VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF TECHNOLOGY
FACULTY OF APPLIED SCIENCE

PROBABILITY & STATISTICS (MT2013)

Semester: 231

Central Processing Units (CPUs)

Advisor:	Phan Thị Khánh Vân,	FAS - HCMUT
Students:	Lê Nguyễn Gia Bảo	- 2210216.
	Trần Đình Đăng Khoa	- 2211649.
	Bùi Vũ Thiên Đăng	- 2252151.
	Trần Tuấn Minh Khoa	- 2252365.
	Nguyễn Hữu Trí	- 2252842.

HO CHI MINH CITY, APRIL 2024

Contents

1 Introduction

Overview

The dataset in question is a comprehensive collection of detailed technical specifications, release dates, and pricing information for a vast range of central processing units (CPUs) utilized in computer systems. The data is organized in a structured format, such as a comma-separated values (CSV) file, facilitating efficient data processing and analysis.

Central processing units, commonly referred to as CPUs, are the primary computational engines within a computer system. They are responsible for executing instructions, performing calculations, and coordinating the various components and peripherals of the system. CPUs are considered the brain of a computer, playing a crucial role in determining its overall performance and capabilities.

The CPU market is dominated by a few major semiconductor manufacturers, with Intel and AMD being the most prominent players. These companies have established themselves as industry leaders, continuously pushing the boundaries of CPU design and performance through innovative architectures, manufacturing processes, and feature enhancements.

The dataset likely encompasses a comprehensive array of attributes and characteristics pertaining to the CPUs, encompassing various essential metrics. These attributes may include, but are not limited to, clock speed, number of cores and threads, cache sizes, supported instruction sets, manufacturing process technology, thermal design power (TDP), and release or launch dates. Additionally, the dataset may contain information on initial retail pricing, socket or platform compatibility, and other relevant technical specifications.

By leveraging this extensive dataset, researchers, analysts, and industry professionals can conduct in-depth analyses and comparisons of CPU performance, efficiency, and pricing trends across different manufacturers and product generations. Such analyses can provide valuable insights into the evolution of CPU technology over time, enabling informed decision-making processes for hardware procurement, optimal resource allocation, and identifying potential areas for technological advancements or performance optimizations.

Furthermore, the dataset can serve as a valuable resource for academic research, enabling investigations into various aspects of CPU design, architecture, and performance optimization techniques. It can also facilitate the development and benchmarking of CPU-intensive applications, algorithms, and computational models across diverse domains, fostering interdisciplinary collaborations and driving innovation within the field of high-performance computing.

By combining this dataset with other relevant data sources, such as system benchmarks, power consumption measurements, and real-world application performance metrics, researchers and developers can gain a comprehensive understanding of the intricate relationships between CPU specifications, system performance, and energy efficiency, ultimately leading to more informed decisions and optimizations in the design and deployment of computer systems.

2 Background Knowledge

2.1 Hypothesis Testing

Hypothesis testing is a fundamental statistical procedure used to make inferences about population parameters based on sample data. It is widely employed in various fields, including scientific research, quality control, and decision-making processes. The primary objective of hypothesis testing is to evaluate the plausibility of a specific claim or hypothesis concerning a population parameter, such as the mean, proportion, or variance.

In hypothesis testing, two mutually exclusive hypotheses are formulated: the null hypothesis (H_0) and the alternative hypothesis (H_a). The null hypothesis typically represents the status quo, the baseline assumption, or the claim that the researcher wishes to test against. The alternative hypothesis represents the opposite or the alternative claim that the researcher aims to support or conclude if the null hypothesis is rejected.

The process of hypothesis testing involves the following steps:

1. Formulate the null hypothesis (H_0) and the alternative hypothesis (H_a).
2. Specify the significance level (α), which is the probability of rejecting the null hypothesis when it is true (Type I error).
3. Calculate the test statistic from the sample data.
4. Determine the critical region or the critical value(s) based on the significance level and the chosen test.
5. Compare the test statistic with the critical region or critical value(s).
6. Make a decision: Reject or fail to reject the null hypothesis.

The decision to reject or fail to reject the null hypothesis is based on the comparison between the test statistic and the critical region or critical value(s). If the test statistic falls within the critical region, the null hypothesis is rejected in favor of the alternative hypothesis. If the test statistic does not fall within the critical region, the null hypothesis is not rejected.

There are several types of hypothesis tests, including:

- Tests for means (one-sample, two-sample, and paired data)
- Tests for proportions
- Tests for variances
- Tests for correlation and regression coefficients
- Goodness-of-fit tests
- Non-parametric tests

The choice of the appropriate hypothesis test depends on the nature of the data, the research question, and the assumptions underlying the statistical model.

It is important to note that hypothesis testing is subject to two types of errors: Type I error (rejecting the null hypothesis when it is true) and Type II error (failing to reject the null hypothesis when it is false). The significance level (α) controls the probability of committing a Type I error, while the power of the test (1 - β) represents the probability of correctly rejecting the null hypothesis when it is false, where β is the probability of committing a Type II error.

Hypothesis testing is a powerful tool for making statistical inferences and drawing conclusions from data. However, it is crucial to carefully interpret the results, consider the practical implications, and recognize the limitations and assumptions underlying the chosen statistical test.

2.2 Linear Regression

Linear regression is a fundamental statistical technique used to model the relationship between a dependent variable and one or more independent variables. It is widely employed in various fields, including economics, finance, engineering, and social sciences, to analyze and make predictions based on observed data.

The primary objective of linear regression is to find the best-fitting straight line that describes the relationship between the dependent variable (also known as the response variable) and the independent variable(s) (also known as predictor variables or explanatory variables). This line is represented by a linear equation, which takes the following form:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

Where:

- Y is the dependent variable
- β_0 is the intercept (the value of y when all independent variables are zero)
- $\beta_1, \beta_2, ..., \beta_n$ are the coefficients (slopes) associated with the respective independent variables
- $x_1, x_2, ..., x_n$ are the independent variables
- ε is the error term, representing the difference between the observed values and the predicted values

The process of linear regression involves estimating the values of the coefficients (β_0, β_1, β_2, ..., β_n) using a set of observed data points. This estimation is typically performed using the method of least squares, which aims to minimize the sum of squared differences between the observed values and the predicted values obtained from the linear equation. Linear regression models can be classified into two main types:

- Simple Linear Regression: This model involves only one independent variable and is represented by the equation $Y = \beta_0 + \beta_1 x + \varepsilon$.
- Multiple Linear Regression: This model involves two or more independent variables and is represented by the equation $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$

Once the linear regression model is fitted to the data, it can be used for various purposes, such as:

1. Prediction: The model can be used to predict the value of the dependent variable based on new values of the independent variables.

2. Inference: Statistical tests can be performed to assess the significance of the independent variables and the overall model fit.

3. Interpretation: The coefficients of the independent variables can be interpreted to understand the magnitude and direction of their impact on the dependent variable.

It is important to note that linear regression models make several assumptions, including linearity, normality of residuals, homoscedasticity (constant variance of residuals), and independence of observations. Violations of these assumptions can lead to biased or inefficient estimates and invalid statistical inferences.

Additionally, linear regression models are susceptible to issues such as multicollinearity (high correlation among independent variables) and outliers, which can influence the model's performance and interpretability. Various diagnostic techniques and model validation methods are employed to assess the reliability and robustness of the linear regression model.

Linear regression serves as a foundation for more advanced regression techniques, such as logistic regression (for binary or categorical dependent variables), nonlinear regression, and time series analysis, among others. Its simplicity, interpretability, and widespread applicability make linear regression a fundamental tool in statistical modeling and data analysis.

3 Data Pre-processing

3.1 Load Data

```
# Importing data
intel_cpu <- read.csv("~/Downloads/archive/Intel_CPUs.csv")
```

This line of R code is used to import a dataset from a Comma-Separated Values (CSV) file into the R environment. The `read.csv()` function is a built-in function in R that reads a CSV file and creates a data frame object from its contents. A data frame is a two-dimensional tabular data structure in `R`, where each column represents a variable, and each row represents an observation.

In this specific code: "**~/Downloads/archive/Intel_CPUs.csv**" is the file path that specifies the location and name of the CSV file to be imported, `intel_cpu` is the name assigned to the data frame object that will store the imported data.

After executing this line of code, the contents of the "**Intel_CPUs.csv**" file will be read and stored in the `intel_cpu` data frame within the `R` environment. The data frame will have the same structure as the CSV file, with columns representing variables and rows representing observations.

3.2 Explore Data

```
# The head() function is used to preview the first few rows
# of the data frame
head(intel_cpu)
```

This line calls the `head()` function and passes the `intel_cpu` data frame as an argument. By default, the `head()` function prints the first *six* rows of the given data frame or matrix.

The `head()` function is a valuable tool for data exploration and validation, especially when working with large datasets. By previewing the initial rows, you can quickly assess the structure of the data, check the column names, and ensure that the data has been imported correctly.

Inspecting the first few rows can reveal potential issues or anomalies in the data, such as missing values, incorrect data types, or unexpected values. It also provides an initial glimpse into the content and format of the data, which can inform subsequent data cleaning, transformation, or analysis steps.

By executing `head(intel_cpu)`, the output will display the first *six* rows of the `intel_cpu` data frame, allowing you to visually inspect the data and make informed decisions about the next steps in the data analysis workflow.

3.3 Handle Missing Values

3.4 Handle Outliers

3.5 Feature Scaling/Normalization

4 Conclusion

References

[1] Peter Bruce and Andrew Bruce. *Practical Statistics for Data Scientists*. O'Reilly, 2017.

[2] Nguyễn Tiến Dũng (chủ biên) and Nguyễn Đình Huy. *Xác suất - Thống kê & Phân tích số liệu*. 2019.

[3] Douglas C Montgomery and George C Runger. *Applied Statistics and Probability for Engineers*. Wiley, 6th edition, 2006.

[4] Sheldon M Ross. *Introduction to Probability and Statistics for Engineers and Scientists*. Academic Press, 2020.