ENM 531: Data-driven Modeling and Probabilistic Scientific Computing

Lecture #2: Primer on Probability and Statistics

Paris Perdikaris
January 26, 2021

Penn
UNIVERSITY of PENNSYLVANIA

Discrete random variables

- A ***discrete random variable*** is one which may take on only a countable number of distinct values such as 0,1,2,3,4,........ Discrete random variables are usually (but not necessarily) counts. If a random variable can take only a finite number of distinct values, then it must be discrete. Examples of discrete random variables include the number of children in a family, the Friday night attendance at a cinema, the number of patients in a doctor's surgery, the number of defective light bulbs in a b
- The ***probability distribution*** of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function.

Continuous random variables

- A ***continuous random variable*** is one which takes an infinite number of possible values. Continuous random variables are usually measurements. Examples include height, weight, the amount of sugar in an orange, the time required to run a mile.
- A continuous random variable is not defined at specific values. Instead, it is defined over an *interval* of values, and is represented by the ***area under a curve*** (in advanced mathematics, this is known as an *integral*). The probability of observing any single value is equal to 0, since the number of values which may be assumed by the random variable is infinite.

- Suppose a random variable X may take all values over an interval of real numbers. Then the probability that X is in the set of outcomes A, $P(A)$, is defined to be the area above A and under a curve. The curve, which represents a function $p(x)$, must satisfy the following:
- ***1:*** *The curve has no negative values ($p(x) \geq 0$ for all x)*
- ***2:*** *The total area under the curve is equal to 1.*
- A curve meeting these requirements is known as a ***density curve***.

Probability density functions

$$\Pr[a \le X \le b] = \int_a^b f_X(x)\,dx.$$

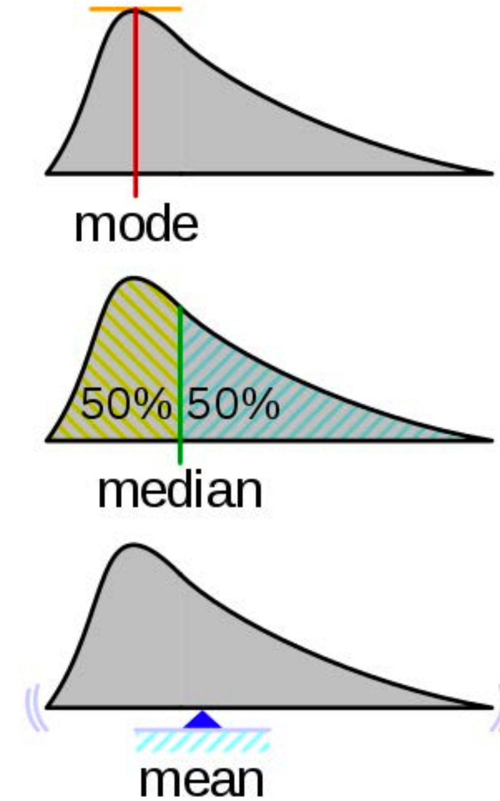
Hence, if F_X is the cumulative distribution function of X, then:

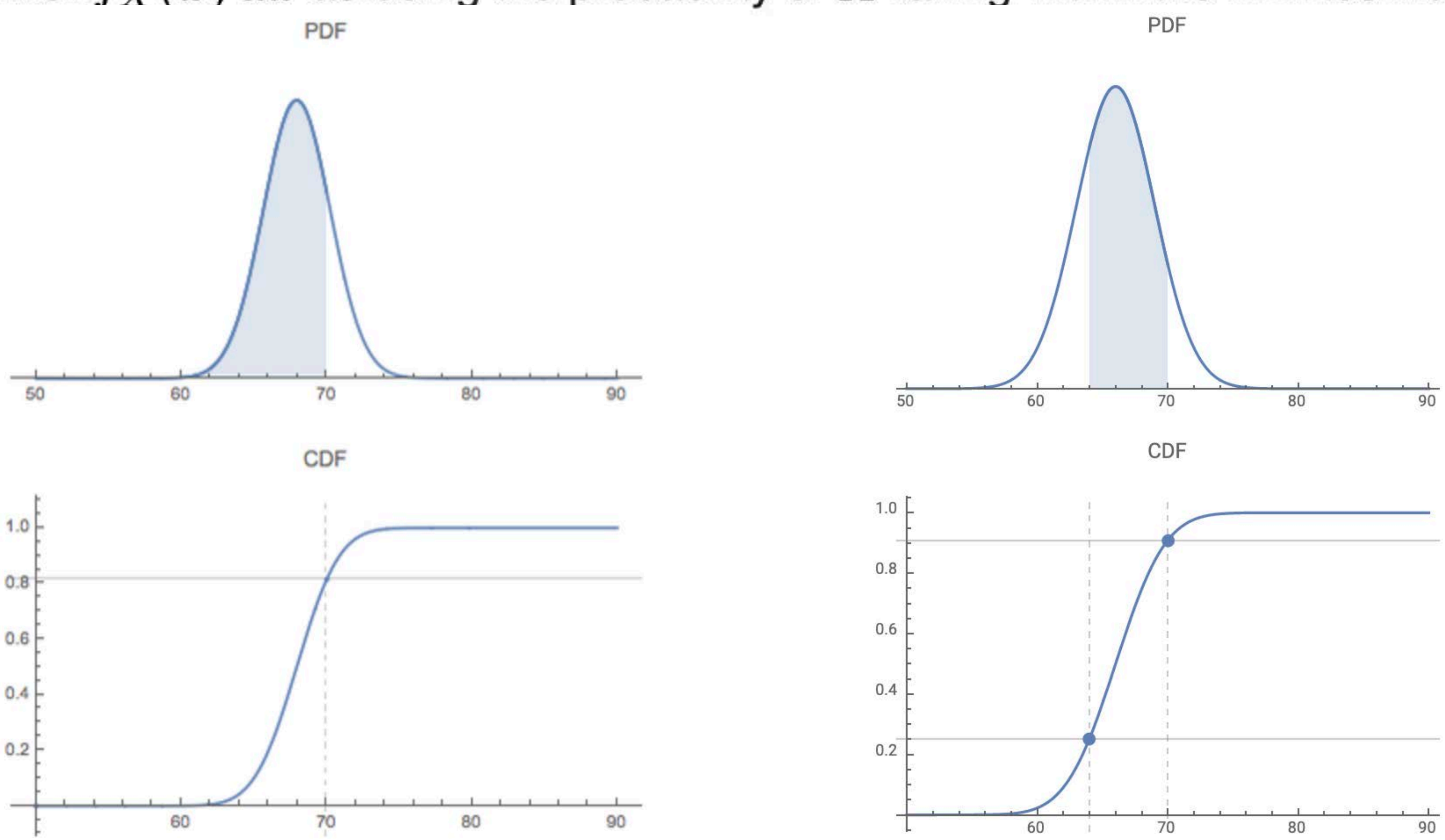
$$F_X(x) = \int_{-\infty}^x f_X(u)\,du,$$

and (if f_X is continuous at x)

$$f_X(x) = \frac{d}{dx}F_X(x).$$

Intuitively, one can think of $f_X(x)\,dx$ as being the probability of X falling within the infinitesimal interval $[x, x + dx]$.



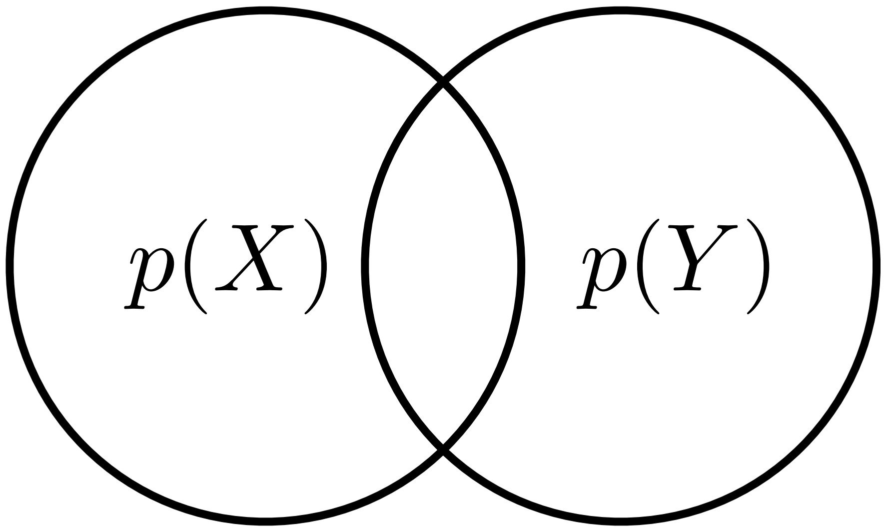


Basic rules of probability

Sum rule $$p(X) = \sum_Y p(X, Y)$$

Product rule $$p(X, Y) = p(Y|X)p(X)$$

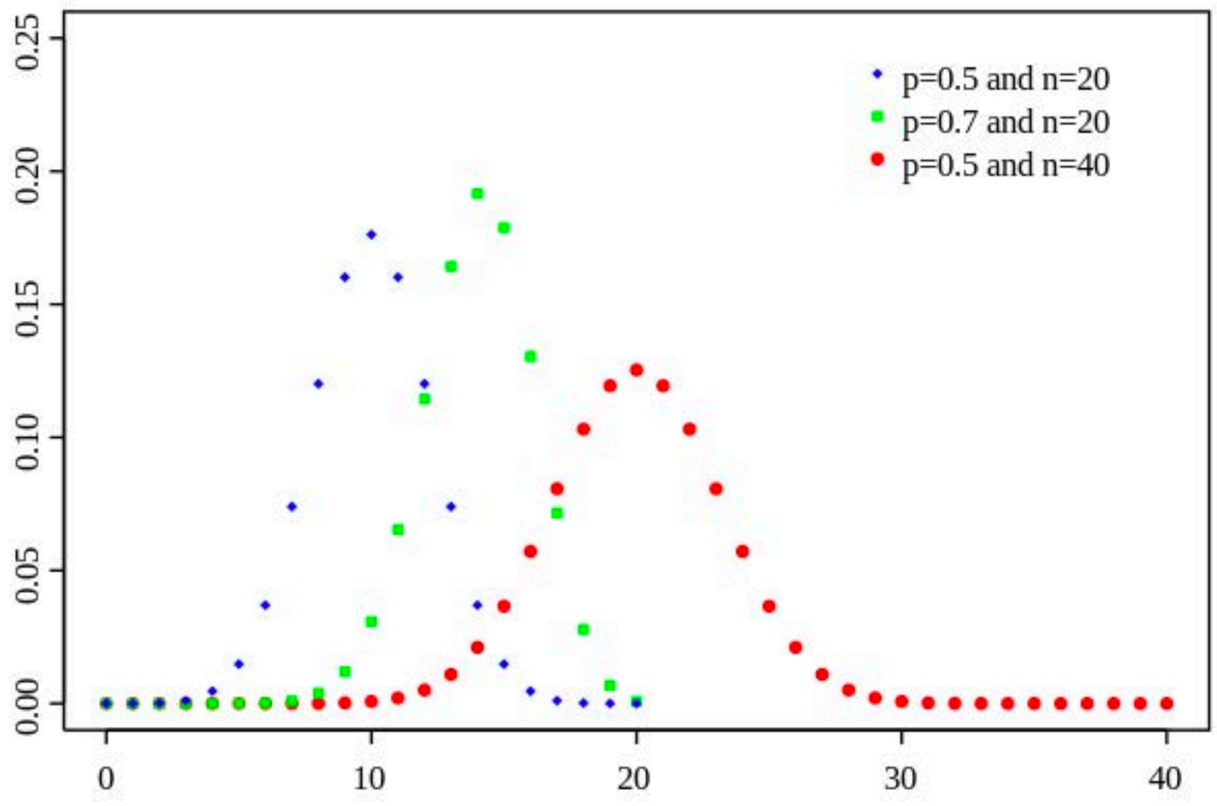
Bayes rule $$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

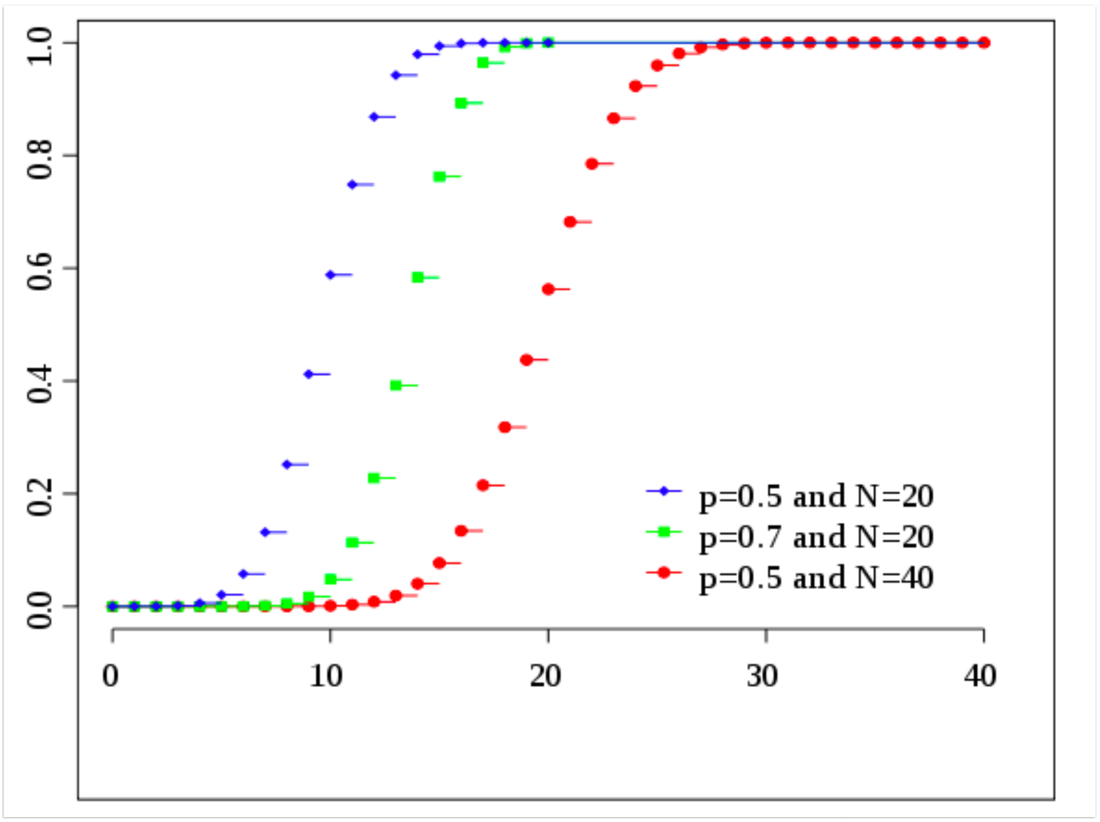


Venn diagrams

Reading: Chapters 1,2 from Bishop "Pattern Recognition and Machine Learning

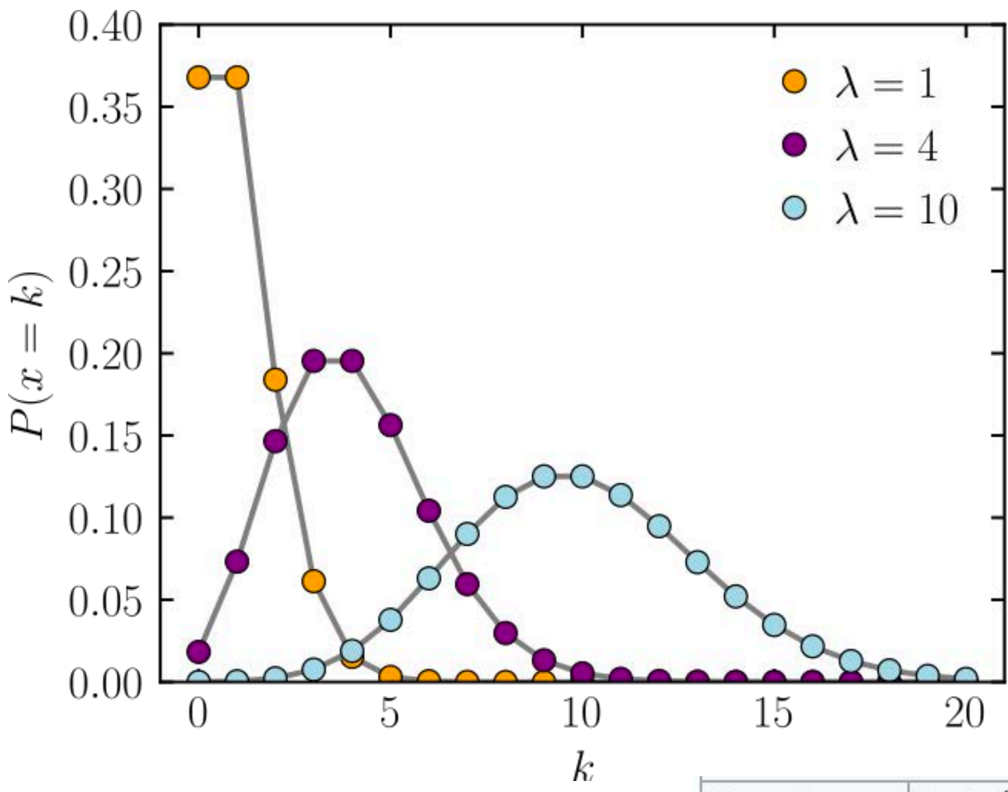
The binomial distribution

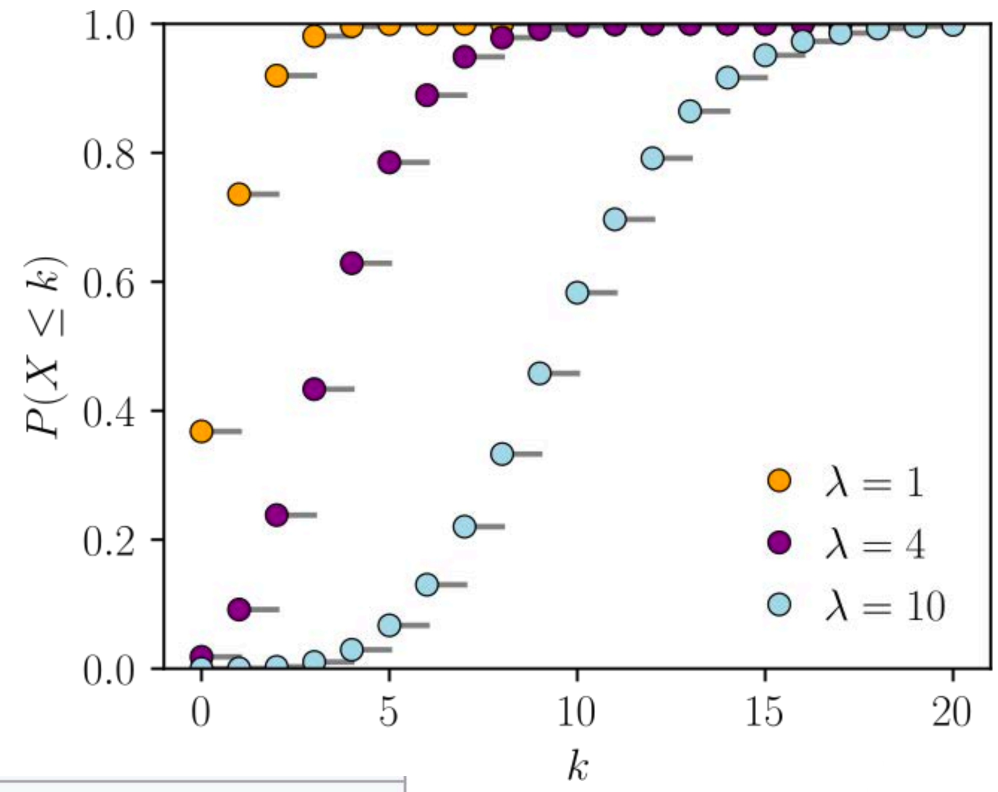




Notation	$B(n,p)$
Parameters	$n \in \{0,1,2,\ldots\}$ – number of trials $p \in [0,1]$ – success probability for each trial $q = 1 - p$
Support	$k \in \{0,1,\ldots,n\}$ – number of successes
PMF	$\binom{n}{k} p^k q^{n-k}$
CDF	$I_q(n-k, 1+k)$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	npq

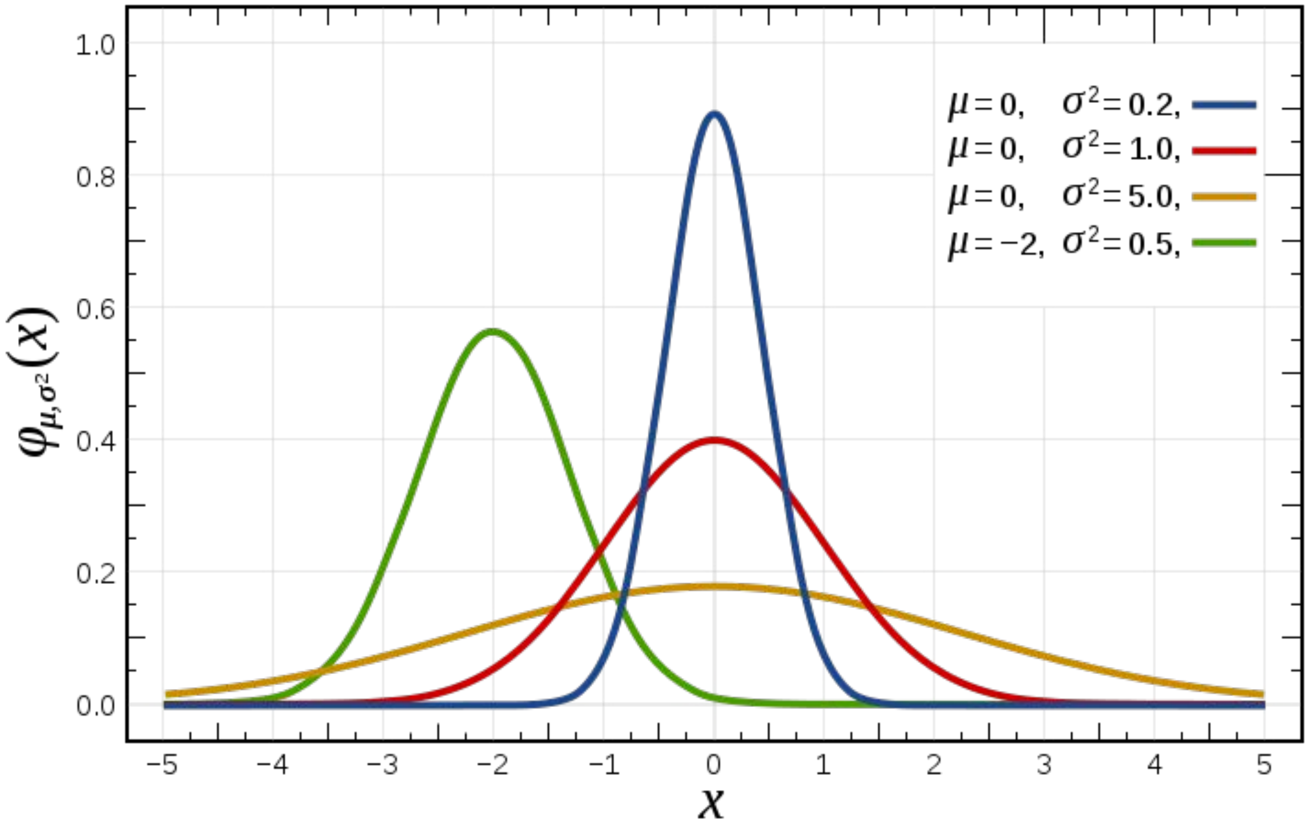
The Poisson distribution

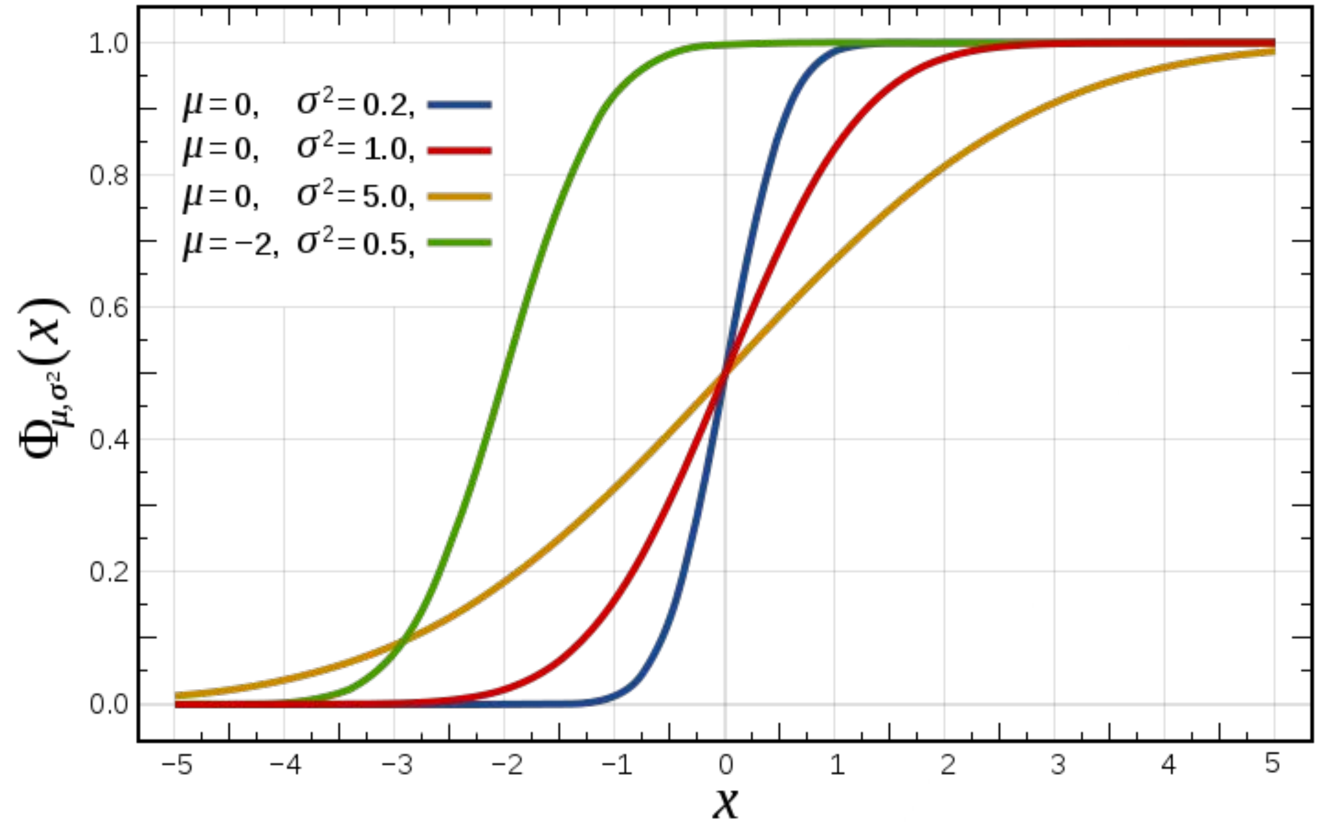




Notation	$\mathrm{Pois}(\lambda)$
Parameters	$\lambda \in (0, \infty)$ (rate)
Support	$k \in \mathbb{N}_0$ (Natural numbers starting from 0)
PMF	$\dfrac{\lambda^k e^{-\lambda}}{k!}$
CDF	$\dfrac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$, or $e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \dfrac{\lambda^i}{i!}$, or $Q(\lfloor k+1 \rfloor, \lambda)$ (for $k \geq 0$, where $\Gamma(x, y)$ is the upper incomplete gamma function, $\lfloor k \rfloor$ is the floor function, and Q is the regularized gamma function)
Mean	λ
Median	$\approx \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$
Mode	$\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$
Variance	λ

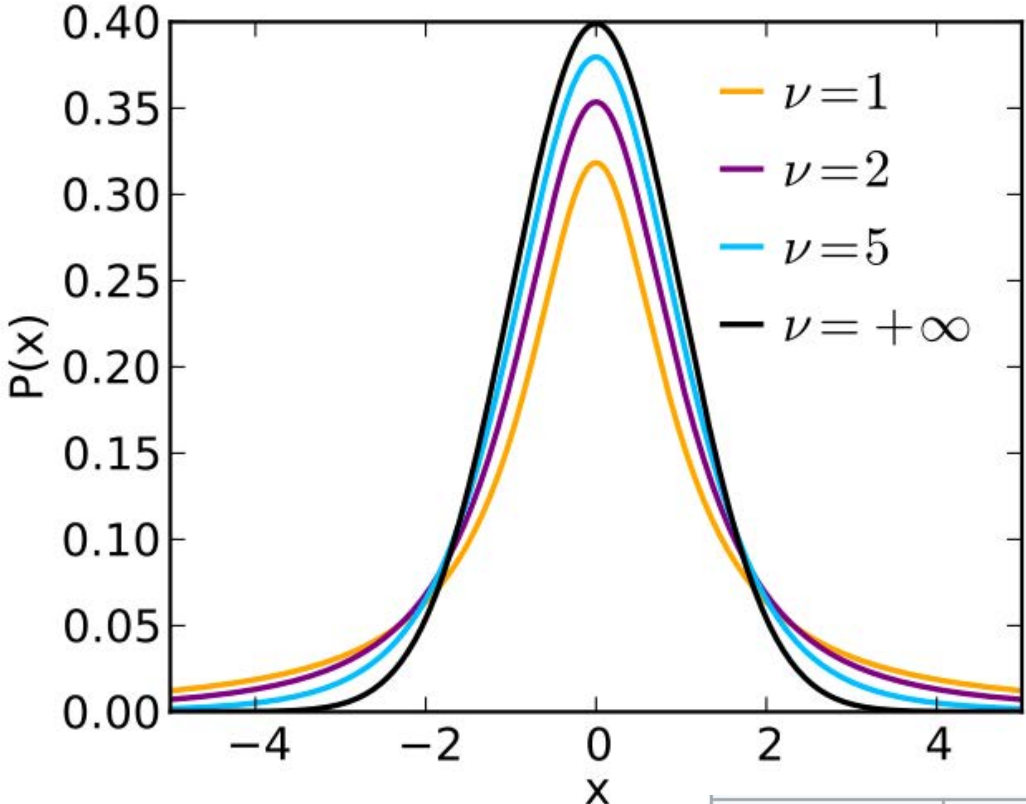
The Gaussian distribution

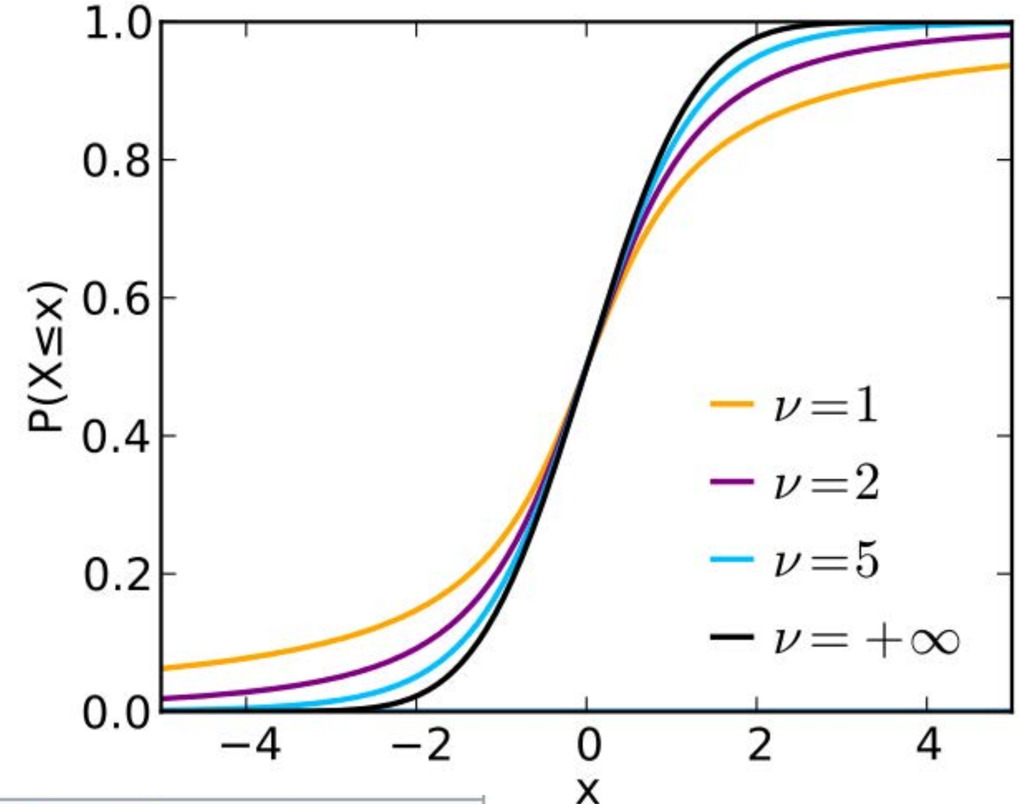




Notation	$\mathcal{N}(\mu,\, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 > 0$ = variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
Quantile	$\mu + \sigma\sqrt{2}\operatorname{erf}^{-1}(2F-1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2

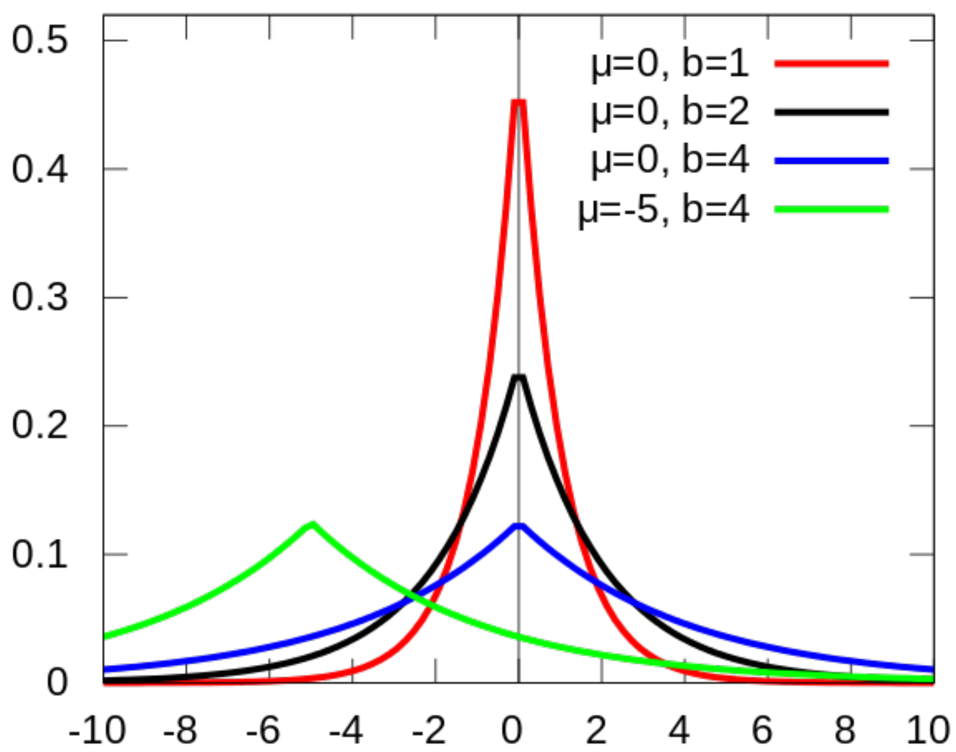
The Student-t distribution

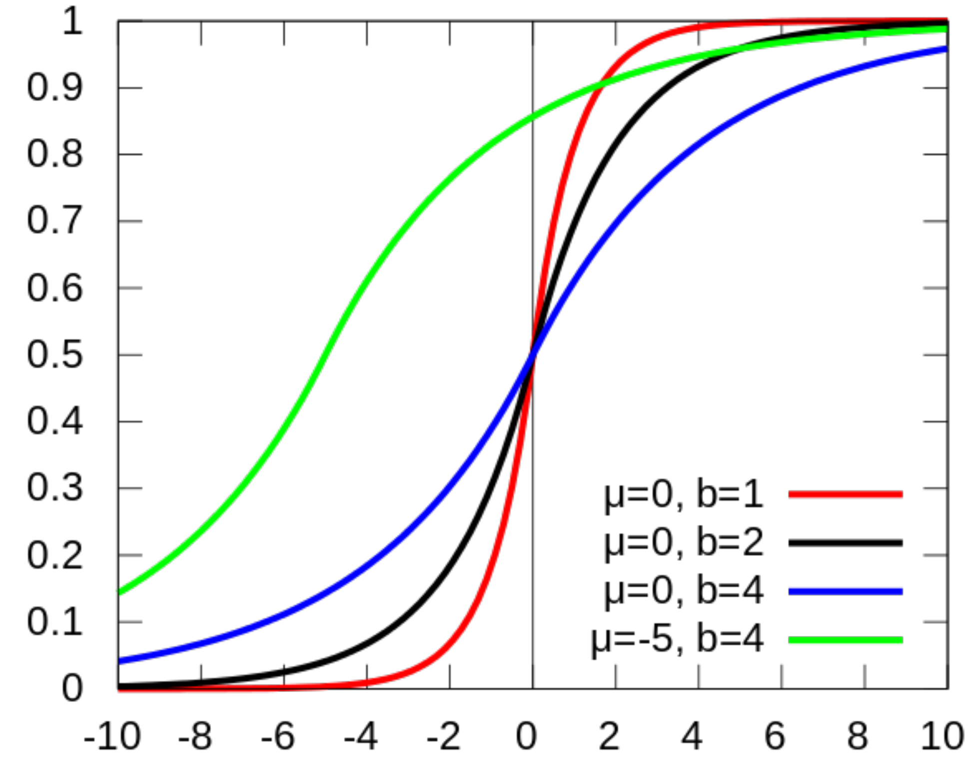




Parameters	$\nu > 0$ degrees of freedom (real)
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$\frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \times \frac{{}_2F_1\left(\frac{1}{2},\frac{\nu+1}{2};\frac{3}{2};-\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}$ where ${}_2F_1$ is the hypergeometric function
Mean	0 for $\nu > 1$, otherwise undefined
Median	0
Mode	0
Variance	$\frac{\nu}{\nu-2}$ for $\nu > 2$, ∞ for $1 < \nu \leq 2$, otherwise undefined

The Laplace distribution





Parameters	μ location (real) $b > 0$ scale (real)
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{1}{2b} \exp\left(-\frac{\lvert x-\mu \rvert}{b}\right)$
CDF	$\begin{cases} \frac{1}{2}\exp\left(\frac{x-\mu}{b}\right) & \text{if } x < \mu \\ 1 - \frac{1}{2}\exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu \end{cases}$
Mean	μ
Median	μ
Mode	μ
Variance	$2b^2$

Gaussian vs Student-t vs Laplace

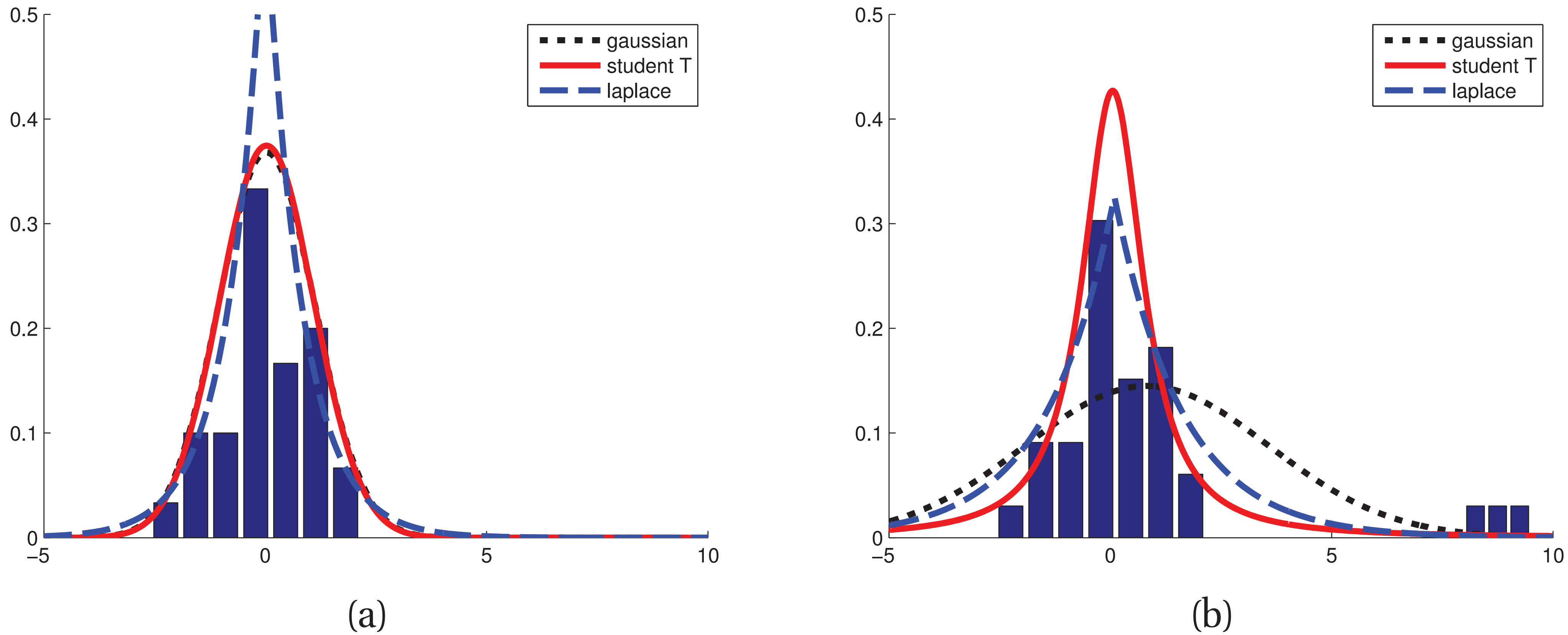
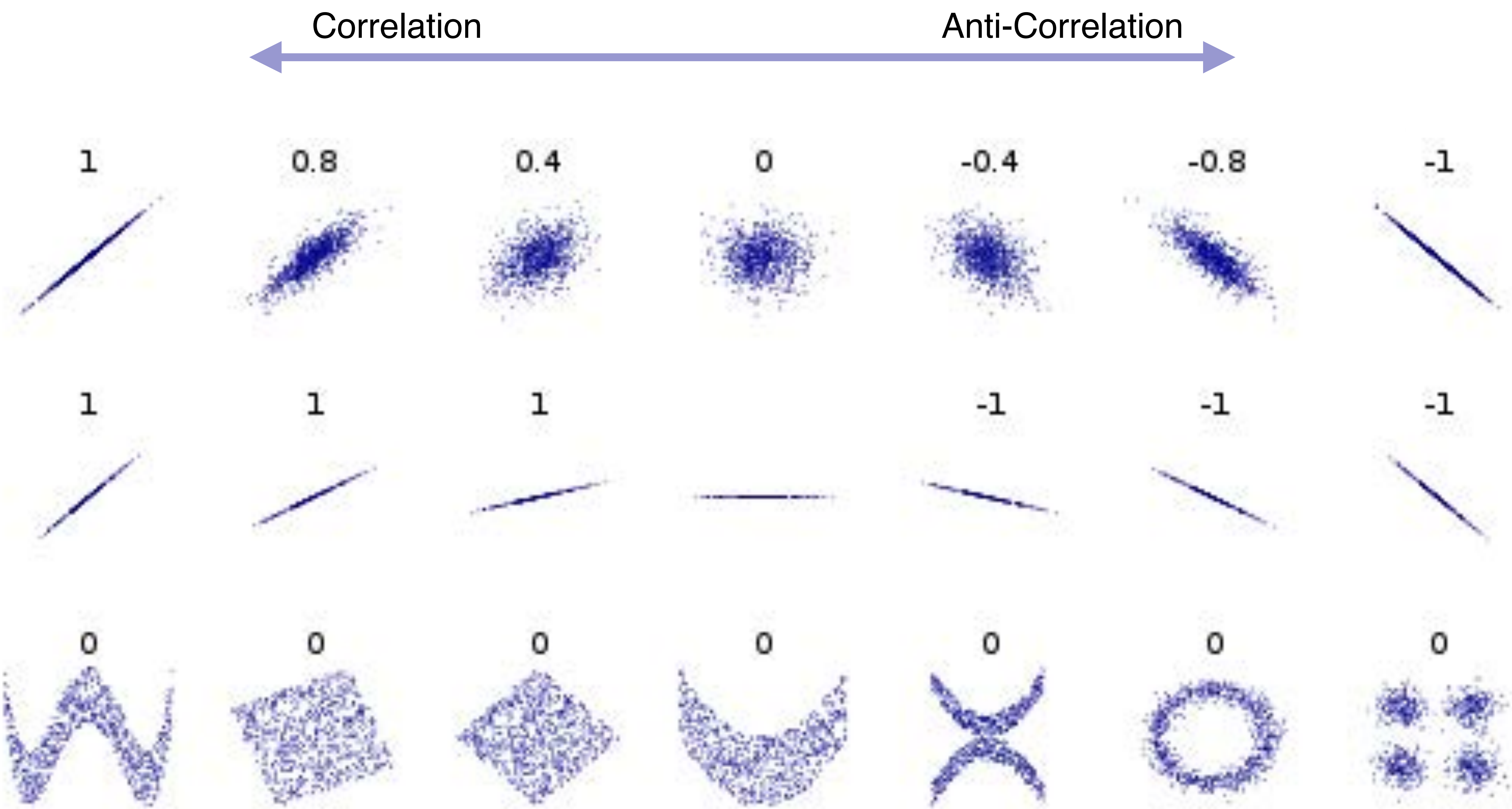
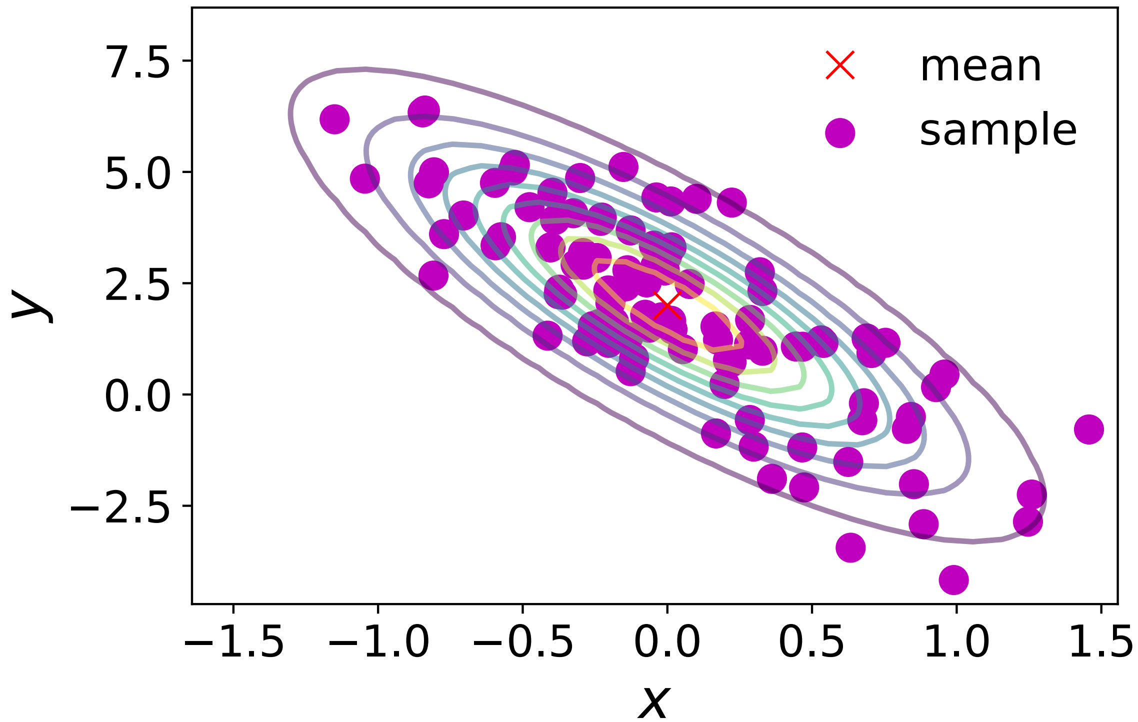


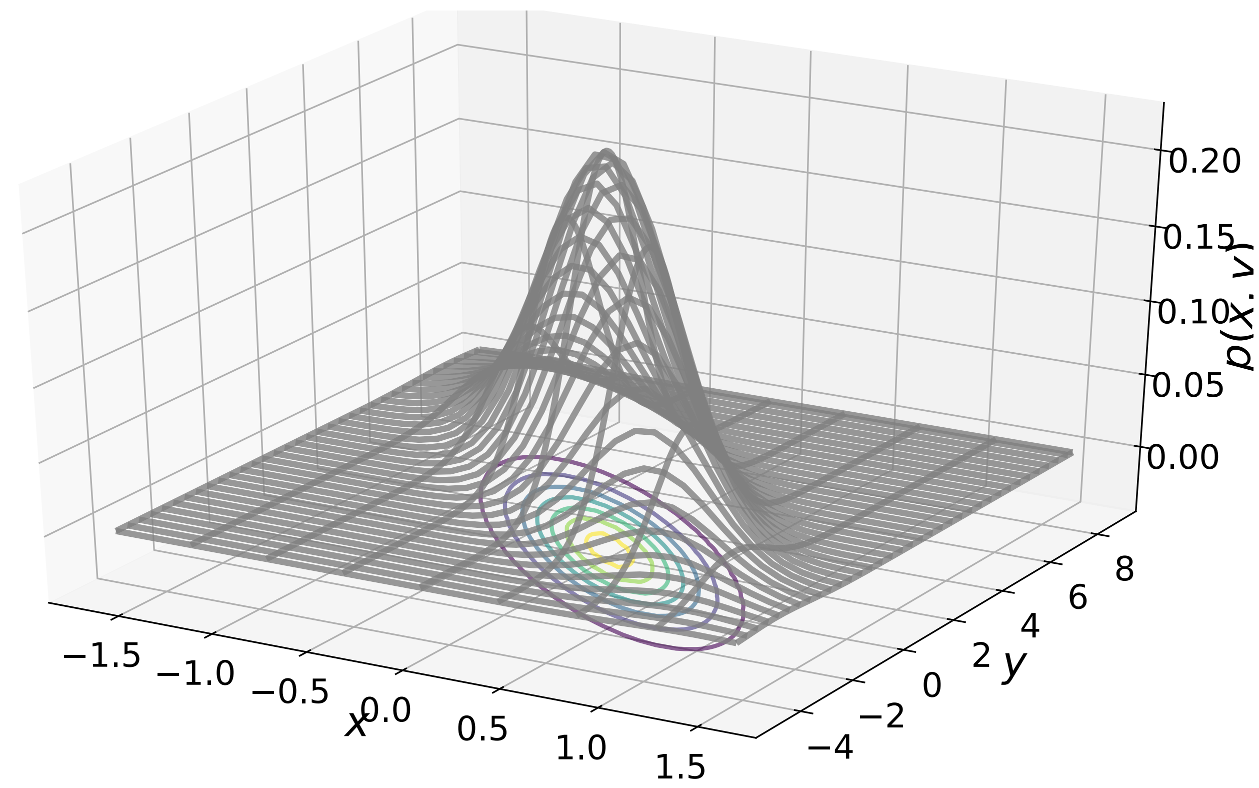
Figure 2.8 Illustration of the effect of outliers on fitting Gaussian, Student and Laplace distributions. (a) No outliers (the Gaussian and Student curves are on top of each other). (b) With outliers. We see that the Gaussian is more affected by outliers than the Student and Laplace distributions. Based on Figure 2.16 of (Bishop 2006a). Figure generated by `robustDemo`.

Correlation and linear dependence



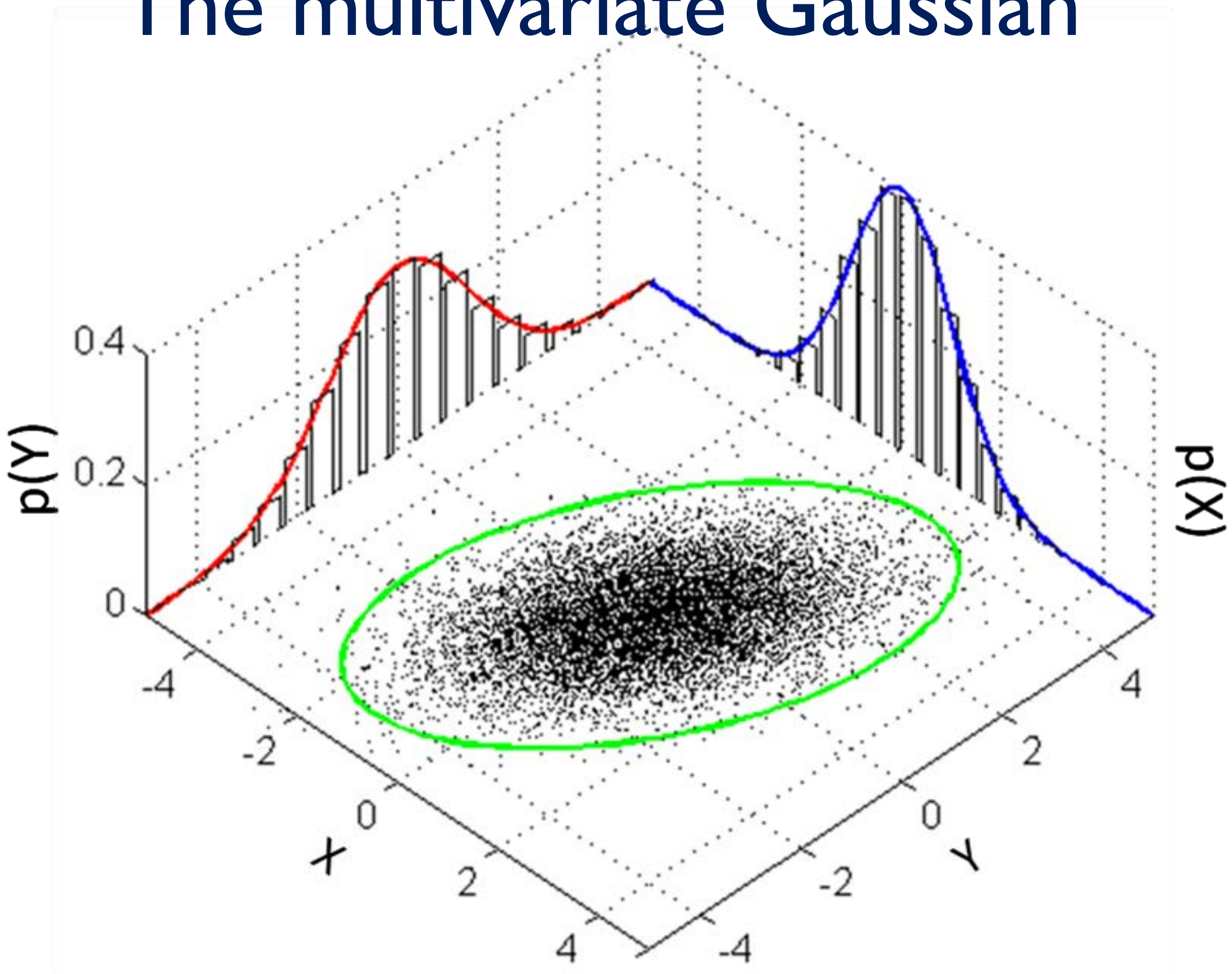
The multivariate Gaussian





$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\big(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big)$$

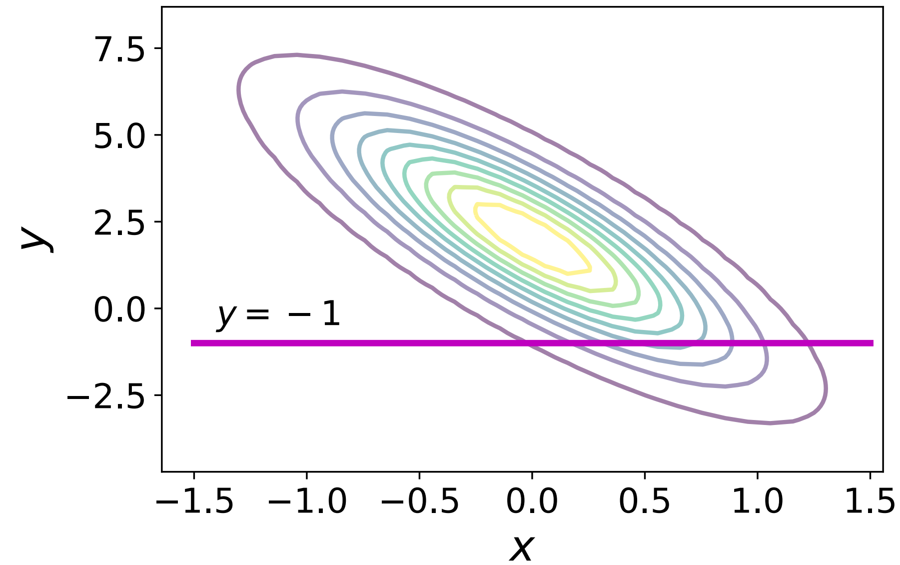
The multivariate Gaussian



Notation	$\mathcal{N}(\boldsymbol{\mu},\, \boldsymbol{\Sigma})$
Parameters	$\boldsymbol{\mu} \in \mathbf{R}^k$ — location $\boldsymbol{\Sigma} \in \mathbf{R}^{k \times k}$ — covariance (positive semi-definite matrix)
Support	$\boldsymbol{x} \in \boldsymbol{\mu} + \text{span}(\boldsymbol{\Sigma}) \subseteq \mathbf{R}^k$
PDF	$\det(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$, exists only when $\boldsymbol{\Sigma}$ is positive-definite
Mean	$\boldsymbol{\mu}$
Mode	$\boldsymbol{\mu}$
Variance	$\boldsymbol{\Sigma}$

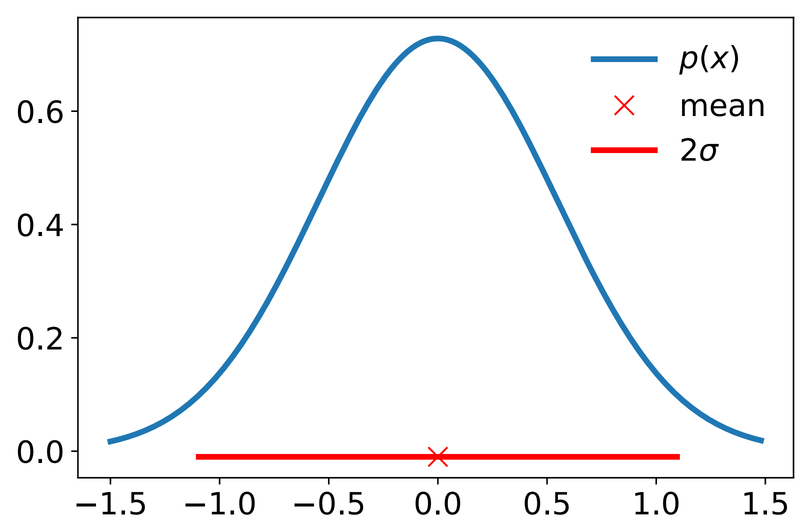
Marginals and conditionals of a Gaussian

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right)$$



Marginal distribution

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} = \mathcal{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx})$$

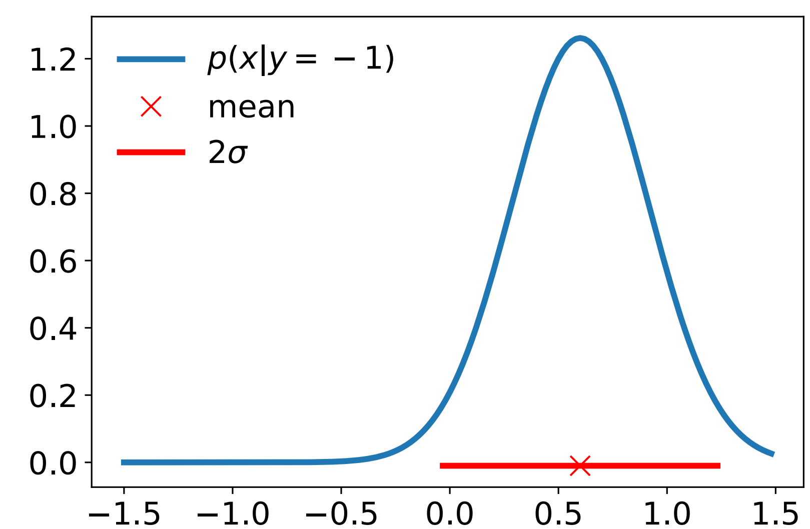


Conditional distribution

$$p(\boldsymbol{x} \,|\, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_{x\,|\,y}, \boldsymbol{\Sigma}_{x\,|\,y})$$

$$\boldsymbol{\mu}_{x\,|\,y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y)$$

$$\boldsymbol{\Sigma}_{x\,|\,y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx} \,.$$



These are unique properties that make the Gaussian distribution very simple and attractive to compute with! It is essentially our main building block for computing under uncertainty.