

Document de Treball 5

Objectiu

L'objectiu d'aquesta sessió és:

1. Anàlisi d'una variable categòrica dicotòmica.
2. Anàlisi d'una variable categòrica en tres o més categories.
3. Exemple d'una variable categòrica dicotòmica.
4. Exemple d'una variable categòrica en tres o més categories.

1. Anàlisi d'una variable categòrica dicotòmica

Material de l'aulavirtual:

Teoria	Activitats amb R	Activitats sense R	Indestap
Secció: Unitat 2. Anàlisi estadística d'una variable. Pdf: Tema 7. Inferència sobre una proporció. Pàgines: 1 a 19	Document: Exercicis unitat 2. Dades: Dades unitat 3. Exercicis: 1,2,16,17,18 i 19.	Secció: Problemes Carpeta: Categòriques.	Document: Indestap Secció del document: 2.6- Subsecció:2.6a
Unitat 2. Anàlisi estadística d'una variable  Tema 4 Inferència sobre una mitjana poblacional Pujat 20/02/2020 10:30  Una mostra_Unitat 2  Tema 7 Inferència sobre una proporció Pujat 27/01/2020 10:16  Exercicis unitat 2 Pujat 27/01/2020 10:17  dades unitat 2 Pujat 27/01/2020 10:17 Problemes  Probabilitat  Una mostra  Dues mostres  k mostres  Categòriques Indestap  Instal·lació Rcmdr Pujat 3/02/2020 09:40  Instal·lació Rcmdr en MacOS Pujat 3/02/2020 09:40  Resolució Problemes Rcmdr MacO Pujat 3/02/2020 09:41  Indestap Pujat 3/02/2020 09:42  DietasDieabetes Pujat 3/02/2020 09:43			

Destresa:

- a) Conèixer el tipus de variable que anem a estudiar, i la seua distribució.
Estudiarem variables categòriques dicotòmiques, que són aquelles que sols poden agafar dos possibles resultats. La categoria d'interés s'anomena èxit i l'altra categoria fracàs ($X = \text{èxit}$ o $X = \text{fracàs}$)

Tota variable categòrica dicotòmica segueix un comportament Bernoulli, $X \sim Be(\pi)$, on π representa la probabilitat d'èxit en la població. Notar que la probabilitat de fracàs serà $1 - \pi$.

El nostre objectiu en aquest cas és estudiar $\pi = \Pr(\text{èxit})$ i podem estar interessats en calcular l'estimació puntual de π , l'interval de confiança per a π i contrast d'hipòtesi sobre π .

Exemple: Consum d'aliments amb soia

Per estimar el percentatge d'individus de certa població que consumeixen productes amb soia s'ha elegit una mostra de grandària 100; s'ha observat que 18 dels individus de la mostra eren consumidors d'aquests productes, la resta no ho eren.
Què podem dir sobre el percentatge de consumidors en la població? Podria ser del 30%?

El paràmetre d'interès és la proporció d'individus en la població que consumeixen productes de soia: π

La nostra variable X mesura si un individu consumeix o no soia, de tal forma que serà èxit si sí consumeix i fracàs si no consumeix.

Si tenim una mostra de grandària n (en el nostre cas una mostra de $n = 100$), estem repetint l'experiment anterior n voltes.

Quina és la distribució d'una variable r que mesura el nombre d'èxits d'un total de n ?

Solució: $r \sim Bi(n = 100, \pi)$

- b) Estimació puntual de π (pàgina 11 document teoria), valor que s'estima a partir de la mostra. L'estimació puntual és la probabilitat d'èxit en la mostra, és a dir:

$$\hat{\pi} = \frac{\text{nombre d'èxits de la mostra}}{\text{grandària mostral}}$$

A la funció de les dades que utilitzem per estimar, $\hat{\pi} = r / n$, se li anomena **estimador puntual**. El valor concret observat amb unes dades, $\hat{\pi} = 18/100 = 0.18$ (en l'exemple de consum de soia), és l'**estimació puntual** de π .

- c) Interval de confiança per a π (pàgina 12 i 13 document teoria), conjunt de valors possibles per a π .

La distribució d'una variable r que mesura el nombre d'èxits d'un total de n és binomial amb n proves i probabilitat d'èxit π .

En l'exemple:

```
> binom.test(rbind(.Table), alternative='two.sided', p=.3, conf.level=.95)
Exact binomial test

data:  rbind(.Table)
number of successes = 18, number of trials = 100, p-value = 0.00849
alternative hypothesis: true probability of success is not equal to 0.3
95 percent confidence interval:
 0.1103112 0.2694771
sample estimates:
probability of success
              0.18
```

Interval de confiança

- d) Contrast d'hipòtesi sobre π : (observada una única mostra)

$H_0: \pi = p$	$H_0: \pi = p$	$H_0: \pi = p$
$H_1: \pi \neq p$	$H_1: \pi > p$	$H_1: \pi < p$

La utilització del p-valor (regla de decisió) sempre és la mateixa, sota l'assumpció de que hipòtesi nul·la és certa:

Si $p\text{-valor} \leq \alpha$, rebutgem la hipòtesi nul·la

Si $p\text{-valor} > \alpha$, no podem rebutjar la hipòtesi nul·la

En l'exemple, Estem interessats en una pregunta concreta sobre el paràmetre d'interès: Podria ser cert que $\pi = 0.3$?

El contrast d'hipòtesi serveix per a resoldre aquest tipus de qüestions. Per a respondre a eixes qüestions ens plantegem dues hipòtesis contraposades. En aquest cas les dues hipòtesis seran:

Hipòtesis Nul·la $H_0: \pi = 0.3$

Hipòtesis Alternativa $H_A: \pi \neq 0.3$

Si les dades observades, en aquest cas $r = 18$ amb $n = 100$, no són compatibles amb H_0 rebutjarem aquesta hipòtesi i optarem per H_A . El p-valor és $0.00849 < 0.05$. Rebutgem H_0 .

```
> binom.test(rbind(.Table), alternative='two.sided', p=.3, conf.level=.95)
```

Exact binomial test

```
data:  rbind(.Table)
number of successes = 18, number of trials = 100, p-value = 0.00849
alternative hypothesis: true probability of success is not equal to 0.3
95 percent confidence interval:
 0.1103112 0.2694771
sample estimates:
probability of success
      0.18
```

$H_A: \pi \neq 0.3$

Els tres càlculs (apartats b), c) y d)) els realitzem mitjançant el test Binomial (`binom.test`). Si volem calcular un interval de confiança hem de comprovar `conf.level` = confiança de l'interval que volem i `alternative` = 'two.sided'. Si estem interessats en resoldre un contrast d'hipòtesi comprovem que `p` = π_0 (valor que volem contrastar) i `alternative` = 'two.sided', o 'less', o 'greater', segons siga la hipòtesi alternativa.

Destresa	Pàgina del document on es troba la resposta	Exemples en el document
a)	5 (repassar també Unitat 1. Tema 3 Probabilitat.)	
b)	10 i 11	14 i 15
c)	12 i 13	14 i 15
d)	3 a 9 (bilateral) i 16 a 19 (unilateral)	3 a 9 i 16 a 19

Activitat

Recordeu que podeu realitzar totes les activitats esmenades en l'apartat de material de l'aula virtual.

2. Anàlisi d'una variable categòrica en tres o més categories

Material de l'aulavirtual:

Teoria	Activitats amb R	Activitats sense R	Indestap
Secció: Unitat 2. Anàlisi estadística d'una variable. Pdf: Tema 7. Inferència sobre una proporció. Pàgines: 1 a 19	Document: Exercicis unitat 2. Dades: Dades unitat 3. Exercicis: 1,2,16,17,18 i 19.	Secció: Problemes Carpeta: Categòriques.	Document: Indestap Secció del document: 2.6- Subsecció:2.6b
Unitat 2. Anàlisi estadística d'una variable  Tema 4 Inferència sobre una mitjana poblacional Pujat 20/02/2020 10:30  Una mostra_Unitat 2  Tema 7 Inferència sobre una proporció Pujat 27/01/2020 10:16  Exercicis unitat 2 Pujat 27/01/2020 10:17  dades unitat 2 Pujat 27/01/2020 10:17		Problemes  Probabilitat  Una mostra  Dues mostres  k mostres  Categòriques	Indestap  Instal·lació Rcmdr Pujat 3/02/2020 09:40  Instal·lació Rcmdr en MacOS Pujat 3/02/2020 09:40  Resolució Problemes Rcmdr Mac Pujat 3/02/2020 09:41  Indestap Pujat 3/02/2020 09:42  DietasDiabetes Pujat 3/02/2020 09:43

Destresa:

- a) Conèixer el tipus de variable que anem a estudiar.
És una variable categòrica amb k categories, (k és tres o més). Les categories $C_1, C_2, \dots, C_k$. Podem considerar la població dividida en categories. Cada categoria representa una proporció del total.

Exemple:

Les llavors de lli són un potent anti-cancerigen, contenen una gran quantitat de fibra i àcids grassos omega 3 i tenen un efecte beneficiós en moltes malalties: inflamatòries, cardiovasculars, al·lèrgies... S'ha realitzat un estudi sobre una mutació de llavors de lli amb la que s'espera obtenir oli per un us medicinal. La quantitat d'àcid palmític en la llavor {baix, mitjà, alt} i el color de la llavor {marró, jaspiada} eren factors importants de la investigació. D'acord amb les lleis de Mendel, les sis combinacions possibles haurien d'ocórrer en una proporció 3:6:3:1:2:1, és a dir sumen 16.

La variable indica el tipus de llavors de lli que poden ser de 6 formes: {marró-baix, marró-mitjà, marró-alt, jaspiat-baix, jaspiat-mitjà, jaspiat-baix}

- b) L'únic estudi que anem a realitzar en aquest cas és un contrast d'hipòtesi de la forma

$$H_0: \pi_1 = \pi_1^0, \dots, \pi_k = \pi_k^0$$

$$H_1: \text{Alguna de les igualtats anteriors no es compleix}$$

On π_i^0 = probabilitat de pertànyer a la categoria i , $i=1, \dots, k$

Per a resoldre este contrast utilitzarem el test de la Xi-quadrat (**chisq.test**). En les opcions del test hem de comprobar que el vector de valors observats es correspon amb el nombre de voltes que

hem observat cada categoria en la mostra i que el vector de probabilitats coincideix en els valors que volem contrastar $\pi_1^0, \dots, \pi_k^0$.

En l'exemple, Tenim una mostra aleatòria de grandària 72 d'una població de llavors de lli que pertanyen a una de les 6 categories en les que hem dividit la població: $C_1, C_2, \dots, C_6$. Volem contrastar una hipòtesis nul·la que especifique les probabilitats poblacionals d'aquestes categories.

En particular, contrastar si les probabilitats poblacionals venen donades per les lleis de Mendel: $\pi_1^0 = 3/16, \pi_2^0 = 6/16, \pi_3^0 = 3/16, \pi_4^0 = 1/16, \pi_5^0 = 2/16$ i $\pi_6^0 = 1/16$.

$H_0: \pi_1^0 = 3/16, \pi_2^0 = 6/16, \pi_3^0 = 3/16, \pi_4^0 = 1/16, \pi_5^0 = 2/16$ i $\pi_6^0 = 1/16$

H_A : alguna proporció és distinta.

Per calcular l'estadístic de la Chi-quadrada cal determinar la proporció esperada de cada categoria: $E_i = 72 * \pi_i^0$, per a cada $i=1,2,\dots,6$.

color	Nivell àcid	Observat (mostra)	Proporció esperada	Total*Proporció esperada=	Esperat
Marró	baix	15	3/16	72*3/16=	13.5
Marró	mitjà	26	6/16	72*6/16=	27
marró	alt	15	3/16	72*3/16=	13.5
jaspiat	baix	0	1/16	72*1/16=	4.5
jaspiat	mitjà	8	2/16	72*2/16=	9
jaspiat	alt	8	1/16	72*1/16=	4.5
TOTAL		72			72

Calculem el valor de l'estadístic, que en aquest cas es distribueix com una variable χ^2 amb 5 graus de llibertat = nombre de categories – 1.

$$\chi_s^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \frac{(15 - 13.5)^2}{13.5} + \frac{(26 - 27)^2}{27} + \dots + \frac{(8 - 4.5)^2}{4.5} = 7.71$$

Si no tenim les dades guardades en un fitxer, en la finestra d'instruccions podem escriure la següent instrucció, on indiquem els valors observats en cada categoria i les probabilitats a contrastar

```
> chisq.test(c(15,26,15,0,8,8),p=c(3/16,6/16,3/16,1/16,2/16,1/16))
```

Chi-squared test for given probabilities

data: c(15, 26, 15, 0, 8, 8)

X-squared = 7.7037, df = 5, p-value = 0.1733

L'ordre és important

El p-valor és $0.1733 > 0.05$. No rebutgem H_0 . No hi ha evidència significativa de que les dades siguin inconsistentes amb el model mendelià.

Els valors esperats de l'exemple de les llavors de lli també es poden obtenir amb el R:

```
> chisq.test(c(15,26,15,0,8,8), p=c(3/16,6/16,3/16,1/16,2/16,1/16))$expected
```

[1] 13.5 27.0 13.5 4.5 9.0 4.5

- c) Validesa del mètode. El mètode serà vàlid si tots els valors esperats són majors o igual que 5.

Destresa	Pàgina del document on es troba la resposta	Exemples en el document
a)	20	20, 24
b)	21 al 32	24 al 29

Activitat

Recordeu que podeu realitzar totes les activitats esmenades en l'apartat de material de l'aula virtual.

3. Exemple variable categòrica dicotòmica

Exemple:

Un nou medicament es considerarà efectiu en la curació de certa malaltia si hi ha evidència que la curació és major que amb els tractaments actuals, que s'estima que és del 30% en les persones que pateixen la malaltia. S'administra aquest medicament a 50 malalts i no es curen 14. Tenim evidència de l'efectivitat del medicament?

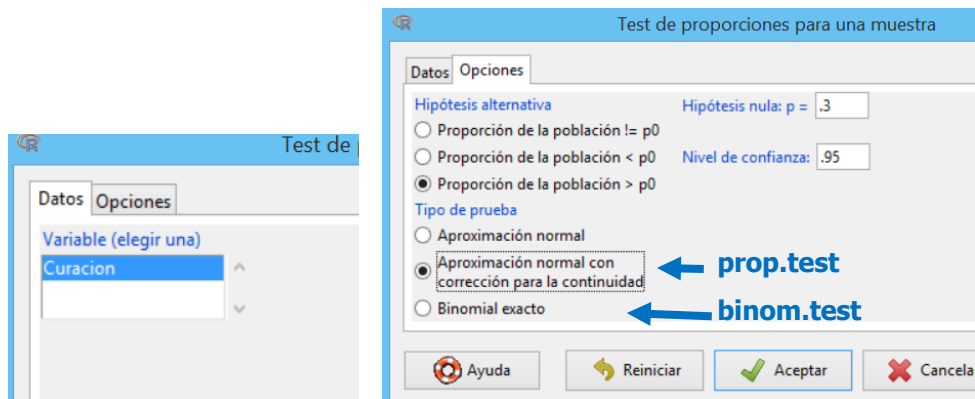
L'única forma en què podem donar resposta a aquesta qüestió és mitjançant un contrast d'hipòtesis respecte de la probabilitat de curació:

H_0 : El medicament no és eficaç ($p = 0.3$)

H_A : El medicament és eficaç ($p > 0.3$)

que podem resoldre utilitzant el menú:

Estadístics -> Proporcions -> Test de proporcions per a una mostra



Elegim el **binom.test** perquè **sempre és vàlid**, mentre que el **prop.test**, en basar-se en l'aproximació normal de la binomial, **requereix que $np_0 \geq 5$ i $n(1-p_0) \geq 5$** .

```
Frequency counts (test is for first level):
Curacion
No SI
14 36
```

```
Exact binomial test
```

```
data: rbind(.Table)
number of successes = 14, number of trials = 50, p-value = 0.6721
alternative hypothesis: true probability of success is greater than 0.3
95 percent confidence interval:
 0.1779032 1.0000000
sample estimates:
probability of success
              0.28
```

Cura! Com es diu en la primera línia de la eixida, en tractar-se d'una variable dicotòmica s'ha calculat la proporció de la primera categoria del fitxer (en aquest cas, el No), Per tant, la sortida correspon al contrast:

$$H_0: p_{No} = 0.3$$

$$H_A: p_{No} > 0.3$$

que no és el que volíem resoldre. Perquè l'eixida siga correcta, haurem d'utilitzar primer el menú **Dades -> Modificar variables... -> Reordenar nivells de factor per indicar que la primera categoria siga el Si.**

Noteu que al fer un contrast direccional ('greater' = **major que**) ens retorna un interval de confiança de la forma [LI, 1], mentre que per als direccionals ('less' = **menor que**) ens donaria un interval [0, LS], respectivament, on LI és el límit inferior i LS és el límit superior.

També tenim la possibilitat d'utilitzar directament el test de proporcions per calcular intervals de confiança al 95% mitjançant la instrucció següent:

binom.test(num_èxits, num_total)

on **num_èxits** és el nombre d'èxits observats i **num_total** és la grandària de la mostra. Així, no seria necessari crear un conjunt de dades 'adhoc'. Ara bé, si volem calcular un interval de confiança amb un altre nivell de confiança, per exemple al 90% seria necessari afegir a la instrucció l'opció: **conf.level=0.90**. En aquest cas obtindríem:

```
> binom.test( 36, 50, conf.level=.90 )

Exact binomial test

data: 36 and 50
number of successes = 36, number of trials = 50, p-value = 0.002602
alternative hypothesis: true probability of success is not equal to 0.5
90 percent confidence interval:
 0.5973826 0.8220968
sample estimates:
probability of success
      0.72
```

que ens proporciona l'interval (0.5973826,0.8220968) i comprovem que s'ha realitzat, per defecte, el contrast d'hipòtesi bilateral amb una probabilitat 0.5 d'èxit en la població:

$$H_0: p = 0.5$$

$$H_A: p \neq 0.5$$

Per obtenir els resultats que volem per a l'exemple actual (amb un valor de p concret i una alternativa direccional):

$$H_0: p = 0.3$$

$$H_A: p > 0.3$$

hauríem d'incloure també aquestes dues noves opcions:

binom.test(num_èxits, num_total, p=0.3, alternative="greater")

```
> binom.test( 36, 50, p=0.3, alternative='greater' )

Exact binomial test

data: 36 and 50
number of successes = 36, number of trials = 50, p-value = 1.136e-09
alternative hypothesis: true probability of success is greater than 0.3
95 percent confidence interval:
 0.5973826 1.0000000
sample estimates:
probability of success
      0.72
```

El **p-valor** és **1.136 10⁻⁹**, per la qual cosa rebutgem la hipòtesi nul·la i concloem que **hi ha evidència estadística que el nou medicament és millor que els tractaments actuals**. A més a més, l'estimació puntual d'eficàcia és 0.72.

4. Exemple variable categòrica amb k o més categories

Exemple: Es denomina epistaxis a les hemorràgies amb origen en les fosses nasals. Normalment es tracta d'un procés que remiteix espontàniament. La primera mesura és baixar-li el cap al malalt i indicar-li que es pressioni fortament el nas amb dos dits durant cinc minuts cronometrats. Si després de cinc minuts de compressió persisteix el sagnat, l'epistaxi ha de ser avaluada per un metge. Quan els sagnats són recurrents poden causar anèmia a causa d'una deficiència de ferro.

En un estudi retrospectiu es van analitzar les dades de 200 malalts adults amb diagnòstic d'epistaxis ingressats en un hospital de referència entre els anys 1997 i 2004 tots dos inclosos (8 anys)

Els responsables de l'estudi van classificar l'anèmia en tres nivells segons la gravetat (lleu, moderada i severa). Del total de pacients, 46 malalts van tenir anèmia, i va haver-hi:

Grau d'anèmia	Nombre de malalts
Lleu	12
Moderada	19
Severa	15

Podem assumir que els tres graus d'anèmia tenen igual probabilitat d'ocórrer?

Com tenim el resum de freqüències absolutes observades és més senzill executar directament el test de la Bondat d'Ajust mitjançant la instrucció:

```
chisq.test( c(O1,O2,...,Ok), p=c(p1,p2,...,pk) )
```

on O_i són les freqüències observades de les k categories i i p_i les probabilitats de cadascuna d'elles. En aquest cas: (12,19,15) i (1/3,1/3,1/3), respectivament.

Si volem obtenir la taula de valors esperats, hem d'usar:

```
chisq.test( c(O1,O2,...,Ok), p=c(p1,p2,...,pk) )$expected
```

En el nostre exemple, el contrast de bondat d'ajust és:

$H_0: p_{\text{lleu}} = 1/3, p_{\text{moderada}} = 1/3, p_{\text{severa}} = 1/3$

H_A : No totes les igualtats es compleixen.

```
> chisq.test( c(12,19,15), p=c(1/3,1/3,1/3) )
```

```
Chi-squared test for given probabilities
```

```
data:  c(12, 19, 15)
```

```
X-squared = 1.6087, df = 2, p-value = 0.4474
```

Com el **p-valor** és **0.01685**, rebutgem la hipòtesi nul·la d'igualtat, per **la qual cosa hi ha evidència estadística que tots els graus d'anèmia no tenen la mateixa probabilitat d'ocórrer**.

Recordeu que si tots els valors esperats són majors o iguals a 5, el mètode és vàlid. En aquest cas:

```
> chisq.test( c(12,19,15), p=c(1/3,1/3,1/3) )$expected
[1] 15.33333 15.33333 15.33333
```