Example : Chris, an Environmental Engineer, wants to know the Hg concentration of fish in the Eno river. She takes a rod and reel to a spot in Duke Forest to try to catch some fish. From past experience she knows that the average number of fishes she catches per hour is $\lambda = 6$. Let y denote the number of fish she catches in one hour of fishing. What is the distribution for y?

Divide the time interval (an hour) into n subintervals (small enough, say, a second) and assume

- the probability of catching more than one fish in a subinterval is (almost) zero
- the probability of catching a fish is the same for all subintervals
- the number of fish caught in each subinterval is independent of other subintervals

Then y has *approximately* a binomial distribution with mean $\mu = n\,p = \lambda$ (**Why?**), so the probability of catching a fish in one subinterval must be $p = \lambda/n$ and the probability distribution for y must be approximately

$$
\begin{aligned}
p(y) &\approx \frac{n!}{y!\,(n-y)!}(\lambda/n)^y\,(1-\lambda/n)^{n-y} \\
&= \frac{n(n-1)(n-2)\ldots(n-y+1)}{y!}\frac{\lambda^y}{n^y}\,(1-\lambda/n)^{n-y} \\
&= \frac{\lambda^y}{y!}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-y+1}{n}\right)\frac{(1-\lambda/n)^n}{(1-\lambda/n)^y} \\
&\to \frac{\lambda^y}{y!}\,e^{-\lambda} \qquad \text{for } y = 0, 1, \ldots \text{ as } n \to \infty,
\end{aligned}
$$

the Poisson distribution with mean parameter $\lambda = 6$.

Let y_t denote the number of fishes she catches in t hours.

Q: *What's the distribution for y_t?*

A: Poisson with mean parameter λt.

Q: *What's the probability she catches exactly one fish in twenty minutes?*

A: $2e^{-2} \approx 0.2707$ ($\lambda t = 6 * 1/3 = 2$; remember units!).

Let z_1 denote the waiting time (in hours) before Chris catches the first fish.

What is the distribution of z_1?

- Note that the waiting time is shorter than any number t **if and only if** Chris catches at least one fish in the first t hours... so

$$\begin{aligned} F(t) &= P(z_1 \leq t) \\ &= P(y_t \geq 1) \\ &= 1 - P(y_t = 0) \\ &= 1 - \frac{(\lambda t)^0}{0!} e^{-\lambda t} \\ &= 1 - e^{-\lambda t}, \qquad t \geq 0. \end{aligned}$$

The density function is:

$$\begin{aligned} f(t) &= F'(dt) \\ &= \lambda e^{-\lambda t}, \qquad t \geq 0. \end{aligned}$$

This is called the **exponential distribution**.

Let z_α denote the waiting time (in hours) before Chris catches the α^{th} fish, for $\alpha = 1, 2, ...$

What is the distribution of z_α?

- Note that the waiting time does not exceed any number t **if and only if** Chris catches **at least** α fish in the first t hours... so

$$\begin{aligned} F(t) &= P(y_t \geq \alpha) \\ &= \sum_{j=\alpha}^{\infty} \frac{(\lambda t)^j}{j!} e^{-\lambda t}, \qquad t \geq 0 \\ &= \texttt{gammainc}(\texttt{l} * \texttt{t}, \texttt{a}) \text{ in MatLab.} \end{aligned}$$

  Differentiating term-by-term leads to some cancellation (try it!), with:

$$\begin{aligned} f(t) &= F'(dt) \\ &= \frac{\lambda^\alpha t^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda t}, \qquad t \geq 0. \end{aligned}$$

This is called the **gamma distribution**.

Aside on the Gamma Function

The Gamma distribution is named after the function $\Gamma(\alpha)$ in the numerator,

$$
\begin{aligned}
\Gamma(\alpha) &= \int_0^\infty t^{\alpha-1} e^{-t}\, dt \\
&= \texttt{gamma(a)} \text{ in MatLab} \\
\Gamma(\alpha) &= \int_0^\infty (\alpha-1) t^{\alpha-2} e^{-t}\, dt \text{ (integrate by pts)} \\
&= (\alpha-1)\Gamma(\alpha-1) \\
&= (\alpha-1)(\alpha-2)\Gamma(\alpha-2) \\
&= (\alpha-1)(\alpha-2)\cdots 3\cdot 2\cdot 1\cdot \Gamma(1) \\
&= (\alpha-1)! \quad \text{for positive integers } \alpha \\
\Gamma(1) &= 0! = 1 = 10^0, \\
\Gamma(10) &= 9! = 362880 \approx 3.6\times 10^5, \\
\Gamma(100) &\approx 10^{156}, \\
\Gamma(1000) &\approx 10^{2565}, \\
\Gamma(\alpha) &\approx \alpha^\alpha e^{-\alpha}\sqrt{2\pi/\alpha} \text{ (Stirling's Approx)} \\
\ln\Gamma(\alpha) &= \texttt{gammaln(a)} \text{ in MatLab} \\
& \quad \text{(otherwise it gets HUGE!)}
\end{aligned}
$$

Exponential and Gamma Distributions

The Exponential distribution is the special case of the Gamma distribution, with $\alpha = 1$. Both distributions are sometimes parametrized by $\beta = 1/\lambda$, the average time-per-event, and sometimes by $\lambda = 1/\beta$, the average rate of events.

The mean and variance for the **exponential distribution** are

$$\begin{aligned}
\mu &= \int_0^\infty t\lambda e^{-\lambda t}\,dt \\
&= \lambda^{-1} = \beta \qquad \text{(Why?)} \\
\sigma^2 &= \int_0^\infty (t - \lambda^{-1})^2 \lambda e^{-\lambda t}\,dt \\
&= \lambda^{-2} = \beta^2.
\end{aligned}$$

and, for the **gamma distribution**,

$$\begin{aligned}\mu &= \int_0^\infty t\frac{\lambda^\alpha t^{\alpha-1}}{\Gamma(\alpha)}e^{-\lambda t}\,dt \\ &= \alpha\lambda^{-1} = \alpha\beta \qquad \text{(Why?)} \\ \sigma^2 &= \int_0^\infty (t-\lambda^{-1})^2\frac{\lambda^\alpha t^{\alpha-1}}{\Gamma(\alpha)}e^{-\lambda t}\,dt \\ &= \alpha\lambda^{-2} = \alpha\beta^2.\end{aligned}$$

Recall that the **Standard Normal Distribution** has mean $\mu = 0$, variance $\sigma^2 = 1$, and density function

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

It turns out (we'll see why later) that the *square* of a standard normal $y = z^2$ has a gamma distribution with $\alpha = 1/2$ and $\beta = 2$. This turns out to be important in statistics when we consider the sum of squared errors in regression.

Chi-Square Distribution

- A **chi-square (χ^2) random variable** is a gamma-type random variable with $\alpha = \nu/2$ and $\beta = 2$ (or $\lambda = 1/2$)

$$f(\chi^2) = c\ (\chi^2)^{(\nu/2)-1}\ e^{-\chi^2/2} \quad \chi^2 \geq 0$$

where

$$c = \frac{1}{2^{\nu/2}\Gamma(\frac{\nu}{2})}$$

- Mean and Variance

$$\mu = \nu \quad \sigma^2 = 2\nu$$

- The parameter ν is called the **number of degrees of freedom** for the chi-square distribution.

- Important in statistics: $\chi^2 = {z_1}^1 + ... + {z_\nu}^2$, each z_i a standard normal distribution.

Failure Time Distributions

- The **reliability** of a product is the probability that the product will meet a set of specifications for a given period of time.

- The **failure time** of a product is the length of time that the product performs according to specifications.

- The **failure time distribution** for a product is the density function of the failure time t, denoted by $f(t)$.

- Called **survival time** in medical applications (e.g. clinical trials).

- The probability that the product (or subject) will fail before any fixed time t_0 is

$$F(t_0) = \int_0^{t_0} f(t)\ dt$$

- Call a product *reliable* if it survives until time t_0. Then the **reliability** of the product (i.e., the probability that it will survive at least time t_0) is

$$R(t_0) = 1 - F(t_0),$$

also called the **Survival function** $S(t_0)$.

Hazard Rates

Given that a product has lasted at least time t, what is the probability that it will fail in the next short time period of length dt? Does this failure rate increase over time, decrease, or stay the same? What does it look like for human lifetimes? For computer chips? For automobile bearings?

- Consider the two events

  A : Item fails in the interval $(t, t + dt]$

  B : Item survives until t

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} \\ &\approx \frac{f(t)dt}{1 - F(t)} = \frac{f(t)dt}{R(t)} \end{aligned}$$

- The **hazard rate** for a product is defined to be

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{R(t)}$$

where $f(t)$ is the density function of the product's failure time distribution.

- Note $h(t) = -[\ln\Big(1 - F(t)\Big)]'$ (chain rule), so

$$F(t) = 1 - e^{-\int_0^t h(s)\,ds}, \qquad t > 0$$

and we may recover the distribution function from the hazard function.

- Example: If $h(t) \equiv \lambda$ for all $t > 0$, then

$$F(t) = 1 - e^{-\int_0^t \lambda\,ds} = 1 - e^{-\lambda t}, \qquad t > 0$$

the Exponential Distribution with rate λ (or mean $\beta = 1/\lambda$).

- **Example 17.1** The exponential distribution is often used in industry to model the failure time distribution of a product.

  Find the hazard rate for the exponential distribution.

$$F(t) = \int_{-\infty}^{t} f(y)\ dy = \int_{0}^{t} \lambda e^{-\lambda y}\, dy = 1 - e^{-\lambda t}, \quad t > 0.$$

  Then the hazard rate is

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{\lambda e^{-\lambda t}}{1 - (1 - e^{-\lambda t})} = \lambda.$$

- Constant hazard rate implies that the product does not wear out, that is, it is just as likely to survive one more hour at any age (Lack of Memory Property)

- In some applications the failure rate may **increase** (why?) or **decrease** (how is *that* possible?).

- The **Weibull distribution** allows $h(t) \propto t^{\alpha-1}$ for $\alpha > 1$ (increasing), $\alpha < 1$ (decreasing), or $a = 1$ (exponential).

Weibull Distribution

- Density Function

$$f(y) = \alpha\lambda y^{\alpha-1}e^{-\lambda y^{\alpha}}, \quad y > 0$$

Two parameters: $\alpha, \lambda > 0$

- Mean and Variance

$$\mu = \lambda^{-1/\alpha}\Gamma(\frac{\alpha+1}{\alpha})$$

$$\sigma^2 = \lambda^{-2/\alpha}\Big[\Gamma(\frac{\alpha+2}{\alpha}) - \Gamma(\frac{\alpha+1}{\alpha})^2\Big]$$

- Reliability and Hazard Rate

$$\text{CDF } F(t) = \int_0^t \alpha\lambda y^{\alpha-1}e^{-\lambda y^{\alpha}}\, dy = 1 - e^{-\lambda t^{\alpha}}, \quad t > 0$$

$$\text{Reliability } R(t) = 1 - F(t) = e^{-\lambda t^{\alpha}}$$

$$\text{Hazard Rate } h(t) = \frac{f(t)}{R(t)} = \alpha\lambda t^{\alpha-1}$$

Hazard Rate for Weibull Distribution

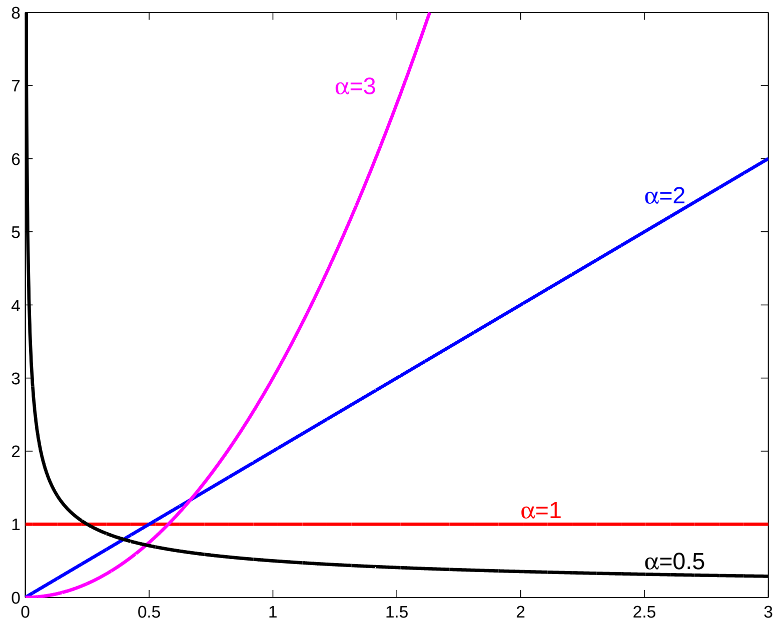
$$h(t) = \frac{f(t)}{R(t)} = \alpha\lambda t^{\alpha-1}, \qquad t > 0$$

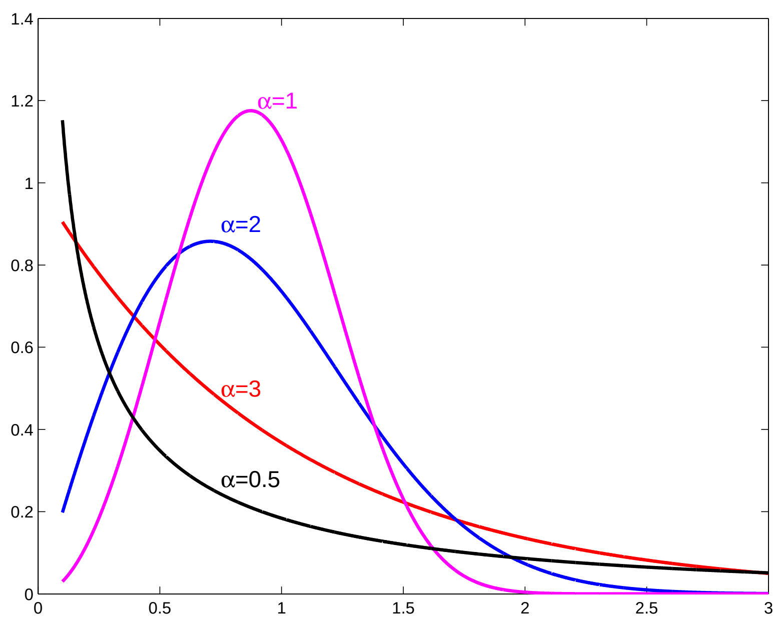
Weibull distribution provides a great deal of flexibility to model the system when the hazard rate

- increases with time (bearing wear): $\alpha > 1$
- decreases with time (semiconductors): $\alpha < 1$
- constant with time (failures caused by external shocks): $\alpha = 1$

Note x has the Weibull distribution if and only if x^α has the exponential distribution with rate λ.

Plotting of Weibull distribution's hazard rates and density functions with $\beta = 1$ and various α values.





Beta Distribution

- The probability density function for a **beta-type random variable** is given by

$$f(y) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 < y < 1$$

for parameters $\alpha, \beta > 0$ where

$$\begin{aligned} B(\alpha,\beta) &= \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\, dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \\ &= \texttt{beta(a,b)} \text{ in MatLab} \end{aligned}$$

- Recall that

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

and $\Gamma(\alpha) = (\alpha-1)!$ when α is a positive integer.

- Mean and Variance are

$$\mu = \frac{\alpha}{\alpha+\beta} \quad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The cumulative distribution function (CDF) of the beta distribution is called **incomplete beta function**.

$$
\begin{aligned}
F(p) &= \int_0^p \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}\, dy \\
&= \texttt{betainc(p, a, b)} \text{ in MatLab} \\
&= \sum_{j=\alpha}^{n} p(j) \text{ if } \alpha \text{ and } \beta \text{ are integers,}
\end{aligned}
$$

where $p(j)$ is a binomial probability distribution with parameters p and $n = (\alpha + \beta - 1)$.

Discrete Distributions

Bi(p): Bernoulli: Independent zero-one values with same probability p of success.

Bi(n,p): Binomial: Number of successes in a fixed number n of independent trials with the same probability p of success; also number of successes when sampling a population with replacement.

HG(n,A,B): Hypergeometric: Number of successes in a fixed number n of samples without replacement from a finite population of A successes, B failures.

MN($n,\vec{p}$): Multinomial: Numbers of each of k possible outcomes in a fixed number n of independent trials with the same outcome probability vector $\vec{p} = \{p_i\}$.

Ge(p)**:** Geometric: Number of independent trials needed for one success.

NB(p,α)**:** Negative Binomial: Number of independent trials needed for α successes.

Po(λ)**:** Poisson: Number of events in a fixed period if events in different periods are independent with constant rate λ.

Un(n)**:** Uniform: Finite number n of equally-likely outcomes.

Continuous Distributions

Un(S)**:** Uniform: Density function constant on some set S.

No(μ, σ^2)**:** Normal: Sum or average of large number of independent quantities.

Ex(β)**:** Exponential: Failure time if hazard is constant $1/\beta$; time-to-first-event distribution for Poisson with rate $\lambda = 1/\beta$.

We(α, λ)**:** Weibull: Failure time if hazard is power $\alpha - 1$.

Ga(α, β)**:** Gamma: Time-to-α^{th}-event distribution for Poisson with rate $\lambda = 1/\beta$.

Be(α, β)**:** Beta: Order statitics: α^{th} largest ($= \beta^{th}$ smallest) of $n = \alpha + \beta - 1$ indep. uniforms.