Prova de síntesi. 13 de juny de 2014
MÈTODES NO PARAMÈTRICS I DE REMOSTREIG. Grau en Estadística. Curs 2013-14

estudiant	probabilitats	discretes
1	65	52
2	88	57
3	83	78
4	92	76
5	50	30
6	67	67
7	100	96
8	100	74
9	73	65
10	90	87
11	83	78
12	94	89

Problema 1. Les dades de l'esquerra corresponen a l'avaluació contínua de 12 estudiants d'un mateix curs d'introducció a l'estadística matemàtica. La columna "probabilitats" correspon a una avaluació de la part "càlcul de probabilitats", mentre que la columna "discretes" correspon a una avaluació similar de la part "variables aleatòries i distribucions discretes" feta posteriorment als mateixos estudiants. Són notes sobre 100.

Respon les següents qüestions, utilitzant els llistats del final de l'enunciat de l'examen quan ho creguis convenient. En tot moment considerarem un nivell de significació de 0.05 o un nivell de confiança de 0.95. Quan realitzis una prova d'hipòtesis has d'expressar clarament les hipòtesis nul·la i alternativa.

1) Indica el nom d'una prova d'hipòtesis basada en rangs que sigui adequada per a intentar demostrar que la nota mediana ha variat de l'avaluació de probabilitats a la de discretes. Indica les condicions de validesa de la prova que has triat. Indica a quina prova paramètrica normal (en aquest cas, per les mitjanes) seria comparable.

Es tracta clarament d'una situació de dades aparellades, dues avaluacions diferents però pel mateix estudiant. Per tant la prova de rangs candidata més clara seria la prova de Wilcoxon dels rangs amb signe.

Aquesta prova requereix treballar amb les diferències dins cada subjecte o bloc, la variable diferència ha de tenir distribució contínua i ***simètrica****, encara que no cal suposar-li cap forma distribucional concreta, en particular* ***no cal suposar homogeneïtat entre les dispersions*** *de les notes a ambdues proves, aquí la variable que compta és la variable diferència.*

Seria l'equivalent de rangs a la típica prova t de Student per dades aparellades.

2) Indica justificadament les hipòtesis, el resultat i la conclusió final d'aquesta prova. Què demostra aquest resultat?

Si δ indica la mediana de les diferències, tenim una prova ***bilateral*** *(no es vol demostrar que la nota hagi pujat –o baixat–, només si hi ha diferències sense indicar-ne el sentit) $H_0 : \delta = 0$ vs $H_1 : \delta \neq 0$. Els llistats utilitzant la funció 'wilcox.test' de R no serveixen ja que corresponen al cas no aparellat, de dues mostres independents. Cal*

NOM I COGNOMS: **FIRMA:**

basar-se en els càlculs pas a pas realitzats, per altra banda molt detallats. La suma de rangs positius de les diferències, R^+, val 77, i la de rangs negatius 0, òbviament (no n'hi ha cap), per tant tenim que l'estadístic de Wilcoxon és $W = \min\{R^+, R^-\} = 0$, molt més extrem (menor) que el valor que trobaríem a les taules per la prova bilateral sota un nivell 0.05 ($w_{0.05}(12) = 13$). Es rebutjaria H_0, hi ha hagut variació en la mediana.

3) Indica justificadament el valor de l'estimació puntual i l'interval de confiança per al canvi experimentat en la mediana de les diferències de notes.

La mediana de totes les diferències (6.5) té sentit com a estimació puntual de l'efecte estudiat. Però l'estimació puntual associada al test de Wilcoxon per dades aparellades és la mediana de les semisumes de diferències, és a dir 10.5. Restant aquest valor a les diferències uniformitzaríem al màxim la mostra, en el sentit de maximitzar el p-valor o equilibrar els rangs de diferències negatives i positives.

L'interval de confiança bilateral associat al test anterior correspon a $[D_{(\lambda)}, D_{(\upsilon)}]$ on $D_{()}$ és el vector de diferències ordenat i les posicions υ i λ es determinen com:

$$\upsilon^* = \frac{n+1}{2} + \frac{1}{2} z_\alpha \sqrt{n} = \frac{12+1}{2} + \frac{1}{2} 1.96\sqrt{12} = 9.89$$

$$\upsilon = \left\{ \begin{array}{l} \upsilon^* \text{ si } \upsilon^* \text{ és enter} \\ \left[\upsilon^* + 1\right] \text{ en cas contrari} \end{array} \right\} = 10 \quad \lambda = n - \upsilon + 1 = 12 - 10 + 1 = 3$$

$$z_\alpha \text{ valor t.q. } Pr\left(|Z| \le z_\alpha\right) = 1 - \alpha,\ Z \sim N(0,1)$$

Per tant l'interval de confiança seria $[D_{(3)}, D_{(10)}]=[4, 20]$. No seria gaire precís ja que es basa en un resultat asimptòtic i aquí la mostra és bastant petita.

4) Els resultats anteriors no són exactes (per exemple, el p-valor que s'obtindria no seria del tot correcte), per quina raó? Per a la mateixa situació, seria exacta una prova de permutacions? Per quina raó?

Els resultats anteriors no són exactes a causa de la presència d'empats. En llenguatge no tècnic, substituir les dades originals pels seus rangs permet construir la taula del test, els rangs són sempre 1, 2, …, n, de manera que es poden enumerar totes les possibilitats i tabular la

distribució de l'estadístic de test sota H_0. Quan no hi ha empats tenim una prova exacta (excepte si s'utilitza una aproximació asimptòtica). Quan hi ha empats les configuracions possibles es disparen (caldria una taula per cada configuració possible d'empats), tasca impossible. Per tant no estem fent una prova exacta, en certa manera es podria dir que estem utilitzant la taula equivocada, una taula que només val pel cas sense empats.

Una prova de permutacions basada en enumerar totes les permutacions possibles seria exacta, a base de força bruta computacional, un cop disposem de les dades construïm la taula del test exacta, al marge de si hi ha empats. Un test de permutacions de Monte Carlo, on generem una mostra gran de les permutacions possibles (però no totes) no seria exacte.

5) Realitza un prova de permutacions per intentar demostrar que la nota **mitjana** ha variat entre l'avaluació de probabilitats i la de discretes. Indica clarament i de forma justificada les hipòtesis d'aquest test, el p-valor obtingut i la conclusió final.

Els llistats ens permeten fer un test de permutacions exacte, enumerant totes les permutacions, que no són gaires: $2^{12} = 4096$. Si ara δ designa la mitjana poblacional de les diferències, les hipòtesis plantejades tindrien la mateixa forma que a la pregunta 2 (test bilateral, etc.). Segons els llistats, la mitjana mostral de les diferències sobre la mostra original és a la variable 'm.d'. Aquest valor no es mostra als llistats, però sí que s'indica quantes vegades ha passat que la mitjana sobre les diferències de dades permutades ha estat superior o igual a aquest valor (2 vegades ha passat), ha estat inferior o igual (sempre ha passat, 4096 vegades) o bé ha estat tant o més extrema (en negatiu o en positiu):

```
> sum(abs(m.perm) >= abs(m.d))
[1] 4
```

Aquest és el valor que cal utilitzar per calcular el p-valor ja que el test plantejat és bilateral. El p-valor exacte és 4 / 4096 = 0.00098 (arrodonint a 5 decimals) i per tant rebutgen la hipòtesi nul·la, la conclusió final és que hi ha hagut variació de la nota mitjana.

Fixeu-vos que aquí no tindria sentit fer (4 + 1) / (4096 + 1), no estem estimant el p-valor sinó que l'estem calculant de forma exacta, enumerant totes les possibles permutacions.

Problema 2. Per les mateixes dades del problema anterior, atès que cada parella de notes es refereix al mateix alumne, seria d'esperar que hi hagués un cert grau de dependència entre les variables *X* = 'probabilitats' i *Y* = 'discretes'.

Respon les següents qüestions, utilitzant els llistats del final de l'enunciat quan ho creguis convenient. En tot moment considerarem un nivell de significació de 0.05 o un nivell de confiança de 0.95. Quan realitzis una prova d'hipòtesis has d'expressar clarament les hipòtesis nul·la i alternativa:

1) Ignorant el empats, calcula el coeficient de correlació de Kendall entre *X* i *Y* i determina si és significatiu. Explica el significat del valor obtingut (que no vol dir el mateix que "ser significatiu estadísticament") d'aquest coeficient.

El nombre de concordances entre diferències dels valors de X i de Y és $n_C = 51$ i el de discordances $n_D = 13$. El total de parelles de valors de X o de Y a formar és de $n(n-1)/2 = 66$, per tant el coeficient de correlació de Kendall mostra seria $\hat{\tau} = (51 - 13) / 66 = 0.576$. (Com que hi ha empats, és possible que algú hagi dividit per 51 + 13 = 64, amb resultat 0.594; també ho donem per bo.) En realitat tots aquests càlculs no són del tot correctes i caldria calcular l'anomenat tau-B o tau-C, estimacions alternatives suposadament millors en presència d'empats. Si algú ho fa, perfecte (donaria un valor de l'ordre de 0.59), però dono per bo l'anterior.

Per estudiar-ne la significació es plantejaria el contrast H_0 : "X i Y estocàsticament independents" vs $H_1 : \tau(X,Y) \neq 0$, "el coeficient de correlació de Kendall poblacional no és nul". Rebutjarem H_0 si $|\hat{\tau}| > \tau_\alpha(n)$ on $\tau_{0.05}(12) = 0.455$ correspon al valor crític a les taules de la prova bilateral pel nivell 0.05 i mida mostral 12. Com que 0.576 > 0.455, podem rebutjar H_0, ens decantem a favor de la hipòtesi que afirma que la correlació de Kendall poblacional no és nul·la.

El valor obtingut del coeficient, positiu, indica que predominen les concordances entre X i Y, si d'un alumne "a" a un alumne "b" la nota de probabilitats ha estat millor, la tendència és que això també passi per discretes, etc.

2) Indica el valor del coeficient de correlació de Spearman.

Del resultat de `cor.test(rank(x), rank(y))`*, se'n dedueix que el valor demanat és 0.7275934.*

Aquesta pregunta solament es considerarà ben contestada si l'elecció d'aquest valor es justifica, és a dir si diu clarament que el coeficient de correlació de Spearman correspon al coeficient de correlació de Pearson però ***entre els rangs.***

3) Realitza una prova de permutacions per a determinar si el coeficient de correlació lineal de Pearson és significatiu.

H_0 : "X i Y estocàsticament independents" vs H_1 : $\rho(X,Y) \neq 0$, "el coeficient de correlació de Pearson poblacional no és nul". Si H_0 fos certa, la mostra observada seria una més de les possibles permutacions a les quals deixem igual un dels vectors d'observacions (indistintament, de X o de Y) i permutem l'altre. Aquí el nombre de permutacions possibles de les observacions Y (o X) és molt gran, 12!, és casi impossible fer una prova de permutacions exacta, enumerant totes les permutacions possibles. Utilitzarem l'aproximació de Monte Carlo: generem una mostra de 9999 permutacions aleatòries (permutant els valors de Y, `sample(y, replace = FALSE)`*) i sobre cadascuna calculem el coeficient de correlació de Pearson mostral. D'aquestes 9999 mostres permutades, solament 12 proporcionen un valor de correlació mostral més extrem (per negatiu o positiu, no oblidem que tenim una alternativa bilateral:* `sum(abs(r.perms) >= abs(r))`*) que el valor de la correlació sobre la mostra original: $r = 0.837$. Per tant, l'estimador de Dwass del p-valor, (12 + 1) / (9999 + 1) = 0.0013 permet rebutjar H_0.*

4) Calcula l'interval de confiança bootstrap percentil pel coeficient de correlació de Pearson (bootstrap no paramètric).

Correspon directament als quantils mostrals 0.025 i 0.0975 del vector de B = 10000 valors de la correlació mostral de Pearson obtinguts a partir d'altres tantes remostres bootstrap:

```
quantile(r.boots["r*",], probs = c(0.025, 0.975))
     2.5%     97.5%
0.4268142 0.9788047
```

5) Calcula l'interval de confiança bootstrap-t per al coeficient de correlació de Pearson (bootstrap no paramètric).

En principi es tractaria de l'interval de confiança bootstrap-t "no simetritzat", de cues iguals (totes dues amb probabilitat bootstrap 0.025)

$$\left[r - t^*_{0.975}\widehat{SE}_r,\ r - t^*_{0.025}\widehat{SE}_r\right]$$

A partir de:

```
quantile(t.boots, probs = c(0.975, 0.025))
    97.5%      2.5%
10.123830 -1.505648
```

Tenim que $t^*_{0.975} = 10.124$, $t^*_{0.025} = -1.506$ de manera que l'interval seria $[0.837 - 10.124\ 0.0997,\ 0.837 + 1.506\ 0.0997] = [-0.172,\ 0.987]$.

Si algú calcula l'interval "simetritzat" també es pot considerar bé, bastaria substituir els valors crítics t^* per ± 5.685

.

LLISTATS R

```
> # ****************************************************************
> #      PROBLEMA 1
> # ****************************************************************
> avaluacions = read.table("avaluacions.txt", header = TRUE)
>
> wilcox.test(avaluacions$probabilitats, avaluacions$discretes,
+   paired = FALSE, correct = FALSE, conf.int = TRUE)

        Wilcoxon rank sum test

data:  avaluacions$probabilitats and avaluacions$discretes
W = 99, p-value = 0.1186
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -3.000033 24.000009
sample estimates:
difference in location
               11.2736

Warning messages:
1: In wilcox.test.default(avaluacions$probabilitats, avaluacions$discretes,  :
  cannot compute exact p-value with ties
2: In wilcox.test.default(avaluacions$probabilitats, avaluacions$discretes,  :
  cannot compute exact confidence intervals with ties
> wilcox.test(avaluacions$probabilitats, avaluacions$discretes,
+   paired = FALSE, alternative = "greater", correct = FALSE, conf.int = TRUE)

        Wilcoxon rank sum test

data:  avaluacions$probabilitats and avaluacions$discretes
W = 99, p-value = 0.05932
alternative hypothesis: true location shift is greater than 0
95 percent confidence interval:
 -0.9999481        Inf
sample estimates:
difference in location
               11.2736

Warning messages:
1: In wilcox.test.default(avaluacions$probabilitats, avaluacions$discretes,  :
  cannot compute exact p-value with ties
2: In wilcox.test.default(avaluacions$probabilitats, avaluacions$discretes,  :
  cannot compute exact confidence intervals with ties
>
> # Diferències:
> d = avaluacions$probabilitats - avaluacions$discretes
> d
 [1] 13 31  5 16 20  0  4 26  8  3  5  5
> n = length(d)
> n
[1] 12
> # Valors absoluts de les diferències:
> abs.d = abs(d)
> abs.d
 [1] 13 31  5 16 20  0  4 26  8  3  5  5
> #
> # Rangs dels valors absoluts de les diferències:
> rabs.d = rank(abs.d)
> rabs.d
 [1]  8 12  5  9 10  1  3 11  7  2  5  5
> #
> # Suma de rangs de diferencies positives:
> r.plus = sum(rabs.d[d > 0])
> r.plus
```

```
[1] 77
> # Suma de rangs de diferencies negatives:
> r.minus = sum(rabs.d[d < 0])
> r.minus
[1] 0
> #
> #
> # Mediana de totes les diferències:
> median(d)
[1] 6.5
> # Mediana de totes les semisumes entre diferències:
> sSums = outer(d, d, "+") / 2
> median(sSums[lower.tri(sSums, diag = TRUE)])
[1] 10.5
>
> # Permutacions sobre el vector de 12 diferències
> # ==============================================
> # Enumeració de TOTES les permutacions possibles, maneres segons les quals
> # podem permutar DINS cada parella de valors (probabilitats, discretes).
> # En altres paraules, maneres possibles segons les quals podem donar un
> # signe - o + a les diferències:
> sgn = c(-1, +1)
> signsTab = expand.grid(as.data.frame(matrix(rep(sgn, n), ncol = n)))
> signsTab = apply(signsTab, 1, "*", abs.d)
> # Cada columna de 'signsTab' conté les diferències sobre una permutació
> # possible. Per exemple les 10 primeres:
> signsTab[,1:10]
    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
V1   -13   13  -13   13  -13   13  -13   13  -13    13
V2   -31  -31   31   31  -31  -31   31   31  -31   -31
V3    -5   -5   -5   -5    5    5    5    5   -5    -5
V4   -16  -16  -16  -16  -16  -16  -16  -16   16    16
V5   -20  -20  -20  -20  -20  -20  -20  -20  -20   -20
V6     0    0    0    0    0    0    0    0    0     0
V7    -4   -4   -4   -4   -4   -4   -4   -4   -4    -4
V8   -26  -26  -26  -26  -26  -26  -26  -26  -26   -26
V9    -8   -8   -8   -8   -8   -8   -8   -8   -8    -8
V10   -3   -3   -3   -3   -3   -3   -3   -3   -3    -3
V11   -5   -5   -5   -5   -5   -5   -5   -5   -5    -5
V12   -5   -5   -5   -5   -5   -5   -5   -5   -5    -5
> # Nombre de permutacions possibles:
> nperm = ncol(signsTab)
> nperm
[1] 4096
> #
> # Estimació de la mitjana de les diferències sobre cada possible permutació:
> m.perm = apply(signsTab, 2, mean)
> #
> # La mitjana de les diferències a la mostra original és:
> m.d = mean(d)
> #
> sum(m.perm >= m.d)
[1] 2
> sum(abs(m.perm) >= abs(m.d))
[1] 4
> sum(m.perm <= m.d)
[1] 4096

> # ****************************************************************
> #      PROBLEMA 2
> # ****************************************************************
> x = avaluacions$probabilitats
> y = avaluacions$discretes
>
> # Taula amb totes les possibles diferències entre x[i] i x[j]:
> difs.x = outer(x, x, "-")
> # Descartem les diferències de la diagonal (i == j) i de la meitat triangular
> # superior:
> difs.x = difs.x[ltri <- lower.tri(difs.x)]
```

```
> # Totes les possibles diferències entrey y[i] i y[j]:
> difs.y = outer(y, y, "-")[ltri]
> #
> # Nombre de concordances:
> concor = sum(sign(difs.x)*sign(difs.y) > 0)
> concor
[1] 51
> # Nombre de discordances:
> discor = sum(sign(difs.x)*sign(difs.y) < 0)
> discor
[1] 13
> #
> cor.test(rank(x), rank(y))

        Pearson's product-moment correlation

data:  rank(x) and rank(y)
t = 3.354, df = 10, p-value = 0.007316
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2638789 0.9181189
sample estimates:
      cor
0.7275934

> #
> # Coeficient de correlació lineal de Pearson sobre la mostra original:
> r = cor(x,y)
> r
[1] 0.8372688
> # En tot el problema, com a estimació de l'error estàndard del coeficient
> # de correlació de Pearson mostral farem servir l'estimació paramètrica normal:
> se.r = (1 - r*r) / sqrt(n - 3)
> se.r
[1] 0.09966033
>
> # Una permutació aleatòria del vector 'y':
> y.perm = sample(y, replace = FALSE)
> # 9999 permutacions aleatòries i càlcul de la correlació:
> nperm = 9999
> set.seed(5719)
> r.perms = replicate(nperm, cor(x, sample(y, replace = FALSE)))
> #
> sum(r.perms >= r)
[1] 12
> sum(abs(r.perms) >= abs(r))
[1] 12
> sum(r.perms <= r)
[1] 9987
>
> # 1 remostra bootstrap:
> # Determino quins estudiants formaran part de la mostra aleatòria i amb
> # reemplaçament:
> i.boot = sample(1:n, replace = TRUE)
> i.boot
 [1] 1 1 5 8 6 7 6 7 6 6 5 7
> # Remostra bootstrap:
> x[i.boot]
 [1]  65  65  50 100  67 100  67 100  67  67  50 100
> y[i.boot]
 [1] 52 52 30 74 67 96 67 96 67 67 30 96
> # Correlació sobre la remostra bootstrap:
> r.boot = cor(x[i.boot], y[i.boot])
> r.boot
[1] 0.914613
> # Error estàndar d'aquesta estimació de la correlació:
> se.boot = (1 - r.boot*r.boot) / sqrt(n - 3)
> se.boot
[1] 0.05449437
```

```
> # 10000 rèpliques bootstrap no paramètric de r i del seu error estàndard:
> nboot = 10000

> set.seed(5719)
> r.boots = replicate(nboot,
+ {
+   i.boot = sample(1:n, replace = TRUE)
+   r.boot = cor(x[i.boot], y[i.boot])
+   se.boot = (1 - r.boot*r.boot)
+   c(r.boot, se.boot)
+ }
+ )
> rownames(r.boots) = c("r*", "se*")
> # Faltava dividir per sqrt(n - 3):
> r.boots[2,] = r.boots[2,] / sqrt(n - 3)
> # r.boots és una matriu de 2 files i 10000 columnes,
> # la primera fila són les répliques bootstrap de la correlació,
> # la segona fila son els corresponents errors estàndard.
> # Les 5 primeres rèpliques bootstrap:
> r.boots[,1:5]
         [,1]      [,2]      [,3]       [,4]       [,5]
r*  0.5389235 0.7971575 0.7980396 0.98421245 0.85728882
se* 0.2365205 0.1215133 0.1210443 0.01044195 0.08835196
> # Valors "estudentitzats":
> t.boots  = (r.boots["r*",] - r) / r.boots["se*",]
> # Els 5 primers:
> t.boots[1:10]
 [1] -1.2613929 -0.3300982 -0.3240894 14.0724303  0.2265941
> #
> # Alguns quantils:
> quantile(r.boots["r*",], probs = c(0.025, 0.975))
     2.5%     97.5%
0.4268142 0.9788047
> quantile(t.boots, probs = c(0.975, 0.025))
    97.5%      2.5%
10.123830 -1.505648
> quantile(abs(r.boots["r*",]), probs = 0.95)
      95%
0.9655576
> quantile(abs(t.boots), probs = 0.95)
     95%
5.684999
```