F is finite; it says that for n sufficiently large,

$$\sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2).$$

The famous normal approximation to the binomial distribution is but one example, for a binomial rv can be written as the sum of i.i.d. Bernoulli rvs, and thus the CLT applies.

Sometimes, the CLT is written in terms of the sample average,

$$\overline{X}(n) = \frac{1}{n}\sum_{j=1}^{n} X_j, \tag{12}$$

in which case it becomes

$$Z_n \stackrel{\text{def}}{=} \frac{\sqrt{n}}{\sigma}(\overline{X}(n) - \mu) \Longrightarrow N(0,1).$$

If $\mu = 0$ and $\sigma^2 = 1$, then the CLT simplifies to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \Longrightarrow N(0,1),$$

or equivalently

$$\sqrt{n}\ \overline{X}(n) \Longrightarrow N(0,1).$$

The CLT yields the theoretical justification for the construction of *confidence intervals*, allowing one to say, for example, that

"I am 95% confident that the true mean μ lies within the interval $[22.2, 22.4]$".

We briefly review this next.

1.6 *Confidence intervals for estimating an unknown mean $\mu = E(X)$*

In statistics, we estimate an unknown mean $\mu = E(X)$ of a distribution by collecting n iid samples from the distribution, $X_1, \ldots, X_n$ and using as our approximation the sample mean

$$\overline{X}(n) = \frac{1}{n}\sum_{j=1}^{n} X_j. \tag{13}$$

But how good an approximation is this? The CLT helps answer that question:

Letting $\sigma^2 = Var(X)$ denote the variance of the distribution, we conclude that

$$Var(\overline{X}(n)) = \frac{\sigma^2}{n}. \tag{14}$$

The central limit theorem asserts that as $n \to \infty$, the distribution of

$Z_n \stackrel{\text{def}}{=} \frac{\sqrt{n}}{\sigma}(\overline{X}(n) - \mu)$ tends to $N(0,1)$, the unit normal distribution. Letting Z denote a $N(0,1)$ rv, we conclude that for n sufficiently large,

$Z_n \approx Z$ in distribution. From here we obtain for any $z \geq 0$,

$$P(|\overline{X}(n) - \mu| > z\frac{\sigma}{\sqrt{n}}) \approx P(|Z| > z) = 2P(Z > z).$$

(We can obtain any value of $P(Z > z)$ by referring to tables, etc.)

For any $\alpha > 0$ no matter how small (such as $\alpha = 0.05$), letting $z_{\alpha/2}$ be such that $P(Z > z_{\alpha/2}) = \alpha/2$, we thus have

$$P(|\overline{X}(n) - \mu| > z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \approx \alpha,$$

which can be rewritten as

$$P(\overline{X}(n) - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}(n) + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \approx 1 - \alpha,$$

which implies that the unknown mean μ lies within the interval $\overline{X}(n) \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ with (approximately) probability $1 - \alpha$.

We have thus constructed a *confidence interval* for our estimate:

> *we say that the interval* $\overline{X}(n) \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ *is a* $100(1-\alpha)\%$ *confidence interval for the mean* μ.

Typically, we would use (say) $\alpha = 0.05$ in which case $z_{\alpha/2} = z_{0.025} = 1.96$, and we thus obtain a 95% confidence interval $\overline{X}(n) \pm (1.96)\frac{\sigma}{\sqrt{n}}$.

The length of the confidence interval is $2(1.96)\frac{\sigma}{\sqrt{n}}$ which of course tends to 0 as the sample size n gets larger.

The main problem with using such confidence intervals is that we would not actually know the value of σ^2; it would be unknown (just as μ is). But this is not really a problem: we instead use an estimate for it, the *sample variance* $s^2(n)$ defined by

$$s^2(n) = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2.$$

It can be shown that $s^2(n) \to \sigma^2$, with probability 1, as $n \to \infty$ and that $E(s^2(n)) = \sigma^2,\ n \ge 2$.

So, in practice we would use $s(n)$ is place of σ when constructing our confidence intervals. For example, a 95% confidence interval is given by $\overline{X}(n) \pm (1.96)\frac{s(n)}{\sqrt{n}}$.

The following recursions can be derived; they are useful when implementing a simulation requiring a confidence interval:

$$\overline{X}_{n+1} = \overline{X}_n + \frac{X_{n+1} - \overline{X}_n}{n+1},$$

$$S^2_{n+1} = \Big(1 - \frac{1}{n}\Big)S^2_n + (n+1)(\overline{X}_{n+1} - \overline{X}_n)^2.$$

1.7 Multivariate random variables and joint distributions

With $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}} \in \mathcal{R}^{\mathrm{n}}$, the distribution of a (continuous) random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)^{\mathrm{T}}$ can be described by a joint density function

$$f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n).$$

If the rvs are independent, then this f decomposes into the product of the individual density functions,

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2)\cdots f_n(x_n),$$

where f_i denotes the density function of X_i. In general, however, correlations exist among the n rvs and thus such a simple product form does not hold. The joint cdf is given by

$$F(\mathbf{x}) = P(X_1 \le x_1, \ldots, X_n \le x_n).$$

The moment generating function of a random vector is defined as follows: For $\theta = (\theta_1, \ldots, \theta_n)^{\mathrm{T}}$

$$M_{\mathbf{X}}(\theta) = E(e^{\theta^{\mathrm{T}}\mathbf{X}}) = E(e^{\theta_1 X_1 + \cdots + \theta_n X_n}). \tag{15}$$

A very important class of random vectors appearing in applications are multivariate normals (*Gaussian*), for then the joint distribution is completely determined by the vector of means

$$\mu = (\mu_1, \mu_2, \ldots, \mu_n)^{\mathrm{T}} = E(\mathbf{X}) = (E(X_1), E(X_2), \ldots, E(X_n))^{\mathrm{T}},$$

and the $n \times n$ matrix of covariances,

$$\mathbf{\Sigma} = (\sigma_{\mathbf{i,j}}),$$

where $\sigma_{i,j} = Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$, with $0 < \sigma_{ii} = \sigma_i^2 = Var(X_i),\ i, j \in \{1, 2 \ldots, n\}$. We denote such a normal vector by $\mathbf{X} \sim N(\mu, \mathbf{\Sigma})$. Such a Σ is *symmetric*,

$$\mathbf{\Sigma}^{\mathrm{T}} = \mathbf{\Sigma},$$

and *positive semidefinite*,

$$\mathbf{x}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{x} \geq \mathbf{0},\ \mathbf{x} \in \mathcal{R}^{\mathrm{n}},$$

and any $n \times n$ matrix with those two properties defines a covariance matrix for a multivariate normal distribution. *Positive semidefinite is equivalent to all eigenvalues of* $\mathbf{\Sigma}$ *being non-negative.* The moment generating function (15) has an elegant form analogous to the one-dimensional normal case:

$$M_{\mathbf{X}}(\theta) = E(e^{\theta^{\mathrm{T}}\mathbf{X}}) = e^{\mu^{\mathrm{T}}\theta + \frac{1}{2}\theta^{\mathrm{T}}\mathbf{\Sigma}\theta}. \tag{16}$$

A very nice feature of a Gaussian vector $\mathbf{X}$ is that it can always be expressed as a linear transformation of n iid $N(0,1)$ rvs. If $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)^{\mathrm{T}}$ are iid $N(0,1)$, then there exists a linear mapping (matrix) $\mathbf{A}$ (from $\mathcal{R}^{\mathrm{n}}$ to $\mathcal{R}^{\mathrm{n}}$), such that

$$\mathbf{X} = \mathbf{A}\mathbf{Z} + \mu. \tag{17}$$

In this case $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$. Conversely, *any linear transformation of a Gaussian vector is yet again Gaussian:* If $\mathbf{X} \sim N(\mu, \mathbf{\Sigma})$ and $\mathbf{B}$ is a $m \times n$ matrix, then

$$\mathbf{Y} = \mathbf{B}\mathbf{X} \sim N(\mathbf{B}\mu, \mathbf{B}\mathbf{\Sigma}\mathbf{B}^{\mathrm{T}}) \tag{18}$$

is an m dimensional Gaussian where $m \geq 1$ can be any integer.

For purposes of constructing a desired $\mathbf{X} \sim N(\mu, \mathbf{\Sigma})$ using (17) a solution $\mathbf{A}$ to $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$ is not unique (in general), but it is easy to find one such solution by *diagonalization*:

Because $\mathbf{\Sigma}$ is symmetric with real elements, it can be re-written as $\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$, where $\mathbf{D}$ is the diagonal matrix of the n eigenvalues, and $\mathbf{U}$ is an *orthogonal* matrix ($\mathbf{U}^{\mathrm{T}} = \mathbf{U}^{-\mathbf{1}}$) with columns as the associated eigenvectors. Because $\mathbf{\Sigma}$ is also positive semidefinite, all the eigenvalues must be non-negative yielding the existence of $\sqrt{\mathbf{D}}$ and thus a solution

$$\mathbf{A} = \mathbf{U}\sqrt{\mathbf{D}}. \tag{19}$$

Thus finding an $\mathbf{A}$ reduces to finding the $\mathbf{U}$ and the $\mathbf{D}$. There also exists a *lower triangular* solution to $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$ called the *Cholesky decomposition* of $\mathbf{\Sigma}$. Such a decomposition is more attractive computationally (for algebraically computing $\mathbf{X}$ from $\mathbf{Z}$) since it requires less additions and multiplications along the way. If the matrix $\mathbf{\Sigma}$ has full rank (equivalently, $\mathbf{\Sigma}$ is positive definite (all n eigenvalues are positive) as opposed to only being positive semidefinite), then the Cholesky decomposition can be easily computed iteratively by solving for it directly: Assuming $\mathbf{A}$ is lower triangular, solve the set of equations generated when setting $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$. We will study this in more detail later when we learn how to efficiently simulate multivariate normal rvs.