8 Ehdollinen jakauma

Tämän kappaleen tärkeitä käsitteitä:

- Ehdollinen jakauma; ehdollinen ptnf/tf.
- Kertolaskusääntö eli ketjusääntö yhteisjakauman esittämiseksi.
- Ehdollinen odotusarvo ja ehdollinen varianssi.
- Yhteisjakauman määrittely hierarkkisesti kertolaskusäännön avulla.

8.1 Ehdolliset jakaumat

- Jos sv:lla (X, Y) on diskreetti jakauma, niin sm:n X ehdollinen ptnf ehdolla $Y = y$ määritellään ehdollisen todennäköisyyden kaavan avulla,

$$f_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \qquad \text{kun } f_Y(y) > 0. \tag{1}$$

- Funktio $f_{X|Y}(\cdot \mid y)$ on satunnaismuuttujan X ptnf, kun tiedetään, että $Y = y$.
- Ehdollinen ptnf $f_{Y|X}(y \mid x)$ määritellään analogisesti.
- Kaavaa (1) pidetään mallina, kun määritellään ehdollinen tiheysfunktio jatkuvan yhteisjakauman tapauksessa.

Ehdollinen tiheysfunktio, kun yhteisjakauma on jatkuva

Määritelmä

Olkoon (X, Y):llä jatkuva jakauma tiheysfunktiolla $f_{X,Y}$. Tällöin sm:n X ehdollinen tiheysfunktio ehdolla $Y = y$ on

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \qquad x \in \mathbb{R},$$

kun y on sellainen, että $f_Y(y) > 0$. Vastaavasti määritellään

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)}, \qquad y \in \mathbb{R},$$

kun $f_X(x) > 0$.

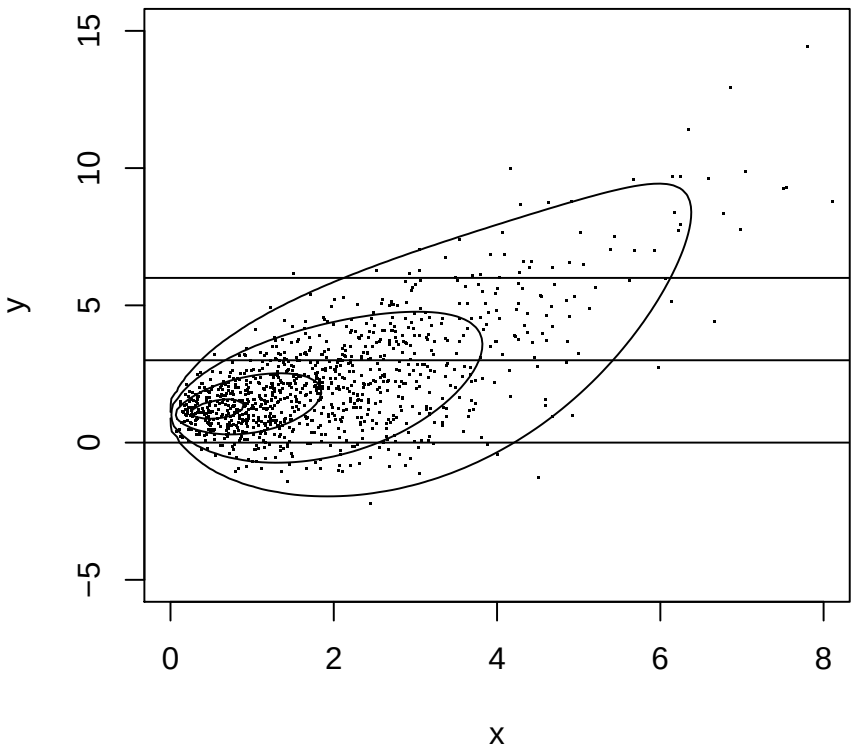
Huomautuksia ehdollisen tiheysfunktion määritelmästä

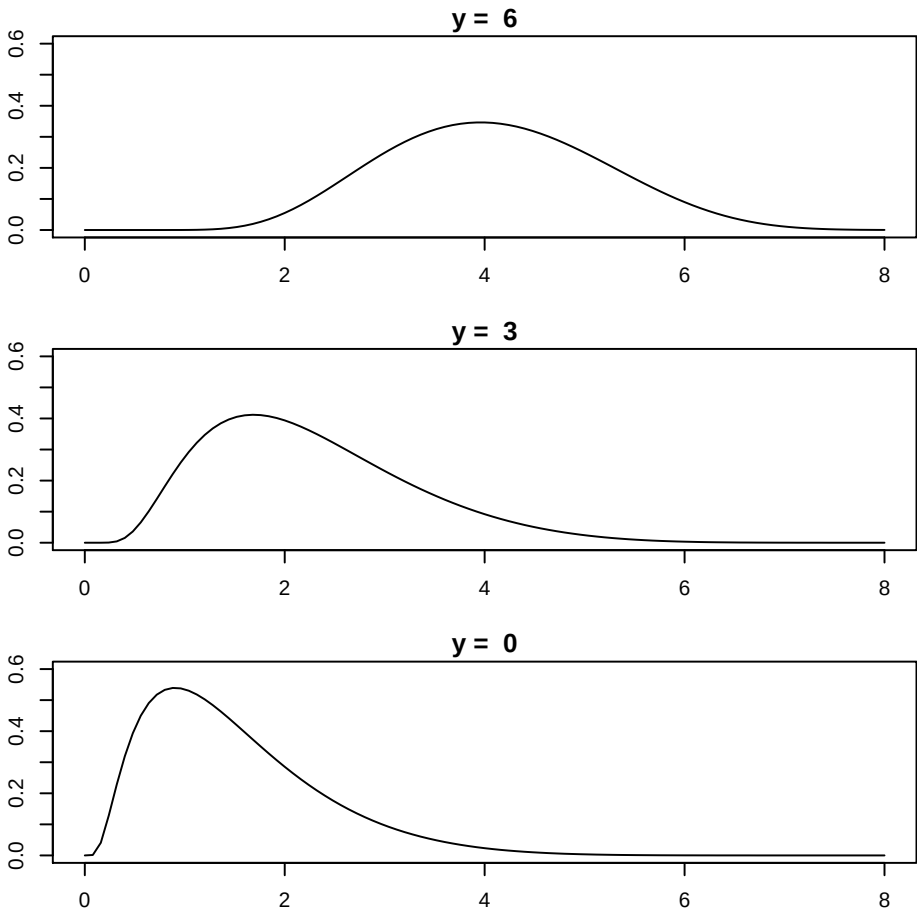
- Jos $f_Y(y) > 0$, niin ehdollinen tiheysfunktio $x \mapsto f_{X|Y}(x \mid y)$ on tiheysfunktio, sillä se on ei-negatiivinen ja

$$\int_{-\infty}^{\infty} f_{X|Y}(x \mid y)\,\mathrm{d}x = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x = 1.$$

- Geometrinen tulkinta:
  - Funktio $x \mapsto f_{X|Y}(x \mid y)$ saadaan yhteistiheysfunktiosta $f_{X,Y}$ normalisoimalla sen vaakasuora "leike" $x \mapsto f_{X,Y}(x,y)$ tiheysfunktioksi.
  - Funktio $y \mapsto f_{Y|X}(y \mid x)$ saadaan normalisoimalla pystysuora leike $y \mapsto f_{X,Y}(x,y)$ tiheysfunktioksi.

Ytf ja ehdollisia tiheysfunktioita $x \mapsto f_{X|Y}(x \mid y)$





Onko ehdollinen tiheysfunktio hyvin määritelty?

- Jatkuvan yhteisjakauman voi esittää useamman kuin yhden tiheysfunktion avulla kunhan ne vain yhtyvät melkein kaikkialla.
- Ehdollisen tf:n määrittelyyn saadaan käyttää mitä tahansa ytf:n versiota, joten ehdollinen tf ei ole yksikäsitteinen funktio.
- Monikäsitteisyydestä ei synny ongelmia, minkä takia sitä ei oteta huomioon puhetavassa: puhutaan ehdollisesta tiheysfunktiosta (eikä esim. ehdollisen tiheysfunktion tietystä versiosta).
- Jatkuvan yhteisjakauman tapauksessa ehdon $\{Y = y\}$ todennäköisyys on aina nolla. Tämän takia ehdolliselle tiheysfunktiolle ei voida antaa suoraan tulkintaa ehdollisen todennäköisyyden avulla, vaan ehdollisten todennäköisyyksien raja-arvojen kautta.

Ehdollisen tf:n tulkinta ehdollisten tn:ien raja-arvojen avulla

- Teemme uskottavaksi, että jos ytf $f_{X,Y}$ on riittävän sileä ja $f_Y(y) > 0$, niin tällöin

$$P(X \in A \mid y \leq Y \leq y + h) \xrightarrow[h \to 0+]{} \int_A f_{X|Y}(x \mid y)\, \mathrm{d}x, \quad \text{kaikilla } A \subset \mathbb{R}. \tag{2}$$

- Ts. pienillä $h > 0$ ehdollinen todennäköisyys $P(X \in A \mid y \leq Y \leq y + h)$ saadaan osapuilleen selville integroimalla ehdollista tiheysfunktiota $f_{X|Y}(x \mid y)$ muuttujan x suhteen joukon A yli mielivaltaiselle $A \subset \mathbb{R}$.
- Jatkuvan sm:n arvo havaitaan vain tietyllä tarkkuudella, joten sovelluksissa tehdään tämän tapainen tulkinta, kun käsitellään ehdollista tf:a ehdolla $Y = y$, jossa y on havaittu arvo.

Raja-arvotuloksen perustelu

- Lähdetään liikkeelle esityksestä

$$P(X \in A \mid y \le Y \le y+h) = \frac{\frac{1}{h} P(X \in A, y \le Y \le y+h)}{\frac{1}{h} P(y \le Y \le y+h)}$$
$$= \frac{\int_A \mathrm{d}x \, \frac{1}{h} \int_y^{y+h} f_{X,Y}(x,t)\,\mathrm{d}t}{\frac{1}{h} \int_y^{y+h} f_Y(u)\,\mathrm{d}u}$$

- Sekä osoittajassa että nimittäjässä esiintyy muotoa

$$\frac{1}{h} \int_y^{y+h} g(u)\,\mathrm{d}u$$

olevia integraalikeskiarvoja, jotka lähestyvät arvoa $g(y)$ aika yleisillä oletuksilla, kun $h \to 0+$.

Miksi integraalikeskiarvo suppenee

- Oletetaan esimerkin vuoksi, että funktio g on jatkuva pisteessä y ja integroituva jossakin sen ympäristössä.
- Olkoon $\epsilon > 0$ mielivaltainen.
- Valitaan (jatkuvuuden nojalla) $\delta > 0$ siten, että $|g(u) - g(y)| \leq \epsilon$ kaikilla u, joille $|u - y| < \delta$.
- Jos $0 < h < \delta$, niin

$$g(y) - \epsilon = \frac{1}{h} \int_y^{y+h} (g(y) - \epsilon) \, \mathrm{d}u \leq \frac{1}{h} \int_y^{y+h} g(u) \, \mathrm{d}u \leq \frac{1}{h} \int_y^{y+h} (g(y) + \epsilon) \, \mathrm{d}u = g(y) + \epsilon.$$

- Tämä todistaa, että integraalikeskiarvo lähestyy raja-arvoa $g(y)$, kun $h \to 0+$.

Perustelu päättyy

- Sovelletaan integraalikeskiarvon raja-arvoa kaavassa

$$P(X \in A \mid y \leq Y \leq y+h) = \frac{\int_A \mathrm{d}x \, \frac{1}{h} \int_y^{y+h} f_{X,Y}(x,t)\,\mathrm{d}t}{\frac{1}{h}\int_y^{y+h} f_Y(u)\,\mathrm{d}u}$$

- Vaihdetaan osoittajassa huolettomasti rajankäynnin ja integroinnin järjestys

$$P(X \in A \mid y \leq Y \leq y+h) \xrightarrow[h\to 0+]{} \frac{\int_A f_{X,Y}(x,y)\,\mathrm{d}x}{f_y(y)} = \int_A \frac{f_{X,Y}(x,y)}{f_Y(y)}\,\mathrm{d}x.$$

- Mitkä olisivat tyylikkäät oletukset, joilla kaikki tämän perustelun askeleet saataisiin vietyä läpi täsmällisesti?

8.2 Kertolaskusääntö eli ketjusääntö

- Ehdollisen ptnf/tf:n määritelmästä saadaan kertolaskusääntö eli ketjusääntö,

$$f_{X,Y}(x,y) = f_X(x)\, f_{Y|X}(y \mid x) = f_Y(y)\, f_{X|Y}(x \mid y), \quad \text{kaikilla } x, y. \tag{3}$$

- Mitä tämä tarkoittaa jatkuvan jakauman tapauksessa, jossa ytf ei ole yksikäsitteinen?
- Jos reunatiheysfunktio $f_X(x)$ ja ehdollinen tf $f_{Y|X}(y \mid x)$ johdetaan lähtemällä liikkeelle joistakin ytf:n (mahdollisesti toisistaan eriävistä) versioista, niin niiden tulo kelpaa yhteistiheysfunktioksi.

Ongelma kertolaskusäännön $f_X(x)\, f_{Y|X}(y \mid x)$ soveltamisessa

- Ehdollinen tiheys $f_{Y|X}(y \mid x)$ ei ole välttämättä määritelty kaikilla x.
- Kertolaskusäännön yhteydessä ehdollisen tf:n tai ptnf:n määritelmää laajennetaan (tarvittaessa) jollakin tavalla niihin pisteisiin, joissa ehtomuuttujan tiheys on nolla, esim. sopimalla, että

$$f_{Y|X}(y \mid x) = \begin{cases} \dfrac{f_{X,Y}(x,y)}{f_X(x)}, & \text{kun } f_X(x) > 0, \\ 0, & \text{muuten.} \end{cases} \tag{4}$$

- Toiselle ehdolliselle tf:lle $f_{X|Y}(x \mid y)$ käytetään samanlaista laajennusta.
- Tämän jälkeen kertolaskusääntö pitää ongelmattomasti paikkansa kaikilla $x, y \in \mathbb{R}$.
- Nämä komplikaatiot eivät aiheuta todellisia ongelmia käytännön laskuissa.

Bayesin kaava (tiheysfunktioille)

- Kertolaskusäännöstä voidaan ratkaista toinen ehdollisista tiheysfunktioista, jos reunatiheydet sekä toinen ehdollinen tf tunnetaan.
- Esim., kun $f_X(x) > 0$ (ja muut x-arvot eivät ole relevantteja), on

$$f_{Y|X}(x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_Y(y)\, f_{X|Y}(x \mid y)}{f_X(x)}. \tag{5}$$

- Tämä on Bayesin kaava tiheysfunktioille.

Bayesin kaava tulkittuna verrannollisuustulokseksi

- Bayesin kaavasta seuraa, että kiinteällä x

$$f_{Y|X}(y \mid x) = \frac{1}{f_X(x)} f_Y(y)\, f_{X|Y}(x \mid y)$$
$$\propto f_Y(y)\, f_{X|Y}(x \mid y) = f_{X,Y}(x, y).$$

- Verrannollisuus $g(y) \propto h(y)$ tarkoittaa sitä, että

$$g(y) = c\, h(y)$$

  jollakin verrannollisuusvakiolla c (joka saa riippua parametreista, ja muista muuttujista paitsi muuttujasta y).
- Tämän ansiosta ehdollinen jakauma on toisinaan helppo tunnistaa tarkastelemalla ytf:n lauseketta.

Esimerkki verrannollisuustarkastelusta

- Esimerkissä 7.2 (tasajakauma funktion $h(x) = \exp(-\frac{1}{2}\,x^2)$ kuvaajan alla) ytf:llä on lauseke

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}}\,1\{0 < y < \exp(-\frac{1}{2}x^2)\}.$$

- Ehdolliset tf:t voidaan laskea joko jakamalla ytf esimerkissä johdetuilla reunatf:oilla, tai seuraavalla päättelyllä.

$f_{Y|X}$ esimerkissä

- Kiinteällä x on ytf:llä $f_{X,Y}(x, y)$ nollaa suurempi vakioarvo, kun $0 < y < \exp(-\frac{1}{2}x^2)$ ja muuten arvo nolla. Tämän takia ehdollinen jakauma on tasajakauma,

$$Y \mid (X = x) \sim \mathrm{U}(0, \exp(-\frac{1}{2}x^2)).$$

- Kun muistetaan, että reunajakaumassaan $X \sim N(0, 1)$, niin yhteisjakauma voidaan esittää muodossa

$$Y \mid X \sim \mathrm{U}(0, \exp(-\frac{1}{2}X^2)),$$
$$X \sim N(0, 1).$$

- Tämä on esimerkki yhteisjakauman hierarkkisesta määrittelystä (reunajakauma f_X ja ehdollinen jakauma $f_{Y|X}$).

$f_{X|Y}$ esimerkissä

- Kiinteällä $0 < y < 1$ on ytf:llä nolla suurempi vakioarvo tietyllä x-akselin välillä, joka saadaan ratkaisemalla epäyhtälö

$$0 < y < \exp(-\frac{1}{2}x^2)$$

  muuttujan x suhteen. Tämä lasku ratkaistiin jo aiemmin.
- Tuloksesta tunnistetaan, että ehdollinen jakauma on tasajakauma

$$X \mid (Y = y) \sim \mathrm{U}(-\sqrt{-2\ln y}, \sqrt{-2\ln y}), \qquad 0 < y < 1.$$

- Tämä tulos ja Y:n reunajakauma (ks. esimerkki 7.2) kertovat yhteisjakaumalle toisen hierarkkisen esityksen (f_Y ja $f_{X|Y}$).

8.3 Diskreetin ja jatkuvan muuttujan yhteisjakauma

- Tarkastellaan sm:ien X ja Y yhteisjakaumaa siinä tapauksessa, kun X:n jakauma on diskreetti Y:n jatkuva.
- Olkoon X:n mahdolliset arvot $x_1, x_2, \ldots$ ja olkoon sen (reuna-)ptnf f_X.
- Voidaan osoittaa, että tällöin on olemassa ehdolliset tiheysfunktiot $y \mapsto f_{Y|X}(y \mid x_i), i \geq 1$ siten, että

$$P(X = x_i, Y \in B) = f_X(x_i) \int_B f_{Y|X}(y \mid x_i)\,\mathrm{d}y,$$

kaikilla i ja kaikilla $B \subset \mathbb{R}$,

mutta todistuksessa tarvitaan mittateoriaa (tulos seuraa ns. Radonin-Nikodymin lauseesta).

Diskreetin X ja jatkuvan Y yhteisjakauma

- Yhteisjakauman esittää funktio

$$f_{X,Y}(x,y) = f_X(x)\, f_{Y|X}(y \mid x).$$

- Seuraavantyyppiset todennäköisyydet saadaan laskettua summaamalla diskreetin muuttujan ja integroimalla jatkuvan muuttujan suhteen,

$$P(X \in A, Y \in B) = \sum_{x \in A} \int_B f_{X,Y}(x,y)\, \mathrm{d}y, \qquad A, B \subset \mathbb{R}\,.$$

- Voimme myös tässä tapauksessa kutsua yhteisjakauman esitystä $f_{X,Y}$ tiheydeksi tai tiheysfunktioksi, mutta on tärkeää pitää mielessä, että yhden muuttujan suhteen summataan ja toisen suhteen integroidaan.

Ehdolliset jakaumat, kun X diskreetti ja Y jatkuva

- Jos $P(X = x_i) > 0$, niin jatkuvan muuttujan Y ehdollinen tiheys on $f_{Y|X}(y \mid x_i)$.
- Diskreetin muuttujan X ptnf ehdolla $Y = y$ määritellään Bayesin kaavalla, siis

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)},$$

  kun $f_Y(y) > 0$.
- Tämän kaavan voisi myös motivoida rajankäynnin kautta samaan tapaan kuin jatkuvan yhteisjakauman tapauksessa tehtiin.
- Edellä $f_Y(y) = \sum_x f_{X,Y}(x, y)$.

Diskreetti X ja jatkuva Y

- Kertolaskusääntö ja Bayesin kaava ovat voimassa myös tässä tapauksessa.
- Tarvittaessa ehdollisen tiheyden $f_{Y|X}(y \mid x)$ ja ehdollisen ptnf:n $f_{X|Y}(x \mid y)$ määritelmää voidaan laajentaa myös sellaisille argumenteille, joilla ne eivät vielä tulleet määriteltyä.

Diskreetti X ja jatkuva Y: tiedostamattoman tilastotieteilijän laki

Lause

Olkoon (X, Y) sv, jossa X:llä on diskreetti ja Y:llä jatkuva jakauma, ja olkoon $Z = g(X, Y)$ jokin sen reaaliarvoinen muunnos. Tällöin

$$EZ = \sum_x \int g(x,y)\, f_{X,Y}(x,y)\, \mathrm{d}y = \int \sum_x g(x,y)\, f_{X,Y}(x,y)\, \mathrm{d}y$$

mikäli

$$\sum_x \int |g(x,y)|\, f_{X,Y}(x,y)\, \mathrm{d}y < \infty.$$

Diskreetin muuttujan suhteen summataan ja jatkuvan suhteen integroidaan.

8.4 Ehdollinen odotusarvo

Määritelmä (Ehdollinen odotusarvo ehdolla sm:n arvo)

Olkoon (X, Y) sv, $g(X, Y)$ jokin sen muunnos, ja oletetaan, että osaamme määritellä sm:n Y ehdollisen jakauman ehdolla $X = x$. Sm:n $g(X, Y)$ ehdollinen odotusarvo ehdolla $X = x$,

$$E(g(X, Y) \mid X = x),$$

on satunnaismuuttujan $g(x, Y)$ odotusarvo, kun Y:n jakaumana käytetään sen ehdollista jakaumaa ehdolla $X = x$.

Ts.

$$E(g(X, Y) \mid X = x) = \begin{cases} \int g(x, y)\, f_{Y|X}(y \mid x)\, \mathrm{d}y, & Y \text{ jatkuva} \\ \sum_y g(x, y)\, f_{Y|X}(y \mid x), & Y \text{ diskreetti.} \end{cases}$$

Tunnetut tekijät voidaan vetää ulos

- Jos funktio g on muotoa

$$g(x, y) = g_1(x)\, g_2(x, y),$$

niin tietenkin

$$E(g_1(X)\, g_2(X, Y) \mid X = x) = g_1(x)\; E(g_2(X, Y) \mid X = x). \tag{6}$$

- Tämä voidaan ilmaista sanomalla, että tunnetut tekijät saadaan vetää ulos ehdollisesta odotusarvosta.
- Jos ehtona on $X = x$, niin kaikki yksinomaan sm:sta X riippuvat tekijät ovat tunnettuja.

Regressiofunktio

Määritelmä (Y:n ehdollinen odotusarvo ehdolla $X = x$; regressiofunktio)

Sm:n Y ehdollinen odotusarvo ehdolla $X = x$ on sen ehdollisen jakauman odotusarvo, $E(Y \mid X = x)$. Funktiota

$$x \mapsto E(Y \mid X = x)$$

kutsutaan Y:n *regressiofunktioksi* X:n suhteen (engl. *regression function of Y on X*).

Regressiofunktio saadaan kaavoilla

$$E(Y \mid X = x) = \begin{cases} \int y \, f_{Y|X}(y \mid x) \, \mathrm{d}y, & \text{jos } Y \text{ jatkuva,} \\ \sum_y y \, f_{Y|X}(y \mid x), & \text{jos } Y \text{ diskreetti.} \end{cases}$$

Ehdollinen varianssi

Määritelmä (Ehdollinen varianssi ehdolla sm:n arvo)

Sm:n $g(X, Y)$ ehdollinen varianssi ehdolla $X = x$,

$$\operatorname{var}(g(X, Y) \mid X = x),$$

on satunnaismuuttujan $g(x, Y)$ varianssi, kun Y:n jakaumana käytetään sen ehdollista jakaumaa ehdolla $X = x$.

Kun merkitään $m(x) = E(g(X, Y) \mid X = x)$, niin

$$\operatorname{var}(g(X, Y) \mid X = x) = \begin{cases} \int [g(x, y) - m(x)]^2 \, f_{Y|X}(y \mid x) \, \mathrm{d}y, & Y \text{ jva,} \\ \sum_y [g(x, y) - m(x)]^2 \, f_{Y|X}(y \mid x), & Y \text{ diskr.} \end{cases}$$

Toinen kaava ehdolliselle varianssille

- Ts. kun merkitään $m(x) = E(g(X, Y) \mid X = x)$, niin

$$\operatorname{var}(g(X, Y) \mid X = x) = E[(g(X, Y) - m(X))^2 | X = x]$$

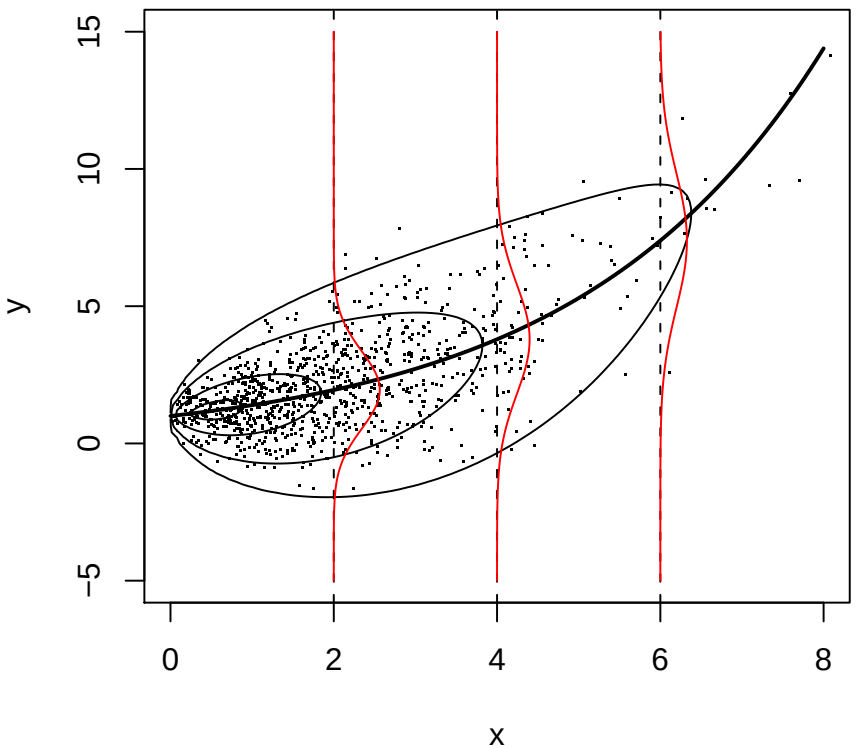
- Jos tässä binomi kerrotaan auki, lasketaan odotusarvo ja järjestellään termejä, nähdään että

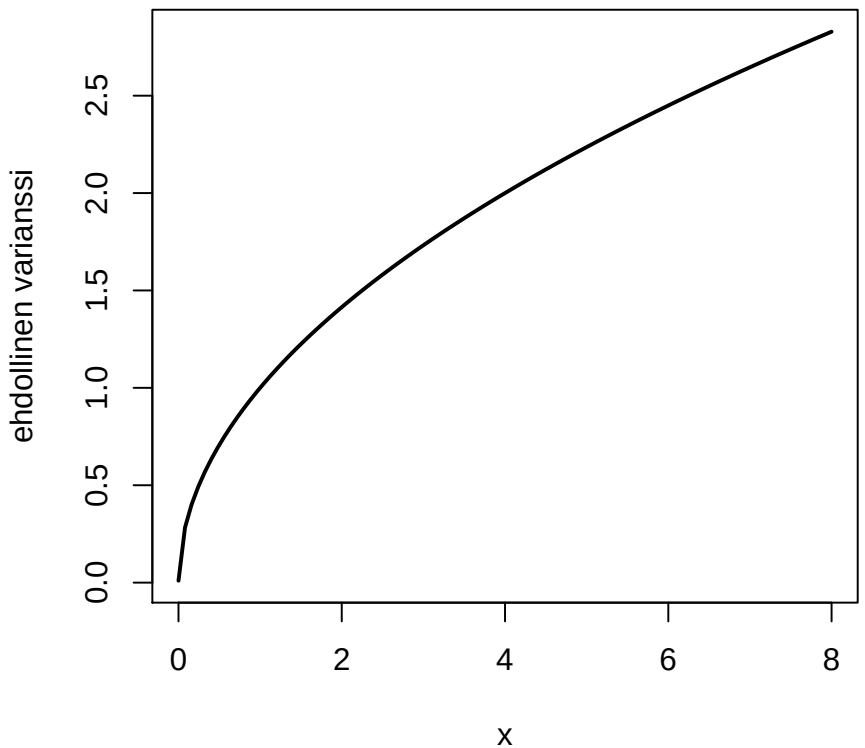
$$\operatorname{var}(g(X, Y) \mid X = x) = E[(g(X, Y))^2 \mid X = x] - [m(x)]^2. \tag{7}$$

- Tämä on tutun kaavan $\operatorname{var} Z = EZ^2 - (EZ)^2$ vastine ehdolliselle varianssille.

Ehdollinen odotusarvo ja ehdollinen varianssi esimerkissä

1. Yhteisjakauma, ehdollisia tiheysfunktioita $y \mapsto f_{Y|X}(y \mid x)$ ja ehdollinen odotusarvo $E(Y \mid X = x)$.
2. Ehdollinen varianssi $\mathrm{var}(Y \mid X = x)$.





Ehdollinen odotusarvo satunnaismuuttujana

Määritelmä (Ehdollinen odotusarvo ehdolla satunnaismuuttuja)

Merkitään väliaikaisesti

$$m(x) = E(g(X, Y) \mid X = x).$$

Sovitaan, että $m(x) = 0$ niillä x, joilla $m(x)$ se ei muuten tule määritellyksi, eli joilla ehdollinen jakauma $Y \mid (X = x)$ ei ole määritelty. Tämän jälkeen voidaan vapaasti puhua satunnaismuuttujasta $m(X)$. Sitä kutsutaan sm:n $g(X, Y)$:n ehdolliseksi odotusarvoksi ehdolla sm X.
Käytetään merkintää

$$E(g(X, Y) \mid X) = m(X).$$

Huomautuksia ehdollisesta odotusarvosta $E(g(X, Y) \mid X)$

- Määrittelimme $E(g(X, Y) \mid X) = m(X)$, jossa $m(x) = E[g(X, Y) \mid X = x]$.
- Mieleen saattaa juolahtaa käyttää merkintää $E(g(X, Y) \mid X = X)$, mutta se olisi järjetön.
- Ts. vakiintunut merkintä $E[g(X, Y) \mid X = x]$ on ongelmallinen, koska siihen ei saa sijoittaa symbolin x tilalle satunnaismuuttujaa X.
- Ehdollinen odotusarvo $E(g(X, Y) \mid X)$ on sellainen sm, joka saa arvon $E(g(X, Y) \mid X = x)$ (todennäköisyydellä yksi) silloin, kun X saa arvon x.

Ehdollinen odotusarvo on paras ennuste

Lause

Jos $E[g(X,Y)^2] < \infty$, niin $E(g(X,Y) \mid X)$ on keskineliövirheen mielessä paras sm:n $g(X,Y)$ ennuste sm:n X funktion avulla, ts.

$$E[(g(X,Y) - E(g(X,Y) \mid X))^2] \leq E[(g(X,Y) - h(X))^2]$$

valitaan funktio $h : \mathbb{R} \to \mathbb{R}$ miten tahansa.

- Todistus on harjoitustehtävä.
- Jos tehtävänä on ennustaa sm:n Y arvo jollakin sm:n X arvon funktiolla, niin keskineliövirheen mielessä paras mahdollinen ennuste on $m(x)$, jossa m on regressiofunktio $m(x) = E[Y \mid X = x]$.

Odotusarvon laskeminen iteroituna odotusarvona

Lause

Odotusarvo voidaan laskea iteroituna odotusarvona, eli

$$Eg(X,Y) = EE(g(X,Y) \mid X),$$

mikäli odotusarvo $Eg(X,Y)$ on olemassa laajennettuna reaalilukuna.

Todistus Esitetään perustelu diskreetissä tapauksessa.

$$\begin{aligned} Eg(X,Y) &= \sum_{x,y} g(x,y)\, f_{X,Y}(x,y) = \sum_{x,y} g(x,y)\, f_X(x)\, f_{Y|X}(y \mid x) \\ &= \sum_x f_X(x) \sum_y g(x,y)\, f_{Y|X}(y \mid x) \\ &= EE(g(X,Y) \mid X). \end{aligned}$$

Esimerkki iteroidusta odotusarvosta

- Tarkastellaan yhteisjakaumaa

$$X \mid Y \sim \mathrm{Bin}(Y, \theta),$$
$$Y \sim \mathrm{Poi}(\lambda)$$

jossa $\lambda > 0$ ja $0 < \theta < 1$.

- Nyt

$$E(X \mid Y) = Y\,\theta,$$

joten

$$EX = EE(X \mid Y) = E(Y\,\theta) = \theta\,EY = \theta\lambda.$$

Ehdollinen varianssi satunnaismuuttujana

- Kuten ehdollinen odotusarvo, myös ehdollinen varianssi voidaan laskea ehtona satunnaismuuttuja X (eikä ehdolla sen arvo $X = x$).
- Ensin määritellään funktio

$$v(x) = \operatorname{var}(g(X, Y) \mid X = x),$$

ja määritelmää jatketaan koko reaaliakselille sopimalla, että $v(x) = 0$ niillä argumenteilla, joilla ehdollinen jakauma $Y \mid (X = x)$ ei ole luonnostaan määritelty.

- Tämän jälkeen määritellään, että

$$\operatorname{var}(g(X, Y) \mid X) = v(X).$$

Kaava varianssille

Lause

Sm:n varianssi on yhtä kuin sen ehdollisen varianssin odotusarvon sekä ehdollisen odotusarvon varianssin summa, eli

$$\operatorname{var}(g(X,Y)) = E \operatorname{var}(g(X,Y) \mid X) + \operatorname{var} E(g(X,Y) \mid X). \quad (8)$$

Todistus: Harjoitustehtävä.

8.5 Yhteisjakauman määrittely hierarkkisesti

- Tilastollisia malleja spesifioidaan usein kertomalla, mikä on yhden sm:n reunajakauma ja toisen ehdollinen jakauma.
- Tällöin yhteisjakauma saadaan kertolaskusäännöllä, esim.

$$f_{X,Y}(x, y) = f_X(x)\, f_{Y|X}(y \mid x).$$

- Tällöin voidaan puhua hierarkkisesta mallista.

Esimerkki 1 (kaksi diskreettiä sm:a)

- Hyönteisen munimien munien lkm $Y \sim \mathrm{Poi}(\lambda)$, jossa $\lambda > 0$.
- Kukin munista kehittyy toukaksi toisistaan riippumatta tn:llä $0 < \theta < 1$.
- Olkoon X toukaksi kehittyvien munien lkm. Tällöin

$$X \mid (Y = y) \sim \mathrm{Bin}(y, \theta).$$

- Yhteisjakauma on diskreetti, ja sen yptnf on

$$f_{X,Y}(x, y) = f_Y(y)\, f_{X|Y}(x \mid y) = \mathrm{e}^{-\lambda}\, \frac{\lambda^y}{y!} \binom{y}{x} \theta^x (1-\theta)^{y-x},$$

jossa x ja y ovat kokonaislukuja $0, 1, 2, \ldots$, ja $x \leq y$.

Esimerkki 1 jatkuu

- Tämä malli voidaan määritellä myös sanomalla, että

$$X \mid Y \sim \mathrm{Bin}(Y, \theta),$$
$$Y \sim \mathrm{Poi}(\lambda)$$

  jossa $\lambda > 0$ ja $0 < \theta < 1$ ovat vakioita.
- Voidaan esim. kysyä, mikä on X:n reunajakauma. Sen ptnf voidaan esittää summana

$$f_X(x) = \sum_y f_{X,Y}(x, y).$$

- Joidenkin laskujen jälkeen havaitaan, että reunajakauma on $X \sim \mathrm{Poi}(\lambda\theta)$. (Edellä laskettiin EX iteroidun odotusarvon kaavalla, ja silloin saatiin $EX = \lambda\,\theta$.)

Esimerkki 2 (jatkuvan ja diskreetin sm:n yhteisjakauma)

- Olkoon sm:lla Θ betajakauma $\mathrm{Be}(\alpha, \beta)$, jossa $\alpha, \beta > 0$ ovat vakioita. Kolikkoa heitetään n kertaa, ja lasketaan kuinka monta kertaa saadaan klaava, kun klaavan todennäköisyys on Θ.
- Ehdolla $\Theta = \theta$ klaavojen lukumäärällä X on jakauma $\mathrm{Bin}(n, \theta)$.
- Yhteisjakauma voidaan esittää funktiolla

$$f_{\Theta,X}(\theta, x) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \binom{n}{x} \theta^x (1-\theta)^{n-x},$$

jossa $0 < \theta < 1$ ja $x = 0, 1, \ldots, n$.

Esimerkki 2 jatkuu

- Tämä malli voitaisiin spesifioida myös sanomalla, että

$$X \mid \Theta \sim \mathrm{Bin}(n, \Theta),$$
$$\Theta \sim \mathrm{Be}(\alpha, \beta)$$

  jossa $\alpha, \beta > 0$ ja $n \geq 0$ ovat vakioita.
- Mikä on sm:n Θ jakauma, kun havaitaan, että $X = x$?
- Kysymys saadaan ratkaistua tarkastelemalla hetken yhteisjakauman esitystä muuttujan θ funktiona, minkä jälkeen on selvää, että

$$\Theta \mid (X = x) \sim \mathrm{Be}(\alpha + x, \beta + n - x).$$

- Tämä on esimerkki Bayes-päättelystä.

Esimerkki 3 (kahden jatkuvan muuttujan yhteisjakauma)

- Useasta ohjelmistoista löytyy satunnaislukugeneraattori normaalijakaumalle.
- Tarkastellaan seuraavaa algoritmia, jossa $\sigma_X > 0$, μ_X ja $\sigma_Z > 0$ ovat annettuja lukuja ja m on jokin funktio, joka palauttaa reaaliluvun reaalilukuargumentilla.
  1. Simuloi $X \sim N(\mu_X, \sigma_X^2)$.
  2. Simuloi $Z \sim N(0, \sigma_Z^2)$.
  3. Aseta $Y = m(X) + Z$.
- Kuvaile sm:ien X ja Y yhteisjakauma hierarkkisesti.
- Kun satunnaislukugeneraattoria kutsutaan monta kertaa, niin eri kerroilla palautettavia lukuja voidaan pitää keskenään riippumattomien satunnaismuuttujien arvoina, sillä todelliset satunnaislukugeneraattorit toimivat tällä tavoin.

Esimerkki 3: yhteisjakauman kuvailu hierarkkisella mallilla

- Yhteisjakauma voidaan esittää kaavoilla

$$Y \mid X \sim N(m(X), \sigma_Z^2)$$
$$X \sim N(\mu_X, \sigma_X^2).$$

- Ehdollinen jakauma $[Y \mid X = x]$ nähdään siitä, että simuloinnissa $X \perp\!\!\!\perp Z$, joten X:n arvolla x ehdollistaminen ei muuta Z:n jakaumaa.
- Kyseisessä ehdollisessa jakaumassa sm:n X arvo on vakio x, ja lopuksi vakion x:n tilalle kirjoitetaan satunnaismuuttuja X.

Esimerkki 3: yhteistiheysfunktio

- Koska X:n jakauma on jatkuva, ja Y:n ehdollinen jakauma ehdolla $X = x$ on jatkuva kaikilla x, on myös yhteisjakauma jatkuva.
- Ytf on

$$
\begin{aligned}
f_{X,Y}(x,y) &= f_X(x)\, f_{Y|X}(y \mid x) \\
&= \frac{1}{\sigma_X \sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2}\right) \\
&\quad \frac{1}{\sigma_Z \sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(y-m(x))^2}{\sigma_Z^2}\right)
\end{aligned}
$$

Esimerkki 4: jatkoa esimerkille 3

- Valitaan regressiofunktiolle m edellisessä esimerkissä lineaarinen muoto, $m(x) = \alpha + \beta\,(x - \mu_X)$.
- Tällöin $m(X)$ antaa keskineliövirheen mielessä parhaan ennusteen satunnaismuuttujan Y arvolle.
- Koska $m(X)$ on lineaarinen, sen täytyy myös olla keskineliövirheen mielessä paras lineaarinen ennuste.

Esimerkki 4 jatkuu

- Jakson 7.6 mukaan täytyy olla

$$m(x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X),$$

  jossa μ_Y ja $\sigma_Y > 0$ (oletus) ovat sm:n Y odotusarvo ja keskihajonta ja $-1 < \rho < 1$ (oletus) on X:n ja Y:n korrelaatiokerroin.
- Lisäksi $\sigma_Z^2 = \sigma_Y^2 (1 - \rho^2)$.
- Vektorille (X, Y) syntyy ns. kaksiulotteinen normaalijakauma, jonka tiheysfunktion lauseke löytyy luentomonisteesta.
- Moniulotteista normaalijakaumaa käsitellään tarkemmin kappaleessa 10.