Магадлал Статистик

2025/8/28

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

Лекц I
Санамсаргүй хувьсагч, түүний тархалт

> "Статистик сэтгэлгээ нь нэг л өдөр идэвхтэй иргэдийн хувьд уншиж бичих чадвар шиг зайлшгүй шаардлагатай болно." гэж H.G. Wells-ийн хэлсэн нь зөв бололтой! — Уилкс 1951

1 Удиртгал буюу машин сургалт дахь магадлал, статистикийн нэг хэрэглээ

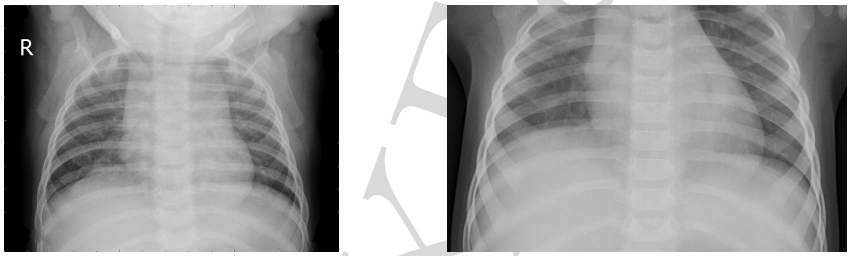
Сэдэлжүүлэх жишээ: Машин сургалт

Машин сургалт нь компьютерийн ухаан, математик болон статистик гэсэн шинжлэх ухааны салбаруудын огтлолцол дээр оршдог. Ингээд Байесын зарчим хэмээх статистикийн нэг арга барилд тулгуурласан машин сургалтын жишээ авч үзье.

Хүний цээжний рентген зургаар түүний хатгаатай юу, эрүүл үү гэдгийг таньж ялгах машин сургалтын загвар зохио.

Иймэрхүү асуудлуудыг ихэвчлэн хяналттай машин сургалтын алгоритмуудаар шийддэг. Хяналттай машин сургалт гэдэг нь тавьж буй асуудалтай уялдуулан урьдаас ангилсан өгөгдөл дээр машин сургалтын алгоритмыг хэрэглэхийг явдал юм.

Хатгаатай юу, эрүүл үү гэдгийг таних хяналттай машин сургалтын загвар зохиоход эрүүл ба хатгаатайгаар нь ангилсан рентген зургийн мэдээлэл хэрэгтэй. Үүнд зориулан www.kaggle.com/paultimothymooney/chest-xray-pneumonia веб хуудас дээрх уушгины рентген зургуудыг татан авч ашиглав. Тус зургуудыг эрүүл ба хатгаатай гэсэн оношоор нь ангилсан байна.



(a) эрүүл (b) бактерын гаралтай хатгаа

Зураг 1: Цээжний рентген зураг

Өгөгдөл бэлдэх

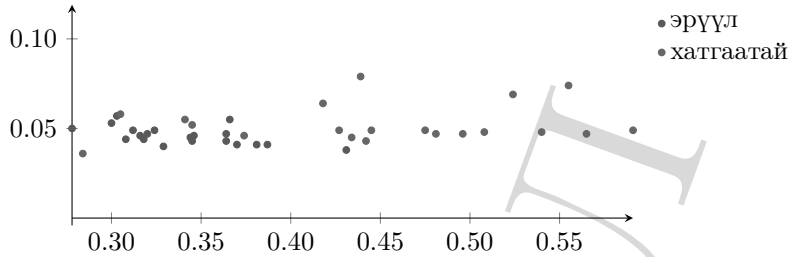
- IM-0115-0001.jpeg
- IM-0117-0001.jpeg
- person1_bacteria_1.jpeg
- person100_virus_184.jpeg

$$\underbrace{\text{хар цагаан зураг}}_{\text{матриц}} \rightarrow \text{сингуляр утгын задаргаа} \rightarrow \underbrace{\text{сингуляр утгууд}}_{\text{вектор}}$$

 $\longrightarrow$ $(0.565, 0.047, \text{bacteria})$

№	SV1	SV2	type
1	0.3664534	0.05499858	normal
21	0.5650811	0.04686256	bacteria
41	0.4338590	0.04457886	virus

Хүснэгт 1: Бэлдсэн өгөгдлийн зарим мөр



Зураг 2: Бэлдсэн өгөгдөл

Бэлдсэн өгөгдөл

Энд зөвхөн зарим зургийн мэдээллийг л ашиглав.

Приор магадлал

Байесын зарчим ашиглахад приор магадлалуудыг нь өгөх хэрэгтэй болдог. Манай бодлогын хувьд энэ нь эрүүл ба хатгаатай хүмүүсийн эзлэх хувь байх юм. Гэвч үүнийг одоогоор мэдэхгүй тул приор магадлалуудыг адил тэнцүү

$$P(\text{эрүүл}) = P(\text{хатгаатай}) = 0.5$$

гэж үзнэ. Үүний дагуу эрүүл болон хатгаатай гэсэн хоёр бүлгийн хэмжээ 50:50 харьцаатай байхаар тооцож бүлэг тус бүрээс 20 хүний цээжний рентген зураг түүвэрлэн авсан.

Үнэний хувь

Үнэний хувь олохын тулд хувьсагч бүр ямар тархалттай байхыг тогтоосон байх шаардлагатай. Гэнэн Байесын алгоритмыг практикт хэрэглэхдээ тасралтгүй хувьсагчдын хувьд хэвийн тархалтыг түгээмэл ашигладаг. Өөрөөр хэлбэл бэлдсэн өгөгдөл дэх SV1, SV2 хувьсагчдыг хэвийн тархалттай гэж үзнэ.

Хэвийн тархалтын параметрүүдийг бүлэг тус бүрээр, бэлдсэн өгөгдөлдөө тулгуурлан үнэлбэл дундаж нь

$$\hat{\mu}_{\text{эрүүл}} \approx (0.340, 0.046) \text{ ба } \hat{\mu}_{\text{хатгаатай}} \approx (0.449, 0.053)$$

харин дисперс нь

$$\hat{\sigma}^2_{\text{эрүүл}} \approx (0.001, 2 \cdot 10^{-5}) \text{ ба } \hat{\sigma}^2_{\text{хатгаатай}} \approx (0.007, 10^{-4})$$

байна.

Постериор магадлал буюу машинаар оношлогоо

Байесын зарчмаар тухайн нэг хүнийг рентген зургаар нь оношлоно гэвэл

$$P(\text{эрүүл}|\text{зураг}) \qquad \text{ба} \qquad P(\text{хатгаатай}|\text{зураг})$$

постериор магадлалуудыг олж аль нь их вэ гэдгийг ажиглана. Ингээд хэрэв эрүүл гэсэн постериор магадлал их байвал эрүүл гэсэн оношлогоо тавина. Энд рентген зургийг түүнд харгалзах (SV1, SV2) вектор төлөөлөх буюу тооцооллыг эдгээр тоон утгууд дээр хийнэ.

$P(\text{эрүүл}|\text{SV1}, \text{SV2})$ ба $P(\text{хатгаатай}|\text{SV1}, \text{SV2})$ постериор магадлалуудыг жишихэд зөвхөн үнэний хувь болон приор магадлал хоёрын үржвэрийг л эрүүл ба хатгаатай гэсэн хоёр тохиолдолд олоход хангалттай. Цаашилбал приор магадлалуудыг тэнцүү гэж тооцсон тул ердөө үнэний хувийг л тооцоолох шаардлагатай байна. Түүнчлэн үнэний хувийг олохдоо өмнө дурдсанчлан SV1, SV2 хувьсагчдыг хэвийн тархалттай гэж үзнэ.

Жишээлэн нэг хүний хувьд $(\text{SV1}, \text{SV2}) = (0.401, 0.049)$ байсан гэе. Тэгвэл

$$\begin{aligned} P(\text{эрүүл}|\text{SV1}, \text{SV2}) &\sim P(\text{SV1}|\text{эрүүл})P(\text{SV2}|\text{эрүүл}) \\ &= \Phi_{\text{эрүүл}}(\text{SV1}) \cdot \Phi_{\text{эрүүл}}(\text{SV2}) \\ &= \Phi_{\text{эрүүл}}(0.401) \cdot \Phi_{\text{эрүүл}}(0.049) \\ &\approx 0.973 \cdot 0.749 \approx 0.729 \end{aligned}$$

ба үүнтэй төстэйгээр

$$P(\text{хатгаатай}|\text{SV1}, \text{SV2}) \sim 0.283 \cdot 0.344 \approx 0.097$$

буюу эрүүл гэсэн постериор магадлал их гарсан тул тус хүнд машинаар тавих оношь нь "эрүүл" байх юм.

2 Яагаад магадлал, статистик гэж?

Яагаад магадлал, статистик гэж?

Детерминистик бус өгөгдлөөс прогноз гаргах, предикт хийхэд статистик загвар ашигладаг. Ерөнхийдөө детерминистик гэж тодорхой байдлыг хэлдэг. Харин үүний эсрэг нь стохастик буюу санамсаргүй байдал билээ.

Сэдэлжүүлэх жишээ дэх өгөгдөл бол санамсаргүй өгөгдөл юм. Яагаад гэвэл тухайн нэг зургийг тоо руу хувиргахад гарах утга нь түүний эрүүл ба хатгаатай аль бүлэгт харьяалагдахыг тооцсон ч тогтмол биш буюу хувьсаж байна. Цаашилбал хувьсах хувьсахдаа урьдчилан тооцоолохын аргагүй буюу санамсаргүй байна.

Харин шинжлэх ухаанд санамсаргүй өөрчлөгддөг зүйлсийн хувьсах зүй тогтлыг судалдаг суурь онол нь магадлалын онол юм. Статистик, стохастик процессийн онол, мэдээллийн онол, квант механик зэрэг нь магадлалын онол дээр суурилан хөгжсөн.

Магадлалын онолд судлагдахууны хувьсах зүй тогтлыг яг л "бурхан" мэт нэгд нэггүй мэднэ гэж үздэг. Жишээлбэл зоос орхиход сүлд буух боломжийг 50 хувь гэж тооцдог. Гэтэл заримдаа тэгж нэгд нэггүй хэлэхэд бэрх байдаг. Ийм үед судлагдахууны зарим хэсгийг түүвэрлэн авч өгөгдөл бүрдүүлээд улмаар түүнээсээ мөнөөх зүй тогтлыг олдог. Сонирхуулбал 1894 онд Английн математикч, статистикч Карл Пирсон оюутнуудаараа 24000 удаа зоос орхих туршилт хийлгэхэд 12012 удаад нь сүлд буужээ. Үүн шиг түүврийн аргаар цуглуулсан өгөгдлийг ашиглан судлагдахууны хувьсах зүй тогтлын талаар дүгнэлт гаргадаг шинжлэх ухааны чиглэлийг статистик гэнэ.

Сэдэлжүүлэх жишээнд авч ашигласан рентген зургууд бол түүврийн аргаар цуглуулсан өгөгдөл билээ. Ингээд оношь тавихыг автоматжуулах ажлыг статистик өгөгдөлд тулгуурлан хийх болсон тул үүнд статистикийн загвар ашиглах нь гарцаагүй.

Эцэст нь магадлалын онол ба статистик хоёрыг шатраар зүйрлүүлбэл магадлалын онол нь ноён, статистик нь бэрс болно. Ноёнгүй бэрс байхгүй. Өөрөөр хэлбэл статистикаас магадлалын онолыг салгах аргагүй. Мөн үүн шиг машин сургалт, өгөгдлийн ухаан зэргээс статистикийг салгах арга үгүй.

Детерминизм буюу бас ахин яагаад магадлал, статистик гэж?

Яг үнэндээ санамсаргүй зүйл гэж үгүй. Ердөө л бид хэтийн төлвийг нь урьдчилан яг таг тооцоолж чадахгүй байгаа юм.

Дээрх санаа нь шинжлэх ухаан болон философид өргөн хэлцэгдсэн детерминизмийн гол агуулгатай нийцдэг бөгөөд энэ талаар алдар суутнууд ч үгээ хэлжээ.

- Пьер-Симон Лаплас (1776) Хэрэв орон зай дахь бүх биеийн байрлал, хөдөлгөөний тухай бүрэн мэдлэг бидэнд байсан бол ирээдүй, өнгөрсөнийг төгс тооцоолох боломжтой.
- Альберт Эйнштейн (1926) Квант механикийн магадлалын тайлбарт эргэлзэн, "Бурхан шоо хаядаггүй" (God does not play dice) хэмээн тэмдэглэсэн нь байгаль дахь санамсаргүй мэт үзэгдэл ч далд хууль дүрмээр удирдагдана гэсэн итгэл үнэмшлийг харуулдаг.
- Стивен Хокинг (1988) Бидний үйлдэл детерминистик хуульд захирагдлаа ч бид түүнийг урьдчилан тооцоолж чадахгүй. Тиймээс л бидэнд чөлөөт сонголттой мэт санагддаг.

3 Магадлалын онол байгуулах суурь

Нэр томьёо

Эх олонлог Судалгаанд хамрагдах бүх объектуудын олонлог

Хувьсагч Эх олонлогийн элементүүдийн шинж чанарыг илэрхийлэгч Хувьсагчийг авах утгаар нь *тоон* ба *чанарын* гэж ангилдаг.

Тархалт Эх олонлогт хувьсагчийн утга тус бүр хэдэн хувийг эзэлж буйг хувьсагчийн эх олонлог дахь тархалт буюу товчоор тархалт гэнэ.

Параметр Хувьсагчийн тархалтын шинж чанарыг тодорхойлох тоо

Туршилт Эх олонлогийн элементийн шинж чанарыг ажиглах

Үзэгдэл Туршилтаас гарах үр дүн Үзэгдэл нь өөр үзэгдлүүдэд задрахгүй бол түүнийг *энгэл үзэгдэл* гэнэ. Магадлалын онолд санамсаргүй үр дүнтэй туршилтыг авч үздэг. Өөрөөр хэлбэл энд санамсаргүй үзэгдэл яригдаж байна.

Магадлал Үзэгдлийн ажиглагдах боломжийг илэрхийлэх сөрөг бус тоо

(үзэгдэл, хувьсагч, магадлал) гурвал

Үзэгдэл нь эх олонлогийн тухайн нэг элементийн шинж чанарыг нэг янзаар илэрхийлэх тул үүнийг боловсруулан хувьсагч хэлбэрт буулгаж болно.

$$X = X(\omega)$$

Энд ω нь үзэгдэл, X нь хувьсагч юм. Түүнчлэн ω нь санамсаргүй үзэгдэл байх тул X хувьсагчийн утга мөн санамсаргүй байна.

Үзэгдлийг хувьсагч руу буулгасан тул үзэгдлийн ажиглагдах боломжийг илэрхийлдэг магадлал ч хувьсагчтай харгалзана.

$$P(X = x) = P(X(\omega) = x) = P(\omega) = p$$

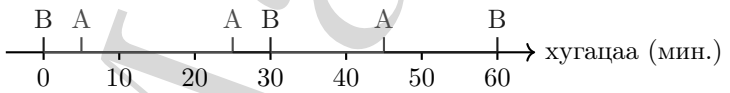
Энд P нь магадлалын хэмжээс, x нь ω үзэгдэл дээрх X хувьсагчийн утга, p нь ω үзэгдлийн магадлал юм.

Хялбар жишээ

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 2/3 хувийг А програм боловсруулж байгааг ажиглажээ. Энэ зөв үү?



Зураг 3: Жишээ бодлогын зураглал

Санамсаргүй үзэгдэл

Тодорхойлолт 1. Туршилтаар ажиглагдах бүх эгэл үзэгдлийг хамтат нь *эгэл үзэгдлийн огторгуй* гэж нэрлээд Ω үсгээр тэмдэглэн бичнэ.

Жишээний хувьд дараах эгэл үзэгдлүүд байна.

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

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

Эгэл үзэгдлүүд нийлэн янз бүрийн үзэгдэл болно.

$\{\text{хүсэлтийг В програм боловсруулах}\} = \{0\} \cup (25, 30] \cup (45, 60]$

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

Эгэл үзэгдлүүдээр зохиож болох бүх үзэгдлийг хамтат нь $\mathcal{F}$ гэе.

Үзэгдлийн магадлал

Тодорхойлолт 2. $\mathcal{F}$ нь үзэгдлүүдийн олонлог байг. Тэгвэл

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

функцийг *магадлал* гэнэ.

P функцийг хэрэв Ω олонлог

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

гэх мэтчилэн янз бүрийн аргаар тодорхойлдог.

Жишээний хувьд P функцийг дараах байдлаар тодорхойлж болно.

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

Нэг үр дүнгийн магадлал

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

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

байна. Гэвч 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 аксиом

Тодорхойлолт 3. $(\Omega, \mathcal{F}, P)$ бүлийг *магадлалын огторгуй* гэнэ.

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

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

4 Санамсаргүй хувьсагч

Санамсаргүй хувьсагч

Тодорхойлолт 4. Эгэл үзэгдлийн огторгуй дээр тодорхойлсон, бодит тоон утгатай $X : \Omega \to \mathbb{R}$ функцийг *санамсаргүй хувьсагч* гэнэ.

Жишээний хувьд X нь хүсэлт боловсруулагдах хүртэл хүлээх 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.

Судалгааны зорилгоос хамааран эх олонлог дээр янз бүрийн санамсаргүй хувьсагч авч үзэж болдог.

- Y нь серверт хүсэлт ирэх үеийн цаг хугацааны момент бол

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

- Хүсэлтийг А 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.

Яагаад санамсаргүй хувьсагч гэж?

Санамсаргүй хувьсагч нь тоон утгатай билээ. Иймээс судалгаанд тооны ухаан буюу математикийн арга техник хэрэглэх нөхцөл бүрдэнэ.

Санамсаргүй хувьсагчийн ангилал

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

5 Санамсаргүй хувьсагчийн тархалт

Дискрет санамсаргүй хувьсагчийн тархалт

Тодорхойлолт 5. X санамсаргүй хувьсагч дискрет буюу

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

үед

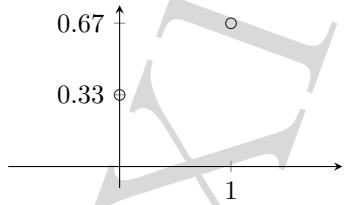
$$f_X(x) = \begin{cases} P(X = x_i), & x = x_i \\ 0, & x \neq x_i \end{cases}$$

функцийг тус *дискрет санамсаргүй хувьсагчийн магадлалын нягтын (массын) функц* гэнэ.

I_A дискрет санамсаргүй хувьсагчийн утгуудын магадлал

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

тул магадлалын массын функц нь дараах хэлбэртэй байна.



Зураг 4: I_A дискрет санамсаргүй хувьсагчийн магадлалын нягтын (массын) функц

$$f_{I_A}(x) = \begin{cases} \frac{1}{3}, & x = 0 \\ \frac{2}{3}, & x = 1 \\ 0, & \text{бусад} \end{cases}$$

Нийлмэл үзэгдлийн магадлал олох

X санамсаргүй хувьсагч дискрет буюу $X \in \{x_1, x_2, \ldots\}$ бол $\{a \leq X \leq b\}$ үзэгдлийн магадлалыг

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

гэж олно.

Тасралтгүй санамсаргүй хувьсагчийн тархалт

X нь тасралтгүй санамсаргүй хувьсагч бол $P(X = x_i) = 0$ буюу туршилтын нэг үр дүнгийн магадлал тэгтэй тэнцүү тул $\{a \leq X \leq b\}$ нийлмэл үзэгдлийн магадлалыг утга нэг бүрийн магадлалуудын ердийн нийлбэрээр олох нь утгагүй. Харин үүний оронд интеграл тооллын аргыг ашиглан $[a, b]$ интервал дахь x утгуудын магадлалыг нийлбэрчилдэг. Энэ дашрамд тасралтгүй санамсаргүй хувьсагчийн тархалтыг илэрхийлэх нягтын функцийн тодорхойлолт гардаг.

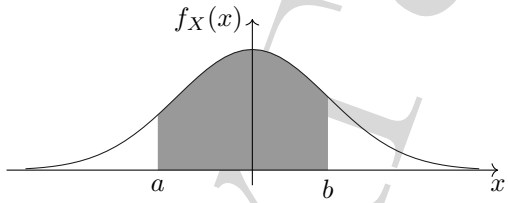
Тодорхойлолт 6. X нь тасралтгүй санамсаргүй хувьсагч бөгөөд хэрэв $a \leq b$ бүрийн хувьд

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

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

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

Иймд $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ байх нь илэрхий юм.



Зураг 5: X тасралтгүй санамсаргүй хувьсагчийн магадлалын нягтын функцийн график ба $a \leq X \leq b$ үзэгдлийн магадлалыг хэвийн тархалт гэдэг тархалтаар жишээлсэн дүрслэл

Дискрет ба тасралтгүй хувьсагчдын тархалтын ялгаа

> ⚠ Тасралтгүй санамсаргүй хувьсагчийн хувьд
>
> $$f_X(x) \neq P(X = x)$$
>
> байна.

Үнэндээ

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

байна. Иймээс тасралтгүй санамсаргүй хувьсагчийн хувьд

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

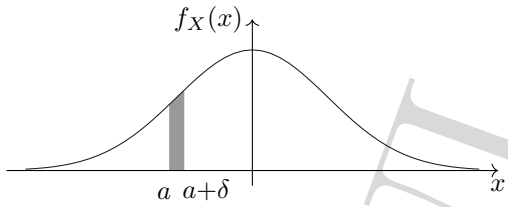
гэж үзнэ.

Тасралтгүй санамсаргүй хувьсагчийн $P(X = a)$ магадлал

Хэрэв $f_X(x)$ функц a цэг дээр тасралтгүй бол өчүүхэн 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 6: X тасралтгүй санамсаргүй хувьсагчийн магадлалын нягтын функцийн график ба $a < X < a + \delta$ үзэгдлийн магадлалыг хэвийн тархалт гэдэг тархалтаар жишээлсэн дүрслэл

Жигд тархалт

Санамсаргүй хувьсагчийн утгуудын ажиглагдах боломж ижил бол уг тархалтыг *жигд тархалт* гэнэ. Жишээний хувьд $Y \sim U(0, 60)$ буюу Y хувьсагч

(a) Тасралгүй жигд тархалтын магадлалын нягтын функцийн график

(b) Дискрет жигд тархалтын магадлалын нягтын (массын) функцийн график

Зураг 7: Жигд тархалтын ерөнхий дүр төрх

$(0, 60)$ завсарт жигд тархалттай.

Лекц II
Дундаж болон дундаж квадрат хазайлт

> Магадлал бол хэсэгхэн мэдлэгээс төрсөн хүлээлт юм. Үзэгдэл явагдахад нөлөөлж болох бүх нөхцөл байдалтай төгс танилцах нь хүлээлтийг баттай болгож, магадлалын онолын хэрэгцээ шаардлага ба орон зайг үгүй болгодог. — Жорж Бүүл

1 Хамааралгүйн чанар

Статистикийн тусламжтай хариулах зарим асуултууд

Санамсаргүй хувьсдаг шинж чанар бүхий юмсийг судлахад статистик хэрэг болдог билээ. Хэрэв нэг л юм судалж байгаа бол түүний шинж чанарыг

илэрхийлэх санамсаргүй хувьсагчийн тархалтыг олоход хангалттай. Мөн тухайн шинж чанар үнэхээр санамсаргүй юу гэдэг ч статистикийн нэг асуудал болно. Харин олон юм авч үзэхэд дараах хоёр асуулт гардаг.

1. Эдгээр юмс ижил шинж чанартай юу?
2. Юмсийн нэг нь нөгөөгийнхөө шинж чанарт нөлөөлдөг үү? Нөлөөтэй бол хэрхэн нөлөөлдөг вэ?

Энд олон юмс яригдаж буй тул мэдээж олон хувьсагч авч үзнэ. Тэгэхээр дээрх асуултуудын эхнийх нь хувьсагчид ижил тархалттай юу эсвэл тархалтуудынхаа зарим параметрээр ижил үү гэсэн утгатай болно. Харин хоёр дахь асуултыг хувьсагчид хамааралтай юу, хамаарал нь хэр хүчтэй вэ бас хамаарал нь ямар зүй тогтолтой вэ гэж тавина. Одоо хамааралтай юу гэсэн асуултыг тархалтын зүгээс авч үзсэнтэй танилцана.

Хамааралгүйн чанар

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

бол X болон Y санамсаргүй хувьсагчдыг *хамааралгүй* гэнэ. Хувьсагчид дискрет үед үүнийг $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$ гэж томьёолсон ч болно. Санамсаргүй хувьсагч нь үзэгдэлтэй холбогдох тул хамаарлын талаарх ойлголт үзэгдэлд ч хамаатай. Ийнхүү

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

бол эдгээр үзэгдлүүдийг хамааралгүй гэнэ.

Энд буй $f_{X,Y}(x,y)$ нь X ба Y хувьсагчдын хамтын нягтын функц бөгөөд хамтын тархалт болон санамсаргүй хувьсагчдын хамаарлын талаар дараа дэлгэрэнгүй авч үзнэ.

2 Пуассоны тархалт

Хамааралгүй туршилтын дараалал

Ямар нэг туршилт явуулж байгаа гэж үзье. Туршилтаас гарах үр дүнгүүд буюу үзэгдлүүдийн аль нэгийг "амжилт" хэмээн онцлоод A гэж тэмдэглэнэ. Ийнхүү үр дүнг нь хоёр ангилсан туршилтыг *Бернуллийн туршилт* гэдэг. $P(A)$ буюу амжилтын магадлалыг p бас туршилтыг давтан явуулах тоог n гэж тэмдэглэнэ. Мөн туршилтын хоёр өөр давталтын үр дүнгүүд хамааралгүй байг. Тэгвэл үүнтэй холбогдуулан практикт өргөн тохиолдох янз бүрийн санамсаргүй хувьсагч авч үзэж болдог.

Пуассоны тархалт

Нэгж хугацаанд төгсгөлгүй олон "амжилт" гарах боломжтой туршилтын "амжилт"-ын тоо гэсэн дискрет санамсаргүй хувьсагчийн нягтын функц

$$f_X(x) = \begin{cases} \frac{\lambda^x}{x!}e^{-\lambda}, & x \in \{0,1,2,\dots\} \\ 0, & \text{бусад} \end{cases}$$

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

Пуассоны тархалтын тусламжтай шийдэх бодлого

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

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

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

байна.

Пуассоны тархалт ба бином тархалт

Эдгээр тархалтуудын хувьд санамсаргүй хувьсагч нь адилхан бөгөөд туршилтаар илрэх амжилтын тоог илэрхийлдэг. Харин туршилтын тоо талаас зарчмын ялгаатай.

Пуассоны тархалтын хувьд тодорхой хугацаанд төгсгөлгүй олон амжилт гарах боломжтой туршилт авч үзсэн. Төгсгөлгүй олон амжилт гаргахын тулд туршилтыг хязгааргүй олон удаа давтах болно. Ийнхүү туршилтын тоо хязгааргүй их буюу $n \to \infty$ болно. Гэтэл бином тархалтын хувьд туршилтын тоо төгсгөлөг байдаг.

Бином тархалтыг ЕБС-ийн математикийн хичээлээр үздэг ба үүний нягтын функц нь

$$f_X(x) = \begin{cases} C_n^x p^x (1-p)^{n-x}, & x \in \{0, 1, 2, \ldots, n\} \\ 0, & \text{бусад} \end{cases}$$

хэлбэртэй байдаг билээ.

Пуассоны тархалтын нягтын гаргалгаа

Дараах бодолтын үеэр $\lim\limits_{n\to\infty} np$ хязгаар оршин байхыг шаардахаас өөр аргагүй болно. Үүний тулд $p \to 0$ буюу амжилт нь ховор тохиолддог үзэгдэл байх хэрэгтэй. Ийнхүү $\lambda_n = np$, $\lim\limits_{n\to\infty} \lambda_n = \lambda$ гэе.

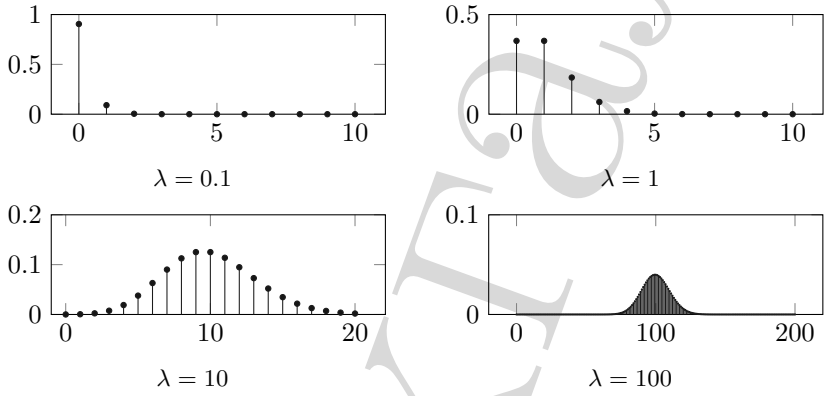
$$\begin{aligned} f_X(x) &= P(n \text{ туршилтад } x \text{ удаа амжилт илрэх}) \\ &= C_n^x p^x (1-p)^{n-x} \qquad \text{бином тархалтын нягт} \\ &= \frac{n(n-1)\ldots(n-x+1)}{x!} \left(\frac{\lambda_n}{n}\right)^x \left(1 - \frac{\lambda_n}{n}\right)^{n-x} \end{aligned}$$

[1] Poisson distribution

$$= \frac{\lambda_n^x}{x!}\left(1+\frac{-\lambda_n}{n}\right)^n 1\left(1-\frac{1}{n}\right)\dots\left(1-\frac{x-1}{n}\right)\left(1-\frac{\lambda_n}{n}\right)^{-x}$$

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

Пуассоны тархалт, параметрийн янз бүрийн утгад



Зураг 8: Пуассоны тархалт, параметрийн янз бүрийн утгад

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

Пуассоны томьёо

Бином тархалтын хувьд туршилтын тоо n ихсэхэд нягтыг нь тооцоолоход бэрхшээлтэй болдог. Иймд үед хэрэв амжилтын магадлал бага бол бином тархалтын нягтыг Пуассоны тархалтыг ашиглан ойролцоолон тооцоолж болох нь Пуассоны тархалтын нягтын гаргалгаанаас харагдана. Ийнхүү бином тархалтын нягтыг n их, p бага үед

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

гэж ойролцоо бодно. Үүнийг *Пуассоны томьёо* гэдэг. Харин амжилтын магадлал p нь эсрэгээрээ их үед амжилтын эсрэг үзэгдлийг амжилт гэж үзээд тус томьёог хэрэглэнэ.

3 Дундаж

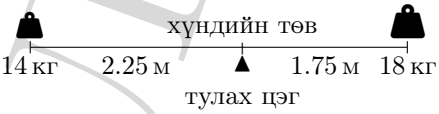
Санамсаргүй хувьсагчийн математик дундаж

Тодорхойлолт 7. Санамсаргүй хувьсагчийн утгуудыг тэдгээрийн нягтаар жинлэсэн дунджийг *математик дундаж*[2] гэнэ.

[2]expectation

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

$$-2 \cdot \tfrac{14}{32} + 2 \cdot \tfrac{18}{32} \neq 0$$



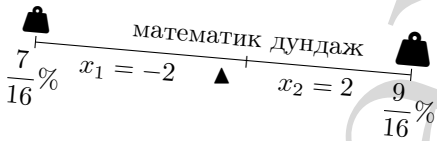
$$-2.25 \cdot \tfrac{14}{32} + 1.75 \cdot \tfrac{18}{32} = 0$$

Энд хөшүүргийн хүчний мөрийг тулах цэгээс хэмжсэн.

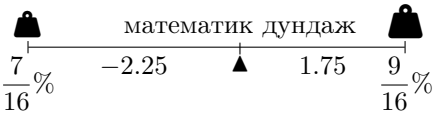
Дискрет санамсаргүй хувьсагчийн математик дундаж

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

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



$$E(X) = -2 \cdot \frac{7}{16} + 2 \cdot \frac{9}{16} = \frac{1}{4}$$



$$E(X) = -2.25 \cdot \frac{7}{16} + 1.75 \cdot \frac{9}{16} = 0$$

Энд ▲ нь координатын эхийг төлөөлнө.

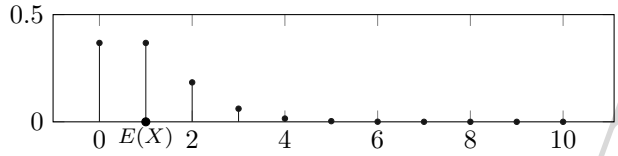
Пуассоны тархалттай санамсаргүй хувьсагчийн математик дунджийг ол.

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

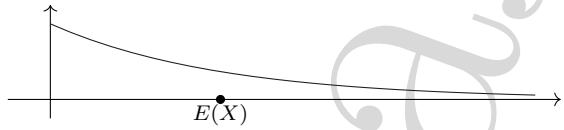
Тасралтгүй санамсаргүй хувьсагчийн математик дундаж

Тасралтгүй санамсаргүй хувьсагчийн математик дунджийг дараах томъёогоор олно.

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



Зураг 9: Пуассоны тархалттай санамсаргүй хувьсагчийн математик дундаж, $\lambda = 1$ үед



Зураг 10: Тасралтгүй хувьсагчийн математик дундаж, илтгэгч тархалтын тохиолдолд

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.

$\lambda > 0$ параметр бүхий илтгэгч тархалт буюу

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

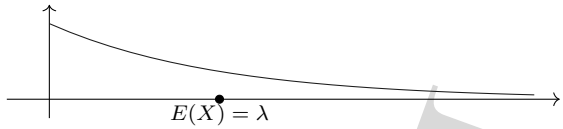
нягттай X санамсаргүй хувьсагчийн математик дунджийг ол.

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

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 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

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



Зураг 11: $\lambda = 1$ параметр бүхий илтгэгч тархалтын нягтын муруй ба математик дундаж

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.

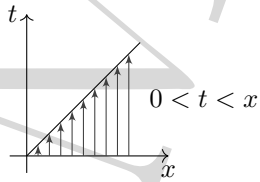
$$E(X) = \int_0^\infty P(X \geq t)\, dt$$

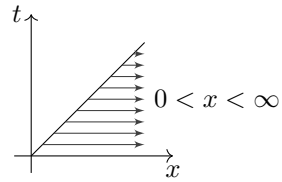
Илтгэгч тархалттай санамсаргүй хувьсагчийн хувьд дээрх чанар биелж буйг харуул.

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

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

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

□

Чанар 3 (Ухамсаргүй статистикчийн хууль[3]). $g : \mathbb{R} \to \mathbb{R}$ функцийн хувьд

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

байна.

Үүний 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

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

Пуассоны тархалттай санамсаргүй хувьсагчийн момент үүсгэгч функцийг ол.

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

[3]The law of the unconscious statistician

Стандарт хэвийн тархалттай санамсаргүй хувьсагчийн момент үүсгэгч функцийг ол.

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

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

гэж олдоно.

Момент үүсгэгч функцийн хэрэглээ

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

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

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

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

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

болно.

4 Геометр тархалт

Геометр тархалт

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

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

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

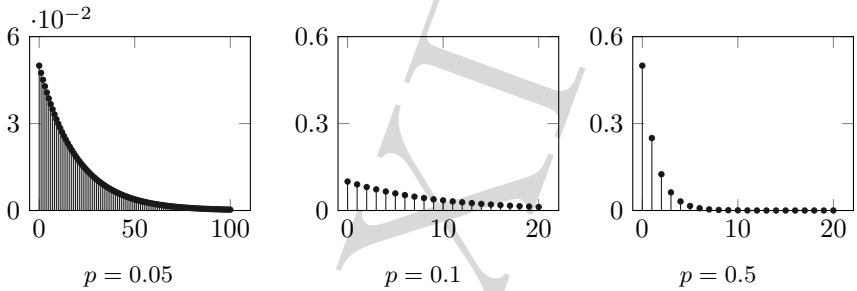
Үүнээс гадна санамсаргүй хувьсагчийг амжилт илэртэл явуулах туршилтын тоо буюу $\{1, 2, \dots\}$ утгатайгаар авах явдал бий. Өөрөөр хэлбэл геометр тархалтыг хоёр янзаар авч үздэг.

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

Геометр тархалтын нягтын гаргалгаа

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

Геометр тархалт, параметрийн янз бүрийн утгад



Зураг 12: 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 8. 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 болон Y хамааралгүй үед $D(X + Y) = D(X) + D(Y)$

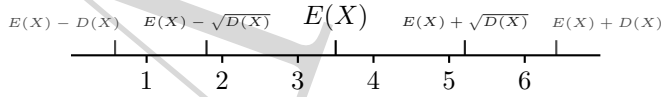
Стандарт хазайлт

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

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

$f_X(x) = \frac{1}{6}$, $x \in \{1, 2, \ldots, 6\}$ 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 13: Жишээ бодлогын бодолтын үр дүн

Чебышевийн тэнцэл биш

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

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

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

байна.

[a]2021 оны судалгаагаар тогтоосон монгол эмэгтэйчүүдийн дундаж наслалт

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

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

□

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

Наслалтын стандарт хазайлт 13.5 жил байг.

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

Лекц III

Тархалтын функц

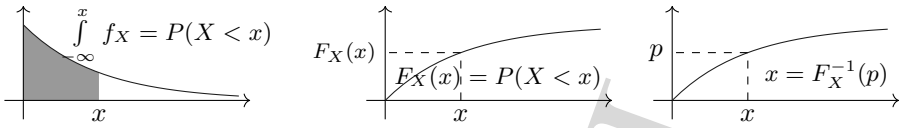
Статистик бол шинжлэх ухааны судалгааны арга зүйн үндсэн элемент юм. — Карл Пирсон

1 Тархалтын функц

Санамсаргүй хувьсагчийн тархалт, түүнийг илэрхийлэх функцүүд

Тархалтын функцийн чанарууд

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

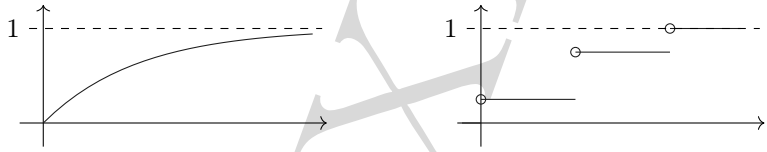


(a) Нягтын функц (b) Тархалтын функц (c) Квантил

Зураг 14: Тасралтгүй санамсаргүй хувьсагчийн тархалт, илтгэгч тархалтын тохиолдолд

(a) Нягт (b) Тархалтын функц

Зураг 15: Дискрет санамсаргүй хувьсагчийн тархалт, бином тархалтын тохиолдолд

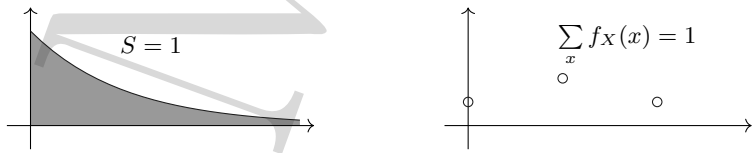


Тасралтгүй санамсаргүй хувьсагч Дискрет санамсаргүй хувьсагч

Зураг 16: 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 17: Нягтын нийлбэр

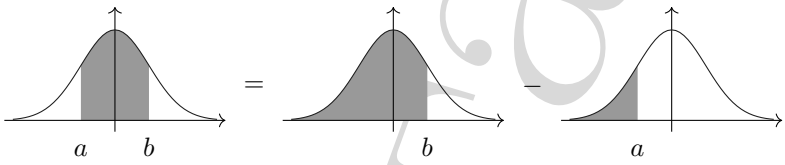
Тасралтгүй санамсаргүй хувьсагчийн тархалтын функц болон нягтын функцийн холбоо, үзэгдлийн магадлал

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

Туршилтын тархалтын функц

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

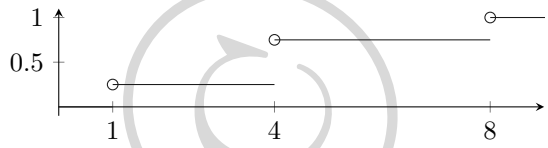
$$F_n(x) = \frac{\mu_n(x)}{n}, \quad x \in \mathbb{R}, \quad \mu_n(x) = |\{i : X_i < x,\, i = 1, \ldots, n\}|$$

функцийг *туршилтын тархалтын функц* гэдэг.

Туршилтын тархалтын функц олох

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

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



Зураг 19: Туршилтын тархалтын функцийн график

Туршилтын тархалтын функцийн чанар

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

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

чанар биелнэ.

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

2 Дунджийн чанар

Ухамсаргүй статистикчийн хуулийн баталгаа

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

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

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

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

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

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

- $f_Y(y)$:

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

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

3 Илтгэгч тархалт

Илтгэгч тархалт

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

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

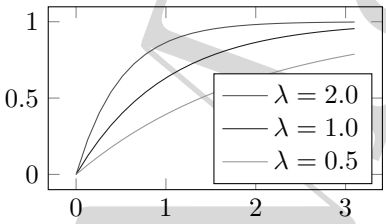
Илтгэгч тархалтын функцийн гаргалгаа

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

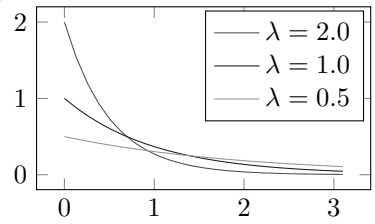
$$\begin{aligned} F_X(x) &= P(X < x) = P(x \text{ хугацааны дотор } A \text{ үзэгдэл явагдах}) \\ &= 1 - P(X \geq x) \\ &= 1 - P(x \text{ хугацааны дотор } A \text{ үзэгдэл явагдахгүй байх}) \\ &= 1 - P(\mu_n = 0, n \to \infty, p \to 0) \\ &= 1 - \frac{(\lambda x)^0}{0!} e^{-\lambda x} \quad \text{Пуассоны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

Илтгэгч тархалт, параметрийн янз бүрийн утгад

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



(a) Тархалтын функц



(b) Нягтын функц

Зураг 20: Илтгэгч тархалт, параметрийн янз бүрийн утгад

Илтгэгч тархалтын зарим чанар

1. Геометр тархалтын тасралтгүйн аналог
2. $X_i \sim \text{Exp}(\lambda_i)$[5] $(i = 1, \ldots, n)$ ба *хамтдаа хамааралгүй*[6] бол

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

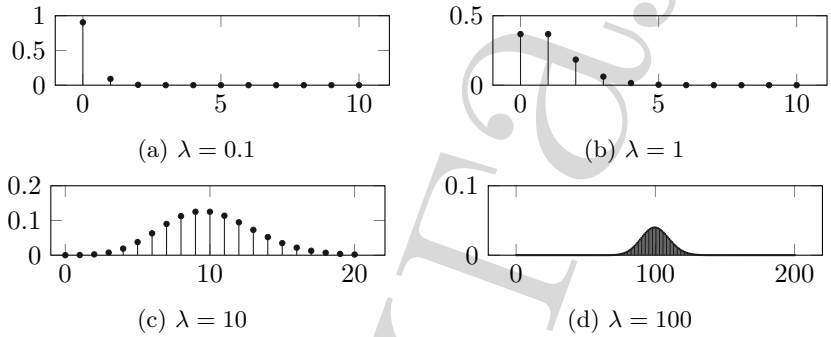
[5] илтгэгч тархалт
[6] энэ талаар хожим үзнэ

3. $\forall x, y \geq 0$ бүрийн хувьд

$$P(X \geq x + y | X \geq y) = P(X \geq x)^7$$

4 Хэвийн тархалт

Пуассоны тархалт параметрийн их утгад



(a) $\lambda = 0.1$ (b) $\lambda = 1$

(c) $\lambda = 10$ (d) $\lambda = 100$

Зураг 21: Пуассоны тархалтын параметр ба нягтын хэлбэр

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 $\ln[(1+\delta)^{\lambda(1+\delta)+1/2}] = [\lambda(1+\delta)+1/2]\ln(1+\delta) = (\lambda+1/2+\lambda\delta)(\delta - \delta^2/2 + O(\delta^3)) \approx \lambda\delta + \lambda\delta^2/2 + O(\delta^3)$ 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.

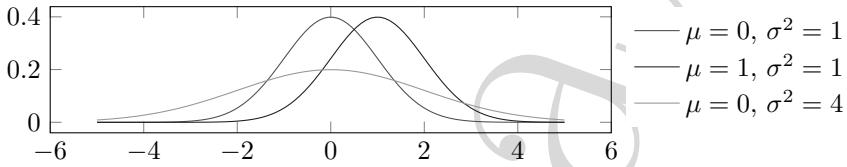
[7] илтгэгч тархалтын санамжгүй байдал; энэ талаар хожим үзнэ

"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 22: Хэвийн тархалтын нягтын муруй параметрийн янз бүрийн утгад

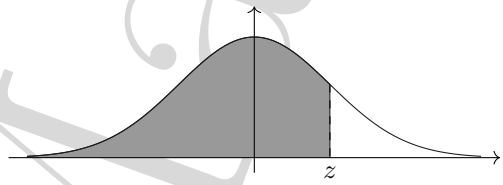
тархалтыг бас Гауссын тархалт ч гэдэг. X хувьсагч хэвийн тархалттай гэхийг $X \sim N(\mu, \sigma^2)$ гэж тэмдэглэнэ.

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 23: $\Phi(z)$ 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

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

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

$$Z = \frac{X - \mu}{\sigma}^{8} \sim N(0, 1)$$

болно.

[8] үүнийг стандарт хувиргалт гэнэ

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 байдаг[9]. Жишээлбэл $X_1, \ldots, X_n \sim \text{Ber}(p)$[10] ба хамааралгүй үед $X = X_1 + \ldots + X_n \sim B(n,p)$[11] байдаг тул

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

буюу

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

байна.

Туршилтын тоо их үед бином тархалтын нягтыг ойролцоо бодох

Бином тархалтын нягтыг n их, p бага үед

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

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

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

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

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

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

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

[9] Хязгаарын гол теорем гэдэг нэрээр дараа үзнэ.
[10] Бернуллийн тархалт
[11] бином тархалт

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: $x = \underline{\text{хот явах оршин суугчдын тоо}}$ бөгөөд энэ нь рейлбусын суудлын тоотой холбогдоно. Мөн энэ нь мэдэгдэхгүй буюу манай олох зүйл байна. • Ийнхүү $X = \underline{\text{хот явах оршин суугчдын тоо}}$ гэсэн санамсаргүй хувьсагч авч үзнэ.

Сард (30 хоног) нэгээс ихгүй удаа зорчигчид багтахгүй байх

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

Бином тархалтаар халз бодох гэвэл

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

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

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

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

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.

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

Лекц IV
Наслалтын тархалт

Амьдрал бол магадлалын сургууль юм. — Волтер Багехот

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 гэж ойлгодог. Иймд уг магадлалыг

(a) $P(AB)$ магадлал (b) $P(B|A)$ нөхцөлт магадлал

Зураг 24: A үзэгдлээс хамаарсан B үзэгдлийн нөхцөлт магадлал

олохдоо B үзэгдлийг зөвхөн A үзэгдлийн хүрээнд авч үзэх юм уу

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

томьёо ашигладаг.

Нөхцөлт магадлалтай холбогдох зарим санамж

⚠
1. $P(B|A)$ бол нөхцөлт магадлал харин $P(B \setminus A)$ бол $P(B - A)$ буюу үзэгдлүүдийн ялгавар юм.
2. $P(\underbrace{C+D}_{\text{үзэгдэл}} | \underbrace{A+B}_{\text{нөхцөл}}) = \frac{P((A+B)(C+D))}{P(A+B)}$
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.

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.

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

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

$$\begin{aligned} P_k(n) = \Bigg[&\sum_{i=1}^{k-1} P(i \text{ эрийг сонгох}|i \text{ эр хамгийн сайн нь}) \\ &+ \sum_{i=k}^{n} P(i \text{ эрийг сонгох}|i \text{ эр хамгийн сайн нь}) \Bigg] P(i \text{ эр хамгийн сайн нь}) \end{aligned}$$

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

$$= \left[\sum_{i=k}^{n}\frac{k-1}{i-1}\right]\frac{1}{n} = \frac{k-1}{n}\sum_{i=k}^{n}\frac{1}{i-1} \quad (k>1)$$

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

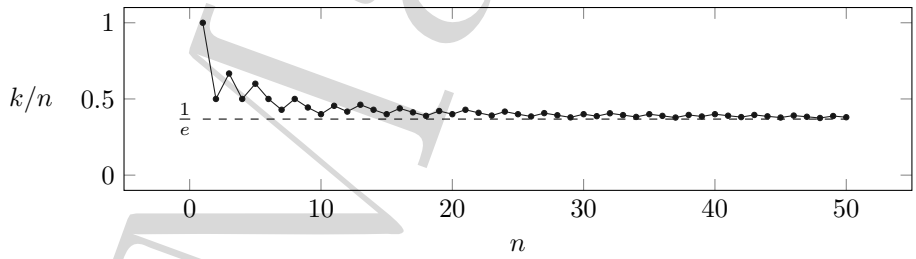
$n=1,\ldots,10$ үед хамгийн оновчтой k буюу хэд дэх эр дээр хамгийн сайн сонголт таарахыг R програмын тусламжтай бодож гаргав.

```
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.



Зураг 25: Оновчтой сонголтын бодлогын k аргумент ба n параметрийн харьцаа

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

Баталгаа $n\to\infty$ үед $P_k(n)=\frac{k-1}{n}\sum\limits_{i=k}^{n}\frac{1}{i-1}$ магадлал дотор $x=\lim\limits_{n\to\infty}\frac{k-1}{n}$ 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 $\lim\limits_{n\to\infty} \frac{1}{n} = 0$ тул дээрх орлуулга ёсоор $\lim\limits_{n\to\infty} \frac{k}{n} = \frac{1}{e}$ болно. □

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

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

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

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

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

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

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

хэлбэртэй болно.

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

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

$f_{X|A}(x) = F'_{X|A}(x)$ ерөнхий томьёо ба $A = \{a \le X < b\}$ үзэгдлээс хамаарсан нөхцөлт тархалтын функцийн томьёонуудаас

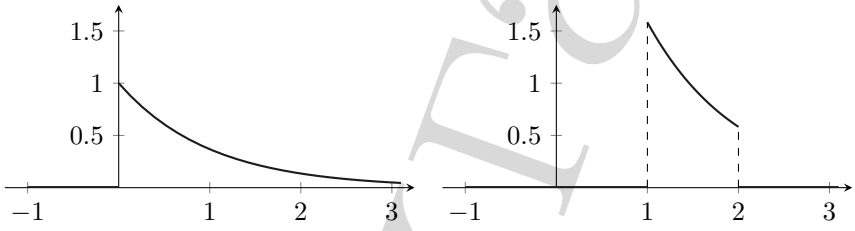
$$f_{X|a\le X<b}(x) = \begin{cases} \frac{f_X(x)}{P(a\le X<b)}, & a < x \le b \\ 0, & \text{бусад} \end{cases}$$

томьёо гарна. 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 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 функцийн графикийг нөхцөлт бус буюу ердийн нягтын функцийн графиктай харьцуулж харуулав. Энд $1 < x < 2$



(a) Нөхцөлт бус нягтын функц (b) Нөхцөлт нягтын функц

Зураг 26: 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

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 томьёолж болох буюу наслалтыг 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.

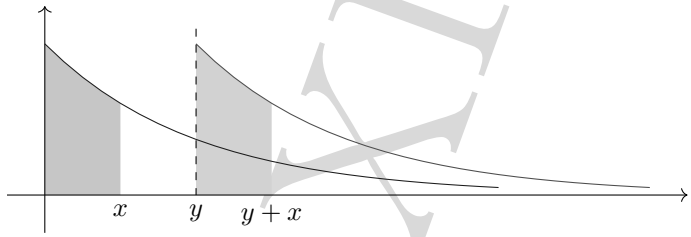
Баталгаа

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

□

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 27: Илтгэгч тархалтын санамжгүй байдал

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

4 Саатлын эрчим

Саатлын эрчим буюу мөхлийн эрчим

Тодорхойлолт 9. 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*[12] буюу *мөхлийн эрчим*[13] гэнэ.

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

[12] failure rate
[13] hazard rate

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*[14] 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.

Тогтмол саатлын эрчимтэй тархалт

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

$$\begin{aligned} h_X(x) = \frac{f_X(x)}{1 - F_X(x)} &= k \\ \int_0^x \frac{f_X(t)}{1 - F_X(t)} dt &= \int_0^x k dt \\ -\int_0^x \frac{1}{1 - F_X(t)} d(1 - F_X(t)) &= k \int_0^x dt \\ -\ln(1 - F_X(t)) \mid_0^x &= kt \mid_0^x \\ -\ln(1 - F_X(x)) &= kx \\ F_X(x) &= 1 - e^{-kx} \end{aligned}$$

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

[14] rate parameter

Тогтмол мөхлийн эрчим – санамжгүй байдал – хамааралгүй туршилт

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 $n \to \infty$ гэхэд хүргэнэ. Энэ тохиолдолд амжилт нь тоологдохуйц байхын тулд $p \to 0$ 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 функц[15]

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

Kaplan-Meier-ын муруй

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

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

гэж уялддаг тул Kaplan-Meier-ын муруйд харгалзах функцийг

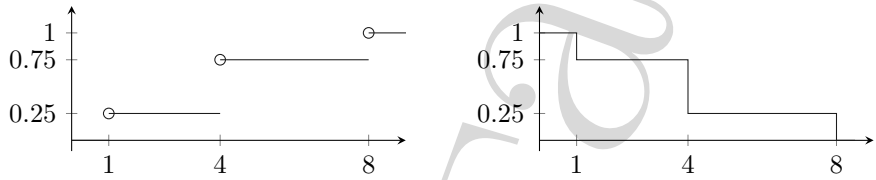
$$R_n(x) = 1 - F_n(x)$$

гэж олно.

[15] Reliability function буюу Survival function

Наслалтын мэдээлэл бүхий 1, 4, 4, 8 өгөгдлөөр Kaplan-Meier-ын муруй байгуул.

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



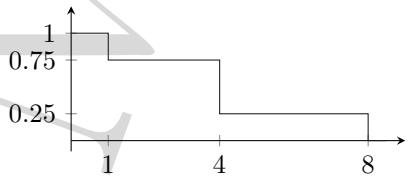
Зураг 28: Туршилтын тархалтын функцийн график ба Kaplan-Meier-ын муруй

Kaplan-Meier-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	$\mathbf{3} = 4 - 1$	$0.75 = 3/4$
4	*2*	$1 = \mathbf{3} - 2$	$0.25 = 1/4$
8	1	$0 = 2 - 1$	$0.00 = 0/4$

Хүснэгт 2: Kaplan-Meier-ын муруй байгуулахад шаардлагатай найдварын функцийн утга олох нь



Зураг 29: Kaplan-Meier-ын муруй

Цензуртай өгөгдөл ба Kaplan-Meier-ын муруй

Судалгаанд хамруулж буй объектуудын зарим нь судалгаа дуусахад саатаагүй 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 цензуртай өгөгдлөөр зохиох найдварын функцийг үлдэж буй хугацааны доторх бүрэн тодорхой мэдээлэлд л үндэслэж зохиодог.

Цензуртай өгөгдөл

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

Хүснэгт 3: Цензуртай өгөгдөл

Цензуртай өгөгдлөөр Kaplan-Meier-ын муруй байгуулах

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 4: Цензуртай өгөгдлөөр Kaplan-Meier-ын муруй байгуулахад шаардагдах тооцоо

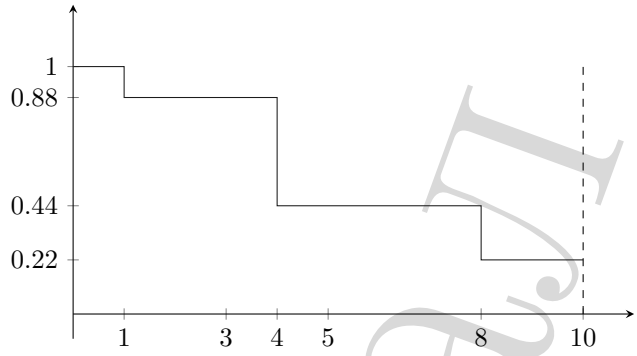
Энд $\mu_n(x)$ болон $\nu_n(x)$ нь хугацааны x эгшинд харгалзах үзэгдлийн давтамжийг цензуртай эсэхээр ялгаж олсоныг илэрхийлнэ.

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.

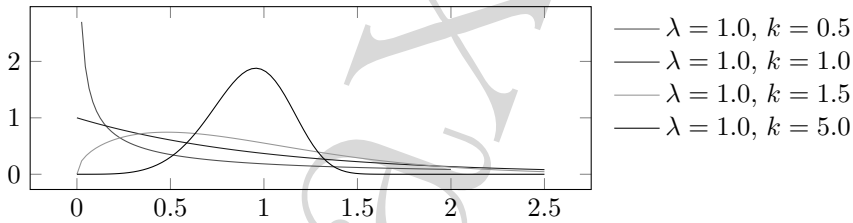
$$f_X(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x > 0 \\ 0 & x \le 0 \end{cases}$$



Зураг 30: Цензуртай өгөгдлөөр байгуулсан Kaplan-Meier-ын муруй

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



Зураг 31: Вейбуллын тархалтын нягтын функц параметрийн янз бүрийн утгад

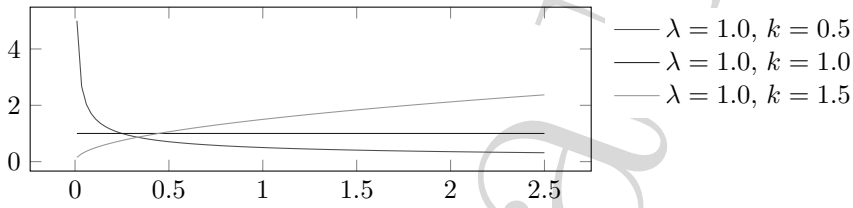
Тархалтын функц

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

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

Вейбуллын тархалтын 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 32: 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 холбоо

Weib(λ, k) буюу

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

нь $k = 1$ үед

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

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

Лекц V
Бернуллийн процесс

> Магадлал заримдаа гарцаагүйд маш ойролцоо байдаг ч тэр нь хэзээ ч үнэхээр гарцаагүй байдаггүй. — Мюррей Гелл-Манн

1 Урвасан бином тархалт

Урвасан бином тархалтаар шийдэх бодлого

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*[16] гээд $NB(r, p)$ гэж тэмдэглэнэ. Энд p нь амжилтын магадлал юм.

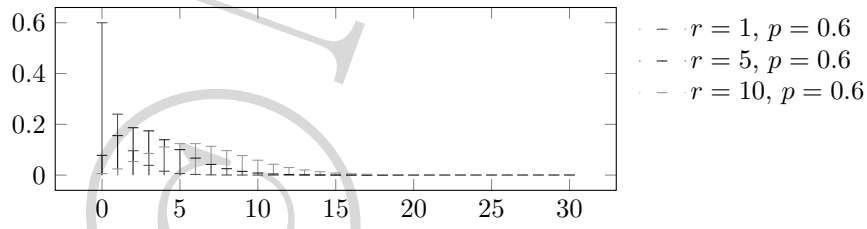
Амжилтыг A, бүтэлгүйтлийг $\bar{A}$ гэж тэмдэглээд улмаар туршилтууд хамааралгүй гэдгийг анхаарч эцэст нь биномын коэффициентийн чанар ашиглавал r ширхэг амжилтаас өмнө x удаа бүтэлгүйтэх магадлал буюу урвасан бином тархалтын нягтын илэрхийлэл дараах байдлаар олдоно.

$$P(X=x) = P(\overbrace{\underbrace{\bar{A}\cdot\bar{A}\cdot A\cdot\bar{A}\cdot\ldots\cdot\bar{A}}_{A \text{ нь } r \text{ ширхэг, } \bar{A} \text{ нь } x \text{ ширхэг}}\cdot\overbrace{A}^{+1}}^{x+r-1 \text{ ширхэг туршилт}} + \ldots + \overbrace{\underbrace{\bar{A}\cdot\bar{A}\cdot\bar{A}\cdot A\cdot\ldots\cdot\bar{A}}_{A \text{ нь } r \text{ ширхэг, } \bar{A} \text{ нь } x \text{ ширхэг}}\cdot\overbrace{A}^{+1}}^{x+r-1 \text{ ширхэг туршилт}})$$

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

(C_{x+r-1}^{r-1} ширхэг ялгаатай үр дүн)

Урвасан бином тархалтын нягтын хэлбэр



Зураг 33: Урвасан бином тархалт, параметрийн янз бүрийн утгад

[16] Negative binomial distribution

Бодлогын шийдийг урвасан бином тархалтын тусламжтай олсон нь

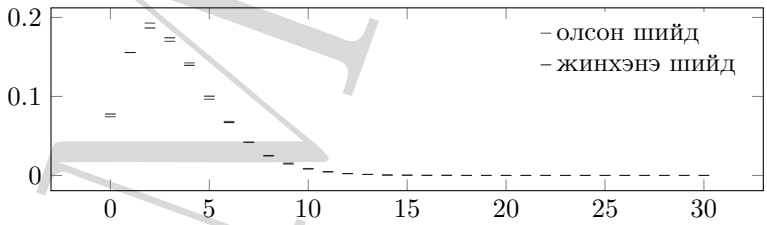
Бодлогын шийдийг урвасан бином тархалтын тусламжтай R програм дээр дараах тушаалаар олж болно.

```
P_nb <- dnbinom(x = 0:30, prob = 0.6, size = 5)
P_nb <- P_nb/sum(P_nb)  # X ≤ 30 буюу тархалт хэрчигдсэн тул
    нормчилно.
print(P_nb)
```

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)
```

Симуляц ба урвасан бином тархалтаар олсон хоёр шийдийн харьцуулалт



Зураг 34: x ширхэг айл юу ч хялайлгалгүй явуулах магадлал

Урвасан бином тархалтын зарим чанар

- $E(X) = \dfrac{(1-p)r}{p}$ Жишээний хувьд $E(X) = \dfrac{0.4 \cdot 5}{0.6} \approx 3.333$ байна.
- $D(X) = \dfrac{(1-p)r}{p^2}$ Жишээний хувьд $D(X) = \dfrac{0.4 \cdot 5}{0.6^2} \approx 5.555$ байна.

- Түүврийн дисперс нь түүврийн дунджаасаа их үед Пуассоны тархалт[17]-ын оронд ашигладаг.
- $NB\left(r, \frac{\lambda}{r+\lambda}\right) \xrightarrow{r\to\infty} \text{Pois}(\lambda)$
- $NB(r=1,p) = \text{Geom}(p)$

2 Бернуллийн процесс

Бернуллийн процесс

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 гэе. Тэгвэл үүнтэй холбогдуулан практикт өргөн тохиолдох янз бүрийн санамсаргүй хувьсагч зохиож болно.

Тодорхойлолт 10. $X_i \sim \text{Ber}(p)$ буюу $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$ бас $X_1, X_2, \ldots$ хамааралгүй байг. Тэгвэл $X_1, X_2, \ldots$ санамсаргүй хувьсагчдын төгсгөлөг болон төгсгөлгүй дарааллыг *Бернуллийн процесс* гэнэ.

Бернуллийн процесстой холбогдох хялбар бодлого

Ану, Бат, Вандан, Гэрэл, Дөлгөөн нар нэг зоосон мөнгө олоод түүнийгээ хэн нь авах вэ гэдгээ шодохоор шийджээ. Гэтэл Ану "Нэрсийнхээ дарааллаар зоосоо хаяад хамгийн эхэлж тоотой талаараа буулгасан нь зоосоо авъя. Хэрвээ хэн нь ч тоотой талаараа буулгахгүй бол зоосоо буяны санд хандивлая." гэсэн санал гаргав. Анугийн санал шударга уу?
Эх сурвалж: www.slideshare.net/Erdenetsagaanaa/ss-40157371

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

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

геометр тархалт гарах бөгөөд $p = 0.5$ байна.

Бернуллийн процессын санамжгүй байдал

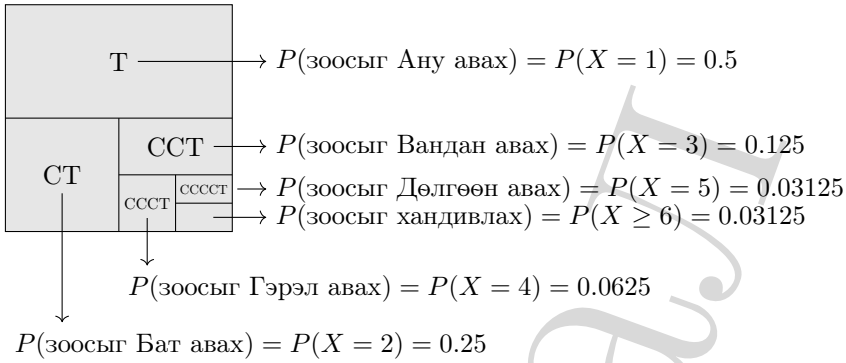
$X_1, X_2, \ldots$ хувьсагчид хамааралгүй тул энэ процесс санамжгүй юм. Иймд энэ процессын өнгөрсөнөөс ирээдүйг урьдчилан хэлэх боломжгүй.

Бернуллийн процесстой холбоотой зарим тархалт

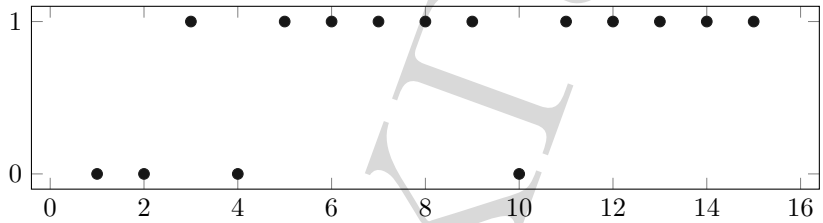
Энэ процессын хувьд дараах санамсаргүй хувьсагчдыг өмнө авч үзсэн.

1. Дараалсан n туршилтад илрэх амжилтын тоо. $B(n,p)$ буюу бином тархалт.

[17] дундаж ба дисперс нь тэнцүү байдаг



Зураг 35: Бернуллийн процесстой холбогдох бодлогын зураглал



Зураг 36: $p = 0.6$ үед симуляцалсан Бернуллийн процессын нэг тохиолдол

2. Нэг амжилт илрэх хүртэлх бүтэлгүйтлийн тоо. Geom(p) буюу геометр тархалт.
3. Нэг амжилт илрэх хүртэлх туршилтын тоо. Geom(p) буюу геометр тархалт.
4. r удаа амжилт илрэх хүртэлх бүтэлгүйтлийн тоо. $NB(r, p)$ буюу урвасан бином тархалт.

Бернуллийн процесстой холбоотой тархалтуудын уялдаа

Геометр тархалт ба Урвасан бином тархалт

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

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

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

$$\underbrace{0\,0\,0\,0\,1\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,1}$$

B : n = нийт туршилтын тоо = $s + r$: NB

B : r = амжилтын тоо = r : NB

B : амжилтын тоо = хувьсагч = бүтэлгүйтлийн тоо : NB

$NB(r,p)$ 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.

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

Харин шинээр үүсэх X болон Y процессууд хамааралгүй байж чадахгүй.

Бадарчинтай жишээ ба Бернуллийн процессын k дугаар амжилтад харгалзах туршилтын дугаар

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

Үүнийг олох нь угтаа $k = 5$ 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 олъё. 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 $\{1, 2, \ldots\}$ утгууд авдаг

$$X = \text{"амжилт илэртэл явуулах туршилтын тоо"}$$

гэсэн хувьсагчтай

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

хувилбарыг ашиглана.

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

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

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

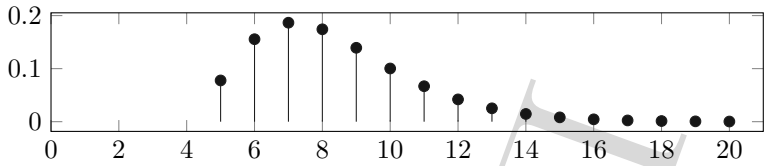
байна.

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

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

$p = 0.6$ үед Y_5 хувьсагчийн хамгийн өндөр магадлалтай утга буюу моод нь $x = 7$ байх тул жишээ бодлогын хариу 7 болно.

Бернуллийн процесс дахь дараалсан амжилтын тооны тархалт



Зураг 37: Y_5 хувьсагчийн нягт, $p = 0.6$ үед

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

Бернуллийн процесс дахь дараалсан амжилтын тоо нь угтаа "бүтэлгүйтэх хүртэл явуулах туршилтын тоо" байх тул $x \in \{1, 2, \ldots\}$ утга бүхий хувьсагчтай 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

> Амьдралд санамсаргүй зүйл гэж үгүй. Болж буй бүх зүйл болох ёстойдоо л болсон. Процесст итгээрэй. — Нэр нь тодорхойгүй

1 Гамма тархалт

Гамма функц

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

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

Гамма функцээс гамма тархалт

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

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

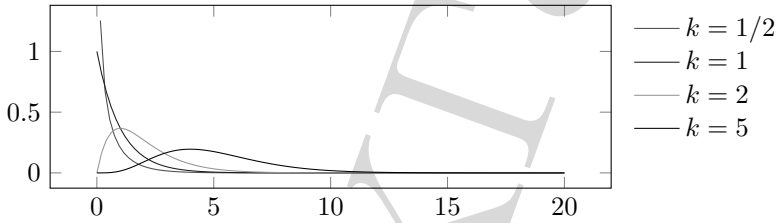
гэж авч болно.

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

Иймэрхүү байдлаар ямар ч функцийг нягтын функц буюу магадлалын хэмжээс болгон авч болдог. Гамма тархалт хэмээх энэхүү шинэ тархалт нь энэ сэдвээр үзэх Пуассоны процесс гэсэн санамсаргүй процесст хэсэг хугацаанд ямар нэг үзэгдэл дор хаяж төчнөөн удаа ажиглагдах магадлалыг олоход хэрэг болно.

Тархалтын хэлбэрийн параметр

Гамма функцийн аргументаас уламжилсан k параметр нь тархалтын хэлбэрийг тодорхойлдог.



Зураг 38: $f_X(x) = \frac{1}{\Gamma(k)} x^{k-1} e^{-x}$ функц k параметрийн янз бүрийн утгад

Тархалтын масштабын буюу эрчмийн параметр

Тархалтад масштабын буюу эрчмийн параметр оруулбал тус тархалт нь урт, богино янз бүрийн хугацаанд үргэлжлэх процессыг загварчлахад ашиглах боломж нээгдэнэ.

Ийнхүү масштаб оруулж ирэх буюу эрчмийн параметр нэмэхийн тулд $x := \lambda x$[18] орлуулга хийнэ. Гамма тархалтад хамаатай $\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx$ интеграл дахь функц тус орлуулгаар

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

болох тул уг эрчмийн параметр бүхий гамма тархалтын нягтын функц

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

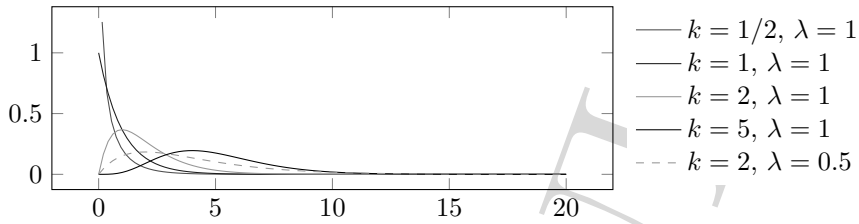
хэлбэртэй болно.

Гамма тархалтын нягтын функц ба нягтын муруй

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

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

[18] Хэрэв масштабын агуулгатай параметр нэмье гэвэл $x := x/\lambda$ орлуулга хийнэ.



Зураг 39: Гамма тархалтын нягтын муруй параметрийн янз бүрийн утгад

Гамма тархалтын чанар

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

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

Гамма тархалт бусад тархалттай холбогдох нь

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

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

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

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

  байна. Энэ чанарын гаргалгаатай дараагийн хэсэгт танилцана.

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

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

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

2 Пуассоны процесс

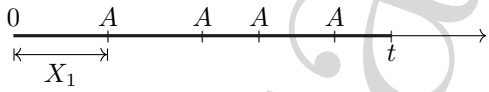
Илтгэгч тархалт ба Пуассоны тархалтын уялдаа холбоо

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

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 40: Пуассоны процессын эхний амжилт илрэх хүртэлх хугацаа

Удаах амжилт илэртэл хүлээх хугацааны тархалт

Зураг 41: 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. Иймхүү $X_2 \sim \text{Exp}(\lambda)$ боллоо. Дараагийн хувьсагчдын хувьд ч ийм байдлаар $X_3, X_4, \ldots \sim \text{Exp}(\lambda)$ гэсэн дүгнэлт гарна.

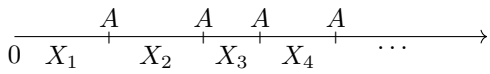
Нөгөө талаас $X_2 \sim \text{Exp}(\lambda)$ гэвэл илтгэгч тархалтын санамжгүй чанараар

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

буюу өмнөхтэй ижил үр дүнд хүрнэ. Иймхүү амжилтын эрчим тогтмол буюу λ параметрийн утга тогтмол үед удаах амжилт илэртэл хүлээх хугацаа өнгөрсөнөөс хамаарахгүй тул эхний амжилт хүртэлх хугацаа болон амжилт хоорондын хугацаагаар тодорхойлогдох санамсаргүй хувьсагчдын дараалал нь ой санамжгүй процесс үүсгэнэ.

Пуассоны процесс[19]

[19] Poisson process



Зураг 42: Пуассоны процессын зураглал

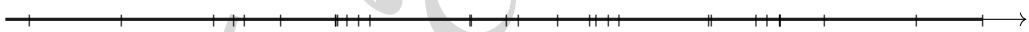
Тодорхойлолт 11. X_1 нь эхний амжилт илэртэл хүлээх хугацаа, X_i $(i = 2, 3, \ldots)$ нь дараагийн амжилт хоорондын хугацаа ба $X_1, X_2, \ldots$ хувьсагчид нэг ижил $\text{Exp}(\lambda)$ тархалттай бөгөөд хамааралгүй байг. Тэгвэл $X_1, X_2, \ldots$ санамсаргүй хувьсагчдын дарааллыг *нэгэн төрлийн Пуассоны процесс* гэнэ.

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

Пуассоны процессын жишээ

> ”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.

Энд рекорд амжилт хоорондын хугацаа үнэхээр Пуассоны процесс, цаашилбал нэгэн төрлийн Пуассоны процесс болох уу гэсэн асуулт зайлшгүй тавигдана.



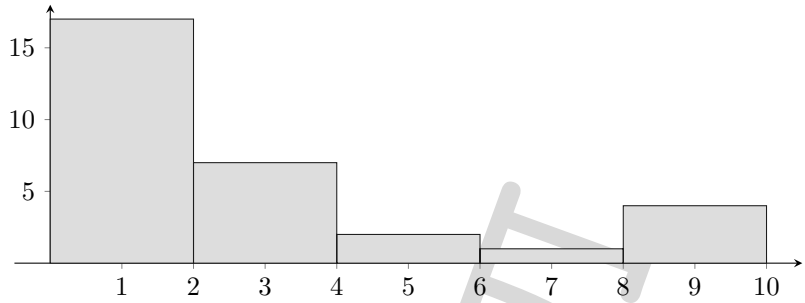
Зураг 43: Рекорд амжилт хэдийд бүртгэгдсэнийг хугацаан цацраг дээр дүрсэлсэн нь

Энэхүү процессыг нэгэн төрлийн Пуассоны процесс гэх үндэслэл гаргаж болно. Үүний тулд статистикийн зохих аргууд ашиглана. Эхлээд өгөгдлийг R програмд дараах байдлаар оруулна.

```
X <- c(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)
```

Уг өгөгдлийн хувьд бүлгийн шинжүүрээр санамсаргүй гэдэг тэг таамаглалыг хоёр талт өрсөлдөгч таамаглалын эсрэг `randtests::runs.test(X)` гэж шалгахад 0.710 гэсэн магадлалын утга гарсан бас `acf(X)` тушаалаар олдох тус хугацаан цувааны автокорреляцид үндэслэн рекорд амжилт хоорондын зай нь санамсаргүй бөгөөд санамжгүй процесс юм хэмээн дүгнэж болно.

Дараах гистограммаас рекорд амжилт хоорондын хугацаа илтгэгч тархалттай гэсэн таамаг төсөөлөл гарна. Үүн шиг гистограмм байгуулахын тулд `hist(X)` тушаал өгнө.



Зураг 44: Рекорд амжилт хоорондын хугацааны тархалтыг харуулсан гистограмм

Нэгэн төрлийн Пуассоны процесс дахь санамсаргүй хувьсагчийн тархалтын параметрийг үнэлэх

Илтгэгч тархалтын хувьд $E(X) = 1/\lambda$ ба тоон өгөгдлөөс олдох түүврийн дундаж ойролцоогоор 2.779 тул $\lambda = 1/2.779 \approx 0.3598$ гэж "үнэлнэ".

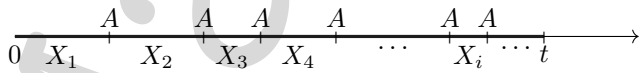
Тархалтын параметр олох иймэрхүү аргыг моментын арга гэдэг бөгөөд хожим үзнэ.

Ийнхүү Пуассоны процесс үүсгэх илтгэгч тархалттай санамсаргүй хувьсагчийн эрчмийн параметрийн утгыг олсон тул зарим үзэгдлийн магадлалыг тооцоолж болно.

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

Цаашилбал уг процесст ажиглагдах бусад үзэгдлийн магадлалыг ч олж болох бөгөөд үүнд Пуассоны тархалт болон гамма тархалт ашиглана.

Нэгэн төрлийн Пуассоны процесс симуляцлах



Зураг 45: t хугацаанд үргэлжлэх Пуассоны процесс

```
X := 0                          # процессын үргэлжилсэн хугацаа
REPEAT                          # процессыг эхлүүлэх
  X_i := GENERATE_EXP(lambda)   # илтгэгч тархалттай санамсаргүй
    утга
  X := X + X_i                  # X = X_1 + X_2 + ... + X_i үргэлжилсэн
    хугацаа
  IF (X > t)                    # процесс нь заасан хугацааны
    заагийг давах
    BREAK                       # давталтыг зогсоох буюу процессыг
    дуусгах
```

```
  PRINT X_i                          # санамсаргүй процесс үүсгэх утга
ENDREPEAT                            # давтагдах бүлэг тушаалын төгсгөл
```

Нэгэн төрлийн Пуассоны процесс симуляцлах алгоритмыг R хэлээр дараах байдлаар програмчилж болно.

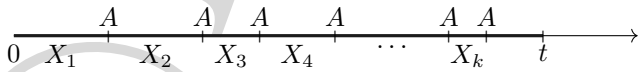
```
X <- 0                               # процессын үргэлжилсэн хугацаа
repeat {                             # процессыг эхлүүлэх
  X_i <- rexp(1, lambda)             # илтгэгч тархалттай санамсаргүй
    утга
  X <- X + X_i                       # X = X_1 + X_2 + ... + X_i үргэлжилсэн
    хугацаа
  if (X > t)                         # процесс нь заасан хугацааны
    заагийг давах
    break                            # давталтыг зогсоох буюу процессыг
    дуусгах
  print(X_i)                         # санамсаргүй процесс үүсгэх утга
}                                    # давтагдах бүлэг тушаалын төгсгөл
```

Нэгэн төрлийн бус Пуассоны процесс симуляцлах

Нэгэн төрлийн бус Пуассоны процесс симуляцлахад илтгэгч тархалтын эрчмийн параметрийг процессын үргэлжилсэн хугацаанаас хамааруулан тухай бүр олж шинэчлэх гэсэн нэмэлт ажил гарна. Энэ нь процесс ой санамжтай болсон буюу санамсаргүй хувьсагчдын хооронд цаг хугацаа гэх хөндлөнгийн нөлөөнөөс үүдэх холбоо хамаарал үүссэнийг илтгэнэ.

```
X := 0
REPEAT
  lambda := ESTIMATE_RATE(X)         # эрчмийн параметрийн утга олох
  X_i := GENERATE_EXP(lambda)
  X := X + X_i
  IF (X > t)
    BREAK
  PRINT X_i
ENDREPEAT
```

t хугацаанд амжилт яг k удаа илрэх магадлал ба Пуассоны тархалт



Зураг 46: Пуассоны процесст t хугацаанд амжилт яг k удаа илрэх

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

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

гэж гарна.

Жишээний хувьд

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

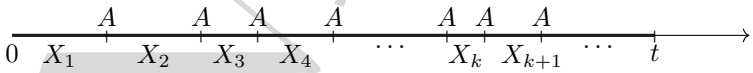
Жишээ бодлогыг симуляцийн аргаар бодсон нь

4 жилд яг нэг шинэ рекорд гарах нь эхний рекорд 4 жилийн дотор гараад харин удаах рекорд 4 жил дотор багтахгүйтэй тэнцүү чанартай юм. Иймд симуляцийн туршилтын үеэр $\{X_1 \leq 4\}\{X_1 + X_2 > 4\}$ үзэгдлийн давтамжийг олно.

```
set.seed(1)
n <- 0                                  # үзэгдлийн тоо буюу давтамж
for (i in 1:1000) {                     # туршилтыг 1000 удаа давтана
  X1 <- rexp(n = 1, rate = lambda)      # эхний рекорд амжилт хүртэлх
    хугацаа
  X2 <- rexp(n = 1, rate = lambda)      # удаах рекорд амжилт хүртэлх
    хугацаа
  if (X1 <= 4 && X1 + X2 > 4)           # 4 жилд яг нэг шинэ рекорд
    гарах
    n <- n + 1                          # үзэгдэл тоолох
}
n / 1000                                # үзэгдлийн давтамж

0.34                                    # гарсан хариу
```

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



Зураг 47: Пуассоны процесст t хугацаанд амжилт дор хаяж k удаа илрэх

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

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

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

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

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

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

болно.

Жишээ бодлогыг симуляцийн аргаар бодсон нь

```
set.seed(1)
n <- 0                                  # үзэгдлийн тоо буюу давтамж
for (i in 1:10000) {                    # туршилтыг 10000 удаа давтана
  X1 <- rexp(n = 1, rate = lambda)      # эхний рекорд амжилт хүртэлх
    хугацаа
  if (X1 <= 4) {                        # 4 жилд дор хаяж нэг шинэ
    рекорд гарах
    n <- n + 1                          # үзэгдэл тоолох
  }
}
n / 10000                               # үзэгдлийн давтамж
```

```
| 0.765                    # гарсан хариу
```

Лекц VII
Санамсаргүй хувьсагчийн хувиргалт

> Амьдралын хамгийн чухал асуултууд нь ихэнх тохиолдолд үнэндээ зөвхөн магадлалын асуудлууд байдаг. — Пьер-Симон Лаплас

1 Санамсаргүй хувьсагчийн хувиргалт

Санамсаргүй хувьсагчийн хувиргалт

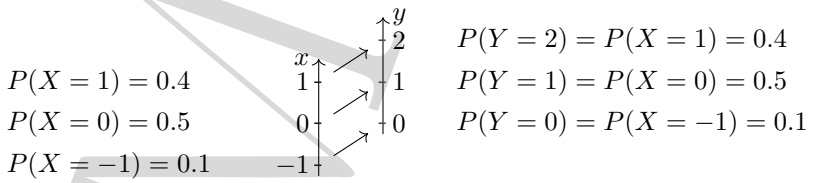
Тодорхойлолт 12. Санамсаргүй хувьсагчийн утгыг ямар нэг функцээр үйлчлэн өөрчлөхийг *санамсаргүй хувьсагчийн хувиргалт* гэнэ.

Дискрет санамсаргүй хувьсагчийн 1:1 хувиргалт

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

Хүснэгт 5: X санамсаргүй хувьсагчийн тархалтын хүснэгт

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



Зураг 48: Дискрет санамсаргүй хувьсагчийн 1:1 чанартай хувиргалт

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

Хүснэгт 6: $Y = X + 1$ санамсаргүй хувьсагчийн тархалтын хүснэгт

Бодолт дараах байдалтай байсан.

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

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

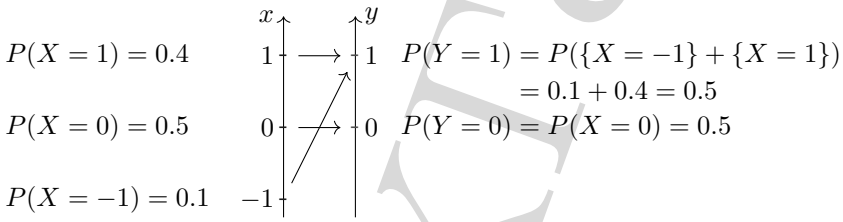
Ийнхүү 1:1 чанартай $g(\cdot)$ функцээр тодорхойлогдох $Y = g(X)$ дискрет санамсаргүй хувьсагчийн нягт олох

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

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

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

Хүснэгт 7: $Y = |X|$ санамсаргүй хувьсагчийн тархалтын хүснэгт

Y хувьсагчийн тархалтыг олохын тулд

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

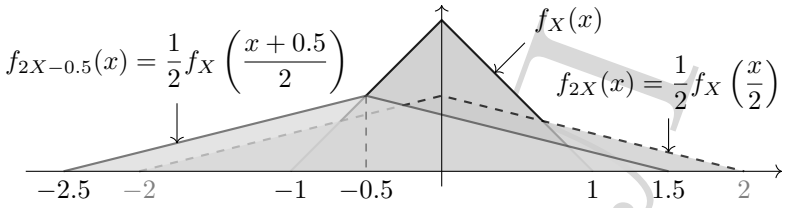
бодолт хийсэн. Ийнхүү 1:1 нөхцөл үл хангах $g(\cdot)$ функцээр тодорхойлогдох $Y = g(X)$ дискрет санамсаргүй хувьсагчийн нягт олох дараах томьёо зохиож болно.

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

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

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



Зураг 50: $Y = 2X - 0.5$ 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

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

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

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

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

$Y = g(X)$ 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

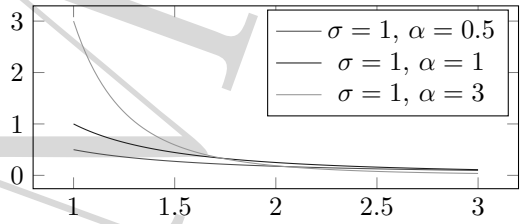
$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|$$

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

Энд $X \sim \mathrm{U}(0,1)$, $\sigma > 0$, $\alpha > 0$ 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

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

буюу I төрлийн Парето тархалт гарна.



Парето тархалтын нягт, параметрийн янз бүрийн утгад

Парето тархалтын хэрэглээ ба өргөн сүүлтэй тархалт

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

1. Элсний ширхэгийн хэмжээ

2. Солирын хэмжээ
3. Сүлжээгээр дамжих файлын хэмжээ
4. Суперкомпьютерт өгөх ажлын хэмжээ
5. Хот, сууринги йн хэмжээ
6. Эрсдэл, гамшиг

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

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

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

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

Тасралтгүй санамсаргүй хувьсагчийн $Y = X^2$ хувиргалт

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

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

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

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

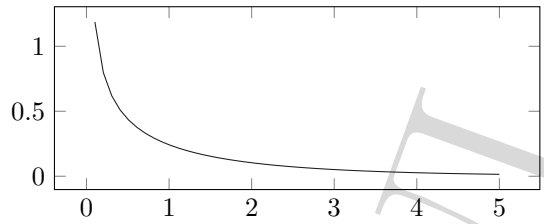
болно.

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

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

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

буюу 1 чөлөөний зэрэгтэй хи-квадрат тархалт гарна.



Зураг 51: $f_Y(y)$ функцийн график буюу 1 чөлөөний зэрэгтэй хи-квадрат нягтын муруй

Геометр тархалттай хувьсагчийн $Y = X^2$ хувиргалт

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

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

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

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

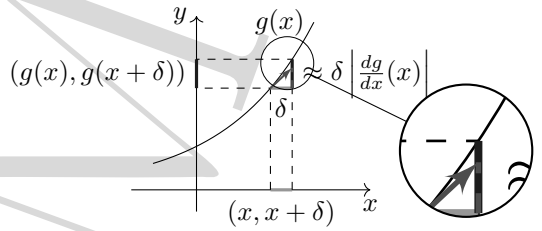
томьёогоор тархалтыг нь олж болно. Иймхүү Y хувьсагчийн нягт

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

болно.

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

$g(\cdot)$ функц эрс монотон байг.



Зураг 52: Хувиргалтаар үүсэх хувьсагчийн нягт олох томьёоны гаргалгаа

$$\begin{array}{ccc} P(x<X<x+\delta)) = P(\underbrace{g(x)}_{y}<Y<g(x+\delta))) \approx P(y<Y<y+\delta|\frac{dg}{dx}(x)|)) \\ \| & & \| \\ \delta f_X(x) & & \delta\left|\frac{dg}{dx}(x)\right| f_Y(y) \end{array}$$

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

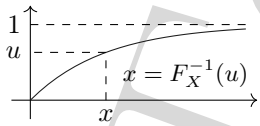
2 Санамсаргүй хувьсагч симуляцлах урвуу хувиргалтын арга

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

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

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



Зураг 53: Урвуу хувиргалтын аргын санаа ба тархалтын квантил

Урвуу хувиргалтын аргаар Парето тархалттай санамсаргүй тоо үүсгэх

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

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

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

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

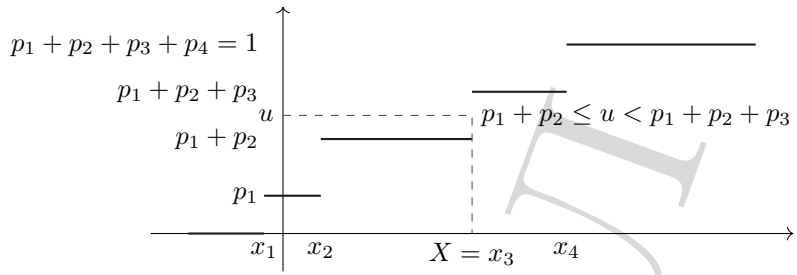
Тус томьёо болон урвуу хувиргалтын аргын дагуу дараах програм бичиж болно. Програмын кодыг R хэлээр бичив.

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

Урвуу хувиргалтын аргаар дискрет тархалт симуляцлах

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

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



Зураг 54: Урвуу хувиргалтын аргаар дискрет санамсаргүй хувьсагч симуляцлах зарчим

Урвуу хувиргалтын аргаар дискрет хувьсагч симуляцлах алгоритм

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

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

Хүснэгт 8: Дискрет санамсаргүй хувьсагчийн тархалтын хүснэгт

Хүснэгтээр өгсөн тархалттай X санамсаргүй хувьсагчийг урвуу хувиргалтын аргаар симуляцал.

```
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

50-50-90-ийн дүрэм: Таны сонголт 90 хувийн магадлалтайгаар буруу болох байлаа ч танд сонголтоо зөв хийх 50-50 хувийн боломж үргэлж бий. — Энди Рүүни

1 Санамсаргүй вектор, түүний тархалт

Санамсаргүй вектор, хамтын тархалт

Тодорхойлолт 13. Нэгээс олон санамсаргүй хувьсагчдыг хамтад нь *санамсаргүй вектор*, $(X_1, \ldots, X_p)$ 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 9: Хамтын тархалт, статистик хэмжээс ашиглаж олсон

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

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

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

Зураг 55: Тасралтгүй санамсаргүй векторын хамтын нягт ба санамсаргүй түүвэр

Хамтын тархалт ба үзэгдлийн магадлал

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

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

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

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

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

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

Тухайн тархалт

Тодорхойлолт 14. 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		$\sum$
	1	0	
1	**0.2**	0.3	0.5
0	**0.1**	0.4	0.5
$\sum$	**0.3**	0.7	1

Хүснэгт 10: Хамтын тархалт ба тухайн тархалт

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

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

Зураг 56: Тасралтгүй санамсаргүй векторын хамтын болон тухайн нягт

Хамтын тархалт ба санамсаргүй хувьсагчдын хамааралгүй байдал

Өмнө $\forall(x,y) \in \mathbb{R}^2$ бүрийн хувьд $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ бол X болон Y хувьсагчдыг хамааралгүй гэнэ хэмээн тодорхойлж байснаа эргэн авч үзье.

X болон Y хувьсагчид хамааралгүй бол $\forall(x,y) \in \mathbb{R}^2$ бүрийн хувьд $\{X < x\}$ болон $\{Y < y\}$ үзэгдлүүд хамааралгүй байна. Иймд

$$F_{X,Y}(x,y) = P(X < x, Y < y) = P(X < x)P(Y < y) = F_X(x)F_Y(y)$$

болно. Тархалтын функцийг дифференциалчилбал нягтын функц гардагийг санавал $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ нөхцөл мөрдөн гарна.

X болон Y хувьсагчдыг хамааралгүй гэдгийг $X \perp\!\!\!\perp Y$ байдлаар тэмдэглэнэ.

Нөхцөлт тархалт

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

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

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

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

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

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

Дискрет санамсаргүй хувьсагчийн нөхцөлт тархалтыг дараах байдлаар олно.

$$f_{X_1|X_2}(X_1 = 1|X_2 = 1) = P(X_1 = 1|X_2 = 1) = P(\text{эр}|\text{солгой})$$

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

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

Хүснэгт 11: Хамтын, тухайн болон нөхцөлт тархалт

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

Мөн жишээний хувьд

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

тул X_1 ба X_2 хувьсагч хамааралтай.

$f_{X,Y}(x,y) = \frac{3x^2}{16} + \frac{y}{2}$, $0 < x < 2$, $0 < y < 1$ 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 тавигдахааргүй байгаа явдал нь тус хувьсагчдыг хамааралгүй байж чадахгүйг харуулж байна.

2 Бүтэн магадлалын томьёо

Үржүүлэх дүрэм ба бүтэн магадлалын томьёо

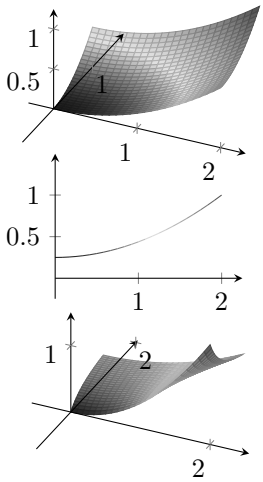
X хувьсагч Y хувьсагчаас хамаардаг гэж үзье.

Үржүүлэх дүрэм

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

Бүтэн магадлалын томьёо

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



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

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

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

Зураг 57: Тасралтгүй санамсаргүй хувьсагчийн хамтын тархалт, тухайн тархалт, нөхцөлт тархалт

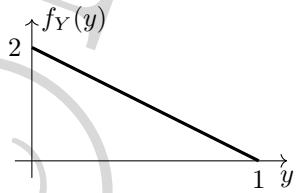
Богд Жавзандамба хутагтад сүсэгтэн олноос өргөдөг өргөл барьцын хэмжээ $k = 3$ хэлбэрийн параметр болон $1/\lambda = 1$ масштабын параметр бүхий гамма тархалттай байв. Харин Данигай сойвон өргөл барьцын Y хувийг Богдын санд бүртгээд бусдыг нь хувьдаа завшдаг бол санд орох өргөл барьцын хэмжээний тархалтыг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

Тунгалаг тамир романаас сэдэвлэв.

Өргөл барьцын анхны хэмжээг X гэвэл бодлогын нөхцөл ёсоор

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

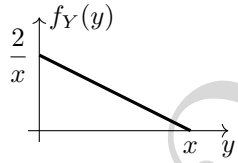
болно. Харин санд орох өргөл барьцын хэмжээ $Z = X \cdot Y$ буюу X хувьсагчаас хамаарна. Тэгэхээр үүнийг Y хувьсагчийн хувиргалт гэж болно.



Зураг 58: Y санамсаргүй хувьсагчийн тархалт

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

Үржүүлэх дүрмээр $f_{X,Z}(x,z) = f_{Z|X}(z|x)f_X(x) = (x-z)e^{-x}$, $x > 0$, $0 < z < x$ болно. Зургаас харвал $0 < Z < \infty$ ба $Z < X < \infty$ байна. Иймд бүтэн



Зураг 59: X хувьсагчийн нөхцөл дэх Z санамсаргүй хувьсагчийн тархалт

магадлалын томьёогоор дараах илтгэгч тархалт олдоно.

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

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

Үржүүлэх дүрэм

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

Бүтэн магадлалын томьёо

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

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

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

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

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

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 томьёо* олгодог. Холбоо хамаарлыг үүн шиг нөгөө талаас нь авч үзэхийг *Байесын зарчим* харин уг зарчимд тулгуурласан статистик шинжилгээг *Байесын шинжилгээ* гэдэг.

Байесын томьёогоор $f_{Y|X}(y|x)$ 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 томьёо

Байесын томьёо

$$
\begin{aligned}
f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{f_X(x)}\frac{f_Y(y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}\frac{f_Y(y)}{f_X(x)} \\
&= \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int\limits_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)dy} \\
P(B_i|A) &= \frac{P(B_iA)}{P(A)} = \frac{P(B_iA)}{P(A)}\frac{P(B_i)}{P(B_i)} = \frac{P(AB_i)}{P(B_i)}\frac{P(B_i)}{P(A)} \\
&= \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum\limits_{i=1}^{k} P(A|B_i)P(B_i)}
\end{aligned}
$$

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

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

Тодорхойлолт 15. $\forall(x,y,z) \in \mathbb{R}^3$ хувьд

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

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

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

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

2 дугаар чанарын баталгаа

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

□

Үржүүлэх дүрэм ба нөхцөлт үл хамаарал

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

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

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

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

болно.

Нөхцөлт болон нөхцөлт бус үл хамаарал

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

боловч үүний урвуу өгүүлбэр нь ерөнхийдөө худал өөрөөр хэлбэл

$$(X \perp\!\!\!\perp Y) \mid Z \not\Rightarrow X \perp\!\!\!\perp Y$$

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

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

$X \leftarrow Z \rightarrow Y$ **тохиолдол**

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

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

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

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

Энэ нь ерөнхийдөө $f_X(x) f_Y(y)$ байж чадахгүй тул $X \not\perp\!\!\!\perp Y$ байна. Харин одоо Z хувьсагчийг нөхцөлд авъя.

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

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

$X \rightarrow Z \rightarrow Y$ **тохиолдол**

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

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

Энд бүтэн магадлалын томьёо ашиглав. Сүүлийн илэрхийлэл $f_X(x) f_Y(y)$ биш байгаа тул $X \not\perp\!\!\!\perp Y$ гэж дүгнэнэ. Одоо Z хувьсагчийг нөхцөлд авъя. Байесын томьёо ашиглавал

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

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

$X \to Z \leftarrow Y$ **тохиолдол**

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

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

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

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

болно. Эндээс z аргументаар интеграл авлаа ч $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ буюу $X \perp\!\!\!\perp Y$ чанар хадгалагдана. Харин одоо Z хувьсагч дээр нөхцөл тавих буюу утгыг нь бэхэлье.

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

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

5 Нөхцөлт математик дундаж

Нөхцөлт математик дундаж

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

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

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

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

болно.

Санамсаргүй тоо ширхэг бүхий санамсаргүй хувьсагчдын нийлбэрийн математик дундаж

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

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

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

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

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

болно.

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

Бүтэн дисперсийн томьёо дараах хэлбэртэй байна.

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

Үүний тусламжтай санамсаргүй тоо ширхэг бүхий санамсаргүй хувьсагчдын нийлбэрийн дисперсийг, тухайлбал Тунгалаг тамир романаас сэдэвлэсэн жишээ дэх Богдын сангийн орлогыг илэрхийлэх $S = Z_1 + \ldots + Z_T$ хувьсагчийн дисперсийг

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

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

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

Лекц IX

Олон хэмжээст хэвийн тархалт ба шугаман загвар

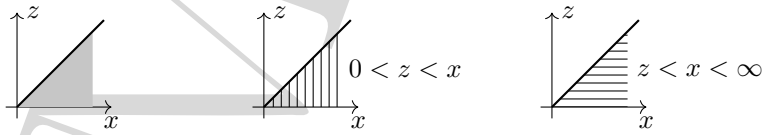
> Таны загварчилж буй үзэгдэл шугамантай ойролцоо эсэх талаар бодохгүйгээр шугаман регрессийг ашиглаж болох хэдий ч та тэгэх ёсгүй. — Жордан Элленберг

1 Санамсаргүй векторын математик дундаж ба ковариацийн матриц

Санамсаргүй векторын математик дундаж

$X = (X_1, \ldots, X_p)^{\mathrm{T}}$ буюу p хэмжээст санамсаргүй вектор авч үзье. Энд $^{\mathrm{T}}$ нь матриц, векторын хөрвүүлэх үйлдлийг илэрхийлнэ. X санамсаргүй векторын математик дунджийг дараах байдлаар тодорхойлно.

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



Зураг 60: (X, Z) санамсаргүй векторын авах утга

Хүснэгтээр өгсөн тархалттай дискрет санамсаргүй векторын математик дунджийг ол.

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

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

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

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

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

Хоёр санамсаргүй хувьсагчийн ковариац ба корреляц

X болон Y санамсаргүй хувьсагчдын ковариацийн коэффициентыг

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

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

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

Харин корреляцийн коэффициентыг дараах байдлаар тодорхойлдог.

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

Ковариац болон корреляцийн чанар

Чанар 8. 1. $\operatorname{cov}(X,X)=D(X)$

2. $\operatorname{cov}(X,Y)=E(XY)-E(X)E(Y)$
3. X болон Y хамааралгүй бол $\operatorname{cov}(X,Y)=0$ байна.
4. $-1\le\rho(X,Y)\le1$
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.

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

Үүнийг олоход ковариацийн 2 дугаар чанар болон ухамсаргүй статистикчийн хууль хэрэг болно.

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

Тасралтгүй тохиолдолд зарчмын хувьд өмнөхтэй төстэй байдлаар

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

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

Санамсаргүй векторын ковариацийн матриц

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

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

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

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

нь $p\times q$ хэмжээст матриц байна.

Ковариацийн матриц ба корреляцийн коэффициент

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

олно.

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

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

байна. Корреляцийн коэффициентүүдээр матриц байгуулж болох бөгөөд жишээний хувьд

$$\mathcal{R} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

байна.

2 Олон хэмжээст хэвийн тархалт

Олон хэмжээст хэвийн тархалт

$$f_X(x) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right\} \quad x = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$$

нягттай тархалтыг *олон хэмжээст хэвийн тархалт* гэнэ. X санамсаргүй векторыг μ дундаж утгын вектор болон Σ ковариацийн матриц гэсэн параметрүүд бүхий олон хэмжээст хэвийн тархалттай гэдгийг $X \sim N(\mu, \Sigma)$ гэж тэмдэглэнэ.

Хоёр хэмжээст хэвийн тархалт

ρ корреляцтай X_1 ба X_2 санамсаргүй хувьсагчдаас тогтох $X = (X_1, X_2)^{\mathrm{T}}$ 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

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

болно.

$\mu = (5, -4)^{\mathrm{T}}$ 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.

$\mu = (\mu_1, \mu_2)^{\mathrm{T}} = (5, -4)^{\mathrm{T}}$ 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 $\sigma_1 = 2$, $\sigma_2 = 1$, $\rho = -1/2$ гэж олдоно. Иймд нягт нь

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

болно. Харин графикийг нь дараагийн слайд дээр байгуулж үзүүлнэ.

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

X_1	X_2	
	-1	1
0	0.1	0.0
1	0.5	0.4

(a) Хамтын тархалт

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

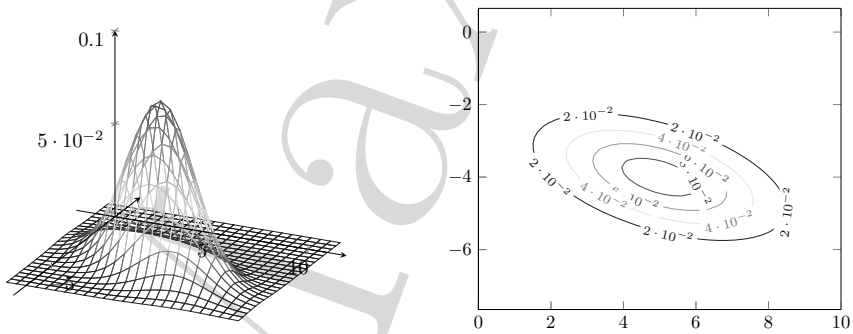
(b) Хамтын тархалт ба тухайн тархалт

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

(c) Тухайн тархалт

Хүснэгт 12: Хамтын тархалт ба тухайн тархалт

Ковариацийн матриц тархалтад нөлөөлөх байдал



нягтын функцийн график нягтын функцийн түвшний шугам

Зураг 61: Хоёр хэмжээст хэвийн тархалтын нягт, хувьсагчид хамааралтай бөгөөд дисперс нь ялгаатай үед

Тухайн тархалт

X_1 болон X_2 санамсаргүй векторууд хамтдаа хэвийн тархалттай бол

$$X_1 \sim N(\mu_1, \Sigma_{11}) \quad X_2 \sim N(\mu_2, \Sigma_{22})$$

буюу олон хэмжээст хэвийн тархалтын тухайн тархалт нь мөн л хэвийн тархалт байна. Энд

$$\mu_1 = E(X_1) \qquad \Sigma_{11} = \mathrm{cov}(X_1, X_1)$$
$$\mu_2 = E(X_2) \qquad \Sigma_{22} = \mathrm{cov}(X_2, X_2)$$

Нөхцөлт тархалт

$X_2 = x_2$ үеийн X_1 санамсаргүй векторын нөхцөлт тархалт нь

$$(X_1|X_2 = x_2) \sim N[\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}]$$

буюу

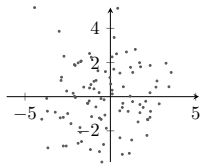
$$E(X_1|X_2 = x_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

нөхцөлт математик дундаж болон

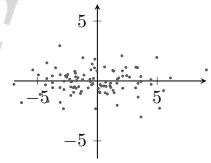
$$\text{cov}(X_1|X_2 = x_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

нөхцөлт ковариацийн матриц бүхий хэвийн тархалт байна. Дээрх томьёонуудад $\Sigma_{12} = \text{cov}(X_1, X_2)$ гэж тэмдэглэсэн бөгөөд ковариацийн матрицын чанар ёсоор $\Sigma_{21} = \Sigma_{12}^{\text{T}}$ юм.

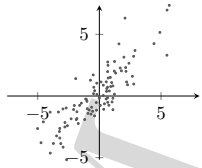
$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.8 \\ -0.8 & 2 \end{pmatrix}\right]$ бол X_2 санамсаргүй хувьсагчийн нөхцөл дэх X_1 санамсаргүй хувьсагчийн тархалтыг ол.



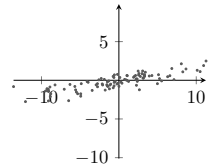
$D(X_1) = D(X_2)$, $\rho(X_1, X_2) = 0$



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



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



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

Зураг 62: Хоёр хэмжээст хэвийн тархалттай санамсаргүй утгууд

Жишээний хувьд $\mu_1 = 0$, $\mu_2 = 0$, $\Sigma_{11} = 1$, $\Sigma_{12} = \Sigma_{21} = -0.8$, $\Sigma_{22} = 2$ ба нөхцөлт математик дундаж нь

$$E(X_1|X_2 = x_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) = -0.8 \cdot \frac{1}{2} \cdot x_2 = -0.4x_2$$

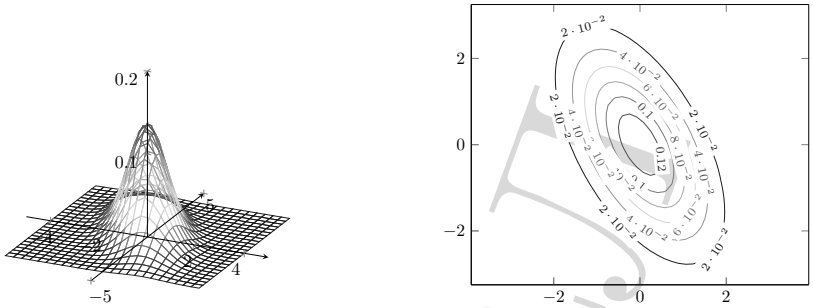
харин нөхцөлт ковариац нь

$$\text{cov}(X_1|X_2 = x_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 1 - (-0.8)\frac{1}{2}(-0.8) = 0.68$$

гэж олдоно. Иймд $X_2 = x_2$ үеийн X_1 хувьсагчийн нөхцөлт тархалт нь $N(-0.4x_2, 0.68)$ буюу

$$f(x_1|x_2) = \frac{1}{\sqrt{2\pi 0.68}} \exp\left\{-\frac{(x_1 + 0.4x_2)^2}{2(0.68)}\right\}$$

болно.



нягтын функцийн график нягтын функцийн түвшний шугам

Зураг 63: X_1 ба X_2 санамсаргүй хувьсагчдын хамтын тархалт

3 Регрессийн шугаман загвар

Нөхцөлт математик дундаж ашигласан аппроксимац

X_2 санамсаргүй векторын тусламжтай X_1 санамсаргүй векторын утгыг олох

$$X_1 = h(X_2) + U$$

аппроксимац авч үзье. Энд U нь аппроксимацийн "алдаа" буюу X_2 хувьсагчаар үл тайлбарлагдах үлдэгдэл юм. Хэрэв ийм аппроксимацийн алдааг

$$\text{MSE} = E\left\{[X_1 - h(X_2)]^{\mathrm{T}}[X_1 - h(X_2)]\right\}$$

дундаж квадрат алдаагаар хэмжинэ гэвэл нөхцөлт математик дундаж ашигласан

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

аппроксимац хамгийн сайн нь болдог.

Чанар 10. $E(X_1|X_2)$ нь X_2 хувьсагчийн тусламжтай X_1 хувьсагчийг аппроксимацлах бүх $h(X_2)$ функц дундаас хамгийн бага дундаж квадрат алдаатай нь юм.

Баталгаа $\{X_2 = x_2\}$ нөхцөлд $h(X_2)$ функц ердийн тогтмол болно. Иймд $\text{MSE} = E\left\{[X_1 - h(X_2)]^{\mathrm{T}}[X_1 - h(X_2)]\right\}$ дундаж квадрат алдааг минимумчлах бодлого нь $\text{MSE} = E\left\{[X_1 - c]^{\mathrm{T}}[X_1 - c]\right\}$ алдааг хамгийн бага болгох c тогтмолыг олох гэсэн бодлого руу шилжинэ. Үүнийг c аргументын хувьд дифференциалчлаад тэгтэй тэнцүүлбэл $E\{2[X_1 - c]\} = 0$ тэгшитгэл зохиогдох бөгөөд шийд нь $c = E(X_1)$ болно. Эцэст нь $\{X_2 = x_2\}$ нөхцөл тавьсанаа анхаарвал $h(X_2) = E(X_1|X_2)$ болно. □

Нөхцөлт математик дундаж ашигласан аппроксимацийн бусад чанар

$X_1 = E(X_1|X_2) + U$ аппроксимац дараах чанартай.

Чанар 11.

1. $E(U|X_2) = 0$

2. $E(U) = 0$
3. $\text{cov}(E(X_1|X_2), U) = 0$

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

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

чанар ашиглавал

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

болж тус чанар батлагдана. □

Регрессийн шугаман загвар

Өмнө үзсэнчлэн хэрэв X_1 ба X_2 санамсаргүй векторууд олон хэмжээст хэвийн тархалттай бол $E(X_1|X_2 = x_2)$ нь X_2 хувьсагчийн x_2 утгаас шугаман байдлаар хамаарсан функц болно. Иймд X_1 хувьсагчийг X_2 аргументтай шугаман функцээр аппроксимацлах боломжтой. Ийм загварыг *регрессийн шугаман загвар* гэнэ.

$$\begin{aligned}
X_1 &= E(X_1|X_2) + U \\
&= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) + U \\
&= \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12}\Sigma_{22}^{-1}X_2 + U \\
&= \beta_1 + BX_2 + U
\end{aligned}$$

Энд $B = \Sigma_{12}\Sigma_{22}^{-1}$, $\beta_1 = \mu_1 - B\mu_2$ гэж тэмдэглэв. Хэрэв X_1 ба X_2 нь скаляр бол тус загвар энгийн $X_1 = \beta_1 + \beta_2 X_2 + U$ хэлбэртэй болно. Харин илүү олон хувьсагчтай жишээлбэл $X = (X_1, X_2, X_3)^{\mathrm{T}}$ үед $X_1 = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + U$ загвар гарна.

Детерминацийн коэффициент

Регрессийн шугаман загварын зүгээс харвал X_1 скаляр хувьсагчийн дисперс дараах байдлаар задарна.

$$\underbrace{D(X_1)}_{\text{нийт дисперс}} = \underbrace{D(\beta_1 + BX_2)}_{\text{тайлбарлагдах дисперс}} + \underbrace{D(U)}_{\text{үл тайлбарлагдах дисперс}}$$

Регрессийн шугаман загварын хамааран хувьсагчийн дисперсэд эзлэх загвараар тайлбарлах дисперсийн хувийг тус загварын *детерминацийн коэффициент* гэдэг.

$$\rho^2 = \frac{D(\beta_1 + BX_2)}{D(X_1)} = \frac{D(BX_2)}{D(X_1)} = \frac{\operatorname{cov}(BX_2, BX_2)}{D(X_1)} = \frac{B\operatorname{cov}(X_2, X_2)B^{\mathrm{T}}}{D(X_1)}$$
$$= \frac{B\Sigma_{22}B^{\mathrm{T}}}{D(X_1)} = \frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}[\Sigma_{12}\Sigma_{22}^{-1}]^{\mathrm{T}}}{D(X_1)} = \frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{12}^{\mathrm{T}}}{D(X_1)} = \frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{D(X_1)}$$

$\Sigma = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 2 \end{pmatrix}$ бол $X_1 = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + U$ шугаман загварын детерминацийн коэффициентыг ол.

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

буюу $\Sigma_{11} = D(X_1) = 1$, $\Sigma_{12} = \begin{pmatrix} 2 & 1 \end{pmatrix}$, $\Sigma_{21} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, $\Sigma_{22} = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$ болж улмаар

$$\rho^2 = \frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{D(X_1)} = \frac{5}{6} \approx 0.833$$

утга олдоно.

Нөхцөлт корреляц ба нөхцөлт үл хамаарал

Сая үзсэн

$$\operatorname{cov}(X_1|X_2 = x_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

нөхцөлт ковариацийн матриц нь ковариацийн матриц л тул үүнээс корреляцийн коэффициент олж улмаар корреляцийн матриц байгуулж болно. Энэхүү нөхцөлт ковариацийн матрицаас олдох *нөхцөлт корреляц* нь нөхцөлд авсан хувьсагчийн нөлөөг тооцсон үеийн нөгөө хоёр хувьсагчийн корреляц юм. Хэрэв тус хоёр хувьсагч угтаа хамааралгүй буюу хөндлөнгийн бусад хувьсагчдын нөлөөгөөр холбогдож байсан бол нөхцөлт корреляц нь тэгтэй тэнцүү гарна. Иймхүү ердийн корреляц нь тэгээс ялгаатай атлаа нөхцөлт корреляц тэгтэй тэнцүү гарах нь өмнө үзсэн нөхцөлт үл хамаарал байгааг илтгэж буй явдал болно.

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

X_1 ба X_2 хувьсагчдын нөхцөлт бус ердийн корреляц $\rho(X_1, X_2) = \dfrac{2}{\sqrt{3 \cdot 5}} \approx 0.516$

буюу тэгээс ялгаатай тул эдгээр нь хамааралтай. Харин нөхцөлт корреляц нь

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \left(\begin{array}{cc|c} 3 & 2 & 2 \\ 2 & 5 & 2 \\ \hline 2 & 2 & 2 \end{array}\right) \quad \Sigma_{11} = \begin{pmatrix} 3 & 2 \\ 2 & 5 \end{pmatrix} \quad \Sigma_{12} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \quad \Sigma_{22} = (2)$$

$$\operatorname{cov}(X_1|X_2 = x_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

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

Тухайн корреляц

Хоёр хувьсагч дээрх хөндлөнгийн өөр бусад хувьсагчдын нийтлэг нөлөөг зайлуулсаны дараах үлдэгдэл хэсэг хоорондын корреляцийг *тухайн корреляц* гэдэг.

$$\rho(X_1, X_2 \| X_3) = \rho(U_1, U_2)$$

Энд $U_1 = X_1 - (\beta_1 + B_1 X_3)$ ба $U_2 = X_2 - (\beta_2 + B_2 X_3)$ буюу шугаман загварын үлдэгдэл юм.

Нөхцөлт корреляц ба тухайн корреляц

Тухайн корреляц дахь хөндлөнгийн нийтлэг нөлөөг зайлуулах үйлдэлд шугаман загвар ашигладаг. Үүнээс улбаалан хэрэв санамсаргүй хувьсагчдын нөхцөлт математик дундаж шугаман хэлбэртэй бол тэдгээрийн тухайн корреляц нь нөхцөлт корреляцийн дундаж утга[20]тай тэнцүү болдог.

Олон хэмжээст хэвийн тархалтын хувьд нөхцөлт математик дундаж нь шугаман хэлбэртэй тул тус тархалттай санамсаргүй хувьсагчдын нөхцөлт корреляц ба тухайн корреляц хоёр тэнцүү байна. Ерөнхийдөө нөхцөлт математик дундаж нь шугаман хэлбэртэй байдаггүйг санавал энэхүү хоёр корреляц тархалт бүрийн хувьд эквивалент байх албагүй юм.

Ковариацийн матрицын урвуу ашиглаж тухайн корреляц олох

Тухайн корреляцийг ковариацийн матрицын урвуугийн тусламжтай олж болдог. Хамтдаа олон хэмжээст хэвийн тархалттай $X_1, \ldots, X_p$ санамсаргүй хувьсагчдын i болон j дүгээр хувьсагчдын хоорондох тухайн (i болон j дүгээр хувьсагчдаас бусад хувьсагчдын нөлөөг зайлуулсан) корррреляцийг олохдоо $X_1, \ldots, X_p$ хувьсагчдын ковариацийн матрицын урвуугаас шууд олсон корреляцийн коэффициентыг эсрэг тэмдэгтэйгээр авна.

$$\Sigma = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 2 & 5 & 2 & -1 \\ 1 & 2 & 2 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix}$$

[20]Ерөнхийдөө нөхцөлт ковариац нь нөхцөлт математик дундаж шиг нөхцөлд буй хувьсагчаас хамаарсан санамсаргүй хувьсагч юм. Харин олон хэмжээст хэвийн тархалтын хувьд энэ нь тогтмол байдаг.

тохиолдолд $\rho(X_1, X_2 \| X_3, X_4)$ тухайн корреляцийг ковариацийн матрицын урвуу ашиглаж ол.

$$\Sigma^{-1} = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 2 & 5 & 2 & -1 \\ 1 & 2 & 2 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix}^{-1} = \frac{1}{2} \begin{pmatrix} 21 & -7 & -5 & -3 \\ -7 & 3 & 1 & 1 \\ -5 & 1 & 3 & 1 \\ -3 & 1 & 1 & 1 \end{pmatrix}$$

$$\rho(X_1, X_2 \| X_3, X_4) = -\frac{-\frac{7}{2}}{\sqrt{3/2}\sqrt{21/2}} = \frac{\sqrt{7}}{3} \approx 0.882$$

Лекц X

Хамааралгүй санамсаргүй хувьсагчдын нийлбэрийн тархалт

Түүх бол амьдарч буй хүмүүсийнх нь нийлбэр юм. — Бен Орлин

1 Хамааралгүй санамсаргүй хувьсагчдын нийлбэрийн тархалт

Хамааралгүй хувьсагчдын нийлбэрийн тархалт

$Z = X + Y$ гэе.

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

Бүтэн магадлалын томьёоны янз бүрийн хэлбэрүүд

Бүтэн магадлалын томьёоны дараах хэлбэрүүдийг өмнө үзсэн.

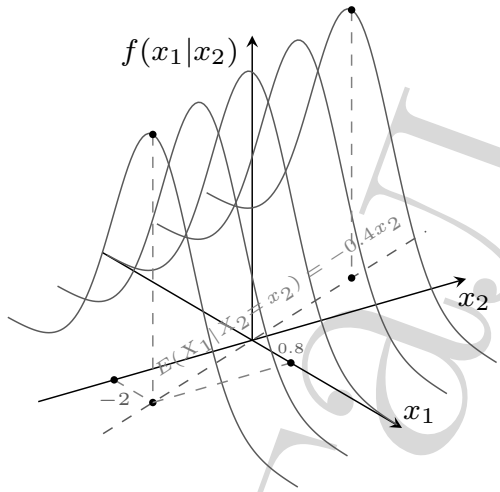
- $f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) dy$
- $P(A) = \sum_i P(A|B_i) P(B_i)$

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

Хамааралгүй хувьсагчдын нийлбэрийн тархалт ба хуниас

X болон Y хамааралгүй байг.

$$F_{X+Y}(z) = P(X + Y < z)$$



Зураг 64: $f(x_1|x_2) = \dfrac{1}{\sqrt{2\pi 0.68}} \exp\left\{-\dfrac{(x_1+0.4x_2)^2}{2(0.68)}\right\}$ нөхцөлт нягтын муруй

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

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

Үүнийг f_X болон f_Y тархалтуудын *хуниас* гэнэ.

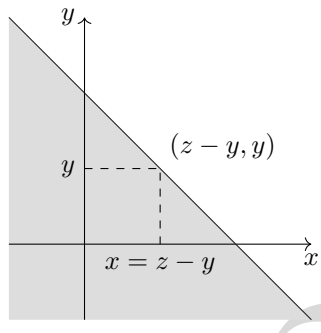
X болон Y хувьсагчид $(0,1)$ завсарт жигд тархалттай бөгөөд хамааралгүй бол $X+Y$ нийлбэрийн тархалтыг олъё.

- $0 \le Z \le 1$ тохиолдол

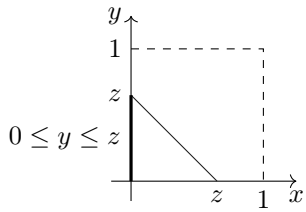
$$f_{X+Y}(z) = \int_0^z \underbrace{f_X(z-y)}_{1}\underbrace{f_Y(y)}_{1}dy = z$$

- $1 \le Z \le 2$ тохиолдол

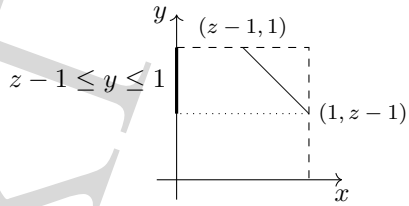
$$f_{X+Y}(z) = \int_{z-1}^{1} 1dy = 2-z$$



Зураг 65: $Z = X + Y$ санамсаргүй хувьсагчийн тархалтын функцтэй холбогдох үзэгдэл



(a) $0 \le Z \le 1$ байх тохиолдол



(b) $1 \le Z \le 2$ байх тохиолдол

Зураг 66: Жигд тархалттай хувьсагчдын $Z = X + Y$ нийлбэр

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

$0 \le X + Y \le z$ байх тул

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

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

Дискрет тархалтуудын хуниас

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

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

$$= \sum_k f_X(z-k) f_Y(k)$$

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

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

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

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

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

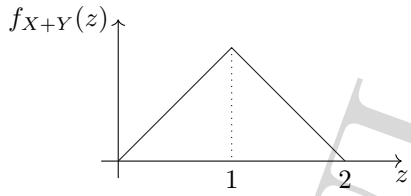
Мөн $n \to \infty$ үед энд ямар нэг асимптот буюу хязгаарын тархалт оршин байх уу?

Жигд тархалттай, хамааралгүй хувьсагчдын нийлбэр

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

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

[21] independent identically distributed

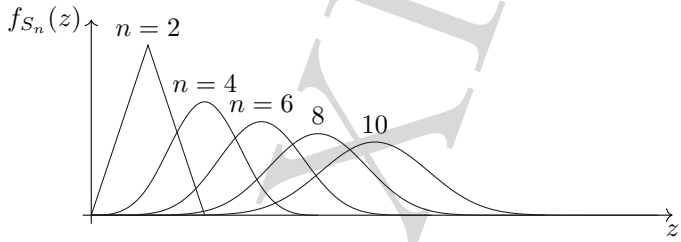


Зураг 67: $(0,1)$ завсарт жигд тархалттай хамааралгүй хувьсагчдын $X+Y$ нийлбэрийн тархалт

Илтгэгч тархалттай, хамааралгүй хувьсагчдын нийлбэр

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

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



Зураг 68: $U(0,1)$ тархалттай, хамааралгүй хувьсагчдын нийлбэрийн тархалт

Процесс ба санамсаргүй түүвэр

$$X_1, \ldots, X_n$$

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

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

дундаж авч үзье. $X = (X_1, \ldots, X_n)$ түүврийн хувьд $\bar{S}_n$ нь түүврийн дундаж гэх $\bar{X}$ статистиктай адил юм. Ийнхүү $n \to \infty$ үед статистик түүврийн дундаж бас S_n хувьсагчийн тархалтыг олох буюу X_i хувьсагчдын (нэг ижил) тархалтын n давхар хуниас олох явдал нь $\bar{S}_n$ хувьсагчийн асимптот буюу хязгаарын тархалтыг олох уруу шилжинэ.

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

$\bar{S}_n$ хувьсагчийн математик дундаж

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

$$= \frac{E(X_1) + \ldots + E(X_n)}{n} = \frac{\mu + \ldots + \mu}{n} = \mu$$

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

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

Энд нийлбэрийн дундаж квадрат хазайлтыг задлахдаа хувьсагчид хамааралгүй болохыг ашиглав.

$\bar{S}_n$ хувьсагчийн асимптот тархалт

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

Дараах функцийг момент үүсгэгч функц гэдэг.

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

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

Момент үүсгэгч функцийн зарим чанар

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

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

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

болно. Үлдэх 3 дугаар чанарыг ашиглавал

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

хувьсагчийн момент үүсгэгч функц

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

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

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

$$M_{X_1}(t) = E[e^{tX_1}] = 1 + tE(X_1) + \frac{t^2}{2}E(X_1^2) + t^2 h(t)$$

$$= 1 + \frac{t^2}{2} + t^2 h(t), \qquad \text{энд } t \to 0 \text{ үед } h(t) \to 0$$

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

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

үр дүнд хүрнэ.

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

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

Хязгаарын гол теорем

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

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

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

байна.

Лекц XI
Хамааралтай хувьсагчдын дараалал, Марковын хэлхээ

> Бид санамсаргүй процессын үр дүнд тогтмол байдал бий болно гэж хүлээдэггүй бөгөөд зүй тогтол мэт зүйлийг олж илрүүлсэнээр уг процесс үнэхээр санамсаргүй гэсэн санааг тэр дор нь үгүйсгэж орхидог. Санамсаргүй процесс нь хүмүүст уг процесс санамсаргүй биш гэсэн итгэл төрүүлэхүйц олон дараалллыг бий болгодог. — Даниел Канеман

1 Хамааралтай хувьсагчдын дараалал

Хамааралтай хувьсагчдын дараалал

Бернуллийн болон Пуассоны процесс "ой санамжгүй" өөрөөр хэлбэл өнгөрсөн үеийн мэдээллээс хэтийн ирээдүйг урьдчилан прогнозлох боломжгүй юм. Үүнийг санамсаргүй хувьсагчдын дараалал дахь дурын X_i болон X_j хоёр хувьсагч хамааралгүй гэж томьёолж байсан. Гэвч практикт жишээлбэл автомат удирдлага, харилцаа холбоо, дохио боловсруулалт, аж үйлдвэр, эдийн засаг дахь стохастик динамик системүүдийг илэрхийлэх X_{n+1} хувьсагч өмнөх

$X_n, X_{n-1}, \ldots, X_0$ 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 $1, \ldots, n, \ldots$ 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 $S = \{1, \ldots, m\}$ болно. Иймд $X_n \in S$ байна.

[Бямбажав Д., Магадлалын онол, математик статистик, 1999] Жил бүрийн хур тунадасны хэмжээ харилцан адилгүй байдаг. Хур тунадасны хамгийн бага түвшинг 1 дүгээр төлөв, удаахыг 2 дугаар төлөв гэх мэтчилэн тэмдэглэе. Ингэвэл 1 дүгээр төлөв нь гантай жилийг харин 4 дүгээр төлөв нь усархаг жилийг заана.

Шилжилтийн магадлал ба Марковын нөхцөл

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

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

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

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

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 магад-*

лалын матриц гэнэ.

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

Чанар 12. Дурын i бүрийн хувьд $\sum_{j=1}^{m} p_{ij} = 1$ буюу тусдаа тархалт байна.

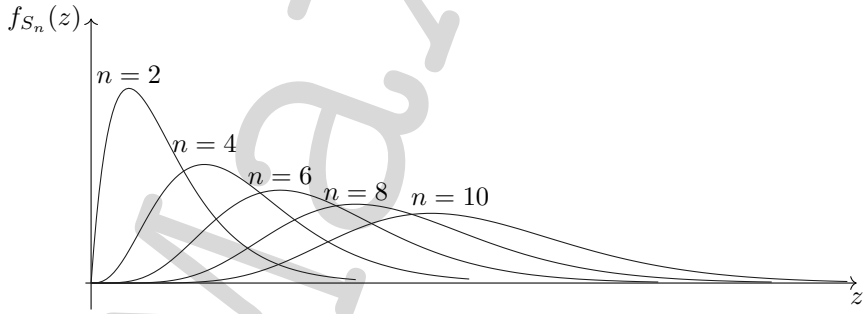
[Бямбажав Д., Магадлалын онол, математик статистик, 1999] Практикаас харахад гантай жилээс усархаг жилд, усархаг жилээс гантай жилд шууд шилждэггүй байна. Тийнхүү дараах шилжилтийн магадлалын матриц олджээ.

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

Жишээний хувьд тухайлбал $i = 1$ дүгээр мөрийн хувьд

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

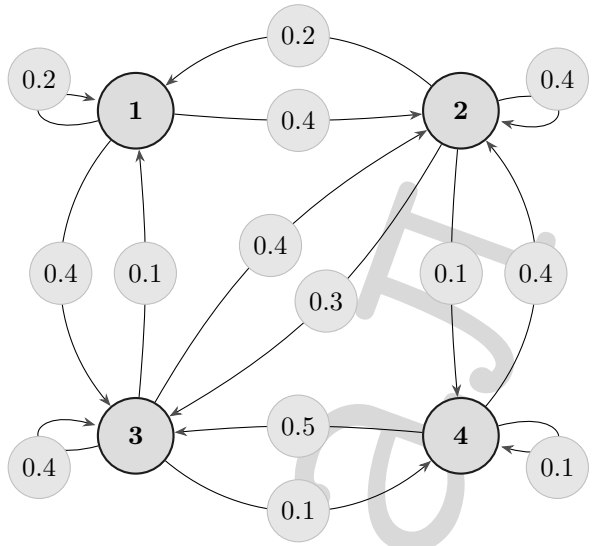
байна.



Зураг 69: Exp(λ) тархалттай, хамааралгүй хувьсагчдын нийлбэрийн тархалт

[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



Зураг 70: Жишээгээр өгсөн шилжилтийн магадлалын матрицад харгалзах чиглэлт граф

нь шилжилтийн магадлал өөрчлөгдөхгүй буюу хэлхээг нэгэн төрлийн гэж үзэх үндэс болно.

(1) INPUT P := (p_{ij})

(2) current_state := 1; the_number_of_drought_years := 0

(3) FOR year FROM 1 TO the_number_of_years

    (a) current_state := new_state(P, current_state)

    (b) IF current_state == 1 THEN the_number_of_drought_years ++

    ENDFOR

(4) RETURN the_number_of_years / the_number_of_drought_years

```
P <- matrix(data = c(0.2, 0.4, 0.4, 0.0, 0.2, 0.4, 0.3, 0.1, 0.1,
    0.4, 0.4, 0.1, 0.0, 0.4, 0.5, 0.1), nrow = 4, byrow = TRUE)
new_state <- function (P, current_state) {
  new_state <- 0; u <- runif(n = 1); cum_prob <- 0
  while (cum_prob < u) {
    new_state <- new_state + 1
    cum_prob <- cum_prob + P[current_state, new_state]
  }
  return(new_state)
}
set.seed(0)
current_state <- 1; n_drought_years <- 0
```

```
for (year in 1:{n_years <- 1000}) {
  current_state <- new_state(P, current_state)
  if (current_state == 1) n_drought_years = n_drought_years + 1
}
cat("Frequency = ", n_years / n_drought_years, "\n")
cat("P(drought) = ", n_drought_years / n_years, "\n")
```

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

Заасан төлвүүдийг дайрах магадлал

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

Дараах харьцааг үржүүлэх дүрэм гэдэг. $P(AB) = P(A|B)P(B)$

Үржүүлэх дүрэм болон Марковын нөхцөл ёсоор

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

болно.

Өмнөхтэй адилаар цааш үргэлжлүүлбэл

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

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

Тодорхой тооны шилжилтээр заасан төлөвт очих боломж

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

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

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

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

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

байдаг бөгөөд үүнийг бүтэн магадлалын томьёо гэдэг.

Бүтэн магадлалын томьёо, нөхцөлт магадлалын томьёо болон Марковын нөхцөл ёсоор

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

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

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

Энэ тохиолдолд өмнөх томьёо

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

хэлбэртэй болно. Шилжилтийн магадлалуудыг орлуулж бодвол

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

үр дүн гарна.

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

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

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

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

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

Жишээний хувьд

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

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

3 Марковын хэлхээний төлвийн стационар тархалт

Төлвийн стационар болон асимптот стационар тархалт

$j = 1, \ldots, m$ төлөв бүрийн хувьд $P(X_n = j) = \pi_j$, өөрөөр хэлбэл систем j төлөвт байх магадлал n буюу хугацаанаас хамаараагүй, эсвэл $\lim\limits_{n\to\infty} P(X_n = j) = \pi_j$ буюу хугацаа өнгөрөх тусам систем j төлөвт байх магадлал тогтворжих явдал заримдаа тохиолддог. Марковын хэлхээний ийм $(\pi_1, \ldots, \pi_m)$ тархалтуудыг харгалзан *стационар тархалт* болон *асимптот стационар тархалт* гэнэ.

Төлвийн стационар тархалт олох

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

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

Жишээний хувьд стационар тархалтыг нь олохын тулд

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

тэгшитгэл бичнэ. Үүнийг бодвол ойролцоогоор

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

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

Төлвийн асимптот стационар тархалт олох

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

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

өөрөөр хэлбэл анхны төлвөөс үл хамаарсан асимптот стационар тархалт олдоно.

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

Тэгэхээр дээрх нөхцөл биелэх үед

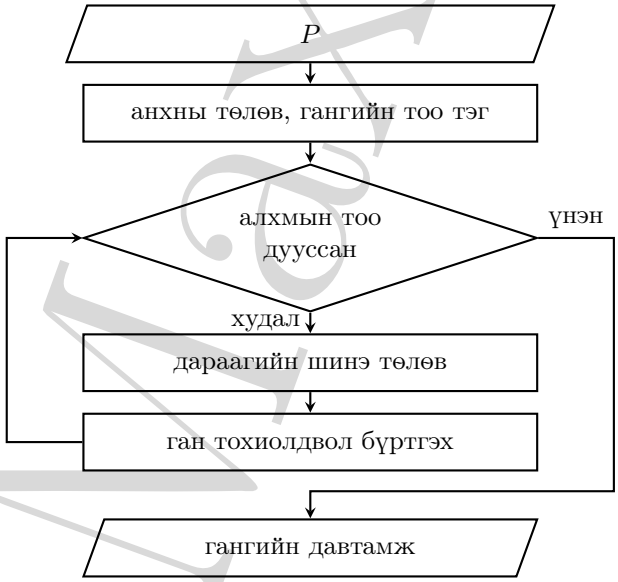
$$P(n) = P^n$$

томьёо ашиглаж асимптот стационар тархалт олох боломжтой.

Жишээний хувьд P^2 матрицын бүх элемент эерэг байсан тул

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

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



4 Марковын хэлхээний төлвийн ангилал

Төлвийн ангилал

Тодорхой алхмын дараа i төлвөөс j төлөвт шилжих боломжтой өөрөөр хэлбэл $p_{ij}(n) > 0$ байх n олддог бол i төлвөөс j төлөв мөрдөнө гээд $i \to j$ гэж тэмдэглэдэг. Харин i болон j төлвүүд бие биеэсээ мөрдөх бол тэдгээрийг *харилцан мөрдөх төлвүүд* гээд $i \leftrightarrow j$ гэж тэмдэглэнэ. Өөр хоорондоо харилцан

мөрдөх төлвүүдийг хамтад нь *үл задрах анги* гэдэг. Марковын хэлхээг үл задрах ангиудад хувааж болох бөгөөд хэрэв хэлхээ ганц үл задрах ангиас тогтож байвал түүнийг *үл задрах хэлхээ* гэнэ.

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

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

буюу i төлөвт төгсгөлөг хугацааны дараа баталгаатай эргэн ирдэг бол тус төлвийг *рекуррент* харин эсрэг тохиолдолд *транзиент* гэнэ. Систем i төлөвт эхний удаа эргэж ирэх дундаж хугацаа буюу

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

нь төгсгөлөг байх албагүй юм.

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

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

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
$$a_i = \sum_{j=1}^{m} p_{ij} \cdot a_j, \quad i = 1, \ldots, m$$
систем тэгшитгэлээр нэг утгатай тодорхойлогдоно.

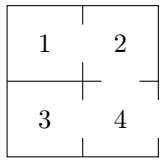
Эцэст нь систем шингээгч төлөвт шилжих дундаж хугацааг авч үзье. Төлөв бүрт харгалзах тус дундаж хугацааг $\mu_1, \ldots, \mu_m$ гэвэл эдгээр нь дараах тэгшитгэлүүдээр нэг утгатай тодорхойлогдоно.

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

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

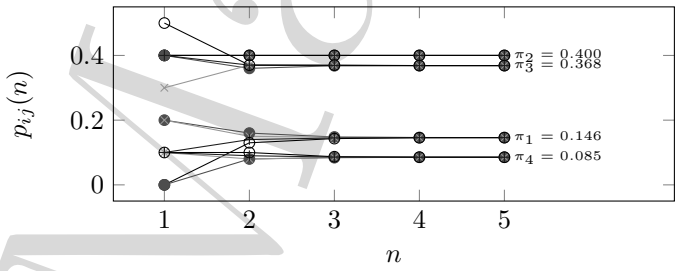
Хулгана хаалгануудын аль нэгийг тэнцүү магадлалтай сонгоно. Муур өрөөнөөсөө гарахгүй. Хэрэв хулгана мууртай өрөөнд орвол муур түүнийг барьж иднэ. Хулгана байшингаас гарч чадвал эргэж орохгүй.

"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 71: Марковын хэлхээний асимптот стационар тархалтын нийлэлт

Шингээгч төлвүүд нь 1 болон 5 дугаар төлөв бөгөөд

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

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. Зогсолтын моментын дээрх тодорхойлолтыг зөвхөн Марковын хэлхээний хувьд бичсэн. Өөрөөр хэлбэл зогсолтын момент нь санамсаргүй процессын онцлог болон процессын зогсох нөхцлөөс шалтгаалан янз бүрээр тодорхойлогдож болно.

Лекц XII
Тархалтын параметрийн статистик үнэлэлт

> Өгөгдөлтэй болохоосоо өмнө онолдох нь гол алдаа юм. — Артур Конан Дойл

1 Тархалтын загварын тухай таамаглал дэвшүүлэх

Тархалтын загварын тухай таамаглал дэвшүүлэх

Тархалт нь ұл мэдэгдэх X санамсаргүй хувьсагчийн эх олонлогоос авсан $X_1, \ldots, X_n$ түүврээр гаргаж авсан өгөгдөлд үндэслэн уг санамсаргүй хувьсагчийн эх олонлогийн тархалтын тухай таамаглал хэрхэн дэвшүүлэхийг авч үзье. Ийнхүү санамсаргүй хувьсагчийн тархалт болон тархалтынх нь шинж чанарыг өгөгдөлд тулгуурлан олж тогтоох нь магадлалаас статистик уруу шилжиж буй явдал юм.

Жишээ буюу бодлого

> 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.

Симуляцын аргаар гарган авсан өгөгдөл

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

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

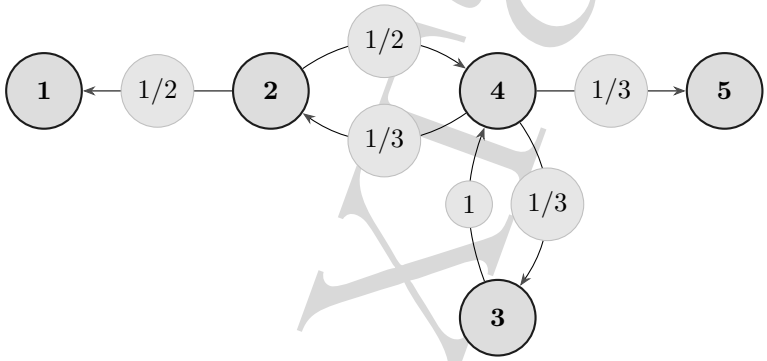
```
print(round(x = Z, digits = 2))
```

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

Гистограмм ба тархалтын тухай таамаглал

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

```
hist(Z)
```



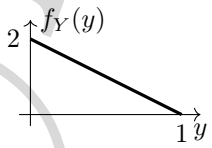
Зураг 72: "Муур ба хулгана" жишээний шилжилтийн магадлалын матрицад харгалзах чиглэлт граф

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 73: Y санамсаргүй хувьсагчийн тархалт

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 шаардлагатай. Статистикт параметрийн үл мэдэгдэх утга олохыг параметр үнэлэх гэдэг. Тархалтын параметр үнэлэхэд хэрэглэдэг моментын арга, хамгийн их үнэний хувийн арга гэх мэтчилэн аргууд байдаг.

Моментын арга

Илтгэгч тархалттай X санамсаргүй хувьсагчийн математик дундаж буюу нэг дүгээр эрэмбийн анхны момент[22] 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.

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

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.

[22] $\alpha_k = E[X^k]$ энд $k = 1$

⚠ Заримдаа параметрийг түүнээс хамаарсан функцээр дамжуулан үнэлдэг.

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

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

буюу $1/\hat{\lambda} = \bar{Z}$ нь *хазайлтгүй үнэлэлт*[23] юм. Иймхүү 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 $Z_1, \ldots, Z_n$ нь энгийн санамсаргүй түүвэр тул эдгээр хувьсагчид хамааралгүй бөгөөд бүгд нэг ижил илтгэгч тархалттай. Иймд үнэлэлтийн дундаж квадрат алдаа

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

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

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

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

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

> 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

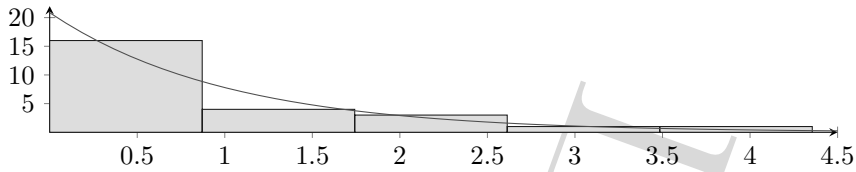
$$P\left(\chi^2_{1-\alpha/2,2n} < 2\lambda n\bar{Z} < \chi^2_{\alpha/2,2n}\right) = 1 - \alpha$$

буюу

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

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

```
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)
```



Зураг 74: Симуляцын аргаар гарган авсан өгөгдлийн гистограмм

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

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

томьёогоор

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

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

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

томьёотой 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 томьёо олдоно.

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

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

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

харьцаагаар хэмждэг.

Өмнөх хоёр тэгшитгэлээс дараах томьёо гарна.

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

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

> 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 $\frac{\chi^2_{\alpha/2,2n}}{\chi^2_{1-\alpha/2,2n}} \approx 1.998$ 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

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

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

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

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 олъё. 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
Байесын үнэлэлт, Хамгийн их үнэний хувьтай үнэлэлт

> Шүүмжлэлт сэтгэлгээ бол идэвхтэй бөгөөд тасралтгүй үйл явц юм. Энэ нь бид бүгдээрээ Байесчууд шиг сэтгэж, шинэ мэдээлэл ирэх тутамд мэдлэгээ шинэчлэхийг шаарддаг. — Даниел Левитин

1 Байесын үнэлэлт

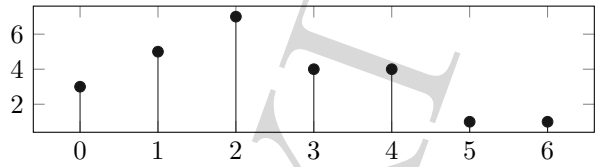
Тархалтын параметр санамсаргүй хувьсагч байж болох нь

Өмнө үзсэн үнэлэлтүүдийн хувьд тархалтын параметрийг тогтмол гэж үзэж байсан. Гэвч бодит байдал дээр тархалтын параметр нь тогтмол бус буюу хувьсдаг цаашилбал санамсаргүйгээр хувьсдаг байж болно. Жишээлбэл нэгэн төрлийн бус Пуассоны процессын эрчмийн параметр нь цаг хугацаанаас яг таг детерминистек байдлаар хамаардаг байх албагүй юм.

Тэгвэл ийм санамсаргүй хувьсдаг параметрийг хэрхэн үнэлэх вэ? Хэрэв параметр санамсаргүй хувьсагч юм бол уг параметрийг судлахын тулд түүний тархалтыг авч үзэх нь зайлшгүй юм. Нөгөө талаас параметрийг судлах ганц барьц буй нь түүвэр билээ. Иймд Θ параметрийн тархалтыг $X = (X_1, \ldots, X_n)$ түүврээс хамааруулан $f_{\Theta|X}(\theta|x)$ байдлаар авч үзнэ.

Параметрийн статистик үнэлэлт

Санамсаргүй хувьсагчийн тархалтын параметрийг үнэлэхийн тулд тус хувьсагчийн эх олонлогоос түүвэр авдаг. Түүврийг хэрэгжүүлж параметрийн тухай мэдээлэл агуулсан өгөгдөлтэй болох бөгөөд өгөгдөл гарган авах арга зам бол туршилт явуулах явдал юм. Туршилт бол судалж буй зүйлийн шинж чанарынх нь тодорхойгүй байдлыг тодорхой болгох зорилготой ажиглалт, хэмжилт зэрэг ямар нэг үйлдэл билээ. Үүний дараа өгөгдөл дээр зохих үнэлэлт ажиллуулж параметрийн утгыг олдог.



Зураг 75: Дискрет хувьсагчийн эх олонлогоос авсан өгөгдлийн давтамж

Санамсаргүй хувьсдаг параметрийг үнэлэх нь

Өмнө тэмдэглэсэн

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

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

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

хэлбэрт шилжүүлээд 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.

Байесын зарчим

Параметр ба түүвэр хоёрынх шиг холбоо хамаарлынх нь эсрэг чиглэлд тавигдсан статистикийн асуудлыг шийдэх арга зам бол Байесын томьёо ашиглах явдал юм.

Дараах томьёог Байесын томьёо гэдэг.

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

Үнэндээ X хувьсагч Y хувьсагчаас хамаардаг боловч Y хувьсагчийг X хувьсагчаас хамааруулан авч үзэх шаардлага цөөнгүй тулгардаг. Холбоо хамаарлын чиглэлийнх нь эсрэг тавигдсан ийм асуудлыг өмнө дурдсанчлан Байесын томьёоны тусламжтай шийддэг. Судалгааны ийм аргачлалыг *Байесын зарчим* гэдэг.

Параметрийн үнэлэлтийн хувьд Байесын зарчим ашиглавал $f_\Theta(\theta)$ приор тархалтыг "сонгох" шаардлага тулгарахыг өмнө дурдсан. Энэхүү сонголтыг мэдээж параметрийн талаарх тухайн үеийн мэдлэгт үндэслэж гаргана. Харин Байесын томьёоны тусламжтай олдох $f_{\Theta|X}(\theta|x)$ постериор тархалт нь параметрийн тархалтын талаарх мэдлэгийг түүвэрт үндэслэн шинэчилж өгөх юм.

Мөн приор тархалтыг сонгож заах нь яг үнэндээ параметрийн тархалтыг таамаглаж буй явдал бөгөөд постериор тархалтыг олох нь уг таамаглалыг бодит баримтын зүгээс шинэчилж буй хэрэг юм.

$$P(\text{таамаглал}|\text{баримт}) = \frac{P(\text{баримт}|\text{таамаглал})P(\text{таамаглал})}{P(\text{баримт})}$$

Цаашилбал энэхүү шинэчлэгдсэн таамаглал нь зөвхөн тухайн үеийн хуучин таамаглал ба баримт хоёрын хүрээнд л хүчин төгөлдөр байх тул хэрэв шинэ баримт гарч ирвэл уг шинэчилсэн таамаглалаа хуучин таамаглал гэж үзээд ахин шинэчлэх хэрэгтэй юм. Чухам иймээс "Өнөөгийн постериор тархалт нь маргаашийн приор болно." гэж Линдлей хэлсэн ажээ.

Байесын үнэлэлт

Байесын зарчимд тулгуурласан үнэлэлтийг *Байесын үнэлэлт* гэдэг. Энэ хичээлээр дараах Байесын үнэлэлтүүдийг үзнэ. Үүнд:

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

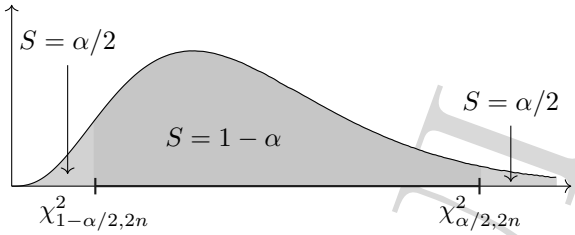
Хамгийн их постериорын үнэлэлт

Өгсөн түүврийн хувьд параметрийн утгыг түүний хамгийн өндөр нягттай буюу хамгийн үнэмшилтэй утгаар авч болох бөгөөд ийм үнэлэлтийг *хамгийн их постериорын үнэлэлт* гэнэ.

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

Хамгийн бага дундаж квадрат алдаатай үнэлэлт

$E(X_2|X_1 = x_1)$ нөхцөлт математик дундаж нь X_1 хувьсагч ашиглаж X_2



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

Зураг 76: Илтгэгч тархалтын эрчмийн параметрийн $1-\alpha$ итгэх магадлалтай итгэх завсар

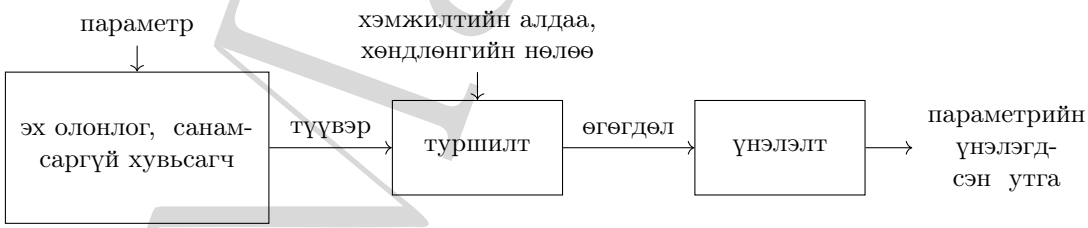
> хувьсагчийг прогнозлох бүх $h(X_1)$ функц дундаас хамгийн бага дундаж квадрат алдаатай нь юм.

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

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

буюу Θ параметрийн бүхий л $h(X)$ үнэлэлт дундаас "хамгийн сайн" нь буюу хамгийн бага дундаж квадрат алдаатай үнэлэлт болно. $\hat{\Theta} = E(\Theta|X = x)$ нь X түүврийн бэхлэгдсэн утга буюу x өгөгдлөөс хамаарсан функц байх тул өгөгдөл ямар байхаас шалтгаалж параметрийн үнэлэгдсэн утга янз бүр болно.

> $\Theta \sim U(0,1)$, $X = 3\Theta + U$, $U \sim U(-1,1)$, $\text{cov}(U,\Theta) = 0$ загварын хувьд $\hat{\Theta} = E(\Theta|X = x)$ үнэлэлтийг ол.



Зураг 77: Параметрийн статистик үнэлэлт

Эхлээд (X,Θ) санамсаргүй векторын хамтын тархалтыг олъё. 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. Үлмаар 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 функцийн чанар ашиглавал

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

байдлаар $c = 0.5$ гэж олдоно. Иймхүү

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

боллоо.

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

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

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

гэж олдоно. Тэгвэл

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

нөхцөлт нягт олдоно. Улмаар

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

үнэлэлт олдоно.

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

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

гэж олдоно. Тэгвэл

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

нөхцөлт нягт олдоно. Улмаар

$$\hat{\Theta}=E(\Theta|X=x)=\int_{x/3-1/3}^{x/3+1/3}\theta\frac{3}{2}d\theta$$
$$=\frac{3}{2}\left[\frac{\theta^2}{2}\right]_{x/3-1/3}^{x/3+1/3}=\frac{x}{3}\quad 1\le x\le 2$$

үнэлэлт олдоно.

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

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

гэж олдоно. Тэгвэл

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

нөхцөлт нягт олдоно. Улмаар

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

үнэлэлт олдоно.

Олсон үнэлэлтээ зураг дээр (тахир шугам) зурж харуулав.

Одоо харин үнэлэлтийг $\hat{\Theta}=E(\Theta|X)=a+bX$ буюу шугаман хэлбэртэй гэвэл чухам ямар үр дүн гарахыг харъя. $Y=a+bX$ шугаман загварын хувьд

$$a=E(Y)-bE(X)$$
$$b=\frac{\mathrm{cov}(X,Y)}{D(X)}$$

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

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

тэгшитгэл бичигдэнэ.

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

$$E(X)=E(3\Theta+U)=3E(\Theta)+E(U)=\frac{3}{2}$$

$$D(X) = D(3\Theta + U) = 9D(\Theta) + D(U) + 2 \cdot 3\,\mathrm{cov}(\Theta, U)$$
$$= 9 \cdot \frac{1}{12} + \frac{2^2}{12} + 0 = \frac{13}{12}$$

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

$$\mathrm{cov}(X, \Theta) = E(X\Theta) - E(X)E(\Theta) = E[(3\Theta + U)\Theta] - \frac{3}{2} \cdot \frac{1}{2}$$
$$= 3E(\Theta^2) + E(U\Theta) - \frac{3}{4} = 3[D(\Theta) + (E(\Theta)^2)] - \frac{3}{4}$$
$$= 3\left[\frac{1}{12} + \frac{1}{4}\right] - \frac{3}{4} = \frac{1}{4}$$

болно.

Ийнхүү

$$\hat{\Theta} = \frac{1}{2} + \frac{3}{13}\left[X - \frac{3}{2}\right] = \frac{2}{13} + \frac{3}{13}X \approx 0.154 + 0.231X$$

хэлбэртэй шугаман үнэлэлт олдлоо. Үүнийг зураг дээр нэмж зурав. Зургаас шугаман загвараар олдсон үнэлэлт хамгийн бага дундаж квадрат алдаатай үнэлэлтээс хэр зэрэг зөрж байгааг харж болно.

Эцэст нь дээрх бодолтыг симуляцын туршилтаар шалгая. Эхлээд авч үзэж буй загварт харгалзах санамсаргүй өгөгдөл гаргаж авна.

```
set.seed(0)
n <- 5000
Theta <- runif("n" = n, "min" = 0, "max" = 1)
U <- runif("n" = n, "min" = -1, "max" = 1)
```

Энэхүү өгөгдөл нь жишээ бодлого дахь $\Theta \sim \mathrm{U}(0, 1)$ ба $U \sim \mathrm{U}(-1, 1)$ нөхцлийг `hist(Theta); hist(U)` харин $\mathrm{cov}(U, \Theta) = 0$ нөхцлийг `cov(U, Theta)` тушаалаар гарах үр дүнг ажиглан шалгаж болно.

```
X <- 3 * Theta + U
```

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 гэдгийг харуулна.

Одоо $-1 \le x \le 1$ үед олдох $\hat{\Theta} = E(\Theta|X = x) = \frac{x+1}{6}$ үнэлэлт үнэн зөв болохыг туршилтаар шалгана. Үүний тулд эхлээд R програмаар байгуулах диаграмм дээр нэмж зураг графикуудыг зөвхөн $-1 \le x \le 1$ бас $0 \le \Theta \le 1$ мужид л зурна гэж хязгаарлах шаардлагатай.

```
clip(-1,1,0,1)
```

$\hat{\Theta} = E(\Theta|X = x) = \frac{x+1}{6}$ үнэлэлтийг өмнөх диаграмм дээр давхарлаж зурахын тулд дараах тушаал өгнө.

```
lines(x = {x <- c(-1,1)}, y = {x+1}/6, col = "blue")
```

Энэ нь $-1 \leq x \leq 1$ нөхцөл хангах өгөгдөл дээрх $\Theta = a + bX$ шугаман загварт харгалзах регрессийн шулуунтай бараг давхцаж буйг

```
abline(reg = lm(formula = Theta ~ X, subset = X < 1))
```

байдлаар дүрслэн харж болно.

Харин $1 \leq x \leq 2$ үед олдох $\hat{\Theta} = E(\Theta|X = x) = \frac{x}{3}$ үнэлэлт зөв гэдгийг шалгахын тулд өмнөхтэй төстэй дараах тушаалуудыг дэс дараалан өгч гаргах үр дүнг нь ажиглана.

```
clip(1,2,0,1)
lines(x = {x <- c(1,2)}, y = x/3, col = "blue")
abline(reg = lm(formula = Theta ~ X, subset = X > 1 & X < 2))
```

Эцэст нь $2 \leq x \leq 4$ үед олдох $\hat{\Theta} = E(\Theta|X = x) = \frac{x+2}{6}$ үнэлэлтийг шалгахад дараах код бичиж ажиллуулна.

```
clip(2,4,0,1)
lines(x = {x <- c(2,4)}, y = {x+2}/6, col = "blue")
abline(reg = lm(formula = Theta ~ X, subset = X > 2 & X < 4))
```

Мөн аналитик тооцооны төгсгөлд олсон параметрийн шугаман үнэлэлтийн коэффициентуудыг ч туршилтаар нягталж болно.

```
fit <- lm(formula = Theta ~ X)
print(coefficients(fit))
```

Эндээс шулууны коэффицентууд ойролцоогоор 0.149 болон 0.232 гэж олдож буй нь өмнө олсон $a = \frac{2}{13} \approx 0.154$ болон $b = \frac{3}{13} \approx 0.231$ аналитик утгуудтай ойролцоо байна. Эцэст нь эдгээр шулуунуудыг өмнө байгуулсан диаграмм дээр нэмж зураад харьцуулан харъя. Үүний тулд дараах хэлбэртэй тушаал өгч болно.

```
clip(-1,4,-1,2)
abline(a = 2/13, b = 3/13, col = "red")
abline(reg = fit)
```

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. $\operatorname{cov}(\tilde{\Theta}, \hat{\Theta}) = 0$
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

$$E(\tilde{\Theta}|X) = E(\hat{\Theta} - \Theta|X) = E(\hat{\Theta}|X) - E(\Theta|X) = E(\hat{\Theta}|X) - E(\Theta|X)$$
$$= E(E(\Theta|X)|X) - E(\Theta|X) = E(\Theta|X) - E(\Theta|X) = \hat{\Theta} - \hat{\Theta} = 0$$

2. Бүтэн дунджийн томьёо болон 1 дүгээр чанар ашиглавал $E(\tilde{\Theta}) = E(E(\tilde{\Theta}|X)) = E(0) = 0$
3. Бүтэн дунджийн томьёо болон 1 дүгээр чанар ашиглавал

$$E(\tilde{\Theta}g(X)) = E[E(\tilde{\Theta}g(X)|X)] = E[g(X)E(\tilde{\Theta}|X)] = 0$$

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

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

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

5. 4 дүгээр чанарыг ашиглавал

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

2 Кошийн тархалт

Кошийн тархалт

Балчир хүү аавтайгаа хамт хөл бөмбөг тоглож байв. Хүү ааваас нь γ зайд байх бөмбөгийг аав руугаа 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.

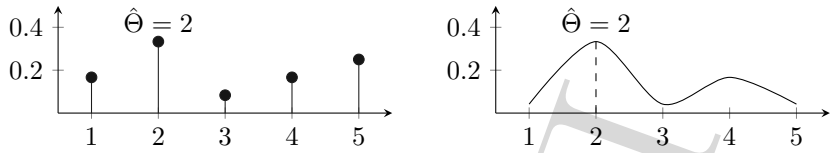
Бодлогын нөхцөл ёсоор $U \sim U(-\pi/2, \pi/2)$ болох бөгөөд X санамсаргүй хувьсагчийн утга U хувьсагчийн утгаас хамаарна. Өөрөөр хэлбэл $X = g(U)$ хувиргалт өгчээ.

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

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

Одоо X хувьсагчийн тархалтын функцийг олъё.

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



(a) Дискрет үед (b) Тасралтгүй үед

Зураг 78: Хамгийн их постериорын үнэлэлт

байх тул

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

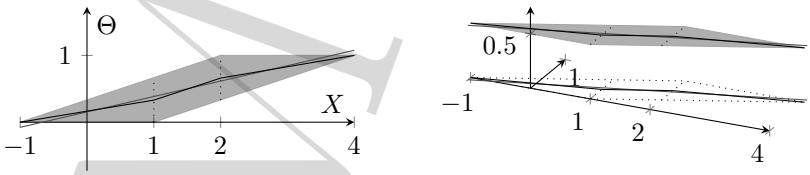
гэж олдоно. Уламжлал авбал дараах хэлбэртэй нягтын функц олдоно.

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

Кошийн тархалт, түүний параметрүүд

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

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



(a) (X, Θ) векторын авах утга (b) (X, Θ) векторын хамтын нягт

Зураг 79: (X, Θ) векторын хамтын тархалт ба $\hat{\Theta} = E(\Theta|X = x)$ үнэлэлт

Кошийн тархалтын зарим чанар болон хэрэглээ

Чанар

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

4. $X, Y \sim N(0,1)$ бөгөөд хамааралгүй бол $\dfrac{X}{Y} \sim \text{Cauchy}(0,1)$

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

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

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

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.

3 Хамгийн их үнэний хувьтай үнэлэлт

Хамгийн их үнэний хувьтай үнэлэлт

Тархалтын параметрийг санамсаргүйгээр хувьсдаг гэж үзээд улмаар Байесын зарчмын тусламжтай судлахад дараах харьцаа хэрэг болсон.

$$\overbrace{f_{\Theta|X}(\theta|x)}^{\substack{\text{постериор}\\ \text{нягт}}} = \frac{\overbrace{f_{X|\Theta}(x|\theta)}^{\substack{\text{үнэний}\\ \text{хувь}}} \overbrace{f_{\Theta}(\theta)}^{\substack{\text{приор}\\ \text{нягт}}}}{f_X(x)}$$

Харин Θ параметрийг буцаагаад тогтмол хэмжигдэхүүн гэж үзвэл энэ нь $P(\Theta = \theta) = 1$ бөхсөн тархалттай "санамсаргүй хувьсагч" болно. Тэгвэл $f_\Theta(\theta)$ приор нягт параметрийн жинхэнэ утга дээр 1, бусад утга дээр 0 утгатай байна. Тэгэхээр постериор нягтыг максимумчлах нь үнэний хувийг максимумчлахтай тэнцүү чанартай буюу хамгийн их постериорын аргаарх үнэлэлт нь хамгийн их үнэний хувьтай үнэлэлт болно. Түүнчлэн параметрийг тогтмол гэсэн тул үүнийг θ гэж тэмдэглэнэ.

Хэрэв θ параметр бүхий $f_X(x,\theta)$ нягттай X санамсаргүй хувьсагчийн эх олонлогоос авсан $X_1, \ldots, X_n$ түүвэр нь энгийн санамсаргүй түүвэр бол $f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n$

үнэний хувь дараах хэлбэртэй болох ба үүнийг $L(X,\theta)$ гэж тэмдэглээд *үнэний хувийн функц* гэнэ.

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

Энд $X_1,\ldots,X_n$ түүврийг энгийн санамсаргүй түүвэр гэсэн нь туршилтуудыг хамааралгүйгээр явуулахыг илтгэнэ. Улмаар

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

буюу үнэний хувийн функцийг хамгийн их утгад нь хүргэх параметрийн утгыг *хамгийн их үнэний хувьтай үнэлэлт* гэнэ.

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 санамсаргүй хувьсагчийн эх олонлогоос авсан $X_1,\ldots,X_n$ түүврийн хувьд үнэний хувийн функц дараах хэлбэртэй байна.

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

$\gamma = 10$ үед дараах байдлаар симуляцлан гаргаж авсан өгөгдөлд тулгуурлан γ параметрийн утгыг буцаан үнэлье.

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

$X_1,\ldots,X_n$ 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.

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

Одоо дээрх логарифм-үнэний хувийн функцээс γ параметрээр уламжлал авч тэгтэй тэнцүүлэн Лагранжийн нөхцөл бичнэ.

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

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

Өмнөх нөхцөлийг дараах байдлаар бичиж болно.

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

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

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

нөхцөл хангана.

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

```
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.

Лекц XIV

Статистик таамаглал шалгах, тархалтын загварын тохирцыг тогтоох

> Би статистикаар үнэнээс бусад бүхнийг баталж чадна. — Жорж Канинг

1 Таамаглал шалгах

Статистик таамаглал

Тархалтын параметрийн талаарх таамаг төсөөллийг *статистик таамаглал* гэнэ. Статистикт нэг нь нөгөөгөө үгүйсгэсэн хоёр таамаглалыг зэрэг авдаг. Тэдний нэгийг тэг, нөгөөг өрсөлдөгч таамаглал гээд харгалзан H_0, H_1 гэж тэмдэглэнэ. Практикт ихэвчлэн тархалтын үл мэдэгдэх параметрийн тухай дараах гурван таамаглалын аль нэгийг авч үздэг.

- Хоёр талт таамаглал

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

- Нэг талт таамаглал

$$H_0 : \theta = \theta_0 \text{ ба } H_1 : \theta < \theta_0 \text{ зүүн өрсөлдөгч таамаглал}$$
$$H_0 : \theta = \theta_0 \text{ ба } H_1 : \theta > \theta_0 \text{ баруун өрсөлдөгч таамаглал}$$

Энд θ нь үл мэдэгдэх параметр, θ_0 нь таамаглаж буй утга юм.

H_0 таамаглалын зүгээс харвал эх олонлог дараах байдлаар үл огтлолцох хоёр хэсэгт хуваагдана.

$$\{\text{Эх олонлог}\} = \{H_0 \text{ үнэн байх олонлог}\} \cup \{H_0 \text{ худал байх олонлог}\}$$

H_0 таамаглалыг шалгахын тулд бидэнд тархалтын эх олонлогийн талаарх мэдээлэл шаардлагатай бөгөөд тийм мэдээлэлтэй болохын тулд түүвэр авдаг билээ. Хэрэв H_0 худал байх олонлог мэдэгдэх бол дараах байдлаар таамаглалд хариулт өгнө.

$$H_0 \text{ таамаглал} \begin{cases} \text{худал} & \text{хэрэв түүвэр} \in \{H_0 \text{ худал байх олонлог}\} \\ \text{үнэн} & \text{хэрэв түүвэр} \notin \{H_0 \text{ худал байх олонлог}\} \end{cases}$$

Гэвч практикт H_0 худал байх олонлог үргэлж мэдэгдэх албагүй. Иймд тус олонлогийг үнэлж олох хэрэгтэй.

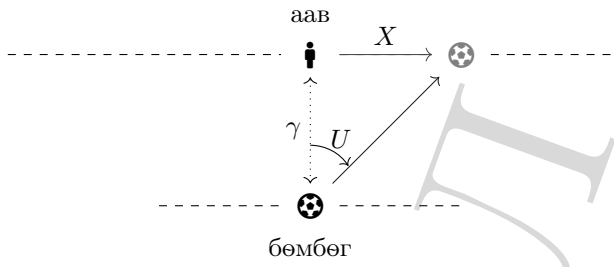
Статистик шинжүүр

Тэг таамаглалыг хүлээн авах эсвэл няцаах шийдвэр гаргах дүрмийг *шинжүүр* гэдэг. Шинжүүрийн зүгээс тэг таамаглал худал байх олонлогийг *шинжүүрийн няцаах муж* гэдэг. Иймд таамаглал шалгахын тулд шинжүүрийн няцаах мужийг олох хэрэгтэй. Шинжүүрийн няцаах муж олдсон үед тэг таамаглалд дараах байдлаар хариулт өгнө.

$$H_0 \text{ таамаглал} \begin{cases} \text{худал} & \text{хэрэв түүвэр} \in \text{шинжүүрийн няцаах муж} \\ \text{үнэн} & \text{хэрэв түүвэр} \notin \text{шинжүүрийн няцаах муж} \end{cases}$$

Таамаглал шалгахад гарах алдаа

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.



Зураг 80: Жишээ бодлогын зураглал

Шинжүүрийн статистик

Шинжүүрийн статистик бол тэг таамаглалын үнэн эсэхийг шалгахаар зохиосон статистик юм.

$$\text{Статистик}: \quad \underset{\Downarrow}{\left\{\begin{matrix}\text{Эх}\\ \text{олонлог}\end{matrix}\right\}} = \underset{\Downarrow}{\left\{\begin{matrix}H_0 \text{ үнэн}\\ \text{байх олонлог}\end{matrix}\right\}} \cup \underset{\Downarrow}{\left\{\begin{matrix}H_0 \text{ худал}\\ \text{байх олонлог}\end{matrix}\right\}}$$

$$\begin{matrix}\mathbb{R} & \text{интервал} & \text{интервал}\end{matrix}$$

түүвэр $\rightarrow$ тоон шулууны цэг

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

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

Тунгалаг тамир романаас сэдэвлэсэн жишээг эргэн авч үзье. Тухайн үед Богдын санд орох өргөл барьцын хэмжээг илэрхийлэх санамсаргүй хувьсагчийг $\lambda = 1$ эрчмийн параметр бүхий илтгэгч тархалттай гэсэн хариу гаргаж байсан. Дараа нь тус аналитик бодолтыг шалгах зорилгоор уг хувьсагчийг симуляцлан 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 өгөгдөл үүсгэсэн. Одоо үүнд тулгуурлаж дээрх хариуг нягтлах буюу 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. Дээрх өгөгдлийн хувьд уг параметрийн үнэлэлт $\hat{\lambda} \approx 0.98$ гэж олдсон нь $\lambda_0 \equiv 1$ утгаас бага тул ийнхүү нэг талт зүүн өрсөлдөгч таамаглалыг тэг таамаглалын эсрэг дэвшүүлэв.

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.

> Exp(λ) 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 $P(X^2 \leq \chi^2_{1-\alpha/2,2n}$ эсвэл $X^2 \geq \chi^2_{\alpha/2,2n}) \leq \alpha$
- Нэг талт зүүн өрсөлдөгчтэй үед $P(X^2 \leq \chi^2_{1-\alpha,2n}) \leq \alpha$
- Нэг талт баруун өрсөлдөгчтэй үед $P(X^2 \geq \chi^2_{\alpha,2n}) \leq \alpha$

хэлбэрт шилжинэ. Иймд шинжүүрийн няцаах муж нь харгалзан

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

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

Шинжүүрийн няцаах муж бэлэн болсон тул таамаглалаа шалгая. Бодлогын нөхцөлд өгсөн 0.68, 0.56, 0.70, 0.14, 4.36, 0.71, 2.09, 0.39, 0.26, 0.45, 1.38, 1.53, 0.28, 2.10, 2.83, 1.03, 1.80, 0.59, 0.06, 1.70, 0.48, 0.71, 0.11, 0.28, 0.17 түүврийн хувьд хэмжээ нь $n = 25$, дундаж нь $\bar{X} \approx 1.016$ тул шинжүүрийн статистикийн туршилтын утга

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

болно. $X^2 = 50.8 \not\leq \chi^2_{1-\alpha,2n} = \chi^2_{0.95,50} \approx 34.764$ буюу шинжүүрийн статистикийн туршилтын утга шинжүүрийн няцаах мужид унахгүй байгаа тул тэг

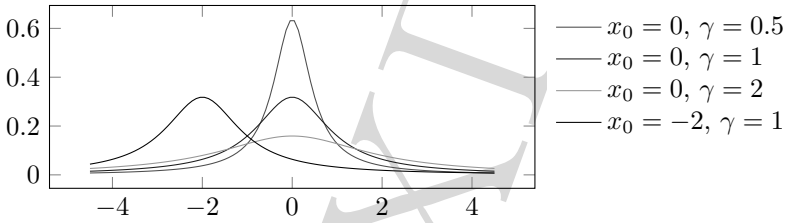
таамаглалыг үл няцаана. Өөрөөр хэлбэл илтгэгч тархалтын эрчмийн параметрийн тухай

$$H_0 : \lambda = \lambda_0 \equiv 1$$
$$H_1 : \lambda < \lambda_0$$

нэг талт зүүн өрсөлдөгчтэй таамаглалыг $\alpha = 0.05$ итгэх түвшинд хүлээн авна.

p-утга

Тэг таамаглалыг үнэн гэж тооцсон тохиолдолд тус таамаглалыг шалгах үед байсантай харьцуулахад тэг таамаглалд илүү эсрэг тэсрэг үр дүн гарах магадлалыг *магадлалын утга* буюу *p-утга* гэнэ. p-утга $< \alpha$ бол H_0 таамаглалыг худал, эсрэг тохиолдолд үнэн гэж дүгнэнэ. Тэг таамаглалыг няцаах сонирхол-



Зураг 81: Кошийн тархалтын нягтын муруй параметрийн янз бүрийн утгад

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

Шинжүүрийн чадал

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

		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 13: Таамаглал шалгахад гарах алдаа

Шинжүүрийн няцаах муж ба итгэх завсар

Хоёр талт өрсөлдөгчтэй параметрийн таамаглалын α итгэх түвшинтэй шинжүүрийн няцаах мужийн гүйцээлт нь тус параметрийн $1 - \alpha$ итгэх магадлал бүхий итгэх завсартай давхацдаг.

Илтгэгч тархалтын эрчмийн параметрийн итгэх завсар дараах хэлбэртэй.

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

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

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

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.

$$H_0 : F_X(x) = F_0(x)$$
$$H_1 : F_X(x) \neq F_0(x)$$

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	$\ldots$	x_k	Нийлбэр
Давтамж	n_1	n_2	$\ldots$	n_k	n

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

- Шинжүүрийн статистик

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

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

$$X^2_{k-1} \sim \chi^2_{k-1}$$

байна.

- Шинжүүрийн няцаах муж, шинжүүрийн няцаах утга

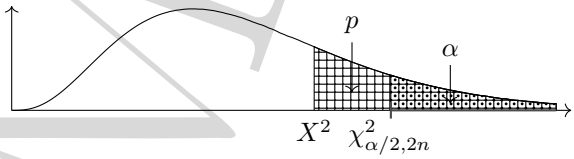
$$X^2_{k-1} \geq \chi^2_{\alpha,k-1}$$

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

"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.



Зураг 82: Илтгэгч тархалтын эрчмийн параметрийн α итгэх түвшинтэй, нэг талт баруун өрсөлдөгчтэй тэг таамаглалын шинжүүрийн няцаах муж ба p-утга

$$X^2_3 = \frac{(134 - 364 \cdot 0.4)^2}{364 \cdot 0.4} + \frac{(196 - 364 \cdot 0.5)^2}{364 \cdot 0.5} + \frac{(24 - 364 \cdot 0.05)^2}{364 \cdot 0.05} + \frac{(10 - 364 \cdot 0.05)^2}{364 \cdot 0.05}$$
$$\approx 7.544$$

Энэ нь $\chi^2_{3,0.05} = 7.815$ буюу шинжүүрийн няцаах утгаас их биш байгаа тул тэг таамаглалыг $\alpha = 0.05$ итгэх түвшинд үл няцаана.

Тус таамаглалыг R програм дээр дараах байдлаар шалгах боломжтой.

```
chisq.test(x = c(134,196,24,10), p = c(0.4,0.5,0.05,0.05))
```

Үр дүн

```
Chi-squared test for given probabilities

data:  c(134, 196, 24, 10)
X-squared = 7.544, df = 3, p-value = 0.05644
```

Дээрх p-утга дараах байдлаар гарна.

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

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.

> 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 $[0, 1]$, $(1, 2]$, $(2, \infty)$ гурван интервалд хуваааж бүлэглэе. Тэгвэл 16, 5, 4 давтамж олдоно. Одоо дээрх интервалуудад харгалзах магадлалуудыг олъё. Тус магадлалуудыг тэг таамаглалын нөхцөлд өөрөөр хэлбэл $\lambda = 1$ эрчмийн параметр бүхий илтгэгч тархалтаар олно.

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

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.

$$X_2^2 = \frac{(16 - 25 \cdot 0.632)^2}{25 \cdot 0.632} + \frac{(5 - 25 \cdot 0.232)^2}{25 \cdot 0.232} + \frac{(4 - 25 \cdot 0.135)^2}{25 \cdot 0.135} \approx 0.229$$

Энэ нь $\chi^2_{2,0.05} = 5.991$ 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

> Онолд зөвхөн зөв эсвэл буруу байх гэсэн хувилбар л бий. Харин загварт гурав дахь боломж байдаг: энэ нь зөв байж болох хэдий ч ач холбогдолгүй.
> — Манфред Айген

1 Үнэний хувийн харьцаат шинжүүр

Үнэний хувийн харьцаат шинжүүр

Байесын зарчимд суурилсан хамгийн их постериорын үнэлэлт нь хамгийн их постериор нягттай буюу хамгийн үнэмшилтэй утгыг тархалтын параметрийн утга гэж үздэг. Статистик таамаглалыг ч ийм байдлаар шалгаж болно. Тухайлбал

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

параметрийн энгийн таамаглалуудын хувьд хамгийн их постериор магадлалтай таамаглалыг үнэн гэх юм. Өөрөөр хэлбэл, $X = (X_1, \ldots, X_n)$ түүврээр цуглуулсан $x = (x_1, \ldots, x_n)$ өгөгдлийн хувьд

$$P(H_1|X = x) > P(H_0|X = x)$$

бол H_0 таамаглалыг няцааж, H_1 таамаглалыг хүлээн зөвшөөрнө.

Байесын зарчмаар

$$P(H_1|X=x) > P(H_0|X=x)$$

нь

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

эсвэл үүнтэй тэнцүү чанартай

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

тэнцэл биш рүү шилжих бөгөөд энэ нь биелж байвал H_0 таамаглалыг няцааж, H_1 таамаглалыг хүлээн зөвшөөрнө.

Харин Байесын бус хувилбар нь

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

үед H_0 таамаглалыг няцааж, H_1 таамаглалыг хүлээн авна. Энд шинжүүрийн няцаах утга c хэдтэй тэнцүү байхаас аль төрлийн алдаа ямар хэмжээгээр гарах нь шалтгаална. c нь Байесын хувилбарт приор магадлалуудын харьцаагаар тодорхойлогдож байсан. Харин Байесын бус хувилбарт

$$P(\text{I төрлийн алдаа}) = P(H_1|H_0) = q\,P(LR=c|H_0) + P(LR>c|H_0) = \alpha$$

харьцаагаар тодорхойлогдоно. Энд α нь итгэх түвшин юм. Мөн энд нэмж хэлбэл $LR=c$ үед H_0 таамглалыг q магадлалтайгаар няцаадаг.

$N(\theta,\sigma^2)$ 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 тооцно.

Таамаглал шалгахын тулд $N(\theta,\sigma^2)$ тархалттай санамсаргүй хувьсагчийн эх олонлогоос $X=(X_1,\ldots,X_n)$ түүвэр авсан гэвэл шинжүүрийн няцаах муж дараах хэлбэртэй болно.

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

Үргэлжлүүлэн хувиргавал дараах тэнцэл биш гарна.

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

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

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

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

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

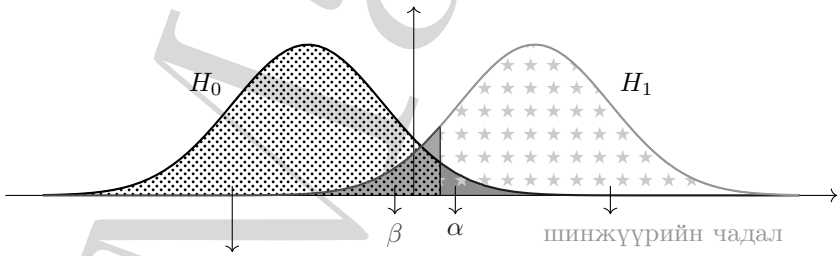
хэлбэртэй шинжүүрийн няцаах муж олдлоо.

Үнэний хувийн харьцаат шинжүүрийн алдааны магадлал

Ерөнхийдөө үнэний хувийн харьцаат шинжүүрийн хувьд I төрлийн алдаа багасахад II төрлийн алдаа ихсэж харин эсрэгээрээ I төрлийн алдаа ихсэхэд II төрлийн алдаа багасна. Үүнийг тухайлан $N(\theta, \sigma^2)$ хэвийн тархалтын дундаж утгын параметрийн тухай

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

таамаглалын хувьд дараах зурагт дүрсэлж үзүүлэв.



H_0 таамаглалыг хүлээн авах магадлал

$\alpha = P(\text{I төрлийн алдаа})$, $\beta = P(\text{II төрлийн алдаа})$

Зураг 83: Нэг талт шинжүүрийн I болон II төрлийн алдаа ба чадал

Үнэний хувийн харьцаат шинжүүрийн асимптот няцаах муж

$H_0 : \theta = \theta_0$ параметрийн таамаглал шалгах байг. Энд $\theta_0 = (\theta_{1,0}, \ldots, \theta_{k,0})$ буюу нийтдээ k ширхэг үл мэдэгдэх параметрийн тухай таамаглал шалгана. Тэгвэл түүврийн хэмжээ n хүрэлцээтэй их бол шинжүүрийн няцаах муж дараах хэлбэртэй байна.

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

Энд $\chi^2_{\alpha,k}$ нь k чөлөөний зэрэгтэй хи-квадрат тархалтын $1-\alpha$ эрэмбийн квантилын утга буюу α хэмжээтэй талбай бүхий тархалтын баруун сүүлний утга юм. Дээрх шинжүүрийн няцаах мужийн хувьд параметрийн хамгийн их үнэний хувийн үнэлэлтийг цор ганц оршин байхыг шаардсан нэмэлт нөхцөл тавьдаг.

2 Регрессийн шугаман загвар

Корреляцын коэффициент ба шугаман хамаарал

Хувьсагчид шугаман хамааралтай үед корреляцын коэффициент үнэмлэхүй утгаараа нэгтэй тэнцүү байдаг.

Олон хэмжээст хэвийн тархалт, корреляцын коэффициент

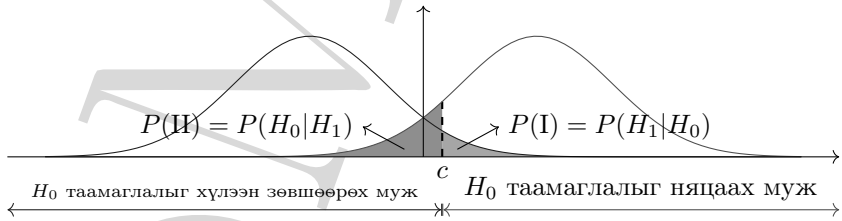
Корреляцын коэффициент тэгээс ялгаатай, үнэмлэхүй утгаараа нэгд ойр үед олон хэмжээст хэвийн тархалттай санамсаргүй хувьсагчдын хамаарлыг шугаман функцээр илэрхийлэх боломжтой.

Шашин	Будда	Христ	Ислам	Бусад	Нийлбэр
Давтамж	134	196	24	10	364

Хүснэгт 14: Сүм хийдийн тоо, шашны төрлөөр, 2018 оны эцэст, YCX

Регрессийн шугаман загвар

X_1 болон X_2 хувьсагчид хамтдаа олон хэмжээст хэвийн тархалттай үед $E(X_2|X_1 = x_1)$ нь X_1 хувьсагчийн x_1 утгаас шугаман байдлаар хамаарсан функц байдаг тул X_2 хувьсагчийг X_1 хувьсагчийн шугаман эвлүүлгийн тусламжтай прогнозлох боломжтой. Иймм загварыг *регрессийн шугаман загвар* гэнэ.



Зураг 84: Үнэний хувийн харьцаат шинжүүрийн алдааны магадлал

Нөхцөлт математик дундаж ашигласан загварын чанар

$X_2 = E(X_2|X_1) + U$ загвар дараах шинж чанартай болохыг өмнө үзсэн.

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 прогнозлох бүх $h(X_1) : \mathbb{R}^r \to \mathbb{R}^{p-r}$ функц дундаас хамгийн бага дундаж квадрат алдаатай ($\text{MSE} = E\left\{(X_2 - h(X_1))^T (X_2 - h(X_1))\right\}$) нь юм.

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

Детерминацын коэффициент

$E(X_2|X_1 = x_1)$ нөхцөлт математик дунджийг $X_2 = E(X_2|X_1 = x_1) + U$ гэж X_2 хувьсагчийг прогнозлоход ашигладаг бол харин $D(X_2|X_1 = x_1)$ нөхцөлт дисперсийг

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

буюу X_2 хувьсагчийн дисперсэд эзлэх тус нөхцөлт дисперсийн хувиар уг прогнозын илэрхийлэх чадварыг хэмжихэд ашиглах бөгөөд ρ^2 хэмжигдэхүүнийг *детерминацын коэффициент* гэнэ. Мөн дээрх прогнозын загварын хувьд дараах харьцаа хүчинтэй байдаг.

$$\underbrace{D(X_2)}_{\text{нийт дисперс}} = \underbrace{D(X_2|X_1 = x_1)}_{\text{тайлбарлагдах дисперс}} + \underbrace{D(U)}_{\text{үл тайлбарлагдах дисперс}}$$

$X_2 = E(X_2|X_1) + U$ загварын хувьд

$$\underbrace{D(X_2)}_{\text{нийт дисперс}} = \underbrace{D(X_2|X_1 = x_1)}_{\text{тайлбарлагдах дисперс}} + \underbrace{D(U)}_{\text{үл тайлбарлагдах дисперс}}$$

байх тул детерминацын коэффициентыг

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

гэж олно.

Регрессийн шугаман загварын параметрийн үнэлэлт

Загварын параметрийг MSE буюу дундаж квадрат алдааг минимумчлах байдлаар үнэлдэг.

Өмнө дурдсанчлан $E(X_2|X_1)$ бол хамгийн бага дундаж квадрат алдаатай.

X_1 болон X_2 хувьсагчид хамтдаа олон хэмжээст хэвийн тархалттай бол $E(X_2|X_1 = x_1) = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 + \Sigma_{21}\Sigma_{11}^{-1}x_1$ байна.

Иймд $X_2 = E(X_2|X_1) + U = a + bX_1 + U$ загварын параметрүүдийн дараах үнэлэлт гарна.

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

Шинэ тэмдэглэгээ
Тэмдэглэгээг хялбаршуулахын тулд X_2 болон X_1 хувьсагчдыг харгалзан Y болон X гэе.

Тэгвэл регрессийн шугаман загварыг дараах байдлаар бичнэ.

$$Y = a + bX + U$$

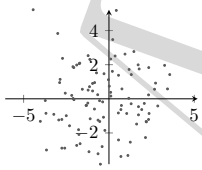
Мөн параметрийн үнэлэлтийн томьёо дараах хэлбэртэй болно.

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

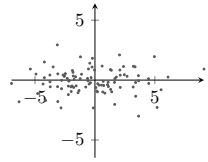
Түүврийн хувьд дээрх томьёонд буй моментуудыг харгалзах түүврийн моментуудаар үнэлж болох тул

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

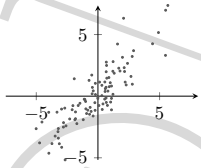
болно. Энд $\bar{X}$ болон $\bar{Y}$ нь түүврийн дундаж, $S^2(X)$ нь түүврийн дундаж квадрат хазайлт, $S^2(X, Y)$ нь түүврийн ковариацын коэффициент юм.



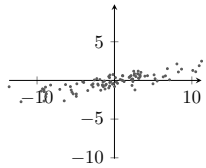
$D(X_1) = D(X_2),\ \rho(X_1, X_2) = 0$



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



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



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

Зураг 85: Хоёр хэмжээст хэвийн тархалттай санамсаргүй утгууд

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

шугаман загварын параметрийн үнэлэлтийн томьёог алдааны квадратуудын нийлбэрийг минимумчлах байдлаар бодож олъё.

$$SSE = (n-1)S^2(U) = \sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} (Y_i - (a + bX_i))^2$$

a болон b хувьсагчдаар тухайн уламжлал авч тэгтэй тэнцүүлбэл

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

систем тэгшитгэл гарна.

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

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

Ингээд үнэлэлтийн дараах томьёо гарна.

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

Ийнхүү бодох аргыг *хамгийн бага квадратын арга* гэнэ.

R ашиглаж регрессийн шинжилгээ хийх

Өгөгдөл оруулах байдал

```
data <- data.frame(
  education = c(11, 12, 11, 15, 8, 10, 11, 12, 17, 11),
  annual.income = c(25, 33, 22, 41, 18, 28, 32, 24, 53, 26)
)
```

Загварын параметр үнэлэх

```
fit <- lm(formula = annual.income ~ education, data = data)
```

Прогноз

```
predict(object = fit, newdata = data.frame(education = c(16, 20)))
```

Нэмэлт шинжилгээ

```
summary(fit)
```

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

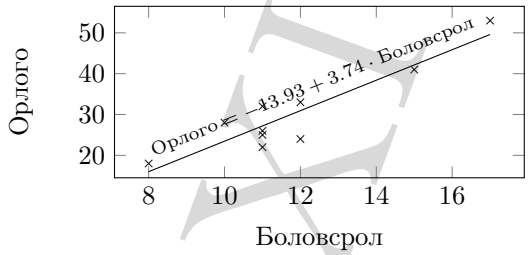
Лекц XVI

Хяналттай машин сургалтад ашиглах зарим статистик арга техник

> Ухаалаг хиймэл оюун гэхээс илүү ёс зүйтэй хиймэл оюун ухаанд анхаарлаа төвлөрүүлээрэй. — Абхижит Наскар

1 Регрессийн шугаман загвар

Өгөгдөл ба бодлого



Зураг 86: Боловсролд зарцуулсан хугацаа (жил) ба жилийн орлого (мян.$)

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")
```

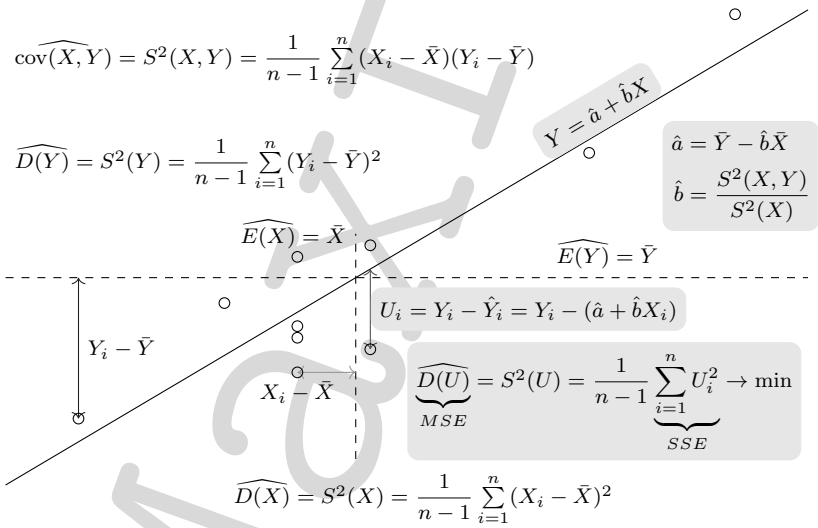
R програм дээрх загварын үнэлгээ болон прогноз

$\log_2(\text{урсгал}) = a + b \cdot \text{он}$ 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))])
```



Зураг 87: Регрессийн шугаман загвар

Прогноз "залгаас" дээрээ огцом үсрэлттэй байгаа тул

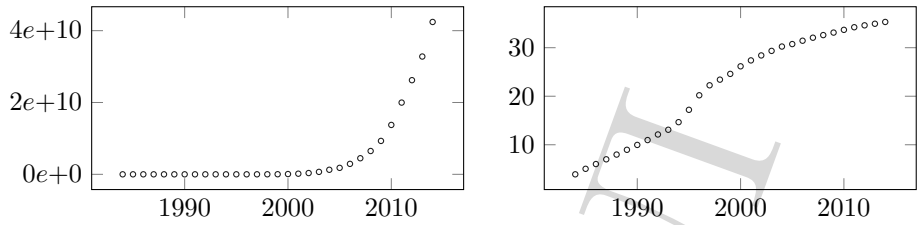
$$\log_2(\text{урсгал}) = a + b \cdot \text{он}$$

загвар хэтийн прогноз гаргахад тохиромж муутай гэж үзнэ. Иймд өөр загвар авч үзье.

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



(a) Анхдагч тоон мэдээ (b) $\log_2$ хувиргалт

Зураг 88: Интернет сүлжээний урсгал, GB/cap, 1984-өөс 2014 он

```
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.

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

загварыг p эрэмбийн *авторегрессийн загвар* гэнэ. Энд U_t нь загварын алдаа юм.

Загварын хэрэглээг өмнөх хэсэгт ашигласан интернет урсгалын мэдээнд тулгуурлан үзье.

Авторегрессийн загварын эрэмбэ тогтоох

Загварын эрэмбэ тогтооход тухайн автокорреляц ашиглана.

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

Авторегрессийн загварын параметрүүдийг үнэлэх

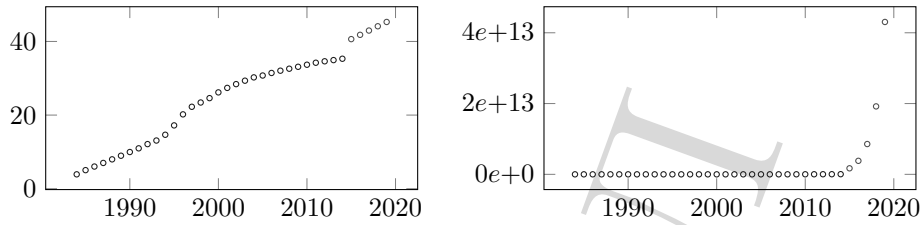
Загварыг хамгийн бага квадратын аргаар яаж үнэлэхийг харуулав.

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

Ийнхүү

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

загвар гарав.



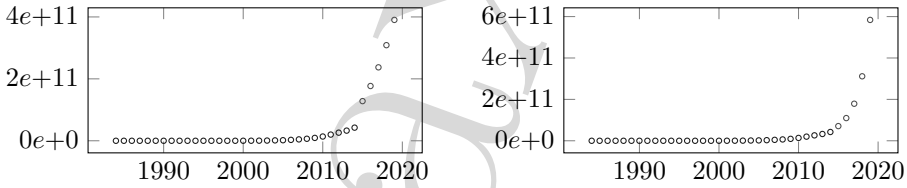
(a) $\log_2$ хувиргалттай (b) хувиргалтгүй

Зураг 89: Интернет сүлжээний урсгал, $\log_2(\text{урсгал}) = a + b \cdot \text{он}$ загварын тусламжтай гаргасан прогноз

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))])
```



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

Зураг 90: Интернет сүлжээний урсгалын хэтийн төлөвийг олон гишүүнт бүхий загваруудаар прогнозлосон нь

3 Ложистик регресс

Ложистик регресс

$Y \sim \text{Ber}(p)$ байх Y хувьсагчийн утгыг X хувьсагчийн тусламжтай прогнозлох зорилго тавья. Хэрэв

$$p = P(Y = 1) = E(Y|X) = a + bX$$

загвар авч үзвэл тэгээс бага эсвэл нэгээс их утгатай "буруу" прогноз гарах боломжтой. Үүнээс зайлсхийхийн тулд $(0, 1)$ завсарт утгатай

$$p = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

ложистик функц ашигладаг.

Ложит загвар

Өмнөх загварыг $\frac{p}{1-p} = e^{a+bX}$ байдлаар бичиж болно. $\frac{p}{1-p}$ *магадлалын харьцаа* нь $(0, \infty)$ засварт утгатай байх бөгөөд 0 болон ∞ нь p магадлалын бага болон их утгад харгалзана. Иймхүү 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(0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,
    0,1,0,0,0,0,1,1,1,0,0,0,1,1,1,0,0,0),
  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.

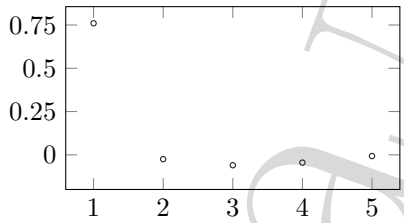
Иймхүү эцэстээ

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

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 91: Интернет сүлжээний урсгал хувьсагчийн автокорреляц

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

$$\underset{k}{\operatorname{argmax}}\, P(C_k|X_1,\ldots,X_p)$$

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

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

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

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

гэж үздэг.

Ийнхүү

$$\underset{k}{\operatorname{argmax}}\, P(C_k) \prod_{i=1}^{p} P(X_i|C_k)$$

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
```