Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona
MÈTODES NO PARAMÈTRICS I DE REMOSTREIG. Grau en Estadística, curs 2017-18
Prova de síntesi, 12 de juny de 2016.
Respon als mateixos fulls de l'examen. Cada pregunta val 1 punt.

New	Old
46	44
32	31
42	25
45	22
37	30
44	30
38	32
47	19
49	40
41	

PROBLEMA 1. Una companyia cosmètica realitzà un petit assaig sobre una nova crema pel tractament de les taques de la pell. Mesurà l'efectivitat de la nova crema (New) comparada amb la crema dominant en el mercat (Old). Una mostra de 20 persones es va aleatoritzar en dos grups de 10, de manera que els integrants de cada grup van rebre durant l'assaig una, i solament una, de les cremes. En el grup Old al final va fallar una persona, però per causes totalment alienes a l'experiment. La taula adjunta mostra el percentatge de taques eliminat per cada tractament, en cadascuna de les 19 persones de les quals finalment es van tenir dades.

Utilitzant quan calgui els llistats finals, i considerant sempre un nivell de significació de 0,05 o un nivell de confiança de 0,95, respon raonadament les qüestions següents:

1) Es volia determinar si per terme mitjà la nova crema reduïa més les taques que la crema dominant al mercat. Atès que la normalitat d'aquestes dades és dubtosa, es va decidir realitzar una prova de permutacions. Indica les hipòtesis implicades en aquesta prova, l'estadístic de test i el seu valor, el p-valor i la conclusió final.

Es tracta d'un cas clar de comparació de dues mostres independents. Si indiquem com a F_{New} la distribució de la variable aleatòria Y = "reducció de taques" quan s'aplica la nova crema i com F_{Old} la distribució sota la crema dominant al mercat, les hipòtesis implicades son

1. Hipòtesis:

$H_0: F_{New} = F_{Old}$

$H_1: \mu_{New} > \mu_{Old}$

on μ aquí significa la mitjana poblacional associada a aquestes distribucions.

2. Valor estadístic mostral

Segons els llistats, l'estadístic en el que s'ha basat la prova de permutacions és la **diferència de mitjanes mostrals, $\bar{Y}_{New} - \bar{Y}_{Old}$, que sobre la mostra original té el valor 11,76667.**

3. Nº permutacions:

Tractant-se de la comparació de dues mostres independents, cal permutar lliurement tot el vector de 19 observacions o bé, equivalentment, permutar les etiquetes "New" i "Old". De permutacions possibles n'hi ha un nombre enorme, 19!, però permutacions

COGNOMS, NOM:	**SIGNATURA:**

equiprobables que realment puguin proporcionar un valor de diferència de mitjanes diferent n'hi ha moltes menys, **19! / (10! 9!) = 92378.** S'ha fet la llista de totes les permutacions amb repetició possibles i

4. Càlcul del pvalor "exacte": Atès que l'alternativa era unilateral, $\mu_{New} > \mu_{Old}$)s'ha comptat el nombre de permutacions amb repetició que donaven **un valor de diferència de mitjanes superior o igual a l'observat sobre la mostra original.** En total són **104**, de manera que el **p-valor exacte és 104 / 92378 = 0,00113**,

5. Conclusions: El pvalor és inferior al nivell de significació utilitzat, de manera que rebutjarem H_0. Deduïm que hi ha evidències a favor de pensar que la nova crema redueix, per terme mitjà, més les taques que la dominant al mercat.

2) Com a alternativa a la prova anterior, també es va considerar la possibilitat de fer una prova de rangs per intentar demostrar que la reducció de taques mediana era més gran en "New" que en "Old". Indica quina prova triaries i les seves condicions de validesa, les hipòtesis contrastades en aquesta prova, el valor de l'estadístic de test i la conclusió final. En la resposta a aquesta pregunta pots ignorar els empats.

1. **Identificar el test i escriure el contrast a realitzar;**

   En aquesta situació de dues mostres independents, la prova de rangs a considerar és la de U-Mann-Whitney

   Les hipòtesis considerades serien ara

   H_0 : $\mu_{New} = \mu_{Old}$ en front a

   H_1 : $\mu_{New} > \mu_{Old}$, on μ aquí significa la mediana.

2. **Condicions de validesa**

   Aquesta prova és vàlida per comparar dues mostres independents d'una variable aleatòria contínua amb qualsevol distribució, però que sigui la mateixa sota cadascuna de les situacions comparades (incloent els paràmetres de dispersió, que han de ser iguals) llevat possiblement de les respectives medianes. (Un percentatge, com el de la resposta observada aquí, en realitat és una variable discreta, però s'acostuma a considerar vàlid utilitzar el model continu si està calculat sobre un total prou gran.)

3. **Llistat R**

4. No cal realitzar manualment cap càlcul, segons els llistats s'ha utilitzat correctament la funció wilcox.test (on paired=FALSE, opció per defecte):

```
wilcox.test(reducTaques ~ crema, alternative = "greater",
            conf.int = TRUE, correct = FALSE)
```

que ha proporcionat el valor de l'estadístic

5. **Estadístic de test, regió crítica** $W = \sum R_{Old} - \left(n_{Old}\left(n_{Old}+1\right)\right)/2 = 81$ que sota la Hipòtesis nul·la segueix una distribució exacte U-Mann Whitney o bé podem realitzar una aproximació asintòtica a la Normalitat si hi ha empats (com és el cas i és el que utilitza l'R en aquesta situació) associat a **un p-valor unilateral** de 0,001624

   R_old indica els rangs que és la posició que li correspon al valor de la mostra després d'endreçar tots valors old i new. Nold és el tamany mostral de old que en aquest cas és de 9.

6. **Conclusió**

   Rebutjarem si el pvalor < 0.05 que en aquest cas ens porta a rebutjar H_0, de manera que la conclusió és la mateixa que a la pregunta anterior, però ara amb la mediana.

3) Estima l'efecte de substituir una crema per l'altra (és a dir, la diferència de medianes) associat a la prova anterior, en forma d'estimació puntual i també com a interval de confiança **bilateral**.

1. **Calcular una estimació per a la mediana de les diferències que és en realitat la millor estimació.**

   L'estimació de l'efecte ja la proporcionen els llistats basats en wilcox.test. Com sabem correspon a la mediana de totes les possibles diferències entre valors New i Old i segons els llistats és 12,99998. També es podria obtenir de forma molt directa a partir dels $10 \times 9 = 90$ valors de diferència ja ordenats que es proporcionen als llistats. Normalment s'agafaria com a mediana mostral la mitjana dels valors (del vector de diferències ordenades) que es troben a les posicions 45 i 46, és a dir el valor 13. Ambdues possibilitats són perfectament vàlides, només responen a formes diferents d'estimar la mediana.

2. **Llistats de R**
   a. basats en wilcox.test per trobar l'estimació anterior
   b. `dij = outer(reducTaques[1:n1], reducTaques[-(1:n1)], "-")`, per tal de calcular totes les diferències possibles
   c. `sort(dij)` endreçar el vector diferència

   Noteu; Els intervals de confiança que proporcionaven els llistats no són adequats ja que són unilaterals i aquí es demana un interval bilateral

3. **Construir el Interval de Confiança**
   a. Obtenir totes les diferències i endreçar-les
   b. Trobar les posicions λ i ν dins del vector de les 90 diferències ordenades
      Recordem que aquests posicions s'obtindrien com: $\lambda = u_{0,05}(10,9) + 1 = 20 + 1 = 21$ i $\nu = 10 \times 9 - 21 + 1 = 70$, on $u_{0,05}(10,9)$ correspon al valor crític bilateral de l'estadístic U de la prova de Mann-Whitney-Wilcoxon, per un nivell de significació 0,05 i per mides mostrals 10 i 9.
      La forma d'obtenir l'interval anterior és preferible a la basada en les fórmules per mostres grans, ja que aquí les mostres no ho són gaire (de grans). Amb l'aproximació asimptòtica, si es calcula bé, al final s'obté el mateix interval. Però en alguns casos, degut a errors de vegades numèrics i de vegades més de concepte, utilitzant l'enfoc asimptòtic el resultat ha estat diferent del correcte.

4. **Donar un Interval de Confiança vàlid**

Cal trobar les posicions λ i ν dins del vector de les 90 diferències ordenades:. Dins el vector de les 90 diferències ordenades, a la posició 21 hi ha el valor 5 i a la posició 70 hi ha el valor 19, de manera que l'interval de confiança demanat és [5, 19].

5. **Interpretació basada en l'Interval**

Si haguéssim de concloure amb aquest interval bilateral (que no seria el correcte pel contrast efectuat anteriorment) veuríem que no inclou el zero per tant ens dóna a pensar que existeixen diferències entre tractaments

En realitat a l'experiment anterior hi van participar 30 subjectes que es van aleatoritzar en 3 grups. A més dels tractaments que ja coneixem, 10 subjectes més (grup "Control") van rebre una crema hidratant, que en principi no tenia cap relació amb la reducció de taques. També per causes totalment atzaroses, alienes a l'experiment, finalment no es van obtenir dades de 2 subjectes d'aquest grup. Les dades completes van ser les indicades a l'esquerra. A partir d'ara considerarem aquestes dades completes.

New	Old	Control
46	44	26
32	31	49
42	25	33
45	22	19
37	30	31
44	30	38
38	32	44
47	19	50
49	40	
41		

4) Realitza una prova de permutacions per a intentar demostrar que hi ha alguna diferència entre les mitjanes dels tres grups. Indica les hipòtesis contrastades en aquesta prova, el valor de l'estadístic de test i la conclusió final.

Ara tindríem una hipòtesi nul·la de total igualtat distribucional

1. Hipòtesis:

H_0 : $F_{New} = F_{Old} = F_{Control}$

H_1 : $\mu_i \neq \mu_j$, per $i \neq j$, i,j = New, Old, Control,

on μ ara significa la mitjana poblacional.

2. Valor de l'estadístic

L'estadístic de test utilitzat **ha estat el mateix que a l'ANOVA d'un factor,** F, que sobre la mostra original ha donat el valor $F = 4{,}890214$.

3. Nº permutacions:

El mètode de permutació adient és el mateix que a la pregunta 1), és a dir, permutar lliurement el vector de les 27 observacions. Fins i tot considerant les permutacions amb repetició, el nombre de possibilitats és enorme, $27! / (10!\ 9!\ 8!) > 2 \times 10^{11}$ **de manera que sembla molt més realitzable un enfoc de prova de permutacions de Monte Carlo.**

4. Càlcul del pvalor "mètode montecarlo":

Els llistats corresponen a 19999 permutacions aleatòries. De les 19999 permutacions generades, **334 han proporcionat un valor d'F igual o superior a 4,890214.** Segons el mètode de Dwass, el **p-valor estimat és** $(334 + 1) / (19999 + 1) = 0{,}01675$

5. Conclusions:

El pvalor és inferior al nivell de significació utilitzat, de manera que rebutjarem H_0. Deduïm que hi ha evidències a favor de pensar que hi ha alguna diferència de mitjanes.

LLISTATS DEL PROBLEMA 1

```
> reducTaques = c(46, 32, 42, 45, 37, 44, 38, 47, 49, 41,
+                 44, 31, 25, 22, 30, 30, 32, 19, 40)
> crema = factor(rep(c("New", "Old"), c(10,9)), levels = c("New", "Old"))
> n1 = 10
> n2 = 9
> n = n1 + n2
>
> # Aquesta funció calcula la diferència de mitjanes mostrals entre
> # dos grups de dades. Cal pasar-li els índexs dels elements del
> # primer grup i el vector de totes les dades juntes:
> difMitjanes = function(indexsPrimerGrup, totesLesDades) {
+   mean(totesLesDades[indexsPrimerGrup]) - mean(totesLesDades[-indexsPrimerGrup])
+ }
>
> # A les dades reals, els primers n1 valors són del primer grup (New):
> dif.observada = difMitjanes(1:n1, reducTaques)
> dif.observada
[1] 11.76667
>
# Valor de n! / (n1! * n2!):
> choose(n, n1)
[1] 92378
>
> # Diferència de mitjanes mostrals per cadascuna de les 92378 permutacions
> # (amb repetició) posibles:
> dif.perms = combn(n, n1, difMitjanes, totesLesDades = reducTaques)
> # Diferència de mitjanes per les 10 primeres permutacions:
> dif.perms[1:10]
 [1] 11.766667 12.400000  9.655556  8.388889  7.755556  9.444444  9.444444  9.866667
 [9]  7.122222 11.555556
>
> sum(dif.perms >= dif.observada)
[1] 104

> wilcox.test(reducTaques ~ crema, alternative = "greater",
+             conf.int = TRUE, correct = FALSE)

        Wilcoxon rank sum test

data:  reducTaques by crema
W = 81, p-value = 0.001624
alternative hypothesis: true location shift is greater than 0
95 percent confidence interval:
 6.000026      Inf
sample estimates:
difference in location
              12.99998

Warning messages:
1: In wilcox.test.default(x = c(46, 32, 42, 45, 37, 44, 38, 47, 49,  :
  cannot compute exact p-value with ties
2: In wilcox.test.default(x = c(46, 32, 42, 45, 37, 44, 38, 47, 49,  :
  cannot compute exact confidence intervals with ties
> wilcox.test(reducTaques ~ crema, alternative = "less",
+             conf.int = TRUE, correct = FALSE)

        Wilcoxon rank sum test

data:  reducTaques by crema
W = 81, p-value = 0.9984
alternative hypothesis: true location shift is less than 0
95 percent confidence interval:
     -Inf 17.00002
sample estimates:
difference in location
              12.99998

Warning messages:
1: In wilcox.test.default(x = c(46, 32, 42, 45, 37, 44, 38, 47, 49,  :
  cannot compute exact p-value with ties
2: In wilcox.test.default(x = c(46, 32, 42, 45, 37, 44, 38, 47, 49,  :
  cannot compute exact confidence intervals with ties
> ks.test(reducTaques[1:n1], reducTaques[-(1:n1)], alternative = "greater")
```

```
        Two-sample Kolmogorov-Smirnov test

data:  reducTaques[1:n1] and reducTaques[-(1:n1)]
D^+ = 1.1102e-16, p-value = 1
alternative hypothesis: the CDF of x lies above that of y

Warning message:
In ks.test(reducTaques[1:n1], reducTaques[-(1:n1)], alternative = "greater") :
  cannot compute exact p-value with ties
> ks.test(reducTaques[1:n1], reducTaques[-(1:n1)], alternative = "less")

        Two-sample Kolmogorov-Smirnov test

data:  reducTaques[1:n1] and reducTaques[-(1:n1)]
D^- = 0.67778, p-value = 0.01288
alternative hypothesis: the CDF of x lies below that of y

Warning message:
In ks.test(reducTaques[1:n1], reducTaques[-(1:n1)], alternative = "less") :
  cannot compute exact p-value with ties


> dij = outer(reducTaques[1:n1], reducTaques[-(1:n1)], "-")
> sort(dij)
 [1] -12  -8  -7  -6  -3  -3  -2  -2   0   0   1   1   1   2   2   2   2   3   4   5   5   5
[23]   6   6   6   7   7   7   7   7   8   8   9   9  10  10  10  11  11  11  12  12  12  12
[45]  13  13  13  13  14  14  14  14  15  15  15  15  15  16  16  16  16  16  17  17  17  17
[67]  18  18  19  19  19  19  19  20  20  21  22  22  22  23  23  24  24  25  25  26  27  27
[89]  28  30


> reducTaques = c(46, 32, 42, 45, 37, 44, 38, 47, 49, 41,
+                 44, 31, 25, 22, 30, 30, 32, 19, 40,
+                 26, 49, 33, 19, 31, 38, 44, 50)
> crema = factor(rep(c("New", "Old", "Control"), c(10,9,8)),
                 levels = c("New", "Old", "Control"))

> oneway.test(reducTaques ~ crema, var.equal = TRUE)

        One-way analysis of means

data:  reducTaques and crema
F = 4.8902, num df = 2, denom df = 24, p-value = 0.01654

> kruskal.test(reducTaques ~ crema)

        Kruskal-Wallis rank sum test

data:  reducTaques by crema
Kruskal-Wallis chi-squared = 7.9349, df = 2, p-value = 0.01892


> f.obs = oneway.test(reducTaques ~ crema, var.equal = TRUE)$statistic
> f.obs
       F
4.890214
>
> nperm = 19999
> set.seed(1277)
> f.perms = replicate(nperm,
      oneway.test(sample(reducTaques) ~ crema, var.equal = TRUE)$statistic)
> sum(f.perms >= f.obs)
[1] 334
```

PROBLEMA 2. Les següents dades corresponen al temps, en segons, que 9 estudiants van ser capaços d'aguantar la respiració, és a dir són temps d'apnea. Per cada estudiant, la dada "Normal" correspon al temps d'apnea en condicions normals (més exactament, havent fet un minut de relaxació just després d'una classe) mentre que la dada "Hipervent" correspon al mateix temps però després de fer uns exercicis d'hiperventilació.

Subjecte	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Normal	56	56	65	65	50	25	87	44	35
Hipervent	87	91	85	91	75	28	122	66	58

Utilitzant quan calgui els llistats finals, i considerant sempre un nivell de significació de 0,05 o un nivell de confiança de 0,95, respon raonadament les qüestions següents:

1) Estima el coeficient de correlació τ de Kendall entre "Normal" i "Hipervent" i realitza un test per determinar si es pot afirmar que és diferent de zero. Pots ignorar els empats.

1. **Construcció del contrast**
   H_0: (X,Y) són independents i H_1: $\tau \neq 0$
2. **Càlcul del coeficient de correlació; fixa't que en aquest cas coincideix amb l'estadístic de test**

Segons els llistats hi ha 31 concordances i 2 discordances entre parelles de valors de Normal (X) i Hipervent (Y), aquestes es calculen fent totes les diferències possibles entre els parells dels valors Normal i realitzant les diferències sobre les mateixes posicions;

Nombre concordants = nc = #{(Xi–Xj)*(Yi–Yj)>0} (per i diferent de j)

Nombre discordants = nd = #{(Xi–Xj)*(Yi–Yj)<0} (per i diferent de j)

sobre $9(9-1)/2 = 36$ emparellaments possibles. Ignorant els empats, es podria estimar τ com $(31-2)/36 = 0{,}8056$.

3. **Cal comparar el valor de l'estadístic de test amb el valor de les taules de la tau de kendall** i aquest valor és 0,556 ja que és un test bilateral
4. **Definir la regió crítica donades les taules**
   Donades les hipòtesis, atès que el valor crític bilateral per l'estimació de τ que permetria rebutjar H_0 és 0,556, i $|0{,}8056| > 0{,}556$, podem rebutjar H_0. I per tant deduïm que hi ha evidències per considerar que no hi ha independència i per tant podem considerar que el valor tau de kendall és diferent de zero

2) Realitza una prova de permutacions per determinar si el coeficient de correlació de Pearson, ρ, entre "Normal" i "Hipervent" és positiu. Indica clarament les hipòtesis, l'estadístic de test, el p-valor i la conclusió final. Tingues en compte que la correlació mostral *r* entre aquestes variables és *r* = 0,9655.

1. Contrarst d'hipòtesis:

H_0: $\rho(X,Y)= 0$

H_1: $\rho(X,Y) > 0$.

2. Estadístic de test,

la correlació mostral **r, observada sobre les dades reals és 0,9655.**

3. Nº permutacions:

En la present situació, que la hipòtesi nul·la sigui certa implica permutar un dels vectors de dades, deixant fix l'altre. El nombre de permutacions possibles és **9! = 362888,** gran però avui en dia computacionalment assequible.

4. Càlcul pvalor "exacte"

S'han enumerat totes aquestes permutacions i sobre cadascuna s'ha calculat l'estadístic r. Hi ha **24 permutacions tals que la correlació mostral és superior o igual a l'observat 0,9655.** Per tant, el **p-valor exacte és 24 / 362888 = 0,0000661,**

5. Conclusions:

Es pot rebutjar H_0, aquestes dades sustenten una suposició de correlació de Pearson positiva entre ambdues variables.

3) Calcula els intervals de confiança bootstrap-t i bootstrap-t simetritzat (no paramètrics) de ρ.

1. Expressió Interval de confiança bootstrap-t (oc cues equiprobables)

L'interval de confiança bootstrap-t demanat es defineix com:

$$\left[r - t^*_{0,975} \hat{SE}_r,\ \ r - t^*_{0,025} \hat{SE}_r \right]$$

2. Estimacions mostrals necessàries

r= 0.9655, i $\hat{SE}_r = \left(1 - r^2\right) / \sqrt{n-3}$, aquí $\hat{SE}_r = \left(1 - 0{,}9655^2\right) / \sqrt{9-3} = 0{,}0277$, és una estimació de l'error estàndard de r.

3. Generació bootstrap: Obtenció del quartils adients

Per obtenir els t^* cal haver obtingut B ("**nboot" als llistats**) rèpliques bootstrap de t i estimar-ne els corresponents quantils, per tant els quantils adients són els obtinguts a partir de `quantile(t.boots, probs = c(0.025, 0.975))`. Finalment l'interval demanat és:

4. Resultat:

$$\left[0{,}9655 - 5{,}3276 \times 0{,}0277,\ \ 0{,}9655 - \left(-0{,}9430\right) \times 0{,}0277\right] = \left[0{,}8179,\ \ 0{,}9916\right].$$

1. Expressió Interval de confiança bootstrap-t simetritzat

$a\ r \pm t^*_{[0{,}95]} \hat{S}E_r$ on $t^*_{[0{,}95]}$

2. Estimacions mostrals necessàries

r= 0.9655, i $\hat{S}E_r = \left(1 - r^2\right) / \sqrt{n-3}$, aquí $\hat{S}E_r = \left(1 - 0{,}9655^2\right) / \sqrt{9-3} = 0{,}0277$, és una estimació de l'error estàndard de r.

3. Generació bootstrap: Obtenció del quartils adients

L'interval de confiança bootstrap-t simetritzat correspon vol dir **el quantil 0,95 de $|t|$. Per tant aquest valor l'obtenim de** `quantile(abs(t.boots), probs = 0.95)` i

4. Resultat:

Finalment l'interval demanat és $0{,}9655 \pm 3{,}1958 \times 0{,}0277 = \left[0{,}8770,\ 1\right]$, substituint per 1 el valor de l'extrem superior que realment s'obtindria, 1,054, i que no té sentit com a valor de correlació.

Calcula l'interval de confiança percentil no paramètric de ρ.

Directament de `quantile(r.boots, probs = c(0.025, 0.975))` tenim que l'interval **demanat és [0,8759, 0,9935].**

4) En un segon experiment es va voler estudiar la possible relació entre el BMI (Body Mass Index que resulta de dividir el pes en kg per l'alçada al quadrat en m) i temps d'apnea en condicions normals (més exactament, havent fet un minut de relaxació just després d'una classe) en una mostra de 40 alumnes. Les dades i algunes comandes de R que us poden ser útils es troben al final del llistat. Indiqueu quin dels mètodes que hem vist a classe (Rangs i regressió no paramètrica/Robusta) seria més adient en aquesta situació i per què. Expliqueu en que es basa el mètode que heu triat (contrast hipòtesis, estimació,...)

1. **Triar la proba adequada**

   Ja que es vol estudiar si existeix i com és la possible relació, ens decantaríem per una regressió no paramètrica d'ajustos locals lineals ja que en el gràfic es veu que no hi ha monotonicitat entre les dues variables (apnea i bmi), hi ha una primera part que la funció creix i una última part en que la funció decreix, no queda clar a simple vista veure que passa entremig. Per tant es descarta la correlació de pearson i de kendall

2. **El contrast seria equivalent al que s'efectua en regressió lineal**

   H_0: $y=m(x,\theta)+\varepsilon$

   H_1: $y!=m(x,\theta)+\varepsilon$

   En aquest cas $y=m(x,\theta)+\varepsilon=\beta_0+\beta_1 x+\varepsilon$ i l'equació anirà canviant segons els suport i l'amplada de banda triats

   El mètode triat es basa en ajustar la millor recta possible utilitzant la minimització dels errors al quadrat afegint un pes (w) per cada observació que varia segons el suport i segons l'amplada de banda triada,

   Min $\Sigma[w_i*(y_i-m(x_i,\text{estimacions}(\theta)))^2$, i =1..., 40

3. **El llistat de R triat seria**

```
>np1<-npregbw(apnea~bmi,regtype="ll")

>np1.2<-npreg(np1, regtype="ll")
```

Que permet fer una regressió no paramètrica amb ajustos locals lineals; (regtype="ll")

4. **La interpretació del resultats seria**

Els resultats ens indiquen de la amplada de banda que s'ha triat, en aquest cas 1.475986. En aquest cas, l'assignació dels pesos s'ha realitzat mitjançant una funció kernel gaussiana.

El grau d'ajust de la funció ajustada té un $R^2=0.5027938$ que es calcula de manera equivalent que en cas de la regressió lineal, la seva interpretació és equivalent i ens indica que hi ha un ajust moderadament bo. Cal assenyalar que a

quest resultats és molt diferent de l'obtingut a `cor.test(bmi, apnea, method="pearson")` que té un resultat de -0.3982831, si calculessim R ≈ 0,709

LLISTATS DEL PROBLEMA 2.

```
> normal = c(56, 56, 65, 65, 50, 25, 88, 44, 35)
> hipervent = c(87, 91, 85, 91, 75, 28, 122, 66, 58)

> n = length(normal)
> n
[1] 9

> cor (normal,hipervent)
[1] 0.9655208

> normal - hipervent
[1] -31 -35 -20 -26 -25  -3 -34 -22 -23

> # Taula amb totes les possibles diferències entre
> # normal[i] i normal[j]:
> difs.normal = outer(normal, normal, "-")
> # Descartem les diferències de la diagonal (i == j) i de la meitat
> # triangular superior:
> difs.normal = difs.normal[ltri <- lower.tri(difs.normal)]
> # Totes les possibles diferències entre hipervent[i] i hipervent[j]:
> difs.hipervent = outer(hipervent, hipervent, "-")[ltri]
> #
> # Nombre de concordances:
> concor = sum(sign(difs.normal)*sign(difs.hipervent) > 0)
> concor
[1] 31
> # Nombre de discordances:
> discor = sum(sign(difs.normal)*sign(difs.hipervent) < 0)
> discor
[1] 2

> factorial(n)
[1] 362880
> require(gtools)
> # Llista de totes les permutacions possibles de 1,2,...,n:
> perms = permutations(n,n)
> # 'perms' és una matriu amb n! files,
> # cada fila representa una permutació
> dim(perms)
[1] 362880      9
>
> # Correlacions entre el vector 'normal' i totes les permutacions
> # possibles del vector 'hipervent':
> r.perms = apply(perms, 1,
                  function(iperm) cor(normal, hipervent[iperm]))
> sum(r.perms <= r)
[1] 362868
> sum(abs(r.perms) >= abs(r))
[1] 32
> sum(r.perms >= r)
[1] 24


> # 20000 rèpliques bootstrap no paramètric de les dades
> # i càlcul de la correlació mostral sobre cadascuna (valors r*):
> nboots = 20000
> r.boots = replicate(nboots, {
+   iboots = sample.int(n, replace = TRUE)
+   cor(normal[iboots], hipervent[iboots])
+ })
> # Les primeres 20 correlacions bootstrap:
> r.boots[1:20]
 [1] 0.9535545 0.9795684 0.9583672 0.9948724 0.9876154 0.9817911 0.9992018 0.9654787
 [9] 0.9158267 0.9685632 0.9655725 0.9732719 0.9726834 0.9910551 0.9666103 0.9718700
[17] 0.9606061 0.8804907 0.9849044 0.9806722
>
> # Estimació de l'error estàndard de cadascuna d'aquestes
> # correlacions mostrals r*:
> se.boots = (1 - r.boots^2) / sqrt(n - 3)
> # Els primers 20 valors de l'error estàndard de r*, SEr*:
```

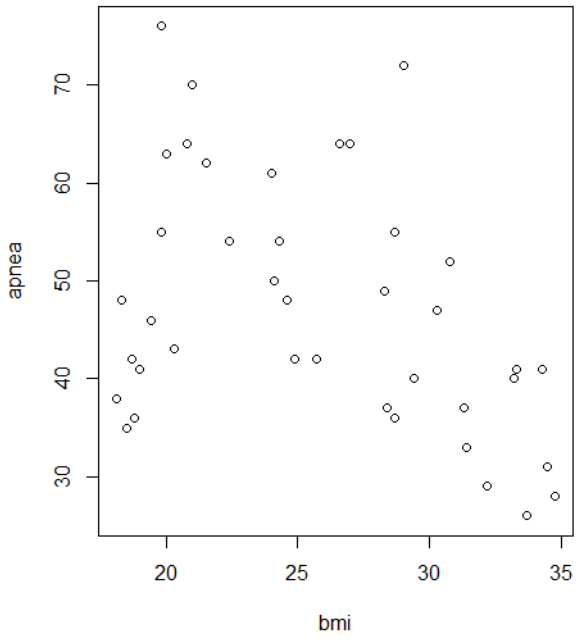
```
> se.boots[1:20]
 [1] 0.0370419471 0.0165118712 0.0332853906 0.0041759614 0.0100493285 0.0147321345
 [7] 0.0006514794 0.0276999968 0.0658346976 0.0252646104 0.0276260582 0.0215317696
[13] 0.0219993028 0.0072708292 0.0268074353 0.0226449911 0.0315314507 0.0917481182
[19] 0.0122324460 0.0156285427
>
> # 20000 valors de l'estadístic estudentitzat t* = (r* - r) / SEr*
> t.boots = (r.boots - r) / se.boots
> # Els primers 20 valors de t*:
> t.boots[1:20]
 [1] -0.323047361  0.850761917 -0.214914993  7.028702681  2.198621805  1.104411389
 [7] 51.699275388 -0.001518544 -0.754830742  0.120420840  0.001872346  0.359984550
[13]  0.325582704  3.511883992  0.040642452  0.280381665 -0.155866541 -0.926776992
[19]  1.584610426  0.969474041
>
> # Diversos quantils d'aquests valors bootstrap:
> quantile(r.boots, probs = c(0.025, 0.975))
     2.5%     97.5%
0.8758946 0.9935383
> quantile(abs(r.boots), probs = c(0.025, 0.975))
     2.5%     97.5%
0.8758946 0.9935517
> quantile(abs(r.boots), probs = 0.95)
      95%
0.9903969
>
> quantile(se.boots, probs = c(0.025, 0.975))
       2.5%       97.5%
0.005248068 0.095043709
> quantile(abs(se.boots), probs = c(0.025, 0.975))
       2.5%       97.5%
0.005248068 0.095043709
> quantile(abs(se.boots), probs = 0.95)
       95%
0.06899512
>
> quantile(t.boots, probs = c(0.025, 0.975))
      2.5%      97.5%
-0.9429992  5.3276456
> quantile(abs(t.boots), probs = c(0.025, 0.975))
      2.5%      97.5%
0.02246474 5.38547212
> quantile(abs(t.boots), probs = 0.95)
     95%
3.195762
```

```
 ##################

bmi <- c(18.3, 30.8, 34.8, 19.8, 20.3, 28.4, 31.4, 30.3, 24.3, 24.0, 29.0, 24.9,
 22.4, 25.7, 33.2, 20.0, 20.8, 18.7, 28.3, 27.0, 33.7, 19.8, 32.2, 28.7,
 18.5, 19.0, 18.1, 24.6, 29.4, 34.5,  21.5, 21.0, 19.4, 18.8, 34.3, 33.3,
 26.6, 28.7, 31.3, 24.1)

 apnea<-c(48, 52, 28, 55, 43, 37,  33, 47, 54, 61, 72, 42,
 54 ,42, 40, 63, 64, 42, 49, 64, 26, 76, 29, 55,
 35, 41, 38, 48, 40, 31, 62, 70, 46, 36, 41, 41,
 64, 36, 37, 50)

plot(bmi, apnea)
```




```
> cor.test(bmi, apnea, method="kendall", exact=FALSE)

        Kendall's rank correlation tau

data:  bmi and apnea
z = -2.3682, p-value = 0.01787
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.2631331

> cor.test(bmi, apnea, method="pearson")

        Pearson's product-moment correlation

data:  bmi and apnea
t = -2.6766, df = 38, p-value = 0.01092
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.63144922 -0.09906457
sample estimates:
       cor
-0.3982831

require(np)

>np1<-npregbw(apnea~bmi,regtype="ll")

>np1.2<-npreg(np1, regtype="ll")

> summary(np1.2)

Regression Data: 40 training points, in 1 variable(s)
                 bmi
Bandwidth(s): 1.475986

Kernel Regression Estimator: Local-Linear
Bandwidth Type: Fixed
Residual standard error: 8.952328
R-squared: 0.5027938

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```