OINARRIZKO ESTATISTIKA:
R PRAKTIKAK

MARÍA MERINO MAESTRE
USUE MORI CARRASCAL

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

CIP. Unibertsitateko Biblioteka

Merino Maestre, María

Oinarrizko estatistika [Recurso electrónico] : R praktikak / María Merino Maestre, Usue Mori Carrascal. – Datos. - Bilbao : Universidad del País Vasco / Euskal Herriko Unibertsitatea, Argitalpen Zerbitzua = Servicio Editorial, [2017]. – 1 recurso en línea : PDF (228 p.).
Modo de acceso: World Wide Web
ISBN: 978-84-9082-636-2.

1. Estadística - Problemas y ejercicios. 2. R (Lenguaje de programación). I. Mori Carrascal, Usue, coaut.

(0.034)519.2(076)
(0.034)681.3.06R

María Merino Maestre
maria.merino@ehu.eus
http://www.ehu.es/mae/html/prof/Maria.html
Matematika Aplikatua eta Estatistika eta Ikerkuntza Operatiboa
Zientzia eta Teknologia Fakultatea

Usue Mori Carrascal
usue.mori@ehu.eus
http://www.sc.ehu.es/ccwbayes/members/umori/home/
Matematika Aplikatua eta Estatistika eta Ikerkuntza Operatiboa
Zientzia eta Teknologia Fakultatea

UPV/EHUko Euskara Zerbitzuak sustatua eta zuzendua, Euskarazko ikasmaterialgintza sustatzeko deialdiaren bitartez.

ISBN: 978-84-9082-636-2

Aurkibide orokorra

Hitzaurrea

Zalantza barik, arlo anitzetan agertzen den interes handiko irakasgaia da Estatistika. Are gehiago, jakintza-arlo guztietan hartzen da aintzat: Zientziak, Osasun Zientziak, Ingeniaritza eta Arkitektura, Gizarte eta Lege Zientziak, eta Giza Zientziak, hain zuzen ere.

Gaur egun, gauza jakina da ezinbestekoa dela softwarea erabiltzea heziketa-prozesuan. Ikasmaterial honetan, Estatistika lantzeko hamazazpi praktika adierazten dira *R* programaren bidez. *R* da analisi estatistiko eta grafikoetara bideraturiko programa eta programazio-hizkuntza. Proiektua irekia eta doan banatzen dena da (GNU General Public Licence deritzon litzentziak ezarritako irizpideen arabera), eta Linux, Windows eta Macintosh sistemetarako aurki daiteke. Aucland Unibertsitateko Ross Ihaka eta Robert Gentleman irakasleek sortu zuten 1992. urtean; egun, *R*-ren Garapenerako Talde Nuklear («R Development Core Team» ingelesez) izenarekin ezagutzen diren estatistikan aditu batzuk, «developers» deritzenak, arduratzen dira programaren banaketaz eta garapenaz. Gero eta maizago erabiltzen ari den programa da, bai irakaskuntza-mailan (ASA saria jaso du), bai ikerkuntzarako, eta bai enpresa-arloan ere. Izan ere, *R* edozein erabiltzailek heda dezake haiek sortutako *pakete* edo eraskinen bidez. Ikasmaterial hau idazteko momentuan, eskuragai dauden pakete kopurua ia ia 10.000koa da, bere hazkundea esponentziala izanik. Programaren inguruko hasierako ikaskuntza/ezagutza fasea gainditu ostean, *R*-ren erabilera erraza da. Gainera, *Rcmdr* paketea instalatuz, interfaze grafiko bat erabiltzeko aukera dugu, programa komertzialen itxura lortuz. EHUko Medikuntza Fakultateko José Ramón Ruedari eta Borja Santosi esker eskuragarri dugu, egun, *Rcmdr* paketearen euskarazko bertsio bat, eta, ikasmaterial honetan, bertsio hori erabiliko dugu.

R-ri buruzko bibliografia `http://www.r-project.org/doc/bib/R-publications.html` helbidean kontsulta daiteke; bereziki, merezi du honako dokumentazio hau aipatzea: [1, 5, 7, 9, 11, 12, 14, 15, 19, 28, 29]. Nahiz eta Estatistika irakasgaiari buruzko material ugari egon, tamalez, urria da euskarazko bibliografia, bereziki aplikazio informatikoei buruzkoa. Materiala azken ikasturteetan osatu dugu, Zientzia eta Teknologia Fakultatean Estatistikako irakasle-lanetan aritu garen bitartean. Honela, Euskara eta Etengabeko Prestakuntzaren arloko Errektoreordetzaren Sare argitalpenean azaltzen diren *Estatistika: SPSS praktikak* (`http://testubiltegia.ehu.es/Estatistika-spss-praktikak`) eta *Estatistika: R praktikak* (`http://testubiltegia.ehu.es/Estatistikak-r-praktikak`) ikasmaterialen jarraipen moduan aurkezten da. Gainera, lehenengo atalean, aipatutako bi ikasmaterial horien egitura berbera mantendu dugu, *SPSS* software komertzialetik *Rcmdr* paketerako bidea errazteko. Bigarren atalean, *R* programa irekiaren erabilera zabalago bat eskaintzen da, eranskinean eduki osagarri gehiago erakutsiz.

Estatistika arduratzen da datuen bilketaz, antolaketaz eta interpretazioaz. Horretarako, arlo guztiak jorratzen ditu: datu-bilketaren planifikazioatik hasita, esperimentuen diseinuan eta laginketan ere parte hartzen du. Iragarpenak egiteko ere erabil daitezke estatistikan oinarritutako ereduak, eta gaur egun arlo askotan aplikatzen dira. Populazio bat aztertu nahi denean, askotan, ale guztiak aztertzea posible ez denez, populazio osoaren azpimultzo adierazgarri bat hartzen da, lagin bat. Lagin hori aztertzerako garaian, egungo ordenagailuei esker, lehen eskuz eginezinak ziren metodoak erraz aplikatu ahal ditugu gaur egun. Hori horrela, laginetik ateratako informazioa erabilita, bi motatako analisi estatistikoak egin daitezke: deskribatzailea eta inferentziala.

Lehenengoa, estatistika deskribatzailea, 1. praktikan lantzen da. Datuen laburpena egiteko erabiltzen da; adibidez, esperimentu baten emaitzen azterketa eta deskribapen orokor bat egiteko. Datuen deskribapena bai grafikoki bai zenbaki bidez egin daiteke, horretarako batezbestekoak, desbideratze estandarrak, ehunekoak eta maiztasunak erabiliz, besteak beste. Bigarrena, inferentzia estatistikoa, hirugarren praktikatik aurrera garatu eta landuko dugu. Arlo honen funtsezko teoria, estatistika matematikoa izenekoa, probabilitatearen teorian oinarritzen da. Probabilitatearen teoria XVII. mendean jaio zela esaten da, zorizko jokoen inguruko Blaise Pascal eta Pierre de Fermat matematikarien arteko posta-trukeari esker. Inferentziaren oinarria izanik, probabilitatearekin lotutako kontzeptu batzuk birpasatzen ditugu 2. praktikan, eta teorema batzuen emaitzak ulertzen saiatuko gara grafikoen bidez eta esperimentazioaren bidez. Ondoren, inferentzia estatistikoa arloan zentratuko gara. Arlo honen helburu nagusia da, lagin batzuen informazioa ezaguna izanik, populazio osorako ondorioak ateratzea. Zehazki, jasotako datuak zorizko behaketak direnean, populazio osorako inferentziak egiteko erabil daitezke. Batzuetan, datuen hainbat ezaugarri zenbatetsi nahi izaten dira, eta, beraz, 3. praktikan konfiantza-tarteak agertzen dira. Beste batzuetan, datuetan oinarrituta zehaztutako hipotesi bat baztertzen den ala ez aztertu nahi izaten da, eta, horregatik, hipotesi-kontrasteak 4., 5. eta 6. praktiketan aztertzen dira. Eredugintza estatistikoari buruzko lehen zertzeladak 7. praktikan azaltzen dira, bereziki, aldagaien arteko erlazioak eta ereduak aztertuz, korrelazio-analisien eta erregresioaren bidez. Azkenik, kalitatearen kontrol estatistikoa egiteko grafiko interesgarri batzuk lantzen dira 8. praktikan.

Ikasmaterial honek Estatistikaren oinarrizko kontzeptuak jorratzen ditu R softwarearen bidez. Hiru ataletan dago antolatuta; lehenengo eta bigarren ataletan, bederatzi gai lantzen dira berrogeita hamar adibide baino gehiago erabiliz. Zehazki, **I. atalean**, *Rcmdr* paketea erabiliz, zortzi *Rcmdr praktika* azaltzen dira. Atal hau oinarrizko ikastaroei zuzendua dago, batik bat. **II. atalean**, R programazio lengoaia erabiliz, bederatzi *R praktika* garatu ditugu, eduki konplexuagoetara atea irekiz. Noski, praktika bakoitzaren hasieran, helburuak azaldu ditugu laburki; ondoren, kontzeptuak birpasatu ditugu hainbat adibideren laguntzaz. **III. atalean** lau eranskin gehitu ditugu. A eranskinean, ikasmaterial honetan erabilitako datu-baseak azaltzen dira, eta fitxategiak `https://ehubox.ehu.eus/index.php/s/6C3gBHFiTME7OTx` web orrialdean jarri ditugu eskuragai. B eranskinean, erabilitako R paketeen laburpena egin dugu. C eranskinean, bederatzi gaien autoebaluaziorako hogeita hamar ariketa proposatu ditugu, eta D eranskinean azalduko dira horien emaitzak. Ariketa horien emaitzak lortzeko *Rscript* fitxategiak ere aurretik aipatutako web orrialdean daude eskuragai. Azkenik, ikasmaterialaren bukaeran, bibliografia dago.

I. praktikak: *Rmcdr* paketea erabiliz

0. ***Rmcdr* praktika.** *Rcmdr* paketearen sarrera. Instalazioari eta datu-baseen erabilerari buruzko azalpenak topa ditzakegu.

1. ***Rmcdr* praktika.** Estatistika deskribatzailea jorratzen da aldagai baterako eta birako: alde batetik, maiztasun-taulak nola adierazi; bestalde, estatistikoen kalkulua nola gauzatu eta, azkenik, grafiko batzuk egiteko jarraitu beharreko pausoak azaltzen ditugu.

2. ***Rmcdr* praktika.** Probabilitatearen teoriaren banaketen kontzeptu batzuk birpasatzen ditugu; esate baterako, zorizko aldagai jarraien eta diskretuen oinarrizko koantilak, funtzioak, grafikoak eta zorizko laginak nola sortu azaltzen da.

3. ***Rmcdr* praktika.** Hasiera ematen dio inferentzia estatistikoari, populazio-parametro ohikoak zenbatesteko konfiantza-tarteen kalkulua eta interpretazioa nola egin azalduz.

4. ***Rmcdr* praktika.** Hipotesi-kontraste parametrikoak lantzen dira, hau da, batezbestekoak, proportzioak eta bariantzak aztertzeko eta konparatzeko test estatistikoak.

5. ***Rmcdr* praktika.** Hipotesi-kontraste ez-parametriko batzuen erabilera birpasatzen da. Bereziki, doikuntza-egokitasunerako probak, independentzia- eta homogeneotasun-probak.

6. ***Rmcdr* praktika.** Bariantza-analisia garatzen da. Bereziki, faktore bakarreko kasua azaltzen da. Bariantzen berdintasunerako probak, ANOVA izeneko taulak, eta konparaketa anizkoitzak egiteko metodoak azaltzen dira.

7. ***Rmcdr* praktika.** Erregresioa lantzen da; zehazki linealak edo linealizatu daitezkeen erregresio bakunak konparatzen dira: hiperbolikoa, potentziala, esponentziala eta koadratikoa. Iragarpenak nola egin ere aipatzen da.

II. praktikak: *R* lengoaian programatuz

0. ***R* praktika.** *R* programaren sarrera. *R* eta *Rstudio*ren instalazioa nola egin azaltzen da. Komando batzuk, objetu mota ohikoenak, eragile logikoak eta zikloei buruzko oinarriak agertzen dira.

1. ***R* praktika.** Aldagai baterako eta bitarako estatistika deskribatzailea lantzen da: maiztasun-taulak, hainbat grafiko mota, eta estatistikoen zerrenda zehaztua erakutsiz. Azkenik, normaltasuna aztertzeko metodo deskribatzaileak ere aipatzen dira.

2. ***R* praktika.** Probabilitatearen teoriaren eta zorizko aldagaien banaketekin lan egiteko oinarrizko funtzioak, koantilak, grafikoak eta laginak zoriz sortzeko era azaltzen dira. Gainera, teorema batzuk enpirikoki eta grafikoki aztertzen dira.

3. ***R* praktika.** Hasiera ematen dio inferentzia estatistikoari populazio-parametro ohiko eta ez hain ohikoak zenbatesteko konfiantza-tarteak kalkulatuz eta interpretatuz, eta laginaren tamaina ere kalkulatuz.

4. ***R* praktika.** Hipotesi-kontraste parametriko ohiko eta bereziagoak nola gauzatu azaltzen da; hau da, batezbestekoak, proportzioak eta bariantzak aztertzeko eta konparatzeko test estatistikoak. Bi motatako erroreen kalkulua, testaren ahalmena eta laginaren tamainarena aztertzen dira.

5. ***R* praktika.** Hipotesi-kontraste ez-parametriko batzuen erabilera birpasatzen da. Doikuntza-egokitasunerako probak (Pearsonen khi karratu probak, normaltasuna aztertzeko hainbat test, eta Box-Cox-en transformazioa); independentzia- eta homogeneotasun-probak; zorizkotasun-probak; eta, azkenik, populazioak konparatzeko proba ez-parametrikoak lagin bakar baten kasuan, eta bi lagin edo gehiagotarako.
6. ***R* praktika.** Bariantza-analisia garatzen da faktore bakarrerako eta bi faktoretarako, interakzioarekin edo interakziorik gabe. Bariantzen berdintasunerako probak, ANOVA izeneko taulak, eta konparaketa anizkoitzak egiteko metodoak erakusten dira.
7. ***R* praktika.** Korrelazioa eta erregresioa gaiak lantzen dira. Alde batetik, erregresio bakuna zein anizkoitza: lineala edo linealiza daitezkeen ereduak. Bestalde, parametroak nola zenbatetsi, eredu osoa eta koaldagai bakoitzaren adierazgarritasuna nola aztertu, korrelazioa burutzea, diagnosia nola egin, iragarpenenak zenbatestea, grafikoki adieraztea erregresio-kurba eta iragarpenak azaltzen dira.
8. ***R* praktika.** Kalitate-kontrolaren oinarrizko grafikoak kalkulatzeko eta interpretatzeko bidea ikusten da.

III. Eranskinak

A eranskina. Datu-base guztien laburpena.

B eranskina. Erabilitako *R* paketeen azalpena.

C eranskina. Autoebaluaziorako ariketak.

D eranskina. Autoebaluaziorako ariketen emaitzak.

Osatutako materiala lagungarria izan daiteke edozein fakultate eta eskolatan, estatistikarekin lotutako irakasgaietan lantzeko; esate baterako, Estatistika Deskribatzailea, Inferentzia Estatistikoa, Matematika II eta Estatistika, Estatistika Aplikatua, Bioestatistika, Estatistika, Matematika eta Estatistika, besteak beste.

Azkenik, eskerrak eman nahi dizkiogu UPV/EHUko Euskara eta Etengabeko Prestakuntzaren arloko Errektoreordetzari, ikasmaterialgintzako proiektu honen hizkuntza egokitzeko emandako laguntzagatik.

Leioan, 2017ko urtarrilean.

Usue eta María

I. atala

Praktikak: *Rcmdr* paketea erabiliz

0. *Rcmdr* praktika

Sarrera

Helburua

Lehenengo praktika honek helburu bikoitza du. Alde batetik, *R* programa instalatzeko eta erabiltzen hasteko oinarrizko pausoak azalduko dira, *Rcmdr* paketean oinarrituz. Bestetik, *Rcmdr*n datu-baseen irakurketa nola egin daitekeen komentatuko da, eta baita oinarrizko datu-baseen kudeaketa nola gauzatu ere.

Kontuan izan hurrengo kapituluetan erabiliko diren datu guztiak `https://ehubox.ehu.eus/index.php/s/6C3gBHFiTME7OTx` orrialdean daudela eskuragai.

0.1. Instalazioa

R open source code edo kode ireki motako software estatistikoa instalatzeko, jo honako helbide honetara:

`http://cran.es.r-project.org`

eta, adibidez, Windows sistema operatiboan instalatzeko, jarraitu honako pauso hauek (bertsio berririk eskuragai baldin badago, jaitsi azkena):

`Download R for Windows`

`base` → `download R.3.3.2 for Windows`

Azkenik, programa instalatzeko, exekutatu deskargatutako fitxategi exekutagarria.

Era antzekoan, *R* Linux edo (Mac) OS X sistema operatiboetan instalatzeko aukera ere badago.

0.2. *Rcmdr* paketea

Esan bezala, *R* programazio-lengoaia bat da, eta bere ohiko erabilera kode bidezkoa da. Alabaina, *R*ko oinarrizko komandoak erabili beharrean, praktiken ikasmaterial honen lehenengo atalean *Rcmdr* izeneko paketea erabiliko dugu gehienbat; izan ere, pakete honek *R*-ren interfaze grafiko bat eskaintzen digu, eta aukera ematen digu komandorik ezagutu gabe analisi batzuk gauzatzeko.

Beraz, *Rcmdr* paketea instalatzeko:

`Paquetes` → `Instalar Paquetes` → `Spain(Madrid)` → `Rcmdr`

Prozesu hori behin baino ez da egin behar.

Gainera, badago *Rcmdr*-ren euskarazko bertsio bat gaur egun. Aktibatzeko, erabili nahi dugun bakoitzean honako agindu hauek exekutatu behar ditugu R-ren terminalean edo orrialde nagusian:

```
> Sys.setenv(LANG = "eu")
> library(Rcmdr)
```

Mac OS X motako sistema eragilea duzuenok instalatu XQuartz programa *Rcmdr* instalatu baino lehen: http://xquartz.macosforge.org.

Informazio gehiagorako:

http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/installation-notes.html

Hemendik aurrera, liburuaren lehen atal honetan, *Rcmdr* paketean oinarrituz emango ditugu azalpenak. Hala ere, jarduera eta analisi hauek guztiak komando bidez nola egin daitezkeen ikasteko, dokumentu honen bigarren atalera jo dezakegu.

Gainera, *R*-ren komando bidezko erabilera aurreratuan interesatuta dagoen irakurleak argibide gehiago aurki ditzake honako web orrialdeetan:

- Quick R: `http://www.statmethods.net/`
- J. Baronen laburpena: `http://www.psych.upenn.edu/%7Ebaron/refcard.pdf`
- T. Shorten laburpena: `http://cran.r-project.org/doc/contrib/Short-refcard.pdf`
- V. Ricciren erregresioaren laburpena: `http://cran.r-project.org/doc/contrib/Ricci-refcard-regression.pdf`
- *R* seek: `http://www.rseek.org/`

0.3. Datu-baseen irakurketa

R programan sortutako berezko datu-fitxategiek *.Rda* edo *.RData* hedapena dute. Beraz, horrelako fitxategi bat kargatzeko, jarraitu *Rcmdr* interfazean honako argibide hauek:

`Datuak` → `Kargatu datuak` → `... datu multzotik`

Hala ere, *R* gai da hedapen anitzeko fitxategiak irekitzeko: testu soila, URL, SPSS, Minitab, STATA, Excel, Access eta dbase motakoak, besteak beste. Honelako datu-baseak kargatzeko, honako pauso hauek jarrai daitezke *Rcmdr* interfazetik:

`Datuak` → `Inportatu datuak` → `... datu multzotik`

0.4. Datu-baseen aldaketak

Datu-base batean aldagai berriak sor daitezke, edota dauden aldagaiei aldaketak egin dakizkieke. Horretarako:

`Datuak` → `Kudeatu datu multzo aktiboko aldagaiak`

Ikus ditzagun adibide batzuk **Altuerak.RData** fitxategia erabiliz. Hasteko, kargatu fitxategia:

`Datuak` → `Kargatu datu multzoa` → `Altuerak.RData`

Zenbakizko aldagaiak kualitatibo bihurtzen

`Ikusi datu multzoa` sakatzen badugu, ikus dezakegu *jaioterria* aldagaia kodifikatu gabe dagoela. Horrek esan nahi du jaiotze-lurralde bakoitza zenbaki baten bidez adierazita dagoela, aldagai numeriko bat balitz bezala. Aldagai hori kualitatibo edo, *R*-ren nomenklatura erabiliz, faktore bihurtzeko:

`Datuak` → `Kudeatu datu multzo aktiboko aldagaiak` → `Zenbakizko aldagaiak faktore bihurtu` → `Aldagaiak:` *jaioterria*. → `Faktore-mailak: ipini maila-izenak. Izen berria :` *jaioterriak*. `1` *Araba*, `2` *Bizkaia*, `3` *Gipuzkoa*.

Aldagaiak zatikatzea

Demagun *pisua* aldagaia 4 zatitan zatikatu nahi dugula, datu elkartu moduan adieraziz. Hiru segmentazio-metodo ezberdinen bidez egin daiteke hori.

- *pisua* aldagaia zabalera berdineko 4 tartetan zatikatzeko:

  `Datuak` → `Kudeatu datu multzo aktiboko aldagaiak` → `Zatitu zenbakizko aldagai bat` → `Zatikatzeko aldagaia:` *pisua*. `Aldagai berriaren izena:` *pisua*1. `Klase kopurua: 4. Maila-izenak: heinak. Segmentazio-metodoa: zabalera berdineko zatiak.`

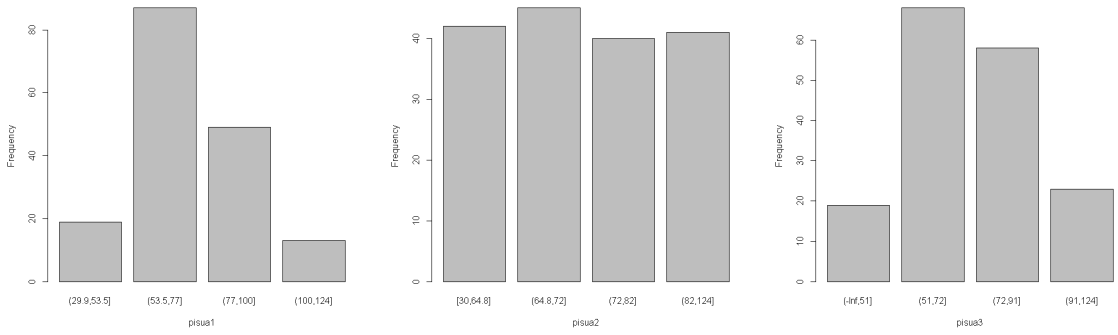
- *pisua* aldagaia elementu kopuru edo maiztasun berdina duten 4 tartetan zatikatzeko:

  `Datuak` → `Kudeatu datu multzo aktiboko aldagaiak` → `Zatitu zenbakizko aldagai bat` → `Zatikatzeko aldagaia:` *pisua*. `Aldagai berriaren izena:` *pisua*2. `Klase kopurua: 4. Maila-izenak: heinak. Segmentazio-metodoa: kopuru bereko zatiak.`

- *pisua* aldagaia $K - means$ bezalako algoritmo baten bidez zatikatzeko, datuetan dauden multzo «naturalak» edo berezkoak identifikatuz:

  `Datuak` → `Kudeatu datu multzo aktiboko aldagaiak` → `Zatitu zenbakizko aldagai bat` → `Zatikatzeko aldagaia:` *pisua*. `Aldagai berriaren izena:` *pisua*3. `Klase kopurua: 4. Maila-izenak: heinak. Segmentazio-metodoa: segmentu naturalak.`

Honako barra diagrama hauetan hiru zatikatze mota horiekin lortutako tarteak eta haien maiztasunak ikus ditzakegu:



Aldagai berria

Jarraian, *imc (indice masa corporal)* izeneko aldagai berria sortuko dugu honako kalkulu hau eginez:

$$imc = \frac{pisua(Kg)}{altuera^2(m^2)}.$$

*Rcmdr*n hori egiteko:

`Datuak` → `Kudeatu datu multzo aktiboko aldagaiak` → `Kalkulatu aldagai berri bat` → `Aldagai berriaren izena: imc. Konputatzeko adierazpena:` $pisua/(altuera/100)^2$

Aldagaiak birkodetzea

imc indizean oinarrituta, obesitate-maila aldagaia (*maila*) eraikiko dugu SEEDO izeneko Obesitatearen Ikerketarako Sozietatearen sailkapenari jarraituz:

$$\text{obesitate-maila} = \begin{cases} \text{gutxiegizko pisua}, & \text{baldin } imc < 18,5 \text{ bada}, \\ \text{pisu normala}, & \text{baldin } 18,5 \leq imc < 25 \text{ bada}, \\ \text{gehiegizko pisua}, & \text{baldin } 25 \leq imc < 30 \text{ bada}, \\ \text{obesitatea}, & \text{baldin } imc \geq 30 \text{ bada} \end{cases}$$

Rcmdr erabiliz, *imc* aldagaia birkodetu eta *maila* aldagai berria sortzeko:

`Datuak` → `Kudeatu datu multzo aktiboko aldagaiak` → `Berkodetu aldagaiak` → `Berkodetu behar diren aldagaiak:` *imc*. `Aldagai-izen berria:` *maila*. `Sartu birkodetzeko jarraibideak:`

```
0:18.499="gutxiegizkoa";
18.499:24.999="normala";
24.999:29.999="gehiegizkoa";
29.999:100="obesitatea"
```

1. *Rcmdr* praktika

Estatistika deskribatzailea

Helburua

Praktika honen xedea izango da jasotako datu esperimentalak aztertzea, laburtzea eta deskribatzea, bai metodo grafikoen bidez eta bai zenbakizkoen bidez.

Bi kasu banatuko ditugu: aldagai bakarraren analisi deskribatzailea eta bi aldagairen aldibereko azterketa, bi kasuetan maiztasun-taulen eta grafiko ezberdinen eraikuntza aztertuz. Noski, bi kasuetan bai aldagai kuantitatiboen eta bai kualitatiboen azterketa landuko dugu. Azkenik, aldagai bakarraren kasuan, horiek deskribatzeko estatistiko ezberdinen kalkuluari buruzko azalpenak ere emango ditugu.

1.1. Adibidea

Praktika honetako helburuak lantzeko, honako adibide honetan oinarrituko gara.

1. adibidea. *Gizabanako baten altuera eta erlazionatutako beste zenbait aldagai aztertu nahi ditugu estatistika deskribatzailea erabiliz. Aztertutako aldagaiak honako hauek dira: gizabanakoaren altuera, aitaren altuera, amaren altuera, jaioterria, sexua eta pisua.*

Aurreko atalean azaldu bezala, **Altuerak.RData** fitxategian dauzkazue gordeta datu horiek. Beraz, hasteko, datuak kargatuko ditugu:

`Datuak` → `Kargatu datu multzoa` → `Altuerak.RData`

Gogoratu hasi baino lehen *jaioterria* aldagaia faktore bihurtu behar dugula, aurreko praktikan ikasi dugun moduan.

1.2. Aldagai baten analisi deskribatzailea

1.2.1. Maiztasun-taulak

Rcmdr paketeak soilik aldagai kualitatiboen maiztasun-taulak eraikitzen ditu. Gure aldagaia kuantitatibo (numeriko) gisa kodifikatuta badago, faktore motara pasatu beharko dugu lehenik maiztasun-taula eraiki nahi badugu:

`Datuak` → `Kudeatu datu multzo aktiboko aldagaiak` → `Zenbakizko aldagaia faktore bihurtu` → `Aldagaiak:` *amarena*. `Faktore mailak: erabili zenbakiak. Aldagai-izen berria:` *amarena.f*

Kontuan izan kodifikatutako aldagai berri hau `amarena.f` aldagaian gorde dugula. Orain, maiztasun-taula eraikitzeko:

`Estatistikoak` → `Laburpenak` → `Maiztasun-banaketak` → *amarena.f*

Lortutakoa *Emaitza* izeneko leihoan ikus dezakegu. Kontuan izan hemendik aurrera *Rcmdr*-ren Emaitzak leihoko irteerak honako formatu honetan adieraziko ditugula:

```
counts:
amarena.f
153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171
  1   3   6  10  12  16  14  15  24  19  12   7   8  11   4   1   1   3   2

percentages:
amarena.f
  153   154   155   156   157   158   159   160   161   162   163   164   165
 0,59  1,78  3,55  5,92  7,10  9,47  8,28  8,88 14,20 11,24  7,10  4,14  4,73
  166   167   168   169   170   171
 6,51  2,37  0,59  0,59  1,78  1,18
```

Hau da, f maiztasun absolutuen eta $100 \cdot h$ ehuneko erlatiboen taula agertzen zaigu.

Hemendik, moda $Mo = 161$ dela ondoriozta daiteke, eta pertzentilak kalkulatzeko aukera dugu.

1.2.2. Grafikoak

Rcmdr erabiliz, grafiko mota ezberdinak egin daitezke. Kontuan izan aldagaia zein motatakoa (kualitatiboa, kuantitatiboa, eta diskretua edo jarraikia) den begiratu beharko dugula, grafiko mota aukeratzeko garaian:

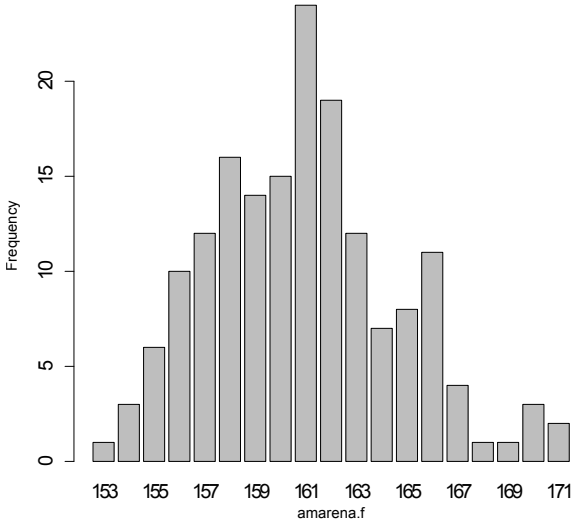
`Grafikoak` → `Barra-diagrama` → *amarena.f*

`Grafikoak` → `Zurtoin-eta hosto-diagrama` → *amarena*

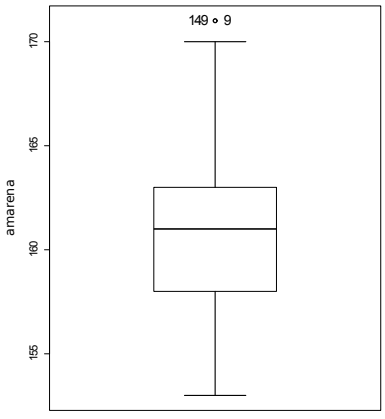
`Grafikoak` → `Kutxa-diagrama` → *amarena*

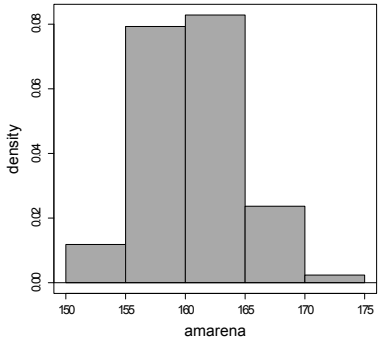
`Grafikoak` → `Histograma` → `Datuak:` *amarena*. `Aukerak. Klase kopurua: 5, Ardatzen eskala: dentsitateak.`

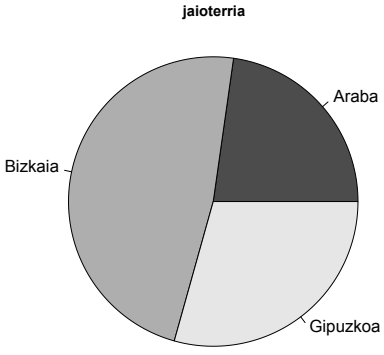
`Grafikoak` → `Sektore-grafikoa` → `Aldagaia:` *jaioterria*




```
      n: 169
  1   153 | 0
  4   154 | 000
 10   155 | 000000
 20   156 | 0000000000
 32   157 | 000000000000
 48   158 | 0000000000000000
 62   159 | 00000000000000
 77   160 | 000000000000000
(24)  161 | 000000000000000000000000
 68   162 | 000000000000000000
 49   163 | 000000000000
 37   164 | 0000000
 30   165 | 00000000
 22   166 | 00000000000
 11   167 | 0000
  7   168 | 0
  6   169 | 0
  5   170 | 000
```







1.2.3. Estatistikoak

Estatistikoak dira aldagaiak era laburrean eta zehatzean adierazten dituzten zenbakizko balioak. *Rcmdr* erabiliz, datu-baseko aldagai guztien oinarrizko estatistiko batzuk honela lor ditzakegu:

`Estatistikoak` → `Laburpenak` → `Datu multzo aktiboa`

```
aitarena        amarena             sexua
 Min.   :157.0   Min.   :153.0   gizonezkoa :89
 1st Qu.:169.0   1st Qu.:158.0   emakumezkoa:82
 Median :173.0   Median :161.0
 Mean   :172.1   Mean   :160.9
 3rd Qu.:175.0   3rd Qu.:163.0
 Max.   :185.0   Max.   :171.0
 NA's   :1       NA's   :2
    jaioterria    altuera          pisua
 Araba   :38   Min.   :159.0   Min.   : 30.00
 Bizkaia :80   1st Qu.:165.0   1st Qu.: 64.75
 Gipuzkoa:49   Median :167.0   Median : 72.00
 NA's    : 4   Mean   :166.7   Mean   : 72.48
               3rd Qu.:169.0   3rd Qu.: 82.00
               Max.   :175.0   Max.   :124.00
               NA's   :3       NA's   :3
```

Ikusten dugunez, minimoa, 1. kuartila, mediana, batezbestekoa, 3. kuartila eta maximoa kalkulatzen ditu aldagai kuantitatiboentzat, hots, $\{min,\ q_1,\ Me,\ \overline{x},\ q_3,\ max\}$ estatistikoak lortu ditugu bide honi jarraituz. Aldiz, aldagai kualitatiboentzat maila bakoitzaren maiztasunak ematen dizkigu.

Aldagai kuantitatiboentzat, estatistiko sorta zabalago bat lortu nahi badugu:

`Estatistikoak` → `Laburpenak` → `zenbakizko laburpenak` → `Aldagaiak:` *amarena*.
`Estatistikoak: batezbestekoa, desbideratze estandarra, kuartilarteko heina, aldakuntza-koefizientea, asimetria, kurtosia, koantilak`

```
    mean     sd     IQR   cv      skewness   kurtosis    0% 25% 50% 75% 100% n NA
 160.8994 3.712343  5 0.02307244 0.3981746 -0.05915588 153 158 161 163 171 169 2
```

Honela, batezbestekoa (*mean*), lagin-kuasidesbideratze estandarra (*sd*), kuartilarteko heina (*IQR*), aldakuntza-koefizientea (*cv*), asimetria-koefizientea (*skewness*), kurtosia (*kurtosis*), mi-

nimoa (%0), 1. kuartila (%25), mediana (%50), 3. kuartila (%75), maximoa (%100), laginaren tamaina (n) eta balio galduen kopurua (NA) lortu ditugu, hurrenez hurren.

1.3. Bi aldagaien analisi deskribatzailea

Aldagai bakar bat aztertzeaz gain, askotan, bi aldagai aldi berean aztertzeko beharra suertatzen zaigu. Aldagai bikote motaren arabera, tratamentu mota ezberdina eman beharko diegu; honako konbinazio hauek aztertuko ditugu jarraian:

- Aldagai kualitatibo bat eta aldagai kuantitatibo bat.
- Bi aldagai kualitatibo.
- Bi aldagai kuantitatibo.

1.3.1. Maiztasun-taulak

Atal honen helburua izango da bi aldagai aldi berean taula bidez aztertzea. *Rcmdr*n hau soilik bi aldagaiak kualitatiboak direnean egingo dugu. Kasu honetan, **kontingentzia-taula** edo bi norabideko taulak erabiliko ditugu. Adibidez, **Altuerak.RData** fitxategiko *sexua* eta *jaioterria* aldagaiak honela adierazi ditzakegu kontingentzia-taula baten bitartez:

`Estatistikoak` → `Kontingentzia-taulak` → `Sarrera bikoitzeko taula` → `Errendakako aldagaia:` *sexua*, `Zutabeko aldagaia:` *jaioterria*

```
Frequency table:
             jaioterria
sexua         Araba Bizkaia Gipuzkoa
  gizonezkoa     18      43       26
  emakumezkoa    21      37       24
```

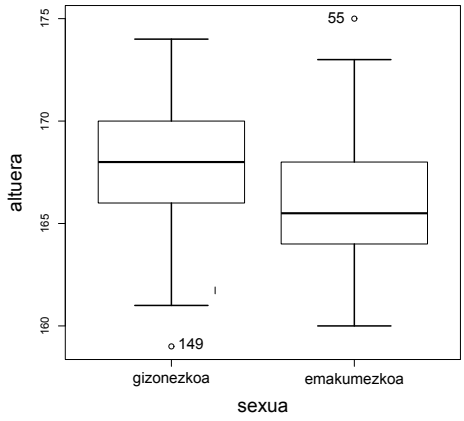
Adibidez, ikus dezakegu laginean Arabako emakumezkoen maiztasuna 21 pertsonakoa dela. Aldiz, Gipuzkoako gizonezkoena 26 da.

Honelako taulak aurrerago sakonago aztertuko ditugu 5. *Rcmdr* praktikan, independentzia- eta homogeneotasun-probak aztertzen ditugunean.

1.3.2. Grafikoak

• Aldagai kuantitatibo eta kualitatibo pare bat elkarrekin irudikatzeko, aldagai kuantitatiboentzat azaldutako ohiko grafikoak erabiliko ditugu (zurtoin-eta hosto-diagramak, kutxa-diagramak, histogramak eta abar), baina aldagai kualitatiboaren maila bakoitzerako grafiko ezberdin bat eginez. Adibide gisa, irudika dezagun altuera sexuarekiko, kutxa-diagramak erabiliz:

`Grafikoak` → `Kutxa-diagrama` → `Datuak:` *altuera*`. Marraztu taldearen arabera:` *sexua*

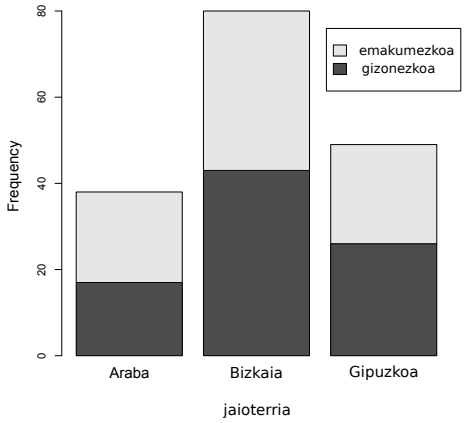


55. eta 149. behaketak bi balio arraro (outlierak) dira. Izan ere, 55. behaketa emakume altuenari dagokio, eta 149. behaketa gizon baxuenari.

Grafiko honen bitartez, bi laginen (emakumezkoen) eta gizonezkoen banaketak konpara ditzakegu grafikoki. Adibidez, ikus dezakegu lagin honetan gizonezkoen altueraren mediana handiagoa dela, eta, are gehiago, gizonezkoen 1. kuartila emakumezkoen 2. kuartila baino handiagoa da. Aldiz, sakabanapen aldetik, bi laginak antzekoak dira. Honelako konparaketak kuantitatiboki nola egin sakonago aztertuko dugu 4. (bi populazioen konparaketak) eta 6. (2 populazio baino gehiagoren konparaketak) *Rcmdr* praktiketan.

• Bi aldagai kualitatiboen kasuan, barra-diagrama berezi batzuk eraiki ditzakegu, aldagai kualitatiboetako bat bestearen arabera irudikatuko dutenak. Adibidez, `jaioterria` `sexua`-rekiko adierazi nahi badugu barra-diagramak erabiliz:

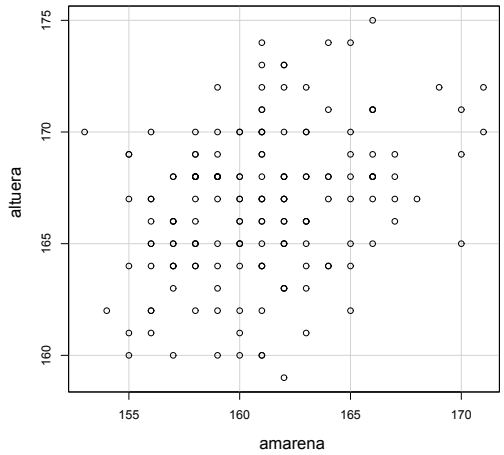
`Grafikoak` → `Barra-diagrama` → `Datuak:` *jaioterria*`. Marraztu taldearen arabera:` *sexua*



Grafiko honetan, jaioterria sexuaren arabera nola aldatzen den ikus dezakegu, eta bi aldagaien arteko menpekotasuna azter dezakegu grafikoki. Menpekotasun hauek kuantitatiboki aztertzeko, 5. *Rcmdr* praktikan azalduko ditugun independentzia- eta homogeneotasun-probak erabiliko ditugu.

• Bi aldagaiak kuantitatibo jarraituak badira, elkarrekin irudikatzeko erabiliko dugun grafiko mota barreiadura-diagrama edo puntu-hodeia da. Adibidez, demagun amaren altuera eta altuera aldagaiak irudikatu nahi ditugula. *Rcmdr* erabiliz:

`Grafikoak` → `Barreiadura-diagrama` → `x aldagaia:` *amarena*, `y aldagaia:` *altuera*



Grafiko honetan, bi aldagai jarraituen arteko erlazioa nolakoa den azter dezakegu. Erlazio hauek, ondoren, eredu ezberdinen bidez kuantifika ditzakegu. Eredu horiei erregresio-eredu deritzegu, eta 7. gaian aztertuko ditugu sakonago.

1.3.3. Estatistikoak

Aldagai jarraiki bakar baten estatistikoak lortzen genituen moduan, aldagai jarraitu bat eta kualitatibo bat badauzkagu, estatistiko berdinak taldeen arabera kalkula ditzakegu. Adibidez, batezbestekoa, kuasi-desbiderapen estandarra eta koantilak lortzeko:

`Estatistikoak` → `Laburpenak` → `zenbakizko laburpenak` → `Aldagaiak:` *amarena*. → `Laburbildu taldeka:` *sexua* → `Estatistikoak: Batezbestekoa, Desbideratze estandarra , Koantilak.`

```
              mean       sd    0% 25% 50% 75%  100%   n  NA
emakumezkoa 161.2250 3.518396 153 159 161 163  171   80   2
gizonezkoa  160.6067 3.874796 154 158 160 163  171   89   0
```

2. *Rcmdr* praktika

Probabilitatearen teoria

Helburua

Praktika honen jomuga da probabilitatearen teorian landutako kontzeptu batzuk *Rcmdr* paketea erabiliz jorratzea. Adibidez, zorizko aldagaien probabilitatearen legea, edo dentsitate-funtzioa eta banaketa-funtzioa kalkulatzea eta irudikatzea, pertzentilak lortzea edo zorizko laginak sortzea.

2.1. Banaketak

Zorizko aldagai jarraituak erabiliz, zenbait ekintza gauza ditzakegu *Rcmdr*-ren bidez honako agindu hauek erabiliz:

`Banaketak` → `Banaketa jarraituak` → `{Banaketa normala, t banaketa, Khi karratuaren banaketa, F banaketa, banaketa esponentziala, banaketa uniformea,...` → `{koantilak, probabilitateak, marraztu banaketa, atera lagina banaketatik}`

Ikusten dugun moduan, banaketa bakoitza erabiliz, lau motatako emaitzak atera daitezke:

- `Pertzentilak (koantilak)`: $\alpha \in (0,1)$ probabilitate guztietarako p_α estatistikoa ematen digu, $P(X \leq p_\alpha) = \alpha$ betetzen duena, baldin `ezkerrerako isatsa` hautatzen badugu edo $P(X > p_\alpha) = \alpha$ betetzen duena, `eskuinerako isatsa` hautatzekotan.

  Adibidez $X : \mathcal{N}(0,1)$ banaketa erabiliz, $p_{0,05}$ balioa kalkulatuko dugu, non $P(X > p_{0,05}) = 0,05$:

  `Banaketak` → `Banaketa jarraituak` → `Banaketa normala` → `Koantil normalak` → `Probabilitateak: 0.05, Batezbestekoa: 0, Desbideratze estandarra: 1, eskuinerako isatsa`

```
[1] 1.644854
```

$X : t_8$ banaketa erabiliz, $p_{0,025}$ balioa kalkulatuko dugu, non $P(X \leq p_{0,025}) = 0,025$:

`Banaketak` → `Banaketa jarraituak` → `t banaketa` → `t koantilak` → `Probabilitateak: 0.025, Askatasun-graduak: 8, ezkerrerako isatsa`

```
[1] -2.306004
```

- `Banaketa-funtzioa (probabilitateak)`: $a \in \mathbb{R}$ zenbaki guztietarako banaketa-funtzioa (probabilitate metatua) ematen digu, $P(X \leq a)$, baldin `ezkerrerako isatsa` hautatzen badugu edo $P(X > a)$, `eskuinerako isatsa` hautatzekotan.

  Adibidez $X : \mathcal{F}_{2,3}$ banaketa erabiliz, $P(X > 1,5)$ kalkulatuko dugu:

  `Banaketak` → `Banaketa jarraituak` → `F banaketa` → `F probabilitateak` → `Aldagaiaren balioa: 1.5, Zenbakitzailearen askatasun-graduak: 2, Izendatzailearen askatasun-graduak: 3, eskuinerako isatsa`

```
[1] 0.3535534
```

  Bestalde, $X : \mathcal{N}(10, 2)$ banaketa erabiliz, $P(X \leq 1,7)$ kalkulatuko dugu:

  `Banaketak` → `Banaketa jarraituak` → `Banaketa normala` → `Probabilitate normalak` → `Aldagaiaren balioa: 1.7, Batezbestekoa: 10, Desbideratze estandarra: 2, ezkerrerako isatsa`

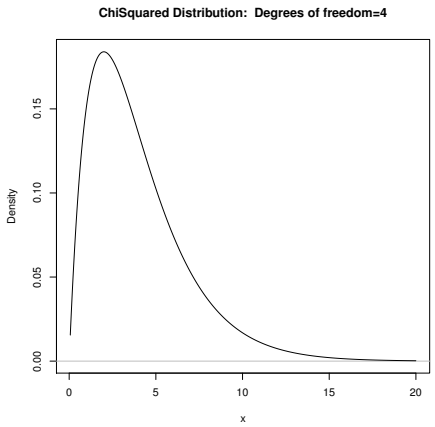
```
[1] 1.662376e-05
```

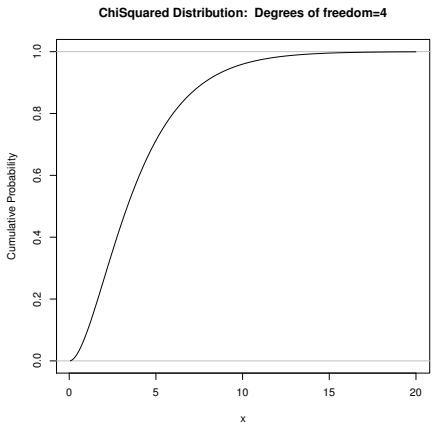
- f `edo` F`-ren grafikoa`: behin banaketaren parametroak finkatuta, f dentsitate-funtzioaren edo F banaketa-funtzioaren grafikoa ematen digu.

  Adibidez $X : \chi_4$ banaketaren f dentsitate-funtzioaren irudikapena egiteko:

  `Banaketak` → `Banaketa jarraituak` → `Chi-karratuaren banaketa` → `Marraztu Chi-karratuaren banaketa Askatasun-graduak: 4, Marraztu dentsitate-funtzioa`

  Era berean, bere banaketa-funtzioa irudika dezakegu, azken aukera aldatuz.





- `Laginak`: behin banaketa eta bere parametroa(k), lagin kopurua eta laginen tamaina finkatuta, ezaugarri horiekin zoriz sortutako lagina(k) ematen di(zki)gu.

  Adibidez, $X : t_8$ banaketatik $n = 5$ tamainako zorizko lagin bat ateratzeko:

  `Banaketak` → `Banaketa jarraituak` → `t banaketa` → `Atera lagina t-banaketatik` → `Askatasun-graduak: 8, Lagin kopurua=1, Behaketa kopurua=5`

  Honek **tSamples** izeneko datu-basea definituko du, eta bertan gordeko da sortu berri dugun lagina. Gainera, datu-base hau aktibo jarriko du, eta jarraian has gaitezke berarekin lanean.

Zorizko aldagai diskretuen kasuan, honako aukera hauek eskaintzen ditu *Rcmdr* paketeak:

`Banaketak` → `Banaketa diskretuak` → `{Banaketa binomiala, Poisson-en banaketa, ...}` → `{koantilak, probabilitate metatuak, probabilitateak, marraztu banaketa, atera lagina banaketatik}`

Kasu honetan, banaketa bakoitzetik bost motatako emaitzak atera daitezke: aurreko guztiak eta, gainera, bi probabilitate mota kalkula ditzakegu:

- F `banaketa-funtzioa (probabilitate metatuak)` $F(a) = P(X \leq a)$ `ezkerrerako isatsa`-ri dagokio, edo $P(X > a)$ `eskuinerako isatsa`-ri dagokio.

  Adibidez, $X : Bin(10,\ 0,2)$ banaketa erabiliz, $P(X > 2)$ kalkulatzeko aurreko kasuan bezala egingo dugu:

  `Banaketak` → `Banaketa diskretuak` → `Banaketa binomiala` → `Probabilitate binomial metatuak` → `Aldagaiaren balioa: 1.7, Entsegu binomialak: 10, Arrakasta-probabilitatea: 0.2, eskuinerako isatsa`

```
[1] 0.3222005
```

- f `probabilitate-legea` $f(a) = P(X = a)$ emaitza lortzeko balio du.
  Adibidez, $X : \mathcal{P}(1, 5)$ banaketa erabiliz:

  `Banaketak` $\rightarrow$ `Banaketa diskretuak` $\rightarrow$ `Poisson-en banaketa` $\rightarrow$ `Poisson-en probabilitateak` $\rightarrow$ `Batezbestekoa: 1.5`

```
   Probability
0 0.2231301601
1 0.3346952402
2 0.2510214302
3 0.1255107151
4 0.0470665182
5 0.0141199554
6 0.0035299889
7 0.0007564262
```

  Beraz, adibidez $P(X = 5) = 0,0141199554$.

3. *Rcmdr* praktika

Konfiantza-tartezko zenbatespena

Helburua

Praktika honen bidez, ikusiko dugu nola kalkulatu populazio-parametroak zenbatesteko konfiantza-tarte ohikoenak *Rcmdr* erabiliz. Zehazki:

- Lagin bakarreko batezbestekoaren $I_{\mu}^{1-\alpha}$ konfiantza-tartea, σ ezezagunerako.
- Bi bariantzaren zatidurarako $I_{\sigma_1^2/\sigma_2^2}^{1-\alpha}$ konfiantza-tartea.
- Bi populazio askeren batezbestekoen kendurarako $I_{\mu_1-\mu_2}^{1-\alpha}$ konfiantza-tartea.
- Binakako datuen batezbestekoen kendurarako $I_{\mu_d}^{1-\alpha}$ konfiantza-tartea.
- Lagin bakarrerako proportziorako $I_{p}^{1-\alpha}$ konfiantza-tartea.

3.1. Lagin bakar baten konfiantza-tarteak

3.1.1. Batezbestekorako konfiantza-tarteak

Atal honetan, bariantza ezezaguneko populazio normalen batezbestekorako konfiantza-tarteak kalkulatzen ikasiko dugu, lagin bakarreko t testa erabiliz.

2. adibidea. *Hiri txiki batean, ur-erabilerari buruzko ikerkuntza batean, 25 etxebizitzatako zorizko lagina atera da. Banaketa normalari darraion X aldagaia aztertuko da: asteko erabilitako ur litro kopurua. Zoriz aukeraturiko aste batean, honako balio hauek lortu ziren:*

175 185 186 118 158 150 190 178 137 175
180 200 189 200 180 172 145 192 191 181
183 169 172 178 210

Informazio hori erabiliz, zenbatets itzazu puntualki μ, σ^2 eta σ. Lor ezazu μ-rako % 90eko konfiantza-tartea. Hiriko ur-depositua aski handia da asteko 160 litroko batez besteko kontsumoa baimentzeko. Uste duzu hirian ur faltaren arazorik egon daitekeenik? Azal ezazu erantzuna, lortutako konfiantza-tartean oinarriturik.

`Ura` izeneko aldagaian, adibidean emandako 25 datuak eskuz sartu edo **praktikadatuak.xls** fitxategiko **ur** izeneko orritik inportatu ostean, μ, σ^2 eta σ puntualki estimatzeko:

`Estatistikoak` → `Zenbakizko laburpenak` → `Aldagaiak:` *ura.* → `Estatistikoak: Batezbestekoa, Desbideratze estandarra` → `Onartu`

```
   mean       sd  n
 175.76 20.79319 25
```

Beraz, $\hat{\mu} = 175,76$ litro/aste, $\hat{\sigma^2} = 20,79319^2$ (litro/aste)^2 eta $\hat{\sigma} = 20,79319$ litro/aste.

Ondoren, batezbestekoaren konfiantza-tartea lortzeko, *Rcmdr*n jarraitu honako pauso hauek:

`Estatistikoak` → `Batezbestekoak` → `Lagin bakarreko t testa` → `Populazioaren batezbestekoa !=` *mu*`0. Hipotesi nulua: mu=[0.0]. Konfiantza-maila [0.90]` → `Onartu`

```
        One Sample t-test
data:  ura$ura
t = 42.2638, df = 24, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 168.6451 182.8749
sample estimates:
mean of x
   175.76
```

`t` estatistikoa, `df` (degree freedom) askatasun-graduak, `p-value` p-balioa eta `alternative hypothesis` balioak hipotesi-kontraste parametrikoen gaiarekin erlazionatuta daude [1]. Jarraian

[1] Hemendik aurrera eta gai honetan, ez ditugu lehen lerro horiek erakutsiko, 4. *Rcmdr* praktikan azalduko baitira berriro.

`90 percent confidence interval` lerroaren azpian, populazioaren batez besteko ur litroaren %90eko konfiantza-tartea ikus dezakegu: $I_\mu^{0,90} = (168,6451,\ 182,8749)$. Honekin ondoriozta dezakegu proposatutako ur-depositua (160 litrokoa) ez dela nahikoa izango. Azkenik, azken lerroan, batezbestekoaren estimazio puntuala agertzen zaigu: $\hat{\mu} = 175,76$ litro/aste.

3. adibidea. *Honako taula honetan, lan zehatz bat gauzatzeko enpresa bateko langileen denbora-tarteak (segundotan) adierazi dira. Demagun populazio normalak direla.*

Emakumezkoa	*103*	*94*	*110*	*87*	*98*		
Gizonezkoa	*97*	*82*	*123*	*92*	*175*	*88*	*118*

a) Kalkula ezazu batezbestekoaren puntu-zenbatespena eta % 90 mailako konfiantza-tartea.

b) Kalkula itzazu gizonezkoenak eta emakumezkoenak.

c) Kalkula ezazu % 90eko bariantzen zatidurarako konfiantza-tartea. Zer esan daiteke?

d) Kalkula ezazu berriro populazioaren batezbestekoen arteko diferentziarako % 90 mailako konfiantza-tartea. Zein ondorio atera dezakegu?

`denbora` eta `sexua` (1=emakumezkoa, 2=gizonezkoa) aldagaiek `denbora` izeneko datu-basea osatzen dute. Beraz, hasteko, datu guztiak sartu/inportatu (**praktikadatuak.xls** fitxategiko **denbora** orrialdean daude eskuragai).

Ondoren, 3. adibideko (a) atalari erantzuteko, jarraitu aurreko ataleko pausoei, denbora izeneko aldagai kuantitatiborako batezbestekoaren puntu-zenbatespena eta % 90 mailako konfiantza-tartea kalkulatzeko [2]

```
        One Sample t-test
data:  denbora$denbora
90 percent confidence interval:
  92.54004 118.62662
sample estimates:                mean of x
    105.5833
```

Batezbestekoaren puntu-zenbatespena eta % 90eko konfiantza-tartea lortu ditugu, $\hat{\mu} = 105,5833$ eta $I_\mu^{0,90} = (92,54004,\ 118,62662)$, non μ denbora-tartearen batezbestekoa baita.

Orain, (b) atala ebazteko, kalkula ditzagun banan-banan emakumezkoen eta gizonezkoen batezbestekoen puntu-zenbatespenak eta % 90eko konfiantza-tarteak. Horretarako, bihur dezagun *sexua* aldagai kualitatibo (factor):

[2] Ohartu programak kontuan hartuko duela etiketen zenbakien ordena, diferentziak edo zatidurak aztertzerakoan.

`Datuak` → `Kudeatu datu multzo aktiboko aldagaiak` → `Zenbakizko aldagaiak faktore bihurtu` → `Ipini maila-izenak. Izen berria[sexua]` → `1[emakumezkoa] 2[gizonezkoa]`

Orain, emakumezkoen konfiantza-tartea lortzeko, lehenik datu aktiboak filtratu behar ditugu:

`Datuak` → `Datu multzo aktiboa` → `Filtratu datu multzo aktiboa` → `Sartu aldagai guztiak. Aldagaiak:` *sexua*. `Filtratu adierazpena [sexua==1]. Datu multzo berriaren izena[den.em]`

Kontuan izan *sexua* faktore moduan egotekotan, adierazpena `sexua=="emakumezkoa"` idatzi beharko genukeela, `sexua==1` beharrean.

Orain, soilik emakumezkoen datuak daudenez aktibatuta, t-testa eska daiteke aurreko adibidean bezalaxe, batezbestekoaren konfiantza-tartea eta zenbastespen puntuala lortzeko. Gainera, prozesu bera jarraituz, gizonezkoen tartea kalkula daiteke, baina, gizonezkoen datuak filtratu baino lehenago, jatorrizko datuak aktibatu behar dira:

`Datuak` → `Datu multzo aktiboa` → `Aukeratu datu multzo aktiboa. Datu multzoa:[denbora]`

Hona hemen lortutako emaitzak:

```
        One Sample t-test
data:  den.em$denbora
90 percent confidence interval:
  90.07214 106.72786
sample estimates: mean of x
     98.4
```

$I_{\mu_1}^{0,90} = (90,7214,\ 106,72786)$, emakumezkoen batez besteko denbora zenbatesteko % 90eko konfiantza-tartea da $\hat{\mu}_1 = 98,4$.

```
        One Sample t-test
data:  den.giz$denbora
90 percent confidence interval:
  87.07556 134.35301
sample estimates: mean of x
 110.7143
```

$I_{\mu_2}^{0,90} = (87,07556,\ 134,35301)$ gizonezkoen batez besteko denbora zenbatesteko % 90eko konfiantza-tartea da eta $\hat{\mu}_2 = 110,7143$.

Baina, nola ondoriozta dezakegu sexuak denboran eragina duela? Horretarako, beharrezkoa da lagin birako zenbatespen-tarteak kalkulatzea.

3.2. Lagin birako zenbatespen-tarteak

Amai dezagun 3. adibidea: $I_{\mu_1-\mu_2}^{0,90}$ tartea kalkulatuz (ohartu laginak askeak direla). Tarte egokia aukeratzeko, lehenik populazio-bariantzen arteko erlazioa aztertu behar da (c) atalean galdetzen diguten moduan.

3.2.1. Bariantzen zatidurarako konfiantza-tarteak

Rcmdr programak bariantzen zatidura aztertzeko hiru aukera ematen dizkigu: F, Bartlett eta Levene testak, hain zuzen ere. Beraz,

`Estatistikoak` → `Bariantzak` → `F, Bartlett edo Levene`

Banaketa normalari jarraitzen dioten bi lagin askeen kasuan, F testa erabiliko dugu. Esaterako, 3. adibideko (c) atalari erantzuna emateko:

```
        F test to compare two variances
data:  denbora by sexua
90 percent confidence interval:
 0.01624629 0.45394809
sample estimates:                ratio of variances
    0.07365542
```

$I^{0,90}_{\sigma_1^2/\sigma_2^2} = (0,01624629,\ 0,45394809)$ dela ikus dezakegu, eta 1 tartean ez dagoenez, orduan $\sigma_1^2 \neq \sigma_2^2$ ondorioztatzen da % 90eko konfiantza-mailarekin. Are gehiago, $I^{0,90}_{\sigma_1^2/\sigma_2^2} \subset (0,\ 1)$ denez, $\sigma_1^2 < \sigma_2^2$ ondorioztatzen dugu.

3.2.2. Bi populazio askeren batezbestekoen diferentziarako konfiantza-tarteak

Behin bariantza ezezagunen arteko erlazioa aztertuta, has gaitezen 3. adibideko (d) atalarekin, eta kalkula dezagun batezbestekoen diferentziarako konfiantza-tartea:

`Estatistikoak` → `Batezbestekoak` → `Lagin independenteko t testa` → `Taldeak[sexua] Mendeko aldagaia [denbora]. Hipotesi alternatiboa [alde bikoa]. Konfiantza-maila [0.90]. Bariantzak berdinak direla jo? [ez]` → `Onartu`

```
        Welch Two Sample t-test
data:  denbora by sexua
90 percent confidence interval:
 -36.42670  11.79813
sample estimates:
mean in group emakumezkoa  mean in group gizonezkoa
                  98.4000                  110.7143
```

Beraz, $I^{0,90}_{\mu_1-\mu_2} = (-36,42670,\ 11,79813)$ da batezbestekoen diferentziarako % 90eko konfiantza-mailako zenbatespen-tartea. $0 \in I^{0,90}_{\mu_1-\mu_2}$ dagoenez, ezin da ondorioztatu batezbesteko bat bestea baino handiagoa denik; ezin da baztertu emakumezko eta gizonezkoen batez besteko denboren berdintasuna.

Kontuan izan bariantzak berdinak direla ondorioztatu izan bagenu, honako agindu hauek erabili beharko genituzkeela:

`Estatistikoak` → `Batezbestekoak` → `Lagin independenteko t testa` → `Taldeak [sexua] Mendeko aldagaia [denbora]. Hipotesi alternatiboa [alde bikoa]. Konfiantza-maila [0.90]. Bariantzak berdinak direla jo? [bai]` → `Onartu`

3.2.3. Binakako datuen batezbestekoen diferentziarako konfiantza-tartea

4. adibidea. *Espektofotometriaren bidez, tomate freskoen eta ontziraturiko tomateen nahitaezko elementuak ikertu dira. Horretarako, kobre kopurua konparatu da tomate freskoetan eta tomate berberetan, haiek ontziratu ondoren. Datuak honako hauek dira:*

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Freskoa	*0,066*	*0,079*	*0,069*	*0,076*	*0,071*	*0,087*	*0,071*	*0,073*	*0,067*	*0,062*
Latakoa	*0,085*	*0,088*	*0,091*	*0,096*	*0,093*	*0,095*	*0,079*	*0,078*	*0,065*	*0,068*

Demagun populazioak normalak direla. Erantzun honako galdera hauei:

(a) Kalkula ezazu populazioaren batezbestekoen arteko diferentziarako % 98 mailako KT. Ondoriozta daiteke diferentzia adierazgarririk dagoela?
(b) Diferentziak badaude, ondoriozta dezakegu tomate freskoen kobre kopurua latakoena baino baxuagoa dela?
(c) Ondorioztatu daiteke diferentzia gutxienez 0,003 dela?

Hasteko, galdera hauei guztiei erantzuteko, $I^{0,98}_{\mu_1-\mu_2}$ tartea kalkulatu behar dugu, baina kasu honetan datuak binakakoak dira. `Freskoa` eta `Latakoa` aldagaiek `tomate` datu-basea osatzen dute, eta datuak sartu edo **praktikadatuak.xls** fitxategiko **tomate** orrialdetik inportatu ostean, honela egin dezakegu batezbestekoen diferentziarako konfiantza-tartearen kalkulua:

`Estatistikoak` → `Batezbestekoak` → `T test binakatua` → `Lehen aldagaia[freskoa] Bigarren aldagaia[latakoa]. Hipotesi alternatiboa[alde bikoa]. Konfiantza-maila [0.98].` → `Onartu`

```
        Paired t-test
data:  tomate$freskoa and tomate$latakoa
98 percent confidence interval:
 -0.019189074 -0.004210926
sample estimates:        mean of the differences
      -0.0117
```

$\overline{d} = -0,0117$ da batezbestekoen diferentziaren zenbatespen puntuala. Bi populazioen kobre kopuruaren batezbestekoen arteko diferentzia $I_{\mu_D}^{0,98} = I_{\mu_1-\mu_2}^{0,98} = (-0,019189074,\ -0,004210926)$ tartean dago, % 98ko konfiantza-mailarekin; (a) galderari erantzunez, 0 tartearen barne ez dagoenez, bi batezbestekoen artean diferentzia adierazgarriak daudela esan dezakegu. Gainera, (b) atalari erantzuteko, ohartu, tartea negatiboa denez, $\mu_1 - \mu_2 < 0$; hau da, ondoriozta daiteke tomate freskoen kobre kopurua lataraturikoena baino baxuagoa dela % 98ko konfiantza-mailaz. Azkenik, (c) galderari erantzuteko, erreparatu $-0,003 \in I_{\mu_D}^{0,98}$ dagoenez, ondoriozta daitekeela diferentzia 0,003 izatea posible dela.

3.3. Populazio binomialetarako konfiantza-tarteak

3.3.1. Proportziorako konfiantza-tarteak

5. adibidea. *Suzirien aireratze-instalazio berri bat ikertzen ari dira. Dagoen sisteman, $p = 0,8$ da jaurtiketa batean arazorik ez egoteko probabilitatea. Sistema berriarekin 40 jaurtiketa esperimental egiten dira; horietatik 34tan ez da arazorik egon.*

(a) Kalkula ezazu % 95eko p-ren konfiantza-tartea.
(b) Ondoriozta daiteke sistema berria hobea dela? Eta okerragoa dela? Esan dezakegu diferentziak adierazgarriak direla?

Hasteko, (a) atalari erantzuteko, $I_p^{0,95}$ tartea kalkulatu behar dugu. Horretarako, jaurtiketak adieraziko dituen aldagai bitar bat sortuko dugu:

`Datuak` → `Datu multzo berria` → `Sartu datu multzoaren izena:` *jaurtiketak* → `Onartu`

Orduan, *var*1 aldagaian, 34 aldiz 1 eta 6 aldiz 0 sartuko dugu. Ondoren, faktore bihurtuko dugu:

`Datuak` → `Kudeatu datu multzo aktiboko aldagaiak` → `Zenbakizko aldagaiak faktore bihurtu` → `Aldagaiak:` *var*1. `Faktore-mailak: ipini maila-zenak. Aldagai-izen berria:` *jaurtiketa* → `Onartu. 0` *arazoak*, `1` *arazorik ez*

Zenbaki txikiena, hau da, 0, *arrakasta moduan* hartuko du *Rcmdr*k. Guk, arrakasta «arazorik ez» gertaera izatea nahi dugunez, kasu honetan, faktorearen ordena aldatu behar dugu [3]:

`Datuak` → `Aldatu datu multzo aktiboko aldagaiak` → `Berrantolatu faktore-mailak` → `Faktorea:` *jaurtiketa*. `Izena faktorearentzat:` *ordenatua*. `Egin faktore ordenatua.` `Onartu` → `Ordena berria: arazoak 2,arazorik ez 1` → `Onartu`

Prozesu hau guztia ekiditeko, datuak **praktikadatuak.xls/jaurtiketa** orrialdetik inporta ditzakegu zuzenean.

Orain, proportziorako konfiantza-tartea kalkulatuko dugu banaketa binomial zehatza erabiliz (hau da, banaketa binomiala banaketa normalaren bidez, hurbildu gabe):

`Estatistikoak` → `Proportzioak` → `Lagin bakarreko proportzio-testa. Datuak:` *ordenatua*. `Aukerak: Hipotesi alternatiboa: Populazioaren proportzioa` $!= p0$. `Konfiantza-maila=` 0.95. `Test mota: binomial zehatza`

```
        Exact binomial test
data:  rbind(.Table)
number of successes = 34, number of trials = 40, p-value = 0.5541
alternative hypothesis: true probability of success is not equal to 0.8
95 percent confidence interval:
 0.7016473 0.9428977
sample estimates:
probability of success
        0.85
```

Beraz, arrakasta-proportzioaren puntu-estimazioa: $\hat{p} = \frac{34}{40} = 0,85$ eta arrakasta-proportzioaren zenbatespen-tartea $I_p^{0,95} = (0,7016,\ 0,9429)$ da;

(b) atalari erantzuteko, erreparatu $0,8$ arrakasta-proportzioa tartearen barnean dagoela; beraz, % 95eko konfiantza-mailarekin, ezin da ondorioztatu sistema berria hobea denik, ezta txarragoa denik ere; izan ere, berdintasuna ezin da baztertu eta ezin dugu esan ezberdintasun adierazgarririk dagoenik.

[3] Baliokideki, *var*1 aldagaia definitzeko 34 aldiz a (arrakasta) eta 6 aldiz p (porrota) sar dezakegu. Horrela, ez da beharrezkoa faktore bihurtzea, zenbakizkoa ez delako, eta ez da beharrezkoa berrordenatzea, ordena alfabetikotzat hartzen baitu.

4. *Rcmdr* praktika

Hipotesi-kontraste parametrikoak

Helburua

Praktika honetan, *Rcmdr* erabiliz bost hipotesi-kontraste ohikoenak nola burutu ikusiko dugu:

- `Lagin bakarreko t testa`: $H_0 : \mu = \mu_0$
- `Bi bariantza konparatzeko F testa`: $H_0 : \sigma_1^2 = \sigma_2^2$
- `Bi lagin independenteko t testa`: $H_0 : \mu_1 = \mu_2$ (lagin askeak).
- `Datu binakatuentzako t testa`: $H_0 : \mu_1 = \mu_2$ (binakako datuak).
- `Lagin bakarrerako proportzio-testa`: $H_0 : p = p_0$ (binomial zehatza)

Kontuan izan test hauek gauzatzeko aurreko praktikan ikasitako gauza asko berrerabiliko ditugula.

4.1. Lagin bakar baten hipotesi-kontrasteak

4.1.1. Batezbestekorako hipotesi-kontrasteak

Konfiantza-tarteen kasuan bezala, batezbesteko baten test ohikoena gauzatuko dugu: bariantza ezezaguneko populazio normal batentzako t-testa.

6. adibidea. *Ebatz dezagun 2. adibidea, %10eko esangura-mailako hipotesi-kontrastea planteatuz. Hiriko ur-depositua aski handia bada asteko 160 litroko batez besteko kontsumoa baimentzeko, hirian ur falta arazoak egon daitezke?*

Aurreko galderari erantzuteko, alde bateko hipotesi nulua $H_0 : \mu \leq 160$ egin nahiko genuke. `ura` izeneko aldagaiaren 25 datuak sartu edo inportatu ostean, *Rcmdr*n, jarraitu honako pauso hauek:

`Estatistikoak` → `Batezbestekoak` → `Lagin bakarreko t testa` → `Populazioaren batezbestekoa` > *mu*`0. Hipotesi nulua: mu=[160]. Konfiantza-maila [0.90]` → `Onartu`

```
data:  Datos$ura
t = 3.7897, df = 24, p-value = 0.0004474
alternative hypothesis: true mean is greater than 160
sample estimates:         mean of x    175.76
```

Aurreko emaitzan, `t` estatistikoa, `df` (degree freedom) Student-en askatasun-graduak, `p-value` p-balioa eta `alternative hypothesis` $H_1 : \mu > 160$ hipotesi alternatiboa dira.

1. Batezbestekoaren honako hipotesi-kontraste hau planteatu dugu:
   $\begin{cases} H_0 : & \mu \leq 160 \\ H_1 : & \mu > 160 \end{cases}$
2. X :='asteko kontsumitutako ur litro kopurua': $\mathcal{N}(\mu, \sigma)$, non σ ezezaguna baita. Beraz, estatistiko egokia $t_p = \frac{\overline{x}-\mu_0}{S/\sqrt{n}} = 3,7897$. Eta $T_p = \frac{\overline{X}-\mu_0}{s/\sqrt{n}} : t_{n-1} = t_{24}$
3. Alde bateko eskuinetiko kontrastea denez, p-balioa kalkulatzeko:
   $p = P(t_{24} > 3,7897) = 0,0004474$
4. Ondorioa: $p < \alpha = 0,10$ denez, populazioaren batez besteko ur litro kopurua 160 baino handiagoa izatea onartzen da % 10eko esangura-mailarekin; beraz, ur-depositua (160 litrokoa) ez da nahikoa izango.

4.2. Lagin birako hipotesi-kontrasteak

7. adibidea. *Har itzazu berriro ere 3. adibideko datuak; egin ezazu, hasteko, bi populazio-bariantzak konparatzeko hipotesi-kontrastea, eta, ondoren, bi populazio-batezbestekoen arteko erlazioa aztertzeko testa.*

Hasteko, `denbora` eta `sexua` (1=emakumezkoa, 2=gizonezkoa) aldagaiekin `denbora` izeneko datu-basea osatuko dugu, aurreko praktikan bezala.

4.2.1. Bi bariantza konparatzeko hipotesi-kontrasteak

Lehenik, populazio-bariantzen arteko erlazioa aztertu behar da; hipotesi nulua $H_0 : \sigma_1^2 = \sigma_2^2$ da, eta *Rcmdr* programak hiru aukera ematen dizkigu: F, Bartlett eta Levene testak. Beraz:

`Estatistikoak` → `Bariantzak` → `F, Bartlett edo Levene`

F testa apunteetan ikusitako metodoari dagokio:

```
        F test to compare two variances
data:  denbora by sexua
F = 0.0737, num df = 4, denom df = 6, p-value = 0.02468
alternative hypothesis: true ratio of variances is not equal to 1
```

Aurreko emaitzan, `F` estatistikoa (0,0737), `num df` eta `denom df` Fisher banaketaren askatasun-graduak (4 eta 6), `p-value` p-balioa (0,02468) eta `alternative hypothesis` $H_1 : \sigma_1 \neq \sigma_2$ hipotesi alternatiboa dira. Hori horrela, $p < \alpha \Rightarrow \sigma_1^2 \neq \sigma_2^2$ ondorioztatzen da % 10eko esangura-mailarekin.

4.2.2. Bi populazio askeren batezbestekoak konparatzeko hipotesi-kontrastea

Behin bariantzen arteko erlazioa aztertuta, batezbestekoen diferentzia nulua den ala ez kontrasta dezakegu:

`Estatistikoak` → `Batezbestekoak` → `Lagin independenteko t testa` → `Taldeak` [*sexua*] `Mendeko aldagaia` [*denbora*]`. Hipotesi alternatiboa [alde bikoa]. Konfiantza-maila [0.90]. Bariantzak berdinak direla jo? [ez]` → `Onartu`

```
        Welch Two Sample t-test

data:  denbora by sexua
t = -0.9638, df = 7.187, p-value = 0.3664
alternative hypothesis: true difference in means is not equal to 0
```

Aurreko emaitzan, `t` estatistikoa (-0,9638), `num df` Student banaketaren askatasun-graduak (7,187), `p-value` p-balioa (0,3664) eta `alternative hypothesis` $H_1 : \mu_1 - \mu_2 \neq 0$ hipotesi alternatiboa dira. Horrela, $p > \alpha = 0,10$ denez, ezin da ondorioztatu batezbesteko bat bestea baino handiagoa denik; ezin da baztertu sexuen arteko batez besteko denboren berdintasuna.

4.2.3. Binakako datuen batezbestekoak konparatzeko hipotesi-kontrastea

8. adibidea. *Har ezazu berriro 4. adibidea, eta plantea itzazu honako galderei erantzuteko hipotesi kontraste egokiak:*

(a) % 5eko esangura mailaz, esan dezakegu diferentzia adierazgarririk dagoenik tomate fresko eta ontziratutakoen batezbesteko kobre kopuruaren artean?
(b) % 5eko esangura mailaz, esan dezakegu tomate ontziratutakoen batezbesteko kobre kopurua, tomate freskoena baino handiagoa dela?
(c) % 5eko esangura mailaz, esan dezakegu tomate fresko eta ontziratutakoen batezbesteko kobre kopuruaren diferentzia, gutxienez 0,003 dela?

Hasteko, (a) atalari erantzuteko, har dezagun honako hipotesi nulu hau $H_0 : \mu_1 = \mu_2$. Binakako datuen `tomate` datu-basea osatu ondoren (`freskoa` eta `latakoa` aldagaiekin):

`Estatistikoak` → `Batezbestekoak` → `T test binakatua` → `Lehen aldagaia [`*freskoa*`] Bigarren aldagaia [`*latakoa*`]. Hipotesi alternatiboa [alde bikoa]. Konfiantza-maila [0.98].` → `Onartu`

```
        Paired t-test
data:  tomate$freskoa and tomate$latakoa
t = -4.4079, df = 9, p-value = 0.001701
alternative hypothesis: true difference in means is not equal to 0
```

$p < \alpha = 0,02$ denez, ondoriozta daiteke tomate fresko eta ontziratutakoen kobre kopuruaren batezbestekoen arteko diferentzia nabarmena dela, % 2ko esangura-mailarekin.

Orain, (b) atalari erantzuteko, azter dezagun honako hipotesi nulu hau: $H_0 : \mu_1 \geq \mu_2$:

`Estatistikoak` → `Batezbestekoak` → `T test binakatua` → `Lehen aldagaia [`*freskoa*`] Bigarren aldagaia [`*latakoa*`]. Hipotesi alternatiboa [diferentzia<0]. Konfiantza-maila[0.98].` → `Onartu`

```
        Paired t-test
data:  tomate$freskoa and tomate$latakoa
t = -4.4079, df = 9, p-value = 0.0008504
alternative hypothesis: true difference in means is less than 0
```

$p < \alpha = 0,02$ denez, ondoriozta daiteke tomate freskoen kobre kopurua lataraturikoena baino baxuagoa dela, % 2ko esangura-mailarekin.

Azkenik, (c) atalari erantzuna emateko, azter dezagun honako hipotesi nulu hau: $H_0 : \mu_1 + 0,003 \leq \mu_2$. Lehenik, $X_1' = X_1 + 0,003$ aldagai berria, `freskoa003` izenekoa, kalkulatu behar da:

`Datuak` $\rightarrow$ `Kudeatu datu multzo aktiboko aldagaiak` $\rightarrow$ `Kalkulatu aldagai berria` $\rightarrow$ `Aldagai berriaren izena [`$freskoa003$`]. Konputatzeko adierazpena [`$freskoa + 0.003$`].` $\rightarrow$ `Onartu`

Jarraian,

```
data:  tomate$freskoa003 and tomate$latakoa
t = -3.2776, df = 9, p-value = 0.9952
alternative hypothesis: true difference in means is greater than 0
```

$p > \alpha = 0,02$ denez, ondoriozta daiteke ontziraturikoen batez besteko kobre kopurua tomate freskoena baino gutxienez 0,003 altuagoa dela, % 2ko esangura-mailarekin.

Gainera, eskuz egiten dugun moduan, $D = X_1 - X_2$ diferentzia aldagaia sor dezakegu. Ondoren, 4.1.1. atalean ikusitakoa aplika daiteke, diferentzia aldagaia erabiliz eta μ_0 0,003 balioan finkatuz.

`Datuak` $\rightarrow$ `Kudeatu datu multzo aktiboko aldagaiak` $\rightarrow$ `Kalkulatu aldagai berria` $\rightarrow$ `Aldagai berriaren izena [`$diferentzia$`]. Konputatzeko adierazpena [`$freskoa - latakoa$`].` $\rightarrow$ `Onartu`

Behin diferentzia aldagaia sortuta, batezbestekoa eta desbiderapena puntualki estima ditzakegu:

`Estatistikoak` $\rightarrow$ `Laburpenak` $\rightarrow$ `Zenbakizko laburpenak` $\rightarrow$ `Datuak: [`$diferentzia$`]` $\rightarrow$ `Onartu`

mean	sd	IQR	0%	25%	50%	75%	100%	n
-0.0117	0.008393781	0.01325	-0.022	-0.01975	-0.0085	-0.0065	0.002	10

Beraz, $\hat{\mu}_D = \overline{d} = -0,0117$ eta $\hat{\sigma}_D = s_{n-1,D} = 0,008394$.

Gainera, $H_0 : \mu_1 - \mu_2 \leq -0,003$ hipotesi-kontrastea egiteko lagin bakarreko t testa erabiliko dugu:

`Estatistikoak` $\rightarrow$ `Batezbestekoak` $\rightarrow$ `Lagin bakarreko t testa` $\rightarrow$ `Aldagaia: [`$diferentzia$`]. Populazio batezbestekoa >` $mu0$`. Hipotesi nulua: mu` $\leq$ `[-0.003]. Konfiantza-maila [0.98]` $\rightarrow$ `Onartu`

Emaitza:

```
        One Sample t-test
data:  diferentzia
t = -3.2776, df = 9, p-value = 0.9952
alternative hypothesis: true mean is greater than -0.003
95 percent confidence interval:
 -0.01656572         Inf
sample estimates:
mean of x
  -0.0117
```

1. Batezbestekoen diferentziaren honako hipotesi-kontraste hau planteatu dugu:
   $\begin{cases} H_0: & \mu_1 - \mu_2 \leq -0,003 \\ H_1: & \mu_1 - \mu_2 > -0,003 \end{cases}$
2. Estatistiko egokia $t_p = \frac{\overline{d}-(-0,003)}{S_D/\sqrt{n}} = -3,2776$.
3. Alde bateko eskuinetiko kontrastea denez, p-balioa kalkulatzeko:
   $p = P(t_9 > -3,2776) = 0,9952$
4. Ondorioa: $p > \alpha = 0,02$ denez, ezin dugu baztertu populazioaren batez besteko kobre kopuruaren diferentzia -0,003 baino txikiagoa ala berdina dela, %2ko esangura-mailaz. Beraz, balio absolutuan, ezin dugu baztertu diferentzia gutxienez 0,003koa denik.

4.3. Populazio binomialetarako hipotesi-kontrasteak

4.3.1. Proportziorako hipotesi-kontrasteak

9. adibidea. *Har ezazu berriro ere 5. adibidea eta erantzun itzazu bertan planteatutako hiru galderak hipotesi-kontraste parametrikoak erabiliz:*

(a) Ondoriozta daiteke sistema berria hobea dela?
(b) Eta okerragoa dela?
(c) Esan dezakegu diferentziak adierazgarriak direla?

Hasteko, `Jaurtiketa` izeneko aldagaia `(0: arazorik ez, 1:arazoak)` eraiki edo **praktika-datuak.xls/jaurtiketa** orrialdetik inportatu.

Ondoren, (a) atalari erantzungo diogu eskuinetiko hipotesi-kontrastea eginez ($H_1 > 0,8$):

`Estatistikoak` $\rightarrow$ `Proportzioak` $\rightarrow$ `Lagin bakarreko proportzio-testa` $\rightarrow$ `Aukerak:`
`Hipotesi alternatiboa: Populazioaren proportzioa` $> p0$`. Hipotesi nulua:` $p = 0.8$`.`
`Konfiantza-maila=` 0.95`. Test mota: binomial zehatza`

```
number of successes = 34, number of trials = 40, p-value = 0.2859
alternative hypothesis: true probability of success is greater than 0.8
95 percent confidence interval:
 0.7252555 1.0000000
```

p-balioa $p = 0,2859 > \alpha = 0,05$ denez, H_1 baztertzen da; hots, $p \not> 0,8$. Ezin da esan sistema berria hobea denik.

Segidan, (b) ataleko galderari erantzuteko, $H_1 : p < 0,8$ hipotesi alternatiboko alde bakarreko kontrastea egingo dugu:

`Estatistikoak` $\rightarrow$ `Proportzioak` $\rightarrow$ `Lagin bakarreko proportzio-testa` $\rightarrow$ `Aukerak:`
`Hipotesi alternatiboa: Populazioaren proportzioa` $< p0$`. Hipotesi nulua:` $p = 0.8$`.`
`Konfiantza-maila=` 0.95`. Test mota: binomial zehatza`

```
number of successes = 34, number of trials = 40, p-value = 0.8387
alternative hypothesis: true probability of success is less than 0.8
95 percent confidence interval:
 0.000000 0.932599
```

Ezkerraldetiko alde bateko kontraste honi dagokionez, p-balioa $p = 0,8387 > \alpha = 0,05$ denez, H_1 baztertzen da; hots, $p \not< 0,8$ eta ezin da esan sistema berria okerragoa denik ere.

Azkenik, (c) galderari erantzuteko $H_0 : p = 0,8$ hipotesi nulua duen hipotesi-kontrastea ebatziko dugu *Rcmdr* erabiliz:

`Estatistikoak` $\rightarrow$ `Proportzioak` $\rightarrow$ `Lagin bakarreko proportzio-testa` $\rightarrow$ `Aukerak:`
`Hipotesi alternatiboa: Populazioaren proportzioa` $!= p0$`. Hipotesi nulua:` $p = 0.8$`.`
`Konfiantza-maila=` 0.95`. Test mota: binomial zehatza`

```
data:  rbind(.Table)
number of successes = 34, number of trials = 40, p-value = 0.5541
alternative hypothesis: true probability of success is not equal to 0.8
95 percent confidence interval:
 0.7016473 0.9428977
sample estimates:
probability of success
      0.85
```

Aurreko emaitzan, `number of successes` arrakasta kopurua (34), `number of trials` proba kopuru osoa (40), `p-value` p-balioa (0,5541) eta `alternative hypothesis` $H_1 : p \neq 0,8$ hipotesi alternatiboa dira. Alde biko kontraste honi dagokionez, p-balioa $p = 0,5541 > \alpha = 0,05$ denez, H_0 ez da errefusatzen. Hots, $p = 0,8$ ezin dugu baztertu, edo, beste era batera esanda, ezin da baztertu sistema berriaren arrakasta % 80koa denik. Beraz, sistema berria eta zaharraren artean ez daude ezberdintasun adierazgarririk.

5. *Rcmdr* praktika

Hipotesi-kontraste ez-parametrikoak

Helburua

Praktika honetan, hipotesi-kontraste ez-parametriko mota batzuk *Rcmdr* bidez nola egin erakutsiko dugu, zehazki, doikuntza-egokitasunerako probak eta independentzia- eta homogeneotasun-probak.

5.1. Doikuntza-egokitasunerako kontrasteak

5.1.1. Pearsonen khi karratu kontrastea

10. adibidea. *Mendel-en legeak esperimentalki egiaztatu nahi ditugu. Horretarako, 500 landare gurutzatu ditugu, eta, teoriaren arabera, lore gorri, arrosa, hori eta zuriko landare kopuruek, hurrenez hurren, proportzionalak izan beharko lukete 8, 12, 10 eta 20 zenbakiekin. Lorturiko datuak 70, 126, 96 eta 208 izan ziren, hurrenez hurren.*

Gogoratu mota honetako hipotesi-kontrasteetan, honako hipotesi nulu hau honela definitzen dela:

$$H_0 : (p_1, p_2, \ldots, p_k) = (p_1^0, p_2^0, \ldots, p_k^0)$$

Adibidearen kasuan, beraz, honako hipotesi nulu hau dugu:

$$H_0 : (p_1, p_2, p_3, p_4) = (0,16,\ 0,24,\ 0,20,\ 0,40)$$

Hipotesi-kontraste hau gauzatzeko, lehenengoz datu-basea eraiki behar dugu:

`Datuak` → `datu multzo berria` → `Datuak`

500 behaketa direnez, datu guztiak eskuz sartzea nahiko neketsua izan daiteke. Beraz, *R*ko kodea erabiliko dugu. Horretarako, idatzi honako hau *R Script* izeneko leihoan, eta, bi lerroak hautatu ostean, sakatu exekutatu:

```
> balioak <- c(rep(1,70),rep(2,126),rep(3,96),rep(4,208))
> Datuak <- data.frame(var1=balioak)
```

Orain, sortu berri dugun zenbakizko aldagai hau ($var1$) faktore bihurtzeko:

`Datuak` → `Aldatu datu multzo aktiboko aldagaiak` → `Zenbakizko aldagaiak faktore bihurtu` → $var1$. `Faktore-mailak: eman izena mailei. 1:` *gorria*, `2:` *arrosa*, `3:` *horia*, `4:` *zuria*

Prozesu hau ekidin nahi badugu, datu hauek berak **praktikadatuak.xls** fitxategiko **mendel** fitxategian daude gordeta.

Orain, *Rcmdr*n bidez, Pearsonen χ^2 proba egiteko:

`Estatistikoak` → `Laburpenak` → `Maiztasun-banaketak` → `Aldagaia:` $var1$. `Doikuntza-egokitasunerako Khi karratuaren proba (aldagai batentzako bakarrik)` → `a Onartu` → `Hypothesized Probabilitites {0.16, 0.24, 0.20, 0.40}` → `Onartu`

```
counts:
var1
gorria arrosa  horia  zuria
    70    126     96    208

percentages:
var1
gorria arrosa  horia  zuria
  14.0   25.2   19.2   41.6

        Chi-squared test for given probabilities

data:  .Table
X-squared = 2.03, df = 3, p-value = 0.5662
```

1. $X :=$«landareen kolorea», non p_1, p_2, p_3 eta p_4, gorria, arrosa, horia eta zuria izateko proportzioak baitira, hurrenez hurren. Doikuntza-egokitasunerako kontrastea planteatu dugu:
$$\begin{cases} H_0: & (p_1, p_2, p_3, p_4) = (0,16,\ 0,24,\ 0,20,\ 0,40) \\ H_1: & (p_1, p_2, p_3, p_4) \neq (0,16,\ 0,24,\ 0,20,\ 0,40) \end{cases}$$

2. $e_i > 5$ baldintzak betetzen dira. Beraz, estatistiko egokia $\chi^2_p = 2,03$. Eta $\chi^2_p \sim \chi^2_3$
3. p-balioa kalkulatzeko:
   $p = P(\chi^2_3 > 2,03) = 0,5662$
4. Ondorioa: $p > \alpha = 0,05$ denez, H_0 ezin da baztertu, hots, landare-koloreen kopuruak 8, 12, 10 eta 20 zenbakiekiko proportzionalak dira.

5.1.2. Normaltasunerako kontrasteak

Khi karratuaren doikuntza-egokitasun testa datuen normaltasuna kontrastatzeko erabil daiteke. Alabaina, badaude normaltasuna kontrastatzeko doikuntza-egokitasun proba espezifiko egokiago batzuk. Hori horrela, *Rcmdr*n normaltasuna aztertzeko proba bat baino gehiago ditugu eskuragai. Jarraian, bi adibide azalduko ditugu.

Shapiro-Wilk proba (n <50 denean)

Shapiro-Wilk normaltasun-testa erabiliko dugu, gure lagina txikia denean ($n < 50$) normaltasuna kontrastatzeko.

11. adibidea. *Kontrasta ezazu honako datu hauek banaketa normaletik ateratakoak diren hipotesia: 20, 22, 24, 30, 31, 32, 38.*

Lehenengoz, datu-basea eraikiko dugu *Rcmdr*n balioak banan-banan sartuz:

`Datuak` → `datu multzo berria` → `Datuak`

`Estatistikoak` → `Laburpenak` → `Normaltasun-testa(Test of normality)` → `var1` → `Shapiro-Wilk` → `Ados`

```
        Shapiro-Wilk normality test
data:  Dato$var1
W = 0.9478, p-value = 0.7096
```

1. Doikuntza-egokitasunerako kontrastea planteatu dugu:
   $\begin{cases} H_0: & X \approx \mathcal{N}(28,1429,\ 6,3882) \\ H_1: & X \not\approx \mathcal{N}(28,1429,\ 6,3882) \end{cases}$
2. Beraz, Shapiro-Wilken estatistiko egokia $W = 0,9478$.
3. p-balioaren kalkulua $p = 0,7096$

4. Ondorioa: $p > \alpha = 0,05$ denez, H_0 ezin da baztertu; hots, ezin da baztertu aldagaiaren normaltasuna.

Lilliefors (Kolmogorov-Smirnov) testa ($n \geq 50$ denean)

Normaltasun-test hau erabiliko dugu, soilik, oso lagin handiak dauzkagunean ($n \geq 50$).

12. adibidea. *Birus baten latentzia-aldia ikertzeko, 90 txitari inokulatu zitzaien birusa. Bakoitzarengan gaixotasunaren lehenengo sintomak agertu arte pasatutako egun kopurua aztertu zen. Beraz, X = egun kopurua izeneko aldagaia da. Honako hauek ziren lortutako datuak:*

8	10	8	14	16	9	12	13	9	12	12	10	15	8	6
5	9	11	13	5	9	12	13	8	14	8	5	14	6	13
7	8	12	12	8	6	8	9	9	15	8	9	8	13	7
9	12	8	6	9	14	13	8	12	9	11	8	16	10	6
10	13	6	5	14	12	14	6	11	12	10	12	6	7	10
6	15	7	9	5	9	7	10	7	10	8	11	11	14	15

Azter ezazu ea datu hauek banaketa normala jarraitzen duten.

Lehenengoz, datuak eskuz sartu edo **praktikadatuak.xls** fitxategiko **birusa** orritik inportatuko ditugu. Ondoren,

`Estatistikoak` → `Laburpenak` → `Normaltasun-testa (Test of normality)` → `birusa` → `Lilliefors (Kolmogorov-Smirnov)` → `Ados`

```
        Lilliefors (Kolmogorov-Smirnov) normality test
data:  birusa
D = 0.12728, p-value = 0.00104
```

1. Doikuntza-egokitasunerako kontrastea planteatu dugu:
$$\begin{cases} H_0: & X \approx \mathcal{N}(9,8778,\ 2,9711) \\ H_1: & X \not\approx \mathcal{N}(9,8778,\ 2,9711) \end{cases}$$

2. Beraz, Lilliefors (Kolmogorov-Smirnov) estatistiko egokia $D = 0,12728$.

3. p-balioaren kalkulua $p = 0,00104$.

4. Ondorioa: $p < \alpha = 0,05$ denez, H_0 baztertzen dugu; hots, aldagaiaren normaltasuna baztertzen dugu.

5.2. Independentzia-kontrastea eta homogeneotasun-kontrastea

13. adibidea. *Urin gastrikoaren azidotasun-maila desberdinen (baxua: aklorhidria edo hipoklorhidria; eta altua: normal edo hiperklorhidria) eta gaixotasun motaren (ultzera gastrikoa eta minbizia) artean menpekotasunik dagoen aztertu nahi dugu. Lorturiko emaitzak honako 2×4 kontingentzia-taulan adierazten dira:*

X / Y	*Aklorhidria*	*Hipoklorhidria*	*Normal*	*Hiperklorhidria*
Ultzera gastrikoa	*3*	*7*	*35*	*9*
Minbizia	*22*	*2*	*6*	*0*

Erantzun itzazu honako galdera hauek:

a) Kontrasta ezazu, azidotasun gastrikoaren eta gaixotasun motaren arteko independentzia, % 1eko esangura-mailarekin.

b) Laginean, zein da minbizia pairatzen dutenen ehunekoa?

c) Laginean, zein da minbizia pairatzen dutenen eta azidotasun-maila normala edo altua dutenen ehunekoa?

d) Minbizia pairatzen dutenen artean, zein da azidotasun normala edo altua dutenen ehunekoa?

e) Azidotasun normala edo altua dutenen artean, zein da minbizia dutenen ehunekoa?

Behatutako maiztasun batzuk txikiak dira; beraz, hasteko, egiazta dezagun $e_{ij} > 5$ baldintza betetzen dela i eta j guztietarako. Horretarako:

`Estatistikoak` → `Kontingentzia-taulak` → `Sartu eta analizatu sarrera bikoitzeko taula. Errenkada kopurua: 2. Zutabe kopurua: 4. Maiztasunak sartu:`

3	7	35	9
22	2	6	0

`Estatistikoak. Kalkulatu ehunekoak: ehunekorik ez. Hypothesis test: Independentziaren khi karratuaren proba. Inprimatu espero izandako maiztasunak.`

```
# Expected Counts
          1        2        3        4
1 16.071429 5.785714 26.35714 5.785714
2  8.928571 3.214286 14.64286 3.214286
```

Izan ere, itxarondako maiztasunen taula honako hau da:

e_{ij}	Aklorhidria	Hipoklorhidria	Normal	Hiperklorhidria
Ultzera gastrikoa	16.071429	5.785714	26.35714	5.785714
Minbizia	8.928571	3.214286	14.64286	3.214286

Aurreko taulan ikusten denez, e_{22} eta e_{24} itxarondako maiztasunak 5 baino txikiagoak direnez, beste taula bat sortu behar da. Taula hori eratzeko, lehenengo zutabea bigarrenarekin (azidotasun-maila baxua) eta laugarren zutabea hirugarrenarekin (azidotasun-maila altua) bilduko ditugu, honako 2x2 dimentsioko kontingentzia-taula hau lortuz:

o_{ij}	Baxua	Altua
Ultzera gastrikoa	10	44
Minbizia	24	6

Orain, berriro ere, hipotesi nulua duen honako test hau egin nahi dugu *Rcmdr* erabiliz: H_0 : azidotasun-maila askea da gaixotasunarekiko.

`Estatistikoak` → `Kontingentzia-taulak` → `Sartu eta analizatu sarrera bikoitzeko taula. Errenkada kopurua: 2. Zutabe kopurua: 2. Maiztasunak sartu:`

10	44
24	6

. `Estatistikoak. Kalkulatu ehunekoak: ehunekorik ez. Hypothesis test: Independentziaren khi karratu proba.`

```
# Counts
   1  2
1 10 44
2 24  6
```

```
        Pearson's Chi-squared test
data:  .Table
X-squared = 30.2576, df = 1, p-value = 3.783e-08
```

Egiazta dezagun determinantea $|o_{11}o_{22} - o_{12}o_{21}| > \frac{n}{2}$ denez, Yatesen zuzenketa beharrezkoa dela. *Rcmdr*k zuzenketa hori defektuz egiten ez duenez, Rscript fitxategian `correct=TRUE` jarri beharko dugu modu honetan:

```
> .Table <- matrix(c(10,44,24,6), 2, 2, byrow=TRUE)
> rownames(.Table) <- c('ultzera', 'minbizia')
> colnames(.Table) <- c('baxua', 'altua')
> .Table  # Counts
> .Test <- chisq.test(.Table, correct=TRUE)
> .Test
```

Kode zati hori hautatu ostean, **Exekutatu** botoia sakatu eta Yatesen zuzenketa aplikatuko dugu, honako emaitzak hauek lortuz:

```
        Pearson's Chi-squared test with Yates' continuity correction
data:  .Table
X-squared = 27.7595, df = 1, p-value = 1.374e-07
```

Orain, testaren emaitzak interpretatuko ditugu:

1. X :='gaixotasun mota' eta Y :='azidotasun-maila'. Independentzia-kontrastea planteatu dugu:
   $\begin{cases} H_0: & X \text{ eta } Y \text{ askeak} \\ H_1: & X \text{ eta } Y \text{ mendekoak} \end{cases}$
2. Beraz, estatistiko egokia $\chi^2_Y = 27,7595$. Eta $\chi^2_Y := \chi^2_1$
3. p-balioa kalkulatzeko:
   $p = P(\chi^2_1 > 27,7595) = 1,374 \cdot 10^{-7}$
4. Ondorioa: $p < \alpha = 0,05$ denez, H_0 errefusatzen da; hots, menpekoak dira.

Bestalde, itxarondako maiztasunak eta hondarrak ere eska daitezke.

`Estatistikoak` → `Kontingentzia-taulak` → `Sartu eta analizatu sarrera bikoitzeko taula. Errenkada kopurua: 2. Zutabe kopurua: 2. Maiztasunak sartu:`

10	44
24	6

.

`Estatistikoak. Kalkulatu ehunekoak: ehunekorik ez. Hypothesis test: Independentziaren khi karratu proba. Khi karratuaren estatistikoaren osagaiak. Inprimatu espero izandako maiztasunak.`

```
# Expected Counts
                baxua    altua
u. gastrikoa 21.85714 32.14286
minbizia     12.14286 17.85714

# Chi-square Components
      1    2
1  6.43 4.37
2 11.58 7.87
```

Hau da,

Itxarondako maiztasunak		
e_{ij}	Baxua	Altua
Ultzera gastrikoa	21.85714	32.14286
Minbizia	12.14286	17.85714

Hondarrak		
$(o_{ij} - e_{ij})^2/e_{ij}$	Baxua	Altua
Ultzera gastrikoa	6.43	4.37
Minbizia	11.58	7.87

Azkenik, datuen ehunekoak era errazean kalkulatzeko:

(i) Errenkaden ehunekoak:

`Estatistikoak` → `Kontingentzia-taulak` → `Sartu eta analizatu sarrera bikoitzeko taula. Errenkada kopurua: 2. Zutabe kopurua: 2. Maiztasunak sartu:` $\begin{array}{|c|c|}\hline 10 & 44 \\ \hline 24 & 6 \\ \hline \end{array}$. `Estatistikoak. Kalkulatu ehunekoak: errenkaden ehunekoak.`

```
# Row Percentages
     1    2 Total Count
1 18.5 81.5   100    54
2 80.0 20.0   100    30
```

(ii) Zutabeen ehunekoak:

`Estatistikoak` → `Kontingentzia-taulak` → `Sartu eta analizatu sarrera bikoitzeko taula. Errenkada kopurua: 2. Zutabe kopurua: 2. Maiztasunak sartu:` $\begin{array}{|c|c|}\hline 10 & 44 \\ \hline 24 & 6 \\ \hline \end{array}$. `Estatistikoak. Kalkulatu ehunekoak: zutabeen ehunekoak.`

```
# Column Percentages
          1   2
1      29.4  88
2      70.6  12
Total 100.0 100
Count  34.0  50
```

(iii) Guztirakoen ehunekoak:

`Estatistikoak` → `Kontingentzia-taulak` → `Sartu eta analizatu sarrera bikoitzeko taula. Errenkada kopurua: 2. Zutabe kopurua: 2. Maiztasunak sartu:` $\begin{array}{|c|c|}\hline 10 & 44 \\ \hline 24 & 6 \\ \hline \end{array}$. `Estatistikoak. Kalkulatu ehunekoak: guztirakoen ehunekoak.`

```
# Percentage of Total
          1    2 Total
1     11.9 52.4  64.3
2     28.6  7.1  35.7
Total 40.5 59.5 100.0
```

Honekin, adibideko azken lau galderak erantzun ditzakegu:

(b) %35,7, (c) %7,1, (d) %20, (e) %12.

14. adibidea. *EAEn egindako azterketa soziologiko batean, ikertu nahi dute galdera baten erantzunak probintziaka banatzen direnentz. Taulan agertzen den maiztasun-banaketa lortu zen. Onar daiteke erantzuna probintziaren menpe dagoela?*

	A	*B*	*G*
Alde	*11*	*13*	*9*
Kontra	*32*	*28*	*27*
Abstentzioa	*7*	*9*	*14*

Adibide honetan, datuak 3x3 kontingentzia-taulan agertzen dira, eta ez da beharrezkoa datuak biltzea, itxarondako maiztasunak 5 baino handiagoak direlako. Hipotesi nulua H_0 : erantzuna eta probintzia askeak izatea da.

```
        Pearson's Chi-squared test
data:  .Table
X-squared = 3.81, df = 4, p-value = 0.4323


# Expected Counts
            Araba Bizkaia Gipuzkoa
alde           11      11       11
kontra         29      29       29
abstentzioa    10      10       10
```

$\chi^2_p = 3,81$ estatistikoa da, $p - balioa = 0,4323$ da. Interpretazioa: $p > \alpha = 0,05$ denez, H_0 onartzen da, hots, askeak dira.

5.3. Populazioak konparatzeko kontraste ez-parametrikoak

Normaltasuna eta beste zenbait baldintza betetzen ez direnerako, *Rcmdr*n populazioak konparatzeko honako test hauek egin ditzakegu:

- **Wilcoxonen lagin bakarreko testa:** populazio bakar baten batezbestekoaren inferentzia egiteko.
- **Bi lagineko Wilcoxonen testa:** bi lagin askeren batezbestekoak konparatzeko testa.
- **Lagin parekatuen Wilcoxonen testa:** binakako datuen batezbestekoak konparatzeko testa.
- **Kruskal-Wallis testa:** bi populazio aske baino gehiagoren batezbestekoak konparatzeko testa.
- **Friedman-en hein-baturaren testa:** bi populazio erlazionatu (binakako) baino gehiagoren batezbestekoak konparatzeko testa.

Test hauek, honako agindu hauei jarraituz egin ditzakegu:

`Estatistikoak` → `Test ez-parametrikoak`

Argibide gehiago eta adibideak ikusteko, joan 5. *R* praktikara.

6. *Rcmdr* praktika

Bariantza-analisia

Helburua

Praktika honen jomuga da ANOVA edo bariantza-analisia aztertzea. Faktore bakarreko analisia hartuko da populazioaren batezbestekoak konparatzeko asmoz. Diferentziak adierazgarriak diren kasuetan, batezbestekoen binakako berdintasuna adieraziko da.

6.1. Faktore bakarreko bariantza-analisia

15. adibidea. *Hiru herrialdeetako (A, B, C) kutsadura-maila alderatu nahi da. Horretarako, diametroa 10 mm baina txikiagoa duten aireko partikula-kontzentrazioaren (PM10 partikulak) neurketa bat egiten da herrialde bakoitzeko 15 neurketa-estaziotan:*

A herrialdea	*30,36*	*28,19*	*28,16*	*28,10*	*30,42*	*32,55*	*30,78*	
	27,83	*32,16*	*32,42*	*26,36*	*29,85*	*28,54*	*32,56*	*28,26*
B herrialdea	*30,91*	*32,47*	*33,75*	*26,78*	*27,64*	*28,76*	*33,36*	
	27,80	*32,54*	*32,73*	*30,69*	*30,66*	*30,19*	*31,76*	*31,32*
C herrialdea	*49,28*	*50,57*	*50,92*	*52,09*	*50,59*	*53,77*	*46,79*	
	47,04	*49,86*	*47,90*	*50,53*	*48,80*	*49,66*	*51,40*	*47,78*

Hipotesi nulua $H_0 : \mu_1 = \mu_2 = \mu_3$ da; PM10 partikulen kontzentrazioa herrialdearekiko independentea izatea da. Bi aldagai dauzkagu, beraz, kualitatibo bat (*herrialdea*: A, B, C) eta kuantitatibo bat (*kontzentrazioa*).

Datuak inportatu (**praktikadatuak.xls** fitxategiko **kutsadura** orrialdean) edo eskuz *kutsadura* izeneko datu-basean gorde ostean, lehenik, zorizkotasuna, askatasuna, normaltasuna eta bariantzen homogeneotasunaren baldintzak egiaztatu beharko genituzke.

Lehen bi baldintzak laginketatik suposatuko ditugu. Aldiz, normaltasuna aztertzeko, $n_i < 50$ denez kasu guztietan, Shapiro-Wilk proba erabiliko dugu (ikus 5. *Rcmdr* praktikako, 5.1.2. atala). Datuak herrialde bakoitzarekiko filtratuz, probak eskatuko ditugu. Ikus dezagun, adibidez, A herrialderako proba:

```
        Shapiro-Wilk normality test
data:  kutsaduraA$kontzentrazioa
W = 0.90834, p-value = 0.1277
```

1. Normaltasuna kontrastatzeko planteamendua dugu:
   $\begin{cases} H_0: & X_A \approx \mathcal{N}(29,7693,\ 2,0197) \\ H_1: & X_A \not\approx \mathcal{N}(29,7693,\ 2,0197) \end{cases}$
2. Beraz, Shapiro-Wilken estatistiko egokia: $W = 0,90834$.
3. p-balioaren kalkulua: $p = 0,1277$.
4. Ondorioa: $p > \alpha = 0,05$ denez, H_0 ezin da baztertu, hots, ezin da A herrialdeko P10 partikulen kontzentrazioaren normaltasuna errefusatu.

Eta horrela, beste bi herrialdeen datuei Shapiro-Wilk proba eginez, $p = 0,3685$ eta $p = 0,8978$ balioak lortuko ditugu, hurrenez hurren. Beraz, ezin dugu baztertu $X_B : \mathcal{N}(30,7573,\ 2,1661)$ eta $X_C : \mathcal{N}(49,7987,\ 1,9315)$ betetzen denik.

Bi populazio baino gehiagoren bariantzen homogeneotasuna aztertzeko, Bartletten hipotesi-kontrastea eska daiteke (ikus 4. praktika, 4.2.1. atala).

`Estatistikoak` → `Bariantzak` → `Bartlett testa`

```
        Bartlett test of homogeneity of variances
data:  kontzentrazioa by herrialdea
Bartlett's K-squared = 0.18242, df = 2, p-value = 0.9128
```

1. Bariantzen homogeneotasunaren kontrastea planteatu dugu:
   $\begin{cases} H_0: & \sigma_A^2 = \sigma_B^2 = \sigma_C^2 \\ H_1: & \textit{kontrakoa} \end{cases}$
2. Beraz, Bartlett-en estatistiko egokia: $K - \textit{squared} = 0,18242$.
3. p-balioa: $p = 0,9128$ da.

4. Ondorioa: $p > \alpha = 0,05$ denez, ezin da hipotesi nulua baztertu, hots, bariantzen berdintasuna ezin dugu errefusatu % 5eko esangura-mailarekin.

Kontuan izan Bartlett testak normaltasuna dagoenean ematen dituela emaitzarik fidagarrienak. Normaltasuna ez bada betetzen, egokiagoa da Levene testa erabiltzea.

Orain, baldintzak betetzen direla egiaztatu ostean, *Rcmdr* erabiliz, bariantza-analisia egin dezakegu:

`Estatistikoak` → `Batezbestekoak` → `Faktore bakarreko bariantza-analisia`

```
            Df  Sum Sq Mean Sq F value Pr(>F)
herrialdea   2   3824  1911.8   458.8 <2e-16 ***
Residuals   42    175     4.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Batezbestekoen berdintasunerako ANOVA kontrastea planteatu dugu:
   $\begin{cases} H_0: & \mu_A^2 = \mu_B^2 = \mu_C^2 \\ H_1: & \textit{kontrakoa} \end{cases}$

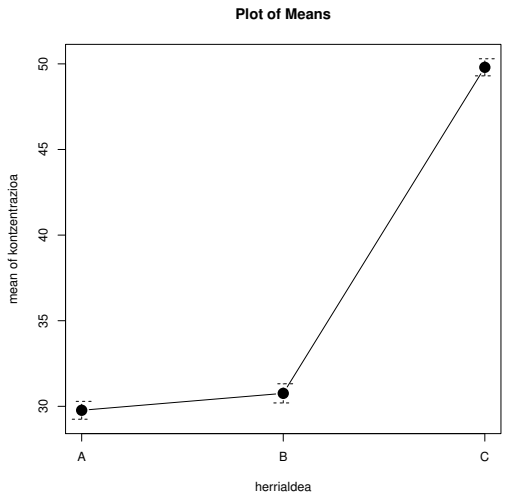
2. Beraz, estatistiko egokia: $F = 458,8$.

3. p-balioaren kalkulua: $p < 2 \cdot 10^{-16}$.

4. Ondorioa: $p < \alpha = 0,05$ denez, ezin da ondorioztatu batezbesteko guztiak berdinak direnik.

Batezbestekoen % 95eko konfiantza-tarteen grafikoa egiteko:

`Grafikoak` → `Marraztu batezbestekoen grafikoa` → `Datuak: Faktoreak [`*herrialdea*`]. Mendeko aldagaiak [`*kontzentrazioa*`] Aukerak. Errore-barrak: konfiantza-tarteak. Konfiantza-maila: 0.95`



Baina zer ondoriozta daiteke herrialde ezberdinen kutsadura-mailen arteko diferentziei buruz? Konparaketa anitzak egin beharko ditugu Tukey-ren metodoa erabiliz. Horretarako, *Rcmdr*n honako pauso hauek jarraitu:

`Estatistikoak` → `Batezbestekoak` → `Faktore bakarreko bariantza-analisia` → `Batezbestekoen alderaketa binakatuak`

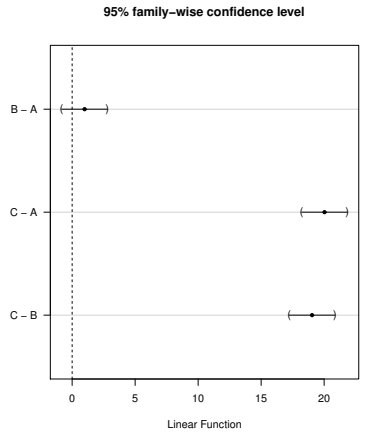
Lehenengo atalean, Tukeyren hipotesi-kontrasteak agertzen dira, herrialdeen binakako batez besteko kontzentrazioen berdintasuna aztertuz:

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = kontzentrazioa ~ herrialdea, data = kutsadura)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
B - A == 0   0.9880     0.7454   1.325    0.389
C - A == 0  20.0293     0.7454  26.870   <1e-04 ***
C - B == 0  19.0413     0.7454  25.545   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Bigarren atalean, Tukeyren %95eko konfiantza-tarteak agertzen dira, herrialdeen binakako batez besteko altueren diferentzia zenbatesteko.

```
         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = kontzentrazioa ~ herrialdea, data = kutsadura)

Quantile = 2.4289
95% family-wise confidence level


Linear Hypotheses:
           Estimate lwr     upr
B - A == 0  0.9880  -0.8225  2.7985
C - A == 0 20.0293  18.2188 21.8399
C - B == 0 19.0413  17.2308 20.8519
```

Konfiantza-tarte horien adierazpen grafiko bat ere automatikoki agertuko zaigu:



Eta hirugarren atalean, multzo homogeneoen sailkapena agertzen da.

```
A   B   C
"a" "a" "b"
```

Tukeyren alde biko kontraste edo konfiantza-tarteetan ikus daitekeenez, ez dago ezberdintasun adierazgarririk % 5eko esangura-mailarekin A eta B herrialdeen batezbestekoen artean (p-balioa=0,389>0,05 eta $0 \in I_{\mu_A-\mu_B} = (-0,8225,\ 2,7985)$). Aldiz, C herrialdeak beste bi herrialdeek baino PM10 partikula-kontzentrazio altuagoa dauka %5eko esangura-mailaz (p-balioak <1e-04 eta bi konfiantza-tarteak positiboak dira).

Gainera, **bi multzo homogeneo** osatzen dira: (a) kutsadura baxua duten herrialdeak (A, B) (b) kutsadura altua duen herrialdea (C).

6.2. Faktore biko bariantza-analisia

Faktore biko bariantza-analisia egiteko (interakzioarekin, $n > 1$):

`Estatistikoak` → `Batezbestekoak` → `Faktore anizkoitzeko bariantza-analisia` → `Faktoreak[faktorea1, faktorea2]. Mendeko aldagaia[aldagaia]`

Informazio gehiago eta adibideak ikusteko, ikusi 6. *R* praktika.

7. *Rcmdr* praktika

Erregresioa

Helburua

Praktika honen helburua da erregresio linealean eta anizkoitzean aplikatzen diren teknikak *Rcmdr* erabiliz nola aplikatu azaltzea.

- Lehenik, populazio-ereduaren itxura proposatzeko grafikoak eta teknikak deskribatuko ditugu, parametroak zenbatesteko bidea adieraziz.
- Bigarrenik, erregresio-ereduaren erabilgarritasuna nola aztertu zehaztuko dugu, bai adierazgarritasun orokorra, bai koaldagai bakoitzaren garrantzia aztertuz.
- Azkenik, iragarpenak egiteko prozedura azalduko dugu.

7.1. Adibidea

Praktika honetako helburuak lantzeko, honako adibide honetan oinarrituko gara.

16. adibidea. *Elektrizitate-konpainia batean, X etxearen neurriaren (oin karratutan) eta Y etxebizitzaren hileko energia-kontsumoa (kwh-tan) aztertu nahi da. Horretarako, 10 etxebizitza aztertu dira honako datuak lortuz:*

Neurria	*1 290*	*1 350*	*1 470*	*1 600*	*1 710*	*1 840*	*1 980*	*2 230*	*2 400*	*2 930*
Kontsumoa	*1 182*	*1 172*	*1 264*	*1 493*	*1 571*	*1 711*	*1 804*	*1 840*	*1 956*	*1 954*

Etxearen neurriaren menpeko hileko energia-kontsumoaren (kwh-tan) erregresio-eredu egoki bat doitu nahi dugu.
(a) Proposatu eta grafikoki aztertu erregresio-eredu batzuk.
(b) Kalkulatu erregresio-eredu horien koefizienteak eta aztertu haien doikuntza-egokitasuna, metodo

deskribatzaileak erabiliz.
(c) Erabili metodo inferentzialak erregresio-ereduen erabilgarritasuna aztertzeko eta ko-aldagaien adierazgarritasuna aztertzeko.
(d) Konparatu ereduak haien artean, eta aukeratu egokiena.
(e) Erabili eredua iragarpenak egiteko. Zein da 1.500 oin karratu dituen etxe baten itxarondako energia-kontsumoa, erregresio-eredu ezberdinen arabera?
(f) Egin iragarpenen inferentzia. Kalkulatu 1.500 oin karratu dituen etxe baten itxarondako energia-kontsumoaren banakako konfiantza-tartea, eredu ezberdinak erabiliz.

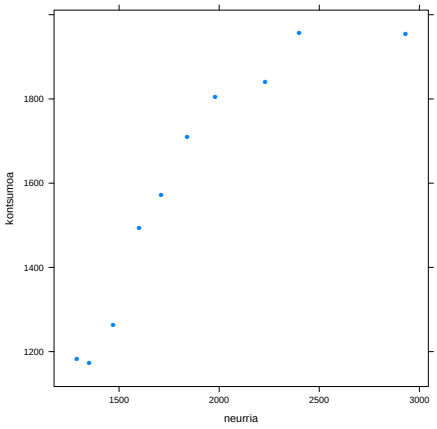
Galdera hauek guztiak hurrengo ataletan erantzungo ditugu. Zehazki, (a) galderari 7.2. atalean erantzungo diogu; (b) eta (c) atalei 7.3. atalean erantzungo diegu, ereduak banan-banan aztertuz, bai modu deskribatzailean eta bai metodo inferentzialak erabiliz; (d) atalari 7.4. atalean emango zaio erantzuna; azkenik, (e) eta (d) atalak 7.5. atalean aztertuko dira.

Hasteko, eraiki dezagun `Datos` izeneko lagina, eskuz ala datuak **praktikadatuak.xls** fitxategiko **energia** orritik inportatuz. Bi aldagai izango ditugu: bat *neurria* eta bestea *kontsumoa*.

7.2. Populazio-eredua proposatzea: $\mathbf{Y = f(X_1, X_2, \ldots, X_k) + \epsilon}$

Azter dezagun grafikoki lagina. Horretarako, aldagai azaltzaile bakoitza menpeko aldagaiarekiko irudikatuko dugu, puntu-hodeia erabiliz. Kasu honetan, $k = 1$ denez (aldagai azaltzaile bakarra dugu), hodei-puntu bakarra egin dezakegu. Bestela, k grafiko egin beharko genituzke: X_j *vs* Y grafiko bat aldagai aske bakoitzerako.

`Grafikoak` → `XY grafikoa` → `Aldagai azaltzailea:` *neurria*. `Menpeko aldagaiak:` *kontsumoa*



Itxura honekin, erregresio-eredu ezberdinak egokiak izan daitezkeela ikusten dugu. Beraz, hauetako batzuk aztertu eta ebaluatuko ditugu hurrengo ataletan.

7.3. Erregresio-ereduen analisi kuantitatiboa

Esan bezala, hodei-puntuaren grafikoari begiratuz, zenbait erregresio-eredu egokiak izan daitezkeela dirudi. Guk honako oinarrizko erregresio-eredu hauek aztertuko ditugu:

Erregresio eredua	**Adierazpen matematikoa**	**Adierazpen linealizatua**	**Mendeko aldagaia**	**Aldagai azaltzailea**
Lineala	$\hat{Y} = b_0 + b_1 X$		Y	X
Hiperbolikoa	$\hat{Y} = b_0 + \frac{b_1}{X}$		Y	$1/X$
Potentziala	$\hat{Y} = b_0 X_1^b$	$\log \hat{Y} = \log b_0 + b_1 \log X$	$logY$	$logX$
Esponentziala	$\hat{Y} = b_0 e^{b_1 X}$	$\log \hat{Y} = \log b_0 + b_1 X$	$logY$	X
Koadratikoa	$\hat{Y} = b_0 + b_1 X + b_2 X^2$		Y	X, X^2

7.3.1. Erregresio-eredu lineal bakuna

Erregresio lineala

$$\hat{Y} = b_0 + b_1 X \quad \Leftrightarrow \quad \hat{Kontsumoa} = b_0 + b_1 Neurria$$

Rcmdr erabiliz aztertzeko, jarraitu honako pauso hauei:

`Estatistikoak` → `Doitu ereduak` → `Erregresio lineala` → `Idatzi ereduaren izena:` *RegModel.*1. `Mendeko aldagaia:` *kontsumoa*. `Aldagai azaltzailea:` *neurria*

```
Call:
lm(formula = kontsumoa ~ neurria, data = Datos)

Residuals:
    Min      1Q  Median      3Q     Max
-208.02 -105.36   52.89   77.29  155.27

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 578.92775  166.96806   3.467 0.008476 **
neurria       0.54030    0.08593   6.288 0.000236 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133.4 on 8 degrees of freedom
Multiple R-squared:  0.8317,        Adjusted R-squared:  0.8107
F-statistic: 39.54 on 1 and 8 DF,  p-value: 0.0002359
```

Analisi deskribatzailea

Hasteko, $\hat{Y} = 578,92775 + 0,54030X$ erregresio-zuzenaren adierazpena da. Ohartu, gainera, eraikitako azken eredua eredu aktibo gisa gordeta geratzen dela *Rcmdr*n, haren gainean beste eragiketaren bat egin nahiko bagenu erabili ahal izateko.

Ereduaren doikuntza-egokitasuna era deskribatzailean aztertzeko, honako informazio hau dugu: hondarren errore estandarra $S = 133,4$ da, determinazio-koefizientea $R^2 = 0,8317$. Azken balio hori, 1etik hurbil dagoenez, erregresio-ereduaren menpekotasun lineala handia dela esan dezakegu, eta kontsumoaren aldakortasunaren %83,17 neurriagatik azalduta dagoela.

Analisi inferentziala

• Ereduaren doikuntza-egokitasuna: koaldagai guztiak nuluak direla esaten duen hipotesi nulua ($H_0 : B_1 = 0$, kasu honetan) duen hipotesi-kontrastearen estatistikoa $F = 39,54$ da, eta p-balioa $p = 0,0002359 < 0,001$ denez, erregresio-eredua guztiz erabilgarria da.

• Koefizienteen adierazgarritasuna: koefiziente konstantea B_0 nahiko adierazgarria da; $\hat{B}_0 = b_0 = 578,92775$ ez-nulutzat har daiteke, bere p-balioa $p = 0,008476 < 0,01$ baita. $X = neurria$ aldagaia guztiz adierazgarria da $Y = kontsumoa$ azaltzeko; $\hat{B}_1 = b_1 = 0,54030$ koefizientea ez-nulutzat har dezakegu, bere p-balioa $p = 0,000236 < 0,001$ baita.

• Koefizienteen konfiantza-tarteak: B_0 eta B_1 parametroen konfiantza-tarteak lortzeko:

`Ereduak` → `Konfiantza-tarteak` → `0.95`

```
               Estimate       2.5 %      97.5 %
(Intercept) 578.9277515 193.8987213 963.9567817
neurria       0.5403044   0.3421499   0.7384589
```

Horrela, $I^{0,95}_{B_0} = (194,\ 964)$ eta $I^{0,95}_{B_1} = (0,3421,\ 0,7385)$ direla ondorioztatzen dugu. Berriz ere, argi dago bi parametroak ez-nuluak direla onar dezakegula. Gainera, B_1-ren tartea positiboa denez, menpekotasun gorakorra dugu; hau da, zenbat eta handiagoa izan etxebizitzaren neurria, orduan eta handiagoa izango da kontsumoaren energia, espero moduan. Bereziki, etxebizitzaren oin karratu bakoitzagatik energiaren igoera 0,3421 eta 0,7385 kwh-ko tartean egotea espero dugu, % 95eko konfiantzaz.

7.3.2. Erregresio-eredu hiperbolikoa (alderantzizkoa)

Erregresio hiperbolikoa

$$\hat{Y} = b_0 + \frac{b_1}{X} \quad \Leftrightarrow \quad \hat{Kontsumoa} = b_0 + b_1 \cdot \frac{1}{Neurria}$$

aztertzeko, lehenengoz, X aldagaiaren alderantzizkoa, $\frac{1}{X}$, aldagai berria eraiki behar dugu:

`Datuak` $\rightarrow$ `Kudeatu datu multzo aktiboko aldagaiak` $\rightarrow$ `Kalkulatu aldagai berri bat` $\rightarrow$ `Aldagai berriaren izena:` *alderantzizkoa*. `Konputatzeko adierazpena:` $1/neurria$

Eta erregresio-eredu hiperbolikoa garatzeko:

`Estatistikoak` $\rightarrow$ `Doitu ereduak` $\rightarrow$ `Erregresio lineala` $\rightarrow$ `Idatzi ereduaren izena:` $RegModel.2$. `Mendeko aldagaia:` *kontsumoa*. `Aldagai azaltzailea:` *alderantzizkoa*

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.787e+03  9.537e+01   29.22 2.04e-09 ***
alderantzizkoa -2.106e+06  1.639e+05  -12.85 1.27e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 69.92 on 8 degrees of freedom
Multiple R-squared:  0.9538,        Adjusted R-squared:  0.948
F-statistic: 165.2 on 1 and 8 DF,  p-value: 1.27e-06
```

Analisi deskribatzailea

Hots, erregresio-eredu hiperbolikoaren adierazpena $\hat{Y} = 2787 - 2,106 \cdot 10^6/X$ da.

Gainera, hondarren errore estandarra $S = 69,92$ da, eta determinazio-koefizientea $R^2 = 0,9538$. Beraz, R^2 1etik oso hurbil dagoenez, erregresio-ereduaren menpekotasun hiperbolikoa oso handia dela esan dezakegu, eta kontsumoaren aldakortasunaren $\%95,38$ neurriaren alderantzizkoagatik azalduta dagoela.

Analisi inferentziala

• Doikuntza-egokitasuna: $H_0 : B_1 = 0$ (eredu ez-erabilgarria) hipotesi nulutzat duen hipotesi-kontrastearen estatistikoa $F = 165,2$ da, eta p-balioa $p < 0,001$ denez, erregresio-eredu hiperbolikoa guztiz erabilgarria da.

• Aldagaien adierazgarritasuna: koefiziente konstantea B_0 guztiz adierazgarria da, bere p-balioa $p < 0,001$ baita. $X = 1/Neurria$ alderantzizko aldagaia ere guztiz adierazgarria da $Y = kontsumoa$ azaltzeko, bere p-balioa $p < 0,001$ baita.

• Koefizienteen konfiantza-tarteak: B_0 eta B_1 parametroen konfiantza-tarteak lortzeko:

`Ereduak` → `Konfiantza-tarteak` → `0.95`

```
                  Estimate       2.5 %       97.5 %
(Intercept)        2786.87     2566.957     3006.783
alderantzizkoa -2105963.35 -2483854.856 -1728071.848
```

Horrela, $I_{B_0}^{0,95} = (2566,957,\ 3006,783)$ eta $I_{B_1}^{0,95} = (-2483854,856,\ -1728071,848)$ direla ondorioztatzen dugu. Berriz ere, argi dago bi parametroak ez-nuluak direla onar dezakegula. Gainera, ikus dezakegu neurriaren alderantzizkoa igotzen den heinean, kontsumoa jaitsi egiten dela.

7.3.3. Erregresio-eredu potentziala

Erregresio potentziala

$$\hat{Y} = b_0 + X^{b_1} \quad \Leftrightarrow \quad log(\hat{Y}) = log(b_0) + b_1 log(X) \quad \Leftrightarrow \quad log(\hat{kontsumoa}) = log(b_0) + b_1 log(neurria)$$

aztertzeko, X eta Y aldagaien logaritmo nepertarrak, $\log(X)$ eta $\log(Y)$, aldagai berriak sortu behar ditugu:

`Datuak` → `Aldatu datu multzo aktiboko aldagaiak` → `Kalkulatu aldagai berri bat` → `Aldagai berriaren izena:` $logX$`. Konputatzeko adierazpena:` $log(neurria)$

eta

`Datuak` → `Aldatu datu multzo aktiboko aldagaiak` → `Kalkulatu aldagai berri bat` → `Aldagai berriaren izena:` $logY$`. Konputatzeko adierazpena:` $log(kontsumoa)$

Eta erregresio-eredu potentziala garatzeko:

`Estatistikoak` → `Doitu ereduak` → `Erregresio lineala` → `Idatzi ereduaren izena:` $RegModel.3$`. Mendeko aldagaia:` $logY$`. Aldagai azaltzailea:` $logX$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.98454    0.69826   2.842   0.0217 *
logX         0.71564    0.09296   7.698 5.75e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07353 on 8 degrees of freedom
Multiple R-squared:  0.8811,        Adjusted R-squared:  0.8662
F-statistic: 59.26 on 1 and 8 DF,  p-value: 5.754e-05
```

Analisi deskribatzailea

Hots, $log(\hat{Y}) = 1,98454 - 0,71564 \cdot logX$ erregresio-eredu potentzial linealizatuaren adierazpena denez, $\hat{Y} = e^{1,98454} \cdot X^{-0,71564} = 7,2757 \cdot X^{-0,71564}$.

Gainera, hondarren errore estandarra $S = 0,07353$ eta determinazio-koefizientea $R^2 = 0,8811$ da. R^2 1etik hurbil dagoenez, erregresio-ereduaren menpekotasun potentziala handia da; kontsumoaren logaritmoaren %88,11 neurriaren logaritmoaren bidez azaltzen da.

Analisi inferentziala

• Doikuntza-egokitasuna: $H_0 : B_1 = 0$ (eredu ez-erabilgarria) hipotesi nulua duen hipotesi-kontrastearen estatistikoa $F = 59,26$ da, eta p-balioa $p < 0,001$ denez, erregresio-eredu potentziala guztiz erabilgarria da.

• Koefizienteen adierazgarritasuna: koefiziente konstantea B_0 adierazgarria da, bere p-balioa $p < 0,05$ baita. Era berean, $logX$ aldagaia guztiz adierazgarria da $logY$ azaltzeko, bere p-balioa $p < 0,001$ baita.

• Koefizienteen konfiantza-tarteak: B_0 eta B_1 parametroen konfiantza-tarteak lortzeko:

`Ereduak` → `Konfiantza-tarteak` → `0.95`

```
             Estimate    2.5 %    97.5 %
(Intercept) 1.9845414 0.3743509 3.5947320
logX        0.7156374 0.5012628 0.9300121
```

Horrela, $I^{0,95}_{log(B_0)} = (0,3744,\ 3,5947)$ eta $I^{0,95}_{B_1} = (0,5013,\ 0,9300)$ direla ondorioztatzen dugu. Berriz ere, argi dago bi parametroak ez-nuluak direla onar dezakegula. Dagokion transformazioa eginez, $I^{0,95}_{B_0} = (e^{0,3744},\ e^{3,5947})$.

7.3.4. Erregresio-eredu esponentziala

Erregresio esponentziala

$$\hat{Y} = b_0 \cdot e^{b_1 X} \quad \Leftrightarrow \quad \hat{log(Y)} = log(b_0) + b_1 \cdot X \quad \Leftrightarrow \quad log(\hat{kontsumoa}) = log(b_0) + b_1 \cdot neurria$$

aztertzeko, Y aldagaiaren logaritmo nepertarra, $\log(Y)$, aldagaia berriro behar da erregresio-eredu esponentziala garatzeko:

`Estatistikoak` → `Doitu ereduak` → `Erregresio lineala` → `Idatzi ereduaren izena:` $RegModel.4$. `Mendeko aldagaia:` $logY$. `Aldagai azaltzailea:` $neurria$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.706e+00  1.205e-01  55.625 1.21e-11 ***
neurria     3.464e-04  6.204e-05   5.584  0.00052 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09634 on 8 degrees of freedom
Multiple R-squared:  0.7958,        Adjusted R-squared:  0.7703
F-statistic: 31.18 on 1 and 8 DF,  p-value: 0.00052
```

Analisi deskribatzailea

Hots, $\hat{logY} = 6,706 + 0,0003464X$ erregresio-eredu esponentzial linealizatuaren adierazpena denez, $\hat{Y} = e^{6,706} \cdot e^{0,0003464X} = 817 \cdot e^{0,0003464X}$.

Gainera, hondarren errore estandarra $S = 0,09634$ da, determinazio-koefizientea $R^2 = 0,7958$ izanik. R^2 1etik hurbil dagoenez, erregresio-ereduaren menpekotasun esponenziala handia da; kontsumoaren logaritmoaren %79,58 neurriaren bidez azaltzen da.

Analisi inferentziala

- Doikuntza-egokitasuna: $H_0 : B_1 = 0$ (eredu ez-erabilgarria) hipotesi nulutzat duen hipotesi-

kontrastearen estatistikoa $F = 31,18$ da, eta p-balioa $p < 0,001$ denez, erregresio-eredu esponentziala guztiz erabilgarria da.

• Koefizienteen adierazgarritasuna: koefiziente konstantea B_0 guztiz adierazgarria da, bere p-balioa $p < 0,001$ baita. X aldagaia guztiz adierazgarria da $logY$ azaltzeko, bere p-balioa $p < 0,001$ baita.

• Koefizienteen konfiantza-tarteak: B_0 eta B_1 parametroen konfiantza-tarteak lortzeko:

`Ereduak` → `Konfiantza-tarteak` → `0.95`

```
              Estimate        2.5 %       97.5 %
(Intercept) 6.7055321963 6.4275468779 6.9835175147
neurria     0.0003464129 0.0002033482 0.0004894775
```

Horrela, $I^{0,95}_{log(B_0)} = (6,4275,\ 6,9835)$ eta $I^{0,95}_{B_1} = (0,0002,\ 0,0005)$ direla ondorioztatzen dugu. Berriz ere, argi dago bi parametroak ez-nuluak direla onar dezakegula. Dagokion transformazioa eginez, $I^{0,95}_{B_0} = (e^{6,4275},\ e^{6,9835})$.

7.3.5. Erregresio-eredu koadratikoa

Erregresio koadratikoa

$$\hat{Y} = b_0 + b_1 \cdot X + b_2 \cdot X^2 \quad \Leftrightarrow \quad \hat{kontsumoa} = b_0 + b_1 \cdot neurria + b_2 \cdot neurria^2$$

aztertzeko, X aldagaiaren karratua, X^2, aldagaia sortu behar da:

`Datuak` → `Aldatu datu multzo aktiboko aldagaiak` → `Kalkulatu aldagai berri bat` → `Aldagai berriaren izena:` $neurria2$. `Konputatzeko adierazpena:` $neurria * neurria$

eta erregresio-eredu koadratikoa garatzeko:

`Estatistikoak` → `Doitu ereduak` → `Erregresio lineala` → `Idatzi ereduaren izena:` $RegModel.5$. `Mendeko aldagaia:` $kontsumoa$. `Aldagai azaltzailea:` $neurria, neurria2$

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.216e+03  2.428e+02  -5.009 0.001550 **
neurria      2.399e+00  2.458e-01   9.758 2.51e-05 ***
neurria2    -4.500e-04  5.908e-05  -7.618 0.000124 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.8 on 7 degrees of freedom
Multiple R-squared:  0.9819,        Adjusted R-squared:  0.9767
F-statistic: 189.7 on 2 and 7 DF,  p-value: 8.001e-07
```

Analisi deskribatzailea

Hots, $\hat{Y} = -1216 + 2,399X - 0,00045X^2$ erregresio-eredu koadratikoaren adierazpena da.

Gainera, hondarren errore estandarra $S = 46,8$ da, eta determinazio-koefizientea $R^2 = 0,9819$. Azken balio hori, 1etik oso hurbil dagoenez, erregresio-ereduaren menpekotasun koadratikoa oso handia da; kontsumoa %98,19 neurriaren eta neurriaren karratuaren bidez azaltzen da.

Analisi inferentziala

• $H_0 : B_1 = B_2 = 0$ (eredu ez-erabilgarria) hipotesi nulutzat duen hipotesi-kontrastearen estatistikoa $F = 189,7$ da, eta p-balioa $p < 0,001$ denez, erregresio-eredu koadratikoa guztiz erabilgarria da.

• Koefizienteen adierazgarritasuna: koefiziente konstantea B_0 nahikoa adierazgarria da, bere p-balioa $p < 0,01$ baita; eta X aldagaia guztiz adierazgarria da Y azaltzeko, bere p-balioa $p < 0,001$ baita. Azkenik, X^2 aldagaia ere guztiz adierazgarria da Y azaltzeko, bere p-balioa $p < 0,001$ baita.

• Koefizienteen konfiantza-tarteak: B_0, B_1 eta B_2 parametroen konfiantza-tarteak lortzeko:

`Ereduak` → `Konfiantza-tarteak` → `0.95`

```
                Estimate         2.5 %         97.5 %
(Intercept) -1.216144e+03 -1.790290e+03 -6.419981e+02
neurria      2.398930e+00  1.817621e+00  2.980239e+00
neurria2    -4.500402e-04 -5.897342e-04 -3.103462e-04
```

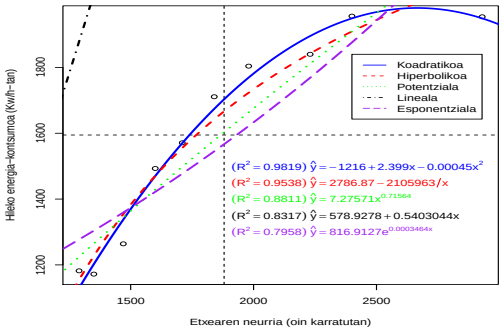
Horrela, $I_{B_0}^{0,95} = (-1,7903 \cdot 10^3,\ -6,42 \cdot 10^2)$, $I_{B_1}^{0,95} = (1,8176,\ 2,9802)$ eta $I_{B_2}^{0,95} = (-5,8973 \cdot 10^4,\ -3,1035 \cdot 10^4)$ direla ondorioztatzen dugu. Berriz ere, argi dago hiru parametroak ez-nuluak direla onar dezakegula.

7.4. Erregresio-ereduen konparaketa

Erregresio-eredu erabilgarriak R^2 parametroaren arabera ordenatzen baditugu, eredu egokienetik eredu desegokienera honako sailkapen hau izango genuke:

Eredu koadratikoa	$\hat{Y} = -1216 + 2,399X - 0,0004500X^2$	$(R^2 = 0,9819)$	egokiena
Eredu hiperbolikoa	$\hat{Y} = 2787 - 2,106 \cdot 10^6/X$	$(R^2 = 0,9538)$	
Eredu potentziala	$\hat{Y} = 7,2757 \cdot X^{-0,71564}$	$(R^2 = 0,8611)$	
Eredu lineala	$\hat{Y} = 578,92775 + 0,54030X$	$(R^2 = 0,8317)$	
Eredu esponentziala	$\hat{Y} = 817 \cdot e^{0,0003464X}$	$(R^2 = 0,7958)$	desegokiena

Bukatzeko, eredu guztiak grafikoki datuekin batera irudikatu nahi baditugu, ikusi 7. *R* praktika.



7.5. Iragarpenak

Behin erregresio-eredua erabilgarria dela ziurtatu dugunean, zenbatespenak eta iragarpenak egiteko erabil dezakegu. *Rcmdr*k ez digu hori zuzenean egiteko aukera ematen, beraz, *R*ko kodea erabili beharko dugu.

Hurrengo ataletan, taula grisetan eta > sinboloaz hasitako lerroak **R script-a** izeneko leihoan idatzi behar ditugu. Aginduak exekutatzeko, hautatu idatzitako lerroak, eta sakatu ondoren **Exekutatu** botoia. Emaitza, betiko moduan, **Emaitzak** leihoan agertuko zaigu, urdinez.

Analisi deskribatzailea

Kalkula dezagun, 16. adibideko (e) atalean galdetu bezala, 1.500 oin karratu dituen etxe baterako eredu ezberdinak iragarritako energia-kontsumoa.

Erregresio-eredu lineala erabiliz, iragarpenak lortzeko honako kode hau erabili beharko genuke:

```
> predict(RegModel.1, data.frame(neurria=1500))
```

```
       1
1389.384
```

Beraz, eredu linealak esaten du 1500 oin karratu dituen etxe baten kontsumoa 1389, 384 dela esaten du. Kontuan izan datuak formatu berezi batean sartu ditugula: `data.frame` moduan.

Era berean joka dezakegu beste erregresio-eredu guztiekin, kasu bakoitzean dagokion aldagaia sartuz.

Analisi inferentziala

Orain, har 16. adibideko (f) atala eta kalkula ditzagun adibidez 1.500 oin karratu dituen etxe baterako eredu ezberdinak iragarritako energia-kontsumoren banakako konfiantza-tarteak.

Erregresio-eredu lineala erabiliz iragarpenaren banakako konfiantza-tarteak lortzeko, kode hau erabili beharko genuke:

```
> predict(RegModel.1, data.frame(neurria=1500), interval="prediction")
```

Emaitza, betiko moduan, **Emaitza** izeneko leihoan agertuko zaigu, urdinez:

```
       fit     lwr      upr
1 1389.384 1057.99 1720.779
```

Hots, erregresio-eredu linealaren arabera, iragarpen-zenbatespena $\hat{y}(1500) = 1389.384$ kwh da, eta banakako iragarpen-tartea $I^{1-\alpha}_{y_0|x_0} = I^{0,95}_{y_0|1500} = (1057,99,\ 1720,779)$.

Erregresio-eredu hiperbolikoa erabiliz, era antzekoan jokatuko genuke:

```
> predict(RegModel.2, data.frame(alderantzizkoa=1/1500), interval="prediction")
```

```
       fit      lwr      upr
1 1382.894 1209.581 1556.207
```

Hots, erregresio-eredu linealaren arabera, iragarpen-zenbatespena $\hat{y}(1500) = 1382,894$ kwh da eta banakako iragarpen-tartea $I^{1-\alpha}_{y_0|x_0} = I^{0,95}_{y_0|1500} = (1209,581,\ 1556,207)$.

Erregresio-eredu potentziala erabiliz:

```
> predict(RegModel.3, data.frame(logX=log(1500)), interval="prediction")
```

```
       fit      lwr      upr
1 7.218156 7.035541 7.40077
```

Kontuan izan kasu honetan, ereduaren itxura $log(\hat{kontsumoa}) = log(b_0) + b_1 log(neurria)$ zela; beraz, iragartzen ari garen balioa da, kontsumoaren logaritmo nepertarra. Hots, kontrako transformazioa eginez, erregresio-eredu linealaren arabera, iragarpen-zenbatespena $\hat{y}(1500) = e^{7,218156} = 1363,972$ kwh da, eta iragarpen-tartea $I^{1-\alpha}_{y_0|x_0} = I^{0,95}_{y_0|1500} = (e^{7,035541},\ e^{7,40077}) = (1136,309,\ 1637,245)$.

Erregresio-eredu esponentziala erabiliz:

```
> predict(RegModel.4, data.frame(neurria=1500), interval="prediction")
```

```
       fit      lwr      upr
1 7.225151 6.985889 7.464414
```

Hots, erregresio-eredu linealaren arabera, iragarpen-zenbatespena $\hat{y}(1500) = e^{7,225151} = 1373,546$ kwh da, eta iragarpen-tartea $I^{1-\alpha}_{y_0|x_0} = I^{0,95}_{y_0|1500} = (e^{6,985889},\ e^{7,464414}) = (1081,267,\ 1744,833)$.

Azkenik, erregresio-eredu koadratikoa erabiliz:

```
> predict(RegModel.5,data.frame(neurria=1500,neurria2=1500*1500),
+         interval="prediction")
```

```
       fit      lwr      upr
1 1369.661 1250.317 1489.005
```

Hots, erregresio-eredu linealaren arabera, iragarpen-zenbatespena $\hat{y}(1500) = 1369,661$ kwh da, eta iragarpen-tartea $I^{1-\alpha}_{y_0|x_0} = I^{0,95}_{y_0|1500} = (1250,317,\ 1489,005)$.

II. atala

Praktikak: *R* lengoaian programatuz

0. *R* praktika

Sarrera

Helburua

R programazio-lengoaia bat da, eta, beraz, harekin lan egiteko eta analisi estatistikoak egiteko, komando edo agindu ezberdinak erabiltzea da ohikoena. Atal honetan, *R*-ren komando bidezko erabileraren sarrera bat emango dugu.

Kontuan izan hurrengo kapituluetan erabiliko diren datu guztiak
`https://ehubox.ehu.eus/index.php/s/6C3gBHFiTME7OTx` orrialdean daudela eskuragai.

0.1. *R* eta RStudio-ren instalazioa

R open source code edo kode ireki motako software estatistikoa instalatzeko, jo honako helbide honetara:

`http://cran.es.r-project.org`

Helbide honetan, sistema eragile ezberdinetarako instalazio-fitxategiak dituzue eskuragai. Behin hori instalatuta, dagoeneko *R*n programatzen has gaitezke kontsola edo *R*-ren leiho nagusia erabiliz.

Alabaina, azken urteetan RStudio izeneko kode editoreaz baliatzen hasi dira *R*-ren erabiltzaile asko eta asko. Editore honek, aukera ematen du kodea, emaitzak, irudiak eta beste hainbat elementu batera bistaratzeko; *R*ko kodea hobeto antolatu eta zuzenean paketeak sortzeko erramintak eskuratzen ditu, eta, orohar, kodea sortzeko garaian erraztasun ugari ematen ditu. Editore honen doako bertsioa instalatzeko, honako helbide honetara jo dezakegu:

`https://www.rstudio.com/`

0.2. Oinarrizko informazioa

Esan bezala, *R* programazio-lengoaia bat da; beraz, hemendik aurrera, *R*ko komandoak honako formatu honen bidez adieraziko ditugu, beste azalpenetatik bereiziz:

```
> komandoak
```

Gainera, kodeak edo egin nahi dugun eragiketak emaitzaren bat ematen badu, emaitza segidan agertuko da, honela:

```
> 2+2

[1] 4
```

Hala, dokumentu honetan agertuko diren komando guztiak *.R* hedapeneko script batean gorde daitezke nahi izanez gero, norberak bere kasa komandoak exekutatu eta proba ezberdinak egiteko.

Programazioarekin hasteko, kontuan izan *R* kalkulagailu moduan ere erabil dezakegula leiho nagusian (*Console* izenekoan) nahi ditugun eragiketak idatziz:

```
> 2^3+4/3

[1] 9.333333

> log(2)

[1] 0.6931472
```

Noski, erabilera oinarrizko horretaz gain, gauza askoz ere interesgarri eta konplexuagoak egin ditzakegu *R*n. Hasteko, aldagaiei esleipenak egiteko «=» edo «< −» operadoreak erabiliko ditugu:

```
> aldagaia1 <- 2
> aldagaia2 = 2
> aldagaia1

[1] 2

> aldagaia2

[1] 2
```

sinboloaz hasitako lerroak komentarioak dira, eta ez dute inolako ekintzarik gauzatzen:

```
> #Hemen gure kodeari buruzko azalpenak jar ditzakegu
> aldagaia3 <- 3^3+1
```

Bestalde, *R*ko oinarrizko komando hauek ezagutzea oso lagungarria egingo zaigu:

```
> #Momentuan erabilgarri dauden paketeak erakusteko:
> library()
```

```
> #Momentuan erabilgarri dauden datuak erakusteko:
> data()
```

```
> #Zein direktoriotan kokatuta gauden jakiteko:
> getwd()
```

```
> #Direktorioz aldatzeko:
> setwd("direktorioaren helbidea")
```

```
> #Definituta dauden objektuen zerrenda ikusteko:
> ls()
```

```
> #Definituta dauden objektuen zerrenda ezabatzeko:
> rm()
```

```
> #Funtzio edo objektu bati buruzko informazioa lortzeko:
> help("izena")
```

```
> #Sortutako objektu guztiak ezabatzeko:
> rm(list=ls())
```

```
> #Rko pakete bat instalatzeko
> install.packages("Paketearen izena")
```

```
> #Rko pakete bat eguneratzeko
> update.packages("Paketearen izena")
```

0.3. *R*ko oinarrizko objektuak

R lengoaia objektuetan oinarritutako lengoaia bat da. Era honetan, ezinbestekoa da *R*n berez definitutako objektu batzuen definizioa eta erabilera apur bat ezagutzea.

Atal honetan, *R*n erabiltzen diren oinarrizko objektuei buruzko sarrera bat emango dugu.

0.3.1. Bektoreak

*R*ko objektu mota garrantzitsuena bektorea da, eta era ezberdinetan defini daiteke oinarrizko bektore bat:

```
> #Zenbaki jakin batzuekin osatutako bektorea sortzeko:
> x <- c(1, 2, 3, 4, 5)
> x

[1] 1 2 3 4 5

> #1etik 6rako zenbaki osoen sekuentzia bat sortzeko:
> x2 <- c(1:6)
> x2

[1] 1 2 3 4 5 6

> #Zenbaki-sekuentzia bat sortzeko beste modu bat:
> x3 <- seq(0, 1, 0.2)
> x3

[1] 0.0 0.2 0.4 0.6 0.8 1.0
```

Alabaina, bektoreek mota askotako elementuak gorde ditzakete, ez bakarrik zenbakiak:

```
> y <- c(TRUE, FALSE, TRUE, FALSE)
> z <- c("bai", "ez", "bai", "bai")
```

Bektore batek zein motatako elementuak gordetzen dituen jakiteko, `class()` funtzioa erabil dezakegu:

```
> class(x)

[1] "numeric"

> class(y)
```

```
[1] "logical"

> class(z)

[1] "character"
```

Ikusten dugu x elementu numerikoz osaturiko bektore bat dela, y elementu boolearrez (logical) eta z karakterez. Ikusten dugunez, bektore bakoitzean soilik mota bateko elementuak gorde daitezke.

Aipatuko dugun azken elementu mota faktoreak dira. Hauek balio kualitatiboak adierazteko balio digute, maila edo balio posible batzuk zehaztuz:

```
> k <- c("ilehoria", "ilegorria", "ilehoria", "beltzarana", "beltzarana")
> f <- factor(k, levels=c("ilehoria", "ilegorria", "beltzarana"))
> f

[1] ilehoria   ilegorria  ilehoria   beltzarana beltzarana
Levels: ilehoria ilegorria beltzarana
```

Ikusten dugunez, 5 elementuko bektore bat sortu dugu, `factor` motakoa. Elementu hauetako bakoitzak, nahitaez, «ilehoria», «ilegorria», «beltzarana» balioetako bat hartu behar du.

Edozein motatako bektoreen luzera jakiteko, `length()` funtzioa erabil dezakegu:

```
> length(x)

[1] 5
```

Gainera, bektore bateko elementu zehatz bat atzitzeko, [ ] kortxeteak erabiliko ditugu, indizeak 1 zenbakitik hasten direlarik:

```
> x[3]

[1] 3

> y[2]

[1] FALSE

> z[1]

[1] "bai"
```

Azkenik, gogoratu elementu bakarreko objektuak ere bektore bezala ulertzen dituela Rk.

0.3.2. Matrizeak

Matrizeak bi dimentsiotako bektoreak dira, eta `matrix()` funtzioaren bidez sortzen dira era honetan:

```
> m <- matrix(c(1, 4, 5, 2, 1, 2), nrow=2, ncol=3)
> m

     [,1] [,2] [,3]
[1,]    1    5    1
[2,]    4    2    2
```

Matrize baten errendaka eta zutabe kopurua jakiteko, `dim()` funtzioa erabiliko dugu:

```
> dim(m)

[1] 2 3
```

Azkenik, matrize baten elementu zehatz bat atzitzeko, [ ] kortxeteak erabiliko ditugu bektoreekin bezala, baina bi dimentsioak zehaztuz:

```
> m[1, 2]

[1] 5
```

Kontuan izan matrize batean elementu guztiek mota berekoak izan behar dutela.

0.3.3. *Data frame*ak

*Data frame*ak datuak gordetzeko objektu mota ohikoenak dira. Zutabe bakoitzean aldagai baten balioak gordetzen dira, eta errenkada bakoitzak lagineko elementu bati egiten dio erreferentzia. Kontuan izan matrizeekin ez bezala, *data frame* objektuen zutabeek mota ezberdinetako datuak gorde ditzaketela.

Datu baseen irakurketa

*R*ko berezko datu fitxategien extensioak *.Rda* edo *.RData* dira. Mota honetako datu-fitxategi bat kargatzeko:

```
> load("Altuerak.RData")
> head(altuerak, 4)
```

```
  aitarena amarena       sexua jaioterria altuera pisua
1      174     156 emakumezkoa          2     165    65
2      177     159  gizonezkoa          2     170    67
3      173     161  gizonezkoa          1     168    51
4      174     156  gizonezkoa          2     167    69
```

Kontuan izan `head(datuak, n)` funtzioak datu-basearen lehenengo n lerroak erakusten dizkigula. Gainera, `load` funtzioa erabiltzen dugunean, zuzenean Rn objektu berri bat kargatuko da, datuak gordeko dituena. Aurreko kasuan, adibidez, `altuerak` izeneko objektua datuak gordetzen dituen `data.frame` bat da.

Bestalde, Rn beste formatuetako datu-fitxategiak ere karga ditzakegu. Adibidez, testu edo .csv formatuko fitxategiak kargatzeko [1]:

```
> tomate <- read.table("tomate.txt", header=TRUE)
> head(tomate, 3)

  freskoa latakoa
1   0.066   0.085
2   0.079   0.088
3   0.069   0.091
```

Testu-fitxategi bateko datuak kargatzen ditugunean, objekturen batean gorde behar ditugu, ondoren erabili ahal izateko. Adibidez, aurreko kasuan, `tomate` izeneko objektuan gorde ditugu.

Gainera, Excel fitxategiak irekitzeko **readxl** paketea instalatu behar dugu aurrenik. Ondoren, adibidez, **praktikadatuak.xls** fitxategiko **tomate** izeneko orria kargatuko dugu, eta `tomate2` izeneko objektuan gorde:

```
> library(readxl)
> tomate2 <- read_excel("praktikadatuak.xls", sheet='tomate')
> head(tomate2, 3)

  freskoa latakoa
1   0.066   0.085
2   0.079   0.088
3   0.069   0.091
```

Azkenik, SPSSk sortutako .sav bukaerako fitxategiak irekitzeko **foreign** paketea instalatu behar dugu. Kargatu ostean, era honetan inporta ditzakegu datuak:

[1] Kontuan izan fitxategia uneko direktorioan kokatuta ez badago, helbide osoa jarri beharko dugula.

```
> library(foreign)
> altuerak <- read.spss("Altuerak.sav", to.data.frame=TRUE)
> head(altuerak, 3)

  AITARENA AMARENA       SEXUA JAIOTERRIA ALTUERA PISUA
1      174     156 emakumezkoa          2     165    65
2      177     159  gizonezkoa          2     170    67
3      173     161  gizonezkoa          1     168    51
```

Komando hauek guztiek aukera eta parametro anitz dauzkate, datu-fitxategi mota ezberdinen ezaugarriei aurre egiteko. Horiei buruzko argibide gehiago lortzeko, *R*-ren laguntza erabil dezakegu.

Datu-base bateko elementuak eta aldagaiak eskuratzea

Data frame baten elementu zehatz bat atzitzeko, [ ] kortxeteak erabiliko ditugu matrizeekin bezala:

```
> #2. errenkada eta 3. zutabeko balioa atzitzeko
> altuerak[2, 3]

[1] gizonezkoa
Levels: gizonezkoa emakumezkoa
```

Era berean, errenkada oso bat edo zutabe oso bat eskuratzeko:

```
> #1. errendaka
> altuerak[1, ]
> #1. zutabea
> altuerak[, 1]
```

Honekin batera, erabilgarria izan daiteke `names()` funtzioa, datu-base baten aldagai guztien izenak zerrendatzen dituena:

```
> names(altuerak)

[1] "AITARENA"   "AMARENA"    "SEXUA"      "JAIOTERRIA" "ALTUERA"
[6] "PISUA"
```

Kasu honetan, aldagaien izenak maiuskulaz daude, eta minuskulak erabiltzea erosoagoa denez, aldaketa honela egin dezakegu:

```
> colnames(altuerak) <- tolower(colnames(altuerak))
```

Behin hori jakinda, aldagai bat bere izena erabiliz atzitzeko $ ikurra erabiliko dugu:

```
> altuerak$amarena
```

Azkenik, baldintza zehatz batzuk betetzen dituzten lagineko elementuak aukeratzeko (iragazkiak):

```
> #Emakumezkoen datuak soilik aukeratu
> em.datuak <- altuerak[altuerak$sexua=="emakumezkoa", ]
>
> #170 cm-tik gorako indibiduoen datuak aukeratzeko
> altuenak.datuak <- altuerak[altuerak$altuera > 170, ]
>
> #170 cm-tik gorako emakumeen datuak aukeratzeko
> emaltuenak.datuak <- altuerak[altuerak$altuera > 170 &
+                                 altuerak$sexua == "emakumezkoa", ]
>
> #170 cm-tik gorako EDO 160 cm-tik beherako
> #indibiduoen datuak aukeratzeko
> emaltuenak.datuak <- altuerak[altuerak$altuera > 170 |
+                                 altuerak$altuera < 160, ]
```

Aldagai motak definitu eta aldatzea

Aldagai baten mota ezagutzeko, *class*() funtzioa erabil dezakegu, lehen azaldu moduan:

```
> class(altuerak$altuera)

[1] "numeric"

> class(altuerak$jaioterria)

[1] "numeric"
```

Aurreko kodean ikus dezakegu adibidez, *jaioterria* aldagaiaren mota gaizki definituta dagoela. Izan ere, aldagai kualitatiboa izanda, faktore motakoa izan beharko litzakete, eta ez numerikoa. Aldaketa egiteko:

```
> altuerak$jaioterria <- factor(altuerak$jaioterria, levels=c(1, 2, 3),
+                              labels=c("Araba", "Bizkaia", "Gipuzkoa"))
> class(altuerak$jaioterria)

[1] "factor"

> head(altuerak$jaioterria)

[1] Bizkaia  Bizkaia  Araba    Bizkaia  Araba    Gipuzkoa
Levels: Araba Bizkaia Gipuzkoa
```

1 balioa «Araba» balioarekin aldatu dugu, 2 balioa «Bizkaia»rekin eta 3 balioa «Gipuzkoa»rekin. Aldaketa egin ostean, ikus dezakegu *jaioterria* aldagaia orain ondo kodifikatuta dagoela. Era berean, `as.numeric()` funtzioak aldagai bat numeriko modura aldatzeko balio digu.

Aldagai berriak sortzea

Demagun *imc (indice masa corporal)* izeneko aldagai berria sortu nahi dugula, non $imc = \frac{pisua(Kg)}{altuera^2(m^2)}$ baita. Horretarako, honako kode hau exekuta dezakegu:

```
> altuerak$imc <- altuerak$pisua/(altuerak$altuera/100)^2
```

Era berean, edozein aldagai sor dezakegu.

Aldagaiak birkodetzea

Jarraian, *imc* indizean oinarrituta, obesitate-maila aldagaia (*maila*) eraikiko dugu SEEDO izeneko Obesitatearen Ikerketarako Sozietatearen kodeketari jarraituz:

$$\text{obesitate-maila} = \begin{cases} \text{gutxiegizko pisua}, & \text{baldin } imc < 18,5 \text{ bada}, \\ \text{pisu normala}, & \text{baldin } 18,5 \leq imc < 25 \text{ bada}, \\ \text{gehiegizko pisua}, & \text{baldin } 25 \leq imc < 30 \text{ bada}, \\ \text{obesitatea}, & \text{baldin } imc \geq 30 \text{ bada} \end{cases}$$

Horretarako, `cut` funtzioa erabiliko dugu:

```
> altuerak$maila <- cut(altuerak$imc, c(0, 18.5, 25, 30, Inf))
```

Horrek nire aldagaia $(0, 18.5]$, $(18.5, 25]$, $(25, 30]$ eta $(30, \infty)$ tarteetan banatuko du. Tartearen beheko-muga itxia eta goiko muga irekia izatea nahi badugu, SEEDO kodeketan bezala:

```
> altuerak$maila <- cut(altuerak$imc, c(0, 18.5, 25, 30, Inf), right = FALSE)
```

Aldagaiak diskretizatzea

Demagun pisua aldagaia 4 tartetan diskretizatu nahi dugula. Hori egiteko metodo ezberdinak daude eskuragai, **RcmdrMisc** paketeaz baliatuz.

Gogoratu paketea kargatu aurretik instalatu egin behar dugula beti:

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

Ondoren, paketea kargatu eta erabiltzen hasi gaitezke:

```
> library("RcmdrMisc")
```

- Zabalera berdineko zatiak eraiki nahi baditugu:

```
> altuerak$pisumaila1 <- bin.var(altuerak$pisua, bins = 4,
+                                method = c("intervals"))
```

- Maiztasun bereko zatiak eraiki nahi baditugu:

```
> altuerak$pisumaila2 <- bin.var(altuerak$pisua, bins = 4,
+                                method = c("proportions"))
```

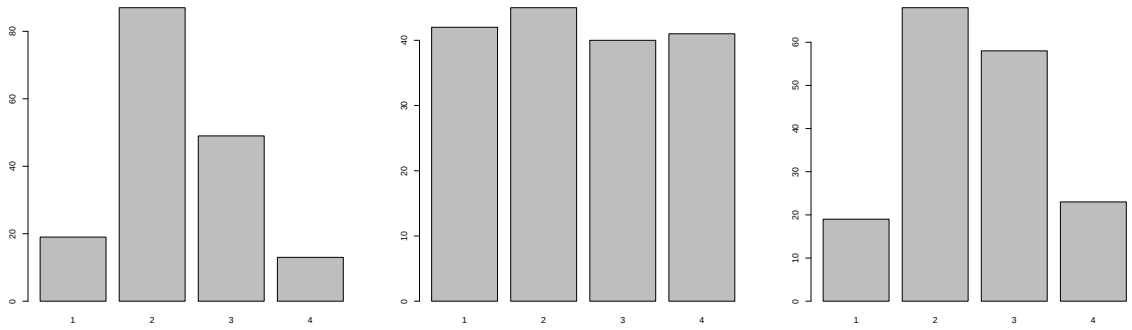
- Algoritmo automatiko bat erabiliz, segmentu naturalak eraiki nahi baditugu:

```
> altuerak$pisumaila3 <- bin.var(altuerak$pisua, bins = 4,
+                                method = c("natural"))
```

Hiru aldagai berri horien barra-diagrama egin dezakegu orain:

```
> #Hiru grafiko bata bestearen alboan erakusteko.
> par(mfrow=c(1,3))
>
> #Hiru aldagaien maiztasun-taulak lortu
> counts1 <- table(altuerak$pisumaila1)
> counts2 <- table(altuerak$pisumaila2)
> counts3 <- table(altuerak$pisumaila3)
```

```
>
> #Barra-diagramak egin
> barplot(counts1)
> barplot(counts2)
> barplot(counts3)
```



0.3.4. Funtzioak

Funtzioak *R*ko beste oinarrizko objektu batzuk dira, eta, bertan defektuz eraikita dauden funtzioez gain, erabiltzaileak beste funtzio batzuk eraiki ditzake. Funtzio baten oinarrizko estruktura honako hau da:

```
> funtzioa <- function(arg1, arg2, ...) {
+  aginduak
+  return(objektua)
+ }
```

Kontuan izan funtzioek edozein motatako objektuak itzul ditzaketela. Adibidez, bi zenbakiren batez-bestekoa kalkulatzen duen funtzioa honela eraikiko genuke:

```
> batezbestekoa <- function(x1, x2) {
+  bb <- (x1 + x2) / 2
+  return(bb)
+ }
```

Ondoren, funtzio hau erabil dezakegu honela:

```
> batezbestekoa(10, 15)

[1] 12.5
```

0.4. Operadore logikoak eta zikloak

0.4.1. `if`, `else if` eta `else` aginduak

Gure exekuzioa baldintza batzuen arabera baldintzatu nahi dugunean, `if`, `else if` eta `else` aginduak erabiliko ditugu. Agindu horiek argumentu bakarra dute: `TRUE` edo `FALSE` itzuliko duen adierazpen bat:

```
> if (baldintza1){
+   aginduak
+ } else if (baldintza2) {
+   aginduak
+ } else {
+   aginduak
+ }
```

Adibidez:

```
> # Baldintza ezberdinen arabera, mezu ezberdinak pantailaratu
> baldintzazko.funtzioa <- function(adina){
+ if (adina < 4){
+  print("Haurrak 4 urte baino gutxiago ditu.")
+ }else if (adina <= 7){
+  print("Haurrak 4 eta 7 urte arteko adina du.")
+ }else{
+  print("Haurrak 7 urte baino gehiago dauzka.")
+ }
+ }
>
> #Funtzioa exekutatu balio ezberdinetarako
> baldintzazko.funtzioa(3)

[1] "Haurrak 4 urte baino gutxiago ditu."

> baldintzazko.funtzioa(5)

[1] "Haurrak 4 eta 7 urte arteko adina du."

> baldintzazko.funtzioa(10)

[1] "Haurrak 7 urte baino gehiago dauzka."
```

Kontuan izan baldintzetako bat betetzen denean, hurrengo baldintzak ez direla egiaztatuko, eta `else` aginduaren bitartez aurreko bi baldintzak betetzen ez dituzten kasuez arduratu garela.

0.4.2. `for` zikloak

Iterazio kopurua ezaguna den begizta bat eraiki nahi badugu, `for` agindua erabiliko dugu honako modu honetan:

```
> for(aldagaia in zerrenda){
+   aginduak
+ }
```

Adibidez:

```
> for (adina in c(1:4)){
+ print(paste("Haurraren adina", adina, "da."))
+ }

[1] "Haurraren adina 1 da."
[1] "Haurraren adina 2 da."
[1] "Haurraren adina 3 da."
[1] "Haurraren adina 4 da."
```

Edo,

```
> for (adina in c(1, 3, 5, 7)){
+   print(paste("Haurraren adina", adina, "da."))
+ }

[1] "Haurraren adina 1 da."
[1] "Haurraren adina 3 da."
[1] "Haurraren adina 5 da."
[1] "Haurraren adina 7 da."
```

Kontuan izan `aldagaia in zerrenda` adierazpen hori erabiltzen dela aldagai edo kontagailu batek begiztan hartuko dituen balioen zerrenda definitzeko.

0.4.3. `while` aginduak

Iterazio kopuru ezezaguna duen begizta bat eraiki nahi dugunean, `while` agindua erabiliko dugu. Honela, baldintza bat betetzen den bitartean, agindu sorta bat exekutatuko da:

```
> while (baldintza){
+  aginduak
+ }
```

Adibidez:

```
> adina <- 0
> while(adina < 3) {
+   print(paste("Haurraren adina", adina, "da."))
+   adina <- adina + 1
+ }

[1] "Haurraren adina 0 da."
[1] "Haurraren adina 1 da."
[1] "Haurraren adina 2 da."
```

1. *R* praktika

Estatistika deskribatzailea

Helburua

Praktika honen xedea da jasotako datu esperimentalak aztertzea, laburtzea eta deskribatzea, bai metodo grafiko eta bai zenbakizkoen bidez. Honako hauek egiten ikasiko dugu: (1) Maiztasun-taulak, (2) Grafikoak eta (3) Estatistikoak. Gainera, aldagai bakar baten analisia egiteaz gain, bi aldagai aldi berean nola aztertu ere ikusiko dugu.

1.1. Adibidea

Praktika honetako helburuak lantzeko, honako adibide honetan oinarrituko gara:

17. adibidea. *Aurreko ataletan erabilitako* ***Altuerak.RData*** *fitxategian oinarrituz, aztertu bertako aldagaiak estatistika deskribatzailea erabiliz.*

Hasteko, datu-basea era honetan kargatuko eta atonduko dugu:

```
> load("Altuerak.RData")
> #Jaioterria aldagaia faktore bihurtu
> altuerak$jaioterria <- factor(altuerak$jaioterria, levels=c(1, 2, 3),
+                               labels=c("Araba", "Bizkaia", "Gipuzkoa"))
```

Datu-base honen bidez, gizabanako baten altuera eta beste zenbait aldagai aztertu nahiko ditugu estatistika deskribatzailea erabiliz. Aztertuko ditugun aldagaiak hauek dira: gizabanakoaren altuera, aitaren altuera, amaren altuera, jaioterria, sexua eta pisua.[1]

[1] Gogoratu `jaioterria` aldagaia kuantitatibo gisa dagoela kodifikatuta, eta, beraz, faktore modura aldatu behar dugula hasi aurretik, aurreko atalean ikasi bezala.

Atal guztian zehar datu-base honekin lan egingo dugunez, `attach` funtzioa erabiliko dugu aurreko praktikan azaldu bezala:

```
> #Datu-base honekin praktika guztian zehar lan egin behar dugunez
> attach(altuerak)
```

Honekin, aldagaian zuzenean atzitu ahalko ditugu, $ ikurra etengabe erabili gabe.

1.2. Maiztasun-taulak

Aldagai bakar bat aztertzean, `table` komandoarekin f maiztasun absolutuak lor ditzakegu. Adibidez, `amarena` aldagaiaren maiztasunak era honetan lortuko genituzke:

```
> table(amarena)

amarena
153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171
  1   3   6  10  12  16  14  15  24  19  12   7   8  11   4   1   1   3   2
```

Beste maiztasunak F, h eta H lortzeko, honako komando hauek dituen programa bat eraiki dezakegu:

```
> taula <- table(amarena) # x balioak eta f maiztasunak
> x <- names(taula) # xi balioak
> k <- length(taula) # xi ezberdin kopurua
> f <- taula # f maiztasun absolutuak
> n <- sum(f) # lagin tamaina
> h <- 100*taula/n # h portzentajeak
> F <- cumsum(f) # F maiztasun absolutu metatuak
> H <- cumsum(h) # H maiztasun absolutu metatuak
> names(F) <- names(f)
> names(H) <- names(h)
> taula <- as.data.frame(matrix(nrow=k,ncol=4))
> taula[,1]<- f
> taula[,2]<- F
> taula[,3]<- h
> taula[,4]<- H
> colnames(taula) <- c('f','F','h','H')
> rownames(taula) <- x
```

```
> #Maiztasun-taula:
> taula

     f   F         h          H
153  1   1  0.591716   0.591716
154  3   4  1.775148   2.366864
155  6  10  3.550296   5.917160
156 10  20  5.917160  11.834320
157 12  32  7.100592  18.934911
158 16  48  9.467456  28.402367
159 14  62  8.284024  36.686391
160 15  77  8.875740  45.562130
161 24 101 14.201183  59.763314
162 19 120 11.242604  71.005917
163 12 132  7.100592  78.106509
164  7 139  4.142012  82.248521
165  8 147  4.733728  86.982249
166 11 158  6.508876  93.491124
167  4 162  2.366864  95.857988
168  1 163  0.591716  96.449704
169  1 164  0.591716  97.041420
170  3 167  1.775148  98.816568
171  2 169  1.183432 100.000000
```

Kontuan izan `table` komandoak bi norabideko taulak eraikitzeko ere balio duela, bai aldagai kuantitatibo bat eta aldagai kualitatibo bat dugunean, eta bai bi aldagai kualitatiborentzat (kontingentzia-taulak):

```
> table(aitarena, jaioterria)

        jaioterria
aitarena Araba Bizkaia Gipuzkoa
     157     0       0        1
     160     1       0        0
     162     1       1        0
     163     0       2        1
     164     2       1        0
     165     2       2        2
     166     0       6        2
     167     0       3        3
     168     0       4        4
     169     4       5        4
     170     0       3        2
     171     4       4        3
     172     4       7        3
```

```
    173     3       4        5
    174     5      15        7
    175     2       5        6
    176     1       4        1
    177     0       6        2
    178     2       2        0
    179     0       1        0
    180     3       1        2
    181     0       2        1
    182     1       0        0
    183     2       0        0
    185     0       2        0
```

```
> table(altuerak$sexua, altuerak$jaioterria)


              Araba Bizkaia Gipuzkoa
  gizonezkoa     17      43       26
  emakumezkoa    21      37       23
```

Aurrerago, honelako taulak gehiago landuko ditugu 5. *R* praktikan.

1.3. Estatistikoak

Hasteko, oinarrizko estatistiko batzuk era honetan kalkula ditzakegu:

```
> summary(aitarena)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  157.0   169.0   173.0   172.1   175.0   185.0       1

> summary(sexua)

 gizonezkoa emakumezkoa
         89          82
```

Kontuan izan funtzio honek trataera ezberdina ematen diela aldagai kualitatibo eta kuantitatiboei.

Gainera, **fBasics** izeneko paketea instalatzen eta kargatzen badugu, `basicStats` komandoa erabil dezakegu estatistikoen zerrenda luzeago bat lortzeko:

```
> library(fBasics)
> basicStats(aitarena)

                aitarena
nobs          171.000000
NAs             1.000000
Minimum       157.000000
Maximum       185.000000
1. Quartile   169.000000
3. Quartile   175.000000
Mean          172.141176
Median        173.000000
Sum         29264.000000
SE Mean         0.376596
LCL Mean      171.397739
UCL Mean      172.884614
Variance       24.110129
Stdev           4.910207
Skewness       -0.073548
Kurtosis        0.157704
```

Kontuan izan `var` eta `sd` balioak S^2_{n-1} eta S_{n-1} estatistikoei dagozkiela. Beraz, oraindik, RI, S^2_n, S_n, CV eta ν estatistikoak falta dira, baina erraz lor daitezke honako kode hau erabiliz:

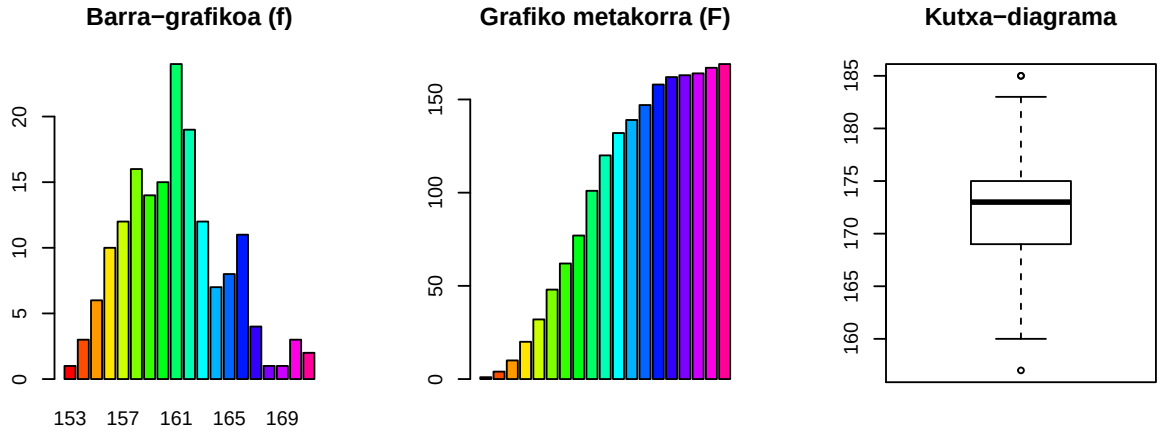
```
> # Errenkada kopurua edo laginaren tamaina
> n <- dim(altuerak)[1]
> # Moda
> moda <- as.numeric(x[which(f==max(f))])
> # 3. kuartila
> q3 <-quantile(aitarena,probs=0.75, na.rm=TRUE)
> # 1. kuartila
> q1 <-quantile(aitarena,probs=0.25, na.rm=TRUE)
> # Kuartilarteko heina
> RI <- q3 - q1
> # Bariantza
> varn <- var(aitarena, na.rm=TRUE)*(n-1)/n
> # Desbiderapen estandarra
> sdn <- sqrt(varn)
> # Alborapena
> nu <- (mean(aitarena, na.rm=TRUE)-moda)/sdn
> # Pearsonen aldakuntza-koefizientea
> cv <- 100*sdn/mean(aitarena, na.rm=TRUE)
```

Ohartu, baita ere, `na.rm=TRUE` komandoak balio galduak kontutan ez hartzeko adierazten diela funtzio ezberdinei.

1.4. Grafikoak

Elkartu gabeko datuen kasuan, honako komando hauek jarrai daitezke barra-grafikoa, grafiko metakorra eta kutxa-diagrama izeneko grafikoak ateratzeko.

```
> #Grafikoak 1x3 dimentsioko matrize batean adieraziko ditugu.
> par(mfrow=(c(1,3)))
> barplot(taula$f, col=rainbow(20), main="Barra-grafikoa (f)")
> barplot(taula$F, col=rainbow(20), main="Grafiko metakorra (F)")
> boxplot(aitarena, main="Kutxa-diagrama", xlab="", ylab="")
```



Kontuan izan kutxa-diagraman balio galduak ez datozela berez identifikatuta. Balio hauek identifikatzeko:

```
> # Identifikatu outlierren y balioak
> outlier.y <- boxplot(aitarena, plot=FALSE)$out
>
> # Lortu outlier-en x balioak
> outlier.x <- which(aitarena %in% outlier.y)
> outlier.x

[1]  55 117 149
```

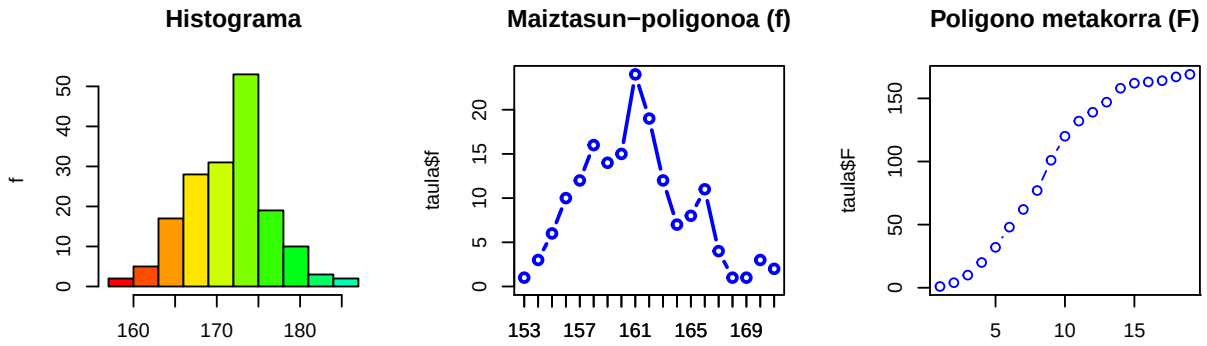
Beraz, balioa arraroak dituzte 55. 117. eta 149. erregistroak.

Gainera, zurtoin-eta hosto-diagrama ere egin dezakegu `stem` funtzioa erabiliz:

```
> stem(aitarena)


  The decimal point is at the |

  156 | 0
  158 |
  160 | 0
  162 | 00000
  164 | 000000000
  166 | 00000000000000
  168 | 00000000000000000000
  170 | 00000000000000000
  172 | 000000000000000000000000000
  174 | 0000000000000000000000000000000000000000000
  176 | 000000000000000
  178 | 00000
  180 | 000000000
  182 | 000
  184 | 00
```

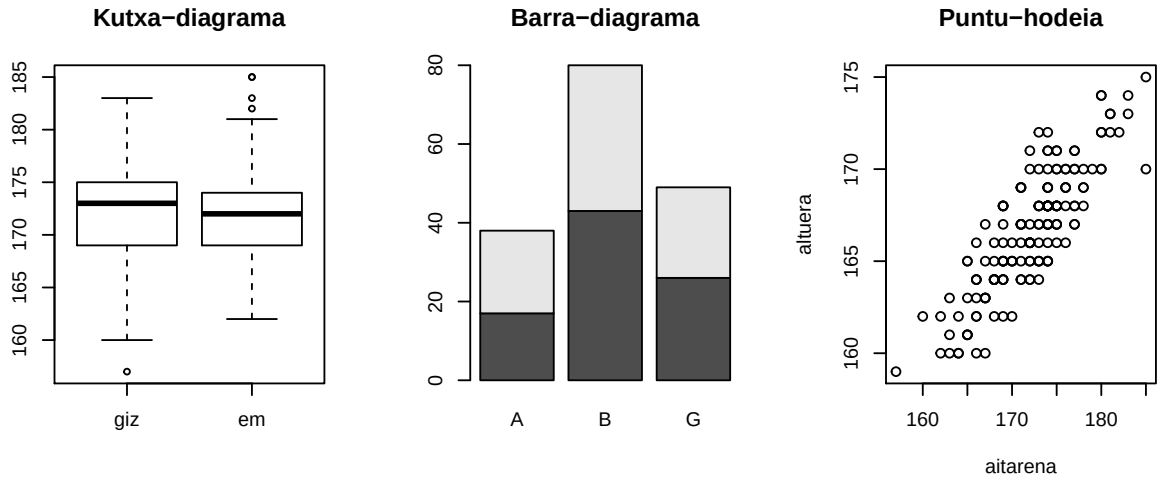
Datu elkartuak direnean, histograma, maiztasun-poligonoa eta poligono metakorra egin daitezke:

```
> par(mfrow=c(1,3))
> tarteak <- c(157, 160, 163, 166, 169, 172, 175, 178, 181, 184, 187)
> hist(aitarena, freq=TRUE, breaks=tarteak,xlab="", ylab="f",
+      main="Histograma",col=rainbow(20))
> plot(taula$f,col="blue",main="Maiztasun-poligonoa (f)",type="b",xlab=" ")
> axis(1,names(taula$f))
> plot(taula$F,col="blue",main="Poligono metakorra (F)",type="b",xlab=" ")
> axis(1,names(taula$f))
```



Bi aldagai ditugunean, aukera ezberdinak aztertuko ditugu aldagai moten arabera. Hauek dira grafiko posible batzuk:

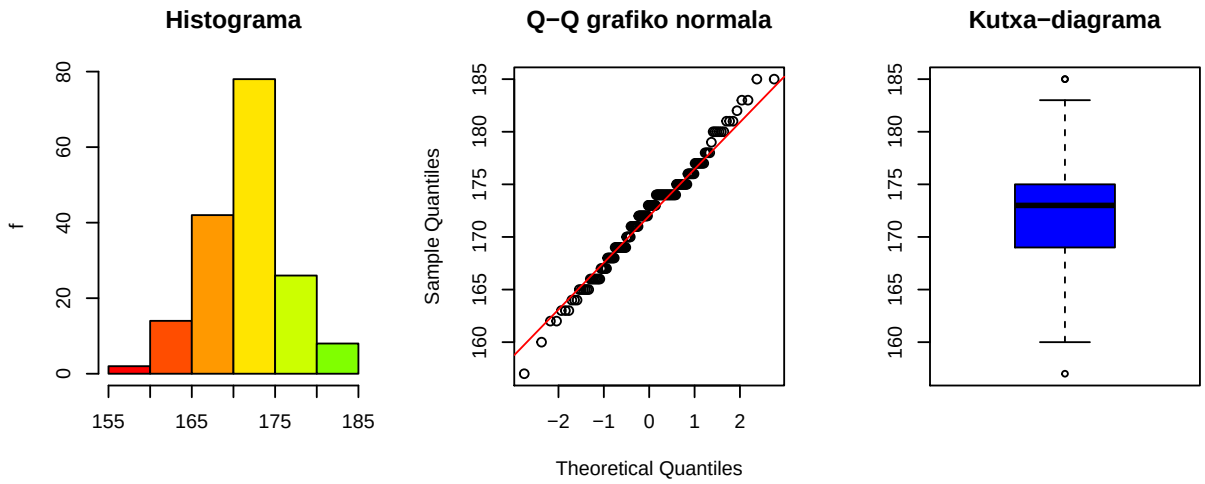
```
> par(mfrow=c(1,3))
> #Aldagai kualitatibo bat eta kuantitatibo bat: kutxa-diagrama
> boxplot(aitarena~sexua, main="Kutxa-diagrama", names=c("giz", "em"))
>
> #Bi aldagai kualitatibo: barra-diagrama
> barplot(table(sexua, jaioterria), main="Barra-diagrama",
+         names.arg=c("A", "B", "G"))
>
> #Bi aldagai kuantitatibo: puntu-hodeia
> plot(altuera~aitarena, main="Puntu-hodeia")
```



1.5. Normaltasuna aztertzeko metodo deskribatzaileak

Normaltasuna aztertzeko, honako grafiko hauek azter ditzakegu: histograma, probabilitate normalaren `qqnorm` grafikoa eta kutxa-diagrama.

```
> par(mfrow=c(1,3))
> hist(aitarena, xlab="", ylab="f",
+      main="Histograma",col=rainbow(20))
> qqnorm(aitarena, main="Q-Q grafiko normala")
> qqline(aitarena, col=2)
> boxplot(aitarena, main="Kutxa-diagrama", xlab="", ylab="",col="blue")
```



Ikusten dugunez, datuak ez dira asko aldentzen banaketa normaletik. 5. praktikan ikusiko dugu normaltasuna kuantitatiboki nola aztertu.

Azkenik, adibidea bukatu dugu eta datu-basea askatuko dugu `detach` funtzioa erabiliz:

```
> detach(altuerak)
```

2. *R* praktika

Probabilitatearen teoria

Helburua

Praktika honen jomuga da probabilitatearen teorian landutako kontzeptu batzuk jorratzea. Adibidez, zorizko aldagaien probabilitatearen legea, edo dentsitate-funtzioa, banaketa-funtzioa, pertzentilak kalkulatzea edo zorizko laginak sortzea eta grafikoki irudikatzea eta teorema batzuen aplikazioak zenbakiz eta grafikoki egiaztatzea.

2.1. Banaketak

Zorizko aldagai ezagun gehienen kasuan, pertzentilak, probabilitatearen legea, banaketa-funtzioa eta zorizko laginak era errazean kalkula ditzakegu, *R*ko berezko funtzio sorta bat erabiliz. Funtzio hauen izenak honako taula honetako aurrizki eta atzizki bat konbinatuz sortzen dira:

Aurrizkiak		Atzizkiak			
`q`	pertzentilak	`binom`	Binomiala	`norm`	Normala
`d`	probabilitatearen legea/dentsitate-funtzioa	`poisson`	Poisson	`chisq`	Khi karratua
`p`	banaketa-funtzioa	`unif`	Uniformea	`t`	Student-en t
`r`	zorizko lagina sortzea	`exp`	Esponentziala	`f`	Fisher-Snedecor-en F

Jarraian, aurrizki bakoitzaren azalpen bat emango dugu, hainbat adibide erakutsiz:

• `pertzentilak (koantilak)`: $\alpha \in (0,1)$ probabilitate guztietarako p_α balioa ematen digute funtzio hauek; $P(X \leq p_\alpha) = \alpha$ betetzen duena, baldin `lower.tail=TRUE` hautatzen badugu, eta $P(X > p_\alpha) = \alpha$ betetzen duena, `lower.tail=FALSE` hautatzekotan. Noski, banaketaren parametroak eta α-ren balioa (`p`) zehaztu beharko ditugu kasu bakoitzean. Adibidez,

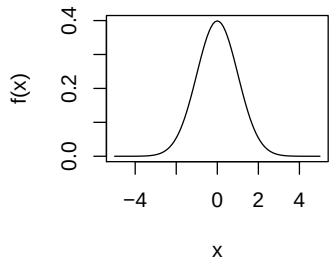
```
> #Banaketa normala N(0,1)-erako eta alpha=0,05 eskuinera uzten duen balioa
> qnorm(p=0.05, mean=0, sd=1, lower.tail=FALSE)

[1] 1.644854

> #5 askatasun-graduko t banaketa eta alpha=0,05 ezkerrera uzten duen balioa
> qt(p=0.025, df=5, lower.tail=TRUE)

[1] -2.570582
```

• `banaketa-funtzioa`. $a \in \mathbb{R}$ zenbaki guztietarako banaketa-funtzioa (probabilitate metatua) ematen digu, $P(X \leq a)$, baldin `lower.tail=TRUE` hautatzen badugu, edo $P(X > a)$, `lower.tail=FALSE` hautatzekotan. Noski, banaketaren parametroak eta a-ren balioa (`q`) zehaztu beharko ditugu kasu bakoitzean. Adibidez,

```
> #2 eta 3 askatasun-graduetako Fisher-en banaketa batentzat P(F > 1,5)
> pf(q=1.5, df1=2, df2=3, lower.tail=FALSE)

[1] 0.3535534

> #X:Bin(10, 0,4) banaketa batentzat P(X<=4) probabilitatea
> pbinom(q=4, size=10, prob=0.4, lower.tail=TRUE)

[1] 0.6331033
```

• f `edo dentsitate-funtzioa/probabilitatearen legea.` Behin banaketaren parametroak finkatuta, f dentsitate-funtzioaren edo banaketa-funtzioaren balioak ematen dizkigu. Aldagai jarraituen kasuan, funtzio hauek soilik erabiliko ditugu dentsitate-funtzioa irudikatu nahi badugu; aldiz, aldagai diskretuen kasuan, $P(X = x)$ kalkulatzeko erabil dezakegu:

```
> #N(0,1) banaketaren dentsitate-funtzioa
> x <- seq(-5,5,0.1)
> fx <- dnorm(x, mean=0, sd=1)
> plot(x, fx, xlab="x", ylab="f(x)", type="l")
```



```
> #X:P(3) poisson banaketa izanik, P(X=2):
> dpois(x=2, lambda=3)

[1] 0.2240418

> #X:Bin(10, 0,2) banaketa binomiala izanik, P(X=3):
> dbinom(x=3, size=10, prob=0.2)

[1] 0.2013266
```

• `Lagina`. Banaketa zehatz batetik ateratako zorizko laginak ematen dizkigu. Horretarako, banaketa mota eta bere parametroa(k) eta laginaren tamaina finkatu behar ditugu:

```
> #2 askatasun-graduko khi karratu batetik, 10 elementuko zorizko lagina:
> rchisq(n=5, df=2)

[1] 0.06150303 0.20230790 3.56524267 1.57064821 0.82919803

> #X:U(0,1) banaketa uniformetik, 10 elementuko zorizko lagina:
> runif(n=5, min=0, max=1)

[1] 0.01592651 0.16120601 0.49691595 0.14959513 0.16447929
```

Gogoratu funtzio hauei guztiei buruzko informazioa R-ren laguntzan daukazuela eskuragai.

2.2. Probabilitatearen teoriaren oinarrizko teorema batzuk

Azter ditzagun, grafikoki, banaketa batzuen hurbilketen inguruko emaitza teoriko batzuk.

2.2.1. Banaketa binomialaren eta Poisson-en banaketen arteko erlazioa

Teoriatik, badakigu $n > 50$ eta $p < 0,1$ badira, $Bin(n,\ p) \approx \mathcal{P}(np)$ betetzen dela. Ikus dezagun, adibidez, $X : Bin(50,\ 0,01) \approx Y : \mathcal{P}(0,5)$ betetzen dela, f probabilitatearen legeak eta F banaketa-funtzioak grafikoki konparatuz.

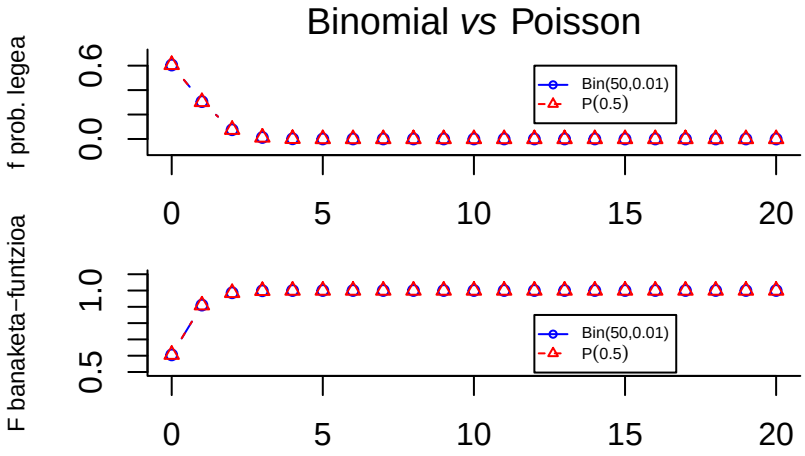
Hasteko, gure probarako x balio batzuk aukeratuko ditugu, eta horien probabilitatearen legearen balioak eta banaketa-funtzioen balioak kalkulatuko ditugu, bai X banaketa erabiliz eta bai Y banaketa erabiliz.

```
> # x balio posibleak
> x1 <- c(0:20)
```

```
> #x horietarako probabilitatearen legearen balioak, X:Bin(50, 0,01) erabiliz.
> binf <- dbinom(x1, size=50, prob=0.01)
> #x horietarako probabilitatearen legearen balioak, Y:P(0,5) erabiliz.
> poissonf <- dpois(x1, lambda=0.5)
> #x horietarako banaketa-funtzioaren balioak, X:Bin(50, 0,01) erabiliz.
> binF <- pbinom(x1, size=50, prob=0.01)
> #x horietarako banaketa-funtzioaren balioak, Y:P(0,5) erabiliz.
> poissonF <- ppois(x1, lambda=0.5)
```

Jarraian, lortutako balioak grafikoki konparatuko ditugu; alde batetik, f probabilitatearen legearen balioak, eta bestetik, F banaketa-funtzioaren balioak:

```
> #Irudian marjinak aldatu
> par(mfrow=c(2,1), mar = c(1.5,4,1.5,2),  oma = c(1,1,1,1), cex.lab=0.75)
> #1-Irudia
> plot(binf~x1, type="b", ylab="f prob. legea", xlab=" ",
+      ylim=c(-0.1, 0.7), col="blue", pch=1, lty=1, bty='l', cex=0.75)
> title(substitute(paste("Binomial ", italic('vs'), " Poisson")))
> lines(poissonf~x1,col="red",type="b", pch=2, lty=2, cex=0.75)
> #Legenda
> legend(12, 0.6, c("Bin(50,0.01)", expression(P(0.5))),
+        col=c("blue","red"), pch=c(1,2), lty=c(1,2), cex=0.5)
> #2-Irudia
> plot(binF~x1, type="b", ylab="F banaketa-funtzioa", xlab=" ",
+      ylim=c(0.5, 1.1), col="blue", pch=1, lty=1, bty='l', cex=0.75)
> lines(poissonF~x1, col="red", type="b", pch=2, lty=2, cex=0.75)
> #Legenda
> legend(12, 0.85, c("Bin(50,0.01)", expression(P(0.5))),
+        col=c("blue","red"), pch=c(1,2), lty=c(1,2), cex=0.5)
```



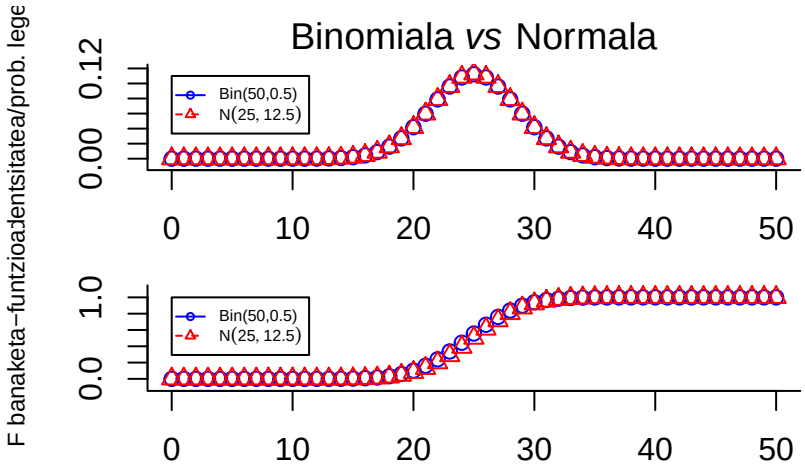
Ohartu `legend` komandoan lehenik legendaren kokapena adierazi dugula, eta, gero, agertuko den testua, koloreak, ikur eta kurba motak daudela adierazita. Bestalde, `plot` aginduko aukerak zer adierazten duten ikusteko, `help(par)` sakatu eta «Graphical Parameters» atalera jo.

2.2.2. Banaketa binomialaren eta normalaren arteko erlazioa (Moivre-ren teorema)

Gogoratu baldin $n > 30$ eta $0,1 < p < 0,9$ badira, $Bin(n,\ p) \approx \mathcal{N}(np,\ \sqrt{npq})$ dela (Moivre-ren teorema). Adibidez, azter dezagun grafikoki $Bin(50,\ 0,5) \approx \mathcal{N}(25,\ \sqrt{12,5})$ kasua. Honako bektore hauek definituz hasiko gara:

```
> # x balio posibleak
> x2 <- c(0:50)
> #x horietarako probabilitatearen legearen balioak, X:Bin(50, 0,01) erabiliz.
> binf <- dbinom(x2, size=50, prob=0.5)
> #x horietarako probabilitatearen legearen balioak, Y:N(25, sqrt(12,5)) erabiliz.
> normalf <- dnorm(x2, mean=25, sd=sqrt(12.5))
> #x horietarako banaketa-funtzioaren balioak, X:Bin(50, 0,01) erabiliz.
> binF <- pbinom(x2, size=50, prob=0.5)
> #x horietarako banaketa-funtzioaren balioak, Y:N(25, sqrt(12,5)) erabiliz.
> normalF <- pnorm(x2, mean=25, sd=sqrt(12.5))
```

Ondoren, aurreko ataleko prozedura berdina jarraituz egingo dugu irudikapena, honako hau lortuz:

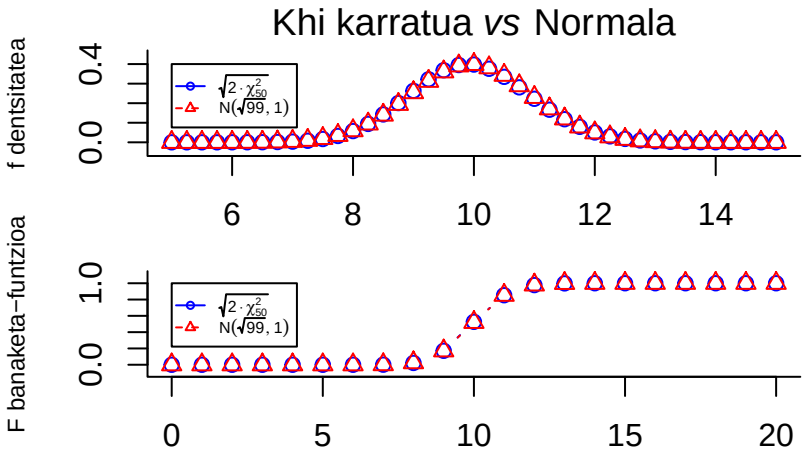


2.2.3. Khi karratu banaketaren eta banaketa normalaren arteko erlazioa

Kontuan izan baldin $n > 30$ bada, $\chi^2_{\alpha;n} \approx \frac{1}{2}(z_\alpha + \sqrt{2n-1})^2$ dela. Baliokideki, $\sqrt{2 \cdot \chi^2_{\alpha;n}} \approx z_\alpha + \sqrt{2n-1}$. Beraz, $\sqrt{2 \cdot \chi^2_n} \approx \mathcal{N}(\sqrt{2n-1}, 1)$ denez, $P(\chi^2_n \leq x^2/2) = P(\mathcal{N}(\sqrt{2n-1}, 1) \leq x)$; hau da, $F_\chi(x^2/2) = F_\mathcal{N}(x)$ eta deribatuz, $x \cdot f_\chi(x^2/2) = f_\mathcal{N}(x)$.

Adibidez, ikus dezagun $\sqrt{2 \cdot \chi^2_{50}} \approx \mathcal{N}(\sqrt{2 \cdot 50 - 1}, 1)$. Horretarako, honako bektore hauek kalkulatuko ditugu, x balio ezberdinen dentsitate- eta banaketa-funtzioen balioak kalkulatuz eta jarraian irudikatuz:

```
> x3 <- seq(5, 15, 0.25)
> chi2f <- x3*dchisq(x3*x3/2, df=50)
> normalf <- dnorm(x3, mean=sqrt(2*50-1), sd=1)
> x4 <- c(0:20)
> chi2F <- pchisq(x4*x4/2, df=50, lower.tail=TRUE)
> normalF <- pnorm(x4, mean=sqrt(2*50-1), sd=1)
```

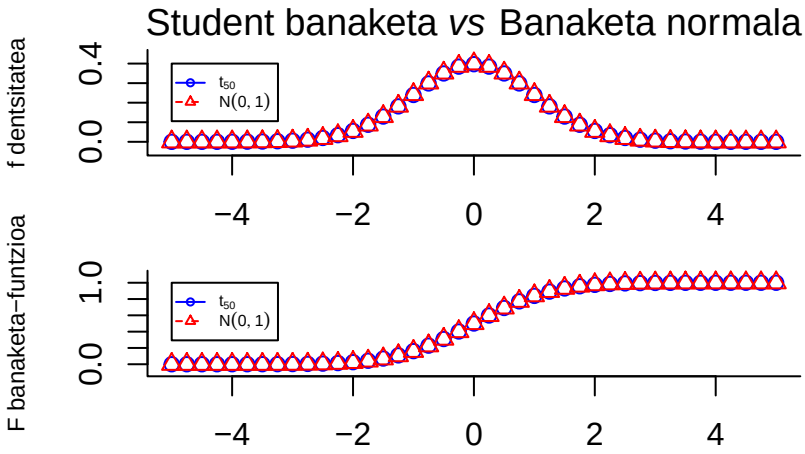


2.2.4. Student-en banaketaren eta banaketa normalaren arteko erlazioa

Baldin $n > 30$ bada, $t_n \approx \mathcal{N}(0,\ 1)$.

```
> x5 <- seq(-5, 5, 0.25)
> tf <- dt(x5, df=50)
> normalf <- dnorm(x5, mean=0, sd=1)
> tF <- pt(x5, df=50)
> normalF <- pnorm(x5, mean=0, sd=1)
```

Adibidez, ikus dezagun grafikoki $t_{50} \approx N(0,\ 1)$ dela:



3. *R* praktika

Konfiantza-tartezko zenbatespena

Helburua

Praktika honetan ikusiko dugu nola egin lagin bakar baterako eta bi laginen konparaziorako konfiantza-tarteak eta laginaren tamainaren kalkulua *R*ko komando bidez. Zehazki, honako konfiantza-tarte hauek nola kalkulatu ikasiko dugu:

- Lagin bakarreko batezbestekoaren $I_{\mu}^{1-\alpha}$ konfiantza-tartea, σ ezezagun eta ezagunerako.
- Lagin bakarreko bariantzarako $I_{\sigma^2}^{1-\alpha}$ konfiantza-tartea.
- Bi bariantzaren zatidurarako $I_{\sigma_1^2/\sigma_2^2}^{1-\alpha}$ konfiantza-tartea.
- Bi populazio askeren batezbestekoen kendurarako $I_{\mu_1-\mu_2}^{1-\alpha}$ konfiantza-tartea.
- Binakako datuen batezbestekoen kendurarako $I_{\mu_d}^{1-\alpha}$ konfiantza-tartea.
- Lagin bakarrerako proportziorako $I_{p}^{1-\alpha}$ konfiantza-tartea.
- Bi laginen proportzioen kendurarako $I_{p_1-p_2}^{1-\alpha}$ konfiantza-tartea.

3.1. Lagin bakar baten konfiantza-tarteak

18. adibidea. *Har ezazu berriro ere 3. Rcmdr praktikako 2. adibidea. Bertan, hiri txiki batean, ur-erabilerari buruzko ikerkuntza baterako, 25 etxebizitzatako zorizko lagina ateratzen zela esaten zen. Banaketa normalari darraion X aldagaia aztertzen zen:* asteko erabilitako ur litro kopurua. *Lagin hau erabiliz, kalkula ditzagun:*

(a) Batezbestekorako %90eko konfiantza-tartea, σ ezezaguna izanik. Hiriko ur-depositua aski handia al da asteko 160 litroko batez besteko kontsumoa baimentzeko?

(b) Batezbestekorako %90eko konfiantza-tartea, $\sigma^2 = 400$ bariantza ezaguna izanik. Egoera hau aintzat hartuz, hiriko ur-depositua aski handia al da asteko 160 litroko batez besteko kontsumoa baimentzeko?
(c) Bariantzarako %90eko konfiantza-tartea.

Hasteko, karga ditzagun datuak bektore batean:

```
> ura <- c(175, 185, 186, 118, 158, 150, 190, 178, 137, 175,
+ 180, 200, 189, 200, 180, 172, 145, 192, 191, 181, 183, 169,
+ 172, 178, 210)
```

Kontuan izan datu hauek **praktikadatuak.xls** fitxategiko **ur** orrialdean daudela eskuragai.

3.1.1. Batezbestekorako konfiantza-tarteak

(a) atalari erantzunez, kasurik ohikoenean, hau da, **bariantza ezezaguneko** batezbestekorako konfiantza-tarteak eraikitzeko, `t.test` funtzioa erabiliko dugu era honetan:

```
> t.test(ura, conf.level=.90)


One Sample t-test

data:  ura
t = 42.264, df = 24, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 168.6451 182.8749
sample estimates:
mean of x
   175.76
```

Ikusten dugu $I_\mu^{0.90} = (168.6451,\ 182.8749)$ dela, beraz, ur-depositua ez da nahikoa handia.

Aldiz, **bariantza ezaguna** denean, `t.test` funtzioak ez digu nahi dugun tartea ematen. Alabaina, badakigu kasu honetan konfiantza-tartearen formula honako hau dela: $I_\mu^{1-\alpha} = \left(\bar{x} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$. Honetan oinarrituz, eraiki dezagun guk geuk funtzio orokor bat:

```
> #x: lagina
> #sigma2: populazioaren bariantza
> #km: konfiantza-tartea
> z.konfiantza.tartea <- function(x, sigma2, km){
+ lbb <- mean(x) # Lagin-batezbestekoa
```

```
+ n <- length(x) # Laginaren tamaina
+ alpha <-1 - km
+ errorea <- qnorm(1-alpha/2)*sqrt(sigma2)/sqrt(n)
+ behe <- lbb-errorea #beheko muga
+ goi <- lbb+errorea #goiko muga
+ return(c(behe, goi))
+ }
```

`z.konfiantza.tartea` funtzioak, datu lagin bat `x`, bariantza ezagunaren balioa `sigma2` eta konfiantza-maila `km` parametroak hartzen ditu sarrera parametro bezala, eta bariantza ezaguneko batezbestekoaren konfiantza-tartea itzultzen du. Kontuan izan `qnorm(alpha)` funtzioak $z_{1-\alpha}$ balioa itzultzen digula.

Orain, 18. adibideko (b) atalean, bariantza ezaguna dela esaten digute, hots $\sigma^2 = 400$. Beraz, konfiantza-tartea eraikitzeko:

```
> z.konfiantza.tartea(x=ura, sigma2=400, km=0.9)

[1] 169.1806 182.3394
```

Ikusten dugu $I_{\mu}^{0.90} = (169.1806,\ 182.3394)$ dela; beraz ur-depositua ez da nahikoa handia.

3.1.2. Bariantzarako konfiantza-tarteak

Era berean, normaltasunaren menpeko bariantzarako zenbatespen-tartea kalkulatzeko formula $I_{\sigma^2}^{1-\alpha} = \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2;n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2;n-1}}\right)$ dela kontuan hartuta, honako funtzio hau programa dezakegu berau kalkulatzeko:

```
> #x: lagina
> #sigma2: populazioaren bariantza
> #km: konfiantza-maila
> bar.konfiantza.tartea <- function(x, km){
+   n <- length(x) # Laginaren tamaina
+   alpha <- 1 - km
+   lbar <- var(x) # Lagin-kuasibariantza
+   behe <- (n-1)*lbar /qchisq(1-alpha/2,df=n-1)
+   goi  <- (n-1)*lbar /qchisq(alpha/2,df=n-1)
+   return(c(behe, goi)) #Bariantzarako konfiantza-tartea
+ }
```

`bar.konfiantza.tartea` funtzioak datu-lagin bat `x` eta konfiantza-maila `km` parametroak hartzen ditu sarrera-parametro bezala, eta bariantzarako konfiantza-tartea itzultzen du.

Funtzioaren exekuzioa honako modu honetan gauzatuko dugu 18. adibidearen (c) atala ebazteko:

```
> bar.konfiantza.tartea(ura, 0.95)

[1] 263.6048 836.7417
```

Ikusten dugunez, `qchisq(alpha,df=n)` komandoak $\chi^2_{1-\alpha;n}$ itzultzen digu, banaketa-funtzioa erabiliz.

Desbiderapen estandarraren konfiantza-tartea lortzeko, erro karratua aplikatu beharko diogu aurreko konfiantza-tarteari:

```
> sqrt(bar.konfiantza.tartea(ura, 0.95))

[1] 16.23591 28.92649
```

Beraz, $I_\sigma = (16.23591,\ 28.92649)$.

Bestalde, bariantzarako konfiantza-tartea **EnvStats** paketeko `varTest` funtzioarekin ere egin dezakegu. Adibidez, aurreko datuak erabiliz, emaitza baliokideak lortuko ditugu `90\% Confidence Interval` atalean:

```
> library(EnvStats)
> varTest(ura, alternative = "two.sided", conf.level=0.9)
```

3.2. Lagin birako zenbatespen-tarteak

19. adibidea. *Har itzazu berriro ere 3. adibideko datuak. Bertan, lan zehatz bat egiteko enpresa bateko langileen denbora-tarteak (segundotan) adierazten ziren. Konpara itzazu emakumezkoen eta gizonezkoen bariantzak eta batezbestekoak, dagozkion konfiantza-tarteak erabiliz.*

Defini ditzagun, hasteko, datuak:

```
> emakumezkoa <- c(103, 94, 110, 87, 98)
> gizonezkoa <- c(97, 82, 123, 92, 175, 88, 118)
```

Datu horiek **praktikadatuak.xls** fitxategiko **denbora** orrialdean daude gordeta.

3.2.1. Bariantzen zatidurarako konfiantza-tarteak

Hasteko, bariantzen zatidurarako konfiantza-tarteak `var.test` komandoa erabiliz kalkulatuko ditugu, era honetan:

```
> var.test(emakumezkoa, gizonezkoa, conf.level=0.9)


F test to compare two variances

data:  emakumezkoa and gizonezkoa
F = 0.073655, num df = 4, denom df = 6, p-value = 0.02468
alternative hypothesis: true ratio of variances is not equal to 1
90 percent confidence interval:
 0.01624629 0.45394809
sample estimates:
ratio of variances
        0.07365542
```

Ikusten dugu $I^{0,90}_{\sigma_1^2/\sigma_2^2} = (0,01624629,\ 0,45394809)$. 1 balioa tartearen barnean ez dagoenez, $\sigma_1^2 \neq \sigma_2^2$ ondorioztatzen da % 90eko konfiantza-mailarekin. Gainera, tarteko balio guztiak 1 baino txikiagoak direnez, $\sigma_1^2 < \sigma_2^2$ dela esan dezakegu % 90eko konfiantza-mailarekin.

3.2.2. Bi populazio askeren batezbestekoen diferentziarako konfiantza-tarteak

Aurreko ataleko emaitza erabiliz, batezbestekoen diferentziarako konfiantza-tartea eraikiko dugu `t.test` funtzioa erabiliz:

```
> t.test(emakumezkoa, gizonezkoa, conf.level=0.9, var.equal=FALSE)


Welch Two Sample t-test

data:  emakumezkoa and gizonezkoa
t = -0.9638, df = 7.1866, p-value = 0.3664
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -36.42670  11.79813
sample estimates:
mean of x mean of y
  98.4000  110.7143
```

`var.equal=FALSE` aukera zehaztuz, bariantza ezberdineko populazioak direla zehazten dugu. Kontrako kasuan, `var.equal=TRUE` jarri beharko genuke.

$I^{0,90}_{\mu_1-\mu_2} = (-36,42670,\ 11,79813)$ dela ikus dezakegu, eta 0 tartearen barne dagoenez, $\mu_1 = \mu_2$ dela ezin dugu baztertu.

3.2.3. Binakako datuen batezbestekoen diferentziarako konfiantza-tartea

20. adibidea. *Har dezagun berriro 4. adibidea. Gogoratu, bertan, tomate freskoen eta ontziratuen kobre kopurua konparatu nahi genuela, %98ko konfiantza-tartea eraikiz.*

Datuak kargatzen ditugu:

```
> library(readxl)
> tomate <- read_excel("praktikadatuak.xls", sheet='tomate')
```

Konfiantza-tartea eraikitzen dugu. Horretarako, kontuan izan, binakako datuekin `paired=FALSE` komandoa gehitu behar dugula.

```
> t.test(tomate$freskoa, tomate$latakoa, conf.level=0.98, var.equal=FALSE)


Welch Two Sample t-test

data:  tomate$freskoa and tomate$latakoa
t = -2.817, df = 15.478, p-value = 0.0127
alternative hypothesis: true difference in means is not equal to 0
98 percent confidence interval:
 -0.0224698151 -0.0009301849
sample estimates:
mean of x mean of y
   0.0721    0.0838
```

3.3. Populazio binomialetarako konfiantza-tarteak

3.3.1. Proportziorako konfiantza-tarteak

21. adibidea. *Har ezazu berriro ere 3.. Rcmdr praktikako 5. adibidea. Bertan, suzirien aireratze-instalazio berri bat ikertzen ari direla diote. Dagoen sisteman, $p = 0,8$ da jaurtiketa batean arazorik ez egoteko probabilitatea. Sistema berriarekin 40 jaurtiketa esperimental egiten dira; horietatik 34tan ez da arazorik egon. Kalkula ezazu % 95eko p-ren konfiantza-tartea. Ondoriozta daiteke sistema berria hobea dela?*

% 95eko p-ren konfiantza-tartea kalkulatzeko eskatzen digute. Hasteko, banaketa binomial zehatza erabiliz kalkula dezakegu konfiantza-tartea `binom.test` funtzioa erabiliz, eta arrakasta kopurua (x), laginaren tamaina (n) eta konfiantza-maila (conf.level) zehaztuz:

```
> binom.test(x=34, n=40, conf.level=0.95)


Exact binomial test

data:  34 and 40
number of successes = 34, number of trials = 40, p-value = 8.365e-06
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.7016473 0.9428977
sample estimates:
probability of success
                 0.85
```

$I_p^{0,95} = (0,7016473,\ 0,9428977)$ da, eta $0,8 \in I_p^{0,95}$ denez, ezin dugu esan sistema berria zaharra baino hobea denik.

Laginaren tamaina handia denean, binomiala gutxi gorabehera normal estandarra denez, proportziorako honako konfiantza-tarte hau $I_p^{1-\alpha} = \left(\hat{p} \mp z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$ erabil dezakegu. Hori kalkulatzeko, programa dezagun guk geuk bi aukerak kontuan hartzen dituen funtzio bat:

```
> #x: arrakasta kopurua
> #n: laginaren tamaina
> #km: konfiantza-maila
> pop.konfiantza.tartea <- function(x, n, km){
+   lagina <- rep(c(0,1),c((n-x),(x)))
+   alpha <- 1 - km
+   p <- mean(lagina) # p parametroaren zenbatespen puntuala
+   q <- 1 - p
+   if(n*p>5 & n*q>5){
+     errorea <- qnorm(1-alpha/2) * sqrt(p*(1-p)/n)
+     behe <- p - errorea
+     goi <- p + errorea
+     return(c(behe,goi)) # Proportzioarako konfiantza-tartea
+   }else{
+     t <- binom.test(x, n, km)
+     return(t$conf.int[1:2])
+   }
+ }
```

Funtzio honek `x` arrakasta kopurua, `n` laginaren tamaina eta `km` konfiantza-maila hartzen ditu sarrera-parametro gisa, eta aipatutako konfiantza-tartea itzultzen du.

Gure kasuan:

```
> pop.konfiantza.tartea(x=34, n=40, km=0.95)

[1] 0.7393445 0.9606555
```

Beraz, arrakasta-proportzioaren zenbatespen-tartea $I_p^{0,95} = (0,7393,\ 0,9607)$ da, eta $0,8$ arrakasta-proportzioa bere barnean dago % 95eko konfiantza-mailarekin.

3.3.2. Bi proportzioen diferentziarako konfiantza-tarteak

22. adibidea. *Artikulu mota bat saltzen duen enpresa batek baieztatzen du A markakoak B markakoak baino gehiago saltzen direla, eta diferentzia % 8koa dela. 200 bezeroren artean 42k nahiago dute A markako artikulua, eta 150 bezeroren artean 18k B markakoa. Kalkula ezazu % 94ko konfiantza-tartea, bi marken proportzioen arteko diferentziarako. Erabaki ezazu ea baliozkotzat har daitekeen % 8ko diferentziaren baieztapena.*

$I_{p_1-p_2}^{0,94}$ tartea kalkulatu behar dugu. Laginaren tamainak handiak badira, hurbilketa normala erabil daitekeenez, diferentziarako konfiantza-tartearen adierazpena honako hau da:

$$I_{p_1-p_2}^{1-\alpha} = \left(\hat{p}_1 - \hat{p}_2 \mp z_{\alpha/2} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \right)$$

Programa dezagun:

```
> x1 <- 42 # arrakasta kopurua A markan
> n1 <- 200  # A markako produktuen laginaren tamaina
> x2 <- 18  # arrakasta kopurua B markan
> n2 <- 150 # B markako produktuen laginaren tamaina
> km <- 0.94 # konfiantza-maila
> alpha<- 1 - km
> p1 <- x1/n1 # p1-en zenbatespen puntuala
> p2 <- x2/n2 # p2-ren zenbatespen puntuala
> d <- p1-p2 # (p1-p2)-ren zenbatespen puntuala
> errorea <- qnorm(1-alpha/2) * sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2)
> behe <- d - errorea
> goi <- d + errorea
>
> #estimazio puntuala eta (p1-p2)-ren konfiantza-tartea
> c(d, behe, goi)

[1] 0.09000000 0.01634826 0.16365174
```

Era berean, funtzio orokor bat ere egin daiteke.

3.4. Laginaren tamaina

R programarekin, oso erraza da laginaren tamaina kalkulatzea.

23. adibidea. *1.000 etxetako zorizko lagin batean, ikusi da 228tan butanoa erabiltzen dela. Zein izan behar da laginaren tamaina, % 99ko konfiantza-mailarekin, laginaren proportzioaren eta populazioaren proportzioaren arteko errorea % 5 baino txikiagoa izateko?*

$n = \frac{z^2_{\alpha/2}\hat{p}\hat{q}}{e^2}$ laginaren tamainaren adierazpena dela kontuan hartuta:

```
> # x: arrakasta kopurua
> # n: laginaren tamaina
> # km: konfiantza-maila
> lagin.tamaina.kt <- function(x, n, km, errorea){
+   alpha <- 1 - km
+   p <- x/n # p-ren puntu zenbatespena
+   q <- 1 - p
+   n <- qnorm(1-alpha/2)**2 * p * q / errorea**2
+   return(ceiling(n))
+ }
```

Gure kasurako exekuzioa:

```
> lagin.tamaina.kt(x=228, n=1000, km=0.99, errorea=0.05)

[1] 468
```

Beraz, laginaren tamainak gutxienez 468koa izan behar du.

4. *R* praktika

Hipotesi-kontraste parametrikoak

Helburua

Praktika honetan ikasiko dugu nola egin hipotesi-kontraste parametrikoak *R*ko kodea erabiliz. Zehazki:

- `Lagin bakarreko batezbestekorako testak` σ `ezezagun eta ezagunerako`: $H_0 : \mu = \mu_0$
- `Lagin bakarreko bariantzarako testa`: $H_0 : \sigma = \sigma_0$
- `Bi bariantza konparatzeko F testa`: $H_0 : \sigma_1^2 = \sigma_2^2$
- `Bi lagin independenteko t testa`: $H_0 : \mu_1 = \mu_2$ (lagin askeak).
- `Datu binakatuentzako t testa`: $H_0 : \mu_1 = \mu_2$ (binakako datuak).
- `Lagin bakarrerako proportzio-testa`: $H_0 : p = p_0$ (binomial zehatza)
- `Bi laginen proportzioen konparaketarako testa`: $H_0 : p_1 = p_2$

Gainera, horietako zenbait hipotesi-kontrastetarako erroreen kalkulua eta laginaren tamainaren kalkuluak ere programatuko ditugu.

4.1. Lagin bakar baten hipotesi-kontrasteak

24. adibidea. *Har ezazu berriro ere 18. adibidea, eta egin ditzagun, kodea erabiliz,* μ*-rako eta* σ*-rako test ezberdinak honako galdera hauei erantzuteko:*

(a) % 5eko esangura-mailaz esan dezakegu hiriko batezbesteko ur-kontsumoa 170 litro baino handiagoa dela?

(b) Populazioaren desbiderapena 20 dela onartzen badugu, % 5eko esangura-mailaz esan dezakegu hiriko batezbesteko ur-kontsumoa 170 litro baino handiagoa dela?

Gogoratu datuak **praktikadatuak.xls** fitxategiko **ur** orrialdean daudela eskuragai.

4.1.1. Batezbestekorako hipotesi-kontrasteak

Ohikoena da **bariantza ezezaguneko** populazio normal baten batezbestekorako hipotesi-kontrastea egin nahi izatea. Horretarako, aurreko gaian azaldutako `t.test` funtzioa erabiliko dugu. Adibidez, adibideko (a) atalari erantzuteko, $H_1 : \mu > 170$ hipotesi alternatiboa duen testa gauzatuko dugu:

```
> t.test(ura, alternative="greater", mu=170)


One Sample t-test

data:  ura
t = 1.3851, df = 24, p-value = 0.08939
alternative hypothesis: true mean is greater than 170
95 percent confidence interval:
 168.6451      Inf
sample estimates:
mean of x
   175.76
```

Ikusten dugunez, estatistikoaren balioa $t = 1,3851$ da eta $p = 0,08939$. Hori horrela, $\alpha = 0,05$ esangura-maila hartuz, $(p > \alpha)$ esan dezakegu hipotesi nulua ezin dugula baztertu; hau da, $\mu \leq 170$ ezin dugu baztertu.

Kontuan izan kontrako hipotesi-kontrastea egiteko ($H_1 : \mu < 170$) `alternative="less"` jarri beharko genukeela, eta alde biko testa egiteko, aldiz, ($H_1 : \mu \neq 170$), `alternative="two.sided"`.

Bariantza ezaguna den kasurako `t.test` funtzioak ez digu balio; beraz, guk geuk programatuko dugu hipotesi-kontrastea. Horretarako, batezbestekorako eskuinetiko testa egiten duen funtzio bat eraikiko dugu, bariantza ezaguna dela suposatuz:

```
> #x: lagina
> #sigma: populazioaren desbiderapen estandarra
> #mu0: batezbestekoaren erreferentziazko balioa
>
> eskuinetiko.z.test <- function(x, sigma, mu0, alpha){
+   n <- length(ura) # laginaren tamaina
+   bb <- mean(ura) # lagin-batezbestekoa
+   s <- sd(ura) # laginaren desbiderapen estandarra
```

```
+   z <- (bb-mu0)/(sigma/sqrt(n)) # Estatistikoa
+   oe <- c(-Inf, qnorm(1-alpha)) # Onarpen-eskualdea
+   p <- 1 - pnorm(z) #p-balioa
+   return(list(estatistikoa=z, onarpen.eskualdea=oe, p.balioa=p))
+ }
```

Ikusten dugunez, funtzio honek lagina (`x`), populazioaren desbideratze estandarraren ezagunaren balioa (`sigma`), kontrastatu nahi dugun batezbestekoaren balioa (`mu0`) eta esangura-maila (`alpha`) hartzen ditu sarrera-parametro bezala, eta estatistikoa, onarpen-eskualdea eta p-balioa itzultzen dituen zerrenda bat ematen du.

Adibidez, adibidearen (b) atalari erantzuteko, egin dezagun berriro ere $H_1 : \mu > 170$ hipotesi alternatibodun testa gure funtzioa erabiliz:

```
> eskuinetiko.z.test(x=ura, sigma=20, mu0=170, alpha=0.05)

$estatistikoa
[1] 1.44

$onarpen.eskualdea
[1]     -Inf 1.644854

$p.balioa
[1] 0.0749337

> #Emaitzako atal ezberdinak atzitu ditzakegu $ erabiliz:
> eskuinetiko.z.test(x=ura, sigma=20, mu0=170, alpha=0.05)$estatistikoa

[1] 1.44
```

Ikusten dugunez, estatistikoaren balioa $z = 1,44$ da, onarpen-eskualdea $(-\infty,\ 1,644854)$ eta $p = 0,0749337$. Hori horrela, $\alpha = 0,05$ esangura-maila hartuz, $(p > \alpha)$ esan dezakegu hipotesi nulua ezin dugula baztertu; hau da, $\mu \leq 170$ ezin dugu baztertu. Era berean, formulak ezagutuz, beste edozein test programa dezakegu.

4.1.2. Bariantzarako hipotesi-kontrasteak

25. adibidea. *Autoko bateriak ekoizten dituen fabrikatzaile batek baieztatu du baterien erdibizitza 0,9 urteko desbideratze estandarra duen banaketa normalari darraiola. 10 tamainako honako zorizko lagin honetan oinarrituz, 0,05 esangura-mailarekin onar daiteke $\sigma > 0,9$ urte dela?*

0,82 0,84 1,05 2,93 1,61 1,61 2,46 2,79 0,73 2,89

$H_0 : \sigma \leq 0,9$ hipotesi nulua kontrastatzeko, dagokion funtzioa eta exekuzioa ikus ditzakegu jarraian:

```
> eskuinetiko.bar.test <- function(x, sigma0, alpha){
+   n <- length(x) # lagin tamaina
+   s <- sd(x)
+   chi2p <- (n-1)*s^2/sigma0^2 # Estatistikoa
+   oe <- c(0, qchisq(1-alpha/2, df=n-1)) #Onarpen-eskualdea
+   p <- 1 - pchisq(chi2p, df=n-1)
+   return(list(estatistikoa=chi2p, onarpen.eskualdea=oe, p.balioa=p))
+ }
```

`eskuinetiko.bar.test` funtzioak lagina (`x`), kontrastatu nahi dugun desbideratze estandarraren balioa (`sigma0`) eta esangura-maila (`alpha`) hartzen ditu sarrera-parametro bezala, eta estatistikoa, onarpen-eskualdea eta p-balioa itzultzen dituen zerrenda bat ematen du.

Orain, funtzioa exekutatuz:

```
> #Exekuzioa
> x <- c(0.82 , 0.84 , 1.05 , 2.93 , 1.61 , 1.61 , 2.46 , 2.79 , 0.73 , 2.89)
> eskuinetiko.bar.test(x, sigma0=0.9, alpha=0.05)

$estatistikoa
[1] 9.302481

$onarpen.eskualdea
[1]  0.00000 19.02277

$p.balioa
[1] 0.409834
```

$\chi^2_p = 9,302481$ estatistikoa $S_0 = (0,\ 19,02277)$ onarpen-eskualdean dagoenez, edo baliokideki p-balioa $0,409834 > \alpha$ denez, $H_0 : \sigma \leq 0,9$ hipotesia ezin da baztertu % 5eko esangura-mailarekin.

Bestalde, bariantzaren testa, **EnvStats** paketeko `varTest` funtzioarekin ere egin dezakegu, aurreko gaian aipatu dugun moduan.

4.2. Lagin birako hipotesi-kontrasteak

26. adibidea. *Har ditzagun berriro 19. adibideko datuak; konpara itzazu emakumezkoen eta gizonezkoen denboren bariantzak eta, ondoren, batezbestekoak.*

Gogoratu datuak **praktikadatuak.xls** fitxategiko **denbora** orrialdean daudela:

```
> #Datuak inportatu
> library(readxl)
> denbora <- read_excel("praktikadatuak.xls", sheet="denbora")
> #Sexua aldagaia faktore bihurtu
> denbora$sexua <- factor(denbora$sexua, levels=c(1, 2),
+                         labels=c("emakumezkoa", "gizonezkoa"))
```

4.2.1. Bi bariantzen hipotesi-kontrasteak

Bi populazio-bariantza konparatzeko hipotesi-kontrastea `var.test` funtzioarekin gauzatuko dugu.

```
> var.test(denbora~sexua, data=denbora, alternative="two.sided")


F test to compare two variances

data:  denbora by sexua
F = 0.073655, num df = 4, denom df = 6, p-value = 0.02468
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.01182809 0.67743181
sample estimates:
ratio of variances
        0.07365542
```

Lortutako emaitza 3.2.1. atalean lortutakoaren berdina da. Ikusten dugunez, $F = 0,073655$ da estatistikoaren balioa, 4 eta 6 askatasun-graduko Fisher banaketa jarraitzen duena. $p = 0,02468$ da, beraz, $\alpha = 0,05$ hartuta, $p < \alpha$, eta, ondorioz, hipotesi nulua errefusatzen dugu eta $\sigma_1^2 \neq \sigma_2^2$ onartu.

4.2.2. Bi populazio askeren batezbestekoak konparatzeko hipotesi-kontrastea

Bi batezbestekoak konparatzeko hipotesi-kontraste gehienak jada ezaguna dugun `t.test` funtzioaren bitartez egin ditzakegu.

Hasteko, har dezagun berriro ere 26. adibidea. Kasu honetan, bariantza ezezaguneko bi populazio aske dauzkagu (emakumezkoena eta gizonezkoena), eta populazio bakoitzetik lagin bat, $n_1 = 5$ eta $n_2 = 7$ izanik. Gainera, aurreko atalean egindako bariantzak konparatzeko testarekin ondorioztatu dugu $\sigma_1^2 \neq \sigma_2^2$. Orain, $H_1 : \mu_1 > \mu_2$ hipotesi alternatiboa duen testa egin nahi dugu [1]:

[1] Bariantzak berdinak balira, honako aldaketa hau egin beharko genuke kodean: `var.equal = FALSE`. Gainera, `alternative` parametroa `less` edo `two.sided` jarriz, test ezberdinak egin ditzakegu.

```
> t.test(denbora~sexua, data=denbora, alternative = "greater", var.equal=FALSE)


Welch Two Sample t-test

data:  denbora by sexua
t = -0.9638, df = 7.1866, p-value = 0.8168
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -36.4267      Inf
sample estimates:
mean in group emakumezkoa  mean in group gizonezkoa
                  98.4000                   110.7143
```

Ikusten dugu $t = -0,9638$ dela estatistikoaren balioa, eta $p = 0,8168$ dela. 0,05-eko esanguramaila hartuz, $p > \alpha$, eta, beraz, hipotesi nulua $H_1 : \mu_1 \leq \mu_2$ ezin dugu baztertu.

4.2.3. Binakako datuen batezbestekoak konparatzeko hipotesi-kontrastea

Datuak binakakoak badira ere, `t.test` funtzioa erabil dezakegu aldaketa txiki batekin.

27. adibidea. *20. adibideko tomate freskoen eta ontziratutakoen kobre kopuruaren problema berriro hartzen dugu, eta $H_1 : \mu_1 < \mu_2$ hipotesi alternatiboa duen kontrastea egin nahi dugu.*

```
> freskoak <- c(0.066, 0.079, 0.069, 0.076, 0.071,0.087, 0.071,
+               0.073, 0.067, 0.062)
> ontziratutakoak <- c(0.085, 0.088, 0.091, 0.096, 0.093, 0.095,
+                      0.079, 0.078, 0.065, 0.068)
> t.test(freskoak, ontziratutakoak, alternative = "less", paired=TRUE)


Paired t-test

data:  freskoak and ontziratutakoak
t = -4.4079, df = 9, p-value = 0.0008504
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
         -Inf -0.006834283
sample estimates:
mean of the differences
                -0.0117
```

$t = -4,4079$ eta $p = 0,0008504$ denez, $\alpha = 0,05$ hartuz, H_0 errefusatzen dugu eta $\mu_1 < \mu_2$ onartuko dugu; hau da, tomate freskoen kobre kopurua ontziratutakoena baino txikiagoa da.

`t.test` funtzioaren bitartez gauzatu ezin dugun kasu bakarra bi populazio-bariantzak ezagunak direneko kasua dugu. Kasu honetan, hipotesi-kontrastea aurreko ataletan ikasi dugun bezala programatu beharko genuke.

4.3. Populazio binomialetarako hipotesi-kontrasteak

4.3.1. Proportzio baterako hipotesi-kontrastea

Demagun 21. adibideko egoeran gaudela, eta $H_1 : p \neq 0,8$ hipotesi alternatiboa duen testa gauzatu nahi dugula. Banaketa binomial zehatza erabiltzen duen testa gauzatu nahi badugu, `binom.test` funtzioa erabiliz eta arrakasta kopurua (x), laginaren tamaina (n) eta arrakastaren probabilitatea (p) zehaztuz:

```
> binom.test(x=34, n=40, p=0.8, alternative="two.sided")


Exact binomial test

data:  34 and 40
number of successes = 34, number of trials = 40, p-value = 0.5541
alternative hypothesis: true probability of success is not equal to 0.8
95 percent confidence interval:
 0.7016473 0.9428977
sample estimates:
probability of success
                  0.85
```

$p = 0,5541$ da eta, beraz, $\alpha = 0,05$ hartuz, ezin dugu hipotesi nulua baztertu.

Laginaren tamaina handia denean, binomiala gutxi gorabehera normal estandarra da. Beraz, aurreko kasu batzuetan ikusi bezala eta dagozkion formulak erabiliz, programatu egin beharko genuke.

4.3.2. Bi proportzioen diferentziarako hipotesi-kontrasteak

Egin dezagun 22. adibidea, hipotesi-kontraste egokia erabiliz. Kontrastatu nahi dugun hipotesi nulua $H_0 : p_1 - p_2 = 0,08$ da. Programa dezagun, hasteko, mota honetako testak gauzatuko dituen funtzio bat:

```
> #x1: 1. lagineko arrakasta kopurua
> #n1: 1. laginaren tamaina
> #x2: 2. lagineko arrakasta kopurua
> #n1: 2. laginaren tamaina
> #km: konfiantza-maila
> #d0: diferentziarako erreferentzia-balioa
> #mota: test mota "alde.bikoa", "eskuinetikoa", "ezkerretikoa"
>
> biprop.test <- function(x1, n1, x2, n2, d0, mota){
+ p1 <- x1/n1
+ p2 <- x2/n2
+ zp <- (p1-p2-d0)/sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2)
+ if (mota=="alde.bikoa") {
+ p.bal <- 2*(1-pnorm(abs(zp))) # p-balioa
+ } else if (mota=="eskuinetikoa"){
+ p.bal <- 1-pnorm(zp)
+ } else {
+ p.bal <- pnorm(zp)
+ }
+ return(list(estatistikoa=zp, pbalioa=p.bal))
+ }
```

Kontuan izan funtzio hau erabiltzeko zehaztu egin behar ditugula bi laginen tamainak (`n1` eta `n2`) eta arrakasta kopuruak (`x1` eta `x2`), `d0` proportzioen diferentziaren kontrastatu nahi dudan balioa, eta, azkenik, kontraste mota (`alde.bikoa`, `eskuinetikoa` edo `ezkerretikoa`). Bueltan, estatistikoa eta p-balioa jaulkiko dira.

Orain, adibidea ebazteko beharrezko testa gauzatzeko, funtzio hau erabiliko dugu:

```
> biprop.test(x1=42, n1=200, x2=18, n2=150, d0=0.08, mota="alde.bikoa")

$estatistikoa
[1] 0.2553631

$pbalioa
[1] 0.7984427
```

$z_p = 0,2554$ eta $p-balioa = 0,7984 > \alpha = 0,06$ denez, H_0 ez da errefusatzen; hots, diferentzia, $p_1 - p_2 = 0,08$. Hau da, ezin da baztertu % 6ko esangura-mailarekin A markako saldutako artikulu-proportzioaren eta B markakoenaren arteko diferentzia % 8koa denik.

4.4. Erroreak eta laginaren tamaina

R erabiliz, erraza da α I motako errorea, β II motako errorea, $1-\alpha$ konfiantza-maila, $1-\beta$ ahalmena eta n laginaren tamaina kalkulatzea. Horretarako, haien definizioetatik abiatuz, formulak besterik ez ditugu inmplementatu behar. Izan ere,

$$\begin{aligned}
\alpha &= P(H_0 \textit{ baztertu} \mid H_0 \textit{ egiazkoa}),\\
1-\alpha &= P(H_0 \textit{ onartu} \mid H_0 \textit{ egiazkoa}),\\
\beta &= P(H_0 \textit{ onartu} \mid H_0 \textit{ gezurrezkoa}),\\
1-\beta &= P(H_0 \textit{ baztertu} \mid H_0 \textit{ gezurrezkoa}).
\end{aligned}$$

4.4.1. I eta II motako erroreak eta ahalmena

I eta II motako erroreak, konfiantza-maila zein ahalmena kalkula ditzagun, kodea erabiliz eta adibide honetan oinarrituz:

28. adibidea. *Enpresa bateko kafe-makina automatikoak zerbitzu bakoitzean ematen duen likido kopurua, gutxi gorabehera, banaketa normalari darraion aldagaia da, batezbestekoa 200 ml eta desbideratze estandarra 15 ml izanik. Aldizka, makina berraztertu egiten da bederatzi kaferen lagin baten batez besteko likido kopurua neurtuz. Baldin laginaren batezbestekoa 191 ml eta 209 ml tartean badago, makinak ondo funtzionatzen duela onartzen da; bestela $\mu \neq 200$ ml onartzen da.*

a) Kalkula ezazu I motako errorea egiteko probabilitatea, baldin $\mu = 200$ ml.

b) Kalkula ezazu II motako errorea egiteko probabilitatea, baldin $\mu = 215$ ml.

(a)

$$\begin{aligned}
\alpha &= P(H_0 \textit{ baztertu} \mid H_0 \textit{ egia}) \Leftrightarrow\\
\Leftrightarrow 1-\alpha &= P(191 \leq \overline{X} \leq 209 \mid \mu = 200) = [\overline{X} : \mathcal{N}(200, 5)] =\\
&= P(-9/5 \leq Z \leq 9/5) = P(Z \leq 9/5) - P(Z \leq -9/5)
\end{aligned}$$

(b)

$$\begin{aligned}
\beta &= P(H_0 \textit{ onartu} \mid H_1 \textit{ egia}) = P(191 \leq \overline{X} \leq 209 \mid \mu = 215) =\\
&= P(6/5 \leq Z \leq 24/5) = P(Z \leq 24/5) - P(Z \leq 6/5).
\end{aligned}$$

Hasteko, behar izango ditugun konstante batzuk definituko ditugu:

```
> mu0 <- 200
> mu1 <- 215
> sigma <- 15
> n <- 9
> sd <- sigma/sqrt(n)
> behe <- 191
> goi <- 209
```

Orain, α (I motako errorea) eta konfiantza-maila kalkulatuko ditugu:

```
> #Konfiantza-maila eta alpha (I motako errorea)
> z1 <- (behe-mu0)/sd
> z2 <- (goi-mu0)/sd
> km <- pnorm(z2)-pnorm(z1)
> km

[1] 0.9281394

> alfa <- 1-km
> alfa

[1] 0.07186064
```

Azkenik, β (II motako errorea) eta $1-\beta$ (ahalmena) kalkulatzeko:

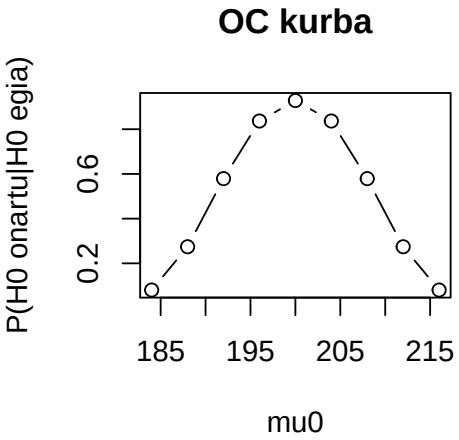
```
> #1-beta eta beta (II motako errorea)
> z1 <- (behe-mu1)/sd
> z2 <- (goi-mu1)/sd
> beta <- pnorm(z2)-pnorm(z1)
> beta

[1] 0.1150689

> ahalmena <- 1-beta
> ahalmena

[1] 0.8849311
```

29. adibidea. *H_0 ez baztertzeko probabilitatea H_0 egiazkoa denean OY ardatzean kokatzen badugu eta OX ardatzean dagozkien μ_0 balio posible batzuk kokatzen baditugu, bikote guztiak kurba baten bidez lotuz,* ***eragiketa-kurba karakteristikoa*** *izenekoa lortzen dugu:* **OC kurba**. *Kurba horiek maiz erabiltzen dira aplikazio industrialetan hipotesi-kontrastearen doitasuna bisualki aztertzeko.*

Aurreko adibidea kontuan hartuta, kalkula ezazu H_0 ez baztertzeko probabilitatea honako μ_0 balio hauentzat: 184, 188, 192, 196, 200, 204, 208, 212 eta 216. Egin ezazu OC kurba.

$$1-\alpha = P(191 \leq \overline{X} \leq 209 \mid \mu = \mu_0) = [\overline{X} : \mathcal{N}(\mu_0, s/\sqrt{n})] = P(\tfrac{191-\mu_0}{s/\sqrt{n}} \leq Z \leq \tfrac{209-\mu_0}{s/\sqrt{n}})$$

```
> muset <- seq(184, 216, 4) #mu0 balioak
> sigma <- 15 #Populazioaren des.est.
> n <- 9  #Laginaren tamaina
> bb.sigma <- sigma/sqrt(n) #Bb-aren des.est.
> behe <- 191
> goi <- 209
> z1 <- (behe-muset)/bb.sigma
> z2 <- (goi-muset)/bb.sigma
> km <- pnorm(z2)-pnorm(z1)
> #Irudikatu OC kurba
> plot(muset,km,xlab="mu0", main="OC kurba",
+ ylab="P(H0 onartu|H0 egia)", type="b")
```

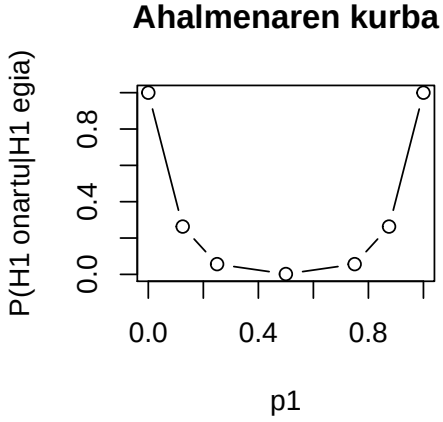


30. adibidea. *Bola zuri eta gorrien kopuru handi bat gordetzen duen kutxa daukagu. Bola zurien proportzioa p da, eta gorrien proportzioa $1-p$. Honako kontraste hau egin nahi dugu: $H_0 : \ p = 1/2$. Horretarako, n bola ateratzen ditugu zoriz, eta honako irizpide hau erabiltzen dugu: H_0 errefusatuko dugu, soilik 0 edo n bola zuri agertzen badira. $n = 10$ suposatuz, kalkula ezazu $1-\beta(p)$,* **testaren ahalmen-funtzioa**. *Kalkula itzazu $1-\beta(0), 1-\beta(1), 1-\beta(1/2), 1-\beta(3/4), 1-\beta(1/4), 1-\beta(1/8)$ eta $1-\beta(7/8)$. Egin ezazu p balio alternatibo ezberdinetarako ahalmen-funtzioaren adierazpen grafikoa.*

$$1-\beta = 1 - P(H_0 \text{ onartu} \mid H_1 \text{ egiazkoa}) = P((X=0) \cup (X=n) \mid p = p_1) = (1-p_1)^{10} + p_1^{10}.$$

Izan ere, H_1-en menpe, $X : Bin(10, p_1)$ baita.

```
> n <- 10
> p1set <- c(0, 1/8, 1/4, 1/2, 3/4, 7/8, 1)
> ahalmena <- p1set^n+(1-p1set)^n
> plot(p1set,ahalmena,xlab="p1",
+ ylab="P(H1 onartu|H1 egia)",
+ type="b",main="Ahalmenaren kurba")
```



Test ezberdinen ahalmenaren kalkuluaren inguruan gehiago sakondu nahi duenak `pwr` paketera jo dezake.

4.4.2. Laginaren tamaina

31. adibidea. *Edari-makina baten 36 zerbitzutako zorizko lagin batean, 21,9 dl da batez besteko edukia. Demagun populazioa 1,42 desbideratze estandarra duen banaketa normalari darraiola. $\mu \geq 22,2$ dl hipotesi nulua $\mu < 22,2$ dl hipotesi alternatiboaren aurka kontrastea egiteko % 5eko esangura-mailarekin, zein izan behar da laginaren tamaina testaren ahalmena 0,90 izateko, batez-besteko alternatiboa 21,3 dl izanik?*

Laginaren tamaina kalkulatzen duen funtzio bat programatzeko aukera ere badago. $n = \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu_1}\right)^2$ adierazpenari jarraituz:

```
> lagin.tamaina.test <- function(alpha, ahalmena, mu0, mu1, sigma){
+ dif <- mu0-mu1
+ n <- (sigma*(qnorm(1-alpha)+qnorm(ahalmena))/dif)^2
+ return(ceiling(n))
+ }
```

Funtzio honek `alpha` esangura-maila, `ahalmena` ahalmenaren maila, `mu0` batezbesteko nulua, `mu1` batezbesteko alternatiboa eta `sigma` populazioaren desbideratze estandarra jasotzen ditu parametro bezala. Irteeran, errore horiek lortzeko beharrezko laginaren tamaina minimoa itzuliko digu, laginaren tamainaren formulatik lortutako n balioa goruntz biribilduz.

Adibidea ebazteko funtzioa ebaluatzen dugu, beharrezko parametro-balioekin:

```
> lagin.tamaina.test(alpha=0.05, ahalmena=0.90, mu0=22.2, mu1=21.3, sigma=1.42)

[1] 22
```

Beraz, laginaren tamaina gutxienez 22koa izan beharko da.

5. *R* praktika

Hipotesi-kontraste ez-parametrikoak

Helburua

Praktika honetan aztertuko dugu nola ebatzi hipotesi-kontraste ez-parametriko batzuk *R*ko kodea erabiliz.

5.1. Doikuntza-egokitasunerako kontrasteak

5.1.1. Pearsonen khi karratu kontrasteak

32. adibidea. *Har ezazu berriro ere 10. adibidea, non Mendelen legeak esperimentalki egiaztatu nahi baikenituen. Horretarako, 500 landare gurutzatzen genituen, eta, teoriaren arabera, lore gorri, arrosa, hori eta zuriko landare kopuruek, hurrenez hurren, proportzionalak izan beharko lukete 8, 12, 10 eta 20 zenbakiekin. Lorturiko datuak 70, 126, 96 eta 208 izan ziren, hurrenez hurren.*

Honako hipotesi-kontraste hau definitzen dugu:

$$H_0 : (p_1, p_2, p_3, p_4) = (0,16,\ 0,24,\ 0,20,\ 0,40)$$

*R*n doikuntza-egokitasunerako khi karratu test bat egiteko, `chisq.test` funtzioa erabiliko dugu, bertan behatutako maiztasunak (`x`) eta hipotesi nuluko probabilitateak (`lore.prob`) zehaztuz:

```
> lore.beh <- c(70, 126, 96, 208)
> lore.prob<- c(8/50, 12/50, 10/50, 20/50)
> chisq.test(x=lore.beh, p=lore.prob)
```

```
Chi-squared test for given probabilities

data:  lore.beh
X-squared = 2.03, df = 3, p-value = 0.5662
```

$\chi^2_p = 2,03$ estatistikoa da eta $p - balioa = 0,5662$ da. Interpretazioa: $p > \alpha = 0,05$ denez, H_0 onartzen da.

Era berean, edozein banaketarekin honako modu honetan egin dezakegu khi karratu testa:

33. adibidea. *Honako taula honetan, urmael batetik ateratako 100 laginen organismo kopurua adierazten da. Froga ezazu datu horiek Poisson banaketa batetik aterata daudela.*

Organismo kopurua	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Lagin kopurua	*15*	*30*	*25*	*20*	*5*	*4*	*1*	*0*

Lagin horri lotutako datuak **praktikadatuak.xls** fitxategiko **organismo** orrialdean daude, inportatu nahi izanez gero. Hipotesi nulua $H_0 : X = \mathcal{P}(\hat{\lambda})$ da.

$\hat{\lambda} = 1,86$ zenbatetsi ostean, kalkula ditzagun itxarondako probabilitateak, eta meta ditzagun 5 baino txikiagoak direnak. Horretarako, 2. R praktikan azaltzen diren funtzio batzuk erabiliko ditugu:

```
> n <- 100
> k <- 7
> #Behatutako maiztasunak
> org.beh <- c(15, 30, 25, 20, 5, 4, 1, 0)
> #Lagina
> org <- rep(0:k, org.beh)
> #Estimatu lambda
> lambda <- mean(org)
> #Kalkulatu P(X=x) balioak x=0,1,2,3,4,5,6 balioentzat
> prop0 <- dpois(0:(k-1),lambda)
> #Kalkulatu 1-P(X<=6)=P(X>=7)
> prop0[8] <- 1-ppois(k-1,lambda)
> #Itxarondako maiztasunak
> org.itx <- prop0*n
> #Itxarondako maiztasunak 5 baino txikiagoak direnez, metatu
> org.itx.met <- c(org.itx[1:4], sum(org.itx[5:8]))
> org.beh.met <- c(org.beh[1:4], sum(org.beh[5:8]))
> #Metatu ondoren, itxarondako probabilitate berriak
> prop0.met <- org.itx.met/n
```

Azkenik, egin dezagun khi karratuaren testa:

```
> chisq.test(x=org.beh.met, p=prop0.met)


Chi-squared test for given probabilities

data:  org.beh.met
X-squared = 1.1404, df = 4, p-value = 0.8878
```

$\chi^2_p = 1,1404$ estatistikoa da eta $p - balioa = 0,8878$ da. Interpretazioa: $p > \alpha = 0,05$ denez, H_0 onartzen da. Parametro bat zenbatetsi denez, askatasun-graduak $df = 3$ izan beharko lirateke, eta $p - balioa =$`1-pchisq(1.1404,3)`$= 0,7673$

5.1.2. Kolmogorov-Smirnov kontrastea

Doikuntza-egokitasunerako beste test bat Kolmogorov-Smirnov testa da. Kontraste honekin, ez da beharrezkoa klaseen bilketa egitea esperotako maiztasunak txikiak direnean. Test hori *R*n egiteko, `ks.test` funtzioa erabiliko dugu lagina (`x`), kontrastatu nahi dugun banaketa (`dist`) eta bere parametroak zehaztuz.

Egin dezagun 33. adibidea, metodo hau erabiliz.

```
> org <- rep(0:k, org.beh)
> ks.test(x=org, dist=pois, lambda)


Two-sample Kolmogorov-Smirnov test

data:  org and lambda
D = 0.55, p-value = 0.9255
alternative hypothesis: two-sided
```

Interpretazioa: $p = 0,9255 > \alpha = 0,05$ denez, H_0 onartzen da, hots, $X = \mathcal{P}(1,86)$.

5.1.3. Normaltasunerako kontrasteak

Banaketa normala da banaketa guztietatik garrantzitsuenetakoa. Hori horrela, bai teorian eta bai *R*n normaltasuna aztertzeko proba ugari existitzen dira. Guk bi aurkeztuko ditugu:

Shapiro-Wilk proba (n <50 denean)

34. adibidea. *Kontrasta ezazu honako datu hauek banaketa normaletik ateratakoak diren hipotesia: 20, 22, 24, 30, 31, 32, 38.*

```
> x <- c(20, 22, 24, 30, 31, 32, 38)
> shapiro.test(x)


Shapiro-Wilk normality test

data:  x
W = 0.9478, p-value = 0.7096
```

Kontraste honen estatistikoa $0,1858$ da, eta p-balioa$= 0,6577$. Beraz, ezin da baztertu aldagaiaren normaltasuna.

Lilliefors (Kolmogorov-Smirnov) testa ($n \geq 50$ denean)

Har dezagun berriro ere 12. adibidea.

35. adibidea. *Birus baten latentzia-aldia ikertzeko, 90 txitari inokulatu zitzaien birusa, gaixotasunaren lehenengo sintomak agertu arte pasatutako egun kopurua aztertuz. Analiza dezagun, kodea erabiliz, birusen latentzia-aldiaren normaltasuna.*

Gogoratu datuak **praktikadatuak.xls** fitxategiko **birusa** orrian ditugula eskuragai.

```
> library(readxl)
> birusa <- read_excel("praktikadatuak.xls", sheet='birusa')
> library(nortest)
> lillie.test(birusa$birusa)


Lilliefors (Kolmogorov-Smirnov) normality test

data:  birusa$birusa
D = 0.12728, p-value = 0.00104
```

Kontraste horren estatistikoa $0,12728$ da eta p-balioa$= 0,00104$. Beraz, ezin da baztertu aldagaiaren normaltasuna.

5.1.4. Normaltasunerako Box-Cox-en transformazioa

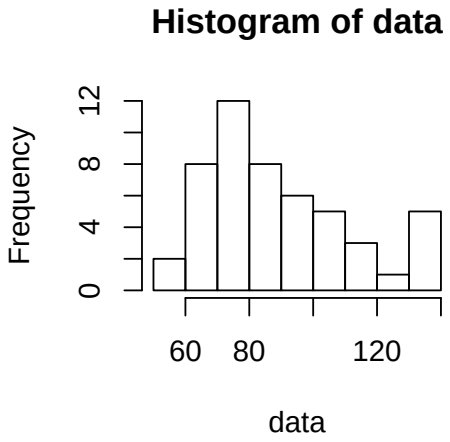
36. adibidea. *Demagun 1974. urteko per capita errenta aztertu nahi dugula. Datuak honako taula honetan adierazten dira:*

x_i	55	65	75	85	95	105	115	125	135
f_i	2	8	12	8	6	5	3	1	5

Banaketa normala jarraitzen duela ondoriozta al daiteke? Ezezko kasuan, bilatu normaltasuna lortzea ahalbidetuko duen transformazio egoki bat.

Lehenik, datuak eskuz sartu edo inportatu ostean, (**praktikadatuak.xls/errenta**) histograma eraikiko dugu, datuen banaketaren itxura aztertzeko:

```
> xi <- seq(55, 135, 10)
> fi <- c(2, 8, 12, 8, 6, 5, 3, 1, 5)
> data <- rep(xi, fi)
> hist(data)
```



Histograma ikusita, nabarmena da normaltasunik ez dagoela. Gainera, Lilliefors (Kolmogorov-Smirnov) testa erabil dezakegu ($n \geq 50$) hau beste modu batera ziurtatzeko:

```
> lillie.test(data)


Lilliefors (Kolmogorov-Smirnov) normality test

data:  data
D = 0.17313, p-value = 0.0006917
```

Ikusten dugun moduan, p-balioa oso txikia da; beraz, normaltasuna errefusatzen dugu.

Honelako kasuetan, ohikoa da, datuei transformazioak aplikatzea, normaltasuna lortzeko helburuarekin; eta transformazio mota ohikoenak Box-Cox-enak dira:

$$X(\lambda) = \frac{X^\lambda - 1}{\lambda}.$$

Beraz, aurreko datuei Box-Cox-en transformazioa aplikatuko diegu, eta, Lilliefors (Kolmogorov-Smirnov) kontrastearen bidez, normaltasuna aztertuko dugu. Hasteko, Box-Cox-en transformazioa definituko duen funtzioa definituko dugu:

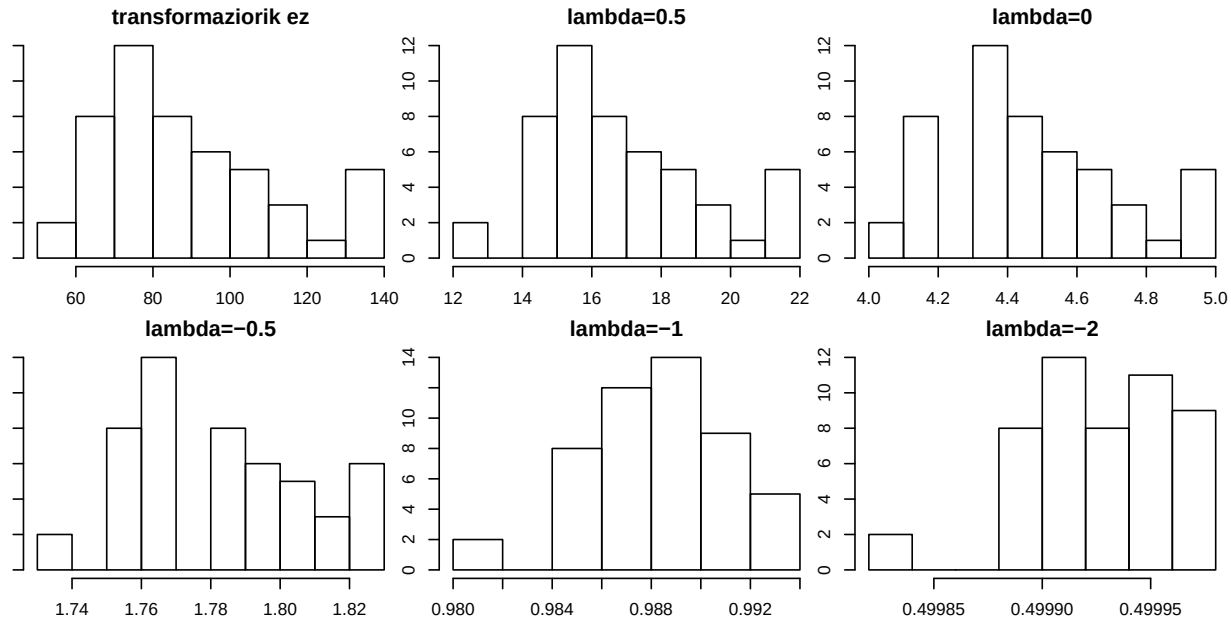
```
> box.cox <- function(data,l){
+   if(l == 0) log(data)
+   else (data^l-1)/l
+ }
```

Orain, histograma eskuinalderantz alboratua denez, $\lambda < 1$ parametro batzuk aukeratuko ditugu, eta definitu dugun funtzioa erabiliko dugu transformazioak gauzatzeko; zehazki, $\lambda \in \{0,5,\ 0,\ -0,5,\ -1,\ -2\}$:

```
> data.p05 <- box.cox(data, 0.5)
> data.0 <- box.cox(data, 0)
> data.m05 <- box.cox(data, -0.5)
> data.m1 <- box.cox(data, -1)
> data.m2 <- box.cox(data, -2)
```

Tranformatutako datu horien histogramak irudikatzen baditugu, hasierako histogramarekin batera:

```
> par(mfrow=c(2, 3), mar=c(2, 1, 2, 1))
> hist(data, main="transformaziorik ez") #transformazio gabe
> hist(data.p05, main="lambda=0.5") #lambda=0.5
> hist(data.0, main="lambda=0") #lamda=0
> hist(data.m05, main="lambda=-0.5") #lambda=-0.5
> hist(data.m1, main="lambda=-1") #lambda=-1
> hist(data.m2, main="lambda=-2") #lambda=-2
```



Azkenik, Lilliefors (Kolmogorov-Smirnov) kontrastearen p-balioak kalkulatuko ditugu transformazio bakoitzetik lortutako laginetarako:

$$\begin{cases} H_0: \ X(\lambda) \approx \mathcal{N}(\overline{x}_{X(\lambda)}, s_{X(\lambda)}) \\ H_1: \ X(\lambda) \not\approx \mathcal{N}(\overline{x}_{X(\lambda)}, s_{X(\lambda)}) \end{cases}$$

```
> lillie.test(data.p05)
> lillie.test(data.0)
> lillie.test(data.m05)
> lillie.test(data.m1)
> lillie.test(data.m2)
```

Honako hauek dira emaitzak:

$$\begin{aligned} X(0,5) = 2(\sqrt{X} - 1) &\Rightarrow p = 0,001276 \\ X(0) = \ln X &\Rightarrow p = 0,002956 \\ X(-0,5) = -2(1/\sqrt{X} - 1) &\Rightarrow p = 0,008234 \\ X(-1,0) = -(1/X - 1) &\Rightarrow p = 0,0257 \\ X(-2,0) = -0.5(1/X^2 - 1) &\Rightarrow p = 0,01458 \end{aligned}$$

$\alpha = 0,01$ hartzen badugu, $\lambda = -1,0$ eta $\lambda = -2,0$ kasuetan ezin da baztertu normaltasuna. $p-$*balio* handiena $\lambda = -1.0$ Box-Cox-en transformazioari dagokio, grafikoki itxura egokiena daukana, hain zuzen.

5.2. Independentzia- eta homogeneotasun-probak

Har dezagun berriro ere 13. adibidea, honako datu hauetatik abiatuta gaixotasuna eta azidotasun-maila desberdinen arteko independentzia edo menpekotasuna aztertzea proposatzen ziguna:

X / Y	Aklorhidria	Hipoklorhidria	Normal	Hiperklorhidria
Ultzera gastrikoa	3	7	35	9
Minbizia	22	2	6	0

*R*n independentzia- eta homogeneotasun-probak `chisq.test` funtzioaren bidez egin ditzakegu:

```
> kont.taula <- matrix(c(3, 7, 35, 9, 22, 2, 6, 0), nrow=2, ncol=4, byrow=TRUE)
> chisq.test(kont.taula)

Warning in chisq.test(kont.taula):  Chi-squared approximation may be incorrect


Pearson's Chi-squared test

data:  kont.taula
X-squared = 43.417, df = 3, p-value = 2.007e-09
```

Kontuan izan funtzioak mezu bat jaulki digula, itxarondako maiztasun batzuk 5 baino txikiagoak direla izanik arrazoia. Hauek bistaratzeko:

```
> chisq.test(kont.taula)$expected

Warning in chisq.test(kont.taula):  Chi-squared approximation may be incorrect

          [,1]     [,2]     [,3]     [,4]
[1,] 16.071429 5.785714 26.35714 5.785714
[2,]  8.928571 3.214286 14.64286 3.214286
```

Beraz, 1. eta 2. zutabeak elkartu beharko ditugu, eta baita 3. eta 4. zutabeak ere:

```
> kont.taula.met <- matrix(nrow=2, ncol=2)
> kont.taula.met[,1] <- kont.taula[,1] + kont.taula[,2]
> kont.taula.met[,2] <- kont.taula[,3] + kont.taula[,4]
> kont.taula.met

     [,1] [,2]
[1,]   10   44
[2,]   24    6
```

Orain, independentzia-proba gauzatu behar dugu; baina 2x2ko taula bat izanik, Yates-en zuzenketa aplikatu behar dugun aztertu behar dugu. Kodea erabiliz, funtzio bat eraikiko dugu, 2x2ko taula bakoitzerako test egokia gauzatzen duena:

```
> chisq.test.22 <- function(kont.taula){

+   if(dim(kont.taula)[1]!=2 | dim(kont.taula)[2]!=2){
+     stop('Funtzio hau soilik 2x2-ko taulentzat da')
+   } else {
+     if(abs(det(kont.taula))>sum(kont.taula)/2){
+       #Yates-en zuzenketa beharrezkoa da
+       chisq.test(kont.taula, correct=TRUE)
+     }else{
+       #Yates-en zuzenketa ez da beharrezkoa
+       chisq.test(kont.taula, correct=FALSE)
+     }
+   }
+ }
```

Ikusten dugunez, funtzioak 2x2 dimentsioko kontingentzia-taula bat hartzen du sarrera-parametro gisa. Ondoren, Yatesen zuzenketa aplikatu behar duen aztertzen du, kasu bakoitzean dagokion kontrastea aplikatuz. Kontuan izan `correct=TRUE` aginduak, `chisq.test` funtzioan, Yatesen zuzenketaren aplikazioa ziurtatzen duela.

Gure kontingentzia-taulari funtzio hau aplikatzen badiogu:

```
> chisq.test.22(kont.taula.met)


Pearson's Chi-squared test with Yates' continuity correction

data:  kont.taula
X-squared = 27.76, df = 1, p-value = 1.374e-07
```

Ikusten dugunez, kasu honetan, Yatesen zuzenketa aplikatu da. Estatistikoak $27,76$ balio du eta p balioak 1,374e-07; beraz, H_0 errefusatu egingo dugu.

37. adibidea. *Azkenik, har ezazu berriro ere 22. adibidea, eta konpara ezazu, metodo ez-parametrikoak erabiliz, bi marka ezberdinetako artikuluen salmenta/arrakasta-maila.*

Bi populazio binomial ezberdinen arrakasta-probabilitateak konparatzeko, `prop.test` funtzioa ere erabili daiteke modu honetan:

```
> lagin.tamainak <- c(200, 150)
> arrakastak <- c(42, 18)
> prop.test(arrakastak, lagin.tamainak, conf.level=0.94)


2-sample test for equality of proportions with continuity correction

data:  arrakastak out of lagin.tamainak
X-squared = 4.2748, df = 1, p-value = 0.03868
alternative hypothesis: two.sided
94 percent confidence interval:
 0.01051492 0.16948508
sample estimates:
prop 1 prop 2
  0.21   0.12
```

Kasu honetan, estatistikoaren balioa $X_p^2 = 4,2748$ da, eta p-balioa=0,03868; beraz, % 5eko esangura-mailaz, H_0 errefusatzen dugu bi proportzioen ezberdintasuna onartuz.

Kontraste honek ere, azken finean, khi karratu test bat egiten du. Kontuan izan funtzio honek defektuz Yatesen zuzenketa aplikatzen duela behar denean, baina hori ekidin dezakegu `correct=FALSE` aukera gehituz.

5.3. Zorizkotasun-kontrastea

38. adibidea. *Inkesta batean 60 pertsona hautatu dira, gazteak (G) eta helduak (H) honela ordenatuta agertuz. Ondoriozta daiteke zoriz hautatu direla?:*

H G H H H G G G H H G H H H H G G G G H G H G G H H H G G H

G G G G H H G H H G G G G H H G H H H H G G H H G H G G G H H

Adibide honetan, honako hau da kontrastatu nahi dugun hipotesi nulua:
H_0: gazteak eta helduak zoriz hautatuak izan dira.
Kontuan izan kasu honetan, eskuartean darabilgun aldagaia kualitatibo dikotomikoa dela.

Teorian oinarrituz, honela programa dezakegu zorizkotasun-proba:

```
> #lagina: faktore moduko bektore bat
>
> zorizkotasun.proba <- function(lagina){
+ n <- length(lagina) #Laginaren tamaina
+ n1 <- length(which(lagina==levels(lagina)[1])) #1. taldeko elementuen kop.
```

```
+ n2 <- length(which(lagina==levels(lagina)[2])) #2. taldeko elementuen kop.
+ #Ondoz ondoko elementuen kenketak egin: aldaketak[i]=lagina[i+1]-lagina[i]
+ aldaketak <- diff(as.numeric(lagina))
+ #Boladak kalkulatu
+ R <- 1 + length(which(aldaketak!=0))
+ #Estatistikoa kalkulatu
+ ER <- 2*n1*n2/n+1
+ VarR <- 2*n1*n2*(2*n1*n2-n)/(n*n*(n-1))
+ #p-balioa kalkulatu
+ zp <- (R-ER)/sqrt(VarR)
+ p <- 2*(1-pnorm(abs(zp)))
+ return(list(boladak=R, estatistikoa=zp, pbalioa=p))
+ }
```

Funtzio honek lagina hartzen du sarrera-parametro gisa, eta irteeran bolada kopurua eta zorizkotasun-testaren estatistikoa eta p-balioa gordetzen dituen zerrenda bat itzultzen digu.

Adibidea ebazteko, sartu edo inportatu datuak **praktikadatuak.xls** fitxategiko **gh** orrialdetik, eta, ondoren, aplikatu eraikitako funtzioa:

```
> zorizkotasun.proba(gazte.helduak)

$boladak
[1] 29

$estatistikoa
[1] -0.5127295

$pbalioa
[1] 0.6081406
```

Horrela, bolada kopurua $\hat{R} = 29$, estatistikoa $z_p = -0,5127295$ eta p-balioa $p = 0,6081406 > \alpha$; beraz, ezin dugu hipotesi nulua (lagina zoriz aukeratu dela) errefusatu.

39. adibidea. *Kalitate-ikuskatzaile batek, langileek egiten duten piezen kopurua ikertzeko asmoz, 50 egunean zehar langile batek egindako kopurua jaso du (10 unitateka neurtua), eta honako emaitza hauek lortu ditu:*

100 110 80 75 130 95 105 125 140 85 115 120 150 60 77,5 92 112

83 136 65 72,5 89 160 90 114 155 55 124 92,5 50 115 120 150 60

77,5 92 112 83 136 65 72,5 89 160 90 114 155 55 124 92,5 50

Ondoriozta daiteke zoriz aukeratu dela lagina?

Adibide honetan, aldagaia kuantitatiboa denez, esate baterako, medianarekiko egin dugu kontrastea. H_0: zorizkotasuna.

```
> #Datuak inportatu
> library(readxl)
> piezak <- read_excel("praktikadatuak.xls", sheet="pieza")
> #Konparatu lagina medianarekiko eta bihurtu faktore:
> lagina <- as.factor(piezak$pieza.akastunak>median(piezak$pieza.akastunak))
> #Egin zorizkotasun-proba
> zorizkotasun.proba(lagina)

$boladak
[1] 28

$estatistikoa
[1] 0.5715476

$pbalioa
[1] 0.5676285
```

Horrela, bolada kopurua 28 da, estatistikoa $z_p = 0,5715476$ eta p-balioa $p = 0,5676285 > \alpha$; beraz, ezin dugu hipotesi nulua errefusatu.

Aldagai jarraitua denean, bolada-testa egiteko `runs.test` komandoa ere erabil daiteke, behin `lawstat` paketea instalatuta:

```
> library(lawstat)
> runs.test(piezak$pieza.akastunak)


Runs Test - Two sided

data:  piezak$pieza.akastunak
Standardized Runs Statistic = 0.57155, p-value = 0.5676
```

Ikusten dugun moduan, emaitza berdinak lortu ditugu.

5.4. Populazioak konparatzeko kontraste ez-parametrikoak

5.4.1. Bi lagin askeren konparaketa

40. adibidea. *Klase batean adimen-proba bat egin da, (1) ikasle berrien eta (2) errepikatzaileen artean emaitzak berdin banatzen diren ala ez aztertzeko asmoz. 30 ikasleren notak aztertu dira,*

puntuazio maximoa 100 dela. Zer ondoriozta daiteke?

B	92	12	83	36	65	72,5	89	60	90	14	55	55	24	92,5	50
E	100	10	80	75	30	95	5	25	40	85	15	20	50	60	77,5

Adibide hau ebazteko, honako hipotesi nulu hau formulatuko dugu.

H_0: ikasle berrien eta errepikatzaileen batez besteko puntuazioak berdinak dira.

Demagun, hipotesi hau kontrastatzeko, **Mann-Whitney testa** egin nahi dugula. Hasteko, datuak eskuz sartu edo inportatuko ditugu:

```
> #Datuak inportatu
> library(readxl)
> adimena <- read_excel("praktikadatuak.xls", sheet="adimen")
```

Ondoren, `wilcox.test` funtzioa erabiliko dugu, modu honetan:

```
> wilcox.test(Adimena ~ Ikaslea, alternative='two.sided', exact=FALSE,
+ correct=FALSE, data=adimena)


Wilcoxon rank sum test

data:  Adimena by Ikaslea
W = 128, p-value = 0.5201
alternative hypothesis: true location shift is not equal to 0
```

Horrela, p-balioa $p = 0,5201 > \alpha$; beraz, ezin dugu berdintasuna errefusatu.

5.4.2. Bi lagin aske baino gehiagoren konparaketa

41. adibidea. *Pedagogo batek konparatu nahi ditu irakasteko honako hiru metodo hauek: (1) online, (2) erdipresentziazkoa eta (3) presentziazkoa. Horretarako, ikasturte batean zehar hiru metodoekin irakasten den ikasgai bat hautatu du; metodo bakoitzarekin zorizko lagin bakun bat aukeratu du, eta amaierako notak honako hauek izan dira:*

Online	78	80	65	57	89				
Erdipresentziala	74	88	82	93	55	70			
Presentziala	68	83	50	91	84	77	94	81	92

Zer ondoriozta daiteke % 90eko konfiantza-mailarekin?

Adibidea hau ebazteko, honako hipotesi nulu hau definituko dugu:

H_0: hiru metodoekin berdinak dira batez besteko notak.

Hori ebazteko, demagun **Kruskal-Wallis testa** egin nahi dugula. Hori, Rko `kruskal.test` funtzioarekin egin dezakegu.

```
> metodo <- read_excel("praktikadatuak.xls", sheet="metodo")
> metodo$metodoa <- factor(metodo$metodoa)
> kruskal.test(puntuazioa ~ metodoa, data=metodo)


Kruskal-Wallis rank sum test

data:  puntuazioa by metodoa
Kruskal-Wallis chi-squared = 1.1451, df = 2, p-value = 0.5641
```

Estatistikoa $K = 1,1451$ eta p-balioa $p = 0,5641 > \alpha$; beraz, ezin dugu berdintasuna errefusatu.

Ondoren, binakako konparaketak egin nahiko bagenitu, konparaketa anitzak egiteko zuzenketa mota ugari aplika ditzakegu Rn `pairwise.wilcox.test` funtzioa erabiliz. Holmen metodoa da test nahiko ohiko bat, datuen banaketaren inguruan suposiziorik egiten ez duena:

```
> pairwise.wilcox.test(x=metodo$puntuazioa, g=metodo$metodoa, data=Datos,
+                      p.adjust.method="holm")


Pairwise comparisons using Wilcoxon rank sum test

data:  metodo$puntuazioa and metodo$metodoa

       erdipre online
online 1.00    -
presen 1.00    0.89

P value adjustment method: holm
```

Ikusten dugunez, p-balio guztiak oso handiak dira, espero bezala; beraz, ez dago ezberdintasun adierazgarririk bikotekoen artean.

5.4.3. Binakako datuen bi laginen konparaketa

42. adibidea. *Enpresa baten arabera, haiek ekoitzitako iragazkia baliagarria da erregai-kontsumoa murrizteko, autoen karburagailuaren hasieran kokatuta. Informazioa kontrastatzeko, 30 auto hautatu*

ziren, eta bakoitzaren kontsumoa neurtu zen, iragazki gabe (ez) eta iragazkiarekin (bai). Behatutako datuak (l/100 km-tan neurtuak) honako taula honetan agertzen direnak badira, zer ondoriozta daiteke?

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Ez	6,8	7,0	7,2	9,0	9,1	10,0	9,2	8,5	8,0	8,9	9,3	10,1	6,5	7,8	6,9
Bai	6,4	6,5	7,3	8,8	8,8	9,0	9,4	8,1	7,5	8,9	9,2	10,5	6,4	8,0	6,5

	16	17	18	19	20	21	22	23	24	25	26	27	38	29	30
Ez	7,4	8,7	9,3	8,2	8,0	7,0	9,3	7,0	6,9	10,0	9,4	8,0	7,8	9,0	9,5
Bai	7,1	9,0	9,7	8,0	7,4	6,7	9,9	6,6	7,0	9,0	8,6	7,1	8,8	8,3	8,2

Honako hipotesi nulu hau planteatuko dugu:

H_0: iragazkirik gabeko batez besteko kontsumoa ez da iragazkiarekikoa baino handiagoa.

Datu-basea inportatu edo sortu ondoren, `wilcox.test` funtzioa erabil dezakegu `paired=TRUE` zehaztuz:

```
> ezbai <- read_excel("praktikadatuak.xls", sheet="ezbai")
> wilcox.test(ezbai$ez, ezbai$bai, alternative='two.sided',
+             correct=FALSE, exact=FALSE, paired=TRUE)


Wilcoxon signed rank test

data:  ezbai$ez and ezbai$bai
V = 325.5, p-value = 0.01946
alternative hypothesis: true location shift is not equal to 0
```

Horrela, heinen batura $T = 325,5$ eta $p = 0,01946 < \alpha$; beraz, ezin dugu onartu hipotesi nulua, hots, ezin dugu baieztapena errefusatu.

5.4.4. Binakako datuen bi lagin baino gehiagoren konparaketa

43. adibidea. *Demagun hiru katalizatzaileen efizientziak konparatu nahi ditugula. Horretarako, 8 egunean zehar proba batzuk egin ditugu, honako efizientzia hauek lortuz:*

	1.*eguna*	2.*eguna*	3.*eguna*	4.*eguna*	5.*eguna*	6.*eguna*	7.*eguna*	8.*eguna*
A	84,5	102,8	99,1	80,2	92,2	100,3	81,0	101,3
B	78,4	79,1	78,0	76,0	65,2	82,5	76,3	60,2
C	63,1	79,9	67,8	52,9	60,2	76,2	57,3	76,2

Suposatzen badugu egun bakoitzeko baldintzak bereziak direla, eta, beraz, egun bakoitzeko datuak ezin direla elkarren artean konparatu (hau da, egun bakoitzeko baldintzak bereziak dira), kontrastatu ea hiru katalizatzaileen batezbesteko efizientziak berdinak direnentz.

Adibidea ebazteko, honako hipotesi nulu hau planteatuko dugu: H_0: hiru katalizatzaileen batezbesteko efizientziak berdinak dira. Jo dezagun hipotesi-kontraste parametriko bat gauzatzeko beharrezko baldintzak (adib. normaltasuna) ez direla betetzen. Kasu honetan, Friedman testa egiteko `friedman.test` funtzioa erabiliko dugu, era honetan:

```
> #Datuak inportatu
> efizientzia <- read_excel("praktikadatuak.xls", sheet="efizientzia")
> efizientzia <- as.matrix(efizientzia)
> friedman.test(efizientzia)


Friedman rank sum test

data:  efizientzia
Friedman chi-squared = 13, df = 2, p-value = 0.001503
```

Estatistikoaren balioa 13 denez, eta p-balioa 0,001503, $\alpha = 0,05$ esangura-mailaz, hipotesi nulua errefusatuko dugu.

Azkenik, **scmamp** paketean duzue eskuragai test ez-parametriko sorta zabal bat.

6. *R* praktika

Bariantza-analisia

Helburua

Praktika honen jomuga da ANOVA edo bariantza-analisia R bidez nola egin azaltzea. Faktore bakarra eta bi faktore kontuan hartzen dituzten analisiak egingo ditugu, bi populazio baino gehiagoren batezbestekoak konparatzeko asmoz. Gainera, diferentziak adierazgarriak diren kasuetan, batezbestekoen binakako konparaketak nola egin ere ikusiko dugu.

6.1. Faktore bakarreko bariantza-analisia

44. adibidea. *Har ezazu berriro ere 15. adibidea, eta konparatu hiru herrialdeetako batez besteko kutsadurak, horretarako, herrialde bakoitzeko 15 neurketa-estaziotan PM10 partikulen neurketak erabiliz.*

Beraz, honako hipotesi nulu hau definituko dugu:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

Ohartu kasu honetan faktore bat (*herrialdea*) eta aldagai jarraitu bat (*konzentrazioa*) ditugula. Alabaina, faktore bakarreko bariantza analisia aplikatu ahal izateko, beharrezko baldintzak betetzen diren egiaztatu behar dugu. Zorizkotasuna eta askatasuna laginketa-prozesutik suposatuko ditugu. Aldiz, bariantzen homogeneotasuna eta populazioaren normaltasuna bermatzeko, dagozkien hipotesi-kontrasteak gauzatu behar ditugu:

Hasteko, normaltasunerako Shapiro-Wilk proba egingo dugu, laginak txikiak direlako:

```
> library(readxl)
> kutsadura<- read_excel("praktikadatuak.xls", sheet='kutsadura')
> kutsadura$herrialdea <- factor(kutsadura$herrialdea)
>
> #Normaltasun-testak
> A <- kutsadura[kutsadura$herrialdea=="A",]
> B <- kutsadura[kutsadura$herrialdea=="B",]
> C <- kutsadura[kutsadura$herrialdea=="C",]
> shapiro.test(A$kontzentrazioa)


Shapiro-Wilk normality test

data:  A$kontzentrazioa
W = 0.90834, p-value = 0.1277

> shapiro.test(B$kontzentrazioa)


Shapiro-Wilk normality test

data:  B$kontzentrazioa
W = 0.93886, p-value = 0.3683

> shapiro.test(C$kontzentrazioa)


Shapiro-Wilk normality test

data:  C$kontzentrazioa
W = 0.97285, p-value = 0.8978
```

Ikusten dugunez, hiru p-balioak $\alpha = 0,05$ baino handiagoak dira ($p_A = 0,1277, p_B = 0,3685$, $p_C = 0,8978$); beraz, normaltasuna betetzen denik ezin dugu baztertu. Orain, bariantzen homogeneotasuna aztertzeko Bartlett testa gauzatuko dugu:

```
> #Bariantzen homogeneotasun-testak
> bartlett.test(kutsadura$kontzentrazioa, kutsadura$herrialdea)


Bartlett test of homogeneity of variances

data:  kutsadura$kontzentrazioa and kutsadura$herrialdea
Bartlett's K-squared = 0.18242, df = 2, p-value = 0.9128
```

Bartlett probaren p-balioa= $0,9128$ da. Beraz, ezin da baztertu hipotesi nulua; hots, bariantzen berdintasuna ezin dugu baztertu % 5eko esangura-mailarekin.

Kontuan izan Bartlett testak normaltasuna dagoenean ematen dituela emaitzarik fidagarrienak. Normaltasuna ez bada betetzen, egokiagoa da Levene testa erabiltzea, **car** paketeko `LeveneTest` funtzioan inplementatuta dagoena.

Azkenik, baldintza guztiak betetzen direla ikusita, faktore bakarreko bariantza-analisia gauza dezakegu, `aov` funtzioa erabiliz:

```
> AnovaModel.1 <- aov(kontzentrazioa ~ herrialdea, data=kutsadura)
> summary(AnovaModel.1)

            Df Sum Sq Mean Sq F value Pr(>F)
herrialdea   2   3824  1911.8   458.8 <2e-16 ***
Residuals   42    175     4.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estatistikoa $F = 458,8$ da, eta haren p-balioa $p < 2 \cdot 10^{-16}$. Beraz, batezbestekoen berdintasuna baztertuko dugu.

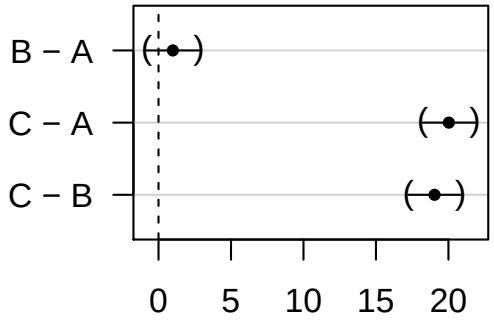
Orain, binakako konparaketak egiteko, metodo ezberdinen artean aukera dezakegu. Tukeyren metodoa erabiltzeko, **multcomp** paketeko `glht` funtzioa erabiliko dugu:

```
> library(multcomp)
> binakakoak.1 <-glht(AnovaModel.1, linfct = mcp(herrialdea = "Tukey"))
```

Binakako diferentzietarako konfiantza-tarteak lor ditzakegu, bai numerikoki eta bai grafiko bidez:

```
> confint(binakakoak.1)


 Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = kontzentrazioa ~ herrialdea, data = kutsadura)

Quantile = 2.4289
95% family-wise confidence level


Linear Hypotheses:
```

```
           Estimate lwr     upr
B - A == 0  0.9880  -0.8225  2.7985
C - A == 0 20.0293  18.2188 21.8399
C - B == 0 19.0413  17.2308 20.8519

> plot(binakakoak.1, main="", xlab="")
```



Azkenik, multzo homogeneoen metodoa erabiltzeko:

```
> cld(binakakoak.1)

  A   B   C
"a" "a" "b"
```

Ikusten dugun moduan, kasu honetan, $C-A$ eta $C-B$ bikoteetarako konfiantza-tarteak positiboak dira; beraz, bikote horien artean diferentziak daudela esan dezakegu. $B-A$ bikoteari dagokionez, 0 tarte barruan dago, eta, ondorioz, ezin dugu bi herrialde horien batez besteko kutsaduraren berdintasuna baztertu. Zehazki, $\mu_A = \mu_B < \mu_C$ ondoriozta daiteke, %95eko konfiantza-mailaz. Multzo homogeneoen kasuan, metodoak bi multzo homogeneo bereizten ditu, A eta B herrialdeak multzo batean egokituz eta C herrialdea beste batean (kutsadura handiagokoa).

6.2. Faktore biko bariantza-analisia interakzioarekin (n > 1)

45. adibidea. *Zirkulazio-istripuetan droga batzuek izaniko eragina ikertu nahi da. Horretarako, gidatzeko simulagailu batean jarri ziren bederatzi emakume eta bederatzi gizon, bakoitza tratamendu baten mende: marihuanaren eraginpean, alkoholaren eraginpean, eta drogarik gabe. Simulagailuak 0 eta 35 puntu artean ematen ditu, eta puntuazio altuenak gidatzeko egoera onenekin erlazionatuta daude. Honako datu hauek lortu ziren:*

Pertsona	*Marihuana*	*Alkohola*	*Droga barik*
E	*19, 18, 25*	*8, 10, 10*	*21, 31, 26*
G	*20, 17, 21*	*18, 7, 16*	*28, 14, 24*

Sexuaren arabera, ondoriozta al daiteke hiru drogen eragina berdina dela? Zer esan daiteke sexuaren eta hartutako drogaren arteko interakzioari buruz?

Kasu honetan, hiru aldagai dauzkagu: bi kualitatibo: sexua (1=emakumezkoa, 2=gizonezkoa) eta tratamendu mota (1=marihuana, 2=alkohola, 3=droga gabekoa) eta kuantitatibo bat, puntuazioa. Beraz, bi faktoreko bariantza-analisia egingo dugu, berriro ere `aov` funtzioa erabiliz eta aldagaien arteko interakzioak `droga*sexua` adierazpenaren bidez kodetuz. Suposatuko dugu behar diren baldintzak betetzen direla.

Hipotesi nulua $H_0 : \mu_{11} = \mu_{12} = \mu_{13} = \mu_{21} = \mu_{22} = \mu_{23}$ da.

```
> drogasexua <- read_excel("praktikadatuak.xls", sheet='drogasexua')
>
> #Aldagai kualitatiboak faktore bihurtu
> drogasexua$sexua <- factor(drogasexua$sexua)
> drogasexua$droga <- factor(drogasexua$droga)
>
> #Bi faktoreko bariantza-analisia
> AnovaModel.2 <- aov(puntuazioa ~ droga*sexua, data=drogasexua)
> summary(AnovaModel.2)

            Df Sum Sq Mean Sq F value  Pr(>F)
droga        2  489.0  244.50  11.170 0.00182 **
sexua        1    0.5    0.50   0.023 0.88238
droga:sexua  2   54.3   27.17   1.241 0.32365
Residuals   12  262.7   21.89
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

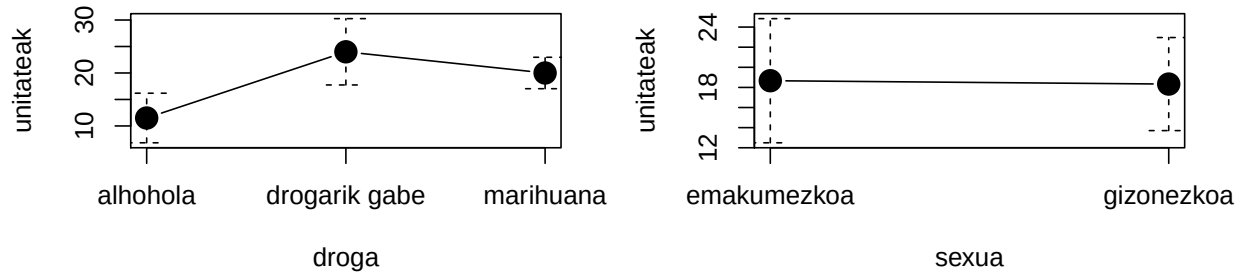
Aurreko taulan, faktore biko bariantza-analisiaren laburpena dugu. Droga mota faktorearekiko estatistikoa $F' = 11,170$ eta p-balioa $p' = 0,0018$ denez, adierazgarria da faktore horren eragina puntuazioetan. Hala ere, sexuak ez du eraginik emaitzetan ($p'' = 0,8824$), eta ez dago interakziorik ($p''' = 0,3236$).

Jarraian, puntuazioen batezbestekoak eta desbideratzeak kalkula ditzakegu:

```
> tapply(drogasexua$puntuazioa, list(droga=drogasexua$droga,
+ sexua=drogasexua$sexua), mean, na.rm=TRUE) # means

               sexua
droga           emakumezkoa gizonezkoa
  alhohola         9.333333   13.66667
  drogarik gabe   26.000000   22.00000
  marihuana       20.666667   19.33333

> tapply(drogasexua$puntuazioa, list(droga=drogasexua$droga,
+  sexua=drogasexua$sexua), sd, na.rm=TRUE) # std. dev.

               sexua
droga           emakumezkoa gizonezkoa
  alhohola         1.154701   5.859465
  drogarik gabe    5.000000   7.211103
  marihuana        3.785939   2.081666
```

Gainera, % 95eko batezbestekoen konfiantza-tarteak grafikoki irudikatzeko eska diezaiokegu, **RcmdrMisc** paketeko `plotMeans` funtzioa erabiliz:

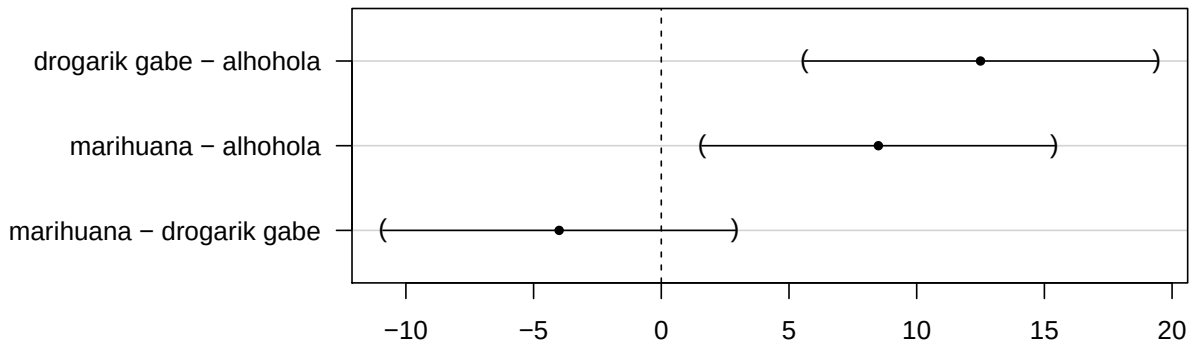
```
> library(RcmdrMisc)
> par(mfrow=c(1,2))
> plotMeans(drogasexua$puntuazioa, drogasexua$droga, main="",
+           error.bars="conf.int", level=0.95, ylab="unitateak", xlab="droga")
> plotMeans(drogasexua$puntuazioa, drogasexua$sexua,  main="",
+           error.bars="conf.int", level=0.95, ylab="unitateak", xlab="sexua")
```



Azkenik, azter ditzagun droga faktorearekiko batez besteko puntuazioen konparaketa anitzak. Horretarako, lehenengo efektua zuen faktore bakarrarekin (*droga*) anova eredua eraikiko dugu:

```
> AnovaModel.3 <- aov(puntuazioa ~ droga, data=drogasexua)
> binakakoak.2 <- glht(AnovaModel.3, linfct = mcp(droga = "Tukey"))
> binakakoak.2


 General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Linear Hypotheses:
                               Estimate
drogarik gabe - alhohola == 0      12.5
marihuana - alhohola == 0           8.5
marihuana - drogarik gabe == 0     -4.0
```

```
> par(mar = c(2,11,4,2) + 0.1)
> plot(binakakoak.2, main="")
```




```
> cld(binakakoak.2)

     alhohola drogarik gabe     marihuana
          "a"           "b"           "b"
```

Tukeyren alderaketa binakatuen arabera, % 5eko esangura-mailarekin bi bikoteren artean diferentzia adierazgarriak ikusten dira: $\mu_{drogarikgabe} > \mu_{alkohola}$ eta $\mu_{marihuana} > \mu_{alkohola}$. Ondorio berbera ikusten da batezbestekoen diferentzien % 95eko konfiantza-tarteen grafikoan. Tukeyren alderaketa binakatuen arabera, **bi multzo homogeneo** osatzen dira; batean alkohola dago, puntuazio baxuenari dagokionean, eta beste biak beste multzo batean daude.

6.3. Faktore biko bariantza-analisia interakziorik gabe (n = 1)

46. adibidea. *Termometroak ekoizten dituen lantegi batean, marka desberdineko lau makina erabiltzen dira, eta makina bakoitza lau lanaldi ezberdinetako (A, B, C eta D) langileek erabiltzen dute. Honako hau da egun batean lanaldi bakoitzean eta makinako ekoitzitako unitate kopurua:*

Unitate kopurua	*Makina*			
Lanaldi	*A*	*B*	*C*	*D*
1	*14*	*9*	*7*	*8*
2	*12*	*11*	*10*	*9*
3	*16*	*8*	*8*	*11*
4	*14*	*8*	*6*	*10*

Egin ezazu dagokion bariantza-analisia, suposizioak eta ondorioak zehaztuz.

Hipotesi nulua $H_0 : \mu_{11} = \ldots = \mu_{44}$ da.

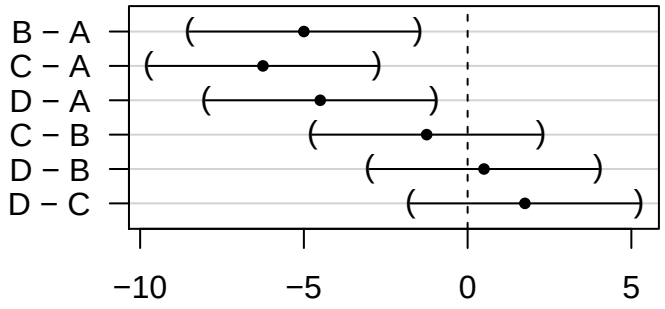
Kasu honetan ere hiru aldagai ditugu, bi kualitatibo (lanaldia eta makina) eta kuantitatibo bat (unitate kopurua). n=1 denez, ez dugu interakziorik, eta, beraz, faktoreen arteko biderkadura (`lanaldia*makina`) erabili beharrean, faktoreen arteko batura (`lanaldia+makina`) erabili behar dugu `aov` funtzioan:

```
> lanaldimakina <- read_excel("praktikadatuak.xls", sheet='lanaldimakina')
> #Aldagai kualitatiboak faktore bihurtu
> lanaldimakina$lanaldia <- factor(lanaldimakina$lanaldia)
> lanaldimakina$makina <- factor(lanaldimakina$makina)
> AnovaModel.4 <- aov(unitateak ~ lanaldia+makina, data=lanaldimakina)
> summary(AnovaModel.4)

            Df Sum Sq Mean Sq F value  Pr(>F)
lanaldia     3   5.19   1.729    0.69 0.58082
makina       3  89.19  29.729   11.86 0.00176 **
Residuals    9  22.56   2.507
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lanaldiaren faktorearekiko estatistikoa $F' = 0,6898$ eta p-balioa $p' = 0,580816$ direnez, faktore horrek ez du eraginik errendimenduan. Hala ere, makinaren arabera, aldatu egiten da errendimendua ($F'' = 11,8587$ eta $p'' = 0,001765$).

Jarraian, konparaketa anitzak aztertuko ditugu: `Comparaciones dos a dos de medias` sakatuz `ANOVA de un factor` lekutik.

```
> bikoteak.3 <- glht(AnovaModel.4, linfct = mcp(makina = "Tukey"))
> plot(bikoteak.3, main="")
```




```
> confint(bikoteak.3)


 Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = unitateak ~ lanaldia + makina, data = lanaldimakina)

Quantile = 3.1224
95% family-wise confidence level


Linear Hypotheses:
           Estimate lwr     upr
B - A == 0 -5.0000  -8.4958 -1.5042
C - A == 0 -6.2500  -9.7458 -2.7542
D - A == 0 -4.5000  -7.9958 -1.0042
C - B == 0 -1.2500  -4.7458  2.2458
D - B == 0  0.5000  -2.9958  3.9958
D - C == 0  1.7500  -1.7458  5.2458

> cld(bikoteak.3)

  A   B   C   D
"b" "a" "a" "a"
```

Tukeyren arabera, ondoriozta daiteke, makina mota faktorean, % 5eko esangura-mailarekin 1. makinaren etekina desberdina dela gainerakoetatik. Eta berriro **bi multzo homogeneo** ditugu: alde batetik, 1. makina, etekin handiena ematen duena, eta, bestaldetik, 2., 3. eta 4. makinak.

7. *R* praktika

Erregresioa

Helburua

Praktika honen helburua da erregresio linealean eta anizkoitzean aplikatzen diren teknikak R bidez nola aplikatu ikastea.

- Lehenik, populazio-ereduaren itxura proposatzeko grafikoak eta teknikak deskribatzen dira, parametroak zenbatesteko bidea adieraziz.
- Bigarrenik, erregresio-ereduaren erabilgarritasuna nola aztertu azalduko da, bai adierazgarritasun orokorra planteatuz, bai koaldagai bakoitzaren garrantzia ikusiz.
- Hirugarrenik, aldagaien arteko korrelazio lineal bakuna, anizkoitza eta partziala nola kalkulatu ikertuko dugu.
- Laugarrenik, diagnosian, hipotesien eta hondarren azterketa nola egin komentatzen da.
- Azkenik, iragarpenak egiteko prozedura azalduko dugu.

7.1. Adibidea

Praktika honetan ere, 16. adibideko datuetan oinarrituko gara helburuak lantzeko:

47. adibidea. *Elektrizitate-konpainia batean, X etxearen neurriaren (oin karratutan) eta Y etxebizitzaren hileko energia-kontsumoa (kwh-tan) aztertu nahi zen. Horretarako, 10 etxebizitza aztertu dira:*

(a) Irudikatu datuak eta aztertu grafikoki zein erregresio-eredu izan daitezkeen egokiak.
(b) Proposatu eta eraiki erregresio-eredu batzuk, eta, aztertu ostean, aukeratu egokiena.

(c) Aztertu erregresio-eredu egokienaren erabilgarritasuna, doikuntza-egokitasuna eta parametroen inferentzia eginez.
(d) Egin korrelazioaren analisia.
(e) Komentatu diagnosia. Zein da hondar txikiena duen behaketa? Eta haren doitutako balioa?
(f) Erabili eredua iragarpenak egiteko. Zein da 1.500 oin karratu dituen etxe baten itxarondako energia-kontsumoa? Kalkulatu iragarpen- eta konfiantza-tarteak.

Galdera hauek gauztiak hurrengo ataletan erantzungo ditugu. Zehazki, (a) galderari 7.2. atalean erantzungo diogu; (b) galderari 7.3. atalean erantzungo diogu, ereduak banan-banan aztertuz, bai grafikoki eta bai metodo estatistikoak erabiliz; (c) puntuari 7.4. atalean emango zaio erantzuna, eta (d) puntuari 7.5. atalean. Diagnosia, hau da, (e) atala 7.6. atalean landuko da. Azkenik, 7.7. atalean iragarpenak nola egin ikusiko dugu, (f) galderari erantzuna emanez.

Hasteko, datuak definituko ditugu:

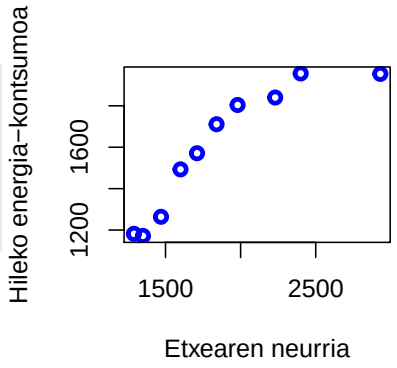
```
> kontsumoa <- c(1182, 1172, 1264, 1493, 1571, 1711, 1804, 1840, 1956, 1954)
> neurria <- c(1290, 1350, 1470, 1600, 1710, 1840, 1980, 2230, 2400, 2930)
> energiadatuak <- data.frame(kontsumoa=kontsumoa, neurria=neurria)
```

Nahi izanez gero, datuak eskuragai daude **praktikakdatuak.xls** fitxategian, **energia** izeneko orrian.

7.2. Datuen irudikapena

Adibideko (a) atalari erantzuna emanez, datuak irudikatu egingo ditugu. Zehazki, baldin $k = 1$ bada, kasu honetan bezala, hodei-puntua egin dezakegu:

```
> plot(energiadatuak$neurria,
+      energiadatuak$kontsumoa,
+      col="blue", lwd=3,
+      xlab="Etxearen neurria",
+      ylab="Hileko energia-kontsumoa")
```



Itxura honekin, hainbat eredu mota izan daitezke egokiak. Beraz, hurrengo ataletan kuantitatiboki aztertuko dugu eredu ezberdinen doikuntza-egokitasuna.

$k > 1$ bada, k grafiko egin beharko ditugu: X_j vs Y bikote bakoitzarentzat bat. Hori egiteko, adibidez, `pairs` funtzioa erabil dezakegu.

7.3. Populazio-eredua proposatzea: $\mathbf{Y = f(X_1, X_2, \ldots, X_k) + \epsilon}$

Atal honetan, erregresio-ereduak deskribatu eta nola eraiki ikasiko dugu, aurreko adibidean oinarrituz. Has gaitezen erregresio-eredu linealarekin: $\hat{Y} = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \ldots + b_k \cdot X_k$ edo, gure kasuan, erregresio-eredu lineal bakuna:

$$\hat{Kontsumoa} = b_0 + b_1 \cdot Neurria$$

Erregresio linealak doitzeko *R*ko komandoa `lm` (*lineal model*) da. Funtzio horren parametro gisa, aldagaiak menpekoak eta askeak definitzen dituen erregresio-formula bat adierazi behar dugu, eta baita zein datu base erabiliko dugun ere eredua doitzeko (`data`). Adibidez:

```
> erreg.lin <- lm(kontsumoa~neurria, data=energiadatuak)
```

Aurreko komandoak, `erreg.lin` objektuan, $\hat{Kontsumoa} = b_0 + b_1 \cdot Neurria$ erregresio lineal doituari buruzko informazioa gordeko du. Kontuan izan erregresio anizkoitzaren kasuan, `lm` agindua erabil dezakegula aldagai aske guztiak zerrendatuz:

```
> erreg.lin2 <- lm(y~x1+x2+x3+...+xk, data=datubasea)
```

Objektu honek gordetzen duen informazioaren laburpena ikusteko, `summary` funtzioa erabiliko dugu:

```
> summary(erreg.lin)


Call:
lm(formula = kontsumoa ~ neurria, data = energiadatuak)

Residuals:
    Min      1Q  Median      3Q     Max
-208.02 -105.36   52.89   77.29  155.27

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 578.92775  166.96806   3.467 0.008476 **
```

```
neurria       0.54030    0.08593   6.288 0.000236 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133.4 on 8 degrees of freedom
Multiple R-squared:  0.8317,Adjusted R-squared:  0.8107
F-statistic: 39.54 on 1 and 8 DF,  p-value: 0.0002359
```

Erregresio-eredu doitua gorde dugun objektuak (`erreg.lin`) zein informazio mota gordetzen duen ikusteko, `names` funtzioa erabil dezakegu:

```
> names(erreg.lin)

 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

Ikusten dugunez, `erreg.lin` objektuak 12 atal ditu, eta bakoitzak informazio zati bat gordetzen du.

Objektu honen laburpena ere beste objektu batean gordetzen badugu, era berean joka genezake:

```
> lab.lin <- summary(erreg.lin)
> names(lab.lin)

 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

Azkenik, objektu hauetatik balio konkreturen bat lortzeko, $ operadorea erabiliko dugu. Adibidez:

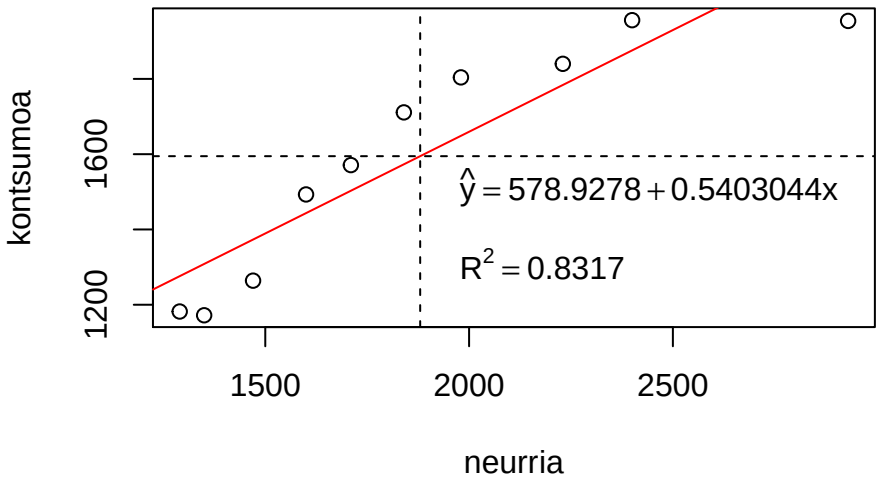
```
> erreg.lin$coefficients #Koefizienteak lortzeko

(Intercept)     neurria
578.9277515   0.5403044

> lab.lin$coefficients #Koefizienteak eta haien inguruko inferentzia

               Estimate   Std. Error  t value     Pr(>|t|)
(Intercept) 578.9277515 166.96805715 3.467296 0.0084763572
neurria       0.5403044   0.08592981 6.287741 0.0002358846

> lab.lin$r.squared #R^2 balioa lortzeko
```

```
[1] 0.8317054

> lab.lin$adj.r.squared #R^2 balioa lortzeko

[1] 0.8106686

> lab.lin$sigma #S hondarren errore estandarra

[1] 133.4377

> lab.lin$fstatistic #F estatistikoa

   value    numdf    dendf
39.53569  1.00000  8.00000
```

Laburpen moduan, $\hat{Y} = 578,92775 + 0,54030X$ da erregresio-eredu linealaren formula. Xren koefizientea guztiz adierazgarria ($p = 0,0002358846$) da, eta konstantearena ($p = 0,0084763572$) oso adierazgarria. $S = 133,4, R^2 = 0,8317, \overline{R}^2 = 0,8107, F = 39,54$ eta, beraz, eredua guztiz adierazgarria da. Balio horiek zehatzago aztertuko ditugu ondoko ataletan.

Irudika dezagun lortutako erregresio-eredu lineala datuekin batera:

```
> #x eta y aldagaien batezbestekoak
> mx <- mean(energiadatuak$neurria)
> my <- mean(energiadatuak$kontsumoa)
>
> #Puntu-hodeia
> plot(energiadatuak$neurria, energiadatuak$kontsumoa,
+      xlab="neurria", ylab="kontsumoa")
>
> #Erregresio zuzena gorriz
> abline(erreg.lin, col="red")
>
> #Batezbestekoen lerro horizontal eta bertikalak
> abline(h=my,lty=2)
> abline(v=mx,lty=2)
>
> #Komentarioak gehitu
> text(mx+50, my-100,
+ expression(hat(y)==578.9278 + 0.5403044 * x), pos=4)
> text(mx+50,my-300, expression(R^2== 0.8317), pos=4)
```



Adibide honetan, erregresio zuzenak ez dirudi eredu egokiena, ez baita datuetara oso ondo doitzen. Beraz, azter ditzagun lineal bihur daitezkeen oinarrizko beste erregresio mota batzuk:

Lineala	$\hat{Y} = b_0 + b_1 X$		
Hiperbolikoa	$\hat{Y} = b_0 + \frac{b_1}{X}$		
Potentziala	$\hat{Y} = b_0 X_1^b$	$\Leftrightarrow$	$\ln \hat{Y} = \ln b_0 + b_1 \ln X$
Esponentziala	$\hat{Y} = b_0 e^{b_1 X}$	$\Leftrightarrow$	$\ln \hat{Y} = \ln b_0 + b_1 X$
Koadratikoa	$\hat{Y} = b_0 + b_1 X + b_2 X^2$		

7. *Rcmdr* praktikan, erregresio mota horiek nola linealizatu aztertu dugu. Beraz, zuzenean R kodearen bitartez nola eraikiko genituzkeen ikusiko dugu:

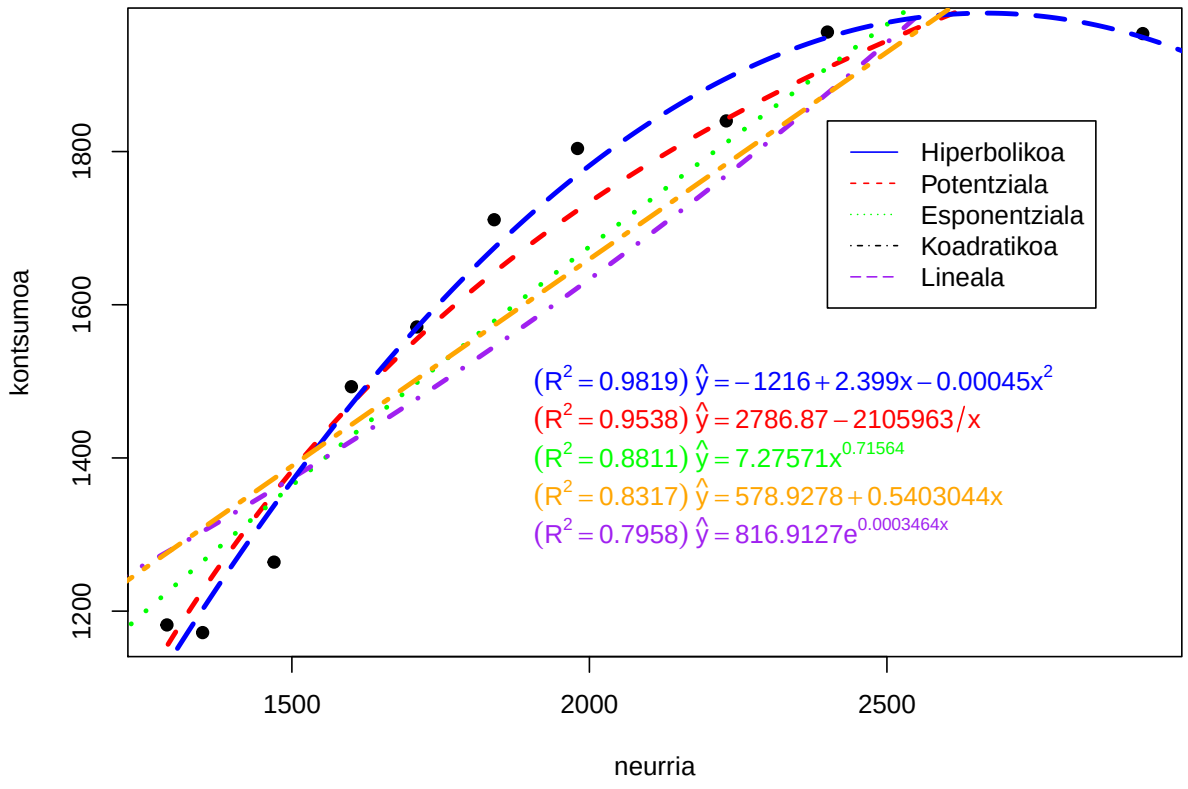
```
> # Erregresio hiperbolikoa
> energiadatuak$hipneurria <- 1/energiadatuak$neurria # Koaldagaia (1/X)
> erreg.hip <- lm(kontsumoa ~ hipneurria, data=energiadatuak)
>
> # Erregresio potentziala
> energiadatuak$lnkontsumoa <- log(energiadatuak$kontsumoa) # Ald. mendekoa (ln Y)
> energiadatuak$lnneurria <- log(energiadatuak$neurria) # Koaldagaia (ln X)
> erreg.pot <- lm(lnkontsumoa~lnneurria, data=energiadatuak)
>
> # Erregresio esponentziala
> erreg.esp <- lm(lnkontsumoa~neurria, data=energiadatuak)
>
> # Erregresio koadratikoa
```

```
> energiadatuak$neurria2 <- energiadatuak$neurria^2 # Koaldagaia (X^2)
> erreg.koad <- lm(kontsumoa~neurria + neurria2, data=energiadatuak)
```

Datuekin batera, doitutako erregresio-eredu hauek guztiak irudika ditzakegu. Gainera, gogoratu R^2 balioa erabil genezakeela ereduak egokienetik desegokienera ordenatzeko. Beraz, bakoitzaren R^2 balioa eta formula komentario gisa grafikoan adieraziko ditugu:

```
> #Puntu-hodeia
> plot(energiadatuak$neurria, energiadatuak$kontsumoa,
+      xlab="neurria", ylab="kontsumoa",
+      pch=21, bg="black")
>
> #Neurri-balio posibleen sekuentzia bat (lerroa leuna geratzeko)
> neurriseq<- seq(1000, 3000, by=1)
>
> #Erregresio hiperbolikoa erabili balio horien iragarpenak egiteko eta irudikatu
> koaldagaiak1 <- data.frame(hipneurria=1/neurriseq)
> yhip <- predict(erreg.hip, koaldagaiak1)
> lines(neurriseq, yhip, type="l", col="red", lty=2, lwd=3)
>
> #Erregresio potentziala erabili balio horien iragarpenak egiteko eta irudikatu
> koaldagaiak2 <- data.frame(lnneurria=log(neurriseq))
> ypot <- exp(predict(erreg.pot, koaldagaiak2))
> lines(neurriseq, ypot, type="l", col="green", lty=3, lwd=3)
>
> #Erregresio esponentziala erabili balio horien iragarpenak egiteko eta irudikatu
> koaldagaiak3 <- data.frame(neurria=neurriseq)
> yesp <- exp(predict(erreg.esp, koaldagaiak3))
> lines(neurriseq, yesp, type="l", col="purple", lty=4, lwd=3)
>
> #Erregresio koadratikoa erabili balio horien iragarpenak egiteko eta irudikatu
> koaldagaiak4 <- data.frame(neurria=neurriseq, neurria2=neurriseq^2)
> ykoad <- predict(erreg.koad, koaldagaiak4)
> lines(neurriseq, ykoad, col="blue", lty=5, lwd=3)
>
> #Erregresio zuzena erabili balio horien iragarpenak egiteko eta irudikatu
> koaldagaiak5 <- data.frame(neurria=neurriseq)
> ykoad <- predict(erreg.lin, koaldagaiak5)
> lines(neurriseq, ykoad, col="orange", lty=6, lwd=3)
>
> #Komentarioak gehitu
> text(mx, my-150,expression((R^2==0.9538)~hat(y)==2786.87-2105963/x),
+      pos=4,col="red")
> text(mx, my-200,expression((R^2==0.8811)~hat(y)==7.27571*x^{0.71564}),
+      pos=4,col="green")
```

```
> text(mx, my-300,expression((R^2==0.7958)~hat(y)==816.9127*e^{0.0003464*x}),
+      pos=4,col="purple")
> text(mx, my-100,expression((R^2==0.9819)~hat(y)==-1216+2.399*x-0.00045*x^2),
+                            pos=4, col="blue")
> text(mx, my-250,expression((R^2==0.8317)~hat(y)==578.9278+0.5403044*x),
+      pos=4, col="orange")
>
> #Legenda gehitu
> legend(2400, 1840,
+        c("Hiperbolikoa","Potentziala","Esponentziala", "Koadratikoa","Lineala"),
+        lty=c(1,2,3,4,5), col=c("blue","red","green","black","purple"))
```



Kontuan izan `predict` funtzioak, erregresio-eredu zehatz bat erabiliz, zehaztutako balio batzuen iragarpenak egiten dituela. Kapitulu honetako azken atalean zehaztuko dugu, xehetasun gehiagorekin, funtzio horren erabilera.

Nahiz eta bost ereduak guztiz erabilgarriak izan, R^2-ren arabera ordenatuz eta baita irudiari

dagokionez ere, egokiena eredu koadratikoa dela ondorioztatzen da:

$$\hat{Y} = b_0 + b_1 X + b_2 X^2 = -1216 + 2,399X - 0,0004500X^2$$

7.4. Ereduaren erabilgarritasuna

Erregresio koadratikoa oinarritzat hartuz, ereduaren erabilgarritasuna nola aztertu azalduko dugu jarraian.

7.4.1. Doikuntza-egokitasuna

Doikuntza-egokitasun globala aztertzeko, honako hipotesi hau definituko genuke:

$$\begin{cases} H_0: & B_1 = B_2 = 0 \\ H_1: & \exists i : B_i \neq 0 \end{cases}$$

```
> summary(erreg.koad)


Call:
lm(formula = kontsumoa ~ neurria + neurria2, data = energiadatuak)

Residuals:
    Min      1Q  Median      3Q     Max
-73.792 -22.426   5.886  31.689  52.436

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.216e+03  2.428e+02  -5.009 0.001550 **
neurria      2.399e+00  2.458e-01   9.758 2.51e-05 ***
neurria2    -4.500e-04  5.908e-05  -7.618 0.000124 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.8 on 7 degrees of freedom
Multiple R-squared:  0.9819,Adjusted R-squared:  0.9767
F-statistic: 189.7 on 2 and 7 DF,  p-value: 8.001e-07
```

Determinazio-koefizientea $R^2 = 0,9819$ da eta zuzendutakoa $\overline{R}^2 = 0,9767$. Azkenik, zenbatespen-errore estandarra $S = 46,8$ da.

F estatistikoak $F = 189,7$ balio du, eta dagokion p-balioa $p = 8 \cdot 10^{-7}$ denez, eredua guztiz erabilgarria dela esan dezakegu.

Hondar-bariantza, bariantza azaldua eta azaltzeko bariantza kalkulatzeko `anova` komandoa erabil daiteke:

```
> anova(erreg.koad)

Analysis of Variance Table

Response: kontsumoa
          Df Sum Sq Mean Sq F value    Pr(>F)
neurria    1 703957  703957 321.388  4.15e-07 ***
neurria2   1 127112  127112  58.032 0.0001244 ***
Residuals  7  15333    2190
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Horrela, $S^2_{\hat{y}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{n} = \frac{703957+127112}{10} = 83106,9$, $S^2_e = \frac{\sum e_i^2}{n} = \frac{15333}{10} = 1533,3$ eta $S^2_y = \frac{\sum(y_i - \bar{y})^2}{n} = \frac{831069-15333}{10} = 84640,2$. Izan ere, $R^2 = \frac{S^2_{\hat{y}}}{S^2_y} = 0,981885$.

7.4.2. Parametroen inferentzia

$H_0 : B_i = 0$, (hots, X_i aldagaia ez da adierazgarria) motako hipotesi nulua duten testen emaitzak ikusteko:

```
> summary(erreg.koad)$coefficients

                Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) -1.216144e+03 2.428064e+02 -5.008698 1.550025e-03
neurria      2.398930e+00 2.458356e-01  9.758270 2.513355e-05
neurria2    -4.500402e-04 5.907662e-05 -7.617907 1.244152e-04
```

X eta X^2-ren koefizienteak guztiz adierazgarriak dira ($p = 2,513355 \cdot 10^{-5}$ eta $p = 1,244152 \cdot 10^{-4}$, hurrenez hurren), eta konstatearena oso adierazgarria da ($p = 1,550025 \cdot 10^{-3}$).

Parametroen konfiantza-tarteak lortzeko, `confint` komandoa erabil dezakegu:

```
> confint(erreg.koad, level=0.95)

                   2.5 %        97.5 %
(Intercept) -1.790290e+03 -6.419981e+02
neurria      1.817621e+00  2.980239e+00
neurria2    -5.897342e-04 -3.103462e-04
```

Horrela, $I_A^{0,95} = (-1790,290,\ -641,9981)$, $I_{B_1}^{0,95} = (1,817621,\ 2,980239)$, eta $I_{B_2}^{0,95} = (-0,00058973424,\ -0,0003103462)$ direla ikus dezakegu.

Gainera, datuak tipifikatzen baditugu, tipifikatutako beta koefizienteak lor ditzakegu:

```
> energiadatuak.tip <- energiadatuak
> energiadatuak.tip$kontsumoa <- (energiadatuak$kontsumoa-
+                                   mean(energiadatuak$kontsumoa))/
+                                   sd(energiadatuak$kontsumoa)
> energiadatuak.tip$neurria <- (energiadatuak$neurria-
+                                 mean(energiadatuak$neurria))/
+                                   sd(energiadatuak$neurria)
> energiadatuak.tip$neurria2 <- (energiadatuak$neurria2-
+                                  mean(energiadatuak$neurria2))/
+                                   sd(energiadatuak$neurria2)
> erreg.koad.tip <- lm(kontsumoa ~ neurria + neurria2, data=energiadatuak.tip)
> summary(erreg.koad.tip)


Call:
lm(formula = kontsumoa ~ neurria + neurria2, data = energiadatuak.tip)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24062 -0.07313  0.01919  0.10333  0.17099

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.721e-16  4.826e-02   0.000 1.000000
neurria      4.049e+00  4.149e-01   9.758 2.51e-05 ***
neurria2    -3.161e+00  4.149e-01  -7.618 0.000124 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1526 on 7 degrees of freedom
Multiple R-squared:  0.9819,Adjusted R-squared:  0.9767
F-statistic: 189.7 on 2 and 7 DF,  p-value: 8.001e-07
```

$\beta_1 = 4,049$ eta $\beta_2 = -3,161$ direnez, etxearen neurria da gehien azaltzen duen aldagaia.

7.5. Korrelazioa

S^2 kobariantzak eta r Pearsonen korrelazio-indizeak kalkulatzeko:

```
> energiadatuak.koad <- subset(energiadatuak,
+                              select=c(kontsumoa, neurria, neurria2))
> cov(energiadatuak.koad)

             kontsumoa      neurria     neurria2
kontsumoa     94044.68     144765.6 5.664858e+08
neurria      144765.56     267933.3 1.106541e+09
neurria2  566485813.33 1106540666.7 4.639648e+12
```

Beraz, bariantzak: $s_y^2 = 94044,68$, $s_1^2 = 267933,3$ eta $s_2^2 = 4,639648\ 10^{12}$, kobariantzak: $s_{y1} = 144765,6$, $s_{y2} = 566485813,33$ eta $s_{12} = 1106540666,7$ eta Pearsonen korrelazio bakunak: $r_{y1} = 0,9119788$, $r_{y2} = 0,8575895$ eta $r_{12} = 0,9924566$.

Korrelazioen testak egiteko ($H_0 : \rho_{ij} = 0$), liburutegiko **Hmisc** paketea instalatu ondoren:

```
> library(Hmisc)
> energiadatuak.koad <- as.matrix(energiadatuak.koad)
> rcorr(energiadatuak.koad, type="pearson")

          kontsumoa neurria neurria2
kontsumoa      1.00    0.91     0.86
neurria        0.91    1.00     0.99
neurria2       0.86    0.99     1.00

n= 10


P
          kontsumoa neurria neurria2
kontsumoa           0.0002  0.0015
neurria   0.0002            0.0000
neurria2  0.0015    0.0000
```

Beraz, ondorioztatzen da aldagai-bikote guztien korrelazioak ez direla nuluak; hau da, ez dira koerlaziorik gabekoak, baizik eta koerlazionatuak. Emaitza horrek egiaztatzen du X eta X^2 koaldagaien artean multikolinealtasuna dagoela.

Korrelazio partzialak lortzeko, liburutegiko **ggm** paketea instalatu eta kargatu behar da:

```
> library(ggm)
> pcor(c("kontsumoa", "neurria", "neurria2"), var(energiadatuak.koad))

[1] 0.9651543

> pcor(c("kontsumoa", "neurria2", "neurria"), var(energiadatuak.koad))

[1] -0.9446489

> pcor(c("neurria", "neurria2", "kontsumoa"), var(energiadatuak.koad))

[1] 0.9969376
```

Horrela, $r_{y1\cdot 2} = 0,9651543$, $r_{y2\cdot 1} = -0,9446489$ eta $r_{12\cdot y} = 0,9969376$.

7.6. Diagnosia

Hasteko, hondarrak eta doitutako balioak ateratzeko:

```
> hondarrak <- residuals(erreg.koad) # hondarrak
> hondar.tipifikatuak <- rstandard(erreg.koad) # hondar tipifikatuak
> doitutako.balioak <- fitted(erreg.koad) # doitutako balioak
```

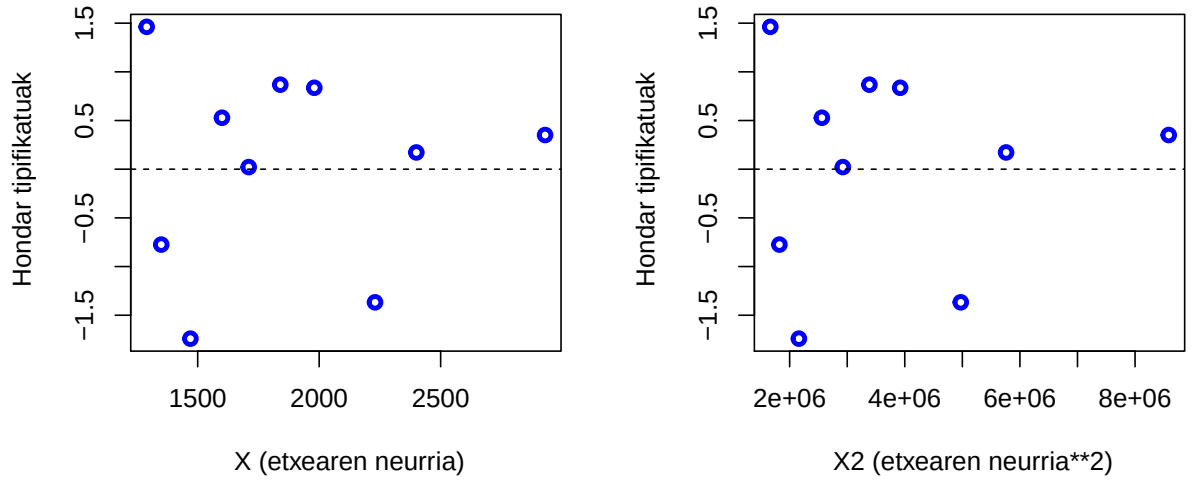
Era honetan, kalkula dezakegu, adibidez, 47. adibideko (e) atalari erantzunez, errorerik txikiena duen balioa zein den:

```
> # Errore minimoko behaketaren indizea
> behaketa <- which(hondarrak==min(abs(hondarrak)))
> # Behaketa honen errorea
> hondarrak[behaketa]

        5
0.9358866

> # Behaketa honen doitutako balioa
> fitted(erreg.koad)[behaketa]

       5
1570.064
```

Beraz, errore txikiena duen behaketa bosgarrena da, bere errorea $e_5 = 0,9359$ izanik eta bere doitutako balioa $\hat{y}_5 = 1570,064$.

Orain, erroreen analisia egiten has gaitezke:

```
> par(mfrow=c(1,2))
>
> #Irudikatu X vs. errore tipikoak
> plot(energiadatuak$neurria, hondar.tipifikatuak, col="blue", lwd=3,
+ xlab=" X (etxearen neurria)", ylab=" Hondar tipifikatuak")
> abline(h=0,lty=2)
>
> #Irudikatu X^2 vs. errore tipikoak
> plot(energiadatuak$neurria2, hondar.tipifikatuak, col="blue", lwd=3,
+ xlab=" X2 (etxearen neurria**2)", ylab=" Hondar tipifikatuak")
> abline(h=0,lty=2)
```



Puntuak zorizkoak direnez (forma zehatzik ez dute), horiek egitean **zehaztatze-akatsik ez** dagoela ondorioztatzen dugu, hau da, ereduaren itxura ondo aukeratu dugula.

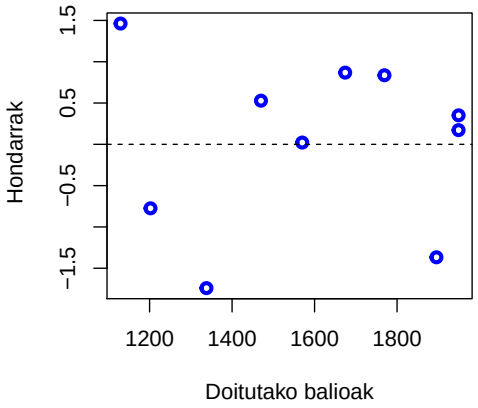
• Orain, erroreak autokorrelatuak dauden ala ez aztertuko dugu. Horretarako, aurreko grafikoek uhin itxura duten ala ez begiratuko dugu, eta, gainera, Durbin-Watson testa egingo dugu, `car` paketeko `durbinWatsonTest` funtzioa erabiliz:

```
> library(car)
> durbinWatsonTest(erreg.koad)

 lag Autocorrelation D-W Statistic p-value
   1      -0.1298873      2.078928   0.504
 Alternative hypothesis: rho != 0
```

Aurreko grafikoak uhin-itxurakoak ez direnez eta Durbin-Watson koefizientea (2,078928) 2tik hurbil dagoenez, hondarren artean **autokorrelaziorik ez** dagoela ondorioztatzen dugu (0tik edo 4tik hurbil egonez gero autokorrelazio negatiboa edo positiboa dagoela esaten da, hurrenez hurren)

• Orain, erroreen homozedastizitatea aztertuko dugu $\hat{Y}^{tip}$ *vs* e^{tip} grafikoa aztertuz:

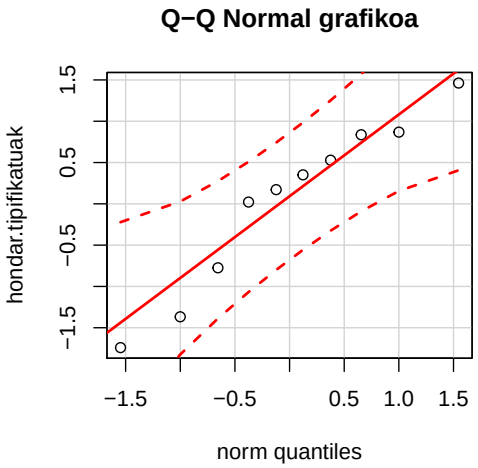
```
> plot(doitutako.balioak, hondar.tipifikatuak,
+      col="blue", lwd=3,
+      xlab=" Doitutako balioak ",
+      ylab=" Hondarrak")
> abline(h=0,lty=2)
```



Grafikoa $e^{tip} = 0$ ardatzean zentratutako banda horizontal batean dagoenez (ez da elipse edo triangelu itxurakoa, besteak beste), hondarren **homozedastizitatea** onartzen dugu.

• Azkenik, e hondar-aldagaian $(-3,\ 3)$ tartetik kanpoko daturik ez dagoela, eta, aldagaiaren normaltasuna azter dezakegu normal Q-Q grafiko bat eta Kolmogorov-Smirnov proba erabiliz.

```
> qqPlot(hondar.tipifikatuak,
+        main="Q-Q Normal grafikoa")
```



Alde batetik, e_i balio guztiak $(-3,\ 3)$ tartean daudenez, ez dago tartetik kanpoko daturik. Beste aldetik, normal Q-Q grafikoaren puntuak 1. koadrantearen erdikaritik hurbil daudenez, normaltasuna onartzen da. Gainera, Kolmogorov-Smirnoven kontrastea e^{tip} aldagaian aplikatuz ($e \approx \mathcal{N}(0, S)$, hots, $e^{tip} = e/S \approx \mathcal{N}(0,1)$):

```
> ks.test(hondar.tipifikatuak, dist="pnorm", 0, 1)


Two-sample Kolmogorov-Smirnov test

data:  hondar.tipifikatuak and 0
D = 0.7, p-value = 0.7273
alternative hypothesis: two-sided
```

$p - balioa = 0,7273$ denez, ezin da hondarren normaltasuna errefusatu % 95eko konfiantza-mailaz.

• Azkenik, multikolinealtasuna aztertu beharko genuke. Kasu honetan, **multikolinealtasuna** dagoen susmoa dugu, X eta X^2 aldagaien arteko korrelazioa adierazgarria delako (izan ere, bi aldagai horien arteko korrelazio nuluarean testaren $p - balioa = 1,403848 \cdot 10^{-8}$ da, lehen ikusi dugun moduan).

7.7. Iragarpena

Behin erregresio-eredua ontzat emanda, zenbatespenak eta iragarpenak egiteko erabil dezakegu. Horretarako, `predict` funtzioa erabiliko dugu, eredua eta iragarpenak lortu nahi ditugun koaldagaien balioak zehaztuz. Adibidez, 1500 neurriko etxe baten kontsumoa erregresio koadratikoaren bidez iragartzeko:

```
> datu.berriak <- data.frame(neurria=1500, neurria2=1500^2)
> predict(erreg.koad, datu.berriak)

       1
1369.661
```

Kalkula dezagun, orain, dagokion iragarpen-tartea:

```
> predict(erreg.koad, datu.berriak, interval="prediction")

       fit      lwr      upr
1 1369.661 1250.317 1489.005
```

Hots, $I^{1-\alpha}_{y_0|x_0} = I^{0,95}_{y_0|1500} = (1250,317,\ 1489,005)$

Azkenik, kalkula dezagun batezbestekorako konfiantza-tartea:

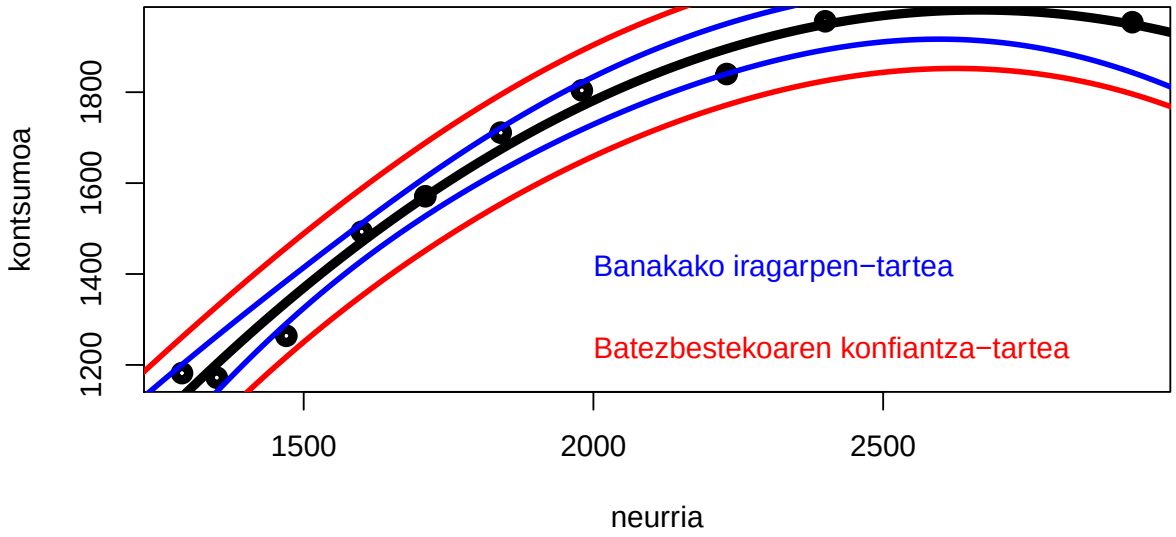
```
> predict(erreg.koad, datu.berriak, interval="confidence")

       fit      lwr      upr
1 1369.661 1324.989 1414.333
```

Hots, $I^{1-\alpha}_{\mu_y|x_0} = I^{0,95}_{\mu_y|1500} = (1324,989,\ 1414,333)$.

Irudika ditzagun erregresio-eredu koadratikoan banakako iragarpenerako eta batezbestekorako konfiantza-tarteak.

```
> #Balio berriak definitu
> xseq=seq(1000, 3000, 1)
> datu.berriak <- data.frame(neurria=xseq, neurria2=xseq**2)
>
> #Iragarpen-tarteak kalkulatu
> irag.bat <- predict(erreg.koad, datu.berriak, interval="confidence")
> doituak <- irag.bat[,1]
> lwi <- irag.bat[,2]
> uwi <- irag.bat[,3]
>
> #Konfiantza-tarteak kalkulatu
> irag.giz <- predict(erreg.koad, datu.berriak, interval="prediction")
> lwm <- irag.giz[,2]
> uwm <- irag.giz[,3]
```

```
> #Irudikapena
> plot(energiadatuak$neurria, energiadatuak$kontsumoa, lwd=5,
+      xlab="neurria", ylab="kontsumoa")
> lines(xseq, doituak, col="black", lty=1, pch=1, lwd=5)
> text(2000, my-200,"Banakako iragarpen-tartea", col="blue", adj=c(0,0))
> text(2000, my-380,"Batezbestekoaren konfiantza-tartea", col="red", adj=c(0,0))
> lines(lwi~xseq, col="blue", lwd=3)
> lines(uwi~xseq, col="blue", lwd=3)
> lines(lwm~xseq, col="red", lwd=3)
> lines(uwm~xseq, col="red", lwd=3)
```



Aurreko pauso guztiak beste edozein erregresio eredurentzat jarrai ditzakegu, era baliokide batean.

8. *R* praktika

Kalitatearen kontrol estatistikoa

Helburua

Praktika honen xede nagusia zera da: kalitatearen kontrol estatistikoan erabiltzen diren oinarrizko kontrol-grafikoak R erabiliz eraikitzea. Alde batetik, aldagaien eta atributuen grafikoak nola eraiki azalduko dugu; bestalde, Paretoren diagrama nola sortu erakutsiko dugu.

8.1. Aldagaien grafikoak

Atal honetan azalduko da nola eraiki $\overline{x}$ batez besteko balioen, R heinen eta S desbideratze estandarren kontrol-grafikoak, R erabiliz.

48. adibidea. *Zementu-fabrika batean, 9 ordu eta erdian zehar, ordu erdian behin lau zaku zementu hautatu dira zoriz, eta hauek izan dira haien pisuak:*

lagina	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ordua	7	$7\frac{1}{2}$	8	$8\frac{1}{2}$	9	$9\frac{1}{2}$	10	$10\frac{1}{2}$	11	$11\frac{1}{2}$	12	$12\frac{1}{2}$	13	$13\frac{1}{2}$	14	$14\frac{1}{2}$	15	$15\frac{1}{2}$	16	$16\frac{1}{2}$
1	53	50	52	46	51	48	51	48	47	49	54	49	50	53	50	51	47	52	50	49
2	51	49	50	47	48	49	46	47	52	50	55	45	48	54	49	50	51	51	53	50
3	52	50	53	52	49	51	49	48	46	50	53	46	51	52	48	49	48	50	48	51
4	49	49	51	50	50	50	50	49	51	51	52	48	47	50	51	52	49	49	49	52

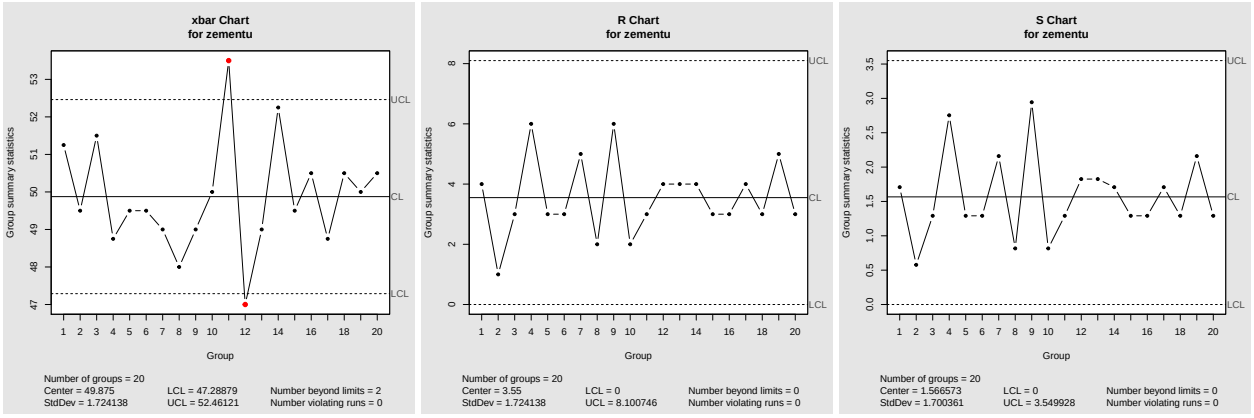
Egin itzazu batezbestekoen, heinen eta desbideratze estandarraren kontrol-grafikoak. Zer ondoriozta daiteke?

Datuak **praktikadatuak.xls** fitxategiko **zementu** orrialdetik kargatu ostean, ikus dezakegu bi aldagaiez osatutako datu-base batean gordeta daudela. Alde batetik, *lagina* aldagaiak 1etik 20rako balioak hartuko ditu, lagina zein ordutan jaso den adieraziz. Bestalde, *pisua* aldagaian hautatutako zaku zementuen pisuak gordeko dira:

```
> zementu<- read_excel("praktikadatuak.xls", sheet='zementu')
> #Aldagai kualitatiboa faktore bihurtu
> zementu$lagina <- factor(zementu$lagina)
```

Ondoren, **qcc** paketea instalatu eta kargatu ostean, honela egingo genituzke hiru grafikoak:

```
> library(qcc)
> zementu <- qcc.groups(zementu$pisua, zementu$lagina)
> #Batez besteko balioak
> obj.x<- qcc(zementu,type="xbar", plot=FALSE)
> #Heina
> obj.r<- qcc(zementu,type="R", plot=FALSE)
> #Desbiderapen estandarrak
> obj.s<- qcc(zementu,type="S", plot=FALSE)
>
> #Irudikapena
> dev.new(width=5, height=5)
> plot(obj.x)
> plot(obj.r)
> plot(obj.s)
```



Gogoratu lerro zentralak zenbatesten ari garen parametroaren zenbatespen puntuala adierazten digula. Ondoren, lerro etenen bidez, kontrol-mugak adierazten dira. Ikus dezakegunez, aldakortasuna kontrol-mugen artean dagoen arren (hots, batezbestekoen inguruan homogeneotasun handia), kontrol-mugetatik kanpo daude 11. eta 12. behaketak (gorriz adierazita daude). Beraz, sistema kontroletik kanpo dagoela ondorioztatuko genuke.

Grafikoarekin batera estatistiko batzuk inprimatu nahiko bagenitu, `add.stats=TRUE` jarri beharko genuke. Edo bestela, `summary` funtzioa erabili; adibidez:

```
> summary(obj.x)


Call:
qcc(data = zementu, type = "xbar", plot = FALSE)

xbar chart for zementu

Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  47.00   49.00   49.50   49.88   50.50   53.50

Group sample size:  4
Number of groups:  20
Center of group statistics:  49.875
Standard deviation:  1.724138

Control limits:
      LCL      UCL
 47.28879 52.46121
```

Beraz, batez besteko balioen zenbatespen puntuala 49,875 da, eta kontrol-mugak 47,2888 eta 52,4612. Gainera, batez besteko akats kopuru minimoa 47 da, eta maximoa 53,50, bi balio horiek kontrol-mugetatik kanpo geratzen direlarik.

8.2. Atributuen grafikoak

8.2.1. Akastun unitateen ehunekoa (p kontrol-grafikoa)

49. adibidea. *Bonbillak ekoizten dituen enpresa batean, 70 bonbillako 30 zorizko lagin hautatu ondoren, honako hau izan zen akastun bonbillen kopurua, hurrenez hurren: 1, 2, 0, 3, 2, 0, 1, 3, 1, 2, 1, 1, 0, 1, 0, 1, 0, 2, 1, 3, 1, 0, 1, 0, 2, 0, 4, 1, 2, 0. Zenbatets ezazu ekoitzitako akastun bonbillen ehunekoa, eta egin ezazu akastun bonbilla-portzentajearen kontrol-grafikoa.*

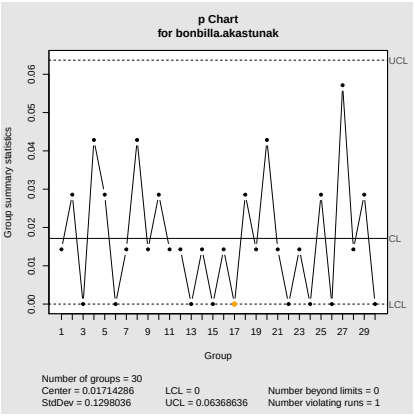
Hasteko, 30 luzeradun *akastunak* izeneko aldagaia osatuko dugu, eta 70 balio konstantea duen *tamaina* izeneko aldagaia.

```
> bonbilla.akastunak <- c(1, 2, 0, 3, 2, 0, 1, 3, 1, 2, 1, 1, 0, 1, 0, 1, 0,
+                2, 1, 3, 1, 0, 1, 0, 2, 0, 4, 1, 2, 0)
> n <- length(bonbilla.akastunak)
> k <- 70
> bonbillak.tamaina <- rep(k,n)
```

Inportatu nahi izanez gero, datu hauek eskuragai daude **praktikadatuak.xls** fitxategiko **bonbillak** orrialdean.

Ondoren, `qcc` funtzioa erabiliz eta akastun bonbilla-portzentajearen kontrol-grafikoa egiteko:

```
> bonbillak.p<- qcc(bonbilla.akastunak, sizes=bonbillak.tamaina, type="p")
> plot(bonbillak.p)
```



Ekoitzitako akastun bonbillak % 1,71 dira, goiko kontrol-muga % 6,37 izanik. Grafikoari begiratuz, bonbillen ekoizpenaren prozesua kontrolen menpe dagoela ondoriozta daiteke.

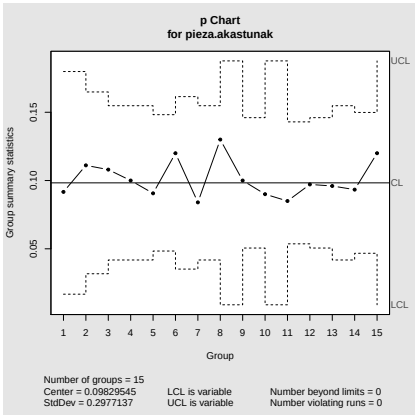
50. adibidea. *Produkzio-prozesu batean, hainbat tamainatako 15 lagin hartu dira, eta honako hau izan da akastun piezen kopurua:*

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
n_i	120	180	250	250	320	200	250	100	350	100	400	350	250	300	100
x_i	11	20	27	25	29	24	21	13	35	9	34	34	24	28	12

Zenbatetsi ekoitzitako akastun piezen ehunekoa, eta egin ezazu akastunen ehunekoaren kontrol-grafikoa.

Berriro, *akastunak* eta *tamaina* izeneko aldagaiak osatuko ditugu. Kontuan izan kasu honetan, laginaren tamaina ez dela konstantea:

```
> pieza.akastunak <- c(11, 20, 27, 25, 29, 24, 21, 13, 35, 9, 34, 34, 24, 28, 12)
> piezak.tamaina <- c(120, 180, 250, 250, 320, 200, 250, 100, 350,
+              100, 400, 350, 250, 300, 100)
> piezak.p<- qcc(pieza.akastunak, sizes=piezak.tamaina, type="p")
> plot(piezak.p)
```

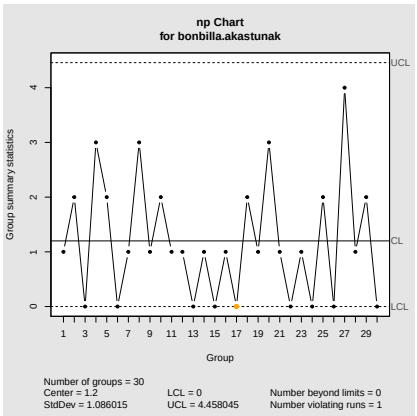


Tamainak konstanteak ez direnez, kontrol-mugak ere ez. Grafikoari begiratuz, piezen ekoizpenaren prozesua kontrolen mende dagoela ondoriozta daiteke, eta ekoitzitako akastun piezen portzentajea % 9,83 da.

8.2.2. Akastun unitateen kopurua (*np* kontrol-grafikoa)

Egin dezagun 49. adibideari lotutako *np* kontrol-grafikoa, tamaina konstantea dela kontuan izanik:

```
> bonbillak.np<- qcc(bonbilla.akastunak, sizes=bonbillak.tamaina, type="np")
> plot(bonbillak.np)
```



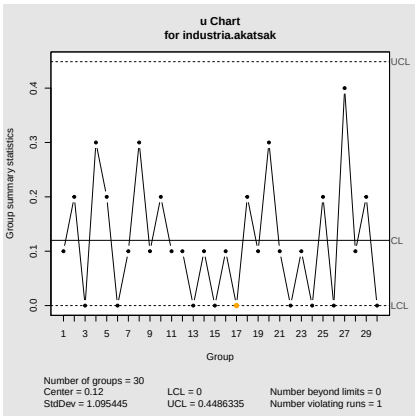
Grafikoari begiratuz, berriz egiaztatzen da bonbillen ekoizpenaren prozesua kontrolen mende dagoela. Ohartu, 49. adibideko grafikoarekin alderatuta, eskala baino ez dela aldatu. Izan ere, grafiko mota honek akastun unitate osoa aztertzen du, n konstantea denean.

8.2.3. Batez besteko akatsen kopurua unitateko (u kontrol-grafikoa)

51. adibidea. *Industria-zentro batean artikulu berezi baten ekoizpena kontrolatzeko, bi orduan behin, 10 unitateko loteak hautatu eta aztertzen dira. Horrela, 30 lagin zoriz hautatu dira, akatsen kopurua honako hau izanik: 1, 2, 0, 3, 2, 0, 1, 3, 1, 2, 1, 1, 0, 1, 0, 1, 0, 2, 1, 3, 1, 0, 1, 0, 2, 0, 4, 1, 2, 0. Zer ondoriozta daiteke batez besteko akatsen kopuruaren inguruan?*

Hasteko, *akatsak* izeneko aldagaia sortuko dugu, eta 10 balio konstanteko *tamaina* aldagaia ere bai. Nahi izanez gero, datu horiek eskuragai daude **praktikadatuak.xls** fitxategiko **bonbillak** orrialdean, eta inporta daitezke. Ondoren, `qcc` funtzioa erabili dugu `type="u"` zehaztuz, batez besteko akats kopurua aztertzeko:

```
> industria.akatsak <- c(1,2,0,3,2,0,1,3,1,2,1,1,0,1,0,
+                        1,0,2,1,3,1,0,1,0,2,0,4,1,2,0)
> n <- length(industria.akatsak)
> k <- 10
> tamaina.industria <- rep(k,n)
> akatsak.u<- qcc(industria.akatsak, sizes=tamaina.industria, type="u")
> plot(akatsak.u)
```



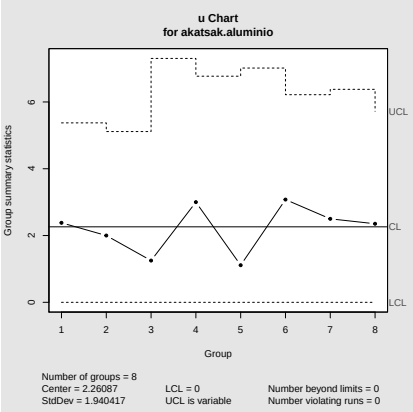
Unitateko batez besteko akats kopurua 0,12 da, eta, behaketa guztiak goi-muga baino baxuagoak direnez, prozesua kontrolaren menpe dagoela esan dezakegu.

52. adibidea. *Aluminio anodizatua ekoizten duen enpresa batek prozesua ikuskatzean metro karratuko akatsak zenbatu ditu. Prozesua kontrolpean dagoela ondoriozta daiteke?*

Lagina	1	2	3	4	5	6	7	8
Ordua	7	8	9	10	11	12	13	14
Behatutako m^2 kopurua	2,1	2,5	0,8	1	0,9	1,3	1,2	1,7
Behatutako akatsen kopurua	5	5	1	3	1	4	3	4

Aurreko kasuetan bezala, bi aldagai sortuko ditugu: lehenengoa 8 luzeradun *akatsak* izenekoa, akatsen kopuruak adierazten dituena, eta bigarrena *tamaina* deritzona, behaketa bakoitzean aztertutako azalera zehaztuko diguna [1]. Ondoren, `qcc` funtzioa erabiliko dugu batez besteko akats kopurua aztertzeko:

```
> akatsak.aluminio <- c(5, 5, 1, 3, 1, 4, 3, 4)
> tamaina.aluminio <- c(2.1, 2.5, 0.8, 1, 0.9, 1.3, 1.2, 1.7)
> aluminio.u<- qcc(akatsak.aluminio, sizes=tamaina.aluminio, type="u")
> plot(aluminio.u)
```



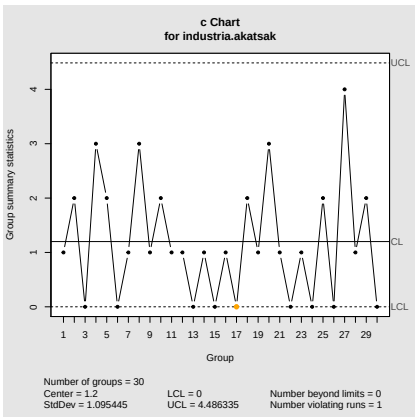
Lerro zentralak adierazten digu $\overline{u} = 2,26$ akats metro karratuko daudela. Gainera, behaketa guztiak goi-muga baino baxuagoak direnez, prozesua kontrolaren mende dagoela esan dezakegu.

8.2.4. Akatsen kopurua artikuluko (c kontrol-grafikoa)

Laginaren tamaina konstantea denean, c kontrol-grafikoak artikulu bakoitzeko akats kopurua aztertzen laguntzen digu. Egin dezagun, adibidez, 51. adibideari lotutako c kontrol-grafikoa, tamaina konstantea baita.

```
> akatsak.c<- qcc(industria.akatsak, type="c")
> plot(akatsak.c)
```

[1] Datu hauek eskuragai daude **praktikadatuak.xls** fitxategiko **aluminio** orrialdean.



Lote bakoitzeko batez besteko akatsen kopurua 1,2 da, eta, behaketa guztiak goi-muga baino baxuagoak direnez, prozesua kontrolaren mende dago.

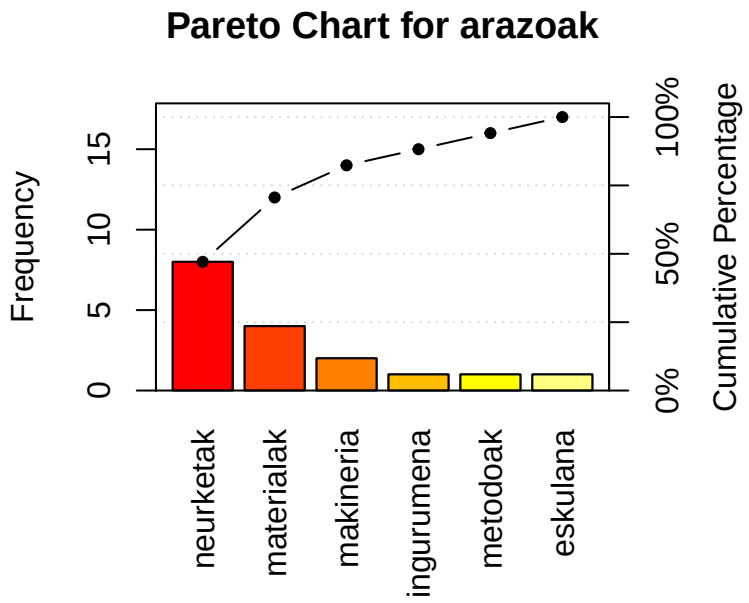
8.3. Paretoren diagrama

Pareto diagrama, aldagai kualitatiboak aztertzeko grafiko berezi bat da. Kalitate-kontrolean maiz erabiltzen da, produkzio-prozesuan sortutako arazo moten maiztasunak aztertzeko.

53. adibidea. *Laboreak ekoizten dituen enpresa baten salmenta-sailean bezeroen kexak aztertzen ari dira. Horretarako, lantalde batek bukatutako produktuen arazoen kausa-efektuak ikertu ondoren, honela sailkatu dira arazo nagusiak: (1) makineria (zaku-betetzailea, etiketatzailea, paletizatzailea), (2) ingurumena (kutsadura, hautsa), (3) materialak (lehengaiak, gaizki ixtea edo irekitzea), (4) erabilitako metodoak (zaku eta paleten manipulazioa), (5) egindako neurketak (pisatzea, nahasturaren neurria), (6) eskulana (makineriaren manipulazioa, zaku eta paleten garraioa). Demagun aste batean atzemandako arazoak honako hauek izan direla: 2 makineria motakoak, 1 ingurumen motakoa, 4 materialengatik, 1 metodoagatik, 8 neurketengatik eta 1 eskulana motakoa. Egin ezazu Paretoren diagrama. Grafikoari begiratuz, zein izango da lehenengo helburua?*

Hasteko, eraiki dezagun *arazoak* izeneko aldagaia, dagozkion izenak esleituz. Paretoren grafikoa `pareto.chart` funtzioaren bidez lor daiteke.

```
> arazoak <- c(2, 1, 4, 1, 8, 1)
> names(arazoak) <- c("makineria", "ingurumena", "materialak",
+ "metodoak", "neurketak", "eskulana")
> pareto.chart(arazoak)
```




```
Pareto chart analysis for arazoak
             Frequency Cum.Freq. Percentage Cum.Percent.
  neurketak          8         8  47.058824     47.05882
  materialak         4        12  23.529412     70.58824
  makineria          2        14  11.764706     82.35294
  ingurumena         1        15   5.882353     88.23529
  metodoak           1        16   5.882353     94.11765
  eskulana           1        17   5.882353    100.00000
```

Paretoren diagraman bi grafiko ikusten dira batera. Barra beherakorrek (bere eskala bertikalki ezkerraldean dagoelarik) maiztasun absolutuak (f_i) adierazten dituzte. Kurba metatuan (bere eskala bertikalki eskuinaldean dagoelarik) ehuneko metatuak adierazten dira (H_i). Porroten % 71a lehenengo bi kausek eragin dute: neurketek eta materialekin lotutako arazoek. Lehenengo helburua, beraz, neurketen arloko alderdi guztiak hobetzea izango litzateke; izan ere, arazoen % 47 arlo honetakoa da.

III. atala

Eranskinak

A. eranskina

Datu-baseen laburpena

Eranskin honetan, ikasmaterialean zehar ariketetan eta adibideetan erabilitako datu-baseei buruzko argibide batzuk emango ditugu. Aurretik aipatu dugun moduan, datu base hauek `https://ehubox.ehu.eus/index.php/s/6C3gBHFiTME70Tx` web orrialdean daude eskuragai.

Hasteko, praktiketako adibideetan erabiltzen diren datu-baseen laburpena daukagu honako taula honetan, bakoitzaren izena, aldagai kualitatibo eta kuantitatibo kopurua, laginaren tamaina, eta agertzen diren adibideen erreferentziak adieraziz. Kontuan izan datu-baseak alfabetikoki ordenatuta agertzen direla:

Izena	**Aldagai kualitatiboak**	**Aldagai kuantitatiboak**	**Lagin tamaina**	**Adibideak**
Altuerak.sav	2	4	171	Sarrera
Altuerak.RData	2	4	171	1., 17.
praktikadatuak.xls/adimen	1	1	30	40.
praktikadatuak.xls/aluminio	0	4	8	52.
praktikadatuak.xls/birusa	0	1	90	12., 35.
praktikadatuak.xls/bonbillak	0	1	30	49.
praktikadatuak.xls/denbora	1	1	12	3., 19., 7., 26.
praktikadatuak.xls/drogasexua	2	1	18	45.
praktikadatuak.xls/efizientzia	0	3	8	43.
praktikadatuak.xls/energia	0	2	10	16., 47.
praktikadatuak.xls/errenta	0	1	50	36.
praktikadatuak.xls/ezbai	0	2	30	43.
praktikadatuak.xls/gh	1	0	60	38.
praktikadatuak.xls/jaurtiketa	1	0	40	5., 21.
praktikadatuak.xls/kutsadura	1	1	45	15., 44.
praktikadatuak.xls/lanaldimakina	2	1	16	46.
praktikadatuak.xls/mendel	1	0	500	10., 32.
praktikadatuak.xls/metodo	1	1	20	41.
praktikadatuak.xls/organismo	0	1	100	33.
praktikadatuak.xls/piezak	0	1	50	39.
praktikadatuak.xls/tomate	0	2	10	4., 20., 8., 27.
praktikadatuak.xls/ur	0	1	25	2., 18., 6., 24.
praktikadatuak.xls/zementu	1	1	80	48.

Aipatu bezala, datu-base hauek `https://ehubox.ehu.eus/index.php/s/6C3gBHFiTME7OTx` estekan daude eskuragai, erabili nahi dituen edonorentzat. Beraz, jarraian, datu-base hauetako bakoitza nola irakurri eta, behar denean, nola atondu ikusiko dugu. Gainera, datu-baseari buruzko informazio apur bat gehiago ere emango dugu.

Altuerak.sav

Gizabanako baten altuera eta erlazionatutako beste zenbait aldagai aztertzen ditu datu-base honek, zehazki, pisua, sexua, jatorrizko probintzia eta gurasoen altuerak. Datu-basea irakurri eta prestatzeko, erabili honako kode hau:

```
> library(foreign)
> altuerak <- read.spss("Altuerak.sav", to.data.frame=TRUE)
> #Aldagaien izenak minuskulaz idatzi
> colnames(altuerak) <- tolower(colnames(altuerak))
> #Jaioterria aldagaia faktore bihurtu
> altuerak$jaioterria <- factor(altuerak$jaioterria, levels=c(1, 2, 3),
+                               labels=c("Araba", "Bizkaia", "Gipuzkoa"))
```

Datu-basearen lehen lerroak ikusteko:

```
> head(altuerak, 4)

  aitarena amarena       sexua jaioterria altuera pisua
1      174     156 emakumezkoa    Bizkaia     165    65
2      177     159  gizonezkoa    Bizkaia     170    67
3      173     161  gizonezkoa      Araba     168    51
4      174     156  gizonezkoa    Bizkaia     167    69
```

praktikadatuak.xls/adimen

Klase batean adimen-proba bat egin da bi aldagai jasoz: lortutako puntuazioa (0-100 balioak) eta ikaslea errepikatzailea ala berria den gordetzen duen aldagai bat. Datu-base hau irakurtzeko:

```
> #Datuak inportatu
> library(readxl)
> adimena <- read_excel("praktikadatuak.xls", sheet="adimen")
```

Eta lehen lerroak ikusteko:

```
> head(adimena, 4)

  Ikaslea Adimena
1  berria      92
2  berria      12
3  berria      83
4  berria      36
```

praktikadatuak.xls/aluminio

Aluminio anodizatua ekoizten duen enpresa batek, prozesua ikuskatzean, metro karratuko akatsak zenbatu ditu.

```
> #Datuak inportatu
> library(readxl)
> aluminio <- read_excel("praktikadatuak.xls", sheet="aluminio")
```

```
> head(aluminio, 4)

  lagina ordua neurria industria.akatsak
1      1     7     2.1                 5
2      2     8     2.5                 5
3      3     9     0.8                 1
4      4    10     1.0                 3
```

praktikadatuak.xls/birusa

Birus baten latentzia-aldia ikertzeko, 90 txitari inokulatu zitzaien birusa. Bakoitzarengan, gaixotasunaren lehenengo sintomak agertu arte pasatutako egun kopurua aztertu zen.

Datu-basea kargatu eta lehen lerroak ikusteko:

```
> #Datuak inportatu
> library(readxl)
> birusa <- read_excel("praktikadatuak.xls", sheet="birusa")
> head(birusa, 4)

  birusa
1      8
```

```
2      5
3      7
4      9
```

praktikadatuak.xls/bonbillak

Bonbillak ekoizten dituen enpresa batean, 70 bonbillako 30 zorizko lagin hautatu ondoren, akastun bonbillen kopurua zenbatu du lote bakoitzean.

Datu-basea kargatu eta lehen lerroak ikusteko:

```
> #Datuak inportatu
> library(readxl)
> bonbillak <- read_excel("praktikadatuak.xls", sheet="bonbillak")
> head(bonbillak, 4)

  bonbilla.akastunak
1                  1
2                  2
3                  0
4                  3
```

praktikadatuak.xls/denbora

Lan zehatz bat egiteko, enpresa bateko langileen denbora-tarteak (segundotan) adierazi dira, haien sexua ere erregistratuz. Kontuan izan sexua aldagaia zenbakizko aldagai moduan dagoela definituta, eta faktore izatera pasatu behar dugula erabiltzen hasi baino lehen:

```
> #Datuak inportatu
> library(readxl)
> denbora <- read_excel("praktikadatuak.xls", sheet="denbora")
> #Aldagai kualitatiboak faktore bihurtu
> denbora$sexua <- factor(denbora$sexua, labels=c("emakumezkoa", "gizonezkoa"))
> head(denbora, 4)

  denbora       sexua
1     103 emakumezkoa
2      94 emakumezkoa
3     110 emakumezkoa
4      87 emakumezkoa
```

praktikadatuak.xls/drogasexua

Zirkulazio-istripuetan droga batzuek izaniko eragina ikertzeko, gidatzeko simulagailu batean jarri ziren hiru emakume eta hiru gizon, bakoitza tratamendu baten mende: marihuanaren eraginpean, alkoholaren eraginpean eta drogarik gabe. Simulagailuak 0 eta 35 puntu artean ematen ditu, eta puntuazio altuenak gidatzeko egoera onenekin erlazionatuta daude. Datu-baseak indibiduoaren sexua, jaso duen tratamendua, eta lortutako puntuazioa gordetzen ditu:

```
> drogasexua <- read_excel("praktikadatuak.xls", sheet='drogasexua')
> #Aldagai kualitatiboak faktore bihurtu
> drogasexua$sexua <- factor(drogasexua$sexua)
> drogasexua$droga <- factor(drogasexua$droga)
```

Lehen lerroak ikusteko:

```
> head(drogasexua, 4)

       sexua         droga puntuazioa
1 emakumezkoa     marihuana         19
2 emakumezkoa      alhohola          8
3 emakumezkoa drogarik gabe         21
4 emakumezkoa     marihuana         18
```

praktikadatuak.xls/efizientzia

8 egunetan zehar egindako probetan, A, B eta C katalizatzaileek izandako efizientzia neurtzen da:

```
> #Datuak inportatu
> efizientzia <- read_excel("praktikadatuak.xls", sheet="efizientzia")
> head(efizientzia, 4)

      A    B    C
1  84.5 78.4 63.1
2 102.8 79.1 79.9
3  99.1 78.0 67.8
4  80.2 76.0 52.9
```

praktikadatuak.xls/energia

Elektrizitate-konpainia batean, etxearen neurriaren (oin karratutan) eta etxebizitzaren hileko energia-kontsumoa (kwh-tan) jaso dira 10 etxetan:

```
> #Datuak inportatu
> energia <- read_excel("praktikadatuak.xls", sheet="energia")
> head(energia, 4)

  neurria kontsumoa
1    1290      1182
2    1350      1172
3    1470      1264
4    1600      1493
```

praktikadatuak.xls/errenta

1974. urteko per capita errenta jasotzen da:

```
> #Datuak inportatu
> errenta <- read_excel("praktikadatuak.xls", sheet="errenta")
> head(errenta, 4)

  errenta
1      55
2      55
3      65
4      65
```

praktikadatuak.xls/ezbai

Enpresa baten arabera, haiek ekoitzitako iragazkia baliagarria da erregai-kontsumoa murrizteko, autoen karburagailuaren hasieran kokatuta. Informazioa kontrastatzeko, 30 auto hautatu ziren eta bakoitzaren kontsumoa neurtu zen l/100 km-tan, iragazkirik gabe (ez) eta iragazkiarekin (bai).

```
> #Datuak inportatu
> ezbai <- read_excel("praktikadatuak.xls", sheet="ezbai")
> head(ezbai, 4)

   ez bai
1 6.8 6.4
2 7.0 6.5
3 7.2 7.3
4 9.0 8.8
```

praktikadatuak.xls/gh

Inkesta batean 60 pertsona hautatu dira, gazteak (G) eta helduak (H). Datu-base honek hautaketaren ordena adierazten du:

```
> #Datuak inportatu
> gh <- read_excel("praktikadatuak.xls", sheet="gh")
> gh$gazteheldu <- factor(gh$gazteheldu)
> head(gh, 4)

  gazteheldu
1          H
2          G
3          H
4          H
```

praktikadatuak.xls/jaurtiketa

Suzirien aireratze-instalazio berri bat ikertzen ari dira, eta 40 jaurtiketa esperimental egin dira, horietatik arazorik zeinek ez duten izan (a, arrakasta) eta arazoak zeinek izan dituzten (p, porrota) adieraziz.

```
> #Datuak inportatu
> jaurtiketa <- read_excel("praktikadatuak.xls", sheet="jaurtiketa")
> jaurtiketa$arrakasta <- factor(jaurtiketa$arrakasta)
> head(jaurtiketa, 4)

  arrakasta
1         a
2         a
3         a
4         a
```

praktikadatuak.xls/kutsadura

Hiru herrialdetako (A, B, C) kutsadura-maila alderatu nahi da. Horretarako, diametroa 10 mm baina txikiagoa duten aireko partikula kontzentrazioaren (PM10 partikulak) neurketa bat egin da herrialde bakoitzeko 15 neurketa-estaziotan:

```
> #Datuak inportatu
> kutsadura <- read_excel("praktikadatuak.xls", sheet="kutsadura")
```

```
> kutsadura$herrialdea <- factor(kutsadura$herrialdea)
> head(kutsadura, 4)

  kontzentrazioa herrialdea
1          30.36          A
2          28.19          A
3          28.16          A
4          28.10          A
```

praktikadatuak.xls/lanaldimakina

Termometroak ekoizten dituen lantegi batean, marka desberdineko lau makina erabiltzen dira, non makina bakoitza lau lanalditan erabiltzen baitute. Datu-baseak egunero lanaldiko eta makinako ekoitzitako unitate kopurua gordetzen du.

```
> langilemakina <- read_excel("praktikadatuak.xls", sheet='lanaldimakina')
> #Aldagai kualitatiboak faktore bihurtu
> lanaldimakina$lanaldia <- factor(lanaldimakina$lanaldia)
> lanaldimakina$makina <- factor(lanaldimakina$makina)
> head(lanaldimakina, 4)

  lanaldia makina unitateak
1        1      A        14
2        1      B         9
3        1      C         7
4        1      D         8
```

praktikadatuak.xls/mendel

Mendelen legeak esperimentalki egiaztatzeko, 500 landare gurutzatu genituen, bakoitzaren kolorea gordez. Kolore posibleak gorria, arrosa, horia eta zuria dira:

```
> mendel <- read_excel("praktikadatuak.xls", sheet='mendel')
> #Aldagai kualitatiboak faktore bihurtu
> mendel$kolorea <- factor(mendel$kolorea)
> head(mendel, 4)

  kolorea
1  gorria
2  gorria
3  gorria
4  gorria
```

praktikadatuak.xls/metodo

Pedagogo batek konparatu nahi ditu irakasteko honako hiru metodo hauek: (1) online, (2) erdipresentziazkoa eta (3) presentziazkoa. Horretarako, ikasturte batean zehar hiru metodoekin irakasten den ikasgai bat hautatu du; eta metodo bakoitzarekin zorizko lagin bakun bat aukeratu du, amaierako notak gordez:

```
> #Datuak inportatu
> library(readxl)
> metodo <- read_excel("praktikadatuak.xls", sheet="metodo")
> metodo$metodoa <- factor(metodo$metodoa)
> head(metodo, 4)

  puntuazioa metodoa
1         78  online
2         80  online
3         65  online
4         57  online
```

praktikadatuak.xls/organismo

Urmael batetik ateratako 100 laginen organismo kopurua adierazten da:

```
> #Datuak inportatu
> library(readxl)
> organismo <- read_excel("praktikadatuak.xls", sheet="organismo")
> head(organismo, 4)

  organismo
1         0
2         0
3         0
4         0
```

praktikadatuak.xls/pieza

Kalitate-ikuskatzaile batek, langileek egiten dituzten piezen kopurua ikertzeko asmoz, 50 egunean zehar langile batek egindako kopurua jaso du (10 unitateka neurtua):

```
> #Datuak inportatu
> library(readxl)
```

```
> piezak <- read_excel("praktikadatuak.xls", sheet="pieza")
> head(piezak, 4)

  pieza.akastunak
1             100
2             110
3              80
4              75
```

praktikadatuak.xls/tomate

Espektofotometriaren bidez, tomate freskoen eta ontziraturiko tomateen nahitaezko elementuak ikertu dira. Horretarako, kobre kopurua konparatu da tomate freskoetan eta tomate berberetan, haiek ontziratu ondoren:

```
> library(readxl)
> tomate <- read_excel("praktikadatuak.xls", sheet='tomate')
> head(tomate, 4)

  freskoa latakoa
1   0.066   0.085
2   0.079   0.088
3   0.069   0.091
4   0.076   0.096
```

praktikadatuak.xls/ur

Hiri txiki batean, ur-erabilerari buruzko ikerkuntza batean, 25 etxebizitzatako zorizko lagina atera da, *asteko erabilitako ur litro kopurua* aldagaia jasoz.

```
> library(readxl)
> ur <- read_excel("praktikadatuak.xls", sheet='ur')
> head(ur, 4)

  ura
1 175
2 185
3 186
4 118
```

praktikadatuak.xls/zementu

Zementu-fabrika batean, 9 ordu eta erditan zehar, ordu erdian behin lau zaku zementu hautatu dira zoriz, eta haien pisuak erregistratu dira.

```
> library(readxl)
> zementu <- read_excel("praktikadatuak.xls", sheet='zementu')
> #Aldagai kualitatiboa faktore bihurtu
> zementu$lagina <- factor(zementu$lagina)
> head(zementu, 4)

  lagina pisua
1      1    53
2      1    51
3      1    52
4      1    49
```

Azkenik, autoebaluazioan erabiliko diren datu-baseen laburpen-taula bat emango dugu:

Izena	Aldagai kualitatiboak	Aldagai kuantitatiboak	Lagin tamaina	Adibideak
Altuerak.RData	2	4	171	9., 13., 16., 21.
praktikadatuak.xls/AUTOegurra	0	2	24	28.
praktikadatuak.xls/AUTOespeziea	1	1	30	20.
praktikadatuak.xls/AUTOkutsadura	0	1	57	1.
praktikadatuak.xls/AUTOlekugari	2	1	27	25.
praktikadatuak.xls/AUTOlurongarri	2	1	20	24.
praktikadatuak.xls/AUTOmatraze	0	2	25	27.
praktikadatuak.xls/AUTOnitrogeno	1	1	20	2.
praktikadatuak.xls/AUTOpozoia	1	1	15	23.
praktikadatuak.xls/AUTOpresioa	0	2	5	26.
praktikadatuak.xls/AUTOsasoia	1	1	24	19.
praktikadatuak.xls/AUTOzuhaitzak	0	1	10	8.

B. eranskina

Erabilitako *R* paketeak

Honako taula honetan, ikasmaterialean zehar erabilitako *R*ko paketeen laburpen bat ematen da, izena, berau erabiltzen edo aipatzen den praktiken zerrenda, eta pakete horri emango diezaiokegun erabileraren azalpen bat emanez.

Izena	**Praktikak**	**Erabileraren laburpena**
car	7. *R* praktika	Erregresiorako tresnak eskaintzen dizkigu. Besteak beste, ereduen diagnosirako Durbin-Watson testa dauka inplementatuta.
EnvStats	3. *R* praktika	Bariantzarako konfiantza-tartea eraikitzeko funtzio bat dauka eskuragai (`varTest`).
fBasics	1. *R* praktika	Oinarrizko estatistiko deskribatzaile zerrenda bat lortzeko `basicStats` funtzioa dauka.
foreign	0. eta 1. *R* praktikak	*R*-ren berezko formatuetan ez dauden datu-fitxategiak irekitzeko balio du, besteak beste, SPSSn sortutako fitxategiak.
ggm	7. *R* praktika	Aldagaien arteko korrelazio partzialak lortzeko erabil dezakegu.
Hmisc	7. *R* praktika	Korrelazioen testak egiteko erabil dezakegu.
lawstat	5. *R* praktika	Bolada-testa egiteko `runs.test` funtzioa dauka eskuragai.
nortest	5. *R* praktika	Normaltasun-test aukera zabala eskaintzen du, besteak beste, Lilliefors (Kolmogorov-Smirnov) testa.
multcomp	6. *R* praktika	Bi populazio baino gehiagoren batezbestekoak konparatzeko bariantza-analisia egin ostean, binakako konparaketak egiteko tresnak eskaintzen ditu.
pwr	4. *R* praktika	Hipotesi-kontrasteen ahalmena aztertzeko tresnak eta funtzioak ditu.
qcc	8. *R* praktika	Kalitate-kontrola egiteko tresnak inplementatzen ditu.
Rcmdr	I. atala	*R* software-aren interfaze grafiko bat eskaintzen du.
RcmdrMisc	6. *R* praktika	*Rcmdr*ko zenbait funtzio berezi gordetzen ditu. Besteak beste, batezbestekoen konfiantza-tarteak irudikatzeko funtzio bat eskaintzen digu (`plotMeans`).
readxl	II. atala	Microsoft Excel-en sortutako fitxategiak *R*n ireki eta kudeatzeko tresnak eskaintzen ditu.
scmamp	5. *R* praktikan	Hipotesi-kontraste ez-parametrikoen aukera zabala eskaintzen du.

C. eranskina

Autoebaluaziorako ariketak

Ondorengo ataletan, autoebaluazio-ariketak zerrendatzen dira, gaika. Kontuan izan [Z] sinboloa duten ariketak zailtasun-maila handiagokoak direla, eta ikasmaterial honen bigarren ataleko ezagutzak erabiltzea eskatzen dutela.

C.1. Estatistika deskribatzailea

1. ariketa. *Airearen kutsadura aztertzeko, 57 hiritan aztertu zen esekidura-materiaren partikula-edukia (mikrogramo/m^3-tan), eta honako emaitza hauek lortu ziren:*

68	*63*	*42*	*27*	*30*	*36*	*28*	*32*	*79*	*27*	*22*	*23*	*24*	*25*	*44*
65	*43*	*25*	*74*	*51*	*36*	*42*	*28*	*31*	*28*	*25*	*45*	*12*	*57*	*51*
12	*32*	*49*	*38*	*42*	*27*	*31*	*50*	*38*	*21*	*16*	*24*	*69*	*47*	*23*
22	*43*	*27*	*49*	*28*	*23*	*19*	*46*	*30*	*43*	*49*	*12*			

Gogoratu datu hauek ***praktikadatuak.xls*** *fitxategiko* ***AUTOkutsadura*** *orrialdean daudela eskuragai. Datuak inportatu edo eskuz sartu ostean:*

a) Irudika ezazu aireko esekidura-materiaren partikula-edukien banaketa, histograma (14 tarte) eta kutxa-diagrama erabiliz. Grafikoak ikusi ondoren, zer esan daiteke simetriari buruz? Ondorio berberera heltzen al gara?

b) Zein da simetria aztertzeko estatistikoa? Zenbat balio du? Zer ondorioztatzen da?

c) Balio arrarorik badago? Zein?

d) Kalkula eta interpreta itzazu joera zentraleko estatistikoak.

e) Zein da behaketa guztien tartearen luzera? Nola deritzo estatistikoari?

f) Zein da behaketa zentralen tartearen luzera? Nola deritzo estatistikoari?

g) Zein da laginaren kuasi-desbiderapen estandarra eta kuasibariantza? Zertarako dira?

h) Zein balioren artean egongo da hiri kutsatuenetariko %25a? Zer erabili duzu erantzuteko?

i) Zein balioren artean egongo da gutxien kutsatutako hirien %10a? Zer erabili duzu erantzuteko?

2. ariketa. *Lur mota berberean hogei zuhaitz landatu ziren, eta egoera berdinetan zaindu (zuhaitz bakoitzak eguzki eta ur kantitate berbera jaso zuen). Aldiz, landatzean, zuhaitzen erdiak ez zuen nitrogenorik jaso (kontrol moduan) eta beste erdiak bai. 140 egun igaro ondoren, honako hauek izan ziren enborren pisuen balioak (gramotan). Egin ezazu grafiko bat nitrogenoa jaso ez zutenen eta jaso zutenen enborren batez besteko pisuak konparatzeko (jo dezagun banaketa normalak zirela).*

Nitrogeno gabe	*0,32*	*0,53*	*0,28*	*0,37*	*0,47*	*0,43*	*0,36*	*0,42*	*0,38*	*0,43*
Nitrogenoarekin	*0,26*	*0,43*	*0,47*	*0,49*	*0,52*	*0,75*	*0,79*	*0,86*	*0,62*	*0,46*

Datuak inportatu nahi baditugu, ***praktikadatuak.xls*** *fitxategiko* ***AUTOnitrogeno*** *orrialdera jo.*

C.2. Probabilitatearen teoria

3. ariketa. *Aseguru-etxe bateko langile batek bost aseguru-poliza saldu dizkie adin berdineko bost gizabanakori. Taula aktuarialen arabera, adin horretako gizabanako batek 30 urte baino gehiago bizitzeko probabilitatea 3/5ekoa da. Kalkula itzazu (a) 30 urte barru gutxienez hiru pertsona bizirik egoteko probabilitatea, eta (b) 30 urte barru gehienez bi pertsona bizirik egoteko probabilitatea.*

4. ariketa. *Gasolina-zerbitzugune batera iristen diren automobil kopurua, batez beste, orduko 204koa da. Baldin zerbitzugune horrek gehienez minutuko hamar automobil zerbitza baditzake, kalkula ezazu minutu zehatz batean zerbitza daitezkeenak baino automobil gehiago iristeko probabilitatea.*

5. ariketa. *Populazio batean, % 0,004koa da 12.000 euro baino gehiago kobratzen duen gizabanako kopurua. Kalkula ezazu, aztertutako 5.000 gizabanakoren artean, gehienez bi pertsonak aipatutako kantitatea kobratzeko probabilitatea, kontsultatutako guztiek erantzuten dutela jorik.*

6. ariketa. *Fabrikatze-prozesu batean, ezaguna da eguneroko akastun unitate kopurua 10 parametroko poisson banaketari darraiola. Kalkula ezazu (a) 150 egunetan akastun unitate kopurua 1.480 baino handiagoa izateko probabilitatea, eta (b) kopuru hori 1.480 eta 1.520 artean egotekoa.*

7. ariketa. *Izan bedi $X : \mathcal{N}(1500,\ 38,7)$ zorizko aldagaia. (a) Zein da aldagaiaren balioa, non banaketa-funtzioa 0,5 baita? (balio horri mediana esaten zaio) eta (b) Zein da aldagaiaren balioa, non banaketa-funtzioa 0,25 baita? (balio horri 1. kuartila esaten zaio).*

C.3. Konfiantza-tartezko zenbatespena

8. ariketa. *10 tamainako lagin batean, honako hauek dira zuhaitzen diametroak (cm-tan): 97, 117, 140, 78, 99, 148, 108, 135, 126, 121. Kalkula ezazu zuhaitzen batez besteko diametroa, % 95eko konfiantza-mailarekin, populazioaren banaketa normala dela suposatuz (datuak eskuragai daude* ***praktikadatuak.xls/AUTOzuhaitzak*** *fitxategian).*

9. ariketa. *Ireki ezazu* `Altuerak.RData` *fitxategia.*

a) Zein da batez besteko pisuaren %95eko konfiantza-tartea? Ondoriozta al daiteke batez besteko pisua 68 kg baino altuagoa dela?

b) Zein da emakumezkoen batez besteko pisuaren %95eko konfiantza-tartea? Ondoriozta al daiteke emakumezkoen batez besteko pisua 68 kg baino altuagoa dela?

c) Ondoriozta al daiteke gizonezkoen eta emakumezkoen pisuaren bariantzen arteko diferentzia dagoela %95eko konfiantza-mailaz?

d) Ondoriozta al daiteke gizonezkoen eta emakumezkoen pisuaren batezbestekoen arteko diferentzia dagoela %95eko konfiantza-mailaz?

e) Zein da emakumezkoen proportziorako %95eko konfiantza-tartea? Ondoriozta al daiteke proportzio berdinean daudela emakumezkoak eta gizonezkoak?

f) Ondoriozta al daiteke aiten batez besteko altuera haien umeena baino baxuagoa dela, %95eko konfiantza-mailaz?

10. ariketa. *[Z] Enpresa batek, instalazioak gehitzeko beharra ikertzeko asmoz, jasotzea espero duen eskaria zenbatetsi nahi du. Horretarako, ohiko bezeroen artean hamaika hautatzen ditu, eta honako taula honetan adierazten dira azken urtean haiek eskatutako unitate kopuruak:*

unitate kopurua (x_i)	*1000*	*1002*	*1004*	*1006*	*1008*	*1010*	*1012*
bezero kopurua (f_i)	*1*	*2*	*1*	*3*	*1*	*2*	*1*

Laginean oinarrituta, sortu beharrezko datu-basea, eta, suposatuz eskaria banaketa normalari darraiola, erantzun honako galdera honi:
(a) Ondoriozta ezazu ea % 90eko konfiantza-tartean posiblea den populazioaren batezbestekoa 1005 *unitate izatea.*
(b) Ondoriozta ezazu % 95eko konfiantza-tartean oinarrituz ea posiblea den bariantza 5 unitate baino txikiagoa izatea.

11. ariketa. *[Z] 1.000 etxeko zorizko lagin batean, ikusi da 228tan butanoa erabiltzen dela. Kalkula ezazu butanoa erabiltzen duten etxe-proportzioaren % 99 mailako konfiantza-tartea, banaketa binomial zehatza erabiliz.*

12. ariketa. *[Z] Probintzia batean zentral nuklear bat eraikitzearen aldeko iritzia ezagutzeko, honako datu hauek aurkitu ziren: kostaldeko 400 biztanleren artean 168 zentralaren alde daude; barrualdeko 500 biztaleren artean, berriz, 145 dira aldekoak. Kalkula ezazu % 95eko konfiantza-tartea kostaldean eta barrualdean, zentral nuklearraren aldeko biztanle-proportzioen arteko diferentziarako. Zein ondorio atera dezakegu?*

C.4. Hipotesi-kontraste parametrikoak

13. ariketa. *Ireki ezazu* `Altuerak.RData` *fitxategia.*

a) Ondoriozta al daiteke batez besteko pisua 68 kg baino altuagoa dela?

b) Ondoriozta al daiteke batez besteko pisua zehatz-mehatz 68 kg dela?

c) Ondoriozta al daiteke emakumezkoen batez besteko pisua 68 kg-tik gorakoa dela?

d) Ondoriozta al daiteke gizonezkoen eta emakumezkoen pisuaren bariantzen arteko diferentzia dagoela, %5eko esangura-mailaz?

e) Ondoriozta al daiteke gizonezkoen eta emakumezkoen pisuaren batezbestekoen arteko diferentzia dagoela, %5eko esangura-mailaz?

f) Ondoriozta al daiteke proportzio berdinean daudela emakumezkoak eta gizonezkoak, %5eko esangura-mailaz?

g) Emakumezkoen ehunekoa % 60 baino txikiagoa dela ondoriozta daiteke, %5eko esangura-mailaz?

h) Gizonezkoen ehunekoa % 70 baino handiagoa edo berdina dela ondoriozta daiteke, %5eko esangura-mailaz?

i) Ondoriozta al daiteke aiten batez besteko altuera haien umeena baino baxuagoa dela, %5eko esangura-mailaz?

j) [Z] Onar al daiteke Gipuzkoan eta Araban emakumezkoen proportzioa berdina denik, %5eko esangura mailaz?

C.5. Hipotesi-kontraste ez-parametrikoak

14. ariketa. *Zoriz aukeraturiko 200 haziri aplikatutako 3 tratamentu kimiko probatzean, hozitze-probak eginez, honako* 3×2 *kontingentzia-taula honetan adierazitako emaitzak lortu ziren.*

X / Y	*A*	*B*	*D*
Hozitu zirenak	*190*	*170*	*180*
Hozitu ez zirenak	*10*	*30*	*20*

Kontrasta ezazu tratamentu kimikoak hozitzean eragina daukan ala ez.

15. ariketa. *Bi marka desberdinetako enpresetan ekoizten dituzten patata frijituen poltsen kalitate-kontrola egin ondoren, honako datu hauek lortu ziren. Onar daiteke, %1eko esangura-mailarekin, marka mota kalitatearekiko askea dela?*

X / Y	*1*	*2*
Kalitate-kontrola ez gainditu	*89*	*37*
Kalitate-kontrola gainditu	*297*	*197*

(a) Azter ezazu hipotesi-kontrastea, khi karratuaren testa erabiliz
(b) [Z] Azter ezazu hipotesi-kontrastea, proportzioen metodoa erabiliz.

16. ariketa. *Ireki ezazu* `Altuerak.RData` *izeneko fitxategia.*
(a) Onar daiteke % 5eko esangura-mailarekin, 3 lurraldeetako emakumezko eta gizonezkoen proportzioak berdinak direnik? Zein proba egin duzu? Betetzen al dira beharrezko baldintzak?
(b) Zein dira Gipuzkoako emakumeen behatutako eta itxarondako maiztasunak? Zenbat pertsona dira Gipuzkoakoak?
(c) Gipuzkoako biztanleen artean zein ehuneko dira emakumezkoak? Eta Bizkaiko biztanleen artean?
(d) Emakumezkoen artean, zein ehuneko dira arabarrak?
(e) Eta onar daiteke hipotesi nulua $H_0 : (p_1, p_2, p_3) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $p_1 =$ *Arabakoa izatea,* $p_2 =$ *Bizkaikoa izatea eta* $p_3 =$ *Gipuzkoakoa izatea direlarik?*
(f) Onar daiteke, % 5eko esangura-mailarekin, altuera *aldagaia banaketa normalari darraiola?*

17. ariketa. *[Z] Ebatz ezazu 3. adibidea, kontraste ez-parametrikoak erabiliz.*

18. ariketa. *[Z] Ebatz ezazu 4. adibidea, kontraste ez-parametrikoak erabiliz.*

19. ariketa. *[Z] 1964-1969 urteetan jaso ziren honako datu hauek. Urteka, heriotza-indizeak independenteak direla jotzen da. Zehaztu ezazu urte-sasoiaren arabera heriotza-indizeen desberdintasuna adierazgarria den ala ez, %5eko esangura-mailaz.*

Urte-sasoia	*Behaketak*
Udaberria	*9, 9,3, 9,3, 9,2, 9,4, 9,1*
Uda	*8,8, 8,7, 8,8, 8,6, 8,7, 8,3*
Udazkena	*9,4, 9,4, 10,3, 9,8, 9,4, 9,6*
Negua	*10,6, 9,8, 10,9, 10,2, 9,7, 9,9*

*Gogoratu datu hauek **praktikadatuak.xls** fitxategiko **AUTOsasoia** orrialdean daudela eskuragai. Datuak inportatu edo eskuz sartu ostean, ebatz ezazu ariketa, kontraste ez-parametrikoak erabiliz.*

20. ariketa. *[Z] Pinu multzo baten altuera (metrotan) eta espeziea ezagutuz, altueraren arabera pinu guztiak berdintzat har daitezke?*

Pinea	*8,52*	*Pinaster*	*8,52*	*Pinea*	*8,13*
Pinaster	*6,45*	*Pinea*	*6,43*	*Halapensis*	*7,17*
Silvestris	*7,41*	*Pinea*	*6,21*	*Pinaster*	*8,40*
Pinea	*7,15*	*Halapensis*	*7,07*	*Silvestris*	*8,87*
Pinaster	*8,73*	*Pinaster*	*8,83*	*Pinea*	*6,12*
Laricio	*7,55*	*Pinaster*	*8,53*	*Pinaster*	*8,91*
Halapensis	*6,54*	*Laricio*	*7,84*	*Silvestris*	*8,81*
Laricio	*7,74*	*Silvestris*	*8,59*	*Laricio*	*7,40*
Silvestris	*8,65*	*Laricio*	*7,41*	*Pinaster*	*8,19*
Silvestris	*8,81*	*Pinaster*	*8,94*	*Pinaster*	*8,56*

*Gogoratu datu hauek **praktikadatuak.xls** fitxategiko **AUTOespeziea** orrialdean daudela eskuragai. Datuak inportatu edo eskuz sartu ostean, ebatz ezazu kontrastea, metodo ez-parametrikoak erabiliz.*

21. ariketa. *[Z] Ireki ezazu `Altuerak.RData` izeneko fitxategia. Friedmann testaren bidez, azter ezazu ea gizabanakoen batez besteko altuera haien amen eta aiten altueraren berdina den.*

C.6. Bariantza-analisia

22. ariketa. *19. ariketako datuetan oinarrituz, erantzun honako galdera hauei:*

(a) Azter ezazu indizeen normaltasuna urte-sasoiaren arabera. Zer ondoriozta daiteke %1eko esangura-mailaz?

(b) Azter ezazu bariantzen homogenotasuna %1eko esangura-mailaz.

(c) Zehaztu ezazu ea urte-sasoiaren arabera heriotza-indizeen desberdintasuna adierazgarria den ala ez, %5eko esangura-mailaz.

(d) Zer ondoriozta daiteke, %5eko esangura-mailaz, batezbestekoen alderaketa binakatuei buruz?

23. ariketa.

Hiru arrazatako arratoiak ditugu: A, B eta C. Jakin nahi dugu ea pozoi baten aurrean erresistentzia berdina duten ala ez. Erantzun-aldagaia zera da: pozoia jaso duten 100 arratoiren artean hildako kopurua. Arraza mota bakoitzerako 5 neurketa egin dira.

Arraza mota		
A	*B*	*C*
30	*85*	*40*
20	*73*	*28*
35	*92*	*39*
42	*86*	*41*
60	*75*	*50*

Gogoratu datu hauek ***praktikadatuak.xls*** *fitxategiko* ***AUTOpozoia*** *orrialdean daudela eskuragai.*

(a) Azter ezazu erresistentziaren normaltasuna, arraza motaren arabera. Zer ondoriozta daiteke %5eko esangura-mailaz?

(b) Azter ezazu bariantzen homogenotasuna, %5eko esangura-mailaz.

(c) Zehaztu ezazu ea arraza motaren arabera erresistentziaren desberdintasuna adierazgarria den ala ez, %5eko esangura-mailaz.

(d) Zer ondoriozta daiteke, %5eko esangura-mailaz, batezbestekoen alderaketa binakatuei buruz? Zein(tzuk) d(ir)a erresistentzia handiena dauka(te)n arraza mota(k)

24. ariketa. *[Z] Patata-ekoizpenarekiko lau ongarrien arteko desberdintasunak ikertzeko asmoz, bost finka hartu ziren. Finka bakoitza tamaina eta mota bereko lau lurzatitan azpizatitu zen. Zoriz, finka bakoitzeko lurzati bakoitzean banatu ziren ongarriak. Ongarrien arabera edo finken arabera, esan daiteke desberdintasunak adierazgarriak direla? Honako hauek izan ziren etekinen datuak (tonatan):*

Etekina	*Ongarria*			
Lurzatia	*1*	*2*	*3*	*4*
1	*2,1*	*2,2*	*1,8*	*2,1*
2	*2,2*	*2,6*	*1,9*	*2,0*
3	*1,8*	*2,7*	*1,6*	*2,2*
4	*2,0*	*2,5*	*2,0*	*2,4*
5	*1,9*	*2,8*	*1,9*	*2,1*

Gogoratu datu hauek ***praktikadatuak.xls*** *fitxategiko* ***AUTOlurongarri*** *orrialdean daudela eskuragai.*

25. ariketa. *[Z] Hiru gari mota landatu dira hiru lekuetan. Hazi eta gero, mota bakoitzeko lagin bat hartu da (guztira 9), irina lortzeko. Lagin bakoitzeko irinarekin, hiruna ogi-barra egin dira. Hurrengo taulan agertzen dira lortutako 27 ogi-barren bolumenak (unitate estandarretan). % 1eko esangura-mailarekin, kontrasta ezazu ea gari motak edo lekuak eraginik badaukan ogi-barren batez besteko bolumenean. Esan daiteke bi faktoreen arteko interakzioarik dagoenik?*

Bolumena	*Gari mota*		
Lekua	*1*	*2*	*3*
1	*15,2 13,8 14,3*	*13,4 16,5 15,2*	*20,4 18,2 16,3*
2	*7,6 4,8 3,9*	*4,8 2,7 3,9*	*12,2 11,8 13,4*
3	*19,2 17,5 18,4*	*11,3 13,4 12,6*	*22,3 25,1 24,2*

Gogoratu datu hauek ***praktikadatuak.xls*** *fitxategiko* ***AUTOlekugari*** *orrialdean daudela eskuragai.*

C.7. Erregresioa

26. ariketa. *Material isolatzaile berri baten asmatzaileak presio batzuen eraginpean egongo den bi hazbeteko lodieradun ale baten konpresioa neurtu nahi du. Horretarako, bost ale hartzen dira kontuan. Honako taula honetan agertzen dira presioak (hazbete koadroko 10 libratan neurtuta) eta lortutako konpresioak (0,1 hazbetetan):*

Ale	*1*	*2*	*3*	*4*	*5*
Presioa	*1*	*2*	*3*	*4*	*5*
Konpresioa	*1*	*1*	*2*	*2*	*4*

Gogoratu datu hauek ***praktikadatuak.xls*** *fitxategiko* ***AUTOpresioa*** *orrialdean daudela eskuragai.*

(a) Plantea itzazu presioarekiko konpresioaren eredu lineala, hiperbolikoa, koadratikoa, potentziala eta esponentziala. % 5eko esangura-mailarekin, zein eredu dira erabilgarriak?
(b) Ordena itzazu erabilgarriak direnak, doikuntza-egokitasunaren arabera; 1.a izango da egokiena.
(c) Idatz itzazu erregresio-koefizienteak eredu linealean eta esponentzialean.
(d) Har ezazu $Y = Ae^{BX} + \epsilon$ eredu esponentziala. Azter ezazu doikuntza-egokitasuna.
(e) Zein da Bren % 95eko konfiantza-tartea? Zer ondoriozta daiteke?
(f) Zein da Arena? Ondoriozta daiteke $A = 1$ dela hipotesi-kontraste baten bidez?
(g) Ereduaren arabera, zein behaketari dagokio ondoen dagoen iragarritako balioa? Zenbat da?
(h) Ereduaren arabera, zein da 4. behaketaren iragarritako balioarena, % 95eko konfiantza-mailarekin?

C.8. Kalitatearen kontrol estatistikoa [Z]

27. ariketa. *Produkzio-prozesu batean, 100 matrazetako 25 lagin hartu dira, eta honako hauek dira akastun matrazeen kopuruak: 14, 11, 13, 12, 5, 8, 14, 21, 13, 2 ,6, 1, 17, 13, 18, 5, 7, 5, 4, 13, 2, 14, 4, 11, 1. Zenbatets itzazu ekoitzitako akastun matrazeen ehunekoa eta akastun matrazeen kopuru osoa, eta egin eta interpreta itzazu kontrol-grafikoak.*

Gogoratu datu hauek ***praktikadatuak.xls*** *fitxategiko* ***AUTOmatraze*** *orrialdean daudela eskuragai.*

28. ariketa. *Egurrezko enpresa batean, ekoitzitako oholtzarren akatsen kalitate-kontrola egin nahi da. 24 oholtzarreko lagina hautatu ondoren, honako hau izan da pieza bakoitzaren akatsen kopurua: 7, 6, 8, 10, 24, 6, 5, 4, 8, 11, 15, 8, 4, 16, 11, 12, 8, 6, 5, 9, 7, 14, 8, 21. Egin ezazu oholtzarreko akatsen kopururako grafikoa.*

Gogoratu datu hauek ***praktikadatuak.xls*** *fitxategiko* ***AUTOegurra*** *orrialdean daudela eskuragai.*

29. ariketa. *Estatu Batuetan, Lan Sailean urtero porrot egiten duten enpresak aztertzen eta sailkatzen dira honako kategoria hauetan: (1) produzkioan esperientzarik eza, (2) gerentzian esperientzarik eza, (3) esperientzia desorekatua, (4) gaitasunik eza, (5) beste kausa batzuk (adibidez, arduragabekeria, iruzurra, hondamen naturala) eta (6) kausa ezezagunak. Eraiki ezazu Paretoren diagrama, 1.463 enpresa eraikitzaileren porrotak aztertzean honako emaitza hauek lortu badira:*

Kategoria mota	(1)	(2)	(3)	(4)	(5)	(6)
Kasu kopurua	113	234	317	695	19	85

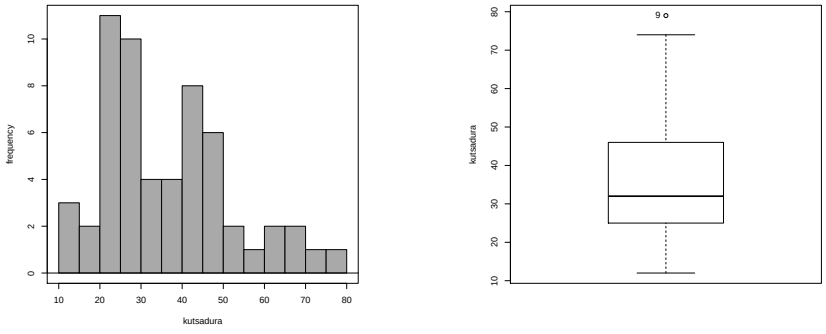
D. eranskina

Autoebaluaziorako ariketen emaitzak

Eranskin honetan, gai bakoitzerako autoebaluaziorako ariketen emaitzak agertzen dira. Kontuan hartu autoebaluazio hauek gauzatzeko beharrezko R kodea eskuragai dagoela `https://ehubox.ehu.eus/index.php/s/6C3gBHFiTME70Tx` orrialdean.

D.1. Estatistika deskribatzailea

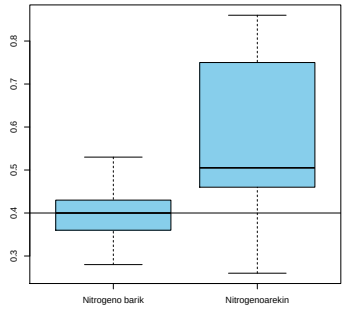
1. ariketa



(a) Eskuinalderantz asimetrikoa.
(b) Asimetria-koefizientea $0,761 > 0$, (**fBasics** paketeko `basicStats` funtzioa erabiliz, $0,721667 > 0$) eskuinalderantz asimetrikoa.
(c) 9. behaketa, hau da, $79\mu g/m^3$.
(d) $\overline{x} = 36,7193\mu g/m^3$, $Me = 32\mu g/m^3$, $Mo = 27\mu g/m^3$ eta $28\mu g/m^3$.
(e) $79 - 12 = 67\mu g/m^3$, heina.
(f) $RI = 21\mu g/m^3$, kuartilarteko heina.
(g) $s_{n-1}^2 = 251,706(\mu g/m^3)^2$ eta $s_{n-1} = 15,86523\mu g/m^3$.

(h) $p_{75} = 46\mu g/m^3$ eta $max = p_{100} = 79\mu g/m^3$ artean.
(i) $min = p_0 = 12\mu g/m^3$ eta $= p_{10} = 21,6\mu g/m^3$ artean.

2. ariketa



D.2. Probabilitatearen teoria

3. ariketa Izan bedi $X : Bin(5,\ 0,6)$ z.a.d,
(a) $P(X \geq 3) = P(X > 2) = 0,68256$
(b) $P(X \leq 2) = F(2) = 0,31744$

4. ariketa Izan bedi $X : \mathcal{P}(3,4)$ z.a.d,
$P(X > 10) = 0,0008101702$

5. ariketa Izan bedi $X : Bin(5000,\ 0,00004)$,
$P(X \leq 2) = 0,9988521$
Hurbilketa: $Bin(5000,\ 0,00004) \approx \mathcal{P}(0,2)$ denez,
$P(X \leq 2) \approx 0{,}9988515$

6. ariketa Izan bedi $\Sigma_n : \mathcal{N}(1500,\ 38,7298)$ z.a.j,
(a) $P(\Sigma_n > 1480) = 0{,}6972118$
(b) $P(1480 < \Sigma_n < 1520) = P(\Sigma_n > 1480) - P(\Sigma_n \geq 1520) = 0{,}3944237$

7. ariketa Izan bedi $X : \mathcal{N}(1500,\ 38,7)$ z.a.j,
(a) $F(a) = 0,5 \Rightarrow a = F^{-1}(0,5) = 1500$
(b) $F(b) = 0,25 \Rightarrow b = F^{-1}(0,25) = 1473{,}897$

D.3. Konfiantza-tartezko zenbatespena

8. ariketa $I_\mu^{0,95} = (101,3788,\ 132,4212)$

9. ariketa
a) $I_{mu}^{0,95} = (69,725,\ 75,227)$ Bai
b) $I_{mu_e}^{0,95} = (61,530,\ 69,038)$ Ez
c) $I_{\sigma_g^2/\sigma_e^2}^{0,95} = (0,609,\ 1,450)$ Ez
d) $I_{\mu_g-\mu_e}^{0,95} = (8,792,\ 18,985)$ Bai, emakumezkoak pisu txikiagoa.
e) $I_p^{0,95} = (0,403,\ 0,557)$ Bai
f) $I_{\mu_a-\mu_u}^{0,95} = (4,983,\ 5,831)$ Ez, kontrakoa.

10. ariketa
(a) $I_{\mu}^{0,90} = (1003,692,\ 1008,580)$ Bai
(b) $I_{\sigma^2}^{0,95} = (7,03,\ 44,35)$. Ez

11. ariketa $I_p^{0,99} = (0,1947,\ 0,2639)$

12. ariketa $I_{p_1-p_2}^{0,95} = (0,0674,\ 0,1926),\ p_1 > p_2$

D.4. Hipotesi-kontraste parametrikoak

13. ariketa

a) $H_1 : \mu > 68$, $t = 3,2121$, $a.g. = 167$, $p - balioa = 0,0007904 < \alpha \Rightarrow \mu > 68$, bai

b) $H_1 : \mu = 68$, $t = 3,2121$, $a.g. = 167$, $p - balioa = 0,001581 < \alpha \Rightarrow \mu \neq 68$, ez

c) $H_1 : \mu_e > 68$, $t = -1,4398$, $a.g. = 80$, $p - balioa = 0,9231 \geq \alpha \Rightarrow \mu_e \geq 68$, ez

d) $H_0 : \sigma_g^2 = \sigma_e^2$, $F = 0,9413, a.g. = (86,80)$, $p - balioa = 0,7815 \geq \alpha$, berdintasuna

e) $H_0 : \mu_g = \mu_e$, $t = 5,3807$, $a.g. = 166$, $p - balioa = 2,492\ 10^{-07}$, diferenteak

f) $H_0 : p = 0,5$, $x = 89$, $n = 171$, $p - balioa = 0,6465 \geq \alpha$, bai

g) $H_0 : p \geq 0,6$, $x = 82$, $n = 171$, $p - balioa = 0,0009485 < \alpha$, bai

h) $H_0 : p \geq 0,7$, $x = 89$, $n = 171$, $p - balioa = 6,271e-07 < \alpha$, ez

i) $H_1 : \mu_a < \mu_u$, $t = 25,1609$, $a.g. = 166$, $p - balioa = 1$, ez, kontrakoa.

j) $H_0 : p_{araba} = p_{gipuzkoa}$, $z = 0,7732673$, $p - balioa = 0,4393642$, berdinak

D.5. Hipotesi-kontraste ez-parametrikoak

14. ariketa $\chi^2_p = 11,111$, a.g. $= 2$, $p = 0,003866 < \alpha$, menpekoak.

15. ariketa
(a) $\chi^2_p = 4,2857$, a.g. $= 2$, $p = 0,03843 > \alpha$, askatasuna ezin errefusatu.
(b) $H_0 = p_1 = p_2$, $z_p = 2,25913$, $p = 0,02387528 > \alpha$, beraz, proportzioak berdinak direnik ezin da baztertu.

16. ariketa
(a) Independentzia-testa. $e_{ij} > 5$, baldintzak betetzen dira. $\chi^2_p = 0,90584$, a.g. $= 2$, $p = 0,6358 > \alpha$, askatasuna ezin errefusatu.
(b) $o_{2,3} = 23$, $e_{2,3} = 23,76647$. 49 gipuzkoar daude.
(c) % 0,4693878, % 0,4625000.
(d) % 0,2592593.
(e) Doikuntza-egokitasunerako proba. $\chi^2_p = 17,042$, a.g. $= 2$, $p = 0,0001992 < \alpha$, beraz, ez dira proportzio berdinetan agertzen.
(f) Kolmogorov-Lilliefors testa. $D = 0,083995$, $p = 0,005668 < \alpha$, normaltasuna baztertzen da.

17. ariketa H_0 : emakumezkoen eta gizonezkoen batez besteko denbora-tarteak berdinak dira. Horrela, p-balioa $p = 0,8763 > \alpha$; beraz, ezin dugu berdintasuna errefusatu.

18. ariketa H_0 : tomate fresko eta ontziraturikoen batez besteko kobre kopurua berdina da. Horrela, p-balioa (zuzendu gabe) $p = 0,006874 < \alpha$; beraz, ezin dugu onartu hipotesi nulua.

19. ariketa H_0 : batez besteko heriotza-tasa berdina da urteko sasoi guztietan. Estatistikoa $K = 20,2047$ eta p-balioa $p = 0,0001539 < \alpha$; beraz, ezin dugu berdintasuna onartu.

20. ariketa H_0 : batez besteko altuerak berdinak dira espezieekiko. Estatistikoa $K = 15,971$ eta p-balioa $p = 0,003058 < \alpha$; beraz, ezin dugu berdintasuna onartu.

21. ariketa Estatistikoa=276,29, a.g. $= 2$, $p < 2,2e-16$. Beraz, berdintasuna baztertzen da.

D.6. Bariantza-analisia

22. ariketa
(a) Urte-sasoiaren araberako heriotza-indizeen normaltasunerako Shapiro-Wilk hipotesi-kontrasteak %1eko esangura-mailaz:
$H_0 : X_{udaberria} = \mathcal{N}(9,2167,\ 0,1472)$, $W = 0,9580$, $p = 0,8043$; beraz, normaltasuna ezin da baztertu.
$H_0 : X_{uda} = \mathcal{N}(8,65,\ 0,1871)$, $W = 0,8149$, $p = 0,09764$; beraz, normaltasuna ezin da baztertu.
$H_0 : X_{udazkena} = \mathcal{N}(9,65,\ 0,3564)$, $W = 0,7908$, $p = 0,04845$; beraz, normaltasuna ezin da baztertu.
$H_0 : X_{negua} = \mathcal{N}(10,1833,\ 0,4792)$, $W = 0,9107$, $p = 0,4407$; beraz, normaltasuna ezin da bazter-

tu.

(b) Urte-sasoiaren araberako heriotza-indizeen bariantzen homogenotasunaren hipotesi-kontrasteak, $H_0 : \sigma^2_{udaberria} = \sigma^2_{uda} = \sigma^2_{udazkena} = \sigma^2_{negua}$ %1eko esangura-mailaz:
Barlett: $K = 7,5688$, $p = 0,05582$; beraz, bariantzen homogenotasuna ezin da baztertu.
Levene: $F = 3,2943$, $p = 0,04166$; beraz, bariantzen homogenotasuna ezin da baztertu.

(c) Urte-sasoiaren araberako heriotza-indizeen batezbestekoen berdintasunaren ANOVA hipotesi-kontrasteak, $H_0 : \mu^2_{udaberria} = \mu^2_{uda} = \mu^2_{udazkena} = \mu^2_{negua}$ %5eko esangura-mailaz,
$F = 24,57$ eta $p = 6,5e-07 < \alpha \Rightarrow$ adierazgarria da urte-sasoiaren arabera heriotza-indizeen desberdintasuna.

(d) Tukeyren arabera, %5eko esangura-mailaz binakako konparaketak: $\mu_{uda} < \mu_{negua}, \mu_{udaberria} < \mu_{negua}$ eta $\mu_{udazkena} > \mu_{uda}$ guztiz adierazgarriak dira, eta $\mu_{udazkena} < \mu_{negua}$ eta $\mu_{udaberria} > \mu_{uda}$ adierazgarriak dira. (Hiru multzo homogeneo daude: $\mu_{uda} < \mu_{udaberria} = \mu_{udazkena} < \mu_{negua}$)

23. ariketa
(a) Arraza motaren araberako erresistentziaren normaltasunerako Shapiro-Wilken hipotesi-kontrasteak %5eko esangura-mailaz:
$H_0 : X_A = \mathcal{N}(37,4,\ 14,9599)$, $W = 0,9709$, $p = 0,8812$; beraz, normaltasuna ezin da baztertu.
$H_0 : X_B = \mathcal{N}(82,2,\ 7,9812)$, $W = 0,9167$, $p = 0,5088$; beraz, normaltasuna ezin da baztertu.
$H_0 : X_C = \mathcal{N}(39,6,\ 7,8294)$, $W = 0,9305$, $p = 0,5997$; beraz, normaltasuna ezin da baztertu.

(b) Arraza motaren araberako erresistentziaren bariantzen homogenotasunaren hipotesi-kontrasteak, $H_0 : \sigma^2_A = \sigma^2_B = \sigma^2_C$ %5eko esangura-mailaz:
Barlett: $K = 2,1139$, $p = 0,3477$; beraz bariantzen homogenotasuna ezin da baztertu.
Levene: $F = 1,2255$, $p = 0,3279$; beraz bariantzen homogenotasuna ezin da baztertu.

(c) Arraza motaren araberako erresistentziaren batezbestekoen berdintasunaren ANOVA hipotesi-kontrasteak, $H_0 : \mu^2_A = \mu^2_B = \mu^2_C$ %5eko esangura-mailaz,
$F = 27,43$ eta $p = 3,34\ 10^{-5} < \alpha \Rightarrow$ diferentziak guztiz adierazgarriak dira.

(d) Tukeyren arabera, %5eko esangura-mailaz binakako konparaketak: $\mu_B > \mu_A$ eta $\mu_C < \mu_B$ guztiz adierazgarriak dira. (Bi multzo homogeneo daude: $\mu_A = \mu_C < \mu_B$). Erresistentzia handiena A eta C arraza motako arratoiek izan dute.

24. ariketa $F'(L) = 0,6471$ eta $p' = 0,6395716 > \alpha \Rightarrow$ lurzatien arabera desberdintasunak ez dira adierazgarriak. $F''(O) = 14,0392$ eta $p'' = 0,0003137 < \alpha \Rightarrow$ ongarrien arabera desberdintasunak adierazgarriak dira. % 5eko esangura-mailarekin desberdintasunak adierazgarriak dira.

25. ariketa $F'(L) = 164,3671$ eta $p' = 2,738e-12 < \alpha \Rightarrow$ lekuak eragina du barren bolumena azaltzeko. $F''(G) = 78,2467$ eta $p'' = 1,323e-09 < \alpha \Rightarrow$ gari motak ere bai. $F''' = 8,3692$ eta $p''' = 0,0005356 < \alpha \Rightarrow$ lurzatien eta gari motaren artean interazkioa dago. % 1eko esangura-mailarekin, desberdintasunak adierazgarriak dira.

D.7. Erregresioa

26. ariketa
(a) Erabilgarriak: lineala ($p = 0,03535$), potentziala ($p = 0,04837$) eta esponentziala ($p = 0,01539$). Ez-erabilgarriak: hiperbolikoa ($p = 0,1951$) eta koadratikoa ($p = 0,07619$).

(b) Esponentziala ($R^2 = 0,8929$) > lineala ($R^2 = 0,8167$) > potentziala ($R^2 = 0,7762$)

(c) Lineala: $\hat{Y} = -0,1 + 0,7X$ eta esponentziala: $\hat{Y} = 0,6155722e^{0,34657X}$

(d) Ereduaren erabilgarritasuna aztertzeko estatistikoa: $F = 25$ eta beraren p-balioa $p = 0,01539 < 0,05$ denez, eredua erabilgarria da %5eko esangura-mailarekin. Gainera, eredua oso egokia da korrelazio-koefizienteak (balio absolutuan) 1etik hurbil daudelako, $r^2 = 0,8929, \overline{r} = 0,8571$. Bestalde, zenbatespen-errore estandarra $S = 0,2192$

(e) $I_B^{0,95} = (0,1259832,\ 0,567164)$, 0 barne ez dagoenez, X presioa adierazgarria da Y konpresioa azaltzeko. Unitate bat gehitzean presioan, konpresioa 0,126 eta 0,567 unitate artean aldatzea espero dugu.

(f) $a' = \ln a = -0,485 \Rightarrow a = e^{a'} = e^{-0,48520} = 0,6155722$ eta $I_{A'}^{0,95} = (-1,2168185,\ 0,2464125) \Rightarrow I_A^{0,95} = (e^{-1,2168185}, e^{0,2464125}) = (0,2961709,\ 1,279427)$. Bai, $p = 0,1253 > 0,05 \Rightarrow a' = 0 \Leftrightarrow a = e^0 = 1$

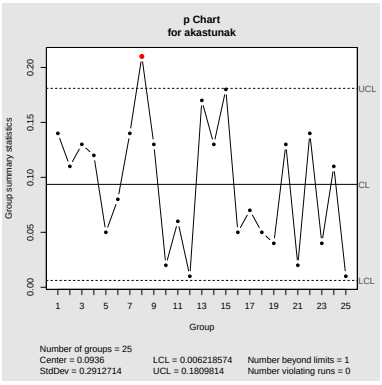
(g) Lehenengo behaketa, hondarra $e = 0,1294494$ eta $\hat{y}_1 = 0,8705506$

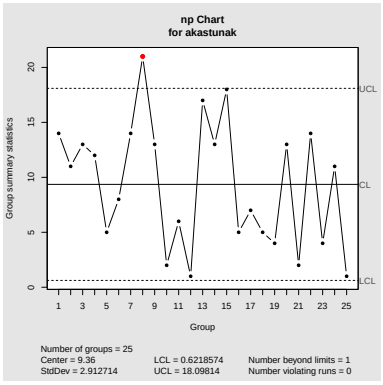
(h) $\hat{y}_4 = 2,462289$ eta $I_{y_4|x_4} = (1,111534,\ 5,454501)$

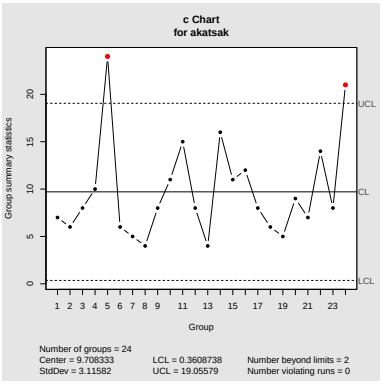
D.8. Kalitatearen kontrol estatistikoa

27. ariketa Akastun matrazeen proportzioa: $\hat{p} = 0,0936,\ LCI = 0,0062,\ LCS = 0,1810$ eta akastun matrazeen kopuru osoa: $n\hat{p} = 9,3600,\ LCI = 0,6219,\ LCS = 18,0981$. 8. behaketa goiko kontrol-mugatik at dagoenez, kontroletik kanpo dago prozesua.

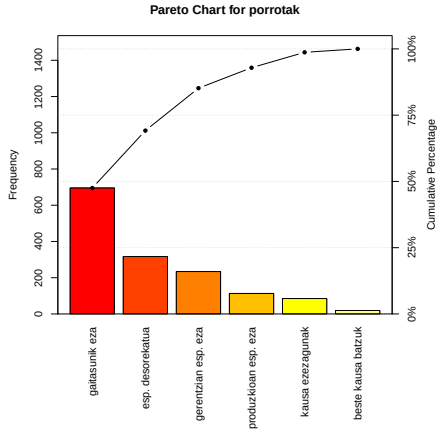
28. ariketa Marra zentralean ikusten denez, akatsen kopurua oholtzarreko 9,7083 da. Bi behaketa 19,0558 goi-muga baino handiagoak dira; izan ere, 5. eta 24. behaketak (akatsak 21 eta 24 izanik, hurrenez hurren). Beraz, prozesua kontroletik kanpo dago.







29. ariketa



Bibliografia

[1] A.J. Arriaza, F. Fernandez, M.A. Lopez, M. Muñoz, S. Perez, and A. Sanchez. *Estadística básica con R y R commander.* UCA, 2008.

[2] B. Calvo and G. Santafe. scmamp: Statistical comparison of multiple algorithms in multiple problems. *The R Journal*, 8(1), 2015.

[3] J. M. Casas-Sánchez. *Inferencia estadística.* Centro de Estudios Ramón Areces, 1997.

[4] G. Casella and R.L. Berger. *Statistical Inference.* Duxbury Press, 1990.

[5] J. C. Correa and N. Gonzalez. *Graficos Estadisticos con R.* Universidad Nacional-Sede Medellin, 2002.

[6] J. L. Devore. *Probabilidad y Estadística para Ingeniería y Ciencias.* International Thomson, 2001.

[7] J. J. Faraway. *Linear Models with R.* Chapman & Hall/CRC, 2004.

[8] J. M. Vilar Fernández. *Modelos Estadísticos Aplicados.* Servicio de Publicaciones de la Universidad de la Coruña, 2003.

[9] G. J. Kerns. *Introduction to Probability and Statistics Using R.* 2010.

[10] J. Kickinson and S. Chakraborti. *Non parametric statistical inference.* Editorial Dekker Inc, 1992.

[11] P. Kuhnert and B. Venables. *An Introduction to R: Software for Statistical Modelling & Computing.* CSIRO Mathematical and Informatical Sciences, Cleveland, Australia, 2005.

[12] J. H. Maindonald. *Using R for Data Analysis and Graphics. Introduction, Code and Commentary.* Centre for Bioinformation Science, Australian National University, 2004.

[13] W. Mendenhall and T. Sincich. *Probabilidad y Estadística para Ingeniería y Ciencias.* Prentice Hall Hispanoamericana, 1997.

[14] M. Merino. *Estatistika: SPSS praktikak.* UPV/EHU, 2011.

[15] M. Merino. *Estatistika: R praktikak.* UPV/EHU, 2012.

[16] J.N. Millar and J.C. Millar. *Estadística y Quimiometría para Química Analítica.* Prentice Hall, Pearson Educación, S.A. Madrid, 2002.

[17] I. Miller and J.E. Freund. *Probabilidad y Estadística para Ingeniería y Ciencias.* Prentice Hall, 1997.

[18] P. Elosua Oliden. *R Gizarte-zientzietarako. Datuen eta eskalen analisia Rcommander-ekin.* UPV/EHUko Argitalpen Zerbitzua, 2009.

[19] E. Paradis. *R Hasiberrientzat.* UEU, 2005.

[20] D. Peña. *Estadística modelos y métodos.* Alianza Editorial, 1987.

[21] S. Ríos. *Ejercicios de Estadística.* Paraninfo, 1989.

[22] V.K. Rohatgi. *Statistical inference.* Editorial Wiley, 1984.

[23] V.K. Rohatgi. *An introduction to probability theory and mathematical statistics.* John Wiley and sons, 2000.

[24] L. Ruiz-Maya. *Problemas de Estadística.* Editorial AC, 1989.

[25] R.L. Scheaffer and J.T. McClave. *Probabilidad y Estadística para Ingeniería.* Iberoamericana, 1993.

[26] M. R. Spiegel. *Estadística.* MacGraw-Hill, 2002.

[27] G. Velasco and P.M. Wisniewski. *Probabilidad y estadística para Ingeniería y Ciencias.* Thomson Learning, 2001.

[28] B. Venables, D. Smith, R. Gentleman, R. Ihaka, M. Machler, A. Gonzalez, and S. Gonzalez. *Notas sobre R. Un entorno de programación para Análisis de Datos y Gráficos.* Department of Statistics, University of Adelaide, 2000.

[29] J. Verzani. *simpleR. Using R for Introductory Statistics.* CSI Math department, 2002.

[30] R.E. Walpole. *Probabilidad y Estadística para Ingeniería y Ciencias.* Pearson Educacion, 2007.

[31] R.E. Walpole and Myers. *Probabilidad y Estadística.* Mc Graw Hill, 1992.

UNIBERTSITATEKO ESKULIBURUAK
MANUALES UNIVERSITARIOS

INFORMAZIOA ETA ESKARIAK • INFORMACIÓN Y PEDIDOS

UPV/EHUko Argitalpen Zerbitzua • Servicio Editorial de la UPV/EHU
argitaletxea@ehu.eus • editorial@ehu.eus
1397 Posta Kutxatila - 48080 Bilbo • Apartado 1397 - 48080 Bilbao
Tfn.: 94 601 2227 • www.ehu.eus/argitalpenak

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea