Random Variable (RV):

For a realization ω from the space of all possible realizations Ω, a random variable X takes the value X_0 in the subset A:

$$X(\omega) = X_0 \ , \ X_0 \in \mathrm{A}$$

If X is a discrete RV → A is a countable (or denumerable) set of values
If X is a continuous RV → A is a subset of $\mathbb{R}$

The support of the RV is the volume that is associated with the RV

Measurement of temperature:	support = point.
Measurement of incoming solar radiation:	support = surface.
Measurement of a pollutant concentration:	support = volume.

Examples

Throwing a dice → X_0 = 1 or 2 or 3 or 4 or 5 or 6
Temperature in the room → $X_0 \in$ {10, .. , 30} (°C)

Cumulative density function (CDF):

A random variable is fully described by its associated CDF F_x:

$$P[X(\omega) \leq a] = F_X(a)$$

F_X must satisfy the 3 following properties:

(1) $\lim_{x \to -\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x \to +\infty} F_X(x) = 1$

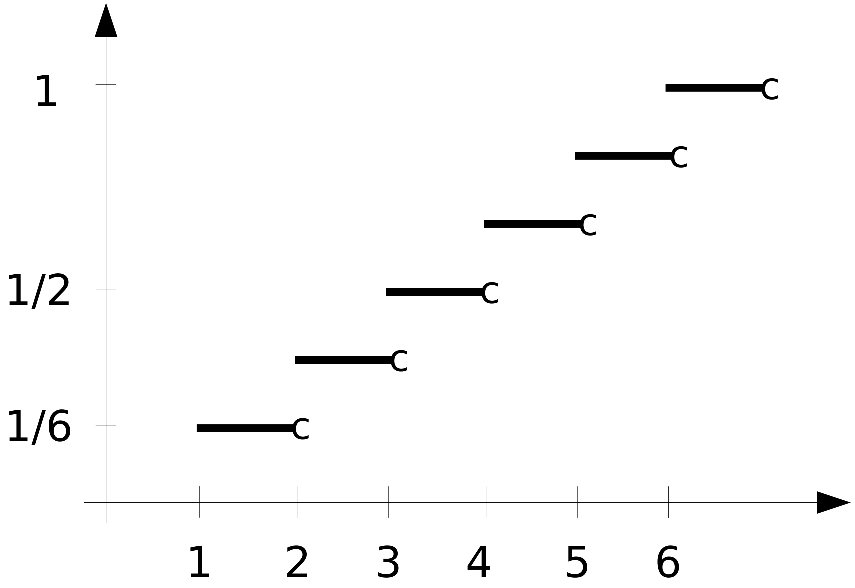
(2) F_X is a monotonic, nondecreasing function, that is $F_X(a) \leq F_X(b) \; \forall \, a < b$

(3) F_X is continuous from the right, that is $\lim_{h \to 0^+} F_X(x+h) = F_X(x)$

Examples of CDF:

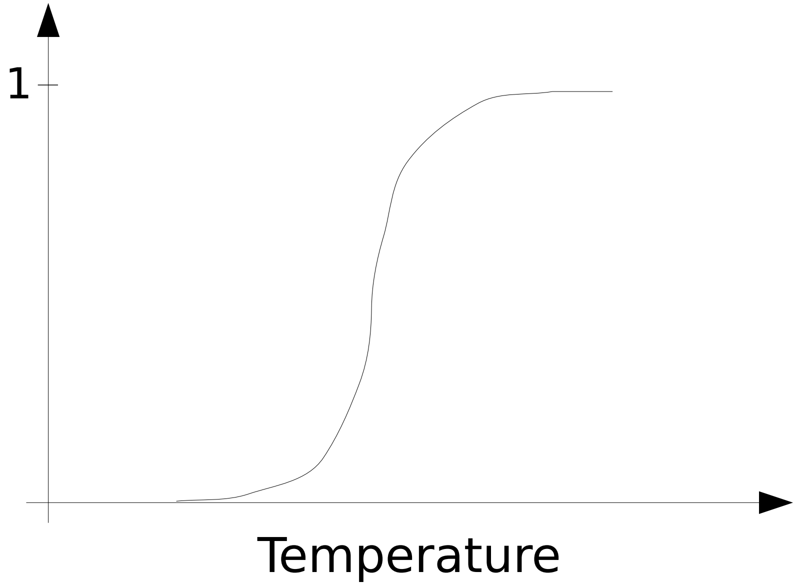
If X is a discrete RV:

X = outcome of a dice



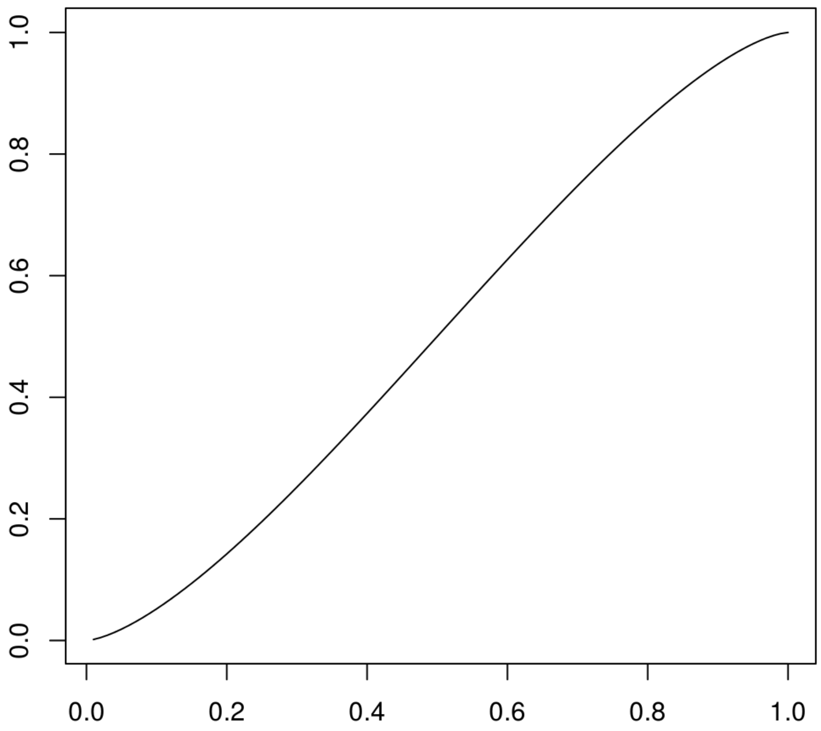
If X is a continuous RV:

X = temperature of the room

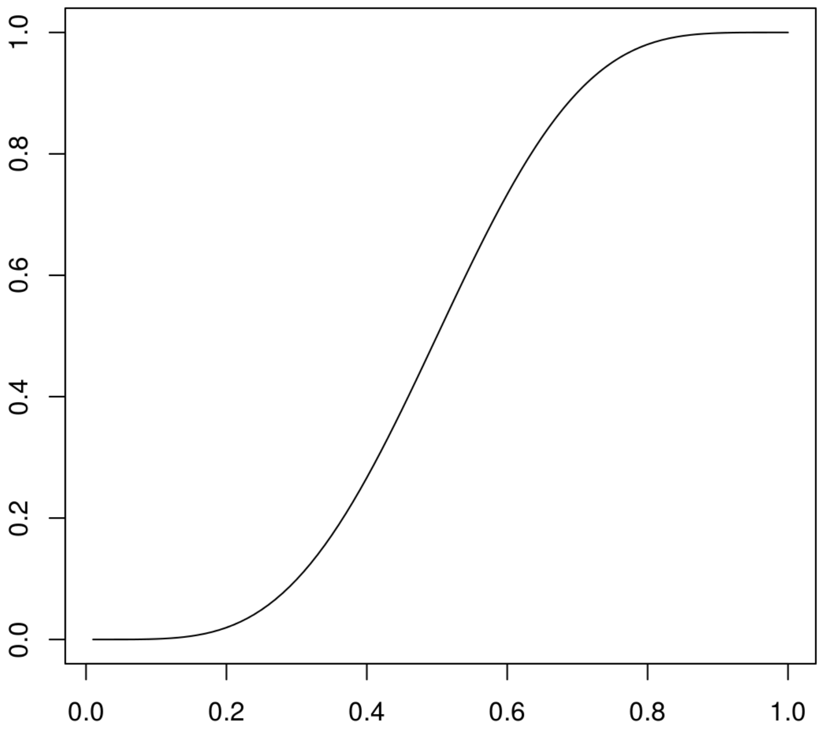


Which CDF corresponds to the most uniform distribution?

1



2



Probability density function (PDF):

A random variable can also be fully described by its associated PDF

If X is a discrete RV:
$$f_X(x) = \begin{cases} P[X = x_i] & \text{if } x = x_i \text{ , } i=1..n \\ 0 & \text{if } x \neq x_i \end{cases}$$

$$\sum_i f_X(x_i) = 1$$

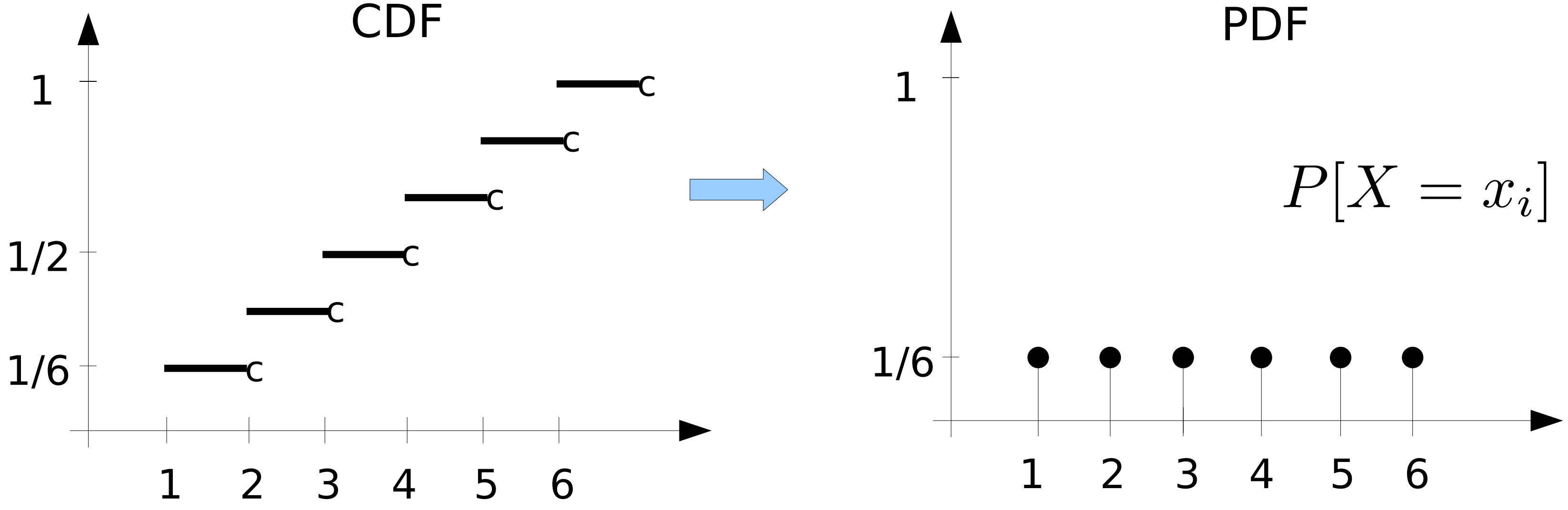
If X is a continuous RV:
$$f_X(x) = \frac{dF_X(x)}{dx}$$

$$P[x < X(\omega) \leq (x + \mathrm{dx})] = f_X(x)\mathrm{dx}$$

$$\int_A f_X(x)\,dx = 1$$

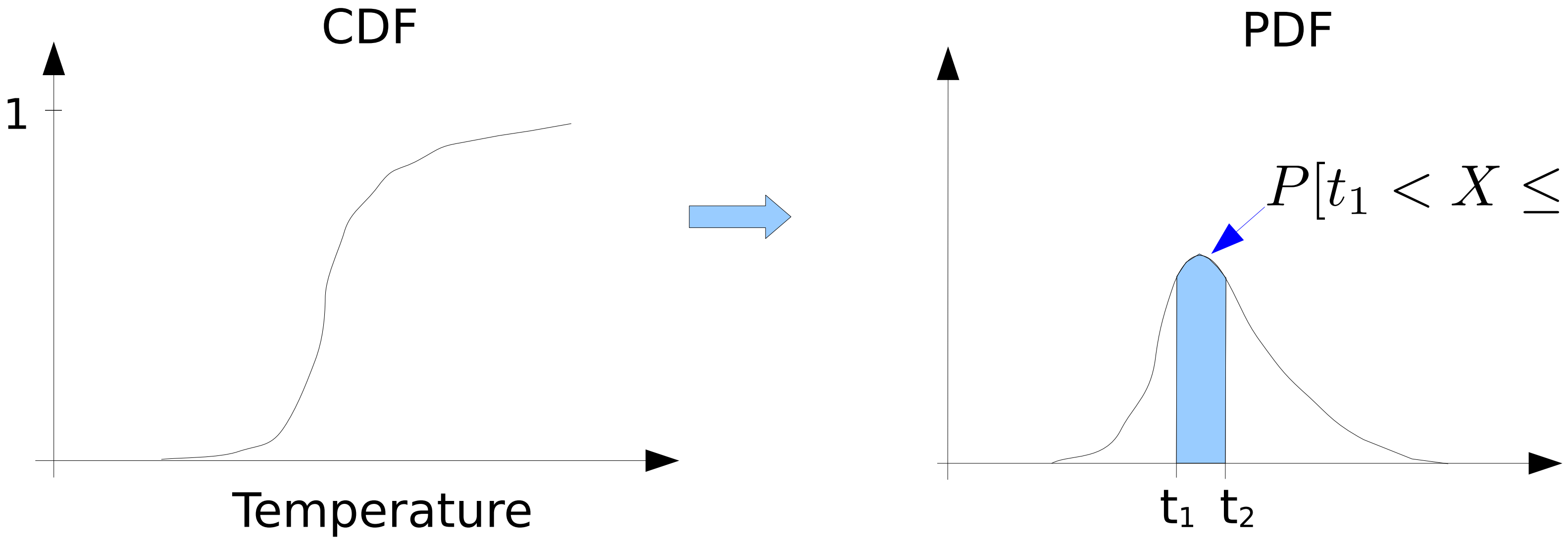
Examples of PDF:

If X is a discrete RV:



$$P[X = x_i] = f_X(x_i)$$

If X is a continuous RV:



$$P[t_1 < X \leq t_2] = \int_{t_1}^{t_2} f_X(x)\, dx$$

Descriptors of a RV

Expected value of a discrete RV:

$$\mathrm{E}[X] = \sum_i f_X(x_i)x_i$$

Expected value of a function g of X:

$$\mathrm{E}[g(X)] = \sum_i f_X(x_i)g(x_i)$$

Expected value of a continuous RV:

$$\mathrm{E}[X] = \int_A x f_X(x)\, dx$$

Expected value of a function g of X:

$$\mathrm{E}[g(X)] = \int_A g(x) f_X(x)\, dx$$

Expectation is a linear operator:

$$\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$$

Descriptors of a RV

Raw moments of a RV

$$\mu'_n(X) = \mathrm{E}[X^n]$$

Central moments of a RV

$$\mu_1(X) = \mu'_1(X)$$

$$\mu_n(X) = \mathrm{E}\left[(X - E[X])^n\right]$$

$$\mu_n(X) = \sum_{k=0}^{n} (-1)^{n-k} \tbinom{n}{k} \mu'_k \mu_1^{n-k}$$

Binomial series identity:

$$(a+b)^n = \sum_{k=0}^{n} \tbinom{n}{k} a^k b^{n-k} \qquad \tbinom{n}{k} = \frac{n!}{k!(n-k)!}$$

Moment generating function $M_X(t)$

$$M_X(t) = \mathrm{E}[e^{tX}]$$

$$M_X(t) = \mathrm{E}[1 + Xt + \frac{1}{2!}(Xt)^2 + \frac{1}{3!}(Xt)^3 + ...]$$

$$M_X(t) = 1 + \mu'_1 t + \frac{1}{2!}\mu'_2 t^2 + \frac{1}{3!}\mu'_3 t^3 + ...$$

$$M_X(t) = \sum_{i=0}^{\infty} \frac{1}{i!}\mu'_i t^i$$

Therefore $\boxed{\frac{d^n M_X}{dt^n}(0) = \mu'_n}$

For a discrete RV: $$M_X(t) = \sum_i e^{tx_i} f_X(x_i)$$

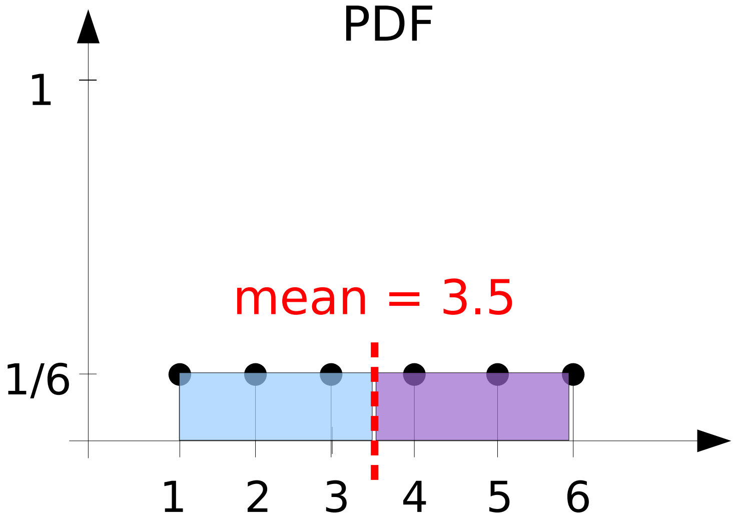
For a continuous RV: $$M_X(t) = \int_A e^{tx} f_X(x)\, dx$$

To describe the “center” of a distribution

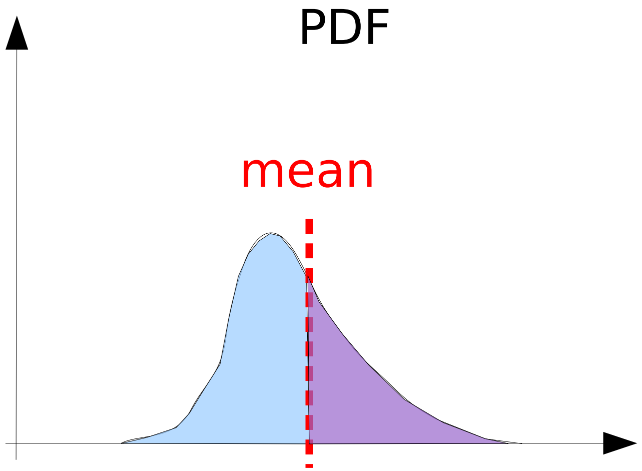
(using mainly 1st order moments)

(1) Mean $\mu(X) = \mathrm{E}[X]$ Mean = x value of the center of gravity of PDF

Discrete RV:



Continuous RV:

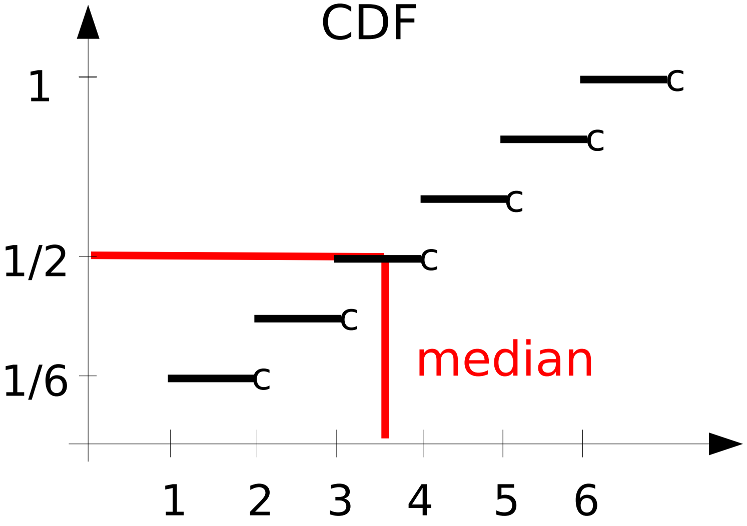


(2) Median

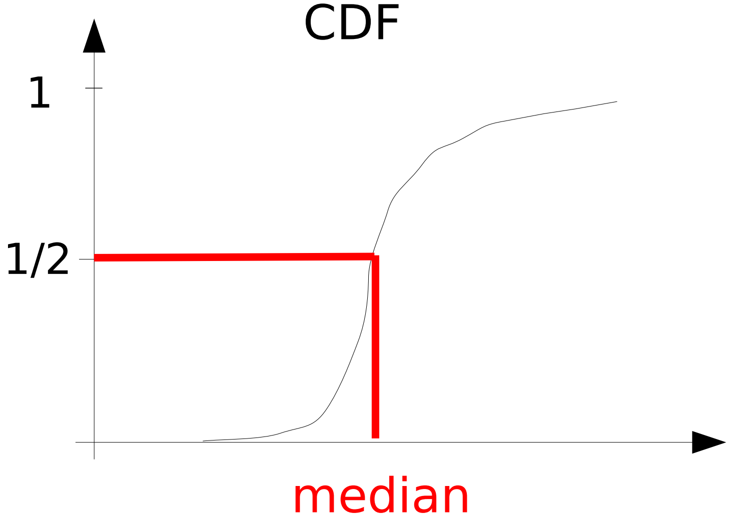
quantile y% means y% of the values are below q.
median : 50% of the values are below the median.
median is more robust than mean.

X in {xi,..,x_n} $\begin{cases} \text{if n odd} & \text{median} = x_{(n+1)/2} \\ \text{if n even} & \text{median} = 1/2(x_{n/2} + x_{1+n/2}) \end{cases}$

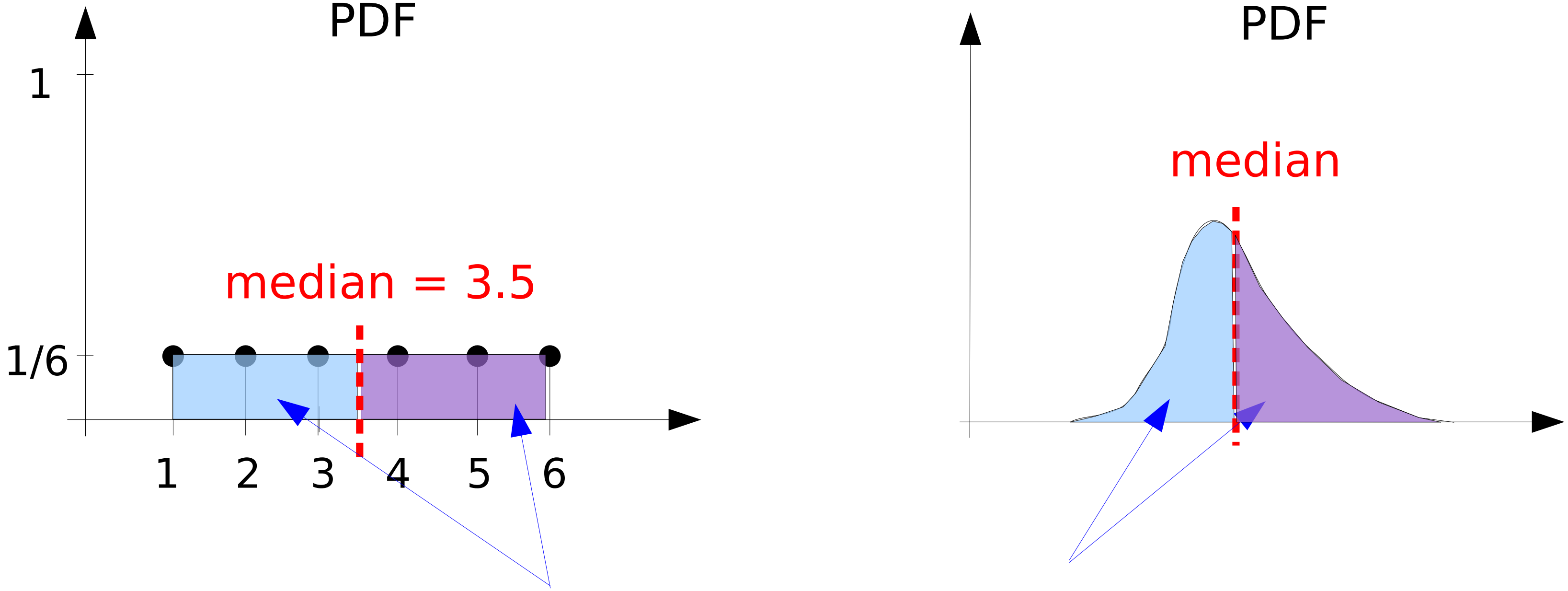
Discrete RV:



Continuous RV:



Discrete RV: Continuous RV:

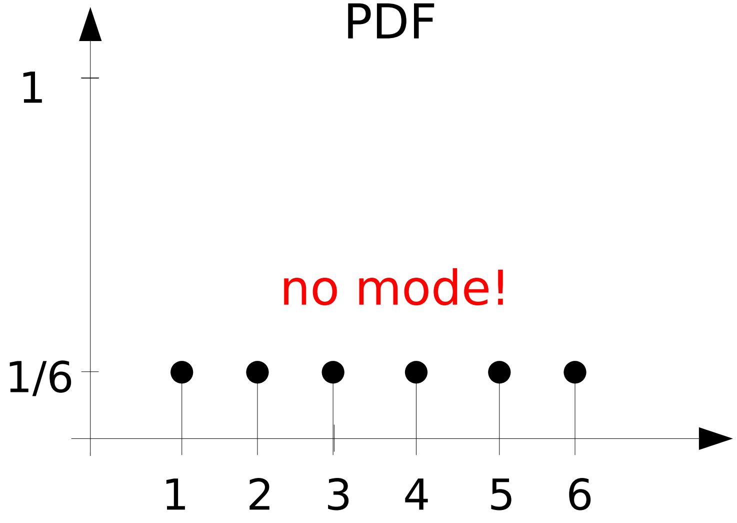


Integrated areas below and above median are equal

(3) Mode

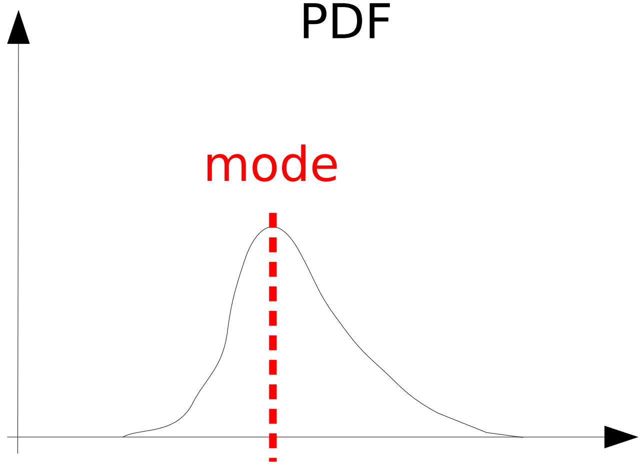
mode = most probable value = value at which f_X is max (if any).

Discrete RV:

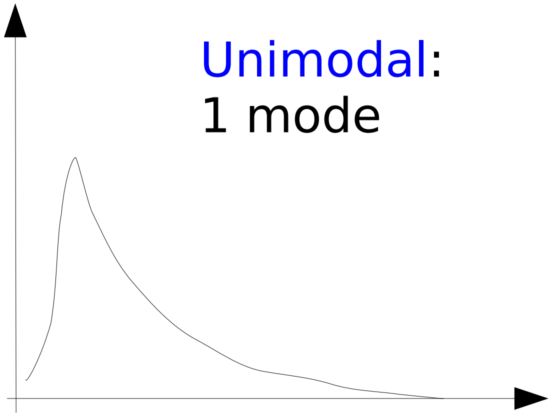


no mode!

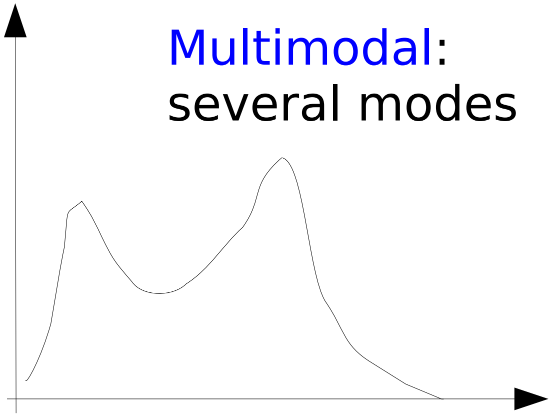
Continuous RV:



mode



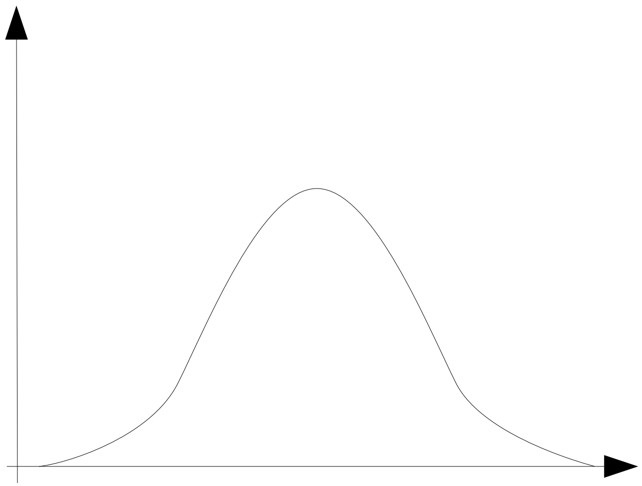
Unimodal: 1 mode

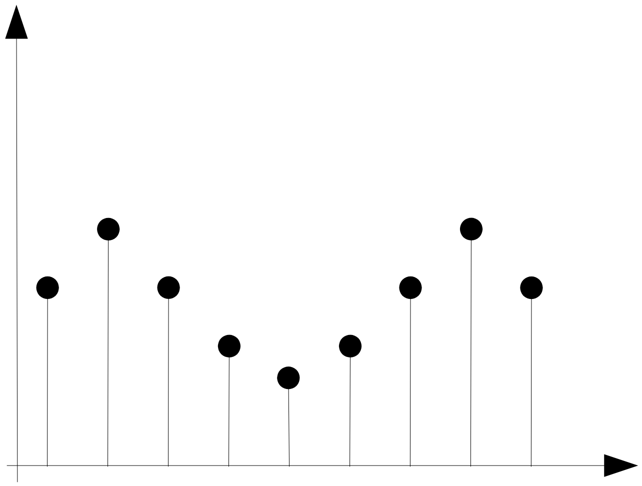


Multimodal: several modes

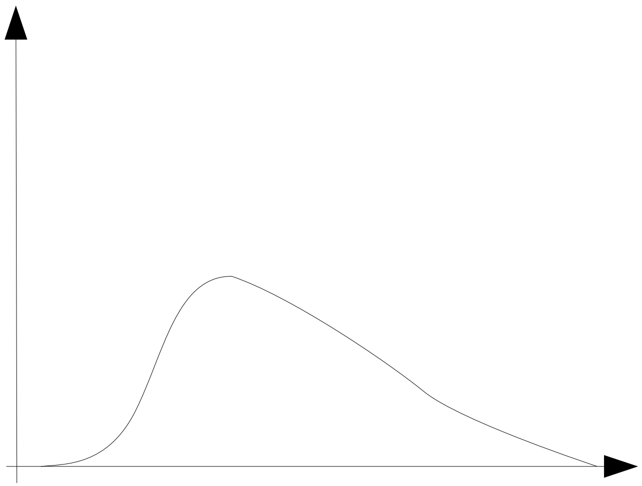
Indicate the relative locations of the mean, median and mode.

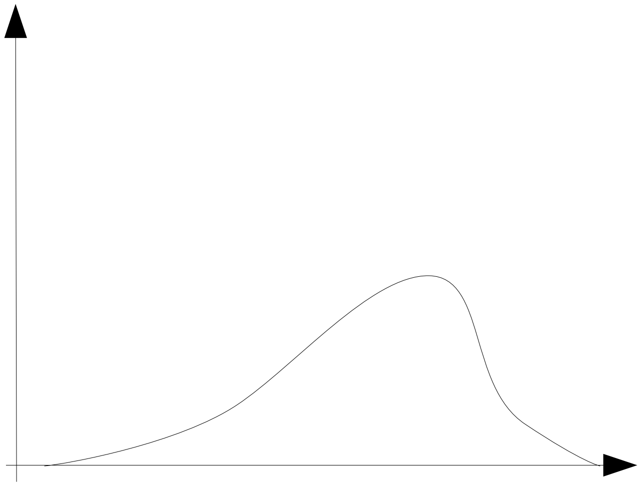
If PDF is symmetrical





If PDF is asymmetrical





To describe the “dispersion” of a distribution

(using moments order > 1)

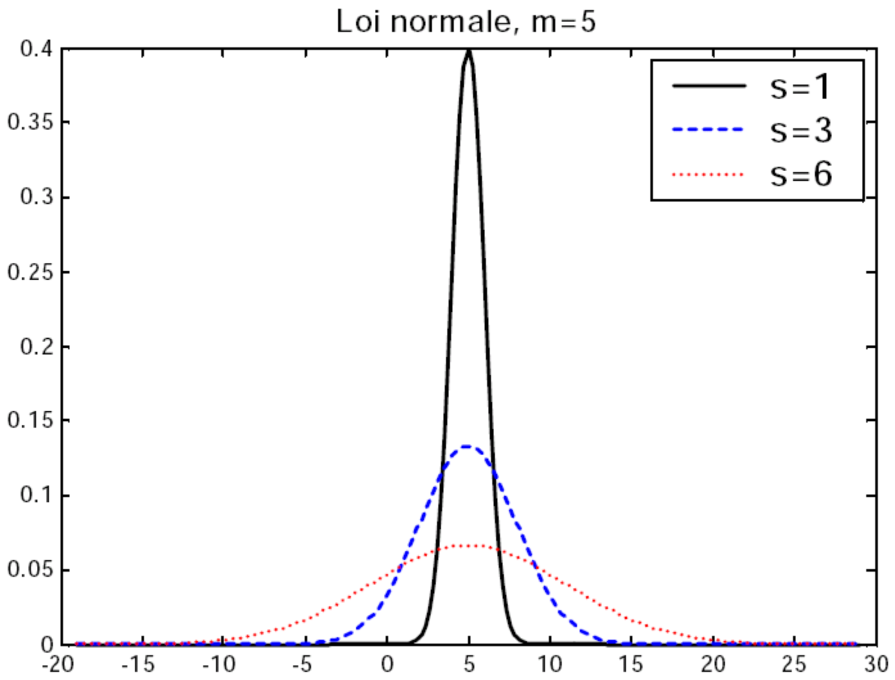
Variance

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

Standard deviation

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

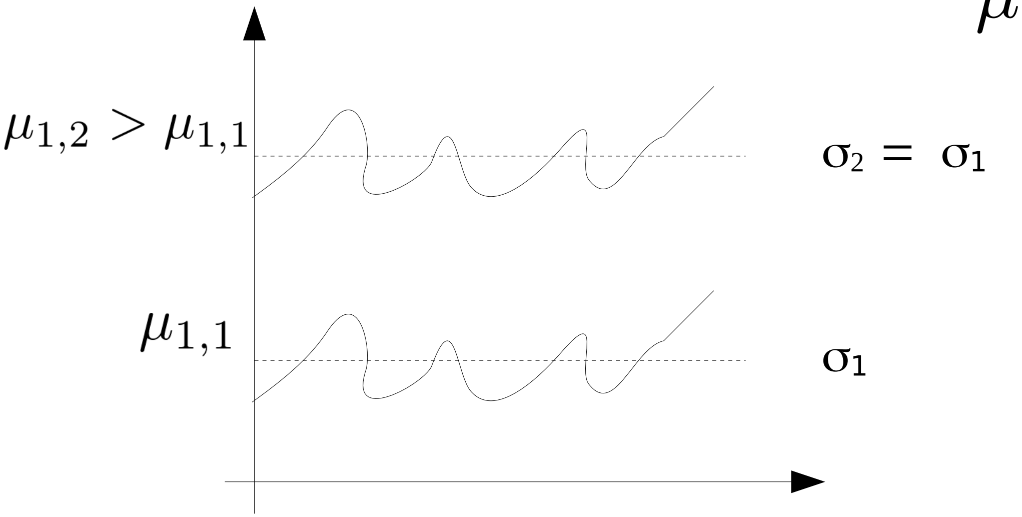
Quantify spread of the distribution around the mean



Coefficient of variation

$$CV = \frac{\sigma}{\mu_1}$$

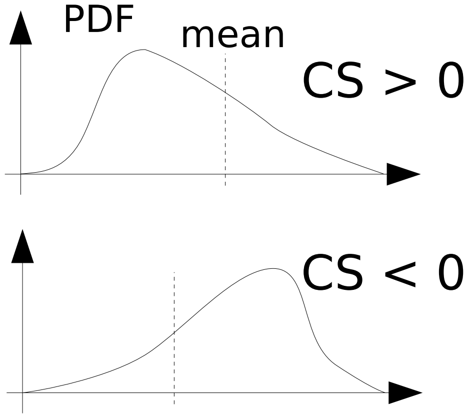
Quantifies variability
normalized by the mean



$CV_1 > CV_2$

Coefficient of skewness

$$CS = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}$$

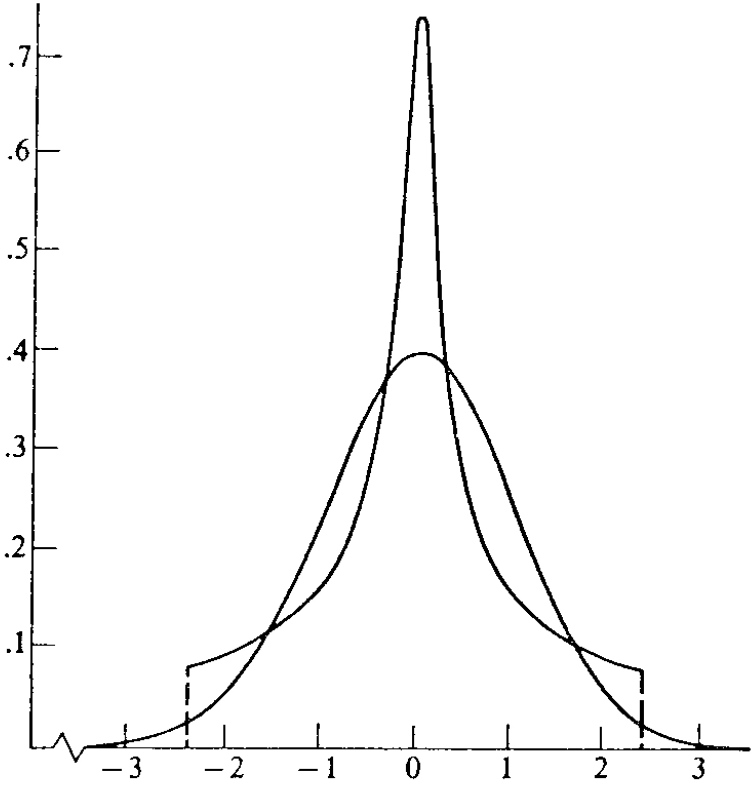


Coefficient of kurtosis

$$CK = \frac{\mu_4}{\sigma^4} - 3$$

CK > 0 → dist. more peaked than Normal dist.
CK < 0 → dist. less peaked than Normal dist.

Distributions can have same first 4 moments, but still be different!!!

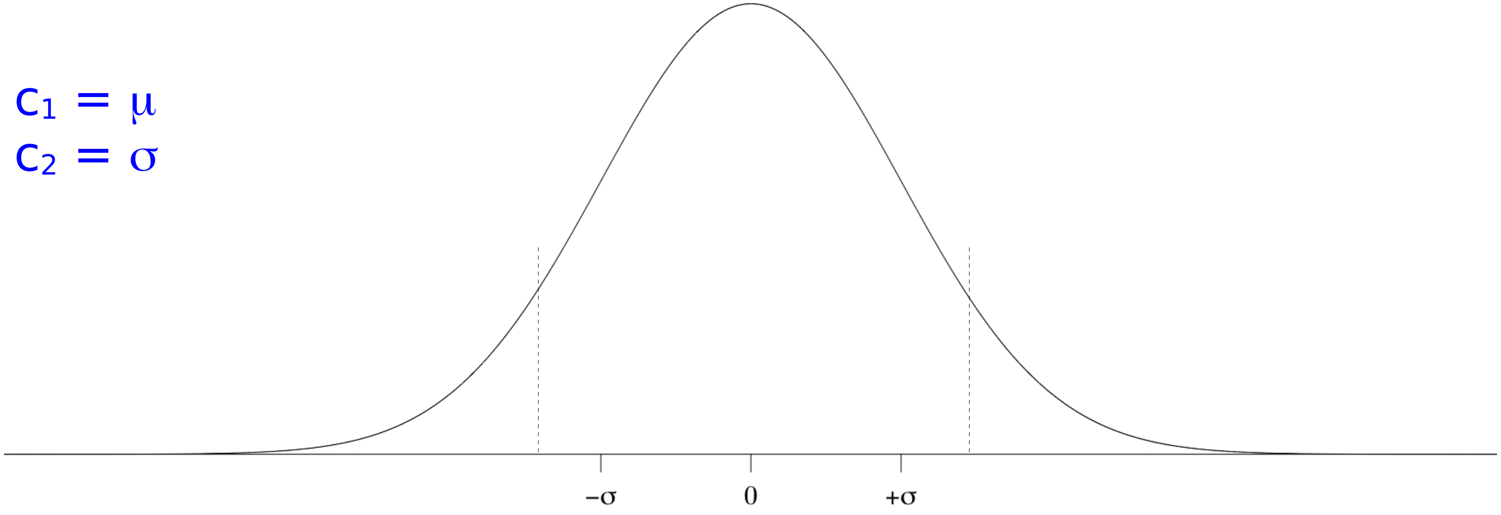


From Mood *et al.*, Introduction to the theory of Statistics, 1974

The Normal distribution

$$f(x) = \frac{1}{c_2\sqrt{2\pi}} \exp\left\{-\frac{(x-c_1)^2}{2c_2^2}\right\}$$

$c_1 = \mu$
$c_2 = \sigma$



Normal Distribution

$$\left.\begin{aligned} q_{10} &= \mu - 1.28\sigma \\ q_{90} &= \mu + 1.28\sigma \end{aligned}\right\}$$ 80% of the values in $[\mu - 1.28\sigma, \mu + 1.28\sigma]$

Normal (or Gaussian) dist. plays a major role because the dist. of the sum of n RV tends toward a normal distribution (central limit theorem).

Let X be a RV with a Normal distribution

$$M_X(t) = \int_{-\infty}^{+\infty} e^{tx} f_X(x)\, dx = \frac{1}{c_2\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{tx} e^{-\frac{(x-c_1)^2}{2c_2^2}}\, dx$$

$$M_X(t) = e^{tc_1 + \frac{t^2 C_2^2}{2}}$$

Hence

$$\mu_1 = \frac{dM_X}{dt}(0) = c_1$$

$$\sigma^2 = \frac{d^2 M_X}{dt^2}(0) - \mu_1^2 = c_2^2$$

Considering 2 RV X and Y, one possible way to characterize their relationship is to calculate their covariance or their correlation coefficient.

Covariance:
$$\begin{aligned}\mathrm{Cov}[X,Y] &= E[(X-E[X])(Y-E[Y])]\\ &= E[XY]-\mathrm{E}[X]\mathrm{E}[Y]\end{aligned}$$

Correlation coefficient:
$$\rho_{XY} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X\sigma_Y}$$
(ρ = Cov for RV with std = 1)

Both coefficients measures the linear relationship between X and Y.

Correlation coef. is better in this sense because it is dimensionless and normalized, so it eases comparison.

Properties of variance

$$\mathrm{Var}[aX+b] = a^2\mathrm{Var}[X]$$

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$$

useful in geostat

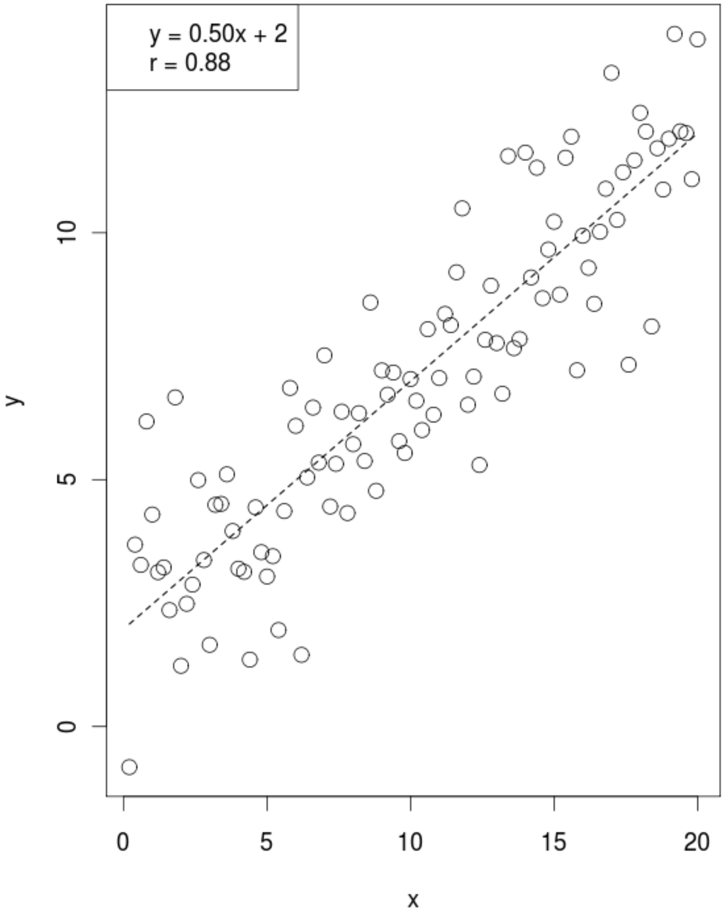
$$\mathrm{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{Cov}[X_i, X_j]$$

$$\mathrm{Var}\left[\int_A cZ(x)\,dx\right] = c^2 \int_A \int_A \mathrm{Cov}[Z(x), Z(y)]\,dx\,dy$$

Correlation and linear regression

Linear regression of Y as function of X

$$\hat{Y} = aX + b$$



Best line Ŷ that fits XY cloud.

Least squares: G = E[(Ŷ-Y)²] is minimum

$$\frac{\partial G}{\partial a} = 0 \quad \text{and} \quad \frac{\partial G}{\partial b} = 0$$

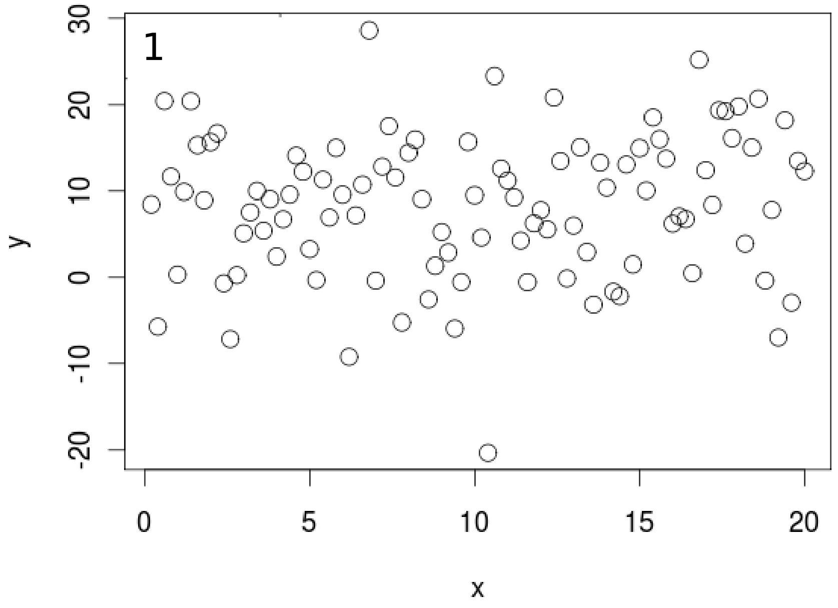
$$\begin{cases} a = \dfrac{\mathrm{Cov}[X,Y]}{\sigma_X^2} = \dfrac{\boxed{\rho_{XY}}\,\sigma_Y}{\sigma_X} \\ b = \mathrm{E}[Y] - a\mathrm{E}[X] \;\Rightarrow\; \mathrm{E}[\hat{Y}] = \mathrm{E}[Y] \end{cases}$$

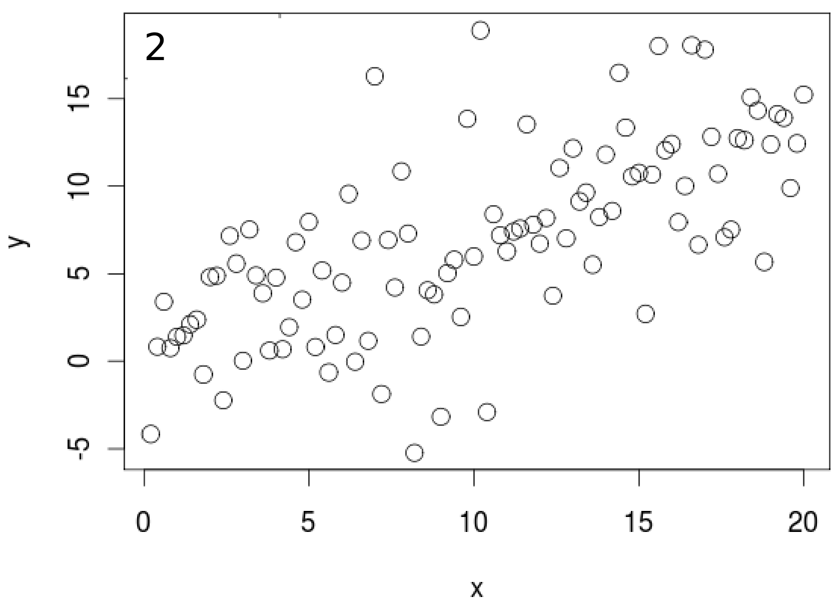
(line passes by gravity center)

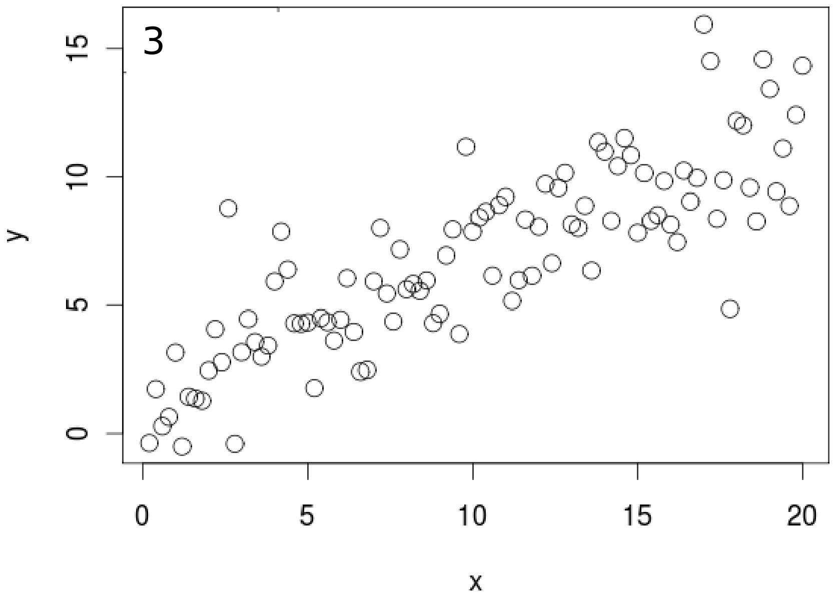
$$\mathrm{Var}[\hat{Y} - Y] = \mathrm{Var}[Y](1 - \rho_{XY}^2)$$

→ ρ_{XY}^2 represents the percentage of variance of Y explained by Ŷ.

Scatter plots corresponding to different correlation coefficients







Sort graphs by decreasing correlation coefficient

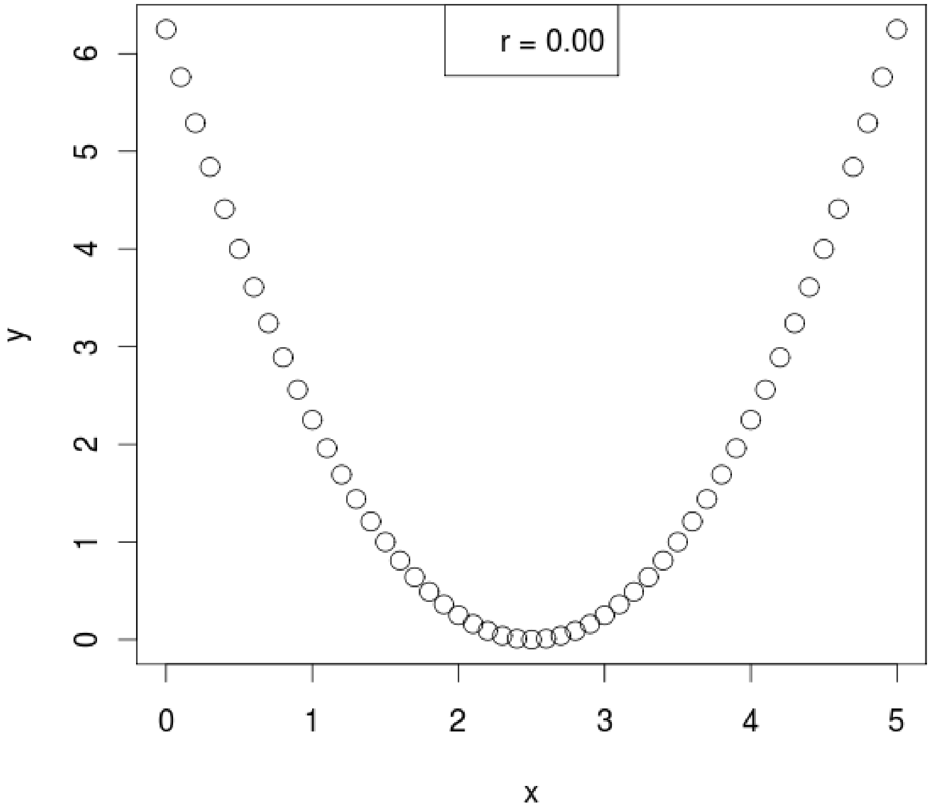
Limitations and pitfalls when using (linear) correlation coefficient

1. Correlation = 0 does not mean independence

$y = (x - 2.5)^2$

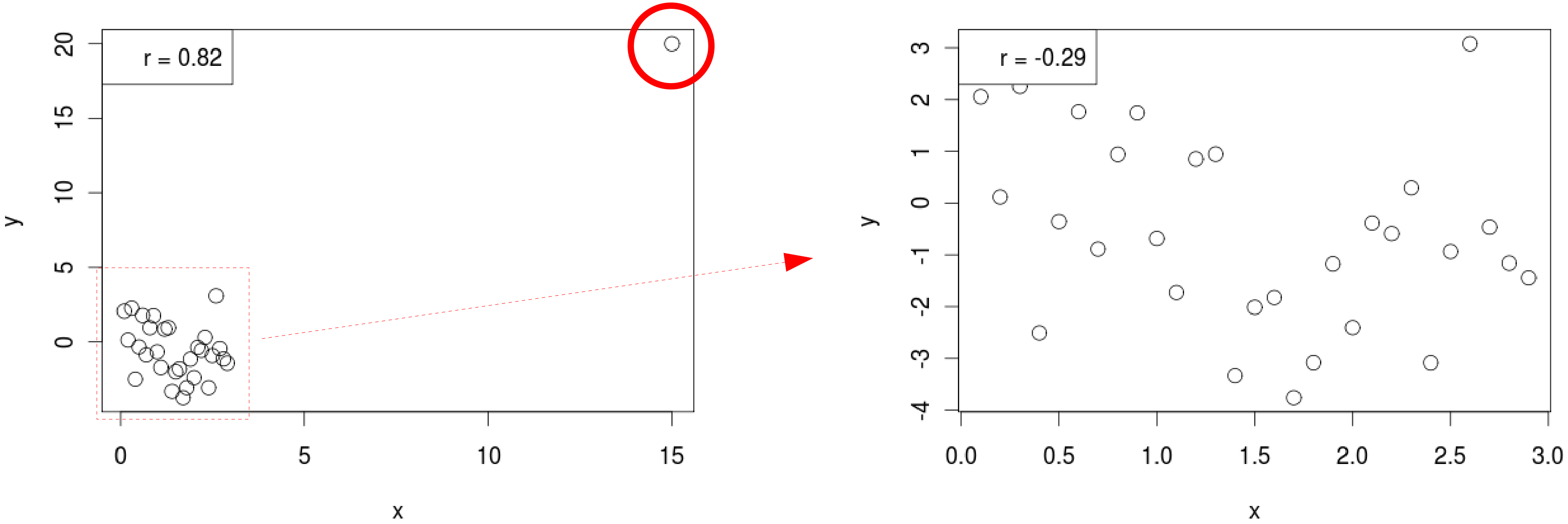
→ deterministic relationship
but $\rho = 0$!

Coef. correlation quantifies the “amount of LINEARITY” between 2 random variables



Limitations and pitfalls when using (linear) correlation coefficient

2. Correlation can be biased by isolated values (outliers)

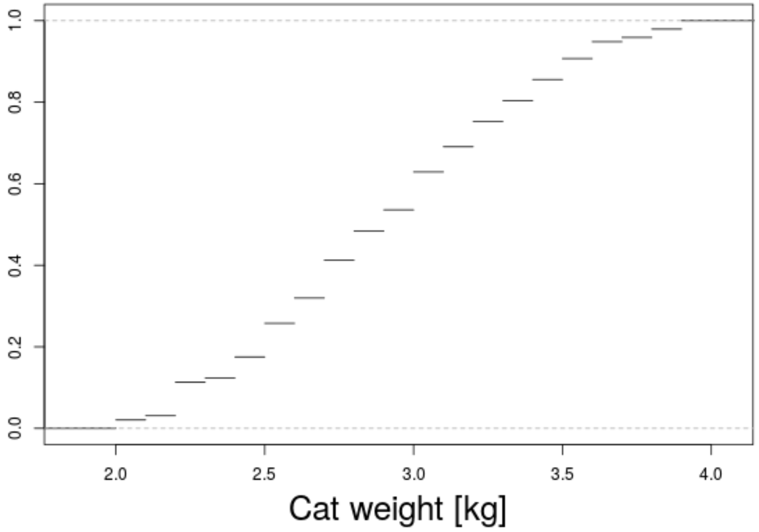


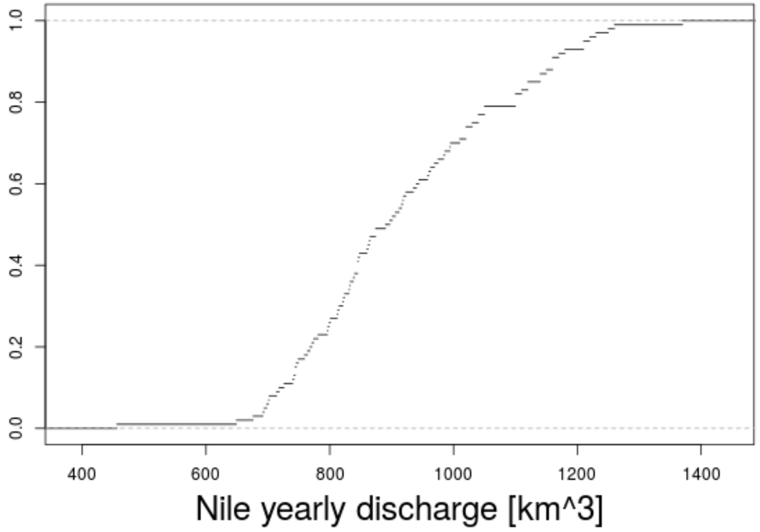
Limitations and pitfalls when using (linear) correlation coefficient

3. High correlation does not mean causality.

- “Hidden” variable: energy used for heating and number of deaths by cold in winter are correlated but no causality!

- Co-fluctuation: 2 CDF will have high correlation, whatever they represent.





Sampling and estimation

So far, RV with known distribution (CDF or PDF).

However, in practice, only access to a SAMPLE of the population.

Only access to estimates of descriptors (moments, quantiles), ex:

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

can and usually will be different from μ

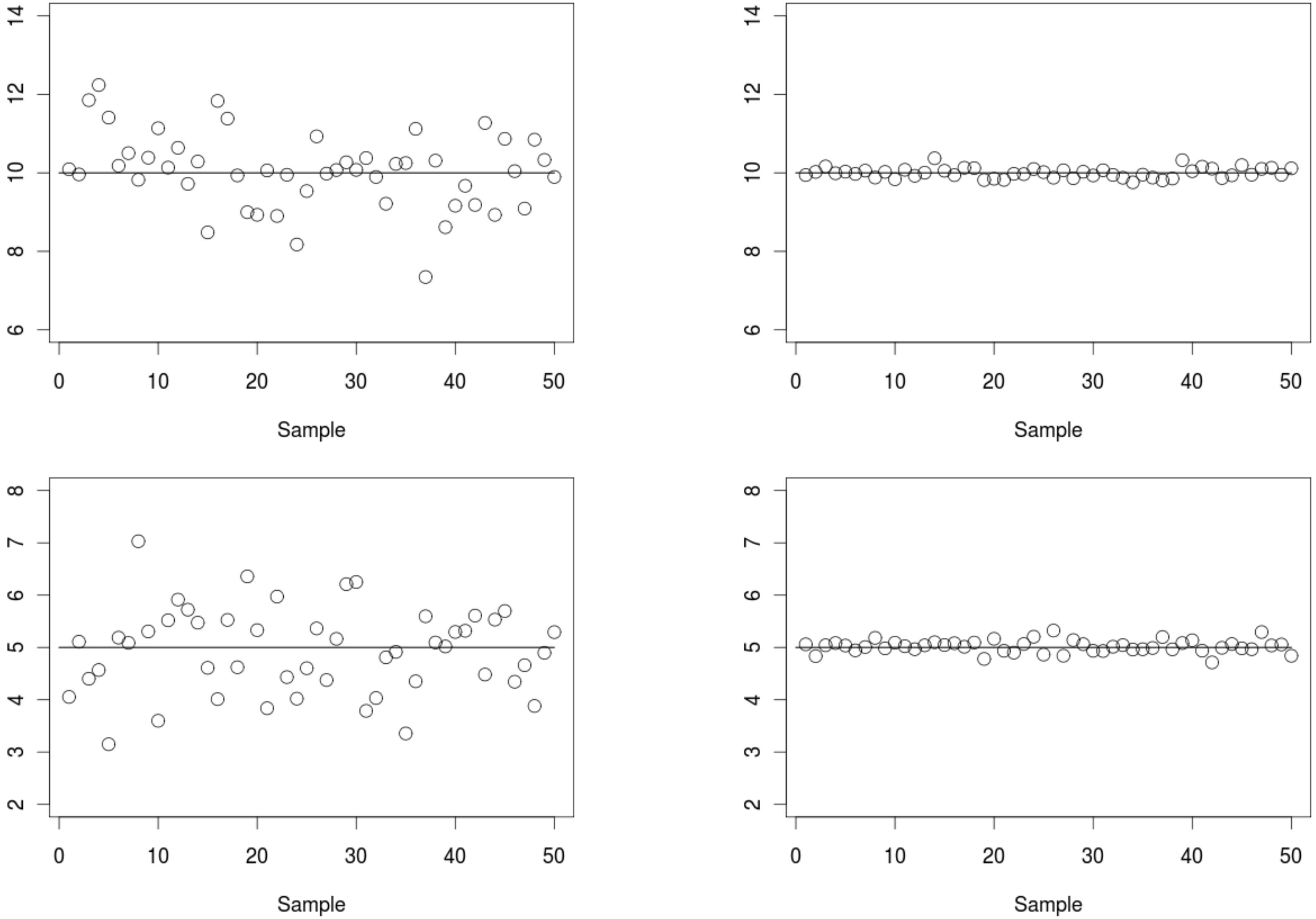
$$s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - m)^2$$

can and usually will be different from σ

Usually, greek letters will denote population moments while latin letters will denote sample moments.

Sampling and estimation

Example: generate 50 sets of 20 and 1000 normally distributed values. Mean = 10 and std = 5 for the population.



20 values

1000 values

Mean

Std