STATISTICAL TESTING RISES TO THE CHALLENGE OF MULTIPLICITY IN GENETIC STUDIES

STATISTICAL SIGNIFICANCE

Statistical tests are invaluable tools used across different fields of scientific research. In many such fields, a critical issue arises when multiple statistical tests need to be simultaneously conducted, such as when seeking to identify interesting genes that can provide insights on certain cancers from among many thousand genes. Statisticians work towards addressing this multiplicity problem wherever it occurs, as a crucial part of the broader effort towards ensuring that statistical tests are used correctly across the sciences.

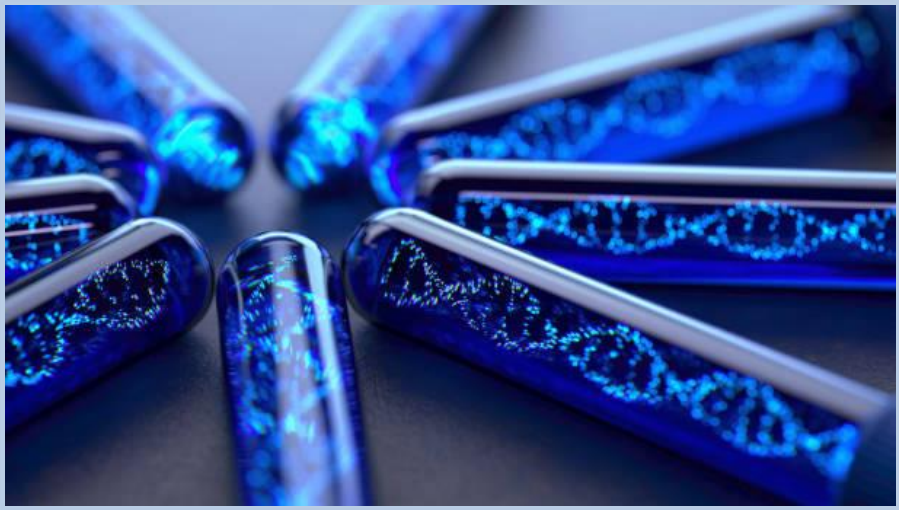
Multiplicity Problem in the Digital Age

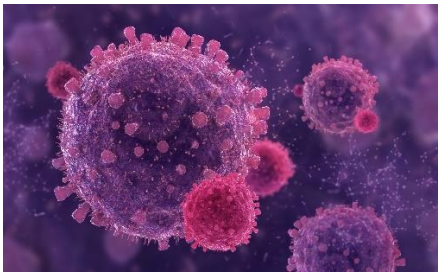
A statistical test is a tool scientists use to inform them about whether a prevailing assumption (the null hypothesis) is reasonable or not given data gathered about it. It does this using the notion of "confidence level," realized through the computation of p-values that allow the scientist to claim, for example, 95% confidence in their decision to reject their null hypothesis. Without going deeper into what confidence means (which is, itself, quite a rich source of discourse), a peculiar thing happens when a scientist performs multiple hypothesis tests at the same level of confidence. What happens is that the true level of confidence for each test is actually lower than that at which the scientist conducted them. This poses a significant problem (pun intended), since it becomes misleading for the scientist to continue to claim that they are rejecting their null hypothesis at their predefined level of confidence. This is known as the multiplicity problem.

The multiplicity problem is not new, and various solutions have been developed to address it, but the problem worsens as the number of hypotheses increases, and for a very large number of hypotheses, traditional solutions become very impractical. Enter the digital age; human civilization is mapping genomes, examining brain structures, and doing many other amazing scientific endeavors that typically involve the conduct of thousands of hypothesis tests!

Differentially Expressed Genes

One research area in genetics is the identification of differentially expressed genes. A gene is differentially expressed if considerable change is observed in it between two experimental conditions. Finding these genes provides crucial insight to scientists, such as narrowing down where to look in explaining the development of particular cancers. Such endeavors require testing thousands of genes, thereby presenting a large-scale multiple testing problem for statisticians to solve.





Customizing Solutions

Oftentimes, statisticians need to come up with solutions that are specifically customized to the research they are collaborating on. There are innumerable possible studies on differential gene expression, but available modern methods that have been shown to work well in some studies may not necessarily work just as well when applied to others. Working with biologists, statisticians develop methods to address the multiplicity problem in gene expression studies by designing methods that control the rate of false discoveries.

The Search for Truth

At the core of every scientific endeavor is the pursuit of truth. Various challenges, such as the multiplicity problem, can make this truth quite elusive. As such, statisticians work with other scientists to construct and optimize tools for analyzing scientific data so that they may accurately and efficiently lead investigators to correct conclusions.

ASA