

P5 - Bootstrap

Demagun gure populazioa adierazten duela X zorizko aldagaiak. Gure helburua banaketa honen parametro ezberdinen estimazioa egitea izango da eta estimatzaile horien banaketak eta propietateak ezagutzen saiatzea teknika ezberdinak erabiliz.

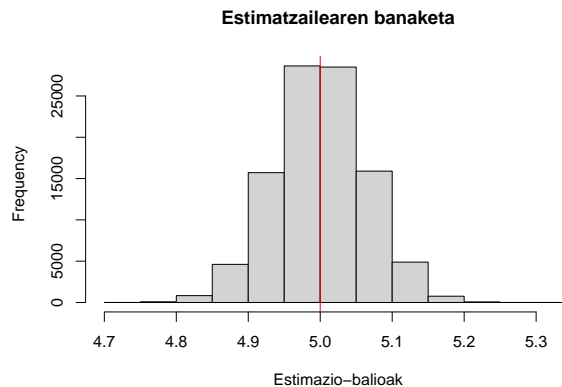
Estimatzailearen banaketa

Klasean azaldu moduan, estimatzaileak zorizko aldagaiak dira, lagin ezberdinen arabera balio ezberdinak (estimazioak) hartzen dituztenak. Hau hala izanik, estimatzailearen banaketa ezagutzeak bere propietateei buruzko informazioa emango digu. Estimatzailearen banaketa lortzeko bide ezberdin batzuk azter ditzagun:

Banaketa teorikoa

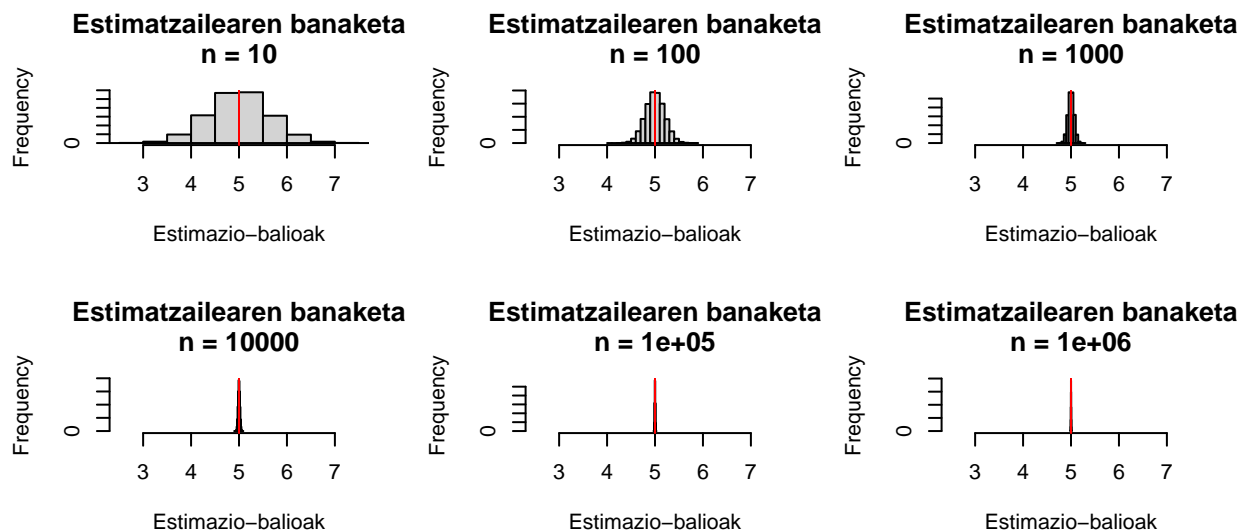
Demagun X en banaketa ezaguna dela, adibidez, $X \sim \mathcal{N}(5, 2^2)$. $EX = \mu$ parametroaren estimatzaile gisa $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ hartzen badugu, $\hat{\mu} = \bar{X} \sim \mathcal{N}(5, \frac{2}{\sqrt{n}})$ dela dakigu limite zentralaren teorematik. $n = 1000$ laginaren tamaina izanik, irudikatu dezagun estimatzailearen banaketa, laginduz eta histograma bat eginez:

```
n <- 1000
hist(rnorm(100000, 5, 2/sqrt(n)), main="Estimatzailearen banaketa", xlab="Estimazio-balioak")
abline(v=5, col="red")
```



Ariketa: Demagun $X \sim \mathcal{N}(5, 2^2)$ banaketa dugula, irudikatu $\hat{\mu} = \bar{X}$ estimatzailearen banaketa n balio ezberdinetarako (batzuk handiak eta besteak txikiak). Zer esan dezakezu?

```
par(mfrow=c(3, 3))
for (n in c(10, 100, 1000, 10000, 100000, 1000000)){
  hist(rnorm(100000, 5, 2/sqrt(n)), main=paste("Estimatzailearen banaketa\nn =", n), xlab="Estimazio-balioak")
  abline(v=5, col="red")
}
```



Soluzioa: lagin tamaina handitu ahala bariantza txikitzen da.

OHARRA: Estimatzaileraren banaketa teorikoa oso kasu gutxietan ezagutuko dugu, izan ere, X aldagaiaren banaketa ezaguna izatea ez da ohikoa, eta gainera, ezaguna izanda ere, estimatzaile guztientzat ez da tribiala banaketa analitikoki kalkulatzeko.

Lagin banaketa

Demagun X ren banaketa ezaguna dela, adibidez, $X \sim \mathcal{N}(5, 2^2)$ eta demagun banaketaren mediana (θ) estimatzeko lagin mediana erabiliko dugula: $\hat{\theta} = \text{Median}\{X_1, X_2, \dots, X_n\}$. Aurreko atalean ez bezala, kasu honetan ez da erraza estimatzaileraren benetako probabilitate banaketa zein den jakitea. Egoera honetan, X -ren banaketatik lagin ezberdin asko ateraz, $\hat{\theta}$ estimatzaileraren banaketa hurbildu dezakegu. Lehenik, banaketaren lagingak sortuko ditugu.

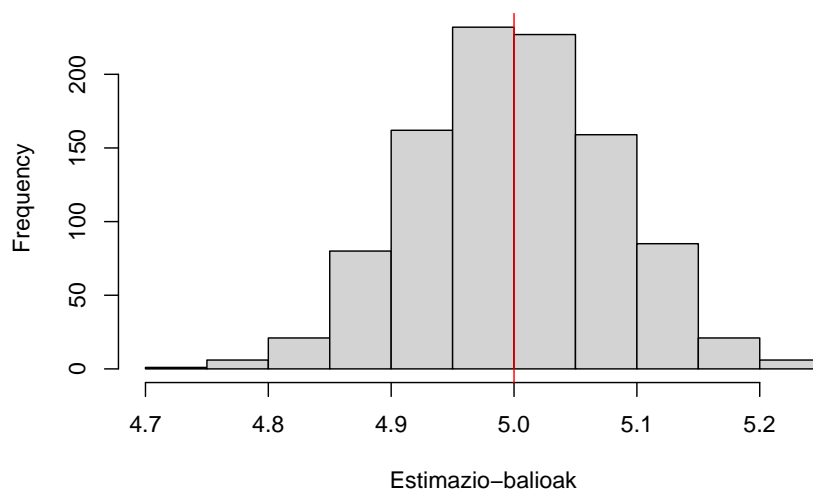
```
banaketa.laginak <- function(lagin.kopurua, lagin.tamaina, mu, sigma){
  lagingak <- matrix(rnorm(lagin.kopurua*lagin.tamaina, mu, sigma),
                    nrow=lagin.kopurua, ncol=lagin.tamaina)
  return(lagingak)
}

laginak <- banaketa.laginak(1000, 1000, 5, 2)
```

Ondoren, lagin bakoitzeko estimazioa lortuko dugu (mediana aplikatuz) eta emaitza (estimatzaileraren lagina) histograma baten bidez irudikatuko dugu.

```
estimazioak <- apply(laginak, 1, median)
hist(estimazioak, main="Estimatzaileraren banaketa", xlab="Estimazio-balioak")
abline(v=5, col="red")
```

Estimatzailearen banaketa

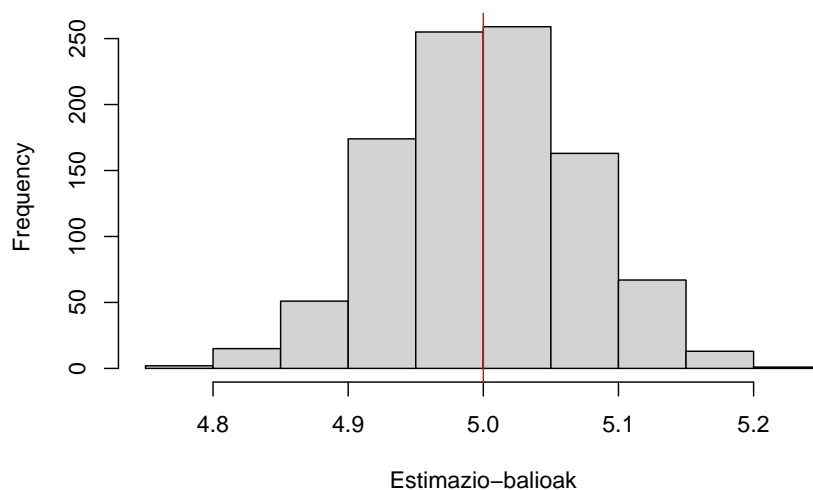


OHARRA: Egoera errealean populazioaren banaketa ez da ezaguna izaten, beraz, banaketa hau ez dugu eskura izaten eta ezin dugu lagindu.

Ariketa: Atal honetako populazio berdina hartuz, irudikatu $\hat{\theta}$ moztutako batezbestekoaren lagin-banaketa. Moztutako batezbestekoa lagineko datuen %50 zentralen batezbestekoa da eta estatistiko hau egoera batzuetan mediana eta batezbestekoa baino egokiagoa da banaketaren "zentroa" adierazteko.
OHARRA: Moztutako batezbestekoa inplementatzeko, aztertu **mean** funtzioaren **trim** argumentua.

```
estimazioak <- apply(laginak, 1, mean, trim=0.25)
hist(estimazioak, main="Estimatzailearen banaketa", xlab="Estimazio-balioak")
abline(v=5, col="red")
```

Estimatzailearen banaketa



Bootstrap banaketa

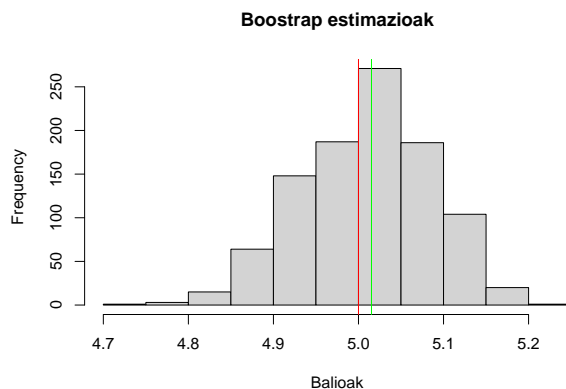
Suposatu dezagun X ren banaketa ez dugula ezagutzen eta dugun informazio bakarra banaketa honetatik ateratako $n = 1000$ tamainako lagin bat dela (lagina objetuan gordeta)¹. Mediana estimatzailearen ($\hat{\theta}$) banaketa, hurbildu nahi dugu baina kasu honetan populazioaren banaketa ez dugu ezagutzen, beraz ezin ditugu bertatik laginak atera, soilik daukagu lagin bat. Beraz, estimatzailearen banaketa lortzeko bootstrap ez parametrikoa erabiliko dugu. Gure laginetik abiatuz, itzuleradun laginketa erabiliz $B = 1000$ bootstrap lagin sortuko ditugu, baita ere $n = 1000$ tamainakoak. Kontuan izan n eta B k ez dutela zertan beti zenbaki berdinak izan.

```
bootstrapLaginakLortu <- function(hasierako.lagina, boot.lagin.kopurua, boot.lagin.tamaina){
  boot.laginak <- matrix(
    sample(hasierako.lagina, boot.lagin.kopurua*boot.lagin.tamaina, replace=TRUE),
    nrow=boot.lagin.kopurua, ncol=boot.lagin.tamaina)
  return(boot.laginak)
}

boot.laginak <- bootstrapLaginakLortu(lagina, 1000, 1000)
```

Behin aldagaiaren (bootstrap) laginketa dugula, lagin bakoitzari estimatzailea aplikatu eta emaitza histograma baten bidez irudikatuko dugu. Horrez gain bi marra bertikal marraztuko ditugu, bat benetako parametroaren balioa (gorriz) adierazteko eta bestea lagina laginaren bidez lortutako estimazioa (berdez) adierazteko.

```
boot.estimazioak <- apply(boot.laginak, 1, median)
hist(boot.estimazioak, main="Bootstrap estimazioak", xlab="Balioak")
estimazioa <- median(lagina)
abline(v=5, col="red") # Hau egoera errealetan ez dugu ezagutuko!
abline(v=estimazioa, col="green")
```



OHARRA: Azken egoera hau da errealistena, populazioaren eta estimatzailearen banaketak ez dira ezagunak. Hala ere, bootstrap-banaketa erabili dezakegu lagin-banaketa hurbiltzeko eta gure estimatzailearen hainbat propietate ezagutzeko edo konfiantza-tarteak eraikitzeko. Ikusten denez, bootstrap metodoak ez du balio estimazioa hobetuko, izan ere, lagin-banaketa benetako parametroan dago zentratuta eta, aldiz, bootstrap-banaketa laginaren estimazioan.

Ariketa: Errepikatu aurreko kode blokeetan egindakoa baina kasu honetan $\hat{\theta}$ moztutako batezbestekoaren bootstrap banaketa irudikatuz. Gainera, egin proba ezberdinak hasierako laginaren tamaina eta bootstrap laginen kopurua aldatuz.

¹Tutorialeko ariketak egiteko, lagina artifizialki sortuko dugu `lagina <- rnorm(1000, 5, 2)` eginez

Alborapena eta errore estandarra

Estimatzaillearen banaketatik informazio ezberdina atera dezakegu. Tinkotasunari buruzko informazioa lortzeko adibidez alborapena eta SE hurbildu ditzakegu.

Ariketa: Bete itzazu hurrengo bi funtzioetan falta diren kode-lerroak.

Implementatuko dugun lehenengo funtzioak estimatzaillearen alborapena hurbiltzen du. Alborapena $b(\hat{\theta}) = E(\hat{\theta}) - \theta$ da. Hau hurbiltzeko bootstrap metodoa eta plug-in printzipioa aplikatuko ditugu, itxaropena bootstrap estimazioen batezbestekoarengatik ordezkatzuz ($E(\hat{\theta}) = \hat{\theta}^*$) eta benetako parametroaren balioa estimazioa erabiliz ordezkatzuz ($\theta \sim \hat{\theta}$).

```
estimatuAlborapena <- function(boot.estimazioak, estimazioa){  
  #Kalkulatu bootstrap estimazioak erabiliz alborapena (44. gardenkia)  
  return()  
}
```

Orain hurbildu dezagun estimatzaillearen errore estandarra:

```
estimatuSE <- function(boot.estimazioak){  
  #Kalkulatu bootstrap estimazioak erabiliz SE (45. gardenkia)  
  return()  
}
```

Ariketa: Aurreko ataleko 'lagina' objektua erabiliz eta bootstrap laginketa ezberdinak erabiliz (lagin kopurua eta lagin tamainak aldatuz), hurbildu itzazu $\hat{\mu} = \bar{X}$ eta $\hat{\theta} = \text{Median}\{X_1, X_2, \dots, X_n\}$ estimatzaileen alborapena eta errore estandarra. Zer esan dezakezu?

Konfiantza-tarteak

Konfiantza tarte teorikoa

Konfiantza tarte teorikoak estimatzaillearen menpeko estatistiko jakin baten banaketa ezaguna denean erabili ditzakegu. Adibidez, hasierako populazioa normala (edo nahikoa handia) denean eta batezbestekoaren estimatzaile moduan $\hat{\mu} = \bar{X}$ hartuta, errore estandarrean oinarritutako konfiantza tarteak eraiki ditzakegu μ parametroarentzat era honetan:

$$\frac{\bar{X} - \mu}{\frac{S_{n-1}}{\sqrt{n}}} : t_{n-1} \Rightarrow KT = (\bar{x} - t_{\alpha/2; n-1} \frac{S_{n-1}}{\sqrt{n}}, \bar{x} + t_{\alpha/2; n-1} \frac{S_{n-1}}{\sqrt{n}})$$

Ariketa: Goiko formularen oinarrituz, bete kode-hutsuneak hurrengo funtzioan batezbestekorako ohiko t-konfiantza tarteak kalkulatzeko.

```
getCI <- function(x, alpha= 0.05)  
{
```

```

m <- mean(x)
se <- sd(x)/sqrt(length(x))
df <- #Bete t banaketaren askatasun graduekin
t <- qt(1-alpha/2, df)
d <- # Bete formula egokiarekin
ci <- c(m-d, m+d)
names(ci) <- paste("%", c((alpha/2)*100, (1-alpha/2)*100), sep="")
return(ci)
}

```

Ariketa: Erabili aurreko funtzioa 'lagina' objetutik batezbestekorako %95ko konfiantza-tartea kalkulatzeko. Kontuan izan $0.95 = 1 - \alpha$ dela.

Honelako konfiantza tartek t.test funtzioarekin ere kalkulatu ditzakegu honela:

```
t.test(lagina, conf.level=0.95)$conf.int
```

```
## [1] 4.838818 5.083120
## attr(,"conf.level")
## [1] 0.95
```

OHARRA: Ezin badugu lortu estatistikoaren benetako banaketa bootstrap konfiantza-tartek erabili ditzakegu.

Pertzentiletan oinarritutako bootstrap konfiantza tartek

Honelako konfiantza tartek eraikitzeke, estimatzailearen bootstrap banaketaren pertzentilak erabiliko ditugu zuzenean. Adibidez, %95 konfiantza mailarako, 2.5. eta 97.5. pertzentilak erabiliko ditugu:

$$KT = (\hat{\theta}_{0.025}, \hat{\theta}_{0.975})$$

Ariketa: Pertzentiletan oinarritutako bootstrap konfiantza tartek $\hat{\mu}$ parametroarentzat honako kodea erabiliz eraiki ditzakegu. Bete kode-hutsuneak 'quantile' funtzioa erabiliz eta erabili 'lagina' objetutik batezbestekorako %95ko konfiantza-tartea kalkulatzeko.

```

getCIBootQuantile <- function(x, B=1000, alpha=0.05)
{
  n <- length(x)
  boot.laginak <- #SORTU B=1000 BOOTSTRAP LAGIN
  boot.estimazioak <- #LAGIN BAKOITZERAKO LORTU ESTIMAZIOA
  ci <- #BETE boot.estimazioak BALIOEN BI PERTZENTIL EGOKIEKIN (bi posizioko bektore bat)
  names(ci) <- paste("%", c((alpha/2)*100, (1-alpha/2)*100), sep="")
  return(ci)
}

getCIBootQuantile(lagina, alpha=0.05)

```

Ariketa: Orokortu aurreko funtzioa, $\hat{\mu}$ parametroarentzat erabili nahi den estimatzailea (adibidez, 'mean', 'median',...) funtzioari argumentu moduan zehaztuz. Erabili funtzio berria mediana erabiliz.

OHARRA Aurreko atalean kalkulaturako tartea sinpleak dira, baina kasu askotan ez dituzte emaitza onak ematen. Hurrengo atalean tarte apur bat sofistikatuagoak nola eraiki ikusiko dugu.

Bootstrap t-tarteak

Azkenik, populazioaren banaketa ez bada normala edo ez badugu ezagutzen, ohiko konfiantza tartean agertzen den t-banaketa ez dugu izango, baina benetako banaketa hurbiltzen saiatu gaitzke bootstrap teknika erabiliz. Ondoren, banaketa hau erabiliz, konfiantza tarte honela eraiki dezakegu:

$$KT = \left(\hat{\theta} - t_{1-\alpha/2; n-1}^* \cdot SE(\hat{\theta}), \hat{\theta} - t_{\alpha/2; n-1}^* \cdot SE(\hat{\theta}) \right)$$

Ariketa: $\hat{\mu}$ parametrorako bootstrap t-konfiantza tartea honako kodea erabiliz eraiki ditzakegu. Kontuan izan, *batezbestekoa* estimatzailearen errore estandarrak (SE) formula itxia duela. Zein da? Zer egin dezakegu errore estandarra ezin badugu modu itxian kalkulatu? Bete kode-hutsunak:

```
getCIBootT <- function(x, alpha=0.05, B=1000, closed=TRUE)
{
  estimazioa <- estim_fun(x)
  n <- length(x)
  boot.laginak <- #SORTU B=1000 BOOTSTRAP LAGIN
  boot.estimazioak <- #LAGIN BAKOITZERAKO LORTU ESTIMAZIOA
  se <- #BETE PARAMETROAREN ERRORE ESTANDARRAREKIN
  boot.t <- approxDistributionTBoot(estimazioa, boot.laginak, boot.estimazioak)
  ci_inf <- estimazioa - quantile(boot.t, 1-alpha/2)*se
  ci_sup <- # BETE KONFIANTZA TARTEAREN GOIKO MUGAREKIN
  ci <- c(ci_inf, ci_sup)
  names(ci) <- paste("%", c((alpha/2)*100, (1-alpha/2)*100), sep="")
  return(ci)
}

approxDistributionTBoot <-function(estimation, boot.samples, boot.estimations, closed=closed){
  n <- ncol(boot.samples)
  B <- nrow(boot.samples)
  boot.se <- # DEFINITU BOOTSTRAP ESTIMATZAILEEN ERRORE ESTANDARRAK
  t.balioak <- # SORTU BOOTSTRAP t BANAKETA (Ikus 50. gardenkia)
}
```

Ariketak

1. **ariketa** Demagun honako bi laginak ditugula:

```
lagina1 <- c(11.8, 9.3, 11.4, 20.4, 15.2, 17.8, 15.2, 13.5, 15.1, 10.6, 18.4, 7.5, 17.7, 19.9,
            12.4, 14.6, 21.2, 10.1, 10.2, 13.7, 15.8, 18.1, 16.2, 19.8, 14.2, 14.1, 11.3,
            11.1, 16.8, 14.8, 15.8, 14.6, 15.0, 15.7, 10.3, 8.8, 17.0, 15.7, 18.6, 12.1, 16.4,
            19.3, 9.9, 15.9, 19.7, 17.6, 15.9, 16.2, 11.1, 15.6)

lagina2 <- c(0.10, 0.03, 0.03, 1.11, 0.50, 0.59, 0.22, 0.26, 3.19, 0.49, 1.14, 0.36, 0.38,
            1.10, 0.02, 0.18, 0.51, 0.41, 0.01, 0.51, 0.15, 0.36, 0.09, 1.50, 0.51, 0.57,
            0.66, 0.16, 1.10, 0.39, 0.01, 2.52, 1.94, 0.68, 1.06, 0.01, 0.09, 0.74, 0.07,
            1.39, 0.66, 0.92, 0.07, 0.13, 0.42, 0.02, 0.21, 0.47, 0.78, 1.71, 0.01, 0.44,
            0.31, 0.28, 0.26, 0.55, 1.44, 0.17, 0.20, 0.27, 0.07, 0.12, 0.33, 1.02, 1.54,
            0.43, 0.47, 1.99, 0.96, 0.52, 0.02, 0.65, 0.76, 0.03, 0.95, 0.36, 1.33, 0.69,
            0.15, 0.00, 0.46, 1.05, 1.23, 0.84, 0.49, 0.04, 0.04, 0.09, 0.42, 0.14, 0.20,
            0.01, 0.21, 1.88, 0.39, 0.16, 0.10, 0.07, 0.06, 0.17)
```

- Irudikatu laginen histogramak. Zer esan dezakegu banaketei buruz?
- Batezbestekoaren bootstrap banaketa irudikatu eta kalkulatu alborapena eta errore estandarra bi kasuetan.
- Bariantzaren bootstrap banaketa irudikatu eta kalkulatu alborapena eta errore estandarra bi kasuetan.
- Kalkulatu batezbestekorako:
 - Ohiko konfiantza tarteak, asumituz hasierako populazioak banaketa normala jarraitzen duela.
 - Pertzentiletan oinarritutako bootstrap konfiantza tarteak.
 - Bootstrap t-konfiantza tarteak.
- (ARIKETA EXTRA) Kalkulatu bariantzarako:
 - Pertzentiletan oinarritutako bootstrap konfiantza tarteak.
 - Bootstrap t-konfiantza tarteak (OHARRA: Kasu honetan estimatzailearen benetako SE ez da ezaguna, ikus 53. gardenkia).

2. **ariketa** $\mathcal{N}(20, 5^2)$ populazioan oinarrituz, aipaturiko hiru motako μ rako konfiantza-tarteak (KT) adierazitako konfiantza maila betetzen duten ala ez aztertu. Horretarako oinarritu konfiantza tartearen interpretazioan:

- Sor ditzagun zoriz $n = 100$ tamainako $N = 10000$ lagina gure populaziotik ($\mathcal{N}(20, 5^2)$).
- Lagin bakoitzerako kalkula dezagun μ rako %95eko konfiantza-tartea eta begira dezagun ia parametroaren benetako balioa tartean dagoen.
- Konfiantza-tartea %95ekoa bada, parametroa tartean egon beharko luke 0.95 probabilitatearekin. Konprobatu ea hau betetzen den.

3. **ariketa** Errepikatu aurreko ariketa baina hasierako populazioa banaketa esponentzial bat dela asumituz ($X \sim \text{Exp}(\lambda = 2)$). Zer ondorio atera dituzu?