Statistics I: Chapter 6&7 - Special random variables and distribution of sums of independent random variables

Carlos Oliveira

Office: 511, Quelhas 6
E-mail: carlosoliveira@iseg.ulisboa.pt

ISEG - Lisbon School of Economics and Management

A random variable X has a *discrete uniform distribution* and it is referred to as a discrete uniform random variable if and only if , its probability function is given by

$$f_X(x_j) = \frac{1}{k},\ j = 1, 2, 3, ..., k$$

where $x_j \neq x_i$ for $i \neq j$ $\ D_X = \{x_1, x_2, ..., x_k\}$

Properties: Assuming that $x_1 = 1,\ x_2 = 2,\ \cdots,\ x_k = k$, then

1. $\mu_X = E(X) = \frac{k+1}{2}$
2. $Var(X) = \frac{k^2-1}{12}$
3. $M_X(t) = \sum_{i=1}^{k} e^{tx_i}/k$

Example: Let X be the random variable that represents the number of dots when one rolls a die. Then X follows a discrete uniform distribution taking values $\{1, 2, 3, 4, 5, 6\}$. Its probability function is given by

$$P(X = x) = \begin{cases} \frac{1}{6}, & x = 1, 2, 3, 4, 5, 6 \\ 0, & \text{otherwise} \end{cases}.$$

Expected value:

$$E(X) = \frac{6+1}{2} = \frac{7}{2}$$

Variance:

$$Var(X) = \frac{6^2 - 1}{12} = \frac{35}{12}$$

Moment generating function:

$$M_X(t) = E(e^{tX}) = \frac{1}{6} \sum_{i=1}^{6} e^{i \times t}.$$

The *Bernoulli random variable* takes the value 1 with probability p and the value 0 with probability $1-p$, where $p \in (0,1)$, that is

$$X = \begin{cases} 1 & \text{where } P(X=1) = p \\ 0 & \text{where } P(X=0) = 1-p \end{cases}$$

the probability function is given by

$$f_X(x) = P(X=x) = \begin{cases} p^x(1-p)^{1-x}, & x = 0,1 \\ 0, & \text{otherwise} \end{cases}$$

Properties:

1. $E(X) = p$
2. $Var(X) = p(1-p)$
3. $M_X(t) = (1-p) + pe^t$.

Remark: This random variable is used when the result of the experiment is a success or a failure.

- The *Binomial random variable* is defined as the number of successes in n trials, each of which has the probability of success p.

- *The Binomial random variable:* $X =$ number of successes in n trials. One can show that the probability function is given by

$$f_X(x) = \binom{n}{x} \times p^x(1-p)^{n-x}, \, x = 0, 1, 2, \cdots, n$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

is the number of x combinations from a set with n elements and $k! = k \times (k-1) \times \ldots \times 2 \times 1$.

Remark:

- The parameters of the random variable are n and p.
- If X is a Binomial random variable with parameters n and p we write $X \sim B(n, p)$.
- In the case of the Bernoulli random variable $X \sim B(1, p)$.

Properties:

1. $E(X) = np$,
2. $Var(X) = np(1-p)$
3. $M_X(t) = [(1-p) + pe^t]^n$
4. If $X_i \sim B(1, p)$ and the X_i are independent random variables $\sum_{i=1}^{n} X_i \sim B(n, p)$, that is the sum of n independent Bernoulli random variables with parameter p is a Binomial random variable with parameters r and p.
5. If $X_1 \sim B(n_1, p)$ and $X_2 \sim B(n_2, p)$ and X_1 and X_2 are independent, then $X_1 + X_2 \sim B(n_1 + n_2, p)$

Example: In a given factory, 2% of the produced products have a failure. Let X be the random variable that represents the number of products produced with a failure in a sample of 5 products. Then

$$X \sim B(n = 5, p = 0.02).$$

Expected value:

$$E(X) = n \times p = 5 \times 0.02 = 0.1$$

Variance:

$$Var(X) = n \times p \times (1 - p) = 5 \times 0.02 \times 0.8 = 0.08$$

Probability:

$$P(X = 1) = \binom{5}{1} \times 0.02^1 \times 0.98^4 = 0.092$$

Example: Let X be a random variable that represents the lifetime in hours of a bulb lamp with density function

$$f_X(x) = \frac{1}{100} e^{-x/100}, \text{ for } x > 0$$

Compute the probability that 3 bulb lamps in a sample of 5 have a lifetime smaller 100 hours.

Solution: We can start by computing the probability that 1 bulb lamp has lifetime smaller 100 hours.

$$P(X < 100) = \int_0^{100} \frac{1}{100} e^{-x/100} dx = 1 - e^{-1} = \approx 0.63.$$

Let Y be the random variable that represents the number of bulb lamps in a sample of 5 with a lifetime smaller 100 hours.

$$Y \sim B(n = 5, p = 0.63)$$

The required probability is

$$P(Y = 3) = \binom{5}{3} \times 0.63^3 \times 0.37^2 = 0.3423$$

Example: Suppose that in a group of 1000 computers 50 have a problem in the hardware system. We pick randomly a sample of 100 computers. Let X be the random variable that counts the number of computers with hardware problems.

a) What is the distribution of X if the experiment is done with replacement?

   Answer: X counts the number of successes in a set of 100 computers. The selection is made with replacement.
   $X \sim Bin(100, 1/20)$

   What is the distribution of X if the experiment is done without replacement?

   Answer: X counts the number of successes in a set of 100 computers, but the selection is made without replacement. This means that X does not follow a binomial distribution. Indeed,

$$P(X=x)=\frac{\binom{1000-50}{100-x}\binom{50}{x}}{\binom{1000}{100}}.$$

Consider a finite population of size N that contains exactly M objects with a specific feature. The hyper-geometric distribution is a discrete probability distribution that describes the probability of k successes in n draws (to get k objects with the referred feature), without replacement.

$$X \sim Hypergeometric(N, M, n)$$

$$P(X = k) = \begin{cases} \frac{\binom{N-M}{n-k}\binom{M}{k}}{\binom{N}{n}}, & k = \max\{0, n-(N-M)\}, \cdots, \min\{n, M\} \\ 0, & \text{otherwise} \end{cases}$$

Properties: If $X \sim Hypergeometric(N, M, n)$, then

1. $E(X) = n \times \frac{M}{N}$
2. $Var(X) = n\frac{M}{N}(1 - \frac{M}{N})\frac{N-n}{N-1}$
3. There is no closed form solution for $M_X(t)$

Example: Assume that there are 20 balls in a box, where 2 are green, 8 are blue, 5 are red and 5 yellow. If someone chooses randomly and without replacement 3 balls from the box. Compute the probability that 1 of them is blue.

Solution: Let X be the random variable that counts the number of blue balls in a set of 3 when the experiment is made without replacement.

$$X \sim Hypergeometric(N = 20, M = 8, n = 3)$$

The probability function of this random variable is

$$P(X = x) = \begin{cases} \frac{\binom{8}{x}\binom{12}{2}}{\binom{20}{3}}, & x = 0, 1, 2, 3 \\ 0, & \text{otherwise} \end{cases}$$

The required probability is

$$P(X = 1) = \frac{\binom{8}{1}\binom{12}{2}}{\binom{20}{3}} = \frac{44}{95}.$$

- In connection with repeated Bernoulli trials, we are sometimes interested in the number of the trial on which the k^{th} success occurs.

Assume that $\mathbf{k = 1}$.

- Each trial has two potential outcomes called "success"and "failure". In each trial the probability of success is p and of failure is $(1-p)$.
- We are observing this sequence until the first success has occurred.
- If X is the random variable that counts the number of trials until a success! and the 1^{st} success occurs on the x^{th} trial (the first $x-1$ trials are failures), then X follows a geometric distribution with a probability of success p.

$$X \sim Geo(p)$$

$$P(X = x) = (1-p)^{x-1}p, \quad x = 1, 2, 3, \cdots$$

X is the random variable that counts the number of trials until a success!

Useful Result:

$$P(X > n) = (1-p)^n, \quad n \in \mathbb{N}$$

Therefore, $F_X(n) = 1 - (1-p)^n$, for $n \in \mathbb{N}$

Memoryless property:

$$P(X > n + m \,|\, X > m) = P(X > n), \quad n \in \mathbb{N}$$

Remark: U is a random variable taking values in $\mathbb{N}$ that satisfies the memoryless property iff U has a geometric distribution.

- Assume now that we are observing a sequence of Bernoulli trials until a predefined number k of successes has occurred.
- If the k^{th} success is to occur on the x^{th} trial, there must be $k-1$ successes on the first $x-1$ trials, and the probability for this is

$$P(X=x)=\binom{x-1}{k-1}p^k(1-p)^{x-k}, \quad x=k,k+1,k+2,\cdots$$

where X follows a negative binomial distribution with parameters k and p

$$X \sim NB(k,p)$$

Properties:

1. $E(X)=\frac{k}{p}$
2. $Var(X)=\frac{k}{p}\left(\frac{1}{p}-1\right)$
3. $M_X(t)=\left(\frac{pe^t}{1-e^t(1-p)}\right)^k$

The Poisson random variable is a discrete rv that describes the number of occurrences within a randomly chosen unit of time or space. For example, within a minute, hour, day, kilometer.

The Poisson probability function is a discrete function defined for non-negative integers. If X is a Poisson random variable with parameter λ, we write $X \sim Poisson(\lambda)$. The Poisson distribution with parameter $\lambda > 0$, it is defined by

$$f_X(x) = P\left(X = x\right) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, 2, ..$$

Properties:

1. $E(X) = Var(X) = \lambda$.
2. $M_X(X) = e^{\lambda\left(e^t - 1\right)}$.
3. If $X_i \sim Poisson(\lambda_i)$ and the X_i are independent random variables, then $\sum_{i=1}^n X_i \sim Poisson\left(\sum_{i=1}^n \lambda_i\right)$.

Example: Assume that the number of people that take bus n1 in small city follows a Poisson distribution with $Var(X) = 3$.

Question: What is the probability that in a random day 5 people take bus n1?
Solution: Firstly we should notice that

$$X \sim Poi(\lambda) \text{ and } Var(X) = 3$$

Then $\lambda = 3$. Now we have to compute the probability

$$P(X = 5) = \frac{e^{-3} \times 3^5}{5!} \approx 10\%$$

Question: What is the probability that 5 people take bus n1 in two days?

Solution: Let X_i be the rv that represents the number of people that take bus n1 with $i = 1, 2$. Then

$$X_i \sim Poi(3) \text{ and } X_1 + X_2 \sim Poi(6)$$

Now we have to compute the probability

$$P(X_1 + X_2 = 5) = \frac{e^{-6} \times 6^5}{5!} \approx 16\%$$

The probability density function and the cumulative distribution function of an *exponential random variable* with parameter λ are respectively

$$f_X(x) = \begin{cases} 0 & \textit{if} \quad x < 0 \\ \lambda e^{-\lambda x} & \textit{if} \quad x \geq 0 \end{cases} \qquad F_X(x) = \begin{cases} 0 & \textit{if} \quad x < 0 \\ 1 - e^{-\lambda x} & \textit{if} \quad x \geq 0 \end{cases}$$

Remark: If X is an exponential random variable with parameter λ we write $X \sim Exp(\lambda)$.

Properties: Let X be an exponential random variable. Then,

1. Moment Generating Function $M_X(t) = (1 - t/\lambda)^{-1}$ $t < \lambda$.
2. $E(X) = 1/\lambda$ and $Var(X) = 1/\lambda^2$.
3. Lack of memory: $P(X > x + s | X > x) = P(X > s)$ for any $x \geq 0$ and $s \geq 0$.
4. Let $X_i \sim Exp(\lambda_i)$, $i = 1, 2, ..., k$, be independent random variables, then $Y = \min\{X_1, X_2, ..., X_k\} \sim Exp(\sum_{i=1}^{k} \lambda_i)$

Example: Let X be a random variable that represents the lifetime of an electronic component in years. It is known that X follows an exponential distribution such that

$$P(X > 1) = e^{-1/3}.$$

Question: Knowing that $X \sim Exp(\lambda)$, what is the value of λ?
Solution: Taking into account that

$$P(X > 1) = e^{-1/3} \quad \text{and} \quad P(X > 1) = 1 - F_X(1) = e^{-\lambda}$$

then $\lambda = 1/3$.
Question: What is the probability that the lifetime of the component electronic is grater than 3 years knowing that it is grater than 1 year?
Solution: Given the memoryless property we have that

$$P(X > 3 \,|\, X > 1) = P(X > 2) = e^{-2/3}.$$

Question: Assume that one has 3 similar electronic components that are independent. What is the probability that the lowest lifetime of these electronic components is lower than 2 years?

Solution: Since we have 3 independent and identical components, than we must have 3 random variables X_1, X_2, X_3 representing respectively the lifetime of the electronic component $1, 2, 3$. The lowest lifetime is the random variable

$$Y = \min(X_1, X_2, X_3)$$

According to the properties we have that

$$Y \sim Exp(3 \times 1/3).$$

Therefore,

$$P(Y < 2) = 1 - e^{-2}.$$

Poisson Process:

- N_t represents the number occurrences in the interval $(0, t]$, where $t > 0$. The collection of random variables $\{N_t,\ t > 0\}$ is called a Poisson process with intensity λ if
  a) the number of occurrences in disjoint intervals are independent random variables;
  b) the number of occurrences in intervals of the same size are random variavles with the same distribution are independent random variables;
  c) $N_t \sim Poi(\lambda \times t)$.

Relationship between the Poisson and Exponential distribution:
Let

- N_t be the number occurrences in the interval $(0, t]$, where $t > 0$.
- X_i be the time spent between the two consecutive occurrences $(i - 1)$ and i of the event.

If the collection of random variables $\{N_t,\ t > 0\}$ is a Poisson process, then

$$X_i \sim Exp(\lambda).$$

Example: Assume that X_t represents the number of clients that go to a store in t hours. The average number of clients in two hours is 5. X_t follows a Poisson process.

(1) Compute the probability that in 1 hour at least 2 clients go to the store.

Answer: Firstly, we can notice that $X_t \sim Poisson(\lambda t)$ and $E(X_t) = \lambda t$. Additionally, $E(X_t) = 2\lambda = 5$, meaning that $\lambda = 5/2$. The requested probability is

$$P(X_1 \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - e^{-5/2} - e^{-5/2} 5/2 \approx 0.7127$$

(2) Compute the probability that 5 clients go to the store in 1 hour and a half knowing that no clients went there in the first 30 minutes.

Answer: The requested probability is

$$P(X_{1.5} = 5 | X_{0.5} = 0) = \frac{P(X_{1.5} = 5, X_{0.5} = 0)}{P(X_{0.5} = 0)}$$

Example:

$$\begin{aligned} P(X_{1.5}=5|X_{0.5}=0) &= \frac{P(X_{1.5}=5, X_{0.5}=0)}{P(X_{0.5}=0)} \\ &= \frac{P(X_{1.5}-X_{0.5}=5-0, X_{0.5}=0)}{P(X_{0.5}=0)} \\ &= \frac{P(X_{1.5}-X_{0.5}=5)P(X_{0.5}=0)}{P(X_{0.5}=0)} \\ &= P(X_{1.5}-X_{0.5}=5) = P(X_1=5) \\ &= \frac{e^{-5/2}(5/2)^5}{5!} = 0.0668 \end{aligned}$$

(iii) Compute the probability that the first client arrives 45 minutes after the opening hour.

Let Y be the rv that represents the time spent until the first client arrives. $Y \sim Exp(5/2)$

$$P(Y \geq 3/4) = 1 - F_Y(3/4) = 1 - (1 - e^{-5/2*3/4}) \approx 0.1534.$$

Question: How to model the time between two or three or more occurrences in a Poisson Process?

Gamma distribution: The *gamma cumulative distribution* function is defined for $x > 0$, $a > 0$, $b > 0$, by the integral

$$F_X(x) = \frac{1}{b^a \Gamma(a)} \int_0^x u^{a-1} e^{-\frac{u}{b}} du$$

where $\Gamma(t) = \int_0^\infty e^{-u} u^{t-1} du$ is the Gamma function. The parameters a and b are called the shape parameter and scale parameter, respectively.

The probability density function for the gamma distribution is

$$f_X(x) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-\frac{x}{b}}$$

Remarks:

1. If X is a gamma random variable with parameters a and b we write $X \sim Gamma(a, b)$
2. if $a = 1$ and $\frac{1}{b} = \lambda$, $X \sim Exp(\lambda) = Gamma(1, \frac{1}{\lambda})$.
3. **Important case:** When $a = v/2$ and $b = 2$ we have the chi-squared distribution which has the notation $\chi^2(v) = Gamma(v/2, 2)$. v is known as degrees of freedom.

Relationship between the Poisson and Gamma distribution: Let

- N_t be the number occurrences in the interval $(0, t]$, where $t > 0$.
- X_i be the time spent between the two consecutive occurrences $(i-1)$ and i
- $Y_{i,n}$ be the time spent between the occurrences $(i-n)$ and i of the event.

If the collection of random variables $\{N_t,\ t > 0\}$ is a Poisson process, then

$$X_i \sim Exp(\lambda) = Gamma\left(1, \frac{1}{\lambda}\right) \quad \text{and} \quad Y_{i,n} \sim Gamma\left(n, \frac{1}{\lambda}\right).$$

Properties: Let X be a Gamma distribution with parameters a and b.

1. The Moment generating function of the Gamma distribution is given by: $M_X(t) = (1 - bt)^{-a}$ for $t < 1/b$
2. $E(X) = ab$.
3. $Var(X) = ab^2$.
4. Let X_1, X_2,..., X_n be independent random variables with Gamma distribution $X_i \sim Gamma(a_i, b)$, $i = 1, ..., n$, then $\sum_{i=1}^{n} X_i \sim Gamma(\sum_{i=1}^{n} a_i, b)$.
5. If $n \in \mathbb{N}$, then $X \sim Gamma(n, b)$, then $2X/b \sim \chi^2(2n)$

In the case of the chi-squared random variables we have:

1. $E(X) = v$.
2. $Var(X) = 2v$.
3. Let X_1, X_2,...,X_k be independent random variables with Chi-squared distribution $X_1 \sim \chi^2(v_1)$ and $X_2 \sim \chi^2(v_2), ..., X_k \sim \chi^2(v_k)$, then $\sum_{i=1}^{k} X_i \sim \chi^2\left(\sum_{i=1}^{k} v_i\right)$.

Exercise 13: Compute the following probabilities:

1. If Y is distributed $\chi^2(4)$ find $P(Y \leq 7.78)$.
2. If Y is distributed $\chi^2(10)$ find $P(Y > 18.31)$.
3. If Y is $\chi^2(1)$ find $P(Y \leq 3.\,8416)$.

Exercise 14: Using the moment generating function, show that if $X \sim Gamma(a, b)$ and $Y = 2X/b$, then $Y \sim \chi^2(2a)$.

The probability density function of the *uniform random variable* on an interval (a,b), where $a<b$, is the function

$$f_X(x)=\begin{cases} 0 & if \quad x\leq a \\ \frac{1}{b-a} & if \quad a<x<b \\ 0 & if \quad b\leq x \end{cases}$$

The cumulative distribution function is the function

$$F_X(x)=\begin{cases} 0 & if \quad x\leq a \\ \frac{x-a}{b-a} & if \quad a\leq x\leq b \\ 1 & if \quad b\leq x \end{cases}$$

Remark: If X is a *uniform random variable* in the interval (a,b) we write $X\sim U(a,b)$.

1. The moment generating function

$$M_X\left(t\right)=\begin{cases}\frac{e^{tb}-e^{ta}}{t(b-a)} & if \quad t\neq 0\\ 1 & if \quad t=0.\end{cases}$$

   (The moment-generating function is not differentiable at zero, but the moments can be calculated by differentiating and then taking $\lim_{t\to 0}$)

2. Moments about the origin

$$E(X^k)=\frac{b^{k+1}-a^{k+1}}{(b-a)(k+1)},k=1,2,3,\ldots$$

3. $E(X)=\left(a+b\right)/2.$
4. $Var(X)=(b-a)^2/12.$
5. $Skewness=\gamma_1=0.$

Example: Let X be a continuous uniform random variable in the interval $(2, 10)$. Compute the following probabilities:

Questions: Compute the following probabilities:

- $$P(X > 5) = 1 - F_X(5) = 1 - \frac{5-2}{10-2} = \frac{5}{8}$$
- $$P(X < 9 | X > 5) = \frac{P(5 < X < 9)}{1 - F_X(5)} = \frac{F_X(9) - F_X(5)}{1 - F_X(5)}$$ $$= \frac{7/8 - 3/8}{5/8} = \frac{2}{5}$$
- $$E(X) = \frac{10+2}{2} = 6 \quad Var(X) = \frac{(10-2)^2}{12} = \frac{16}{3}$$

Inverse transform sampling: Important result in simulation. This result shows us that, in certain conditions, $Y = F_X(X) \sim U(0,1)$ and if $Y \sim U(0,1)$ then $F_X^{-1}(Y) \sim F_X(x)$.

Example: Assume that X follows an exponential distribution with parameter 1. Find the distribution of $Y = F_X(X)$

The most famous continuous distribution is the *normal distribution* (introduced by Abraham de Moivre, 1667-1754). The normal probability density function is given by

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

The cumulative distribution function does not have a close form solution:

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

When a random variable X follows a normal distribution with parameters μ and σ^2 we write $X \sim N(\mu, \sigma^2)$.

Properties:

1. Moment generating function $M_X(t) = e^{\left(\mu t + 0.5\sigma^2 t^2\right)}$
2. $E(X) = \mu$.
3. $Var(X) = \sigma^2$

There is no closed form solution to the CDF of a normal distribution, which means that one has to use an adequate software to compute the probabilities. Alternatively, one may use the tables with probabilities for the normal distribution with mean equal to 0 and variance equal to 1. To use this strategy one has to notice that

$$X \sim N(\mu, \sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$$

When $\mu = 0$ and $\sigma^2 = 1$, the distribution is denoted as standard normal distribution.

The probability density function of the standard normal distribution is denoted $\phi(z)$ and it is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

The standard normal cumulative distribution function is denoted as

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \phi(t)\, dt.$$

Properties of the standard normal cumulative distribution function:

- $P(Z > z) = 1 - \Phi(z)$.
- $P(Z < -z) = P(Z > z)$.
- $P(|Z| > z) = 2[1 - \Phi(z)]$, for $z > 0$.

Examples: Assume that the weight of a certain population is modeled by a normal distribution with a mean 50 Kg and a standard deviation 5kg.

Question: What is the probability that someone weighs more than 65kg?

Solution: Let X be the random variable that represents the weight of a certain person in the given population. The required probability is

$$P(X > 65) = P\left(\frac{X-\mu}{\sigma} > \frac{65-\mu}{\sigma}\right) = P(Z > 3) = 1 - \Phi(3)$$
$$= 1 - 0.9987 = 0.0013 = 0.13\%,$$

where

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1).$$

Question: What is the weight that is exceeded by 80% of the population?

Solution: We want to find the level W such that $P(X > W) = 80\%$.

$$0.80 = P(X > W) = P\left(\frac{X-\mu}{\sigma} > \frac{W-\mu}{\sigma}\right) = P\left(Z > \frac{W-50}{5}\right).$$

Now, taking into account the shape of the normal density function we know that the threshold $\frac{W-50}{5} < 0$. Therefore, noticing that

$$0.80 = P\left(Z > \frac{W-50}{5}\right) \Leftrightarrow 0.20 = P\left(Z < \frac{W-50}{5}\right)$$

$$0.20 = P\left(Z > -\frac{W-50}{5}\right)$$

one may easily check at the tables that

$$-\frac{W-50}{5} = 0.842 \Leftrightarrow W = 45.79$$

Exercise 20: A baker knows that the daily demand for a specific type of bread is a random variable X such that $X \sim N(\mu = 50, \sigma^2 = 25)$. Find the demand which has probability 1% of being exceeded.

Theorem: (Linear combinations of Normal random variables): Let X and Y be two independent random variables such that $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$. Let $V = aX + bY + c$, then

$$V \sim N(\mu_V, \sigma_V^2)$$

where

$$\begin{aligned} \mu_V &= a\mu_X + b\mu_Y + c \\ \sigma_V^2 &= a^2\sigma_X^2 + b^2\sigma_Y^2. \end{aligned}$$

Remarks:

- A special case is obtained when $b = 0$, if $V = aX + c$, then $V \sim N(\mu_V, \sigma_V^2)$ where $\mu_V = a\mu_X + c$, $\sigma_V^2 = a^2\sigma_X^2$.
- if $X \sim N\left(\mu, \sigma^2\right)$, $Z = \frac{X-\mu}{\sigma} \sim N\left(0, 1\right)$.

Example: Let X and Y be two independent random variables such that

$$X \sim N(\mu = 10, \sigma^2 = 4) \quad \text{and} \quad Y \sim N(\mu = 12, \sigma^2 = 5).$$

Question: Compute the following probability

$$P(X + Y > 19).$$

Solution: Firstly, we notice that

$$X + Y \sim N(22, 9).$$

Therefore,

$$P\left(\frac{X + Y - 22}{3} > \frac{19 - 22}{3}\right) = P(Z > -1) = \Phi(1) = 0.8413,$$

where

$$Z = \frac{X + Y - 22}{3} \sim N(0, 1).$$

Theorem: If the random variable $X_i, i = 1, ..., n$ have a normal distribution, $X_i \sim N\left(\mu_i, \sigma_i^2\right)$, and are independent, then

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

- Assuming that $\mu_i = \mu_X$ and $\sigma_i^2 = \sigma_X^2$, for $i = 1, ..., n$ we have

$$\sum_{i=1}^{n} X_i \sim N(n\mu_X, n\sigma_X^2).$$

Thus

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu_X, \sigma_X^2/n).$$

If we standardize we have

$$Z = \frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}} \sim N(0, 1)$$

We have seen the following result:

If $X_1, X_2, \cdots X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$ then the following holds true:

i)

$$\sum_{i=1}^{n} X_i \sim N(\mu n, \sigma^2 n) \quad \text{or equivalently} \quad \overline{X} \sim N(\mu, \sigma^2/n)$$

ii)

$$\frac{\sum_{i=1}^{n} X_i - \mu n}{\sigma\sqrt{n}} \sim N(0,1) \quad \text{or equivalently} \quad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

However, what happens if the $X_i's$ are not normally distributed? The answer is given by the Central Limit Theorem:

Theorem: (*The Central Limit Theorem* - Lindberg-Levy)

Assume that X_i, $i = 1, ..., n$ are independent, $E(X_i) = \mu_X$, and $Var(X_i) = \sigma_X < +\infty$, then the distribution of

$$Z = \frac{\sum_{i=1}^{n} X_i - n\mu_X}{\sigma_X \sqrt{n}} = \frac{\sqrt{n}\left(\overline{X} - \mu_X\right)}{\sigma_X}$$

converges to a standard normal distribution as n tends to infinity. We write $Z \overset{a}{\sim} N(0,1)$ where the symbol $\overset{a}{\sim}$ reads "distributed asymptotically"

Remarks:

- This means that if the sample size is large enough ($n \geq 30$), then the distribution of Z is close to the standard normal.
- The previous result is useful when X_i, with $i = 1, ..., n$ do not follow a normal distribution (in this case we know the exact distribution of Z).

Assume that X represents the profit of a store in thousands of euros in a random day. The density function of X is given by

$$f_X(x) = \begin{cases} x, & 0 < x < 1 \\ 2 - x, & 1 < x < 2 \end{cases}$$

Compute the probability that the store has a profit greater than 29 thousands of euros in a month (30 days).

Solution: We start by noticing that

$$E(X) = \int_0^1 x^2 dx + \int_1^2 2x - x^2 dx = 1$$
$$E(X^2) = \int_0^1 x^3 dx + \int_1^2 2x^2 - x^3 dx = 7/6$$
$$Var(X) = 7/6 - 1 = 1/6$$

Solution: Assume that X_i represents the profit of a store in thousands of euros in day $i = 1, 2, \cdots, 30$, then we want to compute the following probability:

$$P\left(\sum_{i=1}^{30} X_i > 29\right)$$

By using the central limit theorem we know that

$$Z = \frac{\sum_{i=1}^{30} X_i - 30}{\sqrt{30/6}} \overset{a}{\sim} N(0,1)$$

Therefore

$$P\left(\sum_{i=1}^{30} X_i > 29\right) = P\left(\frac{\sum_{i=1}^{30} X_i - 30}{\sqrt{30/6}} > \frac{29-30}{\sqrt{30/6}}\right)$$
$$= P(Z > -0.45) \approx \Phi(0.45) = 0.6736$$

A special case of the Central Limit Theorem of Lindberg-Levy is the Central Limit Theorem of De Moivre-Laplace, which corresponds to the case that each X_i is Bernoulli with parameter $p = P(X_i = 1)$.

Theorem: (The Central Limit Theorem - De Moivre-Laplace) If the X_i, $i = 1, ..., n$ are independent Bernoulli random variables with $p = P(X_i = 1) \in (0, 1)$ then

$$Z = \frac{\sqrt{n}\left(\overline{X} - p\right)}{\sqrt{p(1-p)}}$$

converges to a standard normal distribution as n tends to infinity. We write $Z \overset{a}{\sim} N(0, 1)$.

Example: Assume that a person is infected with a virus with probability 0.05. If we analyze 100 people, what is the probability that at least 7 are infected?

Solution: Let X_i be a random variable defined by

$$X_i = \begin{cases} 1, & \text{if person } i \text{ is infected} \\ 0, & \text{otherwise} \end{cases}$$

with $i = 1, \cdots, 100$. Therefore, X_i is a Bernoulli random variable. Assuming independence between the rv, we have that

$$\sum_{i=1}^{100} X_i \overset{a}{\sim} N(5, 4.75)$$

Therefore,

$$P\left(\sum_{i=1}^{100} X_i \geq 7\right) = P\left(\frac{\sum_{i=1}^{100} X_i - 5}{\sqrt{4.75}} \geq \frac{7-5}{\sqrt{4.75}}\right) \approx 1 - \Phi(0.92)$$

$$= 1 - 0.8212 = 0.1788$$

Exercise 21: Assume that X_i, with $i = 1, 2, 3$ represent the profit, in million of euros, of 3 different companies located in 3 different countries. If

$$X_1 \sim N(1, 0.01), \quad X_2 \sim N(1.5, 0.03), \quad X_3 \sim N(2, 0.06)$$

1. Which company is more likely to have a profit greater than 1.5 millions?
2. What is the probability of the profit of these 3 companies does not exceed 4 millions of euros? (Assume independence.)