Tutorial: Addressing researcher degrees of freedom through minP adjustment

Maximilian Mandl

18-06-2024

1 General idea and intuition: minP

The minP procedure has the major advantage that it has a relatively intuitive principle, as illustrated by the following example. In a comment on a study by Mathews et al. (2008) claiming that breakfast cereal intake before pregnancy is positively associated with the probability to conceive a male fetus, Young et al. (2009) reinterpret the small p-value of 0.0034 obtained in the original article. They notice that Mathews et al. (2008) did not only analyse the association between fetal sex and the consumption of breakfast cereals, but also many other food items—a typical case of multiple testing. Based on the analysis of permuted data (i.e. data with randomly shuffled fetal sex status), Young et al. (2009) argue that "one would expect to see a p-value as small as 0.0034 approximately 28 percent of the time when nothing is going on". Implicitly, they apply the minP procedure for adjusting the smallest raw p-value of 0.0034 to 0.28 in this context where multiple tests are performed to investigate multiple food items. Our suggestion consists of translating this approach into the context of the analytical researcher degrees of freedom towards addressing the statistical factors of the replication crisis.

To make use of the minP adjustment method, we basically need two main ingredients: **(1)** the original unadjusted p-values based on the different analytical specifications and **(2)** the p-values obtained from the permuted datasets. If we then want to adjust the minimal p-value obtained from the different specifications, we count how often the minimal p-value is smaller for the permuted data sets compared to the unadjusted minimal p-value. This proportion is our final result and thus our adjusted p-value.

2 Toy example

Just imagine a researcher wants to perform a paired two-sided hypothesis test on two variables X_1 and X_2. X_2 includes missing values. It could be reasonable to either drop or mean impute the missing observations. Subsequently, we can choose between a paired t-test or a nonparametric paired Wilcoxon-test. We are left with four analytical specifications in total. Note that in reality the number of analytical specifications can be much higher and that the degree of adjustment is a function of the total number of specifications and the dependency between the hypotheses.

2.1 Create data

In the following we create a dataframe with two normally distributed variables ($n = 200$) and randomly sample 40 observations to create missing values using the *sample* function.

```
set.seed(1)
x1 = rnorm(n = 200, mean = 0, sd = 0.1)
x2 = rnorm(n = 200, mean = 0.05, sd = 0.1)
data = data.frame(x1 = x1, x2 = x2)
ind = sample(1:200,40)
data[ind,"x2"] = NA
```

We now either drop the missing observations or impute via the mean.

```
data.drop = data[complete.cases(data), ]
data.impute = data
data.impute$x2[is.na(data.impute$x2)] = mean(data.impute$x2,na.rm = TRUE)
```

2.2 Run initial analyses

Now, we apply a paired t-test and a paired Wilcoxon-test to both data sets resulting in four different initial p-values.

```
result1 = t.test(data.drop$x1,data.drop$x2,conf.level = 0.95, paired = TRUE)

result2 = wilcox.test(data.drop$x1, data.drop$x2, alternative = "two.sided",
                           conf.level = 0.95, paired = TRUE)

result3 = t.test(data.impute$x1,data.impute$x2,conf.level = 0.95, paired = TRUE)

result4 = wilcox.test(data.impute$x1, data.impute$x2, alternative = "two.sided",
                      conf.level = 0.95, paired = TRUE)

result1$p.value;result2$p.value;result3$p.value;result4$p.value
```

```
## [1] 8.677233e-06

## [1] 1.894192e-05

## [1] 1.614232e-07

## [1] 3.969446e-07
```

```
# save minimal p-value for comparison
p.minimal = result3$p.value
```

2.3 Permutation

Subsequently, we want to apply the minP-adjustment method. We run a simple for-loop with 1000 iterations to shuffle one of the two variables for every iteration i of the loop – in this case we shuffle the variable X_1. Here, we again use the *sample* function to randomly shuffle X_1 without replacement. Then we apply the hypothesis tests to each shuffled data set. We save the results for every iteration in a single vector. In every iteration i we get a single p-value for each specification resulting in four p-values per loop.

```
p = c()
for (i in 1:1000){
  set.seed(i)
  df1 = data.drop
  df1$x1 = df1$x1[sample(nrow(df1))]

  df2 = data.impute
  df2$x1 = df2$x1[sample(nrow(df2))]

  result1 = t.test(df1$x1,df1$x2,conf.level = 0.95,paired = TRUE)

  result2 = wilcox.test(df2$x1, df2$x2, alternative = "two.sided",
                        conf.level = 0.95, paired = TRUE)

  result3 = t.test(df2$x1,df2$x2,conf.level = 0.95, paired = TRUE)
```

```r
  result4 = wilcox.test(df2$x1, df2$x2, alternative = "two.sided",
                        conf.level = 0.95, paired = TRUE)

  p = c(p,result1$p.value,result2$p.value,result3$p.value,result4$p.value)
}
```

2.4 Extract minimal p-values

Note that we have to reshape this single vector into a matrix. The rows of the matrix represent our analytical strategies while the columns represent the number of shuffled datasets.

```r
p.permuted = as.data.frame(matrix(p, ncol = 1000,  byrow = FALSE), stringsAsFactors = FALSE)
```

Subsequently, we extract the minimal p-values for every permutation.

```r
min.permuted = apply(p.permuted,2,min)
```

2.5 Adjust p-values

If we want to adjust the smallest p-value from our initial analyses ($p = 1.614232 * 10^{-7}$) we simply count how often the minimal p-values from the permuted data sets are smaller than the original p-value.

```r
sum(min.permuted < p.minimal)/1000
```

```
## [1] 0.829
```

In our simple toy example the original p-value gets adjusted to $p = 0.829$ which makes sense in terms of our data generation.

3 References

Mathews, F., Johnson, P.J., Neil, A.: You are what your mother eats: evidence for maternal preconception diet influencing foetal sex in humans. Proceedings of the Royal Society B: Biological Sciences 275(1643), 1661–1668 (2008).

Young, S.S., Bang, H., Oktay, K.: Cereal-induced gender selection? most likely a multiple testing false positive. Proceedings of the Royal Society B: Biological Sciences 276(1660), 1211–1212 (2009).