Statistics and Probability Notes

Devansh Shah

May 2022

Contents

Chapter 1

Basic Concepts on Probability

1.1 Introduction

Definition 1.1.1 (Observation)**.** *Any recording of information whether it is numerical or categorical.*

Definition 1.1.2 (Experiment)**.** *Any procedure that generates observations.*

Definition 1.1.3 (Sample Space)**.** *The set of all possible outcomes of an experiment. It is denoted by the symbol S.*

Definition 1.1.4 (Sample Points)**.** *Every outcome in a sample space is called an element of the sample space or simply a sample point.*

Definition 1.1.5 (Event)**.** *An event is a subset of a sample space.*

Definition 1.1.6 (Simple Event)**.** *An event is said to be simple if it consists of exactly one sample point*

Definition 1.1.7 (Compound Event)**.** *An event it said to be compound if it consists of more than one sample point*

Note 1.1.8.

The sample space is itself an event and is usually called a sure event.

Note 1.1.9.

A subset of S that contains no elements at all is the empty set, denoted by ϕ, and is usually called the null event.

1.2 Operations with Events

Definition 1.2.1 (Union)**.** *The union of two events A and B, denoted by $A \cup B$, is the event containing all the elements that belong to A or B or to both. Mathematically,*

$$A \cup B = \{x : x \in A \; or \; x \in B\}$$

Definition 1.2.2 (Intersection)**.** *The intersection of two events A and B, denoted by $A \cap B$, is the event containing all elements that are common to A and B. Mathematically,*

$$A \cap B = \{x : x \in A \; and \; x \in B\}$$

Definition 1.2.3 (Complement)**.** *The complement of event A with respect to the sample space S, denoted by $A^{'}$ or A^{C}, is the set of all elements of S that are not in A. Mathematically,*

$$A^{'} = \{x : x \in S \; and \; x \notin A\}$$

Definition 1.2.4 (Mutually Exclusive Events)**.** *Events A and B are mutually exclusive or mutually disjoint if $A \cap B = \phi$, i.e, if A and B have no elements in common*

Theorem 1.2.5. *Some basic properties of operations events are as follows:*

1. $A \cap A^{'} = \phi$
2. $A \cap \phi = \phi$
3. $A \cup A^{'} = S$
4. $(A^{'})^{'} = A$
5. $(A \cup B)^{'} = A^{'} \cap B^{'}$ *(De Morgan's Law)*
6. $(A \cap B)^{'} = A^{'} \cup B^{'}$ *(De Morgan's Law)*
7. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ *(Distributive Law)*
8. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ *(Distributive Law)*
9. $A \cup B = A \cup (B \cap A^{'})$
10. $A = (A \cap B) \cup (A \cap B^{'})$
11. $A \cup B = (A \cap B^{'}) \cup (A \cap B) \cup (A^{'} \cap B)$

The last property is useful to express a union of two sets as the union of three disjoint events.

Note 1.2.6 (De Morgan's Laws)**.**

For n events $A_1, A_2, ..., A_n$,

$$(A_1 \cup A_2 \cup \cdots \cup A_n)^{'} = A_1^{'} \cap A_2^{'} \cap \cdots \cap A_n^{'}$$

$$(A_1 \cap A_2 \cap ... \cap A_n)^{'} = A_1^{'} \cup A_2^{'} \cup \cdots \cup A_n^{'}$$

Definition 1.2.7 (Contained)**.** *If all the elements in event A are also in event B, then event A is contained in event B, denoted by $A \subset B$.*

Note 1.2.8.

$A = B \iff A \subset B \; and \; B \subset A$

1.3 Counting Methods

Theorem 1.3.1. *Multiplication Principle*

If an operation can be performed in n_1 ways and if for each of these ways, a second operation can be performed in n_2 ways, and for each of the first two ways, a third operation can be performed in n_3 ways, and so forth, then the sequence of k operations can be performed in $n_1 n_2 n_3 \dots n_k$ ways

Theorem 1.3.2. *Addition Principle*

If there are k procedures and the i^{th} procedure may be performed in n_i ways where i = 1,2,...,k, then the number of ways in which we may perform procedure 1 or procedure 2 or ... or procedure k is given by $n_1 + n_2 + ... + n_k$ assuming that no two procedures may be performed together.

Definition 1.3.3 (Permutation)**.** *An arrangement of r objects from a set of n objects, where r $\leq$ n in which order matters. It is denoted by ${}^nP_r$.*

Note 1.3.4.

The number of permutations of n distinct objects taken r at a time is given by

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

In particular, one can arrange n distinct objects in n! ways.

Note 1.3.5.

If there are n_1 objects of the first kind, n_2 objects of the second kind, ..., n_k objects of the k^{th} kind, then the number of permutations of all these objects is given by

$$ {}^nP_{n_1,n_2,\dots,n_k} = \frac{n!}{n_1! n_2! \dots n_k!} $$

Note 1.3.6 (Circular Permutations)**.**

The number of permutations of n distinct objects arranged in a circle is $(n-1)!$. This is because the rotation of the circle does not affect the relative positions of the objects and since there are n possible rotationally symmetric states of the circle, the total permutations is $\frac{n!}{n}$, which is equal to $(n-1)!$.

Definition 1.3.7 (Combination)**.** *A selection of r objects from a set of n objects, where r $\leq$ n, in which the order of objects does not matter. It is denoted by ${}_nC_r$ or ${}^nC_r$ or $\binom{n}{r}$.*

Note 1.3.8.

The number of combinations of n distinct objects taken r at a time is given by

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

Note 1.3.9 (Multiset Problem)**.**

The number of ways to select r objects from n different types of objects (where all objects of the same type are indistinguishable) is given by $\binom{n+r-1}{r}$ or equivalently, $\binom{n+r-1}{n-1}$. Here, since repetition is allowed, it is possible than r $>$ n. We use the stars and bars method to solve such problems.

Example: The number of solutions to the equation $x_1 + x_2 + x_3 = 5$ where x_1, x_2, x_3 are non-negative integers is equal to $\binom{5+3-1}{3-1}$ = $\binom{7}{2}$ = 21

1.4 Relative Frequency and Probability

Definition 1.4.1 (Relative Frequency)**.** *Let E be an experiment and let A be an event associated with E. Suppose we repeat the experiment n times and the event A occurs n_A times. Then $f_A = \frac{n_A}{n}$ is called the relative frequency of the event A in the n repetitions of E.*

Theorem 1.4.2. *Properties of relative frequency:*

1. $0 \leq f_A \leq 1$
2. f_A = 1 *if, and only if, A occurs every time among the n repetitions.*
3. f_A = 0 *if, and only if, A never occurs among the n repetitions.*
4. *If A and B are mutually exclusive events and if $f_{A\cup B}$ is the relative frequency associated with the event $A \cup B$, then $f_{A\cup B} = f_A + f_B$.*
5. *As the experiment is repeated more and more times, the value of f_A approaches a stable value. In particular, if n is the number of times the experiment is repeated,*

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

That is, the relative frequency of occurrence of an event approaches its theoretical probability as the experiment is repeated a large number of times.

Theorem 1.4.3. *Axioms of Probability*

Consider an experiment whose sample space is S, and let A be an event associated with the experiment. Then,

1. $0 \leq P(A) \leq 1$
2. $P(S) = 1$
3. *If $A_1, A_2, \ldots$ are mutually exclusive events then,*

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i)$$

4. $P(\phi) = 0$
5. $P(A') = 1 - P(A)$
6. *If* $A \subset B$, *then* $P(A) \leq P(B)$

Theorem 1.4.4. *Inclusion-Exclusion Principle*

$$P(A_1 \cup A_2 \cup ... \cup A_n) = \sum_{i=1}^{n} P(A_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P(A_i \cap A_j) + \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap ... \cap A_n)$$

In particular,

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
2. $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C)$

Note 1.4.5 (Some Interesting Problems)**.**

Click on the links below:

1. The Birthday Problem
2. The Airplane Problem
3. Boy or Girl Paradox
4. Another Boy-Girl Paradox

Note 1.4.6.

Do not assume that each of the outcomes is equally likely in all cases. For example, consider the case of buying a lottery ticket. There are two possible scenarios - either you win the lottery, or you don't. But this does not mean that $P(win) = P(lose) = \frac{1}{2}$.

1.5 Conditional Probability

Definition 1.5.1 (Conditional Probability)**.** *Let A and B be two events associated with an experiment E. We denote the conditional probability of the event A, given that the event B has already occurred by* $P(A|B)$.

Note 1.5.2 (Formula for Conditional Probability)**.**

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ where } P(B) \neq 0$$

Theorem 1.5.3. *Multiplication Rule of Probability*

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1) \dots P(A_n|A_{n-1} \cap A_{n-2} \cap \cdots \cap A_1),$$

provided that $P(A_1 \cap \cdots \cap A_{n-1}) > 0$

Theorem 1.5.4. *Law of Total Probability*
Let $E_1, E_2, \dots, E_n$ *be a partition of the sample space S. In other words,* $E_1, E_2, \dots, E_n$ *are mutually exclusive and exhaustive events such that* $E_i \cap E_j = \phi$ *for all* $i \neq j$ *and* $\bigcup_{i=1}^{n} E_i = S$*. Then for any event A,*

$$P(A) = \sum_{i=1}^{n} P(A \cap E_i) = \sum_{i=1}^{n} P(E_i)P(A|E_i)$$

Note 1.5.5.

Recall that in set theory, a partition of a set A is a set of mutually disjoint non-empty subsets of A whose union is A. In other words, B is a partition of A if B is a set whose elements are non-empty subsets of A, and every element of A is in exactly one element of B.

Theorem 1.5.6. *Bayes' Theorem*
Combining the formula for conditional probability and the law of total probability, we come up with a powerful way to find the probability of a cause of an event, given that the event occurred. This is called Bayes' Theorem.
Let $A_1, A_2, \dots, A_n$ *be a partition of the sample space S. Then,*

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$$

Note 1.5.7 (Some interesting links for Bayes' Theorem)**.**

1. The Monty Hall Problem
2. Bayes' Theorem Explained
3. The Medical Test Paradox
4. Example from Thinking Fast and Slow
5. Base Rate Neglect Problem

1.6 Independent Events

In general, knowing that an event A has occurred gives a different view on the chance of an event B's occurrence. But, sometimes the probability occurrence of two events do not depend on each other. In other words, events A and B are said to be independent if the occurrence (or non-occurrence) of one event does not in any way influence the occurrence (or non-occurrence) of the other event.

Definition 1.6.1 (Independent Events)**.** *Mathematically, two events A and B are said to be independent if, and only if,* $P(A \cap B) = P(A)P(B)$

Alternatively, if A and B are independent events, then $P(A|B) = P(A)$ and $P(B|A) = P(B)$

Note 1.6.2 (Properties of Independent Events)**.**

Suppose $P(A) > 0, P(B) > 0$.

1. If A and B are independent events, then they cannot be mutually exclusive.
2. If A and B are mutually exclusive, then they cannot be independent.
3. The sample space S as well as the empty set ϕ are independent of any event.
4. If $A \subset B$, then A and B are dependent unless $B = S$

However, (1) and (2) does not imply that any two events with non-zero probability must be either independent or mutually exclusive.

Note 1.6.3.

The properties of independence, unlike the mutually exclusive property, cannot be shown on a Venn diagram. In general, it is not always easy to deduce whether two events are independent without finding their probabilities. Moreover, there is no relation between mutually exclusive events and independent events other than the fact that two events with non-zero probabilities cannot be both mutually exclusive and independent simultaneously.

Theorem 1.6.4. *If A and B are independent events, then so are the following:*

1. A *and* B'
2. A' *and* B
3. A' *and* B'

The above theorem is true because the occurrence or non-occurrence of A and B should not have any affect on the occurrence of non-occurrence of the other, by the definition of independent events.

Definition 1.6.5 (Pairwise independent Events)**.** *A set of events* $A_1, A_2, ..., A_n$ *are said to be pairwise independent if, and only if,*

$$P(A_i \cap A_j) = P(A_i)P(A_j) \qquad \textit{for all } i \neq j \textit{ and } i, j = 1, 2, ..., n$$

Definition 1.6.6 (Mutually Independent Events)**.** *The events* $A_1, A_2, ..., A_n$ *are said to be mutually independent (or simply independent) events if, and only if, for any subset* $\{A_{i_1}, A_{i_2}, \cdots, A_{i_k}\}$ *of* $\{A_1, A_2, \cdots, A_n\}$*,*

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_k}) \qquad \textit{for all distinct i's}$$

Note 1.6.7 (Remarks on Independent Events)**.**

1. There are a total of $2^n - n - 1$ different cases in case of n mutually independent events, and only $\binom{n}{2}$ different cases in case of n pairwise independent events. This is because, for mutually independent events, you can select any number of events (between 2 and n) at a time, while you can only select 2 events at a time for pairwise independent events. (Recall that $\binom{n}{0} + \binom{n}{1} + ... + \binom{n}{n} = 2^n$. So, $\binom{n}{2} + \binom{n}{3} + ... + \binom{n}{n} = 2^n - n - 1$)
2. Mutual independence implies pairwise independence, but pairwise independence does not necessarily imply mutually independence (except in the case where there are only two events, in which case, both mean exactly the same thing). This shows that mutual independence is a stronger condition than pairwise independence.
3. Suppose the events $A_1, A_2, \cdots, A_n$ are mutually independent. Let

$$B_i = A_i \text{ or } A_i^{'}, \text{ where } i = 1, 2, \cdots, n$$

Then, $B_1, B_2, \cdots, B_n$ are also mutually independent events.

Note 1.6.8 (Terminology)**.** *There are some important terminologies that you should be familiar with:*

1. *Sensitivity* = $P(+|\textit{disease})$ *(also called Recall in Machine Learning)*
2. *Specificity* = $P(-|\textit{no disease})$
3. *Prevalence* = $P(\textit{disease})$
4. *Positive Predicted Value* = $P(\textit{disease}|+)$ *(also called Precision in Machine Learning)*

Apart from independence, there's another property of events known as **conditional independence**. Two events, A and B, are said to be conditionally independent given another event C if $P(A \cap B|C) = P(A|C)P(B|C)$. Equivalently, $P(A|B \cap C) = P(A|C)$ and $P(B|A \cap C) = P(B|C)$.

Note that independence does not imply conditional independence or vice versa!

Chapter 2

Random Variables

2.1 Introduction

Definition 2.1.1 (Random Variable)**.** *Let S be a sample space associated with the experiment E. A function X, which assigns a number to every element $s \in S$ is called a random variable.*

Note 2.1.2.

1. X is a real-valued function, ie, the range of X is a subset of $\mathbb{R}$.
2. The range of X is the set of real numbers, $R_X = \{x \mid x = X(s), s \in S\}$. Each possible value x of X represents an event that is a subset of the sample space S.
3. If S has elements that are themselves real numbers, we take $X(s) = s$. In this case, $R_X = S$.

Definition 2.1.3 (Equivalent Events)**.** *Let E be an experiment and S its sample space. Let X be a random variable defined on S and R_X be its range space. That is, $X : S \to \mathbb{R}$. Let B be an event with respect to R_X, i.e., $B \subset R_X$ Suppose that A is defined as $A = \{s \in S | X(s) \in B\}$. In other words, A consists of all the sample points, s, in S for which $X(s) \in B$. In this case we say that A and B are equivalent events and $P(A) = P(B)$.*

2.2 Discrete Probability Distributions

Definition 2.2.1 (Discrete Random Variable)**.** *Let X be a random variable. If the number of possible values of X (i.e, the range space) is finite or countable infinite, we call X a discrete random variable. That is, the possible values of X may be listed as $x_1, x_2, x_3, \cdots$.*

Definition 2.2.2 (Probability Function and Distribution)**.** *For a discrete random variable, each value of X has a certain probability $f(x)$. Such a function $f(x)$ is called the probability function (p.f.) or probability mass function (p.m.f.). The collections of pairs $(x_i, f(x_i))$ is called the probability distribution of X.*

Note 2.2.3.

The probability of $X = x_i$ denoted by $f(x_i)$ (i.e., $f(x_i) = P(X = x_i)$) must satisfy the following two conditions.

1. $\forall x_i \quad f(x_i) \geq 0$
2. $\sum_{i=1}^{\infty} f(x_i) = 1$

2.3 Continuous Probability Distributions

Definition 2.3.1 (Continuous Random Variable)**.** *Suppose that R_X, the range space of a random variable, X, is an interval or a collection of intervals. Then we say that X is a continuous random variable.*

Definition 2.3.2 (Probability Density Function)**.** *Let X be a continuous random variable. The probability density function (p.d.f.) $f(x)$ is a function satisfying the following conditions:*

1. $\forall x \in R_X \quad f(x) \geq 0$
2. $\int_{R_X} f(x)dx = 1$ *or equivalently,* $\int_{-\infty}^{\infty} f(x)dx = 1$ *since* $f(x) = 0$ *for* $x \notin R_X$
3. *For any* c *and* d *such that* $c < d$ *and* $(c, d) \subset R_X$, $P(c \leq X \leq d) = \int_c^d f(x)dx$

Note 2.3.3.

1. $P(c \leq X \leq d) = \int_c^d f(x)dx$ represents the area under the graph of the p.d.f. $f(x)$ between $x = c$ and $x = d$.
2. For any specified value of X, say x_0,
$$P(X = x_0) = \int_{x_0}^{x_0} f(x)dx = 0$$
Hence in case of a continuous probability distribution, the probability that X equals to a fixed value is always 0 and $P(c \leq X \leq d) = P(c \leq X < d) = P(c < X \leq d) = P(c < X < d)$. Therefore, whenever dealing with continuous distributions, $<$ and $\leq$ can be used interchangeably.
3. $A = \phi \implies P(A) = 0$ but the converse is not necessarily true, i.e., $P(A) = 0 \not\implies A = \phi$. For example, A can be a non-empty set whose values all lie outside the sample space. Consider the event of obtaining a number in $A = \{7\}$ when a single die is thrown. Clearly, $P(A) = 0$ but $A \neq \phi$
4. If X assumes values only in some interval $[a, b]$, we may simply set $f(x) = 0$ for all X outside the interval.
5. If X is a continuous random variable, and $f(x)$ is the probability density function, then it is possible that $f(x) > 1$ for some values of x. This is because $f(x) \neq P(X = x)$ (In fact, $P(X = x) = 0$ for all values of x in case of continuous distribution). $f(x)$ gives the probability **density** (and not the actual probability). In other words,
$$f(x) = \lim_{\delta x \to 0} \frac{P(x \leq X \leq x + \delta x)}{\delta x}$$
6. The probability density function of any distribution can never be negative. This can be proven as follows:
Assume that the pdf is negative over the interval (a, b) (the pdf cannot be negative simply at a single point because it must be continuous). Then, $P(a \leq X \leq b) = \int_a^b f(x)dx$. Since the pdf is negative, the area under the pdf curve is below the x-axis and hence the integral will be negative. But probability can never be negative. Hence, pdf can also never be negative. Moreover, another definition of pdf is that it is the derivative of the cumulative distribution function. We know that the CDF is always non-decreasing and so the derivative cannot be negative. (Recall that the derivative of a function gives the slope or gradient of the function)

2.4 Cumulative Distributive Function

Definition 2.4.1 (Cumulative Distributive Function)**.** *Let X be a random variable - discrete or continuous. We define $F(x)$ to be the cumulative distribution function (c.d.f.) of the random variable X where*

$$F(x) = P(X \leq x)$$

2.4.1 CDF for Discrete Random Variables

If X is a discrete random variable, then

$$F(x) = \sum_{t \leq x} f(t) = \sum_{t \leq x} P(X = t)$$

The c.d.f of a discrete random variable is a step function.
For any two numbers a and b with $a \leq b$,

$$\begin{aligned} P(a \leq X \leq b) &= P(X \leq b) - P(X < a) \\ &= F(b) - F(a^-) \end{aligned}$$

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

$$P(a \leq X \leq b) = P(X = a \text{ or } a+1 \text{ or } \ldots \text{or } b)$$

Also, $P(a \leq X \leq b) = F(b) - F(a-1)$.
When $b = a, P(X = a) = F(a) - F(a-1)$.

2.4.2 CDF for Continuous Random Variables

If X is a continuous random variable, then

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

For a continuous random variable X,

$$f(x) = \frac{d}{dx}F(x)$$

provided the derivative exists.
Also, $P(a \leq X \leq b) = P(a < X \leq b) = F(b) - F(a)$.

Note 2.4.2.

1. $F(x)$ is a non-decreasing function, i.e., $x_1 < x_2 \implies F(x_1) \leq F(x_2)$
2. $0 \leq F(x) \leq 1$

2.5 Mean and Variance of a Random Variable

Definition 2.5.1 (Expected value)**.**

1. *If X is a discrete random variable taking on values $x_1, x_2, \cdots$ with probability function $f_X(x)$, then the mean or **expected value** or (mathematical expectation) of X, denoted by $E(X)$ (or $\mathbb{E}[X]$) as well as by μ_X is defined by*

$$\mu_X = E(X) = \sum_i x_i f_X(x_i) = \sum_x x f_X(x)$$

2. *If X is a continuous random variable with probability density function $f_X(x)$, the mean of X is defined by*

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

Note 2.5.2.

1. The expected value is not necessarily a possible value of the random variable X
2. In other words, the expected value is the weighted average of all the possible values of the random variable (weighted by their probabilities)
3. In the discrete case, if $f_X(x) = \frac{1}{N}$ for each of the N values of x, then the mean, $E(X) = \sum_i x_i f(x_i) = \frac{1}{N}\sum_i x_i$ becomes the average of the N items.

Definition 2.5.3 (Expectation of a Function of a Random Variable)**.** *For any function $g(X)$ of a random variable X with p.f. (or p.d.f) $f_X(x)$,*

- $E[g(X)] = \sum_x g(x) f_X(x)$ *if X is a discrete Random Variable and provided the sum exists*
- $E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$

Note 2.5.4.

Linearity of Expectation

A very important property of expectation is that it is distributive over linear functions of random variables, i.e., $E[X + Y] = E[X] + E[Y]$ where X, Y are random variables. In general, $E[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = E[a_1X_1] + E[a_2X_2] + \cdots + E[a_nX_n] = a_1E[X_1] + a_2E[X_2] + \ldots a_nE[X_n]$ where the $X_i's$ are random variables. This is always true! (no constrains/requirements on the random variables)

A special case of expectation arises when $g(x) = (x - \mu_X)^2$. This leads to the definition of variance of a random variable X.

Definition 2.5.5 (Variance)**.** *Let X be a random variable with p.f. (or p.d.f.) $f(x)$, then the variance of X is defined as*

$$\begin{aligned}\sigma_X^2 = V(X) &= E[(X - E(X))^2] = E[(X - \mu_X)^2] \\ &= \begin{cases} \sum_x (x - \mu_X)^2 f_X(x) & \text{, if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx & \text{, if } X \text{ is continuous} \end{cases}\end{aligned}$$

Variance is the expectation of the squared deviation of a random variable from its expected value. Variance is a measure of dispersion, meaning it is a measure of how far a set of numbers is spread out from their average value.

Note 2.5.6 (Properties of Variance of R.V.)**.**

- $V(X) \geq 0$ since it is equal to the sum of the squares of numbers (which must be non-negative).
- $V(X) = E(X^2) - [E(X)]^2$. Then, we can write

$$V(X) = \sum_i x_i^2 f_X(x) - \left(\sum_i x_i f_X(x)\right)^2$$

  and similarly for the continuous case.
- The positive square root of the variance is called the **standard deviation** of X, i.e., $\sigma_X = \sqrt{V(X)}$. We often use standard deviation instead of variance since the unit of standard deviation is the same as that of the random variable, but the unit of variance is the square of the unit of the random variable.
- If $V(X) > 0$, then $\forall x \in \mathbb{R}\ P(X = x) < 1$. This is because, the variance is zero if, and only if, the random variable is actually a constant, i.e., $\exists c\ P(X = c) = 1$. (Proof by contraposition)
- If the variance of X is c, then the variance of $X + b$ is c. In other words, if each of the possible values a random variable can assume is increased/decreased by a constant, the variance remains unaffected.
- If the variance of X is c, then the variance of bX is b^2c. In other words, if each of the possible values a random variable can assume is multiplied by a constant, the variance is multiplied by the square of that constant.
- Combining the above 2 properties, $V(aX + b) = a^2V(X)$

Note 2.5.7 (Properties of Expectation)**.**

1. Expectation gives us the "population mean" (or intuitively, the central location of the possible values) of X. Therefore, $E(X)$ is also called the location parameter in some literature.
2. The expected value of a constant is the constant itself.
3. $g(x) = x^k$. Then $E(X^k)$ is called the **k-th moment of X**. $E[(X - \mu_X)^k]$ is called the **k-th central moment of X** or the **k-th moment about the mean** (Since μ_X gives a measure of centre of a distribution).
4. $E(X - \mu_X) = 0$ (Because $E(X) = \mu_X = E(\mu_X)$ since μ_X is a constant). Therefore, the first central moment is always 0.
5. $E[(X - E(X))^3]$ measures the degree of symmetry of the distribution. If it is close to zero, the distribution is more symmetric.
6. $E(X^2) \geq (E(X))^2$, and the inequality holds trivially if, and only if, X is a constant (i.e., non-random). More specifically, $\exists c\ P(X = c) = 1$. In such a case, $V(X) = 0$ (i.e., there is no variability for X, or in other words, X is not actually a "random" variable)

7. $E(aX+b) = aE(X)+b$ where a and b are arbitrary constants. We prove the property for the discrete case:

$$\begin{aligned}
E(aX+b) &= \sum_x (ax+b) f_X(x) \\
&= \sum_x ax f_X(x) + \sum_x b f_X(x) \\
&= a\left(\sum_x x f_X(x)\right) + b\left(\sum_x f_X(x)\right) \text{ Note: } \sum_x f_X(x) = 1 \text{ as it is the probability function} \\
&= aE(X) + b
\end{aligned}$$

8. $V(X) = E(X^2) - [E(X)]^2$. The proof is as follows:

$$\begin{aligned}
V(X) &= E[(X-\mu_X)^2] \\
&= E[X^2 - 2X\mu_X + \mu_X^2] \\
&= E(X^2) - 2\mu_X E(X) + E(\mu_X^2) \text{ (Note that } \mu_X = E(X) \text{ is a constant)} \\
&= E(X^2) - 2\mu_X^2 + \mu_X^2 \\
&= E(X^2) - \mu_X^2 \\
&= E(X^2) - [E(X)]^2
\end{aligned}$$

9. $V(aX+b) = a^2V(X)$ where a and b are arbitrary constants. The proof is as follows:

$$\begin{aligned}
V(aX+b) &= E[(aX+b)^2] - [E(aX+b)]^2 \\
&= E(a^2X^2 + 2abX + b^2) - (a\mu_X + b)^2 \\
&= a^2E(X^2) + 2abE(X) + b^2 - (a^2\mu_X^2 + 2ab\mu_X + b^2) \\
&= a^2E(X^2) = a^2\mu_X^2 \text{ (Because } E(X) = \mu_X\text{)} \\
&= a^2[E(X^2) - (E(X))^2] \\
&= a^2V(X) \text{ (By property 2)}
\end{aligned}$$

10. In general, $E[g(X)] \neq g(E(X))$. The equality only holds in the case of linear combinations of X.

Note 2.5.8.

It is important to remember that if 2 random variables X and Y are equal, then they follow the same distribution. But if two random variables follow the exact same distribution, it does not mean that they are equal. A distribution only gives us information about how a random variable behaves and the probability associated which each value being assumed by the random variable. For example, if X refers to the number of heads when I toss a coin, and Y refers to the number of heads when you toss a coin, both X and Y have the same distribution. But if I get a head, it does not mean that you get a head. Recall that we say that $X = Y \iff \forall s \in S\ X(s) = Y(s)$ (the value of the random variable is same for all the sample points).
The above definition of equality of random variables is actually far more general and can be applied to functions too (since a random variable is just a function). Two functions f and g are said to be equal, i.e., $f = g$, if, and only if,

1. Domain of f = Domain of g = A
2. Codomain of f = Codomain of g = B
3. $\forall x \in A\ f(x) = g(x)$. That is, for every possible input, they give the same output.

2.6 Chebyshev's Inequality

If we know the probability distribution of some random variable X, then we can easily compute $E(X)$ and $V(X)$. However, the converse is not true, i.e., it is not possible to reconstruct the probability distribution of X given $E(X)$ and $V(X)$. Hence, we cannot compute quantities such as $P(|X - E(X)| \leq c)$ for some positive constant c.
Nonetheless, Russian Mathematician Pafnuty Chebyshev gave a very useful upper (or lower) bound to such a probability. The result is called Chebyshev's inequality.

Theorem 2.6.1. *Chebyshev's Inequality*
Let X be a random variable (discrete or continuous) with $E(X) = \mu$ and $V(X) = \sigma^2$. The for any positive number k,

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

In other words, the probability that the value of X lies at least k standard deviations away from its mean is at most $\frac{1}{k^2}$, Alternatively,

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

since the two events are complementary to each other.

Note 2.6.2.

- $|x| \leq k \iff -k \leq x \leq k$
- The value of k in Chebyshev's Inequality can be any positive number
- This inequality is true for all distributions with finite mean and variance
- The theorem gives a lower bound on the probability of $|X - \mu| < k\sigma$. There is no guarantee that this lower bound is close to the exact probability.

A slightly more convenient form of Chebyshev's Inequality is given as follows:

$$P(|X - \mu| \geq c) \leq \frac{V(X)}{c^2}$$
$$P(|X - \mu| < c) \geq 1 - \frac{V(X)}{c^2}$$

for any positive constant c. Note that this is exactly the same as the formula given in the theorem. We can obtain these formulae by setting $c = k\sigma$ and using $k^2 = \frac{c^2}{\sigma^2} = \frac{c^2}{V(X)}$.

When we need to find $P(a < X < b)$ when (a, b) is not symmetric about μ, we can still use Chebyshev's inequality by changing the interval and the inequalities accordingly.
For example, if we want to calculate $P(-2\sigma < X - \mu < 2.5\sigma)$, we may do it as follows:

$$\begin{aligned} P(-2\sigma < X - \mu < 2.5\sigma) &\geq P(-2\sigma < X - \mu < 2\sigma) \\ &= P(|X - \mu| < 2\sigma) \\ &\geq 1 - \frac{1}{2^2} = \frac{3}{4} \end{aligned}$$

Important: The chain of inequality must be the same throughout the argument. Otherwise, the equation gives no information whatsoever. For example, the following would not make much sense (although the equation is perfectly correct, we cannot draw any conclusion regarding what we actually want to calculate):

$$\begin{aligned} P(-2\sigma < X - \mu < 2.5\sigma) &\leq P(-2.5\sigma < X - \mu < 2.5\sigma) \\ &= P(|X - \mu| < 2.5\sigma) \\ &\geq 1 - \frac{1}{2.5^2} = 0.84 \end{aligned}$$

Some examples of Chebyshev's Inequality:

- $P(X \geq \mu + 1) \leq Var(X)$. The proof is as follows:

$$\begin{aligned} P(X - \mu \geq 1) &\leq P(|X - \mu| \geq 1) \\ &= P(|X - \mu| \geq k\sigma) \qquad \text{for } k = \frac{1}{\sigma} \\ &\leq \frac{1}{k^2} = \sigma^2 = Var(X) \end{aligned}$$

- If $\mu = 0$, then $\forall k > 0, k \in \mathbb{Z}, \quad P(X^{2k} \geq 1) \leq Var(X)$. The proof is as follows:

$$\begin{aligned} X^{2k} \geq 1 \iff X^2 \geq 1^{\frac{1}{k}} = 1 \iff |X| \geq 1 \\ \therefore P(X^{2k} \geq 1) &= P(|X| \geq 1) \\ &= P(|X - \mu| \geq 1) \qquad \text{where } \mu = 0 \\ &\leq Var(X) \qquad \text{from our first example} \end{aligned}$$

Chapter 3

2-Dimensional Random Variables and Conditional Probability Distributions

3.1 Introduction

Definition 3.1.1 (2-D Random Variable)**.** *Let E be an experiment and S be a sample space associated with E. Let X and Y be two functions each assigning a real number to each $s \in S$. Then, we call (X, Y) a two dimensional random variable or a random vector.*

Definition 3.1.2 (Range Space)**.** $R_{X,Y} = \{(x, y) \mid x = X(s), y = Y(s), s \in S\}$

Definition 3.1.3 (n-Dimensional Random Vector)**.** *Let $X_1, X_2, ..., X_n$ be n functions each assigning a real number to every outcome $s \in S$. We call $(X_1, X_2, ..., X_n)$ an n-dimensional random variable or an n-dimensional random vector*

Definition 3.1.4 (Discrete 2-D Random Variable)**.** *(X, Y) is a 2-D discrete random variable if the possible values of $(X(s), Y(s))$ are finite or countable infinite. That is, the possible values of $(X(s), Y(s))$ may be represented as*

$$(x_i, y_j), \text{ for } i = 1, 2, 3, \ldots; j = 1, 2, 3, \ldots$$

Definition 3.1.5 (Continuous 2-D Random Variable)**.** *(X, Y) is a 2-D continuous random variable if the possible values of $(X(s), Y(s))$ can assume all values in some region of the Euclidean plane $\mathbb{R}^2$*

3.2 Joint Probability Density Function

3.2.1 Joint Probability Function for Discrete RVs

Let (X, Y) be a 2-D Discrete random variable defined on the sample space of an experiment. With each possible value (x_i, y_j), we associate a number $f_{X,Y}(x_i, y_j)$ representing $P(X = x_i, Y = y_j)$ and satisfying the following conditions:

1. $f_{X,Y}(x_i, y_j) \geq 0 \quad \forall (x_i, y_j) \in R_{X,Y}$
2. $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f_{X,Y}(x_i, y_j) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} P(X = x_i, Y = y_j) = 1$

The second condition can be rewritten simply as the summation over all $f(x_i, y_i) > 0$ equals to 1. That is,

$$\sum_{(x_i, y_i): f_{X,Y}(x_i, y_i) > 0} f_{X,Y}(x_i, y_i) = 1$$

The function $f_{X,Y}(x_i, y_j)$ defined for all pairs of values $(x_i, y_j) \in R_{X,Y}$ is called the joint probability function of (X, Y).
Let A be any set consisting of pairs of (x, y) values. Then the probability $P((X, Y) \in A)$ is defined by the summation of the joint probability function over all the pairs in A. That is,

$$P((X,Y) \in A) = \sum_{(x,y)\in A} f_{X,Y}(x,y)$$

3.2.2 Joint Probability Density Function for Continuous RVs

Let (X, Y) be a 2-D continuous random variable assuming all values in some region R of the Euclidean plane $\mathbb{R}^2$. Then, $f_{X,Y}(x, y)$ is called a joint probability density function if it satisfies the following 2 conditions:

1. $f_{X,Y}(x,y) \geq 0 \quad \forall (x,y) \in R_{X,Y}$
2. $$\iint_{(x,y)\in R_{X,Y}} f_{X,Y}(x,y)\ dx\ dy = 1$$

   or alternatively,

   $$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\ dx\ dy = 1$$

Moreover, if the Cumulative Density function of a distribution with 2 random variables is $F(x, y)$ where $F(x, y) = P(X \leq x, Y \leq y)$, then we can obtain the joint probability density function by taking the derivative of $F(x, y)$. However, since there are two variables, we can take the partial derivative with respect to x and y successively. In other words,

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) \\ &= \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y} F_{X,Y}(x,y)\right) \\ &= \frac{\partial}{\partial y}\left(\frac{\partial}{\partial x} F_{X,Y}(x,y)\right) \end{aligned}$$

We know by Clairaut's theorem (mixed derivative theorem) that the order of mixed partial derivatives does not matter.

It is also necessary to remember that the joint probability density function does not give the actual probability - it only indicates the density of probability in that region. So, it is possible (and allowed) for the value of $f_{X,Y}(x, y)$ to be greater than 1. (The only restriction on $f_{X,Y}(x, y)$ is that it should always be non-negative and its integral over the 2-D plane must be equal to 1.)

3.3 Marginal and Conditional Probability Distributions

Definition 3.3.1 (Marginal Probability Distribution)**.** *Let (X, Y) be a 2-D random variable with joint probability function (joint probability density function in case of continuous RV) $f_{X,Y}(x, y)$. The marginal probability distributions of X and Y are respectively given by:*

- *For discrete RV,*

$$f_X(x) = \sum_y f_{X,Y}(x,y) \qquad \textit{and} \qquad f_Y(y) = \sum_x f_{X,Y}(x,y)$$

- *For continuous RV,*

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\ dy \qquad \textit{and} \qquad f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\ dx$$

Note 3.3.2.

The practical interpretation of the marginal distribution of X is as follows: we are focusing on viewing the distribution of X by ignoring the presence of Y. Note that

- $f_X(x)$ should not involve y
- $f_X(x)$ is a pdf/pmf and so it must have all the properties of a pdf/pmf

If (X, Y) is discrete, then the marginals are also discrete. Similarly, if (X, Y) is continuous, the marginals must also be continuous. Note that it is possible for X to be discrete and Y to be continuous but we do not consider such cases here.

Note 3.3.3.

Note that $f(x, y)$ not only tells us how X and Y behave independently (through the marginal), but it also tells us how X and Y behave jointly and how they affect each other. So, you can derive the marginal from the joint distribution but you cannot derive the joint distribution given the 2 marginal distributions (in general). However, in special cases (where the variables are independent) it is possible to reconstruct the joint distribution given the marginals.

Definition 3.3.4 (Conditional Probability Distribution)**.** *Let (X, Y) be a discrete (or continuous) 2-D random variable with joint probability function (or joint p.d.f.) $f_{X,Y}(x, y)$. Let $f_X(x)$ and $f_Y(y)$ be the marginal probability functions of X and Y respectively. Then, the conditional distribution of Y given that $X = x$ is given by*

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \qquad \textit{provided } f_X(x) > 0$$

for each x within the range of X.
Similarly, the conditional distribution of X given that $Y = y$ is given by

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \qquad \textit{provided } f_Y(y) > 0$$

for each y within the range of Y.

In other words, the conditional probability density function of Y given X is simply the joint probability density of (X, Y) divided by the marginal distribution of X.

Note 3.3.5 (Remarks)**.**

1. The conditional p.f.'s (p.d.f.'s) satisfy all the requirements for a 1-D p.f. (p.d.f.). Thus, we have
   (a) For a fixed y, $f_{X|Y}(x|y) \geq 0$ and for a fixed x, $f_{Y|X}(y|x) \geq 0$
   (b) For discrete RVs,
   $$\sum_x f_{X|Y}(x|y) = 1 \qquad \text{and} \qquad \sum_y f_{Y|X}(y|x) = 1$$
   For continuous RVs,
   $$\int_{-\infty}^{+\infty} f_{X|Y}(x|y)\ dx = 1 \qquad \text{and} \qquad \int_{-\infty}^{+\infty} f_{Y|X}(y|x)\ dy = 1$$

2. For $f_X(x) > 0$,
$$f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y|x)$$
Similarly, for $f_Y(y) > 0$,
$$f_{X,Y}(x,y) = f_Y(y) f_{X|Y}(x|y)$$

Note 3.3.6.

Consider $f_{Y|X}(y|x)$ for the following points:

- The conditional distribution is similar in meaning to the conditional probability. It is the distribution of the RV Y when the RV X is fixed at a certain value x.
- It is important to remember that $f_{Y|X}(y|x)$ is a distribution for y (and NOT x), so it must satisfies all the properties of a pdf/pmf of the argument y for every x that it is defined.
- $f_{Y|X}(y|x)$ may or may not be a function of x. But it is defined only when x satisfies $f_X(x) > 0$. If $f_{Y|X}(y|x)$ does not depend on x, then X and Y are independent.
- $f_{Y|X}(y|x)$ is NOT a pdf/pmf for x. So, there is no requirement that $\int_{-\infty}^{+\infty} f_{Y|X}(y|x)dx$ = 1 when Y is continuous or $\sum_x f_{Y|X}(y|x) = 1$ when Y is discrete.

Note 3.3.7 (Conditional Probability and Conditional Expectation)**.**

Both the conditional probability and conditional expectation are established on the conditional distribution. In particular, if (X,Y) is a continuous random vector, for any x and y,
$$P(Y \le y | X = x) = \int_{-\infty}^{y} f_{Y|X}(t|x)dt$$
$$E(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy,$$
where the former depends on both y and x but the latter depends only on x. If (X,Y) is a discrete random variable, the integration is replaced with summation.
Alternatively, you can also use
$$E(Y|X) = \int_{-\infty}^{\infty} y f_{Y|X}(y|X)dy,$$
which is a function of the random variable X.

Note 3.3.8 (Uniform Distribution)**.**

(X,Y) is uniformly distributed if its pdf/pmf is in the form
$$f_{X,Y}(x,y) = \begin{cases} c & (x,y) \in A \\ 0 & \text{elsewhere} \end{cases},$$
where c is a real number independent of x and y. In fact if (X,Y) is continuous, $c = \dfrac{1}{area(A)}$; if (X,Y) is discrete, $c = \dfrac{1}{\#A}$ where #A represents the number of elements in A.
(X,Y) is uniform does not imply that X and/or Y is uniform. Likewise, "both X and Y are uniformly distributed" does not imply that (X,Y) is uniformly distributed.
It is important to note that when the joint probability density is uniform inside a region A_1 and zero otherwise, then the probability of $(X,Y) \in A_2$ is simply $\dfrac{area(A_2 \cap A_1)}{area(A_1)}$. In other words, the probability of is proportional to the area of the region (where is it non-zero).

Note 3.3.9.

When we say that the double integral of the pdf over the cartesian plane must be equal to 1, we mean that the volume bounded by the pdf over the entire plane is 1. When the joint pdf is uniform over a bounded region (because if the region is unbounded, $f_{X,Y}(x,y)$ would have to be 0 as the area is infinite), the volume bounded by the pdf over that region is simply the area of the region times the height (which is the constant value of pdf). This is why, probability is proportional to the area in which we are calculating the probability (NOT the area over which the pdf is uniform and non-zero)

3.4 Independent Random Variables

Definition 3.4.1 (Independent RV)**.** *Random variables X and Y are said to be independent if, and only if,*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \qquad \textbf{\textit{for all}}\ x,y$$

This can be extended to n random variables - $X_1, X_2, ..., X_n$ are independent if, and only if,

$$f_{X_1,X_2,...,X_n}(x_1,x_2,...,x_n) = f_{X_1}(x_1)f_{X_2}(x_2)...f_{X_n}(x_n) \qquad \textbf{\textit{for all}}\ x_i$$

The interpretation of independent random variables is similar to the interpretation of independent events. In this case, if X and Y are 2 independent random variables, then knowing the value of one does not affect the distribution of the other random variable. It practically means that the value of one random variable is not related to that of the other.

Note 3.4.2 (Checking independence)**.**

There are several ways to define/check the independence of random variables. They are:

- Random variables X, Y are independent if, and only if, for **arbitrary** sets $A, B \subset \mathbb{R}$,

  $$P(X \in A; Y \in B) = P(X \in A)P(Y \in B),$$

  which is quite similar to the definition of independent events.

- Random variables X, Y are independent if, and only if, for any $x, y \in \mathbb{R}$,

  $$P(X \leq x; Y \leq y) = P(X \leq x)P(Y \leq y),$$

  which can be rewritten as $F_{X,Y} = F_X(x)F_Y(y)$

- Random variables X, Y are independent if, and only if, for **any** functions g_1 and g_2, $E(g_1(X)g_2(Y))$ = $E(g_1(X)) \times E(g_2(Y))$

Note 3.4.3.

Note that in general, $f_{Y|X}(y|X) \neq f_Y(y)$ for a particular value of Y. They are equal for all values of X, Y if, and only if, X, Y are independent events. This follows from the interpretation of independent events - the knowledge about X does not affect the distribution of Y. This is analogous to the probability case, in which, $P(A|B) = P(A)$ when A, B are independent.

Note 3.4.4.

In the conditional distribution $f_{Y|X}$, there can be 2 variables x, y. However, the random variable is only y since we need to fix the value of x before solving. Furthermore, if X and Y are independent, then $f_{Y|X}$ only involves y. This is because for independent events, the conditional distribution is equal to the marginal distribution (irrespective of the the value of x.) In other words $Y|X$ only equals to Y if X and Y are independent.

Note 3.4.5 (Product Space)**.**

The product of 2 positive functions $f_X(x)$ and $f_Y(y)$ means a function that is positive on a product space. In other words, if $f_X(x) > 0$ for $x \in A_1$ and $f_Y(y) > 0$ for $y \in A_2$, then $f_X(x)f_Y(y) > 0$ for $(x, y) \in A_1 \times A_2$. An example of a product space is:

$$(x, y) \in [a, b] \times [c, d] \iff a \leq x \leq c,\ b \leq y \leq d$$

"$f_{X,Y}(x, y)$ is positive in a product space" is a necessary (but not sufficient) condition so that two random variables are independent. It can be used to assert that two random variables are not independent if this condition is not met, but it cannot be used to claim that two random variables are independent if they satisfy this condition.
The logic behind this is very simple - if the region in which $f_{X,Y}(x, y)$ is positive is not a rectangle (say, its a triangle), then you can find a point outside the triangle such that it lies within both the intervals under which x and y are constrained. But since $f_{X,Y}(x, y)$ would be 0 there and the marginals $f_X(x)$ and $f_Y(y)$ would be non-zero, $f_{X,Y}(x, y) \neq f_X(x)f_Y(y)$

- If X, Y are continuous random variables, for them to be independent, we need that $A = \{(x, y)|f_{X,Y}(x, y) > 0\}$ can be written in the form $(\bigcup_{i=1}^{\infty}[a_i, b_i]) \times (\bigcup_{j=1}^{\infty}[c_j, d_j])$. An even quicker view is that at least it must be a union of a countable number of rectangles.
- If X, Y are discrete random variables, for them to be independent, we need that for every $x \in A_1,\ y \in A_2,\ f_{X,Y}(x, y) > 0$

If you can decompose $f_{X,Y}(x, y)$ into 2 factors - one involving only x (and some constants) and the other involving only y (and some constants), then they are independent random variables (although the two factors themselves may not be $f_X(x)$ and $f_Y(y)$). On the other hand, if you cannot decompose the joint p.d.f. $f_{X,Y}(x, y)$ into 2 factors by separating x and y, they're not independent. This provides an easy way to check if 2 random variables are independent or not. More formally, if we can rewrite the joint p.d.f $f_{X,Y}(x, y)$ as the product of two functions $g(x)$ and $h(y)$, then X, Y are independent. Otherwise, they are not independent. It is important to remember that $g(x)$ is not necessarily the marginal distribution of x and similarly, $h(y)$ is not necessarily the marginal distribution of y.

3.5 Expectation of Random Variables

Definition 3.5.1 (Expectation)**.** *The expectation of $g(X, Y)$ is defined as*

$$E[g(X, Y)] = \begin{cases} \sum_x \sum_y g(x, y)f_{X,Y}(x, y), & \textit{for Discrete RV} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f_{X,Y}(x, y)dxdy, & \textit{for Continuous RV} \end{cases}$$

Note 3.5.2 (Expectation of Independent RV)**.**

If X, Y are independent random variables, then $E[XY] = E[X]E[Y]$. More generally, $E[g_1(X)g_2(Y)]$ = $E[g_1(X)] \times E[g_2(Y)]$ holds for **any** functions g_1 and g_2 **if, and only if**, X and Y are independent.

Note 3.5.3 (Some interesting questions)**.**

1. What is the meaning of $E(E(Y|X))$? How would you evaluate it?
   Ans: $E(E(Y|X))$: The inner expectation refers to the expected value of y over the conditional distribution of $f_{Y|X}(y|x)$. So, $E(Y|X) = \int_{-\infty}^{\infty} yf_{Y|X}(y|X)dy$. The result would be a function of X. Then, the expectation of this would be taken over the marginal distribution of x (since only x is the random variable now). So,

$$E(E(Y|X)) = \int_{-\infty}^{\infty} x\left(\int_{-\infty}^{\infty} yf_{Y|X}(y|X)dy\right)dx$$

2. What is $E(Y)$? How would you evaluate it?
   Ans: There are two ways to interpret and evaluate $E(Y)$.

   (a) Using the marginal distribution of $f_Y(y)$,

   $$E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy$$

   (b) Using the joint distribution of X, Y,

   $$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x,y) dx dy$$

   In other words, for every value of (x, y), just look at the value of y and multiply it by its probability density - take the sum over all possible values of (x, y).

   It is clear that both methods give the same answer (as expected - pun intended)

3. How would you compute $E[\alpha g_1(X)+\beta g_2(Y)]$ and $E[\alpha g_1(X)\beta g_2(Y)]$, where α, β are real numbers and g_1, g_2 are arbitrary but fixed functions?
   Ans:

   $$\begin{aligned} E[\alpha g_1(X) + \beta g_2(Y)] &= \alpha E[g_1(X)] + \beta E[g_2(Y)] \\ &= \alpha \left(\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g_1(X) f_{X,Y}(x,y) dx\ dy \right) + \beta \left(\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g_2(Y) f_{X,Y}(x,y) dx\ dy \right) \end{aligned}$$

   $$\begin{aligned} E[\alpha g_1(X)\beta g_2(Y)] &= \alpha\beta E[g_1(X) g_2(Y)] \\ &= \alpha\beta \left(\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g_1(X) g_2(Y) f_{X,Y}(x,y) dx\ dy \right) \end{aligned}$$

   You cannot simplify it further unless you know that X, Y are independent, in which case $E[\alpha g_1(X)\beta g_2(Y)] = \alpha E[g_1(X)] \times \beta E[g_2(Y)]$

3.5.1 Covariance as a Special Case of Expectation

Let $g(X, Y) = (X - \mu_X)(Y - \mu_Y)$. This leads to the definition of covariance between two random variables.

Definition 3.5.4 (Covariance)**.** *Let (X, Y) be a bivariate random vector with joint p.f. (or p.d.f.) $f_{X,Y}(x, y)$. Then, the covariance of (X, Y) is defined as*

$$Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

For discrete case,

$$\begin{aligned} Cov(X,Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f_{X,Y}(x,y) \end{aligned}$$

For continuous case,

$$\begin{aligned} Cov(X,Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{X,Y}(x,y) dx dy \end{aligned}$$

Note 3.5.5 (Practical Interpretation of Covariance)**.**

In probability theory and statistics, covariance is a measure of the joint variability of two random variables.If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values (that is, the variables tend to show similar behavior), the covariance is positive. In the opposite case, when the greater values of one variable mainly correspond to the lesser values of the other, (that is, the variables tend to show opposite behavior), the covariance is negative. The sign of the covariance therefore shows the tendency in the linear relationship between the variables. The magnitude of the covariance is not easy to interpret because it is not normalized and hence depends on the magnitudes of the variables. The normalized version of the covariance, the correlation coefficient, however, shows by its magnitude the strength of the linear relation.
Covariance shows the joint behaviour of X and Y. If covariance is positive, we can infer that whenever $X > \mu_X$, it is likely that $Y > \mu_Y$ (try to understand this using the definition of covariance). Similarly, whenever $X < \mu_X$, it is likely that $Y < \mu_Y$ (because the value of $(X - \mu_X)(Y - \mu_Y)$ would still be positive). On the other hand, if the covariance is negative, we can infer that whenever one of the random variables (say, X) is more than its expected value (μ_X), the other random variable (Y) is likely to be less than its expected value μ_Y.

Note 3.5.6 (Relation between Covariance and Independence)**.**

If X and Y are independent, the covariance is zero (because there is no relation between the 2 random variables). But the converse is not necessarily true: covariance = 0 does not imply the independence of the random variables.

Note 3.5.7 (Remarks on Covariance)**.**

- $Cov(X, Y) = E(XY) - \mu_X\mu_Y$ (i.e., $Cov(X, Y) = E(XY) - E(X)E(Y)$
- If X and Y are independent, then $Cov(X, Y) = 0$. However, $Cov(X, Y) = 0$ **does not imply** that X and Y are independent.
- $Cov(X + a, Y + b) = Cov(X, Y)$
- $Cov(aX, bY) = ab\ Cov(X, Y)$
- Combining the above two properties, $Cov(aX + b, cY + d) = ac\ Cov(X, Y)$
- $V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab\ Cov(X, Y)$. The term $2ab\ Cov(X, Y)$ tells us how X and Y behave jointly. In particular, when X, Y are independent, then $V(aX+bY) = a^2V(X)+b^2V(Y)$
- Using $V(\alpha X) = \alpha^2V(X)$ for any real number α, we can simplify the above to be: $V(X + Y) = V(X)+V(Y)+2Cov(X, Y)$. This can also be extended to multiple random variables, namely, $V(X_1 + X_2 + \cdots + X_n)$. This leads to the sum of n variance terms and $\binom{n}{2}$ co-variance terms. However, if the random variables are uncorrelated, this formula can be greatly simplified as all the covariance terms disappear; so we have if $X_1, X_2, \ldots, X_n$ are pairwise uncorrelated,

  $$V(X_1 \pm X_2 \pm X_3 \pm \cdots \pm X_n) = V(X_1) + V(X_2) + \cdots + V(X_n)$$

  Note the "$\pm$" on the left side of the equality and the "+" on the right side. This is because $(-1)^2 = 1$
- The variance is a special case of the covariance in which the two variables are identical (that is, in which one variable always takes the same value as the other)

  $$Cov(X, X) = var(X) \equiv \sigma^2(X) \equiv \sigma_X^2$$

- For any constant α, $Cov(X, \alpha) = 0$
- Covariance depends on the degree of association between X and Y, and also on the magnitudes of X and Y. So, larger covariance does not necessarily imply larger degree of association (i.e., the magnitude of X and Y can be big). Thus, we use correlation coefficient to eliminate the impact of the magnitude of X and Y. We adjust it using σ_X and σ_Y, since they tell us about the magnitude of X and Y.

Covariance measures the **linear** relationship between 2 variables. So, if two random variables are uncorrelated (i.e., their covariance is 0) it simply means that there is no linear relationship between them. It is possible that they are very closely associated by a non-linear relationship (eg. logarithmic, quadratic, trigonometric, etc.) and hence, they may not be independent. So, independence is a much stronger (and stricter) condition than uncorrelatedness.

3.5.2 Correlation coefficient

Definition 3.5.8 (Correlation coefficient)**.** *Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. It is denoted by* ρ*.*

$$\begin{aligned}\rho &= \frac{Cov(X,Y)}{\sqrt{V(X)}\,\sqrt{V(Y}} \\ &= \frac{\mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))]}{\sigma_x\ \sigma_y} \\ &= \frac{\mathbb{E}[XY]-\mathbb{E}[X]\ \mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2]-(\mathbb{E}[X])^2}\ \sqrt{\mathbb{E}[Y^2]-(\mathbb{E}[Y])^2}}\end{aligned}$$

Note 3.5.9 (Remarks)**.**

- Although $-\infty < Cov(X,Y) < \infty$, $-1 < \rho_{X,Y} < 1$. This is because the correlation coefficient is "normalised".
- $\rho_{X,Y}$ is a measure of the degree of **linear** relationship between X and Y.
- If X and Y are independent, then $\rho_{X,Y} = 0$. On the other hand, the converse is not necessarily true.
- The sign of ρ indicates whether the correlation is positive or negative. For example, if $\rho > 0$ then it means that whenever $X > \mu_X$ it is more likely that $Y > \mu_Y$ also. Similarly, if $X < \mu_X$, it is more likely that $Y < \mu_Y$ when the correlation coefficient is positive. If the correlation coefficient is negative, it means that as X increases, Y tends to decrease, and vice versa.
- The magnitude of ρ indicates how strongly the 2 random variables are linearly associated. The closer the magnitude is to 1, the greater the association/correlation. If $|\rho| = 1$, there is a perfect linear correlation between the two variables - when plotted, all the values of (x, y) lie on a straight line.

Chapter 4

Special Probability Distributions

4.1 Discrete Uniform Distribution

Definition 4.1.1 (Discrete Uniform Distribution)**.** *If the random variable X assumes the values $x_1, x_2, \ldots x_k$ with equal probability, then the random variable X is said to have a discrete uniform distribution and the probability function is given by*

$$f_X(x) = \begin{cases} \frac{1}{k}, & \text{if } x = x_1, x_2, \ldots x_k \\ 0, & \text{otherwise} \end{cases}$$

Theorem 4.1.2. *Mean and Variance of Discrete Uniform Distribution*

$$\mu = E(X) = \sum_{all\ x} x f_X(x) = \sum_{i=1}^{k} x_i \frac{1}{k} = \frac{1}{k} \sum_{i=1}^{k} x_i$$

$$\sigma^2 = V(X) = \sum_{all\ x} (x-\mu)^2 f_X(x) = \frac{1}{k} \sum_{i=1}^{k} (x_i - \mu)^2$$

Alternatively,

$$\sigma^2 = E(X^2) - \mu^2 = \frac{1}{k} \left(\sum_{i=1}^{k} x_i^2 \right) - \mu^2$$

4.2 Bernoulli and Binomial Distribution

Definition 4.2.1 (Bernoulli Experiment)**.** *A Bernoulli experiment is a random experiment with only 2 possible outcomes, say "success" or "failure" (e.g. head or tail, defective or non-defective, boy or girl, yes or no). It is convenient to code the 2 outcomes as 1 and 0.*

Definition 4.2.2 (Bernoulli Distribution)**.** *A random variable X is defined to have a Bernoulli distribution if the probability function of X is given by*

$$f_X(x) = \begin{cases} p^x (1-p)^{1-x}, & x = 0, 1 \\ 0, & \text{otherwise} \end{cases}$$

where the parameter p satisfies the $0 < p < 1$

Note 4.2.3.

- $(1-p)$ is often denoted by q.
- $P(X=1)=p$ and $P(X=0)=1-p=q$
- We code any outcome that we consider to be a success to be 1, and all other outcomes (that we consider to be failures) to be 0. This is why the only possible values of x are 0 and 1. For example, if we are interested in finding the number of red balls being drawn from a box then we will code any selection that results in a red ball being drawn to be 1, and all non-red balls to be 0.

Note 4.2.4 (Parameter and Family of Distributions)**.**

Suppose $f_X(x)$ 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. In the Bernoulli Distribution, p is the parameter. The collection of all probability distributions for different values of the parameter is called a **family** of probability distributions.

Theorem 4.2.5. *Mean and Variance of Bernoulli Distribution*

$$\mu = E(X) = p$$
$$\sigma^2 = V(X) = p(1-p) = pq$$

Definition 4.2.6 (Binomial Distribution)**.** *A random variable X is defined to have a binomial distribution with 2 parameters n and p (i.e. $X \sim B(n,p)$ - to be read as "X follows a binomial distribution with parameters n and p"), if the probability function of X is given by,*

$$P(X=x) = f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} p^x q^{n-x}$$

*for $x = 0, 1, \ldots, n$, where p satisfies $0 < p < 1$, $q = 1-p$, and n ranges over the positive integers. Here, X is the **number of successes that occur in n independent Bernoulli trials***

Note 4.2.7.

When $n=1$, the probability distribution of X becomes $f_X(x) = p^x(1-p)^{1-x},\ x = 0, 1$, which is identical to Bernoulli Distribution. Hence, we can say that Bernoulli Distribution is a special case of the binomial distribution.
A random variable X is a $B(n,p)$ random variable **if, and only if,** $X = X_1 + X_2 + \cdots + X_n$ where $X_1, \ldots, X_n$ are independent random variables, each of which follows the same Bernoulli distribution with the success probability p. By convention, we say "$X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) $Bernoulli(p)$ random variables".
This is particularly useful to derive the expectation and variance of X, as given below.

Theorem 4.2.8. *Mean and Variance of Binomial Distributions If X has a binomial distribution with parameters n and p (i.e., $X \sim B(n,p)$), then the mean and variance of X are*

$$\mu = E(X) = np$$
$$\sigma^2 = V(X) = np(1-p) = npq$$

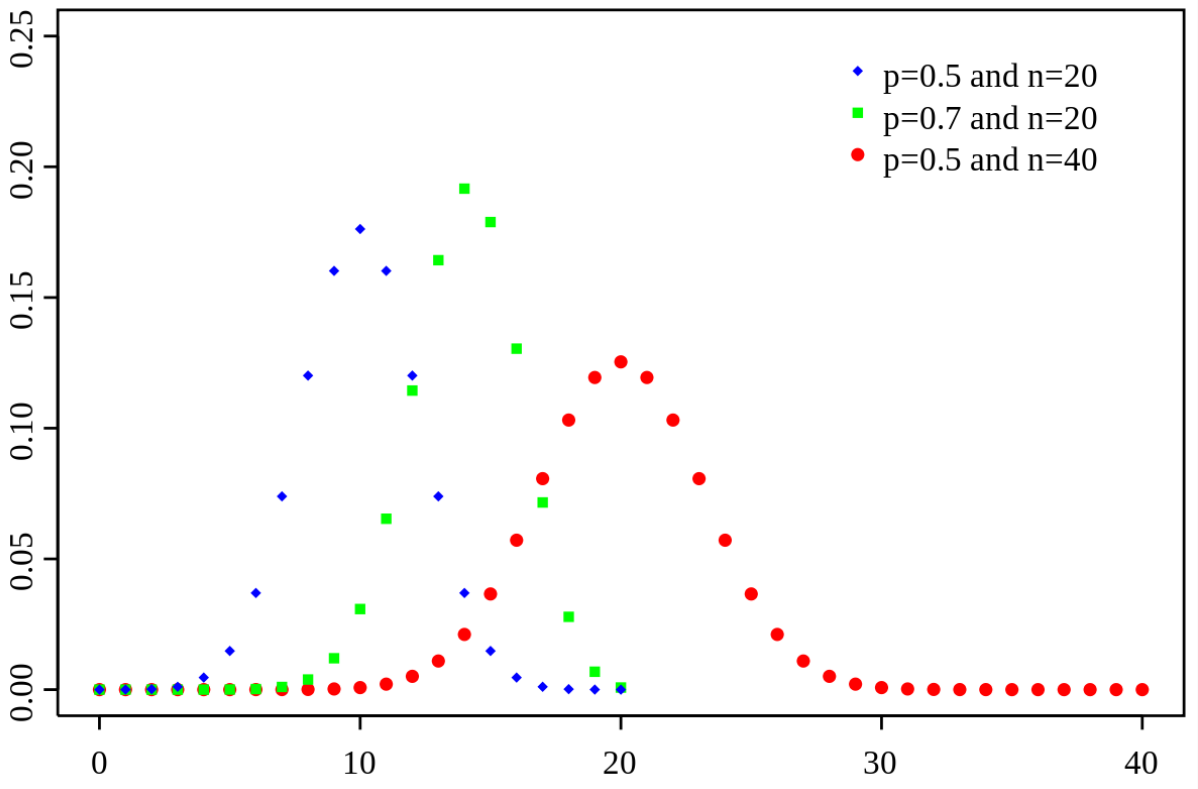


Figure 4.1: Examples of Binomial Distributions

The above formula can be derived by calculating the expected value and variance for each of the individual Bernoulli trial and then adding all of them up. It is always true that if $X = X_1 + X_2 + \cdots + X_n$, then $E[X] = E[X_1] + E[X_2] + \cdots + E[X_n]$. Moreover since we know that all of X_1 through X_n are independent, the covariance of any two of them is zero and so, $V(X) = V(X_1) + V(X_2) + \cdots + V(X_n)$. We already know that the expectation for any one trial is p and the variance of any one trial is pq (using Bernoulli trial), hence we easily observe that for n trials, the expectation is np and the variance is npq.

Note 4.2.9 (Conditions for a Binomial Experiment)**.**

1. It consists of n repeated Bernoulli trials
2. Only 2 possible outcomes: success and failure in each trial
3. $P(success) = p$ is the same constant in each trial, i.e., the number of trials previously performed or the outcomes of earlier trials should not affect the probability of success of any trial (e.g. Any experiment "without replacement" is likely to be not a Bernoulli distribution since the successive trials need not be independent).
4. Trials are independent
5. The random variable X is the number of successes among the n trials in a binomial experiment

Only if all the above conditions are met, $X \sim B(n, p)$.

Note 4.2.10 (Derivation of pmf of Binomial distribution)**.**

Consider a specific realization of $X_1, X_2, \ldots, X_n$ namely $x_1, x_2, \ldots, x_n$ such that $\sum_{i=1}^{n} x_i = x$. Note the independence of $X_1, X_2, \ldots, X_n$ and that they are all $Bernoulli(p)$ random variables. So, we

have,

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) &= P(X_1 = x_1)P(X_2 = x_2)\ldots P(X_n = x_n) \\ &= \prod_{i=1}^{n} p^{x_i} q^{1-x_i} = p^{\sum_{i=1}^{n} x_i} q^{n-\sum_{i=1}^{n} x_i} \\ &= p^x q^{n-x} \end{aligned}$$

$\sum_{i=1}^{n} x_i = x$ on the one hand means that the realized value for the corresponding X is x; on the other hand, it means that out of n trials, we get x successes. There are $\binom{n}{x}$ number of such sequences, as we can think of it as choosing x positions to take value 1 out of a length n sequence, and other positions will be 0. As a consequence by nothing that for different choices of $x_1, x_2, \ldots, x_n$, $\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$ are sets of mutually exclusive events, we have

$$\begin{aligned} P(X = x) &= P\left(\bigcup_{x_1,\ldots,x_n : \sum x_i = x} \{X_1 = x_1, X_2 = x_2 \ldots, X_n = x_n\} \right) \\ &= \sum_{x_1,\ldots,x_n : \sum x_i = x} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &= \sum_{x_1,\ldots,x_n : \sum x_i = x} p^x q^{n-x} = \binom{n}{x} p^x q^{n-x} \end{aligned}$$

Note 4.2.11.

If X and Y are 2 **independent** random variables that follow Binomial distributions with the same probability of success but possibly differing number of trials, $X + Y$ is also a binomial distribution with the same probability of success but the number of trials equal to the sum of the trials of the two random variables. That is, if $X \sim B(n, p), Y \sim B(m, p)$ then, $X+Y \sim B(n+m, p)$. (Intuitively this should make sense since they are independent)

Note 4.2.12 (Hypergeometric Distribution)**.**

The hypergeometric distribution is a discrete probability distribution that describes the probability of k successes (random draws for which the object drawn has a specified feature) in n draws, without replacement, from a finite population of size N that contains exactly K objects with that feature, wherein each draw is either a success or a failure. In contrast, the binomial distribution describes the probability of k successes in n draws with replacement.
A random variable X follows the hypergeometric distribution if its probability mass function (pmf) is given by:

$$p_X(k) = P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where,

- N is the population size
- K is the number of success states in the population
- n is the number of draws (quantity each trial)
- k is the number of observed successes

4.2.1 Negative Binomial Distribution

Let us consider an experiment where the properties are the same as those listed for a binomial experiment with the exception that the trials will be repeated until a fixed number of successes occur.
We are interested in the probability of the k^{th} success occurring on the x^{th} trial where x is the random variable. Notice how this is different from a binomial distribution in which case we are interested in the probability of x successes in n trials.

Definition 4.2.13 (Negative Binomial Distribution)**.** *Let X be a random variable representing the number of trials to produce the k successes in a sequence of independent Bernoulli trials. The random variable X is said to follow a Negative Binomial Distribution with parameters k and p (i.e. $NB(k,p)$). The probability function of X is given by:*

$$P(X=x) = f_X(x) = \begin{cases} \binom{x-1}{k-1} p^x q^{x-k}, & \text{for } x = k, k+1, k+2, ... \\ 0, & \text{otherwise} \end{cases}$$

An example of negative binomial distribution would be to find the probability that the 5th success occurs in the 7th trial. So, for this to occur, we need to find the probability of getting 4 successes in the first 6 trials and then multiply that by the probability of getting a success (for the 7th trial).

Theorem 4.2.14. *Mean and Variance of Negative Binomial Distribution*
If $X \sim NB(k,p)$, i.e., if X follows a negative binomial distribution with parameters k and p, then,

$$\mu = E[X] = \frac{k}{p}$$
$$\sigma^2 = Var(X) = \frac{(1-p)k}{p^2}$$

4.2.2 Geometric Distribution

A geometric distribution is a special kind of negative binomial distribution in which we find the number of trials required to have the first success. In other words, we set the parameter $k = 1$. So, whenever we talk about a geometric distribution, there is only 1 parameter p, which refers to the probability of success for any individual trial. We write $X \sim Geom(0.4)$ to denote that X follows a geometric distribution with probability of success = 0.4.
Geometric distribution represents the probability of the number of successive failures before a success is obtained in a Bernoulli trial. That is, in a geometric distribution, a Bernoulli trial is repeated until a success is obtained and then stopped.
The probability and cumulative probability functions for a geometric distribution are:

$$P(X=x) = p(1-p)^{x-1}$$
$$P(X \leq x) = 1 - (1-p)^x$$

An easy way to understand the cumulative probability function is to think of the complement event: What is the probability that all the first x trials result in failure? That would be $(1-p)^x$. So, the probability that at least 1 success occurs in the first x trials is just $1-(1-p)^x$
Obviously, the mean of the geometric distribution is $\frac{1}{p}$ and the variance is $\frac{1-p}{p^2}$ since $k = 1$.

Note 4.2.15.

Another way to describe a negative binomial distribution is to consider it as the sum of many geometric random variables. For example, $Y \sim NB(k,p)$ and $X_i = Geom(p), i = 1,2,3,\ldots,k$. Then, it is possible to write $Y = X_1 + X_2 + \cdots + X_k$. So, it follows that $E[Y] = E[X_1] + E[X_2] + \ldots E[X_n] = \frac{1}{p} + \frac{1}{p} + \cdots + \frac{1}{p} = \frac{k}{p}$. Also, $Var(Y) = Var(X_1) + Var(X_2) + \cdots + Var(X_k) = \frac{1-p}{p^2} + \frac{1-p}{p^2} + \cdots + \frac{1-p}{p^2} = \frac{(1-p)k}{p^2}$

4.3 Poisson Distribution

Definition 4.3.1 (Poisson Experiments)**.** *Experiments yielding numerical values of a random variable X,* ***the number of successes occurring during a given time interval or in a specified region****, are called Poisson experiments. They must satisfy the following conditions:*

1. *The number of successes occurring in one time interval or specified region are* ***independent*** *of those occurring in any other disjoint time interval or region of space*
2. *The* ***probability of a single success*** *occurring during a very short time interval or in a small region is* ***directly proportional to the length of the time interval*** *or the size of the region and does not depend on the number of successes occurring outside this time interval or region.*
3. *The* ***probability of more than one success*** *occurring in such a short time interval or falling in such a small region is* ***negligible****.*

The given time interval, t, may be of any length, such as a minute, a day, a week, or even a year

Examples of Poisson experiment: generate observations for the random variable X representing the number of telephone calls in an hour received by an office, or the number of postponed games due to rain during a football season.
The specified region could be a line segment, an area, a volume, or perhaps a piece of material. In this case, X might represent the number of mushrooms in a plot of land, the number of bacteria in a given culture, or the number of typing errors in a page.

Definition 4.3.2 (Poisson Distribution)**.** *The number of successes X in a Poisson experiment is called a Poisson random variable. The probability distribution of the Poisson random variable X, is called the Poisson distribution and the probability function is given by*

$$f_X(x) = P(X = x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!}, \text{ for } x = 0,1,2,3,... \\ 0, \text{ otherwise} \end{cases}$$

where λ is the average number of successes occurring in the given time interval or specified region and $e \approx 2.71828$ is called Euler's number.

We write $X \sim P(\lambda)$ to indicate that X follows a Poisson Distribution with parameter λ.

Theorem 4.3.3. *Mean and Variance of Poisson Random Variable*
IF X has a Poisson distribution with parameter λ, then

$$\mu = E[X] = \lambda$$
$$\sigma^2 = V(X) = \lambda$$

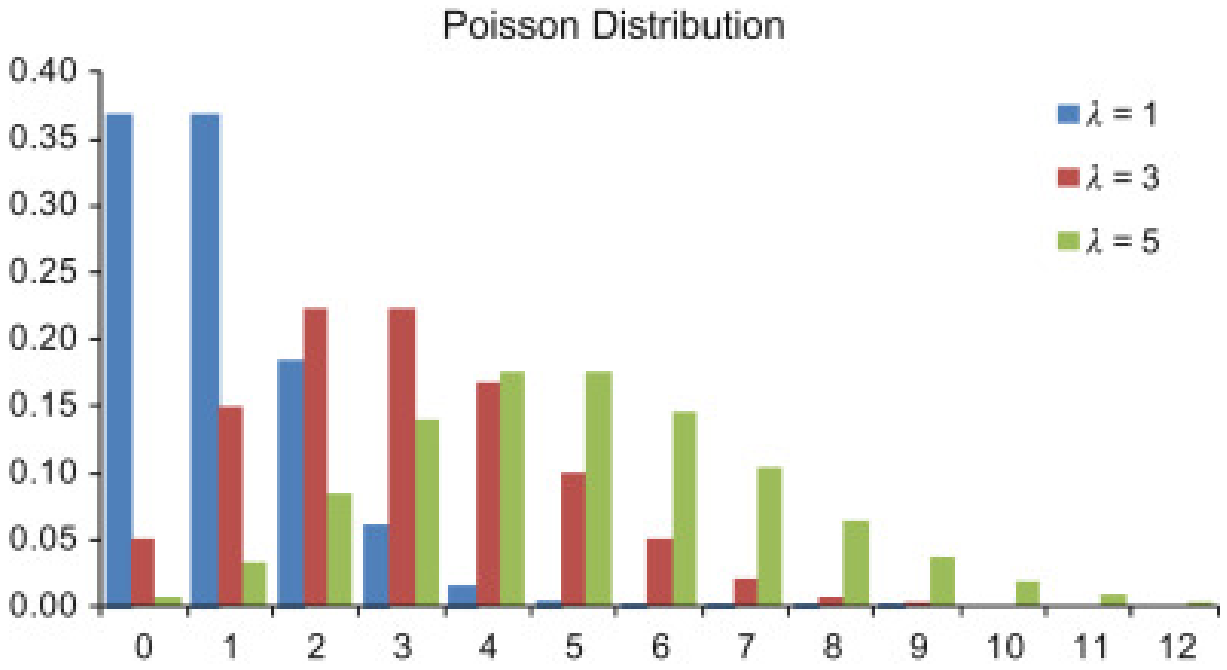


Figure 4.2: Example of some Poisson Distributions

Note 4.3.4 (Density Manipulation)**.**

The method for deriving these results is called density manipulation. The fundamental idea is simple: for any arbitrary probability function $f(x)$,

- if the distribution is discrete, $\sum_{x\in\{x|f(x)>0\}} f(x) = 1$
- if the distribution is continuous $\int_{\infty}^{\infty} f(x) = 1$

Both of the above follow directly from the axioms of probability (since the total probability must be 1).
Proof of mean and variance of Poisson Random Variable:

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} \\ &= \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} \\ &= \lambda\sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} \qquad \text{where } y = x-1 \\ &= \lambda \qquad \text{because } \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} = \sum_{y=0}^{\infty} f_Y(y) = 1, \text{where } Y \sim P(\lambda) \end{aligned}$$

In other words, $\sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} = 1$ because it is a probability function and its sum over all possible values must be 1.

The proof of variance of Poisson Distribution is obtained by first finding $E[X(X-1)]$.

$$\begin{aligned}
E[X(X-1)] &= \sum_{x=0}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-2)!} \\
&= \lambda^2 \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} \qquad \text{where } y = x-2 \\
&= \lambda^2
\end{aligned}$$

Now,

$$\begin{aligned}
V(X) &= E[X^2] - [E(X)]^2 \\
&= E[X(X-1)] + E[X] - [E(X)]^2 \qquad \text{due to linearity of expectation} \\
&= \lambda^2 + \lambda - \lambda^2 \\
&= \lambda
\end{aligned}$$

Note 4.3.5 (Properties of the Poisson Distribution)**.**

- Let X follows $Poisson(\lambda_1)$ distribution. Let Y follows $Poisson(\lambda_2)$ distribution. If X and Y are independent, then $X + Y \sim Poisson(\lambda_1 + \lambda_2)$.
  For example, if the average number of robberies in a day is 4, then the average number of robberies in 2 days will be 8, assuming that robberies occurring on different days are independent.
- Let X be the number of occurrences of an event in a period of time T; it has the $Poisson(\lambda)$ distribution. If Y is the number of occurrences of the event in a period of time tT, then $Y \sim Poisson(t\lambda)$.

Note 4.3.6 (Comparison between Distributions)**.**

Binomial Distribution, Negative Binomial distribution, and the Poisson distribution are all founded on Bernoulli trials. Their corresponding random variables X, however are defined differently:

- For Binomial distribution, X is defined to be the number of successes out of n independent Bernoulli trials with p constant for all trials.
- For Negative Binomial distribution, X is defined to be the number of trials needed so that we achieve k successes.
- For Poisson distribution, X is defined to be the number of successes in a period of time or in a specific region.

4.4 Poisson Approximation to the Binomial Distribution

Theorem 4.4.1. *Let X be a **Binomial** random variable with parameters n and p. That is,*

$$P(X = x) = f_X(x) = \binom{n}{x} p^x q^{n-x}, \text{ where } q = 1 - p$$

Suppose that $n \to \infty$ and $p \to 0$ in such a way that $\lambda = np$ remains a constant as $n \to \infty$. Then, X will have approximately a Poisson distribution with parameter np. That is,

$$\lim_{p \to 0,\ n \to \infty} P(X = x) = \frac{e^{-np}(np)^x}{x!}$$

Note 4.4.2.

If p is close to 1, we can still use Poisson distribution to approximate binomial probabilities by interchanging what we have defined to be a success and a failure so that p becomes a value close to zero.

Note 4.4.3.

The reason we can approximate a Binomial distribution using a Poisson distribution because for small values of p, the binomial distribution is right skewed (with the right-tail being longer). This makes intuitive sense because the lower the probability of success, the smaller the expected number of successes. Moreover, even the Poisson distribution is right-skewed. On the other hand, when the probability of success is close to $\frac{1}{2}$, the binomial distribution becomes nearly symmetric, and hence the normal distribution can better approximate it.

4.5 Continuous Uniform Distribution

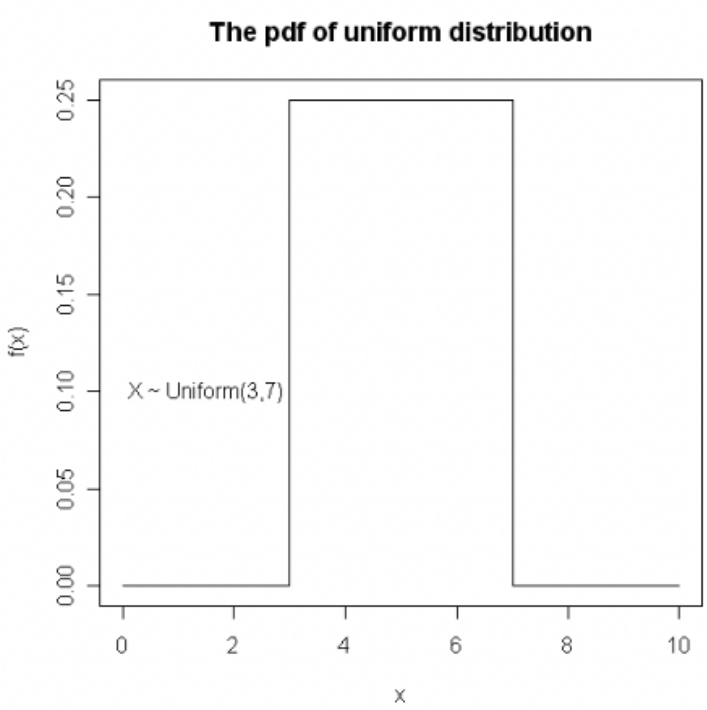
Definition 4.5.1 (Continuous Uniform Distribution)**.** *A continuous random variable is said to have a uniform distribution over the interval $[a, b]$, $-\infty < a < b < \infty$, denoted by $U(a, b)$, if its probability density function is given by,*

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & \text{for } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

This distribution is also referred to as rectangular distribution because of the rectangular shape of the p.d.f.

Theorem 4.5.2. *Mean and Variance of Continuous Uniform Distribution*
If X is uniformly distributed over $[a, b]$, then

$$E[X] = \frac{a+b}{2}$$
$$V(X) = \frac{(b-a)^2}{12}$$



Proof:

$$\begin{aligned} E[X] &= \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a} \left[\frac{x^2}{2}\right]_{x=a}^{x=b} \\ &= \frac{1}{b-a} \frac{b^2-a^2}{2} \\ &= \frac{a+b}{2} \end{aligned}$$

Also,

$$\begin{aligned} E[X^2] &= \int_a^b x^2 \frac{1}{b-a} dx == \frac{1}{b-a} \left[\frac{x^3}{3}\right]_{x=a}^{x=b} \\ &= \frac{1}{b-a} \frac{b^3-a^3}{3} \\ &= \frac{a^2+ab+b^2}{3} \end{aligned}$$

Then,

$$\begin{aligned} V(X) &= E[X^2] - (E[X])^2 \\ &= \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

Keep in mind that these formulae are applicable only when the distribution is defined on a single interval.

Note 4.5.3 (c.d.f. of a uniformly distributed random variable)**.**

Let X be a uniformly distributed random variable between $[a, b]$. Then,

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(t)dt \\ &= \begin{cases} \int_{-\infty}^{x} 0dt, \text{ for } x < a \\ \int_{-\infty}^{a} 0dt + \int_{a}^{x} \frac{1}{b-a}dt, \text{ for } a \leq x \leq b \\ \int_{-\infty}^{a} 0dt + \int_{a}^{b} \frac{1}{b-a}dt + \int_{b}^{x} 0dt, \text{ for } b < x \end{cases} \\ &= \begin{cases} 0, \text{ for } x < a \\ \frac{x-a}{b-a} \text{ for } a \leq x \leq b \\ 1, \text{ for } b < x \end{cases} \end{aligned}$$

4.6 Exponential Distribution

Definition 4.6.1 (Exponential Distribution)**.**

A continuous random variable X assuming all non-negative values is said to have an exponential distribution with parameter $\alpha > 0$ if its probability density function is given by

$$f_X(x) = \begin{cases} \alpha e^{-\alpha x}, \text{ for } x > 0 \\ 0, \text{ otherwise} \end{cases}$$

Note that $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

Theorem 4.6.2. *Mean and Variance of Exponential RV*
If X has an exponential distribution with parameter $\alpha > 0$, then

$$\begin{aligned} E[X] &= \frac{1}{\alpha} \\ V(X) &= \frac{1}{\alpha^2} \end{aligned}$$

The proof is very similar to that of continuous uniform distribution and is left an exercise to the reader.

Note 4.6.3.

The p.d.f. can be written in the form $f_X(x) = \frac{1}{\mu} e^{\frac{-x}{\mu}}$, for $x > 0$ and 0 otherwise. Then, $E[X] = \mu$ and $V(X) = \mu^2$.

Theorem 4.6.4. *No Memory Property of Exponential Distribution*
Suppose that X has an exponential distribution with parameter $\alpha > 0$. Then, for any 2 positive numbers s, t, we have

$$P(X > s + t | X > s) = P(X > t)$$

It can be proved using the formulae for conditional probability distributions and is a good exercise for the reader to revise.
The above theorem states that the exponential distribution has no memory in the following sense: Let X denote the life length of a bulb. Given that the bulb has lasted s time units (i.e. $X > s$) then the probability that it will last for the next t units (i.e. $X > s + t$) is the same as the probability that it would have lasted for the first t units if it were brand new.
This property usually refers to the cases when the distribution of a "waiting time" until a certain event does not depend on how much time has elapsed already. To model memoryless situations accurately, we must constantly 'forget' which state the system is in: the probabilities would not be influenced by the history of the process. For example, assume the bus frequency follows an exponential distribution. If you have been waiting at a bus stop for 10 minutes, what is the probability that you have to wait for 5 more minutes? This probability will be the same for someone who just arrived at the bus stop 10 minutes after you came. So, the expected amount of waiting time does not depend on the time for which you have already been waiting for.
Apart from the exponential distribution, the only other distribution which has this memoryless property is the geometric distribution (which makes sense because your odds of getting a success do not increase even though you have performed many trials before it - for example, you toss a coin 3 times and you get tails all 3 times (which you consider to be a failure). Your chances of getting a head (success) remain the same the next time you throw it too).

Note 4.6.5 (c.d.f of the exponential distribution)**.**

Let X be a continuous random variable following an exponential distribution with parameter α. Then, for $x \geq 0$,

$$F_X(x) = P(X \leq x) = \int_0^x \alpha e^{-\alpha t} dt = \left[-e^{-\alpha t}\right]_{t=0}^{t=x} = 1 - e^{-\alpha x},$$

and 0 otherwise.
Hence, $P(X > x) = e^{-\alpha x}$, for $x > 0$.

Exponential distribution is popularly used to model the survival (recovery) time of a patient in the medical research, where $P(X > t) = 1 - F_X(t)$ is called the **survival function** It is the probability that the survival (recovery) time of a patient is greater than t.

Note 4.6.6 (Application of the Exponential Distribution)**.**

The exponential distribution is frequently used as a model for the distribution of times between the occurrence of successive events such as customers arriving at a service facility or calls coming into a switchboard.

4.7 Normal Distribution

Definition 4.7.1 (Normal Distribution)**.** *The random variable X assuming all real values, $-\infty < x < \infty$ has a normal (or Gaussian) distribution if its probability density is given by*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \text{ for } -\infty < x < \infty$$

where $-\infty < \mu < \infty$ and $\sigma > 0$. It is denoted by $N(\mu, \sigma^2)$, and μ and σ^2 are called the parameters of the normal distribution.

4.7.1 Properties of the Normal Distribution

1. The graph of this distribution is of bell-shaped and called the normal curve. It is symmetrical about the vertical line $x = \mu$.
2. The mean, median, and mode of the normal distribution are the same and equal to μ.
3. The maximum point occurs at $x = \mu$ and its value is $\frac{1}{\sigma\sqrt{2\pi}}$.
4. The normal curve approaches the horizontal axis asymptotically as we proceed in either direction away from the mean.
5. The total area under the curve and above the horizontal axis is equal to 1 (since it is a probability density function)
6. It can be shown that $E[X] = \mu$ and $V(X) = \sigma^2$.
7. 2 normal curves are identical in shape if they have the same σ^2. But they are centered at different positions when their means are different.
8. As σ increases, the curve flattens, and as σ decreases, the curve steepens/sharpens.
9. If $X \sim N(\mu, \sigma^2)$, and if $Z = \frac{X-\mu}{\sigma}$, then $Z \sim N(0, 1)$. That is, Z follows the $N(0, 1)$ distribution. Then, $E[Z] = 0$ and $V(Z) = 1$. We say that Z has a standardized normal distribution. That is, the p.d.f. of Z may be written as: $f_Z(z) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{z^2}{2}\right)$.
   The importance of the standardized normal distribution is the fact that it is tabulated. Whenever X has distribution $N(\mu, \sigma^2)$, we can always simplify the process of evaluating the values of $P(x_1 < X < x_2)$ by using the transformation $Z = \frac{X-\mu}{\sigma}$. Then, $x_1 < X < x_2$ is equivalent to $\frac{x_1-\mu}{\sigma} < Z < \frac{x_2-\mu}{\sigma}$.
   Let $z_1 = \frac{x_1-\mu}{\sigma}$ and $z_2 = \frac{x_2-\mu}{\sigma}$. Then, $P(x_1 < X < x_2) = P(z_1 < Z < z_2)$
10. The density is symmetric about μ, which is the expectation and median of the distribution. One direct consequence is that $P(X \leq \mu) = P(X \geq \mu) = 0.5$. μ is also called the location parameter, which determines the location of the center of the distribution.
11. $\sigma^2 = V(X)$ is the shape parameter (also called the dispersion parameter in literature), which determines the shape of the density function.
12. For any value of μ, σ^2, the density if positive for all $x \in \mathbb{R}$. It get closer and closer to (but never equal to) 0, when x approaches ∞ or $-\infty$.
13. The standardization $Z = \frac{X-\mu}{\sigma}$ is very important. The density becomes symmetric about 0. That is, for any $z \in \mathbb{R}$, $P(Z \leq -z) = P(Z \geq z)$. $E[Z] = 0$ and $V(Z) = 1$. With this standardization, for $x_1 < x_2$, $P(x_1 < X < x_2)$ (with μ and σ^2 being any given values) can always be obtained from the table for Z.
14. For any normal random variable X, the probability that X is within c standard deviations from the mean value is always deterministic, where $c > 0$ is a known constant. In particular, if $X \sim N(\mu, \sigma^2)$,

$$P(\mu - c\sigma < X < \mu + c\sigma) = P(-c < \frac{X-\mu}{\sigma} < c) = P(|Z| < c),$$

which does not depend on μ and σ. Observe that $\frac{X-\mu}{\sigma}$ gives you exactly the number of standard deviations that the random variable is away from its expected value.
For example, the probability that a normal random variable is within 1 standard deviation of its mean is 68.27%, within 2 standard deviations is 95.45%, and within 3 standard deviations is 99.73%. So, if the IQ scores of a population follow a normal distribution with mean = 100 and standard deviation = 15, then we can say that 68.27% of the population has an IQ score between 85 and 115, and 99.73% of the population has an IQ score between 55 and 145.

15. For 2 **independent normal random variables** $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, then $X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$. Also, $X-Y \sim N(\mu_1-\mu_2, \sigma_1^2+\sigma_2^2)$

Note 4.7.2.

It is important to remember that the last point is not true for any kind of general distribution. For example, if $X \sim exp(\lambda)$ and $Y \sim exp(\lambda)$ and X and Y are independent, $X+Y$ does not follow an exponential distribution.

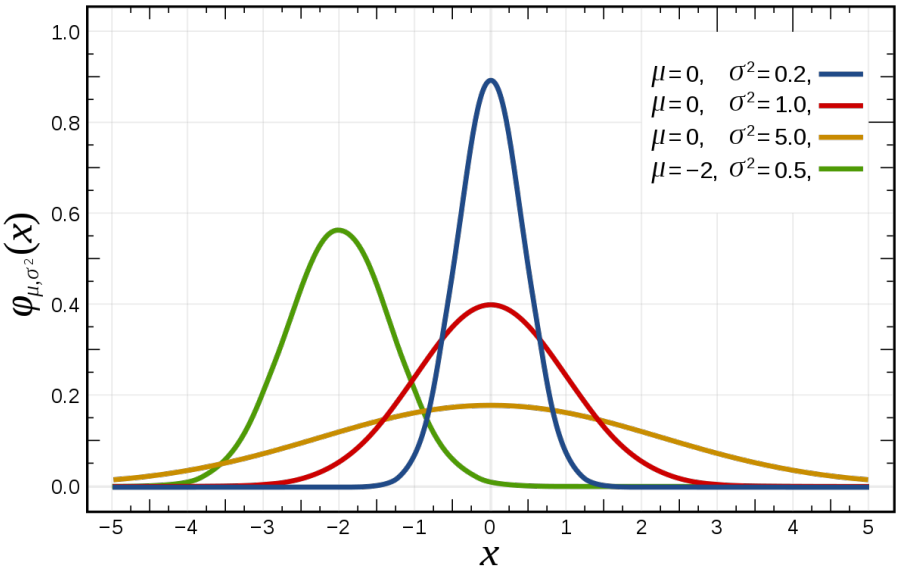


Figure 4.3: Examples of Normal Distribution

4.7.2 Application of Standardisation

It is very important to standardise the values before comparing two distributions that follow a normal distribution. The value of Z tells us how far the data point is from the mean (in terms of the number of standard deviations).
For example, consider 2 classes - A and B - which had an examination. The distribution of marks in A followed a normal distribution with mean = 50 and standard deviation = 5. Similarly, in class B, the mean was 65 and the standard deviation was 8. Then if student p (from class A) got 60 marks and student q (from class B) got 65 marks, can we say that q is better than p?
No! We cannot compare how well p and q did directly without standardising their scores. It may be the case that class B had an easier paper in general or the teacher was lenient. We want to know how well p and q did compared to their class, and use that to determine who did better. So, we find the Z values for both of them,

$$Z_p = \frac{60-50}{5} = 2, \; Z_q = \frac{65-65}{8} = 0$$

This can be interpreted as follows: p is 2 standard deviations above the class average (which means he did better than 95% of his class) while q is average in his class. Hence, we can conclude that p did better than q (assuming that students in both classes are randomly assigned).

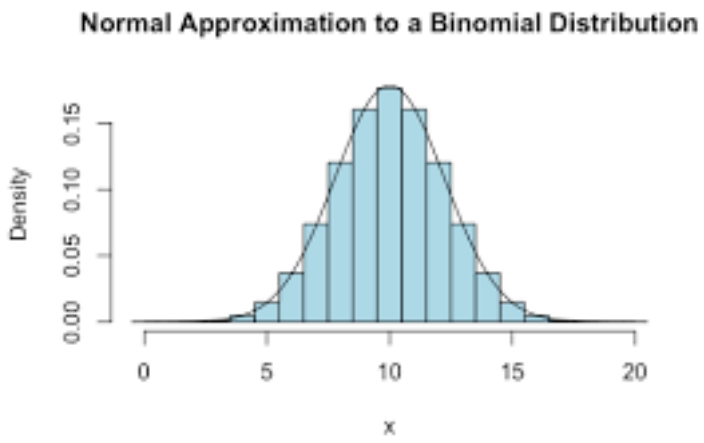
4.7.3 Statistical Tables

Statistical tables give the values $\Phi(z)$ for a given z, where $\Phi(z)$ is the cumulative distribution function of a standardized Normal random variable Z. It follows that $1 - \Phi(z)$ is the upper cumulative probability for a given z. Thus, $\Phi(z) = P(Z \leq z)$ and $1 - \Phi(z) = P(Z > z)$.
Some statistical tables give the 100α percentage points, z_α, of a standardized Normal distribution, where

$$\alpha = P(Z \geq z_\alpha) = \int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{z^2}{2}\right) dz$$

Since the p.d.f. of Z is symmetrical about 0, $P(Z \geq z_\alpha) = P(Z \leq -z_\alpha) = \alpha$

4.8 Normal Approximation to the Binomial Distribution



When $n \to \infty$ and $p \to 0$, we may use Poisson distribution to approximate a Binomial Distribution.
Similarly, when $n \to \infty$ and $p \to \frac{1}{2}$, we can use normal distribution to approximate the binomial distribution. In fact, even when n is small and p is not extremely close to 0 or 1, the approximation is fairly good. We need p to be close to $\frac{1}{2}$ because then the binomial distribution will be close to symmetric and hence, we can approximate it using normal distribution (which is always symmetric). If p is close to 0 or 1, then the binomial distribution is skewed and we cannot use normal distribution to approximate it. Thus, we use Poisson (not symmetric) approximation in such cases.
A good rule of thumb is to use the normal approximation only when $np > 5$ and $n(1-p) > 5$.

Notice that we use the Poisson approximation and normal approximation in completely different cases, depending on the value of p. If p is close to 0 (or 1), then we use the Poisson approximation. On the other hand, if p is close to $\frac{1}{2}$, we use the normal approximation. It is important to keep in mind that these are just approximations and they couldn't give you the exact value. Roughly speaking, how good the approximation is depends on how the corresponding conditions are satisfied.

Theorem 4.8.1. *If X is a binomial random variable with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$, then as $n \to \infty$,*

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \text{ is approximately } \sim N(0,1)$$

. This is equivalent to saying that X approximately follows $N(np, np(1-p))$ (in the above equation, we have standardized X directly).

An easy way to remember is that X follows $N(\mu, \sigma^2)$ where $\mu = np$ and $\sigma^2 = np(1-p)$ in the case of binomial distribution. The 2 parameters of the normal distribution always refer to the mean and the variance respectively.

4.8.1 Continuity Correction

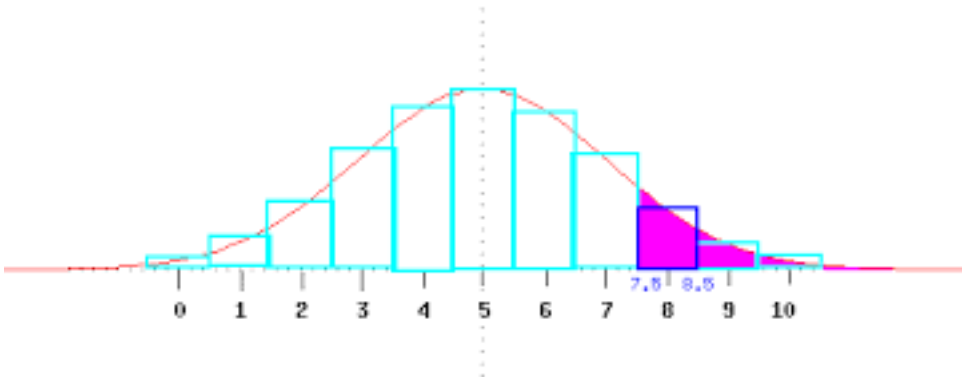


Figure 4.4: Continuity Corrections

When we use normal approximation for a binomial distribution to calculate probabilities, we run into some problems. For example, how would we calculate $P(X = k)$ for some value k? We know that for a continuous random variable, the probability that X takes on a fixed value is 0. But this is not true for a binomial distribution. Hence, we need to approximate $P(X = k)$ as being equal to $P(k - \frac{1}{2} < X < k + \frac{1}{2})$. It might be helpful to draw the bars of binomial distribution and the corresponding normal distribution to understand.
A continuity correction is the name given to adding or subtracting 0.5 to a discrete x-value. For example, suppose we would like to find the probability that a coin lands on heads less than or equal to 45 times during 100 flips. That is, we want to find $P(X \leq 45)$. To use the normal distribution to approximate the binomial distribution, we would instead find $P(X \leq 45.5)$ because 45 is the midpoint of the bar which ranges from 44.5 to 45.5, and since we want to include the entire bar (since we are including 45), we need to consider the upper bound. You don't need to memorize this continuity corrections if you understand when to include the point or not and use that to determine whether to add 0.5 or subtract 0.5. We can use the following approximations in general (only when using normal distribution to approximate binomial distribution):

1. $P(X = k) \approx P(k - \frac{1}{2} < X < k + \frac{1}{2})$
2. $P(a \leq X \leq b) \approx P(a - \frac{1}{2} < X < b + \frac{1}{2})$
   $P(a < X \leq b) \approx P(a + \frac{1}{2} < X < b + \frac{1}{2})$
   $P(a \leq X < b) \approx P(a - \frac{1}{2} < X < b - \frac{1}{2})$
   $P(a < X < b) \approx P(a + \frac{1}{2} < X < b - \frac{1}{2})$

3. $P(X \leq c) = P(0 \leq X \leq c) \approx P(\frac{-1}{2} < X < c + \frac{1}{2})$ (because the minimum value of X in a binomial distribution is 0)

4. $P(X > c) = P(c < X \leq n) \approx P(c + \frac{1}{2} < X < n + \frac{1}{2})$ (because the maximum value of X in a binomial distribution is n)

Chapter 5

Sampling and Sampling Distributions

5.1 Population and Sample

Definition 5.1.1 (Population)**.** *The totality of all possible outcomes or observations of a survey or experiment is called a population.*

A **sample** is a subset of a population.
Every outcome or observation can be recorded as a numerical or categorical value. Thus, each member of a population is a value of a random variable. There are two kinds of populations, namely, finite and infinite populations. As the names suggest, a finite population consists of a finite number of elements (for example, the number of species of dogs) whereas an infinite population is one that consists of an infinitely (countable and uncountable) large number of elements (for example, the results of all possible rolls of a pair of dice).

Note 5.1.2.

Some finite populations are so large that in theory we assume them to be infinite, such as the population lives of a certain type of storage battery being manufactured from a factory.

5.2 Random Sampling

5.2.1 Simple Random Sampling

A set of n members taken from a given population is called a **sample** of size n.

Definition 5.2.1 (Simple Random Sample)**.** *A simple random sample of n members is a sample that is chosen in such a way that every subset of n observations of the population has the same probability of being selected. (Or, in other words, each sample point in the population has equal probability of being selected)*

If the process of selecting a sample is **random**, it is necessarily independent of other samples being drawn.

5.2.2 Sampling without Replacement

In general, there are $\binom{N}{n}$ samples of size n that can be drawn from a finite population of size N without replacement (we disregard the order of selection). Each sample has an equal chance of being selected. Hence, each sample has a probability of $\frac{1}{\binom{N}{n}}$ of being selected.

5.2.3 Sampling with Replacement

Here, the order of elements is taken into consideration (since 2 elements can appear in the outcome but be chosen in different orders). In general, there are N^n samples of size n that can be drawn from a finite population of size N with replacement.Hence, each sample has a probability of $\frac{1}{N^n}$ of being selected

5.2.4 Sampling from an Infinite Population (with/without replacement)

We would be sampling from an infinite population if we sample with replacement from a finite population, and our sample would be random if

1. in each draw, all the elements of the population have the same probability of being selected, and
2. successive draws are independent.

Even if we sample without replacement from an "infinite" population, removing one element from an infinite population does not affect the population. For example, if you have an equal number of infinite red and blue balls. If you draw a red ball in your first draw, the probability of drawing a red ball still remains equal to the blue ball (because you still have infinite red balls!).

Note 5.2.2.

It is important to remember that within a population, there can be many repeated values and each is considered a separate data point. For example, consider the population of all NUS students. Say, we are interested in calculating their CAP. We know that the range of CAP is from 0 - 5 but we don't know the exact CAP of any student prior to sampling. Hence, this is a random variable. Further, multiple students may (in fact by Pigeonhole Principle, must) have the same CAP. We still consider them separate data points since they represent the CAP of different students. Just because the "values" are the same, it does not mean that the data points themselves are the same.

Note 5.2.3.

Selecting a sample without replacement from a finite population cannot be considered equivalent to sampling from an infinite population since the elements will get exhausted after a finite number of draws. In contrast, for an infinite population or finite population with replacement, you are able to draw a sample as large as you want.

Definition 5.2.4 (Random Sample)**.** *Let X be a random variable with certain probability distribution, $f_X(x)$. Let $X_1, X_2, \ldots, X_n$ be n independent random variables each having the same distribution as X, then $(X_1, X_2, \ldots, X_n)$ is called a random sample of size n from a population with distribution $f_X(x)$.*
The joint p.f. (or p.d.f.) of $(X_1, X_2, \ldots, X_n)$ is given by $f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\ldots f_{X_n}(x_n)$, where $f_X(x)$ is the p.f. (or p.d.f.) of the population.

We say that $(X_1, X_2, \ldots, X_n)$ is a random sample from a distribution with probability function $f_X(x)$. Here, a "random sample" is equivalent to "$X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.)". So, a random sample of size n can be represented by n independent random variables that follow the same distribution.
Keep in mind that $X_i's$ follow the same distribution (i.e., identically distributed) does not mean that $X_1 = X_2 = \cdots = X_n$ (i.e., identical random variables).
You cannot say that $X + Y = 2X$ just because X and Y follow the same distribution. The actual realization of the random variables X and Y may be different even though they follow the same

distribution! Two random variables are equal if, and only if, they map **each** sample point to the same value. If Alice throws a die and Bob also throws a die, both the outcomes follow the same distribution. But, they are different random variables (if we are interested in the number showing up)! If Alice gets a 4, it does not mean that Bob also gets a 4.

Note 5.2.5 (Motivation behind random sample, parameter, and statistic)**.**

1. The population parameter, say, μ, is **not observed**; but the values of $X_1, X_2, \ldots, X_n$ are observed as $x_1, x_2, \ldots, x_n$. So we hope to establish an "estimation rule" to estimate the unknown μ using the observed $X_1 = x_1, X_2 = x_2 \ldots, X_n = x_n$; such a rule is called a statistic (say, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$). In literature, $\bar{X}$ is also called an "estimator" for μ. $\bar{x}$ is **computed/observed** based on the sample, and is called an estimate.

2. Note that μ is an unknown constant - it is not a random variable. The population mean is fixed and constant - we just don't know the value because we haven't got data from the entire population.

3. The statistic is a function of the sample; the fundamental requirement is that it does not depend on any unknown parameters. for example, $g_1(X_1, \ldots, X_n) = 1$ is a statistic since it maps every input of samples to 1. But it is not a useful statistic.

4. The statistical performance of the statistic decides which one is better to use in practice. In other words, we would be interested in studying the distribution of a statistic, called "sampling distribution". The sampling distribution is also crucial for the subsequent inference for the corresponding parameter.

5. The sampling distribution for a statistic, say $\bar{X}$, is **not** how the sample $(X_1, X_2, \ldots, X_n)$ performs. Instead, it is the distribution for the corresponding statistic (say, $\bar{X}$). So, constructing the histogram based on the observed $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ to have a view of the sampling distribution of $\bar{X}$ is absolutely meaningless. Instead we should do the following:

   (a) Get the first sample $(X_1^{(1)}, X_2^{(1)}, \ldots, X_n^{(1)})$ from the distribution $f_X(x)$, and compute the sample mean $\bar{X}^{(1)}$.

   (b) Get the second sample $(X_1^{(3)}, X_2^{(2)}, \ldots, X_n^{(2)})$ from the distribution $f_X(x)$, and compute the sample mean $\bar{X}^{(2)}$.

   (c) Continue this procedure many times

   (d) Get the K^{th} sample $(X_1^{(K)}, X_2^{(K)}, \ldots, X_n^{(K)})$ from the distribution $f_X(x)$, and compute the sample mean $\bar{X}^{(K)}$.

   (e) Draw the histogram of $\bar{X}^{(1)}, \bar{X}^{(2)}, \ldots, \bar{X}^{(n)}$

6. In general, $\bar{X} \neq \mu$. $\bar{X}$ is a random variable and can have different realizations depending on the random sample drawn. But, $E[\bar{X}] = \mu$. It is important to understand this difference.

5.3 Sampling Distribution of the Sample Proportion

Let's say we are interested in finding out the proportion of males in a given population. Gender is a categorical variable with 2 possible values: male and female.

Let $X_1, X_2, \cdots, X_n$ be the observations of gender of a random sample of size n. Then, each of the $X_i \sim Bernoulli(p)$ where p is the population proportion of males.

So, $X_1 + X_2 + \cdots + X_n \sim Bin(n, p)$ (by definition of Binomial distribution - assuming that the population is large enough such that the value of p remains the same while selecting different people (since we are doing this selection without replacement).

Sample proportion, $\hat{p} = \frac{1}{n}\sum_i X_i$. So, $\hat{p} \sim \frac{1}{n} Bin(n, p)$

Obviously, the value of p is unknown (since we are trying to estimate it). So, in practice, we use $\hat{p}$ is our estimate for p.

If $n\hat{p}(1-\hat{p}) \geq 5$, then we can approximate the binomial distribution using the normal distribution. Then, $\hat{p} \sim N(p, \frac{p(1-p)}{n})$.

So, the standard error of $\hat{p}$ is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

5.4 Sampling Distributions of Sample Mean

Our main purpose in selecting random variables is to elicit information about the unknown population parameters.

5.4.1 Statistic and Sampling Distribution

A function of a random sample $(X_1, X_2, \ldots, X_n)$ (which does not depend on any unknown quantities) is called a **statistic**. For example, $\bar{X}$ is a statistic as $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.
Another example is X_{median} or X_{min}.
But $V = \sum_{i=1}^n \frac{(X_i - \mu)^2}{n-1}$ is not a statistic if we do not know the population mean.
Hence, **a statistic is a random variable**. It is meaningful to consider the probability distribution of a statistic. The probability distribution of a statistic is called a **sampling distribution**.

Sample Mean

If $X_1, X_2, \ldots, X_n$ represent a random sample of size n, then the sample mean is defined by the statistic

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

If the values in a random sample are observed and they are $x_1, x_2, \ldots, x_n$, then the **realization** of the statistic $\bar{X}$ is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We denote the population mean by μ_X.

Theorem 5.4.1. *Sampling Distribution of the sample mean*
*For random samples of size n taken from an infinite population or from a **finite population with replacement** having population mean μ and population standard deviation σ, the sampling distribution of the sample mean $\bar{X}$ has its mean and variance given by*

$$\mu_{\bar{X}} = \mu_X \qquad \text{and} \qquad \sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n}$$

, where n is the sample size. In other words,

$$\mathbb{E}[\bar{X}] = \mathbb{E}[X] \qquad \textit{and} \qquad V(\bar{X}) = \frac{V(X)}{n}$$

Theorem 5.4.2. *Law of Large Numbers (LLN)*
Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population having any distribution with mean μ and finite population variance σ^2. Then for any $\epsilon \in \mathbb{R}$,

$$P(|\bar{X} - \mu| > \epsilon) \to 0 \textit{ as } n \to \infty$$

.

This says that as the sample size increases, the probability that the sample mean differs from the population mean goes to zero. Another way of looking at this is that as n gets larger, it is increasing likely that $\bar{X}$ is close to μ. (The error between our estimated mean and the actual population mean gets smaller and smaller as we conduct sampling with larger and larger samples)

Note 5.4.3.

The strong law of large numbers (also called Kolmogorov's law) states that the sample average converges almost surely to the expected value. That is, $P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$

Note 5.4.4 (Proof of LLN)**.**

With Chebyshev's inequality, this theorem can be easily proved. Note that ϵ is an arbitrary but fixed constant. So, we have

$$0 \leq P(|\bar{X} - \mu| > \epsilon) \leq \frac{V(\bar{X})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \qquad \text{as } n \to \infty$$

The variance of the population should be finite for the expectation of the random variable $\bar{X}$ to converge to a fixed value. Obviously, μ must be finite since it is a fixed unknown constant. If variance were infinite, $\bar{X}$ could assume infinitely many values (realizations) for different random samples and there is no guarantee that the expectation of all these values will be equal to the population mean.

Theorem 5.4.5. *Central Limit Theorem*
Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from. a population having any distribution with mean μ and finite population variance σ^2. The sampling distribution of the sample mean $\bar{X}$ is ***approximately normal*** *with mean μ and variance $\frac{\sigma^2}{n}$ if n is* ***sufficiently large****. Hence,*

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \textit{ follows approximately } N(0,1)$$

Note that the variance of the sampling distribution of the sample mean is at most equal to the population variance (since $n \geq 1$).
The above theorem is powerful in the sense that it tells us that the distribution of the sampling distribution will be normal if n is large **no matter what the shape of the population distribution is**.

1. If for all $i = 1, 2, \ldots, n$, X_i are $N(\mu, \sigma^2)$, then $\bar{X}$ follows $N(\mu, \frac{\sigma^2}{n})$ regardless of the sample size n
2. Similarly, if for all $i = 1, 2, \ldots, n$, X_i are approximately $N(\mu, \sigma^2)$, then $\bar{X}$ approximately follows $N(\mu, \frac{\sigma^2}{n})$ regardless of the sample size n

Note that the above 2 points have nothing to do with the central limit theorem. With the normality assumption, the distribution of $\bar{X}$ is exactly normal. More generally, the linear combination of independent normal random variables with the same parameters is also a normal random variable (but obviously the parameters change). If $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, then $a_1X_1 + a_2X_2 + \cdots + a_nX_n \sim N(\mu \sum_i^n a_i, \sigma^2 \sum_i^n a_i^2)$. Then, if you choose each of the a_i to be equal to $\frac{1}{n}$, you get the first result.

Note 5.4.6.

Central limit theorem only gives us the approximate distribution for $\bar{X}$, not the exact distribution. More rigorously, central limit theorem says that, for any $z \in \mathbb{R}$,

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

where $\Phi(z)$ is the c.d.f. for the standard normal distribution. So, practically, when n is large, the distribution of $\frac{\bar{X} - \mu)}{\sigma/\sqrt{n}}$ is similar to that of a standard normal random variable. We normally use central limit theorem when $n \geq 30$ but this is not a fixed rule. If the actual distribution is quite symmetric, we can use it even for smaller values of n. The smaller the n, the less accurate the approximation.

Note 5.4.7.

You must ensure that the distribution for which central limit theorem is being applied has a finite variance. For example, a Cauchy distribution has no mean or variance and so you cannot apply the central limit theorem. Instead, any linear combination of Cauchy variables has a Cauchy distribution (so that the mean of a random sample of observations from a Cauchy distribution has a Cauchy distribution).

Note 5.4.8 (Central Limit Theorem vs LLN)**.**

Central Limit Theorem provides an approximate distribution of $\bar{X}$ (much more information). LLN provides a likely good estimate of μ based on $\bar{X}$ from a large sample but says nothing about the distribution.

Note 5.4.9.

It is the distribution of $\bar{X}$ that follows approximately a normal distribution if n is large. The underlying distribution of the population does not follow a normal distribution even if n is large, i.e., $\bar{X}$ follows approximately normal as n is large but X does not follow approximately normal even if the population size is very large. Each individual sample still follows the same underlying distribution that the population follows. For example, consider tossing a coin. Let the sample

size be 1000. So, one sample consists of tossing a coin 1000 times and noting the number of heads and tails. Find the mean of the sample (assign 1 to head and 0 to tail). Now, repeat this procedure multiple times to get multiple sample means. When looking at the distribution of all the sample means, we expect it to follow a normal distribution but we don't expect each individual sample to follow a normal distribution. In particular, each sample is a Bernoulli distribution (with n = 1000 and p = 0.5). The sample mean is a normal distribution with mean = 0.5 (you are more likely to get 500 heads than 900 heads out of 1000 tosses).

5.5 Sampling Distribution of the Difference of 2 Sample means

Theorem 5.5.1.
If independent samples of sizes n_1 ($\geq$ 30) and n_2 ($\geq$ 30) are drawn from two populations, with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2 respectively, then the sampling distribution of the differences of the sample means, $\bar{X}_1$ and $\bar{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2 \qquad \textit{and} \qquad \sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

The proof the above theorem is relatively straightforward:

$$\mu_{\bar{X}_1-\bar{X}_2} = E[\bar{X}_1 - \bar{X}_2] = E[\bar{X}_1] - E[\bar{X}_2] = \mu_1 - \mu_2$$

and

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2)$$

since $\bar{X}_1$ and $\bar{X}_2$ are independent. Therefore,

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Since $\bar{X}_1$ and $\bar{X}_2$ are approximately normally distributed, therefore $\bar{X}_1 - \bar{X}_2$ is also approximately normally distributed.

Note 5.5.2.

1. Note that if both $n_1, n_2 \geq 30$, the normal approximation for the distribution of $\bar{X}_1 - \bar{X}_2$ is very good regardless of the shapes of the two population distributions.

2. Recall that if a random variable $Y \sim N(\mu, \sigma^2)$, then $\dfrac{Y-\mu}{\sigma} \sim N(0,1)$. Similarly, here we have

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \text{ approximately } \sim N(0,1)$$

We denote the sample standard deviation as S and the sample variance as S^2. Further, we denote the population standard deviation as σ and population variance as σ^2.

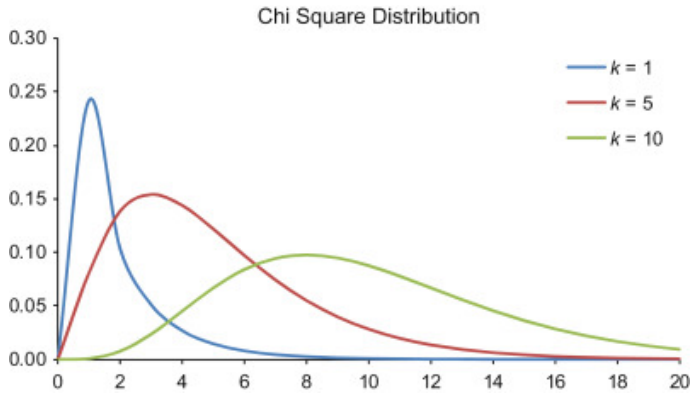


Figure 5.1: Chi-squared distribution

5.6 Chi-Square Distribution

Definition 5.6.1 (Chi-Square Distribution)**.** *If Y is a random variable with p.d.f.*

$$f_Y(y) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1} e^{-y/2}, \text{ for } y > 0 \\ 0, \text{ otherwise} \end{cases}$$

then Y is said to have a chi-squared distribution with n degrees of freedom, denoted by $\chi^2(n)$, where n is a positive integer, and $\Gamma(.)$ is the gamma function.

The gamma function, $\Gamma(.)$ is defined by

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

for $n = 1, 2, 3, \ldots$

Note 5.6.2 (Some Properties of Chi-square distributions)**.**

1. If $Y \sim \chi^2(n)$, then $E[Y] = n$ and $V(Y) = 2n$.
2. For large n, $\chi^2(n)$ approx $\sim N(n, 2n)$.
3. If $Y_1, Y_2, \ldots, Y_k$ are **independent** chi-squared random variables with $n_1, n_2, \ldots, n_k$ degrees of freedom respectively, then $Y_1 + Y_2 + \cdots + Y_K$ has chi-square distribution with $n_1 + n_2 + \cdots + n_k$ degrees of freedom. That is, $\sum_{i=1}^k Y_i \sim \chi^2\left(\sum_{i=1}^k n_i\right)$

Theorem 5.6.3.

1. *If $X \sim N(0,1)$, then $X^2 \sim \chi^2(1)$*
2. *Let $X \sim N(\mu, \sigma^2)$, then $[(X-\mu)/\sigma]^2 \sim \chi^2(1)$*
3. *Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ, and variance σ^2. Define*

$$Y = \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2}$$

Then $Y \sim \chi^2(n)$

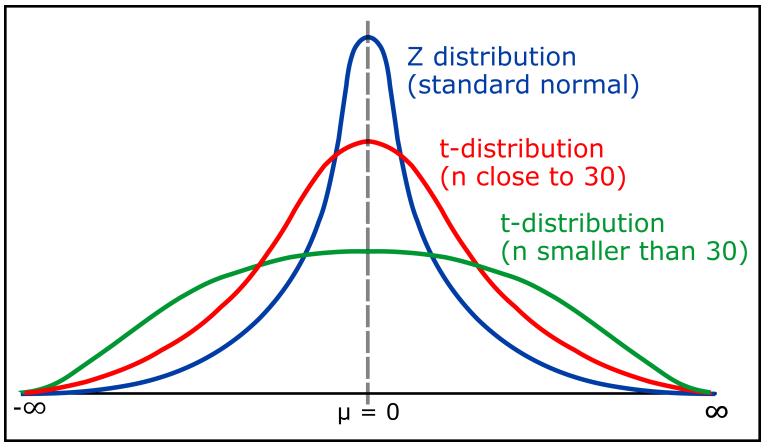


Figure 5.2: t-distribution

Let c be a constant satisfying $P(Y \geq c) = \int_c^\infty f_Y(y)dy = \alpha$, where $Y \sim \chi^2(n)$. We use the notation $\chi^2(n;\alpha)$ to denote this constant c. That is, $P(Y \geq \chi^2(n;\alpha)) = \int_{\chi^2(n;\alpha)}^\infty f_Y(y)dy = \alpha$.
Similarly, $\chi^2(n;1-\alpha)$ is the constant satisfying $P(Y \leq \chi^2(n;1-\alpha)) = \int_0^{\chi^2(n;1-\alpha)} f_Y(y)dy = \alpha$

5.7 The Sampling Distribution of $(n-1)S^2/\sigma^2$

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population. Then the statistic

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is called the sample variance. The sampling distribution of the random variable S^2 has little practical application in statistics.
Instead, we shall consider the sampling distribution of the random variable $\frac{(n-1)S^2}{\sigma^2}$ when $X_i \sim N(\mu, \sigma^2)$ for all i.

Theorem 5.7.1.
*If S^2, is the variance of a random sample of size n taken from a **normal** population having the variance σ^2, then the random variable $\frac{(n-1)S^2}{\sigma^2}$ has a **chi-squared distribution with** $n-1$ **degrees of freedom.** That is, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$*

5.8 The t-distribution

Definition 5.8.1 (t-distribution). *Suppose $Z \sim N(0,1)$ and $U \sim \chi^2(n)$. If Z and U are **independent**, and let $T = \frac{Z}{\sqrt{U/n}}$ then the random variable T follows the t-distribution with n degrees of freedom. That is, $\frac{Z}{\sqrt{U/n}} \sim t(n)$*

If T follows a t-distribution with n degrees of freedom, then its p.d.f. is given by

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma(n/2)}\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}, \qquad -\infty < t < \infty$$

where the gamma function is defined as earlier.

Note 5.8.2.

1. The graph of the t-distribution is symmetric about the vertical axis and resembles the graph of the standard normal distribution.
2. It can be shown that the p.d.f. of t-distribution with n d.f. (degrees of freedom) is approaching to the p.d.f. of standard normal distribution when $n \to \infty$. That is,

   $$\lim_{n\to\infty} f_T(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$$

   as $n \to \infty$
3. The values of $P(T \geq t) = \int_t^\infty f_T(x)dx$ for selected values of n and t are given in a statistical table.
4. If $T \sim t(n)$, then $E[T] = 0$ and $V(T) = \frac{n}{n-2}$ for $n > 2$.

Note 5.8.3.

If the random sample was selected from a normal population, then

$$Z = \frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} \sim N(0,1)$$

and

$$U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

It can be shown that $\bar{X}$ and S^2 are independent, and so are Z and U. Therefore,

$$\begin{aligned} T &= \frac{\bar{X}-\mu}{S/\sqrt{n}} \\ &= \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} \sim t_{n-1} \end{aligned}$$

That is, T has a t-distribution with $n-1$ d.f. (degrees of freedom). Observe the relation between Z, U, T closely. If you replace the σ (Population standard deviation) in Z with S (Sample standard deviation), the distribution changes from standard normal to t-distribution with $n-1$ degrees of freedom.

5.9 The F-distribution

Definition 5.9.1 (The F-distribution)**.** *Let U and V be independent random variables having $\chi^2(n_1)$ and $\chi^2(n_2)$ respectively. Then, the distribution of the random variable $F = \frac{U/n_1}{V/n_2}$, is called a F-distribution with (n_1, n_2) degrees of freedom.*

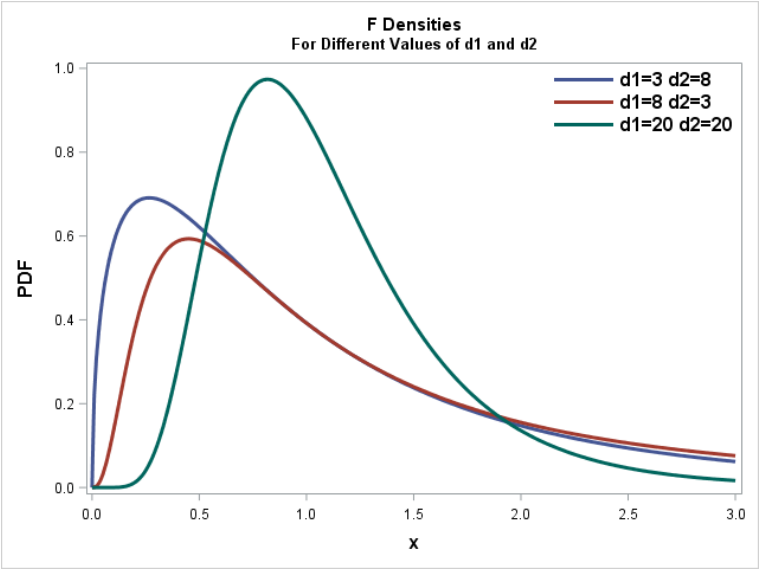


Figure 5.3: F-distribution

Observe that F is the ratio of 2 χ^2 random variables, each adjusted by its respective degrees of freedom. Moreover, the parameters of the F-distribution are precisely the degrees of freedom of the 2 χ^2 random variables (first of the random variable in the numerator, then that of the one in the denominator)
The p.d.f of F is given by

$$f_F(x) = \begin{cases} \dfrac{n_1^{n_1/2} n_2^{n_2/2} \Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \dfrac{x^{n_1/2-1}}{(n_1x+n_2)^{(n_1+n_2)/2}}, & \textit{for } x > 0 \\ 0, & \textit{otherwise} \end{cases}$$

It can be shown that

$$E[X] = \frac{n_2}{n_2-2}, \text{ with } n_2 > 2$$
$$V(X) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}, \text{ with } n_2 > 4$$

Theorem 5.9.2.
If $F \sim F(n,m)$, then $\dfrac{1}{F} \sim F(m,n)$

This theorem follows immediately from the definition of F-distribution. Recall that $F = \dfrac{U/n_1}{V/n_2}$. Then, $\dfrac{1}{F} = \dfrac{V/n_2}{U/n_1}$ follows $F(n_2, n_1)$.
Values of the F-distribution can be found in the statistical tables. The table gives the values of $F(n_1, n_2; \alpha)$ such that $P(F > F(n_1, n_2; \alpha)) = \alpha$. For example, $F(5, 4; 0.05) = 6.26$ means that $P(F > 6.26) = 0.05$, where $F \sim F(5, 4)$.

Theorem 5.9.3.

$$F(n_1, n_2; 1-\alpha) = \frac{1}{F(n_2, n_1; \alpha)}$$

Here is a short proof: Let $F \sim F(n_1, n_2)$, then $\frac{1}{F} \sim F(n_2, n_1)$, Based on the definition of $F(n_1, n_2; 1-\alpha)$,

$$P(F > F(n_1, n_2; 1-\alpha)) = 1-\alpha,$$

which leads to

$$P(F < F(n_1, n_2; 1-\alpha)) = 1 - P(F > F(n_1, n_2; 1-\alpha)) = \alpha$$

That is,

$$P\left(\frac{1}{F} > \frac{1}{F(n_1, n_2; 1-\alpha)}\right) = \alpha,$$

which together with the fact that $\frac{1}{F} \sim F(n_2, n_1)$ implies

$$\frac{1}{F(n_1, n_2; 1-\alpha)} = F(n_2, n_1, \alpha)$$

Note 5.9.4.

Checking the definition of t and F distributions, we find one important connection between them: If $Y \sim t(n)$, then $Y^2 \sim F(1, n)$

5.10 Summary of Sampling Distributions

1. If $X_1, X_2, \ldots, X_n$ are $N(\mu, \sigma^2)$, then $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ follows $N(0,1)$ regardless of the sample size n. If $X_1, X_2, \ldots, X_n$ have mean μ and finite variance σ^2 and n is sufficiently large, then $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ approximately follows the $N(0,1)$ standard normal distribution.
2. If $X_1, X_2, \ldots, X_n$ are $N(\mu, \sigma^2)$, then $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, where S^2 is the sample variance of the random sample.
3. If $X_1, X_2, \ldots, X_n$ are $N(\mu, \sigma^2)$, then $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$.
4. Let $X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$ be $N(\mu_1, \sigma_1^2)$ and let $X_{2,1}, X_{2,2} \ldots, X_{2,n_2}$ be $N(\mu_2, \sigma_2^2)$. Denote by S_1^2 and S_2^2 the sample variances of $X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$ and $X_{2,1}, X_{2,2} \ldots, X_{2,n_2}$ respectively. Then, we have $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$.

Chapter 6

Estimation Based on Normal Distribution

6.1 Point Estimation of Mean and Variance

6.1.1 Introduction

Assume that some characteristics of the elements in a population can be represented by a random variable X whose p.d.f. (or p.f.) is $f_X(x;\theta)$, where the form of the probability density function (or probability function) is assumed known except that it contains an unknown parameter θ. Further assume that the values $x_1, x_2, \ldots, x_n$ of a random sample $X_1, X_2, \ldots, X_n$ from $f_X(x;\theta)$ can be observed. On the basis of the observed sample values $x_1, x_2, \ldots, x_n$,, it is desired to estimate the value of the unknown parameter θ.

6.1.2 Estimation

The estimation can be made in 2 ways: **Point estimation** and **Interval estimation**.
Point estimation is to let the value of some statistic, say

$$\hat{\Theta} = \hat{\Theta}(X_1, X_2, \ldots, X_n),$$

to estimate the unknown parameters θ.
Such a statistic $\hat{\Theta}(X_1, X_2, \ldots, X_n)$ is called a **point estimator**.
Recall that a statistic is a function of the random sample which does not depend on any unknown parameters. Examples of statistic include $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ or $X_{(n)} = max(X_1, X_2, \ldots, X_n)$.
Let $W = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$. Then W is not a statistic if μ is not known. However, W is a statistic if μ is known.

6.1.3 Point Estimate of Mean

Suppose μ is the population mean. The statistic that one uses to obtain a point estimate is called an estimator. For example, $\bar{X}$ is an estimator of μ. The value of $\bar{X}$, denoted by $\bar{x}$, is an estimate of μ.
It should not come as a surprise that different random samples give different point estimates of μ.

Note 6.1.1.

It is important to distinguish clearly these three concepts: an estimator/statistic (e.g. $\bar{X}$), an estimate (e.g. $\bar{x}$), and a population parameter (e.g.μ). An estimator/statistic is a computational rule. It is also a random variable. When the data (random sample) are available, it tells us how to compute. An estimate is a computed value of the estimator based on the observed data (random

sample). It is not a random variable - it is a particular realization of the random variable. A population parameter is something about the population - it is not a random variable. Even if you do not know the population parameter, it is an unknown constant, NOT a random variable.

6.1.4 Interval Estimation

Interval estimation is to define two statistics, say, $\hat{\Theta_L}$ and $\hat{\Theta_U}$, where $\hat{\Theta_L} < \hat{\Theta_U}$ so that $(\hat{\Theta_L}, \hat{\Theta_U})$ constitutes a random interval for which the probability of containing the unknown parameter θ can be determined.
For example, suppose σ^2 is known. Let

$$\hat{\Theta_L} = \bar{X} - 2\frac{\sigma}{\sqrt{n}} \text{ and } \hat{\Theta_R} = \bar{X} + 2\frac{\sigma}{\sqrt{n}}$$

Then, $\left(\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right)$ is an interval estimator for μ.

6.1.5 Biased and Unbiased Estimators

Definition 6.1.2 (Unbiased estimator)**.** *A statistic $\hat{\Theta}$ is said to be an unbiased estimator of the parameter θ if $E[\hat{\Theta}] = \theta$.*

For example, $\bar{X}$ is an unbiased estimator of μ. That is, $E[\bar{X}] = \mu$.
Also, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimator of σ^2. That is, $E[S^2] = \sigma^2$.
However, $T = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a **biased** estimator of σ^2. It can be shown that $E[T] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$

Note 6.1.3.

Here is a short derivation for $E[S^2] = \sigma^2$:

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + \bar{X}^2) \\
&= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \text{ (because } \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \text{)} \\
&= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \\
Now, E[S^2] &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] \text{ (based on the result above)} \\
&= \frac{1}{n-1}\left(E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\bar{X}^2]\right) \text{ (linearity of expectation)} \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} E[X_i^2] - nE[\bar{X}^2]\right) \text{ (linearity of expectation)} \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}[V(X_i) + (E[X])^2] - n[V(\bar{X}) + (E[\bar{X}])^2]\right) \text{ (because } V(Y) = E[Y^2] - (E[Y])^2 \text{)} \\
&= \frac{1}{n-1}\left(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right) \text{ (because } E[X] = \mu, V(X) = \sigma^2, E[\bar{X}] = \mu, V(\bar{X}) = \frac{\sigma^2}{n} \text{)} \\
&= \frac{1}{n-1}(n-1)\sigma^2 \\
&= \sigma^2
\end{aligned}$$

Note 6.1.4.

Let $X_1, X_2, \ldots, X_n$ be a random sample. Then, observe that $E(X_1) = \mu$, i.e., X_1 is also an unbiased estimator for μ. So, why do we bother to use $\bar{X}$ instead of simply using X_1?
The answer is quite simple. The variance of X_1 is σ^2. The variance of $\bar{X}$ is $\frac{\sigma^2}{n}$ where n is the sample size. This means that $\bar{X}$ varies less than X_1 and so it makes more sense to use that to estimate the population mean. This also provides some intuition as to why the variance of $\bar{X}$ is $\frac{\sigma^2}{n}$. We know that μ is a constant and has variance = 0. So, the larger the sample size, the better the estimate (closer to constant μ) and hence, lower the variability. We're taking the mean of n random variables and so, the variation reduces too.

Note 6.1.5.

Our aim is to find an estimator which is unbiased and has a low variance. But even if a biased estimator has a very low variance, it is not a good estimator. Being unbiased matters more than having low variance. Accuracy (how close your estimate is to the actual value) is more important than precision (how close your estimates are to each other). Consistency matters only when you're actually estimating well. In practice, we choose the estimator with the lowest MSE (Mean Square Error) - We try to minimize the mean of the squares of the error between our estimate and the actual value. For an unbiased estimator, the MSE = variance.

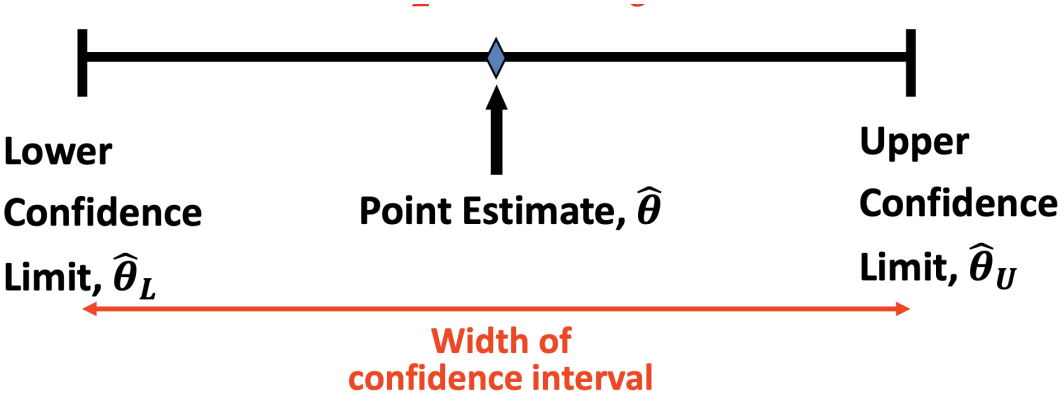


Figure 6.1: Interval Estimation

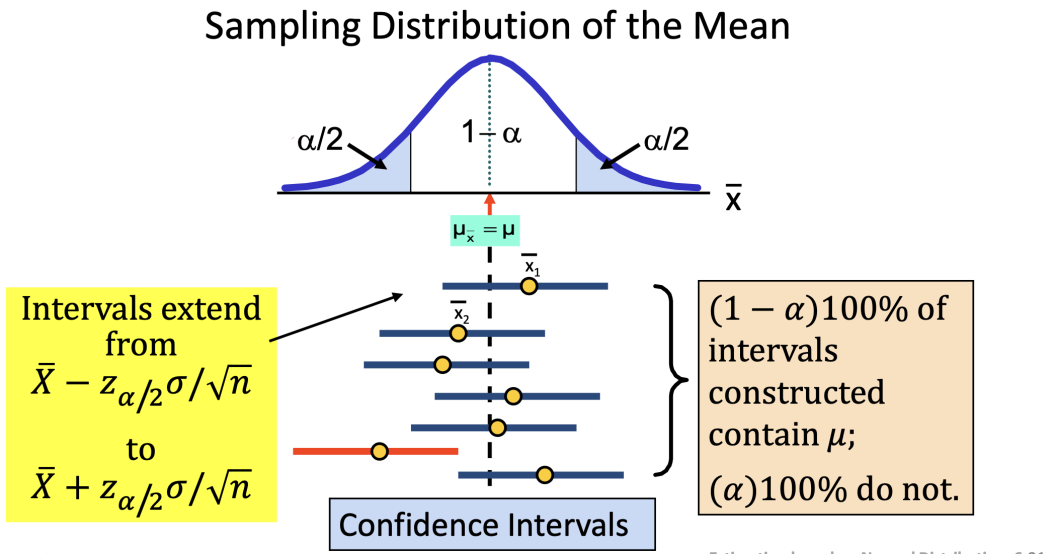


Figure 6.2: Sampling Distribution of the Mean

6.2 Interval Estimation

An interval estimate of a population parameter θ is an interval of the form $\hat{\theta_L} < \theta < \hat{\theta_U}$, where $\hat{\theta_L}$ and $\hat{\theta_U}$ depend on

1. The value of the statistic $\hat{\Theta}$ for the particular sample, and
2. The sampling distribution of $\hat{\Theta}$

Since different samples will generally yield different values of $\hat{\Theta}$, and therefore different values of $\hat{\theta_L}$ and $\hat{\theta_U}$, these end points of the interval are values of corresponding random variables $\hat{\Theta_L}$ and $\hat{\Theta_U}$.
These intervals may not contain the parameter θ as $\hat{\theta_L}$ and $\hat{\theta_U}$ vary.
We shall seek a random interval $(\hat{\Theta_L}, \hat{\Theta_U})$ containing θ with a given probability $1-\alpha$. That is, $P(\hat{\Theta_L} < \theta < \hat{\Theta_U}) = 1-\alpha$. Then the interval $\hat{\theta_L} < \theta < \hat{\theta_U}$ computed from the selected sample is called a $(1-\alpha)100\%$ confidence interval for θ and the fraction $(1-\alpha)$ is called the **confidence coefficient** or **degree of confidence**, and the endpoints $\hat{\theta_L}$ and $\hat{\theta_U}$ are called the lower and upper confidence limits respectively.
This means that if samples of the same size n are taken, then in the long run, $(1-\alpha)100\%$ of the intervals will contain the unknown parameter θ, and hence with a confidence of $(1-\alpha)100\%$, we can say that the interval covers θ.

Note 6.2.1.

It is important to differentiate the random interval from the interval estimate. Once you have realized the values of the upper and lower bounds of the interval, it does not make sense to talk about the probability of a population parameter lying within the interval. The population parameter is an unknown constant. It either does or does not lie within the computed interval - there is no notion of probability involved. Practically, for a computed C.I., we can only claim that with a certain confidence the interval will cover the true value. This is the reason we choose to say "confidence" rather than "probability".

Note 6.2.2. *The general form a confidence interval of a population parameter is given by:*

$$\textbf{\textit{point estimate}} \pm (\textbf{\textit{multiplier}} \times \textbf{\textit{standard error}})$$

where the standard error is the standard deviation of the point estimate and the multiplier is a constant that depends on the level of confidence.

6.3 Confidence Interval (C.I.) for Population Proportion

We have shown that when n is sufficiently large such that $n\hat{p}(1-\hat{p}) \geq 5$, then $\hat{p} \sim N(\hat{p}, \sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$.

So, if $\hat{p}$ is the sample proportion taken from a sample of size n, a $(1-\alpha)100\%$ CI for population proportion p is given by:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Recall that $P(Z \geq z_\alpha) = \alpha$ as per our definition of z_α.

6.4 Confidence Interval (C.I.) for the Mean

6.4.1 Known Variance Case

Assume that we know the population variance and we are trying to estimate the population mean. Further, we also know that the population distribution is normal or n is sufficiently large (say $n \geq 30$).

Then, by the Central Limit Theorem, we can expect that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Therefore, $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Hence,

$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

or

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Hence, if $\bar{X}$ is the mean of a random sample of size n from a population with known variance σ^2, a $(1-\alpha)100\%$ confidence interval for μ is given by

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

6.4.2 Sampling Size for Estimating μ

Most of the time, $\bar{X}$ will not be exactly equal to μ and the point estimate is in error. The size of this error will be $|\bar{x} - \mu|$. We know that $P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$. In other words,

$$P\left(|\bar{X} - \mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Let e denote the margin of error. We want the error $|\bar{X} - \mu|$ does not exceed the margin of error, e, with a probability larger than $1 - \alpha$. That is,

$$P(|\bar{X} - \mu| \leq e) \geq 1 - \alpha$$

Since $P\left(|\bar{X} - \mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$, therefore

$$e \geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Hence for a given margin of error e, the sample size is given by

$$n \geq \left(z_{\alpha/2}\frac{\sigma}{e}\right)^2$$

Note 6.4.1.

It is important to understand the implications of the above formula. The required sample size depends on:

1. $z_{\alpha/2}$ (and hence on α). In particular, lower the value of α, higher the value of $z_{\alpha/2}$. So, for a higher degree of confidence, we need to have a larger sample size (as expected)
2. σ - If the underlying distribution shows higher variation, it is necessary to have a larger sample size to have the same margin of error and degree of confidence.
3. e - To have a lower margin of error, you need to have a higher sample size.

6.4.3 Unknown Variance Case

We now try to find the confidence interval for mean with:

1. unknown population variance
2. the population is normal or very close to normal distribution
3. the sample size is small (so we cannot use central limit theorem)

Let $T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$ where S^2 is the sample variance. We know that $T \sim t_{n-1}$. Hence,

$$\begin{aligned}
P\left(-t_{n-1;\alpha/2} < T < t_{n-1;\alpha/2}\right) &= 1 - \alpha \\
P\left(-t_{n-1;\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{n-1;\alpha/2}\right) &= 1 - \alpha \\
P\left(-t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} < \bar{X} - \mu < t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right) &= 1 - \alpha \\
P\left(\bar{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right) &= 1 - \alpha
\end{aligned}$$

So, If $\bar{X}$ and S are the sample mean and the standard deviation of a random sample of size $n < 30$ from an approximate normal population with unknown variance σ^2, a $(1-\alpha)100\%$ confidence interval for μ is given by

$$\bar{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}$$

For large n (say, $n > 30$), the t-distribution is approximately the same as the $N(0,1)$ distribution. Hence, we can replace $t_{n-1;\alpha/2}$ by $z_{\alpha/2}$. So, when σ^2 is unknown, population is normal and $n > 30$, a $(1-\alpha)100\%$ confidence interval is given by

$$\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}$$

6.5 Confidence Intervals (C.I.) for the Difference between 2 Means

If we have 2 populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively, then $\bar{X}_1 - \bar{X}_2$ is the point estimator of $\mu_1 - \mu_2$.

6.5.1 Known Variances

If σ_1^2 and σ_2^2 are known and not equal, and the two populations are normal, or when σ_1^2 and σ_2^2 are known and not equal but n_1, n_2 are sufficiently large ($n_1 \geq 30, n_2 \geq 30$), then we know that

$$(\bar{X}_1 - \bar{X}_2) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

We can assert that

$$P\left(-z_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{\alpha/2}\right) = 1 - \alpha$$

which leads to the following $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

6.5.2 Large Sample Confidence Interval (C.I.) for Unknown Variances

We use this when:

1. σ_1^2 and σ_2^2 are unknown
2. n_1, n_2 are sufficiently large ($n_1 \geq 30, n_2 \geq 30$)
3. We may replace σ_1^2 and σ_2^2 by their estimates S_1^2 and S_2^2

Then, a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

6.5.3 Unknown but Equal Variances

We use this when:

1. σ_1^2 and σ_2^2 are unknown but equal
2. The two populations are normal
3. Sample sizes are small ($n_1 \leq 30, n_2 \leq 30$)

Let $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$(\bar{X}_1 - \bar{X}_2) \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Therefore we obtain a standard normal random variable in the form

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

We can estimate σ^2 by the pooled sample variance (essentially just taking the weighted average):

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

where S_1^2 and S_2^2 are the sample variances of the first and second samples respectively.
Remember that S_p^2 is an estimator for the population variance, and NOT an estimator for the variance of the difference of the means. To get an estimate for the variance of the difference of the means, you still need to multiply S_p^2 by $\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

Note 6.5.1.

The above formula should intuitively make sense: every observation contributes equally in the estimation of their common variance σ^2. In terms of samples, it is the weighted average of the 2 sample variances with the weights being one less than the sample sizes.

Note that if the two populations are normal with the same variance σ^2, then $\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2_{n_1 - 1}$ and $\frac{(n_2 - 1)S_1^2}{\sigma^2} \sim \chi^2_{n_2 - 1}$.
Hence,

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_1 + n_2 - 2}$$

Substituting S_p^2 for σ^2, we obtain the statistic:

$$T = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2}$$

We can assert that

$$P\left(-t_{n_1+n_2-2;\alpha/2} < \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < t_{n_1+n_2-2;\alpha/2}\right) = 1 - \alpha$$

Therefore a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{X}_1 - \bar{X}_2) - t_{n_1+n_2-2;\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{n_1+n_2-2;\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where S_p is the pooled estimate of the population standard deviation and $t_{n_1+n_2-2;\alpha/2}$ is the value from the t-distribution with the degrees of freedom $n_1 + n_2 - 2$ leaving an area of $\alpha/2$ to the right. In other words, $P(W > t_{n_1+n_2-2;\alpha/2}) = \alpha/2$ where $W \sim t_{n_1+n_2-2}$.

6.5.4 C.I. for the difference between 2 means for paired data (dependent data)

Say for example, we run a test on a new diet using 15 individuals, the weights before (x_i) and after (y_i) the completion of the diet form our two samples. Observations in the two samples made on the same individual are related and hence, form a pair. To determine if the diet is effective, we must consider the differences $d_i = x_i - y_i$ of paired observations.
These differences are the values of the random sample $d_1, d_2, \ldots, d_n$ from a population that we shall assume to be normal with mean μ_D and unknown variance σ_D^2. In fact, $\mu_D = \mu_1 - \mu_2$ (by linearity) and the point estimate of μ_D is given by

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)$$

. The point estimate of σ_D^2 is given by

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

For a small sample and approximately normal population, a $(1-\alpha)100\%$ confidence interval for μ_D can be established as follows:

$$P(-t_{n-1;\alpha/2} < T < t_{n-1;\alpha/2}) = 1 - \alpha$$

where $T = \dfrac{\bar{d} - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$
Therefore, a $(1-\alpha)100\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is given by:

$$\bar{d} - t_{n-1;\alpha/2}\frac{S_D}{\sqrt{n}} < \mu_D < \bar{d} + t_{n-1;\alpha/2}\frac{S_D}{\sqrt{n}}$$

For a large sample ($n > 30$), we may replace $t_{n-1;\alpha/2}$ by $z_{\alpha/2}$ and a $(1-\alpha)100\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is given by:

$$\bar{d} - z_{\alpha/2}\frac{S_D}{\sqrt{n}} < \mu_D < \bar{d} + z_{\alpha/2}\frac{S_D}{\sqrt{n}}$$

Note 6.5.2.

Here is a general strategy for constructing mean related confidence intervals. Suppose we are to construct a $(1-\alpha)$ confidence interval for mean related parameter θ (e.g. θ could be μ, $\mu_1 - \mu_2$, or other possible combinations of the population means). Then, the following are the steps you need to follow:

1. Look for an estimator $\hat{\theta}$ for θ, e.g. $\bar{X}$ for μ, $\bar{X}_1 - \bar{X}_2$ for $\mu_1 - \mu_2$.

2. Derive the variance $V(\hat{\theta})$.
3. Construct $(1-\alpha)$ C.I. to be $\hat{\theta} \pm M\sqrt{V}$ where M is called the multiplier, and V is related to $V(\hat{\theta})$. The following is how they are determined:
   (a) If $V(\hat{\theta})$ does not depend on any other parameter (e.g. in the case σ^2 is known, $V(\bar{X}) = \sigma^2/n$), $V = V(\bar{\theta})$, and $M = z_{\alpha/2}$. Here we may need the condition that the data are normal and/or the sample size n is big.
   (b) If the derived $V(\hat{\theta})$ contains some unknown parameter, e.g., σ^2, we replace the parameter with its estimator, e.g. we use S^2 to replace σ^2; this results in $\hat{V}(\hat{\theta})$. Then, we use $V = \hat{V}(\hat{\theta})$, however M has 2 possibilities:
      i. If the sample size n is sufficiently large, $M = z_{\alpha/2}$.
      ii. If the sample size n is not large, but the data are normally distributed, $M = t(df, \alpha/2)$. Here df = degrees of freedom, which is the d.f. of the estimator for the parameter contained in $V(\hat{\theta})$.

6.6 C.I. for Variances and Ratio of Variances

6.6.1 C.I. for a variance of a normal population

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a approximately normal $N(\mu, \sigma^2)$ distribution. Then the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$

is a point estimate of σ^2

Case 1: When μ is known

When μ is known, we have $\dfrac{X_i - \mu}{\sigma} \sim N(0,1)$ for all i. Also, $\left(\dfrac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(1)$ for all i. Hence, $\sum_{i=1}^{n} \dfrac{(X_i - \mu)^2}{\sigma^2} \sim \chi^n_2$.
Therefore,

$$P\left(\chi^2_{n;1-\alpha/2} < \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} < \chi^2_{n;\alpha/2}\right) = 1 - \alpha$$

Rearranging the inequalities with σ^2 on one side, we get

$$P\left(\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{n;\alpha/2}} < \sigma^2 < \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{n;1-\alpha/2}}\right) = 1 - \alpha$$

Note that $\chi^2_{n;\alpha/2}$ satisfies $P(W > \chi^2_{n;\alpha/2}) = \dfrac{\alpha}{2}$, where $W \sim \chi^2(n)$.
Therefore, a $(1-\alpha)100\%$ confidence interval for σ^2 of $N(\mu, \sigma^2)$ population with μ known is

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{n;\alpha/2}} < \sigma^2 < \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{n;1-\alpha/2}}$$

Case 2: μ is unknown

When μ is unknown, we have

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

The above result is true for both small and large n. Therefore,

$$P\left(\chi^2_{n-1;1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1;\alpha/2}\right) = 1-\alpha$$

and hence,

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}\right) = 1-\alpha$$

Therefore, a $(1-\alpha)100\%$ confidence interval for σ^2 for $N(\mu, \sigma^2)$ a population with μ unknown is

$$\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}$$

where S^2 is the sample variance

Note 6.6.1.

A $(1-\alpha)100\%$ confidence interval for σ is obtained by taking the square root of each end point of the interval for σ^2.
Therefore, when μ is known, a $(1-\alpha)100\%$ C.I. for σ is

$$\sqrt{\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n;\alpha/2}}} < \sigma < \sqrt{\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n;1-\alpha/2}}}$$

When μ is unknown, a $(1-\alpha)100\%$ C.I. for σ is

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}}$$

Notice that the parameter (or the degrees of freedom) of the χ^2-distribution changes from n to $n-1$ when μ is unknown.

6.6.2 C.I. for the ratio of 2 variances of normal population with unknown means

Let $X_1, X_2, \ldots, X_{n_1}$ be a random sample of size n_1 from a (or approximately) normal $N(\mu_1, \sigma_1^2)$ population and $Y_1, Y_2, \ldots, Y_{n_2}$ be a random sample of size n_2 from a (or approximately) normal $N(\mu_2, \sigma_2^2)$ population.
Then, $\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1)$ and $\frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$ where $S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\bar{X})^2$ and $S_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_i-\bar{Y})^2$. Hence

$$F = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$

We can then assert that

$$P\left(F_{n_1-1,n_2-1;1-\alpha/2} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{n_1-1,n_2-1;\alpha/2}\right) = 1-\alpha$$

Therefore,

$$P\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1;\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1;1-\alpha/2}}\right) = 1-\alpha$$

where $P(F_{n_1-1,n_2-1} \geq F_{n_1-1,n_2-1;\alpha/2}) = \alpha/2$ with F_{n_1-1,n_2-1} denotes a random variable following an F-distribution with parameters (n_1-1) and (n_2-1).

Hence, a $(1-\alpha)100\%$ confidence interval for the ratio $\frac{\sigma_1^2}{\sigma_2^2}$ when μ_1 and μ_2 are unknown is

$$\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1;\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}F_{n_2-1,n_1-1;1-\alpha/2}$$

since $F_{n_1-1,n_2-1;1-\alpha/2} = \dfrac{1}{F_{n_2-1,n_1-1;\alpha/2}}$

Note 6.6.2.

A $(1-\alpha)100\%$ confidence interval for $\frac{\sigma_1}{\sigma 2}$ is obtained by taking the square of each end point of the interval for $\frac{\sigma_1^2}{\sigma_2^2}$.

So, when μ_1 and μ_2 are unknown, a $(1-\alpha)100\%$ confidence interval for $\frac{\sigma_1}{\sigma 2}$ is

$$\sqrt{\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1;\alpha/2}}} < \frac{\sigma_1}{\sigma_2} < \sqrt{\frac{S_1^2}{S_2^2}F_{n_2-1,n_1-1;1-\alpha/2}}$$

Chapter 7

Hypothesis Testing based on Normal Distribution

7.1 Null and Alternative Hypotheses

Definition 7.1.1 (Statistical Hypothesis)**.** *A statistical hypothesis is an assertion or conjecture concerning one or more populations*

It is important to understand that the rejection of a hypothesis is to conclude that it is false, while the acceptance of a hypothesis merely implies that we have insufficient evidence to believe otherwise. Because of this terminology, the statistician or experimenter will often chose to state the hypothesis in a form that will hopefully be rejected.

Definition 7.1.2 (Null Hypothesis)**.** *Hypothesis that we formulate with the hope of rejecting, denoted by H_0. A null hypothesis concerning a population parameter will always be stated to specify an exact value of the parameter*

Definition 7.1.3 (Alternative Hypothesis)**.** *The rejection of H_0 leads to the acceptance of an alternative hypothesis, denoted by H_1. It allows for the possibility of several values.*

For example, if we wish to determine whether the mean IQ of the pupils of a certain school is different from 100, we could set $H_0 : \mu = 100$ against $H_1 : \mu \neq 100$. This is called a two-sided alternative. The test is called a two-sided (or two-tailed) test. We may like to test whether the mean IQ of the pupils is greater than 100 (or less than 100). This is called a one-sided alternative and the test is called a one-sided (or one-tailed) test. That is, $H_0 : \mu = 100$ against $H_1 : \mu > 100$ or $H_0 : \mu = 100$ against $H_1 : \mu < 100$.
The procedure for statistical inference is quite simple. We choose a random sample and calculate the sample parameter. Then, we calculate how likely it was possible to obtain a value that was as favourable to our alternative hypothesis assuming that the null hypothesis was true. If the probability of this occurring is low, then we can reject the null hypothesis. Typically, we set the value of statistical significance at 5%. That is, if the probability of obtaining such a result (or more favourable is less than 5%, we can reject the null hypothesis.
The whole framework of hypothesis testing is to look for evidence to reject the null hypothesis. (Instead, if the framework was to find evidence to reject the alternative hypothesis and you were not able to find such evidence, you might conclude that the alternative hypothesis is true. But others may argue that you simply did not dig deep enough for evidence.)

Note 7.1.4.

It is worth understanding the difference between "accept" and "do not reject" in statistics. There are only two kinds of theories in the world - false theories, and theories that are yet to be proved false. For example, you hypothesise that there are no black swans on Earth. Obviously if

you see a single black swan, your hypothesis immediately fails. On the other hand, if you don't see any black swan for 10 years, you still cannot be sure that there is no black swan (someone may argue that you did not search hard enough). In this sense, we can say that we do not have sufficient evidence to reject the hypothesis that there are no black swans on Earth (however, we have not yet established its truth).
Hence, we set our null hypothesis to be the status quo (what is currently believed) and pit it against our alternative hypothesis (what we wish to establish).

Note 7.1.5.

By convention (and theoretically supported), we usually write the null hypothesis in the "equal" form, i.e., $H_0 : \theta = \theta_0$, with θ_0 being a fixed constant. So the hypotheses have three possible forms:

- $H_0 : \theta = \theta_0$ versus $H_0 : \theta > \theta_0$; in this case $H_0 : \theta = \theta_0$ in fact means $\theta \leq \theta_0$
- $H_0 : \theta = \theta_0$ versus $H_0 : \theta < \theta_0$; in this case $H_0 : \theta = \theta_0$ in fact means $\theta \geq \theta_0$
- $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$

So, we need to check the form of H_1 to ensure the meaning of $H_0 : \theta = \theta_0$ in practice.

7.1.1 Types of Error

There are two possible true answers (we call it the state of nature) which we don't know, i.e., we don't know if H_0 is true or H_1 is true.
A **Type I** error is when you reject H_0 even when H_0 is actually true. It is considered a serious

	State of Nature	
Decision	H_0 is true	H_0 is false
Reject H_0	**Type I error** Pr(Reject H_0 given that H_0 is true) = α	Correct decision Pr(Reject H_0 given that H_0 is false) = 1 – β
Do not reject H_0	Correct decision Pr(Do not reject H_0 given that H_0 is true) = 1 – α	**Type II error** Pr(Do not reject H_0 given that H_0 is false) = β

Figure 7.1: Types of Error

type of error. Many statisticians and experimenters fabricate data just to be able to reject the null hypothesis (and persuade people that their theory is correct).
A **Type II** error is not rejecting H_0 when H_0 is false. When we say that H_0 is false, we actually mean that H_1 is true.
Here, α is the level of significance. α is equal to the probability of making a type I error, i.e., the probability of rejecting H_0 even when it is true. Formally, $\alpha = P(\text{reject } H_0 | H_0 \text{ is true})$.
α is set by the researcher in advance (it is usually set at 5% or 1%).
β is the probability of committing a type II error, i.e., the probability of not rejecting H_0 even when H_0 is false. Formally, β = $P(\text{do not reject } H_0 | H_1)$.

$$1 - \beta = \textbf{Power of a test} = P(\textit{reject } H_0 | H_1)$$

Note that it is not possible to determine the probability of committing a type II error, denoted by β, unless we have a specific alternative hypothesis.

Note 7.1.6.

Type I error is usually treated more seriously (since it causes a change in the status quo); so we need to control it in the first place. We set a significance level, called α, and require that the decision rule satisfy $P(\text{reject } H_0|H_0 \text{ is true}) \leq \alpha$. Once we have controlled our type I error (upper bound it), we try to minimize type II error (so that our test is powerful) and so, we typically require $P(\text{reject } H_0|H_0 \text{ is true}) = \alpha$. This is so that the type I error is controlled but maximized (up to an acceptable level) to reduce the type II error. Notice that when establishing a testing rule, type I and type II errors trade off each other. This similar to the trade-off between the confidence level and the width of a confidence interval - it is possible to gain confidence by expanding our interval but then it becomes meaningless. For example, if you set $\alpha = 0.0001$, i.e., your chances of making a type I error is 0.01% but then your decision rule will commit a type II error with large probability.

Note 7.1.7.

Observe that when a test is performed, one can make at most one type of error since if H_0 is rejected, type I error is possible and if H_0 is not rejected, then type II error is possible. These are disjoint events.

Note 7.1.8.

Hypothesis testing cannot be used to determine the truth value of any statement. It is a decision making process. Even if you reject the null hypothesis, it does not mean that the null hypothesis is false. It simply means that the evidence suggested that it was unlikely for the null hypothesis to be true. In any case, you might have made an error (which you will not find out until new evidence comes to light).

Note 7.1.9.

The hypotheses should not contain any random variables since it must be a statement with a fixed truth value. For example, $H_0 : \bar{X} < 10$ is not a valid null hypothesis.

Note 7.1.10. *The 5 steps involved in hypothesis testing are:*

1. *State assumptions*
   For population proportion:
   (a) *variable measured is categorical*
   (b) *data is obtained from randomisation*
   (c) n *is sufficiently large such that* $np_0(1-p_0) \geq 5$ *where* $H_0 : p = p_0$

   For population mean(s),
   (a) *variable measured is quantitative*
   (b) *data is obtained from randomisation*
   (c) *population distribution is approximately normal*
2. *State null and alternative hypotheses*
3. *State test statistic, null distribution and observed value*
4. *State p-value and interpret*
5. *Conclusion to reject/not reject* H_0 *at* α

Statistical significance refers to the claim that a result from data generated by testing or experimentation is not likely to occur randomly or by chance but is instead likely to be attributable to a specific cause. A p-value less than 0.05 is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis.

7.1.2 Acceptance and Rejection Regions

To test a hypothesis about a population parameter, we first select a suitable test statistic for the parameter under hypothesis. Once the significance level, α, is given, a decision rule can be found such that it divides the set of all possible values of the test statistic into 2 regions, one being the rejection region (or critical region) and the other the acceptance region.
Once a sample is taken, the value of the test statistic is obtained. If the test statistic assumes a value in the rejection region, the null hypothesis is rejected; otherwise it is not rejected.
The value that separates the rejection and acceptance regions is called the critical value.

7.2 Hypothesis Testing Concerning Mean

7.2.1 Known Variance

Consider the problem of testing the hypothesis concerning the mean, μ, of a population with:

1. Variance, σ^2, known
2. Underlying distribution is normal or n is sufficiently large (say $n > 30$)

Two-sided Test

Test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.
When the population is normal or the sample size is large (then by the central limit theorem), we can expect that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Hence under $H_0 : \mu = \mu_0$, we have $\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$.

Critical Value Approach

By using a significance level of α, it is possible to find two critical values $\bar{x_1}$ and $\bar{x_2}$ such that the interval $\bar{x_1} < \bar{X} < \bar{x_2}$ defines the acceptance region and the two tails of the distribution $\bar{X} < \bar{x_1}$ and $\bar{X} > \bar{x_2}$ constitute the critical (or rejection region). Hence, in this case there are two cut-off values (critical values), defining the regions of rejection and acceptance. (Observe that the acceptance and rejection regions are mutually exclusive and exhaustive. This ensures that we make one of two decisions: reject H_0 or do not reject H_0.)

The critical region can be given in terms of z values by means of the transformation $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$ (Note that μ_0 is the value of μ under H_0).
Therefore,

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

or

$$P\left(\mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Hence, $\bar{x_1} = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $\bar{x_2} = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ From the population, we select a random sample of size n and compute the sample mean. If $\bar{X}$ falls in the acceptance region $\bar{x_1} < \bar{X} < \bar{x_2}$, we

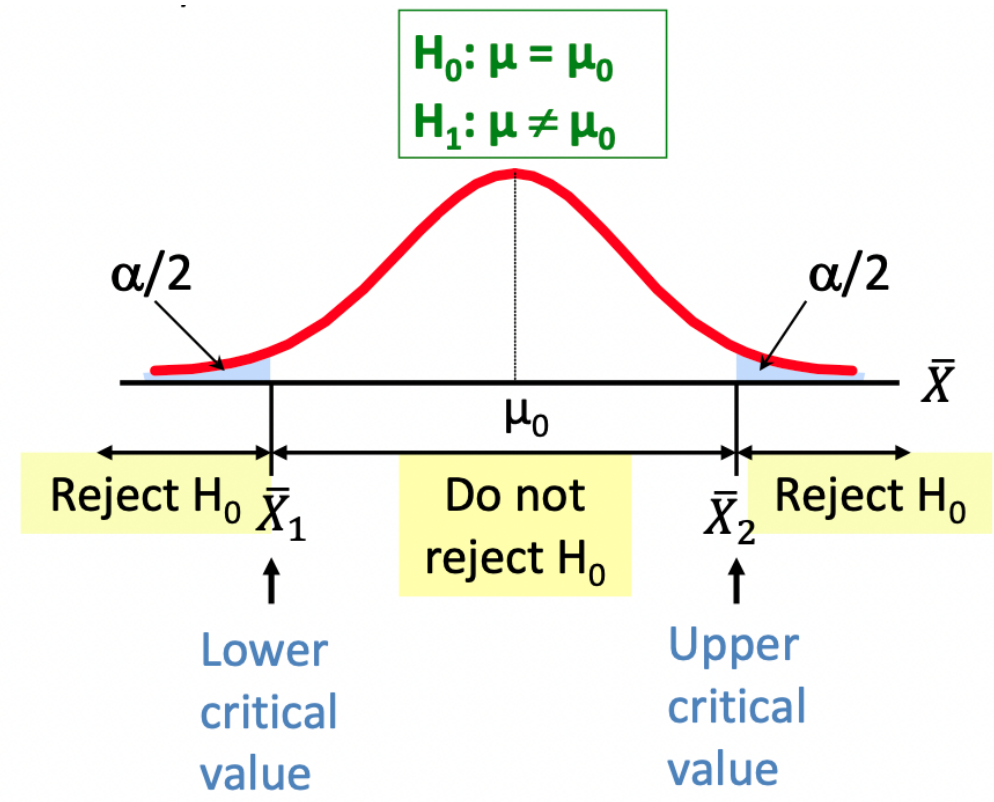


Figure 7.2: Two-sided test

conclude that $\mu = \mu_0$; otherwise we reject H_0 and accept that $H_1 : \mu \neq \mu_0$.
Since $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$, therefore $\bar{x_1} < \bar{X} < \bar{x_2}$ is equivalent to $-z_{\alpha/2} < Z < z_{\alpha/2}$. The critical region is usually stated in terms of Z rather than $\bar{X}$.

Relationship between two-sided test and Confidence Interval

The two-sided test procedure just described is equivalent to finding a $(1-\alpha)100\%$ confidence interval for μ. Then, H_0 is accepted if the confidence interval covers μ_0. If the C.I. does not cover μ_0, we reject H_0 in favour of the alternative $H_1 : \mu \neq \mu_0$ since

$$P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \iff P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Note 7.2.1.

Confidence interval is equivalent to the hypothesis testing if, and only if, the latter is two-sided; this is applicable not only for the mean related inference but also for the variance related inference.

p-value Approach to Testing

Definition 7.2.2 (p-value)**.** *p-value is defined as the probability of obtaining a test statistic more extreme (in favour of the alternative hypothesis) ($\leq$ or $\geq$) than the observed sample given the H_0 is true. It is also called the observed level of significance.*

1. Convert a sample statistic to a test statistic
2. Obtain the p-value
3. Compare the p-value with α. If p-value ¡ α, reject H_0. If p-value $\geq \alpha$, do not reject H_0.

Note 7.2.3.

Using p-value or the rejection region approach to perform the test must result in exactly the same conclusion/decision in terms of reject or do not reject H_0.
That is, p-value $< \alpha \iff$ the test statistic is in the rejection region.

Note 7.2.4.

To get the p-value for a two-sided test,

1. Take the minimum of the probability of obtaining a test statistic larger than observed value, and lower than observed value (because you don't know which is actually true: $\theta > \theta_0$ or $\theta < \theta_0$. So the more extreme one will have lower probability. Also, exactly one of $\theta > \theta_0$ or $\theta < \theta_0$ is true if the alternative hypothesis is true).
2. Multiply this by 2 to get the p-value.

One-sided Test

(a) Test $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$.
Let $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$. Then, H_0 is rejected if the observed values of Z, say z, is greater than z_α. Note that there is only one critical value since the rejection area is only one tail (in this case, the upper tail of the distribution).
(b) Test $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$.
Let $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$. Then, H_0 is rejected if the observed values of Z, say z, is greater than $-z_\alpha$. Note that there is only one critical value since the rejection area is only one tail (in this case, the lower tail of the distribution).

7.2.2 Unknown Variance

Consider the problem of testing the hypothesis concerning the mean, μ, of a population with:

1. Variance unknown
2. Underlying distribution is normal

(1) Two-sided Test
Test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Let $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$ where S^2 is the sample variance. Then, H_0 is rejected if the observed value of T, say t, $> t_{n-1;\alpha/2}$ or $< -t_{n-1;\alpha/2}$.
(2) One-sided Test
Test $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$. Then H_0 is rejected if $t > t_{n-1;\alpha}$ Test $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$. Then H_0 is rejected if $t < -t_{n-1;\alpha}$

7.3 Hypotheses Testing Concerning Difference Between 2 Means

7.3.1 Known Variances

1. Variances σ_1^2 and σ_2^2 are known
2. Underlying distributions are normal or both n_1 and n_2 are sufficiently large (greater than 30)

We know that the difference of two normal distributions follows a normal distribution. Then, we can proceed as before using the concepts of p-value or acceptance/rejection regions.

7.3.2 Large Sample Testing with Unknown Variances

1. Variances σ_1^2 and σ_2^2 are unknown
2. Both n_1 and n_2 are sufficiently large (greater than 30)

7.3.3 Unknown but Equal Variances

1. Variances σ_1^2 and σ_2^2 are unknown but equal
2. Both n_1 and n_2 are small (less than 30)

7.3.4 Paired Data

7.4 Hypothesis Testing Concerning Variance

7.4.1 One Variance Case

Assume that the underlying distribution is normal. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n drawn from a (approximate) $N(\mu, \sigma^2)$ distribution, where σ^2 is unknown.
We wish to test the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$. We know that $\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$.
Hence $H_0 : \sigma^2 = \sigma_0^2$ is rejected if the observed χ^2-value lies in the critical region (here, our test statistic is $\dfrac{(n-1)S^2}{\sigma_0^2}$). The critical region depends on the alternative hypothesis, and is summarised as follows: where $P(W > \chi^2_{n-1;\alpha}) = \alpha$ with $W \sim \chi^2(n-1)$

H_1	**Critical Region**
$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_{n-1;\alpha}$
$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{n-1;1-\alpha}$
$\sigma^2 \neq \sigma_0^2$	$\chi^2 < \chi^2_{n-1;1-\alpha/2}$ or $\chi^2 > \chi^2_{n-1;\alpha/2}$

Table 7.1: Critical Regions for Different Alternative Hypotheses

7.4.2 Ratio of Variances

Let us assume that the underlying distribution is normal and the means are unknown. When we are comparing the precision of one measuring device with that of another, or the variability in grading practices of one teacher with that of another, or the consistence of one production process with that of another, we are testing about the comparison between two population variances (or standard deviations).
We know that when two independent samples of sizes n_1 and n_2 are randomly selected from two normal populations, then $F = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$. Under $H_0 : \sigma_1^2 = \sigma_2^2$, $F = \dfrac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$.
Our test statistic is $F = \dfrac{S_1^2}{S_2^2}$. Hence $H_0 : \sigma_1^2 = \sigma_2^2$ is rejected if the observed F-value lies in the critical region . The critical region depends on the alternative hypothesis, and is summarised as follows: where $P(W > F_{v_1,v_2;\alpha}) = \alpha$ with $W \sim F(v_1, v_2)$

Note 7.4.1.

H_1	**Critical Region**
$\sigma_1^2 > \sigma_2^2$	$F > F_{n_1-1,n_2-1;\alpha}$
$\sigma_1^2 < \sigma_2^2$	$F < F_{n_1-1,n_2-1;1-\alpha}$
$\sigma_1^2 \neq \sigma_2^2$	$F < F_{n_1-1,n_2-1;1-\alpha/2}$ or $F > F_{n_1-1,n_2-1;\alpha/2}$

Table 7.2: Critical Regions for Different Alternative Hypotheses

Test statistic plays a key role in performing hypothesis testing. Test statistic must be a function of the sample, e.g., $X_1, X_2, \ldots, X_n$ and does not rely on any unknown parameter. Similar to the construction of the confidence interval, we can summarize the procedure for constructing the test statistic for mean-related hypothesis tests as follows. Denote by θ the parameter of interest. We consider the hypotheses:

$$H_0 : \theta = \theta_0 \qquad \textit{versus} \qquad H_1 : \ldots$$

where θ_0 is a given value which is assumed to be the true value of θ in developing the distribution of the test statistic. H_1 could be $\theta < \theta_0, \theta > \theta_0$, or $\theta \neq \theta_0$. The following are the steps:

1. Look for an estimator $\hat{\theta}$ for θ, e.g., $\bar{X}$ for μ.
2. Derive the formula for $V(\hat{\theta})$.
3. The test statistic is constructed to be $T = \dfrac{\hat{\theta} - \theta}{\sqrt{V}}$.
   (a) If $V(\hat{\theta})$ does not depend on any unknown parameter, e.g. when σ^2 is known, $V(\bar{X}) = \sigma^2/n$, we set $V = V(\hat{\theta})$. Then T follows approximately $N(0,1)$ when the data are normal or the sample size is sufficiently large.
   (b) If $V(\hat{\theta})$ contains some other unknown parameters, e.g. σ^2, we replace the parameter, e.g. S^2 can be used to replace σ^2 and result in $\hat{V}(\hat{\theta})$. We set $V = \hat{V}(\hat{\theta})$. Then, the distribution of T has two possibilities:
      i. The sample size is sufficiently large, then $T \sim N(0,1)$ approximately.
      ii. If the sample size is small but the observations are normally distributed, then $T \sim t(df)$ where df is the degrees of freedom of the parameter estimated in $V(\hat{\theta})$.

Note that the above strategy is not applicable to construct test statistic for the variance related tests.

7.5 Important Assumptions

1. To use the fact that $\dfrac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$, we need all the individual $X_i's$ to be normally distributed. This is an exact distribution and not an approximation - we don't require n to be large to use this but we do require the normality assumption. In other words, our test statistic follows a t-distribution only when our underlying population distribution is normally distributed and we are using sample variance as an estimate for population variance .
2. On the other hand, for $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, we must have either the individual X_i's being normally distributed (in which case this is exactly true) or n being large (in which case

we can approximate this using classical central limit theorem). Typically, if $n > 30$, we can approximate $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ even if the normality assumption does not hold. If σ is unknown, we can estimate it using s, the sample variance. But this will not change the distribution to t-distribution. We are just trying to estimate the population variance by using the sample variance as our best guess.

7.5.1 Assumptions for 2 Samples Independent t-test

1. Random samples - the samples are obtained from randomisation.
2. The 2 samples are independent and within each group, the observations are independent and identically distributed.
3. Response variable is quantitative
4. Equal variances for the 2 groups ($\sigma_1^2 = \sigma_2^1 = \sigma^2$)
5. Individual normality - normal data for each group.

Only if all the above assumptions are satisfied,, $T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$ where $s_p^2 = \frac{(n_1-1)S_1^2 - (n_2-1)S_2^2}{n_1+n_2-2}$

7.6 Summary

3 conditions

A. Normal Distributions
B. Parameters Known
C. Large Sample Size

One Sample

Estimation of the population mean, μ

Conditions (A, B, C)	Pivotal Quantity	Distribution	$100(1-\alpha)\%$ Confidence Interval
(Y, Y(σ^2), --)	$T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	$N(0,1)$	$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
(N, Y(σ^2), Y)	$T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	Approximate $N(0,1)$ (CLT)	$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
(Y, N, --)	$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$	$t(n-1)$	$\bar{X} \pm t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}$
(N, N, Y)	$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$	Approximate $N(0,1)$ (CLT & LLN)	$\bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$

* Y = Yes, N = No, - = It does not matter

CLT: Central Limit Theorem, LLN = Law of Large Numbers

$\Pr(Z > z_{\alpha/2}) = \alpha/2$ with $Z \sim N(0,1)$, $\Pr(T > t_{v;\,\alpha/2}) = \alpha/2$ with $T \sim t(v)$

Figure 7.3: Confidence Interval Summary

Hypothesis testing on the population mean. $H_0: \mu = \mu_0$

Conditions (A, B, C)	Test Statistic	Alternative	Reject H_0 if
(Y, Y(σ^2), --)	$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$H_1: \mu \neq \mu_0$	$T < -z_{\alpha/2}$ or $T > z_{\alpha/2}$
		$H_1: \mu > \mu_0$	$T > z_\alpha$
		$H_1: \mu < \mu_0$	$T < -z_\alpha$
(N, Y(σ^2), Y)	$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$H_1: \mu \neq \mu_0$	$T < -z_{\alpha/2}$ or $T > z_{\alpha/2}$
		$H_1: \mu > \mu_0$	$T > z_\alpha$
		$H_1: \mu < \mu_0$	$T < -z_\alpha$
(Y, N, --)	$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$	$H_1: \mu \neq \mu_0$	$T < -t_{n-1;\,\alpha/2}$ or $T > t_{n-1;\,\alpha/2}$
		$H_1: \mu > \mu_0$	$T > t_{n-1;\,\alpha}$
		$H_1: \mu < \mu_0$	$T < -t_{n-1;\,\alpha}$
(N, N, Y)	$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$	$H_1: \mu \neq \mu_0$	$T < -z_{\alpha/2}$ or $T > z_{\alpha/2}$
		$H_1: \mu > \mu_0$	$T > z_\alpha$
		$H_1: \mu < \mu_0$	$T < -z_\alpha$

* Y = Yes, N = No, - = It does not matter

Figure 7.4: Hypothesis Testing Summary

Chapter 8

Regression

Consider a quantitative response variable Y and an explanatory variable X. Our goal is to estimate Y using X.
For example, we could consider estimating the price of a house (Y) with the size of the house (X).
We also call the **response** variable, the **dependent** variable or **target** variable or **output** variable.
We also call the **explanatory** variable, the **independent** variable or the **regressor** or the **input** variable or the **covariate** or the **predictor** variable.

Definition 8.0.1 (Regression)**.** *A regression of the response variable Y on the regressor X is a mathematical relationship between the mean of Y and different values of X.*

Linear regression means that the regression is linear, of the form: $Y = \beta_0 + \beta_1 X + \epsilon$.
ϵ is a random variable. It has variance σ^2. β_0 is the Y-intercept, and β_1 is the slope of the line, known as coefficients or parameters of the model.
The word "linear" refers to **linearity in the parameters**. The following are still linear regression models:

- $Y = \beta_0 + \beta_1 sin(X) + \epsilon$
- $Y = \beta_0 + \beta_1 log(X) + \epsilon$
- $Y = \beta_0 + \beta_1 e^X + \epsilon$

However, the following are not linear regression models:

- $Y = \beta_0 sin(\beta_1 X) + \epsilon$
- $Y = \beta_0 e^{\beta_1 X} + \epsilon$

The word "simple" refers to only one regressor in the model. When a model has more than one regressor, we call it "multiple linear regression".

8.1 Simple Linear Regression

8.1.1 Assumptions

The assumptions for a simple linear regression model $Y \sim X$ are:

1. Data were obtained by randomization.
2. The relationship between X and Y is linear.

3. The error term $\epsilon \sim N(0, \sigma^2)$ where σ is a constant.

We do not check these assumptions before building a model. Instead, we check them after fitting the model.
The implications of the model assumptions are:

- For any particular X value, the response is a variable that has a normal distribution: $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$.
- For any particular X value, the mean of variable Y is $\beta_0 + \beta_1 X$.
- For any value of X, the variance of Y is always the same: σ^2

Here, β_0, β_1 and σ^2 are the parameters to be estimated.

8.1.2 Estimation

As an example, we use Ordinary Least Squares (OLS) Estimation (although in practice, the most commonly used estimation is the Maximum Likelihood Estimation). In OLS, we consider all possible candidate lines. For each line, we compute the sum of the squared residual $e_i's$ where each residual is the difference of the predicted value (by the model) and the actual y value. That is, we pick the line which minimises $\sum_i e_i^2$ where $e_i = \hat{y} - y$ and we call it the line of best-fit.
There are 2 types of estimation: point estimate and interval estimate (just as we saw in the case of estimating the population mean).

Point Estimate

The point estimates of β_0 and β_1 can be calculated from the OLS estimates of the model. The reason why these are *estimates* and not the actual values of β_0 and β_1 is because the sample itself is chosen randomly from the population. So, if we obtain another random sample and find the best-fit-line (using OLS) on that sample, we'll get different values of the slope and intercept. It is important to understand why the values we obtain are estimates and not the exact measures of the slope and intercept. This is also why we can talk about interval estimates of β_0 and β_1: A 95% confidence interval for β_0 means that if we continually select random samples and obtain 95% confidence interval estimates of β_0, about 95% of them will contain the true β_0 (which is an unknown constant).
Another fact is that, given a value of X, $\hat{Y}$ is a point estimate of the mean selling price. This is by definition of our regression line. In other words, the predicted output value by a model is an estimator for the **mean** output value at that given input value.

Definition 8.1.1 (Interpolation)**.** *Interpolation refers to the estimation of the mean response for an X value that had not been observed, but is* ***within the range of observed values****.*

Definition 8.1.2 (Extrapolation)**.** *Extrapolation refers to the estimation of the mean response for an X value that is* ***outside the range of observed values****.*

For extrapolation, we do not know the form of the relationship outside of our sampled values, hence it's better to avoid.

Estimating σ^2
σ^2 gives us an idea of the variability of the response values around the fitted line. Remember our assumption for simple linear regression is that the variability is the same for all values of X. (Note that we are referring to the variability of Y and since Y is assumed to be some linear function of X, the variability of Y can potentially depend on the value of X.)
The point estimate of σ^2 is computed using the residuals. Till now, we had only seen the case

where we estimate the population variance using the sample variance. Using the residuals is another way we can estimate the variance. The residuals are computed as: $e_i = Y_i - \hat{Y}_i$ and we can obtain the residual standard error directly from R.

Interval Estimate

For different sample data (taken from the same population), we will get different point estimates of β_0 and β_1. Hence, it is sometimes preferable to report a confidence interval for the parameters β_0 and β_1 instead of the point estimates.
We can obtain a 95% CI for each coefficient (β_0 an β_1) from the R.

8.1.3 Testing Hypotheses

After building a model, we may ask:

- Is the model statistically significant?
- Is the explanatory variable included in the model statistically significant? (That is, does the response variable actually depend on the explanatory variable?)

To answer these questions, we can perform few tests.
There are 2 kinds of tests that can be conducted using the R-output:

- t-test: for testing the significance of one regressor (or one coefficient)
- F test: for testing the significance of the whole model.

In simple linear regression, there is only one explanatory variable, and so the F-test is equivalent to the t-test (as if the model is statistically significant, it is due to the only explanatory variable and vice versa).
In multiple linear regression (at least 2 coefficients excluding β_0), there are more than one t-test possible, and we have to select the right one to make the desired inference.

t-tests for β_1

The 5 steps involved in performing t-test are as follows:

1. Assumptions: same as assumptions of the model.
2. The null and alternative hypotheses are

   $$H_0 : \beta_1 = 0 \quad \text{or} \quad H_0 : \text{regressor X is not significant}$$

   $$H_1 : \beta_1 \neq 0 \quad \text{or} \quad H_1 : \text{regressor X is significant}$$

   Observe that if the coefficient of a regressor is 0, it is equivalent to removing that regressor from the model entirely.
   Note that one-sided tests are also possible.
3. Test statistic is a t-statistic (i.e., the test statistic follows a t-distribution). In our case, $T = \dfrac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$, which can be found from the R-output. Here, $SE(\hat{\beta}_1)$ is the standard error of the slope coefficient.
   For a simple model, the null distribution of T is t_{n-2} where n is the number of observations.
4. Derive the p-value (from R-output)
5. Conclude whether the slope β_1 is significantly different from 0 at a pre-specified α-level

8.1.4 F-tests

To test if the whole model is significant or not, we use F-test. The null hypothesis is that the model is not significant, while the alternative hypothesis is that the model is significant. Equivalently,

H_0 : all the coefficients excluding the intercept are zero
H_1 : at least one of the coefficients, excluding the intercept, is non-zero

In a simple linear regression model, the hypothesis of the F-test actually are $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, which are the same as for a t-test for β_1.

Note 8.1.3. *In the simple model, the t-test to test the significance of the slope and the F-test to test the significance of the model have the same p-value.*

What does it mean if we do not reject H_0 of the F-test?
It means the regressor(s) used in the model is/are not significant. It suggests a new model without any regressors, as follows: $Y = \beta_0 + \epsilon$ (here Y refers to the actual values). Then, the new fitted model reduces to: $\hat{Y} = \hat{\beta}_0$, which is called intercept model (since the only term is the Y-intercept). This suggests that Y does not depend on any regressor.

Note 8.1.4. *In the intercept model, the value of $\hat{\beta}_0$ must be the sample mean of Y. This should be intuitive because if we don't have any information regarding Y, our best estimate of Y, for any X, is given by the mean of Y.*

8.2 Regression Diagnostics

Recall our assumptions for the linear regression model:

1. Randomisation (the sample data must be obtained through randomisation)
2. Linearity (Y must be linearly related to X)
3. Normality (the error term must follow a standard normal distribution)
4. Constant variance (homoscedasticity) - the variance of Y is the same for all values of X.

After we have built the model, we need to verify if our assumptions were valid. We do this by:

1. Randomisation - from the steps of data collection
2. Linearity - using scatter plot between Y and X.
3. Normality - checked using the residuals of the built model.
4. Constant variance (homoscedasticity) - checked using the residuals of the built model.

If we observe one or more observations are violated, we can fix our model in the following way:

1. If the linearity assumption is violated, one possible fix is to add higher order terms (e.g. X^2) to the model.
2. Variance not constant - transform the response by taking $ln(Y)$, square root ($\sqrt{Y}$) or the reciprocal ($\frac{1}{Y}$, to be the response of the model. Note that such a transformation will change the interpretation of the coefficient β_1.

Recall that the raw residuals are defined to be $e_i = Y_i - \hat{Y}_i$ for each observation. The raw residuals are our best estimates of what the $\epsilon_i's$ are.
Residual plots are used to:

- Check the normality assumption.
- Check for non-constant variance and the need to transform Y
- Check the need to add higher order terms in X.

We use **Standardized Residuals (SR)** (r_i) (although other kinds of residuals such as studentized residuals and deleted residuals yield the same information)

$$\text{standardized residual} = \frac{Y - \hat{Y}}{\text{standard error of } (Y - \hat{Y})}$$

We can plot:

- SR on the y-axis against the $\hat{Y}_i's$ on the x-axis.
- SR on the y-axis against X on the x-axis.
- Histogram of the SR
- QQ-plot of the SR.

We expect the plot of SR against $\hat{Y}$ and SR versus X to have points scattered randomly about 0, within the interval (-3, 3). For the histogram and QQ plot of SR, we expect a normal distribution.

Note 8.2.1. *The SR from a fitted model are not exactly independent, but when sample size n is large, we can expect the SR show randomness.*

8.2.1 Outliers and Influential Points

An **outlier** of a model is identified by the residuals. It is a point that is far from the rest of the data points in absolute value. We should investigate these points to see if they should be dropped.
We can go by the rule (heuristic) that an outlier has standardized residuals greater than 3 or less than -3.
An **influential point** is one that affects the parameter estimates greatly. An outlier may or may not be influential. The influence of a point can be measured using Cook's distance. Cook's distance measures the effect of deleting a given observation. Points with a large Cook's distance could be influential points (we can use 1 as the threshold)

8.3 Coefficient of Determination: R^2

The coefficient of determination of a linear model, R^2, is a value of a statistic which helps to check the goodness of fit of the model.
It is interpreted as the proportion of total variation of the response (about the sample mean $\bar{Y}$ that is explained by the model .
It takes on a value between 0 and 1.
If there are repeated X values with different Y values in the dataset, R^2 can never be 1.
In case of a simple model, $r = \sqrt{R^2} = |Cor(X, Y)|$.
If $\hat{\beta}_1 < 0$, then $Cor(X, Y) = -R$.

One weakness of R^2 is that is doesn't account for the number of regressors in the model. Increasing the number of regressors will always result in R^2 increasing so how do we compare models with different numbers of regressors?
We use adjusted R^2 to compare such models, which accounts for the differing number of regressors.

8.4 Multiple Linear Regression

The t-test of each regressor in a multiple linear regression model has $n-p-1$ degrees of freedom, where p is the number of predictors (or non-intercept terms in the regression model) and n is the sample size or number of observations. So, each regressor causes the loss of a degree of freedom.

Similarly, the F-test statistic of the overall model follows an F-distribution with degrees of freedom p and $n-p-1$.

8.4.1 Adjusted R-squared

R^2 has the deficiency that the more variables we get, the larger it is. Hence, for the same response, it is not reasonable to compare the fit of models using R^2.
The measure of fit is the adjusted R^2, which takes into account the number of regressors included:

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

where p is the number of predictors/regressors in the model.
So, for the same response, we use the adjusted R^2 to compare the fit of 2 models.

8.4.2 Indicator Variables

An indicator variable for a categorical variable with 2 categories takes on the value 1 if the category is observed, and 0 otherwise.
For instance, with variable gender X_2 having two categories (male and female), an indicator variable for category male would be:

$$I(\text{gender = male}) = \begin{cases} 1, \text{if } X_2 = \text{male} \\ 0, \text{if } X_2 = \text{female} \end{cases}$$

Then, a model could look something like: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 I(X_2 = male)$.

Recall that $\hat{Y}$ is the predicted value of Y and Y is the actual ground-truth response value (that we are trying to predict).

It is possible that there is an interaction between two variables and it affects the response. A possible model is: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 I(X = male) + \beta_3 X_1 \times I(X_2 = male)$.

Note 8.4.1. *When you decide to keep an interaction term (a term involving more than one variable) in the model equation, you should keep all the individual variable terms that make up the interaction term too, even if they are not statistically significant in the presence of the interaction term.*

Chapter 9

Describing Numerical Data

Definition 9.0.1 (Parameter)**.** *A parameter is a numerical summary of the population. It is unknown.*

Definition 9.0.2 (Statistic)**.** *A statistic is a summary of a sample taken from the population. We compute it based on the data in our sample.*

There are 2 kinds of statistics:

- Descriptive statistics
- Inferential statistics

We use the statistics, computed using the sample, to make inferences about a population parameter. For example, we use the sample mean to estimate the population mean.

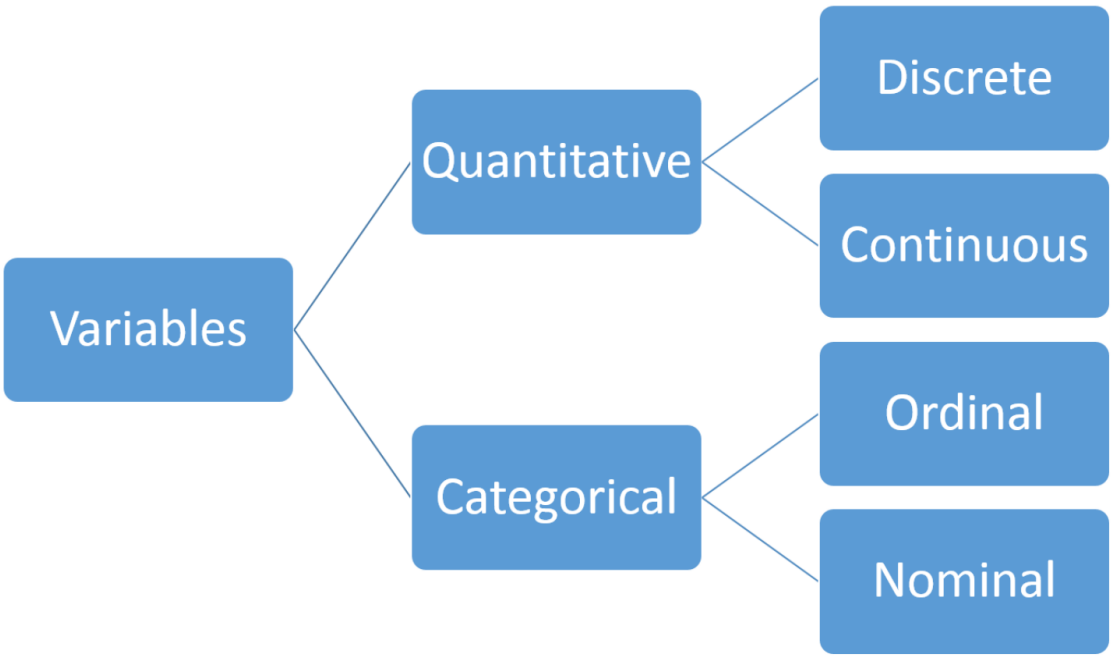


Figure 9.1: Types of data

In this chapter, we cover the descriptive statistics: numerical and graphical summaries for single quantitative variable and then summaries or the association between two variables (one categorical and one quantitative or both quantitative) in the data.

Inferential statistics will be covered later.
There are two major ways of describing numerical data:

1. Numerical summaries/descriptive measures: number of observations (sample size), location, variability, and other measures
2. Graphical summaries: histogram, boxplot, QQ plot (for checking normality), scatter plot for bivariate data.

All relevant commands can be found in the chapters on R and Python. Here, we only provide their explanations and relevance.

9.1 Single Quantitative Variable

Definition 9.1.1 (Skewness)**.** *If a distribution is unimodal (has just one peak), beside the location and variability, we would check whether it is symmetric or skewed to one side.*
If the bulk of the data is at the left, and the right tail is longer, we say that the distribution is ***skewed right or positively skewed****.*
If the peak is towards the right and the left tail is longer, we say that the distribution is ***skewed left or negatively skewed****.*
Given a sample size of n, the sample skewness is given by

$$\frac{\sqrt{n(n-1)}}{n-2} \times \frac{m_3}{(m_2)^{3/2}}$$

where

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \textit{and} \quad m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

The skewness value represents the amount and direction of skew.

- If skewness = 0, the data is perfectly symmetrical, but it's very rare
- If $-0.5 <$ skewness < 0.5, the distribution is approximately symmetric
- If skewness is between -1 and - 0.5, or between 0.5 and 1, the distribution is moderately skewed.
- If $|\text{skewness}| > 1$, the distribution is highly skewed.

Beside skewness, **kurtosis** is another numerical measure of shape of a distribution. Higher value of kurtosis indicate a higher, sharper peak; lower values indicate a lower, less distinct peak.

Definition 9.1.2 (kurtosis)**.** *Given a sample of size n, the sample kurtosis, or actually excess kurtosis is*

$$\frac{n-1}{(n-2)(n-3)}\left[\frac{(n+1)m_4}{m_2^2} - 3(n-1)\right]$$

where

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \textit{and} \quad m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

Kurtosis tells you how sharp the central peak is, relative to a standard bell curve.

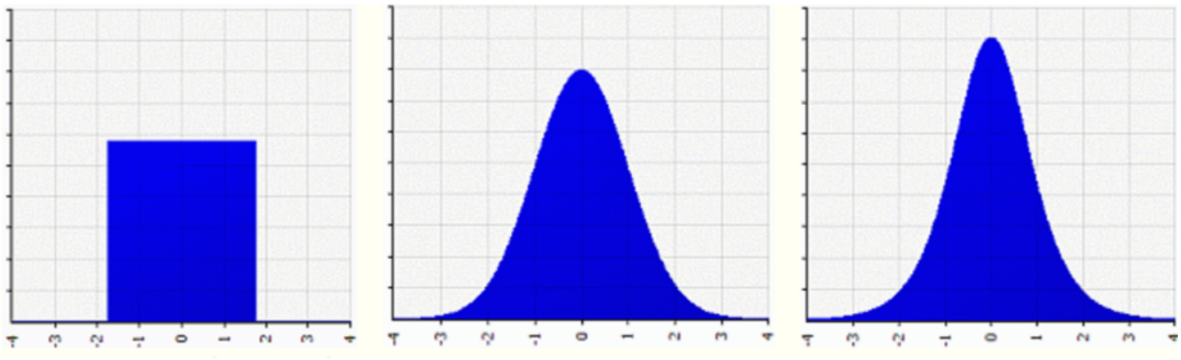


Figure 9.2: 3 distributions with mean = 0, sd = 1, skewness = 0 but different kurtosis

Note 9.1.3.

- *For a dataset, when the mean is the same (or approximately) the same as median, then the data is close to symmetric.*
- *Mean is sensitive to outliers while median is not*
- *When mean is much larger than median, data is right (positively) skewed; when mean is much smaller than median, then the data is left (negative) skewed.*

But, numerical summaries are not enough: no matter how many summary measures we report, nothing beats a picture.

We use boxplots, QQ plots, histograms and density plots for single quantitative variable analysis.

A **histogram** is a graph that uses bars to portray the frequencies of relative frequencies of the possible outcomes for a quantitative variable. In a histogram, we look for whether the data is unimodal or bimodal or multimodal. Is it symmetric of skewed? Are there any suspected outliers?

Density plots can be thought of as plots of smoothed histograms (the probability density plot of histogram is different from the standard normal density plot). While the normal density plot follows a bell standard normal distribution with mean and variance equal to that of the distribution, the density plot actually follows the distribution closely (i.e., the shape of the histogram).

Boxplots provide a skeletal representation of a distribution, and they are very well suited for showing distributions for multiple variables.

The box is made of Q_1 and Q_3. The max-whisker reach is determined by: $Q_3 + 1.5IQR$ and the min-whisker reach is given by $Q_1 - 1.5IQR$. Any data point that is out of the range from the min to max whisker reach is defined to be an outlier. Except the outliers, the maximum point determines the upper whisker and the minimum points determine the lower whisker of a boxplot.

A boxplot might have mild outlier and extreme outlier. An outlier is defined to be extreme if it's larger than $Q_3 + 3IQR$ or smaller than $Q_1 - 3IQR$.

A given boxplot helps us identify median, lower and upper quantiles, and outliers.

The purpose of plotting a **QQ plot** of a dataset is to see if the data follows (approximately) a normal distribution or not.

A QQ-plot plots the standardized sample quantiles against the theoretical quantiles of a N(0, 1) distribution. If they fall on a straight line, then we can say that there is evidence that the data came from a normal distribution.

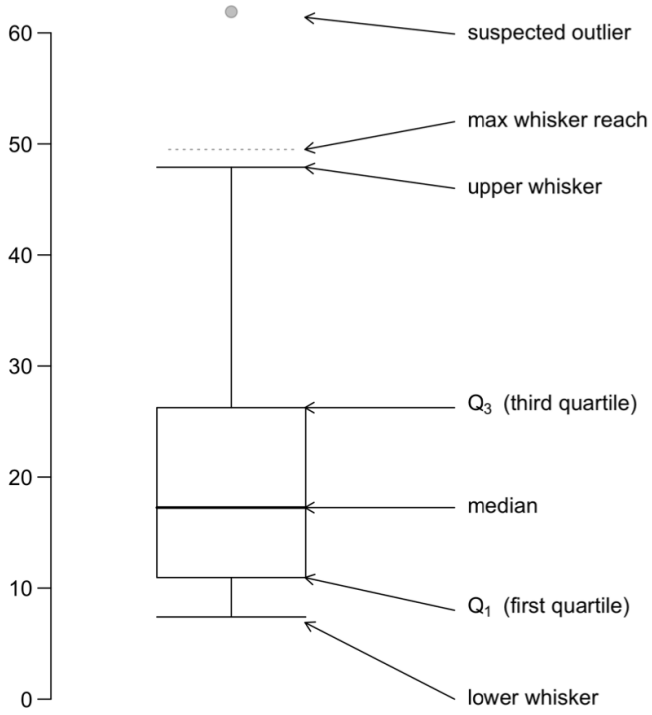


Figure 9.3: Caption

From the points on the plot, we can usually tell whether our sample data has longer or shorter tail than normal, and on which side of the mean it occurs.

Note 9.1.4. *The comments below are for QQ plot when the theoretical quantiles are on the X-axis and the sample quantiles are on the Y-axis:*

- *Right tail is below the straight line → shorter than normal*
- *Right tail is above the straight line → longer than normal*
- *Left tail is below the straight line → longer than normal*
- *Left tail is above the straight line → shorter than normal*

The comments should change accordingly if the sample quantiles were on the X-axis and theoretical quantiles are on the Y-axis (for example, if the right tail is below the straight line, then it is longer than normal)

9.2 Robust Estimators for Location and Scale Parameters

What is the motivation for using robust statistics, i.e., why do we need to use robust estimators?

To understand about a population, we cannot assess it all, but just through a representative of a population, which is a sample/dataset collected randomly from the population. We make some assumptions about the underling distribution (population distribution) based on our domain knowledge or past samples. However, outlier(s) may appear in the sample, hence the sample distribution may depart from the underlying distribution assumptions. Therefore, the conclusions derived from this sample using some statistical methods might not be reliable if these statistical methods are not robust.

Definition 9.2.1 (Robust)**.** *A statistical method is robust if the statistic is insensitive to slight departures from the assumptions that justify the use of the statistic. In other words, a statistical method is said to be robust with. respect to a particular assumption if it performs adequately even when that assumption is modestly violated.*

The robustness of a robust statistic can be measured by measures such a breakdown point, influence curve, and gross error sensitivity.

Consider the formula to calculate a 95% confidence interval for the population mean:

$$\bar{X} \pm t_{n-1,0.975} \times \frac{s}{\sqrt{n}}$$

where $t_{n-1,0.975}$ corresponds to the 0.975-th quantile of a t-distribution with $n-1$ degrees of freedom (some may use the z-score instead but that's not relevant here). Recall that the assumptions under which the above formula holds are:

- The sample must be obtained from randomization, either by a random sample or a randomized experiment
- The distribution of the data should be approximately normal.

The assumption that data are obtained through randomization is crucial. The procedure for generating the confidence interval is not robust to this assumption.

The assumption that the data from a normal population is not crucial. That is, the procedure is **robust** to this assumption. We only need to ensure that there are no extreme outliers in the dataset and sample size n is relatively large enough, then we can proceed.

9.2.1 Robust Estimation of Location

The sample mean $\bar{x}$ is an estimator of the location parameter μ and it is the **Maximum Likelihood Estimator (MLE)** when the underlying distribution is normal. This can be proven by considering the loglikelihood function and setting its partial derivative w.r.t. μ to be zero (to find the maximum).

The definition of MLE for a location parameter is the value that maximises the joint conditional pdf $f(x_1, x_2, \ldots, x_n|\mu)$ where f is the probability density function. That is, it is the value of the location parameter that makes it most likely to obtain the sample data.

For example, if $X_1, \ldots, X_n$ are IID $N(\mu, \sigma^2)$ (and we don't know μ), then the joint pdf is:

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}$$

Notice that now the values of X's are fixed (we have observed the sample) and we're trying to predict μ, σ^2. In particular, right now we're interested in minimising μ so we can treat σ as a constant. In any case, the likelihood function depends on μ and σ.

To minimise f, we can also minimise $\log(f)$ since $\log$ is monotonically increasing. Then, the loglikelihood function is obtained by taking $\log$ on both sides,

$$l(\mu, \sigma^2) = -n\log(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

To find the value of μ (as a function of x_i's) that maximises loglikelihood, we can take the derivative w.r.t. μ and set it equal to 0. Then, we get:

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

Solving the equation above, we get $\mu = \frac{\sum_{i=1}^{n} x_i}{n}$, which is the sample mean. Hence, the sample mean is the most likely value of the population mean given the sample observations, and hence we call it the maximum likelihood estimator.

Notice that we use the fact that the observations are IID in the proof. Hence, the assumption of data obtained through randomization is crucial to the reliability of this statistic.

However, the mean is not robust to outliers - i.e., it is affected significantly by the presence of outliers. Hence, we need robust estimators of a location parameter. Some commonly used ones are:

1. Trimmed mean
2. Winsorized mean
3. Huber's M-Estimates
4. Tukey's bisquare estimator
5. Humpe's M-estimator

Trimmed Mean

Trimmed mean is a simple robust estimator of location.

Definition 9.2.2 (Trimmed mean)**.** *The $100\alpha\%$ trimmed mean is calculated by discarding the lowest $100\alpha\%$ and the highest $100\alpha\%$ and taking the arithmetic mean of the remaining data.*

It is recommended that we choose α from 0.1 to 0.2.

The higher the value of α, the lower the effective sample size and so, the less accurate our estimate of location parameter would be (since we're discarding away a lot of data by treating it as outliers/noise). If we set $\alpha = 0.5$, we're throwing away all the observations. As $\alpha \to 0.5$, the trimmed mean approaches the median (since we're only considering the central observation in the end).

Winsorized Mean

Trimmed mean is the mean of trimmed data. Winsorized man is the mean of trimmed and replaced data. Winsorization replaces extreme data values with less extreme values.

Definition 9.2.3 (Winsorized mean)**.** *Let $[a]$ denote the nearest integer of number a. Let dataset of observations $x_1, \ldots, x_n$ be sorted to give $x_{(1)}, \ldots, x_{(n)}$. Then, the winsorized mean is computed after all the $[n\alpha]$ smallest observations are replaced by $x_{([n\alpha]+1)}$ and the $[n\alpha]$ largest observations are replced by $x_{(n-[n\alpha])}$.*

It is recommended that we choose α from 0.1 to 0.2.

Huber's M-estimates

We know that for a location parameter with underlying distribution approximately normal, location parameter μ is estimated well by sample mean, $\bar{x}$.
$\bar{x}$ is the MLE of μ, where $\bar{x}$ is found by minimising $\sum_{i=1}^{n}(x_i - \mu)^2$, which is called sum of squared errors.

Huber (1964) proposed that when estimating the location parameter μ, one can obtain more robustness by another function of error than the sum of their squares.

Hence, instead of minimizing the sum of squared error, Huber proposed we can find the estimator denoted by T - which is a function of $x_1, \ldots, x_n$ where T is the minimizer of

$$\sum_{i=1}^{n} \rho(x_i - T)$$

where ρ is a non-constant function that is meaningful. So, by varying the function ρ, we can get different estimators of location (sine we're trying to minimise different metrics/objective functions). Note that we need ρ to be non-constant, because otherwise we cannot minimise it by varying T.

For any specific function ρ, we simply need to differentiate it w.r.t. T (since we're assuming that the observations are fixed, and we need to find a T that minimises the error function) The class of M-estimator proposed by Huber contains the sample mean, sample median and all the maximum likelihood estimators.

1. If we set function $\rho(x) = x^2$, the minimiser of $\sum_{i=1}^{n} \rho(x_i - T)$ = $\sum_{i=1}^{n}(x_i - T)^2$ is $\bar{x}$
2. If we set function $\rho(x) = |x|$, the minimiser of $\sum_{i=1}^{n} \rho(x_i - T)$ = $\sum_{i=1}^{n} |x_i - T|$ is the sample median.
3. If we set function $\rho(x) = -\log f(x)$ where f is the assumed density function, the minimiser of $\sum_{i=1}^{n} \rho(x_i - T) = \sum_{i=1}^{n} -\log f(x_i - T)$ is the maximum likelihood estimator (MLE).
4. If we set function $\rho(x)$ as
$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{for } |x| \leq k \\ k|x| - \frac{1}{2}x^2 & \text{for } |x| > k \end{cases}$$
then the estimator corresponds to a winsorized mean.
5. If we set function $\rho(x)$ as
$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{for } |x| \leq k \\ \frac{1}{2}k^2 & \text{for } |x| > k \end{cases}$$
then the estimator corresponds to a trimmed mean.

Note 9.2.4 (Minimiser)**.** *When we say that $x = x_0$ is a minimiser of function $f(x)$ it means that $f(x)$ has the minimum value at $x = x_0$. That is, $\forall x, f(x) \geq f(x_0)$.*
Clearly, at $f(x = x_0)$, the gradient is zero (by definition) and so we can find x_0 by setting the derivative to be zero.

9.2.2 Robust Estimators of Scale

The sample standard deviation which usually is denoted as s is a commonly used estimator of the population scale parameter, σ. However, the usual sample standard deviation is not robust, it is sensitive to outliers and may not remain bounded even when a single data point is replaced by an arbitrary number.

With robust scale estimators, the estimates remain bounded even when a portion of the data points are replaced by arbitrary numbers.

Interquartile Range (IQR)

The interquartile range (IQR) can be used to estimate or measure the scale parameter, though it is not a robust estimator of σ (in the sense that if more than 25% of the upper or lower portion of the data is changed, it will be affected significantly.

IQR is defined by $IQR = Q_3 - Q_1$ where Q_1 and Q_3 are the first and third quartiles respectively.

Note 9.2.5. *For a normal distribution, the standard deviation σ can be estimated by dividing the interquartile range by 1.35.*
Proof:

1. *In general, we can write $X = \mu + \sigma Z$ where $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0, 1)$ is the standard normal distribution.*
2. *Then, $IQR(X) = \sigma \times IQR(Z)$ (obviously IQR won't change if a constant, μ is added, and $IQR(aX) = a \times IQR(X)$ since each observation is scaled by a constant factor of a).*
3. *The values of Q_1 and Q_3 for standard normal distribution are tabulated and can be found to be -0.675 and 0.675 respectively. Thus, $IQR(Z) = 1.35$*
4. *From (2), $\sigma = \dfrac{IQR(X)}{IQR(Z)} = \dfrac{IQR(X)}{1.35}$*

Median Absolute Deviation (MAD)

The most popular robust estimator of scale parameter is MAD.

Definition 9.2.6 (MAD). *$MAD = median_i(|x_i - median_j(x_j)|)$, where the inner median, $median_j(x_j)$ is the median of n observations, and the outer median, $median_i$ is the median of the n absolute values of the deviations about the median.*

Note 9.2.7. *For a normal distribution, $\sigma = 1.4826 \times MAD$ can be used to estimate the standard deviation σ.*
The proof is similar to that of the IQR case, i.e., we can show that $MAD(X) = \sigma_X \times MAD(Z)$. Therefore, $\sigma = \dfrac{MAD(X)}{MAD(Z)}$, and theoretically, we can get $MAD(Z) = \dfrac{1}{1.4826}$

Gini's Mean Difference

The Gini's mean difference equals the mean of all the mutual differences of any 2 observations of the sample. Since we have $n(n-1)/2$ terms while the terms with outliers may be much less than $n(n-1)/2$ so that a robust estimator of σ may be expected using the Gini's mean difference.

Definition 9.2.8 (Gini's Mean Difference)**.**

$$G = \frac{1}{\binom{n}{2}} \sum_{i<j} |x_i - x_j|$$

If the observations are from a normal distribution, then $\frac{G\sqrt{\pi}}{2}$ *is an unbiased estimator if the standard deviation* σ *(because it can be shown that* $G(X) = \sigma G(Z)$ *and* $G(Z) = 2/\sqrt{\pi}$ *asymptotically (as* $n \to \infty$*).*

Chapter 10

Categorical Data Analysis

10.1 Introduction

A variable is called **categorical** if each observation belongs to one of a set of categories. Examples of categorical variables are gender, religion, race, type of residence.

How do we distinguish between quantitative and categorical variables: simply ask if there is a meaningful distance between any 2 points in the data (or whether finding the average or median or any mathematical operations makes sense). If such a distance is meaningful, then you have quantitative data. If not, categorical data. For example, it makes sense to compute the difference in systolic blood pressure between subjects but it does not make sense to consider the mathematical operation (”smoker” - ”non-smoker”).

It is crucial to identify which type of data you have (quantitative or categorical), as it affects the exploration techniques that you can apply.

A categorical variable is **ordinal** if the observations can be ordered (e.g. low, medium, high), but do not have specific quantitative values, i.e, there is some intrinsic ordering in the categories.
A categorical variable is **nominal** if the observations can be classified into categories, but the categories have no specific ordering (e.g. blue eyes, black eyes, brown eyes).

10.2 Single Categorical Variable

For a single categorical variable, we can use **frequency table** (which also can produce the proportion or percentage of categories) as numerical summaries. The category with the highest frequency is the **modal category**.

A common graphical to display a categorical variable is **bar plot**.

10.3 Two Categorical Variables

There are different methods to explore the association of 2 categorical variables: **contingency table** (useful for comparing proportions and calculating odds ratio), **Chi-square test** χ^2 (to check association) and charts (for easy visualisation).

10.3.1 Contingency Tables

As an example, let’s consider the following contingency table: Based on the above table, we’re

	Chest Pain	**No Chest Pain**	**Total**
Male	8.8%	91.2%	100%
Female	6.7%	93.3%	100%

Table 10.1: Contingency Table

interested in answering the following questions:

- Is the probability of having chest pain in males the same as the probability of having chest pain in females?
- Are the 2 variables independent (or associated)?
- Can we infer the causality (gender causing different probability of having chest pain)?

It is important to identify which variable is the response variable and which one is the explanatory variable, so that the conditional proportion is calculated correctly for making inference.

Lets assume that we have response Y with 2 outcomes in the columns (success and failure) and explanatory X has 2 levels/categories in rows. In row 1 and row 2, let π_1, π_2 denote the probabilities of a success, respectively. Let p_1 and p_2 denote the sample proportion of successes in row 1 and row 2.

The sample difference $p_1 - p_2$ is used to estimate the real difference $\pi_1 - \pi_2$. If this difference is significant, we can infer the association between X and Y.

Relative Risk: The ratio p_1/p_2 that is used to estimate π_1/π_2 is called relative risk. If this RR is significantly different from 1, we can also infer the association between X and Y.

Another term that can help us explore the association in a 2-way table is **odds ratio**.

For a probability of success π, the odds of success is defined as $odds = \pi/(1-\pi)$.

In a 2-way contingency table, the odds ratio (OR) is defined as the ratio of 2 odds: $\theta = \dfrac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$, whereas relative risk is the ratio of 2 probabilities.

Given a dataset (like the chest pain example), the sample odds ratio is: $\dfrac{p_1/(1-p_1)}{p_2/(1-p_2)}$.

Odds ratio can be any non-negative number.

Properties of Odds Ratio

1. When X and Y are independent, in the population, we have $\pi_1 = \pi_2$, so $\theta = 1$. This value is a baseline for comparison.
2. The further values of θ from 1 in a given direction, the stronger association it represents.
3. If the order of the rows or the order of the columns is reversed (but not both), the new value of θ is the inverse of the original value. This ordering is usually arbitrary, so whether we get $\theta = 4$ or $\theta = 0.25$ is simply a matter of how we label the rows and columns.

4. The odds ratio does not change when the table orientation reverses so that the rows become the columns and the columns become the rows. This also means that the OR takes the same value when it is defined using the conditional distribution of X given Y as it does when defined using the distribution of Y given X. That is, it treats the variables symmetrically. This is the big difference between OR and RR or OR and the difference of proportions.

Confidence Intervals for Odds Ratio

A 2×2 table for 2 variables X and Y has 4 cell counts $n_{11}, n_{12}, n_{21}, n_{22}$, the sample OR is

$$\hat{\theta} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}}$$

A $100\%(1-\alpha)$ confidence interval for the real odds ratio θ is formed by

$$exp\{\log(\hat{\theta}) \pm z_{\alpha/2} \times ASE(log\hat{\theta})\}$$

where ASE is called Alpha's Standard Error and is given by:

$$ASE(\log \hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

If we want to get a 95% Ci for the population OR then we'll use $\alpha = 0.05$, and $z_{\alpha/2} = 1.96$.

If the CI contains 1, that means the population OR might be 1, hence the two variables X and Y might be independent.

Prospective Versus Retrospective Studies

	Positive Outcome (Disease)	**Negative Outcome (No Disease)**
Exposure	a	b
No Exposure	c	d

Table 10.2: Suitable Methods to Assess the Association

Consider a 2×2 contingency table as shown above, where X is the exposure variable with 2 outcomes (exposure and no exposure) and Y with 2 outcomes (yes or no). There are at least 3 measures of association available:

1. Difference between the two conditional proportions: $p_1 - p_2 = \frac{a}{a+b} - \frac{c}{c+d}$
2. Relative risk: p_1/p_2
3. Odds Ratio: $\hat{\theta} = \frac{a/b}{c/d} = \frac{ad}{bc}$

However, not all the samples are suitable (or valid) to use all the 3 measures. Only the prospective studies can use all these measures to quantify the association between X and Y, whereas the retrospective studies can use odds ratio only.

Prospective Study:

- Sample subjects randomly from a population.

- Either randomly assign the exposure variable to the subjects or record their exposure variable status.
- Follow the subjects over time to see if they develop the disease.

Main advantage: We can obtain valid estimate of p_1 and p_2 from the 2×2 table. Hence, all the 3 measures of association are valid for this kind of study.

Retrospective Study (or Case-Control Study):

- Sample a group of cases (people with the disease)
- Sample a group of controls (people without the disease)
- Check each subject to see if they were exposed or not.

Although it is cheap, quick and involves fewer subjects (especially if the disease is rare), the huge disadvantage is that we **cannot** obtain valid estimate of π_1 and π_2 from the table since we need the estimate of `P(disease|exposure)` whereas from the retrospective study we got `P(exposure|disease)`

10.3.2 Chi-Squared (χ^2) Test for $r \times c$ Tables

Definition 10.3.1 (Independence and Dependence (Association))**.** *Two categorical variables are **independent** if the population conditional distributions for one of them are identical at each category of the other.*
*The variables are **dependent**, or associated, if the conditional distributions are not identical.*

We now learn about a hypothesis test (χ^2 test) for association between two categorical variables.

The chi-square test has the following hypotheses:

$$H_0 : \text{The two variables are independent}$$

$$H_1 : \text{The two variables are dependent}$$

In order to find the evidence against null hypothesis H_0, we'll assume X and Y are independent as stated under H_0 and calculate what are the cell counts should be, called **expected counts**. The more these expected counts are similar as the observed counts, the weaker the evidence (against H_0); if these expected counts are very different from the observed counts then we got strong evidence against H_0.

For a particular, cell, the expected count is

$$\text{Expected count} = \frac{\text{Row total} \times \text{Column Total}}{\text{Total Sample Size}}$$

In short, test statistic is the evidence that the data provide. It will summaries how far the observed counts are from the expected counts. A larger value of the test statistic will give stronger evidence against the null hypothesis.

The formula for test statistic χ^2 (with continuity correction) is:

$$\chi^2 = \sum \frac{(|\text{observed count - expected count}| - 0.5)^2}{\text{expected count}}$$

Note that there are variations on this formula in other books/documents.

Expected counts are not necessarily integers. If all the expected counts are larger than 5, it's suitable to use chi-square test to test the independence. **If there is at least one expected count lesser than 5, we say sample is of small size, and there is another test for this situation.**

p-value helps to quantifies how strong the evidence against H_0 is. p-value is calculated from χ^2 distribution with degree of freedom 1, which is the the area in the right of the test statistic value. The smaller p-value the stronger the evidence against H_0 is.

Conclusion: If p-value is small (< 0.05) we can say: data provide strong evidence against H_0. If p-value is moderately small (between 0.05 and 0.2) we say data provide evidence against H_0 but not strong. If p-value is large (> 0.2) we say data do not provide enough evidence against H_0.

Given a pre-specified significance level α (the common α is 0.05), if p-value $< \alpha$ we'll reject H_0; Otherwise we do not reject H_0.

If any of the expected counts in the 2×2 table is lesser than 5, we should use Fisher exact test. This test use the same hypotheses as the chi-square test, the procedures to perform the test is similar, however the test statistic is obtained differently and the way p-value is calculated is also different.

Dependent Samples in 2×2 Tables

Consider this example: There are 50 students taking a statistical course. The lecturer gave a test on R at the beginning of the course. There were 26 students who passed and 24 of them failed. After taking the course which can help students improve their ability working with R, another test was given. This time, 42 students passed and 8 of them failed. Does taking this course help students to improve their ability working with R?

The samples for Before and for After are the collected from the same set of 50 students. Hence, the two samples are dependent. Thus, the chi-square test or Fisher exact test should not be used for this case.

	Pass Test 2)	**Fail Test 2**
Pass Test 1	25	1
Fail Test 1	17	7

Table 10.3: Dependent Samples Contingency Table

McNemar's test should be used in this case. If the table is having the cell counts denoted as a, b, c, d then the idea is: to see if the statistical course effectively improve students' ability in using R, we need to check the count of b, c. 24 students who failed in Before, now 17 of them move to Pass under After ($c = 17$), whereas 26 who passed under Before now 1 of them move to Fail under After ($b = 1$). The moving of these numbers 17 and 1 is just a randomness or is it because of the dependence? The hypotheses are:

H_0 : The results of Test 1 and Test 2 are independent, i.e., course has no effect

H_1 : The results of Test 1 and Test 2 are dependent, i.e., the course has an effect

The test statistic (used for the case when sample is large enough):

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

or when sample has a small cell count, we can use the test statistic with correction:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c}$$

The test statistic above follows a χ^2_1 distribution.

10.4 Chi-Squared χ^2 Test for $r \times c$ Tables

It's very common that we want to check the association between two nominal variables where one of them or both have more than 2 outcomes. The more special case is one or both variables are ordinal.

Note 10.4.1. *We are still checking the association between **2** variables - the only difference now is that each of the variables can have an arbitrary number of categories.*

As such, the χ^2 test can be extended to tables larger than 2 by 2. In general suppose that we have r rows and c columns that define two categorical random variables. Expected value in each cell is computed exactly the same way as for the 2×2 table. The only difference is that the χ^2 distribution to use is the $\chi^2_{(r-1)(c-1)}$ distribution, i.e., the degree of freedom is $(r-1)(c-1)$.

Standardized Residuals

Test statistics and p-value describe the evidence against H_0. A cell-by-cell comparison of observed and estimated expected frequencies is necessary to help us to understand better the nature of evidence.

Definition 10.4.2 (Standardized or Adjusted Residual)**.**

$$r_{ij} = \frac{n_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}(1-p_{i+})(1-p_{+j})}}$$

where n_{ij} is the observed count in row i and column j (cell ij); μ_{ij} is the expected count for cell ij under H_0; p_{i+} is the marginal probability of row i and p_{+j} is the marginal probability of column j.

$\sqrt{\mu_{ij}(1-p_{i+})(1-p_{+j})}$ is the estimated standard error of $(n_{ij} - \mu_{ij})$ under H_0.

The residuals r_{ij} can be derived by the output of the chi-squared test.

When H_0 is true, each r_{ij} has a large-sample standard normal distribution. If $|r_{ij}| > 2$ in any cell, it indicates a lack of fit of H_0 in that cell.

A large positive residual means that the observed count is much higher than the expected count whereas a large negative residual means that the observed count is much smaller than the expected count under H_0.

Some Comments about Chi-Squared Tests

1. Pearson's χ^2 test only indicate the degree of evidence for an association, but they usually can not answer other questions about dataset. It is often better to study the nature of the association, rather than relying solely on these tests.

2. For any $r \times c$ table where $r \geq 2, c \geq 2$, χ^2 test is not always applicable since they require large samples (so that the sampling distribution of χ^2 statistic can be closer to chi-squared distribution. **The approximation is poor when more than 25% of the cell counts have expected values less than 5 or when** $n/(IJ) < 5$. So, we do not perform chi-squared test if the first condition is violated.

3. Another test that is equivalent to the chi-squared test is the likelihood ratio test.

4. χ^2 does not depend on the order in which the rows and columns are listed. Thus they ignore some information when there is ordinal variable.

Table With Ordinal Variable

It is quite common to assume that as the levels of X increases, responses on Y tend to increase or to decrease towards higher levels of X.

To detect a trend association, most simple and common analysis assigns scores to categories and measures the degree of linear trend or correlation.

Let $u_1, \ldots, u_I$ denote scores for the rows and let $v_1, \ldots, v_J$ denote scores for the columns. The scores have the same ordering as the category level, they should reflect distances between categories (greater distances between categories regarded as further apart).

A test for the association of 2 ordinal variables is linear-by-linear test. Its null hypothesis is: two variables are independent; the alternative hypothesis is: two variables are dependent.
The test statistic is computed by

$$M^2 = (n-1)r^2$$

where r is the sample correlation between X and Y.

The formula for r is given by:

$$r = \frac{\sum_{i,j}(u_i - \bar{u})(v_j - \bar{v})p_{ij}}{\sqrt{[\sum_i (u_i - \bar{u})^2 p_{i+}]\left[\sum_j (v_j - \bar{v})^2 p_{+j}\right]}}$$

where $\bar{u} = \sum_i u_i p_{i+}$ is the sample mean of row scores, $\bar{v} = \sum_j v_j p + +j$ is the sample mean of column scores.
Note that: $p_{ij} = n_{ij}/n; p_{i+} = n_{i+}/n; p_{+j} = n_{+j}/n$

For large samples, the test statistic M^2 has approximately a chi-squared distribution with 1 degree of freedom.

One issue with this is that the choice of scores assigned to each category may affect the result (i.e,. the p-value). So, it is usually better to use one's own judgement by selecting scores that reflect distances between categories.

Chapter 11

R commands

The most important function to remember is `help`. It takes in another command/function as argument and gives a description of what the function is, and how to use it. For example, `help(hist)` explains how to plot histograms using `hist`.

11.1 Matrices and Vectors

```
## to create a vector
v <- c(1, 2, 3) # c stands for "concatenate"
v <- c(T, T, F, T) # T = True, F = False
v <- numeric(3) # v = (0, 0, 0)

## using the rep function (rep = "replicate")
v <- rep(c(1, 2), 3) # v = (1, 2, 1, 2, 1, 2)
v <_ rep(c(6, 3), c(2, 4)) # 6 and 3 are replicated 2, 4 times, respectively.
                           # v = (6, 6, 3, 3, 3, 3)

## using the seq function (seq = "sequence")
v <- seq(from=1, to=10, by=2) # v = (1, 3, 5, 7, 9)
v <- seq(from=2, to=10, length=5) # v = (2, 4, 6, 8, 10)
v <- seq(10) # v = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

## an alternative (shorthand) for seq is:
v <- 1:5 # v = (1, 2, 3, 4, 5)
v <- 1:5*2 # v = (2, 4, 6, 8, 10)

## appending item(s) to an existing vector
numbers <- c(2, 4, 6, 8, 10)
strings <- c("hello", "world")
c(numbers, 12, 14) # a vector of numbers (2, 4, 6, 8, 10, 12, 14)
c(strings, 12, 14) # a vector of strings, ("hello", "world", "12", "14)

## to create a matrix from a vector (filled column-wise)
# 1 3 5
# 2 4 6
v <- matrix(c(1, 2, 3, 4, 5, 6), nrow=2, ncol=3)

## to create a matrix filled row-wise
# 1 2 3
```

```r
# 4 5 6
v <- matrix(c(1, 2, 3, 4, 5, 5), nrow=2, ncol=3, byRow=T)

## using dim function
v <- c(1:6)
dim(v) <- c(2, 3) # v = matrix of 2 rows, 3 columns (filled column-wise)
                  # 1 3 5
                  # 2 4 6
dim(v)=NULL # convert v back to vector

## concatenate matrices/vectors

# stacks one on top of the other
# 1 2 3
# 4 5 6
rbind(c(1, 2, 3), c(4, 5, 6))

## stacks one beside the other
# 1 4
# 2 5
# 3 6
cbind(c(1, 2, 3), c(4, 5, 6))

## row and column names
rownames(matrix) <- c("Male", "Female")
colnames(matrix) <- c("Rich", "Poor")

## numerical analysis
rowMeans(matrix) # means of every row
colMeans(matrix) # means of every column
max(x) # maximum value of vector x
min(x) # minimum value of vector x
sum(x) # \sum_{i=1}^{n} x_i
mean(x) # \frac{1}{n}\sum_{i=1}^{n} x_i
range(x) # max(x) - min(x)
sort(x) # sorted version of x
var(x) # variance \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})
sd(x) # standard deviation \sigma_x
cov(x, y) # covariance of x and y = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}
cor(x, y) # correlation coefficient \rho_{x,y} = \frac{Cov(x,y)}{\sigma_x\sigma_y}
```

11.2 Dataframes

R handles data in objects known as dataframes. A dataframe is an object with rows and columns. The rows contain different observations or measurements and the columns contain the value of different variables.
All the values of the same variable must go in the same column

```r
## converting vector/matrix to dataframe
v <- c(1:6)*2 # v = (2, 4, 6, 8, 10, 12)
m <- matrix(v, 2, 3) # matrix of 2 rows, 3 columns, filled column-wise
```

```r
df <- data.frame(m)
names(df) # [1] "X1" "X2" "X3" (these are default column names)

## converting collection of vectors to dataframe (as columns)
a <- c(11, 12)
b <- c(13, 14)
df <- data.frame(a, b)
names(df) # [1] "a" "b" (vector names are used as column names)

## reading a csv file
data <- read.csv(filename, sep=",", heeader=False)
attach(data) # allows us to access columns without specifying the dataframe

## names of columns of dataframe
names(data) # [1] "color" "weight" "species"
colnames(data) <- c("color", "weight (in kg)", "species")

## row names are automatically assigned and labelled as "1", "2", ...
row.names(data) = c("Row 1", "Row 2") # can be re-named if desired

## Relabelling column data
Gender <- ifelse(Gender == "0", "M", "F")
Cost <- strtoi(gsub("[,\$]", "", Cost), 10) # removes comma and dollar sign, and converts
                                            # string to integer

# accessing dataframes
dataframe[1, 2] # returns the value of the 1st observation in the 2nd column
dataframe[1:5, ] # returns all columns corresponding to first 5 observations
dataframe[gender == "M", ] # returns all observations whose gender is "M"
dataframe[Gender == "F" & CA2 > 90, ] # returns all observations whose gender is M and
                                      # CA2 > 90
which(gender == "M") # indices of observations whose gender value is "M"
subset(data, gender == "M")
nrow(data[which(gender == "M"), ]) # number of rows satisfying condition

## combining dataframes by rows
# creates a dataframe from first 3 rows of data
data1 <- data.frame(data[1:3,])
# creates a dataframe from other 3 rows of data
data2 <- data.frame(data[4:6,])
# combining two dataframes above
data <- rbind(data1, data2)
```

Note that if the variables (names of columns) are not the same in both dataframes, an error message will be displayed when you try to use `rbind`.

```r
## combining dataframes by variables (columns)
data <- cbind(data4, data5)
```

Note that if the variables (names of columns) do not have the same length in both dataframes, an error message will be displayed when you try to use `cbind`.

```r
## Use merge command to merge 2 dataframes by a common variable
merge(data1, data2, by="Subject", all=TRUE)
```

The `merge` command in R is quite similar to the `JOIN` function in SQL. To see more details of `merge`, read the documentation here.

```
## Sorting a dataframe
data[order(CA1), ] # sorts by CA1 in ascending order
data[rev(order(CA2)), c("Subject", "CA1", "CA2")] # sorts by CA2 in descending order
                                                  # and only selects 3 columns specified.
```

11.3 Reading Data Files

There are several ways of reading/importing data files into R:

1. `scan` offers a low-level reading facility: read data into a vector or list from the console or file
2. `read.table` can be used to read dataframes from free format text files (.txt files)
3. `read.csv` can be used to read dataframes from files using comma to separate values (.csv files)
4. `read.fwf` can be used to read files that have a fixed width format.

When reading from an Excel file, a simple method is to save each worksheet separately as a csv file and use `read.csv` on each saved csv file.

When the first line of a file contains the names of variables, then we use: `header = TRUE`.

```
## using the scan() function
> v <- scan()
1: 1 2 3 4 5
6: 4 3 2 1
10: 1 1 1 1
14:
Read 13 times
> v
  [1] 1 2 3 4 5 4 3 2 1 1 1 1 1

## importing a free format data file
data <- read.table("C:/Data/crab.txt", header = TRUE)

## storing variable names when header = FALSE
varnames <- c("Subject", "Gender", "CA1", "CA2", "HW")
data <- read.table("C:/Data/ex.txt", header = FALSE, col.names = varnames)
```

Note that missing values of variable names are denoted by NA.

```
## importing csv data
data <- read.table("C:/Data/ex.txt", sep = ",") # csv files don't have to be .csv
data <- read.csv("C:/Data/ex.txt") # alternative to above
```

When importing a Fixed Width Format (FWF) file, we can specify the widths of variables in a vector and use the `read.fwf` function.

```
data <- read.fwf("C:/Data/ex.txt", width=c(2, 1, 3, 3, 1))
```

11.3.1 Importing Binary Files

Binary data generated from other statistical software can be read into R (but it should be avoided).

The R package `foreign` provides import facilities for some other statistical software.

Activate the package by typing `library(foreign)`

The following functions are available:

1. `read.spss` reads in SPSS files.
2. `read.mtp` imports Minitab worksheets.
3. `read.xport` reads in SAS files in TRANSPORT format.
4. `read.S` reads in binary objects produced by S-plus.

11.4 Loops in R

There are 2 kinds of loops in R: `while` loops and `for` loops. We will just take a look at examples to understand how to use them as they're quite simple (and very similar to other programming languages).

```
## find the sum of the first 10 integers
x <- 0
S <- 0
while (x <= 10) {
    S <- S + x
    x <- x + 1
}
S

## to print all the squares of integers from 1 to 5
x = 0
test = TRUE # or: test = 1
while (test > 0) {
    x = x + 1
    test = isTRUE(x < 5) # or test = x < 5
    cat(x^2, "\n")
}
```

Note:

- We use `isTRUE` to check if an expression evaluates to TRUE
- We use `cat` to print to screen. It is a variadic function (accepts variable number of arguments) and merges all of them using a separator (default: space-separated)

```
## find the sum of first i numbers from i = 1 to 10
x = numeric(10)
for (i in 1:10) {
    s = 0
    for (j in 1: i) {
```

```
        s = s + j
    }
    x[i] = s
    cat("The sum of the first", i, "numbers = ", x[i], "\n")
}
```

11.5 Redirecting Output in R

We use the `sink` function to send objects and text to a file. This is useful when we want to keep a copy of the output in a file or when the contents of an object or function are too big to display on screen.

```
sink("C:/Data/datasink_ex.txt")
x = numeric(10)
for (i in 1:10) {
    s = 0
    for (j in 1: i) {
        s = s + j
    }
    x[i] = s
    cat("The sum of the first", i, "numbers = ", x[i], "\n")
}
sink()
```

The function `write.table` or `write.csv` can be used to write dataframes to a file.

```
write.table(data, "C:/Data/ex.txt")
```

11.6 User-defined Functions

Characteristics of a function in R:

- Has a name
- Has parameters (0 or more)
- Has a body
- Returns something

Note that R returns the last output of a function by default, even if you do not explicitly provide a return statement.

Some examples:

```
## function to find standard error of the mean of a vector: SE = σ/√n

se <- function(x) {
    sqrt(var(x)/length(x))
}

## function to find the median of a given vector
my_median = function(x) {
  y = sort(x)
```

```
  m = NULL
  mid = length(x) / 2
  if (length(x) %% 2 == 1) {
    m = y[mid + 1]
  } else{
    m = (y[mid] + y[mid + 1])/2
  }
  m
}
```

Note:

- `sort` does not modify the original vector, it returns a sorted version
- `%%` is the modulo operator in R
- The return value of the function body is the return value of the last statement executed.

11.7 Matrix Operations

```
## Let A, B, C be matrices such that
dim(A) # m n
dim(B) # n p
dim(C) # m n

## matrix operations
A + C # adding two matrices
A - C # subtracting two matrices
A * C # element-wise product (Hadamard product)
A / C # element-wise division
A %*% B # matrix multiplication of A and B
t(A) # transpose of A
crossprod(A, B) # equivalent to t(A) %*% B
det(A) # determinant of matrix
solve(A) # inverse of matrix
diag(A) # diagonal elements of square matrix
diag(c(1, 2, 3)) # diagonal matrix with elements 1, 2, 3
diag(n) # identity matrix of order n
eigen(A)$values # eigenvalues in decreasing order
eigen(A)$vectors # corresponding eigenvectors
svd(A) # signular value decomposition of matrix
```

11.8 Single Categorical Variable Analysis

```
table(columnName) # generates frequency table on a categorical column
prop.table(table(columnName)) # returns a proportion table
barplot(table(gender), ylab="Frequency", xlab="Gender", col=c(2, 5),
        main="Gender Frequency Bar Plot") # creates bar plot
pie(table(columnName)) # creates pie chart
```

11.9 Single Quantitative Variable

```r
summary(marks) # returns 5 number summary, and the mean
quantile(marks, 0.8) # returns the q^th quantile
IQR(marks) # Interquantile Range = Q_3 - Q_1
var(marks) # variance
sd(marks) # standard deviation
marks[order(marks)[1:5]] # smallest 5 observations
size <- length(marks) # sample size
mark[order(mark)[(size-4):size]] # 5 largest observations

# function to calculate sample skewness
skew <- function(x) {
    n <- length(x)
    m3 <- mean((x - mean(x))^3)
    m2 <- mean((x - mean(x))^2)
    sk = m3/m2^(3/2) * sqrt(n*(n-1))/(n-2)
    return (sk)
}

# function to calculate kurtosi
kurt <- function(x) {
    n = lengthx(x)
    m4 = mean((x - mean(x))^4)
    m2 = mean((x - mean(x))^2)
    k = (n-1)/((n-2)*(n-3)) * ((n+1)*m4/(m2^2) - 3*(n-1))
    return(k)
}

hist(marks, prob=F, col=2, xlab="Marks", ylab="Number of students",
        main="Histogram of marks") # density histogram


# full statistical summary - 5-number summary, number of observations, outliers
boxplot.stats(marks, coef=1.5)
boxplot(marks) # plots a boxplot

# using ggplot2 for boxplots
ggplot(data) + geom_boxplot(aes(y = mark))

# QQ plot
qqnorm(data, xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",
    main = "QQ Plot", pch = 20)
qqline(data, col = "red")

# using ggplot2 for QQ plots
ggplot(data) + geom_qq(aes(sample = mark)) + geom_qq_line(aes(sample = mark), color = "red")

# using ggplot2 for plotting histogram
library(ggplot2)
ggplot(data) +
    geom_histogram(aes(x = mark), data = data, binwidth = 3, fill = "grey", color = "red")
```

```r
# histogram with density plot overlay
hist(mark, freq=FALSE, xlab="mark", ylab="density", axes=TRUE, nclass=10)
x <- seq(0, 30, length.out=98) # should match sample size of mark
y <-dnorm(x, mean(mark), sd(mark))
lines(x, y, col = "red")

# using ggplot2 for histogram with density plot overlay
p <- ggplot(data) +
    geom_histogram(aes(x = mark, y = ..density..),
                   binwidth = 3, fill = "grey", color = "black")
x <- seq(0, 30, length.out=98)
y = dnorm(x, mean(mark), sd(mark))
df <- data.frame(x = x, y = y)
p + geom_line(data = df, aes(x = x, y = y), color = "red")
```

Note 11.9.1. *To add a histogram on top of another, you can specify* `add=TRUE` *as an argument of the* `hist` *function.*
To specify the number of bins of a histogram, you can use the `breaks` *argument of* `hist`*.*

Robust Estimators of Location and Scale

```r
# 20% trimmed mean
mean(x, trim = 0.2)

# winsorized mean and variance
winsor <- function(x, alpha = 0.2) {
    n = length(x)
    xq = n * alpha
    x = sort(x)
    m = x[round(xq) + 1]
    M = x[n - round(xq)]
    x[which(x < m)] = m
    x[which(x > M)] = M
    return(c(mean(x), var(x)))
}
winsor(x)

# Median Absolute Deviation (MAD)
median(abs(x - median(x)))
mad(x) # estimate of σ = 1.4826 × MAD

library(MASS)
hubers(x, k=0.84) # this gives the 20% winsorized mean
```

11.10 Two Categorical Variables

```r
table(rowCategory, columnCategory) # contigency table
prop.table(contigencyTable, margin=1) # row-wise proportion table.
                                      # Use margin=2 for column-wise proportion table
barplot(contingencyTable, beside=F,
```

```
       legend=rownames(contingencyTable)) # Use beside=T for creating separate bar
                                          # chart based on row-values

# example:
chest.pain<-matrix(c(46,474,37,516), ncol=2, byrow=2)
dimnames(chest.pain)<-list(Gender=c("Male", "Female"), CP=c("Yes","No"))

## 2-sample test for equality of proportions without continuity correction:
test<-prop.test(chest.pain,correct=FALSE)
# relative risk
RR<-(test$estimate[1])/(test$estimate[2])
# odds
odds<-test$estimate/(1- test$estimate)
# odds ratio
OR<-odds[1]/odds[2]

## function to compute OR and CI of OR (returns OR, ASE, CI and confidence level)
OR <- function(x, pad.zeros = FALSE, conf.level = 0.95) {
    if(pad.zeros) {
        if(any(x==0)) {x <- x + 0.5}
    }
    theta <- x[1,1]*x[2,2]/(x[2,1]*x[1,2])
    ASE<-sqrt(sum(1/x))
    CI<-exp(log(theta) +c(-1,1)*qnorm(0.5*(1+conf.level))*ASE)
    list(estimator=theta, ASE=ASE,conf.interval=CI,conf.level=conf.level)
}
# chi-square test (works even for > 2 categories)
chisq.test(chest.pain)
chisq.test(table)$stdres # gives the standard residuals of every cell

## fisher exact test
fisher.test(matrix, alternative = "two.sided")

## McNemar Test
mcnemar.test(x, correct = TRUE)

## linear-by-linear test (built-in functions)
## example of infant malformation and mother's alcohol consumption
library(coin) # COIN : COnditional INference
Input =(
"MI         Absent    Present
Alcohol
Zero         17066         48
Below.1      14464         38
1-2            788          5
3-5            126          1
>6              37          1
")
set = as.table(read.ftable(textConnection(Input)))
test = lbl_test(set,scores = list(MI = c(0,1), Alcohol = c(0,0.5,1.5,4,7)))

## manually performing linear-by-linear test
```

```
nc1<-c(17066,14464,788,126,37); # Column 1
nc2<-c(48,38,5,1,1);            # Column 2

rsum<-nc1+nc2;                  # Row sums
csum<-c(sum(nc1),sum(nc2));     # Column sums
n<-sum(csum)                    # total cell counts

rowp<-rsum/n                    # margin prob for rows
colp<-csum/n                    # margin prob for columns

pc1<-rsum*csum[1]/n;            # prediction of Column 1
pc2<-rsum*csum[2]/n;            # prediction of Column 2

v<-c(0,1);                      # specifying the scores for columns
u<-c(0,.5,1.5,4,7.0);           # specifying the scores for rows

ubar=sum(u*rowp);               # weighted average scores for rows
vbar<-sum(v*colp);              # weighted average scores for columns
CV<-sum(c(sum((u-ubar)*nc1/n),sum((u-ubar)*nc2/n))*(v-vbar))  # weighted covariance

V1<-sum((u-ubar)^2*rsum/n);     # weighted variance for rows' scores
V2<-sum((v-vbar)^2*csum/n);     # weighted variance for columns' scores

r<-CV/sqrt(V1*V2)               # weighted correlation

M<-sqrt(n-1)*r                  # Normalized test statistic

# one-sided p-value
1-pnorm(abs(M))                 # or pnorm(abs(M), lower.tail = FALSE)

M^2 # test statistic

# Test P-value (2 sided p-value)
1-pchisq(M^2,1)                 # or pchisq(M^2,1, lower.tail = FALSE)
```

11.11 One Categorical and One Quantitative Variable

```
bp <- boxplot(quant~categorical)
bp$out # values of outlier
grp = bp$group # outliers in each group
which(grp == 1) # outlier indices in group 1
bp$out[which(grp == 1)]

## histograms of multiple groups in R
# To specify that 3 graphs in one column in one page
par(mfrow=c(2,2))
# Histogram for the energy of type 1
hist(energy[which(type ==1)], include.lowest = TRUE,freq=FALSE,
col="grey",xlab = "Type 1",main ="Histograms of Energy by types")
hist(energy[which(type ==2)], include.lowest = TRUE, freq=FALSE,
col="grey", xlab = "Type 2",main = "")
```

```r
hist(energy[which(type ==3)], include.lowest = TRUE, freq=FALSE,
col="grey", xlab = "Type 3",main = "")
# To get back to 1 graph in one page.
par(mfrow=c(1,1))
```

11.12 Two Quantitative Variables

```r
plot(x, y, xlim=c(-10, 10), ylim=c(-10, 10)) # scatter plot with given axis limits
cor(x, y) # correlation coefficient
abline(lm((x, y)) # draw a regression line on top of a plot

# scatter plot of two quant variables grouped by categorical variable
plot(age, time,
     main = "Survival Time from Malignant Melanoma",
     xlab = "Age (in years)",
     ylab = "Survival Time (in days)",
     col = ifelse(Melanoma$sex == "1", "blue", "red"))
legend("topleft",
       pch = c(1, 1),
       c("Female", "Male"),
       col = c("red", "blue"))

# scatter plot of x10 and y, grouped by a categorical variable, x11
ggplot(data=data, aes(x=x10, y=y, color=x11)) +
  geom_point(size=2) +
  xlab("Weight of the car") +
  ylab("Gasoline mileage performance (mpg)") +
  ggtitle("Scatter plot of weight of car vs. mpg, classified by transmission type")
```

11.13 Random Variables and Sampling Distributions

```r
# generating 10 random numbers following a given distribution
rbinom(10, size=8, prob=0.5)
rnorm(10, mean=9, sd=5)
rt(10, df=14)
rexp(10, rate=1/100)
rpois(10, lambda=3)

# P(X ≤ q) for different distributions of X
pnorm(80, mean=100, sd=15) # P(X ≤ 80) where X ~ N(100, 225)
pbinom(10, size=14, prob=0.3, lower.tail=F) # P(X > 10) where X ~ Bin(14, 0.3)
pt(2, df=5) # P(X ≤ 2) where X ~ t_5

# finding q_0.9 for different distributions
qnorm(0.9, mean=10, sd=5)
qbinom(0.9, size=10, prob=0.4)
qt(0.9, df=10)
```

11.14 Hypothesis Testing

```r
## One sample independent t-test
# alternative = c("two.sided", "less", "greater")
t.test(x=data, mu=0, alternative="two.sided", var.equal=F, conf.level=0.95)

# Welch Two Sample t-test
t.test(x, y, mu=0, alternative="less", var.equal=F)

# Paired (Dependent) Samples t-test
t.test(x, y, mu=0, alternative="greater", paired=T)
t.test(x - y, mu=0, alternative="greater")

# Equal Variance F test
var.test(x, y) # H0: equal variance

# Normality test
# H0: sample is from a normal distribution
# H1: sample is not from a normal distribution
# test returns small value -> not from normal distribution
shapiro.test(x)

# Test associaiton of categorical variables
chisq.test(x, y) # H0: not associated
```

11.15 Linear Regression

```r
model = lm(y~x1 + x2 + x3) # fit linear regression model
summary(model) # get model statistics
confint(model, level=0.95) # confidence interval of coefficients
abline(model) # add regression line to plot

# Predicting using model
predict(model, c(10, 20, 30))
predict(model, c(10, 20, 30), interval="confidence", level=0.95)

# Checking model assumptions
rawRes = model$res
SR = rstandard(model)
which(SR > 3 | SR < -3) # indices of outliers
which(cooks.distance(model) > 1) # indices of influential points

# Plotting residual plots
par(mfrow=c(2, 2))
plot(model)
dev.off()
```

Chapter 12

Python Commands

Normally, jupyter notebook is used to run python code snippets because of ease of use and convenience (since you can run specific blocks of code). It is very popular in data science applications and machine learning.

There are many libraries available that we will be using to perform data manipulation and statistical analysis. For example, `numpy`, `pandas`, `matplotlib`, etc.

12.1 Vectors and Matrices

```
import numpy as np

# np.arange(start, stop, step)
np.arange(5) # [0, 1, 2, 3, 4]
np.arange(1, 5) # [1, 2, 3, 4]
np.arange(1, 10, 2) # [1, 3, 5, 7, 9]

# np.linspace(start, stop, num=50)
x = np.linspace(2, 3, 5) # [2.0, 2.25, 2.5, 2.75, 3.0]

np.ones(5) # [1, 1, 1, 1, 1]
np.zeroes(5) # [0, 0, 0, 0, 0]
```

Alternatively, you can also just create your own array in python and convert it to a numpy array using `np.array`

```
# np.asmatrix(array of arrays)
matrix = np.asmatrix([[1, 2, 3], [4, 5, 6]])
matrix.shape # (2, 3)

# creating identity matrices
np.identity(3) # [[1, 0],
               #  [0, 1]]

# creating diagonal matrices, or extracting diagonal from matrix
np.diag([3, 4])# [[3, 0],
               #  [0, 4]]
np.diag([[1, 2, 3], [4, 5, 6], [7, 8, 9]]) # [1, 5, 9]
```

Note that you can directly perform element-wise operations on numpy arrays and matrices (as opposed to regular arrays in python). So, you don't need to use `for` loops!

```
x = np.array([1, 2, 3, 4])
x * 2 # 2, 4, 6, 8
x + 1 # 2, 3, 4, 5
```

Matrix operations like transpose and inverse are also quite easy to perform:

```
matrix = np.asmatrix([[1, 2, 3], [4, 5, 6], [7, 8, 9]])
matrix.T # transpose of matrix
matrix.I # inverse of matrix
```

`numpy` also provides various methods to combine vectors and matrices, and perform operations along rows/columns (called axes). Other powerful functions that are commonly used are also listed below:

```
# stack rows on top of each other (vertically)
np.vstack([[1, 2, 3], [4, 5, 6]]) # [[1, 2, 3],
                                  #  [4, 5, 6]]

# stack columns side by side (horizontally)
np.hstack([[1, 2, 3], [4, 5, 6]]) # [1, 2, 3, 4, 5, 6]

# np.concatenate: a more general combining function
a = np.array([[1, 2], [3, 4]])
b = np.array([[5, 6]])
np.concatenate((a, b), axis=0) # [[1, 2],
                               #  [3, 4],
                               #  [5, 6]]
np.concatenate((a, b.T), axis=1) # [[1, 2, 5],
                                 #  [3, 4, 6]]

# sum of elements (can be axis-wise too)
a = np.array([[1, 2, 3], [4, 5, 6]])
np.sum(a) # 21
np.sum(a, axis=0) # [5, 7, 9] -> column-wise sum
np.sum(a, axis=1) # [6, 15] -> row-wise sum

# index of maximum value
np.argmax(a, axis=0) # [1, 1, 1] -> column-wise index of max
np.argmax(a, axis=1) # [2, 2] => row-wise index of max

# np.where(condition x, y): if true, yield x, otherwise yield y
a = [1, 2, 3, 4, 5, 6]
np.where(a < 4, a, 10*a) # [1, 2, 3, 40, 50, 60]

# returns an array of distinct elements in the array
np.unique([1, 2, 2, 1, 3, 1, 3]) # [1, 2, 3]

# returns the set difference of two arrays
np.setdiff1d([1, 2, 3, 4], [1, 3]) # [2, 4]

# returns the setwise intersection of two arrays
np.intersect1d([1, 2, 3, 4], [1, 3]) # [2, 4]
```

```
# cumulative sum of elements in the array/matrix
np.cumsum([1, 2, 3, 4, 5]) # [1, 3, 6, 10, 15]

# difference of consecutive elements
np.diff([1, 2, 4, 10, 20]) # [1, 2, 6, 10]
```

12.2 Useful Pandas functions

```
import pandas as pd

df = pd.read_csv("marks.csv")
df['midterm'] # accessing a particular columns
df['final'] = [100, 90, 99, 94, 80] # adding another column

df[df['gender'] == 'F'] # accessing rows based on column-value
df.loc[df['gender'] == 'F', :] # alternative using loc

pd.cut(df.finals, bins=[0, 50, 60, 70, 80, 101], labels=["F", "D", "C", "B", "A"],
    right=False) # convert marks to grade based on intervals

pd.concat([df1, df2], axis=1) # concatenate the two dataframes column-wise (side by side)

pd.concat([df1, df2], axis=0) # concatenate two dataframes row-wise (on top of each other)

df.rename({'midterm': 'Midterm Marks', 'finals': 'Final Marks'} # rename column names
df.columns = ['Midterm Marks', 'Final Marks'] # alternative way to rename column names

df.drop("Midterm Marks", axis=1) # drop column from table

df['marks'].replace(100, 99) # replace all 100 with 99 in marks column

df.sort_values('marks') # dataframe sorted by marks

pd.merge(df1, df2, on="id") # inner-join dataframes based on id column

df.to_csv("marks.csv") # writing to file
```

12.3 Single Quantitative Variable

Numerical analysis:

```
import numpy as np
import pandas as pd
import statistics as st
from statistics import mean
from statistics import median
from statistics import variance
data = pd.read_csv (r"C:\Data\midterm_marks")
data.columns = ['Obs', 'mark'] # changing the columns' name

mean(data['mark'])
```

```
median(data['mark'])
variance(data['mark'])
np.quantile(data['mark'],0.25)
np.quantile(data['mark'],0.5)
np.quantile(data['mark'],0.75)
min(data['mark'])
max(data['mark'])
variance(data['mark'])
st.stdev(data['mark']) #standard deviation
q75, q25 = np.percentile((data['mark']), [75 ,25])
iqr = q75 - q25
```

We can create user-defined functions to calculate the skewness and kurtosis too.

For all robust statistics, we can define our own functions (e.g. Gini's mean difference, trimmed mean, winsorized mean). As an example, we show how to find Gini's mean difference below:

```
# G = 1/(n choose 2) * sum_{i<j} |x_i - x_j|
import math
def gini(lst):
    res = 0
    length = len(lst)
    for i in range(length):
        for j in range(length):
            if i >= j:
                continue
            res += abs(lst[i] - lst[j])
    return res / math.comb(length, 2)
```

Note 12.3.1. *There's a slight rounding error/difference in the way* `scst.quantiles` *and* `np.quantile` *calculate the quantiles. It is* ***recommended to use*** `np.quantile` *for NUS modules.*

Graphical analysis (boxplots, histograms, QQ plots in Python)

```
import matplotlib.pyplot as plt
import pylab
import scipy.stats as scst

# histogram
plt.hist(data['mark'], bins=None, range=None, density=True, color='C1')
plt.title('Histogram of marks')
plt.xlabel('Values')
plt.ylabel('Probability Density')
plt.show()

# histogram with normal density overlaying
l = list(np.arange(0,30,0.5))
x = data['mark']
y = scst.norm.pdf(l,loc = mean(x),scale = st.stdev(x)) # this equivalent to qnorm in R
plt.plot(l, y)
plt.hist(data['mark'], bins=None, range=None, density=True, color='C2')
plt.show()
```

```
# histogram with its corresponding density overlaying
data['mark'].plot(kind = 'density', xlim = [0, 28])
plt.hist(data['mark'], bins=None, range=None, density=True, color='C2')
plt.show()

# boxplot
plt.boxplot(data['mark'])
plt.show()

# QQ-plot
scst.probplot(x, dist="norm", plot=pylab)
pylab.title('QQ Plot')
pylab.show()
```

12.4 Single Categorical Variable

```
# frequency table - can be generalised to pivot table
import pandas as pd
data = pd.read_csv("Data/bats.csv")
tab = pd.crosstab(index=data["type"], columns="count")
print(count)

# bar plot
import matplotlib.pyplot as plt
fig = plt.figure()
ax = fig.add_axes([0, 0, 1, 1])
types = ['1', '2', '3']
counts = [4, 12, 4]
ax.bar(types, counts)
plt.xlabel("type")
plt.yabel("frequency")
plt.show()
```

12.5 Association Between 2 Variables

12.5.1 Both Quantitative Variables

```
import numpy as np
import matplotlib.pyplot as plt

# pearons's correlation coefficient of 2 quantitative varisbles
scst.pearsonr(data['Midterm Marks'], data['Final Marks'])

# scatterplot to visualize associations
plt.scatter(midterm['M'], final['F'],label='Scatter Plot 1',color='b')
plt.xlabel('Midterm')
plt.ylabel('Final')
plt.title('Scatter Plot of Midterm and Final Marks')
plt.show()
```

```
# scatter plot of x10 and y, grouped by a categorical variable, x11
fig, ax = plt.subplots()
for car_type, color in zip([0, 1], ["red", "blue"]):
    subset = data.loc[data['x11'] == car_type]
    ax.scatter(subset['x10'], subset['y'], color=color,
        label= "automatic" if car_type == 1 else "manual")
ax.legend()
plt.title('Scatter plot of weight of car vs. mpg, classified by transmission type')
plt.show()
```

Given a scatterplot, you can mention the following characteristics:

- Is there any possible relationship between the 2 variables?
- Is the association positive or negative?
- Is the association linear or non-linear?
- Are some observations unusual, departing from the overall trend?
- Is the variance of the y-variable stable when the value of the x-variable is changed? (constant variance - homoscedasticity)
- Does the association depend on another categorical variable (which might indicate a possibility of Simpson's paradox)

12.5.2 One Categorical and One Quantitative

We can form boxplots or histograms grouped by the quantitative variable.

```
import pandas as pd
import matplotlib.pyplot as plt

# boxplot of energy consumption of bats grouped by type
bats = pd.read_csv('C:/Data/bats.csv')
fig, ax = plt.subplots(figsize=(7,5))
bats.boxplot(column=['energy'], by='type', ax=ax, color = 'b')
plt.show()

# histograms of multiple groups
fig, axes = plt.subplots(1, 3, figsize=(8,3), dpi=100, sharex=True, sharey=True)
colors = ['tab:red', 'tab:blue', 'tab:orange']
for i, (ax, type) in enumerate(zip(axes.flatten(), bats.type.unique())):
    x = bats.loc[bats.type==type, 'energy']
    ax.hist(x, alpha=0.5, bins=30, density=True, stacked=True,
        label=str(type), color=colors[i])
    ax.set_title(type)
plt.suptitle('Probability Histogram of Energy by types', y=1.05, size=16)
ax.set_xlim(0, 50); ax.set_ylim(0, 1);
plt.tight_layout();
```

12.5.3 Both Categorical Variables

```
# contingency table, OR, RR
data = {'Yes': [46,37], 'No': [474,516]}
df = pd.DataFrame(data, columns =['Yes', 'No'])

tab = [ df['Yes']/(df['Yes'] + df['No']),  df['No']/(df['Yes'] + df['No']) ]
tab = np.asmatrix([tab[0],tab[1]])
tab = np.transpose(tab)
print('Conditional Probabilities ', '\n', tab)

prob = df['Yes']/(df['Yes'] + df['No'])
print('RR = ', prob[0]/prob[1] ) # relative risk

odds = prob/(1-prob) # the odds of 'Yes'
print('OR = ', odds[0]/odds[1] ) # odds ratio

# chi-square test
import scipy.stats as scst
obs = np.array([[46, 474], [37, 516]])
scst.chi2_contingency(obs, correction = True) # returns [test statistic, p-value,
                                              #           degrees of freedom]
scst.fisher_exact(table, alternative = "two-sided") # [odds ratio, p-value]

# McNemar Test
from statsmodels.stats.contingency_tables import mcnemar
test1 = mcnemar(table, exact=False, correction=True) # approximate p-value
test2 = mcnemar(table, exact=True, correction=True)  # exact p-value, used when table
                                                     # has small cell count


# linear-by-linear built-in
import statsmodels.api as sm
ct = sm.stats.Table(np.asarray(table))
print(ct.test_ordinal_association(row_scores=row_scores, col_scores=col_scores))

# linear by linear test (user-defined)
def linear_by_linear_test(col1, col2, u, v):
    matrix = [col1, col2]
    row_total = [a + b for a,b in zip(col1, col2)]
    col_total = [sum(col1), sum(col2)]
    n = sum(row_total)

    row_p = [x / n for x in row_total]
    ubar = sum([a*b for a,b in zip(row_p, u)])
    col_p = [x / n for x in col_total]
    vbar = sum([a*b for a,b in zip(col_p, v)])

    numerator = 0
    for i in range(len(u)):
        for j in range(len(v)):
            ui = u[i]
            vj = v[j]
```

```
            pij = matrix[j][i] / n
            numerator += (ui - ubar) * (vj - vbar) * pij


    denominator = (sum([row_p[i] * (u[i] - ubar)**2 for i in range(len(u))]) * \
                   sum([col_p[j] * (v[j] - vbar)**2 for j in range(len(v))]))**0.5

    r = numerator / denominator
    M = (n-1) * r**2
    p_value = 1- scst.chi2.cdf(M, 1)
    return "M is {} and two sided p-value is {}".format(M,p_value)
```