

# MATH-650 Assignment 4

Saket Choudhary (USCID: 2170058637) ([skchoudh@usc.edu](mailto:skchoudh@usc.edu))

09/24/2015

## Problem 4.14[O-ring study]

```
problem4 <- read.csv('problem_4.csv')
below65 <- problem4[problem4$Launch == 'Cool',]$Incidents
above65 <- problem4[problem4$Launch == 'Warm',]$Incidents
n1 <- length(below65)
n2 <- length(above65)
n <- n1+n2
```

### t-statistic [no permutation]

```
ttest.nopermutation <- t.test(below65, above65, var.equal = T)
```

Observed t-statistic: 3.8876498

p-value:  $7.9291726 \times 10^{-4}$

### t-statistics [permutation testing]

```
library(coin)
```

```
## Loading required package: survival
```

```
combined_data = c(below65, above65)
labels <- factor(rep(c("below65", "above65"), c(length(below65), length(above65))))
ttest.withpermutation <- oneway_test(combined_data ~ labels,
                                     alternative="less",
                                     distribution=approximate(B=9999))
```

```
## Note we need to use "alternative=less" as the null hypothesis tested
## here is H_0: theta >=0 and theta here is below65-above65 rather than
## other way round
```

Rather than an exact calculation we use an “approximate” pvalue by Monte carlo sampling 10000 times. So the textbook states the oneway p-value for permutation test to be 0.00988 while, we get 0.0108011 which is quite close to the stated value in textbook. However it is different from the p-value obtained with no permutation test  $7.9291726 \times 10^{-4}$ . In this case, where the data cannot be assumed to follow a t-distribution, the p-value with no permutation test is incorrect.

## Problem 4.15

```
library(combinat)
```

```
##  
## Attaching package: 'combinat'  
##  
## The following object is masked from 'package:utils':  
##  
##      combn
```

```
problem15.group1 <- c(1,5)  
problem15.group2 <- c(4,8,9)  
problem15.data <- c(problem15.group1,problem15.group2)  
diff.original <- sum(problem15.group1)-sum(problem15.group2)  
all_combinations <- combn(problem15.data, length(problem15.group1), simplify = T)  
count <- 0  
for (i in 1:dim(all_combinations)[2]){  
  sety1 <- all_combinations[,i]  
  sety2 <- setdiff(problem15.data, sety1)  
  diff <- sum(sety1)-sum(sety2)  
  if (diff<=diff.original){  
    count <- count+1  
    print(sety1)  
    print(sety2)  
  }  
}
```

```
## [1] 1 5  
## [1] 4 8 9  
## [1] 1 4  
## [1] 5 8 9
```

```
pvalue <- count/dim(all_combinations)[2]  
pvalue
```

```
## [1] 0.2
```

p-value for  $\bar{Y}_1 - \bar{Y}_2 = 0.2$

## Problem 4.31

```
problem31 <- read.csv('ex0431.csv')  
problem31.control <- problem31[problem31$Group=='Control',]  
problem31.therapy <- problem31[problem31$Group=='Therapy',]  
wtest <- wilcox.test(problem31.therapy$Survival, problem31.control$Survival, paired = F, alternative =  
wtest
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data:  problem31.therapy$Survival and problem31.control$Survival
## W = 479, p-value = 0.1325
## alternative hypothesis: true location shift is greater than 0
```

Since the data was 'censored', Wilcoxon rank sum test was one of the natural choices for testing the null hypothesis that the therapy patients live longer than the control patients. Thus, from the analysis of control and therapy patients, there is very little statistical significance of the therapy patients having more survival rate than the control patients (p-value of 0.1324824). Since the evidence is not statistically significant it is not possible to rule out that the both sets of patients have similar survival time and hence we do not go further with the additive model to find the 95% CI for  $\delta$ .