MS-A0504 Todennäköisyyslaskennan ja tilastotieteen peruskurssi

3A Satunnaismuuttujien summa ja keskihajonta

Lasse Leskelä

Matematiikan ja systeemianalyysin laitos
Perustieteiden korkeakoulu
Aalto-yliopisto

Lukuvuosi 2018–2019
Periodi IV

Sisältö

Kahden satunnaismuuttujan summa

Kahden satunnaismuuttujan summa $X + Y$ on satunnaismuuttuja, jonka 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 ovat stokastisesti riippumattomat:

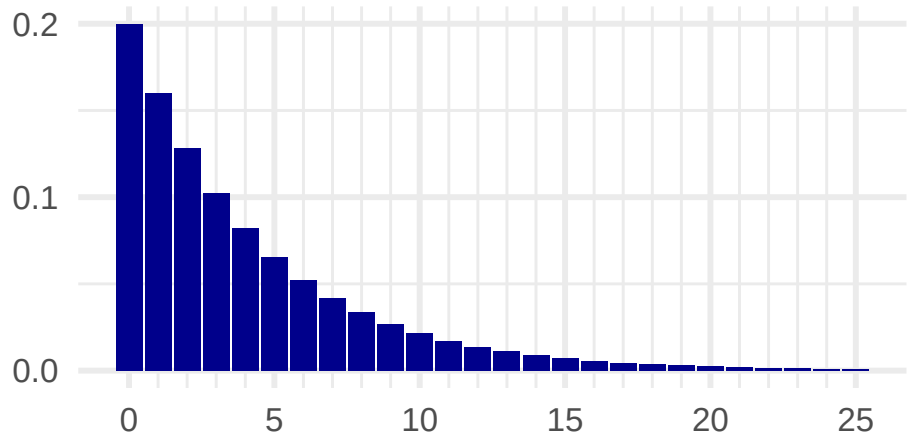
$$f_{X+Y}(s) = \sum_x f_X(x) f_Y(s - x)$$

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

Kahden satunnaismuuttujan summa

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

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



Määritä satunnaismuuttujan $X_1 + X_2$ jakauma.

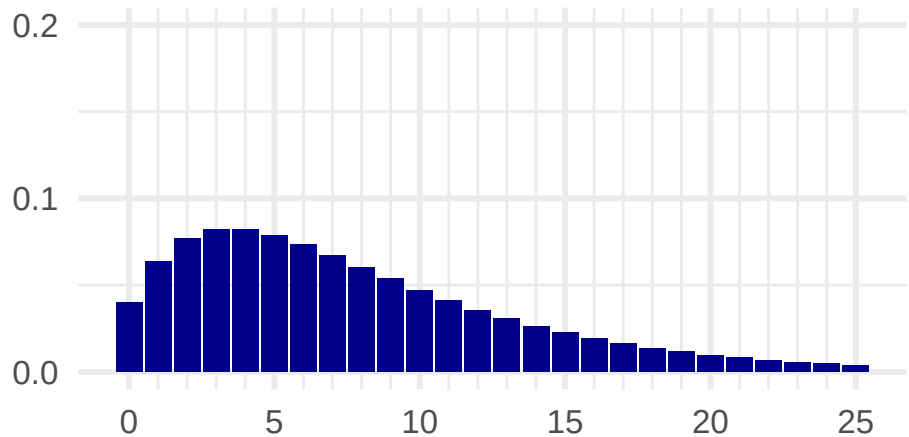
Kahden satunnaismuuttujan summa

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

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



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: Yleinen tapaus

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. □

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 $\rho_{i,j} = \mathrm{Cor}(X_i, X_j)$:

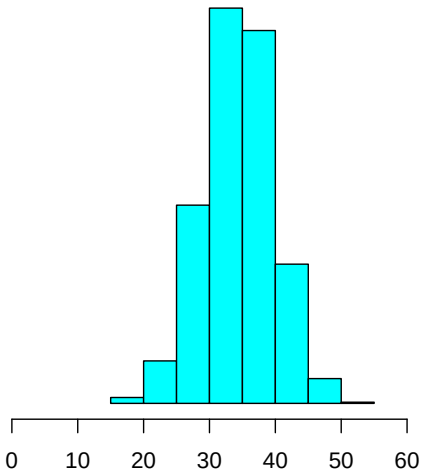
Summan termit	$\mathbb{E}(\sum_i X_i)$	$\mathrm{SD}(\sum_i X_i)$
Yleiset	$\sum_i \mu_i$	$\sqrt{\sum_i \sigma_i^2 + \sum_i \sum_{j \neq i} \sigma_i \sigma_j \rho_{i,j}}$
Riippumattomat	$\sum_i \mu_i$	$\sqrt{\sum_i \sigma_i^2}$
Riippumattomat ja samoin jakautuneet	μn	$\sigma \sqrt{n}$

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

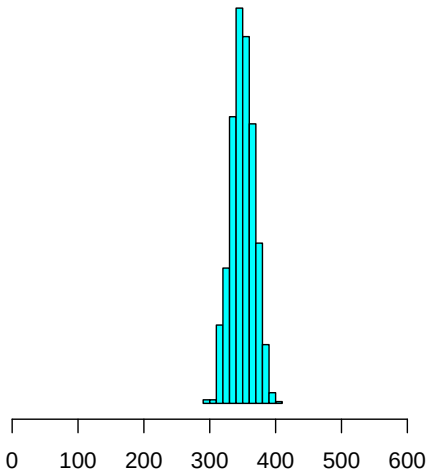
$$\sigma = \sqrt{\mathbb{E}(X_i^2) - \mu^2} = \sqrt{\tfrac{1}{6}(1^2 + \cdots + 6^2) - (3.5)^2} \approx 1.7.$$

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



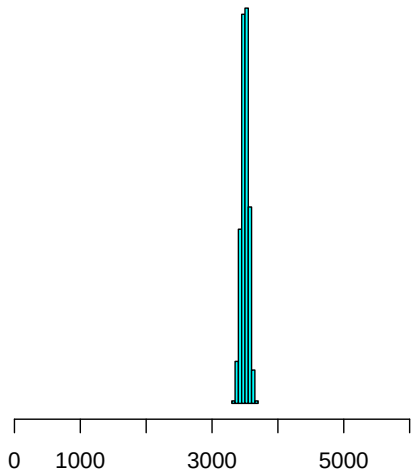
$\mathbb{E}(S_{10}) = 35$

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



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

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

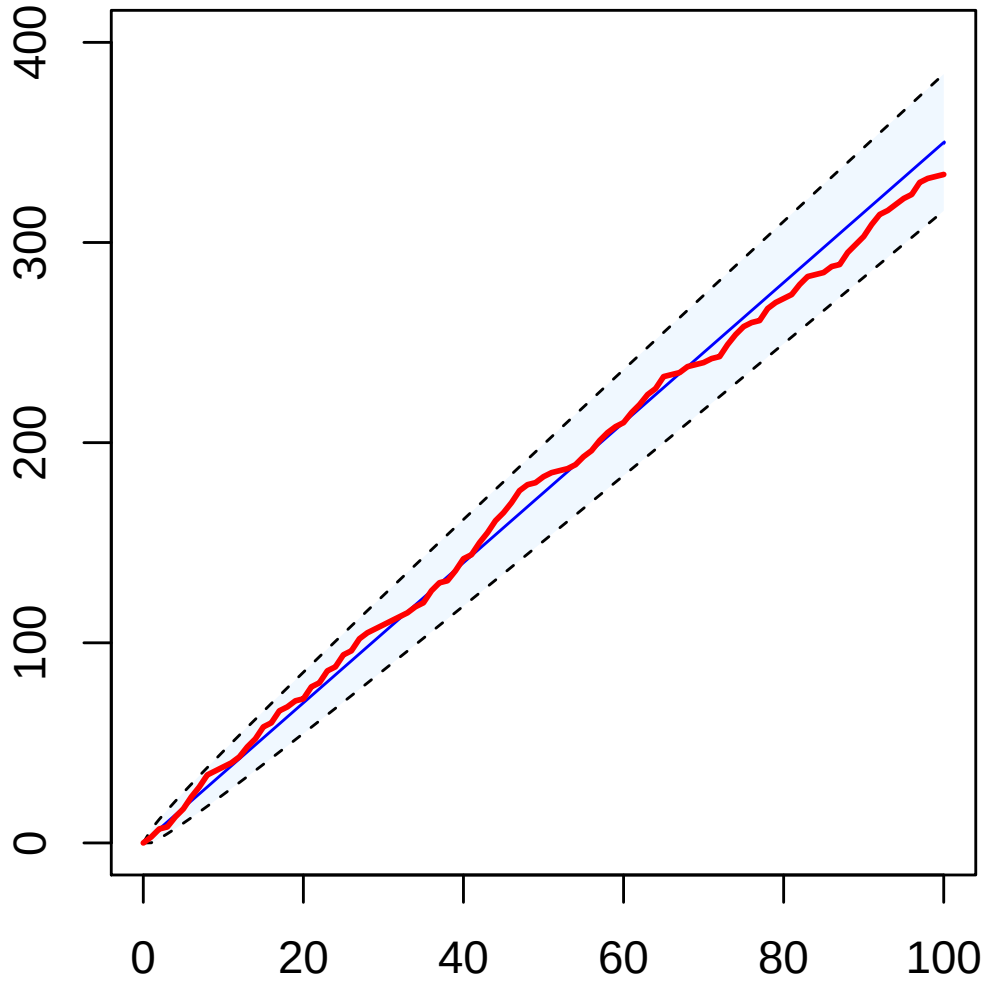


$\mathbb{E}(S_{1000}) = 3500$

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

Esim. Noppapeli

100 kierroksen tuotto on odotusarvoltaan $\mu \times 100 = 350$ ja keskihajonnaltaan $\sigma \times \sqrt{100} \approx 17$



Chebyshevin epäyhtälö: $\mathbb{P}(S_{100} = 350 \pm 34) \geq 75\%$

Noppapelin 100 pelikierroksen tuotto on siis melko todennäköisesti (tn $\geq 75\%$) välillä 316–384 EUR.

Esimerkki: Lentoyhtiö

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 $\sigma_N = \sigma_X \times \sqrt{300} \approx 3.8$.

Chebyshevin 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: Tarkka jakauma

Millainen on lennolle saapuvien lukumäärän N tarkka jakauma?

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

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

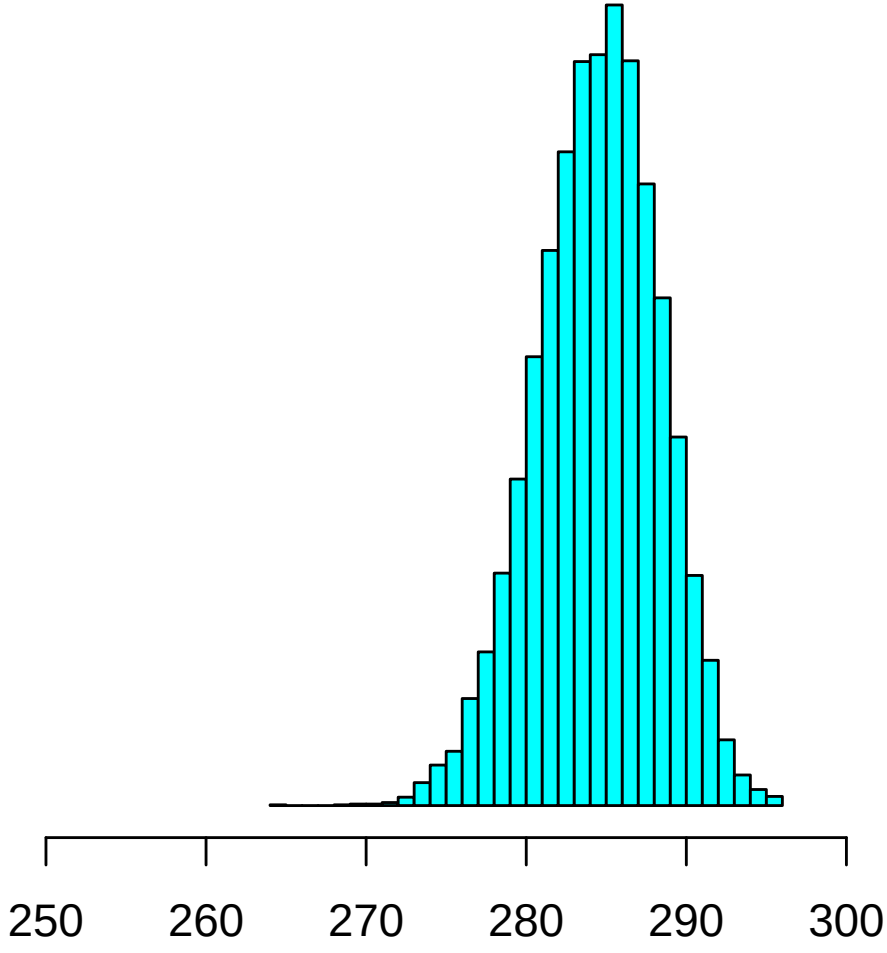
$$\mathbb{P}(N = 0) \;=\; (1 - 0.95)^{300} \;\leq\; 0.1^{300} \;=\; 10^{-300}$$

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

- 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%.

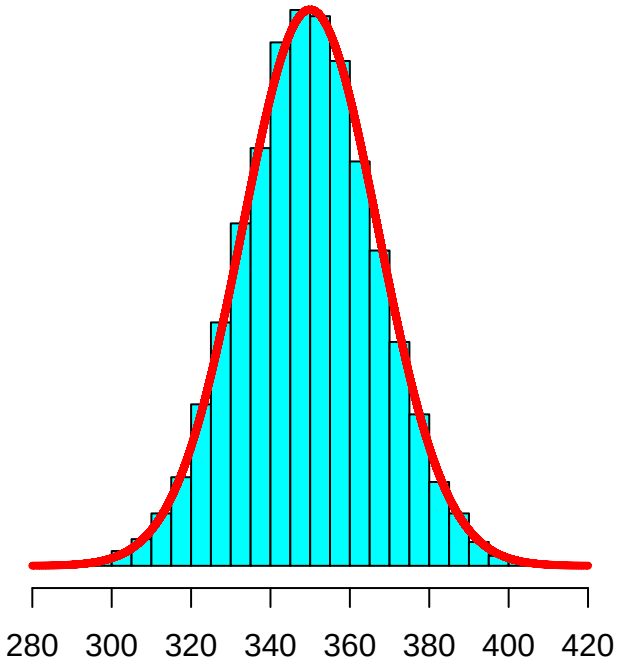
Lentoyhtiö: Simuloitu jakauma

Simuloidaan N:n tavoin $\mathrm{Bin}(300, 0.95)$-jakautuneita satunnaislukuja 10000 kappaletta ja piirretään histogrammi.

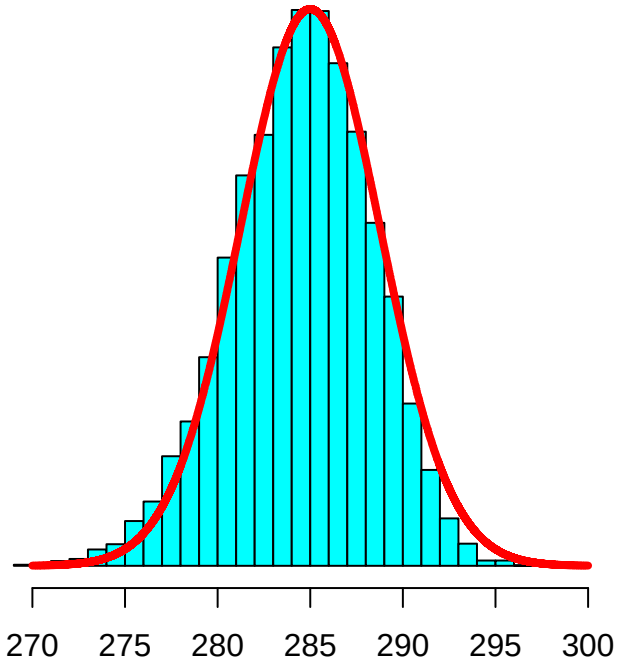


100 noppaa vs. 300 lentolippua

Nopanheittojen summan ja saapuvien lentomatkustajien lukumäärän jakaumat ovat likimain samanmuotoiset:



100 riippumattoman
nopanheiton summa



300 riippumattoman
indikaattorimuuttujan summa

Tämä ei ole sattumaa!

Sisältö

Normaaliapproksimaatio

Fakta (Keskeinen raja-arvolause)

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

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

noudattaa suurilla n likimain normitettua normaalijakaumaa tiheysfunktiona

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

Huom

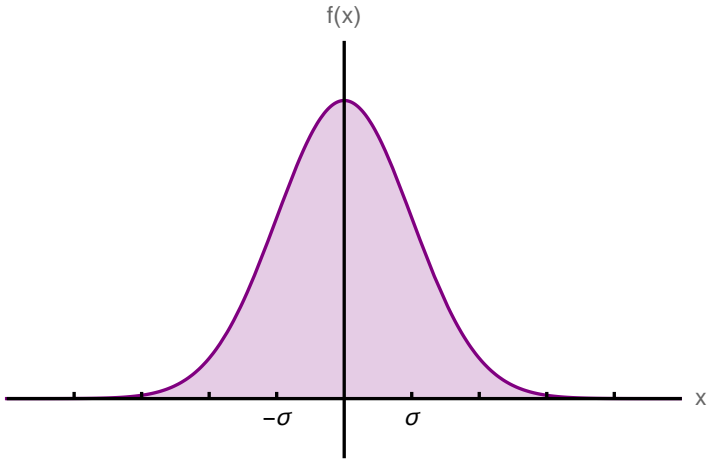
Tämä on universaali luonnonlaki, sillä X_i:n jakauman luonteesta (diskreetti/jatkuva, symmetrinen/vino) ei tarvitse olettaa mitään.

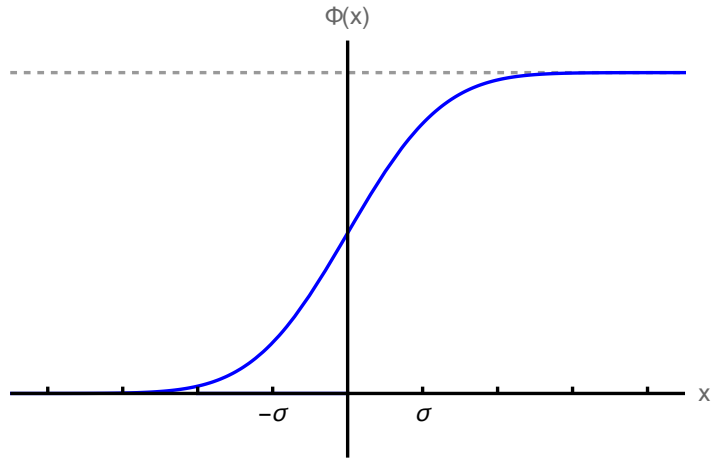
de Moivre 1733, Laplace 1812, Lyapunov 1911, Lindeberg 1922, Turing 1934

Normaalijakauma

Satunnaisluku Z noudattaa normaalijakaumaa odotusarvona μ ja keskihajontana σ, jos sillä on tiheysfunktio

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





Normitetun normaalijakauman ($\mu = 0$ ja $\sigma = 1$) kertymäfunktio on

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

Normaalijakauman normittaminen

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 normitettua normaalijakaumaa ($\mu_Z = 0$ ja $\sigma_Z = 1$).

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?

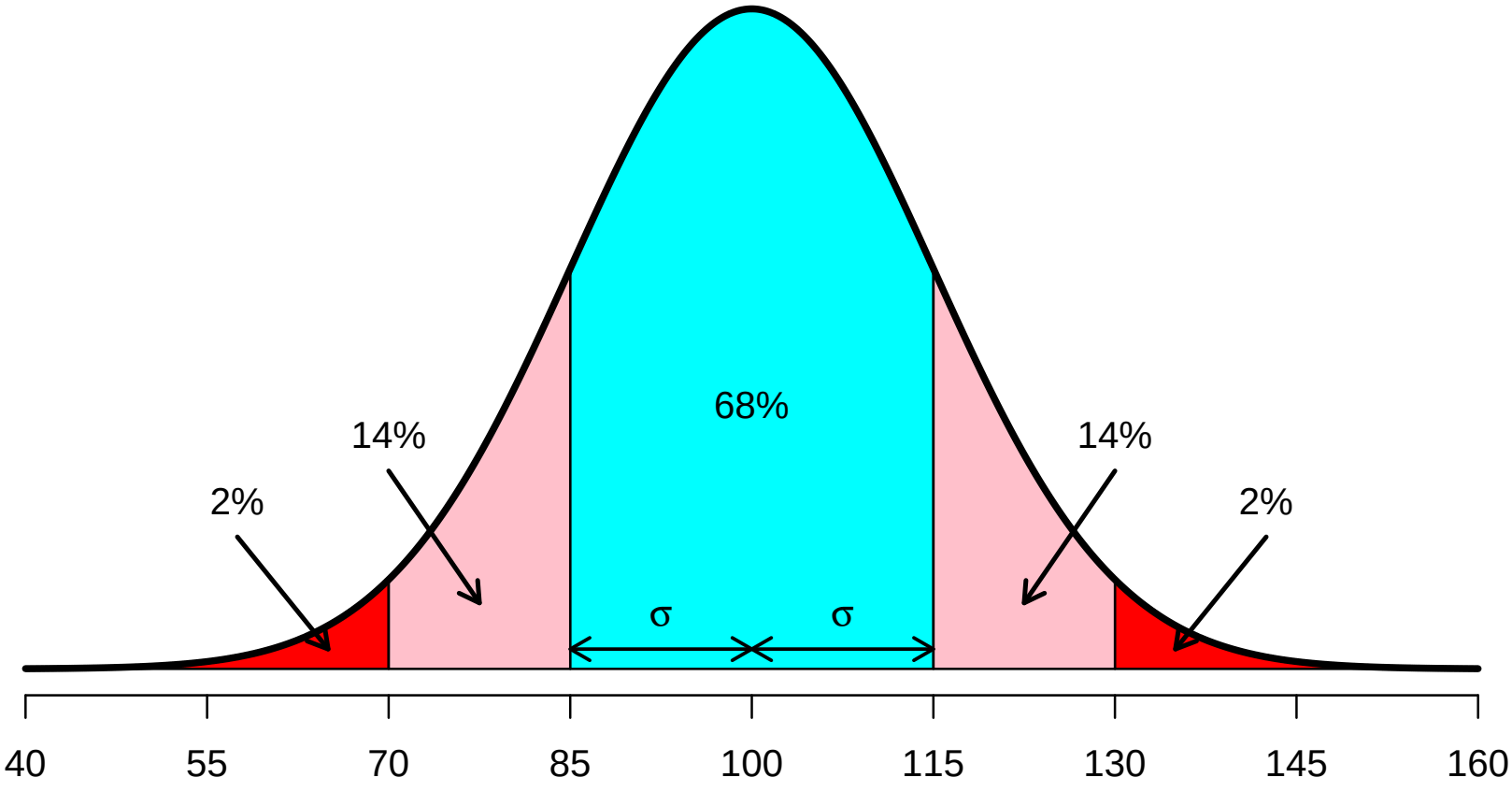
$$\begin{aligned}\mathbb{P}(X > 130) &= \mathbb{P}\left(\frac{X-\mu}{\sigma} > \frac{130-100}{15}\right) \\ &= \mathbb{P}(Z > 2) = \mathbb{P}(Z \leq -2) \approx 2.3\%.\end{aligned}$$

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

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

Esimerkki: Älykkyysosamäärä

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

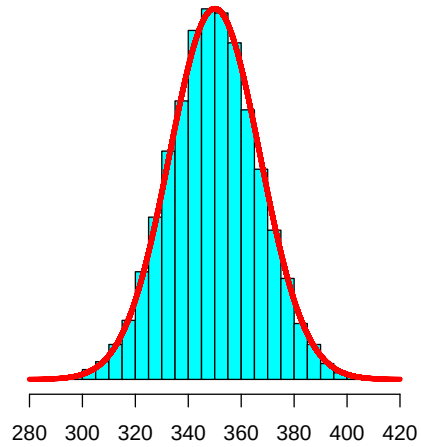


Esimerkki: Noppapeli

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:

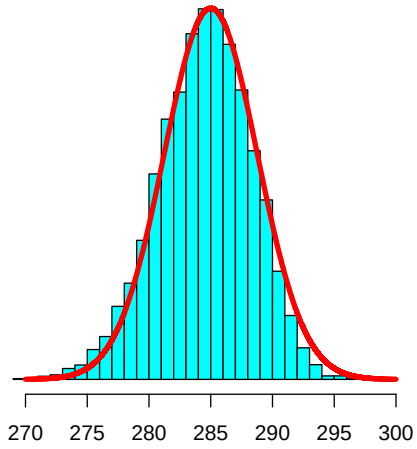
$$\frac{S_{100} - 350}{17} = \frac{S_{100} - 100\mu_X}{\sqrt{100}\sigma_X} \overset{\mathrm{d}}{\approx} Z.$$

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

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

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:

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

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

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

Seuraavalla kerralla puhutaan empiiristä jakaumista . . .