Kolmogorov-Smirnov proba

Josemari Sarasola

Kolmogorov-Smirnov proba

- Kolmogorov-Smirnov proba datu-multzo bat probabilitate-banaketa batekin bat etor daitekeen erabakitzeko proba estatistiko bat da.
- Horretarako probabilitate banaketaren banaketa-funtzioa datuei dagokien banaketa funtzioarekin alderatzen du, eta bi funtzioen arteko D_{max} distantzia maximoa kalkulatu.
- Distantzia maximoa handiegia bada, α esangura-maila baterako, kasualitatez ezin izan dela gertatu ondorioztatzen dugu, eta beraz datuak emandako probabilitate-banaketarekin bat ez datozela erabaki.
- α esangura-maila kasualitateaz ez fidatzeko aurrez ezartzen dugun probabilitatea edo portzentajea da, nolabait esateko aurrez finkatzen dugun mesfidantza-maila bat.

Banaketa-funtzio enpirikoa

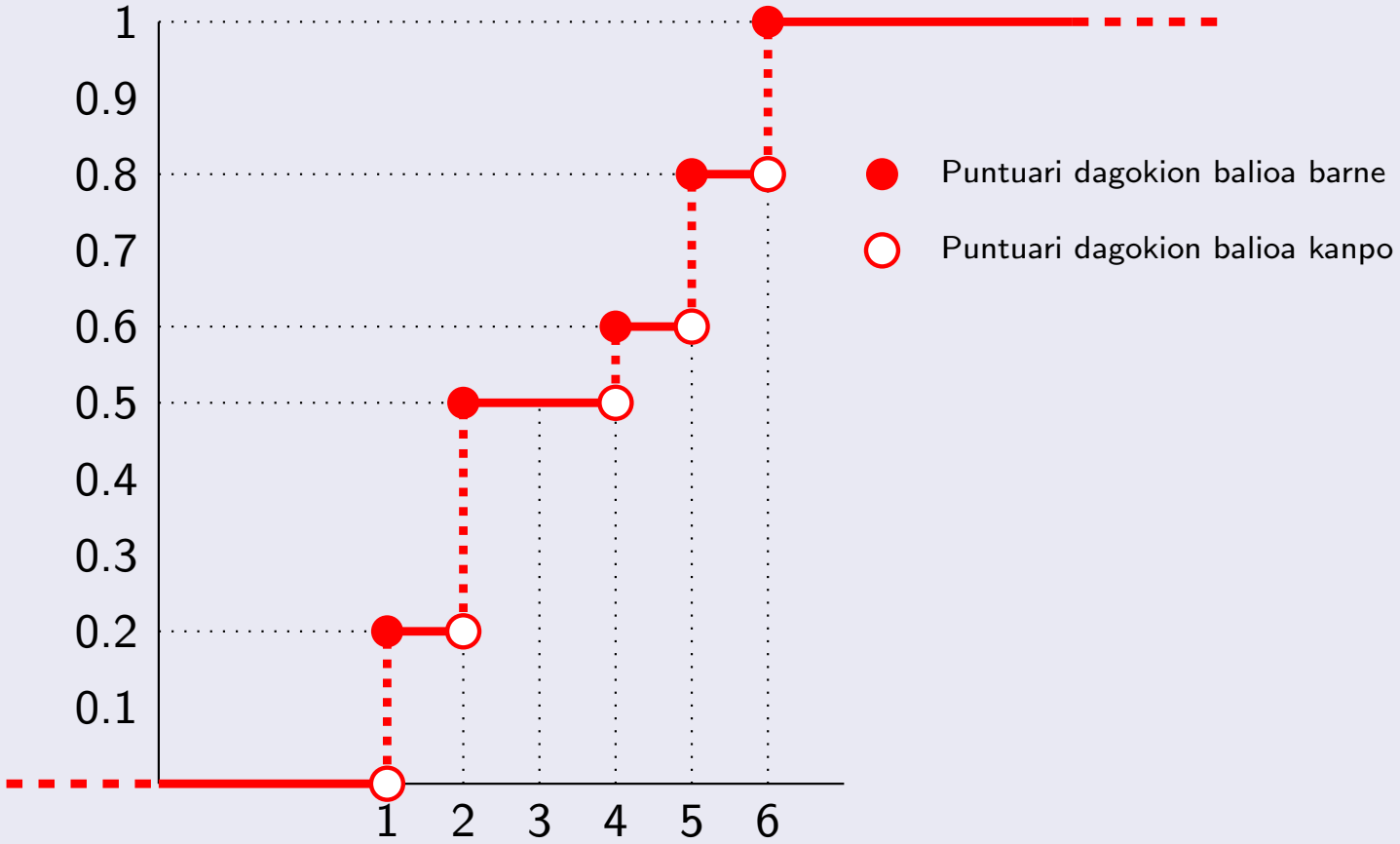
Banaketa-funtzio enpirikoa datuei dagokien banaketa-funtzioa da, datuei dagozkien maiztasun erlatibo metatuak probabilitatetzat hartuz.

Adibidez, dado bat 10 aldiz bota eta 1-1-2-2-2-4-5-5-6-6 emaitzak lortu baditugu, hauek dira F maiztasun erlatibo metatuak, banaketa-funtzioaren baliotzat hartuko ditugunak:

x	1	2	3	4	5	6
$f(x)$	0.2	0.3	0	0.1	0.2	0.2
$\hat{F}(x)$	0.2	0.5	0.5	0.6	0.8	1

Banaketa-funtzio enpirikoa

Grafikoki, banaketa-funtzio enpirikoak eskailera moduan irudikatzen dira beti:



Banaketa-funtzio teorikoa

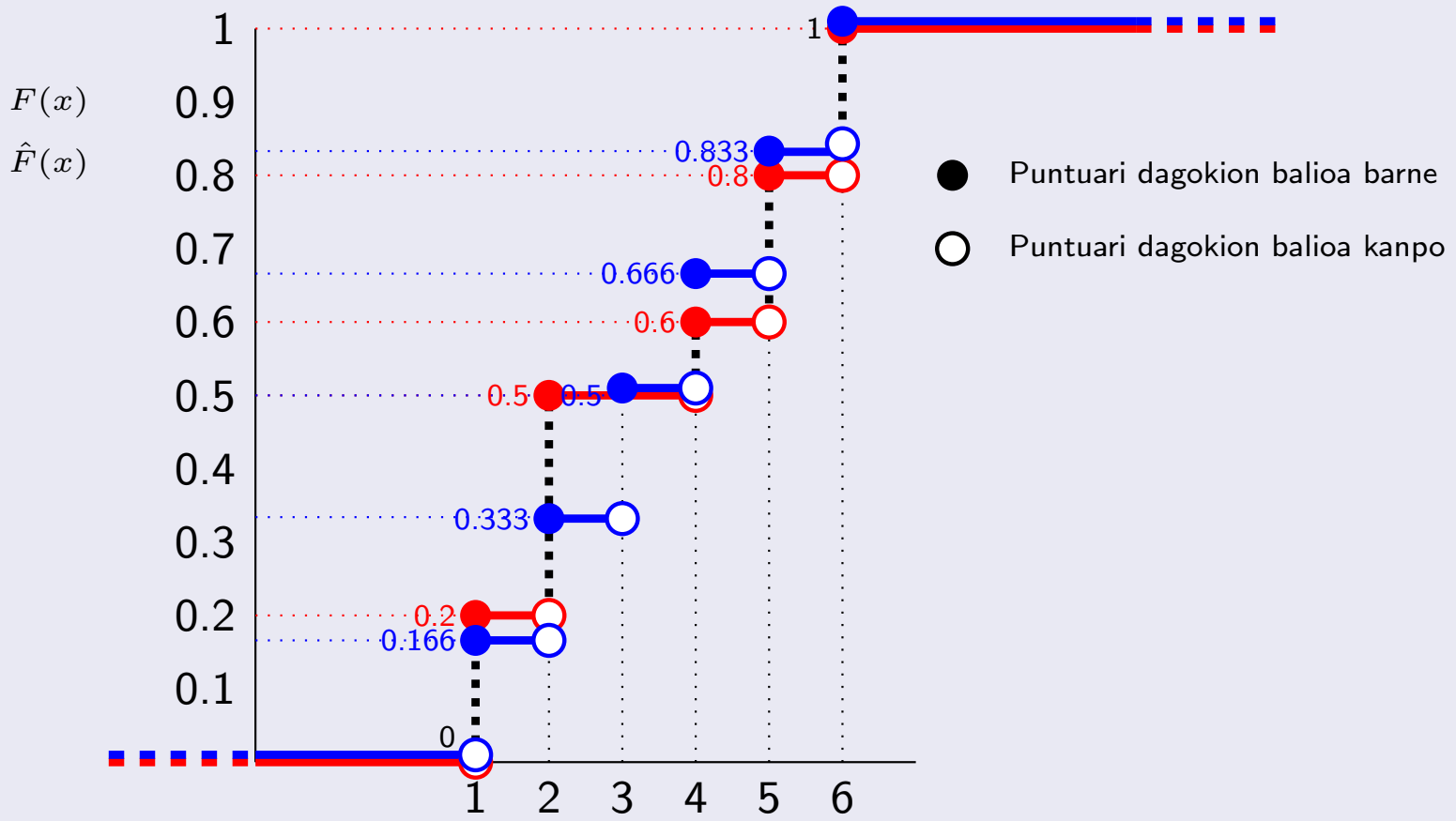
Banaketa-funtzio enpirikoa banaketa-funtzio teoriko batekin batera jarri behar dugu, biak alderatzeko. Dadoaren adibidean, dadoa ongi egina dagoen eta beraz alde guztietako probabilitateak 1 diren erabaki nahi dugu. beraz, banaketa teorikoa hau izango da:

x	1	2	3	4	5	6
p(x)	1/6	1/6	1/6	1/6	1/6	1/6
F(x)	1/6	2/6	3/6	4/6	5/6	1

Banaketa-funtzio enpirikoa eta teorikoa behar bezala bereizteko, notazio hau erabiliko dugu hemendik aurrera: $F(x)$ banaketa-funtzio teorikoa izango da, eta $\hat{F}(x)$ banaketa-funtzio enpirikoa.

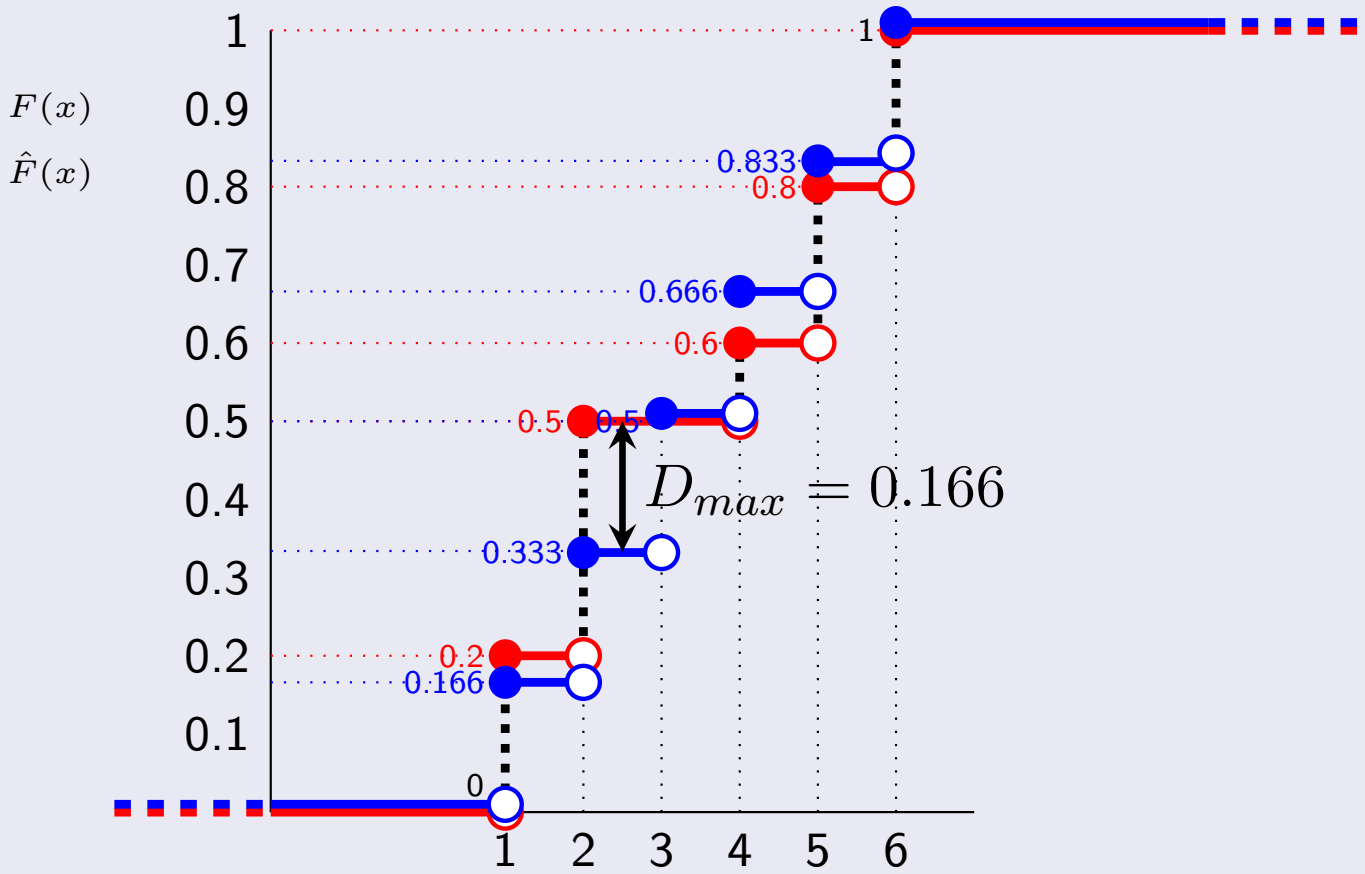
Banaketa-funtzio teorikoa

Banaketa-funtzio enpirikoarekin batera grafikatzen dugu:



Distantzia maximoa

Eta ondoren bi eskailera edo banaketa-funtzioen arteko balio maximoa bilatu eta kalkulatzen dugu:



Distantzia maximoa taula bidez

Arestikoa adibide sinplea izan arren, distantzia maximoa bilatzea ez da erraza, eta gainera eskuz egiteko, zehaztasun handia beharko genuke. Baina, (eskerrak!), kalkula daiteke taula bidez ere:

x	$\hat{F}(x)$	$F(x)$	$D = \lvert\hat{F}(x) - F(x)\rvert$
1^-	0	0	0
1	0.2	0.166	0.033
2^-	0.2	0.166	0.033
2	0.5	0.333	0.166
3^-	0.5	0.333	0.166
3	0.5	0.5	0
4^-	0.5	0.5	0
4	0.6	0.666	0.066
5^-	0.6	0.666	0.066
5	0.8	0.833	0.033
6^-	0.8	0.833	0.033
6	1	1	0

Esango duzu: zertarako kalkulatu beharra (-) puntuetan ere? Orain, badirudi ez dela beharrezkoa, baina banaketa teorikoa jarraitua denean, ikusiko duzu ezinbestekoa dela.

Azken erabakia: esangura maila

Orain, D_{max} distantzia banaketa teorikoa datu horietarako egokia ez dela baieztatzeko aski handia den erabaki behar dugu. Printzipioz (estatistikan, hipotesi nulupean diogu) sinesten dugu banaketa teorikoa egokia dela datuetarako, eta portzentaje txiki batez ez gara fidatzen horrekin. Banaketa teorikoarekiko mesfidantza-portzentaje horri α esangura-maila edo adierazgarritasun-maila deritzo, eta 0.2, 0.1, 0.5, 0.02 edo 0.01 izaten da. Ikertzaileak erabakitzen du maila hori aldez aurretik: zenbat eta sinistuagoa izan banaketarekin, orduan eta α txikiagoa jarriko du.

Azken erabakia: taula eta distantzia kritikoa

α esangura-maila bakoitzeko (hobe litzateke agian mesfidantza-maila deitzea), distantzia *kritiko* desberdinak ditugu. Suertatu zaigun distantzia hori, taulak ematen duen balio kritikoa baino handiagoa bada, distantzia aski handia dela erabaki eta banaketa teorikoa baztertu egiten dugu, hots datuetarako egokia ez dela erabaki. Gurean, $n = 10$ eta $\alpha = \%5 = 0.05$ finkatzen badugu, distantzia kritikoa 0.409 da. 0.166 suertatu zaigu, ez da hartara balio kritikoraino iristen, eta beraz banaketa teorikoa datuetarako egokia dela erabakitzen dugu (dadoa ongi eraturik dagoela alegia, probabilitate berdinekin).

n	$\alpha = 0.20$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.02$	$\alpha = 0.01$
9	0,339	0,387	0,43	0,48	0,513
10	0,323	0,369	0,409	0,457	0,489
11	0,308	0,352	0,391	0,437	0,468

Taula eskuragarri hemen: https://gizapedia.hirusta.io/pdf-printer-friendly-statistical-table-for-the-one-sample-kolmogorov-smirnov-test/.

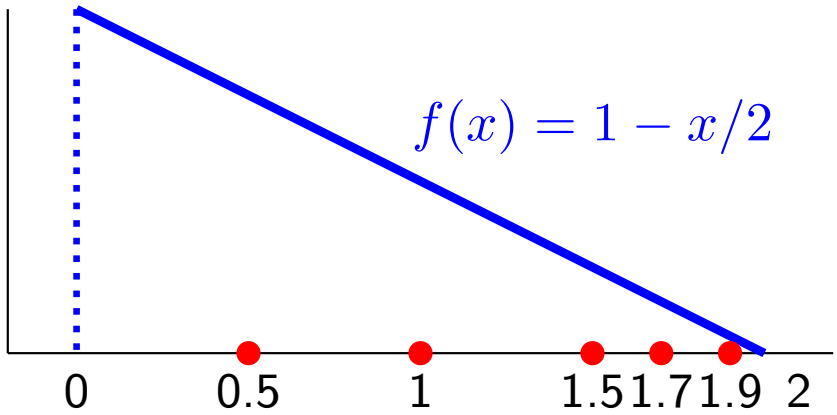
Azken ohar bat α-ri buruz

Ohartu α zenbat eta txikiagoa izan, orduan eta distantzia kritiko handiagoa dugula. Normala da: zenbat eta α txikiagoa izan, orduan eta konfiantza handiagoa dugu banaketa teorikoarekin, eta distantzia kritikoa handiago jarri behar da banaketa baztertzeko.

K-S proba banaketa teorikoa jarraitua denean

0.5-1-1.5-1.7-1.9 lagin datuetarako banaketa-funtzio teoriko gisa $F(x) = x - x^2/4; 0 \leq x \leq 2$ ezarri da. Banaketa teoriko egokia dela esan al daiteke? $\alpha = 0.10$.

K-S proba garatu aurretik, halako esplorazio bat egin dezakegu galdetzen duenari buruz, dentsitate-funtzioa datuekin batera irudikatuz:



Ikusten den den bezala, ereduan ezartzen da 2 inguruko balioak direla probabilitate txikienekoak, eta 0 ingurukoak handienekoak. Datuak, berriz, ugaldu egiten dira 2 baliora gerturatu ahala. Beraz, ez dirudi eredua oso egokia denik.

K-S proba banaketa teorikoa jarraitua denean

0.5-1-1.5-1.7-1.9 lagin datuetarako banaketa-funtzio teoriko gisa $F(x) = x - x^2/4; 0 \leq x \leq 2$ ezarri da. Banaketa teoriko egokia dela esan al daiteke? $\alpha = 0.10$.

x	$\hat{F}(x)$	$F(x)$	$D = \lvert\hat{F}(x) - F(x)\rvert$
0.5^-	0	0.4375	0.4375
0.5	0.2	0.4375	0.2375
1^-	0.2	0.75	0.55
1	0.4	0.75	0.35
1.5^-	0.4	0.9375	0.5375
1.5	0.6	0.9375	0.3375
1.7^-	0.6	0.9775	0.3775
1.7	0.8	0.9775	0.1775
1.9^-	0.8	0.9975	0.1975
1.9	1	0.9975	0.0025

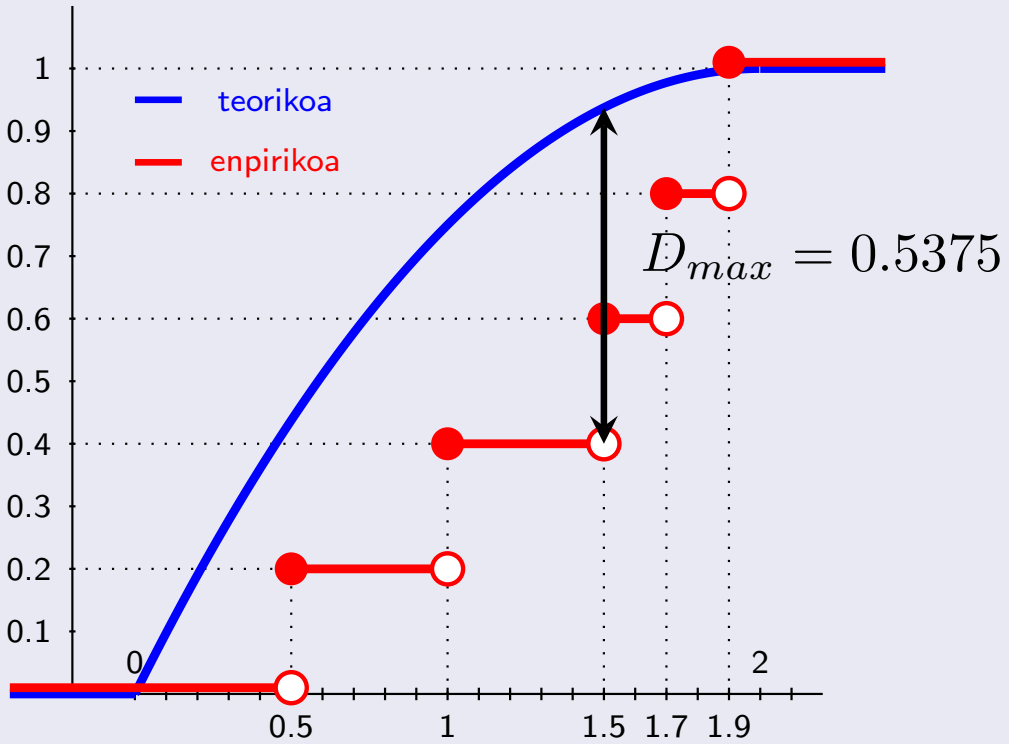
Banaketa teorikoaren eta enpirikoaren arteko distantzia maximoa $D_{max} = 0.5375$ da.

K-S proba, banaketa teorikoa jarraitua denean

$n = 5$ datuetarako eta $\alpha = \%10$ harturik, distantziaren balio kritikoa, hortik gora banaketa teorikoa baztertzera daramana, $D^* = 0.509$ da. $D_max = 0.5375 > 0.509$. Beraz, hipotesi nulua, banaketa teorikoa datuetarako egokia dela alegia, baztertu behar da.

K-S proba, banaketa teorikoa jarraitua denean: grafikoki

$F(x)$ (teorikoa), enuntziatuan ematen den $F(x)$ funtzioaren adierazpen grafikoa baino ez da, balio posibleen tartean, 0tik 2ra alegia.



Ariketa I

Ikasle batzuek suspenditutako irakasgai kopurua jaso da: 1-1-2-3-4. Eredu gisa, suspenditutako irakasgai kopurua 1, 2, 3 edo 4 izan daitekeela planteatu da, irakasgai bat gehiago suspenditzeko probabilitatea aurrekoa baino 0.1 txikiagoa delarik. Eredu egokia dela esan al daiteke? $\alpha = 0.05$. Grafikoki nahiz analitikoki ebatzi.

Ariketa II

Eguneko salmenta hauek jaso dira enpresa batean: 0.8-1.4-2.8-3.4. Salmentetarako eredu hau ezarri da: $f(x) = 0.25; 0.5 < x < 4.5$. Eredua egokia dela esan al daiteke? $\alpha = \%1$ Grafikoki nahiz analitikoki ebatzi.