RESEARCH NOTE

REVISED Multiple statistical tests: Lessons from a d20 [version 2; referees: 3 approved]

Christopher R. Madan

Department of Psychology, Boston College, Chestnut Hill, MA, 02467, USA

v2 **First published:** 02 Jun 2016, **5**:1129 (doi: 10.12688/f1000research.8834.1)
Latest published: 07 Sep 2016, **5**:1129 (doi: 10.12688/f1000research.8834.2)

Abstract

Statistical analyses are often conducted with α= .05. When multiple statistical tests are conducted, this procedure needs to be adjusted to compensate for the otherwise inflated Type I error. In some instances in tabletop gaming, sometimes it is desired to roll a 20-sided die (or 'd20') twice and take the greater outcome. Here I draw from probability theory and the case of a d20, where the probability of obtaining any specific outcome is $^1/_{20}$, to determine the probability of obtaining a specific outcome (Type-I error) at least once across repeated, independent statistical tests.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
REVISED version 2 published 07 Sep 2016	✓		
version 1 published 02 Jun 2016	✓ report	✓ report	✓ report

1 **Jens Foell**, Florida State University USA

2 **Matthew Wall**, Imperial College London UK

3 **Steven R Shaw**, McGill University Canada

Discuss this article

Comments (0)

Corresponding author: Christopher R. Madan (madanc@bc.edu)

How to cite this article: Madan CR. **Multiple statistical tests: Lessons from a d20 [version 2; referees: 3 approved]** *F1000Research* 2016, **5**:1129 (doi: 10.12688/f1000research.8834.2)

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

First published: 02 Jun 2016, **5**:1129 (doi: 10.12688/f1000research.8834.1)

REVISED **Amendments from Version 1**

Based on reviewer comments, a few minor edits have been made: (1) fixed minor phrasing issues (e.g., 'regular' vs. 'conventional'); (2) added the reference discussing methods for correcting for multiple comparisons; and (3) removed the 'across n die' fragment.

See referee reports

Introduction

In scientific research, it is important to consider the issue of conducting multiple statistical tests and the likelihood of spuriously obtaining a 'significant' effect. Within a null-hypothesis significance testing (NHST) framework, statistical tests are usually conducted with $\alpha = .05$, i.e., the likelihood of falsely rejecting the null hypothesis as .05. Interestingly, this value coincides with the probability of obtaining a specific outcome on a 20-sided dice (or 'd20'), as $\frac{1}{20} = .05$. In the current (fifth) edition of Dungeons & Dragons, a tabletop game, many in-game events are determined based on the outcome of a d20. However, to make some events more likely, there are times when players roll a d20 'with advantage', meaning that they roll the d20 twice and take the greater value[1]. (There are also instances where a d20 is rolled 'with disadvantage', where the lesser value is taken, but here I will only focus on the former case.) This parallels the use of NHST without any correction for multiple comparisons, as it is more likely to get a significant effect due to chance (i.e., Type-I error) if many tests are conducted without a correction for multiple comparisons.

Here I wondered how much the probability of obtaining a 20, on a d20, would increase due to multiple tests–i.e., obtaining at least one 20 across n die. This approach assumes that each statistical test is wholly independent from each other, and thus is likely to over-estimate the effect related to conducting multiple statistical tests using variations in how the measures are calculated or the use of different, but correlated, measures. Nonetheless, this exploration is based in probability theory and mathematical derivations, rather than computational simulations, and can serve as an comprehensible primer in understanding the relationship between repeated statistical tests and probability distributions.

Developing an intuition of statistics and probability distributions is of particular importance as most people, both laymen[2] and scientists[3,4], have misconceptions about NHST. This is further compounded by critics of NHST, which often over-emphasize the limitations of the approach, e.g., see 4–6. By providing a comprehensible example of how repeated statistical tests can inflate chance likelihoods, I hope that these demonstrations can improve researchers' intuitions regarding NHST. This approach is not contrary to those suggested by the use of confidence intervals and Bayesian statistics—which have become increasingly adopted across the life sciences, from medicine to psychology[7,8]—but rather to improve comprehension of the characteristics of NHST.

Mathematical derivations

The probability that of a specific outcome occurring on each on n die, each with d sides is:

$$P(d,n) = 1 - \left(\frac{d-1}{d}\right)^n$$

The probability of obtaining a specific outcome across n rolls of a d-sided die are listed in Table 1.

To develop some intuition of the effect of multiple die rolls, several simple cases can be considered.

For $d = 2$, i.e., a coin, the probability of obtaining at a heads when flipping one coin ($n = 1$) is $\frac{1}{2}$. The probability of obtaining a heads twice (with two coins, $n = 2$) is $\left(\frac{1}{2}\right)^2$ or $\frac{1}{4}$. In contrast, the probability of obtaining at least one heads when flipping two coins is $\frac{3}{4}$, as there are four possible outcomes ({HH, HT, TH, TT}) and three of them satisfy the criteria of 'at least one heads' ({HH, HT, TH}) and only one outcome does not ({TT}). This can more clearly be considered as the complementary event, where the probability is $1 - \frac{1}{4}$, which resolves to $\frac{3}{4}$.

For $d = 6$, i.e., a 'conventional' six-sided die, the probability of obtaining any specific outcome is $\frac{1}{6}$. When considering multiple dice, it is again important to differentiate the probability of obtaining 'obtaining the same specific outcome multiple times', e.g., the

Table 1. Probability of obtaining a specific outcome at least once, using a *d*-sided die rolled *n* times.

n	*d* = 2	*d* = 6	*d* = 20	*d* = 100	*d* = 1000
1	.5000	.1667	.0500	.0100	.0010
2	.7500	.3056	.0975	.0199	.0020
3	.8750	.4213	.1426	.0297	.0030
4	.9375	.5177	.1855	.0394	.0040
5	.9688	.5981	.2262	.0490	.0050
6	.9844	.6651	.2649	.0585	.0060
7	.9922	.7209	.3017	.0679	.0070
8	.9961	.7674	.3366	.0773	.0080
9	.9980	.8062	.3698	.0865	.0090
10	.9990	.8385	.4013	.0956	.0100
20	1.0000	.9739	.6415	.1821	.0198
50	1.0000	.9999	.9231	.3950	.0488
100	1.0000	1.0000	.9941	.6340	.0952
500	1.0000	1.0000	1.0000	.9934	.3936
1,000	1.0000	1.0000	1.0000	1.0000	.6323
10,000	1.0000	1.0000	1.0000	1.0000	1.0000

probability of obtaining two sixes with two dice is $\left(\frac{1}{6}\right)^2 = \frac{1}{36}$, from the case of 'obtaining at least one specific outcome across multiple dice'. To determine the probability of obtaining a specific out-come on *any* of multiple dice, the complementary event should again be considered, i.e., the probability of not obtaining that outcome on any of the die. For $n = 1$, the probability of *not* obtaining a specific outcome is $\frac{5}{6}$. Following from this, the probability of obtaining that specific outcome is $1 - \frac{5}{6}$ or $\frac{1}{6}$. When $n = 2$, the probability of not obtaining a six on either of the dice is $\left(\frac{5}{6}\right)^2$, which resolves to $\frac{25}{36}$. The complementary event of obtaining 'at least one six' is $1 - \frac{25}{36}$ or $\frac{11}{36}$. Here we can see that with two dice, the probability of obtaining at least one six (or any other specific outcome) is nearly doubled, from $\frac{6}{36}$ (i.e., $\frac{1}{6}$ with a single die).

For $d = 20$, i.e., a 20-sided die, the probability of obtaining any specific outcome is $\frac{1}{20}$ or .05. If $n = 2$ dice are rolled, the probability of obtaining at least one 20 is $\frac{39}{400}$ or .0975. If $n = 10$ dice are rolled, the probability of obtaining at least one 20 is ≈ .4013. With $n = 20$ dice, this increases further to ≈ .6415.

We can also consider a more general problem, the probability of obtaining an outcome of o or greater, on at least one of n d-sided die:

$$P(d,n,o) = 1 - \left(\frac{d-(d-o+1)}{d}\right)^n$$

For instance, when rolling a six-sided die, the probability of obtaining a five or higher is $\frac{2}{6}$ (equivalent to $\frac{12}{36}$). Following from the same approach of calculating the complementary event, the probability of obtaining not obtaining any two specific outcomes across multiple dice is $1 - \left(\frac{4}{6}\right)^2$, which resolves to $\frac{20}{36}$. Figure 1 and Table 2 show the probability of obtaining at least o on a $d = 20$ die, across n dice.

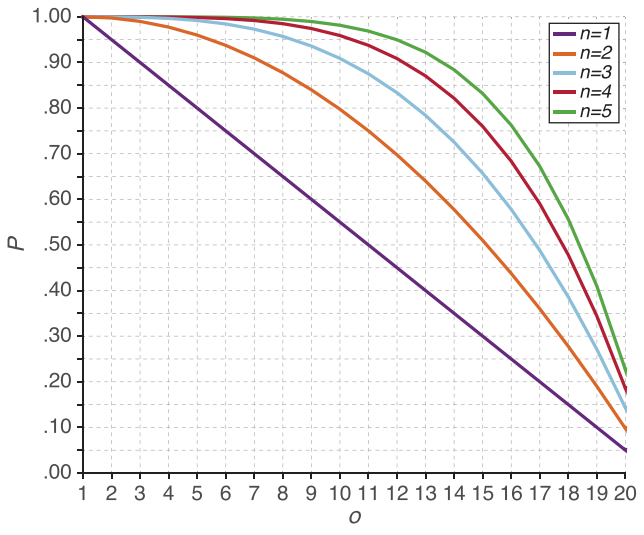

Figure 1. Probability (*P*) of obtaining at least an outcome *o* once, across *n* d20 die.

Table 2. Probability of obtaining at least an outcome *o* once, across *n* d20 die.

o	*n* = 1	*n* = 2	*n* = 3	*n* = 4	*n* = 5
1	1.0000	1.0000	1.0000	1.0000	1.0000
2	.9500	.9975	.9999	1.0000	1.0000
3	.9000	.9900	.9990	.9999	1.0000
4	.8500	.9775	.9966	.9995	.9999
5	.8000	.9600	.9920	.9984	.9997
6	.7500	.9375	.9844	.9961	.9990
7	.7000	.9100	.9730	.9919	.9976
8	.6500	.8775	.9571	.9850	.9947
9	.6000	.8400	.9360	.9744	.9898
10	.5500	.7975	.9089	.9590	.9815
11	.5000	.7500	.8750	.9375	.9688
12	.4500	.6975	.8336	.9085	.9497
13	.4000	.6400	.7840	.8704	.9222
14	.3500	.5775	.7254	.8215	.8840
15	.3000	.5100	.6570	.7599	.8319
16	.2500	.4375	.5781	.6836	.7627
17	.2000	.3600	.4880	.5904	.6723
18	.1500	.2775	.3859	.4780	.5563
19	.1000	.1900	.2710	.3439	.4095
20	.0500	.0975	.1426	.1855	.2262

Discussion

While it is widely understood that multiple comparisons need to be corrected for, many would underestimate the degree of inflation in Type-I error associated with additional, uncorrected statistical tests. Critically, statistical procedures have been developed to correct for multiple comparisons (e.g., Bonferroni, Tukey's HSD), see 9 for a detailed review. Nonetheless, the mathematical derivations presented here clearly illustrate the influence of multiple statistical tests on the likelihood of obtaining a specific outcome due to chance alone. These derivations and examples should be useful in providing a concrete example of the problem associated with uncorrected multiple comparisons and may prove useful as a pedagogical tool.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Grant information

The author declared that no grants were involved in supporting this work.

Acknowledgments

I would like to thank Critical Role (http://www.geekandsundry.com/shows/critical-role/) for introducing me to the concept of rolling a die 'with advantage'.

References

1. **D&D Player's Basic Rules**. Version 0.2. 2014.
Reference Source
2. Tromovitch P: **The lay public's misinterpretation of the meaning of 'significant': A call for simple yet significant changes in scientific reporting.** *J Res Practice.* 2015; **11**(1): P1.
Reference Source
3. Gliner JA, Leech NL, Morgan GA: **Problems with null hypothesis significance testing (NHST): what do the textbooks say?** *J Exp Educ.* 2002; **71**(1): 83–92.
Publisher Full Text
4. Nickerson RS: **Null hypothesis significance testing: a review of an old and continuing controversy.** *Psychol Methods.* 2000; **5**(2): 241–301.
PubMed Abstract | Publisher Full Text
5. Abelson RP: **On the surprising longevity of flogged horses: Why there is a case for the significance test.** *Psychol Sci.* 1997; **8**(1): 12–15.
Publisher Full Text
6. Cortina JM, Dunlap WP: **On the logic and purpose of significance testing.** *Psychol Methods.* 1997; **2**(2): 161–172.
Publisher Full Text
7. Fidler F, Cumming G, Burgman M, *et al.*: **Statistical reform in medicine, psychology and ecology.** *J Socio Econ.* 2004; **33**(5): 615–630.
Publisher Full Text
8. Fidler F: **Ethics and statistical reform: Lessons from medicine**. In A Panter & S Sterba (Eds.) *Handbook of Ethics in Quantitative Methodology.* Routledge: New York; 2011.
Publisher Full Text
9. Curran-Everett D: **Multiple comparisons: philosophies and illustrations.** *Am J Physiol Regul Integr Comp Physiol.* 2000; **279**(1): R1–8.
PubMed Abstract

Open Peer Review

Current Referee Status: ✓ ✓ ✓

Version 2

Referee Report 07 September 2016

doi:10.5256/f1000research.10322.r16168

Jens Foell
Department of Psychology, Florida State University, Tallahassee, FL, USA

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 05 July 2016

doi:10.5256/f1000research.9509.r14147

Steven R Shaw
Department of Educational and Counselling Psychology (ECP), McGill University, Montreal, QC, Canada

The author provides an excellent foundation for developing an intuitive understanding of null hypothesis significant testing. The concept of using a 20 sided die to assist graduate students and new researchers to better understand what exactly is meant by .05 and how multiple comparisons have a dramatic influence on interpretation is an interesting one — and I believe novel. This intuitive approach can also be used to improve understanding for the general public and overall science communication. Although this report includes a mathematical derivation, which may be too advanced for new scholars or science communication, it provides an excellent rationale for the use of 20 sided die.

When can also easily imagine that this rationale can be used better understand robustness of the outcomes of studies that may be influenced by experiment-wise error rates, such as in the cases of multiple attempts at replication or multiple trials of a specific experiment. The 20 sided die provides a concrete and real world method of communicating the complexities of multiple comparisons that is far more user-friendly than random number generators and variations of Monte Carlo studies.

I am curious as to whether the exactitude in engineering and manufacturing a 20 sided die will result in exactly equal probability of each number appearing. A six sided die is created with right angles and is relatively easy to create an equal probability of landing on each side. Obviously, this makes no difference

or changes the point of the paper; yet, it may add error should anyone actually attempt to roll the 20 sided dice multiple times. Just a thought as I am unsure on this issue. I suppose should a 20 sided die found to contain significant error or even bias, then there would be scandal in the Dungeons & Dragons world.

The author also deserves credit for establishing high levels of nerd credibility. When Dungeons & Dragons, mathematical derivations, and useful statistical communication methods are combined into a single published scientific paper; the trifecta of nerd credibility has been achieved.

Overall, this is a well-written report and the mathematics is correct. I am hopeful that the author continues to elaborate on this concept and develops other uses of a 20 sided die for communication and lesson plans in courses on research design and basic applied statistics. This will also be helpful in explaining experiment wise error rate/multiple comparisons in the cases of science communication as well.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 08 June 2016

doi:10.5256/f1000research.9509.r14150

Matthew Wall
Division of Brain Sciences, Imperial College London, London, UK

This short report provides a simple and concise illustration of some of the issues surrounding multiple comparisons in statistical testing. I see nothing wrong with the logic or the mathematics, and can see that this would make a valuable contribution as an assigned text on an introductory statistics course.

The only suggestion I have for material that could be added would be to include a citation to additional material on the topic of multiple-comparisons correction for the interested reader. This would make the piece more valuable as a teaching aid. I'd suggest adding a couple of sentences to the discussion along those lines and citing a review paper on the topic, such as Curran-Everett, 2000 [1]

Couple of minor points:

1. The paragraph after table 1 starts 'For intuition,'. This seems an odd phrase to me. Maybe replace with 'Intuitively' or 'For the purposes of familiarity...' or something like that.

2. Just before the general equation the phrase 'across n die' appears on its own, as a separate paragraph. An error? Or is this supposed to be a subtitle?

References

1. Curran-Everett D: Multiple comparisons: philosophies and illustrations.*Am J Physiol Regul Integr Comp Physiol*. 2000; **279** (1): R1-8 PubMed Abstract

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 03 June 2016

doi:10.5256/f1000research.9509.r14149

Jens Foell
Department of Psychology, Florida State University, Tallahassee, FL, USA

The author provides a novel demonstration of the increasing probability of spurious research results when performing multiple tests: noticing the equivalence of a p = .05 statistical threshold and a d20 die (as used in popular games), the author goes on to describe changes in probabilities when using multiple dice, or when allowing multiple sides of the die to count as a correct result. The aim of this article, as I understand it, is to demonstrate that these changes have a surprisingly large influence on statistical hypothesis testing, and at the same time to provide a hands-on example that many readers might be able to relate to (in the form of the d20 die).

The article is well-written and in my opinion fulfills both of these goals. Its rationale and the mathematical derivations seem to be sound and correct. I can easily see this article being used by educators to teach the topic of spurious statistical results and to make the topic more accessible.

I have some minor edits/recommendations to increase the overall clarity and readability of the article:

- When the game name "Dungeons & Dragons" is first mentioned in the introduction, it should be followed by a citation of the game's publisher, year of publication, and, if necessary/appropriate, copyright information.

- A d6 die is described as "regular" in the text. This term seems ambiguous to me and I recommend replacing it with a term such as "cube-shaped," "classical," or "conventional."

- The current version of the manuscript appears to contain a printing error: before the last paragraph of the "mathematical derivations" section, the sentence fragment "across n die." is printed without context.

I recommend the indexing of this article after these minor issues have been addressed, and I hope the author will continue to produce research notes that highlight statistical issues in an approachable manner.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.