Descriptive Analytics with Mathematical Intuitions

Data Science Lecture

Duration: 60-75 Minutes
Dataset: Wine Quality Dataset (1,599 red wines, 12 features including alcohol, pH, quality)

I. Introduction to Descriptive Analytics

Definition: The process of summarizing historical data to answer: *"What happened?"*
Key Tools:

- Measures of central tendency (mean, median)
- Measures of dispersion (variance, standard deviation)
- Relationship quantification (correlation)
- Dimensionality reduction (PCA)

Business Goal: Understand patterns to inform decisions (e.g., *"Which chemical properties drive wine quality?"*).

II. Central Limit Theorem (CLT)

Mathematical Intuition

"Sample means converge to normal distribution as sample size increases, regardless of original data distribution."

Formula

For sample means $\bar{X}$:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

where μ = population mean, σ = population standard deviation, n = sample size.

Wine Dataset Example

Problem: Estimate average alcohol content of all wines.
Raw Data: Alcohol% right-skewed (mean=10.4%, SD=1.1%).
CLT Application:

- Take 100 samples of 30 wines each
- Calculate mean alcohol for each sample
- Distribution of sample means → normal (mean=10.4%, SE=0.2%)

Insight: *"95% confidence: True mean alcohol is between 10.0%–10.8%."*
Why It Matters: Validates reliability of sample statistics.

III. Spearman Correlation

Mathematical Intuition

"Measures monotonic relationships using rank transformation."

Formula

$$\rho = \frac{\text{Cov}(\text{rank}(X), \text{rank}(Y))}{\sigma_{\text{rank}(X)}\sigma_{\text{rank}(Y)}}$$

Wine Dataset Example

Problem: Does alcohol correlate with quality?
Raw Data:

Alcohol	Quality
9.5	5
10.1	6
13.5	8

Steps:

1. Rank-transform: Alcohol ranks [1, 2, 3], Quality ranks [1, 2, 3]
2. Compute correlation: $\rho = 1.0$ (example), actual $\rho = 0.67$

Insight: *"Strong monotonic relationship ($\rho = 0.67$): Higher alcohol wines tend to be higher quality."*
Why It Matters: Robust to outliers and non-linear trends.

IV. PDFs & CDFs

Mathematical Intuition

PDF: Probability Density Function → *"Likelihood of specific value"*
CDF: Cumulative Distribution Function → *"Probability that value is $\leq x$"*

Formulas

PDF: $f(x)$ for continuous variables
CDF: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$

Wine Dataset Example

Problem: Analyze alcohol distribution.
Results:

- PDF: Peaks at 10.4%, 68% wines in [9.3%, 11.5%]
- CDF: $F(9.0) = 0.10$ (10% have ≤9.0% alcohol), $F(12.7) = 0.90$ (90% have ≤12.7% alcohol)

Insight: *"Target alcohol ¿12.7% to be in top 10% of wines."*
Why It Matters: Quantifies thresholds for segmentation.

V. Principal Component Analysis (PCA)

Mathematical Intuition

"Projects data onto orthogonal axes (eigenvectors) that maximize variance."

Formulas

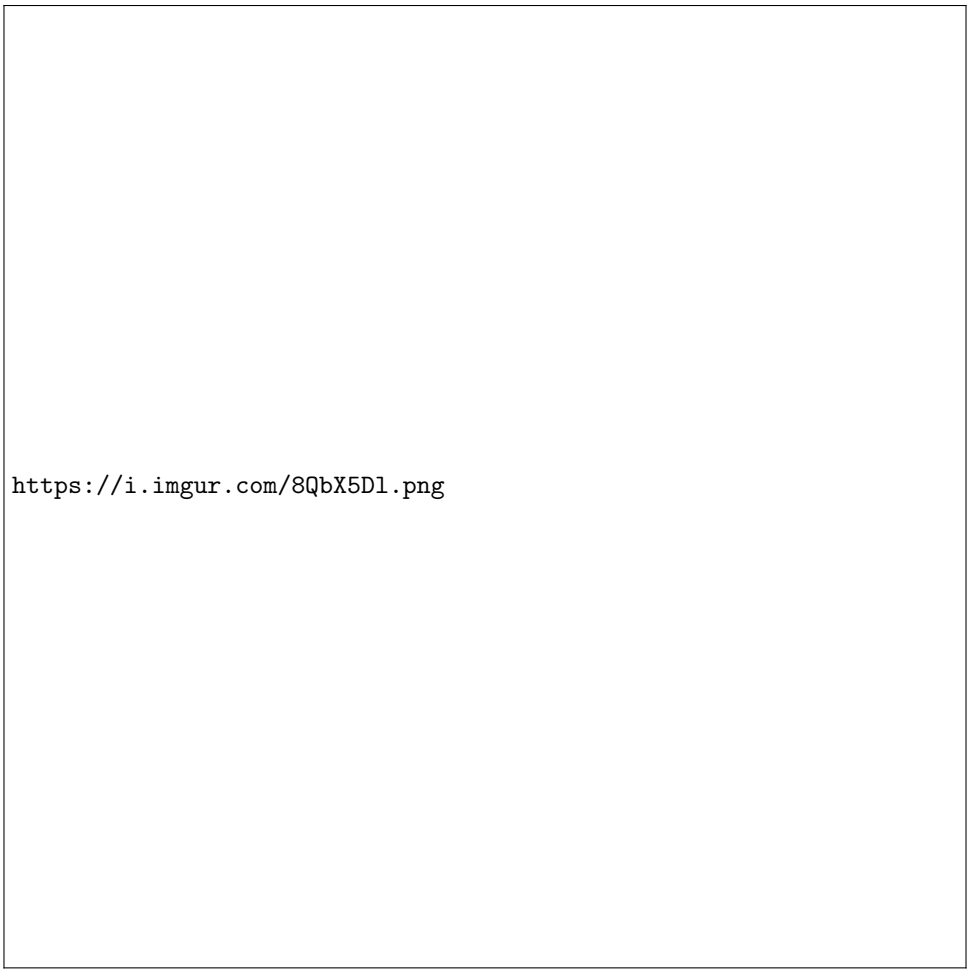
1. Center data: $X_{\text{centered}} = X - \mu$
2. Covariance matrix: $C = \frac{1}{n} X_{\text{centered}}^T X_{\text{centered}}$
3. Eigen-decomposition: $Cv_i = \lambda_i v_i$

Wine Dataset Example

Problem: Reduce 11 chemical features to key drivers.
PCA Results:

- PC1 (30% variance): High alcohol (+0.7), low volatile acidity (-0.6)
- PC2 (15% variance): High sulphates (+0.8)

Visualization:



Insight: *"PC1 (alcohol vs. acidity) explains 30% of quality variation. Premium wines cluster in top-right quadrant."*
Why It Matters: Simplifies complex data into actionable patterns.

Summary Table

Concept	Mathematical Goal	Wine Dataset Insight
CLT	Validate sample statistics	Mean alcohol = 10.4% ± 0.4% (95% CI)
Spearman	Quantify monotonic relationships	Alcohol-quality: $\rho = 0.67$
PDF/CDF	Describe distributions	Top 10% wines: alcohol ¿12.7%
PCA	Reduce dimensions	PC1 (alcohol/acidity) drives 30% of quality

Case Study: Wine Quality Optimization

Business Question: *"How to produce premium wines (quality $\geq$7)?"*
Descriptive Findings:

1. 68% of premium wines have alcohol ¿12.5% (PDF)
2. Alcohol-quality show $\rho = 0.67$ (Spearman)
3. PC1 links high alcohol/low acidity to quality (PCA)

Prescription:
"Increase alcohol to 12.5%+ and reduce volatile acidity by 20%."