COLUMBIA | ENGINEERING
EXECUTIVE EDUCATION

# Week 5
## Statistical Distributions — The Shape of Data

Applied Data Science

Columbia University - Columbia Engineering

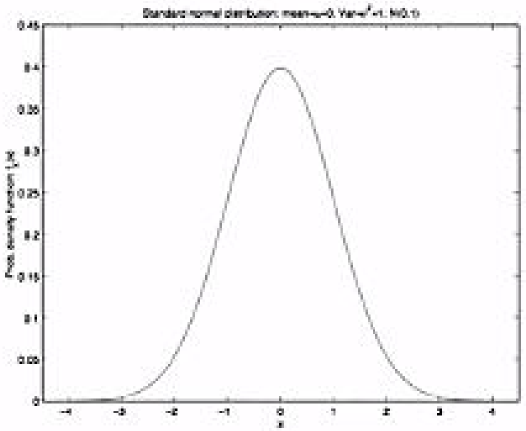
# Course Agenda

COLUMBIA | ENGINEERING
EXECUTIVE EDUCATION

- Week 1: Python Basics: How to Translate Procedures into Codes
- Week 2: Intermediate Python – Data Structures for Your Analysis
- Week 3: Relational Databases – Where Big Data is Typically Stored
- Week 4: SQL – Ubiquitous Database Format/Language
- **Week 5: Statistical Distributions – The Shape of Data**
- Week 6: Sampling – When You Can't or Won't Have ALL the Data
- Week 7:Hypothesis Testing – Answering Questions about Your Data
- Week 8: Data Analysis and Visualization – Using Python's NumPy for Analysis
- Week 9: Data Analysis and Visualization – Using Python's Pandas for Data Wrangling
- Week 10: Text Mining – Automatic Understanding of Text
- Week 11: Machine Learning – Basic Regression and Classification
- Week 12: Machine Learning – Decision Trees and Clustering

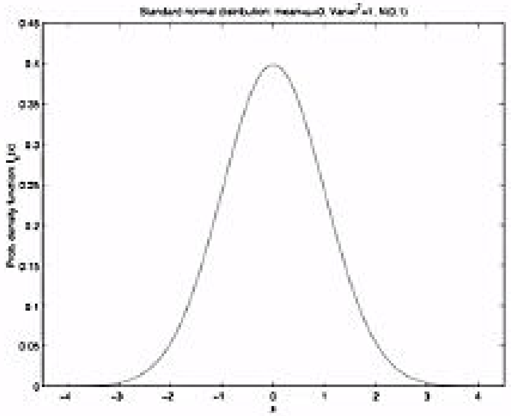
## The Normal distribution

- Most important & popular distribution in statistics.
- Many problems can be (very well) approximated & solved using the normal distribution.
- Very good approximation for sum of large number of uncertain quantities



**Notation:** $N(\mu, \sigma^2)$; in figure: $\mu = 0$, $\sigma^2 = 1$.

## Characteristics of normal distributions



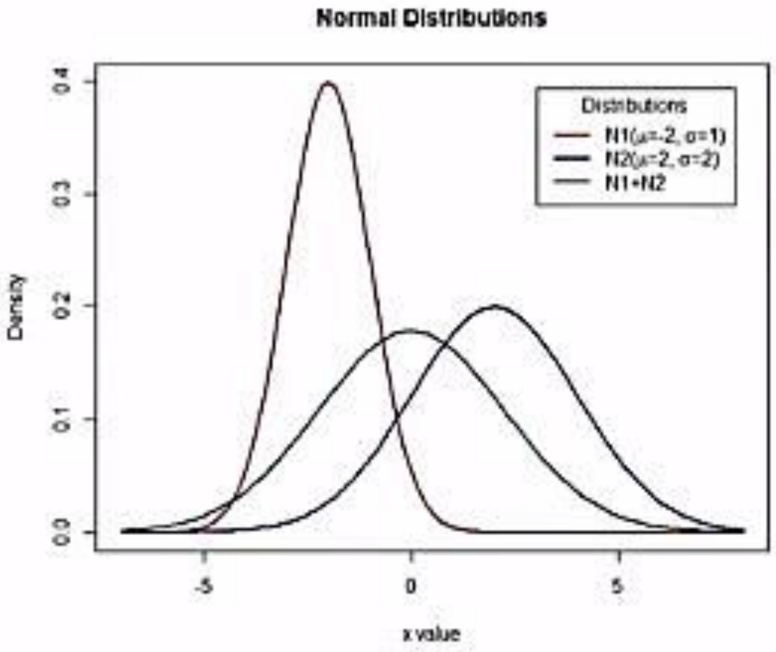
- Continuous data
- Interpretation:
  - $P(X \in [x, x + dx]) \simeq f_X(x)dx$
  - $f_X(\cdot)$ is the probability density function
  - $P(a \leq X \leq b) =$ area under the curve between $a, b$.

# Normally Distributed Random Variables- Distribution of Sum

## Distribution of sums of Normal random variables is Normal

**Fact:** If $X, Y$ are normally distributed and *independent* then

- $aX + b$ is normal; i.e., linear transformation of normal is normal
- $Z = aX + bY$ is normal; sum of independent normals is normal
  - $Z \sim N(a\mu_X + b\mu_Y,\ a^2\sigma_X^2 + b^2\sigma_y^2)$



# Types of Portfolios (1/2)

## Joint Distributions

- Joint density function: $f : \mathbb{R}^2 \to \mathbb{R}$
- Interpretation:

$$P(X \in [x, x+dx], Y \in [y, y+dy]) \simeq f(x,y)dx \cdot dy \qquad \text{for all } (x,y)$$

- Properties:

$$f_{X,Y}(x,y)) \geq 0 \text{ for all } (x,y),$$
$$\int_x \int_y f_{X,Y}(x,y)dydx = 1$$

- Probability of any event

$$P((X,Y) \in B) = \int\int_{(x,y)\in B} f_{X,Y}(x,y)dydx$$

- *Marginal* density function of X is defined as:

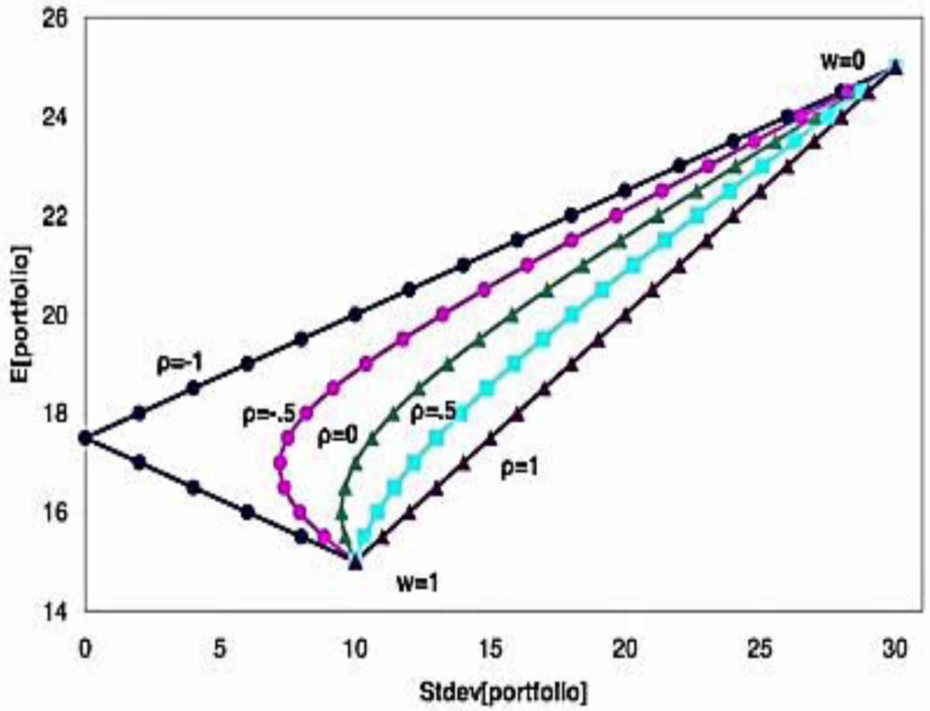
$$f_X(x) = \int_y f_{X,Y}(x,y)dy$$

- If $X$ and $Y$ are independent:

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) \qquad \text{(product of marginal densities)}$$

10

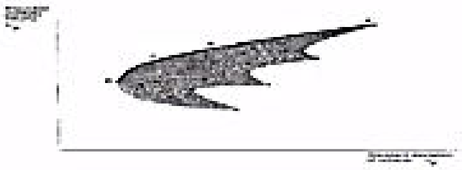
**Portfolio with Correlated Stocks**



# Types of Portfolios (2/2)

## Portfolio returns with multiple stocks

- With multiple stocks, the best portfolio is more difficult to compute
- Basically, any point in region represents a portfolio
- *Efficient frontier*: first defined by Markowitz in his influential '52 paper that launched portfolio theory
  (he got the Nobel prize for that paper!)



## Value-at-Risk (VaR)

The 99% Value-at-Risk of an investment is the amount $x$, such that the returns from that investment over a fixed time period will be $\leq x$ with probability 1%.

What is the 99% VaR over one year for the S&P 500?
(Annual rate of return of S&P 500 is normal with $\mu = 8.79\%$ and $\sigma = 15.75\%$.)

# Other Types of Distribution (1/3)

COLUMBIA | ENGINEERING
EXECUTIVE EDUCATION

## Bernoulli Distribution

- Discrete distribution with two possible outcomes
- $$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } (1-p) \end{cases}$$

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1-p & \text{if } 0 \leqslant x < 1 \\ 1 & \text{if } x \geqslant 1 \end{cases}$$

$$E(X) = p$$

$$Var(X) = p(1-p)$$


- **Examples**
  - probability of click in Display advertising
  - probability of stock price going up or down in a period

## Binomial Distribution

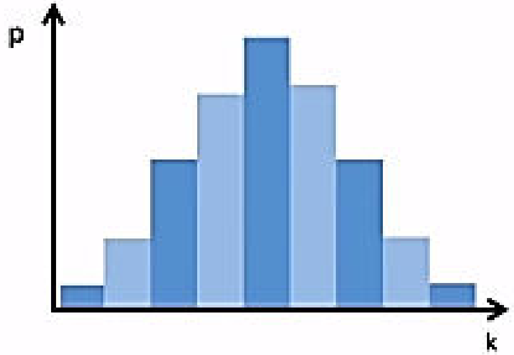
- $k$ success in $n$ independent trials

$$\text{Per trial} \begin{cases} \text{success (e.g. purchase) with probability } p \\ \text{failure (e.g. no purchase) with probability } 1-p \end{cases}$$

- $$\begin{aligned} p_X(k) &= Pr(\text{k success in n trials}) \\ &= \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

$$E(X) = np$$

$$Var(X) = np(1-p)$$



# Other Types of Distribution (2/3)

## Geometric Distribution

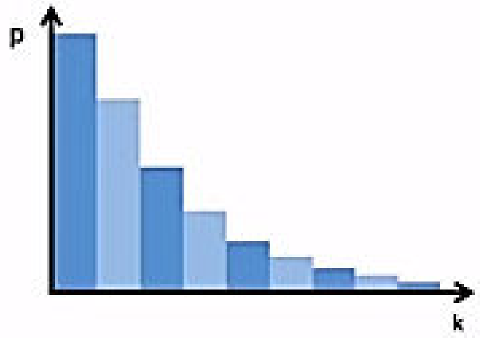
- Number of trials until first success
-

$$p_X(k) = p(1-p)^{k-1}$$

$$F_X(k) = 1-(1-p)^k$$

$$E(X) = \frac{1}{p}$$

$$Var(X) = \frac{1-p}{p^2}$$




- **Example**: A certain basketball player has a 60% chance of making a free throw. Assume all free throws are independent. What is the probability that he makes his first free throw on the $3^{rd}$ try?

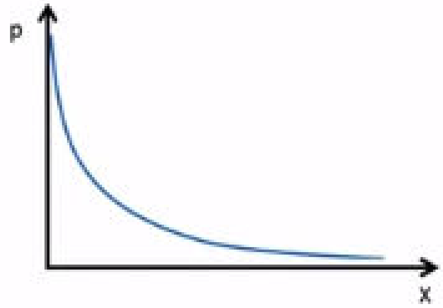
## Exponential Distribution

$$f_X(x) = \lambda e^{-\lambda x} \quad x \geqslant 0$$

$$F_X(x) = 1 - e^{-\lambda x} \quad x \geqslant 0$$

$$E(X) = \frac{1}{\lambda}$$

$$Var(X) = \frac{1}{\lambda^2}$$



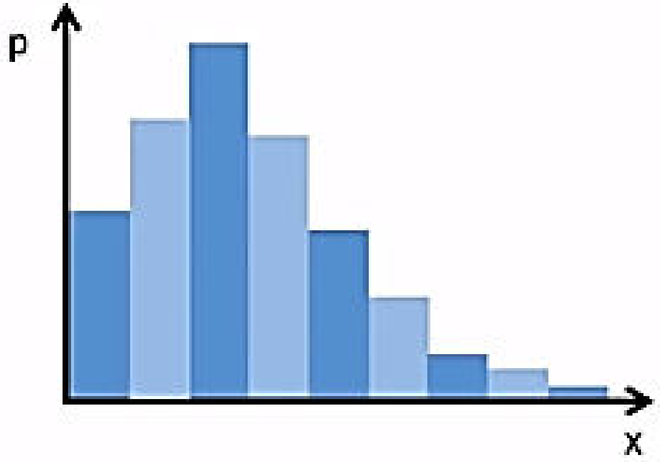
## Poisson Distribution

- Probability of a given number of events occurring in a fixed interval of time and/or space
- 

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

$$E(X) = \lambda$$

$$Var(X) = \lambda$$

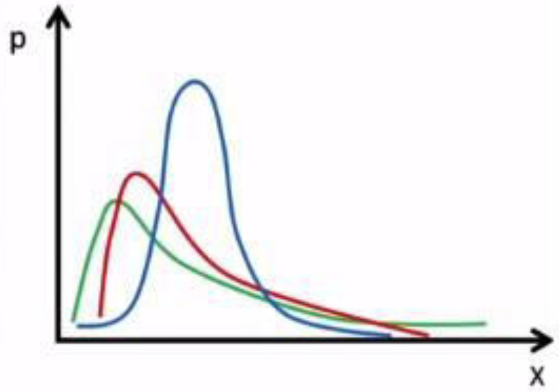



## Lognormal Distribution

- If $ln(x)$ is normally distributed, x is lognormally distributed.
- $ln(X) \sim N(\mu, \sigma^2)$

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

$$F(x) = \Phi\left(\frac{ln(x)-\mu}{\sigma}\right)$$




- Consequence of CLT on the logarithm of product of independent random variables
- Arises in many natural phenomenon. For instance:
  - Biological processes: size of a living tissue, blood pressure in adult human
  - Epidemic or rumor spreading: number of affected nodes

GLOBAL IVY EDUCATION

www.emeritus.org