DATA ANALYSIS
Year 2019–2020

#3 Statistical Methods

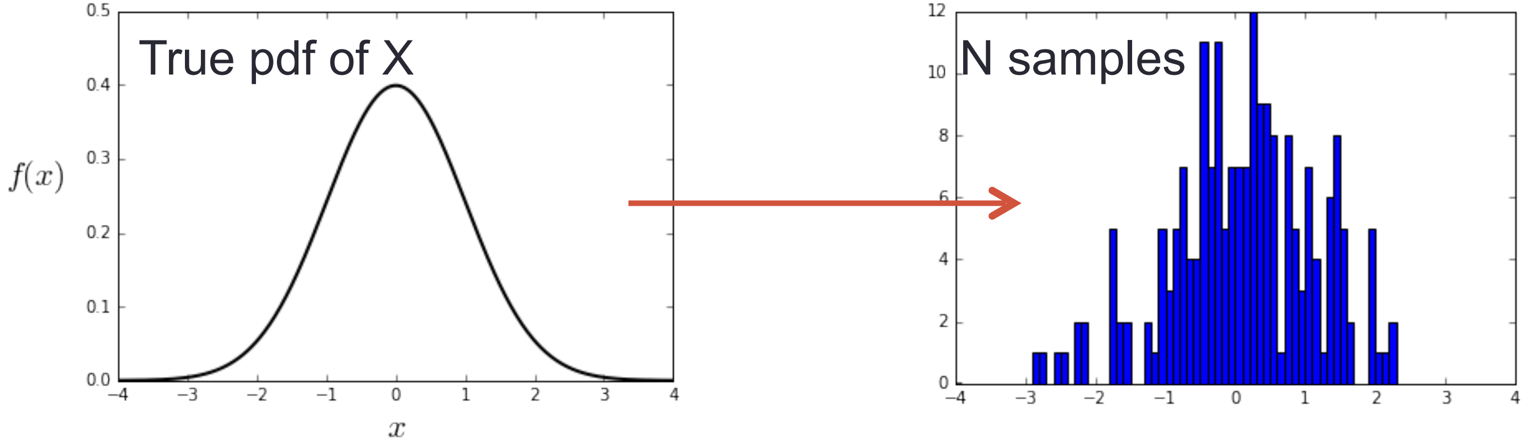
Master M1–Marine Physics
Université de Bretagne Occidentale

Jonathan GULA
[gula@univ-brest.fr]

Estimators

In practice, if X is a random variable, we will deal with a finite number N of empirical realizations of the random variables :

$$x_k \text{ for } k = 1..N$$

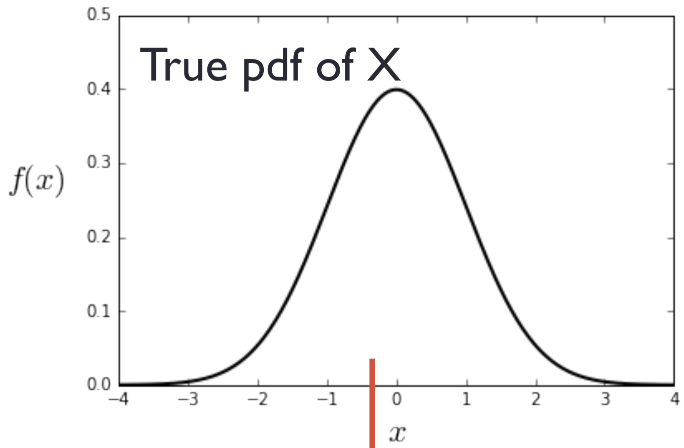


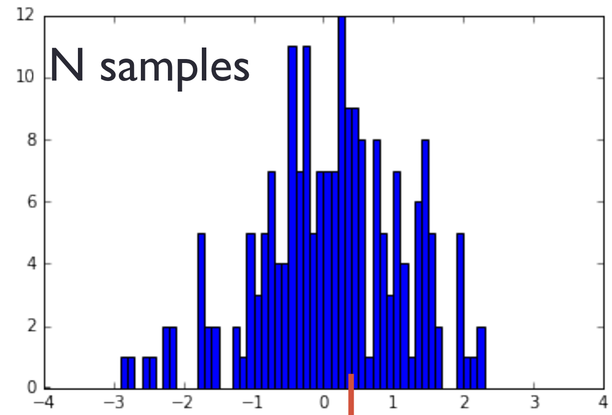
In practice we never know the true pdf but we can **estimate** it using the N samples.

We have access to the properties of X only via the empirical N samples.

Estimators

$$x_k \text{ for } k = 1..N$$





True population mean:

$$\mu = \int x f(x) dx$$

True population variance:

$$\sigma^2 = \int (x-\mu)^2 f(x)\, dx$$

Mean estimator (= sample mean)

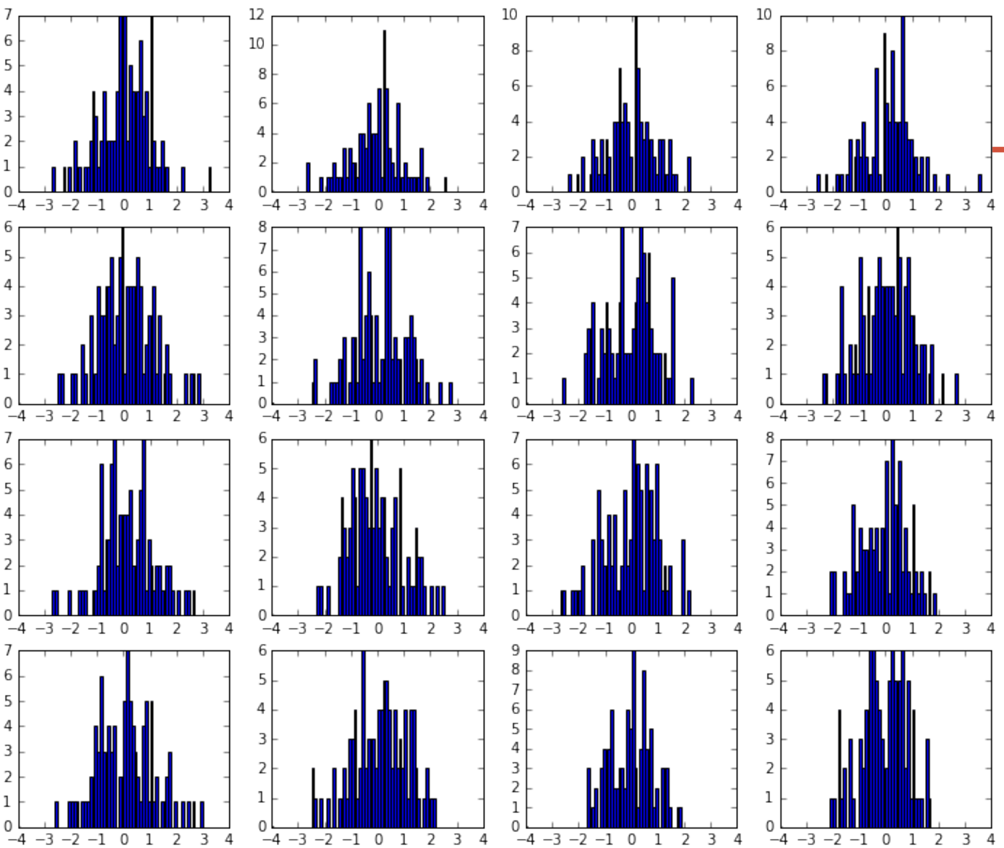
$$\hat{\mu}(x) = \frac{1}{N} \sum_k x_k$$

Variance estimator:

$$s^2 = \frac{1}{N-1} \sum_k (x_k - \hat{\mu})^2$$

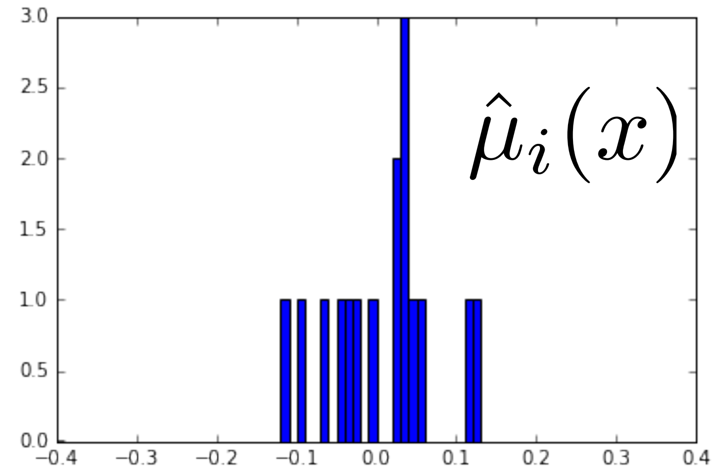
Central limit theorem

Statistics computed from random variables (mean, variance, etc.) are themselves random variables.



For each sample we compute the mean:

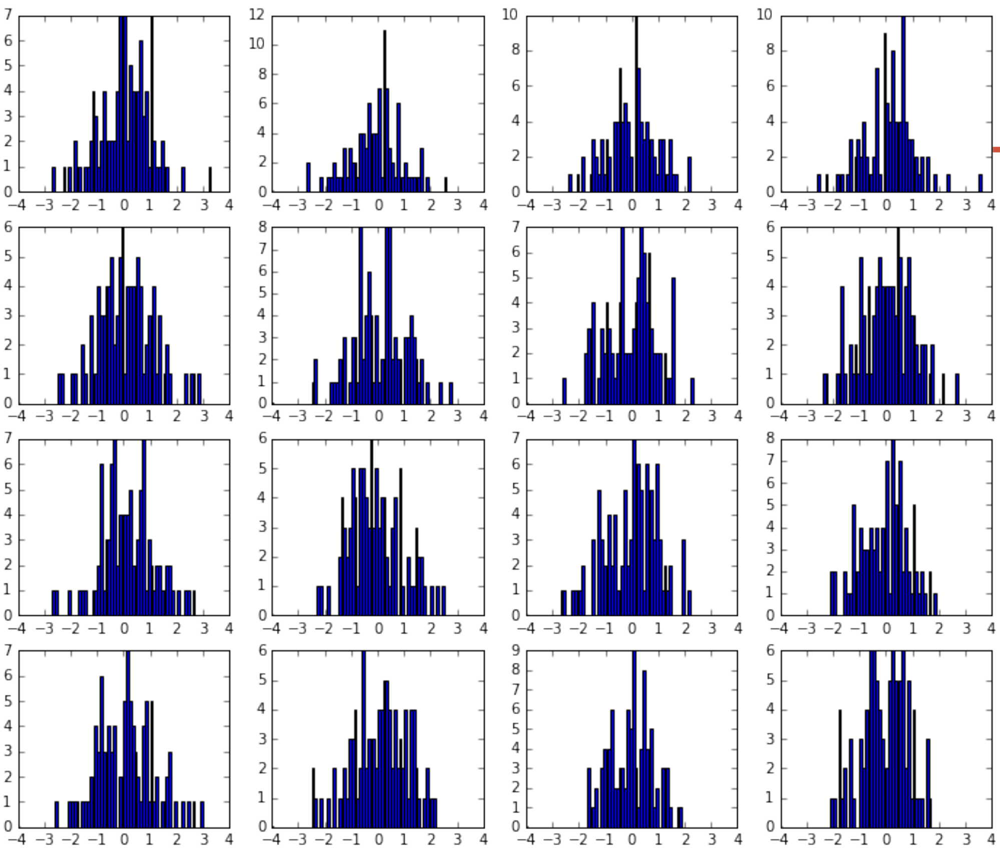
$$\hat{\mu}_i(x) = \frac{1}{N}\sum_k x_k$$



With m = 16 samples

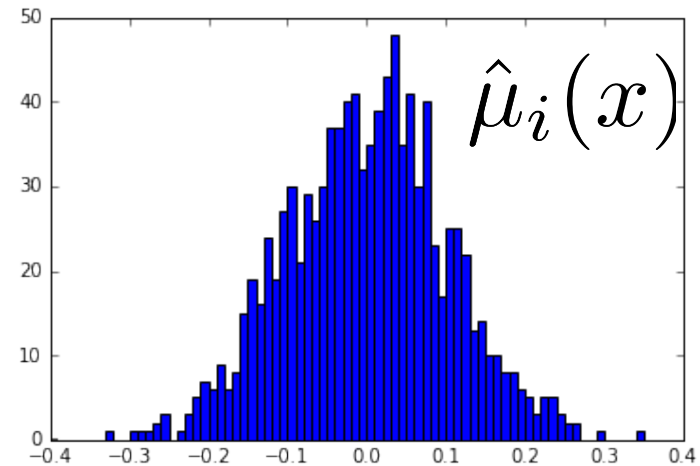
Central limit theorem

Statistics computed from random variables (mean, variance, etc.) are themselves random variables.



For each sample we compute the mean:

$$\hat{\mu}_i(x) = \frac{1}{N}\sum_k x_k$$



With m = 1000 samples

Central limit theorem

The **Central limit theorem** states that the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value (true population mean) and finite variance, will be approximately normally distributed, regardless of the underlying distribution.

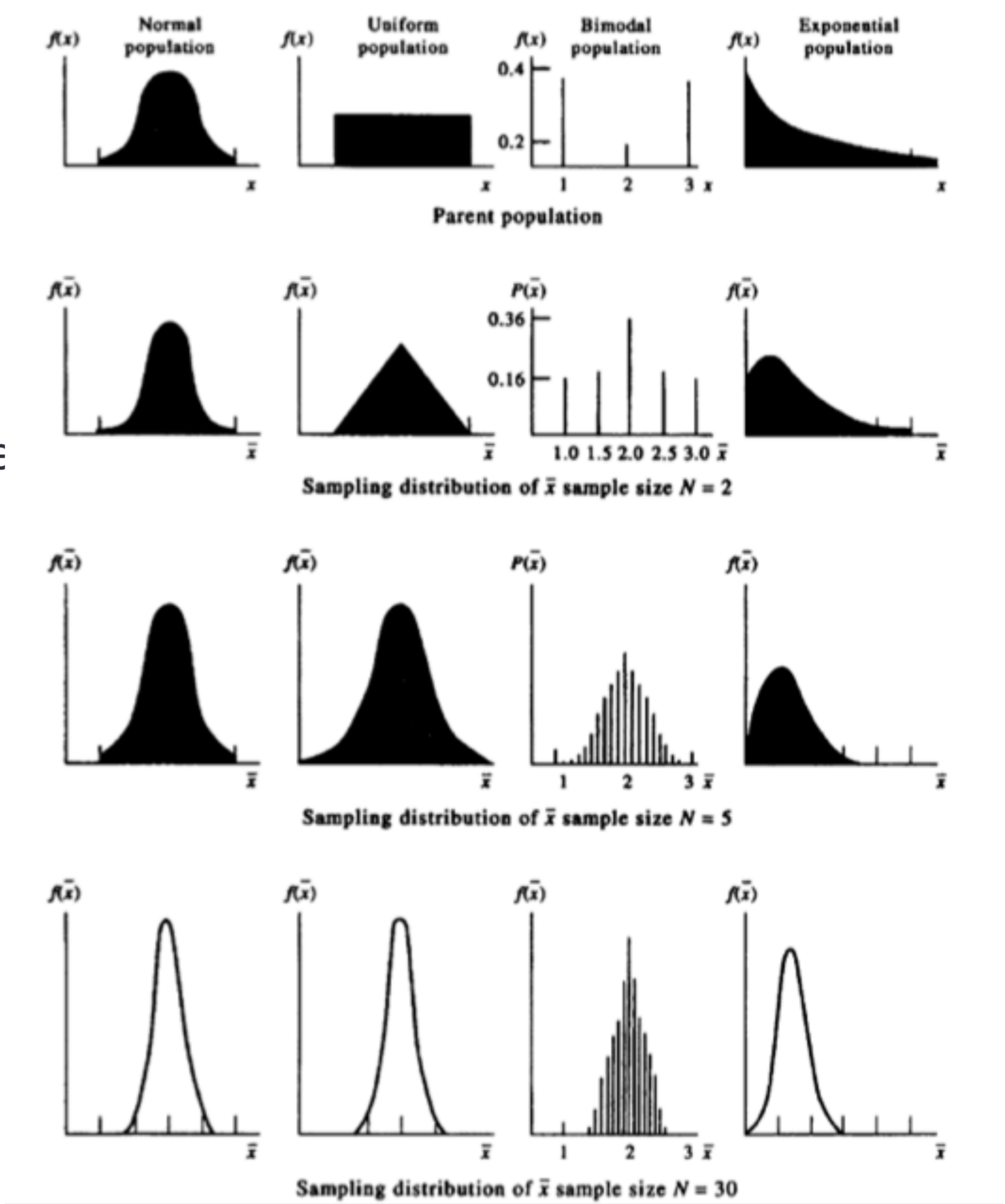
Let $X_i, i = 1..\mathrm{Ns}$ be a sequence of independent random variables (each containing N values) drawn from distributions with true mean μ and variance σ^2. Then as *Ns* becomes large, the distribution of the mean values $\hat{\mu}_i$ of each sample X_i approaches the normal distribution with mean μ and variance σ^2/N.

$$\hat{\mu}(x) \sim \mathcal{N}(\mu, \sigma/\sqrt{N})$$

(Regardless of the distribution of the original population variable from which the samples were drawn).

Central limit theorem

The fact that the $X_i, i = 1..N$ may have any kind of distribution is the reason for the importance of the normal distribution in probability theory and why the CLT is key in probability theory.



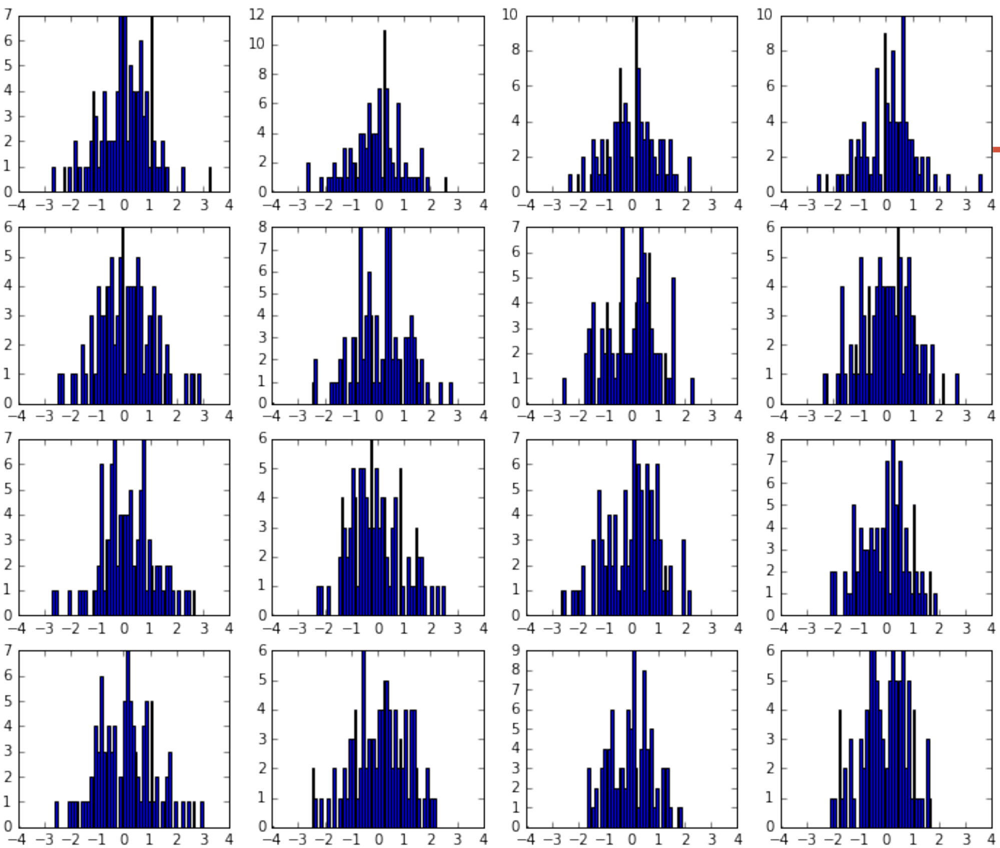
Central limit theorem

It has important implications in geophysics where you constantly average values in space and time.

For example, data from high-resolution CTD systems are generally vertically averaged (or averaged over some set of cycles in time), thus approaching a normal PDF for the data averages, via the central limit theorem.

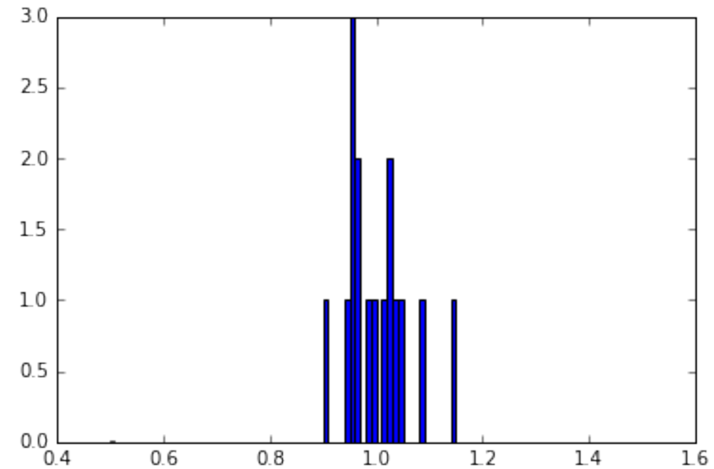
Variance as a random variable

Let's apply the same idea to the variance estimate.



For each sample we compute the estimated variance:

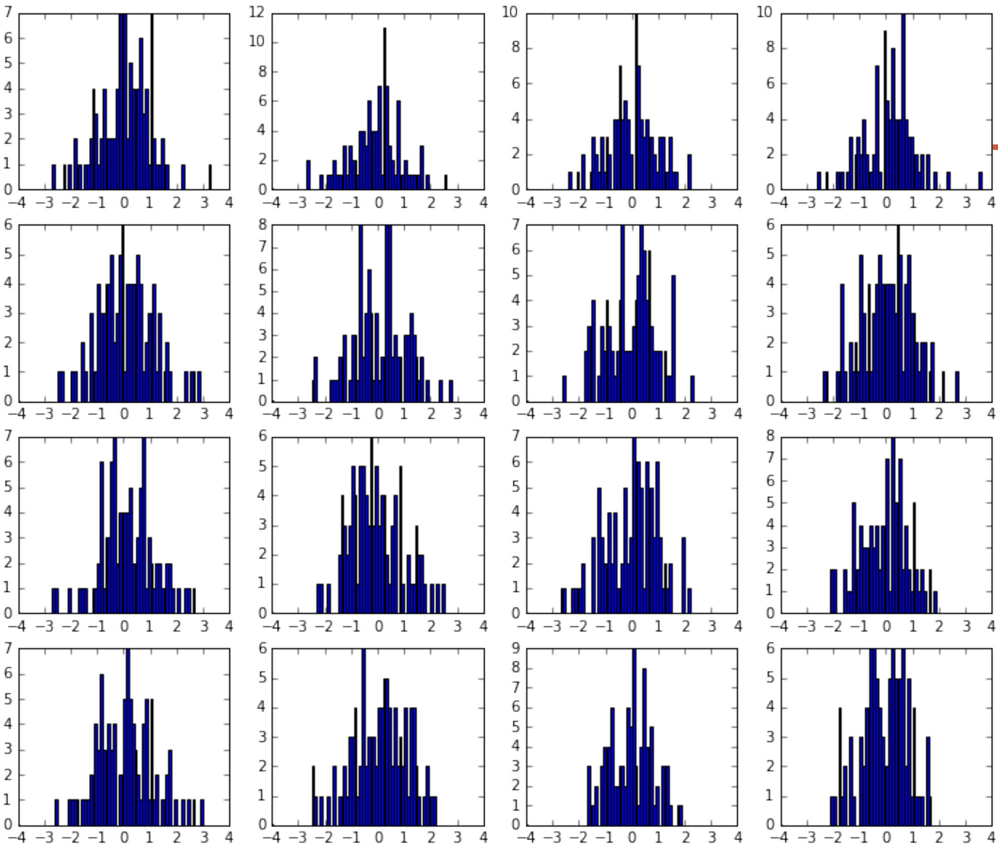
$$s^2 = \frac{1}{N-1}\sum_k (x_k - \hat{\mu})^2$$



With m = 16 samples

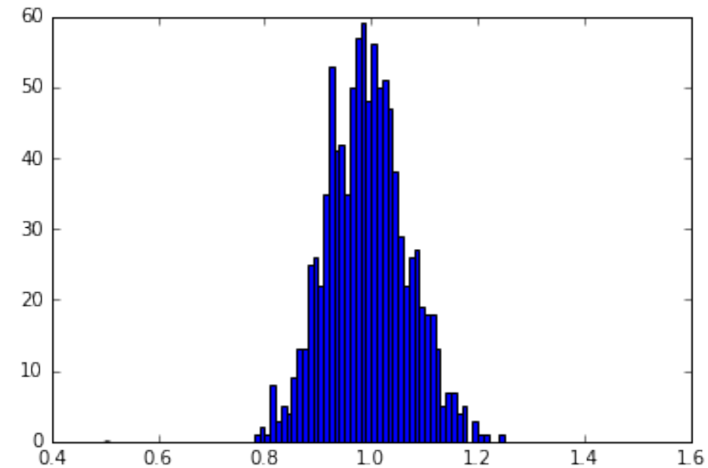
Variance as a random variable

Let's apply the same idea to the variance estimate.



For each sample we compute the estimated variance:

$$s^2 = \frac{1}{N-1} \sum_k (x_k - \hat{\mu})^2$$

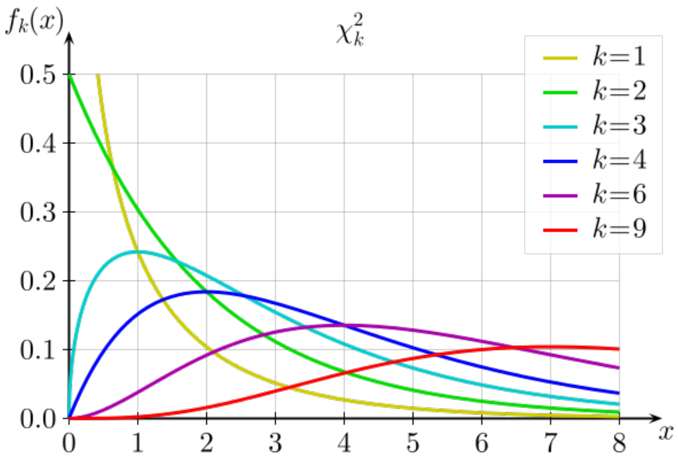


With m = 1000 samples

Variance as a random variable

Let $X_i, i = 1..N$ be a sequence of independent random variables drawn from a **normal distribution** with variance σ^2. Then as N becomes large, the distribution of the estimated variance values s^2 of each sample X_i approaches **a chi-square distribution** with N-1 degrees of freedom.

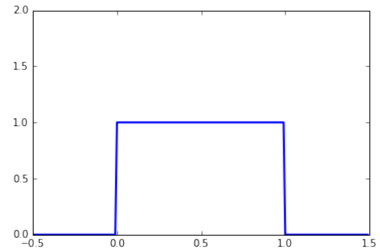
$$\frac{1}{\sigma^2}\sum_{i=1}^{N}(X_i - \overline{X})^2 = \frac{(N-1)s^2}{\sigma^2} = \chi_\nu^2$$



Generating a random variable with a given pdf

A commonly used technique is called the **Inverse transform technique**.

Let *Y* be a uniform random variable in the range [0,1].

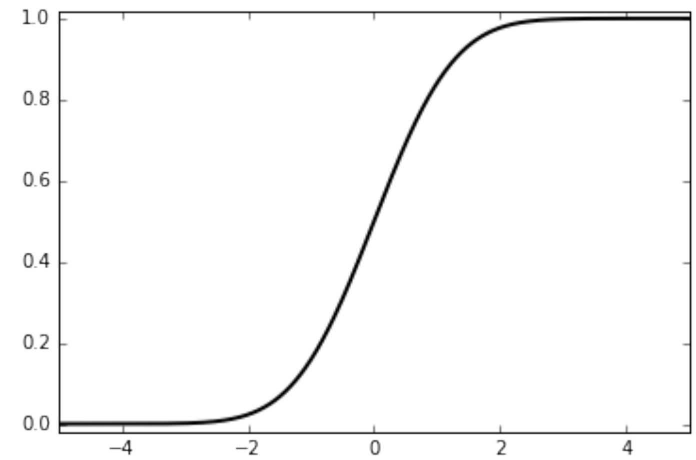


If $X = F^{-1}(Y)$, then X is a random variable with a CDF $F(X)$

Therefore if we have a random number generator to generate numbers according to the uniform distribution, we can generate any random variable with a known distribution, if we can invert the function giving the CDF of the distribution.

Generating a random variable with a given pdf

Example: The normal distribution

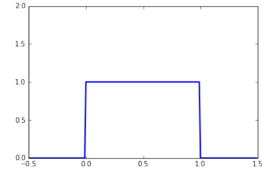


The CDF is $$F(x) = \frac{1}{2}\left[1 + \mathrm{erf}(\frac{\mathrm{x}-\mu}{\sigma\sqrt{2}})\right]$$

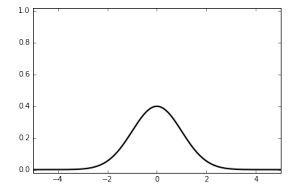
[with the error function $\mathrm{erf}(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x} e^{-t^2}\,\mathrm{d}t$]

The inverse function of the CDF is $$F^{-1}(y) = \mu + \sqrt{2}\sigma\mathrm{erf}^{-1}(2\mathrm{y}-1)$$

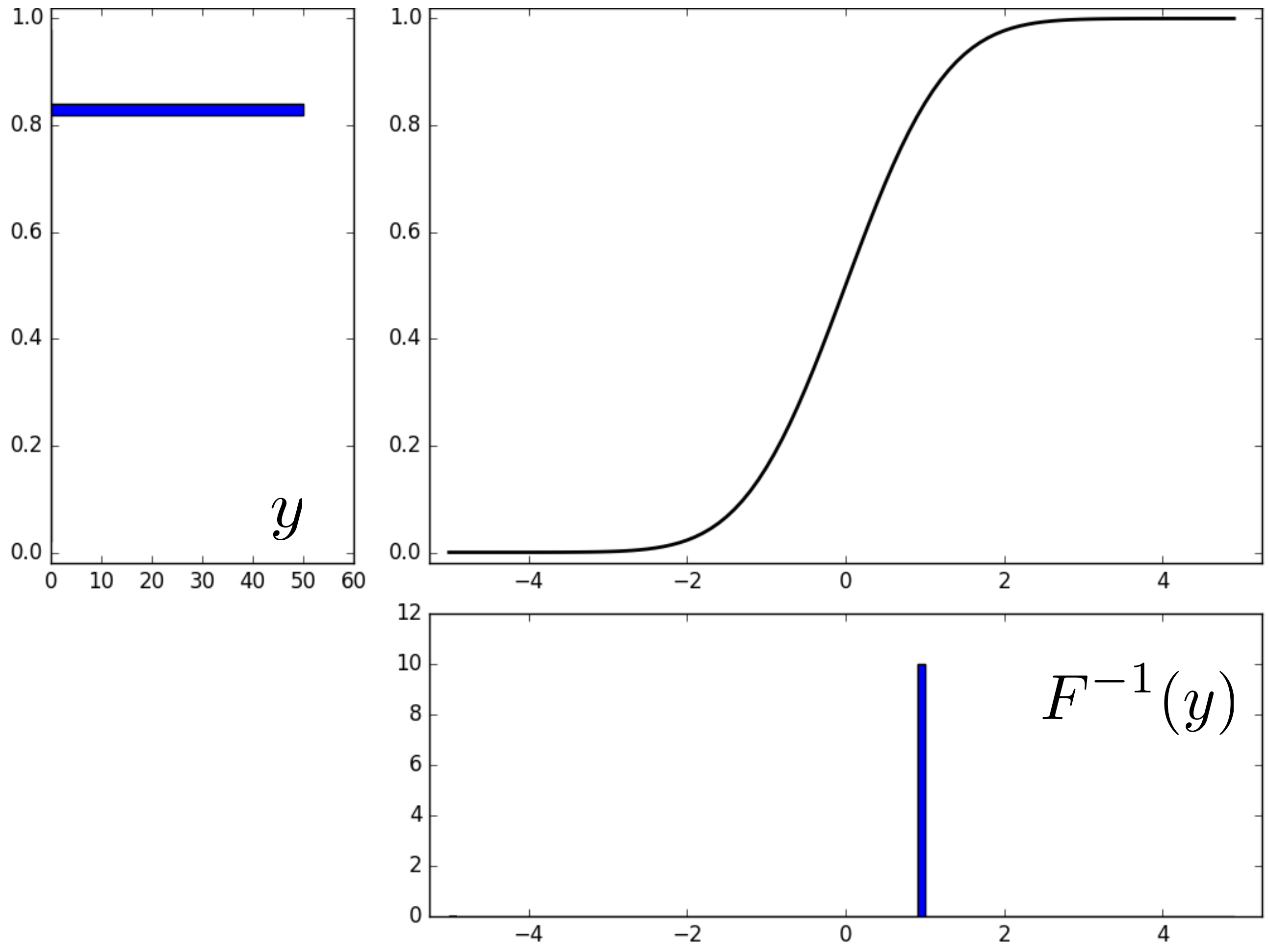
So if y is a uniform random variable in the range [0,1].



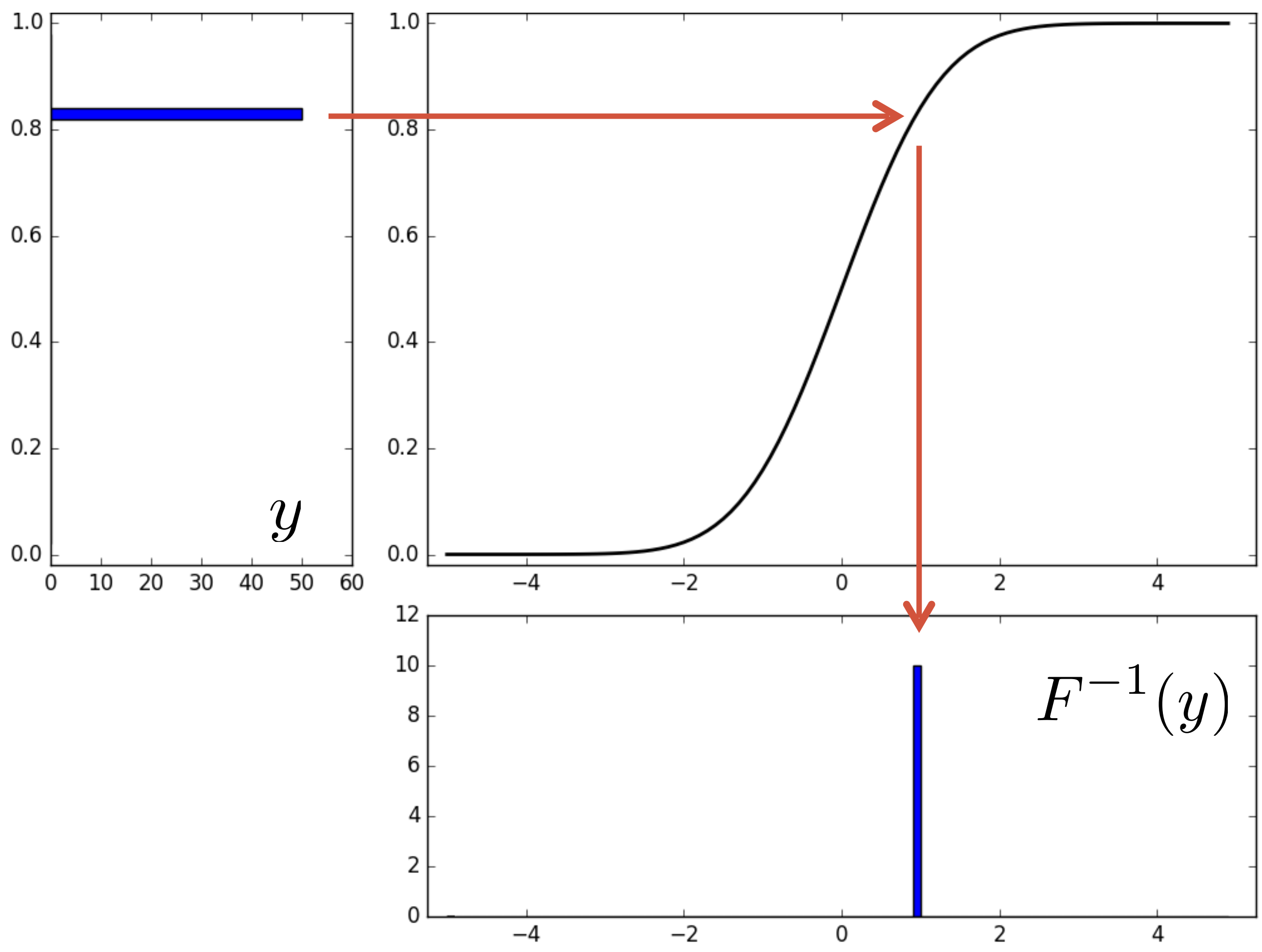
$F^{-1}(y)$ is a random variable following a normal distribution



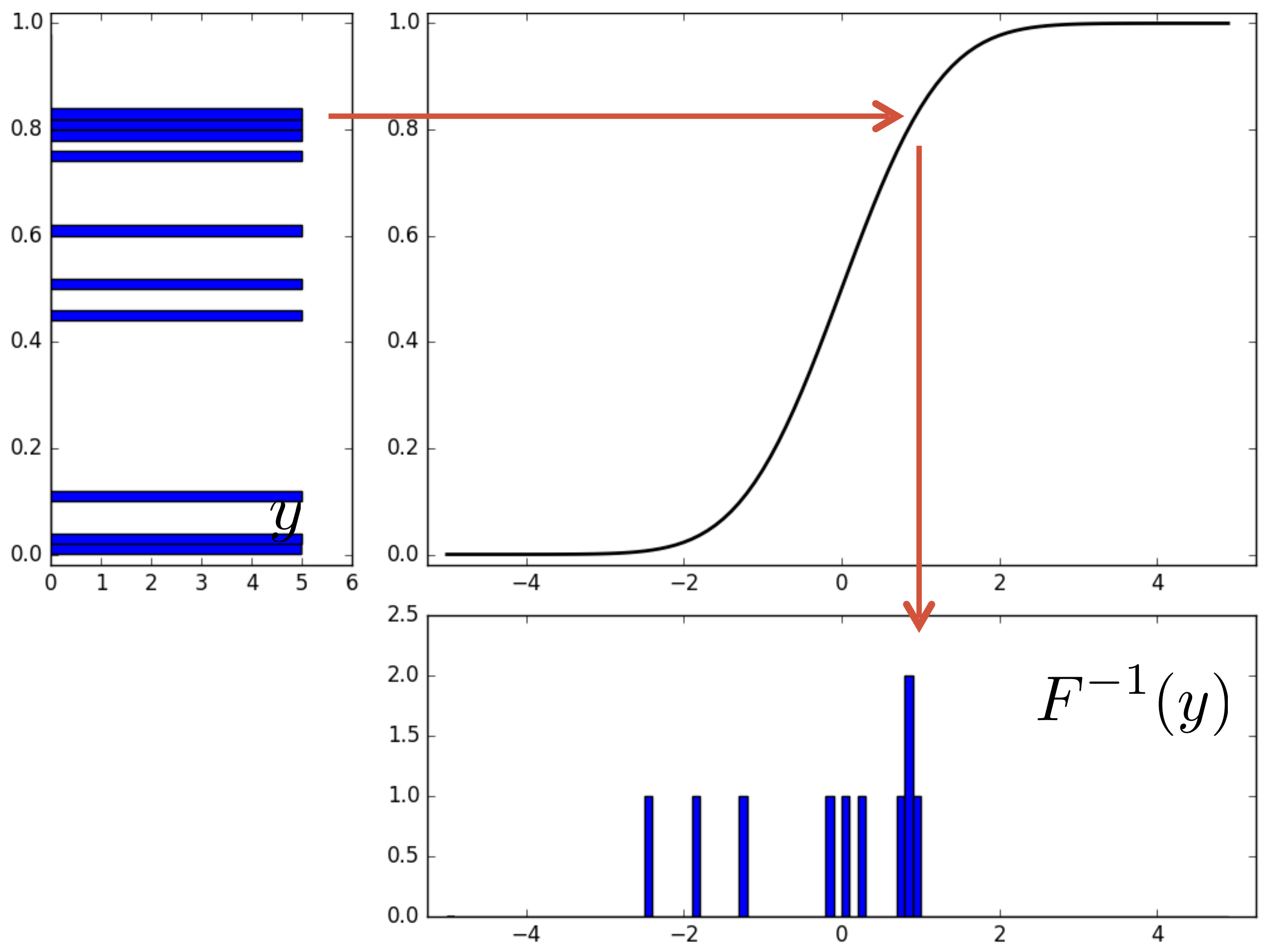
Generating a random variable with a given pdf



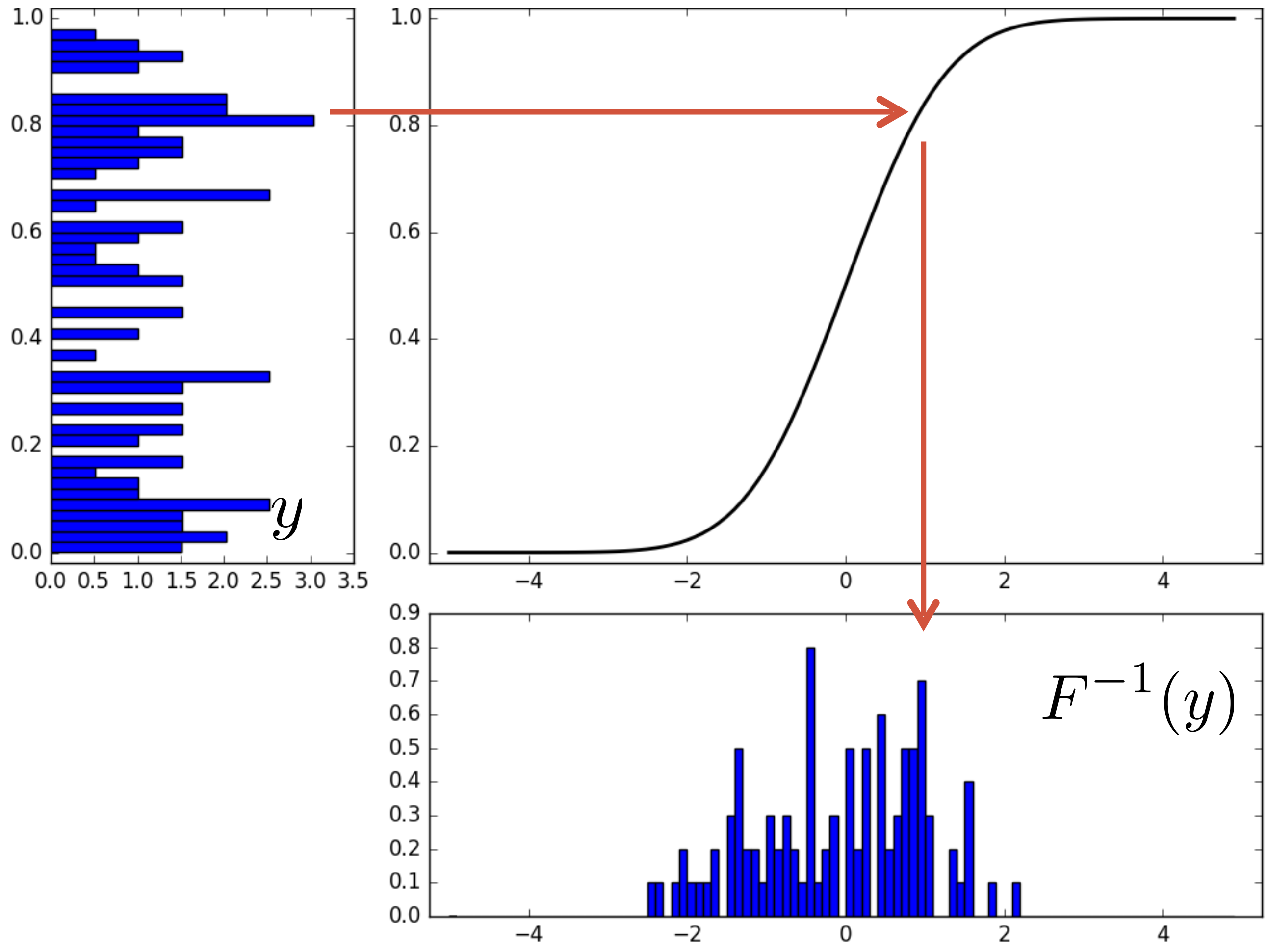
Generating a random variable with a given pdf



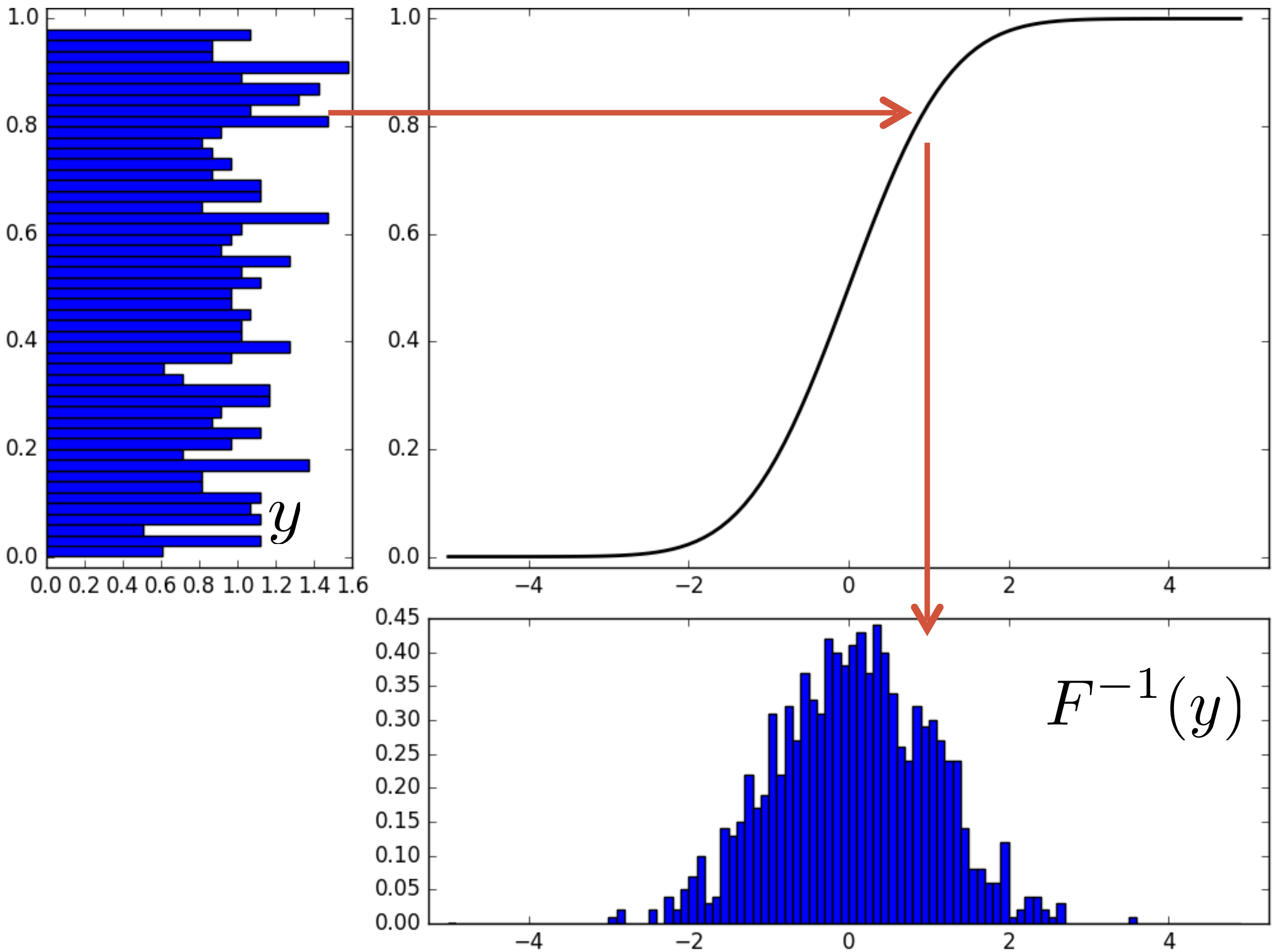
Generating a random variable with a given pdf



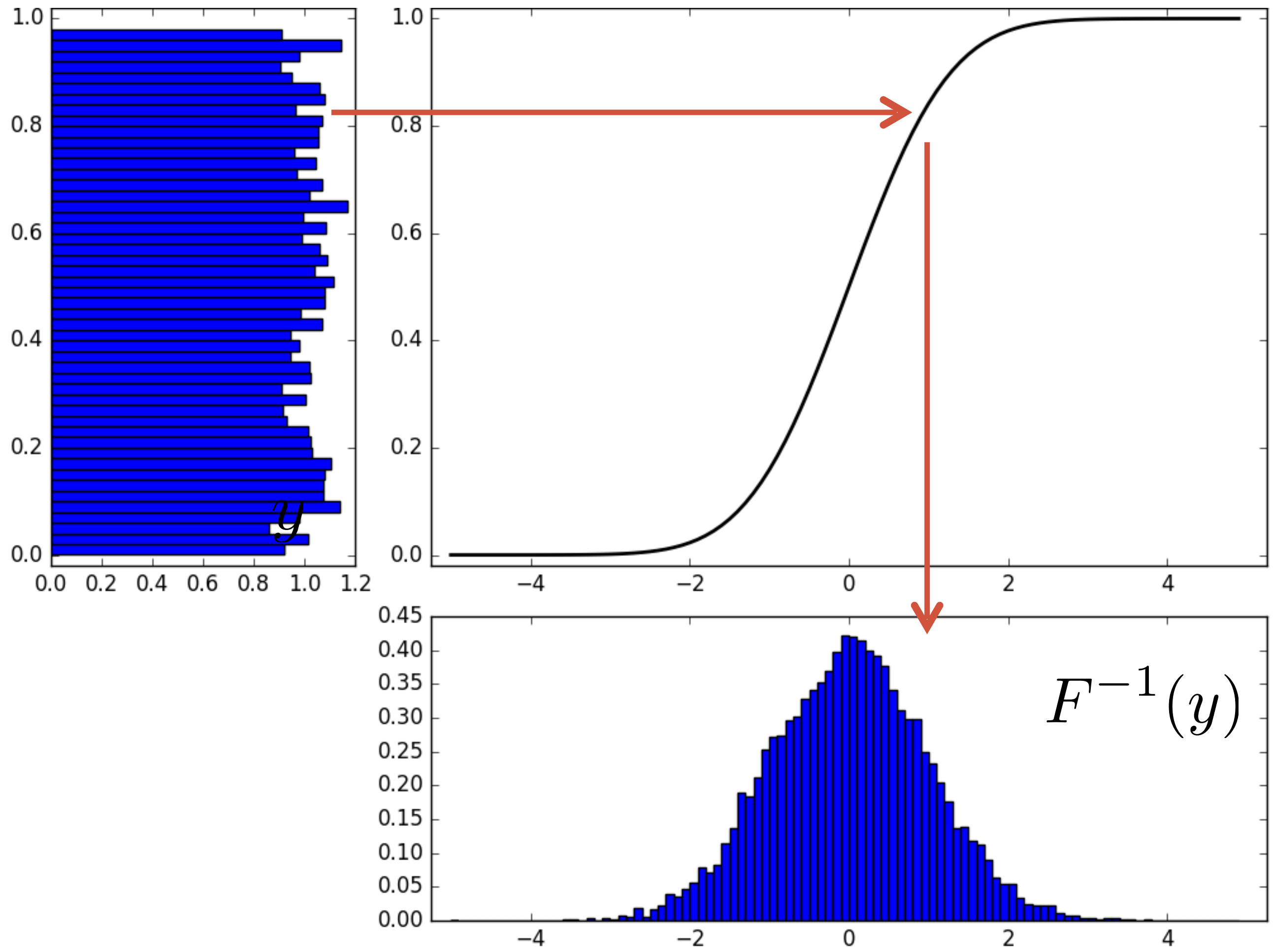
Generating a random variable with a given pdf



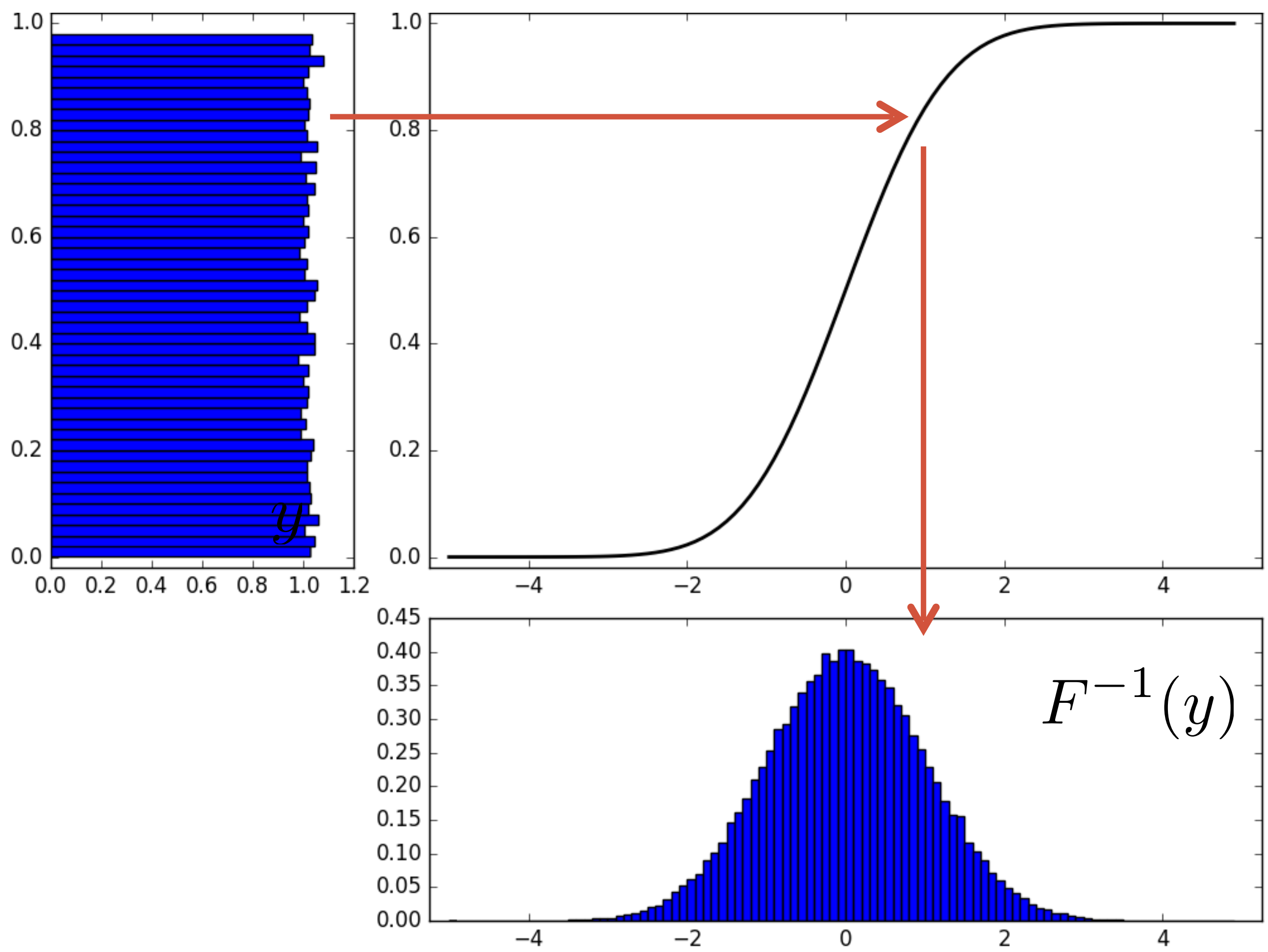
Generating a random variable with a given pdf



Generating a random variable with a given pdf



Generating a random variable with a given pdf



Moments and estimators

- See TD2 – Mean as a random variable (#1)