Review of Probability
(for Lossless Section)

For Details See:

Appendix A in text book

Ch. 10 in Lathi's Book

Probability

Motivate with Frequency of Occurrence Viewpoint

Consider N Events: $\omega_1, \omega_2, \ldots \omega_N$

Conduct Experiment n_T times…
and let n_i = # of times event ω_i occurred.

Then we can "roughly define" the probability as $P(\omega_i) = \dfrac{n_i}{n_T}$

We know that the law of large numbers implies that this rough definition will converge to the true probability as $n_T \rightarrow \infty$

Example: 6-sided Die $\omega_1 = 1, \omega_2 = 2, \ldots \omega_6 = 6$

From classic Prob. Theory we know that $P(\omega_i) = 1/6$

Also… for sets of events: $P(\omega_i \leq 3) = 1/2$

Axioms of Probability

Rules probability <u>must</u> follow.

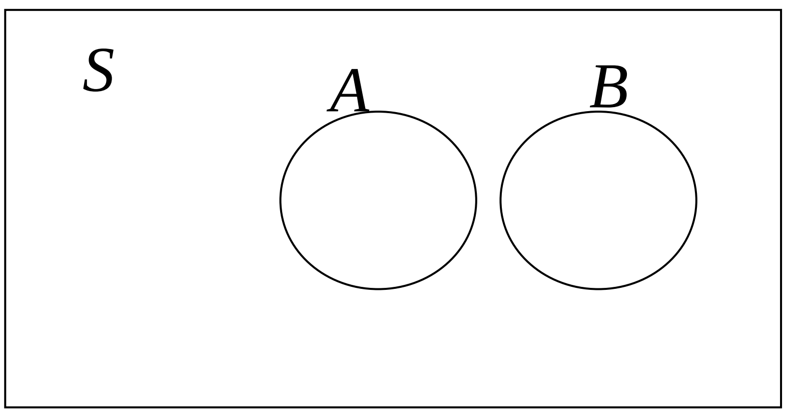
Let S be the set of all possible events

A1: For any event set A, $P(A) \geq 0$

A2: $P(S) = 1$

A3: If $A \cap B = \varnothing, \quad then \quad P(A \cup B) = P(A) + P(B)$

From these:
$0 \leq P(A) \leq 1$



Examples of A3 for 6-sided Die

1. A = {1,2} B = {3}

$$P(A \cup B) = P(\omega_i \leq 2) + P(\omega_i = 3)$$

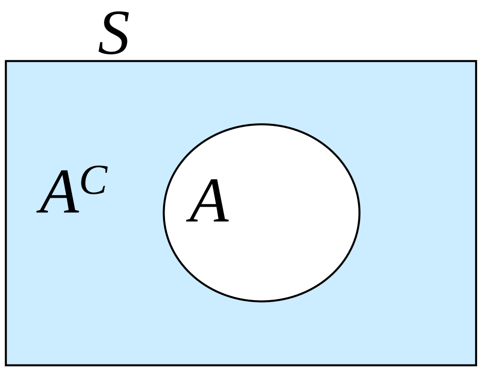
$$= \frac{2}{6} + \frac{1}{6} = \frac{1}{2} \quad \text{as before}$$

2. A = {1,2} B = {3}

$$\underbrace{P(A \cup B)}_{=P(A)} \neq P(A) + P(B)$$

$$P(A) \neq P(A) + \frac{1}{6}$$

Some Properties of Probability



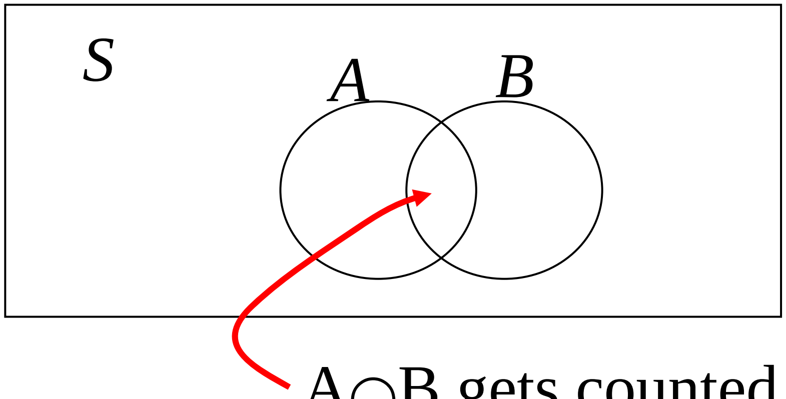
$P1: \quad P(A^C \cup A) = 1$ (Because $S = A \cup A^C$)

$P2: \quad P(A^C \cup A) = P(A^C) + P(A)$

Follows from A3 because $A \cap A^C = \varnothing$

P1 & P2 together give $P(A^C) = 1 - P(A)$

$P3: \quad If \; A \cap B \neq \varnothing, \; then \; P(A \cup B) = P(A) + P(B) - P(A \cap B)$



A∩B gets counted twice…
so subtract off one

Joint Probability

Consider two separate "experiments":

The probability that… the 1^{st} experiment had outcome A

…***AND***… the 2^{nd} experiment had outcome B

is denoted as $P(A,B)$

Often *A* & *B* come from a single experiment having multiple observations…

Experiment: Randomly choose a person

Observations: Height & Weight of chosen person

$P(H > 6', W > 170 \text{ lbs})$ = prob the selected person is taller than 6' ***AND*** weighs more than 170 lbs

Conditional Probability & Independence

Consider two separate observations (from 1 or 2 experiments)

Given that you know what was observed for one of the outcomes, what is the probability that you will get the other outcome??

$P(A|B)$ = probability that you observe A given that B has occured

$$(\star)\quad P(A\,|\,B) = \frac{P(A,B)}{P(B)}$$

"Prob of *A* given *B*"

Note that $P(A|B) \geq P(A,B)$ because $P(B) \leq 1$

Independence: If B provides no information about A, then knowledge of B does not change the probability of observing A:

$$P(A\,|\,B) = P(A)$$

In this case, A & B are called independent events

If *A* & *B* are independent then $P(A,B) = P(A)P(B)$

"Proof ": from $(\star)$ $\quad P(A,B) = \underbrace{P(A\,|\,B)}_{\substack{=P(A) \\ \text{by Indep}}} P(B) = P(A)P(B)$

Prob. vs. Conditional Prob. vs. Joint Prob.

These measure <u>single</u> events — This measures <u>multiple</u> events

We know that $P(A|B) \geq P(A,B)$

What about $P(A)$ vs. $P(A|B)$?

$$P(A) \overset{?}{\substack{> \\ = \\ <}} P(A\,|\,B)$$

We know that if A & B are independent then "="

Otherwise, $P(A|B)$ could be higher or lower than $P(A)$ depending on how B restricts the occurance.

$P(W > 100 \text{ lbs} \mid H < 2')$ is smaller than $P(W > 100 \text{ lbs})$

$P(W > 100 \text{ lbs} \mid H > 7')$ is larger than $P(W > 100 \text{ lbs})$

Example: Prob of Characters in English Text

1. What is the prob. of getting a specific letter?

Most probable letter is e: $P(e) \approx 0.127$

q and z are least probable: $P(q) \approx P(z) \approx 0.001$

2. If you know the current letter… What is the prob. of getting a specific letter in the next position?

Say current letter is q:

Just Guesses!

$P(u|q) \approx 0.99$ $P(e|q) \approx 0.001$

Note:	$P(e\|q) < P(e)$	prior info decreases prob
	$P(u\|q) > P(u)$	prior info increases prob

Knowing the current letter completely redistributes the probability of the next letter (i.e., sequential letters are not independent)

Random Variables (RVs)

Mathematical tool to assign #'s to events

Note: a problem may provide a natural assignment

To each outcome ω_i assign a number $X(\omega_i)$

Examples:
- ASCII code for symbols
- Letter grades get mapped to {0, 1, 2, 3, 4}

Purpose: to allow numerical analyses such as…
- Plots…
- Sums (means, variances)…
- Sets define via inequalities…
- Prob. Functions…
- Etc.

<u>Discrete RVs</u>

For now we will limit ourselves to Discrete RVs

(Later for lossy compression we will need Continuous RVs)

A Discrete RV *X* can take on values only from

- A finite set
- A countably infinite set (e.g., the integers but not the reals)

The finite-set case is the more important one here

<u>Examples</u>

- *X* can take only values in the set {0, 0.5, 1, 1.5, … , 9.5, 10}
- *X* can take only values in the set {0, 1, 2, 3, … }
- An RV *X* that can take any value in the interval [0, 1] is <u>NOT discrete</u>; it is continuous

Probability Function for Discrete RVs

For discrete RV X the probability function is $f_X(x)$, defined as:

$$f_X(x) = P(X = x)$$

$$\sum_x f_X(x) = 1$$

RV symbol… upper case

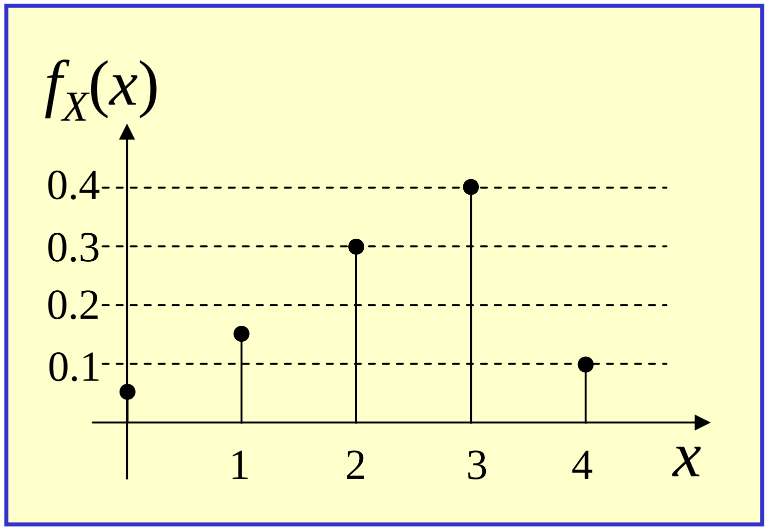
Dummy Variable… lower case

"Prob. That RV X takes on value x"

Example: Let events be letter grades… A, B, C, D, F
RV X maps these to numbers: 4, 3, 2, 1, 0

Assume these probabilities:

$P(X = 0) = 0.05$ $f_X(0) = 0.05$
$P(X = 1) = 0.15$ $f_X(1) = 0.15$
$P(X = 2) = 0.3$ $f_X(2) = 0.3$
$P(X = 3) = 0.4$ $f_X(3) = 0.4$
$P(X = 4) = 0.1$ $f_X(4) = 0.1$



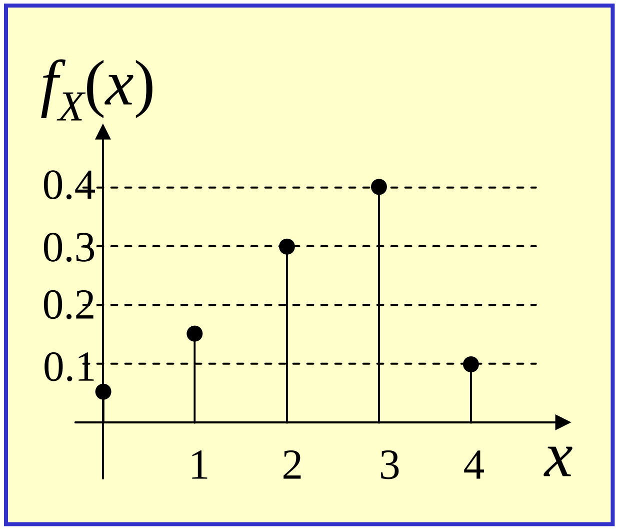
Cumulative Distribution Function (CDF)

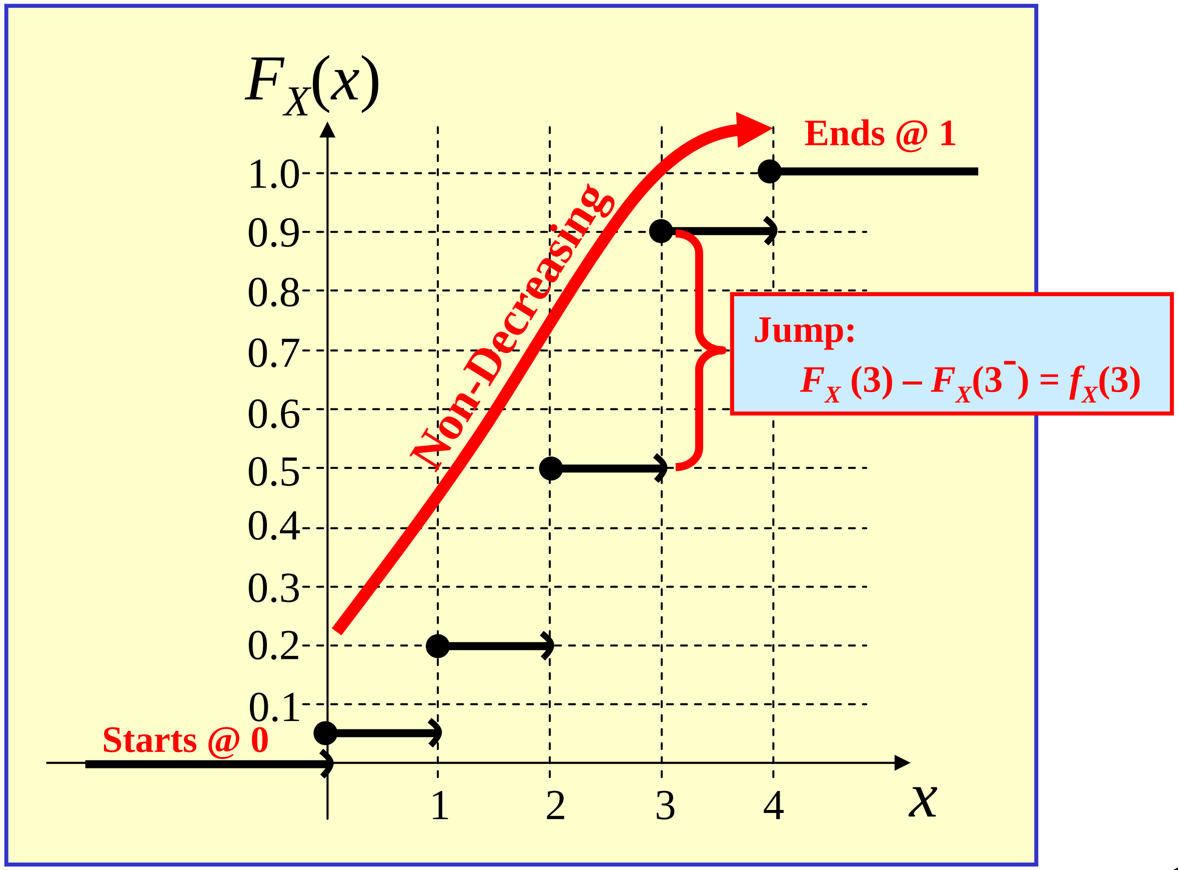
For RV X the CDF $F_X(x)$ is defined as: $F_X(x) = P(X \le x)$

For a discrete RV the CDF and PF are related by: $F_X(x) = \sum_{y=x_{\min}}^{x} f_X(y)$

For Our Example:

$f_X(0) = 0.05$
$f_X(1) = 0.15$
$f_X(2) = 0.3$
$f_X(3) = 0.4$
$f_X(4) = 0.1$





Mean of RV

Mean = Average = Expected Value

Call it E{X}

Motivation First w/ Data Analysis View

Consider RV X = Score on a test   Data: $X_1, X_2, \ldots X_N$

Possible values of X : $V_0 \;\; V_1 \;\; V_2 \ldots \;\; V_{100}$

$0 \quad 1 \quad 2 \;\ldots\; 100$

Test Average

$$= \frac{\sum_{i=1}^{N} X_i}{N} = \frac{N_0 V_0 + N_1 V_1 + N_2 V_2 + \ldots N_n V_{100}}{N} = \sum_{i=0}^{100} V_i \frac{N_i}{N}$$

$\frac{N_i}{N} \approx P(X=V_i)$

N_i = # of scores of value V_i

$N = \sum_{i=1}^{n} N_i$ (Total # of scores)

This is called Data Analysis or Empirical View — Statistics

Theoretical View of Mean

Data Analysis View leads to Probability Theory:

- For Discrete random Variables :

$$\mathrm{E}\{X\} = \sum_{n=1}^{n} x_i \, f_X(x_i)$$

Probability

Notation: $\mathrm{E}\{X\} = \bar{X} = m_X$

Property: $E\{aX + b\} = aE\{X\} + b$

where X is an RV and a and b are just numbers

Aside: Probability vs. Statistics

Probability Theory
- » Given a PDF Model
- » Predict how the data will behave

Statistics
- » Given a set of data
- » Determine how the data did behave

$$E\{X\} = \sum_i x_i \; f_X(x_i)$$

Dummy Variable

PDF

$\approx$

"Law of Large Numbers"

$$Avg = \frac{1}{N}\sum_{i=1}^{n} X_i$$

Data

There is no DATA here!!!
The PDF models how data will behave

There is no PDF here!!!
The Statistic measures how the data did behave

Variance of RV

There are similar Data vs. Theory Views here… Let's go to the theory

Variance measures extent of Deviation Around the Mean

Variance: $\sigma^2 = E\{(X - m_x)^2\}$

$$= \sum_i (x_i - m_x)^2 f_X(x_i)$$

Can show that: $\sigma^2 = E\{X^2\} - \bar{X}^2$

Note : If zero mean… $\sigma^2 = E\{X^2\}$

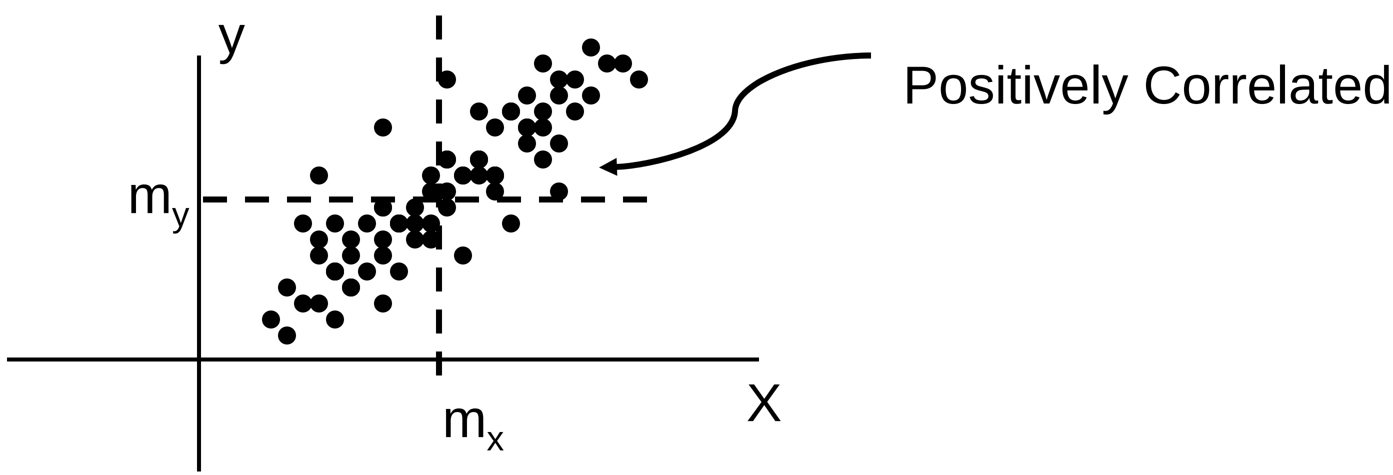
$$= \sum_i x_i^2 f_X(x)$$

Correlation Between RV's

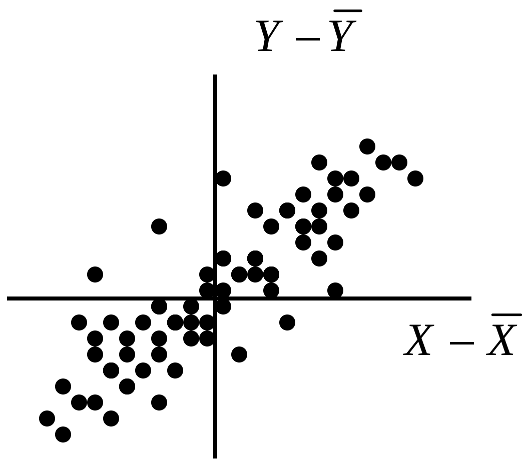
Motivation First w/ Data Analysis View

Consider a random experiment with two outcomes

⇒ 2 RVs X and Y of height and weight respectively

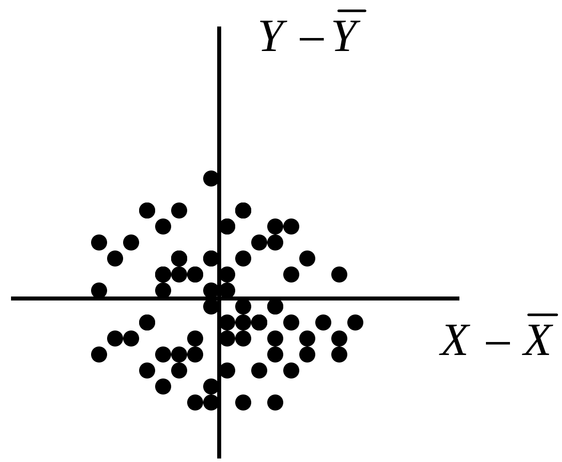


Three main Categories of Correlation



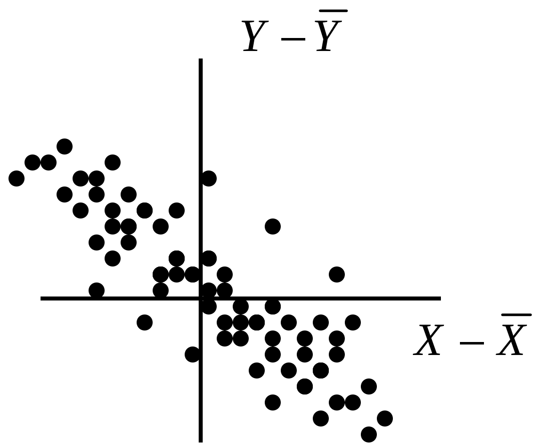
Positive correlation
“Best Friends”

Height
&
Weight



Zero Correlation
i.e. uncorrelated
“Complete Strangers”

Height
&
$ in Pocket



Negative Correlation
“Worst Enemies”

Student Loans
&
Parents’ Salary

Now the Theory…

To capture this, define Covariance :

$$\sigma_{XY} = E\{(X - \overline{X})(Y - \overline{Y})\}$$

$$\sigma_{XY} = \sum_i \sum_j (x_i - \overline{X})(y_j - \overline{Y}) p_{XY}(x_i, y_j)$$

If the RVs are both Zero-mean : $\sigma_{XY} = \mathrm{E}\{XY\}$

If X = Y: $\sigma_{XY} = \sigma_X^2 = \sigma_Y^2$

If X & Y are independent, then: $\sigma_{XY} = 0$

If $\sigma_{XY} = E\{(X - \overline{X})(Y - \overline{Y})\} = 0$

Say that X and Y are "uncorrelated"

If $\sigma_{XY} = E\{(X - \overline{X})(Y - \overline{Y})\} = 0$

Then $\underbrace{E\{XY\}} = \overline{X}\,\overline{Y}$

Called "Correlation of X &Y"

So… RVs X and Y are said to be uncorrelated

if $E\{XY\} = E\{X\}E\{Y\}$

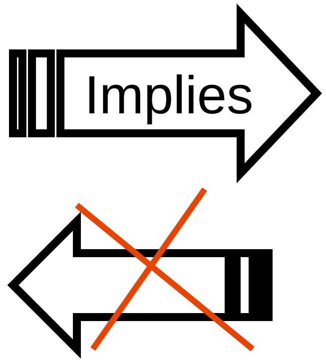
Independence vs. Uncorrelated

X & Y are Independent

$f_{XY}(x,y)$

$= f_X(x) f_Y(y)$

PDFs Separate

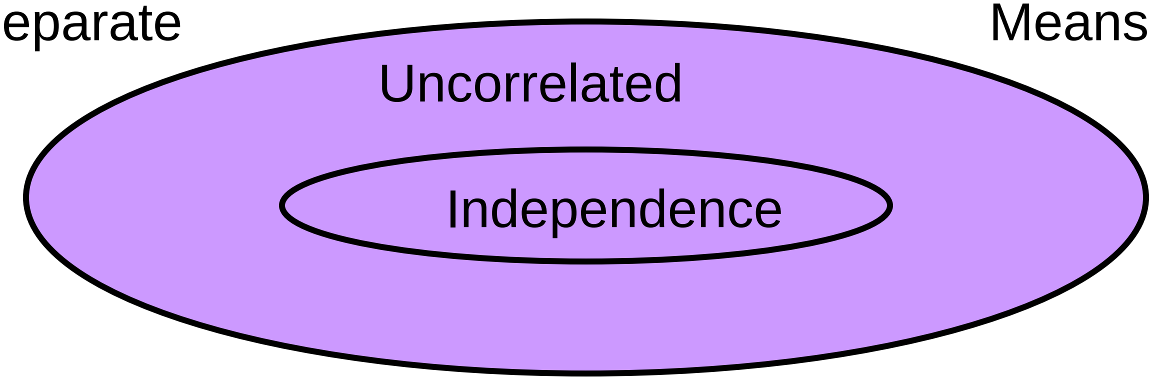


X & Y are Uncorrelated

$E\{XY\}$

$= E\{X\}E\{Y\}$

Means Separate



INDEPENDENCE IS A STRONGER CONDITION !!!!

Confusing Terminology...

Covariance : $$\sigma_{XY} = E\{(X - \overline{X})(Y - \overline{Y})\}$$

Correlation : $$E\{XY\}$$

Same if zero mean

Correlation Coefficient : $$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$$-1 \leq \rho_{XY} \leq 1$$

For Random Vectors…

$$\mathbf{x} = [X_1 \; X_1 \; \cdots \; X_N]^T$$

Correlation Matrix :

$$\mathbf{R_x} = E\{\mathbf{x}\mathbf{x}^T\} = \begin{bmatrix} E\{X_1 X_1\} & E\{X_1 X_2\} & \cdots & E\{X_1 X_N\} \\ E\{X_2 X_1\} & E\{X_2 X_2\} & \cdots & E\{X_2 X_N\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{X_N X_1\} & E\{X_N X_2\} & \cdots & E\{X_N X_N\} \end{bmatrix}$$

Covariance Matrix :

$$\mathbf{C_x} = E\{(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^T\}$$