Conformal Prediction on Prevalence

2023-07-05

Creating prediction intervals on prevalence using data from 2020 - we report the case for males

```
y_f_list=c(rawdata_curr[rawdata_curr$YEAR..DISPLAY.=='2020',2:37])
y_f<-unlist(y_f_list)
n<-length(y_f)

#Selecting grid parameters
grid_factor = 1.5
n_grid = 400
alpha = 0.1

#Creating the test grid
test_grid = seq(0, +grid_factor * max(abs(y_f)),
                length.out = n_grid)


########Utils
#Plot p-value function
plot_pval = function(test_grid, pval_fun, pred, alpha) {
  plot(
    test_grid,
    pval_fun,
    type = 'l',
    main = "p-value function",
    xlab = "Test grid",
    ylab = "p-value function"
  )
  abline(h = alpha, lty = 2)
  abline(v = pred, col = 'red')
}
########End utils
```

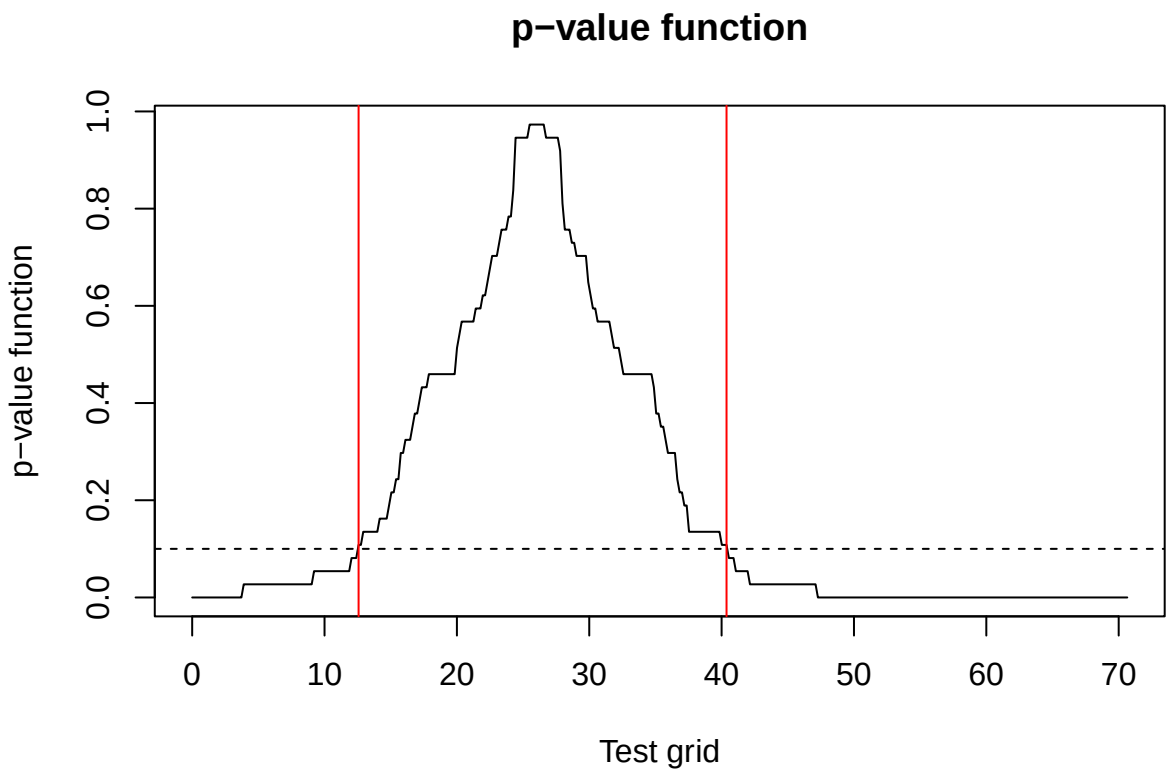
T-prediction intervals

```
wrapper_full = function(grid_point) {
  aug_y = c(grid_point, y_f)
  mu = mean(aug_y)
  ncm = abs(mu - aug_y)###
  sum((ncm[-1] >= ncm[1])) / (n + 1)
}

pval_fun = sapply(test_grid, wrapper_full)
index_in = pval_fun > alpha
pred_t_interval = range(test_grid[index_in])
pred_t_interval
```

```
## [1] 12.57180 40.37143
```

```
plot_pval(test_grid, pval_fun, pred_t_interval, alpha)
```



KNN prediction intervals using k=0.3*n

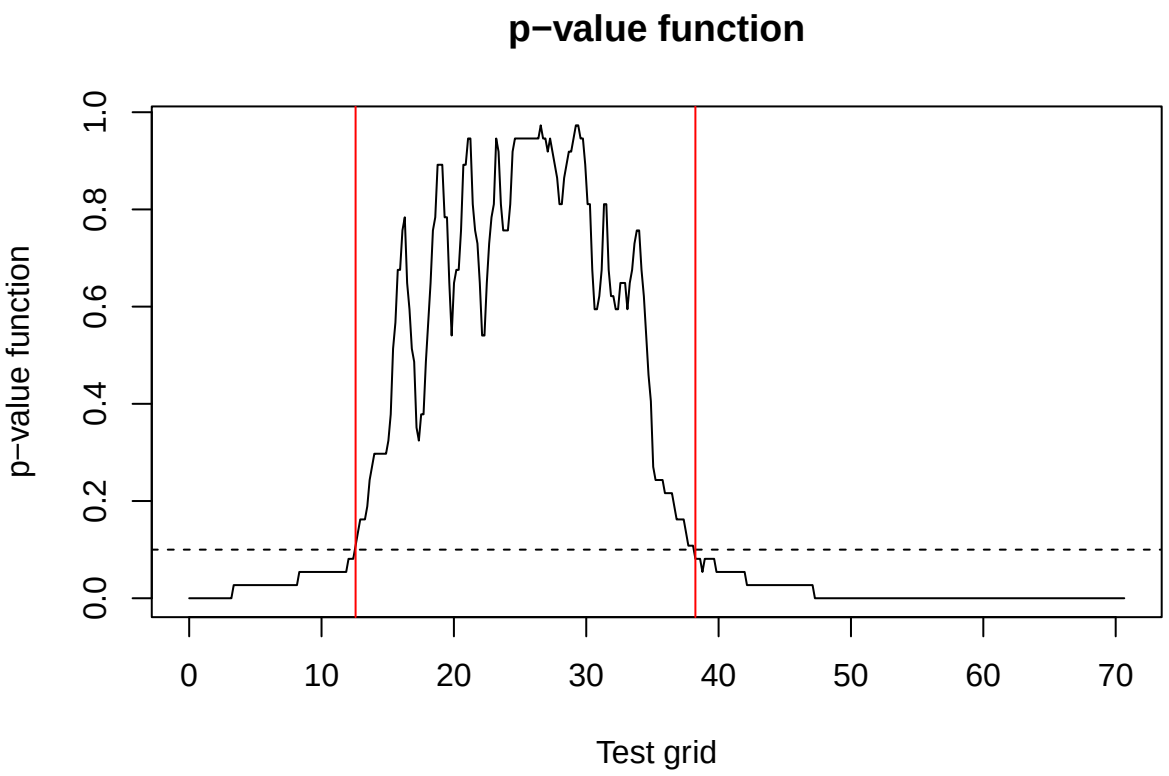
```
pval_fun = numeric(n_grid)
k = 0.3

wrapper_knn = function(grid_point) {
  aug_y = c(grid_point, y_f)
  ncm = kNNdist(matrix(aug_y), k * n)
  sum((ncm[-1] >= ncm[1])) / (n + 1)
}

pval_fun = sapply(test_grid, wrapper_knn)
index_in = pval_fun > alpha
pred_knn = test_grid[as.logical(c(0, abs(diff(index_in))))]
pred_knn
```

```
## [1] 12.57180 38.24662
```

```
#Plot p-value function
plot_pval(test_grid, pval_fun, pred_knn, alpha)
```



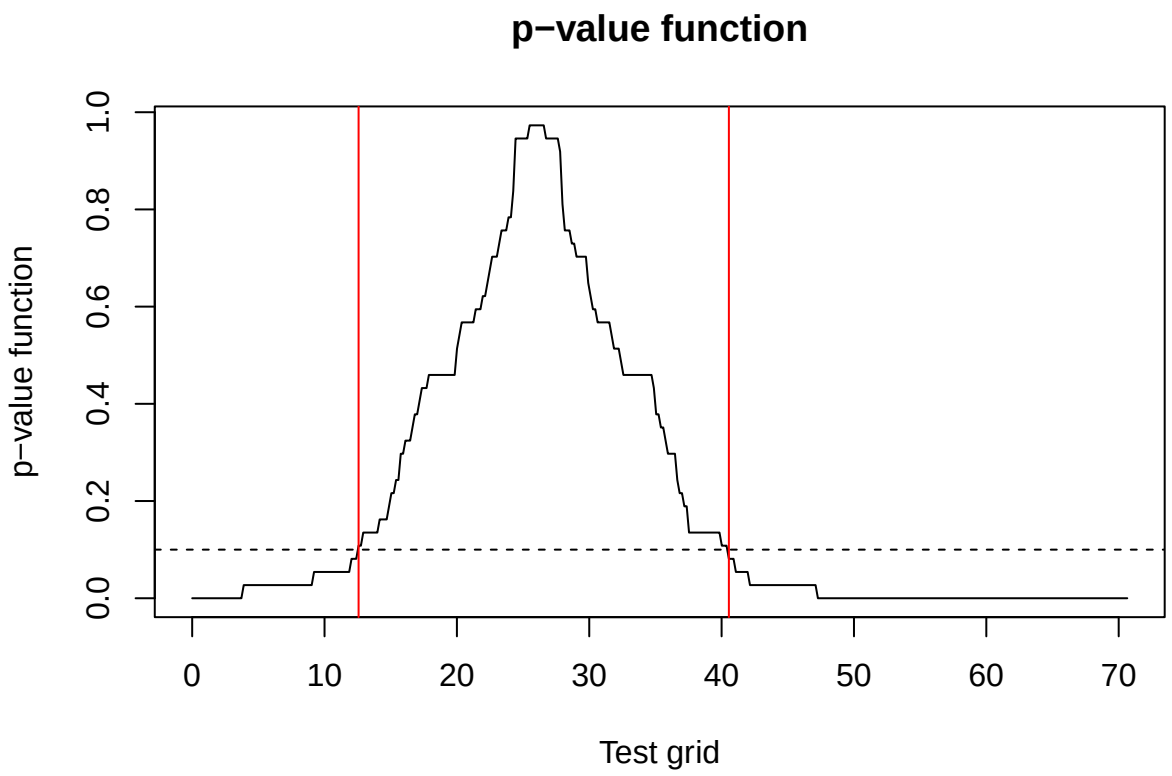
Using Mahalanobis distance

```
pval_fun = numeric(n_grid)
wrapper_mal = function(grid_point) {
  aug_y = c(grid_point, y_f)
  ncm = mahalanobis(matrix(aug_y), colMeans(matrix(aug_y)), cov(matrix(aug_y)))
  sum((ncm[-1] >= ncm[1])) / (n + 1)
}

pval_fun = sapply(test_grid, wrapper_mal)
index_in = pval_fun > alpha
pred_mahalanobis = test_grid[as.logical(c(0, abs(diff(index_in))))]
pred_mahalanobis
```

```
## [1] 12.5718 40.5485
```

```
#Plot p-value function
plot_pval(test_grid, pval_fun, pred_mahalanobis, alpha)
```

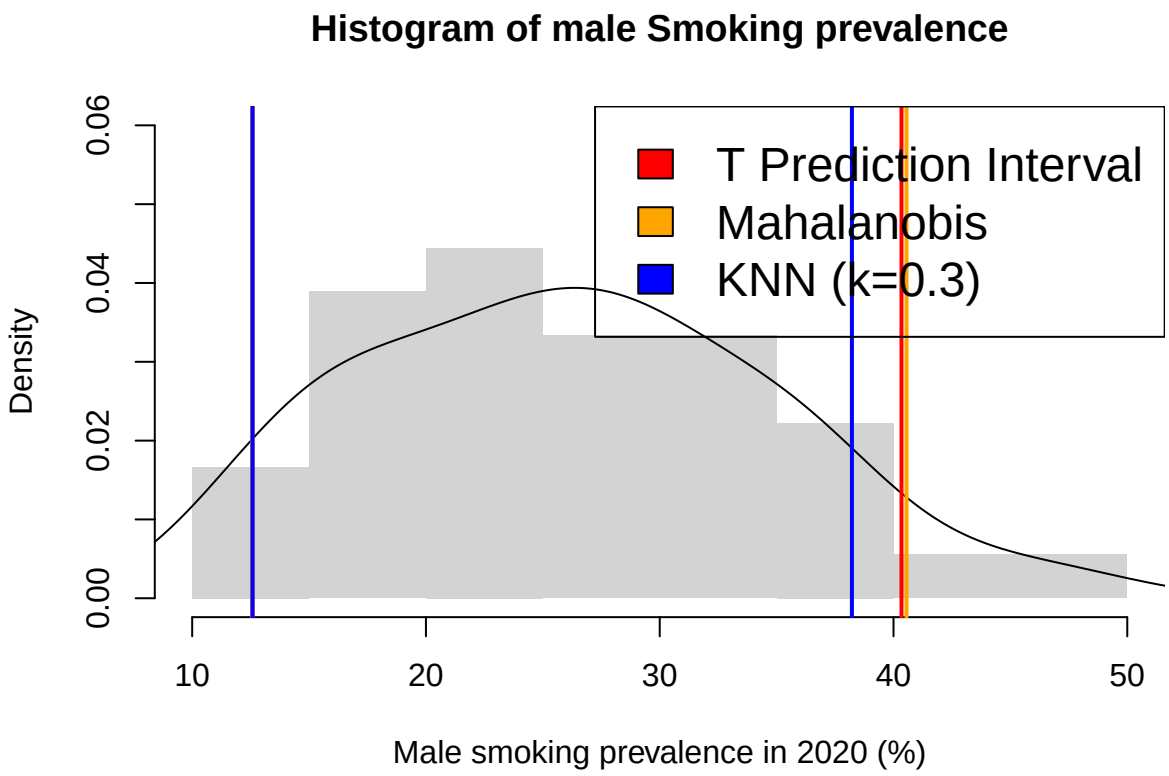


Comparison of the intervals

```
hist(
  y_f,
  breaks = 12,
  freq = FALSE,
  main = 'Histogram of male Smoking prevalence',
  xlab = 'Male smoking prevalence in 2020 (%)',
  ylim= c(0,0.06),
  border = NA
)
lines(density(y_f))

abline(v = jitter(pred_t_interval, amount=0.03), col = 'red', lwd = 2)
abline(v = jitter(pred_mahalanobis, amount=0.03), col = 'orange', lwd = 2)
abline(v = jitter(pred_knn, amount=0.03), col = 'blue', lwd = 2)

legend("topright",
       legend = c("T Prediction Interval", "Mahalanobis", "KNN (k=0.3)"),
       fill = c("red", "orange", "blue"),
       cex=1.5)
```


Histogram of male Smoking prevalence
T Prediction Interval
Mahalanobis
KNN (k=0.3)
Density
0.00
0.02
0.04
0.06
10
20
30
40
50
Male smoking prevalence in 2020 (%)