E & C
I
T
IIT Guwahati

Advanced Certification Programme in Data Science Business Analytics

Math/Stats for Data Science

Topics Covered

- Probability Distributions
- Case Study
- Analysing House Price Data with R
- Q and A

Probability Distributions

Random Variable

Represents Different Values Based on a Random Experiment

- **Definition:** A variable that takes values based on the outcomes of a random experiment

Types of random variables:

- **Discrete:** Takes only specific, finite values
- **Continuous:** Takes any value within a range

Probability Distributions

Defines how Probabilities are Assigned to Possible Values of a Random Variable

- **Definition:** Describes the distribution of probabilities across a random variable's values
- **Purpose:** Helps model and understand uncertainty in real-world scenarios
- **Applications:** Used in finance, biology, engineering and social sciences

Types of Probability Distributions

Categorised into Discrete and Continuous Distributions

Discrete probability distributions

- Used for countable values
- Each value has a specific probability

Continuous probability distributions

- Used for values within a range
- Probabilities assigned over intervals

Discrete Probability Distribution

Applies to Countable Values with Specific Probabilities

- **Example:** Number of people watching a movie at a multiplex per day
- **Key insight:** The exact number varies daily, making it a discrete random variable
- **Observation:** Records show attendance ranges from 200 to 215

Discrete Probability Distribution

Frequency Distribution of Attendees Over the last 100 Days

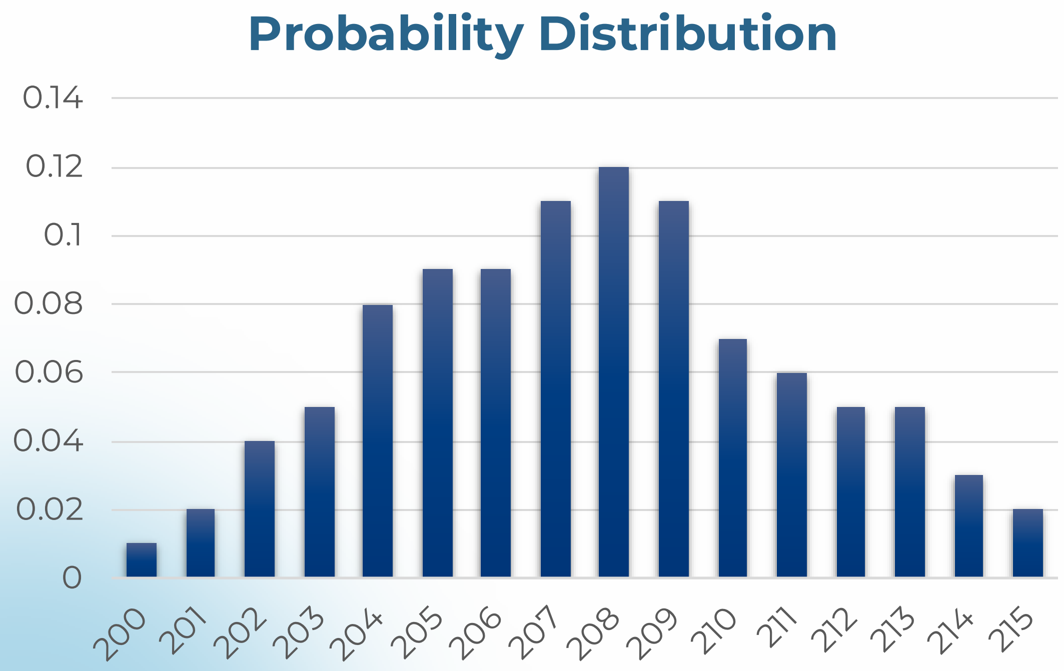
No Of People	Number Of Days This Event Was Observed
200	1
201	2
202	4
203	5
204	8
205	9
206	9
207	11

No Of People	Number Of Days This Event Was Observed
208	12
209	11
210	7
211	6
212	5
213	5
214	3
215	2
Total	**100**

- **Data representation:** Frequency distribution over 100 days

Discrete Probability Distribution

Below is the Probability Distribution for the Discrete Random Variable ‘No of People’.



- The probability distribution for a random variable provides a probability for each possible value and these probabilities must sum to 1

- i.e. $\Sigma p(x) = 1$

- where p(x) is probability of event x

Discrete Distribution: Binomial Distribution

A Probability Distribution Based on Bernoulli’s Experiment

- **Bernoulli’s experiment:** A process with two possible outcomes per trial

Conditions

- A sample of n experimental units is selected
- Each unit has two possible outcomes (success or failure)
- Probability of success (p) remains constant
- Outcomes are independent of each other

Formula for a Binomial Distribution

Calculates the Probability of ‘x’ Successes in ‘n’ Trials

- The probability of obtaining x successes in n trials with a probability p of success in each trial can be calculated using the formula;

$$p(X=x) = \left(\frac{n!}{x!(n-x)!}\right).p^{x}\ .q^{n-x}$$

- (where q = 1 - p)

Applications of Binomial Distribution

Used in Various Fields for Probability-Based Analysis

- **Product life cycle:** Survival age analysis
- **Risk management:** Quantifying operational risks in banking
- **Healthcare:** Assessing the risk of life-threatening diseases
- **Email analytics:** Predicting the probability of an email being read
- **Medical research:** Analysing allergy relief effectiveness

Poisson Probability Distribution

Models the Probability of Events Occurring in a Fixed Interval

- **Definition:** A discrete probability distribution introduced by Simeon Denis Poisson
- **Random variable (X):** Represents the number of times an event occurs in a given time or space

Poisson Probability Distribution

Examples of Poisson Probability Distribution

- **Typos per printed page** (*space-based*)
- **Cars passing an intersection per minute** (*time-based*)
- **Alaskan salmon caught in a driftnet** (*space-based*)
- **Customers at an ATM in 10-minute intervals** (*time-based*)
- **Students arriving during office hours** (*time-based*)

Poisson Probability Distribution

Determines the Probability of Event Occurrences in a Fixed Interval

- **Probability density function:**

$$f(x) = \frac{e^{-\lambda}\,\lambda^{x}}{x!}$$

For $x = 0, 1, 2, \ldots$ and $\lambda > 0$, where λ is both the mean and the variance of X

- **Single parameter:** λ determines the probability of an event

Applications of Poisson Distribution

Models Rare or Random Events Over Time or Space

- **Historical use:** Deaths by horse kicks in the prussian army
- **Medical applications:** Birth defects, genetic mutations and rare diseases
- **Traffic analysis:** Car accidents and traffic flow optimisation

Applications of Poisson Distribution

Models Rare or Random Events Over Time or Space

- **Quality Control:** Typing errors per page, foreign objects in food
- **Environmental Studies:** Spread of endangered animals
- **Equipment Maintenance:** Machine failures per month

Normal Distribution

The Most Important Continuous Probability Distribution

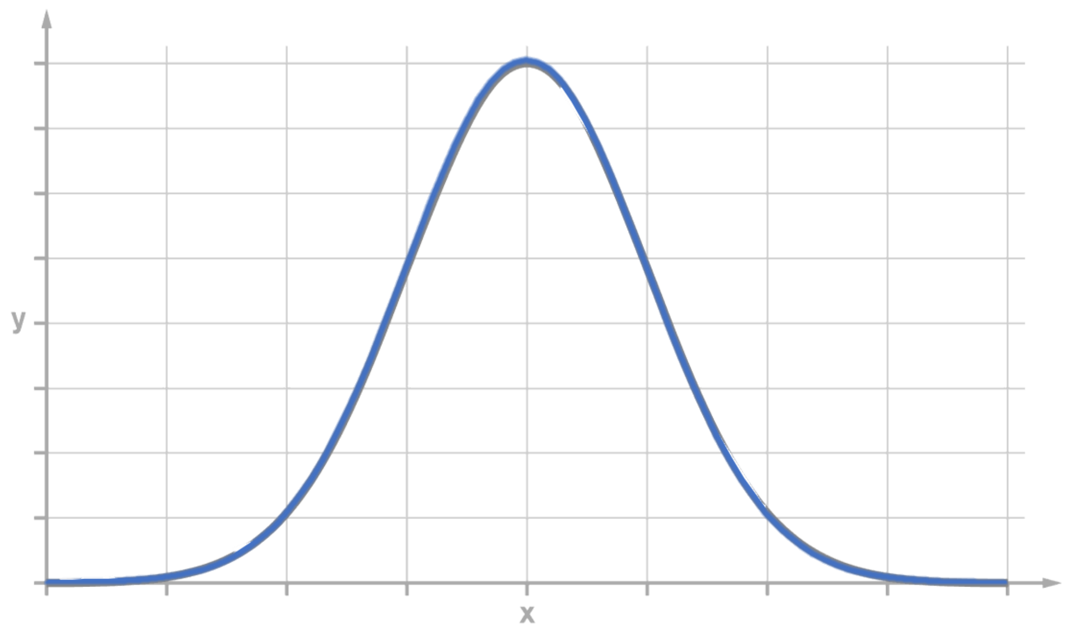
- **Definition:** A key distribution in probability and statistics

Significance:

- Used to draw conclusions from sample data
- Essential in statistical process control

Normal Distribution: Characteristics

Key Properties that define its Statistical Importance



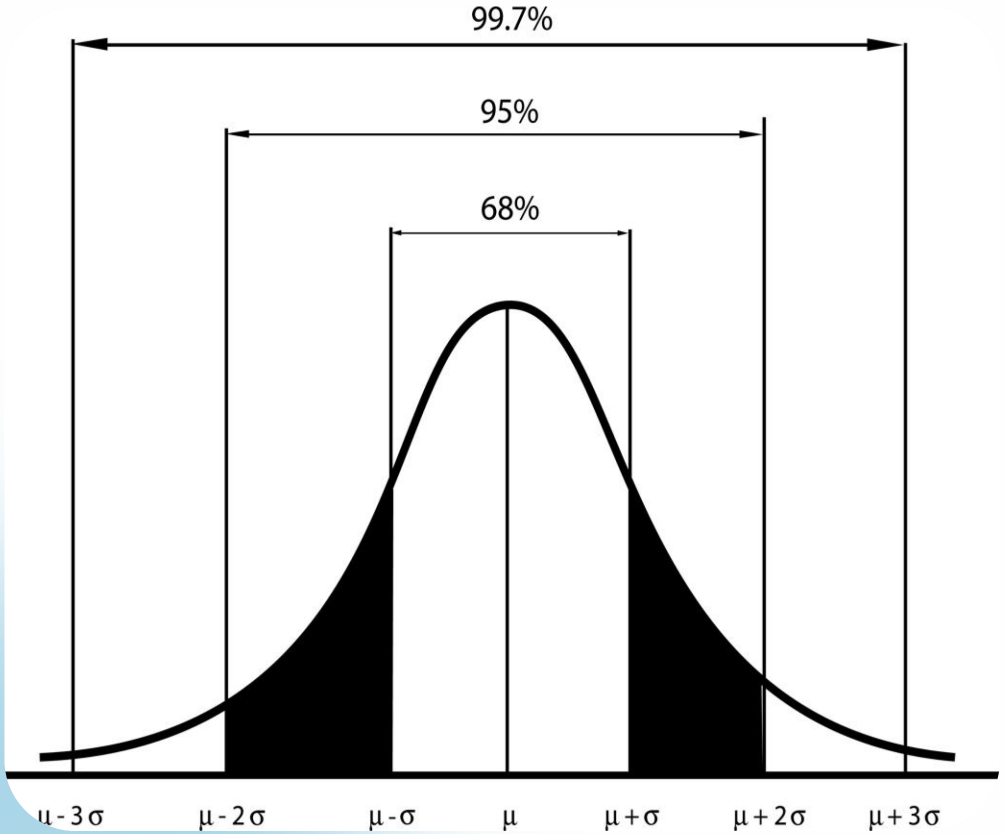
- **Shape:** Bell-shaped and symmetrical.
- **Mean = Median = Mode:** The distribution is centred around the mean

Observation Spread:

- Most values cluster near the mean
- Fewer values appear further from the mean

Normal Distribution: Characteristics

Key Properties That Define Its Statistical Importance



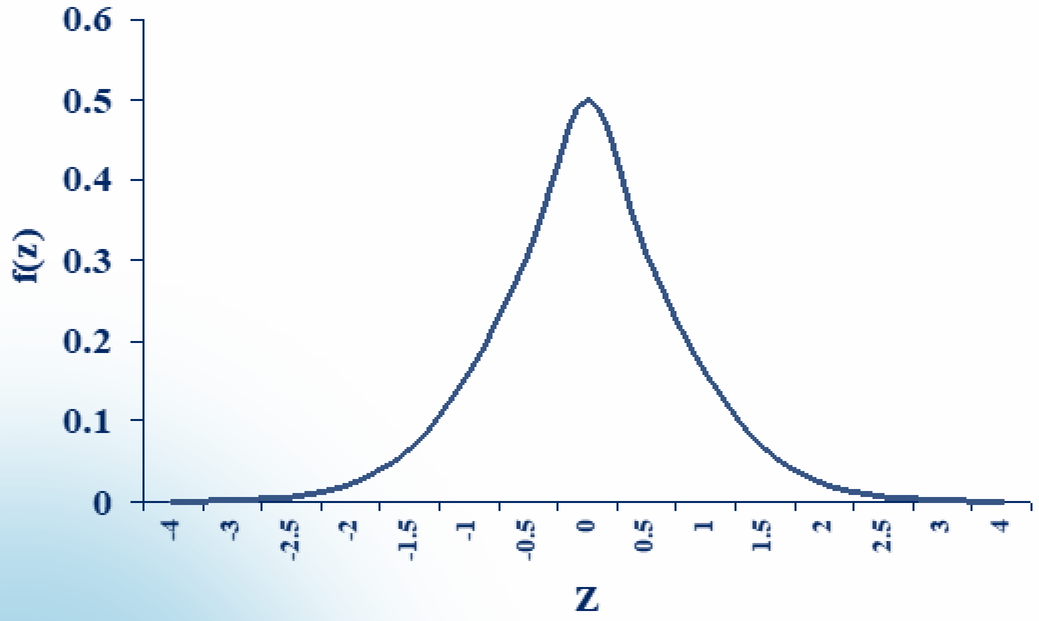
- **Defined by**: Mean (μ) and Standard Deviation (σ)

Empirical Rule:

- 68.3% of values within $\mu \pm 1\sigma$
- 95.4% of values within $\mu \pm 2\sigma$
- 99.7% of values within $\mu \pm 3\sigma$

Standard Normal Distribution

A Normal Distribution With Fixed Parameters



- **Definition:** A special case of the normal distribution
- **Mean (μ):** 0
- **Standard Deviation (σ):** 1
- **Standard Random Variable:** Denoted by .

Standard Normal Distribution

Converting Any Normal Distribution to Standard Form

- **Transformation process:** Adjust mean to 0 and standard deviation to 1

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

Steps:

- Subtract the mean (μ) from each observation
- Divide by the standard deviation (σ)

Case Study

Analysing House Price Data with R

Loading and Examining the Dataset

Importing, Displaying and Summarising Data

Load libraries:

- **dplyr –** Data manipulation
- **ggplot2 –** Data visualisation
- **tidyr –** Data tidying
- **knitr –** Creating tables

Load dataset:

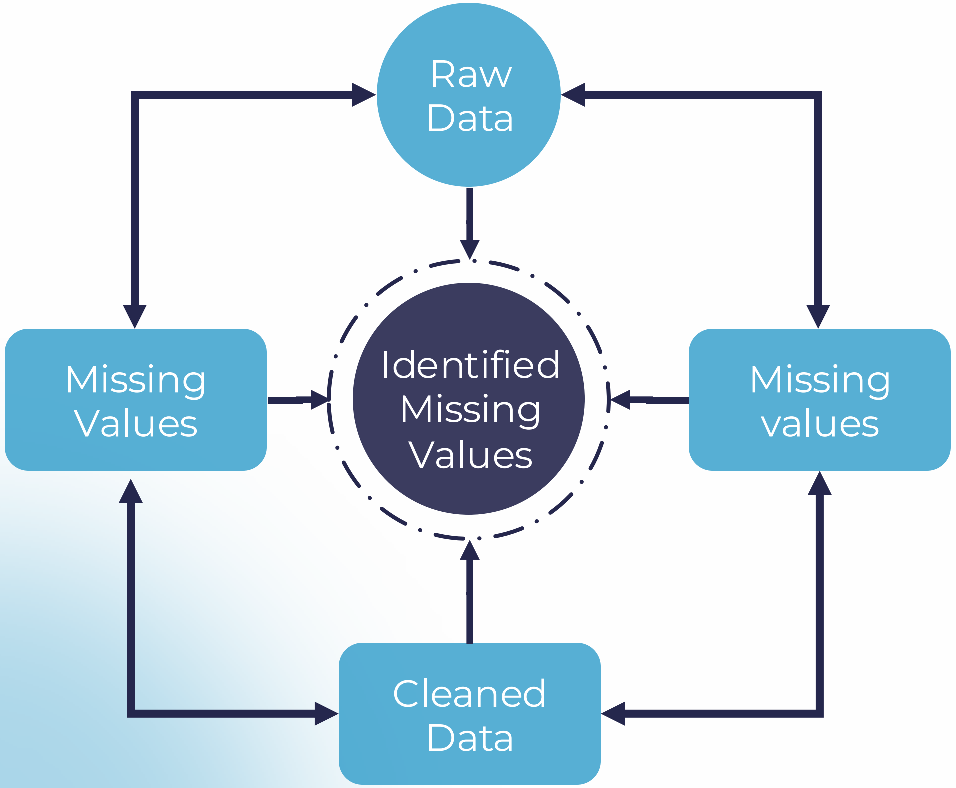
- data <- read.csv("houseprice.csv")

Display first few rows: head(data)

Dataset summary: summary(data)

Handling Missing Values

Identifying and Filling Missing Data



Check for missing values:

- missing_values <- sapply(data, function(x) sum(is.na(x)))
- missing_values

Fill missing values:

- data$LotFrontage[is.na(data$LotFrontage)] <- mean(data$LotFrontage, na.rm = TRUE)

Descriptive Statistics

Measuring Distance From the Mean

mean_LotFrontage	median_LotFrontage	sd_LotFrontage
mean_LotArea	median_LotArea	sd_LotArea
mean_OverallQual	median_OverallQual	sd_OverallQual

Visualising Lot Area Distribution

Creating and Saving a Histogram

Create histogram

```
ggplot(data, aes(x = LotArea)) +
geom_histogram(binwidth = 500, fill = "blue", color = "white") +
labs(title = "Distribution of Lot Area", x = "Lot Area", y = "Frequency") +
theme_minimal()
```

Save plot:

```
ggsave("lot_area_histogram.png")
```

Analysing Overall Quality vs Lot Area

Visualising the Relationship using a Boxplot

Create boxplot:

```
ggplot(data, aes(x = factor(OverallQual), y = LotArea)) +
geom_boxplot(fill = "lightgreen") +
labs(title = "Boxplot of Lot Area by Overall Quality",
     x = "Overall Quality",
     y = "Lot Area") + theme_minimal()
```

- **Interpretation:** Shows how lot sizes vary across different quality ratings

Save plot:

```
ggsave("overall_quality_boxplot.png")
```

Correlation Analysis

Measuring Relationships Between Numerical Variables

Calculate correlation matrix:

```
correlation_matrix <- cor(data %>% select_if(is.numeric), use = "complete.obs")
print(correlation_matrix)
```

Create heatmap:

```
library(reshape2)
correlation_melted <- melt(correlation_matrix)
ggplot(correlation_melted, aes(Var1, Var2, fill = value)) + geom_tile() +
  scale_fill_gradient2(midpoint = 0, low = "red", mid = "white", high = "blue", limit = c(-1, 1))
+  theme_minimal() +  labs(title = "Correlation Heatmap", x = "", y = "")
```

Save plot:

```
ggsave("correlation_heatmap.png")
```

Saving Results

Exporting Statistics, Plots and Reports

Save summary statistics:

```
write.csv(descriptive_stats, "descriptive_statistics_summary.csv", row.names = FALSE)
```

Save plots:

```
ggsave("lot_area_histogram.png")
ggsave("overall_quality_boxplot.png")
ggsave("correlation_heatmap.png")
```

Generate report:

Use the saved CSV file and images for reporting house price analysis

Q & A

Thank you