https://doi.org/10.1038/s41587-023-01794-9

How the Monty Hall problem is similar to the false discovery rate in high-throughput data analysis

The Monty Hall problem is arguably one of the most well-known probability problems in the public domain[1,2]. The problem was named after Monty Hall, the host of the American television game show *Let's Make a Deal*. The game has three doors, with a car behind one door and a goat behind each of the other two doors. The contestant does not know which door the car is behind and thus randomly chooses a door. This is where the situation becomes interesting. The host, who can see what is behind the two unchosen doors, opens one unchosen door with a goat behind it and asks the contestant if they would like to switch their already chosen door with the unopened unchosen door. The question is: would switching increase the contestant's chance of winning?

The Monty Hall problem became famous as a brain teaser. An overwhelming majority of people's first guess is that switching would not increase the chance of winning because it is impossible to be certain about which of the two unopened doors has the car behind it. However, the fact that the car can be behind either unopened door does not mean that the two unopened doors are equally likely to be hiding the car. What is the reason? The order of actions matters.

Let us have a thought experiment with two scenarios (Fig. 1a).

Scenario 1 is the Monty Hall problem: first, the contestant chooses a door; second, the host opens an unchosen door with a goat behind.

In Scenario 2, we switch the action order and let the host choose first. That is, the host first randomly opens one of the two doors with a goat behind.

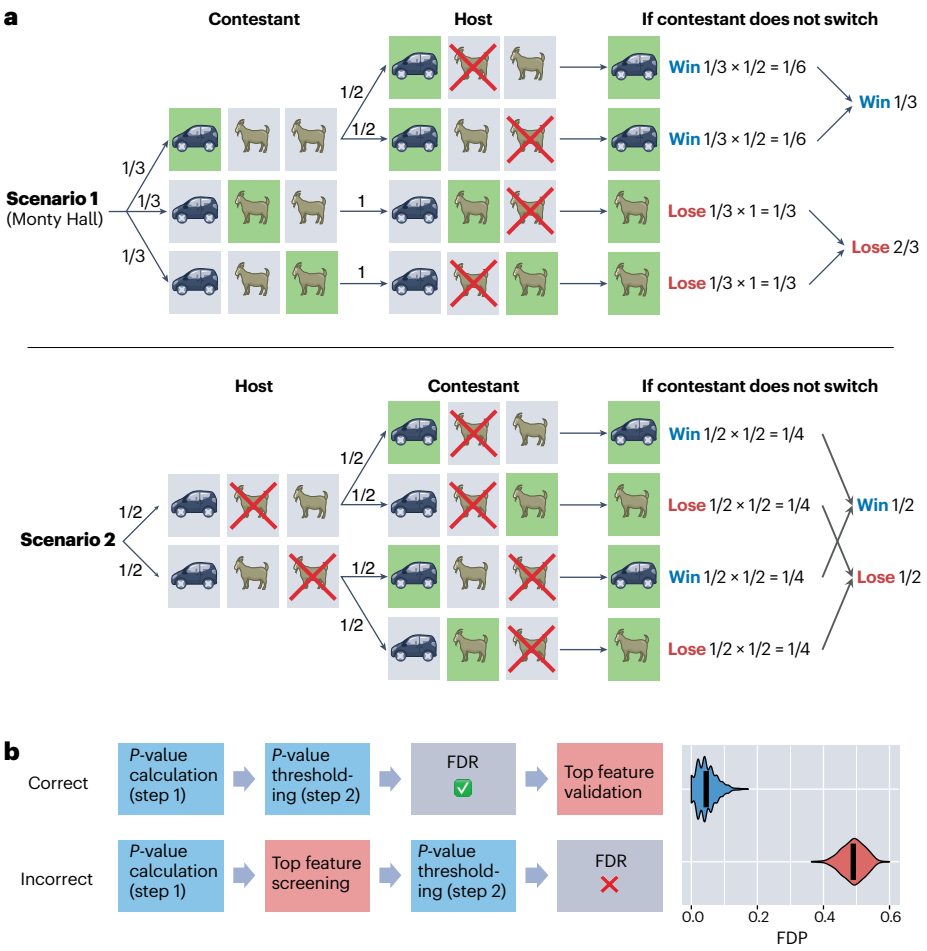

Fig. 1 | The order of actions matters in probability calculation: from the Monty Hall problem to the false discovery rate in high-throughput data analysis. a, The Monty Hall problem and two scenarios. Scenario 1 is the Monty Hall problem, in which the contestant first chooses a door and then the host opens an unchosen door with a goat behind it. Scenario 2 switches the order and lets the host first open a door with a goat behind. Although the contestant is left to choose between two doors in both scenarios, the winning probabilities are different. **b**, Two analysis procedures for identifying 'interesting' features, each of which receives a *P*-value. In the correct procedure (top), *P*-value thresholding is performed on all features' *P*-values, resulting in a valid control of the FDR. In the incorrect procedure (bottom), a feature screening step is added, so only the features with the smallest *P*-values are retained before the *P*-value thresholding step, leading to failed FDR control. The violin plots on the right show the distributions of the false discovery proportions (FDPs) of the two procedures in a simulation study (Zenodo https://doi.org/10.5281/zenodo.7809547) in which the target FDR is 0.05. Note that the FDR is defined as the expectation (that is, average) of the FDP distribution; only the correct procedure controls the FDR to be under 0.05.

In both scenarios, the contestant is left to choose between two unopened doors, one of which has the car behind it. However, the contestant's two choices have different chances of winning under the two scenarios (Fig. 1a). In Scenario 1, the contestant has only a 1/3 chance of winning if not switching the choice. In contrast, in Scenario 2, the contestant has a 1/2 chance of winning regardless of the choice. Many people surprised at the Monty Hall problem in fact have Scenario 2 in mind, thus thinking the two unopened doors are equally likely to have the car behind it.

The Monty Hall problem is an example that demonstrates how the order of actions can influence the final probability calculation. An interesting connection between the Monty Hall problem and scientific research is the calculation of the false discovery rate (FDR), the most widely used criterion in high-throughput data analysis where thousands of features (for example, genes) are examined simultaneously. Technically, the FDR is defined as the expected proportion of false discoveries among the discoveries.

In bioinformatics analysis, two steps are typically taken to identify 'interesting' features (Fig. 1b, top). In step 1, a *P*-value is calculated for every feature (usually, a smaller *P*-value means the feature is more likely interesting). In step 2, a *P*-value threshold is determined by a statistical procedure (for example, the Benjamini–Hochberg procedure[3] or Storey's *q*-value procedure[4]) to control the FDR to a target level (for example, 5%). After the two steps, a feature is identified as a discovery if its *P*-value is under the threshold.

In practice, most researchers do not validate all discoveries but only the features with the smallest *P*-values (Fig. 1b, top). This 'top feature validation' is a reasonable strategy, given a limited amount of resources. However, if this strategy is not used in the last step but performed as 'top feature screening' before step 2, then it would break down the theoretical guarantee of the statistical procedure for *P*-value thresholding in step 2, resulting in an inflated FDR (Fig. 1b, bottom). This phenomenon is well-known to statisticians and often referred to as "double dipping"[5,6] because the same set of *P*-values is used twice: first to screen for the top features and second to find the *P*-value threshold. This double dipping issue will cause step 2 to fail to control the FDR to the target level.

To make the discussion more concrete, imagine that we have RNA-seq samples from a wild-type condition and a gene knockdown condition. Our goal is to find the differentially expressed (DE) genes — that is, the interesting features whose expression levels changed significantly after the gene knockdown. The standard practice is to calculate a *P*-value for each gene (step 1) and find the *P*-value threshold corresponding to the 5% FDR (step 2). Then the genes with *P*-values below the threshold will be identified as DE genes. In the correct approach, the *P*-values are used only once, to determine the threshold, and if the *P*-values are valid (that is, *P*-values of true non-DE genes should be uniformly distributed between 0 and 1), the identified DE genes should satisfy the target 5% FDR (Fig. 1b, top). However, if the *P*-values are used twice — first to screen for the genes with small *P*-values after step 1 and second to find the *P*-value threshold based on only these genes in step 2 — then, in this incorrect approach, the identified DE genes may have an actual FDR far exceeding the target 5% (Fig. 1b, bottom).

A practical strategy for avoiding the possible failure of FDR control is to use in silico negative controls, such as permuted data[7] or simulated data[8] that is expected to contain no interesting features, to verify that the *P*-values before thresholding (step 2) approximately follow a uniform distribution between 0 and 1 (ref. 9). This sanity check is essential but largely neglected in data analysis. Another strategy is to avoid the complexity of *P*-value calculation and directly control the FDR[10], but a sanity check is still needed.

In summary, how to calculate probability correctly is a challenging question in many real-world problems, ranging from the Monty Hall problem in mass media to the high-throughput data analysis problem in scientific research. The order of actions taken is a critical but often ignored factor that determines the validity of probability calculation. As a result, to ensure the transparency and reproducibility of statistical analysis results in research papers, researchers should precisely record all data analysis procedures, including but not limited to the selection of data points and features, in their exact order. To help researchers implement this practice, research journals may add to the reporting summary document a table that streamlines the analysis procedures (Table 1 is an example).

Table 1 | Example table of data analysis procedures

Step	Samples	Features	Operations
0	10 wild-type samples 10 knockdown samples	20,000 genes	
1	Same as above	2,000 genes	Retain the genes with the largest variance at the log(transcripts per million) scale.
2	Same as above	250 genes	1. Calculate *P*-values for the 2,000 genes using the two-sided Wilcoxon rank-sum test. 2. Adjust the *P*-values using the Benjamini-Hochberg procedure. 3. Retain the genes with adjusted *P*-values under 0.05.

Jingyi Jessica Li [1,2] ✉

[1]Department of Statistics, University of California, Los Angeles, Los Angeles, CA, USA. [2]Radcliffe Institute of Advanced Study, Harvard University, Cambridge, MA, USA.
✉e-mail: lijy03@g.ucla.edu

Published online: 17 May 2023

References

1. Selvin, S. *Am. Stat.* **29**, 67, https://doi.org/10.1080/00031305.1975.10479121 (1975).
2. Rosenhouse, J. *The Monty Hall Problem: The Remarkable Story of Math's Most Contentious Brain Teaser* (Oxford Univ. Press, 2009).
3. Benjamini, Y. & Hochberg, Y. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
4. Storey, J. D. *Ann. Stat.* **31**, 2013–2035 (2003).
5. Benjamini, Y. *Biom. J.* **52**, 708–721 (2010).
6. Taylor, J. & Tibshirani, R. J. *Proc. Natl Acad. Sci. USA* **112**, 7629–7634 (2015).
7. Li, Y., Ge, X., Peng, F., Li, W. & Li, J. J. *Genome Biol.* **23**, 79 (2022).
8. Song, D., Wang, Q., Yan, G., Liu, T. & Li, J. J. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01772-1 (2023).
9. Song, D. & Li, J. J. *Genome Biol.* **22**, 124 (2021).
10. Ge, X. et al. *Genome Biol.* **22**, 288 (2021).

Acknowledgements

The author appreciates comments and feedback from Wei Li at the University of California, Irvine, Chongzhi Zang at the University of Virginia, and the author's PhD student Guanao Yan and postdoc Xinzhou Ge at UCLA. The author was supported by the following grants: National Science Foundation DBI-1846216 and DMS-2113754, NIH/NIGMS R35GM140888, a Johnson & Johnson WiSTEM2D Award, Sloan Research Fellowship, UCLA David Geffen School of Medicine W. M. Keck Foundation Junior Faculty Award, and the Chan-Zuckerberg Initiative Single-Cell Biology Data Insights Grant. The author was a fellow at the Radcliffe Institute for Advanced Study at Harvard University in 2022–2023 while writing this paper.

Competing interests

The author declares no competing interests.