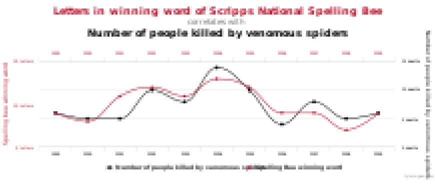
Wikipedia

Multiple comparisons problem

In statistics, the **multiple comparisons**, **multiplicity** or **multiple testing problem** occurs when one considers a set of statistical inferences simultaneously[1] or infers a subset of parameters selected based on the observed values.[2] In certain fields it is known as the look-elsewhere effect.



An example of data produced by data dredging, showing a correlation between the number of letters in a spelling bee's winning word and the number of people in the United States killed by venomous spiders. The clear similarity in trends is a coincidence. If many data series are compared, similarly convincing but coincidental data may be obtained.

The more inferences are made, the more likely erroneous inferences are to occur. Several statistical techniques have been developed to prevent this from happening, allowing significance levels for single and multiple comparisons to be directly compared. These techniques generally require a stricter significance threshold for individual comparisons, so as to compensate for the number of inferences being made.

Contents

History

The interest in the problem of multiple comparisons began in the 1950s with the work of Tukey and Scheffé. Other methods, such as the closed testing procedure (Marcus et al., 1976) and the Holm–Bonferroni method (1979), later emerged. In 1995, work on the false discovery rate began. In 1996, the first conference on multiple comparisons took place in Israel. This was followed by conferences around the world, usually taking place about every two years.[3]

Definition

Multiple comparisons arise when a statistical analysis involves multiple simultaneous statistical tests, each of which has a potential to produce a "discovery", of the same dataset or dependent datasets. A stated confidence level generally applies only to each test considered individually, but often it is desirable to have a confidence level for the whole family of simultaneous tests.[4] Failure to compensate for multiple comparisons can have important real-world consequences, as illustrated by the following examples:

- Suppose the treatment is a new way of teaching writing to students, and the control is the standard way of teaching writing. Students in the two groups can be compared in terms of grammar, spelling, organization, content, and so on. As more attributes are compared, it becomes increasingly likely that the treatment and control groups will appear to differ on at least one attribute due to random sampling error alone.
- Suppose we consider the efficacy of a drug in terms of the reduction of any one of a number of disease symptoms. As more symptoms are considered, it becomes increasingly likely that the drug will appear to be an improvement over existing drugs in terms of at least one symptom.

In both examples, as the number of comparisons increases, it becomes more likely that the groups being compared will appear to differ in terms of at least one attribute. Our confidence that a result will generalize to independent data should generally be weaker if it is observed as part of an analysis that involves multiple comparisons, rather than an analysis that involves only a single comparison.

For example, if one test is performed at the 5% level and the corresponding null hypothesis is true, there is only a 5% chance of incorrectly rejecting the null hypothesis. However, if 100 tests are conducted and all corresponding null hypotheses are true, the expected number of incorrect rejections (also known as false positives or Type I errors) is 5. If the tests are statistically independent from each other, the probability of at least one incorrect rejection is 99.4%.

Note that of course the multiple comparisons problem arises not in every situation where several hypotheses are empirically tested, be that sequentially or in parallel (concurrent).[5] Roughly speaking, the multiple comparisons problem arises whenever multiple hypotheses are tested on the same dataset (or datasets that are not independent) or whenever one and the same hypothesis is tested in several datasets.

The multiple comparisons problem also applies to confidence intervals. A single confidence interval with a 95% coverage probability level will contain the population parameter in 95% of experiments. However, if one considers 100 confidence intervals simultaneously, each with 95% coverage probability, the expected number of non-covering intervals is 5. If the intervals are statistically independent from each other, the probability that at least one interval does not contain the population parameter is 99.4%.

Techniques have been developed to prevent the inflation of false positive rates and non-coverage rates that occur with multiple statistical tests.

Classification of multiple hypothesis tests

The following table defines the possible outcomes when testing multiple null hypotheses. Suppose we have a number m of null hypotheses, denoted by: $H_1, H_2, ..., H_m$. Using a statistical test, we reject the null hypothesis if the test is declared significant. We do not reject the null hypothesis if the test is non-significant. Summing each type of outcome over all H_i yields the following random variables:

	Null hypothesis is true (H_0)	**Alternative hypothesis is true (H_A)**	**Total**
Test is declared significant	V	S	R
Test is declared non-significant	U	T	$m-R$
Total	m_0	$m-m_0$	m

- m is the total number hypotheses tested
- m_0 is the number of true null hypotheses, an unknown parameter
- $m-m_0$ is the number of true alternative hypotheses
- V is the number of false positives (Type I error) (also called "false discoveries")

- S is the number of true positives (also called "true discoveries")
- T is the number of false negatives (Type II error)
- U is the number of true negatives
- $R = V + S$ is the number of rejected null hypotheses (also called "discoveries", either true or false)

In m hypothesis tests of which m_0 are true null hypotheses, R is an observable random variable, and S, T, U, and V are unobservable random variables.

Controlling procedures

If *m* independent comparisons are performed, the *family-wise error rate* (FWER), is given by

$$\bar{\alpha} = 1 - \left(1 - \alpha_{\{\text{per comparison}\}}\right)^m.$$

Hence, unless the tests are perfectly positively dependent (i.e., identical), $\bar{\alpha}$ increases as the number of comparisons increases. If we do not assume that the comparisons are independent, then we can still say:

$$\bar{\alpha} \le m \cdot \alpha_{\{\text{per comparison}\}},$$

which follows from Boole's inequality. Example: $0.2649 = 1 - (1 - .05)^6 \le .05 \times 6 = 0.3$

There are different ways to assure that the family-wise error rate is at most $\bar{\alpha}$. The most conservative method, which is free of dependence and distributional assumptions, is the Bonferroni correction $\alpha_{\{\text{per comparison}\}} = \alpha/m$. A marginally less conservative correction can be obtained by solving the equation for the family-wise error rate of m independent comparisons for $\alpha_{\{\text{per comparison}\}}$. This yields $\alpha_{\{\text{per comparison}\}} = 1 - (1 - \alpha)^{1/m}$, which is known as the Šidák correction. Another procedure is the Holm–Bonferroni method, which uniformly delivers more power than the simple Bonferroni correction, by testing only the lowest p-value ($i = 1$) against the strictest criterion, and the higher p-values ($i > 1$) against progressively less strict criteria.[6] $\alpha_{\{\text{per comparison}\}} = \alpha/(m - i + 1)$.

Multiple testing correction refers to re-calculating probabilities obtained from a statistical test which was repeated multiple times. In order to retain a prescribed family-wise error rate α in an analysis involving more than one comparison, the error rate for each comparison must be more stringent than *α*. Boole's inequality implies that if each of *m* tests is performed to have type I error rate *α*/*m*, the total error rate will not exceed *α*. This is called the Bonferroni correction, and is one of the most commonly used approaches for multiple comparisons.

In some situations, the Bonferroni correction is substantially conservative, i.e., the actual family-wise error rate is much less than the prescribed level *α*. This occurs when the test statistics are highly dependent (in the extreme case where the tests are perfectly dependent, the family-wise error rate with no multiple comparisons adjustment and the per-test error rates are identical). For example, in fMRI analysis,[7][8] tests are done on over 100,000 voxels in the brain. The Bonferroni method would require p-values to be smaller than .05/100000 to declare significance. Since adjacent voxels tend to be highly correlated, this threshold is generally too stringent.

Because simple techniques such as the Bonferroni method can be conservative, there has been a great deal of attention paid to developing better techniques, such that the overall rate of false positives can be maintained without excessively inflating the rate of false negatives. Such methods can be divided into general categories:

- Methods where total alpha can be proved to never exceed 0.05 (or some other chosen value) under any conditions. These methods provide "strong" control against Type I error, in all conditions including a partially correct null hypothesis.
- Methods where total alpha can be proved not to exceed 0.05 except under certain defined conditions.
- Methods which rely on an omnibus test before proceeding to multiple comparisons. Typically these methods require a significant ANOVA, MANOVA, or Tukey's range test. These methods generally provide only "weak" control of Type I error, except for certain numbers of hypotheses.
- Empirical methods, which control the proportion of Type I errors adaptively, utilizing correlation and distribution characteristics of the observed data.

The advent of computerized resampling methods, such as bootstrapping and Monte Carlo simulations, has given rise to many techniques in the latter category. In some cases where exhaustive permutation resampling is performed, these tests provide exact, strong control of Type I error rates; in other cases, such as bootstrap sampling, they provide only approximate control.

Large-scale multiple testing

Traditional methods for multiple comparisons adjustments focus on correcting for modest numbers of comparisons, often in an analysis of variance. A different set of techniques have been developed for "large-scale multiple testing", in which thousands or even greater numbers of tests are performed. For example, in genomics, when using technologies such as microarrays, expression levels of tens of thousands of genes can be measured, and genotypes for millions of genetic markers can be measured. Particularly in the field of genetic association studies, there has been a serious problem with non-replication — a result being strongly statistically significant in one study but failing to be replicated in a follow-up study. Such non-replication can have many causes, but it is widely considered that failure to fully account for the consequences of making multiple comparisons is one of the causes.[9]

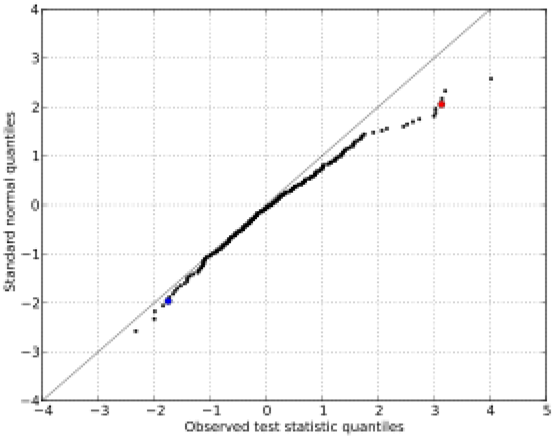
In different branches of science, multiple testing is handled in different ways. It has been argued that if statistical tests are only performed when there is a strong basis for expecting the result to be true, multiple comparisons adjustments are not necessary.[10] It has also been argued that use of multiple testing corrections is an inefficient way to perform empirical research, since multiple testing adjustments control false positives at the potential expense of many more false negatives. On the other hand, it has been argued that advances in measurement and information technology have made it far easier to generate large datasets for exploratory analysis, often leading to the testing of large numbers of hypotheses with no prior basis for expecting many of the hypotheses to be true. In this situation, very high false positive rates are expected unless multiple comparisons adjustments are made.

For large-scale testing problems where the goal is to provide definitive results, the familywise error rate remains the most accepted parameter for ascribing significance levels to statistical tests. Alternatively, if a study is viewed as exploratory, or if significant results can be easily re-tested in an independent study, control of the false discovery rate (FDR)[11][12][13] is often preferred. The FDR, loosely defined as the expected proportion of false positives among all significant tests, allows researchers to identify a set of "candidate positives" that can be more rigorously evaluated in a follow-up study.[14]

The practice of trying many unadjusted comparisons in the hope of finding a significant one is a known problem, whether applied unintentionally or deliberately, is sometimes called "p-hacking."[15][16]

Assessing whether any alternative hypotheses are true

A basic question faced at the outset of analyzing a large set of testing results is whether there is evidence that any of the alternative hypotheses are true. One simple meta-test that can be applied when it is assumed that the tests are independent of each other is to use the Poisson distribution as a model for the number of significant results at a given level α that would be found when all null hypotheses are true. If the observed number of positives is substantially greater than what should be expected, this suggests that there are likely to be some true positives among the significant results. For example, if 1000 independent tests are performed, each at level α = 0.05, we expect 0.05 × 1000 = 50 significant tests to occur when all null hypotheses are true. Based on the Poisson distribution with mean 50, the probability of observing more than 61 significant tests is less than 0.05, so if more than 61 significant results are observed, it is very likely that some of them correspond to situations where the alternative hypothesis holds. A drawback of this approach is that it over-states the evidence that some of the alternative hypotheses are true when the test statistics are positively correlated, which commonly occurs in practice.. On the other hand, the approach remains valid even in the presence of correlation among the test statistics, as long as the Poisson distribution can be shown to provide a good approximation for the number of significant results. This scenario arises, for instance, when mining significant frequent itemsets from transactional datasets. Furthermore, a careful two stage analysis can bound the FDR at a pre-specified level.[17]



A normal quantile plot for a simulated set of test statistics that have been standardized to be Z-scores under the null hypothesis. The departure of the upper tail of the distribution from the expected trend along the diagonal is due to the presence of substantially more large test statistic values than would be expected if all null hypotheses were true. The red point corresponds to the fourth largest observed test statistic, which is 3.13, versus an expected value of 2.06. The blue point corresponds to the fifth smallest test statistic, which is -1.75, versus an expected value of -1.96. The graph suggests that it is unlikely that all the null hypotheses are true, and that most or all instances of a true alternative hypothesis result from deviations in the positive direction.

Another common approach that can be used in situations where the test statistics can be standardized to Z-scores is to make a normal quantile plot of the test statistics. If the observed quantiles are markedly more dispersed than the normal quantiles, this suggests that some of the significant results may be true positives.

See also

Key concepts

- Familywise error rate
- False positive rate
- False discovery rate (FDR)
- False coverage rate (FCR)
- Interval estimation
- Post-hoc analysis
- Experimentwise error rate

General methods of alpha adjustment for multiple comparisons

- Closed testing procedure
- Bonferroni correction
- Boole–Bonferroni bound
- Duncan's new multiple range test
- Holm–Bonferroni method
- Harmonic mean p-value procedure

Related concepts

- Testing hypotheses suggested by the data
- Texas sharpshooter fallacy
- Model selection
- Look-elsewhere effect
- Data dredging

References

1. Miller, R.G. (1981). *Simultaneous Statistical Inference 2nd Ed*. Springer Verlag New York. ISBN 978-0-387-90548-8.
2. Benjamini, Y. (2010). "Simultaneous and selective inference: Current successes and future challenges". *Biometrical Journal*. **52** (6): 708–721. doi:10.1002/bimj.200900299 (https://doi.org/10.1002%2Fbimj.200900299). PMID 21154895 (https://pubmed.ncbi.nlm.nih.gov/21154895).
3. [1] (http://www.mcp-conference.org)
4. Kutner, Michael; Nachtsheim, Christopher; Neter, John; Li, William (2005). *Applied Linear Statistical Models*. pp. 744–745.
5. Georgiev, Georgi (2017-08-22). "Multivariate Testing – Best Practices & Tools for MVT (A/B/n) Tests" (http://blog.analytics-toolkit.com/2017/multivariate-testing-practices-tools-mvt-abn-tests/). *Blog for Web Analytics, Statistics and Data-Driven Internet Marketing | Analytics-Toolkit.com*. Retrieved 2020-02-13.
6. Aickin, M; Gensler, H (May 1996). "Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1380484). *Am J Public Health*. **86** (5): 726–728. doi:10.2105/ajph.86.5.726 (https://doi.org/10.2105%2Fajph.86.5.726). PMC 1380484 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1380484). PMID 8629727 (https://pubmed.ncbi.nlm.nih.gov/8629727).
7. Logan, B. R.; Rowe, D. B. (2004). "An evaluation of thresholding techniques in fMRI analysis". *NeuroImage*. **22** (1): 95–108. CiteSeerX 10.1.1.10.421 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.421). doi:10.1016/j.neuroimage.2003.12.047 (https://doi.org/10.1016%2Fj.neuroimage.2003.12.047). PMID 15110000 (https://pubmed.ncbi.nlm.nih.gov/15110000).
8. Logan, B. R.; Geliazkova, M. P.; Rowe, D. B. (2008). "An evaluation of spatial thresholding techniques in fMRI analysis". *Human Brain Mapping*. **29** (12): 1379–1389. doi:10.1002/hbm.20471 (https://doi.org/10.1002%2Fhbm.20471). PMID 18064589 (https://pubmed.ncbi.nlm.nih.gov/18064589).
9. Qu, Hui-Qi; Tien, Matthew; Polychronakos, Constantin (2010-10-01). "Statistical significance in genetic association studies" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3270946). *Clinical and Investigative Medicine. Medecine Clinique et Experimentale*. **33** (5): E266–E270. ISSN 0147-958X (https://www.worldcat.org/issn/0147-958X). PMC 3270946 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3270946). PMID 20926032 (https://pubmed.ncbi.nlm.nih.gov/20926032).

10. Rothman, Kenneth J. (1990). "No Adjustments Are Needed for Multiple Comparisons". *Epidemiology*. **1** (1): 43–46. doi:10.1097/00001648-199001000-00010 (https://doi.org/10.1097%2F00001648-199001000-00010). JSTOR 20065622 (https://www.jstor.org/stable/20065622). PMID 2081237 (https://pubmed.ncbi.nlm.nih.gov/2081237).
11. Benjamini, Yoav; Hochberg, Yosef (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society, Series B*. **57** (1): 125–133. JSTOR 2346101 (https://www.jstor.org/stable/2346101).
12. Storey, JD; Tibshirani, Robert (2003). "Statistical significance for genome-wide studies" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC170937). *PNAS*. **100** (16): 9440–9445. Bibcode:2003PNAS..100.9440S (https://ui.adsabs.harvard.edu/abs/2003PNAS..100.9440S). doi:10.1073/pnas.1530509100 (https://doi.org/10.1073%2Fpnas.1530509100). JSTOR 3144228 (https://www.jstor.org/stable/3144228). PMC 170937 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC170937). PMID 12883005 (https://pubmed.ncbi.nlm.nih.gov/12883005).
13. Efron, Bradley; Tibshirani, Robert; Storey, John D.; Tusher, Virginia (2001). "Empirical Bayes analysis of a microarray experiment". *Journal of the American Statistical Association*. **96** (456): 1151–1160. doi:10.1198/016214501753382129 (https://doi.org/10.1198%2F016214501753382129). JSTOR 3085878 (https://www.jstor.org/stable/3085878).
14. Noble, William S. (2009-12-01). "How does multiple testing correction work?" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2907892). *Nature Biotechnology*. **27** (12): 1135–1137. doi:10.1038/nbt1209-1135 (https://doi.org/10.1038%2Fnbt1209-1135). ISSN 1087-0156 (https://www.worldcat.org/issn/1087-0156). PMC 2907892 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2907892). PMID 20010596 (https://pubmed.ncbi.nlm.nih.gov/20010596).
15. Young, S. S., Karr, A. (2011). "Deming, data and observational studies" (http://www.niss.org/sites/default/files/Young%20Karr%20Obs%20Study%20Problem.pdf) (PDF). *Significance*. **8** (3): 116–120. doi:10.1111/j.1740-9713.2011.00506.x (https://doi.org/10.1111%2Fj.1740-9713.2011.00506.x).
16. Smith, G. D., Shah, E. (2002). "Data dredging, bias, or confounding" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1124898). *BMJ*. **325** (7378): 1437–1438. doi:10.1136/bmj.325.7378.1437 (https://doi.org/10.1136%2Fbmj.325.7378.1437). PMC 1124898 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1124898). PMID 12493654 (https://pubmed.ncbi.nlm.nih.gov/12493654).
17. Kirsch, A; Mitzenmacher, M; Pietracaprina, A; Pucci, G; Upfal, E; Vandin, F (June 2012). "An Efficient Rigorous Approach for Identifying Statistically Significant Frequent Itemsets". *Journal of the ACM*. **59** (3): 12:1–12:22. arXiv:1002.1104 (https://arxiv.org/abs/1002.1104). doi:10.1145/2220357.2220359 (https://doi.org/10.1145%2F2220357.2220359).

Further reading

- F. Betz, T. Hothorn, P. Westfall (2010), *Multiple Comparisons Using R*, CRC Press
- S. Dudoit and M. J. van der Laan (2008), *Multiple Testing Procedures with Application to Genomics*, Springer
- Farcomeni, A. (2008). "A Review of Modern Multiple Hypothesis Testing, with particular attention to the false discovery proportion". *Statistical Methods in Medical Research*. **17**: 347–388. doi:10.1177/0962280206079046 (https://doi.org/10.1177%2F0962280206079046).
- Phipson, B.; Smyth, G. K. (2010). "Permutation P-values Should Never Be Zero: Calculating Exact P-values when Permutations are Randomly Drawn". *Statistical Applications in Genetics and Molecular Biology*. doi:10.2202/1544-6155.1585 (https://doi.org/10.2202%2F1544-6155.1585).
- P. H. Westfall and S. S. Young (1993), *Resampling-based Multiple Testing: Examples and Methods for p-Value Adjustment*, Wiley

- P. Westfall, R. Tobias, R. Wolfinger (2011) *Multiple comparisons and multiple testing using SAS*, 2nd edn, SAS Institute
- A gallery of examples of implausible correlations sourced by data dredging (http://www.tylervigen.com/spurious-correlations)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Multiple_comparisons_problem&oldid=956165775"