Aalto-yliopisto

Keskeinen raja-arvolause ja sen sovellukset liiketoiminnassa

https://github.com/perej1/clt-teaching

Jaakko Pere

25. maaliskuuta 2025

Kohdeyleisö

- Ensimmäisen ja toisen vuoden kandiopiskelijat (lukion matematiikka).

Kohdeyleisö

- Ensimmäisen ja toisen vuoden kandiopiskelijat (lukion matematiikka).
- Kurssilla on jo käsitelty seuraavat käsitteet:
  - Satunnaismuuttuja
  - Odotusarvo
  - Varianssi
  - Estimaattori
  - Suurten lukujen laki

Luennon sisältö

Moraali

Normaalijakauma

Keskeinen raja-arvolause

Sovellukset

Table of Contents

Suurten lukujen laki

Olkoon $X_1, \ldots, X_n$ riippumattomia ja samoin jakautuneita satunnaismuuttujia niin, että $\mu = \mathbb{E}(X_1) \in (-\infty, \infty)$.

Suurten lukujen laki

Olkoon $X_1, \ldots, X_n$ riippumattomia ja samoin jakautuneita satunnaismuuttujia niin, että $\mu = \mathbb{E}\,(X_1) \in (-\infty, \infty)$. Tällöin suurten lukujen laki kertoo, että suurilla n

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathbb{E}\,(X_1)\,.$$

Suurten lukujen laki

Olkoon $X_1, \ldots, X_n$ riippumattomia ja samoin jakautuneita satunnaismuuttujia niin, että $\mu = \mathbb{E}\left(X_1\right) \in (-\infty, \infty)$. Tällöin suurten lukujen laki kertoo, että suurilla n

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathbb{E}\left(X_1\right).$$

Toisin sanoen keskiarvo estimoi odotusarvoa.

Suurten lukujen laki

Olkoon $X_1, \ldots, X_n$ riippumattomia ja samoin jakautuneita satunnaismuuttujia niin, että $\mu = \mathbb{E}(X_1) \in (-\infty, \infty)$. Tällöin suurten lukujen laki kertoo, että suurilla n

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathbb{E}(X_1).$$

Toisin sanoen keskiarvo estimoi odotusarvoa.

Tarkemmin ilmaistuna $\lim_{n\to\infty} \mathbb{P}\left(\left|\bar{X} - \mu\right| > \varepsilon\right) = 0$ kaikilla $\varepsilon > 0$.

Ongelma

- Suurten lukujen laki ei kerro mitään siitä kuinka hyvin keskiarvo estimoi odotusarvoa.

Ongelma

- Suurten lukujen laki ei kerro mitään siitä kuinka hyvin keskiarvo estimoi odotusarvoa.
- Voidaanko keskiarvon jakaumasta sanoa jotain edes suurilla otosko'oilla n?

Esimerkki: luottokorttipetokset

- Otos sisältää luottokorttitapahtumia Euroopasta syyskuulta 2013. Tapahtumia kirjattiin kahdelta päivältä (Lähde: Kaggle).

Esimerkki: luottokorttipetokset

- Otos sisältää luottokorttitapahtumia Euroopasta syyskuulta 2013. Tapahtumia kirjattiin kahdelta päivältä (Lähde: Kaggle).
- Luottokorttitapahtumia on yhteensä 284 807, joista 492 luokiteltiin petoksiksi.

Esimerkki: luottokorttipetokset

- Otos sisältää luottokorttitapahtumia Euroopasta syyskuulta 2013. Tapahtumia kirjattiin kahdelta päivältä (Lähde: Kaggle).
- Luottokorttitapahtumia on yhteensä 284 807, joista 492 luokiteltiin petoksiksi.
- Estimaatti luottokorttipetoksen todennäköisyydelle on $492/284807 \approx 0.0017$.

Esimerkki: luottokorttipetokset

- Otos sisältää luottokorttitapahtumia Euroopasta syyskuulta 2013. Tapahtumia kirjattiin kahdelta päivältä (Lähde: Kaggle).
- Luottokorttitapahtumia on yhteensä 284 807, joista 492 luokiteltiin petoksiksi.
- Estimaatti luottokorttipetoksen todennäköisyydelle on $492/284807 \approx 0.0017$.
- Kuinka varma voin olla siitä, että saatu piste-estimaatti on lähellä populaatisuuretta?

Simulaatio

1. Simuloi otos riippumattomia ja samoin jakauneita havaintoja $x_1, \ldots, x_n$ (otoskoko on n). Laske keskiarvo

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Simulaatio

1. Simuloi otos riippumattomia ja samoin jakauneita havaintoja $x_1, \ldots, x_n$ (otoskoko on n). Laske keskiarvo

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

2. Toista yllä oleva toimenpide $m = 1000$ kertaa. Näin meillä on m keskiarvoa.

Simulaatio

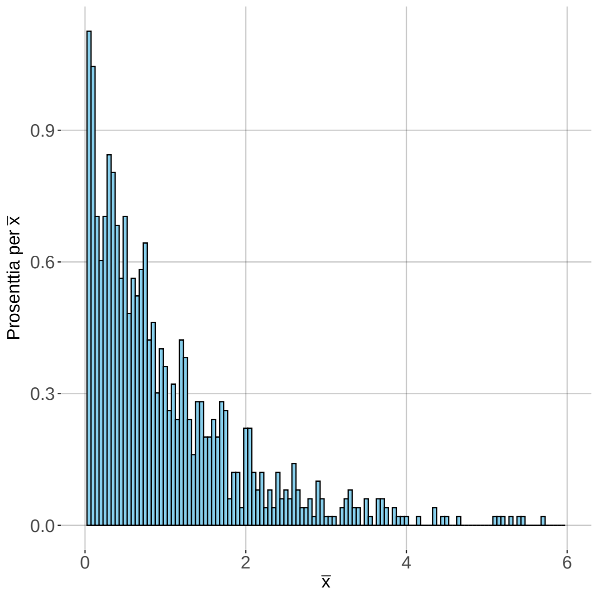
1. Simuloi otos riippumattomia ja samoin jakauneita havaintoja $x_1, \ldots, x_n$ (otoskoko on n). Laske keskiarvo

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

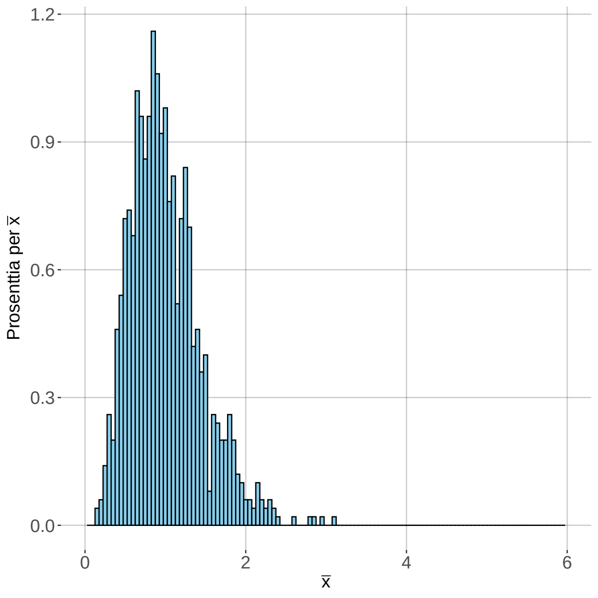
2. Toista yllä oleva toimenpide $m = 1000$ kertaa. Näin meillä on m keskiarvoa.
3. Piirrä histogrammi keskiarvoista.

- Seuraavaksi simuloimme otoksia eksponenttijakaumasta $\mathrm{Exp}(1)$ skaalaparametrilla $\lambda = 1$ (jakauman odotusarvo on 1). Kyseisen eksponentijakauman tiheysfunktio on

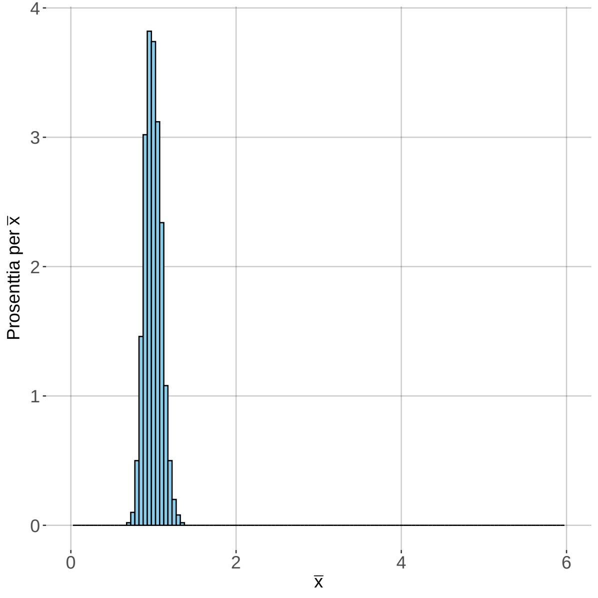
$$f(x) = \begin{cases} e^{-x}, & \text{kun} \quad x \in [0, \infty), \\ 0, & \text{muulloin.} \end{cases}$$



Kuva: Eksponenttijakauma $\mathrm{Exp}\,(1)$, $n = 1$ ja $m = 1000$.



Kuva: Eksponenttijakauma $\mathrm{Exp}\,(1)$, $n = 5$ ja $m = 1000$.



Kuva: Eksponenttijakauma $\mathrm{Exp}\,(1)$, $n = 100$ ja $m = 1000$.

Opetustavoitteet

Luennon jälkeen osaamme

1. approksimoida todennäköisyyksiä klassisen keskeisen raja-arvolauseen avulla ja

Opetustavoitteet

Luennon jälkeen osaamme

1. approksimoida todennäköisyyksiä klassisen keskeisen raja-arvolauseen avulla ja
2. muodostaa likiarvoisen luottamusvälin odotusarvolle perustuen keskeiseen raja-arvolauseeseen.

Table of Contents

Normaalijakauma on absoluuttisesti jatkuva jakauma. Merkitsemme normaalijakaumaa parametrein $\mu \in (-\infty, \infty)$ ja $\sigma^2 \in (0, \infty)$ notaatiolla $\mathrm{N}\left(\mu, \sigma^2\right)$. Jakauman $\mathrm{N}\left(\mu, \sigma^2\right)$ tiheysfunktio on

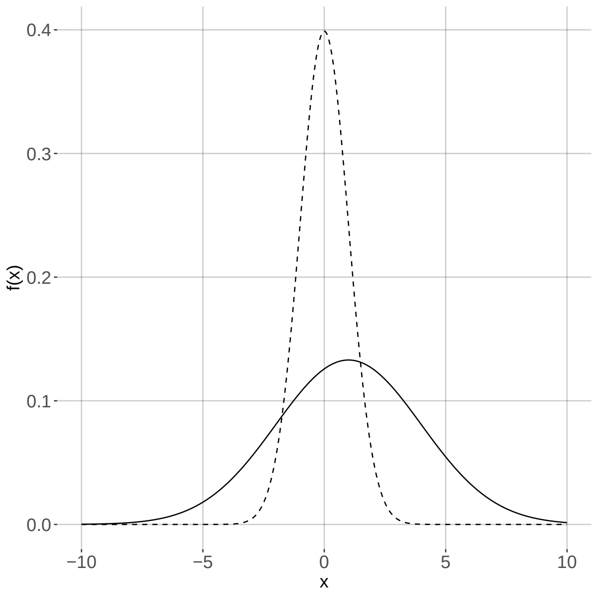
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Normaalijakauma on absoluuttisesti jatkuva jakauma. Merkitsemme normaalijakaumaa parametrein $\mu \in (-\infty, \infty)$ ja $\sigma^2 \in (0, \infty)$ notaatiolla $\mathrm{N}\left(\mu, \sigma^2\right)$. Jakauman $\mathrm{N}\left(\mu, \sigma^2\right)$ tiheysfunktio on

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Tällöin normaalijakauman kertymäfunktio voidaan esittää tiheysfunktion avulla

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t.$$



Kuva: Normaalijakaumat $\mathrm{N}(0, 1)$ (katkoviiva) ja $\mathrm{N}(1, 9)$ (jatkuva viiva).

Table of Contents

Lause

Olkoon $X_1, X_2, \ldots, X_n$ *riippumattomia ja samoin jakautuneita satunnaismuuttujia niin, että* $0 < \mathrm{Var}\,(X_1) < \infty$. *Merkitsemme* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\mu = \mathbb{E}\,(X_1)$ *ja* $\sigma = \sqrt{\mathrm{Var}\,(X_1)}$. *Tällöin, kun otoskoko* n *on suuri, niin*

$$\tilde{X} = \sqrt{n}\frac{\bar{X} - \mu}{\sigma}$$

noudattaa likimain standardinormaalijakaumaa $\mathrm{N}\,(0, 1)$,

Lause

Olkoon $X_1, X_2, \ldots, X_n$ riippumattomia ja samoin jakautuneita satunnaismuuttujia niin, että $0 < \mathrm{Var}\left(X_1\right) < \infty$. Merkitsemme $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\mu = \mathbb{E}\left(X_1\right)$ ja $\sigma = \sqrt{\mathrm{Var}\left(X_1\right)}$. Tällöin, kun otoskoko n on suuri, niin

$$\tilde{X} = \sqrt{n}\frac{\bar{X} - \mu}{\sigma}$$

noudattaa likimain standardinormaalijakaumaa $\mathrm{N}\left(0, 1\right)$,

$$\mathbb{P}\left(\tilde{X} \leq x\right) \approx \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d}t =: \Phi\left(x\right).$$

Table of Contents

Todennäköisyyksien approksimointi

Keskeisen raja-arvolauseen oletusten pätiessä $\tilde{X}$ noudattaa likimain normaalijakaumaa, jossa

$$\tilde{X} = \sqrt{n}\frac{\frac{1}{n}\overbrace{\sum_{i=1}^{n} X_i}^{=S_n} - \mu}{\sigma} = \frac{S_n - n\mu}{\sqrt{n}\sigma}.$$

Keskeisen raja-arvolauseen oletusten pätiessä $\tilde{X}$ noudattaa likimain normaalijakaumaa, jossa

$$\tilde{X} = \sqrt{n}\frac{\frac{1}{n}\overbrace{\sum_{i=1}^{n} X_i}^{=S_n} - \mu}{\sigma} = \frac{S_n - n\mu}{\sqrt{n}\sigma}.$$

Tällöin

$$\mathbb{P}\left(S_n \leq x\right) = \mathbb{P}\left(\underbrace{\frac{S_n - n\mu}{\sqrt{n}\sigma}}_{=\tilde{X}} \leq \frac{x - n\mu}{\sqrt{n}\sigma}\right) \approx \Phi\left(\frac{x - n\mu}{\sqrt{n}\sigma}\right).$$

Esimerkki

- $X_1, \ldots, X_n$ ovat riippumattomia satunnaismuuttujia Bernoullin jakaumasta, $\mathbb{P}(X_1 = 1) = p$ ja $\mathbb{P}(X_1 = 0) = 1 - p$.

Esimerkki

- $X_1, \ldots, X_n$ ovat riippumattomia satunnaismuuttujia Bernoullin jakaumasta, $\mathbb{P}(X_1 = 1) = p$ ja $\mathbb{P}(X_1 = 0) = 1 - p$.
- Voimme laskea, että $\mu = \mathbb{E}(X_1) = p$ ja $\sigma^2 = \mathrm{Var}(X_1) = p(1 - p)$.

Esimerkki

- $X_1, \ldots, X_n$ ovat riippumattomia satunnaismuuttujia Bernoullin jakaumasta, $\mathbb{P}(X_1 = 1) = p$ ja $\mathbb{P}(X_1 = 0) = 1 - p$.
- Voimme laskea, että $\mu = \mathbb{E}(X_1) = p$ ja $\sigma^2 = \mathrm{Var}(X_1) = p(1-p)$.

Saamme approksimaation

$$\mathbb{P}(S_n \leq x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right).$$

Esimerkki: luottokorttipetokset

- Oletetaan, että tiedämme petoksen todennäköisyyden olevan $p = 0.0016$.
- Mikä on todennäköisyys havaita vähintään $x = 492$ petosta, kun luottokorttitapahtumia oli kokonaisuudessaan $n = 284807$?
- Approksimoimme

$$\mathbb{P}(S_n \geq x) \approx 1 - \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right).$$

Likiarvoinen luottamusväli

Luottamusvälin määritelmä odotusarvolle

- Oletetaan, että X on satunnismuuttuja odotusarvolla $\mu = \mathbb{E}(X) \in (-\infty, \infty)$. Olkoon $\boldsymbol{X} = (X_1, \ldots, X_n)$ satunnaismuuttujan X havaintoja.

Luottamusvälin määritelmä odotusarvolle

- Oletetaan, että X on satunnismuuttuja odotusarvolla $\mu = \mathbb{E}(X) \in (-\infty, \infty)$. Olkoon $\boldsymbol{X} = (X_1, \ldots, X_n)$ satunnaismuuttujan X havaintoja.
- Luottamustason $1 - \alpha$ luottamusväli on satunnaisväli $[\ell(\boldsymbol{X}), u(\boldsymbol{X})]$, jolle pätee

$$\mathbb{P}(\mu \in [\ell(\boldsymbol{X}), u(\boldsymbol{X})]) \geq 1 - \alpha.$$

Luottamusvälin approksimointi (1)

Oletetaan, että satunnaismuuttujan X varianssi $0 < \sigma^2 < \infty$ on tiedossa. Tällöin keskeisen raja-arvolauseen mukaan

$$\mathbb{P}\left(z_\ell \leq \sqrt{n}\frac{\mu - \bar{X}}{\sigma} \leq z_u\right) \approx 1 - \alpha,$$

Luottamusvälin approksimointi (1)

Oletetaan, että satunnaismuuttujan X varianssi $0 < \sigma^2 < \infty$ on tiedossa. Tällöin keskeisen raja-arvolauseen mukaan

$$\mathbb{P}\left(z_\ell \leq \sqrt{n}\frac{\mu - \bar{X}}{\sigma} \leq z_u\right) \approx 1 - \alpha,$$

jossa väli $[z_\ell, z_u]$ valitaan niin, että

$$\Phi(z_u) - \Phi(z_\ell) = 1 - \alpha.$$

Yllä F on jakauman $\mathrm{N}\,(0, 1)$ kertymäfunktio.

Luottamusvälin approksimointi (2)

Valitsemalla esimerkiksi $z_u = z_{1-\alpha/2}$ ja $z_\ell = -z_u$ päädymme seuraavaan luottamusväliin

$$\left[\bar{X} - \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}, \bar{X} + \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}\right].$$

Luottamusvälin approksimointi (2)

Valitsemalla esimerkiksi $z_u = z_{1-\alpha/2}$ ja $z_\ell = -z_u$ päädymme seuraavaan luottamusväliin

$$\left[\bar{X} - \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}, \bar{X} + \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}\right].$$

Käytännössä yleensä (tuntematon) keskihajonta korvataan vielä otoskeskihajonnalla $\hat{\sigma}$, sillä suurten lukujen lain perusteella $\hat{\sigma} \approx \sigma$. Päädymme luottamusväliin

$$\left[\bar{X} - \frac{\hat{\sigma} z_{1-\alpha/2}}{\sqrt{n}}, \bar{X} + \frac{\hat{\sigma} z_{1-\alpha/2}}{\sqrt{n}}\right].$$

Esimerkki: luottokorttipetokset

- Otos sisältää luottokorttitapahtumia syyskuulta 2013. Tapahtumia kirjattiin kahdelta päivältä (Lähde: Kaggle).

Esimerkki: luottokorttipetokset

- Otos sisältää luottokorttitapahtumia syyskuulta 2013. Tapahtumia kirjattiin kahdelta päivältä (Lähde: Kaggle).
- Luottokorttitapahtumia on yhteensä 284 807, joista 492 luokiteltiin petoksiksi.

Esimerkki: luottokorttipetokset

- Otos sisältää luottokorttitapahtumia syyskuulta 2013. Tapahtumia kirjattiin kahdelta päivältä (Lähde: Kaggle).
- Luottokorttitapahtumia on yhteensä 284 807, joista 492 luokiteltiin petoksiksi.
- Tavoite: Estimoidaan 95% luottamusväli petoksien osuudelle luottokorttitapahtumista.

Malli

- Satunnaismuuttuja X noudattaa Bernoullijakaumaa tuntemattomalla parametrilla p, jossa p on petoksen todennäköisyys.
- $\mathbb{P}(X = 1) = p$ ja $\mathbb{P}(X = 0) = 1 - p$, jossa tapahtuma $\{X = 1\}$ indikoi petosta ja $\{X = 0\}$ vastaa normaalia luottokorttitapahtumaa.
- Havainnot $x_1, \ldots x_{284807}$ ovat binäärisiä (0 tai 1). Oletamme, että havainnot ovat riippumattomia ja samoin jakautuneita.

Ratkaisu

- Voimme laskea $\mathbb{E}(X) = p$ ja $\mathrm{Var}(X) = p(1-p)$.
- Suurten lukujen lain mukaan $p \approx \hat{p} = \frac{1}{n}\sum_{i=1}^{284807} x_i$.
- Joten voimme approksimoida 95% luottamusvälin seuraavasti

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

Ratkaisu (2)

Sijoittamalla lukuarvot

- $\hat{p} \approx 0.0017$, (R komento `mean(data)`) ja
- $z_{1-\alpha/2} \approx 1.96$ (R komento `qnorm(1 - 0.05 / 2)`)

saamme luottamusväliksi

$$\approx [0.0016, 0.0019] \, .$$

Ratkaisu (2)

Sijoittamalla lukuarvot

- $\hat{p} \approx 0.0017$, (R komento `mean(data)`) ja
- $z_{1-\alpha/2} \approx 1.96$ (R komento `qnorm(1 - 0.05 / 2)`)

saamme luottamusväliksi

$$\approx [0.0016, 0.0019] \, .$$

Luottamusvälin laskemiseen liittyvät yksityiskohdat löytyvät skriptistä `credit.R`.

Tiivistelmä

Sovelsimme keskeistä raja-arvolausetta luottokorttipetoksiin

1. approksimoimalla petoksien määrän todennäköisyyttä ja
2. laskemalla likiarvoistettuja luottamusvälejä petoksen todennäköisyydelle.