University of Washington Box 351525
Psychology Writing Center psywc@uw.edu
http://www.psych.uw.edu/psych.php#p=339 (206) 685-8278

Reporting Results of Common Statistical Tests in APA Format

The goal of the results section in an empirical paper is to report the results of the data analysis used to test a hypothesis. The results section should be in condensed format and lacking interpretation. Avoid discussing why or how the experiment was performed or alluding to whether your results are good or bad, expected or unexpected, interesting or uninteresting. This document is specifically about how to report statistical results. Refer to our handout "Writing an APA Empirical (lab) Report" for details on writing a results section.

Every statistical test that you report should relate directly to a hypothesis. Begin the results section by restating each hypothesis, then state whether your results supported it, then give the data and statistics that allowed you to draw this conclusion.

If you have multiple numerical results to report, it's often a good idea to present them in a figure (graph) or a table (see our handout on APA table guidelines).

In reporting the results of statistical tests, report the descriptive statistics, such as means and standard deviations, as well as the test statistic, degrees of freedom, obtained value of the test, and the probability of the result occurring by chance (p value). Test statistics and p values should be rounded to two decimal places. All statistical symbols that are not Greek letters should be italicized (*M, SD, N, t, p,* etc.).

When reporting a significant difference between two conditions, indicate the direction of this difference, i.e. which condition was more/less/higher/lower than the other condition(s). Assume that your audience has a professional knowledge of statistics. Don't explain how or why you used a certain test unless it is unusual.

p values

There are two ways to report p values. One way is to use the alpha level (the a priori criterion for the probablility of falsely rejecting your null hypothesis), which is typically .05 or .01. Example: $F(1, 24) = 44.4$, $p < .01$. You may also report the exact p value (the a posteriori probability that the result that you obtained, or one more extreme, occurred by chance). Example: $t(33) = 2.10$, $p = .03$. If your exact p value is less than .001, it is conventional to state merely $p < .001$. If you report exact p values, state early in the results section the alpha level used as a significance criterion for your tests. Example: "We used an alpha level of .05 for all statistical tests."

EXAMPLES

Reporting a significant single sample t-test ($\mu \neq \mu_0$):

Students taking statistics courses in psychology at the University of Washington reported studying more hours for tests ($M = 121$, $SD = 14.2$) than did UW college students in in general, $t(33) = 2.10$, $p = .034$.

Reporting a significant t-test for dependent groups ($\mu_1 \neq \mu_2$):

Results indicate a significant preference for pecan pie ($M = 3.45$, $SD = 1.11$) over cherry pie ($M = 3.00$, $SD = .80$), $t(15) = 4.00$, $p = .001$.

Reporting a significant t-test for independent groups ($\mu_1 \neq \mu_2$):

UW students taking statistics courses in Psychology had higher IQ scores ($M = 121$, $SD = 14.2$) than did those taking statistics courses in Statistics ($M = 117$, $SD = 10.3$), $t(44) = 1.23$, $p = .09$.

Over a two-day period, participants drank significantly fewer drinks in the experimental group ($M= 0.667$, $SD =$

1.15) than did those in the wait-list control group (M= 8.00, SD= 2.00), $t(4) = -5.51$, p=.005.

Reporting a significant omnibus F test for a one-way ANOVA:

An analysis of variance showed that the effect of noise was significant, $F(3,27) = 5.94$, $p = .007$. Post hoc analyses using the Scheffé post hoc criterion for significance indicated that the average number of errors was significantly lower in the white noise condition ($M = 12.4$, $SD = 2.26$) than in the other two noise conditions (traffic and industrial) combined ($M = 13.62$, $SD = 5.56$), $F(3, 27) = 7.77$, $p = .042$.

Reporting tests of a priori hypotheses in a multi-group study:

Tests of the four a priori hypotheses were conducted using Bonferroni adjusted alpha levels of .0125 per test (.05/4). Results indicated that the average number of errors was significantly lower in the silence condition ($M = 8.11$, $SD = 4.32$) than were those in both the white noise condition ($M = 12.4$, $SD = 2.26$), $F(1, 27) = 8.90$, $p = .011$ and in the industrial noise condition ($M = 15.28$, $SD = 3.30$), $F(1, 27) = 10.22$, $p = .007$. The pairwise comparison of the traffic noise condition with the silence condition was non-significant. The average number of errors in all noise conditions combined ($M = 15.2$, $SD = 6.32$) was significantly higher than those in the silence condition ($M = 8.11$, $SD = 3.30$), $F(1, 27) = 8.66$, $p = .009$.

Reporting results of major tests in factorial ANOVA; non-significant interaction:

Attitude change scores were subjected to a two-way analysis of variance having two levels of message discrepancy (small, large) and two levels of source expertise (high, low). All effects were statistically significant at the .05 significance level.

The main effect of message discrepancy yielded an F ratio of $F(1, 24) = 44.4$, $p < .001$, indicating that the mean change score was significantly greater for large-discrepancy messages ($M = 4.78$, $SD = 1.99$) than for small-discrepancy messages ($M = 2.17$, $SD = 1.25$). The main effect of source expertise yielded an F ratio of $F(1, 24) = 25.4$, $p < .01$, indicating that the mean change score was significantly higher in the high-expertise message source ($M = 5.49$, $SD = 2.25$) than in the low-expertise message source ($M = 0.88$, $SD = 1.21$). The interaction effect was non-significant, $F(1, 24) = 1.22$, $p > .05$.

Reporting results of major tests in factorial ANOVA; non-significant interaction:

A two-way analysis of variance yielded a main effect for the diner's gender, $F(1, 108) = 3.93$, $p < .05$, such that the average tip was significantly higher for men ($M = 15.3\%$, $SD = 4.44$) than for women ($M = 12.6\%$, $SD = 6.18$). The main effect of touch was non-significant, $F(1, 108) = 2.24$, $p > .05$. However, the interaction effect was significant, $F(1, 108) = 5.55$, $p < .05$, indicating that the gender effect was greater in the touch condition than in the non-touch condition.

Reporting the results of a chi-square test of independence:

A chi-square test of independence was performed to examine the relation between religion and college interest. The relation between these variables was significant, X^2 $(2, N = 170) = 14.14$, $p < .01$. Catholic teens were less likely to show an interest in attending college than were Protestant teens.

Reporting the results of a chi-square test of goodness of fit:

A chi-square test of goodness-of-fit was performed to determine whether the three sodas were equally preferred. Preference for the three sodas was not equally distributed in the population, X^2 $(2, N = 55) = 4.53$, $p < .05$.

Thanks to Laura Little, Ph.D., UW Department of Psychology, for providing the examples reported here.

