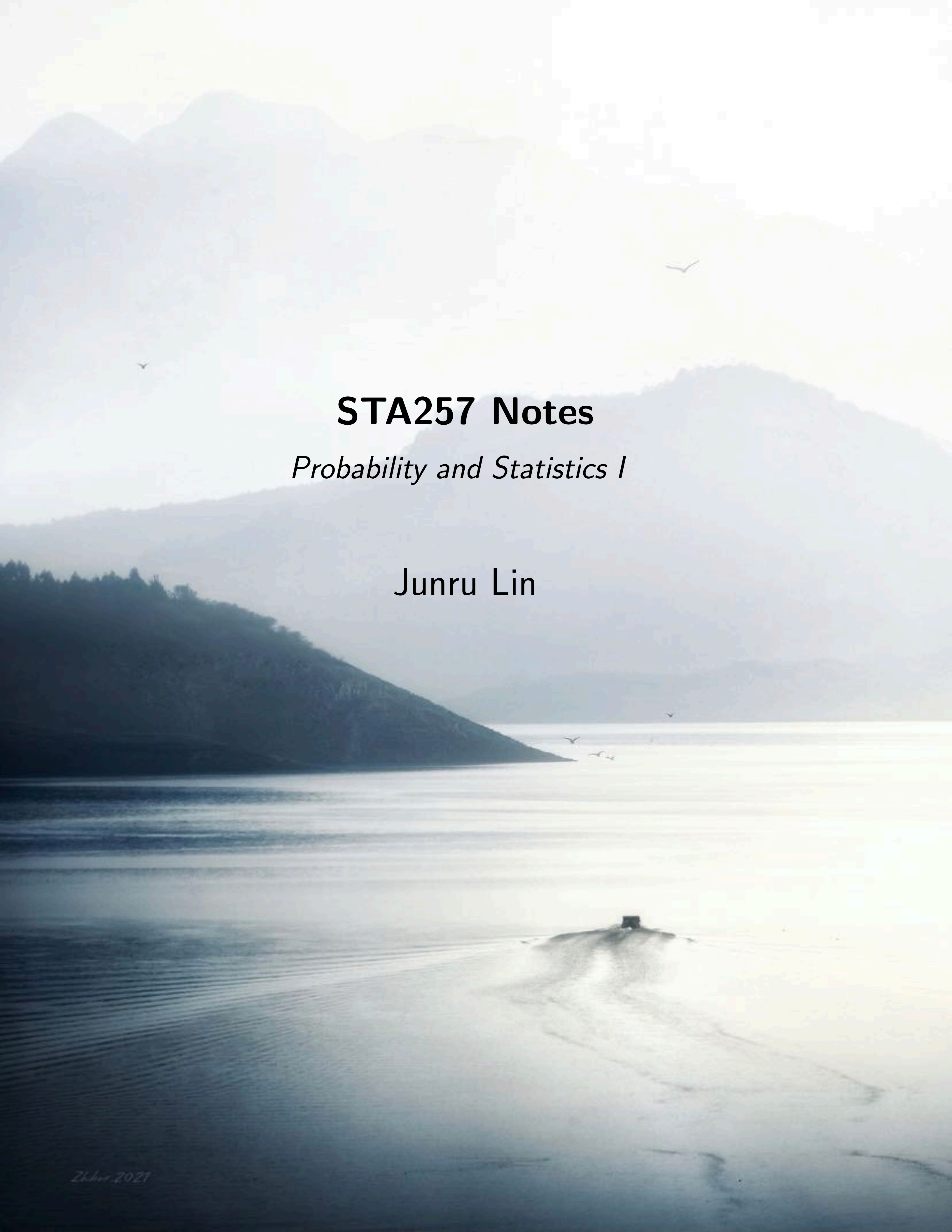

STA257 Notes

Probability and Statistics I

Junru Lin

Contents

Chapter 1

Probability

1.1 Sample Spaces and Events

Sample Space
The **sample space** of an experiment, denoted by S, is the set of all possible outcomes of that experiment.

Event
An **event** is any collection (subset) of outcomes contained in the sample space S. An event is said to be **simple** if it consists of exactly one outcome and **compound** if it consists of more than one outcome.

Complement
The **complement** of an event A, denoted by A', is the set of all outcomes in S that are *not* contained in A.

Intersection
The **intersection** of two events A and B, denoted by $A \cap B$ and read A *and* B, is the event consisting of all outcomes that are in *both* A *and* B.

Union
The **union** of two events A and B, denoted by $A \cup B$ and read A *or* B, is the event consisting of all outcomes that are *either* in A *or* in B or in both events (so that the union includes outcomes for which both A and B occur as well as outcomes for which exactly one occurs) — that is, all outcomes in at least one of the events.

Disjoint / Mutually Exclusive
When A and B have no outcomes in common, they are said to be **disjoint** or **mutually exclusive** events. Mathematicians write this compactly as $A \cap B = \varnothing$, where $\varnothing$ denotes the event consisting of no outcomes whatsoever (the null or empty event)

De Morgan's Laws
Let A and B be two events in the sample space of some experiment. Then

$$(A \cup B)' = A' \cap B'$$

$$(A \cap B)' = A' \cup B'$$

Venn Diagrams
Venn diagrams are often used to visually represent samples spaces and events. To construct a Venn diagram, draw a rectangle whose interior will represent the sample space S. Then any event A is represented as the interior of a closed curve (often a circle) contained in S.

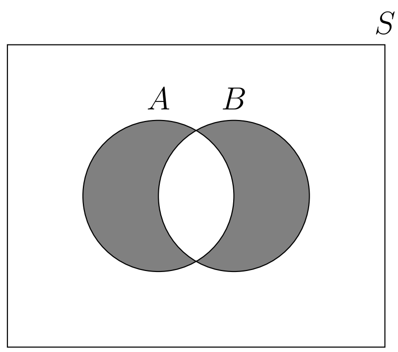

Figure 1.1: Venn Diagram Example

1.2 Axioms, Interpretations, and Properties of Probability

Axiom 1
For any event A, $P(A) \geq 0$.

Axiom 2
$P(S) = 1$.

Axiom 3
If A1, A2, A3, . . . is an infinite collection of disjoint events, then

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$$

Proposition
$P(\varnothing) = 0$, where $\varnothing$ is the null event. This, in turn, implies that the property contained in Axiom 3 is valid for a *finite* collection of disjoint events.

Relative Frequency
The ratio $n(A)/n$ is called the relative frequency of occurrence of the event A in the sequence of n replications.

Complement Rule
For any event A, $P(A) = 1 - P(A')$.

Proposition
For any event A, $P(A) \leq 1$.

Addition Rule
For any event A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Equally Likely Outcomes
In an experiment consist of N outcomes with equal probabilities $p = P(E_i)$ for each i, $p = \frac{1}{N}$.
For an event A containing $N(A)$ outcomes, $P(A) = \frac{N(A)}{N}$.

1.3 Counting Methods

Proposition
If the first element or object of an ordered pair can be selected in n_1 ways, and for each of these n1ways the second element of the pair can be selected in n_2 ways, then the number of pairs is $n_2 n_2$.

k-tuple
An ordered collection of k objects is a k-tuple (so a pair is a 2-tuple and a triple is a 3-tuple)

Fundamental Counting Principle
Suppose a set consists of ordered collections of k elements (k-tuples) and that there are n_1 possible choices for the first element; for each choice of the first element, there are n_2 possible choices of the second element;. . .; for each possible choice of the first $k-1$ elements, there are n_k choices of the kth element. Then there are $n_1 n_2 \cdots n_k$ possible k-tuples.

Tree Diagrams
A tree diagram can be used to represent pictorially all the possibilities.
Starting from a point on the left side of the diagram, for each possible first element of a pair a straight-line segment emanates rightward. Each of these lines is referred to as a first-generation branch. Now for any given first-generation branch we construct another line segment emanating from the tip of the branch for each possible choice of a second element of the pair. Each such line segment is a second-generation branch.

Permutation

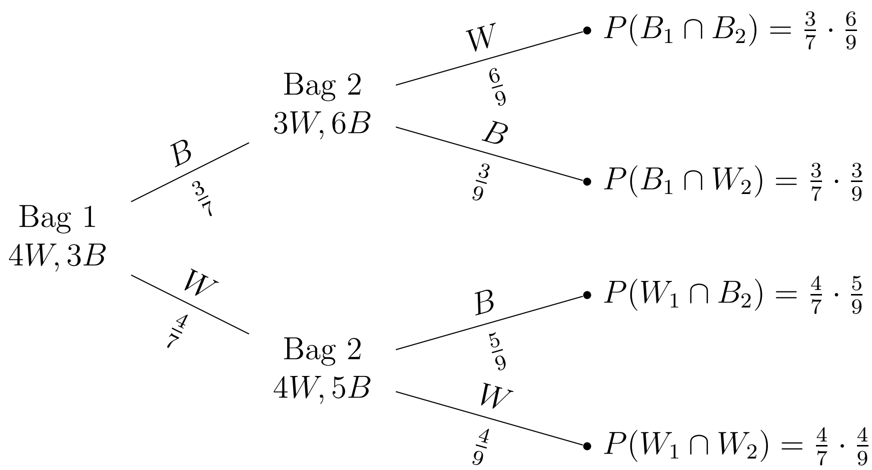

Figure 1.2: Tree Diagram Example

Any ordered sequence of k objects taken without replacement from a set of n distinct objects is called a permutation of size k of the objects. The number of permutations of size k that can be constructed from the n objects is denoted by ${}_nP_k$.

$$
{}_nP_k = n(n-1)(n-2)\cdots(n-k+2)(n-k+1) = \frac{n!}{(n-k)!}
$$

Combination
Given a set of n distinct objects, any unordered subset of size k of the objects is called a combination. The number of combinations of size k that can be formed from n distinct objects will be denoted by $\binom{n}{k}$ or ${}_nC_k$.

$$
{}_nC_k = \binom{n}{k} = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!}
$$

1.4 Conditional Probability

Conditional Probability
For any two events A and B with $P(B) > 0$, the **conditional probability of A given that B has occurred**, denoted $P(A|B)$, is defined by

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

Multiplication Rule
$P(A \cap B) = P(A|B) \cdot P(B)$ and $P(A \cap B) = P(B|A) \cdot P(A)$.

Exhaustive
The events are exhaustive if $A_1 \cup \cdots \cup A_k = S$, so that one A_i must occur.

Law of Total Probability
Let $A_1, \ldots, A_k$ be mutually exclusive and exhaustive events. Then for any other event B,

$$\begin{aligned} P(B) &= P(B|A_1) \cdot P(A_1) + \cdots + P(B|A_k) \cdot P(A_k) \\ &= \sum_{i=1}^{k} P(B|A_i) \cdot P(A_i) \end{aligned}$$

Bayes Theorem
Let $A_1, \ldots, A_k$ be a collection of mutually exclusive and exhaustive events with $P(A_i) > 0$ for $i = 1, \ldots k$. Then for any other event B for which $P(B) > 0$,

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j) \cdot P(A_j)}{\sum_{i=1}^{k} P(B|A_i) \cdot P(A_i)} \qquad j = 1, \ldots, k$$

1.5 Independence

Independent
Two events A and B are independent if $P(A|B) = P(A)$ and are dependent otherwise.

Proposition
If two events are mutually exclusive, they cannot be independent.

Proposition
A and B are independent if and only if $P(A \cap B) = P(A) \cdot P(B)$.

Mutually Independent
Events $A_1, \ldots, A_k$ are mutually independent if for every $k (k = 2, 3, \ldots, n)$ and every subset of indices $i_1, i_2, \ldots, i_k$,

$$P(A_{i1} \cap A_{i2} \cdots \cup A_{ik}) = P(A_{i1}) \cdot P(A_{i2}) \cdots P(A_{ik})$$

Chapter 2

Discrete Random Variables and Probability Distributions

2.1 Random Variables

Random Variable
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. In mathematical language, a random variable is a function whose domain is the sample space and whose range is some subset of real numbers.
Note: Random variables are customarily denoted by uppercase letters, such as X and Y, near the end of our alphabet. We will use lowercase letters to represent some particular value of the corresponding random variable. The notation $X(s) = x$ means that x is the value associated with the outcome s by the rv X.

Bernoulli random variable
Any random variable whose only possible values are 0 and 1 is called a Bernoulli random variable.

Discrete
A discrete random variable is an rv whose possible values constitute either a finite set or a countably infinite set (e.g., the set of all integers, or the set of all positive integers).

Continuous
A random variable is continuous if both of the following apply:

1. Its 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 ∞) or all numbers in a disjoint union of such intervals (e.g., $[0, 10] \cup [20, 30]$).

2. No possible value of the variable has positive probability, that is, $P(X = c) = 0$ for any possible value c.

2.2 Probability Distributions for Discrete Random Variables

Probability Mass Function
The **probability distribution** or **probability mass function** (pmf) of a discrete rv is defined for every number x by

$$p(x) = P(X = x) = P(\text{all } s \in S : X(s) = x)$$

Parameter and Family
Suppose $p(x)$ depends on a quantity that can be assigned any one of a number of possible values, with each different value determining a different probability distribution. Such a quantity is called a **parameter** of the distribution. The collection of all probability distributions for different values of the parameter is called a **family** of probability distributions.

Cumulative Distribution Function
The cumulative distribution function (cdf) F(x) of a discrete rv 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 any number x, F(x) is the probability that the observed value of X will be at most x.

Proposition
For any two numbers a and b with $a \leq b$,

$$P(a \leq X \leq b) = F(b) - F(a-)$$

where "$a-$" represents the largest possible X value that is strictly less than a. In particular, if the only possible values are integers and if a and b are integers, then

$$\begin{aligned} P(a \leq X \leq b) &= P(X = a \text{ or } a + 1 \ldots \text{ or } b) \\ &= F(b) - F(a - 1) \end{aligned}$$

Taking $a = b$ yields $P(X = a) = F(a) - F(a - 1)$ in this case.

2.3 Expected Value and Standard Deviation

2.3.1 Expected Value

Expected Value / Mean Value
Let X be a discrete rv with set of possible values D and pmf $p(x)$. The **expected value** or **mean value** of X, denoted by $E(X)$ or μ_X or just μ, is

$$E(X) = \mu_X = \mu = \sum_{x \in D} x \cdot p(x)$$

Expected Value of a Function (Law of the Unconscious Statistician)
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)] or h(X), is computed by

$$E[h(X)] = \sum_{D} h(x) \cdot p(x)$$

Linearity
For any functions $h_1(X)$ and $h_2(X)$ and any constants a_1, a_2, and b,

$$E[a_1 h_1(X) + a_2 h_2(X) + b] = a_1 E[h_1(X)] + a_2 E[h_2(X)] + b$$

In particular, for any linear function $aX + b$,

$$E(aX + b) = a \cdot E(x) + b$$

(or, using alternative notation, $\mu_{aX+b} = a \cdot \mu_X + b$).

2.3.2 Variance and Standard Deviation

Variance
Let X have pmf $p(x)$ and expected value μ. Then the **variance** of X, denoted by $\mathrm{Var}(X)$ or σ_X^2 or just σ^2, is

$$\mathrm{Var}(X) = \sum_{D} [(x - \mu)^2 \cdot p(x)] = E[(x - \mu)^2]$$

Standard Deviation
The **standard deviation** (SD) of X, denoted by $\mathrm{SD}(X)$ or σ_X or just σ, is

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

Chebyshev's Inequality
Let X be a discrete rv with mean μ and standard deviation σ. Then, for any $k \geq 1$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

That is, the probability X is at least k standard deviations away from its mean is at most $1/k^2$.

Variance Shortcut Formula
$\text{Var}(X) = \sigma^2 = E(X^2) - \mu^2$.

Properties of Variance

$$\text{Var}(aX + b) = \sigma^2_{aX+b} = a^2 \cdot \sigma^2_X \text{ and } \sigma_{aX+b} = |a| \cdot \sigma_X$$

In particular,

$$\sigma_{aX} = |a| \cdot \sigma_X \text{ and } \sigma_{X+b} = \sigma_X$$

2.4 The Binomial Distribution

Binomial Experiment
An experiment for which Conditions 14 are satisfieda fixed number of dichotomous, independent, homogeneous trialsis called a binomial experiment.

1. The experiment consists of a sequence of n smaller experiments called trials, where n is fixed in advance of the experiment.

2. Each trial can result in one of the same two possible outcomes (dichotomous trials), which we denote by success (S) or failure (F).

3. The trials are independent, so that the outcome on any particular trial does not influence the outcome on any other trial.

4. The probability of success is constant from trial to trial (homogeneous trials); we denote this probability by p.

Rule
Consider sampling without replacement from a dichotomous population of size N. If the sample size (number of trials) n is at most 5% of the population size, the experiment can be analyzed as though it were exactly a binomial experiment.

Binomial Random Variable X
Given a binomial experiment consisting of n trials, the binomial random variable X associated with this experiment is defined as

$$X = \text{the number of successes among the } n \text{ trials}$$

Notation
We will write $X \backsim \text{Bin}(n, p)$ to indicate that X is a binomial rv based on n trials with success probability p. Because the pmf of a binomial rv X depends on the two parameters n and p, we denote the pmf by $b(x; n, p)$.

Theorem

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

Notation
For $X \backsim \text{Bin}(n, p)$, the cdf will be denoted by

$$B(x; n, p) = P(X \leq x) = \sum_{y=0}^{x} b(y; n, p) \qquad x = 0, 1, \ldots, n$$

Proposition
If $X \backsim \text{Bin}(n, p)$, then

$$\begin{aligned} E(X) &= np \\ \text{Var}(X) &= np(1-p) = npq \\ \text{SD}(X) &= \sqrt{npq} \end{aligned}$$

where $q = 1 - p$

Binomial Probability Calculations in Matlab and R

Function	pmf	cdf
Notation	$b(x; n, p)$	$B(x; n, p)$
Matlab	binopdf(x, n, p)	binocdf(x, n, p)
R	dbinom(x, n, p)	pbinom(x, n, p)

2.5 The Poisson Distribution

Poisson Distribution
A random variable X is said to have a Poisson distribution with parameter $\mu (\mu > 0)$ if the pmf of X is

$$p(x; \mu) = \frac{e^{-\mu} \mu^x}{x!} \qquad x = 0, 1, 2, \ldots$$

Poisson Distribution as a Limit
Suppose that in the binomial pmf $b(x; n, p)$ we let $n \to \infty$ and $p \to 0$ in such a way that np approaches a value $\mu > 0$. Then $b(x; n, p) \to p(x; \mu)$.

Mean and Variance
If X has a Poisson distribution with parameter μ, then

$$E(X) = \mathrm{Var}(X) = \mu$$

Poisson Process
If the number of events occurring during a fixed time interval of length t has a Poisson distribution with parameter $\mu = \lambda t$, then this process is called a **Poisson Process**, and λ is called the rate of the process.

Poisson Probability Calculations in Matlab and R

Function	pmf	cdf
Notation	$p(x;\mu)$	$P(x;\mu)$
Matlab	poisspdf(x,μ)	poisscdf(x,μ)
R	dpois(x,μ)	ppois(x,μ)

2.6 Hypergeometric Distribution

Assumptions

1. The population or set to be sampled consists of N individuals, objects, or elements (a finite population).
2. Each individual can be characterized as a success (S) or a failure (F), and there are M successes in the population.
3. A sample of n individuals is selected without replacement in such a way that each subset of size n is equally likely to be chosen.

The random variable of interest is X = the number of S's in the sample. The probability distribution of X depends on the parameters n, M, and N, so we wish to obtain the pmf $P(X = x) = h(x; n, M, N)$.

Hypergeometric Distribution
If X is the number of S's in a random sample of size n drawn from a population consisting of M S's and $(N - M)$ F's, then the probability distribution of X, called the hypergeometric distribution, is given by

$$P(X = x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

for x an integer satisfying $\max(0, n - N + M) \le x \le \min(n, M)$.

Mean and Variance
The mean and variance of the hypergeometric rv X having pmf $h(x; n, M, N)$ are

$$E(X) = n \cdot \frac{M}{N} = np$$
$$\text{Var}(X) = \frac{N-n}{N-1} \cdot n \cdot \frac{M}{N}(1 - \frac{M}{N}) = \frac{N-n}{N-1} \cdot np(1-p)$$

where $p = \frac{M}{N}$.

Finite Population Correction Factor
It is defined as $\frac{N-n}{N-1}$ (in the formula of Variance).

2.7 Negative Binomial Distributions

Conditions

1. The experiment consists of a sequence of independent trials.
2. Each trial can result in either a success (S) or a failure (F).
3. The probability of success is constant from trial to trial, so $P(S \text{ on trial } i) = p$ for $i = 1, 2, 3 \ldots$.
4. The experiment continues (trials are performed) until a total of r successes has been observed, where r is a specified positive integer.

Negative Binomial Random Variable
The random variable of interest is X = the number of trials required to achieve the rth success, and X is called a **negative binomial random variable**.

Notation
Let $nb(x; r, p)$ denote the pmf of X. Since trials are independent,

$$nb(x; r, p) = P(X = x) = P(r-1 \; S\text{'s on the first } x-1 \text{ trials }) \cdot P(S)$$

PMF
The pmf of the negative binomial rv X with parameters r = desired number of S's and $p = P(S)$ is

$$nb(x; r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \qquad x = r, r+1, r+2, \ldots$$

Mean and Variance
If X is a negative binomial rv with parameters r and p, then

$$E(X) = \frac{r}{p}$$
$$\mathrm{Var}(X) = \frac{r(1-p)}{p^2}$$

Hypergeometric and Negative Binomial Calculations in Matlab and R

	Hypergeometric	Negative Binomial
Function	pmf	pmf
Notation	$h(x; n, M, N)$	$nb(x; r, p)$
Matlab	hygepdf(x, N, M, n)	nbinpdf($x - r, r, p$)
R	dhyper($x, M, N - M, n$)	dnbino($x - r, r, p$)

2.8 Geometric Distributions

Geometric Distribution
In special case of $r = 1$ in *Negative Binomial Distributions*, the pmf becomes

$$nb(x; 1, p) = (1-p)^{x-1}p \qquad x = 1, 2, \ldots$$

which is called the **geometric distribution**.

Geometric Random Variable
The random variable X = number of trials required to achieve one success is a **geometric random variable**.

2.9 Moments and Moment Generating Functions

2.9.1 Moment

Moment
The **kth moment** of a random variable X is $E(X^k)$, while the **kth moment about the mean** (or **kth central moment**) of X is $E[(X -)^k]$, where $= E(X)$.

2.9.2 Skewness

Skewness Coefficient
The **skewness coefficient** is defined to be

$$\frac{E[(X-\mu)^3]}{\sigma^3} = E[(\frac{X-\mu}{\sigma})^3]$$

Negatively Skewed
When the skewness coefficient is negative, we say that the distribution is **negatively skewed** or that it is **skewed to the left**.

Positively Skewed
When the skewness coefficient is positive, we we say that the distribution is **positively skewed** or that it is **skewed to the right**.

2.9.3 Moment Generating Function

Moment Generating Function
The moment generating function (mgf) of a discrete random variable X is defined to be

$$M_X(t) = E(e^{tX}) = \sum_{x \in D} e^{tX} p(x)$$

where D is the set of possible X values. The moment generating function exists iff $M_X(t)$ is defined for an interval that includes zero as well as positive and negative values of t.

MGF Uniqueness Theorem
If the mgf exists and is the same for two distributions, then the two distributions are identical. That is, the moment generating function uniquely specifies the probability distribution; there is a one-to-one correspondence between distributions and mgfs.

Theorem
If the mgf of X exists, then $E(X^r)$ is finite for all positive integers r, and

$$E(X^r) = M_X^{(r)}(0)$$

Proposition
Let X have the mgf $M_X(t)$ and let $Y = aX + b$. Then $M_Y(t) = e^{bt} M_X(at)$.

2.10 Simulation of Discrete Random Variables

Sample Mean
For a set of numerical values $x_1, \ldots, x_n$, the **sample mean**, denoted by $\bar{x}$, is

$$\bar{x} = \frac{x_1 + \cdots x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample Standard Deviation
The **sample standard deviation** of these numerical values, denoted by s, is

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Estimation
If $x_1, \ldots, x_n$ represent simulated values of a random variable X, then we may estimate the expected value and standard deviation of X by $\hat{\mu} = \bar{x}$ and $\hat{\sigma} = s$, respectively.

Chapter 3

Continuous Random Variables and Probability Distributions

3.1 Probability Density Functions and Cumulative Distribution Functions

3.1.1 Probability Density Functions

Probability Density Function
Let X be a continuous rv. Then a probability distribution or probability density function (pdf) of X is a function $f(x)$ such that for any two numbers a and b with $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

That is, the probability that X takes on a value in the interval $[a, b]$ is the area above this interval and under the graph of the density function. The graph of $f(x)$ is often referred to as the *density curve*.
For f(x) to be a legitimate pdf, it must satisfy the following two conditions:

1. $f(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} f(x)dx =$ [area under the entire graph of $f(x)$] $= 1$

Uniform Distribution
A continuous rv X is said to have a uniform distribution on the interval $[A, B]$ if the pdf of X is

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

Notation
The statement that X has a uniform distribution on $[A, B]$ will be denoted $X \backsim \text{Unif}[A, B]$.

3.1.2 Cumulative Distribution Functions

Cumulative Distribution Function
The cumulative distribution function $F(x)$ for a continuous rv X is defined for every number x by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

For each x, $F(x)$ is the area under the density curve to the left of x.

Proposition
Let X be a continuous rv with pdf $f(x)$ and cdf $F(x)$. Then for any number a,

$$P(X > a) = a - F(a)$$

and for any two numbers a and b with $a < b$,

$$P(a \leq X \leq b) = F(b) - F(a)$$

Proposition
If X is a continuous rv with pdf $f(x)$ and cdf $F(x)$, then at every x at which the derivative $F'(x)$ exists, $F'(x) = f(x)$.

3.1.3 Percentile

Percentile
Let p be a number between 0 and 1. The $(100p)$th percentile of the distribution of a continuous rv X, denoted by η_p, is defined implicitly by the equation

$$p = F(\eta_p) = \int_{-\infty}^{\eta_p} f(y)dy$$

Assuming we can find the inverse of F(x), this can also be written as

$$\eta_p = F^{-1}(p)$$

In particular, the median of a continuous distribution is the 50th percentile, $\eta_{.5}$or $F^{-1}(.5)$. We will occasionally denote the median of a distribution simply as η (i.e., without the .5 subscript).

3.2 Expected Values and Moment Generating Functions

Expected Value
The **expected value** or **mean value** of a continuous rv X with pdf $f(x)$ is

$$\mu = \mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

Law of the Unconscious Statistician
If X is a continuous rv with pdf $f(x)$ and $h(X)$ is any function of X, then

$$\mu_{h(X)} = E[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot f(x)dx$$

Variance
The variance of a continuous random variable X with pdf $f(x)$ and mean value μ is

$$\sigma_X^2 = \mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx = E[(X-\mu)^2]$$

Standard Deviation
The standard deviation of X is

$$\sigma_X = \mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

Proposition
Let X be a continuous rv with pdf $f(x)$, mean μ, and standard deviation σ. Then the following properties hold.

1. (variance shortcut) $\mathrm{Var}(X) = E(X^2) - \mu^2 = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx - (\int_{-\infty}^{\infty} x \cdot f(x)dx)^2$
2. (Chebyshevs inequality) For any constant $k \geq 1$,
$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$
3. (linearity of expectation) For any functions $h_1(X)$ and $h_2(X)$ and any constants a_1, a_2, and b,
$$E[a_1h_1(X) + a_2h_2(X) + b] = a_1E[h_1(X)] + a_2E[h_2(X)] + b$$
4. (rescaling) For any constant a and b,
$$E(aX+b) = a\mu + b \quad \mathrm{Var}(aX+b) = a^2\sigma^2 \quad \sigma_{aX+b} = |a|\sigma$$

Moment Generating Function
The **moment generating function** (mgf) of a continuous random variable X is

$$M_X(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

As in the discrete case, the moment generating function exists iff $M_X(t)$ is defined for an interval that includes zero as well as positive and negative values of t.

3.3 The Normal (Gaussian) Distribution

3.3.1 Normal (Gaussian) Distribution

Normal (Gaussian) Distribution
A continuous rv X is said to have a **normal distribution** (or **Gaussian distribution**) with parameters μ and σ, where $-\infty < \mu < \infty$ and $\sigma > 0$, if the pdf of X is

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

Notation
The statement that X is normally distributed with parameters μ and σ is often abbreviated $X \backsim N(\mu, \sigma)$.

MGF
The moment generating function of a normally distributed random variable X is

$$M_X(t) = e^{\mu t + \sigma^2 t^2/2}$$

3.3.2 Standard Normal Distribution

Standard Normal Distribution
The normal distribution with parameter values $\mu = 0$ and $\sigma = 1$ is called the **standard normal distribution**.

Standard Normal Distribution Variable
A random variable that has a standard normal distribution is called a **standard normal random variable** and will be denoted by Z. The pdf of Z is

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad -\infty < z < \infty$$

The cdf of Z is $P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$, which we will denote by $\Phi(z)$.

Proposition
If $X \backsim N(\mu, \sigma)$, then the standardized rv Z defined by

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

has standard normal distribution. Thus,

$$P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$$

$$P(X \leq a) = \Phi(\frac{a-\mu}{\sigma}) \qquad P(X \geq b) = 1 - \Phi(\frac{a-\mu}{\sigma})$$

and the 100pth percentile of the $N(\mu, \sigma)$ distribution is given by

$$\eta_p = \mu + \Phi^{-1}(p) \cdot \sigma$$

Conversely, if $X \backsim N(0,1)$ and μ and σ are constants (with $\sigma > 0$), then the "un-standardized" rv $X = \mu + \sigma Z$ has a normal distribution with mean μ and standard deviation σ.

Empirical Rule
If the population distribution of a variable is (approximately) normal, then

1. Roughly 68% of the values are within 1 SD of the mean.
2. Roughly 95% of the values are within 2 SDs of the mean.
3. Roughly 99.7% of the values are within 3 SDs of the mean.

Proposition
Let $X \backsim N(\mu, \sigma)$. Then for ant constants a and b with $a \neq 0$, $aX + b$ is also normally distributed. That is, any linear rescaling of a normal rv is normal.

Approximating the Binomial Distribution
Let X be a binomial rv based on n trials with success probability p. Then if the binomial probability histogram is not too skewed, X has approximately a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$. In particular, for $x =$ a possible value of X,

$$\begin{aligned} P(X \leq x) = B(x; n, p) \\ &\approx \text{(area under the normal curve to the left of } x + .5) \\ &= \Phi(\frac{x + .5 - np}{\sqrt{npq}}) \end{aligned}$$

In particular, the approximation is adequate provided that both $np \geq 10$ and $nq \geq 10$.

Normal probability and quantile calculations in Matlab and R

Function	cdf	quantile (the (100p)th percentile)
Notation	$\Phi(\frac{x-\mu}{\sigma})$	$\eta_p = \mu + \Phi^{-1}(p) \cdot \sigma$
Matlab	normcdf(x, μ, σ)	norminv(p, μ, σ)
R	pnorm(x, μ, σ)	qnorm(p, μ, σ)

3.4 The Exponential Distribution

Exponential Distribution
X is said to have an **exponential distribution** with parameter $\lambda(\lambda > 0)$ if the pdf of X is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

CDF
Let X be an exponential variable with parameter λ. Then the cdf of X is

$$F(x;\lambda) = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-\lambda x} & x > 0 \end{cases}$$

Mean and Standard Deviation
The mean and standard deviation of X are both equal to $1/\lambda$.

3.5 The Gamma Distribution

3.5.1 Gamma Function

Gamma Function
For $\alpha > 0$, the **gamma function** $\Gamma(\alpha)$ is defined by

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

Propositions

1. For any $\alpha > 1$,
$$\Gamma(\alpha) = (\alpha - 1) \cdot \Gamma(\alpha - 1)$$
(via integration by parts)

2. For ant positive integer n,
$$\Gamma(n) = (n-1)!$$

3. $$\Gamma(\frac{1}{2}) = \sqrt{\pi}$$

4. For any $\alpha, \beta > 0$,
$$\int_0^\infty x^{\alpha-1} e^{-x/\beta} dx = \beta^\alpha \Gamma(\alpha)$$

3.5.2 Gamma Distribution

Gamma Distribution

A continuous random variable X is said to have a **gamma distribution** if the pdf of X is

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

where the parameters α and β satisfy $\alpha>0,\beta>0$.

Mean and Variance

The mean and variance of a gamma random variable are

$$E(X)=\mu=\alpha\beta$$
$$\mathrm{Var}(X)=\sigma^2=\alpha\beta^2$$

CDF

Let X have a gamma distribution with parameters α and β. Then for any $x>0$, the cdf of X is given by

$$P(X\le x)=G(\frac{x}{\beta};\alpha)$$

the incomplete gamma function evaluated at α/β.

MGF

The moment generating function of a gamma random variable is

$$M_X(t)=\frac{1}{(1-\beta t)^\alpha}\qquad t<1/\beta$$

3.5.3 Other Distributions

Standard Gamma Distribution

When $\beta=1$, X is said to have a **standard gamma distribution**, and its pdf may be denoted $f(x;\alpha)$.

Erlang Distribution

In the special case where the shape parameter α is a positive integer, n, the gamma distribution is sometimes rewritten with the substitution $\lambda=1/\beta$, and the resulting pdf is

$$f(x;n,1/\lambda)=\frac{\lambda^n}{(n-1)!}x^{n-1}e^{-\lambda x},\qquad x>0$$

Incomplete Gamma Function

When X is a standard gamma rv, the cdf of X, which for $x>0$ is

$$G(x;\alpha)=P(X\le x)=\int_0^x\frac{1}{\Gamma(\alpha)}y^{\alpha-1}e^{-y}dy$$

is called the **incomplete gamma function**.

Gamma and Exponential Calculations in Matlab and R

	Gamma	Exponential
Function	cdf	cdf
Notation	$G(x/\beta;\alpha)$	$F(x;\lambda) = 1 - e^{-\lambda x}$
Matlab	gamcdf(x,α,β)	expcdf$(x,1/\lambda)$
R	pgamma$(x,\alpha,1/\beta)$	pexp(x,λ)

3.6 The Weibull Distribution

Weibull Distribution
A random variable X is said to have a **Weibull distribution** with parameters α and β $(\alpha > 0, \beta > 0)$ if the pdf of X is

$$f(x;\alpha,\beta) = \begin{cases} \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Mean and Variance

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

CDF

$$F(x;\alpha,\beta) = \begin{cases} 0 & x < 0 \\ 1 - e^{-(x/\beta)^\alpha} & x \geq 0 \end{cases}$$

3.7 The Lognormal Distribution

Lognormal Distribution
A nonnegative rv X is said to have a **lognormal distribution** if the rv $Y = \ln(X)$ has a normal distribution. The resulting pdf of a lognormal rv when $\ln(X)$ is normally distributed with parameters μ and γ is

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

Mean and Variance

$$E(X) = e^{\mu+\sigma^2/2}$$

$$\mathrm{Var}(X) = e^{2\mu+\sigma^2} \cdot \left(e^{\sigma^2} - 1\right)$$

CDF

$$\begin{aligned} F(x;\mu,\sigma) &= P(X \le x) \\ &= P[\ln(X) \le \ln(x)] \\ &= P\left[\frac{\ln(X)-\mu}{\sigma} \le \frac{\ln(x)-\mu}{\sigma}\right] \\ &= P\left[Z \le \frac{\ln(x)-\mu}{\sigma}\right] \\ &= \Phi\left[\frac{\ln(x)-\mu}{\sigma}\right] \end{aligned}$$

3.8 The Chi-Squared Distribution

Chi-Squared Distribution

Let v be a positive integer. Then a random variable X is said to have a **chi-squared distribution** with parameter v if the pdf of X is the gamma density with $\alpha = v/2$ and $\beta = 2$. The pdf of a shi-squared rv is thus

$$f(x;v) = \begin{cases} \frac{1}{2^{v/2}\Gamma(v/2)} x^{(v/2)-1} e^{-x/2} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

The parameter v is called the number of degrees of freedom (df) of X. The symbol χ often used in place of chi-squared.

Mean and Variance

$$\mu = \alpha\beta = (\frac{v}{2}) \cdot 2 = v$$

$$\sigma^2 = \alpha\beta^2 = (\frac{v}{2}) \cdot 4 = 2v$$

MGF

$$M_X(t) = (1-2t)^{-v/2}, \qquad t < \frac{1}{2}$$

3.9 The Beta Distribution

Beta Distribution
A random variable X is said to have a beta distribution with parameters α, β (both positive), A, and B if the pdf of X is

$$f(x;\alpha,\beta,A,B) = \begin{cases} \frac{1}{B-A} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(B)} \left(\frac{x-A}{B-A}\right)^{\alpha-1} \left(\frac{B-x}{B-A}\right)^{\beta-1} & A \leq x \leq B \\ 0 & \text{otherwise} \end{cases}$$

Mean and Variance

$$\mu = A + (B-A) \cdot \frac{\alpha}{\alpha+\beta}$$
$$\sigma^2 = \frac{(B-A)^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Standard Beta Distribution
$A = 0$, $B = 1$.

Chapter 4

Joint Probability Distributions and Their Applications

4.1 Jointly Distributed Random Variables

4.1.1 Probability Mass Function

Joint Probability Mass Function
Let X and Y be two discrete rvs defined on the sample space S of an experiment. The **joint probability mass function** $p(x, y)$ is defined for each pair of numbers (x, y) by

$$p(x, y) = P(X = x \text{ and } Y = y)$$

A function $p(x, y)$ can be used as a joint pmf provided that $p(x, y) \geq 0$ for all x and y and $\sum_x \sum_y p(x, y) = 1$.

Generalization
If $X_1, X_2, \ldots X_n$ are all discrete random variables, the **joint pmf** of the variables is the function

$$p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1 \cap X_2 = x_2 \cap \ldots \cap X_n = x_n)$$

Marginal Probability Mass Functions
The **marginal probability mass functions** of X and of Y, denoted by $p_X(x)$ and $p_Y(y)$, respectively, are given by

$$p_X(x) = \sum_y p(x, y)$$

$$p_Y(y) = \sum_x p(x, y)$$

4.1.2 Probability Density Function

Joint Probability Density Function
Let X and Y be continuous rvs. Then $f(x, y)$ is the **joint probability density function** for X and Y if for any two-dimensional set A,

$$P((X, Y) \in A) = \iint_A f(x, y)dxdy$$

In particular, if A is the two-dimensional rectangle $\{(x, y) : a \leq x \leq b, c \leq y \leq d\}$, then

$$((X, Y) \in A) = P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y)dydx$$

Generalization
If the variables are continuous, the **joint pdf** of $X_1, X_2, \ldots X_n$ is the function $f(x_1, x_2, \ldots, x_n)$ such that for any n intervals $[a_1, b_1], \ldots, [a_n, b_n]$,

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, x_2, \ldots, x_n)dx_n \ldots dx_1$$

Marginal Probability Density Functions
The **marginal probability density functions** of X and Y, denoted by $f_X(x)$ and $f_Y(y)$, respectively, are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy \qquad \text{for} -\infty < x < \infty$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx \qquad \text{for} -\infty < y < \infty$$

4.1.3 Independent and Dependent

Independent
Two random variables X and Y are said to be **independent** if for every pair of x and y values,

$$p(x, y) = p_X(x) \cdot p_Y(y) \text{ when } X \text{ and } Y \text{ are discrete}$$

or

$$f(x, y) = f_X(x) \cdot f_Y(y) \text{when } X \text{ and } Y \text{ are continuous}$$

Dependent
If the equation above is not satisfied for all (x, y), then X and Y are said to be **dependent**.

4.2 Expected Values, Covariance, and Correlation

4.2.1 Expected Values

Law of the Unconscious Statistician
Let X and Y be jointly distributed rvs with pmf $p(x,y)$ or pdf $f(x,y)$ according to whether the variables are discrete or continuous. Then the expected value of a function $h(X,Y)$, denoted by $E[h(X,Y)]$ or $\mu_{h(X,Y)}$, is given by

$$E[h(X,Y)] = \begin{cases} \sum_x \sum_y h(x,y) \cdot p(x,y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) \cdot f(x,y) dx dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

Linearity of Expectation
Let X and Y be random variables. Then, for any functions $h_1, h2$ and any constants a_1, a_2, b,

$$E[a_1 h_1(X,Y) + a_2 h_2(X,Y) + b] = a_1 E[h_1(X,Y)] + a_2 E[h_2(X,Y)] + b$$

Theorem
Let X and Y be independent random variables. If $h(X,Y) = g_1(X) \cdot g_2(Y)$, then

$$E[h(X,Y)] = E[g_1(X) \cdot g_2(Y)] = E[g_1(X)] \cdot E[g_2(Y)]$$

4.2.2 Covariance

Covariance
The **covariance** between two rvs X and Y is

$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &\begin{cases} \sum_x \sum_y (x-\mu_X)(y-\mu_Y) p(x,y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y) dx dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases} \end{aligned}$$

Proposition
For any two random variables X and Y,

1. $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
2. $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$

3. (Covariance shortcut formula) $\text{Cov}(X, Y) = E(XY) - \mu_X \cdot \mu_Y$
4. (Distributive property of covariance) For any rv Z and any constants, a, b, c,
$$\text{Cov}(aX + bY + c, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$$

4.2.3 Correlation

Correlation
The **correlation coefficient** of X and Y, denoted by $\text{Corr}(X, Y)$, or $\rho_{X,Y}$, or just ρ, is defined by
$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Proposition
For any two rvs X and Y,

1. $\text{Corr}(X, Y) = \text{Corr}(Y, X)$
2. $\text{Corr}(X, X) = 1$
3. (Scale invariance property) If a, b, c, d are constants and $ac > 0$,
$$\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$$
4. $-1 \leq \text{Corr}(X, Y) \leq 1$
5. If X and Y are independent, then $\rho = 0$, but $\rho = 0$ does not imply independence.
6. $\rho = 0$ or -1 iff $Y = aX + b$ for some numbers a and b with $a \neq 0$.

Uncorrelated
When $\rho = 0$, X and Y are said to be **uncorrelated**.

Proposition
Two rvs X and Y are uncorrelated if, and only if, $E[XY] = \mu_X \cdot \mu_Y$.

4.3 Properties of Linear Combinations

4.3.1 Expected Value and Variance

Theorem
Let the rvs $X_1, X_2, ..., X_n$ have mean values $\mu_1, ..., \mu_n$ and standard deviations $\sigma_1, ..., \sigma_n$, respectively.

1. Whether or not the X_is are independent,

$$\begin{aligned} E(a_1X_1 + \cdots + a_nX_n + b) &= a_1E(X_1) + \cdots + a_nE(X_n) + b \\ &= a_1\mu_1 + \cdots + a_n\mu_n + b \end{aligned}$$

   and

$$\begin{aligned} \text{Var}(a_1X_1 + \cdots + a_nX_n + b) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} a_i^2\sigma_i^2 + 2\sum\sum_{i<j} a_ia_j\text{Cov}(X_i, X_j) \end{aligned}$$

2. If $X_1, X_2, ..., X_n$ are independent,

$$\begin{aligned} \text{Var}(a_1X_1 + \cdots + a_nX_n + b) &= a_1^2\text{Var}(X_1) + \cdots + a_n^2\text{Var}(X_n) \\ &= a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2 \end{aligned}$$

   and

$$\text{SD}(a_1X_1 + \cdots + a_nX_n + b) = \sqrt{a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2}$$

Corollary
For any two rvs X_1 and X_2, and any constants a_1, a_2, b,

$$E(a_1X_1 + a_2X_2 + b) = a_1E(X_1) + a_2E(X_2) + b$$

and

$$\text{Var}(a_1X_1 + a_2X_2 + b) = a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2) + 2a_1a_2\text{Cov}(X_1, X_2)$$

In particular,

$$\begin{aligned} E(X_1 + X_2) &= E(X_1) + E(X_2) \\ E(X_1 - X_2) &= E(X_1) - E(X_2) \end{aligned}$$

If X_1 and X_2 are independent,

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$$

$$\text{Var}(X_1 - X_2) = \text{Var}(X_1) + \text{Var}(X_2)$$

4.3.2 Probability Distribution Function of a Sum

Theorem
Suppose X and Y are independent, continuous rvs with marginal pdfs $f_X(x)$ and $f_Y(y)$, respectively. Then the pdf of the rv $W = X + Y$ is given by

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x)f_Y(w - x)dx$$

4.3.3 Moment Generating Functions for Linear Combinations

Proposition
Let $X_1, X_2, \ldots, X_n$ be independent random variables with moment generating functions $M_{X_1}(t), M_{X_2}(t), \ldots, M_{X_n}(t)$, respectively. Then the moment generating function of the linear combination $Y = a_1X_1 + a_2X_2 + \cdots a_nX_n + b$ is

$$M_Y(t) = e^{bt} M_{a_1X_1}(t) \cdot M_{a_2X_2}(t) \cdot \cdots \cdot M_{X_n}(a_nt)$$

In the special case that $a_1 = a_2 = \cdots = a_n = 1$ and $b = 0$, so $Y = X_1 + \cdots + X_n$,

$$M_Y(t) = M_{X_1}(t) \cdot M_{X_2}(t) \cdot \cdots \cdot M_{X_n}(t)$$

That is, the mgf of a sum of independent rvs is the *product* of the individual mgfs.

4.3.4 Linera Combinations of different distributions

Normal Distribution
If $X_1, X_2, ..., X_n$ are independent, normally distributed rvs (with possibly different means and/or sds), then any linear combination of the X_is also has a normal distribution. In particular, the sum of independent normally distributed rvs itself has a normal distribution, and the difference $X_1 - X_2$ between two independent, normally distributed variables is itself normally distributed.

Poisson Distribution
Suppose $X_1, X_2, ..., X_n$ are independent Poisson random variables, where X_i has mean μ_i. Then $Y = X_1 + \cdots + X_n$ also has a Poisson distribution, with mean $\mu_1 + \cdots + \mu_n$.

Exponential Distribution
Suppose $X_1, X_2, ..., X_n$ are independent exponential random variables with common parameter λ. Then $Y = X_1 + \cdots + X_n$ has a gamma distribution, with parameters $\alpha = n$ and $\beta = 1/\lambda$ (aka the Erlang distribution).
Note: This proposition requires the X_i to have the same rate parameter λ, i.e., the X_is must be independent and identically distributed.

4.4 Conditional Distributions and Conditional Expectation

4.4.1 Conditional Distributions and Independence

Conditional Probability Mass Function
Let X and Y be two discrete random variables with joint pmf $p(x, y)$ and marginal X pmf

$p_X(x)$. Then for any x value such that $p_X(x) > 0$, the **conditional probability mass function of Y given $X = x$** is

$$p_{Y|X}(y|x) = \frac{p(x,y)}{p_X(x)}$$

Conditional Probability Density Function

Let X and Y be two continuous random variables with joint pdf $f(x,y)$ and marginal X pdf $f_X(x)$. Then for any x value such that $f_X(x) > 0$, the **conditional probability density function of Y given $X = x$** is

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

Independency

In the discrete case above, X and Y are independent iff

$$p_{Y|X}(y|x) = \frac{p(x,y)}{p_X(x)} = \frac{p_X(x)p_Y(y)}{p_X(x)} = p_Y(y)$$

4.4.2 Conditional Expectation and Variance

Conditional Expectation

Let X and Y be two discrete random variables with conditional probability mass function $p_{Y|X}(y|x)$. Then the **conditional expectation** (or **conditional mean**) of Y given $X = x$ is

$$\mu_{Y|X=x} = E(Y|X = x) = \sum_y y \cdot p_{Y|X}(y|x)$$

Analogously, for two continuous rvs X and Y with conditional probability density function $f_{Y|X}(y|x)$,

$$\mu_{Y|X=x} = E(Y|X = x) = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y|x)dy$$

More generally, the conditional mean of any function $h(Y)$ is given by

$$E(h(Y)|X = x) = \begin{cases} \displaystyle\sum_y h(y) \cdot p_{Y|X}(y|x) & \text{(discrete case)} \\ \displaystyle\int_{-\infty}^{\infty} h(y) \cdot f_{Y|X}(y|x)dy & \text{(continuous case)} \end{cases}$$

Conditional Variance

The **conditional variance of Y given $X = x$** is

$$\begin{aligned} \sigma^2_{Y|X=x} &= \mathrm{Var}(Y|X = x) \\ &= E[(Y - \mu_{Y|X=x})^2|X = x] \\ &= E(Y^2|X = x) - \mu^2_{Y|X=x} \end{aligned}$$

4.4.3 The Laws of Total Expectation and Variance

Law of Total Expectation
For any two random variables X and Y,

$$E[E(Y|X)] = E(Y)$$

(This is sometimes referred to as computing $E(Y)$ by means of *iterated expectation*.)

Law of Total Variance
For any two random variables X and Y,

$$\mathrm{Var}(Y) = \mathrm{Var}[E(Y|X)] + E[\mathrm{Var}(Y|X)]$$

4.5 Limit Theorems (What Happens as n Gets Large)

4.5.1 Random Sample

Random Sample
The rvs $X_1, X_2, ..., X_n$ are said to be **independent and identically distributed (iid)** if

1. The X_is are independent rvs.
2. Every X_i has the same probability distribution.

Such a collection of rvs is also called a (simple) **random sample** of size n.

4.5.2 Sample Total

Sample Total

$$T = X_1 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

Proposition
Suppose $X_1, X_2, ..., X_n$ are idd with common mean μ and common standard deviation σ. T has the following properties:

1. $E(T) = n\mu$
2. $\mathrm{Var}(T) = n\sigma^2$ and $\mathrm{SD}(T) = \sqrt{n}\sigma$
3. If the X_is are normally distributed, then T is also normally distributed.

4.5.3 Sample Mean

Sample Mean

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{T}{n}$$

Proposition
Suppose $X_1, X_2, ..., X_n$ are idd with common mean μ and common standard deviation σ. $\bar{X}$ has the following properties:

1. $E(\bar{X}) = \mu$
2. $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$ and $\mathrm{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
3. If the X_is are normally distributed, then $\bar{X}$ is also normally distributed.

4.5.4 The Central Limit Theorem

Central Limit Theorem
Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with mean μ and standard deviation σ. Then, in the limit as $n \to \infty$, the standardized versions of T and $\bar{X}$ have the standard normal distribution. That is,

$$\lim_{n\to\infty} P\left(\frac{T - n\mu}{\sqrt{n}\sigma} \leq z\right) = P(Z \leq z) = \Phi(z)$$

and

$$\lim_{n\to\infty} P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) = P(Z \leq z) = \Phi(z)$$

where Z is a standard normal rv.
It is customary to say that T and $\bar{X}$ are **asymptotically normal**.
Thus when n is sufficiently large, the sample total T has approximately a normal distribution with mean $\mu_T = n\mu$ and standard deviation $\sigma_T = \sqrt{n}\sigma$
Equivalently, for large n the sample mean $\bar{X}$ has approximately a normal distribution with mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$.

4.5.5 Applications of the Central Limit Theorem

Corollary
Consider an event A in the sample space of some experiment with $p = P(A)$. Let $X =$ the number of times A occurs when the experiment is repeated n independent times, and define

$$\hat{P} = \hat{P}(A) = \frac{X}{n}$$

Then

1. $\mu_{\hat{P}} = E(\hat{P}) = p$
2. $\sigma_{\hat{P}} = \mathrm{SD}(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$
3. As n increases, the distribution of $\hat{P}$ approaches a normal distribution. In practice, it is taken to say that $\hat{P}$ is approximately normal, provided that $np \geq 10$ and $n(1-p) \geq 10$.

Normal Distributions
CLT justifies normal approximations to the following distributions:

1. Poisson, when μ is large
2. Negative binomial, when r is large
3. Gamma, when α is large

Lognormal Distribution
Let $X_1, X_2, ..., X_n$ be a random sample from a distribution for which only positive values are possible $[P(X_i > 0) = 1]$. Then if n is sufficiently large, the product $Y = X_1\ X_2\ \cdots X_n$ has approximately a lognormal distribution; that is, $\ln(Y)$ has approximately a normal distribution.

4.5.6 The Law of Large Numbers

Law of Large Numbers
If $X_1, X_2, ..., X_n$ is a random sample from a distribution with mean μ and finite variance, then $\bar{X}$ converges to μ

1. In mean square: $E\left[(\bar{X} - \mu)^2\right] \to 0$ as $n \to \infty$
2. In probability: $P(|\bar{X} - \mu| \geq \varepsilon) \to 0$ as $n \to \infty$ for any ε.

4.6 Transformations of Jointly Distributed Random Variables

Transformation Theorem ((Bivariate Case)
Suppose that the partial derivative of each $v_i(y_1, y_2)$ with respect to both y_1 and y_2 exists

and is continuous for every $(y_1, y_2) \in T$. Form the 2×2 matrix

$$M = \begin{bmatrix} \frac{\partial v_1(y_1,y_2)}{\partial y_1} & \frac{\partial v_1(y_1,y_2)}{\partial y_2} \\ \frac{\partial v_2(y_1,y_2)}{\partial y_1} & \frac{\partial v_2(y_1,y_2)}{\partial y_2} \end{bmatrix}$$

The determinant of this matrix, called the *Jacobian*, is

$$\det(M) = \frac{\partial v_1}{\partial y_1} \cdot \frac{\partial v_2}{\partial y_2} - \frac{\partial v_1}{\partial y_2} \cdot \frac{\partial v_2}{\partial y_1}$$

The joint pdf for the new variables is

$$g(y_1, y_2) = f(v_1(y_1, y_2), v_2(y_1, y_2)) \cdot |\det(M)|$$