FSAN/ELEG815: Statistical Learning

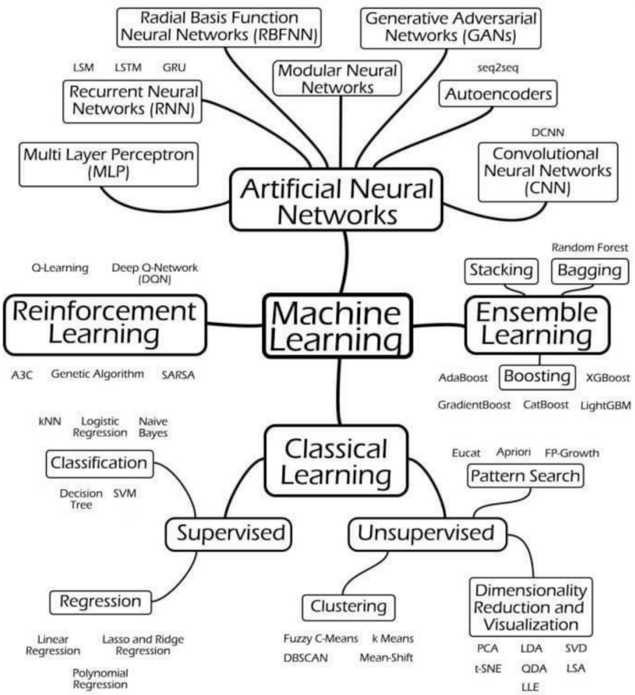
Gonzalo R. Arce

Department of Electrical and Computer Engineering
University of Delaware

I: Review of Probability

Outline of the Course

1. Review of Probability and Stationary processes
2. Eigen Analysis, Singular Value Decomposition (SVD) Principal Component Analysis (PCA) and Matrix Completition
3. The Learning Problem
4. Training vs Testing
5. The Linear Model
6. Overfitting and Regularization (Ridge Regression)
7. Lasso Regression
8. Suport Vector Machines (SVM)
9. Neural Networks
10. Convolutional Neural Networks


Radial Basis Function Neural Networks (RBFNN)
Generative Adversarial Networks (GANs)
LSM
LSTM
GRU
Modular Neural Networks
seq2seq
Recurrent Neural Networks (RNN)
Autoencoders
DCNN
Multi Layer Perceptron (MLP)
Convolutional Neural Networks (CNN)
Artificial Neural Networks
Random Forest
Q-Learning
Deep Q-Network (DQN)
Stacking
Bagging
Reinforcement Learning
Machine Learning
Ensemble Learning
A3C
Genetic Algorithm
SARSA
AdaBoost
Boosting
XGBoost
GradientBoost
CatBoost
LightGBM
kNN
Logistic Regression
Naive Bayes
Classical Learning
Eucat
Apriori
FP-Growth
Classification
Pattern Search
Decision Tree
SVM
Supervised
Unsupervised
Regression
Clustering
Dimensionality Reduction and Visualization
Linear Regression
Lasso and Ridge Regression
Polynomial Regression
Fuzzy C-Means
k Means
DBSCAN
Mean-Shift
PCA
LDA
SVD
t-SNE
QDA
LSA
LLE

Random Variables

Definition

For a space S, the subsets, or events of S, have associated probabilities.

- To every event δ, we assign a number $x(\delta)$, which is called a *R.V.*
- The distribution function of x is

$$\Pr\{x \leq x_0\} = F_x(x_0) \qquad -\infty < x_0 < \infty$$

Properties:

1. $F(+\infty) = 1, \quad F(-\infty) = 0$
2. $F(x)$ is continuous from the right

$$F(x^+) = F(x)$$

3. $\Pr\{x_1 < x \leq x_2\} = F(x_2) - F(x_1)$

Example

Fair toss of two coins: H=heads, T=Tails

Define numerical assignments:

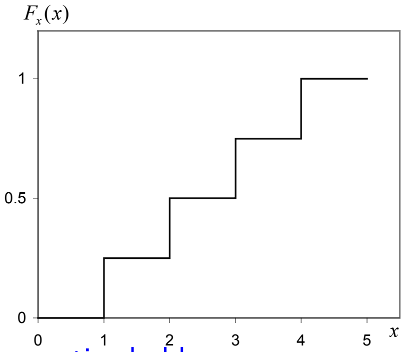
Events(δ)	Prob.	X(δ)	Y(δ)
HH	1/4	1	-100
HT	1/4	2	-100
TH	1/4	3	-100
TT	1/4	4	500

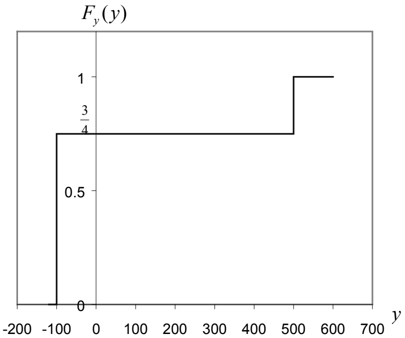
This assignments yield different distribution functions

$$F_x(2) = \Pr\{HH, HT\}$$
$$F_y(2) = \Pr\{HH, HT, TH\}$$

How do we attain an intuitive interpretation of the distribution function?

Distribution Plots





Note properties hold:

1. $F(+\infty) = 1, \quad F(-\infty) = 0$
2. $F(x)$ is continuous from the right

$$F(x^+) = F(x)$$

3. $\Pr\{x_1 < x \leq x_2\} = F(x_2) - F(x_1)$

Definition

The probability density function is defined as,

$$f(x) = \frac{dF(x)}{dx}$$

$$\text{or} \quad F(x) = \int_{-\infty}^{x} f(x)dx$$

$$\text{Thus} \quad F(\infty) = 1 \Rightarrow \int_{-\infty}^{\infty} f(x)dx = 1$$

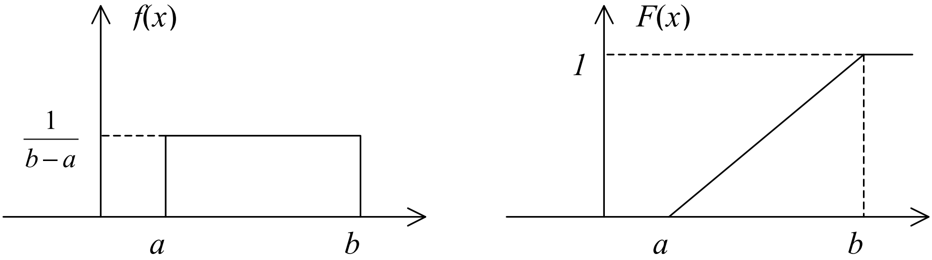
Types of distributions:

- Continuous: $\Pr\{x = x_0\} = 0 \quad \forall x_0$
- Discrete: $F(x_i) - F(x_i^-) = \Pr\{x = x_i\} = P_i$
  - In which case $f(x) = \sum_i P_i \delta(x - x_i)$
- Mixed: discontinuous but not discrete

Distribution examples

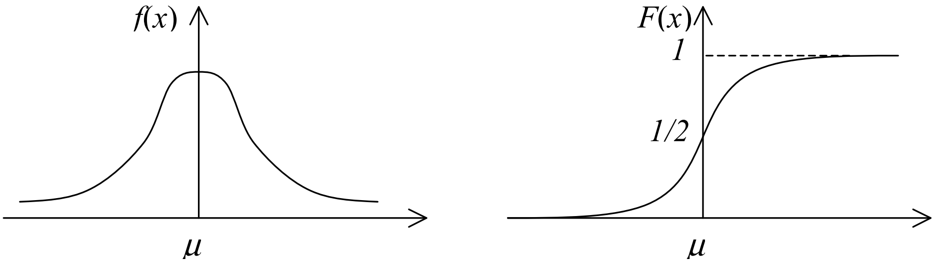
Uniform: $x \sim U(a,b) \quad a < b$

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{else} \end{cases}$$



Gaussian: $x \sim N(\mu, \sigma)$

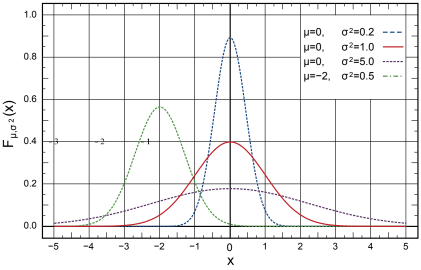
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

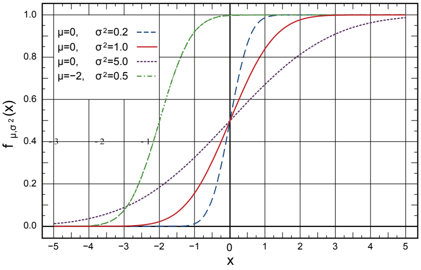


Gaussian Distribution Example

Example

Consider the Normal (Gaussian) distribution PDF and CDF for $\mu = 0, \sigma^2 = 0.2, 1.0, 5.0$ and $\mu = -2, \sigma^2 = 0.5$





Binomial: $x \sim B(p,q) \quad p+q=1$

Example

Toss a coin n times. What is the probability of getting k heads?

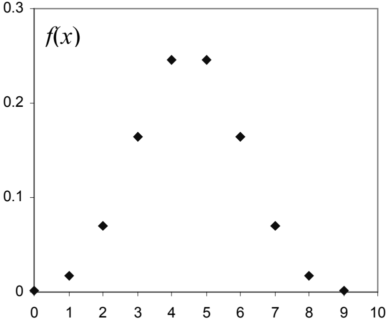
For $\quad p+q=1$, where q is probability of a tail, and p is the probability of a head:

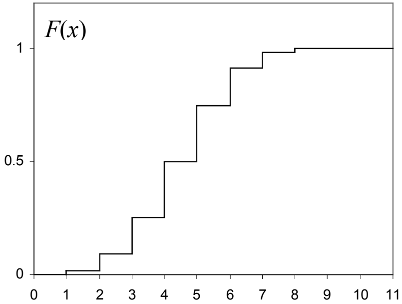
$$\begin{aligned}
\Pr\{x=k\} &= \binom{n}{k} p^k q^{n-k} \qquad \left[\text{NOTE: } \binom{n}{k} = \frac{n!}{(n-k)!k!}\right] \\
\Rightarrow f(x) &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \delta(x-k) \\
\Rightarrow F(x) &= \sum_{k=0}^{m} \binom{n}{k} p^k q^{n-k} \quad m \le x < m+1
\end{aligned}$$

Binomial Distribution Example I

Example

Toss a coin n times. What is the probability of getting k heads? For $n = 9, p = q = \frac{1}{2}$ (fair coin)

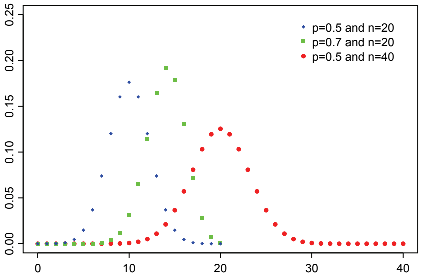


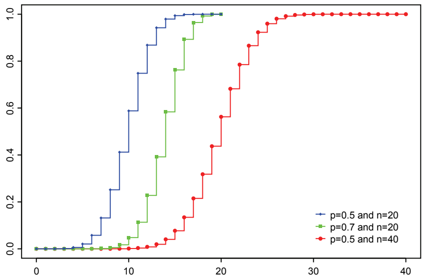


Binomial Distribution Example II

Example

Toss a coin n times. What is the probability of getting k heads? For $n = 20, p = 0.5, 0.7$ and $n = 40, p = 0.5$.





Conditional Distributions

Definition

The conditional distribution of x given event "M" has occurred is

$$F_x(x_0|M) = \text{Pr}\{x \leq x_0|M\} = \frac{\text{Pr}\{x \leq x_0, M\}}{\text{Pr}\{M\}}$$

Example

Suppose $M = \{x \leq a\}$, then

$$F_x(x_0|M) = \frac{\text{Pr}\{x \leq x_0, M\}}{\text{Pr}\{x \leq a\}}$$

If $x_0 \geq a$, what happens?

Special Cases

Special Case: $x_0 \geq a$

$$\Pr\{x \leq x_0, x \leq a\} = \Pr\{x \leq a\}$$

$$\Rightarrow F_x(x_0|M) = \frac{\Pr\{x \leq x_0, M\}}{\Pr\{x \leq a\}} = \frac{\Pr\{x \leq a\}}{\Pr\{x \leq a\}} = 1$$

Special Case: $x_0 \leq a$

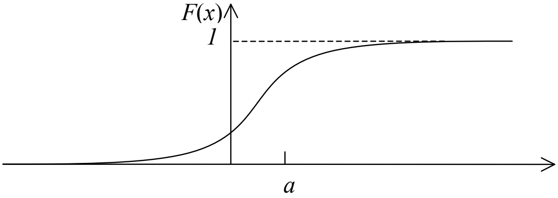
$$\Rightarrow F_x(x_0|M) = \frac{\Pr\{x \leq x_0, M\}}{\Pr\{x \leq a\}} = \frac{\Pr\{x \leq x_0\}}{\Pr\{x \leq a\}}$$

$$= \frac{F_x(x_0)}{F_x(a)}$$

Conditional Distribution Example

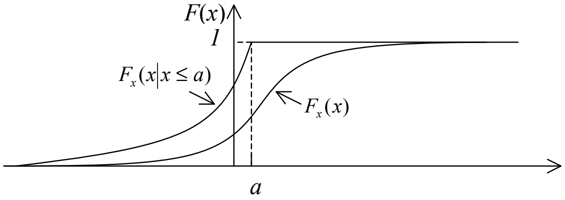
Example

Suppose



What does $F_x(x|M)$ look like? Note $M = \{x \leq a\}$.

$$\Rightarrow F_x(x_0|M) = \begin{cases} \frac{F_x(x_0)}{F_x(a)} & x \leq a \\ 1 & a \leq x \end{cases}$$



- Distribution properties hold for conditional cases:
  - Limiting cases: $F(\infty|M) = 1$ and $F(-\infty|M) = 0$
  - Probability range: $\Pr\{x_0 \leq x \leq x_1|M\} = F(x_1|M) - F(x_0|M)$
  - Density–distribution relations:

$$f(x|M) = \frac{\partial F(x|M)}{\partial x}$$

$$F(x_0|M) = \int_{-\infty}^{x_0} f(x|M)dx$$

Example (Fair Coin Toss)

Toss a fair coin 4 times. Let x be the number of heads. Determine $\Pr\{x=k\}$.

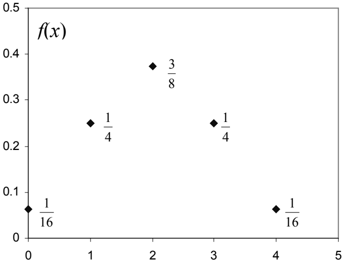
Recall

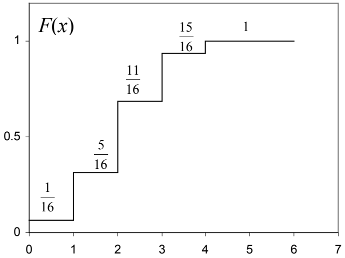
$$\Pr\{x=k\} = \binom{n}{k} p^k q^{n-k}$$

In this case

$$\begin{aligned}
\Pr\{x=k\} &= \binom{4}{k}\left(\frac{1}{2}\right)^4 \\
\Pr\{x=0\} &= \Pr\{x=4\} = \frac{1}{16} \\
\Pr\{x=1\} &= \Pr\{x=3\} = \frac{1}{4} \\
\Pr\{x=2\} &= \frac{3}{8}
\end{aligned}$$

Density and Distribution Plots for Fair Coin ($n = 4$) Ex.





What type of distribution is this? Discrete. Thus,

$$F(x_i) - F(x_i^-) = \Pr\{x = x_i\} = P_i$$

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{x} \sum_i P_i \delta(x - x_i)dx$$

Conditional Case

Example (Conditional Fair Coin Toss)

Toss a fair coin 4 times. Let x be the number of heads. Suppose $M =$ [at least one flip produces a head]. Determine $\Pr\{x = k|M\}$.

Recall,

$$\Pr\{x = k|M\} = \frac{\Pr\{x = k, M\}}{\Pr\{M\}}$$

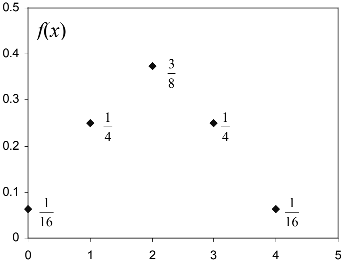
Thus first determine $\Pr\{M\}$

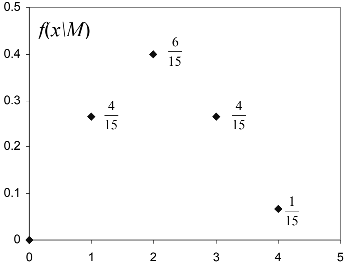
$$\begin{aligned}\Pr\{M\} &= 1 - \Pr\{\text{No heads}\} \\ &= 1 - \frac{1}{16} \\ &= \frac{15}{16}\end{aligned}$$

Next determine $\Pr\{x = k|M\}$ for the individual cases, $k = 0, 1, 2, 3, 4$

$$
\begin{aligned}
\Pr\{x = 0|M\} &= \frac{\Pr\{x = 0, M\}}{\Pr\{M\}} = 0 \\
\Pr\{x = 1|M\} &= \frac{\Pr\{x = 1, M\}}{\Pr\{M\}} \\
&= \frac{\Pr\{x = 1\}}{\Pr\{M\}} = \frac{1/4}{15/16} = \frac{4}{15} \\
\Pr\{x = 2|M\} &= \frac{\Pr\{x = 2\}}{\Pr\{M\}} = \frac{3/8}{15/16} = \frac{6}{15} \\
\Pr\{x = 3|M\} &= \frac{4}{15} \\
\Pr\{x = 4|M\} &= \frac{1}{15}
\end{aligned}
$$

Conditional and Unconditional Density Functions





Are they proper density functions?

Functions of a *R.V.*

Problem Statement

Let x and $g(x)$ be *RVs* such that

$$y = g(x)$$

Question: How do we determine the distribution of y?

Note

$$\begin{aligned} F_y(y_0) &= \Pr\{y \leq y_0\} \\ &= \Pr\{g(x) \leq y_0\} \\ &= \Pr\{x \in R_{y_0}\} \end{aligned}$$

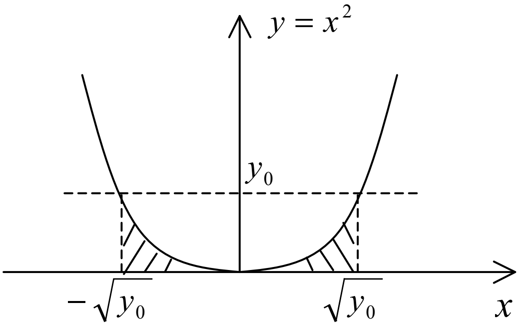
where

$$R_{y_0} = \{x : g(x) \leq y_0\}$$

Question: If $y = g(x) = x^2$, what is R_{y_0} ?

Example

Let $y = g(x) = x^2$. Determine $F_y(y_0)$.



Note that

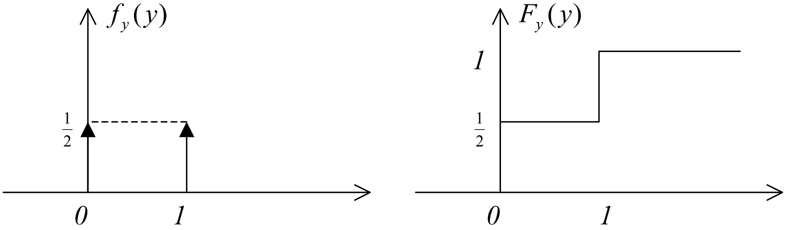
$$\begin{aligned} F_y(y_0) &= \text{Pr}(y \leq y_0) \\ &= \text{Pr}(-\sqrt{y_0} \leq x \leq \sqrt{y_0}) \\ &= F_x(\sqrt{y_0}) - F_x(-\sqrt{y_0}) \end{aligned}$$

Example

Let $x \sim N(\mu, \sigma)$ and

$$y = U(x) = \begin{cases} 1 & \text{if} \quad x > \mu \\ 0 & \text{if} \quad x \leq \mu \end{cases}$$

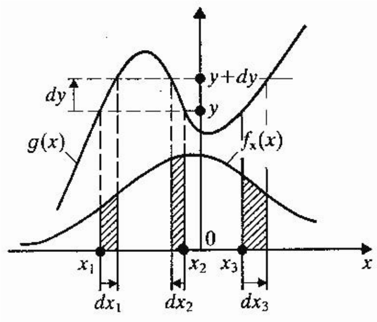
Determine $f_y(y_0)$ and $F_y(y_0)$.



General Function of a Random Variable Case

Additional Notes

To determine the density of $y = g(x)$ in terms of $f_x(x_0)$, look at $g(x)$



$$
\begin{aligned}
f_y(y_0)dy_0 &= \Pr(y_0 \le y \le y_0 + dy_0) \\
&= \Pr(x_1 \le x \le x_1 + dx_1) + \Pr(x_2 + dx_2 \le x \le x_2) \\
&\quad + \Pr(x_3 \le x \le x_3 + dx_3)
\end{aligned}
$$

Additional Notes

$$\begin{aligned} f_y(y_0)dy_0 &= \Pr(x_1 \leq x \leq x_1 + dx_1) + \Pr(x_2 + dx_2 \leq x \leq x_2) \\ &\quad + \Pr(x_3 \leq x \leq x_3 + dx_3) \\ &= f_x(x_1)dx_1 + f_x(x_2)|dx_2| + f_x(x_3)dx_3 \qquad (*) \end{aligned}$$

Note that

$$dx_1 = \frac{dx_1}{dy_0} dy_0 = \frac{dy_0}{dy_0/dx_1} = \frac{dy_0}{g'(x_1)}$$

Similarly

$$dx_2 = \frac{dy_0}{g'(x_2)} \quad \text{and} \quad dx_3 = \frac{dy_0}{g'(x_3)}$$

Thus $(*)$ becomes

$$f_y(y_0)dy_0 = \frac{f_x(x_1)}{g'(x_1)} dy_0 + \frac{f_x(x_2)}{|g'(x_2)|} dy_0 + \frac{f_x(x_3)}{g'(x_3)} dy_0$$

or

$$f_y(y_0) = \frac{f_x(x_1)}{g'(x_1)} + \frac{f_x(x_2)}{|g'(x_2)|} + \frac{f_x(x_3)}{g'(x_3)}$$

Function of a R.V. Distribution General Result

Set $y = g(x)$ and let $x_1, x_2, \ldots$ be the roots, i.e.,

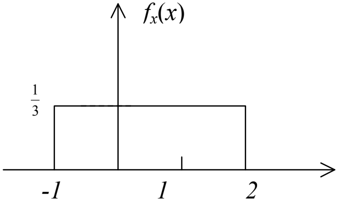
$$y = g(x_1) = g(x_2) = \ldots$$

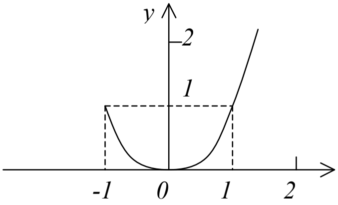
Then

$$f_y(y) = \frac{f_x(x_1)}{|g'(x_1)|} + \frac{f_x(x_2)}{|g'(x_2)|} + \ldots$$

Example

Suppose $x \sim U(-1,2)$ and $y = x^2$. Determine $f_y(y)$.





Note that

$$g(x) = x^2 \Rightarrow g'(x) = 2x$$

Consider special cases separately:

Case 1: $0 \leq y \leq 1$

$$y = x^2 \Rightarrow x = \pm\sqrt{y}$$

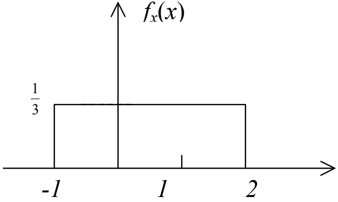
$$\begin{aligned} f_y(y) &= \frac{f_x(x_1)}{|g'(x_1)|} + \frac{f_x(x_2)}{|g'(x_2)|} \\ &= \frac{1/3}{|2\sqrt{y}|} + \frac{1/3}{|-2\sqrt{y}|} = \frac{1/3}{\sqrt{y}} \end{aligned}$$

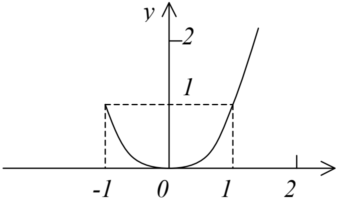
Case 2: $1 \leq y \leq 4$

$$y = x^2 \Rightarrow x = \sqrt{y}$$

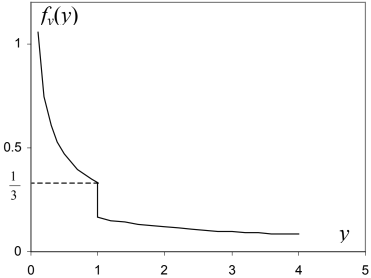
$$f_y(y) = \frac{f_x(x_1)}{|g'(x_1)|} = \frac{1/3}{2\sqrt{y}} = \frac{1/6}{\sqrt{y}}$$

Result: For $x \sim U(-1,2)$ and $y = x^2$





$$f_y(y) = \begin{cases} \frac{1/3}{\sqrt{y}} & 0 \leq y \leq 1 \\ \frac{1/6}{\sqrt{y}} & 1 < y \leq 4 \end{cases}$$



Example

Let $x \sim N(\mu, \sigma)$ and $y = e^x$. Determine $f_y(y)$.

Note $g(x) \geq 0$ and $g'(x) = e^x$

Also, there is a single root (inverse solution):

$$x = \ln(y)$$

Therefore,

$$f_y(y) = \frac{f_x(x)}{|g'(x)|} = \frac{f_x(x)}{e^x}$$

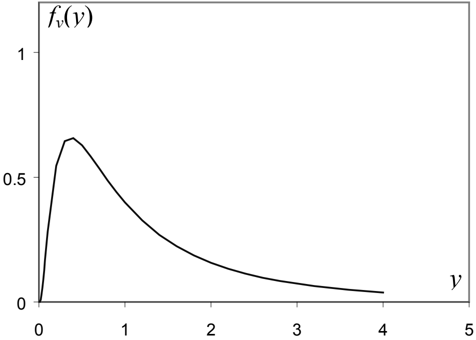
Expressing this in terms of y through substitution yields:

$$f_y(y) = \frac{f_x(\ln(y))}{e^{\ln(y)}} = \frac{f_x(\ln(y))}{y}$$

Note that x is Gaussian:

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

$$\Rightarrow f_y(y) = \frac{1}{\sqrt{2\pi}y\sigma} e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}}, \quad \text{for } y > 0$$



Log normal density

Mean, Median and variance

Definitions

$$\text{Mean} \quad E\{x\} = \int_{-\infty}^{\infty} x f(x) dx$$

$$\text{Conditional Mean} \quad E\{x|M\} = \int_{-\infty}^{\infty} x f(x|M) dx$$

Example

Suppose $M = \{x \geq a\}$. Then

$$E\{x|M\} = \int_{-\infty}^{\infty} x f(x|M) dx = \frac{\int_a^{\infty} x f(x) dx}{\int_a^{\infty} f(x) dx}$$

For a function of a *RV*, $y = g(x)$,

$$E\{y\} = \int_{-\infty}^{\infty} y f_y(y) dy = \int_{-\infty}^{\infty} g(x) f_x(x) dx$$

Example

Suppose $g(x)$ is a step function: Determine $E\{g(x)\}$.

$$E\{g(x)\} = \int_{-\infty}^{\infty} g(x) f_x(x) dx = \int_{-\infty}^{x_0} f_x(x) dx = F_x(x_0)$$

Median

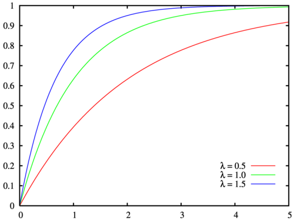
Definitions

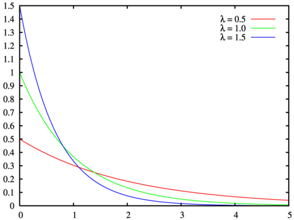
$$\text{Median}=\ m \quad \int_{-\infty}^{m} f(x)dx = \int_{m}^{\infty} f(x)dx = \frac{1}{2}$$

$$\text{Median} \quad Pr\{x \leq m\} = Pr\{x \geq m\}$$

Example

Let $x \sim \lambda \exp^{-\lambda x} U(x)$. Then $m = \frac{\ln(2)}{\lambda}$





Definition (Variance)

$$\text{Variance} \quad \sigma^2 = \int_{-\infty}^{\infty} (x-\eta)^2 f(x)dx$$

where $\eta = E\{x\}$. Thus,

$$\sigma^2 = E\{(x-\eta)^2\} = E\{x^2\} - E^2\{x\}$$

Example

For $x \sim N(\eta, \sigma^2)$, determine the variance.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\eta)^2}{2\sigma^2}}$$

Note: $f(x)$ is symmetric about $x = \eta \Rightarrow E\{x\} = \eta$

Also

$$\int_{-\infty}^{\infty} f(x)dx = 1 \Rightarrow \int_{-\infty}^{\infty} e^{-\frac{(x-\eta)^2}{2\sigma^2}} dx = \sqrt{2\pi}\sigma$$

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\eta)^2}{2\sigma^2}} dx = \sqrt{2\pi}\sigma$$

Differentiating w.r.t. σ:

$$\Rightarrow \int_{-\infty}^{\infty} \frac{(x-\eta)^2}{\sigma^3} e^{-\frac{(x-\eta)^2}{2\sigma^2}} dx = \sqrt{2\pi}$$

Rearranging yields

$$\int_{-\infty}^{\infty} (x-\eta)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\eta)^2}{2\sigma^2}} dx = \sigma^2$$

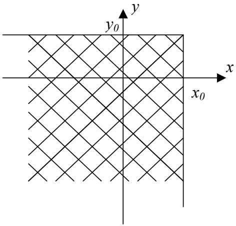
or

$$E\{(x-\eta)^2\} = \sigma^2$$

Bivariate Statistics

Given two *RVs*, x and y, the bivariate (joint) distribution is given by

$$F(x_0, y_0) = \Pr\{x \leq x_0, y \leq y_0\}$$

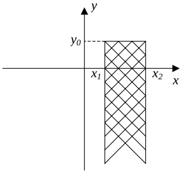


Properties:

- $F(-\infty, y) = F(x, -\infty) = 0$
- $F(\infty, \infty) = 1$
- $F_x(x) = F(x, \infty), \quad F_y(y) = F(\infty, y)$

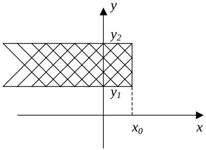
Special Cases

Case 1: $M = \{x_1 \le x \le x_2, y \le y_0\}$



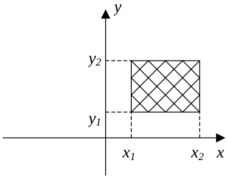
$$\Rightarrow \Pr\{M\} = F(x_2, y_0) - F(x_1, y_0)$$

Case 2: $M = \{x \le x_0, y_1 \le y \le y_2\}$



$$\Rightarrow \Pr\{M\} = F(x_0, y_2) - F(x_0, y_1)$$

Case 3: $M = \{x_1 \leq x \leq x_2, y_1 \leq y \leq y_2\}$ Then



and

$$\Pr\{M\} = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + \underbrace{F(x_1, y_1)}_{\downarrow}$$

Added back because this region was subtracted twice

Definition (Joint Statistics)

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}$$

and

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(\alpha,\beta) d\alpha d\beta$$

In general, for some region M, the joint statistics are

$$\text{Pr}\{(x,y) \in M\} = \iint_M f(x,y) dx dy$$

Marginal Statistics: $F_x(x) = F(x,\infty)$ and $F_y(y) = F(\infty,y)$

$$\Rightarrow f_x(x) = \int_{-\infty}^{\infty} f(x,y) dy$$

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

Independence

Definition (Independence)

Two *RVs* x and y are statistically independent if for arbitrary events (regions) $x \in A$ and $y \in B$,

$$\Pr\{x \in A, y \in B\} = \Pr\{x \in A\}\Pr\{y \in B\}$$

Letting $A = \{x \leq x_0\}$ and $B = \{y \leq y_0\}$, we see x and y are independent iff

$$F_{x,y}(x,y) = F_x(x)F_y(y)$$

and by differentiation

$$f_{x,y}(x,y) = f_x(x)f_y(y)$$

Joint Moments

For *RVs* x and y and function $z = g(x, y)$

$$
\begin{aligned}
E\{z\} &= \int_{-\infty}^{\infty} z f_z(z) dz \\
E\{g(x,y)\} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f(x,y) dx dy
\end{aligned}
$$

Definition (Covariance)

For *RVs* x and y,

$$
\begin{aligned}
C_{xy} &= \mathrm{Cov}(x,y) \\
&= E[(x-\eta_x)(y-\eta_y)] \\
&= E[xy] - \eta_x E[y] - \eta_y E[x] + \eta_x \eta_y \\
&= E[xy] - \eta_x \eta_y
\end{aligned}
$$

Definition (Correlation Coefficient)

The correlation coefficient is given by

$$r = \frac{C_{xy}}{\sigma_x \sigma_y}$$

Note that

$$\begin{aligned} 0 &\leq E\{[a(x-\eta_x)+(y-\eta_y)]^2\} \\ &= E\{(x-\eta_x)^2\}a^2 + 2E\{(x-\eta_x)(y-\eta_y)\}a + E\{(y-\eta_y)^2\} \\ &= \sigma_x^2 a^2 + 2C_{xy}a + \sigma_y^2 \end{aligned}$$

This is a positive quadratic function of a

$\Rightarrow$ Roots are imaginary and discriminant is non-positive

$$\begin{aligned} \sqrt{4C_{xy}^2 - 4\sigma_x^2\sigma_y^2} &\rightarrow \text{imaginary} \\ \Rightarrow 4C_{xy}^2 - 4\sigma_x^2\sigma_y^2 &\leq 0 \\ \Rightarrow C_{xy}^2 &\leq \sigma_x^2\sigma_y^2 \end{aligned}$$

Thus,

$$|C_{xy}| \leq \sigma_x\sigma_y \quad \text{and} \quad |r| = \frac{|C_{xy}|}{\sigma_x\sigma_y} \leq 1$$

Definition (Uncorrelated)

Two *RV*s are uncorrelated if their covariance is zero

$$\begin{aligned} C_{xy} &= 0 \\ \Rightarrow r &= \frac{C_{xy}}{\sigma_x\sigma_y} = 0 \\ &= \frac{E\{xy\} - E\{x\}E\{y\}}{\sigma_x\sigma_y} = 0 \\ \Rightarrow E\{xy\} &= E\{x\}E\{y\} \end{aligned}$$

Thus

$$C_{xy} = 0 \Leftrightarrow E\{xy\} = E\{x\}E\{y\}$$

Result

If x and y are independent, then

$$E\{xy\} = E\{x\}E\{y\}$$

and x and y are uncorrelated

Note: Converse is not true (in general)

- Converse only holds for Gaussian *RVs*
- Independence is a stronger condition than uncorrelated

Definition (Orthogonality)

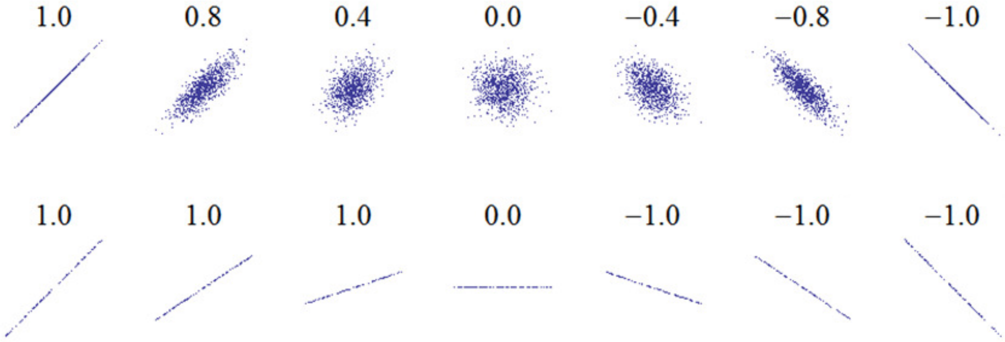
Two *RVs* are orthogonal if

$$E\{xy\} = 0$$

Note: If x and y are correlated, they are not orthogonal

Example

Consider the correlation between two *RV*'s, x and y, with samples shown in a scatter plot



Sequences and Vectors of Random Variables

Definition (Vector Distribution)

Let $\{x\}$ be a sequence of *RVs*. Take N samples to form the random vector

$$\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$$

Then the vector distribution function is

$$\begin{aligned} F_{\mathbf{x}}(\mathbf{x}^0) &= \Pr\{x_1 \le x_1^0, x_2 \le x_2^0, \ldots, x_N \le x_N^0\} \\ &\triangleq \Pr\{\mathbf{x} \le \mathbf{x}^0\} \end{aligned}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} x_{r1} \\ x_{r2} \\ \vdots \\ x_{rN} \end{bmatrix}$$

The distribution in the complex case is defined as

$$\begin{aligned} F_{\mathbf{x}}(\mathbf{x}^0) &= \Pr\{\mathbf{x}_r \leq \mathbf{x}_r^0, \mathbf{x}_i \leq \mathbf{x}_i^0\} \\ &\triangleq \Pr\{\mathbf{x} \leq \mathbf{x}^0\} \end{aligned}$$

The density function is given by

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{\partial^N F_{\mathbf{x}(\mathbf{x})}}{\partial x_1 \partial x_2 \ldots \partial x_N}$$

$$F_{\mathbf{x}}(\mathbf{x}^0) = \int_{-\infty}^{\mathbf{x}_1^0} \int_{-\infty}^{\mathbf{x}_2^0} \cdots \int_{-\infty}^{\mathbf{x}_N^0} f_{\mathbf{x}}(\mathbf{x}) dx_1 dx_2 \ldots dx_N$$

Properties:

$$
\begin{aligned}
F_{\mathbf{x}}([\infty, \infty, \cdots, \infty]^T) &= 1 \\
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} &= 1 \\
F_{\mathbf{x}}([x_1, x_2, \cdots, -\infty, \cdots, x_N]^T) &= 0
\end{aligned}
$$

Also

$$
\begin{aligned}
F([\infty, x_2, x_3, \cdots, x_N]^T) &= F([x_2, x_3, \cdots, x_N]^T) \\
\int_{-\infty}^{\infty} f([x_1, x_2, x_3, \cdots, x_N]^T) dx_1 &= f([x_2, x_3, \cdots, x_N]^T)
\end{aligned}
$$

- Setting $x_i = \infty$ in the cdf eliminates this sample
- Integrating over $(-\infty, \infty)$ along x_i in the pdf eliminates this sample

Expectations & Moments

Objective: Obtain partial description of process generating $\mathbf{x}$

Solution: Use moments

The first moment, or mean, is

$$\mathbf{m_x} = E\{\mathbf{x}\} = [m_1, m_2, \ldots, m_N]^T = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \mathbf{x} f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}$$

$$\Rightarrow m_k = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_k f_{\mathbf{x}}(\mathbf{x}) dx_1 dx_2 \cdots dx_N = \int_{-\infty}^{\infty} x_k \underbrace{f_{x_k}(x_k)} dx_k$$

$\Uparrow$ marginal distribution of x_k

Definition (Correlation Matrix)

A complete set of second moments is given by the correlation matrix

$$\begin{aligned}\mathbf{R_x} &= E\{\mathbf{x}\mathbf{x}^H\} = E\{\mathbf{x}\mathbf{x}^{*^T}\} \\ &= \begin{bmatrix} E\{|x_1|^2\} & E\{x_1x_2^*\} & \cdots & E\{x_1x_N^*\} \\ E\{x_2x_1^*\} & E\{|x_2|^2\} & \cdots & E\{x_2x_N^*\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{x_Nx_1^*\} & E\{x_Nx_2^*\} & \cdots & E\{|x_N|^2\} \end{bmatrix}\end{aligned}$$

Result

The correlation matrix is Hermitian symmetric

$$\begin{aligned}(\mathbf{R_x})^H &= (E\{\mathbf{x}\mathbf{x}^H\})^H \\ &= E\{(\mathbf{x}\mathbf{x}^H)^H\} \\ &= E\{\mathbf{x}\mathbf{x}^H\} = \mathbf{R_x}\end{aligned}$$

Definition (Covariance Matrix)

The set of second central moments is given by the covariance

$$\begin{aligned}\mathbf{C_x} &= E\{(\mathbf{x}-\mathbf{m_x})(\mathbf{x}-\mathbf{m_x})^H\} \\ &= E\{\mathbf{x}\mathbf{x}^H\} - \mathbf{m_x}E\{x^H\} - E\{\mathbf{x}\}\mathbf{m_x}^H + \mathbf{m_x}\mathbf{m_x}^H \\ &= \mathbf{R_x} - \mathbf{m_x}\mathbf{m_x}^H\end{aligned}$$

Result

The covariance is Hermitian symmetric

$$\mathbf{C_x} = \mathbf{C_x}^H$$

Result

The correlation and covariance matrices are positive semi-definite

$$\mathbf{a}^H\mathbf{R_x a} \geq 0 \quad \mathbf{a}^H\mathbf{C_x a} \geq 0 \quad (\forall \mathbf{a})$$

To prove this, note

$$\begin{aligned}
\mathbf{a}^H\mathbf{R_x a} &= \mathbf{a}^H E\{\mathbf{x}\mathbf{x}^H\}\mathbf{a} \\
&= E\{\mathbf{a}^H\mathbf{x}\mathbf{x}^H\mathbf{a}\} \\
&= E\{(\mathbf{a}^H\mathbf{x})(\mathbf{a}^H\mathbf{x})^H\} \\
&= E\{|\mathbf{a}^H\mathbf{x}|^2\} \geq 0
\end{aligned}$$

For most cases, $\mathbf{R}$ and $\mathbf{C}$ are positive define

$$\mathbf{a}^H\mathbf{R_x a} > 0 \quad \mathbf{a}^H\mathbf{C_x a} > 0$$

$\Rightarrow$ no linear dependencies in $\mathbf{R_x}$ or $\mathbf{C_x}$

Definitions (Cross-Correlation and Cross-Covariance)

For random vectors $\mathbf{x}$ and $\mathbf{y}$,

$$\begin{aligned} \text{Cross-correlation} \triangleq \mathbf{R_{xy}} &= E\{\mathbf{x}\mathbf{y}^H\} \\ \text{Cross-covariance} \triangleq \mathbf{C_{xy}} &= E\{(\mathbf{x}-\mathbf{m_x})(\mathbf{y}-\mathbf{m_y})^H\} \\ &= \mathbf{R_{xy}} - \mathbf{m_x}\mathbf{m_y}^H \end{aligned}$$

Definition (Uncorrelated Vectors)

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated if

$$\mathbf{C_{xy}} = \mathbf{R_{xy}} - \mathbf{m_x}\mathbf{m_y}^H = 0$$

or equivalently

$$\mathbf{R_{xy}} = E\{\mathbf{x}\mathbf{y}^H\} = \mathbf{m_x}\mathbf{m_y}^H$$

Note that as in the scalar case

$$\begin{aligned} \text{independence} &\Rightarrow \text{uncorrelated} \\ \text{uncorrelated} &\not\Rightarrow \text{independence} \end{aligned}$$

Also, $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if

$$\mathbf{R}_{\mathbf{xy}} = E\{\mathbf{x}\mathbf{y}^H\} = \mathbf{0}$$

Example

Let $\mathbf{x}$ and $\mathbf{y}$ be the same dimension. If

$$\mathbf{z} = \mathbf{x} + \mathbf{y}$$

find $\mathbf{R}_{\mathbf{z}}$ and $\mathbf{C}_{\mathbf{z}}$

By definition

$$\begin{aligned}\mathbf{R_z} &= E\{(\mathbf{x}+\mathbf{y})(\mathbf{x}+\mathbf{y})^H\} \\ &= E\{\mathbf{x}\mathbf{x}^H\}+E\{\mathbf{x}\mathbf{y}^H\}+E\{\mathbf{y}\mathbf{x}^H\}+E\{\mathbf{y}\mathbf{y}^H\} \\ &= \mathbf{R_x}+\mathbf{R_{xy}}+\mathbf{R_{yx}}+\mathbf{R_y}\end{aligned}$$

Similarly

$$\mathbf{C_z} = \mathbf{C_x}+\mathbf{C_{xy}}+\mathbf{C_{yx}}+\mathbf{C_y}$$

Note: If $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated,

$$\mathbf{R_z} = \mathbf{R_x}+\mathbf{m_x}\mathbf{m_y}^H+\mathbf{m_y}\mathbf{m_x}^H+\mathbf{R_y}$$

and

$$\mathbf{C_z} = \mathbf{C_x}+\mathbf{C_y}$$

Definition (Multivariate Gaussian Density)

For a N dimensional random vector $\mathbf{x}$ with covariance $\mathbf{C_x}$, the multivariate Gaussian pdf is

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{N}{2}}|\mathbf{C_x}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m_x})^H \mathbf{C_x}^{-1}(\mathbf{x}-\mathbf{m_x})}$$

Note the similarity to the univariate case

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}}$$

Example

Let $N = 2$ (bivariate case) and $\mathbf{x}$ be real. Then

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{and} \quad \mathbf{m_x} = E\{\mathbf{x}\} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$$

$$\begin{aligned}
\mathbf{C_x} &= E\{(\mathbf{x}-\mathbf{m_x})(\mathbf{x}-\mathbf{m_x})^T\} \\
&= E\{\mathbf{x}\mathbf{x}^T\} - \mathbf{m_x}\mathbf{m_x}^T \\
&= E\left\{\begin{bmatrix} x_1^2 & x_1x_2 \\ x_2x_1 & x_2^2 \end{bmatrix}\right\} - \begin{bmatrix} m_1^2 & m_1m_2 \\ m_2m_1 & m_2^2 \end{bmatrix} \\
&= \begin{bmatrix} E\{x_1^2\} - m_1^2 & E\{x_1x_2\} - m_1m_2 \\ E\{x_2x_1\} - m_2m_1 & E\{x_2^2\} - m_2^2 \end{bmatrix}
\end{aligned}$$

Recall that

$$\sigma_x^2 = E\{x^2\} - E^2\{x\}$$

and

$$r = \frac{E\{x_1x_2\} - m_1m_2}{\sigma_{x_1}\sigma_{x_2}}$$

Rearranging: $\mathbf{C_x} = \begin{bmatrix} \sigma_{x_1}^2 & r\sigma_{x_1}\sigma_{x_2} \\ r\sigma_{x_1}\sigma_{x_2} & \sigma_{x_2}^2 \end{bmatrix}$

Also,

$$\begin{aligned} \mathbf{C_x}^{-1} &= \frac{1}{\sigma_{x_1}^2\sigma_{x_2}^2 - r^2\sigma_{x_1}^2\sigma_{x_2}^2} \begin{bmatrix} \sigma_{x_2}^2 & -r\sigma_{x_1}\sigma_{x_2} \\ -r\sigma_{x_1}\sigma_{x_2} & \sigma_{x_1}^2 \end{bmatrix} \\ &= \frac{1}{\sigma_{x_1}^2\sigma_{x_2}^2(1-r^2)} \begin{bmatrix} \sigma_{x_2}^2 & -r\sigma_{x_1}\sigma_{x_2} \\ -r\sigma_{x_1}\sigma_{x_2} & \sigma_{x_1}^2 \end{bmatrix} \end{aligned}$$

Substituting into the Gaussian pdf and simplifying

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}) &= \frac{1}{2\pi|\mathbf{C_x}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m_x})^T \mathbf{C_x}^{-1}(\mathbf{x}-\mathbf{m_x})} \\ &= \frac{1}{2\pi\sigma_{x_1}\sigma_{x_2}(1-r^2)^{\frac{1}{2}}} e^{-\frac{1}{2(1-r^2)}\left[\frac{(x_1-m_1)^2}{\sigma_{x_1}^2} - 2r\frac{(x_1-m_1)(x_2-m_2)}{\sigma_{x_1}\sigma_{x_2}} + \frac{(x_2-m_2)^2}{\sigma_{x_2}^2}\right]} \end{aligned}$$

Note: If uncorrelated, $r = 0$

$$\Rightarrow f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{2\pi\sigma_{x_1}\sigma_{x_2}} e^{-\frac{1}{2}\left[\frac{(x_1-m_1)^2}{\sigma_{x_1}^2}+\frac{(x_2-m_2)^2}{\sigma_{x_2}^2}\right]}$$
$$= f_{x_1}(x_1) f_{x_2}(x_2)$$

Gaussian special case result:

$$\text{uncorrelated} \Rightarrow \text{independent}$$

Example

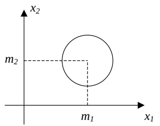
Examine the contours defined by

$$(\mathbf{x}-\mathbf{m}_{\mathbf{x}})^T \mathbf{C}_{\mathbf{x}}{}^{-1}(\mathbf{x}-\mathbf{m}_{\mathbf{x}}) = \text{constant}$$

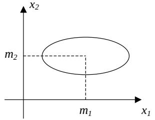
Why? For all values on the contour

$$f_{\mathbf{x}}(\mathbf{x}) = \text{constant}$$

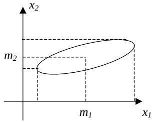
$r = 0 \quad \sigma_{x_1} = \sigma_{x_2}$



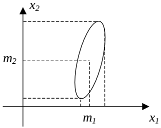
$r = 0 \quad \sigma_{x_1} > \sigma_{x_2}$



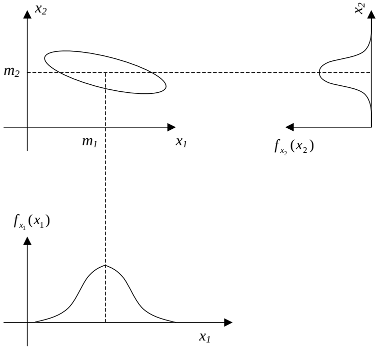
$r > 0 \quad \sigma_{x_1} > \sigma_{x_2}$



$r > 0 \quad \sigma_{x_1} < \sigma_{x_2}$

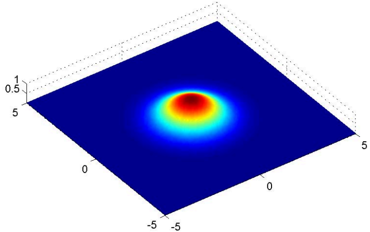


$r < 0 \quad \text{and} \quad \sigma_{x_1} > \sigma_{x_2}$

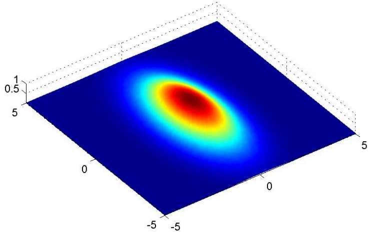


- Integrating over x_2 yields $f_{x_1}(x_1)$
- Integrating over x_1 yields $f_{x_2}(x_2)$

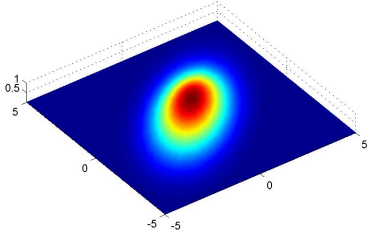
Additional Gaussian (surface) examples:



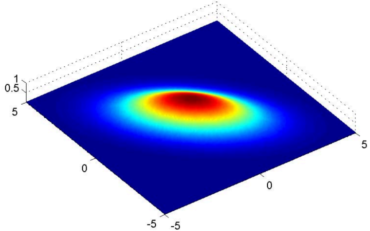
$r = 0 \quad \sigma_{x_1} = \sigma_{x_2}$



$r = 0 \quad \sigma_{x_1} < \sigma_{x_2}$



$r > 0 \quad \sigma_{x_1} < \sigma_{x_2}$



$r < 0 \quad \sigma_{x_1} < \sigma_{x_2}$

Tchebycheff Inequality

For any $\epsilon > 0$,

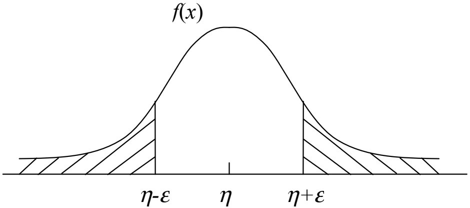
$$\Pr(|x-\eta| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

To prove this, note

$$\begin{aligned} \Pr(|x-\eta| \geq \epsilon) &= \int_{-\infty}^{\eta-\epsilon} f(x)dx + \int_{\eta+\epsilon}^{\infty} f(x)dx \\ &= \int_{|x-\eta|\geq\epsilon} f(x)dx \end{aligned}$$

Also note that

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} (x-\eta)^2 f(x)dx \\ &\geq \int_{|x-\eta|\geq\epsilon} (x-\eta)^2 f(x)dx \end{aligned}$$



$$\sigma^2 \geq \int_{|x-\eta|\geq\epsilon} (x-\eta)^2 f(x)dx$$

Using the fact that $|x-\eta| \geq \epsilon$ in the above gives

$$\begin{aligned}\sigma^2 &\geq \epsilon^2 \int_{|x-\eta|\geq\epsilon} f(x)dx \\ &= \epsilon^2 \mathrm{Pr}\{|x-\eta| \geq \epsilon\}\end{aligned}$$

Rearranging gives the desired result

$$\Rightarrow \mathrm{Pr}\{|x-\eta| \geq \epsilon\} \leq \left(\frac{\sigma}{\epsilon}\right)^2$$

QED

Markov's Inequality

If x is a non-negative RV, then for all $a > 0$

$$\Pr\{x \geq a\} \leq \frac{E\{x\}}{a}.$$

Proof:

$$\begin{aligned}
\Pr\{x \geq a\} &= \int_a^\infty f(x)dx \\
&\leq \int_a^\infty \frac{x}{a} f(x)dx \qquad \text{since } x \geq a \\
&\leq \frac{1}{a}\int_0^\infty x f(x)dx \\
&= \frac{E\{x\}}{a}.
\end{aligned}$$

Chernoff's Bounding Method

Let x be a RV on $\mathbb{R}$. Then for all $\epsilon > 0$

$$\text{Pr}\{x \geq \epsilon\} \leq \min_{s>0} \quad e^{-s\epsilon} E\{e^{sx}\}.$$

To prove this for any $s > 0$:

$$\begin{aligned} \text{Pr}\{x \geq \epsilon\} &= \text{Pr}\{sx \geq s\epsilon\} \\ &= \text{Pr}\{e^{sx} \geq e^{s\epsilon}\} \end{aligned}$$

Using Markov's Inequality:

$$\begin{aligned} \text{Pr}\{x \geq \epsilon\} = \text{Pr}\{e^{sx} \geq e^{s\epsilon}\} &\leq \frac{E\{e^{sx}\}}{e^{s\epsilon}} \\ &= e^{-s\epsilon} E\{e^{sx}\}. \end{aligned}$$

Hoeffding's Inequality

Consider $S_N = \sum_{i=1}^{N} x_i$ where $x_1, \ldots, x_N$ are independent RV's on $\mathbb{R}$ such that $a_i \leq x_i \leq b_i$. Then, for any $\epsilon > 0$

$$\Pr\{|S_N - E\{S_N\}| \geq \epsilon\} \quad \leq \quad 2e^{-2\epsilon^2/\sum(b_i - a_i)^2}$$

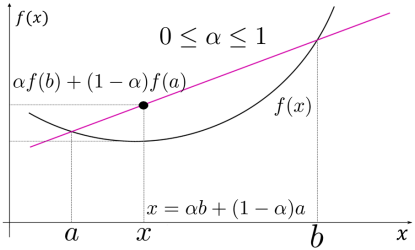
Proof:

First, demonstrate that if $E\{x\} = 0$ then $E\{e^{sx}\} \leq e^{s^2(b-a)^2/8}$ for any $s > 0$.
If $x \in [a, b]$ then the convexity of the function $f(x) = e^{sx}$ implies that

$$e^{sx} \leq \alpha f(b) + (1-\alpha) f(a),$$

$$e^{sx} \leq \alpha e^{sb} + (1-\alpha)e^{sa}, \quad \text{since } \alpha = \frac{x-a}{b-a},$$

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}$$



Hoeffding's Inequality

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}$$

Using the fact that $E\{x\} = 0$ we obtain:

$$\begin{aligned} E\{e^{sx}\} &\leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}, \\ &= e^{sa}\left(\frac{b}{b-a} - \frac{a}{b-a}e^{s(b-a)}\right), \text{ since } y = e^{\ln(y)} \\ &= e^{\ln\left[e^{sa}\left(\frac{b-ae^{s(b-a)}}{b-a}\right)\right]} \end{aligned}$$

Thus,

$$E\{e^{sx}\} \leq e^{g(s)}$$

where $g(s) = sa + \ln(b - ae^{s(b-a)}) - \ln(b-a)$.

Hoeffding's Inequality

$$g(s) = sa + \ln(b - ae^{s(b-a)}) - \ln(b-a)$$

By Taylor's theorem:

$$g(s) = g(0) + g'(0)s + \frac{1}{2!}g''(\xi)s^2, \quad 0 \leq \xi \leq s$$

$$g(0) = 0, \qquad g'(0) = 0, \qquad g''(\xi) \leq \frac{(b-a)^2}{4}$$

Substituting, we get: $g(s) \leq \frac{s^2(b-a)^2}{8}$.

Substituting in previous demonstration (i.e. $E\{e^{sx}\} \leq e^{g(s)}$) :

$$\Longrightarrow \boxed{E\{e^{sx}\} \leq e^{s^2(b-a)^2/8}} \qquad (*)$$

Hoeffding's Inequality

Second, apply Chernoff's bounding method i.e.:

$$\Pr\{x \geq \epsilon\} \leq \min_{s>0} \quad e^{-s\epsilon} E\{e^{sx}\})$$

to the random variable: $S_N - E\{S_N\}$,

$$\begin{aligned}
\Pr\{S_N - E\{S_N\} \geq \epsilon\} &\leq \min_{s>0} \quad e^{-s\epsilon} E\left\{e^{s(S_N - E\{S_N\})}\right\} \\
&\leq \min_{s>0} \quad e^{-s\epsilon} E\left\{e^{s\left(\sum_{i=1}^{N}(x_i - E\{x_i\})\right)}\right\} \\
\text{since the } x_i \text{ are independent} \qquad &\leq \min_{s>0} \quad e^{-s\epsilon} \prod_{i=1}^{N} E\{e^{s(x_i - E\{x_i\})}\}
\end{aligned}$$

Applying our first result $(*)$ to $y_i = x_i - E\{x_i\}$ where $E\{y_i\} = 0$:

$$E\{e^{s(x_i - E\{x_i\})}\} \leq e^{s^2(b_i - a_i)^2/8}$$

Hoeffding's Inequality

Substitute $E\{e^{s(x_i-E\{x_i\})}\} \leq e^{s^2(b_i-a_i)^2/8}$ in the previous Chernoff's bound:

$$\Pr\{S_N - E\{S_N\} \geq \epsilon\} \quad \leq \quad \min_{s>0} \quad e^{-s\epsilon} \prod_{i=1}^{N} E\{e^{s(x_i-E\{x_i\})}\}$$

we get:

$$\begin{aligned}\Pr\{S_N - E\{S_N\} \geq \epsilon\} \quad &\leq \quad \min_{s>0} e^{-s\epsilon} \prod_{i=1}^{N} e^{s^2(b_i-a_i)^2/8} \\ &= \quad \min_{s>0} e^{-s\epsilon+\sum_{i=1}^{N}(s^2/8)(b_i-a_i)^2}\end{aligned}$$

It can be shown that the minimum is at $s = 4\epsilon/\sum(b_i-a_i)^2$.

Hoeffding's Inequality

$$\Pr\{S_N - E\{S_N\} \geq \epsilon\} \leq e^{-s\epsilon + \sum_{i=1}^{N}(s^2/8)(b_i - a_i)^2}$$

Substituting the minimum ($s = 4\epsilon / \sum_{i=1}^{N}(b_i - a_i)^2$):

$$\Pr\{S_N - E\{S_N\} \geq \epsilon\} \quad \leq \quad e^{-2\epsilon^2/\sum_{i=1}^{N}(b_i - a_i)^2}$$

If we consider $-x_1, ..., -x_N$ instead, we obtain:

$$\Pr\{S_N - E\{S_N\} \leq -\epsilon\} \leq e^{-2\epsilon^2/\sum_{i=1}^{N}(b_i - a_i)^2}$$

By combining the two bounds, we finish the proof:

$$\Pr\{|S_N - E\{S_N\}| \geq \epsilon\} \quad \leq \quad 2e^{-2\epsilon^2/\sum_{i=1}^{N}(b_i - a_i)^2}$$

Hoeffding's Inequality

Example:

Find the Hoeffding's Inequality of a random variable $x_i \overset{\text{i.i.d.}}{\sim}$ Ber(p).

Solution:

Consider the Hoeffding's Inequality:

$$\Pr\{|S_N - E\{S_N\}| \geq \epsilon\} \leq 2e^{-2\epsilon^2/\sum_{i=1}^{N}(b_i-a_i)^2}$$

Since $x_i \sim$ Ber(p), then $a_i = 0$, $b_i = 1$, $S_N = \sum_{i=1}^{N} x_i \sim Bin(N,p)$, and $E\{S_N\} = Np$. Taking $\epsilon = N\delta$ and applying Hoeffding's Inequality:

$$\Pr\left\{\left|\sum_{i=1}^{N} x_i - Np\right| \geq N\delta\right\} \leq 2e^{-2(N\delta)^2/\sum_{i=1}^{N}(1-0)^2}$$

$$\Pr\left\{\left|\frac{1}{N}\sum_{i=1}^{N} x_i - p\right| \geq \delta\right\} \leq 2e^{-2N\delta^2}$$

Hoeffding's Inequality

$$\Pr\left\{\left|\frac{1}{N}\sum_{i=1}^{N} x_i - p\right| \geq \delta\right\} \leq 2e^{-2N\delta^2}$$

$$\Pr\{|\nu - \mu| \geq \delta\} \leq 2e^{-2N\delta^2}$$