Test d'hipòtesis i comparació de mitjanes

Pablo Sánchez
Tardor 2023

Test d'hipòtesi

El test d'hipòtesi és un grup de teories, mètodes i tècniques per comparar poblacions. S'utilitza habitualment en recerca científica i moltes indústries. Per exemple, una empresa pot tenir una teoria que augmentar el preu del seu producte augmentarà els ingressos, o què canviar el nom d'un lloc web pot augmentar el trànsit. Fins i tot podem utilitzar el test d'hipòtesi per analitzar si un medicament és eficaç en el tractament de condicions de salut específiques. No només s'està provant hipòtesis a tot el nostre voltant, sinó que també és una disciplina ben establerta. Els primers orígens es remunten al segle XVIII, quan l'anàlisi dels registres de naixement va mostrar que per cada naixement hi ha una probabilitat una mica més gran de ser home que dona.

En el test d'hipòtesi sempre comencem amb la suposició que no hi ha cap diferència entre les poblacions. Ho fem per reduir el risc d'introduir qualsevol biaix als nostres tests. Això s'anomena hipòtesi nul · la (H_0). Podem ampliar l'exemple de la proporció de naixements entre homes i dones per veure la influència dels suplements de vitamina C. La nostra hipòtesi nul · la podria ser que no hi ha cap diferència en la proporció del gènere en els naixements dels nadons de dones que prenen i dones què no prenen suplements de vitamina C. Aleshores creem una hipòtesi alternativa (H_1), que normalment pot adoptar una de dues formes. Podem dir que hi ha una diferència entre els naixements masculins i femenins entre les dones que prenen suplements de vitamina C i les que no ho fan. O podem explicitar la direcció de la diferència, per exemple, que la població que pren suplements de vitamina C té més naixements femenins que les que no prenen els suplements.

Hi ha moltes maneres de realitzar proves d'hipòtesis, però un flux de treball general és: en primer lloc, decidim de quines poblacions volem analitzar la diferència, en aquest cas dones adultes que fan servir o no suplements de vitamina C. Aleshores, desenvolupem hipòtesis nul · les i alternatives: que els naixements tenen la mateixa probabilitat de ser homes o dones en ambdues poblacions, o que els nadons són més propensos a ser nenes en dones que prenen suplements de vitamina C. Ara recollim les nostres dades mostrals. Concretament, apuntem el gènere dels nadons nascuts en ambdues poblacions. A continuació, realitzem proves estadístiques sobre les dades de la mostra. Finalment, utilitzem els resultats per extreure conclusions sobre la població que representa la mostra.

L'objectiu d'un test d'hipòtesi no és demostrar que la hipòtesi alternativa és (probablement) certa; l'objectiu és demostrar que la hipòtesi nul · la és (probablement) falsa.

Una manera de pensar-ho és imaginar que un test d'hipòtesi és un judici penal on s'està jutjant a la hipòtesi nul · la. La hipòtesi nul · la és l'acusada, la persona investigadora fent el test és el fiscal, i la pròpia prova estadística és el jutge. Igual que a un judici penal, hi ha una presumpció d'innocència: es considera que la hipòtesi nul · la és certa tret que l'investigador pugui demostrar més enllà de qualsevol dubte raonable que és falsa.

Abans d'entrar en detalls sobre com es construeix una prova estadística, és útil entendre la filosofia que hi ha darrere. Idealment ens agradaria construir la nostra prova per tal què mai cometem errors. Malauradament, com que el món és desordenat, això generalment no és possible. De vegades només teniu mala sort: per exemple, suposem que tireu una moneda 10 vegades seguides i surt cara les 10 vegades. Això sembla ser una evidència molt forta que la moneda està esbiaixada, però per descomptat, hi ha una possibilitat d'1 en 1024 que això passi encara que la moneda fos totalment equilibrada. Com a conseqüència, l'objectiu del test d'hipòtesis estadístiques no és eliminar els errors, sinó minimitzar-los.

En aquest punt, hem de ser una mica més precisos sobre què entenem per "errors". En primer lloc, diguem una obvietat: o bé la hipòtesi nul · la és certa, o bé és falsa; i la nostra prova rebutjarà la hipòtesi nul · la o l'acceptarà. Així, com il · lustra la taula següent, després d'executar la prova i fer la nostra elecció, podria haver passat una d'aquestes quatre coses:

	Acceptar H_0	Rebutjar H_0
H_0 és certa	Decisió correcta	error tipus I
H_0 és falsa	error tipus II	Decisió correcta

Com a conseqüència, en realitat hi ha dos tipus diferents d'error. Si rebutgem una hipòtesi nul · la que és realment certa, hem comès un error de tipus I. D'altra banda, si conservem la hipòtesi nul · la quan de fet és falsa, hem comès un error de tipus II.

Recordeu el símil comparant les proves estadístiques amb un judici? Si a un judici es requereix que s'estableixi "més enllà de qualsevol dubte raonable" que l'acusat va fer la cosa de què se'l acusa, totes les regles per aportar proves estan dissenyades per garantir que no hi hagi (gairebé) cap possibilitat de condemnar injustament un acusat innocent. El judici està dissenyat per protegir els drets d'un acusat (com a la frase "és millor que s'escapin deu culpables que què pateixi un innocent"). En altres paraules, un judici no tracta els dos tipus d'error de la mateixa manera i castigar els innocents es considera molt pitjor que deixar lliures els culpables. Amb una prova estadística passa pràcticament el mateix: el principi de disseny més important de la prova és controlar la probabilitat d'un error de tipus I, per mantenir-la per sota d'algun valor de probabilitat fix que considerem suficient. Aquesta probabilitat, que es denota α, s'anomena **nivell de significació** de la prova. Es diu que un test d'hipòtesi té un nivell de significació α si la taxa d'error de tipus I no és més gran que α. Per convenció, els valors d'α es fixen en 0.05, 0.1 o 0.001. Un nivell de significació d' α=0.05 vol dir que tenim un 5% de probabilitat de cometre un error de tipus I. És a dir, tenim un 5% de probabilitat de rebutjar la hipòtesi nul · la sent aquesta certa.

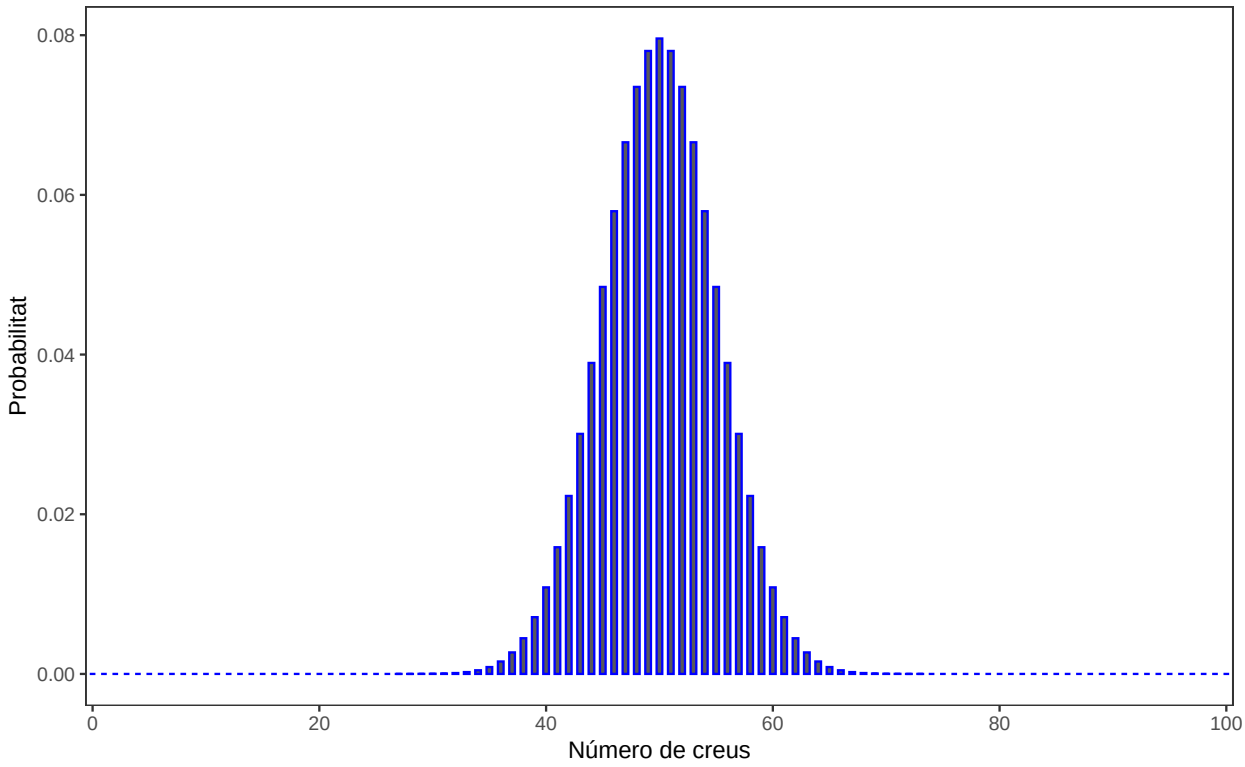
En aquest punt hem de començar a parlar de com es construeix un test d'hipòtesi. Per a això, pensem en un exemple senzill. Tenim una moneda i un amic ens diu que les monedes equilibrades no existeixen i que una moneda qualsevol sempre tindrà més números de cares o creus en un número prou gran de tirades. Òbviament, això no ens ho creiem, per què sabem que les monedes tenen una probabilitat de 0.5 de sortir cara o creu. Però la nostra creença no és suficient i no convencerem al nostre amic. Hem de demostrar-ho estadísticament. Ignorem les dades reals que podem obtenir simplement llençant la moneda de moment, i pensem en l'estructura de l'experiment que faríem per provar-ho. Independentment de quins són els números reals, la forma de les dades és que X de N tirades seran creu. A més, suposem de moment que la hipòtesi nul · la és realment certa: què totes les monedes estan equilibrades i la probabilitat de treure creu llençant una moneda sigui força propera al 50%. Per descomptat, amb el que ja hem vist fins ara, no esperem que aquesta probabilitat sigui exactament 0.5: si, per exemple, féssim 100 tirades i 53 ($X = 53$) de les tirades fossin creu, probablement ens veurem obligats a admetre que les dades són bastant coherents amb la hipòtesi nul · la. D'altra banda, si 99 de les tirades sortissin creu ($X = 99$), ens sentiríem bastant segurs que la hipòtesi nul · la és incorrecta. De la mateixa manera, si només 3 tirades fossin creu ($X = 3$), estaríem igualment segurs que la hipòtesi nul · la és incorrecte. Siguem una mica més tècnics sobre això: tenim una quantitat X que podem calcular mirant les nostres dades; després de mirar el valor d'X, prenem una decisió sobre si creiem que la hipòtesi nul · la és correcta o si rebutgem la hipòtesi nul · la a favor de l'alternativa. El nom d'aquesta cosa que calculem per guiar les nostres eleccions és l'**estadístic de contrast**.

Després d'haver escollit l'estadístic de contrast, el següent pas és indicar amb precisió quins valors de l'estadístic provocarien què és rebutgi la hipòtesi nul · la i quins valors farien que la mantinguem. Per fer-ho, hem de determinar quina seria la distribució de mostreig de l'estadístic si la hipòtesi nul · la fos realment certa. Aquesta distribució ens indica exactament quins valors d'X podríem esperar sota la nostra hipòtesi nul · la. Per tant, podem utilitzar aquesta distribució com a eina per avaluar fins a quin punt la hipòtesi nul · la coincideix amb les nostres dades.

Com determinem la distribució mostral de l'estadístic de contrast? Per a molts test d'hipòtesis, aquest pas és realment força complicat. Tanmateix, de vegades és molt fàcil. I, per sort per a nosaltres, el nostre exemple de la moneda ens proporciona un dels casos més fàcils. El que volem comprovar és si el nostre amic pot treure més cares que creus (probabilitat teòrica de 0.5), el nostre estadístic de contrast serà el nombre X de creus que ha tret d'una mida mostral d'N tirades. Hem vist una distribució com aquesta abans: així és exactament com es defineix la distribució binomial. Per tant, diríem que la hipòtesi nul · la prediu que X està distribuït binomialment, i s'escriu

$$X \sim Binomial(\theta, N)$$

Com que la hipòtesi nul · la estableix que $\theta = 0{,}5$ i el nostre experiment té $N = 100$ tirades de moneda, tenim la distribució de mostreig que necessitem. Aquesta distribució mostral es representa a la següent figura.



No hi ha sorpreses realment: la hipòtesi nul · la diu que $X = 50$ és el resultat més probable, i diu que gairebé segur que veurem entre 40 i 60 creus al nostre experiment.

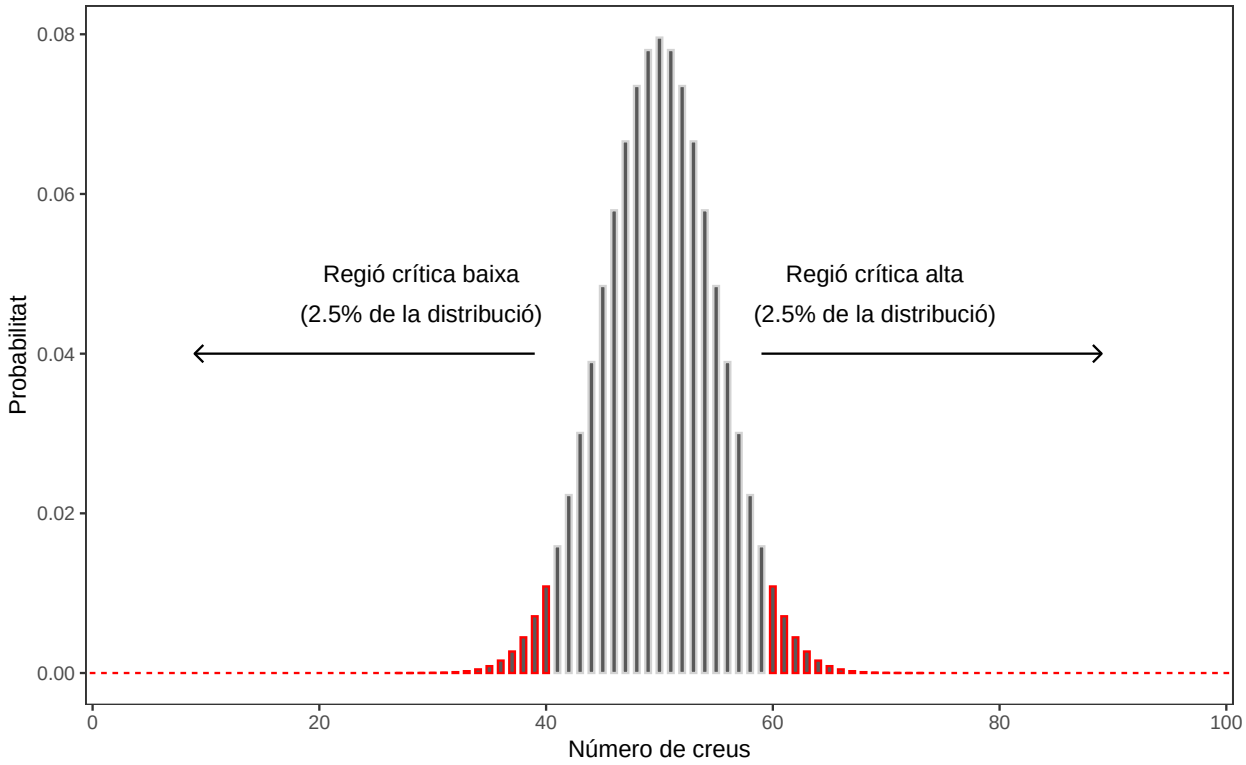
Presa de decisions

Hem construït un estadístic de contrast (X) i hem escollit aquest estadístic de tal manera que estem bastant segurs que si X és a prop de $N/2$, llavors hauríem d'acceptar la hipòtesi nul · la, i si no, l'hauríem de rebutjar. La pregunta que queda és la següent: exactament quins valors de l'estadístic de contrast hem d'associar amb la hipòtesi nul · la, i quins valors exactament van amb la hipòtesi alternativa? Per exemple, hem fet que el nostre amic llenci la moneda 100 cops i hem observat un valor de $X = 62$. Quina decisió he de prendre? Hem de triar creure la hipòtesi nul · la o la hipòtesi alternativa?

Per respondre a aquesta pregunta, hem d'introduir el concepte de regió crítica per a l'estadístic de contrast X. La regió crítica del test correspon a aquells valors de X que ens portarien a rebutjar la

hipòtesi nul · la (per això la regió crítica també és de vegades anomenada regió de rebuig). Com trobem aquesta regió crítica? Bé, considerem el que sabem:

- X ha de ser molt gran per rebutjar la hipòtesi nul · la.
- Si la hipòtesi nul · la és certa, la distribució mostral de X és Binomial(0.5,N).
- Si $\alpha = 0.05$, la regió crítica ha de cobrir el 5% d'aquesta distribució de mostreig.



Aquesta figura mostra la regió crítica associada al test d'hipòtesi que hem plantejat, amb un nivell de significació d'$\alpha = 0.05$. El mateix gràfic mostra la distribució mostral de X sota la hipòtesi nul · la: les barres grises corresponen a aquells valors de X per als quals mantindríem la hipòtesi nul · la. Les barres vermelles mostren la regió crítica: aquells valors de X per als quals rebutjaríem la hipòtesi nul · la. Com que la hipòtesi alternativa és bilateral (és a dir, permet tant $\theta < 0.5$ com $\theta > 0.5$, la regió crítica cobreix les dues cues de la distribució. Per garantir un nivell α de 0.05, hem d'assegurar-nos que cadascuna de les dues regions abasta el 2.5% de la distribució del mostreig.

És important assegurar-se d'entendre aquest darrer punt: la regió crítica correspon als valors de X per als quals **rebutjaríem la hipòtesi nul · la**, i la distribució mostral en qüestió descriu la probabilitat que obtinguéssim un valor particular de X si la hipòtesi nul · la fos realment certa. Ara, suposem que hem escollit una regió crítica que cobreix el 20% de la distribució del mostreig i suposem que la hipòtesi nul · la és realment certa. Quina seria la probabilitat de rebutjar incorrectament la hipòtesi nul · la? La resposta és, per descomptat, el 20%. I per tant, hauríem construït una prova que tingués un nivell α de 0.2. Si volem $\alpha = 0.05$, la regió crítica només pot cobrir el 5% de la distribució de mostreig del nostre estadístic de contrast.

La nostra regió crítica consta dels valors més extrems, coneguts com les cues de la distribució. Com a resultat, si volem $\alpha = 0.05$, aleshores les nostres regions crítiques corresponen a $X \leq 40$ i $X \geq 60$ És a dir, si el nombre de creus està entre 41 i 59, llavors hauríem d'acceptar la hipòtesi nul · la. Si el nombre està entre 0 i 40 o entre 60 i 100, hauríem de rebutjar la hipòtesi nul · la (o acceptar la hipòtesi alternativa). Els números 40 i 60 sovint s'anomenen valors crítics, ja que defineixen les vores de la regió crítica.

En aquest punt, el nostre test d'hipòtesi està essencialment complert: (1) escollim un nivell α (per exemple, $\alpha = 0.05$, (2) obtenim un estadístic de contrast (per exemple, X) que ens permet de comparar H_0 amb H_1, (3) esbrineu la distribució de mostreig de l'estadístic de contrast en el supòsit que la hipòtesi nul · la sigui certa (en aquest cas, binomial) i després (4) calculeu la regió crítica que produeix un nivell α adequat. (0-40 i 60-100). Tot el que hem de fer ara és calcular el valor de l'estadístic de contrast per a les dades reals (per exemple, $X = 62$) i després comparar-lo amb els valors crítics per prendre la nostra decisió. En aquest cas, 62 és més gran que el valor crític de 60, rebutjaríem la hipòtesi nul · la o, per formular-la una mica diferent, diem que la prova ha donat un **resultat significatiu**.

La diferència entre els tests unilaterals i bilaterals

Tornem a visitar l'exemple que acabem de fer. La hipòtesi nul · la ha quedat força clara:

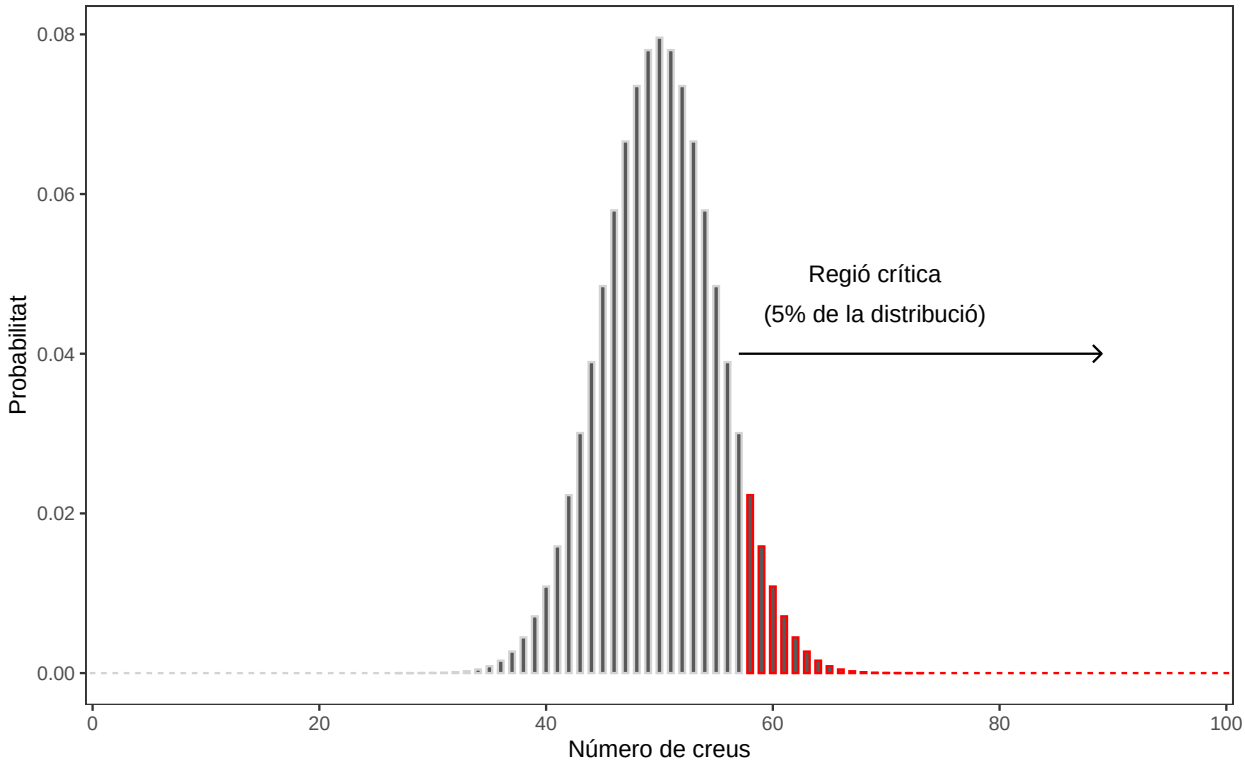
$$H_0 : \theta = 0.5$$

És a dir, la nostra hipòtesi nul · la, la què ha de ser demostrat que no és correcta, és que la moneda té la mateixa probabilitat de caure de cara que caure de creu. I com definiríem la nostra hipòtesi alternativa? En aquest cas, com que només volem testar si la probabilitat de creu és *més gran de* 0.5, aquesta prova serà *bilateral*. S'anomena així perquè la hipòtesi alternativa cobreix l'àrea dels dos "costats" de la hipòtesi nul · la i, com a conseqüència, la regió crítica de la prova cobreix les dues cues de la distribució de mostreig (2,5% a banda i banda si $\alpha = 0.05$) i matemàticament l'expressarem així:

$$H_1 : \theta \neq 0.5$$

Si per contra, el nostre amic ens hagués dit que les monedes sempre cauen més de creu què de cara, hauríem de reformular les hipòtesis nul · la i alternativa i tindríem un test *unilateral*. Les dues hipòtesis ara es formularien així:

$$H_0 : \theta \leq 0.5$$
$$H_1 : \theta > 0.5$$



En aquest cas, la hipòtesi alternativa és que $\theta > 0.05$, de manera que només rebutjaríem la hipòtesi nul · la per a valors grans de X. Com a conseqüència, la regió crítica només cobreix la cua superior de la distribució de mostreig; concretament el 5% superior de la distribució.

El p-valor

Definim el **p-valor** com la probabilitat d'observar els resultats de l'estudi o d'altres més allunyats de la hipòtesi nul · la, si la hipòtesi nul · la fos certa.

Si el p-valor compleix amb la condició de ser menor que un nivell de significació imposat arbitràriament (α), aquest es considera un resultat **estadísticament significatiu** i, per tant, permet rebutjar la hipòtesi nul · la.

Els p-valors quantifiquen quanta evidència hi ha per a la hipòtesi nul · la. Els p-valors grans ens porten a no tenir prou evidència per a acceptar la hipòtesi alternativa, mantenint-nos en canvi amb la suposada hipòtesi nul · la. Els p-valors petits ens fan dubtar d'aquesta hipòtesi original a favor de la hipòtesi alternativa.

Comparació de mitjanes

El test de la t de Student

Anem a posar en pràctica un test d'hipòtesis. Aquí, compararem els estadístics de contrast entre dos grups d'una variable. Al conjunt de dades `incidentsUSC1`, `Incidents` és una variable numèrica que representa el número d'incidents policials per dia. `diaSetmana` és una variable de tipus caràcter, una variable categòrica amb 7 nivells: Dilluns, Dimarts, Dimecres, Dijous, Divendres, Dissabte i Diumenge, que descriuen el dia de la setmana al que fa referència els recomptes d'incidents policials. Podem fer-nos

preguntes sobre les diferències en el número d'incidents entre diferents dies de la setmana. Per exemple, hi ha més incidents un dissabte què un divendres?

El càlcul de l'estadística descriptiva per a cada grup és senzill.

–> Calculeu la mitjana i la desviació estàndard d'incidents pels divendres i els dissabtes de la taula d'incidents policials

Amb aquestes informacions tendiríem a dir que un dia hi ha més incidents que l'altre, però no podem fer-ho sense fer un test estadístic per a fer el contrast d'hipòtesis. Podem fer servir diversos tipus de tests estadístics per a comparar mitjanes. El primer què veurem és el test de la t de Student. Aquest test el podem fer servir quan les nostres dades no són suficients per a tenir un estadístic que segueixi la distribució normal. El test de la t de Student ens permet comparar la mitjana de dos grups. Deixarem fora d'aquest manual tota la matemàtica associada a aquests tests, per a centrar-nos en les funcions de R que ens permeten fer-los i en la interpretació dels resultats.

De totes maneres, sí que citarem les assumpcions que aquest test fa sobre les nostres dades:

- **Normalitat**: El test assumeix que les nostres dades es distribueixen normalment.
- **Independència**: El test assumeix que les observacions s'han agafat de manera independent.
- **Homogeneïtat de les variàncies**: També coneguda com homoscedasticitat, aquesta assumpció diu que les desviacions estàndard d'ambdós grups són iguals.

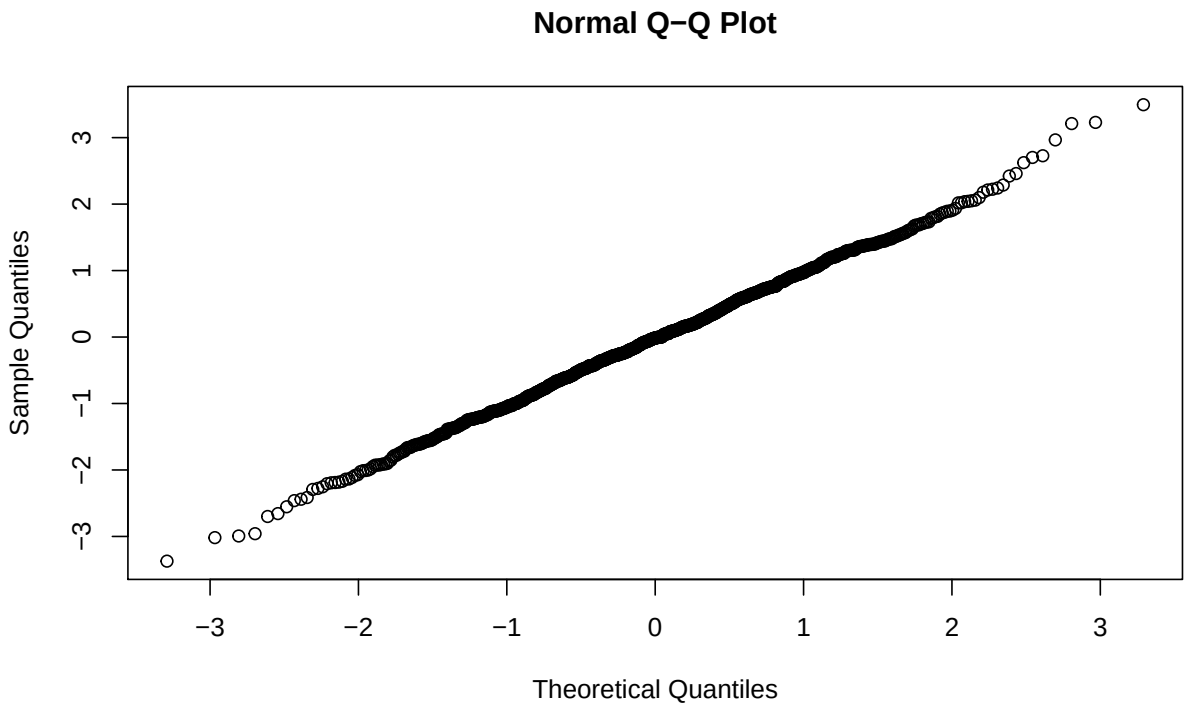
La tercera assumpció és important, però per ara la descartarem, ja què R, per defecte, fa servir una variant del test de la t de Student què es diu test de Welch. Aquest test ja es fa càrrec de les desviacions estàndards diferents.

La primer assumpció, però, és important i ens hem d'assegurar que les nostres dades es distribueixen de manera normal.

–> Visualitzeu les distribucions amb un histograma. Expliciteu què l'eix X estigui entre 40 i 200 incidents a les dues gràfiques. Quin tipus de distribució us sembla què segueixen les dades?

Podem anar més enllà per a comprovar la normalitat i fer servir una altra representació gràfica, els gràfics quantil-quantil o **QQ plots**. Aquests gràfics us permet comprovar visualment si observeu alguna infracció sistemàtica de la normalitat. En una gràfica QQ, cada observació es representa com un únic punt. l'Eix x és el quantil teòric en el qual hauria de caure l'observació si les dades es distribueixen normalment (amb la mitjana i la variància estimades a partir de la mostra). A l'eix y hi ha el quantil real de les dades dins de la mostra. Si les dades són normals, els punts haurien de formar una línia recta. Per exemple:

```r
set.seed(42)
unaDistribucioNormal <- rnorm(1000, 0, 1)
qqnorm(y=unaDistribucioNormal)
```



–> Feu servir la funció `qqnorm` per a representar el gràfic quantil-quantil de cadascún dels grups de dades que voleu comparar

–> Diríeu què són dades normals?

Doncs amb aquestes informacions estem en disposició de formular la nostra hipòtesi.

La hipòtesi nul · la és que la mitjana mostral dels dos grups (dissabte i divendres) és la mateixa, i la hipòtesi alternativa és que la mitjana del número d'incidents dels dissabtes és més gran que la dels divendres. Aquesta suposició la fem en base a què la mitjana dels dissabtes sembla ser més gran, però sense haver-la calculat prèviament, podríem tenir la creença que durant el cap de setmana hi ha més gent al carrer, per exemple, el què dona peu a més possibles interaccions que acabin en un incident policial. Podem escriure aquestes hipòtesis mitjançant equacions. μ representa una mitjana de població desconeguda i utilitzem subíndex per indicar a quin grup pertany aquesta mitjana mostral. Una manera alternativa d'escriure les equacions és comparar les diferències de mitjanes mostrals amb zero.

Hipòtesis

- H_0: El número mitjà d'incidents policials són **els mateixos** els dissabtes i els divendres.

$$H_0 : \mu_{Dissabte} = \mu_{Divendres}$$
$$H_0 : \mu_{Dissabte} - \mu_{Divendres} = 0$$

- H_1: El número mitjà d'incidents policials és **més alt** els dissabtes què els divendres.

$$H_1 : \mu_{Dissabte} > \mu_{Divendres}$$
$$H_1 : \mu_{Dissabte} - \mu_{Divendres} > 0$$

Com què sembla que les nostres dades assumeixen les condicions per a fer el test de la t de Student, (al menys normalitat i independència, la desviació estàndard l'hem deixat de banda per ara), podem fer servir la funció `t.test` del paquet `base` (no hem de carregar res).

La funció `t.test` té diversos argument. Pot acceptar només una mostra (per a fer un test de si el valor de la mitjana és igual a un valor proporcionat per l'usuari), o acceptar dues mostres i comparar la seva mitjana. Us deixo la comanda què heu de fer servir omplint els buits:

```
t.test(x = ______, y = _______, alternative = ________)
```

Una vegada obtingut el resultat, comentem-lo:

```
##
##  Welch Two Sample t-test
##
## data:  dissabte and divendres
## t = 3.3161, df = 280.29, p-value = 0.0005166
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  5.5361    Inf
## sample estimates:
## mean of x mean of y
##  120.4028  109.3819
```

- t: és el valor de l'estadístic de contrast.
- df: són els graus de llibertat. En un test t ordinari (no el test Welch), aquest número és el número d'observacions menys 1. Com més gran sigui aquest número, més semblant a una distribució normal.
- p-valor: és la probabilitat que comentem un error de tipus I (rebutjar la hipòtesi nul · la sent aquesta certa). Com més petit sigui, menys probabilitat de cometre aquest error.
- Hipòtesi alternativa: ens diu que en cas què el p-valor sigui més petit que el nostre valor de significació α, la diferència entre les mitjanes és més gran que zero.
- 95 percent confidence interval: Intèrval en el qual es trobaria la diferència de mitjanes el 95% de les comparacions.
- Mitjanes dels dos grups comparats.

El test *Shapiro-Wilk* per a comprovar la normalitat d'una mostra.

Ja hem vist que veure si les nostres dades estan distribuïdes normalment és molt important per a dur a terme un test de la t de Student, ja què és una de les seves assumpcions. El mateix passarà per a fer el següent tipus de test, l'anàlisi de la variància o ANOVA. Veure les distribucions gràficament, amb histogrames o QQ plots, és sempre una bona idea, però també tenim un test d'hipòtesi implementat en R per a comprovar formalment si unes dades segueixen la distribució normal.

Aquest test és el de **Shapiro-Wilk**. La hipòtesi nul · la d'aquesta prova és que la població es distribueix normalment. Així, si el p-valor és inferior al nivell α escollit, aleshores la hipòtesi nul · la es rebutja i hi ha evidència que les dades no es distribueixen normalment. D'altra banda, si el p-valor és més gran que el nivell α escollit, aleshores la hipòtesi nul · la (que les dades provenen d'una població distribuïda normalment) no es pot rebutjar (per exemple, per a un nivell α de 0.05, un conjunt de dades amb un p-valor inferior a 0.05 rebutja la hipòtesi nul · la que les dades provenen d'una població distribuïda normalment; en conseqüència, un conjunt de dades amb un p-valor superior al valor α de 0.05 no rebutja la hipòtesi nul · la que les dades provenen d'una població normalment distribuïda).

$$H_0 : Distribució(X) \text{ és } normal$$
$$H_1 : Distribució(X) \text{ no és } normal$$

Per exemple, per a una normal amb mitjana 0 i desviació estàndard 1, el test **Shapiro-wilk** amb un nivell α de 0.05 seria:

```
shapiro.test(unaDistribucioNormal)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  unaDistribucioNormal
## W = 0.99882, p-value = 0.767
```

Com què el p-valor és més gran de 0.05, no podem rebutjar la hipòtesi nul · la i per tant concloem què les nostres dades estan distribuïdes normalment.

–> Comproveu què el número d'incidents de divendres i dissabte estan distribuïts normalment tal i com veien amb els QQ plots.

Anàlis de la Variància (ANOVA)

L'anàlisi de la variància, malgrat el seu nom, ens permet comparar mitjanes igual que la t de Student. Però en aquest cas, ens permet comparar més de dos grups alhora. `R` proporciona una funció anomenada `aov()`, que és l'acrònim de "Analysis Of Variance" (anàlisi de la variància). Si escriviu `?aov` i mireu la documentació d'ajuda, veureu que hi ha diversos arguments per a aquesta funció, però els únics dos que ens interessen per ara són `formula` i les `data`. L'argument de la fórmula és el que feu servir per especificar la variable que voleu contrastar en funció de la variable què codifica com les voleu agrupar, i l'argument de dades és el que feu servir per especificar la taula de dades que emmagatzema aquestes variables.

Per exemple, aquí intentarem comparar les mitjanes dels incidents policials de divendres, dissabte i diumenge alhora. Les hipòtesis serien:

$$H_0 : \mu_{Dissabte} = \mu_{Divendres} = \mu_{Diumenge}$$
$$H_1 : \mu_{Dissabte} \neq \mu_{Divendres} \neq \mu_{Diumenge}$$

Les assumpcions de l'ANOVA són les mateixes què les de la t de Student:

- Independència
- Normalitat
- Homoscedasticitat

–> Comproveu la normalitat de les observacions d'incidents per als diumenges

Aquí ja no podem passar de comprovar la homogeneïtat de les variàncies. Per a això farem servir el test de Levene (`leveneTest`) del paquet `car`. El test de Levene fa el següent contrast d'hipòtesis:

$$H_0 : \sigma_{Dissabte} = \sigma_{Divendres} = \sigma_{Diumenge}$$
$$H_1 : \sigma_{Dissabte} \neq \sigma_{Divendres} \neq \sigma_{Diumenge}$$

en el qual la hipòtesi alternativa és què al menys un dels grups a comparar té una variància diferent. Llavors, al test de Levene li hem de donar com arguments primer la variable de la qual volem comparar la variància i en segon lloc la variable d'agrupació.

–> Feu un subset del *data.frame*`incidentsUSC1` que inclogui només els dies divendres, dissabte i diumenge. Anomeneu el nou *data.frame*`incidentsFinde`. NOTA: si volem especificar més d'un grup a l'opció `subset` podem fer-ho separant cada crida a un grup amb l'operador `|`

–> Ara carregueu el paquet `car`. Llegiu el manual de la funció `leveneTest` i feu-la servir sobre la nova taula de dades, indicant el dia de la setmana com a variable d'agrupació.

Ara què hem comprovat que la variància no és diferent per als tres dies, podem fer l'anàlisi de la variància amb la funció `aov()`. Com ja hem dit abans, l'objectiu d'aquest curs no és fer una demostració ni un desenvolupament matemàtic del funcionament de l'anàlisi de la variància, sino donar unes pinzellades de quan aplicar-lo, com fer-ho en R i com interpretar els resultats. Podeu trobar nombrosos recursos en línia si voleu més informació.

La funció `aov()` necessita una fórmula que especifica la variable resposta en funció de la variable agrupació (similar al que hem vist amb el `leveneTest`). La sintaxi de la fórmula és:

```
variableResposta ~ variableAgrupació
```

Després hem d'introduir coma a argument el nom de la taula de dades. D'aquesta manera, no hem de especificar els diferents vectors amb la notació `$`.

–> Feu un ANOVA dels incidents de la taula `incidentsFinde` en funció de la variable agrupació `diaSetmana`. El podeu interpretar?

Com veieu, el resultat ens dona informació de la suma de quadrats (recordeu la distribució χ^2), graus de llibertat i residus, però no ens dona cap informació amb la que puguem acceptar o rebutjar la hipòtesi nul · la.

–> Torneu a cridar la mateixa comanda, però bolqueu el resultat a un objecte què es digui `aovFinde`. Després apliqueu la funció `summary` a aquest nou objecte. Recordeu què `summary` era una funció genèrica i què el seu comportament varia en funció de l'objecte que li passem com a argument.

–>Fixem-nos en el valor de probabilitat (`Pr`). Acceptem o rebutgem la hipòtesi nul · la?

Això està molt bé, però nosaltres tenim la mitjana d'incidents de 3 dies diferents i volem saber com es comparen entre ells dos a dos, no tenir un p-valor per a acceptar o rebutjar la hipòtesi nul · la. R té una funció en el paquet `stats` (actiu per defecte) què ens permet fer aquestes comparacions dos a dos. La funció és `TukeyHSD`, què necessita com a argument un objecte de tipus `aov`.

–> Comproveu amb la funció `class` quin tipus d'objecte és `aovFinde`. Després apliqueu la funció `TukeyHSD`. Quins dies tenen una mitjana d'incidents significativament diferents entre ells?

Una manera gràfica de veure aquests grups és fent servir la funció `boxplot` que ja coneixeu. Si us fixeu bé en la documentació, aquesta funció també accepta una fórmula com a argument. És a dir, que si posem que volem fer la representació gràfica dels incidents en funció del sia de la setmana, tal i com ho vam fer a la funció `aov()`, aconseguirem un boxplot amb una caixa per a cada dia.

–> Feu un boxplot per al data.frame `incidentsFinde`, representant els incidents en funció del dia de la setmana. Modifiqueu els paràmetres gràfic per arreglar els títols dels eixos i poseu un títol a la figura. L'argument `col` accepta un **factor**, de manera que ens permetrà assignar un color diferent a cada nivell d'aquest factor. Quin faríeu servir? Proveu-ho.

Tests no paramètrics

Quan un contrast d'hipòtesi és no paramètric volem dir que la prova no suposa que les vostres dades provenen d'una distribució determinada, com era el cas del test t de Student o l'ANOVA. La prova de Kruskal Wallis (H) és l'alternativa no paramètrica a l'ANOVA què hem vist fins ara, substituint les dades per categories. La prova H s'utilitza quan no es compleixen els supòsits d'ANOVA (com el supòsit de normalitat).

La prova determina si les medianes de dos o més grups són diferents. Com la majoria de proves estadístiques, calculeu un estadístic de contrast i la compareu amb un punt de tall de distribució. El test de Kruskal Wallis planteja les següents hipòtesis:

$$H_0 : \textit{ les medianes de les poblacions son iguals}$$

$$H_1 : \textit{ les medianes de les poblacions no son iguals}$$

El test de Kruskal Wallis us dirà si hi ha una diferència significativa entre els grups. Tanmateix, no us dirà quins grups són diferents. Per això, haureu de fer una prova *Post Hoc*, equivalent al test de Tukey què vam fer amb l'ANOVA.

Tornem al nostre conjunt de dades d'incidents policials per la USC1 dels anys 2021, 2022 i 2023. Volem saber si la mediana d'incidents dels 3 anys és igual o hi ha diferències.

La hipòtesi nul · la és que la mediana dels tres grups (2021, 2022 i 2023) és la mateixa, i la hipòtesi alternativa és que la mediana d'incidents dels tres anys per la USC1 és diferent. Potser a l'any 2023 i 2022 és més gran que al 2021. Aquesta suposició pot venir de què pensem que al 2021 encara hi havia estat d'alarma per la COVID-19 i que no hi havia tant moviment de gent.

–> Calculeu la mitjana, la mediana i la desviació estàndard del nombre d'incidents policials de la USC1 pels anys 2021, 22 i 23.

–> Comproveu la normalitat del nombre d'incidents policials pels anys 2021, 2022 i 2023. Feu-ho analíticament i gràficament. Segueixen tots els anys una distribució normal?

–> Feu una representació gràfica de les dades d'incidents dels tres anys

Aquestes dades presenten una clara tendència, però hem de formalitzar-ho amb un test estadístic. Com què no totes les dades es distribueixen normalment, hem d'aplicar el test no paramètric de Kruskal Wallis. El test a `R` està implementat com a `kruskal.test`, i té la mateixa sintaxi que l'ANOVA: accepta com arguments una fórmula amb la variable resposta i la variable agrupadora, així com la taula de dades a la què fan referència.

–> Feu el test de Kruskal Wallis comparant les medianes del nombre d'incidents per any. Bolqueu el resultat en un nou objecte què es digui `incidentsAny`.

Com passava amb l'ANOVA tenim un p-valor què ens diu que hi ha diferències significatives entre els tres anys, però no sabem entre quins comparats dos a dos. Per a resoldre-ho, tenim el test de Dunn, què a 'R' està implementat com a `dunn_test`. Aquest test, com ja hem vist a diferents tests prèviament, requereix que especifiquem la taula de dades amb la que treballarem i la fórmula relacionant la variable resposta amb la variable agrupadora. Addicionalment podem especificar un mètode d'ajust del p-valor per comparacions múltiples (per defecte fa servir el mètode Holm).

–> Apliqueu el test de Dunn a l'objecte `incidentsAny`. Canvieu el mètode d'ajust del p-valor al mètode `bonferroni`. Es confirma la tendència què vèiem gràficament?

Les comparacions múltiples sorgeixen quan una anàlisi estadística implica múltiples proves estadístiques simultànies, cadascuna de les quals té el potencial de produir un "descobriment". Un nivell de confiança establert generalment només s'aplica a cada prova considerada individualment, però sovint és desitjable tenir un nivell de confiança per a tot un grup de proves simultànies. No compensar les comparacions múltiples pot tenir conseqüències importants en el món real. Per exemple, suposem que considerem l'eficàcia d'un fàrmac en termes de reducció de qualsevol dels diversos símptomes d'una malaltia. A mesura que es consideren més símptomes, és cada cop més probable que el fàrmac sembli provocar una millora respecte als fàrmacs existents pel que fa a almenys un símptoma. A mesura que augmenta el nombre de comparacions, és més probable que els grups que s'estan comparant semblin diferir pel que fa a almenys un atribut. La nostra confiança que un resultat es generalitzarà a dades independents hauria de ser generalment més feble si s'observa com a part d'una anàlisi que implica múltiples comparacions, en lloc d'una anàlisi que només implica una comparació única.

Així, encara que nosaltres només estem comparant 3 anys, si estiguéssim comparant-ne més, simplement per atzar podríem estar rebutjant la hipòtesi nul · la en alguna d'aquestes comparacions. La correcció ens dona un p-valor ajustat (`p.adj`) què en general és més alt que el què haguéssim obtingut sense fer aquestes correccions.