

Aquest entregable és continuació de l'exercici de la pràctica 2 de Descriptiva. Té el mateix enunciat però heu de contestar les preguntes utilitzant les dades que hi ha al fitxer amb el vostre nom de la carpeta "Dades Entregable 1" d'Atenea. Comenceu per crear les noves variables: BMI (índex de massa corporal) i IM (índex de Manning) i fer una nova variable qualitativa, CEat, per classificar les persones en "Joves" si $Edat \leq 25$ anys i "Grans" en cas contrari.

- a) Feu la descriptiva numèrica de la variable BMI amb els factors Sexe i CEat. Utilitzeu la mitjana i la variància com a indicadors de la tendència central i de la dispersió. Compareu el comportament dels 4 tractaments, entenent per tractaments les combinacions dels nivells dels dos factors, és a dir: F-Jove, F-Gran, M-Jove i M-Gran.

1) La mitjana de BMI en el tractament F-Jove és:	
2) La variància de BMI en el tractament F-Jove és:	
3) El tractament en el que BMI ha donat més alt, és:	F-Jove F-Gran M-Jove M-Gran
4) El tractament en el que BMI ha sortit amb més dispersió, és:	F-Jove F-Gran M-Jove M-Gran
5) El Sexe en el que la diferència de BMI entre Joves i Grans ha sortit més alta, és:	F M
6) El grup d'edat en el que la diferència de BMI entre sexes ha sortit més alta, és:	Jove Gran

- b) Suposant que la variable H en la combinació de sexe i grup d'edat M-Gran, segueix una distribució normal d'esperança μ i variància σ^2 :

7) Calculeu la mitjana de H per les persones del segment M-Gran
8) Suposant que $\mu = E[H M-Gran] = 171,2$, quins serien els resultats normals (dues cues, probabilitat = 0.95) de la mitjana de H del segment M-Gran? Doneu-ne els extrems de l'interval (valors crítics):
9) Continuant el dos punts anteriors (7 i 8), quina és la probabilitat que la mitjana $H M-Gran$ surti tant o més estranya del que ha sortit?
10) Calculeu la variància de H per les persones del segment M-Gran
11) Suposant que $\sigma^2 = V(H M-Gran) = 78$, quins serien els resultats normals (dues cues, probabilitat = 0.95) de la variància de H del segment M-Gran? Doneu-ne els valors crítics
12) Continuant el dos punts anteriors (10 i 11), quina és la probabilitat que la variància S^2 de $H M-Gran$ surti tant o més estranya del que ha sortit?

- c) Descriviu el comportament conjunt de les variables Aspecte i Vidre. Contesteu:

13) De les persones que han respost la opció: Espai d'Aspecte, quina proporció (tant per u) ha dit: Sí a Vidre?	
14) En quina opció d'Aspecte, les proporcions de "Sí" i de "No" a Vidre són més diferents?	E F M T
15) En quina opció d'Aspecte, les proporcions de "Sí" i de "No" a Vidre s'assemblen més a les proporcions de "Sí" i "No" de totes les persones juntes?	Espai Futur Manera Temps

Al qüestionari d'Atenea escriviu les respostes numèriques amb 5 (o més) xifres significants. Si convé es pot utilitzar la notació científica, per exemple: $1.2345e-9$

Suposeu que X és una variable aleatòria de distribució Binomial Negativa, amb domini $\{0, 1, 2, \dots, k, \dots\}$ i paràmetres $r > 0$ i $p \in (0, 1)$ desconeguts. Aquesta és la parametrització utilitzada a la taula de distribucions habituals. S'ha obtingut una mostra aleatòria de X . La seva taula de freqüències és:

X	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Freq	3	14	36	60	120	142	198	191	223	197	177	146	132	105	69

X	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Freq	56	45	27	25	9	7	3	7	6	1	0	1	0	0	0

Per comprovar si heu entrat bé les dades, la suma de les X^2 és: 193736. En totes les respostes feu servir 5 (o més) xifres significants, i quan convingui es poden escriure en format científic.

- a) Estimeu pel mètode dels moments els paràmetres p i r , utilitzant els dos primers moments. Per veure si l'estimació ajusta bé, dibuixeu i compareu la funció de distribució estimada $F_{mm}(x)$ amb l'empírica $F_{fr}(x)$ i contesteu les preguntes següents:

1) El valor estimat de p és: $\hat{p} =$
2) El valor estimat de r és: $\hat{r} =$
3) El valor estimat de $E(X)$ és: $\hat{\mu} =$
4) El valor estimat de $V(X)$ és: $\widehat{\sigma^2} =$
5) $\max(F_{mm}(x) - F_{fr}(x)) =$

- b) Estimeu pel mètode de màxima versemblança els paràmetres p i r . Per veure si l'estimació ajusta bé, dibuixeu i compareu la funció de distribució estimada $F_{mv}(x)$ amb l'empírica $F_{fr}(x)$ i contesteu les preguntes següents:

6) El valor estimat de p és: $\hat{p} =$
7) El valor estimat de r és: $\hat{r} =$
8) El valor estimat de $E(X)$ és: $\hat{\mu} =$
9) El valor estimat de $V(X)$ és: $\widehat{\sigma^2} =$
10) $\max(F_{mv}(x) - F_{fr}(x)) =$
11) Assumint que els valors dels paràmetres de X fossin els valors estimats per màxima versemblança, trobeu $[a, b]$, la regió de predicció de X (amb $\text{prob} \geq 0,99$ i dues cues equiprobables):
12) De la regió de l'apartat 11) calculeu $\Pr(X \in [a, b]) =$

- c) En el supòsit que considerem que p és una v.a. amb funció de densitat proporcional a $f(p) \propto p^a(1-p)^b$, amb $a=0,6$ i $b=2,7$ amb el valor de r conegut i igual a 11. Un cop coneguda la vostra mostra, calculeu la funció de densitat a posteriori de p , dibuixeu-la i contesteu les preguntes següents:

13) L'esperança de la distribució a posteriori de p , és:
14) La variància de la distribució a posteriori de p , és:
15) A posteriori, l'esperança de X és:
16) A posteriori, la variància de X és:

Considerem la variable aleatòria biparamètrica X amb funció de densitat:

$$f(x; \lambda, \eta) = \frac{\eta^\lambda}{\Gamma(\lambda)} x^{-\lambda-1} e^{-\frac{\eta}{x}}$$

per $x > 0, \lambda > 0$ i $\eta > 0$

Comproveu que és una família exponencial amb els estadístics $T(x) = (\log(x), 1/x)$, trobeu-ne el paràmetre canònic $\theta = (\theta_1, \theta_2)$, i la constant $C(\theta)$. En algun apartat caldrà fer servir la funció digamma $\psi(x)$ que correspon a la derivada del logaritme de la funció Gamma.

A) En el cas que sabéssim que $E[T(X)] = (1,85, 0,18)$, contesteu les preguntes següents:

1) Quin seria el valor de λ ?
2) Quin seria el valor de η ?
3) Quin seria el valor de $V(\log(X))$?
4) Quin seria el valor de $V(1/X)$?
5) Quin seria el valor de $Cov(\log(X), 1/X)$?

B) En la situació habitual en la que no coneixem el valor dels paràmetres, però tenim una mostra: Si la grandària és 329, la mitjana dels logaritmes de les X ha donat 1,84599 i la de les inverses $1/x$ ha donat 0,17652, estimeu per màxima versemblança els paràmetres de la distribució λ i η . Calculeu la matriu d'Informació en els paràmetres canònics, avalueu la matriu en el màxim versemblant, i contesteu:

Nota: Feu servir la Suficiència, no obstant, les dades són a Dades entregable 3 d'Atenea amb el vostre nom.

6) L'estimador m.v. de λ ha donat:
7) L'estimador m.v. de η ha donat:
8) El terme [1,1] de la matriu d'informació obtinguda és:
9) El terme [2,2] de la matriu d'informació obtinguda és:
10) El terme [1,2] de la matriu d'informació obtinguda és:

C) En la mateixa situació de l'apartat B), calculeu la fita de Cramer-Rao de l'estimador màxim versemblant de $T(X) = (E[\log(X)], E[1/X])$. Utilitzant el teorema central del límit, calculeu l'interval de confiança asimptòtic (95% dues cues) del paràmetre λ , i també el de η . Contesteu:

11) La fita de Cramer-Rao de l'estimador m.v. de $E[\log(X)]$ és:
12) La fita de Cramer-Rao de l'estimador m.v. de $E[1/X]$ és:
13) L'interval de confiança de λ és:
14) L'interval de confiança de η és:

Al qüestionari d'Atenea escriviu les respostes numèriques amb 5 (o més) xifres significants.

Entregable 4: Proves d'hipòtesi i Intervalls de confiança

Llegeix el següent article i respon a les qüestions plantejades:

ARTICLE

'Efectividad de un programa de juego basado en realidad virtual para la mejora cognitiva en la esquizofrenia'

O.Lopez-Martin, A.Segura Fragoso, M.Rodriguez Hernández, I.Dimbwadyo Terrorer, B.Polonio-López

Gaceta Sanitaria. 2016; 30(2):133-136

<http://gacetasanitaria.org/es/pdf/S0213911115002083/S300/>

Un grup d'investigació en aquest tema està interessat en replicar l'estudi amb dades de pacients que han recollit. En particular, volen analitzar si les conclusions referides als diferents Scores confirmen el resultats presentats a l'article. Com no coneixen les metodologies estadístiques de test d'hipòtesi, us demanen que els hi ajudeu a aplicar-les a les seves dades.

La base de dades que us faciliten inclouen dades de 70 pacients i té el següent contingut:

- **ID:** Identificador de l'individu
- **Grupo:** Grup al que es va assignar aleatòriament (GC="Grup Control", GE="Grup Experimental")
- **Sexo:** Sexe de l'individu
- **Edad:** Edat en anys de l'individu
- **NivEstud:** Nivell d'estudis de l'individu (PI="Primaris incomplerts", PC="Primaris complerts", S="Secundaris")
- **EstCiv:** Estat civil de l'individu (S="Soltero/a", O="Altres")
- **Diag:** Diagnòstic principal (TP="Tipus Paranoide", OT="Altres Tipus")
- **AnEvol:** Anys d'evolució de la malaltia
- **VelocProcB:** T-Score Basal de velocidad de procesamiento (abans de la intervenció)
- **AtenVigB:** T-Score Basal d'atenció i vigilància (abans de la intervenció)
- **MemorTrabB:** T-Score Basal de memoria de trabajo (abans de la intervenció)

- **AprenVerbB**: T-Score Basal d'aprenentatge verbal (abans de la intervenció)
- **AprenVisB**: T-Score Basal d'aprenentatge visual (abans de la intervenció)
- **RazonResB**: T-Score Basal de raonament i resolució de problemes (abans de la intervenció)
- **VelocProcP**: T-Score Post-intervenció de velocidad de procesamiento
- **AtenVigP**: T-Score Post-intervenció d'atenció i vigilància
- **MemorTrabP**: T-Score Post-intervenció de memòria de trabajo
- **AprenVerbP**: T-Score Post-intervenció d'aprenentatge verbal
- **AprenVisP**: T-Score Post-intervenció d'aprenentatge visual
- **RazonResP**: T-Score Post-intervenció de raonament i resolució de problemes

En tots els test, feu servir el nivell de significació $\alpha = 0.05$ i considereu les hipòtesis alternatives de tipus bilateral.

Resoleu els apartats següents:

1. Inicialment, tot i que l'assignació dels grups Control i Experimental s'hagi fet a l'atzar, és important establir que no hi hagi diferències en les característiques dels participants en ambdós grups. Així doncs, heu de fer els tests necessaris per veure que la distribució de *Sexe*, *Nivell d'estudis*, *Estat civil* i *Diagnòstic principal* és homogènia en tots dos grups. Així mateix, les mitjanes d'*Edat* i d'*Anys d'evolució* dels dos grups no han de ser significativament diferents. Si es detecta alguna diferència significativa, s'hauria de reportar als experts per si pugués ser motiu de confusió en les anàlisis de la intervenció.

Reprodueix el contingut de la taula 1 de l'article (excepte la grandària de l'efecte de Cohen) fent els tests corresponents amb les dades del fitxer per determinar si podem considerar que els grups tenen aquestes sis característiques semblants.
2. Ara centrat en les variables T-Score, el primer que s'ha de comprovar és si la situació Basal és similar en els dos grups. Per això, cal que apliqueu el test corresponent per comparar en cada T-Score Basal si hi ha diferències entre els dos grups (Taula 2, segona columna). Una primera aproximació a poder establir la efectivitat de la intervenció seria comprovar si hi ha diferències entre els dos grups amb els valors dels T-Score Post-intervenció. Fes el test corresponent per decidir aquesta situació (Taula 2, tercera columna)

3. Una forma de quantificar la variació del T-Score en cada grup després de la intervenció consisteix en fer un interval de confiança al 95% per a la variació del T-Score (Postintervenció - Basal). Calcula aquest interval de confiança dels dos T-Scores per a cada grup (Taula 2, quarta columna).
4. De forma equivalent als intervals de confiança, podem fer un test per comparar si hi ha diferència entre el valor de l'Score post-intervenció i el basal, dins de cada grup. Això implica que per cada individu tindrem el valor basal i post-intervenció i es tracta de determinar si la mitjana de les diferències és diferent de zero. Tot i que en l'article no ho fan, aplica el test corresponent.
5. Finalment, s'ha de determinar si hi ha diferències en les variacions en els T-Scores (Postintervenció - Basal) quan comparem els grups. En l'article s'ha aplicat la tècnica de l'ANOVA i els estadístics i els seus p-valors apareixen en les columnes 5 i 6 de la taula 2. En el vostre cas heu d'aplicar el test t de dues mostres corresponent per comparar les mitjanes de les variacions (Postintervenció - Basal) entre els dos grups i reportar l'estadístic t i el seu p-valor.

Al qüestionari d'Atenea escriviu les respostes numèriques amb 5(o més) xifres significants.

COMENTARIS:

- La implementació en R dels test de la chi-quadrat per a taules de contingència, per defecte fa una correcció en el càlcul de l'estadístic. Per als resultats que es demanen no s'ha d'aplicar aquestes correccions i per tant, seleccioneu el paràmetre `correct=FALSE` de la funció `chisq.test`
- En els casos en que s'ha d'aplicar el test de la t-Student per a mostres independents, és necessari verificar si les variàncies són o no iguals aplicant el test corresponent. En funció del resultat, s'ha de modificar el paràmetre `var.equal` en la funció `t.test`.
- La tècnica de l'ANOVA (ANalysis Of VAriance) per a variables contínues que es fa servir en l'article correspon a un test de comparació de mitjanes de dos o més grups amb disseny aleatoritzat (independent). En aquest cas, com hi ha dos grups (control i experimental) es pot resoldre el test de forma equivalent fent la prova de la t-Student corresponent.
- La tècnica de l'ANOVA de mesures repetides també correspon a un test de comparació de mitjanes de dos o més grups quan hi ha més d'una mesura en cada unitat experimental. En aquest cas, com hi ha dos grups (control i experimental) i es fan dues mesures per a cada individu (basal i post-intervenció), es pot resoldre el test fent una prova de la t-Student per a dos grups on les dades corresponen a les diferències individuals (post-basal).

Les preguntes que haureu de contestar al qüestionari d'Atenea, les teniu a continuació agrupades pels apartats de l'exercici.

1.

1) Estadístic de la prova per comparar els anys d'evolució en els dos grups:
2) P-valor de la prova per comparar els anys d'evolució en els dos grups:
3) Estadístic de la prova per comparar el nivell d'estudis en els dos grups:
4) P-valor de la prova per comparar el nivell d'estudis en els dos grups:

2.

5) Estadístic de la prova per comparar el T-Score de Memòria de Treball Basal dels dos grups:
6) P-valor de la prova per comparar el T-Score de Memòria de Treball Basal dels dos grups:
7) Estadístic de la prova per comparar el T-Score de Memòria de Treball Post-Intervenció dels dos grups:
8) P-valor de la prova per comparar el T-Score de Memòria de Treball Post-Intervenció dels dos grups:

3.

9) Límit Inferior de l'interval de confiança del 95% per a la diferència del T-Score de Memòria de Treball (Post-Intervenció-Basal) en el Grup Control:
10) Límit Superior de l'interval de confiança del 95% per a la diferència del T-Score de Memòria de Treball (Post-Intervenció-Basal) en el Grup Control:
11) Límit Inferior de l'interval de confiança del 95% per a la diferència del T-Score de Memòria de Treball (Post-Intervenció-Basal) en el Grup Experimental:
12) Límit Superior de l'interval de confiança del 95% per a la diferència del T-Score de Memòria de Treball (Post-Intervenció-Basal) en el Grup Experimental:

4.

13) Estadístic de la prova per comparar el T-Score de Memòria de Treball Basal vs. Post-Intervenció en el grup Experimental:
14) P-valor de la prova per comparar el T-Score de Memòria de Treball Basal vs. Post-Intervenció en el grup Experimental:
15) Estadístic de la prova per comparar el T-Score de Memòria de Treball Basal vs. Post-Intervenció en el grup Control:
16) P-valor de la prova per comparar el T-Score de Memòria de Treball Basal vs. Post-Intervenció en el grup Control:

5.

17) Estadístic de la prova per comparar la diferència del T-Score de Memòria de Treball (Post-Intervenció-Basal) en els dos grups:
18) P-valor de la prova per comparar la diferència del T-Score de Memòria de Treball (Post-Intervenció-Basal) en els dos grups:

Entregable 5: Model Lineal General

El fitxer de dades conté estimacions del percentatge de greix corporal determinat pesant sota l'aigua, a més de diferents mesures de circumferències corporals per 200 individus. Una mesura acurada del greix corporal és costosa i és desitjable disposar de mètodes d'estimació amb menys cost, basats en mesures indirectes.

D'altra banda, aquests individus, que 3 mesos abans tenien un percentatge de greix similar, havien estat assignats a quatre grups: un de control i tres tipus de tractaments que combinen dieta amb pautes d'entrenament físic de diferents nivell. Al cap de 3 mesos, es recull el seu estat actual.

Els camps són:

- **Density:** Densitat determinada a partir del pesatge submarí (g/cm^3)
- **bodyfat:** Percentatge de greix corporal de l'equació de Siri (1956)
- **Age:** Edat (anys)
- **Weight:** Pes (lliures)
- **Height:** Alçada (polzades)
- **Neck:** Circumferència del coll (cm)
- **Chest:** Circumferència toràcica (cm)
- **Abdomen:** Circumferència abdominal (cm)
- **Hip:** Circumferència de maluc (cm)
- **Thigh:** Circumferència de la cuixa (cm)
- **Knee:** Circumferència de genoll (cm)
- **Ankle:** Circumferència del turmell (cm)
- **Biceps:** Circumferència de bíceps (estesa) (cm)
- **Forearm:** Circumferència de l'avantbraç (cm)
- **Wrist:** Circumferència de canell (cm)
- **Diet:** Grup de tractament (A-(No Diet),B,C,D)

Hi ha dos objectius:

1. Trobar una model estadístic que permeti predir la mesura del greix corporal (**Density** o **bodyfat**) en base a algunes de les mesures indirectes recollides.
2. Comparar els grups de tractament en quant a la mesura de greix corporal.

1 Anàlisis a realitzar

1. Per a la variable resposta proposada en l'apartat 1) de l'enunciat individual i amb el predictors indicats:
 - (a) Anomenem *M1* el model complet. Pren nota del valor de la *F* i del p-valor del test Omnibus d'aquest model, així com dels coeficients estimats i de la R^2 del model.
 - (b) Analitza la possible existència de multicolinealitat a partir dels VIFs del model
 - (c) Simplifica el model eliminant les variables predictores no significatives una a una ($\alpha = 0.05$). Anomenem aquest model *M2*.
 - (d) Valida les premises del model *M2*, analitzant els residus estudentitzats per veure si hi ha tendència i/o valors estranys. Analitzeu també els valors influents.
 - (e) Fent servir el model *M2*, calcula la predicció puntual i els intervals de predicció i confiança pels valors de les covariables indicats en l'enunciat individual.
2. Considerem ara el model per a la variable indicada en l'apartat 2) de l'enunciat individual respecte només al factor tractament **Diet**
 - (f) Fent servir el contrast de tipus baseline **contr.treatment**, calcula el model associat. Anomenem aquest model *M3*.
3. Considerem ara el model en el que s'inclou només la covariable **Weight** i el factor **Diet** amb la seva interacció
 - (g) Calcula el model associat. L'anomenem *M4*
4. Simplifica el model *M4* traient els termes no significatius. Anomenem aquest model *M5*

- z) Per a la regressió Lineal Múltiple, considera la variable resposta **bodyfat** fent servir com a variables predictores:
Chest, Hip, Knee, Forearm, Wrist

El cas proposat per predir és:

(Chest=100,6, Hip=101,2, Knee=38,7, Forearm=28,7, Wrist=18,2)

Per a la segona part, considera la variable resposta bodyfat respecte el factor tractament Diet i la covariable Weight

Les preguntes que haureu de contestar al qüestionari d'Atenea, les teniu a continuació agrupades pels apartats de l'exercici.

1.

1) Coeficient de determinació R^2 del model total (M1):
2) Variància residual del model M1:
3) Valor de l'ordenada a l'origen (intercept) del model M1:
4) Estadístic F del test òmnibus del model M1:
5) P-valor de la taula de coeficients (summary) per a la primera variable explicativa (Chest) en el model M1:
6) Màxim dels VIFs en el model M1:
7) Número de variables predictores en el model M2:
8) P-valor del test que compara els models M1 i M2:
9) Percentatge de residus estudentitzats amb valor absolut superior a 2 en el model M2:
10) Valor màxim de les distàncies de Cook del model M2:
11) Límit inferior de la regió de predicció al 95% del cas proposat (model M2):
12) Límit superior de la regió de predicció al 95% del cas proposat (model M2):
13) Límit inferior de l'interval de confiança al 95% del cas proposat (model M2):
14) Límit superior de l'interval de confiança al 95% del cas proposat (model M2):

2.

15) Valor de l'intercept del model només amb la variable Diet amb contrast tipus treatment (model M3):
16) Coeficient de determinació R^2 del model M3:
17) Error Estàndard dels coeficients de les variables indicadores (dummy) per a les dietes en el model M3
18) Límit inferior de l'interval de confiança al 95% per la mitjana del grup Diet="A" (emmeans en el model M3):
19) Límit superior de l'interval de confiança al 95% per la mitjana del grup Diet="A" (emmeans en el model M3):
20) Desviació estàndard residual del model M4:
21) R^2 Ajustada pel model M4:
22) Estadístic F per decidir la significació de la interacció en el model M5: