ĐỀ CƯƠNG ÔN TẬP

KINH TẾ LƯỢNG VÀ SPSS CAO HỌC

Dành cho học viên cao học và nghiên cứu sinh chưa từng học kinh tế lượng hay SPSS, cần hiểu để thi và để tự phân tích dữ liệu

Econometrics and SPSS Study Guide for Master's Students – a bilingual guide

SPSS cơ bản · thống kê mô tả · ước lượng · kiểm định giả thuyết

Hồi quy OLS · vi phạm giả định · biến công cụ · 2SLS · GMM · ML · logit · Tobit

Chuỗi thời gian · ARMA · GARCH · VAR · đồng liên kết · mô phỏng

Song ngữ Việt – Anh · Vietnamese – English

Phiên bản 1.0 – Tháng 10 năm 2026

https://tailieuso.vin

MỤC LỤC

CONTENTS

Chương 1. Làm quen SPSS và chuẩn bị dữ liệu – *Chapter 1. Getting Started with SPSS and Preparing Data* 7

1.1. Giao diện SPSS và lần chạy phân tích đầu tiên – *1.1. The SPSS Interface and Your First Analysis* 7

1.2. Đọc dữ liệu từ Excel, tệp văn bản và cơ sở dữ liệu – *1.2. Reading Data from Excel, Text Files and Databases* 10

1.3. Nhập số liệu và khai báo biến trong Data Editor – *1.3. Entering Data and Defining Variables in the Data Editor* 12

1.4. Giá trị khuyết và cách xử lý – *1.4. Handling Missing Values* 15

1.5. Biến đổi dữ liệu: tính biến mới, mã hoá lại, chọn ca – *1.5. Transforming Data: Computing, Recoding and Selecting Cases* 17

1.6. Các loại thang đo và loại dữ liệu – *1.6. Levels of Measurement and Types of Data* ..20

1.7. Kinh tế lượng là gì và các bước xây dựng một mô hình – *1.7. What Econometrics Is and the Steps of Building a Model* 24

1.8. Tỷ suất sinh lợi và cách dùng trong mô hình tài chính – *1.8. Returns and How They Enter Financial Models* 26

1.9. Bài tập vận dụng có lời giải – *1.9. Applied exercises with solutions* 28

1.10. Câu hỏi tự luận và dàn ý trả lời – *1.10. Essay questions and answer outlines* 30

1.11. Câu hỏi trắc nghiệm – *1.11. Multiple-choice questions* 32

Chương 2. Mô tả và khám phá dữ liệu – *Chapter 2. Describing and Exploring Data* .. 35

2.1. Bảng tần số và biểu đồ cho biến phân loại – *2.1. Frequency Tables and Charts for Categorical Variables* 35

2.2. Trung bình, trung vị, độ lệch chuẩn và độ méo – *2.2. Mean, Median, Standard Deviation and Skewness* 38

2.3. Explore: ngoại lai, biểu đồ hộp và so sánh nhóm – *2.3. Explore: Outliers, Boxplots and Group Comparison* 40

2.4. Bảng chéo Crosstabs và kiểm định chi bình phương – *2.4. Crosstabs and the Chi-Square Test* 43

2.5. Phân phối chuẩn và kiểm tra tính chuẩn của dữ liệu – *2.5. The Normal Distribution and Checking Normality* 45

2.6. Biểu đồ phân tán và nhận diện quan hệ giữa hai biến – *2.6. Scatterplots and Spotting Relationships Between Two Variables* 48

2.7. Codebook và báo cáo mô tả gọn cho luận văn – *2.7. Codebook and Compact Descriptive Reporting* 50

2.8. Lỗi hay gặp khi mô tả dữ liệu và ôn tập – *2.8. Common Mistakes in Descriptive Work and Review* 52

2.9. Bài tập vận dụng có lời giải – *2.9. Applied exercises with solutions* 55

2.10. Câu hỏi tự luận và dàn ý trả lời – *2.10. Essay questions and answer outlines* 57

2.11. Câu hỏi trắc nghiệm – *2.11. Multiple-choice questions* 59

Chương 3. Xác suất, lấy mẫu và ước lượng – *Chapter 3. Probability, Sampling and Estimation* 62

3.1. Biến ngẫu nhiên, kỳ vọng và phương sai – *3.1. Random Variables, Expectation and Variance* 62

3.2. Các phân phối chuẩn, t, chi bình phương và F – *3.2. The Normal, t, Chi-Square and F Distributions* 64

3.3. Phân phối của trung bình mẫu và định lý giới hạn trung tâm – *3.3. Sampling Distribution of the Mean and the Central Limit Theorem* 67

3.4. Ước lượng điểm: không chệch, vững và hiệu quả – *3.4. Point Estimation: Unbiased, Consistent and Efficient* 69

3.5. Khoảng tin cậy cho trung bình và tỷ lệ – *3.5. Confidence Intervals for Means and Proportions* 72

3.6. Khoảng tin cậy tiệm cận và dựng khoảng cho tham số chưa biết – *3.6. Asymptotic Confidence Intervals for Unknown Parameters* 75

3.7. Cỡ mẫu và phân tích power – *3.7. Sample Size and Power Analysis* 77

3.8. Mô phỏng Monte Carlo và bootstrap – *3.8. Monte Carlo Simulation and the Bootstrap* 79

3.9. Bài tập vận dụng có lời giải – *3.9. Applied exercises with solutions* 82

3.10. Câu hỏi tự luận và dàn ý trả lời – *3.10. Essay questions and answer outlines* 84

3.11. Câu hỏi trắc nghiệm – *3.11. Multiple-choice questions* 86

Chương 4. Kiểm định giả thuyết và so sánh trung bình – *Chapter 4. Hypothesis Testing and Comparing Means* 88

4.1. Logic của kiểm định: H0, H1, mức ý nghĩa, sai lầm loại I và II – *4.1. The Logic of Testing: H0, H1, Significance Level, Type I and II Errors* 88

4.2. Kiểm định t một mẫu và cách đọc p-value – *4.2. One-Sample t Test and Reading the p-value* 90

4.3. So sánh hai trung bình độc lập: phương sai bằng và khác nhau – *4.3. Comparing Two Independent Means: Equal and Unequal Variances* 93

4.4. So sánh trung bình mẫu ghép cặp – *4.4. Paired-Sample Comparisons* 95

4.5. Kiểm định tỷ lệ và phân tích thí nghiệm xã hội – *4.5. Testing Proportions and Analysing Social Experiments* 98

4.6. Thiết kế thí nghiệm: chọn tỷ lệ nhóm xử lý để sai số nhỏ nhất – *4.6. Experiment Design: Choosing the Treated Share to Minimise Variance* 100

4.7. Giả thuyết ngầm và kiểm định tiệm cận – *4.7. Implicit Hypotheses and Asymptotic Tests* 102

4.8. Phân tích phương sai ANOVA một nhân tố – *4.8. One-Way Analysis of Variance* 105

4.9. Bài tập vận dụng có lời giải – *4.9. Applied exercises with solutions* 107

4.10. Câu hỏi tự luận và dàn ý trả lời – *4.10. Essay questions and answer outlines* 109

4.11. Câu hỏi trắc nghiệm – *4.11. Multiple-choice questions* 112

Chương 5. Tương quan và hồi quy tuyến tính – *Chapter 5. Correlation and Linear Regression* 114

5.1. Hệ số tương quan và ý nghĩa thống kê – *5.1. Correlation Coefficients and Their Significance* 114

5.2. Hồi quy đơn và phương pháp bình phương nhỏ nhất – *5.2. Simple Regression and Ordinary Least Squares* 116
5.3. Đọc bảng kết quả hồi quy trong SPSS – *5.3. Reading SPSS Regression Output* 119
5.4. Hồi quy bội và cách diễn giải hệ số – *5.4. Multiple Regression and Interpreting Coefficients* 121
5.5. Biến giả, biến tương tác và dạng hàm – *5.5. Dummy Variables, Interactions and Functional Form* 124
5.6. Kiểm định hệ số và kiểm định F cho nhóm biến – *5.6. Testing Coefficients and Joint F Tests* 126
5.7. Chọn mô hình, so sánh mô hình và dự báo – *5.7. Model Selection, Comparison and Prediction* 128
5.8. Lỗi hay gặp khi diễn giải hồi quy và ôn tập chương – *5.8. Common Interpretation Errors and Chapter Review* 131
5.9. Bài tập vận dụng có lời giải – *5.9. Applied exercises with solutions* 134
5.10. Câu hỏi tự luận và dàn ý trả lời – *5.10. Essay questions and answer outlines* 137
5.11. Câu hỏi trắc nghiệm – *5.11. Multiple-choice questions* 139

Chương 6. Vi phạm giả định của OLS và cách xử lý – *Chapter 6. Violations of OLS Assumptions and Remedies* 142
6.1. Các giả định Gauss-Markov và hệ quả khi bị vi phạm – *6.1. The Gauss-Markov Assumptions and What Breaks* 142
6.2. Phương sai sai số thay đổi: phát hiện và chữa – *6.2. Heteroskedasticity: Detection and Remedies* 145
6.3. Tự tương quan phần dư và kiểm định Durbin-Watson – *6.3. Autocorrelation and the Durbin-Watson Test* 148
6.4. Đa cộng tuyến và chỉ số VIF – *6.4. Multicollinearity and the VIF* 151
6.5. Phần dư không chuẩn, điểm ảnh hưởng và ngoại lai – *6.5. Non-Normal Residuals, Influential Points and Outliers* 153
6.6. Sai dạng hàm, bỏ sót biến và sai số đo – *6.6. Misspecification, Omitted Variables and Measurement Error* 155
6.7. Kiểm định ràng buộc phi tuyến và thay đổi cấu trúc – *6.7. Testing Nonlinear Restrictions and Structural Change* 158
6.8. Quy trình chẩn đoán trọn vẹn và ôn tập – *6.8. A Full Diagnostic Routine and Review* 161
6.9. Bài tập vận dụng có lời giải – *6.9. Applied exercises with solutions* 163
6.10. Câu hỏi tự luận và dàn ý trả lời – *6.10. Essay questions and answer outlines* 165
6.11. Câu hỏi trắc nghiệm – *6.11. Multiple-choice questions* 167

Chương 7. Biến nội sinh, biến công cụ, 2SLS và GMM – *Chapter 7. Endogeneity, Instrumental Variables, 2SLS and GMM* 170
7.1. Nội sinh là gì và vì sao OLS bị chệch – *7.1. What Endogeneity Is and Why OLS Is Biased* 170
7.2. Biến công cụ: hai điều kiện và cách chọn – *7.2. Instrumental Variables: Two Conditions and How to Choose* 172

7.3. Bình phương nhỏ nhất hai giai đoạn (2SLS) – *7.3. Two-Stage Least Squares (2SLS)* 175
7.4. Sai số chuẩn của 2SLS: làm tay khác phần mềm ở đâu – *7.4. 2SLS Standard Errors: Manual Versus Software* 177
7.5. Công cụ yếu và kiểm định quá đồng nhất – *7.5. Weak Instruments and Overidentification Tests* 180
7.6. Kiểm định Hausman về tính ngoại sinh – *7.6. The Hausman Test for Exogeneity* 182
7.7. Phương pháp GMM tổng quát – *7.7. The Generalized Method of Moments* 184
7.8. Hệ phương trình đồng thời: cung cầu, nhận dạng và 3SLS – *7.8. Simultaneous Equations: Supply and Demand, Identification and 3SLS* 186
7.9. Bài tập vận dụng có lời giải – *7.9. Applied exercises with solutions* 189
7.10. Câu hỏi tự luận và dàn ý trả lời – *7.10. Essay questions and answer outlines* 192
7.11. Câu hỏi trắc nghiệm – *7.11. Multiple-choice questions* 194

Chương 8. Khả năng cực đại và mô hình biến phụ thuộc bị giới hạn – *Chapter 8. Maximum Likelihood and Limited Dependent Variable Models* 197
8.1. Ước lượng khả năng cực đại – *8.1. Maximum Likelihood Estimation* 197
8.2. Ba kiểm định cổ điển: Wald, tỷ số hợp lý và nhân tử Lagrange – *8.2. The Three Classical Tests: Wald, Likelihood Ratio and Lagrange Multiplier* 199
8.3. Khả năng cực đại giả, sai số chuẩn vững và kiểm định điều kiện moment – *8.3. Quasi-Maximum Likelihood, Robust Standard Errors and Conditional Moment Tests* 202
8.4. Logit và probit cho biến phụ thuộc nhị phân – *8.4. Logit and Probit for Binary Outcomes* 205
8.5. Mô hình đa ứng: biến phân loại nhiều mức – *8.5. Multi-Response Models* 207
8.6. Dữ liệu đếm: Poisson và các mở rộng – *8.6. Count Data: Poisson and Extensions* 209
8.7. Mô hình Tobit và các mở rộng – *8.7. The Tobit Model and Extensions* 212
8.8. Sai lệch chọn mẫu và tác động xử lý – *8.8. Sample Selection Bias and Treatment Effects* 214
8.9. Mô hình thời lượng – *8.9. Duration Models* 217
8.10. Bài tập vận dụng có lời giải – *8.10. Applied exercises with solutions* 219
8.11. Câu hỏi tự luận và dàn ý trả lời – *8.11. Essay questions and answer outlines* 222
8.12. Câu hỏi trắc nghiệm – *8.12. Multiple-choice questions* 224

Chương 9. Chuỗi thời gian đơn biến – *Chapter 9. Univariate Time Series* 226
9.1. Chuỗi thời gian, tính dừng và hàm tự tương quan – *9.1. Time Series, Stationarity and the Autocorrelation Function* 226
9.2. Tiến trình trung bình trượt MA – *9.2. Moving Average Processes* 228
9.3. Tiến trình tự hồi quy AR – *9.3. Autoregressive Processes* 230
9.4. Nghiệm đơn vị và bước đi ngẫu nhiên – *9.4. Unit Roots and Random Walks* 233
9.5. Ước lượng mô hình ARMA theo phương pháp Box-Jenkins – *9.5. Estimating ARMA Models with the Box-Jenkins Method* 235
9.6. Dự báo bằng ARMA và đánh giá sai số dự báo – *9.6. Forecasting with ARMA and Evaluating Forecast Errors* 238
9.7. Mùa vụ trong dữ liệu tài chính – *9.7. Seasonality in Financial Data* 240

9.8. Bài tập vận dụng có lời giải – *9.8. Applied exercises with solutions* 243
9.9. Câu hỏi tự luận và dàn ý trả lời – *9.9. Essay questions and answer outlines* 245
9.10. Câu hỏi trắc nghiệm – *9.10. Multiple-choice questions* 247

Chương 10. Mô hình biến động và hệ số tương quan – *Chapter 10. Volatility Models and Correlations* 249
10.1. Mô hình ARCH và GARCH cho biến động thay đổi – *10.1. ARCH and GARCH for Changing Volatility* 249
10.2. Mở rộng bất cân xứng: GJR, EGARCH và GARCH-in-mean – *10.2. Asymmetric Extensions: GJR, EGARCH and GARCH-in-Mean* 252
10.3. Kiểm định bất cân xứng: sign bias và đường tác động tin tức – *10.3. Testing for Asymmetry: Sign Bias Tests and the News Impact Curve* 254
10.4. Mô hình biến động ngẫu nhiên – *10.4. Stochastic Volatility Models* 256
10.5. Dự báo hiệp phương sai và GARCH đa biến – *10.5. Forecasting Covariances and Multivariate GARCH* 259
10.6. Kiểm định ràng buộc phi tuyến trong mô hình biến động và ôn tập – *10.6. Nonlinear Restriction Tests in Volatility Models and Review* 261
10.7. Bài tập vận dụng có lời giải – *10.7. Applied exercises with solutions* 264
10.8. Câu hỏi tự luận và dàn ý trả lời – *10.8. Essay questions and answer outlines* 266
10.9. Câu hỏi trắc nghiệm – *10.9. Multiple-choice questions* 268

Chương 11. Chuỗi thời gian đa biến, đồng liên kết và mô hình chuyển đổi – *Chapter 11. Multivariate Time Series, Cointegration and Switching Models* 271
11.1. Mô hình VAR: ước lượng và chọn độ trễ – *11.1. VAR Models: Estimation and Lag Selection* 271
11.2. VAR có biến ngoại sinh và tương tác giữa lợi suất tài sản với biến vĩ mô – *11.2. VAR with Exogenous Variables: Asset Returns and Macro Variables* 274
11.3. Kiểm định nhân quả Granger, hàm phản ứng đẩy và phân rã phương sai – *11.3. Granger Causality, Impulse Responses and Variance Decomposition* 276
11.4. Kiểm định đồng liên kết Engle-Granger dựa trên phần dư – *11.4. The Engle-Granger Residual-Based Cointegration Test* 279
11.5. Mô hình hiệu chỉnh sai số – *11.5. Error Correction Models* 281
11.6. Kiểm định Johansen dựa trên VAR – *11.6. The Johansen Test Based on VARs* 283
11.7. Ứng dụng: ngang giá sức mua, trái phiếu quốc tế và cấu trúc kỳ hạn – *11.7. Applications: Purchasing Power Parity, International Bonds and the Term Structure* 285
11.8. Mô hình chuyển đổi Markov và tự hồi quy ngưỡng – *11.8. Markov Switching and Threshold Autoregressive Models* 288
11.9. Ôn tập cuối học phần: câu hỏi tổng hợp và đề mẫu – *11.9. Final Review: Integrated Questions and Sample Exam* 290
11.10. Bài tập vận dụng có lời giải – *11.10. Applied exercises with solutions* 293
11.11. Câu hỏi tự luận và dàn ý trả lời – *11.11. Essay questions and answer outlines* 296
11.12. Câu hỏi trắc nghiệm – *11.12. Multiple-choice questions* 298

Phụ lục A. Đáp án trắc nghiệm – *Appendix A. Answers to the multiple-choice questions* 301

Phụ lục B. Bảng thuật ngữ Việt – Anh – *Appendix B. Vietnamese–English glossary*. 315

Phụ lục C. Tài liệu tham khảo – *Appendix C. References* 321

Chương 1. Làm quen SPSS và chuẩn bị dữ liệu – *Chapter 1. Getting Started with SPSS and Preparing Data*

Mục tiêu chương – *Chapter objectives*

- Mở dữ liệu từ Excel hoặc tệp văn bản, rồi khai báo tên, loại, nhãn và giá trị khuyết cho từng biến trong SPSS
  Open data from Excel or text files, then define the name, type, labels and missing values of each variable in SPSS
- Tính biến mới, mã hoá lại và chọn ca bằng Compute, Recode, Select Cases, rồi đọc kết quả ở cửa sổ Output Viewer
  Create new variables, recode, and select cases with Compute, Recode and Select Cases, then read the results in the Output Viewer
- Phân loại thang đo và kiểu dữ liệu, tính tỷ suất sinh lợi đơn giản và log, và kể đúng thứ tự các bước xây dựng một mô hình kinh tế lượng
  Classify measurement scales and data types, compute simple and log returns, and list the steps of building an econometric model in the correct order

1.1. Giao diện SPSS và lần chạy phân tích đầu tiên – *1.1. The SPSS Interface and Your First Analysis*

SPSS là phần mềm thống kê. Bạn đưa số liệu vào, chọn phép phân tích bằng menu, rồi nhận kết quả mà không phải tự tính tay. Nó giống máy tính bỏ túi, nhưng xử lý được cả bảng số hàng nghìn dòng.

SPSS is statistics software. You load your data, pick an analysis from the menus, and get results without calculating by hand. Think of a pocket calculator that can handle a table with thousands of rows.

Hai cửa sổ bạn dùng nhiều nhất – *The two windows you will use most*

Data Editor (trình soạn dữ liệu) là nơi chứa số liệu, trông như một bảng Excel. Mỗi hàng là một ca (case), tức một đối tượng được khảo sát, ví dụ một học sinh. Mỗi cột là một biến (variable), tức một đặc điểm đo được, ví dụ điểm Toán.

The Data Editor is where your data lives, and it looks like an Excel sheet. Each row is a case, meaning one surveyed unit such as one student. Each column is a variable, meaning one measured feature such as a maths score.

Data Editor có hai thẻ ở góc dưới. Data View cho bạn xem số liệu. Variable View liệt kê tên, kiểu và nhãn của từng biến, mỗi biến một hàng.

The Data Editor has two tabs at the bottom-left corner. Data View shows the values. Variable View lists the name, type and label of each variable, one variable per row.

Viewer (cửa sổ kết quả) tự mở ra ngay khi bạn chạy phân tích đầu tiên. Bên trái là khung mục lục liệt kê từng kết quả, bên phải là các bảng và biểu đồ. Bấm một mục bên trái thì SPSS nhảy tới kết quả đó.

The Viewer (output window) opens by itself the first time you run an analysis. The left pane is an outline listing each result, and the right pane holds the tables and charts. Click an item on the left and SPSS jumps to that result.

Mở một tệp dữ liệu mẫu – *Opening a sample data file*

Tệp dữ liệu của SPSS có đuôi .sav. Phần mềm đi kèm một số tệp mẫu để tập, trong đó có demo.sav. Bạn nên tập trên tệp mẫu trước khi động vào số liệu luận văn của mình.

SPSS data files end in .sav. The software ships with sample files for practice, including demo.sav. Practise on a sample file before you touch your own thesis data.

1. Trong Data Editor, vào menu File > Open > Data...
   In the Data Editor, choose File > Open > Data... from the menu.
2. Trong hộp thoại hiện ra, tìm thư mục chứa tệp mẫu của SPSS trên máy bạn, rồi chọn demo.sav. Mặc định hộp thoại chỉ liệt kê tệp .sav.
   In the dialog box, browse to the folder holding SPSS sample files on your computer and select demo.sav. By default the dialog lists only .sav files.
3. Bấm Open. Số liệu hiện ra trong Data View. Bấm thẻ Variable View để xem tên và nhãn các biến.
   Click Open. The data appears in Data View. Click the Variable View tab to see variable names and labels.

Chạy thủ tục đầu tiên – *Running your first procedure*

Thủ tục đơn giản nhất là Descriptives, cho ra số ca, giá trị nhỏ nhất, lớn nhất, trung bình và độ lệch chuẩn. Độ lệch chuẩn (standard deviation) cho biết các giá trị lệch khỏi trung bình bao xa, tính trung bình. Điểm cả lớp gần nhau thì con số này nhỏ.

The simplest procedure is Descriptives. It gives the number of cases, minimum, maximum, mean and standard deviation. The standard deviation tells you how far values typically sit from the mean. If a class scores close together, this number is small.

Giả sử bạn có tệp điểm Toán của 5 học sinh, biến tên là Toan. An được 6, Bình 7, Chi 8, Dũng 9, Hà 10. Bạn đã nhập tệp này vào Data Editor.

Suppose you have a file with the maths scores of 5 students, in a variable named Toan. An scored 6, Binh 7, Chi 8, Dung 9 and Ha 10. You have already entered this file in the Data Editor.

1. Vào menu Analyze > Descriptive Statistics > Descriptives...
   Choose Analyze > Descriptive Statistics > Descriptives... from the menu.
2. Chọn biến Toan ở danh sách bên trái, bấm nút mũi tên để chuyển sang ô Variable(s).
   Select the variable Toan in the left list and click the arrow button to move it into the Variable(s) box.
3. Bấm OK. Viewer mở ra với bảng Descriptive Statistics.
   Click OK. The Viewer opens with a Descriptive Statistics table.

Đọc kết quả – *Reading the output*

Bảng kết quả có một dòng cho biến Toan. Đọc từng cột như sau:

The table has one row for the variable Toan. Read each column as follows:

Cột trong bảng *Table column*	**Giá trị** *Value*	**Cách hiểu** *Meaning*
N	5	Có 5 học sinh có điểm hợp lệ. *Five students have a valid score.*
Minimum	6	Điểm thấp nhất lớp là 6. *The lowest score in the class is 6.*
Maximum	10	Điểm cao nhất lớp là 10. *The highest score in the class is 10.*
Mean	8,00 *8.00*	Trung bình là (6+7+8+9+10) chia 5 = 40 chia 5 = 8. *The mean is (6+7+8+9+10) divided by 5 = 40 divided by 5 = 8.*
Std. Deviation	1,581 *1.581*	Bình phương các độ lệch: 4+1+0+1+4 = 10. Chia cho 4 được 2,5, khai căn ra 1,581. *Squared deviations sum to 4+1+0+1+4 = 10. Divide by 4 to get 2.5, then take the square root: 1.581.*

Bạn tự kiểm lại bằng tay như trên là cách tốt nhất để tin vào kết quả. Chú ý SPSS chia cho N trừ 1 (ở đây là 4), không chia cho N.

Checking by hand as above is the best way to trust the output. Note that SPSS divides by N minus 1 (here 4), not by N.

VÍ DỤ / EXAMPLE – Kết luận cho lớp: điểm Toán trung bình 8,00, thấp nhất 6, cao nhất 10, độ lệch chuẩn 1,58. Điểm tập trung quanh 8, không có em nào quá xa.
Conclusion for the class: the mean maths score is 8.00, the lowest is 6, the highest is 10, and the standard deviation is 1.58. Scores cluster around 8 and no student is far off.

LỖI HAY GẶP / COMMON MISTAKE – Quên lưu cửa sổ kết quả: tệp .sav chỉ lưu số liệu, còn kết quả nằm ở tệp Viewer riêng. Đóng SPSS mà không lưu Viewer là mất bảng.
Forgetting to save the output window: the .sav file stores only data, while results live in a separate Viewer file. Close SPSS without saving the Viewer and the tables are gone.

LỖI HAY GẶP / COMMON MISTAKE – Nhập điểm vào hai cột cho cùng một biến, hoặc để mỗi học sinh một cột. SPSS cần mỗi học sinh một hàng, mỗi biến một cột.
Typing scores into two columns for one variable, or giving each student a column. SPSS needs one row per student and one column per variable.

LỖI HAY GẶP / COMMON MISTAKE – Không thấy biến trong danh sách của hộp thoại. Thường là biến kiểu chữ (string), mà Descriptives chỉ nhận biến số. Kiểm tra cột Type trong Variable View.
Not finding your variable in the dialog list. It is often a string variable, and Descriptives accepts only numeric ones. Check the Type column in Variable View.

PHẢI THUỘC / MEMORISE – Data Editor chứa số liệu (hàng là ca, cột là biến); Viewer chứa kết quả; đường dẫn chạy Descriptives là Analyze > Descriptive Statistics > Descriptives.
The Data Editor holds the data (rows are cases, columns are variables); the Viewer holds the results; the path to Descriptives is Analyze > Descriptive Statistics > Descriptives.

1.2. Đọc dữ liệu từ Excel, tệp văn bản và cơ sở dữ liệu – *1.2. Reading Data from Excel, Text Files and Databases*

Dữ liệu thật hiếm khi gõ tay vào SPSS. Thường bạn nhận một tệp Excel, một tệp văn bản tải từ sàn giao dịch, hoặc một cơ sở dữ liệu của cơ quan. SPSS lưu dữ liệu thành bảng: mỗi hàng là một ca (một quan sát, như một phiên giao dịch), mỗi cột là một biến (như giá đóng cửa). Khi nhập, bạn chỉ cần giữ đúng cấu trúc đó.
Real data is rarely typed into SPSS by hand. Usually you receive an Excel file, a text file downloaded from an exchange, or a database from an organisation. SPSS stores data as a table: each row is a case (one observation, such as one trading session) and each column is a variable (such as the closing price). When importing, your job is to keep that structure intact.

Đọc tệp Excel – *Reading an Excel file*

Trước khi mở, đóng tệp trong Excel và dọn bảng: dòng 1 là tên cột, không trộn ô, mỗi cột chỉ một loại dữ liệu, không có dòng tổng ở cuối.
Before opening, close the file in Excel and tidy the sheet: row 1 holds column names, no merged cells, one data type per column, and no totals row at the bottom.

1. Chọn File > Open > Data..., rồi ở ô Files of type chọn Excel (.xls, .xlsx, ∗.xlsm).
   Choose File > Open > Data..., then set Files of type to Excel (.xls, .xlsx, ∗.xlsm).
2. Chọn tệp và bấm Open. Hộp thoại nhập Excel hiện ra.
   Select the file and click Open. The Excel import dialog appears.
3. Giữ dấu tích ở "Read variable names from the first row of data" nếu dòng 1 là tên cột.
   Keep "Read variable names from the first row of data" ticked if row 1 holds column names.
4. Ở Worksheet, chọn đúng trang tính. Nếu chỉ cần một vùng, gõ vào ô Range, ví dụ A1:D6.
   Under Worksheet, pick the right sheet. To read only part of it, type a block in Range, for example A1:D6.
5. Bấm OK, rồi mở Variable View và kiểm tra cột Type của từng biến.
   Click OK, then open Variable View and check the Type column of each variable.

VÍ DỤ / EXAMPLE – Tệp GiaCP.xlsx, trang Gia, có bốn cột: Ngay, Ma, GiaDong, KhoiLuong. Năm hàng dữ liệu là 02/01/2025 VNM 62,5 1200000; 03/01/2025 VNM 63,0 950000; 06/01/2025 VNM 62,0 1100000; 02/01/2025 FPT 118,0 800000; 03/01/2025 FPT 119,5 870000. Nhập với Range A1:D6 và dấu tích tên biến, bạn được 5 ca và 4 biến. Ngay có kiểu Date, Ma kiểu String, GiaDong và KhoiLuong kiểu Numeric. Nếu hàng 7 của Excel có dòng "Tổng", Range A1:D7 sẽ biến chữ "Tổng" thành một ca thứ 6 và làm cả cột bị coi là chuỗi chữ.
The file GiaCP.xlsx, sheet Gia, has four columns: Ngay, Ma, GiaDong, KhoiLuong. The five data

> *rows are 02/01/2025 VNM 62.5 1200000; 03/01/2025 VNM 63.0 950000; 06/01/2025 VNM 62.0 1100000; 02/01/2025 FPT 118.0 800000; 03/01/2025 FPT 119.5 870000. Importing with Range A1:D6 and the variable-name box ticked gives 5 cases and 4 variables. Ngay is Date, Ma is String, GiaDong and KhoiLuong are Numeric. If row 7 in Excel held a "Total" line, a range of A1:D7 would turn that label into a sixth case and force the numeric column to read as text.*

Đọc tệp văn bản – *Reading a text file*

Tệp .txt hoặc .csv chỉ là chữ thuần, nên SPSS phải được chỉ cách cắt chữ thành cột. Công cụ làm việc này là Text Wizard (trình hướng dẫn nhập văn bản).

A .txt or .csv file is plain text, so SPSS must be told how to cut it into columns. The tool for this is the Text Wizard.

1. Chọn File > Import Data > Text Data... rồi mở tệp.
   Choose File > Import Data > Text Data... and open the file.
2. Ở bước hỏi cách sắp xếp biến, chọn Delimited nếu các cột cách nhau bằng dấu phẩy, tab hay chấm phẩy. Chọn Fixed width nếu mỗi cột chiếm số ký tự cố định.
   At the step about how variables are arranged, choose Delimited if columns are separated by commas, tabs or semicolons. Choose Fixed width if each column occupies a fixed number of characters.
3. Trả lời "Yes" cho câu hỏi tên biến nằm ở đầu tệp, nếu dòng đầu là tên cột.
   Answer "Yes" to the question about variable names at the top of the file if the first line holds column names.
4. Chọn đúng dấu ngăn cách (delimiter) và xem khung xem trước. Cột nào dính vào nhau nghĩa là chọn sai dấu.
   Pick the right delimiter and watch the preview. Columns glued together mean the wrong delimiter.
5. Ở bước cuối, gán kiểu dữ liệu cho từng cột, rồi bấm Finish.
   At the later step, assign a data type to each column, then click Finish.

Đọc từ cơ sở dữ liệu – *Reading from a database*

SPSS đọc được cơ sở dữ liệu nào có trình điều khiển ODBC (Open Database Connectivity, chuẩn để chương trình khác nhau nói chuyện với kho dữ liệu). Bạn phải cài đúng trình điều khiển trước. Trình điều khiển Access chỉ chạy với bản SPSS 32 bit.

SPSS can read any database that has an ODBC (Open Database Connectivity) driver, a standard that lets different programs talk to a data store. Install the right driver first. The Access driver works only with the 32-bit version of SPSS.

1. Chọn File > Import Data > Database > New Query...
   Choose File > Import Data > Database > New Query...
2. Chọn nguồn dữ liệu trong danh sách rồi bấm Next. Nhập mật khẩu nếu được hỏi.
   Pick the data source from the list and click Next. Enter a password if asked.
3. Chọn bảng, kéo các trường cần dùng sang khung bên phải.
   Select the table and drag the fields you need to the right-hand pane.

4. Nếu chỉ cần một phần, đặt điều kiện lọc ở bước giới hạn ca, rồi bấm Finish.
If you need only part of the data, set a filter at the step that limits cases, then click Finish.

LỖI HAY GẶP / COMMON MISTAKE – Dòng 1 không phải tên cột nhưng vẫn để dấu tích tên biến, nên một giá trị thật bị nuốt thành tên. Ngược lại, bỏ tích khi dòng 1 là tên cột thì cả cột thành chuỗi chữ.
Leaving the variable-name box ticked when row 1 is real data swallows a value as a name. Unticking it when row 1 is a header turns the whole column into text.

LỖI HAY GẶP / COMMON MISTAKE – Số kiểu Việt viết 62,5 nhưng máy đặt dấu thập phân là chấm, khiến SPSS đọc sai hoặc coi là chữ. Kiểm tra cài đặt vùng của máy hoặc đổi dấu trong Excel trước khi nhập.
Vietnamese-style numbers are written 62,5 while the machine expects a decimal point, so SPSS misreads them or treats them as text. Check the regional settings or fix the separator in Excel before importing.

LỖI HAY GẶP / COMMON MISTAKE – Cột ngày bị đọc thành số hoặc chuỗi. Khi đó bạn không sắp xếp theo thời gian được. Đặt lại Type thành Date trong Variable View.
A date column is read as a number or a string, so you cannot sort by time. Reset Type to Date in Variable View.

MẸO / TIP – Nhập xong, lưu ngay bằng File > Save As thành tệp .sav, để lần sau mở nhanh và giữ nguyên khai báo biến.
Right after importing, save with File > Save As as a .sav file so it opens fast next time and keeps the variable settings.

1.3. Nhập số liệu và khai báo biến trong Data Editor – *1.3. Entering Data and Defining Variables in the Data Editor*

Data Editor là cửa sổ bảng tính của SPSS, nơi bạn gõ và sửa số liệu. Nó có hai tab ở góc dưới. Data View là bảng số liệu: mỗi hàng là một đối tượng khảo sát, mỗi cột là một biến. Variable View là bảng khai báo: mỗi hàng mô tả một biến. Hãy hình dung Data View như tờ phiếu điền, còn Variable View như phần chú thích ghi cột nào nói về cái gì.
The Data Editor is the spreadsheet-style window where you type and edit data in SPSS. It has two tabs at the bottom. Data View holds the data: each row is one case, each column is one variable. Variable View holds the setup: each row describes one variable. Think of Data View as the filled-in forms and Variable View as the notes saying what each column means.

Ví dụ xuyên suốt: khảo sát 3 hộ gia đình với sáu biến. maho (mã hộ, chữ), tuoi (tuổi chủ hộ), gioitinh (giới tính chủ hộ), songuoi (số người trong hộ), thunhap (thu nhập bình quân tháng, triệu đồng) và khuvuc (khu vực sống).
Running example: a survey of 3 households with six variables. maho (household code, text), tuoi (age of the head), gioitinh (head's gender), songuoi (household size), thunhap (monthly income, million VND) and khuvuc (area of residence).

Nhập số liệu số và chuỗi – *Entering numeric and string data*

Số liệu số là thứ bạn tính được, như tuổi hay thu nhập. Số liệu chuỗi là chữ, như mã hộ H001; SPSS không cộng trừ được chuỗi. Mã nhóm như 1 và 2 vẫn là số, vì sau này bạn gắn nhãn cho chúng.

Numeric data are values you can calculate with, such as age or income. String data are text, such as the code H001; SPSS cannot add or subtract strings. Group codes such as 1 and 2 are still numbers, because you will attach labels to them later.

1. Mở SPSS, chọn tab Data View, bấm vào ô đầu tiên của cột đầu.
   Open SPSS, select the Data View tab and click the first cell of the first column.
2. Gõ giá trị rồi nhấn Tab để sang ô bên phải, hoặc Enter để xuống ô dưới.
   Type a value, then press Tab to move right or Enter to move down.
3. Nhập hàng thứ nhất: H001, 45, 1, 4, 18.5, 1. Hàng hai: H002, 38, 2, 3, 12.0, 2. Hàng ba: H003, 52, 1, 5, 9.8, 2.
   Enter row one: H001, 45, 1, 4, 18.5, 1. Row two: H002, 38, 2, 3, 12.0, 2. Row three: H003, 52, 1, 5, 9.8, 2.

Khi bạn gõ ô đầu tiên, SPSS tự đặt tên tạm cho cột là VAR00001 và coi là số. Hãy sang bước đặt tên ngay để khỏi nhầm cột.

When you type the first cell, SPSS names the column VAR00001 and treats it as numeric. Go straight to naming so you do not mix up columns.

Khai báo trong Variable View – *Defining variables in Variable View*

Chuyển sang tab Variable View. Mỗi cột của bảng này là một thuộc tính của biến. Ba thuộc tính cần làm trước là Name, Label và Values.

Switch to the Variable View tab. Each column here is one property of a variable. Three properties come first: Name, Label and Values.

1. Cột Name: gõ tên biến vào từng hàng (maho, tuoi, gioitinh, songuoi, thunhap, khuvuc). Tên phải bắt đầu bằng chữ cái, không có dấu cách, không trùng nhau. Nên dùng chữ không dấu.
   Name column: type each variable name (maho, tuoi, gioitinh, songuoi, thunhap, khuvuc). A name must start with a letter, contain no spaces and be unique. Plain letters without accents are safer.
2. Cột Label: gõ mô tả đầy đủ, ví dụ "Thu nhập bình quân tháng của hộ (triệu đồng)". Nhãn dài được, tên biến thì ngắn. Nhãn này xuất hiện trong bảng kết quả và biểu đồ.
   Label column: type a full description, for example "Monthly household income (million VND)". Labels can be long while names stay short. The label appears in output tables and charts.
3. Cột Values: bấm ô của gioitinh rồi bấm nút ba chấm bên phải. Gõ 1 vào Value, "Nam" vào Label, bấm Add. Gõ 2 và "Nữ", bấm Add, rồi OK.
   Values column: click the gioitinh cell, then the button at its right. Type 1 in Value and "Male" in Label, click Add. Type 2 and "Female", click Add, then OK.
4. Làm tương tự cho khuvuc: 1 là "Thành thị", 2 là "Nông thôn".
   Repeat for khuvuc: 1 is "Urban", 2 is "Rural".

5. Chọn View > Value Labels để Data View hiện chữ "Nam", "Nữ" thay cho số. Dữ liệu bên dưới vẫn là 1 và 2.
   Choose View > Value Labels so Data View shows the words instead of numbers. The stored data remain 1 and 2.

Kiểu biến và định dạng – *Variable type and format*

Cột Type cho biết SPSS đọc biến là loại nào. Phổ biến nhất là Numeric (số) và String (chuỗi); còn có ngày tháng, tiền tệ và vài kiểu khác. Bấm ô Type rồi bấm nút bên phải để mở hộp thoại chọn.

The Type column shows how SPSS reads a variable. The most common types are Numeric and String; there are also date, currency and others. Click the Type cell, then the button at its right, to open the dialog.

1. Với maho: chọn String, đặt Width đủ dài cho mã dài nhất, ví dụ 8.
   For maho: choose String and set Width long enough for the longest code, for example 8.
2. Với thunhap: giữ Numeric, đặt Width 8 và Decimals 1 để hiện 18.5 thay vì 19. Với tuoi và songuoi đặt Decimals 0.
   For thunhap: keep Numeric, set Width 8 and Decimals 1 so it shows 18.5 rather than 19. For tuoi and songuoi set Decimals 0.

Decimals chỉ đổi cách hiển thị. Giá trị gõ vào vẫn được giữ nguyên để tính toán.

Decimals only changes how the value is displayed. The value you typed is kept for calculation.

VÍ DỤ / EXAMPLE – Sau khi khai báo, bảng Data View hiện hàng một là H001, 45, Nam, 4, 18.5, Thành thị. Khi chạy thống kê cho thunhap, kết quả ghi tên "Thu nhập bình quân tháng của hộ (triệu đồng)" thay vì "thunhap". Trung bình của ba hộ là (18.5 + 12.0 + 9.8) / 3 = 40.3 / 3 ≈ 13.4 triệu đồng.

After setup, row one in Data View reads H001, 45, Male, 4, 18.5, Urban. When you run statistics for thunhap, the output shows the full label instead of "thunhap". The mean of the three households is (18.5 + 12.0 + 9.8) / 3 = 40.3 / 3 ≈ 13.4 million VND.

LỖI HAY GẶP / COMMON MISTAKE – Đặt tên biến có dấu cách hoặc bắt đầu bằng số, SPSS báo lỗi và không nhận.

Naming a variable with a space or a leading digit makes SPSS show an error and reject the name.

LỖI HAY GẶP / COMMON MISTAKE – Đặt Width của biến chuỗi quá ngắn, mã dài bị cắt cụt và hai mã khác nhau có thể thành giống nhau.

Setting a string Width too short truncates long codes, and two different codes can end up identical.

LỖI HAY GẶP / COMMON MISTAKE – Gõ chữ "Nam" thẳng vào cột số. SPSS không nhận. Hãy mã hoá bằng 1, 2 rồi gắn nhãn giá trị.

Typing the word "Male" into a numeric column fails. Code it as 1 and 2, then add value labels.

LỖI HAY GẶP / COMMON MISTAKE – Quên nhãn giá trị, kết quả chỉ hiện 1 và 2, người đọc không biết nhóm nào là nhóm nào.

Forgetting value labels leaves only 1 and 2 in the output, and readers cannot tell which group is which.

> **PHẢI THUỘC / MEMORISE –** Name ngắn, không dấu cách, bắt đầu bằng chữ. Label là mô tả dài. Values gắn chữ cho mã số. Type và Decimals quyết định kiểu và cách hiển thị.
> *Name is short, with no spaces, and starts with a letter. Label is the long description. Values attaches words to numeric codes. Type and Decimals set the kind of data and how it is displayed.*

1.4. Giá trị khuyết và cách xử lý – *1.4. Handling Missing Values*

Giá trị khuyết là chỗ trong bảng dữ liệu mà đáng lẽ phải có một con số hoặc một chữ nhưng lại không có. Ví dụ phiếu khảo sát có câu "Thu nhập hằng tháng", người trả lời bỏ trống vì ngại nói. Ô đó chưa phải là 0 đồng, nó chỉ là thông tin chưa biết.

A missing value is a spot in the data table where a number or a word should be but is not. Say a survey asks about monthly income and the respondent leaves it blank because they would rather not say. That cell is not 0 dong. It is simply unknown.

Bỏ sót giá trị khuyết làm kết quả lệch vì phần mềm sẽ coi mã đặt tạm là dữ liệu thật. Giả sử 10 phiếu có thu nhập (triệu đồng): 8, 12, 15, 9, trống, 20, trống, 11, 14, 10. Tám giá trị có thật cộng lại bằng 99, trung bình là 99 / 8 = 12,375.

Missing values distort results when the software treats a placeholder code as real data. Suppose 10 forms show income (million dong): 8, 12, 15, 9, blank, 20, blank, 11, 14, 10. The eight real values add up to 99, so the mean is 99 / 8 = 12.375.

Nếu người nhập gõ 0 vào hai ô trống, SPSS tính 99 / 10 = 9,9, thấp hơn thực tế. Nếu người nhập gõ mã 999 cho "từ chối trả lời" mà không khai báo, tổng thành 99 + 1998 = 2097 và trung bình thành 209,7 triệu, một con số vô lý. Mô hình hồi quy sau này cũng bị kéo lệch theo cách tương tự.

If someone types 0 into the two blanks, SPSS computes 99 / 10 = 9.9, which is too low. If someone types the code 999 for "refused to answer" and never declares it, the sum becomes 99 + 1998 = 2097 and the mean becomes 209.7 million, which is absurd. A regression model later on gets pulled off in the same way.

Hai loại giá trị khuyết trong SPSS – *Two kinds of missing values in SPSS*

Loại thứ nhất là khuyết hệ thống (system-missing): ô số để trống, SPSS tự hiện dấu chấm "." và tự bỏ qua khi tính. Loại thứ hai là khuyết do người dùng khai báo (user-missing): bạn đã nhập một mã như 999 hay 99, rồi báo cho SPSS biết mã đó nghĩa là "không có thông tin".

The first kind is system-missing: a numeric cell is left empty, SPSS shows a dot "." and skips it in calculations. The second kind is user-missing: you typed a code such as 999 or 99 and told SPSS that this code means "no information".

Mã người dùng có ích khi bạn muốn phân biệt lý do. Ví dụ 999 là từ chối trả lời, 998 là không biết. Chọn mã nằm ngoài khoảng giá trị có thể có của biến, để không lẫn với dữ liệu thật.

User codes help when you want to tell the reasons apart, for example 999 for refused and 998 for don't know. Pick codes outside the range the variable can really take, so they never mix with real data.

Cách khai báo cho biến số – *Declaring missing values for a numeric variable*

1. Mở cửa sổ Data Editor, bấm thẻ Variable View ở góc dưới bên trái.
   Open the Data Editor and click the Variable View tab at the bottom left.
2. Tìm dòng của biến thu nhập, bấm vào ô thuộc cột Missing, rồi bấm nút ba chấm xuất hiện bên phải ô.
   Find the row of the income variable, click its cell in the Missing column, then click the three-dot button that appears on the right.
3. Chọn Discrete missing values, gõ 999 vào ô thứ nhất. Cần thêm mã khác thì gõ tiếp vào ô thứ hai hoặc thứ ba, tối đa ba mã.
   Choose Discrete missing values and type 999 in the first box. For another code, use the second or third box, up to three codes.
4. Muốn khai báo cả một khoảng, chọn Range plus one optional discrete missing value, nhập cận dưới và cận trên, ví dụ từ 990 đến 999.
   To declare a whole range, choose Range plus one optional discrete missing value and enter the lower and upper limits, for example 990 to 999.
5. Bấm OK. Chạy Analyze > Descriptive Statistics > Frequencies và đọc dòng Valid (hợp lệ) với Missing (khuyết) để kiểm tra số ca đã khớp.
   Click OK. Run Analyze > Descriptive Statistics > Frequencies and read the Valid and Missing rows to check that the counts match.

Cách khai báo cho biến chuỗi – *Declaring missing values for a string variable*

Biến chuỗi là biến chứa chữ, như nghề nghiệp hay tên tỉnh. Ô chuỗi để trống không tự động bị coi là khuyết hệ thống, SPSS xem đó là một giá trị rỗng như mọi giá trị khác. Vì vậy bạn phải khai báo.

A string variable holds text, such as occupation or province. An empty string cell is not treated as system-missing automatically. SPSS sees it as one more value, an empty one. So you must declare it.

Làm giống các bước trên nhưng chỉ dùng được Discrete missing values, không có khoảng. Gõ đúng mã như trong dữ liệu, ví dụ KB cho "không biết", và nhớ rằng chữ hoa chữ thường được xét khác nhau. Nếu muốn coi ô trống là khuyết, gõ một khoảng trắng vào ô khai báo.

Follow the same steps, but only Discrete missing values is available, with no range. Type the code exactly as it appears in the data, for example KB for "don't know", and remember that upper and lower case count as different. To treat blank cells as missing, type a single space in the declaration box.

> **VÍ DỤ / EXAMPLE –** Bảng 10 phiếu ở trên, hai ô thu nhập để trống. Nếu gõ 999 vào hai ô đó rồi khai báo 999 là giá trị khuyết, Frequencies báo Valid = 8, Missing = 2, và Descriptives cho trung bình 12,375. Nếu quên khai báo, trung bình thành 209,7. Chỉ cần nhìn con số đã thấy sai.
> *Take the 10 forms above, with two income cells empty. Type 999 in both and declare 999 as missing, and Frequencies reports Valid = 8, Missing = 2, while Descriptives gives a mean of 12.375. Skip the declaration and the mean becomes 209.7. The number alone shows something is wrong.*

Xử lý sau khi khai báo – ***What to do after declaring***

Mặc định, nhiều thủ tục chỉ dùng các ca đủ dữ liệu ở mọi biến đang phân tích (loại theo ca, listwise). Cách khác là mỗi cặp biến dùng mọi ca có đủ hai giá trị đó (loại theo cặp, pairwise). Hai cách có thể cho cỡ mẫu khác nhau, nên luôn ghi cỡ mẫu thực dùng vào báo cáo.

By default, many procedures use only the cases that have data on every variable in the analysis (listwise deletion). The alternative is that each pair of variables uses all cases having both values (pairwise deletion). The two can give different sample sizes, so always report the sample size actually used.

SPSS cũng có lệnh Transform > Replace Missing Values để điền thay thế, như dùng trung bình của chuỗi. Cách này làm giảm độ phân tán thật của dữ liệu, nên chỉ dùng khi bạn giải thích được lý do. Khi tỷ lệ khuyết nhỏ và ngẫu nhiên, loại bỏ ca thường là lựa chọn an toàn.

SPSS also has Transform > Replace Missing Values to fill gaps, for example with the series mean. This shrinks the real spread of the data, so use it only when you can justify it. When missing cases are few and random, dropping them is usually the safe choice.

> **LỖI HAY GẶP / COMMON MISTAKE –** Gõ 0 vào ô trống thu nhập, hoặc để mã 999 mà không khai báo, hoặc khai báo ở biến này nhưng quên biến khác dùng cùng mã. Một lỗi nữa là nhầm khi người từ chối trả lời có đặc điểm khác hẳn nhóm còn lại, ví dụ nhóm thu nhập cao hay bỏ trống, thì việc loại họ làm lệch kết quả dù đã khai báo đúng.
> *Typing 0 into blank income cells, leaving 999 undeclared, or declaring it for one variable and forgetting another that uses the same code. One more trap: if people who refuse to answer differ systematically from the rest, for example high earners who skip the income question, dropping them biases the results even when the declaration is correct.*

> **PHẢI THUỘC / MEMORISE –** Ô trống số là khuyết hệ thống, mã như 999 phải tự khai báo ở cột Missing. Biến chuỗi chỉ khai báo được từng giá trị rời, không khai báo khoảng. Luôn kiểm tra dòng Valid và Missing trong Frequencies sau khi khai báo.
> *A blank numeric cell is system-missing, while a code like 999 must be declared in the Missing column. String variables accept only individual values, not ranges. Always check the Valid and Missing rows in Frequencies after declaring.*

1.5. Biến đổi dữ liệu: tính biến mới, mã hoá lại, chọn ca – *1.5. Transforming Data: Computing, Recoding and Selecting Cases*

Dữ liệu thô hiếm khi dùng được ngay. Bạn thường phải tạo thêm biến, gom nhóm lại hoặc chỉ giữ một phần quan sát. SPSS có bốn công cụ cho việc này: Compute (tính biến mới), Recode (mã hoá lại), Select Cases (chọn ca) và Split File (chia tệp). Nhóm lệnh tính và mã hoá nằm trong menu Transform, nhóm lệnh chọn và chia nằm trong menu Data.

Raw data is rarely ready to use. You often need to add variables, collapse values into groups, or keep only part of the observations. SPSS has four tools for this: Compute, Recode, Select Cases

and Split File. Compute and Recode live in the Transform menu; Select Cases and Split File live in the Data menu.

Trước khi bắt đầu, hãy lưu một bản sao tệp dữ liệu. Phần lớn thao tác ở đây ghi thẳng vào tệp đang mở, và bạn khó quay lại nếu lưu đè.
Before you start, save a copy of your data file. Most of these actions write straight into the open file, and it is hard to go back once you save over it.

Tạo biến mới bằng phép tính – *Creating a New Variable with a Calculation*

Compute giống việc thêm một cột công thức trong Excel. Bạn đặt tên cột mới, rồi viết công thức dùng các biến có sẵn. Cách này hay dùng để lấy logarit, bình phương, hay tính tỷ số giữa hai biến. Trong kinh tế lượng, logarit thu nhập giúp kéo gần các giá trị rất lớn và rất nhỏ, và cho phép đọc hệ số theo phần trăm.
Compute works like adding a formula column in Excel. You name the new column, then write a formula using existing variables. It is used for logarithms, squares, or ratios of two variables. In econometrics, the log of income pulls very large and very small values closer together and lets you read coefficients in percentage terms.

1. Vào Transform, chọn Compute Variable.
   Open Transform, then choose Compute Variable.
2. Ở ô Target Variable, gõ tên biến mới, ví dụ `ln_thunhap`. Tên không có dấu cách.
   In the Target Variable box, type the new name, for example `ln_thunhap`. No spaces in the name.
3. Ở ô Numeric Expression, gõ LN(thunhap). Bạn cũng có thể chọn hàm LN trong danh sách Function group rồi bấm đúp.
   In the Numeric Expression box, type LN(thunhap). You can also pick LN from the Function group list and double-click it.
4. Bấm OK. Một cột mới xuất hiện ở cuối Data View.
   Click OK. A new column appears at the end of Data View.

VÍ DỤ / EXAMPLE – Bốn người có thu nhập 5, 10, 20 và 50 triệu đồng một tháng. LN cho lần lượt khoảng 1,609; 2,303; 2,996 và 3,912. Thu nhập tăng gấp mười lần (từ 5 lên 50) mà logarit chỉ tăng từ 1,609 lên 3,912. Hiệu 3,912 trừ 1,609 bằng 2,303, đúng bằng LN(10).
Four people earn 5, 10, 20 and 50 million dong a month. LN gives about 1.609, 2.303, 2.996 and 3.912. Income grows tenfold (5 to 50), yet the log rises only from 1.609 to 3.912. The difference 3.912 minus 1.609 is 2.303, which equals LN(10).

LỖI HAY GẶP / COMMON MISTAKE – Logarit chỉ tính được cho số dương. Nếu thu nhập bằng 0 hoặc âm, SPSS để ô trống (system-missing) mà không báo lỗi lớn. Hãy đếm số ca có giá trị trong biến mới và so với biến gốc.
Logs exist only for positive numbers. If income is 0 or negative, SPSS leaves the cell blank (system-missing) without a loud warning. Count the valid cases in the new variable and compare with the original.

Mã hoá lại nhóm tuổi – *Recoding Age into Groups*

Recode biến một biến số thành các nhóm. Nên chọn Recode into Different Variables để giữ nguyên biến tuổi gốc và tạo biến nhóm riêng. Lựa chọn Into Same Variables sẽ ghi đè lên biến gốc.

Recode turns a numeric variable into groups. Choose Recode into Different Variables so the original age stays intact and the groups go into a separate variable. Into Same Variables overwrites the original.

1. Vào Transform, chọn Recode into Different Variables. Chuyển biến tuoi sang ô bên phải.
   Open Transform and choose Recode into Different Variables. Move tuoi to the right-hand box.
2. Ở Output Variable, gõ tên nhomtuoi, bấm Change.
   Under Output Variable, type nhomtuoi and click Change.
3. Bấm Old and New Values. Chọn Range, LOWEST through value, nhập 29; ở New Value gõ 1, bấm Add.
   Click Old and New Values. Choose Range, LOWEST through value, enter 29; under New Value type 1 and click Add.
4. Chọn Range, nhập 30 đến 44, giá trị mới 2, bấm Add. Chọn Range, value through HIGHEST, nhập 45, giá trị mới 3, bấm Add.
   Choose Range, enter 30 through 44, new value 2, click Add. Choose Range, value through HIGHEST, enter 45, new value 3, click Add.
5. Bấm Continue, rồi OK. Sang Variable View, khai báo nhãn giá trị: 1 là dưới 30, 2 là 30–44, 3 là từ 45.
   Click Continue, then OK. In Variable View, add value labels: 1 is under 30, 2 is 30–44, 3 is 45 and over.

VÍ DỤ / EXAMPLE – Tuổi 24, 30, 44, 45, 61 được mã thành 1, 2, 2, 3, 3. Người 44 tuổi thuộc nhóm 2, người 45 tuổi sang nhóm 3.
Ages 24, 30, 44, 45 and 61 become 1, 2, 2, 3, 3. The 44-year-old falls in group 2, and the 45-year-old moves to group 3.

LỖI HAY GẶP / COMMON MISTAKE – Khoảng bị hở. Nếu tuổi có số lẻ như 29,5 mà bạn chỉ khai 29 và 30 thì ca đó rơi ra ngoài. Dùng LOWEST và HIGHEST ở hai đầu, và kiểm tra bằng bảng tần số xem còn ca nào trống nhóm.
Gaps between ranges. If age has decimals like 29.5 and you only declared 29 and 30, that case falls through. Use LOWEST and HIGHEST at the ends and run a frequency table to check for cases left without a group.

Lọc quan sát theo điều kiện – *Selecting Cases by Condition*

Select Cases cho SPSS bỏ qua những ca không thoả điều kiện khi phân tích. Giả sử giới tính mã 1 là nam, 2 là nữ, và bạn muốn chỉ phân tích nữ.

Select Cases makes SPSS ignore the cases that fail a condition during analysis. Suppose gender is coded 1 for male and 2 for female, and you want to analyse women only.

1. Vào Data, chọn Select Cases.
   Open Data, then choose Select Cases.
2. Chọn If condition is satisfied, bấm If, gõ gioitinh = 2, bấm Continue.
   Choose If condition is satisfied, click If, type gioitinh = 2, click Continue.
3. Ở mục Output, giữ Filter out unselected cases, bấm OK. Các dòng bị loại bị gạch chéo ở số thứ tự.
   Under Output, keep Filter out unselected cases and click OK. Excluded rows show a diagonal strike on the row number.
4. Chạy phân tích cần làm, ví dụ thống kê mô tả `ln_thunhap`. Xong việc, mở lại Select Cases, chọn All cases, bấm OK.
   Run the analysis you need, for example descriptives of `ln_thunhap`. When done, reopen Select Cases, choose All cases and click OK.

> **LỖI HAY GẶP / COMMON MISTAKE –** Quên chọn lại All cases là lỗi phổ biến nhất. Mọi kết quả sau đó chỉ tính cho nữ mà bạn không hay biết. Nếu chọn Delete unselected cases rồi lưu, các ca bị xoá mất vĩnh viễn.
> *Forgetting to switch back to All cases is the most common mistake. Every later result covers women only and you may not notice. If you choose Delete unselected cases and then save, those cases are gone for good.*

Chia tệp theo nhóm – *Splitting the File by Group*

Split File không loại ca nào. Nó chạy cùng một phân tích riêng cho từng nhóm và đặt kết quả cạnh nhau, tiện để so nam với nữ.
Split File removes no cases. It runs the same analysis separately for each group and places the results side by side, which makes comparing men and women easy.

1. Vào Data, chọn Split File.
   Open Data, then choose Split File.
2. Chọn Compare groups, đưa gioitinh vào ô Groups Based on, bấm OK.
   Choose Compare groups, move gioitinh into Groups Based on, and click OK.
3. Chạy phân tích. Kết quả hiện hai phần, một cho mỗi giới. Khi xong, quay lại Split File và chọn Analyze all cases, do not create groups.
   Run the analysis. The output shows two parts, one per gender. When finished, return to Split File and choose Analyze all cases, do not create groups.

> **PHẢI THUỘC / MEMORISE –** Chọn ca (Select Cases) loại bỏ nhóm khỏi phân tích; chia tệp (Split File) giữ mọi nhóm và phân tích riêng từng nhóm. Cả hai đều phải tắt thủ công khi dùng xong.
> *Select Cases removes groups from the analysis; Split File keeps every group and analyses each one separately. Both must be switched off manually when you finish.*

1.6. Các loại thang đo và loại dữ liệu – *1.6. Levels of Measurement and Types of Data*

Mỗi biến trong bộ dữ liệu là một cách ghi lại thế giới bằng con số hoặc nhãn. Thang đo cho biết con số đó mang được bao nhiêu thông tin. Biết thang đo thì bạn biết được phép

tính nào có nghĩa và kỹ thuật nào dùng được. Cùng lúc, bạn cần biết dữ liệu được thu theo kiểu nào: một thời điểm, một chuỗi thời gian hay cả hai.

P: Every variable records something about the world as a number or a label. Its level of measurement tells you how much information those values carry, and so which calculations make sense and which techniques you can use. You also need to know how the data were collected: at one point in time, over time, or both.

Bốn thang đo – *Four levels of measurement*

Hãy đi từ thang yếu nhất đến mạnh nhất. Mỗi thang mới giữ mọi khả năng của thang trước và thêm một khả năng mới.

P: Go from the weakest level to the strongest. Each level keeps everything the previous one allows and adds one more ability.

- Thang định danh (nominal): con số hay chữ chỉ là nhãn để gọi tên nhóm. Ví dụ: ngành nghề, tỉnh thường trú, giới tính. Bạn chỉ đếm và so sánh được "giống hay khác". Mã 1 = nam, 2 = nữ không có nghĩa nữ "lớn hơn" nam.
  GACH: Nominal scale: numbers or words are only labels for groups, such as occupation, home province or gender. You can count and check "same or different". The codes 1 = male and 2 = female do not mean female is "larger".
- Thang thứ bậc (ordinal): các nhóm xếp được theo thứ tự, nhưng khoảng cách giữa các bậc không đo được. Ví dụ: học lực giỏi, khá, trung bình; mức hài lòng từ 1 đến 5. Bạn biết "cao hơn", không biết cao hơn bao nhiêu.
  GACH: Ordinal scale: categories can be ranked, but the gaps between ranks are not measured. Examples: grades of excellent, good, average, or a satisfaction score from 1 to 5. You know "higher", not "higher by how much".
- Thang khoảng (interval): khoảng cách giữa các giá trị có nghĩa, nhưng điểm 0 chỉ là quy ước. Ví dụ: nhiệt độ theo độ C. Chênh 10°C giữa 20°C và 30°C bằng chênh giữa 30°C và 40°C. Nhưng 20°C không "nóng gấp đôi" 10°C, vì 0°C không phải là hết nhiệt.
  GACH: Interval scale: distances between values are meaningful, but zero is only a convention. Example: temperature in °C. The gap from 20°C to 30°C equals the gap from 30°C to 40°C. Yet 20°C is not "twice as hot" as 10°C, because 0°C does not mean no heat.
- Thang tỷ lệ (ratio): có khoảng cách có nghĩa và điểm 0 thật. Ví dụ: thu nhập, chiều cao, số năm đi học, giá cổ phiếu. Thu nhập 20 triệu gấp 2,5 lần thu nhập 8 triệu, vì 20 : 8 = 2,5.
  GACH: Ratio scale: meaningful distances plus a true zero. Examples: income, height, years of schooling, share price. An income of 20 million is 2.5 times 8 million, because 20 / 8 = 2.5.

Bảng sau đây gom lại để tra nhanh:

P: Use this table for quick lookup:

Thang đo *Level*	**Phép làm được** *What you can do*	**Số đo mô tả hợp lý** *Sensible summary*	**Kỹ thuật thường dùng** *Typical techniques*
Định danh *Nominal*	Đếm, so sánh giống hay khác *Count, check same or different*	Tần số, tỷ lệ, mốt *Frequencies, proportions, mode*	Bảng chéo, chi bình phương, logit cho biến phụ thuộc nhị phân *Crosstabs, chi-square, logit for a binary outcome*
Thứ bậc *Ordinal*	Xếp thứ tự *Rank*	Trung vị, tần số *Median, frequencies*	Kiểm định phi tham số, tương quan hạng *Nonparametric tests, rank correlation*
Khoảng *Interval*	Cộng, trừ khoảng cách *Add, subtract distances*	Trung bình, độ lệch chuẩn *Mean, standard deviation*	Kiểm định t, tương quan, hồi quy *t-tests, correlation, regression*
Tỷ lệ *Ratio*	Thêm phép chia, so sánh gấp bao nhiêu lần *Also divide, compare by "times as much"*	Như thang khoảng, thêm tỷ số và logarit *As interval, plus ratios and logarithms*	Hồi quy, ANOVA, mô hình dạng logarit *Regression, ANOVA, log-form models*

Xác định thang đo của một biến – *Finding the level of a variable*

1. Hỏi: các giá trị chỉ là nhãn nhóm? Nếu đúng, đó là định danh.
   BUOC: Ask: are the values just group labels? If yes, it is nominal.
2. Nếu không, hỏi: xếp được thứ tự nhưng khoảng cách giữa hai bậc liền nhau không đều hoặc không đo được? Nếu đúng, đó là thứ bậc.
   BUOC: If not, ask: can they be ranked, but are the gaps between neighbouring ranks unequal or unmeasured? If yes, it is ordinal.
3. Nếu khoảng cách có nghĩa, hỏi: giá trị 0 có nghĩa là "không có gì" không? Có thì là tỷ lệ, không thì là khoảng.
   BUOC: If gaps are meaningful, ask: does 0 mean "none at all"? If so it is ratio, otherwise interval.
4. Trong SPSS, mở Variable View, vào cột Measure của biến và chọn Nominal, Ordinal hoặc Scale. Scale gộp cả khoảng và tỷ lệ.
   BUOC: In SPSS, open Variable View, go to the Measure column of the variable and pick Nominal, Ordinal or Scale. Scale covers both interval and ratio.

> **VÍ DỤ / EXAMPLE –** Biến "số con trong hộ" nhận giá trị 0, 1, 2, 3. Khoảng cách đều nhau và 0 nghĩa là không có con, nên đây là thang tỷ lệ: hộ 4 con gấp đôi hộ 2 con. Biến "hạng nhà trường" (1 = top đầu, 2 = top giữa, 3 = top cuối) là thứ bậc, nên tính trung bình hạng chỉ là ước chừng thô.
> *VIDU: The variable "number of children in a household" takes values 0, 1, 2, 3. Gaps are equal and 0 means no children, so it is ratio: a household with 4 children has twice as many as one with 2. The variable "school tier" (1 = top, 2 = middle, 3 = bottom) is ordinal, so an average tier is only a rough guide.*

Ba kiểu cấu trúc dữ liệu – ***Three data structures***

Cắt ngang (cross-section) là ảnh chụp nhiều đối tượng tại một thời điểm. Ví dụ: thu nhập của 50 hộ gia đình trong năm 2024, mỗi hộ một dòng. Chuỗi thời gian (time series) theo dõi một đối tượng qua nhiều kỳ. Ví dụ: doanh thu của một cửa hàng trong 36 tháng liên tiếp. Dữ liệu bảng (panel) theo dõi nhiều đối tượng qua nhiều kỳ. Ví dụ: 50 hộ đó được khảo sát lại mỗi năm từ 2020 đến 2024, tạo ra 250 quan sát.

P: Cross-section data is a snapshot of many units at one moment, such as the income of 50 households in 2024, one row per household. A time series follows one unit over many periods, such as a shop's revenue over 36 consecutive months. Panel data follows many units over many periods, such as the same 50 households surveyed each year from 2020 to 2024, giving 250 observations.

Cách chọn kỹ thuật phụ thuộc vào cấu trúc này. Dữ liệu cắt ngang thường lo vấn đề phương sai sai số thay đổi. Chuỗi thời gian phải kiểm tra tính dừng và tự tương quan, vì giá trị hôm nay thường dính với hôm qua. Dữ liệu bảng cần mô hình riêng, vì cùng một hộ lặp lại nhiều năm thì các quan sát không độc lập.

P: The structure drives the choice of method. Cross-section data usually raises concerns about changing error variance. Time series need checks for stationarity and autocorrelation, since today's value tends to stick to yesterday's. Panel data needs its own models, because repeated rows from the same household are not independent.

LỖI HAY GẶP / COMMON MISTAKE – Coi điểm Likert 1–5 là thang tỷ lệ rồi nói "hài lòng 4 gấp đôi hài lòng 2". Strictly, đó là thứ bậc. Nhiều nghiên cứu vẫn tính trung bình và chạy hồi quy trên thang này, nhưng bạn nên nói rõ giả định đó. Một lỗi khác: để mã số điện thoại hay mã tỉnh ở dạng Scale rồi tính trung bình, kết quả vô nghĩa.

LOI: Treating a 1–5 Likert score as ratio and saying "satisfaction 4 is twice satisfaction 2". Strictly it is ordinal. Many studies still average it and run regressions on it, but you should state that assumption openly. Another slip: leaving phone numbers or province codes as Scale and averaging them gives meaningless results.

LỖI HAY GẶP / COMMON MISTAKE – Gộp dữ liệu của nhiều năm vào một bảng rồi coi là cắt ngang. Nếu cùng một đối tượng xuất hiện nhiều lần, đó là dữ liệu bảng.

LOI: Stacking several years into one table and calling it cross-section. If the same unit appears more than once, it is panel data.

PHẢI THUỘC / MEMORISE – Định danh chỉ có nhãn; thứ bậc có thứ tự; khoảng có khoảng cách nhưng 0 là quy ước; tỷ lệ có 0 thật. Cắt ngang là nhiều đối tượng, một thời điểm; chuỗi thời gian là một đối tượng, nhiều kỳ; bảng là nhiều đối tượng, nhiều kỳ.

THUOC: Nominal has labels only; ordinal adds order; interval adds distances but zero is a convention; ratio has a true zero. Cross-section is many units at one time; a time series is one unit over many periods; a panel is many units over many periods.

1.7. Kinh tế lượng là gì và các bước xây dựng một mô hình – *1.7. What Econometrics Is and the Steps of Building a Model*

Kinh tế lượng là việc dùng số liệu thật và công cụ thống kê để đo và kiểm tra các quan hệ mà lý thuyết kinh tế nêu ra. Ví dụ: chủ quán trà sữa tin rằng giảm giá thì bán được nhiều hơn. Kinh tế lượng trả lời câu hỏi cụ thể hơn: giảm 1.000 đồng một ly thì doanh số tăng khoảng bao nhiêu ly, và con số đó tin được đến đâu.

Econometrics uses real data and statistical tools to measure and test the relationships that economic theory proposes. Say the owner of a bubble tea shop believes a lower price sells more cups. Econometrics asks something sharper: if the price drops by 1,000 dong per cup, how many more cups are sold, and how far can you trust that number?

Kinh tế và tài chính dùng chung bộ công cụ, nhưng khác ở loại dữ liệu và câu hỏi. Kinh tế thường có số quan sát vừa phải, như khảo sát hộ gia đình hoặc số liệu theo năm. Tài chính thường có dữ liệu dày, như giá cổ phiếu theo ngày hoặc theo giờ. Dữ liệu tài chính nhiễu nhiều, và mức dao động của nó đổi theo thời gian.

Economics and finance share the same toolbox, but differ in the data and the questions. Economics usually works with a moderate number of observations, such as household surveys or yearly figures. Finance usually has dense data, such as stock prices by day or by hour. Financial data is noisy, and how much it swings changes over time.

Tiêu chí *Criterion*	**Kinh tế** *Economics*	**Tài chính** *Finance*
Dữ liệu thường gặp *Typical data*	Khảo sát hộ, số liệu vĩ mô theo năm hoặc quý, số quan sát không nhiều. *Household surveys and macro figures by year or quarter, with relatively few observations.*	Giá và lợi suất theo ngày hoặc theo giờ, số quan sát lớn. *Prices and returns by day or hour, with many observations.*
Câu hỏi hay đặt ra *Typical question*	Một yếu tố tác động lên kết quả bao nhiêu, như học thêm một năm thì lương tăng bao nhiêu. *How much one factor affects an outcome, such as the pay gain from one more year of schooling.*	Lợi suất, rủi ro và giá trị trong tương lai sẽ biến động ra sao. *How returns, risk and future values will move.*
Khó khăn điển hình *Typical difficulty*	Sai số khi đo, biến bị bỏ sót, hai biến cùng ảnh hưởng lẫn nhau. *Measurement error, omitted variables, and two variables affecting each other.*	Độ dao động đổi theo thời gian, có những ngày giá nhảy rất xa. *Volatility that changes over time, with some days of very large price jumps.*

Ở học phần này, bạn sẽ thấy cả hai nhánh. Phần đầu nghiêng về kinh tế (hồi quy, biến nội sinh). Phần sau nghiêng về tài chính (chuỗi thời gian, mô hình biến động).

In this course you will meet both branches. The early chapters lean toward economics (regression, endogenous variables). The later ones lean toward finance (time series, volatility models).

Chuỗi bước xây dựng một mô hình – *The steps of building a model*

1. Đặt câu hỏi và giả thuyết. Lấy từ lý thuyết hoặc nghiên cứu trước, rồi viết rõ dấu mong đợi. Ví dụ: giá giảm thì lượng bán tăng, nên hệ số của giá phải âm.
   State the question and hypothesis. Take it from theory or earlier studies, and write down the expected sign. For example, a lower price means higher sales, so the price coefficient should be negative.
2. Viết mô hình. Chọn biến phụ thuộc (kết quả cần giải thích), các biến giải thích, dạng hàm và thành phần sai số. Bước này nên làm trên giấy trước khi mở SPSS.
   Write the model. Choose the dependent variable (the outcome to explain), the explanatory variables, the functional form and the error term. Do this on paper before opening SPSS.
3. Thu thập và làm sạch dữ liệu. Xác định loại dữ liệu (cắt ngang, chuỗi thời gian hay bảng), nguồn, đơn vị đo. Kiểm tra giá trị khuyết và số liệu bất thường.
   Collect and clean the data. Identify the data type (cross-section, time series or panel), the source and the units. Check for missing and odd values.
4. Ước lượng. Chọn phương pháp, thường bắt đầu bằng bình phương nhỏ nhất (OLS, tìm đường thẳng sát các điểm dữ liệu nhất). Trong SPSS: Analyze, Regression, Linear.
   Estimate. Pick a method, usually starting with ordinary least squares (OLS, which finds the line closest to the data points). In SPSS: Analyze, Regression, Linear.
5. Kiểm định. Xem hệ số có khác 0 một cách có ý nghĩa thống kê không, mô hình giải thích được bao nhiêu phần biến thiên, và các giả định của phương pháp có đứng vững không.
   Test. Check whether the coefficients differ from zero in a statistically meaningful way, how much of the variation the model explains, and whether the method's assumptions hold.
6. Sửa và ước lượng lại nếu cần. Giả định bị vi phạm hoặc dấu hệ số ngược lý thuyết thì quay về bước 2: thêm biến, đổi dạng hàm hoặc đổi phương pháp.
   Revise and re-estimate if needed. If an assumption fails or a sign contradicts theory, go back to step 2: add a variable, change the functional form or change the method.
7. Diễn giải và sử dụng. Nói bằng đơn vị của bài toán, nêu phạm vi áp dụng, rồi dự báo hoặc đưa ra khuyến nghị.
   Interpret and use. Speak in the units of the problem, state the range where the result applies, then forecast or make recommendations.

VÍ DỤ / EXAMPLE – Giả thuyết: thu nhập hộ tăng thì chi tiêu tăng, hệ số nằm giữa 0 và 1. Mô hình: Chi = $\beta_0 + \beta_1 \cdot$Thunhap + u, đơn vị triệu đồng một tháng. Dữ liệu: 200 hộ. Giả sử OLS cho $\beta_0 = 1{,}8$ và $\beta_1 = 0{,}62$ với p nhỏ hơn 0,001. Dấu dương và giá trị 0,62 nằm giữa 0 và 1, khớp giả thuyết. Diễn giải: hộ có thu nhập cao hơn 1 triệu thì chi tiêu cao hơn khoảng 0,62 triệu. Dự báo cho hộ thu nhập 20 triệu: $1{,}8 + 0{,}62 \times 20 = 1{,}8 + 12{,}4 = 14{,}2$ triệu đồng. Con số này chỉ đáng tin trong khoảng thu nhập mà mẫu có.
Hypothesis: higher household income means higher spending, with a coefficient between 0 and 1. Model: Spend = $\beta_0 + \beta_1 \cdot$Income + u, in million dong per month. Data: 200 households. Suppose OLS gives $\beta_0 = 1.8$ and $\beta_1 = 0.62$ with p below 0.001. The positive sign and a value between 0 and 1 match the hypothesis. Interpretation: a household earning 1 million more spends about 0.62 million more.

Forecast for a household earning 20 million: 1.8 + 0.62 × 20 = 1.8 + 12.4 = 14.2 million dong. This figure is only reliable within the income range the sample covers.

LỖI HAY GẶP / COMMON MISTAKE – Chạy nhiều mô hình rồi chỉ giữ cái cho kết quả đẹp. Giả thuyết phải có trước khi nhìn kết quả, nếu không p-value mất giá trị.
Running many models and keeping only the one with a pretty result. The hypothesis must come before you see the output, or the p-value loses its meaning.

LỖI HAY GẶP / COMMON MISTAKE – Đọc tương quan thành nhân quả. Hệ số dương cho thấy hai biến đi cùng nhau trong mẫu, chưa chứng minh biến này gây ra biến kia.
Reading correlation as causation. A positive coefficient shows two variables move together in the sample, not that one causes the other.

LỖI HAY GẶP / COMMON MISTAKE – Dừng ngay sau khi ước lượng. Chưa kiểm tra giả định thì sai số chuẩn và p-value có thể sai.
Stopping right after estimation. Without checking the assumptions, standard errors and p-values may be wrong.

PHẢI THUỘC / MEMORISE – Giả thuyết, mô hình, dữ liệu, ước lượng, kiểm định, sửa mô hình, diễn giải. Kinh tế hay hỏi tác động của một yếu tố, tài chính hay hỏi lợi suất, rủi ro và dự báo.
Hypothesis, model, data, estimation, testing, revision, interpretation. Economics often asks about the effect of one factor; finance often asks about returns, risk and forecasts.

1.8. Tỷ suất sinh lợi và cách dùng trong mô hình tài chính – *1.8. Returns and How They Enter Financial Models*

Tỷ suất sinh lợi (lợi suất) là phần trăm lãi hoặc lỗ bạn có được khi giữ một tài sản trong một khoảng thời gian. Mua cổ phiếu giá 50.000 đồng, hôm sau thành 55.000 đồng thì lãi 10%. Trong kinh tế lượng tài chính, người ta ít khi phân tích giá, mà phân tích lợi suất.
P: A return is the percentage gained or lost from holding an asset over a period. Buy a share at 50,000 dong and see it reach 55,000 dong the next day, and you have earned 10%. In financial econometrics, analysts rarely model prices directly. They model returns.

Hai cách tính lợi suất – *TIEUDE: Two Ways to Compute a Return*

Cách thứ nhất là lợi suất đơn (còn gọi lợi suất số học): lấy chênh lệch giá chia cho giá kỳ trước. Cách thứ hai là lợi suất logarit: lấy logarit tự nhiên (ln) của tỷ số giá hai kỳ liên tiếp. Với biến động nhỏ, hai số gần bằng nhau.
P: The first is the simple return: the price change divided by the previous price. The second is the log return: the natural logarithm (ln) of the ratio of two consecutive prices. For small moves, the two numbers are close.

```
Lợi suất đơn: R(t) = (P(t) - P(t-1)) / P(t-1)
Lợi suất logarit: r(t) = ln( P(t) / P(t-1) )
```

Trong đó P(t) là giá đóng cửa ngày t, P(t−1) là giá đóng cửa phiên liền trước. Nếu cổ phiếu có trả cổ tức D trong kỳ, lợi suất đơn tính tổng là (P(t) + D − P(t−1)) / P(t−1).

P: Here P(t) is the closing price on day t and P(t−1) is the close of the previous session. If the stock pays a dividend D in the period, the total simple return is (P(t) + D − P(t−1)) / P(t−1).

Vì sao đổi chuỗi giá sang chuỗi lợi suất – *TIEUDE: Why Convert Prices to Returns*

- Lợi suất không có đơn vị, nên so sánh được giữa cổ phiếu giá 10.000 đồng và cổ phiếu giá 200.000 đồng.
  GACH: A return has no unit, so you can compare a 10,000-dong stock with a 200,000-dong stock.
- Chuỗi giá thường đi lên hoặc xuống theo xu hướng dài, trung bình và độ dao động đổi theo thời gian. Chuỗi lợi suất dao động quanh một mức khá ổn định, gọi là chuỗi dừng. Các mô hình ở chương sau cần tính dừng để cho kết quả đáng tin.
  GACH: Price series usually drift up or down over long stretches, so their mean and spread change over time. Return series fluctuate around a fairly steady level, which is called a stationary series. The models in later chapters need stationarity to give reliable results.
- Lợi suất logarit cộng được theo thời gian: tổng lợi suất log của nhiều ngày bằng ln(giá cuối / giá đầu). Lợi suất đơn thì phải nhân dồn.
  GACH: Log returns add up over time: the sum of daily log returns equals ln(final price / starting price). Simple returns must be compounded by multiplication.
- Lợi suất đơn cộng được giữa các tài sản trong danh mục: lợi suất danh mục bằng tổng các lợi suất đơn nhân tỷ trọng. Lợi suất log không có tính chất này.
  GACH: Simple returns add across assets in a portfolio: the portfolio return is the weighted sum of simple returns. Log returns do not have this property.

Tính trong SPSS – *TIEUDE: Computing in SPSS*

1. Mở dữ liệu, kiểm tra cột ngày đã sắp từ cũ đến mới. Nếu chưa, chọn Data, Sort Cases, rồi sắp theo cột ngày tăng dần.
   BUOC: Open the data and check that the date column runs from oldest to newest. If not, choose Data, Sort Cases, and sort by date in ascending order.
2. Mở cửa sổ Syntax (File, New, Syntax). Giả sử cột giá đóng cửa tên là gia.
   BUOC: Open a Syntax window (File, New, Syntax). Assume the closing price column is named gia.
3. Gõ các lệnh sau rồi bôi đen và bấm nút Run:
   BUOC: Type the following commands, select them, and click Run:

```
COMPUTE rdon = (gia - LAG(gia)) / LAG(gia).
COMPUTE rlog = LN(gia / LAG(gia)).
EXECUTE.
```

1. Mở Data View. Dòng đầu của hai cột mới để trống, vì ngày đầu chưa có giá hôm trước. Đó là bình thường.
   BUOC: Open Data View. The first row of both new columns is empty because the first day has no previous price. That is expected.

VÍ DỤ / EXAMPLE – Giá đóng cửa 4 phiên (nghìn đồng): 50; 55; 52,25; 54. Lợi suất đơn: (55−50)/50 = 0,10; (52,25−55)/55 = −0,05; (54−52,25)/52,25 = 0,0335. Lợi suất log: ln(1,10) = 0,0953; ln(0,95) = −0,0513; ln(54/52,25) = 0,0329. Tổng log = 0,0953 − 0,0513 + 0,0329 =

0,0770, đúng bằng ln(54/50) = ln(1,08). Cộng ba lợi suất đơn ra 0,0835, lệch so với mức tăng thật 8%. Phải nhân dồn: 1,10 × 0,95 × 1,0335 = 1,08.
VIDU: Closing prices over 4 sessions (thousand dong): 50; 55; 52.25; 54. Simple returns: (55−50)/50 = 0.10; (52.25−55)/55 = −0.05; (54−52.25)/52.25 = 0.0335. Log returns: ln(1.10) = 0.0953; ln(0.95) = −0.0513; ln(54/52.25) = 0.0329. The log sum is 0.0953 − 0.0513 + 0.0329 = 0.0770, which equals ln(54/50) = ln(1.08). Adding the three simple returns gives 0.0835, which misses the true 8% gain. You must compound: 1.10 × 0.95 × 1.0335 = 1.08.

VÍ DỤ / EXAMPLE – Bài luyện. Giá đóng cửa: 20; 22; 21,12; 22. Tính ba lợi suất đơn và ba lợi suất log, rồi kiểm tra tổng log. Đáp án: lợi suất đơn 0,10; −0,04; 0,0417. Lợi suất log 0,0953; −0,0408; 0,0408. Tổng log = 0,0953, bằng ln(22/20) = ln(1,1).
VIDU: Practice. Closing prices: 20; 22; 21.12; 22. Compute the three simple returns and three log returns, then check the log sum. Answer: simple returns 0.10; −0.04; 0.0417. Log returns 0.0953; −0.0408; 0.0408. The log sum is 0.0953, equal to ln(22/20) = ln(1.1).

LỖI HAY GẶP / COMMON MISTAKE – Sắp dữ liệu từ mới đến cũ làm LAG lấy nhầm giá, ra lợi suất đảo dấu. Dùng giá chưa điều chỉnh khi cổ phiếu chia tách hoặc trả cổ phiếu thưởng tạo ra cú nhảy giá giả. Đưa giá bằng 0 hoặc âm vào LN gây giá trị khuyết. Lẫn đơn vị cũng hay gặp: 0,10 là 10%, không phải 0,10%.
LOI: Sorting data from newest to oldest makes LAG pick the wrong price and flips the sign. Using unadjusted prices around a stock split or bonus issue creates a fake price jump. Feeding a zero or negative price to LN produces missing values. Unit mix-ups are also common: 0.10 is 10%, not 0.10%.

PHẢI THUỘC / MEMORISE – Lợi suất đơn = (P(t) − P(t−1)) / P(t−1). Lợi suất log = ln(P(t)/P(t−1)). Log cộng được theo thời gian, đơn cộng được giữa các tài sản. Chuỗi lợi suất thường dừng, chuỗi giá thường không.
THUOC: Simple return = (P(t) − P(t−1)) / P(t−1). Log return = ln(P(t)/P(t−1)). Log returns add across time, simple returns add across assets. Return series are usually stationary, price series usually are not.

1.9. Bài tập vận dụng có lời giải – *1.9. Applied exercises with solutions*

Bài 1. Tính tỷ suất sinh lợi của một cổ phiếu có cổ tức – *Exercise 1. Computing the return of a dividend-paying stock*

Cổ phiếu X có giá cuối tháng 1 là 50.000 đồng, cuối tháng 2 là 54.000 đồng (trong tháng 2 nhận cổ tức 1.000 đồng), cuối tháng 3 là 51.300 đồng (không có cổ tức). Tính tỷ suất đơn giản từng tháng 2 và 3, tỷ suất đơn giản cộng dồn hai tháng, và tỷ suất log từng tháng
Stock X closed at 50,000 dong at the end of month 1 and 54,000 dong at the end of month 2 (a 1,000 dong dividend was paid in month 2). It closed at 51,300 dong at the end of month 3 with no dividend. Compute the simple return for months 2 and 3, the two-month cumulative simple return, and the log return for each month

1. Tháng 2: lãi giá là 54.000 − 50.000 = 4.000 đồng, cộng cổ tức 1.000 đồng thành 5.000 đồng. Chia cho giá đầu kỳ: 5.000 / 50.000 = 0,10 = 10%
   Month 2: the price gain is 54,000 − 50,000 = 4,000 dong; adding the 1,000 dong dividend gives 5,000 dong. Divide by the starting price: 5,000 / 50,000 = 0.10 = 10%

2. Tháng 3 không cổ tức: (51.300 − 54.000) / 54.000 = −2.700 / 54.000 = −0,05 = −5%
   Month 3 has no dividend: (51,300 − 54,000) / 54,000 = −2,700 / 54,000 = −0.05 = −5%
3. Cộng dồn hai tháng phải nhân các hệ số: 1,10 × 0,95 = 1,045, tức tỷ suất 4,5%. Không được lấy 10% − 5% = 5%
   Cumulating over two months means multiplying the growth factors: 1.10 × 0.95 = 1.045, a return of 4.5%. Simply subtracting 10% − 5% = 5% is wrong
4. Tỷ suất log tháng 2 = ln(55.000 / 50.000) = ln(1,1) ≈ 0,0953. Tháng 3 = ln(51.300 / 54.000) = ln(0,95) ≈ −0,0513
   Log return for month 2 = ln(55,000 / 50,000) = ln(1.1) ≈ 0.0953. For month 3 = ln(51,300 / 54,000) = ln(0.95) ≈ −0.0513
5. Kết quả: tỷ suất đơn giản 10% và −5%, cộng dồn 4,5%; tỷ suất log 0,0953 và −0,0513, tổng 0,0440 đúng bằng ln(1,045)
   Result: simple returns are 10% and −5%, cumulative 4.5%; log returns are 0.0953 and −0.0513, and their sum 0.0440 equals ln(1.045)

Bài 2. Biến đổi dữ liệu hộ gia đình bằng Compute, Recode và Select Cases – ***Exercise 2. Transforming household data with Compute, Recode and Select Cases***

Sáu hộ có (tổng thu nhập tháng, triệu đồng; số người): hộ 1 (24; 4), hộ 2 (30; 3), hộ 3 (18; 2), hộ 4 (45; 5), hộ 5 (27; 3), hộ 6 (36; 4). Tính biến thu nhập bình quân đầu người. Mã hoá thành nhóm: dưới 9 là 1, từ 9 đến dưới 10 là 2, từ 10 trở lên là 3. Sau đó chọn các hộ có thu nhập bình quân từ 9 trở lên và tính trung bình của biến này trong nhóm được chọn

Six households have (monthly total income, million dong; household size): H1 (24; 4), H2 (30; 3), H3 (18; 2), H4 (45; 5), H5 (27; 3), H6 (36; 4). Compute per-capita income. Recode it into groups: under 9 is 1, 9 to under 10 is 2, 10 or more is 3. Then select households with per-capita income of 9 or more and find the mean of this variable in the selected group

1. Biến mới (Transform, Compute Variable) là tổng thu nhập chia số người: 24/4 = 6; 30/3 = 10; 18/2 = 9; 45/5 = 9; 27/3 = 9; 36/4 = 9
   The new variable (Transform, Compute Variable) is total income divided by household size: 24/4 = 6; 30/3 = 10; 18/2 = 9; 45/5 = 9; 27/3 = 9; 36/4 = 9
2. Mã hoá lại (Transform, Recode into Different Variables, để giữ biến gốc): hộ 1 có 6 nên mã 1; hộ 2 có 10 nên mã 3; bốn hộ còn lại có 9 nên mã 2. Số ca theo mã: 1 hộ, 4 hộ, 1 hộ
   Recode (Transform, Recode into Different Variables, which keeps the original): H1 has 6 so code 1; H2 has 10 so code 3; the other four households have 9 so code 2. Counts by code: 1, 4, 1
3. Chọn ca (Data, Select Cases, điều kiện bình quân ≥ 9) giữ hộ 2 đến hộ 6, tức 5 ca; hộ 1 bị loại khỏi phân tích
   Select Cases (Data, Select Cases, condition per-capita ≥ 9) keeps households 2 to 6, which is 5 cases; household 1 is excluded from the analysis
4. Kết quả: trung bình trong 5 hộ được chọn = (10 + 9 + 9 + 9 + 9) / 5 = 46 / 5 = 9,2 triệu đồng/người
   Result: the mean over the 5 selected households = (10 + 9 + 9 + 9 + 9) / 5 = 46 / 5 = 9.2 million dong per person

Bài 3. Giá trị khuyết mã 999 làm lệch trung bình – ***Exercise 3. A missing-value code of 999 distorts the mean***

Biến thu nhập (triệu đồng/tháng) của 10 người gồm: 12, 15, 999, 18, 14, 999, 16, 13, 17, 15, trong đó 999 là mã người trả lời từ chối khai. Tính trung bình khi chưa khai báo 999 là khuyết, tỷ lệ khuyết, và trung bình sau khi khai báo đúng

The income variable (million dong per month) for 10 people is: 12, 15, 999, 18, 14, 999, 16, 13, 17, 15, where 999 is the code for refusing to answer. Compute the mean before declaring 999 as missing, the share of missing data, and the mean after declaring it correctly

1. Chưa khai báo: SPSS coi 999 là số thật. Tổng = 12 + 15 + 18 + 14 + 16 + 13 + 17 + 15 + 999 + 999 = 120 + 1.998 = 2.118
   Before declaring: SPSS treats 999 as a real number. The sum = 12 + 15 + 18 + 14 + 16 + 13 + 17 + 15 + 999 + 999 = 120 + 1,998 = 2,118
2. Trung bình sai = 2.118 / 10 = 211,8 triệu đồng, vô lý so với các giá trị còn lại
   The wrong mean = 2,118 / 10 = 211.8 million dong, which is absurd next to the other values
3. Tỷ lệ khuyết = 2 / 10 = 20%. Khai báo trong Variable View, cột Missing, giá trị rời rạc 999
   Missing share = 2 / 10 = 20%. Declare it in Variable View, column Missing, as the discrete value 999
4. Kết quả: còn 8 giá trị hợp lệ, tổng 120, trung bình đúng = 120 / 8 = 15 triệu đồng/tháng, khác hẳn 211,8
   Result: 8 valid values remain with sum 120, so the correct mean = 120 / 8 = 15 million dong per month, far from 211.8

1.10. Câu hỏi tự luận và dàn ý trả lời – *1.10. Essay questions and answer outlines*

Câu 1. Trình bày các bước xây dựng một mô hình kinh tế lượng và nêu ví dụ cho từng bước

Question 1. *Describe the steps of building an econometric model and give an example for each step*

- Bước đầu: nêu vấn đề và giả thuyết từ lý thuyết kinh tế, ví dụ chi tiêu tăng theo thu nhập
  First: state the problem and hypothesis from economic theory, for example spending rises with income
- Xác định dạng mô hình và các biến, rồi thu thập dữ liệu phù hợp (chéo, chuỗi thời gian hoặc bảng)
  Specify the model form and variables, then collect suitable data (cross-section, time series or panel)
- Ước lượng tham số bằng phần mềm như SPSS, sau đó kiểm định xem mô hình và các hệ số có ý nghĩa không
  Estimate the parameters with software such as SPSS, then test whether the model and coefficients are meaningful

- Dùng mô hình đã qua kiểm định để dự báo hoặc đề xuất chính sách, và quay lại sửa mô hình nếu kết quả vô lý
  Use the tested model for forecasting or policy advice, and go back to revise it if results make no sense

Câu 2. Phân biệt bốn thang đo (định danh, thứ bậc, khoảng, tỷ lệ) và nêu ví dụ đời thường cho mỗi loại

Question 2. *Distinguish the four measurement scales (nominal, ordinal, interval, ratio) and give an everyday example of each*

- Định danh chỉ để phân loại, không có thứ tự, ví dụ nghề nghiệp hay nơi ở
  Nominal only labels categories with no order, for example occupation or place of residence
- Thứ bậc có thứ tự nhưng khoảng cách không chắc bằng nhau, ví dụ mức hài lòng từ 1 đến 5
  Ordinal has an order but the gaps may not be equal, for example satisfaction from 1 to 5
- Khoảng có khoảng cách bằng nhau nhưng không có số 0 thật, ví dụ nhiệt độ độ C
  Interval has equal gaps but no true zero, for example temperature in degrees Celsius
- Tỷ lệ có số 0 thật nên so sánh được bằng phép chia, ví dụ thu nhập, giá cổ phiếu
  Ratio has a true zero so division makes sense, for example income or a share price

Câu 3. Giải thích vì sao phải khai báo giá trị khuyết và nêu cách xử lý trong SPSS

Question 3. *Explain why missing values must be declared and how to handle them in SPSS*

- Mã như 999 hay 99 nếu không khai báo sẽ bị tính như số thật, làm lệch trung bình và các thống kê khác
  A code such as 999 or 99, if undeclared, is counted as a real number and distorts the mean and other statistics
- Cách khai báo: vào Variable View, cột Missing, nhập mã khuyết cho từng biến
  How to declare: go to Variable View, column Missing, and enter the missing code for each variable
- Cần nhìn tỷ lệ khuyết của từng biến và hiểu rằng nhiều phân tích sẽ bỏ cả ca có một ô khuyết, làm mẫu hao đi
  Check the missing share of each variable, and note that many analyses drop a whole case if one cell is missing, shrinking the sample
- Có thể bỏ ca khuyết hoặc thay bằng giá trị hợp lý, nhưng phải ghi rõ cách làm
  You may drop incomplete cases or replace values sensibly, but the method must be stated clearly

Câu 4. Nêu các cách đưa dữ liệu vào SPSS và những việc cần làm trong Data Editor sau khi nhập

Question 4. *List the ways to get data into SPSS and what to do in the Data Editor afterwards*

- Mở trực tiếp tệp Excel, tệp văn bản hoặc kết nối cơ sở dữ liệu, hoặc gõ tay vào Data View
  Open an Excel file, a text file or a database connection directly, or type values into Data View

- Kiểm tra dòng đầu có phải tên cột không và kiểu dữ liệu từng cột (số hay chữ) có đọc đúng không
  Check whether the first row holds column names and whether each column's type (number or text) was read correctly
- Trong Variable View, đặt tên biến, loại, nhãn biến, nhãn giá trị (ví dụ 1 là nam, 2 là nữ) và thang đo
  In Variable View, set the variable name, type, label, value labels (for example 1 is male, 2 is female) and measurement scale
- Khai báo giá trị khuyết, rồi chạy thống kê mô tả đơn giản để phát hiện số lạ trước khi phân tích
  Declare missing values, then run simple descriptive statistics to spot odd numbers before analysis

1.11. Câu hỏi trắc nghiệm – *1.11. Multiple-choice questions*

1. Biến "mức hài lòng" với các lựa chọn 1 rất thấp, 2 thấp, 3 vừa, 4 cao, 5 rất cao thuộc thang đo nào?
Which scale does the variable "satisfaction level" (1 very low, 2 low, 3 medium, 4 high, 5 very high) belong to?

A. Khoảng / *Interval*
B. Định danh / *Nominal*
C. Thứ bậc / *Ordinal*
D. Tỷ lệ / *Ratio*

2. Số liệu GDP hằng năm của một nền kinh tế trong 20 năm liên tiếp là loại dữ liệu nào?
What type of data is the annual GDP of one economy over 20 consecutive years?

A. Dữ liệu chéo / *Cross-sectional data*
B. Chuỗi thời gian / *Time series data*
C. Dữ liệu bảng / *Panel data*
D. Dữ liệu định tính / *Qualitative data*

3. Giá cổ phiếu đầu kỳ là 80.000 đồng, cuối kỳ là 88.000 đồng, không có cổ tức. Tỷ suất sinh lợi đơn giản là bao nhiêu?
A share costs 80,000 dong at the start and 88,000 dong at the end, with no dividend. What is the simple return?

A. 10%
B. 12,5% / *12.5%*
C. 9%
D. 8%

4. Trong SPSS, kết quả của một phép phân tích như bảng thống kê mô tả hiện ở cửa sổ nào?
In SPSS, in which window does the result of an analysis, such as a descriptive statistics table, appear?

A. Syntax Editor
B. Data View
C. Output Viewer
D. Variable View

5. Bước nào đứng đầu trong quy trình xây dựng một mô hình kinh tế lượng?
Which step comes first when building an econometric model?

A. Dự báo giá trị tương lai / *Forecast future values*
B. Nêu vấn đề và giả thuyết dựa trên lý thuyết kinh tế / *State the problem and hypothesis based on economic theory*
C. Ước lượng các tham số bằng phần mềm / *Estimate the parameters with software*
D. Kiểm định ý nghĩa của các hệ số / *Test the significance of the coefficients*

6. Cột thu nhập có mã 999 nghĩa là "từ chối khai". Nếu không khai báo 999 là giá trị khuyết thì điều gì xảy ra khi tính trung bình?
The income column uses 999 to mean "refused to answer". If 999 is not declared as missing, what happens when the mean is computed?

A. Trung bình bị kéo lệch vì 999 được tính như một số thật / *The mean is pulled off because 999 is counted as a real number*
B. Trung bình không đổi vì 999 chỉ là một mã / *The mean is unchanged because 999 is only a code*
C. SPSS tự nhận ra 999 là bất thường và bỏ qua / *SPSS spots 999 as abnormal and skips it*
D. Cột thu nhập bị xoá khỏi tệp dữ liệu / *The income column is deleted from the data file*

7. Sau khi dùng Select Cases để lọc theo điều kiện (giữ cách lọc mặc định, không xoá), các ca không thoả điều kiện sẽ ra sao?
After using Select Cases to filter by a condition (keeping the default filter option, not deleting), what happens to cases that fail the condition?

A. Bị đổi thành giá trị khuyết trong dữ liệu gốc / *They are turned into missing values in the original data*
B. Vẫn được tính vào mọi kết quả phân tích / *They are still counted in every analysis result*
C. Bị xoá vĩnh viễn khỏi tệp ngay lập tức / *They are permanently removed from the file at once*
D. Bị loại khỏi phân tích nhưng vẫn còn trong tệp / *They are left out of the analysis but remain in the file*

8. Giá một cổ phiếu tăng từ 100 lên 110, không cổ tức. Tỷ suất sinh lợi log xấp xỉ bằng bao nhiêu?

A share price rises from 100 to 110 with no dividend. About how large is the log return?

A. 0,110 / *0.110*

B. 0,905 / *0.905*

C. 0,100 / *0.100*

D. 0,095 / *0.095*

Chương 2. Mô tả và khám phá dữ liệu – *Chapter 2. Describing and Exploring Data*

Mục tiêu chương – *Chapter objectives*

- Tính và đọc được tần số, phần trăm, trung bình, trung vị, độ lệch chuẩn và độ méo, rồi chọn thước đo phù hợp với hình dạng dữ liệu
  Calculate and read frequencies, percentages, mean, median, standard deviation and skewness, then choose the measure that suits the shape of the data
- Dùng Explore, Crosstabs, kiểm định Shapiro-Wilk và biểu đồ phân tán để phát hiện ngoại lai, so sánh nhóm, kiểm tra tính chuẩn và nhận ra quan hệ giữa hai biến
  Use Explore, Crosstabs, the Shapiro-Wilk test and scatter plots to spot outliers, compare groups, check normality and recognise the relationship between two variables
- Lập codebook và viết đoạn báo cáo mô tả gọn cho luận văn, tránh các lỗi mô tả thường gặp
  Build a codebook and write a short descriptive report for a thesis, avoiding common description mistakes

2.1. Bảng tần số và biểu đồ cho biến phân loại – *2.1. Frequency Tables and Charts for Categorical Variables*

Biến phân loại là biến chỉ cho biết một người thuộc nhóm nào, ví dụ ngành nghề, giới tính, tỉnh cư trú. Với loại biến này, trung bình cộng không có nghĩa. Việc đầu tiên cần làm là đếm: mỗi nhóm có bao nhiêu người và chiếm bao nhiêu phần của mẫu. Bảng tần số làm đúng việc đó.

A categorical variable only tells you which group someone belongs to, such as occupation, gender or place of residence. An average makes no sense for it. The first job is counting: how many people are in each group and what share of the sample they make up. A frequency table does exactly that.

Tần số (frequency) là số quan sát rơi vào một nhóm. Tần suất, hay phần trăm, là tần số chia cho tổng rồi nhân 100. Nói "64 người" kém dễ hiểu hơn "gần một phần ba mẫu". Biểu đồ cột và biểu đồ tròn vẽ cùng những con số đó thành hình để người đọc nhìn một lần là thấy.

Frequency is the number of observations that fall in one group. Relative frequency, or percent, is the frequency divided by the total, times 100. "64 people" says less than "almost a third of the sample". Bar charts and pie charts draw the same numbers as pictures, so a reader sees the pattern at a glance.

Chuẩn bị: một tệp dữ liệu đã mở trong SPSS, trong đó có biến phân loại (ở đây là biến nghenghiep) đã khai báo thang đo Nominal (định danh) và đã gán nhãn giá trị, ví dụ 1 = Công nhân, 2 = Nhân viên văn phòng. Chưa gán nhãn thì bảng chỉ hiện số 1, 2, 3, rất khó đọc.

You need a data file open in SPSS with a categorical variable (here nghenghiep) declared as Nominal and given value labels, for example 1 = Worker, 2 = Office staff. Without labels the table shows only 1, 2, 3, which is hard to read.

Các bước lập bảng và biểu đồ – *Steps to build the table and charts*

1. Vào menu Analyze, chọn Descriptive Statistics, rồi Frequencies.
   Open the Analyze menu, choose Descriptive Statistics, then Frequencies.
2. Chuyển biến nghenghiep từ danh sách bên trái sang ô Variable(s) bằng mũi tên giữa hai khung.
   Move nghenghiep from the left list into the Variable(s) box with the arrow between the panes.
3. Giữ dấu chọn ở Display frequency tables để SPSS in bảng tần số.
   Keep Display frequency tables ticked so SPSS prints the frequency table.
4. Bấm nút Charts. Chọn Bar charts hoặc Pie charts. Ở mục Chart Values, chọn Frequencies để trục hiện số người, hoặc Percentages để trục hiện phần trăm. Bấm Continue.
   Click Charts. Choose Bar charts or Pie charts. Under Chart Values pick Frequencies to show counts, or Percentages to show percents. Click Continue.
5. Bấm OK. Kết quả hiện trong cửa sổ Output, gồm bảng tần số và biểu đồ.
   Click OK. The results appear in the Output window: the frequency table and the chart.

Ví dụ: cơ cấu ngành nghề của mẫu khảo sát – *Example: occupation mix of a survey sample*

Khảo sát 210 người lao động, trong đó 10 người bỏ trống câu hỏi nghề nghiệp. Bảng tần số SPSS in ra như sau:

A survey covers 210 workers, and 10 of them left the occupation question blank. SPSS prints this frequency table:

Nghề nghiệp *Occupation*	**Tần số** *Frequency*	**Phần trăm** *Percent*	**Phần trăm hợp lệ** *Valid Percent*	**Phần trăm cộng dồn** *Cumulative Percent*
Công nhân *Worker*	64	30,5	32,0	32,0
Nhân viên văn phòng *Office staff*	52	24,8	26,0	58,0
Kinh doanh tự do *Self-employed*	40	19,0	20,0	78,0
Giáo viên *Teacher*	24	11,4	12,0	90,0
Khác *Other*	20	9,5	10,0	100,0
Giá trị khuyết (System) *Missing (System)*	10	4,8	–	–

Nghề nghiệp *Occupation*	**Tần số** *Frequency*	**Phần trăm** *Percent*	**Phần trăm hợp lệ** *Valid Percent*	**Phần trăm cộng dồn** *Cumulative Percent*
Tổng *Total*	210	100,0	–	–

Cách đọc: cột Frequency đếm người. Cột Percent chia cho cả 210 người, kể cả người bỏ trống: 64 chia 210 được 30,5%. Cột Valid Percent chỉ chia cho 200 người đã trả lời: 64 chia 200 được 32,0%. Khi báo cáo cơ cấu ngành nghề, bạn dùng Valid Percent, vì người bỏ trống không thuộc nhóm nào.

How to read it: Frequency counts people. Percent divides by all 210, including blanks: 64 divided by 210 gives 30.5%. Valid Percent divides only by the 200 who answered: 64 divided by 200 gives 32.0%. When you report the occupation mix, use Valid Percent, because blank answers belong to no group.

Kết luận cho ví dụ: công nhân là nhóm đông nhất (32,0%), tiếp đến nhân viên văn phòng (26,0%). Hai nhóm này gộp lại chiếm 58,0% số người trả lời. Biểu đồ cột sẽ cho thấy cột công nhân cao nhất và cột "Khác" thấp nhất. Biểu đồ tròn chia hình thành năm miếng theo đúng các tỷ lệ trên.

Conclusion for the example: workers are the largest group (32.0%), followed by office staff (26.0%). Together they make up 58.0% of respondents. The bar chart shows the worker bar tallest and the "Other" bar shortest. The pie chart splits into five slices in the same proportions.

Chọn biểu đồ nào – *Which chart to choose*

Biểu đồ cột hợp với hầu hết trường hợp, nhất là khi có nhiều nhóm hoặc cần so sánh độ cao. Biểu đồ tròn chỉ nên dùng khi có tối đa khoảng năm, sáu nhóm và bạn muốn nhấn mạnh "phần của tổng thể". Trục tung của biểu đồ cột phải bắt đầu từ 0, nếu không cột sẽ trông chênh lệch hơn thực tế.

A bar chart suits most cases, especially with many groups or when you compare heights. Use a pie chart only with about five or six groups at most, and when you want to stress the "share of the whole". The vertical axis of a bar chart must start at 0, otherwise the bars look more different than they are.

> **PHẢI THUỘC / MEMORISE –** Công thức tần suất: phần trăm của một nhóm = tần số nhóm ÷ tổng số × 100. Báo cáo cơ cấu thì dùng Valid Percent (chia cho số người đã trả lời).
> *Relative frequency: percent of a group = group frequency ÷ total × 100. Report the mix with Valid Percent (divided by the number of people who answered).*

> **LỖI HAY GẶP / COMMON MISTAKE –** Lấy cột Percent thay cho Valid Percent khi có giá trị khuyết, khiến các tỷ lệ cộng lại không bằng 100% trong bảng báo cáo.
> *Using the Percent column instead of Valid Percent when values are missing, so the shares in your report do not add to 100%.*
>
> **LỖI HAY GẶP / COMMON MISTAKE –** Đọc cột Cumulative Percent của biến định danh. Nghề nghiệp không có thứ tự, nên "58% là công nhân hoặc nhân viên văn phòng" chỉ đúng vì

thứ tự mã hoá tuỳ ý, không mang ý nghĩa gì. Cột này chỉ có ích với biến thứ bậc như mức hài lòng.
Reading Cumulative Percent for a nominal variable. Occupations have no order, so the running total depends on an arbitrary coding order and means nothing. The column is useful only for ordinal variables such as satisfaction level.

LỖI HAY GẶP / COMMON MISTAKE – Quên gán nhãn giá trị, nên bảng và biểu đồ chỉ hiện mã số. Hoặc dùng biểu đồ tròn cho biến có hơn chục nhóm, miếng nhỏ chồng lên nhau không đọc được.
Forgetting value labels, so tables and charts show only code numbers. Or using a pie chart for a variable with a dozen groups, where tiny slices crowd together and cannot be read.

MẸO / TIP – Muốn bảng sắp theo số người từ nhiều đến ít, trong hộp Frequencies bấm nút Format rồi chọn Descending counts.
To sort the table from most to fewest people, click Format in the Frequencies box and choose Descending counts.

2.2. Trung bình, trung vị, độ lệch chuẩn và độ méo – *2.2. Mean, Median, Standard Deviation and Skewness*

Hãy hình dung một phòng có 5 nhân viên, lương tháng (triệu đồng) là 10, 12, 12, 14 và 52. Bốn người lương quanh mức 12, riêng một người là trưởng phòng lương 52. Ba thước đo đầu mô tả "mức lương điển hình" và "độ phân tán" theo những cách khác nhau, còn độ méo cho biết đuôi dữ liệu lệch về phía nào.
Picture a room of 5 employees with monthly salaries (million VND) of 10, 12, 12, 14 and 52. Four people earn around 12, and one manager earns 52. Three measures describe the "typical salary" and the spread in different ways, and skewness tells you which way the tail leans.

Bốn thước đo, bốn câu hỏi – *Four measures, four questions*

Trung bình (mean) là cộng hết rồi chia cho số người. Nó giống điểm cân bằng của một chiếc bập bênh: một giá trị quá xa sẽ kéo cả điểm cân bằng về phía mình. Ở đây trung bình là 100 / 5 = 20 triệu, cao hơn lương của 4 trong 5 người.
The mean is the sum divided by the count. Think of it as the balance point of a seesaw: one distant value drags the whole balance point toward itself. Here the mean is 100 / 5 = 20 million, higher than what 4 of the 5 people earn.

Trung vị (median) là giá trị đứng giữa khi xếp từ nhỏ đến lớn. Với 5 người, đó là người thứ ba: 12 triệu. Dù trưởng phòng lương 52 hay 520, trung vị vẫn là 12. Vì vậy trung vị phản ánh "người bình thường" tốt hơn khi có ngoại lai.
The median is the middle value after sorting from smallest to largest. With 5 people it is the third one: 12 million. Whether the manager earns 52 or 520, the median stays 12. That is why the median describes the "ordinary person" better when outliers are present.

Độ lệch chuẩn (standard deviation) cho biết các giá trị thường cách trung bình bao xa. Nó lấy độ lệch của từng giá trị so với trung bình, bình phương lên, cộng lại, chia cho n – 1 rồi lấy căn bậc hai. Chia cho n – 1 thay vì n vì ta đang dùng mẫu để đoán về tổng thể.

The standard deviation tells you how far values typically sit from the mean. Take each value's deviation from the mean, square it, add them up, divide by n – 1 and take the square root. We divide by n – 1 rather than n because a sample is being used to estimate the population.

Độ méo (skewness) đo mức lệch của phân phối. Dương nghĩa là đuôi dài về phía giá trị lớn, thường gặp ở lương, giá nhà, doanh thu. Âm là đuôi dài về phía giá trị nhỏ. Gần 0 là cân đối.

Skewness measures how lopsided a distribution is. Positive means a long tail toward large values, common for salaries, house prices and revenue. Negative means a long tail toward small values. Near 0 means balanced.

Tính tay rồi đối chiếu SPSS – ***Calculating by hand and checking against SPSS***

1. Nhập 5 giá trị vào một biến tên luong trong Data Editor, thang đo Scale.
   Enter the 5 values in one variable named luong in the Data Editor, measured as Scale.
2. Tính trung bình: (10 + 12 + 12 + 14 + 52) / 5 = 20. Trung vị: dãy đã xếp sẵn, số giữa là 12.
   Compute the mean: (10 + 12 + 12 + 14 + 52) / 5 = 20. For the median, the list is already sorted and the middle number is 12.
3. Tính độ lệch so với 20: –10, –8, –8, –6, +32. Bình phương: 100, 64, 64, 36, 1024. Tổng là 1288.
   Find the deviations from 20: –10, –8, –8, –6, +32. Square them: 100, 64, 64, 36, 1024. The total is 1288.
4. Phương sai mẫu = 1288 / 4 = 322. Độ lệch chuẩn = căn 322 ≈ 17,94.
   Sample variance = 1288 / 4 = 322. Standard deviation = square root of 322 ≈ 17.94.
5. Độ méo theo công thức SPSS: n / ((n – 1)(n – 2)) nhân tổng của (độ lệch / độ lệch chuẩn) lập phương. Lập phương các độ lệch: –1000, –512, –512, –216, 32768, tổng 30528. Chia cho $17{,}94^3 \approx 5778{,}1$ được ≈ 5,283. Nhân 5 / (4 × 3) = 0,4167 ra ≈ 2,20.
   The SPSS skewness formula is n / ((n – 1)(n – 2)) times the sum of cubed (deviation / standard deviation). Cubed deviations: –1000, –512, –512, –216, 32768, summing to 30528. Divide by $17.94^3 \approx 5778.1$ to get ≈ 5.283. Multiply by 5 / (4 × 3) = 0.4167 to get ≈ 2.20.
6. Trong SPSS chọn Analyze > Descriptive Statistics > Descriptives, đưa luong sang ô Variable(s), bấm Options, tích Mean, Std. deviation và Skewness, rồi Continue và OK.
   In SPSS choose Analyze > Descriptive Statistics > Descriptives, move luong into Variable(s), click Options, tick Mean, Std. deviation and Skewness, then Continue and OK.
7. Đọc bảng: Mean = 20,00; Std. Deviation = 17,944; Skewness = 2,201 với sai số chuẩn 0,913. Khớp với tính tay.
   Read the table: Mean = 20.00; Std. Deviation = 17.944; Skewness = 2.201 with a standard error of 0.913. This matches the hand calculation.

> **VÍ DỤ / EXAMPLE –** Bỏ mức 52 đi, bốn giá trị 10, 12, 12, 14 có trung bình 12 và trung vị 12, hai số trùng nhau. Chỉ một người lương cao đã đẩy trung bình từ 12 lên 20, trong khi trung vị đứng yên. Đó là dấu hiệu điển hình của phân phối méo dương.
> *Remove the 52 and the four values 10, 12, 12, 14 have a mean of 12 and a median of 12, the same*

number. One high earner pushed the mean from 12 up to 20 while the median stayed put. This is the classic sign of a positively skewed distribution.

PHẢI THUỘC / MEMORISE – Trung bình lớn hơn trung vị rõ rệt thì thường méo dương; trung bình nhỏ hơn trung vị thì thường méo âm. Độ méo chia cho sai số chuẩn của nó, nếu vượt khoảng 2 (về giá trị tuyệt đối) thì độ lệch đáng để ý. Ở ví dụ: 2,201 / 0,913 ≈ 2,41.
A mean clearly above the median usually means positive skew; a mean below the median usually means negative skew. Divide skewness by its standard error: beyond about 2 in absolute value, the asymmetry deserves attention. In the example: 2.201 / 0.913 ≈ 2.41.

LỖI HAY GẶP / COMMON MISTAKE – Báo cáo mỗi trung bình cho dữ liệu méo, khiến người đọc tưởng ai cũng lương khoảng 20 triệu. Nên báo thêm trung vị.
Reporting only the mean for skewed data, which leads readers to think everyone earns about 20 million. Report the median as well.

LỖI HAY GẶP / COMMON MISTAKE – Tìm trung vị trong Descriptives. Bảng này không có trung vị; hãy dùng Frequencies (nút Statistics) hoặc Explore.
Looking for the median in Descriptives. That table has no median; use Frequencies (Statistics button) or Explore.

LỖI HAY GẶP / COMMON MISTAKE – Tính độ lệch chuẩn bằng cách chia cho n. SPSS chia cho n – 1, nên số của bạn sẽ nhỏ hơn: căn (1288 / 5) ≈ 16,05 chứ không phải 17,94.
Dividing by n when computing the standard deviation. SPSS divides by n – 1, so your figure would come out smaller: square root of (1288 / 5) ≈ 16.05 rather than 17.94.

2.3. Explore: ngoại lai, biểu đồ hộp và so sánh nhóm – *2.3. Explore: Outliers, Boxplots and Group Comparison*

Explore là công cụ SPSS để "soi" một biến số kỹ hơn bảng thống kê thông thường. Nó cho bạn trung bình, trung vị, độ rộng dữ liệu, danh sách giá trị lạ và biểu đồ, lại tách được theo nhóm. Ví dụ đời thường: bạn muốn biết tiền ăn trưa của hai phòng ban, và có người ghi nhầm 5.000.000 thay vì 50.000. Explore chỉ thẳng ra người đó.
Explore is the SPSS tool for taking a closer look at one numeric variable than a plain summary table allows. It gives the mean, median, spread, a list of odd values and charts, and it can split everything by group. Think of lunch spending in two departments, where someone typed 5,000,000 instead of 50,000. Explore points straight at that person.

Ngoại lai (outlier) là giá trị nằm xa hẳn phần còn lại. Nó có thể là lỗi nhập liệu, hoặc là một quan sát có thật nhưng hiếm, như một cổ phiếu tăng vọt vì tin đột xuất. Phân biệt hai trường hợp này quyết định bạn sửa, giữ hay loại.
An outlier is a value that sits far from the rest. It may be a data-entry error, or a real but rare observation, such as a stock that jumps on sudden news. Telling these two cases apart decides whether you fix, keep or drop the value.

Chuẩn bị và các bước chạy – *What to prepare and how to run it*

Bạn cần một biến định lượng (ví dụ lợi suất tháng, đơn vị %) và một biến nhóm (ví dụ Nhóm = 1 là ngân hàng, 2 là bất động sản). Nên có thêm một biến mã định danh như mã cổ phiếu để nhìn ra ngoại lai là mã nào.

You need one numeric variable (for example monthly return in %) and one grouping variable (for example Group = 1 for banks, 2 for real estate). A case-ID variable such as the ticker is also useful, so you can see which stock an outlier is.

1. Vào Analyze, chọn Descriptive Statistics, rồi chọn Explore.
   Go to Analyze, choose Descriptive Statistics, then choose Explore.
2. Đưa biến lợi suất vào ô Dependent List.
   Move the return variable into the Dependent List box.
3. Đưa biến nhóm vào ô Factor List để SPSS tách kết quả theo nhóm.
   Move the group variable into the Factor List box so SPSS splits the output by group.
4. Đưa biến mã cổ phiếu vào ô Label Cases by. Biểu đồ sẽ ghi tên mã cạnh điểm lạ.
   Move the ticker variable into Label Cases by. The chart will then print the ticker beside an odd point.
5. Bấm Statistics, tick Descriptives và Outliers, rồi Continue.
   Click Statistics, tick Descriptives and Outliers, then Continue.
6. Bấm Plots, chọn Factor levels together, tick Stem-and-leaf (bỏ Histogram nếu không cần), tick Normality plots with tests, rồi Continue.
   Click Plots, choose Factor levels together, tick Stem-and-leaf (untick Histogram if you do not need it), tick Normality plots with tests, then Continue.
7. Bấm OK. Đọc bảng Descriptives, bảng Extreme Values, rồi các biểu đồ.
   Click OK. Read the Descriptives table, the Extreme Values table, then the charts.

Đọc biểu đồ hộp (boxplot) – *Reading the boxplot*

Hộp kéo từ tứ phân vị thứ nhất (Q1, mốc mà 25% số liệu nằm dưới) đến tứ phân vị thứ ba (Q3, mốc 75%). Vạch giữa hộp là trung vị. Chiều dài hộp là khoảng tứ phân vị, gọi tắt là IQR. Hai "râu" kéo tới giá trị xa nhất còn không bị coi là ngoại lai.

The box runs from the first quartile (Q1, the point with 25% of the data below it) to the third quartile (Q3, the 75% point). The line inside is the median. The box length is the interquartile range, IQR. The two whiskers reach the farthest value that is not counted as an outlier.

Điểm tròn nhỏ là ngoại lai: cách mép hộp xa hơn 1,5 lần chiều dài hộp. Dấu sao là ngoại lai cực đoan: xa hơn 3 lần chiều dài hộp. Hộp càng cao thì nhóm đó càng phân tán. Nếu vạch trung vị lệch về một đầu hộp, dữ liệu bị méo.

A small circle marks an outlier: more than 1.5 box lengths from the edge of the box. An asterisk marks an extreme outlier: more than 3 box lengths away. A taller box means a more spread-out group. If the median line sits near one end of the box, the data are skewed.

Đọc biểu đồ thân-lá (stem-and-leaf) – *Reading the stem-and-leaf plot*

Đây là cách vẽ histogram bằng chính các con số. "Thân" là phần chữ số đứng trước, "lá" là chữ số cuối. Dòng thân 2 với lá 2, 4, 6, 9 nghĩa là các giá trị 2,2; 2,4; 2,6; 2,9. Hình

dạng cho bạn thấy chỗ tập trung và đuôi dài. SPSS ghi các giá trị cực đoan ở dòng riêng, có chữ Extremes.

It draws a histogram using the numbers themselves. The "stem" is the leading digits, the "leaf" is the last digit. A stem of 2 with leaves 2, 4, 6, 9 means the values 2.2, 2.4, 2.6 and 2.9. The shape shows where data cluster and where the long tail is. SPSS prints very extreme values on a separate line labelled Extremes.

VÍ DỤ / EXAMPLE – Lợi suất tháng (%) của 10 cổ phiếu mỗi nhóm. Nhóm ngân hàng: 1,0; 1,5; 1,8; 2,0; 2,2; 2,4; 2,6; 2,9; 3,1; 8,0. Trung vị = (2,2 + 2,4) / 2 = 2,3. Chia đôi dãy: Q1 = 1,8, Q3 = 2,9, IQR = 1,1. Ngưỡng ngoại lai trên = 2,9 + 1,5 × 1,1 = 4,55; ngưỡng cực đoan = 2,9 + 3 × 1,1 = 6,2. Giá trị 8,0 vượt cả hai ngưỡng, nên hiện dấu sao. Trung bình = 27,5 / 10 = 2,75, nhưng bỏ 8,0 thì còn 19,5 / 9 ≈ 2,17. Nhóm bất động sản: -4,0; -2,5; -1,0; 0,5; 1,5; 2,5; 3,5; 5,0; 6,5; 8,0. Trung vị = 2,0; Q1 = -1,0; Q3 = 5,0; IQR = 6,0; không có ngoại lai; trung bình = 20 / 10 = 2,0. Kết luận: hai nhóm có lợi suất điển hình gần nhau (trung vị 2,3 và 2,0), nhưng bất động sản dao động mạnh hơn nhiều (hộp dài 6,0 so với 1,1). Ở ngân hàng, một mã duy nhất kéo trung bình lên cao. SPSS có thể cho Q1, Q3 lệch chút so với sổ tay tùy cách tính phân vị; cách đọc không đổi.

Monthly returns (%) of 10 stocks per group. Bank group: 1.0, 1.5, 1.8, 2.0, 2.2, 2.4, 2.6, 2.9, 3.1, 8.0. Median = (2.2 + 2.4) / 2 = 2.3. Splitting the sorted list in halves: Q1 = 1.8, Q3 = 2.9, IQR = 1.1. Upper outlier fence = 2.9 + 1.5 × 1.1 = 4.55; extreme fence = 2.9 + 3 × 1.1 = 6.2. The value 8.0 is beyond both, so it shows as an asterisk. Mean = 27.5 / 10 = 2.75, but without 8.0 it is 19.5 / 9 ≈ 2.17. Real estate group: -4.0, -2.5, -1.0, 0.5, 1.5, 2.5, 3.5, 5.0, 6.5, 8.0. Median = 2.0; Q1 = -1.0; Q3 = 5.0; IQR = 6.0; no outliers; mean = 20 / 10 = 2.0. Conclusion: typical returns are close (medians 2.3 and 2.0), but real estate swings far more (box length 6.0 versus 1.1). In banks, one single stock pulls the mean up. SPSS may give slightly different Q1 and Q3 depending on its percentile method; the reading stays the same.

PHẢI THUỘC / MEMORISE – Ngoại lai trong boxplot: xa mép hộp quá 1,5 lần IQR là điểm tròn, quá 3 lần IQR là dấu sao. So sánh nhóm: nhìn trung vị để biết mức, nhìn chiều dài hộp để biết độ phân tán.

Boxplot outliers: more than 1.5 IQR from the box edge is a circle, more than 3 IQR is an asterisk. Comparing groups: use the median for level and the box length for spread.

LỖI HAY GẶP / COMMON MISTAKE – Thấy ngoại lai là xoá ngay. Hãy tra số ca trong bảng Extreme Values, đối chiếu dữ liệu gốc. Nếu là lỗi nhập thì sửa; nếu là quan sát thật thì giữ, và chạy mô hình cả khi có lẫn không có để so sánh.

Deleting an outlier on sight. Look up the case number in the Extreme Values table and check the source data. If it is a typing error, fix it; if it is a real observation, keep it and run your model both with and without it.

LỖI HAY GẶP / COMMON MISTAKE – Chỉ so trung bình giữa các nhóm. Như ví dụ, trung bình ngân hàng 2,75 cao hơn bất động sản 2,0 chỉ vì một mã. Luôn nhìn cả trung vị và hộp.

Comparing group means only. In the example, the bank mean of 2.75 beats the real estate mean of 2.0 only because of one stock. Always look at the median and the box too.

LỖI HAY GẶP / COMMON MISTAKE – Quên biến nhóm trong Factor List nên SPSS gộp hai nhóm làm một biểu đồ duy nhất, che mất khác biệt.

> *Leaving the grouping variable out of Factor List, so SPSS pools both groups into one chart and hides the difference.*

2.4. Bảng chéo Crosstabs và kiểm định chi bình phương – *2.4. Crosstabs and the Chi-Square Test*

Bảng chéo (Crosstabs) xếp hai biến phân loại vào một bảng: biến này làm hàng, biến kia làm cột. Mỗi ô đếm số người rơi vào cả hai nhóm cùng lúc. Giống như đếm trong một lớp học có bao nhiêu bạn nam đi xe buýt, bao nhiêu bạn nữ đi xe buýt. Kiểm định chi bình phương trả lời câu hỏi: hai biến này có liên hệ với nhau, hay chênh lệch chỉ do may rủi khi chọn mẫu?

A crosstab (cross-tabulation) puts two categorical variables in one table: one as rows, one as columns. Each cell counts the people who fall in both groups at once. It is like counting how many boys and how many girls in a class take the bus. The chi-square test asks whether the two variables are linked, or whether the gap is just chance from sampling.

Ý tưởng của kiểm định là so số đếm thực tế với số đếm "lẽ ra phải có" nếu hai biến không liên quan gì. Số đếm thực tế gọi là tần số quan sát. Số đếm lẽ ra phải có gọi là tần số kỳ vọng. Hai số càng lệch nhau thì càng khó tin rằng hai biến độc lập.

The test compares the counts you actually see with the counts you would expect if the two variables had nothing to do with each other. The first are observed frequencies. The second are expected frequencies. The further apart they are, the harder it is to believe the variables are independent.

Cần chuẩn bị gì – *What you need*

- Hai biến phân loại (nhóm, không phải số đo), ví dụ giới tính và kết quả vay vốn.
  Two categorical variables (groups, not measurements), such as gender and loan outcome.
- Mỗi quan sát chỉ nằm trong một ô, và các quan sát độc lập với nhau.
  Each observation falls in exactly one cell, and observations are independent of each other.
- Tần số kỳ vọng đủ lớn: thường yêu cầu không ô nào dưới 5.
  Expected frequencies large enough: the usual rule is no cell below 5.

Các bước làm tay – *Steps by hand*

1. Lập bảng tần số quan sát, rồi cộng tổng từng hàng, từng cột và tổng chung N.
   Build the table of observed counts, then add the row totals, column totals and the grand total N.
2. Tính tần số kỳ vọng của mỗi ô: (tổng hàng × tổng cột) / N.
   Compute the expected frequency of each cell: (row total × column total) / N.
3. Với mỗi ô, tính (quan sát − kỳ vọng)² / kỳ vọng.
   For each cell, compute (observed − expected)² / expected.
4. Cộng giá trị của mọi ô, được thống kê chi bình phương.
   Add the values of all cells to get the chi-square statistic.

5. Tính bậc tự do = (số hàng − 1) × (số cột − 1). Bảng 2 × 2 có bậc tự do bằng 1.
   Compute degrees of freedom = (rows − 1) × (columns − 1). A 2 × 2 table has 1 degree of freedom.
6. So với giá trị tới hạn (bậc tự do 1, mức ý nghĩa 5% là 3,841). Thống kê lớn hơn thì bác bỏ giả thuyết "hai biến độc lập". Hoặc đọc p-value: p < 0,05 thì bác bỏ.
   Compare with the critical value (1 degree of freedom, 5% level: 3.841). If the statistic is larger, reject the hypothesis that the variables are independent. Or read the p-value: p < 0.05 means reject.

VÍ DỤ / EXAMPLE – Một ngân hàng xem 200 hồ sơ vay. Có 120 hồ sơ nam, 80 hồ sơ nữ. Tổng cộng 130 hồ sơ được duyệt, 70 bị từ chối. Quan sát: nam được duyệt 80, nam bị từ chối 40, nữ được duyệt 50, nữ bị từ chối 30. Kỳ vọng: nam được duyệt = 120 × 130 / 200 = 78; nam bị từ chối = 120 × 70 / 200 = 42; nữ được duyệt = 80 × 130 / 200 = 52; nữ bị từ chối = 80 × 70 / 200 = 28. Đóng góp từng ô: 4/78 = 0,051; 4/42 = 0,095; 4/52 = 0,077; 4/28 = 0,143. Tổng chi bình phương = 0,366, bậc tự do 1. Vì 0,366 < 3,841 (p khoảng 0,55) nên không bác bỏ: với mẫu này chưa có bằng chứng giới tính liên quan đến việc được duyệt vay. Tỷ lệ duyệt của nam là 80/120 = 66,7%, của nữ là 50/80 = 62,5%, chênh lệch nhỏ so với dao động ngẫu nhiên.
A bank reviews 200 loan files: 120 from men, 80 from women. In total 130 were approved and 70 rejected. Observed: men approved 80, men rejected 40, women approved 50, women rejected 30. Expected: men approved = 120 × 130 / 200 = 78; men rejected = 120 × 70 / 200 = 42; women approved = 80 × 130 / 200 = 52; women rejected = 80 × 70 / 200 = 28. Cell contributions: 4/78 = 0.051; 4/42 = 0.095; 4/52 = 0.077; 4/28 = 0.143. Chi-square = 0.366 with 1 degree of freedom. Since 0.366 < 3.841 (p about 0.55), do not reject: this sample gives no evidence that gender is linked to approval. The approval rate is 80/120 = 66.7% for men and 50/80 = 62.5% for women, a gap small enough to be random noise.

Làm trong SPSS – *Doing it in SPSS*

1. Mở Analyze, chọn Descriptive Statistics, rồi Crosstabs.
   Open Analyze, choose Descriptive Statistics, then Crosstabs.
2. Đưa biến giới tính vào ô Row(s), biến kết quả vay vào ô Column(s).
   Move the gender variable to Row(s) and the loan outcome variable to Column(s).
3. Bấm Statistics, tích Chi-square, và tích Phi and Cramér's V để đo độ mạnh của liên hệ. Bấm Continue.
   Click Statistics, tick Chi-square, and tick Phi and Cramér's V to measure the strength of the link. Click Continue.
4. Bấm Cells, tích Observed, Expected và Row. Bấm Continue, rồi OK.
   Click Cells, tick Observed, Expected and Row percentages. Click Continue, then OK.
5. Trong bảng Chi-Square Tests, đọc dòng Pearson Chi-Square: cột Value là thống kê, cột df là bậc tự do, cột Asymptotic Significance là p-value.
   In the Chi-Square Tests table, read the Pearson Chi-Square row: Value is the statistic, df is the degrees of freedom, Asymptotic Significance is the p-value.

Với bảng 2 × 2, SPSS còn in thêm dòng hiệu chỉnh liên tục (Continuity Correction) và kiểm định chính xác Fisher. Khi có ô kỳ vọng dưới 5, hãy dùng kết quả Fisher. SPSS cũng in chú thích về số ô có tần số kỳ vọng dưới 5 ngay dưới bảng.

For a 2 × 2 table, SPSS also prints a Continuity Correction row and Fisher's exact test. When any cell has an expected count below 5, use the Fisher result. SPSS also prints a footnote about the number of cells with expected counts below 5.

PHẢI THUỘC / MEMORISE – Chi bình phương = tổng của (quan sát − kỳ vọng)² / kỳ vọng. Kỳ vọng = tổng hàng × tổng cột / N. Bậc tự do = (hàng − 1)(cột − 1). H0: hai biến độc lập.

Chi-square = sum of (observed − expected)² / expected. Expected = row total × column total / N. Degrees of freedom = (rows − 1)(columns − 1). H0: the two variables are independent.

LỖI HAY GẶP / COMMON MISTAKE – Đọc tần số quan sát thay cho tần số kỳ vọng khi kiểm tra điều kiện "ô không dưới 5". Điều kiện này áp dụng cho số kỳ vọng.

Checking the "no cell below 5" rule on observed counts instead of expected counts. The rule applies to expected counts.

LỖI HAY GẶP / COMMON MISTAKE – Kết luận "có liên hệ" rồi suy ra "biến này gây ra biến kia". Chi bình phương chỉ cho biết có liên hệ, không cho biết nguyên nhân và không cho biết liên hệ mạnh hay yếu. Mẫu lớn có thể cho p rất nhỏ dù liên hệ yếu, nên hãy xem thêm Cramér's V.

Concluding "linked" and then "one causes the other". Chi-square only shows an association, not the cause and not its strength. A large sample can give a tiny p even for a weak link, so also check Cramér's V.

LỖI HAY GẶP / COMMON MISTAKE – Đưa số đếm đã gộp nhóm sai, hoặc nhập tỷ lệ phần trăm thay vì số người. Kiểm định cần số đếm gốc.

Feeding in wrongly grouped counts, or entering percentages instead of head counts. The test needs the raw counts.

MẸO / TIP – Khi so sánh, hãy bật phần trăm theo hàng. Đọc 66,7% và 62,5% dễ hơn nhiều so với đọc 80 và 50 khi hai hàng có cỡ khác nhau.

Turn on row percentages when comparing. Reading 66.7% and 62.5% is much easier than 80 and 50 when the two rows differ in size.

2.5. Phân phối chuẩn và kiểm tra tính chuẩn của dữ liệu – *2.5. The Normal Distribution and Checking Normality*

Hãy nghĩ đến chiều cao của một lớp 200 người. Đa số cao quanh mức trung bình, rất ít người cực thấp hoặc cực cao. Vẽ ra, các cột tạo thành hình quả chuông đối xứng. Đó là phân phối chuẩn (normal distribution): đỉnh nằm đúng giữa, hai đuôi thoải dần về hai phía.

P: Think of the heights of a class of 200 people. Most are close to the average, and very few are extremely short or extremely tall. Plotted, the bars form a symmetric bell. That is the normal distribution: the peak sits in the middle and both tails taper off gradually.

Phân phối chuẩn được xác định bởi hai số: trung bình (vị trí tâm chuông) và độ lệch chuẩn (độ bè của chuông). Nhiều kiểm định và mô hình trong các chương sau, như kiểm định t, ANOVA, hồi quy, đều dựa trên giả định này.

P: A normal distribution is fixed by two numbers: the mean (where the bell is centred) and the standard deviation (how wide it spreads). Many tests and models in later chapters, such as the t-test, ANOVA and regression, rely on this assumption.

Quy tắc 68-95-99,7 và điểm z – *The 68-95-99.7 rule and the z-score*

Với dữ liệu chuẩn, tỷ lệ quan sát nằm gần trung bình đã biết trước. Khoảng trung bình ± 1 độ lệch chuẩn chứa khoảng 68%. Khoảng ± 2 độ lệch chuẩn chứa khoảng 95%. Khoảng ± 3 độ lệch chuẩn chứa khoảng 99,7%.

P: For normal data, the share of observations near the mean is known in advance. The mean ± 1 standard deviation holds about 68%. Mean ± 2 standard deviations holds about 95%. Mean ± 3 standard deviations holds about 99.7%.

Điểm z cho biết một giá trị cách trung bình bao nhiêu độ lệch chuẩn. Công thức: z = (x − trung bình) / độ lệch chuẩn. Dấu âm nghĩa là thấp hơn trung bình, dấu dương là cao hơn.

P: The z-score tells you how many standard deviations a value lies from the mean. Formula: z = (x − mean) / standard deviation. A negative sign means below the mean, a positive sign means above it.

> **VÍ DỤ / EXAMPLE –** Điểm thi của một lớp có trung bình 70 và độ lệch chuẩn 10, dạng chuẩn. Khoảng 68% học viên được từ 60 đến 80 điểm, khoảng 95% được từ 50 đến 90, gần như tất cả (99,7%) được từ 40 đến 100. Bạn được 85 điểm thì z = (85 − 70) / 10 = 1,5: cao hơn trung bình một rưỡi độ lệch chuẩn. Bạn bạn được 55 điểm thì z = (55 − 70) / 10 = −1,5. Điểm z lớn hơn 3 hoặc nhỏ hơn −3 là giá trị rất hiếm, nên kiểm tra xem có nhập sai không.
> *VIDU: Exam scores in a class have a mean of 70 and a standard deviation of 10, and are normal. About 68% of students score between 60 and 80, about 95% between 50 and 90, and nearly all (99.7%) between 40 and 100. If you score 85, z = (85 − 70) / 10 = 1.5: one and a half standard deviations above the mean. A friend who scores 55 has z = (55 − 70) / 10 = −1.5. A z-score above 3 or below −3 marks a very rare value, so check whether it was a typing error.*

Kiểm tra tính chuẩn trong SPSS – *Checking normality in SPSS*

Biết lý thuyết chưa đủ, bạn cần xem dữ liệu của mình có gần chuẩn không. Cách gọn nhất là dùng Explore, vì nó cho cả biểu đồ lẫn kiểm định trong một lần chạy.

P: Theory is not enough: you need to see whether your own data are close to normal. The quickest way is Explore, which gives both plots and tests in one run.

1. Mở dữ liệu, chọn Analyze > Descriptive Statistics > Explore.
   Open your data and choose Analyze > Descriptive Statistics > Explore.
2. Đưa biến số cần kiểm tra (ví dụ điểm thi) vào ô Dependent List. Nếu muốn kiểm tra riêng từng nhóm, đưa biến nhóm vào Factor List.
   Move the variable you want to check (for example exam score) into Dependent List. To check each group separately, put the grouping variable in Factor List.

3. Bấm nút Plots, tích ô Normality plots with tests, rồi bấm Continue.
   Click Plots, tick Normality plots with tests, then click Continue.
4. Bấm OK. Đọc bảng Tests of Normality và các biểu đồ Normal Q-Q Plot trong cửa sổ kết quả.
   Click OK. Read the Tests of Normality table and the Normal Q-Q Plot charts in the output window.

Biểu đồ Q-Q (quantile-quantile) xếp các giá trị thực của bạn theo thứ tự, so với giá trị lẽ ra phải có nếu dữ liệu chuẩn. Nếu chuẩn, các chấm nằm sát đường thẳng chéo. Chấm cong thành hình chữ S nghĩa là hai đuôi dày hoặc mỏng hơn chuẩn. Chấm cong thành hình cung nghĩa là dữ liệu lệch về một phía. Vài chấm văng xa ở đầu mút thường là ngoại lai.

P: A Q-Q (quantile-quantile) plot sorts your actual values and compares them with the values you would expect if the data were normal. If they are normal, the dots hug the diagonal line. Dots curving in an S shape mean the tails are heavier or lighter than normal. Dots bending in an arc mean the data are skewed. A few dots far out at the ends are usually outliers.

Bảng Tests of Normality có hai kiểm định: Kolmogorov-Smirnov (kèm hiệu chỉnh Lilliefors) và Shapiro-Wilk. Giả thuyết H0 là dữ liệu có phân phối chuẩn. Nhìn cột Sig.: nếu Sig. lớn hơn 0,05 thì chưa có bằng chứng bác bỏ tính chuẩn. Nếu Sig. nhỏ hơn hoặc bằng 0,05 thì bác bỏ H0, dữ liệu lệch khỏi chuẩn. Với mẫu dưới khoảng 50 quan sát, ưu tiên Shapiro-Wilk.

P: The Tests of Normality table holds two tests: Kolmogorov-Smirnov (with the Lilliefors correction) and Shapiro-Wilk. The null hypothesis H0 says the data are normally distributed. Look at the Sig. column: if Sig. is above 0.05, there is no evidence against normality. If Sig. is 0.05 or below, reject H0 and treat the data as non-normal. For samples under about 50 observations, prefer Shapiro-Wilk.

VÍ DỤ / EXAMPLE – Mẫu 40 học viên cho Shapiro-Wilk có Sig. = 0,318. Vì 0,318 > 0,05, bạn không bác bỏ H0 và có thể coi điểm thi gần chuẩn. Nếu Sig. = 0,012 thì 0,012 < 0,05: bác bỏ H0, cần xem lại Q-Q và biểu đồ hộp để tìm nguyên nhân (thường là ngoại lai hoặc lệch).
VIDU: A sample of 40 students gives a Shapiro-Wilk Sig. of 0.318. Since 0.318 > 0.05, you do not reject H0 and can treat the scores as roughly normal. If Sig. were 0.012, then 0.012 < 0.05: reject H0, and look at the Q-Q plot and boxplot to find the cause (usually outliers or skew).

LỖI HAY GẶP / COMMON MISTAKE – Đọc Sig. lớn là "đã chứng minh chuẩn". Thực ra chỉ là chưa đủ bằng chứng để nói không chuẩn, nhất là khi mẫu nhỏ.
LOI: Reading a large Sig. as "normality proven". It only means there is not enough evidence to call the data non-normal, especially with a small sample.

LỖI HAY GẶP / COMMON MISTAKE – Chỉ nhìn kiểm định mà bỏ qua biểu đồ. Mẫu vài trăm quan sát trở lên thì kiểm định bắt được cả độ lệch rất nhỏ, vô hại, nên Sig. thấp là chuyện thường. Luôn đối chiếu với Q-Q.
LOI: Relying on the test and ignoring the plot. With a few hundred observations or more, the test catches tiny, harmless deviations, so a low Sig. is common. Always compare it with the Q-Q plot.