Probability and Probability Distributions

Probability and Probability Distributions

- Usually we want to do more with data than just describing them!

- We might want to test certain specific inferences about the behavior of the data.

Example

One theory concerning the etiology of breast cancer states that:

- Women aged (over 30) who give birth to their first child are at **greater risk** for eventually developing **breast cancer** than are women who give birth to their first child early in life (before age 20)!
- To test this hypothesis, we might choose 2000 women who are currently ages 45–54 and have never had breast cancer, of whom 1000 had their first child before the age of 20 (call this group A) and 1000 after the age of 30 (group B).

- Follow them for 5 years to assess whether they developed breast cancer during this period
- Suppose there are 4 new cases of breast cancer in group A and 5 new cases in group B.
- Is this evidence enough to confirm a difference in risk between the two groups?!!
- Do we need to increase the sample size? Could this results be due to chance!!
- The problem is that we need a conceptual framework to make these decisions!
- This framework is provided by the underlying concept of probability.

Probability

- A **probability** is a measure of the likelihood that an **event** in the future will happen. It can only assume a value between 0 and 1.
- A value near zero means the event is not likely to happen. A value near one means it is likely.
- There are three ways of assigning probability:
  1. Classical.
  2. Empirical.
  3. Subjective.

Definitions

- An **experiment** is the observation of some activity or the act of taking some measurement.
- An **outcome** is the particular result of an experiment.
- An **event** is the collection of one or more outcomes of an experiment.

Classical Probability

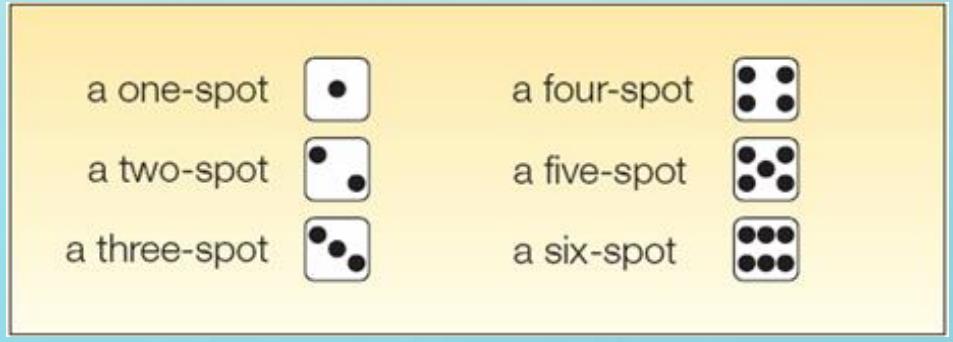
CLASSICAL PROBABILITY

$$\text{Probability of an event} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

Consider an experiment of rolling a **six-sided dice** .

What is the probability of the event “an even number of spots appear face up”?

The possible outcomes are:

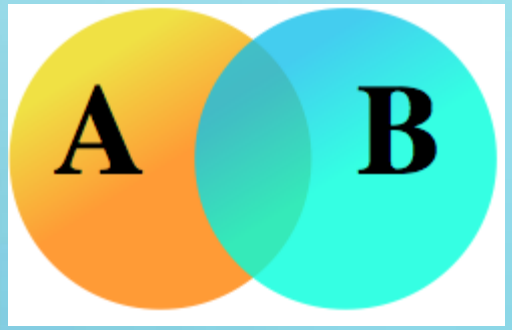


There are **three** “favorable” outcomes (a two, a four, and a six) in the collection of six equally likely possible outcomes.

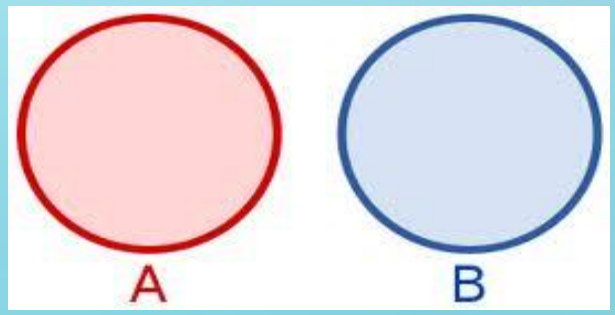
Mutually Exclusive Events

- Events are **mutually exclusive** if the occurrence of any one event means that none of the others can occur at the same time.
- Two events A and B are mutually exclusive if they cannot both happen at the same time.

A, *B NOT* mutually exclusive



A, *B* mutually exclusive



Independent Events

- Events are **independent** if the occurrence of one event does not affect the occurrence of another.

- Two events A and B are called independent events if:

$$Pr(A \cap B) = Pr(A) \cdot Pr(B)$$

Empirical Probability

- The probability of an event is the **relative frequency** of this set of outcomes over an Indefinitely large (or infinite) number of trials.

$$P(E) = f / n$$

- The empirical probability is an **estimate** or estimator of a probability.
- The Empirical probability is based on **observation**.

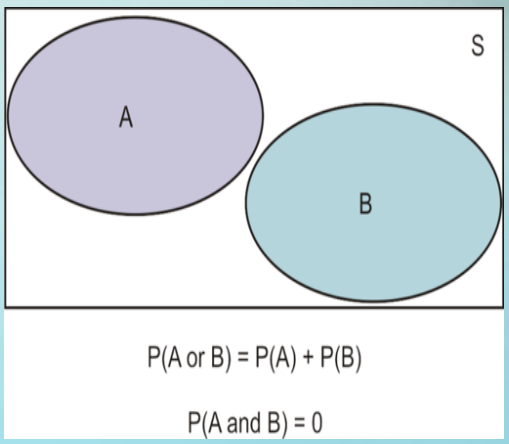
Rules of Probability

Rules of Addition

1. Special Rule of Addition

If two events *A* and *B* are **mutually exclusive**, the probability of one *or* the other event's occurring equals the sum of their probabilities.

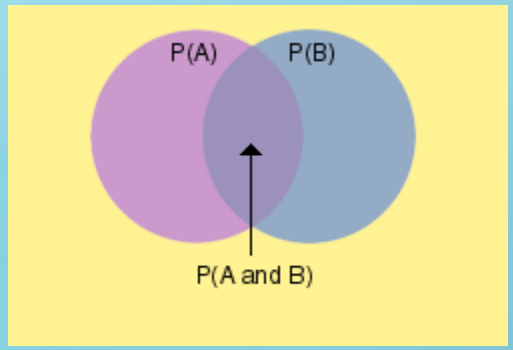
$$P(A \cup B) = P(A) + P(B)$$



2. The General Rule of Addition

If A and B are two events that are not mutually exclusive, then P(A or B) is given by the following formula:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Example

- Let A be the event that a person has normotensive diastolic blood pressure (DBP) readings (DBP $<$ 90), and let B be the event that a person has borderline DBP readings (90 $\leq$ DBP $<$ 95).
- Suppose that $Pr(A) = .7$, and $Pr(B) = .1$.
- Let Z be the event that a person has a DBP $<$ 95. Then

$$Pr(Z) = Pr(A) + Pr(B) = .8$$

- because the events A and B cannot occur at the same time.
- Let X be DBP, C be the event $X \geq 90$, and D be the event $75 \leq X \leq 100$. Events C and D are ***not*** mutually exclusive, because they both occur when $90 \leq X \leq 100$.

Rule of Multiplication

1. Special Rule of Multiplication

- The special rule of multiplication requires that two events *A* and *B* are ***independent***.
- If A and B are independent events, then

$$P(A \cap B) = P(A) \times P(B)$$

2. General rule of multiplication

- Use the general rule of multiplication to find the joint probability of two events when the events are not independent.
- If A and B are NOT independent events, then

$$P(A \cap B) = P(A) \times P(B|A)$$

Conditional Probability

- A **conditional probability** is the probability of a particular event occurring, given that another event has occurred.
- The probability of the event A given that the event B has occurred is written as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Example

Smoking	Lung cancer				Total	
	positive		negative			
	No.	%	No.	%	No.	%
Smoker	15	65.2	8	34.8	23	100
Non smoker	5	13.5	32	86.5	37	100
Total	20	33.3	40	66.7	60	100

If we chose one person at random, what is the probability that he is:

1. A smoker?
2. Has no cancer?
3. Smoker and has cancer?
4. Has cancer given that he is smoker?
5. Has no cancer or non smoker?

Bayes' Rule and Screening Tests

- The predictive value **positive** (*PV+*) of a screening test is the probability that a person has a disease given that the test is positive.

Pr (disease|test+) =

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid \overline{D})P(\overline{D})}$$

Where,

$$P(\overline{D}) = 1 - P(D)$$

$$p(T \mid \overline{D}) = 1 - P(\overline{T} \mid \overline{D})$$

- The predictive value **negative** (*PV*–) of a screening test is the probability that a person does *not* have a disease given that the test is negative.

Pr (no disease | test–)=

$$P(\overline{D} \mid \overline{T}) = \frac{P(\overline{T} \mid \overline{D})P(\overline{D})}{P(\overline{T} \mid D)P(D) + P(\overline{T} \mid \overline{D})P(\overline{D})}$$

Where,

$$p(\overline{T} \mid D) = 1 - P(T \mid D)$$

- The **sensitivity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is present given that the person has a disease.
- The **specificity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is *not* present given that the person does *not* have a disease.
- A **false negative** is defined as a negative test result when the disease or condition being tested for is actually present.
- A **false positive** is defined as a positive test result when the disease or condition being tested for is not actually present.

Prevalence and Incidence

- In clinical medicine, the terms *prevalence* and *incidence* denote probabilities in a special context and are used frequently in this text.
- The prevalence of a disease is the probability of currently having the disease regardless of the duration of time one has had the disease.
- Prevalence is obtained by dividing the number of people who currently have the disease by the number of people in the study population.

- The **cumulative incidence** of a disease is the probability that a person with no prior disease will develop a new case of the disease over some specified time period.
- Incidence should not be confused with prevalence, which is a measure of the total number of cases of disease in a population rather than the rate of occurrence of new cases.

Example

- A medical research team wished to evaluate a proposed screening test for Alzheimer's disease.
- The test was given to a random sample of 450 patients with Alzheimer's disease and an independent random sample of 500 patients without symptoms of the disease.
- The two samples were drawn from populations of subjects who were 65 years or older.
- The results are as follows:

Test Result	Yes (D)	No ($\overline{D}$)	Total
Positive(T)	436	5	441
Negative ($\overline{T}$)	14	495	509
Total	450	500	950

- Compute the **sensitivity** of the symptom

$$P(T \mid D) = \frac{436}{450} = 0.9689$$

- Compute the **specificity** of the symptom

$$P(\overline{T} \mid \overline{D}) = \frac{495}{500} = 0.99$$

- Suppose it is known that the rate of the disease in the general population is **11.3%**. What is the predictive value positive of the symptom and the predictive value negative of the symptom
- The **predictive value positive** of the symptom is calculated as

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid \overline{D})P(\overline{D})}$$

$$= \frac{(0.9689)(0\ .113)}{(0.9689)(0\ .113) + (.01)(1 - 0.113)} = 0.925$$

- The **predictive value negative** of the symptom is calculated as

$$P(\overline{D}\mid\overline{T}) = \frac{P(\overline{T}\mid\overline{D})P(\overline{D})}{P(\overline{T}\mid\overline{D})P(\overline{D}) + P(\overline{T}\mid D)P(D)}$$

$$= \frac{(0.99)(0.887)}{(0.99)(0.887) + (0.0311)(0.113)} = 0.996$$

Random Variables

- A random variable is a function that assigns numeric values to different events in a sample space.
- When the values of a variable (height, weight, or age) can't be predicted in advance, the variable is called a random variable.

RANDOM VARIABLE A quantity resulting from an experiment that, by chance, can assume different values.

DISCRETE RANDOM VARIABLE
A random variable that can assume only certain clearly separated values. It is usually the result of counting something.

CONTINUOUS RANDOM VARIABLE can assume an infinite number of values within a given range. It is usually the result of some type of measurement

EXAMPLES:

1. The number of students in a class.
2. The number of children in a family.
3. The number of cigarettes smoked per day.
4. number of motor-vehicle fatalities in a city during a week

EXAMPLES:

1. Diastolic blood-pressure (DBP) of group of people.
2. The weight of each student in this class.
3. The temperature of a patient.
4. Age of patients.

What is a Probability Distribution?

PROBABILITY DISTRIBUTION:
A listing of all the outcomes of an experiment and the probability associated with each outcome.

CHARACTERISTICS OF A PROBABILITY DISTRIBUTION

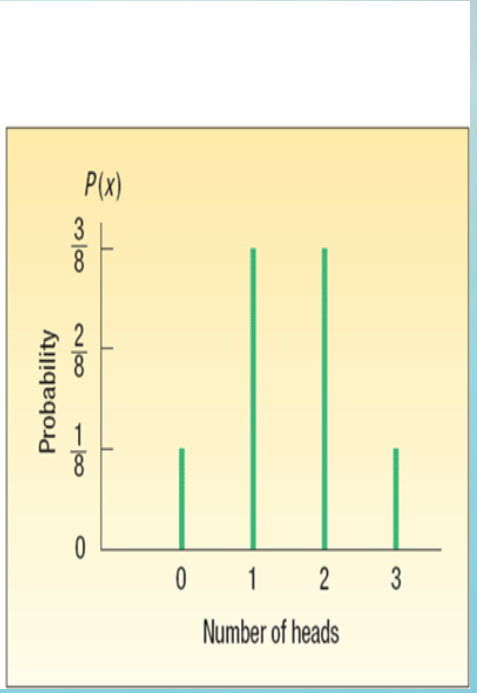
1. The probability of a particular outcome is between 0 and 1 inclusive.
2. The outcomes are mutually exclusive events.
3. The list is exhaustive. So the sum of the probabilities of the various events is equal to 1.

Probability Distributions for Discrete Random Variables

Experiment:
Toss a coin three times. Observe the number of heads. What is the probability distribution for the number of heads**?**

Possible Result	Coin Toss			Number of Heads
	First	Second	Third	
1	T	T	T	0
2	T	T	H	1
3	T	H	T	1
4	T	H	H	2
5	H	T	T	1
6	H	T	H	2
7	H	H	T	2
8	H	H	H	3

Number of Heads, x	Probability of Outcome, $P(x)$
0	$\frac{1}{8} = .125$
1	$\frac{3}{8} = .375$
2	$\frac{3}{8} = .375$
3	$\frac{1}{8} = .125$
Total	$\frac{8}{8} = 1.000$



Example

- Many new drugs have been introduced in the past several decades to bring **hypertension** under control —that is, to reduce high blood pressure to normotensive levels.
- Suppose a physician agrees to use a new antihypertensive drug on atrial basis on the first four untreated hypertensives she encounters in her practice, before deciding whether to adopt the drug for routine use.
- Let X = the number of patients of four who are brought under control. Then X is a **discrete random variable**, which takes on the values 0, 1, 2, 3, 4.

- Suppose from previous experience with the drug, the drug company expects that for any clinical practice the probability that 0 patients of 4 will be brought under control is .008, 1 patient of 4 is .076, 2 patients of 4 is .265, 3 patients of 4 is .411, and all 4 patients is .240.

Probability-mass function for the hypertension-control example

Pr(X = r)	.008	.076	.265	.411	.240
r	0	1	2	3	4

The Mean and Variance of a Discrete Probability Distribution

- If a random variable has a large number of values with positive probability, then the probability-mass function is not a useful summary measure!
- The **expected value** (Mean) of a discrete random variable is defined as:

$$E(X) = \mu = \sum_{i=1}^{R} x_i \Pr(X = x_i)$$

where the X_i's are the values the random variable assumes with positive probability

Example

Find the expected value for the random variable hypertension.

Solution:

$$E(X) = 0(.008) + 1(.076) + 2(.265) + 3(.411) + 4(.240) = 2.80$$

- Thus on average about 2.8 hypertensives would be expected to be brought under control for every 4 who are treated.

- The **Variance** of a Discrete Random Variable:

The variance of a discrete random variable, denoted by *Var* (*X*), is defined by:

$$\text{Var}(X) = \sigma^2 = \sum_{i=1}^{R} (x_i - \mu)^2 P_r(X = x_i)$$

The Cumulative-Distribution Function of a Discrete Random Variable

- Many random variables are displayed in tables or figures in terms of a **cumulative distribution function** rather than a distribution of probabilities of individual values.
- The basic idea is to assign to each individual value the sum of probabilities of all values that are no larger than the value being considered.
- The cumulative-distribution function (cdf) of a random variable X is denoted by $F(X)$ and, for a specific value x of X, is defined by $Pr(X \leq x)$ and denoted by $F(x)$.

Pr(X = r)	.008	.076	.265	.411	.240
r	0	1	2	3	4
F(x)	.008	.084	.349	.76	1.00

Permutations and Combinations

- The number of **permutations** of n things taken k at a time is

$$_nP_k = n(n-1) \times \ldots \times (n - k + 1) = \frac{n!}{(n-k)!}$$

- Where $n!$ = n factorial is defined as $n(n-1) \times \ldots \times 2 \times 1$
- It represents the number of ways of selecting k items out of n, where the order of selection is important.

- The number of **combinations** of n things taken k at a time is:

$$_{n}C_{k} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- It represents the number of ways of selecting k objects out of n where the order of selection does not matter.

The Binomial Distribution

- All examples involving the binomial distribution have a common structure:
- A sample of n independent trials, each of which can have only two possible outcomes, denoted as "success" and "failure."
- The probability of a success at each trial is assumed to be some constant p, and hence the probability of a failure at each trial is $1 - p = q$.
- The term "success" is used in a general way, without any specific contextual meaning.

- The distribution of the number of successes in n statistically independent trials, where the probability of success on each trial is p, is known as the binomial distribution and has a probability-mass function given by:

$$\Pr(X=k) = \binom{n}{k} p^k q^{n-k} \quad , \quad k = 0,1,\ldots,n$$

Example

What is the probability of obtaining 2 boys out of 5 children if the probability of a boy is 0.51 at each birth and the sexes of successive children are considered independent random variables?

$$\Pr(X=2) = \binom{5}{2} (.51)^2 (.49)^3 = .306$$

Expected Value and Variance of the Binomial Distribution

- The expected value and the variance of a binomial distribution are *np* and *npq*, respectively. That is:

$$E(x) = \mu = np$$

$$V(x) = \sigma^2 = npq$$

Practice Problem

- You are performing a cohort study. If the probability of developing disease in the exposed group is .05 for the study duration, then if you (randomly) sample **500** exposed people.
  1. How many do you expect to develop the disease? Give a margin of error (+/- 1 standard deviation) for your estimate.
  2. What's the probability that **at most** 10 exposed people develop the disease?

1. $X \sim \text{binomial}(500, .05)$

$E(X) = 500\,(.05) = 25, \quad Var(X) = 500\,(.05)\,(.95) = 23.75$

$\sigma_x = \sqrt{23.75} = 4.87$

25 ± 4.87

2. $P(X \le 10) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + \ldots + P(X=10) =$

$$\binom{500}{0}(.05)^0(.95)^{500} + \binom{500}{1}(.05)^1(.95)^{499} + \binom{500}{2}(.05)^2(.95)^{498} + \ldots + \binom{500}{10}(.05)^{10}(.95)^{490} < .01$$

The Poisson Distribution

- The Poisson distribution is perhaps the second most frequently used discrete distribution after the binomial distribution. This distribution is usually associated with rare events.
- The probability of *k* events occurring in a time period *t* for a Poisson random variable with parameter λ is given by:

$$P_r(X=k) = \frac{e^{-\mu}\ \mu^k}{k!} \qquad k = 0,1,\ldots$$

where $\mu = \lambda t$ and e is approximately 2.71828.

- Thus the Poisson distribution depends on a single parameter $\mu = \lambda t$.
- Note that the parameter λ represents the *expected number of events per unit time*, whereas the parameter μ represents the *expected number of events over time period t.*
- One important **difference** between the **Poisson** and **binomial** distributions concerns the numbers of trials and events.
  - For a binomial distribution there are a **finite** number of trials *n*, and the number of events can be no larger than *n.*
  - For a Poisson distribution the number of trials is essentially **infinite** and the number of events (or number of deaths) can be indefinitely large, although the probability of *k* events becomes very small as *k* increases.

Expected Value and Variance of the Poisson Distribution

For a Poisson distribution with parameter **μ**, the mean and variance are both equal to **μ**.

- This fact is useful, because if we have a data set from a discrete distribution where the *mean and variance are about the same*

Poisson Approximation to the Binomial Distribution

The binomial distribution with large n and small p can be accurately approximated by a Poisson distribution with parameter $\mu = np$.

Example:

If you have X ~ binomial (500, .05), then $\mu = E(X) = 500(.05) = 25$

$$P(X \leq 10) = \frac{e^{-25}\ 25^{0}}{0!} + \frac{e^{-25}\ 25^{1}}{1!} + \ldots + \frac{e^{-25}\ 25^{10}}{10!}$$

Example

- Assume that in Riyadh region an average of 13 new cases of lung cancer are diagnosed each year. If the annual incidence of lung cancer follows a Poisson distribution, find the probability of that in a given year the number of newly diagnosed cases of lung cancer will be:
  a) Exactly 10.
  b) At lest 8.
  c) No more than 12.
  d) Between 9 and 15.
  e) Fewer than 7.

Solution:

a) $\Pr(X=10) = \frac{e^{-13}\ 13^{10}}{10!} = 0.859$

b) $\Pr(X \geq 8) = 1 - \Pr(X \leq 7) = 1 - 0.054 = 0.946$

c) $\Pr(X \leq 12) = 0.4631$

d) $\Pr(9 \leq X \leq 15) = 0.6639$

e) $\Pr(X < 7) = \Pr(X \leq 6) = 0.0259$

Continuous Probability Distributions

A reminder:

- **Discrete** random variables have a countable number of outcomes
  - Examples: Dead/alive, treatment/placebo, dice, counts, etc.

- **Continuous** random variables have an infinite continuum of possible values.
  - Examples: blood pressure, weight, the speed of a car, the real numbers from 1 to 6.

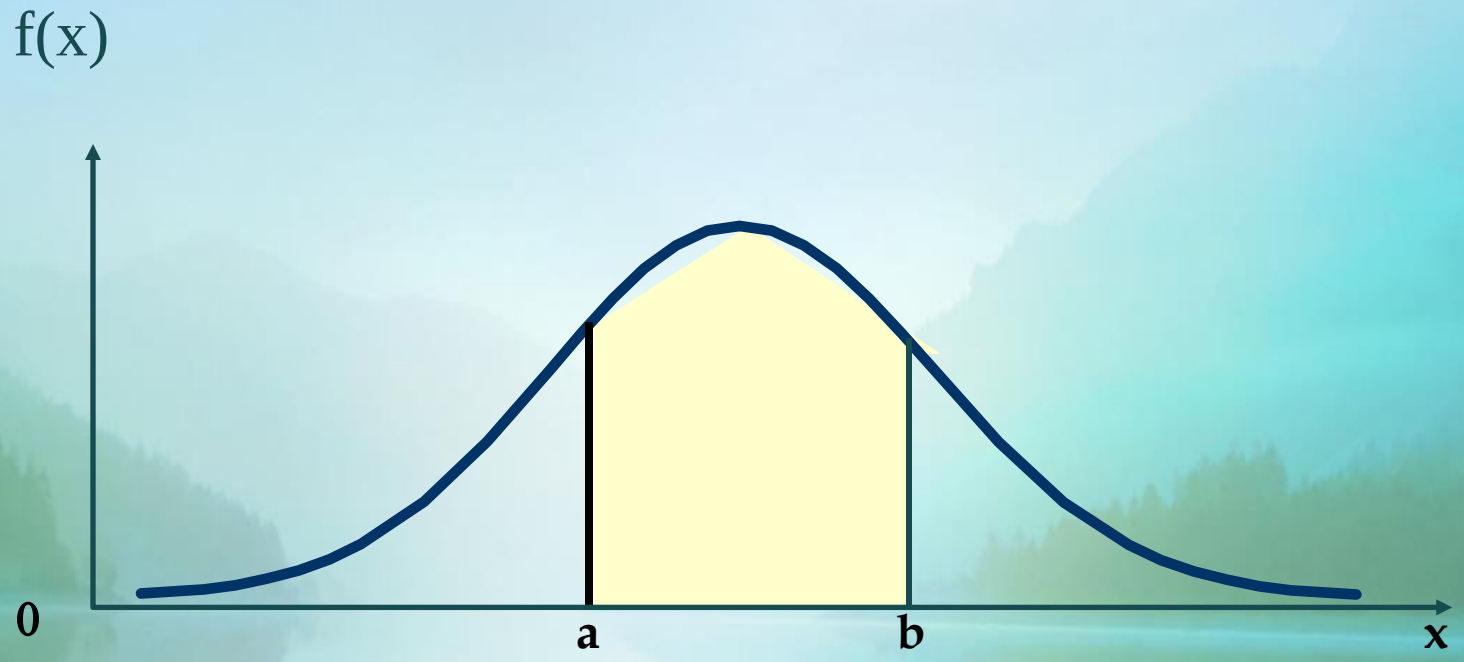
Properties of continuous probability Distributions:

- For continuous random variable, we use the following two concepts to describe the probability distribution:
  1. Probability Density Function (pdf): $f(x)$
  2. Cumulative Distribution Function (CDF): $F(x)=P(X \leq x)$

- Rules governing continuous distributions:
  a) Area under the curve = 1.
  b) $P(X = a) = 0$, where a is a constant.
  c) Area between two points a *and* b = $P(a < x < b)$

Probability Density Function

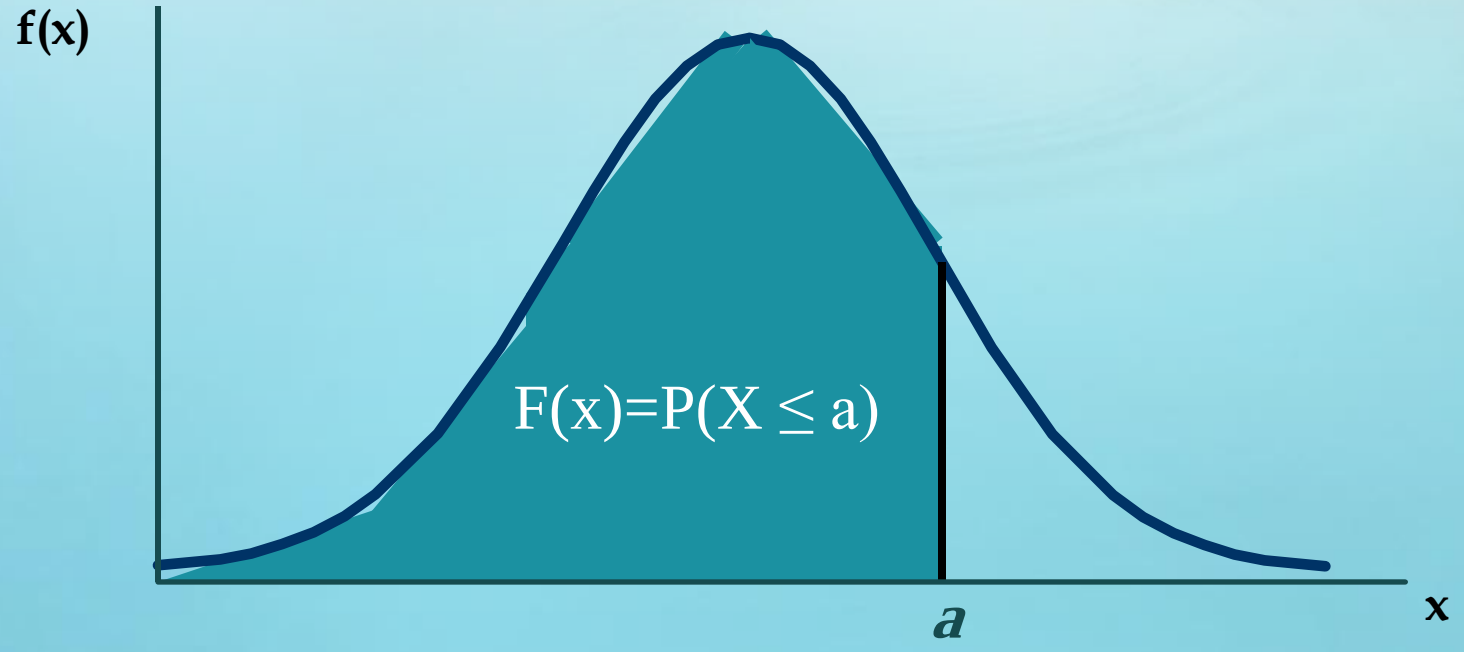
- The probability density function (pdf) of the random variable X is a function such that the area under the density-function curve between any two points a and b is equal to the probability that the random variable X falls between a and b. Thus, the total area under the density-function curve over the entire range of possible values for the random variable is 1.
- The pdf has large values in regions of high probability an small values in regions of low probability.



Probability Density Function shows the probability that the random variable falls in a particular range.

Cumulative Distribution Function

- 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 ≤ *a.* It is represented by the area under the pdf to the left of *a.*



Mean and Variance

- 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.

$$E(X) = \int_{\text{all } x} x_i \, f(x_i) dx$$

- The variance of a continuous random variable X, denoted by $Var(X)$ or σ^2, is the average square distance of each value of the random variable from its expected value, which is given by

$$Var(X) = \int_{\text{all } x} (x_i - \mu)^2 \, f(x_i) dx$$

The Normal Distribution

- The normal distribution is the most widely used continuous distribution. It is also frequently called the Gaussian distribution, after the well-known mathematician Karl Friedrich Gauss.
- The normal distribution is defined by its pdf, which is given as:

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

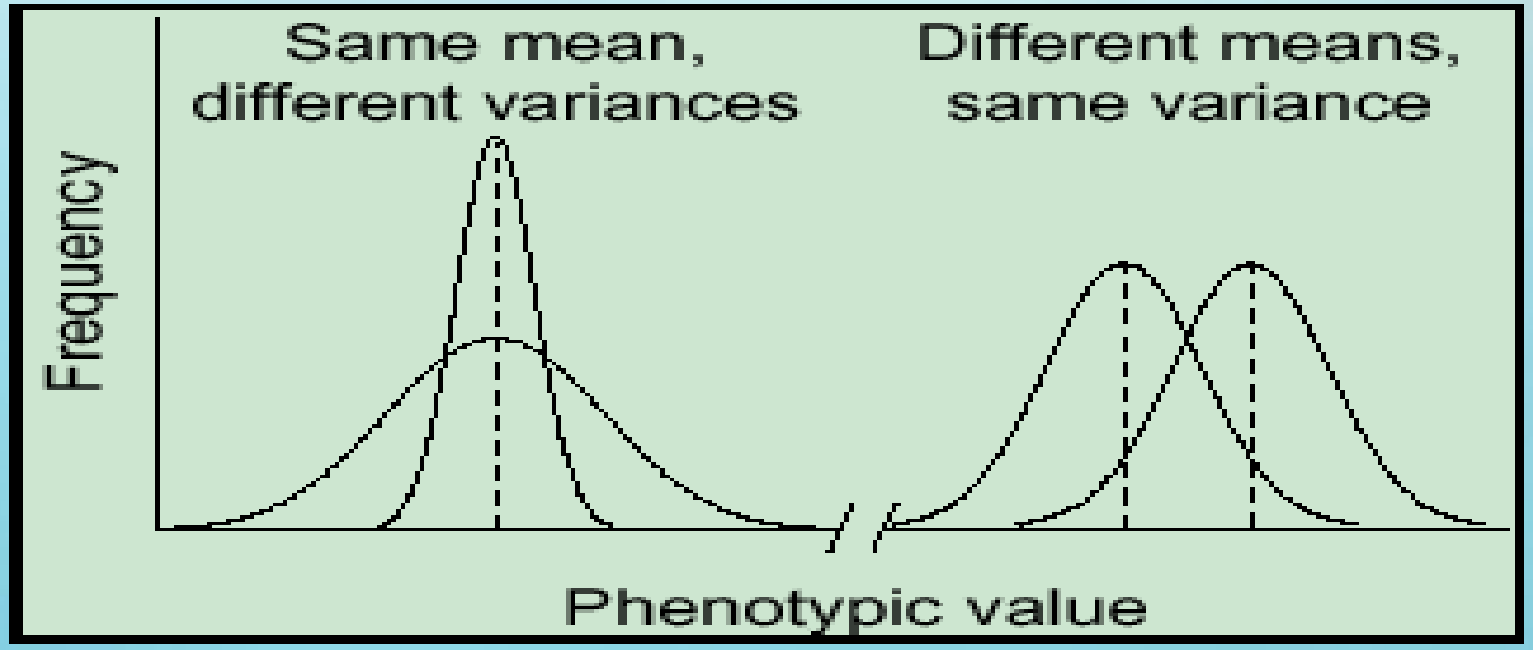
Where:

μ is the mean
σ is the standard deviation
$\pi = 3.1459$
e = 2.71828

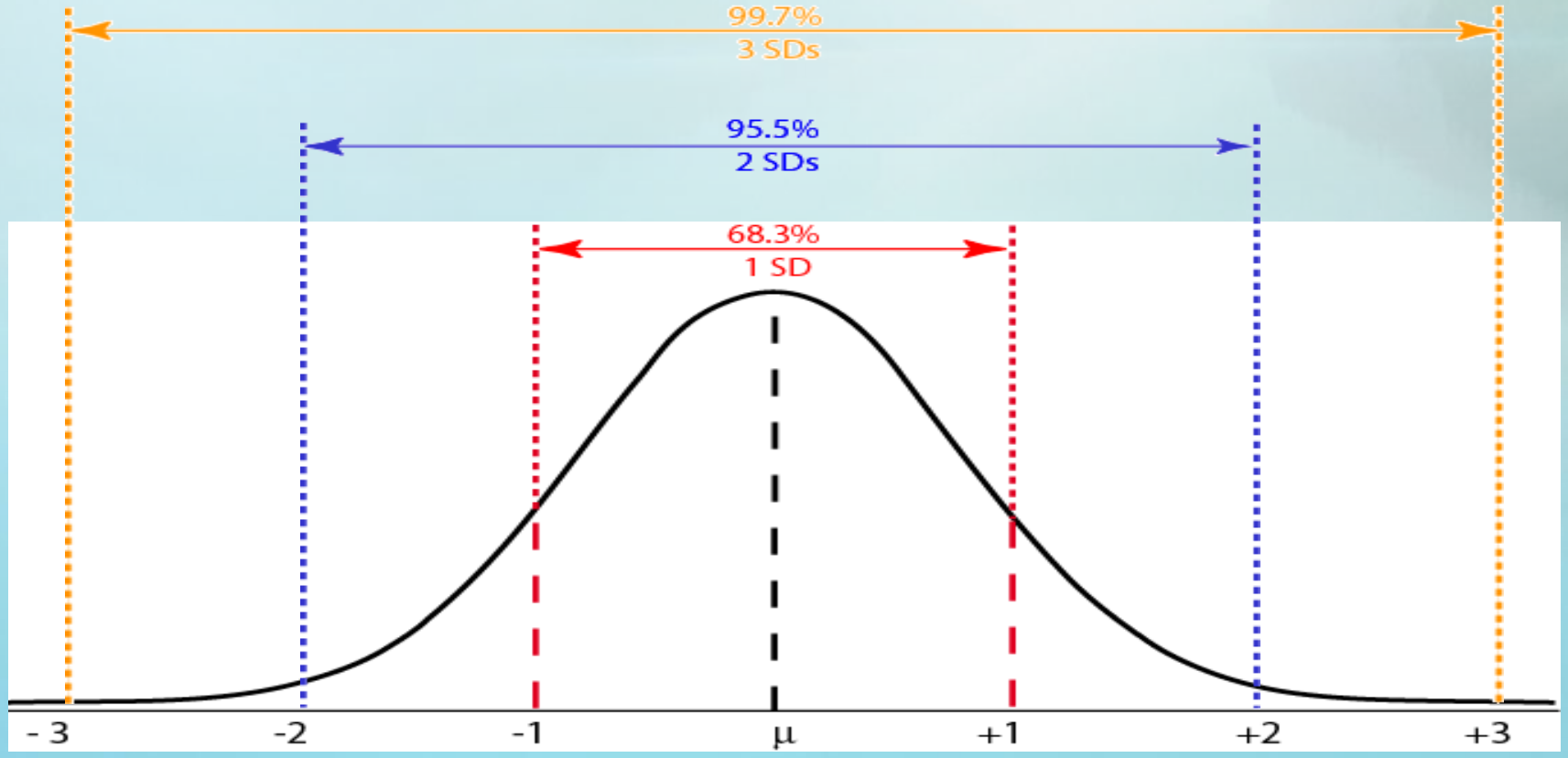
Characteristics of the normal distribution

- The distribution is **symmetric about the mean μ**, and is bell-shaped.
- The mean, the median, and the mode are all equal.
- The total area under the curve above the x-axis is one.
- The normal distribution is completely determined by the parameters **μ** and **σ**.

- The mean can be any numerical value: negative, zero, or positive and determines the location of the curve.
- The standard deviation determines the width of the curve: larger values result in wider, flatter curves.



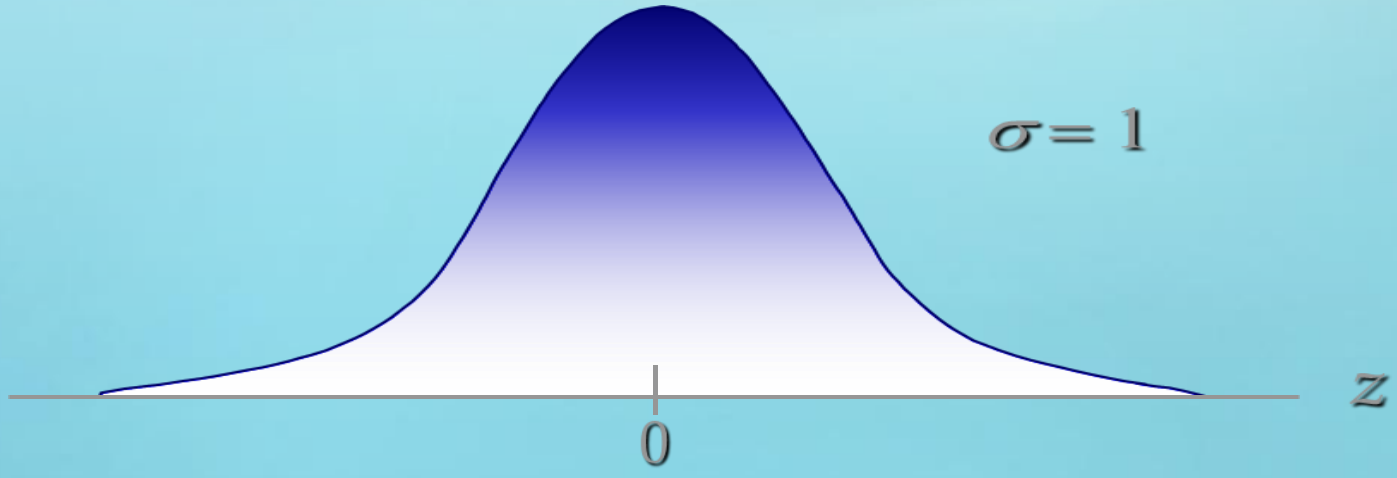
Empirical Rule



- 68% of data lie within 1σ of the mean μ
- 95% of data lie within 2σ of the mean μ
- 99.7% of data lie within 3σ of the mean μ

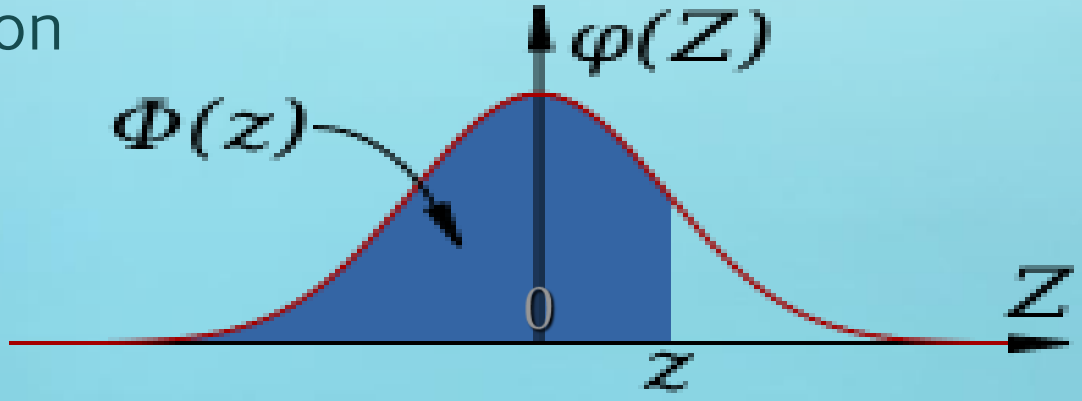
The Standard Normal Distribution

- 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.
- The letter **z** is used to designate the standard normal random variable.



Characteristics of the standard normal distribution

- It is has ZERO mean and One standard deviation and symmetrical about 0.
- The total area under the curve above the x-axis is one.
- We can use table (Z) to find the probabilities and areas.
- Probability that $Z \leq z$ is the area under the curve to the left of z.
- This is called the cdf $[\Phi(z)]$ for a standard normal distribution



How to transform normal distribution (X) to standard normal distribution (Z)

- This is done by the following formula:

$$Z = \frac{X - \mu_X}{\sigma_X}$$

which has a mean 0 and variance 1.

Example:
If X is normal with μ = 3, σ = 2. Find the value of standard normal Z, If X= 6?

Answer:

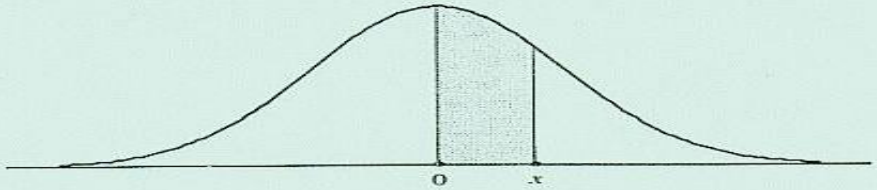
$$z = \frac{x-\mu}{\sigma} = \frac{6-3}{2} = 1.5$$

Example

- 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:

$$\text{If } X \sim N(80,144), \text{ then what is } P_r(90 < X < 100)?$$

Answer in page 121 in the book



The table gives the probability that a standard Normal variable lies between 0 and x (which is equivalent t
shaded area on the figure).

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0754
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2258	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2518	0.2549
0.2	0.2580	0.2612	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.3	0.2881	0.2910	0.2939	0.2967	0.2996	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986

3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9
0.4987	0.4990	0.4993	0.4995	0.4997	0.4998	0.4998	0.4999	0.4999	0.5000

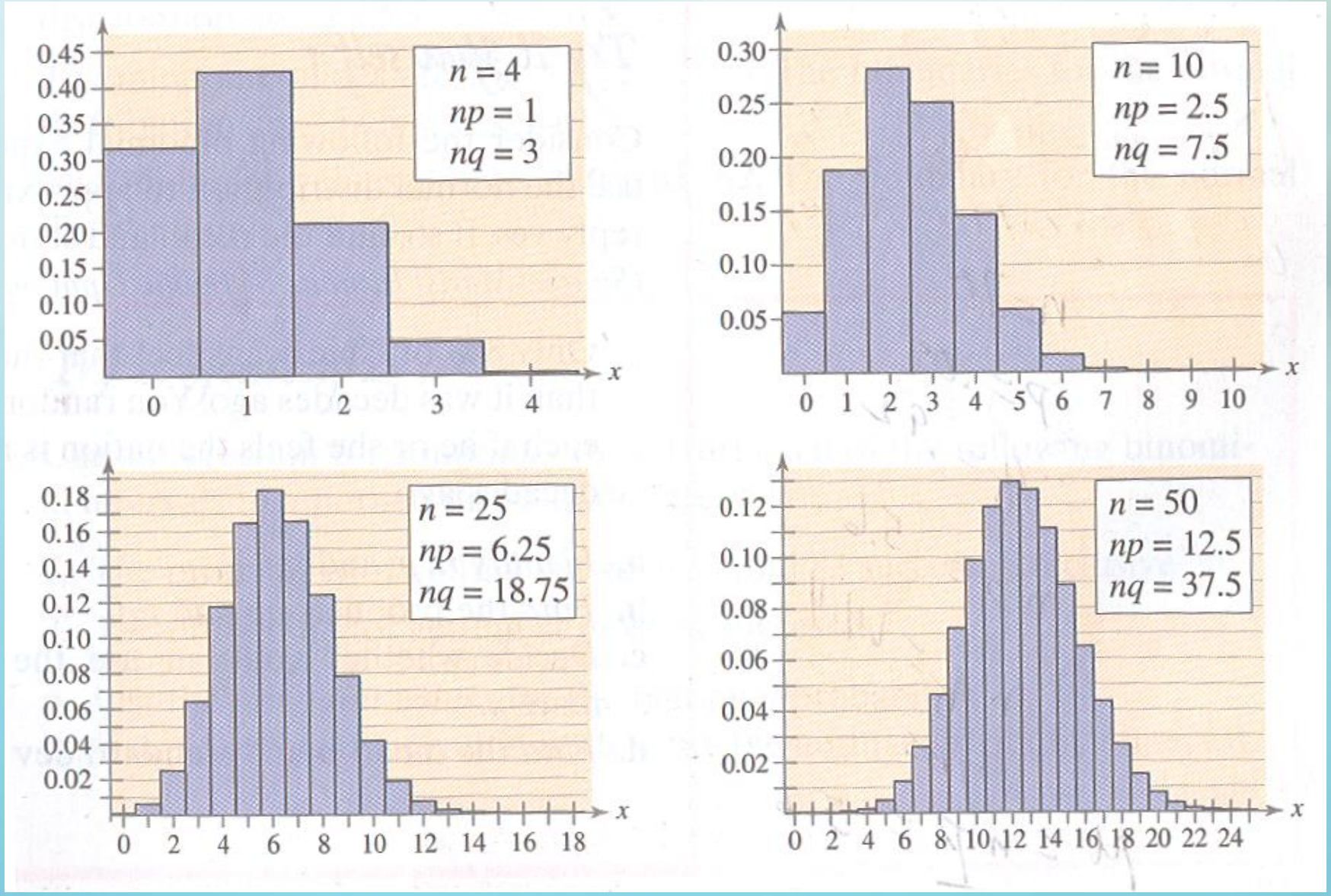
Normal Approximation to the Binomial Distribution

- you learned how to find binomial probabilities. For instance, a surgical procedure has an 85% chance of success and a doctor performs the procedure on 10 patients, it is easy to find the probability of exactly two successful surgeries.
- But what if the doctor performs the surgical procedure on 150 patients and you want to find the probability of fewer than 100 successful surgeries?
- To do this using the techniques described before, you would have to use the binomial formula 100 times and find the sum of the resulting probabilities.
- This is not practical and a better approach is to use a normal distribution to approximate the binomial distribution

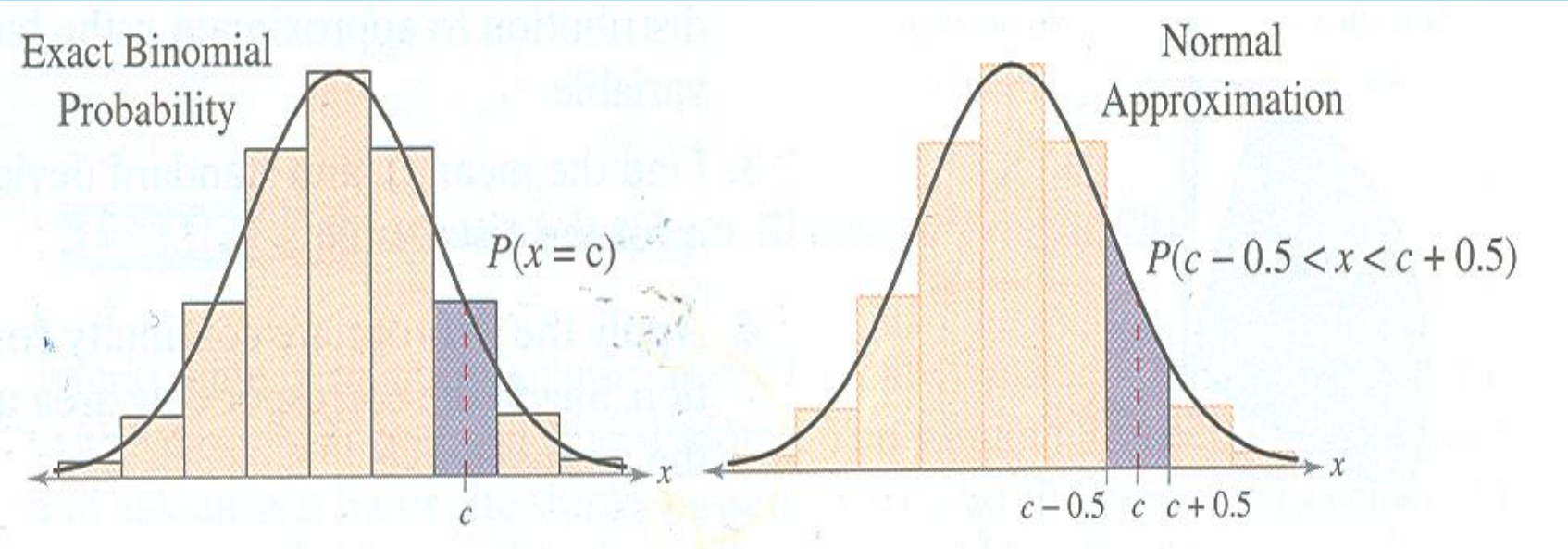
Normal Approximation to a Binomial Distribution

If $np \geq 5$ and $nq \geq 5$, then the binomial random variable x is approximately normally distributed, with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$.

To see why this result is valid, look at the following slide and binomial distributions for p = 0.25 and n = 4, 10, 25 and 50. Notice that as n increases, the histogram approaches a normal curve.


n = 4
np = 1
nq = 3
n = 10
np = 2.5
nq = 7.5
n = 25
np = 6.25
nq = 18.75
n = 50
np = 12.5
nq = 37.5
x

- When you use a continuous normal distribution to approximate a binomial probability, you need to move 0.5 units to the left and right of the midpoint to include all possible x-values in the interval.
- When you do this, you are making a **correction for continuity**.



Example

- Lets go to the previous example where we have

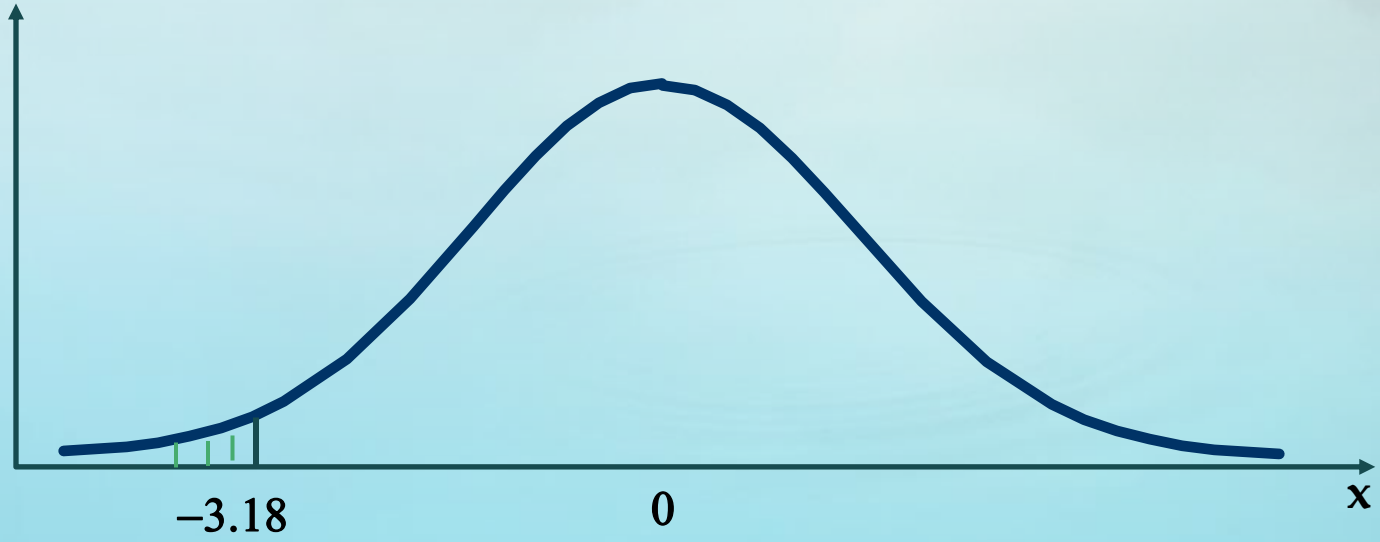
  $X \sim$ binomial (500, .05).

- Find the probability that $P(X \leq 10)$!
- Since np >5 and nq >5, then use normal approximation to do this.
- $E(X) = 500\ (.05) = 25, \quad Var(X) = 500\ (.05)\ (.95) = 23.75$

  $\sigma_x = \sqrt{23.75} = 4.87$

- $P(X \leq 10) = P\ (X < 9.5) = P\ (Z \leq \frac{9.5 - 25}{4.87}) = P\ (Z \leq -3.18)$?

- That is to find this area from Table 3, B.



$$P(Z \leq -3.18) = 0.0007$$