CM 2110 Calculus and Statistical Distributions

Dr. Priyanga D. Talagala

2020-09-18

Contents

Course Syllabus

Pre-requisites

CM 1110

Remark:

This course module contains two main sections: (1) mathematics and (2) statistics. This syllabus is designed for the statistics section. Lectures for mathematics section and statistics section are conducted by two lecturers as two separate sub modules (1.5 hour lectures/Week). End Semester Examination is conducted as a single examination.

Learning Outcomes

On successful completion of this module, students will be able to plan more carefully the design of experiment in advance which provide evidence for or against theories of cause and effect and make inferences about population characteristics based on sample information and thereby solve data analysis problems in different application domains. (R(https://cran.r-project.org/) and RStudio are also freely available to install on your own computer). Get the Open Source Edition of RStudio Desktop. RStudio allows you to run R in a more user-friendly environment.

Outline Syllabus

- Functions of Several Variables
- Linear Algebra
- Coordinate Systems & Vectors
- Differential Equations
- **Statistical Distributions**
- **Estimation**

- **Hypothesis Testing**
- **Contingency tables**
- **Design of Experiments**

Method of Assessment

- Mid-semester examination
- End-semester examination

Recommended Texts

- Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2, pp. 337-472). Pacific Grove, CA: Duxbury.
- Mood, A.M., Graybill, F.A. and Boes, D.C. (2007): Introduction to the Theory of Statistics, 3rd Edn. (Reprint). Tata McGraw-Hill Pub. Co. Ltd.
- Montgomery, D. C. (2017). Design and analysis of experiments. John wiley & sons.

Lecturer

Dr. Priyanga D. Talagala

Schedule

Lectures:

- Friday [9.15 am - 10.45 am]

Tutorial:

- Friday [11.00 am - 12.30 pm]

Consultation time:

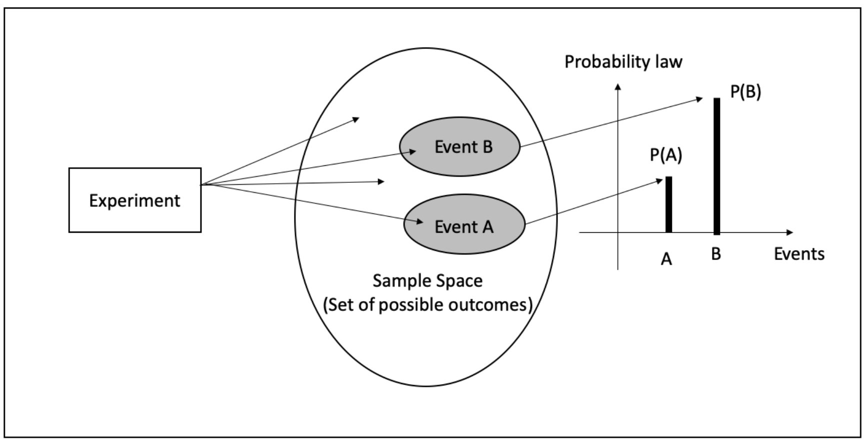
- Friday [8.15 am to 9.00 am]

Chapter 1

Statistical Distributions

Recap: CM 1110-Probability

Axioms of probability



- **Probability** of an event quantifies the **uncertainty**, randomness, or the possibility of occurrence the event.
- The probability of event E is usually denoted by $P(E)$.
- Mathematically, the function $P(.)$ is a set function defined from sample space (Ω) to $[0, 1]$ interval, satisfying the following properties.
- These are called the **'axioms of probability'**.

- **Axiom 1:** For any event A, $P(A) \geq 0$
- **Axiom 2:** $P(\Omega) = 1$
- **Axiom 3:**
  - (a) If $A_1, A_2, \dots, A_k$ is a finite collection of mutually exclusive events, then
    $$P(A_1 \cup A_2 \cup \cdots \cup A_k) = \sum_{i=1}^{k} P(A_i)$$
    (b) If $A_1, A_2, \dots$ is an infinite collection of mutually exclusive events, then
    $$P(A_1 \cup A_2 \cup \dots) = \sum_{i=1}^{\infty} P(A_i)$$

NOTE

- Axioms 1 and 2 imply that for any event E, $0 \leq P(E) \leq 1$.
- $P(E) = 1 \iff$ the event E is certain to occur.
- $P(E) = 0 \iff$ the event E cannot occur.

Methods for determining Probability

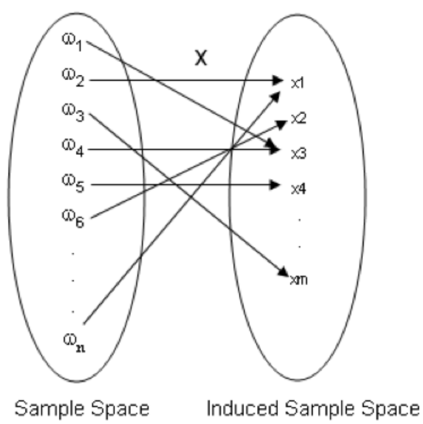
- There are several ways for determining the probability of events.
- Usually we use the following methods to obtain the probability of events.
  - Classical method
  - Relative frequency method (Empirical approach)
  - Subjective method
  - **Using probability models**

1.1 Random Variable

- Some sample spaces contain quantitative (numerical) outcomes, others contain qualitative outcomes.
- Often it is convenient to work with sample spaces containing numerical outcomes.
- A function that maps the original sample space into the real numbers is called a 'random variable'.
- This is more useful when the original sample space contains qualitative outcomes.

Definition 1: Random Variable

Let Ω be a sample space. Let X be a function from Ω to $\Re$ (*i.e.* $X : \Omega \rightarrow \Re$). Then X is called a random variable.



- A random variable assigns a real number to each outcome of a sample space.
- In other words, to each outcome of an experiment or a sample point ω_i, of the sample spaces, there is a unique real number x_i, known as the value of the random variable X.
- The range of the random variable is called the *induced sample space.*
- *A note on notation:* Random variables will always denoted with uppercase letters and the realized values of the random variable (or its range) will be denoted by the corresponding lowercase letters. Thus, the random variable X can take the value x.
- Each outcome of a sample space occurs with a certain probability. Therefore, each possible value of a random variable is associated with a probability.
- Any events of a sample space can be written in terms of a suitably defined random variable.

1.1.1 Types of Random Variables

- A random variable is of two types
  - Discrete Random Variable
  - Continuous Random Variable

1.1.1.1 Discrete Random Variable

- If the induced sample space is discrete, then the random variable is called a **discrete random variable**.

Example 01 Consider the experiment of tossing a coin. Express the following events using a suitably defined random variable

$H =$ *The event of getting a head*

$T =$ *The event of getting a tail*

Example 02

Consider the experiment of rolling of a die. Express the following events using a suitably defined random variable

$A =$ *The event that the number faced up is less than 5*

$B =$ *The event that the number faced up is even*

$C =$ *The event that the number faced up is 2 or 5*

Example 03

Consider the experiment of tossing a coin 10 times. Then the sample space Ω contains $2^{10} = 1024$ outcomes. Each outcome is a sequence of 10 H's and T's.

Express the following events in terms of a suitably defined random variable.

$D =$ *The event that the number of heads is 5*

$E =$ *The event that the number of tails is less than 4*

1.1.1.2 Continuous Random Variable

- If the induced sample space is continuous, then the random variable is called a **continuous random variable.**

Example 04

Consider the experiment of measuring the lifetime (in hours) of a randomly selected bulb. Express the following events in terms of a suitably defined random variable.

$F =$ *The event that the lifetime is less than 300 hours*

$G =$ *The event that the lifetime is 1000 hours*

1.2 Probability Mass Function

Definition 2: Discrete density function of a discrete random variable

If X is a discrete random variable with distinct values $x_1, x_2, \dots, x_n, \dots$, then the function, denoted by $f_X(.)$ and defined by

$$f_X(x) = \begin{cases} P(X = x) & \text{if } x = x_j, j = 1, 2, \dots, n, \dots \\ 0 & \text{if } x \neq x_j \end{cases} \tag{1.1}$$

is defined to be the discrete density function of X.

- The values of a discrete random variable are often called *mass points.*
- $f_X(x)$ denotes the *mass* associated with the *mass point* x_j.
- ***Probability mass function*** *discrete frequency function* and *probability function* are other terms used in place of *discrete density function*
- Probability function gives the measure of probability for different values of X.

1.2.1 Properties of a Probability Mass Function

- Let X be a discrete random variable with probability mass function $f_X(x)$. Then,

1. For any $x \in \Re$, $0 \leq f_X(x) \leq 1$.
2. Let E be an event and $I = \{X(\omega) : \omega \in E\}$. Then $P(E) = P(X \in I) = \sum_{x \in I} f_X(x)$.
3. Let $R = \{X(\omega) : \omega \in \Omega\}$. Then $\sum_{x \in \Re} f_X(x) = 1$.

1.2.2 Representations of Probability Mass Functions

Example 05

Consider the experiment of tossing a fair coin. Let

$$X = \begin{cases} 0 & \text{if the outcome is a Tail} \\ 1 & \text{if the outcome is a Head} \end{cases} \tag{1.2}$$

Find the probability mass function of X. Is X discrete or continuous?

1.2.2.1 Using a table

1.2.2.2 Using a function

1.2.2.3 Using a graph

1.3 Probability Density Function

- Let X be a continuous random variable.

- Then, it is not possible to define a pmf f_x with properties mentioned in Section 1.2. **Why?**

- Instead, we can find a function f_x with the some different properties.

- Probability density function (pdf) of a continuous random variable is a function that describes the relative likelihood for this random variable to occur at a given point.

1.3.1 Properties of a Probability Density Function

Let X be a continuous random variable with probability density function f_x. Then,

1. For any $x \in \Re$, $f_X(x) \geq 0$.
2. Let E be an event and $I = \{X(\omega) : \omega \in E\}$. Then $P(E) = P(X \in I) = \int_I f_X(x)dx$.
3. Let $R = \{X(\omega) : \omega \in \Omega\}$. Then $\int_{\Re} f_X(x)dx = 1$.

1.3.2 Existence of pdf

- To see the existence of such a function, consider a continuous random variable X,
- Suppose that we have a very large number of observations, N, of X, measured to high accuracy (large number of decimal places).
- consider the following grouped frequency table and the histogram constructed from those data.
- The height of the bar on a class interval of this histogram is equal to the relative frequency per unit in that class interval.

Interval	Class boundaries	Class frequency	Height of the bar	Area of the bar
I_1	$x_1 - \delta x/2, x_1 + \delta x/2$	n_1	$\frac{n_1}{\delta x * N}$	$\frac{n1}{N}$
I_2	$x_2 - \delta x/2, x_2 + \delta x/2$	n_2	$\frac{n_2}{\delta x * N}$	$\frac{n2}{N}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_k	$x_k - \delta x/2, x_k + \delta x/2$	n_k	$\frac{n_k}{\delta x * N}$	$\frac{nk}{N}$
Total				

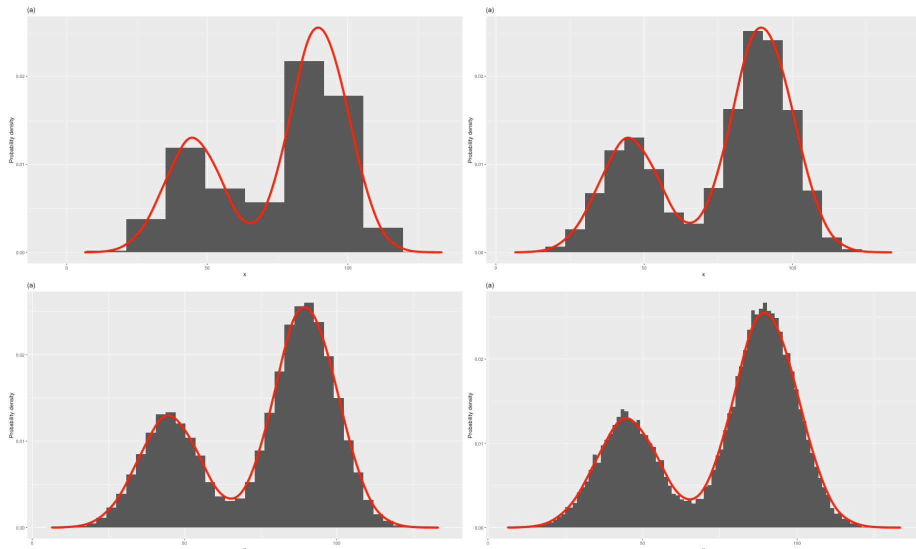


Figure 1.1: Histograms with different class intervals and a possible model for the pdf

- Then, for the i^{th} interval,

$$P(x_i - \frac{\delta x}{2} \leq X \leq x_i + \frac{\delta x}{2}) \approx \text{Area of the bar}$$

and therefore

$$\text{Height of the bar} \approx \frac{\text{Area of the bar}}{\delta x} \approx \frac{P(x_i - \frac{\delta x}{2} \leq X \leq x_i + \frac{\delta x}{2})}{\delta x}$$

- Therefore, the height of a bar represents the *probability density* in that class interval.
- When $\delta x \to 0$, it will allow us to approximate the histogram by a smooth curve as in Figure 1.1 (d).
- As the area under each histogram is 1, the area under the curve is also 1
- For any point x,

$$\text{The height of the curve} \approx \lim_{\delta x \to 0} \frac{P(x_i - \frac{\delta x}{2} \leq X \leq x_i + \frac{\delta x}{2})}{\delta x}$$

will represent the **the probability density at point** x.

- Let the above smooth curve be denoted by f_X.
- Then, f_X has the properties mentioned in Section 1.3.1.
- The function is called the **probability density function of** X.
- **NOTE** Here $f_X(x)$ represents **Probability density** *at point* x. Not *Probability at point* x.

1.3.3 Calculation of Probability using pdf

- Let $c, d \in \Re$ such that $c \leq d$. Then,

$$P(c \leq X \leq d) = \int_c^d f_X(x)dx$$

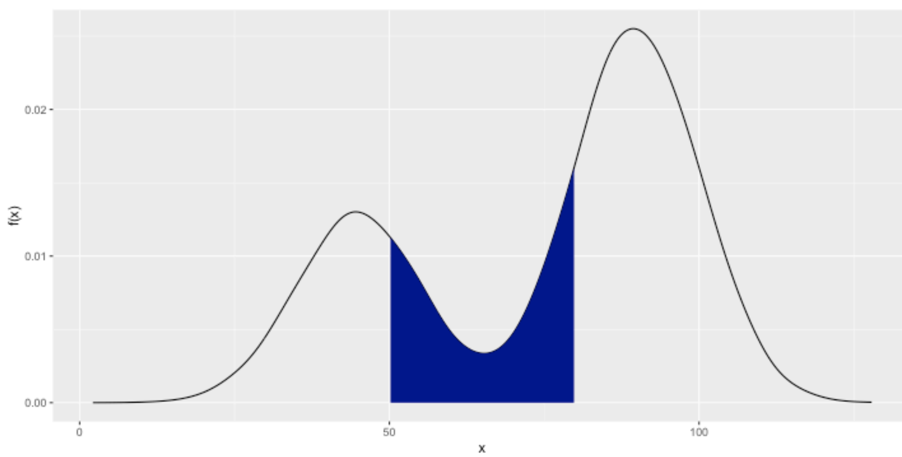


Figure 1.2: $P(c \leq x \leq d = \int_c^d f_X(x)dx$

- **NOTE:** if X is a continuous random variable with the p.d.f f_X, then for any $k \in \Re$,

$$P(X = k) = P(k \leq X \leq k) = \int_k^k f_X(x)dx = 0$$

- Therefore, for a continuous random variable X,

$$P(c < X < d) = P(c \leq X < d) = P(c < X \leq d) = P(c \leq X \leq d) = \int_c^d f_X(x)dx$$

1.4 Cumulative Distribution Function

- There are many problems in which it is of interest to know the probability that the values of a random variable is less than or equal to some real number x.

Definition 3: Cumulative distribution function

The *cumulative distibution function* or cdf of a random variable X, denoted by $F_X(x)$, is defined by

$$F_x(x) = P(X \leq x), \text{ for all } x$$

- Therefore, if X is a discrete random variable, the cdf is given by,

$$F_X(x) = \sum_{t \leq x} f_X(t), \quad -\infty < x < \infty$$

where $f_X(t)$ is the value of the pmf of X at t.

1.4.1 Relationship between cdf and pdf

- If X is a continuous random variable, the cdf is given by,

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \quad -\infty < x < \infty$$

where $f_X(t)$ is the value of the pdf of X at t. (Here t is a dummy integration variable).

- Conversely,

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Example 06

An owner of a software engineering company is interested in knowing how many years his employees stay with his company. Let X be the number of years an employee will stay with the company. Over the years, he has established the following probability distribution:

x	1	2	3	4	5	6	7
$f_X(x) = P(X = x)$	0.1	0.05	0.1	?	0.3	0.2	0.1

1. Find $f_X(4)$

2. Find $P(X < 4)$

3. Find $P(X \leq 4)$

4. Draw the probability mass function of X

5. Draw the cumulative distribution function of X

1.4.2 Properties of a cumulative distribution function of a Discrete random variable

Example 07

$$f_X(x) = \begin{cases} \frac{1}{25}x & 0 \leq x < 5 \\ \frac{2}{5} - \frac{1}{25}x & 5 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases} \tag{1.3}$$

1. Find the CDF of X
2. Find $P(X \leq 8)$
3. Find $P(3 \leq X \leq 8)$

1.

1.5 Expectations and Moments

1.5.1 Expectation

- The expected value, or expectation of a random variable is merely its average value.
- By weighting the values of the random variable according to the probability distribution, we can obtain a number that summarizes a typical or expected value of an observation of the random variable.

Definition 4: Expected value

Let X be a random variable. The *expected value* or *mean* of a random variable $g(X)$, denoted by $E[g(x)]$, is

$$E[g(x)] = \begin{cases} \sum_x g(x)f_X(x) & \text{if } X \text{ is a discrete random variable with pmf } f_X(x) \\ \int_x g(x)f_X(x)dx & \text{if } X \text{ is a continuous random variable with pdf } f_X(x) \end{cases} \tag{1.4}$$

- The mean of a random variable gives a measure of *central location* of the density of X.
- The process of taking expectations is a linear operation.
- For any constants a and b,

$$E(aX + b) = aE(X) + b$$

1.5.1.1 Properties of expected value

Theorem

a. $E(c) = c$ for a constant c
b. $E[cg(X)] = cE[g(X)]$ for a constant c
c. $E[c_1g_1(X) + c_2g_2(X)] = c_1E[g_1(X)] + c_2E[g_2(X)]$
d. If $g_1(x) \geq 0$ for all x, then $E[g_1(X)] \geq 0$
e. If $g_1(x) \geq g_2(x)$ for all x, then $E[g_1(X)] \geq E[g_2(X)]$
f. If $a \leq g_1(x) \leq b$ for all x, then $a \leq E[g_1(X)] \leq b$
g. If X and Y are two **independent** random variables, then $E(X \times Y) = E(X) \times E(Y)$

Example 08

Random variable X has the following pmf

$$f_X(x) = \begin{cases} 0.2 & x = 2 \\ 0.3 & x = 4 \\ 0.4 & x = 5 \\ 0.1 & x = 7 \end{cases} \tag{1.5}$$

1. Find $E(X)$
2. Find $E(X^2)$
3. Find $E\left(\frac{1}{X}\right)$
4. Find $E(2X + 3X^2 - 5)$

1.5.2 Moments

- The various moments of a distribution are an important class of expectation

Definition 5: Moments

If X is a random variable, the *rth moment of* X, usually denoted by μ_r', is defined as

$$\mu_r' = E(X^r).$$

if the expectation exists.

- Note that $\mu_1' = E(X) = \mu$, the mean of X.

Definition 6: Central moments

If X is a random variable, the *rth central moment of X about a* is defined as $E[(x-a)^r]$.

If $a = E(X) = \mu$, we have the *rth central moment of X, about* $E(X)$, denoted by μ_r, which is

$$\mu_r = E[(X - E(X))^r] = E[(X-\mu)^r].$$

- Find μ_1

Definition 7: Variance

If X is a random variable, $Var(X) = E[(X - E(X))^2] = E(X^2) - [E(X)]^2$ provided $E(X^2)$ exists.

- The *variance* of a random variable X is its second central moment, $Var(X) = E[(X - E(X))^2] = E[(X - \mu)^2]$
- The positive square root of $Var(X)$ is the *standard deviation* of X
- The *variance* of a random variable gives a measure of the degree of spread of a distribution around its mean.
- Let X be a random variable, and let μ be $E(X)$. the *variance* of X, denoted by σ^2 of $Var(X)$, is defined by

$$Var(X) = \begin{cases} \sum_x (x-\mu)^2 f_X(x) & \text{if } X \text{ is discrete with mass points } x_!, x_2, \dots, x_j \dots \\ \int_x (x-\mu)^2 f_X(x)dx & \text{if } X \text{ is continuous with probability density function } f_X(x) \end{cases} \tag{1.6}$$

1.5.2.1 Properties of variance of a random variable

Theorem

a. If c is a constant, then $V(cX) = c^2V(X)$
b. $V(c) = 0$, Variance of a constant is zero.
c. If X is a random variable and c is a constant, then $V(c + X) = V(X)$
d. If a and b are constants, then $V(aX + b) = a^2V(X)$
e. If X and Y are two independent random variables, then
    i. $V(X + Y) = V(X) + V(Y)$
    ii. $V(X - Y) = V(X) + V(Y)$

1.6 Models for Discrete Distributions

1.6.1 Discrete Uniform Distribution

A random variable X has a *discrete uniform* $(1, N)$ distribution if

$$f_X(x) = P(X = x) = \frac{1}{N}, \quad x = 1, 2, \dots, N$$

where N is a specified integer.

- The distribution puts equal mass on each of the outcomes $1, 2, \dots, N$.
- If X ha s a discrete uniform distribution, then $E(X) = (N + 1)/2$ and $Var(X) = (N^2 - 1)/12$.

1.6.2 Bernoulli Distribution

Bernoulli Trial

A random experiment of which the outcome can be classified into two categories is called a *Bernoulli trial*

- In general, the results of a Bernoulli Trial are called 'success' and 'failure'. We denote these results by S and F, respectively.
- Consider a Bernoulli trial. Let

$$X = \begin{cases} 0 & \text{if the Bernoulli trial results in a failure} \\ 1 & \text{if the Bernoulli trial results in a success} \end{cases} \tag{1.7}$$

- Suppose that the probability of a ''success' in any Bernoulli trial is θ.
- Then X is said to have a Bernoulli distribution with probability mass function

$$f_X(x) = P(X = x) = \theta^x (1 - \theta)^{1-x}, \quad x = 0, 1$$

- This is denoted as $X \sim Bernoulli(\theta)$.
- If X has a Bernoulli distribution, then $E(X) = \theta$ and $Var(X) = \theta(1 - \theta)$

1.6.3 Binomial Distribution

- A random experiment with the following properties is called a 'Binomial experiment'

1. The random experiment consists of a sequence of n trials, where n is fixed in advance of the random experiment.
2. Each trial can result in one of the same two possible outcomes: "success" (S) or "failure" (F)
3. The trials are independent. Therefore the outcome of any particular trial does not influence the outcome of any other trial.
4. The probability of "success" is the same for each trial. Let this probability is θ.

Binomial distribution

- Consider a binomial experiment with n trials and probability θ of a success.

A random variable X is defined to have a *binomial distribution* if the discrete density function of X is given by

$$f_X(x) = P(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \ \ x = 0, 1, 2, \dots, n$$

- This is denoted as $X \sim Bin(n, \theta)$.
- If X has a binomial distribution, then $E(X) = n\theta$ and $Var(X) = n\theta(1-\theta)$
- The binomial distribution reduces to the Bernoulli distribution when $n = 1$.

1.6.4 Geometric Distribution

- Consider a sequence of independent Bernoulli trials whose probability of "success" for each trial is θ.
- Let X = Number of failures before the first success
- Then, X is said to have a Geometric distribution with parameter θ.
- The probability mass function is given by

$$f_X(x) = P(X = x) = \theta(1-\theta)^x, \ \ x = 0, 1, 2, \dots$$

- This is denoted as $X \sim Geometric(\theta)$.
- If X has a geometric distribution, then $E(X) = (1-\theta)/\theta$ and $Var(X) = (1-\theta)/\theta^2$
- A random variable X that has a geometric distribution is often referred to as a discrete *waiting-time* random variable. It represents how long (in terms of number of failures) one has to wait for a "success."

1.6.5 Negative Binomial Distribution

- Consider a sequence of independent Bernoulli trials whose probability of "success" for each trial is θ.
- Let $X =$ Number of failures before the rth success
- Then, X is said to have a Negative Binomial distribution with parameter θ.
- The probability mass function is given by

$$f_X(x) = P(X = x) = \binom{x+r-1}{r-1}\theta^r(1-\theta)^x, \quad x = 0, 1, 2, ...$$

- This is denoted as $X \sim negbin(r, \theta)$.
- If X has a Negative Binomial distribution, then $E(X) = r(1-\theta)/\theta$ and $Var(X) = r(1-\theta)/\theta^2$
- If in the negative binomial distribution $r = 1$, then the negative binomial density specializes to the geometric density.

1.6.6 Hypergeometric Distribution

- Suppose a population of size N has M individuals of a certain kind ("success").
- A sample of n items is taken from this population without replacement.
- Let X be the number of successes in the sample.
- Then, X is said to have a hypergeometric distribution.
- The probability mass function is given by

$$f_X(x) = P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, 2, \dots, n$$

- Hypergeometric distribution can be used as a model for the number of "successes" in a sample of size n if the sampling is done without replacement from a relatively small population.
- If X has a Hypergeometric distribution, then $E(X) = n.\frac{M}{N}$ and $Var(X) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\left(\frac{N-n}{N-1}\right)$

1.6.7 Poisson Distribution

- The Poisson distribution provides a realistic probability model for the number of events in a given period of time, space, region or length.
- Example:
    - The number of fatal traffic accidents per week in a given city
    - The number of emails per hour coming into the company of a large business
    - the number of defect per unit of some material
- Poisson distribution is suitable if the following conditions hold.
    1. the number of events within non-overlapping time intervals are independent.
    2. Let t be a fixed time point. For a small time interval δt, the probability of exactly one event happening in the interval $[t, t + \delta t]$ is approximately proportional to the length δt of the interval. *i.e.,*

$$\frac{P(\text{exactly one event in } [t, t + \delta t])}{\delta t} \to \text{a positive constant}$$

as $\delta t \to 0$.

3. Let t be a fixed time point. For a small time interval δt, the probability of more than one event happening in the interval $[t, t + \delta t]$ is negligible. *i.e.,*

$$\frac{P(\text{more than one event in } [t, t + \delta t])}{\delta t} \to 0$$

as $\delta t \to 0$.

- Let X be the number of events during a time interval.
- Suppose that the average number of events during the interested time interval is $\lambda (> 0)$.
- Then, the distribution of X can be modeled by a Poisson Distribution with the probability mass function,

$$f_X(x) = P(X = x) = \frac{e^{-\lambda}(\lambda)^x}{x!}, \quad x = 0, 1, 2, ...$$

- This is denoted as $X \sim Poisson(\lambda)$.

- If X has a Poisson distribution, then $E(X) = \lambda$ and $Var(X) = \lambda$
- The Poisson distribution can be used for counts of some sort of a given area, space, region, volume or length as well.

Example 09

Phone calls arrive at a switchboard at an average rate of 2.0 calls per minute.

If the number of calls in any time interval follows the Poisson distribution, then

$X =$ number of phone calls in a given minute. $X \sim Poisson(\ )$

$Y =$ number of phone calls in a given hour. $Y \sim Poisson(\ )$

$W =$ number of phone calls in a 15 seconds. $W \sim Poisson(\ )$

1.7 Models for Continuous Distributions

1.7.1 Uniform Distribution

- The continuous uniform distribution is defined by spreading mass uniformly over an interval $[a, b]$.
- A random variable X is said to have a uniform distribution in (a, b) if its probability density function is given by

$$f_X(x) = \frac{1}{b-a}; \; a \leq x \leq b$$

- This is denoted as $X \sim U(a, b)$ or $X \sim Unif(a, b)$
- It is easy to check $\int_a^b f(x)dx = 1$.
- We also have

$$E(X) = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2}$$

$$Var(X) = \frac{(b-a)^2}{12}$$

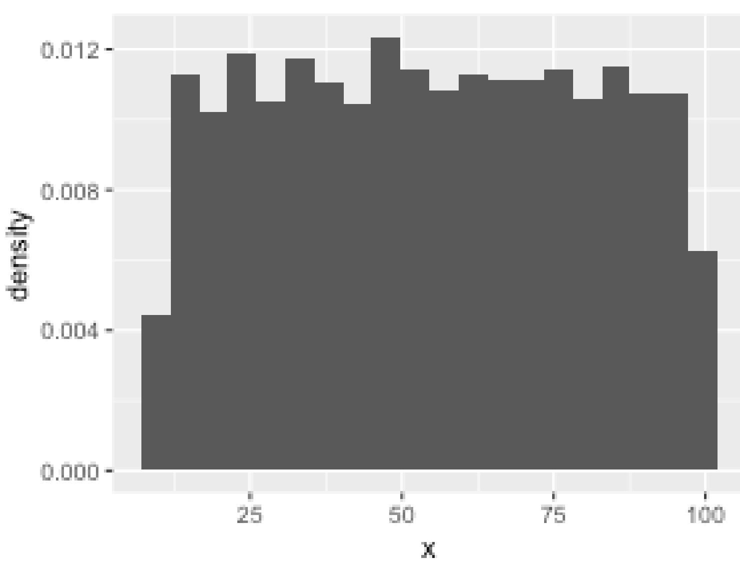


Figure 1.3: Uniform probabiity density

1.7.2 Normal Distribution (Gaussian Distribution)

- One commonly used bell shaped curve is called the normal distribution.
- Many techniques use din applied statistics are based upon the normal distribution.
- The normal distribution has two parameters, usually denoted by μ and σ^2, which are its mean and variance.
- A random variable X is said to have a normal distribution with location parameter μ and scale parameter σ, if its probability density function is given by,

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

- This is denoted by $X \sim N(\mu, \sigma^2)$.
- The normal density function is **symmetric around** the location parameter μ.
- The **dispersion of the distribution** depends on the scale parameter σ.
- If X is a normal random variable, $E(X) = \mu$ and $Vax(X) = \sigma^2$

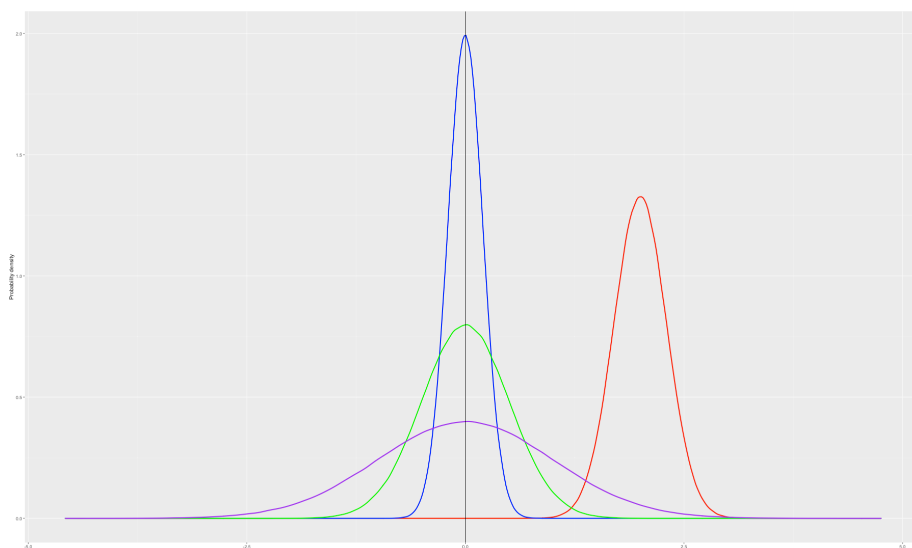


Figure 1.4: Normal distribution for different μ and σ

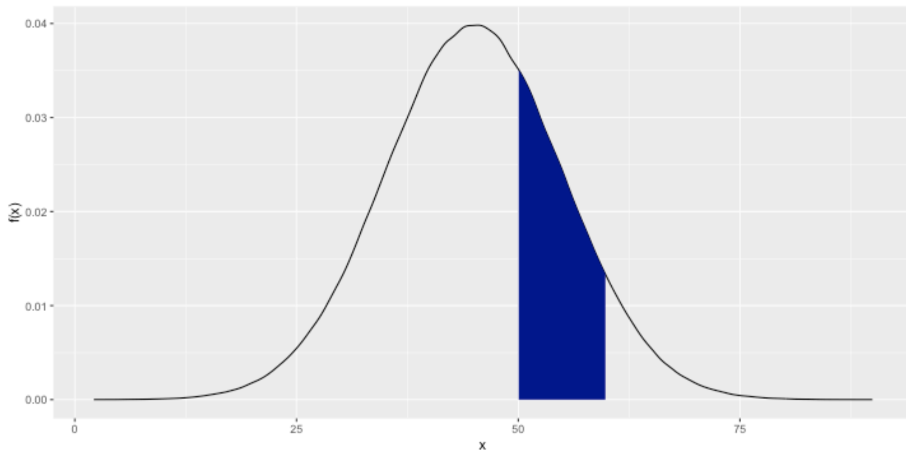


Figure 1.5: $P(a < X < b = \int_a^b f_X(x)dx$

$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

- Evaluating of this integration is somewhat tedious
- When we calculate this type of probabilities of normal distribution manually, it is convenient to use a normal probability table.

1.7.2.1 Standard Normal Distribution

- Normal distribution with $\mu = 0$ and $\sigma = 1$ is called the **standard normal distribution**.
- A random variable with standard normal distribution is usually denoted by Z.
- The probability density function of standard normal distribution is denoted by ϕ
- If $Z \sim N(0, 1)$, then

$$\phi_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}; \quad -\infty < z < \infty$$

- Probabilities related to Z can be found by using standard normal probability table.

Example 10

Let Z be a standard normal random variable. Calculate probabilities given in table below.

No.	Calculate this probability	Answer
1	$P(Z < 0)$	
2	$P(Z < 2.02)$	
3	$P(Z > 0.95)$	
4	$P(Z > -1.48)$	
5	$P(Z < -1.76)$	
6	$P(Z < 1.7)$	
7	$P(Z < -0.33)$	
8	$P(0.94 < Z < 2.41)$	
9	$P(-2.41 < Z < -0.94)$	
10	$P(-2.96 < Z < 1.05)$	

Example contd...

Cumulative Standard Normal Distribution

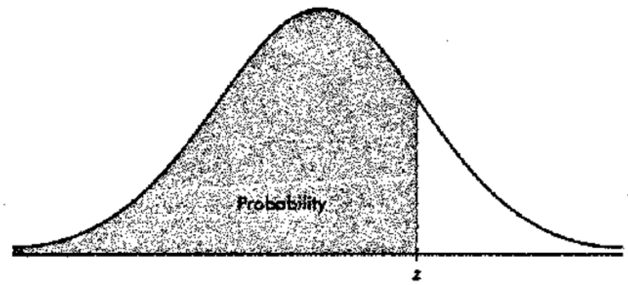


TABLE A: STANDARD NORMAL PROBABILITIES (CONTINUED)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Z_α **Notation**

- Z_α denotes the value such that $P(Z \geq Z_\alpha) = \alpha$
- Here α represents a probability.
- Therfore $0 \leq \alpha \leq 1$

Example 11

Find $Z_{0.025}$"

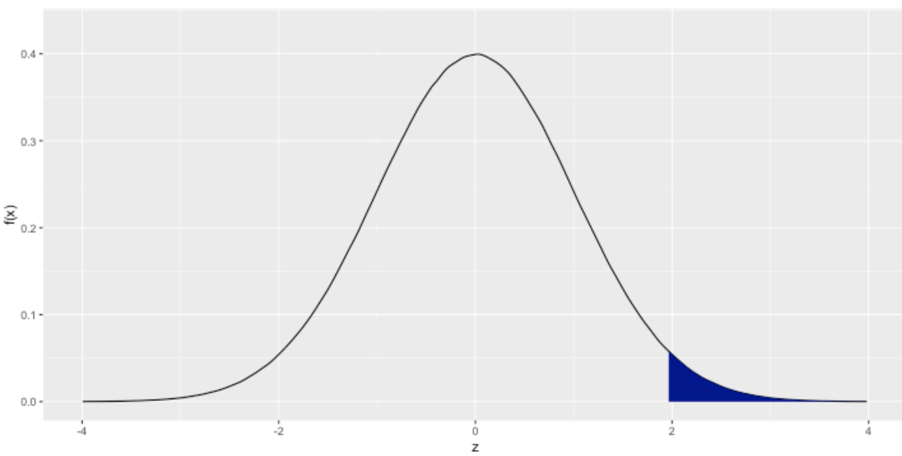


Figure 1.6: $P(a < X < b = \int_a^b f_X(x)dx$

Example 12

Find the following values

1) $Z_{0.01}$
2) $Z_{0.05}$
3) $Z_{0.9}$
4) $Z_{0.975}$
5) $Z_{0.85}$

Example contd ...

1.7.2.2 Calculation of Probabilities of Normal Distribution

- Suppose $X \sim N(\mu, \sigma^2)$.
- Let $Z = \frac{x-\mu}{\sigma}$.
- Then, $Z \sim N(0, 1)$
- This result can be used to find probabilities of any normal distribution.

Example 13

Let $X \sim N(10, 4)$. Calculate $P(X \geq 15)$

Example 14

Calculate the probabilities in the table below

No.	μ	σ	Calculate this probability	Answer
1	95	16	$P(104.92 \leq X \leq 115.16)$	
2	65	15	$P(X \leq 86.66)$	
3	96	20	$P(X < 86.3)$	
4	93	8	$P(91.24 \leq X \leq 109.34)$	
5	63	9	$P(65.55 < X < 76.61)$	
6	102	8	$P(X > 80.55)$	
7	79	18	$P(X < 131.15)$	
8	86	6	$P(X \leq 69.2)$	
9	85	2	$P(X < 86.46)$	
10	100	5	$P(X \leq 112.26)$	
11	58	10	$P(75.19 \leq X \leq 82.1)$	0.0348

No.	μ	σ	Calculate this probability	Answer
12	49	7	$P(X \geq 48.52)$	0.5273
13	103	17	$P(73.97 \leq X \leq 138.28)$	0.9371
14	99	24	$P(X < 82.8)$	0.2498
15	52	10	$P(X \leq 53.58)$	0.5628
16	72	8	$P(70.45 < X \leq 93.5)$	0.5732
17	82	20	$P(48.14 < X < 99.49)$	0.7639
18	94	15	$P(91.93 \leq X \leq 98.55)$	0.1741
19	45	4	$P(42.36 \leq X \leq 50.59)$	0.6643
20	73	1	$P(X \geq 72.38)$	0.7324

Example 15

Calculate the quantiles k in the table below

No.	μ	σ	Calculate k such that	Answer
1	85	6	$P(X < k) = 0.9936$	
2	97	23	$P(X < k) = 0.0694$	
3	77	5	$P(X > k) = 0.0002$	
4	93	12	$P(X > k) = 0.0023$	
5	67	3	$P(X < k) = 0.0197$	
6	59	5	$P(X > k) = 0.9756$	
7	94	13	$P(X > k) = 0.3228$	
8	51	4	$P(X < k) = 0.1515$	
9	49	10	$P(X > k) = 0.9693$	
10	61	13	$P(X < k) = 0.9946$	
11	69	14	$P(X < k) = 0.9357$	
12	85	5	$P(X > k) = 0.008$	
13	96	16	$P(X > k) = 0.0014$	
14	96	7	$P(X < k) = 0.2578$	
15	45	4	$P(X < k) = 0.2578$	

1.7.2.3 Empirical Rule for Normal Distribution

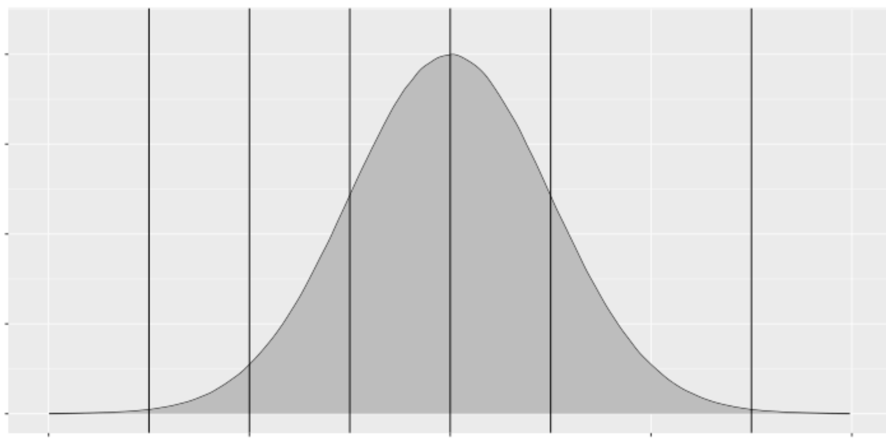


Figure 1.7: Empirical Rule for Normal Distribution

Calculate the following probabilities 1. $P(|x-\mu| \leq \sigma)$ 2. $P(|x-\mu| \leq 2\sigma)$ 3. $P(|x-\mu| \leq 3\sigma)$

Empirical Rule for Normal Distribution

Approximately 68% of the values in any normal distribution lie within one standard deviation, approximately 95% lie within two standard deviations and approximately 99.7% lie within three standard deviations from the mean.

- The normal distribution is somewhat special as its two parameters μ (the mean) and σ^2 (the variance), provide us with complete information about the exact shape and location of the distribution.

- Straightforward calculus shows that the normal distribution has its maximum at $x=\mu$ and inflection point (where the curve changes from concave to convex) at $\mu \pm \sigma$.

1.7.3 Gamma Distribution

- We come across with many practical situations in which the variable of interest has a skewed distribution.

- The gamma family of distributions is a flexible family of distributions on $[0, \infty)$ that yields a wide variety of skewed distributions

A random variable X is said to have a gamma distribution with shape parameter α and scale parameter β if its probability density function is given by

$$f_X(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, \quad 0 < x < \infty, \quad \alpha > 0, \quad \beta > 0$$

Here $\Gamma(\alpha)$ is called the *gamma function*,

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

- The gamma function satisfies many useful relationships, in particular,

1. $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha), \quad \alpha > 0$ (can be verified through integration by parts)
2. $\Gamma(1) = 1$
3. For any positive integer $n(> 0)$, $\Gamma(n) = (n - 1)!$
4. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

- When X has a gamma distribution with shape parameter α and scale parameter β, it is denoted as $X \sim gamma(\alpha, \beta)$
- The parameter α is known as the shape parameter, since it most influences the peakedness of the distibution
- The parameter β is called the scale parameter, since most of its influence is on the spread of the distribution.

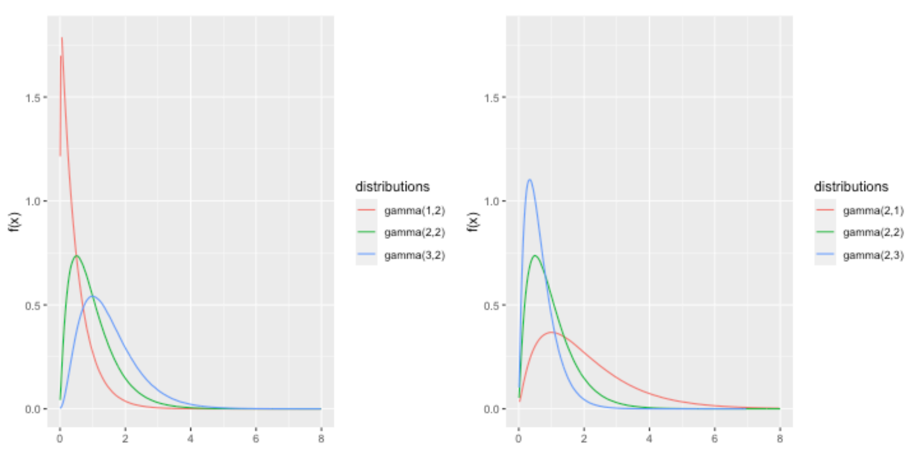


Figure 1.8: Gamma density functions

- If X has a gamma distribution with shape parameter α and scale parameter β, then

- $E(X) = \alpha\beta$
- $Var(X) = \alpha\beta^2$

1.7.4 Exponential Distribution

- This distribution is often used to model lifetime of various items.
- When the number of events in a time interval has a Poisson distribution, the length of time interval between successive events can be modeled by an exponential distribution.
- A random variable X is said to have an exponential distribution with scale parameter β, if its probability density function is given by

$$f_X(x;\beta) = \frac{1}{\beta}e^{-x/\beta}, \quad 0 < x < \infty$$

- This is denoted by $X \sim exponential(\beta)$

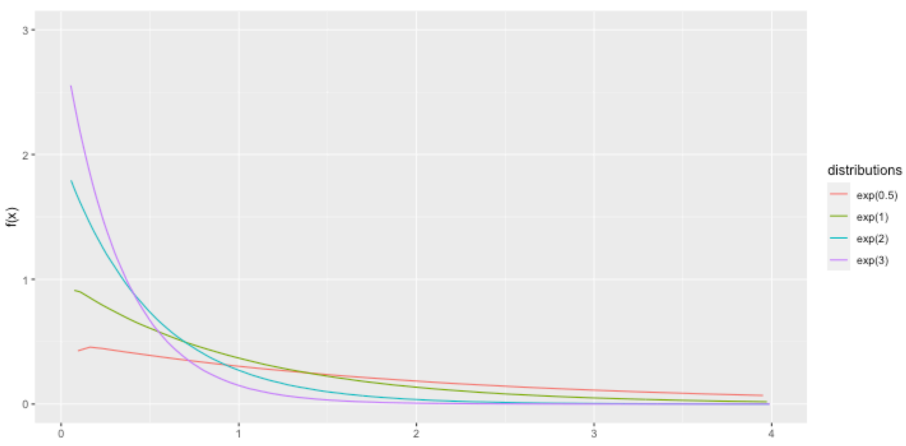


Figure 1.9: Exponential density functions

- Note that exponential distibution is a special case of the gamma distribution.
- It can be easily shown that $X \sim exponential(\beta) \iff X \sim gamma(1.\beta)$
- If X has an exponential distribution, then
  - $E(X) = \beta$
  - $Var(X) = \beta^2$

1.7.5 Beta Distribution

- The beta family of distributions is a continuous family on $(0, 1)$ indexed by two parameters.
- The $beta(\alpha, \beta)$ probability density function is

$$f_X(x; \alpha, beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \;\; 0 < x < 1, \; \alpha > 0, \;\; \beta > 0,$$

where $B(\alpha, \beta)$ denotes the *beta function*,

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx.$$

- The beta function is related to the gamma function through the following identity
$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$
- The beta distibution is one of the few common "named" distributions that give probability 1 to a finite interval, here taken to be $(0, 1)$.
- Therefore, the beta distibution is often used to model proportions, which naturally lie between 0 and 1.

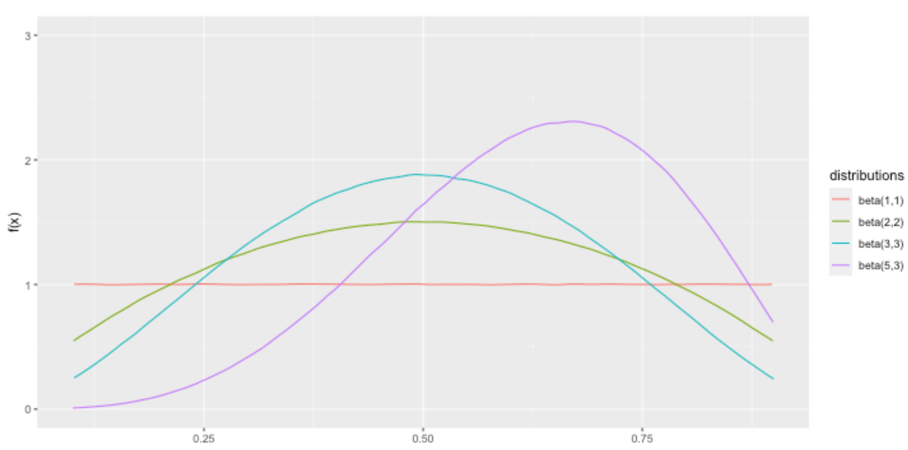


Figure 1.10: Beta density functions

1.8 Approximations

1.8.1 Poisson approximation to Binomial

Suppose $X \sim Bin(n, \theta)$ and n is large and θ is small. Then $X \sim Poisson(n\theta)$ and

$$f_X(x) \approx \frac{e^{-n\theta}(n\theta)^x}{x!}$$

Example 16

Suppose X has a binomial distribution with $n = 40$ and $p = 0.005$. Find $f_X(1)$

1.8.2 Normal approximation to Binomial

Suppose $X \sim Bin(n, \theta)$ and $n\theta \geq 5$ and $n(1-\theta) \geq 5$. Then $X \sim N(n\theta, n\theta(1-\theta))$ and

$$P(X \leq x)_{Binomial} = P(X \leq x + 0.5)_{Normal} \approx P(Z \leq \frac{x + 0.5 - n\theta}{\sqrt{n\theta(1-\theta)}})$$

Note: Since we are approximating a discrete distribution by a continuous distribution, we should apply a *continuity correction*

Example 17

Suppose X has a binomial distribution with $n = 40$ and $p = 0.6$. Find

1. $P(X \leq 20)$
2. $P(X < 25)$
3. $P(X > 15)$
4. $P(X \geq 20)$
5. $P(X = 30)$

1.8.3 Normal approximation to Poisson

Suppose $X \sim Poisson(\lambda)$ and $\lambda > 10$,

$$P(X \leq x)_{Poisson} = P(X \leq x + 0.5)_{Normal} \approx P(Z \leq \frac{x + 0.5 - \lambda}{\sqrt{\lambda}})$$

Note: Since we are approximating a discrete distribution by a continuous distribution, we should apply a *continuity correction*

Example 18

Suppose X has a Poisson distribution with $\lambda = 25$

1. $P(X \leq 20)$
2. $P(X < 25)$
3. $P(X > 15)$
4. $P(X \geq 20)$
5. $P(X = 30)$

1.9 Distribution of Functions of Random Variables

1. **Distribution of the linear transformation of a normal random variable**

Suppose that $X \sim N(\mu, \sigma^2)$. Let $Y = ax + b$, where a and b are constants. Then

$$Y \sim N(a\mu + b, a^2\sigma^2)$$

2. **Standardization of a normal random variable**

Suppose that $X \sim N(\mu, \sigma^2)$. Let $Z = \frac{X-\mu}{\sigma}$. Then

$$Z \sim N(0, 1).$$

3. **Distribution of the square of a standard normal random variable**

Suppose that $Z \sim N(0, 1)$. Let $Y = Z^2$. Then

$$Y \sim \chi^2_1$$

1.10 Distribution of Sum of Independent Random Variables

1. **Distribution of sum of i.i.d Bernoulli random variables**

Suppose that $X_1, X_2, \dots, X_n$ are independent, identically distributed (i.i.d) random variables with $Bernoulli(\theta)$ distribution. Let $Y = X_1 + X_2 + \dots + X_n$. Then

$$Y \sim Bin(n, \theta); \;\; y = 0, 1, 2, \dots, n.$$

2. **Distribution of sum of i.i.d Poisson random variables**

Suppose that $X_1, X_2, \dots, X_n$ are independent, identically distributed (i.i.d) random variables with $Poisson(\lambda)$ distribution. Let $Y = X_1 + X_2 + \dots + X_n$. Then

$$Y \sim Poisson(n\lambda); \;\; y = 0, 1, 2, \dots, .$$

3. **Distribution of sum of independent Poisson random variables**

Suppose that $X_1, X_2, \dots, X_n$ are independent random variables with $Poisson(\lambda_i),\; i = 1, 2, \dots, n$. Let $Y = X_1 + X_2 + \dots + X_n$. Then

$$Y \sim Poisson\left(\sum_{i=1}^{n} \lambda_i\right); \;\; y = 0, 1, 2, \dots, .$$

4. **Distribution of sum of i.i.d Geometric random variables**

Suppose that $X_1, X_2, \dots, X_n$ are independent, identically distributed (i.i.d) random variables with $Geometric(\theta)$ distribution. Let $Y = X_1 + X_2 + \dots + X_n$. Then

$$Y \sim Neg.bin(n, \theta); \;\; y = n, n+1, n+2, \dots, .$$

5. **Distribution of sum of independent Normal random variables**

Suppose that $X_1, X_2, \dots, X_n$ are independent random variables with $X_i \sim N(\mu_i, \sigma_i^2);\; i = 1, 2, \dots, n$. Let $Y = X_1 + X_2 + \dots + X_n$. Then

$$Y \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

6. **Distribution of sum of i.i.d exponential random variables**

Suppose that $X_1, X_2, \dots, X_n$ are independent, identically distributed (i.i.d) random variables with $X_i \sim exp(\lambda);\ i = 1, 2, \dots, n$. Let $Y = X_1 + X_2 + \cdots + X_n$. Then

$$Y \sim gamma(n, \lambda)$$

7. **Distribution of sum of independent gamma random variables**

Suppose that $X_1, X_2, \dots, X_n$ are independent random variables with $X_i \sim gamma(\alpha_i, \lambda);\ i = 1, 2, \dots, n$. Let $Y = X_1 + X_2 + \cdots + X_n$. Then

$$Y \sim gamma\left(\sum_{i=1}^{n} \alpha_i, \lambda\right).$$

1.11 Sampling Distribution

1. **Distribution of sample mean of a normal distribution**

Suppose that $X_1, X_2, \dots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2. Let $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ be the sample mean. Then,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

2. **Distribution of sample variance of a normal distribution**

Suppose that $X_1, X_2, \dots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2. Let $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ be the sample mean and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ be the sample variance. Then,

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

3. **Large sample distribution of sample average - Central Limit Theorem)**

Suppose that $X_1, X_2, \dots, X_n$ is a random sample from any distribution with mean μ and variance σ^2. Let $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ be the sample mean and $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$.

Then, the distribution of Z_n approaches the standard normal distribution as n approaches ∞.

References

Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2, pp. 337-472).Pacific Grove, CA: Duxbury

Mood, A.M., Graybill, F.A. and Boes, D.C. (2007): Introduction to the Theory of Statistics, 3rd Edn. (Reprint). Tata McGraw-Hill Pub. Co. Ltd.

Tutorial

1. Consider the experiment of taking two products randomly form a production line and determine whether each is defective or not. Express the following events using a suitably defined random variable.

$D_0 =$ *The event that both products are non defective*

$D_1 =$ *The event that one product is defective*

$D_2 =$ *The event that both products are defective*

$E =$ *The event that at least one product is defective*

2. Let X be a random variable with the following probability distribution

x	1	1.5	2	2.5	3	other
$f_X(x)$	k	$2k$	$4k$	$2k$	k	0

(a) Find the value of k
(b) Find $P(X = 2.5)$
(c) Calculate $P(X \geq 1.75)$

3. The sample space of a random experiment is $\{a, b, c, d, e, f\}$, and each outcome is equally likely. A random variable X is defined as follows:

Outcome	a	b	c	d	e	f
x	0	0	1.5	1.5	2	3

Determine the probability mass function of X. Use the probability mass function to determine the following probabilities:

(a) $P(X = 1.5)$
(b) $P(0.5 < X < 2.7)$
(c) $P(X > 3)$
(d) $P(0 \leq X < 2)$
(e) $P(X = 0 \text{ or } X = 2)$

4. Verify that the following function is a probability mass function, and determine the requested probabilities.

$$f_X(x) = \frac{8}{7}\left(\frac{1}{2}\right)^x, \text{ x= 1,2,3}$$

(a) $P(X \leq 1)$
(b) $P(X > 1)$
(c) $P(2 < X < 6)$
(d) $P(X \leq 1 \text{ or } X > 1)$

5. A disk drive manufacturer sells storage devices with capacities of one terabyte, 500 gigabytes, and 100 gigabytes with probabilities 0.5, 0.3, and 0.2, respectively. The revenues associated with the sales in that year are estimated to be \$50 million, \$25 million, and \$10 million, respectively. Let X denotes the revenue of storage devices during that year.

(a) Determine the probability mass function of X.
(b) Calculate the probability of getting more than \$20 million of revenue during that year.

(c) Determine the cumulative distribution function of X

6. Consider the following cumulative distribution function:

$$F_X(x) = \begin{cases} 0 & x < -2 \\ 0.2 & -2 \leq x < 0 \\ 0.7 & 0 \leq x < 2 \\ 1 & 2 \leq x \end{cases} \tag{1.8}$$

a) Draw the plot of $F_X(x)$
b) Discuss the properties of $F_X(x)$ (eg: whether it is discrete or continuous, whether it is decreasing or increasing function etc.)
c) Determine the probability mass function of X from the above cumulative distribution function:
d) Plot the probability mass function of X

7. The number of e-mail messages received per hour varies from 10 to 16 with the following probabilities

x = number of messages	10	11	12	13	14	15	16
$P(X = x)$	0.08	0.15	0.1	0.2	0.1	0.07	0.3

a) Let X be the number of e-mail messages received per hour. Find the probability mass function of X
b) Determine the mean and standard deviation of the number of messages received per hour

8. According to past data, twenty percent of all telephones of a certain type are submitted for service while under warranty. Of these, 60% can be repaired whereas the other 40% must be replaced with new units. If a company purchases ten of these telephones, what is the probability that

a) two telephones will be submitted for service under warranty?
b) at most 3 telephones will be submitted for service under warranty?
c) two telephones will end up being replaced under warranty?
d) one telephone will end up being repaired under warranty?

9. Each sample of water has a 10% chance of containing a particular organic pollutant. Assume that the samples are independent with regard to the presence of the pollutant. Find the probability that in the next 18 samples, exactly 2 contain the pollutant

a) Define a suitable random variable for the above question
b) Find the distribution of that random variable
c) Find the probability that in the next 18 samples, exactly 2 contain the pollutant

10. The space shuttle flight control system called Primary Avionics Software Set (PASS) uses four independent computers working in parallel. At each critical step, the computers "vote" to determine the appropriate step. The probability that a computer will ask for a roll to the left when a roll to the right is appropriate is 0.0001. Let X denotes the number of computers that vote for a left roll when a right roll is appropriate.

a) What is the probability mass function of X?
b) What are the mean and variance of X?

11. A University lecturer never finishes his lecture before the end of the hour and always finishes his lectures within 2 minutes after the hour. Let X = the time that elapses between the end of the hour and the end of the lecture and suppose the pdf of X is

$$f_X(x) = \begin{cases} kx^2 & 0 \leq x \leq 2 \\ 0 & otherwise \end{cases} \tag{1.9}$$

a) Find the value of k and draw the density curve.
b) What is the probability that the lecture ends within 1 min of the end of the hour?
c) What is the probability that the lecture continues beyond the hour for between 60 and 90 sec?
d) What is the probability that the lecture continues for at least 90 sec beyond the end of the hour?

12. The daily sales of gasoline are uniformly distributed between 2,000 and 5,000 gallons. Find the probability that sales are:

a) between 2,500 and 3,000 gallon
b) more than 4000 gallons
c) exactly 2500 gallons

13. Suppose X has a continuous uniform distribution over the interval $[-1, 1]$. Determine the following:

(a) Mean, variance, and standard deviation of X
(b) Value for k such that $P(-k < X < k) = 0.90$
(c) Cumulative distribution function

14. Suppose that the time it takes a data collection operator to fill out an electronic form for a database is uniformly between 1.5 and 2.2 minutes.

a) What are the mean and variance of the time it takes an operator to fill out the form?
b) What is the probability that it will take less than two minutes to fill out the form?
c) Determine the cumulative distribution function of the time it takes to fill out the form.

15. An electronic product contains 40 integrated circuits. The probability that any integrated circuit is defective is 0.01, and the integrated circuits are independent. The product operates only if there are no defective integrated circuits. What is the probability that the product operates?

16. A bag of 200 chocolate chips is dumped into a batch of cookies dough. 40 cookies are made from such a batch of dough. What is the probability that a randomly selected cookie has at least 4 chocolate chips?

17. For the case of the thin copper wire, suppose that the number of flaws follows a Poisson distribution with a mean of 2.3 flaws per millimeter.

a) Determine the probability of exactly two flaws in 1 millimeter of wire.

b) Determine the probability of 10 flaws in 5 millimeters of wire

18. Contamination is a problem in the manufacture of magnetic storage disks. Assume that the number of particles of contamination that occur on a disk surface has a Poisson distribution, and the average number of particles per square centimeter of media surface is 0.1. The area of a disk under study is 100 square centimeters.

a) Determine the probability that 12 particles occur in the area of a disk under study.

b) Determine the probability that zero particles occur in the area of the disk under study.

c) Determine the probability that 12 or fewer particles occur in the area of the disk under study

19. The number of surface flaws in plastic panels used in the interior of automobiles has a Poisson distribution with a mean of 0.05 flaw per square foot of plastic panel. Assume that an automobile interior contains 10 square feet of plastic panel.

a) What is the probability that there are no surface flaws in an auto's interior?

b) If 10 cars are sold to a rental company, what is the probability that none of the 10 cars has any surface flaws?

c) If 10 cars are sold to a rental company, what is the probability that at most 1 car has any surface flaws?

20. Cabs pass your workplace according to a Poisson process with a mean of five cabs per hour. Suppose that you exit the workplace at 6:00 p.m. Determine the following:

a) Probability that you wait more than 10 minutes for a cab.

b) Probability that you wait fewer than 20 minutes for a cab.

11. An electronic product contains 40 integrated circuits. The probability that any integrated circuit is defective is 0.01, and the integrated circuits are independent. The product operates only if there are no defective integrated circuits. What is the probability that the product operates?

7. A person must take two buses to go to work. From the past experience, he knows that a bus can come at any time within 6 minutes. Also, the probability that a bus comes within any period of the same length is the same. Hence, it is reasonable to assume the following probability density function for the waiting time X (in minutes) for a bus

$f_X(x) = \frac{1}{6}, \ \ 0 < x < 6.$

Then the total waiting time T at both bus-stops has the following density function

$$f_T(t) = \begin{cases} \frac{t}{36} & 0 \le t \le 6 \\ \frac{1}{3} - \frac{t}{36} & 6 \le t \le 12 \end{cases} \tag{1.10}$$

a) Verify that each of above function is a proper density function
b) What is the probability that the waiting time at the first bus-stop will be less than 2 minutes?
c) What is the probability that the waiting time at the first bus-stop will be less than 2 minutes and the waiting time at the second

Summary

Models for Discrete Distributions

Name	Remarks	Probability mass function	values of X	Parameter Space	Mean	Variance
Discrete Uniform	Outcomes that are equally likely (finite)	$f(x) = \frac{1}{N}$	$x = 1, 2, \ldots, N$	$N = 1, 2, \ldots$	$\frac{N+1}{2}$	$\frac{N^2-1}{12}$
Bernoulli	Bernoulli trial	$f(x) = \theta^x(1-\theta)^{1-x}$	$x = 0, 1$	$0 \leq \theta \leq 1$	θ	$\theta(1-\theta)$
Binomial	X = Number of successes in n fixed trials	$f_X(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$	$x = 0, 1, 2, \ldots, n$	$0 \leq \theta \leq 1$; $n = 1, 2, 3, \ldots$	$n\theta$	$n\theta(1-\theta)$
Geometric	X = Number of failures before the first success	$f_X(x) = \theta(1-\theta)^x$	$x = 0, 1, 2, \ldots$	$0 \leq \theta \leq 1$	$\frac{1-\theta}{\theta}$	$\frac{1-\theta}{\theta^2}$
Negative Binomial	X = Number of failures before the rth success	$f_X(x) = \binom{x+r-1}{r-1}\theta^r(1-\theta)^x$	$x = 0, 1, 2, \ldots$	$0 \leq \theta \leq 1$; $r > 0$	$\frac{r(1-\theta)}{\theta}$	$\frac{r(1-\theta)}{\theta^2}$
Hypergeometric	X = Number of successes in the sample taken without replacement	$f_X(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$	$x = 0, 1, 2, \ldots, n$	$N = 1, 2, \ldots$; $M = 0, 1, \ldots, N$; $n = 1, 2, \ldots, N$	$n\frac{M}{N}$	$n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(\frac{N-n}{N-1}\right)$

Name	Remarks	Probability mass function	values of X	Parameter Space	Mean	Variance
Poisson	X = number of events in a given period of time, space, region or length	$f_X(x) = \frac{e^{-\lambda}(\lambda)^x}{x!}$	$x = 0, 1, 2, ...$	$\lambda > 0$	λ	λ

Models for Continuous Distributions

Name	Probability density function	Values of X	Parameter Space	Mean	Variance
Uniform	$f_X(x) = \frac{1}{b-a}$	$a \leq x \leq b$	$-\infty < a < b < \infty$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal (Gaussian)	$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$-\infty < x < \infty$	$-\infty < \mu < \infty;\ \ \sigma > 0$	μ	σ^2
Gamma	$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$	$0 < x < \infty$	$\alpha > 0;\ \ \beta > 0$	$\alpha\beta$	$\alpha\beta^2$
Exponential	$f_X(x) = \frac{1}{\beta} e^{-x/\beta}$	$0 < x < \infty$	$\beta > 0$	β	β^2
Beta	$f_X(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$0 < x < 1$	$\alpha > 0;\ \ \beta > 0$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

Chapter 2

Estimations

2.1 Statistical Inference

- The process of making educated guess and conclusions regarding a population, using a sample of that population is called **Statistical Inference**.
- Two important problems in statistical inference are **estimation of parameters** and **tests of hypothesis**
- Estimation can be of the form of **point estimation** and **interval estimation**.

2.2 Point Estimation

Main Task

- Assume that some characteristic of the elements in a population can be represented by a random variable X.
- Assume that $X_1, X_2, \dots, X_n$ is a random sample from a density $f(x, \theta)$, where the form of the density is known but the parameter θ is unknown.
- The objective is to construct good estimators for θ or its function $\tau(\theta)$ on the basis of the observed sample values $x_1, x_2, \dots, x_n$ of a random sample $X_1, X_2, \dots, X_n$ from $f(x, \theta)$.

Definition: Statistic

Suppose $X_1, X_2, \dots, X_n$ be n observable random variables. Then, a known function $T = g(X_1, X_2, \dots, X_n)$ of observable random variables $X_1, X_2, \dots, X_n$ is called a **statistic**. A statistic is always a random variable.

Definition: Estimator

Suppose $X_1, X_2, \dots, X_n$ is a random sample from from a density $f(x, \theta)$ and it is desired to estimate θ. Suppose $T = g(X_1, X_2, \dots, X_n)$ is a *statistic* that can be used to determine and approximate value for θ. Then T is called an **estimator** for θ. An estimator is always a random variable.

Definition: Estimate

Suppose $T = g(X_1, X_2, \dots, X_n)$ be an estimator for θ. Suppose that $x_1, x_2, \dots, x_n$ is a set of observed values of the random variable $X_1, X_2, \dots, X_n$. Then *the value* $t = g(x_1, x_2, \dots, x_n)$ obtained by substituting the observed values in the estimator is called an **estimate** for θ.

- Therefore the **estimator** stands for the function of the sample, and the word **estimate** stands for the realized value of that function.
- *Notation:* An estimator of θ is denoted by $\hat{\theta}$. An estimate of θ is also denoted by $\hat{\theta}$. The difference between the two should be understood based on the context.

Parameter	Estimator: Using random sample $(X_1, X_2, \dots, X_n)$	Estimate 1: Using observed sample $(1, 4, 2, 3, 4)$	Estimate 2: Using observed sample $(4, 2, 2, 6, 3)$
μ	$\hat{\mu} = \bar{X}$	$\hat{\mu} =$	$\hat{\mu} =$
σ^2	$\hat{\sigma^2} = S^2$	$\hat{\sigma^2} =$	$\hat{\sigma^2} =$

2.2.1 Methods of finding point estimators

- In some cases there will be an obvious or natural candidate for a point estimator of a particular parameter.
- For example, the sample mean is a good point estimator of the population mean
- However, in more complicated models we need a methodical way of estimating parameters.
- There are different methods of finding point estimators
  - Method of Moments
  - Maximum Likelihood Estimators (MLE)
  - Method of Least Squares
  - Bayes Estimators
  - The EM Algorithm
- However, these techniques do not carry any guarantees with them
- The point estimators that they yeild still must be evaluated before their worth is established

2.2.1.1 Method of Moments

- Let $X_1, X_2, ... X_n$ be a random sample from a population with pdf or pmf $f(x;\theta)$, where $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ and $k \geq 1$.
- Sample moments m' and population moments μ' are defined as follows

Sample moment	Population moment
$m'_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\mu_1 = E(X)$
$m'_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$	$\mu_2 = E(X^2)$
...	...
$m'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$	$\mu_k = E(X^k)$

Each μ'_j is a function θ, i.e. $\mu'_j = \mu'_j(\theta_1, \theta_2, \dots, \theta_k)$ for $j = 1, 2, \dots, k$.

Method of Moments Estimators (MME)

We first equate the first k sample moments to the corresponding k population moments,

$$
\begin{aligned}
m'_1 &= \mu'_1, \\
m'_2 &= \mu'_2, \\
&\dots \\
m'_k &= \mu'_k,
\end{aligned}
$$

Then we solve the resulting systems of simultaneous equations for $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$

Remarks on Method of Moments Estimators

- Very easy to compute
- Always give an estimator to start with
- Generally consistent (Since sample moments are consistent for population moments)
- Not necessarily the best or most efficient estimators

2.2.1.2 Maximum Likelihood Estimators (MLE)

Example

The number of orders per day coming to a certain company seems to have a Poisson distribution with parameter λ.

The number of orders received during 10 randomly selected days are as follows: $12, 14, 15, 12, 13, 10, 11, 15, 10, 6$

Derive an expression for the $P(X_1 = 12,\ X_2 = 14, \dots, X_{10} = 10)$ as a function of λ.

Find the joint probability of the data

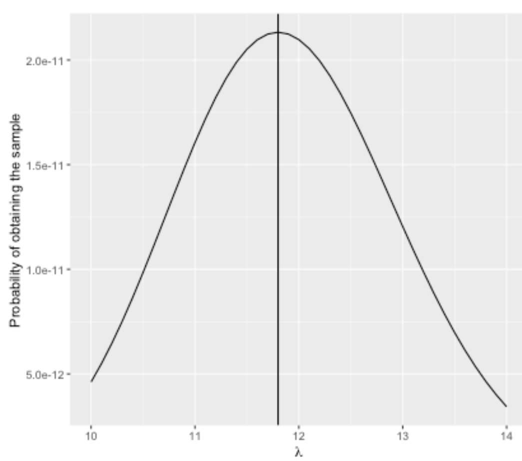


Figure 2.1: Probability of the sample is maximum when $\lambda = 11.8$

- When it is viewed as a function of λ, it is called the **likelihood function of λ for the available data**
- The likelihood for the data is maximum when $\lambda = 11.8$.

- Since these data have already occurred, it is very likely that the data have arisen from a Poisson distribution with $\lambda = 11.8$.
- This estimate for λ is called the **maximum likelihood estimate**
- In order to define maximum-likelihood estimators, we shall first define the likelihood function.

Definition: Likelihood function

Let $x_1, x_2, \dots, x_n$ be a set of observations of random variables $X_1, X_2, \dots, X_n$ with the joint density of n random variables, say $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n;\ \theta)$. This joint density function, which is considered to be a function of θ is called the **likelihood function of θ for the set of observations (sample)** $x_1, x_2, \dots, x_n$.

In particular, if $x_1, x_2, \dots, x_n$ is a random sample from the density $f(x; \theta)$, then the likelihood function is $f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta)$.

Notation

We use the notation $L(\theta;\ x_1, x_2, \dots, x_n)$ for the likelihood function, in order to remind ourselves to think of the likelihood function as a function of θ.

- Likelihood function is seen as a function of θ rather than x
- Likelihood can be viewed as the degree of plausibility.
- An estimate of θ may be obtained by choosing the most plausible value, i.e., where the likelihood function is maximized.

Definition: Maximum Likelihood Estimator

Let $L(\theta) = L(\theta;\ x_1, x_2, \dots, x_n)$ be the likelihood function of θ for the sample $x_1, x_2, \dots, x_n$. Suppose $L(\theta)$ has its maximum when $\theta = \hat{\theta}$.

Then $\hat{\theta}$ is called the **Maximum likelihood estimate of** θ.

The corresponding estimator is called the **Maximum likelihood estimator of** θ.

- Many likelihood functions satisfy regularity conditions; so the maximum likelihood estimator is the solution of the equation

$$\frac{dL(\theta)}{d\theta} = 0$$

Log-likelihood function

Let

$$l(\theta) = ln[L(\theta)].$$

Then, $l(\theta)$ is called the **log-likelihood function**.

- Both $L(\theta)$ and $l(\theta)$ have their maxima at the same value of θ.
- It is sometimes easier tot find the maximum of the logarithm of the likelihood and thereby simplify the calculations in finding the maximum likelihood estimate.

Invariance Property of MLE's

If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

2.2.2 Desirable properties of point estimators

- We discussed several methods of obtaining point estimators.
- It is possible that different methods of finding estimators will lead to same estimator or different estimators.
- In this section we discuss certain properties, which an estimator may or may not posses, that will guide us in deciding whether one estimator is better than another.

2.2.2.1 Unbiasedness

Definition: Unbiased estimator

An estimator $\hat{\theta}$ $(= t(X_1, X_2, \ldots, X_n))$ is defined to be an **unbiased estimator** of θ if and only if

$$E(\hat{\theta}) = \theta$$

- The difference $E(\hat{\theta}) - \theta$ is called as the bias of $\hat{\theta}$ and denoted by

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- An estimator whose bias is equal to 0 is called **unbiased**.

2.2.3 Consistency

Mean-Squared Error

- The *mean-squared error* is a measure of goodness or closeness of an estimator to the target.

Definition: Mean-squared Error (MSE)

The **mean-squared error** of an estimator $\hat{\theta}$ of θ is defined as

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

- The MSE measures the average squared difference between $\hat{\theta}$ and θ.
- The MSE is a function of θ and has the interpretation

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \left[Bias(\hat{\theta})\right]^2$$

- Therefore the MSE incorporates two components, one measuring the variability of the estimator (*precision*) and the other measuring its bias (*accuracy*).
- Small value of MSE implies small combined variance and bias.
- If $\hat{\theta}$ is unbiased, then
$$MSE(\hat{\theta}) = Var(\hat{\theta})$$
- The positive square root of MSE is known as the *root mean squared error*
$$RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})}$$

Consistency

- Estimator $\hat{\theta}$ is said to be consistent for θ if $MSE(\hat{\theta})$ approaches zero as the sample size n approaches ∞.

$$lim_{n\to\infty}E\left[(\hat{\theta}-\theta)^2\right] = 0$$

- Mean-squared error consistency implies that the bias and the variance both approach to zero as n approaches ∞.

2.3 Interval Estimation

- Under point estimation of a parameter θ, the inference is a guess of a **single value** as the value of θ.
- Instead of making the inference of estimating the true value of the parameter to be a point, under interval estimation we make the inference of estimating that the true value of the parameter is contained in **some interval.**

2.3.1 What is gained by using an Interval Estimator?

Example

- For a sample X_1, X_2, X_3, X_4 from a $N(\mu, 1)$, an interval estimator of μ is $[\bar{X} - 1, \bar{X} + 1]$.
- This means that we will assert that μ is in this interval.
- In the previous section (Point estimation) we estimated μ with $\bar{X}$.
- But now we have the less precise estimator $[\bar{X} - 1, \bar{X} + 1]$.
- Under interval estimation, by giving up some precision in our estimate (or assertion about μ), we try to gain some confidence , or assurance that our assertion is correct.

Explanation

- When we estimate μ by $\bar{X}$, the probability that the estimator exactly equaled the value of the parameter being estimated is zero (Why? the probability that a continuous random variable equals any value is 0), *i.e.* $P(\bar{X} = \mu) = 0$.
- However, with an interval estimator, we have a positive probability of being correct.
- The probability that μ is covered by the interval $[\bar{X} - 1, \bar{X} + 1]$ can be calculated as

$$\begin{aligned}
P(\mu \in [\bar{X} - 1, \bar{X} + 1]) &= P(\bar{X} - 1 \leq \mu \leq \bar{X} + 1) \\
&= P(-1 \leq \bar{X} - \mu \leq 1) \\
&= P(-2 \leq \frac{\bar{X} - \mu}{\sqrt{1/4}} \leq 2)
\end{aligned}$$

$$= P(-2 \leq Z \leq 2) \quad \left(\frac{\bar{X} - \mu}{\sqrt{1/4}} \text{ is standard normal} \right)$$
$$= 0.9544.$$

- Therefore now we have over 95% chance of covering the unknown parameter with the interval estimator.
- By moving for a point to an interval we have scarified some precision in our estimate. But it has resulted in increased confidence that our assertion is correct.
- The purpose of using an interval estimator rather than a point estimator is to have some guarantee of capturing the parameter of interest.

2.3.2 Definition of confidence interval

Definition

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ. Let $T_1 = g(X_1, X_2, \ldots, X_n)$, and $T_2 = h(X_1, X_2, \ldots, X_n)$ be two statistics satisfying $T_1 \leq T_2$ for which $P(T_1 < \theta < T_2) = \gamma$, where γ does not depend on θ. Then, the random interval (T_1, T_2) is called a 100γ **percent confidence interval for** θ; γ is called the confidence coefficient; and T_1 T_2 are called the lower and upper confidence limits, respectively, for θ.

Suppose that $x_1, x_2, \ldots, x_n$ is a realization of $X_1, X_2, \ldots, X_n$ and let $t_1 = g(x_1, x_2, \ldots, x_n)$, and $t_2 = h(x_1, x_2, \ldots, x_n)$. Then the *numerical* interval (t_1, t_2) is also called a 100γ **percent confidence interval for** θ.

2.3.3 Interpretation of confidence intervals

- Consider the probability statement $P(\bar{X} - 1.18 \leq \mu \leq \bar{X} + 1.18) = 0.95$.
- The above probability statement implies that the random interval $(\bar{X} - 1.18, \bar{X} + 1.18)$ includes the unknown true mean μ with probability 0.95.

2.3.4 Methods of finding interval estimators

2.3.4.1 Pivotal Quantity Method

Definition: Pivotal Quantity

Let $X_1, X_2, \ldots, X_n$ be a random sample from the density $f(.;\theta)$. Let $Q = q(X_1, X_2, \ldots, X_n; \theta)$; that is, let Q be a function of $X_1, X_2, \ldots, X_n$ and θ. If Q has a distribution that does not depend on θ, then Q is defined to be a *pivotal quantity*

Pivotal Quantity method

If $Q = q(X_1, X_2, \ldots, X_n; \theta)$ is a pivotal quantity and has a probability density function, then for any fixed $0 < \gamma < 1$ there will exist q_1 and q_2 depending on γ such that $P[q_1 < Q < q_2] = \gamma$. Now, if for each possible sample value $(x_1, x_2, \ldots, x_n)$, $q_1 < q(x_1, x_2, \ldots, x_n; \theta) < q_2$ if and only if $t_1(x_1, x_2, \ldots, x_n) < \tau(\theta) < t_2(x_1, x_2, \ldots, x_n)$ for functions t_1 and t_2 (not depending on θ), then (T_1, T_2) is a 100γ percent confidence interval for $\tau(\theta)$, where $T_1 = t_1(X_1, X_2, \ldots, X_n)$ and $T_2 = t_2(X_1, X_2, \ldots, X_n)$.

2.3.5 Methods of evaluating interval estimators

- Coverage probability
- Size (expected length)

Tutorial

1. Let $X_1, X_2, \ldots, X_n \sim iid \quad N(\mu, \sigma^2)$, both μ and σ^2 unknown. Derive a method of moment estimators for μ and σ.

2. Let $X_1, X_2, \ldots, X_n \sim iid \quad Bin(n, \theta)$, both n and θ unknown. Derive a method of moment estimators for n and θ.

3. Let $X_1, X_2, \ldots, X_n \sim iid \quad Unif(\theta_1, \theta_2)$, where $\theta_1 < \theta_2$, both unknown. Derive a method of moment estimators for θ_1 and θ_2.

4. Let $X_1, X_2, \ldots, X_n \sim Poisson(\lambda)$. Derive a method of moment estimators for λ.

5. Let $X_1, X_2, \ldots, X_n \sim iid \quad Gamma(\alpha, \beta)$, both α and β unknown. Derive a method of moment estimators for α and β.

The survival time (in weeks) of 20 randomly selected male mouse exposed to 240 units of certain type of radiation are given below.

$152, 115, 109, 94, 88, 137, 152, 77, 160, 165, 125, 40, 128, 123, 136, 101, 62, 153, 83, 69$

It is believed that the survival times have a gamma distribution. Estimate the corresponding parameters.

6. Let $x_1, x_2, \ldots, x_n$ be n random measurements of random variable X with the density function

$$f_X(x; \lambda) = \lambda x^{\lambda - 1}, \;\; 0 < x < 1, \;\; \lambda > 0$$

   Derive a method of moment estimator for λ.

7. Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a Poisson distribution with parameter λ. Derive the maximum likelihood estimator of λ.

8. Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Derive the maximum likelihood estimators of μ and σ^2.

9. Let $X_1, X_2, \ldots, X_n \sim iid \; Poisson(\lambda)$. Find the MLE of $P(X \leq 1)$

10. Let $X_1, X_2, \ldots, X_n \sim iid \; N(\mu, \sigma^2)$. Find the MLE of μ/σ.

11. Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with the density function

$$f_X(x; \lambda) = \lambda e^{-\lambda x}, \;\; x > 0$$

   . Is the maximum likelihood estimators of λ unbiased?

12. Let $X_1, X_2, \dots, X_n$ be a random sample of size n from a distribution with mean μ and variance σ^2. Show that $\bar{X}$ and $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ are unbiased estimators of μ and σ^2, respectively.

13. Let $x_1, x_2, \dots, x_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Consider the maximum likelihood estimators of σ^2. Show the estimator $\hat{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$ is biased for σ^2, but it has a smaller MSE than $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$.

14. Let $X_1, X_2, \dots, X_n$ be a random sample from some distribution and $E(X) = \mu$. Show that $\bar{X}$ is a better estimator than X_1 and $\frac{X_1 + X_2}{2}$ for μ in terms of MSE.

15. Let $X_1, X_2, \dots, X_n$ be a random sample from a $N(\mu, \sigma^2)$. Let $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ and $T = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$. Show that $\bar{X}$ is consistent for μ and T is consistent for σ^2.

16. Let $X_1, X_2, \dots, X_n$ be a random sample from a $N(\mu, 9)$. Find a 95% confidence interval for μ.

17. Let $X_1, X_2, \dots, X_n$ be a random sample from a $N(\mu, \sigma^2)$, both μ and σ are unknown. Construct a $100(1-\alpha)\%$ $(0 < \alpha < 1)$ confidence interval for μ.

18. Let $X_1, X_2, \dots, X_n$ be a random sample from a $N(\mu, \sigma^2),$. Construct a $100(1-\alpha)\%$ $(0 < \alpha < 1)$ confidence interval for σ^2.

Chapter 3

Hypothesis Testing

- Hypothesis testing about population characteristics is another fundamental aspect od statistical inference.

Definition: Hypothesis

A *hypothesis* is a statement about a **population parameter**.

- In testing hypothesis, we start by making *an assumption* with regard to an unknown population characteristic.
- We then take a random sample from the population, and on the basis of the corresponding sample characteristic, we either accept or reject the hypothesis with a particular degree of confidence.

3.1 Null and alternative hypotheses

- Statistical hypotheses consist of the **null hypothesis** and the **alternative hypothesis,** which together contain all the possible outcomes of the experiment or study.

null hypothesis

- Generally, the *null hypothesis* states that a treatment has *no effect* (and has led to the term "null" hypothesis).

alternative hypothesis

- Usually, alternative hypothesis is the negation, or complement, of the null hypothesis.

- Usually it is the hypothesis that the researcher is interested in proving.
- The null hypothesis is denoted H_0, and the alternative is denoted H_1.
- Let θ is a population parameter.
- the general format of the null hypothesis and alternative hypothesis is

$$H_0 : \theta \in \Theta_o \text{ and } H_1 : \theta \in \Theta_o^c$$

where Θ_o is some subset of the parameter space and Θ_o^c is its complement.

3.2 Different types of Hypotheses

(1) **simple hypotheses:** both H_0 and H_1 consist of only one probability distibution

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta = \theta_1$$

(2) **composite hypothesis:** either H_0 or H_1 has more than one probability distribution

- *one sided*: $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$
- *one sided*: $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$
- *two sided*: $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

3.2.1 Writing H_0 and H_1

1. The equality sign (eq: $=, \leq, or \geq$)) always goes to H_0
2. $H_0 :$ $H_1 :$ which is to be tested (claim to be tested)
3. $H_0 :$ initially favoured one before collecting information $H_1 :$

Example 1

An ideal manufacturing process requires that all products are non-defective. This is very seldom. The goal is to keep the proportion of defective items as low as possible. Let p be the proportion of defective items, and 0.01 be the maximum acceptable proportion of defective items.

Example 2 let θ be the average change in a patient's blood pressure after taking a drug. An experimenter might be interested in testing

3.2.2 Two tailed test and one tailed test

Two tailed test

- Two tailed tests always use $=$ and $\neq$ in the statistical hypotheses.
- Alternative hypothesis allows for either the $>$ or $<$ possibility.
- Here the research is interested in testing deviations from the null in two directions.

One tailed test

- One tailed tests are always directional, and the alternative hypothesis uses either the $>$ or $<$ sign.

3.3 Rejection region

- A hypothesis testing procedure or hypothesis test is a rule that specifies:

  i) for which sample values H_0 is accepted as true
  ii) for which sample values H_0 is rejected and H_1 is accepted as true

- The subset of the sample space for which H_0 will be rejected is denoted as R and called the **rejection** or **critical region**
- The complement set R^c is called the **acceptance** region.
- The rejection region R of a hypothesis test is usually defined through a *test statistic* $W(X)$, a function of the sample.
- For example

$$R = \{X : W(X) > b\}.$$

- If $X \in R$, one rejects H_0
- Otherwise if $X \in R^c$, there is no enough evidence to reject H_0

Figure 3.1: Rejection and nonrejection regions

3.4 Errors in testing hypotheses-type I and type II error

- Samples are used to determine whether to reject H_0 or not.
- Since the decision to reject H_0 or not, is based on incomplete (i.e sample information), there is always a possibility of making an incorrect decision.
- We can make two types errors in testing a hypothesis.
    - *Type I error:* if H_0 is true, but the test incorrectly reject H_0. (reject H_0/ H_0 true)
    - *Type II error:* if H_0 is false, but the test incorrectly accept H_0. (accept H_0/ H_1 true)

	Decision	
Truth	Accept H_0	Reject H_0
H_0	Correct decision	**Type I error** (α)
H_1	**Type II error** (β)	Correct decision (*power*)

Example 3

There is a concern about the perchlorate level found in well water. EPA guidelines suggest that a water supply should have a mean perchlorate level below 4 ppb (parts per billion)

Set up the appropriate hypotheses for this situation by considering the Type I error

3.5 Significance level

- We can control or determine the probability making type I error, α.
- However, by reducing α, we will have to accept a greater probability of making a type II error, β, unless the sample size is increased.
- The probability of committing a type I error is called α or *significance level* (level of significance)
- α equals the area under the curve that is in the rejection region beyond the critical value(s).
- The value of α is always set before the experiment or study is undertaken.
- $1-\alpha$ is called the *level of confidence* of the test.
- Common values of α are 0.05, 0.01, 0.10 and 0.001.

3.6 Power of a test

- The probability of committing a Type II error is β.
- Unlike α, β is not usually stated at the beginning of the hypothesis testing procedure.

Are α and β related?

- α can only be committed when the null hypothesis is rejected and *beta* can only be committed when the null hypothesis is not rejected.

- A researcher cannot commit both a Type I error and Type II error at the same time on the same hypothesis test.

- Generally, α and β are inversely related.

- One way to reduce both errors is to increase the sample size.

- If a larger sample is taken, it is more likely that the sample is representative of the population, which translates into a better chance that a researcher will make the correct choice.

- Statistically, a larger sample will yeild a sample mean closer to the true population mean, thereby indicating better whether the null hypothesis is true or false.

- **Power** of a test is equal to $1-\beta$.

- *i.e.*, the probability of a test rejecting the null hypothesis when the null hypothesis is false.

3.7 Testing hypotheses

- Typically, the hypothesis testing process is presented in terms of an eight-step approach.

Task 1: Hypothesise

Step 1: Establish null and alternative hypotheses.

Task II: Test

Step 2: Determine the appropriate statistical test

Step 3: Set the value of α, the Type I error rate.

Step 4: Establish the decision rule

Step 5: Gather sample data

Step 6: Analyse the data

Task III: Take statistical action

Step 7: Reach a statistical conclusion.

Task IV: Determine the business implications

Step 8: Make a business decision.

- This process of testing hypothesis is referred as the HTAB system, where HTAB is an abbreviation for Hypothesise, Test, Action, Business.

3.8 Methods of testing hypotheses

3.8.1 Using the critical value method to test hypotheses

- A method of testing hypotheses by comparing the sample statistic with the critical value in order to reach a conclusion about rejecting or failing to reject the null hypothesis.
- Critical region: Set of all values of the test statistic that would cause us to reject null hypothesis.

3.8.2 Using the p-value to test hypotheses

- Get the decision by comparing α value with p-value.

- **p-value:** Probability of observing a *sample statistic* at least as extreme as the observed test statistic computed *under the assumption that* H_0 *is true.*

- The p-value defines the smallest value of α for which the null hypothesis can be rejected

  Example

  - If the p-value of a test is 0.038, the null hypothesis cannot be rejected at $\alpha = 0.01$ because 0.038 is the smallest value of alpha for which the null hypothesis can be rejected. However, the null hypothesis can be rejected for $\alpha = 0.05$.
  - If the p-value of a test is 0.0207, , the researcher would reject the null hypothesis for $\alpha = 0.05$ or $\alpha = 0.10$ or any value more than 0.0207.

- Virtually every statistical computer program yields this probability (p-value)

3.8.3 Using confidence intervals to test hypotheses

- Confidence intervals also give some indication about the decision.

Example 4

In order to ensure a good user experience, software companies often thoroughly test their products before releasing them to the public. IMAX Software company conducted a survey several years ago to determine the user-friendliness of one of their products. According to the previous survey the mean customer rating was 4.3 out of 5 (on a scale from 1 to 5, with 1 being low and 5 being high). Suppose a researcher believes that the customer ratings are lower now due to the new additional features of the software product, and he set a new survey in an attempt to prove his claim. Data are gathered and the results are obtained. Use this data to test this claim at 0.05 level of significance. Assume from previous studies that the population standard deviation is 0.574.

$3, 4, 5, 5, 4, 5, 5, 4, 4, 4, 4,$
$4, 4, 4, 4, 5, 4, 4, 4, 3, 4, 4,$
$4, 3, 5, 4, 4, 5, 4, 4, 4, 5$

contd…

1. Hypotheses testing for single population

    - 1.1 Testing hypotheses about a population mean using the z statistic (σ known)
    - 1.2 Testing hypotheses about a population mean using the t statistic (σ unknown)
    - 1.3 Testing hypotheses about a proportion
    - 1.4 Testing hypotheses about a variance

2. Test hypotheses about two populations

    - 2.1 Hypothesis testing for the difference in two means using the z statistic (Population variances known)
    - 2.2 Hypothesis testing for the difference in two means using the t statistic (Population variances unknown)
    - 2.3 Hypothesis testing for the difference in two proportions using the z statistic

Test For	Null Hypothesis (H_0)	Test Statistic	Distribution	Use When
Population mean (μ)	$\mu = \mu_0$	$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	Z	Normal distribution or $n \geq 30$; σ known
Population mean (μ)	$\mu = \mu_0$	$\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	t_{n-1}	Normal distribution and σ unknown
Population proportion (p)	$p = p_0$	$\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Z	$n \geq 20$, $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$
Population variance (σ)	$\sigma = \sigma_0$	$\frac{(n-1)s^2}{\sigma_0^2}$	χ^2_{n-1}	Normal distribution
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	Z	Both normal distributions, or $n_1, n_2 \geq 30$; σ_1, σ_2 known
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$t_{n_1+n_2-2}$	σ_1, σ_2 unknown but assume $\sigma_1 = \sigma_2$, samples are independent, Both normal distributions
Difference of two proportions ($p_1 - p_2$)	$p_1 - p_2 = 0$	$\frac{(\hat{p}_1-\hat{p}_2)-0}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ where $\bar{p} = \frac{x_1+x_2}{n_1+n_2} = \frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$ and $\bar{q} = 1-\bar{p}$	Z	$n\hat{p}, n(1-\hat{p}) \geq 5$ for each group

Cumulative Standard Normal Distribution

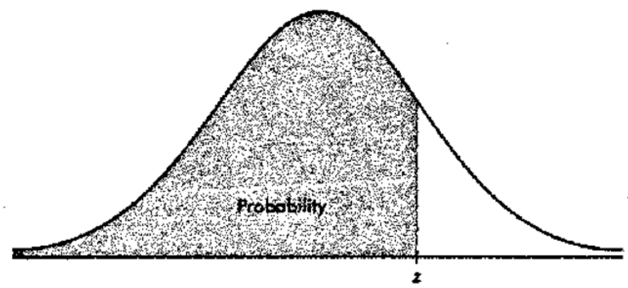


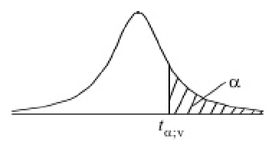
TABLE A: STANDARD NORMAL PROBABILITIES (CONTINUED)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Critical values from the t distribution

Table of the Student's *t*-distribution

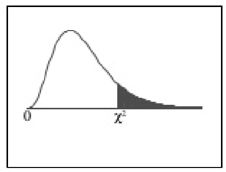
The table gives the values of $t_{\alpha;\nu}$ where
$\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom



ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Critical values from the chi-square distribution

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Tutorial

1. A survey of technical consultants of a computer product s company found that the average net income for a technical consultant is \$74914. Because this survey is now more than 10 years old, an accounting researcher wants to test this figure by taking a random sample of 112 technical consultants to determine whether the net income figure changed. Assume the population standard deviation of net incomes for technical consultants is \$14530. Suppose the 112 technical consultants who respond produce a sample mean of \$78695.

2.

a. Use the data given to test the following hypotheses.

$$H_0 : \mu = 25 \qquad H_1 : \mu \neq 25$$

$$\bar{x} = 28.1 \qquad n = 57 \qquad \sigma = 8.46 \qquad \alpha = 0.01$$

3.

a. Use the data given to test the following hypotheses.

$$H_0 : \mu = 1200 \qquad H_1 : \mu > 1200$$

$$\bar{x} = 1215 \qquad n = 113 \qquad \sigma = 100 \qquad \alpha = 0.10$$

b. Use the $p-$ value to reach a statistical conclusion.

4. A company build fully automated harvesters. For a harvester to be properly balanced when operating , a 25-kg plate is installed on its side. The machine that produces these plates is set to yield plates that average 25-kg. The distribution of plates produced from the machine is normal. However, the supervisor is worried that the machine us out of adjustment and is producing plates that do not average 25 kg. To test this concern, he randomly selects 20 of the plates produced the day before and weighs them. The following data shows the computed sample mean and sample standard deviation

$$\bar{x} = 25.51 \qquad n = 20 \qquad s = 2.1933 \qquad \alpha = 0.05$$

5. Suppose a study reports that the average price for a litre of self -serve regular unleaded petrol is \$1.16. You believe that the figure is higher in your area of the country. You decide to test this claim for your area by randomly calling petrol stations. Your random survey of 25 stations produces the following prices.

$\$1.27, 1.29, 1.16, 1.20, 1.37$
$1.20, 1.23, 1.19, 1.20, 1.24$
$1.16, 1.07, 1.27, 1.09, 1.35$
$1.15, 1.23, 1.14, 1.05, 1.35$
$1.21, 1.14, 1.14, 1.07, 1.10.$

Assume petrol prices for a region are normally distributed. Do the data you obtained provide enough evidence to reject the claim?. Use a 1% level of significance.

6. A manufacturing company believes exactly 8% of its products contain at least one minor flaw. Suppose a company researcher want to test this belief. He randomly selects a sample of 200 products, inspects each item for flaws and determines that 33 items have at least one minor flaw. Use a 0.10 level of significance to test this hypothesis.

7. A survey of the online teaching shows that the primary mode of delivery for 17% of school teachers is google classroom. A researcher believes the figure is higher for western province. To test this idea, he contacts a random sample of 550 teachers asks which primary mode they used to online teaching during the least six months. Suppose 115 replied that google classroom was the primary mode of delivery. Using a level of significance of 0.05, tests idea that the figure is higher for western province.

8. A manufacturing company has been working to implement a just-in-time inventory system for its production line. The final product requires the installation of a pneumatic tube at a particular station on the assembly line. With the just-in-time inventory system, th company's goal is to minimize the number of pneumatic tubes that are piled up at the station waiting to be installed. Ideally, the tubes would arrive just as the operator needs them. However, because of the supplier and the variables involved in getting the tubes to the line, most of the time there will be some build-up of tube inventory. The company expects that, on average, about 20 pneumatic tubes will be at the station. However, the production superintendent does not want the variance of this inventory to be grater than 4. On a given day, the number of pneumatic tubes piled up at the workstation is determined eight different times and the following numbers of tubes are recorded.

$$23, 17, 20, 29, 21, 14, 19, 24$$

a. Use these sample data and $\alpha = 0.05$ to determine whether the variance is greater than 4.

b. Write the assumptions that you make and motivate your answer.

9. A small business business has 37 employees. Because of the uncertain demand for its products, the company usually pays overtime on any given week. The company assumed that about 50 total hours of overtime per week us required and that the variance on this figure is about 25. Company officials want to know whether the variance of overtime hours has changed. Given here is a sample of 16 weeks of overtime data (in hours per week). Assume hours of overtime are normally distributed. Use these data to test the null hypothesis that the variance of overtime data is 25. Let $\alpha = 0.10$

$57, 56, 52, 44$
$46, 53, 44, 44$
$48, 51, 55, 48$
$63, 53, 51, 50$

10. Suppose a researcher want to conduct a hypothesis test to determine whether the average annual wage of an advertising manager is different from the average annual wage of an auditing manager. A random sample of 32 advertising managers and a random sample of 34 auditing manager are taken.They are contacted by telephone and asked their annual salary. The resulting salary data are given below. Suppose $\alpha = 0.05$.

Advertising managers	Auditing managers
$n_1 = 32$	$n_2 = 34$
$\bar{x}_1 = 70.700$	$\bar{x}_2 = 62.187$
$\sigma_1 = 16.253$	$\sigma_2 = 12.900$

11. In a production company new employees are expected to attend a three-day seminar to learn about the company. At the end of the seminar , they are tested to measure their knowledge about the company. The traditional training method has been lecture and a QA session. Management decided to experiment with a different training procedure, which processes new employees in two days by using pre-recorded online lectures and having no QA session. If this procedure works, it could save the company thousands of dollars over a period of several years. However, there is some concern

about the effectiveness of the two-day method, and company managers would like to know whether there is any difference in the effectiveness of two training methods.

To test the difference in the two methods, the managers randomly select one group of 15 newly hired employees to take the three-day seminar (method A) and a second group of 12 new employees for the two-day prerecorded online method (method B). The test scores are given below.

Use $\alpha = 0.05$, determine whether there is a significant difference in the mean scores of the two groups. Assume that the scores for this test are normally distributed and that the population variances are approximately equal.

Method A : $56, 50, 52, 44, 52, 47, 47, 53, 45, 48, 42, 51, 42, 43, 44$

Method B : $59, 54, 55, 65, 52, 57, 64, 53, 53, 56, 53, 57$

12. A study of young entrepreneurs was conducted to determine their definition of success. The entrepreneurs were offered optional choices such as happiness/self-fulfillment, sales/profit and achievement/challenge. The entrepreneurs were divided into groups according to the gross sales of their businesses. A significantly higher portions of entrepreneurs in the \$100000 to \$500000 category than in the less than \$100000 category seemed to rate sales/profit a a definition of success.

Suppose you decide to test this result by taking a survey of your own and identify entrepreneurs by gross sales. You interview 100 young entrepreneurs with gross sales of less than \$100000, and 24 of then define sales/profit as success. You then interview 95 young entrepreneurs with gross sales of \$100000 to \$500000, and 39 cite sales/profit as a definition of success. Use this information to test to determine whether there is a significant difference in the proportions of the two groups that define success as sales/profit. Use $\alpha = 0.01$.

Chapter 4

Contingency Tables

- In Chapters 2 (Estimations) and 3 (Hypothesis Testing) we briefly discussed about the chi-square distribution in the construction of confidence intervals for and the testing of hypotheses concerning a population variance.
- Chi-square distribution, which is one of the most widely used distributions in statistical applications, has many other uses.
- The chi-square distribution is the most frequently employed statistical technique for the analysis of categorical data.

4.1 Types of chi-square tests

Chi square test has two cases:

1) Chi-square goodness of fit test
2) Chi-square test of independence. This is also known as *contingency analysis.*

- In this chapter our focus is on *chi-square test of independence*

4.2 Contingency analysis: Chi-square test of independence

- **Chi-square test of independence** is a statistical test used to analyse the frequencies of two variables with multiple categories to determine whether the two variables are independent.

- For example, a software developer might want to determine whether the user ranking of user-friendliness (*Very Satisfied = 5, Satisfied = 4, Neutral = 3, Dissatisfied = 2, Very Dissatisfied = 1*) of the new software product is independent of the user's education level.

- A market researcher might want to determine whether the type of online communication platforms (eg:Slack, Zula, ezTalks, HipChat, Yammar) preferred by a user is independent of the user's age.
- A software project manager might want to know whether customer overall satisfaction (*Very Satisfied = 5, Satisfied = 4, Neutral = 3, Dissatisfied = 2, Very Dissatisfied = 1* is independent of the experience of the developer (less than 1 year, 1-5 years, more than 5 years)

Example

A software company is planning to develop a new taxi booking and trip scheduling platform that facilitates a real time connection between the commuter and the taxi driver, enabling mutual engagement for the receipt and delivery of a service. Prior to development, the management wants to identify the strength and weakness of the existing services as they may be their potential competitors. Suppose the project manager is interested in determining whether the type of taxi service preferred by a commuter is independent of the commuter's age.

On a questionnaire, the following two questions might be used to measure these aspects.

1. What is your most preferred taxi booking and trip scheduling platform?

a) Uber
b) PickMe

2. Select your age category?

a) < 30 years
b) $30 - 50$ years
c) > 50 years

4.2.1 Contingency table

- Let the number of responses received were 100.
- Now the analyst can *tally the frequencies of responses* to these two questions into a two-way table called **contingency table**

Contingency table

Most preferred taxi booking and trip scheduling platform			
Age	Uber	PickMe	Total
< 30 years	32	25	57
$30 - 50$ years	18	5	23
> 50 years	17	3	20
Total	67	33	100

- Because the chi-square test of independence uses a contingency table, this test is sometimes referred to as **contingency analysis**

Contingency table

	Variable 2			
Variable 1	F	G	H	Total
A			o_{13}	n_A
B				n_B
C				n_C
D				n_D
E				n_E
Total	n_F	n_G	n_H	

- The **observed frequency** for for each cell is denoted as o_{ij}, where i is the row and j is the column.
- The **expected frequencies** are denoted in a similar manner.
- If two variables are independent, they are not related.
- In a sense, the chi-square test of independence is a test of whether the variables are related.
- The null hypothesis for a chi-square test of independence is that the two variables are independent (not related).
- If the null hypothesis is rejected, the conclusion is that the two variables are not independent and are related.
- In general, if the two variables are independent, the expected frequency values of each call can be determined by

$$e_{ij} = \frac{n_i n_j}{N}$$

where:

- $i =$ the row
- $j =$ the column
- $n_i =$ the total of row i
- $n_j =$ the total of column j
- $N =$ the total of all frequencies

- Using these expected frequency values and the observed frequency value, we can compute a chi-square test of independence to determine whether the variables are independent.

4.2.2 Chi-square test of independence

$$\chi^2 = \sum_r \sum_c \frac{(f_0 - f_e)^2}{f_e}$$

where - $df = (r-1)(c-1)$ - $r =$ number of rows - $c =$ number few columns - $\sum_r =$ sum over row - $\sum_c =$ sum over column

- The **null hypothesis** for a chi-square test of independence is that *the two variables are independent.*
- The **alternative hypothesis** for a chi-square test of independence is that *the two variables are not independent*
- This test is one one-tailed.
- The degrees of freedom are $(r-1)(c-1)$

Caution

- In chi-square test of independence, small expected frequencies can lead to inordinately large chi-square values.
- Therefore, contingency tables should not be used with expected cell values of less than 5.
- One way to avoid small expected values is to combine columns or rows whenever possible and whenever doing so make sense.

Example contd…

Suppose the project manager is interested in determining whether the type of taxi service preferred by a commuter is independent of the commuter's age. He conducted a small survey to determine. He received 100 responses. The resulting 3×2 contingency table is given below. Using $\alpha = 0.05$, perform the chi-square test of independence to determine whether preferred mode is independent of age.

	Most preferred taxi booking and trip scheduling platform		
Age	Uber	PickMe	Total
< 30 years	32	25	57
$30 - 50$ years	18	5	23
> 50 years	17	3	20
Total	67	33	100

4.3 Chi-square test of independence with R

- Let us now see how to perform the Chi-square test of independence using R programming language

- Following is the observed table.

```
observed_table <- matrix(c(32, 25, 18, 5, 17, 3), nrow = 3, ncol = 2, byrow = T)
rownames(observed_table) <- c('<30', '30-50', '>50')
colnames(observed_table) <- c('Uber', 'PickMe')
observed_table
```

```
##       Uber PickMe
## <30     32     25
## 30-50   18      5
## >50     17      3
```

```
addmargins(observed_table)
```

```
##       Uber PickMe Sum
## <30     32     25  57
## 30-50   18      5  23
## >50     17      3  20
## Sum     67     33 100
```

- In order to perform the test, we need to apply the `chisq.test()` function to the observed table.

```
output <- chisq.test(observed_table)
output
```

```
##
##  Pearson's Chi-squared test
##
## data:  observed_table
## X-squared = 7.2902, df = 2, p-value = 0.02612
```

- From the above result, we can see that p-value is less than the significance level (0.05). Therefore, we can reject the null hypothesis and conclude that the two variables (age & taxi preference) are not independent

- If we want to see the expected table, we can also do that.

```
output$expected
```

```
##        Uber PickMe
## <30   38.19  18.81
## 30-50 15.41   7.59
## >50   13.40   6.60
```

- Visualizing data of Contingency Table with ggplot2 (Side By Side Bar Graph)

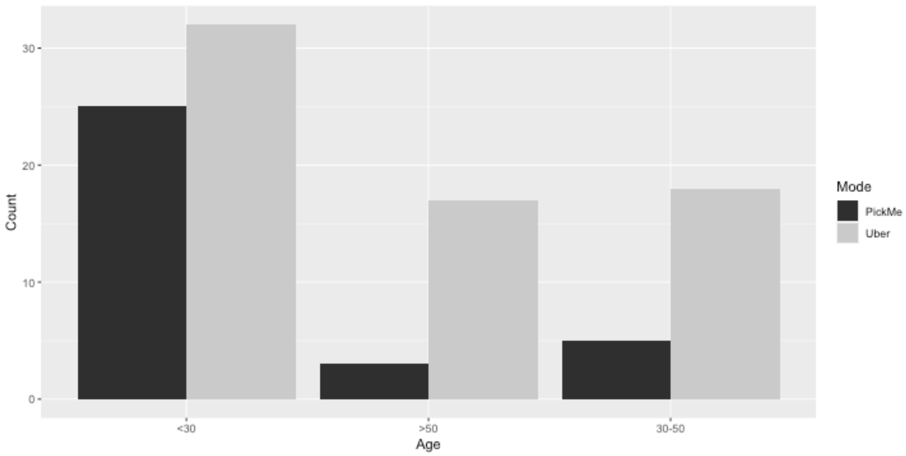
```
library(tidyverse)

table2 <- observed_table %>%
  as.data.frame() %>%
  rownames_to_column() %>%
  gather(Column, Value, -rowname)

colnames(table2) <- c('Age', 'Mode', "Count")
table2
```

```
##     Age   Mode Count
## 1   <30   Uber    32
## 2 30-50   Uber    18
## 3   >50   Uber    17
## 4   <30 PickMe    25
## 5 30-50 PickMe     5
## 6   >50 PickMe     3
```

```
ggplot(data = table2, aes(x = Age, y = Count, fill = Mode)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_grey()
```

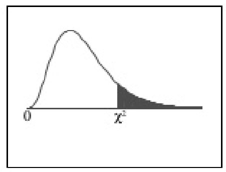


References

- Black, K., Asafu-Adjaye, J., Khan, N., Perera, N., Edwards, P., & Harris, M. (2007). *Australasian business statistics.* John Wiley & Sons.
- Daniel, W. W., & Cross, C. L. (2018). Biostatistics: a foundation for analysis in the health sciences. Wiley.

Critical values from the chi-square distribution

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Tutorial

1. A company maintain an online food ordering service. Suppose market researcher wants to know whether the type of beverage ordered with lunch is independent of the food delivery zone (area). A random sample of 309 lunch customers is selected, resulting in the following contingency table of observed values. Use $\alpha = 0.01$ to determine whether the two variables are independent.

	Preferred beverage		
Zone	Fresh Fruit Juice	Soft drink	Other
Zone 1	26	95	18
Zone 2	41	40	20
Zone 3	24	13	32

2. A computer manufacturer is considering sourcing supplies of an electrical component from four different manufacturers. The director of purchasing asked for samples of 95 from each manufacturer. The numbers of acceptable and unacceptable components from each manufacturer are shown in the following table. Determine whether the quality of the component depends on the type of manufacturer. Let $\alpha = 0.05$

	Manufacturer				
Quality	A	B	C	D	Total
Unacceptable	12	8	5	11	36
Acceptable	83	87	90	84	344
Total	95	95	95	95	380

3. An apparel & clothing market research company conducted a survey in 2019 on social media usage of university students. The frequencies for most frequently used social media platform by male and female students (in thousands) are presented in the following contingency table. Using $\alpha = 0.01$, test the hypothesis that social media preference is independent of gender.

	Social media platform			
Gender	Facebook	Instagram	Twitter	Total
Male	687	327	82	1096
Female	658	304	101	1063
Total	1345	631	183	2159

4. Use R to make a decision for Question 1
5. Use R to make a decision for Question 2
6. Use R to make a decision for Question 3

Chapter 5

Design of Experiments

5.1 Introduction to experimental design

- An experimental design is a plan and a structure to test hypothesis in which the researcher either controls or manipulates one or more variables.
- It contains independent and dependent variables.
- In an experimental design, an independent variable may e either a treatment variable or a classification variable.
- A *treatment variable* is a variable the experimenter controls or modifies in the experiment.
- A *classification variable* is some characteristic of the experimental subjects that was present prior to the experiment and is not a result of the experimenter's manipulations or control.
- Independent variables are sometimes referred to as *factors*.
- Each *independent variable* has two or more *levels*, or classifications.
- *Levels*, or classifications, of independent variables are the subcategories of the independent variables used by the researcher in the experimental design.
- The other type of variable in an experimental design is a *dependent variable*.
- A dependent variable is the repose to the different levels of the independent variables.
- It is the measurement taken under the conditions of the experimental design that reflect the effects of the independent variables(s).

5.2 Analysis of Variance (ANOVA)

- Experimental design in this chapter are analysed statistically by a group of techniques referred to as **analysis of variance** or (**ANOVA**)
- The analysis of variance concepts begin with the notion that individual items being studied, such as employees, machine-produced products, district offices, hospitals ad so on, **are not all the same**.

5.3 The completely randomized design (one-way ANOVA)

- One of the simplest experimental designs is the completely randomized design.
- In the *completely randomized design*, subjects are assigned randomly to treatments.
- The complete randomized design contains only one independent variable, with two or more treatment levels, or classifications.
- If only two treatment levels, or classifications, of the independent variable are present, the design is the same one used to test the difference in means of two independent populations presented in Chapter 3.
- In this section we will focus on completely randomized designs with three or more classification levels.
- Analysis of variance, or ANOVA will be used to analyse the data that result from the treatments.

Example

As an example of a completely randomized design, suppose a researcher decided to analyse the response time of four autonomous mobile robots.

1. What is the independent variable in this design? Autonomous mobile robot

Now the four autonomous mobile robots are the levels of treatment, or classification of the independent variable

2. What is the dependent variable in this design? Response time

3. Is there a **significant** difference in the mean response time of 24 tasks carried out by the four autonomous mobile robots?

```
library(tidyverse)

data <- data.frame(
  Robot = as.factor(c(rep(1,5), rep(2,8), rep(3,7), rep(4,4) )),
  Reponse_time = c(6.33, 6.26, 6.31, 6.29, 6.40, 6.26,
                   6.36, 6.23, 6.27, 6.19, 6.50, 6.19,
                   6.22, 6.44, 6.38, 6.58, 6.54, 6.56,
                   6.34, 6.58, 6.29, 6.23, 6.19, 6.21 ) )

data
```

```
##    Robot Reponse_time
## 1      1         6.33
## 2      1         6.26
## 3      1         6.31
## 4      1         6.29
## 5      1         6.40
## 6      2         6.26
## 7      2         6.36
## 8      2         6.23
## 9      2         6.27
## 10     2         6.19
## 11     2         6.50
## 12     2         6.19
## 13     2         6.22
## 14     3         6.44
## 15     3         6.38
## 16     3         6.58
## 17     3         6.54
## 18     3         6.56
## 19     3         6.34
## 20     3         6.58
## 21     4         6.29
## 22     4         6.23
## 23     4         6.19
## 24     4         6.21
```

```
data %>%
    group_by(Robot) %>%
    summarise(
        count_poison = n())
```

```
## # A tibble: 4 x 2
##   Robot count_poison
##   <fct>        <int>
## 1 1                5
## 2 2                8
## 3 3                7
## 4 4                4
```

4. Compute the mean and standard deviation for each level of the treatment

```
# You can check the level of the robots with the
# following code. You should see three character
# values because we converted it to factor variable

levels(data$Robot)
```

```
## [1] "1" "2" "3" "4"
```

```
data %>%
    group_by(Robot) %>%
    summarise(
        count_robots = n(),
        mean_time = mean(Reponse_time, na.rm = TRUE),
        sd_time = sd(Reponse_time, na.rm = TRUE)
    )
```

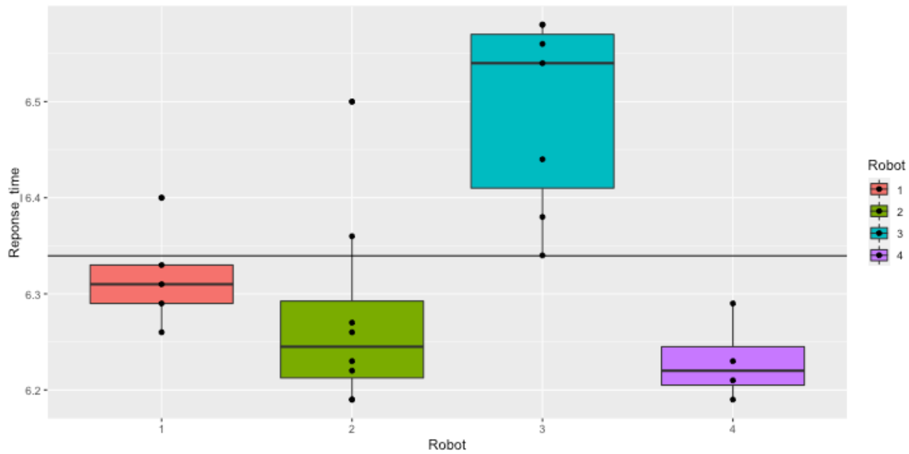
```
## # A tibble: 4 x 4
##   Robot count_robots mean_time sd_time
##   <fct>        <int>     <dbl>   <dbl>
## 1 1                5      6.32  0.0526
## 2 2                8      6.28  0.105
## 3 3                7      6.49  0.101
## 4 4                4      6.23  0.0432
```

5. Graphically check if there is a difference between the distribution.

```
p <- ggplot(data, aes(x = Robot , y = Reponse_time, fill = Robot)) +
    geom_boxplot() +
    geom_point()

overall_mean <-  mean(data$Reponse_time)

p + geom_hline(yintercept = overall_mean)
```



6. Use one-way ANOVA test to determine whether there is a significant difference in the mean response time of the four autonomous mobile robots .

- The basic syntax for an ANOVA test is

  aov(formula, data) Arguments:
  - formula: The equation you want to estimate - data: The dataset used

- The syntax of the formula is:

```
y ~ X1+ X2+...+Xn
```

- `X1+ X2+...+Xn` refers to the independent variables

a) Write the hypothesis to be tested

b) Use the p-value to reach a statistical conclusion

```r
# Run the ANOVA test
anova_one_way <- aov(Reponse_time~Robot, data = data)

# Print the summary of the test
summary(anova_one_way)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Robot        3 0.2366 0.07886   10.18 0.000279 ***
## Residuals   20 0.1549 0.00775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The p-value is lower than the usual threshold of 0.05.

- You are confident to say there is a statistical difference between the groups, indicated by the ∗.

5.4 Pairwise comparison

- The one-way ANOVA test does not inform which robot (group) has a different mean.

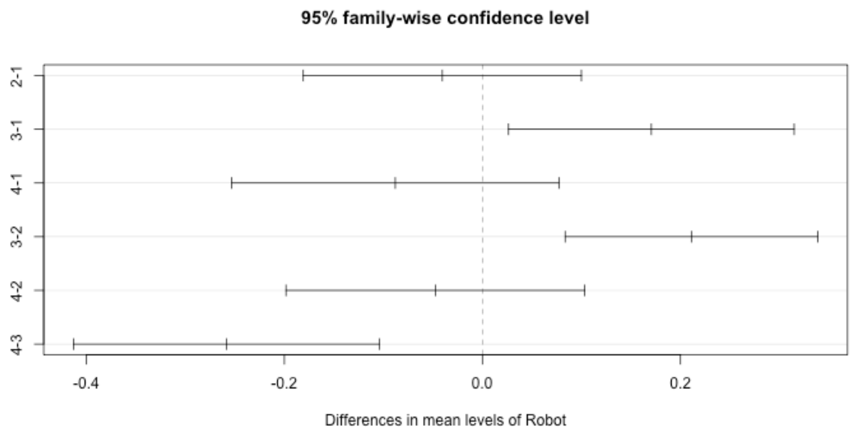
- Instead, we can perform a Tukey test with the function `TukeyHSD()`.

```
comparison <- TukeyHSD(anova_one_way)

#There are print and plot methods for class "TukeyHSD"
print(comparison)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Reponse_time ~ Robot, data = data)
##
## $Robot
##           diff         lwr         upr     p adj
## 2-1 -0.0405000 -0.18093243  0.09993243 0.8502854
## 3-1  0.1705714  0.02633255  0.31481031 0.0169205
## 4-1 -0.0880000 -0.25324639  0.07724639 0.4613461
## 3-2  0.2110714  0.08358107  0.33856179 0.0008519
## 4-2 -0.0475000 -0.19834863  0.10334863 0.8144408
## 4-3 -0.2585714 -0.41296992 -0.10417294 0.0007541
```

```
plot(comparison)
```



Multiple comparison

- Multiple comparisons are to be used only when the analysis of variance yields an overall significant difference in the treatment means.

Tukey's test

- In analysis of variance, this technique is used for pairwise a posterior multiple comparisons to determine if there is a significant differences between the means of any pair of treatment levels in an experimental design.

References

Black, K., Asafu-Adjaye, J., Khan, N., Perera, N., Edwards, P., & Harris, M. (2007). *Australasian business statistics.* John Wiley & Sons.

Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2, pp. 337-472).Pacific Grove, CA: Duxbury

Mood, A.M., Graybill, F.A. and Boes, D.C. (2007): Introduction to the Theory of Statistics, 3rd Edn. (Reprint). Tata McGraw-Hill Pub. Co. Ltd.