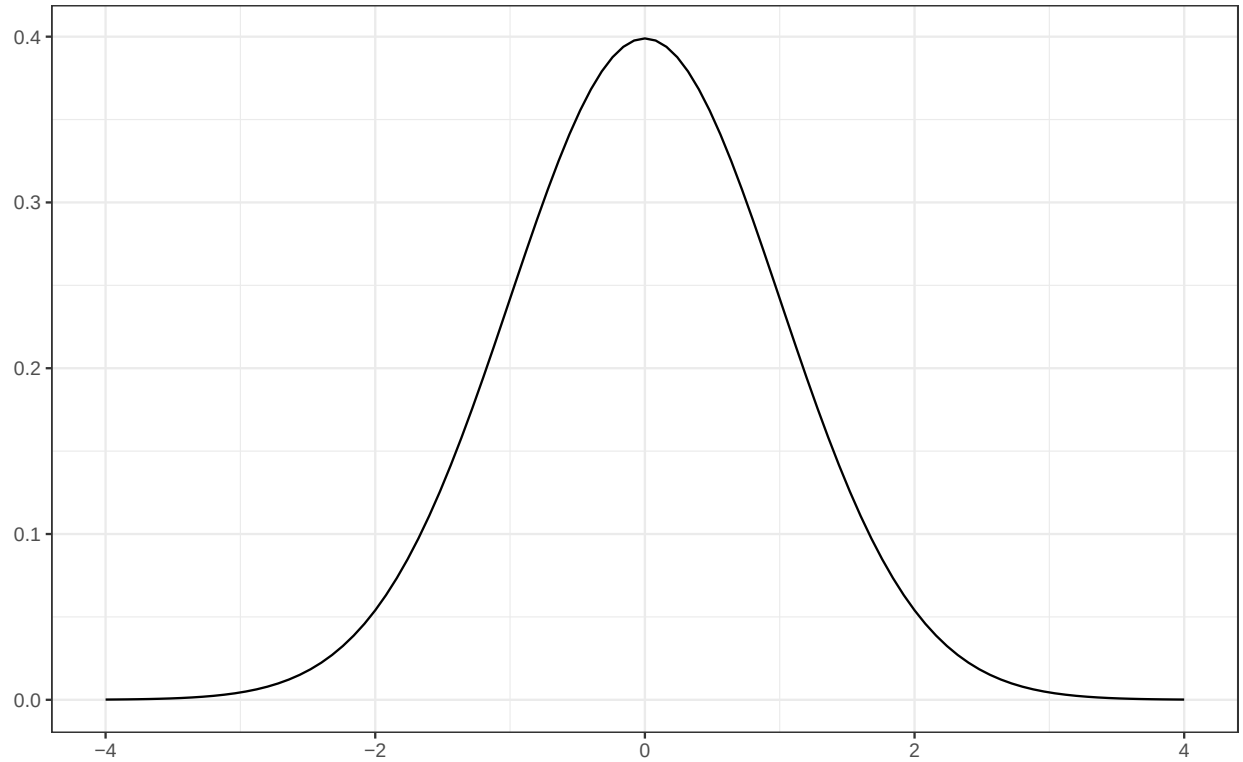
Distribucions 2

Pablo Sánchez
Tardor 2023

La distribució Normal

La següent distribució de probabilitat que tractarem és la **distribució normal**. És una de les distribucions de probabilitat més importants sobre les quals parlarem, ja que una infinitat de mètodes estadístics depenen d'ella i s'aplica a més situacions del món real que les distribucions que hem cobert fins ara. També es coneix com "campana de Gauss". Es descriu una distribució normal utilitzant dos paràmetres, la mitjana de la distribució μ i la desviació estàndard de la distribució σ.

Què significa que una variable es distribueixi normalment?. Si mireu la següent figura, aquesta representa una distribució normal amb mitjana $\mu = 0$ i desviació estàndard $\sigma = 1$. Podeu veure d'on prové el nom "campana de Gauss": sembla una mica una campana. Observeu que, a diferència dels gràfics que hem emprat per il · lustrar la distribució binomial, la imatge de la distribució normal mostra una corba suau en lloc de barres "com a histograma". Aquesta no és una elecció arbitrària: la distribució normal és contínua, mentre que la binomial és discreta. Per exemple, a l'exemple de llançament de daus era possible obtenir un 3 o un 4, però impossible obtenir un 3.9. Als histogrames que hem vist, per exemple, hi ha barres situades a X = 3 i a X = 4, però no hi ha res entremig. Les quantitats contínues no tenen aquesta restricció. Per exemple, suposem que estem parlant del temps. La temperatura en un agradable dia de primavera podria ser de 23 graus, 24 graus, 23,9 graus, o qualsevol cosa intermèdia, ja que la temperatura és una variable contínua, de manera que una distribució normal podria ser molt adequada per descriure les temperatures de la primavera.

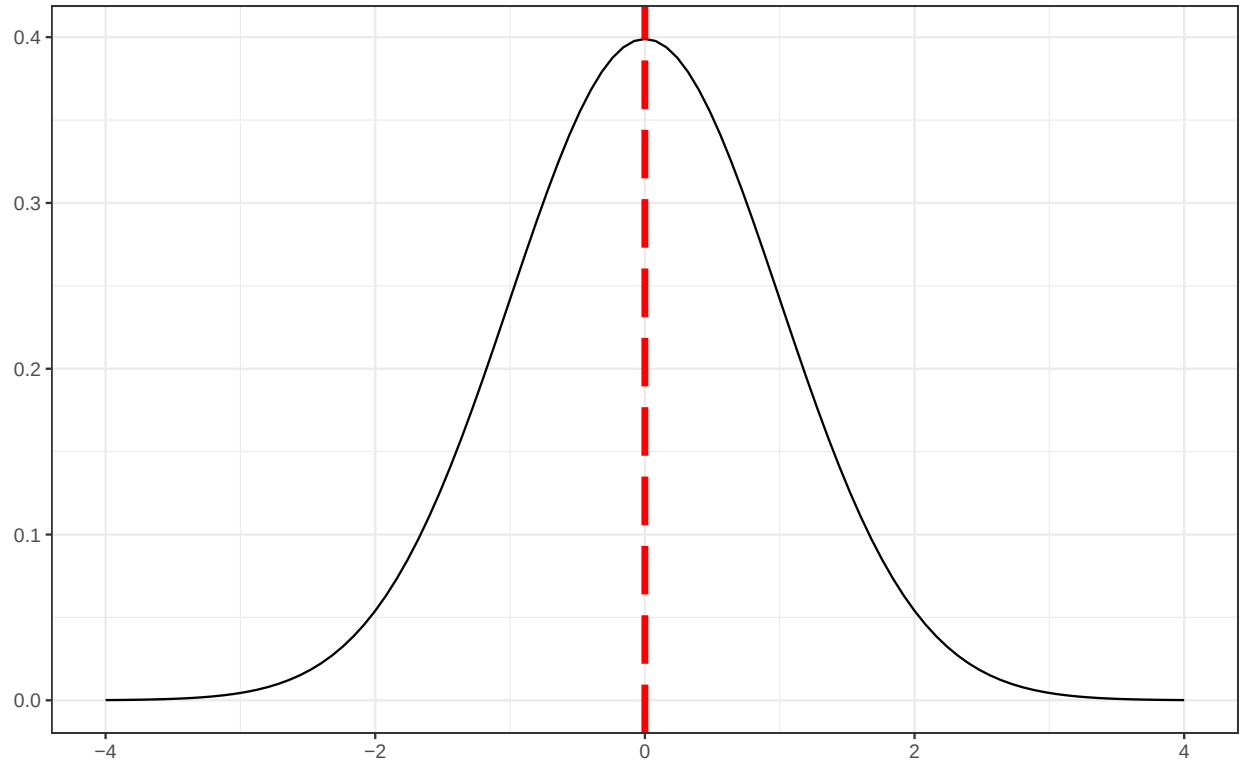


A la pràctica, la distribució normal és tan útil que tendim a utilitzar-la fins i tot quan la variable no és contínua. Sempre que hi hagi prou categories (per exemple, tipologies de fets policials), és una pràctica bastant estàndard utilitzar la distribució normal com a aproximació.

La distribució normal té algunes propietats importants:

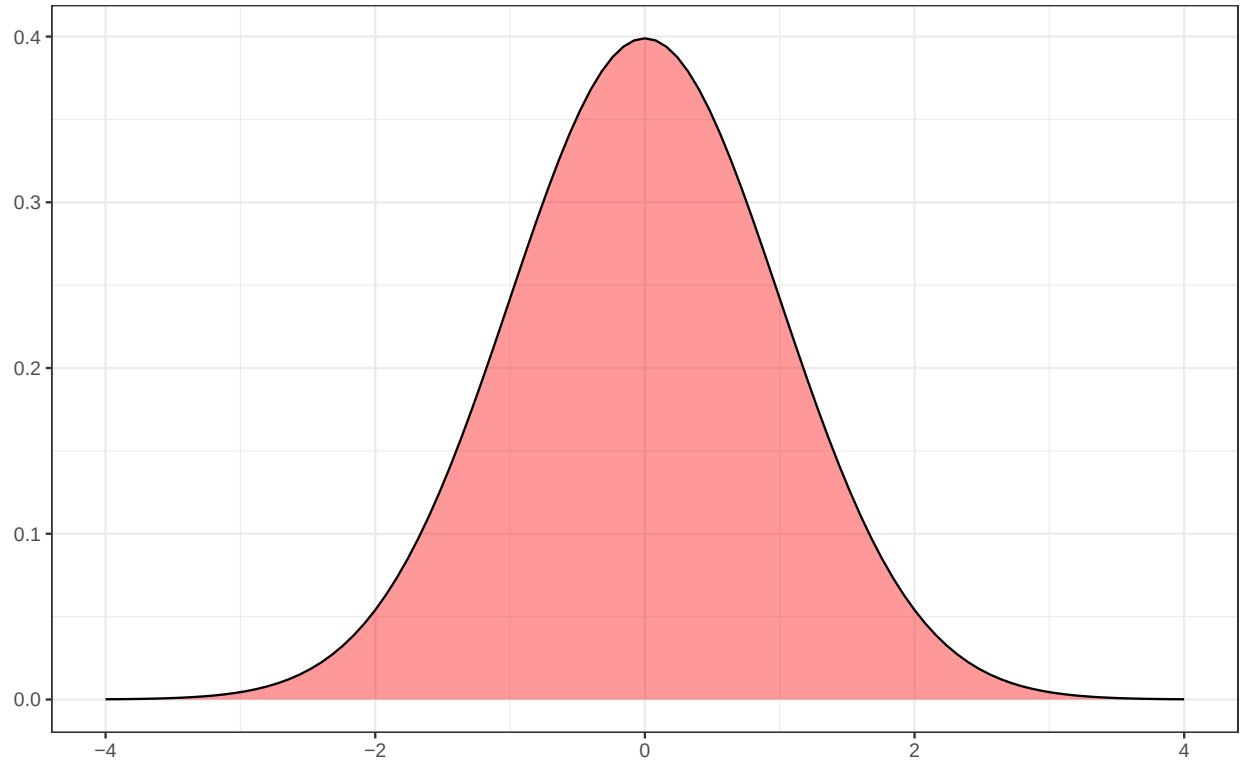
Simetria

La distribució normal és simètrica, així que la part esquerra és la imatge especular de la dreta.



Àrea = 1

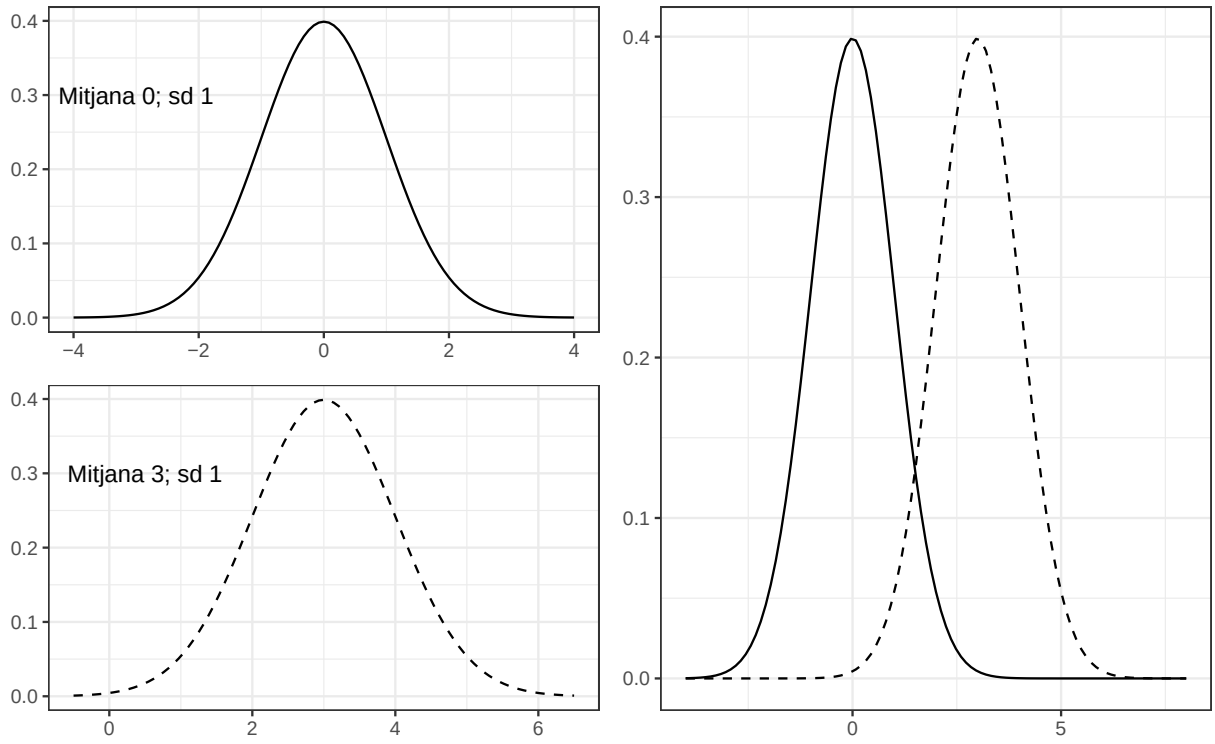
Igual que qualsevol distribució contínua, l'àrea sota la corba és 1.



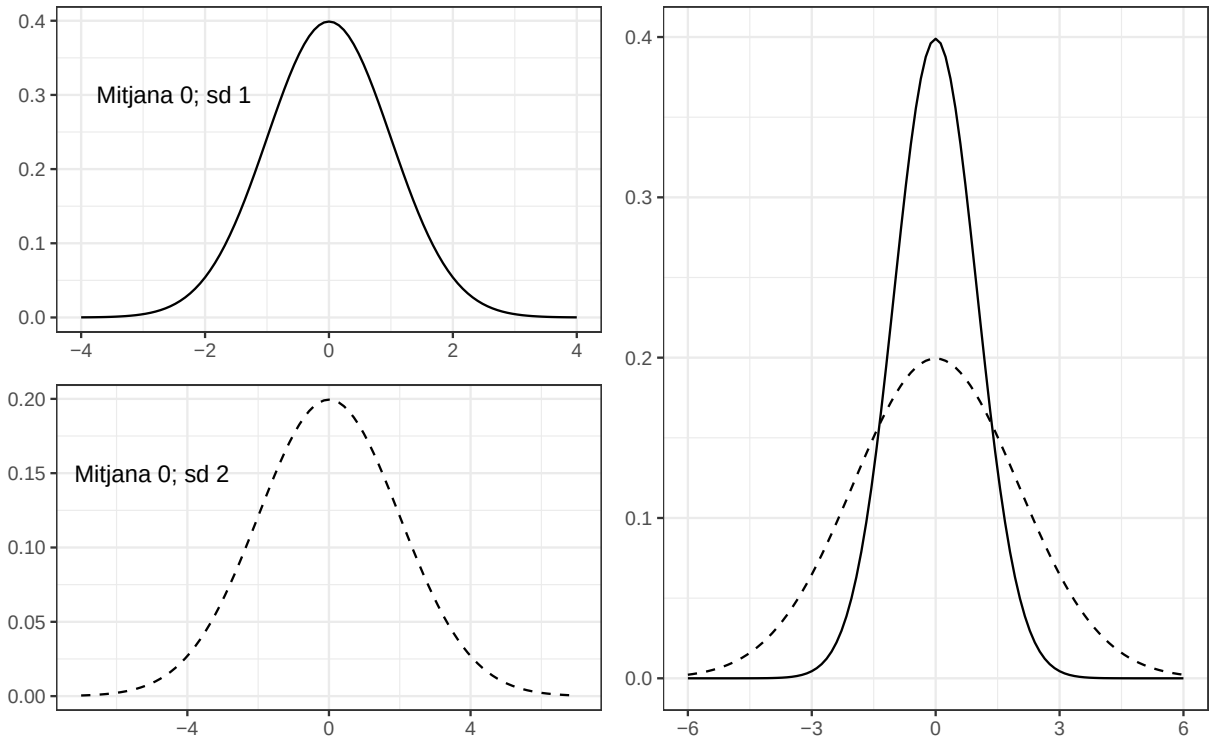
La corba mai arriba a zero

Finalment, la probabilitat no arriba mai a 0, encara que ho sembli als extrems de la cua. Només el 0.006% de la seva àrea està continguda més enllà de les vores d'aquest gràfic.

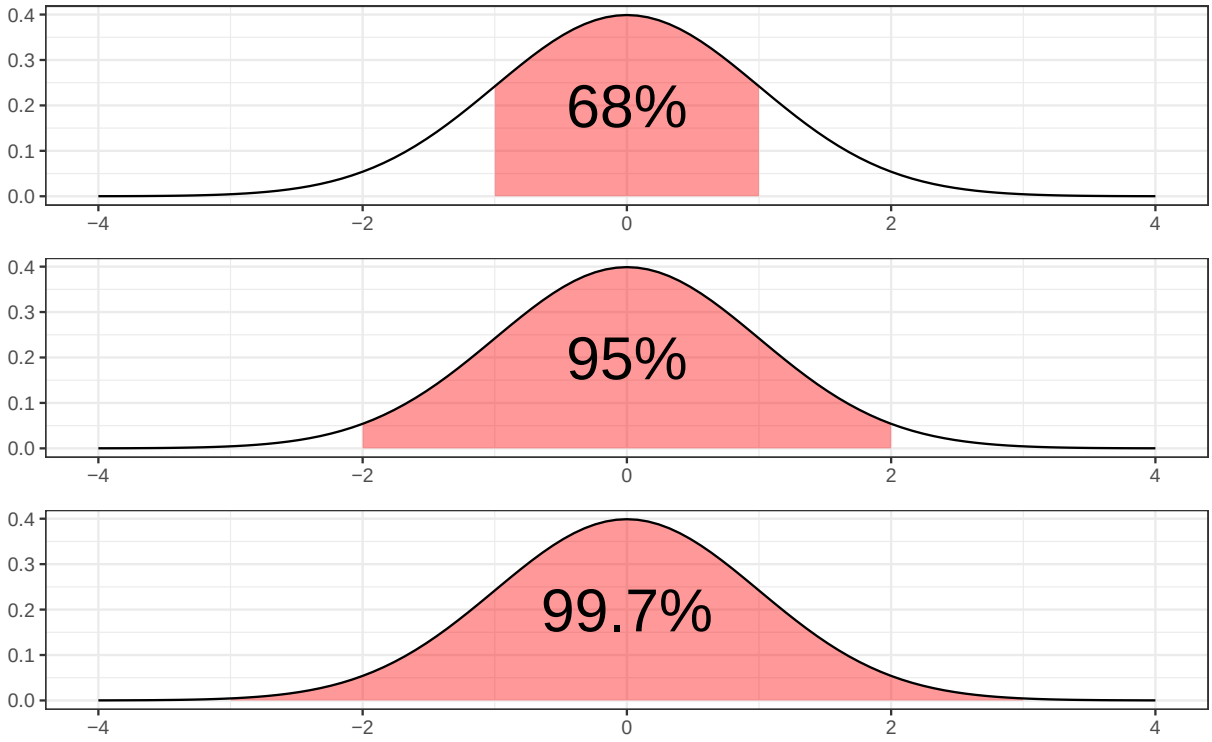
Tenint tot això en compte, a veure si podem tenir una intuïció de com funciona la distribució normal. En primer lloc, fem una ullada a què passa quan juguem amb els paràmetres de la distribució. Amb aquesta finalitat, la següent figura representa dues distribucions normals que tenen mitjanes diferents (0 i 3), però tenen la mateixa desviació estàndard (1). Com és d'esperar, totes aquestes distribucions tenen la mateixa "amplada" i la mateixa forma. L'única diferència entre elles és que s'han desplaçat cap a l'esquerra o cap a la dreta. En tots els altres aspectes són idèntiques. Hem d'esmentar especialment la distribució normal amb mitjana 0 i desviació estàndard 1. Aquesta distribució és la distribució normal estàndard.



En canvi, si augmentem la desviació estàndard mantenint la mitjana constant, el pic de la distribució es manté al mateix lloc, però la distribució s'amplia, com es pot veure a la següent figura. Tingueu en compte, però, que quan ampliem la distribució, l'alçada del pic es redueix. Això ha de passar ja què de la mateixa manera que les altures de les barres que hem utilitzat per dibuixar una distribució binomial discreta han de sumar 1, l'àrea total sota la corba per a la distribució normal ha de ser igual a 1.



Una característica important de la distribució normal és què, independentment de quina sigui la mitjana real i la desviació estàndard, el 68,3% de l'àrea es troba dins d'1 desviació estàndard de la mitjana. De la mateixa manera, el 95,4% de la distribució es troba dins de 2 desviacions estàndard de la mitjana i el 99,7% de la distribució es troba dins de 3 desviacions estàndard.



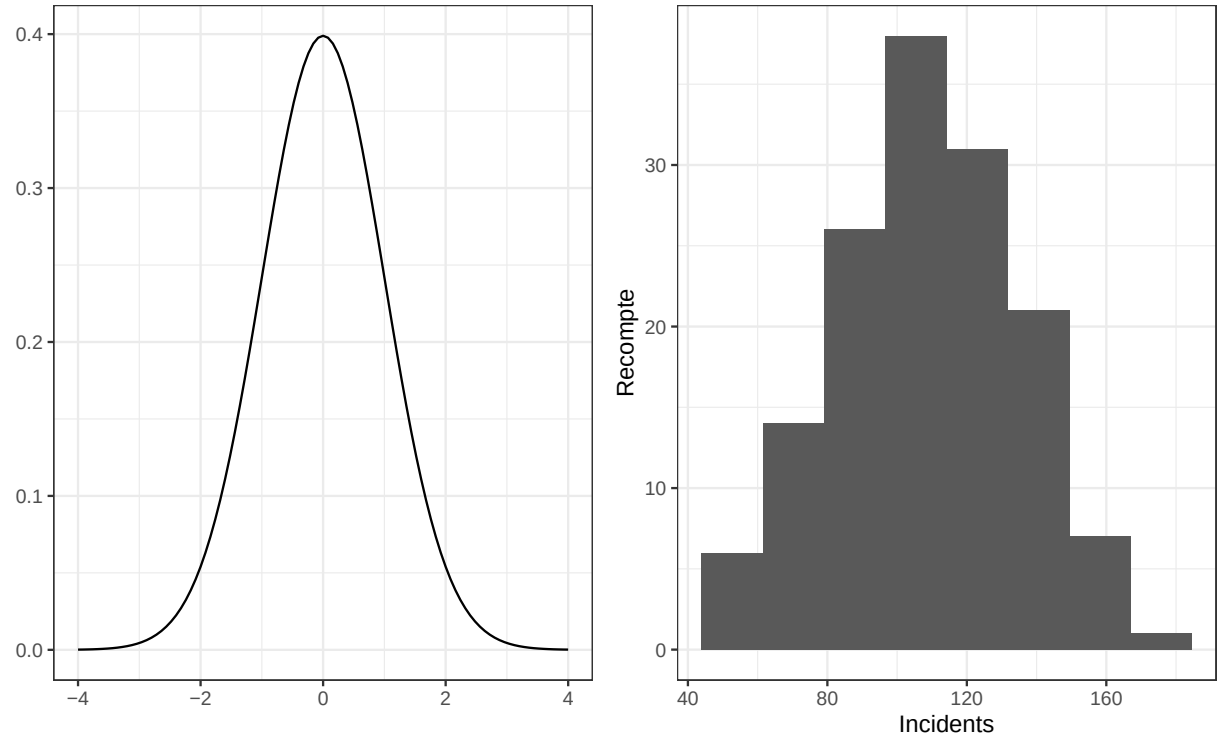
Hi ha molts conjunts de dades que es distribueixen de forma normal.

–> Carregueu la taula `incidents-policials-USC1.tsv` i seleccioneu tots els casos que hagin tingut lloc un divendres, bolcant-los a una nova taula `incidentsDivendres`

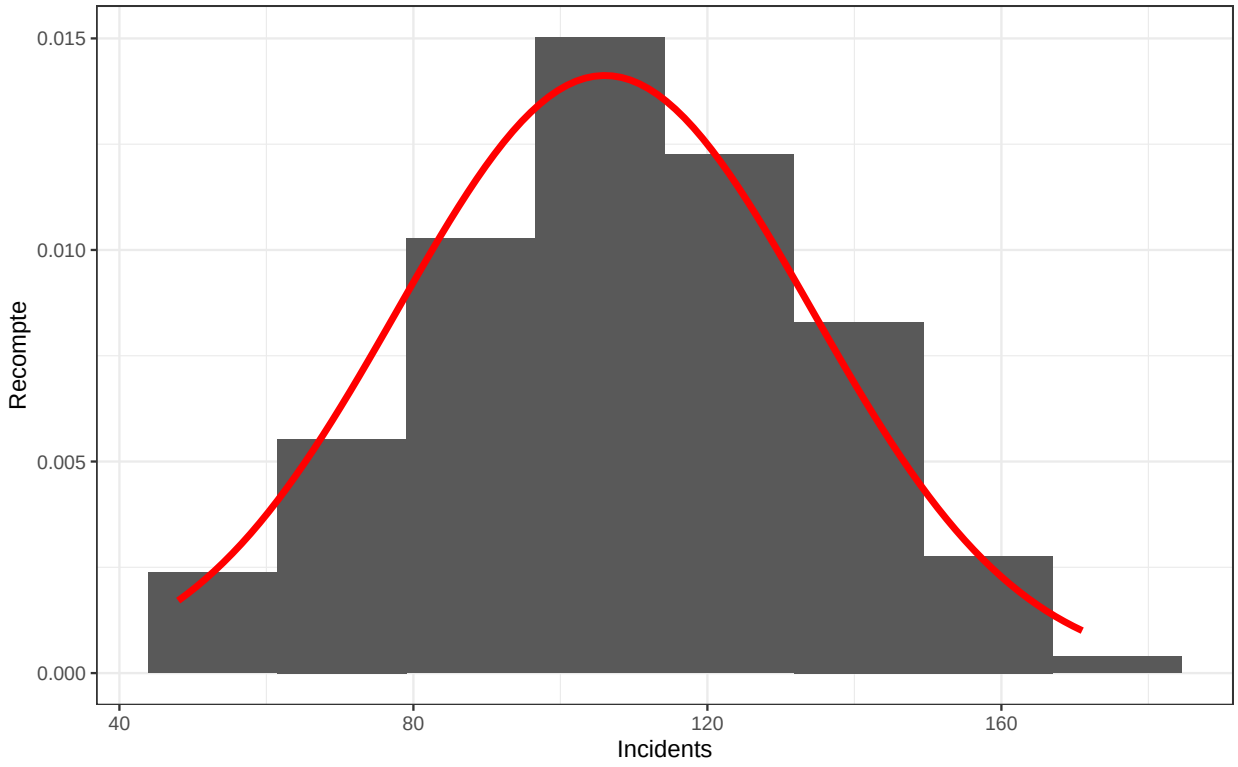
```
incidentsUSC1 <- read_tsv("incidents-policials-USC1.tsv", show_col_types = FALSE)
incidentsDivendres <- incidentsUSC1[incidentsUSC1$diaSetmana=="Divendres",]
```

Per exemple, en el següent gràfic tenim a l'esquerra una distribució normal estàndard (recordeu: mitjana 0; desviació estàndard 1) i a la dreta un histograma que representa la distribució del número de fets policials registrats per l'USC1 durant tots els divendres dels anys 2021, 22 i 23. La mitjana dels incidents dels divendres és de 109.4 i la seva desviació estàndard és de 26.1 incidents.

```
## # A tibble: 1 x 3
##   variable  statistic     p
##   <chr>         <dbl> <dbl>
## 1 Incidents     0.991 0.529
```



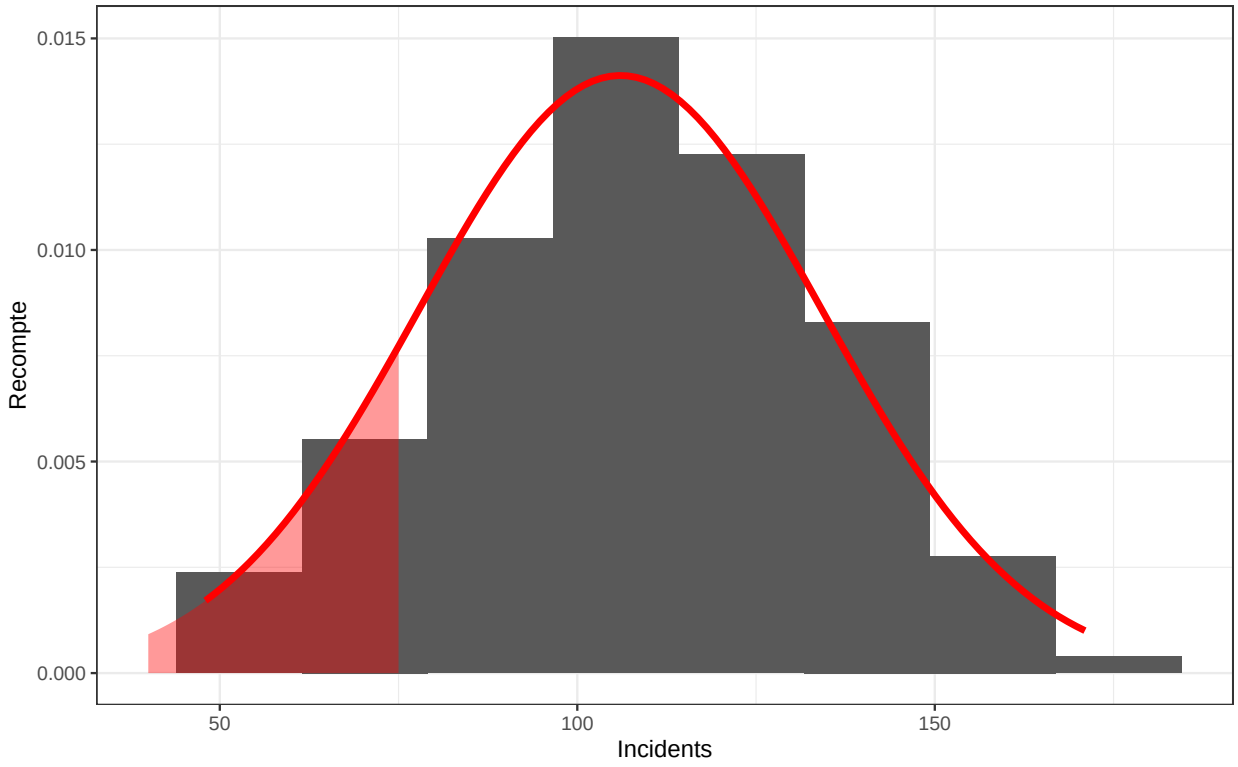
Com que aquestes dades s'assemblen molt a la distribució normal, podem prendre l'àrea sota una distribució normal amb una mitjana de 109 i una desviació estàndard de 26 per aproximar quin percentatge de dies cauen en diferents rangs d'incidents.



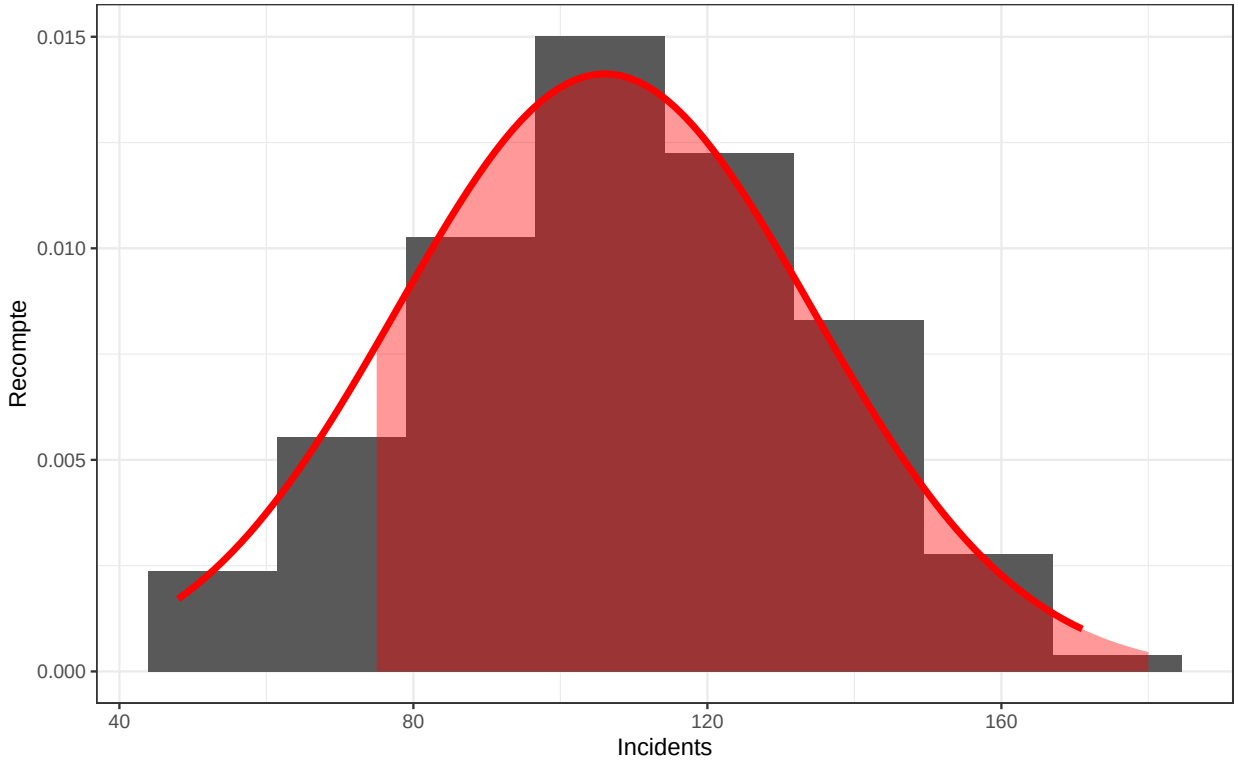
Amb aquestes dades podem preguntar-nos: Quin tant per cent de divendres tenen un número d'incidents menor a 75? Podem respondre-ho mitjançant la funció `pnorm`, que pren l'àrea de la distribució normal què està per sota d'un número. Passem el número d'interès, en aquest cas 75, així com la mitjana i la desviació estàndard de la distribució normal que estem utilitzant. Això ens dóna que al voltant del 10% dels divendres tindrem un número d'incidents per sota de 75.

```
pnorm(75, mean = 109, sd = 26)
```

```
## [1] 0.09548885
```



–> Per trobar el percentatge de divendres amb més de 75 incidents, podem fer servir l'argument `lower.tail` i li donarem un valor igual a FALSE. Així que la funció prendrà l'àrea a la dreta del primer argument.



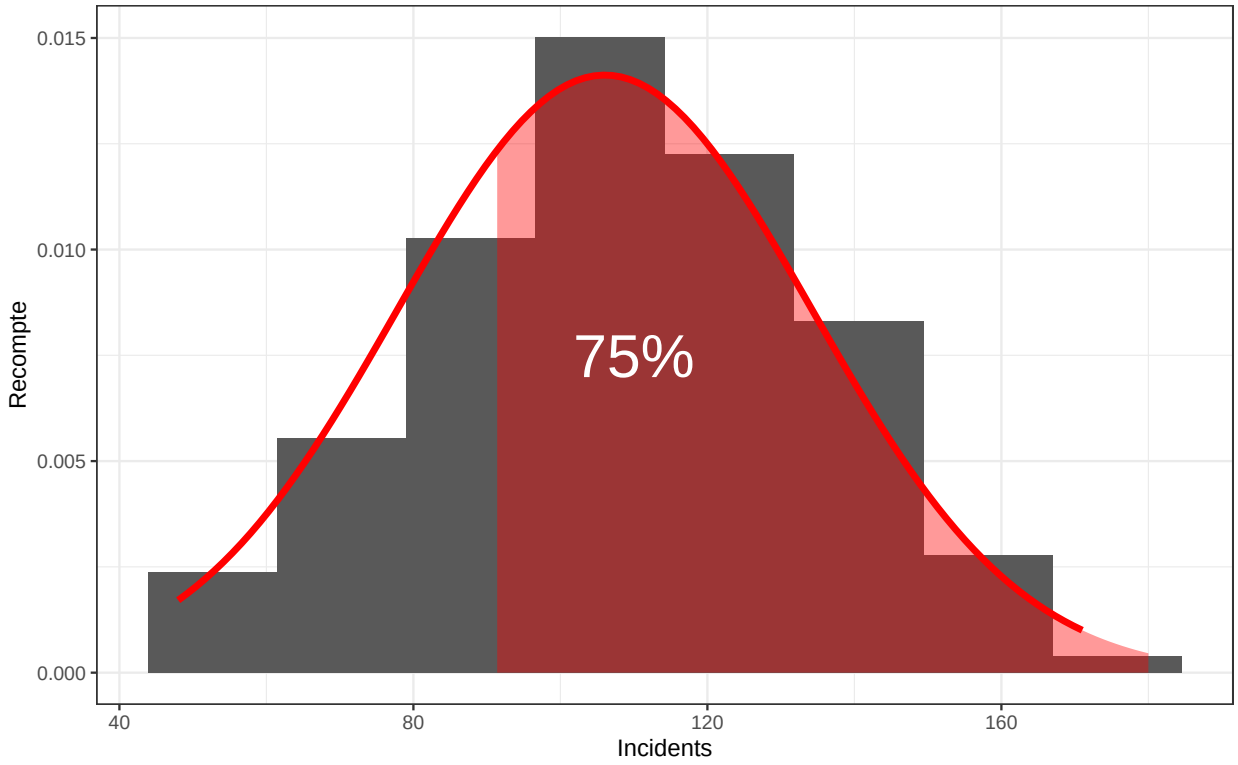
Per tal de trobar el tant per cent de dies que tenen entre 75 i 100 incidents, tal i com ja hem vist amb d'altres distribucions, només hauríem de restar les dos àrees.

–> Quin percentatge de dies estan entre 75 i 100 incidents?

També podem fer l'operació inversa i obtenir el número d'incidents que coincideix amb un quantil concret amb la funció `qnorm`.

–> Quin número d'incidents està al quantil 90?

–> Podem fer servir novament l'argument `lower.tail` per trobar, per exemple, quants incidents hi ha com a mínim el 75% dels dies?



Amb R també podem generar distribucions normals amb números aleatoris mitjançant `rnorm`, passant la mida de la mostra que volem juntament amb la mitjana i la desviació estàndard de la distribució.

–> Feu servir `rnorm` per a generar 10 valors aleatoris d'incidents dels divendres fent servir els valors de mitjana i desviació estàndard que ja hem calculat per a les nostres dades reals. Què podeu comentar dels resultats?

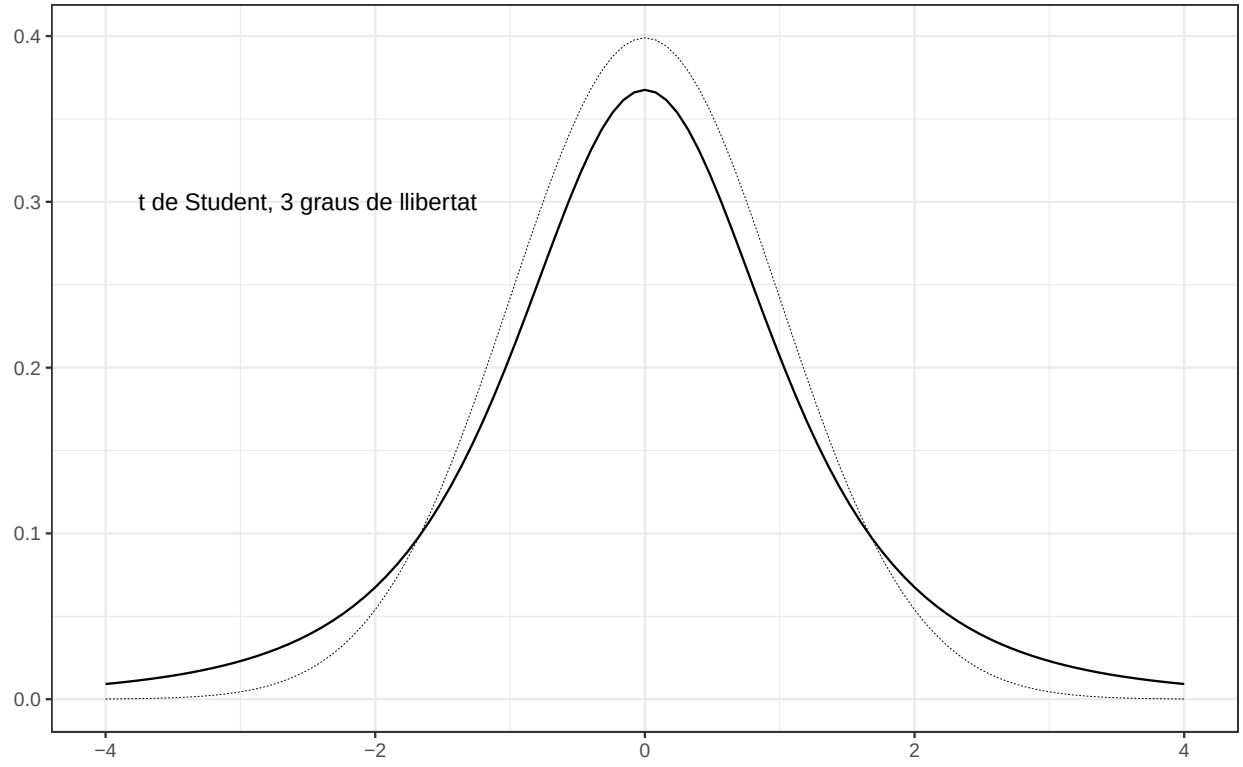
Altres distribucions útils

La distribució normal és la distribució que l'estadística fa més ús, i la distribució binomial és molt útil per a molts propòsits. Però el món de l'estadística està ple de distribucions de probabilitat, algunes de les quals ens trobarem de passada. En particular, tres distribucions amb les que ens podrem trobar són la distribució t, la distribució χ^2 i la distribució F. No donarem fórmules per a cap d'aquestes, ni en parlarem amb detall, però veurem gràficament la pinta que fan.

t de Student

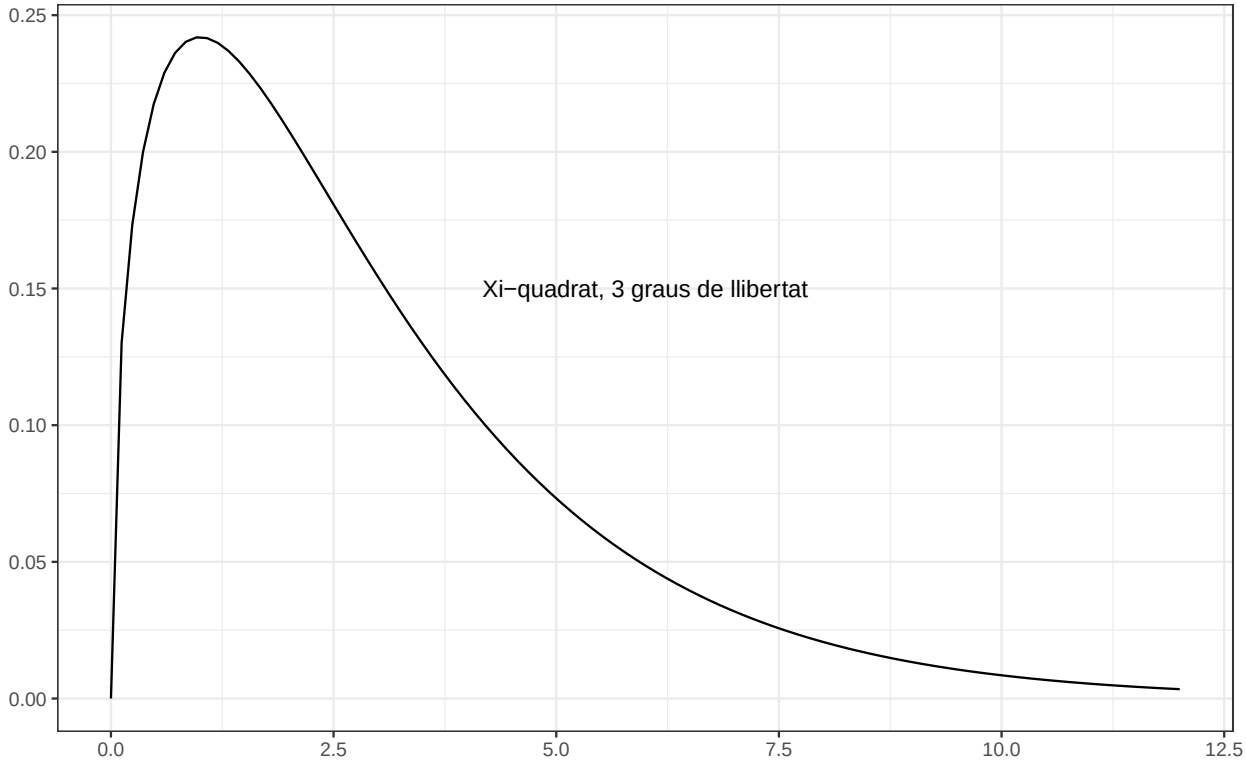
La distribució t és una distribució contínua que és molt semblant a una distribució normal, però les cues s'estenen més cap a fora (veieu figura). Aquesta distribució acostuma a sorgir en situacions en què les dades sembla que segueixen una distribució normal, però no coneixeu la mitjana o la desviació estàndard. Com és d'esperar pel què hem vist amb d'altres distribucions, les funcions rellevants a `R` són `dt()`, `pt()`, `qt()` i `rt()`.

A la figura està representada una distribució t amb 3 graus de llibertat (línia continua). És semblant a una distribució normal, però no és del tot igual. Per a fer la comparació hi ha dibuixada una distribució normal estàndard com a línia discontínua. Tingueu en compte que les "cues" de la distribució t són "més pesades" (és a dir, s'estenen més cap a fora) que les cues de la distribució normal. Aquesta és la diferència important entre les dos distribucions. Com més graus de llibertat tingui la distribució t, més semblant a una normal serà.



Xi-quadrat

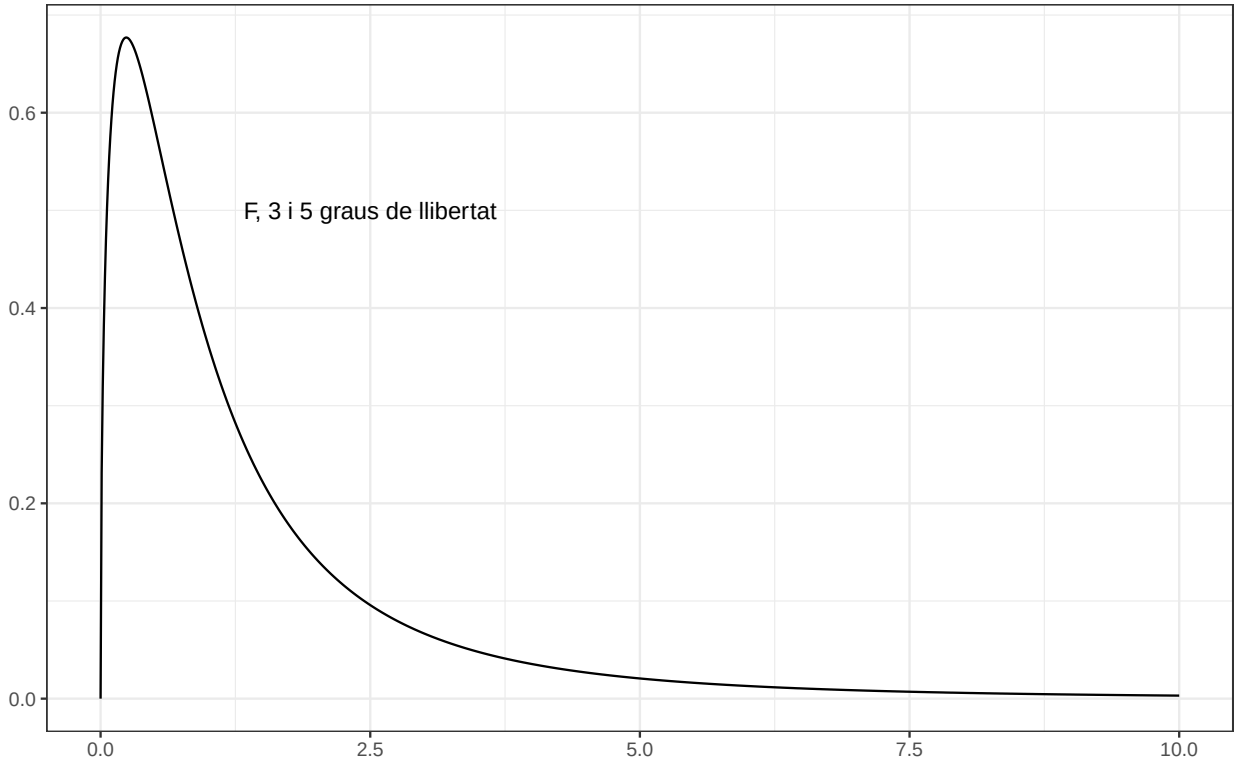
La distribució χ^2 (xi-quadrat) és una altra distribució que apareix a molts llocs. La situació en què la podem trobar és quan fem una anàlisi dades categòriques. La raó principal per la qual la distribució χ^2 apareix sovint és que, si teniu un grup de variables que es distribueixen normalment, eleveu els seus valors al quadrat i després sumeu aquests valors (procediment anomenat com a "suma de quadrats"), aquesta suma té una distribució χ^2. A la figura teniu l'aspecte d'una distribució χ^2. Una vegada més, les comandes a `R` per aquesta distribució són bastant previsibles: `dchisq()`, `pchisq()`, `qchisq()`, `rchisq()`.



A una χ^2 tots els valors són positius i, com es pot veure, té una forma prou esbiaixada.

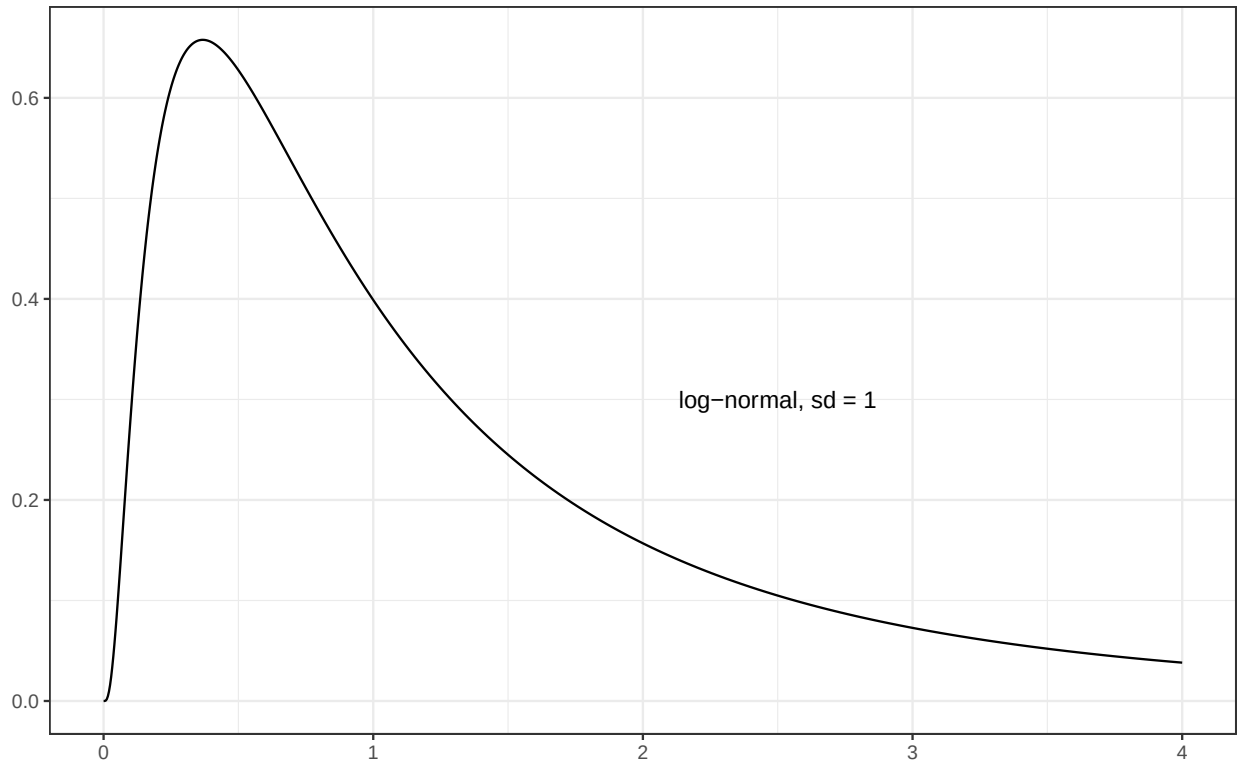
Distribució *F*

La distribució F s'assembla una mica a una distribució χ^2, i sorgeix sempre que necessiteu comparar dues distribucions χ^2 entre si. Com s'ha esmentat, la χ^2 resulta ser la distribució resultant quan prenem una "suma de quadrats". Si voleu comparar dues "sumes de quadrats" diferents, probablement estarem parlant d'alguna cosa que té una distribució F. A la figura teniu l'aspecte d'una distribució F. Les comandes d'`R` per la distribució F són `df()`, `pf()`, `qf()` i `rf()`.



Distribució log-normal

Aquesta distribució representa una variable de la què el seu logaritme segueix una distribució normal. Exemples de fenòmens que segueixen aquesta distribució són la pressió sanguínia, la durada de les partides d'escacs...



En resum, l'objectiu no és que tingueu una comprensió profunda de totes aquestes distribucions diferents, ni que recordeu les relacions precises entre elles. El més important és que entengueu la idea bàsica que totes aquestes distribucions estan profundament relacionades entre si i amb la distribució normal. Més endavant ens trobarem amb dades que es distribueixen normalment, o almenys se suposa que es distribueixen normalment. Si suposeu que les vostres dades es distribueixen normalment, no us hauria de sorprendre veure que les distribucions χ^2, t i F apareixen per tot arreu quan comenceu a fer el vostre anàlisi de dades.