Recognizing Multiplicity in Statistical Testing

By Katie Daisey, Ph.D.

Q In the last installment of Data Points, the authors discussed performing multiple equivalence tests and the need to control alpha risk – the risk of erroneously rejecting the null hypothesis.[1] In which other situations does the alpha risk of multiple tests need to be controlled?

A Alpha risk is the probability that a null hypothesis will be rejected when it is in actuality true – more commonly described as the long-term rate of committing a Type I error. In general, whenever multiple simultaneous hypothesis tests are performed, the alpha risk is elevated and some corrective action must be taken. Think of each test as a roll of the dice. Whatever number you want to get or to avoid will eventually come up.

The previous installment of Data Points discussed the application of the Intersection-Union Test (IUT) principle as a method to control alpha risk, but it may be non-obvious to a non-statistician that multiple tests are occurring. Unfortunately, the mathematical meaning of "simultaneous" differs considerably from the general usage, and testing does not necessarily need to be performed at the same time. Instead, simultaneous refers to any hypothesis testing occurring on the same issue or in the same study, which can be difficult to define and identify. Given here are several broad categories of multiple simultaneous hypothesis tests to help the non-statistician identify simultaneous tests.

MULTIPLE STATISTICS

One type of equivalence test is used to show that two testing procedures are equivalent by showing that several statistics describing the data generated by the two tests are not statistically different. Some common calculated statistics include sample mean, standard deviation, range, and bias. Even though the statistics are only describing two data sets, because multiple parameters are being compared, these are multiple simultaneous hypothesis tests. The more statistics that are compared to provide stronger evidence for test equivalence, the higher the likelihood of a Type I error.

MULTIPLE VARIABLES

The second type of equivalence test is used to show a testing procedure is equivalent between multiple testing labs. Assuming just one statistic is compared (usually the mean), each of these lab-to-lab comparisons will result in a large number of simultaneous hypothesis tests.

A common method of comparing multiple variables is the one-way analysis of variance (ANOVA), which by itself consists of a single hypothesis test. Therefore, an unadjusted alpha level, usually via the F test, can be used to control the risk of a Type I error *(e.g. alpha = 0.05)*. However, if the one-way ANOVA reports a significant difference, often multiple post-hoc (follow-on) tests are performed to identify which treatment groups (factors) differ. These post-hoc tests are simultaneous hypothesis tests, and the risk of a Type I error is again elevated. Most statistical software that automatically perform the post-hoc comparison tests use an adjusted alpha in this situation, but that should be confirmed via software documentation. Of note, Microsoft Excel does not automatically perform post-hoc tests, and an adjusted alpha must be manually calculated when using *p*-values from Excel post-hoc tests.

Multiway analysis of variance (MANOVA) is used to test statistical significance of two or more different factors on multiple outcome variables. When a MANOVA is used to confirm a single, preplanned hypothesis, an unadjusted alpha is appropriate, but when a MANOVA is used to explore any possible effect and/or interaction from the various factors, these are multiple simultaneous hypothesis tests, and the alpha risk is elevated.

SUBGROUP ANALYSES

An interesting type of multiple simultaneous hypothesis test is that of subgroup analyses, sometimes commonly referred to as "data slicing" or "salami slicing."[2] In this approach, subsets of the data, typically based upon a categorical factor, are tested one at a time to check for statistical significance of a hypothesis test. This is commonly seen in surveys and in the medical field, where demographics tend to be an important factor, but it can also be seen in areas such as interlaboratory testing (where each individual lab might constitute a subgroup), multiple instrument/tester situations, and multi-grade testing.

Subgroup analyses can be a statistically valid manner of analyzing data, but like an exploratory MANOVA, exploratory subgroup analyses have a heightened alpha risk that must be corrected for. The risk remains, even if all the possible subgroups are not explicitly tested or if a subgroup analysis is performed significantly after the original analysis of the data.

A not-so-obvious variant of subgroup analyses is repeatedly analyzing the same samples over a period of time. Though the physical testing occurs at distant moments in time, for the purposes of testing multiplicity, these are considered simultaneous tests. Control charting and other monitoring procedures are prime examples of this.

SELECTIVE REPORTING OF SAMPLES

A more insidious type of multiple simultaneous hypothesis test is the inclusion or exclusion of samples, especially when performed post-analysis. Excluding (or including) outliers specifically to reach a particular *p*-value of a statistical test is a very clear misuse of statistics and should not be tolerated. Instead, decisions about the method to identify and reject outliers should be made *a priori,* ideally with the goal of limiting the fraction of samples rejected.

In a similar way, stopping an experiment because enough samples have been collected to reach the desired *p*-value, or conversely, collecting *more* samples because the hypothesis test did not reach the desired *p*-value, is also a situation with increased alpha risk, and alpha should be adjusted accordingly. This can be done correctly using the methods of sequential analysis, but this is quite an advanced topic.

MULTIPLE REGRESSION

The last category of this non-exhaustive list is performing multiple simultaneous hypothesis tests to determine variable importance. This situation is commonly encountered when performing multiple linear regression when the statistical significance of each variable needs to be determined and the variables included or excluded from the regression accordingly. While the alpha risk is not elevated in the first stage to determine if *any* of the variables are significant, it is elevated in the second stage, where it is determined which of the variables (and/or variable interactions) are significant and must be appropriately controlled.

LESS POWER OR MORE SAMPLES?

Adjustment for testing multiplicity always involves reducing alpha for the individual tests, as in a Bonferroni adjustment. However, reducing the alpha of individual tests incurs the penalty of reducing the power to detect deviations from the conditions of *their* respective null hypotheses. As was discussed in another previous Data Points column, this decrease in alpha requires an increase in sample size to keep the same power to detect deviations from the conditions of the null hypotheses.[3] If the individual tests are critical to the overall decision process, the overall sample size must be increased accordingly.

SUMMARY

Testing multiplicity occurs in a very wide variety of statistical decision contexts. Whenever hypothesis testing is applied in an exploratory or sequential manner, the risk of committing a Type I error – the alpha risk – is heightened. To perform an increased number of simultaneous hypothesis tests while keeping the same overall probability of committing a Type I error, increasingly smaller alphas must be used for the individual tests.

Multiplicity correction controls the probability that one or more of the simultaneous tests gives a Type I error at the expense of reducing the power of the individual tests, unless the sample size is increased sufficiently to compensate. While there are methods to adjust for high alpha risk, such as discussed in the previous Data Points installment, it is the author's hope that this column will help identify when these tools might need to be applied. ■

REFERENCES

[1] Dobson, J. and Murphy, T.D., "Controlling the Alpha Risk of Multiple Equivalence Tests Using the Intersection-Union Test (IUT) Principle," *Standardization News* (July-August 2022).

[2] Head, M.L. et al. "The Extent and Consequences of p-Hacking in Science," PLOS Biology Vol. 13, No. 3 (2015). https://doi.org/10.1371/journal.pbio.1002106.

[3] Parendo, C., "Power and Sample Size, Part 1," *Standardization News* (March-April 2022).

Katie Daisey, Ph.D., is a scientist at Arkema Inc., supporting R&D and manufacturing in the areas of statistics, chemometrics, and digital transformation. Dr. Daisey currently serves as the vice chair of the committee on quality and statistics (E11).

John Carson, Ph.D., of P&J Carson Consulting LLC, is the Data Points column coordinator. He is chairman of the subcommittee on statistical quality control (E11.30), part of the committee on quality and statistics (E11), and a member of the committee on environmental assessment, risk management, and cor-rective action (E50).