Hypothesis Testing in AI: Clear Contextual Examples

Introduction

This document presents clear and intuitive examples of hypothesis testing in Artificial Intelligence. Each example includes context, baseline information, sample data, and the correct formulation of null and alternative hypotheses.

1 Model Accuracy Improvement

Context: An image classification model currently achieves an accuracy of 85%. A new CNN architecture is proposed and tested multiple times.

Sample Data (Accuracy in %): 87, 88, 86, 89, 90, 88, 87, 89, 91, 88

Claim: The new model improves accuracy.

$$H_0 : \mu \leq 85$$

$$H_1 : \mu > 85$$

This is a right-tailed test.

2 Training Time Reduction using Larger Batch Size

Context: With batch size 32, the average training time per epoch is 120 seconds. The batch size is increased to 64 to reduce training time.

Sample Data (Time in seconds): 112, 115, 110, 108, 114, 111, 109, 113

Claim: Training time is reduced.

$$H_0 : \mu \geq 120$$

$$H_1 : \mu < 120$$

This is a left-tailed test.

3 Effect of Feature Engineering on Accuracy

Context: A classifier is trained before and after feature engineering.

Sample Data (Accuracy in %):
Before feature engineering: 78, 80, 79, 77, 81, 78, 79
After feature engineering: 82, 84, 83, 85, 81, 84, 83

Claim: Feature engineering changes the model performance.

$$H_0 : \mu_{\text{before}} = \mu_{\text{after}}$$

$$H_1 : \mu_{\text{before}} \neq \mu_{\text{after}}$$

This is a two-tailed test.

4 Reduction in Validation Loss after Hyperparameter Tuning

Context: Baseline validation loss is 0.52. Hyperparameter tuning is applied to reduce the loss.

Sample Data (Loss values): 0.48, 0.50, 0.47, 0.49, 0.46, 0.48

Claim: Validation loss is reduced.

$$H_0 : \mu \geq 0.52$$

$$H_1 : \mu < 0.52$$

This is a left-tailed test.

5 Recall Improvement in Medical Diagnosis Model

Context: A disease detection model has recall 0.78. A new training strategy is introduced.

Sample Data (Recall): 0.81, 0.82, 0.80, 0.83, 0.81, 0.84

Claim: Recall improves.

$$H_0 : \mu \leq 0.78$$

$$H_1 : \mu > 0.78$$

This is a right-tailed test.

6 Optimizer Comparison: Adam vs SGD

Context: Two optimizers are compared for convergence speed.

Sample Data (Epochs):
Adam: 22, 21, 23, 20, 22, 21

SGD: 26, 27, 25, 28, 26, 27
Claim: The convergence speed is different.

$$H_0 : \mu_{\text{Adam}} = \mu_{\text{SGD}}$$
$$H_1 : \mu_{\text{Adam}} \neq \mu_{\text{SGD}}$$

This is a two-tailed test.

7 Effect of Dropout on Test Accuracy

Context: A neural network without dropout has test accuracy of 88%. Dropout is added to improve generalization.
Sample Data (Accuracy in %): 89, 90, 88, 91, 92, 90
Claim: Dropout improves test accuracy.

$$H_0 : \mu \leq 88$$
$$H_1 : \mu > 88$$

This is a right-tailed test.

8 Prediction Error Reduction using Feature Scaling

Context: Mean Absolute Error (MAE) without scaling is 6.5. Feature scaling is applied.
Sample Data (MAE): 6.1, 6.0, 6.2, 5.9, 6.1, 6.0
Claim: Prediction error is reduced.

$$H_0 : \mu \geq 6.5$$
$$H_1 : \mu < 6.5$$

This is a left-tailed test.

9 Bias Detection between Two User Groups

Context: Prediction errors are compared between two user groups.
Sample Data (Error in %):
Group A: 6.2, 6.5, 6.1, 6.4, 6.3
Group B: 5.8, 5.9, 6.0, 5.7, 5.8
Claim: Prediction errors differ between the groups.

$$H_0 : \mu_A = \mu_B$$
$$H_1 : \mu_A \neq \mu_B$$

This is a two-tailed test.

10 Conversion Rate Comparison using A/B Testing

Context: An AI recommendation system is evaluated using A/B testing.
Observed Data:
Control group: 1000 users, 120 conversions Test group: 1000 users, 150 conversions
Claim: The new model increases conversion rate.

$$H_0 : p_T \leq p_C$$

$$H_1 : p_T > p_C$$

This is a right-tailed test.

Key Rule for Students

Always decide the **alternative hypothesis first** based on the claim. The null hypothesis is then written as its logical complement and must include equality.