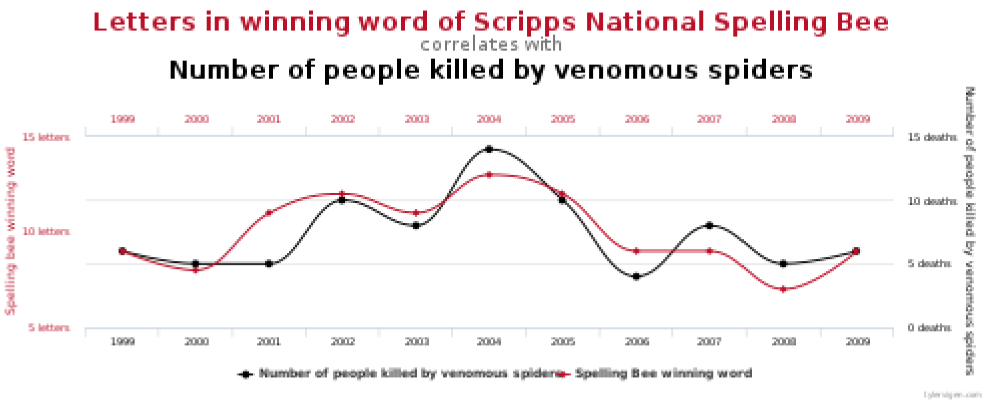
Data dredging



An example of a result produced by data dredging, showing a correlation between the number of letters in Scripps National Spelling Bee's winning word and the number of people in the United States killed by venomous spiders.

Data dredging (or **data fishing**, **data snooping**, **data butchery**), also known as **significance chasing**, **significance questing**, **selective inference**, and ***p*-hacking**[1] is the misuse of data analysis to find patterns in data that can be presented as statistically significant, thus dramatically increasing and understating the risk of false positives. This is done by performing many statistical tests on the data and only reporting those that come back with significant results.[2]

The process of data dredging involves testing multiple hypotheses using a single data set by exhaustively searching—perhaps for combinations of variables that might show a correlation, and perhaps for groups of cases or observations that show differences in their mean or in their breakdown by some other variable.

Conventional tests of statistical significance are based on the probability that a particular result would arise if chance alone were at work, and necessarily accept some risk of mistaken conclusions of a certain type (mistaken rejections of the null hypothesis). This level of risk is called the *significance*. When large numbers of tests are performed, some produce false results of this type; hence 5% of randomly chosen hypotheses might be (erroneously) reported to be statistically significant at the 5% significance level, 1% might be (erroneously) reported to be statistically significant at the 1% significance level, and so on, by chance alone. When enough hypotheses are tested, it is virtually certain that some will be reported to be statistically significant (even though this is misleading), since almost every data set with any degree of randomness is likely to contain (for example) some spurious correlations. If they are not cautious, researchers using data mining techniques can be easily misled by these results.

Data dredging is an example of disregarding the multiple comparisons problem. One form is when subgroups are compared without alerting the reader to the total number of subgroup comparisons examined.[3]

Contents

Drawing conclusions from data

The conventional frequentist statistical hypothesis testing procedure is to formulate a research hypothesis, such as "people in higher social classes live longer", then collect relevant data, followed by carrying out a statistical significance test to see how likely such results would be found if chance alone were at work. (The last step is called testing against the null hypothesis.)

A key point in proper statistical analysis is to test a hypothesis with evidence (data) that was not used in constructing the hypothesis. This is critical because every data set contains some patterns due entirely to chance. If the hypothesis is not tested on a different data set from the same statistical population, it is impossible to assess the likelihood that chance alone would produce such patterns. See testing hypotheses suggested by the data.

Here is a simple example. Throwing a coin five times, with a result of 2 heads and 3 tails, might lead one to hypothesize that the coin favors tails by 3/5 to 2/5. If this hypothesis is then tested on the existing data set, it is confirmed, but the confirmation is meaningless. The proper procedure would have been to form in advance a hypothesis of what the tails probability is, and then throw the coin various times to see if the hypothesis is rejected or not. If three tails and two heads are observed, another hypothesis, that the tails probability is 3/5, could be formed, but it could only be tested by a new set of coin tosses. It is important to realize that the statistical significance under the incorrect procedure is completely spurious – significance tests do not protect against data dredging.

Hypothesis suggested by non-representative data

Suppose that a study of a random sample of people includes exactly two people with a birthday of August 7: Mary and John. Someone engaged in data snooping might try to find additional similarities between Mary and John. By going through hundreds or thousands of potential similarities between the two, each having a low probability of being true, an unusual similarity can almost certainly be found. Perhaps John and Mary are the only two people in the study who switched minors three times in college. A hypothesis, biased by data snooping, could then be "People born on August 7 have a much higher chance of switching minors more than twice in college."

The data itself taken out of context might be seen as strongly supporting that correlation, since no one with a different birthday had switched minors three times in college. However, if (as is likely) this is a spurious hypothesis, this result will most likely not be reproducible; any attempt to check if others with an August 7 birthday have a similar rate of changing minors will most likely get contradictory results almost immediately.

Bias

Bias is a systematic error in the analysis. For example, doctors directed HIV patients at high cardiovascular risk to a particular HIV treatment, abacavir, and lower-risk patients to other drugs, preventing a simple assessment of abacavir compared to other treatments. An analysis that did not correct for this bias unfairly penalised abacavir, since its patients were more high-risk so more of them had heart attacks.[3] This problem can be very severe, for example, in the observational study.[3][2]

Missing factors, unmeasured confounders, and loss to follow-up can also lead to bias.[3] By selecting papers with a significant *p*-value, negative studies are selected against—which is the publication bias. This is also known as "file cabinet bias", because less significant *p*-value results are left in the file cabinet and never published.

Multiple modelling

Another aspect of the conditioning of statistical tests by knowledge of the data can be seen while using the system or machine analysis and linear regression to observe the frequency of data.. A crucial step in the process is to decide which covariates to include in a relationship explaining one or more other variables. There are both statistical (see Stepwise regression) and substantive considerations that lead the authors to favor some of their models over others, and there is a liberal use of statistical tests. However, to discard one or more variables from an explanatory relation on the basis of the data means one cannot validly apply standard statistical procedures to the retained variables in the relation as though nothing had happened. In the nature of the case, the retained variables have had to pass some kind of preliminary test (possibly an imprecise intuitive one) that the discarded variables failed. In 1966, Selvin and Stuart compared variables retained in the model to the fish that don't fall through the net—in the sense that their effects are bound to be bigger than those that do fall through the net. Not only does this alter the performance of all subsequent tests on the retained explanatory model, it may introduce bias and alter mean square error in estimation.[4][5]

Examples in meteorology and epidemiology

In meteorology, hypotheses are often formulated using weather data up to the present and tested against future weather data, which ensures that, even subconsciously, future data could not influence the formulation of the hypothesis. Of course, such a discipline necessitates waiting for new data to come in, to show the formulated theory's predictive power versus the null hypothesis. This process ensures that no one can accuse the researcher of hand-tailoring the predictive model to the data on hand, since the upcoming weather is not yet available.

As another example, suppose that observers note that a particular town appears to have a cancer cluster, but lack a firm hypothesis of why this is so. However, they have access to a large amount of demographic data about the town and surrounding area, containing measurements for the area of hundreds or thousands of different variables, mostly uncorrelated. Even if all these variables are independent of the cancer incidence rate, it is highly likely that at least one variable correlates significantly with the cancer rate across the area. While this may suggest a hypothesis, further testing using the same variables but with data from a different location is needed to confirm. Note that a *p*-value of 0.01 suggests that 1% of the time a result at least that extreme would be obtained by chance; if hundreds or thousands of hypotheses (with mutually relatively uncorrelated independent variables) are tested, then one is likely to obtain a *p*-value less than 0.01 for many null hypotheses.

Remedies

Looking for patterns in data is legitimate. Applying a statistical test of significance, or hypothesis test, to the same data that a pattern emerges from is wrong. One way to construct hypotheses while avoiding data dredging is to conduct randomized out-of-sample tests. The researcher collects a data set, then randomly partitions it into two subsets, A and B. Only one subset—say, subset A—is examined for creating hypotheses. Once a hypothesis is formulated, it must be tested on subset B, which was not used to construct the hypothesis. Only where B also supports such a hypothesis is it reasonable to believe the hypothesis might be valid. (This is a simple type of cross-validation and is often termed training-test or split-half validation.)

Another remedy for data dredging is to record the number of all significance tests conducted during the study and simply divide one's criterion for significance ("alpha") by this number; this is the Bonferroni correction. However, this is a very conservative metric. A family-wise alpha of 0.05, divided in this way by 1,000 to account for 1,000 significance tests, yields a very stringent per-hypothesis alpha of 0.00005. Methods particularly useful in analysis of variance, and in constructing simultaneous confidence bands for regressions involving basis functions are the Scheffé method and, if the researcher has in mind only pairwise comparisons, the Tukey method. The use of Benjamini and Hochberg's false discovery rate is a more sophisticated approach that has become a popular method for control of multiple hypothesis tests.

When neither approach is practical, one can make a clear distinction between data analyses that are confirmatory and analyses that are exploratory. Statistical inference is appropriate only for the former.[5]

Ultimately, the statistical significance of a test and the statistical confidence of a finding are joint properties of data and the method used to examine the data. Thus, if someone says that a certain event has probability of 20% ± 2% 19 times out of 20, this means that if the probability of the event is estimated *by the same method* used to obtain the 20% estimate, the result is between 18% and 22% with probability 0.95. No claim of statistical significance can be made by only looking, without due regard to the method used to assess the data.

Academic journals increasingly shift to the registered report format, which aims to counteract very serious issues such as data dredging and HARKing, which have made theory-testing research very unreliable: For example, Nature Human Behaviour has adopted the registered report format, as it “shift[s] the emphasis from the results of research to the questions that guide the research and the methods used to answer them”.[6] The European Journal of Personality defines this format as follows: “In a registered report, authors create a study proposal that includes theoretical and empirical background, research questions/hypotheses, and pilot data (if available). Upon submission, this proposal will then be reviewed prior to data collection, and if accepted, the paper resulting from this peer-reviewed procedure will be published, regardless of the study outcomes.”[7]

Methods and results can also be made publicly available, as in the open science approach, making it yet more difficult for data dredging to take place.[8]

See also

- Aliasing
- Base rate fallacy
- Bible code
- Bonferroni inequalities
- Cherry picking
- HARKing
- Lincoln–Kennedy coincidences urban legend
- Look-elsewhere effect
- Metascience
- Misuse of statistics
- Overfitting
- Pareidolia
- Post hoc analysis
- Predictive analytics
- Texas sharpshooter fallacy

References

1. Wasserstein, Ronald L.; Lazar, Nicole A. (2016-04-02). "The ASA Statement on p-Values: Context, Process, and Purpose" (https://doi.org/10.1080%2F00031305.2016.1154108). *The American Statistician*. Informa UK Limited. **70** (2): 129–133. doi:10.1080/00031305.2016.1154108 (https://doi.org/10.1080%2F00031305.2016.1154108). ISSN 0003-1305 (https://www.worldcat.org/issn/0003-1305).
2. Davey Smith, G.; Ebrahim, S. (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).
3. 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).
4. Selvin, H.C.; Stuart, A. (1966). "Data-Dredging Procedures in Survey Analysis". *The American Statistician*. **20** (3): 20–23. doi:10.1080/00031305.1966.10480401 (https://doi.org/10.1080%2F00031305.1966.10480401). JSTOR 2681493 (https://www.jstor.org/stable/2681493).
5. Berk, R.; Brown, L.; Zhao, L. (2009). "Statistical Inference After Model Selection" (https://repository.upenn.edu/statistics_papers/540). *J Quant Criminol*. **26** (2): 217–236. doi:10.1007/s10940-009-9077-7 (https://doi.org/10.1007%2Fs10940-009-9077-7). S2CID 10350955 (https://api.semanticscholar.org/CorpusID:10350955).
6. "Promoting reproducibility with registered reports" (https://doi.org/10.1038%2Fs41562-016-0034). *Nature Human Behaviour*. **1** (1): 0034. 10 January 2017. doi:10.1038/s41562-016-0034 (https://doi.org/10.1038%2Fs41562-016-0034). S2CID 28976450 (https://api.semanticscholar.org/CorpusID:28976450).
7. "Streamlined review and registered reports soon to be official at EJP" (https://www.ejp-blog.com/blog/2017/2/3/streamlined-review-and-registered-reports-coming-soon). *ejp-blog.com*.
8. Vyse, Stuart (2017). "P-Hacker Confessions: Daryl Bem and Me" (https://web.archive.org/web/20180805142806/https://www.csicop.org/specialarticles/show/p-hacker_confessions_daryl_bem_and_me). *Skeptical Inquirer*. **41** (5): 25–27. Archived from the original (https://www.csicop.org/specialarticles/show/p-hacker_confessions_daryl_bem_and_me) on 2018-08-05. Retrieved 5 August 2018.

Further reading

- Ioannidis, John P.A. (August 30, 2005). "Why Most Published Research Findings Are False" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327). *PLOS Medicine*. San Francisco: Public Library of Science. **2** (8): e124. doi:10.1371/journal.pmed.0020124 (https://doi.org/10.1371%2Fjournal.pmed.0020124). ISSN 1549-1277 (https://www.worldcat.org/issn/1549-1277). PMC 1182327 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327). PMID 16060722 (https://pubmed.ncbi.nlm.nih.gov/16060722).
- Head, Megan L.; Holman, Luke; Lanfear, Rob; Kahn, Andrew T.; Jennions, Michael D. (13 March 2015). "The Extent and Consequences of P-Hacking in Science" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4359000). *PLOS Biology*. **13** (3): e1002106. doi:10.1371/journal.pbio.1002106 (https://doi.org/10.1371%2Fjournal.pbio.1002106). PMC 4359000 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4359000). PMID 25768323 (https://pubmed.ncbi.nlm.nih.gov/25768323).
- Insel, Thomas (November 14, 2014). "P-Hacking" (https://www.nimh.nih.gov/about/directors/thomas-insel/blog/2014/p-hacking.shtml). *NIMH Director's Blog*.
- Smith, Gary (2016). *Standard Deviations: Flawed Assumptions, Tortured Data, and Other Ways to Lie with Statistics* (https://books.google.com/books?id=B-EoDwAAQBAJ). Gerald Duckworth & Co. ISBN 9780715649749.

External links

- A bibliography on data-snooping bias (http://data-snooping.martinsewell.com/)
- Spurious Correlations (http://www.tylervigen.com/spurious-correlations), a gallery of examples of implausible correlations
- StatQuest: *P*-value pitfalls and power calculations (https://www.youtube.com/watch?v=UFhJefdVCjE) on YouTube
- Video explaining p-hacking (https://www.youtube.com/watch?v=A0vEGuOMTyA) by "Neuroskeptic", a blogger at Discover Magazine
- Step Away From Stepwise (https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0143-6), an article in the Journal of Big Data criticising stepwise regression.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Data_dredging&oldid=1042250972"

This page was last edited on 4 September 2021, at 00:21 (UTC).