

assignment_2

Ted Ladas - s2124289

18/02/2021

Question: In an experiment, the following survival times (in minutes) of the peroneal nerves of four cats and ten rabbits were recorded under a certain condition.

$c = (25, 33, 43, 45)$

$r = (15, 16, 16, 17, 20, 23, 28, 28, 35, 35)$

We want to do a permutation test based on sample means at $\alpha = 0.05$, to investigate whether these samples are from the same distribution, and use $\sum_{i=1}^4 c_i$ as the test statistic.

- (a) Define your null and alternative hypotheses.

We have C, R random variable, with samples c, r defined above.

Suppose that $C \sim F(\lambda_1)$ and $R \sim F(\lambda_2)$, where λ_1, λ_2 are parameters and F distribution.

$H_0 : \lambda_1 = \lambda_2$

$H_1 : \lambda_1 \neq \lambda_2$

We will perform a permutation test to assess this hypothesis.

- (b) Why can we use $\sum_{i=1}^4 c_i$ as the test statistic instead of $d = \bar{c} - \bar{r}$?

We can use the $\sum_{i=1}^4 c_i$ as the test statistic instead of $d = \bar{c} - \bar{r}$, because in our case these two are exactly equivalent. Since we are performing a *permutation test*, whether we measure how many times in each permutation the $\sum_{i=1}^4 c_i$ is bigger than our given $\sum_{i=1}^4 c_i$, or how many times d is bigger than our observed d is the same thing.

- (c) Find the p-value and state clearly what you can conclude from the result of your test.

These are the permutations for which, the $\sum_{i=1}^4 c_i$ is bigger or equal than the given sample.

	x 1	x 2	x 3	x 4	TOTAL
Obs	25	33	43	45	146
1	25	43	45	35	148
2	25	43	45	35	148
3	33	43	45	28	149
4	33	43	45	28	149
5	33	43	45	35	156
6	33	43	45	35	156
7	33	45	35	35	148
8	43	45	28	35	151
9	43	45	28	35	151
10	43	45	28	35	151
11	43	45	28	35	151
12	43	45	35	35	158

There are 4 values in C and 14 possible value in total in C and R , therefore the total number of cases where $\sum_{i=1}^4 c_i$ is smaller than the original is 1001. Therefore the p-value for the one sided test is: $p_0 = \frac{13}{1001} \approx 0.013$. The p-value for our two-sided test is: $p = 2 * p_0 \approx 0.025$. Since $p < \alpha = 0.05$, we can *reject* the Null Hypothesis. That mean that according to our permutation test, C and R , *don't* come from the same distribution.

Now suppose instead we wish to carry out the Mann-Whitney U-test

- (d) Define the test hypotheses.

According to the notation introduced in question (a) the test hypothesis for the Mann-Whitney U-test is:

$$H_0 : F_1 = F_2$$

$$H_1 : F_1 \neq F_2$$

We can carry out the Mann-Whitney U-test, since all its assumptions are fulfilled.

Assumptions:

- C, R samples are independent. This is true because they represent different species.
- The data is at least ordinal. This is true because the data can take continuous values.

- (e) Find the observed value of the test statistic U (without using **R**).

$U = \sum_{i=1}^4 \sum_{j=1}^{10} s(c_i, r_j)$, where

$$s(c_i, r_j) = \begin{cases} 1, & c_i > r_j \\ \frac{1}{2}, & c_i = r_j \\ 0, & c_i \leq r_j \end{cases}$$

Therefore we rank our observations to get the U score manually.

DATA	15	16	16	17	20	23	25	28	28	33	35	35	43	45
Variable	R	R	R	R	R	R	C	R	R	C	R	R	C	C
U-score	0	0	0	0	0	0	6	0	0	8	0	0	10	10

The final U score is the sum of the last row of the table: $\therefore U = 34$

- (f) Use R to find the (approximate) p-value and compare the result with that of part (c) for $\alpha = 0.05$.

```
# helper
m1 = 'The 2-sided p-value is:'
reject_message = '\nTherefore we reject the Null Hypothesis'
accept_message = '\nTherefore we do not have evidence to reject the Null Hypothesis'
alpha = 0.05

# data
c = c(25,33,43,45)
r = c(15,16,16,17,20,23,28,28,35,35)
mw_test = wilcox.test(c,r, alternative = "two.sided")

## Warning in wilcox.test.default(c, r, alternative = "two.sided"): cannot compute
## exact p-value with ties

if(mw_test$p.value<alpha){cat('\n', m1, mw_test$p.value, reject_message, '\n')}
}else{cat('\n', m1, mw_test$p.value, accept_message, '\n')}

##
## The 2-sided p-value is: 0.05542703
## Therefore we do not have evidence to reject the Null Hypothesis

mw_test

##
## Wilcoxon rank sum test with continuity correction
##
## data: c and r
## W = 34, p-value = 0.05543
## alternative hypothesis: true location shift is not equal to 0
```

Comment on results:

The results produced by R and the manual calculation are vastly different since in the first case we rejected the Null hypothesis and in the other we cannot reject it. An important point is the fact that the Mann-Whitney U-test produced an **approximate p-value**, due to the fact that there were ties on the data. That means that it based the p-value calculation on a *normal approximation*. The normal approximation cannot be very accurate in this circumstance, since we only have cumulatively 14 data points. Therefore, we believe that the permutation test is more accurate under the given framework, with a conclusion that the times of survival of Cats and Rabbits come from a **different distribution**. As a second point the below Kernel Density Estimation plot of the data shows that there is a possibility both distributions are bi-modal, since they don't form a clear peak, strengthening the claim that the Normal approximation wouldn't be the correct tool to conclude this analysis. The normal approximation of a distribution is valid and a *good* approximation only around the mean value of the distribution, but we suspect that our distributions might have multiple peaks. We avoided to draw this diagram before the start of the analysis, and we only explored it in the end, to have a conclusive answer, as to not influence ourselves in any way.

The code of the analysis can be found on github at: https://github.com/TedOiler/APS/tree/master/assignment_2.

This is a **private** repository, that I will make public only after the submission deadline of the assessment.