Distribucions de probabilitat i inferència estadística amb R-Commander

Daniel Liviano Solís
Maria Pujol Jover

PID_00208268

UOC
Universitat Oberta de Catalunya

Índex

Introducció

En aquest mòdul s'aborda tota la part de la inferència estadística utilitzant R-Commander. En un primer capítol s'aprendrà a treballar amb variables discretes, les seves distribucions associades i el càlcul de les seves probabilitats. Posteriorment, es farà el mateix amb variables contínues. D'aquesta manera s'estudiaran les distribucions principals de probabilitat tant discretes com contínues que s'utilitzen en la majoria de les assignatures d'estadística de la UOC.

En concret s'aprofundirà en les distribucions següents de probabilitat discretes:

- la distribució binomial
- la distribució geomètrica
- la distribució hipergeomètrica
- la distribució de Poisson

I en les distribucions de probabilitat contínues que s'esmenten a continuació:

- la distribució uniforme
- la distribució exponencial
- la distribució normal
- la distribució t de Student
- la distribució khi quadrat
- la distribució F de Snedecor

Com es podrà veure al final del primer capítol, gairebé totes les distribucions es podran aproximar a una de normal gràcies al teorema del límit central. Aquest aspecte serà molt important per a poder obtenir les distribucions de la mitjana mostral i de la proporció a partir d'una població encara que aquesta no es distribueixi necessàriament segons una de normal.

El segon capítol es dedicarà íntegrament a la inferència estadística sobre els paràmetres principals de les distribucions: la mitjana poblacional (μ), la variància (σ^2) i la proporció (π).

Objectius

L'estudiant ha de ser capaç d'utilitzar R-Commander per al següent:

1. Calcular els valors crítics i les probabilitats associades a una distribució de probabilitat discreta.

2. Calcular els valors crítics i les probabilitats associades a una distribució de probabilitat contínua.

3. Calcular intervals de confiança i contrastar hipòtesis (unilaterals i bilaterals) per a la mitjana poblacional (μ) amb variància poblacional (σ^2) coneguda.

4. Calcular intervals de confiança i contrastar hipòtesis (unilaterals i bilaterals) per a la mitjana poblacional (μ) amb variància poblacional (σ^2) desconeguda.

5. Calcular intervals de confiança i contrastar hipòtesis (unilaterals i bilaterals) per a la variància poblacional (σ^2).

6. Calcular intervals de confiança i contrastar hipòtesis (unilaterals i bilaterals) per a la proporció poblacional (π).

7. Calcular intervals de confiança i contrastar hipòtesis (unilaterals i bilaterals) per a la diferència de mitjanes poblacionals (μ_1 i μ_2) amb variàncies (σ_1^2 i σ_2^2) conegudes.

8. Calcular intervals de confiança i contrastar hipòtesis (unilaterals i bilaterals) per a la diferència de mitjanes poblacionals (μ_1 i μ_2) amb variàncies desconegudes però iguals (σ^2).

9. Calcular intervals de confiança i contrastar hipòtesis (unilaterals i bilaterals) per al quocient de variàncies poblacionals (σ_1^2 i σ_2^2).

10. Calcular intervals de confiança i contrastar hipòtesis (unilaterals i bilaterals) per a la diferència de proporcions poblacionals (π_1 i π_2).

1. Distribucions de probabilitat

1.1. Introducció

Una part fonamental de l'estadística és l'anàlisi de variables aleatòries, les quals es poden distribuir segons diferents funcions de probabilitat. R-Commander ofereix la possibilitat d'analitzar aquestes distribucions de probabilitat de diverses maneres. Específicament, per a cada distribució de probabilitat hi ha quatre opcions de càlcul:

Els quantils ens seran molt útils per a calcular els valors crítics de les distribucions.

1) Quantils. La funció quantil és la inversa de la funció de distribució. Això és, un quantil serà el valor x tal que $P(X \leq x) = p$, essent p la probabilitat que com a tal es troba en l'interval $[0, 1]$.

2) Probabilitats. La probabilitat serà el valor p tal que $P(X \leq x) = p$, és a dir, és la funció inversa a la funció quantil.

La funció de quantia únicament existeix per a les variables aleatòries discretes, i la de densitat per a les contínues. A més, l'àrea sota la corba de totes les funcions de distribució sempre serà la unitat, ja que recull tota la probabilitat acumulada d'una variable aleatòria.

3) Gràfiques. Cada distribució de probabilitat ofereix la possibilitat de calcular els gràfics de la seva funció de densitat o de quantia associats, com també la seva funció de distribució acumulada.

4) Mostres aleatòries. Per a cada distribució es poden generar mostres de n valors aleatoris, a partir d'uns paràmetres inicials determinats.

És important adonar-se que per a variables contínues no té sentit calcular $P(X = x) = p$. Per tant, per a aquest tipus de variables és més correcte calcular $P(X < x) = p$ per a obtenir la probabilitat de la cua de l'esquerra i $P(X > x) = p$ per a la de la dreta.

Respecte al càlcul de quantils i probabilitats, cal especificar que R ofereix la possibilitat de fer càlculs per a totes dues cues (esquerra i dreta), és a dir, tant $P(X \leq x) = p$ com $P(X > x) = p$. És fonamental subratllar el símbol $\leq$ per a la cua esquerra i el símbol $>$ per a la cua dreta, per a evitar confusions en el càlcul de probabilitats. A més, fora d'R-Commander, també és possible calcular amb funcions en codi R el valor de la **funció de densitat** (per a variables contínues) i de la **funció de quantia** (per a variables discretes).

A continuació oferim una relació de les funcions en codi R de les distribucions de probabilitat incorporades en la distribució bàsica d'R i que estan disponibles en el menú desplegable d'R-Commander. En la taula 1 es recullen les distribucions de probabilitat discretes:

Taula 1. Distribucions de probabilitat discretes

Distribució	Funció de quantia	Probabilitats	Quantils	Mostra aleatòria
Binomial	`dbinom`	`pbinom`	`qbinom`	`rbinom`
Poisson	`dpois`	`ppois`	`qpois`	`rpois`
Geomètrica	`dgeom`	`pgeom`	`qgeom`	`rgeom`
Hipergeomètrica	`dhyper`	`phyper`	`qhyper`	`rhyper`
Binomial negativa	`dnbinom`	`pnbinom`	`qnbinom`	`rnbinom`

De la mateixa manera, en la taula 2 s'ofereixen les distribucions de probabilitat contínues i les funcions d'R associades:

Taula 2. Distribucions de probabilitat contínues

Distribució	Funció de densitat	Probabilitats	Quantils	Mostra aleatòria
Normal	dnorm	pnorm	qnorm	rnorm
t Student	dt	pt	qt	rt
khi quadrat	dchisq	pchisq	qchisq	rchisq
F Snedecor	df	pf	qf	rf
Exponencial	dexp	pexp	qexp	rexp
Uniforme	dunif	punif	qunif	runif
Beta	dbeta	pbeta	qbeta	rbeta
Cauchy	dcauchy	pcauchy	qcauchy	rcauchy
Logística	dlogis	plogis	qlogis	rlogis
Log normal	dlnorm	plnorm	qlnorm	rlnorm
Gamma	dgamma	pgamma	qgamma	rgamma
Weibull	dweibull	pweibull	qweibull	rweibull

A continuació oferim diversos exemples pràctics d'anàlisis de variables aleatòries amb diferents distribucions.

1.2. Distribucions de probabilitat discretes

1.2.1. Distribució binomial

La distribució binomial és una de les distribucions principals de probabilitat discretes. Específicament, aquesta mesura el nombre d'èxits en una seqüència de n assajos de Bernoulli independents entre ells, i amb una probabilitat p d'èxit entre els assajos. Això és, definim formalment una variable aleatòria X que es distribueix com una binomial amb paràmetres n i p:

La distribució de Bernoulli

Les variables aleatòries dicotòmiques que prenen únicament el valor 1 quan un assaig o esdeveniment ha tingut lloc amb èxit, i un valor 0 en cas contrari, es distribueixen sota una distribució de Bernoulli. Per tant, la distribució binomial serà una generalització d'una distribució de Bernoulli.

$$X \sim B(n, p)$$

La funció de probabilitat té la forma següent:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Recordeu que per a variables binomials $E(X) = np$ i $V(X) = np(1-p)$.

Essent $\binom{n}{x}$ el quocient binomial, que defineix el nombre de subconjunts de x elements escollits d'un conjunt amb n elements. Aquest pren l'expressió següent:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

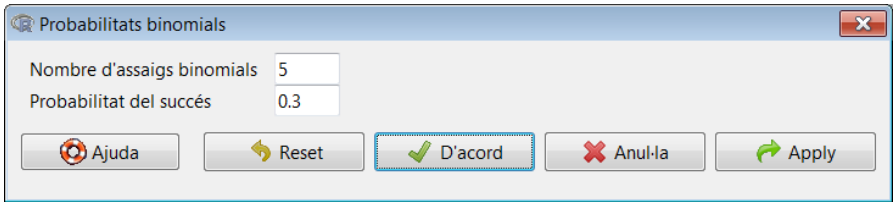
Un quocient binomial no és res més que les combinacions de n elements agafats de x en x.

Per veure com treballem amb variables aleatòries distribuïdes com una binomial, prendrem un exemple fictici. Sabem que la probabilitat que té un jugador de bàsquet determinat d'encistellar un triple és del 30%, és a dir, $p = 0,3$, i volem calcular les

probabilitats que encistelli $1, 2, \ldots$ fins a n triples, essent n el nombre de llançaments (assajos). Si suposem que aquest jugador llança 5 vegades, $n = 5$, el nostre càlcul amb R-Commander el farem de la manera següent:

Distribucions / Distribucions discretes / Distribució binomial / Probabilitats binomials

Veurem que apareixerà una finestra en la qual introduirem la probabilitat d'èxit i el nombre d'assajos:



En prémer *D'acord*, obtindrem el resultat següent:

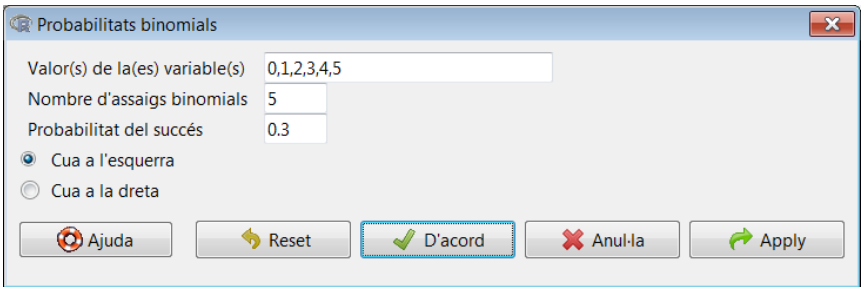
```
> .Table <- data.frame(Pr=dbinom(0:5, size=5, prob=0.3))

> rownames(.Table) <- 0:5

> .Table
       Pr
0 0.16807
1 0.36015
2 0.30870
3 0.13230
4 0.02835
5 0.00243
```

Però, com s'interpreta aquest resultat? Si definim X com el nombre d'èxits, tenim que $P(X = 0) = 0,168$. Això no és més que la probabilitat que en 5 assajos el nostre jugador no encerti cap triple. En l'altre extrem, la probabilitat que encerti 5 triples en 5 intents és de $P(X = 5) = 0,002$. És important observar que la suma d'aquestes cinc probabilitats és igual a u.

Si volguéssim calcular les probabilitats acumulades (també anomenades *de la cua*), la ruta que hauríem de seguir en el menú desplegable seria la següent:

Distribucions / Distribucions discretes / Distribució binomial / Probabilitats binomials acumulades

El quadre de diàleg que ens apareixerà és el següent, on en el valor(s) de la variable hem d'introduir els valors de X que ens interessen (en aquest cas els volem tots), i en les dues opcions següents introduirem el nombre d'assajos i la probabilitat d'èxit, respectivament. A més, hem d'especificar la cua de la distribució que ens interessa. Si activem *Cua a l'esquerra*, estem calculant $P(X \leq x)$:



Al contrari, si volguéssim calcular $P(X > x)$, hauríem, d'activar l'opció *Cua a la dreta*. Seleccionant primer *Cua a l'esquerra* i després *Cua a la dreta* obtenim els resultats següents, respectivament:

```
> pbinom(c(0,1,2,3,4,5), size=5, prob=0.3,
+ lower.tail=TRUE)
[1] 0.16807 0.52822 0.83692 0.96922 0.99757 1.00000

> pbinom(c(0,1,2,3,4,5), size=5, prob=0.3,
+ lower.tail=FALSE)
[1] 0.83193 0.47178 0.16308 0.03078 0.00243 0.00000
```

Fixeu-vos que, si sumem les probabilitats obtingudes per columnes, la suma de cada parell de valors sempre és u. Això és perquè, per definició, sempre es compleix la igualtat següent:

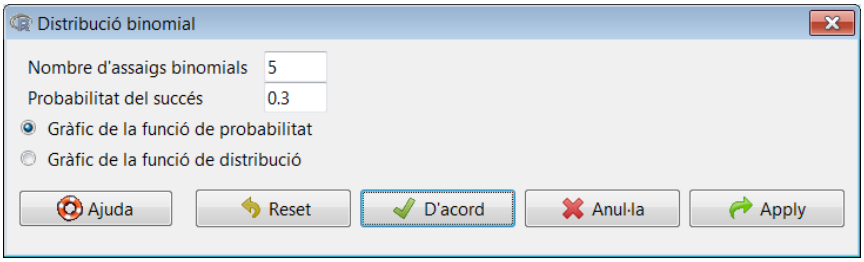
$$P(X \leq x) + P(X > x) = 1.$$

En l'exemple anterior, quina seria la probabilitat d'encistellar més de tres triples? La resposta és $P(X > 3) = 0,03078$. I la probabilitat d'encistellar menys de dos triples? La resposta és $P(X \leq 1) = 0,52822$.

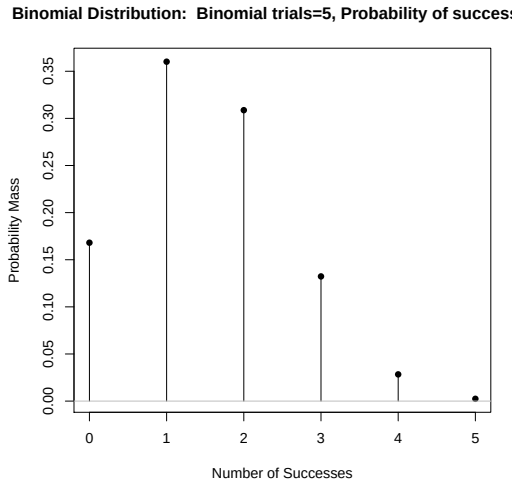
Un últim aspecte que veurem de la distribució binomial és el resultat gràfic de les probabilitats calculades. Hem d'accedir a la ruta següent:

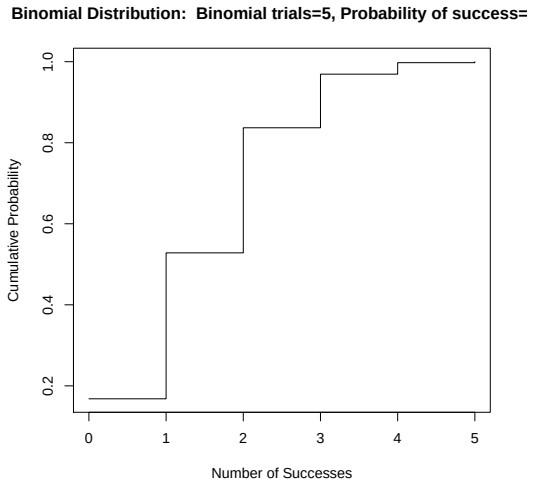
Distribucions / Distribucions discretes / Distribució binomial / Traça una distribució binomial

En el quadre de diàleg resultant, R-Commander ens dóna l'opció d'obtenir dos tipus de gràfics, la *funció de probabilitat* i la *funció de distribució*:



Els gràfics obtinguts sempre es mostraran en la consola d'R, de manera que, per a accedir-hi, haurem d'anar a la barra de tasques i canviar la finestra activa a la d'R. Activant primer la *Gràfic de la funció de probabilitat* i després la *Gràfic de la funció de distribució*, obtindríem els gràfics següents (per separat):





Per acabar, destacarem que en el menú que es desplega en seleccionar *Distribució binomial* apareixen més opcions disponibles, com per exemple el càlcul dels quantils i l'obtenció de mostres aleatòries.

Més endavant ja veurem la utilitat d'aquestes dues opcions.

1.2.2. Distribució geomètrica

La distribució geomètrica està associada a la distribució binomial que acabem de veure. Específicament, aquesta distribució descriu la probabilitat del nombre x d'assajos de Bernoulli necessaris per a obtenir un èxit. Si la probabilitat d'èxit en cada assaig és p, aleshores la probabilitat que x assajos siguin necessaris per a obtenir un èxit queda definida per l'expressió següent:

$$P(X = x) = (1 - p)^{x-1} p.$$

Per a $x = 1, 2, 3, \ldots$, la seqüència de probabilitats que s'obté es denomina *progressió geomètrica*. Així doncs, podem definir la funció de distribució de la manera següent:

$$P(X \leq x) = F(x) = 1 - (1 - 0,3)^x.$$

Reprenent l'exemple anterior del jugador de bàsquet que intenta encistellar triples, en aquest cas podem definir una nova variable aleatòria X, que serà el número de l'intent en què el jugador encistella el primer triple, és a dir, el nombre d'assajos necessaris fins a encistellar. Formalment, diem que X segueix una distribució geomètrica amb paràmetre $p = 0,3$, que es correspon amb la probabilitat d'encistellar un triple, com

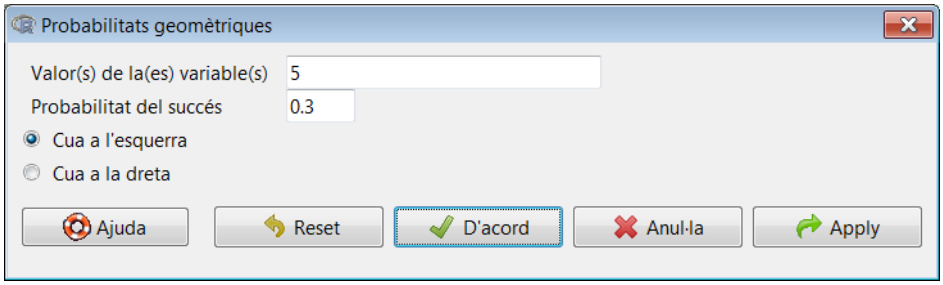
hem vist en l'exemple anterior. Si volem saber la probabilitat que encistelli el primer triple dins dels 6 primers llançaments a cistella, hem de fer el càlcul següent:

$$P(Y \leq 6) = F(6) = 1 - (1 - 0,3)^6 = 0,8823.$$

Per a fer aquest càlcul amb R-Commander, haurem de seguir aquesta ruta:

Distribucions / Distribucions discretes / Distribució geomètrica / Probabilitats geomètriques acumulades

Obtindrem el quadre de diàleg següent, en què introduirem les dades del nostre exemple:



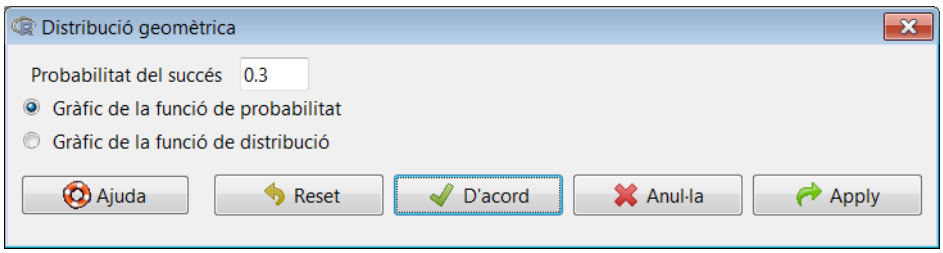
Cal tenir en compte dos aspectes. El primer és que, en l'espai *Valor(s) de la(es) variable(s)*, hem d'introduir **el nombre de fallades** abans del primer encert, això és, $x - 1$. En el nostre exemple, això és $6 - 1 = 5$. El segon aspecte que hem de tenir en compte és que, com que calculem $P(Y \leq x)$, s'ha d'activar l'opció *Cua a l'esquerra*. En la finestra de resultats obtenim el següent:

```
> pgeom(c(5), prob=0.3, lower.tail=TRUE)
[1] 0.882351
```

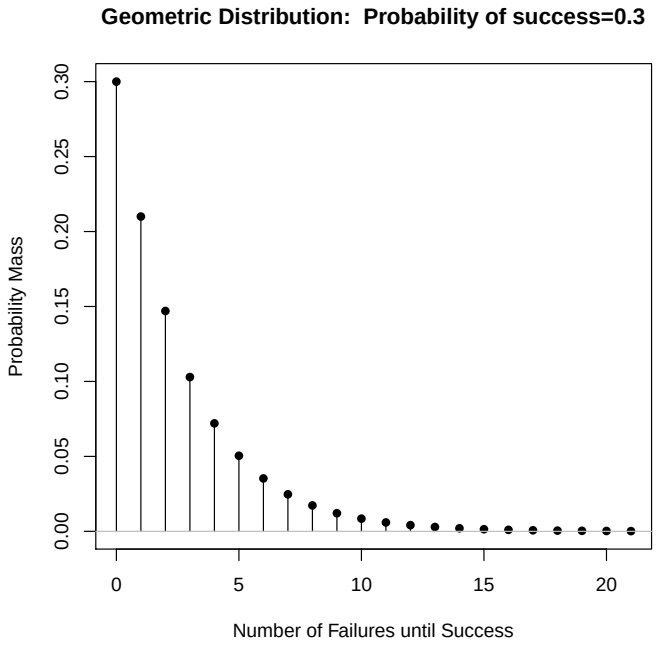
Quant a la distribució geomètrica, és molt interessant veure'n el gràfic de la funció de probabilitat. Això es fa accedint a aquesta ruta del menú desplegable:

Distribucions / Distribucions discretes / Distribució geomètrica / Traça una distribució geomètrica

Ens apareixerà el quadre de diàleg que es mostra a continuació i en el qual seleccionem la funció de probabilitat:



La interpretació del gràfic resultant és molt intuïtiva: en l'eix horitzontal, mostra el nombre de fallades fins a encistellar el primer triple, i en l'eix vertical la probabilitat associada. El primer valor coincideix amb $p = 0,3$, que és la probabilitat d'encistellar un triple, i d'aquí va decreixent fins a zero a mesura que incrementa el nombre d'intents.



1.2.3. Distribució hipergeomètrica

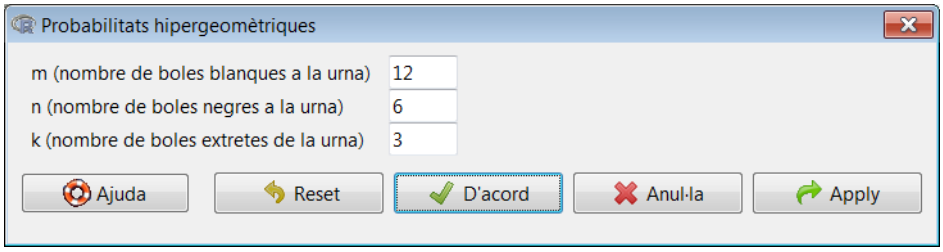
Fins ara hem considerat que les proves o observacions fetes eren independents, però sovint es dóna aquesta condició i l'ús del model binomial pot resultar equivocat quan les poblacions estudiades són petites. Específicament, suposem que tenim una mostra de N boles, de les quals N_1 són verdes i N_2 són vermelles, de manera que $N_1 + N_2 = N$. Davant una extracció de n boles d'aquest conjunt (sense retorn), la variable X serà el nombre de boles verdes obtingudes. Aleshores diem que X segueix una distribució hipergeomètrica tal que:

$$P(X = x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}}.$$

Suposem un exemple empíric. Un estudiant d'oposicions ha preparat dotze de les divuit lliçons de què consta un programa. L'examen consisteix en tres lliçons escollides aleatòriament. Quina probabilitat té l'estudiant de saber els tres temes? Sabem que $N_1 = 12$, $N_2 = 6$ i $N = 18$. A més, l'extracció és $n = 3$. Anem al menú desplegable i accedim a la ruta següent:

Distribucions / Distribucions discretes / Distribució hipergeomètrica / Probabilitats hipergeomètriques

Ens apareixerà el quadre de diàleg següent en què introduirem la informació del nostre exemple:



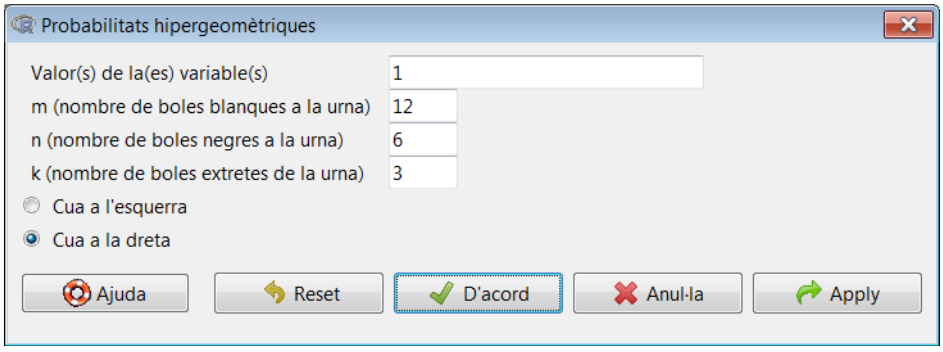
El resultat obtingut és el següent:

```
> .Table <- data.frame(Pr=dhyper(0:3 ,  m = 12 ,  n = 6
   ,  k = 3))

> rownames(.Table) <- 0:3

> .Table
         Pr
0 0.0245098
1 0.2205882
2 0.4852941
3 0.2696078
```

Com veiem, la probabilitat que els tres temes que ha estudiat entrin en l'examen és del 26, 9%. La probabilitat més alta és $P(X = 2) = 0,485$, és a dir, que hagi estudiat dos dels tres temes. Si, per exemple, ens preguntem quina és la probabilitat que sàpiga com a mínim dos temes ($P(X \geq 2)$), necessitem la probabilitat acumulada:

Distribucions / Distribucions discretes / Distribució hipergeomètrica / Probabilitats hipergeomètriques acumulades

És important destacar que en aquest cas necessitem la cua dreta. R calcula $P(X > x)$, de manera que haurem d'introduir en el quadre de diàleg el valor $x = 1$, ja que $P(X > 1) = P(X \geq 2)$:



D'aquesta manera, obtenim el resultat següent:

```
> phyper(c(1), m=12, n=6, k=3, lower.tail=FALSE)
[1] 0.754902
```

És dir, hi ha una probabilitat del 75% que se sàpiga dos o tres temes en l'examen.

1.2.4. Distribució de Poisson

En una distribució de Poisson $E(X) = V(X) = \lambda$.

Una altra de les distribucions discretes importants és la de Poisson, també derivada de la distribució binomial. A continuació prendrem l'exemple d'una variable aleatòria X, definida com el temps d'espera de l'autobús en minuts, i que té com a únic paràmetre λ, que representa tant la mitjana com la variància de la variable aleatòria.

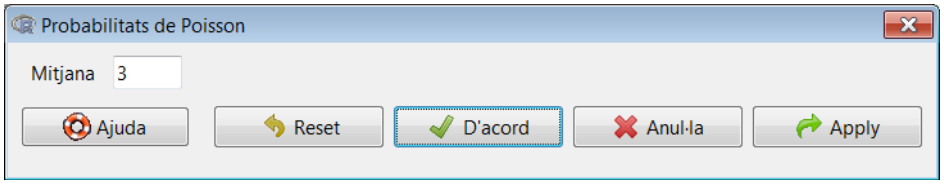
Suposem que el temps mitjà d'espera és de 3 minuts, aleshores $\lambda = 3$ minuts i, per tant, $X \sim Poiss(3)$ o alternativament $X \sim P(3)$. Per a saber la probabilitat que l'autobús tardi 5 minuts, el càlcul que s'ha de fer és el següent:

$$P(X = 5) = \frac{\lambda^k}{k!}e^{-\lambda} = \frac{3^5}{5!}e^{-3} = 0,1008.$$

R-Commander ens proporciona un resultat immediat a aquest càlcul utilitzant la ruta que es mostra a continuació del menú desplegable *Distribucions*:

Distribucions / Distribucions discretes / Distribució de Poisson / Probabilitats de Poisson

Com de costum, introduirem les dades del nostre exemple en el quadre de diàleg que apareixerà:



Per defecte, R-Commander ens dóna els deu primers valors de la probabilitat, és a dir, des de $P(X = 0)$ fins a $P(X = 10)$:

```
> .Table <- data.frame(Pr=dpois(0:10 ,  lambda = 3))

> rownames(.Table) <- 0:10

> .Table
              Pr
```

```
0  0.0497870684
1  0.1493612051
2  0.2240418077
3  0.2240418077
4  0.1680313557
5  0.1008188134
6  0.0504094067
7  0.0216040315
8  0.0081015118
9  0.0027005039
10 0.0008101512
```

De manera similar a les distribucions anteriors que hem descrit en aquest manual, amb la distribució de Poisson també es poden calcular probabilitats acumulades, gràfics i simulacions mostrals.

1.3. Distribucions de probabilitat contínues

1.3.1. Distribució uniforme

La distribució uniforme modelitza variables aleatòries contínues, de tal manera que tots els intervals en el rang de la distribució tenen la mateixa longitud i són igualment probables. El domini d'aquesta distribució està definit pels valors màxim i mínim a i b, respectivament. Una variable aleatòria X distribuïda segons una d'uniforme s'indica amb $X \sim U(a, b)$. La seva funció de densitat està determinada per l'expressió següent:

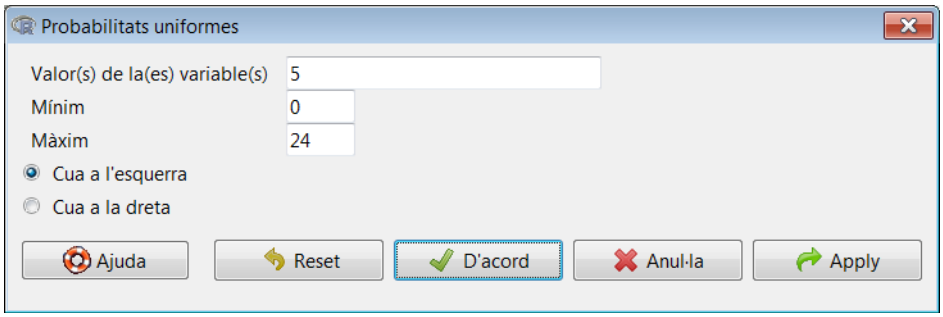
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{per a} \quad a \leq x \leq b, \\ 0 & \text{per a} \quad x < a \ \text{ o } \ x > b. \end{cases}$$

Per a veure com treballem amb variables aleatòries distribuïdes segons una d'uniforme, utilitzarem un exemple. Suposem que una dona està donant a llum un nadó, i l'hora exacta del naixement (X) serà qualsevol moment entre l'hora 0 (ara) i l'hora 24 (la mateixa hora del dia següent), seguint així una distribució uniforme, $X \sim U(0, 24)$.

La probabilitat que la mare doni a llum dins de les primeres cinc hores es calcula de la manera següent:

Distribucions / Distribucions contínues / Distribució uniforme / Probabilitats uniformes

En el quadre de diàleg introduirem el valor de la variable d'interès, el mínim i el màxim:



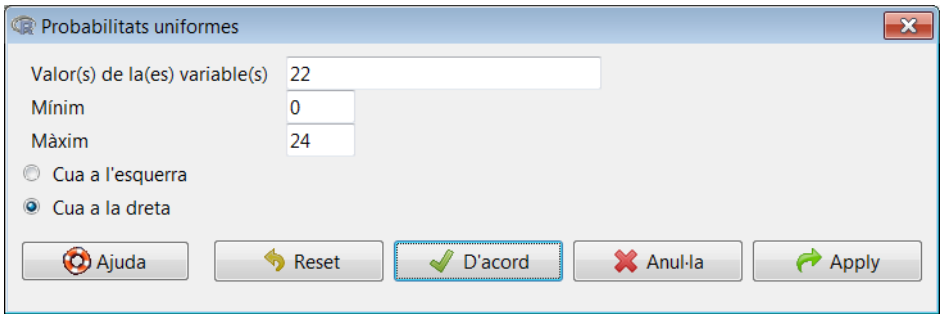
Tingueu present que busquem la probabilitat que queda en la cua de l'esquerra, $P(X < 5)$, i això també ho haurem d'indicar en el quadre de diàleg.

En la finestra de resultats obtindrem la probabilitat següent:

```
> punif(c(5), min=0, max=24, lower.tail=TRUE)
[1] 0.2083333
```

Com s'interpreten aquests resultats? Hem definit X com l'hora del part, i hem obtingut que $P(X < 5) = 0,208$, és a dir, la probabilitat que la mare doni a llum dins de les primeres cinc hores és aproximadament del 21%. És important tornar a insistir que, quan treballem amb distribucions contínues, $P(X < x)$ equival a $P(X \leq x)$, ja que es calculen les àrees de la distribució i no punts exactes. D'aquesta manera, com ja hem dit, en una distribució contínua el càlcul d'una probabilitat del tipus $P(X = x)$ sempre serà nul·la, és a dir, $P(X = x) = 0$.

Cal recordar que podem calcular probabilitats en totes dues cues de la distribució. Aquesta opció està disponible en el quadre de diàleg anterior, en el qual hem hagut d'especificar la cua de la distribució que ens interessa. Fixeu-vos que hem activat l'opció *Cua a l'esquerra*, ja que hem calculat $P(X \leq x)$. Si ara volem calcular la probabilitat que la mare infanti en les últimes dues hores (és a dir, a partir de l'hora 22), estarem calculant $P(X > 22)$, i aleshores haurem d'activar l'opció *Cua a la dreta*:



De manera que obtenim el resultat següent:

```
> punif(c(22), min=0, max=24, lower.tail=FALSE)
[1] 0.08333333
```

1.3.2. Distribució exponencial

La distribució exponencial està relacionada amb la distribució de Poisson. Aquesta distribució modelitza l'interval de temps que transcorre entre dos esdeveniments. Formalment, té un únic paràmetre $\theta = \lambda > 0$, i **està definida per a valors no negatius de la variable aleatòria**. Les funcions de densitat i de distribució prenen la forma següent:

En la distribució Exponencial $E(X) = 1/\theta$ i $V(X) = 1/\theta^2$.

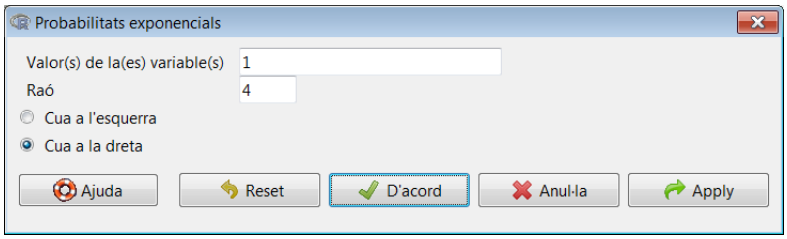
$$f(x) = \theta e^{-\theta x}$$

$$F(x) = P(X \leq x) = 1 - e^{-\theta x}$$

Vegem-ne un exemple: en un gran hospital, el temps que transcorre entre dos parts (mesurat en hores) segueix una distribució exponencial amb un paràmetre $\theta = 4$. Això significa que, atès que en aquesta distribució l'esperança es defineix com a $E(X) = 1/\theta = 0,25$, entre un part i un altre transcorre de mitjana un quart d'hora ($1/4 = 0,25$). Vegem la probabilitat que transcorri una hora o més entre dos parts, és a dir, volem calcular $P(X > 1)$. La ruta que hem de seguir en R-Commander serà:

Distribucions / Distribucions contínues / Distribució exponencial / Probabilitats exponencials

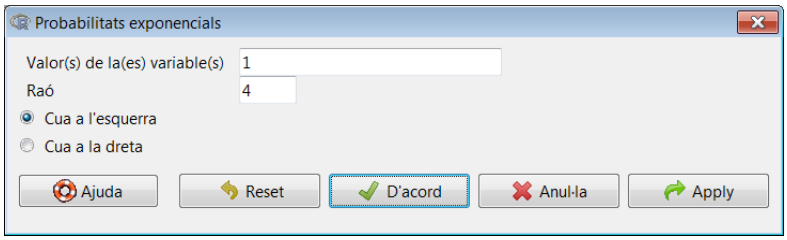
En el quadre de diàleg emergent introduïm la informació següent:



El resultat serà el següent:

```
> pext(c(1), rate=4, lower.tail=FALSE)
[1] 0.01831564
```

És a dir, $P(X > 1) = 0,018$. Això indica que aproximadament hi ha un $1,8\%$ de probabilitats que transcorri una hora o més entre dos parts. Alternativament, podem calcular la probabilitat complementària, és a dir, la probabilitat que transcorri com a màxim una hora entre dos parts. Per a això, caldrà seleccionar l'opció *Cua a l'esquerra* en el menú anterior:



El resultat obtingut serà el següent:

```
> pext(c(1), rate=4, lower.tail=TRUE)
[1] 0.9816844
```

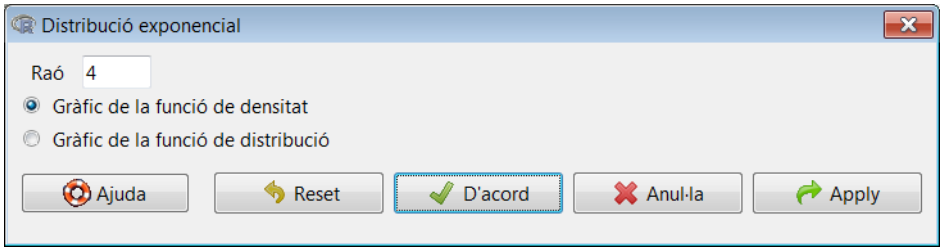
Aquest resultat no és estrany, ja que si sumem les probabilitats que s'han obtingut en les dues cues, la suma sempre és u. Això és així perquè sempre es compleix que:

$$P(X \leq x) + P(X > x) = 1.$$

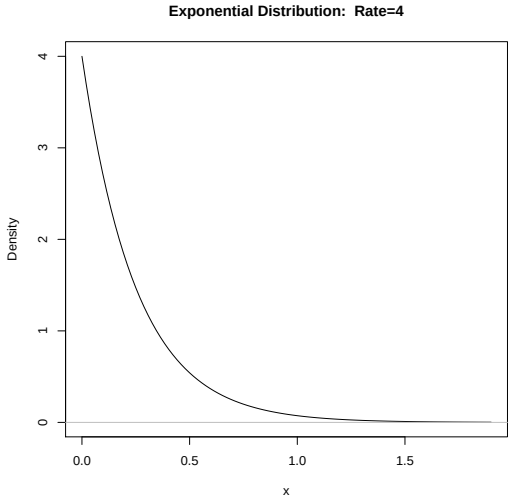
En aquest cas, $P(X \leq 1) = 1 - P(X > 1) = 1 - 0,018 = 0,982$. Un últim aspecte que veurem de la distribució exponencial és el resultat gràfic de les probabilitats calculades. Mitjançant la ruta següent:

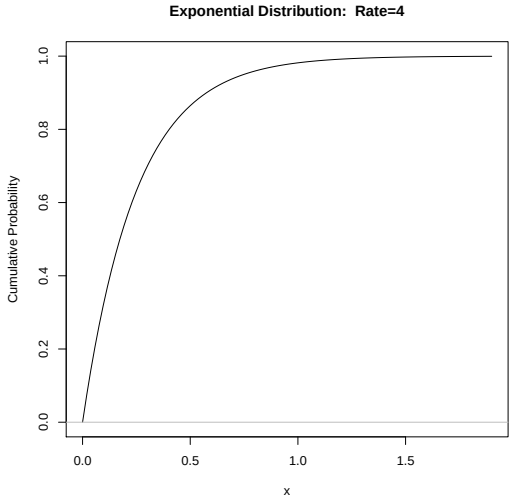
Distribucions / Distribucions contínues / Distribució exponencial / Traça una distribució exponencial

Com amb la resta de distribucions, podem escollir entre la gràfica de la funció de densitat i la de la funció de distribució:



Els gràfics obtinguts sempre es mostraran en la consola d'R, de manera que, per a acceder-hi, haurem d'anar a la barra de tasques i canviar la finestra activa a la d'R. Activant primer la *Gràfic de la funció de densitat* i després la *Gràfic de la funció de distribució*, obtindríem els gràfics següents (per separat):





Igual que amb la resta de distribucions, en el menú desplegable apareixen més opcions disponibles, com per exemple el càlcul dels quantils i l'obtenció de mostres aleatòries.

1.3.3. Distribució normal

La distribució normal o gaussiana és la distribució principal de probabilitat en estadística, i s'utilitza en infinitat d'àmbits. Aquesta distribució es caracteritza per tenir dos paràmetres: la mitjana μ i la desviació típica o estàndard σ. Així, es diu que una variable aleatòria X obeeix a una llei normal si $X \sim N(\mu, \sigma)$. La seva funció de densitat està definida per l'expressió següent:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

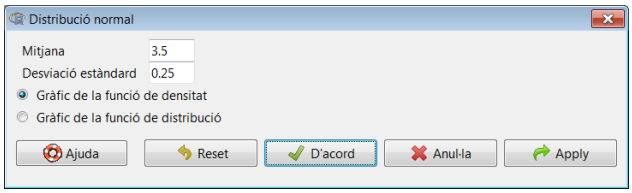
Vegem-ne un primer exemple. A l'hospital de l'exemple anterior, el pes dels nadons en néixer segueix una distribució normal amb una mitjana de $\mu = 3,5$ i una desviació típica $\sigma = 0,25$. Si volem visualitzar la forma de la funció de densitat, seguirem la ruta següent:

La notació de la distribució normal

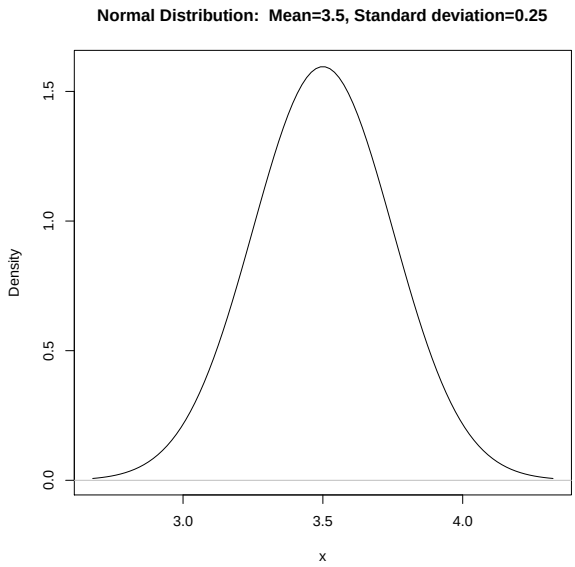
Sovint s'indica com a $X \sim N(\mu, \sigma)$ la variable que es distribueix segons una normal. No obstant això, no hi ha consens sobre aquesta notació, ja que a vegades s'expressa la variància (σ^2) en comptes de la desviació típica (depenent del llibre o manual). Aquí utilitzarem sempre σ per defecte.

Distribucions / Distribucions contínues / Distribució normal / Traça una distribució normal

En el quadre de diàleg resultant introduirem els paràmetres de la distribució i seleccionarem l'opció *Traça una funció de densitat*:



Amb això obtenim el gràfic de la funció de densitat:



A simple vista, veiem com la majoria dels nens en néixer pesen entre 3 i 4 kg. Ara calculem aquesta probabilitat ($P(3 \leq X \leq 4)$) amb R-Commander accedint a la ruta següent:

Distribucions / Distribucions contínues / Distribució normal / Probabilitats normals

El fet que sigui una distribució simètrica ens ofereix diferents opcions per a calcular probabilitats. Prenent el gràfic de dalt, veiem que si tracem una línia vertical en la mitjana aritmètica $(3, 5)$, en totes dues parts de la distribució queda el 50% de la massa probabilística. Això mateix també es pot veure fixant-nos que $\mu = 3,5$ se situa enmig dels valors 3 i 4; aleshores, l'àrea que quedarà a l'esquerra de 3 i l'àrea que quedarà a la dreta de 4 seran iguals gràcies a la simetria de la distribució. Per tant, si volem calcular l'àrea que queda entre 3 i 4, és a dir, $P(3 \leq X \leq 4)$, una opció és calcular-ne una menys les dues cues dels extrems, $P(X \leq 3)$ i $P(X \geq 4)$:

$$P(3 \leq X \leq 4) = 1 - P(X \leq 3) - P(X \geq 4)$$

Com que sabem que aquests dos extrems han de tenir la mateixa àrea o valor, és a dir, $P(X \leq 3) = P(X \geq 4)$, la fórmula anterior es pot simplificar, per exemple, de la manera següent:

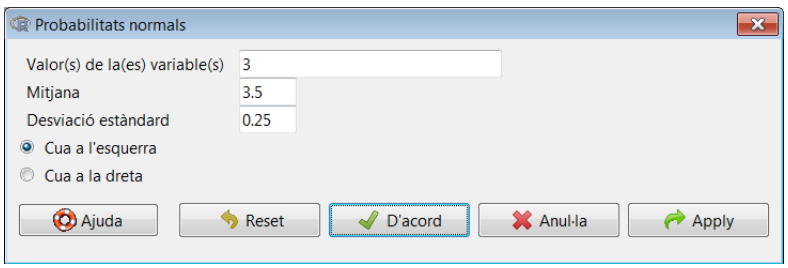
$$P(3 \leq X \leq 4) = 1 - 2\,P(X \leq 3)$$

Altres maneres alternatives d'obtenir aquesta probabilitat són:

$$P(3 \leq X \leq 4) = 1 - 2\,P(X \geq 4)$$

$$P(3 \leq X \leq 4) = P(X \leq 4) - P(X \leq 3)$$

Si seguim la primera opció, la probabilitat $P(X \leq 3)$ amb R-Commander es calcula així:



El resultat obtingut és el següent:

```
> pnorm(c(3), mean=3.5, sd=0.25, lower.tail=TRUE)
[1] 0.02275013
```

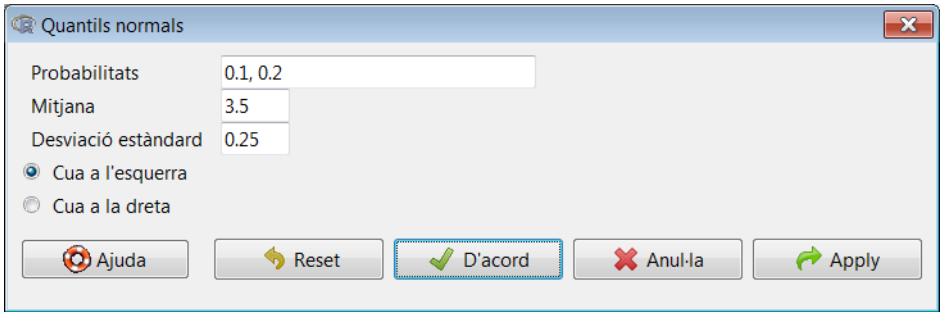
D'aquesta manera, obtenim que $P(3 \leq X \leq 4) = 1 - 2 \cdot 0,023 = 0,954$. És a dir, més del 95% dels nadons naixeran dins de l'interval de pesos [3,4]. **Consell:** en el quadre de diàleg anterior es pot introduir més d'un valor, separant els valors amb comes.

Recomanem a l'estudiant que provi el mateix càlcul en R-Commander seguint les maneres alternatives que hem descrit abans.

Ara vegem un exemple de càlcul de quantils. Suposem que, amb l'objectiu de fer un ús eficient de les incubadores i del personal mèdic per a la vigilància dels nadons, en aquest hospital el cap de pediatria vol separar-los en tres grups: (a) molt poc pes, (b) poc pes i (c) pes normal. El metge decideix que hi haurà un 10% dels nadons en el primer grup, un 20% en el segon, i la resta (un 70%) en el tercer. La pregunta és: quins seran els punts de tall entre els tres grups? Per a respondre a aquesta pregunta, hem de calcular els percentils 0,1 i 0,3, que divideixen l'àrea total en les tres parts que ens interessen (és a dir, 10%, 20% i 70%, en aquest ordre). Fem aquesta operació amb R-Commander:

Distribucions / Distribucions contínues / Distribució normal / Quantils normals

Introduïm els quantils en el quadre de diàleg:



Amb això obtenim els pesos que marquen el límit entre els tres grups:

```
> qnorm(c(0.1,0.2), mean=3.5, sd=0.25, lower.tail=TRUE)
[1] 3.179612 3.289595
```

Aquests percentils indiquen que el primer grup (nadons amb molt poc pes) estarà format pels nadons que pesen menys de 3,18 kg, el segon (nadons amb poc pes) pels que pesen entre 3,18 i 3,29 kg, i el tercer (nadons amb pes normal) pels que pesen més de 3,29 kg.

Distribució de la mitjana mostral

Un cas rellevant en l'estudi de la distribució normal és el de la **distribució de la la mitjana mostral** d'una variable aleatòria. Per a il·lustrar-ho, partim de l'exemple anterior de l'hospital. Els metges volen fer estadístiques de control del pes dels nadons que neixen. Interpretant els n nadons que neixen cada dia com una mostra aleatòria independent, per a cada una d'aquestes mostres s'analitza la mitjana mostral. Com que cada dia (és a dir, a cada mostra) s'obtindrà una mitjana mostral diferent, aquesta

es pot analitzar com una variable aleatòria denominada $\bar{X}$, que té la seva pròpia distribució. A l'hora d'analitzar i calcular probabilitats sobre $\bar{X}$, és imprescindible distingir si coneixem o no la dispersió de la població (σ) de la qual s'obté $\bar{X}$.

Si aquesta dispersió és coneguda (com en l'exemple anterior), $\bar{X}$ es distribueix com una normal, tal que:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Essent n la grandària de la mostra, suposarem que disposem d'una mostra de 150 nadons ($n = 150$), i volem calcular $P(\bar{X} \geq 3,75)$, sabent que $\mu = 3,5$ i *assumint que* σ *és coneguda i igual a* $0,25$.

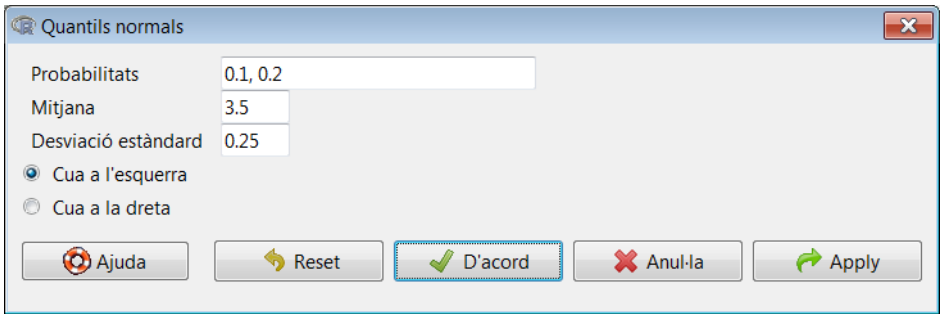
Abans de res, hem de calcular quina és la desviació típica de la variable mitjana mostral $\bar{X}$:

$$\frac{\sigma}{\sqrt{n}} = \frac{0,25}{\sqrt{150}} = 0,02.$$

Així, sabent que $\bar{X} \sim N(3,5;\ 0,02)$, calcularem aquesta probabilitat amb R-Commander accedint a la ruta següent:

Distribucions / Distribucions contínues / Distribució normal / Probabilitats normals

Haurem d'introduir els valors en el quadre de diàleg següent:



El resultat que s'obté és el següent:

```
> pnorm(c(3.75), mean=3.5, sd=0.02, lower.tail=FALSE)
[1] 3.732564e-36
```

Observem que el resultat, en notació científica, és pràcticament zero. Aquest resultat indica que la probabilitat que, en un dia, la mitjana mostral dels $n = 150$ nadons nascuts sigui de $3,75$ kg ($P(\bar{X} \geq 3,75)$) és pràcticament nul·la.

1.3.4. Distribució t de Student

Moltes funcions de distribució deriven de la normal estàndard

A partir de la distribució normal tipificada obtenim la funció khi quadrat $\chi^2_n = \sum_{i=1}^{n} Z_i^2$; la t de Student $t_n = \frac{Z}{\sqrt{\frac{\chi^2_n}{n}}}$ i la F de Snedecor $F_{n,d} = \frac{\frac{\chi^2_n}{n}}{\frac{\chi^2_d}{d}}$.

La distribució t de Student està associada a la distribució normal que acabem de veure. S'utilitza, entre altres casos, quan desconeixem la dispersió de la variable que s'ha d'analitzar. Reprenem l'exemple anterior del pes mitjà dels nadons que neixen a l'hospital esmentat ($\bar{X}$). Suposem que com abans prenem una mostra de $n = 150$ nadons, però en aquest cas només sabem que $\mu = 3,5$; així doncs, ens faltaria un paràmetre per a poder modelitzar $\bar{X}$ segons una distribució normal.

En aquest cas, haurem de fer una estimació de la desviació típica calculant la desviació típica mostral:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

De manera que en els càlculs de l'exercici anterior reemplaçarem σ per s. Aleshores, la distribució mostral de la mitjana ja no és una distribució normal com passava quan coneixíem el vertader valor de la desviació (σ). Si σ és desconeguda i n és la grandària de la mostra, calcularem l'error estàndard mitjançant el quocient següent:

$$\text{Error estàndard} = \frac{s}{\sqrt{n}}$$

Així doncs, si la variable que estudiem segueix una distribució normal amb mitja μ i desviació típica desconeguda, aleshores:

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

segueix una distribució t de Student amb $n-1$ graus de llibertat. En el nostre cas podem definir una nova variable aleatòria $\bar{Y}$, que representarà el pes mitjà d'aquests nadons que neixen a l'hospital esmentat. Formalment establim així aquesta distribució:

$$\bar{Y} \sim t_{n-1}$$

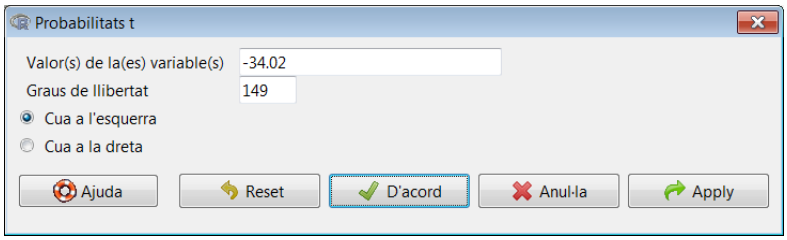
Suposem que volem saber la probabilitat que el pes mitjà dels nadons sigui inferior a 3 kg, és a dir, $P(\bar{X} \leq 3)$, sabent únicament que $\mu = 3,5$, $n = 150$ i $s = 0,18$. El primer pas serà transformar la variable $\bar{X}$ en la variable $\bar{Y}$ amb què poguem operar. En aquest cas, l'objectiu serà calcular $P(\bar{Y} \leq y)$. El valor de y l'obtenim a partir de l'expressió:

$$y = \frac{x - \mu}{\frac{s}{\sqrt{n}}} = \frac{3 - 3,5}{\frac{0,18}{\sqrt{150}}} = -34,02.$$

Un cop ja tenim el valor de la nova variable distribuïda com una t de Student, per a fer el càlcul de $P(Y \leq -34,02)$ amb R-Commander utilitzarem la ruta següent:

Distribucions / Distribucions contínues / Distribució t / Probabilitats t

El quadre de diàleg que hem d'emplenar és el següent:



Cal tenir en compte que hem introduït el valor que ens ha donat de la nova variable a *Valor(s) de la(es) variable(s)*, i a *Graus de llibertat* hem introduït $n-1 = 150-1 = 149$. A més, com que estem calculant $P(Y \leq y)$, hem d'activar l'opció *Cua a l'esquerra*. El resultat és el següent:

```
> pt(c(-34.02), df=149, lower.tail=TRUE)
[1] 1.973326e-72
```

Així doncs, hi ha una probabilitat nul·la que el pes mitjà dels 150 nadons nascuts aquest dia sigui inferior a 3 kg.

1.3.5. Teorema del límit central

El teorema del límit central (TLC) estableix que si una mostra és prou gran ($n > 30$), sigui quina sigui la distribució de la variable d'interès, **la distribució de la mitjana mostral** seguirà aproximadament una distribució normal. A més, la mitjana serà la mateixa que la de la variable d'interès, i la desviació típica de la mitjana mostral serà aproximadament l'error estàndard. El TLC té algunes implicacions, com que una variable aleatòria X distribuïda segons una binomial o una Bernoulli es pot aproximar a una de normal. En el cas de la distribució binomial, quan $n > 30$ i np i $n(1-p)$ tots dos són més grans que 5, la variable X s'aproxima a una normal amb $E(X) = np$ i $Var(X) = np(1-p)$.

El teorema de Moivre-Laplace

Segons aquest teorema una distribució binomial es pot aproximar a una distribució normal sempre que es compleixi que $np(1-p) > 5$.

Reprenem l'exemple que hem vist en el cas de la distribució binomial, el del jugador de bàsquet que encistellava triples. Recordem que $p = 0,3$, però ara suposem que el nombre d'intents (assajos) s'eleva fins a $n = 50$. Volem calcular la probabilitat que el jugador encistelli més de 20 triples. Pel teorema de Moivre-Lapalace i el TLC sabem que:

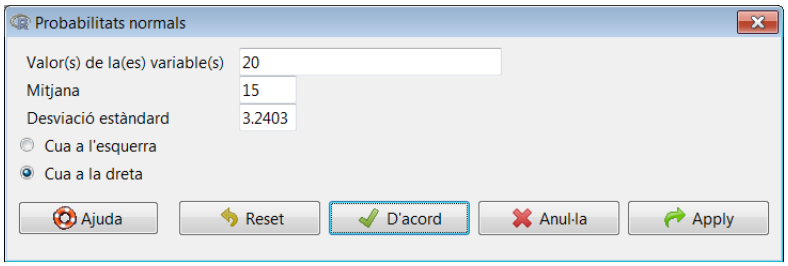
$$E(X) = np = 15$$

$$Var(X) = np(1-p) = 10,5$$

La ruta que hem de seguir en R-Commander per a calcular $P(X > 20)$ serà la següent:

Distribucions / Distribucions contínues / Distribució normal / Probabilitats normals

I el quadre de diàleg que ens apareixerà serà:



El primer que cal destacar és que **no s'ha d'introduir el valor de la variància en l'espai de la desviació típica, primer l'hem de calcular fent l'arrel quadrada**: això és, $\sqrt{10,5} = 3,2403$. El resultat d'aquesta probabilitat és el següent:

```
> pnorm(c(20), mean=15, sd=3.2403, lower.tail=FALSE)
[1] 0.06140726
```

És a dir, el jugador té una mica més del 6% de probabilitats d'encistellar 20 triples en 50 intents.

El segon cas plantejat fa referència a les variables discretes dicotòmiques que només prenen els valors 0 i 1, i que segueixen una distribució de Bernoulli. Lògicament aquestes variables estan estretament vinculades a les proporcions, que de fet són la mitjana de n variables aleatòries de Bernoulli de paràmetre p, en què n és la grandària de la mostra i p, la probabilitat d'exit de cada esdeveniment individual. Suposant la variable aletòria dicotòmica Y, quan la grandària de la mostra n sigui gran, la distribució de la proporció Y serà aproximadament una distribució normal que té com a paràmetres $E(Y) = p$ i $Var(Y) = p(1-p)/n$. Així doncs, en calcular la desviació típica, sempre que parlem de variables de Bernoulli aquesta serà igual a l'error estàndard d'aquesta variable.

En el nostre exemple, teníem que $n = 50$ i $p = 0,3$ i volíem esbrinar la probabilitat que el jugador de bàsquet encistellés més de 20 triples. Si suposem que ara estem aproximant una distribució de Bernoulli a una normal, el que estarem especificant és la proporció de triples que encistella el jugador en 50 intents. Aleshores, si volem calcular la probabilitat que aquesta proporció sigui superior al 40%, els paràmetres de la distribució normal que haurem d'utilitzar seran els següents:

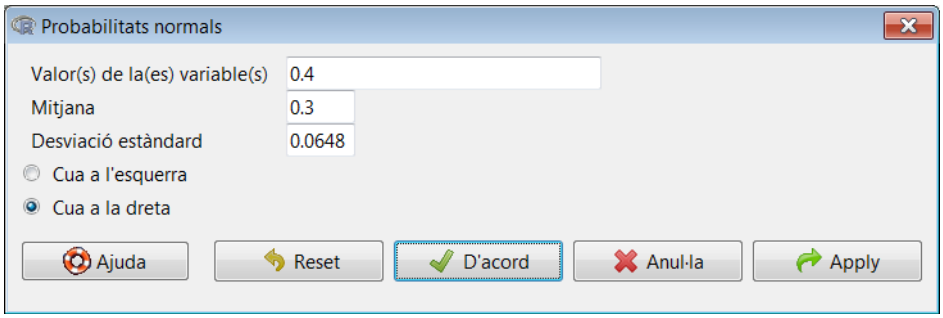
$$E(Y) = p = 0,3$$

$$Var(Y) = \frac{p(1-p)}{n} = 0,0042$$

I la ruta que hem de seguir en el menú desplegable d'R-Commander per a calcular $P(Y > 0,4)$ serà:

Distribucions / Distribucions contínues / Distribució normal / Probabilitats normals

Amb el quadre de diàleg següent:



Com abans, hem de destacar que **no s'ha d'introduir el valor de la variància en l'espai de la desviació típica**, és a dir, s'ha d'introduir $\sqrt{0,0042} = 0,06480$. A més, s'ha de remarcar que el resultat d'aquesta probabilitat és exactament el mateix que la calculada en l'exemple anterior. Això és del tot lògic, ja que una distribució de Bernoulli no és més que un cas particular de la distribució binomial. En el nostre cas, el 40% de triples equival a encistellar 20 triples en 50 intents. El resultat obtingut en R-Commander confirma aquest apunt:

```
> pnorm(c(0.4), mean=0.3, sd=0.06480741, lower.tail=
   FALSE)
[1] 0.0614113
```

És a dir, el jugador té una mica més del 6% de probabilitats d'encistellar el 40% dels triples, o el que és el mateix, d'encistellar 20 triples en 50 intents.

2. Inferència estadística

2.1. Introducció

Aquest capítol està dedicat a l'estudi de les eines principals de la inferència estadística: els intervals de confiança (IC) i els contrastos d'hipòtesis (CH). Normalment disposem d'una sèrie de dades mostrals de les quals ignorem els vertaders paràmetres poblacionals que han generat aquesta mostra. Per a això, els haurem d'estimar. En concret, aprendrem a treballar amb IC i CH, veurem com es calcula un IC i es resol un CH per a la mitjana aritmética, per a la proporció i per a la variància. A més, treballarem la construcció d'IC i la resolució de CH per a la diferència de mitjanes tant per a mostres aparellades com independents, la diferència de proporcions i el quocient de variàncies.

2.2. Inferència sobre la mitjana amb variància poblacional desconeguda

A l'hora de plantejar el càcul d'intervals de confiança per al paràmetre de la mitjana poblacional μ, en cas de desconèixer la variància poblacional (σ^2) haurem d'utilitzar la variància mostral s^2 i la desviació estàndard mostral (s). Com a conseqüència, en utilitzar un paràmetre estimat treballarem amb la distribució t de Student. Així, amb una mostra n i un nivell de significació α, l'IC per al paràmetre μ estarà determinat per l'expressió següent, que s'obté a partir de la mitjana mostral $\bar{X}$:

$$\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}.$$

Vegem un exemple pràctic d'inferència sobre el paràmetre μ amb variància poblacional desconeguda. Suposem que volem estudiar el pes dels estudiants d'una classe discriminant per sexes. Les dades inicials són les següents:

	pes	sexe
1	65	D
2	57	D
3	85	H
4	96	H
5	64	D
6	53	H
7	75	H
8	84	H
9	68	D
10	39	D
11	91	H
12	55	D
13	56	D
14	52	D
15	80	H
16	69	H
17	67	D
18	48	D
19	87	H
20	69	H

En total hi ha $n = 20$ estudiants en la mostra, dels quals 10 són nois i 10 són noies. Veiem que tenim dues variables: *pes* i *sexe*. La primera és una variable numèrica, mentre que la segona és un factor que informa de si l'observació és d'un home (H) o d'una dona (D).

Per començar, vegem l'estadística descriptiva d'aquestes dades accedint a la ruta següent:

Estadístics / Resums / Taula de dades activa

Aquesta instrucció fa que aparegui el resultat següent:

```
      pes   sexe
Min. :  39.00  D:10
1st Qu.:55.75  H:10
Median :67.50
Mean   :68.00
3rd Qu.:81.00
Max.   :96.00
```

Una altra manera alternativa d'obtenir més estadístics descriptius, com hem vist en el mòdul 3, és mitjançant la ruta següent:

Estadístics / Resums / Resums numèrics

Si premem *D'acord* en el menú que ens apareix, obtindrem la informació següent referent a la variable numèrica (pes):

```
mean       sd 0%   25%  50% 75% 100%  n
  68 15.55297 39 55.75 67.5  81   96 20
```

El primer pas serà fer inferència sobre la mitjana poblacional, per a la qual cosa utilitzarem tant el CH com l'IC. Com veurem, tots dos càlculs estan íntimament relacionats. Ens volem preguntar si la mitjana de pes de la classe es correspon amb la de l'escola, que és, suposem, $\mu_0 = 70$. Una primera aproximació consisteix a fer un contrast d'hipòtesi bilateral, amb el qual plantejarem les hipòtesis següents:

$$\begin{array}{ll} H_0: & \mu = 70 \\ H_1: & \mu \neq 70 \end{array}$$

Les hipòtesis mai no s'accepten

És molt important aclarir que, encara que sembli el mateix, estadísticament no té cap sentit **acceptar** una hipòtesi. El que fem és **no rebutjar-la** amb la informació de què disposem i al nivell de confiança donat. Per tant, els nostres resultats poden canviar si ens modifiquen el nivell de confiança o la informació amb la qual treballem.

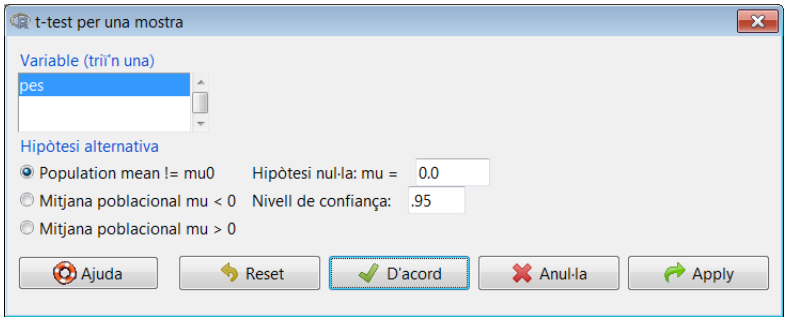
Els possibles resultats d'aquest contrast són dos: o bé no es rebutja H_0 o bé es rebutja H_0. És important destacar que el plantejament d'un contrast requereix un valor μ_0 i un nivell de confiança (o, alternativament, de significació). A més, la H_1 pot estar plantejada a dues cues (bilateral) o a una cua (unilateral). En aquest últim cas, pot ser per l'esquerra ($H_1 : \mu < \mu_0$) o per la dreta ($H_1 : \mu > \mu_0$).

Un interval de confiança parteix d'un raonament similar. La diferència principal és que no s'ha de proporcionar cap valor μ_0, sinó que s'han de buscar els dos valors de la variable aleatòria (en aquest cas, la mitjana poblacional μ) que deixin en les dues cues un percentatge α, que és el nivell de significació. Dit d'una altra manera, la probabilitat que el vertader valor de μ sigui dins de l'interval calculat serà igual al nivell de confiança $1 - \alpha$.

Vegem aquests conceptes en el nostre exemple. Si prenem la ruta següent:

Estadístics / Mitjanes / t-test per a una variable

Obtindrem el quadre de diàleg següent. Aquí hem d'introduir el valor μ_0 (que hem fixat en 70), si volem un contrast unilateral o bilateral, i el nivell de confiança, que és $(1 - \alpha) = 0,95$. Això implica que treballarem amb una significació de $\alpha = 0,05$.



El resultat és el següent:

```
One Sample t-test
data: Dades$pes
t = -0.5751, df = 19, p-value = 0.572
alternative hypothesis: true mean is not equal to 70
95 percent confidence interval:
 60.72099 75.27901
sample estimates:
mean of x
       68
```

Com interpretem aquest resultat? D'una banda, el contrast d'hipòtesi ens dóna un estadístic de $t = -0,5751$ i un *p-valor* de $0,572$. Això ens indica que l'estadístic està situat en la regió d'acceptació, i no en la regió crítica, amb la qual cosa no rebutgem la H_0. Dit d'una altra manera, la mitjana mostral obtinguda no és estadísticament diferent de 70. En general, hem de seguir la regla següent per a resoldre qualsevol contrast:

p-valor $\leq \alpha$	$\Rightarrow$	Rebuig de H_0
p-valor $> \alpha$	$\Rightarrow$	No rebuig de H_0

I, quant a l'interval de confiança, veiem que hem obtingut l'interval $[60,72\ ;\ \ 75,28]$.

En fer inferència amb R-Commander el *p-valor* obtingut ja considera si el contrast és unilateral o bilateral. Per tant, aquest *p-valor* **sempre s'ha de comparar amb** α **i no amb** $\alpha/2$.

Aquest resultat és coherent amb l'anterior, ja que el valor $\mu_0 = 70$ està inclòs en l'interval.

Què passaria si fixéssim un valor $\mu_0 = 80$? A continuació tornarem a fer el test amb aquesta dada (bilateral i amb $\alpha = 0,05$). Els resultats que obtindrem seran:

```
     One Sample t-test

data: Dades$pes
t = -3.4505, df = 19, p-value = 0.00268
alternative hypothesis: true mean is not equal to 80
95 percent confidence interval:
 60.72099 75.27901
sample estimates:
mean of x
       68
```

Veiem que, en aquest cas, rebutgem la H_0, ja que *p-valor* $< 0,05$, i això coincideix amb el fet que 80 no està en l'interval $[60,72\ ;\ \ 75,28]$.

2.3. Inferència sobre la mitjana amb variància poblacional coneguda

A diferència del cas anterior, ara sí coneixem el paràmetre de la variància poblacional (amb desviació estàndard σ), llavors utilitzarem la distribució normal per a calcular l'interval per al paràmetre μ. Específicament, tindrem l'expressió següent per a l'interval de confiança també basat en la mitjana mostral de què disposem ($\bar{x}$):

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

En l'exemple anterior hem assumit que no coneixem la desviació típica poblacional, i per a fer inferència (IC i CH) hem pres la desviació típica mostral, que és $s = 15,55$. Suposem que coneixem la desviació típica poblacional, i que aquesta és $\sigma = 20$. En aquest cas, per al càlcul d'IC i de CH haurem de fer servir la distribució normal, i no la t de Student.

Si ens hi fixem, aquesta opció no està configurada en el menú d'R-Commander i per aquest motiu l'haurem d'introduir manualment. Prèviament, necessitarem instal·lar un paquet estadístic anomenat PASWR (que significa *probability and statistics with R*). Recordeu que tots els paquets addicionals que necessitem només els haurem d'instal·lar una vegada, després ja en podrem disposar sense necessitat de tornar-los a instal·lar. Per a procedir a la instal·lació haurem d'introduir el següent en la finestra d'instruccions:

```
> install.packages("PASWR")
```

Un cop escrit, seleccionarem aquesta línia i premerem *Executar*, de manera que ens sortirà un menú desplegable de repositoris, i en triarem un qualsevol. Un cop instal·lat el paquet PASWR, l'haurem de carregar mitjançant la ruta següent:

Eines / carregar paquet(s) / PASWR

Aquest segon pas sí que l'haurem de fer sempre que volguem utilitzar la funció `z.test`. Plantejarem el contrast següent:

$$\begin{array}{ll} H_0: & \mu = 80 \\ H_1: & \mu \neq 80 \end{array}$$

En la finestra d'instruccions introduirem el test de la manera següent:

```
> z.test(Dades$pes, alternative='two.sided', mu=80,
+ sigma.x=20, conf.level=.95)
```

Quan tinguem escrita aquesta instrucció, s'haurà de seleccionar i prémer l'opció *Executar*. A més de la variable *pes*, colocada en primera posició, vegem quines altres instruccions componen la funció `z.test`:

`alternative`	`'greater'`: unilateral per la dreta `'less'`: unilateral per l'esquerra `'two.sided'`: bilateral
`mu`	Valor de μ_0
`sigma.x`	Valor de σ
`conf.level`	Nivell de confiança $(1-\alpha)$

En la finestra de resultats apareixerà el següent:

```
     One-sample z-Test

data: Dades$pes
z = -2.6833, p-value = 0.00729
alternative hypothesis: true mean is not equal to 80
95 percent confidence interval:
59.23477 76.76523
sample estimates:
mean of x
       68
```

La conclusió que en traiem és que rebutgem H_0, és a dir, el paràmetre μ és estadísticament diferent de 80.

2.4. Inferència sobre la proporció

En aquest cas, estem interessats a fer inferència sobre el paràmetre π, és a dir, la proporció poblacional. Essent $\hat{p}$ la proporció mostral i utilitzant la distribució normal, l'interval de confiança estarà determinat per l'expressió següent:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Alguns autors i R-Commander denominen el paràmetre de proporció poblacional p en comptes de π. Totes dues denominacions són correctes, el que passa és que l'alfabet grec se sol reservar per als paràmetres que es refereixen a la població i per aquest motiu nosaltres hem decidit utilitzar π.

Seguint amb l'exemple anterior, tenim que la mostra de $n = 20$ observacions està dividida entre homes ($n_h = 10$) i dones ($n_d = 10$). Aquest factor ens permet fer inferència sobre les proporcions. Atès que els factors estan ordenats alfabèticament com a H i D, la proporció per defecte serà $p = n_h/n$, és a dir, la proporció d'homes sobre el total. La proporció mostral és de $\hat{p} = 0,5$. Suposem que volem contrastar si la proporció és estadísticament diferent de $\pi = 0,6$. Al 95% de confiança i plantejat bilateralment, aquest contrast es planteja així:

$$H_0: \quad \pi = 0,6$$
$$H_1: \quad \pi \neq 0,6$$

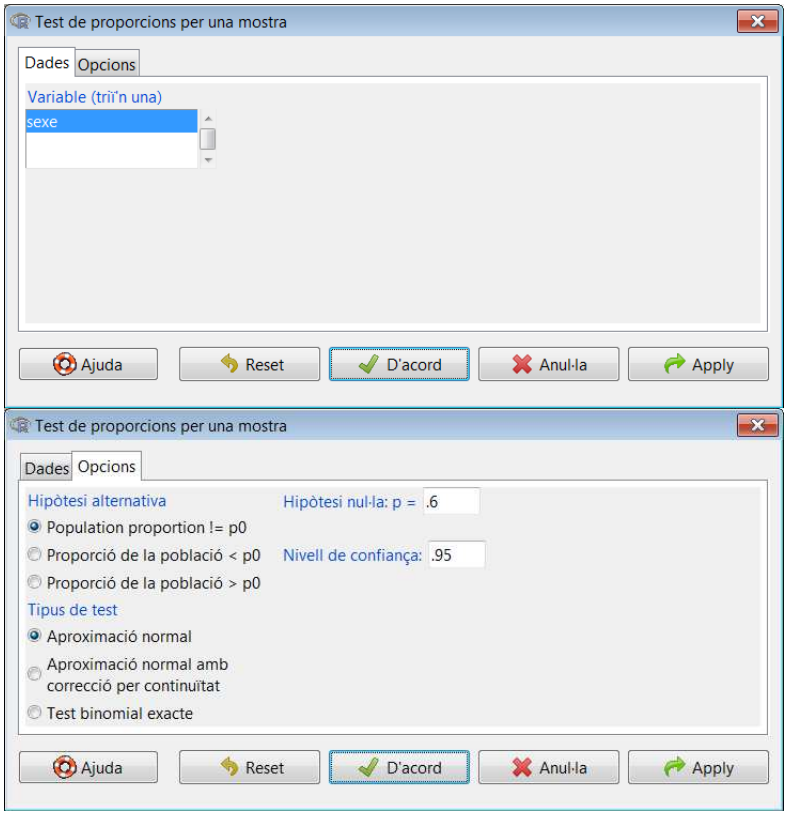
Intervals de màxim marge

De vegades no disposem de prou informació per a poder utilitzar la proporció mostral $\hat{p}$. En aquests casos s'utilitza $\hat{p} = 0,5$, ja que dóna lloc a l'IC més ample possible denominat de màxim marge.

Per fer el test, procedirem de la manera següent:

Estadístics / Proporcions / Test de proporcions per una mostra

Obtenim el quadre de diàleg següent, similar als casos anteriors:



El resultat del contrast serà:

```
  1-sample proportions test without continuity
     correction

data: rbind(.Table), null probability 0.6
X-squared = 0.8333, df = 1, p-value = 0.3613
alternative hypothesis: true p is not equal to 0.6
95 percent confidence interval:
 0.299298 0.700702
sample estimates:
  p
0.5
```

La interpretació d'aquest resultat és anàloga als casos anteriors. El contrast indica un no rebuig de H_0. A més, es comprova que l'interval de confiança és, aproximadament, $[0,3\ ;\ \ 0,7]$. És important destacar que l'IC i el CH són dues cares de la mateixa moneda: si la proporció $\pi = 0,6$ hagués estat fora d'aquest interval, hauríem rebutjat H_0.

2.5. Inferència sobre la variància

En aquesta secció implementarem la inferència amb un CH i un IC basats en la distribució χ^2. Específicament, volem analitzar si el valor de la variància poblacional (σ^2) és estadísticament diferent d'un valor predefinit σ_0^2. Aquest contrast, plantejat bilateralment, pren la forma següent:

$$\begin{array}{ll} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 \neq \sigma_0^2 \end{array}$$

En alguns llibres podeu trobar aquesta relació com a $\frac{(n-1)s^2}{\sigma^2} \sim \chi_n$. En realitat no hi ha consens però a mesura que creix n les diferències entre un i altre càlcul disminueixen.

En l'exemple anterior, hem vist que la distribució estàndard mostral pren el valor $s = 15,55297$, amb la qual cosa la variància mostral és $s^2 = 241,89$. Suposem que volem fer inferència sobre el paràmetre σ^2, de manera que volem calcular un interval al 95% de confiança d'aquest paràmetre. Sabem que l'estadístic pren l'expressió següent:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}$$

Per tant, l'interval que inclou el vertader valor poblacional de σ^2 prendrà la forma següent:

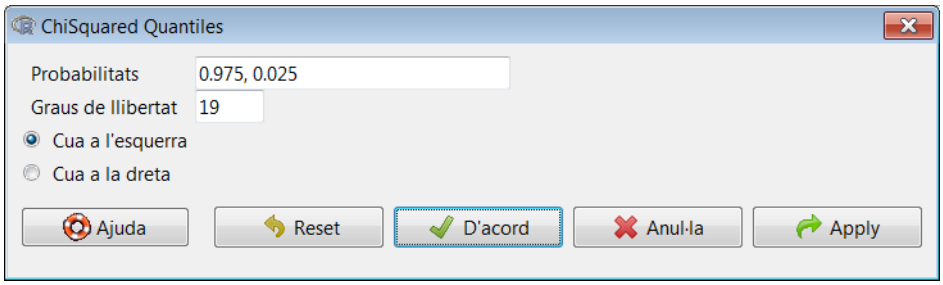
$$P\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2;\,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2;\,n-1}}\right) = 1-\alpha$$

El primer pas serà calcular els valors crítics. Amb un nivell de confiança de $\alpha = 0,05$ i una mostra de $n = 20$, aquests valors seran $\chi^2_{0,975;\,19}$ i $\chi^2_{0,025;\,19}$, els quals equivalen als quantils 0.975 i 0.025 de la distribució χ^2 amb 19 graus de llibertat, respectivament. En R-Commander, ho fem de la manera següent:

Penseu que la distribució χ^2 no és simètrica i, per tant, haurem de buscar els dos valors per a construir l'IC.

Distribucions / Distribucions contínues / Distribució Khi-quadrat / Quantils Khi-quadrat

Obtindrem el quadre de diàleg següent:



Amb això, obtenim valors següents de $\chi^2_{0,975;\,19}$ i $\chi^2_{0,025;\,19}$:

```
> qchisq(c(0.975,0.025), df=19, lower.tail=TRUE)
[1]  32.852327 8.906516
```

Aquests valors marquen els límits dins dels quals l'estadístic cau en la regió de no rebuig de H_0. Per a calcular els valors de l'interval per al paràmetre σ^2, haurem de calcular els extrems de l'interval de l'expressió anterior. Efectuant els càlculs manualment, obtenim el resultat següent:

```
> (20-1)*(15.55297^2)/c(32.852327,8.906516)
[1] 139.8988 516.0270
```

Així, al 95% de confiança, el paràmetre poblacional σ^2 estarà en l'IC = [139, 9 ; 516, 0].

2.6. Inferència sobre la diferència de mitjanes amb mostres independents

L'objectiu d'aquesta anàlisi és comprovar si hi ha diferències estadísticament significatives entre la mitjana d'una mateixa variable (pes) extreta en dues mostres independents diferenciades per la variable (sexe). La primera mostra està composta per homes (H) i la segona per dones (D). Així doncs, per a veure si hi ha diferències en el pes de les dues mostres es planteja el contrast següent:

$$H_0: \quad \mu_d = \mu_h$$

$$H_1: \quad \mu_d \neq \mu_h$$

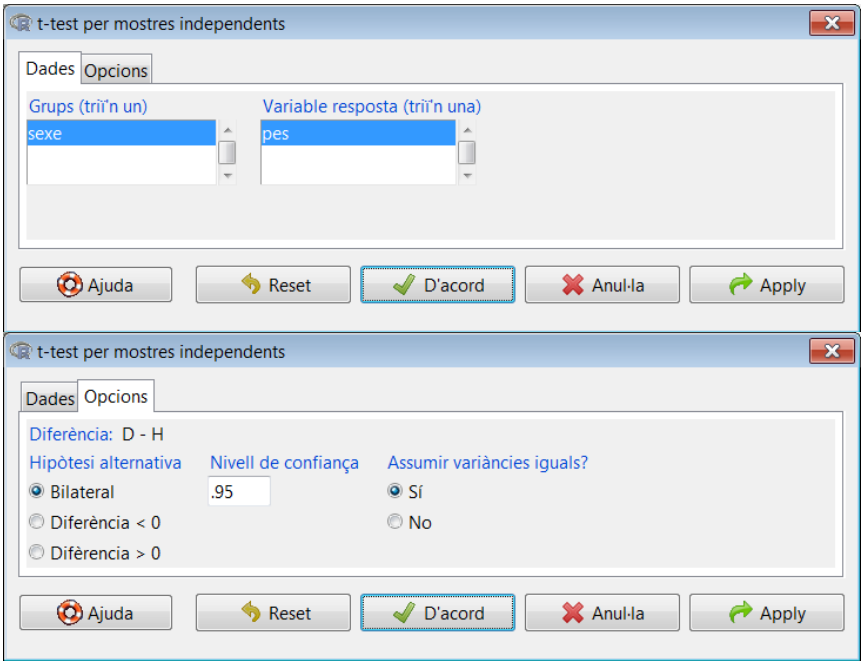
Alternativament, també es pot expresssar de la manera següent:

$$\begin{array}{ll} H_0: & \mu_d - \mu_h = 0 \\ H_1: & \mu_d - \mu_h \neq 0 \end{array}$$

2.6.1. Variàncies poblacionals desconegudes però iguals

En R-Commander utilitzarem la ruta següent i obtindrem el quadre de diàleg que es mostra a continuació:

Estadístics / Mitjanes / t-test per mostres independents



Cal destacar que aquest contrast (*Test t*) assumeix que les variàncies són desconegudes. El quadre de diàleg ens ofereix la possibilitat d'escollir si són iguals o no. En aquest cas, hem suposat que són iguals i el resultat és el següent:

```
  Two Sample t-test

data:  pes by sexe
t = -4.3896, df = 18, p-value = 0.0003535
alternative hypothesis: true difference in means is not
   equal to 0
95 percent confidence interval:
 -32.23387 -11.36613
sample estimates:
mean in group D mean in group H
          57.1            78.9
```

El resultat del contrast ens indica que la H_0 es rebutja, i que les mitjanes són estadísticament diferents. Arribem a aquesta mateixa conclusió considerant l'IC per al paràmetre $\mu_d - \mu_h$, $[-32,23\ ;\ \ -11,37]$. Com que aquest interval no conté el valor zero, el pes mitjà dels homes és estadísticament diferent del de les dones. Encara més, en ser els valors de l'IC positius, podem afirmar que la mitjana del pes dels homes és estadísticament superior a la de les dones.

2.6.2. Variàncies poblacionals conegudes

Aquest contrast varia si coneixem els valors de les variàncies poblacionals de μ_d i μ_h. En aquest cas, haurem d'utilitzar un contrast basat en la distribució normal i no en la distribució t de Student. Com quan es treballa amb una sola mostra, R-Commander no disposa d'aquest contrast en el menú i l'haurem d'introduir manualment. En la finestra d'instruccions, primer crearem dues variables diferenciades, el pes dels homes i el de les dones, cada una amb 10 observacions:

```
> pes_h<-Dades$pes[Dades$sexe=="H"]
> pes_d<-Dades$pes[Dades$sexe=="D"]
```

Un cop creades aquestes variables, tornarem a utilitzar la funció `z.test` del paquet PASWR, aquesta vegada amb dues variables, especificant la desviació estàndard de μ_h i μ_d (`sigma.x` i `sigma.y`, respectivament):

```
> z.test(pes_h, pes_d, alternative='two.sided',
+ sigma.x=4, sigma.y=3, conf.level=0.95)
```

El resultat d'aquest contrast que apareixerà en la finestra de resultats és:

```
    Two-sample z-Test

data: pes_h and pes_d
z = 13.7875, p-value < 2.2e-16
alternative hypothesis: true difference in means is not
   equal to 0
95 percent confidence interval:
 18.70102 24.89898
sample estimates:
mean of x mean of y
     78.9      57.1
```

Observem que, en aquest cas, l'IC és molt més precís, és a dir, l'interval és més estret, perquè coneix les dades poblacionals i utilitza la distribució normal en comptes de la distribució t de Student.

2.7. Inferència sobre la diferència de mitjanes amb mostres aparellades

Una mostra aparellada implica que es prenen dues o més observacions dels mateixos individus. Per a il·lustrar-ho, considerarem unes dades fictícies sobre medicina. Suposem que volem esbrinar si un determinat fàrmac afecta la freqüència cardíaca. Per a això, es dissenya un experiment mèdic amb una mostra de $n = 20$ persones, entre les quals hi ha homes (H) i dones (D). L'experiment consisteix a registrar la freqüència cardíaca abans (*freq_car_i*) i després (*freq_car_f*) de prendre aquest fàrmac, a més de registrar amb una variable dicotòmica si la persona ha tingut taquicàrdia. Les variables de la mostra són les següents:

Dades

	individu	sexe	freq_car_i	freq_car_f	taquicardia
1	1	D	60	68	No
2	2	D	68	80	No
3	3	H	71	75	No
4	4	H	69	79	No
5	5	H	79	85	Sí
6	6	H	84	91	Sí
7	7	H	64	70	No
8	8	H	60	69	No
9	9	H	70	76	No
10	10	H	75	80	No
11	11	D	80	87	Sí
12	12	D	63	68	No
13	13	H	65	71	No
14	14	H	75	83	Sí
15	15	D	72	80	No
16	16	D	77	79	No
17	17	H	78	84	Sí
18	18	D	79	83	Sí
19	19	D	72	81	Sí
20	20	D	71	85	Sí

L'objectiu del test és comprovar si la mitjana de la freqüència cardíaca és estadísticament diferent abans i després de l'experiment. Així doncs, es planteja el contrast següent:

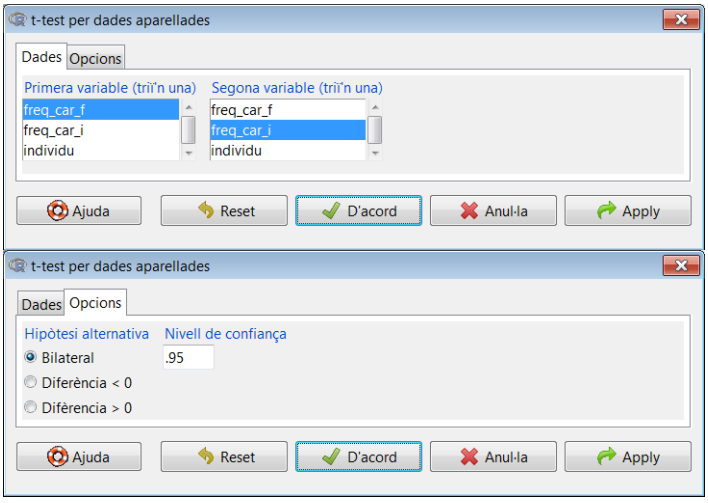
$$H_0: \quad \mu_{final} - \mu_{inicial} = 0$$

$$H_1: \quad \mu_{final} - \mu_{inicial} \neq 0$$

Per efectuar el test accedirem a la ruta següent:

Estadístics / Mitjanes / t-test per dades aparellades

Obtindrem el quadre de diàleg que es mostra a continuació, en què seleccionem totes dues variables:



El resultat és el següent:

```
  Paired t-test

data:  Dades$freq_car_f and Dades$freq_car_i
t = 11.3082, df = 19, p-value = 7.015e-10
alternative hypothesis: true difference in means is not
   equal to 0
95 percent confidence interval:
 5.78587 8.41413
sample estimates:
mean of the differences
                   7.1
```

El resultat del contrast ens indica que la H_0 es rebutja, i que les mitjanes són estadísticament diferents. Una altra manera de veure-ho és considerant l'interval de confiança per al paràmetre $\mu_{final} - \mu_{inicial}$, que és $[5,79\ ;\ \ 8,41]$. Com que aquest interval no conté el valor zero, la mitjana de la freqüència cardíaca dels individus al final de l'experiment és estadísticament diferent de la mitjana registrada al principi. Encara més, atès que els valors de l'IC són positius, podem afirmar que estadísticament la mitjana de la freqüència cardíaca s'ha incrementat amb el fàrmac. És interessant esmentar que la conclusió seria exactament la mateixa, però amb signe oposat, si haguéssim plantejat el contrast de la manera següent:

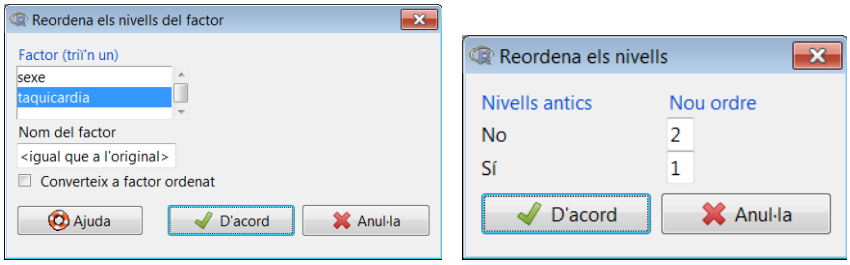
$$\begin{array}{ll} H_0: & \mu_{inicial} - \mu_{final} = 0 \\ H_1: & \mu_{inicial} - \mu_{final} \neq 0 \end{array}$$

2.8. Inferència sobre la diferència de proporcions

En aquesta secció contrastem si dues proporcions diferents són estadísticament iguals o no. Seguint l'exemple anterior, contrastarem si la proporció de taquicàrdies dels homes (π_{th}) és igual a la de les dones (π_{td}) sobre el total d'individus que van participar en l'experiment. R-Commander calcula aquestes proporcions automàticament i en ordre alfabètic a partir dels factors de la mostra. Per això, primer hem de reordenar els factors, de manera que l'ordre del factor sexe sigui *H* i *D*; i el del factor taquicàrdia sigui *Sí* i *No*. Això ho farem de la manera següent:

Dades / Modifica variables de la taula de dades activa / Reordena els nivells d'un factor

Apareixeran els dos quadres de diàleg que es mostren a continuació, en els quals ens preguntarà si volem sobreescriure la variable, i direm que sí:



En aquest últim quadre de diàleg hem invertit l'ordre dels nivells perquè el primer sigui un sí de manera que la proporció sigui sí ha tingut taquicàrdia sobre el total d'individus. En general, farem això per a tots els factors amb un ordre alfabètic que no coincideixi amb el nostre interès.

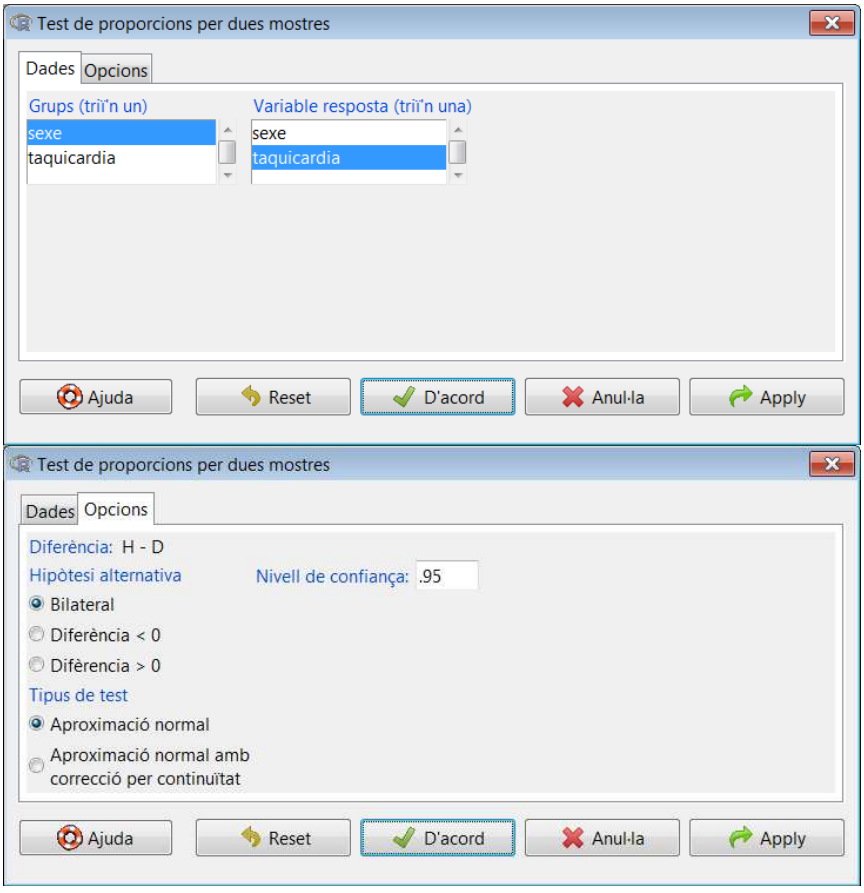
Ara farem el contrast següent:

$$\begin{array}{ll} H_0: & \pi_{th} - \pi_{td} = 0 \\ H_1: & \pi_{th} - \pi_{td} \neq 0 \end{array}$$

Per efectuar aquest contrast, seguirem la ruta següent:

Estadístics / Proporcions / Test de proporcions per dues mostres

I obtindrem el quadre de diàleg que es mostra a continuació:



El resultat és el següent:

```
taquicardia
sexe   Sí   No Total Count
   H 36.4 63.6   100    11
   D 44.4 55.6   100     9

  2-sample test for equality of proportions without
     continuity correction

data:  .Table
X-squared = 0.1347, df = 1, p-value = 0.7136
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.5123192  0.3507031
sample estimates:
   prop 1    prop 2
0.3636364 0.4444444
```

El resultat del contrast ens indica que la H_0 no es rebutja, i que no podem afirmar que les proporcions dels homes i les dones amb taquicàrdia siguin estadísticament diferents. De la mateixa manera, si considerem l'IC per al paràmetre $\pi_{th} - \pi_{td}$, que és $[-0,51\ ;\ \ 0,35]$. Com que aquest interval conté el valor zero, no podem afirmar que la diferència de proporcions sigui estadísticament diferent de zero.

2.9. Inferència sobre el quocient de variàncies

Aquesta secció inclou la inferència basada en la distribució F de Snedecor que es fa per a avaluar si les variàncies de dues mostres són iguals o no. Hem de dir que se suposa que les dues mostres procedeixen de dues variables normals, x_1 i x_2, que a més són independents i estan distribuïdes de manera idèntica. L'objectiu és comprovar si les variàncies d'aquestes variables són iguals.

El contrast que hem de plantejar és el següent:

$$\begin{array}{|ll|}\hline H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \\ \hline\end{array}$$

O bé, si aïllem i tenim en compte que quan parlem d'igualtat de variàncies ens referim al fet que el quocient d'aquestes és la unitat:

$$\begin{array}{|ll|}\hline H_0: & \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1: & \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \\ \hline\end{array}$$

Vegem-ho amb un exemple, suposem que disposem de dades sobre la cotització de dos valors borsaris, V_1 i V_2, amb una cotització diària que s'assumeix independent l'una de l'altra. Els primers valors de les dues variables:

Dades

	Dia	Valor1	Valor2
1	1	9.716769	9.523524
2	2	9.611507	9.402905
3	3	9.608912	9.487459
4	4	10.228090	10.815316
5	5	9.729010	9.906335
6	6	10.313833	11.164754
7	7	11.352083	13.329547
8	8	10.195410	11.103927
9	9	9.883333	10.566861
10	10	10.304893	11.496347

En aquest exemple específic, elaborarem dos gràfics de cotització i els compararem per tenir una primera evidència visual. Basant-nos en la sintaxi d'R que hem après en aquests materials, elaborarem els dos gràfics amb un cert nivell de detall, incorporant títol i bandes de fluctuació que marquen el mínim i el màxim de les cotitzacions.

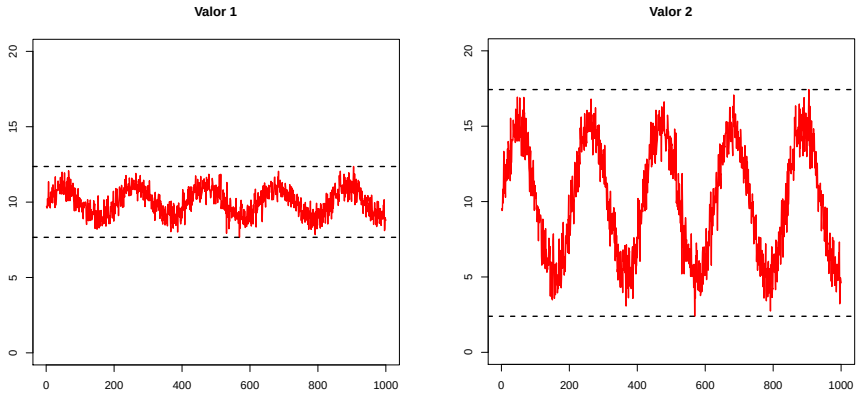
El gràfic per al primer valor es compon a partir de les instruccions següents:

```
> v1 <- Dades$Valor1

> plot(v1,type="l",lwd=2,col="red",main="Valor 1",
+      xlab="",ylab="",ylim=c(0,20))
> abline(h=min(v1),lwd=2,lty=2)
> abline(h=max(v1),lwd=2,lty=2)
```

Anàlogament, per al segon valor tenim les instruccions següents:

```
> v2 <- Dades$Valor2

> plot(v2,type="l",lwd=2,col="red",main="Valor 2",
+      xlab="",ylab="",ylim=c(0,20))
> abline(h=min(v2),lwd=2,lty=2)
> abline(h=max(v2),lwd=2,lty=2)
```

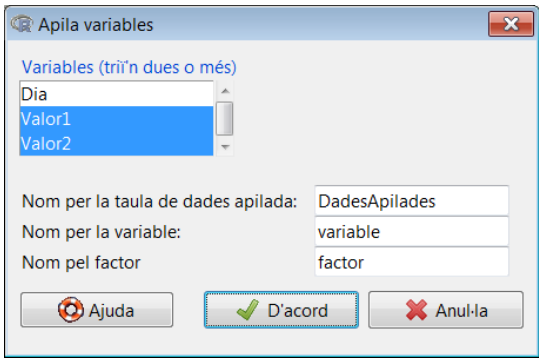
Visualitzant tots dos gràfics conjuntament, podem observar una major variabilitat o dispersió en la cotització del *Valor 2*:



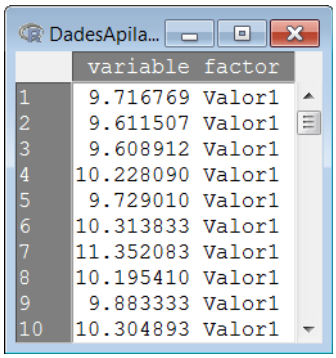
Per a obtenir una certesa estadística sobre la diferència de variàncies entre totes dues variables, haurem d'efectuar un test F de diferència de variàncies. Prèviament hem d'apilar les variables en una sola variable. Com s'ha vist anteriorment, la ruta que hem de seguir és aquesta:

Dades / Taula de dades activa / Apilar variables de la taula de dades activa

En el quadre de diàleg següent apilarem les variables *Valor1* i *Valor2* en una sola variable, i a més crearem un factor que identifiqui, per a cada observació, de quin valor es tracta:



Si visualitzem el nou conjunt de dades obtenim el següent:



El test F es basa en la ràtio de variàncies mostrals, de manera que l'estadístic pren la forma següent:

$$F = \frac{S_1^2}{S_2^2} \sim F_{n-1,m-1}.$$

La variància mostral

Recordem que, per a n observacions d'una variable X, la variància mostral es calcula com a $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

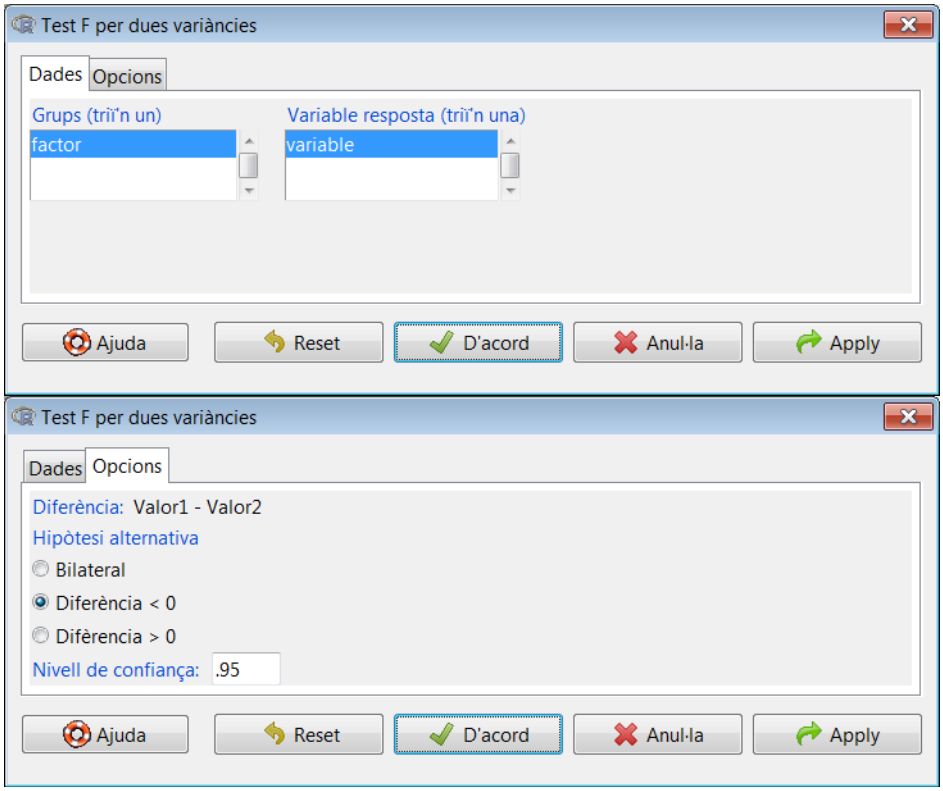
En què n i m és la dimensió de les dues mostres respectivament. En el nostre exemple, com que sospitem que la variància del segon valor és superior, plantejarem el contrast següent:

$$\begin{array}{ll} H_0: & \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1: & \frac{\sigma_1^2}{\sigma_2^2} < 1 \end{array}$$

En R-Commander, seguirem la ruta següent per efectuar aquest contrast, assumint un nivell de significació de $\alpha = 0,05$:

Estadístics / Variàncies / Test F de dues variàncies

En el quadre de diàleg que apareixerà especificarem quina és la hipòtesi alternativa:



El resultat és el següent:

```
> tapply(DadesApilades$Valor, DadesApilades$factor, var,
+ na.rm=TRUE)
    Valor1     Valor2
 0.7700158 13.6700790

> var.test(Valor ~ factor, alternative='less',
+ conf.level=.95, data=DadesApilades)

  F test to compare two variances

data:  Valor by factor
F=0.0563, num df=999, denom df=999, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is less
   than 1
95 percent confidence interval:
 0.0000000 0.0625104
sample estimates:
```

```
ratio of variances
        0.05632856
```

La interpretació és immediata: rebutgem la H_0, és a dir, la variància del *Valor 2* és clarament superior a la del *Valor 1*. Observem que el valor de l'estadístic és aproximadament de $F = 0,05$, la qual cosa ens indica que la variància del segon valor és més o menys 20 vegades superior a la variància del primer valor.

Bibliografia

Gibernans Bàguena, J.; Gil Estallo, À. J.; Rovira Escofet, C. (2009). *Estadística*. Barcelona: Material didàctic UOC.