Data Analytics (CS61061)

Lecture #4

Statistics for Data Analytics

Dr. Debasis Samanta

Professor

Department of Computer Science & Engineering

Quote of the day..

- "I avoid looking forward or backward, and try to keep looking upward."
  - CHARLOTTE BRONTE, an English novelist and poet

Today's discussion…

- Statistics versus Probability
- Concept of random variable
- Probability distribution concept
  - Discrete probability distribution
  - Continuous probability distribution
- Concept of sampling distribution
- Major sampling distributions
- Usage of sampling distributions

Probability and Statistics

Probability is the chance of an outcome in an experiment (also called event).

Event: Tossing a fair coin
Outcome: Head, Tail

Probability deals with **predicting** the likelihood of **future** events.

Statistics involves the **analysis** of the **frequency** of **past** events

Example: Consider there is a drawer containing 100 socks: 30 red, 20 blue and 50 black socks.

We can use probability to answer questions about the selection of a random sample of these socks.

- **PQ1.** What is the probability that we draw two blue socks or two red socks from the drawer?
- **PQ2.** What is the probability that we pull out three socks or have matching pair?
- **PQ3.** What is the probability that we draw five socks and they are all black?

Statistics

Instead, if we have no knowledge about the type of socks in the drawers, then we enter into the realm of statistics. Statistics helps us to infer properties about the population on the basis of the random sample.

Questions that would be statistical in nature are:

- **SQ1**: A random sample of 10 socks from the drawer produced one blue, four red, five black socks. What is the total population of black, blue or red socks in the drawer?
- **SQ2**: We randomly sample 10 socks, and write down the number of black socks and then return the socks to the drawer. The process is done for five times. The mean number of socks for each of these trial is 7. What is the true number of black socks in the drawer?
- etc.

Probability versus Statistics

In other words:

- In probability, we are given a model and asked what kind of data we are likely to see.
- In statistics, we are given data and asked what kind of model is likely to have generated it.

Example 4.1: Measles Study

- A study on health is concerned with the incidence of childhood measles in parents in a city. For each couple, we would like to know how likely, it is that either the mother or father or both have had childhood measles.
- The current census data indicates that 20% adults between the ages 20 and 40 (childbearing ages of parents regardless of sex) have had childhood measles.
  - This give us the probability that an individual in the city has had childhood measles.

Defining Random Variable

Definition 4.1: **Random Variable**

A random variable is a rule that assigns a numerical value to an outcome of interest.

Example 4.2: In "measles Study", we define a random variable X as the number of parents in a married couple who have had childhood measles.

This random variable can take values of $0, 1$ *and* 2.

Note:

- Random variable is not exactly the same as the variable defining a data.
- The probability that the random variable takes a given value can be computed using the rules governing probability.
  - For example, the probability that $X = 1$ means either mother or father but not both has had measles is 0.32. Symbolically, it is denoted as **P(X=1) = 0.32**

Probability Distribution

Definition 4.2: **Probability distribution**

A probability distribution is a definition of probabilities of the values of random variable.

Example 4.3: Given that 0.2 is the probability that a person (in the ages between 20 and 40) has had childhood measles. Then the probability distribution is given by

X	Probability
0	0.64
1	0.32
2	0.04

?

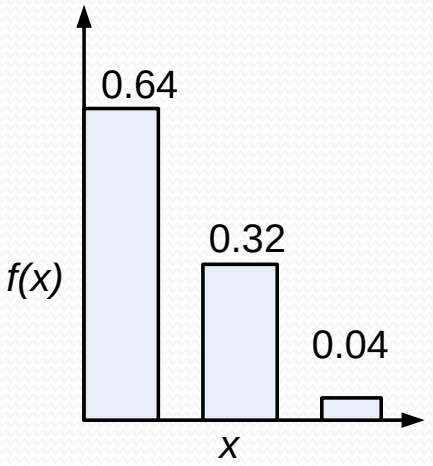
Probability Distribution

- In data analytics, the probability distribution is important with which many statistics making inferences about population can be derived .
  - In general, a probability distribution function takes the following form

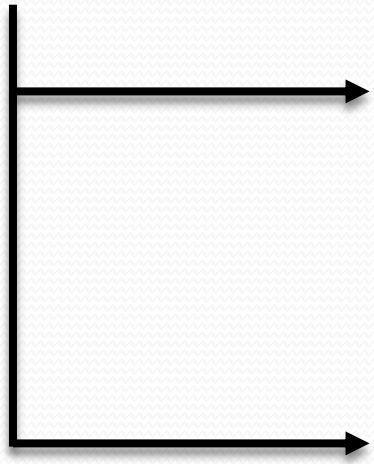
x	x_1	$x_2 \dots \dots \dots \dots \dots x_n$
$f(x) = P(X = x)$	$f(x_1)$	$f(x_2) \dots \dots \dots f(x_n)$

Example: Measles Study

x	0	1	2
$f(x)$	0.64	0.32	0.04



Taxonomy of Probability Distributions



Discrete probability distributions

- Binomial distribution
- Multinomial distribution
- Poisson distribution
- Hypergeometric distribution

Continuous probability distributions

- Normal distribution
- Standard normal distribution
- Gamma distribution
- Exponential distribution
- Chi square distribution
- Lognormal distribution
- Weibull distribution

Usage of Probability Distribution

- Distribution (discrete/continuous) function is widely used in simulation studies.
  - A simulation study uses a computer to simulate a real phenomenon or process as closely as possible.
  - The use of simulation studies can often eliminate the need of costly experiments and is also often used to study problems where actual experimentation is impossible.

Examples 4.4:

1) A study involving testing the effectiveness of a new drug, the number of cured patients among all the patients who use such a drug approximately follows a binomial distribution.

2) Operation of ticketing system in a busy public establishment (e.g., airport), the arrival of passengers can be simulated using Poisson distribution.

Discrete Probability Distributions

Binomial Distribution

- In many situations, an outcome has only two outcomes: success and failure.
  - Such outcome is called dichotomous outcome.
- An experiment when consists of repeated trials, each with dichotomous outcome is called Bernoulli process. Each trial in it is called a Bernoulli trial.

Example 4.5: Firing bullets to hit a target.

- Suppose, in a Bernoulli process, we define a random variable $P(X \equiv$ the number of successes) in trials.
- Such a random variable obeys the binomial probability distribution, if the experiment satisfies the following conditions:

1) The experiment consists of n trials.
2) Each trial results in one of two mutually exclusive outcomes, one labelled a *"success"* and the other a *"failure"*.
3) The probability of a success on a single trial is equal to $\boldsymbol{p}$. The value of p remains constant throughout the experiment.
4) The trials are independent.

Defining Binomial Distribution

Definition 4.3: **Binomial distribution**

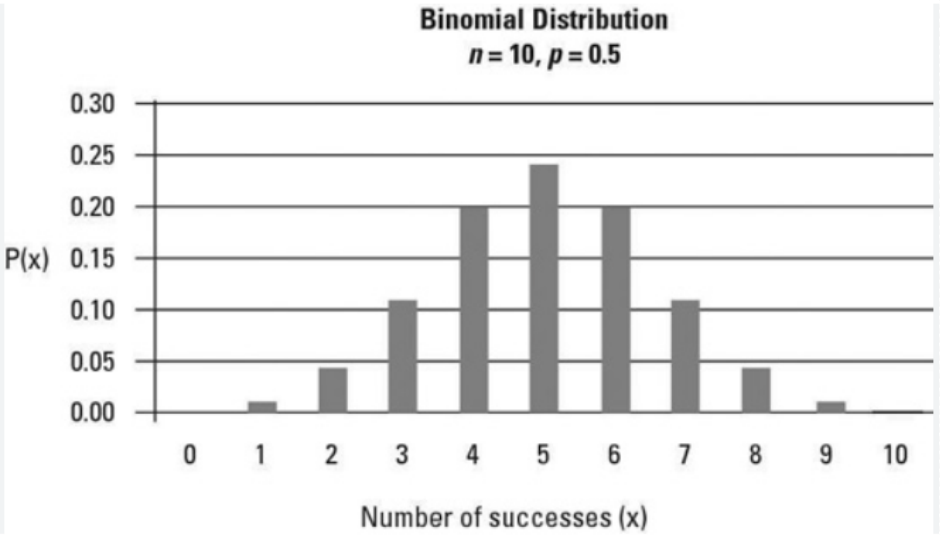
The function for computing the probability for the binomial probability distribution is given by

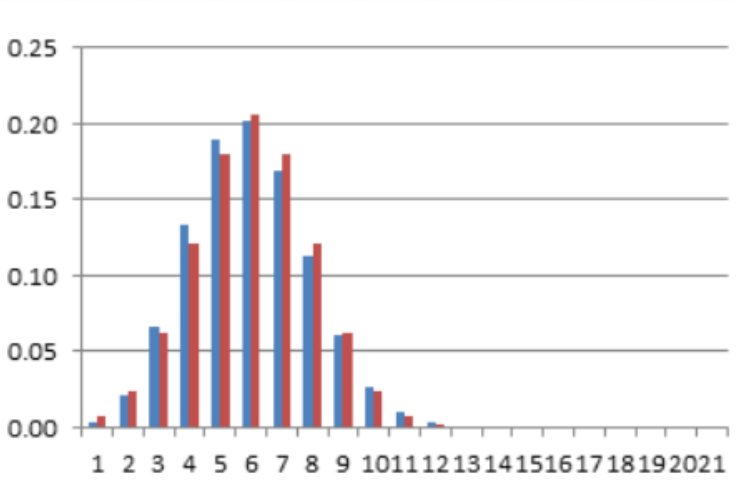
$$f(x) = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

$$for\ x = 0, 1, 2, \ldots, n$$

Here, $f(x) = P(X = x)$, where X denotes "the number of success" and $X = x$ denotes the number of successes is x.

Binomial Distribution Curves





Binomial Distribution

Example 4.6: Measles study

X = having had childhood measles a success

p = 0.2, the probability that a parent had childhood measles

n = 2, here a couple is an experiment and an individual a trial, and the number of trials is two.

Thus,

$$P(x=0) = \frac{2!}{0!(2-0)!}(0.2)^0(0.8)^{2-0} = \mathbf{0.64}$$

$$P(x=1) = \frac{2!}{1!\,(2-1)!}(0.2)^1(0.8)^{2-1} = \mathbf{0.32}$$

$$P(x=2) = \frac{2!}{2!\,(2-2)!}(0.2)^2(0.8)^{2-2} = \mathbf{0.04}$$

X	Probability
0	0.64
1	0.32
2	0.04

Binomial Distribution

Example 4.7: Verify with real-life experiment

Suppose, 10 random numbers each of two digits are generated by a computer (Monte-Carlo simulation method)

15 38 68 39 49 54 19 79 38 14

If the value of the digit is 0 or 1, the outcome is "had childhood measles", otherwise, (digits 2 to 9), the outcome is "did not".

For example, in the first pair (i.e., 15), representing a couple and for this couple, $x = 1$. The frequency distribution, for this sample is

x	0	1	2
$f(x)=P(X=x)$	0.7	0.3	0.0

Note: This has close similarity with binomial probability distribution!

The Multinomial Distribution

The binomial experiment becomes a multinomial experiment, if we let each trial has more than two possible outcome.

Definition 4.4: **Multinomial distribution**

If a given trial can result in the ***k*** outcomes $E_1, E_2, \dots\dots, E_k$ with probabilities $p_1, p_2, \dots\dots, p_k$, then the probability distribution of the random variables $X_1, X_2, \dots\dots, X_k$ representing the number of occurrences for $E_1, E_2, \dots\dots, E_k$ in n independent trials is

$$f(x_1, x_2, \dots\dots, x_k) = \binom{n}{x_1, x_2, \dots\dots, x_k} {p_1}^{x_1} {p_2}^{x_2} \dots\dots {p_k}^{x_k}$$

$$\text{where } \binom{n}{x_1, x_2, \dots\dots, x_k} = \frac{n!}{x_1! x_2! \dots\dots x_k!}$$

$$\sum_{i=1}^{k} x_i = n \text{ and } \sum_{i=1}^{k} p_i = 1$$

The Hypergeometric Distribution

- Collection of samples with two strategies
  - With replacement
  - Without replacement
- A necessary condition of the binomial distribution is that all trials are **independent** to each other.
  - When sample is collected "with replacement", then each trial in sample collection is independent.

Example 4.8:
Probability of observing three red cards in 5 draws from an ordinary deck of 52 playing cards.
  - You draw one card, note the result and then returned to the deck of cards
  - Reshuffled the deck well before the next drawing is made

- The hypergeometric distribution *does not require* independence and is based on the sampling done **without** replacement.

The Hypergeometric Distribution

- In general, the hypergeometric probability distribution enables us to find the probability of selecting x successes in n trials from N items.

Properties of Hypergeometric Distribution

- A random sample of size n is selected without replacement from N items.
- k of the N items may be classified as success and $N-k$ items are classified as failure.

Let X denotes a hypergeometric random variable defining the number of successes.

Definition 4.5: Hypergeometric Probability Distribution

The probability distribution of the hypergeometric random variable X, the number of successes in a random sample of size n selected from N items of which k are labelled success and $N-k$ labelled as failure is given by

$$f(x) = P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

$$\max(0, n-(N-k)) \leq x \leq \min(n, k)$$

Multivariate Hypergeometric Distribution

The hypergeometric distribution can be extended to treat the case where the N items can be divided into k classes $A_1, A_2, \ldots\ldots, A_k$ with a_1 elements in the first class A_1, … and a_k elements in the k^{th} class. We are now interested in the probability that a random sample of size n yields x_1 elements from A_1, x_2 elements from $A_2, \ldots\ldots, x_k$ elements from A_k.

Definition 4.6: Multivariate Hypergeometric Distribution

If N items are partitioned into k classes $a_1, a_2, \ldots\ldots, a_k$ respectively, then the probability distribution of the random variables $X_1, X_2, \ldots\ldots, X_k$, representing the number of elements selected from $A_1, A_2, \ldots\ldots, A_k$ in a random sample of size n, is

$$f(x_1, x_2, \ldots\ldots, x_k) = P(X_1 = x_1, X_2 = x_2, \ldots\ldots, X_k = x_k) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2}\ldots\ldots\binom{a_k}{x_k}}{\binom{N}{n}}$$

$$\text{with } \sum_{i=1}^{k} x_i = n \text{ and } \sum_{i=1}^{k} a_i = N$$

The Poisson Process

There are some experiments, which involve the occurring of the number of outcomes during a given time interval (or in a region of space). Such a process is called the Poisson process.

The Poisson process is one of the most widely used counting processes. It is usually used in scenarios where we are counting the occurrences of certain events that appear to happen at a certain rate, but completely at random.

Example 4.9:

Number of customers visiting a ticket selling counter in a railway station.

The Poisson Process

Properties of Poisson process

- There is a discrete value, say x is the number of times an event occurs in an interval and x can take values 0, 1, 2, ...
- The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
- The average rate at which events occur assumed to be constant.
- Two events cannot occur at exactly the same instant; instead, at each very small sub-interval exactly one event either occurs or does not occur.

If these conditions are true, then x is a Poisson random variable, and the distribution of x is a Poisson distribution.

The Poisson Distribution

Definition 4.7: **Poisson distribution**

The probability distribution of the Poisson random variable X, representing the number of outcomes occurring in a given time interval t, is

$$f(x, \lambda t) = P(X = x) = \frac{e^{-\lambda t}.(\lambda t)^x}{x!}, x = 0, 1, \ldots \ldots$$

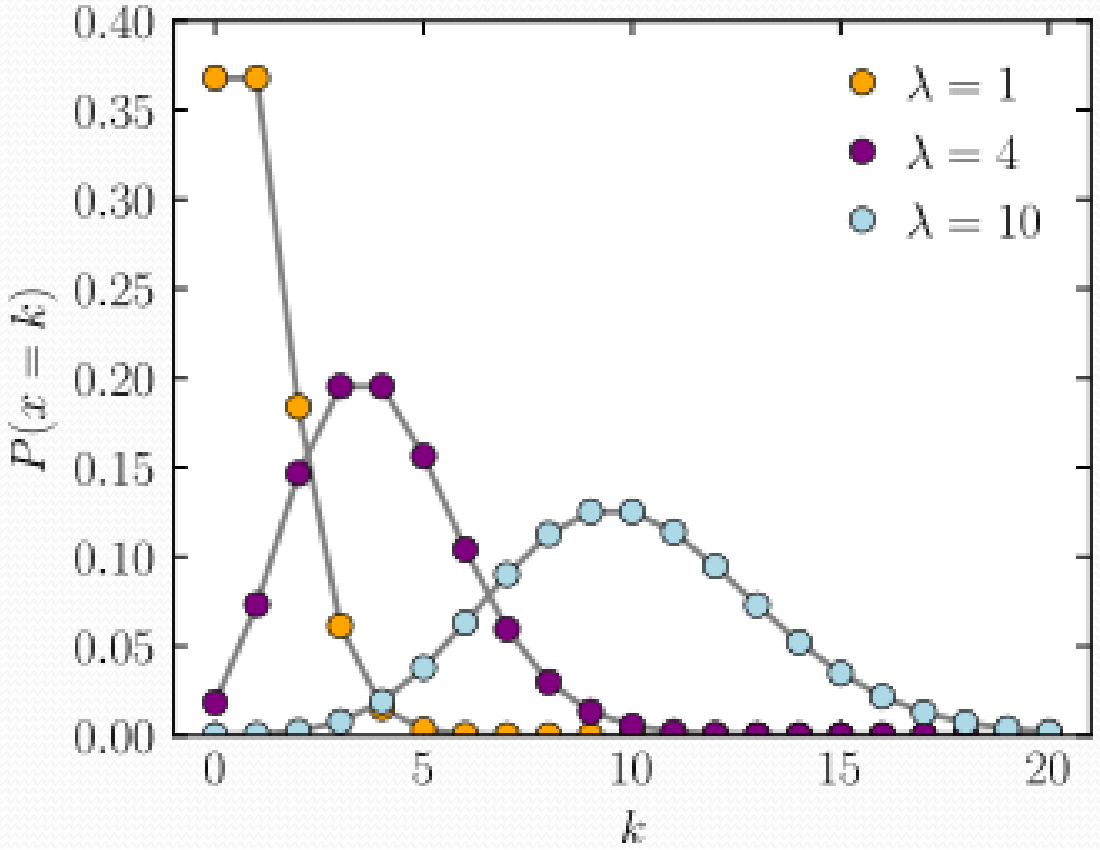
where λ is the average number of outcomes per unit time and $e = 2.71828 \ldots$

What is $P(X = x)$ if $t = 0$?

Example:

- The number of customers arriving at a grocery store can be modelled by a Poisson process with intensity λ=10 customers per hour.
  1. Find the probability that there are 2 customers between 10:00 and 10:20.
  2. Find the probability that there are 3 customers between 10:00 and 10:20 and 7 customers between 10:20 and 11:00.

The Poisson Distribution Curves



Descriptive measures

Given a random variable X in an experiment, we have denoted $f(x) = P(X = x)$, the probability that $X = x$. For discrete events $f(x) = 0$ for all values of x except $x = 0, 1, 2, \ldots\ldots$

Properties of discrete probability distribution

1. $0 \leq f(x) \leq 1$
2. $\sum f(x) = 1$
3. $\mu = \sum x.f(x)$ [is the mean]
4. $\sigma^2 = \sum(x - \mu)^2 . f(x)$ [is the variance]

In $2, 3$ *and* 4, summation is extended for all possible discrete values of x.

Note: For discrete uniform distribution, $f(x) = \frac{1}{n}$ with $x = 1, 2, \ldots\ldots, n$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

Descriptive measures

1. **Binomial distribution**

   The binomial probability distribution is characterized with p (the probability of success) and n (is the number of trials). Then

$$\mu = n.p$$

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

2. **Hypergeometric distribution**

   The hypergeometric distribution function is characterized with the size of a sample (n), the number of items (N) and k labelled success. Then

$$\mu = \frac{nk}{N}$$

$$\sigma^2 = \frac{N-n}{N-1}.\mathrm{n}.\frac{k}{N}\left(1-\frac{k}{N}\right)$$

Descriptive measures

3. **Poisson Distribution**

The Poisson distribution is characterized with λt where $\lambda = the\ mean\ of\ outcomes$ and $t = time\ interval$.

$$\mu = \lambda t$$

$$\sigma^2 = \lambda t$$

Alternative definition of Poisson distribution:

$$P(X = x) = \frac{e^{-\mu} \cdot \mu^x}{x!}$$

Special Case: Discrete Uniform Distribution

- **Discrete uniform distribution**

  A random variable *X* has a discrete uniform distribution if each of the *n* values in the range, say $x_1, x_2, x_3, \ldots, x_n$ has equal probability. That is

$$f(x) = \frac{1}{n}$$

  Where *f(x)* represents the probability mass function.

- **Mean and variance for discrete uniform distribution**

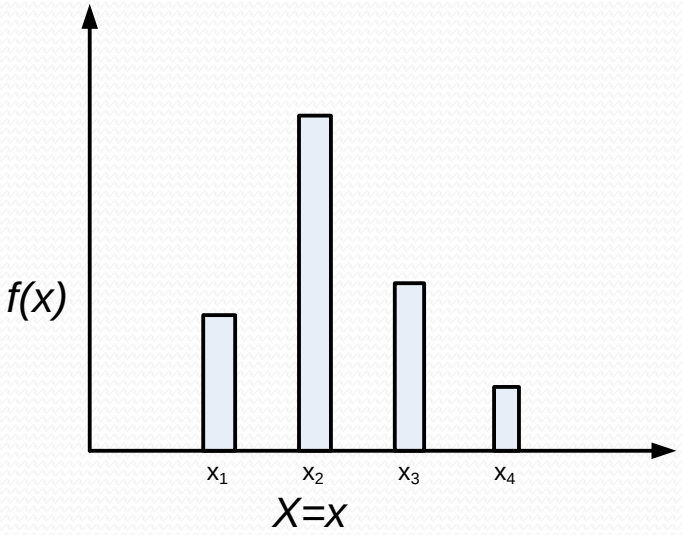
Suppose, *X* is a discrete uniform random variable in the range [*a*,*b*], *such that* $a \leq b$, then

$$\mu = \frac{b+a}{2}$$

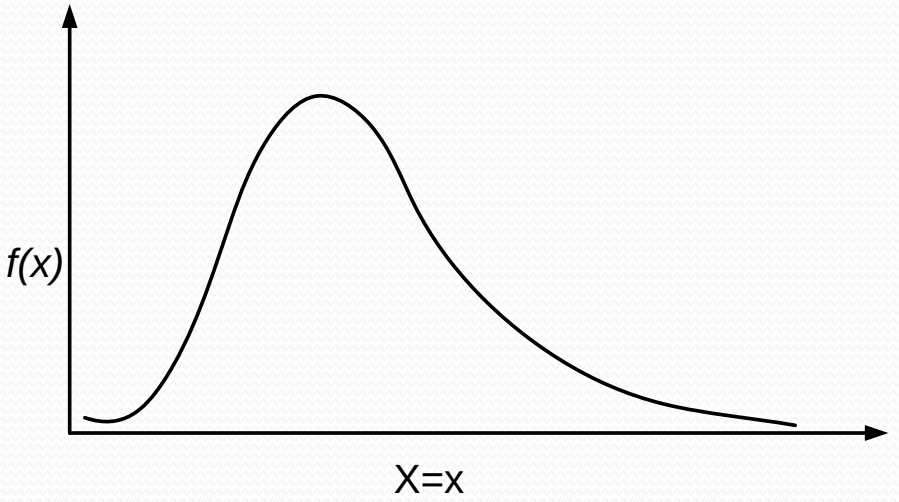
$$\sigma^2 = \frac{(b-a+1)^2-1}{12}$$

Continuous Probability Distributions

Continuous Probability Distributions



Discrete Probability distribution



Continuous Probability Distribution

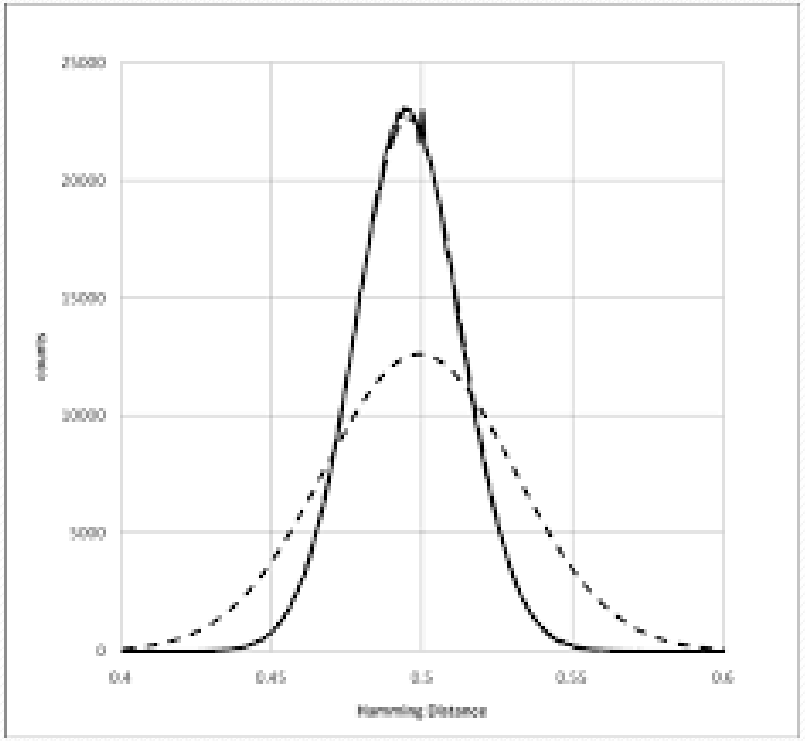
Continuous Probability Distributions

- When the random variable of interest can take any value in an interval, it is called continuous random variable.
  - Every continuous random variable has an infinite, uncountable number of possible values (i.e., any value in an interval)
- Consequently, continuous random variable differs from discrete random variable.

Examples:

1. Tax to be paid for a purchase in a shopping mall. Here, the random variable varies from 0 to $+\infty$
2. Amount of rainfall in *mm* in a region.
3. Earthquake intensity in Richter scale.
4. Height of an earth surface. Here, the random variable varies from $-a$ to $+b, [a, b] \in R, R$ is a set of real numbers.

Continuous Probability Distributions

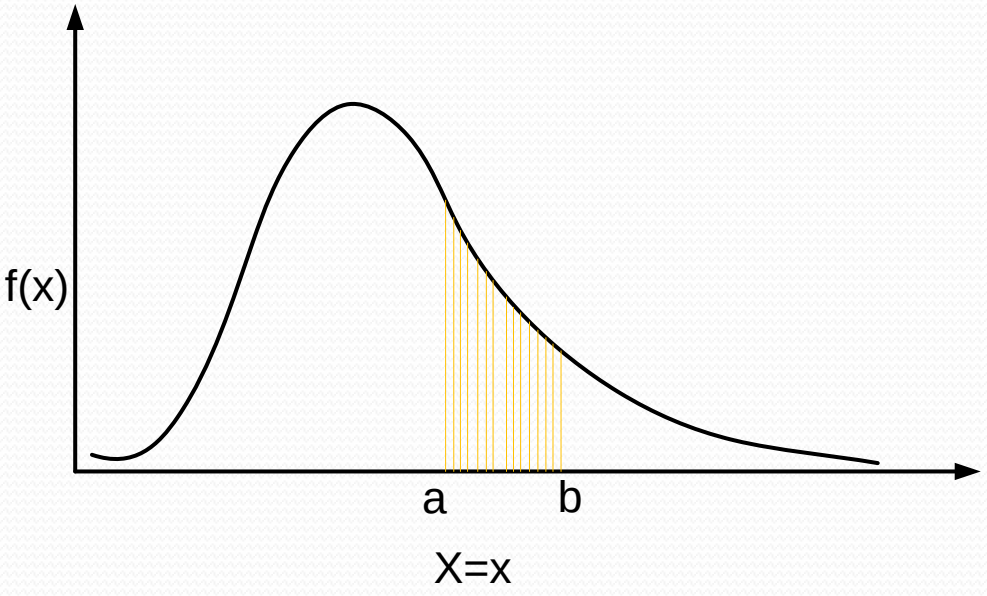


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

Properties of Probability Density Function

The function $f(x)$ is a probability density function for the continuous random variable X, defined over the set of real numbers R, if

1. $f(x) \geq 0$, for all $x \in R$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$
4. $\mu = \int_{-\infty}^{\infty} xf(x)\,dx$
5. $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$



Note: Probability is represented by area under the curve. The probability of a specific value of a continuous random variable will be zero because the area under a point is zero.

Continuous Probability Distributions

- **Example:**
  - Suppose bacteria of a certain species typically live 4 to 6 hours. The probability that a bacterium lives *exactly* 5 hours is equal to zero. A lot of bacteria live for approximately 5 hours, but there is no chance that any given bacterium dies at exactly 5.000000000… hours.
  - However, the probability that the bacterium dies between 5 hours and 5.01 hours is quantifiable.
  - Suppose, the answer is 0.02 (i.e., 2%). Then, the probability that the bacterium dies between 5 hours and 5.001 hours should be about 0.002, since this time interval is one-tenth as long as the previous. The probability that the bacterium dies between 5 hours and 5.0001 hours should be about 0.0002, and so on.

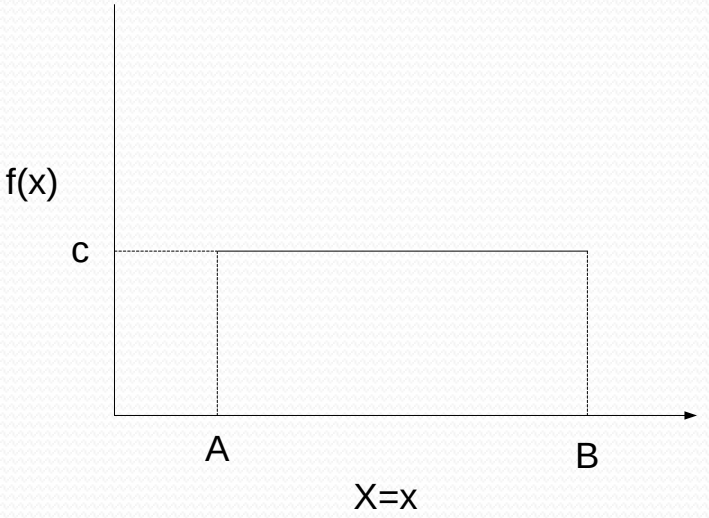
Continuous Probability Distributions

- **Note:**
  - In these three examples, the ratio (probability of dying during an interval) / (duration of the interval) is approximately constant, and equal to 2 per hour (or 2 $hour^{-1}$). For example, there is 0.02 probability of dying in the 0.01-hour interval between 5 and 5.01 hours, and (0.02 probability / 0.01 hours) = 2 $hour^{-1}$. This quantity 2 $hour^{-1}$ is called the probability density for dying at around 5 hours.

  - Therefore, the probability that the bacterium dies at 5 hours can be written as (2 $hour^{-1}$) dt. This is the probability that the bacterium dies within an infinitesimal window of time around 5 hours, where dt is the duration of this window.

Continuous Uniform Distribution

- One of the simplest continuous distribution in all of statistics is the continuous uniform distribution.

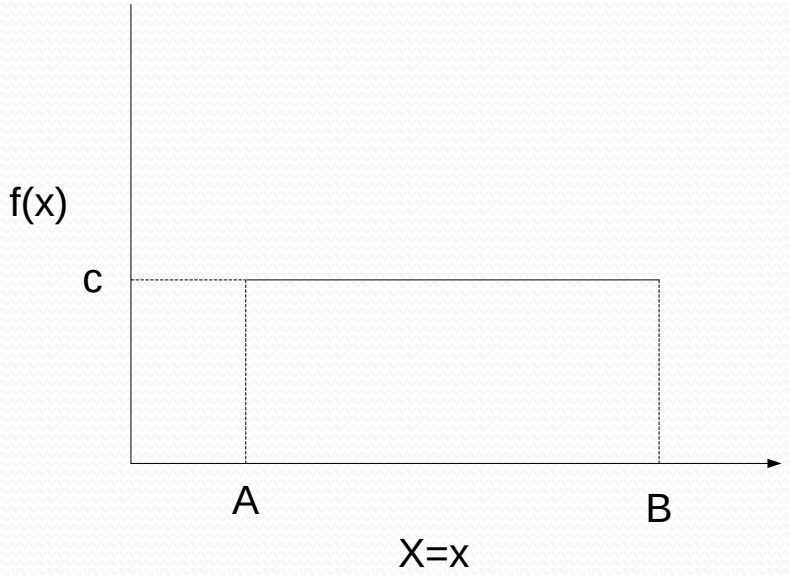


Definition 4.8: Continuous Uniform Distribution

The density function of the continuous uniform random variable X on the interval $[A, B]$ is:

$$f(x: A, B) = \begin{cases} \frac{1}{B-A} & A \leq x \leq B \\ 0 & Otherwise \end{cases}$$

Continuous Uniform Distribution



Note:

a) $\int_{\infty}^{-\infty} f(x)dx = \frac{1}{B-A} \times (B-A) = 1$

b) $P(c < x < d) = \frac{d-c}{B-A}$ where both c and d are in the interval (A,B)

c) $\mu = \frac{A+B}{2}$

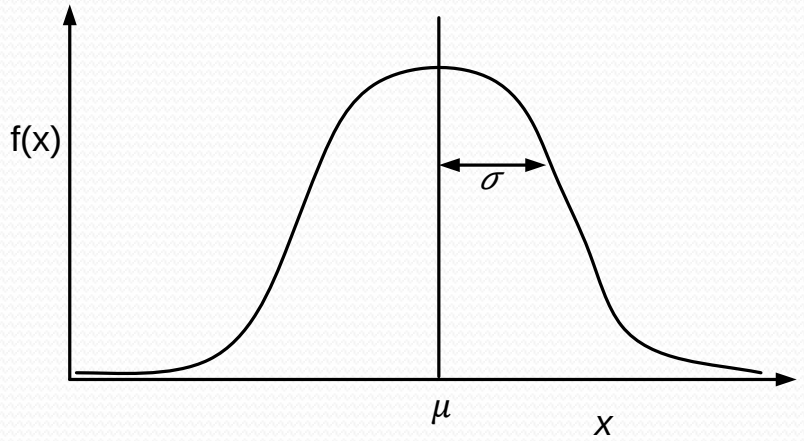
d) $\sigma^2 = \frac{(B-A)^2}{12}$

Normal Distribution

- The most often used continuous probability distribution is the normal distribution; it is also known as **Gaussian** distribution.
- Its graph called the normal curve is the bell-shaped curve.
- Such a curve approximately describes many phenomenon occur in nature, industry and research.
  - Physical measurement in areas such as meteorological experiments, rainfall studies and measurement of manufacturing parts are often more than adequately explained with normal distribution.
- A continuous random variable *X* having the bell-shaped distribution is called a normal random variable.

Normal Distribution

- The mathematical equation for the probability distribution of the normal variable depends upon the two parameters μ and σ, its mean and standard deviation.



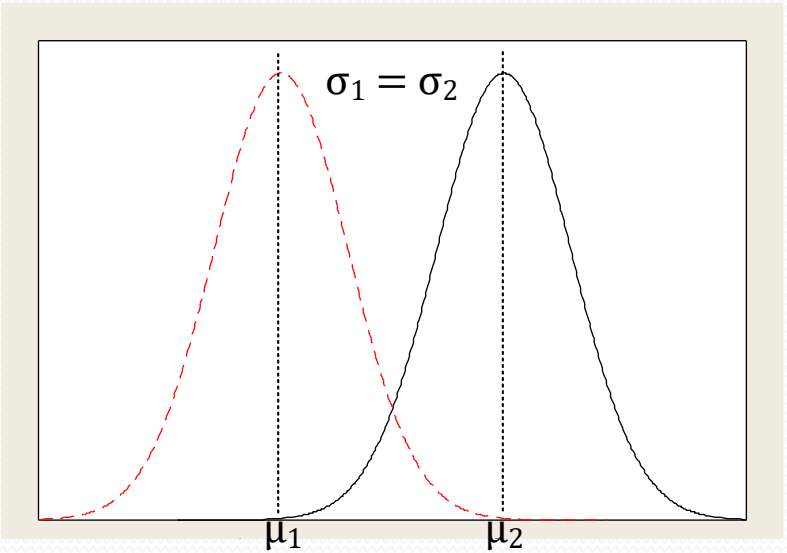
Definition 4.9: Normal distribution

The density of the normal variable x with mean μ and variance σ^2 is

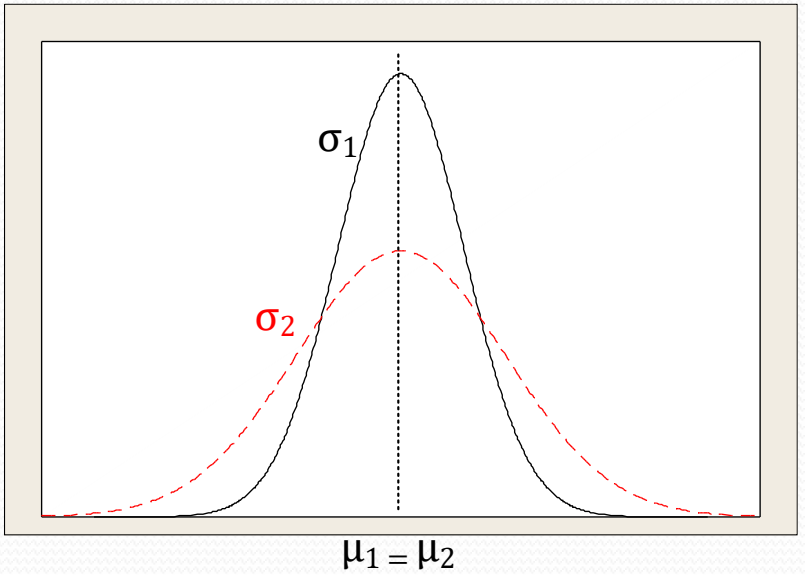
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

where $\pi = 3.14159$ ... and $e = 2.71828$ ……, the Napierian constant

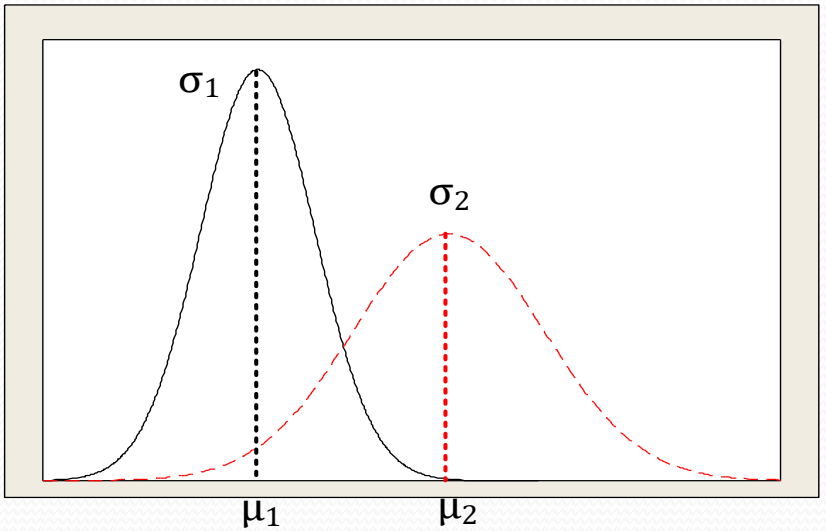
Normal Distribution Curves



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$



Normal curves with $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$

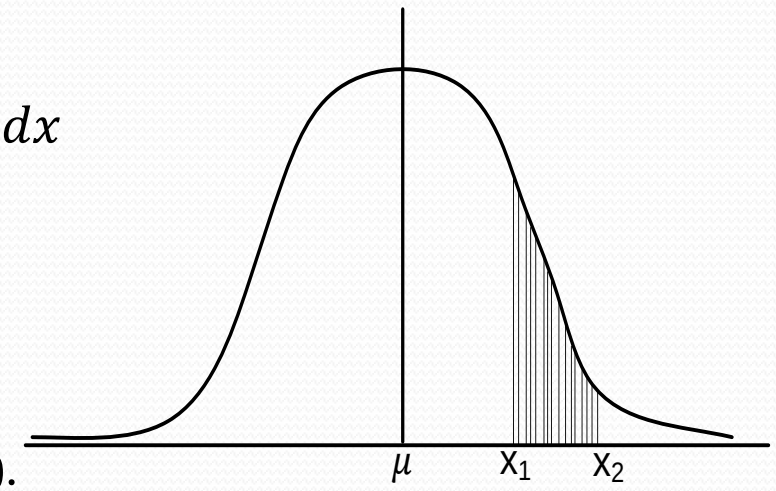
Properties of Normal Distribution

- The curve is symmetric about a vertical axis through the mean μ.
- The random variable x can take any value from $-\infty \; to \; \infty$.
- The most frequently used descriptive parameters define the curve itself.
- The mode, which is the point on the horizontal axis where the curve is a maximum occurs at $x = \mu$.
- The total area under the curve and above the horizontal axis is equal to 1.

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

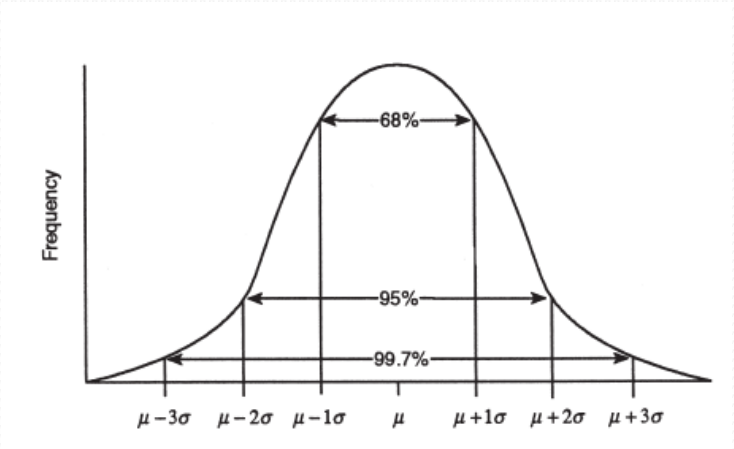
- $\mu = \int_{-\infty}^{\infty} x.f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x.e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$
- $\sigma^2 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2.e^{-\frac{1}{2}[(x-\mu)/\sigma^2]} dx$
- $P(x_1 < x < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$

denotes the probability of x in the interval (x_1, x_2).



Standard Deviation Normal Distribution

- The standard deviation is particularly useful in normal distribution
  - The proportion of elements in the normal distribution (i.e., the proportion of the area under the curve) is a constant for a given number of standard deviations above or below the mean of the distribution



$x = \mu + z.\sigma$

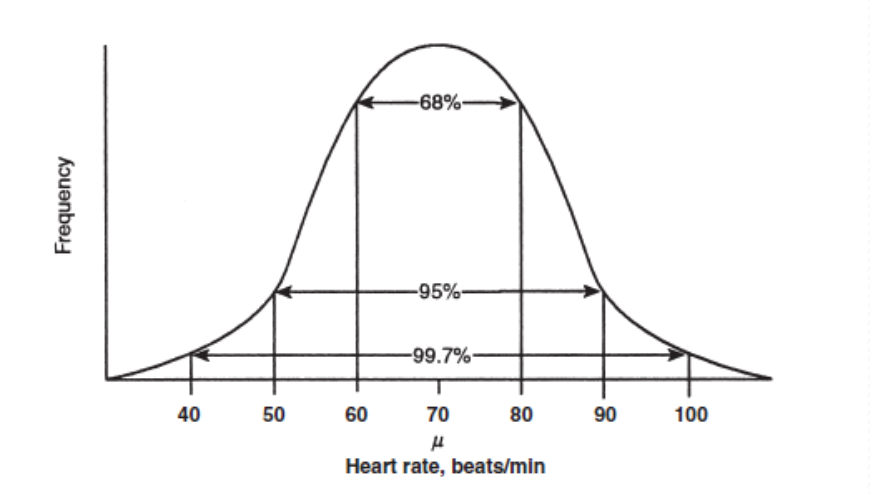
- Approximately 68% of the distribution falls within the ±1 standard deviation of the mean.
- Approximately 95% of the distribution falls within the ±2 standard deviation of the mean.
- Approximately 99.7% of the distribution falls within the ±3 standard deviation of the mean.

Note:

These proportions hold true for every normal distribution

Standard Deviation Normal Distribution

- Example
  - Suppose, a population's resting heart rate is normally distributed with a mean (μ) of 70 and a standard deviation (σ) of 10.



$$x = \mu + z.\sigma$$

  - 68% of the population will have a resting heart rate between 60 and 80.
  - 95% of the population will have a heart rate between approximately 70 ± (2×10), that is, 50 and 90 beats/min.

Z scores: Standard Normal Distribution

- The normal distribution has computational complexity to calculate $P(x_1 < x < x_2)$ for any two (x_1, x_2) and given μ and σ
- To avoid this difficulty, the concept of z-transformation is followed.

$$z = \frac{x-\mu}{\sigma} \qquad \text{[Z-transformation]}$$

- X: Normal distribution with mean μ and variance σ^2.
- Z: Standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.
- Therefore, if *f(x)* assumes a value, then the corresponding value of $f(z)$ is given by

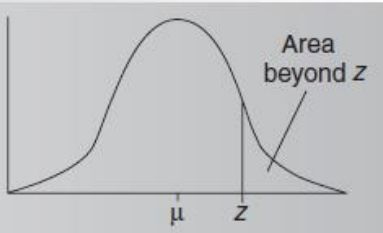
$$f(x:\mu,\sigma): \; P(x_1 < x < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} dz$$

$$= f(z:0,\sigma)$$

Table of z scores

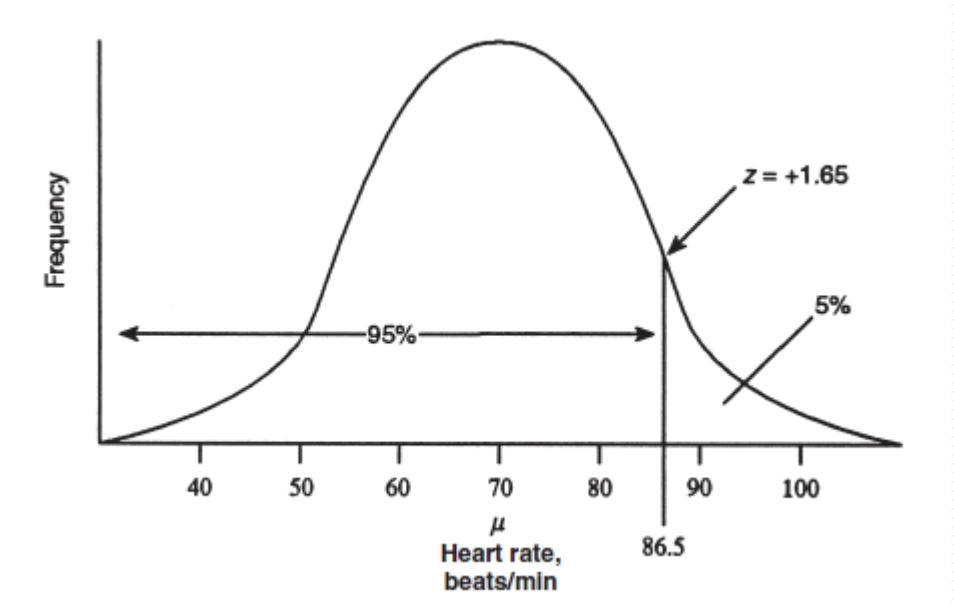
z	Area beyond z	z	Area beyond z
0.00	0.5000	1.65	0.0495
0.05	0.4801	1.70	0.0446
0.10	0.4602	1.75	0.0401
0.15	0.4404	1.80	0.0359
0.20	0.4207	1.85	0.0322
0.25	0.4013	1.90	0.0287
0.30	0.3821	1.95	0.0256
0.35	0.3632	2.00	0.0228
0.40	0.3446	2.05	0.0202
0.45	0.3264	2.10	0.0179
0.50	0.3085	2.15	0.0158
0.55	0.2912	2.20	0.0139
0.60	0.2743	2.25	0.0112
0.65	0.2578	2.30	0.0107
0.70	0.2420	2.35	0.0094
0.75	0.2266	2.40	0.0082
0.80	0.2119	2.45	0.0071
0.85	0.1977	2.50	0.0062
0.90	0.1841	2.55	0.0054
0.95	0.1711	2.60	0.0047
1.00	0.1587	2.65	0.0040
1.05	0.1469	2.70	0.0035
1.10	0.1357	2.75	0.0030
1.15	0.1251	2.80	0.0026
1.20	0.1151	2.85	0.0022
1.25	0.1056	2.90	0.0019
1.30	0.0968	2.95	0.0016
1.35	0.0885	3.00	0.0013
1.40	0.0808	3.05	0.0011
1.45	0.0735	3.10	0.0010
1.50	0.0668	3.15	0.0008
1.55	0.0606	3.20	0.0007
1.60	0.0548	3.30	0.0005



This is a GT table.

Usually the z score values lie in the range -3.3 to +3.3

Table of z scores

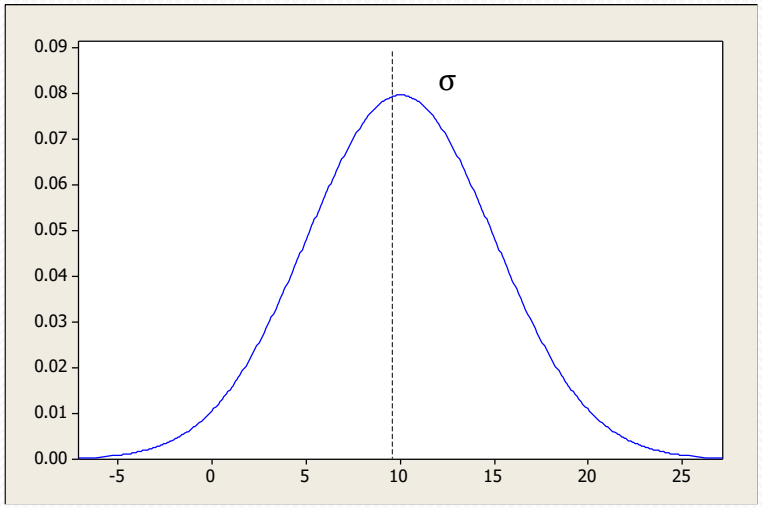


- The nearest figure to 5% (0.05) in the table of the table is 0.0495, the z-sore corresponding to this is 1.65.
- The corresponding heart rate lies 1.65 standard deviations above the mean; that is, it is equal $70+1.65\times10 = 86.5$ We can conclude that 5% of this population has a heart rate above 86.5 beats/min.

Standard Normal Distribution

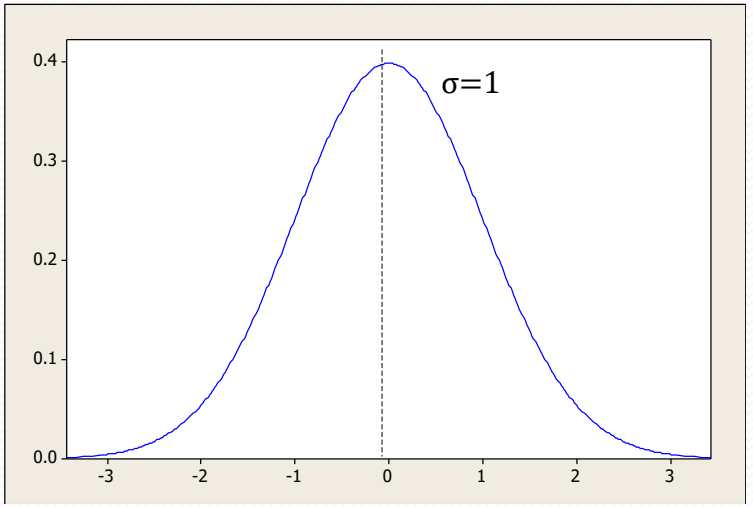
Definition 4.10: Standard normal distribution

The distribution of a normal random variable with mean 0 and variance σ is called a standard normal distribution.



$x=\mu$

$f(x: \mu, \sigma)$



$\mu=0$

$f(z: 0, 1)$

Gamma Distribution

The gamma distribution derives its name from the well known gamma function in mathematics.

Definition 4.11: Gamma Function

$$\Gamma(\alpha) = \int_0^{\alpha} x^{\alpha-1} e^{-x} dx \quad \text{for } \alpha > 0$$

Integrating by parts, we can write,

$$\Gamma(\alpha) = (\alpha - 1) \int_0^{\alpha} x^{\alpha-2} e^{-x} dx$$
$$= (\alpha - 1)\Gamma(\alpha - 1)$$

Thus Γ function is defined as a recursive function.

Gamma Distribution

When $\alpha = n$, we can write,

$\Gamma(n) = (n-1)(n-2) \ldots\ldots\ldots \Gamma(1)$

$= (n-1)(n-2) \ldots\ldots\ldots .3.2.1$

$= (n-1)!$

Further, $\Gamma(1) = \int_0^\infty e^{-x} dx = 1$

Note:

$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ [An important property]

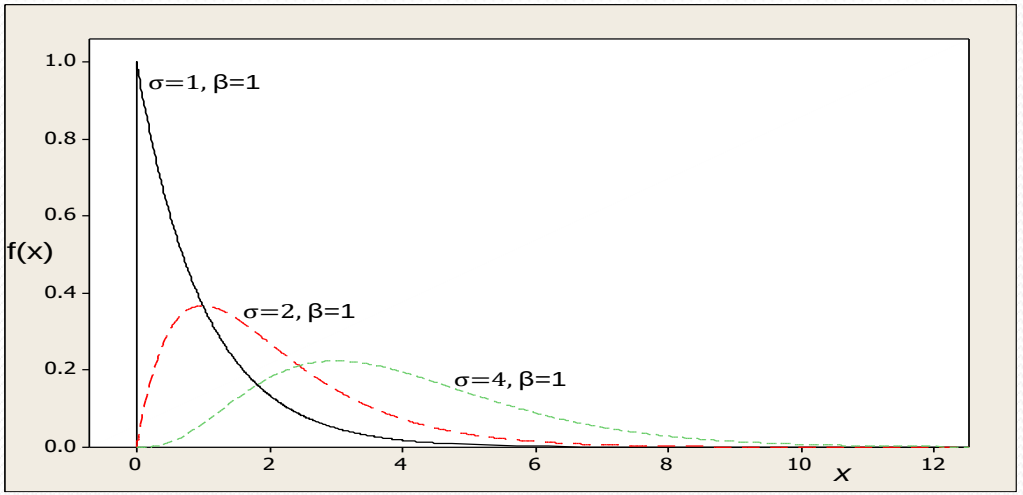
Gamma Distribution

Definition 4.12: **Gamma Distribution**

The continuous random variable x has a gamma distribution with parameters α and β such that:

$$f(x:\alpha,\beta) = \begin{cases} \frac{1}{\beta^{\alpha}\,\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} & x > 0 \\ 0 & Otherwise \end{cases}$$

where $\alpha > 0$ and $\beta > 0$

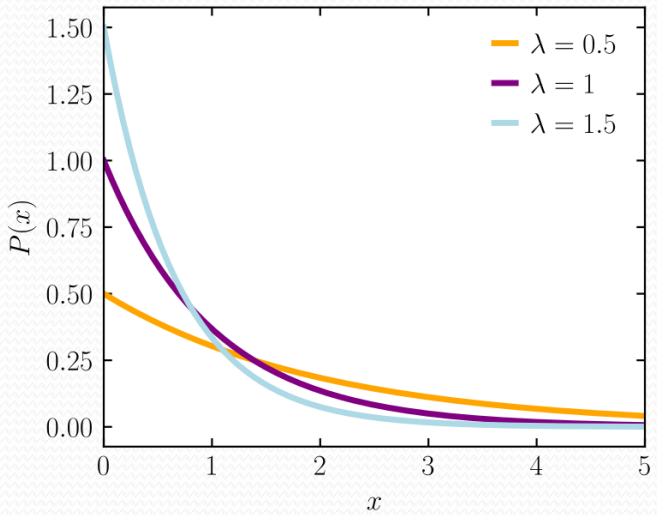


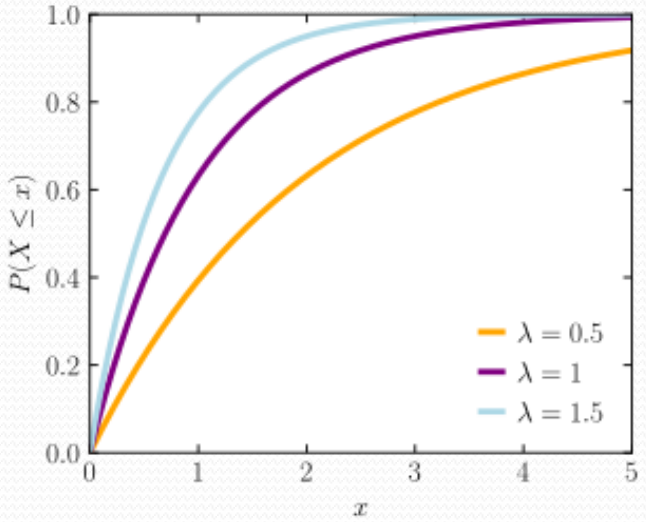
Exponential Distribution

Definition 4.13: **Exponential Distribution**

The continuous random variable x has an exponential distribution with parameter β , where:

$$f(x:\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & \text{where } \beta > 0 \\ 0 \end{cases}$$





Exponential Distribution

Definition 4.13: **Exponential Distribution**

The continuous random variable x has an exponential distribution with parameter β , where:

$$f(x:\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & \text{where } \beta > 0 \\ 0 \end{cases}$$

Note:

1) The mean and variance of gamma distribution are

$$\mu = \alpha\beta$$
$$\sigma^2 = \alpha\beta^2$$

2) The mean and variance of exponential distribution are

$$\mu = \beta$$
$$\sigma^2 = \beta^2$$

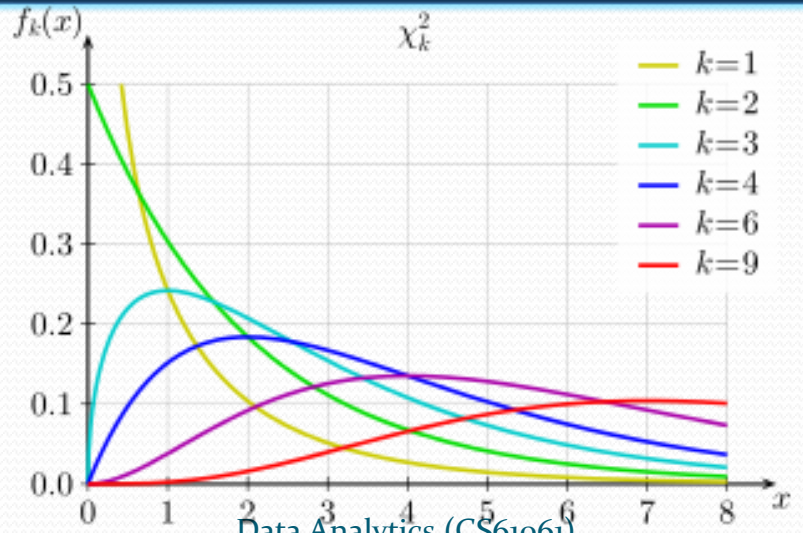
Chi-Squared Distribution

Definition 4.14: Chi-squared distribution

The continuous random variable x has a Chi-squared distribution with v degrees of freedom, is given by

$$f(x:v) = \begin{cases} \dfrac{1}{2^{\frac{v}{2}}\,\Gamma(v/2)}\, x^{v/2-1} e^{-\frac{x}{2}}, & x > 0 \\ 0 & Otherwise \end{cases}$$

where v is a positive integer.



Chi-Squared Distribution

Definition 4.14: Chi-squared distribution

The continuous random variable x has a Chi-squared distribution with v degrees of freedom, is given by

$$f(x:v) = \begin{cases} \frac{1}{2^{\frac{v}{2}}\,\Gamma(v/2)} x^{v/2-1} e^{-\frac{x}{2}}, & x > 0 \\ 0 & Otherwise \end{cases}$$

where v is a positive integer.

- The Chi-squared distribution plays an important role in statistical inference .
- The mean and variance of Chi-squared distribution are:

$$\mu = v \text{ and } \sigma^2 = 2v$$

Lognormal Distribution

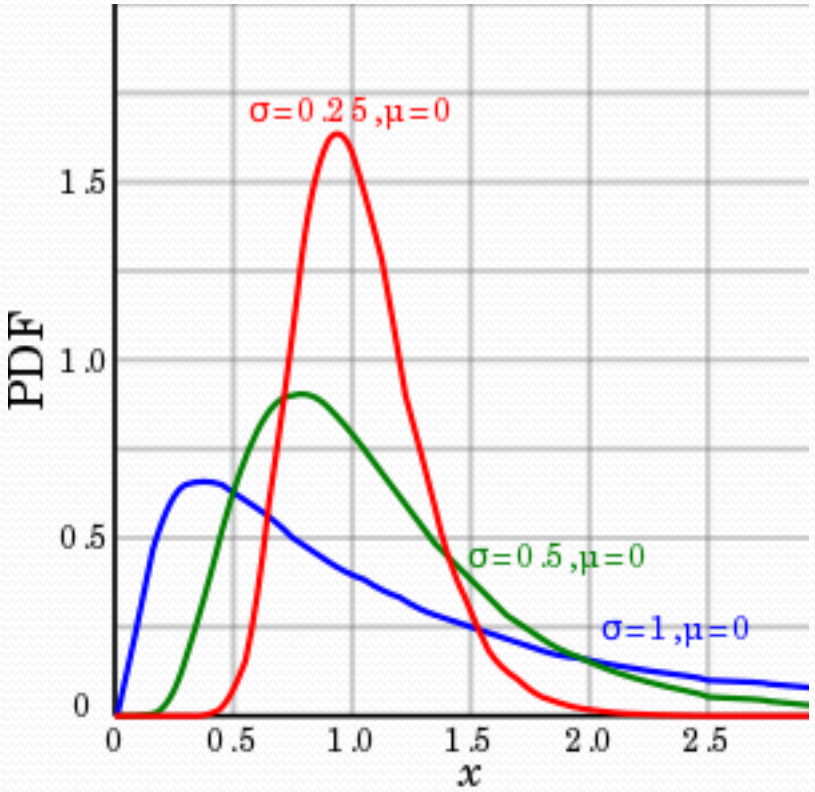
The lognormal distribution applies in cases where a natural log transformation results in a normal distribution.

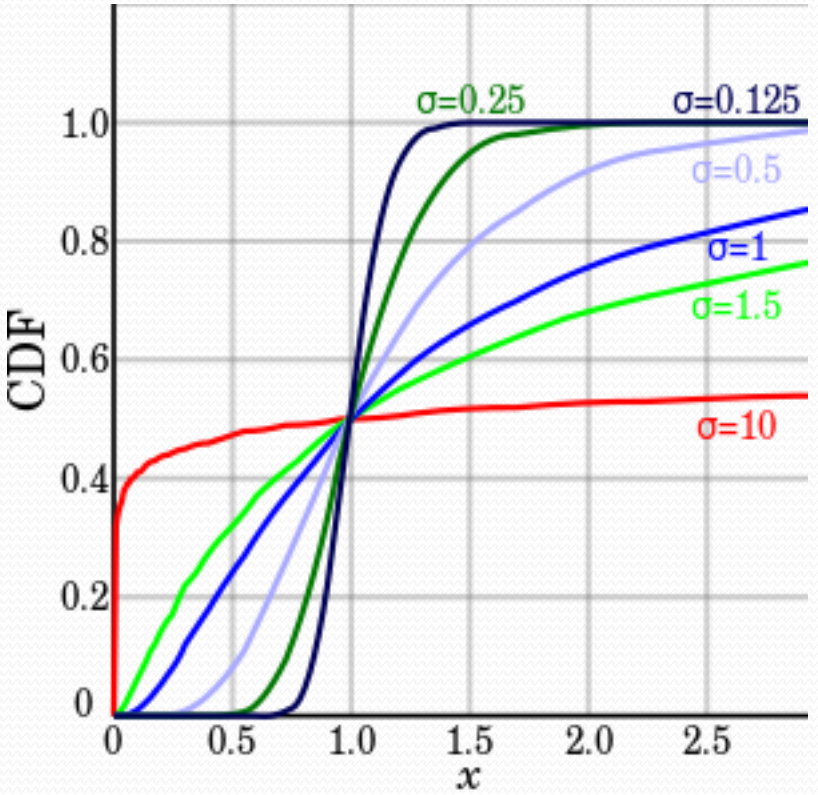
Definition 4.15: Lognormal distribution

The continuous random variable x has a lognormal distribution if the random variable $y = \ln(x)$ has a normal distribution with mean μ and standard deviation σ. The resulting density function of x is:

$$f(x:\mu,\sigma) = \begin{cases} \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Lognormal Distribution





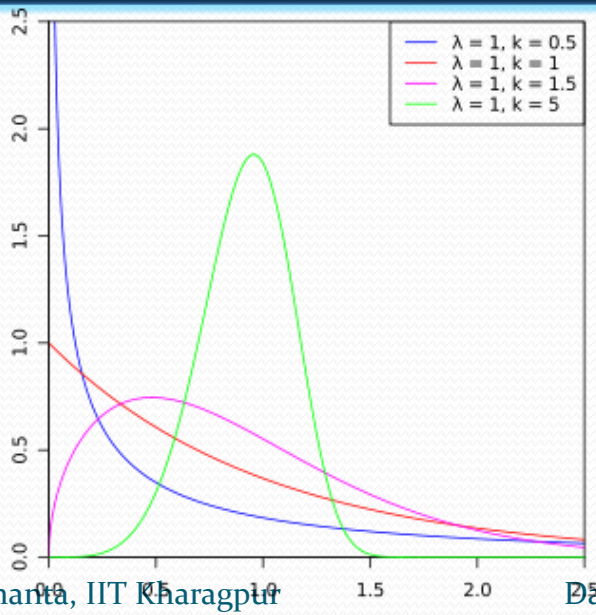
Weibull Distribution

Definition 4.16: Weibull Distribution

The continuous random variable x has a Weibull distribution with parameter α and β such that.

$$f(x:\alpha,\beta) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}} & x > 0 \\ 0 & Otherwise \end{cases}$$

where $\alpha > 0$ and $\beta > 0$



The mean and variance of Weibull distribution are:

$$\mu = \alpha^{-1/\beta}\Gamma\left(1+\frac{1}{\beta}\right)$$

$$\sigma^2 = \alpha^{-2/\beta}\{\Gamma\left(1+\frac{2}{\beta}\right) - [\Gamma(1+\frac{1}{\beta})]^2\}$$

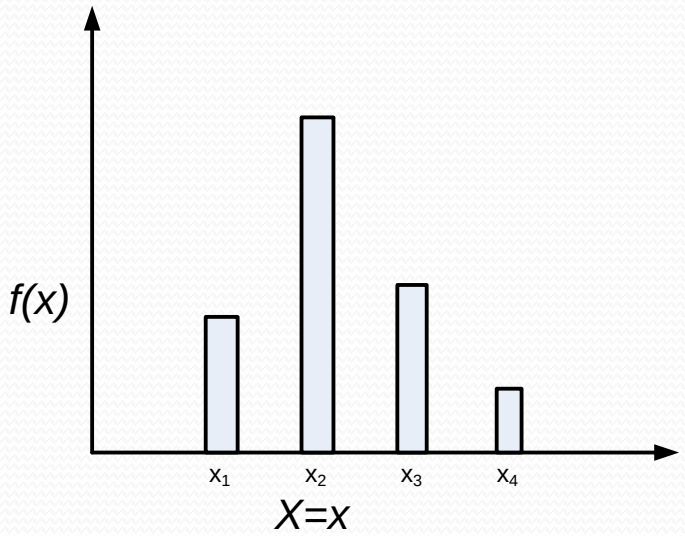
Important note

- Probability mass function
  - A probability mass function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value.
    - Sometimes it is also known as the discrete density function.
- Probability density function
  - A probability density function (PDF) in that is associated with continuous rather than discrete random variables.
- Note:
  - A PDF must be integrated over an interval to yield a probability.

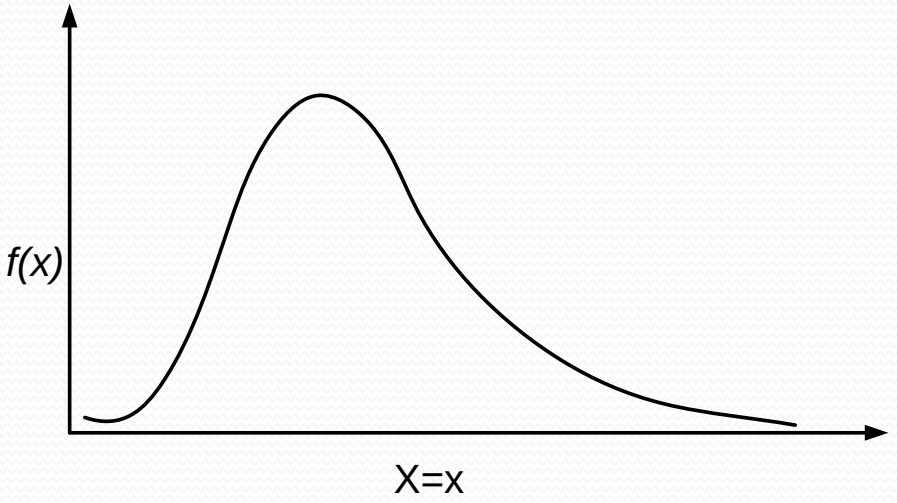
Sample Statistics

Random variable and PDF

$$P(X = x) = f(x)$$



Discrete Probability distribution



Continuous Probability Distribution

In the next part of discussion…

- Basic concept of sampling distribution
- Usage of sampling distributions
- Issue with sampling distributions
- Central Limit Theorem
- Application of Central Limit Theorem
- Major sampling distributions
  - **χ^2 distribution**
  - **t-distribution**
  - **F distribution**

Statistical inference

As a task of statistical inference, we usually follow the following steps:

- **Data collection**
  - Collect a sample from the population.
- **Statistics**
  - Compute a statistics from the sample.
- **Statistical inference**
  - From the statistics we made various statements concerning the values of population parameters.
    - For example, population mean from the sample mean, etc.

Basic terminologies

Some basic terminology which are closely associated to the above-mentioned tasks are reproduced below.

- **Population:** A population consists of the totality of the observation, with which we are concerned.
- **Sample:** A sample is a subset of a population.
- **Random variable:** A random variable is a function that associates a real number with each element in the sample.
- **Statistics:** Any function of the random variable constituting random sample is called a statistics.
- **Statistical inference:** It is an analysis basically concerned with generalization and prediction.

Statistical learning

There are two facts, which are key to statistical inference.

1. Population parameters are fixed number whose values are usually unknown.
2. Sample statistics are known values for any given sample, but vary from sample to sample, even taken from the same population.

- In fact, it is unlikely for any two samples drawn independently, producing identical values of sample statistics.
- In other words, the variability of sample statistics is always present and must be accounted for in any inferential procedure.
- This variability is called sampling variation.

Note:

A sample statistics is random variable and like any other random variable, a sample statistics has a probability distribution.

Note: Probability distribution for random variable is not applicable to sample statistics.

Sampling Distribution

More precisely, sampling distributions are probability distributions and used to describe the variability of sample statistics.

Definition 4.17: Sampling distribution

The sampling distribution of a statistics is the probability distribution of that statistics.

- The probability distribution of sample mean (hereafter, will be denoted as $\bar{X}$) is called the sampling distribution of the mean (also, referred to as the distribution of sample mean).
- Like $\bar{X}$, we call sampling distribution of variance (denoted as S^2).
- Using the values of $\bar{X}$ and S^2 for different random samples of a population, we are to make inference on the parameters μ and σ^2 (of the population).

Sampling Distribution

Example 5.1:

Consider five identical balls numbered and weighting as 1, 2, 3, 4 and 5. Consider an experiment consisting of drawing two balls, replacing the first before drawing the second, and then computing the mean of the values of the two balls.

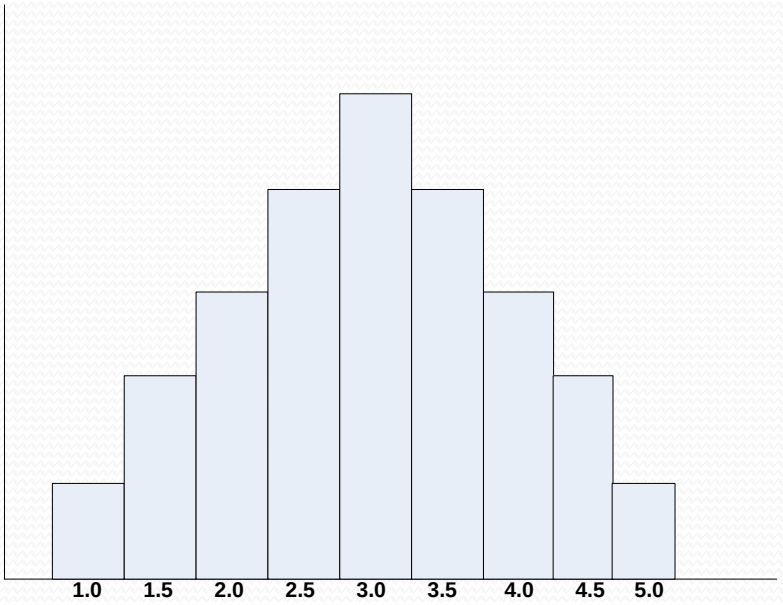
Following table lists all possible samples and their mean.

Sample (X)	Mean ($\bar{X}$)	Sample (X)	Mean ($\bar{X}$)	Sample (X)	Mean ($\bar{X}$)
[1,1]	1.0	[2,4]	3.0	[4,2]	3.0
[1,2]	1.5	[2,5]	3.5	[4,3]	3.5
[1,3]	2.0	[3,1]	2.0	[4,4]	4.0
[1,4]	2.5	[3,2]	2.5	[4,5]	4.5
[1,5]	3.0	[3,3]	3.0	[5,1]	3.0
[2,1]	1.5	[3,4]	3.5	[5,2]	3.5
[2,2]	2.0	[3,5]	4.0	[5,3]	4.0
[2,3]	2.5	[4,1]	2.5	[5,4]	4.5
				[5,5]	5.0

Sampling Distribution

Sampling distribution of means

$\bar{X}$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$f(\bar{X})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{1}{25}$



Issues with Sampling Distribution

1. In practical situation, for a large population, it is infeasible to have all possible samples and hence frequency distribution of sample statistics.

2. The sampling distribution of a statistics depends on
   - the size of the population
   - the size of the samples and
   - the method of choosing the samples.

?

Theorem on Sampling Distribution

Famous theorem in Statistics

Theorem 4.18: Sampling distribution of mean and variance

The sampling distribution of a random sample of size n drawn from a population with mean μ and variance σ^2 will have mean $\bar{X} = \mu$ and variance $S^2 \approx \frac{\sigma^2}{n}$

Example 5.2: With reference to data in Example 5.1

For the population, $\mu = \frac{5+1}{2} = 3$

$$\sigma^2 = \frac{5^2-1}{12} = 2$$

Applying the theorem, we have $\bar{X} = 3 \; and \; S^2 \approx 0.4$

Hence, the theorem is verified!

Central Limit Theorem

The Theorem 4.1 is an amazing result and in fact, also verified that if we sampling from a population with unknown distribution, the sampling distribution of $\bar{X}$ will still be approximately normal with mean μ and variance $\frac{\sigma^2}{n}$ provided that the sample size is large.

This further, can be established with the famous "Central Limit Theorem", which is stated below.

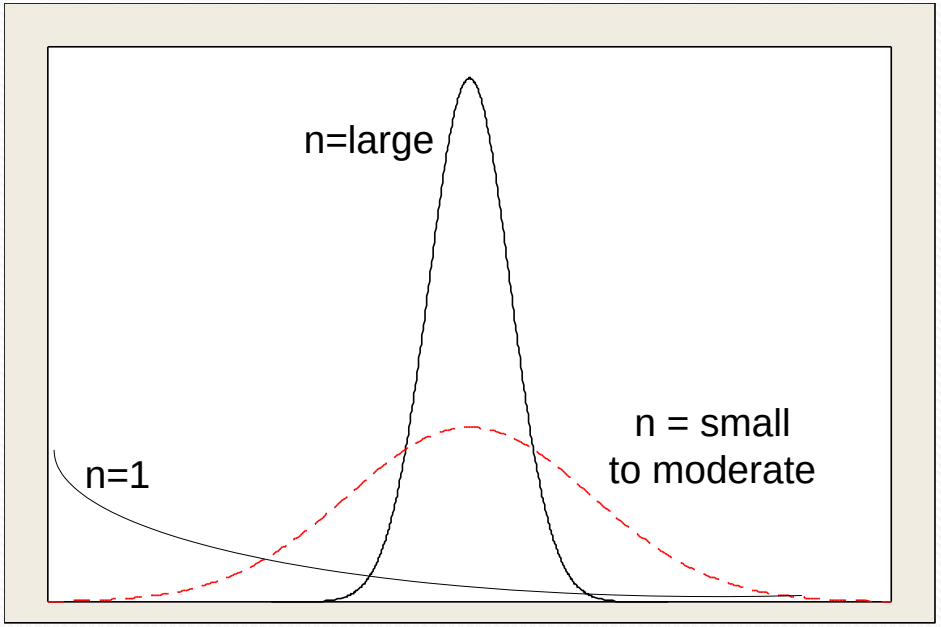
Theorem 4.1: Central Limit Theorem

If random samples each of size n are taken from any distribution with mean μ and variance σ^2, the sample mean $\bar{X}$ will have a distribution approximately normal with mean μ and variance $\frac{\sigma^2}{n}$.

The approximation becomes better as n increases.

Applicability of Central Limit Theorem

- The normal approximation of $\overline{X}$ will generally be good if $n \geq 30$
- The sample size $n = 30$ is, hence, a guideline for the central limit theorem.
- The normality on the distribution of $\overline{X}$ becomes more accurate as n grows larger.



Usefulness of the sampling distribution

- The mean of the sampling distribution of the mean is the population mean.
  - This implies that "on the average" the sample mean is the same as the population mean.
  - We therefore say that the sample mean is an unbiased estimate of the population mean.
- The variance of the distribution of the sample means is σ^2/n.
  - The standard deviation of the sampling distribution (i.e., $\frac{\sigma}{\sqrt{n}}$) of the mean, often called the standard error of the mean.
    - If σ is high then the sample are not reliable, for a very large sample size ($n \to \infty$), standard error tends to zero

Applicability of Central Limit Theorem

- One very important application of the **Central Limit Theorem** is the determination of reasonable values of the population mean μ and variance σ^2 having a sample, that is, a subset of a population.

  - One very important deduction

    For standard normal distribution, we have the z-transformation

    $$z = \frac{x-\mu}{\sigma}$$ (See Slide#43)

    Thus, for a sample statistics

    $$Z = \frac{\overline{X}-\mu}{S} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$$

Applicability of Central Limit Theorem

Example:

- A quiz test for the course CS61061 was conducted and it was found that mean of the scores $\mu = 90$ with standard deviation $\sigma = 20$.
- Now, all students enrolled in the course are randomly assigned to various sections of 100 students in each. A section (X) was checked and the mean score was found as $\bar{X} = 86$.

- **What is the standard error rate?**

  The standard error rate (Central Limit Theorem) = $\frac{\sigma}{\sqrt{n}} = \frac{20}{100} = 2.0$

- **What is the probability of getting a mean of 86 or lower on the quiz test?**

  For standard normal distribution, we have the z-transformation

$$z = \frac{x-\mu}{\sigma}$$

  Thus, for a sample statistics

$$Z = \frac{\bar{X}-\mu}{s} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{86-90}{2} = -2. \qquad P(Z<-2)?$$

Standard Sampling Distributions

- Apart from the normal distribution to describe sampling distribution, there are some other quite different sampling, which are extensively referred in the study of statistical inference.
  - χ^2: Describes the distribution of variance.
  - t: Describes the distribution of normally distributed random variable standardized by an estimate of the standard deviation.
  - F: Describes the distribution of the ratio of two variables.

Chi-square Distribution

The χ^2 Distribution

- A common use of the χ^2 distribution is to describe the distribution of the sample variance.

- Let $X_1, X_2, \ldots, X_n$ be are independent random variables from a normally distributed population with mean = μ and variance = σ^2.
- The χ^2 distribution can be written as

$$\chi^2 = (Z_1)^2 + (Z_2)^2 + (Z_3)^2 + \cdots + (Z_n)^2$$

where $Z_i = \frac{(X_i - \mu)}{S}$.

- This χ^2 is also a random variable of a distribution and is called χ^2-distribution (pronounced as Chi-square distribution).

The χ^2 Distribution

Definition 4.19: χ^2-distribution for Sampling Variance

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistics

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma}\right)^2$$

has a chi-squared distribution with $v = n - 1$ degrees of freedom and variance is $2v$

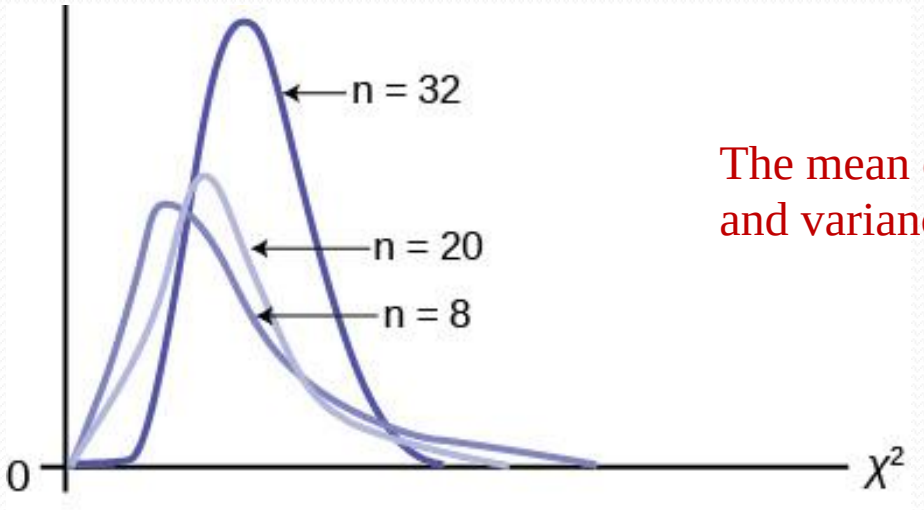
How the χ^2- distribution is used to describe the sampling distribution of S^2?

The χ^2 Distribution

- The χ^2 distribution can be written as

$$\chi^2 = \sum Z^2 = \sum \left(\frac{X-\bar{X}}{\sigma}\right)^2 = \frac{SS}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

- This expression χ^2 describes the distribution (of n samples) and thus having degrees of freedom $\upsilon = n\text{-}1$ and often written as $\chi^2(\upsilon)$, where υ is the only parameter in it.



The mean of this distribution is υ and variance is 2υ

Some facts about χ^2 distribution

- The curves are non symmetrical and skewed to the right.
- χ^2 values cannot be negative since they are sums of squares.
- The mean of the χ^2 distribution is ν, and the variance is 2ν.
- When $\upsilon > 30$, the Chi-square curve approximates the normal distribution. Then, you may write the following

$$Z = \frac{\chi^2 - \upsilon}{\sqrt{2\upsilon}}$$

χ^2 is distribution of sample variances

- A common use of the χ^2 distribution is to describe the distribution of the sample variance. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normally distributed population with mean = μ and variance = σ^2. Then the quantity $(n-1)S^2/\sigma^2$ is a random variable whose distribution is described by a χ^2 distribution with $(n-1)$ degrees of freedom, where S^2 is the usual sample estimate of the population variance. That is

$$S^2 = \frac{(X-\bar{X})^2}{n-1}$$

- In other words, the χ^2 distribution is used to describe the sampling distribution of S^2. Since we divide the sum of squares by degrees of freedom to obtain the variance estimate, the expression for the random variable having a $\chi 2$ distribution can be written

$$\chi^2 = \sum Z^2 = \sum \left(\frac{X-\bar{X}}{\sigma}\right)^2 = \frac{SS}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

Application χ^2 of values

Example: Judging the quality of a machine

A machine is to produce a ball of 100gm. It is desirable to have maximum deviation of 0.01gm (this is the desirable value of σ).

Suppose, 15 balls produced by the machine are select at random and it shows S = 0.0125gm.

What is the probability that the machine will produces an accurate ball?

χ^2 calculation can help us to know this value.

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{14 \times (0.0125)^2}{(0.01)^2} = 21.875$$

This is the χ^2 value with 14 degrees of freedom. The value can be tested with χ^2 table to know the desired probability value.

t-Distribution

The t Distribution

- **The t Distribution**

1. To know the sampling distribution of mean we make use of Central Limit Theorem with $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$
2. This require the known value of σ a priori.
3. However, in many situation, σ is certainly no more reasonable than the knowledge of the population mean μ.
4. In such situation, only measure of the standard deviation available may be the sample standard deviation S.
5. It is natural then to substitute S for σ. The problem is that the resulting statistics not necessarily be normally distributed!
6. The t distribution is to alleviate this problem. This distribution is called $student's\ t$ or simply $t - distribution$.

The t Distribution

- **The t Distribution**

Definition 4.20: t −distribution

The t −distribution with v degrees of freedom actually takes the form

$$t(v) = \frac{Z}{\sqrt{\frac{\chi^2(v)}{v}}}$$

where Z is a standard normal random variable, and $\chi^2(v)$ is χ^2 random variable with v degrees of freedom.

The t Distribution

Corollary: Let $X_1, X_2, \ldots \ldots, X_n$ be independent random variables that are all normal with mean μ and standard deviation σ.

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Using this definition, we can develop the sampling distribution of the sample mean when the population variance, σ^2 is unknown.

That is,

$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has the standard normal distribution.

$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ has the χ^2 distribution with $(n-1)$ degrees of freedom.

Thus, $T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}}$ or $\boxed{T = \frac{\bar{X}-\mu}{S/\sqrt{n}}}$

This is the $t - distribution$ with $(n-1)$ degrees of freedom.

F Distribution

The F Distribution

- The F distribution finds enormous applications in comparing sample variances.

Definition 4.21: F distribution

The statistics F is defined to be the ratio of two independent Chi-Squared random variables, each divided by its number of degrees of freedom. Hence,

$$F(v_1, v_2) = \frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2}$$

Corollary: Recall that $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ is the Chi-squared distribution with $(n-1)$ degrees of freedom.

Therefore, if we assume that we have sample of size n_1 from a population with variance ${\sigma_1}^2$ and an independent sample of size n_2 from another population with variance ${\sigma_2}^2$, then the statistics

$$F = \frac{{S_1}^2/{\sigma_1}^2}{{S_2}^2/{\sigma_2}^2}$$

Summary of sampling distributions

Z – distribution:

- Typically it is used for comparing the mean of a sample to some hypothesized mean for the population in case of large sample, or when **population variance is known**.

t – distribution:

- **population variance is not known.** In this case, we use the variance of the sample as an estimate of the population variance.

χ^2 – distribution:

- It is used for comparing a sample variance to a theoretical population variance.

F – distribution:

- It is used for comparing the variance of two or more populations.

Reference

- The detail material related to this lecture can be found in

Probability and Statistics for Engineers and Scientists (8th Ed.) by Ronald E. Walpole, Sharon L. Myers, Keying Ye (Pearson), 2013.

Any question?

Questions of the day…

1. Give some examples of random variables? Also, tell the range of values and whether they are with continuous or discrete values.

2. In the following cases, what are the probability distributions are likely to be followed. In each case, you should mention the random variable and the parameter(s) influencing the probability distribution function.
   a) In a retail source, how many counters should be opened at a given time period.
   b) Number of people who are suffering from cancers in a town?

Questions of the day…

2. In the following cases, what are the probability distributions are likely to be followed. In each case, you should mention the random variable and the parameter(s) influencing the probability distribution function.
   c) A missile will hit the enemy's aircraft.
   d) A student in the class will secure EX grade.
   e) Salary of a person in an enterprise.
   f) Accident made by cars in a city.
   g) People quit education after i) primary ii) secondary and iii) higher secondary educations.

Questions of the day…

3. How you can calculate the mean and standard deviation of a population if the population follows the following probability distribution functions with respect to an event.
   a) Binomial distribution function.
   b) Poisson's distribution function.
   c) Hypergeometric distribution function.
   d) Normal distribution function.
   e) Standard normal distribution function.

Questions of the day…

4. What are the degrees of freedom in the following cases.

   Case 1: A single number.

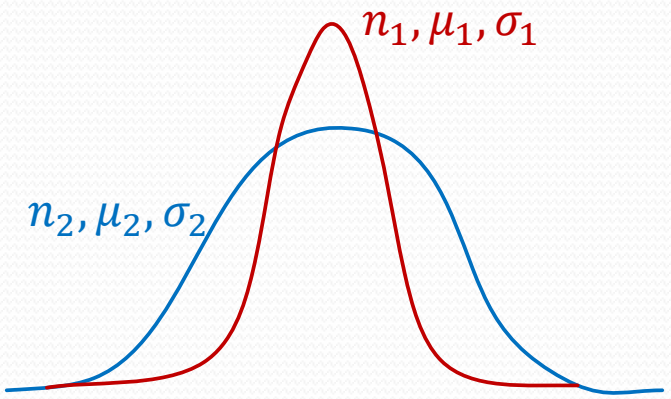
   Case 2: A list of n numbers.

   Case 3: a table of data with m rows and n columns.

   Case 4: a data cube with dimension $m \times n \times p$.

Questions of the day…

5. In the following, two normal sampling distributions are shown with parameters n, μ and σ (all symbols bear their usual meanings).



What are the relations among the parameters in the two?

Questions of the day…

6. Suppose, $\bar{X}$ and S denote the sample mean and standard deviation of a sample. Assume that population follows normal distribution with population mean μ and standard deviation σ. Write down the expression of z and t values with degree of freedom n.