Week 12

Statistical Distributions — The Shape of Data

Applied Data Science

Columbia University - Columbia Engineering

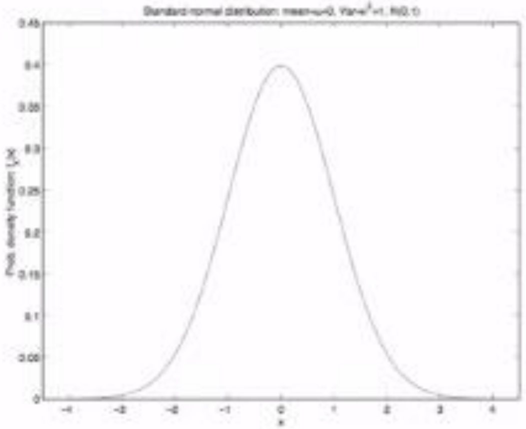
Course Agenda

COLUMBIA | ENGINEERING
EXECUTIVE EDUCATION

- Week 10: Organizing and Analyzing Data with NumPy and Pandas
- Week 11: Cleaning and Visualizing Data with Pandas and Matplotlib
- **Week 12: Statistical Distributions**
- Week 13: Statistical Sampling
- Week 14: Hypothesis Testing
- Week 15: Regression Models in Python
- Week 16: Evaluating Data Models
- Week 17: Classification with K-Nearest Neighbors
- Week 18: Decision Tree Models
- Week 19: Clustering Models
- Week 20: Text Mining in Python -- Analyzing Sentiment
- Week 21: Text Mining in Python -- Topic Modeling

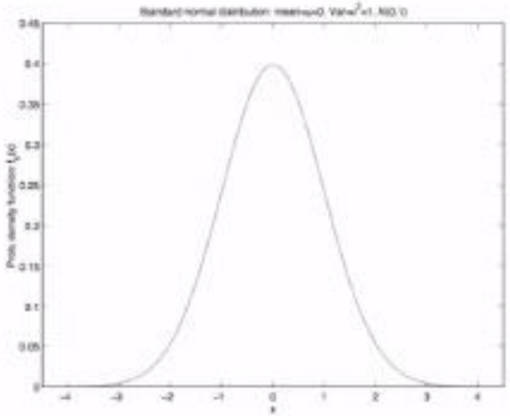
The Normal distribution

- 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



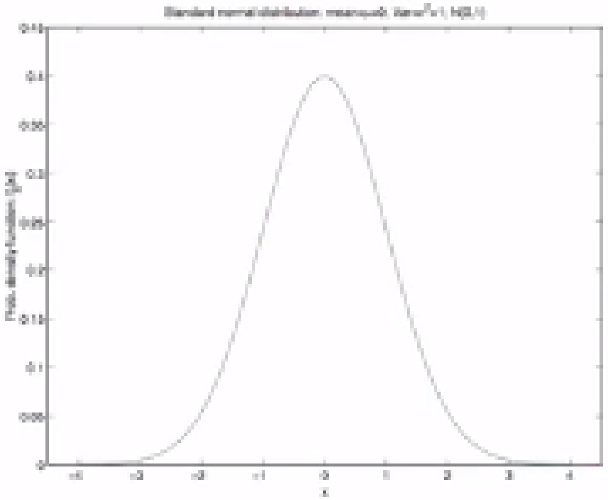
Notation: $N(\mu, \sigma^2)$; in figure: $\mu = 0$, $\sigma^2 = 1$.

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.

Standard normal: $Z \sim N(0,1)$



$P(Z \leq 1.30)$=?

Find z such that $P(Z \leq z) = .95$?

Fact: If $X \sim N(\mu, \sigma^2)$, then

$$\frac{X-\mu}{\sigma} = Z \sim N(0,1),$$

Example 1 – Motivating Example

Columbia Engineering
Executive Education

Consider the stocks:

Stock	Ann.return	Exp.ann.return	Stdev
A	X	$\mu_X = 15\%$	$\sigma_X = 10\%$
B	Y	$\mu_Y = 25\%$	$\sigma_Y = 30\%$

X, Y are Normally distributed. We want to compare two portfolios:

- Safe (S): 70% invested in A and 30% in B
- Risky (R): 30% invested in A and 70% in B

Expected return

Recap of the formula: portfolio standard deviation

$\mathrm{Var}[aX + bY] = ?$

Independent case ($\Rightarrow \rho_{XY} = 0$):

$$\mathrm{Var}[aX + bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y]$$

Correlated case ($\rho_{XY} \neq 0$):

$$\mathrm{Var}[aX + bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y] + 2ab \cdot \mathrm{Cov}[X, Y]$$

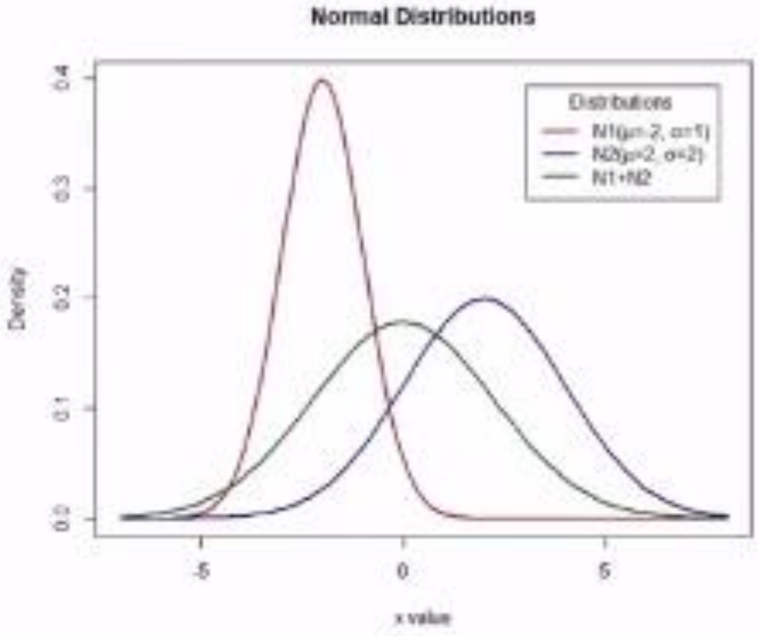
Portfolio standard deviation calculation...

- Recall: $\sigma_X = 10\%$, $\sigma_Y = 30\%$ and X, Y Normal, and $\rho_{XY} = 0$
- $S = 0.7X + 0.3Y$ and $R = 0.3X + 0.7Y$

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
  - $Z \sim N(a\mu_X + b\mu_Y,\ a^2\sigma_X^2 + b^2\sigma_y^2)$



Example 2

Two other portfolios

P_1 : 80% in A and 20% in B

P_2 : 90% in A and 10% in B

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:

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

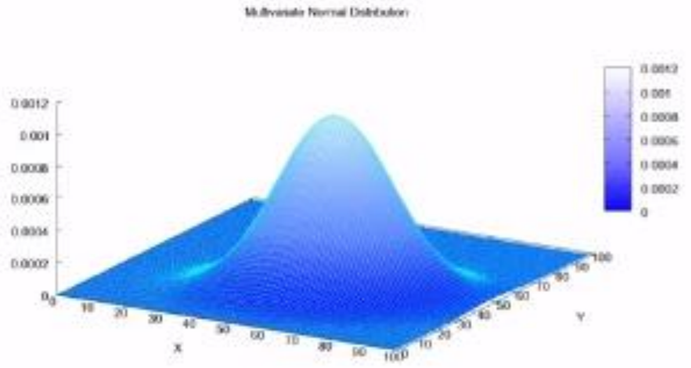
- If X and Y are **independent**:

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

Distribution of sums of Normal random variables is Normal

Fact: If X, Y are *jointly* normally distributed then

- Any linear combination of X, Y also has a normal distribution



Portfolios with Correlated Stocks

Positively correlated stocks: $\rho_{XY} = 0.1$

Q: can we construct better portfolios than S or R?

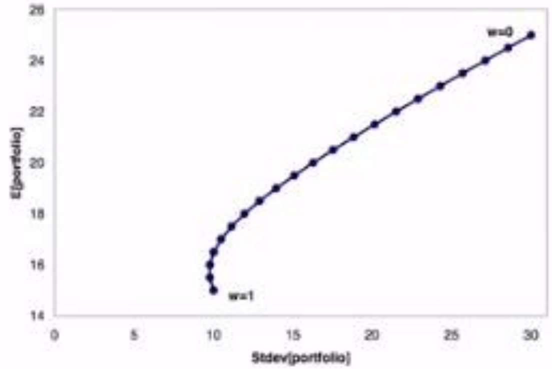
Proposed solution:

- invest fraction w of wealth in A and $(1-w)$ in B
- expected returns? standard deviations?

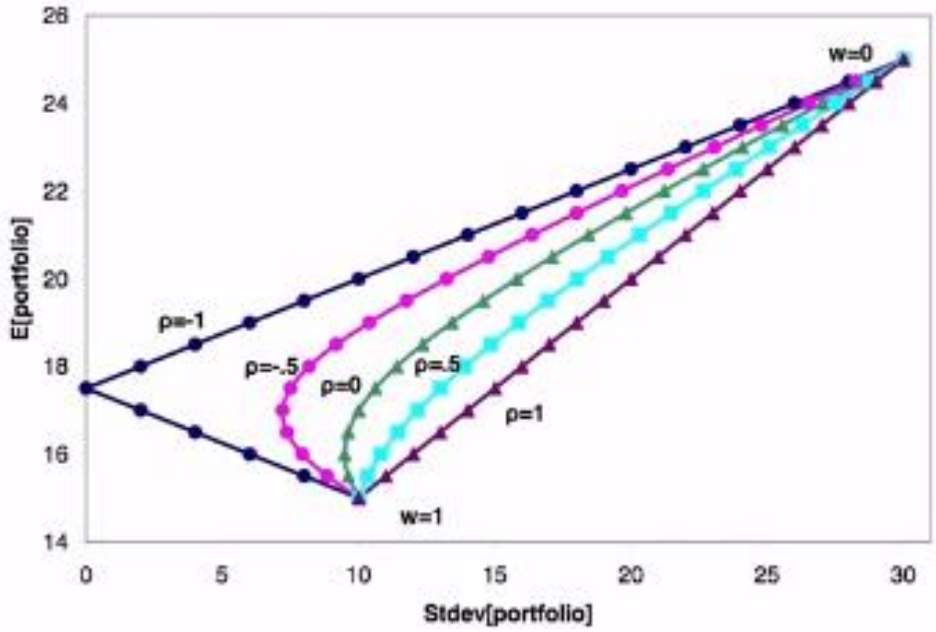
Portfolio diversification: $\rho_{XY} = 0.1$

Let's plot different portfolios:

Each point is a portfolio that invests a fraction w of wealth in A and $(1-w)$ in B

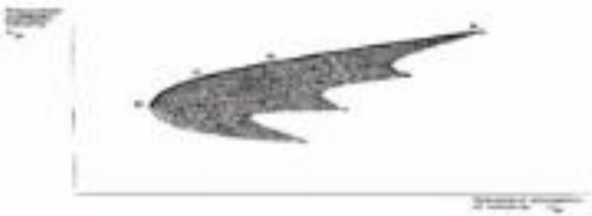


Portfolio returns for variable ρ



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 $\sigma = 15.75\%$.)

Value-at-Risk: a simple example

You are managing a portfolio, say worth $100M, with

average daily payoff $\bar{X} = \$0M$ and standard deviation of daily payoffs $\sigma = \$3M$

What is your 97.5% one-day Value-at-Risk? (Assume returns are Normally distributed.)

1. Plot a histogram of daily payoffs $\bar{X} = \$0M$ and $\sigma = \$3M$
2. From def'n of VaR: we want to find "x" such that 2.5% of days we lose x or more

Summary

1. Standardize:

$$X \to \frac{X-\mu}{\sigma} = Z \sim N(0,1)$$

2. Rephrase question of interest for $X \sim N(\mu, \sigma^2)$ in terms of $Z \sim N(0,1)$; i.e., in # of StdDev. Translate solution back for $X \sim N(\mu, \sigma^2)$
3. **Fact:** If X, Y are jointly normally distributed then
   - $aX + b$ is normal; i.e., linear transformation of normal is normal.
   - $X + Y$ is normal; i.e., sum of jointly normally distributed random variables is normal.
   - $aX + bY$ is normal; combination of the above.
4. Formulas you should know:

$$\boxed{\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]}$$

$$\text{Var}[aX + bY] = a^2\text{Var}[X] + b^2\text{Var}[Y] + 2ab \cdot \text{Cov}[X, Y]$$

or

$$\text{Var}[aX + bY] = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\rho_{XY}\,\sigma_X\sigma_Y$$

Bernoulli Distribution

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

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

$$E(X) = p$$

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

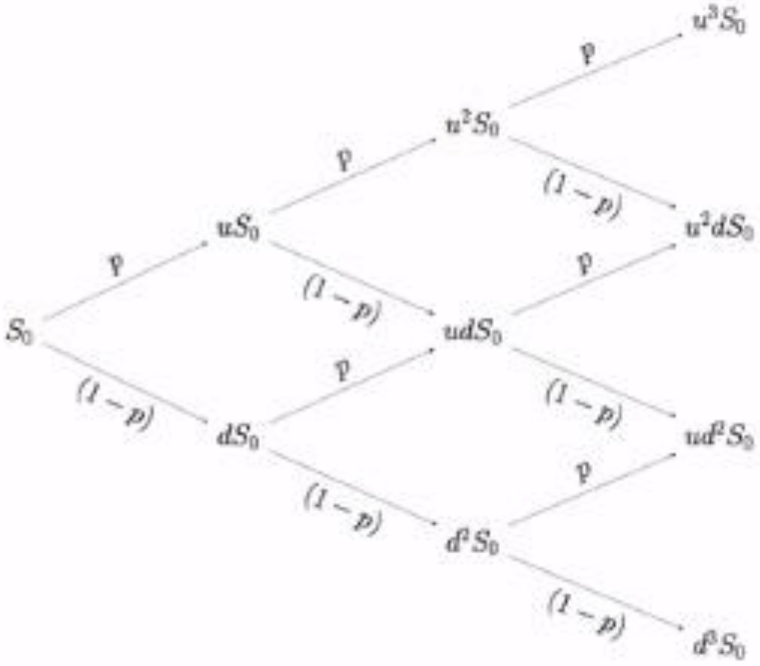
- **Examples**
  - probability of click in Display advertising
  - probability of stock price going up or down in a period

Bernoulli Distribution

- Building block for other richer discrete distributions
  - **Binomial Distribution** - number of successes in n trials (e.g. probability of k clicks out of n ads displayed)
  - **Geometric Distribution** - number of failures before the first success
  - **Negative Binomial Distribution** - number of failures before the x_{th} success

Binomial Distribution

- Example: Binomial Option Pricing Model



Binomial Distribution

- k success in n independent trials

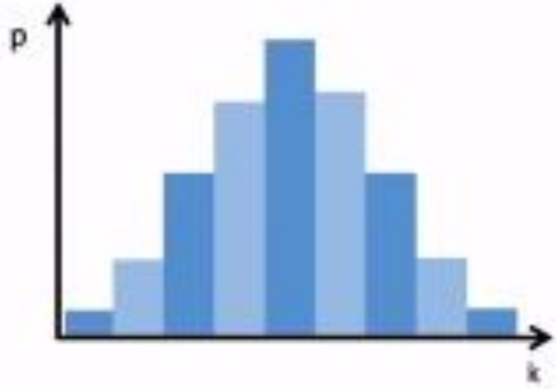
$$\text{Per trial} \begin{cases} \text{success (e.g. purchase) with probability } p \\ \text{failure (e.g. no purchase) with probability } 1-p \end{cases}$$

-

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

$$E(X) = np$$

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



Approximation: If n is large enough, $N(\mu, \sigma^2)$ is a good approximation for $B(n, p)$

- $\mu = np$
- $\sigma^2 = np(1-p)$

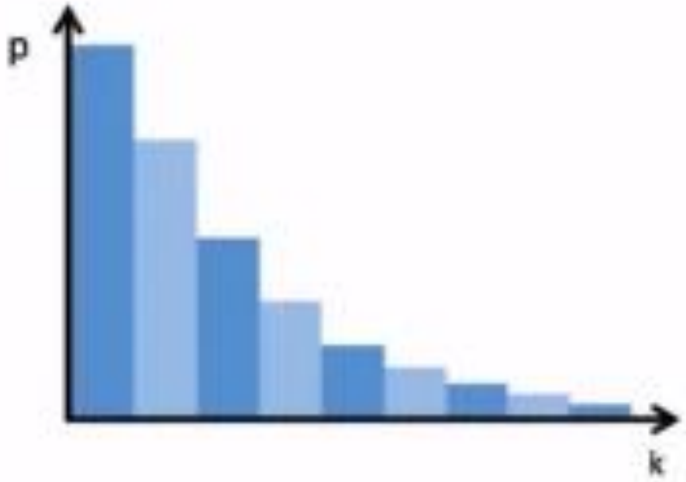
- Number of trials until first success
-

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

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

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

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



- **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 3rd try?

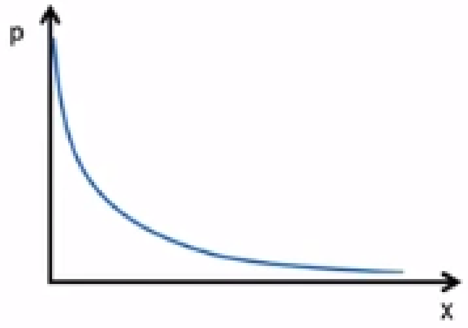
Exponential Distribution

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

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

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

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



Exponential distribution: Properties

- Exponential distribution is the continuous analogue of the geometric distribution

- **Memoryless** - $P(T > s + t | T > s = P(T > t))$

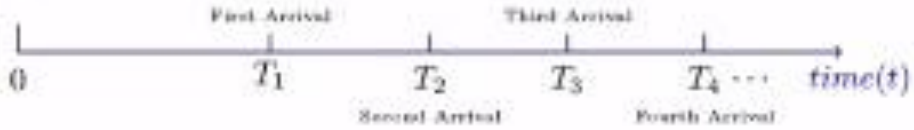
Example of Exponential Distribution

- **Example**: On average number of people arriving at the bus station in an hour is 3. Probability the time till the next person arrive is less than one hour is:

$$F_X(1) = P(X \leqslant 1)$$

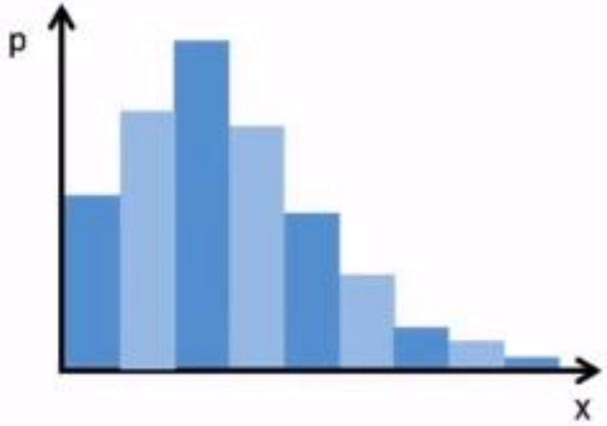
- **Call Center**
Calls arrive at call center an average rate λ per hour. Customers wait in the queue until one of two things happen: an agent is allocated to serve them (through supporting software), or they become impatient and abandon the tele-queue. Service time and customer patience (time to abandonment) are both exponentially distributed.

- **Poisson Process**



Poisson Distribution

- Probability of a given number of events occurring in a fixed interval of time and/or space
- $$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$
  $$E(X) = \lambda$$
  $$Var(X) = \lambda$$



Poisson Process

- The counting process $\{N(t), t > 0\}$ with rates λ, $\lambda > 0$,
  $$P\{N(t) = n\} = \frac{(\lambda t)^n}{n!}e^{-\lambda t}$$
- $$E[N(t)] = \lambda t$$
- The inter-arrival times X_1, X_2, ... are independent and $X_i \sim Exponential(\lambda)$

Example of Poisson Distribution

- **Example**: Which of the following is most likely to be well modeled by a Poisson distribution?
  1. Number of trains arriving at station every hour
  2. Number of lottery winners each year that live in Manhattan
  3. Number of days between solar eclipses
  4. Number of days until a component fails

- **Example**: the mean number of people arriving per hour at a shopping center is 18.What is the probability that the number of customers arriving in an hour is 20.

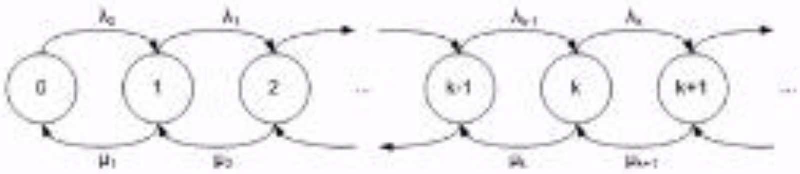
$$P(20) = e^{-18}\frac{18^{20}}{20!}$$

Example of Poisson Process

- **Traffic Model**
  Suppose the time between arrival of buses at the student center is exponentially distributed with a mean of 60 minutes. If we arrive at the student center at a randomly chosen instant, what is the average amount of time that we will have to wait for a bus?

  The Waiting Time Paradox: The memoryless property of the exponential distribution implies that whatever the time at which we arrive , the mean waiting time is the 60 min.

- **Birth and Death Process**



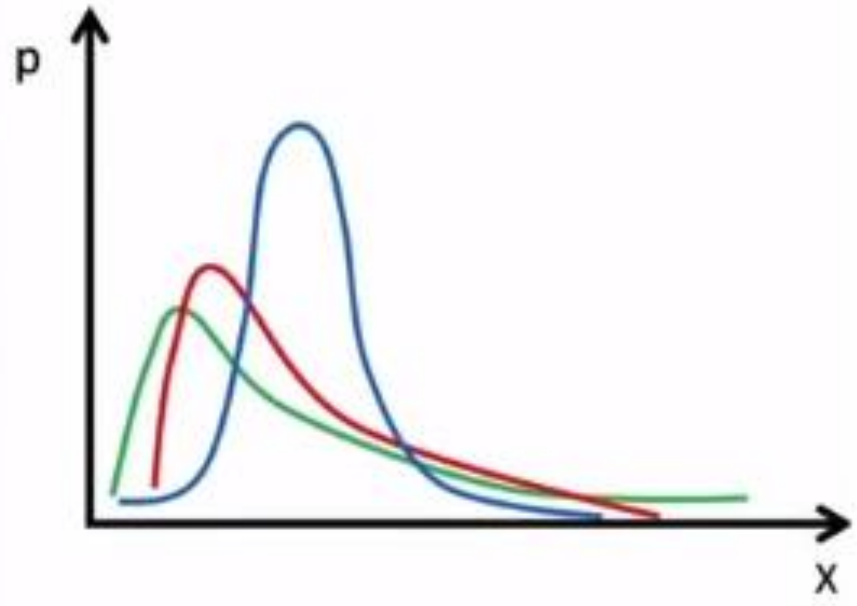
Lognormal Distribution

- 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

EMERITUS