Chapter 1: Classical statistical models and their limitations

Christophe Ley

University of Luxembourg, 2023-2024

Statistical Modelling

Outline

Linear regression

Consider a simple linear regression model of the form

$$Y = a + bX + \epsilon$$

with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $a, b \in \mathbb{R}, \sigma^2 > 0$. We know how to estimate the three parameters involved in this model.

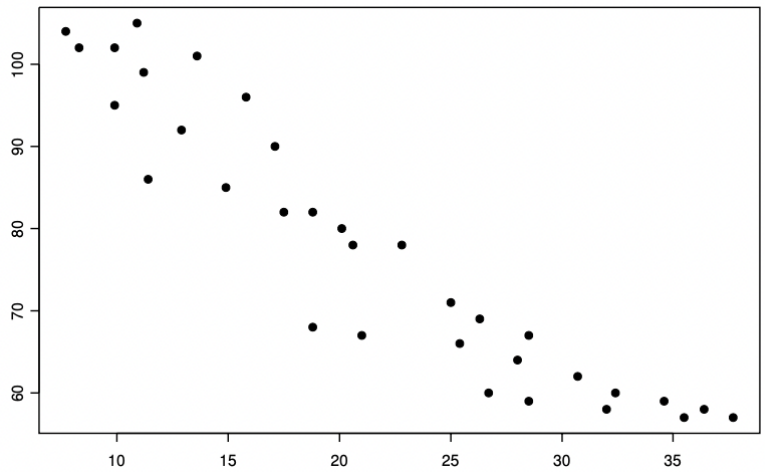


Figure – Scatterplot of amount of snowfall as a function of the distance to the snow belt.

Let's take a step back and think : what assumptions on our data $(Y_i, X_i), i = 1 \ldots, n$, are we making when trying to model the relationship between outcome Y and predictor X using linear regression, or OLS ?

- We assume the residuals $Y_i - a - bX_i$ to be independent
- We assume homoscedasticity
- We assume normality of the residuals ; this implies in particular $Y|X = x \sim \mathcal{N}(a + bx, \sigma^2)$
- We assume a linear variation of Y in the mean : $\mathrm{E}[Y|X = x] = a + bx$

What happens if one or more of these assumptions are not met ? What happens if, in essence, the relationship is not linear ? What use is the R^2 coefficient then ?

- We assume the residuals $Y_i - a - bX_i$ to be independent
- We assume homoscedasticity
- We assume normality of the residuals ; this implies in particular $Y|X = x \sim \mathcal{N}(a + bx, \sigma^2)$
- We assume a linear variation of Y in the mean : $\mathrm{E}[Y|X = x] = a + bx$

What happens if one or more of these assumptions are not met ? What happens if, in essence, the relationship is not linear ? What use is the R^2 coefficient then ?

Albeit being a classical model, linear regression is user-chosen. We decide that the relationship is linear. It greatly simplifies the analysis and the interpretation, at the risk of model misspecification.

The latter risk increases dramatically when two or more predictors are in the model !

Logistic regression

In logistic regression, not the value of Y is predicted, but rather the probability that Y is either YES or NO. This probability, $p(X)$, is given by the logistic function

$$p(X) = \frac{e^{a+bX}}{1+e^{a+bX}}.$$

The parameters are estimated by way of the maximum likelihood approach. The likelihood function takes on the guise

$$\ell(a,b) = \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} (1-p(x_j)) = \prod_{i=1}^{n} p(x_i)^{y_i} (1-p(x_i))^{1-y_i}.$$

because we are in fact considering logistic regression like a binomial model.

Small digression

One easily rewrites the previous expression as

$$\frac{p(X)}{1-p(X)} = e^{a+bX}$$

which we call odds. The odds take any value between 0 and ∞, which respectively indicates very low and very high probabilities. They indicate the ratio of "probability to happen" between two opposed events.

Examples :

- $p(X) = 0.1$, then the odds are $1/9$.
- $p(X) = 0.9$, then the odds are 9.

Odds are quantities used by bookmakers and sport betting companies.

Small digression

One easily rewrites the previous expression as

$$\frac{p(X)}{1-p(X)} = e^{a+bX}$$

which we call odds. The odds take any value between 0 and ∞, which respectively indicates very low and very high probabilities. They indicate the ratio of "probability to happen" between two opposed events.

Examples :

- $p(X) = 0.1$, then the odds are $1/9$.
- $p(X) = 0.9$, then the odds are 9.

Odds are quantities used by bookmakers and sport betting companies. Famous quote : *Beat the odds !*


World Cup
gamblers win
$74,872
betting Uruguay's
Luis Suarez would
BITE an opponent.
Odds:
175 to 1
BBC NEWS
GETTY IMAGES SPORT

Finally, taking the logarithm in $\frac{p(X)}{1-p(X)} = e^{a+bX}$ leads to the log-odds

$$\log\left(\frac{p(X)}{1-p(X)}\right) = a + bX.$$

Log-odds are better known as logit.

Finally, taking the logarithm in $\frac{p(X)}{1-p(X)} = e^{a+bX}$ leads to the log-odds

$$\log\left(\frac{p(X)}{1-p(X)}\right) = a + bX.$$

Log-odds are better known as logit.

A famous alternative proposal to the logit is the probit=probability unit. The modelling goes as follows :

$$P(Y = 1 \,|\, X) = \Phi(a + bX)$$

with Φ the cdf of the standard normal distribution.

Finally, taking the logarithm in $\frac{p(X)}{1-p(X)} = e^{a+bX}$ leads to the log-odds

$$\log\left(\frac{p(X)}{1-p(X)}\right) = a + bX.$$

Log-odds are better known as logit.

A famous alternative proposal to the logit is the probit=probability unit. The modelling goes as follows :

$$P(Y = 1 \,|\, X) = \Phi(a + bX)$$

with Φ the cdf of the standard normal distribution.

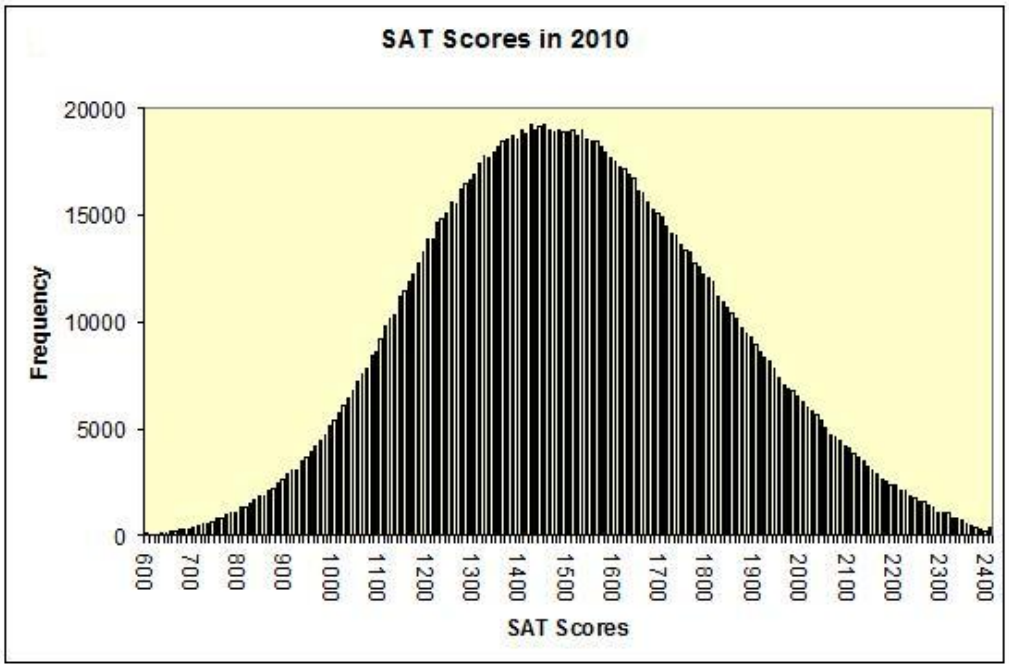
Wrapping up, what modelling assumptions are we making in logistic regression ?

Outline

The basic idea of modelling data

In various situations, describing the data at hand is not sufficient, and you rather wish to have a probability distribution that matches well your data. This allows making much more general statements about the population the data are drawn from.

Everybody knows about the normal distribution and its “universal use”. Think for instance about the IQ points.



In order to better understand the intricacies of the models we have nowadays, their advantages, their limitations, we shall now go back in time and follow the development of modelling from the perspective of the normal distribution.

The name “Gaussian distribution” is a perfect example of Stigler’s law of eponymy : no scientific discovery is named after its original discoverer.

Famous example : Hubble’s law (by Georges Lemaître), and...

In order to better understand the intricacies of the models we have nowadays, their advantages, their limitations, we shall now go back in time and follow the development of modelling from the perspective of the normal distribution.

The name “Gaussian distribution” is a perfect example of Stigler’s law of eponymy : no scientific discovery is named after its original discoverer.

Famous example : Hubble’s law (by Georges Lemaître), and... Stigler’s law itself !

Gauss himself

Carl Friedrich Gauss established the Gaussian bell curve around 1809. He found the formula

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

during his measurements of positions of stars.

$$\text{Final Measurement} = \text{Measurement} + \text{Error}$$

Through the modelling of the error term, Gauss was able to much better understand his measurements.

Belgian influence

The famous Belgian scientific Adolphe Quetelet (Ghent 1796 - Brussels 1874) was a huge fan of the Gaussian distribution !

He went even one step further : all measurements, also in biology and sociology, follow Gauss' distribution. The errors are simply due to the nature making errors !

A star is born

Between 1800 and 1900, the Gaussian distribution was a star ! It got used to model everything.

Around 1880 : the belief in the Gaussian distribution was so strong that the name “normal distribution” got proposed.

A star is born

Between 1800 and 1900, the Gaussian distribution was a star ! It got used to model everything.

Around 1880 : the belief in the Gaussian distribution was so strong that the name “normal distribution” got proposed.

But then, around 1890, something happened that changed the fate of statistics until today...

The influence of the crabs !

Other distributions

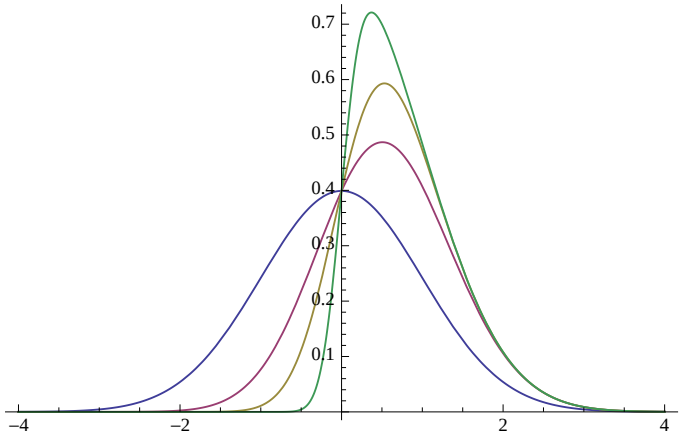
From that moment on, other distributions got proposed under the guidance of Karl Pearson, founder of the journal **Biometrika**.

Other distributions

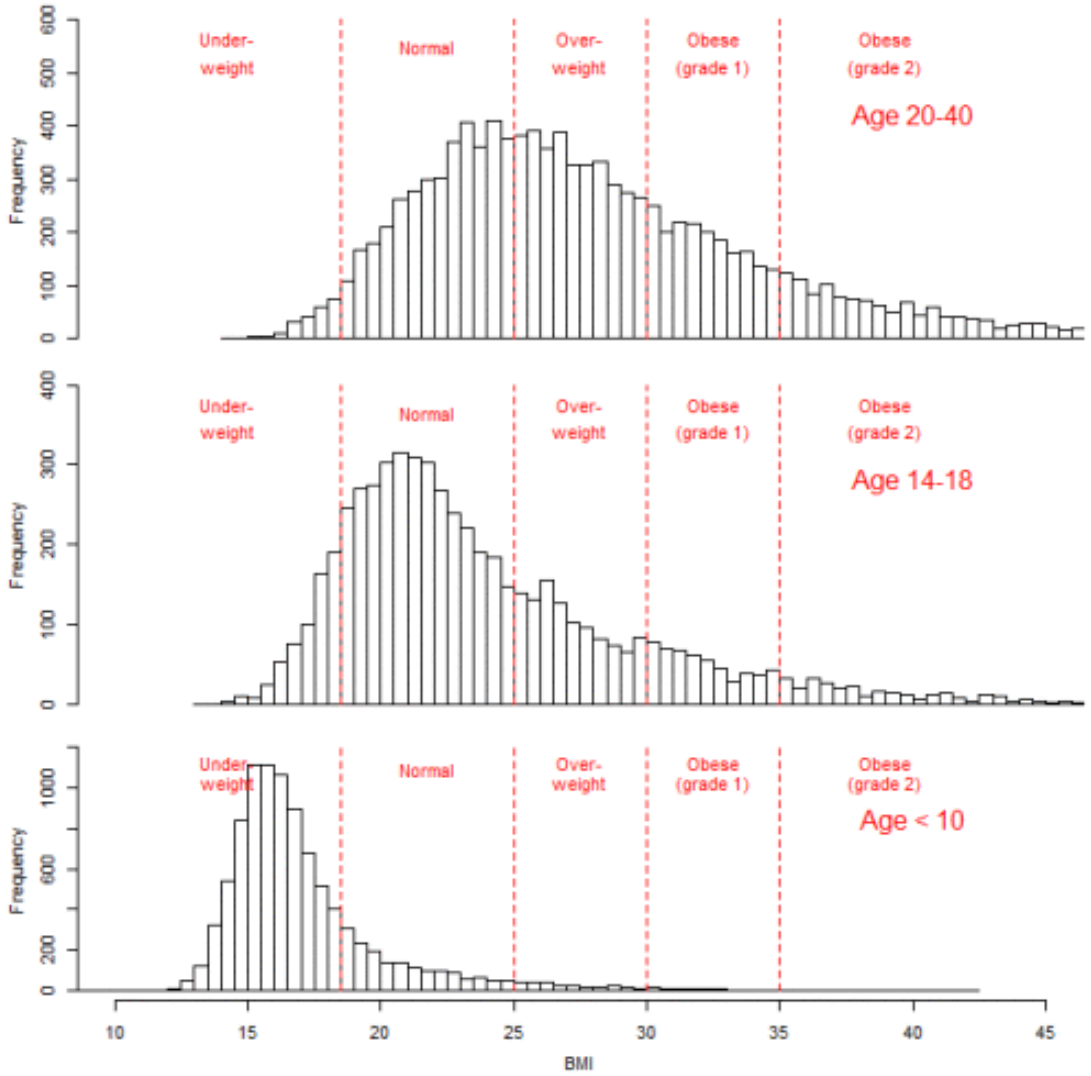
From that moment on, other distributions got proposed under the guidance of Karl Pearson, founder of the journal **Biometrika**.

Example : the skew-normal distribution of de Helgueiro (a fierce opponent to Pearson)

$$2\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\right)\Phi(\delta x)$$



Example : BMI data



Numerous non-normal distributions have been proposed throughout the 20th century. This effort continues since the turn of the century, with more and more non-parametric (especially Machine Learning) approaches seeing the light.

The normal distribution manages to keep its special role nevertheless.

Outline

Distribution on $\mathbb{R}$: The normal distribution

The normal density is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

with $\mu \in \mathbb{R}$ the location and $\sigma^2 > 0$ the scale.

The roles of the two parameters μ and σ become even clearer through the following :

$$E(X) = \mu \qquad \text{and} \qquad Var(X) = \sigma^2.$$

We therefore use the notation $N(\mu, \sigma^2)$. It enjoys numerous tractable properties such as the sum of two independent normals remains normal.

The standard normal

We can turn any $N(\mu, \sigma^2)$ distribution into the standard normal $N(0, 1)$ through the transformation

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

The standard normal has special symbols for the density and distribution function : $\phi(x)$ and $\Phi(x)$, respectively.

Central Limit Theorem

The normal distribution is also so famous because of the, perhaps, most prominent probability theory result : the Central Limit Theorem.

Suppose that the random variables $X_1, \ldots, X_n$ are independent and identically distributed, with common expectation μ and variance σ^2. What can we then say about the sum $S_n = X_1 + X_2 + \cdots + X_n$?

Central Limit Theorem

The normal distribution is also so famous because of the, perhaps, most prominent probability theory result : the Central Limit Theorem.

Suppose that the random variables $X_1, \ldots, X_n$ are independent and identically distributed, with common expectation μ and variance σ^2. What can we then say about the sum $S_n = X_1 + X_2 + \cdots + X_n$?

In formal mathematical words, the CLT states that, as $n \to \infty$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} N(0, \sigma^2).$$

In other words, we have

$$\frac{S_n - n\mu}{\sqrt{n}} \approx N(0, \sigma^2)$$

meaning that, as n becomes large, $S_n/\sqrt{n}$ behaves like a normal distribution $\rightarrow$ clear proof of the central role played by the normal law !

Alternative vision on the CLT

The CLT can also be interpreted from an entropic perspective !

Alternative vision on the CLT

The CLT can also be interpreted from an entropic perspective ! The entropy of a random variable X with density f is defined as

$$H(X) = -E[\log f(X)].$$

It measures the uncertainty of a random variable. The more uncertain it is, the higher the entropy.

The following result is famous.

Theorem

Among all probability distributions on $\mathbb{R}$ with given mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, the entropy is maximized by the $\mathcal{N}(\mu, \sigma^2)$.

Proof

Let g and f be two densities. Note that $\int_{\mathbb{R}} f(x) \log\left(\frac{g(x)}{f(x)}\right) dx \leq 0$ (since $\log(x) \leq x - 1$).
Consequently, $H(f) \leq -\int_{\mathbb{R}} f(x) \log g(x) dx$. Choosing g the normal density, we have

$$\begin{aligned} H(f) &\leq -\int_{\mathbb{R}} f(x) \log\left[(2\pi\sigma^2)^{-1/2} \exp\left(-(x-\mu)^2/(2\sigma^2)\right)\right] dx \\ &= \frac{\log(2\pi\sigma^2)}{2} + \frac{1}{2\sigma^2} \int_{\mathbb{R}} (x-\mu)^2 f(x) dx \\ &= \frac{\log(2\pi\sigma^2)}{2} + \frac{1}{2} \\ &= \frac{\log(2\pi\sigma^2 e)}{2} \\ &= H\left(\mathcal{N}(\mu, \sigma^2)\right), \textit{proof this !} \end{aligned}$$

and hence $H(f) \leq H\left(\mathcal{N}(\mu, \sigma^2)\right)$ for all f, with equality iff f is normal.

What does this result imply for the CLT ? Look at $n^{-1/2} \sum_{i=1}^{n} (X_i - \mu)$. Its mean and variance are the same for all n. And it converges towards the normal law !

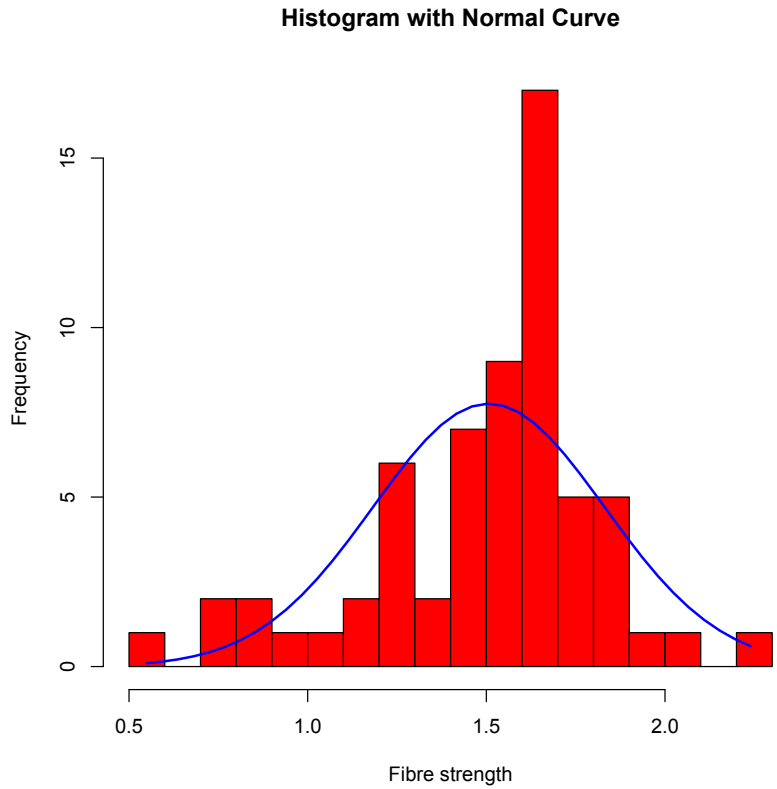
So... we have convergence towards the most uncertain, hence most natural, most normal state !

So, basically, the normal distribution occupies such a central role in probability theory because (i) it has a nice shape, (ii) it has nice mathematical properties and (iii) it is easy to understand and deal with !

Talking about its shape... it is symmetric around its center and has most of the probability mass concentrated around the center ! From a modelling perspective this is to be considered as a serious limitation !

Example of skew data from engineering

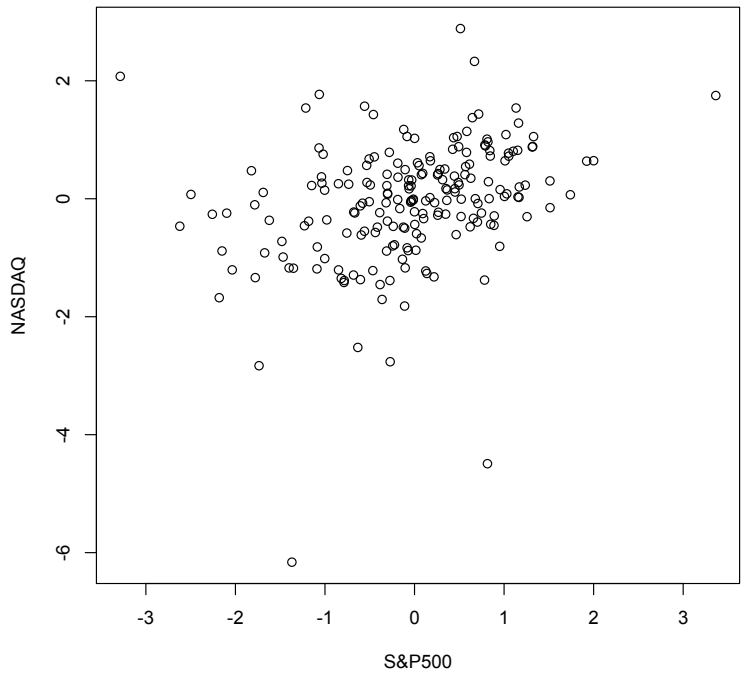
Breaking strength of $n = 63$ glass fibres of length 1.5 cm

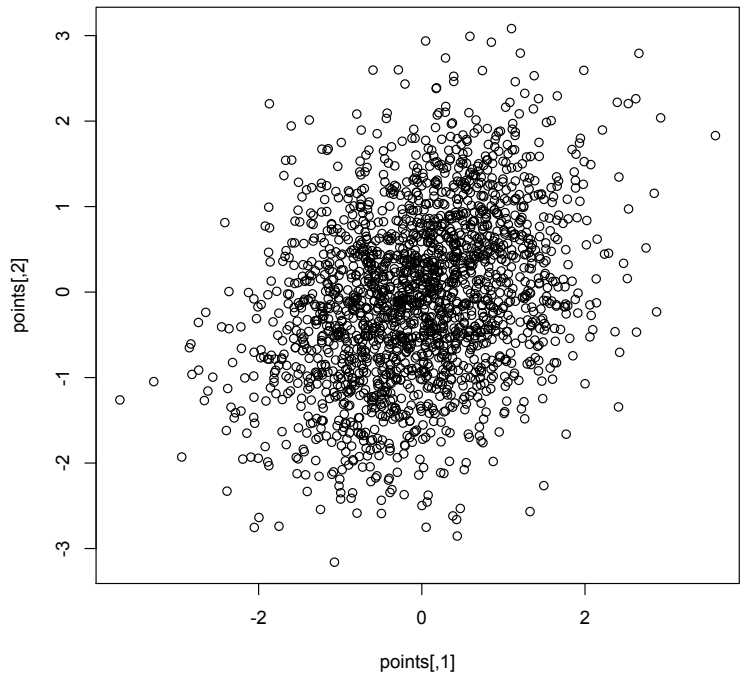


In blue we see the best-fitting normal curve... insufficient !

Example of heavy-tailed data from Finance

200 financial data VS 2000 simulated normal data





The simulated data stem from the best-fitting normal distribution for the left-side data. We clearly see that none of the 2000 simulated points reaches on the X_2 axis such negative values, hence the normal is not capable of predicting such extreme events !

Distribution on $\mathbb{N}$: The Poisson distribution

It counts the number of events happening in a given time interval. The probabilities are given by

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

with k = number of events and λ = the average frequency.

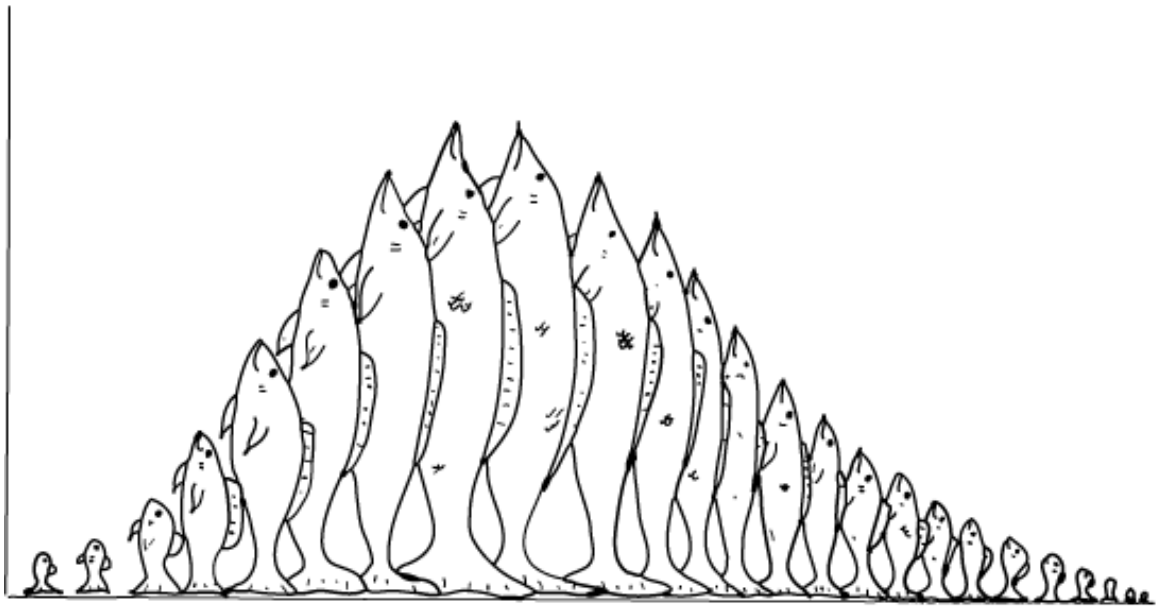
Expectation and variance are extremely simple :

$$E(X) = Var(X) = \lambda$$

Moreover it enjoys nice properties : if $X_1 \sim Po(\lambda_1), X_2 \sim Po(\lambda_2)$ are independent, then $X_1 + X_2 \sim Po(\lambda_1 + \lambda_2)$, while the difference follows the Skellam distribution.

The Poisson is related to various domains of application related to count data, it is related to Poisson processes, it is the limit of the law of rare events (or law of small numbers), etc.

Poisson Distribution



$$P\{x = i\} = e^{-\lambda} \cdot \lambda^i / i!$$

If you aren't laughing now then you have passed an "Im'm not a nerd" test.

Alternatives to the Poisson distribution

What strong limitation do you see in the Poisson distribution ?

Alternatives to the Poisson distribution

What strong limitation do you see in the Poisson distribution ?

A single parameter λ governs both the mean and the spread around the mean ! This can be quite limiting and easily lead to problems of overdispersion (variance larger than mean) or underdispersion (variance smaller than mean).

Alternatives to the Poisson distribution

What strong limitation do you see in the Poisson distribution ?

A single parameter λ governs both the mean and the spread around the mean ! This can be quite limiting and easily lead to problems of overdispersion (variance larger than mean) or underdispersion (variance smaller than mean).

A remedy to overdispersion is the negative binomial distribution whose probabilities are given by

$$\mathrm{P}(X = x) = \binom{x + r - 1}{r - 1}(1 - p)^r p^x$$

with $r \geq 0$ the number of failures before the experiment is stopped, $p \in (0, 1)$ the probability of success, and x represents the number of successes. It is denoted NB(r, p). The famous geometric distribution is a particular case corresponding to NB$(1, 1 - p)$.

Its characteristics are given by

$$\mathrm{E}[X] = \frac{pr}{1-p} \quad \mathrm{Var}[X] = \frac{pr}{(1-p)^2}.$$

We clearly see overdispersion from these expressions.

The NB is a popular choice for modelling (only !) overdispersion in various statistical methods and is well present in many softwares (e.g. SAS, R). This typically happens for meteorological phenomena.

It is to be noted that the negative binomial arises as a Poisson mixture distribution, that is, a Poisson distribution where the parameter λ is itself a random variable following a Gamma distribution.

Alternatives to the Poisson distribution

If we wish to model both over- and underdispersion, then the Conway-Maxwell-Poisson distribution is a good choice. The probabilities are given by

$$\mathrm{P}(X = x) = \frac{\lambda^x}{(x!)^\nu} \frac{1}{\sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}}$$

where $\lambda, \nu > 0$ or $0 < \lambda < 1$ if $\nu = 0$. Obviously, we retrieve the Poisson when $\nu = 1$, when $\nu = 0$ we find the geometric distribution and when $\nu \to \infty$ the limiting distribution is Bernoulli with success parameter $\lambda/(1 + \lambda)$.

The variance can be both larger and smaller than the mean.

A quality check

COM-Poisson Distribution Properties

- Simulation studies demonstrate COM-Poisson flexibility
  - Table II assesses goodness of fit on simulated data of size 500

Table II. True model parameters versus model estimates (and associated goodness-of-fit p-values provided in parentheses) for various assumed distributions

	Estimated parameter			
True distribution	Poisson	Geometric	COM-Poisson	
Poisson($\lambda=10$)	$\lambda=9.986$ (0.9436)	$p=0.091$ (0.0000)	$\lambda=10.244$ (0.8904)	$\nu=1.011$
Geometric($p=0.2$)	$\lambda=3.862$ (0.0000)	$p=0.206$ (0.6220)	$\lambda=0.794$ (0.6220)	$\nu=0.000$
COM-Poisson($\lambda=10,\ \nu=5$)	$\lambda=1.184$ (0.0000)	$p=0.458$ (0.0000)	$\lambda=12.223$ (0.4284)	$\nu=5.267$
COM-Poisson($\lambda=3,\ \nu=0.5$)	$\lambda=9.510$ (0.0000)	$p=0.095$ (0.0000)	$\lambda=3.157$ (0.6604)	$\nu=0.522$

Simulations done by Kimberly F. Sellers in her presentation "A flexible statistical control chart for dispersed count data"

Alternatives to the Poisson distribution

Sometimes it happens that the number of 0 occurrences is much higher than could model the Poisson.

- Example : Counts of movements of fetal lambs in five-second intervals :

No. of movements	0	1	2	3	4	5	6	7
Counts	182	41	12	2	2	0	0	1

- With the Poisson distribution, we would get the following expected counts :

No. of movements	0	1	2	3	4	5	6	7
Counts	167.7	60.1	10.8	1.3	0.1	0	0	0

The chi-square GOF test yields a p-value 0.00025.

- Instead consider the zero-inflated Poisson (ZIP) model :

$$P(Y = y) = pI(y = 0) + (1 - p)f_X(y; \lambda)$$

  with unknown parameters p and λ.
- Using this model, we can calculate expected counts (non-trivial calculations !) :

No. of movements	0	1	2	3	4	5	6	7
Counts	182	36.9	15.6	4.4	0.9	0.2	0	0

Distribution on $\mathbb{R}^+$: The exponential distribution

The exponential distribution is typically used to model times, sizes, time-to-events. Its density is given by

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

with $\lambda =$ frequency, number of events per time unit.

The cumulative distribution function $F(x)$ corresponds to $F(x) = 1 - e^{-\lambda x}$, the expectation to $E(X) = \frac{1}{\lambda}$ (aha ! the meaning of λ becomes clearer) and the variance to $Var(X) = \frac{1}{\lambda^2}$.

The exponential distribution enjoys a unique property : it has no memory !

$$P(X > t + x \,|\, X > t) = P(X > x)!$$

In order to model size type, survival type, or other data on $\mathbb{R}^+$, the exponential is often an oversimplified choice due to its unique parameter λ.

In order to model size type, survival type, or other data on $\mathbb{R}^+$, the exponential is often an oversimplified choice due to its unique parameter λ.

The exponential is strictly decreasing, so cannot model data that first increase and then decrease. This also prevents it from modelling heavy-tailed data.

Its memoryless property also has drawbacks. This can be seen from the concept of failure rate in survival analysis. The failure rate, mostly also known as hazard rate, is defined as

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

where f is the density and $1 - F$ the survival or reliability function. In case of the exponential we obtain

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

A constant failure rate means that time plays no role in the survival. This is of course unrealistic when talking for instance about mortality rates.

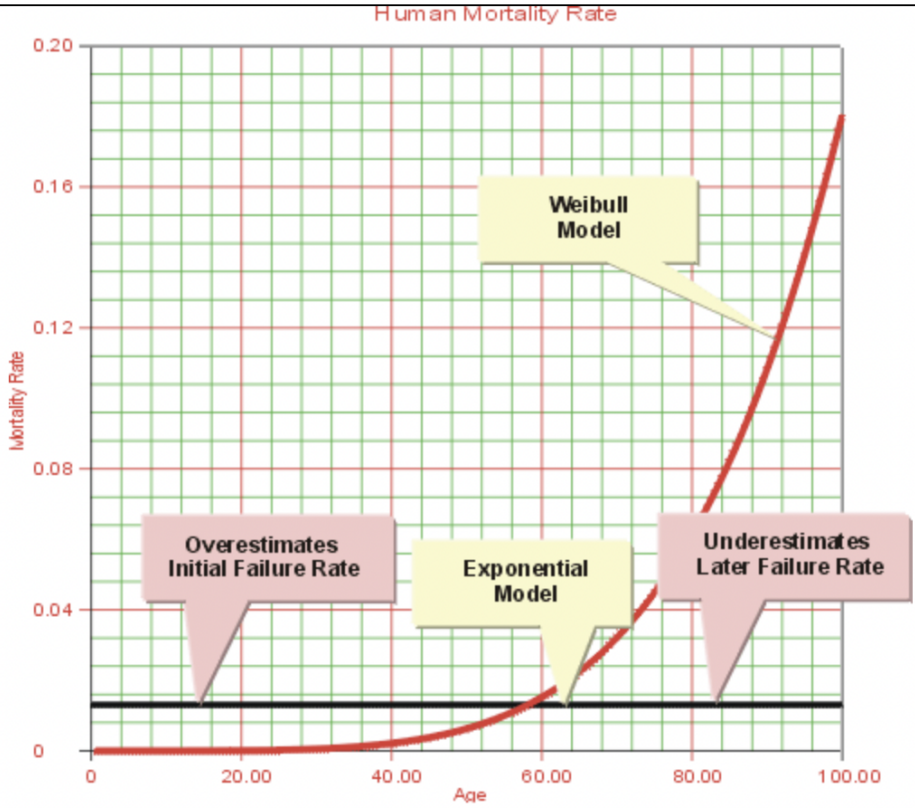


Figure 1: Human mortality rate analyzed with the exponential and Weibull distributions

https ://www.reliasoft.com/resources/resource-center/limitations-of-the-exponential-distribution-for-reliability-analysis

A natural alternative to the exponential (among many !)

Think of a procedure that consists of α independent steps, and each step takes a random exponential time T to be realized. The total time then follows a Gamma distribution.

The Gamma density is

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

with $\lambda =$ frequency, $\alpha =$ shape parameter, and $\Gamma(\alpha)$ the Gamma function. The special case $\alpha = 1$ yields the exponential.

Expectation and variance are respectively given by

$$E(X) = \frac{\alpha}{\lambda} \quad \text{and} \quad Var(X) = \frac{\alpha}{\lambda^2}.$$

Illustration

Gamma density for $\lambda = 0.5$ and $\alpha =$ 0.5 (blue), 1 (red) and 2 (yellow).