Probability Distribution Functions - Continuous variable

Josep L. Carrasco
Bioestadística. Departament de Fonaments Clínics
Universitat de Barcelona

Continuous random variable: functions and properties

The set of possible results is infinite. The probability of a point (specific value) equals 0 $P\left(X = k\right) = 0$.

For a continuous random variable, we have

$$P\left(a \leq X \leq b\right) = P\left(a < X < b\right).$$

Some examples:

- Height. What is the probability of being **exactly** 1.90 meters tall?
- Battery life time. What is the probability for a battery to last **exactly** 48 hours?
- Blood-cholesterol levels. What is the probability of having a cholesterol level of **exactly** 180 mg/dl?

The mass probability function makes no sense for continuous variables because it always takes the value 0.

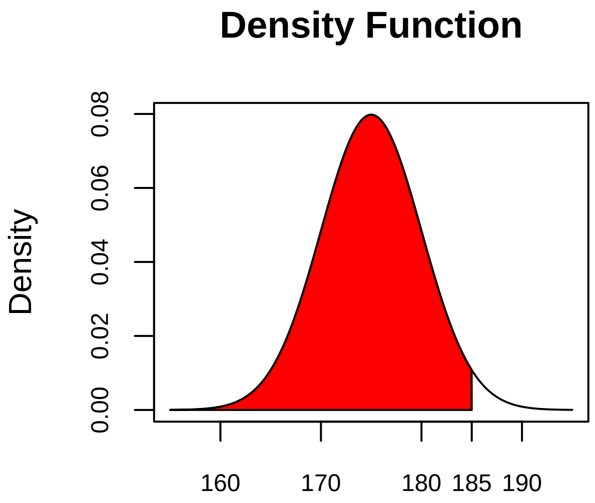
Probability density function

The probability density function is an infinitesimal: it can be interpreted as the probability intensity of a point.

$$f(x) = \lim_{\Delta \to 0} P\left(x - \Delta < X < x + \Delta\right).$$

The probability distribution function is computed by integrating the probability density function:

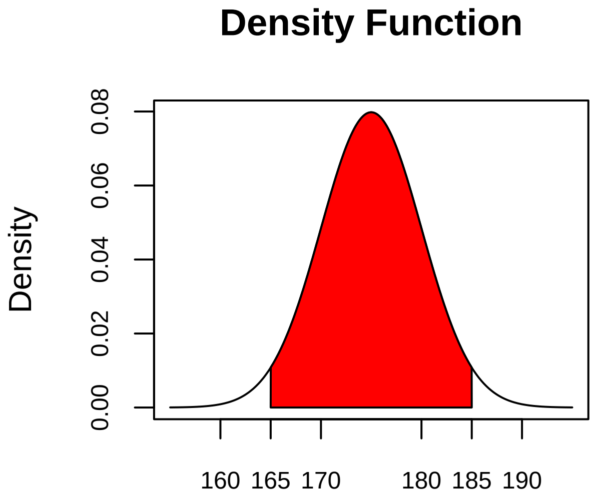
$$F(x) = P\left(X < x\right) = \int_{-\infty}^{x} f(t)\,dt.$$



The red area is the probability of $P\left(X < 185\right)$, i.e. the probability distribution function evaluated at 185.

The probability of an interval is computed as $P\left(a < X < b\right) = P\left(X < b\right) - P\left(X < a\right)$.

For example, the probability of $P\left(165 < X < 185\right)$ is shown in the following plot.



Expectation and Variance

The expectation and variance in a continuous variable are also obtained by integrating the probability density function.

- Expectation

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx.$$

- Variance

$$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2\, f(x)\, dx = \int_{-\infty}^{\infty} x^2 f(x)\, dx - E(X)^2$$

Probability distribution models

Continuous uniform distribution

X: random value in the continuous interval $[a, b]$.

- Probability density function: $f(x) = \frac{1}{b-a}$, $a \leq x \leq b$.
- Probability distribution function:

$$F(x) = \begin{cases} 0 & x < a; \\ \frac{x-a}{b-a} & a < x < b; \\ 1 & x > b. \end{cases}$$

- Expectation and variance:

$$E(x) = \frac{a+b}{2}, \qquad V(X) = \frac{(b-a)^2}{12}.$$

Example. In a neurological exam, the patient is asked to press a button when a light is switched on. This signal can be activated in any moment during a time lapse of 10 seconds. What is the probability that the signal is activated between the seconds 2 and 4?

$$P\left(2 < X < 4\right) = \int_2^4 \frac{1}{10}\, dx = \frac{2}{10} = 0.2.$$

The uniform distribution is commonly used in the generation of random values.

Normal distribution

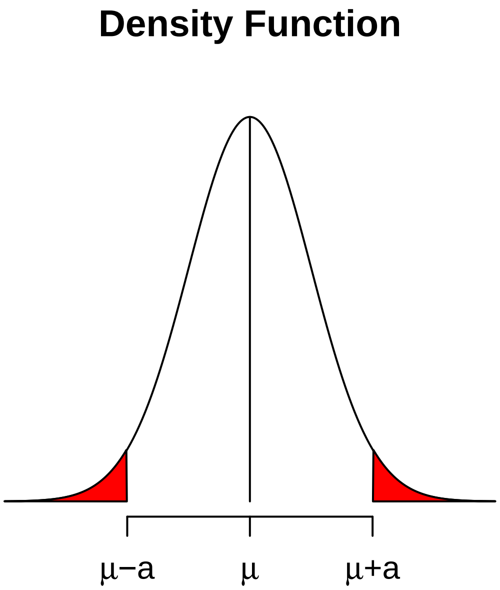
- The most used model in the continuous setting.
- It is determined by the mean (μ) and standard deviation (σ):

$$X \sim N\left(\mu, \sigma\right).$$

- Probability density function: $f(x) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$.

Some properties

- Symmetry: $P\left(X > \mu + a\right) = P\left(X < \mu - a\right)$.



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

Central limit theorem

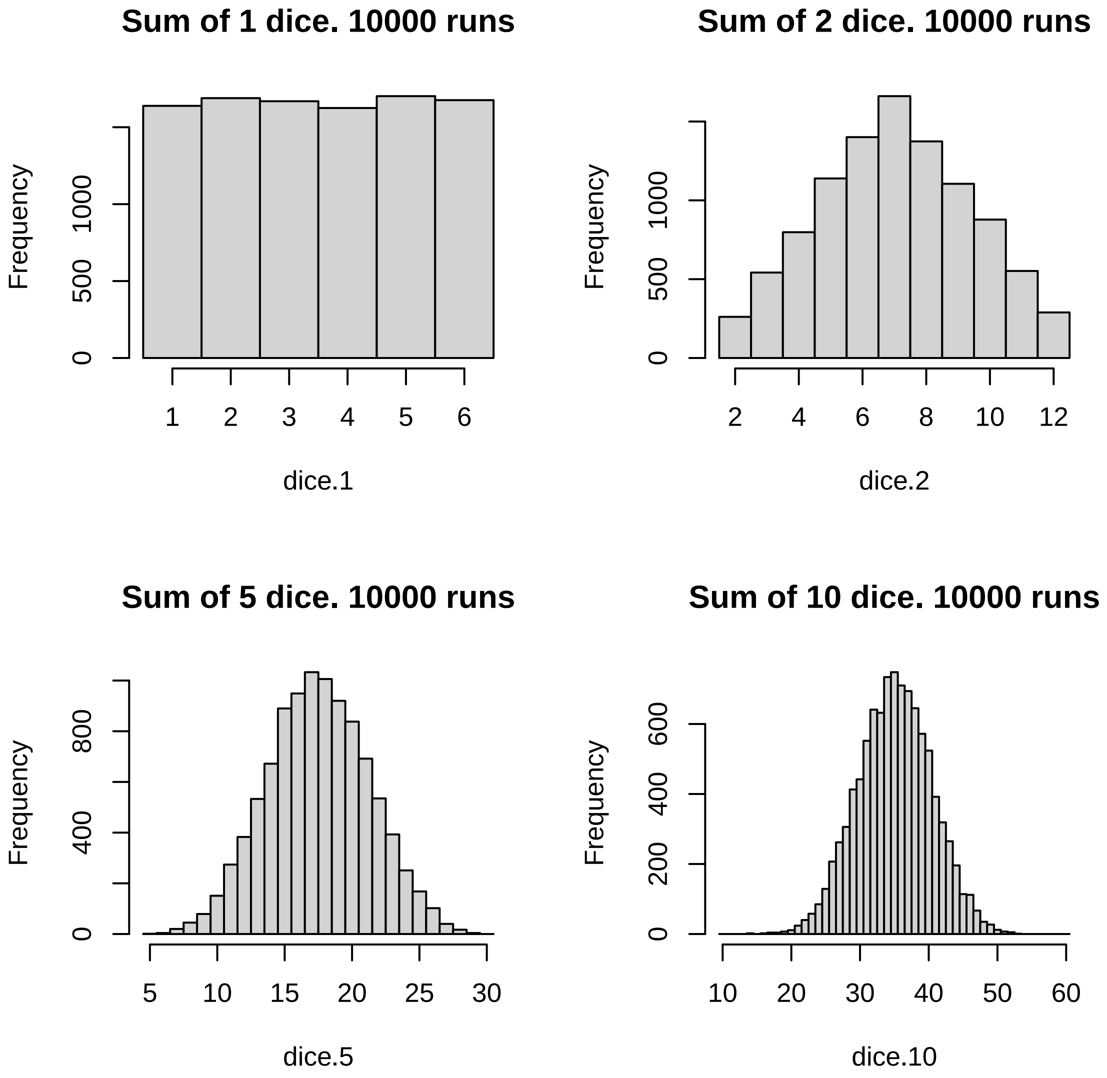
If an experiment is replicated, the sum or mean of its results tend to be normally distributed.

X_i: Independent and identically distributed random variables with mean μ and variance σ^2.

$$S_n = \sum_{i=1}^{n} X_i \sim N\left(n\mu, \sqrt{n}\sigma\right),$$
$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

If X_i are normally distributed, then the result is **exact**. Otherwise, the result is **asymptotic** (with large n).

Example. Throw a dice and take the sum of the results.



Convergence to a normal distribution

In some situations it is possible to approximate the probability of a discrete probability model using the Normal model. Nowadays, from a computational point of view it can be meaningless to use approximations to obtain a specific probability. However, most of the standard statistical procedures are based on this property.

- **Binomial to Normal**. If n is large enough and both p and $q = 1 - p$ are not extreme,

$$B(n, p) \longrightarrow N(np, \sqrt{npq}).$$

The larger n and the less extreme p, the lower the approximation error. The error is said to be acceptable if $np > 5$ and $nq > 5$.

- **Poisson to Normal**. If the mean (λ) is large enough,

$$P(\lambda) \longrightarrow N\left(\lambda, \sqrt{\lambda}\right).$$

The larger λ, the lower the approximation error. The error is said to be acceptable if $\lambda > 5$.

Continuity correction

When we approximate a discrete variable to a continuous one, a continuity correction can be applied to improve the approximation.

The correction is:

$$P\left(X \leq k\right) \approx P\left(Y < k + 0.5\right)$$

where X stands for the discrete variable and Y for the approximation to Normal model.

Example. Binomial with parameters $n = 50$ and $p = 0.2$.

- Probability of observing a result less than 10:

$$P\left(X < 10\right) = P\left(X \leq 9\right) = 0.44374.$$

- Approximation to a Normal distribution: $X \sim B(50, 0.2) \rightarrow Y \sim N(10, 2.8284)$.

  In this case $np = 10 > 5$ and $n\left(1 - p\right) = 40 > 5$, so the approximation error is acceptable.

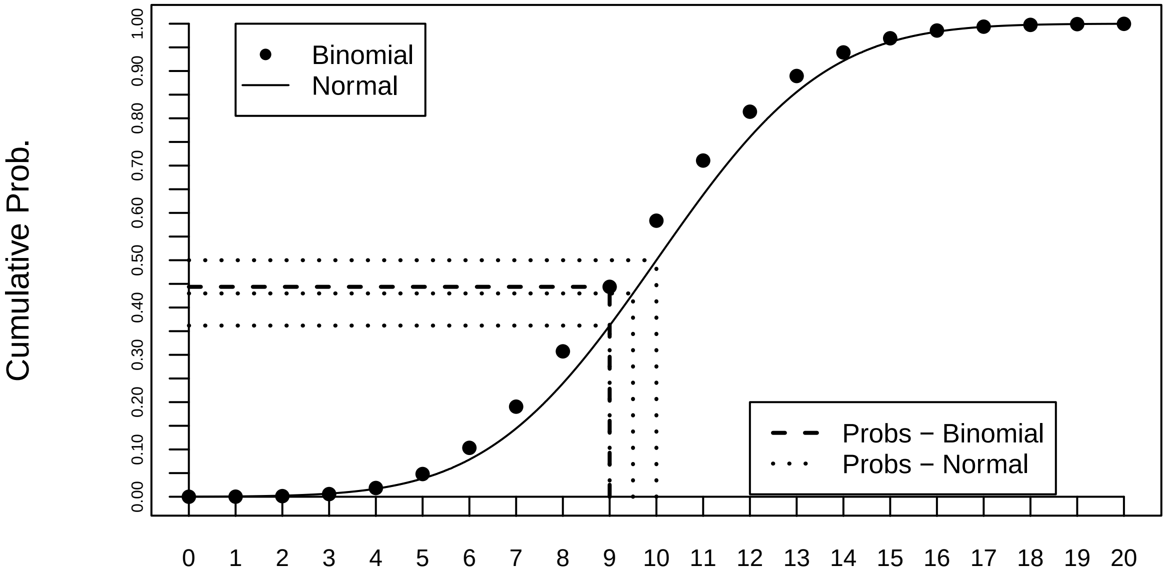
  We have two options to compute the probability using the approximation to the Normal model:

$$P\left(X < 10\right) \approx P\left(Y < 10\right) = 0.5.$$

$$P\left(X < 10\right) = P\left(X \leq 9\right) \approx P\left(Y < 9\right) = 0.3618.$$

- Approximation to a Normal distribution using continuity correction:

$$P\left(X < 10\right) = P\left(X \leq 9\right) \approx P\left(Y < 9.5\right) = 0.4298.$$



Related models

These are models used in data analysis that are derived from the Normal distribution.

- Chi-square
- t-Student
- F-Fisher

We will use them in the *Inference* block.

Log-normal distribution

Let X follow a normal distribution with mean μ and standard deviation σ.

$Y = \exp(X)$ follows a log-normal distribution because $\ln(Y) = X$.

$E(Y) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$, $V(Y) = \exp\left(2\mu + \sigma^2\right)\left(\exp\left(\sigma^2\right) - 1\right)$.

Examples

- Skewed and positive variables as **biochemical parameters** (e.g. blood-cholesterol levels) or a **time-to-event** variable commonly follow a log-normal distribution.

Exponential distribution

X: time to observe one event.

Derivation

Let Y be the count of events with respect to a unit differential (time or space).

For example, number of deaths in a year; number of flaws in 1mm of copper wire.

If Y follows a Poisson distribution with parameter λ, the random variable W= "number of events in x units" follows a Poisson distribution with parameter λx.

We are interested in the variable X="time to one event" (one death, one flaw. . .).

The probability that no event is observed in x units is

$$P(X > x) = P(W = 0) = \frac{e^{-\lambda x}\lambda^0}{0!} = e^{-\lambda x}.$$

The opposite event corresponds to the probability distribution function:

$$F(x) = P\left(X \leq x\right) = 1 - P\left(X > x\right) = 1 - e^{-\lambda x}.$$

The probability density function is obtained by derivation: $f(x) = \lambda e^{-\lambda x}$.

$E(X) = \frac{1}{\lambda}$, $V(X) = \frac{1}{\lambda^2}$.

Poisson process

- $N(t)$: count process in continuous time.

Characteristics:

- $\boldsymbol{N(0) = 0}$.
- **Independence**. The number of events in an interval does not depend on the observed number of events in the rest of the intervals.
- **Stationarity**. The distribution of the number of events only depends on the width of the interval.
- The events can not be simultaneous.

Consequences:

- $N(t)$ follows a **Poisson** distribution.

- The time between events follows an **exponential** distribution.
- $N(t)$ is a **memoryless process**:

$$P\left(X < t_1 + t_2 | X > t_1\right) = P\left(X < t_2\right).$$

Example The number of defective objects produced monthly by a machine follows a Poisson process with mean 60 objects/month (30 days).

- The variable "number of defective objects in a day" follows a Poisson model with mean 2 objects a day.
- The variable "time to next defective object" follows an exponential model with mean 0.5 days.

Gamma distribution

- Probability density function: $f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}$, $x > 0$.
- $E(X) = \frac{r}{\lambda}$, $V(X) = \frac{r}{\lambda^2}$.
- Gamma function:

$$\Gamma(r) = \begin{cases} (r-1)! & \text{when } r \text{ is discrete;} \\ \int_0^\infty t^{r-1} e^{-t}\, dt & \text{when } r \text{ is continuous.} \end{cases}$$

When r is discrete, the gamma distribution is known as **Erlang distribution** in engineering.

The Erlang distribution is the generalization of the exponential distribution: probability of r events in a time interval.

If $r = 1$, the gamma distribution becomes the exponential distribution.

Example A health service is capable of attending 12 patients per hour. However, the service gets collapsed if 3 patients are admitted in less than 5 minutes

The variable "time to the service collapses" follows a Gamma (Erlang) distribution with parameters $r = 3$ and $\lambda = \frac{12 \text{ visits}}{60 \text{ minutes}} = 0.2$.

Weibull distribution

Example. A group of researchers are using computational intensive methods in a computer to perform a simulation study. The computer memory gets overloaded every 200 hours in mean. On the other hand, this overloading increases with the time of use of the computer.

The variable is time to event, time to overload in the example. The exponential model assumes that the rate of the event is constant over time. However, in this example exponential model is not appropriate because the rate increases with time.

Weibull model provides flexibility of the rate over time.

- Probability density function

$$f(x) = \frac{\beta}{\delta}\left(\frac{x}{\delta}\right)^{\beta-1} \exp\left[-\left(\frac{x}{\delta}\right)^{\beta}\right],\ x > 0.$$

- Scale parameter $\delta > 0$ (inverse of the “Poisson’ ’ rate)
- Shape parameter $\beta > 0$.

$$\beta = \begin{cases} < 1 & \text{the failure rate decreases over time} \\ = 1 & \text{the failure rate is constant over time (Weibull = Exponential)} \\ > 1 & \text{the failure rate increases with time} \end{cases}$$

$$E(X) = \delta\Gamma\left(1 + \frac{1}{\beta}\right)$$

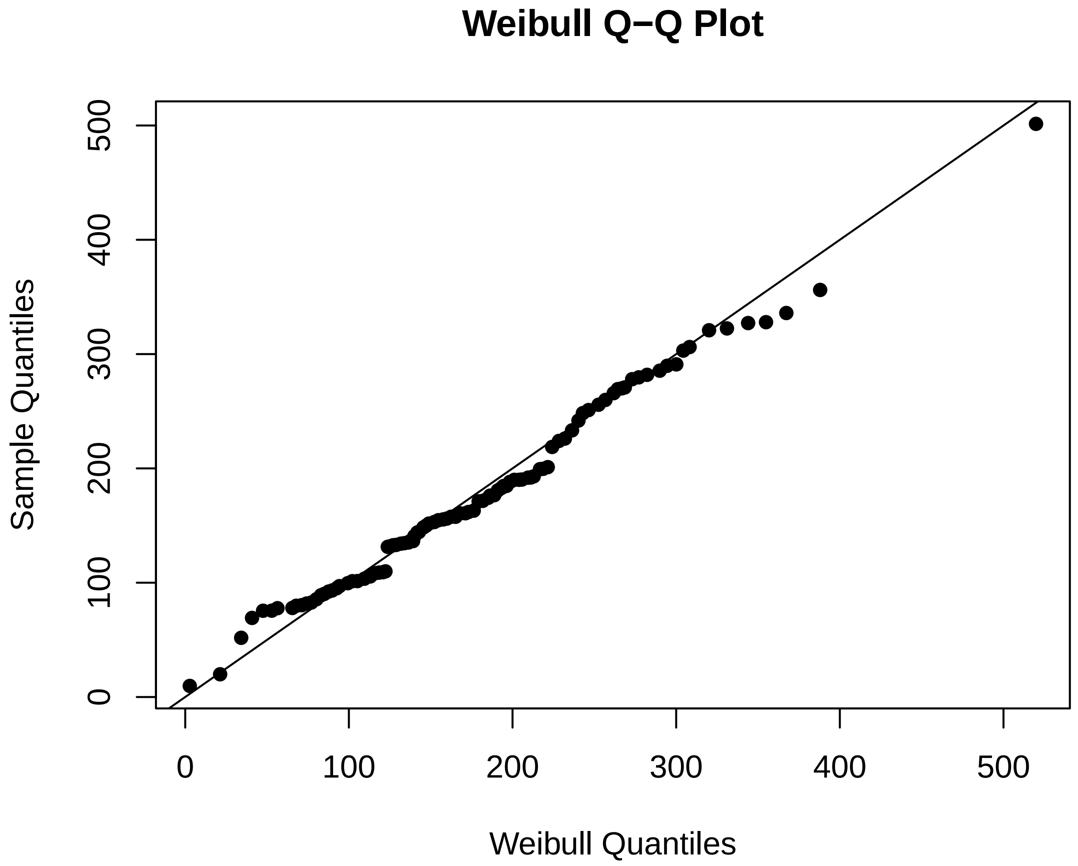
$$V(X) = \delta^2\Gamma\left(1 + \frac{2}{\beta}\right) - \delta^2\left[\Gamma\left(1 + \frac{1}{\beta}\right)\right]^2.$$

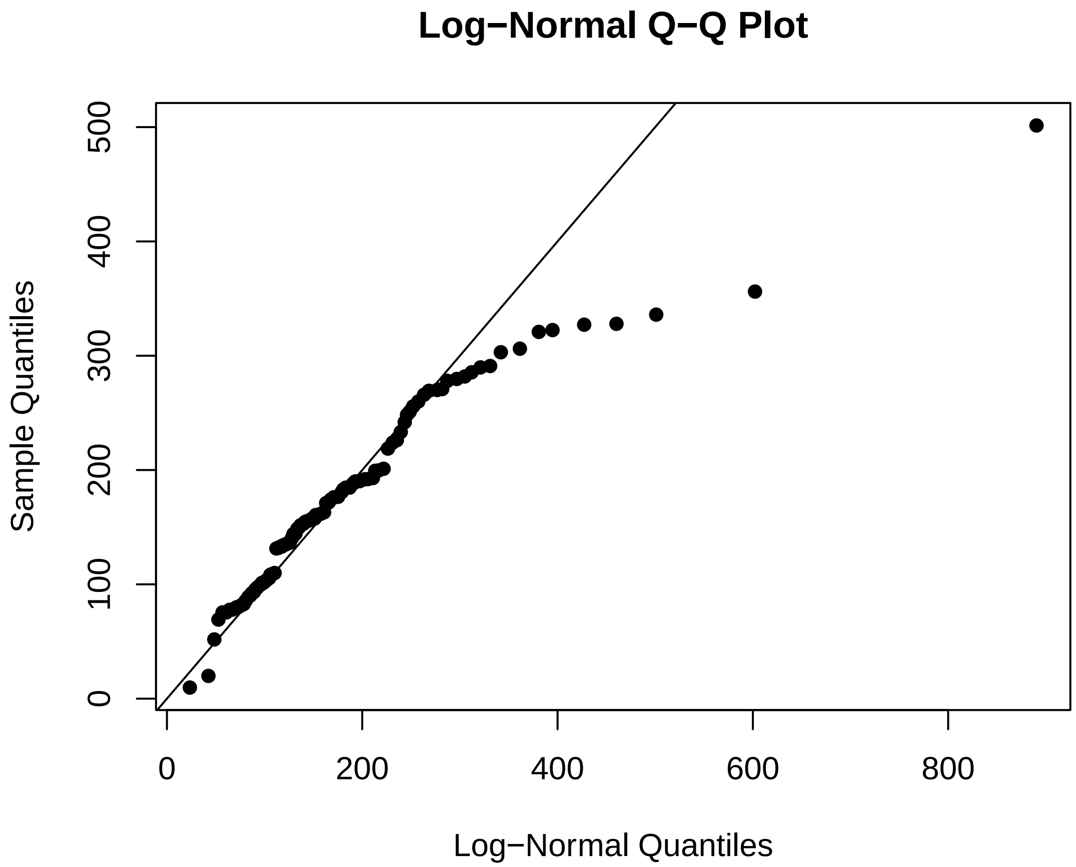
Q-Q plots

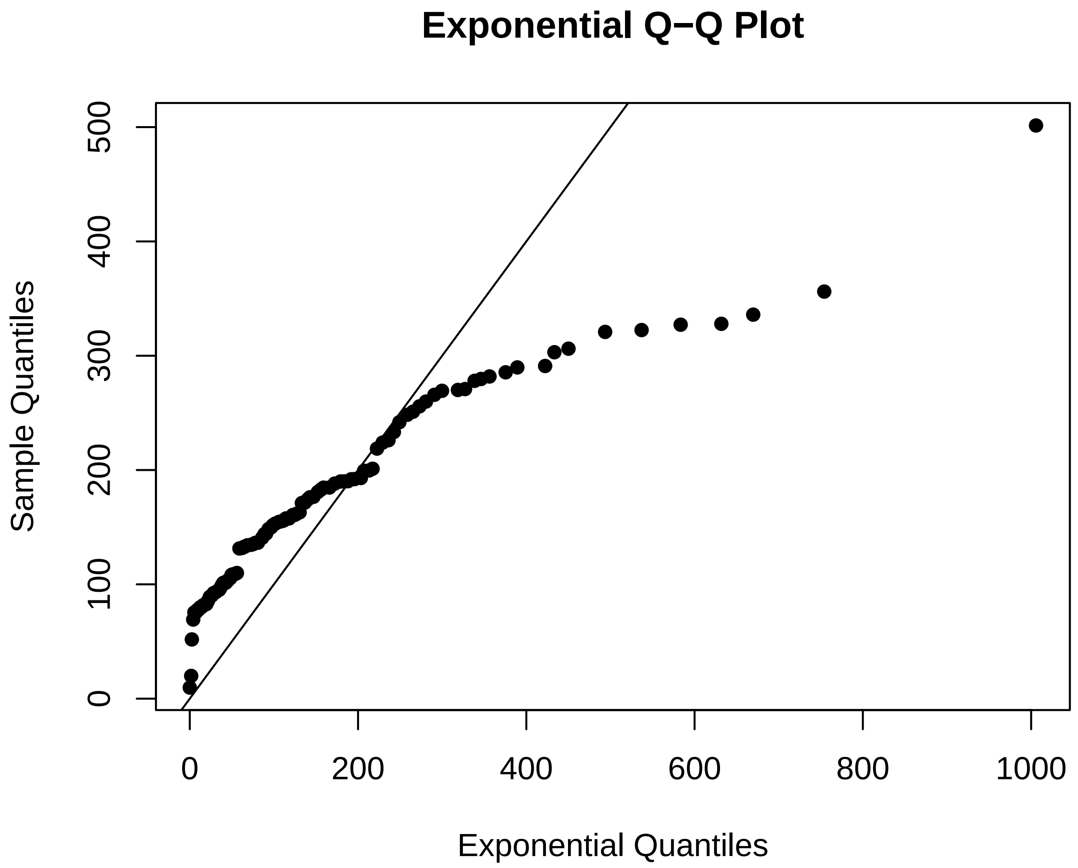
- Descriptive tools to check the goodness of fit of a variable to a specific probability model.
- They are based on comparing the quantiles of the empirical (observed) distribution of the data with those of the theoretical distribution.

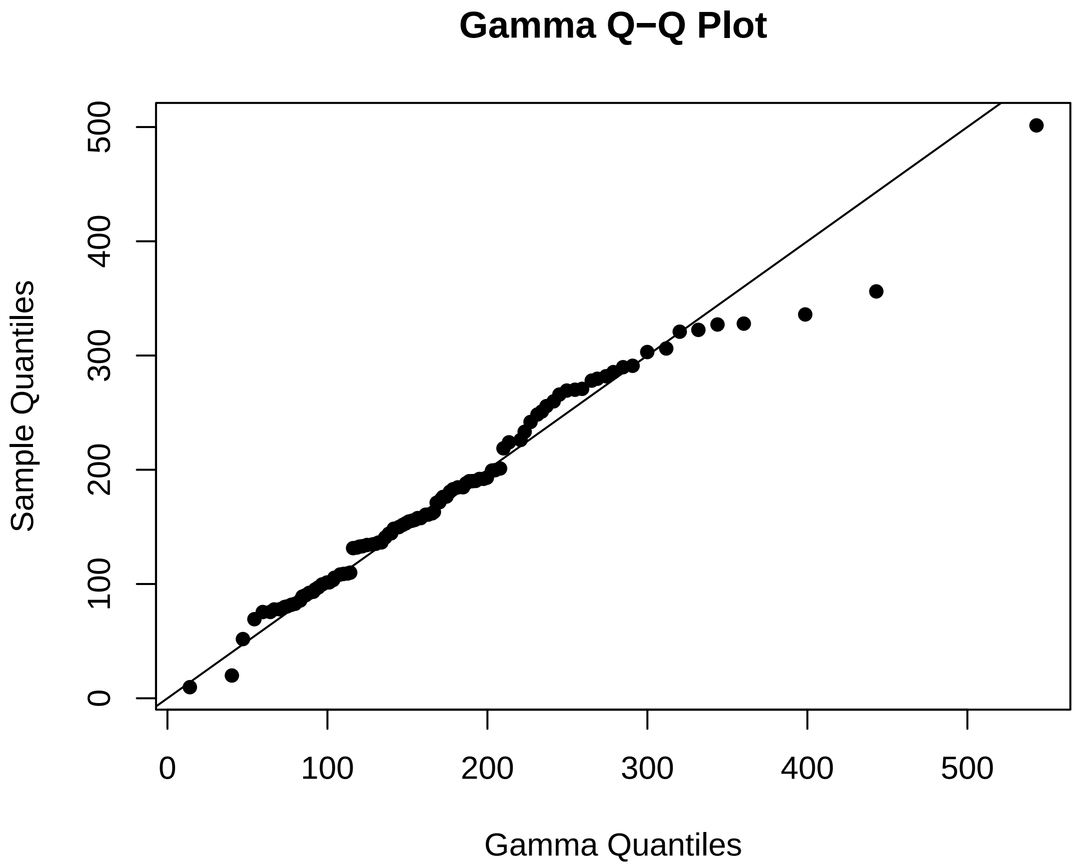
Following the last example of time to overloading, let's suppose we have got the times corresponding to 100 overloadings.

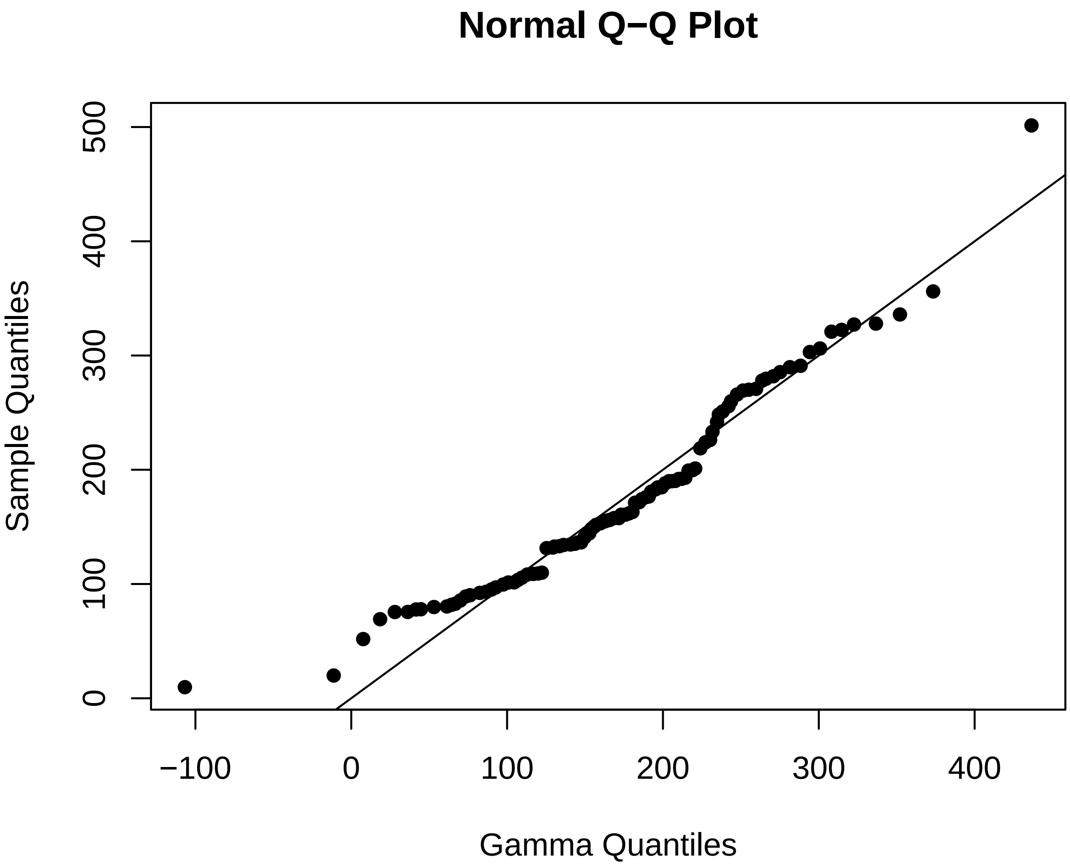
We draw the Q-Q plots for five different models.




Log−Normal Q−Q Plot
Sample Quantiles
0
100
200
300
400
500
0
200
400
600
800
Log−Normal Quantiles


Exponential Q−Q Plot
Sample Quantiles
0
100
200
300
400
500
0
200
400
600
800
1000
Exponential Quantiles


Gamma Q−Q Plot
Sample Quantiles
0
100
200
300
400
500
0
100
200
300
400
500
Gamma Quantiles



The model that best fits the data is the Weibull.