OPERATIONS RESEARCH PROJECT

Real Breast Cancer Dataset

Jahnavi Reddy Ganesina
(Hypothesis - I)

Omkar Deepak Gaikwad
(Hypothesis - II)

Table of Contents

CONTENTS

INTRODUCTION

Robust McNemar's Test for Breast Cancer Dataset

The dataset chosen for the McNemar's Test if Real Breast Cancer dataset. This dataset consists of a group of breast cancer patients, who had surgery to remove their tumour.

The dataset consists of the following variables:

1. Age: age at diagnosis (Years)
2. Tumour_Stage: I, II, III
3. Histology: Infiltrating Ductal Carcinoma, Infiltrating Lobular Carcinoma, Mucinous Carcinoma
4. HER2 status: Positive/Negative
5. Surgery_type: Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy, Other
6. Patient_Status: Alive/Dead [can be null, in case the patient didn't visited again after the surgery and there is no information available whether the patient is alive or dead].

McNemar's Test

Let us consider casual inference for binary outcomes. Here, we will use McNemar's test for matched pairs with binary outcomes. Again, we will consider the overlap problem, where the distribution being considered is the overlap between high-density regions of the treatment and control populations. Again, for the ATT problem, we could introduce balance constraints or simplify the problem by forcing the formulation to use all treatment units. The P-value is 1 – ϕ(z) for sufficiently large samples, and z can be calculated by the following equation, where B and C are counts of discordants/untied responses:

$$z = \left[\frac{B - C - 1}{\sqrt{B + C}}\right]$$

Here, one can think of B as the number of pairs where the outcome from the treated patient was "Yes" and the outcome for the untreated patient was "No." C is the number of pairs where the outcome from the treated patient was "No," and the outcome from the untreated patient was "Yes." The formulation below optimizes the causal effect, measured by how much larger the untied responses B are than the untied responses C, relative to the total number of untied responses B + C. The formulation also maximizes/minimizes the P-value by minimizing/maximizing z. We define the following parameters to formulate the model:

m	the total number of untied responses. We loop over all possible values of m, until the solution becomes infeasible
Ti	the outcome of a treated observation i in the treatment group
Cj	the outcome of a control observation j in the control group
distij	the ijth element of a matrix. It takes the value 1 if the covariates of treated observation i and control observation j are similar enough to be a possible matched pair, otherwise 0.
bij	1 if aij and Cj are equal to 1 and Ti is equal to 0, otherwise 0 (first type of discordant pair)
cij	1 if aij and Ti are equal to 1 and Cj is equal to 0, otherwise 0 (second type of discordant pair)
aij	is a binary variable that is 1 if i and j are in the same pair, otherwise 0

One can show that the number of pairs where the same outcome is realized for treatment and control is irrelevant, so we allow it to be chosen arbitrarily, with no constraints or variables defining it. The total number of pairs n is also not relevant for this test. Therefore, we choose only the total number of untied responses m. Finally, the formulation becomes:

Maximize/Minimize :

$$z(a) = \left[\frac{B - C - 1}{\sqrt{B + C}}\right]$$

Subject to

$$b_{ij} = a_{ij} C_j (1 - T_i) \qquad \forall i, j \quad \text{(Defines } b_{ij}\text{)} \tag{12}$$

$$c_{ij} = a_{ij} T_i (1 - C_j) \qquad \forall i, j \quad \text{(Defines } c_{ij}\text{)} \tag{13}$$

$$\sum_{i \in Q} \sum_{j \in R} b_{ij} = B \qquad \text{(Total number of first type of discordant pairs)} \tag{14}$$

$$\sum_{i \in Q} \sum_{j \in R} c_{ij} = C \qquad \text{(Total number of second type of discordant pairs)} \tag{15}$$

$$B + C = m \qquad \text{(Total number of discordant pairs)} \tag{16}$$

$$\sum_{i \in Q} a_{ij} \leq 1 \qquad \forall j \quad \text{(Choose at most one treatment observation)} \tag{17}$$

$$\sum_{j \in R} a_{ij} \leq 1 \qquad \forall i \quad \text{(Choose at most one control observation)} \tag{18}$$

$$a_{ij} \leq \text{dist}_{ij} \qquad \forall i, j \quad \text{(Choose only pairs that are allowed)} \tag{19}$$

$$a_{ij} \in \{0, 1\} \qquad \forall i, j \quad \text{(Defines binary variable } a_{ij}\text{)} \tag{20}$$

$$\text{(Additional user-defined covariate balance constraints.)} \tag{21}$$

Equations (12) and (13) define variables bij and cij. Equations (14) and (15) are used to define variables B and C. To control the total number of untied responses, we incorporate Equation (16). Equations (17) and (18) confirm that only one treated/control unit will be assigned in a single pair. Equation (19) says that the variable aij can take value 1 only if the value of parameter distij is 1. If we add the constraint Σi∈Q Σj∈R aij =T, where T is the total number of treatment points, we will be in the case of estimating the ATT using all treatment points.

Problem Statement:

The dataset comprises information related to breast cancer patients, encompassing features such as patient demographics, tumor characteristics, treatment details, and follow-up status. Two specific hypotheses have been formulated for investigation: the first examines the potential impact of tumor stage on patient status, and the second explores the relationship between HER2 status and patient outcome. The challenge is to analyze the dataset, implement Robust McNemar's Test, and draw conclusions to validate or refute these hypotheses, ultimately contributing insights to the understanding of factors influencing patient outcomes in breast cancer cases.

In this study, we used RBC (Real Breast Cancer) data. Each row in the dataset represents a patient. Each row of data represents a patient; the outcome is whether a patient is alive or dead. The treatment is the Tumor stage. The covariates are Age, Histology, HER2 status, and Surgery Type.

For Hypothesis 1

Here, distij = 1 whenever the difference between the age of a patient in treated unit i control unit j are all less than or equal to 6, and distij = 0 otherwise. For other covariates like histology, tumor stage, and surgery type, distij = 1 whenever the numeric value of the above 3 covariates in treated unit i and control unit j are the same and 0 otherwise.

For Hypothesis 2

Here, distij = 1 whenever the difference between the age of a patient in treated unit i control unit j is less than or equal to 6, and distij = 0 otherwise. For other covariates like histology, HER2 status, and surgery type, distij = 1 whenever the numeric value of the above 3 covariates in treated unit i and control unit j are the same and 0 otherwise.

HYPOTHESIS TESTING I

Treatment group: Patients with HER2 Status - Negative
Control group: Patients with HER2 Status - Positive.

Other columns like age, tumor stage, surgery type and histology will the covariates for this hypothesis testing.

The covariates in the dataset had data which was in text format, and it was converted to numeric format like below:

Histology		**Tumor Stage**		**Surgery Type**	
Infiltrating Lobular Carcinoma	1	I	1	Simple Mastectomy	1
Infiltrating Ductal Carcinoma	2	II	2	Modified Radical Mastectomy	2
Mucinous Carcinoma	3	III	3	Lumpectomy	3
				Other	4

Table1. Conversion of covariates to numeric form for hypothesis 1

Null Hypothesis H0: Breast cancer patients with a negative HER2 status tend to exhibit a higher survival.

Alternate Hypothesis H1: Breast cancer patients with a positive HER2 status tend to exhibit a higher survival.

Treatment Group: Patients with HER2 Status negative.

Control Group: Patients with HER2 Status positive.

Discordant pairs - 10

Model file –
Using McNemar's test, the model file Stage.mod is utilized to define the discordant pairs. The algorithm searches for one-unit pairings from the Treatment and Control Groups where the values of the covariate variables are similar to both the treatment group and the control group.

Run file –
The run file Stage.run uses the model and data files to calculate the Z values, which allows us to conclude our hypothesis. The allowable difference between the Treatment and Control Group covariate values are determined here. As a result, pairs are created to maintain uniform covariate values.

HYPOTHESIS TESTING I - RESULTS

	min	max	Max *P* -value	Min *P* -value
10	-1.581	1.581	0.943060981	0.05693902
11	-1.206	1.206	0.886091255	0.11390875
12	-0.866	0.866	0.806754919	0.19324508
13	-0.555	0.555	0.71055267	0.28944733
14	-0.267	0.267	0.605265419	0.39473458
15	0	0	0.5	0.5

Figure1. Max and Min P-values for Hypothesis 1

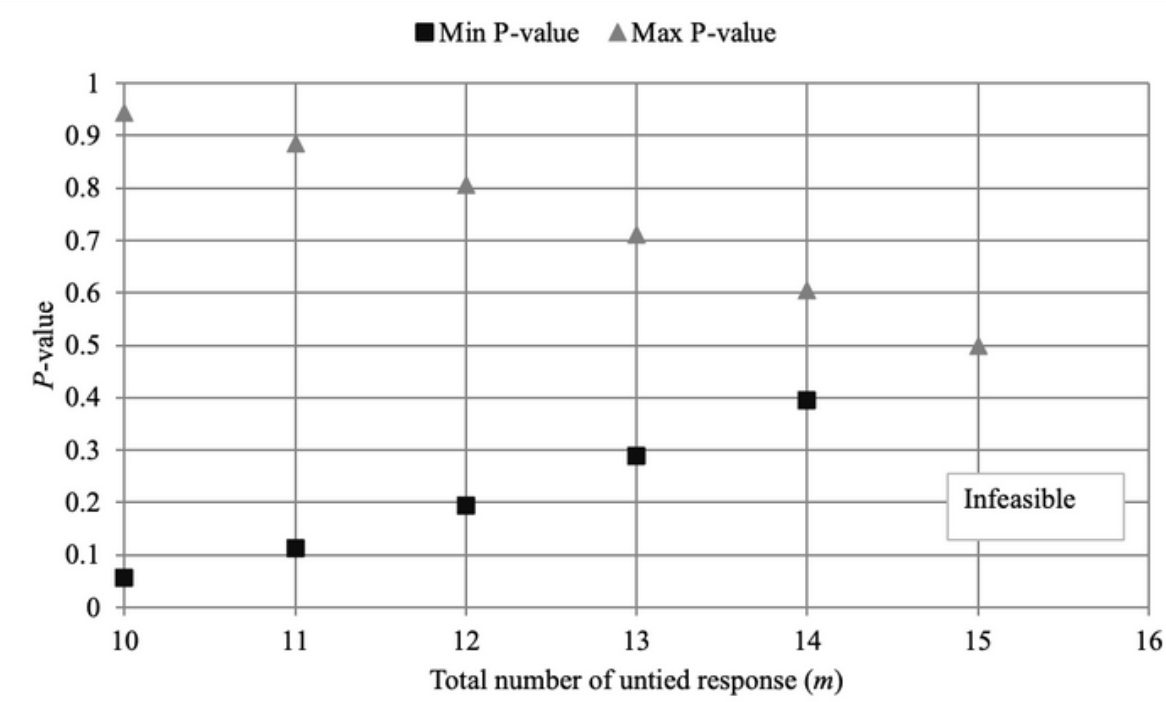

Figure2. P-value vs Total number of untied responses (m) for Hypothesis 1

HYPOTHESIS TESTING II

In this study, we used RBC (Real Breast Cancer) data. Each row in the dataset represents a patient. Each row of data represents a patient; the outcome is whether a patient is alive or dead. The treatment is the Tumor stage. The covariates are Age, Histology, HER2 status, and Surgery Type. Here distij = 1 whenever the difference between age of a patient in treated unit i control unit j are all less than or equal to 6, and distij = 0 otherwise. For other covariates like histology, HER2 status, and surgery type distij = 1 whenever the numeric value of the above 3 covariates in treated unit i and control unit j are the same, and 0 otherwise.

The covariates in the dataset had data which was in text format, and it was converted to numeric format like below:

Histology		**HER2 STATUS**		**Surgery Type**	
Infiltrating Lobular Carcinoma	1	Positive	1	Simple Mastectomy	1
Infiltrating Ductal Carcinoma	2	Negative	0	Modified Radical Mastectomy	2
Mucinous Carcinoma	3			Lumpectomy	3
				Other	4

Table2. Conversion of covariates to numeric form for hypothesis 2

Null Hypothesis (H0): Breast cancer patients who are in Tumor Stage 2 and Tumor Stage 3 tend to have a higher mortality rate.

Alternative Hypothesis (H1): Breast cancer patients who are in Tumor Stage 2 and Tumor Stage 3 do not tend to have a higher mortality rate.

Treatment Group: Individuals who are in Tumor stages 2 and 3.

Control Group: Individuals who are in Tumor stage 1.

Model file –
Using McNemar's test, the model file Stage.mod is utilized to define the discordant pairs. The algorithm searches for one-unit pairings from the Treatment and Control Groups where the values of the covariate variables are similar to both the treatment group and the control group.

Run file –
The run file Stage.run uses the model and data files to calculate the Z values, which allows us to conclude our hypothesis. The allowable difference between the Treatment and Control Group covariate values are determined here. As a result, pairs are created to maintain uniform covariate values.

HYPOTHESIS TESTING II - RESULTS

	min	max	Max *P*-value	Min *P*-value
30	1.278	4.199	0.100624713	1.3405E-05
31	1.437	3.951	0.075358997	3.8913E-05
32	1.591	3.712	0.055804788	0.00010281
33	1.741	3.482	0.040841789	0.00024884
34	1.886	3.258	0.029647477	0.000561
35	2.028	3.043	0.021280124	0.00117116
36	2.167	2.833	0.015117427	0.00230567
37	2.302	2.63	0.010667586	0.00426924
38	2.433	2.433	0.00748715	0.00748715

Figure3. Max and Min P-values for Hypothesis 2

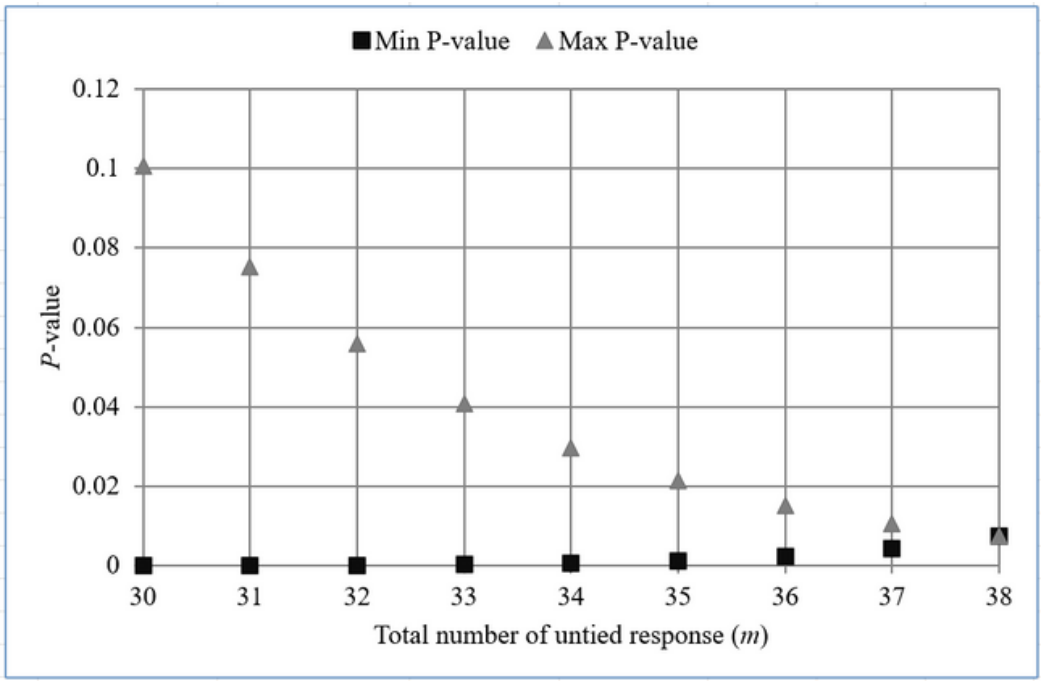

Figure4. P-value vs Total number of untied responses (m) for Hypothesis 2

CONCLUSION

Hypothesis - I

As we can see from the above graph, our P-value is approximately 0.5, which is higher than the significance value of 0.05. Hence, we have enough evidence to accept our Null Hypothesis. Hence, we can conclude that patients with negative HER2 status tend to exhibit a higher survival rate.

Hypothesis - II

As we can see from the above graph, our P-value is approximately 0.007, which is lower than the significance value of 0.05. Hence, we have enough evidence to reject our Null Hypothesis. Hence, we can conclude that patients with Tumor stage 2 and Tumor stage 3 do not have a high mortality rate.

REFERENCES

1) Noor-E-Alam, M. and Rudin, C., "Robust Testing for Causal Inference in Observational Studies", working paper.

2)https://www.kaggle.com/datasets/amandam1/breastcancerdataset