Theoretical foundations of the analysis of large data sets - report 3

Łukasz Rębisz

2024-04-06

Introduction

The aim of this report is the comparison of **Tests Based on Empirical CDF's**:

(a) Higer Criticism Test,

(b) Modified Higer Criticism Test,

(c) Kolmogorov-Smirnov Test,

(d) Anderson-Darling Test.

Definitions

The empirical CDF is the function:

$$F_n(t) = \frac{\sum V_i}{n},$$

where $V_i(t) = 1_{(p_i \leq t)}$.

- Higher Criticism Test (Tukey 1976):

$$HC^* = \max_{1/n<t<1/2} \sqrt{n}\frac{F_n(t)-t}{\sqrt{t(1-t)}}.$$

- Modification of HC Test by Stepanova and Pavlenko (2014):

$$HC_{mod} = \max_{0<t<1} \sqrt{n}\frac{F_n(t)-t}{\sqrt{t(1-t)q(t)}}, \quad q(t) = \log\log\frac{1}{t(1-t)}.$$

Exercise 1

For $n \in \{5000; 50000\}$ we will estimate the probability of the type I error for HC_{mod} using the asymptotic critical value for $\alpha = 0.05$ significance test $C_{crit} = 4.14$.

```
HC_mod <- function(n){
  p <- runif(n)
  k <- 10000
  vec_t <- seq(1/k, 1-1/k, 1/k)
  HC <- max(sqrt(n)*((ecdf(p)(vec_t) - vec_t)/
                      sqrt(vec_t*(1-vec_t)*log(log(1/(vec_t*(1-vec_t)))))))
  return(HC)
}
```

```
vec_HC_mod_5000 <- replicate(1000, HC_mod(5000))
vec_HC_mod_50000 <- replicate(1000, HC_mod(50000))
C_crit <- 4.14

HC_mod_Type_I_error_5000 <- mean(vec_HC_mod_5000 > C_crit)
HC_mod_Type_I_error_50000 <- mean(vec_HC_mod_50000 > C_crit)
```

We get the following results for the probability of Type I Error for the HC_{mod} test:

- P(Type I Error) $= 0.058 \leq 0.05$, when $n = 5000$,
- P(Type I Error) $= 0.045 \leq 0.05$, when $n = 50000$.

In both cases P(Type I Error) is not greater than the significance level 0.05, so the HC_{mod} test is correctly constructed.

Exercise 2

In this paragraph we will estimate (for $n = 5000$) **critical values** of both Higer-Criticism tests at the significance level $\alpha = 0.05$.

```
alpha <- 0.05
n <- 5000

HC <- function(n){
  p <- runif(n)
  k <- 10000
  vec_t <- seq(1/k, 0.5, 1/k)
  HC <- max(sqrt(n)*((ecdf(p)(vec_t) - vec_t)/sqrt(vec_t*(1-vec_t))))
  return(HC)
}

vec_HC_5000 <- replicate(1000, HC(5000))

q_HC_mod <- as.numeric(quantile(vec_HC_mod_5000, 1 - alpha))
q_HC <- as.numeric(quantile(vec_HC_5000, 1 - alpha))
```

The cirtical value ($\alpha = 0.05$) is equal:

- 3.32 for HC test,
- 4.23 for HC_{mod} test (theoretical value from Exercise 1 $C_{crit} = 4.14$).

Exercise 3

Let $n = 5000$ and

$$X_1, ..., X_n \sim N(\mu_i, 1),$$

a) $\mu_1 = 1.2\sqrt{2\log n}, \mu_2 = ... = \mu_n = 0;$

b) $\mu_1 = ... = \mu_{100} = 1.02\sqrt{2\log \frac{n}{200}}, \mu_{101} = ... = \mu_n = 0;$

c) $\mu_1 = ... = \mu_{1000} = 1.002\sqrt{2\log \frac{n}{2000}}, \mu_{1001} = ... = \mu_n = 0.$

We will use the above settings to **compare the power of the following tests**:

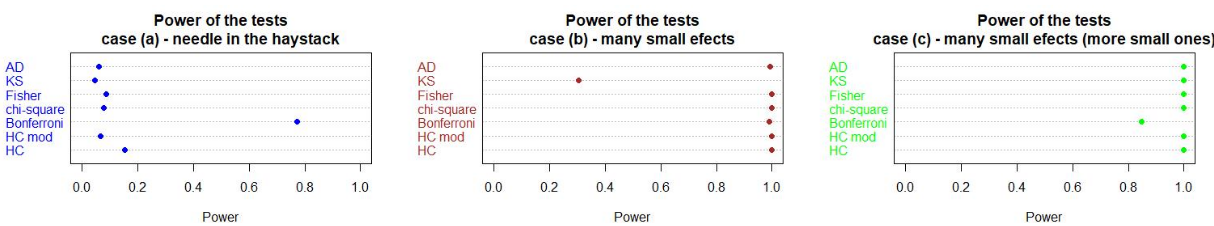
- Higher-Criticism,

- modified Higher-Criticism,
- Bonferroni,
- chi-square,
- Fisher,
- Kolmogorov-Smirnov,
- Anderson-Darling.

```r
k <- 10000

n <- 5000

ex_3 <- function(){
  X_a <- c(rnorm(1, mean = 1.2*sqrt(2*log(n))), rnorm(n-1))
  X_b <- c(rnorm(100, mean = 1.02*sqrt(2*log(n/200))), rnorm(n-100))
  X_c <- c(rnorm(1000, mean = 1.002*sqrt(2*log(n/2000))), rnorm(n-1000))

  p_a <- 2*(1-pnorm(abs(X_a)))
  p_b <- 2*(1-pnorm(abs(X_b)))
  p_c <- 2*(1-pnorm(abs(X_c)))

  # Higher-Criticism
  vec_t <- seq(1/k, 0.5, 1/k)
  HC_a <- max(sqrt(n)*((ecdf(p_a)(vec_t) - vec_t)/sqrt(vec_t*(1-vec_t))))
  decision_HC_a <- mean(HC_a > q_HC)
  HC_b <- max(sqrt(n)*((ecdf(p_b)(vec_t) - vec_t)/sqrt(vec_t*(1-vec_t))))
  decision_HC_b <- mean(HC_b > q_HC)
  HC_c <- max(sqrt(n)*((ecdf(p_c)(vec_t) - vec_t)/sqrt(vec_t*(1-vec_t))))
  decision_HC_c <- mean(HC_c > q_HC)

  # Modified Higher-Critisim
  vec_t <- seq(1/k, 1-1/k, 1/k)
  HC_mod_a <- max(sqrt(n)*((ecdf(p_a)(vec_t) - vec_t)/
                       sqrt(vec_t*(1-vec_t)*log(log(1/(vec_t*(1-vec_t)))))))
  decision_HC_mod_a <- mean(HC_mod_a > q_HC_mod)
  HC_mod_b <- max(sqrt(n)*((ecdf(p_b)(vec_t) - vec_t)/
                       sqrt(vec_t*(1-vec_t)*log(log(1/(vec_t*(1-vec_t)))))))
  decision_HC_mod_b <- mean(HC_mod_b > q_HC_mod)
  HC_mod_c <- max(sqrt(n)*((ecdf(p_c)(vec_t) - vec_t)/
                       sqrt(vec_t*(1-vec_t)*log(log(1/(vec_t*(1-vec_t)))))))
  decision_HC_mod_c <- mean(HC_mod_c > q_HC_mod)

  # Bonferroni
  decision_BF_a <- min(p_a) <= alpha/n
  decision_BF_b <- min(p_b) <= alpha/n
  decision_BF_c <- min(p_c) <= alpha/n

  # Chi-square
  decision_chi_sq_a <- sum(X_a^2) > qchisq(1-alpha, df = n)
  decision_chi_sq_b <- sum(X_b^2) > qchisq(1-alpha, df = n)
  decision_chi_sq_c <- sum(X_c^2) > qchisq(1-alpha, df = n)

  # Fisher
```

```
  decision_Fisher_a <- -2*sum(log(p_a)) > qchisq(1-alpha, df = 2*n)
  decision_Fisher_b <- -2*sum(log(p_b)) > qchisq(1-alpha, df = 2*n)
  decision_Fisher_c <- -2*sum(log(p_c)) > qchisq(1-alpha, df = 2*n)

  # Kolmogorov-Smirnov
  decision_KS_a <- ks.test(p_a, runif(n))$p.value <= alpha
  decision_KS_b <- ks.test(p_b, runif(n))$p.value <= alpha
  decision_KS_c <- ks.test(p_c, runif(n))$p.value <= alpha

  # Anderson-Darling
  decision_AD_a <- ad.test(p_a, "punif")$p.value <= alpha
  decision_AD_b <- ad.test(p_b, "punif")$p.value <= alpha
  decision_AD_c <- ad.test(p_c, "punif")$p.value <= alpha

  return(c(decision_HC_a, decision_HC_b, decision_HC_c,
         decision_HC_mod_a, decision_HC_mod_b, decision_HC_mod_c,
         decision_BF_a, decision_BF_b, decision_BF_c,
         decision_chi_sq_a, decision_chi_sq_b, decision_chi_sq_c,
         decision_Fisher_a, decision_Fisher_b, decision_Fisher_c,
         decision_KS_a, decision_KS_b, decision_KS_c,
         decision_AD_a, decision_AD_b, decision_AD_c))
}
```

The power of the investigating tests



The above results show us that:

- in the case (a) - needle in the haystack - the most powerful test is the Bonferroni: power around 0.8. The other tests have much smaller power in this case ($0 < \text{power} < 0.2$). This result is consistent with the theory, because the Bonferroni test is created for the neddle in the haystack problem.
- In the case (b) be we have 100 efects (total number of statistics $n = 5000$). The Kolmogorov-Smirnov test is powerless in this case (power around 0.3). The other investigating tests have full power which is equal around 1.
- The results for the case (c) - 1000 small effects - are different. In this case the Bonferroni test has smaller power which is equal around 0.85. The other tests have full power.

We can see that the Bonferroni test is the best one in the neddle in the haystack problem.

When we have many small effects we shouldn't use the Bonferroni and Kolmogorov-Smirnov test which don't have the full power unlike tests: Anderson-Darling, Fisher, chi-square, HC and modified HC.

Exercise 4

In this paragraph we will consider the **sparse mixture model**:

$$X_i \sim f(\mu) = (1 - \epsilon)\delta_0 + \epsilon \cdot \delta_\mu, \quad i = 1, ..., n,$$

where $\delta_0 \sim N(0, 1), \quad \delta_\mu \sim N(\mu, 1)$

with $\epsilon = n^{-\beta}$ and $\mu = \sqrt{2r \log n}$.

We are investigating the following values:

a) the **critical values** for the Neyman-Pearson test in the sparse mixture;

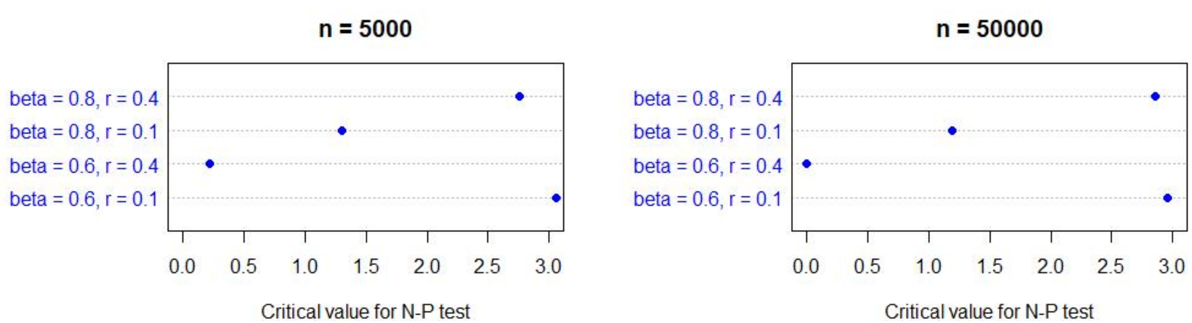
b) the **power** of the Neyman-Pearson test to the power of both versions of HC, Bonferroni, Fisher and chi-square;

for each setting $\beta \in \{0.6; 0.8\}$, $r \in \{0.1; 0.4\}$ and $n \in \{5000; 50000\}$.

a)

```
L <- function(n, beta, r){
  epsilon <- n^(-beta)
  mi <- sqrt(2*r*log(n))
  vec_x <- rnorm(n)
  prod(1-epsilon + epsilon*exp(mi*vec_x - 0.5*mi^2))
}

L_quantile <- function(M, n, beta, r, alpha){
  epsilon <- n^(-beta)
  vec_L <- replicate(M, L(n, beta, r))
  return(as.numeric(quantile(vec_L, 1 - alpha)))
}
```

The critical values for the Neyman-Pearson test in the sparse mixture



The above results show us that:

- There is no significant difference between case for $n = 5000$ and $n = 50000$: the corresponding critical values are similiar.
- When $\beta = 0.8$ the sparse mixture model is close to the neddle in the haystack problem ($\beta = 1, r = 1$). In this case the critical value for the N-P test is higher for the greater parameter $r = 0.4$.
- On the other hand, when $\beta = 0.6$ (many small effects) we get the higher critical value for the smaller $r = 0.1$.

b)

Let's compare the power of the Neyman-Pearson test to the power of both versions of HC, Bonferroni, Fisher and chi-square.

```r
alpha <- 0.05
k <- 10000

ex_2_b <- function(n, beta, r, q_L){
  epsilon <- n^(-beta)
  mi <- sqrt(2*r*log(n))
  vec_epsilon <- rbinom(n, 1, epsilon)
  vec_x <- rnorm(n)*(1-vec_epsilon) + vec_epsilon*rnorm(n, mi, 1)

  # Neyman-Pearson
  L_stat <- prod(1-epsilon + epsilon*exp(mi*vec_x - 0.5*mi^2))
  decision_NP <- L_stat >= q_L

  p <- 2*(1-pnorm(abs(vec_x)))

  # Higher-Criticism
  vec_t <- seq(1/k, 0.5, 1/k)
  HC <- max(sqrt(n)*((ecdf(p)(vec_t) - vec_t)/sqrt(vec_t*(1-vec_t))))
  decision_HC <- mean(HC > q_HC)

  # Modified Higher-Critisim
  vec_t <- seq(1/k, 1-1/k, 1/k)
  HC_mod <- max(sqrt(n)*((ecdf(p)(vec_t) - vec_t)/
                         sqrt(vec_t*(1-vec_t)*log(log(1/(vec_t*(1-vec_t)))))))
  decision_HC_mod <- mean(HC_mod > q_HC_mod)

  # Bonferroni
  decision_BF <- min(p) <= alpha/n

  # Chi-square
  decision_chi_sq <- sum(vec_x^2) > qchisq(1-alpha, df = n)

  # Fisher
  decision_Fisher <- -2*sum(log(p)) > qchisq(1-alpha, df = 2*n)

  return(c(decision_NP, decision_HC, decision_HC_mod, decision_BF,
           decision_chi_sq, decision_Fisher))
}
```

We know that the most powerful Neyman-Pearson test in the sparse mixture model is powerful when:

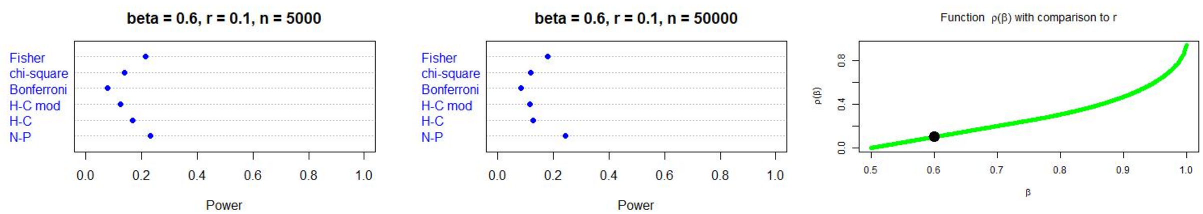
$$r > \rho^*(\beta),$$

where:

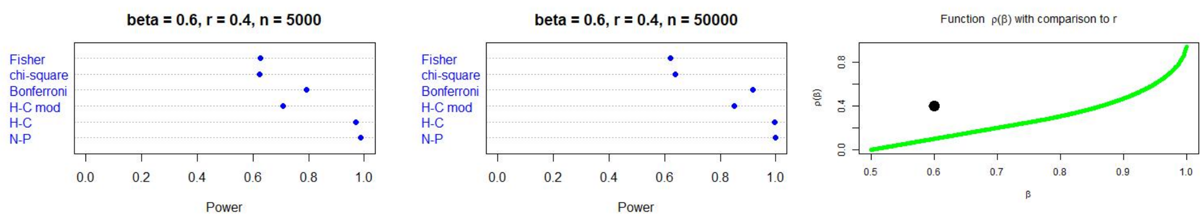
- $\rho^*(\beta) = \beta - 1/2, \quad \frac{1}{2} < \beta \leq \frac{3}{4},$
- $\rho^*(\beta) = (1 - \sqrt{1-\beta})^2, \quad \frac{3}{4} \leq \beta \leq 1.$

```r
vec_beta <- seq(0.5001, 1, 0.001)

rho <- function(beta){
  if(beta>0.5 && beta <= 0.75) return(beta - 0.5)
  if(beta>0.75 && beta <= 1) return((1 - sqrt(1-beta))^2)
  else return(0)
}

rho_vec <- Vectorize(rho)
```

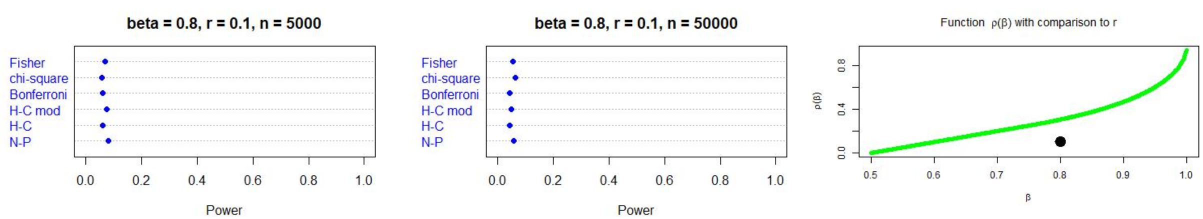
The power of the investigating tests



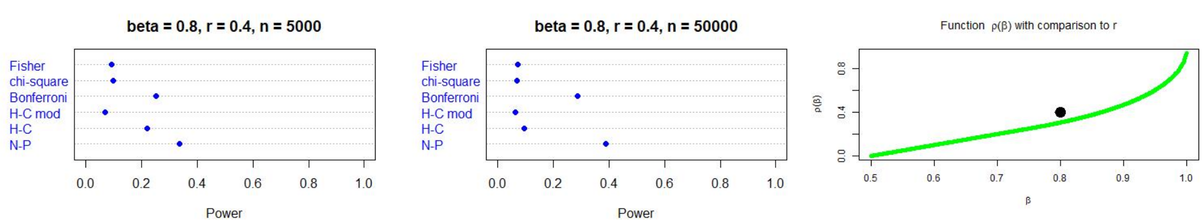
In this case $r = 0.1 = \rho^*(\beta)$. The powers of the corresponding test are similiar for both sizes n. We can see that the most powerful test is the Neyman-Pearson test (consistent with the theory). Also the Fisher test has the power which is close to the power of N-P test. On the other hand: the Bonferroni test is powerless in this case.



We can see that in this case $r > \rho^*(\beta)$, so the N-P test is powerful. Moreover, the power for each test increases with the increase of n. Each of the investigating test has power over 0.5, but only the power for the H-C test reaches value which is close to 1 like the power of N-P test.



This case show us that for $r < \rho^*(\beta)$ the Neyman-Person test (the most powerful test) is powerless and so the each other test is also powerless. Moreover, the powers of tests decrease with the increase of n.



In the last case we have again the situation when $r > \rho^*(\beta)$ and the Neyman-Pearson test should be powerful. We can see that the power of N-P is greater than for the other test, but it reaches the level 0.4. This result is probably caused by the small difference between r and $\rho^*(\beta)$. The theory sais that N-P test is powerful asymptotically when $r > \rho^*(\beta)$ and we can see that indeed the power of N-P (and each other investigating test) increases with the increase of n.