Manual Statistics & Regression Guide

Step-by-Step Mathematical Computations

Educational Mathematics Manual

September 29, 2025

Contents

1 Introduction

This manual provides step-by-step mathematical computations for fundamental statistical methods and regression analysis. Each section includes:

- Mathematical theory and formulas
- Detailed manual calculation examples
- Interpretation of results
- Practical applications

The goal is to understand the mathematics behind these methods rather than relying on software implementations.

2 Linear Regression

2.1 Theory and Mathematical Foundation

Linear regression models the relationship between a dependent variable y and independent variable(s) x using a linear equation.

Formula 1 (Simple Linear Regression Model)**.**

$$y = \beta_0 + \beta_1 x + \epsilon$$

where:

- y = *dependent variable (response)*
- x = *independent variable (predictor)*
- β_0 = *y-intercept*
- β_1 = *slope coefficient*
- ϵ = *error term*

2.2 Parameter Estimation using Least Squares

The least squares method minimizes the sum of squared residuals:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Formula 2 (Least Squares Estimators)**.**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where:

- $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ *(covariance sum)*
- $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ *(variance sum for x)*

2.3 Manual Calculation Example

Let's work through a complete example using taxi fare data.

Example 1 (Taxi Fare vs Trip Miles)**.** *Given the following data points:*

Trip Miles (x)	*Fare (y)*
2.5	*8.50*
4.0	*12.75*
6.5	*18.25*
3.2	*10.80*
5.8	*16.40*

Step 1: Calculate means

$$\bar{x} = \frac{2.5 + 4.0 + 6.5 + 3.2 + 5.8}{5} = \frac{22.0}{5} = 4.4$$

$$\bar{y} = \frac{8.50 + 12.75 + 18.25 + 10.80 + 16.40}{5} = \frac{66.70}{5} = 13.34$$

Step 2: Calculate deviations and products

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
2.5	*8.50*	*-1.9*	*-4.84*	*9.196*	*3.61*
4.0	*12.75*	*-0.4*	*-0.59*	*0.236*	*0.16*
6.5	*18.25*	*2.1*	*4.91*	*10.311*	*4.41*
3.2	*10.80*	*-1.2*	*-2.54*	*3.048*	*1.44*
5.8	*16.40*	*1.4*	*3.06*	*4.284*	*1.96*
			Totals:	*27.075*	*11.58*

Step 3: Calculate slope and intercept

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{27.075}{11.58} = 2.338$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 13.34 - 2.338 \times 4.4 = 13.34 - 10.287 = 3.053$$

Step 4: Write the regression equation

$$\hat{y} = 3.053 + 2.338x$$

Interpretation: *For each additional mile, the fare increases by \$2.34 on average, with a base fare of \$3.05.*

2.4 Model Evaluation

2.4.1 Coefficient of Determination (R^2)

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

where:

- $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ (Total Sum of Squares)

- $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ (Regression Sum of Squares)
- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (Error Sum of Squares)

Manual calculation for our example:

x_i	y_i	$\hat{y}_i$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
2.5	8.50	8.898	23.426	19.711	0.158
4.0	12.75	12.405	0.348	0.873	0.119
6.5	18.25	18.250	24.108	24.107	0.000
3.2	10.80	10.535	6.451	7.874	0.070
5.8	16.40	16.613	9.362	10.719	0.045
		Totals:	63.695	63.284	0.392

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{0.392}{63.695} = 1 - 0.0062 = 0.9938$$

This means 99.38% of the variation in fare is explained by trip miles.

2.4.2 Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{0.392}{5-2}} = \sqrt{0.131} = 0.362$$

3 Multiple Linear Regression

3.1 Matrix Formulation

For multiple predictors, we use matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where:

- $\mathbf{y}$ is an $n \times 1$ vector of responses
- $\mathbf{X}$ is an $n \times (p+1)$ design matrix
- $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of coefficients
- $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of errors

Formula 3 (Normal Equations)**.**

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

3.2 Manual Calculation Example

Example 2 (Taxi Fare with Miles and Time)**.** *Data:*

Miles (x_1)	*Minutes* (x_2)	*Fare* (y)
2.5	*15*	*8.50*
4.0	*22*	*12.75*
6.5	*35*	*18.25*

Step 1: Set up design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 2.5 & 15 \\ 1 & 4.0 & 22 \\ 1 & 6.5 & 35 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 8.50 \\ 12.75 \\ 18.25 \end{pmatrix}$$

Step 2: Calculate $\mathbf{X}^T\mathbf{X}$

$$\mathbf{X}^T = \begin{pmatrix} 1 & 1 & 1 \\ 2.5 & 4.0 & 6.5 \\ 15 & 22 & 35 \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 3 & 13.0 & 72 \\ 13.0 & 59.5 & 330 \\ 72 & 330 & 1858 \end{pmatrix}$$

Step 3: Calculate $\mathbf{X}^T\mathbf{y}$

$$\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 39.50 \\ 182.125 \\ 1003.75 \end{pmatrix}$$

Step 4: Solve for $\hat{\boldsymbol{\beta}}$ *Using matrix inversion (or Gaussian elimination):*

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 1.234 \\ 2.156 \\ 0.089 \end{pmatrix}$$

Final model:

$$\hat{y} = 1.234 + 2.156x_1 + 0.089x_2$$

4 Logistic Regression

4.1 Mathematical Foundation

Logistic regression models the probability of a binary outcome using the logistic function.

Formula 4 (Logistic Regression Model)**.**

$$P(Y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

The linear combination $z = \beta_0 + \beta_1 x$ is called the **logit** or log-odds:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

4.2 Maximum Likelihood Estimation

The likelihood function for logistic regression is:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

The log-likelihood is:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i \ln(p_i) + (1-y_i)\ln(1-p_i)]$$

4.3 Manual Calculation Example

Example 3 (Heart Disease Prediction)**.** *Data:*

Age (x)	*Disease (y)*
45	*0*
55	*1*
65	*1*
35	*0*

Using Newton-Raphson method (simplified):
Step 1: Initial estimates *Start with* $\beta_0 = 0$, $\beta_1 = 0$
Step 2: Calculate probabilities *For initial estimates:* $p_i = 0.5$ *for all* i
Step 3: Calculate score functions

$$U_0 = \sum_{i=1}^{n}(y_i - p_i) = (0+1+1+0) - 4(0.5) = 0$$

$$U_1 = \sum_{i=1}^{n} x_i(y_i - p_i) = 45(0-0.5) + 55(1-0.5) + 65(1-0.5) + 35(0-0.5) = 20$$

Step 4: Information matrix

$$I_{00} = \sum_{i=1}^{n} p_i(1-p_i) = 4(0.5)(0.5) = 1$$

$$I_{11} = \sum_{i=1}^{n} x_i^2 p_i(1-p_i) = (45^2 + 55^2 + 65^2 + 35^2)(0.25) = 3075$$

$$I_{01} = \sum_{i=1}^{n} x_i p_i(1-p_i) = (45+55+65+35)(0.25) = 50$$

Step 5: Update parameters

$$\begin{pmatrix} \beta_0^{(1)} \\ \beta_1^{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & 50 \\ 50 & 3075 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 20 \end{pmatrix}$$

After iteration: $\beta_0 \approx -3.2$, $\beta_1 \approx 0.065$
Final model:

$$P(Disease) = \frac{1}{1 + e^{-(-3.2 + 0.065 \times Age)}}$$

4.4 Odds Ratios

The odds ratio for a one-unit increase in x is:

$$OR = e^{\beta_1}$$

For our example: $OR = e^{0.065} = 1.067$
This means for each additional year of age, the odds of heart disease increase by 6.7%.

4.5 Confusion Matrix and Model Evaluation

The confusion matrix is fundamental for evaluating binary classification models. It provides a detailed breakdown of correct and incorrect predictions.

Definition 1 (Confusion Matrix)**.** *For a binary classification problem with classes 0 (Negative) and 1 (Positive):*

		Predicted	
		Negative (0)	*Positive (1)*
Actual	*Negative (0)*	*TN*	*FP*
	Positive (1)	*FN*	*TP*

where:

- ***TP*** *(True Positives): Correctly predicted positive cases*
- ***TN*** *(True Negatives): Correctly predicted negative cases*
- ***FP*** *(False Positives): Incorrectly predicted as positive (Type I error)*
- ***FN*** *(False Negatives): Incorrectly predicted as negative (Type II error)*

4.5.1 Evaluation Metrics

From the confusion matrix, we can calculate various performance metrics:

Formula 5 (Key Evaluation Metrics)**.**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision\ (PPV) = \frac{TP}{TP + FP} \tag{2}$$

$$Recall\ (Sensitivity,\ TPR) = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity\ (TNR) = \frac{TN}{TN + FP} \tag{4}$$

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

$$False\ Positive\ Rate = \frac{FP}{TN + FP} = 1 - Specificity \tag{6}$$

4.5.2 Manual Calculation Example

Example 4 (Heart Disease Classification Results)**.** *Using our heart disease model with threshold = 0.5:*

Age	*Actual Disease*	*Predicted Probability*
45	*0*	*0.25*
55	*1*	*0.61*
65	*1*	*0.85*
35	*0*	*0.12*

Step 1: Convert probabilities to predictions (threshold = 0.5)

Age	*Actual*	*Predicted*
45	*0*	*0*
55	*1*	*1*
65	*1*	*1*
35	*0*	*0*

Step 2: Create confusion matrix

	Predicted 0	*Predicted 1*
Actual 0	*TN = 2*	*FP = 0*
Actual 1	*FN = 0*	*TP = 2*

Step 3: Calculate metrics

$$\textit{Accuracy} = \frac{2+2}{2+2+0+0} = \frac{4}{4} = 1.00 \tag{7}$$

$$\textit{Precision} = \frac{2}{2+0} = \frac{2}{2} = 1.00 \tag{8}$$

$$\textit{Recall} = \frac{2}{2+0} = \frac{2}{2} = 1.00 \tag{9}$$

$$\textit{Specificity} = \frac{2}{2+0} = \frac{2}{2} = 1.00 \tag{10}$$

$$\textit{F1-Score} = \frac{2 \times 1.00 \times 1.00}{1.00 + 1.00} = 1.00 \tag{11}$$

4.5.3 When to Use Each Metric: Practical Guidelines

The choice of evaluation metric depends on the specific problem context and consequences of different types of errors:

1. Accuracy

- **Use when:** Classes are balanced and false positives/negatives have similar costs
- **Example:** General classification tasks with balanced datasets
- **Limitation:** Misleading with imbalanced classes (e.g., 95% accuracy in cancer screening where 95% are healthy)

2. Precision (Positive Predictive Value)

- **Use when:** False positives are costly or problematic
- **Example:** Email spam detection (don't want important emails marked as spam)
- **Example:** Criminal justice (minimize innocent people wrongly convicted)
- **Focus:** "Of all positive predictions, how many were actually correct?"

3. Recall (Sensitivity, True Positive Rate)

- **Use when:** False negatives are costly or dangerous
- **Example:** Medical diagnosis (don't miss actual diseases)
- **Example:** Fraud detection (don't miss actual fraud cases)
- **Example:** Security threats (don't miss actual threats)
- **Focus:** "Of all actual positive cases, how many did we catch?"

4. Specificity (True Negative Rate)

- **Use when:** Correctly identifying negatives is crucial
- **Example:** Drug testing (important to correctly identify non-users)
- **Example:** Quality control (correctly identify non-defective products)
- **Focus:** "Of all actual negative cases, how many were correctly identified?"

5. F1-Score

- **Use when:** Need balance between precision and recall
- **Example:** Information retrieval systems
- **Example:** Imbalanced datasets where both false positives and false negatives matter
- **Focus:** Harmonic mean provides balanced view when precision/recall trade-off exists

Real-World Scenario Analysis:

Example 5 (COVID-19 Testing)**.** *Consider different testing scenarios:*
Scenario A: Airport Screening

- *Priority: High* ***Recall*** *(catch all infected passengers)*
- *Acceptable: Lower precision (some false positives can be retested)*
- *Rationale: Missing an infected person (FN) could cause outbreak*

Scenario B: Return-to-Work Testing

- *Priority: High* ***Precision*** *(avoid unnecessary quarantine)*
- *Balanced: Reasonable recall (with follow-up testing)*

- *Rationale: False positives cause economic disruption*

Scenario C: Research Study

- *Priority: High* ***Accuracy*** *and* ***Specificity***
- *Rationale: Need reliable population statistics*

Trade-offs and Threshold Selection:
The choice of classification threshold (default 0.5) affects all metrics:

- **Lower threshold** (e.g., 0.3): Higher recall, lower precision
- **Higher threshold** (e.g., 0.7): Higher precision, lower recall
- **ROC curve analysis** helps find optimal threshold for specific requirements

4.6 Youden's J Statistic and Optimal Threshold Selection

When selecting an optimal threshold for binary classification, we need a systematic approach that balances sensitivity and specificity. Youden's J statistic provides an objective method for threshold selection.

Definition 2 (Youden's J Statistic)**.** *Youden's J statistic (also called Youden's Index) is defined as:*

$$J = Sensitivity + Specificity - 1 = TPR + TNR - 1$$

Equivalently:

$$J = Sensitivity - (1 - Specificity) = TPR - FPR$$

where:

- TPR = *True Positive Rate (Sensitivity, Recall)*
- FPR = *False Positive Rate* $= 1 -$ *Specificity*
- J *ranges from 0 (no discriminatory ability) to 1 (perfect discrimination)*

4.6.1 Geometric Interpretation

On the ROC curve, Youden's J statistic represents the **maximum vertical distance** from the ROC curve to the diagonal line (line of no discrimination). This point is furthest from the diagonal and represents the optimal balance between sensitivity and specificity.

Key Properties:

- The diagonal line represents random guessing (TPR = FPR)
- Points above the diagonal indicate better-than-random performance
- The optimal threshold maximizes $J = TPR - FPR$
- This gives equal weight to sensitivity and specificity

4.6.2 Manual Calculation of Optimal Threshold

Example 6 (Heart Disease Threshold Optimization)**.** *Let's expand our heart disease example with more data points and multiple thresholds:*

Extended Dataset:

Patient	*Age*	*Actual Disease*	*Predicted Probability*
1	*35*	*0*	*0.12*
2	*40*	*0*	*0.19*
3	*45*	*0*	*0.25*
4	*50*	*1*	*0.43*
5	*55*	*1*	*0.61*
6	*60*	*0*	*0.72*
7	*65*	*1*	*0.85*
8	*70*	*1*	*0.92*

Step 1: Test multiple thresholds
Threshold = 0.3:

Patient	*Actual*	*Predicted*	*Classification*
1	*0*	*0*	*TN*
2	*0*	*0*	*TN*
3	*0*	*0*	*TN*
4	*1*	*1*	*TP*
5	*1*	*1*	*TP*
6	*0*	*1*	*FP*
7	*1*	*1*	*TP*
8	*1*	*1*	*TP*

Results: TP=4, TN=3, FP=1, FN=0

$$\textit{Sensitivity} = \frac{4}{4+0} = 1.00$$

$$\textit{Specificity} = \frac{3}{3+1} = 0.75$$

$$J_{0.3} = 1.00 + 0.75 - 1 = 0.75$$

Threshold = 0.5:

Patient	*Actual*	*Predicted*	*Classification*
1	*0*	*0*	*TN*
2	*0*	*0*	*TN*
3	*0*	*0*	*TN*
4	*1*	*0*	*FN*
5	*1*	*1*	*TP*
6	*0*	*1*	*FP*
7	*1*	*1*	*TP*
8	*1*	*1*	*TP*

Results: TP=3, TN=3, FP=1, FN=1

$$\textit{Sensitivity} = \frac{3}{3+1} = 0.75$$

$$Specificity = \frac{3}{3+1} = 0.75$$
$$J_{0.5} = 0.75 + 0.75 - 1 = 0.50$$

Threshold = 0.7:

Patient	*Actual*	*Predicted*	*Classification*
1	*0*	*0*	*TN*
2	*0*	*0*	*TN*
3	*0*	*0*	*TN*
4	*1*	*0*	*FN*
5	*1*	*0*	*FN*
6	*0*	*1*	*FP*
7	*1*	*1*	*TP*
8	*1*	*1*	*TP*

Results: TP=2, TN=3, FP=1, FN=2

$$Sensitivity = \frac{2}{2+2} = 0.50$$

$$Specificity = \frac{3}{3+1} = 0.75$$
$$J_{0.7} = 0.50 + 0.75 - 1 = 0.25$$

Step 2: Compare J statistics

Threshold	*Sensitivity*	*Specificity*	*J Statistic*
0.3	*1.00*	*0.75*	*0.75*
0.5	*0.75*	*0.75*	*0.50*
0.7	*0.50*	*0.75*	*0.25*

Conclusion: *The optimal threshold is 0.3, which maximizes Youden's J statistic at 0.75.*

4.6.3 When to Use Youden's J Statistic

Appropriate Scenarios:

- When sensitivity and specificity are equally important
- Initial screening tests where balanced performance is desired
- Research settings where objective threshold selection is needed
- When no domain-specific cost information is available

Limitations:

- Assumes equal costs for false positives and false negatives
- May not be optimal when class imbalance is severe
- Doesn't account for prevalence of the condition
- May not align with clinical or business priorities

4.6.4 Alternative Threshold Selection Methods

While Youden's J is popular, other methods may be more appropriate depending on the context:

1. Cost-Based Optimization: When false positives and false negatives have different costs:

$$\text{Total Cost} = C_{FP} \times FP + C_{FN} \times FN$$

Choose threshold that minimizes total cost.

2. F1-Score Maximization: When precision and recall balance is important:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3. Prevalence-Adjusted Thresholds: When disease prevalence differs from training data, adjust threshold accordingly.

4. Clinical Decision Points: Use domain knowledge to set thresholds based on acceptable miss rates or false alarm rates.

4.6.5 Practical Implementation Steps

Step-by-step process for threshold optimization:

1. Generate predicted probabilities for all test cases
2. Create a list of unique threshold candidates (often all unique predicted probabilities)
3. For each threshold:
   - Convert probabilities to binary predictions
   - Calculate confusion matrix
   - Compute sensitivity, specificity, and J statistic
4. Select threshold with maximum J statistic
5. Validate performance on independent test set
6. Consider domain constraints and adjust if necessary

This systematic approach ensures that the selected threshold optimally balances the model's ability to correctly identify both positive and negative cases, providing a mathematically sound foundation for binary classification decisions.

5 Statistical Tests

6 Correlation Analysis

6.1 Pearson Correlation Coefficient

Formula 6 (Pearson Correlation)**.**

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

6.2 Manual Calculation

Example 7 (House Price vs Rooms)**.** *Data:*

Rooms (x)	*Price (y, $000)*
5.5	*180*
6.2	*220*
4.8	*150*
7.1	*280*
5.9	*200*

Step 1: Calculate means

$$\bar{x} = \frac{5.5 + 6.2 + 4.8 + 7.1 + 5.9}{5} = 5.9$$

$$\bar{y} = \frac{180 + 220 + 150 + 280 + 200}{5} = 206$$

Step 2: Calculate deviations and products

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
5.5	*180*	*-0.4*	*-26*	*10.4*	*0.16*	*676*
6.2	*220*	*0.3*	*14*	*4.2*	*0.09*	*196*
4.8	*150*	*-1.1*	*-56*	*61.6*	*1.21*	*3136*
7.1	*280*	*1.2*	*74*	*88.8*	*1.44*	*5476*
5.9	*200*	*0.0*	*-6*	*0.0*	*0.00*	*36*
			Totals:	*165.0*	*2.90*	*9520*

Step 3: Calculate correlation

$$r = \frac{165.0}{\sqrt{2.90 \times 9520}} = \frac{165.0}{\sqrt{27608}} = \frac{165.0}{166.16} = 0.993$$

Strong positive correlation between rooms and price.

6.3 Hypothesis Testing for Correlation

Test: $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$

Test statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

For our example:

$$t = \frac{0.993\sqrt{5-2}}{\sqrt{1-0.993^2}} = \frac{0.993 \times 1.732}{0.118} = 14.60$$

With $df = 3$, $t_{0.025,3} = 3.182$. Since $|14.60| > 3.182$, we reject H_0.

7 t-Test

7.1 One-Sample t-Test

Tests whether a sample mean differs significantly from a hypothesized population mean.

Formula 7 (One-Sample t-Test)**.**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

where $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

7.2 Manual Calculation

Example 8 (Testing Average House Price)**.** *Test:* $H_0 : \mu = 200$ *vs* $H_1 : \mu \neq 200$
Sample data: 180, 220, 150, 280, 200 (from previous example)
Step 1: Calculate sample statistics

$$\bar{x} = 206, \quad n = 5$$

$$s^2 = \frac{(180-206)^2 + (220-206)^2 + (150-206)^2 + (280-206)^2 + (200-206)^2}{5-1}$$

$$s^2 = \frac{676 + 196 + 3136 + 5476 + 36}{4} = \frac{9520}{4} = 2380$$

$$s = \sqrt{2380} = 48.79$$

Step 2: Calculate test statistic

$$t = \frac{206 - 200}{48.79/\sqrt{5}} = \frac{6}{21.82} = 0.275$$

Step 3: Make decision *With* $df = 4$, $t_{0.025,4} = 2.776$. *Since* $|0.275| < 2.776$, *we fail to reject* H_0.

7.3 Two-Sample t-Test

Compares means of two independent groups.

Formula 8 (Two-Sample t-Test (Equal Variances))**.**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

where $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$ *(pooled standard deviation)*

Example 9 (Comparing Two Groups)**.** *Group 1 (High rooms): 220, 280, 250 Group 2 (Low rooms): 180, 150, 160*

Step 1: Calculate group statistics

$$\bar{x}_1 = 250, \quad s_1^2 = \frac{(220-250)^2 + (280-250)^2 + (250-250)^2}{2} = 900$$

$$\bar{x}_2 = 163.33, \quad s_2^2 = \frac{(180-163.33)^2 + (150-163.33)^2 + (160-163.33)^2}{2} = 255.56$$

Step 2: Calculate pooled variance

$$s_p^2 = \frac{(3-1)\times 900 + (3-1)\times 255.56}{3+3-2} = \frac{1800+511.12}{4} = 577.78$$

$$s_p = 24.04$$

Step 3: Calculate test statistic

$$t = \frac{250 - 163.33}{24.04\sqrt{\frac{1}{3}+\frac{1}{3}}} = \frac{86.67}{19.62} = 4.42$$

With $df = 4$, $t_{0.025,4} = 2.776$. *Since* $4.42 > 2.776$, *we reject* H_0 *(significant difference).*

8 Chi-Square Test of Independence

8.1 Theory

Tests whether two categorical variables are independent.

Formula 9 (Chi-Square Test Statistic)**.**

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

where $E_{ij} = \frac{(row\ total_i)(column\ total_j)}{n}$

8.2 Manual Calculation

Example 10 (Smoking vs Time of Day)**.** *Observed frequencies:*

	Smoker	*Non-smoker*	*Total*
Lunch	*15*	*85*	*100*
Dinner	*25*	*75*	*100*
Total	*40*	*160*	*200*

Step 1: Calculate expected frequencies

$$E_{11} = \frac{100\times 40}{200} = 20, \quad E_{12} = \frac{100\times 160}{200} = 80$$

$$E_{21} = \frac{100\times 40}{200} = 20, \quad E_{22} = \frac{100\times 160}{200} = 80$$

Step 2: Calculate chi-square statistic

$$\chi^2 = \frac{(15-20)^2}{20} + \frac{(85-80)^2}{80} + \frac{(25-20)^2}{20} + \frac{(75-80)^2}{80}$$

$$\chi^2 = \frac{25}{20} + \frac{25}{80} + \frac{25}{20} + \frac{25}{80} = 1.25 + 0.3125 + 1.25 + 0.3125 = 3.125$$

Step 3: Make decision *With* $df = (2-1)(2-1) = 1$, $\chi^2_{0.05,1} = 3.841$. *Since* $3.125 < 3.841$, *we fail to reject* H_0 *(no association).*

9 Conclusion

This manual has provided step-by-step calculations for fundamental statistical methods:

- **Linear Regression**: Understanding how to estimate parameters using least squares and evaluate model fit
- **Logistic Regression**: Maximum likelihood estimation for binary outcomes, odds ratio interpretation, comprehensive evaluation using confusion matrices, and optimal threshold selection using Youden's J statistic
- **Correlation**: Measuring linear relationships between variables
- **t-Tests**: Comparing means within and between groups
- **Chi-Square Tests**: Testing independence in categorical data

Key Takeaways:

1. Manual calculations deepen understanding of underlying mathematics
2. Each method has specific assumptions that must be verified
3. Effect sizes are as important as significance tests
4. Interpretation requires domain knowledge and statistical literacy
5. Model evaluation requires choosing appropriate metrics based on problem context and cost of different error types
6. Threshold selection should be systematic and context-appropriate, with Youden's J statistic providing an objective starting point when equal weight is given to sensitivity and specificity

By working through these calculations by hand, you develop intuition for how these statistical methods work and when to apply them appropriately.