

Applied Statistics

Agenda – Inferential Statistics I

1. Inferential Statistics
2. Some fundamental terms first
 - a. Random Variables
 - b. Distribution and its types
3. Binomial Distribution
4. Uniform Distribution [additional content]
5. Normal Distribution

Inferential Statistics

Descriptive vs. Inferential Statistics

Summaries from data

Summaries give a sense of central tendency, variation, association

Tell a lot about 'what's happening'

Mean, standard deviation, correlation, etc.

Is that enough?

Typically, we work with only 'samples' of data

It is not enough to make generalized statement about the population (what we are actually interested in)

Challenge - how to learn about population from the sample?

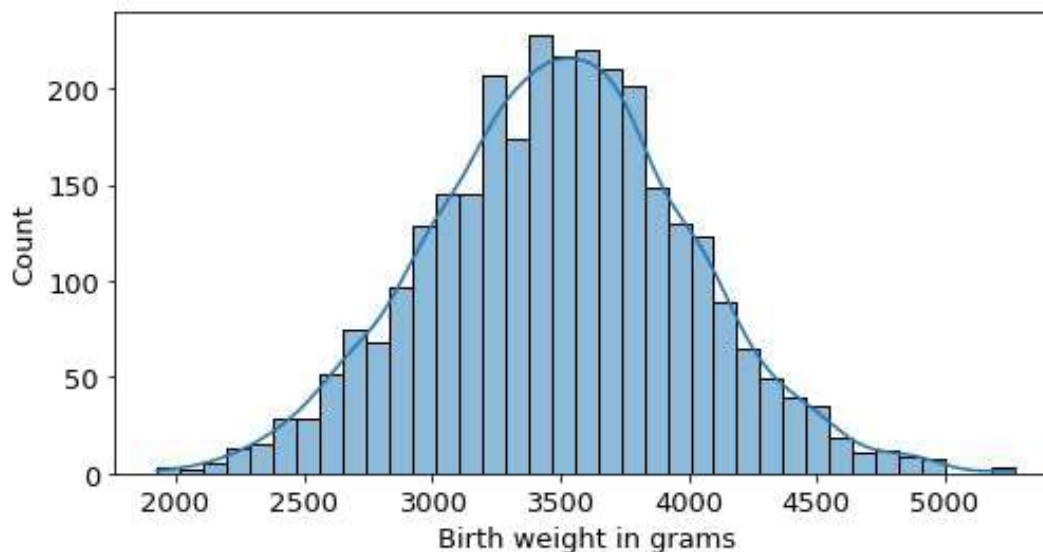
Inferential Statistics

Inferential statistics helps tackle that challenge

There are powerful methods to draw reasonable conclusions about the population from an observed sample

This becomes extremely critical in business decision making

Role of distributions in inferential statistics



What is the chance that birth weight is less than 3000 grams?

Descriptive answer: Count the number of births below 3000 grams from the histogram.

Inferential answer: Consider the underlying distribution and calculate the proportion (area) less than 3000 grams.

Business Problems

Quality Testing

Is the new manufacturing process better/more reliable than the old process?

Meteorology

How likely is it that temp will be more than 20 degree C on a specific day?

Human Resources

Does training the workforce improve sales?

Digital Marketing

What is the chance that the conversion rate on the website will be above x% next month?


...

...


Some fundamental terms

Random Variable

Suppose there are 1000 students in the university. What is the probability that 500 students will pass the upcoming exam?



There is a 50-50 chance that each student will pass or fail



The total number of students who pass can range from 0 to 1000



A random variable assigns a numerical value to each outcome of an experiment. It assumes different values with different probability.

Discrete Random Variable

You work for an Auto insurance company. Suppose the number of insurance claims filed by a driver in a month is a random variable (X) described by

$$X = \begin{cases} 0, & \text{with prob } 0.95 \\ 1, & \text{with prob } 0.04 \\ 2, & \text{with prob } 0.008 \\ 3, & \text{with prob } 0.002 \end{cases}$$

All probabilities must be non-negative and sum to 1.



When all possible values the random variable can take can be listed, we call it a **discrete random variable**

Continuous Random Variable

Suppose the volume of soda in a bottle is described by a random variable.

Can we list all possible values?

498 mL, 499 mL, 500 mL, What about 499.2129415 mL?

Sometimes it is just not possible to list all values a random variable can take

If the random variable can take any value in a given range, we call it a **continuous random variable**

Probability Distribution

Probability Distribution

Describes the values that a random variable can take, along with the probabilities of those values



Discrete Probability Distribution

Arises from discrete random variables

Has an associated **probability mass function**, which gives the probability with which the random variable takes a particular value



Continuous Probability Distribution

Arises from continuous random variables

Has an associated **probability density function**, which helps determine the probability with which the random variable lies between two given numbers

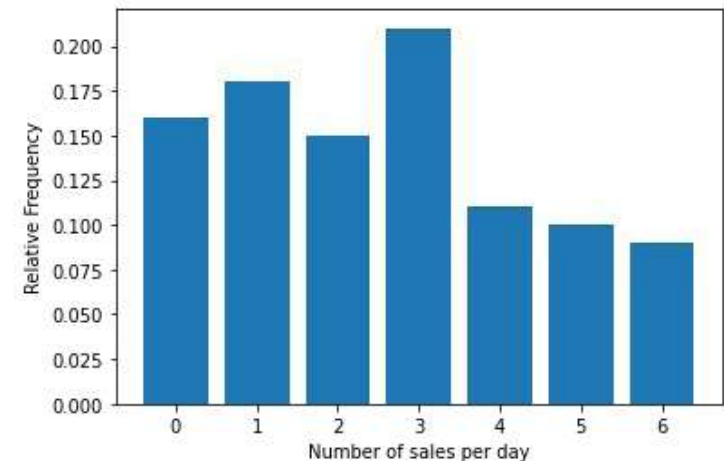
Probability Distribution: Example

A company tracks the number of sales new employees make each day during a 100-day probationary period. The results for one new employee are shown. Construct and plot a probability distribution.

Sales	#Days
0	16
1	18
2	15
3	21
4	11
5	10
6	9



Sales	#Days	Relative Frequency
0	16	0.16
1	18	0.18
2	15	0.15
3	21	0.21
4	11	0.11
5	10	0.10
6	9	0.09



Distributions around us (commonly occurring)

Bernoulli

Company has introduced a new drug to cure a disease, it either cures the disease (it's successful) or it doesn't (it's a failure)

Binomial

The number of defective products in a batch production run

Uniform

The number of microwave ovens sold daily at a busy consumer goods store

Normal

Income distribution of a country on a logarithmic scale

Basic distributions - Binomial

Bernoulli Distribution

Success and failure are non-judgemental. Any one outcome may be termed as success

It has only **two possible outcomes**, namely 1 (success) and 0 (failure), of **one single trial**.

$$X = \begin{cases} 1, & \text{with prob } p \\ 0, & \text{with prob } 1-p \end{cases}$$


Very useful in many scenarios:

Manufacturing defective parts

Outcome of medical test

Binomial Distribution

Suppose we ask any adult who uses the app TikTok if he/she has ever posted a video on the app



The answer can be Yes or No (success or failure)



We can use the Bernoulli distribution to model this scenario



Now let us extend this into a survey of 25 adults chosen at random



We can define **a random variable X which counts the number of successes**
(say, the number of adults who responded Yes)

Binomial Distribution

In many situations an experiment may have only two outcomes - success and failure

These experiments can be modelled using the Binomial probability distribution.

Bernoulli Distribution is a special case of Binomial Distribution with a single trial.

Probability Mass Function

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

Binomial Distribution : Assumptions

Number of trials (n) is fixed.

Each trial is independent of the other trials.

There are only two possible outcomes (success or failure) for each trial.

The probability of a success (p) is the same for each trial.

What happens if these assumptions are violated?

In a month of 30 days, what is the probability that it will rain on more than 10 days, if on average the chance of rain on a given day is 20%?

If we assume that:

1. The event of rain on a particular day is independent of it raining on the previous day.
2. The chance of rain does not increase or decrease over the duration of the month.

Then we can use the binomial distribution with $n=30$ and $p=0.2$ to calculate the probability.

Assumptions 1 and 2 in the example are not strictly valid, but they allow for a direct calculation that may be good enough for practical purposes.

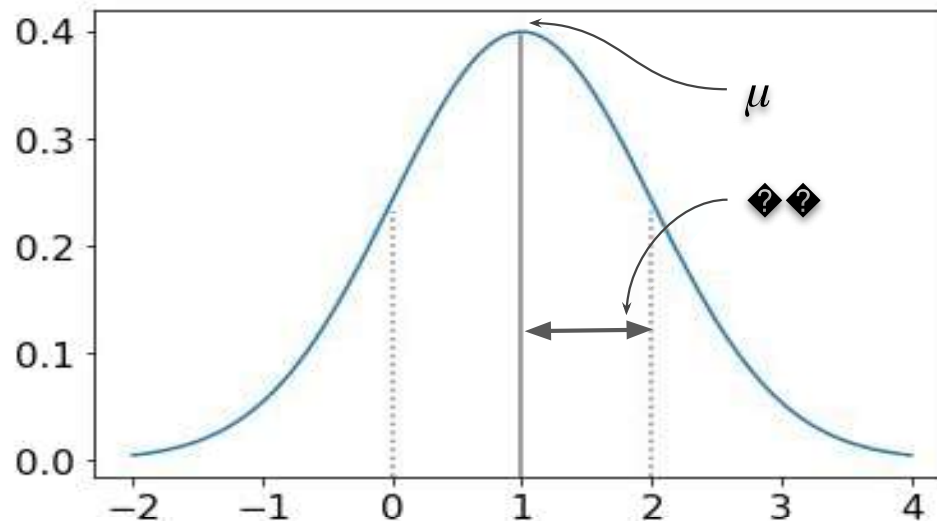
Basic distributions - Normal

Normal Distribution : Introduction

Normal distribution is the most common and useful continuous distribution



It is characterized by a **symmetric bell-shaped** curve having two parameters - mean (μ) and standard deviation (σ).



Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Normal Distribution : Why Normal

Why is it called the normal distribution?



They are commonly found everywhere starting from nature to industry



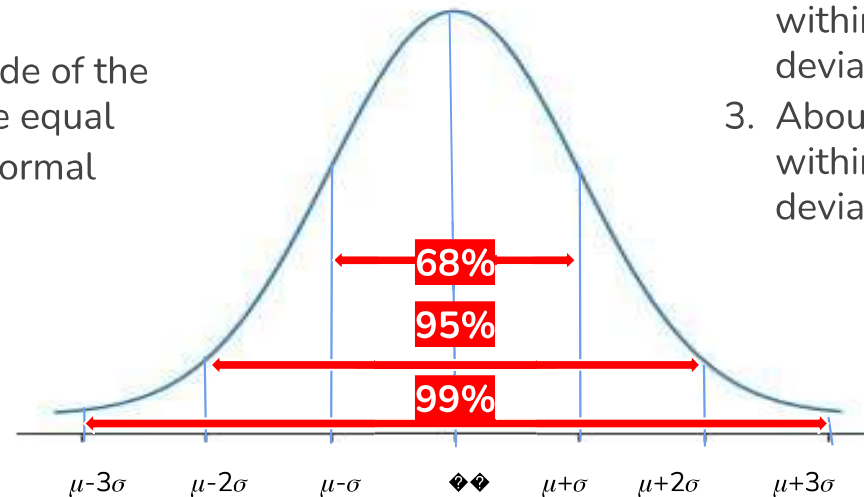
Many useful datasets are approximately normally distributed



For example the height and weight of the adults, IQ scores, measurement errors, quality control test results etc.

Normal Distribution : Properties

1. The graph of the normal distribution is called the normal curve
2. Normal curve is symmetric around the mean
3. Mean, Median and Mode of the normal distribution are equal
4. Total area under the normal curve is 1



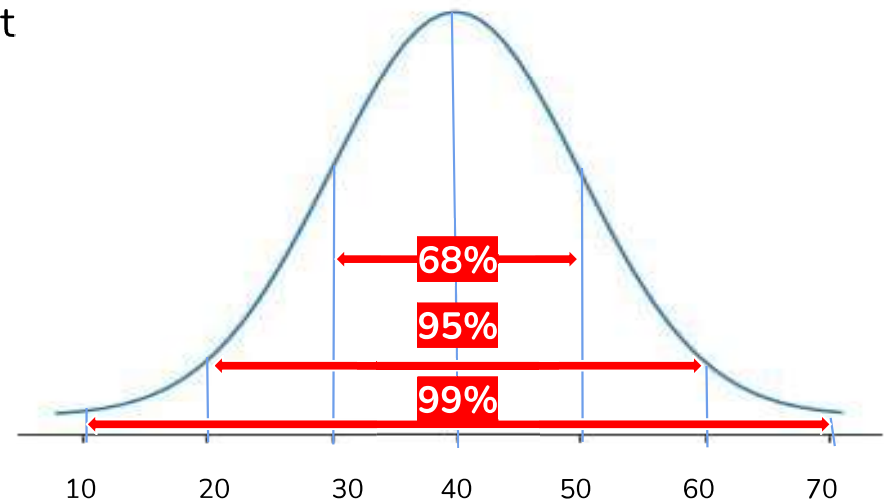
1. About 68% of the data fall within one standard deviation from the mean
2. About 95% of the data fall within two standard deviation from the mean
3. About 99.7% of the data fall within three standard deviation from the mean

Normal Distribution : Example

Assume that a food delivery service provider A has a mean delivery time of 40 minutes and a standard deviation of 10 minutes.

Using the Empirical Rule, we can determine that

- About 68% of the delivery times are between 30-50 minutes (40 ± 10)
- About 95% of the delivery times are between 20-60 minutes ($40 \pm 2(10)$)
- About 99.7% of the delivery times are between 10-70 minutes ($40 \pm 3(10)$)



This property is known as Empirical rule.

Normal Distribution : Area under Density Curve

As with any continuous probability distribution, the area under the density curve between two points indicates the probability that the variable will fall within that interval.



μ and σ are the parameters that decide the center and spread of the normal curve



To find the area, we need Calculus



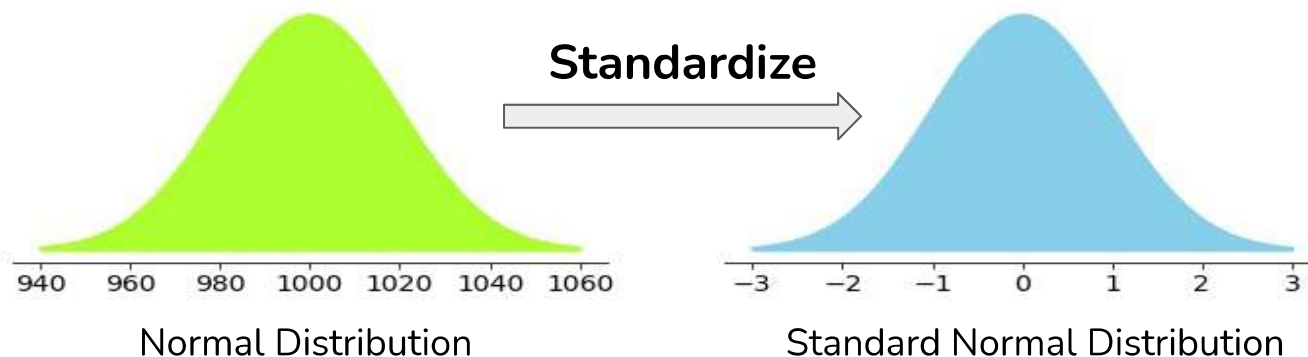
But, there is an easier way to do it in Python (or other softwares).
It provides us the necessary functions to calculate the area.

Normal Distribution : Standard Normal Distribution

A standard normal distribution is used to compare two normal distributions with different parameters (μ , σ)

The standard normal variable is denoted by Z and the distribution is also known as Z distribution

It always has a mean of 0 and standard deviation of 1



Normal Distribution : Z-Score

A normal variable can be converted to standard normal variable by subtracting the mean (μ) and dividing the standard deviation (σ):

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

where,

- X is the observed data point
- Z (**Z-Score/Standard score**) is the measure of the number of standard deviations above or below the mean that data point falls