MS-A0504 Todennäköisyyslaskennan ja tilastotieteen peruskurssi

3A Satunnaismuuttujien summa ja keskiarvo

Jukka Kohonen

Matematiikan ja systeemianalyysin laitos
Perustieteiden korkeakoulu
Aalto-yliopisto

Lukuvuosi 2021–2022
Periodi IV

Sisältö

Kahden s-muuttujan summa: odotusarvo ja keskihajonta

Jos X, Y ovat satunnaismuuttujia ja $S = X + Y$, osaamme jo laskea sen

- odotusarvon: $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- varianssin: $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + 2 \cdot \mathrm{Cov}(X, Y) + \mathrm{Var}(Y)$
- keskihajonnan: $\sqrt{\mathrm{Var}}$

Tunnemme siis $X + Y$:n jakauman *paikan* ja *leveyden*. Mutta emme

vielä tunne summan jakauman *muotoa*. Se voi olla hyvin erilainen kuin X:n ja Y:n jakauman muoto (ks. tulevia esimerkkejä).

Jakauman muodosta olisi paljon hyötyä esim. häntätodennäköisyyksien laskemisessa. (Paljon tiukempia rajoja kuin Tšebyšovilla, vrt. viime luento)

Usean s-muuttujan summa

Ennen muotoihin menemistä, todetaan että kolmen satunnaismuuttujan summalle voimme helposti soveltaa kahden muuttujan kaavoja kahdesti.

$$\begin{aligned}\mathbb{E}(X+Y+Z) &= \mathbb{E}((X+Y)+Z) = \mathbb{E}(X+Y)+\mathbb{E}(Z) \\ &= \mathbb{E}(X)+\mathbb{E}(Y)+\mathbb{E}(Z).\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(X+Y+Z) = &\,\mathrm{Var}(X)+\mathrm{Var}(Y)+\mathrm{Var}(Z) \\ &+2\,\mathrm{Cov}(X,Y)+2\,\mathrm{Cov}(X,Z)+2\,\mathrm{Cov}(Y,Z).\end{aligned}$$

Erityisesti, jos X, Y, Z ovat riippumattomat, niin kovarianssit ovat nollia ja

$$\mathrm{Var}(X+Y+Z) = \mathrm{Var}(X)+\mathrm{Var}(Y)+\mathrm{Var}(Z).$$

Yleistys useammalle muuttujalle menee samaan tapaan (ks. luentomoniste).

Kahden satunnaismuuttujan summa: Muoto

Kahden sm:n summa $S = X + Y$ on satunnaismuuttuja, jonka jakauma voidaan määrittää X:n ja Y:n yhteisjakaumasta $f_{X,Y}(x, y)$. Miten?

Kuten minkä tahansa muunnoksen $g(X, Y)$ jakauma:

1. Tarkastellaan parin (X, Y) yhteisjakaumaa.
2. Tutkitaan, mitä arvoja $g(X, Y)$ voi saada.
3. Kullekin mahdolliselle arvolle s selvitetään, *millä* parin (X, Y) arvoilla) se toteutuu.
4. Lasketaan sellaisten vaihtoehtojen tn:t yhteen.

Kohdassa 3 voidaan käydä vaihtoehdot läpi yksitellen tai yrittää löytää yleisempi sääntö.

Esimerkki

Kahden 100-sivuisen nopan summa S voi saada kokonaislukuarvoja välillä $2 \ldots 200$. Tutkitaan (taululla) niiden todennäköisyydet.

Kahden satunnaismuuttujan summa: Jakauman muoto

Summan $X + Y$ jakauma voidaan määrittää X:n ja Y:n yhteisjakaumasta $f_{X,Y}(x, y)$.

$$f_{X+Y}(s) \;=\; \sum_x f_{X,Y}(x,\; s-x)$$

$$f_{X+Y}(s) \;=\; \int_{-\infty}^{\infty} f_{X,Y}(x,\; s-x)\, dx.$$

Jos summan termit X ja Y ovat stokastisesti riippumattomat:

$$f_{X+Y}(s) \;=\; \sum_x f_X(x)\, f_Y(s-x)$$

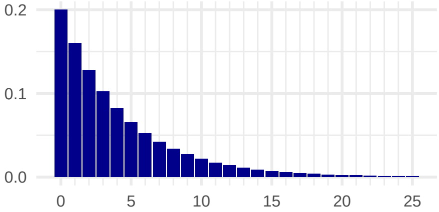
$$f_{X+Y}(s) \;=\; \int_{-\infty}^{\infty} f_X(x)\, f_Y(s-x)\, dx.$$

(Tätä kutsutaan tiheysfunktioiden f_X ja f_Y konvoluutioksi.)

Esimerkki: Kahden geom. sm:n summa

Satunnaismuuttujat X_1 ja X_2 ovat toisistaan riippumattomat ja noudattavat lukujoukon $\{0, 1, 2, \ldots\}$ ns. geometrista jakaumaa parametrina $a = 4/5$ ja tiheysfunktiona

$$f(x) \;=\; (1-a)a^x.$$



(Esimerkki: Heitetään 5-sivuista noppaa kunnes tulee viitonen. “Hukkaheittojen” määrä on geometrisesti jakautunut.)

Määritä satunnaismuuttujan $X_1 + X_2$ jakauma.

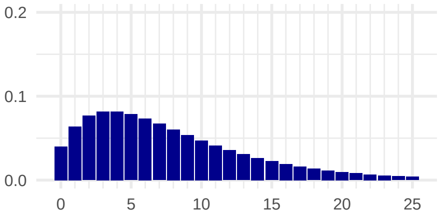
Esimerkki: Kahden geom. sm:n summa

Satunnaismuuttujan $X_1 + X_2$ arvojoukko on $\{0, 1, 2, \dots\}$ ja tiheysfunktio saadaan kaavasta

$$f_{X_1+X_2}(s) = \sum_x f(x) f(s-x) = \sum_{x=0}^{s} (1-a)a^x (1-a)a^{s-x}$$

Summan jakauman tiheysfunktio on

$$f_{X_1+X_2}(s) = (1-a)^2 (s+1) a^s$$



(Esimerkki: Heitetään 5-sivuista noppaa toistuvasti, kunnes on kertynyt *kaksi* viitosta. Yllä ratkaisimme hukkaheittojen määrän jakauman. Se on ns. negatiivinen binomijakauma.)

Usean satunnaismuuttujan summa: muoto

Olkoot X, Y, Z satunnaismuuttujia.
Mikä on summan $X + Y + Z$ jakauma?

Käytetään konvoluutiokaavaa kahdesti.

- Olkoon $U = X + Y$, ja ratkaistaan f_U konvoluutiokaavalla.
- Olkoon $S = U + Z$, ja ratkaistaan f_S konvoluutiokaavalla.

Näin saadaan summan tiheysfunktio täsmälleen. Toistuvat summaukset ja integraalit voivat kyllä olla mutkikkaita.

Monissa erityistapauksissa tämä on kyllä jo tehty, ja summan jakauma on hyvin tiedossa (ja löydät sen kirjallisuudesta tai Wikipediasta). Esimerkkejä:

- geom-jakautuneiden summalla on negatiivinen binomijakauma
- indikaattorimuuttujien summalla on binomijakauma
- eksponenttijakautuneiden summalla on gammajakauma
- ...

Sisältö

Mitä suurten lukujen laki kertoo (ja mitä ei)?

Keskiarvo suuresta määrästä riippumattomia X:n tavoin jakautuneita satunnaislukuja (odotusarvo μ, keskihajonta σ) on suurella todennäköisyydellä likimain

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx \mu.$$

Suurten lukujen laki ei kerro:

- Kuinka tarkka tämä approksimaatio on?
- Miten σ vaikuttaa approksimaation tarkkuuteen?

Approksimaation tarkkuutta voidaan mitata laskemalla

$$\mathrm{SD}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{SD}\left(\sum_{i=1}^{n} X_i\right).$$

Tarvitaan laskukaava summan keskihajonnalle/varianssille.

Summan keskihajonta

Laske $\sigma_{X+Y} = \mathrm{SD}(X+Y)$, kun tunnetaan odotusarvot $\mu_X = 1$ ja $\mu_Y = 1$ sekä keskihajonnat $\sigma_X = 2$ ja $\sigma_Y = 3$.

Ratkaisu

Kovarianssin lineaarisuudesta

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{Cov}(X+Y,\ X+Y) \\
&= \mathrm{Cov}(X,X) + \mathrm{Cov}(Y,X) + \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,Y) \\
&= \mathrm{Var}(X) + 2\,\mathrm{Cov}(X,Y) + \mathrm{Var}(Y),
\end{aligned}$$

joten

$$\mathrm{SD}(X+Y) = \sqrt{\sigma_X^2 + 2\,\mathrm{Cor}(X,Y)\,\sigma_X\sigma_Y + \sigma_Y^2}.$$

Summan keskihajontaa *ei* voi laskea tuntematta korrelaatiota.

- Koska $-1 \le \mathrm{Cor}(X,Y) \le 1$, saadaan yo. kaavasta estimaatit $|\sigma_X - \sigma_Y| \le \mathrm{SD}(X+Y) \le \sigma_X + \sigma_Y$, eli $1 \le \sigma_{X+Y} \le 5$.
- Jos X ja Y ovat riippumattomat, pätee $\mathrm{Cor}(X,Y) = 0$ ja $\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2} = \sqrt{13} \approx 3.6$.

Summan keskihajonta: Monta termiä

Fakta

Satunnaislukujen $X_1, \ldots, X_n$ summan keskihajonta saadaan kaavasta

$$\mathrm{SD}\left(\sum_i X_i\right) = \sqrt{\sum_i \sigma_i^2 + \sum_i \sum_{j \neq i} \sigma_i \sigma_j \rho_{i,j}},$$

missä $\sigma_i = \mathrm{SD}(X_i)$ *ja* $\rho_{i,j} = \mathrm{Cor}(X_i, X_j)$.

Jos $X_1, \ldots, X_n$ ovat riippumattomia ($\rho_{i,j} = 0$) ja samoin jakautuneita ($\mu_i = \mu$ ja $\sigma_i = \sigma$), niin

$$\mathrm{SD}\left(\sum_{i=1}^n X_i\right) = \sqrt{\sum_{i=1}^n \sigma_i^2} = \sqrt{n\sigma^2} = \sigma\sqrt{n}.$$

Summan keskihajonta: Todistus

Kovarianssin lineaarisuudesta

$$
\begin{aligned}
\mathrm{Var}\left(\sum_i X_i\right) &= \mathrm{Cov}\left(\sum_i X_i, \sum_j X_j\right) \\
&= \sum_i \sum_j \mathrm{Cov}(X_i, X_j) \\
&= \sum_i \left(\mathrm{Cov}(X_i, X_i) + \sum_{j \neq i} \mathrm{Cov}(X_i, X_j)\right) \\
&= \sum_i \mathrm{Var}(X_i) + \sum_i \sum_{j \neq i} \mathrm{Cov}(X_i, X_j) \\
&= \sum_i \sigma_i^2 + \sum_i \sum_{j \neq i} \sigma_i \sigma_j \rho_{i,j},
\end{aligned}
$$

joten

$$
\mathrm{SD}\left(\sum_i X_i\right) = \sqrt{\mathrm{Var}\left(\sum_i X_i\right)} = \sqrt{\sum_i \sigma_i^2 + \sum_i \sum_{j \neq i} \sigma_i \sigma_j \rho_{i,j}}.
$$

Summan keskihajonta: Riippumattomat termit

Fakta

Riippumattomien satunnaislukujen $X_1, \ldots, X_n$ summan keskihajonta saadaan kaavasta

$$\mathrm{SD}\left(\sum_{i=1}^{n} X_i\right) \;=\; \sqrt{\sum_{i=1}^{n} \sigma_i^2} \;=\; \sigma\sqrt{n},$$

kun $\sigma_i = \sigma$ kaikilla $i = 1, \ldots, n$.

Todistus.

Tulos seuraa suoraan summan keskihajonnan kaavasta, sillä $\rho_{i,j} = \mathrm{Cor}(X_i, Y_j) = 0$ kaikilla $i \neq j$, kun $X_1, \ldots, X_n$ ovat toisistaan stokastisesti riippumattomat. $\square$

Summan odotusarvo ja keskihajonta: Yhteenveto

Satunnaislukujen $X_1, \ldots, X_n$ summan odotusarvo ja keskihajonta, kun $\mu_i = \mathbb{E}(X_i)$, $\sigma_i = \mathrm{SD}(X_i)$ ja $\rho_{i,j} = \mathrm{Cor}(X_i, X_j)$:

Summan termit	$\mathbb{E}(\sum_i X_i)$	$\mathrm{SD}(\sum_i X_i)$
Yleiset	$\sum_i \mu_i$	$\sqrt{\sum_i \sigma_i^2 + \sum_i \sum_{j \neq i} \sigma_i \sigma_j \rho_{i,j}}$
Riippumattomat	$\sum_i \mu_i$	$\sqrt{\sum_i \sigma_i^2}$
Riippumattomat ja samoin jakautuneet	μn	$\sigma \sqrt{n}$

Välihuomio: SLL:n todistus

Nyt pystymme todistamaan suurten lukujen lain. Tarkastellaan pitkän aikavälin keskiarvoa

$$A_n = (X_1 + X_2 + \ldots + X_n) \;/\; n,$$

missä X_i:t ovat riippumattomia, jokaisella odotusarvo μ ja keskihajonta σ.

Yhdistetään kaksi tietämäämme asiaa:

1. $\mathrm{SD}(A_n)$ ei ole kovin suuri, se on $\sigma/\sqrt{n}$.
2. Tšebyšovin mukaan ei ole todennäköistä, että A_n on kovin monen $\mathrm{SD}(A_n)$:n päässä μ:stä. Siis SLL!

Tarkemmin: Olkoon $r_n = 0.001/\,\mathrm{SD}(A_n)$. Silloin

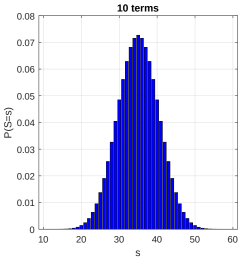
$$\begin{aligned} \mathbb{P}(|A_n - \mu| > 0.001) &\le \frac{1}{r_n^2} && \text{(Tšebyšov)} \\ &= \frac{\sigma^2}{0.001^2} \cdot \frac{1}{n} \\ &\to 0 \qquad \text{kun } n \to \infty. \end{aligned}$$

Esim. Noppapeli

Pelataan n kierrosta noppapeliä. Laske kertyneen tuoton $S_n = X_1 + \cdots + X_n$ odotusarvo ja keskihajonta, $n = 10, 100, 1000$.
Yhden kierroksen tuoton odotusarvo on $\mu = 3.5$ ja keskihajonta

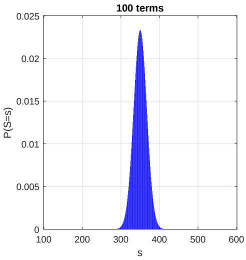
$$\sigma \;=\; \sqrt{\mathbb{E}(X_i^2) - \mu^2} \;=\; \sqrt{\tfrac{1}{6}(1^2 + \cdots + 6^2) - (3.5)^2} \;\approx\; 1.7.$$

Riippumattomat kierrokset $\implies$ $\mathbb{E}(S_n) = \mu n$ ja $\mathrm{SD}(S_n) = \sigma\sqrt{n}$.



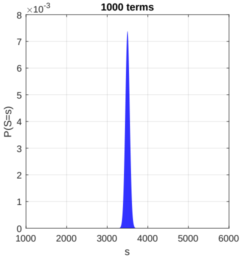
$\mathbb{E}(S_{10}) = 35$

$\mathrm{SD}(S_{10}) \approx 5.4$



$\mathbb{E}(S_{100}) = 350$

$\mathrm{SD}(S_{100}) \approx 17$



$\mathbb{E}(S_{1000}) = 3500$

$\mathrm{SD}(S_{1000}) \approx 54$

Esimerkki: Lentoyhtiö, jakauman muotoa tuntematta

300 lentolippua myydään lennolle, jossa on 290 matkustajapaikkaa. Arviolta 5% lipun ostaneista jää saapumatta lennolle, toisistaan riippumattomasti. Millä tn kaikki saapujat mahtuvat lennolle?

Lennolle saapuvien lukumäärä on $N = X_1 + \cdots + X_{300}$, missä

$$X_i = \begin{cases} 1, & \text{jos lentolipun } i \text{ ostaja saapuu lennolle,} \\ 0, & \text{muuten.} \end{cases}$$

Koska $\mu_X = \mathbb{E}(X_i) = 0.95$ ja $\sigma_X = \mathrm{SD}(X_i) = \sqrt{\mu_X(1-\mu_X)} \approx 0.22$, saadaan $\mu_N = \mu_X \times 300 = 285$ ja $\sigma_N = \sigma_X \times \sqrt{300} \approx 3.8$.

Tšebyšovin epäyhtälö

$$\mathbb{P}(N \in [280, 290]) \approx \mathbb{P}(N = \mu_N \pm 1.32\sigma_N) \geq 1 - \frac{1}{1.32^2} \approx 42.6\%.$$

takaa, että kaikki mahtuvat lennolle vähintään tn:llä 42.6%.
(Tämä kuulostaa pessimistiseltä arviolta?)

Lentoyhtiö: Käyttäen tarkkaa jakaumaa (binomijakauma)

Millainen on lennolle saapuvien lukumäärän N tarkka jakauma?

$$N = X_1 + \cdots + X_{300}$$

Satunnaismuuttujan N arvojoukko on $\{0, 1, 2, \ldots, 300\}$.

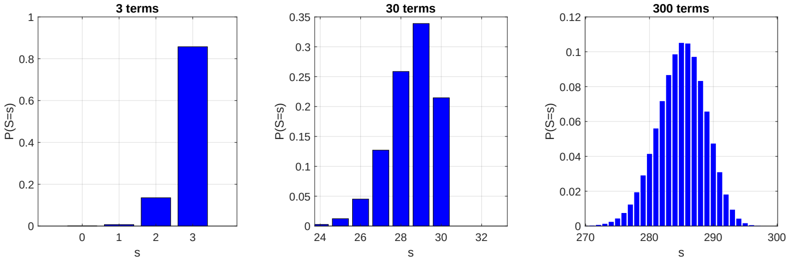
$$\mathbb{P}(N = 0) \; = \; (1 - 0.95)^{300} \; \leq \; 0.1^{300} \; = \; 10^{-300}$$

$$\mathbb{P}(N = k) \; = \; \binom{300}{k}(1 - 0.95)^{300-k} 0.95^k$$

- N noudattaa binomijakaumaa parametrein $n = 300$ ja $p = 0.95$.
- Pienet N:n arvot ovat (yli)tähtitieteellisen epätodennäköisiä
- R:llä saadaan tarkka arvo
  $\mathbb{P}(N \leq 290) \; =$ `pbinom(290,300,0.95)` $\approx$ 93.5% ja
  $\mathbb{P}(N \in [280, 290]) \; =$
  `pbinom(290,300,0.95) - pbinom(279,300,0.95)` $\approx$ 85.7%.

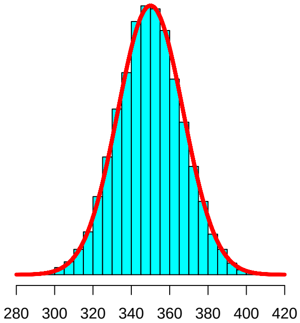
Lentoyhtiö: Tulijoiden määrän jakauma

Jos lippuja myydään n kappaletta, tulijoiden määrä on binomijakautunut parametrein n ja $p = 0.95$. (Jakauman voisi laskea myös indikaattorimuuttujien jakaumasta konvoluutiokaavalla.)

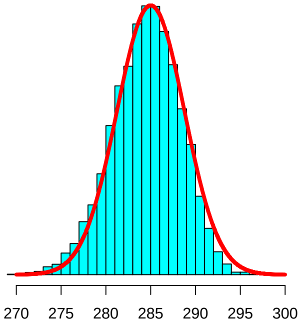


100 noppaa vs. 300 lentolippua

Nopanheittojen summan ja saapuvien lentomatkustajien lukumäärän jakaumat ovat likimain samanmuotoiset (joskin eri paikassa ja eri levyiset).



100 riippumattoman nopanheiton summa



300 riippumattoman indikaattorimuuttujan summa

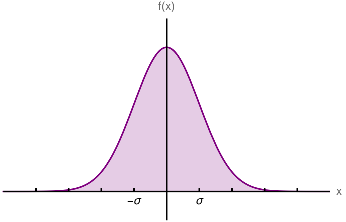
Tämä ei ole sattumaa!

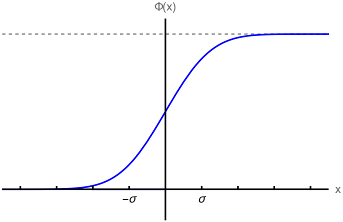
Sisältö

Standardinormaalijakauma

Satunnaisluku Z noudattaa standardinormaalijakaumaa (jolla on odotusarvo $\mu = 0$ ja keskihajonta $\sigma = 1$), jos sillä on tiheysfunktio

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} = \texttt{dnorm(z)} = \texttt{normpdf(z)}$$





Standardinormaalijakauman kertymäfunktio on

$$\Phi(z) = F_Z(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt = \texttt{pnorm(z)} = \texttt{normcdf(z)}$$

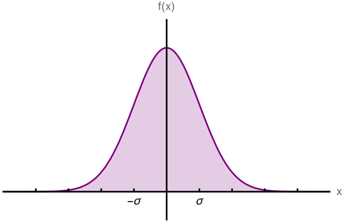
Integraalifunktiota on hankala laskea, mutta se löytyy taulukoista (ks. kurssisivu), laskimista ja tietokoneohjelmista (R, Matlab/Octave).

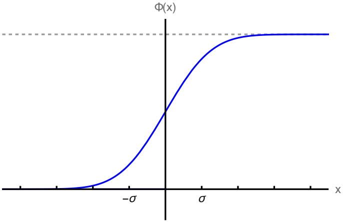
Toinen nimi tälle jakaumalle on normitettu normaalijakauma.

Normaalijakauma (yleinen tapaus)

Satunnaisluku X noudattaa normaalijakaumaa odotusarvolla μ ja keskihajonnalla σ, jos sillä on tiheysfunktio

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \texttt{dnorm(x, mu, sigma)} = \texttt{normpdf(x, mu, sigma)}$$





Yleisen normaalijakauman kertymäfunktio R/Matlab-komennoilla

`pnorm(x, mu, sigma)`, `normcdf(x, mu, sigma)`

Toinen vaihtoehto on palauttaa tapaus standardinormaalijakaumaan skaalaamalla ja siirtämällä, jolloin arvot voi katsoa taulukoista.

Normaalijakaumaa voi skaalata ja siirtää

Fakta

Jos X on normaalijakautunut odotusarvona μ_X ja keskihajontana σ_X, niin tällöin myös $Y = a + bX$ on normaalijakautunut odotusarvona

$$\mu_Y = \mathbb{E}(a + bX) = a + b\mathbb{E}(X) = a + b\mu_X$$

ja keskihajontana

$$\sigma_Y = \mathrm{SD}(a + bX) = |b|\,\mathrm{SD}(X) = |b|\sigma_X$$

Seuraus

Normitettu satunnaisluku $Z = \frac{X - \mu_X}{\sigma_X}$ noudattaa standardinormaalijakaumaa ($\mu_Z = 0$ ja $\sigma_Z = 1$). (Joten sen kertymäfunktiot löytyvät taulukoista.)

Kertymäfunktion laskeminen normittamalla

Jos X on normaali parametrein μ ja σ, niin lineaarimuunnoksella

$$Z = \frac{X - \mu}{\sigma}$$

on standardinormaalijakauma.

Silloin

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) \\ &= \mathbb{P}(\mu + \sigma Z \leq x) \\ &= \mathbb{P}(\sigma Z \leq x - \mu) \\ &= \mathbb{P}\left(Z \leq \frac{x - \mu}{\sigma}\right) \\ &= F_Z\left(\frac{x - \mu}{\sigma}\right). \end{aligned}$$

Lukuarvot $F_z(\ldots)$ eli $\Phi(\ldots)$ voi katsoa taulukoista tai laskea esim. R:llä.

Kertymäfunktio suoraan tai normittamalla

Olkoon X normaalijakautunut odotusarvolla $\mu = 10$ ja keskihajonnalla $\sigma = 3$.

Laske $F_X(16) = \mathbb{P}(X \leq 16)$, eli todennäköisyys että X on enintään kaksi keskihajontaa (2σ) odotusarvoa suurempi?

Tapa 1. Suoraan (R:llä).

```
> pnorm(16,10,3)
[1] 0.9772499
```

Tapa 2. Normitetaan ensin. Koska muuttujalla $Z = (X - 10)/3$ on standardinormaalijakauma, laskemme $F_Z((16 - 10)/3) = F_Z(2)$ näin . . .

```
> pnorm((16-10)/3)
[1] 0.9772499
```

Normaalijakauma: Hyödyllisiä ominaisuuksia

Fakta

Jos X, Y ovat normaalijakautuneita ja riippumattomia, niin $S = X + Y$ on myös normaalijakautunut.

Ylläoleva ei ole mitenkään itsestään selvää! Kaikilla jakaumilla ei suinkaan ole tällaista yhteenlaskuominaisuutta (vrt. esim. tasajakauma).

Fakta

Jos X on normaalijakautunut, niin mikä tahansa lineaarimuunnos $Y = a + bX$ on normaalijakautunut.

Molemmissa tapauksissa siis tiedämme muodon. Jakauman odotusarvon ja keskihajonnan voimme laskea kaavoilla, jotka jo tunnemme (lineaarisuus).

Normaalijakauma: Hyödyllisiä ominaisuuksia II

Fakta

Komplementtisääntö: $\mathbb{P}(X > a) = 1 - \mathbb{P}(X \leq a)$ *(pätee kaikille jakaumille, ei vain normaaljakaumalle).*

Fakta

Normaalijakauma on symmetrinen odotusarvonsa suhteen, eli oikean ja vasemman hännän todennäköisyydet ovat samat (kun mennään saman verran eli a yksikköä μ*:sta vasemmalle ja oikealle).*

$$\mathbb{P}(X > \mu + a) = \mathbb{P}(X < \mu - a).$$

Fakta

Keskiosan todennäköisyys saadaan vähentämällä ykkösestä molemmat hännät.

$$\mathbb{P}(\mu - a < X < \mu + a) = 1 - 2\mathbb{P}(X < \mu - a).$$

Esimerkki: Älykkyysosamäärä

Yhdeksäsluokkalaisten älykkyysosamäärä noudattaa likimain normaalijakaumaa ($\mu = 100$, $\sigma = 15$). Millä tn satunnaisesti valitun yhdeksäsluokkalaisen älykkyysosamäärä on

(a) yli 130?

(b) välillä 85–115?

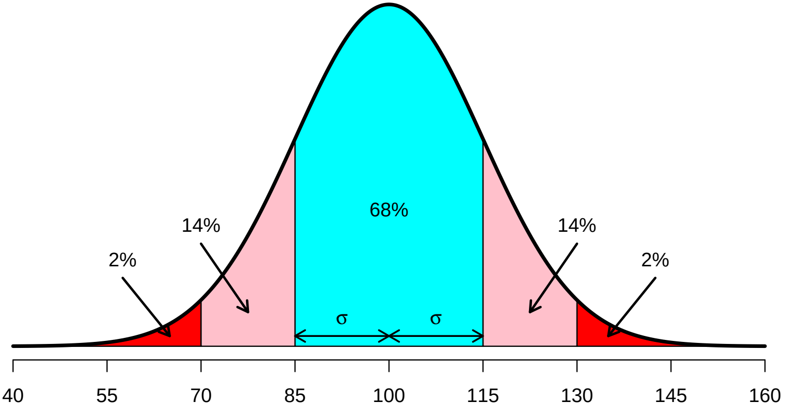
$$\begin{aligned}\mathbb{P}(X > 130) &= \mathbb{P}\left(\frac{X-\mu}{\sigma} > \frac{130-100}{15}\right) \\ &= \mathbb{P}(Z > 2) = \mathbb{P}(Z \le -2) \approx 2.3\%.\end{aligned}$$

$$\begin{aligned}\mathbb{P}(85 \le X \le 115) &= \mathbb{P}\left(\frac{85-100}{15} \le \frac{X-\mu}{\sigma} \le \frac{115-100}{15}\right) \\ &= \mathbb{P}(-1 \le Z \le 1) \\ &= 1 - \mathbb{P}(Z > 1) - \mathbb{P}(Z < -1) \\ &= 1 - 2\mathbb{P}(Z \le -1) \approx 68\%.\end{aligned}$$

R: `pnorm(-2); 1-2*pnorm(-1)`

Esimerkki: Älykkyysosamäärä

Odotusarvo $\mu = 100$, keskihajonta $\sigma = 15$



Sisältö

Normaaliapproksimaatio

Fakta (Keskeinen raja-arvolause)

Jos $X_1, \dots, X_n$ ovat riippumattomia ja samoin jakautuneita satunnaislukuja odotusarvona μ ja keskihajontana σ, niin

$$\frac{\sum_{i=1}^{n} X_i - \mu n}{\sigma \sqrt{n}} \stackrel{\mathrm{d}}{\approx} Z$$

noudattaa suurilla n likimain standardinormaalijakaumaa tiheysfunktiona

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

Huom

Tämä on universaali luonnonlaki, sillä X_i:n jakauman luonteesta (diskreetti/jatkuva, symmetrinen/vino) ei tarvitse olettaa mitään. (Summattavien riippumattomuus sen sijaan on kohtalaisen olennaista.)

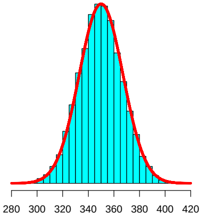
de Moivre 1733, Laplace 1812, Lyapunov 1911, Lindeberg 1922, Turing 1934

Esimerkki: Noppapeli

Millä tn 100 pelikierrokselta kertynyt tuotto on

(a) välillä 316–384?

(b) yli 500 EUR?



Yhden kierroksen tuoton odotusarvo $\mu_X = 3.5$ ja keskihajonta $\sigma_X \approx 1.7$.
Normaaliapproksimaatio:

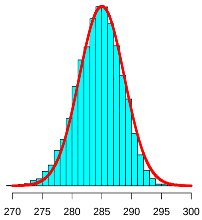
$$\frac{S_{100} - 350}{17} = \frac{S_{100} - 100\mu_X}{\sqrt{100}\sigma_X} \overset{\mathrm{d}}{\approx} Z.$$

$$\begin{aligned}\mathbb{P}(316 \le S_{100} \le 384) &= \mathbb{P}\left(-2 \le \frac{S_{100} - 350}{17} \le 2\right) \\ &\approx \mathbb{P}(-2 \le Z \le 2) = 1 - 2\mathbb{P}(Z \le -2) \approx 95.4\%.\end{aligned}$$

$$\begin{aligned}\mathbb{P}(S_{100} > 500) &= \mathbb{P}\left(\frac{S_{100} - 350}{17} > 8.82\right) \\ &\approx \mathbb{P}(Z > 8.82) = \mathbb{P}(Z \le -8.82) \approx 6 \times 10^{-19}.\end{aligned}$$

Esimerkki: Lentoyhtiö

Millä tn kaikki lipun ostaneet mahtuvat lennolle? (Myyty 300 lippua, lennolla 290 paikkaa.)



Lennolle saapuvien lukumäärä $N = X_1 + \cdots + X_{300}$. Indikaattorin X_i odotusarvo $\mu_X = 0.95$ ja keskihajonta $\sigma_X = 0.22$.
Normaaliapproksimaatio:

$$\frac{N - 285}{3.77} = \frac{N - 300\mu_X}{\sqrt{300}\sigma_X} \stackrel{\mathrm{d}}{\approx} Z.$$

$$\begin{aligned}
\mathbb{P}(N \leq 290) = \mathbb{P}(N \leq 290.5) &= \mathbb{P}\left(\frac{N - 285}{3.77} \leq 1.46\right) \\
&\approx \mathbb{P}(Z \leq 1.46) \\
&= 1 - \mathbb{P}(Z \leq -1.46) \approx 92.8\%.
\end{aligned}$$

(Tarkka tn: `pbinom(290,300,0.95)` = 93.5%)

Sisältö

Eksponenttijakautuneiden summa: muoto

Hyönteisiä osuu tuulilasiin satunnaisesti taajuudella $\lambda = 1/100$ (yksi kpl / 100 sekuntia).

Olkoon $X_i \sim \text{Exp}(\lambda)$ väliaika edellisestä hyönteisestä i:nteen hyönteiseen. Väliajoilla on eksponenttijakauma.

Aika, jossa kertyy n hyönteistä on $S = X_1 + X_2 + \ldots + X_n$. Se *ei ole* eksponenttijakautunut. Kokeile esim. seuraavaa R-koodia lukumäärillä $n = 2$, $n = 5$ or $n = 50$.

```
rate    <- 1/100
repeats <- 1000000
n       <- 5
X       <- matrix(rexp(repeats*n, rate), repeats, n)
S       <- rowSums(X)
hist(S,100)
```

Kyseessä on ns. gammajakauma.

Seuraavalla kerralla puhutaan empiirisistä jakaumista . . .