Probability and Statistics

Pritha Banerjee

University of Calcutta

banerjee.pritha74@gmail.com

March 5, 2025

Overview

1. Probability
2. Probability Distribution
   - Popular Distribution
3. Statistics and their Distribution

Random Phenomena and Probability

- **Deterministic phenomenon:** Phenomenon whose outcome can be predicted with a **very high degree** of confidence. Ex: age from date of birth can be calculated.
- **Stochastic phenomenon:** Phenomenon which can have many possible outcomes for same experimental conditions. Outcomes predicted with **limited** confidence
- Unknown sources of data, data generation process causes errors in data.
- Errors are modeled using probability
- Random phenomenon are of two types:
  - **Discrete** - Finite outcomes. Ex. Tossing a coin.
  - **Continuous** - Infinite number of outcomes. Ex: Body temperature measurement in degree Fahrenheit.

Discrete phenomena - Discrete random variable

- **Sample space** S: Set of all possible outcomes of a random phenomena or experiment. Ex. Two coin toss; $S = \{HH, HT, TH, TT\}$
- **Event** A: Subset of a sample space. Ex: Occurrence of 1 H in first toss of a two coin toss experiment, $A = \{HH, HT\} \subseteq S$
- Each outcome of a sample space is an elementary event.

Probability Measure

- Given an experiment and a sample space S, the objective of probability is to assign to each event A a number $P(A)$, called the probability of the event A, which will give a precise measure of the chance that A will occur.
- All assignment must satisfy the following axioms(basic properties):
  - For any event A, $P(A) \geq 0$.
  - $P(S) = 1$.
  - If $A_1, A_2, \cdots A_n$ is an infinite collection of disjoint events, then $P(A_1 \cup A_2 \cup A_3 \cdots) = \sum_{i=1}^{\infty} P(A_i)$
- $P(\phi) = 0$

Interpretation of Probability Measure (1)

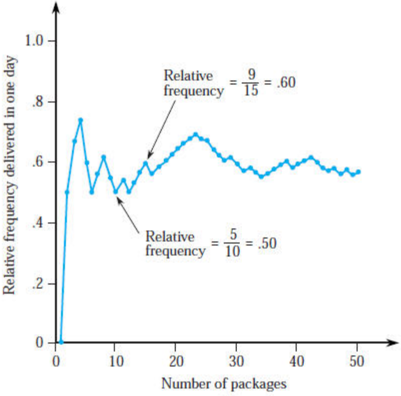
- Consider an experiment is repeatedly performed n times in an identical and independent fashion, and let A be an event consisting of a fixed set of outcomes of the experiment. Ex. $A =$ Obtaining head (H) in tossing a coin.
- Let $n(A)$ is the number of replications (trials) on which A does occur.
- $\frac{n(A)}{n}$ is called the **relative frequency** of occurrence of the event A in the sequence of n replications/ trials.
- As n gets arbitrarily large, relative frequency gets stabilized, i.e, it approaches a limiting value referred to as the limiting (or long-run) relative frequency of the event A;

$$P(A) = \lim_{n\to\infty} \frac{n(A)}{n}$$

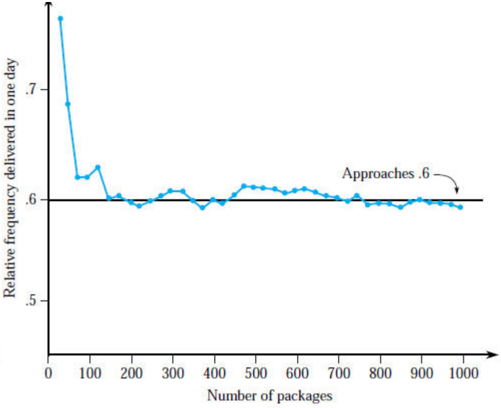
Interpretation of Probability Measure (2)

A be the event that a package sent for 2nd day delivery, actually arrives within one day. The results from sending 10 such packages:

Package #	1	2	3	4	5	6	7	8	9	10
Did A occur?	N	Y	Y	Y	N	N	Y	Y	N	N
Rel. freq. of A $\left(\frac{n(A)}{n}\right)$	0	.5	.667	.75	.6	.5	.571	.625	.556	.5



(a)



(b)

Properties of probability

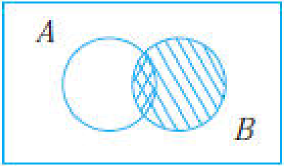
- For any event A, $P(A) + P(A') = 1$, from which $P(A) = 1 - P(A')$.
- For any event A, $P(A) \leq 1$.
- For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- **Independent events**: Two events are independent if occurrence of one has no influence on occurrence of other. Events A and B are independent if and only if $P(A \cap B) = P(A).P(B)$. Ex: Two coin tossing.
- **Mutually exclusive**: Two events are mutually exclusive if occurrence of one implies non occurrence of other event. Events A and B are mutually exclusive iff $P(A \cup B) = P(A) + P(B)$.

Boole's Inequality

- $P(\bigcup\limits_{i=1}^{\infty} A_i) \leq \sum\limits_{i=1}^{\infty} P(A_i)$ for any sets $A_1, A_2, \cdots$
- Upper bound for probability of union of events
- Equality holds when A_is are disjoint.

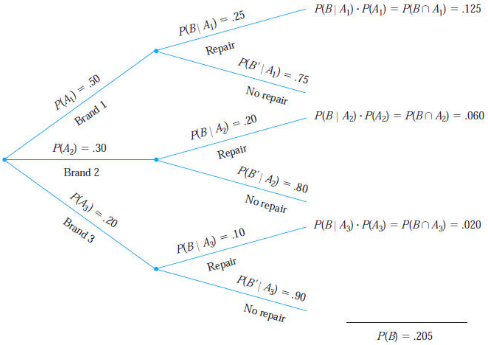
Conditional probability

- For any two events A and B with $P(B) > 0$, the **conditional probability of** A given that B has occurred is defined by $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
- Example with figure:
- **Multiplication Rule** : multiply above by $P(B)$, i.e, $P(A \cap B) = P(A|B).P(B)$.
- **prior and posterior probability of an event**: Before conditional probability is applied, an event has prior probability. With the conditional probability applied an event will get a posterior probability.



Conditional probability: Example

- A video stores sells three different brands of DVD players. Of its DVD player sales, 50% are brand 1 , 30% are brand 2, and 20% are brand 3. It is known that 25% of brand 1's DVD players require warranty repair work, whereas the corresponding percentages for brands 2 and 3 are 20% and 10%, respectively.
- What is the probability that a randomly selected purchaser has bought a brand 1 DVD player that will need repair while under warranty? $P(A_1 \cap B) = P(B|A_1).P(A_1) = .125$
- What is the probability that a randomly selected purchaser has a DVD player that will need repair while under warranty? $P(B) = P[(brand1 \wedge repair) \vee (brand2 \wedge repair) \vee (brand3 \wedge repair)] = .125 + .060 + .020 = .205$
- If a customer returns to the store with a DVD player that needs repair work, what is the probability that it is a brand 1 DVD player? A brand 2 DVD player? A brand 3 DVD player? $P[A_1|B] = \frac{.125}{.205} = .61$, $P[A_2|B) = \frac{.060}{.205} = .29$, $P[A_3|B) = 1 - P[A_1|B) - P[A_2|B] = .10$

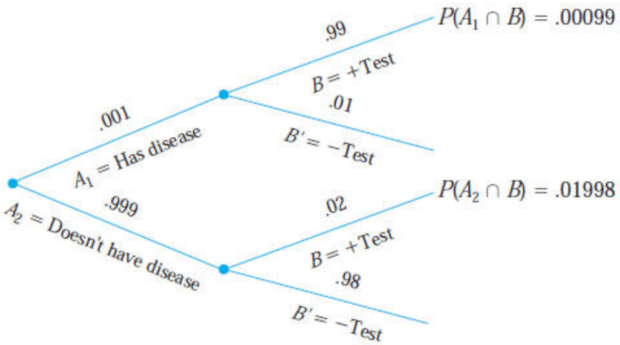


Bayes' Theorem

- **The Law of Total Probability**: Let $A_1, A_2, \cdots, A_k$ be mutually exclusive and exhaustive (one A_i must occur so that $A_1 \cup \cdots \cup A_k = S$) events. Then for any other event B, $P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k) = \sum_{i=1}^{k} P(B|A_i).P(A_i)$
- **Bayes' Theorem**: Let $A_1, A_2, \cdots, A_k$ be a collection of k mutually exclusive and exhaustive events with *prior probabilities* $P(A_i)$, $i = 1, 2, \cdots k$. Then for any other event B for which, the *posterior probability* of A_j given that B has occurred is $P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j).P(A_j)}{\sum_{i=1}^{k} P(B|A_i).P(A_i)}$.
- **Example**:Only 1 in 1000 adults is afflicted with a rare disease for which a diagnostic test is developed. when an individual actually has the disease, a positive result occurs 99% of the time, whereas without the disease shows a positive result only 2% of the time. If a randomly selected individual is tested and the result is positive, what is the probability that the individual has the disease?

Bayes' Theorem: Example

- A_1 = individual has the disease, A_2 = individual does not have the disease, and B = positive test result. Then , $P(A_1) = .001, P(A_2) = .999, P(B|A_1) = .99, P(B|A_2) = .02.$
- $P(B) = .00099 + .01998 = .02097$ [law of total probability : $P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2)$]
- $P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{.00099}{.02097} = .047$



Bayes' Theorem: Generalization

- Let $A_1, A_2, \cdots$ be a partition pf sample space S and let B be any subset of Sample space, then for each $i = 1, 2, \cdots$,
$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum\limits_{j=1}^{\infty} P(B|A_j)P(A_j)}$
- it helps in computing the conditional probability $P(A|B)$ from inverse conditional probability $P(B|A)$

Independence: revisited

- Two events A and B are independent if $P(A|B) = P(A)$ and are dependent otherwise.
- Example: Consider a gas station with six pumps numbered $1, 2, \cdots, 6$ and let E_i denote the simple event that a randomly selected customer uses pump i $(i = 1, 2, \cdots 6)$. Suppose that $P(E_1) = P(E_6) = .10, P(E_2) = P(E_5) = .15, P(E_3) = P(E_4) = .25$ Define events A, B, C by $A = \{2, 4, 6\}, B = \{1, 2, 3\}, C = \{2, 3, 4, 5\}$. We then have $P(A) = .50, P(A \cup B) = .30, andP(A \cup C) = .50$. That is, events A and B are dependent, whereas events A and C are independent. Intuitively, A and C are independent because the relative division of probability among even- and odd-num- bered pumps is the same among pumps $2, 3, 4, 5$ as it is among all six pumps.

Conditional Independence

- Let A, B, C are three events with $P(C) > 0$. Given C, the events A and B are conditionally independent if $P(A \cap B|C) = P(A|C)P(B|C)$ or $P(A|B \cap C) = P(A|C)$

Random Variable

- **Definition**: For a given sample space S of some experiment, a random variable (rv) is any rule that associates a number with each outcome in S. An rv is a function $X : S \to \mathbb{R}$ whose domain is the sample space S and range is the set of real numbers.
- Let sample space $S = Success, Failure$. Random variable $X(Success) = 1, X(Failure) = 0$
- **Bernoulli random variable**: Any random variable whose only possible values are 0 and 1 is called a Bernoulli random variable.

Induced probability function

- Let $S = w_1, w_2 \cdots$ be a sample space and P be a probability measure(function)
- Let X be a random variable with range $X = \{x_1, x_2, \cdots x_m\}$
- Induced probability function P_X on x is $P_X(X = x_i) = P(\{w_j \in S : X(w_i) = x_j\})$
- Example: X : number of heads obtained in three coin tosses.
    - Enumerate the elementary outcomes $w = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$ and $X(w) = \{3, 2, 2, 2, 1, 1, 1, 0\}$
    - Measure the probability of random variable taking on value in its range, i.e, $X = \{0, 1, 2, 3\}$
    - Thus, $P_X(X = x) = \{1/8, 3/8, 3/8, 1/8\}$

Types of Random Variable (RV)

- **Discrete RV**: A discrete random variable is an RV whose possible values either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on ("countably" infinite).
- **Continuous RV**: A random variable is continuous if both of the following apply:
  - set of possible values consists either of all numbers in a single interval on the number line (possibly infinite in extent, e.g., from $-\infty to + \infty$) or all numbers in a disjoint union of such intervals (e.g.,$[0, 10] \cup [20, 30]$).
  - No possible value of the variable has positive probability, that is, $P(X = c) = 0$ for any possible value c.

Probability distribution for Discrete RV (1)

- The probability distribution of X says how the total probability of 1 is distributed among (allocated to) the various possible X values. $P(X = c)$ is denoted as $p(x)$.
- **probability distribution** or **probability mass function (pmf)** of a discrete RV is defined for every number x by $p(x) = P(X = x) = P(\forall s \in S : X(s) = x)$.
- Properties: $p(x) \geq 0, \forall x$ and $\sum xp(x) = 1$
- **Probability histogram**: For each y with $p(y)$, construct a rectangle centered at y. The height of each rectangle is proportional to $p(y)$, and the base is the same for all rectangles.
- $p(x)$ gives a model for the distribution of population values, where population consists of the values of RV X
- Having a population model, use it to compute values of population characteristics (e.g., the mean μ) and make inferences about such characteristics.

Example: Consider 5 Blood donors : a, b, c, d, e. a and b have $O+$. Five blood samples, one from each individual, will be typed in random order until an $O+$ individual is identified. Let RV, X is the number of typings necessary to identify an $O+$ individual. Then the *pmf* of X

$p(1) = P(X = 1) = \frac{2}{5} = 0.4$

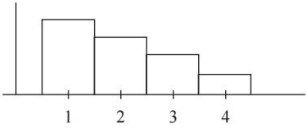
$p(2) = P(X = 2) = P(c \vee d \vee e \text{ first}) \ . \ P(a \vee b \text{ next } |c \vee d \vee e \text{ first})$
$= \frac{3}{5}.\frac{2}{4} = .3$

$p(3) = P(X = 3) = P(c, d \vee e \text{ first and second, then } a \vee b) = \frac{3}{5}.\frac{2}{4}.\frac{2}{3} = .2$

$p(4) = P(X = 4) = P(c, d, \wedge e \text{ first}) = \frac{3}{5}.\frac{2}{4}.\frac{1}{3} = .1$

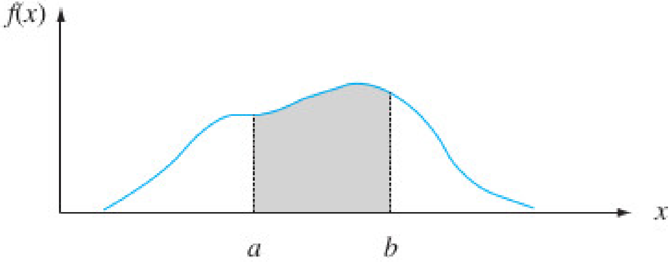
pmf is

$x =$	1	2	3	4
$p(x) =$	0.4	0.3	0.2	0.1



Probability density function for continuous r.v

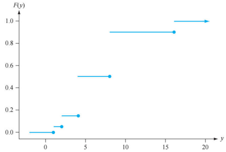
- Probability Density function, PDF, of a continuous r. v is the function $f_X(x)$ that satisfies $F_X(x) = \int\limits_{-\infty}^{x} f_X(t)dt, \forall x$
- Properties: $f_x(x) \geq 0, \forall x$ and $\int\limits_{-\infty}^{\infty} f_X(x)dx = 1$



Cumulative distribution function (cdf) (1)

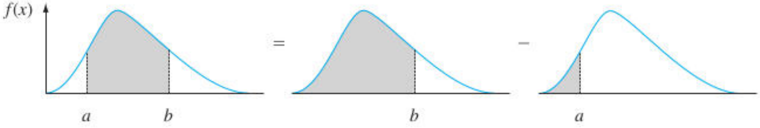
- The cumulative distribution function (cdf) $F(x)$ of a discrete random variable X with pmf $p(x)$ is defined for every number x by $F(x) = P(X \leq x) = \sum_{y:y\leq x} p(y)$.
- For X a discrete rv, the graph of $F(x)$ will have a jump at every possible value of X and will be flat between possible values. Such a graph is called a *step function*.
- For any integer a, b such that $a \leq b$, $P(a \leq X \leq b) = F(b) - F(a^{-})$, i.e $P(a \leq X \leq b) = P(X = a \vee a+1 \vee \cdots \vee b) = F(b) - F(a-1)$
- Example: the probability distribution of r.v $y = \{1, 2, 4, 8, 16\}$ is given as $p(y) = \{.05, .10, .35, .40, .10\}$, then $F(y) = \{.5, .5 + .10 = .15, .15 + .35 = .50, .50 + .40 = .90, .90 + .10 = 1\}$

$$F(y) = \begin{cases} 0, & \text{if } y < 1 \\ 0.05, & \text{if } 1 \leq y < 2 \\ 0.10, & \text{if } 2 \leq y < 4 \\ 0.50, & \text{if } 4 \leq y < 8 \\ 0.90, & \text{if } 8 \leq y < 16 \\ 1, & \text{if } 16 \leq y \end{cases} \tag{1}$$



Cumulative distributions function (cdf) (2)

Computing $P(a \leq X \leq b)$ from cumulative probabilities

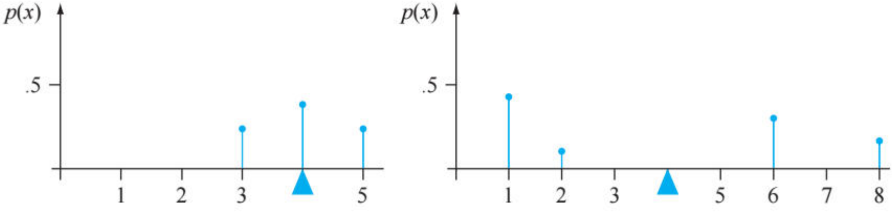


Expected value of X and function $h(X)$

- Let X be a discrete rv with set of possible values D and pmf $p(x)$. The expected value or mean value of X, i.e, $E(X) = \mu_X = \mu = \sum_{x \in D} x.p(x)$
- μ can be interpreted as the long-run average observed value of X when the experiment is performed repeatedly.
- The $E(X)$ describes where the probability distribution is centered.
- If the rv X has a set of possible values D and pmf $p(x)$, then the expected value of any function $h(X)$, denoted by $E[h(X)] = \sum_{x \in D} h(x).p(x)$
- $E(aX + b) = a.E(X) + b$ when $h(X)$ is of the form $aX + b$
- Expectation: $E[x] = \int\limits_{-\infty}^{\infty} x f_X(x) dx$, for probability density function, pdf.

Variance of RV X

- Used to capture the spread or variability in the distribution of X.
- Let X have pmf $p(x)$ and expected value μ. Then the variance of X, $V(X) = \sigma_X^2 = \sigma^2 = \sum_D (x-\mu)^2.p(x) = E[(X-\mu)^2] = E[X^2] - E[X]^2$
- Standard deviation of X is $\sigma_X = \sqrt{\sigma_x^2}$
- σ can be interpreted as the size of a representative deviation from the mean value μ. Example: $\sigma = 10$ means typical deviation from the mean will be something on the order of 10.
- $V(aX+b) = \sigma_{aX+b}^2 = a^2.\sigma_X^2$ and $SD(X) = \sigma_{aX+b} = |a|.\sigma_X$
- Different probability distribution with same $\mu = 4$ but different spread.



(a) (b)

Covariance and Correlation

- The **covariance** of two random variables X and Y is $Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$
- **Covariance** is a measure of how much two random variables changes together
- **Correlation** of two random variables X and Y is $\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X).Var(Y)}}$, thus, ρ lies in $[-1,1]$
- **Correlation** is normalized Covariance.

Joint and Marginal Distribution

- **Joint Distribution:** To capture the properties of two random variables use **Joint** PMF
$$f_{X,Y} = \mathbb{R}^2 \to [0,1]$$
defined by
$$f_{X,Y}(x,y) = P(X = x, Y = y)$$
, where $\sum_X \sum_Y f_{X,Y}(x,y) = 1$
- **Marginal Distribution:** Given Joint PMF $f_{X,Y}(x,y) = P(X = x, Y = y)$, we can obtain the PMF of two random variables:
$$f_X = \sum_y f_{X,Y}(x,y), \text{marginal PMF of } X$$
$$f_Y = \sum_x f_{X,Y}(x,y), \text{marginal PMF of } Y$$
- For **continuous** random variable $\sum$ is replaced by $\int$

Conditional Distribution

- Conditional distribution

$$f_{X|Y} = P(X = x|Y = y)$$

is defined using conditional probability

$$f_{X|Y} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Parameter of Probability distribution : Bernoulli Distribution

- **pmf of Bernoulli RV**: $p(1) = \alpha, p(0) = 1 - \alpha$. pmf is

$$p(x, \alpha) = \begin{cases} 1 - \alpha, & \text{if } x = 0. \\ \alpha, & \text{if } x = 1. \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

- Each choice of **parameter** α gives different pmf. Collection of all probability distributions for different values of the parameter is called a **family of probability distributions.**
- Expectation $E[X] = p$ and $Var[x] = p(1 - p)$

Binomial Distribution (1)

- **Binomial experiment**: An experiment which satisfies the following:
  - The experiment consists of a sequence of n smaller experiments, **trials**, where n is fixed in advance of the experiment.
  - Each trial can result in one of the same two possible outcomes (dichotomous trials), denote by success (S) and failure (F).
  - The trials are independent, so that the outcome on any particular trial does not influence the outcome on any other trial.
  - The probability of success $P(S)$ is constant from trial to trial; we denote this probability by p.
- The binomial random variable X associated with a binomial experiment consisting of n trials is defined as $X =$ the number of Ss among the n trials.

Binomial Probability Distribution (2)

- pmf of binomial RV X is $b(x; n, p) = \{$number of sequences of length n consisting of x Succeses$\}$. $\{$probability of any particular such sequence$\}$, i.e.,

$$b(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, \cdots n. \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

- First factor is the number of ways of choosing x of the n trials to be Successes, i.e, number of combinations of size x that can be constructed from n distinct trials.
- $p^x(1-p)^{n-x}$ is probability of x successes . probability of $n-x$ failures.

Binomial Table

Table A.1 Cumulative Binomial Probabilities

$$B(x; n, p) = \sum_{y=0}^{x} b(y; n, p)$$

a. $n = 5$

		p														
		0.01	**0.05**	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.75**	**0.80**	**0.90**	**0.95**	**0.99**
	0	.951	.774	.590	.328	.237	.168	.078	.031	.010	.002	.001	.000	.000	.000	.000
	1	.999	.977	.919	.737	.633	.528	.337	.188	.087	.031	.016	.007	.000	.000	.000
x	2	1.000	.999	.991	.942	.896	.837	.683	.500	.317	.163	.104	.058	.009	.001	.000
	3	1.000	1.000	1.000	.993	.984	.969	.913	.812	.663	.472	.367	.263	.081	.023	.001
	4	1.000	1.000	1.000	1.000	.999	.998	.990	.969	.922	.832	.763	.672	.410	.226	.049

b. $n = 10$

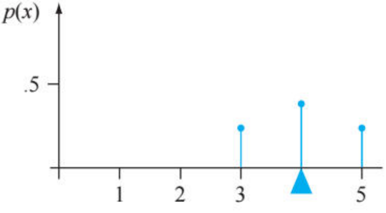
		p														
		0.01	**0.05**	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.75**	**0.80**	**0.90**	**0.95**	**0.99**
	0	.904	.599	.349	.107	.056	.028	.006	.001	.000	.000	.000	.000	.000	.000	.000
	1	.996	.914	.736	.376	.244	.149	.046	.011	.002	.000	.000	.000	.000	.000	.000
	2	1.000	.988	.930	.678	.526	.383	.167	.055	.012	.002	.000	.000	.000	.000	.000
	3	1.000	.999	.987	.879	.776	.650	.382	.172	.055	.011	.004	.001	.000	.000	.000
	4	1.000	1.000	.998	.967	.922	.850	.633	.377	.166	.047	.020	.006	.000	.000	.000
x	5	1.000	1.000	1.000	.994	.980	.953	.834	.623	.367	.150	.078	.033	.002	.000	.000
	6	1.000	1.000	1.000	.999	.996	.989	.945	.828	.618	.350	.224	.121	.013	.001	.000

Using Binomial Table

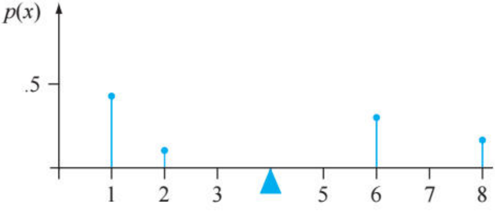
- **Binomial table**: computes binomial probabilities from Binomial table that tabulates cdf $F(X) = P(X \leq x)$ for different n and p values.
- For $X \sim Bin(n, p)$ the cdf is $P(X \leq x) = B(x; n, p) = \sum_{y=0}^{x} b(y; n, p), x = 0, 1, \cdots n$
- **Example :**If 20% of all binding of new book fails binding strength test. Let there by 10 randomly selected books, what is the probability that atmost 5 fail the test?
- X has binomial distribution (Success/ Failure) with $n = 10, p = 0.2$.
- From binomial table see $x = 5, p = 0.2$ column of $n = 10$ table. $B(5, 10, .2) = .994$
- What is the probability that atleast 5 fail the test? $1 - P(X \leq 4) =?$
- What is the probability that between 3 and 5 inclusive fails ? $P(X \leq 5) - P(X \leq 3) =?$

Mean and Variance of Binomial variable X

- if $n = 1$ Binomial distribution becomes Bernoulli distribution.
- Expected value of Bernoulli RV $E(X) = 0.P(X = 0) + 1.P(X = 1) = 0.(1 - p) + 1.p = p = \mu$,as Bernoulli RV.
- $V(X) = E[X^2] - E[X]^2 = p - p^2 = p(1 - p)$
- If $X \sim Bin(n, p)$, $E[X] = \sum_{i=1}^{n} E[X_i] = np$
- If $X \sim Bin(n, p)$, $V[X] = \sum_{i=1}^{n} E[X_i] = n.p(1 - p)$



(a)



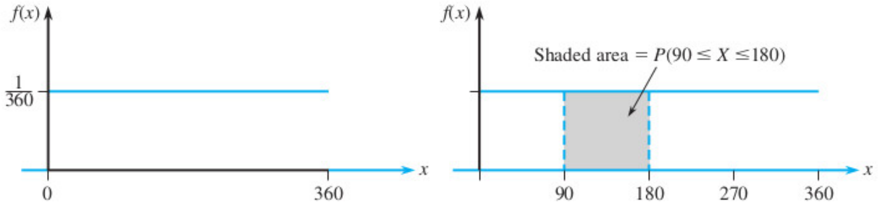
(b)

Geometric Distribution

- Suppose we perform a series of independent Bernoulli trials, each with a probability p of success.
- Let X is number of trials before first success, then
- $P(X = x|p) = (1 - p)^{x-1}p, \text{for} x = 1, 2, \cdots$
- $E[x] = 1/p$, and $Var[x] = (1 - p)/p^2$
- **Example 1**: Suppose you are playing a game of darts. The probability of success is 0.4. What is the probability that you will hit the bullseye on the third try?
- Compute $P[X = 3] = (1 - 0.4)^2.0.4 = 0.144$
- **Example 2**: If a patient is waiting for a suitable blood donor and the probability that the selected donor will be a match is 0.2, then find the expected number of donors who will be tested till a match is found including the matched donor.
- $E(x) = 1/0.2 = 5$

Uniform Distribution

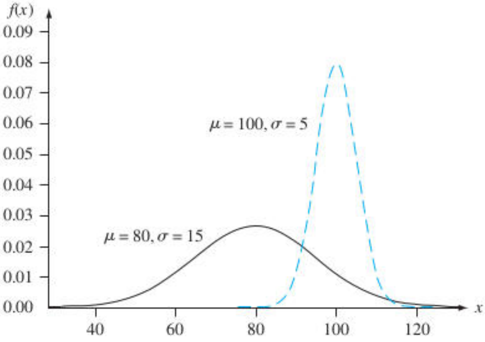
- A continous random variable X is said to be uniformly distributed on an interval $[a, b]$, if its pdf is given as $f_X(x|a, b) = \frac{1}{b-a}$, if $x \in [a, b]$, otherwise, 0.
- $E(X) = (a + b)/2$
- $Var(X) = (b - a)^2/12$



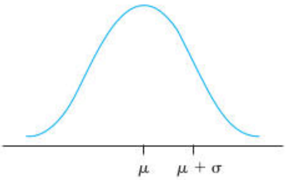
Normal Distribution

- A continuous rv X is said to have a normal distribution with parameters μ and σ (or μ and σ^2), where $-\infty < \mu < \infty$ and $0 < \sigma$, if the PDF of X is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{4}$$



(a)



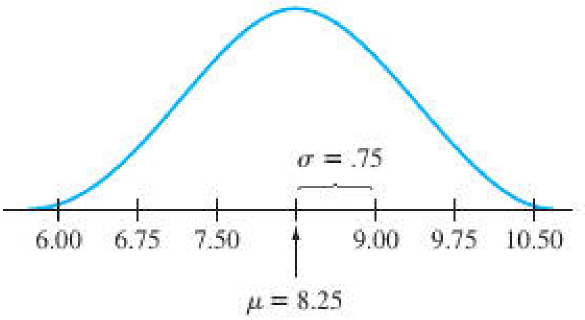
(b)

Statistic

- A statistic is any quantity whose value can be calculated from sample data; mean, standard deviation
- A statistic is a random variable X, x represents an observed/ computed value of X; Mean: $\bar{X}$, S
- The **probability distribution** of a **statistic** is referred to as its **sampling distribution** to emphasize that it describes how the statistic varies in value across all samples that might be selected.
- The random variables $X_1, X_2, \cdots X_n$ are said to form a (simple) random sample of size n (**independent and identically distributed** (iid)) if
  - The X_i's are independent rv's.
  - Every X_i has the same probability distribution.
- Computer simulation is used to obtain information about a statistic's sampling distribution

Simulation of statistic's sampling distribution (1/2)

The **population distribution** for simulation study is **normal** with $\mu = 8.25$ and $\sigma = .75$. For $n = 5, 10, 20, 30$, ie. 4 observations were made generating 500 samples for each n. Sample mean $\bar{x}$ is computed for each sample of n.the results are plotted as histogram.



Simulation of statistic's sampling distribution (2/2)

- For $n = 5, 10, 20, 30$, ie. 4 observations were made generating 500 samples for each n. Sample mean $\bar{x}$ is computed for each sample of n.

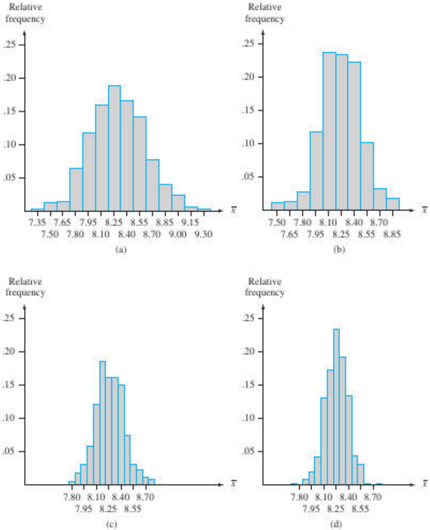
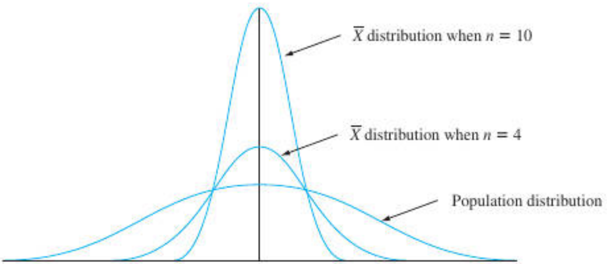


Figure 5.11 Sample histograms for $\bar{x}$ based on 500 samples, each consisting of n observations: (a) $n = 5$; (b) $n = 10$; (c) $n = 20$; (d) $n = 30$

Let $X_1, X_2, \cdots, X_n$ be a random sample from a distribution with mean value μ and standard deviation σ. Then

- $E(\bar{X}) = \mu_{\bar{X}} = \mu$
- $V(\bar{X}) = \sigma^2_{\bar{X}} = \sigma^2/n$ and $\sigma_{\bar{X}} = \sigma/\sqrt{n}$
- In addition, with $T_0 = X_1 + \cdots + X_n$ (the sample total), $E(T_o) = n\mu, V(T_o) = n\sigma^2$, and $\sigma_{T_o} = \sqrt{n}\sigma$

Let $X_1, X_2, \cdots, X_n$ be a random sample from a normal distribution with mean μ and standard deviation σ. Then for any n, X is normally distributed (with mean μ and standard deviation $\sigma/\sqrt{n}$), as is T_o (with mean $n\mu$ and standard deviation $\sqrt{n}\sigma$).



Central Limit Theorem (CLT)

- Let $X_1, X_2, \cdots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then if n is sufficiently large, $\bar{X}$ has approximately a normal distribution with $\mu_{\bar{X}} = \mu$ and $\sigma^2_{\bar{X}} = \sigma^2/n$, and T_o also has approximately a normal distribution with $\mu_{T_o} = n\mu$, $\sigma^2_{T_o} = n\sigma^2$. The larger the value of n, the better the approximation.
- When X_i 's are normally distributed, so is $\bar{X}$ for every sample size n
- For large n a suitable normal curve will approximate the actual distri