Probability and Statistical Distributions

Xuening Bai (白雪宁)

Institute for Advanced Study (IASTU) &

Department of Astronomy (DoA)

清華大學
Tsinghua University

Nov. 5, 2024

Bayes' theorem

From the law of total probability, the Bayes' theorem reads:

Prior probability

$$P(A_i|B) = \frac{P(A_iB)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)}$$

Posterior probability

This is the basis for Bayesian inference, to be covered later in this course.

Given B (data) has occurred, what is the probability of a model?

The denominator does not depend on i, this is usually expressed as

$$P(A_i|B) \propto P(B|A_i)P(A_i)$$

Likelihood Prior

The Error function

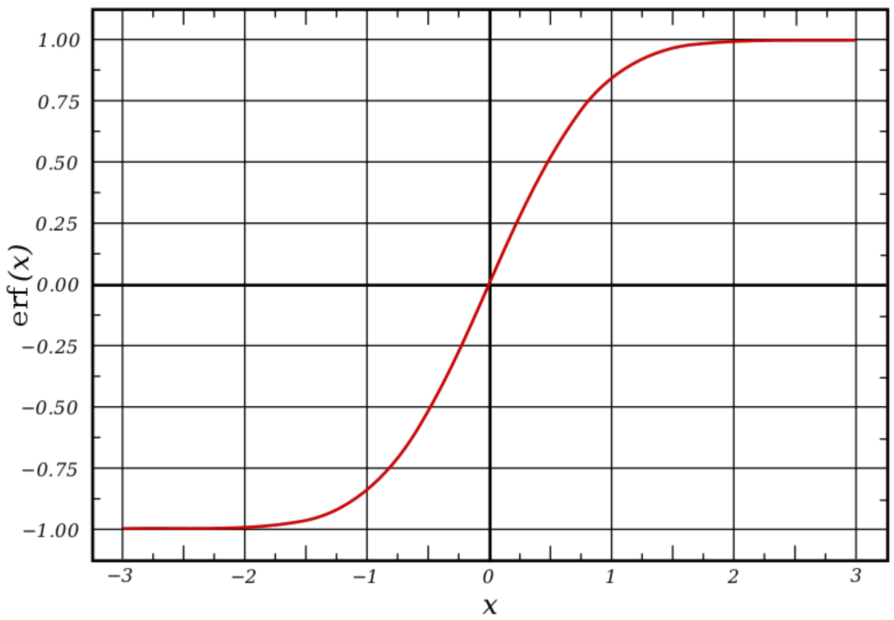
The CDF of a the standard normal distribution :

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

This is related to the error function

$$\mathrm{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt$$

Error function



There is also the complimentary error function:

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2} dt$$

They are related by $\Phi(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right] = \frac{1}{2}\mathrm{erfc}\left(-\frac{x}{\sqrt{2}}\right)$

The Pareto distribution

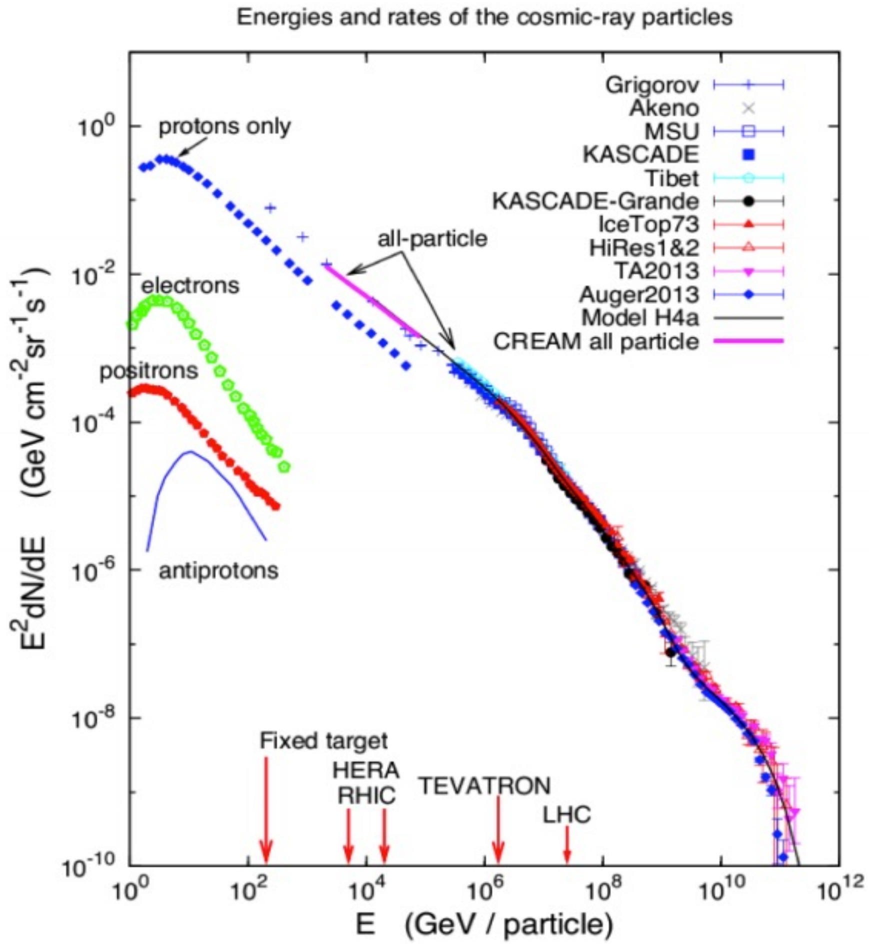
Essentially, a power law:

$$P(X > x) = \left(\frac{x}{x_{\min}}\right)^{-\alpha} \qquad \text{(CDF)}$$

$$0 < x_{\min} < x, \ \alpha > 0 \ .$$

The PDF is

$$f(x) = \begin{cases} 0 & x \le x_{\min} \ , \\ \alpha x_{\min}^{\alpha} x^{-(\alpha+1)} & x > x_{\min} \ . \end{cases}$$



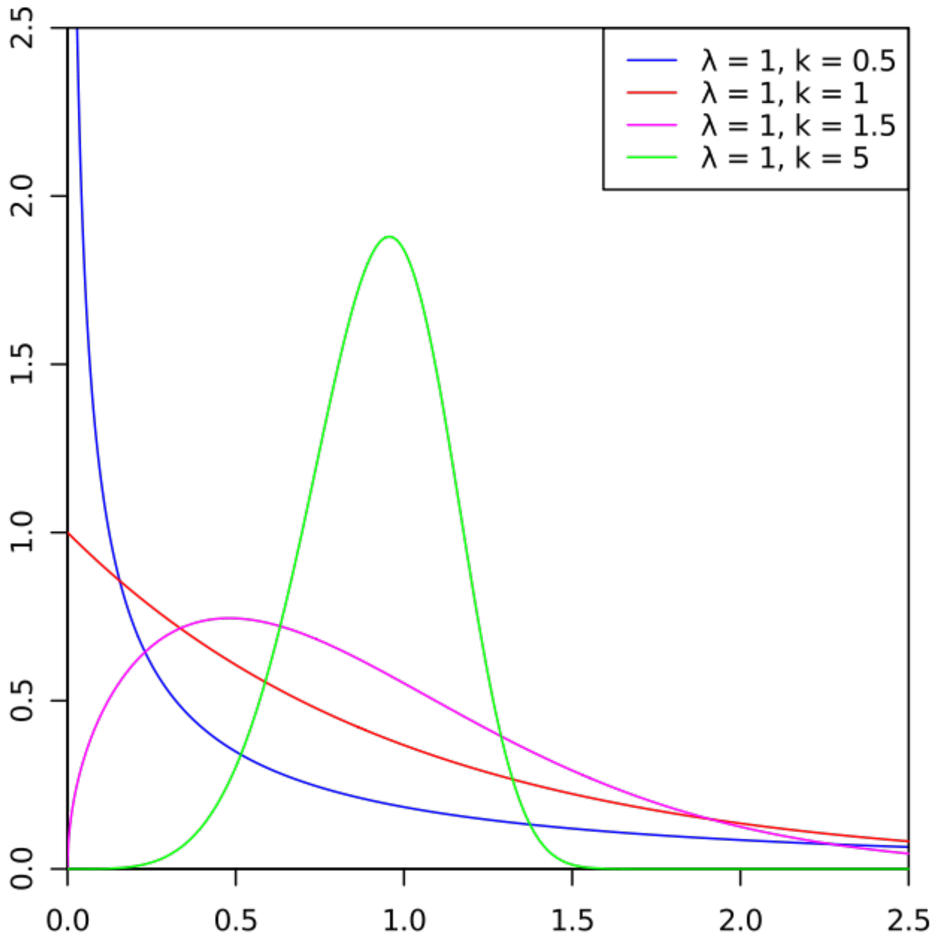
Occurs naturally in nature, especially associated with non-thermal processes that produce energetic particles (cosmic-rays).

Also common to fit data with piecewise power laws.

The Weibull distribution

The Weibull distribution is defined for $x \geq 0$, characterized by the shape parameter k and scale parameter λ:

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$$



It reduces to exponential distribution for k=1, and the Rayleigh distribution for k=2.

Its CDF is fairly simple:

$$F(x) = 1 - e^{-(x/\lambda)^k}$$

And is commonly associated with failure rate, particle size distribution, etc.

Chi squared distribution

If $Z_1,\ldots,Z_k$ are independent, standard normal random variables, then the sum of their squares:

$$\chi_k^2 = \sum_{i=1}^{k} Z_i^2$$

is distributed according to the chi-squared distribution with k degrees of freedom.

Its PDF reads: $$f(x) = \frac{x^{k/2-1} e^{-x/2}}{2^{k/2}\Gamma(k/2)}$$ (x>0)

Mainly used for hypothesis test, particularly the chi-squared test for goodness of fit (will be covered in the next lecture).

It is also customary to define the chi squared distribution per degree of freedom:

$$\chi^2_{\mathrm{dof}} \equiv \chi_k^2/k$$ a.k.a. reduced chi squared

It approaches $\mathcal{N}(1,\sqrt{2/k})$ for large k.

The gamma distribution

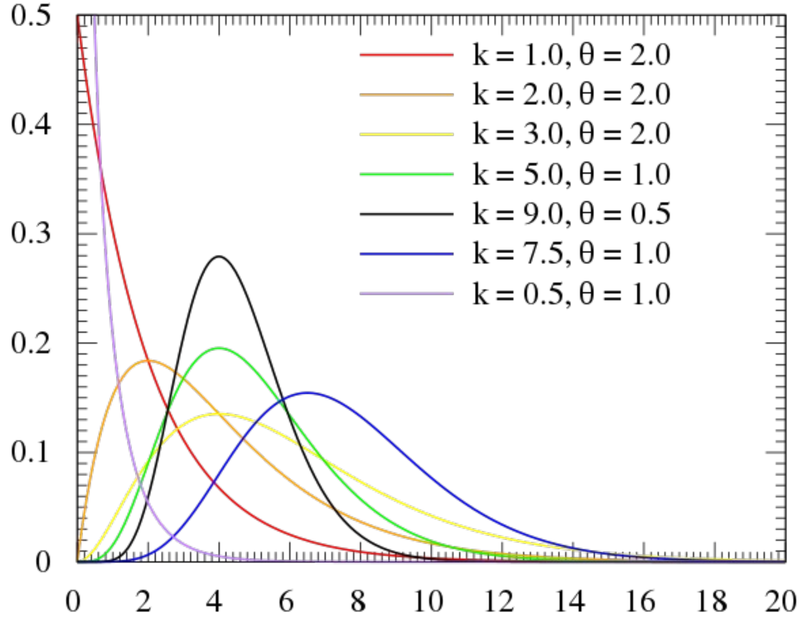
Recall that the Gamma function is defined as $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$

The gamma distribution is a two-parameter family defined as

$$f(x) = \frac{1}{\theta^k} \frac{x^{k-1} e^{-x/\theta}}{\Gamma(k)}$$

and is denoted as $X \sim \Gamma(k, \theta)$

(shape, scale)



The exponential distribution: k=1, θ=1/λ.

The chi squared distribution: k->k/2, θ=2.

The Schechter galaxy luminosity function: k~0.11, θ=1. (Schechter 1976)

It is a conjugate prior to several distributions including the exponential, normal and Poisson distributions (see later in the course).

The beta distribution

The beta distribution is a family of distribution functions defined in [0,1].

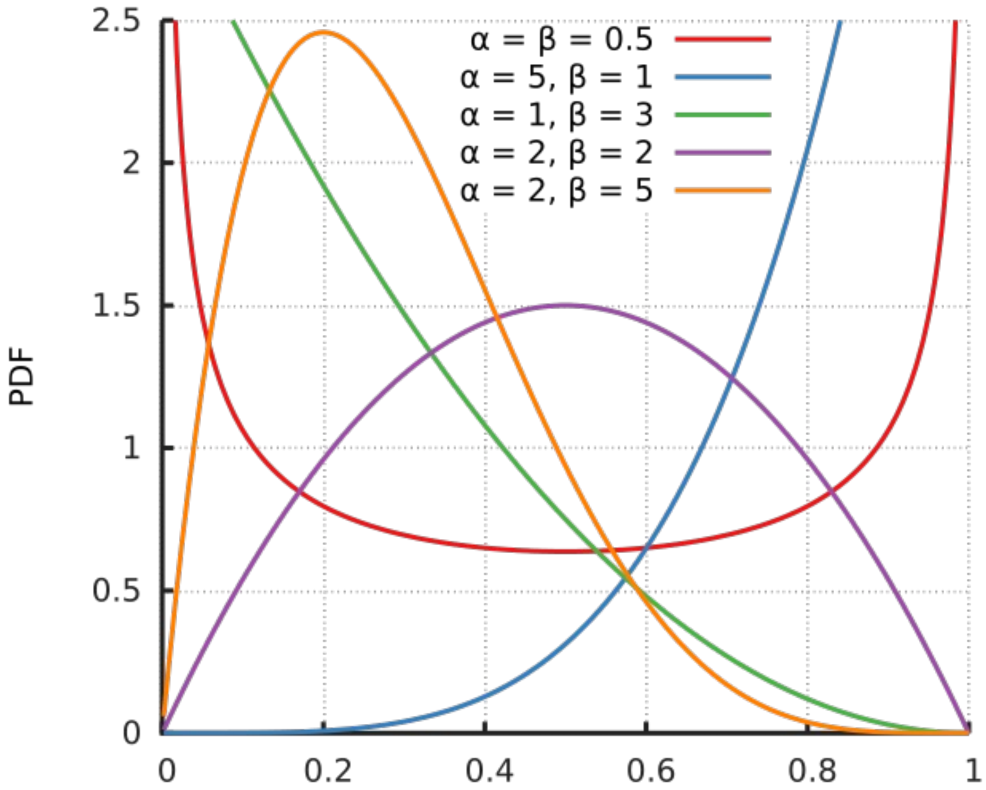
$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

The mean is $\alpha/(\alpha+\beta)$.

Can achieve a variety of shapes via combinations of α and β.

It is a useful distribution for random variables limited to finite intervals.

It is also the conjugate prior for the binomial distribution.

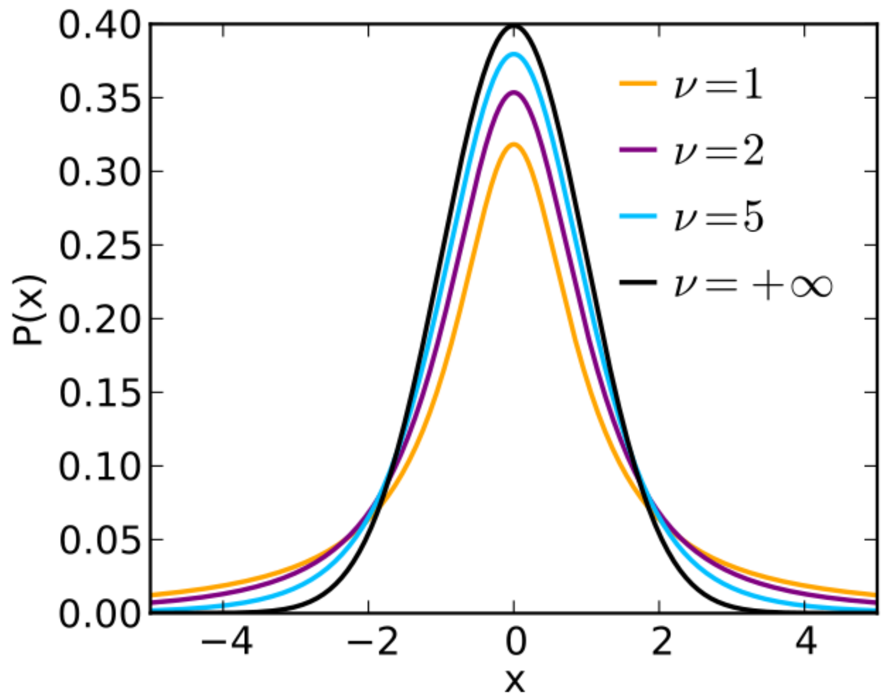


Student's t distribution

The PDF of student's t distribution with ν degrees of freedom reads

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

It approaches standard normal distribution for large ν.



If $X_1,\ldots,X_n$ are independent, standard normal random variables, define

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)\ ,\quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

(unbiased estimates of mean and variance)

Then the variable $T \equiv \frac{\sqrt{n}}{S_n}(\bar{X}_n - \mu)$ satisfies student's t distribution with n-1 degrees of freedom.

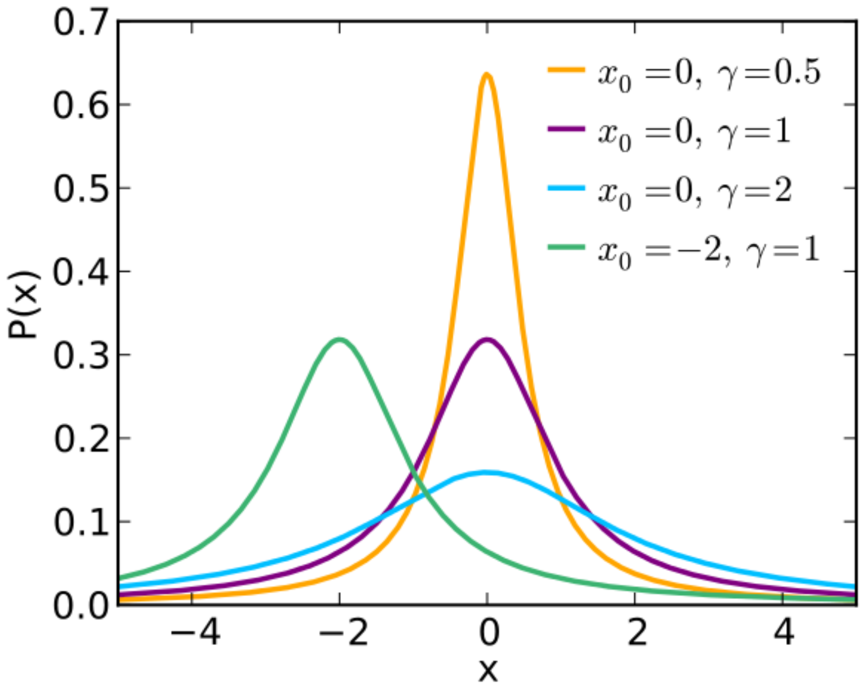
More next lecture.

Lorentzian/Cauchy distribution

The PDF of a Cauchy/Lorentzian distribution reads

$$f(x;\mu,\gamma) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{\gamma^2 + (x-\mu)^2}\right)$$

In standard form (μ=0, γ=1), it coincides with the student's t distribution with one degree of freedom.



A pathological whose mean/variance are undefined (diverge).

In spectroscopy, the shape of spectral lines are subject to several broadening mechanisms, some of which (collisional, natural) yield Lorentzian profiles.

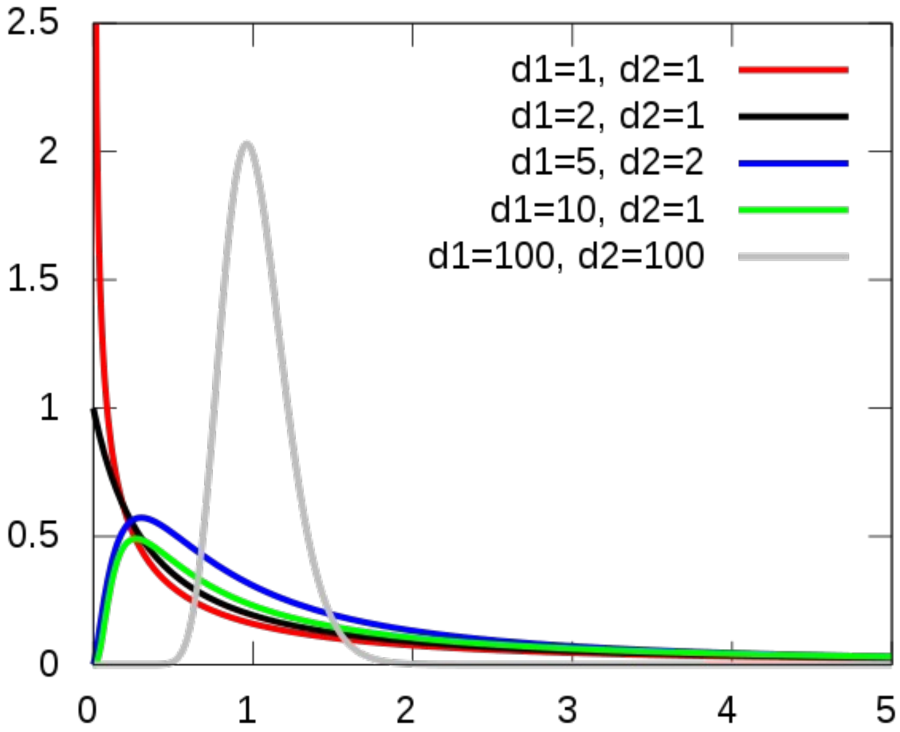
Fisher's F distribution

The Fisher's F distribution is a two-parameter family whose PDF reads

$$f(x) = B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)^{-1} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}}$$

beta function

This is usually denoted by $F(d_1,d_2)$.



This statistics arises from the ratio of two independent reduced chi squared:

$$X_1 \sim \chi^2_{d_1} \; , \; X_2 \sim \chi^2_{d_2}$$

Then $$\frac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2)$$

More next lecture.

Other useful statistics

Skewness: $$\Sigma \equiv \int \left(\frac{x-\mu}{\sigma}\right)^3 f(x)dx$$ How symmetric it is w.r.t. mean.

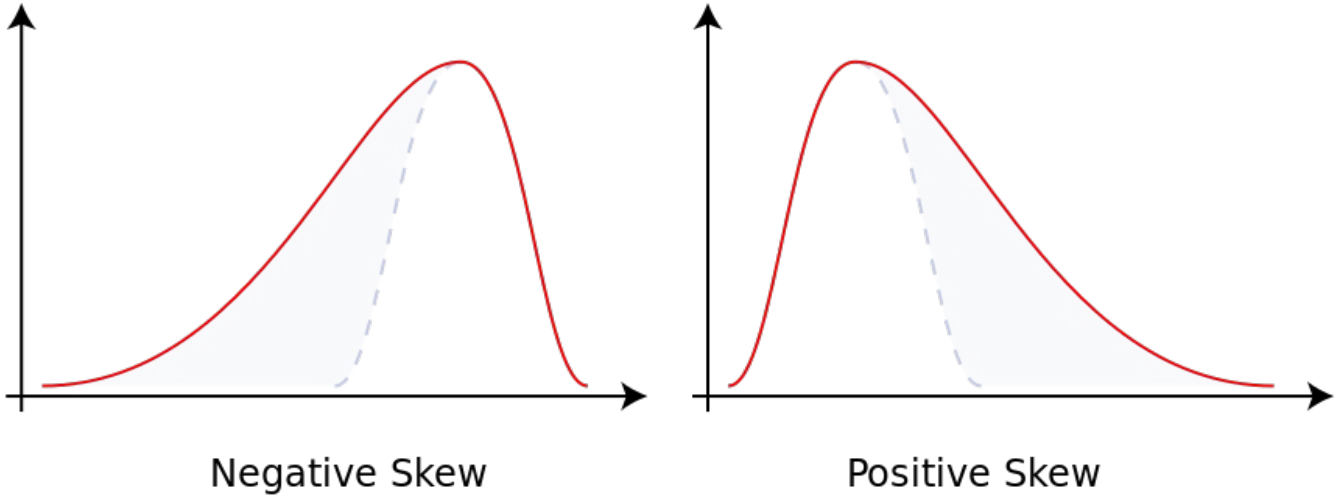
(Excess) Kurtosis: $$K \equiv \int \left(\frac{x-\mu}{\sigma}\right)^4 f(x)dx - 3$$ Dominated by the tail of the PDF

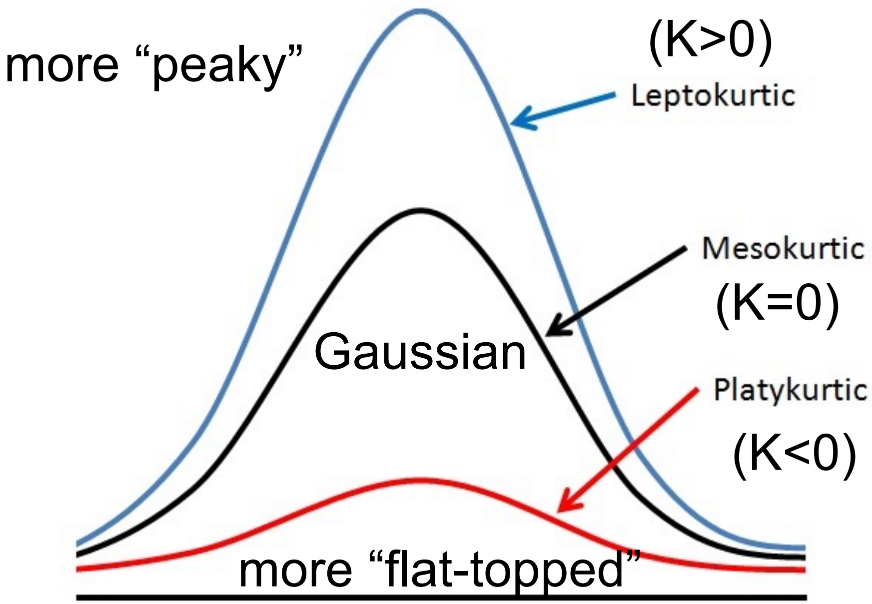
Mode (x_m): Value of x that maximizes f(x) The value that appears most often.

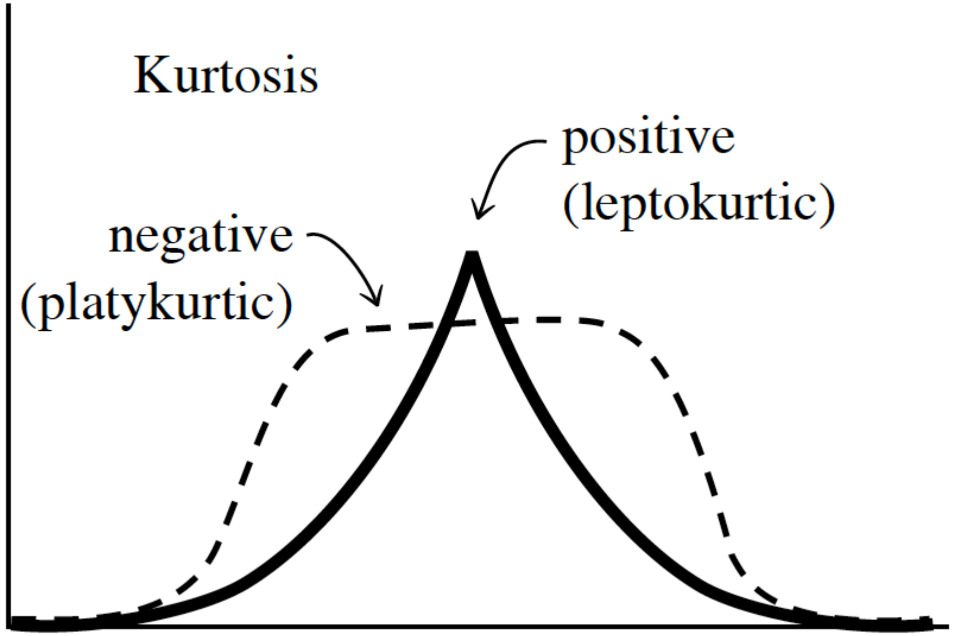
p% quantiles (p is called a percentile), q_p: $$\frac{p}{100} = \int_{-\infty}^{q_p} f(x)dx$$

Values commonly quoted are q_{25}, q_{50} and q_{75}, with q_{50} being the median.

Skewness and Kurtosis







Results for some common distributions

Distribution	Parameters	$E(X)$	q_{50}	σ	Σ	K
Poisson	μ	μ	$\mu - 1/3$	$\sqrt{\mu}$	$1/\sqrt{\mu}$	$1/\mu$
Gaussian	μ, σ	μ	μ	σ	0	0
Exponential	λ	λ^{-1}	$\ln 2/\lambda$	λ^{-2}	2	6
Gamma	k, θ	$k\theta$	no analytic	$k\theta^2$	$2/\sqrt{k}$	$6/k$
Cauchy	μ, γ	N/A	μ	N/A	N/A	N/A
Reduced χ^2	k	1	$(1 - 2/9k)^3$	$\sqrt{2/k}$	$\sqrt{8/k}$	$12/k$
Student's t	ν	0	0	$\nu/(\nu - 2)$	0	$6/(\nu - 4)$

Data-based estimates

Repeated measurements usually yield data that correspond to independent and identically distributed random variables (IID).

Suppose $X_1,\ldots,X_n$ are IIDs. Without knowing their distribution function, we would like to infer some of its basic properties.

Sample mean: $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ unbiased estimate of E(X)

Variance of the mean: $$\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X)}{n}$$

Sample variance: $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$ unbiased estimate of Var(X)

Issue with outliers

Real data may have spurious measurements whose values differ dramatically from others (i.e., outliers).

Median (q_{50}) and interquartile range (q_{75}-q_{25}) are much less affected by the presence of outliers than mean and standard deviation.

Some distributions (e.g., Cauchy) don't have a variance, and interquartile range better quantify the scale parameter.

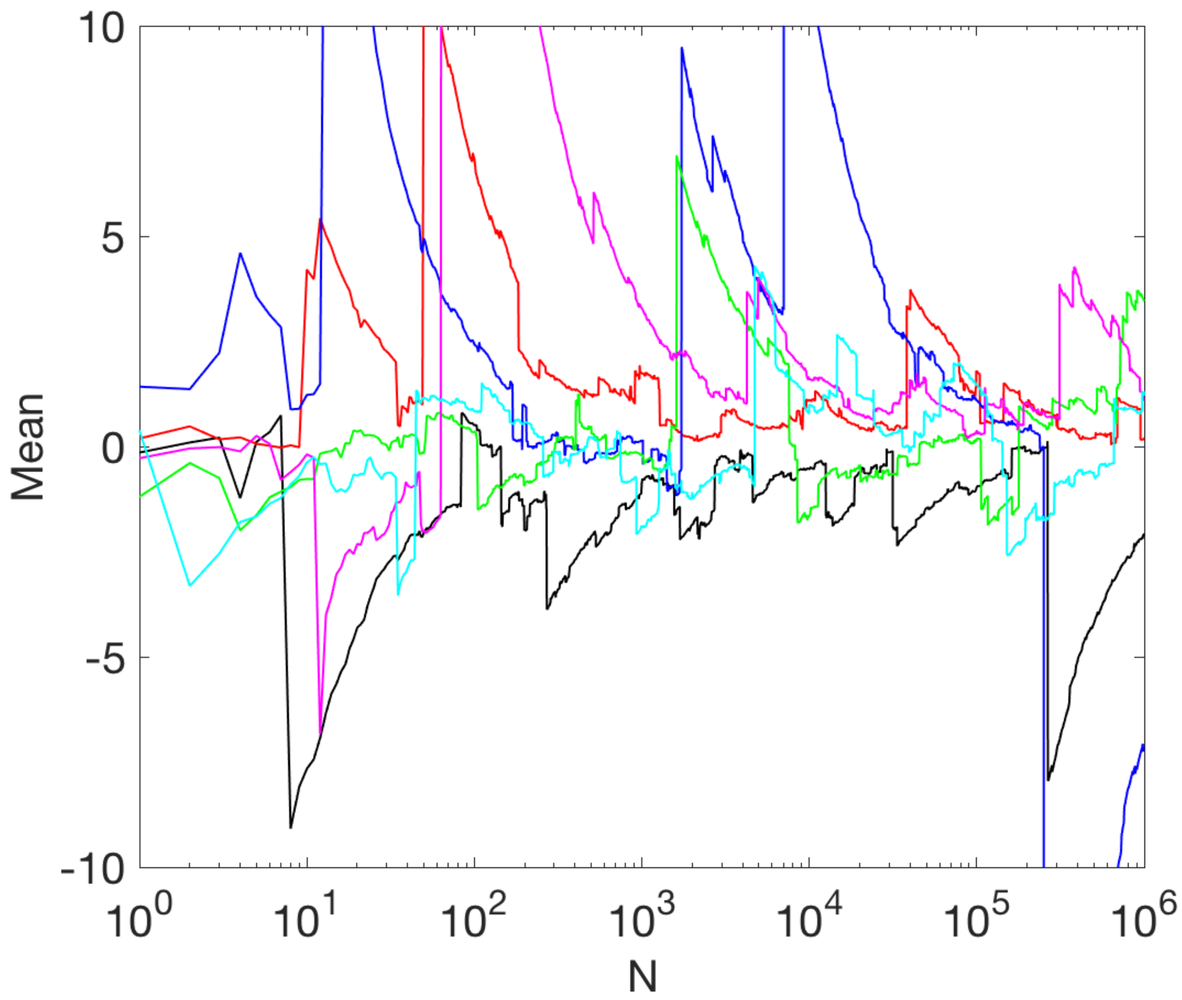
Often, interquartile range is renormalized as

$$\sigma_G \equiv 0.7413(q_{75} - q_{25})$$

which is an unbiased estimator of σ for a Gaussian distribution.

For a Gaussian distribution, the median determined from data shows a scatter around the true mean larger by a factor of $\sqrt{\pi/2} \sim 1.253$ than that determined from $\overline{X}$. This is the price to pay using more robust estimators.

Exceptions: Cauchy distribution



Due to the slowly-decreasing x^{-2} tail, it does not have an expectation, nor variance.

The central limit theorem fails for Cauchy distribution.

Six different realizations of sample mean vs sample size.

Multivariate normal distribution

Let $Z=(Z_1,\ldots,Z_k)^T$ with $Z_1,\ldots,Z_k$ being IIDs, each satisfying the standard normal distribution $\mathcal{N}(0,1)$. The PDF of Z is then given by

$$f(\boldsymbol{z}) = \prod_{i=1}^{k} f(z_i) = \frac{1}{(2\pi)^{k/2}} \exp\left\{ -\frac{1}{2}\sum_{j=1}^{k} z_j^2 \right\} = \frac{1}{(2\pi)^{k/2}} \exp\left\{ -\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z} \right\}$$

We say Z satisfies standard multivariate normal distribution, written as $\mathcal{N}(0, I)$.

More generally, a vector (random variable) $X=(X_1,\ldots,X_k)^T$ has a multivariate normal distribution, denoted by $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, if it has a PDF

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right\}$$

where Σ is a real symmetric and positive-definite matrix.

Multivariate normal distribution

A symmetric positive-definite matrix can be diagonalized and square-rooted:

$$\Sigma = Q^T \Lambda Q = Q^T \Lambda^{1/2} Q Q^T \Lambda^{1/2} Q \equiv (\Sigma^{1/2})^2$$

By defining $Z \equiv \Sigma^{-1/2}(X - \boldsymbol{\mu})$, it is straightforward to show $Z \sim \mathcal{N}(0, I)$.

Given that $X = \Sigma^{1/2} Z + \boldsymbol{\mu}$, it is clear that the marginal distribution for each component of X, namely X_i, is a Gaussian, and

$$E(X) = \Sigma^{1/2} E(Z) + \boldsymbol{\mu} = \boldsymbol{\mu}$$

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= E(X_i X_j) - \mu_i \mu_j \\ &= E\left[(\Sigma^{1/2} Z Z^T \Sigma^{1/2})_{ij} \right] = \Sigma_{ij} \end{aligned}$$

(covariant matrix)

Uncorrelated Gaussian variables are independent.

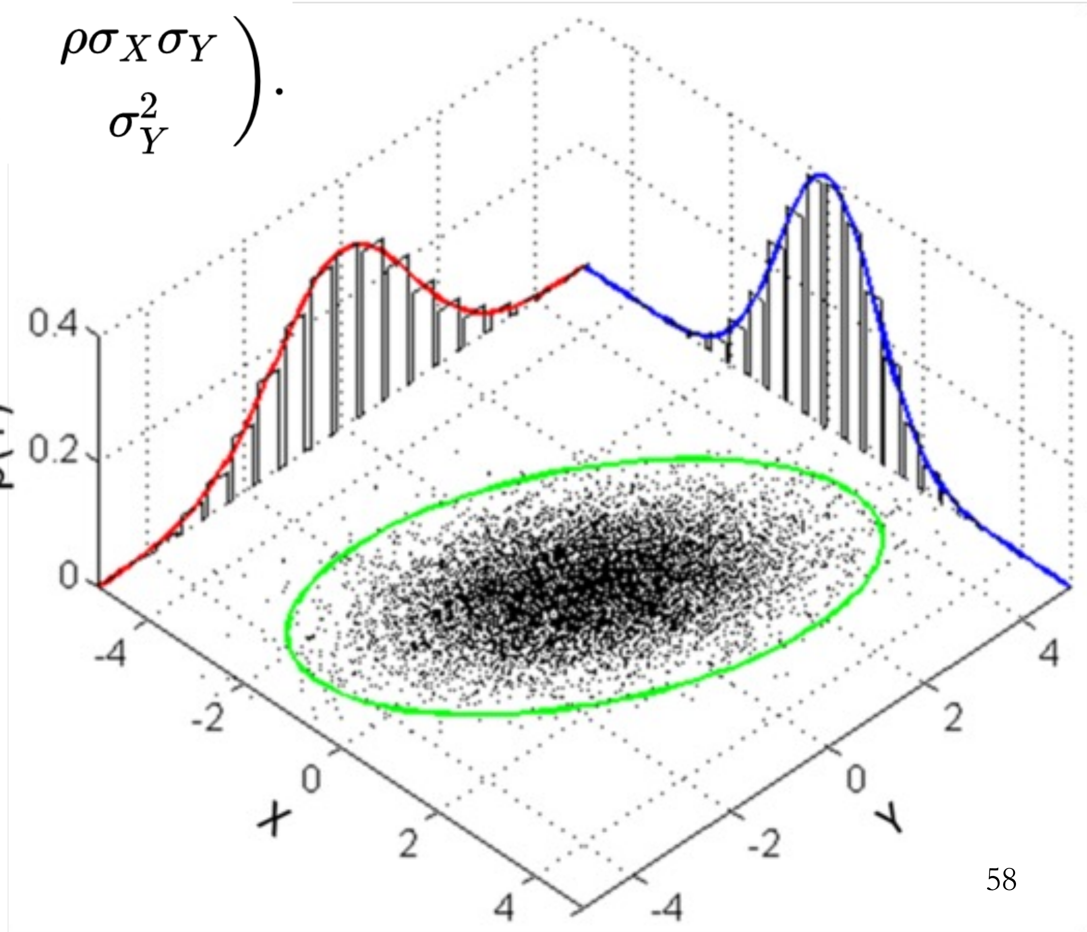
Bivariate normal distribution

In 2D, we have

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right)$$

with $\boldsymbol{\mu} = \begin{pmatrix}\mu_X \\ \mu_Y\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}.$

Usually, contours of constant f is a tilted ellipse.



It is straightforward (but with some algebra) to obtain the orientation of the principle axis, which is related to the diagonalization Σ.

Pearson's sample correlation coefficients

Given N pairs of data (x_i,y_i), the Pearson's sample correlation coefficient is

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

If they are drawn from two uncorrelated Gaussian distributions (i.e., population correlation coefficients ρ=0), then one can show:

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

satisfies Student's t distribution with N-2 degrees of freedom.

If they are drawn from two correlated Gaussian distributions with population correlation coefficient ρ, then the distribution of F:

$$F = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

approximately follows a Gaussian distribution with

$$\mu_F = F(\rho)\ ,\ \sigma_F = (N-3)^{-1/2}$$

Estimate bivariate Gaussian from data

One can estimate four of the five parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y$ the same way as in the univariate case (w. or w/o. outliers). The remaining parameter ρ can be estimated from Pearson's correlation coefficient.

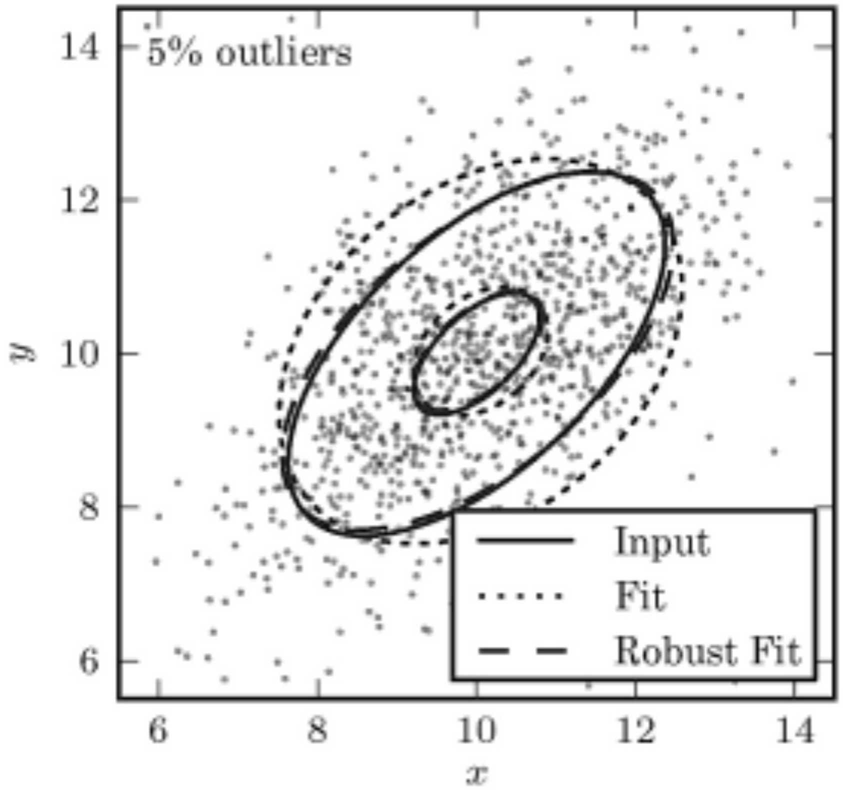
In the presence of outliers, a better way to estimate ρ is as follows:

$$\rho = \frac{V_U - V_W}{V_U + V_W}$$

where $V_{U,W}$ are the variance of U and W, that can be estimated through σ_G, with

$$U = \frac{\sqrt{2}}{2}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right)$$

$$W = \frac{\sqrt{2}}{2}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right)$$



Summary

- Basic probability

  Sample space, (independent) events, (conditional) probability, Bayes' theorem.

- Random variables

  Random variables, CDF, PDF, marginal distribution, random number generator

- Univariate distribution functions

  Binomial, Poisson, normal, exponential, γ, β, Weibull, χ^2, t, Cauchy, *F*…

- Descriptive statistics & data-based estimates

  Expectation, variance, skewness, Kurtosis; estimating E and Var with outliers.

- Law of large numbers and central limit theorem
- Multivariate DFs, correlation and covariance

  Population and sample correlation coefficients, multivariate normal DF, data-based estimates.