

Магадлал Статистик

© 2019 – 2023 Г.Махгал

📅 2023/2/27

Анхаарамж

Энэ нь зөвхөн хичээлийн лекцийн слайд дээрх материалаас тогтох өөрөөр хэлбэл сургалтын бие даасан хэрэглэгдэхүүн үл болох тул Г.Махгалын заах тус хичээлийг судалж буй оюутнуудаас бусад хүмүүс ашиглахад тохиромжгүйг анхаарна уу.

Агуулга

I Санамсаргүй хувьсагч, түүний тархалт	1
1 Магадлалын огторгуй	1
2 Санамсаргүй хувьсагч	3
3 Санамсаргүй хувьсагчийн нягт	4
4 Жигд тархалт	6
II Дундаж болон дундаж квадрат хазайлт	6
1 Пуассоны тархалт	6
2 Хамааралгүйн чанар	8
3 Дундаж	8
4 Геометр тархалт	13
5 Дундаж квадрат хазайлт	13
III Тархалтын функц	15
1 Тархалтын функц	16
2 Дунджийн чанар	18

3 Илтгэгч тархалт	19
4 Хэвийн тархалт	20
5 Хэвийн тархалтын хэрэглээ	22
 IV Наслалтын тархалт	 24
1 Оршил	24
2 Нөхцөлт магадлал ба нөхцөлт тархалт	24
3 Илтгэгч тархалт	28
4 Саатлын эрчим	29
5 Найдварын функц	31
6 Вейбуллын тархалт	34
 V Бернуллийн процесс	 36
1 Урвасан бином тархалт	36
2 Бернуллийн процесс	38
 VI Пуассоны процесс	 42
1 Гамма тархалт	42
2 Пуассоны процесс	44
 VII Санамсаргүй хувьсагчийн хувиргалт	 47
1 Санамсаргүй хувьсагчийн хувиргалт	48
2 Урвуу хувиргалтын арга	53
 VIII Хамтын тархалт ба санамсаргүй хувьсагчдын ха- маарал	 55
1 Санамсаргүй вектор	55
2 Бүтэн магадлалын томъёо	59
3 Байесын зарчим	61

4 Нөхцөлт үл хамаарал	62
5 Нөхцөлт дундаж	64
IX Олон хэмжээст хэвийн тархалт ба шугаман загвар	66
1 Векторын дундаж ба ковариацийн матриц	66
2 Олон хэмжээст хэвийн тархалт	69
3 Регрессийн шугаман загвар	74
X Хамааралгүй санамсаргүй хувьсагчдын нийлбэрийн тархалт	77
1 Хамааралгүй хувьсагчдын нийлбэрийн тархалт	77
2 Хязгаарын гол теорем	80
XI Хамааралтай хувьсагчдын дараалал, Марковын хэлхээ	83
1 Хамааралтай хувьсагчдын дараалал	83
2 Марковын хэлхээ	83
3 Стационар тархалт	88
4 Төлвийн ангилал	90
XII Тархалтын параметрийн статистик үнэлэлт	92
1 Тархалтын загварын тухай таамаглал дэвшүүлэх	93
2 Тархалтын параметрийн үнэлэлт	94
XIII Хамгийн их үнэний хувь бүхий үнэлэлт, Байесын үнэлэлт	100
1 Кошийн тархалт	100
2 Хамгийн их үнэний хувь бүхий үнэлэлт	102
3 Байесын үнэлэлт	104

XIV Статистик таамаглал шалгах, тархалтын загварын тохирцыг тогтоох 110

- | | |
|--|-----|
| 1 Таамаглал шалгах | 110 |
| 2 Тархалтын хэлбэрийн тухай таамаглал шалгах | 114 |

XV Үнэний хувийн харьцаат шинжүүр, регрессийн шугаман загвар 117

- | | |
|----------------------------------|-----|
| 1 Үнэний хувийн харьцаат шинжүүр | 118 |
| 2 Регрессийн шугаман загвар | 120 |

XVI Хяналттай машин сургалтад ашиглах зарим статистик арга техник 125

- | | |
|-----------------------------|-----|
| 1 Регрессийн шугаман загвар | 125 |
| 2 Авторегрессийн загвар | 127 |
| 3 Ложистик регресс | 128 |
| 4 Гэнэн Байесын алгоритм | 130 |

Лекц I

Санамсаргүй хувьсагч, түүний тархалт

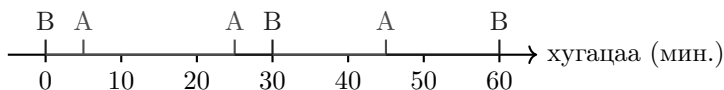
1 Магадлалын огторгуйн математик загвар

Хялбар жишээ

Жишээ 1. Сүлжээгээр дамжин серверт ирэх хүсэлт хугацааны аль ч эгшинд ижил боломжтой. Хүсэлт хүлээн авах портыг 17:00 ба 18:00 цагийн хооронд нээх ба хүсэлтийг боловсруулж хариу өгөх зорилготой А болон В гэсэн хоёр янзын програм ажилладаг.

- А програм 17:05, 17:25, 17:45
- В програм 17:00, 17:30, 18:00

цагт ажиллахаар тохируулжээ. Системийн админ нийт хүсэлтийн $\frac{2}{3}$ хувийг А програм боловсруулж байгааг ажиглажээ. Энэ зөв үү?



Зураг 1: Жишээ бодлогын зураглал

Санамсаргүй үзэгдэл

Тодорхойлолт 1. 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 хугацааг минутаар хэмжинэ гэвэл $\Omega = [0, 60]$ болно.

Тодорхойлолт 2. Эгэл үзэгдлийн огторгуйн дэд олонлогийг *үзэгдэл* гэнэ.

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

Үзэгдлийн магадлал

Тодорхойлолт 3. \mathcal{F} нь үзэгдлүүдийн олонлог байг. Тэгвэл

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

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 $E = (a_1, b_1) \cup \dots \cup (a_k, b_k) \subseteq \Omega$ 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 магадлалын аксиом ёсоор $P(\Omega) = 1$ 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.

Магадлалын огторгуй ба магадлалын үндсэн чанар

Тодорхойлолт 4. Ω нь туршилтаар илэрч болох бүх эгэл үзэгдлүүдийн огторгуй, \mathcal{F} нь байж болох бүх үзэгдлийн цуглуулга буюу Ω эгэл үзэгдлийн огторгуйн бүх дэд олонлогуудын олонлог, P нь тэдгээр үзэгдлийн магадлалыг хэмжих $P : \mathcal{F} \rightarrow [0, 1]$ байдалтай функц байг. Тэгвэл (Ω, \mathcal{F}, P) бүлийг *магадлалын огторгуй* гэнэ.

P функц буюу $P : \mathcal{F} \rightarrow [0, 1]$ магадлалын хэмжээс нь дараах гурван аксиомыг хангана.

1. Дурын E үзэгдлийн хувьд $P(E) \geq 0$ байна.
2. $P(\Omega) = 1$
3. E_1, E_2, \dots харилцан нийцгүй (нэг зэрэг явагдах боломжгүй) үзэгдлүүд бол

$$P(E_1 + E_2 + \dots) = P(E_1) + P(E_2) + \dots$$

(тоологдом аддитив) адилтгал биелнэ.

2 Санамсаргүй хувьсагч**Санамсаргүй хувьсагч**

Тодорхойлолт 5. Эгэл үзэгдлийн огторгуй дээр тодорхойлсон, бодит утгатай $X : \Omega \rightarrow \mathbb{R}$ функцийг *санамсаргүй хувьсагч* гэнэ.

Жишээний хувьд X нь хүсэлтийг боловсруулах хүртэлх хугацаа бол

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

байдлаар тус функцийг бичиж болно.

- 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. $I_A \in \{0, 1\}$

3 Санамсаргүй хувьсагчийн нягт

Дискрет санамсаргүй хувьсагчийн нягт

Тодорхойлолт 6. X санамсаргүй хувьсагч дискрет буюу

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

үед

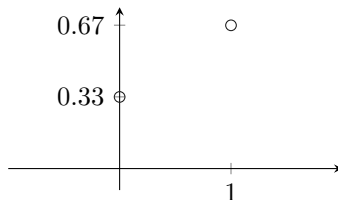
$$P(X = x_i)$$

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. Дурын X *дискрет санамсаргүй хувьсагчийн нягтын функц* дараах хэл-



Зураг 2: I_A дискрет санамсаргүй хувьсагчийн нягт

бэртэй 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

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

буюу үзэгдлийн магадлал нягтын нийлбэртэй тэнцүү байна.

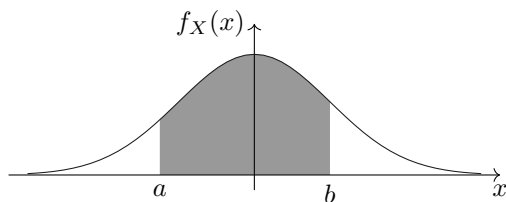
Тодорхойлолт 7. X нь тасралтгүй санамсаргүй хувьсагч бөгөөд хэрэв $a \leq b$ бүрийн хувьд

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

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

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

Иймд



Зураг 3: X хувьсагчийн нягтын муруй ба $a \leq X \leq b$ үзэгдлийн магадлал

1. дурын $x \in \mathbb{R}$ бүрийн хувьд $f_X(x) \geq 0$

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

байх нь гарцаагүй юм.

Дискрет ба тасралтгүй хувьсагчдын нягтын ялгаа

Санамж 1. Ерөнхийдөө тасралтгүй санамсаргүй хувьсагчийн хувьд

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

юм.

Үнэндээ

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

бас

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

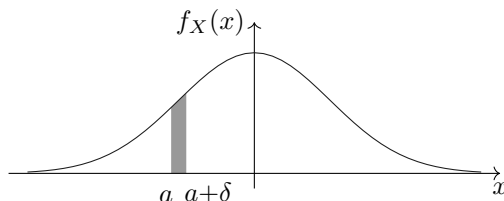
юм.

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.

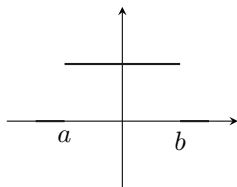


Зураг 4: X санамсаргүй хувьсагчийн нягтын муруй ба $a < X < a + \delta$ үзэгдлийн магадлал

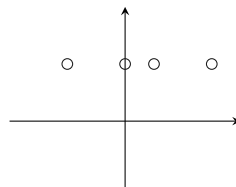
4 Жигд тархалт

Жигд тархалт

Тодорхойлолт 8. Нягт нь тогтмол байх тархалтыг *жигд тархалт* гэнэ.



(a) Тасралгүй жигд тархалт



(b) Дискрет жигд тархалт

Зураг 5: 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

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

Пуассоны тархалтын тусламжтай шийдэх бодлого

Жишээ 2. 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

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

байна.

Пуассоны тархалтын нягтын гаргалгаа

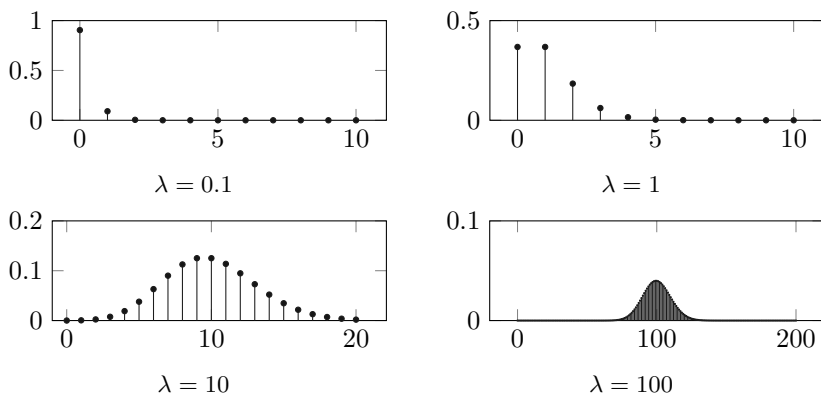
$$\lambda_n = np, \quad \lim_{n \rightarrow \infty} \lambda_n = \lambda \text{ гэе.}$$

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

Одоо хязгаарт шилжвэл $\lim_{n \rightarrow \infty} f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}$ 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 томъёо* гэдэг.



Зураг 6: Пуассоны тархалт, параметрийн янз бүрийн утгад

Пуассоны тархалт, параметрийн янз бүрийн утгад

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

2 Хамааралгүйн чанар

Хамааралгүйн чанар

Тодорхойлолт 9.

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

бол A болон B үзэгдлүүдийг *хамааралгүй* гэнэ.

Тодорхойлолт 10.

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

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.

3 Дундаж

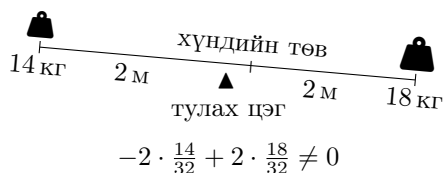
Санамсаргүй хувьсагчийн математик дундаж

Тодорхойлолт 11. Санамсаргүй хувьсагчийн утгуудыг тэдгээрийн магадла-лаар жинлэсэн дунджийг *математик дундаж*² гэнэ.

Математик дундаж нь санамсаргүй хувьсагчийн тархалтын төвийг заадаг. Энэхүү төв нь утгаа сөрөг бус жин бүхий "цэгүүдийн" хүндийн төв юм.

¹Poisson distribution

²expectation

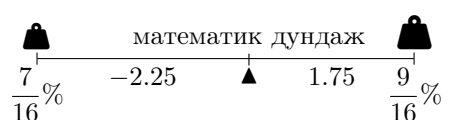
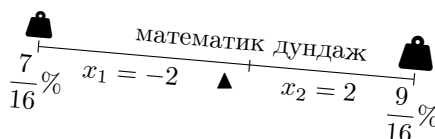


Энд хөшүүргийн хүчний мөрийг тулах цэгээс хэмжсэн.

Дискрет санамсаргүй хувьсагчийн математик дундаж

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

Тулах цэгийг санамсаргүй сонгоно гэвэл хөшүүргийн хүчний мөр дээрх жингийн тархалт нь санамсаргүй хувьсагчийн утгуудын магадлалын тархалттай адил төстэй.



Энд ▲ нь координатын эхийг төлөөлнө.

Жишээ 3. Пуассоны тархалттай санамсаргүй хувьсагчийн математик дундажийг ол.

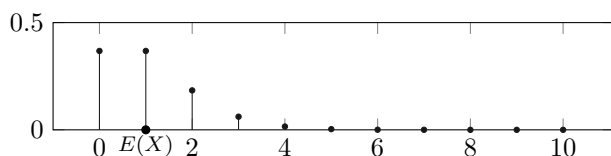
$$\begin{aligned} E(X) &= \sum_{x_i} x_i P(X = x_i) = \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} e^{-\lambda} \\ &= \lambda e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{x!} = \lambda e^{-\lambda} e^{\lambda} = \lambda \end{aligned}$$

Тасралтгүй санамсаргүй хувьсагчийн математик дундаж

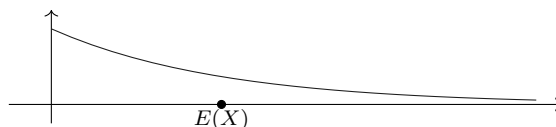
Тасралтгүй санамсаргүй хувьсагчийн математик дундажийг дараах томъёогоор олно.

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

Тасралтгүй санамсаргүй хувьсагчийн утгуудыг түүний нягтаар жинлэсэн нийлбэр олоход ердийн нийлбэр бус интеграл нийлбэр ашиглана. $(E(X), \infty)$ интервал дахь утгууд тоон утгаараа их боловч нягт нь бага харин $(0, E(X))$ интервал дахь утгууд тоон утгаараа бага боловч нягт нь их байна.



Зураг 7: Пуассоны тархалттай санамсаргүй хувьсагчийн математик дундаж, $\lambda = 1$ үед



Зураг 8: Тасралтгүй хувьсагчийн математик дундаж, илтгэгч тархалтын тохиолдолд

Жишээ 4. $\lambda > 0$ параметр бүхий илтгэгч тархалт буюу

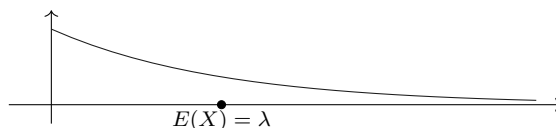
$$f_X(x) = \lambda e^{-\lambda x}, \quad x > 0, \quad \lambda > 0$$

нягттай X санамсаргүй хувьсагчийн математик дунджийг ол.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_0^{\infty} x \lambda e^{-\lambda x} dx \\ &= - \int_0^{\infty} x de^{-\lambda x} = - [x e^{-\lambda x}]_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx \\ &= - \frac{1}{\lambda} \int_0^{\infty} de^{-\lambda x} = \frac{1}{\lambda} \end{aligned}$$

Математик дундаж ба хазайлтын жинлэсэн нийлбэр

Санамсаргүй хувьсагчийн утгаас тус хувьсагчийн математик дундаж хүртэлх зайг түүний *хазайлт* гэдэг. $x > E(X)$ утгуудын дунджаас тооцсон зай



Зураг 9: $\lambda = 1$ параметр бүхий илтгэгч тархалтын нягтын муруй ба математик дундаж

буюу $x - E(X)$ хазайлтын жинлэсэн интеграл нийлбэр нь $x < E(X)$ утгуудын

хувьд дунджаас тооцсон зай буюу $E(X)$ – x хазайлтын жинлэсэн интеграл нийлбэртэй тэнцүү өөрөөр хэлбэл

$$\int_{x>E(X)} [x - E(X)] f_X(x) dx = \int_{x<E(X)} [E(X) - x] f_X(x) dx$$

байдаг. Үүнтэй төстэй чанар дискрет санамсаргүй хувьсагчийн хувьд ч хүчинтэй.

Дунджийн чанар

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

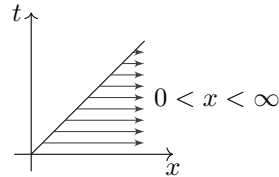
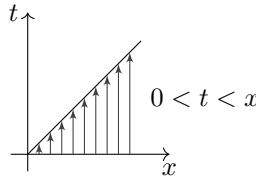
$$E(X) = \int_0^{\infty} P(X \geq t) dt$$

Жишээ 5. Илтгэгч тархалттай санамсаргүй хувьсагчийн хувьд дээрх чанар биелж буйг харуул.

$$\int_0^{\infty} P(X \geq t) dt = \int_0^{\infty} \int_t^{\infty} \lambda e^{-\lambda x} dx dt = - \int_0^{\infty} \int_t^{\infty} de^{-\lambda x} dt = \int_0^{\infty} e^{-\lambda t} dt = \frac{1}{\lambda} = E(X)$$

Баталгаа Интегралчлах эрэмбэ дараа солих замаар шууд батална.

$$\begin{aligned} E(X) &= \int_0^{\infty} x f_X(x) dx = \int_0^{\infty} \left\{ \int_0^x dt \right\} f_X(x) dx \\ &= \int_0^{\infty} \left\{ \int_0^x f_X(x) dt \right\} dx \end{aligned}$$



$$= \int_0^{\infty} \left\{ \int_t^{\infty} f_X(x) dx \right\} dt = \int_0^{\infty} P(X \geq t) dt$$

□

Чанар 2 (Марковын тэнцэл биш). X нь сөрөг бус утгатай санамсаргүй хувьсагч байг. Тэгвэл

$$P(X \geq x_0) \leq \frac{E(X)}{x_0}$$

тэнцэл биш биелнэ.

Баталгаа

$$\begin{aligned} E(X) &= \int_0^{\infty} x f_X(x) dx = \int_0^{x_0} x f_X(x) dx + \int_{x_0}^{\infty} x f_X(x) dx \\ &\geq \int_{x_0}^{\infty} x f_X(x) dx \geq x_0 \int_{x_0}^{\infty} f_X(x) dx = x_0 P(X \geq x_0) \end{aligned}$$

□

Чанар 3 (Ухамсаргүй статистикчийн хууль³). $g : \mathbb{R} \rightarrow \mathbb{R}$ функцийн хувьд

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

байна.

Үүний баталгааг *тархалтын функц* үзсэний дараа хийнэ.

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

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

Жишээ 6. Пуассоны тархалттай санамсаргүй хувьсагчийн момент үүсгэгч функцийг ол.

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

Жишээ 7. Стандарт хэвийн тархалттай санамсаргүй хувьсагчийн момент үүсгэгч функцийг ол.

³The law of the unconscious statistician

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

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

гэж олдоно.

Момент үүсгэгч функцийн хэрэглээ

Момент үүсгэгч функцийг математик дундаж зэрэг момент⁴ олоход ашиглана.

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

Жишээ 8. Пуассоны тархалттай санамсаргүй хувьсагчийн математик дунджийг түүний үүсгэгч функцийн тусламжтай ол.

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

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

болно.

4 Геометр тархалт

Геометр тархалт

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

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

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

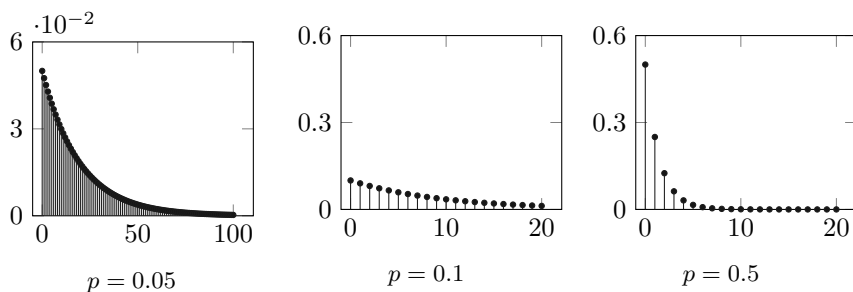
Санамж 2. Үүнээс гадна санамсаргүй хувьсагчийг амжилт илэртэл явуулах туршилтын тоо буюу $\{1, 2, \dots\}$ утгатайгаар авах явдал бий. Өөрөөр хэлбэл геометр тархалтыг хоёр янзаар авч үздэг.

Геометр тархалтын нягтын гаргалгаа

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

⁴ $\alpha_k = E[X^k]$

Геометр тархалт, параметрийн янз бүрийн утгад



Зураг 10: Геометр тархалт, параметрийн янз бүрийн утгад

5 Дундаж квадрат хазайлт

Дундаж квадрат хазайлт буюу дисперс

X санамсаргүй хувьсагчийн утгуудын дунджаасаа хазайх $X - E(X)$ хазайлт дунджаар ямар байхыг хэрхэн олох вэ?

Халз дунджилбал

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

буюу үргэлж тэг гарна. Хазайлтыг абсолют утгаар авч $E|X - E(X)|$ байдлаар дунджилбал санамсаргүй утгатай ялгаврын тэмдгийг яаж тооцох вэ гэсэн хүндрэл учирна. Иймд үүнийг хазайлтын квадратыг дундажлах байдлаар шийдвэрлэдэг.

Тодорхойлолт 12. X санамсаргүй хувьсагчийн утгуудын дунджаасаа хазайх $X - E(X)$ хазайлтын квадратын математик дунджийг *дундаж квадрат хазайлт* буюу дисперс гэдэг.

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

Дундаж квадрат хазайлтын зарим чанар

$$1. D(X) = E(X^2) - [E(X)]^2$$

Энэхүү чанарыг

$$\begin{aligned} D(X) &= E[X - E(X)]^2 \\ &= E[X^2 - 2XE(X) + [E(X)]^2] \\ &= E(X^2) - 2[E(X)]^2 + [E(X)]^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

байдлаар баталж болно.

$$2. D(a + bX) = b^2 D(X)$$

$$3. X \text{ болон } Y \text{ хамааралгүй үед } D(X + Y) = D(X) + D(Y)$$

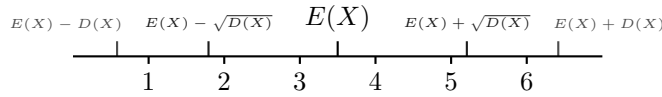
Стандарт хазайлт

Дундаж квадрат хазайлт дахь хазайлтыг квадрат зэрэгт дэвшүүлсэнтэй холбогдох гажилтыг түүнээс квадрат язгуур авах байдлаар засдаг.

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

Жишээ 9. $f_X(x) = \frac{1}{6}$, $x \in \{1, 2, \dots, 6\}$ дискрет жигд тархалтын стандарт хазайлтыг ол.

$E(X) = \sum_{x=1}^6 x \frac{1}{6} = 3.5$ ба $D(X) = \sum_{x=1}^6 (x - 3.5)^2 \frac{1}{6} \approx 2.917$ тул $\sqrt{D(X)} \approx 1.708$ болно.



Зураг 11: Жишээ бодлогын бодолтын үр дүн

Чебышевийн тэнцэл биш

$P(X \geq x_0) \leq \frac{E(X)}{x_0}$ буюу Марковын тэнцэл бишээс практик ач холбогдолгүй үр дүн гарах явдал бий. Тухайлбал $x_0 \leq E(X)$ утга сонговол $P(X \geq x_0) \leq 1$ гэсэн илэрхий үнэлгээ гарна.

Жишээ 10. Хүний дундаж наслалтыг 76.5^5 гэе. Тэгвэл ямар нэг хүний дор хаяж 100 наслах магадлал хамгийн ихдээ

$$P(X \geq 100) \leq \frac{76.5}{100} = 0.765$$

байна.

Энэ дутагдал тус тэнцэл биш зөвхөн математик дундаж гэсэн ганцхан зүйлд тулгуурласантай холбоотой. Иймд математик дундажтай хамт дундаж квадрат хазайлт ашигласан үнэлэлт авч үздэг.

Чанар 4 (Чебышевийн тэнцэл биш).

$$P(|X - E(X)| \geq x_0) \leq \frac{D(X)}{x_0^2}$$

Баталгаа

$$\begin{aligned} P(|X - E(X)| \geq x_0) &= P([X - E(X)]^2 \geq x_0^2) \quad \text{Марковын тэнцэл биш} \\ &\leq \frac{E[X - E(X)]^2}{x_0^2} = \frac{D(X)}{x_0^2} \end{aligned}$$

□

⁵2021 оны судалгаагаар тогтоосон монгол эмэгтэйчүүдийн дундаж наслалт

Өмнөх жишээг Чебышевийн тэнцэл биштэй хамт ахин авч үзье. Энэ тохиолдолд санамсаргүй хувьсагчийн дундаж квадрат хазайлт буюу стандарт хазайлтыг нэмж шаардана.

Жишээ 11. Наслалтын стандарт хазайлт 13.5 жил байг.

$$\begin{aligned} P(X \geq 100) &= P(X \geq 76.5 + 23.5) \leq P(X \geq 76.5 + 23.5) + P(X \leq 76.5 - 23.5) \\ &= P(|X - 76.5| \geq 23.5) \leq \frac{13.5^2}{23.5^2} = 0.33 \end{aligned}$$

Лекц III

Тархалтын функц

1 Тархалтын функц

Санамсаргүй хувьсагчийн тархалт, түүнийг илэрхийлэх функцүүд



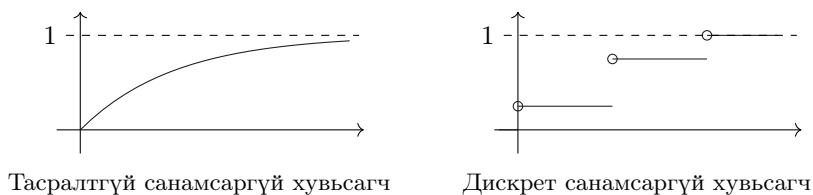
Зураг 12: Тасралтгүй санамсаргүй хувьсагчийн тархалт, илтгэгч тархалтын тохиолдолд



Зураг 13: Дискрет санамсаргүй хувьсагчийн тархалт, бином тархалтын тохиолдолд

Тархалтын функций чанарууд

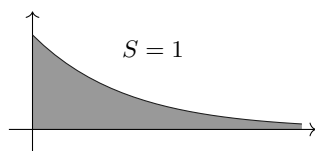
1. Зүүн тасралтгүй
2. Үл буурах
3. $\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1$



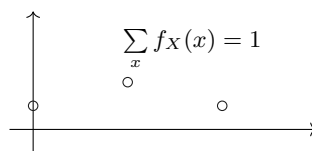
Зураг 14: Тархалтын функц

Нягтын функцийн чанарууд

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



тасралтгүй үед



дискрет үед

Зураг 15: Нягтын нийлбэр

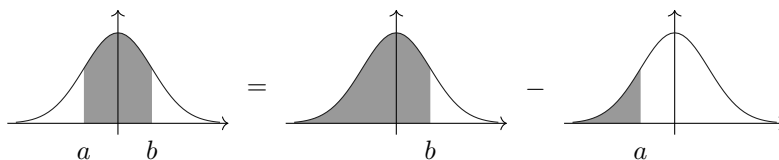
Тасралтгүй санамсаргүй хувьсагчийн тархалтын функц болон нягтын функцийн холбоо, үзэгдлийн магадлал

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

Туршилтын тархалтын функц

$F_X(x)$ тархалттай X санамсаргүй хувьсагчийн эх олонлогоос авсан n хэмжээтэй X_1, \dots, X_n энгийн санамсаргүй түүврээр цуглуулсан өгөгдлийн тархалтыг илэрхийлдэг

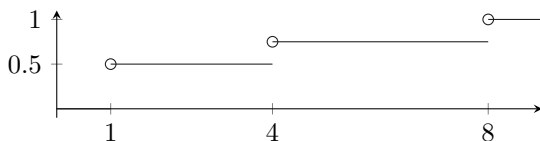
$$F_n(x) = \frac{\mu_n(x)}{n}, \quad x \in \mathbb{R}, \quad \mu_n(x) = |\{i : X_i < x, i = 1, \dots, n\}|$$

функцийг *туришилтын тархалтын функц* гэдэг.

Туршилтын тархалтын функц олох

Жишээ 12. 1, 4, 4, 8 өгөгдөлд харгалзах туршилтын тархалтын функцийг ол.

$$F_4(x) = \begin{cases} 0, & x \leq 1 \\ 2/4, & 1 < x \leq 4 \\ 3/4, & 4 < x \leq 8 \\ 1, & 8 < x \end{cases}$$



Зураг 17: Туршилтын тархалтын функцийн график

Туршилтын тархалтын функцийн чанар

Чанар 5. 1. Дурын $x \in \mathbb{R}$ болон $\epsilon > 0$ бүрийн хувьд

$$\lim_{n \rightarrow \infty} P(|F_n(x) - F_X(x)| < \epsilon) = 1$$

чанар биелнэ.

2. $F_n(x)$ нь тархалтын функцийн чанаруудыг хангана.

2 Дунджийн чанар

Ухамсаргүй статистикчийн хуулийн баталгаа

Сэргээн санах нь 1.

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

$g(\cdot)$ 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 :

$$dy = dg(x) = dg(g^{-1}(y)) = g'(g^{-1}(y))d(g^{-1}(y))$$

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

- $f_Y(y)$:

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

$$f_Y(y) = F'_Y(y) = f_X(g^{-1}(y)) \frac{d}{dy}(g^{-1}(y)) = \boxed{f_X(x) \frac{1}{g'(g^{-1}(y))}}$$

-

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

3 Илтгэгч тархалт

Илтгэгч тархалт

Сэргээн санах нь 2. 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.

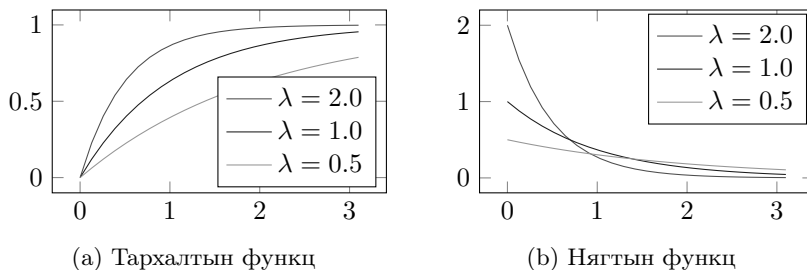
Илтгэгч тархалтын функцийн гаргалгаа

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

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

Илтгэгч тархалт, параметрийн янз бүрийн утгад

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



Зураг 18: Илтгэгч тархалт, параметрийн янз бүрийн утгад

Илтгэгч тархалтын зарим чанар

1. Геометр тархалтын тасралтгүйн аналог
2. $X_i \sim \text{Exp}(\lambda_i)^6$ ($i = 1, \dots, n$) ба *хамтдаа хамааралгүй*⁷ бол

$$\min\{X_1, \dots, X_n\} \sim \text{Exp}(\lambda_1 + \dots + \lambda_n)$$

3. $\forall x, y \geq 0$ бүрийн хувьд

$$P(X \geq x + y | X \geq y) = P(X \geq x)^8$$

4 Хэвийн тархалт

Пуассоны тархалт параметрийн их утгад

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

"Хонх" хэлбэртэй нягтын илэрхийлэл

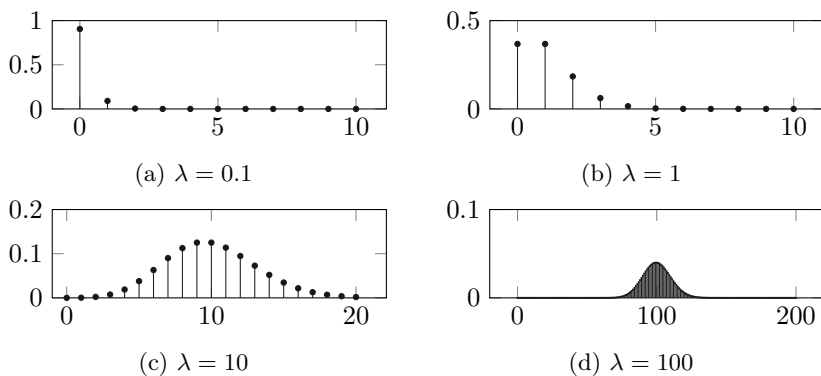
$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \in \{0, 1, 2, \dots\}$$

$E(X) = \lambda$ 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 $\ln[(1 + \delta)^{\lambda(1 + \delta) + 1/2}] =$

⁶илтгэгч тархалт

⁷энэ талаар хожим үзнэ

⁸илтгэгч тархалтын санамжгүй байдал; энэ талаар хожим үзнэ



Зураг 19: Пуассоны тархалтын параметр ба нягтын хэлбэр

$[\lambda(1 + \delta) + 1/2] \ln(1 + \delta) = (\lambda + 1/2 + \lambda\delta)(\delta - \delta^2/2 + O(\delta^3)) \approx \lambda\delta + \lambda\delta^2/2 + O(\delta^3)$
ойролцоо адилтгалыг ашиглаад эцэст нь $\delta = (x - \lambda)/\lambda$ орлуулга хийвэл

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

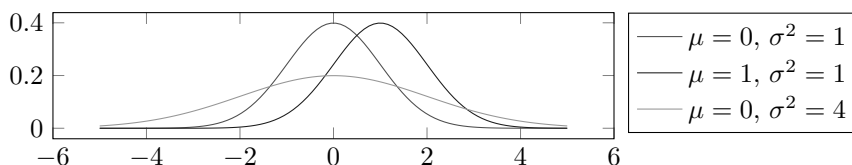
болно.

”Хонх” хэлбэртэй буюу хэвийн тархалт

$f_X(x) = \frac{e^{-(x-\lambda)^2/(2\lambda)}}{\sqrt{2\pi\lambda}}$ 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. Хэвийн



Зураг 20: Хэвийн тархалтын нягтын муруй параметрийн янз бүрийн утгад

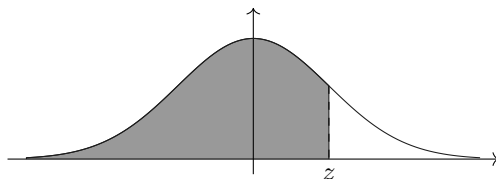
тархалтыг бас Гауссын тархалт ч гэдэг. X хувьсагч хэвийн тархалттай гэхийг $X \sim N(\mu, \sigma^2)$ гэж тэмдэглэнэ.

Стандарт хэвийн тархалтын функц

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

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

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



Зураг 21: $\Phi(z)$ утга буюу z -ээс бага утгуудын хувьд хэвийн тархалтын нягтын муруйн дор байх мужийн талбай

Хэвийн тархалттай санамсаргүй хувьсагчийн шугаман хувиргалт

$X \sim N(\mu, \sigma^2)$ ба

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

бол

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

байдаг бөгөөд гаргалгааг нь хожим үзнэ. Иймд хэрэв $X \sim N(\mu, \sigma^2)$ бол

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

болно.

5 Хэвийн тархалтын хэрэглээ

Хэвийн тархалтын хэрэглээ

Статистик загваруудад хэвийн тархалт гол сонголт нь байдаг. Бас хамааралгүй, олон санамсаргүй хувьсагчдын нийлбэрийн тархалт хэвийн тархалтад ойр байдаг¹⁰. Жишээлбэл $X_1, \dots, X_n \sim \text{Ber}(p)$ ¹¹ ба хамааралгүй үед $X = X_1 + \dots + X_n \sim B(n, p)$ ¹² байдаг тул

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

буюу

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

байна.

⁹үүнийг стандарт хувиргалт гэнэ

¹⁰Хязгаарын гол теорем гэдэг нэрээр дараа үзнэ.

¹¹Бернуллийн тархалт

¹²бином тархалт

Туршилтын тоо их үед бином тархалтын нягтыг ойролцоо бодох

Сэргээн санах нь 3. 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 томъёо* гэдэг.

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

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

n их боловч p нь бага ч биш, их ч биш үед бином тархалттай X санамсаргүй хувьсагчийн $\frac{X-np}{\sqrt{np(1-p)}} \sim N(0, 1)$, $n \rightarrow \infty$ чанарт үндэслэн $P(X = x)$ магадлалыг хэвийн тархалтын нягтаар шууд ойролцоо бодно. Үүнийг *Муавр-Лапласын локал томъёо* гэдэг. Харин $P(a \leq X \leq b) = \sum_{x=a}^b f_X(x)$ магадлалын хувьд n их үед $\frac{X-np}{\sqrt{np(1-p)}} \sim N(0, 1)$, $n \rightarrow \infty$ ёсоор

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

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

Жишээ 13. 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

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

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

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

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

Муавр-Лапласын интеграл томьёоны тасралтгүйн засвар

Бүхэл утга авдаг, бином тархалттай санамсаргүй хувьсагчтай холбогдох үзэгдлийн магадлал олохдоо 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

Наслалтын тархалт

1 Оршил буюу нэр томьёоны тайлбар

Наслалт

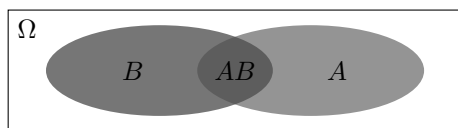
Наслалт гэдэгт ямарваа зүйлийн үргэлжлэх хугацаа эсвэл нөөц чадавхыг хамруулж ойлгоно.

- Биологи, Хүн ам зүй Амьдрах хугацаа
- Инженер техникийн ухаан Ашиглалтын хугацаа ба материалын бат бөхийн нөөц
- Цөмийн физик Цацраг идэвхт бөөмийн задрах хүртэлх хугацаа
- Хүлээлгийн онол буюу үйлчилгээний систем Дараагийн үйлчлүүлэгч ирэх хүртэлх хугацаа
- Эдийн засаг Дампуурах хүртэлх хугацаа
- ...

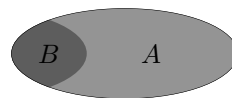
2 Нөхцөлт магадлал ба нөхцөлт тархалт

Нөхцөлт магадлал

B үзэгдлийн магадлалыг A үзэгдлээс хамааруулж олохыг *нөхцөлт магадлал* гээд $P(B|A)$ хэлбэрээр бичнэ. $P(B|A)$ нөхцөлт магадлалыг A үзэгдэл явагдсан үед B үзэгдэл явагдах магадлал ойлгодог. Иймд уг магадлалыг олох-



(a) $P(AB)$ магадлал



(b) $P(B|A)$ нөхцөлт магадлал

Зураг 22: A үзэгдлээс хамаарсан B үзэгдлийн нөхцөлт магадлал

доо B үзэгдлийг зөвхөн A үзэгдлийн хүрээнд авч үзэх юм уу

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

томъёо ашигладаг.

Нөхцөлт магадлалтай холбогдох зарим санамж

Санамж 3. 1. $P(B|A)$ бол нөхцөлт магадлал харин $P(B \setminus A)$ бол $P(B - A)$ буюу үзэгдлүүдийн ялгавар юм.

$$2. \underbrace{P(C + D)}_{\text{үзэгдэл}} | \underbrace{A + B}_{\text{нөхцөл}} = \frac{P((A+B)(C+D))}{P(A+B)}$$

3. $P(B|A)$ ба $P(A|B)$ хоёр нөхцөлт магадлал ерөнхийдөө ялгаатай юм.

Нөхцөлт магадлал ба хамааралгүй үзэгдлүүд

A болон B үзэгдлүүд хамааралгүй үед $P(AB) = P(A)P(B)$ байдаг тул энэ тохиолдолд

$$P(A|B) = P(A) \quad \text{бас} \quad P(B|A) = P(B)$$

байна. Иймд хэрэв

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

бол A үзэгдэл B үзэгдлээс хамааралгүй харин

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

бол B үзэгдэл A үзэгдлээс хамааралгүй байна.

Жишээ 14. Нэг бүсгүйд n эр гэрлэх санал тавьж болзоонд урьжээ. Бүсгүй тэдгээр эрчүүдтэй ээлж дараалан болзоно. Болзоо дээр тухайн эртэй гэрлэх эсэхээ шийдэж хариу өгөх ёстой. Гэрлэнэ гэсэн шийдвэр гаргасан бол удаах эрчүүдтэй болзохгүй. Бас гэрлэх саналыг нь няцаасан эртэйгээ эргэж болзохгүй. Тэгвэл бүсгүй тэдгээр эрчүүдээс хамгийн сайныг нь хэрхэн яаж сонгох вэ?

Эхлээд хамгийн сайн сонголтын магадлалыг авч үзье.

$$\begin{aligned}
 P(\text{хамгийн сайныг нь сонгох}) &= \\
 &= \sum_{i=1}^n P(\{i \text{ дүгээр эрийг сонгох}\} \{i \text{ дүгээр эр хамгийн сайн нь байх}\}) \\
 &= \sum_{i=1}^n P(i \text{ эрийг сонгох} | i \text{ эр хамгийн сайн нь}) P(i \text{ эр хамгийн сайн нь})
 \end{aligned}$$

Үргэлжлүүлэн k дугаар эрийн гэрлэх саналыг хүлээн авлаа гэхэд энэ нь хамгийн сайн сонголт байх магадлалыг олж өгөхөд. Нийтдээ n эр бий тул $k = 1, \dots, n$ утга авна.

$$\begin{aligned}
 P_k(n) &= \left[\sum_{i=1}^{k-1} P(i \text{ эрийг сонгох} | i \text{ эр хамгийн сайн нь}) \right. \\
 &\quad \left. + \sum_{i=k}^n P(i \text{ эрийг сонгох} | i \text{ эр хамгийн сайн нь}) \right] P(i \text{ эр хамгийн сайн нь}) \\
 &= \left[\sum_{i=1}^{k-1} 0 + \sum_{i=k}^n P \left(\begin{array}{l} \text{эхний } i-1 \text{ эрсийн хамгийн сайн нь} \\ \text{эхний } k-1 \text{ эрсийн дотор байх} \end{array} \middle| i \text{ эр хамгийн сайн нь} \right) \right] \frac{1}{n} \\
 &= \left[\sum_{i=k}^n \frac{k-1}{i-1} \right] \frac{1}{n} = \frac{k-1}{n} \sum_{i=k}^n \frac{1}{i-1} \quad (k > 1)
 \end{aligned}$$

$k = 1$ үед эхний эрийг шууд сонгох тул $P_1(n) = \frac{1}{n}$ байна.

$n = 1, \dots, 10$ үед хамгийн оновчтой k буюу хэд дэх эр дээр хамгийн сайн сонголт таарахыг R програмын тусламжтай бодож гаргав.

```

sapply(X = {X <- 1:10; names(X) <- 1:10; X}, FUN = function (n) {
  P_k <- sapply(X = 1:n, FUN = function (k, n) {
    if (k == 1)
      return(1/n)
    {k-1}/n*sum(1/{k-1}:{n-1})
  }, n = n)
  list(k = which.max(P_k), P = round(x = max(P_k), digits = 3))
}) -> X
print(X)

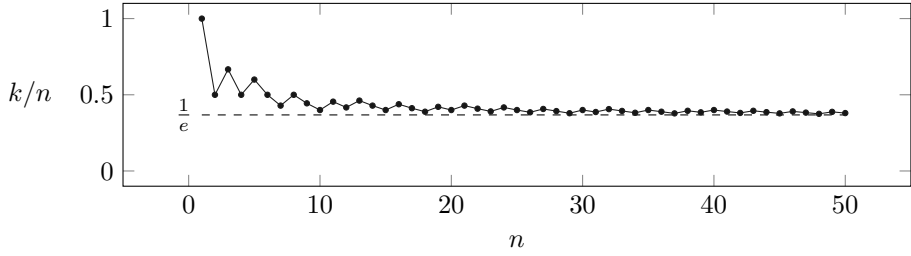
```

	1	2	3	4	5	6	7	8	9	10
k	1	1	2	2	3	3	3	4	4	4
P	1	0.5	0.5	0.458	0.433	0.428	0.414	0.41	0.406	0.399

Энэ нь жишээлбэл $n = 5$ үед $k = 3$ дахь эрийг сонговол оновчтой бөгөөд түүний хамгийн сайн нь байх магадлал $P_3(5) \approx 0.433$ юм.

n эрсийн хувьд k хэд байвал оновчтой вэ гэдгийг n параметрийн их утгад авч үзье.

Чанар 6. Оновчтой сонголтын бодлогын хувьд $\lim_{n \rightarrow \infty} \frac{k}{n} = \frac{1}{e}$ чанар хүчинтэй.



Зураг 23: Оновчтой сонголтын бодлогын k аргумент ба n параметрийн харьцаа

Баталгаа $n \rightarrow \infty$ үед $P_k(n) = \frac{k-1}{n} \sum_{i=k}^n \frac{1}{i-1}$ магадлал дотор $x = \lim_{n \rightarrow \infty} \frac{k-1}{n}$ бас $t = \frac{i-1}{n}$, $dt = \frac{1}{n}$ орлуулга хийе. Мөн n хүрэлцээтэй их тул уг магадлал дахь нийлбэр нь интеграл нийлбэр болно.

$$P(x) = x \int_x^1 \frac{1}{t} dt = -x \ln x$$

Магадлал их байх нь чухал тул үүнийг x аргументаар максимумчилна. x аргументын хувьд Лагранжийн нөхцөл $P'(x) = -\ln x - 1 = 0$ болно. Эндээс $x_{\max} = 1/e$, $P_{\max} = P(1/e) = 1/e$ хариу гарна. Эцэст нь $\lim_{n \rightarrow \infty} \frac{1}{n} = 0$ тул дээрх орлуулга ёсоор $\lim_{n \rightarrow \infty} \frac{k}{n} = \frac{1}{e}$ болно. \square

Ийнхүү n их үед $k \approx \frac{n}{e}$ дугаархыг сонговол энэ нь $\frac{1}{e}$ магадлалтайгаар хамгийн сайн нь байх ажээ.

Жишээ 15.

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

буюу илтгэгч тархалттай X санамсаргүй хувьсагчийн $A = \{a \leq X < b\}$ үзэгдлээс хамаарсан $F_{X|a \leq X < b}(x)$ нөхцөлт тархалтын функц болон $f_{X|a \leq X < b}(x)$ нөхцөлт нягтын функцийг ол. Энд $0 < a < b$ гэж тооцно.

Үзэгдлээс хамаарсан нөхцөлт тархалтын функцийг тодорхойлолтоос $F_{X|a \leq X < b}(x)$ нөхцөлт тархалтын функцийг шууд олох томьёо гарна.

$$F_{X|a \leq X < b}(x) = \frac{P(X < x, a \leq X < b)}{P(a \leq X < b)} = \begin{cases} 0, & x \leq a \\ \frac{F_X(x) - F_X(a)}{F_X(b) - F_X(a)}, & a < x \leq b \\ 1, & x > b \end{cases}$$

Жишээ болгон авсан бодлогод $0 < a < b$ гэснийг анхаарвал дээрх томьёо ёсоор илтгэгч тархалтын хувьд $A = \{a \leq X < b\}$ үзэгдлээс хамаарсан $F_{X|a \leq X < b}(x)$ нөхцөлт тархалтын функц

$$F_{X|a \leq X < b}(x) = \begin{cases} 0, & x \leq a \\ \frac{e^{-\lambda a} - e^{-\lambda x}}{e^{-\lambda a} - e^{-\lambda b}}, & a < x \leq b \\ 1, & x > b \end{cases}$$

хэлбэртэй болно.

Нөхцөлт нягтын функц ба нөхцөлт тархалтын функцийг холбосон дараах ерөнхий томъёог эдгээрийн тодорхойлолт болон уялдаа холбоог нь илэрхийлдэг томъёонд үндэслэн гаргаж болно.

$$\begin{aligned} f_{X|A}(x) &= \lim_{\Delta x \rightarrow 0} \frac{P(X < x + \Delta x | A) - P(X < x | A)}{\Delta x} \\ &= \lim_{\Delta x \rightarrow 0} \frac{P(X < x + \Delta x, A) - P(X < x, A)}{\Delta x P(A)} \\ &= \lim_{\Delta x \rightarrow 0} \frac{F_{X|A}(x + \Delta x) - F_{X|A}(x)}{\Delta x} \\ &= F'_{X|A}(x) \end{aligned}$$

$f_{X|A}(x) = F'_{X|A}(x)$ ерөнхий томъёо ба $A = \{a \leq X < b\}$ үзэгдлээс хамаарсан нөхцөлт тархалтын функцийг томъёонуудаас

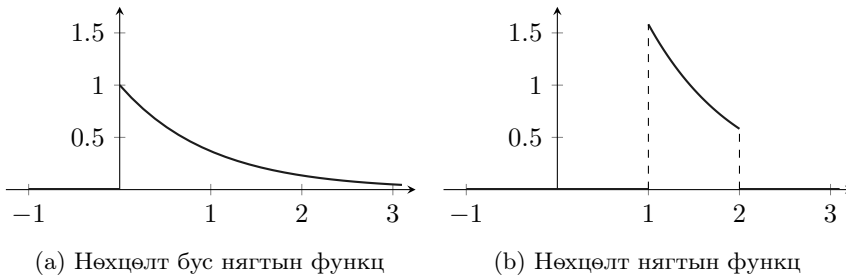
$$f_{X|a \leq X < b}(x) = \begin{cases} \frac{f_X(x)}{P(a \leq X < b)}, & a < x \leq b \\ 0, & \text{бусад} \end{cases}$$

томъёо гарна. Жишээ болгон авсан бодлогод $0 < a < b$ гэснийг анхаарвал дээрх томъёо ёсоор илтгэгч тархалтын хувьд $A = \{a \leq X < b\}$ үзэгдлээс хамаарсан $f_{X|a \leq X < b}(x)$ нөхцөлт нягтын функц

$$f_{X|a \leq X < b}(x) = \begin{cases} \frac{\lambda e^{-\lambda x}}{e^{-\lambda a} - e^{-\lambda b}}, & a < x \leq b \\ 0, & \text{бусад} \end{cases}$$

хэлбэртэй болно.

Жишээний $a = 1$, $b = 2$ тохиолдолд харгалзах $A = \{1 \leq X < 2\}$ үзэгдлээс хамаарсан $f_{X|1 \leq X < 2}(x)$ нөхцөлт нягтын функцийг графикийг нөхцөлт бус буюу ердийн нягтын функцийг графиктай харьцуулж харуулав. Энд $1 < x < 2$



Зураг 24: Илтгэгч тархалтаар жишээлсэн нөхцөлт нягтын функц

утгуудын нөхцөлт нягт өссөн нь $\int_{-\infty}^{\infty} f_X(x) dx = 1$ чанартай холбоотой.

3 Илтгэгч тархалтын санамжгүй байдал

Илтгэгч тархалт

Сэргээн санах нь 4. 1. 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.

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

Энэ сэдэвт ямар нэг зүйлийн үргэлжлэх хугацааг илэрхийлэх санамсаргүй хувьсагчид хамаатай билээ. Үргэлжлэх хугацааны төгсгөлд тодорхой нэг үзэгдэл явагдах бөгөөд үүнийг "амжилт" хэмээн тооцож болно. Тэгэхээр ийм хувьсагчдыг "амжилт" илэртэл хүлээх хугацаа гэж томъёолж болох буюу насалалтыг судлахад илтгэгч тархалтыг хэрэглэж болох юм.

Илтгэгч тархалтын санамжгүй байдал

Чанар 7. $\forall x, y \geq 0$ бүрийн хувьд $P(X \geq x + y | X \geq y) = P(X \geq x)$ байна.

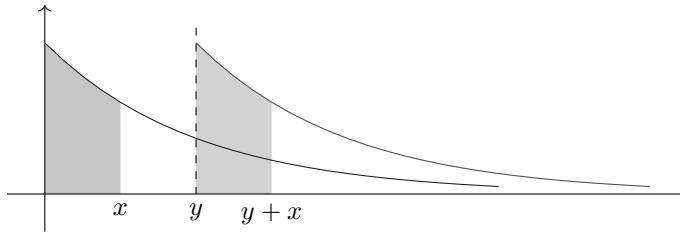
Баталгаа

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

□

Илтгэгч тархалтын санамжгүй байдал

Наслалтын зүгээс $P(X \geq y + x | X \geq y) = P(X \geq x)$ чанарыг тайлбарлавал хэчнээн насалсан нь цааш хэд наслахад нөлөөгүй гэсэн утгатай юм. Энэхүү



Зураг 25: Илтгэгч тархалтын санамжгүй байдал

санамжгүй байдал нь илтгэгч тархалтыг эгэл бөөмсөөс бусад бараг бүхий л зүйлсийн насалтад тохирохооргүй болгоно.

4 Саатлын эрчим

Саатлын эрчим буюу мөхлийн эрчим

Тодорхойлолт 13. Наслалтын 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 *саатлын эрчим*¹³ буюу *мөхлийн эрчим*¹⁴ гэнэ.

Мөхлийн эрчим буюу саатлын эрчим нь хугацааны x эгшин хүртэл насалсан бол яг x эгшин дээрээ шууд үхэх, мөхөх, саатах эрчмийг илэрхийлнэ.

Саатлын эрчим олох томьёо

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 саатлын эрчим

$X \sim \text{Exp}(\lambda)$ 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 саатлын эрчим нь тогтмол байна. Иймд λ параметрийг *эрчмийн параметр*¹⁵ гэдэг. Энэхүү тогтмол саатлын эрчим нь илтгэгч тархалтын санамжгүй байдлын шалтгаан юм. Тогтмол саатлын эрчимтэй, эерэг өөр санамсаргүй хувьсагч байх уу? Үүний хариуг дараагийн слайд дээр авч үзнэ.

¹³failure rate

¹⁴hazard rate

¹⁵rate parameter

Тогтмол саатлын эрчимтэй тархалт

X нь эерэг утгатай, тогтмол саатлын эрчимтэй, тасралтгүй хувьсагч байг.

$$\begin{aligned}
 h_X(x) &= \frac{f_X(x)}{1 - F_X(x)} = k \\
 \int_0^x \frac{f_X(t)}{1 - F_X(t)} dt &= \int_0^x k dt \\
 - \int_0^x \frac{1}{1 - F_X(t)} d(1 - F_X(t)) &= k \int_0^x dt \\
 - \ln(1 - F_X(t)) \Big|_0^x &= kt \Big|_0^x \\
 - \ln(1 - F_X(x)) &= kx \\
 F_X(x) &= 1 - e^{-kx}
 \end{aligned}$$

Энэ нь $X \sim \text{Exp}(k)$ буюу ийм хувьсагч зөвхөн илтгэгч тархалттай байна гэсэн үг юм.

Тогтмол мөхлийн эрчим – санамжгүй байдал – хамааралгүй туршилт

Сэргээн санах нь 5. 1. 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.

2. Амжилт илэртэл хүлээх хугацаа гэсэн санамсаргүй хувьсагчийн магадлалын тархалтыг *илтгэгч тархалт* гэнэ.
3. Амжилт илэртэл хүлээх нь туршилтыг зогсолтгүй явуулах буюу $n \rightarrow \infty$ гэхэд хүргэнэ. Энэ тохиолдолд амжилт нь тоологдохуйц байхын тулд $p \rightarrow 0$ буюу ховор тохиолддог үзэгдэл байна.

Наслалт буюу ямарваа зүйлийн үргэлжлэх хугацаа дуусгавар болж "мөхөх"-ийг "амжилт" хэмээн үзэж болох юм. Мөхлийн эрчим тогтмол байхын уг үндэс нь амжилт гэх үзэгдэл хугацааны аль ч эгшинд ижил боломжтой тохиолдох буюу нэг л туршилтыг хамааралгүйгээр давтан явуулна гэсэн явдал юм. Харин туршилтууд хамааралгүй гэсэн нөхцлийг авч хаявал амжилт буюу мөхлийн магадлал туршилт бүрийн хувьд адил тэнцүү байхаа болих тул мөхлийн эрчим тогтмол бус болно.

5 Найдварын функц ба Kaplan-Meier-ын муруй

Найдварын функц буюу амьдрах чадварын функц¹⁶

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

¹⁶Reliability function буюу Survival function

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

Карлан-Меер-ын муруй

Өгөгдлөөр байгуулах тархалтын функцийг туршилтын тархалтын функц гэдэг билээ. Түүн шиг өгөгдлөөр байгуулах найдварын функцийг графикийг *Карлан-Меер-ын муруй* гэдэг. Найдварын функц нь тархалтын функцтэй

$$R_X(x) = 1 - F_X(x)$$

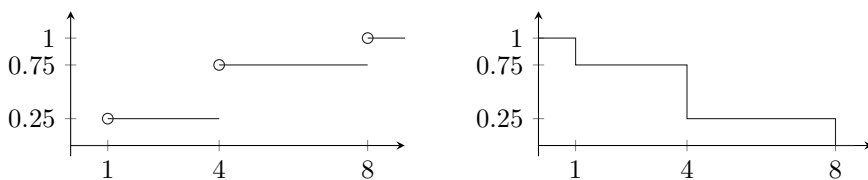
гэж уялддаг тул Карлан-Меер-ын муруйд харгалзах функцийг

$$R_n(x) = 1 - F_n(x)$$

гэж олно.

Жишээ 16. Наслалтын мэдээлэл бүхий 1, 4, 4, 8 өгөгдлөөр Карлан-Меер-ын муруй байгуул.

$$F_4(x) = \begin{cases} 0, & x \leq 1 \\ 1/4, & 1 < x \leq 4 \\ 3/4, & 4 < x \leq 8 \\ 1, & 8 < x \end{cases} \quad R_4(x) = 1 - F_4(x) = \begin{cases} 1, & x \leq 1 \\ 3/4, & 1 < x \leq 4 \\ 1/4, & 4 < x \leq 8 \\ 0, & 8 < x \end{cases}$$



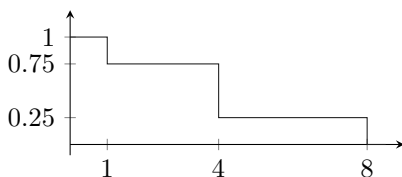
Зураг 26: Туршилтын тархалтын функцийг ба Карлан-Меер-ын муруй

Карлан-Меер-ын муруй байгуулах

1, 4, 4, 8 өгөгдөл 4 элементтэй тул $n = 4$ болно. Бусад тооцоог хүснэгтэлж үзүүлэв.

x	$\mu_4(x)$	$\sum_{z \in [x, \infty)} \mu_4(z)$	$R_4(x)$
0	0	4	$1.00 = 4/4$
1	1	$3 = 4 - 1$	$0.75 = 3/4$
4	2	$1 = 3 - 2$	$0.25 = 1/4$
8	1	$0 = 2 - 1$	$0.00 = 0/4$

Хүснэгт 1: Kaplan-Meier-ын муруй байгуулахад шаардлагатай найдварын функцийн утга олох нь



Зураг 27: Kaplan-Meier-ын муруй

Цензуртай өгөгдөл ба Kaplan-Meier-ын муруй

Судалгаанд хамруулж буй объектуудын зарим нь судалгаа дуусахад саа-таагүй буюу хэвийн байх эсвэл судалгааны дундуур сураггүй алга болох явдал тохиолддог. Ийм мэдээлэлтэй өгөгдлийг *цензуртай өгөгдөл* гэдэг. Цензуртай өгөгдлөөр Kaplan-Meier-ын муруй байгуулахдаа цензурт багтаагүй мэдээллийг түүнд харгалзах хугацаанаас цааш авч ашигладаггүй. Өөрөөр хэлбэл цензур-тай өгөгдлөөр зохиох найдварын функцийг үлдэж буй хугацааны доторх бүрэн тодорхой мэдээлэлд л үндэслэж зохиодог.

Цензуртай өгөгдөл

№	Хугацаа	Цензуртай	Тайлбар
1	1	тийм	сураггүй болсон
2	3	үгүй	
3	4	тийм	
4	4	тийм	сураггүй болсон
5	4	үгүй	
6	5	үгүй	
7	8	тийм	судалгаа дууссан
8	10	үгүй	

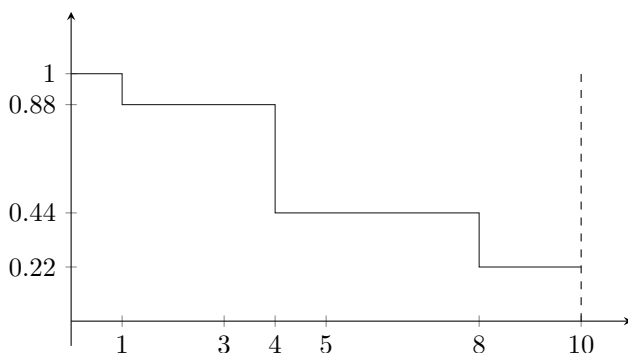
Хүснэгт 2: Цензуртай өгөгдөл

Цензуртай өгөгдлөөр Kaplan-Meier-ын муруй байгуулах

Энд $\mu_n(x)$ болон $\nu_n(x)$ нь хугацааны x эгшинд харгалзах үзэгдлийн дав-тамжийг цензуртай эсэхээр ялгаж олсоныг илэрхийлнэ.

x	$\mu_8(x)$	$\nu_8(x)$	$\sum_{z \in [x, \infty)} (\mu_8(z) + \nu_8(z))$	$R_8(x)$
0	0	0	8	$8/8 = 1$
1	1	0	8	$(8 - 1)/8 \cdot 1 = \mathbf{0.875}$
3	0	1	7	$(\mathbf{7} - \mathbf{0})/\mathbf{7} \cdot \mathbf{0.875} = 0.875$
4	2	1	4	$(4 - 2)/4 \cdot 0.875 = 0.4375$
5	0	1	3	$(3 - 0)/3 \cdot 0.4375 = 0.4375$
8	1	0	2	$(2 - 1)/2 \cdot 0.4375 = 0.21875$
10	0	1	1	$(1 - 0)/1 \cdot 0.21875 = 0.21875$

Хүснэгт 3: Цензуртай өгөгдлөөр Kaplan-Meier-ын муруй байгуулахад шаардагдах тооцоо



Зураг 28: Цензуртай өгөгдлөөр байгуулсан Kaplan-Meier-ын муруй

6 Вейбуллын тархалт

Вейбуллын тархалт

Наслалт буюу ямар нэг зүйлийн үргэлжлэх хугацааг судлахад мөхлийн эрчим нь тогтмол бус өөрөөр хэлбэл цаг хугацааны явцад өөрчлөгддөг тархалт зайлшгүй шаардлагатай. Ийм тархалтуудын нэг бол Вейбуллын тархалт юм.

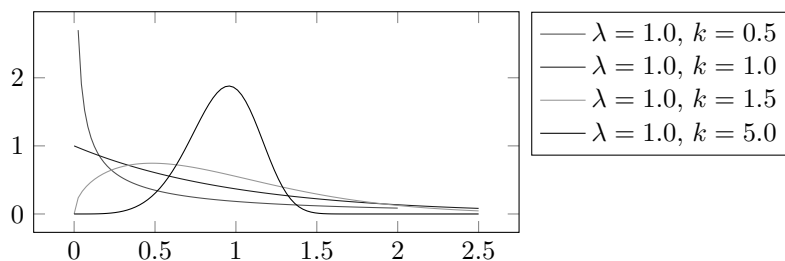
$$f_X(x) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

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

Вейбуллын тархалт, параметрийн янз бүрийн утгад

Тархалтын функц

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$



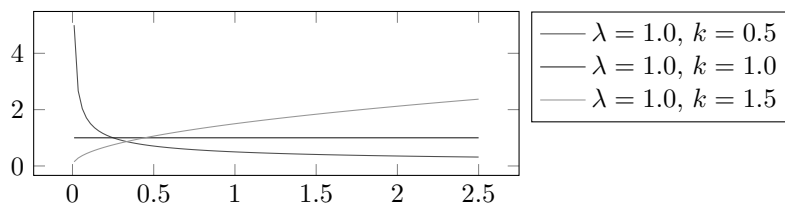
Зураг 29: Вейбуллын тархалтын нягтын функц параметрийн янз бүрийн утгад

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

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

Вейбуллын тархалтын саатлын эрчим буюу 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 30: Вейбуллын тархалтын саатлын 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 саатна.
- $k < 1$ үед $h_X(x)$ буурах буюу шинэ нь хуучнаасаа илүү саатна.

Вейбуллын тархалт болон илтгэгч тархалтын холбооWeib(λ, k) буюу

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

нь $k = 1$ үед

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

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

Лекц V**Бернуллийн процесс****1 Урвасан бином тархалт****Урвасан бином тархалтаар шийдэх бодлого**

Жишээ 17. Нэг бадарчин айл хэсч гуйлга гуйхаар хүрээний нэг гудам уруу оров. Гудамд 30 айл байдаг бөгөөд өнөөх бадарчин маань 5 айлаас юм авахаас нааш 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 r удаагийн амжилтаас өмнөх бүтэлгүйтлийн тооны магадлалын тархалтыг *урвасан бином тархалт*¹⁷ гээд $NB(r, p)$ гэж тэмдэглэнэ. Энд p нь амжилтын магадлал юм.

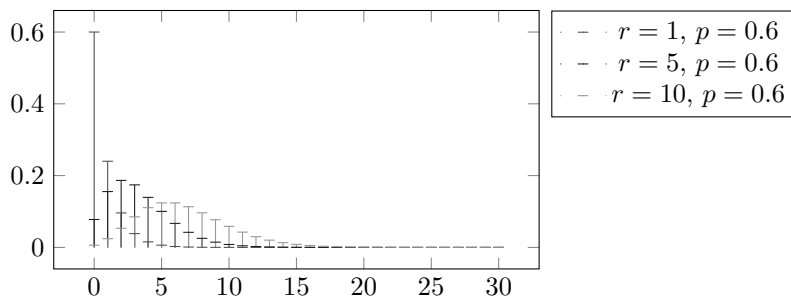
Амжилтыг A , бүтэлгүйтлийг \bar{A} гэж тэмдэглээд улмаар туршилтууд хамааралгүй гэдгийг анхаарч эцэст нь биномын коэффициентийн чанар ашиглавал

¹⁷Negative binomial distribution

r ширхэг амжилтаас өмнө x удаа бүтэлгүйтэх магадлал буюу урвасан бином тархалтын нягтын илэрхийлэл дараах байдлаар олдоно.

$$\begin{aligned}
 P(X = x) &= P(\underbrace{\overbrace{\bar{A} \cdot \bar{A} \cdot A \cdot \bar{A} \cdot \dots \cdot \bar{A}}^{x+r-1 \text{ ширхэг туршилт}} \cdot \overbrace{A}^{+1}}_{A \text{ нь } r \text{ ширхэг, } \bar{A} \text{ нь } x \text{ ширхэг}} + \dots + \underbrace{\overbrace{\bar{A} \cdot \bar{A} \cdot \bar{A} \cdot A \cdot \dots \cdot \bar{A}}^{x+r-1 \text{ ширхэг туршилт}} \cdot \overbrace{A}^{+1}}_{A \text{ нь } r \text{ ширхэг, } \bar{A} \text{ нь } x \text{ ширхэг}}) \\
 &= C_{x+r-1}^{r-1} (1-p)^x p^r = C_{x+r-1}^x (1-p)^x p^r, \quad x \in \{0, 1, 2, \dots\}
 \end{aligned}$$

Урвасан бином тархалтын нягтын хэлбэр



Зураг 31: Урвасан бином тархалт, параметрийн янз бүрийн утгад

Бодлогыг Монте-Карло симуляцын аргаар бодсон нь

```

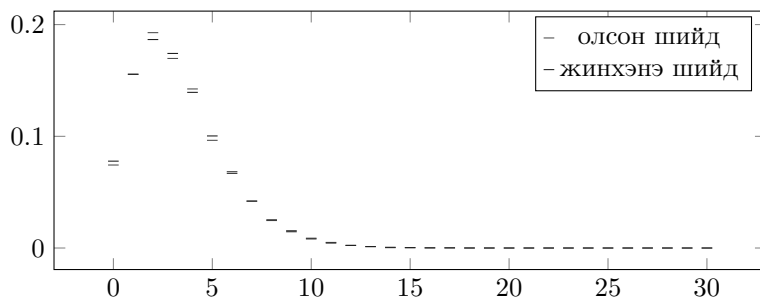
set.seed(0)
N <- vector("mode" = "integer", "length" = 30 + 1)
for (n in 1:10000) {
  attempt <- 0; success <- 0; failure <- 0
  while (success < 5 && attempt < 30) {
    attempt <- attempt + 1
    if (runif(n = 1) < 0.6)
      success <- success + 1
    else
      failure <- failure + 1
  }
  N[failure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

```

Монте-Карло симуляцын аргаар олсон шийд

Жинхэнэ шийдийг R програм дээр дараах тушаалаар олж болно.

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

Зураг 32: x ширхэг айл юу ч хялайлгалгүй явуулах магадлал

Урвасан бином тархалтын зарим чанар

- $E(X) = \frac{(1-p)r}{p}$ Жишээний хувьд $E(X) = \frac{0.4 \cdot 5}{0.6} \approx 3.333$ байна.
- $D(X) = \frac{(1-p)r}{p^2}$ Жишээний хувьд $D(X) = \frac{0.4 \cdot 5}{0.6^2} \approx 5.555$ байна.
- Түүврийн дисперс нь түүврийн дунджаасаа их үед Пуассоны тархалт¹⁸-ын оронд ашигладаг.
- $NB\left(r, \frac{\lambda}{r+\lambda}\right) \xrightarrow{r \rightarrow \infty} \text{Pois}(\lambda)$
- $NB(r=1, p) = \text{Geom}(p)$

2 Бернуллийн процесс

Бернуллийн процесс

Сэргээн санах нь 6. Ямар нэг туршилт авч үзье. Туршилтын үр дүнгээс аль нэг A үзэгдлийг онцгойлон "амжилт" хэмээн авна. Ийнхүү үр дүнг нь хоёр ангилсан туршилтыг *Бернуллийн туршилт* гэдэг. $P(A) = p$ бас туршилтыг өөр хоорондоо хамааралгүй байдлаар n удаа давтсан гээ. Тэгвэл үүнтэй холбогдуулан практикт өргөн тохиолдох янз бүрийн санамсаргүй хувьсагч зохиож болно.

Тодорхойлолт 14. $X_i \sim \text{Ber}(p)$ буюу $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$ бас X_1, X_2, \dots хамааралгүй байг. Тэгвэл X_1, X_2, \dots санамсаргүй хувьсагчдын төгсгөлөг болон төгсгөлгүй дарааллыг *Бернуллийн процесс* гэнэ.

Бернуллийн процесстэй холбогдох хялбар бодлого

Жишээ 18 (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

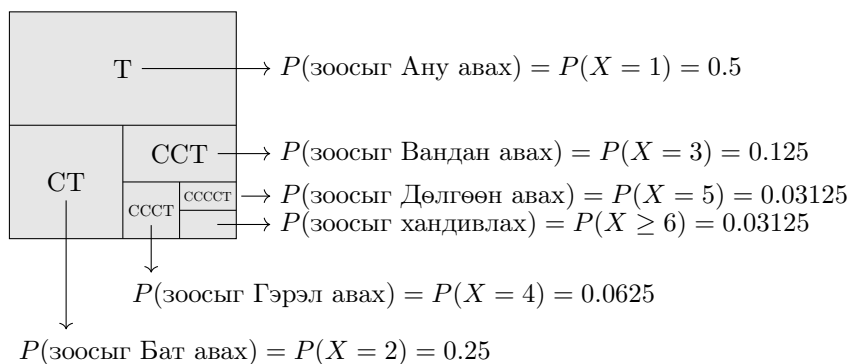
¹⁸ дундаж ба дисперс нь тэнцүү байдаг

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

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

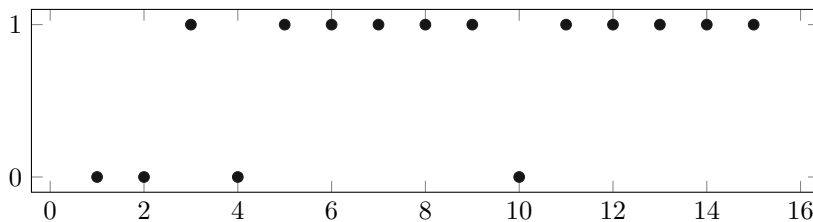
геометр тархалт гарах бөгөөд $p = 0.5$ байна.



Зураг 33: Бернуллийн процесстэй холбогдох бодлогын зураглал

Бернуллийн процессын санамжгүй байдал

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



Зураг 34: $p = 0.6$ үед симуляцалсан Бернуллийн процессын нэг тохиолдол

Бернуллийн процесстэй 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. $\text{Geom}(p)$ буюу геометр тархалт.

3. Нэг амжилт илрэх хүртэлх туршилтын тоо. $\text{Geom}(p)$ 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

$$\underbrace{0000}_\text{Geom}(p) \ 1 \ \underbrace{0000}_\text{Geom}(p) \ 1 \ \underbrace{0000000000}_\text{Geom}(p) \ 1 \ 0$$

$$\underbrace{\hspace{10cm}}_{NB(r=3,p)}$$

Нэг ижил параметр бүхий геометр тархалттай, хамааралгүй хувьсагчдын нийлбэр урвасан бином тархалттай.

Бином тархалт ба Урвасан бином тархалт Бүтэлгүйтэл буюу 0-үүдийн тоог s гээ.

$$\begin{array}{l} \overbrace{0000100001000000100001} \\ B : n = \text{нийт туршилтын тоо} = s + r : NB \\ B : r = \text{амжилтын тоо} = r : NB \\ B : \text{амжилтын тоо} = \text{хувьсагч} = \text{бүтэлгүйтлийн тоо} : NB \end{array}$$

$NB(r, p)$ 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.

$+$	0	1	0	0	1	0	1	0	$X_i \sim \text{Ber}(p)$
	0	0	1	0	1	0	0	0	$Y_i \sim \text{Ber}(q)$
	0	1	1	0	1	0	1	0	$Z_i \sim \text{Ber}(p + q - pq)$

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

$Z \sim \text{Ber}(p)$ процессыг X болон Y хоёр процесст задалъя. Үүний тулд Z процессын амжилтуудад харгалзах $S \sim \text{Ber}(q)$ 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 $X \sim \text{Ber}(pq)$, $Y \sim \text{Ber}(p(1-q))$ байх болно.

Жишээ 19.

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

Харин шинээр үүсэх X болон Y процессууд хамааралгүй байж чадахгүй.

Бадарчинтай жишээ ба Бернуллийн процессын k дугаар амжилтад харгалзах туршилтын дугаар

Жишээ 20. Бадарчинтай жишээг эргэн авч үзье. Уул бадарчин хэд дэх айл дээрээ өөрийн зорьсон 5 дахь хишгээ хүртэх магадлалтай вэ?

Үүнийг олох нь утгаа $k = 5$ дугаар амжилтад харгалзах туршилтын дугаар гэсэн санамсаргүй хувьсагчийн хамгийн өндөр магадлалтай утга буюу моод олох явдал юм. Иймд Бернуллийн процессын k дугаар амжилтад харгалзах туршилтын дугаар гэсэн хувьсагчийн магадлалын тархалтыг олъя. Урвасан бином тархалтын хувьд бүтэлгүйтлийн тоог авч үздэг бол энэ тохиолдолд амжилт болон бүтэлгүйтлийн тооны нийлбэр буюу туршилтын тоог хөндөж байна.

Бернуллийн процессын k дугаар амжилтад харгалзах туршилтын дугаар

Y_k = " k дугаар амжилтад харгалзах туршилтын дугаар" санамсаргүй хувьсагчийн хувьд геометр тархалтын $\{1, 2, \dots\}$ утгууд авдаг

X = "амжилт илэртэл явуулах туршилтын тоо"

гэсэн хувьсагчтай

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

хувилбарыг ашиглана.

$$\begin{array}{cccccccccccccccc}
 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
 \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{3.5cm}} & & & & & & & & & & & & & & & & \\
 X_1=5 & X_2=5 & X_3=9 & & & & & & & & & & & & & & & & \\
 \underbrace{\hspace{10cm}} & & & & & & & & & & & & & & & & & & \\
 Y_3=X_1+X_2+X_3=5+5+9=19
 \end{array}$$

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

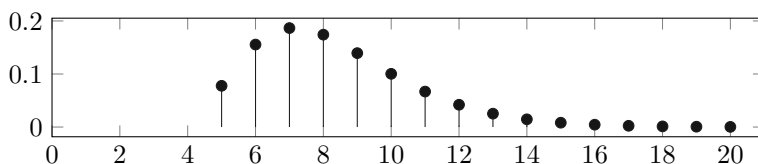
$$E(Y_k) = k/p, \quad D(Y_k) = k(1-p)/p^2$$

байна.

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

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

$p = 0.6$ үед Y_5 хувьсагчийн хамгийн өндөр магадлалтай утга буюу моод нь



Зураг 35: Y_5 хувьсагчийн нягт, $p = 0.6$ үед

$x = 7$ байх тул жишээ бодлогын хариу 7 болно.

Бернуллийн процесс дахь дараалсан амжилтын тооны тархалт

Жишээ 21. Бадарчинтай жишээг эргэн авч үзье. Дараалан хэдэн айл түүнд өглөг өгөх магадлалтай вэ?

Бернуллийн процесс дахь дараалсан амжилтын тоо нь утгаа ”бүтэлгүйтэх хүртэл явуулах туршилтын тоо” байх тул $x \in \{1, 2, \dots\}$ утга бүхий хувьсагчтай $\text{Geom}(1-p)$ тархалтад захирагдана. Геометр тархалтын нягт буурах геометр прогресс байдаг тул 1, 2, 3, 4 болон 5 удаагийн дараалсан амжилтуудаас зөвхөн 1 удаагийн ”дараалсан” амжилт хамгийн өндөр магадлалтай юм.

Лекц VI

Пуассоны процесс

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 $\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx$, $k > 0$ тул нягтын функцийн $\int_{-\infty}^\infty f_X(x)dx = 1$ чанарт нийцүүлэхийн тулд

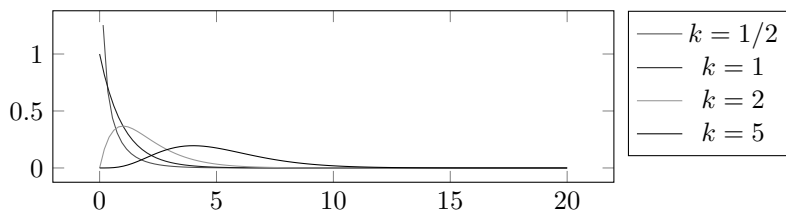
$$f_X(x) = \frac{1}{\Gamma(k)}x^{k-1}e^{-x}$$

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 параметрууд

- Хэлбэрийн параметр



Зураг 36: $f_X(x) = \frac{1}{\Gamma(k)}x^{k-1}e^{-x}$ 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 $x := \lambda x^{19}$ орлуулга хийвэл

$$(\lambda x)^{k-1}e^{-\lambda x}d(\lambda x) = \lambda^k x^{k-1}e^{-\lambda x}dx$$

болох тул

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

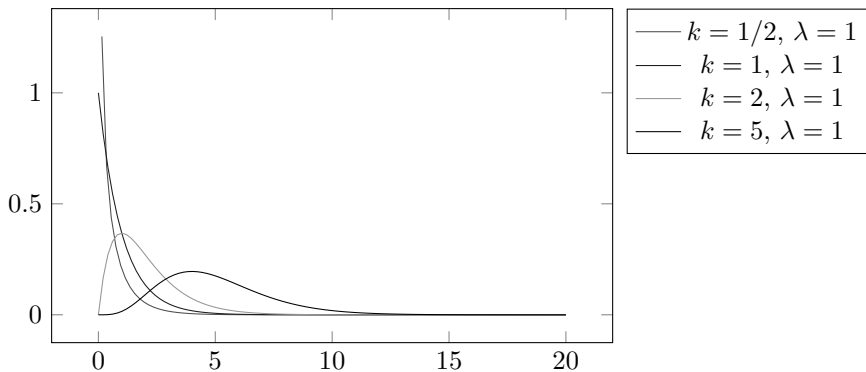
¹⁹ $x := x/\lambda$ орлуулгад харгалзах хувилбарыг масштабын параметр гэдэг.

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 37: 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 $X_1, \dots, X_k \sim \text{Exp}(\lambda)$ хамааралгүй бол

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

байна. Энэ чанарын гаргалгаатай дараагийн хэсэгт танилцана.

- $k \in \mathbb{N}$ бол $\text{Gamma}(\lambda, k)$ нь Эрлангийн тархалттай давхцана.

- Хэрэв $X \sim N(0, 1)$ бол $X^2 \sim \text{Gamma}(1/2, 1/2)$
- Хэрэв X_1, \dots, X_k хувьсагчид хамааралгүй бөгөөд $N(0, 1)$ тархалттай бол

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

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

2 Пуассоны процесс

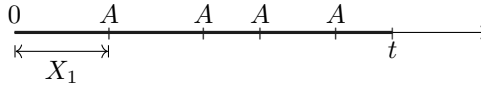
Илтгэгч тархалт ба Пуассоны тархалтын уялдаа холбоо

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

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

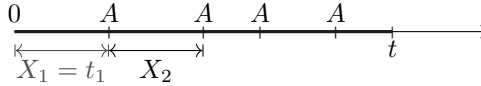
$$P(X \geq 1) = 1 - P(X = 0) = 1 - f_X(0) = 1 - \frac{(\lambda t)^0}{0!} e^{-\lambda t} = 1 - e^{-\lambda t}$$

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 38: Пуассоны процессын эхний амжилт илрэх хүртэлх хугацаа

Удаах амжилт илэртэл хүлээх хугацааны тархалт



Зураг 39: Пуассоны процесс дахь удаах амжилт илрэх хүртэлх хугацаа

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

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

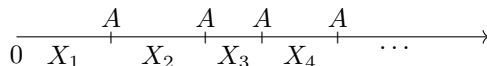
буюу илтгэгч тархалттай "удаах амжилт илэртэл хүлээх хугацаа" $t - t_1$ -ээс бага байх магадлалтай тэнцүү байна. Нөгөө талаас $X_2 \sim \text{Exp}(\lambda)$ гэвэл илтгэгч тархалтын санамжгүй чанараар

$$P(X_2 < t | X_2 \geq t_1) = 1 - P(X_2 \geq t | X_2 \geq t_1)$$

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

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

Пуассоны процесс²⁰



Зураг 40: Пуассоны процессын зураглал

Тодорхойлолт 15. 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 $\text{Exp}(\lambda)$ тархалттай бөгөөд хамааралгүй байг. Тэгвэл X_1, X_2, \dots самсаргүй хувьсагчдын дарааллыг *нэгэн төрлийн Пуассоны процесс* гэнэ.

Ерөнхий тохиолдолд X_1, X_2, \dots хувьсагчдын тархалтын эрчмийн параметрийн утга хугацаанаас хамааран $\lambda(t)$ байдлаар өөрчлөгддөг байж болно. Ийм үед уг процессыг *нэгэн төрлийн бус Пуассоны процесс* гэдэг.

Пуассоны процессын жишээ

Жишээ 22. "12th IAAF World Championships In Athletics: IAAF Statistics Handbook. Berlin 2009" тайлан дахь 1913 оноос 2009 оны хоорондох 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".

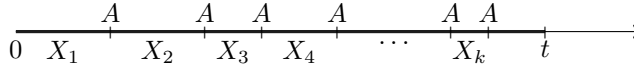
$$\begin{aligned}
P(\text{жилийн дотор шинэ рекорд тогтоох}) &= P(X < 1) = 1 - e^{-0.3598 \cdot 1} \\
&= 0.302
\end{aligned}$$

t хугацаанд амжилт яг k удаа илрэх магадлал ба Пуассоны тархалт

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

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

²⁰Poisson process



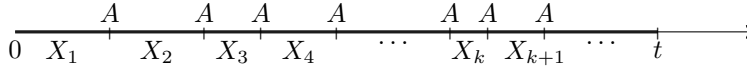
Зураг 41: Пуассоны процесст t хугацаанд амжилт яг k удаа илрэх

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 42: 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 $X_i \sim \text{Exp}(\lambda)$ 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 $\text{Gamma}(\lambda, k)$ тархалттай байна.

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

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

болно.

Лекц VII

Санамсаргүй хувьсагчийн хувиргалт

1 Санамсаргүй хувьсагчийн хувиргалт

Санамсаргүй хувьсагчийн хувиргалт

Тодорхойлолт 16. Санамсаргүй хувьсагчийн утгыг ямар нэг функцээр үйлчлэн өөрчлөхийг *санамсаргүй хувьсагчийн хувиргалт* гэнэ.

Дискрет санамсаргүй хувьсагчийн 1:1 хувиргалт

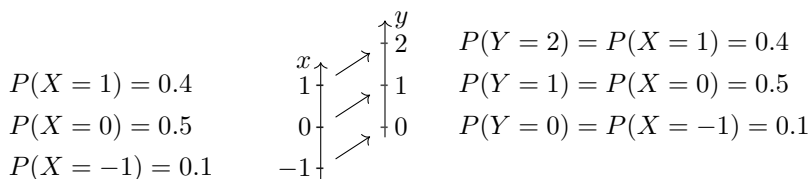
x	-1	0	1
$f_X(x)$	0.1	0.5	0.4

Хүснэгт 4: X санамсаргүй хувьсагчийн тархалтын хүснэгт

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

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

Хүснэгт 5: $Y = X + 1$ санамсаргүй хувьсагчийн тархалтын хүснэгт



Зураг 43: Дискрет санамсаргүй хувьсагчийн 1:1 чанартай хувиргалт

Бодолт дараах байдалтай байсан.

$$\begin{aligned}
 f_Y(2) &= P(Y = 2) = P(X = 1) = f_X(1) = f_X(2 - 1) = f_X(g^{-1}(2)) \\
 f_Y(1) &= P(Y = 1) = P(X = 0) = f_X(0) = f_X(1 - 1) = f_X(g^{-1}(1)) \\
 f_Y(0) &= P(Y = 0) = P(X = -1) = f_X(-1) = f_X(0 - 1) = f_X(g^{-1}(0))
 \end{aligned}$$

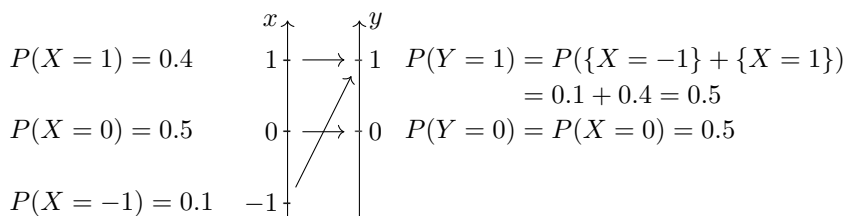
Ийнхүү 1:1 чанартай $g(\cdot)$ функцээр тодорхойлогдох $Y = g(X)$ дискрет санамсаргүй хувьсагчийн нягт олох

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

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



Зураг 44: Дискрет санамсаргүй хувьсагчийн 1:1 нөхцөл үл хангах хувиргалт

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

Хүснэгт 6: $Y = |X|$ санамсаргүй хувьсагчийн тархалтын хүснэгт

Y хувьсагчийн тархалтыг олохын тулд

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

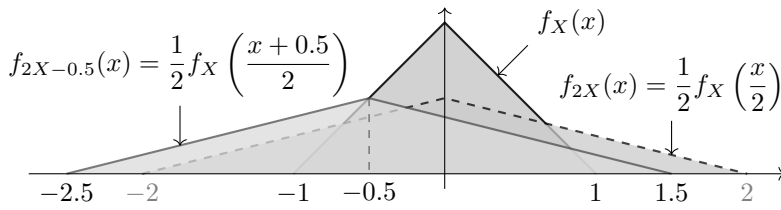
бодолт хийсэн. Ийнхүү 1:1 нөхцөл үл хангах $g(\cdot)$ функцээр тодорхойлогдох $Y = g(X)$ дискрет санамсаргүй хувьсагчийн нягт олох дараах томъёо зохиож болно.

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

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



Зураг 45: $Y = 2X - 0.5$ хувиргалтаар тархалт өөрчлөгдөх байдал

Тархалт 2 дахин "сунасан" тул нягт 2 дахин багасна. Учир нь бүх нягтын нийлбэр 1-тэй тэнцүү. Бас санамсаргүй хувьсагчийн утгуудыг бүгдийг нь -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 24. $X \sim N(\mu, \sigma^2)$ бол $Y = a + bX$ хувьсагчийн тархалтыг ол.

Сэргээн санах нь 7. $X \sim N(\mu, \sigma^2)$ 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

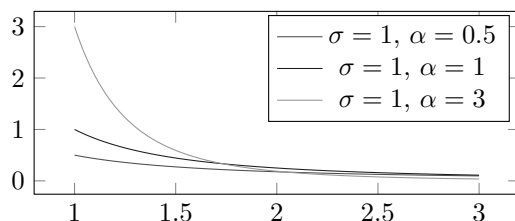
$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

Жигд тархалттай хувьсагчийн $Y = \sigma(1 - X)^{-1/\alpha}$ хувиргалт

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

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

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.

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

$$= \frac{1}{2\sqrt{y}}(f_X(\sqrt{y}) + f_X(-\sqrt{y}))$$

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

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

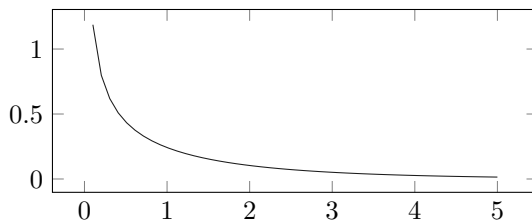
болно.

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

Геометр тархалттай хувьсагчийн $Y = X^2$ хувиргалт

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

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

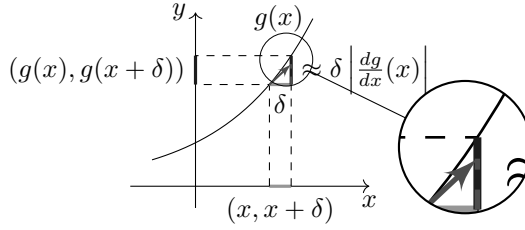
байг. Тэгвэл $Y = X^2$ санамсаргүй хувьсагчийн тархалтыг олж. $Y = g(X) = X^2$ хувиргалтын $g(\cdot)$ функц нь геометр тархалттай санамсаргүй хувьсагчийн авах утгын олонлог $\{0, 1, 2, \dots\}$ дээр 1:1 чанартай буулгалт байна. Иймд

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

томъёогоор 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.



Зураг 47: Хувиргалтаар үүсэх хувьсагчийн нягт олох томьёоны гаргалгаа

$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$ томьёоны өөр гаргалгаа
 $g(\cdot)$ функц эрс монотон байг.

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

$$\parallel \qquad \qquad \qquad \parallel$$

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

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

2 Санамсаргүй хувьсагч симуляцлах урвуу хувиргалтын арга

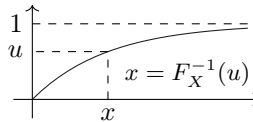
Урвуу хувиргалтын арга

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

$$F_U(u) = P(U < u) = P(F_X(X) < u)$$

$$= P(X < F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u$$

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



Зураг 48: 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

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

томъёо гарна. $1 - U \sim U(0, 1)$ тул үүнийг дараах байдлаар өөрчилж болно.

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

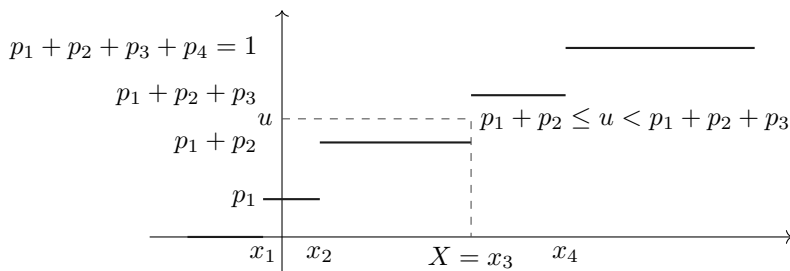
Тус томъёо болон урвуу хувиргалтын аргын дагуу дараах програм бичиж болно. Програмын кодыг R хэлээр бичив.

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

Урвуу хувиргалтын аргаар дискрет тархалт симуляцлах

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

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



Зураг 49: Урвуу хувиргалтын аргаар дискрет санамсаргүй хувьсагч симуляцлах зарчим

Урвуу хувиргалтын аргаар дискрет хувьсагч симуляцлах алгоритм

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

Жишээ 25. Хүснэгтээр өгсөн тархалттай X санамсаргүй хувьсагчийг урвуу хувиргалтын аргаар симуляцалъя.

$$\begin{array}{c|ccc} x & -1 & 0 & 1 \\ f_X(x) & 0.1 & 0.5 & 0.4 \end{array}$$

Хүснэгт 7: Дискрет санамсаргүй хувьсагчийн тархалтын хүснэгт

```
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

1 Санамсаргүй вектор, түүний тархалт

Санамсаргүй вектор, хамтын тархалт

Тодорхойлолт 17. Нэгээс олон санамсаргүй хувьсагчдыг хамтад нь *санамсаргүй вектор*, (X_1, \dots, X_p) санамсаргүй векторын тархалтыг *хамтын тархалт* гэнэ.

Хүйс	Солгой	
	тийм	үгүй
эр	2	3
эм	1	4

(a) Хамтын давтамжийн хүснэгт

X_1	X_2	
	1	0
1	0.2	0.3
0	0.1	0.4

(b) Хамтын тархалтын хүснэгт

Хүснэгт 8: 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 50: Тасралтгүй санамсаргүй векторын хамтын нягт ба санамсаргүй түүвэр

Хамтын тархалт ба үзэгдлийн магадлал

Дискрет санамсаргүй вектор ямар нэг D мужид унах магадлалыг дараах томъёогоор олно.

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

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

Тасралтгүй санамсаргүй вектор ямар нэг D мужид унах магадлалыг дараах томъёогоор олно.

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

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

Тухайн тархалт

Тодорхойлолт 18. 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	Σ
	1 0	
1	0.2 0.3	0.5
0	0.1 0.4	0.5
Σ	0.3 0.7	1

Хүснэгт 9: Хамтын тархалт ба тухайн тархалт

$$P(X_2 = 1) = P(X_1 = 1, X_2 = 1) + P(X_1 = 0, X_2 = 1) \\ = 0.2 + 0.1 = 0.3$$

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



Зураг 51: Тасралтгүй санамсаргүй векторын хамтын болон тухайн нягт

Нөхцөлт тархалт

Сэргээн санах нь 8. 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 9. • $P(AB) = P(A)P(B)$ 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 $X \perp Y$ байдлаар тэмдэглэнэ. Мөн үүнээс дараах нөхцлүүд мөрдөн гарна.

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

X_1	X_2		Σ
	1	0	
1	0.2	0.3	0.5
0	0.1	0.4	0.5
Σ	0.3	0.7	1

X_1	1	0	Σ
$f_{X_1 X_2}(x_1 X_2 = 1)$	2/3	1/3	1

Хүснэгт 10: Хамтын, тухайн болон нөхцөлт тархалт

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

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.

Дискрет санамсаргүй хувьсагчийн нөхцөлт тархалтыг дараах байдлаар олно.

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

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.

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

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

$f_{X|Y}(x|y) = \frac{3x^2 + 8y}{8 + 16y}$ нөхцөлт тархалт нь нөхцөлд буй Y хувьсагчийн утгаас

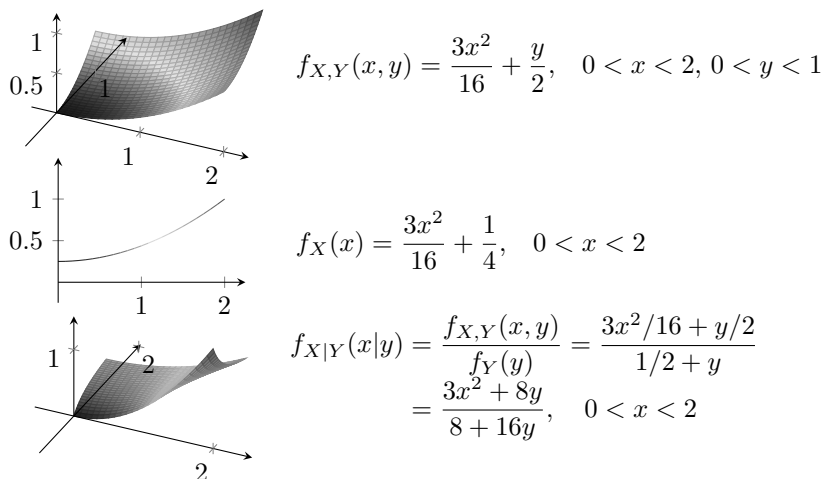
хамаарсан буюу $f_X(x) = \frac{3x^2}{16} + \frac{1}{4}$ тухайн тархалтаас ялгаатай байгаа тул эдгээр хувьсагчид хамааралтай юм. Мөн (X, Y) векторын хувьд $f_{X,Y}(x,y) = \frac{3x^2}{16} + \frac{y}{2} \neq f_X(x)f_Y(y)$ буюу хамтын нягт нь тухайн нягтуудын үржвэрт таамаглаагүй байгаа явдал нь тус хувьсагчдыг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

2 Бүтэн магадлалын томьёо

Үржүүлэх дүрэм ба бүтэн магадлалын томьёо

X хувьсагч Y хувьсагчаас хамаардаг гэж үзье.

Үржүүлэх дүрэм



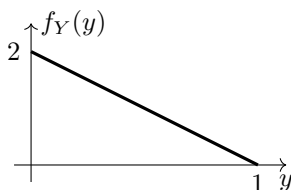
Зураг 52: Тасралтгүй санамсаргүй хувьсагчийн хамтын тархалт, тухайн тархалт, нөхцөлт тархалт

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

Бүтэн магадлалын томьёо

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

Жишээ 26 (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 завшдаг бол санд орох өргөл барьцын хэмжээний тархалтыг ол.

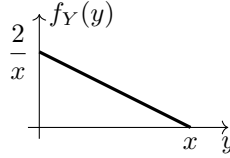


Зураг 53: Y санамсаргүй хувьсагчийн тархалт

Өргөл барьцын анхны хэмжээг X гэвэл бодлогын нөхцөл ёсоор

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

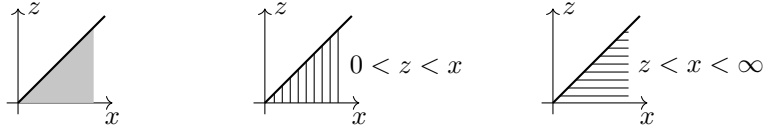
болно. Харин санд орох өргөл барьцын хэмжээ $Z = X \cdot Y$ буюу X хувьсагчаас хамаарна. Тэгэхээр үүнийг Y хувьсагчийн хувиргалт гэж болно.



Зураг 54: X хувьсагчийн нөхцөл дэх Z санамсаргүй хувьсагчийн тархалт

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

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. Зургаас харвал $0 < Z < \infty$ ба $Z < X < \infty$ байна. Иймд бүтэн



Зураг 55: (X, Z) санамсаргүй векторын авах утга

магадлалын томьёогоор дараах илтгэгч тархалт олдоно.

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

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

Үржүүлэх дүрэм

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

Бүтэн магадлалын томьёо

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

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

Жишээ 27. 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.

$$\begin{aligned} P(A) &= P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) \\ &= 1 \cdot \frac{2}{3} + \frac{1}{4} \cdot \frac{1}{3} \\ &= \frac{3}{4} = 0.75 \end{aligned}$$

3 Байесын зарчим

Байесын зарчим

Нөхцөлт нягт дахь хувьсагчдыг бас нөхцөлт магадлал дахь үзэгдлүүдийг жинхэнэ уялдаа холбооных нь эсрэг чиглэлээс авч үзэх өөрөөр хэлбэл X хувьсагч Y хувьсагчаас, A үзэгдэл B үзэгдлээс хамаардаг байтал $f_{Y|X}(y|x)$ нөхцөлт нягт, $P(B|A)$ нөхцөлт магадлал шаардлагатай болох явдал тохиолддог. Ийм үед эдгээрийн байрыг соливол бодлого хөнгөрнө. Ийнхүү байр солих буюу холбоо хамаарлыг жинхэнэ уялдаа холбооных нь зүгээс судлах боломжийг *Байесын томъёо* олгодог. Холбоо хамаарлыг үүн шиг нөгөө талаас нь авч үзэхийг *Байесын зарчим* харин уг зарчимд тулгуурласан статистик шинжилгээг *Байесын шинжилгээ* гэдэг.

Байесын томъёогоор $f_{Y|X}(y|x)$ нөхцөлт нягт ба $P(B|A)$ нөхцөлт магадлалыг $f_Y(y)$ нөхцөлт бус нягт ба $P(B)$ нөхцөлт бус магадлалд тулгуурлаж олно. Энэхүү нөхцөлт бус нягт ба магадлалыг *приор нягт* ба *приор магадлал* гэдэг бол нөхцөлт нягт ба нөхцөлт магадлалыг нь *постериор нягт* ба *постериор магадлал* гэдэг. Байесын шинжилгээний үр дүн буюу постериор нягт ба постериор магадлал ямар байх нь приор нягт ба приор магадлалаас шалтгаалдаг.

Байесын томъёо

Байесын томъёо

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{f_X(x)} \frac{f_Y(y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)} \frac{f_Y(y)}{f_X(x)} \\ &= \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)dy} \\ P(B_i|A) &= \frac{P(B_iA)}{P(A)} = \frac{P(B_iA)}{P(A)} \frac{P(B_i)}{P(B_i)} = \frac{P(AB_i)}{P(B_i)} \frac{P(B_i)}{P(A)} \end{aligned}$$

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

Жишээ 28. Хэрэв оюутан асуултад зөв хариулсан бол тэр уг асуултыг үнэхээр мэддэг байх магадлал ямар байх вэ?

$$\begin{aligned} P(B|A) &= \frac{P(A|B)P(B)}{P(A)} \\ &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})} \\ &= \frac{1 \cdot \frac{2}{3}}{1 \cdot \frac{2}{3} + \frac{1}{4} \cdot \frac{1}{3}} \\ &= \frac{8}{9} \approx 0.889 \end{aligned}$$

4 Нөхцөлт үл хамаарал

Нөхцөлт үл хамаарал

Сэргээн санах нь 10. $\forall (x, y) \in \mathbb{R}^2$ хувьд $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ бол X болон Y санамсаргүй хувьсагчдыг *хамааралгүй* гэнэ.

Тодорхойлолт 19. $\forall (x, y, z) \in \mathbb{R}^3$ 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 $(X \perp Y) \mid Z$ 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 олъя.

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

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 олж.

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

Энд бүтэн магадлалын томьёо ашиглав. Сүүлийн илэрхийлэл $f_X(x)f_Y(y)$ биш байгаа тул $X \not\perp Y$ 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 $(X \perp Y) | Z$ дүгнэлт гарна.

 $X \rightarrow Z \leftarrow Y$ тохиолдол

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

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

улмаар $X \perp Y$ 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 $X \perp Y$ 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 $(X \not\perp Y) | Z$ байна.

5 Нөхцөлт математик дундаж

Нөхцөлт математик дундаж

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

Жишээ 29. Богдын санд орох өргөл барьцын дундаж хэмжээ өргөл барьцын анхны хэмжээнээс хэрхэн хамаарахыг ол.

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

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

□

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 $E(X) = \frac{k}{\lambda}$ 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 математик дундаж

Жишээ 30. 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

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

байна. Энд Z_i нь i дүгээр өргөл барьцаас орох дундаж орлого юм. Бүтэн дундажийн томъёогоор 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 томъёо

Бүтэн дисперсийн томъёо

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

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

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

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

Лекц IX

Олон хэмжээст хэвийн тархалт ба шугаман загвар

1 Санамсаргүй векторын математик дундаж ба ковариацийн матриц

Санамсаргүй векторын математик дундаж

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

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

X_1	X_2		X_1	X_2	Σ	X_1	0	1
	-1	1		-1	1	$f_{X_1}(x_1)$	0.1	0.9
0	0.1	0.0	0	0.1	0.0			
1	0.5	0.4	1	0.5	0.4	X_2	-1	1
			Σ	0.6	0.4	$f_{X_2}(x_2)$	0.6	0.4
					1			

(a) Хамтын тархалт

(b) Хамтын тархалт ба тухайн тархалт

(c) Тухайн тархалт

Хүснэгт 11: Хамтын тархалт ба тухайн тархалт

Жишээ 31. 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 32.

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

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 коварианс ба корреляц

X болон Y санамсаргүй хувьсагчдын ковариансын коэффициент

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

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

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

Харин корреляцын коэффициентг дараах байдлаар тодорхойлдог.

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

Коварианс болон корреляцын чанар

Чанар 8. 1. $\text{cov}(X, X) = D(X)$

2. $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$

3. X болон Y хамааралгүй бол $\text{cov}(X, Y) = 0$ 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 33. Өмнөх жишээнүүдэд авч үзсэн санамсаргүй векторууд дахь хувьсагчдын ковариансын коэффициентг ол.

Үүнийг ковариансын 2 дугаар чанар болон ухамсаргүй статистикчийн хууль ашиглаж олѐ.

$$\begin{aligned} \text{cov}(X_1, X_2) &= E(X_1X_2) - E(X_1)E(X_2) \\ &= \sum_{(x_1, x_2)} x_1x_2f_{X_1, X_2}(x_1, x_2) - E(X_1)E(X_2) \\ &= 0 \cdot (-1) \cdot 0.1 + 0 \cdot 1 \cdot 0.0 + 1 \cdot (-1) \cdot 0.5 + 1 \cdot 1 \cdot 0.4 - E(X_1)E(X_2) \\ &= -0.1 - 0.9 \cdot (-0.2) = 0.08 \end{aligned}$$

Тасралтгүй тохиолдолд зарчмын хувьд өмнөхтэй төстэй байдлаар

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

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

Санамсаргүй векторын ковариацийн матриц

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

Чанар 9. Σ_{XX} тэгш хэмтэй, эерэг тодорхойлогдсон матриц байна.

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

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

нь $p \times q$ хэмжээст матриц байна.

Ковариацийн матриц ба корреляцын коэффициент

Жишээ 34. (X, Y) хоёр хэмжээст санамсаргүй векторын ковариацийн матриц $\begin{pmatrix} 25 & 4 \\ 4 & 1 \end{pmatrix}$ бол $\rho(X, Y)$ корреляцын коэффициент болон корреляцын матрицыг олно.

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

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

байна. Энэ тохиолдолд корреляцын матриц нь

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

болно.

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

$$\Sigma = \begin{pmatrix} \text{cov}(X_1, X_1) & 0 & \dots & 0 \\ 0 & \text{cov}(X_2, X_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \text{cov}(X_p, X_p) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p^2 \end{pmatrix}$$

байх ба улмаар

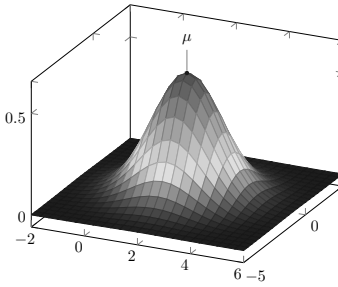
$$f_X(x) = \prod_{i=1}^p f_{X_i}(x_i) = \prod_{i=1}^p \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left\{ -\frac{1}{2\sigma_i^2} (x - \mu_i)^2 \right\}$$

болно.

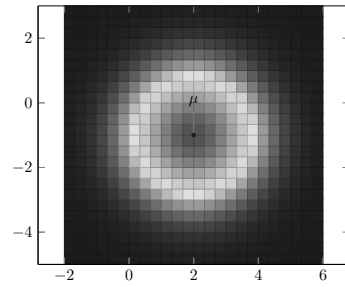
$$\mu = (2, -1)^T, \Sigma = \begin{pmatrix} 2/3 & 0 \\ 0 & 2/3 \end{pmatrix} \text{ параметртэй}$$

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

хэвийн тархалтын нягтыг зурагт дүрслэн харуулав.



(a) хажуугаас



(b) эгц дээрээс

Зураг 56: Хамааралгүй, ижил дисперстэй хувьсагчдын хоёр хэмжээст хэвийн тархалтын нягт

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.

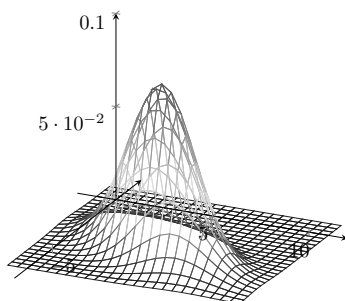
Жишээ 35. $\mu = (5, -4)$ дундаж утгын вектор болон $\Sigma = \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix}$ ковариацийн матриц бүхий хоёр хэмжээст хэвийн тархалтын хамтын нягтын илэрхийллийг бичиж, графикийг нь зур.

$\mu = (\mu_1, \mu_2) = (5, -4)$ 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

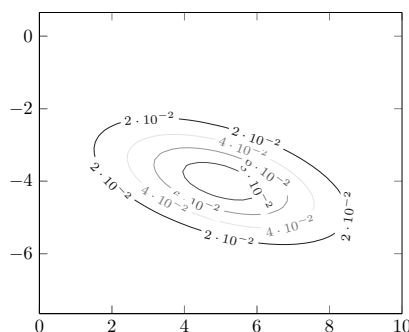
$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{2\sqrt{3}\pi} \exp \left\{ -\frac{1}{6} [(x_1 - 5)^2 + 2(x_1 - 5)(x_2 + 4) + 4(x_2 + 4)^2] \right\}$$

болно. Харин графикийг нь дараагийн слайд дээр байгуулж үзүүлье.

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



нягтын функцийг график

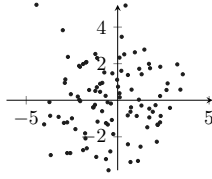


нягтын функцийг түвшний шугам

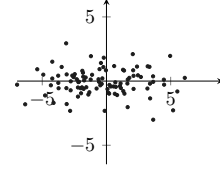
Зураг 57: Хоёр хэмжээст хэвийн тархалтын нягт, хувьсагчид хамааралтай бөгөөд дисперс нь ялгаатай үед

Ковариацийн матриц тархалтад нөлөөлөх байдал

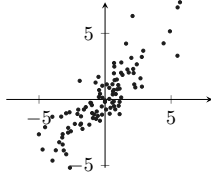
Санамсаргүй векторын нэг хуваалт



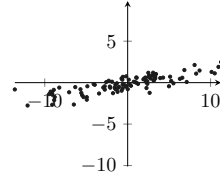
$$D(X_1) = D(X_2), \rho(X_1, X_2) = 0$$



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



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



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

Зураг 58: Хоёр хэмжээст хэвийн тархалттай санамсаргүй утгууд

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

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

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

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

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

Жишээ 36.

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

$r = 2$ 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. Ковариацийн матриц нь дараах блок матриц болно.

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

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

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

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

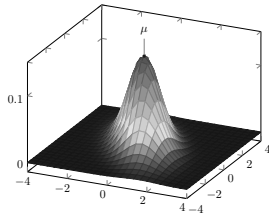
буюу мөн адил олон хэмжээст хэвийн тархалт байна. Харин олон хэмжээст хэвийн тархалттай санамсаргүй хувьсагчдын нөхцөлт математик дундаж болон нөхцөлт ковариацийн матриц дараах хэлбэртэй байдаг.

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

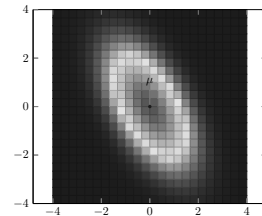
$$\text{cov}(X_2|X_1 = x_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

Жишээ 37. $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 2 \end{pmatrix}$ параметртэй хэвийн тархалтын хувьд $\text{cov}(X_2|X_1 = x_1)$ 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 59: 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

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

бас $\mu_1 = 0$, $\mu_2 = 0$ тул

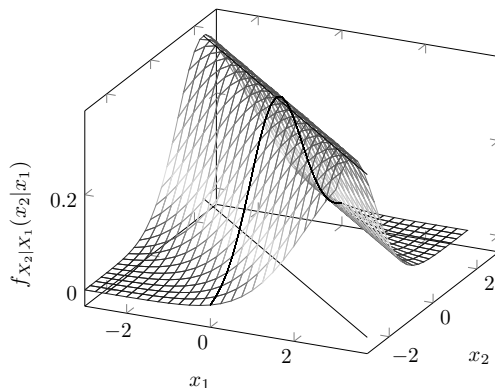
$$E(X_2|X_1 = x_1) = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 + \Sigma_{21}\Sigma_{11}^{-1}x_1 = -0.8x_1$$

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 олддоно.



Зураг 60: $f_{X_2|X_1}(x_2|x_1) = \frac{1}{\sqrt{2\pi}1.36} \exp\left\{-\frac{(x_2 + 0.8x_1)^2}{2(1.36)}\right\}$ нөхцөлт нягтын муруй

3 Регрессийн шугаман загвар

Санамсаргүй хувьсагчийн утгыг нөхцөлт математик дундаж ашиглаж прогнолох нь

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

Энд U нь прогнозын алдаа юм.

Чанар 10.

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

$$\begin{aligned}
 E(X_2|X_1) &= E(E(X_2|X_1) + U|X_1) \\
 E(X_2|X_1) &= E(E(X_2|X_1)|X_1) + E(U|X_1) \\
 E(X_2|X_1) &= E(X_2|X_1)E(1|X_1) + E(U|X_1) \\
 E(X_2|X_1) &= E(X_2|X_1)E(1) + E(U|X_1) \\
 E(X_2|X_1) &= E(X_2|X_1) + E(U|X_1) \\
 E(U|X_1) &= 0
 \end{aligned}$$

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 прогнолох 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

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

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 20 (Детерминацын коэффициент).

$$\begin{aligned}
 \underbrace{\text{cov}(X_2)}_{\text{нийт коварианс}} &= \underbrace{\text{cov}(\beta_0 + BX_1)}_{\text{тайлбарлагдах коварианс}} + \underbrace{\text{cov}(U)}_{\text{үл тайлбарлагдах коварианс}} \\
 \rho^2 &= \frac{\text{тайлбарлагдах коварианс}}{\text{нийт коварианс}}
 \end{aligned}$$

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

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

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 38. $\Sigma = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 2 \end{pmatrix}$ бол $X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$ шугаман загварын детерминацын коэффициенттэйг 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 $\Sigma_{11} = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$, $\sigma_{12} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $\sigma_{21} = (1 \ 2)$, $\sigma_{22} = 2$ 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 ковариацийн матриц, тухайн корреляц ба нөхцөлт үл хамаарал

Өмнө үзсэн

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

нөхцөлт ковариацийн матрицыг $r = p - 2$ үед авч үзье. Энэхүү хоёр хэмжээст ковариацийн 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 39. $\Sigma = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 5 & 2 \\ 2 & 2 & 3 \end{pmatrix}$ бол X_2 ба X_3 хоёр хувьсагч X_1 хувьсагчийн нөхцөлд хамааралгүй болохыг харуул.

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

$$\left. \begin{array}{l} X_1 \\ X_2 \\ X_3 \end{array} \right\} \begin{array}{l} X_1 \\ X_2 \end{array} \quad \Sigma_{11} = (2) \quad \Sigma_{12} = (2 \ 2) \quad \Sigma_{22} = \begin{pmatrix} 5 & 2 \\ 2 & 3 \end{pmatrix}$$

$$\text{cov}(X_2|X_1 = x_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

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

Тухайн корреляц олох бусад аргууд

Тухайн корреляцыг регрессийн шугаман загвар эсвэл ковариацийн матрицын урвуу ашиглан олж болно.

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

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

Жишээ 40. Өмнөх жишээний хувьд тухайн ковариацийн 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 $\rho_{X_1, X_2|X_3} = -\frac{-1/3}{\sqrt{11/6}\sqrt{1/3}} = \sqrt{2/11} \approx 0.4264$ гэж олдоно.

Лекц X

Хамааралгүй санамсаргүй хувьсагчдын нийлбэрийн тархалт

1 Хамааралгүй санамсаргүй хувьсагчдын нийлбэрийн тархалт

Хамааралгүй хувьсагчдын нийлбэрийн тархалт

$Z = X + Y$ гээ.

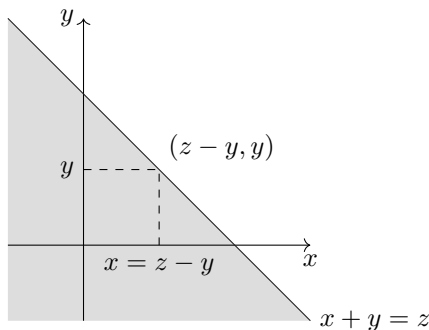
$$F_Z(z) = P(Z < z) = P(X + Y < z) = ?$$

Бүтэн магадлалын томьёоны янз бүрийн хэлбэрүүд

Сэргээн санах нь 11 (Бүтэн магадлалын томьёо).

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

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



Зураг 61: $Z = X + Y$ санамсаргүй хувьсагчийн тархалтын функцтэй холбогдох үзэгдэл

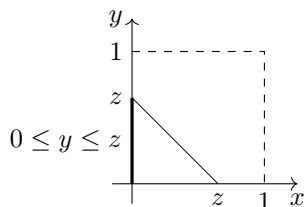
Хамааралгүй хувьсагчдын нийлбэрийн тархалт ба хуниас

X болон Y хамааралгүй байг.

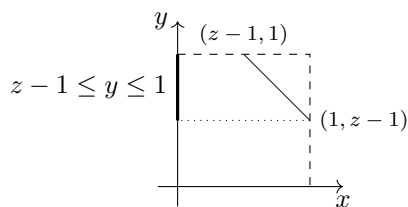
$$\begin{aligned}
 F_{X+Y}(z) &= P(X + Y < z) \\
 &= \int_{-\infty}^{\infty} P(X + Y < z \mid Y = y) f_Y(y) dy && \text{бүтэн магадлалын томьёо} \\
 &= \int_{-\infty}^{\infty} P(X < z - y \mid Y = y) f_Y(y) dy \\
 &= \int_{-\infty}^{\infty} P(X < z - y) f_Y(y) dy && \text{хамааралгүй} \\
 &= \int_{-\infty}^{\infty}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

Үүнийг f_X болон f_Y тархалтуудын *хуниас* гэнэ.

Жишээ 41. X болон Y хувьсагчид $(0, 1)$ завсарт жигд тархалттай бөгөөд хамааралгүй бол $X + Y$ нийлбэрийн тархалтыг олж.



(a) $0 \leq Z \leq 1$ байх тохиолдол



(b) $1 \leq Z \leq 2$ байх тохиолдол

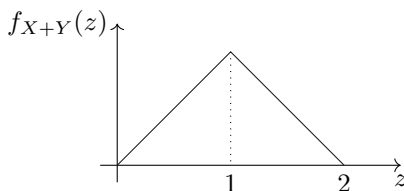
Зураг 62: Жигд тархалттай хувьсагчдын $Z = X + Y$ нийлбэр

- $0 \leq Z \leq 1$ тохиолдол

$$f_{X+Y}(z) = \int_0^z \underbrace{f_X(z-y)}_1 \underbrace{f_Y(y)}_1 dy = z$$

- $1 \leq Z \leq 2$ тохиолдол

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



Зураг 63: $(0, 1)$ завсарт жигд тархалттай хамааралгүй хувьсагчдын $X+Y$ нийлбэрийн тархалт

Жишээ 42. $X, Y \sim \text{Exp}(\lambda)$ хамааралгүй хувьсагчдын $X + Y$ нийлбэрийн тархалтыг олж.

$0 \leq X + Y \leq z$ 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 $\text{Gamma}(k=2, \lambda)$ тархалтын нягтын илэрхийлэл байна. Иймд $X + Y \sim \text{Gamma}(k=2, \lambda)$ болно.

Дискрет тархалтуудын хуниас

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

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

Жишээ 43. X болон Y хувьсагчид хамааралгүй бөгөөд харгалзан λ_X болон λ_Y параметр бүхий Пуассоны тархалттай бол $X + Y$ нийлбэрийн тархалтыг ол.

$X \geq 0$ ба $Y \geq 0$ 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.

2 Хязгаарын гол теорем

Ижил тархалттай, хамааралгүй санамсаргүй хувьсагчид ба тэдгээрийн нийлбэрийн тархалт

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

$$X_1, \dots, X_n \sim IID^{21}(\mu, \sigma^2)$$

байдлаар тэмдэглэж байя. Хэрэв $S_n = X_1 + \dots + X_n$ гэвэл үүний нягт нь n удаа нугалсан хуниас байна.

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

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

Жигд тархалттай, хамааралгүй хувьсагчдын нийлбэр

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

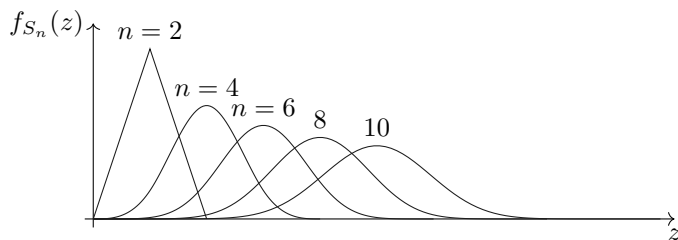
$$f_{S_n}(x) = \begin{cases} \frac{1}{(n-1)!} \sum_{0 \leq j \leq x} (-1)^j C_n^j (x-j)^{n-1}, & 0 < x < n \\ 0, & \text{бусад} \end{cases}$$

Илтгэгч тархалттай, хамааралгүй хувьсагчдын нийлбэр

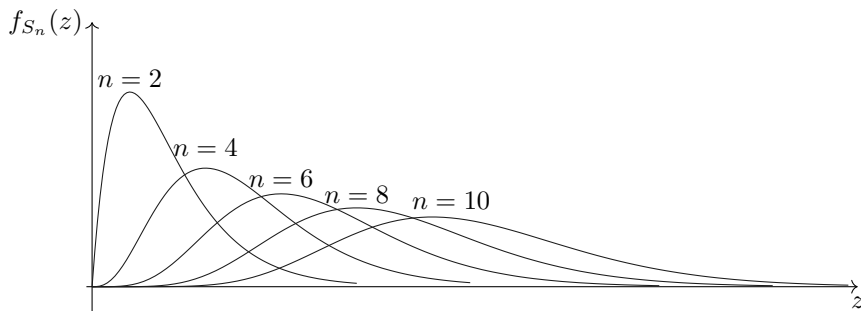
$X_1, \dots, X_n \sim \text{Exp}(\lambda)$ хамааралгүй бол

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

²¹independent identically distributed



Зураг 64: $U(0, 1)$ тархалттай, хамааралгүй хувьсагчдын нийлбэрийн тархалт



Зураг 65: $\text{Exp}(\lambda)$ тархалттай, хамааралгүй хувьсагчдын нийлбэрийн тархалт

Процесс ба санамсаргүй түүвэр

$$X_1, \dots, X_n$$

гэсэн нэг ижил тархалттай, хамааралгүй санамсаргүй хувьсагчид нь Бернуллийн процесс, Пуассоны процесс зэрэг ямар нэг процесс болон санамсаргүй түүврийг төлөөлж чадна. Ингээд $X_1, \dots, X_n \sim IID(\mu, \sigma^2)$ байх үед

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

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

$$\begin{aligned} E(\bar{S}_n) &= E\left[\frac{S_n}{n}\right] = \frac{E(S_n)}{n} = \frac{E(X_1 + \dots + X_n)}{n} \\ &= \frac{E(X_1) + \dots + E(X_n)}{n} = \frac{\mu + \dots + \mu}{n} = \mu \end{aligned}$$

\bar{S}_n хувьсагчийн дундаж квадрат хазайлт

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

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 12 (Момент үүсгэгч функц).

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

Санамж 4. 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.

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

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

болно. Үлдэх 3 дугаар чанарыг ашиглавал

$$S_n^* = \frac{\bar{S}_n - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma}$$

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 $X_i = \frac{X_i - \mu}{\sigma}$ стандарт хувиргалт хийвэл $X_1, \dots, X_n \sim IID(0, 1)$ 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

$$\begin{aligned} M_{X_1}(t) &= E[e^{tX_1}] = 1 + tE(X_1) + \frac{t^2}{2}E(X_1^2) + t^2h(t) \\ &= 1 + \frac{t^2}{2} + t^2h(t), \quad \text{энд } t \rightarrow 0 \text{ үед } h(t) \rightarrow 0 \end{aligned}$$

болно. Улмаар $\lim_{n \rightarrow \infty} \left(1 + \frac{a}{n}\right)^n = e^a$ хязгаар ашиглавал

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

үр дүнд хүрнэ.

Сэргээн санах нь 13. Стандарт хэвийн тархалттай X санамсаргүй хувьсагчийн момент үүсгэгч функц $M_X(t) = e^{t^2/2}$ байдаг.

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

Хязгаарын гол теорем

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

Теорем 1 (Хязгаарын гол теорем). $X_1, \dots, X_n \sim IID(\mu, \sigma^2)$ байг. Тэгвэл n хангалттай их үед

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

байна.

Лекц XI

Хамааралтай хувьсагчдын дараалал, Марковын хэлхээ

1 Хамааралтай хувьсагчдын дараалал

Хамааралтай хувьсагчдын дараалал

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

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 44 (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 томъёолж болно.

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

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.

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

Чанар 11. Дурын i бүрийн хувьд $\sum_{j=1}^m p_{ij} = 1$ буюу тусдаа тархалт байна.

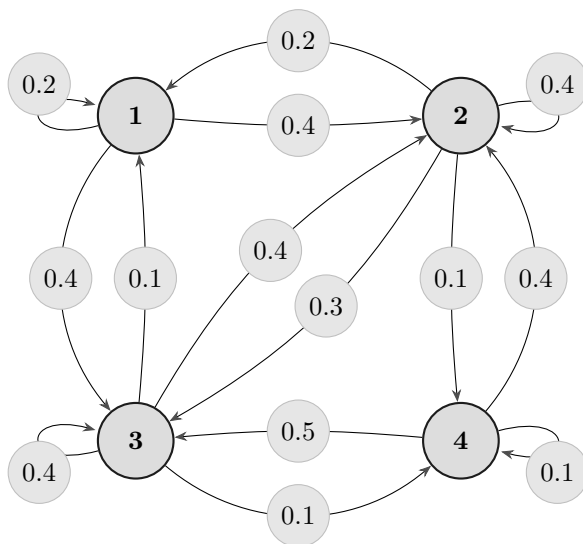
Жишээ 45 (Бямбажав Д., Магадлалын онол, математик статистик, 1999). Практикаас харахад гантай жилээс усархаг жилд, усархаг жилээс гантай жилд шууд шилждэггүй байна. Тийнхүү дараах шилжилтийн магадлалын матриц олджээ.

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

Жишээний хувьд тухайлбал $i = 1$ дүгээр мөрийн хувьд

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

байна.

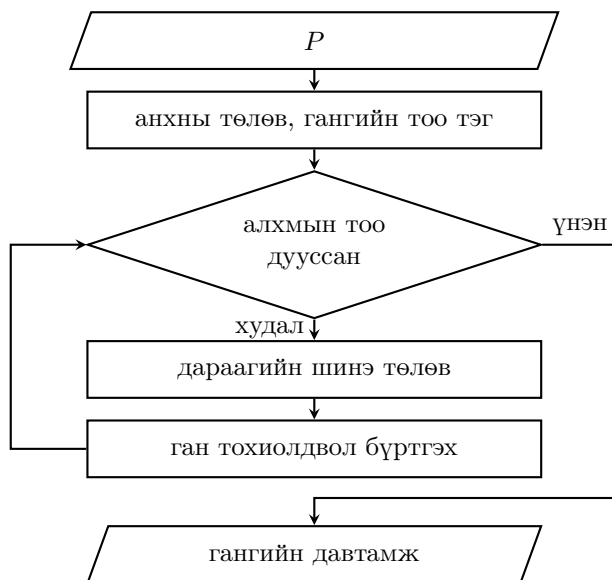


Зураг 66: Жишээгээр өгсөн шилжилтийн магадлалын матрицад харгалзах чиглэлт граф

Жишээ 46 (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.

```
(1) INPUT P := (pij)
(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
```



```

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

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

Сэргээн санал нь 14 (Үржүүлэх дүрэм). $P(AB) = P(A|B)P(B)$

Үржүүлэх дүрэм болон Марковын нөхцөл ёсоор

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

болно.

Өмнөхтэй адилаар цааш үргэлжлүүлбэл

$$\begin{aligned}
 P(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) \\
 &= p_{i_{n-1}i_n} \cdot p_{i_{n-2}i_{n-1}} \cdot \dots \cdot p_{i_0i_1} P(X_0 = i_0) \\
 &= P(X_0 = i_0) \cdot p_{i_0i_1} \cdot \dots \cdot p_{i_{n-2}i_{n-1}} \cdot p_{i_{n-1}i_n}
 \end{aligned}$$

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

Сэргээн санах нь 15 (Бүтэн магадлалын томъёо). B_1, \dots, B_k харилцан нийцгүй, $B_1 + \dots + B_k = \Omega$, $P(B_i) > 0$ бол

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

байна.

Бүтэн магадлалын томъёо, нөхцөлт магадлалын томъёо болон Марковын нөхцөл ёсоор

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

$$= \sum_{k=1}^m p_{ik}(n-1) \cdot p_{kj}$$

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

Жишээ 47. Энэ жил усархаг бол хоёр жилийн дараа ган тохиох магадлалыг ол.

Энэ тохиолдолд өмнөх томъёо

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

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. Ийнхүү

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

томъёо 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.

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 $\lim_{n \rightarrow \infty} P(X_n = j) = \pi_j$ буюу хугацаа өнгөрөх тусам систем j төлөвт байх магадлал тогтворжих явдал заримдаа тохиолддог. Марковын хэлхээний ийм (π_1, \dots, π_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 томъёогоор бодож олдог гэж үзсэн. Тэгвэл төлвийн стационар магадлалууд оршин байвал тэр нь үл хөдлөх буюу хугацаанаас хамаарахгүй тул тус рекуррент томъёо дахь хугацаанаас хамаарсан үл мэдэгдэгчдийн оронд бичигдэнэ. Учир нь өмнө байсан төлвийн магадлал хугацааны нэг алхмын дараа ч өөрчлөгдөлгүй хэвээрээ үлдэх ёстой юм. Ийнхүү төлвийн стационар магадлалуудыг олох дараах систем тэгшитгэл бичиж болно.

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

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

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

өөрөөр хэлбэл анхны төлвөөс үл хамаарсан асимптот стационар тархалт олдоно.

Сэргээн санах нь 16. $P(n) = P^n$ томъёог $p_{ij}(n)$ магадлал олох болон түүний асимптот шинж чанарыг судлахад ашиглана.

Тэгэхээр дээрх нөхцөл биелэх үед

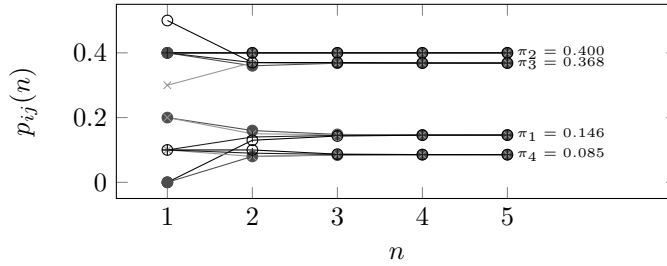
$$P(n) = P^n$$

томъёо ашиглаж асимптот стационар тархалт олох боломжтой.

Жишээний хувьд P^2 матрицын бүх элемент эерэг байсан тул

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

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



Зураг 67: Марковын хэлхээний асимптот стационар тархалтын нийлэлт

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 $i \rightarrow j$ 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 гее. 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 гее. 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

$$(V | X_0 = i) \sim \text{Geom}(1 - P(X_n = i | X_0 = i))$$

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

$$a_i = \sum_{j=1}^m p_{ij} \cdot a_j, \quad i = 1, \dots, m$$

систем тэгшитгэлээр нэг утгатай тодорхойлогдоно.

Эцэст нь систем шингээгч төлөвт шилжих дундаж хугацааг авч үзье. Төлөв бүрт харгалзах тус дундаж хугацааг μ_1, \dots, μ_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 48 (Муур ба хулгана). Муур хулгана хоёр дөрвөн өрөөтэй байшинд зураг дээр үзүүлсэн байдлаар байрлаж байв.

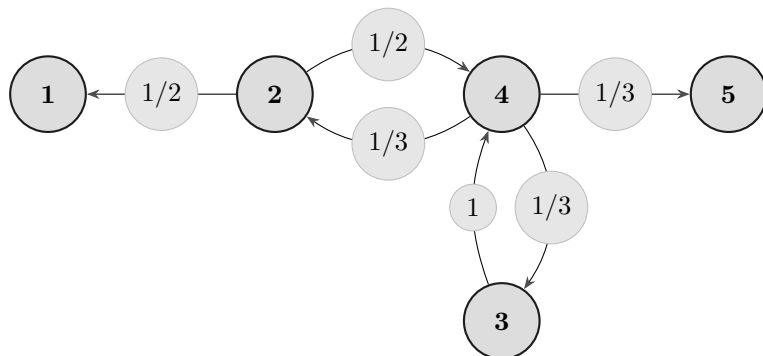


Хулгана хаалгануудын аль нэгийг тэнцүү магадлалтай сонгоно. Муур өрөөнөөсөө гарахгүй. Хэрэв хулгана мууртай өрөөнд орвол муур түүнийг барьж иднэ. Хулгана байшингаас гарч чадвал эргэж орохгүй.

”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 68: ”Муур ба хулгана” жишээний шилжилтийн магадлалын матрицад харгалзах чиглэлт граф

Шингээгч төлвүүд нь 1 болон 5 дугаар төлөв бөгөөд

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

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

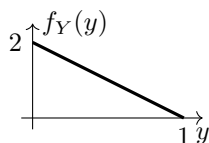
1 Тархалтын загварын тухай таамаглал дэвшүүлэх

Тархалтын загварын тухай таамаглал дэвшүүлэх

Тархалт нь үл мэдэгдэх X санамсаргүй хувьсагчийн эх олонлогоос авсан X_1, \dots, 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 17. 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 завшдаг бол санд орох өргөл



Зураг 69: Y санамсаргүй хувьсагчийн тархалт

барьцын хэмжээний тархалтыг ол.

Симуляцын аргаар гарган авсан өгөгдөл

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

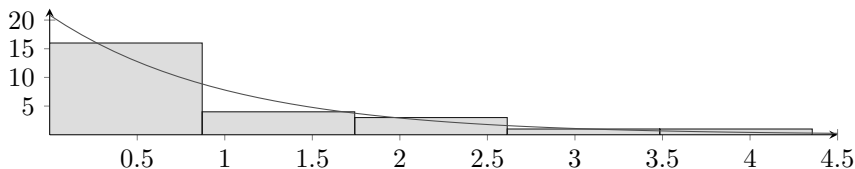
```
set.seed(0)
X <- rgamma(n = 25, shape = 3, rate = 1)
rtriangle <- function(n, a = 0, c = 0, b = 1) {
  U <- runif(n = n)
  ifelse(test = U < (c - a) / (b - a), a + sqrt(U * (b - a) * (c - a)), b - sqrt((1 - U) * (b - a) * (b - c)))
}
Y <- rtriangle(n = length(X))
Z <- X * Y
print(round(x = Z, digits = 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)
```



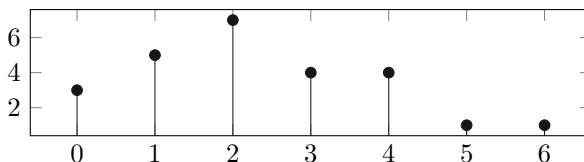
Зураг 70: Симуляцын аргаар гарган авсан өгөгдлийн гистограмм

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

Дискрет хувьсагчийн 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 71: Дискрет хувьсагчийн эх олонлогоос авсан өгөгдлийн давтамж

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

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 момент²² 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.

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}$$

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

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

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

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

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

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

буюу Богдын санд орох өргөл барьцын хэмжээ $\lambda = 0.984$ эрчмийн параметр бүхий илтгэгч тархалттай гэсэн өгүүлбэр боллоо. Бодлогын аналитик шийд $Z \sim \text{Exp}(\lambda = 1)$ гэж олдож байсан. Өөрөөр хэлбэл өгөгдөлд тулгуурлан олсон (үнэлсэн) параметрийн утга нь жинхэнэ утгаас "хазайсан" байна. Энэхүү хазайлтын талаар үргэлжлүүлэн үзэх болно.

Санамж 5. Заримдаа параметрийг түүнээс хамаарсан функцээр дамжуулан үнэлдэг.

²² $\alpha_k = E[X^k]$ энд $k = 1$

Жишээ болгон авч үзэж буй илтгэгч тархалтын параметрийн үнэлэлтийг

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

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

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

буюу $1/\hat{\lambda} = \bar{Z}$ нь *хазайлтгүй үнэлэлт*²³ юм. Ийнхүү бид онолын хувьд хазайлтгүй өөрөөр хэлбэл дундаж алдаа нь тэгтэй тэнцүү байх үнэлэлт ашиглажээ. Гэвч практикт 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, \dots, Z_n нь энгийн санамсаргүй түүвэр тул эдгээр хувьсагчид хамааралгүй бөгөөд бүгд нэг ижил илтгэгч тархалттай. Иймд үнэлэлтийн дундаж квадрат алдаа

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

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.

²³unbiased estimator

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

Сэргээн санах нь 18. $X_1, \dots, X_k \sim \text{Exp}(\lambda)$ хамааралгүй бол

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

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

Сэргээн санах нь 19.

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

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

буюу

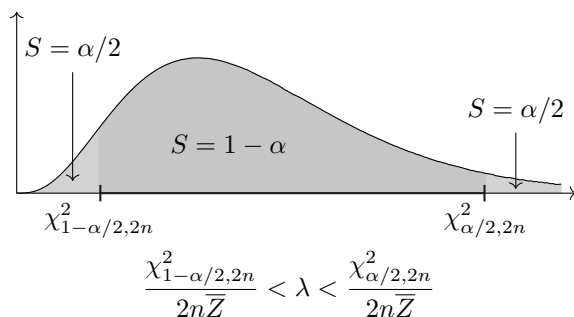
$$\frac{\chi_{1-\alpha/2, 2n}^2}{2n\bar{Z}} < \lambda < \frac{\chi_{\alpha/2, 2n}^2}{2n\bar{Z}}$$

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

```
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

$$\frac{\chi_{1-\alpha/2, 2n}^2}{2n\bar{Z}} < \lambda < \frac{\chi_{\alpha/2, 2n}^2}{2n\bar{Z}}$$



Зураг 72: Илтгэгч тархалтын эрчмийн параметрийн $1 - \alpha$ итгэх магадлалтай итгэх завсар

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.

Тархалтын параметр үнэлэхэд шаардагдах түүврийн минимал хэмжээ

Тархалтын параметрийн завсран үнэлэлтийн нарийвчлал буюу итгэх завсрын өргөнийг бэхэлбэл үүнээс тус параметрийг үнэлэхэд шаардагдах түүврийн минимал хэмжээг олох тэгшитгэл гарч ирдэг. Жишээлбэл $N(\mu, \sigma^2)$ хэвийн тархалтын дундаж буюу μ параметрийн итгэх завсар

$$\bar{X} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \frac{\sigma}{\sqrt{n}}$$

томъёотой байдаг. Энд $\Phi^{-1}()$ бол стандарт хэвийн тархалтын квантилын функц юм. Уг интервалын уртын хагас буюу алдааны хязгаарыг m гэвэл дараах тэгшитгэл гарна.

$$m = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \frac{\sigma}{\sqrt{n}}$$

Эндээс μ параметрийг үнэлэхэд шаардагдах түүврийн хэмжээг заадаг томъёо олдоно.

$$n = \left\lceil \frac{\left[\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sigma\right]^2}{m^2} \right\rceil$$

Илтгэгч тархалтын параметр үнэлэхэд шаардагдах түүврийн хэмжээ

Илтгэгч тархалтын эрчмийн параметрийн итгэх завсар

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

томьёогоор тодорхойлогдсон. Энэ тохиолдолд интервалын өргөнийг тодорхойлоход оролцож буй хи-квадрат тархалтын квантил нь түүврийн хэмжээнээс хамаарах тул уг түүврийн хэмжээг өмнөх шиг шууд олох боломжгүй. Иймд итгэх завсрын өргөн буюу завсран үнэлэлтийн нарийвчлалыг параметрийн жинхэнэ утга ба завсран үнэлэлтийн алдааны хязгаар хоёрын зөрүүг параметрийн жинхэнэ утгад харьцуулсан

$$m = \frac{\lambda - \frac{\chi_{1-\alpha/2, 2n}^2}{2n\bar{X}}}{\lambda} \quad m = \frac{\frac{\chi_{\alpha/2, 2n}^2}{2n\bar{X}} - \lambda}{\lambda}$$

харьцаагаар хэмждэг.

Өмнөх хоёр тэгшитгэлээс дараах томьёо гарна.

$$\frac{\chi_{\alpha/2, 2n}^2}{\chi_{1-\alpha/2, 2n}^2} = \frac{1+m}{1-m}$$

Итгэх магадлал $1 - \alpha$ болон завсран үнэлэлтийн нарийвчлал m өгөгдсөн үед дээрх томьёоноос хи-квадрат тархалтын чөлөөний зэрэг буюу түүврийн хэмжээ n олдоно.

Жишээ 49. Богдын санд орох өргөл барьцын хэмжээг илэрхийлэх илтгэгч тархалттай Z санамсаргүй хувьсагчийн λ параметрийн хувьд олсон 90 хувийн итгэх завсар шиг нарийвчлалтай завсран үнэлэлтэд шаардагдах түүврийн хэмжээг ол.

Энэ тохиолдолд $\alpha = 0.1$ ба $m \approx 0.33$ гэж сонгосон явдал болох тул $\frac{1+m}{1-m} \approx 1.985$ байх бөгөөд $n = 23$ үед үүнтэй хамгийн ойролцоо $\frac{\chi_{\alpha/2, 2n}^2}{\chi_{1-\alpha/2, 2n}^2} \approx 1.998$ харьцаа гарна. Ийнхүү шаардлагатай түүврийн хэмжээ $n = 23$ гэж олдсон нь тус завсрыг байгуулахад ашигласан түүврийн хэмжээ 25-тай ойролцоо байна.

Өмнөх үр дүнг гарган авахад шаардлагатай тооцоог дараах байдлаар хийж болно.

```
alpha <- 0.1; m <- 0.33
ratio <- {1+m}/{1-m}
ratio_chisq <- function(n) {
  qchisq(p = alpha/2, df = 2 * n, lower.tail = FALSE) / qchisq(p
    = 1 - alpha/2, df = 2 * n, lower.tail = FALSE)
}
n <- 1; error <- abs(ratio - ratio_chisq(n))
repeat {
  e <- abs(ratio - ratio_chisq(n + 1))
  if (e > error)
    break
  error <- e; n <- n + 1
}
print(n); print(ratio_chisq(n)); print(ratio)
```

Хи-квадрат тархалтын чөлөөний зэрэг k нь хангалттай их үед $\sqrt{2\chi^2} \sim N(\sqrt{2k-1}, 1)$ байдаг гэсэн чанар буюу Фишерийн дөхөлт ашиглавал өмнө гар-

гасан томьёо дахь хи-квадрат тархалтын сүүлний утгуудын харьцаа

$$\frac{\chi_{\alpha/2, 2n}^2}{\chi_{1-\alpha/2, 2n}^2} = \left(\frac{\frac{\sqrt{2\chi_{\alpha/2, 2n}^2 - \sqrt{4n-1}}}{\sqrt{1}} + \sqrt{4n-1}}{\frac{\sqrt{2\chi_{1-\alpha/2, 2n}^2 - \sqrt{4n-1}}}{\sqrt{1}} + \sqrt{4n-1}} \right)^2 = \left(\frac{\sqrt{4n-1} + \Phi^{-1}(1-\alpha/2)}{\sqrt{4n-1} - \Phi^{-1}(1-\alpha/2)} \right)^2$$

хэлбэртэй болно. Үүнийг мөнөөх томьёоны $\frac{1+m}{1-m}$ илэрхийлэлтэй тэнцүүлж бодвол

$$n = \left\lceil \frac{1}{4} + \left(\frac{\sqrt{\frac{1+m}{1-m}} + 1}{\sqrt{\frac{1+m}{1-m}} - 1} \frac{\Phi^{-1}(1-\alpha/2)}{2} \right)^2 \right\rceil$$

томьёо гарна. Гаргалгаанд нь Фишерийн дөхөлт ашигласан тул уг томьёог их түүвэр авах үед ашиглавал зохино.

Жишээ болгон $1-\alpha = 0.95$ ба $m = 0.1$ сонголтод харгалзах түүврийн хэмжээг сая гаргасан томьёогоор олж. Тус томьёог их түүврийн хувьд хэрэглэх ёстой тул α болон m тооны утгыг ийнхүү багаар сонголоо. Тооцоог R програм дээр дараах байдлаар хийж болно.

```
alpha <- 0.05; m <- 0.1
ratio <- {1+m}/{1-m}
round(1/4 + {{sqrt(ratio)+1}/{sqrt(ratio)-1}*qnorm(p =
1-alpha/2)/2}^2)
```

Эндээс түүврийн хэмжээ $n = 382$ гэсэн үр дүн гарна. Түүнчлэн хи-квадратын харьцаа бүхий анхны томьёогоор үүнтэй ижил хариу гарах бөгөөд үүнийг өмнө өгсөн програмын кодын эхний мөрийг `alpha <- 0.05; m <- 0.1` гэж өөрчлөн ажиллуулах байдлаар шалгаж болно.

Лекц XIII

Хамгийн их үнэний хувь бүхий үнэлэлт, Байесын үнэлэлт

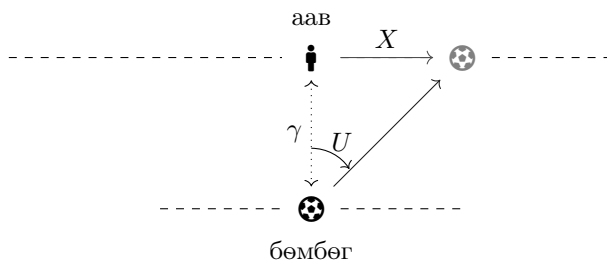
1 Кошийн тархалт

Кошийн тархалт

Жишээ 50. 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.

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.

$$X = \gamma \operatorname{tg}(U)$$



Зураг 73: Жишээ бодлогын зураглал

Энд $X < 0$ утга зураг дээрх чиглэлийн эсрэг чиглэлд гүйх зайг илэрхийлнэ. Одоо X хувьсагчийн тархалтын функцийг олж.

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

байх тул

$$\begin{aligned} F_X(x) &= P(X < x) = P(\gamma \operatorname{tg}(U) < x) = P\left[U < \operatorname{arctg}\left(\frac{x}{\gamma}\right)\right] \\ &= F_U(\operatorname{arctg}(x/\gamma)) = \frac{1}{2} + \frac{1}{\pi} \operatorname{arctg}(x/\gamma) \quad x \in \mathbb{R} \end{aligned}$$

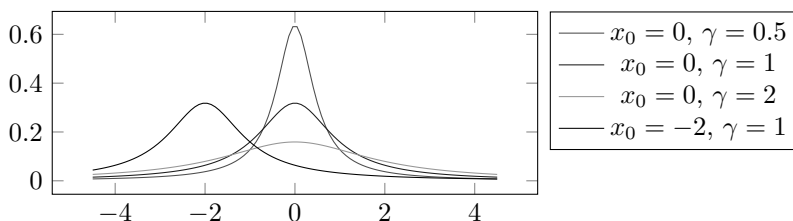
гэж олдоно. Уламжлал авбал дараах хэлбэртэй нягтын функц олдоно.

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

Кошийн тархалт, түүний параметрууд

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

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



Зураг 74: Кошийн тархалтын нягтын муруй параметрийн янз бүрийн утгад

Кошийн тархалтын зарим чанар болон хэрэглээ

Чанар

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

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

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

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.

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

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

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

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

Санамж 6. 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.

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

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

$$\begin{aligned} \frac{d}{d\gamma} \ln L(X, \gamma) &= \frac{d}{d\gamma} \left(-n \ln(\pi\gamma) - \sum_{i=1}^n \ln(1 + (X_i/\gamma)^2) \right) \\ &= -\frac{n}{\gamma} - \sum_{i=1}^n \frac{2(X_i/\gamma)X_i(-1/\gamma^2)}{1 + (X_i/\gamma)^2} \\ &= -\frac{n}{\gamma} + \frac{2}{\gamma} \sum_{i=1}^n \frac{X_i^2}{\gamma^2 + X_i^2} = 0 \end{aligned}$$

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

$$\min_i |X_i| \leq \gamma \leq \max_i |X_i|$$

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.

3 Байесын үнэлэлт

Байесын зарчим

Сэргээн санах нь 20 (Байесын томъёо).

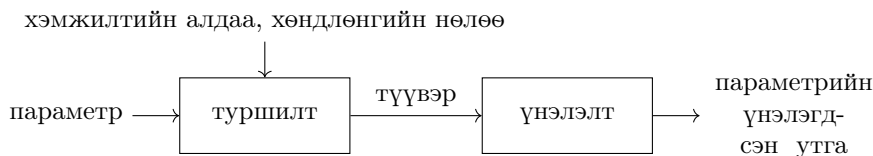
$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}$$

$$P(\text{спам}|\text{үгс}) = \frac{P(\text{үгс}|\text{спам})P(\text{спам})}{P(\text{үгс})}$$

Байесын үнэлэлт

Өмнө үзсэн сонгодог үнэлэлтүүдийн хувьд параметрийг тогтмол гэж тооцож байсан. Харин Байесын үнэлэлтийн хувьд параметрийг санамсаргүй хувьсагч байх тохиолдлыг авч үздэг. 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 $X = (X_1, \dots, X_n)$ түүврээс хамааруулан $f_{\Theta|X}(\theta|x)$ байдлаар авч үздэг.

$$f_{\Theta|X}(\theta|x)$$



Зураг 75: Параметрийн статистик үнэлэлт

нөхцөлт тархалт дахь X түүвэр өөрөө Θ параметрээс хамаарах тул энэ чигээр нь судлахад хүнд юм. Иймд уг тархалтыг Байесын зарчимд тулгуурлан

$$\underbrace{f_{\Theta|X}(\theta|x)}_{\text{постериор тархалт}} = \frac{f_{X|\Theta}(x|\theta) \overbrace{f_{\Theta}(\theta)}^{\text{приор тархалт}}}{f_X(x)}$$

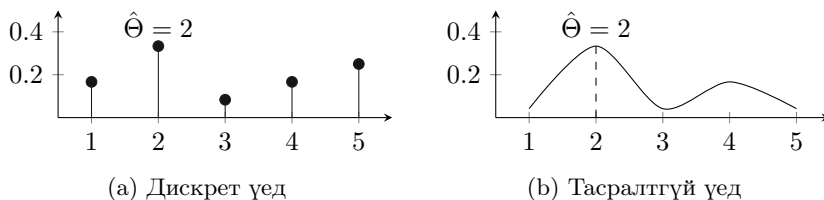
хэлбэрт шилжүүлж судалдаг. Энд $f_{\Theta}(\theta)$ тархалтыг мэдэгддэг гэж тооцно. Иймд $f_{\Theta|X}(\theta|x)$ постериор тархалтаас гарах үнэлэлт $f_{\Theta}(\theta)$ приор тархалтын "сонголтоос" хамаарна. Энэ хичээлээр Байесын зарчимд тулгуурласан дараах үнэлэлтүүдийг авч үзнэ. Үүнд:

- Хамгийн их постериорын үнэлэлт
- Хамгийн бага дундаж квадрат алдаатай үнэлэлт

Хамгийн их постериорын үнэлэлт

$$\hat{\Theta} = \arg \max_{\theta} f_{\Theta|X}(\theta|x)$$

Хамгийн их постериорын үнэлэлтээр ерөнхийдөө параметрийн ганц утга ол-



Зураг 76: Хамгийн их постериорын үнэлэлт

доно.

Хамгийн бага дундаж квадрат алдаатай үнэлэлт

Сэргээн санах нь 21. $E(X_2|X_1 = x_1)$ нөхцөлт математик дундаж нь X_1 хувьсагч ашиглаж X_2 хувьсагчийг прогнолох бүх $h(X_1)$ функц дундаас хамгийн бага дундаж квадрат алдаатай нь юм.

Иймд $\hat{\Theta} = E(\Theta|X = x)$ нь

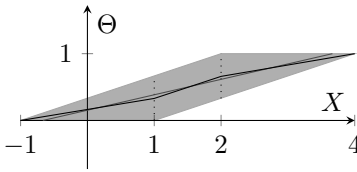
$$E[(\hat{\Theta} - \Theta)^2|X = x] \leq E[(h(x) - \Theta)^2|X = x]$$

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.

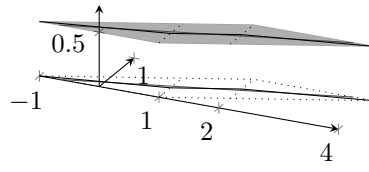
Жишээ 53.

$$\begin{cases} \Theta \sim U(0, 1) \\ X = 3\Theta + U \quad U \sim U(-1, 1) \quad \text{cov}(U, \Theta) = 0 \end{cases}$$

тохиолдолд $\hat{\Theta} = E(\Theta|X = x)$ үнэлэлт олж.



(a) (X, Θ) векторын авах утга



(b) (X, Θ) векторын хамтын нягт

Зураг 77: (X, Θ) векторын хамтын тархалт ба $\hat{\Theta} = E(\Theta|X = x)$ үнэлэлт

Эхлээд (X, Θ) санамсаргүй векторын хамтын тархалтыг олж. Үүний тулд нэн тэргүүнд боломжит утгын олонлогийг нь олох хэрэгтэй. Бодлогын нөхцөлийг харвал $0 \leq \Theta \leq 1$ 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.

$$f_{X,\Theta}(x, \theta) = \begin{cases} c, & (x, \theta) \in \text{параллелограмм} \\ 0, & (x, \theta) \notin \text{параллелограмм} \end{cases}$$

Нягтын функцийг чанар ашиглавал

$$\begin{aligned} 1 &= \int_{\text{параллелограмм}} f_{X,\Theta}(x, \theta) dx d\theta \\ &= c \int_{\text{параллелограмм}} dx d\theta \\ &= c \cdot S_{\text{параллелограмм}} \\ &= 2c \end{aligned}$$

байдлаар $c = 0.5$ гэж олдоно. Ийнхүү

$$f_{X,\Theta}(x, \theta) = \begin{cases} 0.5, & (x, \theta) \in \text{параллелограмм} \\ 0, & (x, \theta) \notin \text{параллелограмм} \end{cases}$$

боллоо.

Эцсийн 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 $-1 \leq x \leq 1$, $1 \leq x \leq 2$ болон $2 \leq x \leq 4$ завсар бүрт өөр өөр байх ажээ. Одоо эдгээрийг тус тусад нь бодож олж.

$-1 \leq x \leq 1$ үед параллелограммын дээд ирмэгт $y = \frac{1}{3}x + \frac{1}{3}$ шулуун харгалзах бөгөөд $0 \leq \Theta \leq x/3 + 1/3$ байх тул X санамсаргүй хувьсагчийн тухайн нягтын илэрхийлэл

$$f_X(x) = \int_0^{x/3+1/3} 0.5d\theta = \frac{x+1}{6} \quad -1 \leq x \leq 1$$

гэж олдоно. Тэгвэл

$$f_{\Theta|X}(\theta|X=x) = \frac{f_{X,\Theta}(x, \theta)}{f_X(x)} = \frac{0.5}{\frac{x+1}{6}} = \frac{3}{x+1} \quad -1 \leq x \leq 1$$

нөхцөлт нягт олдоно. Улмаар

$$\begin{aligned} \hat{\Theta} &= E(\Theta|X=x) = \int_0^{x/3+1/3} \theta \frac{3}{x+1} d\theta \\ &= \frac{3}{x+1} \left[\frac{\theta^2}{2} \right]_0^{x/3+1/3} = \frac{x+1}{6} \quad -1 \leq x \leq 1 \end{aligned}$$

үнэлэлт олдоно.

$1 \leq x \leq 2$ үед $x/3 - 1/3 \leq \Theta \leq x/3 + 1/3$ байх тул X санамсаргүй хувьсагчийн тухайн нягтын илэрхийлэл

$$f_X(x) = \int_{x/3-1/3}^{x/3+1/3} 0.5d\theta = \frac{1}{3} \quad 1 \leq x \leq 2$$

гэж олдоно. Тэгвэл

$$f_{\Theta|X}(\theta|X=x) = \frac{f_{X,\Theta}(x, \theta)}{f_X(x)} = \frac{0.5}{\frac{1}{3}} = \frac{3}{2} \quad 1 \leq x \leq 2$$

нөхцөлт нягт олдоно. Улмаар

$$\hat{\Theta} = E(\Theta|X=x) = \int_{x/3-1/3}^{x/3+1/3} \theta \frac{3}{2} d\theta$$

$$= \frac{3}{2} \left[\frac{\theta^2}{2} \right]_{x/3-1/3}^{x/3+1/3} = \frac{x}{3} \quad 1 \leq x \leq 2$$

үнэлэлт олдоно.

$2 \leq x \leq 4$ үед $x/3 - 1/3 \leq \Theta \leq 1$ байх тул X санамсаргүй хувьсагчийн тухайн нягтын илэрхийлэл

$$f_X(x) = \int_{x/3-1/3}^1 0.5 d\theta = \frac{4-x}{6} \quad 2 \leq x \leq 4$$

гэж олдоно. Тэгвэл

$$f_{\Theta|X}(\theta|X=x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{0.5}{\frac{4-x}{6}} = \frac{3}{4-x} \quad 2 \leq x \leq 4$$

нөхцөлт нягт олдоно. Улмаар

$$\begin{aligned} \hat{\Theta} &= E(\Theta|X=x) = \int_{x/3-1/3}^1 \theta \frac{3}{4-x} d\theta \\ &= \frac{3}{4-x} \left[\frac{\theta^2}{2} \right]_{x/3-1/3}^1 = \frac{x+2}{6} \quad 2 \leq x \leq 4 \end{aligned}$$

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

$$\begin{aligned} a &= E(Y) - bE(X) \\ b &= \frac{\text{cov}(X, Y)}{D(X)} \end{aligned}$$

байдаг (үүнийг хожим үзнэ) тул $Y = E(Y) + \frac{\text{cov}(X, Y)}{D(X)}[X - E(X)]$ буюу тус бодлогын хувьд

$$\hat{\Theta} = E(\Theta) + \frac{\text{cov}(X, \Theta)}{D(X)}[X - E(X)]$$

тэгшитгэл бичигдэнэ.

$$\Theta \sim U(0, 1) \text{ тул } E(\Theta) = \frac{0+1}{2} = \frac{1}{2} \text{ байна.}$$

$$\begin{aligned} E(X) &= E(3\Theta + U) = 3E(\Theta) + E(U) = \frac{3}{2} \\ D(X) &= D(3\Theta + U) = 9D(\Theta) + D(U) + 2 \cdot 3 \text{cov}(\Theta, U) \\ &= 9 \cdot \frac{1}{12} + \frac{2^2}{12} + 0 = \frac{13}{12} \end{aligned}$$

U болон Θ хамааралгүй ($\text{cov}(U, \theta) = 0$) болохыг анхаарвал

$$\text{cov}(X, \Theta) = E(X\Theta) - E(X)E(\Theta) = E[(3\Theta + U)\Theta] - \frac{3}{2} \cdot \frac{1}{2}$$

$$\begin{aligned}
 &= 3E(\Theta^2) + E(U\Theta) - \frac{3}{4} = 3[D(\Theta) + (E(\Theta)^2)] - \frac{3}{4} \\
 &= 3\left[\frac{1}{12} + \frac{1}{4}\right] - \frac{3}{4} = \frac{1}{4}
 \end{aligned}$$

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. $\text{cov}(\tilde{\Theta}, \hat{\Theta}) = 0$
5. $D(\Theta) = D(\hat{\Theta}) + D(\tilde{\Theta})$

Нөхцөлт математик дундаж, шугаман загварын мөн чанарыг сэргээн сануулах үүднээс эдгээр чанарын баталгааг хойно орууллаа.

1. $\hat{\Theta} = E(\Theta|X)$ тул

$$\begin{aligned} E(\tilde{\Theta}|X) &= E(\hat{\Theta} - \Theta|X) = E(\hat{\Theta}|X) - E(\Theta|X) = E(\hat{\Theta}|X) - E(\Theta|X) \\ &= E(E(\Theta|X)|X) - E(\Theta|X) = E(\Theta|X) - E(\Theta|X) = \hat{\Theta} - \hat{\Theta} = 0 \end{aligned}$$

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. Ковариацийг задалсаны дараа дурын $h(X)$ үнэлэлтийн хувьд 3 дугаар чанар бас 2 дугаар чанар ашиглавал

$$\text{cov}(\tilde{\Theta}, h(X)) = E(\tilde{\Theta}h(X)) - E(\tilde{\Theta})E(h(X)) = 0$$

болох тул $\hat{\Theta}$ үнэлэлтийн хувьд ч $\text{cov}(\tilde{\Theta}, \hat{\Theta}) = 0$ байх юм.

5. 4 дүгээр чанарыг ашиглавал

$$D(\Theta) = D(\hat{\Theta} - \tilde{\Theta}) = D(\hat{\Theta}) + D(\tilde{\Theta}) - 2\text{cov}(\hat{\Theta}, \tilde{\Theta}) = D(\hat{\Theta}) + D(\tilde{\Theta})$$

Лекц XIV

Статистик таамаглал шалгах, тархалтын загварын тохирцыг ТОГТООХ

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

$H_0 : \theta = \theta_0$ ба $H_1 : \theta \neq \theta_0$ хоёр талт өрсөлдөгч таамаглал

- Нэг талт таамаглал

$H_0 : \theta = \theta_0$ ба $H_1 : \theta < \theta_0$ зүүн өрсөлдөгч таамаглал

$H_0 : \theta = \theta_0$ ба $H_1 : \theta > \theta_0$ 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 12: 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 шалгах аар зохиосон статистик юм.

$$\begin{array}{ccccc} \left\{ \begin{array}{c} \text{Эх} \\ \text{олонлог} \end{array} \right\} & = & \left\{ \begin{array}{c} H_0 \text{ үнэн} \\ \text{байх олонлог} \end{array} \right\} & \cup & \left\{ \begin{array}{c} H_0 \text{ худал} \\ \text{байх олонлог} \end{array} \right\} \\ \downarrow & & \downarrow & & \downarrow \\ \mathbb{R} & & \text{интервал} & & \text{интервал} \\ \text{түүвэр} \rightarrow & & \text{тоон шулууны цэг} & & \end{array}$$

H_0 худал байх олонлог мэдэгдэхгүй тул түүний оронд шинжүүрийн няцаах муж ашиглана. Шинжүүрийн няцаах мужийг I төрлийн алдаанд тулгуурлаж олно.

$$P(\text{I төрлийн алдаа}) = P(H_1|H_0) \leq \alpha \Rightarrow \text{шинжүүрийн няцаах муж}$$

Жишээ 54. 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

$$P(\text{I төрлийн алдаа}) \leq \alpha$$

буюу

$$P(H_0 \text{ таамаглалыг няцаах} | H_0 \text{ үнэн байх}) \leq \alpha$$

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 22. $\text{Exp}(\lambda)$ тархалттай санамсаргүй хувьсагчийн эх олонлогоос авсан түүврийн хувьд $2\lambda n \bar{X} \sim \chi^2(2n)$ байна.

Ийнхүү жишээнд дэвшүүлсэн таамаглалыг шалгахын тулд

$$X^2 = 2\lambda_0 n \bar{X} \sim \chi^2(2n)$$

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

$$P(\text{I төрлийн алдаа}) = P(H_1|H_0) \leq \alpha$$

тэнцэл биш

- Хоёр талт өрсөлдөгчтэй үед $P(X^2 \leq \chi_{1-\alpha/2, 2n}^2 \text{ эсвэл } X^2 \geq \chi_{\alpha/2, 2n}^2) \leq \alpha$
- Нэг талт зүүн өрсөлдөгчтэй үед $P(X^2 \leq \chi_{1-\alpha, 2n}^2) \leq \alpha$
- Нэг талт баруун өрсөлдөгчтэй үед $P(X^2 \geq \chi_{\alpha, 2n}^2) \leq \alpha$

хэлбэрт шилжинэ. Иймд шинжүүрийн няцаах муж нь харгалзан

- $X^2 \leq \chi_{1-\alpha/2, 2n}^2$ эсвэл $X^2 \geq \chi_{\alpha/2, 2n}^2$
- $X^2 \leq \chi_{1-\alpha, 2n}^2$
- $X^2 \geq \chi_{\alpha, 2n}^2$

болно. Энд $\chi_{\alpha, k}^2$ 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 $\bar{X} \approx 1.016$ тул шинжүүрийн статистикийн туршилтын утга

$$X^2 = 2\lambda_0 n \bar{X} = 2 \cdot 1 \cdot 25 \cdot 1.016 = 50.8$$

болно. $X^2 = 50.8 \not\leq \chi_{1-\alpha, 2n}^2 = \chi_{0.95, 50}^2 \approx 34.764$ буюу шинжүүрийн статистикийн туршилтын утга шинжүүрийн няцаах мужид унахгүй байгаа тул тэг таамаглалыг үл няцаана. Өөрөөр хэлбэл илтгэгч тархалтын эрчмийн параметрийн тухай

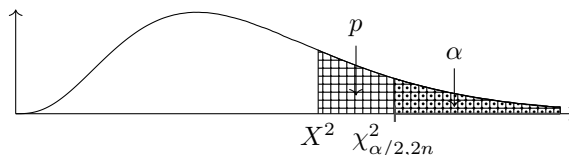
$$H_0 : \lambda = \lambda_0 \equiv 1$$

$$H_1 : \lambda < \lambda_0$$

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. $p\text{-утга} < \alpha$ бол H_0 таамаглалыг худал, эсрэг тохиолдолд үнэн гэж дүгнэнэ. Тэг таамаглалыг няцаах сонирхол-

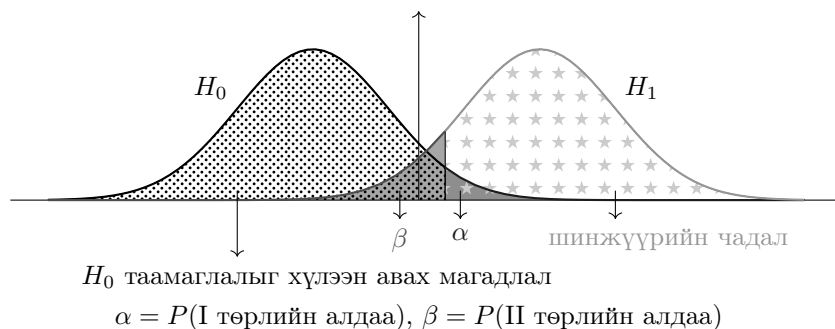


Зураг 78: Илтгэгч тархалтын эрчмийн параметрийн α итгэх түвшинтэй, нэг талт баруун өрсөлдөгчтэй тэг таамаглалын шинжүүрийн няцаах муж ба p -утга

той байгаа үед α итгэх түвшинг *ач холбогдлын түвшин* гэдэг.

Шинжүүрийн чадал

Өрсөлдөгч таамаглал үнэн үед тэг таамаглалыг няцаах магадлалыг *шинжүүрийн чадал* гэнэ.



Зураг 79: 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 давхцдаг.

Сэргээн санах нь 23 (Илтгэгч тархалтын параметрийн итгэх завсар).

$$\frac{\chi_{1-\alpha/2, 2n}^2}{2n\bar{X}} < \lambda < \frac{\chi_{\alpha/2, 2n}^2}{2n\bar{X}}$$

Өмнөх шинжүүрийн хоёр талт шинжүүрийн няцаах мужаас гүйцээлт авах буюу эсрэг нөхцөлийг нь бичээд $2n\bar{X}$ үржвэрт хуваавал дээрх итгэх завсар гарна.

$$\chi_{1-\alpha/2, 2n}^2 < X^2 \equiv 2\lambda n\bar{X} < \chi_{\alpha/2, 2n}^2$$

$$\frac{\chi_{1-\alpha/2, 2n}^2}{2n\bar{X}} < \lambda < \frac{\chi_{\alpha/2, 2n}^2}{2n\bar{X}}$$

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	\dots	x_k
$f_X(x)$	$f_X(x_1)$	$f_X(x_2)$	\dots	$f_X(x_k)$

- Таамаглал

$$H_0 : f_X(x_1) = p_1, \dots, f_X(x_k) = p_k$$

Энд p_1, \dots, p_k бол таамаглаж буй тоонууд юм.

- Түүврийн давтамж

x	x_1	x_2	\dots	x_k	Нийлбэр
Давтамж	n_1	n_2	\dots	n_k	n

Энд n_i бол өгөгдөл дотор байх x_i утгын тоо ширхэг юм.

- Шинжүүрийн статистик

$$X_{k-1}^2 = \sum_{i=1}^k \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i} = \frac{(n_1 - n \cdot p_1)^2}{n \cdot p_1} + \dots + \frac{(n_k - n \cdot p_k)^2}{n \cdot p_k}$$

- Шинжүүрийн статистикийн асимптот тархалт Хэрэв H_0 үнэн бөгөөд түүврийн хэмжээ n хүрэлцээтэй их бол

$$X_{k-1}^2 \sim \chi_{k-1}^2$$

байна.

- Шинжүүрийн няцаах муж, шинжүүрийн няцаах утга

$$X_{k-1}^2 \geq \chi_{\alpha, k-1}^2$$

бол α итгэх түвшинд тэг таамаглалыг няцаана. Энд $\chi_{\alpha, k-1}^2$ нь $k-1$ чөлөөний зэрэгтэй хи-квадрат тархалтын $1-\alpha$ эрэмбийн квантилын утга буюу α хэмжээтэй талбай бүхий тархалтын баруун сүүлний утга юм.

Жишээ 55. "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 13: Сүм хийдийн тоо, шашны төрлөөр, 2018 оны эцэст, УСХ

$$X_3^2 = \frac{(134 - 364 \cdot 0.4)^2}{364 \cdot 0.4} + \frac{(196 - 364 \cdot 0.5)^2}{364 \cdot 0.5} + \frac{(24 - 364 \cdot 0.05)^2}{364 \cdot 0.05} + \frac{(10 - 364 \cdot 0.05)^2}{364 \cdot 0.05} \approx 7.544$$

Энэ нь $\chi_{3,0.05}^2 = 7.815$ 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.

$$p\text{-утга} = P(\chi_3^2 \geq X_3^2) = P(\chi_3^2 \geq 7.544) \approx 0.05644$$

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 56. 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 санд 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, ∞) гурван интервалд хувааж бүлэглэе. Тэгвэл 16, 5, 4 давтамж олдоно. Одоо дээрх интервалуудад харгалзах магадлалуудыг олж. Тус магадлалуудыг тэг таамаглалын нөхцөлд өөрөөр хэлбэл $\lambda = 1$ эрчмийн параметр бүхий илтгэгч тархалтаар олно.

$$\begin{aligned} p_1 &= P(X \in [0, 1]) = P(X < 1) = F_X(1) = 1 - e^{-1 \cdot 1} \approx 0.632 \\ p_2 &= P(X \in (1, 2]) = P(1 < X \leq 2) = F_X(2) - F_X(1) \\ &= (1 - e^{-1 \cdot 2}) - (1 - e^{-1 \cdot 1}) \approx 0.232 \\ p_3 &= P(X \in (2, \infty)) = P(2 < X) = 1 - F_X(2) \\ &= 1 - (1 - e^{-1 \cdot 2}) \approx 0.135 \end{aligned}$$

Ийнхүү санамсаргүй хувьсагчийг $\lambda = 1$ параметр бүхий илтгэгч тархалттай гэсэн тэг таамаглал нь дээрх магадлалуудыг тус тархалтаар тооцоолж буй явдлаар дамжин тусгалаа олж байна.

Улмаар шинжүүрийн статистикийн туршилтын утгыг дараах байдлаар тооцоолж олно.

$$\chi^2 = \frac{(16 - 25 \cdot 0.632)^2}{25 \cdot 0.632} + \frac{(5 - 25 \cdot 0.232)^2}{25 \cdot 0.232} + \frac{(4 - 25 \cdot 0.135)^2}{25 \cdot 0.135} \approx 0.229$$

Энэ нь $\chi^2_{2,0.05} = 5.991$ няцаах утгаас их биш тул тэг таамаглалыг $\alpha = 0.05$ итгэх түвшинд үл няцаана. Үүнийг 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

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

$$P(H_1|X = x) \geq P(H_0|X = x)$$

буюу

$$\frac{P(X = x|H_1)P(H_1)}{P(X = x)} \geq \frac{P(X = x|H_0)P(H_0)}{P(X = x)}$$

эсвэл үүнтэй тэнцүү чанартай

$$LR(X) = \frac{P(X = x|H_1)}{P(X = x|H_0)} \geq \frac{P(H_0)}{P(H_1)} = c$$

нөхцөл биелж байвал H_0 таамаглалыг няцааж, H_1 таамаглалыг хүлээн авна.

Харин Байесын бус хувилбар нь

$$LR = \frac{P(X = x|H_1)}{P(X = x|H_0)} \geq c \quad (\text{дискрет тохиолдолд})$$

$$LR = \frac{f_X(x|H_1)}{f_X(x|H_0)} \geq c \quad (\text{тасралтгүй тохиолдолд})$$

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

$$P(\text{I төрлийн алдаа}) = P(H_1|H_0) = P(LR \geq c|H_0) = \alpha$$

нөхцөлөөр тодорхойлогдоно. Энд α бол итгэх түвшин юм.

Жишээ 57. $N(\theta, \sigma^2)$ 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.

$$\begin{aligned} LR(X) &= \prod_{i=1}^n \frac{f_X(X_i|H_1)}{f_X(X_i|H_0)} = \prod_{i=1}^n \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i - \theta_1)^2/(2\sigma^2)}}{\frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i - \theta_0)^2/(2\sigma^2)}} \\ &= \exp \left\{ \frac{n}{\sigma^2} (\theta_1 - \theta_0) \bar{X} - \frac{n}{2\sigma^2} (\theta_1^2 - \theta_0^2) \right\} \geq c \end{aligned}$$

Үргэлжлүүлэн хувиргавал дараах тэнцэл биш гарна.

$$\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} \geq \frac{\sigma}{\sqrt{n}(\theta_1 - \theta_0)} \ln c + \frac{\sqrt{n}}{2\sigma} (\theta_1 - \theta_0) = c'$$

Тэг таамаглалын нөхцөлд $\bar{X} \sim N(\theta_0, \sigma^2/n)$ буюу $\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} \sim N(0, 1)$ тул

$$\begin{aligned} P(\text{I төрлийн алдаа}) &= P(H_1|H_0) = P(LR \geq c|H_0) \\ &= P\left(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} \geq c'\right) = 1 - F_{\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}}}(c') = 1 - \Phi(c') = \Phi(-c') = \alpha \end{aligned}$$

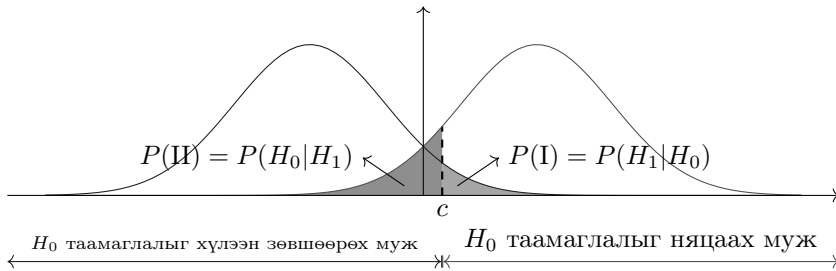
тэгшитгэлээс $c' = -\Phi^{-1}(\alpha) = \Phi^{-1}(1 - \alpha)$ шийд олдоно. Энд $\Phi()$ нь стандарт хэвийн тархалтын функц юм. Ийнхүү

$$\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} \geq \Phi^{-1}(1 - \alpha)$$

хэлбэртэй шинжүүрийн няцаах муж олдоо.

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 багсах бөгөөд энэ нь няцаах утгын сонголтоос шалтгаална.



Зураг 80: 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.

$$-2 \ln LR(X) \geq \chi_{\alpha,k}^2$$

Энд $\chi_{\alpha,k}^2$ 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.

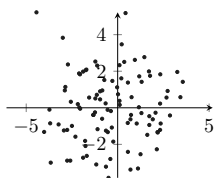
2 Регрессийн шугаман загвар

Корреляцын коэффициент ба шугаман хамаарал

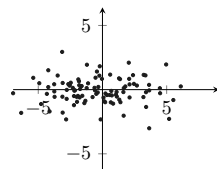
Сэргээн санах нь 24. Хувьсагчид шугаман хамааралтай үед корреляцын коэффициент үнэмлэхүй утгаараа нэгтэй тэнцүү байдаг.

Олон хэмжээст хэвийн тархалт, корреляцын коэффициент

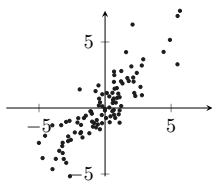
Корреляцын коэффициент тэгээс ялгаатай, үнэмлэхүй утгаараа нэгд ойр үед олон хэмжээст хэвийн тархалттай санамсаргүй хувьсагчдын хамаарлыг шугаман функцээр илэрхийлэх боломжтой.



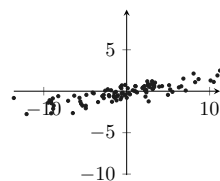
$$D(X_1) = D(X_2), \rho(X_1, X_2) = 0$$



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



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



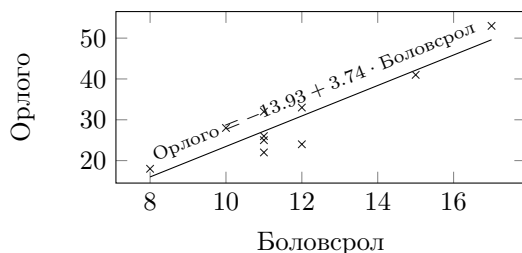
$$D(X_1) > D(X_2), \rho(X_1, X_2) \neq 0$$

Зураг 81: Хоёр хэмжээст хэвийн тархалттай санамсаргүй утгууд

Регрессийн шугаман загвар

Сэргээн санах нь 25. 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 прогнолох боломжтой. Ийм загварыг *регрессийн шугаман загвар* гэнэ.



Зураг 82: 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 26.

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

4-р чанар ёсоор $X_2 = E(X_2|X_1) + U$ бол хамгийн "сайн" загвар юм.

Детерминацын коэффициент

Сэргээн санах нь 27. $E(X_2|X_1 = x_1)$ нөхцөлт математик дунджийг X_2 хувьсагчийг прогнолоход ашиглах бол харин

$$\rho^2 = \frac{\text{cov}(X_2|X_1 = x_1)}{\text{cov}(X_2)} = \frac{D(X_2|X_1 = x_1)}{D(X_2)}$$

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 коэффициент

$$\rho^2 = \frac{D(X_2) - D(U)}{D(X_2)}$$

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 28. 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 параметруудийн дараах үнэлэлт гарна.

$$\begin{aligned}\hat{b} &= \Sigma_{21}\Sigma_{11}^{-1} = \frac{\text{cov}(X_1, X_2)}{D(X_1)} \\ \hat{a} &= \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 = E(X_2) - \hat{b}E(X_1)\end{aligned}$$

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 томъёо дараах хэлбэртэй болно.

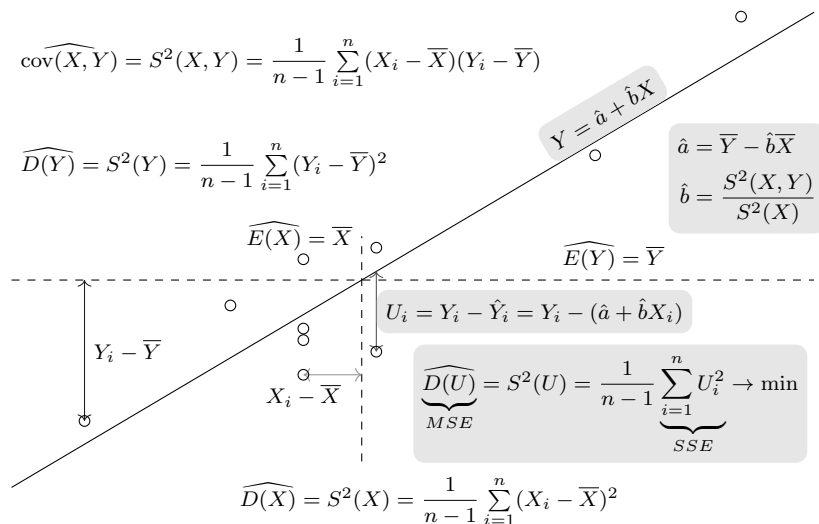
$$\hat{a} = E(Y) - \hat{b}E(X) \quad \hat{b} = \frac{\text{cov}(X, Y)}{D(X)}$$

Түүврийн хувьд дээрх томъёонд буй моментуудыг харгалзах түүврийн моментуудаар үнэлж болох тул

$$\hat{a} = \bar{Y} - \hat{b}\bar{X} \quad \hat{b} = \frac{S^2(X, Y)}{S^2(X)}$$

болно. Энд \bar{X} болон \bar{Y} нь түүврийн дундаж, $S^2(X)$ нь түүврийн дундаж квадрат хазайлт, $S^2(X, Y)$ нь түүврийн ковариацийн коэффициент юм.

Дундаж квадрат алдааг минимумчлах нь алдааны квадратуудын нийлбэрийг минимумчлахаас ялгаагүй юм. Иймд (X, Y) санамсаргүй векторын эх олонлогоос авсан $(X_1, Y_1), \dots, (X_n, Y_n)$ холбоост түүврийн хувьд $Y = a + bX + U$



Зураг 83: Регрессийн шугаман загвар

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

$$\begin{cases} \frac{\partial SSE}{\partial a} = -2 \sum_{i=1}^n (Y_i - (a + bX_i)) = 0 \\ \frac{\partial SSE}{\partial b} = -2 \sum_{i=1}^n (Y_i - (a + bX_i))X_i = 0 \end{cases}$$

систем тэгшитгэл гарна.

Тэгшитгэлүүдийг $-2n$ үржвэрт хувааж түүврийн дундаж, түүврийн дундаж квадрат хазайлт болон түүврийн ковариацийн коэффициент ашиглан хувиргаж бичвэл дараах тэгшитгэл гарна.

$$\begin{cases} \bar{Y} - a - b\bar{X} = 0 \\ \overline{X \cdot Y} - a\bar{X} - b\bar{X}^2 = 0 \end{cases} \quad \begin{cases} a = \bar{Y} - b\bar{X} \\ \underbrace{\overline{X \cdot Y} - \bar{X} \cdot \bar{Y}}_{S^2(X,Y)} - b \underbrace{(\bar{X}^2 - \bar{X}^2)}_{S^2(X)} = 0 \end{cases}$$

Ингээд үнэлэлтийн дараах томьёо гарна.

$$\begin{cases} \hat{a} = \bar{Y} - b\bar{X} \\ \hat{b} = \frac{S^2(X, Y)}{S^2(X)} \end{cases}$$

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)
```

Дээрх тушаалаар дараах үр дүн гарна.

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

- 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.

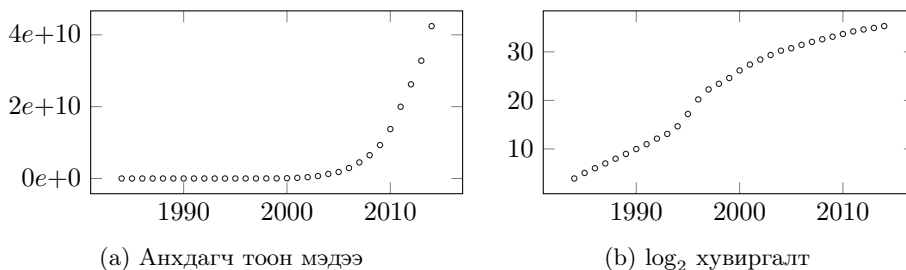
1. U_i хэвийн тархалттай.
2. U_1, \dots, U_n алдаанууд хамааралгүй.
3. $D(U_i)$ тогтмол.
4. $Y = a + b_1X_1 + \dots + b_kX_k + U$ загварын хувьд X_1, \dots, X_k тайлбарлах хувьсагчид хамааралгүй.

Лекц XVI

Хяналттай машин сургалтад ашиглах зарим статистик арга техник

1 Регрессийн шугаман загвар

Өгөгдөл ба бодлого



Зураг 84: 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

$\log_2(\text{урсгал}) = a + b \cdot \text{он загварын үнэлгээ}$

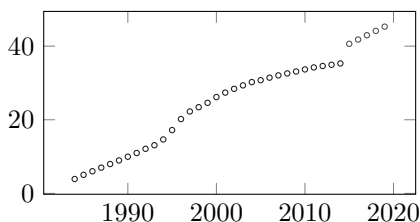
```
fit <- lm(formula = log2(traffic) ~ year, data = data)
```

Прогноз

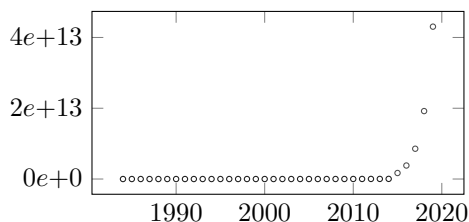
```

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



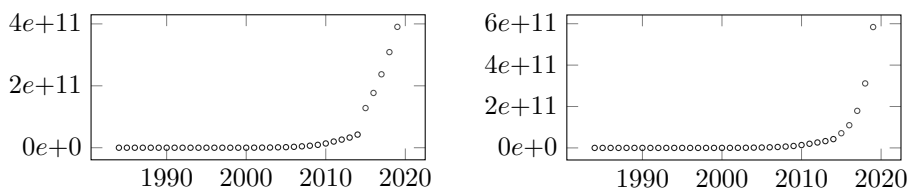
(b) хувиргалтгүй

Зураг 85: 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.



$$(a) \log_2(\text{урсгал}) = a + b \cdot \text{он} + c \cdot \text{он}^2 \quad (b) \log_2(\text{урсгал}) = a + b \cdot \text{он} + c \cdot \text{он}^2 + d \cdot \text{он}^3$$

Зураг 86: 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)
```

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.

$$X_t = b_0 + b_1 X_{t-1} + b_2 X_{t-2} + \dots + b_p X_{t-p} + U_t$$

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

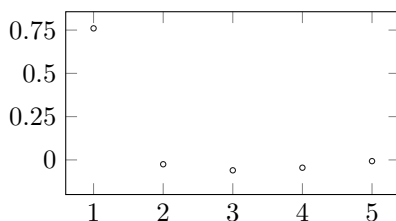
Загварын эрэмбэ тогтооход тухайн автокорреляц ашиглана.

```
| partial_autocorrelations <- acf(x = data$traffic, lag.max = 5,
| type = "partial", plot = TRUE)
| print(partial_autocorrelations)
```

Авторегрессийн загварын параметруудийг үнэлэх

Загварыг хамгийн бага квадратын аргаар яаж үнэлэхийг харуулав.

```
| fit.ar <- ar.ols(x = data$traffic, order.max = 1)
```



Зураг 87: Интернет сүлжээний урсгал хувьсагчийн автокорреляц

Ийнхүү

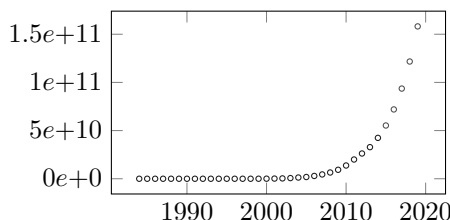
$$X_t = 1.783 \cdot 10^9 + 1.2978X_{t-1}$$

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 88: Интернет сүлжээний урсгалын хэтийн төлвийн прогноз

3 Ложистик регресс

Ложистик регресс

$Y \sim \text{Ber}(p)$ байх Y хувьсагчийн утгыг X хувьсагчийн тусламжтай прогнозолх 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. Ийнхүү 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 үзэхэд

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.

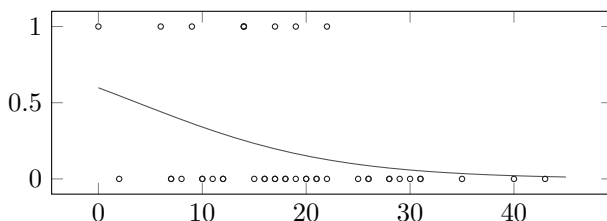
Ийнхүү эцэстээ

$$\ln\left(\frac{p}{1-p}\right) = 0.39805 - 0.10602 \cdot \text{шалгалт}$$

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 89: $p = P(W \text{ үнэлгээний санал өгөх})$ 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

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

$$\operatorname{argmax}_k P(C_k | X_1, \dots, X_p)$$

дугаар ангид хуваарилдаг. Энд C_k нь туршилтын k дугаар үр дүн эсвэл анги, X_1, \dots, X_p нь санамсаргүй хувьсагчид юм. Байесын зарчим ёсоор

$$P(C_k | X_1, \dots, X_p) = \frac{P(X_1, \dots, X_p | C_k) P(C_k)}{P(X_1, \dots, X_p)}$$

болох ба улмаар X_1, \dots, X_p хувьсагчдыг C_k үзэгдлийн нөхцөлд хамааралгүй гэсэн "гэнэхэн" таамаглалд найдаж

$$P(X_1, \dots, X_p | C_k) = P(X_1 | C_k) \cdot \dots \cdot P(X_p | C_k)$$

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)
```

Өгөгдөл цэвэрлэх

```
| tapply(X = data$input16, INDEX = data$class, FUN = var)  
| data$input16 <- NULL
```

`input16` хувьсагчийн хувьд `class` хувьсагчийн 4 гэсэн утгад харгалзах бүлгийн дундаж квадрат хазайлт тэгтэй тэнцүү байсан тул түүнийг зайлуулна.

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	206	2	0	0	0	0	2	0	21	0
1	0	152	38	1	0	11	0	2	0	0
2	0	22	204	0	0	0	0	2	1	0
3	0	16	0	211	0	0	0	0	0	0
4	2	1	0	0	213	4	0	0	8	3
5	0	1	0	21	0	109	0	0	14	57
6	0	0	0	0	2	3	196	0	6	0
7	0	25	1	0	0	2	0	192	5	8
8	7	16	1	0	0	5	0	7	175	0
9	2	5	0	6	0	0	0	0	4	206