**Why Report an Exact *p* Value?**

 Before the arrival of inexpensive high-speed computers, getting an exact *p* value for test statistics such as *t*, *F*, and χ2 required doing integral calculus by hand. To avoid doing the calculus, there were developed tables of critical values for test statistics. The most often applied critical values were those that marked off the most extreme 5% of scores (in two tails for *t*, in one tail for *F* and χ2 ) in the distribution of the test statistic under the null hypothesis. If the obtained (from the sample) absolute value of the test statistic was greater than or equal to the critical value, then *p* was less than or equal to .05, the null hypothesis was rejected, and the effect declared “significant.” If not, then the null was retained and the effect declared “not significant.” Rather than reporting the (unknown) exact value of *p*, the researcher reported either “*p* < .05” or “*p* > .05.” Some used “*p* = n.s.” instead of “*p* > .05.”

 Today the grunt-work of statistical analysis is done on computers with statistical software like SAS, SPSS, and R. In addition to reporting the obtained value of the test statistic, these programs also report an exact *p* value (and other even more interesting statistics, like confidence intervals). Since *p* is a continuous variable (ranging from 0 to 1), it is useful to report its exact value. If *p* is greater than .10, report it to two decimal places. If .001 ≤ *p* < .10, report *p* to three decimal points. If *p* is less than .001, report “*p* < .001.”

 Suppose that Test A produced *t*(19) = 2.063, *p* = .053, *d* = .46, 95% CI [-.006, .92]. Test B produced *t*(19) = 0.472, *p* = .64, *d* = .11, 95% CI [-.34, .54]. The old fashioned way of reporting this would be for Test A, “*t*(19) = 2.063, *p* > .05,” and for Test B, “*t*(19) = 0.472, *p* > .05.” This produces the misimpression that the results of Test A and Test B are equivalent; *p* > .05 in both cases, and the effect not significant in both cases. Giving the exact *p* values makes it obvious that the two results are not equivalent, and the confidence intervals for *d* makes that even more obvious.

 Now suppose that Test C produced *t*(19) = 2.131, *p* = .046, *d* = .48, 95% CI [.008, .93]. The old fashioned report for Test C would be “p < .05” and for Test A would be “p > .05,” making it appear that the two tests produced very different results, but they did not.

 Now suppose that Test D produced *t*(19) = 3.746, *p* < .05. That is significant, just like with Test C, creating the impression that Tests C and D are equivalent – but they are not. The effect revealed by Test D is much larger than that revealed by Test C. For Test D, *t*(19) = 3.746, *p* = .001, *d* = .84, 95% CI [.32, .1.34].

 Test A, *t*(19) = 2.063, *p* = .053, *d* = .46, 95% CI [-.006, .92], is not significant, but it does not indicate very good fit between the null hypothesis and the data. Our confidence interval runs from trivial in one direction to large in the other direction. Suppose that Test E produced *t*(999) = 1.867, *p* = .062, *d* = .06, 95% CI [-.003, .12]. Both tests fall short of significance (*p* > .05), but, Test E, unlike Test A, indicates that there is good fit between the data and the null hypothesis. The confidence interval for *d* not only includes 0 but also indicates that the magnitude of *d* is trivial (not likely greater than .12).

 Now suppose that Test F produced *t*(999) = 1.969, *p* = .049, *d* = .06, 95% CI [.001, .13]. Although “significant” (*p* < .05) this result, like that of “nonsignificant” Test E, can be used to argue that the effect revealed is trivial in magnitude, not likely greater than *d* = .13.

**Reporting *p*: Give an Exact Value**

The American Psychological Association (Publication Manual, sixth edition) says "when reporting *p* values, report exact *p* values (e.g. *p* = .031) to two or three decimal places. However, report *p* values less than .001 as *p* < .001. The tradition of reporting *p* values in the form *p* < .10, *p* < .05, and so forth, was appropriate in a time when only limited tables of critical values were available. However, in tables, the "*p* < " notation may be necessary for clarity."

 Examples:

*p* = .67 two point precision is adequate. If *p* < .10, I recommend three decimal points.

*p* = .053 if you rounded to .05, it would appear to be significant

*p* = .042 three point precision is a good idea if *p* is less than .05, imho

*p* = .007

*p* < .001 Only rarely would I type something like "*p* = .0000000000000000000003962."

 Before the advent of personal computers, researchers reported inexact values for *p* – for example, “*p* < .01,” “*p* < .05,” or “*p* > .05.” It was just too much hassle to do the integral calculus to find exactly what p was, and tables of critical values enabled one to make inexact statements with little effort. Now that we have machines to find exact *p* values for us there rarely is any need to use inexact *p* values. Among the few exceptions are:

* When *p* < .001. We really do not need to state *p* with more than three decimal point precision.
* When using a test statistic for which there is not available a machine that will give us an exact *p*.

Here are some quotes supporting the reporting of exact *p* values.

* Wilkinson, L., & Task Force on Statistical Inference. (1999). [Statistical methods in psychology journals: Guidelines and explanations](http://uregina.ca/~chowsl/Paper_category/apatfsirpt.htm). *American Psychologist*, *54*(8), 594-604. [Read a summary of this article](http://core.ecu.edu/psyc/wuenschk/AP-Aug99.htm).

***Hypothesis tests.***  *It is hard to imagine a situation in which a dichotomous accept–reject decision is better than reporting an actual p value or, better still, a confidence interval. Never use the unfortunate expression "accept the null hypothesis." Always provide some effect-size estimate when reporting a p value.* [Cohen (1994)](http://uregina.ca/~chowsl/Paper_category/apatfsirpt.htm#c13#c13) has written on this subject in this journal. All psychologists would benefit from reading his insightful article.

* Iacobucci, D. (2005). [From the editor – On *p* values](http://jcr.wisc.edu/PDFs/p-values.pdf). *Journal of Consumer Research*, *32*, 1-6

While the decision criterion is indeed binary—that is, “Is my *p*-value less than .05?”—it seems

like valuable information to convey that *p* =.001 or that, while one failed to reject the null,

nevertheless *p* =.06 (Dixon 1998, 391).

Methodologists are increasingly recommending that researchers report precise *p*-values,

for example, *p* =.04 rather than *p* < .05 (Greenwald et al. 1996, 181). To use α = .05 “is an

anachronism. It was settled on when *p*-values were hard to compute and so some specific

values needed to be provided in tables. Now calculating exact *p*-values is easy [i.e., the

computer does it] and so the investigator can report [*p* = .04] and leave it to the reader to

[determine its significance]” (Wainer and Robinson 2003, 26).

* Finch, S., Cumming, G., & Thomason, N. (2001). [Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform](http://www.latrobe.edu.au/psy/cumming/docs/Finch%20et%20al%20JAP%20EPM%2001.pdf). *Educational and Psychological Measurement, 61*, 181-210.

The fourth and current edition (APA, 1994) provided more extensive recommendations. It detailed some statistics that researchers should include when reporting different inferential test results. For the first time, a recommendation to specify the a priori alpha level was made. For the first time, the difference between the alpha level and the p-value was explained, and the reporting of exact p-values suggested.

Comment – the APA has been encouraging the use of EXACT *p* values since 1994 !

In earlier years authors may have had difficulty calculating exact p values, but these are now provided routinely by analysis software.

Reporting of exact p*-*values should help reduce the reliance on statistical tests as dichotomous (or trichotomous) decision criteria. As Rosnow and Rosenthal (1989) wrote: “… there is no sharp line between a ‘significant’ and a ‘nonsignificant’ difference; significance in statistics, like the significance of a value in the universe of values, varies continuously between extremes” (p.1277).

 I could provide dozens of other references supporting the use of exact *p* values, but these I dug up on the Internet should do. I don’t have the most recent APA Publication manual here at home, but will quote from the 1994 edition, Section 1.10:

 ***Statistical significance.*** Two types of probabilities associated with the significance of inferential statistical tests are reported. One refers to the a priori probability that you have selected as an acceptable level of falsely rejecting a given null hypothesis. This probability, called the alpha level, is the probability of a Type I error in hypothesis testing. Commonly used alpha levels are .05 and .01. Before you begin to report specific results, you should routinely state the particular alpha level you selected for the statistical results you conducted:

 An alpha level of .05 was used for all statistical tests.

If you do not make a general statement about the alpha level, specify the alpha level when reporting each result.

 The other kind of a probability refers to the a posteriori likelihood of obtaining a result that is as extreme or more extreme than the actual value of the statistic you obtained, assuming that the null hypothesis is true. For example, given a true null hypothesis, the probability of obtaining the particular value of the statistic you computed might be .008. Many statistical packages now provide these exact values. You can report this distinct piece of information in addition to specifying whether you rejected or failed to reject the null hypothesis using the specified alpha level.

 With an alpha level of .05, the effect of age was statistically significant, F(1, 123) = 7.27, p = .008.

 **or**

The effect of age was not statistically significant, F(1, 123) = 2.45, p = .12.

 **Eff*ect size and strength of relationship.*** Neither of the two types of probability values reflects the importance (magnitude) of an effect of the strength of a relationship because both probability values depend on sample size. You can estimate the magnitude of the effect or the strength of the relationship with a number of measures that do not depend on sample size. Common measures are *r2*, η2, ω2, φ2, Cramér’s *V*, Kendall’s *W*, Cohen’s *d* and …..

 And here is a log of some online discussion of exact *p* values.

Lorri Cerro <cerro@gl.umbc.edu> writes:

I'm about to deal with SPSS t tests in my stats/methodology class,

and I wanted to pose a question that bothered me from last

semester:

I suggest my students report exact p values (since it's more

precise and allowed in 4th ed. APA style), but how do you report a

computer-generated p value of .000? It would not be correct to say

p = .000, right? Is the best option here p < .001?

 -------------------------------------------------------------------------------------------

I have several comments but will respond to your question first. Yes, the best option you have, given your approach, is to report p < .001. Having said that I would like to say the following:

(1) there are two ways of reporting p values: (a) identifying a general level of significance for a test (e.g., p< .05) and then identifying whether a test is significant or not at this level, and (b) reporting the p-value associated with individual tests (as you appear to be doing). The problem with the latter approach is that the p-value is really not of much interest because it says little about the population parameters being tested. Moreover, across replications for a constant sample sizes, we would expect the p-values to vary greatly, from non-significance to some "highly" significant level. When replications are based on varying sample sizes, one can predict quite confidently that the p-values associated with small <oops, that should be “large”) samples will be smaller than the p-values associated with large samples <oops, that should be “small”), all the while the population difference (that is what is being assessed in the t-test) is constant. Focusing on actual p-values is therefore of little value.

(2) there is a common tendency to use p-values as a measure of effect size (i.e., a p < .00000001 is somehow more significant than a p > .05 even though they may be tests of the same population difference) instead of using more appropriate measures of effect size, such as Cohen's d in the two sample case. Focusing on the p-value, as mentioned above, provides little information about the situation in the population(s) which is what the test is all about.

Better to report whether the was significant or not at the p = .05 level, report an effect size for the difference, and a confidence interval for the difference. Power and Meta-analysts in the future will thank you for making their lives so much easier. :-)

-Mike Palij/Psychology Dept/New York University

-------------------------------------------------------------------------------------------

Date: Sun, 24 Mar 96 11:56:53 EST

From: "Karl L. Wuensch" <PSWUENSC@ECUVM.CIS.ECU.EDU>

Subject: Exact p-values

To: ecupsy-l@ECUVM.CIS.ECU.EDU, TIPS@fre.fsu.umd.edu

 -Mike Palij advised us not to report exact p values because:

*(2) there is a common tendency to use p-values as a measure of effect size (i.e., a p< .00000001 is somehow more significant than a p> .05 [sic -- I think Mike meant "p < .05" here] even though they may be tests of the same population difference) instead of using more appropriate measures of effect size, such as Cohen's d in the two sample case. Focusing on the p-value, as mentioned above, provides little information about the situation in the population(s) which is what the test is all about. Better to report whether the [ effect ] was significant or not at >the p > .05 level, report an effect size for the difference, and a confidence interval for the difference.*

 While I agree with most of what Mike has posted (including parts I did not quote), it strikes me that he is suggesting that we not report statistics that are commonly misinterpreted. By that criterion we would report next to nothing. ;-) The exact significance level is no more misleading than is the value of the test statistic (t, F, etc.), should we not report them exactly, just report whether or not they equal or exceed some critical value? I prefer to give the reader more, not less, information, but be sure the reader has all of the appropriate information (not just the exact p, but useful measures of effect size as well). A well written results section will caution the reader when a low p resulted from high power with a trivial effect size. The exact size of the p would seem to be especially important when it is close to that magical .05 level. Consider three p's from studies with equivalent power: p = .045, p = .055, and p = .55. Do we really want to report the first simply as "p < .05" and the latter two as "p > .05"? Certainly the difference between the two "nonsignificant" results is greater than the difference between the "significant" one and the "nearly significant" one.

 My recommendation is to provide an exact p (at least for p's which are > .001) and a measure of strength of effect. With simple comparisons, such as the t-tests mentioned in the original query, why not present confidence intervals for the difference as well (or instead of the "significance testing")? IMHO we might well be better off just dropping "significance testing," at least for the evaluation of effects which can be stated in terms of relatively simple estimators, that is, statistics about which we can easily place confidence intervals (estimates of means, differences between means, correlation coefficients, etc.).

 Karl L. Wuensch, Dept. of Psychology, East Carolina Univ.

-------------------------------------------------------------------------------------------

Date: Sun, 24 Mar 96 14:13:26 EST

To: TIPS@fre.fsu.umd.edu

Subject: Re: Exact p-values

From: palij@xp.psych.nyu.edu

"Karl L. Wuensch" <PSWUENSC@ECUVM.CIS.ECU.EDU> writes:

>-Mike Palij advised us not to report exact p values because:

> [major snippage to save bandwidth]

> While I agree with most of what Mike has posted (including parts

>I did not quote), it strikes me that he is suggesting that we not

>report statistics that are commonly misinterpreted. By that

>criterion we would report next to nothing. ;-)

You know, in some situations that would be an improvement. ;-)

*The exact significance level is no more misleading than is the value of the test statistic (t, F, etc.), should we not report them exactly, just report whether or not they equal or exceed some critical value? I prefer to give the reader more, not less, information, but be sure the reader has all of the appropriate information (not just the exact p, but useful measures of effect size as well).*

I am in agreement. However, recall the original question that was asked: (paraphrasing) "What does one do when the computer output gives 'p < .0000'?" Reporting p < .0001 is not reporting an exact p-value in this case. Moreover, there are still some of us who actually still do some statistical tests with pocket calculators (in my exp psych lab we do this along with computer calculations in order to (a) keep the computer honest ;-) and (b) show that one can still do statistical analyses without a computer based statistical package). In the case of hand calculations, one doesn't know the exact p-value, just whether the test is significant or not but one can still go on and calculate effect size, confidence intervals, and other statistics.

*A well written results section will caution the reader when a low p resulted from high power with a trivial effect size.*

Ah, but the key phrase here is "well written". In too many places, including the journals of several different

areas, the results section, the discussion, and the abstract are not well written. I still cringe when I read an abstract and see "the difference was very significant (p < .00001)" when the person should be presenting an effect size measure as well as a statement about the psychological or practical significance of the result.

*The exact size of the p would seem to be especially important when it is close to that magical .05 level. Consider three p's from studies with equivalent power: p = .045, p = .055, and p = .55. Do we really want to report the first simply as "p < .05" and the latter two as "p > .05"? Certainly the difference between the two "nonsignificant" results is greater than the difference between the "significant" one and the "nearly significant" one.*

Good points. In the second case (p = .055), I tell my students to report that there is a trend toward significance or that the result is marginally significant and that replication with a larger sample size should decide whether the effect is "real" or reliable. I suggest the heuristic of "p > .10" as the criterion for results that are probably not of practical significance (that is, for situation where finding a significant result is not critical [such a critical situation would be finding a new treatment for AIDS]).

 *My recommendation is to provide an exact p (at least for p's which are > .001) and a measure of strength of effect. With simple comparisons, such as the t-tests mentioned in the original query, why not present confidence intervals for the difference as well (or instead of the "significance testing")? IMHO we might well be better off just dropping "significance testing," at least for the evaluation of effects which can be stated in terms of relatively simple estimators, that is, statistics about which we can easily place confidence intervals (estimates of means, differences between means, correlation coefficients, etc.).*

I am pretty much in agreement with you. It should be noted that an APA task force on the use of significance

testing has been or will be empaneled (my memory is a bit vague on this but I remember reading about it in Div. 5's newsletter "The Score"). In the next century we might see a significant change in the presentation of the results in journal articles.

-Mike Palij/Psychology Dept/New York University

-------------------------------------------------------------------------------------------

Date: Mon, 25 Mar 1996 09:08:23 +1000 (EST)

To: TIPS@fre.fsu.umd.edu

From: reece@rmit.edu.au (John Reece)

Subject: Re: APA Style/Probability

*I'm about to deal with SPSS t tests in my stats/methodology class, and I wanted to pose a question that bothered me from last semester: I suggest my students report exact p values (since it's more precise and allowed in 4th ed. APA style), but how do you report a computer-generated p value of .000? It would not be correct to say p = .000, right? Is the best option here p < .001?*

*>*

*Lorri Cerro, Department of Psychology, University of Maryland Baltimore County*

Lorri,

I applaud your instructing your students to report exact p levels. I have long argued against the imprecision of the "one star, two star" approach to reporting significance. My advice to students is exactly as you suggested. When the output is p = .000, report as p < .001. At that level, you're looking at something so small that a precise measurement is relatively meaningless anyway, athough I suppose you could rightly argue a meaningful difference between something that's significant at p = .0009, and something significant at p = .0000000000003. And it's for that very reason that I would suggest instructing your students to report a simple measure of effect size, which for t-test would be Cohen's d.

Hope this helps.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* John Reece, PhD, Department of Psychology & Intellectual Disability Studies ,

\* Royal Melbourne Institute of Technology, \* Bundoora Victoria 3083, AUSTRALIA

-------------------------------------------------------------------------------------------

Date: Sun, 24 Mar 1996 22:50:34 EST

I also prefer exact p-values. I also remember reading an article where the meta-analyst implored researchers to use exact p-values. One worry I have about dropping significance testing and substituting confidence levels has to do with communication. I wonder if reporting confidence intervals wouldn't make it difficult for the reader to follow a results section. Maybe it is just from a lack of practice. However, it might be that the significance test makes it easier to tell a story.

Tony Whetstone ECUPSY-L

-------------------------------------------------------------------------------------------

Date: Tue, 26 Mar 1996 15:16:17 EST

Karl A. Minke opined:

 *At least when one has rejected the null, the exact p-value does convey some information--the probability that one made an error when doing so. The p-value when one fails to reject the null is meaningless, however.*

 Treating the exact p-value as "the probability that one made an error when <rejecting the null hypothesis>" unfortunately is a very common error. One cannot make such an error (a Type I error) unless the null hypothesis is false, so to determine the probability having made such an error one must factor in the probability that the null hypothesis is true. Of course, one is not going to be able to quantify that probability in real situations, but one can expect that psychologists frame their hypotheses such that the probability of the null hypothesis being true (or even near to true) is quite low.

 I recommend the article "On the Probability of Making Type I Errors" by Pollard and Richardson, Psychological Bulletin, 1987, 102: 159-163 for a thorough (but dense) discussion of this problem. They refer to the probability that one has made an error when rejecting a null hypothesis as the "conditional posterior probability of making a Type I error." It is this probability which is commonly but mistakenly assumed to be equal to alpha or p.

 One related confusion is the assertion that using the .05 criterion will result in your making a Type I error 5% of the time you test a null. It should be clear that this is not so, this would be so only if every null hypothesis ever tested was true. Some have even written that the .05 criterion means that 5% of published rejections of the null are Type I errors. While this could be true (one need consider publication policy, the unconditional probability that a null hypothesis is true, and levels of power: under certain circumstances the Type I error rate could equal 5%), it is highly likely that the rate of Type I errors in the literature is extremely small, well below 5%. Of course, one could argue that no point null hypothesis is ever absolutely true, or the probability of such is quite small, but I prefer to think of "range" or "loose" null hypotheses of the form that the effect is zero or so close to zero that it might as well be zero for practical purposes.

 I also disagree with Karl Minke’s statement that p is totally uninformative when its value exceeds .05 or some other magical criterion of "significance." I prefer to treat p as an index of how well the data fit with the null hypothesis. High values of p indicate that the observed data are pretty much what you would expect given the null hypothesis. Low values of p indicate that the obtained sample is unlikely given the null, and thus cast some doubt on the veracity of the null, even if the p is not at or below the criterion of significance. I am reminded of a criminal case on which I was a juror. After evaluating the data my p was above my criterion of significance (I voted 'not guilty'), but not by much -- I thought it more likely that the defendant was guilty than innocent, but not "beyond a reasonable doubt." "Not guilty" is not the same as "innocent." When p = .055 I remain distrustful of the null hypothesis, even if I have not rejected it. When p = .55 I am much more comfortable with the null hypothesis, especially if my power was high.

 Let me share a couple of lines from the excellent article, "Statistical Procedures and the Justification of Knowledge in Psychological Science," by Rosnow and Rosenthal, which appeared in the American Psychologist in October of 1989:

 "surely, God loves the .06 nearly as much as the .05. "Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p?"

 "there is no sharp line between a 'significant' and a 'nonsignificant' difference; significance in statistics, like the significance of a value in the universe of values, varies continuously between extremes."

[Return to Wuensch’s Stat Help Page](http://core.ecu.edu/psyc/wuenschk/StatHelp/StatHelp.htm)

Karl L. Wuensch, January, 2016.