**WHICH CLUSTER OF TECHNIQUES SHOULD I USE TO ANALYZE MY QUANTITATIVE DATA?**

**by Simon Moss**

|  |
| --- |
| **Introduction** |

When conducting quantitative research—research that involves numbers—perhaps one of the most challenging tasks is to decide which techniques you should utilize to analyze data. Researchers have developed hundreds, if not thousands, of statistical techniques—such as ANOVAs, multiple regression, cluster analyses, confirmatory factor analysis, structural equation modeling, nonparametric statistics, proportional hazards models, hierarchical linear modeling, ARIMA, spatial autocorrelation, social network analyses, Bayesian statistics, optimal scaling, deep learning, and data mining. Yet, many of these techniques can be divided into 10 to 20 main clusters. For example

* One cluster of techniques—such as independent t-tests, ANOVAs, and MANOVAs, compare distinct groups of people, organizations, specimens—and so forth.
* A second cluster of techniques—such as ARIMAs and spectral analyses—explore data over time.

This document helps many individuals, from research students to experienced researchers, decide which cluster of techniques they should apply to analyze their quantitative data. The instructions are, perhaps, most useful after researchers have developed a preliminary research question—but have not finalized the research design and methods. Other documents or articles can then help researchers decide precisely which statistical techniques to apply.

|  |
| --- |
| **Phase 1: Extract one research question or hypothesis** |

Many studies comprise more than one research question or hypothesis. For example, consider this abstract. How many research questions did this study test?

|  |
| --- |
| To examine whether exercise improves memory, 20 participants were instructed to exercise more frequently and intensely than usual for a month. Another 20 participants were instructed to exercise less frequently and intensely than usual during this time. After a month, all participants completed a test that gauges memory, called the digit span. Participants also completed questions that gauge their exercise routine over the last year. Compared to the other participants, memory was more proficient in the participants who were instructed to exercise more during the last month. Furthermore, the frequency, but not intensity, of exercise over the last year was positively associated with memory performance. |

At first glance, this research might seem to comprise only one research question. But, actually this research comprises at least three research question. The first research question is whether instructing people to exercise more during a month enhances memory. The second research question is whether the frequency with which people exercised over the year improves memory. The third research question is whether the intensity with which people exercised over the year improves memory.

These research questions might seem similar, but could be subjected to different statistical tests. Therefore, each research question, if possible, should be considered separately. If possible, please choose one research question for this exercise.

|  |
| --- |
| **Phase 2: Specify the unit of analysis** |

Typically, each data point corresponds to one person, organization, nation, plant, field, animal, specimen, or object. In these examples, the person, organization, nation, plant, field, animal, specimen, or object is called the unit of analysis.

To illustrate, consider a researcher who wants to assess whether people who eat carrots are happier than people who do not eat carrots. In this example, the unit of analysis is the individual.

In contrast, consider a researcher who wants to assess whether nations that drive on the left side are more unequal than nations that drive on the right side. In this instance, the nation is the unit of analysis. Each data point—such as the side in which people drive and level of inequality—correspond to one nation

So, before you proceed, in your mind, clarify the unit of analysis for your study. Is the unit of analysis a person, group, specimen, and so forth?

|  |
| --- |
| **Phase 3: Determine the number of random variables** |

In most studies, researchers need to randomly select a sample from a broader population. For example, researchers might need to select a random sample of 100 participants from the population of adults in Australia. Or the researchers might select a random sample of 50 specimens from a random sample of animal droppings, and so forth. In this instance, the participants, specimens, and so forth are referred to a random variable.

In some instances, the research comprises two or more random variables. For example, a researcher might choose a random sample of 20 employees from a random sample of 10 companies. In this instance, both the employees and companies are random variables.

Similar, a researcher might select a random sample of 10 specimens from a random sample of 5 fields. In this instance, both the specimens and fields are random variables.

So, before you proceed, in your mind, clarify the number of random variables in your study. Is the number of random variables one or more than one?

|  |
| --- |
| **Phase 4: Determine which cluster of techniques may be relevant** |

The next series of questions helps you determine which cluster you should consider. The first set of questions are designed to clarify whether your data include complications—complications that can be solved only with a specific cluster of techniques.

|  |  |
| --- | --- |
| Techniques that address particular complications | |
| Question to identify complication | Technique that addresses complication |
| * Do the data include information about the location of specific objects or people—such as coordinates on a map or the postcode? * If so, are objects or people that are close in space likely to be more similar in some way than objects or people that are farther away in space—called spatial autocorrelation? * If so, is the research in a field in which this spatial autocorrelation tends to be considered important, such as geography and environmental science? | If the answer to all three questions is yes, consider geo-statistics, geographic information systems and mapping\* |
| * Are you collecting data from each unit—such as each person, animal, or field—at more than one time? * If so, does the number of times exceed about 50 or so? | If the answer to these two questions is yes, consider time series analyses, such as ARIMA or spectral analyses\*  One technique, called the Granger causality test, is especially useful. This test assesses whether one set of data over time forecasts another set of data over time. |
| * Does the number of random variables, as defined earlier, exceed one? | If the answer to this question is yes, consider multi-level modelling\*. This approach is sometimes called hierarchical linear modelling or mixed effects modelling.  Or, if the outcome measure is not numerical, you might need to consider generalized estimating equations. |
| * Do you want to explore which characteristics, conditions, or circumstances affect when some event is likely to unfold—such as when people with some disease are likely to die or when individuals are likely to complete some task | If the answer to this question is yes, consider survival analyses, such as Cox regression or even multi-state modelling |
| * Does your data include a ranking of various units, such as people, animals, objects, and so forth? For example, your participants might have completed a race, and you have ranked the participants from first to last? | If the answer to this question is yes, consider non-parametric tests, such as Mann-Whitney U tests.  Note that   * Nonparametric statistics are not necessary if you can convert this rank to a more precise number—such as the time at which they completed the race? * You might need to conduct non-parametric tests for other reasons as well. |
| * Does your data include information about the relationship between various units—such as people, animals, or objects? For instance, you might have collected data on the degree to which employees in a company like one another | If the answer to this question is yes, consider social network analyses. |
| * Do you want to assess whether some measure, instrument, or procedure accurately predicts one of two possible outcomes—such as whether some tool accurately diagnoses an illness | If the answer to this question is yes, consider ROC curves |
| * Would you like the results of your analyses to also integrate your past expectations or knowledge about the topic. For example, perhaps you want to estimate the percentage of individuals who read Peppa Pig. However, publishers tell you the percentage is almost definitely between 20% and 30%--and somehow you to incorporate this information in the analyses. | If the answer to this question is yes, consider Bayesian statistics \* |

\*. These clusters of analyses can be challenging, because the researcher needs to reach several decisions or the advice on this topic is not as extensive. The clusters that are assigned asterisks can be especially challenging.

To proceed, consider these alternatives

* If this table indicates that you should consider a specific cluster of techniques, search whether or not this site provides more information about this cluster. Or Google this term.
* If the table indicates that you should consider more than one specific cluster of techniques, proceed to the first cluster that was recommended. After reading about this cluster, if you feel this first cluster is not suitable, consider the second cluster, and so forth.
* If the table indicates that you do not need to consider any of these specific clusters of techniques, proceed to the next set of questions

This second set of questions differentiate the most common cluster of techniques.

|  |  |
| --- | --- |
| Common techniques | |
| Questions to differentiate techniques | Relevant cluster of techniques to consider |
| * Do you want to compare distinct groups of units—such as distinct groups of people, animals, nations, specimens, or drugs.   For example, you might want to compare males and females, blondes and brunettes—or four possible combinations of genders and hair colours—on IQ. Or you might want to compare mice injected with vitamin C with mice injected with vitamin D on activity? | If the answer to this question is yes, consider between-subject analyses such as independent t-tests or between-subject ANOVAs. |
| * Do you want to compare the same units—such as people, animals, nations, specimens, or plants—over time.   For example, you might want to examine the memory of individuals before and after exercise | If the answer to this question is yes, consider within-subject analyses, such as dependent t-tests or repeated-measures ANOVAs |
| * Do you want to compare distinct groups of units--such as distinct groups of people, animals, nations, specimens, or drugs—but over time?   For instance, you might want to compare mice injected with vitamin C with mice injected with vitamin D on three separate occasions? On each occasion, you examine the same mice | If the answer to this question is yes, consider mixed-model analyses, such as mixed-model ANOVAs |
| * Do you want to explore the relationship between numeric variables?   For instance, you might want to examine the association between the self-esteem and income of people. Or you might want to examine the association between the water consumption and height of plants. In these examples, a variable is regarded as numerical if each unit is assigned a number—such as income—rather than a category name. | If the answer to this question is yes, consider correlation and regression analyses |
| * Do you want to identify clusters of similar units, such as sets of people with similar personalities or attitudes? | If the answer to this question is yes, consider cluster analyses or Q methodology |
| * Do you want to explore or test measures or instruments?   For example, you might have asked individuals 20 questions about their personality. You now want to ascertain whether these 20 questions can be divided into sets of items, each of which measures one trait | If the answer to this question is yes, consider analyses of scales—such as Cronbach’s alpha, exploratory factor analysis, and confirmatory factor analysis |
| * Do you want to simultaneously test measures and explore the association between these measures | If the answer to this question is yes, consider structural equation modelling\* |
| * Do you want to examine a sequence of variables?   For example, perhaps you want to assess the sequence in which experiences during childhood shapes personality, and personality shapes income. | If the answer to this question is yes, consider path analyses and structural equation modelling\* |

Other approaches are also available, but more suitable to researchers who have already developed advanced knowledge about data analyses. These techniques include

* Deep learning or machine learning
* Data mining, classification trees, regression trees, and random forests
* Optimal scaling