Banaketa normalak alderatzen: Bhattacharyya koef.

Josemari Sarasola

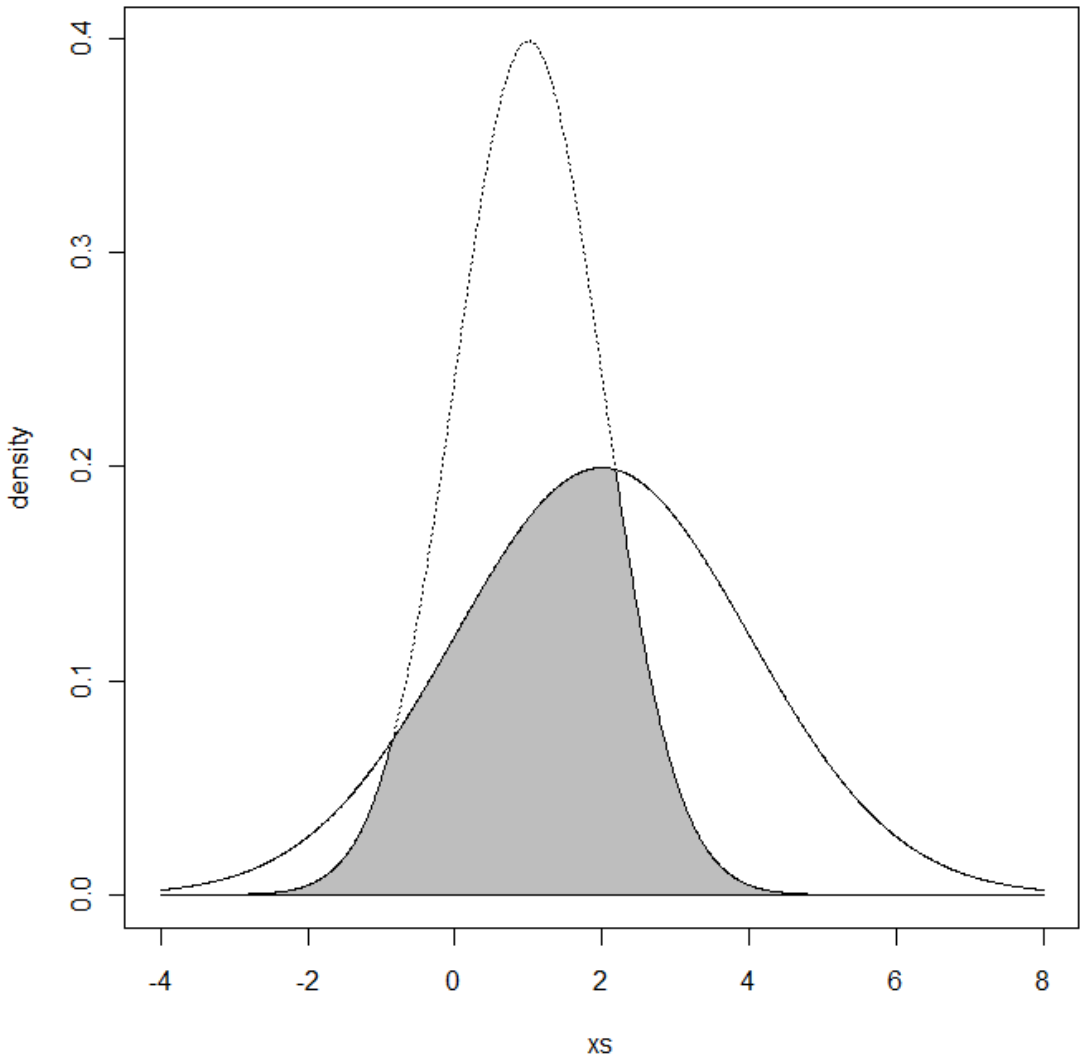
Gizapedia

Banaketa normalak alderatzen

Problema

- Praktikan ohikoa da datu multzoak alderatu nahi izatea, zein zeinekin de antzekoen, eta datu multzoen arteko distantzia handitu den ikusteko.
- Banaketa normala datuetarako maiz erabiltzen den eredua izanik, banaketa normalen arteko distantzia (edo alderantziz, antzekotasuna) nola jaso dezakegun ikasiko dugu.
- Horretarako, bi banaketa normala hartu, parametro ezagunekin, eta biak zenbateraino solapatzen diren neurtu behar dugu.

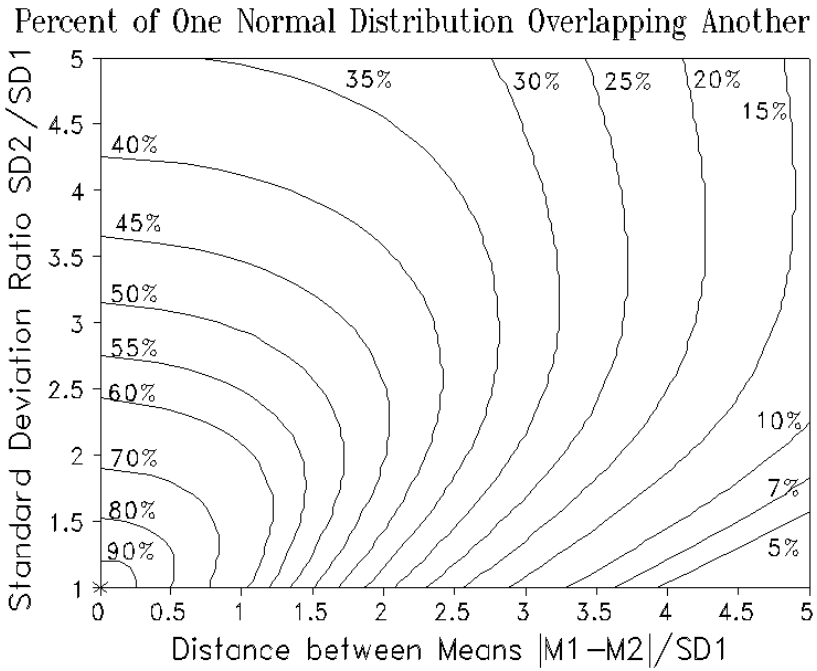
Banaketa normalak alderatzen



Solapamendua (overlapping) zenbat eta handiagoa izan, orduan eta distantzia edo diferentzia handiagoa izango da bi banaketen artean.

Banaketa normalak alderatzen

Bitez $N(\mu_1, \sigma_1)$ eta $N(\mu_2, \sigma_2)$ bi banaketa normal, batezbesteko (mean) eta desbideratze (sd, standard deviation) ezagunekin. Solapamenduzko probabilitateari Bhattacharyya koefizientea deitzen zaio eta nomograma honekin kalkula daiteke, SD1 desbideratze txikiena izanik:



Solapamendua (overlapping) zenbat eta handiagoa izan, orduan eta distantzia edo diferentzxia handiagoa izango da bi banaketen artean.

Banaketa normalak alderatzen

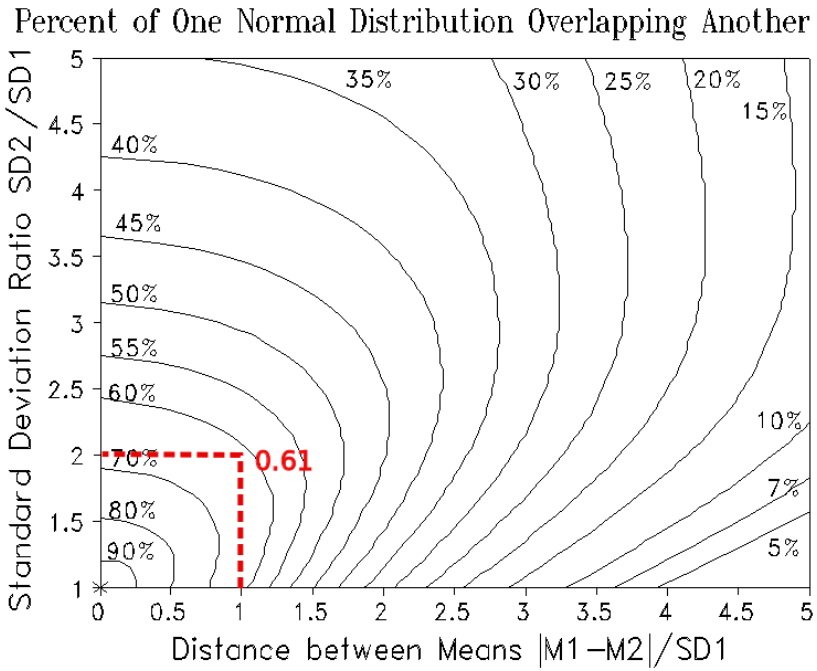
Solapamenduzko probabilitateari Bhattacharyya koefizientea (BC) deitzen zaio, eta bi banaketen arteko antzekotasunaren neurria da. Horren ordez, distantzia edo diferentzia jaso nahi badugu Bhattacharyya distantzia kalkulatzen dugu:

$$D_{BC} = -ln\ BC$$

BC solapamenduak 0-1 bitarteko balioak hartzen ditu eta BC distantziak beti balio positiboak, infinituraino.

Banaketa normalak alderatzen

Adibide bat: zenbateko probabilitatean solapatzen dira $N(0,1)$ eta $N(1,2)$? Eta zenbatean desberdintzen dira?
SD1 desbideratze txikiena da: 1. Beraz, $|M1 - M2|/SD1 = |0-1|/1 = 1$ eta $SD2/SD1 = 2$. Haietatik, nomograma erabiliz, solapamendua gutxi gorabehera 0.62koa estimatu dezakegu, eta horik distantzia $-ln0.61 = 0.49$.



Banaketa normalak alderatzen: ariketak

1: Makina batek egun batean burutzen duen ekoizpena $N(1000, 200)$ banatzen da. Makina horrek ekoizten dituen piezen kalitate-kontrola egiteko beste bi makina ditugu aukeran. A makinak egunean kontrolatzen duen pieza kopurua $N(900, 150)$ banatzen da eta B makinak kontrolatzen dituenak $N(1050, 220)$. Nolabait ekoizpen-kontrol prozesuan geldialdi eta metaketarik ez izateko, zein makina da onena piezak kontrolatzeko?
2: Salmentak sexuaren arabera bereizi eta aztertu ondoren, normal banatzen direla ondorioztatu da. Azken bi urteetako datuetan oinarritua, gizonezkoen batezbesteko erosketa 100 eurokoa da, 25eko desbideratzearekin, lehenengo urtean, eta 130 eurokoa, 30eko desbideratzearekin bigarrenean. Emakumeei buruz, lehenengo urtean, 140 euroko batezbestekoa, 30eko desbideratzearekin, eta bigarren urtean, 160ko batezbestekoa, 30eko desbideratzearekin. Zein urtetan dago distantzia handiena emakume eta gizonen artean?