**ANALYSING SMALL SAMPLES - EXAMINING EACH CASE SEPARATELY**

**by Simon Moss**

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

 Researchers often want to compare an intervention or treatment condition to a control condition. One researcher might, for example, want to explore whether Manuka honey enhances IQ and mood in research candidates. Specifically, this researcher administers a survey in which participants indicate

* whether or not they ever consume Manuka honey
* their perceived IQ
* the extent to which they feel anxious, on a 5-point scale
* the extent to which they feel depressed, on a 5-point scale
* the extent to which they feel satisfied with life on a 5-point scale

The following screen displays the data in SPSS, in which each row corresponds to one research candidate. The top rows correspond to individuals who consume Manuka honey. The bottom rows correspond to individuals who do not consume Manuka honey—the control group. To compare these groups, you might conduct a variety of statistical tests, such as a series of independent t-tests, a logistic regression, and so forth



**Limitations of conventional techniques**

Unfortunately, in some instances, these conventional techniques are unsuitable. In particular, these conventional techniques are unsuitable in at least two circumstances, as the following table outlines

|  |  |
| --- | --- |
| Circumstance in which conventional techniques are unsuitable | Reason conventional techniques are unsuitable |
| The sample size is small; the intervention condition comprises fewer than 8 participants or units for example | * The sample size may be too small to generate significant results.
* Indeed, the sample size may be too small to even identify outliers. So, significant results might be ascribed to outliers rather than genuine effects
 |
| The response to the intervention, such as Manuka honey, varies appreciably across participants, specimens, or whatever the unit is. | * The analysis, therefore, might overlook distinct subgroups or patterns
 |

**An alternative method that treats participants as single cases**

In these circumstances, researchers should utilize another helpful approach instead. The remainder of this document clarifies these phases. In essence, researchers need to complete these activities:

* for each participant, specimen, or unit in the intervention condition, compare every measure to the control group using an adjusted t-test
* apply a Bonferroni adjustment to these t-tests
* subject these t values to a cluster analysis—to identify subsets of participant, specimen, or units that generate overlapping outcomes

|  |
| --- |
| **Phase 1: Convert each score to a z value but…** |

During the first phase, the researcher decides to compare each person in the intervention condition—the individuals who consume Manuka honey—to the control condition—the individuals who do not consume Manuka honey. This section first illustrates, and then justifies, this procedure.

**Calculate the means and standard deviations of the control group**

On each measure, such as IQ and anxiety, the researcher calculates the mean and standard deviation of the control group. The reason will become obvious later. To achieve this goal, you first need to select the control group only. This example illustrates the procedure in SPSS. Even if you do not use SPSS, you should be able to follow the principles and apply similar principles using other software.

* Choose “Select cases” from “Data”
* Choose “If condition is satisfied” and then “If”
* In the empty box at the top, enter “Manuka = 0” to choose only the control group.
* Press “Continue” and then “OK”

Then, calculate the mean and standard deviation of each measure in the control group. In SPSS

* Select “Descriptive Statistics” and then “Descriptives” from “Analyze”
* In the box called “Variables”, specify all the measures—such as IQ, anxiety, depression, and life satisfaction.
* Press OK to generate the following output

|  |
| --- |
| **Descriptive Statistics** |
|  | N | Minimum | Maximum | Mean | Std. Deviation |
| IQ | 7 | 87.00 | 113.00 | 100.2857 | 9.99524 |
| Anxiety | 7 | 1.00 | 5.00 | 3.1429 | 1.57359 |
| Depressed | 7 | 1.00 | 5.00 | 2.7143 | 1.60357 |
| Life\_satisfaction | 7 | 1.00 | 5.00 | 2.8571 | 1.57359 |
| Valid N (listwise) | 7 |  |  |  |  |

**Convert each score in the intervention condition to z scores**

 During the next phase, researchers might convert each score in the intervention condition—such as the IQ, anxiety, depression, and life satisfaction of every participant—to a z score. This value represents the extent to which each of these scores differs from the average in the control group. In particular,

$$z=\frac{X-μ}{σ}$$

* In this formula, X refers to any score in the intervention condition
*  refers to the mean of the control condition
*  refers to the standard deviation of the control condition

To illustrate

* in the previous spreadsheet, the IQ of the first person was 105
* in the previous table of results, the mean and standard deviation of IQ in the control group was 100.29 and 9.995 respectively
* thus z = (105-100.29)/9.995 = .47

So, what does this .47 indicate? In general

* if a z value is close to 0—in particular, between -1.96 and 1.96—the corresponding score does not differ significantly from the mean of the control group
* if a z value exceeds 1.96, the corresponding score exceeds the mean of the control group, as the formula implies
* if a z value is less than 1.96, the corresponding score is less than is the mean of the control group
* in this instance, the IQ of Participant 1 does not differ significantly from the mean of the control group.

To compute these z scores for every score in the intervention group, first select the participants in this group. In SPSS, you would

* Choose “Select cases” from “Data”
* Choose “If condition is satisfied” and then “If”
* In the empty box at the top, enter “Manuka = 1” to choose only the intervention group.
* Press “Continue” and then “OK”

Now, to calculate the z scores associated with IQ in the intervention group, choose “Compute variable” from the “Transform” menu to generate the following screen



 To compute the z scores

* in this box called “Target variable”, enter a name for your new column, such as zIQ.
* in the box called “Numerical expression”, specify the formula for this z score.
* in this example, replace the mean and standard deviation with the values you calculated earlier
* Press OK—and then repeat this procedure to create z scores for the other measures
* These columns should now appear in the data file



|  |
| --- |
| **Phase 1a: Convert each score to an adjusted t value** |

**A problem with the z values**

If you examine the previous formula to calculate the z values, you may uncover a problem. This formula refers to  and , symbols that refer to the mean and standard deviation of a population. However, in this example, the researcher utilized the mean and standard deviation of the control group to estimate the mean and standard deviation of the population. Unfortunately,

* the mean and standard deviation of the control group is not the same as the mean and standard deviation of the population—especially when the number of people in the control group is fewer than 100 or so
* in these instances, the mean and standard deviation of the control group is unstable
* that is, if the study was repeated with another sample of people, the mean and standard deviation of the control group may be quite different.

Because of this complication, these z values do not actually conform to a normal distribution. A z value that exceeds 1.96 might not be significant. Fortunately, Crawford and Garwaithe (2002) developed an adjustment to the formula that can be used instead”



 = individual’s score

= mean of the control sample

 = standard deviation of the control sample

 = control group sample size

Thus, rather than calculate the z values, you should instead calculate these adjusted t values. For example, after selecting only participants in the control group, you could choose “Compute variable” from the “Transform” menu to generate the following screen.



To compute these t scores

* in this box called “Target variable”, enter a name for your new column, such as tIQ.
* in the box called “Numerical expression”, specify the formula for this t score.
* in this example, replace the mean and standard deviation with the values you calculated earlier
* the sample size for the control group was 7
* in SPSS, sqrt refers to square root
* Press OK—and then repeat to create t scores for the other measures
* These columns should now appear in the data file



**Determine whether these t values are significant**

The previous activity converted the original scores to t values. But, what do these t values indicate? Which of these t values are significant? To answer this question, you need to compute the p value associated with each t value. Several methods can be applied to estimate these t values. To illustrate, suppose you want to ascertain whether the first t value—the 1.26—is significant. To achieve this goal

* Open Microsoft Excel
* Enter “=tdist(1.26, 6, 2)” in any cell, but without the quotation marks
* The 1.26 is the t value.
* The 6 is called the degrees of freedom and equals the number of participants or units in the control group minus 1.
* In this example, the value that appears is .25.
* Because this value exceeds 0.05. we conclude the IQ of this person did not differ significantly from the mean IQ of the control group
* If this value was less than 0.05, but the t value was positive, we would conclude the IQ of this person exceeds the mean IQ of the control group
* If this value was less than 0.05, but the t value was negative, we would conclude the IQ of this person is less than is the mean IQ of the control group

You could then repeat this procedure for every other t value. If you can use Excel proficiently, you could probably apply a more efficient procedure. That is, you might

* copy all the t values into Excel
* create the formula for one t value
* then copy this formula to other cells to estimate the p value of every other t value.

Using this information, you could even construct a spreadsheet that resembles the following example. In this example

* the + signs correspond to scores that significantly exceed the mean of the control group
* the – signs correspond to scores that are significantly less than is this mean
* the 0 signs correspond to scores that do not differ significantly from this mean
* this table indicates the intervention sometimes improves IQ or mood and sometimes impairs IQ or mood



|  |
| --- |
| **Family-wise Type I error** |

**Bonferroni adjustments**

The previous example implied the intervention—Manuka honey—might affect IQ and mood in some participants. However, before you reach conclusions, you might need to consider and manage two complications. First, some people feel you should not use a p value of 0.05. Instead, they believe you should:

* Divide .05 by the number of participants in the treatment group.
* For example, if your treatment group comprises 5 people, you would divide .05 by 5 to reach .01.
* Consequently, you would conclude a score differs significantly from the control group only if the p value is less than .01

Indeed, some people feel you should divide .05 by the product of both the number of participants in the treatment group and the number of measures. For example:

* If your treatment group comprises 5 people, and you want to compare the groups on 10 measures, you would divide .05 by (5 x 10) to reach .05/50 or .001
* Consequently, you would conclude a score differs significantly from the control group only if the p value is less than .001

To reach a suitable choice, you need to understand the rationale that underpins this decision, sometimes called a Bonferroni adjustment. This rationale is discussed below.

**The rationale of Bonferroni adjustments: Specific hypotheses or conclusions**

Suppose you want to assess the hypotheses or conclusion that “Manuka honey affects IQ”. If alpha is set to .05, the likelihood that each test of this hypothesis or conclusion generates a false significant result is .05 by definition. Therefore,

* if you test this assumption or conclusion once, the likelihood that you will falsely conclude that Manuka honey affects weight is .05 or 5%--a level that scientists feel is acceptable.
* But, if you test this assumption or conclusion separately for each person, the likelihood that you will falsely conclude that Manuka honey affects weight is appreciably greater than .05 or 5%, as the following table demonstrates. Read each column in sequence

|  |  |
| --- | --- |
| If alpha is set to .05 | If alpha is set to .01 |
| And you test this assumption on five people | And you test this assumption on five people |
| The probability that each test generates a false significant result is .05 | The probability that each test generates a false significant result is .01 |
| The probability that at least one of these tests generates a false significant result is about .05 x 5 = .25 or 25% | The probability that at least one of these tests generates a false significant result is about .01 x 5 = .05 or 5% |
| Therefore, if Manuka honey does not affect weight, the probably of a false conclusion is .25—a level that is not acceptable | Therefore, if Manuka honey does not affect weight, the probably of a false conclusion is .05—a level that is acceptable |

Therefore, to assess the assumption that Manuka honey affects weight, you may need to set alpha to .05 divided by the number of participants in the treatment group. Otherwise, the likelihood that you will incorrectly conclude that Manuka honey affects weight is too high.

**The rationale of Bonferroni adjustments: General hypotheses or conclusions**

If you want to assess a more general hypothesis or conclusion, such as “Manuka honey “Manuka honey affects states of mind”, the problem if even graver. In this example, a false significant result on any of the four measures—IQ, anxiety, depression, and life satisfaction—would support this hypothesis or conclusion. Therefore, as the following table shows, to ensure the likelihood of one false significant is .05, you would need to set alpha to .05 / (number of participants x number of relevant measures).

|  |  |
| --- | --- |
| If alpha is set to .05 | If alpha is set to .05 / 20 or .0025 |
| And you test this assumption on five people with four different measures | And you test this assumption on five people with four different measures |
| The probability that each test generates a false significant result is .05 | The probability that each test generates a false significant result is .0025 |
| The probability that at least one of these tests generates a false significant result is about .05 x 5 x 4 = 1.0 or about 100%.  | The probability that at least one of these tests generates a false significant result is about .0025 x 5 x 4 = .05 or 5% |
| Actually, 100% is an exaggeration, reflecting the formula and reasoning is not quite correct. Regardless, this number implies the probability of a false significant result is very high | Therefore, if Manuka honey does not affect states of mind the probably of a false conclusion is .05—a level that is acceptable |

As this table implies, the Bonferroni adjustment is quite conservative. More accurate adjustments have been developed, such as the Holm, Holland-Copenhaver, Rom, and Hommel procedures (e.g., Hochberg, 1988; Holland , 1981; Holland & Copenhaver, 1987, 1988).

**The Holm procedure**

To illustrate a more powerful variant of the Bonferroni adjustment, consider the Holm procedure. To apply this procedure

* arrange the p values from lowest to highest, as the first and second columns show in the following table
* for this procedure, the adjusted alpha is different for each p value
* in particular, the alpha at each position equals alpha or 0.05 divided by (the number of tests - position in the sequence + 1). These values appear in the third column
* for example, in this instance, for the first p value, the adjusted alpha equals .05/(7 - 1 + 1) = .007
* for the second p value, the adjusted alpha equals .05/(7 - 2 + 1) = .008 and so forth
* only p values that are less than is the corresponding alpha are significant—represented by the bold values

|  |  |  |
| --- | --- | --- |
|  Participant | Corresponding p value for their IQ | alpha |
| John | **.002** | .007 |
| Fred | **.004** | .008 |
| Mary | .06 | .01 |
| Jill | .12 | .013 |
| Kate | .16 | .016 |
| Frank | .18 | .025 |
| Louise | .21 | .05 |

|  |
| --- |
| **Cluster analysis** |

 The previous sections demonstrate how you can determine which specific values differ significantly from the control. These results might indicate the intervention—such as Manuka honey—is helpful for some people, at least on specific measures, but not other people. But often, you would like to characterize the patterns of results more precisely. You might, for example, want to show that

* manuka honey improves mood but not IQ for one subset of individuals
* manuka honey damages mood and IQ for another subset of individuals, and so forth

To identify these subsets, a cluster analysis might be helpful. Specifically, to achieve this goal

* select only the participants in the intervention condition
* choose “Classify” and then “K means cluster” from the “Analyze” menu to generate the following screen



Then

* in the box labelled “Number of clusters”, insert “2”, sometimes the default
* in the box labelled “Variables”, enter your t values
* press “save” and tick “cluster membership”, before choosing “Continue” and
“OK”

The output is not especially useful. Instead, more informative information appears in a new column in the data file, called something like QCL\_1, as shown below.



 This column indicates that Participants 1, 3, 5, 6, and 7 belong to one cluster. Participants 2 and 4 belong to another cluster. To interpret these clusters, you could

* compute the mean and standard deviation of each column of t values for each cluster separately
* even conduct a t-test or ANOVA to compare the clusters on each measure
* you might discover, for example, the first cluster tend to reflect high scores on mood
* the second cluster might reflect low scores on all measures, and so forth
* if your sample size was larger, you might even repeat the analysis, but with more than 2 clusters, such as 3 or 4, and then decide which results are most informative.

|  |
| --- |
| **References** |

Crawford, J. R. & Howell, D. C. (1998). Comparing an individual's test score against norms derived from small samples. The Clinical Neuropsychologist, 12, 482-486.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika, 75, 800-802.

Hochberg, Y., & Benjamini, Y. (1990). More powerful procedures for multiple significance testing. Statistics in Medicine, 9, 811-818.

Holland, B. (1991). On the application of three modified Bonferroni procedures to pairwise multiple comparisons in balanced repeated measures designs. Computational Statistics Quarterly, 3, 219-231.

Holland, B., & Cheung, S. H. (2002). Family size robustness criteria for multiple comparison procedures. Journal of the Royal Statistical Society, B, 54, 63-77.

Holland, B. S., & Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. Biometrics, 43, 417-423.

Holland, B. S., & Copenhaver, M. D. (1988). Improved Bonferroni-type multiple testing procedures. Psychological Bulletin, 104, 145-149.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65-70.