

Probability and Probability Distributions

Probability

- Chance of observing a particular outcome
- Likelihood of an event
- Assumes a “stochastic” or “random” process: i.e.. the outcome is not predetermined - there is an element of chance.
- An outcome is a specific result of a single trial of a probability experiment.

- Probability theory developed from the study of games of chance like dice and cards.
- A process like flipping a coin, rolling a die or drawing a card from a deck are **probability experiments**.

Why Probability in Statistics?

- Results are not certain
- To evaluate how accurate our results are:
 - Given how our data were collected, are our results accurate ?
 - Given the level of accuracy needed, how many observations need to be collected ?

When can we talk about probability ?

- When dealing with a process that has an uncertain outcome
- *Experiment* = any process with an uncertain outcome
- When an experiment is performed, one and only one outcome is obtained

- **Event** = something that may happen or not when the experiment is performed
- An event either occurs or it does not occur
- Events are represented by uppercase letters such as A , B , and C

- Probability of an Event E
= a number between 0 and 1 representing the proportion of times that event E is expected to happen when the experiment is done over and over again under the same conditions.
- Any event can be expressed as a subset of the set of all possible outcomes (S)

S = set of all possible outcomes
 $P(S) = 1$

Why Probability in Medicine?

- Because medicine is an inexact science, physicians seldom predict an outcome with absolute certainty.
- E.g., to formulate a diagnosis, a physician must rely on available diagnostic information about a patient
 - History and physical examination
 - Laboratory investigation, X-ray findings, ECG, etc

- Although no test result is absolutely accurate, it does affect the probability of the presence (or absence) of a disease.
 - Sensitivity and specificity
- An understanding of probability is fundamental for quantifying the uncertainty that is inherent in the decision-making process
- Probability theory is a foundation for statistical inference, &
- Allows us to draw conclusions about a population of patients based on information obtained from a sample of patients drawn from that population.

More importantly probability theory is used to understand:

- About probability distributions: Binomial, Poisson, and Normal Distributions
- Sampling and sampling distributions
- Estimation
- Hypothesis testing
- Advanced statistical analysis

Two Categories of Probability

- Objective and Subjective Probabilities.
- Objective probability
 - 1) Classical probability and
 - 2) Relative frequency probability.

Classical Probability

- Is based on gambling ideas
 - Rolling a die -
 - There are 6 possible outcomes:
 - Total ways = $\{1, 2, 3, 4, 5, 6\}$.
 - Each is equally likely
 - $P(i) = 1/6, i=1,2,\dots,6$.
 - $P(1) = 1/6$
 - $P(2) = 1/6$
 -
 - $P(6) = 1/6$
- SUM = 1

- **Definition:** If an event can occur in N mutually exclusive and equally likely ways, and if m of these posses a characteristic, E , the probability of the occurrence of $E = m/N$.

$P(E)$ = the probability of $E = P(E) = m/N$

- If we toss a die, what is the probability of 4 coming up?

$m = 1$ (which is 4) and $N = 6$

The probability of 4 coming up is $1/6$.

- **Another “equally likely” setting is the tossing of a coin -**

- **There are 2 possible outcomes in the set of all possible outcomes {H, T}.**

- **$P(H) = 0.5$**

- **$P(T) = 0.5$**

- $SUM = 1$**

Relative Frequency Probability

- In the long run process
- The proportion of times the event A occurs — in a large number of trials repeated under essentially identical conditions
- **Definition: If a process is repeated a large number of times (n), and if an event with the characteristic E occurs m times, the relative frequency of E,**

$$\text{Probability of E} = P(E) = m/n.$$

- If you toss a coin 100 times and head comes up 40 times,

$$P(H) = 40/100 = 0.4.$$

- If we toss a coin 10,000 times and the head comes up 5562,

$$P(H) = 0.5562.$$

- ***Therefore, the longer the series and the longer sample size, the closer the estimate to the true value.***

- Since trials cannot be repeated an infinite number of times, theoretical probabilities are often estimated by **empirical probabilities** based on a finite amount of data

- ***Example:***

Of 158 people who attended a dinner party, 99 were ill.

$$**P (Illness) = 99/158 = 0.63 = 63\%.**$$

- In 1998, there were 2,500,000 registered live births; of these, 200,000 were LBW infants.
- Therefore, the probability that a newborn is LBW is estimated by

$$P(\text{LBW}) = 200,000/2,500,000 \\ = 0.08$$

Subjective Probability

- Personalistic (represents one's degree of belief in the occurrence of an event).
- Personal assessment of which is more effective to provide cure – traditional/modern
- Personal assessment of which sports team will win a match.
- Also uses classical and relative frequency methods to assess the likelihood of an event.

- E.g., If someone says that he is 95% certain that a cure for AIDS will be discovered within 5 years, then he means that:

$$***P(\text{discovery of cure for AIDS within 5 years}) = 95\% = 0.95***$$

- ***Although the subjective view of probability has enjoyed increased attention over the years, it has not fully been accepted by scientists.***

Event Relations

Mutually Exclusive Events

- Two events A and B are *mutually exclusive* if they cannot both happen at the same time

$$P(A \cap B) = 0$$

- Example:
 - A coin toss cannot produce heads and tails simultaneously.
 - Weight of an individual can't be classified simultaneously as “underweight”, “normal”, “overweight”.

Event Relations...

Independent Events

- Two events A and B are *independent* if the probability of the first one happening is the same no matter how the second one turns out. OR. *The outcome of one event has no effect on the occurrence or non-occurrence of the other.*

$$P(A \cap B) = P(A) \times P(B) \text{ (Independent events)}$$

$$P(A \cap B) \neq P(A) \times P(B) \text{ (Dependent events)}$$

Example:

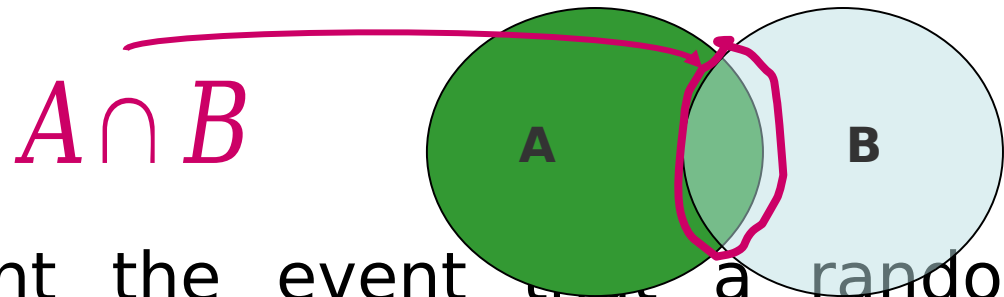
- The outcomes on the first and second coin tosses are independent

Event Relations...

Intersection, and union

- The intersection of two events A and B , $A \cap B$, is the event that A and B happen simultaneously

$$P (A \text{ and } B) = P (A \cap B)$$



- Let A represent the event that a randomly selected newborn is LBW, and B the event that he or she is from a multiple birth
- The intersection of A and B is the event that the infant is both LBW and from a multiple birth

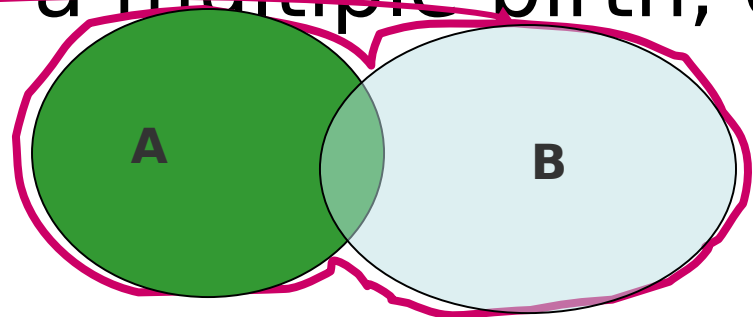
Event Relations...

- The **union** of A and B, $A \cup B$, is the event that either A happens or B happens or they both happen simultaneously

$$P (A \text{ or } B) = P (A \cup B)$$

- In the example above, the union of A and B is the event that the newborn is either LBW or from a multiple birth, or both.

$A \cup B$



Properties of Probability

1. The numerical value of a probability always lies between 0 and 1, inclusive.

$$0 \leq P(E) \leq 1$$

- ✓ *A value 0 means the event can not occur*
- ✓ *A value 1 means the event definitely will occur*
- ✓ *A value of 0.5 means that the probability that the event will occur is the same as the probability that it will not occur.*

2. The sum of the probabilities of all mutually exclusive outcomes is equal to 1.

$$P(E_1) + P(E_2) + \dots + P(E_n) = 1.$$

3. For two mutually exclusive events A and B,

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B).$$

If not mutually exclusive:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

4. The **complement** of an event A , denoted by $\bar{A}$ or A^c , is the event that A does not occur

- Consists of all the outcomes in which event **A** does **NOT** occur

$$P(\bar{A}) = P(\text{not } A) = 1 - P(A)$$

- $\bar{A}$ occurs only when A does not occur.
- These are complementary events.

- In the example, the complement of A is the event that a newborn is not LBW
- In other words, A is the event that the child weighs 2500 grams at birth

$$\mathbf{P(\bar{A}) = 1 - P(A)}$$

$$\begin{aligned} P(\text{not low bwt}) &= 1 - P(\text{low bwt}) \\ &= 1 - 0.076 \\ &= \mathbf{0.924} \end{aligned}$$

Basic Probability Rules

1. Addition rule

- If events **A** and **B** are mutually exclusive:

$$P(A \text{ or } B) = P(A) + P(B)$$

$$P(A \text{ and } B) = 0$$

- **More generally:**

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

P(event A or event B occurs or they both occur)

Example: The probabilities below represent years of schooling completed by mothers of newborn infants

Mother's education	Probability
≤ 8 years	0.056
9 to 11 years	0.159
12 years	0.321
13 to 15 years	0.218
≥ 16 years	0.230
Not reported	0.016

- What is the probability that a mother has completed < 12 years of schooling?

$$P(\leq 8 \text{ years}) = 0.056 \text{ and}$$

$$P(9-11 \text{ years}) = 0.159$$

- Since these two events are mutually exclusive,

$$\begin{aligned} P(\leq 8 \text{ or } 9-11) &= P(\leq 8 \cup 9-11) \\ &= P(\leq 8) + P(9-11) \\ &= 0.056 + 0.159 \\ &= 0.215 \end{aligned}$$

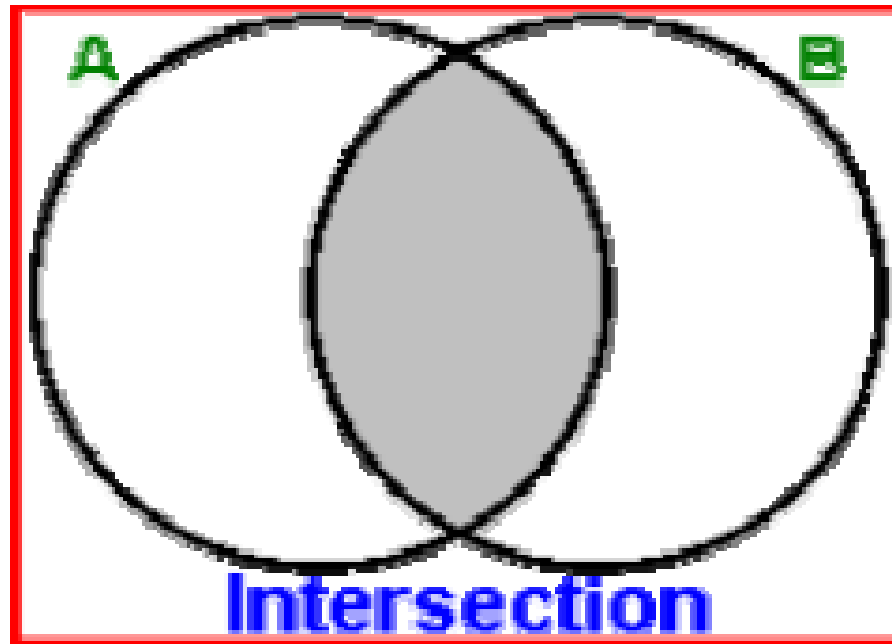
- What is the probability that a mother has completed 12 or more years of schooling?

$$\begin{aligned}P(\geq 12) &= P(12 \text{ or } 13-15 \text{ or } \geq 16) \\&= P(12 \cup 13-15 \cup \geq 16) \\&= P(12) + P(13-15) + P(\geq 16) \\&= 0.321 + 0.218 + 0.230 \\&= 0.769\end{aligned}$$

If A and B are not mutually exclusive events,

then subtract the overlapping:

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



- The following data are the results of electrocardiograms (ECGs) and radionuclide angiocardiograms (RAs) for 19 patients with post-traumatic myocardial effusions.
 - 7 patients developed both ECG and RA abnormality
 - 17 patients developed ECG abnormal
 - 9 patients developed RA abnormal

$$P(\text{ECG abnormal and RA abnormal}) = 7/19 = 0.37$$

$$P(\text{ECG abnormal or RA abnormal}) =$$

$$P(\text{ECG abnormal}) + P(\text{RA abnormal}) - P(\text{Both ECG and RA abnormal}) =$$

$$17/19 + 9/19 - 7/19 = 19/19 = 1.$$

Note: The problem is that the 7 patients whose ECGs and RAs are both abnormal are counted twice

2. Multiplication rule

- If A and B are independent events, then

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

- More generally,

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

$P(A \text{ and } B)$ denotes the probability that **A** and **B** both occur at the same time.

Conditional Probability

- Refers to the probability of an event, given that another event is known to have occurred.
- “What happened first is assumed”
- *Hint - When thinking about conditional probabilities, think in stages. Think of the two events A and B occurring chronologically, one after the other, either in time or space.*

- The *conditional probability* that event B has occurred given that event A has already occurred is denoted $P(B|A)$ and is defined

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

provided that $P(A) \neq 0$.

Example:

A study investigating the effect of prolonged exposure to bright light on retina damage in premature infants.

	Retinopathy YES	Retinopathy NO	TOTAL
Bright light	18	3	21
Reduced light	21	18	39
TOTAL	39	21	60

- The probability of developing retinopathy is:

$$\begin{aligned} P(\text{Retinopathy}) &= \frac{\text{No. of infants with retinopathy}}{\text{Total No. of infants}} \\ &= (18+21)/(21+39) \\ &= 0.65 \end{aligned}$$

- We want to compare the *probability of retinopathy*, given that the infant was *exposed to bright light*, with that the infant was *exposed to reduced light*.
- Exposure to bright light and exposure to reduced light are *conditioning events*, events we want to take into account when calculating ***conditional probabilities***.

- The conditional probability of retinopathy, given exposure to bright light, is:

- $P(\text{Retinopathy/exposure to bright light}) =$

$$\frac{\text{No. of infants with retinopathy exposed to bright light}}{\text{No. of infants exposed to bright light}}$$

$$= 18/21 = 0.86$$

- $P(\text{Retinopathy/exposure to reduced light})$
=

$$\frac{\text{\# of infants with retinopathy exposed to reduced light}}{\text{No. of infants exposed to reduced light}}$$

$$= 21/39 = 0.54$$

- The conditional probabilities suggest that premature infants exposed to bright light have a higher risk of retinopathy than premature infants exposed to reduced light.

- For independent events A and B
$$P(A/B) = P(A).$$
- For non-independent events A and B
$$P(A \text{ and } B) = P(A/B) P(B)$$

(General Multiplication Rule)

Test for Independence

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

$$P(B|A) = P(B)$$

or

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

- Two events A and B are *dependent* if

$$P(B|A) \neq P(B)$$

or

$$P(A \text{ and } B) \neq P(A) \cdot P(B)$$

Example

- In a study of optic-nerve degeneration in Alzheimer's disease, postmortem examinations were conducted on 10 Alzheimer's patients. The following table shows the distribution of these patients according to sex and evidence of optic-nerve degeneration.
- Are the events “patients has optic-nerve degeneration” and “patient is female” independent for this sample of 10 patients?

Optic-nerve Degeneration

Sex

Present

Not Present

Female

4

1

Male

4

1

Solution

• $P(\text{Optic-nerve degeneration}/\text{Female}) =$

$$\frac{\text{No. of females with optic-nerve degeneration}}{\text{No. of females}} = 4/5 = 0.80$$

$$\begin{aligned} P(\text{Optic-nerve degeneration}) &= \\ \frac{\text{No of patients with optic-nerve degeneration}}{\text{Total No. of patients}} &= 8/10 = 0.80 \end{aligned}$$

☞ The events are independent for this sample.

Exercise:

Culture and Gonodectin (GD) test results for 240 Urethral Discharge Specimens

GD Test Result	Culture Result		Total
	Gonorrhea	No Gonorrhea	
Positive	175	9	184
Negative	8	48	56
Total	183	57	240

1. What is the probability that a man has gonorrhea?
2. What is the probability that a man has a positive GD test?
3. What is the probability that a man has a positive GD test and gonorrhea?
4. What is the probability that a man has a negative GD test and does not have gonorrhea
5. What is the probability that a man with gonorrhea has a positive GD test?

6. What is the probability that a man who does not have gonorrhea has a negative GD test?
7. What is the probability that a man does not have gonorrhea has a positive GD test?
8. What is the probability that a man with positive GD test has gonorrhea?

Probability Distributions

- **A probability distribution** is a device used to describe the behaviour that a random variable may have by applying the theory of probability.
- It is the way data are distributed, in order to draw conclusions about a set of data
- **Random Variable** = Any quantity or characteristic that is able to assume a number of different values such that any particular outcome is determined by chance

- **Random variables** can be either discrete or continuous
- **A discrete random variable** is able to assume only a finite or countable number of outcomes
- **A continuous random variable** can take on any value in a specified interval

- With categorical variables, we obtain the *frequency distribution* of each variable
- With numeric variables, the aim is to determine whether or not *normality* may be assumed
 - *If not we may consider transforming the variable or categorize it for analysis (eg age group)*

Therefore, the probability distribution of a random variable is a table, graph, or mathematical formula that gives the probabilities with which the random variable takes different values or ranges of values.

Ways of arranging Objects

- **Factorials:** Given the positive integer n , the product of all the whole numbers from n down through 1 is called n factorial and is written $n!$.

$$\begin{aligned} n! &= n \times (n-1) \times (n-2) \times \dots \times 2 \times 1 \\ &= n \times (n-1)! \end{aligned}$$

By definition; $0! = 1$.

- **Permutation:** An ordered arrangement of objects.

$${}_nP_r = \frac{n!}{(n - r)!}$$

Eg. How many flags of two colours can be formed from a piece of cloth consisting of six different colours?

$$\begin{aligned} {}_6P_2 &= \frac{6!}{(6 - 2)!} \\ &= 30 \text{ different colours} \end{aligned}$$

- **Combinations:** An arrangement of objects without regard to order

$${}^nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

- Eg. How many different committees of 2 matches can be from a group of 6 people?

This is given by

$$\begin{aligned} {}^6C_2 &= \frac{6!}{(6-2)!2!} \\ &= 15 \text{ different committees} \end{aligned}$$

A. Discrete Probability Distributions

- For a discrete random variable, the probability distribution specifies each of the possible outcomes of the random variable along with the probability that each will occur
- Examples can be:
 - Frequency distribution
 - Relative frequency distribution
 - Cumulative frequency

- We represent a potential outcome of the random variable X by x

$$0 \leq P(X = x) \leq 1$$

$$\sum P(X = x) = 1$$

The following data shows the number of diagnostic services a patient receives

x	$P(X = x)$
0	0.671
1	0.229
2	0.053
3	0.031
4	0.010
5	0.006

- What is the probability that a patient receives exactly 3 diagnostic services?

$$P(X=3) = 0.031$$

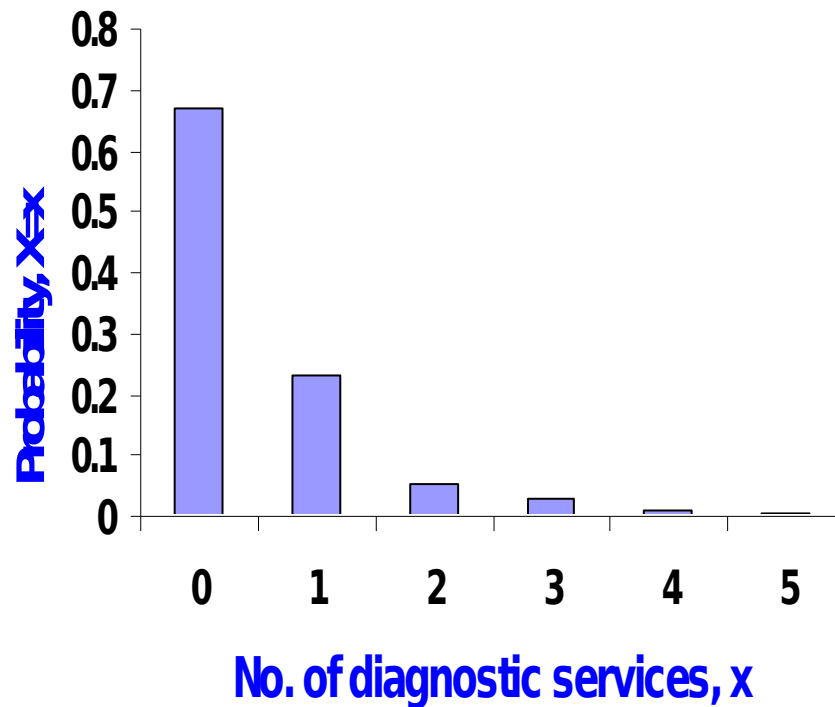
- What is the probability that a patient receives at most one diagnostic service?

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1) \\ &= 0.671 + 0.229 \\ &= 0.900 \end{aligned}$$

- What is the probability that a patient receives at least four diagnostic services?

$$\begin{aligned}P(X \geq 4) &= P(X = 4) + P(X = 5) \\&= 0.010 + 0.006 \\&= 0.016\end{aligned}$$

Probability distributions can also be displayed using a graph



The Expected Value of a Discrete Random variable

- If a random variable is able to take on a large number of values, then a probability mass function might not be the most useful way to summarize its behavior
- Instead, measures of location and dispersion can be calculated (as long as the data are not categorical)

- The average value assumed by a random variable is called its **expected value**, or the **population mean**
- It is represented by $E(X)$ or μ
- To obtain the expected value of a discrete random variable X , we multiply each possible outcome by its associated probability and sum all values with a probability greater than 0.
- **Mean (X) = $\sum (x_i) P(X=x_i)$**

- For the diagnostic service data:

$$\begin{aligned}\text{Mean } (X) &= 0(0.671) + 1(0.229) \\ &\quad + 2(0.053) \\ &\quad \quad + 3(0.031) + 4(0.010) \\ &\quad + 5(0.006) \\ &= 0.498 \approx 0.5\end{aligned}$$

- We would expect an average of 0.5 services for each visit

The Variance of a Discrete Random Variable

- The variance of a random variable X is called the **population variance** and is represented by $\text{Var}(X)$ or σ^2
- It quantifies the dispersion of the possible outcomes of X around the expected value μ .

$$\begin{aligned}
\sigma^2 &= \sum (x_i - \mu)^2 P(X = x_i) \\
&= (0 - 0.5)^2(0.671) + (1 - 0.5)^2(0.229) \\
&\quad + (2 - 0.5)^2(0.053) + (3 - 0.5)^2(0.031) \\
&\quad + (4 - 0.5)^2(0.010) + (5 - 0.5)^2(0.006) \\
&= 0.782
\end{aligned}$$

$$\text{Standard deviation} = \sigma = \sqrt{0.782} = 0.884$$

- Examples of discrete probability distributions are the binomial distribution and the Poisson distribution.

1. Binomial Distribution

- It is one of the most widely encountered discrete probability distributions.
- Consider dichotomous (binary) random variable
- Is based on Bernoulli trial
 - When a single trial of an experiment can result in only **one of two mutually exclusive outcomes** (success or failure; dead or alive; sick or well, male or female)

Example:

- We are interested in determining whether a newborn infant will survive until his/her 70th birthday
- Let Y represent the survival status of the child at age 70 years
- $Y = 1$ if the child survives and $Y = 0$ if he/she does not.

- The outcomes are mutually exclusive and exhaustive
- Suppose that 72% of infants born survive to age 70 years

$$P(Y = 1) = p = 0.72$$

$$P(Y = 0) = 1 - p = 0.28$$

The probability distribution of Y is

y	$P(Y = y)$
0	0.28
1	0.72

A **binomial probability distribution**

occurs when the following requirements are met.

1. The procedure has a fixed number of trials.
2. The trials must be independent.
3. Each trial must have all outcomes that fall into two categories.
4. The probabilities must remain constant for each trial [$P(\text{success}) = p$].

Characteristics of a Binomial Distribution

- The experiment consist of n identical trials.
- Only two possible outcomes on each trial.
- The probability of A (*success*), denoted by p , remains the same from trial to trial. The probability of B (*failure*), denoted by q ,

$$q = 1 - p.$$

- The trials are independent.
- n and π are the parameters of the binomial distribution.
- The mean is $n\pi$ and the variance is $n\pi(1 - \pi)$

- Suppose an event can have only binary outcomes A and B.
- Let the probability of A is π and that of B is $1 - \pi$.
- The probability π stays the same each time the event occurs.

- If an experiment is repeated n times and the outcome is independent from one trial to another, the probability that outcome A occurs exactly x times is:

- $P(X=x) = \binom{n}{x} p^x (1-p)^{(n-x)} \quad , x = 0, 1, 2, \dots, n.$

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

- **n** denotes the number of fixed trials
- **x** denotes the number of successes in the n trials
- **p** denotes the probability of success
- **q** denote probability of failure ($1 - p$)

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

- Represents the number of ways of selecting x objects out of n where the order of selection does not matter.
- where $n! = n(n-1)(n-2)\dots(1)$, and $0! = 1$

Example:

- Suppose we know that 40% of a certain population are cigarette smokers. If we take a random sample of 10 people from this population, what is the probability that we will have exactly 4 smokers in our sample?

- If the probability that any individual in the population is a smoker to be $P=.40$, then the probability that $x=4$ smokers out of $n=10$ subjects selected is:

$$\begin{aligned}
 P(X=4) &= {}^{10}C_4(0.4)^4(1-0.4)^{10-4} \\
 &= {}^{10}C_4(0.4)^4(0.6)^6 = 210(.0256) \\
 &= 0.25
 \end{aligned}$$

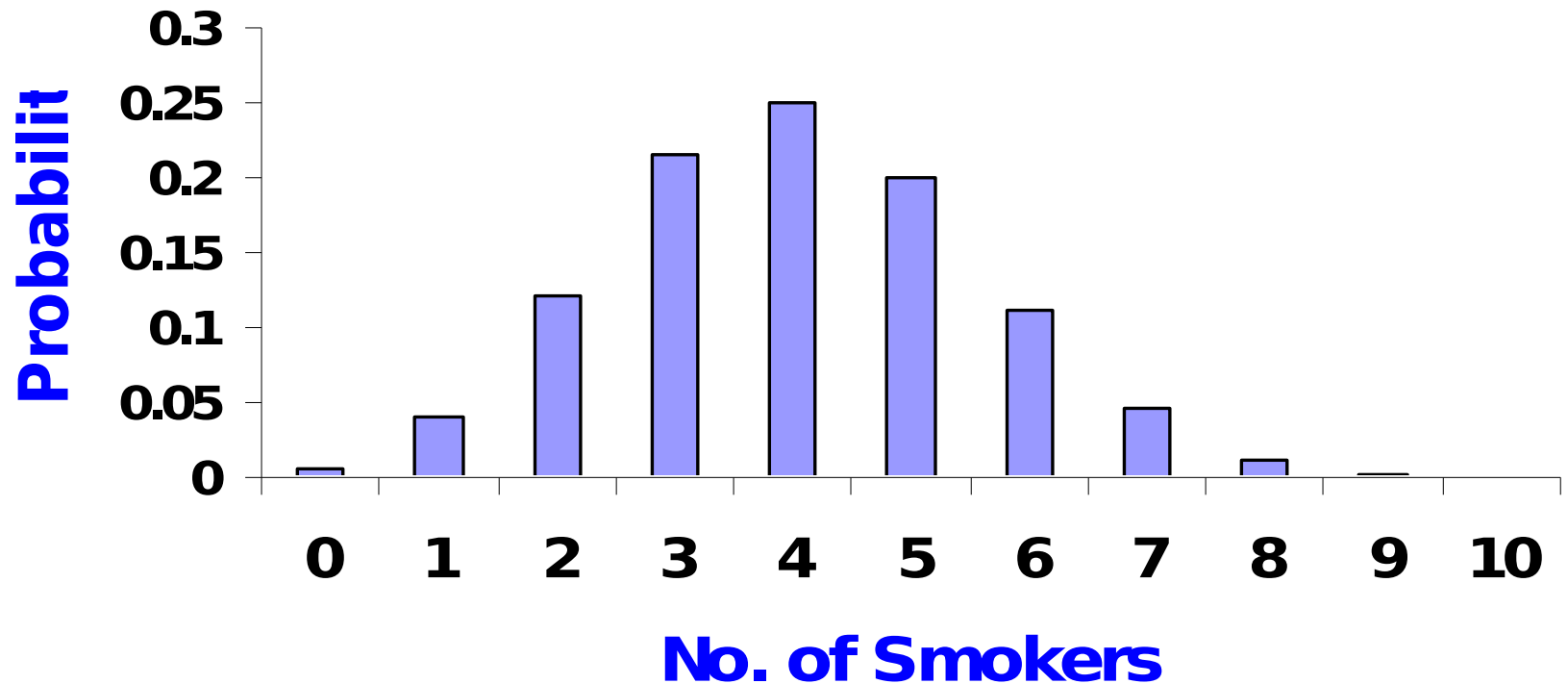
- The probability of obtaining exactly 4 smokers in the sample is about 0.25.

- We can compute the probability of observing zero smokers out of 10 subjects selected at random, exactly 1 smoker, and so on, and display the results in a table, as given, below.
- The third column, $P(X \leq x)$, gives the cumulative probability. E.g. the probability of selecting 3 or fewer smokers into the sample of 10 subjects is
$$P(X \leq 3) = .3823, \text{ or about } 38\%.$$

Binomial Probability Distribution for $n=10$, $p=.4$

x	$P(X=x)$	$P(X \leq x)$
0	.0060	.0060
1	.0404	.0464
2	.1209	.1673
3	.2150	.3823
4	.2508	.6331
5	.2007	.8338
6	.1114	.9452
7	.0452	.9877
8	.0106	.9983
9	.0016	.9999
10	.0001	1.000

The probability in the above table can be converted into the following graph



Exercise

Each child born to a particular set of parents

has a probability of 0.25 of having blood type

O. If these parents have 5 children.

What is the probability that

- a. Exactly two of them have blood type O.
- b. At most 2 have blood type O.
- c. At least 4 have blood type O.
- d. 2 do not have blood type O.

Solution for 'a'

a.)

$$P(x=2) = \binom{5}{2} (0.25)^2 (0.75)^{5-2} \\ = 0.2637$$

The Mean and Variance of a Binomial Distribution

- Once n and P are specified, we can compute the proportion of success,

$$P = x/n$$

- and the mean and variance of the distribution are given by :

$$E(X) = \mu = np, \sigma^2 = npq, \sigma = \sqrt{npq}$$

Example:

- 70% of a certain population has been immunized for polio. If a sample of size 50 is taken, what is the “expected total number”, in the sample who have been immunized?

$$\mu = np = 50(.70) = 35$$

- This tells us that “on the average” we expect to see 35 immunized subjects in a sample of 50 from this population.

- If repeated samples of size 10 are selected from the population of infants born, the mean number of children per sample who survive to age 70 would be

$$\mu = np = (10)(0.72) = 7.2$$

- The variance would be $npq = (10)(0.72)(0.28) = 2.02$ and the SD would be

$$\sqrt{2.02} = 1.42$$

2. The Poisson Distribution

- Is a discrete probability distribution used to model the number of occurrences of an event that takes place infrequently in time or space
- Applicable for counts of events over a given interval of time, for example:
 - number of patients arriving at an emergency department in a day
 - number of new cases of HIV diagnosed at a clinic in a month

- In such cases, we take a sample of days and observe the number of patients arriving at the emergency department on each day,
- or a sample of months and observe the number of new cases of HIV diagnosed at the clinic.
- We are observing a count or number of events, rather than a yes/no or success/failure outcome for each subject or trial, as in the binomial.

- In theory, a random variable X is a count that can assume any integer value greater than or equal to 0

- Suppose events happen *randomly* and *independently* in time at a constant rate. If events happen with rate λ events per unit time, the probability of x events happening in unit time is:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- where $x = 0, 1, 2, \dots, \infty$
- x is a potential outcome of X
- The constant λ (lambda) represents the rate at which the event occurs, or the expected number of events per unit time
- $e = 2.71828$
- It depends up on just one parameter, which is the μ number of occurrences (λ).

Three assumptions must be met for a Poisson distribution to apply:

1. The probability that a single event occurs within a given small subinterval is proportional to the length of the subinterval

$$P(\text{event}) \approx \lambda \Delta t \text{ for constant } \lambda$$

2. The rate at which the event occurs is constant over the entire interval t
3. Events occurring in consecutive subintervals are independent of each other

Example

- The daily number of new registrations of cancer is 2.2 on average.

What is the probability of

- a) Getting no new cases
- b) Getting 1 case
- c) Getting 2 cases
- d) Getting 3 cases
- e) Getting 4 cases

Solutions

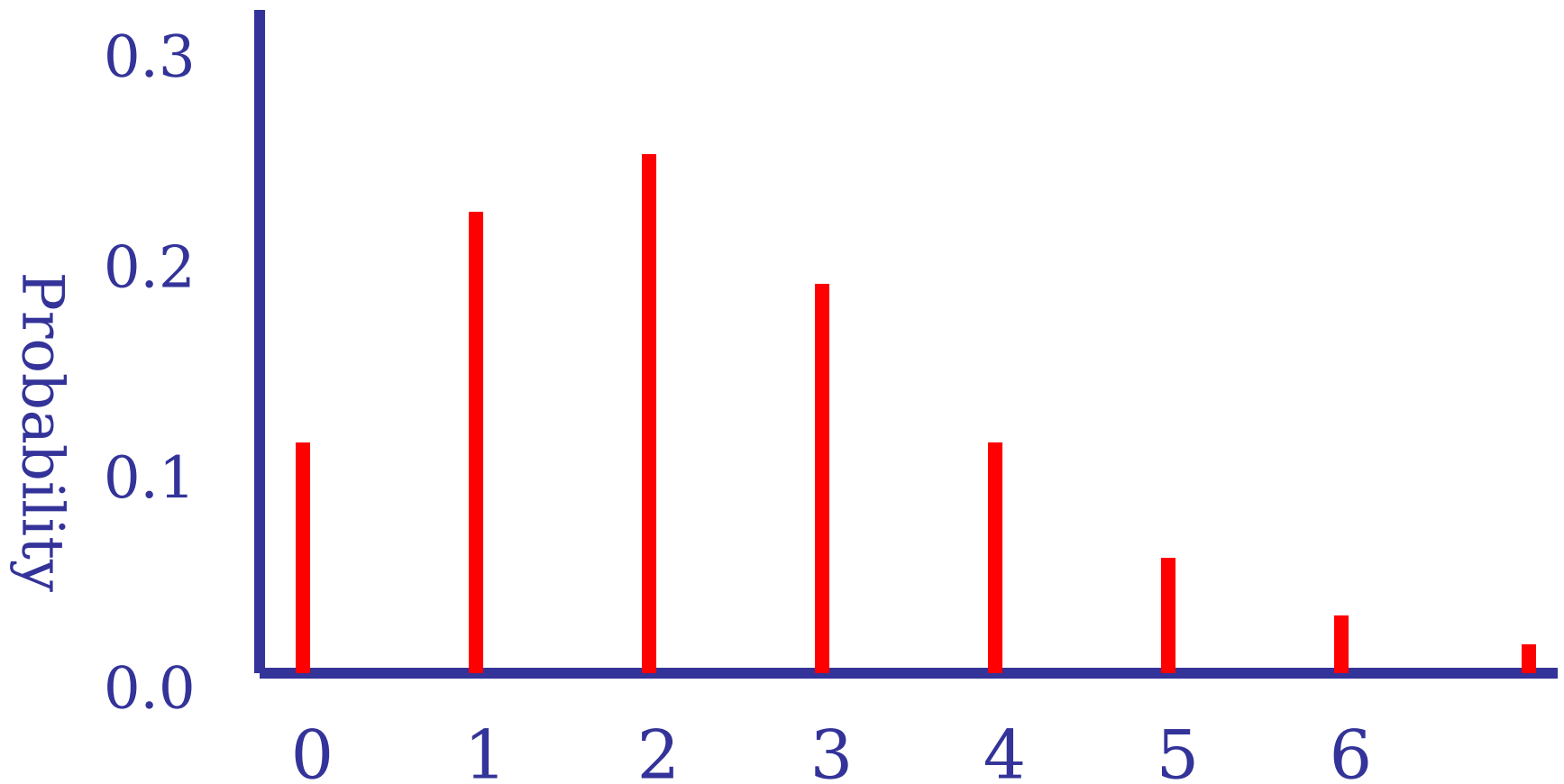
$$\text{a) } P(X=0) = \frac{(2.2)^0 e^{-2.2}}{0!} = 0.111$$

$$\text{b) } P(X=1) = 0.244$$

$$\text{c) } P(X=2) = 0.268$$

$$\text{d) } P(X=3) = 0.197$$

$$\text{e) } P(X=4) = 0.108$$



7
Poisson distribution with mean 2.2

Example:

- In a given geographical area, cases of tetanus are reported at a rate of $\lambda = 4.5/\text{month}$
- What is the probability that 0 cases of tetanus will be reported in a given month?

$$\begin{aligned}P(X = 0) &= \frac{e^{-4.5}(4.5)^0}{0!} \\&= \frac{e^{-4.5}(1)}{(1)} \\&= 0.011\end{aligned}$$

- What is the probability that 1 case of tetanus will be reported?

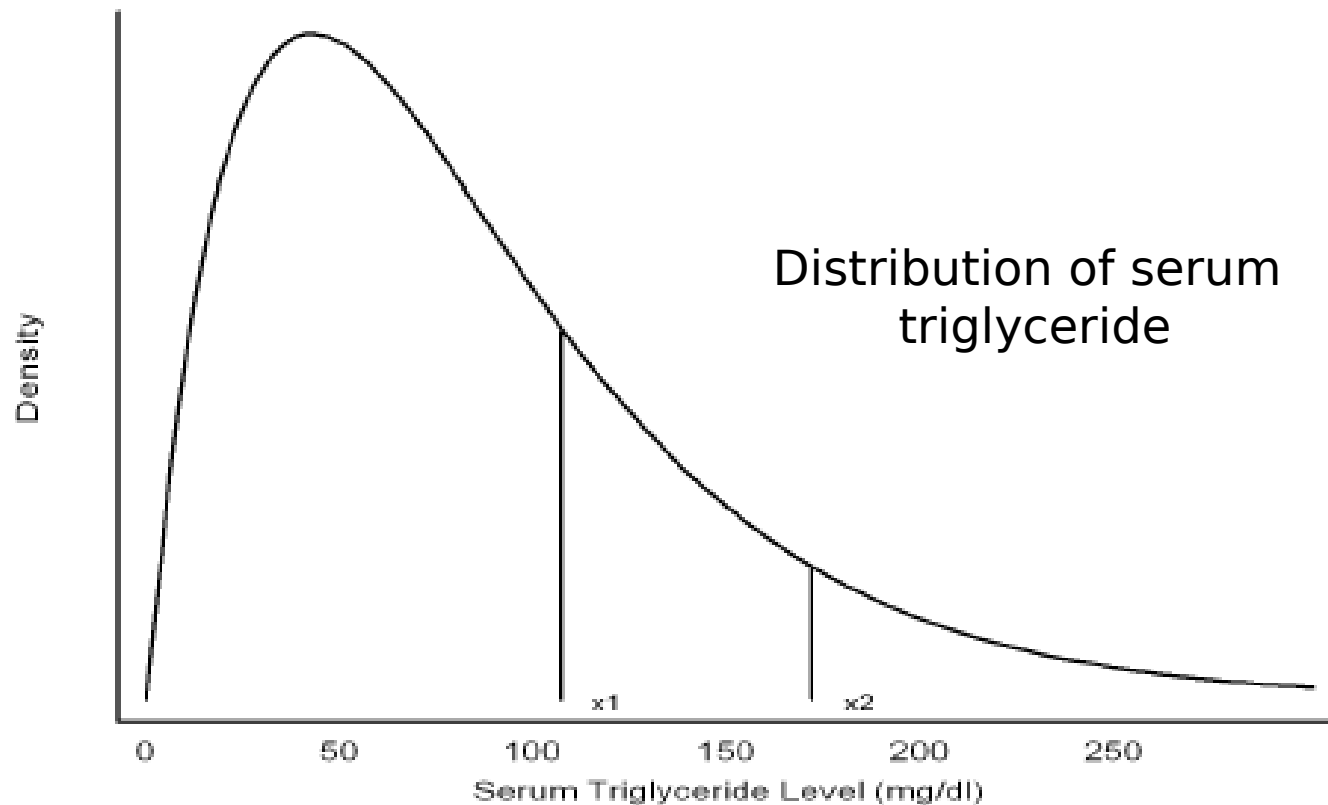
$$\begin{aligned}P(X = 1) &= \frac{e^{-4.5}(4.5)^1}{1!} \\&= \frac{e^{-4.5}(4.5)}{(1)} \\&= 0.050\end{aligned}$$

Characteristics

- The Poisson distribution is very asymmetric when its mean is small
- With large means it becomes nearly symmetric
- It has no theoretical maximum value, but the probabilities tail off towards zero very quickly
- $\forall \lambda$ is the parameter of the Poisson distribution
- The mean is λ and the variance is also λ .

B. Continuous Probability Distributions

- A continuous random variable X can take on any value in a specified interval or range
- With a large number of class intervals, the frequency polygon begins to resemble a smooth curve.
- The probability distribution of X is represented by a smooth curve called a **probability density function**



- The area under the smooth curve is equal to 1
- The area under the curve between any two points x_1 and x_2 is the probability that X takes a value between x_1 and x_2

- Instead of assigning probabilities to specific outcomes of the random variable X , probabilities are assigned to ranges of values
- The probability associated with any one particular value is equal to 0
- Therefore, $P(X=x) = 0$
- Also, $P(X \geq x) = P(X > x)$

- We calculate:
 $\Pr [a < X < b]$, the probability of an interval of values of X .
- For the above reason,

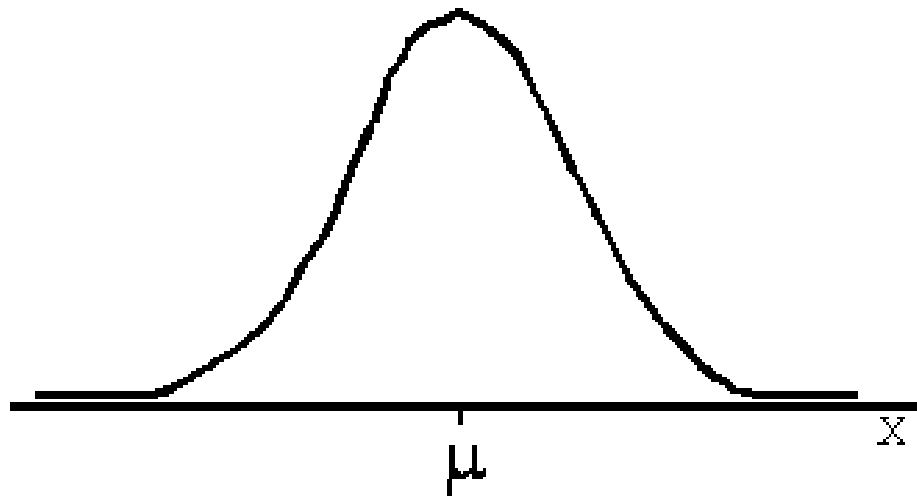
$$\sum_{-\infty}^{\infty} \Pr[X = x]$$

- is also without meaning.

The Normal distribution

- The ND is the most important probability distribution in statistics
- Frequently called the “Gaussian distribution” or **bell-shape** curve.
- Variables such as blood pressure, weight, height, serum cholesterol level, and IQ score — are approximately normally distributed

A random variable is said to have a normal distribution if it has a probability distribution that is symmetric and *bell-shaped*.



- The ND is vital to statistical work, most estimation procedures and hypothesis tests underlie ND
- The concept of “probability of $X=x$ ” in the discrete probability distribution is replaced by the “probability density function $f(x)$ ”
- The ND is also an approximating distribution to other distributions (e.g., binomial)

- A random variable X is said to follow ND, if and only if, its probability density function is:

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

- ✓ π (pi) = 3.14159
- ✓ $e = 2.71828$, x = Value of X
- ✓ Range of possible values of X : $-\infty$ to $+\infty$
- ✓ μ = Expected value of X (“the long run average”)
- ✓ σ^2 = Variance of X .
- ✓ μ and σ are the parameters of the normal distribution — they completely define its shape

The normal distribution with mean μ and variance σ^2 is represented by $N(\mu, \sigma^2)$

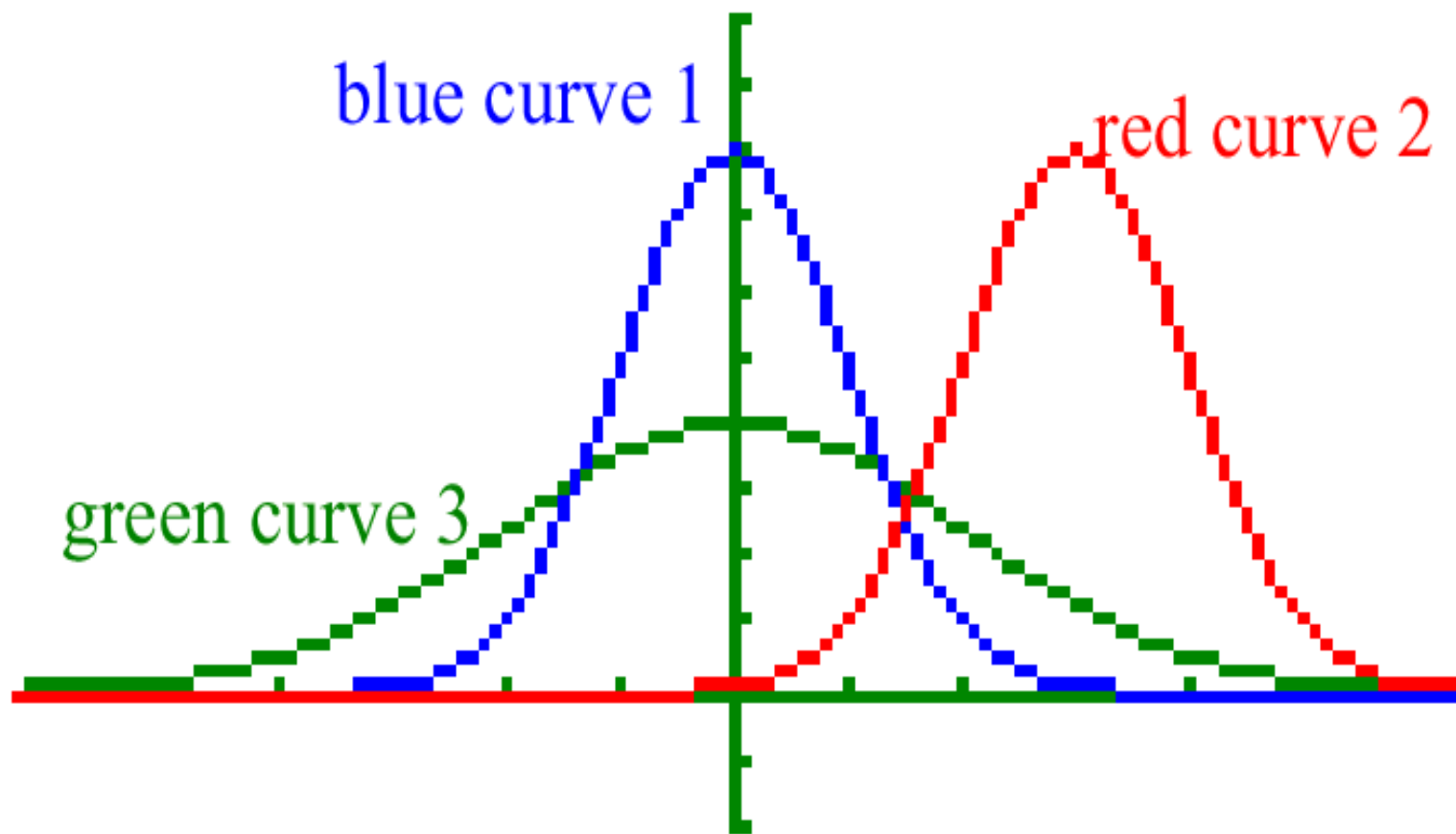
To find $P(X \leq x)$, we would have to draw the probability density function of $N(\mu, \sigma^2)$ and determine the area to the left of x

1. The mean μ tells you about location -

- Increase μ - Location shifts right
- Decrease μ - Location shifts left
- Shape is unchanged

2. The variance σ^2 tells you about narrowness or flatness of the bell -

- Increase σ^2 - Bell flattens. Extreme values are more likely
- Decrease σ^2 - Bell narrows. Extreme values are less likely
- Location is unchanged



$$\mu_1 = \mu_3 < \mu_2 \quad \text{but} \quad \sigma_1 = \sigma_2 < \sigma_3$$

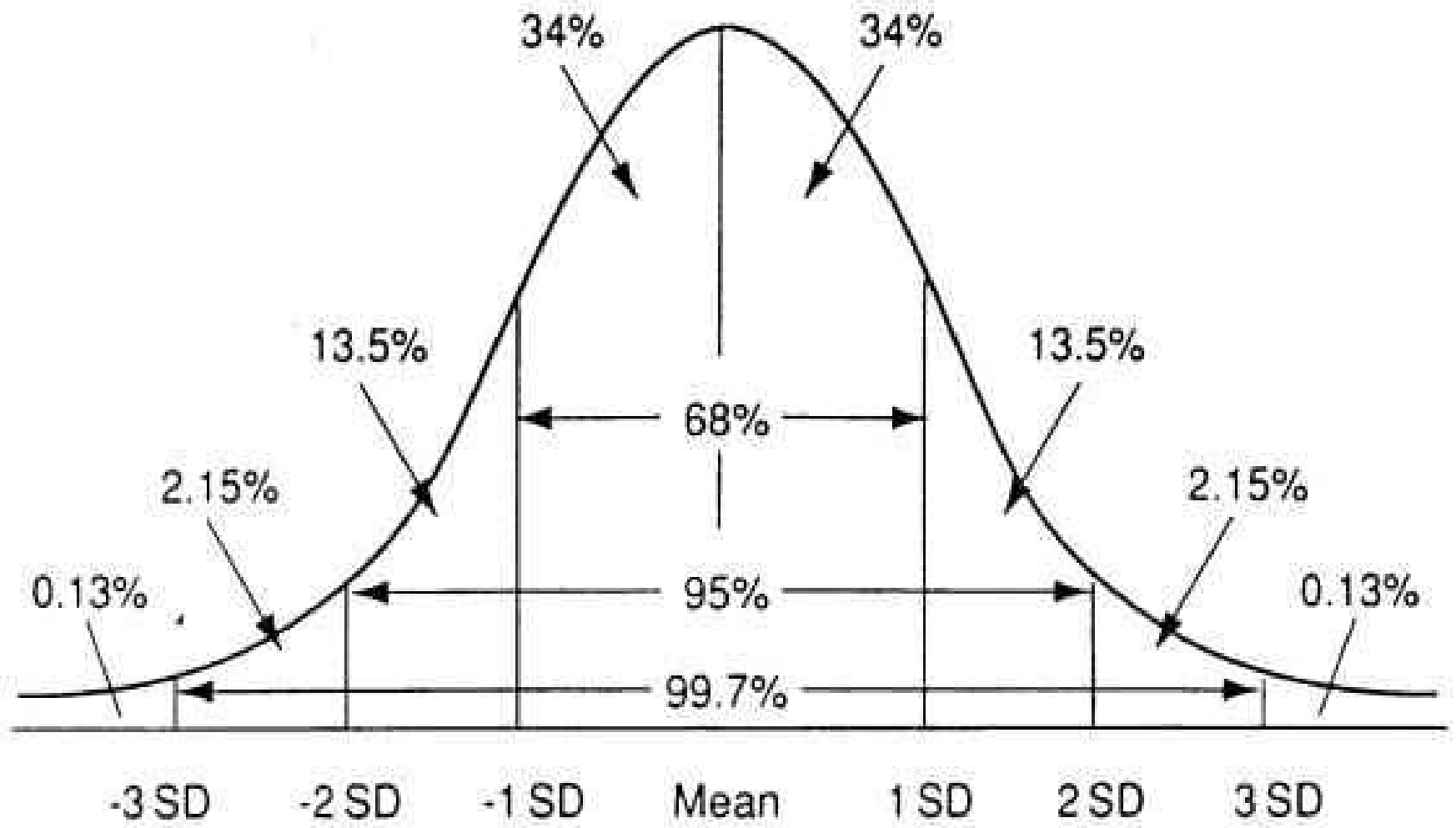
Properties of the Normal Distribution

1. It is symmetrical about its mean, μ .
2. The mean, the median and mode are almost equal. It is unimodal.
3. The total area under the curve about the x-axis is 1 square unit.
4. The curve never touches the x-axis.
5. As the value of σ increases, the curve becomes more and more flat and vice versa.

6. Perpendiculars of:
- $\pm$ SD contain about 68%;
 - ± 2 SD contain about 95%;
 - ± 3 SD contain about 99.7%
of the area under the curve.

Next slide

7. The distribution is completely determined by the parameters μ and σ .

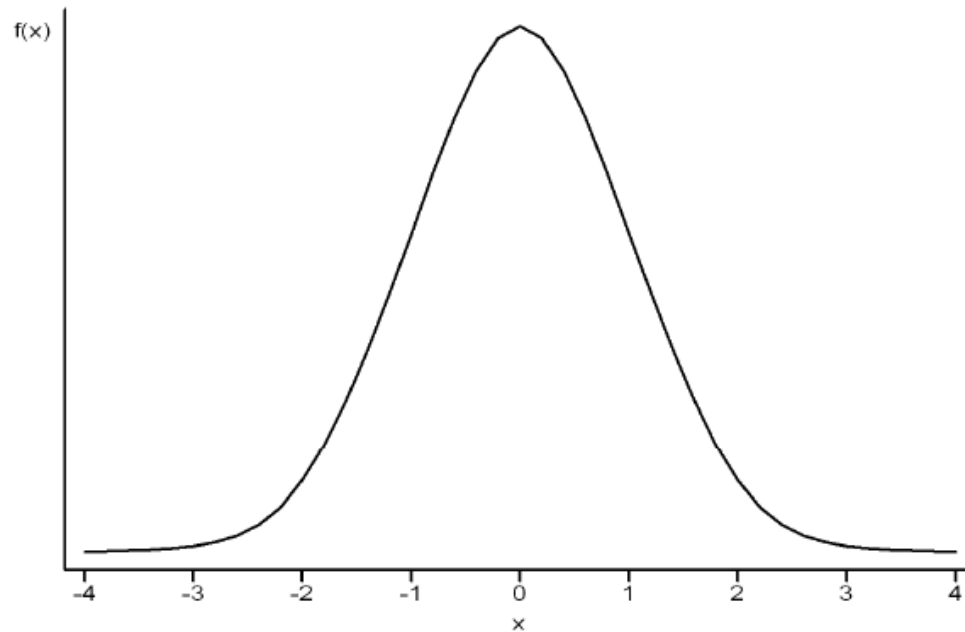


- We have different normal distributions depending on the values of μ and σ^2 .
- We cannot tabulate every possible distribution
- Tabulated normal probability calculations are available only for the ND with $\mu = 0$ and $\sigma^2=1$.

Standard Normal Distribution

- It is a normal distribution that has a mean equal to 0 and a SD equal to 1, and is denoted by $N(0, 1)$.
- The main idea is to standardize all the data that is given by using Z-scores.
- These Z-scores can then be used to find the area (and thus the probability) under the normal curve.

The standard normal distribution has mean 0 and variance 1



- *Approximately 68% of the area under the standard normal curve lies between ± 1 , about 95% between ± 2 , and about 99% between ± 2.5*

Z - Transformation

- If a random variable $X \sim N(\mu, \sigma)$ then we can transform it to a SND with the help of Z-transformation

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

- Z represents the Z-score for a given **x** value

- Consider redefining the scale to be in terms of how many SDs away from mean for normal distribution, $\mu=110$ and $\sigma=15$.

Value

x

50 65 80 95 110 125 140 155

170

-4 -3 -2 -1 0 1 2 3

4

SDs from mean using

$$(x-110)/15 = (x-\mu)/\sigma$$

- This process is known as standardization and gives the position on a normal curve with $\mu=0$ and $\sigma=1$, i.e., the SND, Z .
- A **Z-score** is the number of standard deviations that a given x value is above or below the mean.

Finding normal curve areas

1. The table gives areas between $-\infty$ and the value of z_0 .
2. Find the z value in tenths in the column at left margin and locate its row. Find the hundredths place in the appropriate column.
3. Read the value of the area (P) from the body of the table where the row and column intersect. Values of P are in the form of a decimal point and four places.

Some Useful Tips

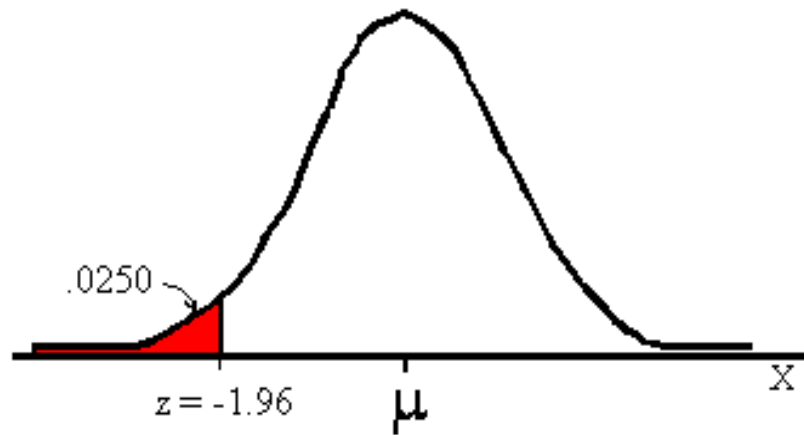
$$\Pr [Z \leq -z] = \Pr [Z \geq +z]$$

$$\Pr [Z < z] + \Pr [Z \geq +z] = 1$$

$$\Pr [Z \geq 0] = .5$$

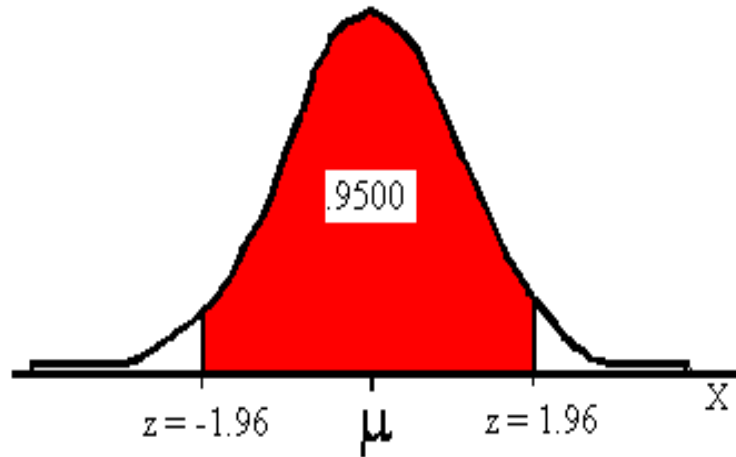
$$\Pr [Z \leq 0] = .5$$

a) What is the probability that $z < -1.96$?



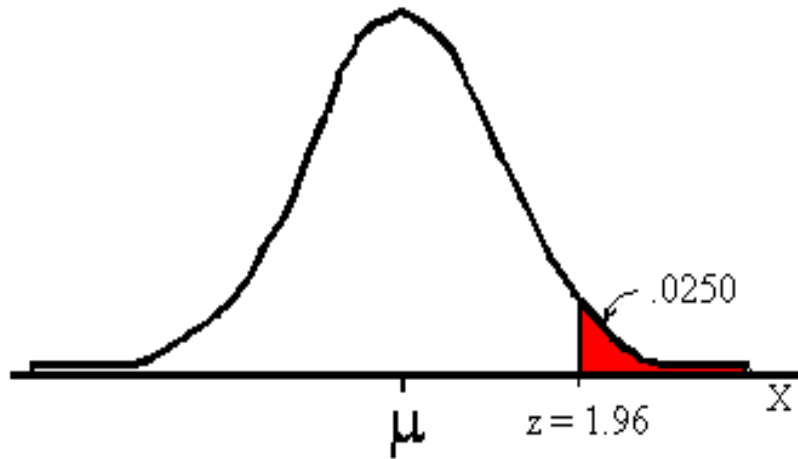
- (1) Sketch a normal curve
- (2) Draw a perpendicular line for $z = -1.9$
- (3) Find the area in the table
- (4) The answer is the area to the left of the line $P(z < -1.96) = 0.0250$

b) What is the probability that $-1.96 < z < 1.96$?

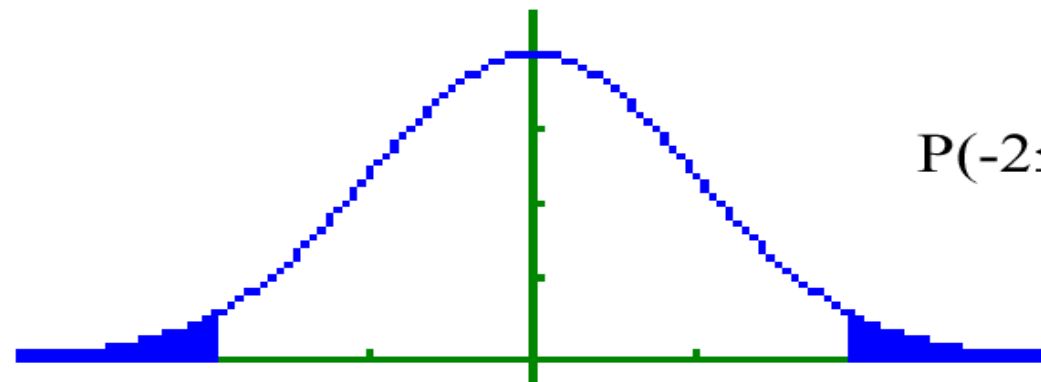


The area between the values $P(-1.96 < z < 1.96) = .9750 - .0250 = .9500$

c) What is the probability that $z > 1.96$?



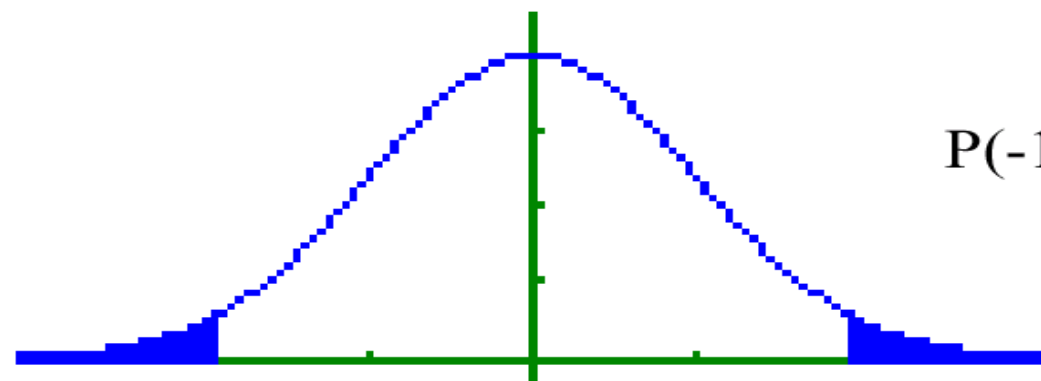
- The answer is the area to the right of the line; found by subtracting table value from 1.0000; $P(z > 1.96) = 1.0000 - .9750 = .0250$



$$P(-2 \leq Z \leq 2) = 0.9544$$

$Z = 2$

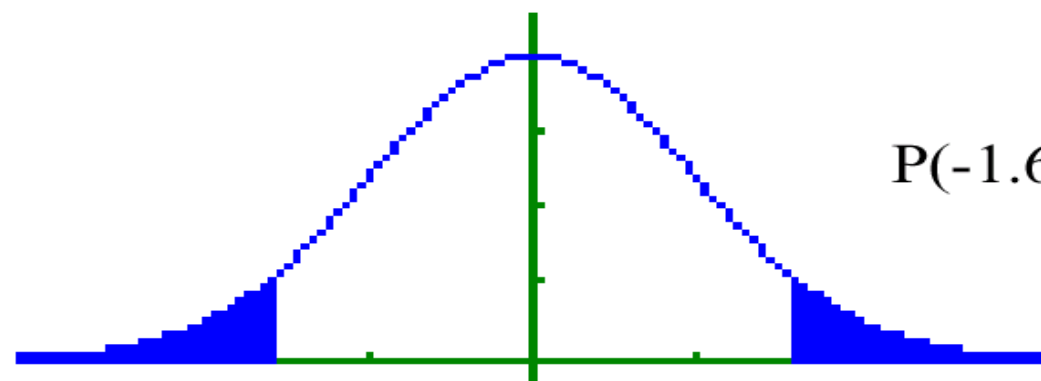
$P = 0.0455003005$



$$P(-1.96 \leq Z \leq 1.96) = 0.9500$$

$Z = 1.96$

$P = 0.0499958428$



$$P(-1.645 \leq Z \leq 1.645) = 0.9000$$

$Z = 1.645$

$P = 0.099969771$

Also $P(-3 \leq Z \leq 3) = 0.9974$ ^{hu} & $P(-3.5 \leq Z \leq 3.5) = 0.9996$ ¹²⁷

Exercise

1. Compute $P(-1 \leq Z \leq 1.5)$

Ans: 0.7745

2. Find the area under the SND from 0 to 1.45

Ans: 0.4265

3. Compute $P(-1.66 < Z < 2.85)$

Ans: 0.9493

Applications of the Normal Distribution

- The ND is used as a model to study many different variables.
- The ND can be used to answer probability questions about continuous random variables.
- Following the model of the ND, a given value of x must be converted to a z score before it can be looked up in the z table.

Example:

- The diastolic blood pressures of males 35–44 years of age are normally distributed with $\mu = 80$ mm Hg and $\sigma^2 = 144$ mm Hg²
 $\sigma = 12$ mm Hg
- Therefore, a DBP of $80 + 12 = 92$ mm Hg lies 1 SD above the mean
- Let individuals with BP above 95 mm Hg are considered to be hypertensive

- a. What is the probability that a randomly selected male has a BP above 95 mm Hg?

$$\begin{aligned}P(X > 95) &= P\left(\frac{X - 80}{12} > \frac{95 - 80}{12}\right) \\&= P(Z > 1.25) \\&= 0.1056\end{aligned}$$

- Approximately 10.6% of this population would be classified as hypertensive

b. What is the probability that a randomly selected male has a DBP above 110 mm Hg?

$$Z = \frac{110 - 80}{12} = 2.50$$

$$P(Z > 2.50) = 0.0062$$

- Approximately 0.6% of the population has a DBP above 110 mm Hg

c. What is the probability that a randomly selected male has a DBP below 60 mm Hg?

$$Z = \frac{60 - 80}{12} = -1.67$$

$$P(Z < -1.67) = 0.0475$$

- Approximately 4.8% of the population has a DBP below 60 mm Hg

d. What value of DBP cuts off the upper 5% of this population?

- Looking at the table, the value $Z = 1.645$ cuts off an area of 0.05 in the upper tail
- We want the value of X that corresponds to $Z = 1.645$

$$Z = \frac{X - \mu}{\sigma}$$

$$1.645 = \frac{X - \mu}{\sigma}$$

$$X = (12)(1.645) + 80$$

$$X = 99.7$$

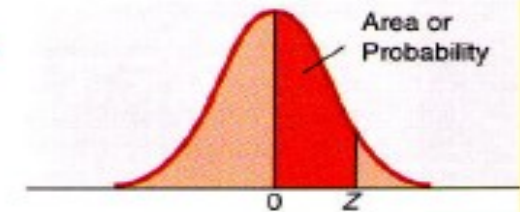
Approximately 5% of the men in this population have a DBP greater than 99.7 mm Hg

Other Distributions

1. Student t-distribution
2. F- Distribution
3. χ^2 -Distribution

Standard Normal Distribution: Z-values

Entries in the table give the area under the curve between the mean and Z standard deviations above the mean. For example, for $Z = 1.25$, the area under the curve between the mean and Z is .3944.



Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4899	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4986	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

TABLE 2 Upper-tail Areas for the Normal Curve



<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.00	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.10	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.20	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.30	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.40	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.50	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.60	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.70	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.80	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.90	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.00	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.10	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.20	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.30	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.40	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.50	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.60	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.70	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.80	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.90	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.00	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.10	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.20	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.30	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.40	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.50	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.60	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.70	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.80	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.90	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.00	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

<i>z</i>	Area
3.500	.00023263
4.000	.00003167
4.500	.00000340
5.000	.00000029

Source: Computed by J. W. Stegeman using SAS.