

Document de Treball 8

Objectiu

L'objectiu d'aquesta sessió és:

1. Realitzar problema 3 de l'examen de Juny 2019
2. Realitzar problema 1 de l'examen de Juny de 2019
3. Realitzar problema 2 de l'examen de Juny de 2019

1. Problema 3 de l'examen Juny 2019

S'ha realitzat un estudi al llarg de 10 anys en el qual s'investiga la relació entre el consum de cafè i el risc de depressió en les persones. Per a això es van recollir dades de 50.739 individus sense símptomes de depressió a l'inici de l'estudi en l'any 1996, i es va fer un seguiment fins al 2006.

Els investigadors van utilitzar uns qüestionaris per recopilar dades sobre el consum de cafè (amb cafeïna) sobre el diagnòstic clínic de la malaltia de la depressió. També es van obtenir dades sobre l'ús de tractaments antidepressius. La taula següent mostra la distribució de la incidència de la malaltia de la depressió classificada segons la quantitat de cafè que solen prendre:

		Consumo de tazas de café				
		Hasta una por semana	De 2 a 6 por semana	1 al día	De 2 a 3 por día	4 o más al día
Diagnóstico de depresión	Sí	670	373	905	564	95
	No	11545	6244	19329	11726	2288

Contesteu a les següents preguntes, justificant totes les respostes:

- a) Calculeu un interval de confiança a l'95% per a la proporció de persones que pateixen depressió. També per a la proporció de persones que no la pateix. Expliqueu el significat de cada un d'ells. Construeixi els mateixos intervals però aquesta vegada per a una confiança el 99%. ¿Observa diferències pel que fa als primers intervals calculats?
- b) Quin tipus de test és l'apropiat per avaluar l'associació estadística entre el consum de cafè i la malaltia de la depressió?
- c) Escriu les hipòtesis del test identificat en l'apartat anterior. Es compleixen les condicions d'aplicabilitat del test que has plantejat? A quina conclusió arribes?
- d) Els investigadors volen també comparar si hi ha diferències entre prendre només una tassa al dia o prendre'n més d'una. Escriu les hipòtesis, verifica les condicions d'aplicabilitat i obté una conclusió per a aquesta comparació.

Solució:

- a) Calculeu un interval de confiança a l'95% per a la proporció de persones que pateixen depressió. També per a la proporció de persones que no la pateix. Expliqueu el significat de cada un d'ells. Construeixi els mateixos intervals però aquesta vegada per a una confiança el 99%. ¿Observa diferències pel que fa als primers intervals calculats?

Definim π = proporció en la població de persones que pateixen depressió.

i definim π_n = proporció en la població de persones que no pateixen depressió (recordeu que aquest valor és $1 - \pi$).

Aleshores com el nombre d'individus de la mostra és 50.739 dels quals si sumem els que tenen depressió fan un total de 2607 sabem que la proporció mostral, és a dir, $p = \text{proporció en la mostra}$ de persones que pateixen depressió y estimació puntual de π , es de $p = 2607/50739$. Llavors hem de mirar les taules 3.3 i 3.4, obtenint els intervals de confiança del 95% i 99% respectivament.

```
> binom.test(2607, 53739, conf.level=.95)

Exact binomial test

data: 2607 and 53739
number of successes = 2607, number of trials = 53739, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.04671089 0.05036257
sample estimates:
probability of success
 0.04851225
```

Tabla 3.3

```
> binom.test(2607, 53739, conf.level=.99)

Exact binomial test

data: 2607 and 53739
number of successes = 2607, number of trials = 53739, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
99 percent confidence interval:
 0.04615576 0.05094910
sample estimates:
probability of success
 0.04851225
```

Tabla 3.4

Anàlogament per al cas de $\pi_n = \text{proporció en la població}$ de persones que no pateixen depressió, amb $p_n = \text{proporció en la mostra}$ de persones que no pateixen depressió y estimació puntual de π_n , $p_n = 51132/50739$

```
> binom.test(51132, 53739, conf.level=.95)

Exact binomial test

data: 51132 and 53739
number of successes = 51132, number of trials = 53739, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.9496374 0.9532891
sample estimates:
probability of success
 0.9514877
```

Tabla 3.5

```
> binom.test(51132, 53739, conf.level=.99)

Exact binomial test

data: 51132 and 53739
number of successes = 51132, number of trials = 53739, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
99 percent confidence interval:
 0.9490509 0.9538442
sample estimates:
probability of success
 0.9514877
```

Tabla 3.6

- b) Quin tipus de test és l'apropiat per avaluar l'associació estadística entre el consum de cafè i la malaltia de la depressió?

Test d'independència. Una única població i dues variables qualitatives:

Variable 1 (X): consum de cafè {fins a un per setmana, dos o més per setmana, un al dia, 2 a 3 al dia, 4 a més al dia}

Variable 2(Y): depressió {sí, no }

- c) Escriu les hipòtesis del test identificat en l'apartat anterior. Es compleixen les condicions d'aplicabilitat del test que has plantejat? A quina conclusió arribes?

H_0 : Independència: El tindre depressió o no és independent del consum de cafè.

H_A : No independència: Tindre depressió està relacionat en el consum de cafè.

Condicions d'aplicabilitat

1. Una única població i dues variables qualitatives
2. Freqüències esperades major o igual a 5.

$$E(\text{Fila } i, \text{Columna } j) = \frac{(\text{Total fila } i) \times (\text{Total columna } j)}{\text{Total general}}$$

Obtenim les següents freqüències esperades que són totes major o igual que 5:

	1/setmana	2 o més/set	1/dia	2 o 3 / dia	4 o més /dia
si	592,6	321,0	981,6	596,2	115,6
no	11622,4	6296,0	19252,4	11693,8	2267,4

Resolem el contrast:

H_0 : Independència: El tindre depressió o no és independent del consum de cafè.

H_A : No independència: Tindre depressió està relacionat en el consum de cafè.

La forma de resoldre estos dos contrastos és la mateixa, mitjançant els test de la Khi-quadrat per a taules de contingència (Pearson's Chi-squared test). Si $p\text{-valor} < \alpha$, aleshores es rebutja H_0 .

```
> .Table # Counts
      Consumo tazas de cafe
Enfermedad Hasta_1_sem De_2_a_6_sem 1_dia De_2_a_3_dia 4_o_mas_dia
Si          670         373   905         564          95
No         11545        6244 19329        11726         2288

> .Test <- chisq.test(.Table, correct=FALSE)
> .Test

      Pearson's Chi-squared test

data: .Table
X-squared = 31.454, df = 4, p-value = 0.000002474
```

Tabla 3.1

Conclusió, existeix suficient evidència estadística per a dir que el consum de cafè i el risc de depressió estan relacionats.

- d) Els investigadors volen també comparar si hi ha diferències entre prendre només una tassa al dia o prendre'n més d'una. Escriu les hipòtesis, verifica les condicions d'aplicabilitat i obté una conclusió per a aquesta comparació.

Per a realitzar per vosaltres. Similar al cas c) una població amb dues variables categòriques

Variable 1 (X): consum de cafè {sí, no}

Variable 2(Y): depressió {sí, no}

S'utilitza la taula 3.2.

```
> .Table # Counts
      Consumo tazas de cafe
Enfermedad Una_al_dia Mas_de_una_al_dia
Si          905         659
No         16329        14014

> .Test <- chisq.test(.Table, correct=FALSE)
> .Test

      Pearson's Chi-squared test

data: .Table
X-squared = 9.8204, df = 1, p-value = 0.001726
```

Tabla 3.2

2. Problema 1 de l'examen Juny 2019

El consum insuficient de calci és un dels principals factors de l'osteoporosi. En un estudi realitzat al País Valencià per avaluar la prevalença de la ingesta de calci en adults, es van prendre mesures a 1206 subjectes, dels quals el 51% eren dones.

Segons l'esmentat estudi, publicat per Jiménez et al (2017), la ingesta diària de calci en adults segueix una distribució aproximadament normal de mitjana 741,4 mg / persona i desviació típica 334,9 mg / persona.

Segons l'IOM (Institute Of Medicine), es considera un consum insuficient de calci si la ingesta diària no supera la quantitat mínima recomanada, fixada en els 800 mg / persona.

- Quin percentatge, en aquesta població, ingereix com a mínim la quantitat de calci recomanada?
- Quin és el percentatge de la població que ingereix diàriament una quantitat superior a 1.000 mg / persona?
- Calcula el percentil 95 i interpreta el resultat.

Es visita una empresa de 10 treballadors. Denotem per Y el nombre d'empleats que tenen una ingesta de calci insuficient.

- Quina és la distribució de Y ?
- Calcula la probabilitat que com a màxim tres empleats tinguin una ingesta de calci insuficient.

Solució:

Dades que ens dona l'enunciat:

La ingesta diària de calci en adults segueix una distribució aproximadament normal de mitjana 741,4 mg / persona i desviació típica 334,9 mg / persona.

X = ingesta diària de calci en adults

$$X \sim N(\mu = 741.4, \sigma = 334.9)$$

Segons l'IOM (Institute Of Medicine), es considera un consum insuficient de calci si la ingesta diària no supera la quantitat mínima recomanada, fixada en els 800 mg / persona.

- a) **Quin percentatge, en aquesta població, ingereix com a mínim la quantitat de calci recomanada?**

L'IOM recomana una quantitat mínima de 800mg/persona, per tant ens pregunten:

La probabilitat que X siga més gran o igual que 800, on sabem que X segueix una distribució normal.

$$P(X \geq 800) = 0.43054862$$

```
> pnorm(c(400,600,800,1000,1200), mean=741.4, sd=334.9, lower.tail=FALSE)
[1] 0.84599553 0.66356617 0.43054862 0.22000664 0.08544276
```

Tabla 1.2

Per tant, el 43.05% de la població.

- b) **Quin és el percentatge de la població que ingereix diàriament una quantitat superior a 1.000 mg / persona?**

Anàlogament a l'apartat anterior:

$$P(X \geq 1000) = 0.22000664$$

```
> pnorm(c(400,600,800,1000,1200), mean=741.4, sd=334.9, lower.tail=FALSE)
[1] 0.84599553 0.66356617 0.43054862 0.22000664 0.08544276
```

Tabla 1.2

- c) Calcula el percentil 95 i interpreta el resultat.

Busquem el valor $q_{0.95}$ tal que $P(X \leq q_{0.95}) = 0.95$, hem de buscar una taula de percentils o quantils (taula 1.3) i per tant obtenim que eixe valor és $q_{0.95} = 1292.2615$

```
> qnorm(c(.01,.05,.50,.95,.99), mean=741.4, sd=334.9, lower.tail=TRUE)
[1] -37.6939 190.5385 741.4000 1292.2615 1520.4939
```

Tabla 1.3

Es visita una empresa de 10 treballadors. Denotem per Y el nombre d'empleats que tenen una ingesta de calci insuficient.

- d) Quina és la distribució de Y?

Y= nombre d'empleats dels n=10 que tenen una ingesta de calci insuficient.

Per a nosaltres, computarem com a èxit aquells empleats que tenen una ingesta de calci insuficient, quina és la probabilitat d'èxit? La probabilitat que una persona tinga una ingesta de calci insuficient, és a dir, menor que 800.

$$P(X \leq 800) = 0.5694514$$

```
> pnorm(c(400,600,800,1000,1200), mean=741.4, sd=334.9, lower.tail=TRUE)
[1] 0.1540045 0.3364338 0.5694514 0.7799934 0.9145572
```

Tabla 1.1

Y segueix una distribució Binomial amb n=10 i probabilitat d'èxit 0.5694514.

- e) Calcula la probabilitat que com a màxim tres empleats tinguin una ingesta de calci insuficient.

$$P(Y \leq 3) = P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3) = 0.08057631$$

```
> .Table <- data.frame(Probability=dbinom(0:10, size=10, prob=0.57))
      Probability
0 0.0002161148
1 0.0028647779
2 0.0170887332
3 0.0604066849
4 0.1401294608
5 0.2229036074
6 0.2462307291
7 0.1865136420
8 0.0927146302
9 0.0273112864
10 0.0036203333
```

Tabla 1.6

3. Problema 2 de l'examen Juny 2019

El peix constitueix una font important d'aliment per la seva aportació nutricional de proteïnes, greixos saludables, vitamines i minerals. No obstant això, sovint els peixos concentren mercuri en els seus organismes provinent de la contaminació de l'aigua. La presència de mercuri en els peixos pot tenir repercussions sobre la salut dels éssers humans, especialment en dones embarassades i nadons.

En l'estudi titulat Fish Intake, contaminants, and human health: Evaluating the risks and the benefits i signat per Mozaffarian et al (2006) s'analitza la concentració de mercuri en els peixos de dos rius de Carolina de Nord (EUA): el riu Lumber i el riu Waccamaw.

Les dades es van recollir en el conjunt de dades "mercuri". Les variables que es van mesurar en els dos grups, definits pels rius, van ser:

Conc: nivells de concentració de mercuri en els peixos (mg / kg)

Long: longituds dels peixos (cm)

- Dóna una estimació puntual de la concentració mitjana de mercuri dels peixos al riu Lumber. Seria correcte calcular un interval de confiança? En cas afirmatiu, calcula un interval de confiança a l'95%.
- Hi ha evidència estadística per afirmar a nivell $\alpha = 0.05$ que el nivell mitjà de concentració al riu Waccamaw és superior a el nivell mitjà en el riu Lumber? Justifica el procediment emprat.

Una part de l'estudi pretén determinar la relació entre la mida dels peixos i el seu nivell de contaminació per mercuri.

- Calcula la recta de regressió adequada per predir la concentració de mercuri a partir de la longitud dels peixos. Quin és el percentatge de la variabilitat de la variable dependent que s'explica per la regressió?
- Quina concentració de mercuri s'esperaria per un peix de 50 cm de longitud?.

Solució:

- Dóna una estimació puntual de la concentració mitjana de mercuri dels peixos al riu Lumber. Seria correcte calcular un interval de confiança? En cas afirmatiu, calcula un interval de confiança a l'95%.

Variable: X_1 = concentració de mercuri dels peixos al riu Lumber.

Paràmetre d'interés: μ_1 = concentració mitjana de mercuri dels peixos al riu Lumber.

Com la grandària de la mostra de concentració de mercuri dels peixos al riu Lumber és de $n_1 = 73$ que és més gran que 30, aleshores sabem pel teorema central del límit que $\bar{X}_1$ segueix una distribució normal amb mitjana μ_1 i desviació típica $\sigma_1/\sqrt{n_1}$. Podem per tant aplicar mètodes paramètrics i és correcte calcular un interval de confiança.

```
> numSummary(mecurio[, "Conc"], groups=mecurio$Rio, statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
      mean      sd   0%   25%   50%   75%  100% data:n
Waccamaw 1.276429 0.8291484 0.11 0.6175 1.05 1.8   3.6   98
Lumber    1.078082 0.6486110 0.27 0.5700 0.91 1.5   3.5   73
```

Tabla 2.1

En concret, l'interval de confiança per a μ_1 al 95% és:

```
> with(Lumber, (t.test(Conc, alternative='two.sided', mu=0.0, conf.level=.95)))

One Sample t-test

data: Conc
t = 14.201, df = 72, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.926750 1.229414
sample estimates:
mean of x
 1.078082
```

Tabla 2.4

- b) Hi ha evidència estadística per afirmar a nivell $\alpha = 0.05$ que el nivell mitjà de concentració al riu Waccamaw és superior a el nivell mitjà en el riu Lumber? Justifica el procediment emprat.

En aquest cas tenim dues mostres independents:

Variable: X_1 = concentració de mercuri dels peixos al riu Lumber.

Paràmetre d'interés: μ_1 = concentració mitjana de mercuri dels peixos al riu Lumber.

Variable: X_2 = concentració de mercuri dels peixos al riu Waccamaw.

Paràmetre d'interés: μ_2 = concentració mitjana de mercuri dels peixos al riu Waccamaw.

1. Normalitat:

Com són mostres independents, per a poder utilitzar test paramètrics hem de veure que $\bar{X}_1$ i $\bar{X}_2$ segueixen les dues una distribució normal. Per l'apartat a) sabem que $\bar{X}_1 \sim N(\mu_1, \sigma_1/\sqrt{n_1})$.

- $\bar{X}_2$ segueix una distribució normal?

També per la taula 2.1 veiem que $n_2 = 98$ és la grandària de la mostra > 30 aleshores $\bar{X}_2$ segueix una distribució normal amb mitjana μ_2 i desviació típica $\sigma_2/\sqrt{n_2}$ pel teorema central de límit (TCL).

Podem aplicar test t per a mostres independents.

2. Homogeneïtat de variàncies: Test de Levene.

Per saber si les variàncies poblacionals són iguals o no utilitzarem el **test de Levene**:

H_0 : Les variàncies dels nivells de mercuri en els dos rius són iguals ($\sigma_1 = \sigma_2$)

H_A : Les variàncies dels nivells de mercuri en els dos rius són diferents ($\sigma_1 \neq \sigma_2$)

Si $p\text{-valor} = 0.02569 \leq \alpha (=0.05)$ Rebutjar la igualtat de variàncies (H_0)

Utilitzarem el test t- per a mostres independents amb variàncies distintes.

```
> leveneTest(Conc ~ Rio, data=mecurio, center="median")
Levene's Test for Homogeneity of Variance (center = "median")
      Df F value    Pr(>F)
group  1  5.0658 0.02569 *
    169
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tabla 2.6

3. Contrast d'hipòtesi:

H_0 : El nivell mitjà de mercuri és el mateix en els dos rius ($\mu_2 = \mu_1$)

H_A : El nivell mitjà de concentració al riu Waccamaw (μ_2) és superior al nivell mitjà en el riu Lumber (μ_1) ($\mu_2 > \mu_1$)

Com el $p\text{-valor} = 0.0405 \leq \alpha (=0.05)$ aleshores rebutgem (H_0). I per tant podem concloure, que existeix suficient evidència estadística per dir a un nivell de significativitat de 0.05 que el El nivell mitjà de concentració al riu Waccamaw és superior al nivell mitjà en el riu Lumber.

```
> t.test(Conc~Rio, alternative='greater', conf.level=.95, var.equal=FALSE, data=mecurio)

Welch Two Sample t-test

data: Conc by Rio
t = 1.7547, df = 168.57, p-value = 0.04057
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.01138394      Inf
sample estimates:
mean in group Waccamaw    mean in group Lumber
      1.276429              1.078082
```

Tabla 2.12

Una part de l'estudi pretén determinar la relació entre la mida dels peixos i el seu nivell de contaminació per mercuri.

- c) Calcula la recta de regressió adequada per predir la concentració de mercuri a partir de la longitud dels peixos. Quin és el percentatge de la variabilitat de la variable dependent que s'explica per la regressió?

Volem predir Y = concentració de mercuri a partir de X = longitud dels peixos.

Aleshores la recta de regressió serà: $y = b_1 x + b_0 = 0.058127 x - 1.131645$

```
Call:
lm(formula = Conc ~ Long, data = mecurio)

Residuals:
    Min       1Q   Median       3Q      Max
-1.1499 -0.3436 -0.1022  0.3123  1.9100

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.131645   0.213615  -5.298 0.000000362 ***
Long         0.058127   0.005228  11.119 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5805 on 169 degrees of freedom
Multiple R-squared:  0.4225, Adjusted R-squared:  0.4191
F-statistic: 123.6 on 1 and 169 DF, p-value: < 2.2e-16
```

Tabla 2.15

El percentatge de la variabilitat de la variable dependent que s'explica per la regressió ens el dona el coeficient de la R^2 que en aquest cas és 0.4225, per tant, el 42.25% de la variabilitat de la concentració de mercuri queda explicada per la longitud dels peixos.

- d) Quina concentració de mercuri s'esperaria per un peix de 50 cm de longitud?

Substituïm en la recta de regressió: $y_{50} = 0.058127 \cdot 50 - 1.131645 = 1.774705$