ASSIGNMENT QUESTION 2

NORMAL DISTRIBUTION & EMPIRICAL RULE

Domain: Human Adult Heights (in cm)

📝 **ASSIGNMENT CONTENT**

1. DOMAIN SELECTION: Human Adult Heights

Why this domain?
- Human heights in a large population follow an approximately normal distribution [[6]]
- Real-world, relatable example for statistical concepts
- Well-documented mean and standard deviation values

Sample Data Parameters:

Parameter	Value (Men)	Value (Women)
Mean (μ)	175 cm	162 cm
Standard Deviation (σ)	7 cm	6.5 cm
Distribution Type	Approximately Normal	Approximately Normal

Source: Global health statistics, WHO anthropometric data

2. THE EMPIRICAL RULE (68-95-99.7 RULE)

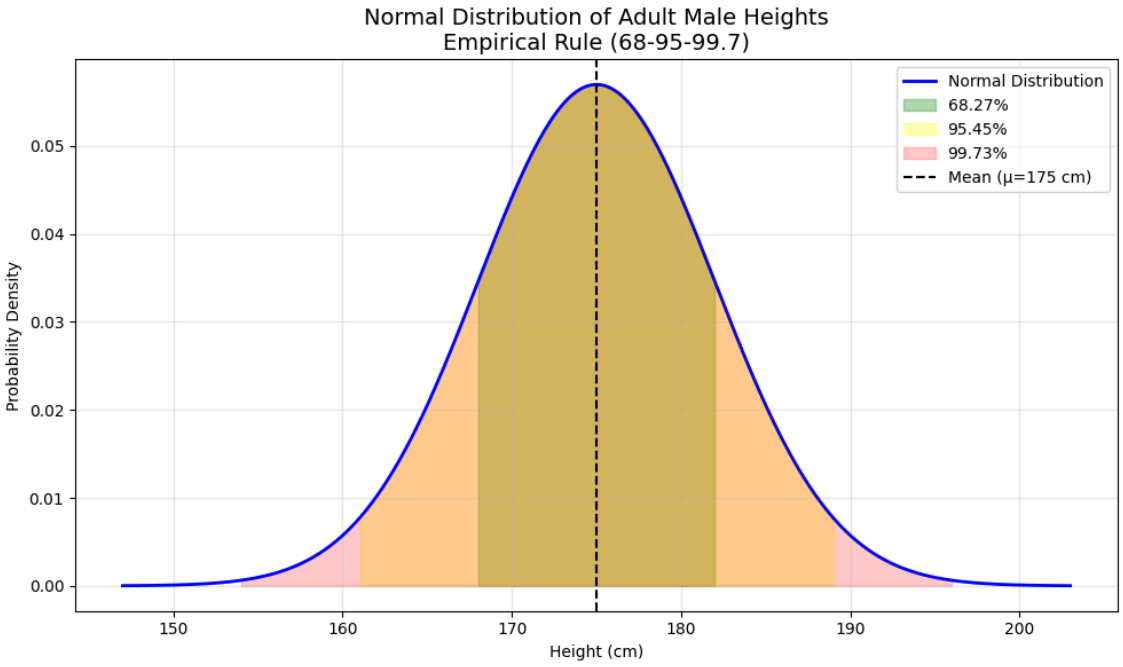
The empirical rule states that for a normal distribution [[8]]:

Standard Deviations from Mean	Percentage of Data	Height Range (Men, μ=175, σ=7)
μ ± 1σ	≈68.27%	168 cm to 182 cm
μ ± 2σ	≈95.45%	161 cm to 189 cm
μ ± 3σ	≈99.73%	154 cm to 196 cm

Mathematical Formulation [[12]]:

```
Pr(μ - 1σ ≤ X ≤ μ + 1σ) ≈ 68.27%
Pr(μ - 2σ ≤ X ≤ μ + 2σ) ≈ 95.45%
Pr(μ - 3σ ≤ X ≤ μ + 3σ) ≈ 99.73%
```

3. VISUAL REPRESENTATION



The sweet spot is at the minimum of the generalization error curve, where bias and variance are balanced optimally.

---

4. MATHEMATICAL FORMULATION

The expected prediction error can be decomposed as:

$$E[(y - \hat{f}(x))^2] = Bias[\hat{f}(x)]^2 + Var[\hat{f}(x)] + \sigma^2$$

Where:
- $Bias[\hat{f}(x)] = E[\hat{f}(x)] - f(x)$
  - Difference between average prediction and true value

- $Var[\hat{f}(x)] = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$
  - Variability of predictions

- σ^2 = Irreducible error (noise)

---

5. PRACTICAL EXAMPLES

Example 1: Linear Regression
- High Bias: Using linear regression on non-linear data
- Low Bias, High Variance: High-degree polynomial regression
- Low Bias, Low Variance: Appropriate degree polynomial with regularization

Example 2: Decision Trees
- High Bias: Very shallow tree (depth = 1 or 2)
- Low Bias, High Variance: Very deep tree (no pruning)
- Low Bias, Low Variance: Optimally pruned tree with cross-validation

Example 3: Neural Networks
- High Bias: Too few layers/neurons
- Low Bias, High Variance: Very large network without regularization
- Low Bias, Low Variance: Appropriate architecture with dropout and regularization

---

6. SUMMARY TABLE

Scenario	Bias	Variance	Training Error	Test Error	Problem
Underfitting	High	Low	High	High	Model too simple
Overfitting	Low	High	Very Low	High	Model too complex
Optimal	Low	Low	Low	Low	Just right ✓
Worst Case	High	High	High	Very High	Both problems

---

7. CONCLUSION

Final Answer: For the best fit model, we should have LOW BIAS AND LOW VARIANCE.

This represents the optimal balance where:
- The model is complex enough to capture the underlying patterns in the data (low bias)
- The model is not so complex that it captures noise and fails to generalize (low variance)
- The model performs well on both training and unseen test data
- The total error is minimized

Achieving this balance requires careful model selection, appropriate regularization, cross-validation, and understanding the tradeoff between model complexity and generalization ability.

The bias-variance tradeoff is one of the most important concepts in machine learning, and mastering it is essential for building models that generalize well to new data.