Cecil Thompson

Omega Graduate School

COM 968-52 Statistics for Social Research III

Prof. Richard Nti

Dr. Sean Talady

January 17, 2024

Assignment #1 – Core Essential Elements

1. Splitting Dependent Variable Data: In the context of the fictional study, "Religiosity and Social Behavior in a Diverse Community," how can researchers split dependent variable data based on independent variables? Discuss the rationale behind this approach and how it allows for meaningful subgroup comparisons.

2. Conducting Correlational Procedures: Explain the process of conducting correlational procedures with statistical software. What statistical measures can researchers employ to determine the strength and direction of relationships between variables? Interpret the findings of correlational analyses from the fictional dataset.

3. Post - hoc Analysis Interpretation : What statistical tests can be employed when conducting post - hoc analyses to compare multiple groups? How do researchers interpret the results of these tests to identify significant pairwise differences? Discuss the importance of post - hoc analyses in exploring additional research questions beyond the

1. Splitting Dependent Variable Data: A certain researcher poses the question of whether there is a relationship between “Religiosity” of people living in a diverse society and their “Social Behavior”. The question is, if there is indeed a relationship between the two, what might the strength of that relationship be? To begin with, the researcher could select his population from an urban center such as New York City, especially because of its extreme diversity. Before selecting his sample, he might choose to narrow down his population choice to the city’s three most populous and most racially diverse boroughs— say, The Bronx, Brooklyn, Queens. He could, then, draw random samples of 30+ participants from each borough (12 in the fictional study). To guide the sampling selection process, the researcher may formulate a major hypothesis with “Religiosity” of people living in the targeted boroughs of this “racially diverse” city as his independent variable and their “Social Behavior” as the dependent variable. As such, the main hypothesis could read, “There is a statistically significant relationship between the “Religiosity” and the “Social Behavior” of people who live in diverse communities” (Fictional Study).

By providing Columns D, E, and F, the fictional study, has, in effect, split the major dependent variable (Social Behavior) into three dependent variables. Since the dependent variable shown in the major hypothesis above is broad enough to cover different categories of “Social Behavior,” it does make sense to treat it as something of a “universal set” containing several “sub-sets” or “categories” of social behaviors. For that reason, the researcher may choose to split up the major dependent variable (Social Behavior) into three or more categories (Fictional Study).

Here, in the fictional study, the dependent variable has been split into three categories. In this way, the strength of the relationship between each of these new subordinate dependent variables and the original independent variable, in the major hypothesis, should be roughly consistent, if the argument has merit. That is to say that the proposition that “Social Behavior” is a function of “Religiosity” should hold even though the social behavior in a particular category, namely, “Community Service,” may not be applicable to many members of this diverse society. Further, by splitting the dependent variable, the door is opened to more detailed investigation of the thesis through the generation of specific minor related hypotheses (Fictional Study).

1. Conducting Correlation Procedures: It appears that Microsoft Excel may be the most accessible and efficient statistical software for conducting correlational procedures. While Spearman and Pearson are the two prominent correlational instruments, “Pearson Product Moment,” frequently referred to as “Pearson r,” is the go-to correlational instrument for most researchers.

First, the fictional dataset is provided in a typical Microsoft Excel, SPSS- adaptable, worksheet format. The headings of Column A through F, range from “Participants ID”, through “Race,” “Community Service (hours),” “Social Justice Attitudes (1-5),” to “Social Cohesiveness Score (1-10),” respectively. Here, I have transcribed the fictional dataset into my own Excel Book 7, worksheet 1. This includes the 12 samples provided for each column in a form that is SPSS software- ready. In order to determine the strength of the relationship between the independent variable (x) and its pairing with each of the three dependent variables (y1, y2, and y3), the following steps based on McCarty’s (2016-2024) video tutorial are taken:

1. Click the “fx” button in the top left of the Excel screen. On the right, a dialog box appears with the heading, “Formula Builder.”
2. Begin to type “Correlation” into the Search bar. Upon typing the first half of the word, “Correl…” in the formula bar, “Correl” will appear on green background in the bar below.
3. By double clicking on “Correl” written on the green background, two parallel bars will appear; one above the other on the right inside in the “Formula Builder” box. At the same time, the formula, “= Correl ( ),” will appear in a cell on the left, just below the columns on the worksheet. In addition, inside the first of the two bars in the box on the right (the first one labeled, “Array 1”) a flashing cursor is also activated.
4. Beginning with the first row under the heading, “Religiosity (Column C),” drag the cursor down to highlight all twelve samples so that the cell to the left, containing the formula, “= Correl ( ),” is populated with the appropriate references as follow:“= Correl (C2:C13).” At the same time, on the right, the identical references will populate “Array 1”.
5. Click “Array 2” in the Formula Builder box on the right to activate the flashing cursor.
6. Go to Column D [“Community Service (hours)]” and repeat step 4 so that the formula, “= Correl ( ),” is populated as “= Correl (D2:D13),” on the worksheet under “Community Service (Column D)”. To the right, the flashing “Array 2” in the box is automatically populated with the same references on the left under D.
7. Click “Done” and the Correlation Coefficient for the statistical relationship between “Religiosity (the independent variable)” and the specific dependent variable splintered from “Social Behavior (the primary dependent variable)” will pop up (McCarty, 2016-2024).

In the effort to “determine the strength and direction of relationship between variables”, researchers frequently rely upon Pearson r and linear regression, respectively. The value of Pearson r must fall somewhere within the interval of the inequality, -1< 0 < 1. The closer the value of r is to -1 0r 1, the stronger the relationship between the independent (x) and dependent (y) variables.

Whenever the correlation coefficient is1 or -1, there is a 100% relationship between x (the independent variable) and y (the dependent variable). There is a common notion that to consider a 100% relationship between two variables is unrealistic. What I can say is that a 100% relationship between two variables is a “certainty”. If a hypothesis, for example, states that “There is a statistically significant relationship between a fertilized ovum and the birth of a human fetus”, the correlation coefficient from Pearson r would be 1. Since a correlation coefficient of “1” or “-1” points to the existence of a “certainty”, wherever r is equal to zero, the existence of any kind of relationship between x and y would simply be an “impossibility”. When the correlation coefficient is r = 0, the relationship between the variables is similar to that which exists between the proverbial apples and oranges.

The fact that Pearson r can only exist within the interval of the interval, (-1 < 0) v (0 < 1), indicates that all values of r on the left side of the inequality falls within the interval of an infinite number of negative values decreasing from 0 until it becomes bound at, and including -1 in the quadrant II of the Cartesian plane. Meanwhile, all possible values of r on the right side of the inequality will be located in quadrat I with a behavior that is increasing in a manner exactly inverse to that of (-1< 0).

With a Correlation Coefficient of r = 0.55, the statistical relationship between “Religiosity” and “Community Service” stands at 55%. In that case, the strength of the relationship between the independent variable (x) and the dependent variable (y) is moderate. At the same time, the extent of the statistical significance in the relationship between “Religiosity” and “Social Justice” is quite strong at r =0.77, or 77%. The correlation in the third pairing—between “Religiosity” and “Social Cohesiveness”—is also considered to have strong statistical significance at r = 0.76, or 76%. The fact that in all three cases the value of r is positive indicates that both the independent variable (x) and dependent variable (y) are positive. And, by that fact “y increases as x.” Whatever the increase of the x-value, the y-value is likely to increase, on the average of 76% of the increase in x. In such a situation, 76% of the points will cluster in a manner that is close to linearity, so that, a line of regression that “best fits” the clustering of points can be drawn. If, however, either x or y is negative, one of the variables will decrease as the other increases.

1. Post-Hoc Analysis Interpretation: In post-hoc analyses, the situation may be such that it becomes reasonable for one to conduct “significance testing”. For instance, a coin may be flipped for gamblers who are calling heads or tails against the consistently disproportionate calls for tails by the House. If the House knows that the coin is *rigged* to fall “tails” at more than a fifty percent rate unbeknown to the gamblers, the statistician could begin to conduct his test by making the assumption that the coin is a *fair* coin. He may, then, proceed to make bets based on a fair coin hypothesis. This, of course, is an experimental or alternative hypothesis (Ha) which follows the assumption that after a large number of tosses, both heads and tails will result roughly fifty percent of the times. Since the “effect” of a one-half probability is expected for landing either heads or tails on a *fair* coin, the assumption of the probable effect of a *rigged* coin is the null hypothesis (H0).

When a large number of “test statistics” (coin tosses) is done to determine whether the outcome is *rigged* or *fair*, the null hypothesis is assumed and the distribution of all those tosses becomes the “null distribution”. If we compute the percentage of the null distribution to be greater than, or equal to our test statistic, we are finding the *p-value*. Should the *p-value* turn out to be less than the *alpha* value, the null hypothesis will be rejected, and the alpha value will support the “statistical significance” of the experimental hypothesis. But, if the *p-value* is large enough (greater than 0.05), the null hypothesis will be accepted. That means that the coin is rigged. The question, then, would be how much? What we can say is that,

wherever “statistical significance” in favor of the experimental hypothesis is supported, based on the difference between 100% and the *p-value* of 0.05 or less, the likelihood that the null hypothesis is upheld diminishes. And here comes ANOVA! What about it?

To spell out the acronym, ANOVA, we could say simply that ANOVA is the test of “*AN*alysis *O*f *VA*riance”. This, unlike a number of the other post-hoc tests, is designed to analyze the variability of multiple groups of data. As such, it is probably the most effective post hoc test for identifying variances among multiple samples, as well as, those within the same samples. ANOVA, unlike the t-test, may be used to compare the variances of three or more groups at the same time, while the t-tests can only be used to compare a couple groups at a time (Kenton, 2023).

Like the t-test, there is a one-way ANOVA that is able to test whether there are any statistically significant differences among the means of three or more independent (unrelated) groups. Logically, the two-way ANOVA test appears to have grown out of the one-way ANOVA test, by which fact, it is somewhat like a t-test + + (Kenton, 2023).

The one-way ANOVA test has the “relation” of a “one-to-one correspondence” in which a single independent variable affects a single dependent variable in t-test fashion. In the case of the two-way ANOVA, though, two independent variables are involved (Kenton, 2023).

Based on the fictional database, we may hypothesize that, “There is statistically significant relationship between people’s race and their Social Justice Attitudes,” based on the twelve samples drawn from a typical racially diverse American population, Pearson Product Moment (PPM) yields a correlation coefficient of r = -0.067651. This result allows us to say with some degree of confidence that this is pretty close to a zero statistical relationship between “Race” and “Social Justice Attitudes”. And the scatter plot and the “best-fit” line of regression would not be in the first or third quadrants in the Cartesian plane. At the same time, one wonders what would the strength of the statistical relationship be, if we were to switch the independent variable back from “race” to “religiosity”, while leaving the dependent variable, “Social Justice Attitudes,” in the coupling situation? We could begin with the hypothesis that, “There is statistically significant relationship between people’s Religiosity and their Social Justice Attitudes in a diverse community.” In this case, the correlation coefficient is, r = 0.769283575.

Because the correlation coefficient, here, is quite strong and positive, at 77% in round numbers, it is reasonable to conclude that, whenever “religiosity” increases, “social justice attitudes” also increase at a rate that is almost proportionate. On the other hand, since the correlation coefficient for the relationship between “race” and “social justice attitudes” is not only negative, but extremely weak, at – 7%, there is an extremely disproportionate, decrease-increase ratio of x and y. But the fact that there is such a huge disparity between the two coefficients, it is reasonable to consider whether there might be sampling errors that could necessitate the use of either the t-tests, the f-test, or ANOVA.

If any of the t-tests becomes the choice, the “Degrees of Freedom (DF)” may become a critical factor. The question is, what is meant by “Degree of Freedom”?

Hantin and Botts (2023), first, describe the process by which the “Degree of Freedom” may be verified. They submit that the combination of values in a dataset in which each set of data can be varied separately without the original mean-value being changed is verification of the validity of the degree of freedom.

For example, given the five numbers15, 32, 25,12,16 in a dataset, their mean would be 20. But actually, finding the degree of freedom, itself, is quite a simple process. Since the formula for degree of freedom is DF = N-1, the degree of freedom, here, would be equal to 4. Therefore, any 4 of the five numbers may be varied so long as the value of their mean remains 20. If the following four of the five numbers, 15, 32, 12, 16 are varied to 17,30, 25,10,18, the mean would still be 20. The logical question, here, is what is the statistical purpose of finding the degree of freedom? (Hantin and Botts, 2023).

By knowing the DF for a population or a sample, it is possible to find the “critical value” from a table or online. The beauty of this is that the critical value found in the table determines the statistical significance of the result (Hantin and Botts, 2023).

Today, the use of new tools in lieu of standard test-statistics are becoming more prevalent as viable ways of detecting errors and variances subsequent to the performance of statistical analyses. The use of “meta-analysis” to access conclusions made in the body of research is well-nigh becoming the standard approach for post hoc evaluation of statistical analyses in medical and psychological sciences. The claim is that findings are currently being made for “meta-analyses” to include more accurate “. . . .estimates of the effect of treatment or disease risk-factor than any individual study contributing to the pooled analysis. The examination of variability or heterogeneity in study-results is also a critical outcome” (Haidich, 2010).

So, meta-analyses, now addressing post hoc analyses

interpretations through a purist “effect size” approach, means that they have been combining their findings with those of other statistical results. The reason for this is that results from test statistics (t-test, F, Pearson r, etc.) can easily be changed into “effect size” analyses (Anderson et al, 2003). What then is “Effect size” all about?

According to *Wikipedia*, “. . . [E]ffect size is a value measuring the strength of a relationship between two variables in a population” (wikipedia.org). Because “effect size” behaves so much like Pearson r, t-test, and ANOVA, there is also a growing tendency, in its use by meta-analyses for post hoc analysis interpretation, to combine their findings with those from test statistics.

Cohen (1988) claims that “Like r2 . . . [effect size] reflects the percentage of variability in the dependent variable that can be explained by the independent variable”. Therefore, it seems correct to say that the main goal of post hoc analysis is to test for errors and variances in datasets. And, when these tests have been conducted for two or more groups, “effect size” or any of a number of different test statistics may be used separately or in combination, based on its aptness to the situation (Cohen, 1988).

Cohen (1988) submits that using r2 as a tool for evaluating the extent of the independent variable’s (x’s) “effect” on the dependent variable (y), subsequent to finding the correlation coefficient (r), is one way of establishing that the dependent variable (y) is a function of the independent variable (x). He says that when the correlation coefficient, r, is squared, the result is equal to the percent of the effect that the independent variable (x) has on the dependent variable (y). Based on that thesis, he offers an example of the relationship between the two variables in an experimental hypothesis which is claimed to show how the square of r is used to isolate the effect of x upon y from other extraneous effects not considered under the assumption in the hypothesis (Cohen, 1988).

To exemplify this, he call attention to a correlational study between temperature (x) and aggression (y) where r = 0.85, and r2 = 0.72. He goes on to explain that this result for the square of r shows that the effect from the temperature (x) “accounts for 72% of [the effect on human] aggression”. The remaining 28 percent effect on aggression (y) in the study, then, is not included in the effect that temperature has on aggression (Cohen, 1988)..

What Cohen (1988) tell us is that by so squaring r (0.85), the 0.72 result means that 72% of the relationship between temperature and aggression is the effect that the independent variable (x) by itself has on the dependent variable (y). So, in fact, what r2 shows is that, although the correlation coefficient accurately shows, in the study, that the relationship between temperature (x) and aggression (y) was 0.85, twenty eight percent of that relationship (“effect”) was indicative of the effect of some other, as yet unknown variables not accounted for under the correlation coefficient derived from the Pearson r calculation by itself. That is to say that 28% of what excites aggression is wholly from some unknown “intervening variables” outside of where the needle is sitting on the Fahrenheit scale on any given day (Cohen, 1988).

Further, in Cohen’s post hoc reference to the statistical relationship between independent (x) and dependent (y) variables as an “effect” under r**2**, it is tantamount to his saying that the square of the correlation coefficient brings us to a place where we are forced to reconsider the relationship between x and y as a Pearson r pathway to “causation”. For, the place of value that **r2** holds as a “causal” tool for post hoc conclusions on the Pearson correlation coefficient may be a source of confusion to those of us who are constantly reminding ourselves of the statistical rule of thumb that “correlation is not causation”. But now, when we contemplate Jason Fernando’s (2023) statement that “. . . the *coefficient of determination* is also known as R-squared”, we recognize that there, in the “effect size” approach to correlation coefficient, is “cause and effect” (Cohen, 1988).

Works Cited

Anderson, C. A., Berkowitz, L., Donnerstein, E., Huesmann, L.R., Johnson, J.D., Linz,

D.G., Malamuth, N. M., & Wartella, E. (2003). The influence of media violence on youth. Psychological Science in the Public Interest, 4(3), 81-110.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.)

Hilldale, NJ: Lawrence Earlbaum Associates

Fernando, Jason (2023, October 18). The correlation coefficient: What it is, what it

tells investors. *Investopedia.* Reviewed by Margret James. Fact

checked by Vikki Velasquez.

Haidich, A. B. (2010 Dec). Meta-analysis I medical research *Hippokratia Quarterly*

*Medical Journal*, 14(Suppl 1), 29-37. PMC.

Hantin, Travis and Botts, Vanessa (2023, November, 21). Degrees of freedom:

Formula, uses and examples. *Investopedia.*

Kenton, Will (2023, June 12). Analysis of Variance (ANOVA) explanations, formula,

and applications.

McCarty, Matt (2016-2024). “Using Excel to calculate correlation coefficient to

interpret relationship between variables.” *Youtube* *Video*.

http://alphabednch.com/data/excel-corr.

Reichard, Joshua (2021). Statistics as a language: Overview of statistical tests and

hypothesis testing (tutorial 2). *Youtube* video