# Lecture 4

## Central Limit Theorem

- For any pdf $h(x)$, the mean of $N$ samples drawn from this distribution will follow a normal distribution $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{N}}\right)$ centered on the true population mean, $\mu$, & std. deviation $\sigma$.

- CAVEAT:- The distribution $h(x)$ must have a non-diverging standard deviation, i.e. the tails of the distribution must fall off faster than $1/x^2$.

- Some observations:
  - As long as the above caveat is satisfied, the CLT applies to any $h(x)$, regardless of its shape.
  - The "accuracy" of the measured mean improves by a factor of $1/\sqrt{N}$ regardless of experimental setup.
  - The CLT is also known as the "weak law of

large numbers": the sample mean converges to the population mean with large number of samples.

- The Cauchy distribution is an example of an $h(x)$ that would not work with the CLT since it does not have a well defined mean or standard deviation.
  - Empirically, if we measure the mean of $N$ samples drawn from a Cauchy dist., we would not get the $1/\sqrt{N}$ improvement predicted by the CLT.

- Note that while the CLT makes this general prediction for the sample mean, it does not imply that this method of calculating the sample mean is the most "efficient."
  - For example, it can be shown that the population mean for a uniform distribution modeled as

$$\tilde{\mu} = \frac{\min(x_i) + \max(x_i)}{2}$$

converges to the population mean as $1/N$, instead of the $1/\sqrt{N}$ efficiency offered by the CLT (see textbook for more discussion).

## Multi-variate distributions

- Often, pdfs are multi-dimensional rather than 1-D as we have seen so far.

- Each variable in the $M$-dimensional space will have its own mean & std. dev as usual:

g. $\mu_x = \int_{-\infty} \int_{-\infty} x\, h(x,y)\, dx\, dy$

and $$V_x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_x)^2\, h(x,y)\, dx\, dy$$

(and same for $\mu_y$ & $V_y$).

• In addition, there will also be the "covariance" between the $M$ parameters:

e.g.
$$V_{xy} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_x)(y-\mu_y)\, h(x,y)\, dx\, dy$$

→ This quantifies how much these variables depend on each other.

→ Also, $V_{xy} \equiv \mathrm{Cov}(x,y)$

• Note that $\sigma_x = \sqrt{V_x}$, $\sigma_y = \sqrt{V_y}$, but

$$\boxed{\sigma_{xy} = V_{xy}}.$$

– A related useful result is if $z = x+y$, then

$$V_z = V_x + V_y + 2V_{xy}$$

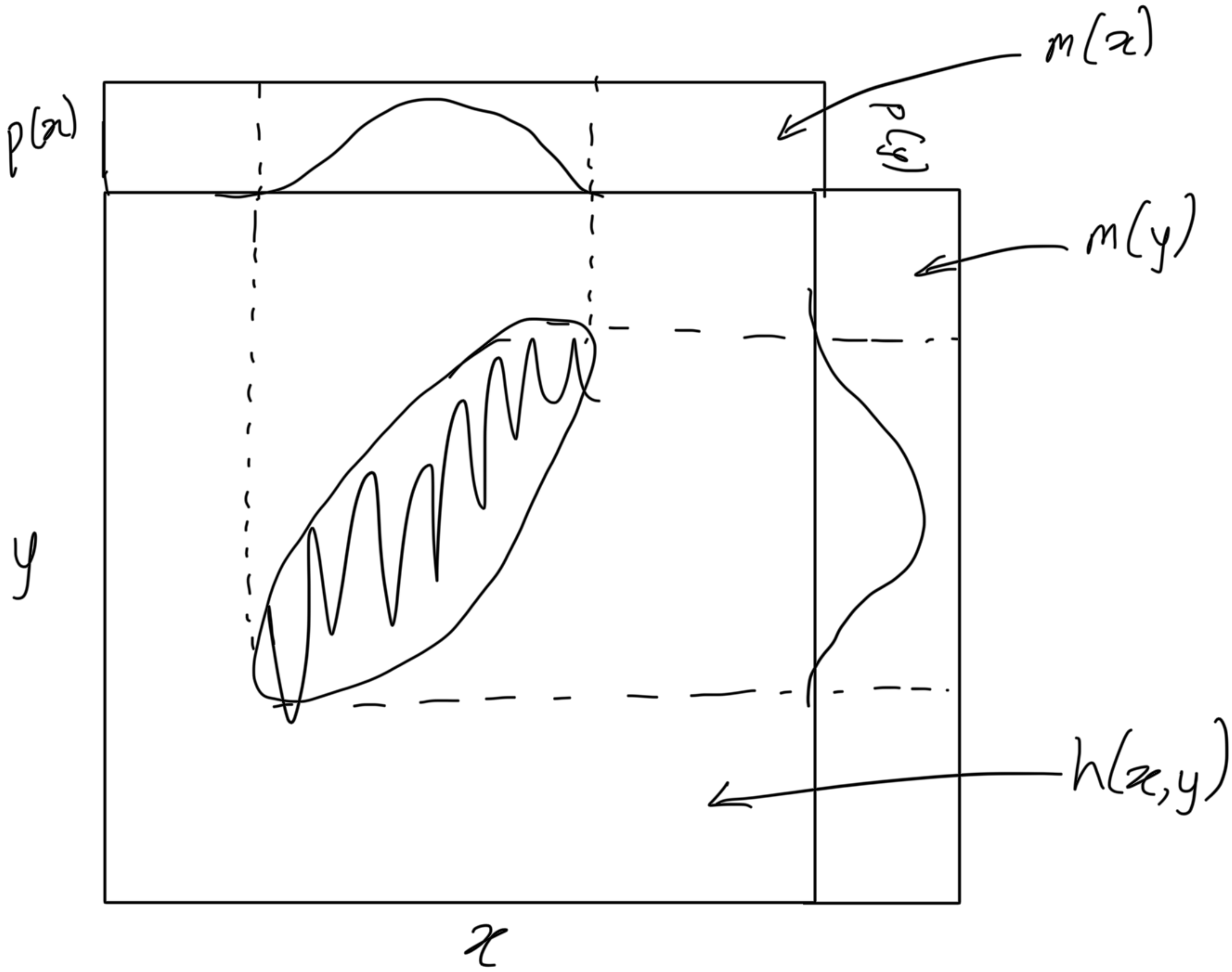
and $w = x-y$,

$$V_w = V_x + V_y - 2V_{xy}$$

– When two variables are "uncorrelated", $\boxed{V_{xy} = 0}$.

- For multi-variate distribution, the "marginal" distribution for one of the variables is,

$$\boxed{m(x) = \int_{-\infty}^{\infty} h(x,y)\, dy}$$




- Multi-variate Gaussian distribution:
- The 1-D Gaussian can be expanded to M-dimensions

$$\boxed{p(\vec{x}|I) = \frac{1}{(2\pi)^{M/2}\sqrt{\det(\overline{C})}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \overline{C}^{-1} (\vec{x}-\vec{\mu})\right)}$$

where,

- The $\vec{x}$ is a column vector & $\vec{x}^T$ is its transposed row vector.

• $C$ is an $M \times M$ symmetric "covariance" matrix, and $\vec{C}^{-1}$ has positive eigenvalues.

$$C_{kj} = \int_{-\infty}^{\infty} x^k x^j \, p(\vec{x}|I) \, d^M x$$

• The argument in the exponent can be explicitly written as,

$$(\vec{x} - \vec{\mu})^T \vec{C}^{-1} (\vec{x} - \vec{\mu}) = \sum_{k=1}^{M} \sum_{j=1}^{M} C^{-1}_{kj} (x^k - \mu^k)(x^j - \mu^j)$$

## Correlation coefficients

- Knowing the covariance between two parameters, $\sigma_{xy}$, & their individual variance, $\sigma_x$ & $\sigma_y$, the correlation coefficient can be written as,

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- Parametric correlation tests:

• When working with samples from two datasets of equal size $N$, $\{x_i\}$ and $\{y_i\}$, the Pearson correlation coefficient is,

$$r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{N} (x_i - \bar{x})^2\right]^{1/2} \left[\sum_{i=1}^{N} (y_i - \bar{y})^2\right]^{1/2}}$$

with $-1 \leq r \leq 1$ & $-1 \rightarrow$ anti-correlated

$1 \rightarrow$ correlated

$0 \rightarrow$ uncorrelated.

- However, this test is susceptible to outliers, and does not offer a way to deal with uncertainties on data.

- Non-parametric tests:

(1) Spearman correlation coefficient:

- Rather than using the samples directly, you sort the samples and assign each sample a "rank", $R_i^x$.
- These ranks are then used in the Pearson formula above.

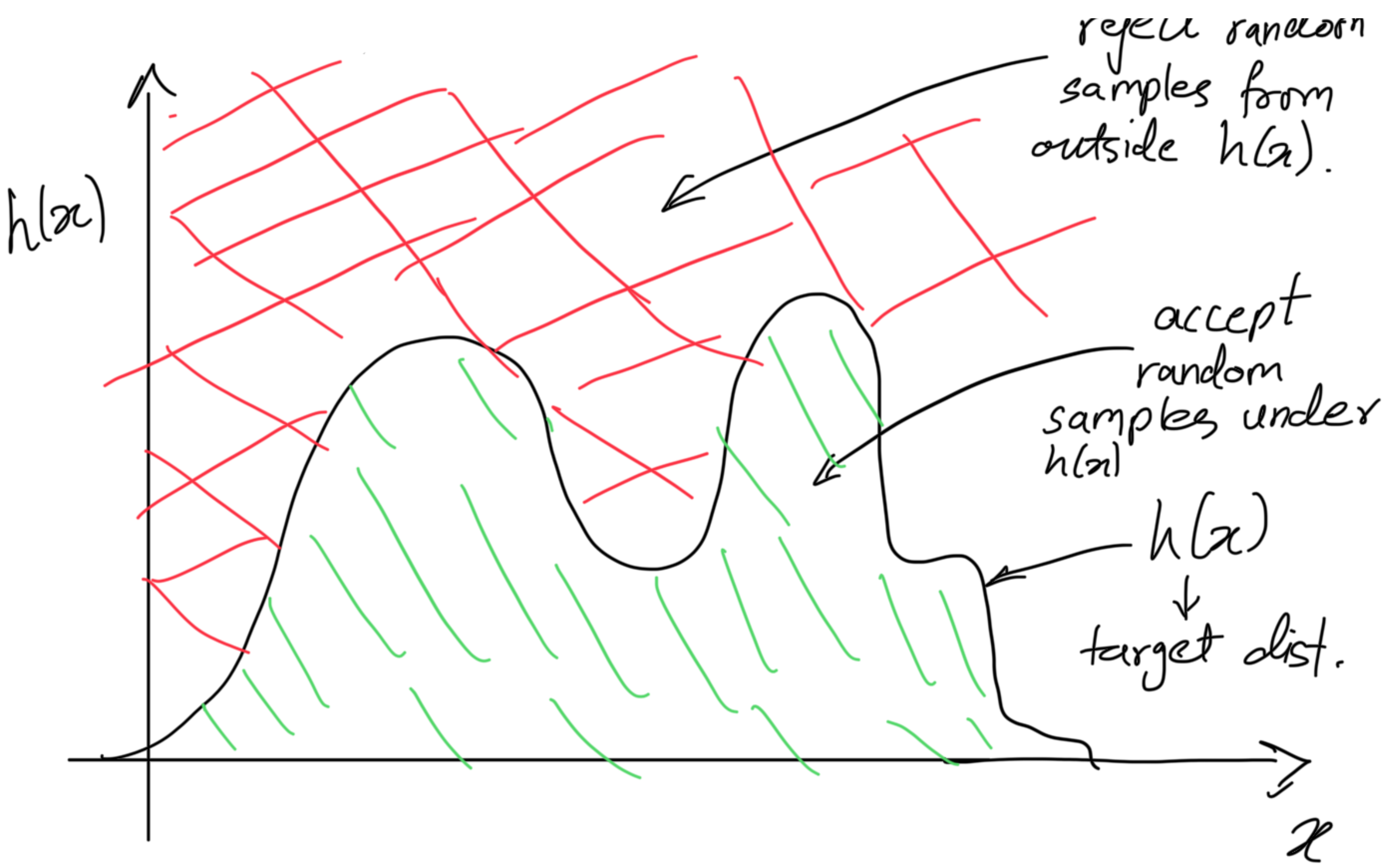
(2) Kendall correlation coefficient:

- With ranks assigned, calculate the number of "concordant" $[(x_j - x_k)(y_j - y_k) > 0]$ and "discordant" $[(x_j - x_k)(y_j - y_k) < 0]$ pairs.

* - Note that in addition to calculating the absolute value of this correlation statistic, we also need to determine how "significant" the measurement is by comparing it to a "null" distribution, which in this case would be the case of uncorrelated variables.

## Random number generation

- One simple way is called "rejection sampling".



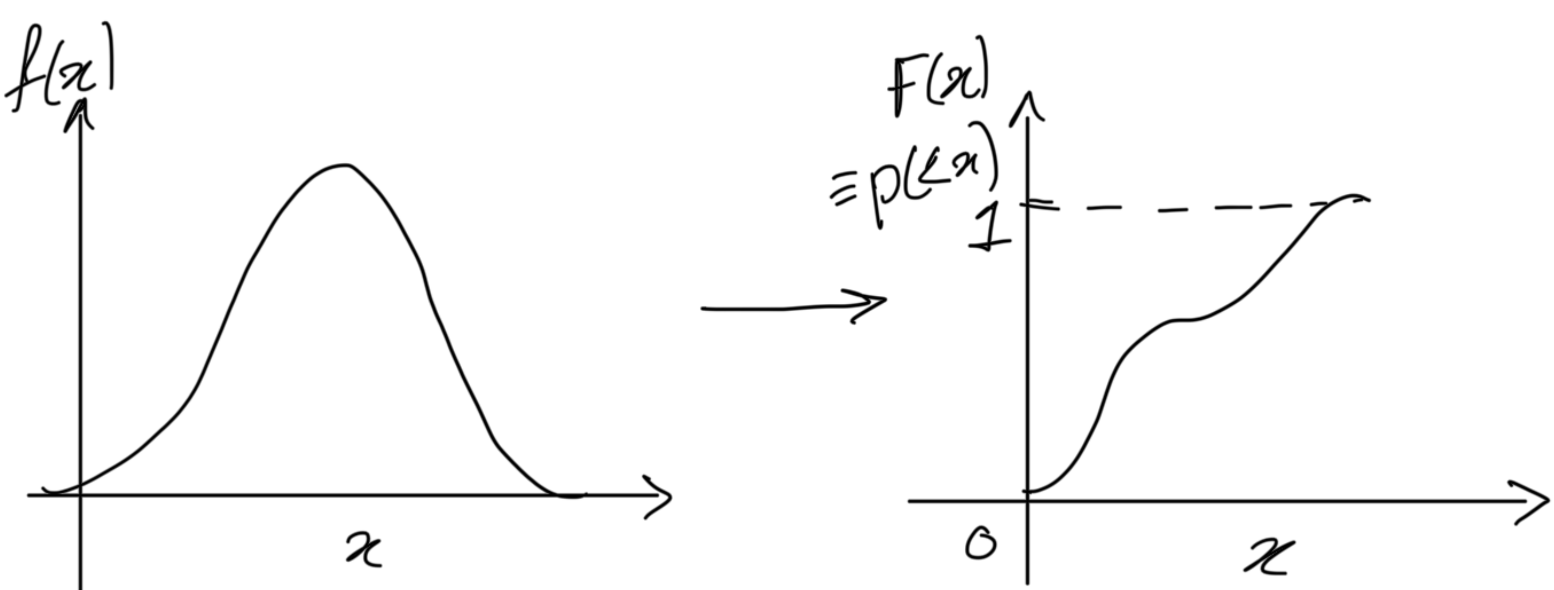

- A more efficient way to draw samples from a distribution by the "transformation method":

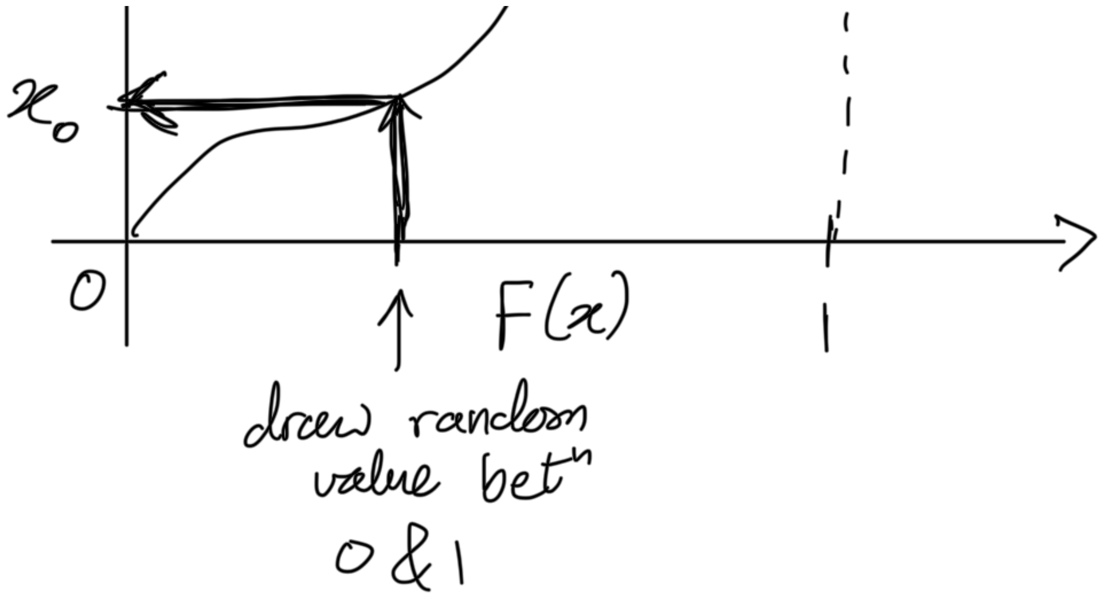
• Given a pdf, $f(x)$, → calculate CDF $F(x)$

↓

Draw a random number from the range $0 \le y \le 1$

→ find $x$ by inverting $F(x) = y$



$x_0$
0
$F(x)$
1
draw random value betn 0 & 1