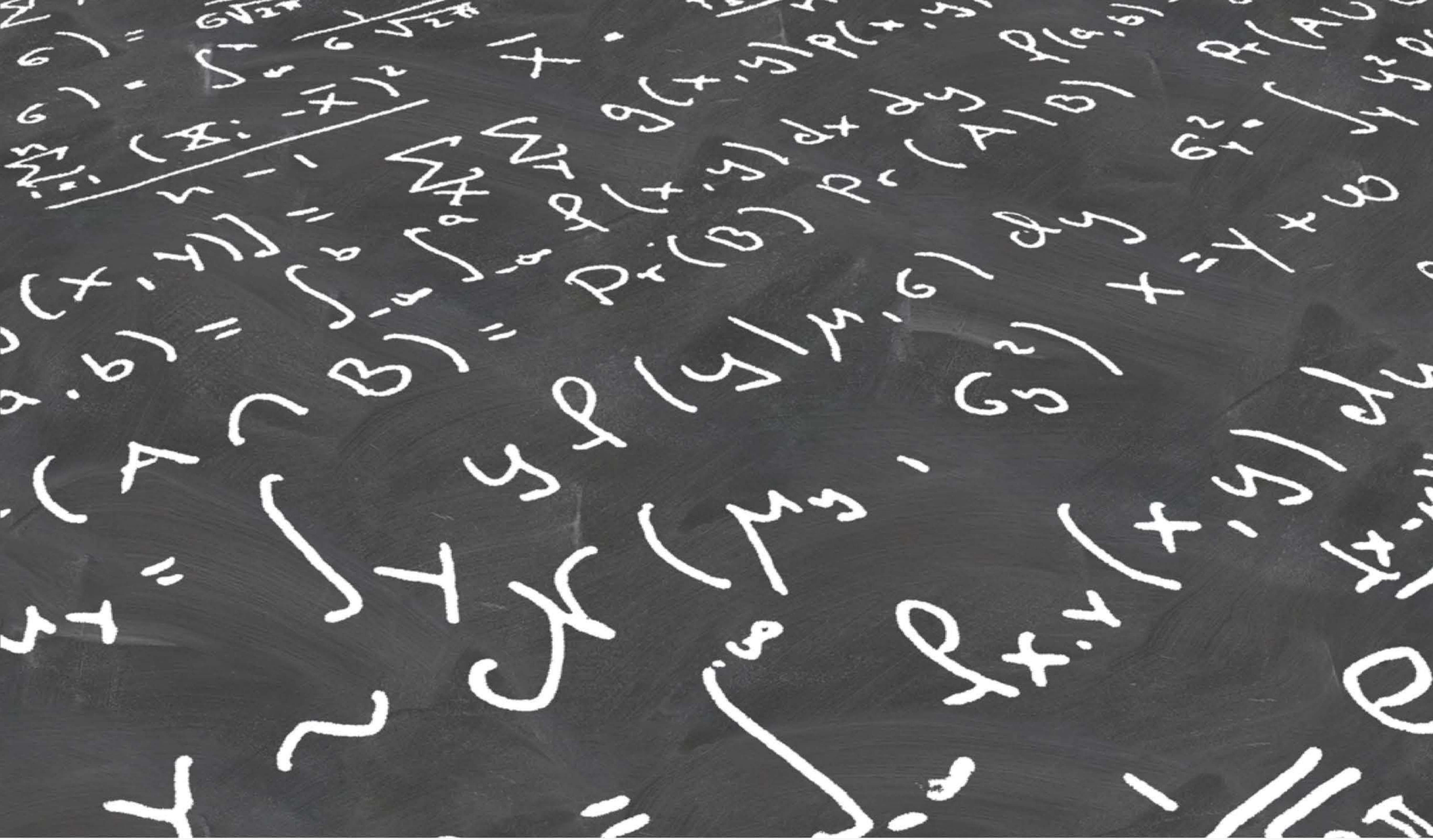


5. Transformations and Multivariate Random Variables

Gunvor Elisabeth Kirkelund
Lars Mandrup

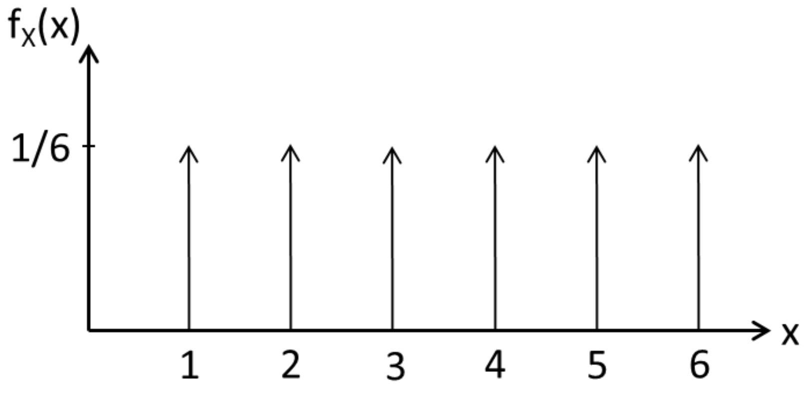
Agenda for Today

- Repetition:
  - One Random Variable
  - Two Random Variables
- Data sampling for test and simulation
- Transformation of random variables
- Sum of two random variables
- Central limit theorem (CLT)

One Stochastic Variable – Discrete

- Probability mass function (pmf):

$$f_X(x) = \begin{cases} Pr(X = x_i) & for\ X = x_i \\ 0 & otherwise \end{cases}$$

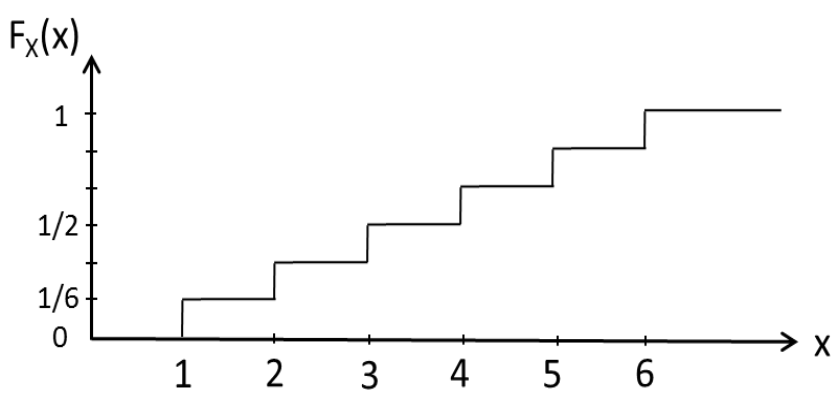


$$0 \leq f_X(x) \leq 1$$

$$\sum_{i=1}^{n} f_X(x_i) = \sum_{i=1}^{n} Pr(X = x_i) = 1$$

- Cumulative distribution function (cdf):

$$F_X(x) = P\,r(X \leq x) = \sum_{i=1}^{n_x} f_X(x_i)$$



$$0 \leq F_X(x) \leq 1$$

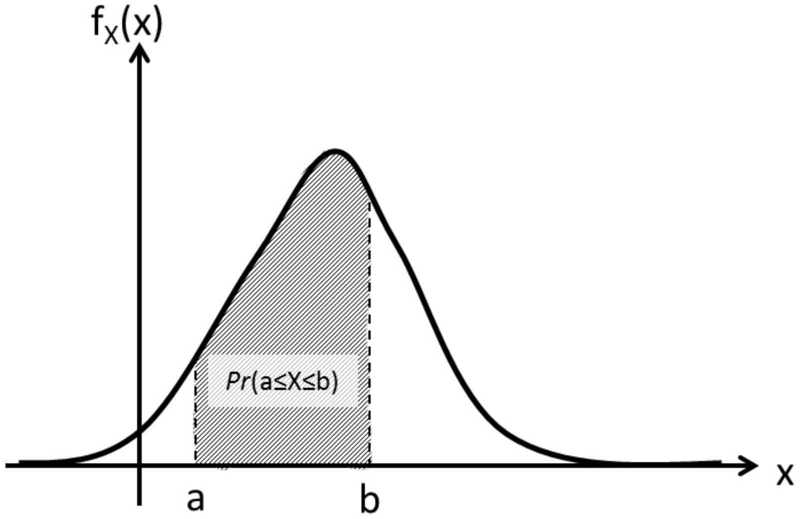
$$\lim_{x \to -\infty} F_X(x) = 0$$

$$\lim_{x \to \infty} F_X(x) = 1$$

One Stochastic Variable – Continuous

- Probability density function (pdf):

$$Pr(a \leq X \leq b) = \int_a^b f_X(x)\, dx$$

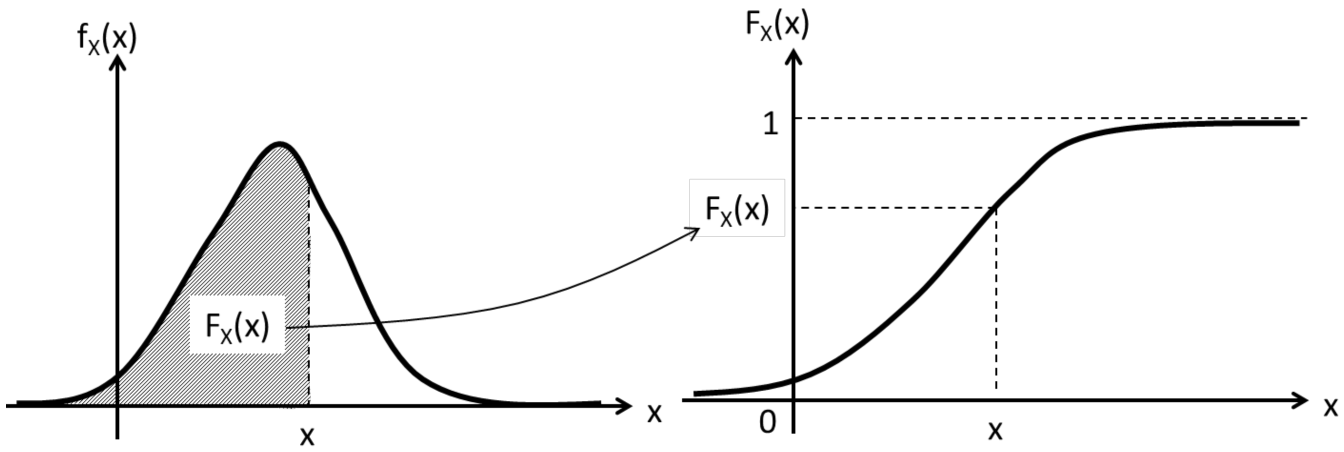


$$f_X(x) \geq 0$$

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

- Cumulative distribution function (cdf):

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du = Pr(X \leq x)$$



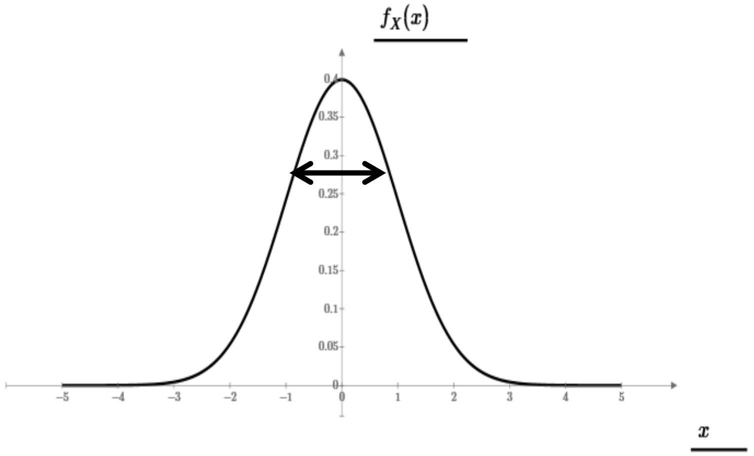
$$0 \leq F_X(x) \leq 1$$

$$\lim_{x \to -\infty} F_X(x) = 0$$

$$\lim_{x \to \infty} F_X(x) = 1$$

Expectations

- Mean value: $E[X] = \bar{X} = \mu_X = \int_{-\infty}^{\infty} x \cdot f_X(x)dx \quad (\sum_{i=1}^{n} x_i\, f_X(x_i))$
- Variance: $Var(\mathrm{X}) = {\sigma_X}^2 = \int_{-\infty}^{\infty} (x - \bar{x})^2 \cdot f_X(x)dx = E\left[X^2\right] - E[X]^2$



- Standard deviation: $\sigma_X = \sqrt{Var(\mathrm{X})}$
- A function: $E[g(X)] = \overline{g(X)} = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)dx \quad (\sum_{i=1}^{n} g(x_i)\, f_X(x_i))$

$$Var\big(g(\mathrm{X})\big) = \int_{-\infty}^{\infty} \left(g(x) - \overline{g(x)}\right)^2 \cdot f_X(x)dx = E\left[g(X)^2\right] - E[g(X)]^2$$

- Linear function: $E[aX + b\ ] = a \cdot E[X] + b$

$$Var[aX + b\ ] = a^2\left(E[X^2] - E[X]^2\right) = a^2 \cdot Var(X)$$

Two Random Variables X, Y

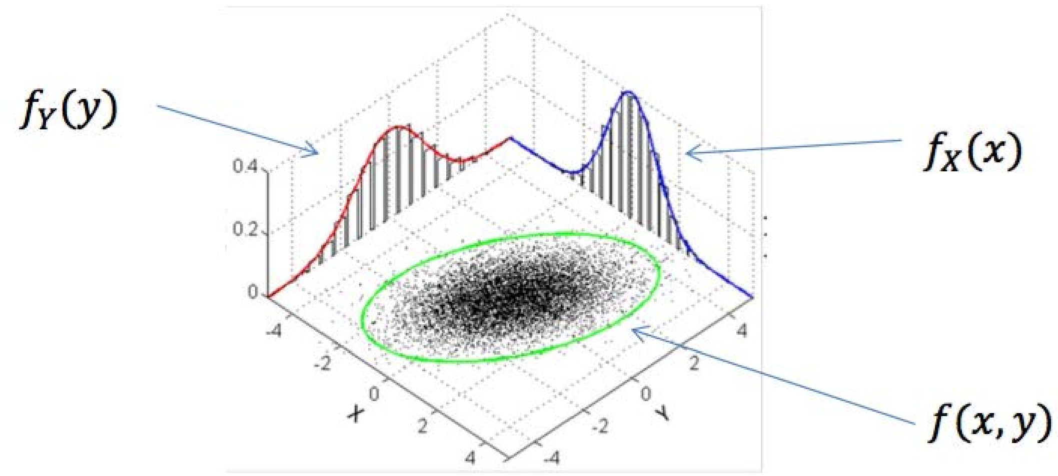
Joint (Simultaneous) pdf: $f_{X,Y}(x,y) \geq 0 \qquad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$

$$Pr\big((a \leq X \leq b) \cap (c \leq Y \leq d)\big) = \int_{c}^{d}\int_{a}^{b} f_{X,Y}(x,y)dxdy$$

Marginals:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx$$



Cumulative Distribution Function cdf:

cdf $\quad F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(x,y)dxdy = Pr(X \leq x \wedge Y \leq y)$

pdf $\quad f_{X,Y}(x,y) = \dfrac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$

Bayes Rule, Conditional PDF and Independence

Bayes rule:

- The joint/simultaneous pmf/pdf for two stochastic variables:

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x)$$

Conditional pdf:

- For a two dimensional pmf/pdf $f_{X,Y}(x,y)$, we can find the conditional pdf with Bayes rule:

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

Independence:

- X and Y are independent if and only if:

$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$ or $f_{X|Y}(x|y) = f_X(x)$ for all x and y

Correlation and Covariance

Correlation tells of the (biased) coupling between variables

- Correlation: $corr(X,Y) = E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot y \cdot f_{X,Y}(x,y)dxdy$

Covariance is without bias from the mean

- Covariance: $cov(X,Y) = E\big[(X-\bar{X})(Y-\bar{Y})\big] = E[XY] - E[X]\cdot E[Y]$

Correlation Coefficient is the normalized Covariance

- Correlation coefficient: $\rho = E\left[\frac{X-\bar{X}}{\sigma_X}\cdot\frac{Y-\bar{Y}}{\sigma_Y}\right] = \frac{E[XY]-E[X]E[Y]}{\sigma_X\cdot\sigma_Y}$

$$-1 \le \rho \le 1$$

- If X and Y are independent:

$E[XY] = E[X]\cdot E[Y]$ and $cov(X,Y) = \rho = 0$

Very important!

i.i.d.: Independent and Identically distributed

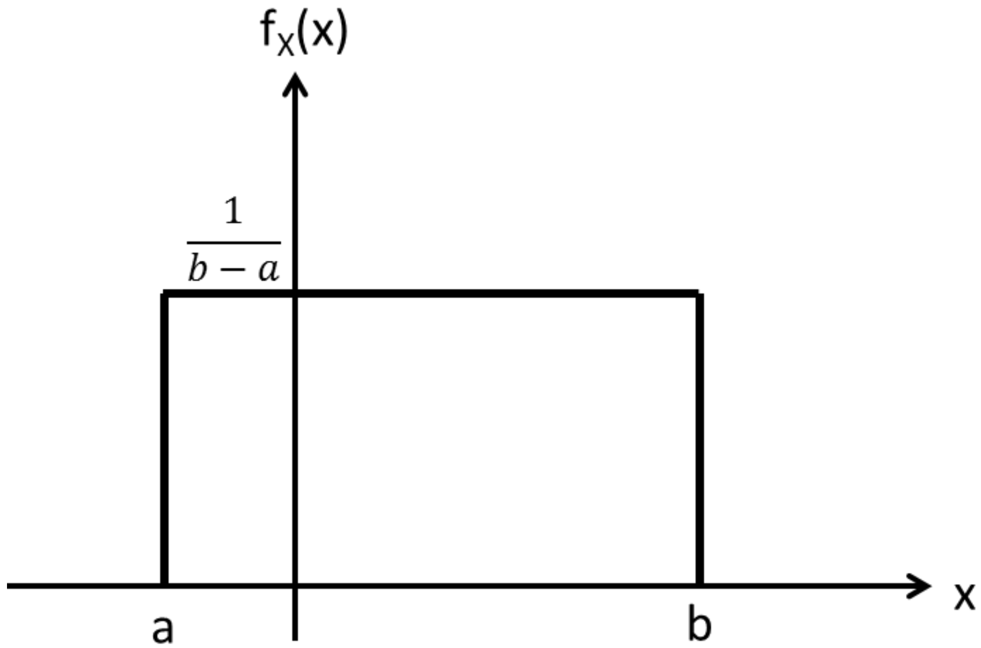
- We define that for series of random variables that is taken from the same distribution (identically distributed), and are sampled independent of each other, that they are i.i.d.

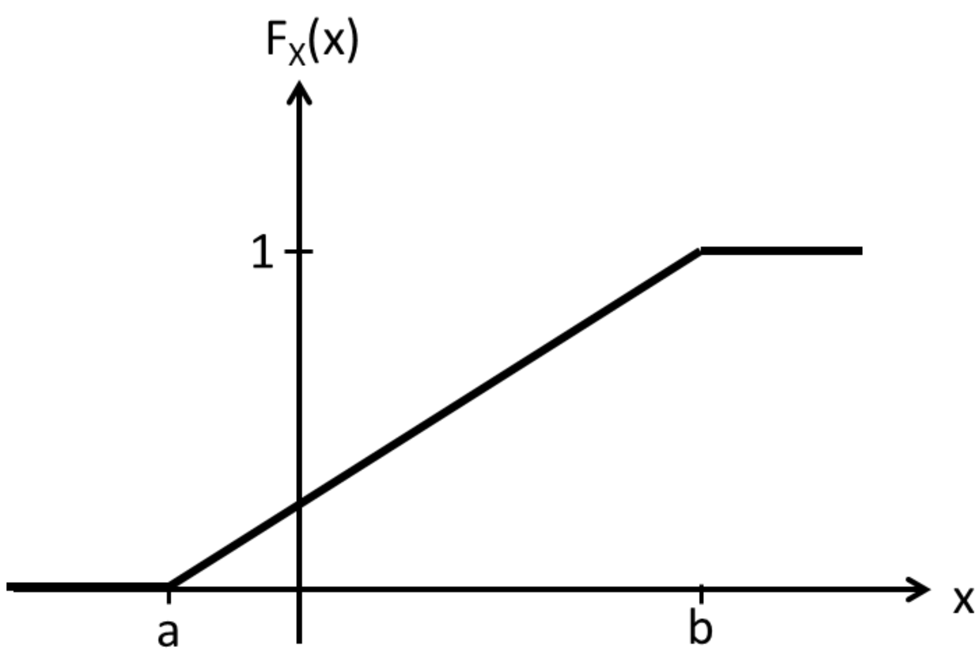
$$\boxed{\text{i.i.d. = Independent and Identically distributed}}$$

- i.i.d. is a very important characteristic in stochastic variable processing and statistics

Uniform Distribution

- $\mathcal{U}$(a,b) (Matlab: *rand*)
- Mean value: $\mu = \frac{a+b}{2}$
- Variance: $\sigma^2 = \frac{1}{12}(b-a)^2$
- pdf: $f_X(x) = \begin{cases} \frac{1}{b-a} & for\ a \leq x \leq b \\ 0 & otherwise \end{cases}$
- cdf: $F_X(x) = \begin{cases} 0 & for\ x \leq a \\ \frac{x-a}{b-a} & for\ a \leq x \leq b \\ 1 & for\ x \geq b \end{cases}$





Gaussian Distribution = Normal Distribution

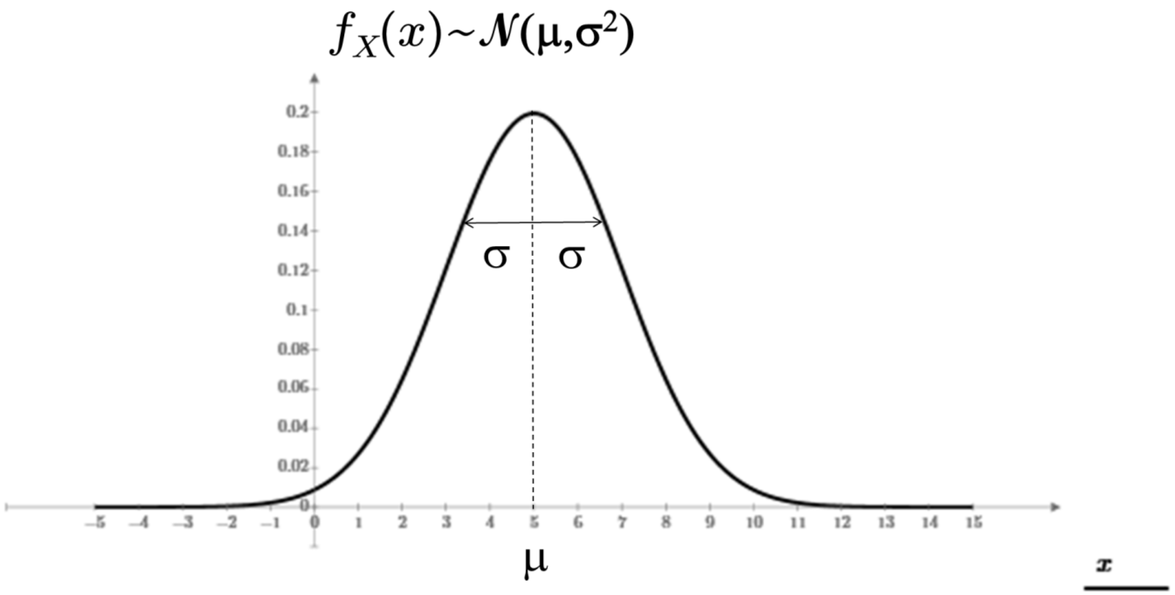
- $\mathcal{N}(\mu,\sigma^2)$
- Mean value: μ
- Variance: σ^2

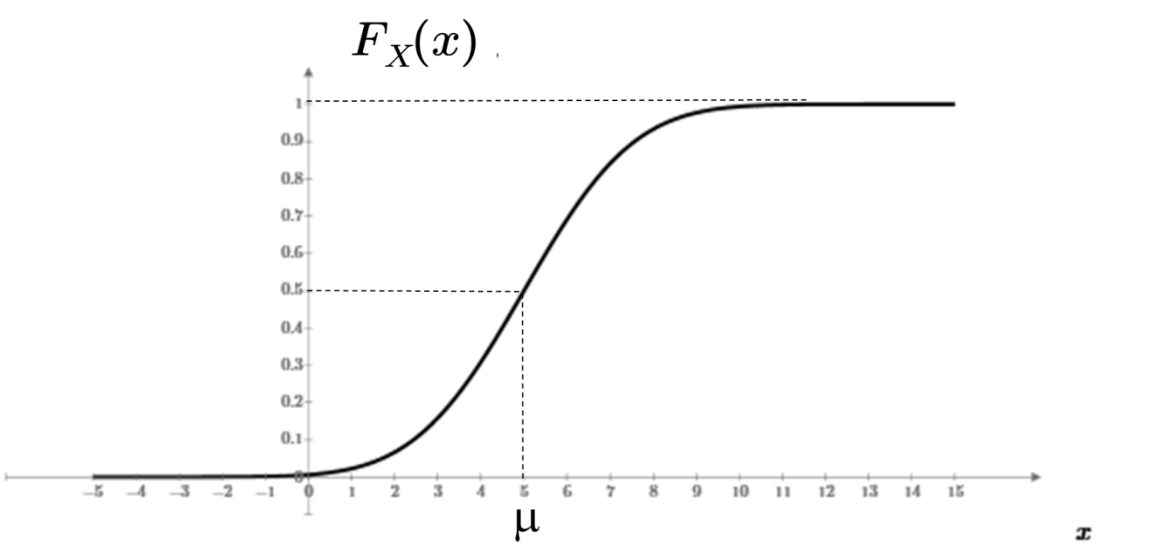
- pdf: $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- cdf: $F_X(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$

No closed expression for the cdf

erf = error-function: $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$





Gaussian Distribution = Normal Distribution

- Beregninger med normalfordelinger: Tabelopslag og Matlab:
- $X \sim \mathcal{N}(\mu, \sigma^2) \;\rightarrow\; Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$ (Standard Normal Distribution)
- $F_X(x) = Pr(X \leq x) = Pr\left(Z \leq \frac{x-\mu}{\sigma}\right) = F_Z(z) = \Phi(z)$ hvor $\mathrm{z} = \frac{x-\mu}{\sigma}$

$Tabel\ 1$ ("$Statistik\ og\ Sandsynlighedsregning$")

- $\Phi(z) = Pr(Z \leq z)$
- $\Phi(-z) = 1 - \Phi(z) = Pr(Z \geq z) = Pr(Z \leq -z)$

Symmetry of Gaussian distribution

- Matlab:

  Obs: Standard deviation

  - $\Pr(X \leq x) = F_X(x) = normcdf(x, \mu, \sigma)$
  - $\Pr(Z \leq z) = F_Z(z) = normcdf(z, 0, 1) = normcdf(z)$

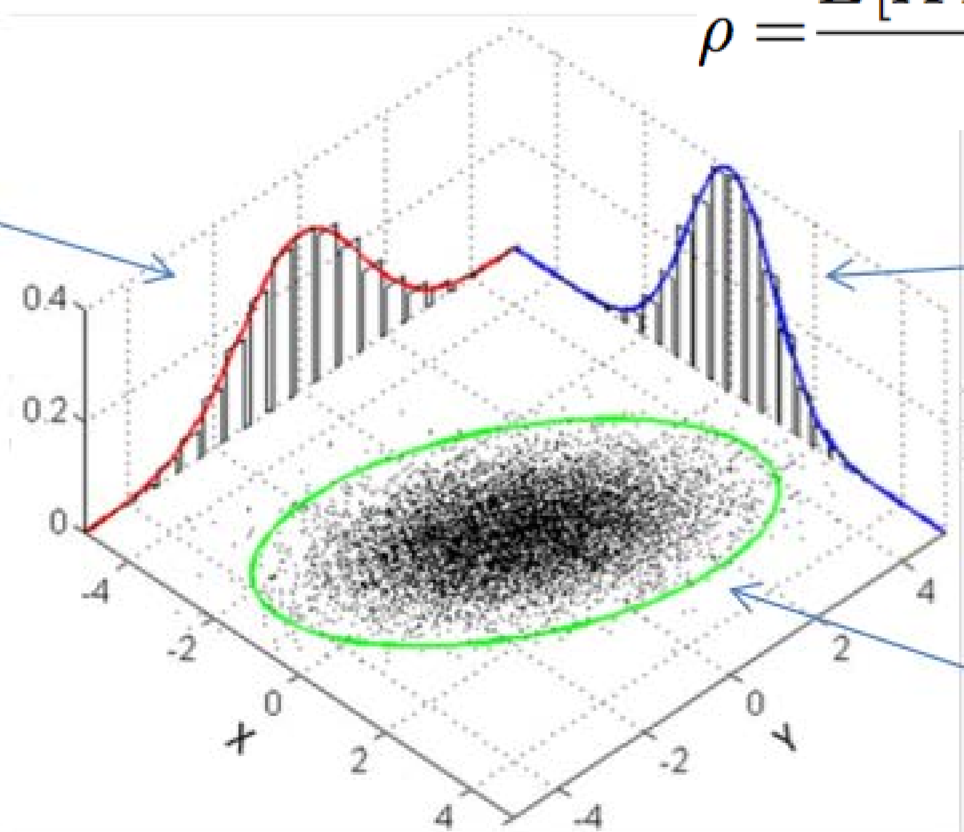
Bivariate (2D) Normal Distribution

Two dimensional Gaussian
$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{z}{2(1-\rho^2)}\right)$$

$$z = \frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}$$

$$\rho = \frac{E[XY]-E[X]E[Y]}{\sigma_X\sigma_Y}$$
Correlation coefficient

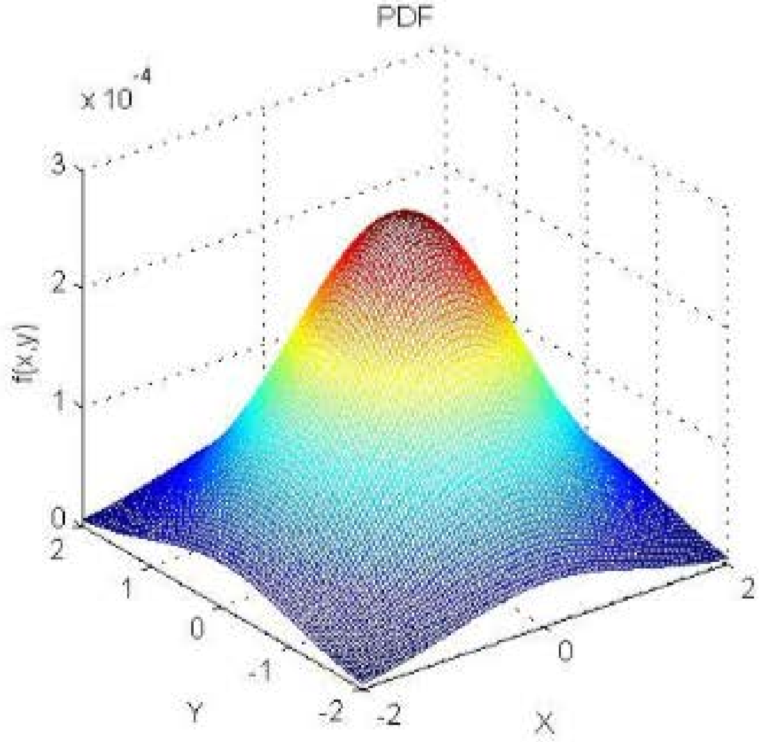
$f_Y(y)$



$f_X(x)$

$f(x,y)$

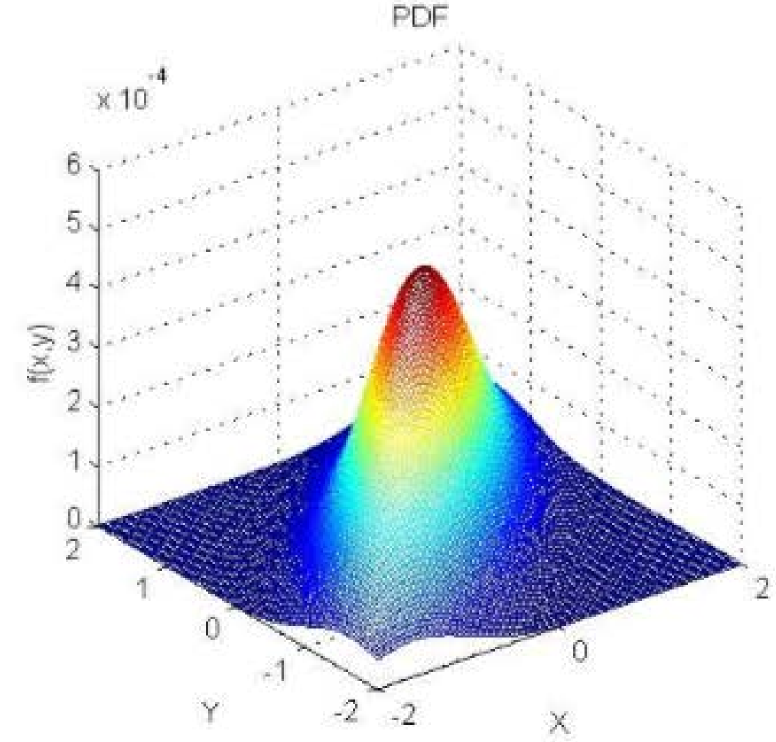
Bivariate Normal Distribution



Symmetric PDF:

$\rho = 0$

X and Y independent



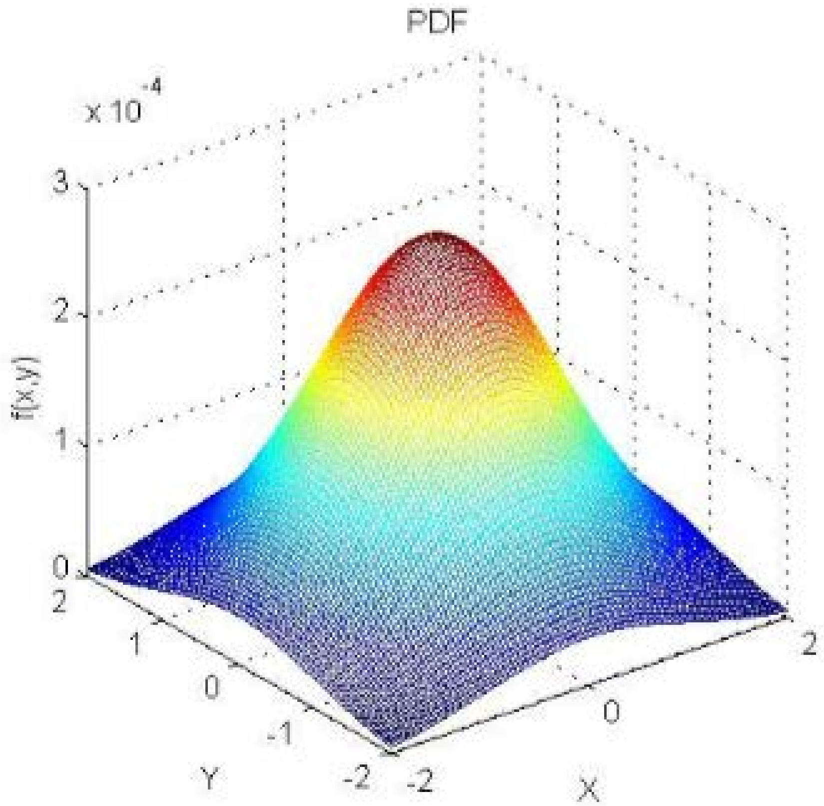
Asymmetric PDF:

$\rho = 0.8$

X and Y **dependent**

Symmetric Case

Bivariate Normal Distribution



Symmetric PDF:
$\rho = 0$

X and Y independent

Because of the independence, we should have

$$f(x|y) = f_X(x)$$

$$f(y|x) = f_Y(y)$$

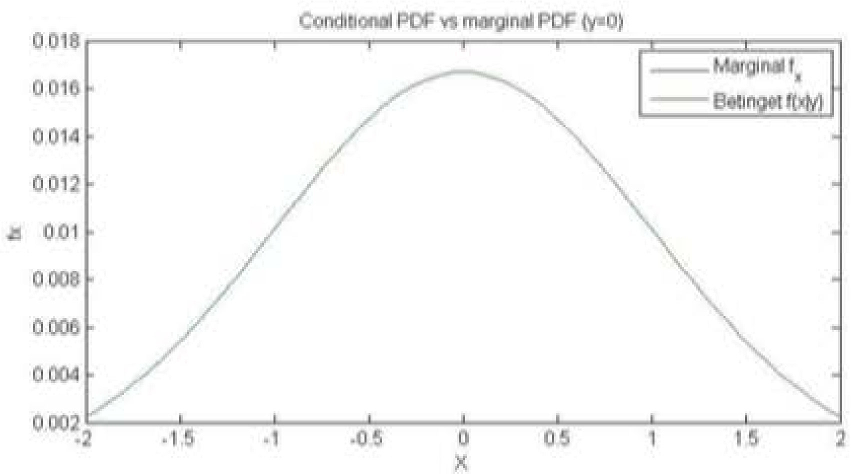
$$f(x, y) = f_X(x) \cdot f_Y(y)$$

Symmetric Case

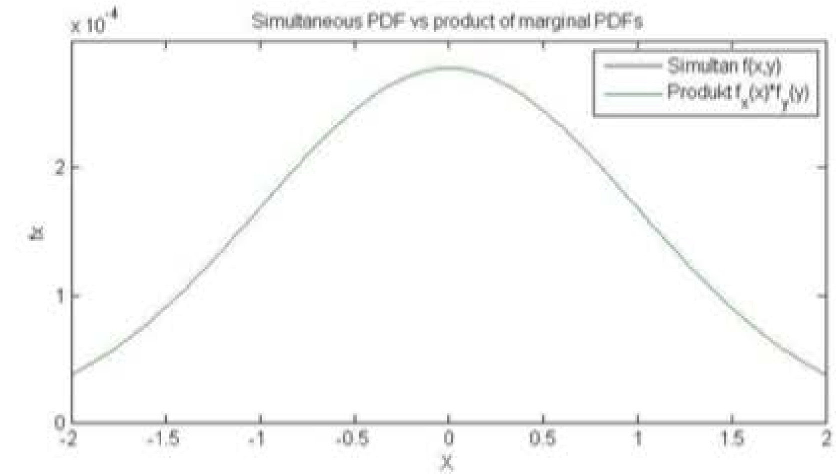
Bivariate Normal Distribution

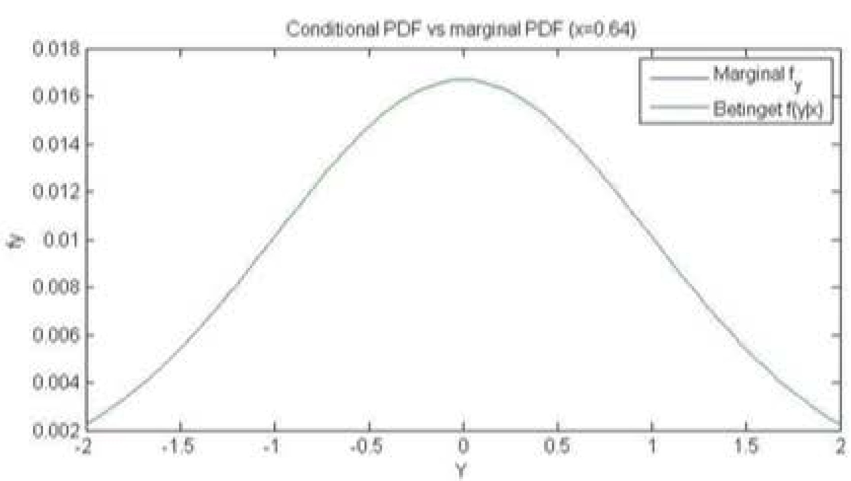
The graphs $(f_{X|Y}(x|y=0), f_{X,Y}(x,y=0))$ and $f_X(x)$ has the same shape (proportional)

$$f(x|y=0) = f_X(x)$$

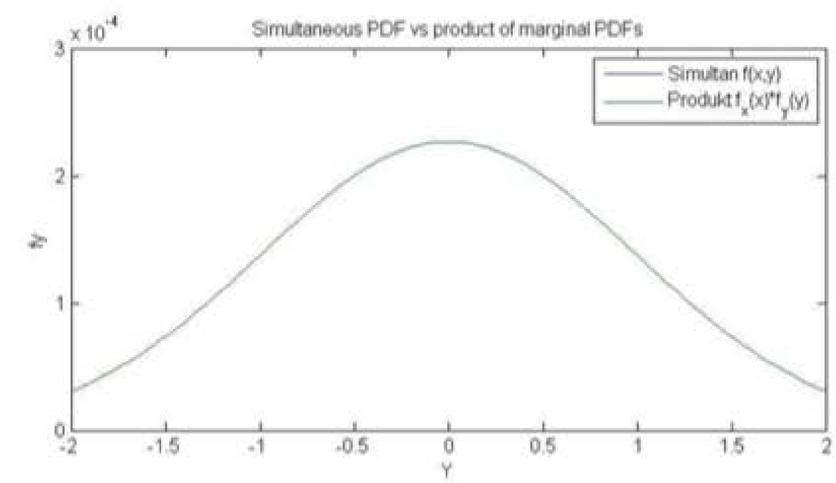


$$f(x,y=0) = f_X(x) \cdot f_Y(y=0)$$





$$f(y|x=0.64) = f_Y(y)$$

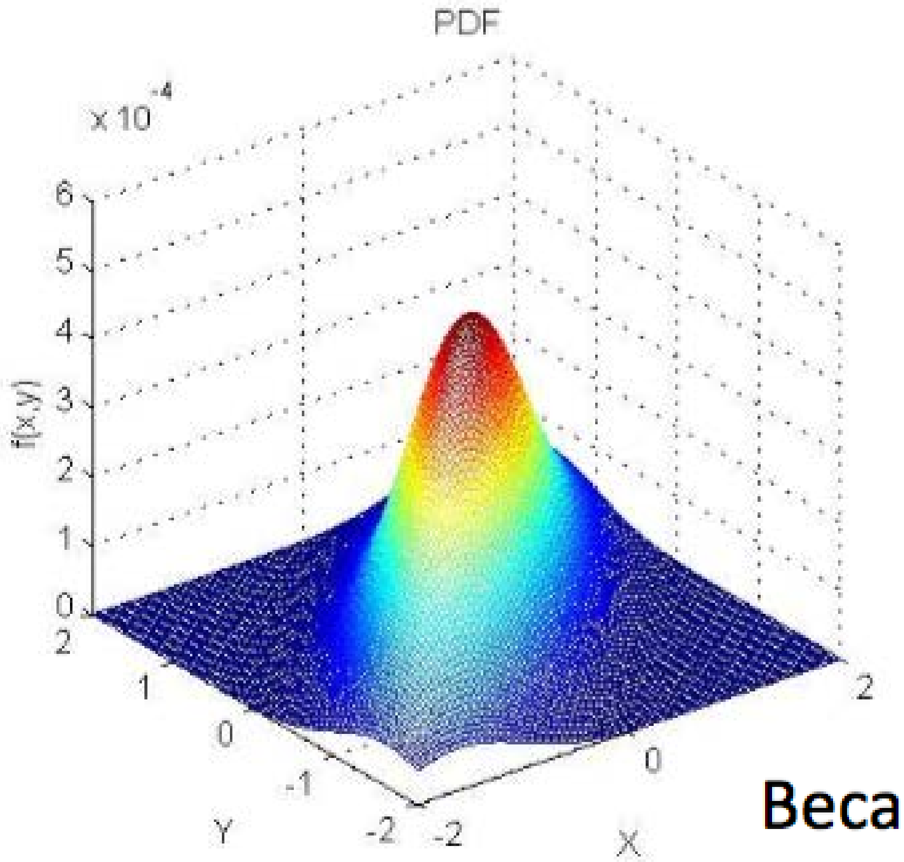


$$f(x=0.64, y) = f_X(x=0.64) \cdot f_Y(y)$$

The graphs $f_{Y|X}(y|x=0.64)$, $f_{X,Y}(x=0.64,y)$ and $f_Y(y)$ has the same shape (proportional)

Asymmetric Case

Bivariate Normal Distribution



Asymmetric PDF:

$\rho = 0.8$

X and Y **dependent**

Because of the dependence, we should have

$$f(x|y) \neq f_X(x)$$

$$f(y|x) \neq f_Y(y)$$

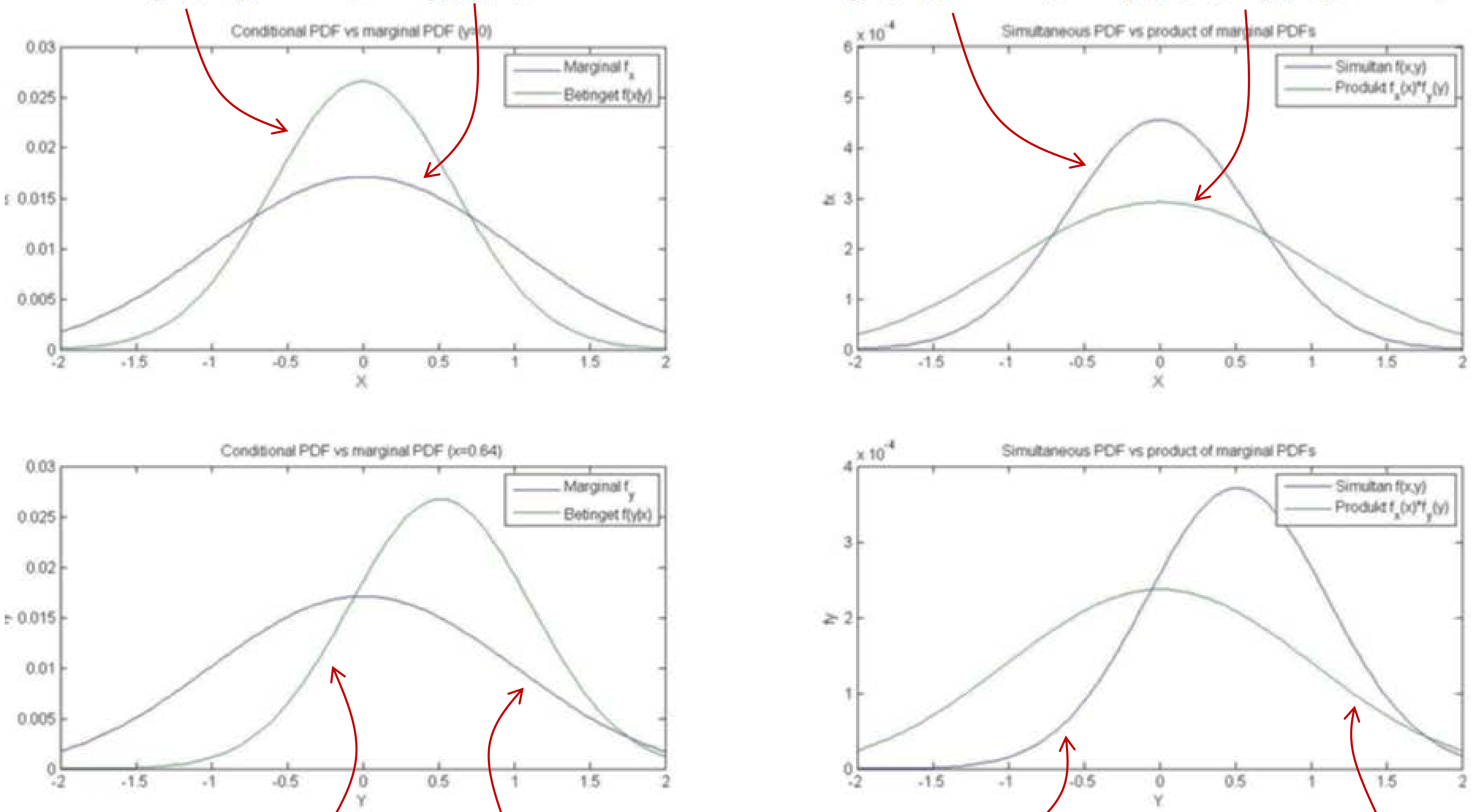
$$f(x,y) \neq f_X(x) \cdot f_Y(y)$$

Asymmetric Case

Bivariate Normal Distribution

The graphs ($f_{X|Y}(x|y=0)$, $f_{X,Y}(x,y=0)$) and $f_X(x)$ do not have the same shapes.

$f(x|y=0) \neq f_X(x)$ $f(x,y=0) \neq f_X(x) \cdot f_Y(y=0)$

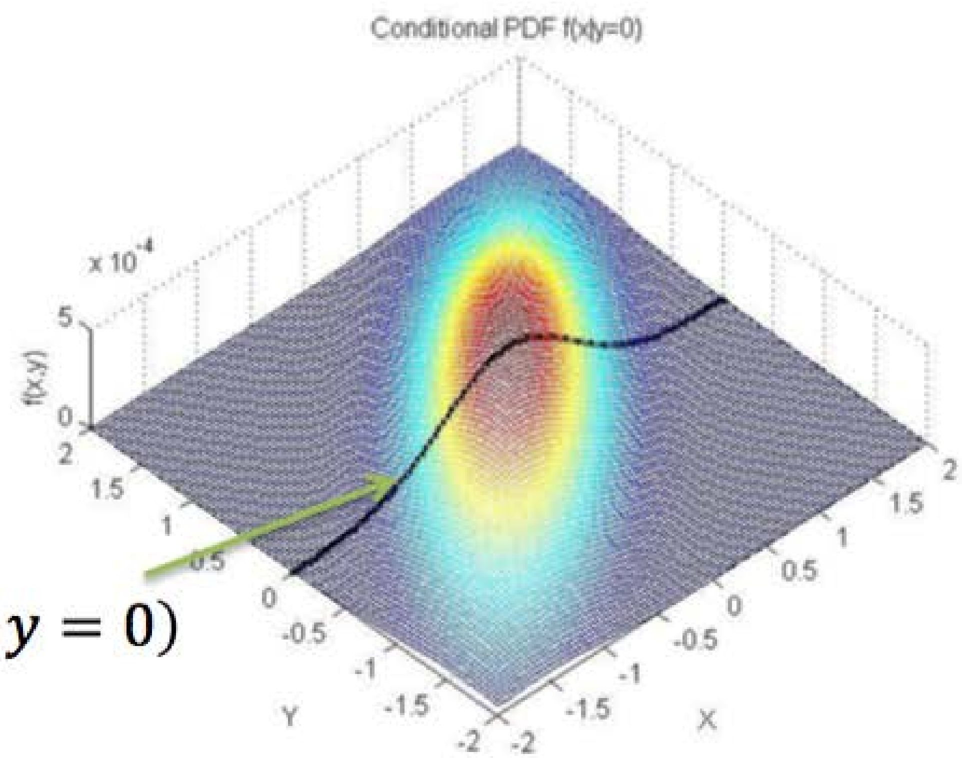


$f(y|x=0.64) \neq f_Y(y)$ $f(x=0.64,y) \neq f_X(x=0.64) \cdot f_Y(y)$

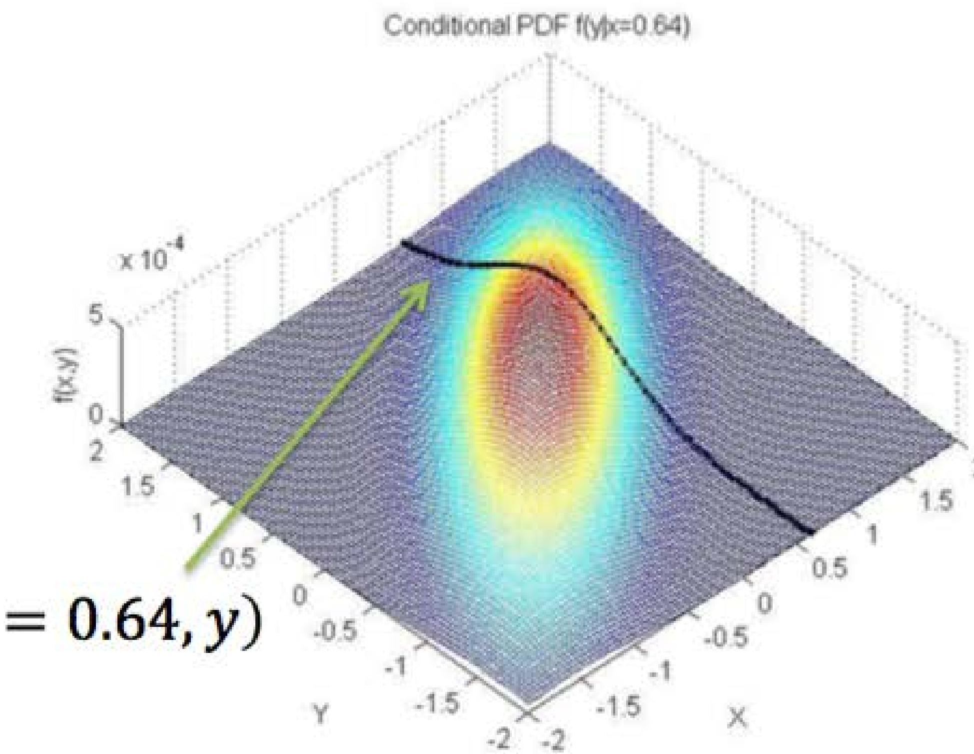
The graphs ($f_{Y|X}(y|x=0.64)$, $f_{X,Y}(x=0.64,y)$) and $f_Y(y)$ do not have the same shapes.

The Conditional pdf's

Bivariate Normal Distribution



$f(x, y = 0)$



$f(x = 0.64, y)$

Area under the curve =

$$\int_{-\infty}^{\infty} f(x, y = 0)dx = f_Y(y = 0)$$

$$f(x|y = 0) = \frac{f(x, y = 0)}{f_Y(y = 0)}$$

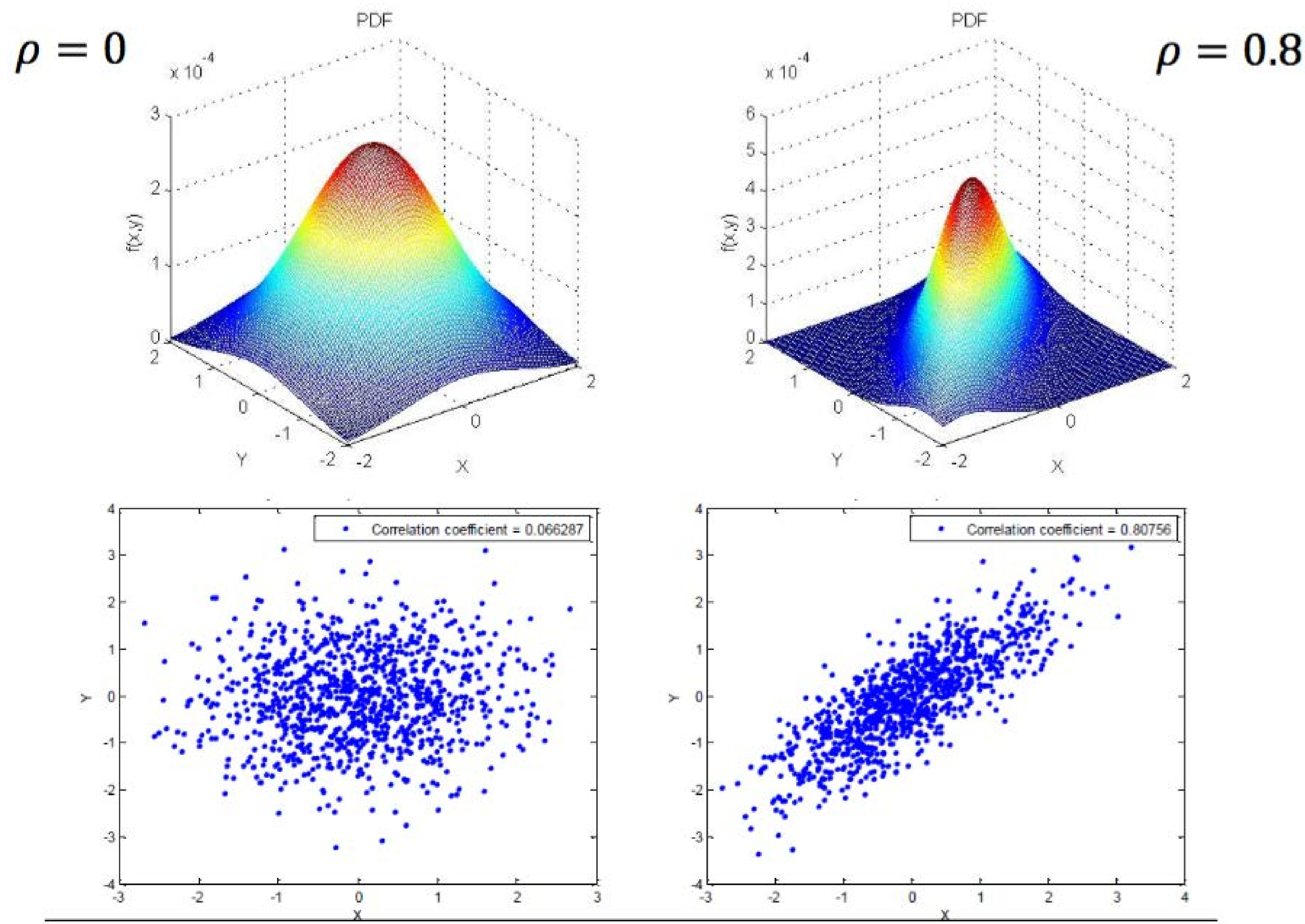
Area under the curve =

$$\int_{-\infty}^{\infty} f(x = 0.64, y)dx = f_X(x = 0.64)$$

$$f(y|x = 0.64) = \frac{f(x = 0.64, y)}{f_X(x = 0.64)}$$

Sampling

Bivariate Normal Distribution



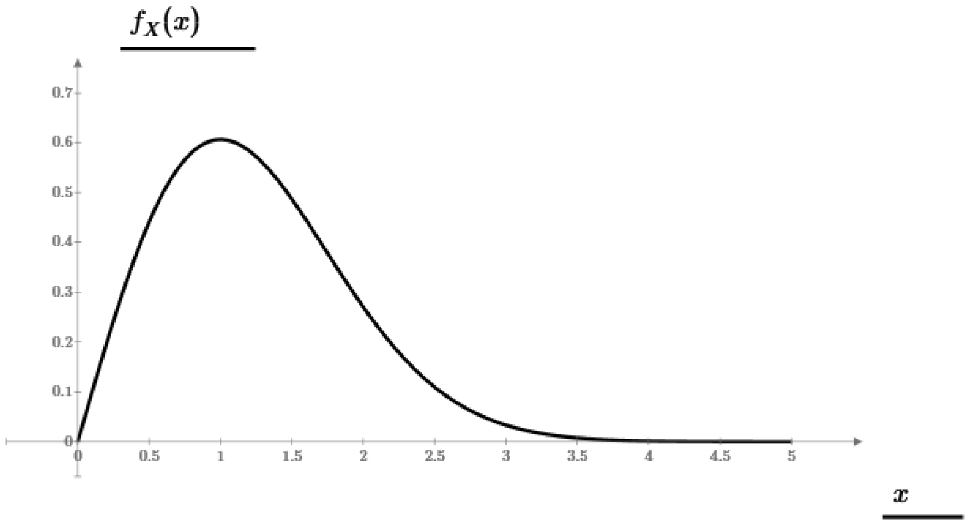
Sampling From Any Distribution

For test or simulation you need testdata (”measurements”) randomly sampled from a given distribution:

- Find the cdf of the distribution: $F_X(x)$
- Find the inverse of the cdf: $y = F_X(x) \;\Rightarrow\; x = F_X^{-1}(y)$
- Draw a ramdom sample: $y \sim \mathcal{U}[0; 1]$
- Insert into the inverse cdf: $x = F_X^{-1}(y)$
- The samples $X = x$ is distributed according to: $F_X(x)$

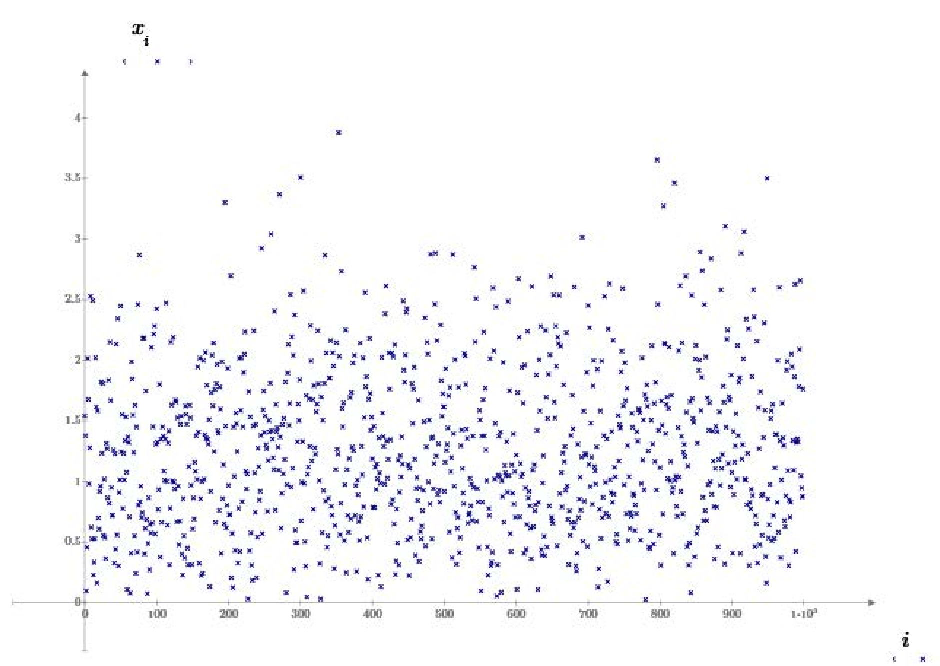
Example – Flight Simulator

- In a flight simulator, the altitude of the plane is simulated to be Rayleigh distributed.
- For a given initial height, draw a Rayleigh distributed sample.



Flight Simulator Example

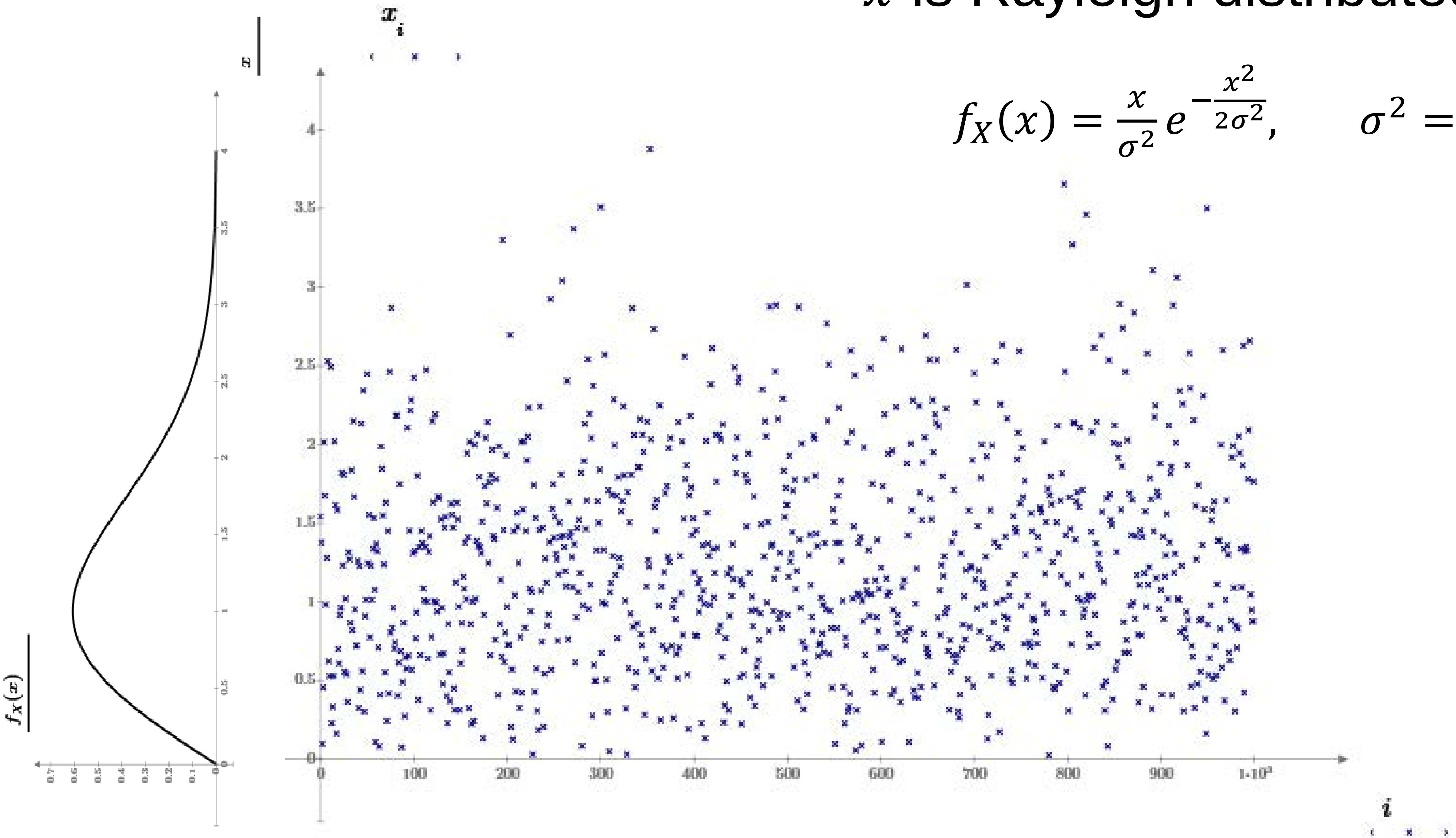
- Rayleigh pdf: $f_X(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}$ for $x \geq 0$
- Rayleigh cdf: $F_X(x) = \int_0^x \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} dx = 1 - e^{-\frac{x^2}{2\sigma^2}}$
- Invers of cdf: $\mathrm{y} = 1 - e^{-\frac{x^2}{2\sigma^2}} \Rightarrow x = \sqrt{-2\sigma^2 \ln(1 - y)}$
- Draw $\mathrm{y} \sim \mathcal{U}[0; 1]$ and insert into $x = \sqrt{-2\sigma^2 \ln(1 - y)}$
- x is Rayleigh distributed



Flight Simulator Example

x is Rayleigh distributed:

$$f_X(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \qquad \sigma^2 = 1$$



Assignment

- Choose an exponential pdf: $f_X(x) = \lambda e^{-\lambda \cdot x}$
- Make a Matlab program that samples from that distribution

Transformation of Variable X to Y

- Given:
  - Pdf: $f_X(x)$
  - Function/Transformation: $\mathrm{Y} = g(X)$
  - Limits: $\mathrm{a} \leq X \leq b$

- Find new pdf: $f_Y(y)$:
  1. Inverse: $x = g^{-1}(y)$
  2. Differentiate: $\frac{dg^{-1}(y)}{dy} = \frac{dx(y)}{dy} = \frac{1}{\frac{dg(x)}{dx}}$
  3. Limits: Find $g(a) = a_Y \leq Y \leq b_Y = g(b)$ based on $\mathrm{a} \leq X \leq b$
  4. New pdf: $f_Y(y) = \sum \left|\frac{dx(y)}{dy}\right| f_X\left(g^{-1}(y)\ \right) = \sum \frac{f_X(x)}{\left|\frac{dy}{dx}\right|}$

Example with Transformation of Random Variable

- We have a random sample x.
- The Noise is known to be Gaussian distributed.
- The signal of the noise is amplified.
- What is the pdf of the amplified noise?

- Given:
  - function: $Y = 2X$
  - pdf: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $\sim \mathcal{N}(\mu, \sigma^2)$
  - Support: $x \in \mathbf{R}$
- Steps:
  1. Inverse: $x = \frac{1}{2}y$
  2. Differentiate: $\frac{d}{dy}\frac{1}{2}y = \frac{1}{2}$
  3. Support: $y \in \mathbf{R}$
  4. New pdf: $f_Y(y) = \frac{1}{2} f_X(\frac{1}{2}y)$.
- Then: $f_Y(y) = \frac{1}{2}\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\frac{y}{2}-\mu)^2}{2\sigma^2}}$

$\sim \mathcal{N}(2\mu, 4\sigma^2)$

Distribution of the Sum of Two Random Variables

- Two random variables X and Y have density functions $f_X(x)$ and $f_Y(y)$.
- If we define a new random variable $Z = X + Y$, and Z have density function $f_Z(z)$.

Convolution of Two functions

- Then $f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx$

Expectation of the Sum of Two Random Variables

- For a random variables $Z = X + Y$.
- X, Y can be both dependent and independent.
- The expectation of Z is:

$$\boxed{E[Z] = E[X] + E[Y]}$$

Proof:

$$\begin{aligned}
E[X+Y] &= \int_x \int_y (x+y) f_{X,Y}(x,y)\,dx\,dy \\
&= \int_x \int_y x f_{X,Y}(x,y)\,dx\,dy + \int_x \int_y y f_{X,Y}(x,y)\,dx\,dy \\
&= \int_x x \int_y f_{X,Y}(x,y)\,dy\,dx + \int_y y \int_x f_{X,Y}(x,y)\,dx\,dy \\
&= \int_x x f_X(x)\,dx + \int_y y f_Y(y)\,dy \\
&= E[X] + E[Y]
\end{aligned}$$

Variance of the Sum of Two Random Variables

- We have $Z = X + Y$.
- For independent random variables X, Y, the variance of Z is:

$$\boxed{var(Z) = var(X) + var(Y).}$$

- For correlated random variables X, Y, the variance of Z is:

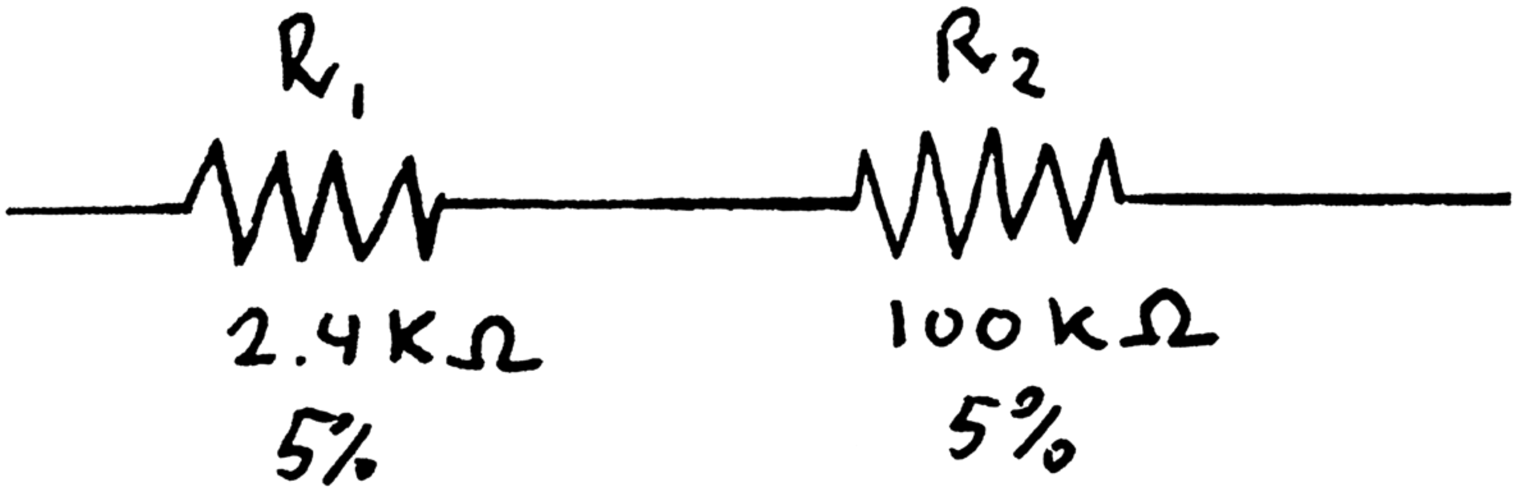
$$\boxed{var(Z) = var(X) + var(Y) + 2cov(X, Y).}$$

where: $cov(X, Y) = E[XY] - E[X]E[Y]$

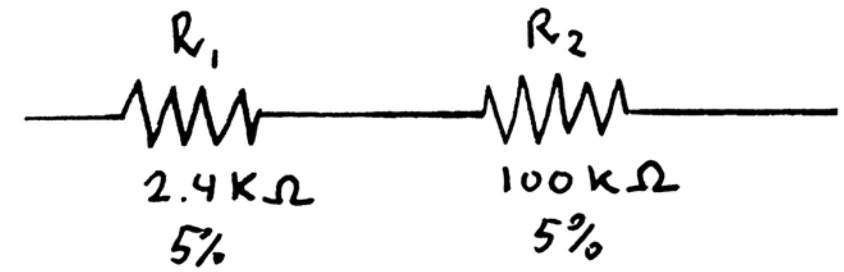
Proof: Similar to the proof of the expectation value

Precision of Resistors in Series

- In a analog filter a resister of size $2.5K\Omega$ is needed.
- We use two 5% resisters of $2.4K\Omega$ and 100Ω respectively.
- What is the resulting uncertainty of the resister?
- X and Y are independent random variables with pdfs: $f_X(x)$ and $f_Y(y)$
- What is the pdf of a random variable Z, where $Z = X + Y$

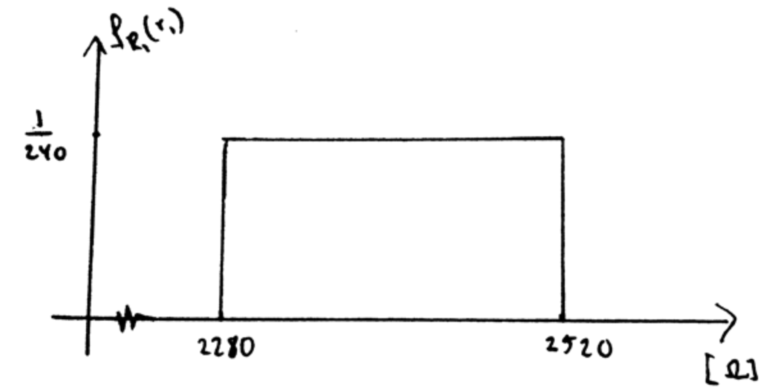


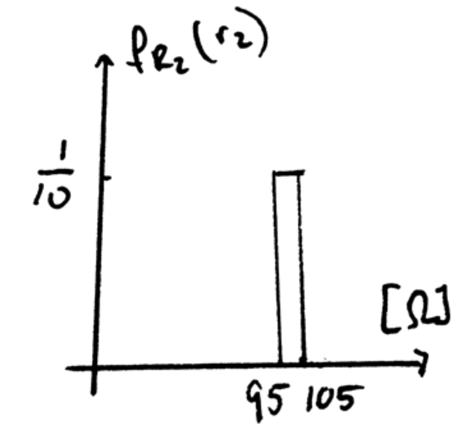
Precision of Resistors in Series

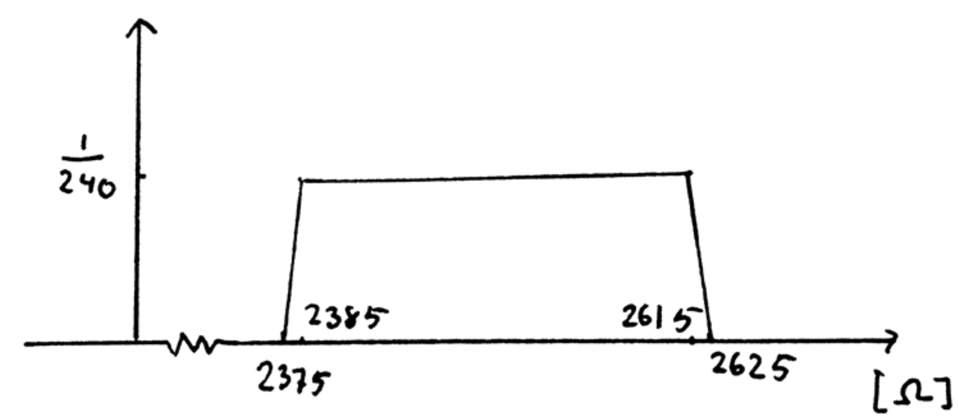


- We assume that the resistance of the resisters are uniformly distributed.
- $R_1 \sim \mathcal{U}[2280; 2520]$
- $R_2 \sim \mathcal{U}[95; 105]$
- The resisters are in series: $R_3 = R_1 + R_2$.
- We have: $f_{R_3}(r_3) = \int_{-\infty}^{\infty} f_X(\rho) f_Y(r_3 - \rho) d\rho$
- We can find that:

$$f_{R_3}(r_3) = \begin{cases} \frac{1}{2400} r_3 - \frac{95}{96} & \text{for } 2375 \leq x < 2385 \\ \frac{1}{240} & \text{for } 2385 \leq x < 2615 \\ -\frac{1}{2400} r_3 + \frac{35}{32} & \text{for } 2615 \leq x < 2625 \\ 0 & \text{otherwise} \end{cases}$$

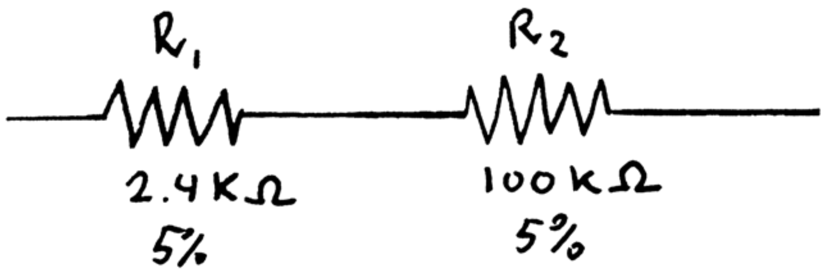






R3 is still a 5% resistor – but no longer uniform distributed!

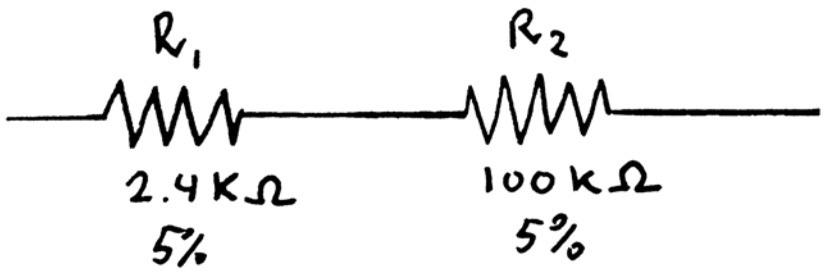
Expected Value of the Resistor



- We assume that R_1 and R_2 are independent
- For a uniform distibution: $\mathrm{E}[R_1] = \frac{1}{2}(2520 + 2280) = 2400\Omega$
- For a uniform distribution: $\mathrm{E}[R_2] = \frac{1}{2}(105 + 95) = 100\Omega$
- For the sum $R_3 = R_1 + R_2$ we have:

$$E[R_3] = E[R_1] + E[R_2] = 2400\Omega + 100\Omega = \underline{2500\Omega}$$

Variance of the Resistor



- We assume that R_1 and R_2 are independent
- For a uniform distibution: $var(R_1) = \frac{1}{12}(2520 - 2280)^2 = 4800$
- For a uniform distribution: $var(R_2) = \frac{1}{12}(105 - 95)^2 = 8{,}333$
- For the sum $R_3 = R_1 + R_2$ we have:

$$var(R_3) = var(R_1) + var(R_2) = 4808 \;\rightarrow\; \sigma_3 = 69\Omega$$

- For one uniform distributed 5%-resistor $R_0 = 2500 \sim \mathcal{U}[2375; 2625]$:

$$var(R_0) = \frac{1}{12}(2625 - 2375)^2 = 5208 \;\rightarrow\; \sigma_0 = 72\Omega$$

- So: $var(R_3) = var(R_1) + var(R_2) < var(R_0)$ $\quad (\sigma_3 < \sigma_0)$

Two Random Variables

Two random variables: X and Y

- Simultaneous pdf: $f_{X,Y}(x,y)$
- Marginal pdf: $f_X(x)$ and $f_Y(y)$
- Conditional pdf: $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$
- Simultaneous cdf: $F_{X,Y}(x,y)$
- Correlation: $corr(X,Y) = E[XY]$
- Covariance: $cov(X,Y) = E[XY] - E[X]E[Y]$
- Correlation coefficient: $\rho = \frac{E[XY]-E[X]E[Y]}{\sigma_X \cdot \sigma_Y}$

- Sum: $\mathrm{Z} = X + Y$
- Expectation: $\mathrm{E[Z]} = E[X] + E[Y]$
- Variance: $\mathrm{Var[Z]} = Var[X] + Var[Y]$ if <u>independent</u>
  $\mathrm{Var[Z]} = Var[X] + Var[Y] + 2cov(X,Y)$ if <u>dependent</u>

Central Limit Theorem

- Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with mean μ and variance σ^2
- Let $\bar{X}$ be the random variable (average):

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- Then in the limit: $n \to \infty$ we have that: $\underline{\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})}$

i.e. in the limit $\bar{X}$ will be normally distributed with mean = μ and variance = $\frac{\sigma^2}{n}$.

The variance is reduced with a factor 1/n

Central Limit Theorem

- Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with mean μ and variance σ^2
- Let X be the random variable:

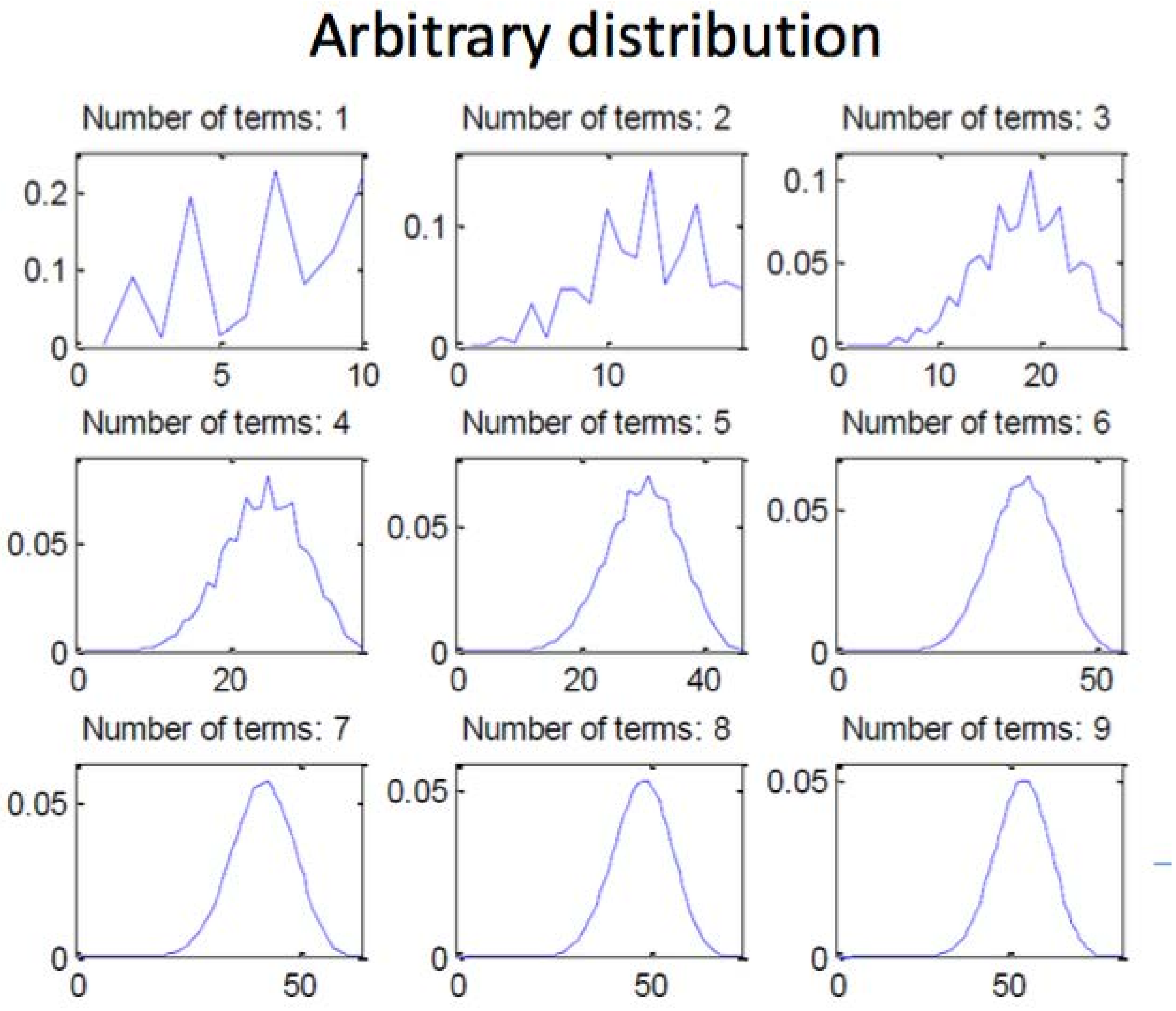
$$X = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} = \frac{\sum_{i=1}^{n} \frac{1}{n} X_i - \mu}{\sqrt{\sigma^2/_n}} = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/_n}}$$

- Then in the limit: $n \to \infty$ we have that: $X \sim \mathcal{N}(0,1)$

i.e. in the limit X will be normally distributed with mean = 0 and variance = 1 (standard normal distributed).

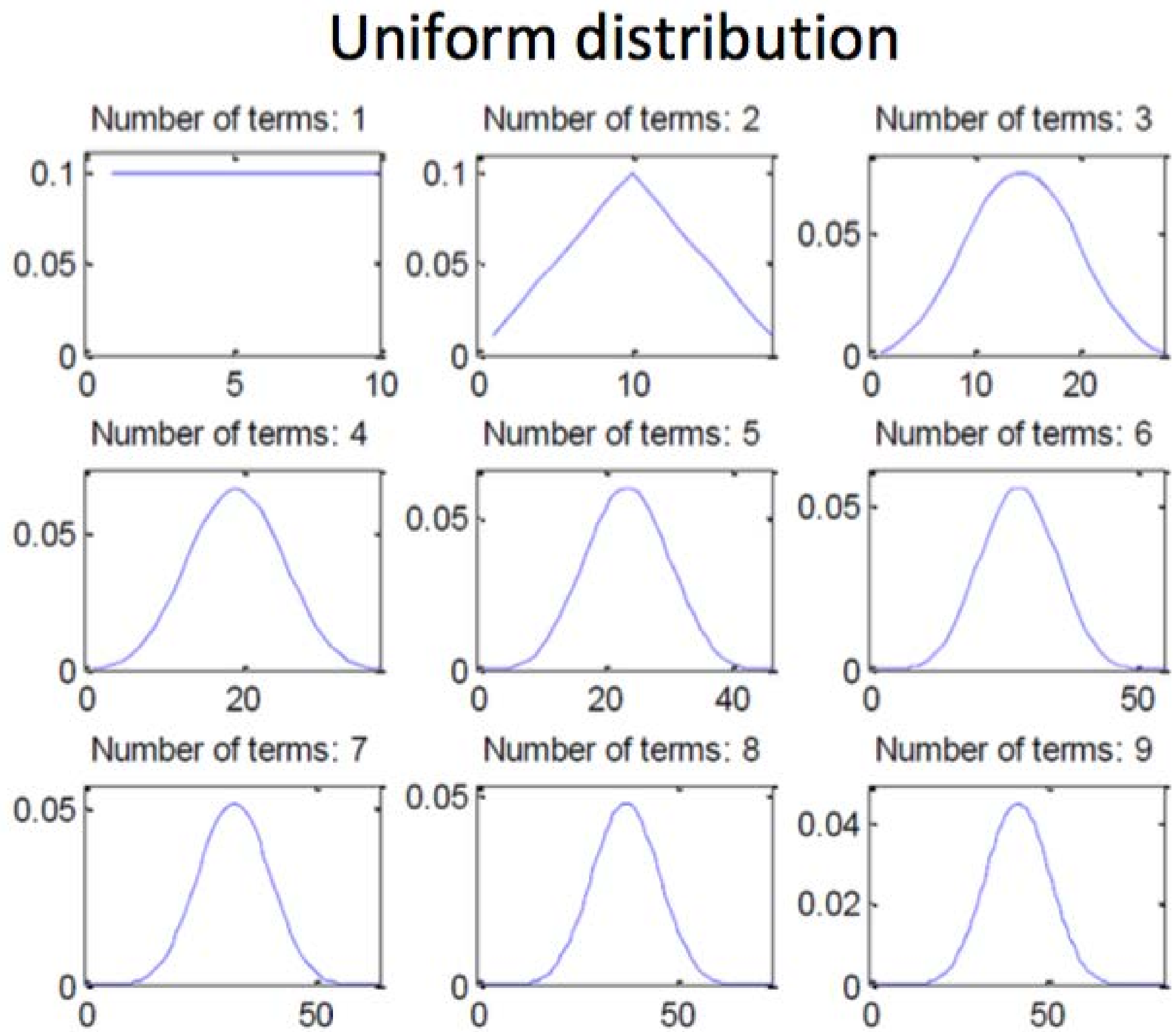
Sum of Random Variables

- The random variables are i.i.d and taken from the same distribution.



Sum of Random Variables

- The random variables are i.i.d and taken from the same uniform distribution.



Words and Concepts to Know

Central Limit Theorem

Convolution

Transformation of stochastic variables

Rayleigh Distribution

Randomly Sampled Data

Bivariate Normal Distribution