6. GAIA:

Oinarrizko estatistika

Matematika Aplikatua,

Estatistika eta Ikerkuntza Operatiboa Saila

Zientzia eta Teknologia Fakultatea

Euskal Herriko Unibertsitatea

Aurkibidea

6. gaia

Oinarrizko estatistika.

Naturako fenomenoetan oinarrituta dauden zenbakizko datuak aztertzeko, eta baieztapen edo lege batzuk onar daitezkeen ebazteko metodoak ematen dituen zientzia da *Estatistika*.
Naturako fenomenotzat honako hauek joko ditugu: alde batetik, natura bizidun eta bizigabean pertsonaren kontrolik gabe gertatzen direnak eta beste aldetik, ikertzaileak partzialki eragindakoak, bere kontrolaren menpean. Azken fenomeno horiei *esperimentu* izena emango diegu. Fenomeno horiek Estatistikaren bidez aztertu ahal izateko, bete behar dituzten baldintzak honako hauek dira:

- Errepikapena: baldintza egonkorren multzo baten menpean hainbat aldiz errepika daitezke.
- Zoria: proba bakoitzean ezin daiteke lortuko den emaitza aurresan.
- Egonkortasuna: egindako proben kopurua emendatuz doan neurrian, emaitza bakoitzaren maiztasun erlatiboak zenbaki batean egonkortasunera jotzen du.

Ikerketa zientifiko esperimental batean ondoko urratsak egon ohi dira:
1) populazio eta esperimentuaren helburua (hipotesien planteamendua) argi eta garbi definitzea;
2) laginketaren prozedura edo esperimentua diseinatzea;
3) datuak jasotzea eta aztertzea;
4) lagin-datuetan oinarrituta, populazioari buruzko inferentzia-metodoak burutzea;
5) aurrreikuspenak egitea, ondorioak eta erabakiak hartzea, fidagarritasuna edo konfiantza-maila zehaztuz.

Estatistika bi arlo nagusitan banatzen da: *Estatistika deskribatzailea* eta *Inferentzia estatistikoa*.

6.1. Estatistika deskribatzailea.

Esperimentuen datuak biltzeaz, aztertzeaz, antolatzeaz eta laburtzeaz arduratzen den Estatistikaren arloa *Estatistika deskribatzailea* da. Horretarako, taulak, parametro estatistikoak (zenbakizko laburpenak) eta grafikoak erabiltzen dira.

Azterketa estatistikoaren helburua den elementu multzoari *populazio* deitzen zaio. Populazio osoa aztertzea zaila denean lagin bat aukeratzen da. Populazioaren elementuen azpimultzoari *zorizko lagin* deitzen zaio, non elementu bakoitzaren aukera askea baita eta aukeratua izateko probabilitatea ezaguna baita. Populazioaren elementu guztiek aukeratuak izateko probabilitate berdina duten zorizko laginari *zorizko lagin bakun* deitzen zaio. *Laginaren tamaina* laginaren elementu-kopurua da, eta n-ren bidez adierazten da.

Izaera aztertzen den populazioaren propietatea da, eta izaerak izan dezakeen era edo egoera bakoitza *modalitatea* da.
Izaerak mota honetakoak izan daitezke:

- *Kualitatiboak:* modalitateak ezin daitezke zenbaki eran adierazi. Bi motakoak izan daitezke: *ordinalak*, modalitateak ordenatu ahal direnean eta *nominalak*, ordenatu ezin direnean.
- *Kuantitatiboak:* modalitateak neurgarriak dira. Bi motakoak ere izan daitezke: *diskretuak*, modalitateek hartzen duten balio-kopurua (finitua edo infinitua) zenbakigarria denean eta *jarraituak*, modalitateek hartzen duten balioa, tarte baten barnean, edozein izan daitekeenean.

Izaeraren modalitateen neurketari dagokion zenbaki multzoari *aldagai estatistiko* deitzen zaio. Esate baterako, kutsadura-maila aldagai kualitatibo ordinala izan daiteke, bere modalitateak: oso altua, altua, ertaina, baxua, oso baxua eta nulua izanik; jaioterria, berriz, aldagai kualitatibo nominala da. Aldagai kuantitatibo diskretu baten adibidea adina da, modalitateak $\{0, 1, 2, 3, \ldots\}$ multzoan daudelako; hala ere, pisua, aldagai kuantitatibo jarraitu moduan kontsidera dezakegu, modalitateak tarte batean hartzen dituztelako balioak.

6.1.1. Taula estatistikoa.

Demagun populazio edo lagin batean aldagai bat aztertu nahi dela. Estatistika deskribatzaileak ematen duen lehenengo urratsa hauxe da: aztertu nahi den fenomenoari buruzko datuak arduraz biltzea. Datuak bildu eta gero, taula estatistiko edo maiztasun-taula izenekoan ordenatuta aurkeztea da bigarren urratsa. Taula estatistikoan datu hauek adierazten dira:

1. zutabean, X aldagaia eta haren balioak, x_i, txikienetik handienera ordenaturik.
2. zutabean, aldagaiaren balio bakoitzari dagokion laginaren ale-kopurua edo *maiztasun absolutua*, f_i. Propietate hauek betetzen dituzte: i) $0 \leq f_i \leq n$, $\forall i = 1, \ldots, k$ eta ii) $\sum_{i=1}^{k} f_i = n$, non k aldagaiaren balio desberdinen kopurua den.
3. zutabean, aldagaiaren balio bakoitzari dagokion *maiztasun metatua*, F_i; hots, txikiagoak diren balio guztien eta horien maiztasun absolutuen batura, $F_i = \sum_{j=1}^{i} f_j$. Propietatea hau betetzen dute: $F_k = n$.
4. zutabean, aldagaiaren balio bakoitzari dagokion *maiztasun erlatiboa*, h_i; hots, maiztasun absolutua eta lagin tamainaren arteko zatidura, $h_i = f_i/n$. Betetzen dituzten propietateak: i) $0 \leq h_i \leq 1$, $\forall i = 1, \ldots, k$ eta ii) $\sum_{i=1}^{k} h_i = 1$. Gainera, $100 \cdot h_i$ aleen ehunekoa da, *portzentajea*.
5. zutabean, aldagaiaren balio bakoitzari dagokion *maiztasun metatu erlatiboa*, H_i; hots, txikiagoak diren balio guztien eta horien maiztasun erlatiboen batura, $H_i = \sum_{j=1}^{i} h_j$. Propietate hau betetzen dute: $H_k = 1$. Gainera, $100 \cdot H_i$ *portzentaje metatua* da.

Askotan taulari bi zutabe gehitzen zaizkio, $x_i f_i$ eta $x_i^2 f_i$, parametro estatistikoen kalkulua errazteko asmoz, azken lerroan datuen batura eta karratuen batura adierazten baitira.

6.1. adibidea. Laborategi batean plastiko berri baten ezaugarriak aztertu nahi dituzte. Horretarako, zorizko lagin bat kontsideratzean propietate batzuk aztertu dira, haien artean apurtze-erresistentzia. Beraz, izan bedi $X =$ *apurtze-erresistentzia* izeneko zorizko aldagaia. Lortutako datuak hauek dira:

5.8 6.2 6.5 6.7 6.9 7.1 6.5 7.2 7.0 6.8
6.6 6.3 6.1 7.1 6.5 6.9 6.7 7.2 4.0 8.6

Datu horien azterketa egin baino lehen, jar ditzagun maiztasun-taula estatistikoan. Aurrerago ikusiko dugunez, datuak txikienetik handienera ordenatzeak estatistiko batzuen kalkulua errazteko balio du. Esate baterako, behaketen % 75ek 7.0 edo gutxiagoko apurtze-erresistentzia dauka. □

Taula 6.1. Maiztasun-taula.

x_i	f_i	F_i	h_i	H_i	$x_i f_i$	$x_i^2 f_i$
4.0	1	1	0.05	0.05	4.00	16.00
5.8	1	2	0.05	0.10	5.80	33.64
6.1	1	3	0.05	0.15	6.10	37.21
6.2	1	4	0.05	0.20	6.20	38.44
6.3	1	5	0.05	0.25	6.30	39.69
6.5	3	8	0.15	0.40	19.50	126.75
6.6	1	9	0.05	0.45	6.60	43.56
6.7	2	11	0.10	0.55	13.40	89.78
6.8	1	12	0.05	0.60	6.80	46.24
6.9	2	14	0.10	0.70	13.80	95.22
7.0	1	15	0.05	0.75	7.00	49.00
7.1	2	17	0.10	0.85	14.20	100.82
7.2	2	19	0.10	0.95	14.40	103.68
8.6	1	20	0.05	1.00	8.60	73.96
$\sum$	20		1		132.70	893.99

Aldagai kuantitatiboetan eta kualitatibo ordinaletan, aipatutako datu guztiak kalkula daitezke. Hala ere, aldagai kualitatibo nominaletan, ordez, soilik dauka zentzurik maiztasun absolutuak eta erlatiboak kalkulatzea.

6.1.2. Parametro estatistikoak.

Aztertutako aldagaiaren propietateak laburki eta zehazki deskribatzen dituen zenbakizko balioari *estatistiko* deitzen zaio.
Estatistikoen sailkapena honako hau da:

- ▷ *Joera zentraleko estatistikoak:* neurriak bere inguruan, nolabait, biltzen dira. Esate baterako, batezbesteko aritmetikoa, mediana eta moda.
- ▷ *Posizio-estatistikoak:* alde batean behaketen ehuneko zehatz bat uzten dutenak. Pertzentilak, koartilak, dezilak.
- ▷ *Sakabanatze-estatistikoak:* zentralizazioarekiko sakabanatzea edo kontzentrazioa adierazten duten neurriak. Adibidez, batezbestekoarekiko desbideratzea, heina, koartilarteko heina, bariantza, desbideratze estandarra, kuasibariantza, kuasidesbideratze estandarra, aldakuntza-koefizientea.
- ▷ *Forma-estatistikoak:* banaketaren itxura adierazten duten zenbakizko balioak. Alborapena eta asimetria koefizientea simetria adierazteko, kurtosia zapaltasuna neurtzeko, besteak beste.

Definizioa: Datu bakoitzaren baturaren eta laginaren gai kopuruaren arteko zatidura batezbestekoa da, $\overline{x}$.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{k} x_i f_i}{n}.$$

Lagin-fluktuazioekiko sentikorra izan daiteke, muturretan jartzen diren balio bitxiek eragina daukatelako batezbestekoan. Oztopo hau ekiditeko, mediana izeneko estatistikoa dugu, banaketa asimetrikoak dituenean, joera zentraleko estatistikorik adierazgarriena dena.

Definizioa: *Mediana*, Me, alde bakoitzean ale-kopuru berdina daukan aldagaiaren balioa da, txikienetatik handienera ordenatuta daudela ale horiek; hau da, alde bakoitzean elementuen % 50 uzten duen aldagaiaren balioa.

Mediana kalkulatzeko taula estatistikoan %50eko portzentaje metatua bilatu. Baldin $H_i = 0.50$, orduan $Me = \frac{x_i + x_{i+1}}{2}$. Bestela, izan bedi H_i, %50eko portzentaje metatutik goitik hurbilen dagoena, hots, $H_{i-1} < 0.50 < H_i$, orduan, $Me = x_i$.

Definizioa: *Moda*, Mo, maiztasun handiena daukan aldagaiaren balioa da; hots, aldi-kopuru gehienetan agertzen den neurria.

Ez du zertan bakarra izan ezta erdialdean egon. Haren kalkuluan lagineko datu bakar bat erabiltzen denez, informazio asko galtzen da.

6.2. adibidea. Kalkula ditzagun adibideko joera zentraleko estatistikoak apurtze-erresistentzia aldagairako.

Em.: Batezbesteko aritmetikoa, $\overline{x} = 132.7/20 = 6.635$. Mediana kalkulatzeko, maiztasun erlatibo metatuen zutabean 0.50 agertzen ez denez, handiagoa den lehenengoa 0.55 da, hots, $H_7 < 0.50 < H_8$ eta horri dagokion aldagaiaren balioa $x_8 = 6.70$ da; alegia, $Me = 6.70$. Azkenik, maiztasun absoluturik handiena 3 da, eta horri dagokion aldagaiaren balioa 6.50; orduan, $Mo = 6.50$. □

Definizioa: Lagin ordenatua 100 zati berdinetan banatzen dituzten aldagaiaren balioak *pertzentilak* deitzen dira, $p_1, p_2, p_3, \ldots, p_{99}, p_{100}$ adierazirik. Orokorrean, j. pertzentila, bere ezkerrean balioen $\%j$-a uzten duen aldagaiaren balioa da.
j. pertzentila kalkulatzeko taula estatistikoan $\%j$-ko portzentaje metatua bilatu. Baldin $H_i = j/100$, orduan $p_j = \frac{x_i + x_{i+1}}{2}$. Bestela, izan bedi H_i, $\%j$-ko portzentaje metatutik goitik hurbilen dagoena, hots, $H_{i-1} < j/100 < H_i$, orduan, $p_j = x_i$.

6.3. adibidea. Kalkula ditzagun adibideko 25. eta 75. pertzentilak.

Em.: Alde batetik, 25. pertzentila kalkulatzeko, maiztasun erlatibo metatuen zutabean $H_5 = 0.25$ agertzen denez,

$$p_{25} = \frac{x_5 + x_6}{2} = \frac{6.30 + 6.50}{2} = 6.40.$$

Beste aldetik, 75. pertzentila kalkulatzeko, maiztasun erlatibo metatuen zutabean $H_{11} = 0.75$ agertzen denez,

$$p_{75} = \frac{x_{11} + x_{12}}{2} = \frac{7.00 + 7.10}{2} = 7.05.$$

Beraz, apurtze-erresistentzia aldagaiaren behaketa zentralen %50a 6.40 eta 7.05 balioen artean dago. □
Aurreko estatistikoek ez digute ematen laginaren kontzentrazioari buruzko informaziorik. Horretarako hurrengo sakabanatze-estatistikoak ikusiko ditugu.

Definizioa: Laginaren *bariantza* batezbestekoarekiko desbideratzeen karratuen batezbestekoa da, eta s_n^2 adierazten da.

$$s_n^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i}{n}.$$

Kalkulua errazteko baliokide den ondoko formula erabil daiteke:

$$s_n^2 = \frac{\sum_{i=1}^{k} x_i^2 f_i}{n} - \overline{x}^2.$$

Definizioa: Laginaren *desbideratze estandarra*, s_n, bariantzaren erro karratu positiboa da. Batzuetan sakabanatze absolutua edota desbideratze tipikoa ere deitzen diogu. $s_n = +\sqrt{s_n^2}$

Definizioa: Laginaren *kuasibariantza* s_{n-1}^2 adierazten da, eta honela kalkulatzen da:

$$s_{n-1}^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i}{n-1}.$$

Bariantzaren eta kuasibariantzaren arteko erlazioa honako hau da:

$$s_{n-1}^2 = \frac{n}{n-1} s_n^2.$$

Definizioa: Laginaren *kuasidesbideratze estandarra*, s_{n-1}, kuasibariantzaren erro karratu positiboa da. $s_{n-1} = +\sqrt{s_{n-1}^2}$

Definizioa: Laginaren *Pearson-en aldakuntza-koefizientea* desbideratze estandarraren eta batezbestekoaren arteko zatidura da. Ehunekotara ekarri ondoren, CV (Coeficiente de Variación) adierazten da. Ezin da erabili batezbestekoa nulua denean:

$$CV = \frac{s_n}{\overline{x}} \cdot 100.$$

Batzuetan sakabanatze erlatiboa ere deitzen diogu, sakabanatzea konparatzeko erabiltzen da eta.

Datuek aldakuntzarik ez daukatenean edozein sakabanatze-estatistikoren balioa zero izango da; beste kasuetan balio positiboak bakarrik hartzen dituzte. Zenbat eta sakabanatze handiagoa izan, estatistikoen balioa handiagoa izango da.

$\overline{x}$, Me, Mo, p_i, s_n eta s_{n-1} aldagaiaren unitatetan adierazten dira; s_n^2 eta s_{n-1}^2 aldagaiaren unitate karratutan eta CV dimentsio gabeko kantitatea da eta ehunekotan adierazten da.

Aldagai kuantitatiboen edozein estatistiko kalkula daiteke. Hala ere, aldagai kualitatibo ordinaletan, soilik Me, Mo eta p_i eta azkenik, aldagai kualitatibo nominaletan Mo kalkulatzea baino ez dauka zentzurik.

6.4. adibidea. Kalkula itzazu aurreko adibideko sakabanatze estatistikoak apurtze-erresistentzia aldagairako.

Em: $s_n^2 = \frac{\sum x_i^2}{n} - \overline{x}^2 = \frac{893.99}{20} - 6.635^2 = 0.6763$ unitate^2.
$s_n = \sqrt{s_n^2} = 0.8224$ unitate.
$s_{n-1}^2 = 0.6763\frac{20}{19} = 0.7119$ unitate^2.
$s_{n-1} = \sqrt{s_{n-1}^2} = 0.8437$ unitate.
$CV = \frac{0.8224}{6.635} \cdot 100 = 12.3943\%$ □

6.1.3. Adierazpen grafikoak.

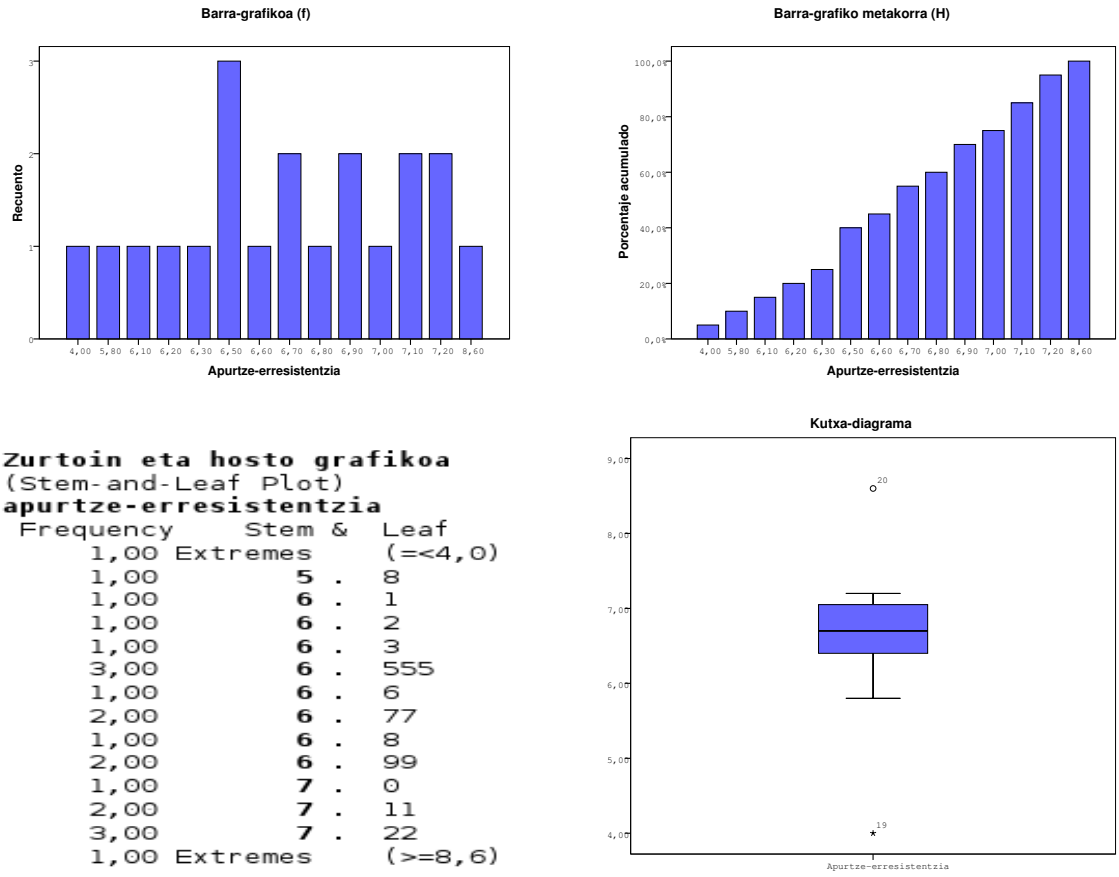
Adierazpen grafikoek informazio orokorra, arina eta ulertzeko erraza eskaintzen digute. Aldagaiaren motaren arabera sailka daitezke.
Aldagai kuantitatibo diskretuetan, hurrengo grafikoak erakitzen dira: barra-grafikoa, grafiko metakorra, zurtoin eta hosto grafikoa eta kutxa-diagrama, besteak beste.

* *Barra-grafikoa* gehien erabiltzen direnetariko bat da. Abzisa-ardatzean aldagaiaren balioak eta ordenatu-ardatzean maiztasun absolutua edo erlatiboak (f_i edo h_i) adierazten dira, eta batzuen eta besteen arteko egokitasuna barra-sistema baten bidez deskribatzen dira. Barra guztiek zabalera berdinekoak izan behar dute. Grafiko honetan, moda aurkitzea begibistakoa da.
* *Grafiko metakor*rean, ordenatu-ardatzean maiztasun metatuak edo metatu erlatiboak (F_i edo H_i) adierazten dira. Grafiko honetan mediana eta edozein pertzentil arin aurki daitezke.
* *Zurtoin eta hosto grafikoa*, grafiko eta taularen nahasketa da; balio bikoitza osagai bitan banandu: hosto-osagaia (azken zifra) eskuinaldeko zutabe batean eta zurtoin-osagaia (gainontzekoa) ezkerraldeko zutabean adierazita.
* *Kutxa-diagrama*, kutxa batean behaketa zentralen %50 adierazten da (p_{25}, p_{50} eta p_{75} erabiliz). Muturretan balio arraroak edo *outlierrak* daude: $(M_1,\ m_1)$ eta $(m_3,\ M_3)$ tarteetan daudenak *outlier moderatuak* deitzen dira, ∘ adierazirik eta $(-\infty,\ M_1)$ eta $(M_3,\ \infty)$ tarteetan daudenak *izugarrizko outlierrak* deitzen dira, * adierazirik; barne hesiak $m_1 = p_{25} - 1.5(p_{75} - p_{25})$ eta $m_3 = p_{75} + 1.5(p_{75} - p_{25})$ dira eta

kanpo hesiak $M_1 = p_{25} - 3(p_{75} - p_{25})$ eta $M_3 = p_{75} + 3(p_{75} - p_{25})$ dira. Behin outlierrak kenduta, balio maximo eta minimoa, kutxatik kanpo marra baten bidez adierazten dira.

6.5. adibidea. Egin itzazu aurreko adibideko apurtze-erresistentzia aldagaiaren grafikoak.

Em.: Aipatutako lau grafikoak irudian ikus daitezke.



6.1. irudia. Adibideko grafikoak.

Kutxa-diagraman behatu ahal denez bi outlier daude: 4.00 balioa (19. datua) izugarrizko outlierra da eta 8.60 balioa (20. datua) outlier moderatua da. Izan ere, $M_1 = 4.45$, $m_1 = 5.425$, $m_3 = 8.025$ eta $M_3 = 9$ direnez, $4.00 \in (-\infty, M_1)$ eta $8.60 \in (m_3, M_3)$ baitira. □

Aldagai kuantitatibo jarraituetan, hurrengo adierazpenak erabiltzen dira: barra-grafikoa eta grafiko metakorra (barra ukitzaileekin), histograma, maiztasun-poligonoa, ojiba edo maiztasun-poligono metakorra eta kutxa-diagrama, besteak beste.

Aldagai kualitatibo nominaletan, soilik maiztasun absolutuak eta erlatiboak kalkulatzen direnez, ondoko grafikoak eraiki daitezke: barra-grafikoa, sektore-grafikoa, piktograma (irudia) eta kartograma (mapa) eta abar.

6.2. Probabilitate-Teoria.

6.2.1. Probabilitatea.

Esperimentua egin aurretik haren emaitza ezagutzea ezinezkoa denean *zorizko* deitzen zaio, teorikoki baldintza berdinen menpean etengabe errepika daiteke eta emaitza posible guztien multzoa ezaguna da. Multzo horri zorizko esperimentuaren *lagin-espazioa* deitzen zaio, eta Ω-ren bidez adierazten da.

Lagin-espazioaren azpimultzoei *gertaera* deitzen zaie, eta letra larrien bidez adierazten dira; haien artean multzo hutsa (*ezinezko gertaera*, $\emptyset$) eta lagin-espazioa bera (*gertaera segurua*, Ω) azpimultzo gisa kontuan izan behar dira. Elementu bakar batek osaturiko gertaerei *oinarrizko gertaera* deitzen zaie, eta elementu batek baino gehiagok osaturikoei *gertaera konposatu*.

6.6. adibidea. Demagun zorizko esperimentua bi dado botatzean datzala. Eraiki ezazu lagin-espazioa eta edozein gertaera.

Em.: Lagin-espazioa: $\Omega = \{(1,1),\ (1,2), \ldots, (6,6)\}$. Gertaera ugari kontsidera daitezke, izan bedi A ='batura 5 izatea' gertaera, orduan $A = \{(1,4),\ (2,3),\ (3,2),\ (4,1)\}$ gertaera konposatua da. □

Definizioak: A gertaera ez jazotzea beste gertaera bat da eta A-ren *kontrako gertaera* deitzen da, $\overline{A}$ adierazirik, $\overline{A} = \{x \in \Omega : x \notin A\}$.
A eta B gertaeren arteko bat gutxienez jazotzea beste gertaera bat da. Gertaera berri honi, A eta B-ren arteko *bildura* deritzogu, $A \cup B$ adierazirik, $A \cup B = \{x \in \Omega : x \in A \vee x \in B\}$.
A eta B gertaerak batera gertatzen direnean, jazotzen den gertaera A eta B-ren *ebakidura* deitzen da, $A \cap B$ adierazten delarik, $A \cap B = \{x \in \Omega : x \in A \wedge x \in B\}$.

A eta B *gertaera bateraezinak* dira baldin eta soilik baldin $A \cap B = \emptyset$ betetzen bada. Beraz, bi gertaera batera jazo ezin direnean, gertaera bateraezinak direla esaten da.
Definizioa: Zorizko esperimentu bati loturiko gertaera guztiek osaturiko multzoari Ω*-ren parteen multzoa* deitzen zaio, hau da, Ω-ren azpimultzo guztiek osaturiko multzoari. $\mathcal{P}(\Omega)$ adierazten da.
Definizioa: Izan bedi zorizko esperimentu bati lotutako Ω lagin-espazioa. $\mathcal{P}(\Omega) \to [0,1]$ definitutako P funtzioa *probabilitatea* deituko diogu:

$$\begin{aligned} P : \mathcal{P}(\Omega) &\to [0,1] \\ A &\to P(A) \end{aligned}$$

ondoko axiomak betetzen baditu: i) $P(\Omega) = 1$ eta ii) $\forall A, B \in \mathcal{P}(\Omega)$ non A eta B bateraezinak diren, $P(A \cup B) = P(A) + P(B)$.

Propietatea: *Izan bedi* $\Omega = A_1 \cup A_2 \cup \ldots \cup A_n$*, non* A_i *oinarrizko gertaera guztiak baterazinak eta ekiprobableak baitira (hots,* $P(A_i) = 1/n$*,* $i = 1, 2, \ldots, n$*). Baldin* $A = A_1 \cup A_2 \cup \ldots \cup A_k$ *bada,* $k \leq n$*, orduan*

$$P(A) = \frac{Kard(A)}{Kard(\Omega)} = \frac{k}{n} = \frac{A\text{-ren aldeko oinarrizko gertaera kopurua}}{\text{oinarrizko gertaera kopurua}}$$

6.7. adibidea. Aurreko zorizko esperimentuan, zein da 5eko batura lortzeko probabilitatea?

Em.: $P(A) = \frac{Kard(A)}{Kard(\Omega)} = \frac{4}{36} = \frac{1}{9} = 0.1111$. Beraz, bi dado airera botatzean, bost lortzeko probabilitatea %11.11-koa da. □

Definizioa: A eta B gertaera *askeak* edo *independenteak* dira baldin $P(A) > 0$ eta $P(B) > 0$ izanik, ondokoa betetzen bada: $P(A \cap B) = P(A) \cdot P(B)$. Independentziaren esanahia bataren jazoerak bestearengan eraginik ez izatea da.

6.2.2. Zorizko aldagaiak.

Zorizko esperimentu bati lotutako ezaugarri bakoitzari *zorizko aldagaia* deitzen zaio, aldagaia balioz aldatzen delako, eta zorizkoa bere portaera zoriaren menpean dagoelako eta ezin daitekeekalo aurresan. Horrela, X zorizko aldagaia Ω lagin-espazioan definituta egonik $\mathbb{R}$-ren balioak hartzen dituen funtzioa da. $X : \Omega \to \mathbb{R}$. Zorizko aldagaiak letra larriz adierazten dira, eta haientzat egiaztaturiko balioak letra txikiz.

Zorizko aldagaiak hartzen duten zenbakizko balio motaren arabera, hurrengo bi multzotan sailkatzen dira. *X zorizko aldagaia diskretua* izango da, baldin hartzen duen zenbakizko balio mota finitua edo infinitu zenbagarria (hau da, zenbaki arrunta) bada. *X zorizko aldagaia jarraitua* izango da, baldin zenbaki erreal positibo guztiak edo zuzenki edo zuzenerdi baten balio guztiak har baditzake.

Zorizko aldagai diskretuak

Definizioa: Zorizko aldagai diskretuaren *probabilitate-legea* aldagaiaren balio bakoitzari bere gertagarritasunaren neurria ematen dion $f : \mathbb{R} \to [0, 1]$ definitutako funtzioa da:

$$\begin{aligned} f : \mathbb{R} &\to [0, 1] \\ x &\to f(x) = P(X = x) \end{aligned}$$

X aldagaiak hartu ahal dituen balioen multzoa *aldagaiaren heina* da, finitua $Irudi(X) = \{x_1, x_2, \ldots, x_n\}$ edo infinitu zenbakigarria $Irudi(X) = \{x_1, x_2, \ldots\}$. $f(x_i) = P(X = x_i)$ adierazten dugu $\forall x_i \in Irudi(X)$, eta aldagaiak hartzen ez dituen x balio guztientzat $f(x) = P(X = x) = 0$ betetzen da.

Propietateak:

1. $f(x) \geq 0, \ \forall x \in \mathbb{R}$
2. $\displaystyle\sum_{x \in \mathbb{R}} f(x) = \sum_{x_i \in Irudi(X)} P(X = x_i) = 1.$

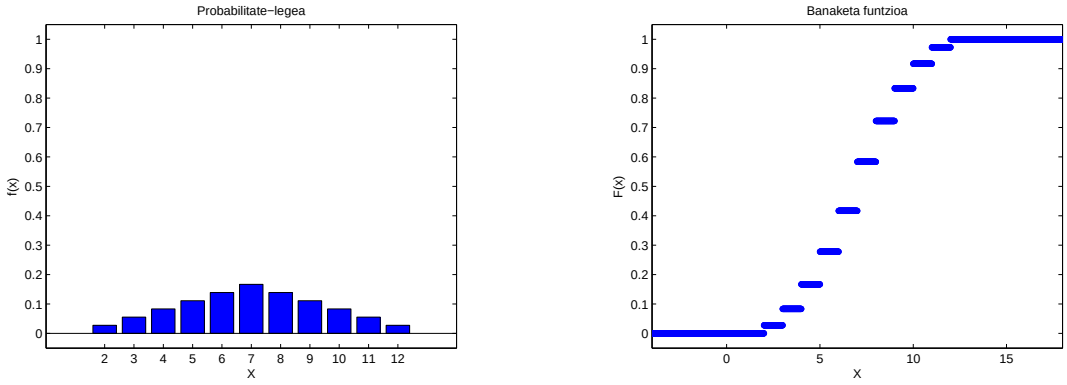
Definizioa: Izan bedi X zorizko aldagai diskretua, X-ren *banaketa-funtzioa* $F : \mathbb{R} \to [0, 1]$ funtzioa da, non $F(x) = P(X \leq x) = \sum_{a \leq x} f(a)$ baita.

6.8. adibidea. Aurreko adibideari lotuta, defini ezazu edozein aldagai diskretu eta interpreta itzazu f eta F funtzioak.

Em.: Defini dezagun $X =$ '*bi dadoen puntuazioaren batura*' z. a. diskretua. Izan ere, aldagaiaren heina $Irudi(X) = \{2,\ 3, \ldots,\ 12\}$ zenbakigarria baita. Orain, *batura 5 izatea* izeneko gertaera, $(X = 5)$ moduan adieraz daiteke. Probabilitatea legeak, $f(x) = P(X = x)$, batura posible bakoitza agertzeko probabilitatea adierazten du eta banaketa

funtzioak, $F(x) = P(X \leq x)$, x balioa edo gutxiagokoa batzeko probabilitatea.

$$f(x) = \begin{cases} 1/36, & \text{baldin } x \in \{2, 12\} \\ 2/36, & \text{baldin } x \in \{3, 11\} \\ 3/36, & \text{baldin } x \in \{4, 10\} \\ 4/36, & \text{baldin } x \in \{5, 9\} \\ 5/36, & \text{baldin } x \in \{6, 8\} \\ 6/36, & \text{baldin } x = 7 \\ 0, & \text{baldin } x \notin I(X) \end{cases} \quad F(x) = \begin{cases} 0, & \text{baldin } x < 2 \\ 1/36, & \text{baldin } 2 \leq x < 3 \\ 3/36, & \text{baldin } 3 \leq x < 4 \\ 6/36, & \text{baldin } 4 \leq x < 5 \\ 10/36, & \text{baldin } 5 \leq x < 6 \\ 15/36, & \text{baldin } 6 \leq x < 7 \\ 21/36, & \text{baldin } 7 \leq x < 8 \\ 26/36, & \text{baldin } 8 \leq x < 9 \\ 30/36, & \text{baldin } 9 \leq x < 10 \\ 33/36, & \text{baldin } 10 \leq x < 11 \\ 35/36, & \text{baldin } 11 \leq x < 12 \\ 1, & \text{baldin } x \geq 12 \end{cases} \quad \square$$



6.2. irudia. f eta F funtzioen adierazpen grafikoak.

Definizioa: X eta Y bi aldagai diskretuak *askeak* dira baldin eta solik baldin $P((X = x_i) \cap (Y = y_j)) = P(X = x_i) \cdot P(Y = y_j)$ bada.

Zorizko aldagai jarraituak

Definizioa: Zorizko aldagai jarraituen *dentsitate-funtzioa* $f : \mathbb{R} \to \mathbb{R}$ funtzioa da, non axioma hauek betetzen baitira:

i) $f(x) \geq 0, \ \forall x \in \mathbb{R}$.

ii) $\int_{-\infty}^{\infty} f(x)dx = 1$. Hau da, f-ren grafikoa eta OX ardatzaren artean geratzen den azalera 1 da.

iii) $\forall a, b \in \mathbb{R}, \ P(a \leq X \leq b) = \int_a^b f(x)dx$. Hau da, X aldagaia a eta b balioen artean egoteko probabilitatea f-ren grafikoaren $x = a$ eta $x = b$ zuzenen eta OX ardatzaren arteko azalera da.

Propietateak:

1. $\forall a \in \mathbb{R},\ P(X = a) = P(a \leq X \leq a) = \int_a^a f(x)dx = 0.$
2. $\forall a, b \in \mathbb{R},\ P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b).$

Definizioa: Izan bedi X zorizko aldagai jarraitua, X-ren *banaketa funtzioa* $F : \mathbb{R} \to [0, 1]$ funtzio hau da:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

Hau da, F-k f-ren grafikoaren eta OX ardatzaren bitartean dagoen azalera adierazten du, x baino txikiagoak diren balioentzat.

6.9. adibidea. Aurreko adibideari lotuta, defini ezazu edozein aldagai jarraitu eta interpreta itzazu f eta F funtzioak.

Em.: Demagun bi dadoak 2 metroko luzera duen mahai baten gainean botatzen direla, defini dezagun $Y =$ '*dado bakoitzetik mahaiaren ertzerainoko distantzien biderkadura (metrotan)*' z. a. Argi dago jarraitua dela, izan ere, aldagaiaren heina $Irudi(Y) = [0,\ 4]$ zenbakigarria ez baita. Defini dezagun ondoko dentsitate funtzioa

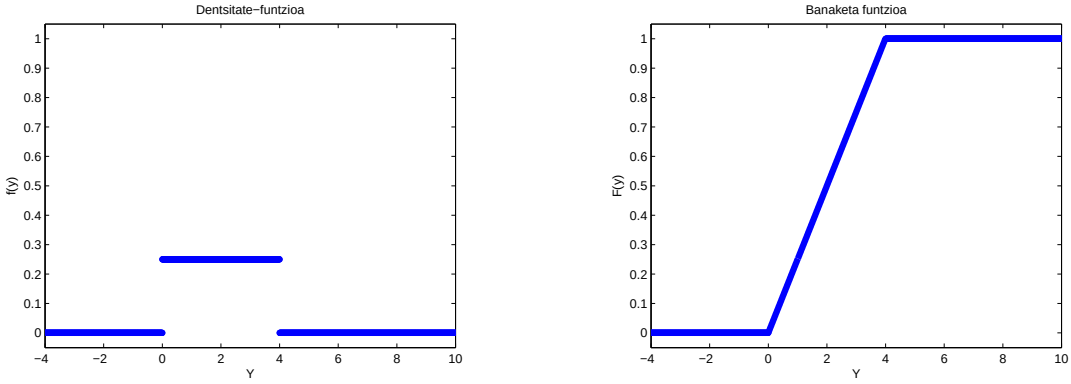
$$f(y) = \begin{cases} \frac{1}{4}, & \text{baldin } y \in [0,\ 4] \\ 0, & \text{baldin } y \notin [0,\ 4] \end{cases}$$

Berez, dentsitate funtzioa da, $f(y) \geq 0,\ \forall y$ eta $\int_{-\infty}^{\infty} f(y)dy = \int_0^4 \frac{1}{4}dy = \frac{1}{4}(4 - 0) = 1$ axiomak betetzenbaitira. Banaketa hau, konstante baten bidez adierazita, *uniformea* deitzen da, edozein bi luzera berbereko azpitartetan jausteko probabilitatea berdina delako; izan ere,

$$\forall a, b \in [0,\ 4],\ P(a < Y < b) = \int_a^b f(y)dy = \frac{b - a}{4}.$$

Banaketa funtzioak, $F(y) = P(Y \leq y)$, distantzien biderkadura y edo gutxiagoko balioa hartzeko probabilitatea adierazten du. Kasu honetan,

$$F(y) = \begin{cases} 0, & \text{baldin } y < 0 \\ \frac{1}{4}y, & \text{baldin } y \in [0,\ 4] \\ 1, & \text{baldin } y > 4 \end{cases} \quad \square$$



6.3. irudia. f eta F funtzioen adierazpen grafikoak.

Definizioa: X eta Y bi aldagai jarraituak *askeak* dira baldin eta solik baldin $P((X \leq x) \cap (Y \leq y)) = P(X \leq x) \cdot P(Y \leq y)$ bada.

6.2.3. Parametroak.

Populazioaren ezaugarri bat adierazten duen zorizko aldagai baten propietateak laburki deskribatzen dituen neurriari *parametro* deitzen zaio. Parametroak letra grekoen bidez adieraziko ditugu.

Definizioa: Izan bedi X zorizko aldagai diskretua, haren probabilitate-legea f izanik. X-ren *batezbestekoa* edo *itxaropen matematikoa* honela definitzen da:

$$\mu = E(X) = \sum_{i=1}^{\infty} x_i \cdot f(x_i)$$

Definizioa: Izan bedi X zorizko aldagai jarraitua eta haren dentsitate-funtzioa f. X-ren *batezbestekoa* edo *itxaropen matematikoa* honela definitzen da:

$$\mu = E(x) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

$E(X)$ geometrikoki irudika daiteke, eta X-ren dentsitate-funtzioaren grafikoan oreka-puntua adierazten du. Probabilitate-masa bere inguruan metatzen den balioa adierazten duen zentralizazio-neurria da. Itxaropen matematikoa aldagaiaren unitate berdinetan adierazten da.

Propietateak:

1. $E(a) = a$
2. $E(aX) = aE(X)$
3. $E(X + Y) = E(X) + E(Y)$. Partikularki, $E(a + X) = a + E(X)$
4. X eta Y aldagai askeak badira, $E(X \cdot Y) = E(X) \cdot E(Y)$

6.10. adibidea. Kalkulatu aurreko adibidearen itxarotako batura eta itxarotako distantzien biderkadura.

Em.: $\mu_X = E(X) = \sum_{x=2}^{12} x \cdot P(X = x) = 2 \cdot P(X = 2) + 3 \cdot P(X = 3) + \ldots + 12 \cdot P(X = 12) = 7$ puntu eta

$$\mu_Y = E(Y) = \int_{-\infty}^{\infty} y \cdot f(y)dy = \int_0^4 y \cdot \frac{1}{4}dy = \frac{1}{4}\frac{4^2}{2} = 2 \text{ metro.} \quad \square$$

Definizioa: Izan bedi X zorizko aldagai diskretua, non $\mu = E(X)$ baita, X-ren *bariantza* honela definitzen da:

$$\sigma^2 = Var(X) = E[(X - \mu)^2] = \sum_{i=1}^{\infty}(x_i - \mu)^2 f(x_i).$$

Definizioa: Izan bedi X zorizko aldagai jarraitua, non $\mu = E(X)$ baita, X-ren *bariantza* honela definitzen da:

$$\sigma^2 = Var(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty}(x - \mu)^2 f(x)dx.$$

Definizioa: Izan bedi X zorizko aldagai diskretu edo jarraitua, X-ren *desbideratze estandarra* bariantzaren erro karratu positibo gisa definitzen da:

$$\sigma = \sqrt{Var(X)}.$$

Batezbestekoa eta desbideratze estandarra aldagaiaren unitate berdinetan adierazten da. Hala ere, bariantzaren unitateak aldagaiaren unitate karratuak dira.

Propietateak:
1. $Var(X) = E(X^2) - (EX)^2$
2. $Var(X) \geq 0$
3. $Var(X) = 0 \Leftrightarrow X$ z.a. konstantea bada $(X = a)$.
4. $Var(a\ X) = a^2\ Var(X)$
5. X eta Y z. aldagai askeak badira, $Var(X + Y) = Var(X) + Var(Y)$ eta $Var(X - Y) = Var(X) + Var(Y)$

6.11. adibidea. Kalkulatu aurreko adibidearen bariantza eta desbideratze estandarra.

Em.: $\sigma_X^2 = Var(X) = E(X^2) - E(X)^2$ adierazpena erabiliko dugu. $E(X^2) = \sum_{x=2}^{12} x^2 \cdot P(X = x) = 2^2 \cdot P(X = 2) + 3^2 \cdot P(X = 3) + \ldots + 12^2 \cdot P(X = 12) = 54.8333$ puntu^2.
Hortaz: $\sigma_X^2 = E(X^2) - E(X)^2 = 54.8333 - 7^2 = 5.8333$ puntu^2 eta $\sigma = 2.4152$ puntu.
$E(Y^2) = \int_{-\infty}^{\infty} y \cdot f(y)dy = \int_0^4 \frac{1}{4}\ y^2\ dy = \frac{4^3}{12} = \frac{16}{3}$ m^2.
Hortaz: $\sigma_Y^2 = E(Y^2) - E(Y)^2 = \frac{16}{3} - 2^2 = \frac{4}{3}$ m^2 eta $\sigma = \frac{2\sqrt{3}}{3}$ m. □

6.2.4. Banaketak.

Zientzia esperimentaletan agertzen diren zorizko fenomeno ugariren eredutzat har daitezkeen banaketak aurkitzea Probabilitate-Teoriaren helburuetariko bat da.

Oinarrizko banaketa diskretuen artean, Bernoulli-rena, binomiala, Poisson-ena, hipergeometrikoa eta multinomiala ditugu, besteak beste. Beharbada gehien erabiltzen dena gertaera baten proben errepikapenari dagokion banaketa da, alegia, banaketa binomiala.

Banaketa binomiala

Demagun zorizko esperimentu bat, non proba bakoitzean *arrakasta* (A gertaera) edo *porrota* ($\overline{A}$ gertaera) lortzen den, $P(A) = p$ eta $P(\overline{A}) = 1-p = q$ direlarik. Horrelako esperimentuari *Bernoulli-ren proba* deitzen diogu.

Demagun $n \in \mathbb{N}$ Bernouilliren proba egiten ditugula, probak askeak izanik. Lagin-espazio honen gainean, $X =$ '*arrakasta-kopurua*', *zorizko aldagai binomiala* definitzen dugu, bere heina $Irudi(X) = \{0, 1, 2, \ldots, n\}$ izanik, $X : Bin(n, p)$ adierazirik eta haren probabilitate-legea honako hau da:

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x}, \qquad \forall x = 0, 1, 2, \ldots, n$$

Propietateak:

1. Itxaropen matematikoa $\mu = E(X) = n \cdot p$ da.
2. Bariantza $\sigma^2 = n \cdot p \cdot q$ eta desbideratze estandarra $\sigma = \sqrt{n \cdot p \cdot q}$.

Oharra: `Excel` programaren bidez, $P(X = x) = \binom{n}{x} p^x q^{n-x}$ probabilitatea kalkula daiteke `DISTR.BINOM(`x`;`n`;`p`;FALSO)` adierazpena erabiliz.

6.12. adibidea. Lantegi batean pieza mota berezi bat egiten da. Kutxa bateko pieza akastunen kopurua $Bin(100,\ 0.01)$ banaketa binomialari darraio. Kutxa bat zoriz irekitzen bada, erantzun itzazu ondoko galderak:

a) zein da pieza akastun bakar bat aurkitzeko probabilitatea?

b) zein da gehienez pieza akastun bat aurkitzekoa?

c) zein da espero dugun pieza akastunen kopurua?

d) zein da desbideratzea?

Em.:

Pieza bakoitza kontsideratzean, interesatzen zaigun gertaera (arrakasta gertaera) $A =$ '*akastuna izatea*' bada, ondoko aldagaia definitu dezakegu: $X =$ '*kutxako pieza akastunen kopurua*': $Bin(n,\ n)$ non $n = 100$ kutxako pieza kopurua den eta $p = 0.01$ pieza bakoitza akastuna izateko probabilitatea den.

a) Orduan, pieza akastun bakar bat aurkitzeko probabilitatea $P(X = 1) = \binom{100}{1} p^1 q^{99} = 100 \cdot 0.01 \cdot 0.99^{99} = 0.3697$ da.

b) Gehienez pieza akastun bat aurkitzeko probabilitatea $P(X \leq 1) = P(X = 0) + (X = 1) = \binom{100}{0} p^0 q^{100} + 0.3697 = 0.99^{100} + 0.3697 = 0.3660 + 0.3697 = 0.7357$ da.

c) Espero dugun pieza akastuen kopurua $\mu = E(X) = 100 \cdot 0.01 = 1$ pieza

d) Desbideratze estandarra $\sigma = \sqrt{Var(X)} = \sqrt{100 \cdot 0.01 \cdot 0.99} = \sqrt{0.99} = 0.9950$ pieza. □

Oinarrizko banaketa jarraituen artean, normala, uniformea, esponentziala, beta, gamma, Pearson-en ji karratu, Student-en t, eta Fisher-Snedecor-en F ditugu, besteak beste.

Banaketa normala

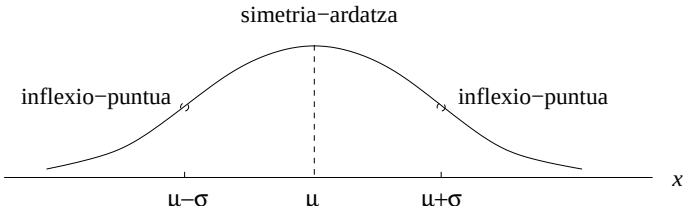
Familia normala zorizko aldagai jarraituen familia garrantzitsuena da. Izan ere, aplikazio praktiko anitz ditu eta interes handiko beste zenbait banaketaren sortzailea da.

Definizioa: Izan bedi X zorizko aldagai jarraitua, non haren heina ardatz erreal osoa baita, batezbestekoa $\mu \in \mathbb{R}$, desbideratze estandarra $\sigma > 0$ eta dentsitate-funtzioa:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \forall x \in \mathbb{R}$$

Orduan X-k *banaketa normalari* jarraitzen dio, $X : \mathcal{N}(\mu, \sigma)$ adierazirik.
Propietateak:

1. Itxaropen matematikoa $E(X) = \mu$.
2. Bariantza $Var(X) = \sigma^2$, eta desbideratze estandarra σ.
3. Banaketa normalaren f dentsitate-funtzioaren grafikoa kanpai itxurako kurba simetrikoa da, eta zentroa μ batezbestekoan dauka, zeina maximoa baita.
4. $x = \mu - \sigma$ eta $x = \mu + \sigma$ abzisako puntuak kurbaren inflexio-puntuak dira. $(\mu-\sigma, \mu+\sigma)$ tartean ahurtasun negatiboa du kurbak, eta handik kanpo, ahurtasun positiboa. Zenbat eta handiagoa izan σ, orduan eta leunagoa da kurba.



6.4. irudia. $\mathcal{N}(\mu, \sigma)$ banaketa normalaren dentsitate-funtzioa.

Gauss-en kanpai guztien artean, $\mu = 0$ eta $\sigma = 1$ baliodunei *kanpai tipiko* deritzegu, eta ordenatu-ardatzarekiko simetrikoa da. Banaketa horrek aparteko garrantzia du, zeren eta beste edozein banaketa normali dagozkion probabilitateak kalkulatzeko erreferentzia gisa erabiliko baita.

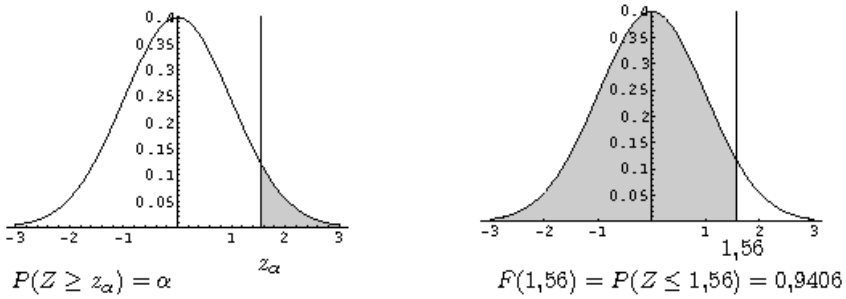
Definizioa: Izan bedi $Z : \mathcal{N}(0, 1)$ z.a. eta $\alpha \in [0, 1]$, $P(Z > z_\alpha) = \alpha$ baldintza betetzen duen balioa, z_α adierazten da eta α *esangura-mailari dagokion puntu kritikoa* deitzen da. Puntu hau aldagaiaren balioa da non balio horren eskuinaldean dentsitate-kurbaren azpian azalera α den. Oharra: `Excel` programaren bidez, $\alpha = P(Z > z_\alpha) = 1 - F(z_\alpha)$ probabilitatea eta $z_\alpha = F^{-1}(1 - \alpha)$ puntu kritikoa kalkula daitezke `1-DISTR.NORM.ESTAND(`z_α`)` eta `DISTR.NORM.ESTAND.INV(`$1-\alpha$`)` adierazpenak erabiliz, hurrenez hurren.

6.13. adibidea. $\mathcal{N}(0,1)$ banaketaren taula erabiliz, kalkulatu a) $P(Z \geq 1.56)$, b) $P(Z < 1.56)$, c) $P(Z \leq -1.56)$, d) $P(Z > -1.56)$, e) $P(-1.72 \leq Z \leq 1.8)$, f) $a \mid P(Z \leq a) = 0.025$, g) $a \mid P(-a \leq Z \leq a) = 0.95$.

Em.:

a) Taulan, puntu batekiko eskuinerantz dagoen azalera agertzen da, 1. zutabean puntu kritikoa, 1. lerroan puntu kritikoaren bigarren hamartarra eta taula barruan azalerak, beraz:

$$P(Z \geq 1.56) = 0.0594$$



6.5. irudia. Banaketa normala: puntu-kritikoa eta banaketa-funtzioa.

Eta banaketaren simetria erabiliz,

b) $P(Z < 1.56) = 1 - P(Z \geq 1.56) = 1 - 0.0594 = 0.9406$.

c) $P(Z \leq -1.56) = P(Z \geq 1.56) = 0.0594$.

d) $P(Z > -1.56) = 1 - P(Z \leq -1.56) = 1 - P(Z \geq 1.56) = 1 - 0.0594 = 0.9406$.

e) $P(-1.72 \leq Z \leq 1.8) = 1 - P(Z > 1.8) - P(Z < -1.72) = 1 - P(Z > 1.8) - P(Z > 1.72) = 1 - 0.0359 - 0.0427 = 0.9214$.

Alderantzizko balioak aurkitzeko:

f) $P(Z \leq a) = 0.025 \Leftrightarrow P(Z \geq -a) = 0.025 \Leftrightarrow z_{0.025} = -a \Leftrightarrow -a = 1.96 \Leftrightarrow a = -1.96$.

g) $P(-a \leq Z \leq a) = 0.95 \Leftrightarrow 1 - 2P(Z > a) = 0.95 \Leftrightarrow P(Z > a) = 0.025 \Leftrightarrow a = z_{0.025} \Leftrightarrow a = 1.96$. □

Demagun $X : \mathcal{N}(\mu, \sigma)$ dugula eta $x_1, x_2 \in \mathbb{R}$, nola kalkula daiteke $P(x_1 < X < x_2)$ probabilitatea? Teorikoki, honela:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

Hala ere, integral horren kalkulua ez da erraza, eta, horretaz gain, μ-ren eta σ-ren balio bakoitzerako kalkulu berriak egin beharko genituzke.

6.1. teorema. *Izan bedi X zorizko aldagaia, non $X : \mathcal{N}(\mu, \sigma)$ baita. Orduan:*

$$Z = \frac{X - \mu}{\sigma}$$

aldagai berriari aldagai normal tipifikatu deritzogu, eta haren banaketa hauxe da: $Z : \mathcal{N}(0, 1)$. Prozesu horren izena tipifikazioa da, eta ondorioz, planteaturiko edozein probabilitate honela kalkula daiteke:

$$P(x_1 < X < x_2) = P\left(\frac{x_1 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{x_2 - \mu}{\sigma}\right) \equiv P(z_1 < Z < z_2).$$

6.14. adibidea. Izan bedi X ='*Kubako tenperatura*' $N(30,\ 5)$ banaketari darraion aldagaia, $\mathcal{N}(0, 1)$ banaketaren taula eta tipifikazioa erabiliz, kalkulatu a) $P(X > 36)$, b) $P(X < 40)$, c) $P(25 < X < 35)$, d) $P(20 < X < 40)$, e) $P(15 < X < 45)$.

Em.:

a) $P(X > 36) = P(\frac{X-30}{5} > \frac{36-30}{5}) = P(Z > 1.20) = 0.1151.$

b) $P(X < 40) = P(\frac{X-30}{5} < \frac{40-30}{5}) = P(Z < 2.00) = 1 - P(Z \geq 2.00) = 1 - 0.0228 = 0.9772.$

c) $P(25 < X < 35) = P(\frac{25-30}{5} < \frac{X-30}{5} < \frac{35-30}{5}) = P(-1.00 < Z < 1.00) = 1 - 2P(Z \geq 1.00) = 1 - 2 \cdot 0.1587 = 0.6826.$

d) $P(20 < X < 40) = P(\frac{20-30}{5} < \frac{X-30}{5} < \frac{40-30}{5}) = P(-2.00 < Z < 2.00) = 1 - 2P(Z \geq 2.00) = 1 - 2 \cdot 0.0228 = 0.9544.$

e) $P(15 < X < 45) = P(\frac{15-30}{5} < \frac{X-30}{5} < \frac{45-30}{5}) = P(-3.00 < Z < 3.00) = 1 - 2P(Z \geq 3.00) = 1 - 2 \cdot 0.00135 = 0.9973.$ □

Propietatea: *Izan bitez $X_1, X_2, \ldots, X_n$ z.a. normalak eta askeak eta $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R}$, orduan, $X = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n$ z.a. banaketa normalari darraio, non $EX = \alpha_1 E(X_1) + \alpha_2 E(X_2) + \ldots +$*

$\alpha_n E(X_n)$ *eta* $Var(X) = \alpha_1^2 Var(X_1) + \alpha_2^2 Var(X_2) + \ldots + \alpha_n^2 Var(X_n)$ *diren.*

Partikularki, izan bitez $X_1, X_2, .., X_n$ *z.a. askeak, non* $X_i : \mathcal{N}(\mu, \sigma)$, $\forall i = 1, 2, \ldots, n$. *Orduan,*

1. $\Sigma_n \equiv X_1 + X_2 + \ldots + X_n : \mathcal{N}(n\mu, \sigma\sqrt{n})$.

   Edo baliokideki, Σ_n *z.a.-ren aldagai tipikikatua* $\dfrac{\Sigma_n - n\mu}{\sigma\sqrt{n}} : \mathcal{N}(0,1)$.

2. $\overline{X} \equiv \frac{X_1+X_2+\ldots+X_n}{n} : \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

   Edo baliokideki, $\overline{X}$ *z.a.-ren aldagai tipikikatua* $\dfrac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} : \mathcal{N}(0,1)$.

Banaketa normalari (doi-doi ala gutxi gorabehera) jarraitzen zaion z.a. jarraien kopurua izugarri handia da. Honen azalpen matematikoa, limitearen teorema zentrala delakoa da, teoria estatistikoaren garrantzitsuenetariko bat. Z.a. ugariak, batzen diren oso antzeko bariantza finituko faktore txiki askeen konkurrentziaz sortzen dira.

6.2. teorema. Limitearen Teorema Zentrala

Izan bitez $X_1, X_2, \ldots, X_n$, μ *batezbesteko eta* $\sigma^2 \neq 0$ *bariantza finituko z.a. aske eta berdinki banatuen segida. Orduan,*

$$\Sigma_n = X_1 + X_2 + \ldots + X_n \quad \textit{z.a. asintotikoki} \quad \mathcal{N}(n\mu, \sigma\sqrt{n}) \quad \textit{da.}$$

Edo era baliokide batera esanda,

$$\frac{\Sigma_n - n\mu}{\sigma\sqrt{n}} \approx \mathcal{N}(0,1)$$

n, aldagai kopurua, zenbat eta handiagoa izan, hurbilketa gero eta hobea izango da. Praktikan, $n \geq 30$ denean, onartuko dugu hurbilketa hau.

Korolarioa: Izan bitez $X_1, X_2, \ldots, X_n$, μ batezbesteko eta $\sigma^2 \neq 0$ bariantza finituko z.a. aske eta berdinki banatuen segida. Orduan,

$$\overline{X} \equiv \frac{X_1 + X_2 + \ldots + X_n}{n} \approx \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Edo era baliokide batera esanda,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx \mathcal{N}(0,1)$$

6.15. adibidea. Demagun aurreko 6.12 adibideari lotuta, 90 kutxa irekitzen direla eta pieza akastunen kopuru osoa zenbatzen dela. Erantzun itzazu ondoko galderak:

a) zein da espero dugun pieza akastunen kopuru osoa?

b) zein da bariantza?

c) zein da pieza akastunen kopuru osoaren banaketa?

d) zein da pieza akastunen kopuru osoa 100 baino txikiagoa izateko probabilitatea?

Em.:

Kutxa bakoitzeko ondoko aldagaia definitzen dugu: $X_i =$ '*i. kutxako pieza akastunen kopurua*': $Bin(100,\ 0.01)$, $\forall i = 1, 2, \ldots, 90$. Orduan, $X_1, X_2, \ldots, X_{90}$ z.a. askeak eta berdinki banatuak dira. Beraz, pieza akastunen kopuru osoa ondoko aldagaia da: $\sum = X_1 + \ldots + X_{90}$.

a) $E(\sum) = 90 \cdot E(X_i) = 90 \cdot 1 = 90$ pieza.

b) $Var(\sum) = 90 \cdot Var(X_i) = 90 \cdot 0.99 = 89.1$ pieza^2.

c) Beraz, limitearen teorema zentrala erabil daiteke, $\sum \approx \mathcal{N}(90,\ 9.44)$

d) Eskatutako probabilitatea ondokoa da: $P(\sum < 100) = [tipifikatuz] = P(\frac{\sum -90}{9.44} < \frac{100-90}{9.44}) = P(Z < 1.06) = 1 - 0.1446 = 0.8554$ □

6.3. teorema. Moivre-ren Teorema

Izan bedi X banaketa binomiala duen aldagaia, $X : Bin(n, p)$, non $p \neq 0, 1$ eta n handia baitira. Orduan,

$$\frac{X - np}{\sqrt{npq}} \approx \mathcal{N}(0, 1)$$

Edo bestela esanda,

$$X \approx \mathcal{N}(np, \sqrt{npq})$$

Praktikan, $n > 50$ eta $p \in (0.1,\ 0.9)$ direnean edo baliokideki, $np > 5$ eta $nq > 5$ direnean, onartzen da hurbilketa hau:

$$Bin(n, p) \approx \mathcal{N}(np, \sqrt{npq})$$

ji karraturen banaketa

Definizioa: Izan bedi $p > 0$, *gamma funtzioa* ondoko aldagai errealeko funtzio erreala da: $\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx$.
Definizioa: X z.a. $n \in \mathbb{N}$ askatasun-gradutako ji karratu banaketari darraio, baldin

$$f(x) = \begin{cases} \frac{(\frac{1}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} \cdot x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}}, & \text{baldin } x > 0 \\ 0, & \text{baldin } x \leq 0 \end{cases}$$

$X : \chi^2_n$ adierazten delarik.
Propietateak:
1. Itxaropen matematikoa $E(X) = n$
2. Bariantza $Var(X) = 2n$ eta desbiderapen estandarra $\sqrt{2n}$
3. Dentsitate funtzioa $[0, +\infty)$ tartean definituta dago, jarraitua da eta ez da simetrikoa.
4. n emendatzerakoan, $n \to \infty$, beraren grafikoa normalarenera hurbilduz doa. Bereziki, $n > 30$ bada, ondoko hurbilketa erabiltzen da: $\chi^2_{\alpha;n} \approx \frac{1}{2}(z_\alpha + \sqrt{2n-1})^2$
5. Izan bitez $Z_1, Z_2, .., Z_n$ z.a. independenteak, guztiak $\mathcal{N}(0,1)$ izanik. Orduan, $Z_1^2 + Z_2^2 + \ldots Z_n^2 : \chi^2_n$.

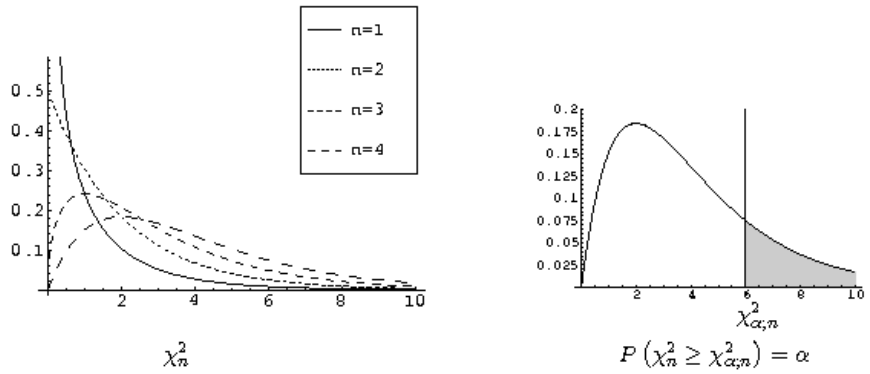
Definizioa: Izan bedi $X : \chi^2_n$ z.a. eta $\alpha \in [0,1]$, $P(X > \chi^2_{\alpha;n}) = \alpha$ baldintza betetzen duen balioa, $\chi^2_{\alpha;n}$ adierazten da eta α *esangura-mailari dagokion puntu kritikoa* deitzen da. Puntu hau aldagaiaren balioa da non balio horren eskuinaldean dentsitate-kurbaren azpian azalera α den. Oharra: `Excel` programaren bidez, $\alpha = P(\chi^2_n > \chi^2_{\alpha;n})$ probabilitatea eta $\chi^2_{\alpha;n}$ puntu kritikoa kalkula daitezke hurrenez hurren ondoko adierazpenak erabiliz: `DISTR.CHI(`$\chi^2_{\alpha;n}$`;`n`)` eta `PRUEBA.CHI.INV(`α`;`n`)`.

6.16. adibidea. χ^2_n banaketaren taula erabiliz kalkula itzazu: a) $P(\chi^2_{10} \geq 2.5582)$, b) $P(\chi^2_{10} \leq 2.5582)$, c) $P(\chi^2_{11} > 20)$, d) $P(3.2470 \leq \chi^2_{10} \leq 20.4832)$, e) $\chi^2_{0.95;16}$, f) $\chi^2_{0.80;10}$, g) $\chi^2_{0.1;100}$, h) $P(\chi^2_{100} > 72)$.

Em.:

a) Taulan puntu batekiko eskuinerantz dagoen azalerak agertzen direnez, (1. lerroan azalerak, 1. zutabean askatasun-graduak eta taula barruan puntu kritikoak), beraz:

$$P(\chi^2_{10} \geq 2.5582) = 0.99$$



6.6. irudia. Ji karraturen banaketa: adibideak eta puntu-kritikoa.

b) $P(\chi^2_{10} \leq 2.5582) = 1 - P(\chi^2_{10} > 2.5582) = 1 - 0.99 = 0.01$

c) $P(\chi^2_{11} > 20)$ kalkulatzeko, 20 zenbakia taulan agertzen ez denez, $P(\chi^2_{11} > 19.6751) = 0.05$ eta $P(\chi^2_{11} > 21.9200) = 0.025$ kontuan hartuta, ontzat hartuko dugun hurrengo hurbilketa lortuko dugu: $P(\chi^2_{11} > 20) \approx 0.0464$

d) $P(3.2470 \leq \chi^2_{10} \leq 20.4832) = P(\chi^2_{10} \geq 3.2470) - P(\chi^2_{10} > 20.4832) = 0.975 - 0.025 = 0.95$

e) $\chi^2_{0.95;16} = 7.9616$

f) $\chi^2_{0.80;10}$ kalkulatzeko, 0.80 probabilitatea taulan agertzen ez denez, $P(\chi^2_{10} > 15.9872) = 0.10$ eta $P(\chi^2_{10} > 4.8652) = 0.90$ kontuan hartuta, interpolatuz $\chi^2_{0.80;10} \approx 6.2555$

Ji-karraturen propietateak erabiliz

g) $\chi^2_{0.1;100} \approx \frac{1}{2}(z_{0.1} + \sqrt{2 \cdot 100 - 1})^2 = \frac{1}{2}(1.28 + \sqrt{199})^2 = 118.3758$

h) $P(\chi^2_{100} > 72) = \alpha \Leftrightarrow \chi^2_{\alpha;100} = 72 \Leftrightarrow 72 \approx \frac{1}{2}(z_\alpha + \sqrt{2 \cdot 100 - 1})^2 \Leftrightarrow 72 \approx \frac{1}{2}(z_\alpha + \sqrt{199})^2 \Leftrightarrow z_\alpha = \sqrt{2 \cdot 72} - \sqrt{199} = -2.1067 \Leftrightarrow \alpha = P(Z > -2.1067) \approx 1 - P(Z > 2.11) = 1 - 0.0174 = 0.9826$ □

Student-en t banaketa

Definizioa: Izan bitez $p, q > 0$, *beta funtzioa* ondoko aldagai biko funtzio erreala da: $\beta(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx$.
Definizioa: X z.a. $n \in \mathbb{N}$ askatasun-gradutako Student-en t banaketari darraio, baldin

$$f(x) = \frac{1}{\beta(\frac{1}{2}, \frac{n}{2})} \cdot \frac{1}{\sqrt{n}} \frac{1}{(1 + \frac{x^2}{n})^{\frac{n+1}{2}}}, \quad \forall x \in \mathbb{R}$$

$X : t_n$ adierazten delarik.

Propietateak:

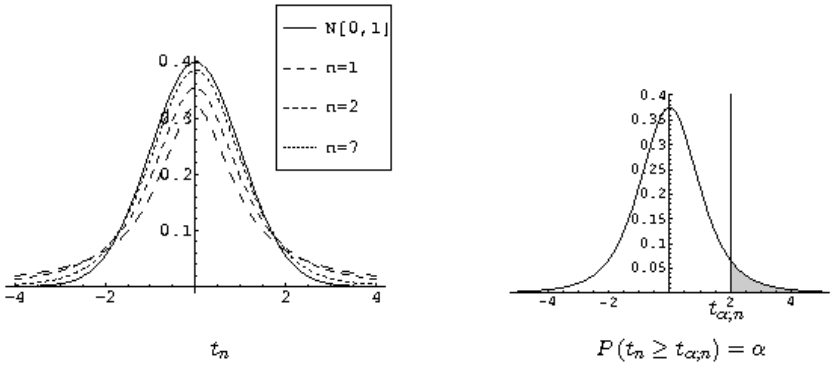
1. Itxaropen matematikoa $E(X) = 0$
2. Bariantza $Var(X) = \frac{n}{n-2} > 1$ (baldin $n > 2$ bada) eta $\lim_{n\to\infty} Var(X) = 1$
3. Dentsitate funtzioa $\mathbb{R}$-n definituta dago, jarraitua eta simetrikoa da
4. Banaketa normalarekin konparatuz, t banaketa erdialdean normal tipikatua baino baxuagoa da eta isatsetan altuagoa da. Baldin $n \to \infty$, orduan $t_n \to \mathcal{N}(0,1)$. Bereziki, $n > 30$ denean, $t_n \approx Z : \mathcal{N}(0,1)$ hurbilketa onartzen da.
5. Izan bitez $Z : \mathcal{N}(0,1)$ z.a. eta $X : \chi^2_n$ n askatasun gradutako ji-karratu z.a. askeak. Orduan, $\frac{Z}{\sqrt{X/n}} : t_n$.

Definizioa: Izan bedi $X : t_n$ z.a. eta $\alpha \in [0,1]$, $P(X > t_{\alpha;n}) = \alpha$ baldintza betetzen duen balioa, $t_{\alpha;n}$ adierazten da eta α *esangura-mailari dagokion puntu kritikoa* deitzen da. Puntu hau aldagaiaren balioa da non balio horren eskuinaldean dentsitate-kurbaren azpian azalera α den. Oharra: `Excel` programaren bidez, $\alpha = P(t_n > t_{\alpha;n})$ probabilitatea eta $t_{\alpha;n}$ puntu kritikoa kalkula daitezke ondoko adierazpenak, hurrenez hurren, erabiliz: `DISTR.T(`$t_{\alpha;n}$`;`n`)` eta `DISTR.T.INV(`$2\cdot\alpha$`;`n`)`.

6.17. adibidea. t_n banaketaren taula erabiliz kalkula itzazu: a) $P(t_{10} > 2)$, b) $P(1 < t_5 < 3)$, c) $t_{0.05;10}$, d) a non $P(t_{10} \le a) = 0.90$, e) b non $P(t_{10} \le b) = 0.20$, f) d non $P(-d \le t_{10} \le d) = 0.95$, g) $t_{0.025;900}$.

Em.:

Taulan puntu batekiko eskuinerantz dagoen azalerak agertzen direnez, (1. lerroan azalerak, 1. zutabean askatasun-graduak eta taula barruan puntu kritikoak), beraz:



6.7. irudia. Student-en t banaketa: adibideak eta puntu-kritikoa.

a) $P(t_{10} > 2)$ kalkulatzeko, 2 zenbakia taulan agertzen ez denez, $P(t_{10} > 1.8125) = 0.05$ eta $P(t_{10} > 2.2281) = 0.025$ kontuan hartuta, interpolatuz ontzat hartuko dugun hurrengo hurbilketa lortuko dugu $P(t_{10} > 2) \approx 0.0387$

b) $P(1 < t_5 < 3) = P(t_5 > 1) - P(t_5 \geq 3)$ eta 1 zein 3 taulan agertzen ez direnez, $P(t_5 > 0.9195) = 0.20$, $P(t_5 > 1.4759) = 0.10$, $P(t_5 > 2.5706) = 0.025$ eta $P(t_5 > 3.3649) = 0.010$ kontuan hartuta, interpolatuz $P(1 < t_5 < 3) = P(t_5 > 1) - P(t_5 \geq 3) \approx 0.1855 - 0.0169 = 0.1686$

c) $P(t_{10} \geq t_{0.05;10}) = 0.05 \Rightarrow t_{0.05;10} = 1.8125$

d) $P(t_{10} \leq a) = 0.90 \Leftrightarrow P(t_{10} \geq a) = 0.10 \Leftrightarrow a = t_{0.10;10} = 1.3722$

e) $P(t_{10} \leq b) = 0.20 \Leftrightarrow P(t_{10} \geq -b) = 0.20 \Leftrightarrow -b = t_{0.20;10} \Leftrightarrow b = -t_{0.20;10} = -0.8791$

f) $P(-d \leq t_{10} \leq d) = 0.95 \Leftrightarrow 1 - 2P(t_{10} > d) = 0.95 \Leftrightarrow P(t_{10} > d) = 0.025 \Leftrightarrow d = t_{0.025;10} = 2.2281$

g) $t_{0.025;900} \approx z_{0.025} = 1.96$ □

Fisher-Snedecor-en banaketa

Definizioa: X z.a. $m \in \mathbb{N}$ eta $n \in \mathbb{N}$ askatasun-graduetako Fisher-Snedecor-en banaketari darraio, baldin

$$f(x) = \begin{cases} \frac{1}{\beta(\frac{m}{2},\frac{n}{2})} \cdot (\frac{m}{n})^{\frac{m}{2}} \frac{x^{\frac{m}{2}-1}}{(1+\frac{m\cdot x}{n})^{\frac{m+n}{2}}}, & \text{baldin } x > 0 \\ 0, & \text{baldin } x \leq 0 \end{cases}$$

$X : \mathcal{F}_{m,n}$ adierazten delarik.

Propietateak:

1. Itxaropen matematikoa $E(X) = \frac{m}{m-2}$ (baldin $m > 2$ bada)
2. Bariantza $Var(X) = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}1$ (baldin $m > 4$ bada)
3. Dentsitate funtzioa $[0, +\infty)$ tartean definituta dago, jarraitua da baina ez da simetrikoa
4. Izan bitez $X : \chi^2_m$ eta $Y : \chi^2_n$ z.a. askeak. Orduan, $\dfrac{X/m}{Y/n} : \mathcal{F}_{m,n}$ z.a. m eta n askatasun gradutako Fisher-Snedecor-en banaketari darraio.
5. $P(\mathcal{F}_{m,n} < a) = P(\mathcal{F}_{n,m} > \frac{1}{a})$, $\forall a > 0$
6. $F_{1-\alpha;m,n} = 1/F_{\alpha;n,m}$

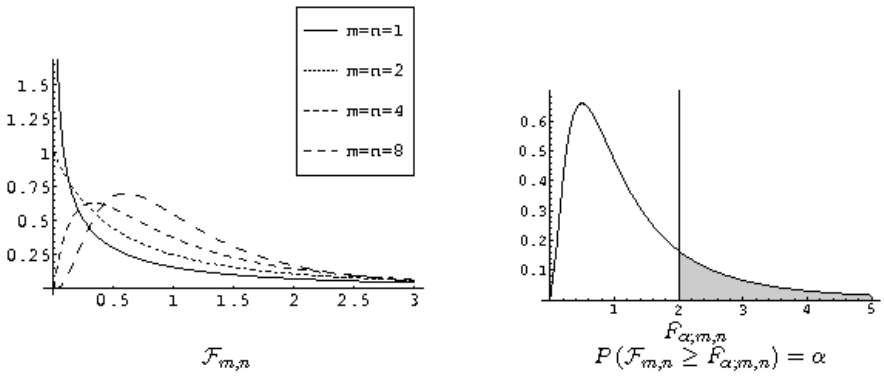
Definizioa: Izan bedi $X : \mathcal{F}_{m,n}$ z.a. eta $\alpha \in [0,1]$, $P(X > F_{\alpha;m,n}) = \alpha$ baldintza betetzen duen balioa, $F_{\alpha;m,n}$ adierazten da eta α *esangura-mailari dagokion puntu kritikoa* deitzen da. Puntu hau aldagaiaren balioa da non balio horren eskuinaldean dentsitate-kurbaren azpian azalera α den.
Oharra: `Excel` programaren bidez, $\alpha = P(\mathcal{F}_{m,n} > F_{\alpha;m,n})$ probabilitatea eta $F_{\alpha;m,n}$ puntu kritikoa kalkula daitezke `DISTR.F(`$F_{\alpha;n}; m; n$`)` eta `DISTR.F.INV(`$\alpha; m; n$`)` adierazpenak erabiliz, hurrenez hurren.

6.18. adibidea. $\mathcal{F}_{m,n}$ banaketaren taula erabiliz kalkula itzazu: a) $F_{0.01;10,12}$, b) $F_{0.95;15,20}$, c) $P(2.5 \leq \mathcal{F}_{10,20} \leq 3)$, d) $a \mid P(\mathcal{F}_{10,15} \leq a) = 0.95$, e) $b \mid P(\mathcal{F}_{10,14} < b) = 0.01$, f) $c \mid P(\mathcal{F}_{10,16} \geq c) = 0.99$.

Em.:

Taulan puntu batekiko eskuinerantz dagoen azalerak agertzen direnez, (1. lerroan zenbakitzaileen a.g, 1. zutabean izendatzaileen a.g., 2. zutabean azalerak eta taula barruan puntu kritikoak), beraz:

a) $F_{0.01;10,12} = 4.2961$



6.8. irudia. Fisher-Snedecor-en banaketa: adibideak eta puntu-kritikoa.

b) $F_{0.95;15,20} = 1/F_{0.05;20,15} = 1/2.3275 = 0.4296$

c) $P(2.5 \leq \mathcal{F}_{10,20} \leq 3) = P(\mathcal{F}_{10,20} \geq 2.5) - P(\mathcal{F}_{10,20} > 3)$ kalkulatzeko $P(\mathcal{F}_{10,20} > 2.3476) = 0.05$, $P(\mathcal{F}_{10,20} > 2.7737) = 0.025$ eta $P(\mathcal{F}_{10,20} > 3.3682) = 0.01$ kontuan hartuta, interpolatuz $P(2.5 \leq \mathcal{F}_{10,20} \leq 3) = P(\mathcal{F}_{10,20} \geq 2.5) - P(\mathcal{F}_{10,20} > 3) \approx 0.0411 - 0.0193 = 0.0218$

d) $P(\mathcal{F}_{10,15} \leq a) = 0.95 \Leftrightarrow P(\mathcal{F}_{10,15} > a) = 0.05 \Leftrightarrow a = F_{0.05;10,15} = 2.5437$

e) $P(\mathcal{F}_{10,14} < b) = 0.01 \Leftrightarrow P(\mathcal{F}_{14,10} > 1/b) = 0.01 \Leftrightarrow 1/b = F_{0.01;14,10} \Leftrightarrow b = 1/4.6074 = 0.2170$

f) $P(\mathcal{F}_{10,16} \geq c) = 0.99 \Leftrightarrow c = F_{0.99;10,16} = 1/F_{0.01;16,10} = 1/4.5276 = 0.2209$ □

6.3. Inferentzia estatistikoa.

Edozein ikerkuntza estatistikoren helburua da aurrez finkaturiko populazioko aleetan agertzen den ezaugarri (aldagai) bat aztertzea. Populazioa osatzen duten ale guztien informazioa ezagutzen denean, *zentsua* egiten dela esaten da. Baina, hori ez da beti posiblea edo egokia izaten. Bai populazioa infinitua delako, edo proba hondagarriak direlako, edo elementu potentzialek osatutako populazio bat aztertzen ari garelako, edo luzeegia edota garestiegia izango litzatekeelako.

Inferentzia estatistikoa, Probabilitate-Teorian oinarritua, ondoko helburuak betetzen dituen metodo multzoa da: emaitzen inferentziak edo orokortzeak egitea eta haien konfiantza-maila neurtzea; eta aurreko etapan lortutako emaitzak interpretatzea eta erabakiak hartzea.

Inferentzia estatistikoa arlo bitan sailka daiteke:

- ▷ *Estimazioa*: laginean lortutako estatistikoen bidez, populazioko parametro ezezagunetara hurbiltzea da.
- ▷ *Hipotesi-kontrastea*: hipotesia planteatu ondoren, lagineko estatistikoen balioen arabera hipotesia onartzeko edo errefusatzeko metodoan datza. Hipotesiak populazioko parametro bati edo batzuei buruzkoak izan daitezke edo doikuntz egokitasunari, askatasunari, homogeneotasunari, zorizkotasunari eta abarreri buruzkoak.

6.3.1. Puntu-estimatzaileak eta konfiantza-tarteak.

Estimazioa populazio parametroak estimatzean datza. Bereziki, populazio-batezbestekoa, μ, eta populazio-bariantza, σ^2, estimatu nahi izango ditugu. Partikularki, X aldagai binomiala den kasuan, bere populazio-proportzioa, p, interesatuko zaigu.
Populazioaren parametro batera hurbiltzeko erabiltzen den estatistikoa *estimatzailea* da. Estimatzaile posible guztien artean hoberenak kontsideratzen dira, hots, estatistikoki zentratuak eta efizienteenak direnak.

Definizioa: Izan bedi μ batezbestekoa eta σ^2 bariantza dituen populaziotik ateratako $X_1, X_2, \ldots, X_n$ zorizko lagin bakuna. Orduan:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n},$$

lagin-batezbestekoa izeneko zorizko aldagaia μ-ren puntu-estimatzailea da.

Definizioa: Izan bedi μ batezbestekoa eta σ^2 bariantza dituen populaziotik ateratako $X_1, X_2, \ldots, X_n$ zorizko lagin bakuna. Orduan,

$$S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

lagin-kuasibariantza izeneko zorizko aldagaia σ^2-ren puntu-estimatzailea da.

Definizioa: Izan bedi X, n eta p parametroetako banaketa binomialari darraion zorizko aldagaia, $X : Bin(n, p)$. Orduan,

$$\hat{P} = \frac{X}{n} = \frac{n \text{ probatan arrakasta-kopurua}}{\text{proba-kopurua}}$$

lagin-proportzioa izeneko zorizko aldagaia populazio-proportzioaren, p, puntu-estimatzailea da.

Estimatzaileak lagin berezi batean hartzen duen zenbakizko balioari parametroaren *estimazioa* deitzen zaio eta letra xehez adierazten da. Laburbilduz, $\hat{\mu} = \overline{x}$, $\hat{\sigma} = s_{n-1}$ eta $\hat{p} = \hat{p}$.
Baina, zein da estimatzailearen doitasuna? Lagin-balioak zein puntutaraino dira populazio osoaren adierazgarriak? Zer konfianza-maila daukate ondorioak? Galdera hauek erantzuteko populazioari buruzko hipotesiak eta lagin-errorea aztertu behar ditugu.
Ikertzaileak *konfiantza-maila* finkatzen du, hots, lagin-balioa populazio-parametrotik errorea baino gehiago ez urruntzeko probabilitatea. Konfiantza-maila %100tik hurbil dagoen portzentajea denez, $\%(1-\alpha)\cdot 100$ moduan adierazten da, non $\alpha \in [0,\ 1]$ balioa nahi duen bezain txikia den. Izan bedi θ populazio-parametroa, orduan $P(B < \theta < G) = 1 - \alpha$ bada, $I_\theta^{1-\alpha} = (B,\ G)$ konfiantza-tartea (KT) da. Esate baterako, maiz erabiltzen den α balioa 0.05 da, hau da, $1 - \alpha = 0.95$ eta %95eko konfiantza mailari dagokio; $P(B < \theta < G) = 0.95 \Rightarrow I_\theta^{0.95} = (B,\ G)$. Ikertzaileak ere *lagin-errorea*, ε, finka dezake, hots, populazio-balioaren

eta zorizko lagin baten bidez estimatutako balioen artean espero izandako desberdintasun handiena. Matematikoki, populazio-batezbestekoa estimatzerakoan, $P(|\mu - \hat{\mu}| < \varepsilon) = P(\hat{\mu} - \varepsilon < \mu < \hat{\mu} + \varepsilon) = 1 - \alpha$ eta konfiantza-tartea $I_{\mu}^{1-\alpha} = (\mu - \varepsilon,\ \mu + \varepsilon)$ da.
Gure helburua μ, σ^2 eta p populazio parametroetarako konfiantza-tarteak eraikitzea da, hots, lagin-erroreak kalkulatzea, lagin bakar baten kasuan, hots, $I_{\mu}^{1-\alpha}$, $I_{\sigma^2}^{1-\alpha}$ eta $I_{p}^{1-\alpha}$. Gainera, bi populazio kasuan, batezbestekoen diferentziarako, barientzen zatidurarako eta proportzioen diferentziarako, hau da, $I_{\mu_1-\mu_2}^{1-\alpha}$, $I_{\sigma_1^2/\sigma_2^2}^{1-\alpha}$ eta $I_{p_1-p_2}^{1-\alpha}$.

6.19. adibidea. Ezaguna denez, automobilen CO_2-aren isurien arabera A letratik G letrarainoko ondoko sailkapena dugu.

A	≤ 100 gr/km
B	101-120 gr/km
C	121-150 gr/km
D	151-165 gr/km
E	166-185 gr/km
F	186-225 gr/km
G	> 225 gr/km

A, B eta C kategoriatan daudenak *garbiak* kontsideratzen dira. Eusko Jaurlaritzako Ingurumen Sailak, emisioak aztertzeko asmoz, talde ikertzaile bati ikerketa bat eskatu zion. Taldeak hurrengo galderak aztertu behar zituen Eukal Autonomia Erkidego osorako:

1. 2004ko autoak garbiak ziren, batez beste?
2. Nolakoa zen 2004ko autoen emisioen sakabanapena?
3. 2004ko diesel motordun autoen %92a eta gasolinazko %40a garbiak ziren?
4. Lur orotako ibilgailu eta monobolumenen batezbesteko emisioen eta gainontzeko autoen arteko diferentzia 30 gramokoa zen 2004 urtean?
5. 2008 urtean CO_2-aren isuriak 10 gramo jaitsi dira 2004 urtearekiko?
6. Bi motako autoen isurien sakabanapena berdina zen 2004 urtean?
7. 2004ko diesel motordun auto garbien eta gasolinazkoen artean, diferentzia esanguratsurik dago?

Talde ikertzaileak galdera hauek aztertzeko, hasierako ikerketa batean 14 autoren lagina kontsideratu zuen, beraien emisioak 2004 urtean eta 2008 urtean neurtuz. Alde batetik, autoak erregaiaren arabera sailkatu

ziren: lehenengo erdia diesel motorduna eta bigarrena gasolinazkoa; bestalde, motaren arabera: I motakoak lur orototako ibilgailu eta monobolumenak izan ziren eta II motakoak gainontzeko ibilgailuak. Ondoko taulan emaitzak adierazten dira kilometroko gramotan neurtuta.

I mota (lur orotako ibilgailu eta monobolumenak)

2004	132	141	165	169	177	183
2008	130	135	158	162	170	181

II mota (gainontzeko ibilgailuak)

2004	105	117	125	137	140	149	146	162
2008	110	115	131	130	125	133	140	152

Banan banan aurreko galderak erantzungo eta ondorioen adierazgarritasuna komentatuko ditugu hurrengo azpiataletan. □

6.3.2. KT populazio-batezbestekorako.

6.4. teorema. (Limitearen Teorema Zentralaren aplikazioa)

Izan bedi μ batezbestekoa eta σ^2 bariantza dituen populaziotik ateratako $X_1, X_2, \ldots, X_n$ zorizko lagin bakuna, $n > 30$ izanik. Orduan,

$$\overline{X} \approx \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Bereziki, populazioaren banaketa normala bada, laginaren tamaina n edozein delarik, $\overline{X}$ zorizko aldagaiaren banaketa zehatz mehatz normala izango da (asintotikoki normala izan beharrean).

Definizioa: Bariantza ezaguna σ^2 duen populazio batetik ateratako n tamainako laginaren batezbestekoa $\overline{x}$ izanik, $\%(1-\alpha)\cdot 100$-*eko μ-ren konfiantza-tartea* hauxe da:

$$I_\mu^{1-\alpha} = (\overline{x} - \varepsilon,\ \overline{x} + \varepsilon)\,,\ \varepsilon = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

Tarte hau, populazioa normala denean edo laginaren tamaina $n > 30$ denean baliozkoa da.
Izan ere,

$$P(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = P(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu - \overline{X} < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) =$$

$$P(z_{\alpha/2}\frac{\sigma}{\sqrt{n}} > \overline{X} - \mu > -z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = P(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}) =$$
$$P(|Z| < z_{\alpha/2}) = 1 - 2P(Z \geq z_{\alpha/2}) = 1 - 2(\alpha/2) = 1 - \alpha$$

6.20. adibidea. σ^2 ezaguna izanik, zein da *% 95eko μ-ren konfiantza-tartea*?

Em.: $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$ eta $z_{\alpha/2} = z_{0.025} = 1.96$; beraz,

$$I_\mu^{0.95} = \left(\overline{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}\ ,\ \overline{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right). \quad \square$$

Interpretazioa: n tamainako laginetan oinarriturik atera daitezkeen % 95eko konfiantza-tarteen % 95ek μ bere barne duela esan daiteke. Hau da, 100 ikertzaileek lagin bana ateratzekotan, gehienez 5en %95eko konfiantza-tarteek ez dute μ bere barnean izango. Beste modu batean esanda, baldin $\overline{x}$ lagin-batezbestekoa μ-ren estimatzaile moduan erabiltzen badugu, ε lagin-errorea baino handiago ez dela izango esan daiteke %95 konfiantza mailaz.

Tartearen luzera bi aldiz ε lagin-errorea da, $\varepsilon = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ delarik. Zenbat eta n lagin-tamaina handiagoa izan, hainbat eta laburragoa izango da konfiantza-tartea. Eta zenbat eta konfiantza-maila handiagoa (α txikiagoa), konfiantza-tartea gero eta luzeagoa izango da.

Laginaren tamainaren kalkulua
Demagun σ^2 ezaguna duen populazioaren batezbestekoa estimatu nahi dela. $\%(1-\alpha)\cdot 100$-eko μ-ren konfiantza-tartearen lagin-errorea ε izatera baldintzatuz. Zein da, baldintza hori bete dezan, beharrezkoa den laginaren tamaina?

$$\varepsilon = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow n = z_{\alpha/2}^2 \cdot \frac{\sigma^2}{\varepsilon^2}.$$

Bariantza ezaguna daukan banaketaren batezbestekorako konfiantza-tartearen kalkulua aztertu dugu. Baina egoera hori artifizial samarra da, zeren eta populazioaren bariantza ez baita eskuarki ezagutzen, eta estimatu behar izaten baita. σ ezezaguna denean, bere balioa, s_{n-1},

kuasibariantzaren erro-karratuaren bidez hurbil dezakegu. Arazoa, kasu honetan, $\dfrac{\overline{X}-\mu}{\sqrt{\frac{S^2_{n-1}}{n}}}$ aldagaiaren banaketa normala ez izatean datza.

6.5. teorema. *Izan bedi μ batezbestekoa eta σ^2 bariantza ezezaguna dituen populaziotik ateratako X_1, X_2, $X_3, \ldots$ X_n zorizko lagin bakuna. Orduan, lagin-batezbestekoa $\overline{X}$ zorizko aldagaiaren banaketa ondokoa da:*

$$\frac{\overline{X}-\mu}{\sqrt{\frac{S^2_{n-1}}{n}}} : t_{n-1}$$

Definizioa: Bariantza ezezaguna σ^2 duen populazio normal batetik ateratako n tamainako laginaren batezbestekoa $\overline{x}$ izanik, $\%100 \cdot (1-\alpha)$ *mailako μ-ren konfiantza-tartea* hauxe da:

$$I_\mu^{1-\alpha} = (\overline{x}-\varepsilon,\ \overline{x}+\varepsilon)\,,\ \varepsilon = t_{\alpha/2;n-1} \cdot \frac{s_{n-1}}{\sqrt{n}}$$

Hain zuzen, Student-en t banaketak ondokoa betetzen du: $P(-t_{\alpha/2;n-1} < t_{n-1} < t_{\alpha/2;n-1}) = 1-\alpha$. Bereziki, $n \geq 30$ denean, gogora ezazu $t_{\alpha/2;n-1} \approx z_{\alpha/2}$ hurbilketa ontzat har dezakegula.

6.21. adibidea. Demagun normaltasuna.

a) Kalkula ezazu 2004ko batezbesteko emisiorako %95eko KT, populazio-desbideratze estandarra 25 gr/km-koa bada.

b) Kalkula ezazu 2004ko batezbesteko emisiorako %95eko KT, populazio-bariantza ezezaguna bada.

c) Erantzun ezazu 1. galdera: 2004ko autoak garbiak ziren, batez beste?

d) Demagun gehienez 3 gr/km-ko lagin-errorea nahi dela, zenbat auto aztertu beharko genuke a) kasuan?

Em.: Izan bedi X ='*2004ko CO_2-aren emisioa*' z.a. Datuetan oinarritua, laginaren tamaina $n = 14$ auto eta ondoko estimazioak kalkula ditzakegu: $\overline{x} = 146.2857$ gr/km, $s_n = 22.0144$ gr/km eta $s_{n-1} = 22.8454$ gr/km.

a) $X : \mathcal{N}(\mu, 25)$ izateagatik, $\overline{X} : \mathcal{N}(\mu, \frac{25}{\sqrt{14}})$. Bestalde, $1 - \alpha = 0.95 \Rightarrow z_{0.025} = 1.96$; beraz, $\varepsilon = 1.96\frac{25}{\sqrt{14}} = 13.0958$ eta %95eko μ-ren konfiantza-tartea ondokoa da:

$$I_{\mu}^{0.95} = (146.2857 \mp 13.0958) = (133.1899,\ 159.3815).$$

b) $X : \mathcal{N}(\mu, \sigma)$ da, σ ezezaguna eta n txikia izanik, beraz, $\varepsilon = t_{0.025;13} \cdot \frac{s_{n-1}}{\sqrt{14}} = 2.160 \cdot \frac{22.8454}{\sqrt{14}} = 13.1883$ eta %95eko μ-ren konfiantza-tartea ondokoa da:

$$I_{\mu}^{0.95} = (146.2857 \mp 13.1883) = (133.0974,\ 159.4740).$$

c) Azkenik, komenta dezagun lehenengo galdera: datuetan oinarritua ezin da ondorioztatu $\mu \leq 150$ denik %95eko konfiantza-mailaz, hau da, ezin da esan garbiak direnik, batez beste. Izan ere, emisioen batezbestekoa 133 gr/km eta 159 gr/km tartean egongo da. Aldakortasuna handitzat har daiteke, lagin tamaina handiagoa kontsideratzekotan ondorioak aldatu ahal izango lirateke.

d) $n = (1.96 \cdot \frac{25}{3})^2 = 266.7778$, hots, 267 auto aztertu beharko genuke populazio-batezbestekoa estimatzeko errorea gehienez 3 gr/km-koa izateko.

Kontuan hartu atal guztietan X-ren normaltasunaren hipotesia egiaztatu beharko genukeela, Estatistikako hipotesi-kontrasteen probak erabiliz. □

6.3.3. KT populazio-bariantzarako.

6.6. teorema. *Izan bedi σ^2 bariantzako populazio normaletik ateratako $X_1, X_2, \ldots, X_n$ zorizko lagin bakuna. Orduan, lagin-kuasibariantza S_{n-1}^2 zorizko aldagairen banaketa ondokoa da:*

$$\frac{(n-1)S_{n-1}^2}{\sigma^2} : \chi_{n-1}^2$$

Definizioa: Populazio normal batetik ateratako n tamainako laginaren kuasibariantza s_{n-1}^2 izanik, $\%100 \cdot (1-\alpha)$ *mailako σ^2-ren konfiantza-tartea*, ondoko tartearen bidez definitzen da:

$$I_{\sigma^2}^{1-\alpha} = \left(\frac{(n-1)s_{n-1}^2}{\chi_{\alpha/2;n-1}^2},\ \frac{(n-1)s_{n-1}^2}{\chi_{1-\alpha/2;n-1}^2} \right)$$

Hain zuzen, χ^2_n banaketak $P\left(\chi^2_{1-\alpha/2;n} < \chi^2_n < \chi^2_{\alpha/2;n}\right) = 1-\alpha$ betetzen du.

6.22. adibidea. Demagun normaltasuna.

a) Kalkula ezazu 2004ko emisioen bariantzarako %90eko konfiantza-tarteak.

b) Kalkula ezazu 2004ko emisioen desbideratze-estandarrerako %90eko konfiantza-tarteak.

c) Aurreko ariketan, a) atalean egindako suposizioa ontzat har daiteke?

d) Erantzun ezazu 2. galdera: Nolakoa zen 2004ko autoen emisioen sakabanapena?

Em.:

a) Populazio-bariantzaren puntu-estimazioa $\hat{\sigma^2} = s^2_{n-1} = 521.9123$ (gr/km)^2 da. $1-\alpha = 0.90 \Rightarrow \chi^2_{0.05;13} = 22.362$ eta $\chi^2_{0.95;13} = 5.892$ eta %90eko σ^2-ren konfiantza-tartea ondokoa da:

$$I^{0.90}_{\sigma^2} = (\frac{13 \cdot s^2_{n-1}}{22.362},\ \frac{13 \cdot s^2_{n-1}}{5.892}) = (303.4102,\ 1151.5377).$$

b) Populazioaren desbideratze estandarraren puntu-estimazioa $\hat{\sigma} = s_{n-1} = 22.8454$ gr/km da eta %90eko σ-ren konfiantza-tartea

$$I^{0.90}_{\sigma} = (\sqrt{303.4102},\ \sqrt{1151.5377}) = (17.4187,\ 33.9343)$$

c) 25 gr/km-ko desbideratze estandarra ontzat har daiteke, $25 \in I^{0.90}_{\sigma}$ baitago.

d) 2. galdera erantzunda dago a) eta b) azpiataletan %90eko konfiantza mailaz.

X-ren normaltasunaren hipotesia ere beharrezkoa da ondorio guztiak ontzat hartzeko. □

6.3.4. KT populazio-proportziorako.

6.7. teorema. (Moivreren teoremaren aplikazioa)

Izan bedi $X : Bin(n, p)$. *Baldin* $n\hat{p} > 5$ *eta* $n\hat{q} > 5$, *orduan, lagin-proportzioaren aldagaiaren banaketa ondokoa da:*

$$\hat{P} \approx \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$$

Definizioa: Populazio binomial batetik ateratako n tamainako laginaren arrakasta-proportzioa $\hat{p}$ izanik, non $\hat{q} = 1 - \hat{p}$, $n\hat{p} > 5$ eta $n\hat{q} > 5$ diren, $\%100 \cdot (1 - \alpha)$ *mailako p-ren konfiantza-tartea* hauxe da:

$$I_p^{1-\alpha} = (\hat{p} - \varepsilon,\ \hat{p} + \varepsilon)\,,\ \varepsilon = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Lagin-tamainaren kalkulua
Demagun p proportzioa estimatu nahi dela, $\%100 \cdot (1 - \alpha)$ mailako konfiantza-tartearen lagin-errorea ε izatera baldintzaturik. Zein izan beharko da laginaren tamaina, baldintza hau bete dadin?

$$\varepsilon = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \Rightarrow n = z_{\alpha/2}^2 \frac{\hat{p}\hat{q}}{\varepsilon^2}$$

6.23. adibidea.

a) Kalkula ezazu 2004ko diesel motordun auto garbien proportziorako %99ko konfiantza-tartea.

b) Kalkula ezazu 2004ko gasolinazko auto garbien proportziorako %99ko konfiantza-tartea.

c) Erantzun ezazu 3. galdera: 2004ko diesel motordun autoen %92a eta gasolinazko %40a garbiak ziren?

d) Zein izan beharko ziren lagin tamainak lagin-erroreak %5 baino handiagoak ez izateko?

Em.:

a) Izan bedi X_D ='*2004 urteko diesel motordun auto garbien kopurua*'. Zorizko aldagaia binomiala dela argi dago, non populazio proportzioa p_D, diesel motordun auto garbien proportzioa den. Datuetan oinarritua ondoko estimazioak ditugu: $\hat{p}_D = \frac{6}{7} = 0.8571$ eta $\hat{q}_D = 1 - \hat{p}_D = 0.1429$. Konfiantza maila %99 eskatzen denez,

$1-\alpha = 0.99 \Rightarrow \alpha = 0.01 \Rightarrow \alpha/2 = 0.005 \Rightarrow z_{0.005} = 2.58$. Hortaz, $\varepsilon = 2.58 \cdot \sqrt{\frac{\hat{p}_D \hat{q}_D}{7}} = 0.3413$ eta %99ko p_D-ren konfiantza-tartea ondokoa da:

$$I_{p_D}^{0.99} = (0.8571 \mp 0.3413) = (0.5158,\ 1).$$

Beraz, diesel motordun auto garbien portzentajea %51 baino handiagoa da.

b) Izan bedi X_G ='*2004 urteko gasolinazko auto garbien kopurua*'. Zorizko aldagaia binomiala dela argi dago, non populazio proportzioa p_G, gasolinazko auto garbien proportzioa den. Datuetan oinarritua ondoko estimazioak ditugu: $\hat{p}_G = \frac{3}{7} = 0.4286$ eta $\hat{q}_G = 1 - \hat{p}_G = 0.5714$. $\varepsilon = 2.58 \cdot \sqrt{\frac{\hat{p}_G \hat{q}_G}{7}} = 0.4286$ eta %99ko p_G-ren konfiantza-tartea ondokoa da:

$$I_{p_G}^{0.99} = (0.4286 \mp 0.4286) = (0,\ 0.8572).$$

Beraz, gasolinazko auto garbien portzentajea %86 baino txikiagoa da.

c) Ikusten dugunez, 3. galdera egiazkoa izan arren, $0.92 \in I_{p_D}^{0.99}$ eta $0.40 \in I_{p_G}^{0.99}$ baitira, konfiantza-tarteak luzeegiak dira oso bestelako gauzak ondorioztatzeko. Izan ere, lagin-erroreak %34 eta %43 dira, hurrenez hurren. Gainera, $n_D \hat{p}_D = 6 > 5$ baina $n_D \hat{q}_D = 1 \not> 5$ eta $n_G \hat{p}_G = 3 \not> 5$, $n_G \hat{q}_G = 4 \not> 5$, $n_D \not> 30$ eta $n_G \not> 30$ izateagatik propotziorako konfiantza-tarteak ezin dira erabili, lagintamainak txikiegiak baitira.

d) $n_D = 2.58^2 \cdot \frac{0.8571 \cdot 0.1429}{0.05^2} = 326.1093$ eta $n_G = 2.58^2 \cdot \frac{0.4286 \cdot 0.5714}{0.05^2} = 652.0664$. Beraz, doitasun handiagoa (lagin-errorea 0.05 baino txikiagoa) lortzeko 327 diesel motordun auto eta 653 gasolinazko auto aztertu beharko genituzke (tamaina hauekin, espero dugu auto garbi eta ez garbi kopurua 5 baino handiagoa izatea eta horrela, konfiantza-tarteak erabili ahal izatea). □

6.3.5. KT bi populazio-batezbestekoen diferentziarako.

Ikusi dugunez, populazio normala denean edo $n \geq 30$, lagin-batezbestekoaren banaketa $\overline{X} \approx \mathcal{N}(\mu, \sigma/\sqrt{n})$ da. Demagun orain, bi populazio aske ditugula, X_1 eta X_2, non batezbestekoak μ_1, μ_2 eta bariantzak σ_1^2, σ_2^2 diren, hurrenez hurren. Populazio bakoitzetik n_1 eta n_2 tamainako

laginak ateratzen baditugu, hurrenez hurren, $\overline{X}_1 \approx \mathcal{N}(\mu_1, \sigma_1/\sqrt{n_1})$ eta $\overline{X}_2 \approx \mathcal{N}(\mu_2, \sigma_2/\sqrt{n_2})$. Azpiatal honetan, bi populazioren batezbestekoen arteko diferentzia, $\mu_1 - \mu_2$, estimatuko dugu.

6.8. teorema. *Izan bedi μ_1 batezbestekoa eta σ_1^2 bariantza ezaguna dituen populaziotik ateratako $X_{11}, X_{12}, X_{13}, \dots X_{1n_1}$ zorizko lagin bakuna eta μ_2 batezbestekoa eta σ_2^2 bariantza ezaguna dituen populaziotik ateratako $X_{21}, X_{22}, X_{23}, \dots X_{2n_2}$ zorizko lagin bakuna, non 1. eta 2. populazioak askeak diren, $n_1 \geq 30$ eta $n_2 \geq 30$ izanik. Orduan, lagin-batezbestekoen arteko diferentzia $\overline{X}_1 - \overline{X}_2$ zorizko aldagaiaren banaketa ondokoa da:*

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx \mathcal{N}(0,1)$$

Bereziki, bi populazioren banaketak normalak badira, bi laginaren tamainak n_1, n_2 edozein izanik, banaketa asintotikoki (gutxigorabehera) normala izan beharrean, zehatz mehatz normala izango da.

Definizioa: Bariantza ezagunetako bi populazio askeren batezbestekoen arteko diferentziarako, $\%100 \cdot (1-\alpha)$ *mailako* $(\mu_1 - \mu_2)$*-ren konfiantza-tartea* hauxe da:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left(\overline{x}_1 - \overline{x}_2 - \varepsilon,\ \overline{x}_1 - \overline{x}_2 + \varepsilon\right),\ \varepsilon = z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Laginaren tamainaren kalkulua
Demagun σ_1^2 eta σ_2^2 ezagunak dituen populazioen batezbestekoen diferentzia estimatu nahi dela. $\%(1-\alpha) \cdot 100$-eko $(\mu_1 - \mu_2)$-ren konfiantza-tartearen lagin-errorea ε eta $n_1 = n_2 \equiv n$ izatera baldintzatuz. Zein da, baldintza hori bete dezan, beharrezkoa den laginaren tamainak?

$$\varepsilon = z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}} \Rightarrow n = z^2_{\alpha/2} \cdot \frac{\sigma_1^2 + \sigma_2^2}{\varepsilon^2}.$$

6.9. teorema. *Izan bedi μ_1 batezbestekoa eta σ_1^2 bariantza ezezaguna dituen populaziotik ateratako $X_{11}, X_{12}, X_{13}, \dots X_{1n_1}$ zorizko lagin bakuna eta μ_2 batezbestekoa eta σ_2^2 bariantza ezezaguna dituen populaziotik ateratako $X_{21}, X_{22}, X_{23}, \dots X_{2n_2}$ zorizko lagin bakuna, non 1. eta 2. populazioak normalak askeak diren eta $n_1 \geq 30, n_2 \geq 30$ izanik. Orduan, baldin $\hat{\sigma}_1^2 = S_{n-1,1}^2$ eta $\hat{\sigma}_2^2 = S_{n-1,2}^2$, lagin-batezbestekoen arteko diferentzia $\overline{X}_1 - \overline{X}_2$ zorizko aldagaiaren banaketa ondokoa da:*

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_{n-1,1}^2}{n_1} + \frac{S_{n-1,2}^2}{n_2}}} \approx \mathcal{N}(0,1)$$

Definizioa: Bariantza ezezagunetako bi populazio askeren batezbestekoen arteko diferentziarako, $\%100 \cdot (1-\alpha)$ *mailako* $(\mu_1 - \mu_2)$*-ren konfiantza-tartea* hauxe da:

$$I_{\mu_1-\mu_2}^{1-\alpha} = \left(\overline{x}_1 - \overline{x}_2 - \varepsilon,\ \overline{x}_1 - \overline{x}_2 + \varepsilon\right),\ \varepsilon = z_{\alpha/2} \cdot \sqrt{\frac{s_{n-1,1}^2}{n_1} + \frac{s_{n-1,2}^2}{n_2}}$$

6.10. teorema. *Izan bedi μ_1 batezbestekoa eta σ_1^2 bariantza (ezezaguna) dituen populaziotik ateratako $X_{11}, X_{12}, X_{13}, \dots X_{1n_1}$ zorizko lagin bakuna eta μ_2 batezbestekoa eta σ_2^2 bariantza (ezezaguna) dituen populaziotik ateratako $X_{21}, X_{22}, X_{23}, \dots X_{2n_2}$ zorizko lagin bakuna, non 1. eta 2. populazioak normalak askeak eta $n_1 < 30$ edo $n_2 < 30$, $\sigma_1 = \sigma_2$ diren. Orduan, baldin $\hat{\sigma}_1^2 = S_{n-1,1}^2$ eta $\hat{\sigma}_2^2 = S_{n-1,2}^2$, lagin-batezbestekoen arteko diferentzia $\overline{X}_1 - \overline{X}_2$ zorizko aldagaiaren banaketa ondokoa da:*

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_{n-1,1}^2 + (n_2-1)S_{n-1,2}^2}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} : t_{n_1+n_2-2}$$

Definizioa: Bariantza ezezagun baina berdineko bi populazio normal eta askeren batezbestekoen arteko diferentziarako, $\%100 \cdot (1-\alpha)$ *mailako* $(\mu_1 - \mu_2)$*-ren konfiantza-tartea* hauxe da:

$$I_{\mu_1-\mu_2}^{1-\alpha} = \left(\overline{x}_1 - \overline{x}_2 - \varepsilon,\ \overline{x}_1 - \overline{x}_2 + \varepsilon\right),\ \varepsilon = t_{\alpha/2;n_1+n_2-2} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{non } s_p = \sqrt{\frac{(n_1-1)s^2_{n-1,1} + (n_2-1)s^2_{n-1,2}}{n_1+n_2-2}} \text{ den}$$

6.11. teorema. *Izan bedi μ_1 batezbestekoa eta σ_1^2 bariantza (ezezaguna) dituen populaziotik ateratako $X_{11}, X_{12}, X_{13}, \dots X_{1n_1}$ zorizko lagin bakuna eta μ_2 batezbestekoa eta σ_2^2 bariantza (ezezaguna) dituen populaziotik ateratako $X_{21}, X_{22}, X_{23}, \dots X_{2n_2}$ zorizko lagin bakuna, non 1. eta 2. populazioak normalak eta askeak diren eta $n_1 < 30$ edo $n_2 < 30$, $\sigma_1 \neq \sigma_2$ izanik. Orduan, $\hat{\sigma}_1^2 = S^2_{n-1,1}$ eta $\hat{\sigma}_2^2 = S^2_{n-1,2}$ badira, lagin-batezbestekoen arteko diferentzia $\overline{X}_1 - \overline{X}_2$ zorizko aldagaiaren banaketa ondokoa da:*

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S^2_{n-1,1}}{n_1} + \frac{S^2_{n-1,2}}{n_2}}} \approx t_g, \ \textit{non } g = \frac{\left(\frac{s^2_{n-1,1}}{n_1} + \frac{s^2_{n-1,2}}{n_2}\right)^2}{\frac{\left(\frac{s^2_{n-1,1}}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s^2_{n-1,2}}{n_2}\right)^2}{n_2-1}} \textit{ den.}$$

g zenbaki osoa ez denean, zenbaki osorik hurbilena erabiliko da.
Definizioa: Bariantza ezezagun eta desberdineko bi populazio normal eta askeren batezbestekoen arteko diferentziarako, $\%100\cdot(1-\alpha)$ *mailako* $(\mu_1 - \mu_2)$*-ren konfiantza-tartea* hauxe da:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left(\overline{x}_1 - \overline{x}_2 - \varepsilon,\ \overline{x}_1 - \overline{x}_2 + \varepsilon\right),\ \varepsilon = t_{\alpha/2;g} \cdot \sqrt{\frac{s^2_{n-1,1}}{n_1} + \frac{s^2_{n-1,2}}{n_2}}$$

Ondorioak $\%100\cdot(1-\alpha)$ konfiantza mailaz: Baldin $0 \in I^{1-\alpha}_{\mu_1-\mu_2}$ bada, orduan $\mu_1-\mu_2 = 0$, hau da, $\mu_1 = \mu_2$, batezbestekoen berdintasuna ezin da errefusatu, hots, desberdintasun adierazgarririk ez dagoela (ezin da onartu bata bestea baino handiagoa edo txikiagoa denik). Hala ere, baldin $I^{1-\alpha}_{\mu_1-\mu_2} \subset (0,\ +\infty)$ betetzen bada, orduan $\mu_1 - \mu_2 > 0$, hau da, $\mu_1 > \mu_2$ eta baldin $I^{1-\alpha}_{\mu_1-\mu_2} \subset (-\infty,\ 0)$ bada, orduan $\mu_1 - \mu_2 < 0$, hau da, $\mu_1 < \mu_2$.

6.24. adibidea. Demagun normaltasuna.

a) Kalkula ezazu 2004ko I eta II auto moten artean batezbesteko emisioen diferentziarako %95eko KT, populazio-desbideratze estandarrak 20 gr/km-koak badira.

b) Kalkula ezazu 2004ko I eta II auto moten artean batezbesteko emisioen diferentziarako %95eko KT, populazioaren berdineko bariantzak ezezagunak badira.

c) Kalkula ezazu 2004ko I eta II auto moten artean batezbesteko emisioen diferentziarako %95eko KT, populazioaren desberdineko bariantzak ezezagunak badira.

d) Erantzun ezazu 4. galdera: Lur orotako ibilgailu eta monobolumenen batezbesteko emisioen eta gainontzeko autoen arteko diferentzia 30 gramokoa zen 2004 urtean?

e) Demagun gehienez 5 gr/km-ko lagin-errorea nahi dela, zenbat auto aztertu beharko genuke a) kasuan?

Em. Izan bitez X_1 ='*2004ko lur orotako ibilgailu eta monobolumenen CO_2-aren emisioa*' eta X_2 ='*2004ko gainontzeko ibilgailuen CO_2-aren emisioa*' z.a. Datuetan oinarritua, laginaren tamainak $n_1 = 6$ I motako auto eta $n_2 = 8$ II motako auto eta ondoko estimazioak kalkula ditzakegu: $\overline{x}_1 = 161.1667$ gr/km, $\overline{x}_2 = 135.1250$ gr/km, $s_{n-1,1} = 20.3019$ gr/km eta $s_{n-1,2} = 18.5121$ gr/km. Estimazio puntuala batezbestekoen diferentziarako: $\hat{\mu}_1 - \hat{\mu}_2 = \overline{x}_2 - \overline{x}_2 = 26.0417$ gr/km da.

a) $X_1 : \mathcal{N}(\mu_1, 20)$ eta $X_2 : \mathcal{N}(\mu_2, 20)$ izateagatik, $\overline{X}_1 - \overline{X}_2 : \mathcal{N}(\mu_1 - \mu_2,\ \sqrt{\frac{20^2}{6} + \frac{20^2}{8}})$. Bestalde, $\varepsilon = 1.96 \cdot \sqrt{\frac{20^2}{6} + \frac{20^2}{8}} = 17.7680$ eta %95eko $\mu_1 - \mu_2$-ren konfiantza-tartea ondokoa da:

$$I^{0.95}_{\mu_1-\mu_2} = (26.0417 \mp 17.7680) = (8.2737,\ 43.8097).$$

b) $X_1 : \mathcal{N}(\mu_1, \sigma_1)$ eta $X_2 : \mathcal{N}(\mu_2, \sigma_2)$ dira, $\sigma_1 = \sigma_2$ ezezagunak eta n_1 eta n_2 txikiak izanik; beraz, $s_p^2 = \frac{5 \cdot s^2_{n-1,1} + 7 s^2_{n-1,2}}{12} = 371.6434$ denez, $\varepsilon = t_{0.025;12} \cdot s_p \cdot \sqrt{\frac{1}{6} + \frac{1}{8}} = 1.782 \cdot 19.2781 \cdot 0.5401 = 18.5544$ eta %95eko $\mu_1 - \mu_2$-ren konfiantza-tartea ondokoa da:

$$I^{0.95}_{\mu_1-\mu_2} = (26.0417 \mp 18.5544) = (7.4873,\ 44.5961).$$

c) $X_1 : \mathcal{N}(\mu_1, \sigma_1)$ eta $X_2 : \mathcal{N}(\mu_2, \sigma_2)$ dira, $\sigma_1 \neq \sigma_2$ ezezagunak eta n_1 eta n_2 txikiak izanik; beraz, $g = \frac{12439.3323}{1205.9344} = 10.3151 \approx 10$ denez, $\varepsilon = t_{0.025;10} \cdot \sqrt{\frac{s^2_{n-1,1}}{n_1} + \frac{s^2_{n-1,2}}{n_2}} = 1.812 \cdot 10.5609 = 19.1363$ eta %95eko $\mu_1 - \mu_2$-ren konfiantza-tartea ondokoa da:

$$I^{0.95}_{\mu_1-\mu_2} = (26.0417 \mp 19.1363) = (6.9054,\ 45.178).$$

d) Azkenik, komenta dezagun 4. galdera: datuetan oinarritua ezin da errefusatu $\mu_1 - \mu_2 = 30$ denik %95eko konfiantza-mailaz, hau da, ezin da deuseztatu bi auto moten artean emisioen batezbesteko diferentzia 30 gr/km-koa denik, izan ere, $30 \in I^{0.95}_{\mu_1-\mu_2}$ baita. Hala ere, ezin da ondorioztatu diferentzia 30 gr/km baino handiagoa edo txikiagoa denik, diferentziarako tartea 30 balioaren eskuinaldean edo ezkerraldean ez baitago.

e) $n_1 = n_2 = 1.96^2 \cdot \frac{20^2+20^2}{5^2} = 62.72$, hots, 63 I motako auto eta 63 II motako auto.

Kontuan hartu, berriro X_1 eta X_2 zorizko aldagaien normaltasunaren hipotesiak egiaztatu beharko genituzkeela. □

Sarritan aurkitzen dugun arazoa laginak askeak ez izatea, baizik eta lagin bateko behaketa bakoitza beste lagineko behaketa bakar batekin lotuta egotea. Adibidez, problema askotan, bi faktoreren eraginak konparatzeko, komenigarria da bi faktoreak elementu bakoitzean aplikatzea, beren eraginen arteko diferentzia behaturik. Esate baterako, bi denboraldi konparatzeko, bi ongarri lurralde berberean, tratamendu barik eta tratamendu batekin eragindako portaera, eta abar.
Diferentzia adierazten duen populazio berria defini dezakegu, $D = X - Y$, eta populazio honetatik ateratako diferentzien zorizko lagina, balioak $d_i = x_i - y_i,\ i = 1, 2, \ldots, n$ izanik. D populazio banaketa normalari darraio baldin X eta Y normalak badira, eta halaber $\mu_D = \mu_X - \mu_Y$ diferentziarako batezbestekoa kalkulatu. Bestalde, diferentzien populazioaren bariantza ezezaguna izango da, nahiz eta X eta Y populazioen bariantzak ezagutu. Beraz, bariantza ezezaguneko populazio normalaren batezbestekorako inferentzia dugu.
Definizioa: Demagun X eta Y populazio normalak direla eta n tamainako binakako lagina ateratzen dela. Populazio normalaren binakako batezbestekoen arteko diferentziarako, $\%100 \cdot (1-\alpha)$ *mailako* $\mu_D = \mu_X - \mu_Y$*-ren konfiantza-tartea* hauxe da:

$$I^{1-\alpha}_{\mu_D} = \left(\overline{d} - \varepsilon,\ \overline{d} + \varepsilon\right),\ \varepsilon = t_{\alpha/2;n-1} \cdot \frac{s_{n-1,D}}{\sqrt{n}}$$

6.25. adibidea. Demagun normaltasuna.

a) Kalkula ezazu 2004 eta 2008 urteen artean batezbesteko emisioen diferentziarako %90eko konfiantza-tartea.

b) Erantzun ezazu 5. galdera: 2008 urtean CO_2-aren isuriak 10 gramo jaitsi dira 2004 urtearekiko?

Em.

a) Izan bitez X ='*2004ko CO_2-aren emisioa*' eta Y ='*2008ko CO_2-aren emisioa*' z.a. Datuetan oinarritua, laginaren tamaina $n = 14$ eta auto berberetan emisioak neurtu direnez, binakako datuak dira. Horregatik, $D = X - Y$ z.a. definituko dugu eta ondoko estimazioak kalkula ditzakegu: $\overline{d} = \frac{76}{14} = 5.4286$ gr/km, $s^2_{n,D} = \frac{922}{14} - \overline{d}^2 = 36.3878$ gr/km, hurrengo taulan oinarritua:

d_i	2	6	7	7	7	2	-5	2	-6	7	15	16	6	10	**76**
d_i^2	4	36	49	49	49	4	25	4	36	49	225	256	36	100	**922**

$X : \mathcal{N}(\mu_X, \sigma_X)$ eta $Y : \mathcal{N}(\mu_Y, \sigma_Y)$ izateagatik, $D = X - Y : \mathcal{N}(\mu_D,\ \sigma_D)$, non $\mu_D = \mu_X - \mu_Y$ eta σ_D ezezaguna den. Beraz, $\varepsilon = t_{0.05;13} \cdot \frac{s_{n-1,D}}{\sqrt{n}} = 1.771 \cdot \frac{s_{n,D}}{\sqrt{n-1}} = 2.9629$ eta %90eko μ_D-ren konfiantza-tartea ondokoa da:

$$I^{0.90}_{\mu_D} = I^{0.90}_{\mu_X - \mu_Y} = (5.4286 \mp 2.9629) = (2.4657,\ 8.3915).$$

b) Azkenik, komenta dezagun 5. galdera: datuetan oinarritua ezin da onartu $\mu_D = \mu_X - \mu_Y = 10$ denik %90eko konfiantza-mailaz, hau da, ezin da onartu jaitsiera 10 gr/km-koa denik, izan ere, $10 \notin I^{0.90}_{\mu_X - \mu_Y}$. Hala ere, $\mu_D > 0$ denez, hots, $\mu_X > \mu_Y$, 2008 urtean emisioa jaitsi da 2004 urtearekiko; are gehiago, tarteari begiratuz jaitsiera 2 gr/km baino handiagoa eta 9 gr/km baino txikiagoa dela ondoriozta daiteke.

Kontuan hartu, berriro X eta Y zorizko aldagaien normaltasunaren hipotesiak egiaztatu beharko genituzkeela. □

6.3.6. KT bi populazio-bariantzen zatidurarako.

6.12. teorema. *Izan bedi μ_1 batezbestekoa eta σ_1^2 bariantza (ezezaguna) dituen populaziotik ateratako X_{11}, X_{12}, $X_{13}, \dots$ X_{1n_1} zorizko lagin bakuna eta μ_2 batezbestekoa eta σ_2^2 bariantza (ezezaguna) dituen populaziotik ateratako X_{21}, X_{22}, $X_{23}, \dots$ X_{2n_2} zorizko lagin bakuna, non 1. eta 2. populazioak askeak diren. Orduan, lagin-bariantzen zatidura zorizko aldagaiaren banaketa ondokoa da:*

$$\frac{\frac{S^2_{n-1,1}}{\sigma_1^2}}{\frac{S^2_{n-1,2}}{\sigma_2^2}} : \mathcal{F}_{n_1-1,n_2-1}$$

Definizioa: σ_1^2 eta σ_2^2 bariantzak (ezezagunak) dituen populazio biren bariantzen arteko zatidurarako, $\frac{\sigma_1^2}{\sigma_2^2}$, $\%100 \cdot (1-\alpha)$ *mailako konfiantza-tartea* hauxe da:

$$I^{1-\alpha}_{\sigma_1^2/\sigma_2^2} = \left(\frac{\frac{s^2_{n-1,1}}{s^2_{n-1,2}}}{F_{\alpha/2;n_1-1,n_2-1}}, \ \frac{\frac{s^2_{n-1,1}}{s^2_{n-1,2}}}{F_{1-\alpha/2;n_1-1,n_2-1}} \right)$$

Gogora ezazu $F_{1-\alpha/2;n_1-1,n_2-1} = 1/F_{\alpha/2;n_2-1,n_1-1}$ betetzen dela.
Ondorioak $\%100 \cdot (1-\alpha)$ konfiantza mailaz: baldin $1 \in I^{1-\alpha}_{\sigma_1^2/\sigma_1^2}$ bada, orduan $\sigma_1^2/\sigma_2^2 = 1$, hau da, $\sigma_1^2 = \sigma_2^2$, bariantzen berdintasuna ezin da errefusatu. Hala ere, baldin $I^{1-\alpha}_{\sigma_1^2/\sigma_1^2} \subset (1, \ +\infty)$ betetzen bada, orduan $\sigma_1^2/\sigma_2^2 > 1$, hau da, $\sigma_1^2 > \sigma_2^2$ eta $I^{1-\alpha}_{\sigma_1^2/\sigma_1^2} \subset (0, \ 1)$ bada, orduan $\sigma_1^2 < \sigma_2^2$.

6.26. adibidea. Demagun normaltasuna.

a) Kalkula ezazu 2004ko bi motako autoen artean bariantzen zatidurarako %95eko konfiantza-tartea.

b) Azkenaurreko adibidean b) edo c) atalen artean, zein da ontzat har daitekeena?

c) Erantzun ezazu 6. galdera: Bi motako autoen isurien sakabanapena berdina zen 2004 urtean?

Em.:

a) %95eko σ_1^2/σ_2^2-ren konfiantza-tartea ondokoa da:

$$I^{0.95}_{\sigma_1^2/\sigma_2^2} = \left(\frac{\frac{s^2_{n-1,1}}{s^2_{n-1,2}}}{F_{0.025;5,7}},\ \frac{s^2_{n-1,1}}{s^2_{n-1,2}} \cdot F_{0.05;7,5} \right) =$$

$\left(\frac{1.2027}{5.29},\ 1.2027 \cdot 6.85\right) = (0.2274,\ 8.2386)$.

b) Azken aurreko adibidearen b) atala ontzar har daiteke eta ez litzateke beharrezkoa izango c) atala ebaztea. Izan ere, $1 \in I^{0.95}_{\sigma_1^2/\sigma_2^2}$ denez, %95eko konfiantza mailaz, bariantzen berdintasuna ezin da errefusatu.

c) Azkeniz, 6. galdera erantzun dugu: bi motako autoen sakabanapena berdina kontsidera daiteke 2004 urtean, izan ere, $1 \in I^{0.95}_{\sigma_1^2/\sigma_2^2}$ betetzen bada, $\sigma_1^2/\sigma_2^2 = 1$ ezin da deuseztatu, hots, $\sigma_1^2 = \sigma_2^2$.

X_1 eta X_2-ren normaltasunaren hipotesiak ere beharrezkoak dira ondorioak ontzat hartzeko. □

6.3.7. KT bi populazio-proportzioen diferentziarako.

6.13. teorema. *Izan bitez X_1 eta X_2 bi populazio binomial askeak, beraien parametroak p_1 eta p_2 direlarik, hurrenez hurren. Hauetatik n_1 eta n_2 tamainetako laginak aukeratzen dira, $n_1\hat{p}_1 > 5$, $n_1\hat{q}_1 > 5$, $n_2\hat{p}_2 > 5$ eta $n_2\hat{q}_2 > 5$ izanik. Orduan, lagin-proportzioen arteko diferentzia $\hat{P}_1 - \hat{P}_2$ zorizko aldagaiaren banaketa ondokoa da:*

$$\hat{P}_1 - \hat{P}_2 \approx \mathcal{N}\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

Definizioa: $X_1 : Bin(n_1, p_1), X_2 : Bin(n_2, p_2)$ askeak eta $n_1\hat{p}_1 > 5$, $n_1\hat{q}_1 > 5$, $n_2\hat{p}_2 > 5$ eta $n_2\hat{q}_2 > 5$ izanik, bi proportzioen arteko diferentziarako, $\%100 \cdot (1-\alpha)$ *$(p_1 - p_2)$-ren mailako konfiantzatartea* hauxe da:

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

Ondorioak $\%100 \cdot (1-\alpha)$ konfiantza mailaz: Baldin $0 \in I^{1-\alpha}_{p_1-p_2}$ bada, orduan $p_1 - p_2 = 0$, hau da, $p_1 = p_2$, proportzioen berdintasuna ezin da errefusatu (ezin da onartu bata bestea baino handiagoa edo txikiagoa denik). Hala ere, $I^{1-\alpha}_{p_1-p_2} \subset (0, +\infty)$ betetzen bada, orduan

$p_1 - p_2 > 0$, hau da, $p_1 > p_2$ eta baldin $I^{1-\alpha}_{p_1-p_2} \subset (-\infty, 0)$ bada, orduan $p_1 - p_2 < 0$, hau da, $p_1 < p_2$.

Lagin-tamainaren kalkulua

Demagun $p_1 - p_2$ proportzioen diferentzia estimatu nahi dela, $\%100 \cdot (1 - \alpha)$ mailako konfiantza-tartearen lagin-errorea ε eta $n_1 = n_2 \equiv n$ izatera baldintzaturik. Zein izan beharko da laginaren tamaina, baldintza hau bete dadin?

$$\varepsilon = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1\hat{q}_1 + \hat{p}_2\hat{q}_2}{n}} \Rightarrow n = z^2_{\alpha/2}\frac{\hat{p}_1\hat{q}_1 + \hat{p}_2\hat{q}_2}{\varepsilon^2}$$

6.27. adibidea.

a) Kalkula ezazu 2004ko diesel eta gasolinazko auto garbien proportzioen diferentziarako %90eko eta %95eko konfiantza-tarteak.

b) Erantzun ezazu 7. galdera: 2004ko diesel motordun auto garbien eta gasolinazkoen artean, diferentzia esanguratsurik dago?

c) Zeintzuk izan beharko ziren lagin tamainak lagin-errorea %5 baino handiagoa izan ez dadin?

Em.: Izan bitez X_D eta X_G lehen bezala definituak.

a) Alde batetik, baldin $1 - \alpha = 0.10$ bada, orduan $z_{0.05} = 1.645$ eta $\varepsilon = 1.645 \cdot 0.2291 = 0.3769$, beraz %90eko $p_D - p_G$-ren konfiantza-tartea ondokoa da:

$$I^{0.90}_{p_D-p_G} = (0.4285 \mp 0.3769) = (0.0516,\ 0.8054).$$

Beste aldetik, baldin $1 - \alpha = 0.05$ bada, orduan $z_{0.025} = 1.96$ eta $\varepsilon = 1.96 \cdot 0.2291 = 0.4490$, beraz %95eko $p_D - p_G$-ren konfiantza-tartea ondokoa da:

$$I^{0.95}_{p_D-p_G} = (0.4285 \mp 0.4490) = (-0.0205,\ 0.8775).$$

b) Orain, 7. galdera erantzutean konfiantza-mailaren arabera ondorioak aldatzen dira. Izan ere, %95eko konfiantza-mailaz ez dago diferentzia esanguratsurik, ezin da ondorioztatu bata bestea baino handiagoa denik ($0 \in I^{0.95}_{p_D-p_G}$ baitago), baina bai %90eko konfiantza-mailaz, diesel auto garbien proportzioa gasolinazkoa baino handiagoa dela esan daiteke ($p_D - p_G > 0$, hots, $p_D > p_G$ baita). Lehen esan genuen bezala, $n_D\hat{p}_D = 6 > 5$ baina $n_D\hat{q}_D = 1 \not> 5$ eta $n_G\hat{p}_G = 3 \not> 5$,

$n_G\hat{q}_G = 4 \not> 5$ izateagatik propotzioen diferentziarako konfiantza-tartea ezin da erabili.

c) $n_D = n_G = 1.645^2 \cdot \frac{0.8571 \cdot 0.1429 + 0.4286 \cdot 0.5714}{0.05^2} = 241.7371$ eta $n_D = n_G = 1.96^2 \cdot \frac{0.8571 \cdot 0.1429 + 0.4286 \cdot 0.5714}{0.05^2} = 288.0272$. Beraz,lagin-errorea 0.05 baino txikiagoa lortzeko, 484 auto (erdia diesel eta erdia gasolina) aztertu beharko genituzke proportzioen diferentzia estimatzeko %90ko konfiantza mailaz eta 578 auto diferentzia estimatzeko %95eko konfiantza-mailaz. □