Statistics and Numerical Method, Tsinghua University

Probability and Statistical Distributions

Xuening Bai (白雪宁)

Institute for Advanced Study (IASTU) &

Department of Astronomy (DoA)

Nov. 5, 2024

Why probability?

Random phenomena do occur in nature.

Turbulence, chaos, quantum systems, etc.

The way we observe/measure events often bare uncertainties.

Our measurements unavoidably have errors.

Lack of understanding of the phenomenon.

Probability gives us a mathematical way to model uncertainty.

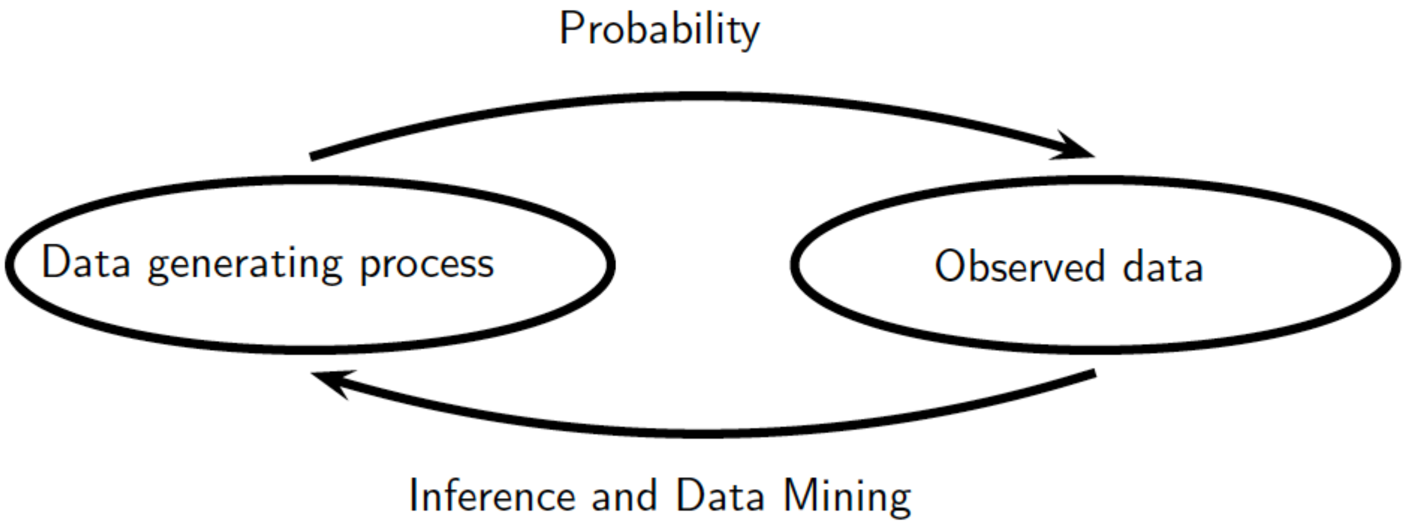
Randomness does not mean unpredictability.

Predictable in probabilistic sense.

Statistics, data mining and machine learning

All concerned with collecting and analyzing data.

Given the data-generating process, what are the properties of the outcomes?



Given the outcomes, what can we say about the process that generated the data?

Outline

- Basic probability
- Random variables
- Univariate distribution functions
- Descriptive statistics & data-based estimates
- Central limit theorem
- Multivariate distribution functions, correlation and covariance

Sample space and events

The sample space Ω is the set of possible outcomes of an experiment.

Points ω in Ω are called sample outcomes/realizations/elements.

Subsets of Ω, usually denoted by *A*, are called events.

Coin toss:

Ω={**H**ead, **T**ail}

Throw a die:

Ω={1,2,3,4,5,6}

Air pollution in Beijing

Ω=[0,∞)

e.g., $A=[\omega\in\Omega \mid \omega<100]$

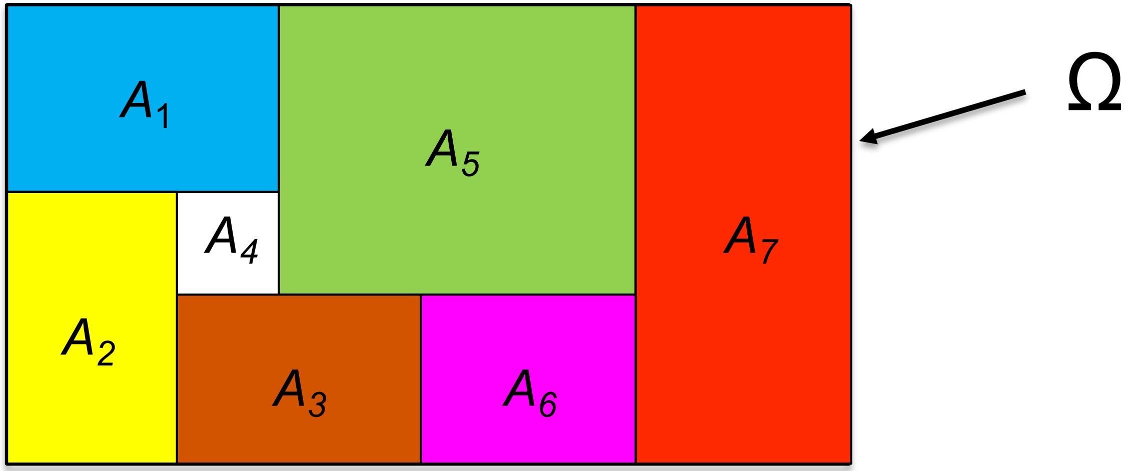
Disjoint events and sample partition

We say events A_1, A_2, … are disjoint or are mutually exclusive if

$$A_i \cap A_j = 0 \quad \text{whenever} \quad i \neq j$$

A partition of Ω is a sequence of disjoint sets A_1, A_2, … such that

$$\cup_i A_i = \Omega$$



Axioms of probability

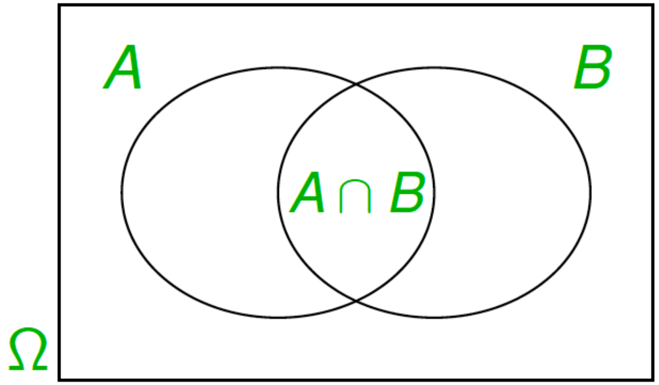
Definition:

A function P that assigns a real number $P(A)$ to each event A is called a probability distribution, or a probability measure, if it satisfies the following:

1. $P(A) \geq 0$ for any event A.
2. $P(\Omega) = 1$.
3. If A_1, A_2, … are disjoint, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

For any events A and B, we have

$$P(A \cup B) = P(A) + P(B) - P(AB)$$



Independent events

Two events A, B are independent if and only if

$$P(A \cap B) = P(A)P(B)$$

Note: this is *P*(*AB*)

Independence is usually assumed based on common sense/intuition, e.g., coin toss, throwing a die.

Disjoint events with positive probability are not independent (because *P*(*AB*)=0).

Conditional probability

Assuming *P*(*B*)>0. The conditional probability of *A* given that *B* has occurred is defined as

$$P(A|B) = \frac{P(AB)}{P(B)}$$

If *A* and *B* are independent events, then $P(A|B) = P(A)$, and vice versa.

More generally, for any pair of events:

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

Note, in general $P(A|B) \neq P(B|A)$

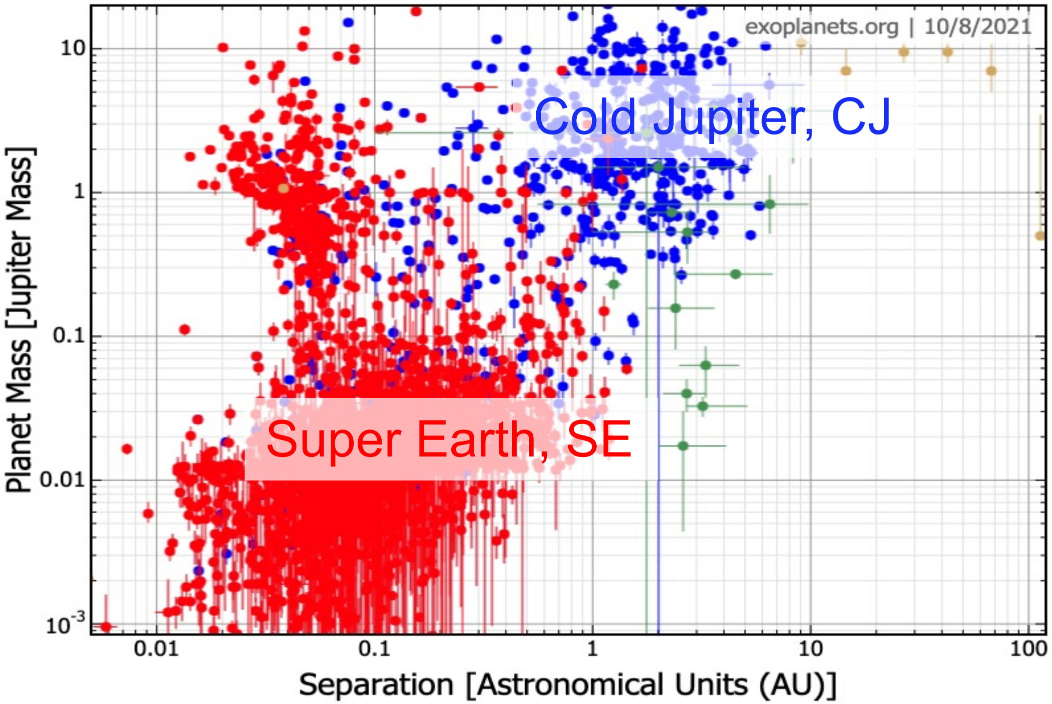
Example: correlation in exoplanetary systems

Around a Sun-like star:

- Probability of having a CJ is ~10% (Cumming+2010).

- Probability of having a SE is ~30% (Zhu+2018).

- Conditional probability of having a CJ when there is a SE is ~30% (Zhu & Wu 2018).



What is the conditional probability of having a SE when there is a CJ?

$$P(\mathrm{SE}) \times P(\mathrm{CJ}|\mathrm{SE}) = P(\mathrm{CJ}) \times P(\mathrm{SE}|\mathrm{CJ}).$$

(Zhu & Wu 2018)

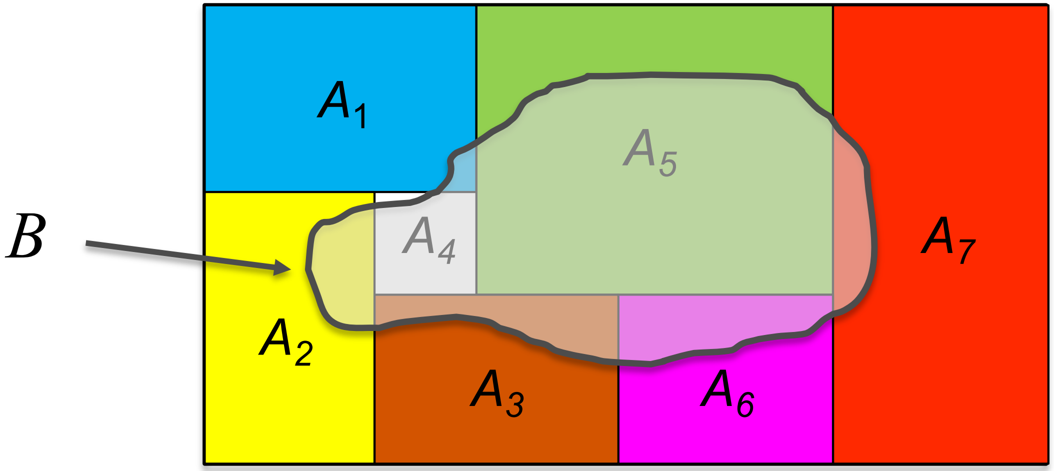
30% 30% 10% 90%!

Law of total probability

Let $A_1, \ldots, A_k$ be a partition of Ω . Then for any event B, we have

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

This is called the law of total probability.



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

Example: medical test

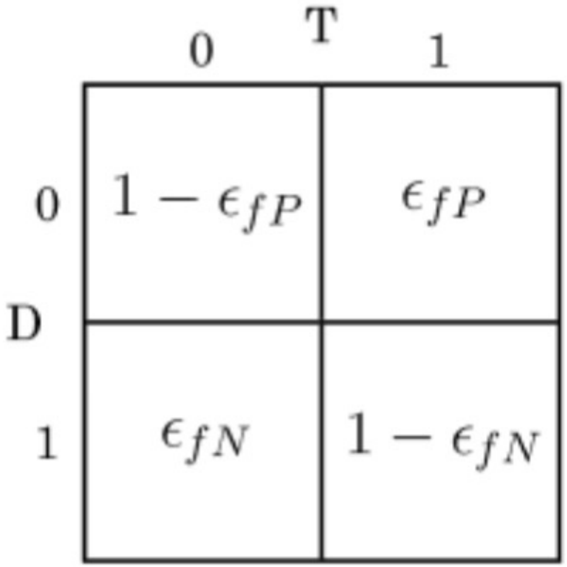
You take a medical test T for some disease D.
There are 4 possibilities:

The test is positive and you are sick (T=1, D=1).

The test is negative and you are sick (T=0, D=1).

The test is positive and you are healthy (T=1, D=0).

The test is negative and you are healthy (T=0, D=0).



		T	
		0	1
D	0	$1-\epsilon_{fP}$	ϵ_{fP}
	1	ϵ_{fN}	$1-\epsilon_{fN}$

Suppose the chance of “false positive” is ε_{fp}, which can be written as

$$P(T=1|D=0)=\frac{P(T=1,D=0)}{P(D=0)}=\epsilon_{\text{fp}}$$

Similarly, the chance of “false negative” is ε_{fn}, which can be written as

$$P(T=0|D=1)=\frac{P(T=0,D=1)}{P(D=1)}=\epsilon_{\text{fn}}$$

Example: medical test

If I take a test and get positive result, what is the chance that I have contracted this disease?

$$P(\underset{\substack{\uparrow\\ \text{hypothesis}}}{D}=1|\underset{\substack{\uparrow\\ \text{data}}}{T}=1)=\frac{P(T=1,D=1)}{P(T=1)}$$

$$=\frac{P(T=1|D=1)P(D=1)}{P(T=1|D=0)P(D=0)+P(T=1|D=1)P(D=1)}$$

Prior probability: based on large population studies, $P(D=1)=\varepsilon_D$.

This leads to

$$P(D=1|T=1)=\frac{(1-\epsilon_{\text{fn}})\epsilon_D}{(1-\epsilon_{\text{fn}})\epsilon_D+\epsilon_{\text{fp}}(1-\epsilon_D)}\approx\frac{\epsilon_D}{\epsilon_D+\epsilon_{\text{fp}}}$$

Therefore, what makes a good medical test requires $\varepsilon_{\text{fp}} \ll \varepsilon_D$.

Outline

- Basic probability
- Random variables
- Univariate distribution functions
- Descriptive statistics & data-based estimates
- Central limit theorem
- Multivariate distribution functions, correlation and covariance

Random variables

Definition: a random variable is a mapping:

$$X : \Omega \to \mathbb{R}$$

which assigns a real number X(ω) to each outcome ω.

Some examples:

- Flip the coin 10 times, number of heads in the sequence is a random variable.
- Let $\Omega = \left\{ (x, y);\ x^2 + y^2 \leq 1 \right\}$. Consider drawing a random point from Ω: ω=(x,y). Random variables can be defined as

  $X(\omega) = x,\ Y(\omega) = y,\ Z(\omega) = x + y,\ W(\omega) = \sqrt{x^2 + y^2}.$ etc.

Cumulative distribution function

The cumulative distribution function (CDF), is a function $F_X : \mathbb{R} \to [0,1]$ defined by

$$F_X(x) = P(X \le x)$$

A random variable is uniquely determined by its CDF.

A CDF is non-decreasing, and $F(-\infty) = 0 \ , \ F(\infty) = 1 \ .$

Depending on whether the CDF takes countably many values as step functions, or continuous/differentiable, one can say a random variable is discrete, or continuous.

Probability distribution function

For discrete random variable, one can define a probability (mass) function f_X:

$$f_X(x) = P(X = x)$$

For a continuous random variable, one can define a probability density function (PDF) f_X, so that for any $a<b$,

$$P(a < X < b) = \int_a^b f_X(x)dx = F(b) - F(a)$$

In other words, $f_X(x) = F'_X(x)$.

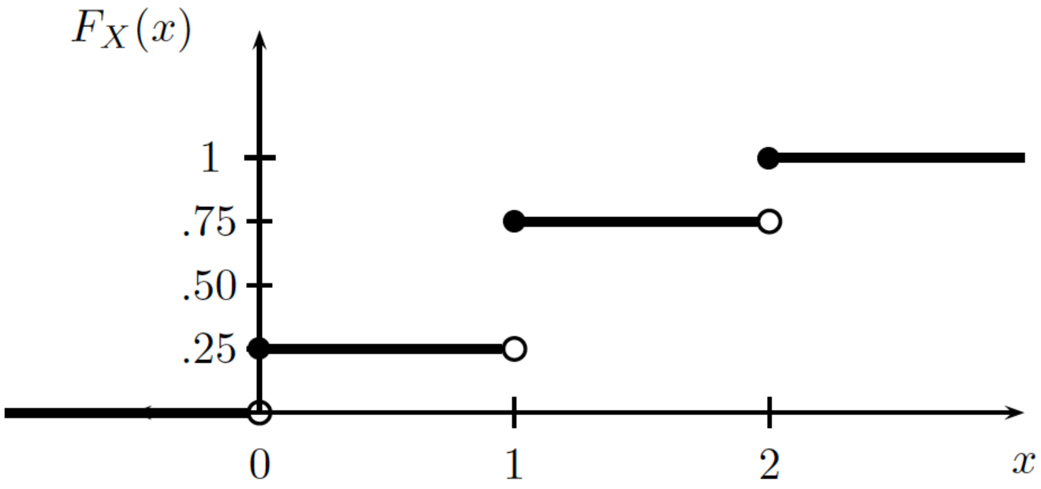
CDF: examples

For a discrete random variable:

$$P(X = 0) = P(X = 2) = 1/4;$$

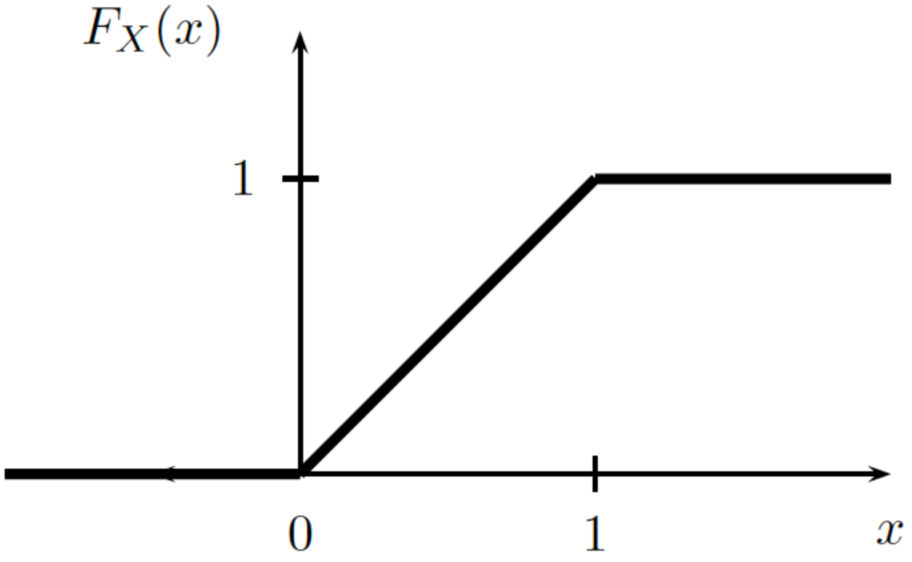
$$P(X = 1) = 1/2$$

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \le x < 1 \\ 3/4 & 1 \le x < 2 \\ 1 & x \ge 2. \end{cases}$$



Uniform distribution in [0 1]:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & x > 1. \end{cases}$$



Transformation of random variables

Given a random variable X, any function of X is also a random variable.

Let the PDF of a random variable *X* be $f(x)$, whereas we are interested in some variable *Y*=**Φ**(*X*). What’s the distribution function of *Y*, denoted by $g(y)$?

Let us assume f is differentiable, we should have

$$g(y)dy = f(x)dx$$

$$\Rightarrow \quad g(y) = f[\Phi^{-1}(y)]\frac{dx}{dy} = f[\Phi^{-1}(y)]\frac{d\Phi^{-1}(y)}{dy}$$

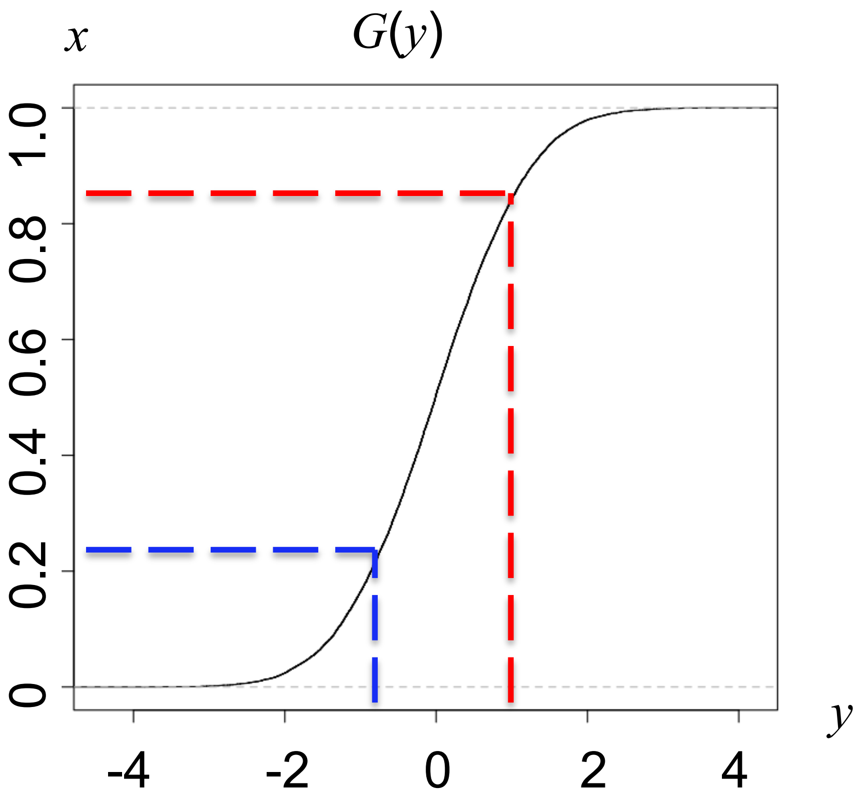
Exercise: random number generator

You are provided with a random number generator, which gives a variable X satisfying uniform distribution in [0,1] (so that $f(x)=1$ for $x \in [0,1]$).

How to use this random number generator to produce a random variable Y, satisfying an arbitrary (continuous) distribution g?

Hint: this can be achieved using the CDF of g, denoted by G.

Solution: it should be easy to verify we can take $Y=G^{-1}(X)$.



Bivariate distribution function

For a pair of random variables X, Y, one can define

Discrete case: the joint mass function

$$f(x, y) = P(X = x, Y = y)$$

Continuous case: (joint) PDF $f\,(x,y)$ should satisfy the following

$$f(x,y) \geq 0\ ,\quad \int dx \int dy f(x,y) = 1\ ,$$ and

For any set $A \in \mathbb{R} \times \mathbb{R}$, $P[(X,Y) \in A] = \int \int_A f(x,y) dx dy$

These are readily generalizable to multi-dimensions.

Marginal distributions

For continuous random variables X, Y with joint PDF $f(x, y)$, their marginal density functions are defined as

$$f_X(x) = \int f(x, y)dy \ , \ f_Y(y) = \int f(x, y)dx$$

Similarly, for discrete variables, there is the marginal mass function.

There is also the conditional probability density function:

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

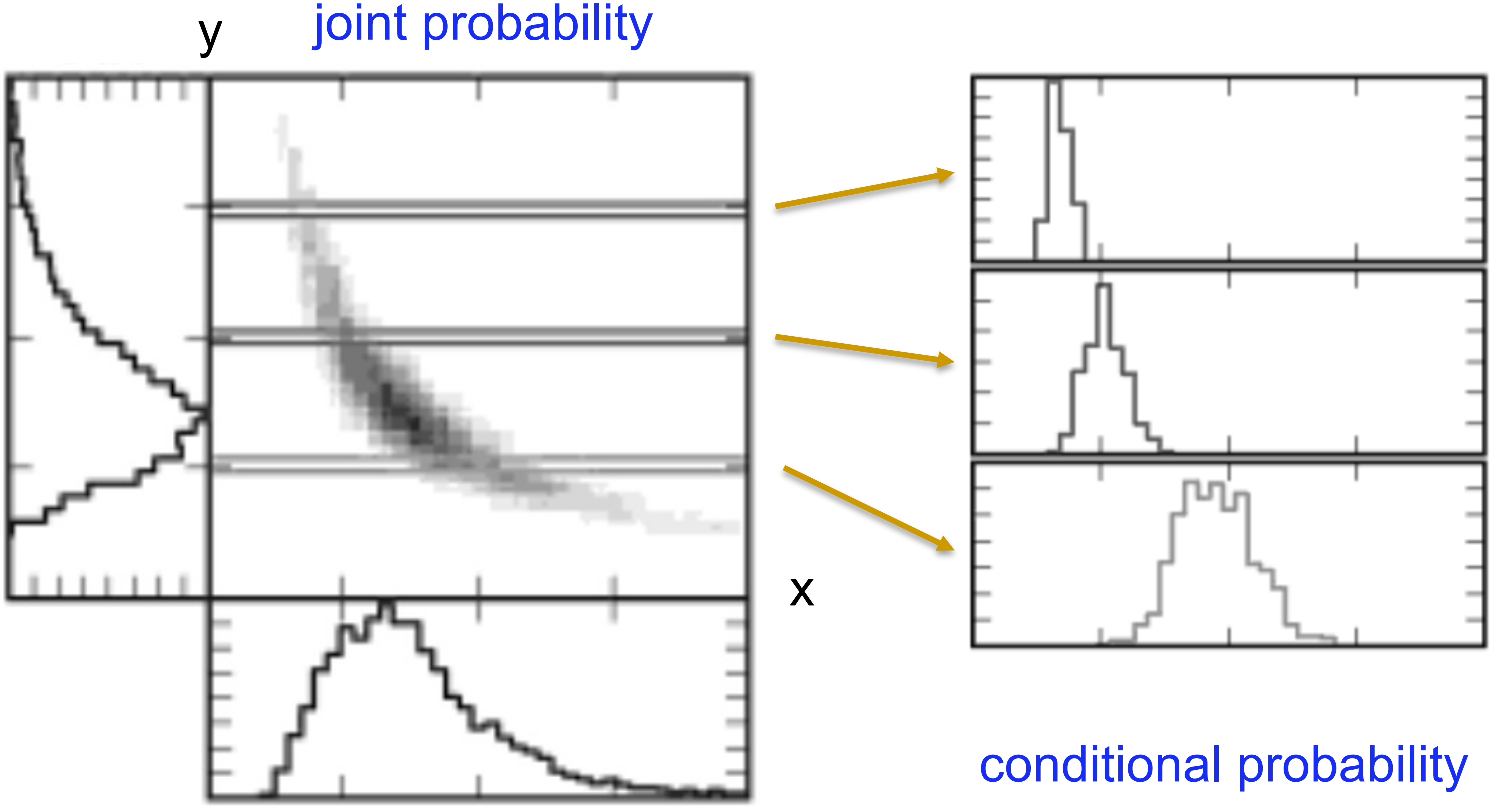
Two random variables X, Y are independent if and only if their respective marginal distribution functions satisfy

$$f(x, y) = f_X(x)f_Y(y)$$

Example: 2D probability distribution

$$f(x,y)$$

joint probability



$$f_Y(x) = \int f(x,y)dx$$

marginalized probability

marginalized probability

$$f_X(x) = \int f(x,y)dy$$

conditional probability

$$f_{X|Y}(x|y_k)$$

Outline

- Basic probability
- Random variables
- Univariate distribution functions
- Descriptive statistics & data-based estimates
- Central limit theorem
- Multivariate distribution functions, correlation and covariance

Important discrete distribution functions

Many of these are related to the outcome of a coin toss that falls heads up with probability p (0<p<1).

The Bernoulli distribution: $f(x) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$

Flip the coin n times and let *X* be the number of heads. Then its distribution function is a binomial distribution:

$$f(x) = \binom{n}{k} p^x (1-p)^{n-x} \quad (x = 0, 1, \ldots, n)$$

(Note n, p are parameters)

Let X be the number of flips needed until the first head, then it satisfies a geometric distribution:

$$f(x) = p(1-p)^{x-1} \quad (x = 1, 2, \ldots)$$

Important discrete distribution functions

Suppose some event on average occurs λ times over some time interval.

Let X be the number of times it actually occurs during such a time interval.

Number of radioactive decay events.

Number of photons a detector receives.

Number of patients in a hospital.

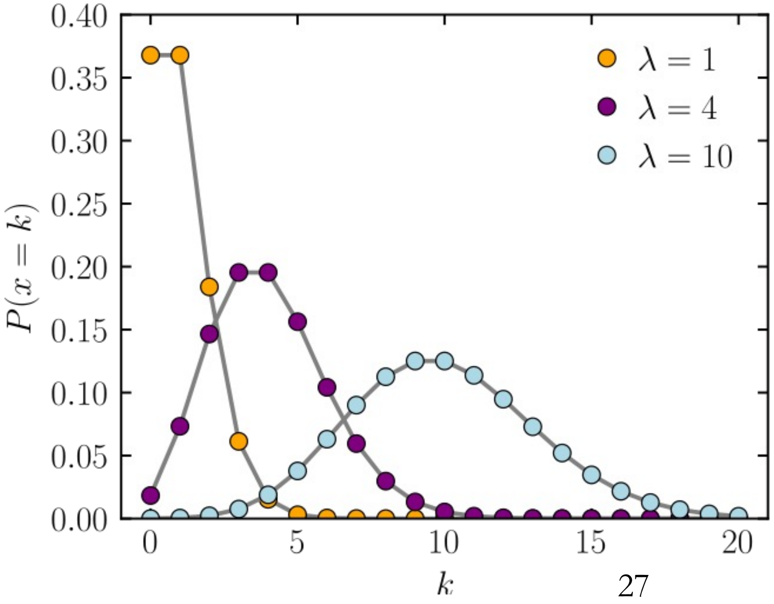
Divide this time interval into n sub-intervals. For sufficiently large n, the probability it occurs in one sub-interval is λ/n<<1. Then, the probability it occurs k times in the whole interval satisfies a binomial distribution:

$$P(X=k)=\frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}$$

Taking the limit of n→∞, it becomes

$$f(k)=\lim_{n\to\infty}P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$$

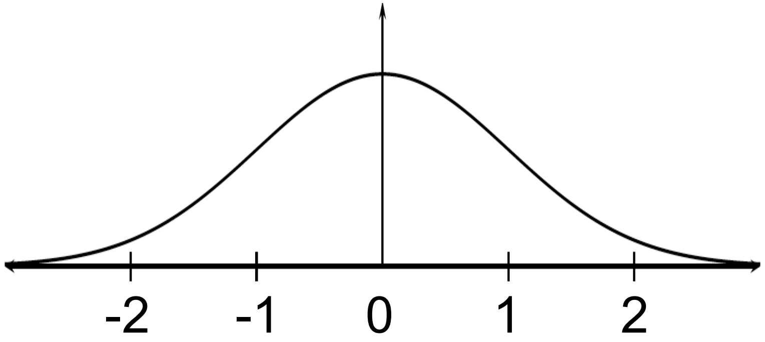
This is known as the Poisson distribution.



Normal/Gaussian distribution

A random variable X has a normal (or Gaussian) distribution with parameters μ and σ, denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$, if

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



It is called "standard normal distribution" if μ=0 and σ=1.

Its importance and wide applications thanks to the "central limit theorem".

If $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, i=1,…,n are independent, then

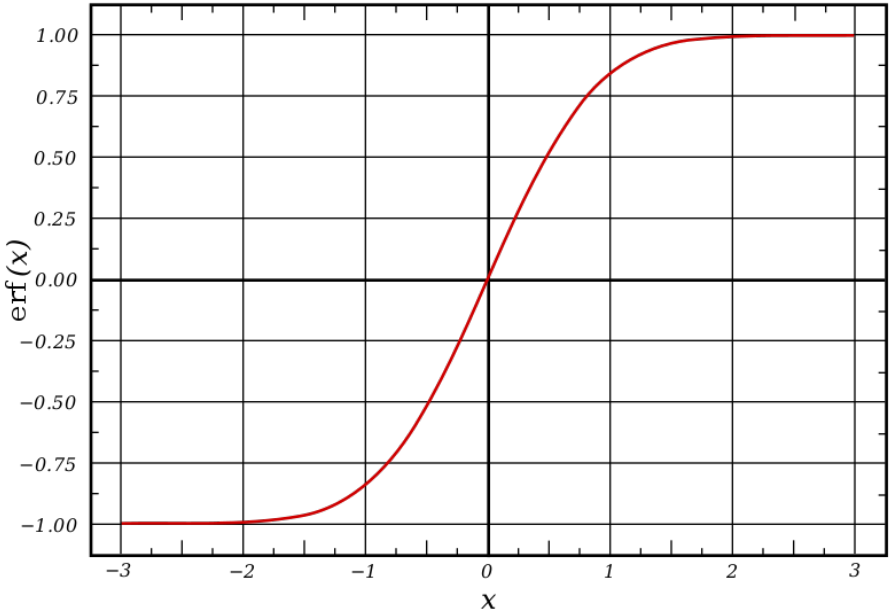
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2 \right)$$

The Error function

The CDF of a the standard normal distribution :

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

Error function



This is related to the error function

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

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

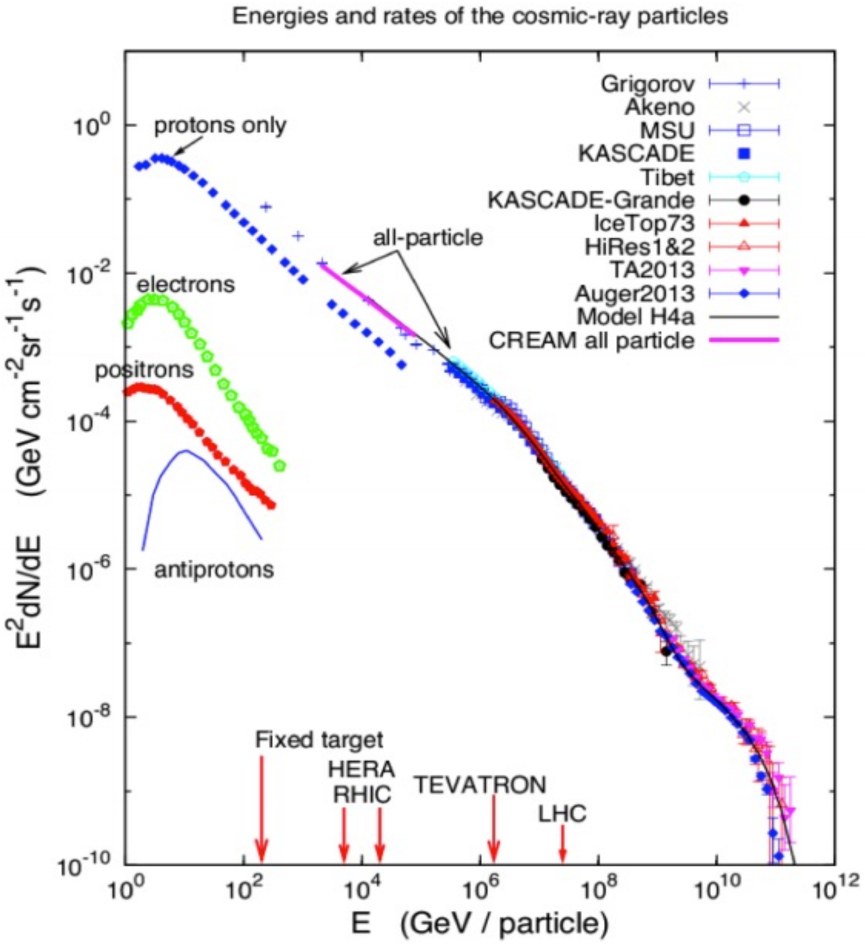
Essentially, a power law:

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

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

The PDF is

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



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 exponential distribution

A random variable *X* has an exponential distribution with parameter λ, denoted by *X*~Exp(λ), if

$$f(x) = \lambda \exp(-\lambda x) \ , \quad (x > 0)$$

Physically, it corresponds to the time interval between individual events in a Poisson process. (being memoryless)

Recall that the Poisson distribution for an event to occur k times over time interval t reads:

$$f(k;t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

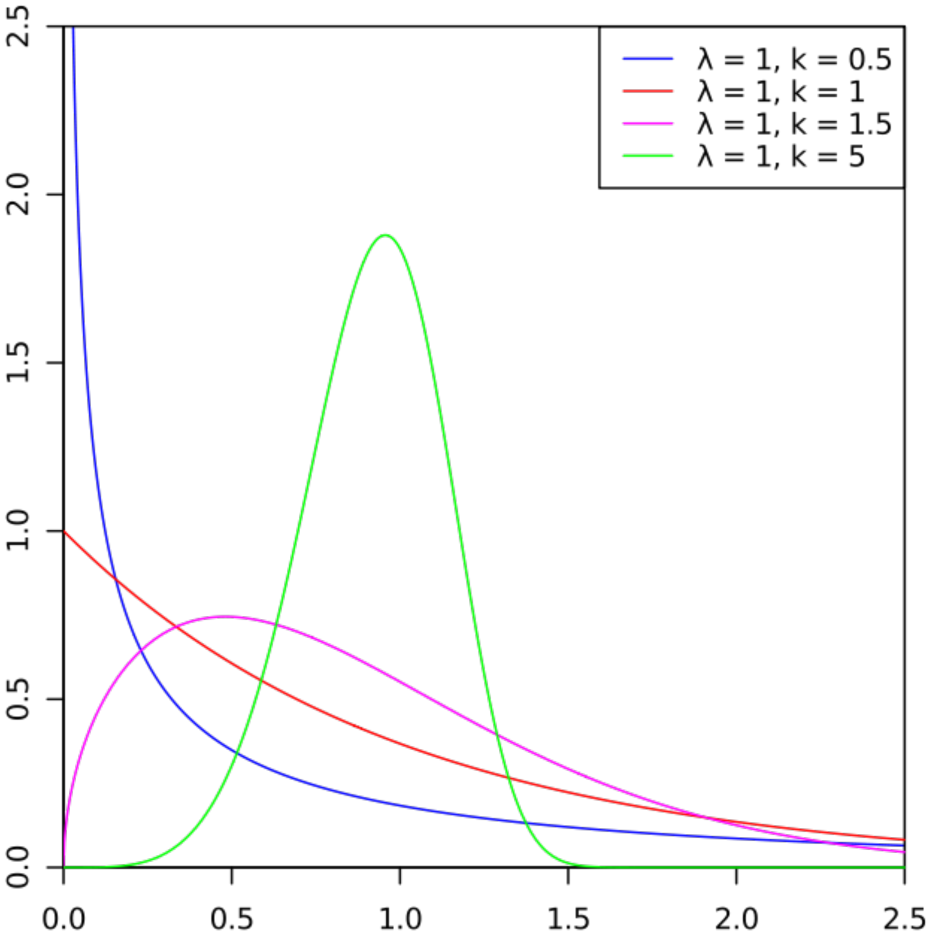
The probability it occurs at least once is: $1 - f(0,t) = 1 - e^{-\lambda t}$

This should be the CDF of our desired distribution function, which is the exponential distribution.

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:

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

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

The gamma distribution

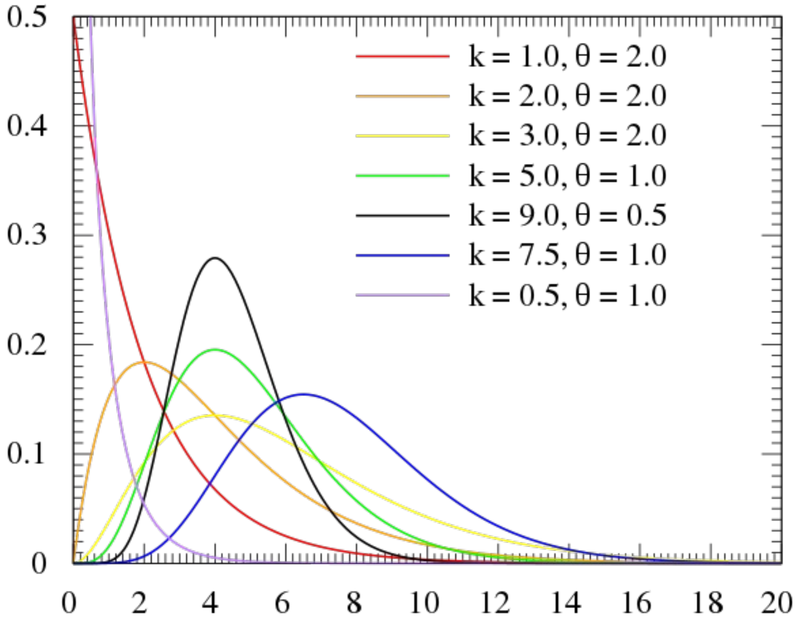
Recall that the Gamma function is defined as $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$

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].

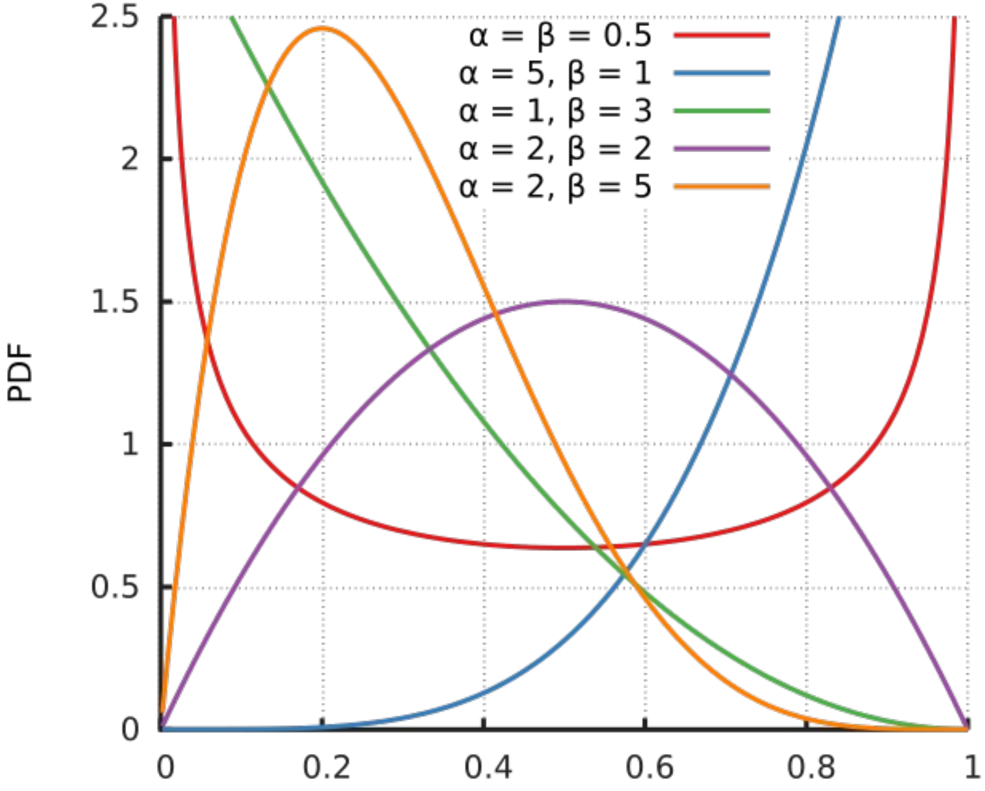
$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

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.

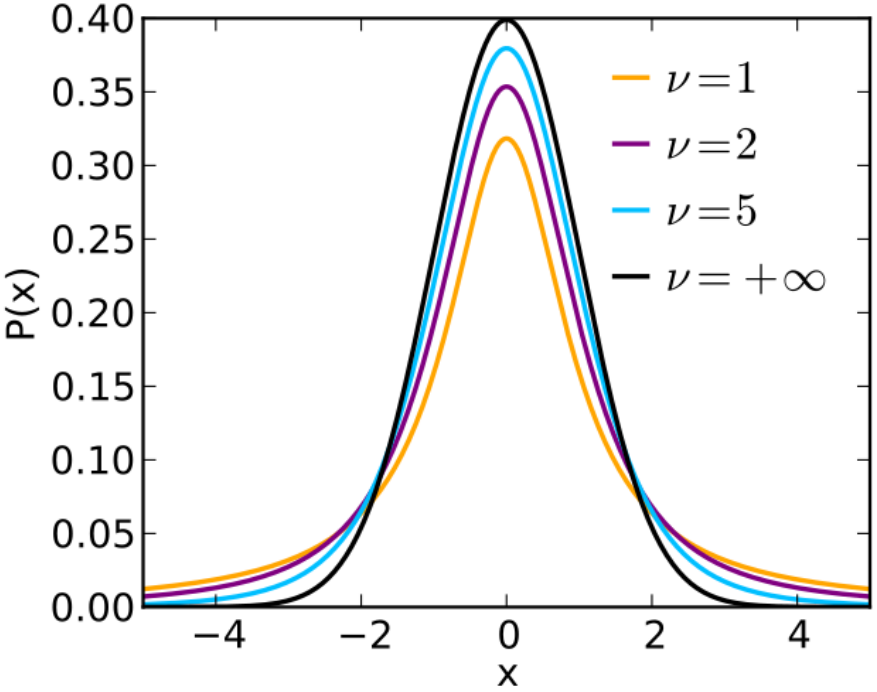


Student's t distribution

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

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

(unbiased estimates of mean and variance)

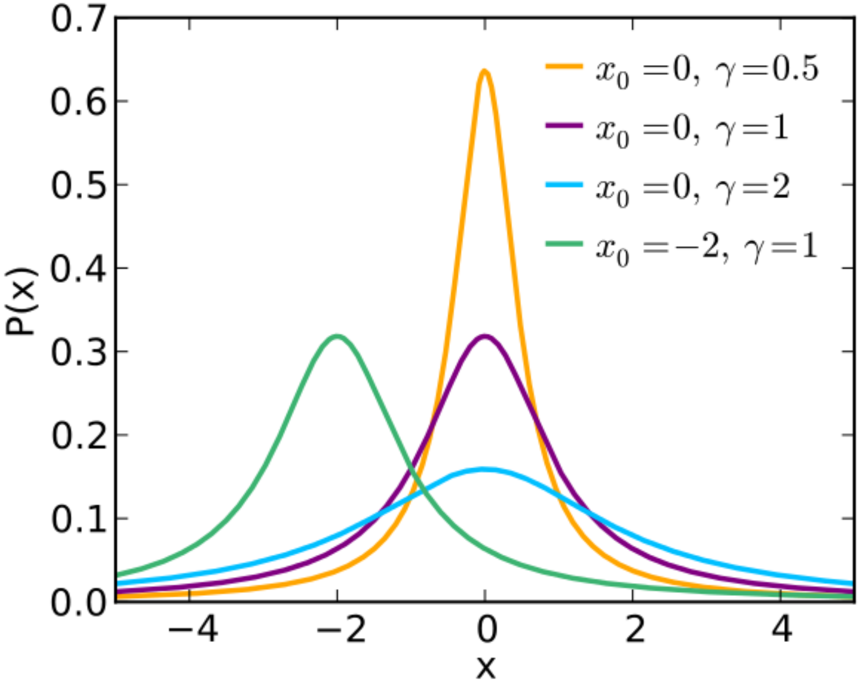
Then the variable $T \equiv \frac{\sqrt{n}}{S_n}(\bar{X}_n-\mu)$ satisfies student's t distribution with n-1 degrees of freedom.

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.

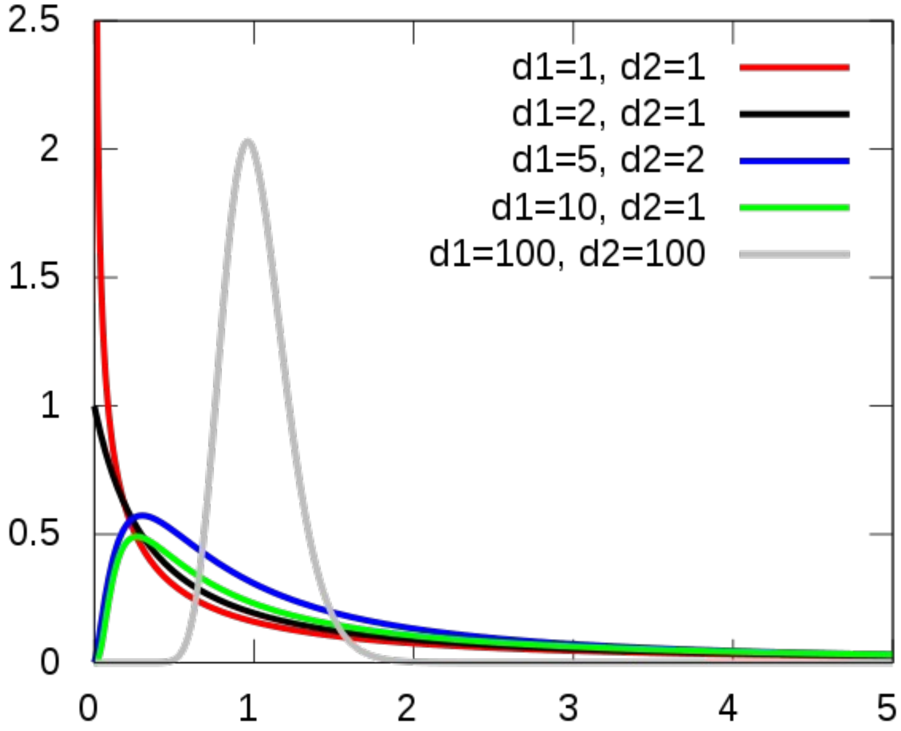
Fisher’s F distribution

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.

Outline

- Basic probability
- Random variables
- Univariate distribution functions
- Descriptive statistics & data-based estimates
- Central limit theorem
- Multivariate distribution functions, correlation and covariance

Expectation value

For notational convenience, now we assume random variables are continuous.

For a random variable X, whose PDF is f(x), its expectation value is

$$E(X) = \int x f(x) dx \equiv \mu$$

To ensure it is well-defined, require $\int |x| f(x) dx$ to be bounded.

For $Y=r(X)$ being another random variable, then

$$E(Y) = E[r(X)] = \int r(x) f(x) dx$$

One can define the kth moment of X as: $E(X^k)$

If $X_1,\ldots,X_k$ are independent, then $E(\prod_{i=1}^{k} X_i) = \prod_{i=1}^{k} E(X_i)$

Variance

The variance of a random variable X is defined as

$$\mathrm{Var}(X) = E\left([X - E(X)]^2\right) = \int (x-\mu)^2 f(x)dx$$

The standard deviation is defined as $\sigma(X) = \sqrt{\mathrm{Var}(X)}$

It is straightforward to show the following properties hold:

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2$$

If $X_1,\ldots,X_k$ are independent, then

$$\mathrm{Var}(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i)$$

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

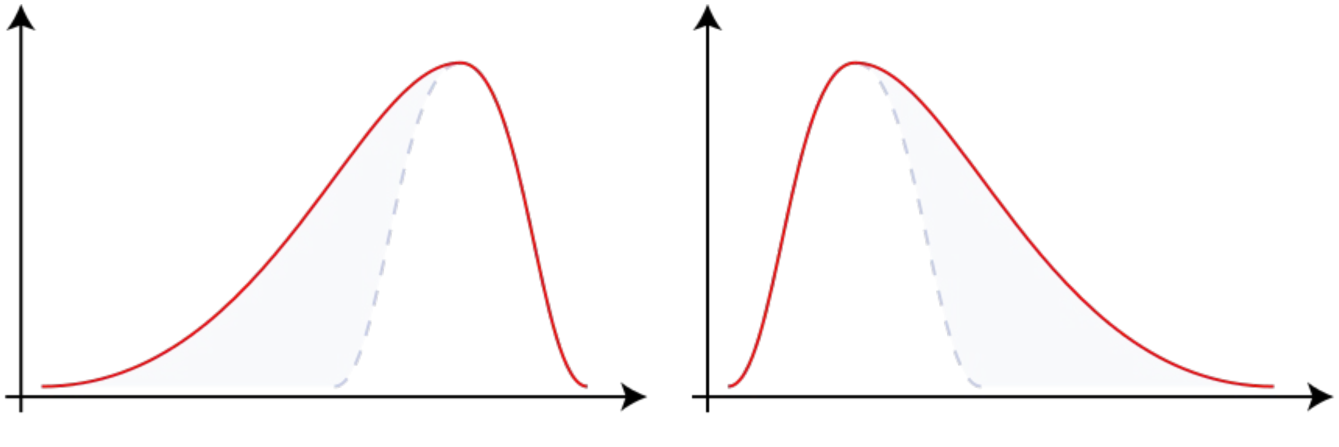
Mode (x_m): Value of x that maximizes f(x)

The value that appears most often.

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

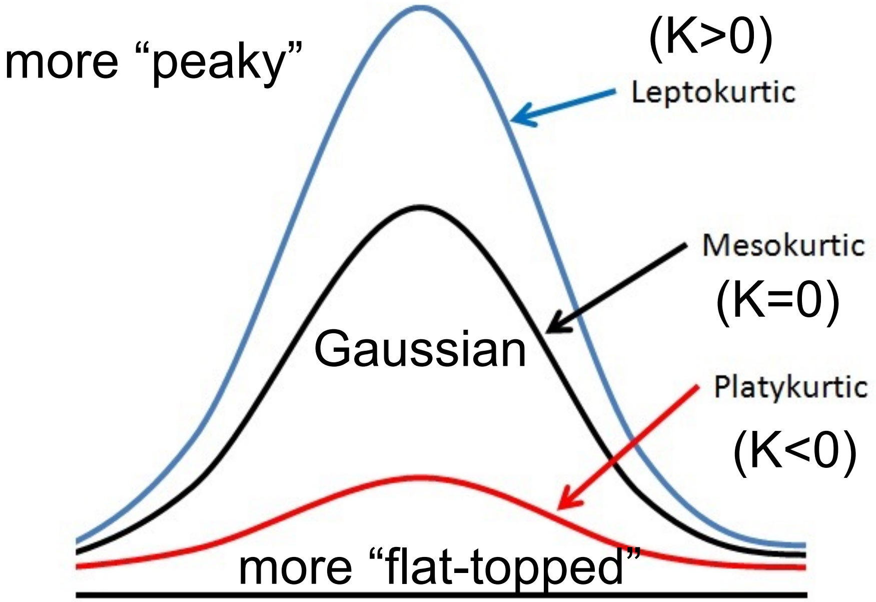
Values commonly quoted are q_{25}, q_{50} and q_{75}, with q_{50} being the median.

Skewness and Kurtosis



Negative Skew

Positive Skew



more “peaky”

(K>0)

Leptokurtic

Mesokurtic

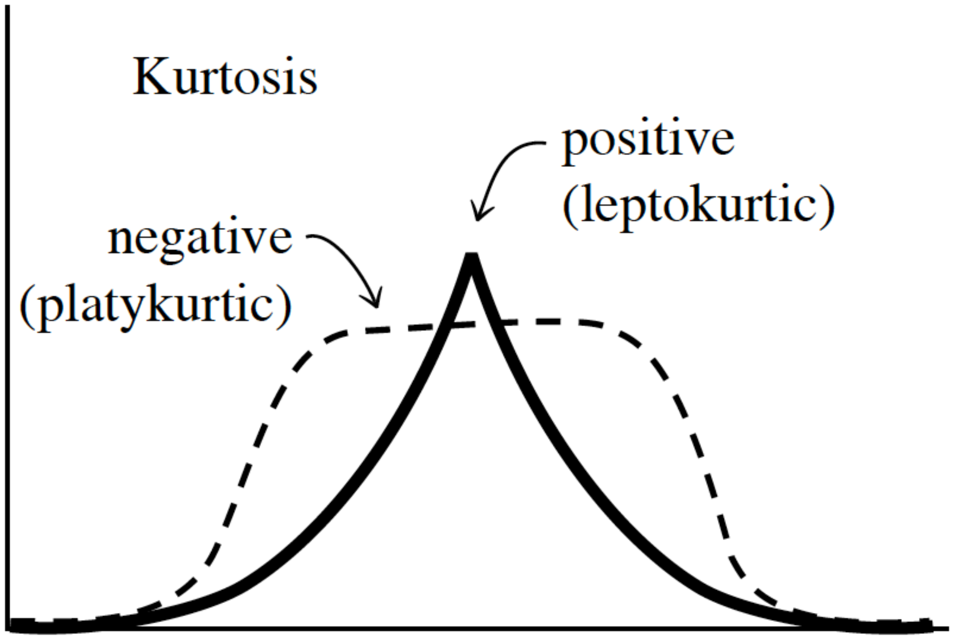
(K=0)

Gaussian

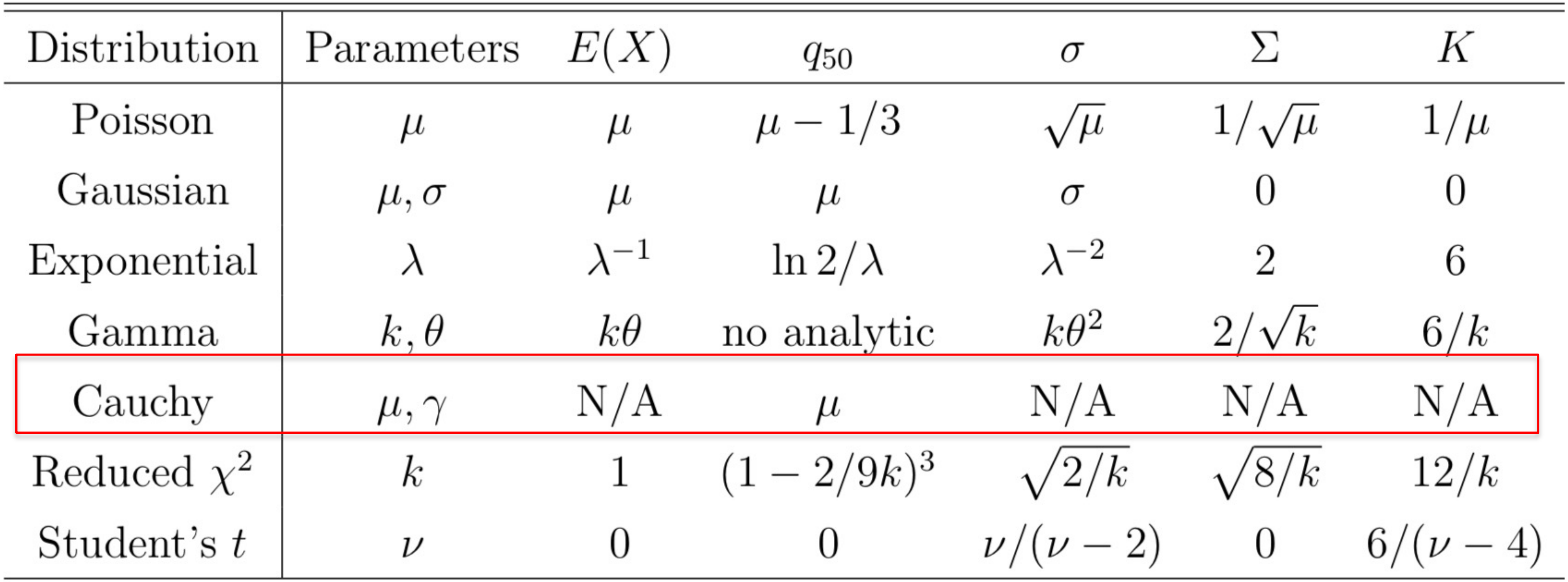
Platykurtic

(K<0)

more “flat-topped”



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.

Outline

- Basic probability
- Random variables
- Univariate distribution functions
- Descriptive statistics & data-based estimates
- **Central limit theorem**
- Multivariate distribution functions, correlation and covariance

Behavior of sums as $n \to \infty$

This arises when

- We repeat an experiment for many times, and the result of one experiment does not influence any others.

  Let X_i be the result of each experiment, subject to measurement errors of some unknown distribution.

  What can we say about the mean of these measurements?

- The outcome is the (additive) accumulation of many small independent actions of the same nature.

  Let X_i be the result of a gambling game (win/lose some money).

  What should one expect for an gambler doing it for a large number of times?

The (strong) law of large numbers

Let $X_1,\ldots,X_n$ be IIDs. If $E|X_1|<\infty$, then:

$$\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$$ converges almost surely to $\mu=E(X_1)$.

which means $$P\left\{\lim_{n\to\infty}\overline{X}_n=\mu\right\}=1\ .$$

There has been several weaker versions of this law, which mostly differ in the conditions and in the form of convergence.

The strong law was first proved by A.N. Komogorov (1930).

The central limit theorem

Let $X_1,\ldots,X_n$ be IIDs, characterized by μ=E(X_1) and Var(X_1)=σ^2, then

$$\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{satisfies:} \qquad \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

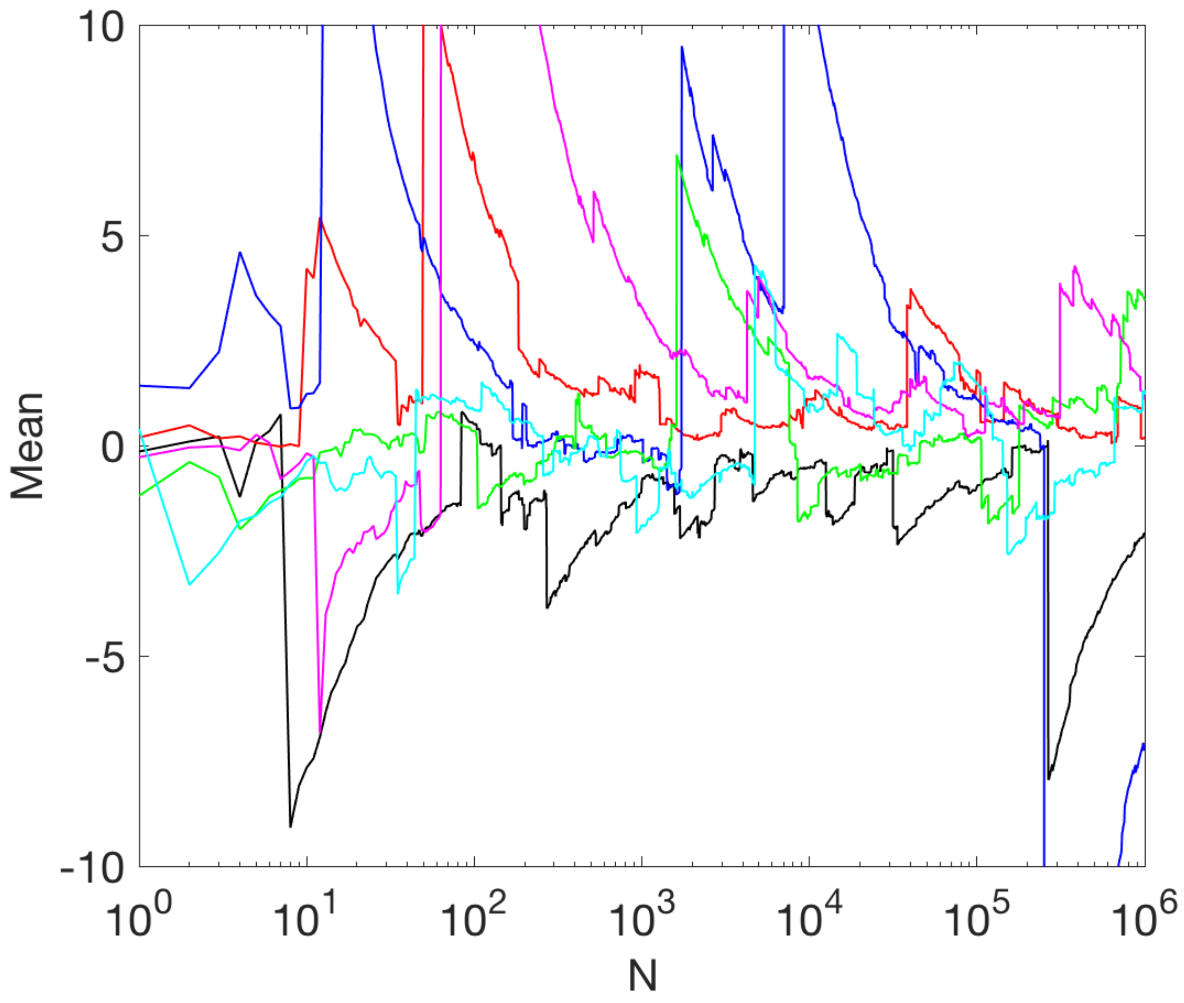
Remarkable result that does not depend on the specific distribution.

Verified in the 1700-1800s, and proved in early 1900s.

The main reason of its success is due to the strong assumption:

Var(X) must exist: the tail of the distribution must decrease faster than x^{-2}.

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.

Example: log-normal distribution

A random variable X satisfies the log-normal distribution with parameters μ and σ if $\ln(X) \sim \mathcal{N}(\mu, \sigma^2)$.

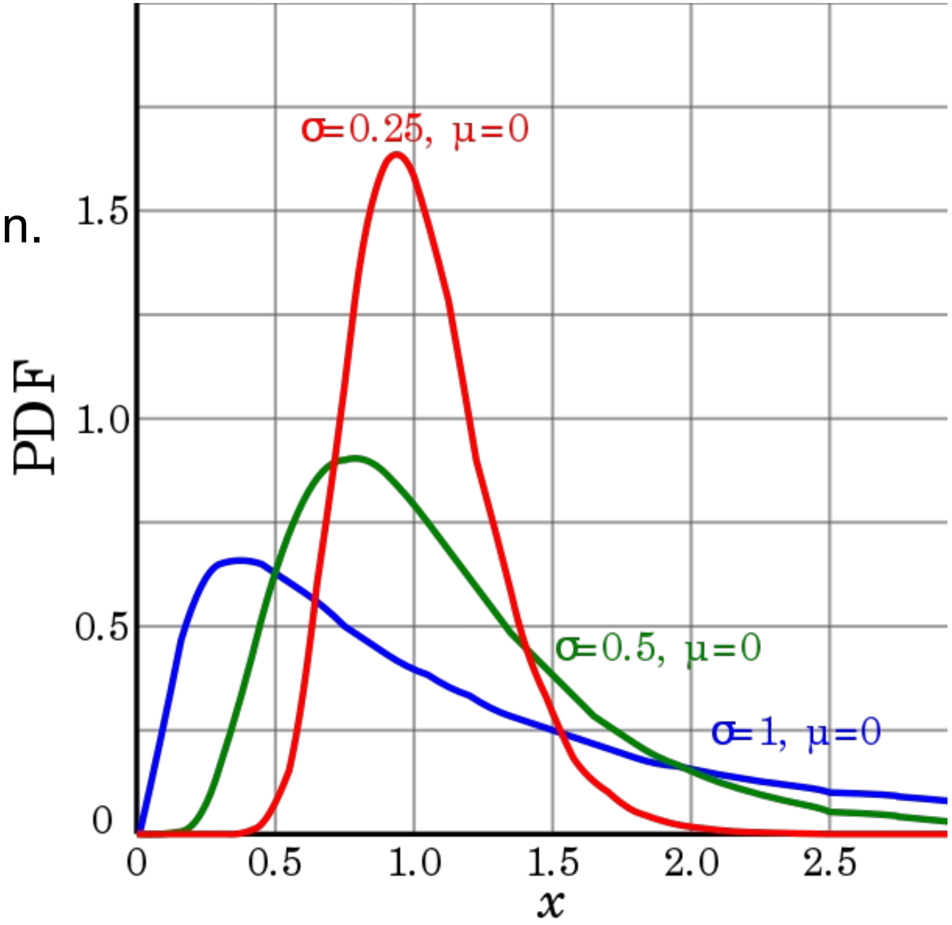
In other words, $X = e^{\mu+\sigma^2 Z}$ with Z satisfying the standard normal distribution.

By the central limit theorem, this naturally arises from a multiplicative (as opposed to additive) process:

Additive process in log space.

This appears in many natural phenomenon, including astrophysics (e.g., the initial core mass function).

log-normal distribution



Outline

- Basic probability
- Random variables
- Univariate distribution functions
- Descriptive statistics & data-based estimates
- Central limit theorem
- Multivariate distribution functions, correlation and covariance

Covariance and correlation

Given two random variables X and Y, how to describe the relation between them?

The simplest statistics that does this is called covariance:

$$\mathrm{Cov}(X,Y) = E\Big((X-\mu_X)(Y-\mu_Y)\Big)$$

(joint variability between X and Y)

One can further define the (population) correlation coefficient as

$$\rho = \rho_{X,Y} = \rho(X,Y) \equiv \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

where

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

$$\sigma_Y = \sqrt{\mathrm{Var}(Y)}$$

If ρ=0, then we say X and Y are uncorrelated.

Covariance and correlation

It is straightforward to see the following:

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$$

$$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X,Y)$$

$$-1 \leq \rho(X,Y) \leq 1$$ (Cauchy-Schwarz inequality)

If X and Y are independent, then then are uncorrelated.

However, the converse is NOT true in general.

Let θ be a uniform distribution between [0, 2π]. Define ξ=cosθ, and ζ=sinθ.

You can find that Cov(ξ,ζ)=0 but clearly they are NOT independent.

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

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

$$\mathrm{Cov}(X_i, X_j) = E(X_i X_j) - \mu_i \mu_j$$

(covariant matrix)

$$= E\left[(\Sigma^{1/2} Z Z^T \Sigma^{1/2})_{ij}\right] = \Sigma_{ij}$$

Uncorrelated Gaussian variables are independent.

Bivariate normal distribution

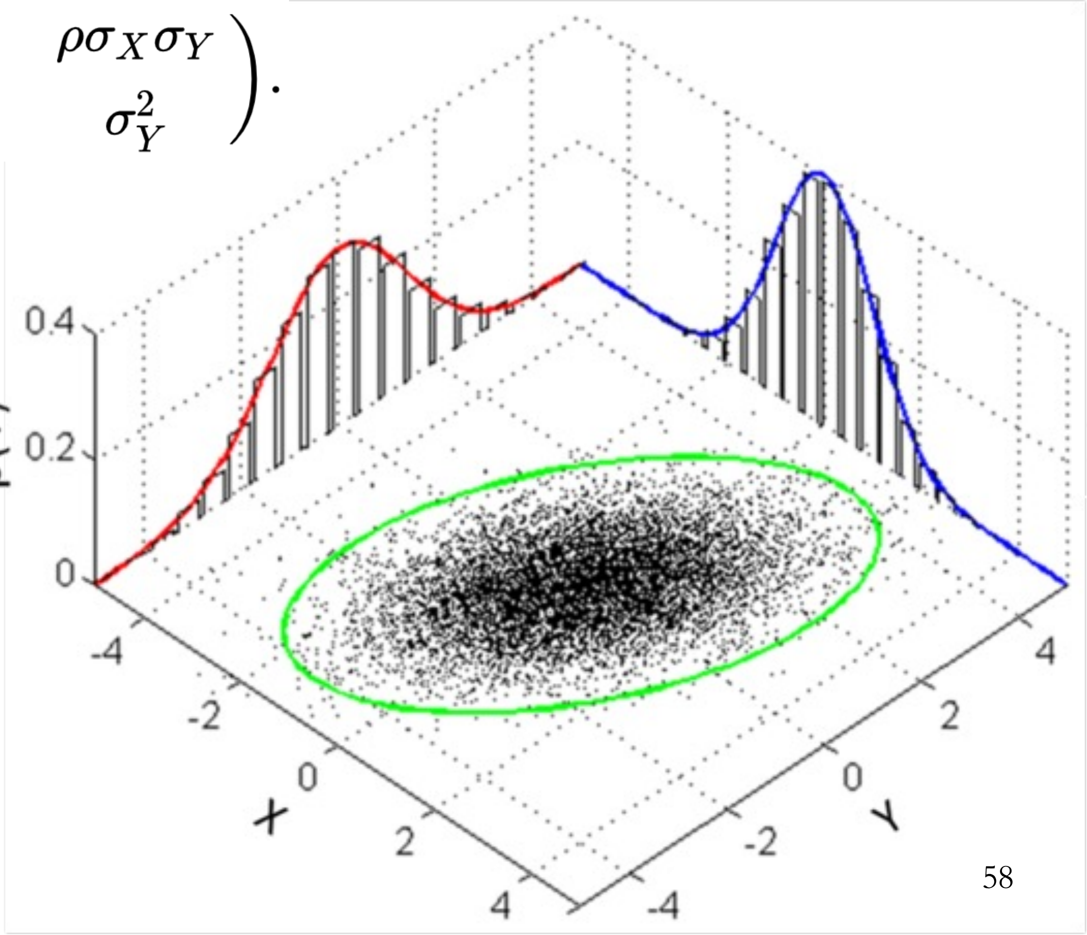
In 2D, we have

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

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.

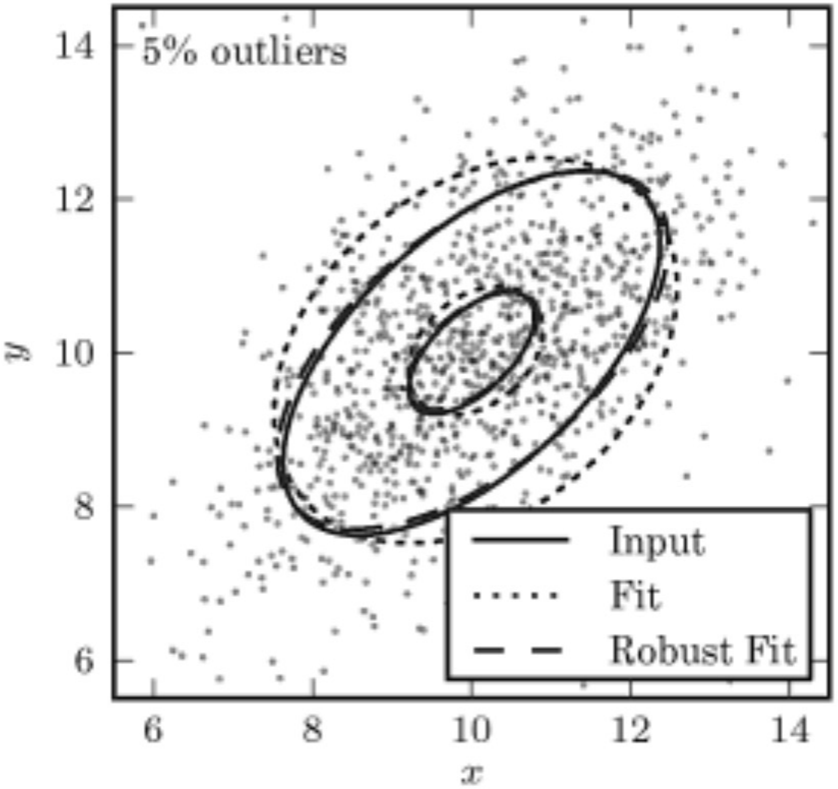
In the presence of outliers, a better way to estimate ρ is as follows:

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

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.