Potència en front de Mixtures

- Resum:

Realització del test de bondat d'ajust de Pearson i el test de Jarque Bera per l'estudi de si una mixtura segueix una distribució normal o no. Aplicarem cada test 10000 cops i calcularem la potència del test en cada cas. Finalment, farem l'estudi corresponent dels resultats veient que són força diferents.

- Abstract:

Realization of the test of Pearson and Jarque-Bera test for the survey of if a Mixture tracks a normal distribution or no. Each test will be applied 10000 times and will compound the potency of the test at each case. Finally, we will do the corresponding survey of the results viewing that they are quite different.

Clara Albert Vilá

Novembre 2019

1. Introducció

En aquest treball calcularem les corbes de potència dels tests de normalitat de Pearson i de Jarque-Bera en front a mixtures de distribucions normals.

Realitzarem tots els càlculs amb Rstudio sota els paràmetres fixos $\pi = 0.5\ i\ \sigma^2 = 1$. A més a més, estudiarem els resultats amb distància entre mitjanes diferents, que variaràn de 1.5 a 3 a intervals de 0.25 i amb mostres de mida 100, 250, 500.

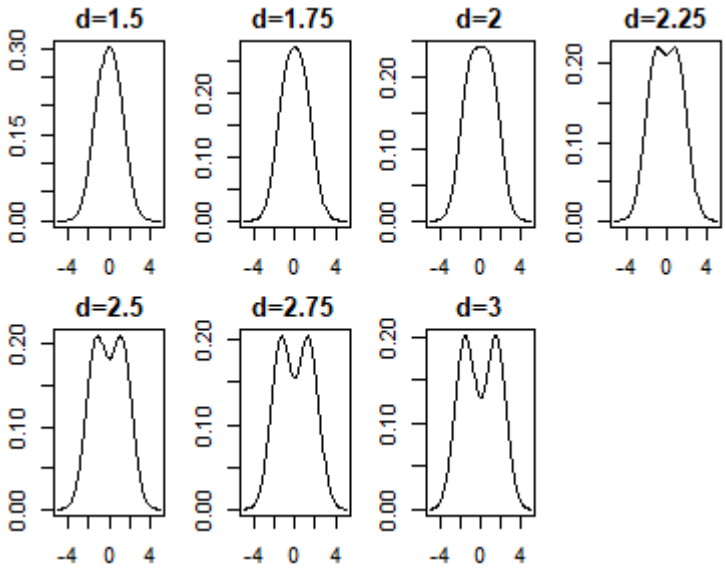
L'objectiu prinical és realitzar un contrast d'hipòtesis per estudiar si la mixtura s'aproxima a una distribució normal (H_0) o, si en canvi, és una mixtura (H_1). S'utilitzarà el nivell de significació estàndard (0.05). Per fer-ho, realitzarem dos tests diferents: el test de bondat d'ajust de Pearson i el test de Jarque Bera. A més a més, una vegada calculat el nombre de vegades que acceptem H_1 calcularem amb quina potència ho fem.

2. Mixtura de normals

La densitat de probabilitat de la mixtura és:

$$p_M(x) = \pi f_N(x; \mu_1, \sigma_1^2) + (1 - \pi) f_N(x; \mu_2, \sigma_2^2)$$

Aquesta densitat l'he utilitzat per calcular diferents mixtures variant la distància entre les seves mitjanes, prenent valors de x entre -5 i 5. Les representacions gràfiques de les funcions de distribució són els següents (per a diferents valors de d):



Podem observar que com més gran és la distància entre mitjanes, més s'accentua que la distribució té dos pics. Per a distàncies petites, la mitxura actua com una normal (forma de campana única) i a partir de distància 2.25 ja s'observen les dues campanes.

- Simulador de mixtures:

Per a que totes les mixtures s'aproximin al comportament d'una distribució normal, he preparat un simulador de mixtures de normals, on a continuació es detalla el codi utilitzat.

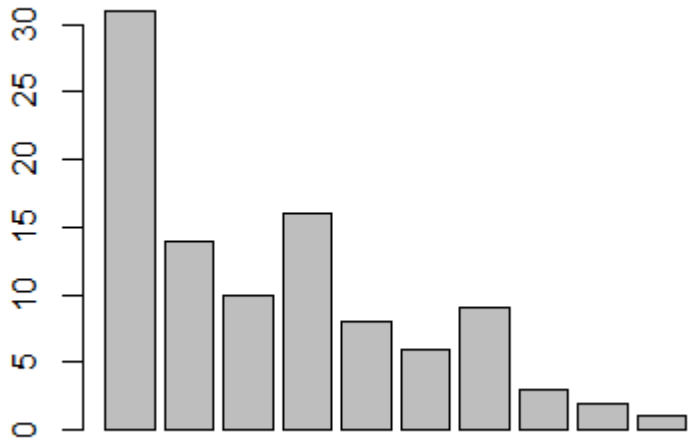
```
sim<-function(p,m1,m2,sg1,sg2){
  ifelse(runif(1)<p,rnorm(1,m1,sg1),rnorm(1,m2,sg2))
}

set.seed(1000)
simix<-function(n,m1,m2,sg1=1,sg2=1,p=0.5){
  mos=numeric(n);
  for(k in 1:n){
  mos[k]<-sim(p,m1,m2,sg1,sg2)
  };mos
}
```

En la primera funció creada, s'utilitza rnorm per generar números aleatoris que segueixin una distribució normal, per tant ens estem asegurant que la mitxura també seguirà aquest tipus de distribució. La segona funció és un replicador de "k" mixtures, que en aquest cas serà 10000.

3. Test de bondat d'ajust de Pearson

El primer test que utilitzarem serà el de bondat d'ajust de Pearson. S'agruparan les dades en intervals equiprobables a partir dels valors de la funció "hist$counts" i a continuació s'aplicarà el test "chisq.test" per a les dades agrupades i s'obtindrà el p-valor.

Per a n=100,mu=-0.75 i sg=1, tenim la següent representació gràfica amb intervals equiprobables.



L'exemple anterior s'aplica només a una repetició del test, però necessitem aplicar el test 10000 vegades, per tant crearem un bucle que faci, per a cada mida de mostra i distància, totes les vegades que volem que es realitzi el test. El bucle és el següent tros de codi, on n=100, mu1=-0.75 i mu2=0.75.

```
for(k in 1:n.rep){
  sample=simix(n1,mu1[1],mu2[1])
  his=hist(sample,breaks=qq,plot = F)
  obs=his$counts
  xi=chisq.test(obs)
  pvs[k]=xi$p.value
}

n.sig=ifelse (pvs<0.05,1,0)

pot=sum(n.sig)/n.rep
```

Una vegada creat el bucle, realitzarem el contrast d'hipòtesi. Comptarem les vegades que es rebutja la hipòtesi nul·la (p-valor<0.05).
A continuació es detalla els resultats en forma de taula obtinguts al realizar el test 10000 diferenciant la mida mostral (N) i la distància entre mitjanes. També s'especifica la potència del test per a cada resultat.

- Resultats test de Pearson:

Distància entre mitjanes	N=100		N=250		N=500	
		Potència		Potència		Potència
1.5	5159	0.5159	9316	0.9316	9996	0.9996
1.75	7854	0.7854	9969	0.9969	10000	1
2	9494	0.9494	10000	1	10000	1
2.25	9957	0.9957	10000	1	10000	1
2.5	9997	0.9997	10000	1	10000	1
2.75	10000	1	10000	1	10000	1
3	10000	1	10000	1	10000	1

Dels resultats obtinguts, s'observa que per a N molt grans i distàncies a partir de 1.75, la potència del test és 1. A més a més, com més és redueix la N, les potències també ho fan, tot i que per a les distàncies més grans (2.75 i 3) la potència continua sent 1.
Cal destacar que totes les potències superen el 0.8 (estadísiticament significatiu) exceptuant quan la distància entre mitjanes és 1.5 i 1.75 per a la mida mostral més petita.

4. Test de Jarque Bera

Per a l'aplicació d'aquest segon test utilitzarem la funció que ens proporciona R per simular el test a unes dades proporcionades. El següent tros de codi representa el test de Jarque-Bera simulat 10000 cops.

```
for(k in 1:n.rep){
  sample=simix(n1,mu1[1],mu2[1])
  taula=jarque.test(sample)
  pvss[k]=taula$p.value
}

n.sig=ifelse (pvs<0.05,1,0)

pot=sum(n.sig)/n.rep
```

Apliquem el test per als diferents valors de mida mostral i mitjanes, on els resultats es recullen a la següent taula:

- Resultats del test de Jarque Bera:

Distància entre mitjanes	N=100		N=250		N=500	
		Potència		Potència		Potència
1.5	119	0.0119	191	0.0191	682	0.0682
1.75	64	0.0064	388	0.0388	2071	0.2071
2	44	0.0044	1146	0.1146	5108	0.5105
2.25	56	0.0056	2953	0.2953	8412	0.8412
2.5	137	0.0137	5807	0.5807	9777	0.997
2.75	407	0.0407	8331	0.8331	9987	0.9987
3	1159	0.1159	9640	0.9640	10000	1

A partir del resulats que hem obtingut veiem com hi ha molt poques potències que arribin al 0.8 (només en 6 ocasions). Com a l'anterior test, com més gran és la distància entre mitjanes, més alta és la potencia, on aquesta última també incrementa quan incrementem la mida mostral.

5. Conclusió

A partir del resultats dels dos tests, afirmem que hi ha grans difèrencies d'acceptació d'H_1 amb un nivell de significació de 0.05. Aquestes diferències són notables quan la mida mostral és 100 per a qualsevol distància, quan N=250 amb distàncies més petites que 2.75 i quan la mida de la mostra és la més gran que hem utilitzat amb 1.5, 1.75 i 2 de distància.

En els resultats del test de Pearson, hem observat com actua millor per a mides petites, ja que la seva potència és "rica" i no cal augmentar la N molt per trobar un bona β. Per un altre costat, el test de Jarque Bera, requereix N molt més gran per a tenir un test amb una bona potència.

Els resultats que coicideixen en els dos tests són que per a distàncies més grans entre mitjanes, la potència del test sempre creix.

6. Webgrafia

https://studylib.es/doc/8463669/1.-generaci%C3%B3n-de-mixturas

https://cran.r-project.org/doc/contrib/Saez-Castillo-RRCmdrv21.pdf

http://www.est.uc3m.es/esp/nueva_docencia/getafe/economia/estadistica_ii/documentacion_transp_archivos/tema2esp.pdf