Auto-Stats

Arne John

2023-07-17

```
if (alternative == "two.sided") {
    alternative_text <- "The report below is for a two-sided test, that is, the alternative hypothesis
} else if (alternative == "greater") {
    alternative_text <- "The report below is for a one-sided greater test, that is, the alternative hypo
} else if (alternative == "less") {
    alternative_text <- "The report below is for a one-sided lesser test, that is, the alternative hypot
}
```

This is an autostat report for an independent samples t-test. Interest centers on the comparison of two groups (i.e., group = Treat versus group = Control) concerning their population means for the dependent variable drp. The t-test assumes that the drp data from each group are continuous and normally distributed. The report below is for a two-sided test, that is, the alternative hypothesis does not state the direction of the effect.

1. Executive Summary of the Default Analysis

The t-test table above summarizes the outcome of the default statistical test. The difference in the two sample means is *9.9545*, with a standard error of 4.3918927. The corresponding value for Cohen's d equals *-0.6841*, with a standard error of *0.3182* and a 95% confidence interval ranging from *-1.2895* to *-0.0710*. This difference is statistically significant at the .05 level: p=0.0286295, t(42) = 2.2665516. We may not reject the null-hypothesis of no population difference between the groups. [NB. this needs to be adjusted for a one-sided test] Note that this does not mean that the data provide evidence for the null hypothesis or provide evidence against the alternative hypothesis; it also does not mean that the null hypothesis is likely to hold. These results also do not identify a likely range of values for effect size. In order to address these questions a Bayesian analysis would be needed. The Vovk-Sellke maximum p-Ratio of *3.6162* indicates the maximum possible odds in favor of H1 over H0, which is not compelling and urges caution. [only include for odds lower than 10] [Note to Arne: these last sentences would clearly be part of a verbose report] [Note to Arne: we could also include a mention of whether the assumptions appear violated, and what a nonparametric test shows]

3. Assumption Checks

In addition to a visual inspection of the raincloud plot, the assumptions of the t-test can also be formally tested.

For group = "Treat", the Shapiro-Wilk test for normality is not [fork: omit the not] statistically significant at the .05 level (i.e., p = *.6517*), and hence we can retain [fork: reject] the hypothesis that the data for group = *"Treat"* are normally distributed. For group = *"Control"*, the Shapiro-Wilk test for normality is not [fork: omit the not] statistically significant at the .05 level (i.e., p = *.7322*), and hence we can retain [fork: reject] the hypothesis that the data for group = *"Control"* are normally distributed. [Note to Arne:

in high-verbose level, we ought to add the reference to Shapiro-Wilk] Note that when the Shapiro-Wilk test is statistically nonsignificant this does not mean that the assumption of normality is met, or that the data support that assertion. Likewise, when the Shapiro-Wilk test is statistically significant this does not mean that the data provide evidence for the assertion that the data are not normally distributed. In order to address these questions a Bayesian analysis would be needed.

The Brown-Forsythe test for equality of variances is *not* [fork: omit the not] statistically significant at the .05 level: F(1,42) = 2.3418, p = .1334. Hence we can retain [fork: reject] the null hypothesis that the variances in both groups are equal. Note that when the Brown-Forsythe test is statistically nonsignificant this does not mean that the assumption of equal variance is met, or that the data support that assertion. Likewise, when the Brown-Forsythe test is statistically significant this does not mean that the data provide evidence for the assertion that groups have different variances. In order to address these questions a Bayesian analysis would be needed.

2. Descriptives

The table below summarizes the observed data for each group separately, followed by a raincloud plot that shows the individual observations.

As can be seen from the table, group = *"Treat"* contains *21* observations and has a mean *drp* of *51.4762* with a standard deviation of *11.0074*; group = *"Control"* contains *23* observations and has a mean *drp* of *41.5217* with a standard deviation of *17.1487*. The observed mean *drp* in group = *"Treat"* is *higher* than the observed mean *drp* in group = *"Control"*.

The raincloud plot shows the individual observations, together with box plots and density estimates (flipped on their side). The raincloud plot allows a visual assessment of (1) the extent to which the data in each group are normally distributed (i.e., are the density estimates symmetric and bell-shaped?); (2) the extent to which the data contain outliers; (3) the extent to which the group variances are equal. When a visual inspection of the raincloud plot suggests non-normality, outliers, or heterogeneity in variance, this may be followed up with assumption tests. One may address each violation separately (i.e., transform the dependent variable to normality, remove the outliers, apply the Welch test instead of the t-test), but in these cases it is generally prudent to conduct a nonparametric test that only takes into account the ranks of the observations. Note that it when reporting the results of a t-test, it is crucial to plot the data.

4. Parameter Estimation: How Strong is the Effect?

```
# Effect size for Cohen's d

# Very small    0.01
# Small 0.20
# Medium    0.50
# Large 0.80
# Very large    1.20
# Huge  2.0

effect_scheme <- "medium to large"

# Sources
 # Cohen, Jacob (1988). Statistical Power Analysis for the Behavioral Sciences. Routledge. ISBN 978-1-1
 # Sawilowsky, S (2009). "New effect size rules of thumb". Journal of Modern Applied Statistical Method
```

As is apparent from the T-test table and the descriptive information, the mean *drp* is observed to be higher for group=*"Treat"* than for group=*"Control"*. The location parameter equals the difference in the two sample means (i.e., *9.9545*), with a standard error of *4.3919*. The corresponding value for *Cohen's d* equals *-0.6841*, with a standard error of *0.3182* and a *95%* confidence interval ranging from *-1.2895* to *-0.0710*. According to Cohen's classification scheme, the value of *-0.6841* corresponds to an observed effect that is *"medium to large"*. [note to Arne: we need to add a reference on this]

The Brown-Forsythe test for equality of variances was *not [fork: omit "not"] significant at the .05 level, but we nevertheless [fork: and this is why we also]* report the results from the Welch test, which assumes that the variances in the two groups are unequal. The location parameter in the Welch test equals the difference in the two sample means and the associated standard error is *4.3076*. The corresponding value for *Cohen's d* equals *-0.6908*, with a standard error of 0.3185 and a 95% confidence interval ranging from *-1.2981* to *-0.0750*. According to Cohen's classification scheme, the value of *-0.6908* corresponds to an observed effect that is *"medium to large"*.

The Shapiro-Wilk test for normality was not [fork: omit "not"] statistically significant at the .05 level, but we nevertheless [fork: and this is why we also] report the result from the Mann-Whitney test, which is based only on the ranks of the observations; therefore, the Mann-Whitney test is relatively robust. The Mann-Whitney location parameter (i.e., the Hodges-Lehmann estimate) equals *-10.0001*. The Mann-Whitney effect size measure is the the rank biserial correlation; here it equals *-0.4410*, with a standard error of *0.1744* and a *95%* confidence interval that ranges from *-0.6745* to *-0.1274*.

For all estimates: the above confidence intervals do not identify a likely range of values for effect size. In order to obtain this information a Bayesian analysis would be needed (e.g., Morey et al., 2016; van den Bergh, 2021).

5. Hypothesis Testing: Is The Effect Absent?

For the standard t-test, the group difference is not statistically significant at the .05 level: p=.0286, t(42) = -2.2666. We may not reject the null-hypothesis of no population difference between the groups. of no population difference between the groups. [NB. this needs to be adjusted for a one-sided test] The Vovk-Sellke maximum p-Ratio of *3.6162* indicates the maximum possible odds in favor of H1 over H0, which is not compelling and urges caution. [only include for odds lower than 10]

For the Welch test, the group difference is not statistically significant at the .05 level: p=*.0264*, *t(37.8554)* = *-2.3109*. We may not reject the null-hypothesis of no population difference between the groups. [NB. this needs to be adjusted for a one-sided test]

For the Mann-Whitney test, the group difference is not statistically significant at the .05 level: p=*.0127*, rank biserial correlation = *-0.4410*. We may not reject the null-hypothesis of no population difference between the groups. [NB. this needs to be adjusted for a one-sided test]

For all tests: the p-value does not quantify evidence for the null hypothesis versus the alternative hypothesis; the p-value also cannot be taken to mean that the null hypothesis is either likely or unlikely to hold, or that the data are more or less likely to occur under the null hypothesis than under the alternative hypothesis. In order to obtain this information a Bayesian analysis would be needed.

6. Sources/References

Fisher, R (1955). "Statistical Methods and Scientific Induction". Journal of the Royal Statistical Society, Series B. 17 (1): 69–78.

Lumley, T., Diehr, P.; Emerson, S., Chen, L. (2002). "The Importance of the Normality Assumption in Large Public Health Data Sets". Annual Review of Public Health. 23 (1): 151–169. doi:10.1146/annurev.publhealth.23.100901.140546. ISSN 0163-7525. PMID 11910059.

Neyman, J, Pearson, E. S. (1933). "On the Problem of the most Efficient Tests of Statistical Hypotheses". Philosophical Transactions of the Royal Society A. 231 (694–706): 289–337.

Tests

Raincloud Plots Paper Kruskal-Wallis test

Mann-Whitney Non-Parametric Test / Wilcoxon Rank Sum David F. Bauer (1972). Constructing confidence sets using rank statistics. Journal of the American Statistical Association 67, 687–690. doi: 10.1080/01621459.1972.10481279.

Myles Hollander and Douglas A. Wolfe (1973). Nonparametric Statistical Methods. New York: John Wiley & Sons. Pages 27–33 (one-sample), 68–75 (two-sample). Or second edition (1999).

Shapiro-Wilk Normality Test Patrick Royston (1982). An extension of Shapiro and Wilk’s WW test for normality to large samples. Applied Statistics, 31, 115–124. doi:10.2307/2347973.

Levene’s test Fox, J. and Weisberg, S. (2019) An R Companion to Applied Regression, Third Edition, Sage.

Student’s t-test “Student” William Sealy Gosset (1908). “The probable error of a mean” (PDF). Biometrika. 6 (1): 1–25. doi:10.1093/biomet/6.1.1. hdl:10338.dmlcz/143545.

Welch’s t-test Welch, B. L. (1947). “The generalization of”Student’s” problem when several different population variances are involved”. Biometrika. 34 (1–2): 28–35. doi:10.1093/biomet/34.1-2.28