ONLINE MASTERS IN **DATA SCIENCE**

DSC 257R - UNSUPERVISED LEARNING

MEASURING DEPENDENCE BETWEEN VARIABLES

SANJOY DASGUPTA, PROFESSOR

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Independent Random Variables

Random vars X, Y are **independent** if $\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y)$.

Independent or not? $X, Y \in \{-1, 0, 1\}$, with these probabilities:

		Y		
		-1	0	1
X	-1	0.4	0.16	0.24
	0	0.05	0.02	0.03
	1	0.05	0.02	0.03

Testing Independence

Suppose you are given samples (X, Y) from a bivariate distribution:

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$$

How would you test whether X and Y are independent?

Dependence

Example: For a person chosen at random from a population, take

$$H = \text{height}$$
$$\mathrm{W} = \text{weight}$$

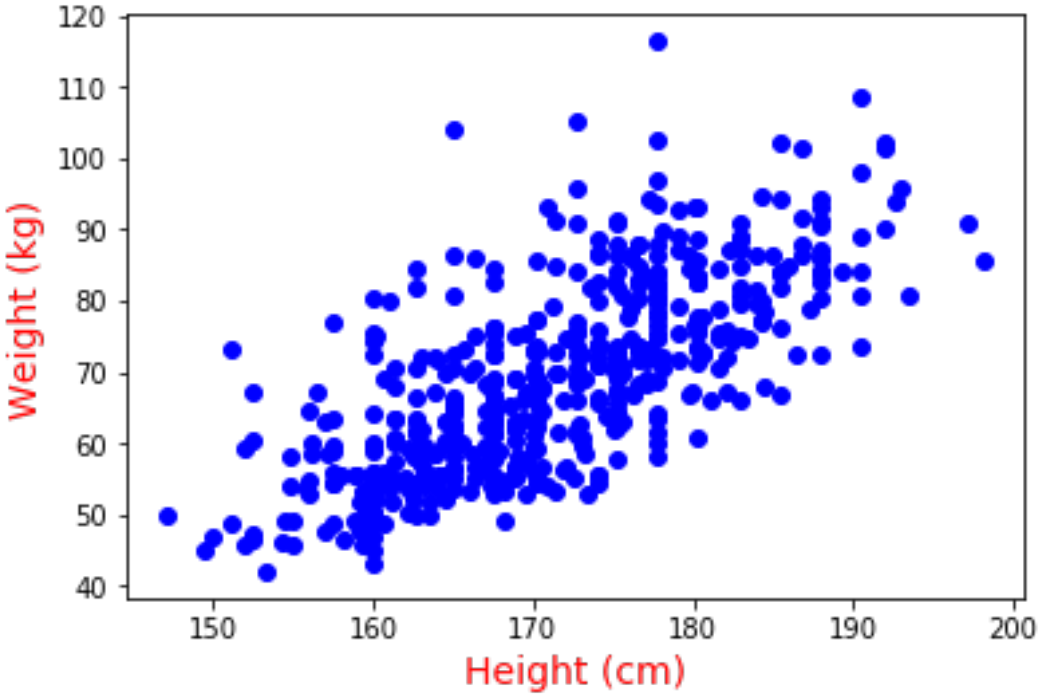
Independence would mean

$$\Pr(H = h, W = w) = \Pr(H = h)\Pr(W = w).$$

This is unlikely to be true. Why?

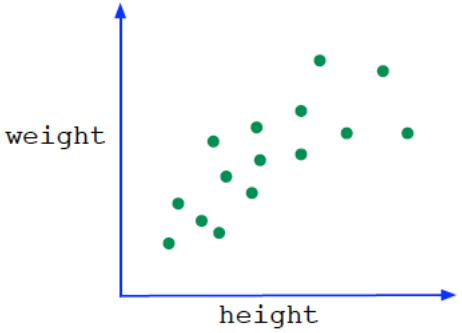
Correlation

Height and weight are **positively correlated.**



Based on body measurements of 507 people at
`https://ww2.amstat.org/publications/jse/datasets/body.txt`

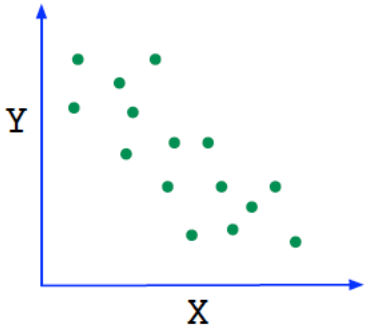
Types of Correlation



H, W **positively correlated**

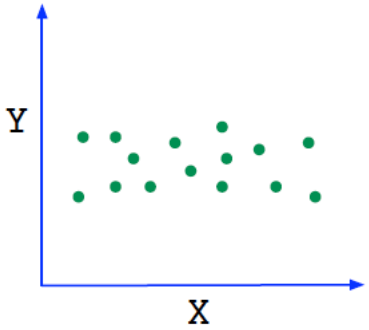
This also implies

$$\mathbb{E}[HW] > \mathbb{E}[H]\ \mathbb{E}[W]$$



X, Y **negatively correlated**

$$\mathbb{E}[XY] < \mathbb{E}[X]\ \mathbb{E}[Y]$$



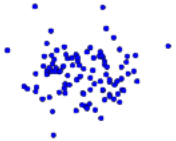
X, Y **uncorrelated**

$$\mathbb{E}[XY] = \mathbb{E}[X]\ \mathbb{E}[Y]$$

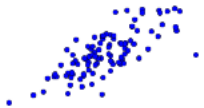
Correlation Coefficient: Pictures

$r = 1$

$r = 0$

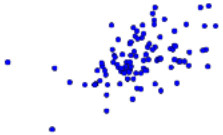


$r = 0.75$

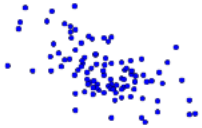


$r = -0.25$

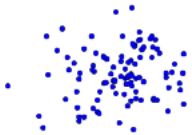
$r = 0.5$



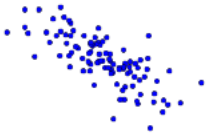
$r = -0.5$



$r = 0.25$



$r = -0.75$



Covariance & Correlation

- **Covariance**

$$cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\ \mathbb{E}[Y]$$

Maximized when $X = Y$, in which case it is $var(X)$.

In general, it is at most $std(X)std(Y)$.

- **Correlation**

$$corr(X, Y) = \frac{cov(X, Y)}{std(X)std(Y)}$$

This is always in the range $[-1, 1]$.

Example

Find $cov(X, Y)$ and $corr(X, Y)$

x	y	$\Pr(x, y)$
1	4	$1/4$
1	-4	$1/4$
-1	4	$1/8$
-1	-4	$3/8$

Independent $\not\equiv$ Uncorrelated