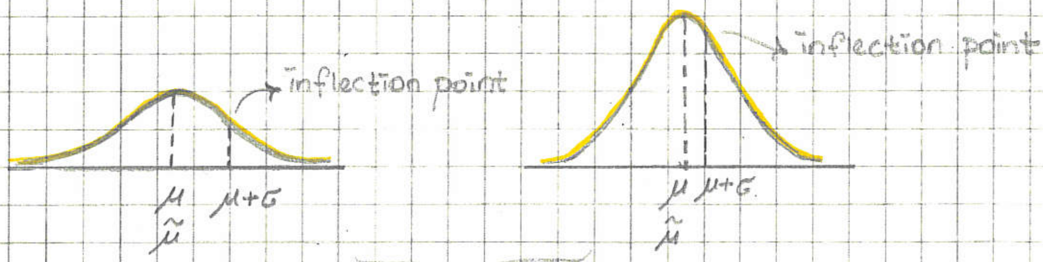
4.3. The Normal Distribution

Definition. A continuous rv X is said to have a normal distribution with parameters μ and σ, where $-\infty < \mu < \infty$ and $\sigma > 0$, if the pdf of X is

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

▽ NOTATION. $X \sim N(\mu,\sigma)$: X has a normal dist. with parameters μ and σ.



Normal density curves

We have

1. $f(x;\mu,\sigma) \geq 0$
2. $\int_{-\infty}^{\infty} f(x;\mu,\sigma)\,dx = 1$
3. $E(X) = \mu$
4. $V(X) = \sigma^2$

▽ One of the first applications of $N(\mu,\sigma)$ was given by Gauss in 1809. Gauss used it to model the errors of observations in astronomy. For this reason, the normal dist. is sometimes called the Gaussian distribution.

The standard normal distribution

$X \sim N(\mu,\sigma)$ with zero mean and unit variance is said to be a standard normal rv, and is denoted by Z.

Definition. The pdf of Z, which is denoted by $\phi(z)$, is

$$\phi(z) = f(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad , \; -\infty < z < \infty .$$

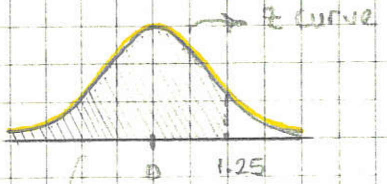
The cdf of Z, which is denoted by $\Phi(z)$, is

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

⊛ The $(100p)$th percentile of the standard normal dist., for any p between 0 and 1, is the solution to the equation $\Phi(z) = p$. Then, we may write $\eta_p = z = \Phi^{-1}(p)$.

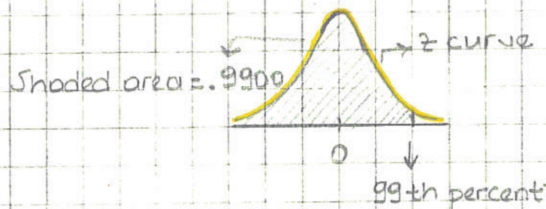
⊛ In order to find the cdf and $(100p)$th percentile, software or a z table can be used.

Example. Consider $P(Z \le 1.25)$. Then, $P(Z \le 1.25) = \Phi(1.25) = .8944$ by z table.



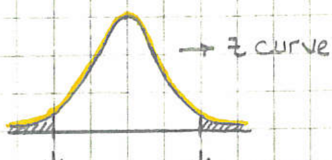
Shaded area = $\Phi(1.25)$

Example. Consider the 99th percentile of the standard normal dist. This means that the area under the curve to the left of the horizontal axis, that is 99th percentile, is .9900.



Now, we have the area, however we do not know the corresponding value of z. Namely, $P(Z \le z) = .9900$. This is called the inverse problem to $P(Z \le z) = ?$

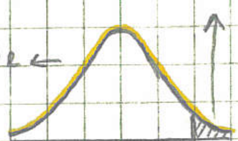
To find z, we can use the z-table again. Therefore, we find that the 99th percentile is (approximately) $z = 2.33$. By symmetry, the 1st percentile is -2.33.



∇_0 If p does not appear inside the z table, the number closest to it is often used.

z_α notation

We use z_α to represent the z-value (or critical value) that corresponds to a specific tail area under the standard normal curve. That is, $z_\alpha = \Phi^{-1}(1-\alpha)$.



For instance, ⊛ $z_{.10} = .10$ (upper tail area is .10)

⊛ $z_{.01} = .01$ (" " " " .01)

The z_α is the $100(1-\alpha)$th percentile of the standard normal distribution.

Nonstandardized normal distributions

Proposition. If $X \sim N(\mu, \sigma)$, then the "standardized" rv Z defined by

$$Z = \frac{X-\mu}{\sigma}$$

$\rightarrow E[Z] = 0 \rightarrow$ centers the dist. at 0

$\rightarrow SD[Z] = 0 \rightarrow$ scales the spread to 1.

has a standard normal distribution. Thus,

$$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right),$$

$$P(X \le a) = \Phi\left(\frac{a-\mu}{\sigma}\right),$$

$$P(X \ge b) = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right),$$

and the $(100p)$th percentile of the $N(\mu, \sigma)$ distribution is given by

$$\eta_p = \mu + \Phi^{-1}(p) \cdot \sigma.$$

Conversely, if $Z \sim N(0,1)$ and μ and σ are constants (with $\sigma > 0$), then the "unstandardized" rv $X = \mu + \sigma \cdot Z$ has a normal distribution with mean μ and standard deviation σ.

Proof. Let $X \sim N(\mu, \sigma)$. Then the cdf of $Z = \frac{X-\mu}{\sigma}$ is given by

$$\begin{aligned}
F_Z(z) &= P(Z \le z) \\
&= P\left(\frac{X-\mu}{\sigma} \le z\right) \\
&= P(X \le \mu + z\sigma) \\
&= \int_{-\infty}^{\mu+z\sigma} f(x; \mu, \sigma)\,dx \\
&= \int_{-\infty}^{\mu+z\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \\
&= \int_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{u^2}{2}} \sigma\,du \\
&= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\,du \\
&= \Phi(z)
\end{aligned}$$

By applying the substitution: $u = \frac{x-\mu}{\sigma}$

$$u = \frac{x-\mu}{\sigma}, \quad du = \frac{du}{\sigma}, \quad \sigma\,du = dx$$

and

$x = \mu + z\sigma \Rightarrow u = z$

$x \to -\infty \Rightarrow u \to -\infty$

This shows that the cdf of $\frac{X-\mu}{\sigma}$ is the standard normal cdf $\Phi(z)$, so $\frac{X-\mu}{\sigma} \sim N(0,1)$.

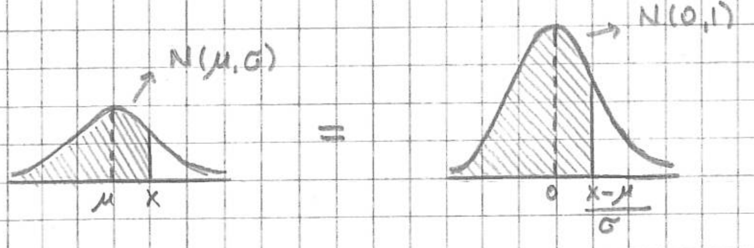
From this main result, we obtain

$$p = P(X \leq \eta_p) = P\left(\frac{X-\mu}{\sigma} \leq \frac{\eta_p - \mu}{\sigma}\right) = \Phi\left(\frac{\eta_p-\mu}{\sigma}\right)$$

$$\Rightarrow \frac{\eta_p - \mu}{\sigma} = \Phi^{-1}(p)$$

$$\Rightarrow \eta_p = \mu + \Phi^{-1}(p)\sigma$$

The converse statement $Z \sim N(0,1) \Rightarrow X = \mu + \sigma Z \sim N(\mu,\sigma)$ is derived similarly.



<u>Example.</u> The return on a diversified investment portfolio is normally distributed. What is the probability that the return is within one standard deviation of its mean value?

<u>Solution.</u> $P\left(\begin{array}{l}X \text{ is within one standard} \\ \text{deviation of its mean}\end{array}\right) = P(\mu - \sigma \leq X \leq \mu + \sigma)$

$$= P\left(\frac{\mu-\sigma-\mu}{\sigma} \leq Z \leq \frac{\mu+\sigma-\mu}{\sigma}\right)$$

$$= P(-1 \leq Z \leq 1)$$

$$= \Phi(1) - \Phi(-1)$$

$$= .8413 - .1587$$
$$= .6826$$

Similarly,

$P\left(\begin{array}{l}X \text{ is within two stand.} \\ \text{dev.s of its mean}\end{array}\right) = P(-2 \leq Z \leq 2)$

$$= \Phi(2) - \Phi(-2)$$

$$= .9544$$

$P\left(\begin{array}{l}X \text{ is within three stand.} \\ \text{dev.s of its mean}\end{array}\right) = P(-3 \leq Z \leq 3)$

$$= \Phi(3) - \Phi(-3)$$

$$= .9973$$

Based on this example, we can give the next rule.

Empirical Rule. If the population dist. of a variable is (approximately) normal, then

1. Roughly 68% of the values are within 1 SD of the mean.
2. Roughly 95% of the values are within 2 SD of the mean.
3. Roughly 99.7% of the values are within 3 SD of the mean.

We will use these results in the development of hypothesis-testing procedures.

The normal MGF

Proposition. The mgf of a normally distributed rv X is

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

Proof. Considering the def. of mgf, we can write

$$M_Z(t) = E(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2-2tz}{2}} dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2-2tz}{2} - \frac{t^2}{2} + \frac{t^2}{2}} dz$$

$$= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2-2tz+t^2}{2}} dz$$

$$= e^{\frac{t^2}{2}} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-t)^2}{2}} dz}_{1}$$

$$= e^{\frac{t^2}{2}}.$$

Now, let X be any normal rv with mean μ and standard deviation σ. We know that $\frac{X-\mu}{\sigma} = Z$, where Z is standard normal. By rewriting it, $X = \mu + Z\sigma$.

Considering this equation and the fact that $M_{aY+b}(t) = e^{bt} M_Y(at)$, we get

$$M_X(t) = M_{\sigma Z+\mu}(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{\frac{\sigma^2 t^2}{2}} = e^{\mu t + \frac{\sigma^2 t^2}{2}}. \blacksquare$$

Proposition. Let $X \sim N(\mu, \sigma)$. Then for any constants a and b with $a \neq 0$, $aX+b$ is also normally distributed. That is, any linear rescaling of a normal rv is normal.

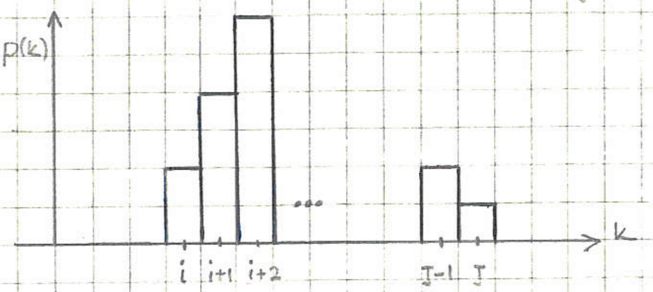
→ If $X \sim N(\mu, \sigma)$ and $Y = aX+b$ $(a \neq 0)$, then Y is also normal with $\mu_Y = a\mu_X + b$ and $\sigma_Y = |a|\sigma_X$.

Subject: Date: ..../..../....

The normal distribution and discrete populations

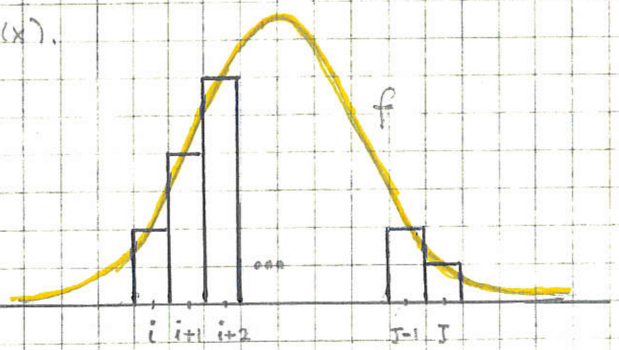
The normal distribution is often used as an approximation to the distribution of values in a discrete population.

Let X be a discrete rv with pmf p(x), and suppose that we want to find $P(i \leq X \leq J)$, $i \leq J$. Consider the figure below.



Here, the base of each rectangle equals 1, and the height with the base midpoint k is p(k), $i \leq k \leq J$. Then, each area is p(k). Then, the sum of the areas of all rectangles is $\sum_{k=i}^{J} p(k)$, which is the exact value of $P(i \leq k \leq J)$.

Now, suppose that pdf f(x) of a continuous rv is a good approximation to p(x).



As this figure shows, $P(i \leq k \leq J)$, the sum of all the areas is approximately the area under f(x) from $i-\frac{1}{2}$ to $J+\frac{1}{2}$ rather than from i to J.

$$P(i \leq k \leq J) \approx \int_{i-1/2}^{J+1/2} f(x)\,dx.$$

The adjustment in the limits of integration is called continuity correction. It is necessary when approximating a discrete dist. with a continuous one.

Example. IQ (as measured by a standard test) is known to be approximately normally distributed with $\mu = 100$ and $\sigma = 15$. What is the probability that a randomly selected individual has an IQ of at least 125?

Solution. Let X denote the IQ of a randomly chosen person. We wish to find $P(X \geq 125)$. Here, the data type (the IQ population) is discrete since IQs are integer-valued. However, we are given it is approx. norm. distributed. Then we are interested in the area from 124.5. By $z = \frac{124.5 - 100}{15} = 1.63$. Thus,

$P(Z \geq 1.63) = P(Z \leq -1.63) = .0516$.

by z table.

Subject: Date: ... / ... / ...

Approximating the binomial distribution

While a normal approximation may be a good approximation to a binomial dist. that is not skewed or only mildly skewed, it is not a good approximation for a binomial distribution that is heavily skewed.

Proposition. Let X be a binomial rv based on n trials with success probability p. Then if the binomial probability histogram is not too skewed, X has approximately a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$. In particular, for x = a possible value of X, $q = 1-p$.

$$P(X \le x) = B(x; n, p) \approx (\text{area under the normal curve to the left of } x+.5)$$

$$= \Phi\left(\frac{x+.5-np}{\sqrt{npq}}\right)$$

In practice, the approximation is adequate provided that both $np \geq 10$ and $nq \geq 10$.

If $np < 10$ or $nq < 10$, the binomial distribution may be too skewed for the normal curve to give accurate approximations.

Example. Suppose that 25% of all licensed drivers in a state do not have insurance. Let X be the number of uninsured drivers in a random sample of size 50 (somewhat perversely, a success is an uninsured driver).

a) Use the normal approximation to the binomial to find $P(X \le 10)$.

b) Use the same approximation to find $P(5 \le X \le 15)$.

c) Compare your results with the exact binomial probabilities.

Solution. We are given $n = 50$ and $p = .25$.

a) We have $\mu = np = 50(.25) = 12.5$, $\sigma = \sqrt{np(1-p)} = \sqrt{50(.25)(.75)} = 3.062$.

Since we found $np = 12.5 > 10$ and $n(1-p) = 50(.75) = 37.5 \geq 10$, the normal approximation can be used.

$$P(X \le 10) = B(10; 50, .25) \approx \Phi\left(\frac{10+.5-12.5}{3.062}\right) = \Phi(-.65) = .2578.$$

b) $P(5 \le X \le 15) = B(15; 50, .25) - B(4; 50, .25) \approx \Phi\left(\frac{15+.5-12.5}{3.062}\right) - \Phi\left(\frac{4+.5-12.5}{3.062}\right) = .8320$

→ NEXT PAGE

Subject: Date: ..../..../....

c) The exact probability of $P(X \leq 10)$ is

$$P(X \leq 10) = \sum_{x=0}^{10} \binom{50}{x} (.25)^x (.75)^{50-x} \underset{\text{by software}}{=} .2622$$

The exact probability of $P(5 \leq X \leq 15)$ is

$$P(5 \leq X \leq 15) = P(X \leq 15) - P(X \leq 4) \underset{\text{by software}}{=} .8348$$

Then, the approximations in (a) and (b) are quite good compared to the exact values above. From $P(5 \leq X \leq 15)$, we see that we can use the continuity correction for both upper and lower limits.

Example. Consider that substrate concentration (mg/cm^3) of influent to a reactor is normally distributed with $\mu = .30$ and $\sigma = .06$.

a) What is the probability that the concentration exceeds $.25$?

b) How would you characterise the largest 5% of all concentration values?

Solution. We are given $X \sim N(.30, .06)$.

a) Here, we wish to find $P(X > .25)$. By the proposition in page 3, we have

$$P(X \geq b) = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right).$$

Then, we obtain

$$P(X > .25) = 1 - \Phi\left(\frac{.25 - .30}{.06}\right) = 1 - \Phi(-.83) \underset{\text{by z table}}{=} 1 - .2033 = .7967.$$

b) We want the 95th percentile:

$$.95 = \Phi\left(\frac{\eta_{.95} - .30}{.06}\right) \Rightarrow \frac{\eta_{.95} - .30}{.06} = -1.645 \Rightarrow \eta_{.95} = .29013.$$ Then, the highest 5% of concentrations are those greater than $.29013$ mg/cm^3.

4.4. The Gamma Distribution and Its Relatives

The gamma family is a class of pdfs that can generate a wide variety of skewed distributional shapes. To define the family of gamma distributions, we first introduce the gamma function.

Definition. For $\alpha > 0$, the gamma function $\Gamma(\alpha)$ is defined by

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

Some properties of the gamma function:

1. For any $\alpha > 1$, $\Gamma(\alpha) = (\alpha-1) \cdot \Gamma(\alpha-1)$
2. For any positive integer, n, $\Gamma(n) = (n-1)!$
3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Proposition. For any $\alpha, \beta > 0$,

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

Proof. Applying the substitution $u = x/\beta \Rightarrow \beta du = dx$, we get

$$\int_0^{\infty} x^{\alpha-1} e^{-x/\beta} dx = \int_0^{\infty} (\beta u)^{\alpha-1} e^{-u} \beta du = \beta^{\alpha} \int_0^{\infty} u^{\alpha-1} e^{-u} du \underset{\text{by the def. of the gamma func.}}{=} \beta^{\alpha} \Gamma(\alpha).$$

The family of gamma distributions

Definition. A cont. rv X is said to have a gamma distribution if the pdf of X is

$$f(x; \alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \quad , x > 0$$

where the parameters α and β satisfy $\alpha > 0$, $\beta > 0$. When $\beta = 1$, X is said to have a standard gamma distribution, and its pdf may be denoted $f(x;\alpha)$.

We have:

1. $f(x;\alpha,\beta) \geq 0$
2. When $\alpha \leq 1 \Rightarrow f(x;\alpha)$ strictly decreases as x increases
   When $\alpha > 1 \Rightarrow f(x;\alpha)$ rises to a maximum and then decreases
3. The parameter β in the definition above is called a scale parameter.

Proposition. The mgf of a gamma rv is

$$M_X(t) = \frac{1}{(1-\beta t)^{\alpha}}.$$

Proof. By definition, the mgf is

$$M_X(t) = E(e^{tX}) = \int_0^{\infty} e^{tx} \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{-x/\beta}\,dx$$

$$= \int_0^{\infty} \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} e^{-x(-t+1/\beta)}\,dx$$

$$= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} x^{\alpha-1} e^{-(-t+1/\beta)x}\,dx$$

By the previous proposition: $\int_0^{\infty} x^{\alpha-1} e^{-\frac{x}{\beta}}\,dx = \beta^{\alpha}\Gamma(\alpha)$.

$$= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \cdot \Gamma(\alpha)\left(\frac{1}{-t+1/\beta}\right)^{\alpha}$$

$$= \frac{1}{(1-\beta t)^{\alpha}}. \blacksquare$$

Proposition. The mean and variance of a rv X having the gamma distribution $f(x;\alpha,\beta)$ are $E(X) = \mu = \alpha\beta$ and $V(X) = \sigma^2 = \alpha\beta^2$.

Definition. When X is a standard gamma rv, the cdf of X, which is

$$G(x;\alpha) = \int_0^{x} \frac{y^{\alpha-1} e^{-y}}{\Gamma(\alpha)}\,dy, \quad x > 0$$

is called the incomplete gamma function.

Proposition. Let X have a gamma distribution with parameters α and β. Then for any $x > 0$, the cdf of X is given by

$$P(X \le x) = G\left(\frac{x}{\beta};\alpha\right),$$

the incomplete gamma function evaluated at x/β.

Example. Let X have a standard gamma distribution with $\alpha = 7$. Find

a) $P(X > 8)$ b) $P(3 \le X \le 8)$ c) $P(X < 4 \text{ or } X > 6)$.

Solution. a) $P(X > 8) = 1 - P(X \le 8) = 1 - G(8;7)$ by Table 4.

b) $P(3 \le X \le 8) = G(8;7) - G(3;7) = .653$

c) $P(X < 4 \text{ or } X > 6) = 1 - P(4 \le X \le 6) = 1 - [G(6;7) - G(4;7)] = .713$

Subject: Date: .../.../...

Example. Suppose the time spent by a randomly selected student at a campus computer laboratory has a gamma dist. with mean 20 min and variance 80 min^2.

a) What are the values of α and β ?

b) What is the probability that a student uses the laboratory for at most 24 min?

Solution. We are given $\begin{cases} E(X) = \mu = \alpha\beta = 20, \\ V(X) = \sigma^2 = \alpha\beta^2 = 80. \end{cases}$

a)
From this, we obtain $\beta = 4$ and $\alpha = 5$. Then, $X \sim Gamma(5,4)$.

b) We wish to find $P(X \leq 24)$.

$$P(X \leq 24) = G\left(\frac{24}{4}; 5\right) = G(6;5) = .715.$$

The exponential function

Definition. X is said to have an exponential distribution with parameter λ ($\lambda>0$) if the pdf of X is

$$f(x;\lambda) = \lambda e^{-\lambda x} \quad , x>0.$$

This pdf is obtained from the pdf of the gamma distribution when $\alpha=1$ and $\beta = 1/\lambda$. Accordingly, the mean and variance are

$$\mu = \alpha\beta = \frac{1}{\lambda} \quad , \quad \sigma^2 = \alpha\beta^2 = \frac{1}{\lambda^2}.$$

by setting $\alpha=1$ and $\beta = 1/\lambda$.

The cdf of X is

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

Example. In studies of anticancer drugs it was found that if mice are injected with cancer cells, the survival time can be modeled with the exponential dist. Without treatment the expected survival time was 10h. What is the prob. that

a) A randomly selected mouse will survive at least 8h? At most 12h?

b) The survival time of a mouse exceeds the mean value by more than 2 SD?

Solution. We are given $\mu = \frac{1}{\lambda} = 10 \Rightarrow \lambda = \frac{1}{10}$.

Let X denote the survival time without treatment.

a) We wish to find $P(X \geq 8)$.

$$P(X \geq 8) = 1 - P(X \leq 8) = 1 - \left(1 - e^{-\frac{1}{10}\cdot 8}\right) = e^{-\frac{8}{10}} = .4493.$$

b) We wish to find $P(X > \mu + 2\sigma)$.

We know that $\mu = \sigma = \frac{1}{\lambda} = 10$.

Then, $P(X > \mu + 2\sigma) = P(X > 30) = 1 - P(X \leq 30) = 1 - \left(1 - e^{-\frac{1}{10}\cdot 30}\right) = e^{-3} = .0498.$

Theorem. If the number of events in $[0,t]$ follows a Poisson dist. with $\mu = \lambda t$, (where λ, the event rate, is the expected number of events per unit time) and if events in disjoint intervals are independent, then the time between two successive events follow an exponential distribution with parameter λ.

Exponential distribution has the memoryless property: By the definition of conditional probability,

$$\begin{aligned} P(X \geq t + t_0 \mid X \geq t_0) &= \frac{P[(X \geq t + t_0) \cap (X \geq t_0)]}{P(X \geq t_0)} \\ &= \frac{P(X \geq t + t_0)}{P(X \geq t_0)} \\ &= \frac{1 - F(t + t_0; \lambda)}{1 - F(t_0; \lambda)} \\ &= \frac{e^{-\lambda(t + t_0)}}{e^{-\lambda t_0}} \\ &= e^{-\lambda t}. \end{aligned}$$

→ It does not depend on t_0.

Here, we suppose that component lifetime is exponentially distributed with parameter λ. The property says that the distribution of remaining lifetime is independent of current age.

For example, suppose that jobs in our system have exponentially distributed service times. If we have a job that's been running for one hour, the probability that a job runs for one additional hour is the same as the prob. that it ran for one hour originally, regardless of how long it's been running.