BME 1132
Probability and Biostatistics

Instructor: Ali AJDER, *Ph.D.*

Week-10

- Introduction
- General Concepts of Continuous Probability Distributions
- and Probability Density Function
- The Normal Distribution
- Properties of the Standard Normal Distribution
- Conversion to the Standard Normal Distribution
- Examples about Continuous and Discrete Probability Distributions

Introduction-1

For discrete random variables, the **PMF** provides the probability of each possible value.
For continuous random variables, the number of possible values is **uncountable**,
and the probability of any specific value is **zero**.

Intuitively, we think about allocating the total probability of 1 among uncountable number of possible values. Therefore, instead of talking about the probability of any specific value x for continuous random variable X, we talk about the probability that the value of the random variable is within a **specific interval** from $x_1 \; to \; x_2$.

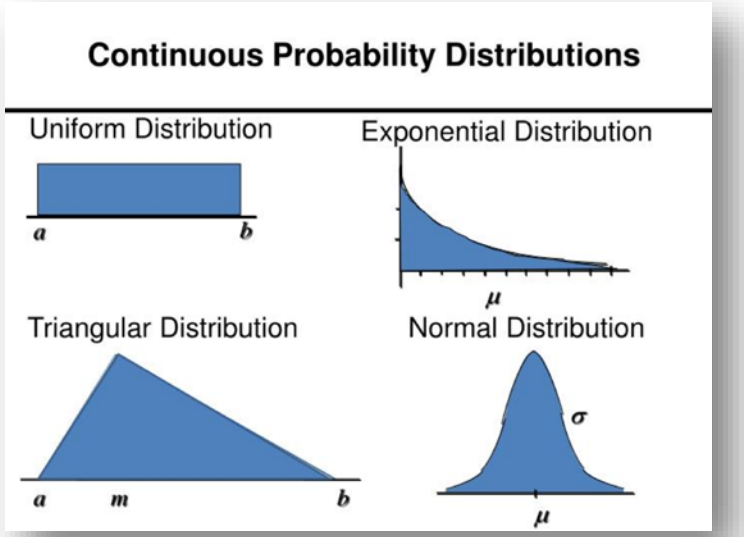
For continuous random variables, we use Probability Density Functions (**PDF**) to specify the distribution. Using the PDF, we can obtain the probability of any interval.

Introduction-2

We will discuss continuous probability distributions in this lecture.
Specifically, ***the normal distribution***—the most widely used distribution in statistical work—

*(**Gaussian*** or "***bell-shaped***," distribution)

Many random variables, such as distribution of birthweights or blood pressures in the general population, tend to *approximately follow a normal distribution*.



In addition, many random variables that are not themselves normal are closely approximated by a normal distribution when summed many times.
In such cases, using the normal distribution is desirable because it is easy to use.

General Concepts & Probability Density Function

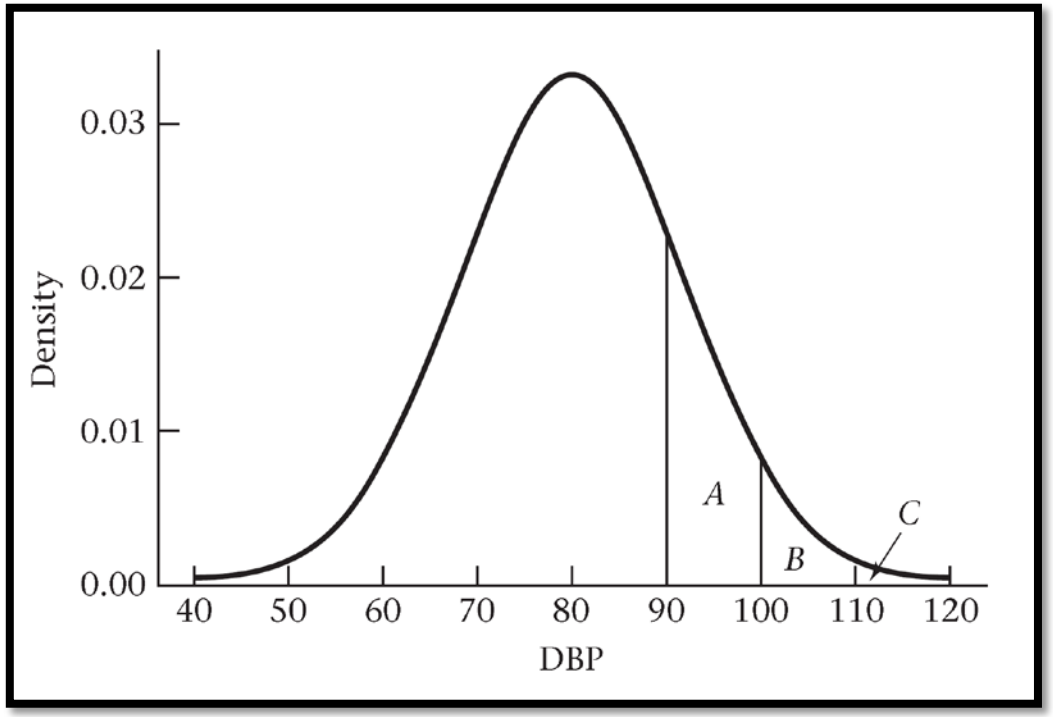
Now, we would like to know which values are more probable than others for a continuous random variable and how probable they are.

Although the probability of exactly obtaining any value is 0, people still have the intuitive notion that certain ranges of values occur more frequently than others. This notion can be quantified using the concept of a **Probability Density Function (PDF).**

The **probability-density function** of the random variable X is a function such that the area under the density-function curve between any two points a and b is equal to the probability that the random variable X falls between a and b. Thus, the total area under the density-function curve over the entire range of possible values for the random variable is 1.

Probability Density Function- Example

Hypertension A PDF for Diastolic Blood Pressure (DBP) in 35- to 44-year-old men is shown in the following Figure. Areas *A*, *B*, and *C* correspond to the probabilities of being mildly hypertensive, moderately hypertensive, and severely hypertensive, respectively. Furthermore, the most likely range of values for DBP occurs around 80 mm Hg, with the values becoming increasingly less likely as we move farther away from 80.

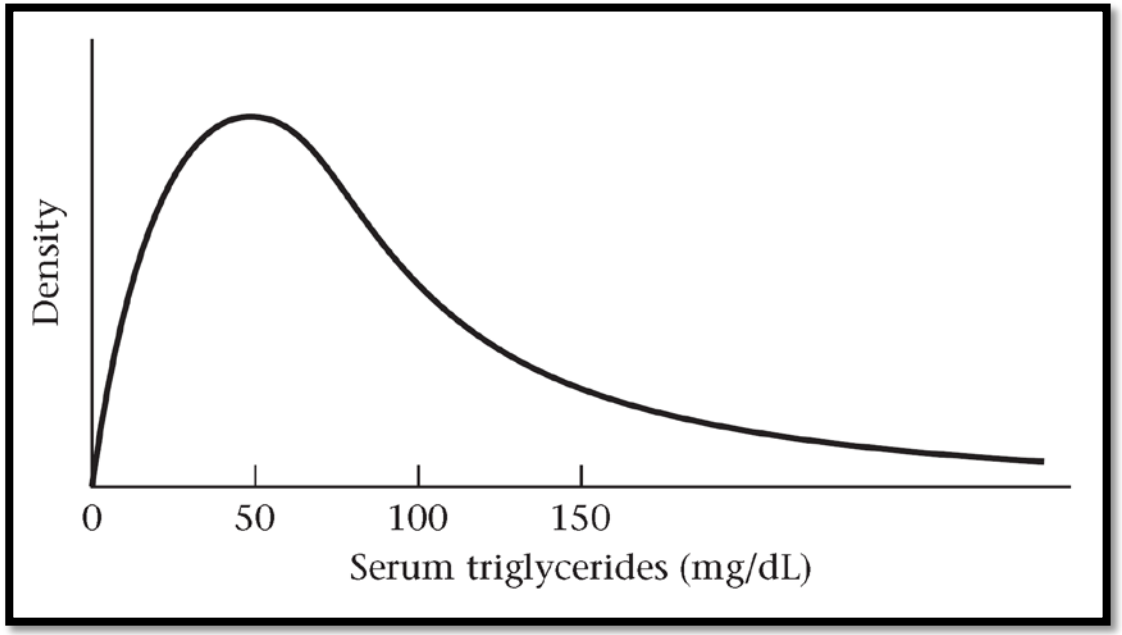


Note:

NOT all continuous random variables have *symmetric bell-shaped distributions* as in Figure.

Probability Density Function- Example

Cardiovascular Disease Serum triglyceride level is an asymmetric, positively skewed, continuous random variable whose PDF appears in the following Figure.



Cumulative Distribution Function

The Cumulative Distribution Function (or CDF) is defined similarly to that for a discrete random variable (see previous Lecture Slides).

The cumulative-distribution function for the random variable X evaluated at the point a is defined as the probability that X will take on values $\leq a$. It is represented by the area under the pdf to the left of a.

The Expected Value and Variance of a Continuous RV

The **expected value** of a continuous random variable X, denoted by $E(X)$, or μ, is the average value taken on by the random variable.

The **variance** of a continuous random variable X, denoted by $Var(X)$ or σ^2, is the average squared distance of each value of the random variable from its expected value, which is given by $E(X - \mu)^2$ and can be re-expressed in short form as $E(X^2) - \mu^2$. The standard deviation, or σ, is the square root of the variance, that is, $\sigma = \sqrt{Var(X)}$.

Example

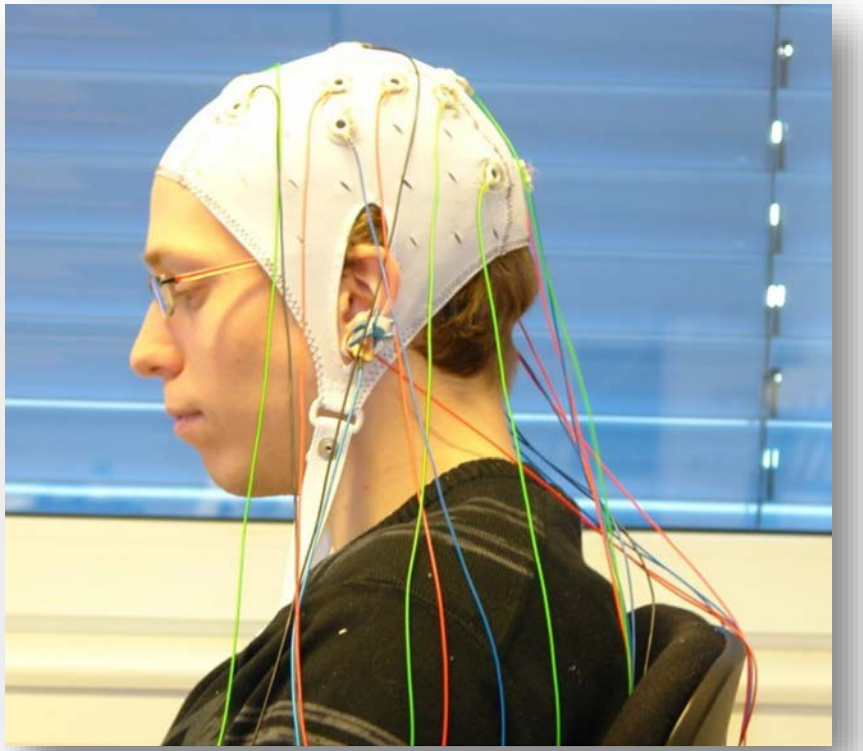
Hypertension The expected value and standard deviation of the distribution of DBP in 35- to 44-year-old men are 80 and 12 mm Hg, respectively.

The Normal Distribution

The normal distribution is the most widely used continuous distribution.

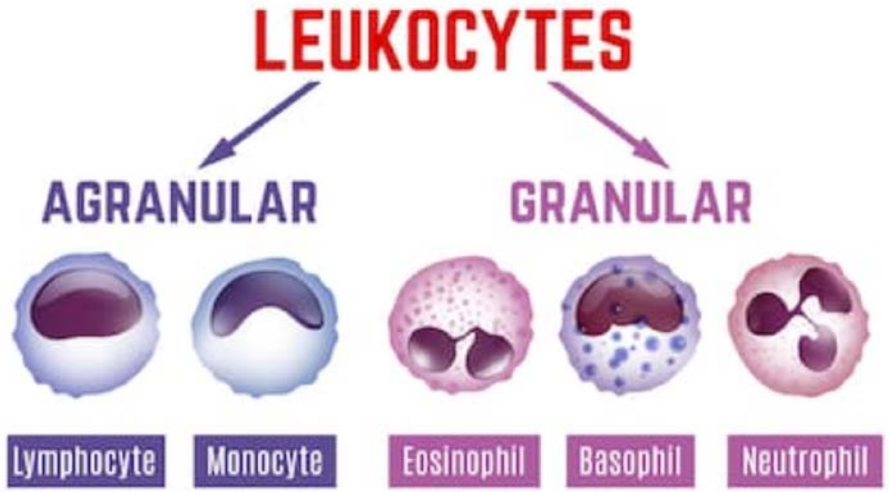
Many other distributions that are not themselves normal can be made approximately normal by transforming the data onto a different scale.

Generally speaking, any random variable that can be expressed as ***a sum of many other random variables*** can be well **approximated by a normal distribution**.
For example, many physiologic measures are determined in part by a combination of several genetic and environmental risk factors and can often be well approximated by a normal distribution.



The Normal Distribution- Example

Infectious Disease The number of lymphocytes in a differential of 100 white blood cells (see previous lecture slides for the definition of a differential) tends to be normally distributed because this random variable is a sum of 100 random variables, each representing whether or not an individual cell is a lymphocyte.



Thus, because of its omnipresence the normal distribution is vital to statistical work, and most estimation procedures and hypothesis tests that we will study assume *the random variable being considered has an underlying normal distribution*.

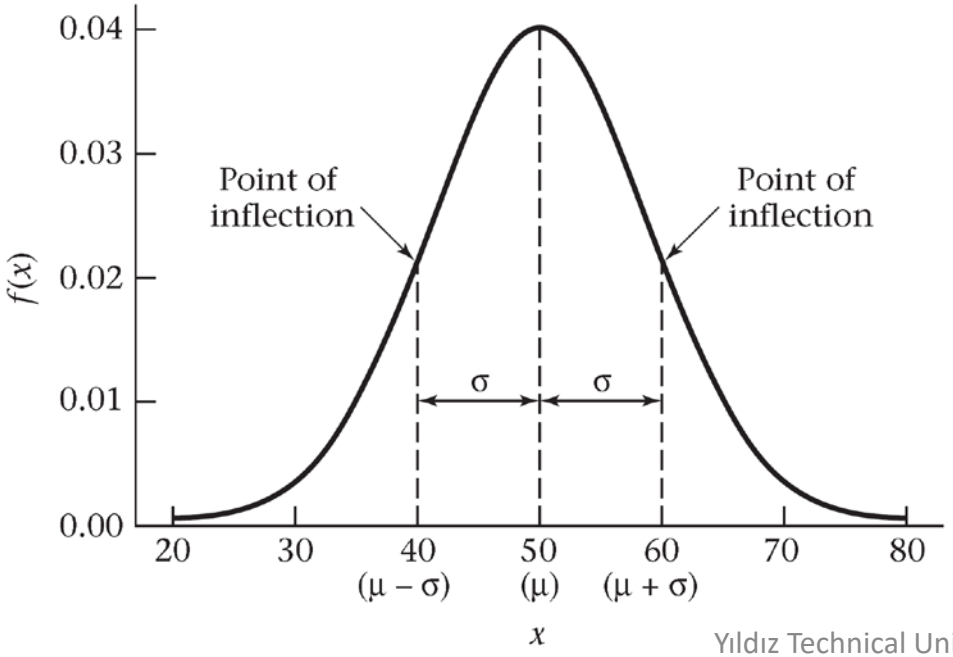
Another important area of application of *the normal distribution is as an approximating distribution to other distributions*. The normal distribution is generally more convenient to work with than any other distribution, particularly in hypothesis testing. Thus, if an accurate normal approximation to some other distribution can be found, we often will want to use it.

The Normal Distribution

The **normal distribution** is defined by its pdf, which is given as

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \quad -\infty < x < \infty$$

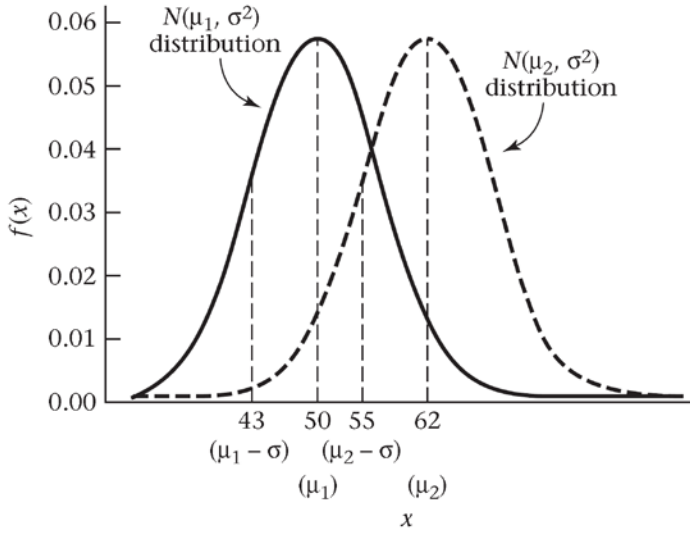
for some parameters μ, σ, where $\sigma > 0$.

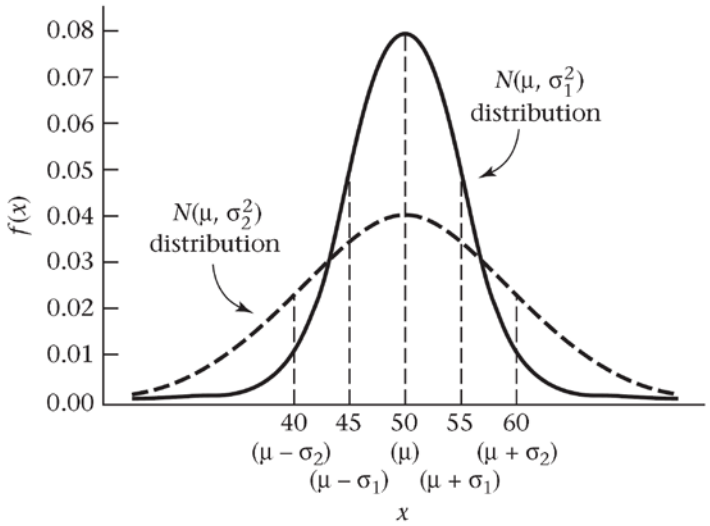


The pdf for a normal distribution with mean μ (50) and variance σ^2(100)

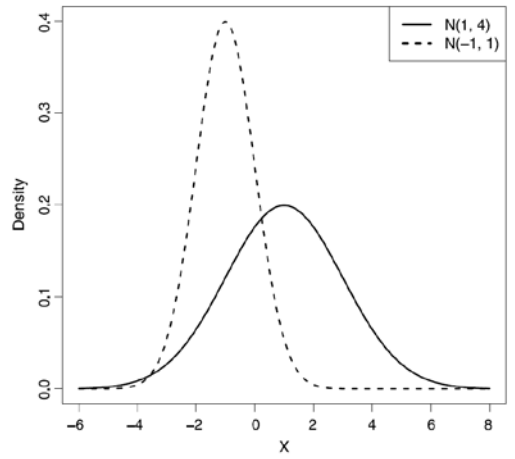
The Normal Distribution

A normal distribution with mean μ and variance σ^2 will generally be referred to as an $N(\mu,\sigma^2)$ distribution.





The entire shape of the normal distribution is determined by the two parameters **μ** and **σ^2**.



Examples of density curves for the normal distribution.

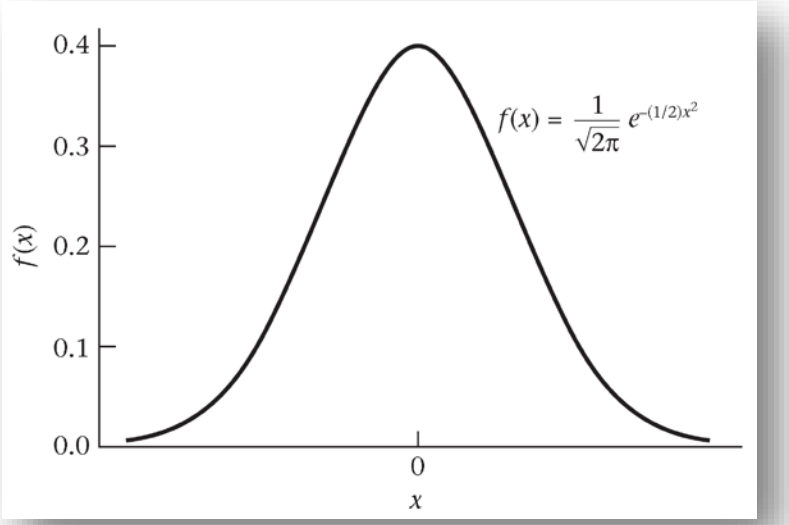
- ✓ The distribution shown by the solid curve has a mean of 1 and variance of 4.
- ✓ The distribution shown by the dashed curve has a mean of −1 and variance of 1.

The Standard Normal Distribution

A normal distribution with mean 0 and variance 1 is called a **standard**, or **unit**, normal distribution. This distribution is also called an $N(0,1)$ distribution.

Properties of the Standard Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{(-1/2)x^2}, \quad -\infty < x < +\infty$$



This distribution is symmetric about 0, because $f(x) = f(-x)$, as shown in Figure.

The Standard Normal Distribution

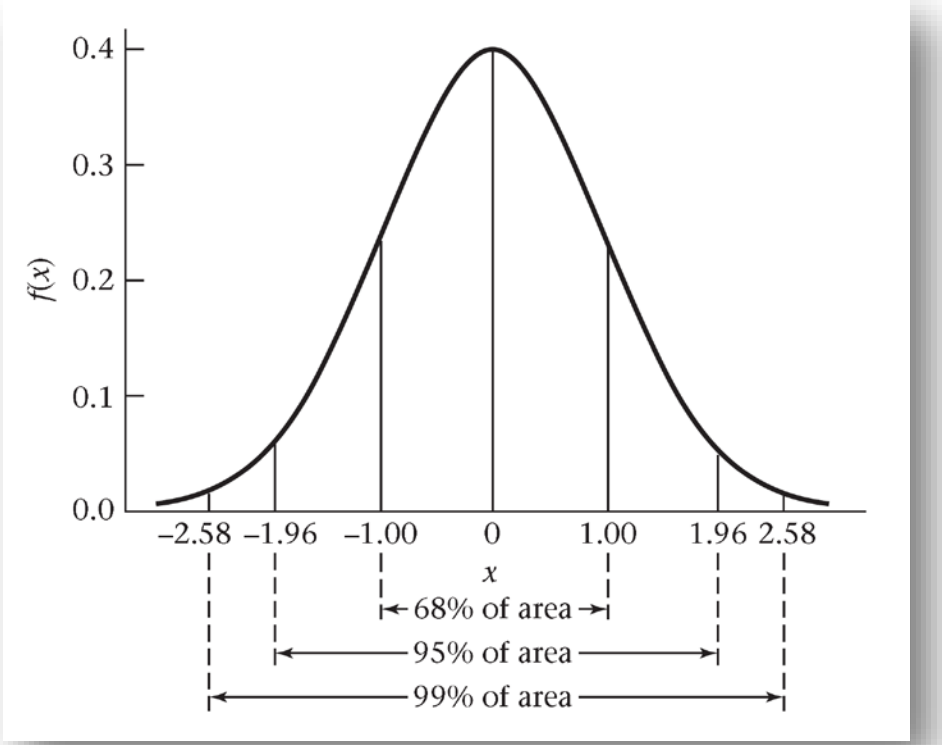
It can be shown that about 68% of the area under the standard normal density lies between +1 and –1, about 95% of the area lies between +2 and –2, and about 99% lies between +2.5 and –2.5.

These relationships can be expressed more precisely by saying that;

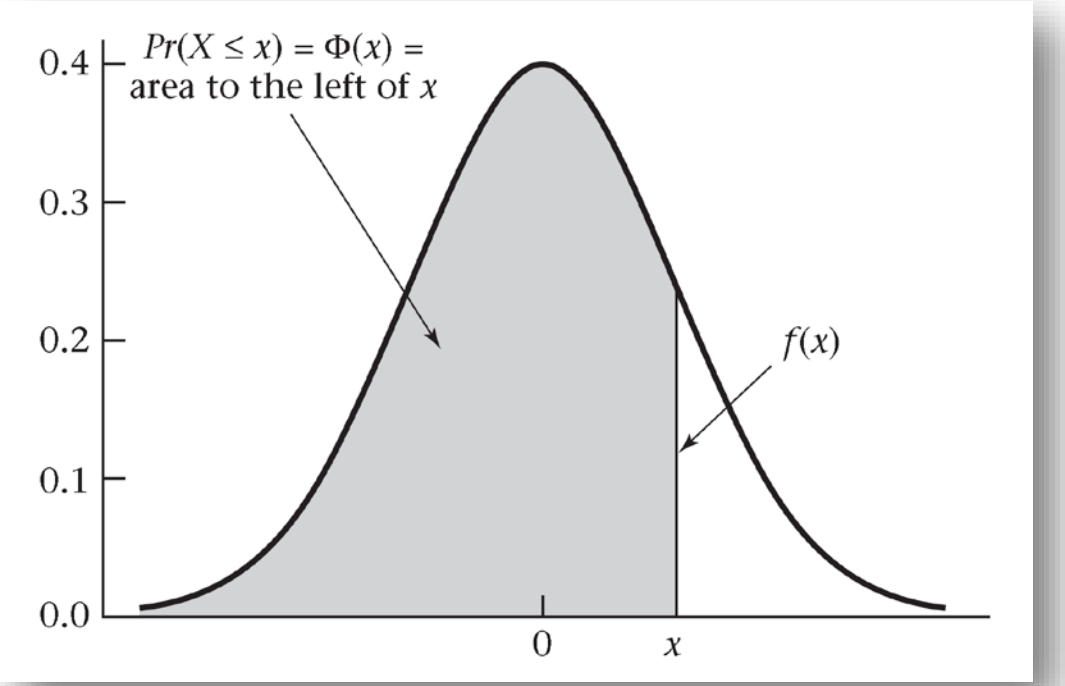
$$\Pr(-1 < X < 1) = 0.6827$$

$$\Pr(-1.96 < X < 1.96) = 0.95$$

$$\Pr(-2.576 < X < 2.576) = 0.99$$



The Standard Normal Distribution



The cumulative-distribution function (cdf) for a standard normal distribution is denoted by

$\Phi(x) = \Pr(X \le x)$

where X follows an $N(0,1)$ distribution.

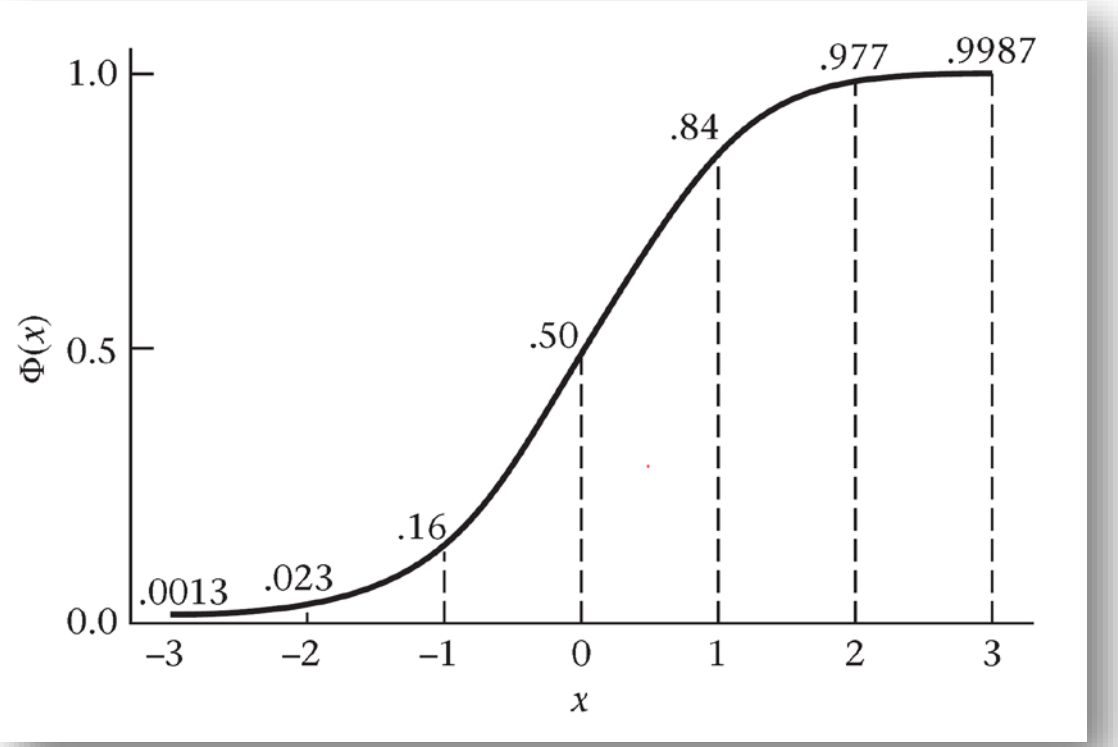
The symbol ~ is used as shorthand for the phrase "is distributed as." Thus, $X \sim N(0,1)$ means that the random variable X is distributed as an $N(0,1)$ distribution.

The Standard Normal Distribution

Symmetry Properties of the Standard Normal Distribution

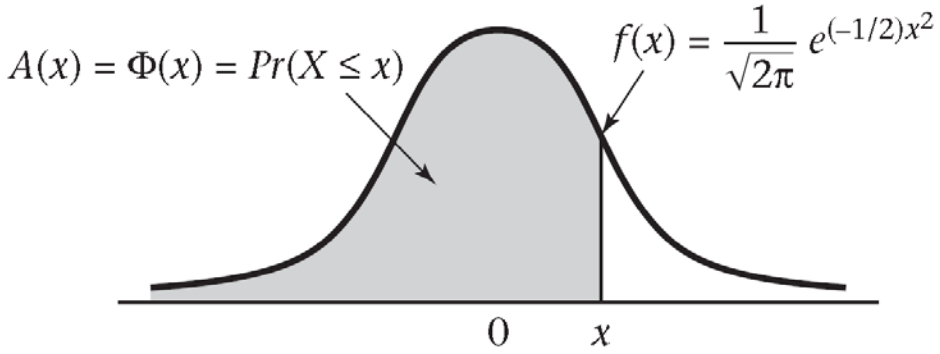
From the symmetry properties of the standard normal distribution,

$$\Phi(-x) = Pr(X \le -x) = Pr(X \ge x) = 1 - Pr(X \le x) = 1 - \Phi(x)$$

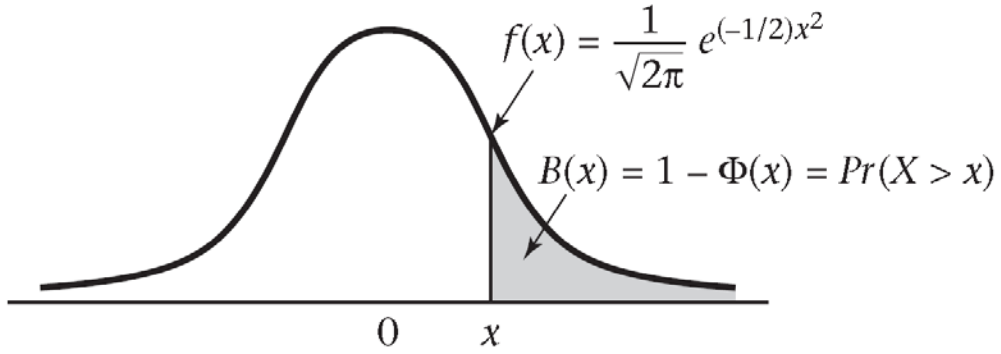


The CDF for a standard normal distribution $[\Phi(x)]$

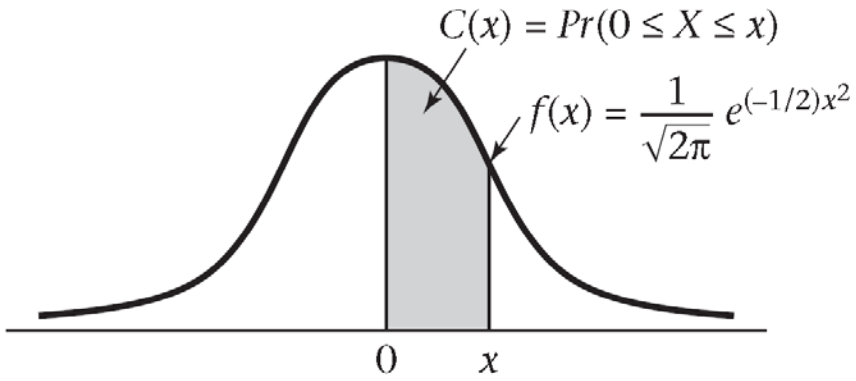
The Normal Distribution



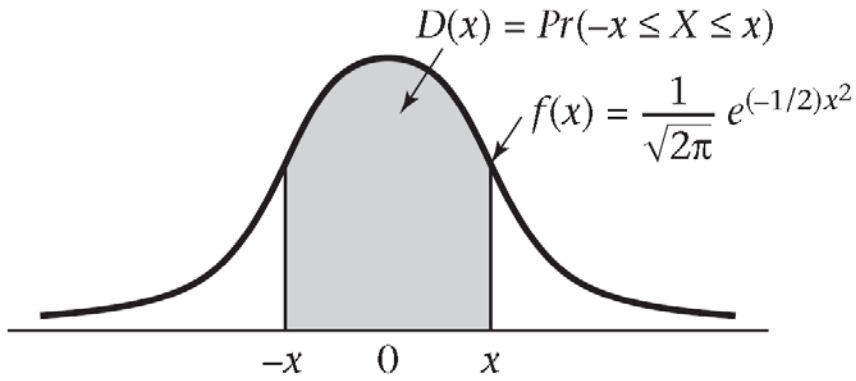
(a)



(b)



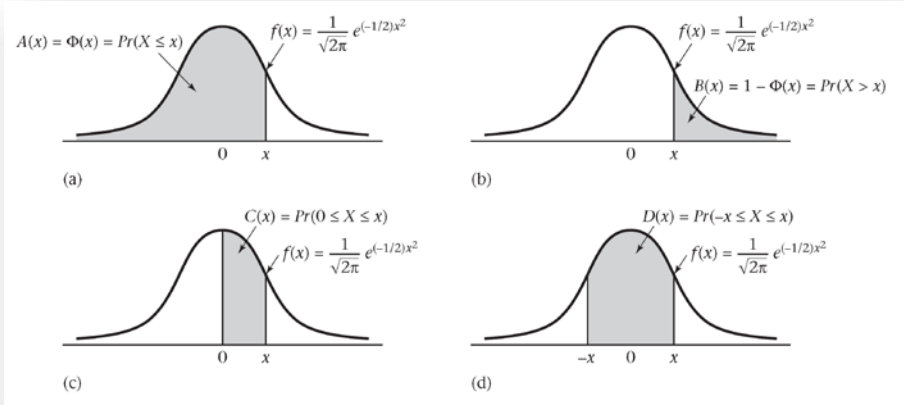
(c)



(d)

The Normal Distribution

x	A[a]	B[b]	C[c]	D[d]
0.0	.5000	.5000	.0	.0
0.01	.5040	.4960	.0040	.0080
0.37	.6443	.3557	.1443	.2886
0.38	.6480	.3520	.1480	.2961
0.76	.7764	.2236	.2764	.5527
0.77	.7793	.2207	.2793	.5587
1.58	.9429	.0571	.4429	.8859
1.59	.9441	.0559	.4441	.8882
2.39	.9916	.0084	.4916	.9832
2.40	.9918	.0082	.4918	.9836



Conversion from an $N(\mu, \sigma^2)$ to $N(1, 0)$

If $X \sim N(\mu,\sigma^2)$ and $Z = (X - \mu)/\sigma$, then $Z \sim N(0,1)$.

Evaluation of Probabilities for Any Normal Distribution via Standardization

If $X \sim N(\mu,\sigma^2)$ and $Z = (X - \mu)/\sigma$

then $$Pr(a < X < b) = Pr\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi[(b-\mu)/\sigma] - \Phi[(a-\mu)/\sigma]$$

The general principle is that for any probability expression concerning normal random variables of the form $Pr(a < X < b)$, the population mean μ is subtracted from each boundary point and divided by the standard deviation σ to obtain an equivalent probability expression for the standard normal random variable Z,

$$Pr[(a-\mu)/\sigma < Z < (b-\mu)/\sigma]$$

The standard normal tables are then used to evaluate this latter probability.

Evaluation of Probabilities for any Normal Distribution using Standardization

Example: Hypertensive

Suppose a mild hypertensive is defined as a person whose DBP is between 90 and 100 mm Hg inclusive, and the subjects are 35- to 44-year-old men whose blood pressures are normally distributed with mean 80 and variance 144. What is the probability that a randomly selected person from this population will be a mild hypertensive?

Table 1. Classifying Hypertension According to the JNC7 Guidelines[5,6]

Normal blood pressure	< 120 / < 80 mm Hg
Prehypertension	120 – 139 / 80 – 89 mm Hg
Stage 1 hypertension	140 – 159 / 90 – 99 mm Hg
Stage 2 hypertension	≥ 160 / ≥ 100 mm Hg

Abbreviation: JNC7, Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure.
Sources: Chobanian et al. *JAMA*. 2003[5]; National High Blood Pressure Education Program. 2004.[6]

Evaluation of Probabilities for any Normal Distribution using Standardization

Example: Hypertensive

Suppose a mild hypertensive is defined as a person whose DBP is between 90 and 100 mm Hg inclusive, and the subjects are 35- to 44-year-old men whose blood pressures are normally distributed with mean 80 and variance 144. What is the probability that a randomly selected person from this population will be a mild hypertensive?

This question can be stated more precisely: If $X \sim N(80,144)$, then what is $Pr(90 < X < 100)$?

$$Pr(90 < X < 100) = Pr\left(\frac{90-80}{12} < Z < \frac{100-80}{12}\right)$$

$$= Pr(0.833 < Z < 1.667) = \Phi(1.667) - \Phi(0.833)$$

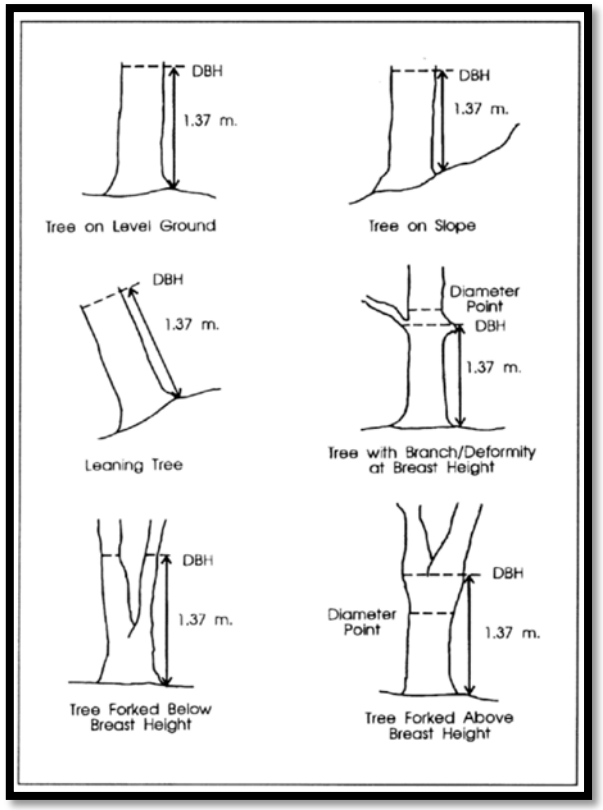
$$= .9522 - .7977 = .155$$

Thus, about 15.5% of this population will have mild hypertension.

Evaluation of Probabilities for any Normal Distribution using Standardization

Example: Botany

Suppose tree diameters of a certain species of tree from some defined forest area are assumed to be normally distributed with mean = 8 in. and standard deviation = 2 in. Find the probability of a tree having an unusually large diameter, which is defined as >12 in.



Evaluation of Probabilities for any Normal Distribution using Standardization

Example: Botany

Suppose tree diameters of a certain species of tree from some defined forest area are assumed to be normally distributed with mean = 8 in. and standard deviation = 2 in. Find the probability of a tree having an unusually large diameter, which is defined as >12 in.

We have $X \sim N(8,4)$ and require

$$Pr(X > 12) = 1 - Pr(X < 12) = 1 - Pr\left(Z < \frac{12-8}{2}\right)$$

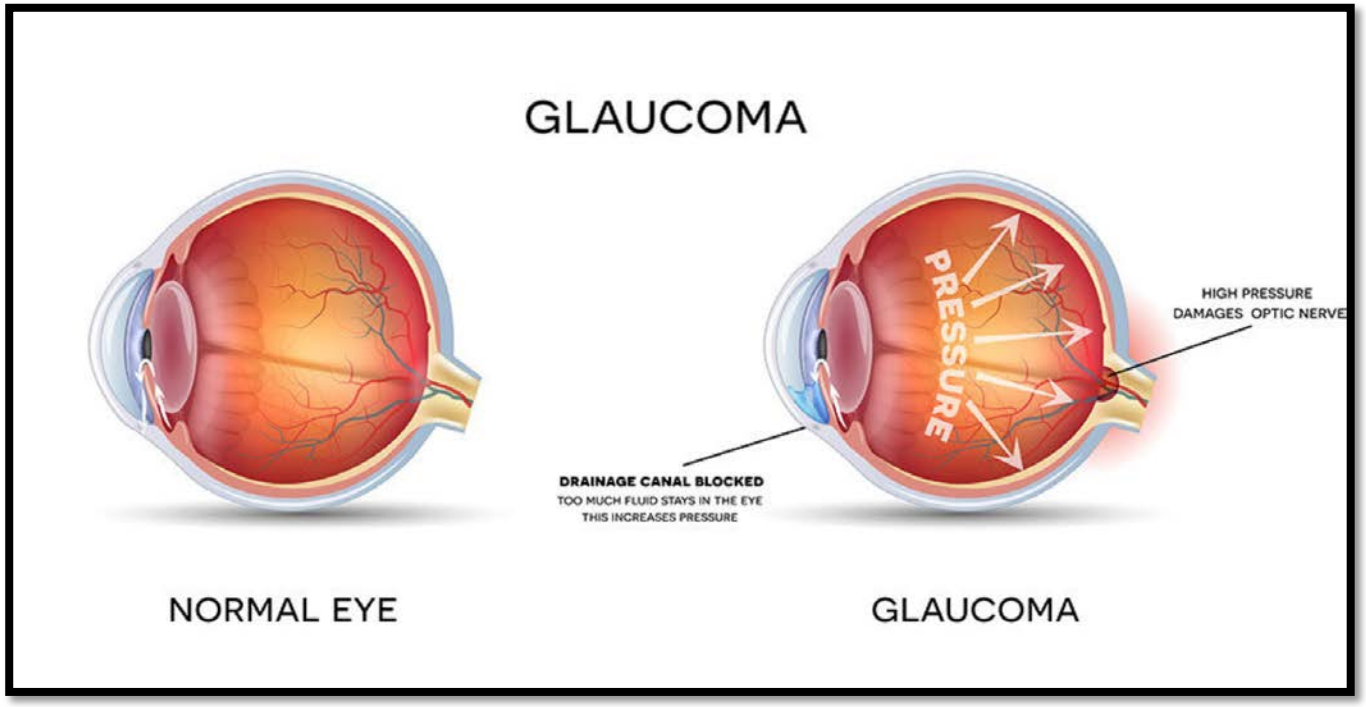
$$= 1 - Pr(Z < 2.0) = 1 - .977 = .023$$

Thus, 2.3% of trees from this area have an unusually large diameter.

Evaluation of Probabilities for any Normal Distribution using Standardization

Example: Ophthalmology

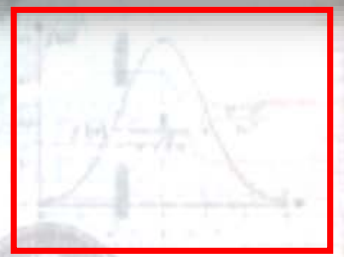
Glaucoma is an eye disease that is manifested by high intraocular pressure (IOP). The distribution of IOP in the general population is approximately normal with mean = 16 mm Hg and standard deviation = 3 mm Hg. If the normal range for IOP is considered to be between 12 and 20 mm Hg, then what percentage of the general population would fall within this range?



Questions?

Johann Carl Friedrich GAUSS
(1777-1855)

Mathematician of the millennium



Did you note the bell-shaped Gaussian curve on this 10DM note?

Bio highlights:

1777: Born in Brunswick
1798: Construction of a regular 17-gon by ruler and compass, first major advance in this field for 2000 years
1799: Dissertation on fundamental theorem of algebra
1801: Gains fame by correctly predicting the position of asteroid Ceres, after developing basics of pattern recognition
1809: Treatise on the motion of celestial bodies
Early 1800s: Non-Euclidean geometry (later publications by *Bolyai*). Discussion of statistical estimators. Geodesy / Heliotrope
1828: Main work on differential geometry; Gaussian curvature
1830s: Theory of magnetism
1855: Dies in Göttingen

BME 1132
Probability and Biostatistics

Examples

The Binomial Distribution

Suppose that we plan to recruit a group of 50 patients with breast cancer and study their survival within five years from diagnosis.
We represent the survival status for these patient by a set of Bernoulli random variables $X_1, \dots, X_{50}$.
(For each patient, the outcome is either 0 or 1.)
Assuming that all patients have the same survival probability, $p = 0.8$, and the survival status of one patient does not affect the probability of survival for another patient, $X_1, \dots, X_{50}$ form a set of 50 Bernoulli trials.
Now we can create a new random variable Y representing the number of patients out of 50 who survive for five years. The number of survivals is the number of 1s in the set of Bernoulli trials. This is the same as the sum of Bernoulli trials, whose values are either 0 or 1:

$$Y = \sum_{i}^{n} X_i$$

$$where\ X_i = 1\ if\ the\ \boldsymbol{i}th\ patient\ survive\ and X_i = 0\ otherwise.$$

Since Y can be any integer number from 0 (no one survives) through 50 (everyone survives), its range is $\{0,1, \dots, 50\}$. The range is a countable set. Therefore, the random variable Y is discrete. The distribution of Y is a **binomial distribution**,

$$Y \sim \boldsymbol{Binomial}(50,0.8)$$

The Binomial Distribution

Suppose that we plan to recruit a group of 50 patients with breast cancer and study their survival within five years from diagnosis.

$$Y = \sum_{i}^{n} X_i$$

$$where\ X_i = 1\ if\ the\ \boldsymbol{i}th\ patient\ survive\ and X_i = 0\ otherwise.$$

$$Y \sim \boldsymbol{Binomial}(50,0.8)$$

a-) Plot PMF of Y.

b-) Find the probability that either 34 or 35 or 36 patients survive.

c-) Calculate the probability that the number of survivals (out of 50) is less than or equal to 36.

The Poisson Distribution

The number of survivals in a group of n = 50 cancer patients in previous example cannot exceed 50.
Now, suppose that we are investigating the number of family doctor visits for each person in one year.
Although very large numbers such as 100 are quite unlikely, there is no theoretical and prespecified upper limit to this random variable.
Theoretically, its range is the set of all nonnegative integers.

$$X \sim \boldsymbol{Poisson}(\lambda)$$

$$Parameter\ \lambda\ is\ interpreted\ as\ the\ rate\ of\ occurence\ within\ a\ time\ period\ or\ space\ limit.\ (\lambda > 0)$$

Assume that the rate of physician visits per year is 2.5: $X \sim Poisson\ (2.5)$.

a-) Find the population mean and variance.

b-) Plot PMF of X.

c-) $P(X = 1) =?$