**📘 PHẦN 5 — SỔ TAY ÔN TẬP TỔNG HỢP / FINAL REVIEW HANDBOOK**

**🧩 I. TỔNG QUAN THƯ VIỆN / LIBRARY OVERVIEW**

| **Thư viện** | **Công dụng chính (Việt)** | **Main Purpose (English)** |
| --- | --- | --- |
| numpy | Tính toán mảng, mô phỏng dữ liệu, sinh ngẫu nhiên | Array operations, numerical simulation |
| pandas | Quản lý, phân tích dữ liệu dạng bảng (Excel-like) | Data manipulation and analysis |
| matplotlib.pyplot | Trực quan hóa dữ liệu | Data visualization and plotting |
| scipy.stats | Hàm xác suất, kiểm định thống kê | Probability distributions, hypothesis testing |
| seaborn *(nếu dùng)* | Vẽ biểu đồ thống kê nâng cao | High-level statistical visualization |

**📄 II. QUY TRÌNH PHÂN TÍCH DỮ LIỆU / DATA ANALYSIS WORKFLOW**

| **Bước** | **Việt** | **English** |
| --- | --- | --- |
| 1️⃣ **Nhập dữ liệu** | Dùng pd.read\_csv(), pd.read\_excel() | Load data using Pandas |
| 2️⃣ **Khám phá dữ liệu** | df.head(), df.info(), df.describe() | Explore data summary |
| 3️⃣ **Làm sạch dữ liệu** | dropna(), fillna(), astype(), drop\_duplicates() | Handle missing, duplicates, type issues |
| 4️⃣ **Phân tích mô tả** | mean(), value\_counts(), groupby() | Compute descriptive statistics |
| 5️⃣ **Trực quan hóa** | plot(), hist(), scatter(), boxplot() | Visualize relationships |
| 6️⃣ **Suy luận thống kê** | scipy.stats.ttest\_\*, norm.cdf(), t.ppf() | Infer population parameters |
| 7️⃣ **Báo cáo kết quả** | Vẽ biểu đồ, ghi nhận xét, lưu file | Plot, interpret, save results |

**🧮 III. CÔNG THỨC THỐNG KÊ CƠ BẢN / BASIC STATISTICAL FORMULAS**

| **Ký hiệu** | **Công thức** | **Ý nghĩa (Việt)** | **Meaning (English)** |
| --- | --- | --- | --- |
| Trung bình mẫu |  | Giá trị trung tâm | Mean |
| Phương sai |  | Độ biến thiên | Variance |
| Độ lệch chuẩn |  | Mức phân tán | Standard deviation |
| Sai số chuẩn |  | Sai số của trung bình | Standard error |
| Khoảng tin cậy 95% (biết σ) |  | Ước lượng trung bình (biết σ) | CI with known σ |
| Khoảng tin cậy 95% (không biết σ) |  | Ước lượng trung bình (dùng t) | CI with unknown σ |
| Phân phối chuẩn PDF |  | Dạng chuẩn xác suất | Normal probability density |
| Định lý giới hạn trung tâm |  | Trung bình mẫu tiệm cận chuẩn | Central Limit Theorem |

**📈 IV. KIỂU DỮ LIỆU & CẤU TRÚC TRONG PANDAS**

| **Đối tượng** | **Giải thích (Việt)** | **English Explanation** |
| --- | --- | --- |
| Series | Một cột dữ liệu (1D) | One-dimensional labeled array |
| DataFrame | Bảng dữ liệu nhiều cột (2D) | Two-dimensional table |
| .index | Dòng đánh số | Row index |
| .columns | Danh sách tên cột | Column names |
| .values | Mảng giá trị thuần | Numpy array of values |

**🧠 V. HÀM THƯỜNG DÙNG TRONG PHÂN TÍCH**

| **Nhóm** | **Lệnh** | **Chức năng (Việt)** | **Function (English)** |
| --- | --- | --- | --- |
| **Khám phá dữ liệu** | head(), info(), describe() | Xem cấu trúc, thống kê tổng quan | Data overview |
| **Lọc dữ liệu** | loc[], iloc[], query() | Lọc dòng theo điều kiện | Filter rows |
| **Nhóm và tổng hợp** | groupby(), agg(), pivot\_table() | Tính toán theo nhóm | Group & aggregate |
| **Xử lý cột** | assign(), drop(), rename() | Thêm, xóa, đổi tên cột | Manage columns |
| **Thống kê** | mean(), sum(), count(), corr() | Tính chỉ số cơ bản | Compute statistics |
| **Trực quan hóa** | plot(), hist(), bar(), scatter() | Vẽ biểu đồ | Visualization |
| **Lưu dữ liệu** | to\_csv(), to\_excel() | Xuất kết quả | Export data |

**🎨 VI. LỆNH VẼ THƯỜNG GẶP TRONG MATPLOTLIB**

| **Lệnh** | **Giải thích (Việt)** | **English Explanation** | **Tham số chính** |
| --- | --- | --- | --- |
| plt.plot(x,y) | Biểu đồ đường | Line chart | color, linestyle, label |
| plt.bar(x,height) | Biểu đồ cột | Bar chart | color, width |
| plt.hist(data,bins) | Biểu đồ tần suất | Histogram | bins, density |
| plt.scatter(x,y) | Biểu đồ phân tán | Scatter plot | alpha, s, color |
| plt.boxplot(data) | Biểu đồ hộp | Boxplot | labels, vert |
| plt.title(), xlabel(), ylabel() | Thêm tiêu đề và tên trục | Titles & labels | — |
| plt.legend(), plt.grid() | Chú thích, lưới | Legend & grid | loc, alpha |
| plt.savefig('fig.png',dpi=300) | Lưu hình | Save figure | dpi, bbox\_inches |

**⚗️ VII. PHÂN PHỐI XÁC SUẤT CƠ BẢN**

| **Phân phối** | **Hàm PDF / PMF** | **Khi nào dùng** | **Ghi chú** |
| --- | --- | --- | --- |
| Chuẩn (Normal) | norm.pdf(x, loc, scale) | Biến liên tục chuẩn | μ: trung bình, σ: độ lệch chuẩn |
| Nhị thức (Binomial) | binom.pmf(k, n, p) | Số thành công trong n thử | p: xác suất |
| Poisson | poisson.pmf(k, mu) | Số sự kiện trong thời gian t | μ: trung bình |
| Mũ (Exponential) | expon.pdf(x, scale) | Thời gian giữa 2 sự kiện | λ = 1/scale |
| Đều (Uniform) | uniform.pdf(x, loc, scale) | Phân bố đồng đều | a=loc, b=loc+scale |

**📊 VIII. MẸO HỌC NHANH / QUICK STUDY TIPS**

| **Chủ đề** | **Mẹo** | **English Note** |
| --- | --- | --- |
| **Pandas** | Dùng groupby + agg thay vì vòng lặp for | Use vectorized ops instead of loops |
| **Matplotlib** | Dùng plt.tight\_layout() khi nhiều subplot | Prevent overlapping labels |
| **SciPy Stats** | Luôn kiểm tra loc, scale khi dùng PDF | Always specify mean and std clearly |
| **Data Cleaning** | Dùng .dropna() cẩn thận — có thể mất dữ liệu | Avoid dropping too much data |
| **Simulation** | Cố định seed: np.random.seed(42) để tái lập kết quả | Fix random seed for reproducibility |
| **Export** | Dùng index=False khi xuất CSV | Avoid adding index column accidentally |

**⚠️ IX. LỖI THƯỜNG GẶP / COMMON MISTAKES**

| **Tình huống** | **Nguyên nhân** | **Cách khắc phục** |
| --- | --- | --- |
| KeyError: 'col' | Sai tên cột | Kiểm tra df.columns |
| ValueError: Length mismatch | Số phần tử không khớp khi gán cột | Đảm bảo độ dài bằng nhau |
| SettingWithCopyWarning | Gán giá trị không đúng cách | Dùng .loc[] khi cập nhật |
| Biểu đồ bị chồng chữ | Không dùng plt.tight\_layout() | Gọi plt.tight\_layout() sau khi vẽ |
| Dữ liệu trống sau dropna() | Có quá nhiều NaN | Dùng fillna() thay thế |
| PDF không trùng với histogram | Sai density=True | Thêm density=True trong plt.hist() |

**🔍 X. CÂU HỎI ÔN TẬP THỰC HÀNH / REVIEW CHECKLIST**

✅ **Pandas**

* Biết đọc / ghi file, xem .info() và .describe()
* Lọc, nhóm, tạo cột mới, xử lý missing
* Biết groupby và pivot\_table

✅ **Matplotlib**

* Vẽ được plot, hist, bar, scatter, boxplot
* Ghi chú, tiêu đề, trục, chú thích

✅ **SciPy Stats**

* Dùng được norm.pdf, cdf, ppf, rvs
* Mô phỏng phân phối, tính xác suất

✅ **Estimation**

* Tính trung bình mẫu, SE, CI với norm và t
* Vẽ đường cong chuẩn & nhận xét kết quả

✅ **Simulation**

* Dùng for + .append() để lặp 100–1000 lần
* Tính trung bình, tỷ lệ, phương sai của mẫu

**📚 XI. KẾT LUẬN CHUNG / FINAL SUMMARY**

| **Chủ đề** | **Cốt lõi (Việt)** | **Core Idea (English)** |
| --- | --- | --- |
| **Pandas** | Làm việc hiệu quả với dữ liệu dạng bảng | Efficient tabular data manipulation |
| **Matplotlib** | Trực quan hóa mọi dạng dữ liệu | Flexible, customizable visualization |
| **SciPy Stats** | Hiểu & mô phỏng phân phối xác suất | Mathematical models for randomness |
| **Estimation** | Suy luận trung bình quần thể từ mẫu | Statistical inference |
| **Simulation** | Thực hành thống kê qua mô phỏng | Understand variability by simulation |