MATH 534 DA Homework 9

Homework Solution

Generate n=200 observations from N(2, 9) with 400 replications. Find out the empirical 90th and 95th percentiles of the ks.test statistics D in the following three cases:

(1) ks.test with parameter being the true value;

(2) ks.test with parameter being the MLE;

(3) use the modified ks.test;

Make a comment on your findings.

Solution:

```
> library(MASS)
> set.seed(1000)
> m=400
> n=100
> s_true <- rep(0,m)
> s_mle <- rep(0,m)
> s_90_modified <- rep(0,m)
> s_95_modified <- rep(0,m)
> for (i in 1:m){
+   #generate 400 data sets
+   data <- rnorm(200,2,3)
+   mle <- fitdistr(data,"normal")
+   s <- rep(0,n)
+   #100 modified ks-test for each data set
+   for (j in 1:n){
+     normal_generated <- rnorm(200,mle$e[1],mle$e[2])
+     mle_mod <- fitdistr(normal_generated,"normal")
+     s[j] <- ks.test(normal_generated,"pnorm",mle_mod$e[1],mle_mod$e[2])$s
+   }
+   s_90_modified[i] <- quantile(s,0.9)
+   s_95_modified[i] <- quantile(s,0.95)
+   s_true[i] <- ks.test(data,"pnorm",2,3)$s
+   s_mle[i] <- ks.test(data,"pnorm",mle$e[1],mle$e[2])$s
+ }
```

```
> quantile(s_true,c(0.9,0.95))
       90%        95%
0.08581135 0.09278481
> quantile(s_mle,c(0.9,0.95))
       90%        95%
0.05952769 0.06545919
> mean(s_90_modified)
[1] 0.0581893
> mean(s_95_modified)
[1] 0.06297491
Summary:
The 90 and 95 percentile of D statistic for modified ks-test
is slightly smaller than ks-test with mle as parameter,
while far smaller than with true parameter.
It is reasonable since the modified ks-test is also based on mle.

Teacher' comments: this is an excellent explaination.
     90%        95%      percentile
 0.08581135 0.09278481   ks.test using use true parameters
 0.05952769 0.06545919   This is just a modified ks.test
 0.0581893  0.06297491   mean of 100 replications of modified ks.test
Thus it is not surprised that the last two rows are very close.
It is smart to take the mean of the 100 preplications of modified ks.test,
and it would be better to compute the SE as well.
Then it clear that the difference between the 2nd and 3rd lines are within 2SE.
Drawback: I asked to do n=200, rather than n=100.

The purpose of this exercise is to let you understand the modified ks.test.
If we use the second approach in the case replacing MLE for the true value,
the ks statstic is much lower than the one ks statistic using the true value,
thus it is likely to believe that  the unknown F is normal distribution.
The modified ks.test adjust the quantile to the proper lower  value, as
                        s=||F(t)-G(t)||.
Thus the probability of rejecting normal assumption if it is indeed from
normal distribution is around 5%.
```

One can replace the command

```
s_mle[i] <- ks.test(data,"pnorm",mle$e[1],mle$e[2])$s
....
s_95_modified[i] <- quantile(s,0.95)
```

by

```
s_mle[i] <- ks.test(data,"pnorm",mle$e[1],mle$e[2])$p
....
s_95_modified[i] <- quantile(s,0.05)
```

Why not quantile(s,0.95) ??

Reject H_o if $s = ||F(t) - G(t)||$ is large, or P-value is small.

For HW10, to test whether data are from normal: is any of the following three correct ?

```
> mle <- fitdistr(X, "normal")     #X is the data
> u <- rep(0, 1000)
> n = 1000
> for (i in 1:n) {
+ normal_generated <- rnorm(200, mle$e[1], mle$e[2])
(1) + mle_mod <- fitdistr(normal_generated, "normal")
+ a <- ks.test(normal_generated, "pnorm", mle_mod$e[1], mle_mod$e[2])
+ u[i] = as.numeric(a$p < 0.05) }
> mean(u)  # =estimate of p-value using I(ks.test with p-value=0.05)
> 0
```

Answer. We see based on the p value of 0 that we fail to reject, and conclude that the estimates are normal as expected from the QQ plot earlier.

Another two approaches:

```
> for (i in 1:n) {
+ normal_generated <- rnorm(200, mle$e[1], mle$e[2])
(2) + mle_mod <- fitdistr(normal_generated, "normal")
+ u[i] <- ks.test(normal_generated, "pnorm", mle_mod$e[1], mle_mod$e[2])$p}
> t= quantile(u,0.05)
> ks.test(X, "pnorm", mle$e[1], mle$e[2])$p     # reject Ho if p<t
Or the (3)rd way:
+ u[i] <- ks.test(normal_generated, "pnorm", mle_mod$e[1], mle_mod$e[2])$s}
> t= quantile(u,0.95)
> ks.test(X, "pnorm", mle$e[1], mle$e[2])$s     # reject Ho if s>t
```

Teacher: Actually, mean(u)=0 means always rejecting H_o using ks.test with MLE and $\alpha = 0.05$.