HYPOTHESIS TESTING ON SUPERSTORE DATA

KARUN RAM

1. Introduction

This short statistical analysis attempts to explain some of the justification behind the features selected for the neural network I built. I attempt to consider features that affect the shipping time of an order. A high level of significance (α) is used to account for the relative homogeneity of the data.

2. Shipping Times Across Categories

Hypotheses:
H_0: The shipping time is equal across product categories.
H_A: The shipping time is not equal across product categories.

To test this, we consider the following distribution (extracted from the data using the code in Stats.py):

Category	Observed Shipping Time (days)	Expected Shipping Time (days)
Bookcases	3.811	3.958
Chairs	3.900	3.958
Labels	4.003	3.958
Tables	3.893	3.958
Storage	3.975	3.958
Furnishings	3.961	3.958
Art	4.054	3.958
Phones	4.001	3.958
Binders	4.022	3.958
Appliances	3.989	3.958
Paper	3.888	3.958
Accessories	3.886	3.958
Envelopes	4.016	3.958
Fasteners	3.977	3.958
Supplies	4.016	3.958
Machines	3.748	3.958
Copiers	3.618	3.958

Given this data, we can perform a chi-squared test for goodness of fit with $df = 16$ with $\alpha = 0.20$. Using the distribution above, we obtain $\chi^2_{df=16} = 0.05672$.

Since $P(\chi^2 > 0.05672) \approx 1 > \alpha$, we fail to reject H_0 and conclude that the assumption equal shipping time across product categories cannot be rejected.

3. Shipping Times Across Regions

Hypotheses:
H_0: The shipping time is equal across regions.
H_A: The shipping time is not equal across regions.

To test this, we consider the following distribution (extracted from the data using the code in Stats.py):

Region	Observed Shipping Time (days)	Expected Shipping Time (days)
South	3.958	3.958
West	3.93	3.958
Central	4.058	3.958
East	3.909	3.958

Given this data, we can perform a chi-squared test for goodness of fit with $df = 3$ with $\alpha = 0.20$. Using the distribution above, we obtain $\chi^2_{df=3} = 0.00333$.

Since $P(\chi^2 > 0.00333) \approx 0.999 > \alpha$, we fail to reject H_0 and conclude that the assumption of equal shipping time across regions cannot be rejected.

4. Shipping Times Across Shipping Modes

Hypotheses:
H_0: The shipping time is equal across shipping modes.
H_A: The shipping time is not equal across shipping modes.

To test this, we consider the following distribution (extracted from the data using the code in Stats.py):

Shipping Mode	Observed Shipping Time (days)	Expected Shipping Time (days)
Second Class	3.238	3.958
Standard Class	5.007	3.958
First Class	2.183	3.958
Same Day	0.044	3.958

Given this data, we can perform a chi-squared test for goodness of fit with $df = 3$ with $\alpha = 0.20$. Using the distribution above, we obtain $\chi^2_{df=3} = 5.0754$.

Since $P(\chi^2 > 5.0754) \approx 0.1664 < \alpha$, we reject H_0 and conclude that the assumption of equal shipping time across shipping modes can be rejected. This, intuitively, is reasonable since shipping modes directly impact shipping speed.