# Replication of Pearson and Spearman Rank Correlations for "Survey data of COVID-19 awareness, knowledge, preparedness, and related behaviors among breast cancer patients in Indonesia" Study

Cicily Balachandar

cicily.balachandar@uconn.edu

## Section 1. Introduction

In the original paper "Survey data of COVID-19 awareness, knowledge, preparedness, and related behaviors among breast cancer patients in Indonesia", the problem being investigated is the association of different factors of COVID-19 in relation to each other. In particular, this paper looks at the association of COVID-19 awareness and knowledge of COVID-19, COVID-19 awareness and preparedness of COVID-19, COVID-19 awareness and related behaviors, knowledge of COVID-19 and preparedness of COVID-19, knowledge of COVID-19 and related behaviors, and preparedness of COVID-19 and related behaviors.

This problem is important because seeing these associations can help determine how to proceed with the education of the pandemic among breast cancer patients in Indonesia, in order to prevent the spread of COVID-19.

The rest of the report will explain what data was obtained and how it was obtained. In addition, the report will detail the statistical methods used, which will be Pearson correlations and Spearman correlations, as well as any assumptions and formulas associated with those statistical methods. Then, the actual analysis of the data will be shown, with the reproduced results. The reproduced results will be compared to the original results and reasons for discrepancies will be mentioned. Furthermore, any assumptions and conclusions from the original paper will be evaluated.

## Section 2. Description of Data

The data was from breast cancer patients in Indonesia and responses were collected through a survey for a sample of 500 patients. In particular, the survey addressed topics related to COVID-19, specifically referencing awareness, preparedness, knowledge, and behaviors.

In the awareness category, the following questions were asked: *"How worried are you about getting the COVID19?", "How worried are you about getting the flu?", "Did you get a flu shot this past year?", "Do you think that you will get sick from the COVID19?*", and *"How likely do you think it is that you or someone you know may get sick from COVID19 this year?"*. For the knowledge category, the corresponding questions included*: "Correctly identified 3 symptoms of the COVID19"* and *"Correctly identified 3 prevention methods of the COVID19"*. The preparedness category included the questions: *"How confident are you that the government can prevent a nationwide outbreak at the COVID19?"* and *"How prepared do you think you are if there were to be a widespread COVID19 outbreak?".* Finally, the related behaviors section included the following questions: *"How much has the COVID19 change your daily routine?"* and *"Are you changing any plans that you have made because at the COVID19?"*.

The study design was cross-sectional with patients selected from three hospitals and the specific sampling used was convenience sampling. Through medical records, the patients selected had breast cancer and have not been infected by COVID-19 (Nindrea et al 3-5).

## Section 3. Methods/Models

The statistical methods used for reproducing the results involve Pearson's correlation coefficient and Spearman's correlation coefficient. First, Pearson's correlation will be discussed. The necessary assumptions for this method are that the two variables are numerical, and with at least one of the variables being normally distributed (Sedgwick). The corresponding equation for calculating this correlation coefficient is as follows (Egghe et al):

$$\mathrm{r} = \frac{n\ \sum_{i=1}^{n} xiyi - (\sum_{i=1}^{n} xi)(\sum_{i=1}^{n} yi)}{\sqrt{n\sum_{i=1}^{n} xi^2 - (\sum_{i=1}^{n} xi)\ ^2}\ \sqrt{n\sum_{i=1}^{n} yi^2 - (\sum_{i=1}^{n} yi\ )\ ^2}}$$

From the above equation, n is the 500 patients, x*i* will correspond to the value of one of the variables x for patient *i* and y*i* will correspond to the variable y (the variable in which x is being compared with) for the same patient *i*. For this case, X is awareness and Y is knowledge for the association of COVID-19 awareness and knowledge of COVID-19, X is awareness and Y is preparedness for the association of COVID-19 awareness and preparedness of COVID-19, and so forth for COVID-19 awareness and related behaviors, knowledge of COVID-19 and preparedness of COVID-19, knowledge of COVID-19 and related behaviors, and preparedness of COVID-19 and related behaviors. For testing the significance of the correlation, the following equation is used (Obilor et al 20):

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

After getting the test statistic from the above equation, if its absolute value is greater than the critical value of the t distribution at a 5% significance level and (n-2) = 498 degrees of freedom, the null hypothesis of ρ = 0 is rejected. Thus, the Pearson correlation coefficient would be significant at the 5% significance level.

Before actually performing these procedures, the data will need to be combined and managed in order to carry out these methods. For this reproduction, the columns that are not explicitly related to COVID were filtered out. In particular, for the COVID-19 Awareness columns, data about the flu and flu shots were not included. The remaining columns ("*How worried are you about getting the COVID19?"* and "*Do you think that you will get sick from the COVID19"*, named (1) and (2), respectively) were treated separately when finding the estimates and conducting significance tests. In the knowledge portion of the data, the two columns ("*Correctly identified 3 symptoms of the COVID19"* and "*Correctly identified 3 prevention methods of the COVID19"*), were combined by creating three levels of very prepared, prepared, and not prepared, where a yes in both columns corresponds to very prepared, only one yes out of both columns corresponds to prepared, and a no in both columns corresponds to not prepared. In the related behaviors portion of the data, only the data from the column "*How much has the COVID19 change your daily routine?"*was used.

After managing the data into a useable format (through R), SAS was used to apply the Pearson correlation method to find the Pearson correlation coefficient and test its significance. Next, the Spearman rank correlation method was performed on the data.

The assumptions for this method include a sample size of N ≥ 20 and at least one of the variables being discrete or being measured based on rank. The equation for the Spearman correlation coefficient is as follows:

$$\mathrm{r_S} = \frac{\sum_{i=1}^{N}(Ri-\bar{R})(Si-\bar{S})}{\sqrt{\sum_{i=1}^{N}(Ri-\bar{R})^2\ \sum_{i=1}^{N}(Si-\bar{S})^2}}$$

where R*i* are the ranks of the corresponding X variable and S*i* are the ranks of the corresponding Y variable. And also $\bar{R} = \sum_{i=1}^{N} Ri/N$ and $\bar{S} = \sum_{i=1}^{N} Si/N$.

The test statistic for significance is below:

$$\mathrm{T} = \frac{r_S\sqrt{N-2}}{\sqrt{1-r_S^2}}$$

where the null hypothesis of $\rho_S = 0$ is rejected, when the absolute value of the test statistic is greater than the critical value for the t distribution at α = 0.05 and 498 degrees of freedom.

SAS was used on the consolidated data to find the Spearman correlation coefficient estimate and to test for its significance.

## Section 4: Analysis of the Data

| X | Y | Pearson r | Pearson p-value | Spearman r | Spearman p-value |
|---|---|---|---|---|---|
| Awareness (1) | Knowledge | 0.57764 | <.0001 | 0.58018 | <.0001 |
| Awareness (2) | Knowledge | 0.33100 | <.0001 | 0.33586 | <.0001 |
| Awareness (1) | Preparedness | 0.06093 | 0.1733 | 0.05273 | 0.2387 |
| Awareness (2) | Preparedness | 0.07101 | 0.1124 | 0.05060 | 0.2583 |
| Awareness (1) | Related Behaviors | 0.35721 | <.0001 | 0.43567 | <.0001 |
| Awareness (2) | Related Behaviors | 0.50424 | <.0001 | 0.54161 | <.0001 |
| Knowledge | Preparedness | 0.53602 | <.0001 | 0.52446 | <.0001 |
| Knowledge | Related Behaviors | 0.66048 | <.0001 | 0.68447 | <.0001 |
| Preparedness | Related Behaviors | 0.57772 | <.0001 | 0.58220 | <.0001 |

From the above table, the corresponding Pearson and Spearman Rank correlation coefficient estimates are shown, along with the p-values for the 5% alpha level tests of significance. Thus, the only nonsignificant correlations are Awareness(1)-Preparedness and Awareness(2)-Preparedness, since their p-values are each greater than 0.05. The strongest correlation coefficient is seen in Knowledge-Related Behaviors, while Awareness(1)-Knowledge, Awareness(2)-Related Behaviors, Knowledge-Preparedness, and Preparedness-Related Behaviors each seem to have a moderately strong correlation.

## Section 5

All of the reproduced Pearson correlation coefficient estimates do not match the original estimates, with each of them, except for Knowledge-Preparedness, being lower than the corresponding estimate from the original study. Also, most of the reproduced correlation coefficients were found to be significant, which was the same as the original study. The only difference is that the reproduced results did not find Awareness(1)-Preparedness and

Awareness(1)-Preparedness to be significant at the 5% significance level. These discrepancies are most likely due to how the data was merged differently compared to the original study.

Looking at the assumptions made for the original study, these assumptions did not justify the use of Pearson's correlation method. There is no evidence that the data is normally distributed and furthermore, the data follows a rank-based system, which makes Spearman's Rank correlation more suitable. Therefore, the conclusions made in the paper are not supported since the wrong methods and tests were used.

## References

Egghe, L., & Leydesdorff, L. (2009, January 29). *The relation between Pearson's correlation coefficient R and Salton's cosine measure*. Association for Information Science & Technology. Retrieved December 4, 2021, from https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.21009.

Nindrea, R. D., Sari, N. P., Harahap, W. A., Haryono, S. J., Kusnanto, H., Dwiprahasto, I., Lazuardi, L., & Aryandono, T. (2020, August 8). *Survey data of covid-19 awareness, knowledge, preparedness and related behaviors among breast cancer patients in Indonesia*. Data in Brief. Retrieved December 4, 2021, from https://www.sciencedirect.com/science/article/pii/S2352340920310398?via%3Dihub.

Obilor, & Amadi. (2018). *Test for Significance of Pearson's Correlation Coefficient (r)*. International Journal of Innovative Mathematics, Statistics & Energy Policies. Retrieved December 4, 2021, from https://www.seahipaj.org/journals-ci/mar-2018/IJIMSEP/full/IJIMSEP-M-2-2018.pdf.

Sedgwick, P. (2012, July 4). *Pearson's correlation coefficient*. The BMJ. Retrieved December 4, 2021, from https://www.bmj.com/content/345/bmj.e4483.full.pdf+html.