Магадлал Статистик

хичээлийн лекц

Г.Махгал

© 2019 – 2021 Г.Махгал

www.galaa.mn

2021/1/28

Сэргээн санах зүйлс

1. МУИС-ийн ерөнхий суурийн ”Математик” хичээл
2. Ерөнхий боловсролын сургуульд үзсэн магадлал ба статистикийн сэдвүүд
   - 2.1 http://econtent.edu.mn/ дээр байрлуулсан ЕБС-ийн сурах бичгүүд дээрх магадлал, статистикийн бүлгүүд
   - 2.2 ”Ерөнхий боловсролын сургуулийн XII ангийн математикийн суралцахуйн удирдамж” http://www.mier.mn/wp-content/uploads/2018/11/Math_XII-angi.pdf 148-р хуудас
3. Хэрэв сонгон судалсан бол МУИС-ийн ерөнхий суурийн ”Магадлал Статистикийн Удиртгал” хичээл

Агуулга

Агуулга

Хичээлийн веб хуудас

хаяг www.magadlal.com/courses/8.html

агуулга хичээлийн агуулга, сэдэвчилсэн төлөвлөгөө, лекцийн эмхтгэл, видео лекц, семинарын бодлого, дүгнэх журам

ФОТО ЗУРАГ БИЧИЖ ТЭМДЭГЛЭХ ЧАДВАРТ СӨРӨГ НӨЛӨӨТЭЙ

Лекц I

Санамсаргүй хувьсагч, түүний тархалт

Санамсаргүй хувьсагч, түүний тархалт сэдвийн агуулга

Хэсэг 1

Магадлалын огторгуйн математик загвар

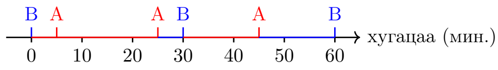
Хялбар жишээ

Жишээ

Сүлжээгээр дамжин серверт ирэх хүсэлт хугацааны аль ч эгшинд ижил боломжтойгоор ирнэ. Хүсэлт хүлээн авах портыг 17:00 ба 18:00 цагийн хооронд нээх ба хүсэлтийг боловсруулж хариу өгөх зорилготой A болон B гэсэн хоёр янзын програм ажилладаг.

- A програм 17:05, 17:25, 17:45
- B програм 17:00, 17:30, 18:00

цагт ажиллахаар тохируулжээ. Системийн админ нийт хүсэлтийн 2/3 хувийг A програм боловсруулж байгааг ажиглажээ. Энэ зөв үү?



Зураг: Жишээ бодлогын зураглал

Санамсаргүй үзэгдэл

Тодорхойлолт

Туршилтаар илэрч болох бүх үр дүнгийн олонлогийг *эгэл үзэгдлийн огторгуй* гээд Ω гэж тэмдэглэнэ.

Жишээний хувьд

$$\omega = \{17{:}00 \text{ цагаас хойших хугацааны аль нэг эгшинд хүсэлт ирэх}\}$$

явдал нь үр дүн юм. Иймд $\Omega = [0, 60]$[1] гэж үзэж болно.

Тодорхойлолт

Эгэл үзэгдлийн огторгуйн дэд олонлогийг *үзэгдэл* гэнэ.

$\{\text{хүсэлтийг B програм боловсруулах}\} = \{0\} \cup (25, 30] \cup (45, 60]$
$\{\text{дор хаяж 8 минутын дараа боловсруулах}\} = (5, 17] \cup (30, 37] \cup (45, 52]$

[1] хугацаа минутаар хэмжье

Үзэгдлийн магадлал

Тодорхойлолт

$\mathcal{F}$ нь үзэгдлүүдийн олонлог байг. Тэгвэл

$$P : \mathcal{F} \to [0, 1]$$

функцийг *магадлал* гэнэ.

P функцийг Ω олонлог

- төгсгөлөг элементтэй бол үр дүнг тоолох
- (хэсэгчилсэн) тасралтгүй (төгсгөлгүй) олон элементтэй бол геометр хэмжээс ашиглах
- статистик өгөгдлөөс тогтож буй бол үр дүнгийн давтамж тооцох

гэх мэтчилэн янз бүрээр тодорхойлдог.

Жишээний хувьд P функцийг дараах байдлаар тодорхойлж болно.

- Хэрэв $E = (a, b) \subseteq \Omega$ бол
$$P(E) = \frac{b - a}{60}$$
- Хэрэв $E = (a_1, b_1) \cup \ldots \cup (a_k, b_k) \subseteq \Omega$ ба (a_i, b_i) интервалууд үл огтлолцох бол
$$P(E) = \sum_{i=1}^{k} \frac{b_i - a_i}{60} = \frac{E \text{ хэсгийн нийт урт}}{60}$$

Нэг үр дүнгийн магадлал

Дурын үр дүн $\omega \in \Omega$ бүрийн хувьд

$$P(\{\omega\}) = \frac{(\omega, \omega) \text{ хэсгийн нийт урт}}{60} = 0$$

байна. Гэвч үргэлж $P(\Omega) = 1$[2] байдаг.
Учир нь Ω тоологдом биш үед

$$P(\Omega) = \sum_{\omega} P(\omega)$$

адилтгал биелдэггүй.
Иймд жишээний хувьд $P([a, b)) = P((a, b)) = P((a, b]) = P([a, b])$ байна.

[2] магадлалын аксиом

Магадлалын огторгуй

Тодорхойлолт

$(\Omega, \mathcal{F}, P)$ бүлийг *магадлалын огторгуй* гэнэ.

Хэсэг 2

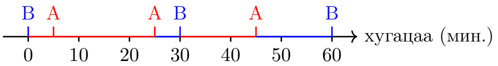
Санамсаргүй хувьсагч

Санамсаргүй хувьсагч

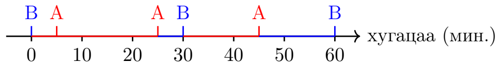
Тодорхойлолт

Эгэл үзэгдлийн огторгуй дээр тодорхойлсон, бодит утгатай $X : \Omega \to \mathbb{R}$ функцийг *санамсаргүй хувьсагч* гэнэ.

Жишээний хувьд X нь хүсэлтийг боловсруулах хүртэлх хугацаа бол



$$X(\omega) = \begin{cases} 0, & \omega = 0 \\ 5-\omega, & 0 < \omega \le 5 \\ 25-\omega, & 5 < \omega \le 25 \\ 30-\omega, & 25 < \omega \le 30 \\ 45-\omega, & 30 < \omega \le 45 \\ 60-\omega, & 45 < \omega \le 60 \end{cases}$$



- Y нь серверт хүсэлт ирэх эгшин бол

$$Y(\omega) = \omega$$

- Хүсэлтийг A програмаар боловсруулах гэсэн үзэгдлийг A гэвэл

$$A = (0, 25] \cup (30, 45]$$

бөгөөд

$$I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}$$

байдлаар *индикатор хувьсагч* тодорхойлж болно.

Санамсаргүй хувьсагчийн ангилал

- *тасралтгүй*: төгсгөлгүй олон утга авна.
  $X \in [0, 20]$ ба $Y \in [0, 60]$
- *дискрет*: төгсгөлөг эсвэл тоологдом олонлогоос утгаа авна.
  $I_A \in \{0, 1\}$

Хэсэг 3

Санамсаргүй хувьсагчийн нягт

Дискрет санамсаргүй хувьсагчийн нягт

Тодорхойлолт

X санамсаргүй хувьсагч дискрет буюу

$$X \in \{x_1, x_2, \dots\}$$

үед

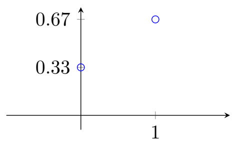
$$P(X = x_i)$$

магадлалуудыг тус *дискрет санамсаргүй хувьсагчийн нягт* гэнэ.

I_A дискрет санамсаргүй хувьсагчийн *нягт*

$$P(I_A = 1) = P(\omega \in A) = \frac{(25-0)+(45-30)}{60} = \frac{2}{3}, \quad P(I_A = 0) = \frac{1}{3}$$

байна.



Зураг: I_A дискрет санамсаргүй хувьсагчийн нягт

Дурын X *дискрет санамсаргүй хувьсагчийн нягтын функц* дараах хэлбэртэй байна.

$$f_X(x) = \begin{cases} P(X = x_i), & x = x_i \\ 0, & \text{бусад} \end{cases}$$

Тасралтгүй санамсаргүй хувьсагчийн нягт

X санамсаргүй хувьсагч дискрет буюу

$$X \in \{x_1, x_2, \dots\}$$

үед

$$P(a \le X \le b) = \sum_{x_i : a \le x_i \le b} f_X(x_i) = \sum_{x_i : a \le x_i \le b} P(X = x_i)$$

буюу үзэгдлийн магадлал нягтын нийлбэртэй тэнцүү байна.

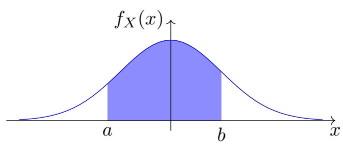
Тодорхойлолт

X нь тасралтгүй санамсаргүй хувьсагч бөгөөд хэрэв $a \le b$ бүрийн хувьд

$$P(a \le X \le b) = \int_a^b f_X(x)dx$$

байх хэсэгчилсэн тасралтгүй функц $f_X : \mathbb{R} \to [0, \infty)$ оршин байвал $f_X(x)$ функцийг X хувьсагчийн *нягтын функц* гэнэ.

$$P(a \le X \le b) = \int_a^b f_X(x)dx = S_{\text{муруй шугаман трапец}}$$



Зураг: X хувьсагчийн нягтын муруй ба $a \le X \le b$ үзэгдлийн магадлал

Иймд

1. дурын $x \in \mathbb{R}$ бүрийн хувьд $f_X(x) \ge 0$
2. $\displaystyle\int_{-\infty}^{\infty} f_X(x) = 1$

байх нь гарцаагүй юм.

Дискрет ба тасралтгүй хувьсагчдын нягтын ялгаа

Санамж

Ерөнхийдөө тасралтгүй санамсаргүй хувьсагчийн хувьд

$$f_X(x) \neq P(X = x)$$

юм.

Үүнээс

$$P(X = a) = \int_a^a f_X(x)dx = 0$$

бас

$$P(a < X < b) = P(a \leq X \leq b)$$

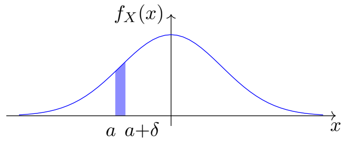
юм.

Тасралтгүй хувьсагчийн $P(X = a)$ магадлал

Хэрэв $f_X(x)$ функц a цэг дээр тасралтгүй бол бага утгатай δ тогтмолын хувьд

$$P(a < X < a + \delta) = \int_a^{a+\delta} f_X(x)dx \approx \delta f_X(a)$$

байна.



Зураг: X санамсаргүй хувьсагчийн нягтын муруй ба $a < X < a + \delta$ үзэгдлийн магадлал

Хэсэг 4

Жигд тархалт

Жигд тархалт

Тодорхойлолт

Нягт нь тогтмол байх тархалтыг *жигд тархалт* гэнэ.

(a) Тасралгүй жигд тархалт (b) Дискрет жигд тархалт

Зураг: Жигд тархалтын нягтын ерөнхий хэлбэр

Жишээний хувьд $Y \sim U(0, 60)$ буюу Y хувьсагч $(0, 60)$ завсар дээр жигд тархалттай юм.

Лекц II

Дундаж болон дундаж квадрат хазайлт

Дундаж болон дундаж квадрат хазайлт сэдвийн агуулга

Хэсэг 1

Пуассоны тархалт

Пуассоны тархалт

Ямар нэг туршилт явуулж байгаа гэж үзье. Туршилтаас гарах үр дүнгийн хувьд ямар нэг үзэгдлийг онцгойлон "амжилт" хэмээн аваад A гэж тэмдэглэе. Ийнхүү үр дүнг нь хоёр ангилсан туршилтыг *Бернуллийн туршилт* гэдэг. $P(A) = p$ бас туршилтыг өөр хоорондоо хамааралгүй байдлаар n удаа давтсан гэе. Тэгвэл үүнтэй холбогдуулан практикт өргөн тохиолдох янз бүрийн санамсаргүй хувьсагч авч үзэж болно.

Нэгж хугацаанд төгсгөлгүй олон "амжилт" гарах боломжтой туршилтын "амжилт"-ын тоо түүний магадлалын тархалт

$$f_X(x) = \frac{\lambda^x}{x!}e^{-\lambda}, \quad x \in \{0, 1, 2, \dots\}$$

бөгөөд үүнийг *Пуассоны*[1] *тархалт* гээд $\text{Pois}(\lambda)$ гэж тэмдэглэнэ. Энд λ нь нэгж хугацаанд гарах амжилтын дундаж тоо юм.

[1] Poisson

Пуассоны тархалтын тусламжтай шийдэх бодлого

Жишээ

Онлайнаар захиалга хүлээн авдаг түргэн хоолны газарт 09:00-өөс 10:00 цагийн хооронд дунджаар 1.5 захиалга ирдэг бол энэ хугацаанд ганц ч захиалга ирэхгүй байх магадлал ямар байх вэ?

Энд санамсаргүй хувьсагч X нь нэг цагт хүлээн авах захиалгын тоо буюу Пуассоны тархалттай санамсаргүй хувьсагч байна. Бодлогын нөхцөл ёсоор $E(X) = \lambda = 1.5$ бөгөөд "ганц ч захиалга хүлээн авахгүй байх" гэсэн үзэгдэлд X хувьсагчийн 0 утга харгалзах тул

$$P(\text{ганц ч захиалга хүлээн авахгүй байх}) = P(X=0) = f_X(0) = \frac{1.5^0}{0!}e^{-1.5} \approx 0.223$$

байна.

Пуассоны тархалтын нягтын гаргалгаа

$\lambda_n = np$, $\lim_{n\to\infty} \lambda_n = \lambda$ гэе.

$$\begin{aligned} f_X(x) &= P(n \text{ туршилтад } x \text{ удаа амжилт илрэх}) \\ &= C_n^x p^x (1-p)^{n-x} \qquad \text{бином тархалтын нягт} \\ &= \frac{n(n-1)\dots(n-x+1)}{x!}\left(\frac{\lambda_n}{n}\right)^x\left(1-\frac{\lambda_n}{n}\right)^{n-x} \\ &= \frac{\lambda_n^x}{x!}\left(1+\frac{-\lambda_n}{n}\right)^n 1\left(1-\frac{1}{n}\right)\dots\left(1-\frac{x-1}{n}\right)\left(1-\frac{\lambda_n}{n}\right)^{-x} \end{aligned}$$

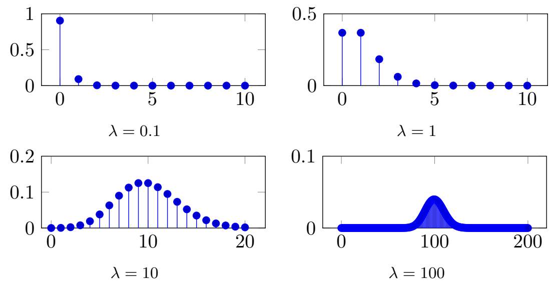
Одоо хязгаарт шилжвэл $\lim_{n\to\infty} f_X(x) = \frac{\lambda^x}{x!}e^{-\lambda}$ болно.

Мөн эндээс бином тархалтын нягтыг n их, p бага үед

$$f_X(x) = P(X=x) \approx \frac{\lambda^x}{x!}e^{-\lambda}, \quad \lambda = np$$

гэж ойролцоо бодохыг харж болно. Үүнийг *Пуассоны томьёо* гэдэг.

Пуассоны тархалт, параметрийн янз бүрийн утгад



Зураг: Пуассоны тархалт, параметрийн янз бүрийн утгад

Пуассоны тархалтын параметр λ нь тодорхой нэг үзэгдлийн нэгж хугацаанд илрэх тооны дундажтай тэнцүү.

Хэсэг 2

Хамааралгүйн чанар

Хамааралгүйн чанар

Тодорхойлолт

$$P(AB) = P(A)P(B)$$

бол A болон B үзэгдлүүдийг *хамааралгүй* гэнэ.

Тодорхойлолт

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad \forall (x,y) \in \mathbb{R}^2$$

бол X болон Y санамсаргүй хувьсагчдыг *хамааралгүй* гэнэ.

Дискрет үед $P(X=x_i, Y=y_j) = P(X=x_i)P(Y=y_j)$ байна.

Хамаарлын талаар дараа дэлгэрэнгүй авч үзнэ.

Хэсэг 3

Дундаж

Санамсаргүй хувьсагчийн математик дундаж

Тодорхойлолт

Санамсаргүй хувьсагчийн утгуудыг тэдгээрийн магадлалаар жинлэсэн дунджийг *математик дундаж*[2] гэнэ.

- Дискрет санамсаргүй хувьсагч

$$E(X) = \sum_{x_i} x_i P(X=x_i)$$

- Тасралтгүй санамсаргүй хувьсагч

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx$$

[2] expectation

Жишээ

Пуассоны тархалттай санамсаргүй хувьсагчийн математик дунджийг ол.

$$\begin{aligned} E(X) &= \sum_{x_i} x_i P(X=x_i) \\ &= \sum_{x=0}^{\infty} x\frac{\lambda^x}{x!}e^{-\lambda} \\ &= \lambda e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x}{x!} = \lambda e^{-\lambda}e^{\lambda} \\ &= \lambda \end{aligned}$$

Дунджийн чанар

Чанар

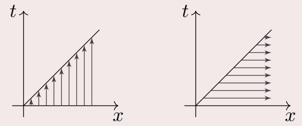
X нь $P(X \geq 0) = 1$ байх тасралтгүй санамсаргүй хувьсагч бөгөөд $E(X) < \infty$ байг. Тэгвэл дараах чанар хүчинтэй.

$$E(X) = \int_0^\infty P(X \geq t)dt$$

Баталгаа

Интегралчлах эрэмбэ дараа солих замаар шууд батална.

$$E(X) = \int_0^\infty x f_X(x)dx = \int_0^\infty \left\{\int_0^x dt\right\} f_X(x)dx$$
$$= \int_0^\infty \int_0^x f_X(x)dtdx$$



$$= \int_0^\infty \int_t^\infty f_X(x)dxdt = \int_0^\infty P(X \geq t)dt$$

Дунджийн чанар

Чанар (Ухамсаргүй статистикчийн хууль[3])

$g : \mathbb{R} \to \mathbb{R}$ функцийн хувьд

$$E(g(X)) = \int_{-\infty}^\infty g(x) f_X(x)dx$$

байна.

Үүний баталгаар *тархалтын функц* үзсэний дараа хийнэ.

[3] The law of the unconscious statistician

Интегралын шугаман чанараас үүдэлтэйгээр ухамсаргүй статистикчийн хуулиас дараах чанарууд шууд гарна.

- $E(aX + b) = aE(X) + b$
- $E[g(X) + h(X)] = E[g(X)] + E[h(X)]$

Санамсаргүй хувьсагчид хамааралгүй бол дараах чанар биелнэ.

$$E(XY) = E(X)E(Y)$$

Момент үүсгэгч функц

$$M_X(t) = E\left[e^{tX}\right], \quad t \in \mathbb{R}$$

Жишээ

Пуассоны тархалттай санамсаргүй хувьсагчийн момент үүсгэгч функцийг ол.

$$M_X(t) = \sum_{x=0}^\infty e^{tx} \frac{\lambda^x}{x!} e^{-\lambda}$$
$$= e^{-\lambda} \sum_{x=0}^\infty \frac{(\lambda e^t)^x}{x!} = e^{-\lambda} e^{\lambda e^t}$$
$$= \exp(\lambda(e^t - 1))$$

Жишээ

Стандарт хэвийн тархалттай санамсаргүй хувьсагчийн момент үүсгэгч функцийг ол.

Стандарт хэвийн тархалтын нягтын функц $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ бас $\int\limits_{-\infty}^{\infty} e^{-x^2/2}dx = \sqrt{2\pi}$ байхыг анхаарвал

$$M_X(t) = E\left[e^{tX}\right] = \int_{-\infty}^\infty e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}dx$$
$$= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-(x-t)^2/2}dx = e^{t^2/2}$$

гэж олдоно.

Момент үүсгэгч функцийн хэрэглээ

Момент үүсгэгч функцийг математик дундаж зэрэг момент[4] олоход ашиглана.

$$M'_X(t)\big|_{t=0} = E[Xe^{tX}]\big|_{t=0} = E(X)$$

Жишээ

Пуассоны тархалттай санамсаргүй хувьсагчийн математик дунджийг түүний үүсгэгч функцийн тусламжтай ол.

$M'_X(t) = (\exp(\lambda(e^t - 1)))' = e^{\lambda(e^t-1)}\lambda e^t$ бөгөөд $t = 0$ үед

$$E(X) = M'_X(t)\big|_{t=0} = \lambda$$

болно.

[4] $\alpha_k = E[X^k]$

Хэсэг 4

Геометр тархалт

Геометр тархалт

"Амжилт" илэртэл туршилт явуулахад тохиолдох бүтэлгүйтлийн тоо болон тэдгээр тоонуудын

$$f(x) = (1-p)^x p, \quad x \in \{0, 1, 2, \dots\}, \quad 0 \le p \le 1$$

магадлалын тархалтыг *геометр тархалт* гээд $\mathrm{Geom}(p)$ гэж тэмдэглэнэ.

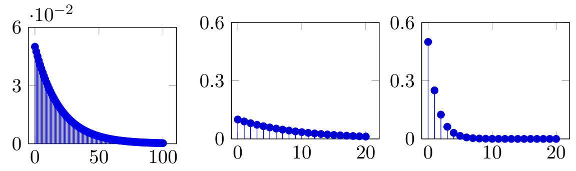
Санамж

Үүнээс гадна санамсаргүй хувьсагчийг амжилт илэртэл явуулах туршилтын тоо буюу $\{1, 2, \dots\}$ утгатайгаар авах явдал бий. Өөрөөр хэлбэл геометр тархалтыг хоёр янзаар авч үздэг.

Геометр тархалтын нягтын гаргалгаа

$$\begin{aligned} f_X(x) &= P(\text{анхны амжилт } x+1 \text{ дүгээр туршилт дээр илрэх}) \\ &= P(\underbrace{\overline{A}\dots\overline{A}}_{x} A) \\ &= \underbrace{P(\overline{A})\dots P(\overline{A})}_{x} P(A) \\ &= (1-p)^x p \end{aligned}$$

Геометр тархалт, параметрийн янз бүрийн утгад



$p = 0.05$ $p = 0.1$ $p = 0.5$

Зураг: Геометр тархалт, параметрийн янз бүрийн утгад

Хэсэг 5

Дундаж квадрат хазайлт

Дундаж квадрат хазайлт буюу дисперс

Тодорхойлолт (Дунджаасаа хазайх хазайлт)

$$X - E(X)$$

X санамсаргүй хувьсагч дунджаасаа дунджаар хэр зэрэг хазайх вэ?

$E[X - E(X)] = E(X) - E[E(X)] = E(X) - E(X) = 0$

$E|X - E(X)|$ санамсаргүй утгатай ялгаврын тэмдгийг яаж тооцох вэ?

Тодорхойлолт (Дундаж квадрат хазайлт)

$$D(X) = E[X - E(X)]^2$$

Дундаж квадрат хазайлтын зарим чанар

1. $D(X) = E[X - E(X)]^2 = E[X^2 - 2XE(X) + [E(X)]^2] = E(X^2) - 2[E(X)]^2 + [E(X)]^2 = E(X^2) - [E(X)]^2$
2. $D(a + bX) = b^2 D(X)$
3. X болон Y хамааралгүй үед $D(X + Y) = D(X) + D(Y)$

Стандарт хазайлт

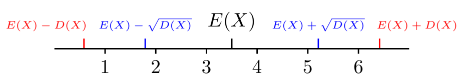
$$\sqrt{D(X)} = \sqrt{E[X - E(X)]^2}$$

Жишээ

$f_X(x) = \frac{1}{6}$, $x \in \{1, 2, \dots, 6\}$ дискрет жигд тархалтын стандарт хазайлтыг ол.

$E(X) = \sum_{x=1}^{6} x\frac{1}{6} = 3.5$ ба

$$D(X) = \sum_{x=1}^{6} (x - 3.5)^2 \frac{1}{6} \approx 2.917 \quad \sqrt{D(X)} \approx 1.708$$



Зураг: Жишээ бодлогын бодолтын үр дүн

Лекц III

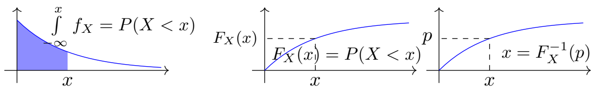
Тархалтын функц

Тархалтын функц сэдвийн агуулга

Хэсэг 1

Тархалтын функц

Санамсаргүй хувьсагчийн тархалт, түүнийг илэрхийлэх функцууд



(a) Нягтын функц (b) Тархалтын функц (c) Квантил

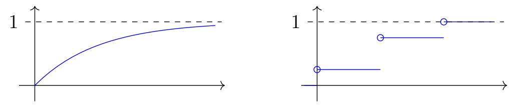
Зураг: Тасралтгүй санамсаргүй хувьсагчийн тархалт, илтгэгч тархалтын тохиодолд

(a) Нягт (b) Тархалтын функц

Зураг: Дискрет санамсаргүй хувьсагчийн тархалт, бином тархалтын тохиолдолд

Тархалтын функцийн чанарууд

1. Зөөн тасралтгүй
2. Үл буурах
3. $\lim_{x\to-\infty} F(x) = 0, \ \lim_{x\to\infty} F(x) = 1$

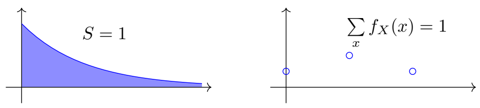


Тасралтгүй санамсаргүй хувьсагч Дискрет санамсаргүй хувьсагч

Зураг: Тархалтын функц

Нягтын функцийн чанарууд

1. $f_X(x) \geq 0$
2. $\int_{-\infty}^{+\infty} f_X(x)dx = 1$



тасралтгүй үед дискрет үед

Зураг: Нягтын нийлбэр

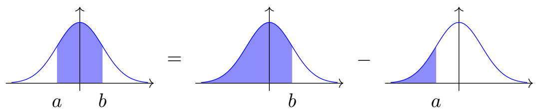
Тасралтгүй санамсаргүй хувьсагчийн тархалтын функц болон нягтын функцийн холбоо, үзэгдлийн магадлал

- Тархалтын функц ба нягтын функцийн уялдаа холбоо

$$F_X(x) = P(X<x) = \int_{-\infty}^{x} f_X(x)dx \quad \text{буюу} \quad F'_X(x) = f_X(x)$$

- Магадлал олоход тархалтын хууль ашиглах нь

$$P(a<X<b) = \int_a^b f_X(x)dx = F_X(b) - F_X(a)$$



Зураг: Тасралтгүй санамсаргүй хувьсагчийн $P(a<X<b)$ магадлал

Хэсэг 2

Дунджийн чанар

Ухамсаргүй статистикчийн хуулийн баталгаа

Сэргээн санах нь

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

$g(\cdot)$ нь дифференциалчлагдах, урвуу нь монотон байх функц гэе. $Y = g(X)$ санамсаргүй хувьсагч авбал

$$E(g(X)) = E(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy$$

болох ба сүүлийн интегралд $y = g(x)$ орлуулга хийж хувьсагч сольё.

- $y = \boxed{g(x)}$
- dy:

$$dy = dg(x) = dg(g^{-1}(y)) = g'(g^{-1}(y))d(g^{-1}(y))$$
$$\frac{d}{dy}(g^{-1}(y)) = \frac{1}{g'(g^{-1}(y))} \quad \Leftrightarrow \quad dx = \boxed{\frac{1}{g'(g^{-1}(y))}dy}$$

- $f_Y(y)$:

$$F_Y(y) = P(Y < y) = P(g(X) < y) = P(X < g^{-1}(y)) = F_X(g^{-1}(y))$$
$$f_Y(y) = F'_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}(g^{-1}(y)) = \boxed{f_X(x)\frac{1}{g'(g^{-1}(y))}}$$

- $$E(g(X)) = \int_{-\infty}^{\infty} y f_Y(y)dy = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

Хэсэг 3

Илтгэгч тархалт

Илтгэгч тархалт

Сэргээн санах нь

Ямар нэг туршилт авч үзье. Туршилтын үр дүнгээс аль нэг A үзэгдлийг онцгойлон "амжилт" хэмээн авна. $P(A) = p$ бас туршилтыг өөр хоорондоо хамааралгүй байдлаар n удаа давтсан гэе. Тэгвэл үүнтэй холбогдуулан практикт өргөн тохиолдох янз бүрийн санамсаргүй хувьсагч зохиож болно.

"Амжилт" илэртэл хүлээх хугацаа гэсэн санамсаргүй хувьсагчийн магадлалын тархалтыг *илтгэгч тархалт* гэнэ.

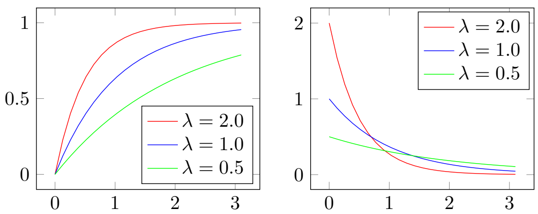
Илтгэгч тархалтын функцийн гаргалгаа

X нь амжилт илрэх хүртэлх хугацаа, $n \to \infty$, $p \to 0$, $\lambda = \lim np$

$$\begin{aligned} F_X(x) &= P(X < x) = P(x \text{ хугацааны дотор } A \text{ үзэгдэл явагдах}) \\ &= 1 - P(X \geq x) \\ &= 1 - P(x \text{ хугацааны дотор } A \text{ үзэгдэл явагдахгүй байх}) \\ &= 1 - P(\mu_n = 0, n \to \infty, p \to 0) \\ &= 1 - \frac{(\lambda x)^0}{0!}e^{-\lambda x} \quad \text{Пуассоны тархалт} \\ &= 1 - e^{-\lambda x}, \quad (x > 0) \end{aligned}$$

Илтгэгч тархалт, параметрийн янз бүрийн утгад

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}, \quad f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$



(a) Тархалтын функц (b) Нягтын функц

Зураг: Илтгэгч тархалт, параметрийн янз бүрийн утгад

Илтгэгч тархалтын зарим чанар

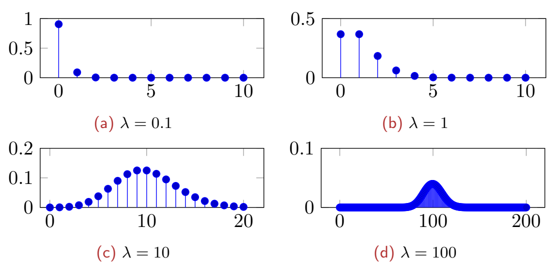
1. Геометр тархалтын тасралтгүйн аналог
2. $X_i \sim \text{Exp}(\lambda_i)$[1] $(i = 1, \ldots, n)$ ба *хамтдаа хамааралгүй*[2] бол
$$\min\{X_1, \ldots, X_n\} \sim \text{Exp}(\lambda_1 + \ldots + \lambda_n)$$
3. $\forall x, y \geq 0$ бүрийн хувьд
$$P(X \geq x + y | X \geq y) = P(X \geq x)$$[3]

[1] илтгэгч тархалт
[2] энэ талаар хожим үзнэ
[3] илтгэгч тархалтын санамжгүй байдал; энэ талаар хожим үзнэ

Хэсэг 4

Хэвийн тархалт

Пуассоны тархалт параметрийн их утгад



Зураг: Пуассоны тархалтын параметр ба нягтын хэлбэр

λ буюу тодорхой нэг үзэгдлийн нэгж хугацаанд илрэх дундаж тоо өсөхөд тархалтын нягт "хонх" хэлбэртэй болж байна. Тэгвэл энэ "хонх" хэлбэртэй нягтын илэрхийлэл ямар байх вэ?

"Хонх" хэлбэртэй нягтын илэрхийлэл

$$P(X=x)=\frac{\lambda^x}{x!}e^{-\lambda},\quad x\in\{0,1,2,\dots\}$$

$E(X)=\lambda$ учраас λ параметрийн утга ихсэхэд өндөр магадлалтай утгууд нь λ орчимд байх тул $x=\lambda(1+\delta)$, $\lambda\gg 1$, $\delta\ll 1$ гэж авъя. Стирлингийн томьёо $n!\approx\sqrt{2\pi n}(n/e)^n$ болон Тейлорын цуваа ашиглавал гарах $\ln[(1+\delta)^{\lambda(1+\delta)+1/2}]=[\lambda(1+\delta)+1/2]\ln(1+\delta)=(\lambda+1/2+\lambda\delta)(\delta-\delta^2/2+O(\delta^3))\approx\lambda\delta+\lambda\delta^2/2+O(\delta^3)$ ойролцоо адилтгалыг ашиглаад эцэст нь $\delta=(x-\lambda)/\lambda$ орлуулга хийвэл

$$\begin{aligned}f_X(x)&=\frac{\lambda^{\lambda(1+\delta)}e^{-\lambda}}{\sqrt{2\pi\lambda(1+\delta)}(\lambda(1+\delta)/e)^{\lambda(1+\delta)}}\\&=\frac{e^{\lambda\delta}(1+\delta)^{-\lambda(1+\delta)-1/2}}{\sqrt{2\pi\lambda}}=\frac{e^{-\lambda\delta^2/2}}{\sqrt{2\pi\lambda}}=\frac{e^{-(x-\lambda)^2/(2\lambda)}}{\sqrt{2\pi\lambda}}\end{aligned}$$

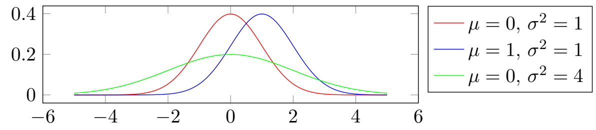
болно.

"Хонх" хэлбэртэй буюу хэвийн тархалт

$f_X(x)=\dfrac{e^{-(x-\lambda)^2/(2\lambda)}}{\sqrt{2\pi\lambda}}$ нь $\mu=\lambda$ ба $\sigma^2=\lambda$ байх үеийн

$$f_X(x)=\frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma},\quad x\in\mathbb{R}$$

нягттай хэвийн тархалт юм. Тэгвэл $\mu=E(X)$ ба $\sigma^2=D(X)$ болно.



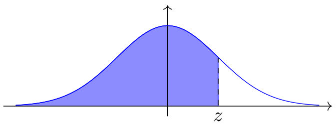
Зураг: Хэвийн тархалтын нягтын муруй параметрийн янз бүрийн утгад

Үүнийг бас Гауссын тархалт ч гэдэг. X хувьсагч хэвийн тархалттай гэхийг $X\sim N(\mu,\sigma^2)$ гэж тэмдэглэнэ.

Стандарт хэвийн тархалтын функц

$N(\mu=0,\sigma^2=1)$ тархалтыг *стандарт хэвийн тархалт* гэнэ.

$$\Phi(z)=P(Z<z)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt=S_{\text{муруй шугаман трапец}}$$



Зураг: $\Phi(z)$ утга буюу z-ээс бага утгуудын хувьд хэвийн тархалтын нягтын муруйн доор байх мужийн талбай

Тархалтын нягт тэгш хэмтэй тул $\Phi(-z)=1-\Phi(z)$ чанар хүчинтэй.

Хэвийн тархалттай санамсаргүй хувьсагчийн шугаман хувиргалт

$X\sim N(\mu,\sigma^2)$ ба

$$Y=a+bX,\quad a,b\in\mathbb{R},\ a\neq 0$$

бол

$$Y\sim N(a+b\mu,b^2\sigma^2)$$

байдаг бөгөөд гаргалгаа нь хожим үзнэ.

Иймд хэрэв $X\sim N(\mu,\sigma^2)$ бол

$$Z=\frac{X-\mu}{\sigma}\,[4]\sim N(0,1)$$

болно.

[4] үүнийг стандарт хувиргалт гэнэ

Хэсэг 5

Хэвийн тархалтын хэрэглээ

Хэвийн тархалтын хэрэглээ

- Статистик загваруудад хэвийн тархалт гол сонголт нь байдаг.
- Хамааралгүй, их олон санамсаргүй хувьсагчдын нийлбэрийн тархалт хэвийн тархалтад ойр байдаг[5].

  Жишээлбэл $X_1,\dots,X_n\sim \mathrm{Ber}(p)$[6] ба хамааралгүй үед $X=X_1+\dots+X_n\sim B(n,p)$[7] байдаг тул

  $$X\sim N(np,np(1-p)),\quad n\to\infty$$

  буюу

  $$\frac{X-np}{\sqrt{np(1-p)}}\sim N(0,1),\quad n\to\infty$$

  байна.

[5] Хязгаарын гол теорем гэдэг нэрээр дараа үзнэ.
[6] Бернуллийн тархалт
[7] бином тархалт

Туршилтын тоо хүрэлцээтэй их үед бином тархалтын нягтыг ойролцоо бодох

Сэргээн санах нь

Бином тархалтын нягтыг n их, p бага үед

$$f_X(x)=P(X=x)=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}\approx\frac{\lambda^x}{x!}e^{-\lambda},\quad\lambda=np$$

гэж ойролцоо бодож болно. Үүнийг *Пуассоны томьёо* гэдэг.

Эсрэгээрээ p их буюу 1-д ойр үед дээрх томьёог $1-p$ магадлал буюу "бүтэлгүйтэл" үзэгдлийн хувьд хэрэглэнэ.

- $p\to 0$ тохиолдолд Пуассоны томьёо
- $p\to 1$ тохиолдолд $p:=1-p$ гэж аваад Пуассоны томьёо
- $0\ll p\ll 1$ тохиолдолд Муавр-Лапласын томьёо (одоо үзнэ)

n их боловч p нь бага ч биш, их ч биш үед бином тархалттай X санамсаргүй хувьсагчийн $\frac{X-np}{\sqrt{np(1-p)}}\sim N(0,1)$, $n\to\infty$ чанарт үндэслэн $P(X=x)$ магадлалыг хэвийн тархалтын нягтаар шууд ойролцоо бодно. Үүнийг *Муавр-Лапласын локал томьёо* гэдэг.

Харин $P(a\le X\le b)=\sum_{x=a}^{b}f_X(x)$ магадлалын хувьд n их үед $\frac{X-np}{\sqrt{np(1-p)}}\sim N(0,1)$, $n\to\infty$ ёсоор

$$P(a\le X\le b)\approx\Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right)-\Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right)$$

болно. Үүнийг *Муавр-Лапласын интеграл томьёо* гэдэг.

Жишээ

Хот орчмын нэг суурин 720 хүн амтай. Оршин суугч бүр бусдаасаа хамааралгүйгээр сард 5 удаа зэргэлдээх хот уруу рэйлбусаар явах бөгөөд хэзээ явах ч нь бусдаас хамаарахгүй. Харин рейлбус өдөрт нэг удаа явдаг. Нэг сард (30 хоног) дунджаар нэгээс ихгүй удаа зорчигчид багтахгүй байж болно гэвэл рейлбус дор хаяж хэдэн хүний суудалтай байх шаардлагатай вэ?

- Энд дараалсан, хамааралгүй туршилтууд яригдаж байна.
- Туршилт: оршин суугч хот явахыг ажиглах; • Амжилт: хот явах;
- Туршилтыг давтах тоо: $n = 720$; • Амжилтын магадлал: $p = 5/30 = 1/6$; • Амжилтын тоо: $x =$ хот явах оршин суугчдын тоо бөгөөд энэ нь рейлбусын суудлын тоотой холбогдоно. Мөн энэ нь мэдэгдэхгүй буюу манай олох зүйл байна.
- Ийнхүү $X =$ хот явах оршин суугчдын тоо гэсэн санамсаргүй хувьсагч авч үзнэ.

Сард (30 хоног) нэгээс ихгүй удаа зорчигчид багтахгүй байх

$$P(X > x) \leq 1/30 \quad \text{буюу} \quad P(0 \leq X \leq x) \geq 29/30$$

Бином тархалтаар халз бодох гэвэл

$$P(X > x) = \sum_{k=x+1}^{720} \frac{720!}{k!(720-k)!} \left(\frac{1}{6}\right)^k \left(1 - \frac{1}{6}\right)^{n-k} \leq 1/30$$

$\frac{X-np}{\sqrt{np(1-p)}} \sim N(0,1)$, $n \to \infty$ буюу Муавр-Лапласын интеграл томьёо ашиглавал

$$P\left(\frac{0 - 720\frac{1}{6}}{\sqrt{720\frac{1}{6}\frac{5}{6}}} \leq \frac{X - 720\frac{1}{6}}{\sqrt{720\frac{1}{6}\frac{5}{6}}} \leq \frac{x - 720\frac{1}{6}}{\sqrt{720\frac{1}{6}\frac{5}{6}}}\right) \geq \frac{29}{30} \quad \Phi\left(\frac{x-120}{10}\right) \geq 0.96$$

$$\frac{x-120}{10} \geq \Phi^{-1}(0.96) \approx 1.75 \qquad x \geq 137.5 \qquad x = 138$$

Муавр-Лапласын интеграл томьёоны тасралтгүйн засвар

Бүхэл утга авдаг, бином тархалттай санамсаргүй хувьсагчтай холбогдох үзэгдлийн магадлалыг бодит тоон утга авдаг, хэвийн тархалттай тасралтгүй санамсаргүй хувьсагчтай холбоотой магадлал ашиглаж байна. Иймд $\{X = x\}$ үзэгдэлд хэвийн тархалттай санамсаргүй хувьсагчийн $\{x - 0.5, x + 0.5\}$ үзэгдэл харгалзуулж болох юм. Ийнхүү Муавр-Лапласын интеграл томьёог дараах байдлаар засварлаж болно.

$$P(a \leq X \leq b) \approx \Phi\left(\frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

Лекц IV

Амьдрах хугацааны тархалт

Амьдрах хугацааны тархалт сэдвийн агуулга

Хэсэг 1

Оршил буюу амьдрах хугацаа

Амьдрах хугацаа

Амьдрах хугацаа гэдэгт ямарваа зүйлийн үргэлжлэх хугацаа эсвэл нөөц чадавхыг хамруулж ойлгоно.

- Биологи, Хүн ам зүй
  Наслалт буюу амьдрах хугацаа
- Инженер техникийн ухаан
  Материалын бат бөхийн нөөц
- Цөмийн физик
  Цацраг идэвхт бөөмийн задрах хүртэлх хугацаа
- Хүлээлгийн онол буюу үйлчилгээний систем
  Дараагийн үйлчлүүлэгч ирэх хүртэлх хугацаа
- Эдийн засаг
  Эрсдэл учрах хүртэлх хугацаа
- . . .

Хэсэг 2

Нөхцөлт магадлал

Нөхцөлт магадлал

B үзэгдэл явагдсан үед A үзэгдэл явагдах магадлалыг дараах байдлаар олно.

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Санамж

$$P(\underbrace{A+B}_{\text{үзэгдэл}}|\underbrace{C+D}_{\text{нөхцөл}}) = \frac{P((A+B)(C+D))}{P(C+D)}$$

A ба B үзэгдэл хамааралгүй үед $P(AB) = P(A)P(B)$ байх тул $P(A|B) = P(A)$ бас $P(B|A) = P(B)$ байна.

Хэсэг 3

Илтгэгч тархалтын санамжгүй байдал

Илтгэгч тархалтын санамжгүй байдал

Чанар

$\forall x, y \geq 0$ бүрийн хувьд $P(X \geq x+y|X \geq y) = P(X \geq x)$ байна.

Баталгаа

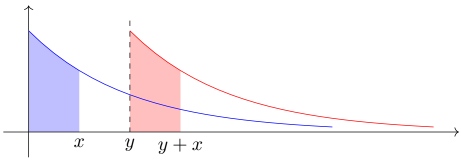
$$\begin{aligned} P(X \geq y+x|X \geq y) &= \frac{P(X \geq y+x, X \geq y)}{P(X \geq y)} = \frac{P(X \geq y+x)}{P(X \geq y)} \\ &= \frac{1-P(X < y+x)}{1-P(X < y)} = \frac{1-F_X(y+x)}{1-F_X(y)} \\ &= \frac{1-(1-e^{-\lambda(y+x)})}{1-(1-e^{-\lambda y})} = e^{-\lambda x} = \\ &= 1-(1-e^{-\lambda x}) = 1-F_X(x) = 1-P(X < x) \\ &= P(X \geq x) \end{aligned}$$

Илтгэгч тархалтын санамжгүй байдал

Амьдрах хугацааны хувьд

$$P(X \geq y+x|X \geq y) = P(X \geq x)$$

чанар нь хэчнээн насалсан нь цааш хэд наслахад нөлөөгүй гэсэн утгатай юм.



Зураг: Илтгэгч тархалтын санамжгүй байдал

Хэсэг 4

Мөхлийн эрчим

Мөхлийн эрчим

Тодорхойлолт

Амьдрах хугацааны тархалтын хувьд

$$h_X(x) = \lim_{\Delta x \searrow 0} \frac{P(x < X < x+\Delta x)|X > x)}{\Delta x}$$

хэмжигдэхүүнийг хугацааны x эгшин дэх *мөхлийн эрчим* гэнэ.

Мөхлийн эрчим нь хугацааны x эгшин хүртэл амьдарсан бол яг x эгшин дээрээ шууд үхэх, мөхөх эрчмийг илэрхийлнэ.

Амьдрах хугацааны тархалтын мөхлийн эрчим

Хэрэв f_X нягт x дээр тасралтгүй бол

$$P(x < X < x+\Delta x) \approx f_X(x)\Delta x$$

байдаг. Мөн $P(X > x) = 1 - F_X(x)$ юм. Иймд

$$\begin{aligned} h_X(x) &= \lim_{\Delta x \searrow 0} \frac{P(x < X < x+\Delta x|X > x)}{\Delta x} \\ &= \lim_{\Delta x \searrow 0} \frac{P(x < X < x+\Delta x, X > x)}{\Delta x P(X > x)} \\ &= \lim_{\Delta x \searrow 0} \frac{P(x < X < x+\Delta x)}{\Delta x(1-F_X(x))} \\ &= \frac{f_X(x)}{1-F_X(x)} \end{aligned}$$

болно.

Илтгэгч тархалтын мөхлийн эрчим

$X \sim \mathrm{Exp}(\lambda)$ санамсаргүй хувьсагчийн хувьд

$$h_X(x) = \frac{f_X(x)}{1-F_X(x)} = \frac{\lambda e^{-\lambda x}}{1-(1-e^{-\lambda x})} = \lambda$$

буюу мөхлийн эрчим нь тогтмол байна. Иймд λ параметрийг *эрчмийн параметр*[1] гэдэг. Энэхүү тогтмол мөхлийн эрчим нь илтгэгч тархалтын санамжгүй байдлын шалтгаан юм.

Тогтмол мөхлийн эрчимтэй, өөрөг өөр санамсаргүй хувьсагч байх уу? Үүний хариуг дараагийн слайд дээр авч үзнэ.

[1] rate parameter

Тогтмол мөхлийн эрчимтэй тархалт

X нь эерэг утгатай, тогтмол мөхлийн эрчимтэй, тасралтгүй с.х. байг.

$$h_X(x) = \frac{f_X(x)}{1 - F_X(x)} = k$$

$$\int_0^x \frac{f_X(t)}{1-F_X(t)}dt = \int_0^x kdt$$

$$-\int_0^x \frac{1}{1-F_X(t)}d(1-F_X(t)) = k\int_0^x dt$$

$$-\ln(1-F_X(t))\,|_0^x = kt\,|_0^x$$

$$-\ln(1-F_X(x)) = kx$$

$$F_X(x) = 1 - e^{-kx}$$

Энэ нь $X \sim \mathrm{Exp}(k)$ болохыг харуулж байна.

Хэсэг 5

Найдварын функц

Найдварын функц[2]

$$\begin{aligned} R_X(x) &= \exp\left\{-\int_0^x h_X(t)dt\right\} \\ &= \exp\left\{-\int_0^x \frac{f_X(t)}{1-F_X(t)}dt\right\} \\ &= \exp\left\{-\int_0^x \frac{1}{1-F_X(t)}d(1-F_X(t))\right\} \\ &= \exp\left\{-\int_0^x d\ln(1-F_X(t))\right\} \\ &= \exp\{\ln(1-F_X(x))\} \\ &= 1 - F_X(x) \end{aligned}$$

[2] Reliability function

Хэсэг 6

Вейбуллын тархалт

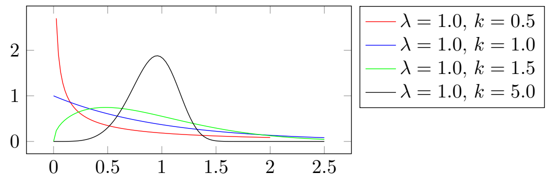
Вейбуллын тархалт

Амьдрах хугацаатай холбоотой тархалтуудын нэг бол Вейбуллын тархалт юм.

$$f_X(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x > 0 \\ 0 & x \le 0 \end{cases}$$

Энд $k > 0$ нь хэлбэрийн параметр, $\lambda > 0$ нь масштабын параметр юм. X санамсаргүй хувьсагч Вейбуллын тархалттай гэхийг $X \sim \mathrm{Weib}(\lambda, k)$ гэж тэмдэглэнэ.

Вейбуллын тархалт, параметрийн янз бүрийн утгад



Зураг: Вейбуллын тархалтын нягтын функц параметрийн янз бүрийн утгад

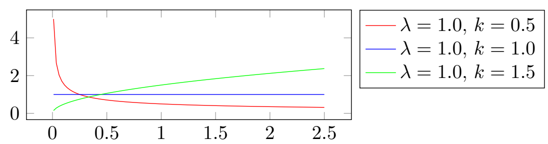
Тархалтын функц

$$\begin{aligned} F_X(x) &= \int_{-\infty}^x f_X(t)dt \\ &= \int_0^x \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} e^{-(t/\lambda)^k}dt \\ &= \int_0^x e^{-(t/\lambda)^k} d(t/\lambda)^k \\ &= e^{-(t/\lambda)^k}\Big|_0^x \\ &= 1 - e^{-(x/\lambda)^k}, \quad \forall x > 0 \end{aligned}$$

$$F_X(x) = \begin{cases} 1 - e^{-(x/\lambda)^k} & x > 0 \\ 0 & x \le 0 \end{cases}$$

Вейбуллын тархалтын мөхлийн эрчим

$$h_X(x) = \frac{f_X(x)}{1-F_X(x)} = \frac{\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}}{1-(1-e^{-(x/\lambda)^k})} = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}$$



Зураг: Вейбуллын тархалтын мөхлийн эрчим

- $k = 1$ үед $h_X(x)$ нь илтгэгч тархалт шиг тогтмол байна.
- $k > 1$ үед $h_X(x)$ өсөх буюу хуучин нь шинээсээ илүү үргэлжлэнэ.
- $k < 1$ үед $h_X(x)$ буурах буюу шинэ нь хуучнаасаа илүү үргэлжлэнэ.

Вейбуллын тархалт болон илтгэгч тархалтын холбоо

Weib(λ, k) буюу

$$f_X(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x > 0 \\ 0 & x \le 0 \end{cases}$$

нь $k = 1$ үед

$$f_X(x) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} & x > 0 \\ 0 & x \le 0 \end{cases}$$

буюу Exp$(1/\lambda)$ өөрөөр хэлбэл $\lambda = 1/\lambda$ параметртэй илтгэгч тархалттай давхцаж байна.

Лекц V

Бернуллийн процесс

Бернуллийн процесс сэдвийн агуулга

Хэсэг 1

Урвасан бином тархалт

Урвасан бином тархалтаар шийдэх бодлого

Жишээ

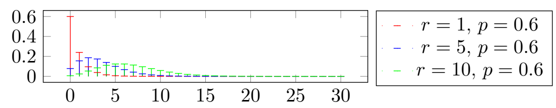
Нэг бадарчин айл хэсч гуйлга гуйхаар хүрээний нэг гудам уруу оров. Гудамд 30 айл байдаг бөгөөд өнөөх бадарчин маань 5 айлаас юм авахаас наш буцахгүй гэж шийдсэн байв. Айл бүрийн хувьд гуйлгачинд юм өгөх магадлал 0.6 бол бадарчинд x ширхэг айл юу ч хялайлгалгүй явуулах магадлал ямар байх вэ?

Туршилт	Айлаас гуйлга гуйх
Амжилт	Айлаас юм авах
Амжилтын магадлал	$p = 0.6$
Санамсаргүй хувьсагч	Юм өгөлгүй явуулсан айлын тоо
Туршилтыг зогсоох нөхцөл	Амжилтын тоо $r = 5$-д хүрэх

Урвасан бином тархалт

Хамааралгүй, нэг ижил тархалттай Бернуллийн туршилтын дараалалд анх өгсөн r тооны амжилтаас өмнөх бүтэлгүйтлийн тооны магадлалын тархалтыг *урвасан бином тархалт*[1] гээд $NB(r, p)$ гэж тэмдэглэнэ. Энд p нь амжилтын магадлал юм.

$$P(X = x) = C_x^{x+r-1}(1-p)^x p^r, \quad x \in \{0, 1, 2, \ldots\}$$



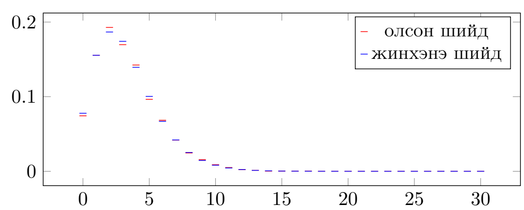
Зураг: Урвасан бином тархалт, параметрийн янз бүрийн утгад

[1] Negative binomial distribution

Бодлогыг Монте-Карло симуляцын аргаар бодсон нь

```
set.seed(0)
N <- vector("mode" = "integer", "length" = 30 + 1)
for (n in 1:10000) {
  attempt <- 0; success <- 0; failure <- 0
  while (success < 5 && attempt < 30) {
    attempt <- attempt + 1
    if (runif(n = 1) < 0.6)
      success <- success + 1
    else
      failure <- failure + 1
  }
  N[failure + 1] <- N[failure + 1] + 1
}
P <- N/10000
print(P)
```


Монте-Карло симуляцын аргаар олсон шийд



Зураг: x ширхэг айл юу ч хялайлгалгүй явуулах магадлал

Жинхэнэ шийдийг R програм дээр дараах тушаалаар олж болно.

```
dnbinom(x = 0:30, prob = 0.6, size = 5)
```

Урвасан бином тархалтын зарим чанар

- $E(X) = \dfrac{(1-p)r}{p}$
  Жишээний хувьд $E(X) = \dfrac{0.4 \cdot 5}{0.6} \approx 3.333$ байна.
- $D(X) = \dfrac{(1-p)r}{p^2}$
  Жишээний хувьд $D(X) = \dfrac{0.4 \cdot 5}{0.6^2} \approx 5.555$ байна.
- Түүврийн дисперс нь түүврийн дунджаасаа их үед Пуассоны тархалт[2]-ын оронд ашигладаг.
- $NB\left(r, \frac{\lambda}{r+\lambda}\right) \xrightarrow{r\to\infty} \mathrm{Pois}(\lambda)$
- $NB(r=1,p) = \mathrm{Geom}(p)$

[2] дундаж ба дисперс нь тэнцүү байдаг

Хэсэг 2

Бернуллийн процесс

Бернуллийн процесс

Сэргээн санах нь

Ямар нэг туршилт авч үзье. Туршилтын үр дүнгээс аль нэг A үзэгдлийг онцгойлон "амжилт" хэмээн авна. Ийнхүү үр дүн нь хоёр ангилсан туршилтыг *Бернуллийн туршилт* гэдэг. $P(A) = p$ бас туршилтыг өөр хоорондоо хамааралгүй байдлаар n удаа давтсан гэе. Тэгвэл үүнтэй холбогдуулан практикт өргөн тохиолдох янз бүрийн санамсаргүй хувьсагч зохиож болно.

Тодорхойлолт

$X_i \sim \mathrm{Ber}(p)$ буюу $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$ бас $X_1, X_2, \ldots$ хамааралгүй байг. Тэгвэл $X_1, X_2, \ldots$ санамсаргүй хувьсагчдын төгсгөлөг болон төгсгөлгүй дарааллыг *Бернуллийн процесс* гэнэ.

Бернуллийн процесстой холбогдох хялбар бодлого

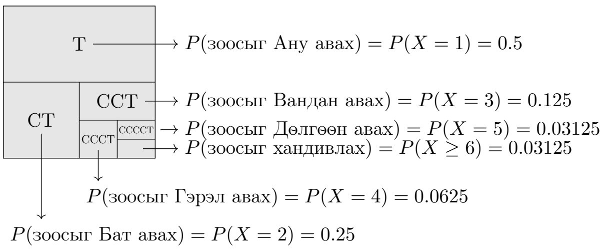
Жишээ (www.slideshare.net/Erdenetsagaanaa/ss-40157371)

Ану, Бат, Вандан, Гэрэл, Дөлгөөн нар нэг зоосон мөнгө олоод түүнийгээ хэн нь авах вэ гэдгээ шодохоор шийджээ. Гэтэл Ану "Нэрсийнхээ дарааллаар зоосоо хаяад хамгийн эхэлж тоотой талаараа буулгасан нь зоосоо авъя. Хэрвээ хэн нь ч тоотой талаараа буулгахгүй бол зоосоо буяны санд хандивлая." гэсэн санал гаргав. Анугийн санал шударга уу?

X нь тоо буутал зоос орхих тоо буюу *амжилт илэртэл явуулах туршилтын тоо* гэвэл

$$P(X = x) = (1-p)^{x-1}p, \quad x = 1, 2, \ldots$$

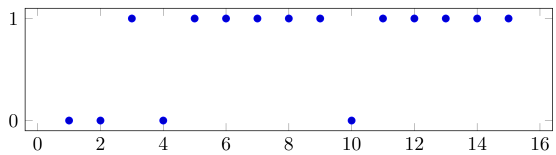
геометр тархалт гарах бөгөөд $p = 0.5$ байна.



Зураг: Бернуллийн процесстой холбогдох бодлогын зураглал

Бернуллийн процессын санамжгүй байдал

$X_1, X_2, \ldots$ хувьсагчид хамааралгүй тул энэ процесс санамжгүй юм. Иймд энэ процессын өнгөрсөнөөс ирээдүйг урьдчилан хэлэх боломжгүй.



Зураг: $p = 0.6$ үед загварчилсан Бернуллийн процессын нэг тохиолдол

Бернуллийн процесстой холбоотой зарим тархалт

Энэ процессын хувьд дараах санамсаргүй хувьсагчдыг өмнө авч үзсэн.

1. Дараалсан n туршилтад илрэх амжилтын тоо.
   $B(n,p)$ буюу бином тархалт.
2. Нэг амжилт илрэх хүртэлх бүтэлгүйтлийн тоо.
   $\mathrm{Geom}(p)$ буюу геометр тархалт.
3. Нэг амжилт илрэх хүртэлх туршилтын тоо.
   $\mathrm{Geom}(p)$ буюу геометр тархалт.
4. r удаа амжилт илрэх хүртэлх бүтэлгүйтлийн тоо.
   $NB(r,p)$ буюу урвасан бином тархалт.

Бернуллийн процесстой холбоотой тархалтуудын уялдаа

- Геометр тархалт ба Урвасан бином тархалт

$$\underbrace{\underbrace{0\,0\,0\,0}_{\mathrm{Geom}(p)}\ 1\ \underbrace{0\,0\,0\,0}_{\mathrm{Geom}(p)}\ 1\ \underbrace{0\,0\,0\,0\,0\,0\,0\,0\,0\,0}_{\mathrm{Geom}(p)}}_{NB(r=3,p)}\ 1\ 0$$

  Өөрөөр хэлбэл нэг ижил геометр тархалттай хамааралгүй хувьсагчдын нийлбэр урвасан бином тархалттай юм.
- Бином тархалт ба Урвасан бином тархалт
  Бүтэлгүйтэл буюу 0-үүдийн тоог s гэе.

$$\overbrace{0\,0\,0\,0\,1\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,1}$$

$$B\ :\ n\ =\ \text{нийт туршилтын тоо}\ =\ s+r\ :\ NB$$
$$B\ :\ r\ =\ \text{амжилтын тоо}\ =\ r\ :\ NB$$
$$B\ :\ \text{амжилтын тоо}\ =\ \text{хувьсагч}\ =\ \text{бүтэлгүйтлийн тоо}:\ NB$$

Ийнхүү $NB(r,p)$ нь $n = s + r$ үед $B(n,p)$ тархалтын "урвуу" юм.

Хамааралгүй Бернуллийн процессуудыг нэгтгэх

X болон Y процессуудын хувьд амжилтын магадлал харгалзан $P(A)=p$ болон $P(B)=q$ байг.

$$\begin{array}{c|l} + \begin{array}{cccccccc} 0&1&0&0&1&0&1&0\\ 0&0&1&0&1&0&0&0 \end{array} & \begin{array}{l} X_i\sim \mathrm{Ber}(p)\\ Y_i\sim \mathrm{Ber}(q)\end{array}\\ \hline \begin{array}{cccccccc} 0&1&1&0&1&0&1&0 \end{array} & Z_i\sim \mathrm{Ber}(p+q-pq) \end{array}$$

Z процессын амжилт болох $A+B$ үзэгдлийн магадлал дараах байдлаар олдоно.

$$\begin{aligned} P(A+B) &= P(A)+P(B)-P(AB)\\ &= P(A)+P(B)-P(A)P(B) \quad /\text{процессууд хамааралгүй}/\\ &= p+q-pq \end{aligned}$$

Бернуллийн процессыг хуваах

$Z\sim \mathrm{Ber}(p)$ процессыг X болон Y хоёр процесст задалъя. Үүний тулд Z процессын амжилтуудад харгалзах $S\sim \mathrm{Ber}(q)$ нэмэлт процесс авч үзнэ.

```
IF Z = 1 THEN
  IF S = 1 THEN
    X := 1
    Y := 0
  ELSE
    X := 0
    Y := 1
  ENDIF
ELSE
  X := 0
  Y := 0
ENDIF
```

Хуваалтын алгоритм ёсоор $X\sim \mathrm{Ber}(pq)$, $Y\sim \mathrm{Ber}(p(1-q))$ байх болно.

Жишээ

$$\begin{array}{cccccccc|l} 0&0&1&0&1&0&1&0 & X_i\sim \mathrm{Ber}(pq)\\ &&\uparrow&&\uparrow&&\uparrow&&\\ 0&1&1&0&1&0&1&0 & Z_i\sim \mathrm{Ber}(p)\\ &\downarrow&&&&&&&\\ 0&1&0&0&0&0&0&0 & Y_i\sim \mathrm{Ber}(p(1-q)) \end{array}$$

Харин шинээр үүсэх X болон Y процессууд хамааралгүй байж чадахгүй.

Бернуллийн процесс дахь k дугаар амжилтад харгалзах туршилтын дугаар

$Y_k=$"k дугаар амжилтад харгалзах туршилтын дугаар" санамсаргүй хувьсагчийн хувьд геометр тархалтын $\{1,2,\ldots\}$ утгууд авдаг $X=$"амжилт илэртэл явуулах туршилтын тоо" гэсэн хувьсагчтай

$$f_X(x)=(1-p)^{x-1}p,\quad x\in\{1,2,\ldots\}$$

хувилбарыг ашиглана.

$$\underbrace{\underbrace{0\,0\,0\,0\,1}_{X_1=5}\;\underbrace{0\,0\,0\,0\,1}_{X_2=5}\;\underbrace{0\,0\,0\,0\,0\,0\,0\,0\,1}_{X_3=9}}_{Y_3=X_1+X_2+X_3=5+5+9=19}\;0$$

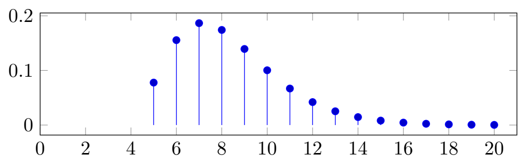
$Y_k=X_1+\ldots+X_k\in\{k,k+1,\ldots\}$ ба $X_1,\ldots,X_k$ хамааралгүй байна. Улмаар $E(X_i)=1/p$, $D(X_i)=(1-p)/p^2$ тул

$$E(Y_k)=k/p,\quad D(Y_k)=k(1-p)/p^2$$

байна.

Бернуллийн процесс дахь "k дугаар амжилтад харгалзах туршилтын дугаар" хувьсагчийн магадлалын тархалт

$$\begin{aligned} P(Y_k=x) &= P(\{\text{эхний } x-1 \text{ туршилтаар } k-1 \text{ амжилт илрэх}\}\\ &\qquad\{x \text{ дүгээр туршилт амжилттай болох}\})\\ &= C_{x-1}^{k-1}p^{k-1}(1-p)^{x-k}p,\quad x\in\{k,k+1,\ldots\} \end{aligned}$$



Зураг: Y_5 хувьсагчийн нягт, $p=0.6$ үед

Бернуллийн процесс дахь дараалсан амжилтын тооны тархалт

Энэ хувьсагч угтаа "бүтэлгүйтэх хүртэл явуулах туршилтын тоо" байх тул $x\in\{1,2,\ldots\}$ утга бүхий хувьсагчтай Geom(1-p) тархалтад захирагдана.

Лекц VI

Пуассоны процесс

Пуассоны процесс сэдвийн агуулга

Хэсэг 1

Гамма тархалт

Гамма функц

$$\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx, \quad k > 0$$

- $\Gamma(k+1) = k\Gamma(k)$
- $\Gamma(1) = 1$
- $\Gamma(1/2) = \sqrt{\pi}$
- $k \in \mathbb{N}$ бол $\Gamma(k) = (k-1)!$

Гамма функцээс гамма тархалт

$x \geq 0$ үед $x^{k-1}e^{-x} > 0$ буюу нягтын функцийн $f_X(x) \geq 0$ чанар биелнэ. Улмаар $\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx, \quad k > 0$ тул нягтын функцийн $\int\limits_{-\infty}^{\infty} f_X(x)dx = 1$ чанарт нийцүүлэхийн тулд

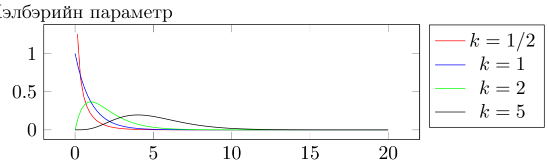
$$f_X(x) = \frac{1}{\Gamma(k)}x^{k-1}e^{-x}$$

гэж авч болно.

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(k)}x^{k-1}e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Гамма тархалтын параметрүүд

- Хэлбэрийн параметр



Зураг: $f_X(x) = \frac{1}{\Gamma(k)}x^{k-1}e^{-x}$ функц k параметрийн янз бүрийн утгад

- Масштаб оруулж ирэх буюу эрчмийн параметр нэмэх
  $\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx$ функцэд $x := \lambda x$[1] орлуулга хийвэл
  $(\lambda x)^{k-1}e^{-\lambda x}d(\lambda x) = \lambda^k x^{k-1}e^{-\lambda x}dx$ болох тул

$$f_X(x) = \frac{1}{\Gamma(k)}\lambda^k x^{k-1}e^{-\lambda x}, \quad x \geq 0$$

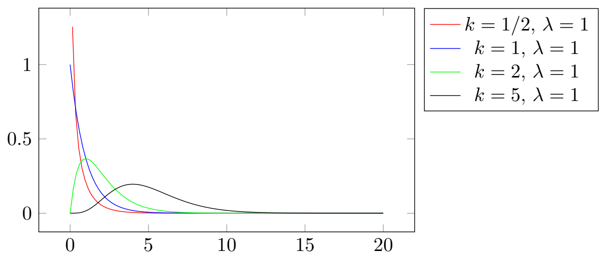
[1] $x := x/\lambda$ орлуулгад харгалзах хувилбарыг масштабын параметр гэдэг.

Гамма тархалт

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(k)}\lambda^k x^{k-1}e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Энд $\lambda > 0$ нь эрчмийн параметр, $k > 0$ нь хэлбэрийн параметр юм.

Гамма тархалт параметрийн янз бүрийн утгад



Зураг: Гамма тархалтын нягтын муруй параметрийн янз бүрийн утгад

Гамма тархалтын чанар

- $E(X) = \frac{k}{\lambda}$
- $D(X) = \frac{k}{\lambda^2}$
- $X \sim \text{Gamma}(\lambda, k)$ бол

$$Y = cX \sim \text{Gamma}(\lambda/c, k)$$

Гамма тархалт бусад тархалттай холбогдох нь

- $\text{Gamma}(\lambda, 1) = \text{Exp}(\lambda)$
- $k_1, k_2 \in \mathbb{N}$ ба $X_1 \sim \text{Gamma}(\lambda, k_1)$, $X_2 \sim \text{Gamma}(\lambda, k_2)$ хувьсагчид хамааралгүй бол

$$X_1 + X_2 \sim \text{Gamma}(\lambda, k_1 + k_2)$$

  буюу $X_1, \ldots, X_k \sim \text{Exp}(\lambda)$ хамааралгүй бол

$$X_1 + \ldots + X_k \sim \text{Gamma}(\lambda, k)$$

  байна. Энэ чанарын гаргалгаатай дараагийн хэсэгт танилцана.
- $k \in \mathbb{N}$ бол $\text{Gamma}(\lambda, k)$ нь Эрлангийн тархалттай давхацна.
- Хэрэв $X \sim N(0,1)$ бол $X^2 \sim \text{Gamma}(1/2, 1/2)$
- Хэрэв $X_1, \ldots, X_k$ хувьсагчид хамааралгүй бөгөөд $N(0,1)$ тархалттай бол

$$X_1^2 + \ldots + X_k^2 \sim \text{Gamma}(1/2, k/2) = \chi^2(k)$$

  $\chi^2(k)$ бол k чөлөөний зэрэгтэй *хи-квадрат тархалт* юм.

Хэсэг 2

Пуассоны процесс

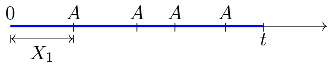
Илтгэгч тархалт ба Пуассоны тархалтын уялдаа холбоо

$$f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \in \{0, 1, 2, \dots\}$$

буюу Пуассоны тархалтын параметр $\lambda > 0$ нь нэгж хугацаанд илрэх амжилтын дундаж тоог илэрхийлдэг. Иймд t хугацаанд амжилт ядаж нэг удаа илрэх үзэгдлийн

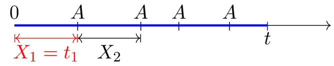
$$P(X \geq 1) = 1 - P(X = 0) = 1 - f_X(0) = 1 - \frac{(\lambda t)^0}{0!} e^{-\lambda t} = 1 - e^{-\lambda t}$$

магадлалыг илтгэгч тархалтын зүгээс харвал энэ нь эхний амжилт илэртэл хүлээх хугацаа буюу илтгэгч тархалттай санамсаргүй хувьсагч X_1 нь t-ээс бага байх $P(X_1 < t)$ магадлалтай тэнцүү байна.



Зураг: Пуассоны процессын эхний амжилт илрэх хүртэлх хугацаа

Удаах амжилт илэртэл хүлээх хугацааны тархалт



Зураг: Пуассоны процесс дахь удаах амжилт илрэх хүртэлх хугацаа

Үлдэж буй $t - t_1$ хугацаанд амжилт ядаж нэг удаа илрэх магадлал

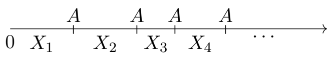
$$1 - \frac{(\lambda(t - t_1))^0}{0!} e^{-\lambda(t-t_1)} = 1 - e^{-\lambda(t-t_1)} = P(X_2 < t - t_1)$$

буюу илтгэгч тархалттай "удаах амжилт илэртэл хүлээх хугацаа" $t - t_1$-ээс бага байх магадлалтай тэнцүү байна. Нөгөө талаас $X_2 \sim \mathrm{Exp}(\lambda)$ гэвэл илтгэгч тархалтын санамжгүй чанараар

$$\begin{aligned} P(X_2 < t | X_2 \geq t_1) &= 1 - P(X_2 \geq t | X_2 \geq t_1) \\ &= 1 - P(X_2 \geq t_1 + (t - t_1) | X_2 \geq t_1) = 1 - P(X_2 \geq t - t_1) \\ &= P(X_2 < t - t_1) = 1 - e^{-\lambda(t-t_1)} \end{aligned}$$

болж байна. Иймд байдлаар $X_3, X_4, \dots \sim \mathrm{Exp}(\lambda)$ болно.

Пуассоны процесс[2]



Зураг: Пуассоны процессын зураглал

Тодорхойлолт

X_1 нь эхний амжилт илэртэл хүлээх хугацаа, X_i $(i = 2, 3, \dots)$ нь дараагийн амжилт хоорондын хугацаа ба $X_1, X_2, \dots$ хувьсагчид нэг ижил $\mathrm{Exp}(\lambda)$ тархалттай бөгөөд хамааралгүй байг. Тэгвэл $X_1, X_2, \dots$ санамсаргүй хувьсагчдын дарааллыг *Пуассоны процесс* гэнэ.

[2] Poisson process

Пуассоны процессын жишээ

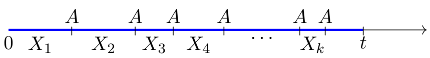
Жишээ

"12th IAAF World Championships In Athletics: IAAF Statistics Handbook. Berlin 2009" тайлан дахь 1913 оноос 2009 оны хооронд хөнгөн атлетикийн эрэгтэйчүүдийн дундын зайн гүйлтийн төрөлд гарсан 32 рекорд амжилтын мэдээллээс гарган авсан X ="удаах рекорд амжилт хүртэлх хугацаа" (жилээр) хувьсагчийн ажиглагдсан утгууд 2.126, 8.11, 8.121, 1.781, 0.921, 3.203, 4.844, 0.025, 0.153, 0.822, 1.049, 0.997, 8.808, 0.126, 3.079, 1.049, 3.479, 2.808, 0.559, 1.104, 0.934, 7.904, 0.238, 3.932, 0.959, 1.134, 0.019, 0.005, 3.915, 8.115, 5.838 байв.

Илтгэгч тархалтын хувьд $E(X) = 1/\lambda$ ба дээрх утгуудын дундаж ойролцоогоор 2.779 тул $\lambda = 0.3598$ гэж "үнэлэв".

$$\begin{aligned} P(\text{жилийн дотор шинэ рекорд тогтоох}) &= P(X < 1) = 1 - e^{-0.3598 \cdot 1} \\ &= 0.302 \end{aligned}$$

t хугацаанд амжилт яг k удаа илрэх магадлал ба Пуассоны тархалт



Зураг: Пуассоны процесст t хугацаанд амжилт яг k удаа илрэх

Өмнөх гаргалгаа, түүний тайлбар болон өмнө үзсэн Пуассоны тархалтын тодорхойлолт зэргээс

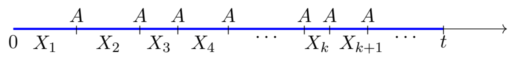
$$P(t \text{ хугацаанд амжилт } k \text{ удаа илрэх}) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

гэж гарна.

Жишээний хувьд

$$\begin{aligned} P(4 \text{ жилд яг нэг шинэ рекорд гарах}) &= \frac{(0.3598 \cdot 4)^1}{1!} e^{-0.3598 \cdot 4} \\ &= 0.341 \end{aligned}$$

t хугацаанд амжилт дор хаяж k удаа илрэх магадлал ба гамма тархалт



Зураг: Пуассоны процесст t хугацаанд амжилт дор хаяж k удаа илрэх

Эхний k амжилт илрэх хугацааг илэрхийлэх $X = X_1 + \dots + X_k$ санамсаргүй хувьсагч авч үзье. Энд $X_i \sim \mathrm{Exp}(\lambda)$ бөгөөд хамааралгүй хувьсагчид юм.

$$\begin{aligned} F_X(t) &= P(X_1 + \dots + X_k < t) \\ &= P(\text{амжилт } k \text{ удаа илрэх}) + P(\text{амжилт } k+1 \text{ удаа илрэх}) + \dots \\ &= \sum_{n=k}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \end{aligned}$$

$$\begin{aligned} f_X(t) = F_X'(t) &= \frac{d}{dt} \sum_{n=k}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \\ &= \sum_{n=k}^{\infty} \left[\frac{n(\lambda t)^{n-1}\lambda}{n!} e^{-\lambda t} + \frac{(\lambda t)^n}{n!} e^{-\lambda t}(-\lambda) \right] \\ &= \sum_{n=k}^{\infty} \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t} - \sum_{n=k}^{\infty} \frac{\lambda^{n+1} t^n}{n!} e^{-\lambda t} \\ &= e^{-\lambda t} \left[\frac{\lambda^k t^{k-1}}{(k-1)!} + \sum_{n=k}^{\infty} \frac{\lambda^{n+1} t^n}{n!} - \sum_{n=k}^{\infty} \frac{\lambda^{n+1} t^n}{n!} \right] \\ &= \frac{1}{(k-1)!} \lambda^k t^{k-1} e^{-\lambda t} = \frac{1}{\Gamma(k)} \lambda^k t^{k-1} e^{-\lambda t}, \quad t \geq 0, \quad k \in \mathbb{N} \end{aligned}$$

буюу нэг ижил илтгэгч тархалттай хамааралгүй хувьсагчдын нийлбэрээр тодорхойлогдох $X = X_1 + \dots + X_k$ хувьсагч $\mathrm{Gamma}(\lambda, k)$ тархалттай байна.

Жишээний хувьд $\lambda = 0.3598$ ба дараах үзэгдлийн хувьд $k = 1$ байх тул

$$\begin{aligned} &P(4 \text{ жилд дор хаяж нэг шинэ рекорд гарах}) \\ &\quad = P(X \leq 4) = F_X(4) \\ &\quad = \int_0^4 f_X(x) dx \\ &\quad = \int_0^4 \frac{1}{\Gamma(1)} 0.3598^1 x^{1-1} e^{-0.3598 \cdot x} dx \\ &\quad = -\int_0^4 e^{-0.3598 \cdot x} d(-0.3598 x) \\ &\quad = -\left[e^{-0.3598 \cdot x} \right]_0^4 \approx 1 - 0.237 \\ &\quad \approx 0.763 \end{aligned}$$

болно.

Лекц VII

Санамсаргүй хувьсагчийн хувиргалт

Санамсаргүй хувьсагчийн хувиргалт сэдвийн агуулга

Хэсэг 1

Санамсаргүй хувьсагчийн хувиргалт

Санамсаргүй хувьсагчийн хувиргалт

Тодорхойлолт

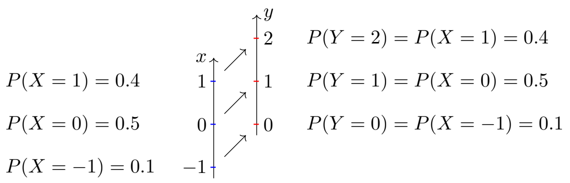
Санамсаргүй хувьсагчийн утгыг ямар нэг функцээр үйлчлэн өөрчлөхийг *санамсаргүй хувьсагчийн хувиргалт* гэнэ.

Дискрет санамсаргүй хувьсагчийн 1:1 хувиргалт

x	-1	0	1
$f_X(x)$	0.1	0.5	0.4

Хүснэгт: X санамсаргүй хувьсагчийн тархалтын хүснэгт

X санамсаргүй хувьсагчийг $g(x) = x + 1$ функцээр хувиргахад үүсэх $Y = g(X) = X + 1$ санамсаргүй хувьсагчийн тархалтыг олъё.



Зураг: Дискрет санамсаргүй хувьсагчийн 1:1 чанартай хувиргалт

y	0	1	2
$f_Y(x)$	0.1	0.5	0.4

Хүснэгт: $Y = X + 1$ санамсаргүй хувьсагчийн тархалтын хүснэгт

Бодолт дараах байдалтай байсан.

$$f_Y(2) = P(Y = 2) = P(X = 1) = f_X(1) = f_X(2 - 1) = f_X(g^{-1}(2))$$
$$f_Y(1) = P(Y = 1) = P(X = 0) = f_X(0) = f_X(1 - 1) = f_X(g^{-1}(1))$$
$$f_Y(0) = P(Y = 0) = P(X = -1) = f_X(-1) = f_X(0 - 1) = f_X(g^{-1}(0))$$

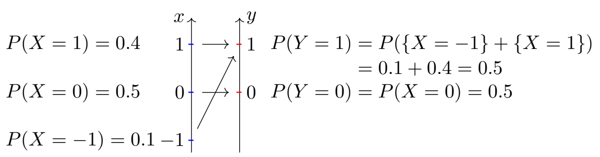
Иймхүү 1:1 чанартай $g(\cdot)$ функцээр тодорхойлогдох $Y = g(X)$ дискрет санамсаргүй хувьсагчийн нягт олох

$$f_Y(y) = f_X(g^{-1}(y))$$

томъёо зохиож болно. Энд $g^{-1}(\cdot)$ нь $g(\cdot)$ функцийн урвуу юм.

Дискрет хувьсагчийн 1:1 нөхцөл үл хангах хувиргалт

$g(\cdot)$ функц 1:1 нөхцөл үл хангах бол $f_Y(y) = f_X(g^{-1}(y))$ томъёо хэрэглэх боломжгүй. 1:1 нөхцөл үл хангах $g(x) = |x|$ функцээр үүсэх $Y = |X|$ хувьсагчийн тархалтыг олъё.



Зураг: Дискрет санамсаргүй хувьсагчийн 1:1 нөхцөл үл хангах хувиргалт

y	0	1
$f_Y(x)$	0.5	$0.1 + 0.4 = 0.5$

Хүснэгт: $Y = |X|$ санамсаргүй хувьсагчийн тархалтын хүснэгт

Y хувьсагчийн тархалтыг олохын тулд

$$\begin{aligned} f_Y(1) &= P(Y = 1) = P(\{X = -1\} + \{X = 1\}) \\ &= P(X = -1) + P(X = 1) = 0.1 + 0.4 \\ &= f_X(-1) + f_X(1) \end{aligned}$$

бодолт хийсэн.

Иймхүү 1:1 нөхцөл үл хангах $g(\cdot)$ функцээр тодорхойлогдох $Y = g(X)$ дискрет санамсаргүй хувьсагчийн нягт олох дараах томъёо зохиож болно.

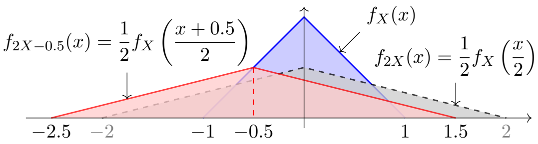
$$f_Y(y) = \sum_{x \in g^{-1}(y)} f_X(x)$$

Жишээ дэх $f_Y(1)$ нягтын хувьд $g^{-1}(1) = \{-1, 1\}$ байсан тул $f_X(-1)$ болон $f_X(1)$ нягтуудыг нэмсэн.

Тасралтгүй санамсаргүй хувьсагчийн хувиргалт

Жишээ

$X \sim \text{tri}(a=-1, b=1, c=0)$ гурвалжин тархалттай хувьсагчийг $Y = 2X - 0.5$ гэж хувиргахад тархалт нь хэрхэн өөрчлөгдөх вэ?



Зураг: $Y = 2X - 0.5$ хувиргалтаар тархалт өөрчлөгдөх байдал

- 2 дахин "сунгасан" тул нягт 2 дахин багасна. Учир нь бүх нягтын нийлбэр 1-тэй тэнцүү.
- -0.5 нэгжээр зөөхөд нягт өөрчлөгдөхгүй.

$Y = a + bX$ шугаман хувиргалт

- $b > 0$ үед

$$F_Y(y) = P(Y < y) = P(a + bX < y) = P\left(X < \frac{y-a}{b}\right) = F_X\left(\frac{y-a}{b}\right)$$

$$f_Y(y) = F_Y'(y) = f_X\left(\frac{y-a}{b}\right)\frac{1}{b} = \frac{1}{b} f_X\left(\frac{y-a}{b}\right)$$

- $b < 0$ үед

$$F_Y(y) = P(a + bX < y) = P\left(X > \frac{y-a}{b}\right) = 1 - F_X\left(\frac{y-a}{b}\right)$$

$$f_Y(y) = F_Y'(y) = -f_X\left(\frac{y-a}{b}\right)\frac{1}{b} = \frac{1}{-b} f_X\left(\frac{y-a}{b}\right)$$

- ерөнхий тохиолдолд

$$f_Y(y) = \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right)$$

Хэвийн тархалттай хувьсагчийн шугаман хувиргалт

Жишээ

$X \sim N(\mu, \sigma^2)$ бол $Y = a + bX$ хувьсагчийн тархалтыг ол.

Сэргээн санах нь

$X \sim N(\mu, \sigma^2)$ ба $Y = a + bX$, $a, b \in \mathbb{R}$, $b \neq 0$ бол $Y \sim N(a + b\mu, b^2\sigma^2)$.

$$\begin{aligned} f_Y(y) &= \frac{1}{|b|} f_X\left(\frac{y-a}{b}\right) = \frac{1}{|b|}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\left(\frac{y-a}{b} - \mu\right)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}|b|\sigma}\exp\left\{-\frac{(y - (a + b\mu))^2}{2b^2\sigma^2}\right\} \end{aligned}$$

$$\mu_Y = a + b\mu,\ \sigma_Y^2 = b^2\sigma^2,\ f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left\{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right\}$$

$Y = g(X)$ хувиргалт

$g(\cdot)$ функц X хувьсагчийн авах утгын олонлог дээр монотон бөгөөд 1:1 байдлаар буулгадаг байг.

- $g(\cdot)$ өсдөг үед $g^{-1}(\cdot)$ бас өсөх ба

$$F_Y(y) = P(Y < y) = P(g(X) < y) = P(X < g^{-1}(y)) = F_X(g^{-1}(y))$$

$$f_Y(y) = F_Y'(y) = f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y)$$

- $g(\cdot)$ буурдаг үед $g^{-1}(\cdot)$ бас буурах ба

$$F_Y(y) = P(Y < y) = P(g(X) < y) = P(X > g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

$$f_Y(y) = F_Y'(y) = -f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y)$$

- ерөнхий тохиолдолд

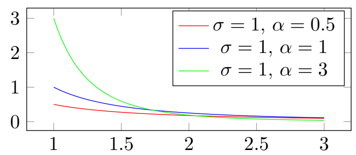
$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|$$

Жигд тархалттай с.х.-ийн $Y = \sigma(1 - X)^{-1/\alpha}$ хувиргалт

Энд $X \sim \mathrm{U}(0, 1)$, $\sigma > 0$, $\alpha > 0$ байна.
$f_X(x) = 1$, $0 < x < 1$ тул $f_X(g^{-1}(y)) = 1$ болно.
$g^{-1}(y) = 1 - \left(\frac{y}{\sigma}\right)^{-\alpha}$ тул $\left|\frac{d}{dy}g^{-1}(y)\right| = \frac{\alpha}{\sigma}\left(\frac{y}{\sigma}\right)^{-\alpha-1}$ улмаар

$$f_Y(y) = \frac{\alpha\sigma^\alpha}{y^{\alpha+1}}, \quad y > \sigma,\ \sigma > 0,\ \alpha > 0$$

буюу I төрлийн Парето тархалт гарна.



Парето тархалтын нягт, параметрийн янз бүрийн утгад

Парето тархалтын хэрэглээ ба өргөн сүүлтэй тархалт

Парето тархалт бол илтгэгч тархалтаас илүү урт сүүлтэй өөрөөр хэлбэл хэт их утгын магадлал илтгэгч тархалтынхаас их юм. Иймэрхүү тархалтуудыг ерөнхийд нь *өргөн сүүлтэй* гэдэг. Өргөн сүүлтэй тархалтууд тохирох зарим тохиолдлыг жишээ болгон дор жагсаав.

1. Элсний ширхэгийн хэмжээ
2. Солирын хэмжээ
3. Сүлжээгээр дамжих файлын хэмжээ
4. Суперкомпьютерт өгөх ажлын хэмжээ
5. Хот, суурингийн хэмжээ
6. Эрсдэл, гамшиг

Мөн энэ тархалттай холбоотой Парето зарчим буюу 80-20-ийн зарчим гэж бий.

$Y = g(X)$ тасралтгүй санамсаргүй хувьсагчийн тархалтыг олох тархалтын функцэд суурилсан арга

Энэ сэдэвт үзэж байгаа аргыг ерөнхийд нь дараах байдлаар алгоритмчилж болно.

1. $\{Y < y\}$ үзэгдлийг X санамсаргүй хувьсагч ашиглаж илэрхийлнэ.
2. $F_Y(y)$ тархалтын функц буюу $P(Y < y)$ магадлалыг X санамсаргүй хувьсагчийн тархалтын функц $F_X(x)$ ашиглаж олно.
3. $F_Y(y)$ функцээс уламжлал авч $f_Y(y)$ нягтын функцийг олно.

Тасралтгүй санамсаргүй хувьсагчийн $Y = X^2$ хувиргалт

$y = g(x) = x^2$ ба $g^{-1}(y) = \sqrt{y}$ буюу 1:1 чанар алдагдсан байна. Энэ тохиолдолд дараах байдлаар Y хувьсагчийн тархалтыг олж болно.

$$\begin{aligned} F_Y(y) &= P(Y < y) = P(X^2 < y) \\ &= P(-\sqrt{y} < X < \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ f_Y(y) &= F_Y'(y) = f_X(\sqrt{y})\frac{d}{dy}(\sqrt{y}) - f_X(-\sqrt{y})\frac{d}{dy}(-\sqrt{y}) \\ &= \frac{1}{2\sqrt{y}}(f_X(\sqrt{y}) + f_X(-\sqrt{y})) \end{aligned}$$

Тухайн тохиолдолд $f_X(x)$ нягт тэгш хэмтэй бол

$$f_Y(y) = \frac{1}{\sqrt{y}} f_X(\sqrt{y})$$

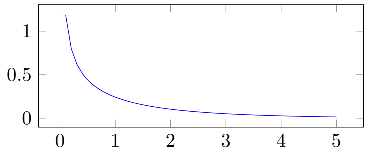
болно.

Стандарт хэвийн тархалттай хувьсагчийн $Y = X^2$ хувиргалт ба хи-квадрат тархалт

$X \sim N(0,1)$ байг. $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ нягт тэгш хэмтэй тул

$$f_Y(y) = \frac{1}{\sqrt{y}} f_X(\sqrt{y}) = \frac{1}{\sqrt{2\pi y}} e^{-y/2}$$

буюу 1 чөлөөний зэрэгтэй хи-квадрат тархалт гарна.



Зураг: $f_Y(y)$ функцийн график буюу 1 чөлөөний зэрэгтэй хи-квадрат нягтын муруй

Геометр тархалттай хувьсагчийн $Y = X^2$ хувиргалт

$X \sim \mathrm{Geom}(p)$ бүр тодруулбал

$$f_X(x) = (1-p)^x p \quad x \in \{0, 1, 2, 3, \dots\}$$

байг. Тэгвэл $Y = X^2$ санамсаргүй хувьсагчийн тархалтыг олъё. $Y = g(X) = X^2$ хувиргалтын $g(\cdot)$ функц нь геометр тархалттай санамсаргүй хувьсагчийн авах утгын олонлог $\{0, 1, 2, \dots\}$ дээр 1:1 чанартай буулгалт байна. Иймд

$$f_Y(y) = f_X(g^{-1}(y))$$

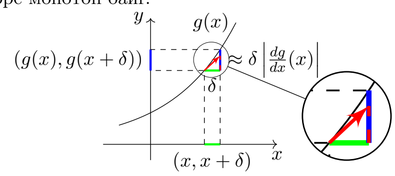
томьёогоор тархалтыг нь олж болно. Ийнхүү Y хувьсагчийн нягт

$$f_Y(y) = f_X(g^{-1}(y)) = (1-p)^{\sqrt{y}} p \quad y \in \{0, 1, 4, 9, \dots\}$$

болно.

$f_Y(y) = f_X(g^{-1}(y)) \left|\frac{d}{dy} g^{-1}(y)\right|$ томьёоны өөр гаргалгаа

$g(\cdot)$ функц эрс монотон байг.



Зураг: Хувиргалтаар үүсэх хувьсагчийн нягт олох томьёоны гаргалгаа

$$P(x < X < x+\delta) = P(\underbrace{g(x)}_{y} < Y < g(x+\delta)) \approx P\left(y < Y < y + \delta\left|\tfrac{dg}{dx}(x)\right|\right)$$

$$\delta f_X(x) \qquad\qquad \delta\left|\tfrac{dg}{dx}(x)\right| f_Y(y)$$

$$f_Y(y) = f_X(x) \frac{1}{\left|\frac{dg}{dx}(x)\right|} = f_X(g^{-1}(y)) \left|\frac{d}{dy} g^{-1}(y)\right|$$

Хэсэг 2

Санамсаргүй хувьсагч загварчлах урвуу хувиргалтын арга

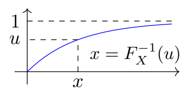
Урвуу хувиргалтын арга

X санамсаргүй хувьсагч тасралтгүй, $F_X(x)$ тархалтын функц X хувьсагчийн боломжит утгын олонлог дээр эрс өсдөг байг. Тэгвэл $F_X(x)$ функц урвуутай байна.

$U = F_X(X)$ гэж авбал $0 \leq U \leq 1$ байна. Харин тархалт нь

$$\begin{aligned} F_U(u) &= P(U < u) = P(F_X(X) < u) \\ &= P(X < F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u \end{aligned}$$

буюу $[0, 1]$ завсар дээрх жигд тархалт байна.



Зураг: Урвуу хувиргалтын аргын санаа ба тархалтын квантил

Урвуу хувиргалтын аргаар Парето тархалттай санамсаргүй тоо үүсгэх

I төрлийн Парето тархалтын функц $F_X(x) = 1 - \left(\frac{x}{\sigma}\right)^{-\alpha}$ тул

$$x = \frac{\sigma}{(1-u)^{1/\alpha}}$$

томьёо гарна. $1 - U \sim \mathrm{U}(0, 1)$ тул үүнийг дараах байдлаар өөрчилж болно.

$$x = \frac{\sigma}{u^{1/\alpha}}$$

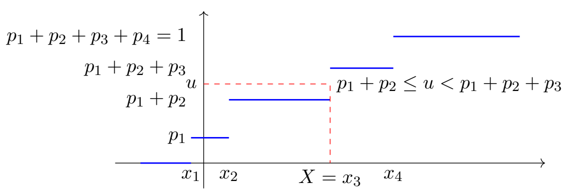
Тус томьёо болон урвуу хувиргалтын аргын дагуу дараах програм бичиж болно. Програмын кодыг R хэлээр бичив.

```
sigma <- 1
alpha <- 3
U <- runif(n = 1000)
X <- sigma / U ** {1 / alpha}
```


Урвуу хувиргалтын аргаар дискрет тархалт загварчлах

$$f_X(x) = P(X = x_i) = p_i, \quad i = 1, 2, \dots, \quad \sum_i p_i = 1, \quad x_1 < x_2 < \dots$$

$$F_X(x) = \sum_{x_i < x} f_X(X = x_i)$$



Зураг: Урвуу хувиргалтын аргаа дискрет санамсаргүй хувьсагч загварчлах зарчим

Урвуу хувиргалтын аргаар дискрет санамсаргүй хувьсагч загварчлах алгоритм

1. $U \sim \mathrm{U}[0, 1]$ санамсаргүй тоо үүсгэнэ
2. $U < F_X(x_k)$ байх хамгийн бага эерэг k тоог хайж олох бөгөөд $X = x_k$ гэж авна

Жишээ

x	-1	0	1
$f_X(x)$	0.1	0.5	0.4

Хүснэгт: Дискрет санамсаргүй хувьсагчийн тархалтын хүснэгт

Хүснэгтээр өгсөн тархалттай X санамсаргүй хувьсагчийг урвуу хувиргалтын аргаар загварчилъя.

```
import random
u = random.random()
if u < 0.1 :
  print -1
elif u < 0.1 + 0.5 :
  print 0
else :
  print 1
```


Лекц VIII

Хамтын тархалт ба санамсаргүй хувьсагчдын хамаарал

Хамтын тархалт ба санамсаргүй хувьсагчдын хамаарал сэдвийн агуулга

Хэсэг 1

Санамсаргүй вектор, түүний тархалт

Санамсаргүй вектор, хамтын тархалт

Тодорхойлолт

Нэгээс олон санамсаргүй хувьсагчдыг хамтад нь *санамсаргүй вектор*, $(X_1, \dots, X_p)$ санамсаргүй векторын тархалтыг *хамтын тархалт* гэнэ.

Хүйс	Солгой: тийм	Солгой: үгүй
эр	2	3
эм	1	4

(a) Хамтын давтамжийн хүснэгт

X_1	$X_2 = 1$	$X_2 = 0$
1	0.2	0.3
0	0.1	0.4

(b) Хамтын тархалтын хүснэгт

Хүснэгт: Хамтын тархалт, статистик хэмжээс ашиглаж олсон

Энд $P(\text{Хүйс} = \text{эр}, \text{Солгой} = \text{тийм}) = P(X_1 = 1, X_2 = 1) = \frac{2}{10} = 0.2$ байна.

$$f_{X,Y}(x,y) = \begin{cases} \frac{3x^2}{16} + \frac{y}{2}, & 0 < x < 2,\ 0 < y < 1 \\ 0, & \text{бусад} \end{cases}$$

хамтын нягттай (X, Y) санамсаргүй вектор авч үзье.

Зураг: Тасралтгүй санамсаргүй векторын хамтын нягт ба санамсаргүй түүвэр

Хамтын тархалт ба үзэгдлийн магадлал

Дискрет санамсаргүй вектор

$$P((X_1, \dots, X_p) \in D) = \sum_{(x_1,\dots,x_p) \in D} f_{(X_1,\dots,X_p)}(x_1, \dots, x_p)$$

$$\begin{aligned} P((X_1 \geq 0, X_2 = 1) &= P(X_1 = 0, X_2 = 1) + P(X_1 = 1, X_2 = 1) \\ &= 0.1 + 0.2 = 0.3 \end{aligned}$$

Тасралтгүй санамсаргүй вектор

$$P((X_1, \dots, X_p) \in D) = \int_D f_{(X_1,\dots,X_p)}(x_1, \dots, x_p) dx_1 \dots dx_p$$

$$P(X < 1, Y < 0.5) = \int_{\{(x,y): 0<x<1; 0<y<0.5\}} \left(\frac{3x^2}{16} + \frac{y}{2} \right) dx dy = \frac{3}{32}$$

Тухайн тархалт

Тодорхойлолт

Санамсаргүй векторын дэд векторын тархалтыг түүний *тухайн тархалт* гэнэ.

Хамтын тархалтаас тухайн тархалт олохдоо зайлуулах гэж буй санамсаргүй хувьсагчийн хувьд гарцаагүй үзэгдэл авна.

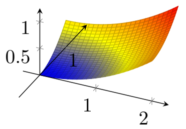
X_1	$X_2 = \mathbf{1}$	$X_2 = 0$	$\sum$
1	**0.2**	0.3	0.5
0	**0.1**	0.4	0.5
$\sum$	**0.3**	0.7	1

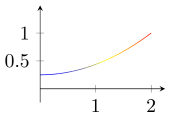
Хүснэгт: Хамтын тархалт ба тухайн тархалт

$$\begin{aligned} P(X_2 = 1) &= P(X_1 = 1, X_2 = 1) + P(X_1 = 0, X_2 = 1) \\ &= 0.2 + 0.1 = 0.3 \end{aligned}$$

$$f_{X,Y}(x,y)=\begin{cases}\dfrac{3x^2}{16}+\dfrac{y}{2}, & 0<x<2,\ 0<y<1\\ 0, & \text{бусад}\end{cases}$$

$$f_X(x)=\int_{-\infty}^{\infty}f_{X,Y}(x,y)dy=\int_0^1\left(\frac{3x^2}{16}+\frac{y}{2}\right)dy=\frac{3x^2}{16}+\frac{1}{4}$$





Зураг: Тасралтгүй санамсаргүй векторын хамтын болон тухайн нягт

Нөхцөлт тархалт

Сэргээн санах нь

B үзэгдэл явагдсан үед A үзэгдэл явагдах магадлал

$$P(A|B)=\frac{P(AB)}{P(B)}$$

Нөхцөлт тархалтыг дараах байдлаар тодорхойлдог.

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

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

Нөхцөлт тархалт ба хамаарал

Сэргээн санах нь

- $P(AB)=P(A)P(B)$ бол A ба B үзэгдлүүдийг хамааралгүй гэнэ.
- $\forall(x,y)\in\mathbb{R}^2$ бүрийн хувьд $f_{X,Y}(x,y)=f_X(x)f_Y(y)$ бол X болон Y хувьсагчдыг хамааралгүй гэнэ.

X болон Y хувьсагчдыг хамааралгүй гэдгийг $X\perp\!\!\!\perp Y$ байдлаар тэмдэглэнэ. Мөн үүнээс дараах нөхцлүүд мөрдөн гарна.

$$f_{X|Y}(x|y)=\frac{f_{X,Y}(x,y)}{f_Y(y)}=\frac{f_X(x)f_Y(y)}{f_Y(y)}=f_X(x)\quad\forall y\in\mathbb{R}$$

$$f_{Y|X}(y|x)=\frac{f_{X,Y}(x,y)}{f_X(x)}=\frac{f_X(x)f_Y(y)}{f_X(x)}=f_Y(y)\quad\forall x\in\mathbb{R}$$

Иймд $f_{X|Y}(x|y)=f_X(x)$ эсвэл $f_{Y|X}(y|x)=f_Y(y)$ нөхцөл биелж байвал хувьсагчдыг хамааралгүй гэж дүгнэнэ.

X_1	X_2: 1	0	$\sum$
1	**0.2**	0.3	0.5
0	0.1	0.4	0.5
$\sum$	**0.3**	0.7	1

X_1	1	0	$\sum$
$f_{X_1\|X_2}(x_1\|X_2=1)$	**2/3**	1/3	1

Хүснэгт: Хамтын, тухайн болон нөхцөлт тархалт

Дискрет санамсаргүй хувьсагчийн нөхцөлт тархалтыг дараах байдлаар олно.

$$f_{X_1|X_2}(X_1=1|X_2=1)=P(X_1=1|X_2=1)=P(\text{эр}|\text{солгой})$$
$$=\frac{P(X_1=1,X_2=1)}{P(X_2=1)}=\frac{P(\text{эр},\text{солгой})}{P(\text{солгой})}=\frac{0.2}{0.3}=\frac{2}{3}$$

Мөн жишээний хувьд

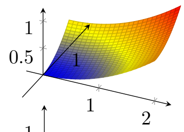
$$f_{X_1|X_2}(X_1=1|X_2=1)=\frac{2}{3}\approx 0.66\neq f_{X_1}(X_1=1)=0.5$$

тул X_1 ба X_2 хувьсагч хамааралтай.

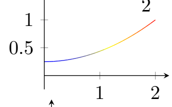
$f_{X,Y}(x,y)=\dfrac{3x^2}{16}+\dfrac{y}{2}$, $0<x<2$, $0<y<1$ хамтын нягтын функцтэй (X,Y) санамсаргүй векторын хувьд $0<x<2$ үеийн $f_{X|Y}(x|y)$ нөхцөлт нягтыг дараах байдлаар олно.

$$f_{X|Y}(x|y)=\frac{f_{X,Y}(x,y)}{f_Y(y)}=\frac{3x^2/16+y/2}{1/2+y}=\frac{3x^2+8y}{8+16y}$$

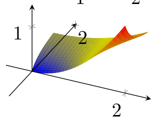
$f_{X|Y}(x|y)=\dfrac{3x^2+8y}{8+16y}$ нөхцөлт тархалт нь нөхцөлд буй Y хувьсагчийн утгаас хамаарсан буюу $f_X(x)=\dfrac{3x^2}{16}+\dfrac{1}{4}$ тухайн тархалтаас ялгаатай байгаа тул эдгээр хувьсагчид хамааралтай юм. Мөн (X,Y) векторын хувьд $f_{X,Y}(x,y)=\dfrac{3x^2}{16}+\dfrac{y}{2}\neq f_X(x)f_Y(y)$ буюу хамтын нягт нь тухайн нягтуудын үржвэрт тавигдахааргүй байгаа явдал нь тус хувьсагчдыг хамааралгүй байж чадахгүйг харуулж байна.



$$f_{X,Y}(x,y)=\frac{3x^2}{16}+\frac{y}{2},\quad 0<x<2,\ 0<y<1$$



$$f_X(x)=\frac{3x^2}{16}+\frac{1}{4},\quad 0<x<2$$



$$f_{X|Y}(x|y)=\frac{f_{X,Y}(x,y)}{f_Y(y)}=\frac{3x^2/16+y/2}{1/2+y}$$
$$=\frac{3x^2+8y}{8+16y},\quad 0<x<2$$

Зураг: Тасралтгүй санамсаргүй хувьсагчийн хамтын тархалт, тухайн тархалт, нөхцөлт тархалт

Хэсэг 2

Бүтэн магадлалын томьёо

Үржүүлэх дүрэм ба бүтэн магадлалын томьёо

X хувьсагч Y хувьсагчаас хамаардаг гэж үзье.

Үржүүлэх дүрэм

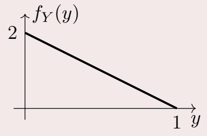
$$f_{X,Y}(x,y)=f_{X|Y}(x|y)f_Y(y)\Longleftarrow f_{X|Y}(x|y)=\frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Бүтэн магадлалын томьёо

$$f_X(x)=\int_{-\infty}^{\infty}f_{X|Y}(x|y)f_Y(y)dy\Longleftarrow\begin{cases}f_{X|Y}(x|y)=\dfrac{f_{X,Y}(x,y)}{f_Y(y)}\\ f_X(x)=\displaystyle\int_{-\infty}^{\infty}f_{X,Y}(x,y)dy\end{cases}$$

Жишээ (Тунгалаг тамир романаас сэдэвлэв.)

Богд Жавзандамба хутагтад сүсэгтэн олноос өргөдөг өргөл барьцын хэмжээ $k = 3$ хэлбэрийн параметр болон $1/\lambda = 1$ масштабын параметр бүхий гамма тархалттай байв. Харин Данигай сойвон өргөл барьцын Y хувийг Богдын санд бүртгээд бусдыг нь хувьдаа



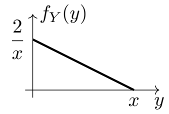
Зураг: Y санамсаргүй хувьсагчийн тархалт

завшдаг бол санд орох өргөл барьцын хэмжээний тархалтыг ол.

Өргөл барьцын анхны хэмжээг X гэвэл бодлогын нөхцөл ёсоор

$$f_X(x) = \frac{1}{2}x^2 e^{-x}, \quad x \geq 0$$

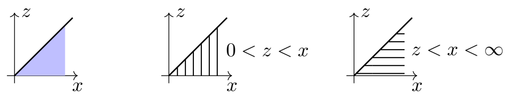
болно. Харин санд орох өргөл барьцын хэмжээ $Z = X \cdot Y$ буюу X хувьсагчаас хамаарна.



Зураг: X хувьсагчийн нөхцөл дэх Z санамсаргүй хувьсагчийн тархалт

$$f_{Z|X}(z|x) = \begin{cases} \dfrac{2(x-z)}{x^2}, & 0 < z < x \\ 0, & \text{бусад} \end{cases}$$

Үржүүлэх дүрмээр $f_{X,Z}(x,z) = f_{Z|X}(z|x) f_X(x) = (x-z)e^{-x}$, $x > 0$, $0 < z < x$ болно.



Зураг: (X, Z) санамсаргүй векторын авах утга

Зургаас харвал $0 < Z < \infty$ ба $Z < X < \infty$ байна. Иймд бүтэн магадлалын томьёогоор дараах илтгэгч тархалт олдоно.

$$f_Z(z) = \int_z^\infty f_{X,Z}(x,z)dx = \int_z^\infty f_{Z|X}(z|x) f_X(x)dx = \int_z^\infty (x-z)e^{-x}dx$$
$$= e^{-z}, \quad z > 0$$

Үржүүлэх дүрэм ба бүтэн магадлалын томьёог үзэгдлүүдийн хувьд дараах байдлаар томьёолдог.

Үржүүлэх дүрэм

$$P(AB) = P(A|B)P(B)$$

Бүтэн магадлалын томьёо

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i) = P(A|B_1)P(B_1) + \ldots + P(A|B_k)P(B_k)$$

Энд $B_1, \ldots, B_k$ харилцан нийцгүй, $B_1 + \ldots + B_k = \Omega$, $P(B_i) > 0$.

Жишээ

Шалгалт зөвхөн нэг нь зөв байдаг дөрвөн хувилбар бүхий сонголттой тест хэлбэртэй асуултуудаас тогтоно. Оюутан шалгалтад бэлдэхдээ хичээлийн сэдвийн 2/3 буюу ойролцоогоор 66 хувийг ойлгож авчээ. Иймд тэр хэрэв мэдэхгүй асуулт таарвал "буудна" гэж шийдэв. Тэгвэл тус оюутан яг одоо тавих асуултад зөв хариулах магадлал ямар байх вэ?

$A = \{\text{зөв хариулах}\}$, $B = \{\text{хариултыг нь мэддэг асуулт таарах}\}$ гэе.

$$P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B})$$
$$= 1 \cdot \frac{2}{3} + \frac{1}{4} \cdot \frac{1}{3}$$
$$= \frac{3}{4} = 0.75$$

Байесын томьёо

Байесын томьёо

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{f_X(x)} \frac{f_Y(y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)} \frac{f_Y(y)}{f_X(x)}$$
$$= \frac{f_{X|Y}(x|y) f_Y(y)}{f_X(x)} = \frac{f_{X|Y}(x|y) f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) dy}$$

$$P(B_i|A) = \frac{P(B_iA)}{P(A)} = \frac{P(B_iA)}{P(A)} \frac{P(B_i)}{P(B_i)} = \frac{P(AB_i)}{P(B_i)} \frac{P(B_i)}{P(A)}$$
$$= \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}$$

Жишээ

Хэрэв оюутан асуултад зөв хариулсан бол тэр уг асуултыг үнэхээр мэддэг байх магадлал ямар байх вэ?

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$
$$= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\overline{B})P(\overline{B})}$$
$$= \frac{1 \cdot \frac{2}{3}}{1 \cdot \frac{2}{3} + \frac{1}{4} \cdot \frac{1}{3}}$$
$$= \frac{8}{9} \approx 0.889$$

Хэсэг 3

Нөхцөлт үл хамаарал

Нөхцөлт үл хамаарал

Сэргээн санах нь

$\forall (x,y)\in\mathbb{R}^2$ хувьд $f_{X,Y}(x,y)=f_X(x)f_Y(y)$ бол X болон Y санамсаргүй хувьсагчдыг *хамааралгүй* гэнэ.

Тодорхойлолт

$\forall (x,y,z)\in\mathbb{R}^3$ хувьд

$$f_{X,Y|Z}(x,y|z)=f_{X|Z}(x|z)f_{Y|Z}(y|z)$$

бол X болон Y санамсаргүй хувьсагчдыг Z хувьсагчийн нөхцөлд хамааралгүй гээд $(X\perp\!\!\!\perp Y)\mid Z$ байдлаар тэмдэглэнэ.

Дараах нөхцлүүд X болон Y санамсаргүй хувьсагч Z хувьсагчийн нөхцөлд хамааралгүй байхтай эквивалент юм.

1. $f_{X|Y,Z}(x|y,z)=f_{X|Z}(x|z)$
2. $f_{X,Y,Z}(x,y,z)=f_X(x)f_{Z|X}(z|x)f_{Y|Z}(y|z)$

2 дугаар чанарын баталгаа

$$\begin{aligned}f_X(x)f_{Z|X}(z|x)f_{Y|Z}(y|z)&=f_X(x)\frac{f_{Z,X}(z,x)}{f_X(x)}f_{Y|Z}(y|z)\\&=f_Z(z)\frac{f_{X,Z}(x,z)}{f_Z(z)}f_{Y|Z}(y|z)=f_Z(z)f_{X|Z}(x|z)f_{Y|Z}(y|z)\\&=f_Z(z)f_{X,Y|Z}(x,y|z)=f_Z(z)\frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)}=f_{X,Y,Z}(x,y,z)\end{aligned}$$

□

Үржүүлэх дүрэм ба нөхцөлт үл хамаарал

Z хувьсагчийн нөхцөлд хамааралгүй X болон Y хоёр хувьсагч авч үзье. Гурван хувьсагчийн хувьд үржүүлэх дүрэм дараах хэлбэртэй.

$$f_{X,Y,Z}(x,y,z)=f_{X|Y,Z}(x|y,z)f_{Y|Z}(y|z)f_Z(z)$$

$(X\perp\!\!\!\perp Y)\mid Z$ үед $f_{X|Y,Z}(x|y,z)=f_{X|Z}(x|z)$ байдаг тул

$$\begin{aligned}f_{X,Y,Z}(x,y,z)&=\underbrace{f_{X|Y,Z}(x|y,z)}_{f_{X|Z}(x|z)}f_{Y|Z}(y|z)f_Z(z)\\&=f_{X|Z}(x|z)f_{Y|Z}(y|z)f_Z(z)\end{aligned}$$

болно.

Нөхцөлт болон нөхцөлт бус үл хамаарал

$$X\perp\!\!\!\perp Y\Rightarrow(X\perp\!\!\!\perp Y)\mid Z$$

боловч үүний урвуу өгүүлбэр нь ерөнхийдөө худал өөрөөр хэлбэл

$$(X\perp\!\!\!\perp Y)\mid Z\nRightarrow X\perp\!\!\!\perp Y$$

юм.

Ингээд гурван хувьсагчийн хувьд нөхцөлт болон нөхцөлт бус үл хамаарлыг дэлгэрэнгүй авч үзье. Гурван хувьсагчийн хувьд холбоо хамаарлын дараах гурван тохиолдол байх боломжтой.

1. $X\longleftarrow Z\longrightarrow Y$
2. $X\longrightarrow Z\longrightarrow Y$
3. $X\longrightarrow Z\longleftarrow Y$

$X\leftarrow Z\rightarrow Y$ тохиолдол

Заасан уялдаа холбоо ёсоор $f_{X,Y,Z}(x,y,z)$ хамтын нягтыг үржүүлэх дүрэм ашиглаж бичвэл

$$f_{X,Y,Z}(x,y,z)=f_{X|Z}(x|z)f_{Y|Z}(y|z)f_Z(z)$$

болно. Одоо z аргументаар интеграл авч $f_{X,Y}(x,y)$ тухайн тархалтыг олъё.

$$f_{X,Y}(x,y)=\int_{-\infty}^{+\infty}f_{X|Z}(x|z)f_{Y|Z}(y|z)f_Z(z)dz$$

Энэ нь ерөнхийдөө $f_X(x)f_Y(y)$ байж чадахгүй тул $X\not\perp\!\!\!\perp Y$ байна.

Харин одоо Z хувьсагчийг нөхцөлд авъя.

$$\begin{aligned}f_{X,Y|Z}(x,y|z)&=\frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)}=\frac{f_{X|Z}(x|z)f_{Y|Z}(y|z)f_Z(z)}{f_Z(z)}\\&=f_{X|Z}(x|z)f_{Y|Z}(y|z)\end{aligned}$$

Эндээс $(X\perp\!\!\!\perp Y)\mid Z$ дүгнэлт гарна.

$X\rightarrow Z\rightarrow Y$ тохиолдол

Заасан уялдааны дагуу $f_{X,Y,Z}(x,y,z)=f_{Y|Z}(y|z)f_{Z|X}(z|x)f_X(x)$ болно. Одоо $f_{X,Y}(x,y)$ тухайн тархалтыг олъё.

$$\begin{aligned}f_{X,Y}(x,y)&=\int_{-\infty}^{+\infty}f_{X,Y,Z}(x,y,z)dz=\int_{-\infty}^{+\infty}f_{Y|Z}(y|z)f_{Z|X}(z|x)f_X(x)dz\\&=f_X(x)\int_{-\infty}^{+\infty}f_{Y|Z}(y|z)f_{Z|X}(z|x)dz=f_X(x)f_{Y|X}(y|x)\end{aligned}$$

Энд бүтэн магадлалын томьёо ашиглав. Сүүлийн илэрхийлэл $f_X(x)f_Y(y)$ биш байгаа тул $X\not\perp\!\!\!\perp Y$ гэж дүгнэнэ.

Одоо Z хувьсагчийг нөхцөлд авъя. Байесын томьёо ашиглавал

$$\begin{aligned}f_{X,Y|Z}(x,y|z)&=\frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)}=\frac{f_{Y|Z}(y|z)f_{Z|X}(z|x)f_X(x)}{f_Z(z)}\\&=f_{X|Z}(x|z)f_{Y|Z}(y|z)\end{aligned}$$

болно. Эндээс $(X\perp\!\!\!\perp Y)\mid Z$ дүгнэлт гарна.

$X\rightarrow Z\leftarrow Y$ тохиолдол

Энэ тохиолдолд Z нь X болон Y хувьсагчдаас хамаарах бөгөөд $X\perp\!\!\!\perp Y$ байна. Үржих дүрмээр

$$f_{X,Y,Z}(x,y,z)=f_{Z|X,Y}(z|x,y)f_{X,Y}(x,y)$$

улмаар $X\perp\!\!\!\perp Y$ болохыг тооцвол

$$f_{X,Y,Z}(x,y,z)=f_{Z|X,Y}(z|x,y)f_X(x)f_Y(y)$$

болно. Эндээс z аргументаар интеграл авлаа ч $f_{X,Y}(x,y)=f_X(x)f_Y(y)$ буюу $X\perp\!\!\!\perp Y$ чанар хадгалагдана.

Харин одоо Z хувьсагч дээр нөхцөл тавих буюу утгыг нь бэхэлье.

$$f_{X,Y|Z}(x,y|z)=\frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)}=\frac{f_{Z|X,Y}(z|x,y)f_X(x)f_Y(y)}{f_Z(z)}$$

Энэ нь ерөнхийдөө $f_X(x)f_Y(y)$ байж чадахгүй тул $(X\not\perp\!\!\!\perp Y)\mid Z$ байна.

Хэсэг 4

Нөхцөлт математик дундаж

Нөхцөлт математик дундаж

$$E(X|Y=y)=\int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx$$

Жишээ

Богдын санд орох өргөл барьцын дундаж хэмжээ өргөл барьцын анхны хэмжээнээс хэрхэн хамаарахыг ол.

Өмнө олсончлон

$$f_{Z|X}(z|x)=\begin{cases}\dfrac{2(x-z)}{x^2}, & 0<z<x\\ 0, & \text{бусад}\end{cases}$$

бас $0<Z\le X$ байх тул

$$E(Z|X=x)=\int_0^x z f_{Z|X}(z|x)dz=\int_0^x z\frac{2(x-z)}{x^2}dz=\frac{x}{3},\quad x>0$$

буюу өргөл барьцын эцсийн хэмжээ нь анхны хэмжээнээс дунджаар 3 дахин багасч байна.

Бүтэн дунджийн томьёо

$$E(E(X|Y))=E(X)$$

Баталгаа

$$\begin{aligned}E(E(X|Y))&=\int E(X|Y)f_Y(y)dy\\&=\int\left(\int xf_{X|Y}(x|y)dx\right)f_Y(y)dy\\&=\int\left(\int xf_{X,Y}(x,y)dy\right)dx\\&=\int x\left(\int f_{X,Y}(x,y)dy\right)dx=\int xf_X(x)dx\\&=E(X)\end{aligned}$$

□

Богдын сангийн жишээг эргэн авч үзье. $f_Z(z)=e^{-z}$, $z>0$ буюу $Z\sim\text{Exp}(\lambda=1)$ илтгэгч тархалттай байсан бас илтгэгч тархалтын математик дундаж нь $1/\lambda$ тул бүтэн дунджийн томьёо ёсоор

$$E(E(Z|X))=E(Z)=1$$

болно. Нөгөө талаас

$$E(Z|X=x)=\frac{x}{3},\quad x>0$$

гэж олдсон, $X\sim\text{Gamma}(\lambda=1,k=3)$ гэж өгсөн бас гамма тархалтын хувьд

$$X\sim\text{Gamma}(\lambda,k)\text{ бол }Y=cX\sim\text{Gamma}(\lambda/c,k)$$

чанар байдаг мөн дундаж нь $E(X)=\dfrac{k}{\lambda}$ зэргийг тооцвол

$$E(E(Z|X=x))=E\left(\frac{X}{3}\right)=\frac{k=3}{\frac{\lambda=1}{c=\frac{1}{3}}}=1$$

болно.

Санамсаргүй тоо ширхэг бүхий санамсаргүй хувьсагчдын нийлбэрийн дундаж

Жишээ

Богдын санд өргөх өргөл барьцын тоо $T\sim\text{Pois}(\lambda=50)$ тархалттай бол сангийн нийт орлогын дундаж утгыг ол.

Өмнө олсончлон нэг өргөл барьцаас орох дундаж орлого $E(Z)=1$ байсан. Иймд өргөл барьцын тооноос хамаарсан нийт орлогын нөхцөлт математик дундаж

$$E(S|T)=E(Z_1+\ldots+Z_T)=E(Z)T$$

байна. Энд Z_i нь i дүгээр өргөл барьцаас орох дундаж орлого юм. Бүтэн дунджийн томьёогоор нийт орлогын математик дундаж

$$E(S)=E(E(S|T))=E(E(Z)T)=E(Z)E(T)=1\cdot 50=50$$

болно.

Санамсаргүй тоо ширхэг бүхий санамсаргүй хувьсагчдын нийлбэрийн дисперс ба бүтэн дисперсийн томьёо

Бүтэн дисперсийн томьёо

$$D(X)=E(D(X|Y))+D(E(X|Y))$$

Богдын сантай жишээнд буй $S=Z_1+\ldots+Z_T$ санамсаргүй хувьсагч буюу ижил тархалттай санамсаргүй тоо ширхэг бүхий хамааралгүй хувьсагчдын нийлбэрийн дисперсийг дараах байдлаар олно.

$$\begin{aligned}D(S)&=E(D(S|T))+D(E(S|T))=E(D(Z_1+\ldots+Z_T))+D(E(Z)T)\\&=E(D(Z_1)+\ldots+D(Z_T))+[E(Z)]^2D(T)\\&=E(D(Z)T)+[E(Z)]^2D(T)=D(Z)E(T)+[E(Z)]^2D(T)\\&=\frac{1}{1^2}\cdot 50+[1]^2 50=100\end{aligned}$$

Энд илтгэгч тархалтын дисперс $1/\lambda^2$ байдгийг ашиглав.

Лекц IX

Олон хэмжээст хэвийн тархалт ба шугаман загвар

Олон хэмжээст хэвийн тархалт ба шугаман загвар сэдвийн агуулга

Хэсэг 1

Санамсаргүй векторын математик дундаж ба ковариацын матриц

Санамсаргүй векторын математик дундаж

$X = (X_1, \ldots, X_p)^T$ буюу p хэмжээст санамсаргүй вектор авч үзье. Тэгвэл түс векторын математик дунджийг дараах байдлаар тодорхойлно.

$$\mu = E(X) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix}$$

Жишээ

X_1 \ X_2	-1	1
0	0.1	0.0
1	0.5	0.4

(a) Хамтын тархалт

X_1 \ X_2	-1	1	$\sum$
0	0.1	0.0	0.1
1	0.5	0.4	0.9
$\sum$	0.6	0.4	1

(b) Хамтын тархалт ба тухайн тархалт

X_1	0	1
$f_{X_1}(x_1)$	0.1	0.9
X_2	-1	1
$f_{X_2}(x_2)$	0.6	0.4

(c) Тухайн тархалт

Хүснэгт: Хамтын тархалт ба тухайн тархалт

Хүснэгтээр өгсөн тархалттай дискрет санамсаргүй векторын математик дунджийг ол.

$$E(X_1) = 0 \cdot 0.1 + 1 \cdot 0.9 = 0.9$$
$$E(X_2) = -1 \cdot 0.6 + 1 \cdot 0.4 = -0.2$$
$$E(X) = \begin{pmatrix} 0.9 \\ -0.2 \end{pmatrix}$$

Жишээ

$$f_{X,Y}(x,y) = \begin{cases} \frac{3x^2}{16} + \frac{y}{2}, & 0 < x < 2,\ 0 < y < 1 \\ 0, & \text{бусад} \end{cases}$$

хамтын нягттай (X, Y) векторын математик дунджийг ол.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_0^1 \left(\frac{3x^2}{16} + \frac{y}{2}\right) dy = \frac{3x^2}{16} + \frac{1}{4}$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_0^2 \left(\frac{3x^2}{16} + \frac{y}{2}\right) dx = y + \frac{1}{2}$$
$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^2 x\left(\frac{3x^2}{16} + \frac{1}{4}\right) dx = \frac{5}{4}$$
$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy = \int_0^1 y\left(y + \frac{1}{2}\right) dy = \frac{7}{12}$$

Хоёр санамсаргүй хувьсагчийн ковариац ба корреляц

X болон Y санамсаргүй хувьсагчдын ковариацын коэффициентыг

$$\operatorname{cov}(X,Y) = E[(X - E(X))(Y - E(Y))]$$

гэж тодорхойлох бөгөөд хувьсагчдын тасралтгүй ба дискрет байдлаас нь хамаарч дараах байдлаар тооцоолно.

$$\operatorname{cov}(X,Y) = \begin{cases} \sum_{(x,y)} (x - E(X))(y - E(Y)) f_{X,Y}(x,y) & \text{дискрет} \\ \int_{\mathbb{R}^2} (x - E(X))(y - E(Y)) f_{X,Y}(x,y)dxdy & \text{тасралтгүй} \end{cases}$$

Харин корреляцын коэффициентыг дараах байдлаар тодорхойлдог.

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{D(X)D(Y)}}$$

Ковариац болон корреляцын чанар

Чанар

1. $\operatorname{cov}(X,X) = D(X)$
2. $\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y)$
3. X болон Y хамааралгүй бол $\operatorname{cov}(X,Y) = 0$ байна.
4. $-1 \leq \rho(X,Y) \leq 1$
5. $Y = a + bX$, $a, b \in \mathbb{R}$ ба $b > 0$ бол $\rho(X,Y) = 1$ харин $b < 0$ бол $\rho(X,Y) = -1$ байна.

Жишээ

Өмнөх жишээнүүдэд авч үзсэн санамсаргүй векторууд дахь хувьсагчдын ковариацын коэффициентыг ол.

X_1 \ X_2	-1	1
0	0.1	0.0
1	0.5	0.4

Үүнийг ковариацын 2 дугаар чанар болон ухамсаргүй статистикчийн хуулъ ашиглаж олъё.

$$\operatorname{cov}(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$$
$$= \sum_{(x_1,x_2)} x_1x_2 f_{X_1,X_2}(x_1,x_2) - E(X_1)E(X_2)$$
$$= 0 \cdot (-1) \cdot 0.1 + 0 \cdot 1 \cdot 0.0 + 1 \cdot (-1) \cdot 0.5 + 1 \cdot 1 \cdot 0.4 - E(X_1)E(X_2)$$
$$= -0.1 - 0.9 \cdot (-0.2) = 0.08$$

$$f_{X,Y}(x,y) = \begin{cases} \frac{3x^2}{16} + \frac{y}{2}, & 0 < x < 2,\ 0 < y < 1 \\ 0, & \text{бусад} \end{cases}$$

Тасралтгүй тохиолдолд зарчмын хувьд өмнөхтэй төстэй байдлаар

$$\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y) = \int_D xy f_{X,Y}(x,y)dxdy - \frac{5}{4} \cdot \frac{7}{12}$$
$$= \int_0^1\int_0^2 xy\left(\frac{3x^2}{16} + \frac{y}{2}\right) dxdy - \frac{35}{48} = \int_0^1 y\left[\frac{3x^4}{64} + \frac{x^2y}{4}\right]_0^2 dy - \frac{35}{48}$$
$$= \int_0^1 \left(\frac{3}{4}y + y^2\right) dy - \frac{35}{48} = \frac{17}{24} - \frac{35}{48} = -\frac{1}{48}$$

гэж бодно. Энд $D = \{(x,y) : 0 < x < 2,\ 0 < y < 1\}$ байна.

Санамсаргүй векторын ковариацын матриц

$$\Sigma_{XX} = D(X) = \operatorname{cov}(X,X) = E[(X-\mu)(X-\mu)^T]$$
$$= \begin{pmatrix} D(X_1) & \operatorname{cov}(X_1,X_2) & \cdots & \operatorname{cov}(X_1,X_p) \\ \operatorname{cov}(X_2,X_1) & D(X_2) & \cdots & \operatorname{cov}(X_2,X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_p,X_2) & \operatorname{cov}(X_p,X_2) & \cdots & D(X_p) \end{pmatrix}$$

Чанар

Σ_{XX} тэгш хэмтэй, эерэг тодорхойлогдсон матриц байна.

X болон Y нь харгалзан p болон q хэмжээст санамсаргүй вектор бол

$$\Sigma_{XY} = \operatorname{cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)^T]$$

нь $p \times q$ хэмжээст матриц байна.

Ковариацын матриц ба корреляцын коэффициент

Жишээ

(X, Y) хоёр хэмжээст санамсаргүй векторын ковариацын матриц $\begin{pmatrix} 25 & 4 \\ 4 & 1 \end{pmatrix}$ бол $\rho(X, Y)$ корреляцын коэффициент болон корреляцын матрицыг олно.

$\begin{pmatrix} 25 & 4 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} D(X) & \operatorname{cov}(X, Y) \\ \operatorname{cov}(Y, X) & D(Y) \end{pmatrix}$ байх тул

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{D(X)D(Y)}} = \frac{4}{\sqrt{25 \cdot 1}} = \frac{4}{5} = 0.8$$

байна. Энэ тохиолдолд корреляцын матриц нь

$$\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

болно.

Хэсэг 2

Олон хэмжээст хэвийн тархалт

Олон хэмжээст хэвийн тархалт

$$f_X(x) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right\}$$

нягттай тархалтыг *олон хэмжээст хэвийн тархалт* гэнэ. Энд $x = (x_1, \dots, x_p)^T \in \mathbb{R}^p$ байна. X санамсаргүй векторыг μ болон Σ параметр бүхий p хэмжээст хэвийн тархалттай гэхийг

$$X \sim N_p(\mu, \Sigma)$$

гэж тэмдэглэнэ.

Хувьсагчид хамааралгүй байх тохиолдол

$X_1, \dots, X_p$ хувьсагчид хамааралгүй бол

$$\Sigma = \begin{pmatrix} \operatorname{cov}(X_1, X_1) & 0 & \dots & 0 \\ 0 & \operatorname{cov}(X_2, X_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \operatorname{cov}(X_p, X_p) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p^2 \end{pmatrix}$$

байх ба улмаар

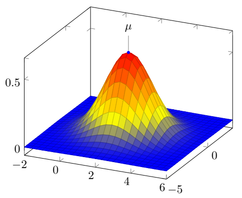
$$f_X(x) = \prod_{i=1}^{p} f_{X_i}(x_i) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{1}{2\sigma_i^2}(x - \mu_i)^2\right\}$$

болно.

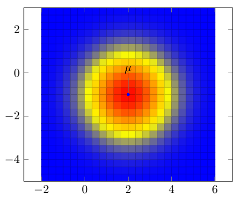
$\mu = (2, -1)^T$, $\Sigma = \begin{pmatrix} 2/3 & 0 \\ 0 & 2/3 \end{pmatrix}$ параметртэй

$$f_{X_1,X_2}(x_1, x_2) = \frac{3}{2\sqrt{2\pi}} \exp\left\{-\frac{(x_1 - 2)^2 + (x_2 + 1)^2}{3}\right\}, \quad (x_1, x_2) \in \mathbb{R}^2$$

хэвийн тархалтын нягтыг зурагт дүрслэн харуулав.



(a) хажуугаас



(b) эгц дээрээс

Зураг: Хоёр хэмжээст хэвийн тархалтын нягт, хувьсагчид хамааралгүй бөгөөд ижил дисперстэй үед

Хоёр хэмжээст хэвийн тархалт

ρ корреляцтай X_1 ба X_2 санамсаргүй хувьсагчдаас тогтох $X = (X_1, X_2)$ санамсаргүй векторын хоёр хэмжээст хэвийн тархалтын тэмдэглэгээг дэлгэрэнгүй бичвэл

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right]$$

хэлбэртэй байна. Харин хамтын нягтын функц нь

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} \exp\left\{-\frac{1}{2(1 - \rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1 - \mu_1}{\sigma_1}\right)\left(\frac{x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2\right]\right\}$$

болно.

Жишээ

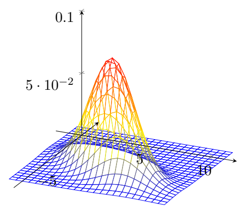
$\mu = (5, -4)$ дундаж утгын вектор болон $\Sigma = \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix}$ ковариацын матриц бүхий хоёр хэмжээст хэвийн тархалтын хамтын нягтын илэрхийллийг бичиж, графикийг нь зур.

$\mu = (\mu_1, \mu_2) = (5, -4)$ ба $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix}$ гэдгээс $\sigma_1 = 2$, $\sigma_2 = 1$, $\rho = -1/2$ болно. Иймд нягт нь

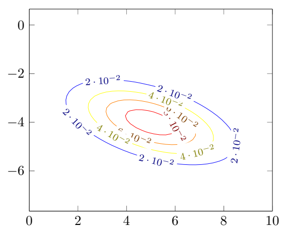
$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\sqrt{3}\pi} \exp\left\{-\frac{1}{6}\left[(x_1 - 5)^2 + 2(x_1 - 5)(x_2 + 4) + 4(x_2 + 4)^2\right]\right\}$$

болно. Харин графикийг нь дараагийн слайд дээр байгуулж үзүүлье.

Жишээ болгон авсан хоёр хэмжээст хэвийн тархалтын нягтын функцийн график, түүний түвшний шугамыг дараах зургаар харууллаа.



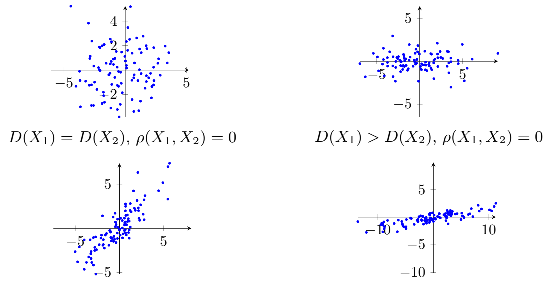
нягтын функцийн график



нягтын функцийн түвшний шугам

Зураг: Хоёр хэмжээст хэвийн тархалтын нягт, хувьсагчид хамааралтай бөгөөд дисперс нь ялгаатай үед

Ковариацын матриц тархалтад нөлөөлөх байдал



$D(X_1) = D(X_2)$, $\rho(X_1, X_2) = 0$ $D(X_1) > D(X_2)$, $\rho(X_1, X_2) = 0$

$D(X_1) = D(X_2)$, $\rho(X_1, X_2) \neq 0$ $D(X_1) > D(X_2)$, $\rho(X_1, X_2) \neq 0$

Зураг: Хоёр хэмжээст хэвийн тархалттай санамсаргүй утгууд

Санамсаргүй векторын нэг хуваалт

$X = (X_1, \dots, X_p)$ санамсаргүй векторыг дараах байдлаар хоёр дэд векторт хуваая.

$$X \left\{ \begin{array}{c} \left. \begin{array}{c} X_1 \\ \vdots \\ X_r \end{array} \right\} X_1 \\ \left. \begin{array}{c} X_{r+1} \\ \vdots \\ X_p \end{array} \right\} X_2 \end{array} \right.$$

Өөрөөр хэлбэл $X_1 = (X_1, \dots, X_r)$ ба $X_2 = (X_{r+1}, \dots, X_p)$ гэе. Тэгвэл ковариацын матриц

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

блок матриц болно. Энд $\Sigma_{11} = \mathrm{cov}(X_1, X_1)$, $\Sigma_{22} = \mathrm{cov}(X_2, X_2)$, $\Sigma_{12} = \mathrm{cov}(X_1, X_2)$, $\Sigma_{21} = \Sigma_{12}^T = \mathrm{cov}(X_2, X_1)$ байна.

Жишээ

$$X \sim N_3\left(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 2 \end{pmatrix} \right)$$

$r = 2$ үед $X = (X_1, X_2, X_3)$ санамсаргүй вектор $X_1 = (X_1, X_2)$ ба $X_2 = (X_3)$ гэсэн хоёр дэд вектор болж задарна.
Ковариацын матриц нь дараах блок матриц болно.

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \left(\begin{array}{cc|c} 1 & 2 & 1 \\ 2 & 5 & 2 \\ \hline 1 & 2 & 2 \end{array} \right)$$

$$\Sigma_{11} = \mathrm{cov}(X_1, X_1) = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix} \quad \Sigma_{12} = \mathrm{cov}(X_1, X_2) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$
$$\Sigma_{21} = \mathrm{cov}(X_2, X_1) = \begin{pmatrix} 1 & 2 \end{pmatrix} \quad \Sigma_{22} = \mathrm{cov}(X_2, X_2) = \begin{pmatrix} 2 \end{pmatrix}$$

Мөн $\mu_1 = (0, 0)$, $\mu_2 = (1)$ болно.

Тухайн тархалт

Олон хэмжээст хэвийн тархалттай санамсаргүй векторыг сая үзсэн шиг хуваавал

$$X_1 \sim N_r(\mu_1, \Sigma_{11}) \quad X_2 \sim N_{p-r}(\mu_2, \Sigma_{22})$$

буюу олон хэмжээст хэвийн тархалтын тухайн тархалт нь мөн адил олон хэмжээст хэвийн тархалт байна.

Нөхцөлт тархалт

Олон хэмжээст хэвийн тархалттай хувьсагчдын нөхцөлт тархалт

$$(X_2 | X_1 = x_1) \sim N_{p-r}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$$

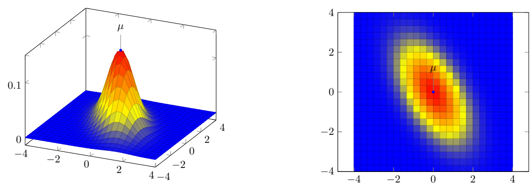
буюу мөн адил олон хэмжээст хэвийн тархалт байна. Харин олон хэмжээст хэвийн тархалттай санамсаргүй хувьсагчдын нөхцөлт математик дундаж болон нөхцөлт ковариацын матриц дараах хэлбэртэй байдаг.

$$E(X_2 | X_1 = x_1) = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 + \Sigma_{21}\Sigma_{11}^{-1}x_1$$
$$\mathrm{cov}(X_2 | X_1 = x_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

Жишээ

$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 2 \end{pmatrix}$ параметртэй хэвийн тархалтын хувьд $\mathrm{cov}(X_2 | X_1 = x_1)$ нөхцөлт дисперс, $E(X_2 | X_1 = x_1)$ нөхцөлт дундаж, $f_{X_1}(x_1)$ тухайн тархалт болон $f_{X_2|X_1}(x_2|x_1)$ нөхцөлт тархалтыг ол.

Энэ тохиолдолд $\rho(X_1, X_2) \approx -0.566$ байна.



(a) хөндлөнгөөс (b) эгц дээрээс

Зураг: Хоёр хэмжээст хэвийн тархалтын нягт

$r = 1$ гэвэл $\Sigma = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 2 \end{pmatrix}$ матрицаас $\Sigma_{11} = 1$, $\Sigma_{22} = 2$, $\Sigma_{12} = \Sigma_{21} = -0.8$ гэж олдох тул

$$\mathrm{cov}(X_2 | X_1 = x_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = 2 - (0.8)^2 = 1.36$$

бас $\mu_1 = 0$, $\mu_2 = 0$ тул

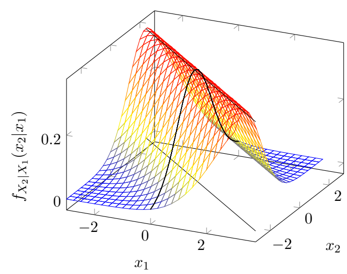
$$E(X_2 | X_1 = x_1) = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 + \Sigma_{21}\Sigma_{11}^{-1}x_1 = -0.8x_1$$

улмаар

$$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2} \right)$$

$$f_{X_2|X_1}(x_2|x_1) = \frac{1}{\sqrt{2\pi 1.36}} \exp\left\{ -\frac{(x_2 + 0.8x_1)^2}{2(1.36)} \right\}$$

гэж олдоно.



Зураг: $f_{X_2|X_1}(x_2|x_1) = \frac{1}{\sqrt{2\pi 1.36}} \exp\left\{ -\frac{(x_2 + 0.8x_1)^2}{2(1.36)} \right\}$ нөхцөлт нягтын муруй

Хэсэг 3

Регрессийн шугаман загвар

Санамсаргүй хувьсагчийн утгыг нөхцөлт математик дундаж ашиглаж прогнозлох нь

$$X_2 = E(X_2|X_1) + U$$

Энд U нь прогнозын алдаа юм.

Чанар

1. $E(U|X_1) = 0$
2. $E(U) = 0$
3. $\operatorname{cov}(E(X_2|X_1), U) = 0$
4. $E(X_2|X_1)$ нь X_1 хувьсагч ашиглаж X_2 хувьсагчийг прогнозлох бүх $h(X_1) : \mathbb{R}^r \to \mathbb{R}^{p-r}$ функц дундаас хамгийн бага дундаж квадрат алдаатай (MSE $= E\left\{(X_2 - h(X_1))^T(X_2 - h(X_1))\right\}$) нь юм.

1 дүгээр чанарын баталгаа

$X_2 = E(X_2|X_1) + U$ прогнозын хоёр талаас нөхцөлт дундаж аваад

1. нөхцөлт математик дунджийн шугаман чанар
2. $E(\varphi(X_1)X_2|X_1) = \varphi(X_1)E(X_2|X_1)$
3. X_1 ба X_2 хамааралгүй бол $E(X_2|X_1) = E(X_2)$
4. c тогтмол бол $E(c) = c$

чанар ашиглавал

$$E(X_2|X_1) = E(E(X_2|X_1) + U|X_1)$$
$$E(X_2|X_1) = E(E(X_2|X_1)|X_1) + E(U|X_1)$$
$$E(X_2|X_1) = E(X_2|X_1)E(1|X_1) + E(U|X_1)$$
$$E(X_2|X_1) = E(X_2|X_1)E(1) + E(U|X_1)$$
$$E(X_2|X_1) = E(X_2|X_1) + E(U|X_1)$$
$$E(U|X_1) = 0$$

болж тус чанар батлагдана. □

Регрессийн шугаман загвар

Өмнө үзсэнчлэн хэрэв X_1 ба X_2 санамсаргүй векторууд олон хэмжээст хэвийн тархалттай бол $E(X_2|X_1 = x_1)$ нь x_1 хувьсагчаас шугаман байдлаар хамаарсан функц болно. Иймд X_2 хувьсагчийг X_1 хувьсагчийн шугаман эвлүүлгийн тусламжтай прогнозлох боломжтой. Ийм загварыг *регрессийн шугаман загвар* гэнэ.

X_1 болон X_2 санамсаргүй векторууд хамтдаа олон хэмжээст хэвийн тархалттай бол

$$X_2 = E(X_2|X_1) + U = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1) + U = \beta_0 + BX_1 + U$$

болно. Энд $B = \Sigma_{21}\Sigma_{11}^{-1}$, $\beta_0 = \mu_2 - B\mu_1$, $U \sim N_{p-r}(0, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$ байна. Уг загвар нь тухайлбал $r = p - 1$ тохиолдолд $X_2 = \beta_0 + \beta^T X_1 + U$ ба энд $\beta^T = B_{(1\times r)}$ нь мөр вектор байх тул

$$X_p = \beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r + U$$

хэлбэртэй болно. Жишээлбэл $X = (X_1, X_2, X_3)$ буюу $p = 3$ үед $r = 3 - 1 = 2$ тул $X_1 = (X_1, X_2)$ ба $X_2 = (X_3)$ болж улмаар

$$X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

загвар гарна.

Шугаман загварын детерминацын коэффициент

Тодорхойлолт (Детерминацын коэффициент)

$$\underbrace{\operatorname{cov}(X_2)}_{\text{нийт ковариац}} = \underbrace{\operatorname{cov}(\beta_0 + BX_1)}_{\text{тайлбарлагдах ковариац}} + \underbrace{\operatorname{cov}(U)}_{\text{үл тайлбарлагдах ковариац}}$$

$$\rho^2 = \frac{\text{тайлбарлагдах ковариац}}{\text{нийт ковариац}}$$

Детерминацын коэффициентыг $r = p - 1$ тохиолдолд олж гаргавал

$$\rho^2 = \frac{\operatorname{cov}(\beta_0 + \beta^T X_1)}{\operatorname{cov}(X_2)} = \frac{\beta^T \operatorname{cov}(X_1)\beta}{\operatorname{cov}(X_2)} = \frac{\sigma_{21}\Sigma_{11}^{-1}\sigma_{12}}{\sigma_{22}}$$

болно. Энд $\sigma_{22} = \Sigma_{22} = DX_2$ нь скаляр, $\sigma_{21} = \Sigma_{21}$ нь r ширхэг компоненттой мөр вектор, $\sigma_{12} = \Sigma_{12}$ нь r ширхэг компоненттой багана вектор байна.

Жишээ

$\Sigma = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 2 \end{pmatrix}$ бол $X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$ шугаман загварын детерминацын коэффициентыг ол.

$r = 2$ тул

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \left(\begin{array}{cc|c} 1 & 2 & 1 \\ 2 & 5 & 2 \\ \hline 1 & 2 & 2 \end{array}\right)$$

буюу $\Sigma_{11} = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$, $\sigma_{12} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $\sigma_{21} = \begin{pmatrix} 1 & 2 \end{pmatrix}$, $\sigma_{22} = 2$ бас

$\Sigma_{11}^{-1} = \begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix}$ болж улмаар

$$\rho^2 = \frac{\sigma_{21}\Sigma_{11}^{-1}\sigma_{12}}{\sigma_{22}} = 0.5$$

үр дүнд хүрнэ.

Нөхцөлт ковариацын матриц, тухайн корреляц ба нөхцөлт үл хамаарал

Өмнө үзсэн

$$\operatorname{cov}(X_2|X_1 = x_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

нөхцөлт ковариацын матрицыг $r = p - 2$ үед авч үзье. Энэхүү хоёр хэмжээст ковариацын матрицаас олдох "нөхцөлт" буюу *тухайн корреляц* нь нөхцөлд авсан хувьсагчдын нөлөөг нөгөө хоёр хувьсагчийн холбоо хамаарлаас зайлуулсан үеийн корреляц юм. Хэрэв тус хоёр хувьсагч угтаа хамааралгүй буюу бусад хувьсагчдын дам нөлөөгөөр холбогдож байсан бол тухайн корреляц нь тэгтэй тэнцүү гарна. Тийнхүү тухайн корреляц тэгтэй тэнцүү гарах нь өмнө үзсэн нөхцөлт үл хамаарал байгааг илтгэж буй явдал юм.

Жишээ

$\Sigma = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 5 & 2 \\ 2 & 2 & 3 \end{pmatrix}$ бол X_2 ба X_3 хоёр хувьсагч X_1 хувьсагчийн нөхцөлд хамааралгүй болохыг харуул.

X_2 ба X_3 хувьсагчдын нөхцөлт бус корреляц $\rho(X_2, X_3) = \dfrac{2}{\sqrt{5 \cdot 3}} \approx 0.516 \neq 0$ гэж олдоно. Харин нөхцөлт буюу тухайн корреляц нь

$$\left.\begin{matrix} X_1 \end{matrix}\right\} X_1 \qquad \Sigma_{11} = (2) \quad \Sigma_{12} = \begin{pmatrix} 2 & 2 \end{pmatrix} \quad \Sigma_{22} = \begin{pmatrix} 5 & 2 \\ 2 & 3 \end{pmatrix}$$

$$\left.\begin{matrix} X_2 \\ X_3 \end{matrix}\right\} X_2 \qquad \operatorname{cov}(X_2 | X_1 = x_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

улмаар $\rho(X_2, X_3 | X_1) = \dfrac{0}{\sqrt{3 \cdot 1}} = 0$ буюу X_2 ба X_3 хоёр хувьсагч X_1 хувьсагчийн нөхцөлд хамааралгүй ажээ.

Тухайн корреляц олох бусад аргууд

1. Регрессийн шугаман загвар ашиглах
2. Ковариацын матрицын урвуу ашиглах

$$\rho_{X_i, X_j | \text{бусад}} = -\frac{p_{ij}}{\sqrt{p_{ii}}\sqrt{p_{jj}}}$$

Энд p_{ij} нь $P = \Sigma^{-1}$ матрицын i дүгээр мөр болон j дүгээр баганын огтлолцолд байх элемент юм.

Жишээ

Өмнөх жишээний хувьд тухайн ковариацын матриц болон $\rho_{X_1, X_2 | X_3}$ тухайн корреляцыг ол.

$$P = \Sigma^{-1} = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 5 & 2 \\ 2 & 2 & 3 \end{pmatrix}^{-1} = \begin{pmatrix} 11/6 & -1/3 & -1 \\ -1/3 & 1/3 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

ба улмаар $\rho_{X_1, X_2 | X_3} = -\dfrac{-1/3}{\sqrt{11/6}\sqrt{1/3}} = \sqrt{2/11} \approx 0.4264$ гэж олдоно.

Лекц X

Хамааралгүй санамсаргүй хувьсагчдын нийлбэрийн тархалт

Хамааралгүй санамсаргүй хувьсагчдын нийлбэрийн тархалт сэдвийн агуулга

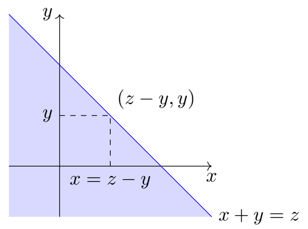
Хэсэг 1

Хамааралгүй санамсаргүй хувьсагчдын нийлбэрийн тархалт

Хамааралгүй хувьсагчдын нийлбэрийн тархалт

$Z = X + Y$ гэе.

$$F_Z(z) = P(Z < z) = P(X + Y < z) = ?$$



Зураг: $Z = X + Y$ санамсаргүй хувьсагчийн тархалтын функцтэй холбогдох үзэгдэл

Бүтэн магадлалын томьёоны янз бүрийн хэлбэрүүд

Сэргээн санах нь (Бүтэн магадлалын томьёо)

- $f_X(x) = \displaystyle\int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) dy$
- $P(A) = \displaystyle\sum_i P(A|B_i)P(B_i)$

$$P(A) = \int_{-\infty}^{\infty} P(A | Y = y) f_Y(y) dy$$

Хамааралгүй хувьсагчдын нийлбэрийн тархалт ба хуниас

X болон Y хамааралгүй байг.

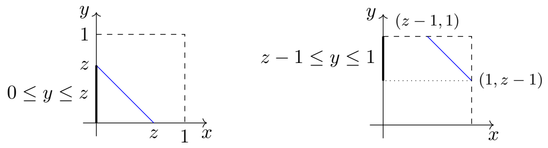
$$\begin{aligned} F_{X+Y}(z) &= P(X + Y < z) \\ &= \int_{-\infty}^{\infty} P(X + Y < z \mid Y = y) f_Y(y) dy && \text{бүтэн магадлалын томьёо} \\ &= \int_{-\infty}^{\infty} P(X < z - y \mid Y = y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} P(X < z - y) f_Y(y) dy && \text{хамааралгүй} \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z-y} f_X(x)\,dx \right] f_Y(y) dy \end{aligned}$$

$$f_{X+Y}(z) = F'_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) dy = f_X \star f_Y$$

Үүнийг f_X болон f_Y тархалтуудын *хуниас* гэнэ.

Жишээ

X болон Y хувьсагчид $(0,1)$ завсарт жигд тархалттай бөгөөд хамааралгүй бол $X+Y$ нийлбэрийн тархалтыг олъё.



(a) $0 \le Z \le 1$ байх тохиолдол (b) $1 \le Z \le 2$ байх тохиолдол

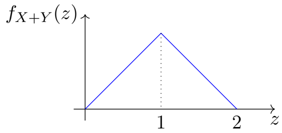
Зураг: Жигд тархалттай хувьсагчдын $Z = X + Y$ нийлбэр

- $0 \le Z \le 1$ тохиолдол

$$f_{X+Y}(z) = \int_0^z \underbrace{f_X(z-y)}_{1}\underbrace{f_Y(y)}_{1}\,dy = z$$

- $1 \le Z \le 2$ тохиолдол

$$f_{X+Y}(z) = \int_{z-1}^1 1dy = 2 - z$$



Зураг: $(0,1)$ завсарт жигд тархалттай хамааралгүй хувьсагчдын $X+Y$ нийлбэрийн тархалт

Жишээ

$X, Y \sim \mathrm{Exp}(\lambda)$ хамааралгүй хувьсагчдын $X+Y$ нийлбэрийн тархалтыг олъё.

$0 \le X + Y \le z$ байх тул

$$\begin{aligned} f_{X+Y}(z) &= \int_0^z f_X(z-y)f_Y(y)dy \\ &= \int_0^z \lambda e^{-\lambda(z-y)}\lambda e^{-\lambda y}dy = \lambda^2 e^{-\lambda z}\int_0^z dy = \lambda^2 z e^{-\lambda z} \end{aligned}$$

гэж гарах ба энэ нь $\mathrm{Gamma}(k=2,\lambda)$ тархалтын нягтын илэрхийлэл байна. Иймд $X+Y \sim \mathrm{Gamma}(k=2,\lambda)$ боллоо.

Дискрет тархалтуудын хуниас

X болон Y хувьсагчид хамааралгүй бөгөөд бүхэл тоон утга авдаг бас $X+Y$ бүхэл тоон утгатай байг.

$$\begin{aligned} f_{X+Y}(z) &= P(X+Y=z) \\ &= \sum_k P(X=z-k, Y=k) \qquad \text{бүтэн магадлалын томьёо} \\ &= \sum_k P(X=z-k)P(Y=k) \qquad \text{хамааралгүй} \\ &= \sum_k f_X(z-k)f_Y(k) \end{aligned}$$

Жишээ

X болон Y хувьсагчид хамааралгүй бөгөөд харгалзан λ_X болон λ_Y параметр бүхий Пуассоны тархалттай бол $X+Y$ нийлбэрийн тархалтыг ол.

$X \ge 0$ ба $Y \ge 0$ болохыг анхаарвал

$$\begin{aligned} f_{X+Y}(z) &= \sum_{k=0}^z f_X(z-k)f_Y(k) \\ &= \sum_{k=0}^z \frac{\lambda_X^{z-k}}{(z-k)!}e^{-\lambda_X}\frac{\lambda_Y^k}{k!}e^{-\lambda_Y} \\ &= e^{-(\lambda_X+\lambda_Y)}\frac{1}{z!}\sum_{k=0}^z C_z^k \lambda_X^{z-k}\lambda_Y^k \\ &= \frac{(\lambda_X+\lambda_Y)^z}{z!}e^{-(\lambda_X+\lambda_Y)} \end{aligned}$$

буюу $\lambda_X + \lambda_Y$ параметр бүхий Пуассоны тархалт олдож байна.

Хэсэг 2

Хязгаарын гол теорем

Ижил тархалттай, хамааралгүй санамсаргүй хувьсагчид ба тэдгээрийн нийлбэрийн тархалт

$X_1, \ldots, X_n$ хувьсагчид нэг ижил μ дундаж, σ^2 дундаж квадрат хазайлттай бас хамааралгүй гэж тооцъё. Хойшид энэ нөхцөлийг

$$X_1, \ldots, X_n \sim IID^{1}(\mu, \sigma^2)$$

байдлаар тэмдэглэж байя.

Хэрэв $S_n = X_1 + \ldots + X_n$ гэвэл үүний нягт нь n удаа нугалсан хуниас байна.

$$f_{S_n}(x) = (f_{X_1} \star \ldots \star f_{X_n})(x)$$

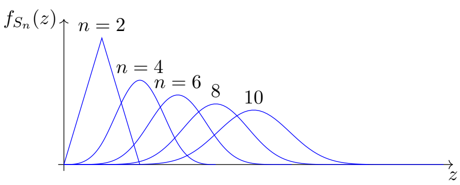
Мөн $n \to \infty$ үед энд ямар нэг асимптот буюу хязгаарын тархалт оршин байх уу?

[1] independent identically distributed

Жигд тархалттай, хамааралгүй хувьсагчдын нийлбэр

$X_1, \ldots, X_n \sim U(0,1)$ хамааралгүй хувьсагчдын $S_n = X_1 + \ldots + X_n$ нийлбэрийн тархалт дараах хэлбэртэй байдаг.

$$f_{S_n}(x) = \begin{cases} \frac{1}{(n-1)!}\sum\limits_{0 \le j \le x}(-1)^j C_n^j (x-j)^{n-1}, & 0 < x < n \\ 0, & \text{бусад} \end{cases}$$

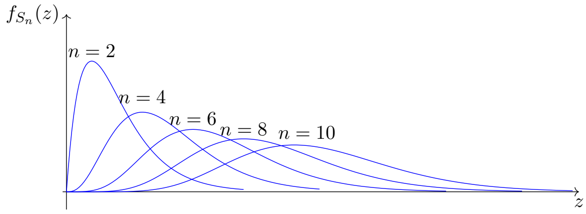


Зураг: $U(0,1)$ тархалттай, хамааралгүй хувьсагчдын нийлбэрийн тархалт

Илтгэгч тархалттай, хамааралгүй хувьсагчдын нийлбэр

$X_1, \ldots, X_n \sim \text{Exp}(\lambda)$ хамааралгүй бол

$$S_n = X_1 + \ldots + X_n \sim \text{Gamma}(k = n, \lambda)$$



Зураг: $\text{Exp}(\lambda)$ тархалттай, хамааралгүй хувьсагчдын нийлбэрийн тархалт

Процесс ба санамсаргүй түүвэр

$$X_1, \ldots, X_n$$

гэсэн нэг ижил тархалттай, хамааралгүй санамсаргүй хувьсагчид нь Бернуллийн процесс, Пуассоны процесс зэрэг ямар нэг процесс болон санамсаргүй түүврийг төлөөлж чадна.

Ингээд $X_1, \ldots, X_n \sim IID(\mu, \sigma^2)$ байх үед

$$\bar{S}_n = \frac{S_n}{n} = \frac{X_1 + \ldots + X_n}{n}$$

дундаж авч үзье. $X = (X_1, \ldots, X_n)$ түүврийн хувьд $\bar{S}_n$ нь түүврийн дундаж гэх $\bar{X}$ статистиктай адил юм.

Ийнхүү $n \to \infty$ үед статистик түүврийн дундаж бас S_n хувьсагчийн тархалтыг олох буюу X_i хувьсагчдын (нэг ижил) тархалтын n давхар хуниас олох явдал нь $\bar{S}_n$ хувьсагчийн асимптот буюу хязгаарын тархалтыг олох уруу шилжинэ.

$\bar{S}_n$ хувьсагчийн дундаж болон дундаж квадрат хазайлт

$\bar{S}_n$ хувьсагчийн математик дундаж

$$E(\bar{S}_n) = E\left[\frac{S_n}{n}\right] = \frac{E(S_n)}{n} = \frac{E(X_1 + \ldots + X_n)}{n} = \frac{E(X_1) + \ldots + E(X_n)}{n} = \frac{\mu + \ldots + \mu}{n} = \mu$$

$\bar{S}_n$ хувьсагчийн дундаж квадрат хазайлт

$$D(\bar{S}_n) = D\left[\frac{S_n}{n}\right] = \frac{D(S_n)}{n^2} = \frac{D(X_1 + \ldots + X_n)}{n^2} = \frac{D(X_1) + \ldots + D(X_n)}{n^2} = \frac{\sigma^2 + \ldots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Энд нийлбэрийн дундаж квадрат хазайлтыг задлахдаа хувьсагчид хамааралгүй болохыг ашиглав.

$\bar{S}_n$ хувьсагчийн асимптот тархалт

Одоо $\bar{S}_n$ хувьсагчийн асимптот тархалтыг олоход анхаарлаа хандуулъя.

Сэргээн санах нь (Момент үүсгэгч функц)

$$M_X(t) = E\left[e^{tX}\right], \quad t \in \mathbb{R}$$

Санамж

Санамсаргүй хувьсагчийн тархалтыг илэрхийлэх бас нэг хэлбэр бол момент үүсгэгч функц юм.

Момент үүсгэгч функцийн зарим чанар

1. X болон Y ижил тархалттай бол $M_X(t) = M_Y(t)$ байна.
2. X болон Y хамааралгүй бол $M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$ байна.
3. $Y = a + bX$ бол $M_Y(t) = e^{at} M_X(bt)$ байна.

$X_1, \ldots, X_n \sim IID(\mu, \sigma^2)$ тул 1 болон 2 дугаар чанар ёсоор

$$M_{X_1 + \ldots + X_n}(t) = M_{X_1}(t) \cdot \ldots \cdot M_{X_n}(t) = [M_{X_1}(t)]^n$$

болно. Үлдэх 3 дугаар чанарыг ашиглавал $S_n^* = \dfrac{\bar{S}_n - \mu}{\sigma/\sqrt{n}} = \dfrac{X_1 + \ldots + X_n - n\mu}{\sqrt{n}\sigma}$ хувьсагчийн момент үүсгэгч функц

$$M_{S_n^*}(t) = e^{-\sqrt{n}\mu t/\sigma}\left[M_{X_1}\left(\frac{t}{\sqrt{n}\sigma}\right)\right]^n$$

гэж олдоно. Нөгөө талаас $X_i = \dfrac{X_i - \mu}{\sigma}$ стандарт хувиргалт хийвэл $X_1, \ldots, X_n \sim IID(0, 1)$ улмаар $M_{S_n^*}(t) = [M_{X_1}(t/\sqrt{n})]^n$ болно.

Стандарт хувиргалт хийсэн тул $E(X_1) = 0$ ба $E(X_1^2) = D(X_1) - [E(X_1)]^2 = 1$ болохыг анхаараад Тейлорын томьёо ашиглавал

$$M_{X_1}(t) = E[e^{tX_1}] = 1 + tE(X_1) + \frac{t^2}{2}E(X_1^2) + t^2 h(t) = 1 + \frac{t^2}{2} + t^2 h(t), \qquad \text{энд } t \to 0 \text{ үед } h(t) \to 0$$

болно. Улмаар $\lim\limits_{n\to\infty}\left(1 + \dfrac{a}{n}\right)^n = e^a$ хязгаар ашиглавал

$$M_{S_n^*}(t) = \left[1 + \frac{t^2/2}{n} + \frac{t^2}{n}h(t/\sqrt{n})\right]^n \to e^{t^2/2}$$

үр дүнд хүрнэ.

Сэргээн санах нь

Стандарт хэвийн тархалттай X санамсаргүй хувьсагчийн момент үүсгэгч функц $M_X(t) = e^{t^2/2}$ байдаг.

Иймд S_n^* хувьсагч $n \to \infty$ үед стандарт хэвийн тархалттай байна.

Хязгаарын гол теорем

Эцэст нь гарсан үр дүнг стандарт хувиргалтаас өмнөх хувьсагчдын хувьд томьёолъё.

Теорем (Хязгаарын гол теорем)

$X_1, \ldots, X_n \sim IID(\mu, \sigma^2)$ байг. Тэгвэл n хангалттай их үед

$$\bar{S}_n = \frac{X_1 + \ldots + X_n}{n} \sim N(\mu, \sigma^2/n)$$

байна.

Лекц XI

Хамааралтай хувьсагчдын дараалал, Марковын хэлхээ

Хамааралтай хувьсагчдын дараалал, Марковын хэлхээ сэдвийн агуулга

Хэсэг 1

Хамааралтай хувьсагчдын дараалал

Хамааралтай хувьсагчдын дараалал

Бернуллийн болон Пуассоны процесс "ой санамжгүй" өөрөөр хэлбэл өнгөрсөн үеийн мэдээллээс хэтийн ирээдүйг урьдчилан прогнозлох боломжгүй юм. Үүнийг санамсаргүй хувьсагчдын дараалал дахь дурын X_i болон X_j хоёр хувьсагч хамааралгүй гэж томьёолж байсан.

Гэвч практикт жишээлбэл автомат удирдлага, харилцаа холбоо, дохио боловсруулалт, аж үйлдвэр, эдийн засаг дахь стохастик динамик системүүдийг илэрхийлэх X_{n+1} хувьсагч өмнөх $X_n, X_{n-1}, \dots, X_0$ хувьсагчдаас эсвэл эдгээрийн заримаас хамааралтай байх явдал өргөн тохиолддог. Иймд хамааралтай санамсаргүй хувьсагчдын дараалал авч үзэх шаардлагатай.

Энэ сэдэвт хамааралтай санамсаргүй хувьсагчдын дарааллын төлөөлөл болгон Марковын хэлхээг авч үзнэ. Магадлалын онолд санамсаргүй процессыг цааш өргөтгөн Марковын процесс, Винерийн процесс гэх мэтчилэн олон талаас нь судалдаг.

Хэсэг 2

Марковын хэлхээ

Марковын хэлхээ

Марковын хэлхээг дискрет хугацаатай, процесс дахь санамсаргүй хувьсагчид дискрет байх үед авч үзнэ. Хугацааг дискрет гэх тул туршилт явуулах үеийн хугацааны эгшнүүдийг $1, \dots, n, \dots$ гэж дугаарлая. Улмаар хугацааны n эгшин дэх системийн төлвийг илэрхийлэх санамсаргүй хувьсагчийг X_n гэе. Тус хувьсагчийн авах утга буюу системийн боломжит төлвүүдээс тогтох төгсгөлөг олонлогийг *төлвийн олонлог* гээд S гэж тэмдэглэе. Төлвүүдийг өөр хооронд нь ялгахын тулд дугаарласан гэвэл $S = \{1, \dots, m\}$ болно. Иймд $X_n \in S$ байна.

Жишээ (Бямбажав Д., Магадлалын онол, математик статистик, 1999)

Жил бүрийн хур тунадасны хэмжээ харилцан адилгүй байдаг. Хур тунадасны хамгийн бага түвшинг 1 дүгээр төлөв, удаахыг 2 дугаар төлөв гэх мэтчилэн тэмдэглэе. Ингэвэл 1 дүгээр төлөв нь гантай жилийг харин 4 дүгээр төлөв нь усархаг жилийг заана.

Шилжилтийн магадлал ба Марковын нөхцөл

Марковын хэлхээний хувьд системийн төлвийг зөвхөн өмнөх төлвөөс хамаарна гэж үздэг бөгөөд үүнийг *Марковын нөхцөл* гэдэг. Иймд системийн шинж чанарыг

$$p_{ij} = P(X_{n+1} = j | X_n = i), \quad i, j \in S$$

хэлбэртэй *шилжилтийн магадлал* болон системийн анхны төлөв байдлын тусламжтай тодорхойлж болно. Марковын нөхцөлийг дараах байдлаар томьёолж болно.

$$\begin{aligned} P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\ = P(X_{n+1} = j | X_n = i) = p_{ij} \end{aligned}$$

Мөн $p_{ij} = P(X_{n+1} = j | X_n = i)$ шилжилтийн магадлал нь хугацаанаас хамаарахгүй байвал тус Марковын хэлхээг *нэгэн төрлийн* гэдэг.

Шилжилтийн магадлалын матриц

$$p_{ij} = P(X_{n+1} = j | X_n = i), \quad i, j \in S$$

шилжилтийн магадлалуудаас тогтох дараах матрицыг *шилжилтийн магадлалын матриц* гэнэ.

$$P = (p_{ij})_{i,j=1,\dots,m} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix}$$

Чанар

Дурын i бүрийн хувьд $\sum_{j=1}^{m} p_{ij} = 1$ буюу тусдаа тархалт байна.

Жишээ (Бямбажав Д., Магадлалын онол, математик статистик, 1999)

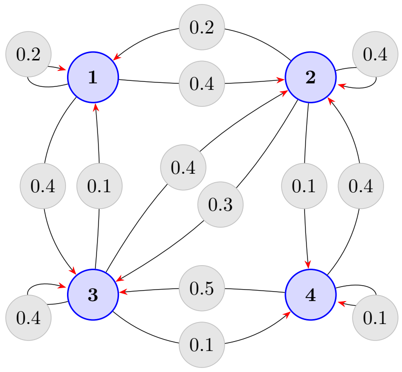
Практикаас харахад гантай жилээс усархаг жилд, усархаг жилээс гантай жилд шууд шилждэггүй байна. Тийнхүү дараах шилжилтийн магадлалын матриц олджээ.

$$\begin{pmatrix} 0.2 & 0.4 & 0.4 & 0 \\ 0.2 & 0.4 & 0.3 & 0.1 \\ 0.1 & 0.4 & 0.4 & 0.1 \\ 0 & 0.4 & 0.5 & 0.1 \end{pmatrix}$$

Жишээний хувьд тухайлбал $i = 1$ дүгээр мөрийн хувьд

$$\sum_{j=1}^{4} p_{1j} = 0.2 + 0.4 + 0.4 + 0 = 1$$

байна.



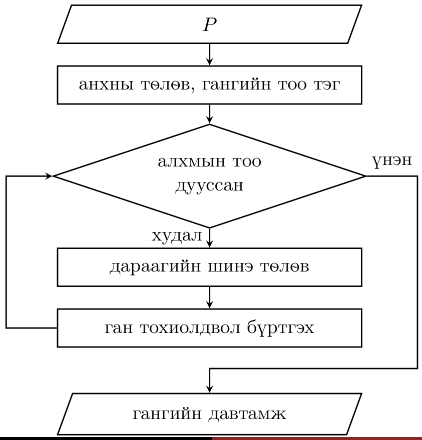
Зураг: Жишээгээр өгсөн шилжилтийн магадлалын матрицад харгалзах чиглэлт граф

Жишээ (Бямбажав Д., Магадлалын онол, математик статистик, 1999)

Судалгаанд хамрагдсан жилүүдэд тус бүс нутгийн уур амьсгал өөрчлөгдөөгүй бол дунджаар хэдэн жилд нэг удаа ган болохыг симуляцын аргаар олж тогтоо.

Үүний тулд бэхэлсэн анхны төлвөөс эхлүүлэн хангалттай олон жил буюу алхам бүхий хийсвэр туршилт явуулж улмаар гантай жилийн давтамж буюу хэдэн жилд нэг удаа ган болохыг тооцож гаргана. Уур амьсгал өөрчлөгдөөгүй гэдэг нь шилжилтийн магадлал өөрчлөгдөхгүй буюу хэлхээг нэгэн төрлийн гэж үзэх үндэс болно.

Програмын алгоритм болон эх кодыг дараагийн слайдаар харуулна.



(1) `INPUT P :=` (p_{ij})
(2) `current_state := 1; the_number_of_drought_years := 0`
(3) `FOR year FROM 1 TO the_number_of_years`
   (a) `current_state := new_state(P, current_state)`
   (b) `IF current_state == 1 THEN the_number_of_drought_years ++`
   `ENDFOR`
(4) `RETURN the_number_of_years / the_number_of_drought_years`


```r
P <- matrix(data = c(0.2, 0.4, 0.4, 0.0, 0.2, 0.4, 0.3, 0.1, 0.1, 0.4,
    0.4, 0.1, 0.0, 0.4, 0.5, 0.1), nrow = 4, byrow = TRUE)
new_state <- function (P, current_state) {
  new_state <- 0; u <- runif(n = 1); cum_prob <- 0
  while (cum_prob < u) {
    new_state <- new_state + 1
    cum_prob <- cum_prob + P[current_state, new_state]
  }
  return(new_state)
}
set.seed(0)
current_state <- 1; n_drought_years <- 0
for (year in 1:{n_years <- 1000}) {
  current_state <- new_state(P, current_state)
  if (current_state == 1) n_drought_years = n_drought_years + 1
}
cat("Frequency = ", n_years / n_drought_years, "\n")
cat("P(drought) = ", n_drought_years / n_years, "\n")
```

```
Frequency =  6.451613
P(drought) =  0.155
```


Заасан төлвүүдийг дайрах магадлал

$P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n)$ буюу систем i_0 төлвөөс эхэлж улмаар $i_1, \ldots, i_n$ төлвүүдийг дэс дараалан дайрах магадлалыг авч үзье.

Сэргээн санах нь (Үржүүлэх дүрэм)

$P(AB) = P(A|B)P(B)$

Үржүүлэх дүрэм болон Марковын нөхцөл ёсоор

$$\begin{aligned}
&P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) \\
&\quad = P(X_n = i_n | X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}) \\
&\qquad \cdot P(X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}) \\
&\quad = P(X_n = i_n | X_{n-1} = i_{n-1}) P(X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}) \\
&\quad = p_{i_{n-1} i_n} P(X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1})
\end{aligned}$$

болно.

Өмнөхтэй адилаар цааш үргэлжлүүлбэл

$$\begin{aligned}
&P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) \\
&\quad = p_{i_{n-1} i_n} \cdot p_{i_{n-2} i_{n-1}} \cdot \ldots \cdot p_{i_0 i_1} P(X_0 = i_0) \\
&\quad = P(X_0 = i_0) \cdot p_{i_0 i_1} \cdot \ldots \cdot p_{i_{n-2} i_{n-1}} \cdot p_{i_{n-1} i_n}
\end{aligned}$$

үр дүнд хүрнэ. Ийнхүү заасан төлвүүдийг дайрах магадлалыг анхны төлвийн магадлал болон шилжилтийн магадлалуудын үржвэрээр илэрхийллээ.

Тодорхой тооны шилжилтээр заасан төлөвт очих боломж

Систем хугацааны эхэнд i төлөвт байснаа хугацааны n алхмын дараа j төлөвт шилжих магадлалыг сонирхоё. Үүнийг

$$p_{ij}(n) = P(X_n = j | X_0 = i)$$

гэж томёолж болно. Энэ тохиолдолд мэдээж $p_{ij}(1) = p_{ij}$ гэж тооцно. $p_{ij}(n)$ магадлалыг олоход бүтэн магадлалын томьёо чухал үүрэгтэй.

Сэргээн санах нь (Бүтэн магадлалын томьёо)

$B_1, \ldots, B_k$ харилцан нийцгүй, $B_1 + \ldots + B_k = \Omega$, $P(B_i) > 0$ бол

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

байна.

Бүтэн магадлалын томьёо, нөхцөлт магадлалын томьёо болон Марковын нөхцөл ёсоор

$$\begin{aligned}
p_{ij}(n) = P(X_n = j|X_0 = i) &= \frac{P(X_n = j, X_0 = i)}{P(X_0 = i)} \\
&= \sum_{k=1}^{m} \frac{P(X_n = j|X_{n-1} = k, X_0 = i) \cdot P(X_{n-1} = k, X_0 = i)}{P(X_0 = i)} \\
&= \sum_{k=1}^{m} P(X_{n-1} = k|X_0 = i) \cdot P(X_n = j|X_{n-1} = k, X_0 = i) \\
&= \sum_{k=1}^{m} P(X_{n-1} = k|X_0 = i) \cdot P(X_n = j|X_{n-1} = k) \\
&= \sum_{k=1}^{m} p_{ik}(n-1) \cdot p_{kj}
\end{aligned}$$

гэсэн рекуррент томьёо гарна. Үүнийг *Колмогоров-Чепмений тэгшитгэл* гэдэг.

Жишээ

Энэ жил усархаг бол хоёр жилийн дараа ган тохиох магадлалыг ол.

Энэ тохиолдолд өмнөх томьёо

$$p_{41}(2) = \sum_{k=1}^{4} p_{4k}(1) \cdot p_{k1}$$

хэлбэртэй болно. Шилжилтийн магадлалуудыг орлуулж бодвол

$$p_{41}(2) = 0 \cdot 0.2 + 0.4 \cdot 0.2 + 0.5 \cdot 0.1 + 0.1 \cdot 0 = 0.13$$

үр дүн гарна.

$P(n) = (p_{ij}(n))_{i,j=1,\dots,m}$ матриц ашиглавал $p_{ij}(n)$ магадлалуудыг дурын i, j хос бүрийн хувьд нийтэд нь илэрхийлж чадах

$$P(n) = P(n-1) \cdot P(1)$$

матрицан тэгшитгэл гарна. Энд $p_{ij}(1) = p_{ij}$ болохыг анхаарвал $P(1) = P$ болно. Энд P бол шилжилтийн магадлалын матриц юм. Иймхүү

$$P(n) = [P(1)]^n = P^n$$

томьёо гарна. Үүнийг $p_{ij}(n)$ магадлал олох болон түүний асимптот шинж чанарыг судлахад ашиглана.

Жишээний хувьд

$$P(2) = P^2 = \begin{pmatrix} 0.16 & 0.4 & 0.36 & 0.08 \\ 0.15 & 0.4 & 0.37 & 0.08 \\ 0.14 & 0.4 & 0.37 & 0.09 \\ 0.13 & 0.4 & 0.37 & 0.10 \end{pmatrix}$$

байна. Эндээс 4 дүгээр мөр, 1 дүгээр баганын 0.13 гэсэн магадлал өмнө олсон $p_{41}(2) = 0.13$ магадлалтай адил байгааг харна уу.

Хэсэг 3

Марковын хэлхээний төлвийн стационар тархалт

Төлвийн стационар болон асимптот стационар тархалт

$j = 1, \dots, m$ төлөв бүрийн хувьд $P(X_n = j) = \pi_j$, өөрөөр хэлбэл систем j төлөвт байх магадлал n буюу хугацаанаас хамаараагүй, эсвэл $\lim_{n\to\infty} P(X_n = j) = \pi_j$ буюу хугацаа өнгөрөх тусам систем j төлөвт байх магадлал тогтворжих явдал заримдаа тохиолддог.

Марковын хэлхээний ийм $(\pi_1, \dots, \pi_m)$ тархалтуудыг харгалзан *стационар тархалт* болон *асимптот стационар тархалт* гэнэ.

Төлвийн стационар тархалт олох

Төлвийн стационар магадлал нь хугацаанаас хамаарахгүй буюу явцын дунд үл хөдлөх магадлал юм. Харин хугацааны тодорхой эгшинд харгалзах төлвийн магадлалыг Колмогоров-Чепмений тэгшитгэл гэгдэх рекуррент томьёогоор бодож олдог гэж үзсэн. Тэгвэл төлвийн стационар магадлалууд оршин байвал тэр нь үл хөдлөх буюу хугацаанаас хамаарахгүй тул тус рекуррент томьёо дахь хугацаанаас хамаарсан үл мэдэгдэгчдийн оронд бичигдэнэ. Учир нь өмнө байсан төлвийн магадлал хугацааны нэг алхмын дараа ч өөрчлөгдөлгүй хэвээрээ үлдэх ёстой юм. Иймхүү төлвийн стационар магадлалуудыг олох дараах систем тэгшитгэл бичиж болно.

$$\begin{cases} \sum_{k=1}^{m} \pi_k \cdot p_{kj} = \pi_j, \quad j = 1, \dots, m & \text{үл хөдлөх буюу стационар} \\ \sum_{k=1}^{m} \pi_k = 1 & \text{тархалт} \end{cases}$$

Жишээний хувьд стационар тархалтыг нь олохын тулд

$$\begin{cases} 0.2\pi_1 + 0.2\pi_2 + 0.1\pi_3 + 0\pi_4 = \pi_1 \\ 0.4\pi_1 + 0.4\pi_2 + 0.4\pi_3 + 0.4\pi_4 = \pi_2 \\ 0.4\pi_1 + 0.3\pi_2 + 0.4\pi_3 + 0.5\pi_4 = \pi_3 \\ 0\pi_1 + 0.1\pi_2 + 0.1\pi_3 + 0.1\pi_4 = \pi_4 \\ \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1 \end{cases}$$

тэгшитгэл бичнэ. Үүнийг бодвол ойролцоогоор

$$\pi = (0.146, 0.400, 0.368, 0.085)$$

шийд олдоно. Тархалтын нөхцөл буюу магадлалуудын нийлбэр нэгтэй тэнцүү бас стационарын $P^T\pi = \pi$ нөхцөл хангах тул энэ нь тус Марковын хэлхээний стационар тархалт мөн. Энд P^T нь хөрвөсөн матриц юм.

Төлвийн асимптот стационар тархалт олох

Хэрэв нэгэн төрлийн Марковын хэлхээний хувьд хугацааны ямар нэг $n > 0$ алхамд харгалзах P^n матрицын бүх элемент эерэг байвал дурын i бүрийн хувьд

$$\lim_{n\to\infty} p_{ij}(n) = \pi_j$$

өөрөөр хэлбэл анхны төлвөөс үл хамаарсан асимптот стационар тархалт олдоно.

Сэргээн санах нь

$P(n) = P^n$ томьёог $p_{ij}(n)$ магадлал олох болон түүний асимптот шинж чанарыг судлахад ашиглана.

Тэгэхээр дээрх нөхцөл биелэх үед

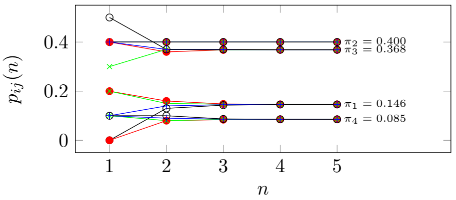
$$P(n) = P^n$$

томьёо ашиглаж асимптот стационар тархалт олох боломжтой.

Жишээний хувьд P^2 матрицын бүх элемент эерэг байсан тул

$$\lim_{n\to\infty} P^n = \begin{pmatrix} 0.146 & 0.400 & 0.368 & 0.085 \\ 0.146 & 0.400 & 0.368 & 0.085 \\ 0.146 & 0.400 & 0.368 & 0.085 \\ 0.146 & 0.400 & 0.368 & 0.085 \end{pmatrix}$$

буюу $\pi = (0.146, 0.400, 0.368, 0.085)$ асимптот стационар тархалт олдоно. Энэ нь өмнө олсон стационар тархалттай давхцаж байна.



Зураг: Марковын хэлхээний асимптот стационар тархалтын нийлэлт

Хэсэг 4

Марковын хэлхээний төлвийн ангилал

Төлвийн ангилал

Тодорхой алхмын дараа i төлвөөс j төлөвт шилжих боломжтой өөрөөр хэлбэл $p_{ij}(n) > 0$ байх n олддог бол i төлвөөс j төлөв мөрдөнө гээд $i \to j$ гэж тэмдэглэдэг. Харин i болон j төлвүүд бие биеэсээ мөрдөх бол тэдгээрийг *харилцан мөрдөх төлвүүд* гээд $i \leftrightarrow j$ гэж тэмдэглэнэ. Өөр хоорондоо харилцан мөрдөх төлвүүдийг хамтад нь *үл задрах анги* гэдэг. Марковын хэлхээг үл задрах ангиудад хувааж болох бөгөөд хэрэв хэлхээ ганц үл задрах ангиас тогтож байвал түүнийг *үл задрах хэлхээ* гэнэ.

Систем i төлөвт эхний удаа буцаж шилжих хугацааг T_i гэе. Тэгвэл

$$P(T_i < \infty | X_0 = i) = 1$$

буюу i төлөвт төгсгөлөг хугацааны дараа баталгаатай эргэн ирдэг бол тус төлвийг *рекуррент* харин эсрэг тохиолдолд *транзиент* гэнэ.

Систем i төлөвт эхний удаа эргэж ирэх дундаж хугацаа буюу

$$E(T_i) = \sum_{n=1}^{\infty} n \cdot P(T_i = n | X_0 = i)$$

нь төгсгөлөг байх албагүй юм.

Систем i төлөвт шилжих нийт тоог V гэе. Тэгвэл тус санамсаргүй хувьсагчийн тархалт ямар байх нь i төлөв рекуррент ба транзиент төлвүүдийн аль нь байхаас шалтгаална.

1. Хэрэв i төлөв рекуррент бол
$$P(V = \infty | X_0 = i) = 1$$
2. Хэрэв i төлөв транзиент бол
$$(V | X_0 = i) \sim \mathrm{Geom}(1 - P(X_n = i | X_0 = i))$$

Мөн i төлөв рекуррент байх зайлшгүй бөгөөд хүрэлцээтэй нөхцөл нь

$$\sum_{n=1}^{\infty} p_{ii}(n) = \infty$$

байх явдал юм.

Хэрэв систем i төлвөөс өөр төлөвт шилжих боломжгүй бол тус төлвийг *шингээгч* гэнэ. Систем шингээгч төлөвт шилжих магадлалыг дараах байдлаар олно.

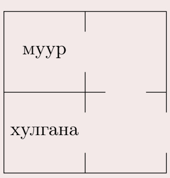
1. Систем шингээгч төлвүүдийн аль нэгд байгаа бол тус төлөвт шингэх магадлал нэгтэй тэнцүү харин бусад төлөвт шингэх магадлал тэгтэй тэнцүү байна.
2. Транзиент i төлвөөс шингээгч төлөвт шилжих магадлал a_i нь
$$a_i = \sum_{j=1}^{m} p_{ij} \cdot a_j, \quad i = 1, \ldots, m$$
систем тэгшитгэлээр нэг утгатай тодорхойлогдоно.

Эцэст нь систем шингээгч төлөвт шилжих дундаж хугацаанд авч үзье. Төлөв бүрт харгалзах тус дундаж хугацааг $\mu_1, \ldots, \mu_m$ гэвэл эдгээр нь дараах тэгшитгэлүүдээр нэг утгатай тодорхойлогдоно.

$$\begin{cases} \mu_i = 0, & i \text{ төлөв шингээгч} \\ \mu_i = 1 + \sum_{j=1}^{m} p_{ij}\mu_j & i \text{ төлөв транзиент} \end{cases}$$

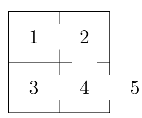
Жишээ (Муур ба хулгана)

Муур хулгана хоёр дөрвөн өрөөтэй байшинд зураг дээр үзүүлсэн байдлаар байрлаж байв.



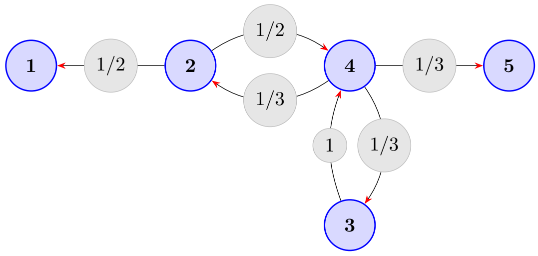
Хулгана хаалгануудын аль нэгийг тэнцүү магадлалтай сонгоно. Муур өрөөнөөсөө гарахгүй. Хэрэв хулгана муууртай өрөөнд орвол муур түүнийг барьж иднэ. Хулгана байшингаас гарч чадвал эргэж орохгүй.

"Муур ба хулгана" жишээ дэх Марковын хэлхээний төлвүүд болон тэдгээрт харгалзах шилжилтийн магадлалын матрицыг дараах байдлаар бичиж болно.



$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Мөн хулгана анх 3 дугаар төлөвт байх тул анхны төлвийн тархалт $(0, 0, 1, 0, 0)$ байна.



Зураг: "Муур ба хулгана" жишээний шилжилтийн магадлалын матрицад харгалзах чиглэлт граф

Шингээгч төлвүүд нь 1 болон 5 дугаар төлөв бөгөөд

$$P(1 \text{ төлөвт шингэх}) = 1/3, \qquad P(5 \text{ төлөвт шингэх}) = 2/3$$

байна.

Зогсолтын момент

Систем i төлөвт анх удаа шилжих хугацааны эгшин буюу

$$\tau_i = \inf\{n \geq 0 : X_n = i\}$$

санамсаргүй хувьсагчийг *зогсолтын момент* гэдэг.

"Муур ба хулгана" жишээний хувьд $E(\tau_1) \approx 5$ ба $E(\tau_5) \approx 4$ байна.

Лекц XII

Тархалтын параметрийн статистик үнэлэлт

Тархалтын параметрийн статистик үнэлэлт сэдвийн агуулга

Хэсэг 1

Тархалтын загвар тавих тухай

Тархалтын загвар тавих тухай

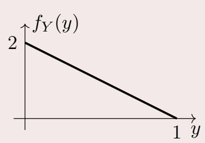
Тархалт нь үл мэдэгдэх X санамсаргүй хувьсагчийн эх олонлогоос авсан $X_1, \ldots, X_n$ түүврээр гаргаж авсан өгөгдөлд үндэслэн уг санамсаргүй хувьсагчийн эх олонлогийн тархалтын тухай таамаглал хэрхэн дэвшүүлэхийг авч үзье.

Ийнхүү санамсаргүй хувьсагчийн тархалт болон тархалтынх нь шинж чанарыг өгөгдөлд тулгуурлан олж тогтоох нь магадлалаас статистик уруу шилжиж буй явдал юм.

Жишээ буюу бодлого

Сэргээн санах нь

Богд Жавзандамба хутагтад сүсэгтэн олноос өргөдөг өргөл барьцын хэмжээ $k = 3$ хэлбэрийн параметр болон $1/\lambda = 1$ масштабын параметр бүхий гамма тархалттай байв. Харин Данигай сойвон өргөл барьцын Y хувийг Богдын санд бүртгээд бусдыг нь хувьдаа



Зураг: Y санамсаргүй хувьсагчийн тархалт

завшдаг бол санд орох өргөл барьцын хэмжээний тархалтыг ол.

Симуляцын аргаар гарган авсан өгөгдөл

Санд орох өргөл барьцын хэмжээг Z гэе. Тэгвэл $Z = X \cdot Y$ байна.

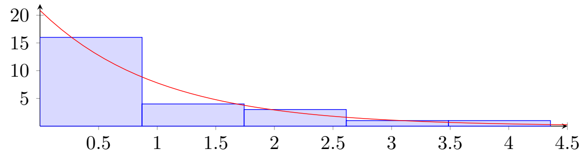
```
set.seed(0)
X <- rgamma(n = 25, shape = 3, rate = 1)

rtriangle <- function(n, a = 0, c = 0, b = 1) {
  U <- runif(n = n)
  ifelse(test = U < (c - a) / (b - a), a + sqrt(U * (b -
    a) * (c - a)), b - sqrt((1 - U) * (b - a) * (b - c)))
}

Y <- rtriangle(n = length(X))
Z <- X * Y
print(round(x = Z, digits = 2))
```

```
0.68 0.56 0.70 0.14 4.36 0.71 2.09 0.39 0.26 0.45 1.38 1.53 0.28
2.10 2.83 1.03 1.80 0.59 0.06 1.70 0.48 0.71 0.11 0.28 0.17
```

Гистограмм ба тархалтын тухай таамаглал

Тасралтгүй санамсаргүй хувьсагчийн тархалтын нягтын хэлбэрийг өгөгдөлд тулгуурлан харахад гистограмм ашигладаг. Иймд гистограммд үндэслэн тархалтын тухай таамаглал дэвшүүлнэ.

```
hist(Z)
```



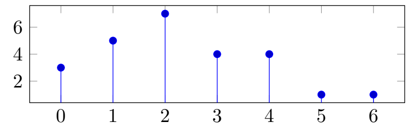
Зураг: Симуляцын аргаар гарган авсан өгөгдлийн гистограмм

Зураг дээрх гистограммаас илтгэгч тархалтын хэлбэр ажиглагдаж байна. Иймд Z хувьсагчийг илтгэгч тархалттай гэж таамаглая.

Дискрет санамсаргүй хувьсагчийн хувьд тархалтын тухай таамаглал дэвшүүлэх

Өгөгдөл дэх утгуудын давтамжаар байгуулах диаграмм ашиглана.

```
X <- c(1, 5, 6, 2, 1, 0, 1, 3, 0, 0, 2, 4, 2, 2, 2, 3, 3,
   3, 4, 1, 2, 4, 1, 4, 2)
plot(x = table(X), ylab = "Frequency")
```



Зураг: Дискрет хувьсагчийн эх олонлогоос авсан өгөгдлийн давтамж

Диаграммыг харвал хувьсагчийг Пуассоны эсвэл бином тархалттай гэсэн таамаглал дэвшүүлэх боломжтой.

Хэсэг 2

Тархалтын параметрийн үнэлэлт

Тархалтын параметрийн үнэлэлт

Өмнөх хэсгийн төгсгөлд Z хувьсагчийг илтгэгч тархалттай гэж таамагласан. Гэтэл илтгэгч тархалт λ гэсэн параметртэй бөгөөд түүний утга мэдэгдэхгүй байна. Иймд таамагласан тархалтаа тодорхой болгохын тулд тус үл мэдэгдэх параметрийн утгыг олох буюу үнэлэх шаардлагатай.

Тархалтын үл мэдэгдэх параметр үнэлэх сонгодог аргуудын төлөөлөл болгон дараах хоёр аргыг авч үзнэ. Үүнд:

1. Моментын арга
2. Хамгийн их үнэний хувь бүхий арга

Моментын арга

Илтгэгч тархалттай X санамсаргүй хувьсагчийн математик дундаж буюу нэг дүгээр эрэмбийн анхны момент[1] нь (эрчмийн) параметрийнхээ урвуутай тэнцүү өөрөөр хэлбэл

$$E(X) = \frac{1}{\lambda}$$

болохыг бид мэднэ. Үүнд харгалзах түүврийн нэг дүгээр эрэмбийн анхны момент бол түүврийн дундаж гэх дараах статистик юм.

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

Иймд $E(X) = \overline{X}$ тэгшитгэлээс дараах үнэлэлт олдоно.

$$\hat{\lambda} = \frac{1}{\overline{X}}$$

[1] $\alpha_k = E[X^k]$ энд $k = 1$

Жишээ бодлогын хувьд симуляцын аргаар гарган авсан өгөгдлийн дундаж нь $\overline{Z} \approx 1.016$ тул Z санамсаргүй санамсаргүй хувьсагчийн (илтгэгч) тархалтын параметрийн утга

$$\hat{\lambda} = \frac{1}{\overline{Z}} \approx \frac{1}{1.016} \approx 0.984$$

гэж олдоно.

Хэрэв тархалт олон параметртэй бол өөр бусад моментуудыг нэмж ашиглана.

R програм дээр гистограмм байгуулах улмаар (илтгэгч) тархалтын нягтын муруй нэмж зурах тушаал дараа байдалтай байна.

```
hist(x = Z, freq = FALSE)
curve(expr = dexp(x, rate = 0.984), add = TRUE)
```


Ийнхүү тархалтын хуулийн тухай анхны таамаглал эцэстээ

$$H_0 : Z \sim \mathrm{Exp}(\lambda = 0.984)$$

буюу Богдын санд орох өргөл барьцын хэмжээ $\lambda = 0.984$ эрчмийн параметр бүхий илтгэгч тархалттай гэсэн өгүүлбэр боллоо.

Бодлогын аналитик шийд $Z \sim \mathrm{Exp}(\lambda = 1)$ гэж олдож байсан. Өөрөөр хэлбэл өгөгдөлд тулгуурлан олсон (үнэлсэн) параметрийн утга нь жинхэнэ утгаас "хазайсан" байна. Энэхүү хазайлтын талаар үргэлжлүүлэн үзэх болно.

Санамж

Заримдаа параметрийг түүнээс хамаарсан функцээр дамжуулан үнэлдэг.

Жишээ болгон авч үзэж буй илтгэгч тархалтын параметрийн үнэлэлтийг

$$\widehat{\varphi(\lambda)} = \frac{1}{\hat{\lambda}} = \overline{Z}$$

байдлаар авъя. Энэ тохиолдолд параметрийн жинхэнэ утгыг ч бас $\frac{1}{\lambda}$ гэж авна.

Үнэлэлтийн хазайлт ба дундаж квадрат алдаа

Жишээний хувьд тархалтын масштабын параметрийн жинхэнэ утга $1/\lambda = 1$ байсан бол статистик үнэлэлтээр $1/\hat\lambda = 1.016$ гэсэн "хазайлттай" утга олдсон. Гэвч онолын хувьд уг үнэлэлтийн хазайлт

$$
\begin{aligned}
b\left(\frac{1}{\hat\lambda}\right) &= E\left(\frac{1}{\hat\lambda} - \frac{1}{\lambda}\right)\\
&= E(\overline{Z}) - \frac{1}{\lambda} = E\left(\frac{Z_1 + \ldots + Z_n}{n}\right) - \frac{1}{\lambda}\\
&= \frac{E(Z_1) + \ldots + E(Z_n)}{n} - 1/\lambda = \frac{n \cdot \frac{1}{\lambda}}{n} - \frac{1}{\lambda} = 0
\end{aligned}
$$

буюу $1/\hat\lambda = \overline{Z}$ нь *хазайлтгүй үнэлэлт* юм. Ийнхүү бид онолын хувьд хазайлтгүй өөрөөр хэлбэл дундаж алдаа нь тэгтэй тэнцүү байх үнэлэлт ашиглажээ. Гэвч практик алдаа ерөнхийдөө байсаар байх тул энэхүү алдааных нь "хэлбэлзлийг" хэмжих шаардлагатай. Тус алдааг дундаж квадрат алдаа гэсэн утгаар хэмждэг.

Үнэлэлтийн дундаж квадрат алдаа ба стандарт алдаа

Жишээний хувьд $b(1/\hat\lambda) = 0$ бас $Z_1, \ldots, Z_n$ нь энгийн санамсаргүй түүвэр тул эдгээр хувьсагчид хамааралгүй бөгөөд бүгд нэг ижил илтгэгч тархалттай. Иймд үнэлэлтийн дундаж квадрат алдаа

$$
\begin{aligned}
SE^2\left(\frac{1}{\hat\lambda}\right) &= E\left(\frac{1}{\hat\lambda} - \frac{1}{\lambda}\right)^2 = D\left(\frac{1}{\hat\lambda} - \frac{1}{\lambda}\right) + \left[E\left(\frac{1}{\hat\lambda} - \frac{1}{\lambda}\right)\right]^2\\
&= \underbrace{D\left(\frac{1}{\hat\lambda}\right)}_{\text{үнэлэлтийн дисперс}} + \underbrace{\left[b\left(\frac{1}{\hat\lambda}\right)\right]^2}_{\text{хазайлтын квадрат}}\\
&= D\left(\frac{Z_1 + \ldots + Z_n}{n}\right) = \frac{D(Z_1) + \ldots + D(Z_n)}{n^2} = \frac{1}{n\lambda^2}
\end{aligned}
$$

болж улмаар үнэлэлтийн стандарт алдаа $SE\left(\frac{1}{\hat\lambda}\right) = \frac{1}{\sqrt{n}\lambda} = \frac{1}{\sqrt{25} \cdot 0.984} \approx 0.203$ гэж олдоно.

Параметрийн завсран үнэлэлт буюу итгэх завсар

Моментын аргаар олсон үнэлэлт параметрийн зөвхөн нэг л утга заадаг. Иймд уг үнэлэлтийг *цэгэн үнэлэлт* гэдэг бөгөөд нөгөө талаас параметрийн *завсран үнэлэлт* буюу *итгэх завсар* авч үздэг.

$$P(T_1 < \lambda < T_2) \geq 1 - \alpha$$

чанартай (T_1, T_2) завсрыг λ параметрийн завсран үнэлэлт буюу $1 - \alpha$ итгэх магадлалтай итгэх завсар эсвэл $(1 - \alpha) \cdot 100\%$ хувийн итгэх завсар гэнэ.

Илтгэгч тархалтын эрчмийн параметрийн итгэх завсар

Итгэх завсрыг параметрийн цэгэн үнэлэлт, түүний тархалтыг ашиглаж хэрхэн олохыг авч үзье.

$\frac{1}{\hat\lambda} = \overline{Z}$ цэгэн үнэлэлтийн тархалтыг олоход $Z_1, \ldots, Z_n$ нь энгийн санамсаргүй түүвэр буюу $Z_1, \ldots, Z_n \sim \text{Exp}(\lambda)$ бөгөөд хамааралгүй хувьсагч гэдгийг ашиглана.

Сэргээн санах нь

$X_1, \ldots, X_k \sim \text{Exp}(\lambda)$ хамааралгүй бол

$$X_1 + \ldots + X_k \sim \text{Gamma}(\lambda, k)$$

Уг чанараар $\frac{n}{\hat\lambda} = n\overline{Z} = Z_1 + \ldots + Z_n \sim \text{Gamma}(\lambda, n)$ болно.

Сэргээн санах нь

1. $X \sim \text{Gamma}(\lambda, k)$ бол
$$Y = cX \sim \text{Gamma}(\lambda/c, k)$$
2. $\text{Gamma}(1/2, k/2) = \chi^2(k)$

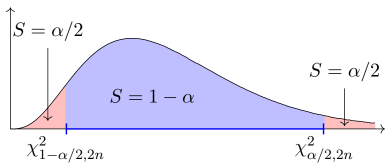
Дээрх чанараар $2\lambda n\overline{Z} = 2\lambda \sum_{i=1}^{n} Z_i \sim \text{Gamma}(\lambda/(2\lambda), 2n/2) = \chi^2(2n)$ болох тул

$$P\left(\chi^2_{1-\alpha/2,2n} < 2\lambda n\overline{Z} < \chi^2_{\alpha/2,2n}\right) = 1 - \alpha$$

буюу

$$\frac{\chi^2_{1-\alpha/2,2n}}{2n\overline{Z}} < \lambda < \frac{\chi^2_{\alpha/2,2n}}{2n\overline{Z}}$$

итгэх завсар олдоно. Энд $\chi^2_{\alpha,k}$ нь k чөлөөний зэрэгтэй хи-квадрат тархалтын $1 - \alpha$ эрэмбийн квантилын утга буюу α хэмжээтэй талбай бүхий тархалтын баруун сүүлний утга юм.



$$\frac{\chi^2_{1-\alpha/2,2n}}{2n\overline{Z}} < \lambda < \frac{\chi^2_{\alpha/2,2n}}{2n\overline{Z}}$$

Зураг: Илтгэгч тархалтын эрчмийн параметрийн $1 - \alpha$ итгэх магадлалтай итгэх завсар

```
alpha <- 0.1
df <- 2 * 25
qchisq(p = alpha/2, df = df, lower.tail = FALSE)
qchisq(p = 1 - alpha/2, df = df, lower.tail = FALSE)
```


Жишээний хувьд Богдын санд орох өргөл барьцын хэмжээг илэрхийлэх Z санамсаргүй хувьсагчийн (илтгэгч) тархалтын λ параметрийн $1 - \alpha = 1 - 0.1 = 0.9$ итгэх магадлалтай өөрөөр хэлбэл 90 хувийн итгэх завсар

$$\frac{\chi^2_{1-\alpha/2,2n}}{2n\overline{Z}} < \lambda < \frac{\chi^2_{\alpha/2,2n}}{2n\overline{Z}}$$

томъёогоор

$$\frac{34.764}{2 \cdot 25 \cdot 1.016} < \lambda < \frac{67.505}{2 \cdot 25 \cdot 1.016}$$

буюу

$$0.684 < \lambda < 1.329$$

гэж олдоно.

Лекц XIII

Хамгийн их үнэний хувь бүхий үнэлэлт, Байесын үнэлэлт

Хамгийн их үнэний хувь бүхий үнэлэлт, Байесын үнэлэлт сэдвийн агуулга

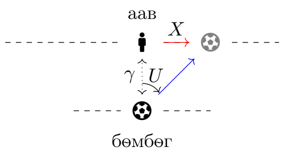
Хэсэг 1

Кошийн тархалт

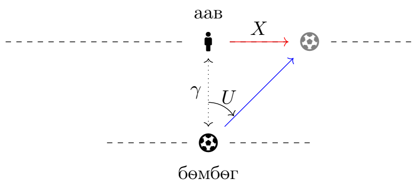
Кошийн тархалт

Жишээ

Балчир хүү аавтайгаа хамт хөл бөмбөг тоглож байв. Хүү ааваас нь γ зайд байх бөмбөгийг аав уруугаа өшиглөж дамжуулах ёстой ч хэт балчир бас туршлагагүй тул бөмбөгийг хаа хамаагүй буюу ижил боломжтойгоор аль ч чиглэлд өшиглөж байв. Харин аав нь хүүгийнхээ өшиглөсөн бөмбөгийг эгц хөндлөн гүйн барьж авч байв. Аавын гүйх зайн тархалтыг ол.



Зураг: Жишээ бодлогын зураглал



Зураг: Жишээ бодлогын зураглал

Бодлогын нөхцөл ёсоор $U \sim U(-\pi/2, \pi/2)$ болох бөгөөд X санамсаргүй хувьсагчийн утга U хувьсагчийн утгаас хамаарна. Өөрөөр хэлбэл $X = g(U)$ хувиргалт өгчээ.

$$X = \gamma\,\mathrm{tg}(U)$$

Энд $X < 0$ утга зураг дээрх чиглэлийн эсрэг чиглэлд гүйх зайг илэрхийлнэ.

Одоо X хувьсагчийн тархалтын функцийг олъё.

$$F_U(x) = \frac{x - \left\{a = -\frac{\pi}{2}\right\}}{\left\{b = \frac{\pi}{2}\right\} - \left\{a = -\frac{\pi}{2}\right\}} = \frac{x + \frac{\pi}{2}}{\pi} = \frac{1}{2} + \frac{1}{\pi}x \qquad -\frac{\pi}{2} \le x \le \frac{\pi}{2}$$

байх тул

$$F_X(x) = P(X < x) = P(\gamma\,\mathrm{tg}(U) < x) = P\left[U < \mathrm{arctg}\left(\frac{x}{\gamma}\right)\right]$$
$$= F_U\left(\mathrm{arctg}(x/\gamma)\right) = \frac{1}{2} + \frac{1}{\pi}\mathrm{arctg}(x/\gamma) \qquad x \in \mathbb{R}$$

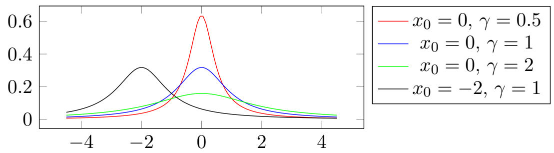
гэж олдоно. Уламжлал авбал дараах хэлбэртэй нягтын функц олдоно.

$$f_X(x) = F_X'(x) = \frac{1}{\pi}\frac{1/\gamma}{1 + (x/\gamma)^2} \qquad x \in \mathbb{R}$$

Кошийн тархалт, түүний параметрүүд

$$f_X(x) = \frac{1}{\pi}\frac{1/\gamma}{1 + \left(\frac{x - x_0}{\gamma}\right)^2} \qquad x \in \mathbb{R}$$

Энд $\gamma > 0$ масштабын параметр, x_0 байршлын параметр юм.



Зураг: Кошийн тархалтын нягтын муруй параметрийн янз бүрийн утгад

Кошийн тархалтын зарим чанар болон хэрэглээ

Чанар

1. $U \sim U(0, 1)$ бол $X = \mathrm{tg}\,(\pi(U - 1/2)) \sim \mathrm{Cauchy}(0, 1)$
2. $X \sim \mathrm{Cauchy}(x_0, \gamma)$ бол $aX + b \sim \mathrm{Cauchy}(ax_0 + b, |a|\gamma)$
3. $X \sim \mathrm{Cauchy}(0, \gamma)$ бол $\frac{1}{X} \sim \mathrm{Cauchy}\left(0, \frac{1}{\gamma}\right)$
4. $X, Y \sim N(0, 1)$ бөгөөд хамааралгүй бол $\frac{X}{Y} \sim \mathrm{Cauchy}(0, 1)$

Хэрэглээ

1. Цацрагийн тархалт
2. Цаг уурын онцгой үзэгдэл: аадар бороо, үер
3. Гэрэлт цамхагийн гэрэл гэх мэт эргэлдэж буй биетийн ажиглагдах байдал
4. …

Кошийн тархалтын зарим чанар

Жишээ

Өмнөх бодлогыг үргэлжлүүлэн авч үзье. Хүүгийн аав дунджаар хэр зайд гүйх вэ?

$$E(X) = \int_{-\infty}^{\infty} x\frac{1}{\pi}\frac{1/\gamma}{1 + (x/\gamma)^2}dx$$
$$= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x/\gamma}{1 + (x/\gamma)^2}dx = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{1}{1 + (x/\gamma)^2}d[1 + (x/\gamma)^2]$$
$$= \frac{1}{2\pi}\left[\ln(1 + (x/\gamma)^2)\right]_{-\infty}^{\infty} = \infty - \infty = \text{тодорхойгүй}$$

Кошийн тархалтын дундаж, дундаж квадрат хазайлт зэрэг моментууд тодорхойгүй, момент үүсгэгч функц нь оршин байдаггүй. Иймд Кошийн тархалтын параметрийг үнэлэхэд моментын арга хэрэглэх боломжгүй юм.

Хэсэг 2

Хамгийн их үнэний хувь бүхий үнэлэлт

Хамгийн их үнэний хувь бүхий үнэлэлт

θ параметр бүхий $f_X(x, \theta)$ нягттай санамсаргүй хувьсагчийн эх олонлогоос авсан $X_1, \dots, X_n$ энгийн санамсаргүй түүврээс хамаарсан

$$L(X, \theta) = f_X(X_1, \theta) \cdot \ldots \cdot f_X(X_n, \theta) = \prod_{i=1}^{n} f_X(X_i, \theta)$$

функцийг *үнэний хувь бүхий функц* гэнэ.

Үнэний хувь бүхий функцийг хамгийн их утгад хүргэх параметрийн утгыг *хамгийн их үнэний хувь бүхий үнэлэлт* гэнэ.

$$\hat{\theta} = \arg\max_{\theta} \; L(X, \theta)$$

Кошийн тархалтын масштабын параметрийн үнэлэлт

$$f_X(x) = \frac{1}{\pi} \frac{1/\gamma}{1 + (x/\gamma)^2} \quad x \in \mathbb{R}$$

Кошийн тархалт ба тус тархалттай эх олонлогоос авсан $X_1, \dots, X_n$ түүврийн хувьд үнэний хувь бүхий функц дараах хэлбэртэй байна.

$$L(X, \gamma) = \prod_{i=1}^{n} f_X(X_i, \gamma) = \prod_{i=1}^{n} \frac{1}{\pi} \frac{1/\gamma}{1 + (X_i/\gamma)^2}$$

Жишээ

$\gamma = 10$ үед дараах байдлаар гарган авсан өгөгдөл дээр тулгуурлан γ параметрийн утгыг буцаан үнэлье.

```
gamma <- 10
set.seed(0)
X <- rcauchy(n = 1000, location = 0, scale = gamma)
```


Санамж

$X_1, \dots, X_n$ тус бүр дээрх $f_X(x, \theta)$ нягтын функцийн утга эерэг бас $\ln(\cdot)$ функц монотон тул $\arg\max_{\theta} \; \ln L(X, \theta) = \arg\max_{\theta} \; L(X, \theta)$ байна.

$$\ln L(X, \gamma) = \sum_{i=1}^{n} \ln f_X(X_i, \gamma) = \sum_{i=1}^{n} \ln \left(\frac{1}{\pi} \frac{1/\gamma}{1 + (X_i/\gamma)^2} \right)$$
$$= -n \ln(\pi\gamma) - \sum_{i=1}^{n} \ln \left(1 + (X_i/\gamma)^2\right)$$

Одоо дээрх логарифм-үнэний хувь бүхий функцээс γ параметрээр уламжлал авч тэгтэй тэнцүүлэн экстремумын нөхцөл бичье.

$$\frac{d}{d\gamma} \ln L(X, \gamma) = -\frac{n}{\gamma} - \sum_{i=1}^{n} \frac{2(X_i/\gamma) X_i (-1/\gamma^2)}{1 + (X_i/\gamma)^2}$$
$$= -\frac{n}{\gamma} + \frac{2}{\gamma} \sum_{i=1}^{n} \frac{X_i^2}{\gamma^2 + X_i^2} = 0$$

Өмнөх нөхцөлийг дараах байдлаар бичиж болно.

$$-\frac{n}{2} + \sum_{i=1}^{n} \frac{X_i^2}{\gamma^2 + X_i^2} = 0$$

тэгшитгэлээс γ ил олдохгүй тул үүнийг тоон аргаар бодно. $\sum_{i=1}^{n} \frac{X_i^2}{\gamma^2 + X_i^2}$ функц γ хувьсагчийнхаа хувьд монотон тул дээрх тэгшитгэлийн шийд

$$\min_i |X_i| \le \gamma \le \max_i |X_i|$$

нөхцөл хангана.

Нэгэнт тоон арга хэрэглэх болсон тул $\ln L(X, \gamma)$ функцийг γ хувьсагчаар шууд максимумчилж бодъё. Үүний тулд R програм дээр дараах тушаал өгч болно.

```
optimize(
  f = function (x, X, n = length(X)) {
    - n * log(pi * x) - sum(log(1 + X ** 2 / x ** 2))
  },
  maximum = TRUE,
  lower = min(abs(X)), upper = max(abs(X)),
  X = X, n = length(X)
)
```

Энд өгөгдөл буюу санамсаргүй хувьсагчдын ажиглагдсан утгуудыг `X` гэсэн вектор байдлаар өгч байна. Ийнхүү дээрх тушаалыг ажиллуулахад

$$\hat{\gamma} \approx 10.086$$

буюу анх авсантай ойролцоо утга бүхий үнэлэлт олдлоо.

Хэсэг 3

Байесын үнэлэлт

Байесын зарчим

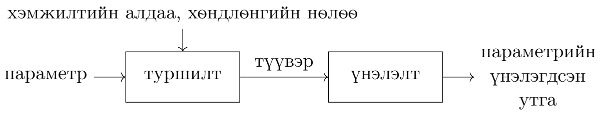
Сэргээн санах нь (Байесын томьёо)

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

$$P(\text{спам}|\text{үгс}) = \frac{P(\text{үгс}|\text{спам}) P(\text{спам})}{P(\text{үгс})}$$

Байесын үнэлэлт

Өмнө үзсэн сонгодог үнэлэлтүүдийн хувьд параметрийг тогтмол гэж тооцож байсан. Харин Байесын үнэлэлтийн хувьд параметрийг санамсаргүй хувьсагч байх тохиолдлыг авч үздэг.



Зураг: Параметрийн статистик үнэлэлт

Параметр санамсаргүй хувьсагч юм бол түүнийг судлахын тулд тархалтыг нь авч үзэх шаардлагатай. Нөгөө талаас түүнийг судлах ганц барьц нь түүвэр юм. Иймд Θ параметрийн тархалтыг $X = (X_1, \ldots, X_n)$ түүврээс хамааруулан $f_{\Theta|X}(\theta|x)$ байдлаар авч үздэг.

$$f_{\Theta|X}(\theta|x)$$

нөхцөлт тархалт дахь X түүвэр өөрөө Θ параметрээс хамаарах тул энэ чигээр нь судлахад хүнд юм. Иймд уг тархалтыг Байесын зарчимд тулгуурлан

$$\underbrace{f_{\Theta|X}(\theta|x)}_{\text{постериор тархалт}} = \frac{f_{X|\Theta}(x|\theta) \quad \overbrace{f_\Theta(\theta)}^{\text{приор тархалт}}}{f_X(x)}$$

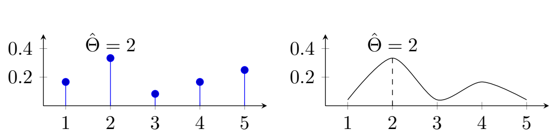
хэлбэрт шилжүүлж судалдаг. Энд $f_\Theta(\theta)$ тархалтыг мэдэгддэг гэж тооцно. Иймд $f_{\Theta|X}(\theta|x)$ постериор тархалтаас гарах үнэлэлт $f_\Theta(\theta)$ приор тархалтын "сонголтоос" хамаарна.

Энэ хичээлээр Байесын зарчимд тулгуурласан дараах үнэлэлтүүдийг авч үзнэ. Үүнд:

- Хамгийн их постериорын үнэлэлт
- Хамгийн бага дундаж квадрат алдаатай үнэлэлт

Хамгийн их постериорын үнэлэлт

$$\hat{\Theta} = \arg\max_{\theta} f_{\Theta|X}(\theta|x)$$



(a) Дискрет үед (b) Тасралтгүй үед

Зураг: Хамгийн их постериорын үнэлэлт

Хамгийн их постериорын үнэлэлтээр ерөнхийдөө параметрийн ганц утга олдоно.

Хамгийн бага дундаж квадрат алдаатай үнэлэлт

Сэргээн санах нь

$E(X_2|X_1 = x_1)$ нөхцөлт математик дундаж нь X_1 хувьсагч ашиглаж X_2 хувьсагчийг прогнозлох бүх $h(X_1)$ функц дундаас хамгийн бага дундаж квадрат алдаатай нь юм.

Иймд $\hat{\Theta} = E(\Theta|X = x)$ нь

$$E[(\hat{\Theta} - \Theta)^2|X = x] \le E[(h(x) - \Theta)^2|X = x]$$

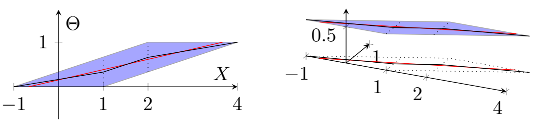
буюу хамгийн бага дундаж квадрат алдаатай үнэлэлт юм.

$\hat{\Theta} = E(\Theta|X = x)$ нь X түүврийн бэхлэгдсэн утга буюу x өгөгдлөөс хамаарсан функц байх тул өгөгдөл ямар байхаас шалтгаалж параметрийн үнэлэгдсэн утга янз бүр болно. Тодруулбал хамгийн бага дундаж квадрат алдаатай үнэлэлт нь түүврээс хамаарсан функц юм.

Жишээ

$$\begin{cases} \Theta \sim U(0,1) \\ X = 3\Theta + U \quad U \sim U(-1,1) \quad \operatorname{cov}(U,\Theta) = 0 \end{cases}$$

тохиолдолд $\hat{\Theta} = E(\Theta|X = x)$ үнэлэлт олъё.



(a) (X, Θ) векторын авах утга (b) (X, Θ) векторын хамтын нягт

Зураг: (X, Θ) векторын хамтын тархалт ба $\hat{\Theta} = E(\Theta|X = x)$ үнэлэлт

$\hat{\Theta} = E(\Theta|X) = a + bX$ хэлбэртэй гэж үзвэл дараах үнэлэлт олдоно.

$$\hat{\Theta} = E(\Theta) + \frac{\operatorname{cov}(X,\Theta)}{D(X)}[X - E(X)] = \frac{1}{2} + \frac{3}{13}\left[X - \frac{3}{2}\right] = \frac{2}{13} + \frac{3}{13}X$$

бие даан судлах слайд

Өмнөх слайд дээрх жишээ бодлогын бодолтыг дэлгэрэнгүй үзье. Эхлээд (X, Θ) санамсаргүй векторын хамтын тархалтыг олъё. Үүний тулд нэн тэргүүнд боломжит утгын олонлогийг нь олох хэрэгтэй. Бодлогын нөхцөлийг харвал $0 \le \Theta \le 1$ харин X нь $X = 3\Theta$ шулуунаас 1 нэгж зайд байх буюу (X, Θ) санамсаргүй векторын авах утгуудын геометр байр зураг дээр үзүүлсэн параллелограмм болно. Улмаар Θ болон U санамсаргүй хувьсагчид жигд тархалттай тул $f_{X,\Theta}(x,\theta)$ буюу (X, Θ) санамсаргүй вектор тус параллелограмм дээр жигд тархана. Тархалт жигд тул дараах нягтын функц нь дараах хэлбэртэй байна.

$$f_{X,\Theta}(x,\theta) = \begin{cases} c, & (x,\theta) \in \text{параллелограмм} \\ 0, & (x,\theta) \notin \text{параллелограмм} \end{cases}$$

бие даан судлах слайд

Нягтын функцийн чанар ашиглавал

$$\begin{aligned} 1 &= \int_{\text{параллелограмм}} f_{X,\Theta}(x,\theta)dxd\theta \\ &= c\int_{\text{параллелограмм}} dxd\theta \\ &= c \cdot S_{\text{параллелограмм}} \\ &= 2c \end{aligned}$$

байдлаар $c = 0.5$ гэж олдоно. Иймхүү

$$f_{X,\Theta}(x,\theta) = \begin{cases} 0.5, & (x,\theta) \in \text{параллелограмм} \\ 0, & (x,\theta) \notin \text{параллелограмм} \end{cases}$$

боллоо.

бие даан судлах слайд

Энэсийн зорилго бол нөхцөлт математик дундаж олох явдал тул нэн тэргүүнд нөхцөлт тархалт олох шаардлагатай. Хамтын тархалтыг харвал

$$f_{\Theta|X}(\theta|X = x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)}$$

нөхцөлт тархалтыг олоход ашиглагдах $f_X(x)$ тухайн нягт $-1 \le x \le 1$, $1 \le x \le 2$ болон $2 \le x \le 4$ завсар бүрт өөр өөр байх ажээ. Одоо эдгээрийг тус тусад нь бодож олъё.

$-1 \le x \le 1$ үед параллелограммын дээд ирмэгт $y = \frac{1}{3}x + \frac{1}{3}$ шулуун харгалзах бөгөөд $0 \le \Theta \le x/3 + 1/3$ байх тул X санамсаргүй хувьсагчийн тухайн нягтын илэрхийлэл

$$f_X(x) = \int_0^{x/3+1/3} 0.5 d\theta = \frac{x+1}{6} \quad -1 \le x \le 1$$

гэж олдоно. Тэгвэл

$$f_{\Theta|X}(\theta|X = x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{0.5}{\frac{x+1}{6}} = \frac{3}{x+1} \quad -1 \le x \le 1$$

нөхцөлт нягт олдоно. Улмаар

$$\hat{\Theta} = E(\Theta|X = x) = \int_0^{x/3+1/3} \theta \frac{3}{x+1} d\theta$$
$$= \frac{3}{x+1}\left[\frac{\theta^2}{2}\right]_0^{x/3+1/3} = \frac{x+1}{6} \quad -1 \le x \le 1$$

үнэлэлт олдоно.

$1 \le x \le 2$ үед $x/3 - 1/3 \le \Theta \le x/3 + 1/3$ байх тул X санамсаргүй хувьсагчийн тухайн нягтын илэрхийлэл

$$f_X(x) = \int_{x/3-1/3}^{x/3+1/3} 0.5 d\theta = \frac{1}{3} \quad 1 \le x \le 2$$

гэж олдоно. Тэгвэл

$$f_{\Theta|X}(\theta|X = x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{0.5}{\frac{1}{3}} = \frac{3}{2} \quad 1 \le x \le 2$$

нөхцөлт нягт олдоно. Улмаар

$$\hat{\Theta} = E(\Theta|X = x) = \int_{x/3-1/3}^{x/3+1/3} \theta \frac{3}{2} d\theta$$
$$= \frac{3}{2}\left[\frac{\theta^2}{2}\right]_{x/3-1/3}^{x/3+1/3} = \frac{x}{3} \quad 1 \le x \le 2$$

үнэлэлт олдоно.

$2 \le x \le 4$ үед $x/3 - 1/3 \le \Theta \le 1$ байх тул X санамсаргүй хувьсагчийн тухайн нягтын илэрхийлэл

$$f_X(x) = \int_{x/3-1/3}^{1} 0.5 d\theta = \frac{4-x}{6} \quad 2 \le x \le 4$$

гэж олдоно. Тэгвэл

$$f_{\Theta|X}(\theta|X = x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{0.5}{\frac{4-x}{6}} = \frac{3}{4-x} \quad 2 \le x \le 4$$

нөхцөлт нягт олдоно. Улмаар

$$\hat{\Theta} = E(\Theta|X = x) = \int_{x/3-1/3}^{1} \theta \frac{3}{4-x} d\theta$$
$$= \frac{3}{4-x}\left[\frac{\theta^2}{2}\right]_{x/3-1/3}^{1} = \frac{x+2}{6} \quad 2 \le x \le 4$$

үнэлэлт олдоно.

Олсон үнэлэлтээ зураг дээр (тахир шугам) зурж харуулав. Одоо харин үнэлэлтийг $\hat{\Theta} = E(\Theta|X) = a + bX$ буюу шугаман хэлбэртэй гэвэл чухам ямар үр дүн гарахыг харъя. $Y = a + bX$ шугаман загварын хувьд

$$a = E(Y) - bE(X)$$
$$b = \frac{\mathrm{cov}(X,Y)}{D(X)}$$

байдаг (үүнийг хожим үзнэ) тул $Y = E(Y) + \frac{\mathrm{cov}(X,Y)}{D(X)}[X - E(X)]$ буюу тус бодлогын хувьд

$$\hat{\Theta} = E(\Theta) + \frac{\mathrm{cov}(X,\Theta)}{D(X)}[X - E(X)]$$

тэгшитгэл бичигдэнэ.

$\Theta \sim U(0,1)$ тул $E(\Theta) = \frac{0+1}{2} = \frac{1}{2}$ байна.

$$E(X) = E(3\Theta + U) = 3E(\Theta) + E(U) = \frac{3}{2}$$
$$D(X) = D(3\Theta + U) = 9D(\Theta) + D(U) + 2 \cdot 3\,\mathrm{cov}(\Theta, U)$$
$$= 9 \cdot \frac{1}{12} + \frac{2^2}{12} + 0 = \frac{13}{12}$$

U болон Θ хамааралгүй ($\mathrm{cov}(U,\theta) = 0$) болохыг анхаарвал

$$\mathrm{cov}(X,\Theta) = E(X\Theta) - E(X)E(\Theta) = E[(3\Theta + U)\Theta] - \frac{3}{2}\cdot\frac{1}{2}$$
$$= 3E(\Theta^2) + E(U\Theta) - \frac{3}{4} = 3[D(\Theta) + (E(\Theta)^2)] - \frac{3}{4}$$
$$= 3\left[\frac{1}{12} + \frac{1}{4}\right] - \frac{3}{4} = \frac{1}{4}$$

болно.

Ийнхүү

$$\hat{\Theta} = \frac{1}{2} + \frac{3}{13}\left[X - \frac{3}{2}\right] = \frac{2}{13} + \frac{3}{13}X \approx 0.154 + 0.231X$$

хэлбэртэй шугаман үнэлэлт олдлоо. Үүнийг зураг дээр нэмж зурав. Зургаас шугаман загвараар олдсон үнэлэлт хамгийн бага дундаж квадрат алдаатай үнэлэлтээс хэр зэрэг зөрж байгааг харж болно.

Эцэст нь дээрх бодолтын зарим хэсгийг симуляцын туршилтаар шалгая. Эхлээд дараах тушаалаар анх авсан загварт тохирох өгөгдөл гаргаж авна.

```
set.seed(0)
n <- 5000
Theta <- runif("n" = n)
X <- 3 * Theta + runif("n" = n, "min" = -1, "max" = 1)
```

Уг өгөгдлөөр цэгэн диаграмм байгуулбал (X, Θ) тархалтын ерөнхий төрх харагдана.

```
plot(x = X, y = Theta, cex = 0.1, col = "gray", asp = 1)
```

Диаграммын төрх нь дээр олсон параллелограмм шиг бас цэгүүдийн тархалт жигд байгаа нь дээрх бодолтын хамтын тархалтад холбогдох хэсэг зөв болохыг харуулж байна.

Мөн сүүлд олсон параметрийн шугаман үнэлэлтийн коэффициентуудыг өгөгдлөө ашиглан олж бас өмнө байгуулсан цэгэн диаграмм дээрээ харгалзах шулууныг нэмж зурахын тулд дараах тушаал өгч болно.

```
coefs <- coefficients(lm(formula = Theta ~ X))
print(coefs)
abline(a = coefs[1], b = coefs[2], col = "blue")
```

Эндээс шулууны коэффициентууд ойролцоогоор 0.149 болон 0.232 гэж олдож байгаа нь өмнө олсон $a \approx 0.154$ болон $b \approx 0.231$ аналитик утгуудтай ойролцоо байна.

Хамгийн бага дундаж квадрат алдаатай үнэлэлт, нөхцөлт математик дундаж, шугаман загварын зарим чанар

Хамгийн бага дундаж квадрат алдаатай үнэлэлтийн алдаа

$$\tilde{\Theta} = \hat{\Theta} - \Theta$$

байх бөгөөд өмнө үзсэн нөхцөлт математик дундаж болон шугаман загварын сэдвийг эргэн санавал тус үнэлэлтийн алдааны хувьд дараах чанарууд илэрхий юм.

1. $E(\tilde{\Theta}|X) = 0$
2. $E(\tilde{\Theta}) = 0$ бөгөөд иймд $\hat{\Theta}$ хазайлтгүй үнэлэлт юм.
3. $\forall g(\cdot)$ функцийн хувьд $E(\tilde{\Theta}g(X)) = 0$
4. $\mathrm{cov}(\tilde{\Theta}, \hat{\Theta}) = 0$
5. $D(\Theta) = D(\hat{\Theta}) + D(\tilde{\Theta})$

Нөхцөлт математик дундаж, шугаман загварын мөн чанарыг сэргээн санулах үүднээс эдгээр чанарын баталгааг хойно оруулав.

1. $\hat{\Theta} = E(\Theta|X)$ тул
$$\begin{aligned} E(\tilde{\Theta}|X) &= E(\hat{\Theta} - \Theta|X) = E(\hat{\Theta}|X) - E(\Theta|X) \\ &= E(\hat{\Theta}|X) - E(\Theta|X) = E(E(\Theta|X)|X) - E(\Theta|X) \\ &= E(\hat{\Theta}|X) - E(\Theta|X) = \hat{\Theta} - \hat{\Theta} = 0 \end{aligned}$$
2. Бүтэн дунджийн томьёо болон 1 дүгээр чанар ашиглавал $E(\tilde{\Theta}) = E(E(\tilde{\Theta}|X)) = E(0) = 0$
3. Бүтэн дунджийн томьёо болон 1 дүгээр чанар ашиглавал
$$E(\tilde{\Theta}g(X)) = E[E(\tilde{\Theta}g(X)|X)] = E[g(X)E(\tilde{\Theta}|X)] = 0$$
4. Ковариацыг задалсаны дараа дурын $h(X)$ үнэлэлтийн хувьд 3 дугаар чанар бас 2 дугаар чанар ашиглавал
$$\mathrm{cov}(\tilde{\Theta}, h(X)) = E(\tilde{\Theta}h(X)) - E(\tilde{\Theta})E(h(X)) = 0$$
болох тул $\hat{\Theta}$ үнэлэлтийн хувьд ч $\mathrm{cov}(\tilde{\Theta}, \hat{\Theta}) = 0$ байх юм.
5. 4 дүгээр чанарыг ашиглавал
$$D(\Theta) = D(\hat{\Theta} - \tilde{\Theta}) = D(\hat{\Theta}) + D(\tilde{\Theta}) - 2\,\mathrm{cov}(\hat{\Theta}, \tilde{\Theta}) = D(\hat{\Theta}) + D(\tilde{\Theta})$$

Лекц XIV

Статистик таамаглал шалгах, тархалтын загварын тохирцыг тогтоох

Статистик таамаглал шалгах, тархалтын загварын тохирцыг тогтоох сэдвийн агуулга

Хэсэг 1

Таамаглал шалгах

Статистик таамаглал

Тархалтын параметрийн талаарх таамаг төсөөллийг *статистик таамаглал* гэнэ.

Статистикт нэг нь нөгөөгөө үгүйсгэсэн хоёр таамаглалыг зэрэг авдаг. Тэдний нэгийг тэг, нөгөөг өрсөлдөгч таамаглал гээд харгалзан H_0, H_1 гэж тэмдэглэнэ.

Практикт ихэвчлэн тархалтын үл мэдэгдэх параметрийн тухай дараах гурван таамаглалын аль нэгийг авч үздэг.

- Хоёр талт таамаглал
$$H_0 : \theta = \theta_0 \text{ ба } H_1 : \theta \neq \theta_0 \text{ хоёр талт өрсөлдөгч таамаглал}$$
- Нэг талт таамаглал
$$H_0 : \theta = \theta_0 \text{ ба } H_1 : \theta < \theta_0 \text{ зүүн өрсөлдөгч таамаглал}$$
$$H_0 : \theta = \theta_0 \text{ ба } H_1 : \theta > \theta_0 \text{ баруун өрсөлдөгч таамаглал}$$

Энд θ нь үл мэдэгдэх параметр, θ_0 нь таамаглаж буй утга юм.

H_0 таамаглалын зүгээс харвал эх олонлог дараах байдлаар үл огтлолцох хоёр хэсэгт хуваагдана.

$$\{\text{Эх олонлог}\} = \{H_0 \text{ үнэн байх олонлог}\} \cup \{H_0 \text{ худал байх олонлог}\}$$

H_0 таамаглалыг шалгахын тулд бидэнд тархалтын эх олонлогийн талаарх мэдээлэл шаардлагатай бөгөөд тийм мэдээлэлтэй болохын тулд түүвэр авдаг билээ. Хэрэв H_0 худал байх олонлог мэдэгдэх бол дараах байдлаар таамаглалд хариулт өгнө.

$$H_0 \text{ таамаглал} \begin{cases} \text{худал} & \text{хэрэв түүвэр} \in \{H_0 \text{ худал байх олонлог}\} \\ \text{үнэн} & \text{хэрэв түүвэр} \notin \{H_0 \text{ худал байх олонлог}\} \end{cases}$$

Гэвч практикт H_0 худал байх олонлог үргэлж мэдэгдэх албагүй. Иймд тус олонлогийг үнэлж олох хэрэгтэй.

Статистик шинжүүр

Тэг таамаглалыг хүлээн авах эсвэл няцаах шийдвэр гаргах дүрмийг *шинжүүр* гэдэг. Шинжүүрийн зүгээс тэг таамаглал худал байх олонлогийг *шинжүүрийн няцаах муж* гэдэг. Иймд таамаглал шалгахын тулд шинжүүрийн няцаах мужийг олох хэрэгтэй.

Шинжүүрийн няцаах муж олдсон үед тэг таамаглалд дараах байдлаар хариулт өгнө.

$$H_0 \text{ таамаглал} \begin{cases} \text{худал} & \text{хэрэв түүвэр} \in \text{шинжүүрийн няцаах муж} \\ \text{үнэн} & \text{хэрэв түүвэр} \notin \text{шинжүүрийн няцаах муж} \end{cases}$$

Таамаглал шалгахад гарах алдаа

		Үнэн бодит байдал	
		H_0 үнэн	H_0 худал
Шинжүүр	H_0 худал	I төрлийн алдаа	зөв шийдвэр
	H_0 үнэн	зөв шийдвэр	II төрлийн алдаа

Хүснэгт: Таамаглал шалгахад гарах алдаа

I төрлийн алдааны магадлал хамгийн ихдээ хэд байж болохыг заасан тоог *итгэх түвшин* гээд α гэж тэмдэглэнэ. Практикт итгэх түвшинг $\alpha = 0.01$, $\alpha = 0.05$ гэх мэтчилэн багаар сонгож авдаг.

Шинжүүрийн статистик

Шинжүүрийн статистик бол тэг таамаглалын үнэн эсэхийг шалгахаар зохиосон статистик юм.

$$\text{Статистик :}\quad \begin{array}{c} \left\{\begin{array}{c}\text{Эх}\\ \text{олонлог}\end{array}\right\} \\ \Downarrow \\ \mathbb{R} \end{array} = \begin{array}{c} \left\{\begin{array}{c}H_0\ \text{үнэн}\\ \text{байх олонлог}\end{array}\right\} \\ \Downarrow \\ \text{интервал} \end{array} \cup \begin{array}{c} \left\{\begin{array}{c}H_0\ \text{худал}\\ \text{байх олонлог}\end{array}\right\} \\ \Downarrow \\ \text{интервал} \end{array}$$

түүвэр $\to$ тоон шулууны цэг

H_0 худал байх олонлог мэдэгдэхгүй тул түүний оронд шинжүүрийн няцаах муж ашиглана. Шинжүүрийн няцаах мужийг I төрлийн алдаанд тулгуурлаж олно.

$$P(\text{I төрлийн алдаа}) = P(H_1|H_0) \le \alpha \Rightarrow \text{шинжүүрийн няцаах муж}$$

Жишээ

Богдын сантай холбоотой жишээг эргэн авч үзье. Тухайн үед симуляцын аргаар гарган авч байсан 0.68, 0.56, 0.70, 0.14, 4.36, 0.71, 2.09, 0.39, 0.26, 0.45, 1.38, 1.53, 0.28, 2.10, 2.83, 1.03, 1.80, 0.59, 0.06, 1.70, 0.48, 0.71, 0.11, 0.28, 0.17 өгөгдөлд тулгуурлаж илтгэгч тархалтын эрчмийн параметрийн тухай

$$H_0 : \lambda = \lambda_0 \equiv 1$$
$$H_1 : \lambda < \lambda_0$$

нэг талт зүүн өрсөлдөгчтэй таамаглалыг $\alpha = 0.05$ итгэх түвшинд шалга.

Шинжүүрийн няцаах муж олохын тулд $P(\text{I төрлийн алдаа}) \le \alpha$ буюу $P(H_0$ таамаглалыг няцаах$|H_0$ үнэн байх$) \le \alpha$ тэнцэл биш бодох тул эхлээд таамаглал шалгахад ашиглах шинжүүрийн статистик зохиох улмаар түүнийхээ тархалтыг H_0 буюу тэг таамаглал үнэн гэсэн нөхцөлд хайх шаардлагатай.

Сэргээн санах нь

$\mathrm{Exp}(\lambda)$ тархалттай санамсаргүй хувьсагчийн эх олонлогоос авсан түүврийн хувьд $2\lambda n\overline{X} \sim \chi^2(2n)$ байна.

Ийнхүү жишээнд дэвшүүлсэн таамаглалыг шалгахын тулд

$$X^2 = 2\lambda_0 n\overline{X} \sim \chi^2(2n)$$

гэсэн статистик авч үзэж болох юм. Нөгөө талаас илтгэгч тархалттай санамсаргүй хувьсагчийн математик дундаж эрчмийн параметрийнхээ урвуутай тэнцүү байдгийг анхаарвал хэрэв H_0 үнэн бол түүврийн дунджийн урвуу ба таамаглаж буй утга хоёр ойролцоо, харин H_0 худал бол тус хоёр утга эсрэгээрээ зөрүүтэй байх тул X^2 нь шинжүүрийн статистик болж чадна.

X^2 шинжүүрийн статистикийн хувьд

$$P(\text{I төрлийн алдаа}) = P(H_1|H_0) \le \alpha$$

тэнцэл биш

- Хоёр талт өрсөлдөгчтэй үед $P(X^2 \le \chi^2_{1-\alpha/2,2n}$ эсвэл $X^2 \ge \chi^2_{\alpha/2,2n}) \le \alpha$
- Нэг талт зүүн өрсөлдөгчтэй үед $P(X^2 \le \chi^2_{1-\alpha,2n}) \le \alpha$
- Нэг талт баруун өрсөлдөгчтэй үед $P(X^2 \ge \chi^2_{\alpha,2n}) \le \alpha$

хэлбэрт шилжинэ. Иймд шинжүүрийн няцаах муж нь харгалзан

- $X^2 \le \chi^2_{1-\alpha/2,2n}$ эсвэл $X^2 \ge \chi^2_{\alpha/2,2n}$
- $X^2 \le \chi^2_{1-\alpha,2n}$
- $X^2 \ge \chi^2_{\alpha,2n}$

болно. Энд $\chi^2_{\alpha,k}$ нь k чөлөөний зэрэгтэй хи-квадрат тархалтын $1-\alpha$ эрэмбийн квантилын утга буюу α хэмжээтэй талбай бүхий тархалтын баруун сүүлний утга юм.

Шинжүүрийн няцаах муж бэлэн болсон тул таамаглалаа шалгая. Бодлогын нөхцөлд өгсөн 0.68, 0.56, 0.70, 0.14, 4.36, 0.71, 2.09, 0.39, 0.26, 0.45, 1.38, 1.53, 0.28, 2.10, 2.83, 1.03, 1.80, 0.59, 0.06, 1.70, 0.48, 0.71, 0.11, 0.28, 0.17 түүврийн хувьд хэмжээ нь $n = 25$, дундаж нь $\overline{X} \approx 1.016$ тул шинжүүрийн статистикийн туршилтын утга

$$X^2 = 2\lambda_0 n\overline{X} = 2 \cdot 1 \cdot 25 \cdot 1.016 = 50.8$$

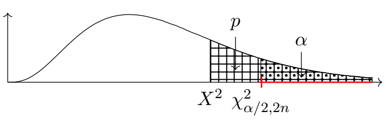
болно. $X^2 = 50.8 \cancel{\le} \chi^2_{1-\alpha,2n} = \chi^2_{0.95,50} \approx 34.764$ буюу шинжүүрийн статистикийн туршилтын утга шинжүүрийн няцаах мужид унахгүй байгаа тул тэг таамаглалыг үл няцаана. Өөрөөр хэлбэл илтгэгч тархалтын эрчмийн параметрийн тухай

$$H_0 : \lambda = \lambda_0 \equiv 1$$
$$H_1 : \lambda < \lambda_0$$

нэг талт зүүн өрсөлдөгчтэй таамаглалыг $\alpha = 0.05$ итгэх түвшинд хүлээн авна.

p-утга

Тэг таамаглалыг үнэн гэж тооцсон тохиолдолд тус таамаглалыг шалгах үед байсантай харьцуулахад тэг таамаглалд илүү эсрэг тэсрэг үр дүн гарах магадлалыг *магадлалын утга* буюу *p-утга* гэнэ. p-утга $< \alpha$ бол H_0 таамаглалыг худал, эсрэг тохиолдолд үнэн гэж дүгнэнэ.

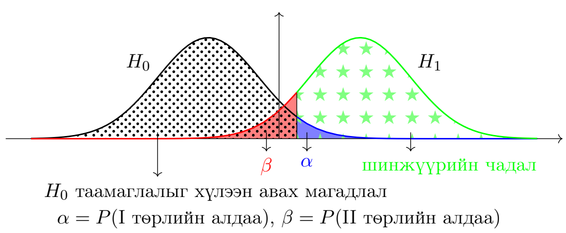


Зураг: Илтгэгч тархалтын эрчмийн параметрийн α итгэх түвшинтэй, нэг талт баруун өрсөлдөгчтэй тэг таамаглалын шинжүүрийн няцаах муж ба p-утга

Тэг таамаглалыг няцаах сонирхолтой байгаа үед α итгэх түвшинг *ач холбогдлын түвшин* гэдэг.

Шинжүүрийн чадал

Өрсөлдөгч таамаглал үнэн үед тэг таамаглалыг няцаах магадлалыг *шинжүүрийн чадал* гэнэ.



Зураг: Нэг талт шинжүүрийн I болон II төрлийн алдаа ба чадал

Шинжүүрийн няцаах муж ба итгэх завсар

Хоёр талт өрсөлдөгчтэй параметрийн таамаглалын α итгэх түвшинтэй шинжүүрийн няцаах мужийн гүйцээлт нь тус параметрийн $1-\alpha$ итгэх магадлал бүхий итгэх завсартай давхацдаг.

Сэргээн санах нь (Илтгэгч тархалтын параметрийн итгэх завсар)

$$\frac{\chi^2_{1-\alpha/2,2n}}{2n\overline{X}} < \lambda < \frac{\chi^2_{\alpha/2,2n}}{2n\overline{X}}$$

Өмнөх шинжүүрийн хоёр талт шинжүүрийн няцаах мужаас гүйцээлт авах буюу эсрэг нөхцөлийг нь бичээд $2n\overline{X}$ үржвэрт хуваавал дээрх итгэх завсар гарна.

$$\chi^2_{1-\alpha/2,2n} < X^2 \equiv 2\lambda n\overline{X} < \chi^2_{\alpha/2,2n}$$

$$\frac{\chi^2_{1-\alpha/2,2n}}{2n\overline{X}} < \lambda < \frac{\chi^2_{\alpha/2,2n}}{2n\overline{X}}$$

Хэсэг 2

Тархалтын хэлбэрийн тухай таамаглал шалгах

Параметрийн бус таамаглал

Статистикт өмнөх хэсэгт үзсэн шиг параметрийн тухай таамаглал авч үзэхийн зэрэгцээ параметрээс бусад зүйлийн талаарх таамаглалуудыг ч авч үздэг. Үүнд:

1. Санамсаргүй хувьсагчийн тархалтын хэлбэрийн тухай
2. Санамсаргүй хувьсагчдын тархалт ижил байх тухай
3. Санамсаргүй хувьсагчид хамааралгүй байх тухай
4. Түүвэр санамсаргүй байх тухай

Эдгээрээс санамсаргүй хувьсагчийн тархалтын хэлбэрийн тухай таамаглалыг энэ хэсэгт авч үзнэ.

Санамсаргүй хувьсагчийн тархалтын хэлбэрийн тухай таамаглал шалгах

X санамсаргүй хувьсагчийн тархалтын хэлбэрийн тухай таамаглал дараах байдалтай байна.

$$H_0 : F_X(x) = F_0(x)$$
$$H_0 : F_X(x) \neq F_0(x)$$

Энд $F_0(x)$ бол таамаглаж буй тархалтын функц юм.

Тус таамаглалыг шалгах олон шинжүүр байдгаас хи-квадрат шинжүүрийг сонгон авч үзнэ.

Тархалтын хэлбэрийн тухай таамаглал шалгах хи-квадрат шинжүүр

- Эх олонлогийн тархалт

x	x_1	x_2	$\ldots$	x_k
$f_X(x)$	$f_X(x_1)$	$f_X(x_2)$	$\ldots$	$f_X(x_k)$

- Таамаглал

$$H_0 : f_X(x_1) = p_1, \ldots, f_X(x_k) = p_k$$

Энд $p_1, \ldots, p_k$ бол таамаглаж буй тоонууд юм.

- Түүврийн давтамж

x	x_1	x_2	$\ldots$	x_k	Нийлбэр
Давтамж	n_1	n_2	$\ldots$	n_k	n

Энд n_i бол өгөгдөл дотор байх x_i утгын тоо ширхэг юм.

- Шинжүүрийн статистик

$$X^2_{k-1} = \sum_{i=1}^{k} \frac{(n_i - n\cdot p_i)^2}{n\cdot p_i} = \frac{(n_1 - n\cdot p_1)^2}{n\cdot p_1} + \ldots + \frac{(n_k - n\cdot p_k)^2}{n\cdot p_k}$$

- Шинжүүрийн статистикийн асимптот тархалт

Хэрэв H_0 үнэн бөгөөд түүврийн хэмжээ n хүрэлцээтэй их бол

$$X^2_{k-1} \sim \chi^2_{k-1}$$

байна.

- Шинжүүрийн няцаах муж, шинжүүрийн няцаах утга

$$X^2_{k-1} \geq \chi^2_{\alpha,k-1}$$

бол α итгэх түвшинд тэг таамаглалыг няцаана. Энд $\chi^2_{\alpha,k-1}$ нь $k-1$ чөлөөний зэрэгтэй хи-квадрат тархалтын $1-\alpha$ эрэмбийн квантилын утга буюу α хэмжээтэй талбай бүхий тархалтын баруун сүүлний утга юм.

Жишээ

"Сүм хийдийн хамаарах шашин" гэсэн дискрет хувьсагчийг

$$H_0 : P(\text{Будда}) = 0.40,\ P(\text{Христ}) = 0.50$$
$$P(\text{Ислам}) = 0.05,\ P(\text{Бусад}) = 0.05$$

тархалттай гэсэн таамаглалыг $\alpha = 0.05$ итгэх түвшинд шалга.

Шашин	Будда	Христ	Ислам	Бусад	Нийлбэр
Давтамж	134	196	24	10	364

Хүснэгт: Сүм хийдийн тоо, шашны төрлөөр, 2018 оны эцэст, ҮСХ

$$X^2_3 = \frac{(134 - 364\cdot 0.4)^2}{364\cdot 0.4} + \frac{(196 - 364\cdot 0.5)^2}{364\cdot 0.5} + \frac{(24 - 364\cdot 0.05)^2}{364\cdot 0.05} + \frac{(10 - 364\cdot 0.05)^2}{364\cdot 0.05} \approx 7.544$$

Энэ нь $\chi^2_{3,0.05} = 7.815$ буюу шинжүүрийн няцаах утгаас их биш байгаа тул тэг таамаглалыг $\alpha = 0.05$ итгэх түвшинд үл няцаана.

Тус таамаглалыг R програм дээр дараах байдлаар шалгах боломжтой.

```
chisq.test(x = c(134,196,24,10), p = c(0.4,0.5,0.05,0.05))
```

Үр дүн

```
Chi-squared test for given probabilities

data:  c(134, 196, 24, 10)
X-squared = 7.544, df = 3, p-value = 0.05644
```

Дээрх p-утга дараах байдлаар гарна.

$$p\text{-утга} = P(\chi^2_3 \geq X^2_3) = P(\chi^2_3 \geq 7.544) \approx 0.05644$$

Үүнийг R програм дээр дараах байдлаар тооцоолж олно.

```
pchisq(q = 7.544, df = 3, lower.tail = FALSE)
```

Тасралтгүй хувьсагчийн тархалтын хэлбэрийн тухай таамаглал шалгахад хи-квадрат шинжүүр ашиглах нь

Тархалтын хэлбэрийн тухай таамаглал шалгахад ашигладаг хи-квадрат шинжүүрийг тасралтгүй санамсаргүй хувьсагчийн хувьд ашиглахдаа тоон өгөгдлийг бүлэглэдэг. Ингээд өгөгдөл дэх тоон утга нэг бүрийг түүний харьяалагдах бүлгээр төлөөлүүлж авдаг.

Жишээ

Богдын сантай холбоотой жишээг эргэн авч үзье. Тухайн үед симуляцын аргаар гарган авч байсан 0.68, 0.56, 0.70, 0.14, 4.36, 0.71, 2.09, 0.39, 0.26, 0.45, 1.38, 1.53, 0.28, 2.10, 2.83, 1.03, 1.80, 0.59, 0.06, 1.70, 0.48, 0.71, 0.11, 0.28, 0.17 өгөгдөлд тулгуурлаж, Богдын сандод орох орлогын эцсийн хэмжээ гэсэн хувьсагчийг $\lambda = 1$ эрчмийн параметр бүхий илтгэгч тархалттай гэсэн таамаглалыг $\alpha = 0.05$ итгэх түвшинд шалга.

Тоон өгөгдлийг R програмын `hist()` функцийнх шиг 0, 1, 2, 3, 4, 5 цэгүүдээр байгуулагдах таван интервалд бүлэглэхэд $(2, 3]$, $(3, 4]$, $(4, 5]$ интервалд харгалзах давтамж буюу тус интервалд харьяалагдах утгуудын тоо 3, 0, 1 байна. Хи-квадрат шинжүүрийн хувьд үүн шиг давтамж багатай интервалуудыг нэгтгэхийг зөвлөдөг. Иймд өгөгдлийг $[0, 1]$, $(1, 2]$, $(2, \infty)$ гурван интервалд хувааж бүлэглэе. Тэгвэл 16, 5, 4 давтамж олдоно.

Одоо дээрх интервалуудад харгалзах магадлалуудыг олъё. Тус магадлалуудыг тэг таамаглалын нөхцөлд өөрөөр хэлбэл $\lambda = 1$ эрчмийн параметр бүхий илтгэгч тархалтаар олно.

$$
\begin{aligned}
p_1 &= P(X \in [0, 1]) = P(X < 1) = F_X(1) = 1 - e^{-1 \cdot 1} \approx 0.632 \\
p_2 &= P(X \in (1, 2]) = P(1 < X \le 2) = F_X(2) - F_X(1) \\
&= (1 - e^{-1 \cdot 2}) - (1 - e^{-1 \cdot 1}) \approx 0.232 \\
p_3 &= P(X \in (2, \infty)) = P(2 < X) = 1 - F_X(2) \\
&= 1 - (1 - e^{-1 \cdot 2}) \approx 0.135
\end{aligned}
$$

$$X_2^2 = \frac{(16 - 25 \cdot 0.632)^2}{25 \cdot 0.632} + \frac{(5 - 25 \cdot 0.232)^2}{25 \cdot 0.232} + \frac{(4 - 25 \cdot 0.135)^2}{25 \cdot 0.135} \approx 0.229$$

Энэ нь $\chi^2_{2, 0.05} = 5.991$ буюу шинжүүрийн няцаах утгаас их биш байгаа тул тэг таамаглалыг $\alpha = 0.05$ итгэх түвшинд үл няцаана.

Эдгээр бүх үйлдлийг R програм дээр дараах байдлаар гүйцэтгэнэ.

```
X <- c(0.68, 0.56, 0.70, 0.14, 4.36, 0.71, 2.09, 0.39,
    0.26, 0.45, 1.38, 1.53, 0.28, 2.10, 2.83, 1.03, 1.80,
    0.59, 0.06, 1.70, 0.48, 0.71, 0.11, 0.28, 0.17)
breaks <- c(0,1,2,Inf)
contingencies <- table(cut(x = X, breaks = breaks))
p <- diff(pexp(q = breaks, rate = 1))
chisq.test(x = contingencies, p = p)
```

Үүгээр дараах үр дүн гарна.

```
Chi-squared test for given probabilities
data:  contingencies
X-squared = 0.2287, df = 2, p-value = 0.8919
```


Лекц XV

Үнэний хувийн харьцаат шинжүүр, регрессийн шугаман загвар

Үнэний хувийн харьцаат шинжүүр, регрессийн шугаман загвар сэдвийн агуулга

Хэсэг 1

Үнэний хувийн харьцаат шинжүүр

Үнэний хувийн харьцаат шинжүүр

Хамгийн их үнэний хувь бүхий үнэлэлт болон хамгийн их постериорын үнэлэлт нь тархалтын үл мэдэгдэх параметрийг хамгийн их нягттай буюу хамгийн үнэмшилтэй утгаар авдаг. Статистик таамаглалыг ч ийм байдлаар шалгаж болно.

Байесын зарчим баримталбал

$$P(H_1 | X = x) \ge P(H_0 | X = x)$$

буюу

$$\frac{P(X = x | H_1) P(H_1)}{P(X = x)} \ge \frac{P(X = x | H_0) P(H_0)}{P(X = x)}$$

эсвэл үүнтэй тэнцүү чанартай

$$LR(X) = \frac{P(X = x | H_1)}{P(X = x | H_0)} \ge \frac{P(H_0)}{P(H_1)} = c$$

нөхцөл биелж байвал H_0 таамаглалыг няцааж, H_1 таамаглалыг хүлээн авна.

Харин Байесын бус хувилбар нь

$$LR = \frac{P(X = x | H_1)}{P(X = x | H_0)} \ge c \quad \text{(дискрет тохиолдолд)}$$

$$LR = \frac{f_X(x | H_1)}{f_X(x | H_0)} \ge c \quad \text{(тасралтгүй тохиолдолд)}$$

үед H_0 таамаглалыг няцааж, H_1 таамаглалыг хүлээн авна.

Энд шинжүүрийн няцаах утга c хэдтэй тэнцүү байхаас аль төрлийн алдаа ямар хэмжээтэй гарах нь шалтгаална. c нь Байесын хувилбарт приор магадлалуудын харьцаагаар тодорхойлогдож байсан. Харин Байесын бус хувилбарт

$$P(\text{I төрлийн алдаа}) = P(H_1 | H_0) = P(LR \ge c | H_0) = \alpha$$

нөхцөлөөр тодорхойлогдоно. Энд α бол итгэх түвшин юм.

Жишээ

$N(\theta, \sigma^2)$ хэвийн тархалт авч үзье. Энд σ^2 мэдэгдэнэ. Тэгвэл тус тархалтын хувьд

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

таамаглал шалгах үнэний хувийн харьцаат шинжүүрийн няцаах мужийг ол. Энд $\theta_1 > \theta_0$ гэж тооцъё.

Таамаглал шалгахын тулд $N(\theta, \sigma^2)$ тархалттай санамсаргүй хувьсагчийн эх олонлогоос $X = (X_1, \dots, X_n)$ түүвэр авсан гэвэл шинжүүрийн няцаах муж дараах хэлбэртэй болно.

$$LR(X) = \prod_{i=1}^{n} \frac{f_X(X_i|H_1)}{f_X(X_i|H_0)} = \prod_{i=1}^{n} \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\theta_1)^2/(2\sigma^2)}}{\frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\theta_0)^2/(2\sigma^2)}}$$
$$= \exp\left\{\frac{n}{\sigma^2}(\theta_1 - \theta_0)\overline{X} - \frac{n}{2\sigma^2}(\theta_1^2 - \theta_0^2)\right\} \ge c$$

Үргэлжлүүлэн хувиргавал дараах тэнцэл биш гарна.

$$\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} \ge \frac{\sigma}{\sqrt{n}(\theta_1 - \theta_0)} \ln c + \frac{\sqrt{n}}{2\sigma}(\theta_1 - \theta_0) = c'$$

Тэг таамаглалын нөхцөлд $\overline{X} \sim N(\theta_0, \sigma^2/n)$ буюу $\frac{\overline{X}-\theta_0}{\sigma/\sqrt{n}} \sim N(0,1)$ тул

$$P(\text{I төрлийн алдаа}) = P(H_1|H_0) = P(LR \ge c|H_0)$$
$$= P\left(\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} \ge c'\right) = 1 - F_{\frac{\overline{X}-\theta_0}{\sigma/\sqrt{n}}}(c') = 1 - \Phi(c') = \Phi(-c') = \alpha$$

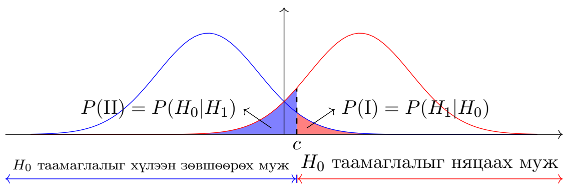
тэгшитгэлээс $c' = -\Phi^{-1}(\alpha) = \Phi^{-1}(1-\alpha)$ шийд олдоно. Энд $\Phi()$ нь стандарт хэвийн тархалтын функц юм. Иймхүү

$$\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} \ge \Phi^{-1}(1-\alpha)$$

хэлбэртэй шинжүүрийн няцаах муж олдлоо.

Үнэний хувийн харьцаат шинжүүрийн алдааны магадлал

Үнэний хувийн харьцаат шинжүүрийн хувьд I төрлийн алдаа багасахад II төрлийн алдаа ихсэж харин эсрэгээрээ I төрлийн алдаа ихсэхэд II төрлийн алдаа багасах бөгөөд энэ нь няцаах утгын сонголтоос шалтгаална.



Зураг: Үнэний хувийн харьцаат шинжүүрийн алдааны магадлал

Үнэний хувийн харьцаат шинжүүрийн асимптот няцаах муж

$H_0 : \theta = \theta_0$ параметрийн таамаглал шалгах байг. Энд $\theta_0 = (\theta_{1,0}, \dots, \theta_{k,0})$ буюу нийтдээ k ширхэг үл мэдэгдэх параметрийн тухай таамаглал шалгана. Тэгвэл түүврийн хэмжээ n хүрэлцээтэй их бол шинжүүрийн няцаах муж дараах хэлбэртэй байна.

$$-2\ln LR(X) \ge \chi^2_{\alpha,k}$$

Энд $\chi^2_{\alpha,k}$ нь k чөлөөний зэрэгтэй хи-квадрат тархалтын $1-\alpha$ эрэмбийн квантилын утга буюу α хэмжээтэй талбай бүхий тархалтын баруун сүүлний утга юм.

Дээрх шинжүүрийн няцаах мужийн хувьд параметрийн хамгийн их үнэний хувь бүхий үнэлэлтийг цор ганц оршин байхыг шаардсан нэмэлт нөхцөл тавьдаг.

Хэсэг 2

Регрессийн шугаман загвар

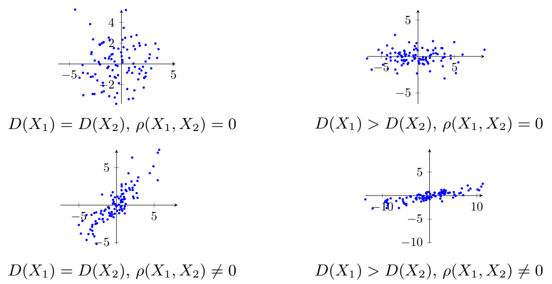
Корреляцын коэффициент ба шугаман хамаарал

Сэргээн санах нь

Хувьсагчид шугаман хамааралтай үед корреляцын коэффициент үнэмлэхүй утгаараа нэгтэй тэнцүү байдаг.

Олон хэмжээст хэвийн тархалт ба корреляцын коэффициент

Корреляцын коэффициент тэгээс ялгаатай, үнэмлэхүй утгаараа нэгд ойр үед олон хэмжээст хэвийн тархалттай санамсаргүй хувьсагчдын хамаарлыг шугаман функцээр илэрхийлэх боломжтой.

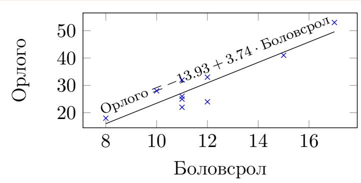


Зураг: Хоёр хэмжээст хэвийн тархалттай санамсаргүй утгууд

Регрессийн шугаман загвар

Сэргээн санах нь

X_1 болон X_2 хувьсагчид хамтдаа олон хэмжээст хэвийн тархалттай үед $E(X_2|X_1 = x_1)$ нь x_1 хувьсагчаас шугаман байдлаар хамаарсан функц байдаг тул X_2 хувьсагчийг X_1 хувьсагчийн шугаман эвлүүлгийн тусламжтай прогнозлох боломжтой. Тийм загварыг *регрессийн шугаман загвар* гэнэ.



Зураг: Боловсролд зарцуулсан хугацаа (жил) ба жилийн орлого (мян.$)

Нөхцөлт математик дундаж ашигласан загварын чанар

$X_2 = E(X_2|X_1) + U$ загвар дараах шинж чанартай болохыг өмнө үзсэн.

Сэргээн санах нь

1. $E(U|X_1) = 0$
2. $E(U) = 0$
3. $\operatorname{cov}(E(X_2|X_1), U) = 0$
4. $E(X_2|X_1)$ нь X_1 хувьсагч ашиглаж X_2 хувьсагчийг прогнозлох бүх $h(X_1) : \mathbb{R}^r \to \mathbb{R}^{p-r}$ функц дундаас хамгийн бага дундаж квадрат алдаатай (MSE $= E\left\{(X_2 - h(X_1))^T(X_2 - h(X_1))\right\}$) нь юм.

4-р чанар ёсоор $X_2 = E(X_2|X_1) + U$ бол хамгийн "сайн" загвар юм.

Детерминацын коэффициент

Сэргээн санах нь

$E(X_2|X_1 = x_1)$ нөхцөлт математик дунджийг X_2 хувьсагчийг прогнозлоход ашиглах бол харин

$$\rho^2 = \frac{\operatorname{cov}(X_2|X_1 = x_1)}{\operatorname{cov}(X_2)} = \frac{D(X_2|X_1 = x_1)}{D(X_2)}$$

буюу X_2 хувьсагчийн дисперсэд эзлэх $D(X_2|X_1 = x_1)$ нөхцөлт дисперсийн хувиар тус прогнозын илэрхийлэх чадварыг хэмжих бөгөөд ρ^2 хэмжигдэхүүнийг *детерминацын коэффициент* гэнэ. Мөн

$$\underbrace{D(X_2)}_{\text{нийт дисперс}} = \underbrace{D(X_2|X_1 = x_1)}_{\text{тайлбарлагдах дисперс}} + \underbrace{D(U)}_{\text{үл тайлбарлагдах дисперс}}$$

харьцаа хүчинтэй байдаг.

$X_2 = E(X_2|X_1) + U$ загварын хувьд

$$\underbrace{D(X_2)}_{\text{нийт дисперс}} = \underbrace{D(X_2|X_1 = x_1)}_{\text{тайлбарлагдах дисперс}} + \underbrace{D(U)}_{\text{үл тайлбарлагдах дисперс}}$$

байх тул детерминацын коэффициентыг

$$\rho^2 = \frac{D(X_2) - D(U)}{D(X_2)}$$

гэж олно.

Регрессийн шугаман загварын параметрийн үнэлэлт

Загварын параметрийг MSE буюу дундаж квадрат алдааг минимумчлах байдлаар үнэлдэг.

Өмнө дурдсанчлан $E(X_2|X_1)$ бол хамгийн бага дундаж квадрат алдаатай.

Сэргээн санах нь

X_1 болон X_2 хувьсагчид хамтдаа олон хэмжээст хэвийн тархалттай бол $E(X_2|X_1 = x_1) = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 + \Sigma_{21}\Sigma_{11}^{-1}x_1$ байна.

Иймд $X_2 = E(X_2|X_1) + U = a + bX_1 + U$ загварын параметрүүдийн дараах үнэлэлт гарна.

$$\hat{b} = \Sigma_{21}\Sigma_{11}^{-1} = \frac{\operatorname{cov}(X_1, X_2)}{D(X_1)}$$
$$\hat{a} = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 = E(X_2) - \hat{b}E(X_1)$$

Шинэ тэмдэглэгээ

Тэмдэглэгээг хялбаршуулахын тулд X_2 болон X_1 хувьсагчдыг харгалзан Y болон X гэе.

Тэгвэл регрессийн шугаман загварыг дараах байдлаар бичнэ.

$$Y = a + bX + U$$

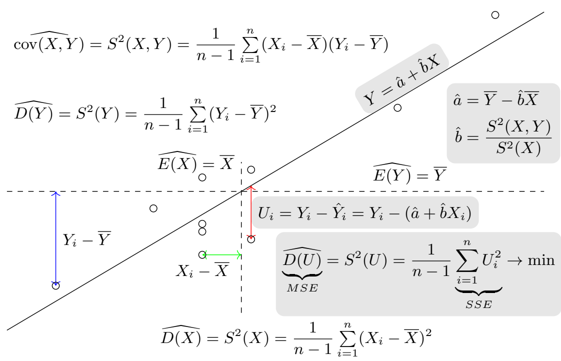
Мөн параметрийн үнэлэлтийн томьёо дараах хэлбэртэй болно.

$$\hat{a} = E(Y) - \hat{b}E(X) \qquad \hat{b} = \frac{\operatorname{cov}(X, Y)}{D(X)}$$

Түүврийн хувьд дээрх томьёонд буй моментуудыг харгалзах түүврийн моментуудаар үнэлж болох тул

$$\hat{a} = \overline{Y} - \hat{b}\overline{X} \qquad \hat{b} = \frac{S^2(X, Y)}{S^2(X)}$$

болно. Энд $\overline{X}$ болон $\overline{Y}$ нь түүврийн дундаж, $S^2(X)$ нь түүврийн дундаж квадрат хазайлт, $S^2(X, Y)$ нь түүврийн ковариацын коэффициент юм.



Зураг: Регрессийн шугаман загвар

Дундаж квадрат алдааг минимумчлах нь алдааны квадратуудын нийлбэрийг минимумчлахаас ялгаагүй юм. Иймд (X, Y) санамсаргүй векторын эх олонлогоос авсан $(X_1, Y_1), \ldots, (X_n, Y_n)$ хослоост түүврийн хувьд $Y = a + bX + U$ шугаман загварын параметрийн үнэлэлтийн томьёог алдааны квадратуудын нийлбэрийг минимумчлах байдлаар бодож олъё.

$$SSE = (n-1)S^2(U) = \sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n}(Y_i - (a + bX_i))^2$$

a болон b хувьсагчдаар тухайн уламжлал авч тэгтэй тэнцүүлбэл

$$\begin{cases} \dfrac{\partial SSE}{\partial a} = -2\displaystyle\sum_{i=1}^{n}(Y_i - (a + bX_i)) = 0 \\ \dfrac{\partial SSE}{\partial b} = -2\displaystyle\sum_{i=1}^{n}(Y_i - (a + bX_i))X_i = 0 \end{cases}$$

систем тэгшитгэл гарна.

Тэгшитгэлүүдийг $-2n$ үржвэрт хувааж түүврийн дундаж, түүврийн дундаж квадрат хазайлт болон түүврийн ковариацын коэффициент ашиглан хувиргаж бичвэл дараах тэгшитгэл гарна.

$$\begin{cases} \overline{Y} - a - b\overline{X} = 0 \\ \overline{X \cdot Y} - a\overline{X} - b\overline{X^2} = 0 \end{cases} \qquad \begin{cases} a = \overline{Y} - b\overline{X} \\ \underbrace{\overline{X \cdot Y} - \overline{X} \cdot \overline{Y}}_{S^2(X,Y)} - b(\underbrace{\overline{X^2} - \overline{X}^2}_{S^2(X)}) = 0 \end{cases}$$

Ингээд үнэлэлтийн дараах томьёо гарна.

$$\begin{cases} \hat{a} = \overline{Y} - b\overline{X} \\ \hat{b} = \dfrac{S^2(X,Y)}{S^2(X)} \end{cases}$$

Ийнхүү бодох аргыг *хамгийн бага квадратын арга* гэнэ.

R ашиглаж регрессийн шинжилгээ хийх

Өгөгдөл оруулах байдал

```
data <- data.frame(
  education = c(11, 12, 11, 15, 8, 10, 11, 12, 17, 11),
  annual.income = c(25, 33, 22, 41, 18, 28, 32, 24, 53, 26)
)
```

Загварын параметр үнэлэх

```
fit <- lm(formula = annual.income ~ education, data = data)
```

Прогноз

```
predict(object = fit, newdata = data.frame(
  education = c(16, 20)
))
```

Нэмэлт шинжилгээ

```
summary(fit)
```



```
Call:
lm(formula = annual.income ~ education, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9479 -1.9583  0.4219  3.0286  4.7917

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.9271     6.7802  -2.054 0.074038 .
education     3.7396     0.5631   6.641 0.000162 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.273 on 8 degrees of freedom
Multiple R-squared:  0.8465,  Adjusted R-squared:  0.8273
F-statistic: 44.11 on 1 and 8 DF,  p-value: 0.0001622
```

Дээрх үр дүнгийн товч тайлбарыг дараагийн слайдаар өгнө.

- Call:
  Функцийг дуудсан байдал буюу загвар, түүнд ашигласан өгөгдөл
- Residuals:
  Алдааны байршлын үзүүлэлтүүд
  хамгийн бага болон их утга, медиан, 25 болон 75 хувийн квантил
- Coefficients:
  Загварын параметрийн үнэлэлт, үнэлэлтийн стандарт алдаа, $H_0 : a = 0$ бас $H_0 : b = 0$ таамаглал шалгах t шинжүүрийн статистикийн туршилтын утга болон p-утга
- Residual standard error:
  Алдааны стандарт алдаа буюу стандарт хазайлт
- R-squared:
  Детерминацын коэффициент, ердийн болон засварласан
- F-statistic:
  H_0 : хувьсах хүчин зүйлс ач холбогдолгүй таамаглал шалгах F шинжүүрийн статистикийн туршилтын утга болон p-утга

Регрессийн шугаман загвар дээрх таамаглалууд

Регрессийн шугаман загварын хувьд дараах таамаглалуудыг хүчинтэй гэж үзэх буюу нэмэлт нөхцөл болгож тавьдаг.

1. U_i хэвийн тархалттай.
2. $U_1, \ldots, U_n$ алдаанууд хамааралгүй.
3. $D(U_i)$ тогтмол.
4. $Y = a + b_1X_1 + \ldots + b_kX_k + U$ загварын хувьд $X_1, \ldots, X_k$ тайлбарлах хувьсагчид хамааралгүй.

Лекц XVI

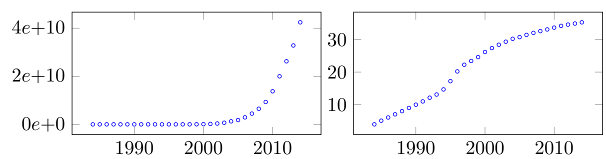
Хяналттай машин сургалтад ашиглах зарим статистик арга техник

Хяналттай машин сургалтад ашиглах зарим статистик арга техник сэдвийн агуулга

Хэсэг 1

Регрессийн шугаман загвар

Өгөгдөл ба бодлого



(a) Анхдагч тоон мэдээ (b) $\log_2$ хувиргалт

Зураг: Интернет сүлжээний урсгал, GB/сар, 1984-өөс 2014 он

Интернет сүлжээний урсгалын прогноз гаргах бодлого авч үзье. Үүний тулд эхлээд дараах шугаман загвар авч үзье.

$$\log_2(\text{урсгал}) = a + b \cdot \text{он}$$

R програмд өгөгдөл оруулах болон цэгэн диаграмм байгуулах байдал

```
data <- data.frame(
  year = 1984:2014,
  traffic = c(15, 33, 65, 128, 252, 498, 1000, 2002, 4444,
    8715, 25830, 150500, 1200000, 5000000, 11200000,
    25500000, 75250000, 175000000, 356000000, 681050000,
    1267800000, 1802745619, 2910579371, 4477367718,
    6491159470, 9301984735, 13751003569, 19974008812,
    26214897380, 32798830927, 42423169029)
)

plot(x = data$year, y = data$traffic, xlab = "Year", ylab
   = "Internet Traffic")
plot(x = data$year, y = log2(data$traffic), xlab = "Year",
   ylab = "Internet Traffic")
```

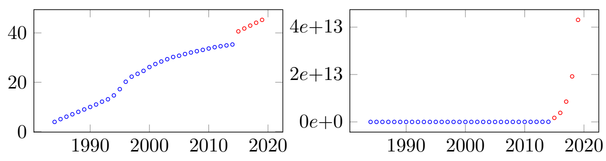

R програм дээрх загварын үнэлгээ болон прогноз

$\log_2(\text{урсгал}) = a + b \cdot \text{он}$ загварын үнэлгээ

```
fit <- lm(formula = log2(traffic) ~ year, data = data)
```

Прогноз

```
forecast <- 2 ** predict(object = fit, newdata =
   data.frame(
  year = 2015:2019
))
plot(x = c(data$year, 2015:2019), y = log2(c(data$traffic,
   forecast)), xlab = "Year", ylab = "Internet Traffic",
   pch = 20, col = c("black", "red")[rep.int(x = 1:2,
   times = c(length(data$year), 5))])
plot(x = c(data$year, 2015:2019), y = c(data$traffic,
   forecast), xlab = "Year", ylab = "Internet Traffic",
   pch = 20, col = c("black", "red")[rep.int(x = 1:2,
   times = c(length(data$year), 5))])
```

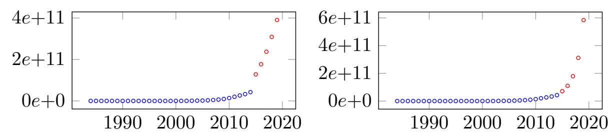
(a) $\log_2$ хувиргалттай (b) хувиргалтгүй

Зураг: Интернет сүлжээний урсгал, $\log_2(\text{урсгал}) = a + b \cdot \text{он}$ загварын тусламжтай гаргасан прогноз

Прогноз "залгаас" дээрээ огцом үсрэлттэй байгаа тул

$$\log_2(\text{урсгал}) = a + b \cdot \text{он}$$

загвар хэтийн прогноз гаргахад тохиромж муутай гэж үзнэ. Иймд өөр загвар авч үзье.



(a) $\log_2(\text{урсгал}) = a + b \cdot \text{он} + c \cdot \text{он}^2$ (b) $\log_2(\text{урсгал}) = a + b\,\text{он} + c\,\text{он}^2 + d\,\text{он}^3$

Зураг: Интернет сүлжээний урсгалын хэтийн төлөвийг олон гишүүнт бүхий загваруудаар прогнозлосон нь

$\log_2(\text{урсгал}) = a + b \cdot \text{он} + c \cdot \text{он}^2$ загварын прогноз бас л "муу" гарсан тул эцэст нь 1995 оноос хойших тоон мэдээг ашигласан

$$\log_2(\text{урсгал}) = a + b \cdot \text{он} + c \cdot \text{он}^2 + d \cdot \text{он}^3$$

загвар авч үзэв. Тус загварыг R дээр дараах байдлаар үнэлнэ.

```
fit <- lm(formula = log2(traffic) ~ poly(x = year, degree
   = 3), data = data, subset = year >= 1995)
```


Хэсэг 2

Авторегрессийн загвар

Авторегрессийн загвар

$Y = a + bX + U$ регрессийн шугаман сонгодог загварын хувьд Y хувьсагчийн утгуудыг хамааралгүй гэж тооцдог. Харин сая авч үзсэн интернет сүлжээний урсгалын хэмжээ гэсэн хувьсагч бол цаг хугацаатай уялдсан хамааралтай юм. Ийм процессыг статистикт *хугацаан цуваа* гэдэг бөгөөд үүнд тохирох олон янзын загвар авч үздэг. Тэдгээр загваруудын нэг бол авторегрессийн загвар юм.

$$X_t = b_0 + b_1 X_{t-1} + b_2 X_{t-2} + \ldots + b_p X_{t-p} + U_t$$

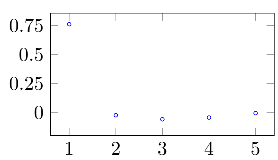
загварыг p эрэмбийн *авторегрессийн загвар* гэнэ. Энд U_t нь загварын алдаа юм.

Загварын хэрэглээг өмнөх хэсэгт ашигласан интернет урсгалын мэдээнд тулгуурлан үзье.

Авторегрессийн загварын эрэмбэ тогтоох

Загварын эрэмбэ тогтооход тухайн автокорреляц ашиглана.

```
partial_autocorrelations <- acf(x = data$traffic, lag.max
   = 5, type = "partial", plot = TRUE)
print(partial_autocorrelations)
```



Зураг: Интернет сүлжээний урсгал хувьсагчийн автокорреляц

Авторегрессийн загварын параметрүүдийг үнэлэх

Загварыг хамгийн бага квадратын аргаар яаж үнэлэхийг харуулав.

```
fit.ar <- ar.ols(x = data$traffic, order.max = 1)
```

Ийнхүү

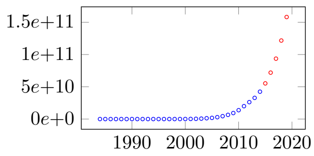
$$X_t = 1.783 \cdot 10^9 + 1.2978 X_{t-1}$$

загвар гарав.

Авторегрессийн загвар ашиглаж прогноз гаргах

Прогнозыг дараах байдлаар гаргана.

```
forecast <- predict(fit.ar, n.ahead = 5)$pred
plot(x = c(data$year, 2015:2019), y = c(data$traffic,
    forecast), xlab = "Year", ylab = "Internet Traffic",
    pch = 20, col = c("black", "red")[rep.int(x = 1:2,
    times = c(length(data$year), 5))])
```



Зураг: Интернет сүлжээний урсгалын хэтийн төлвийн прогноз

Хэсэг 3

Ложистик регресс

Ложистик регресс

$Y \sim \mathrm{Ber}(p)$ байх Y хувьсагчийн утгыг X хувьсагчийн тусламжтай прогнозлох зорилго тавья. Хэрэв

$$p = P(Y = 1) = E(Y|X) = a + bX$$

загвар авч үзвэл тэгээс бага эсвэл нэгээс их утгатай "буруу" прогноз гарах боломжтой. Үүнээс зайлсхийхийн тулд $(0, 1)$ завсарт утгатай

$$p = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

ложистик функц ашигладаг.

Ложит загвар

Өмнөх загварыг $\frac{p}{1-p} = e^{a+bX}$ байдлаар бичиж болно. $\frac{p}{1-p}$ *магадлалын харьцаа* нь $(0, \infty)$ засварт утгатай байх бөгөөд 0 болон ∞ нь p магадлалын бага болон их утгад харгалзана. Иймүү дараах загварын тусламжтай $p = P(Y = 1)$ магадлалыг үнэлж болно.

$$\ln\left(\frac{p}{1-p}\right) = a + bX$$

тэгшитгэлийн зүүн талыг *ложит* гэдэг.

Жишээ: Оюутан W дүн авах санал өгөх магадлал

Оюутан W дүн авах санал өгөх магадлалыг ирц болон явцын шалгалтын онооноос хамааруулан авч үзье. Энэ тохиолдолд

$$\ln\left(\frac{p}{1-p}\right) = a + b_{\text{ирц}} \cdot \text{ирц} + b_{\text{шалгалт}} \cdot \text{шалгалт}$$

загвар зохионо.

```
GP <- data.frame(
  W = c(0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,
    0,1,0,0,0,0,0,1,1,1,0,0,0,1,1,1,0,0,0),
  Attend = c(15,10,12,14,9,14,15,11,14,12,15,10,15,15,15,
    12,14,14,12,15,12,14,15,15,15,12,14,10,15,12,11,15,10,
    15,12,12,12,14,14,14,15,14,10) / 15 * 100,
  Exam = c(11,28,30,19,25,16,26,0,43,12,15,7,31,21,20,40,
    17,16,18,12,31,6,18,10,22,17,9,10,35,20,8,14,19,17,29,
    2,21,22,14,14,28,26,7) / 45 * 100)
```


Загварыг дараах байдлаар үнэлж, нэмэлт шинжилгээ хийнэ.

```
fit <- glm(formula = W ~ Attend + Exam, family =
    binomial(link = "logit"), data = GP)
summary(fit)
```

Эндээс $H_0 : b_{\text{ирц}} = 0$ болон $H_0 : b_{\text{шалгалт}} = 0$ таамаглалын магадлалын утгууд харгалзан 0.279 болон 0.038 гэж олдох тул ирцийг W үнэлгээнд нөлөөгүй гэж үзнэ. Үнэхээр уг хоёр хувьсагчийг W хувьсагчийн утгаар бүлэглээд дунджийг нь олж үзээд

W	Ирц	Шалгалт
0	13.0	20.3
1	13.4	12.8

буюу ирцийн хувьд мэдэгдэхүйц ялгаа ажиглагдахгүй байна. Иймд загварын томьёоллоос `Attend` хувьсагчийг зайлуулаад загвараа ахин үнэлнэ.

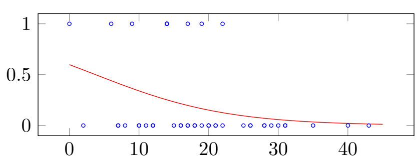
Иймүү эцэстээ

$$\ln\left(\frac{p}{1-p}\right) = 0.39805 - 0.10602 \cdot \text{шалгалт}$$

буюу

$$p = \frac{e^{a+bX}}{1 + e^{a+bX}} = \frac{1}{1 + e^{-(a+bX)}} = \frac{1}{1 + e^{-0.39805 + 0.10602 \cdot \text{шалгалт}}}$$

загвар олдоно.



Зураг: $p = P(\text{W үнэлгээний санал өгөх})$ магадлалын үнэлгээ ба өгөгдөл

Нэмэлт жишээ: Дуу хоолойгоор хүйс таних

www.kaggle.com/primaryobjects/voicegender веб хуудаснаас өгөгдөл татан авч `voice_gender` нэрээр ачаалав.

```
fit <- glm(formula = label ~ ., family = binomial(link =
    "logit"), data = voice_gender)
prop.table(table(
  ifelse(
    test = fitted(fit) > 0.5,
    yes = "male_prediction",
    no = "female_prediction"),
  voice_gender$label
), margin = 2)
```

```
                      female       male
female_prediction 0.97159091 0.02272727
male_prediction   0.02840909 0.97727273
```

Хэсэг 4

Гэнэн Байесын алгоритм

Гэнэн Байесын алгоритм

Гэнэн Байесын алгоритм нь Байесын зарчимд суурилсан ангиллын алгоритм юм. Тус алгоритм юмс үзэгдлийг хамгийн их постериор магадлалтай, өөрөөр хэлбэл

$$\operatorname*{argmax}_k P(C_k|X_1,\ldots,X_p)$$

дугаар ангид хуваарилдаг. Энд C_k нь туршилтын k дугаар үр дүн эсвэл анги, $X_1,\ldots,X_p$ нь санамсаргүй хувьсагчид юм. Байесын зарчим ёсоор

$$P(C_k|X_1,\ldots,X_p) = \frac{P(X_1,\ldots,X_p|C_k)P(C_k)}{P(X_1,\ldots,X_p)}$$

болох ба улмаар $X_1,\ldots,X_p$ хувьсагчдыг C_k үзэгдлийн нөхцөлд хамааралгүй гэсэн "гэнэхэн" таамаглалд найдаж

$$P(X_1,\ldots,X_p|C_k) = P(X_1|C_k)\cdot\ldots\cdot P(X_p|C_k)$$

гэж үздэг.

Ийнхүү

$$\operatorname*{argmax}_k P(C_k)\prod_{i=1}^{p} P(X_i|C_k)$$

бодлогыг бодож юмс үзэгдлийн харьяалагдах ангийн дэс дугаарыг олж тогтоодог. $P(C_k)$ буюу приор магадлалыг түүвэр дэх k дугаар ангийн давтамжаар үнэлдэг бол $P(X_i|C_k)$ магадлал буюу нягтыг X_i хувьсагчийн k дугаар анги дээрх тархалтын нягтын тусламжтай тооцоолдог. X_i хувьсагчийн хувьд хэвийн тархалт, Бернуллийн тархалт, мультиномиал тархалт зэргийг өргөн ашигладаг.

Жишээ: Цифр таних

Шаардлагатай багц суулгах

```
install.packages("klaR")
```

www.datahub.io/machine-learning/pendigits хуудас дээрх өгөгдөл ачаалах

```
data <- read.csv(file = "https://datahub.io/machine-
learning/pendigits/r/pendigits.csv", header = TRUE)
```

Мэдээлэл аль цифрт харгалзахыг илэрхийлсэн `class` хувьсагчийг зохих хэлбэрт оруулах

```
data$class <- as.factor(data$class)
```

Сургалтын болон тестийн өгөгдөл үүсгэх

```
set.seed(0)
subset <- sample.int(n = nrow(data), size = round(nrow(data) *
    0.8), replace = FALSE)
training.dataset <- data[subset,]
testing.dataset <- data[-subset,]
```


Ангиллын зарчим олж тогтоох буюу алгоритмаа сургах

```
classifier <- klaR::NaiveBayes(formula = class ~ ., data =
    training.dataset)
```

Ангиллын зарчим шалгах буюу алгоритмаа тестлэх

```
test <- predict(classifier, testing.dataset)
table("true" = testing.dataset$class, "classifier" =
    test$class)
```

Тестийн үр дүн

```
    classifier
true   0   1   2   3   4   5   6   7   8   9
   0 208   1   0   0   0   0   1   0  21   0
   1   0 149  43   1   0   5   0   2   0   4
   2   0  13 215   0   0   1   0   0   0   0
   3   0   9   0 214   0   0   0   1   0   3
   4   1   3   0   0 226   0   0   1   0   0
   5   0   0   0  35   0 108   0   0  15  44
   6   0   0   0   0   3   0 203   1   0   0
   7   0  22   1   0   0   0   0 198   4   8
   8  21  10   0   0   0   0   0  12 167   1
   9   1   1   0  10   2   0   0   1   0 208
```