Probability and Statistics: Brief Review

Based on Prof. Ramana Grandhi's Seminar Lecture

• Random Variables & Probability Distributions

A **discrete random variable** is a random variable with a finite (or countably infinite) range.

A **continuous random variable** is a random variable with an interval (either finite or infinite) of real numbers for its range.

- Discrete Random Variable

For a discrete random variable X with possible values $x_1, x_2, \ldots, x_n$, a **probability mass function** is a function such that

(1) $f(x_i) \geq 0$

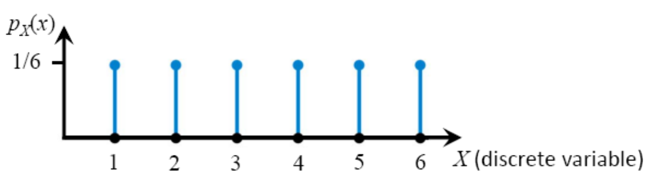
(2) $\sum_{i=1}^{n} f(x_i) = 1$

(3) $f(x_i) = P(X = x_i)$ (3-1)

Probability Mass Function (PMF)

Rolling a Die

outcome	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$	$x=6$	
probability	$p_X(x=1) = 1/6$	$p_X(x=2) = 1/6$	$p_X(x=3) = 1/6$	$p_X(x=4) = 1/6$	$p_X(x=5) = 1/6$	$p_X(x=6) = 1/6$	$\Sigma p_X(x) = 1$



- Discrete Random Variable cont'd

The **cumulative distribution function** of a discrete random variable X, denoted as $F(x)$, is

$$F(x) = P(X \le x) = \sum_{x_i \le x} f(x_i)$$

For a discrete random variable X, $F(x)$ satisfies the following properties.

(1) $F(x) = P(X \le x) = \sum_{x_i \le x} f(x_i)$

(2) $0 \le F(x) \le 1$

(3) If $x \le y$, then $F(x) \le F(y)$ (3-2)

The **mean** or **expected value** of the discrete random variable X, denoted as μ or $E(X)$, is

$$\mu = E(X) = \sum_x x f(x) \tag{3-3}$$

The **variance** of X, denoted as σ^2 or $V(X)$, is

$$\sigma^2 = V(X) = E(X-\mu)^2 = \sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2$$

The **standard deviation** of X is $\sigma = \sqrt{\sigma^2}$.

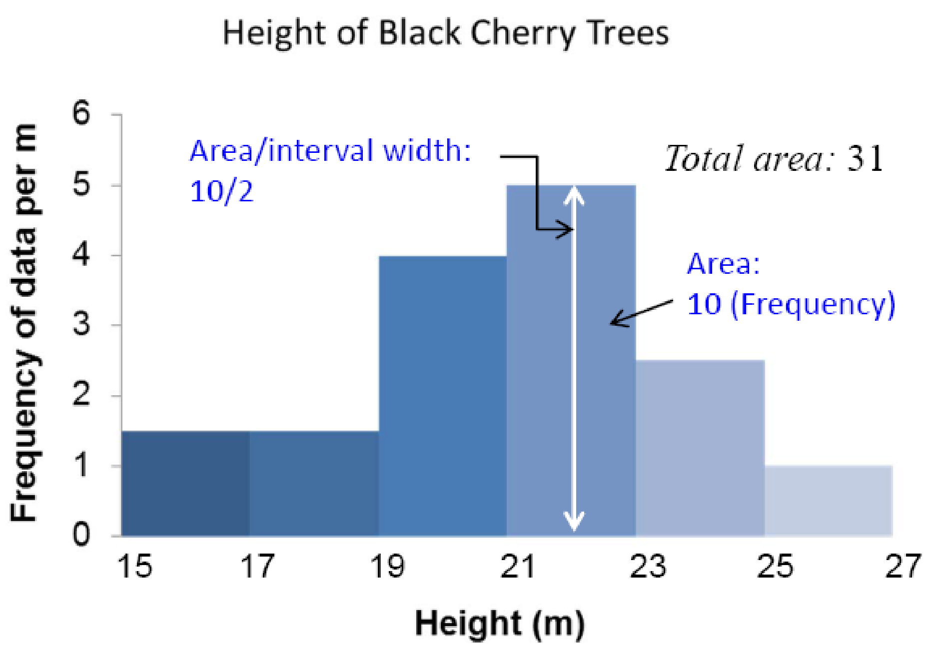
If X is a discrete random variable with probability mass function $f(x)$,

$$E[h(X)] = \sum_x h(x) f(x) \tag{3-4}$$

→ Expected value of a function of a discrete random variable (x)

- Histogram

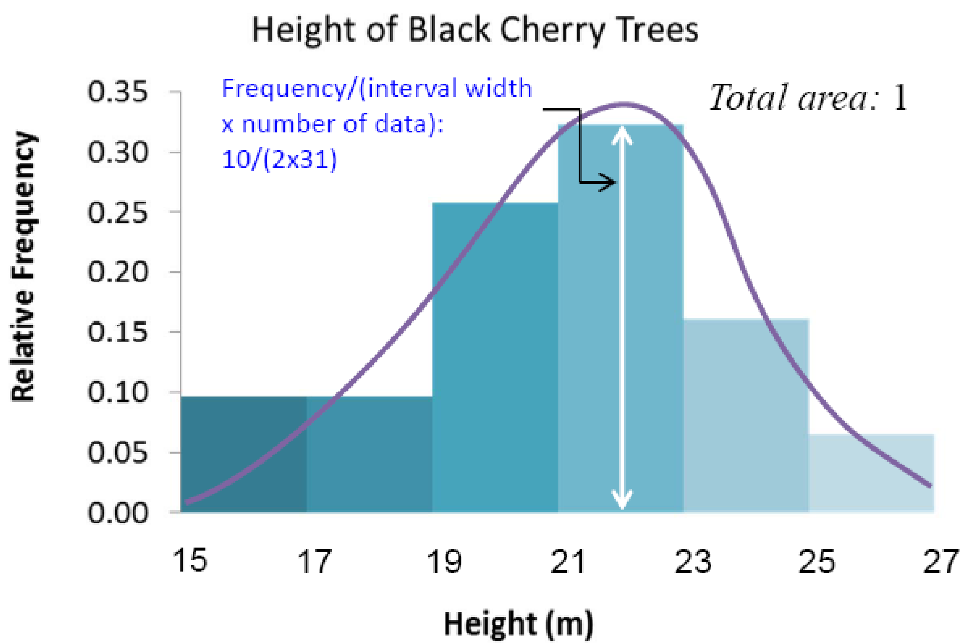
- Graphical representation showing the ***distribution of data***



Height (m)	Frequency of data
15 – 17	3
17 – 19	3
19 – 21	8
21 – 23	**10**
23 – 25	5
25 – 27	2
Σ	*31*

- Normalized Histogram

- A histogram can be ***normalized*** so that the total area of the histogram is 1
- ***Estimate of probability distribution*** of continuous random variable



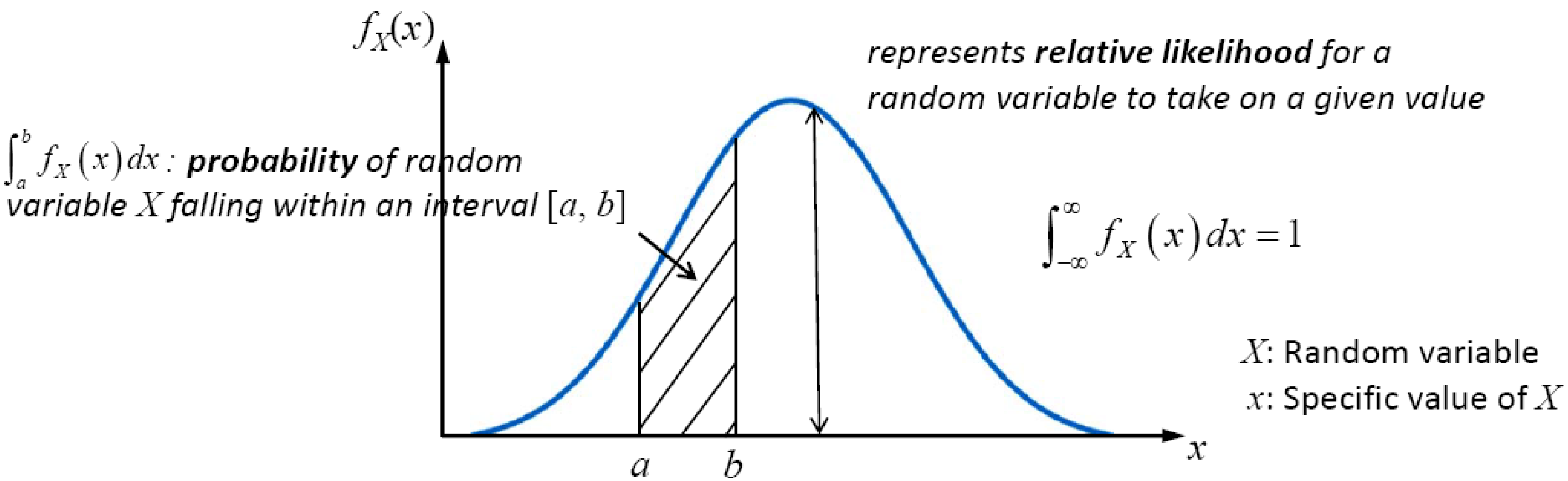
Height (m)	Frequency of data
15 – 17	3
17 – 19	3
19 – 21	8
21 – 23	**10**
23 – 25	5
25 – 27	2
Σ	*31*

If the widths of intervals become ***infinitesimally small*** with an increasing number of data, then a ***continuous curve*** could be drawn

- Continuous Random Variable

Probability Density Function (PDF)

- Function used to describe the probability distribution of a ***continuous random variable***
- The probability that a random variable takes a value over a specific interval is given by the ***integral of PDF***

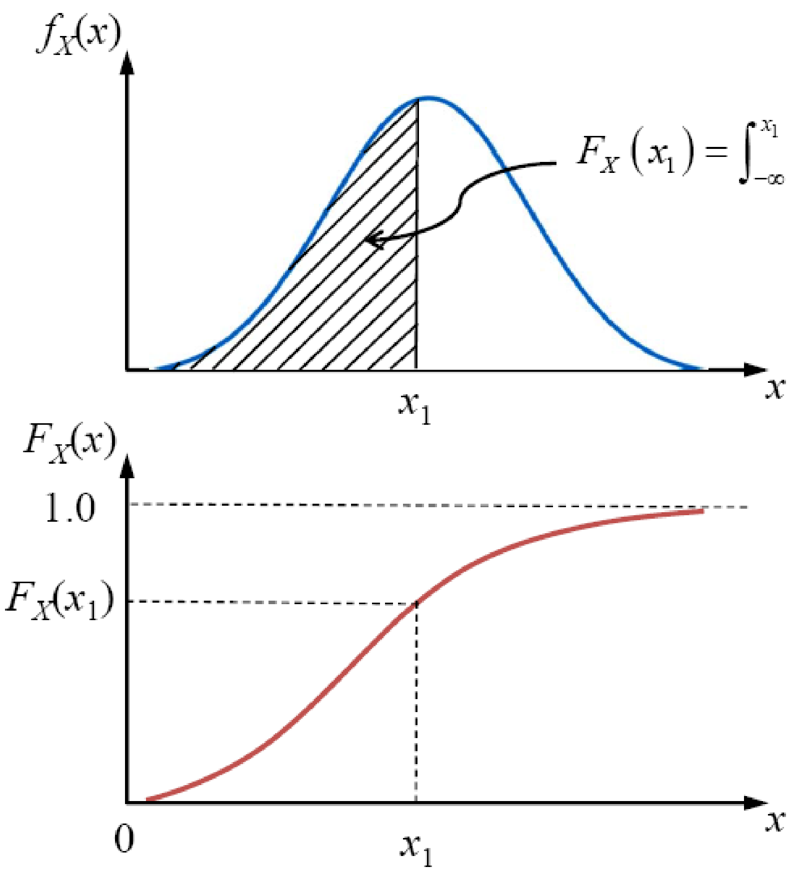


- A specific value of a continuous variable has a relative likelihood
- A specific range of a continuous variable has a probability

- Continuous Random Variable cont'd

Cumulative Distribution Function (CDF)

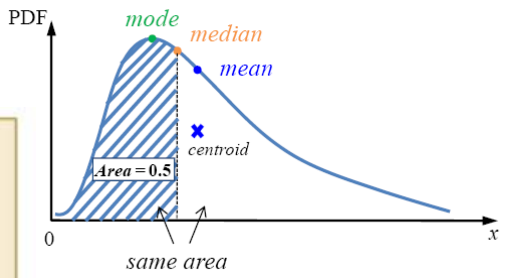
- Describes the probability that a random variable takes on a value less than or equal to a specific value
- Intuitively, a CDF is the "***area so far***" function of the corresponding PDF



PDF: $$f_X(x) = \frac{dF_X(x)}{dx}$$

CDF: $$F_X(x) = \int_{-\infty}^{x} f_X(s)\,ds$$

- Continuous Random Variable cont'd



Mean
Mode
Median
Standard deviation
Skewness,

Suppose that X is a continuous random variable with probability density function $f(x)$. The **mean** or **expected value** of X, denoted as μ or $E(X)$, is

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\, dx \qquad (4\text{-}4)$$

The **variance** of X, denoted as $V(X)$ or σ^2, is

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2$$

The **standard deviation** of X is $\sigma = \sqrt{\sigma^2}$.

If X is a continuous random variable with probability density function $f(x)$,

$$E\left[h(X)\right] = \int_{-\infty}^{\infty} h(x) f(x)\, dx \qquad (4\text{-}5)$$

→ Expected value of a function of a continuous random variable (x)

Chebyshev's inequality:

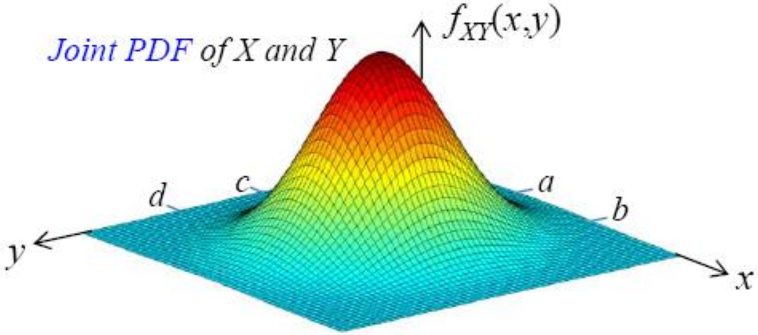
$$P(|X-\mu| > k\sigma) < \frac{1}{k^2}$$

In English: "The probability that the outcome of an experiment with the random variable X will fall more than k standard deviations beyond the mean of X, μ, is less than $\frac{1}{k^2}$."

- Joint PDF and CDF

Joint PDF

- Probability density function of two or more continuous variables

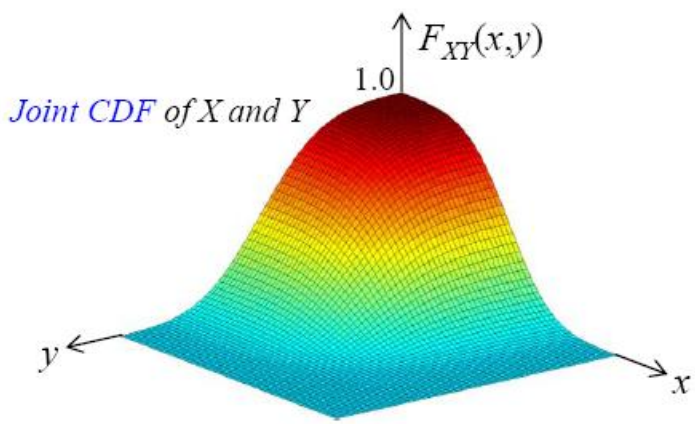


Probability of X and Y falling within a region $[a<X<b,\ c<Y<d]$:

$$P\left[a \le X \le b, c \le Y \le d\right] = \int_c^d \int_a^b f_{XY}(x,y)\,dxdy$$

Joint CDF

- Cumulative distribution function that defines the probability of events defined in terms of two or more variables



Joint CDF of X and Y

$$F_{XY}(x,y) = P\left[-\infty < X \le x, -\infty < Y \le y\right]$$
$$= \int_{-\infty}^{d} \int_{-\infty}^{b} f_{XY}(x,y)\,dxdy$$

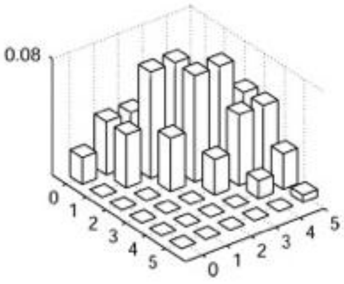
- Joint PDF and CDF cont'd

The **joint probability mass function** of the discrete random variables X and Y, denoted as $f_{xy}(x, y)$, satisfies

(1) $f_{XY}(x,y) \geq 0$

(2) $\sum_X \sum_Y f_{XY}(x,y) = 1$

(3) $f_{XY}(x,y) = P(X = x, Y = y)$ (5-1)



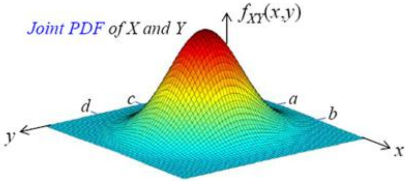
A **joint probability density function** for the continuous random variables X and Y, denoted as $f_{XY}(x,y)$, satisfies the following properties:

(1) $f_{XY}(x,y) \geq 0$ for all x, y

(2) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx\, dy = 1$

(3) For any region R of two-dimensional space,

$$P\left((X,Y) \in R\right) = \iint_R f_{XY}(x,y)\, dx\, dy \qquad (5\text{-}2)$$

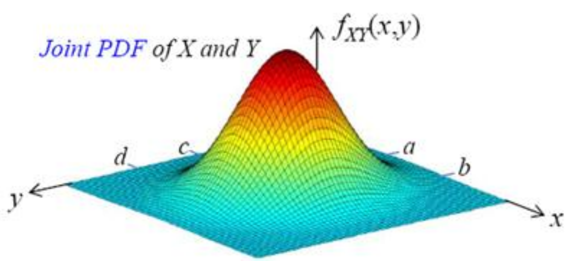


- Marginal & Conditional PDF

If the joint probability density function of random variables X and Y is $f_{XY}(x, y)$, the **marginal probability density functions** of X and Y are

$$f_X(x) = \int f_{XY}(x, y)\,dy \quad \text{and} \quad f_Y(y) = \int f_{XY}(x, y)\,dx \qquad (5\text{-}3)$$

where the first integral is over all points in the range of (X, Y) for which $X = x$ and the second integral is over all points in the range of (X, Y) for which $Y = y$.



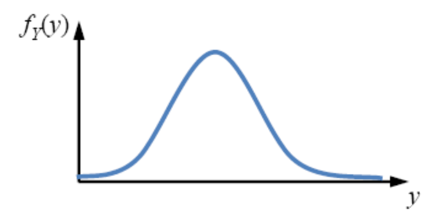
- Given a joint PDF $f_{XY}(x,y)$, marginal PDF of X or Y is obtained by **integrating $f_{XY}(x,y)$ over Y or X**

marginal PDF of X

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

marginal PDF of Y

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



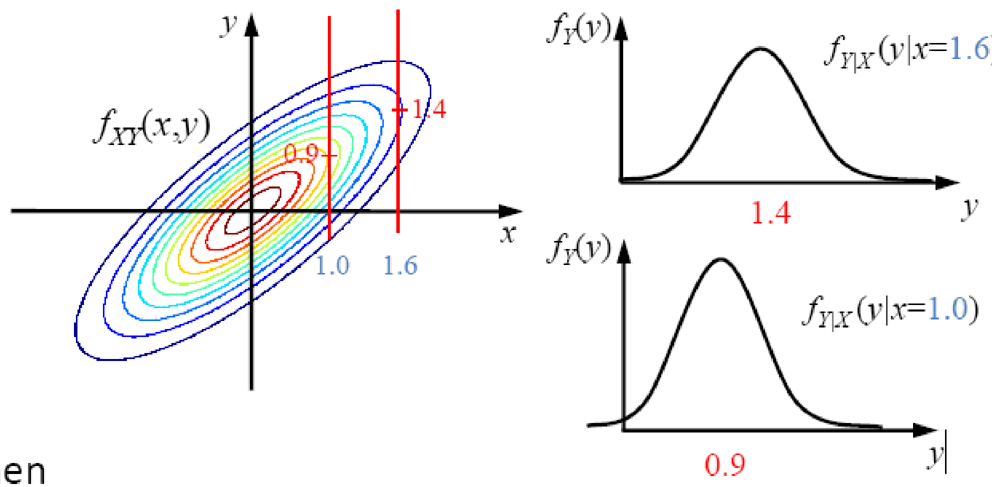
Given continuous random variables X and Y with joint probability density function $f_{XY}(x, y)$, the **conditional probability density function** of Y given $X = x$ is

$$f_{Y|x}(y) = \frac{f_{XY}(x, y)}{f_X(x)} \quad \text{for} \quad f_X(x) > 0 \qquad (5\text{-}4)$$

- Conditional PDF cont'd

- PDF of a variable when other variable(s) are known to have particular value(s)
- Conditional PDF of Y given $X = x$:

$$f_{Y|X}(y|x) = \frac{\overset{\textit{Joint PDF of X and Y}}{\downarrow \atop f_{XY}(x,y)}}{\underset{\textit{Marginal PDF of X}}{f_X(x) \atop \uparrow}}$$



- If X and Y are ***independent***, then

$$f_{X|Y}(x \mid y) = f_X(x) \qquad f_{XY}(x,y) = f_X(x)\, f_Y(y)$$

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

- For n mutually independent random variables, $X_1, X_2, \ldots, X_n$

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) = f_{X_1}(x_1)\, f_{X_2}(x_2) \cdots f_{X_n}(x_n)$$

- Measure of Correlation

Correlation

- tendency of variables to vary together
- If two or more random variables are ***correlated***, they do not satisfy a mathematical condition of probabilistic independence
    - *Covariance* is a measure to describe a ***linear relationship between random variables***

$$\sigma_{XY} = Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

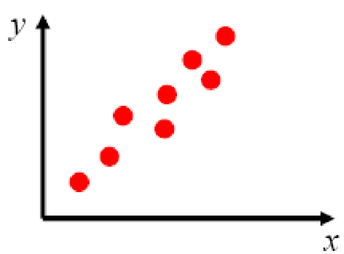
– *Correlation coefficient* is a non-dimensional measure of correlation

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

σ_X, σ_Y: standard deviation of X, Y

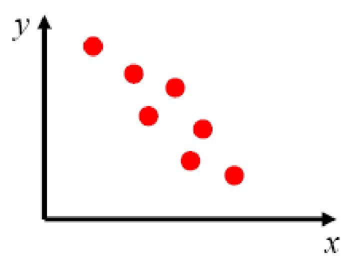
$$-1 \le \rho_{XY} \le 1$$

Positively correlated



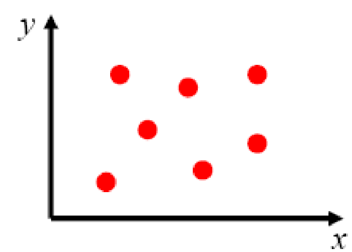
$\rho_{XY} \approx 1$

Negatively correlated



$\rho_{XY} \approx -1$

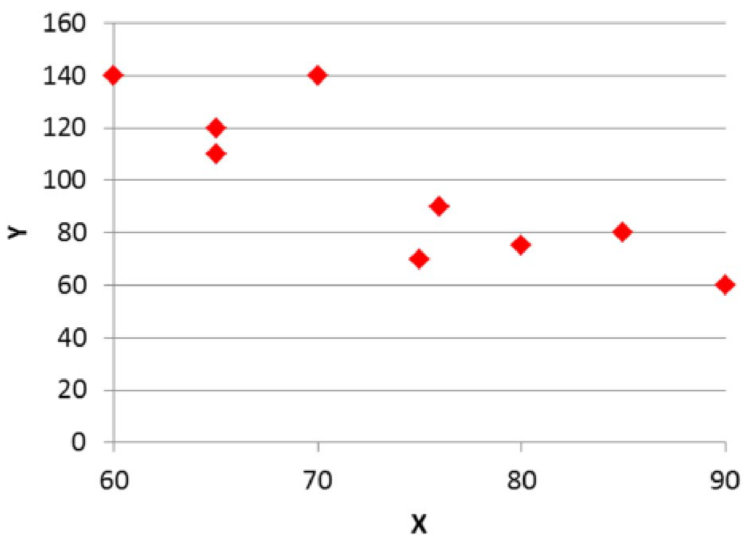
Uncorrelated



$\rho_{XY} = 0$

Correlation Example

X: World oil production (Million barrels/day)	60	65	65	70	75	76	80	85	90
Y: Gasoline price (Dollar/barrel)	140	110	120	140	70	90	75	80	60



Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{60+65+65+\cdots+90}{9} = \frac{666}{9} = 74 \text{ Mbbl/day}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{140+110+120+\cdots+60}{9} = \frac{885}{9} = 98.33 \text{ \$/bbl}$$

Covariance

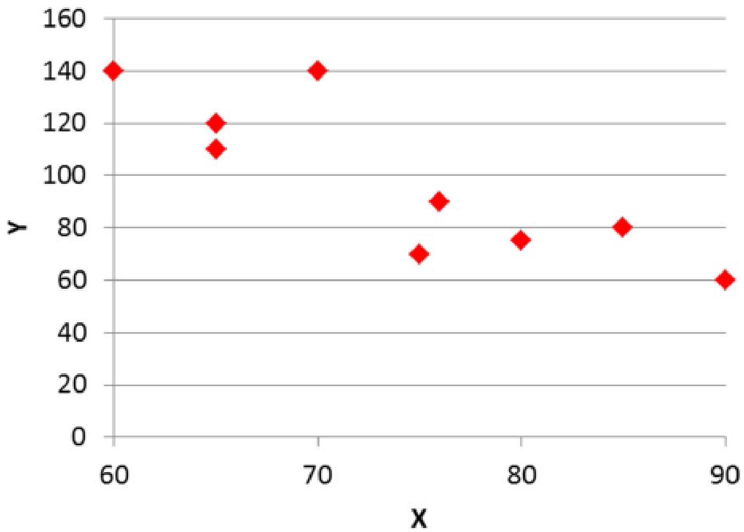
$$\sigma_{XY} = COV(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\sigma_{XY} = \frac{(60-74)(140-98.33)+\cdots+(90-74)(60-98.33)}{8}$$

$$= -256.25$$

Correlation Example

X: World oil production (Million barrels/day)	60	65	65	70	75	76	80	85	90
Y: Gasoline price (Dollar/barrel)	140	110	120	140	70	90	75	80	60



Correlation coefficient

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = -0.85$$

Highly correlated

Standard deviation

$$\sigma_X = \sqrt{V(X)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{792}{8}} = 9.95 \text{ Mbbl/day}$$

$$\sigma_Y = \sqrt{V(Y)} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{7300}{8}} = 30.21 \text{ \$/bbl}$$

- Joint PDF w/ more than two variables

A **joint probability density function** for the continuous random variables $X_1, X_2, X_3, \ldots, X_p$, denoted as $f_{X_1X_2\ldots X_p}(x_1, x_2, \ldots, x_p)$, satisfies the following properties:

(1) $f_{X_1X_2\ldots X_p}(x_1, x_2, \ldots, x_p) \geq 0$

(2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{X_1X_2\ldots X_p}(x_1, x_2, \ldots, x_p)\,dx_1\,dx_2 \ldots dx_p = 1$

(3) For any region B of p-dimensional space,

$$P\left[(X_1, X_2, \ldots, X_p) \in B\right] = \int\int_B f_{X_1X_2\cdots X_p}(x_1, x_2, \ldots, x_p)\,dx_1\,dx_2 \ldots dx_p \qquad (5\text{-}8)$$

If the joint probability density function of continuous random variables $X_1, X_2, \ldots, X_p$ is $f_{X_1X_2\ldots X_p}(x_1, x_2, \ldots, x_p)$, the **marginal probability density function** of X_i is

$$f_{X_i}(x_i) = \int\int\ldots\int f_{X_1X_2\ldots X_p}(x_1, x_2, \ldots, x_p)\,dx_1\,dx_2 \ldots dx_{i-1}\,dx_{i+1} \ldots dx_p \qquad (5\text{-}9)$$

where the integral is over all points in the range of $X_1, X_2, \ldots, X_p$ for which $X_i = x_i$.

$$E(X_i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} x_i f_{X_1X_2\ldots X_p}(x_1, x_2, \ldots, x_p)\,dx_1\,dx_2 \ldots dx_p = \int_{-\infty}^{\infty} x_i\, f_{X_i}(x_i)\,dx_i$$

and (5-10)

$$V(X_i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} (x_i - \mu_{X_i})^2 f_{X_1X_2\ldots X_p}(x_1, x_2, \ldots, x_p)\,dx_1\,dx_2 \ldots dx_p = \int_{-\infty}^{\infty} (x_i - \mu_{X_i})^2 f_{X_i}(x_i)\,dx_i$$

* For independent random variables: