MS-A0502 Todennäköisyyslaskennan ja tilastotieteen peruskurssi

6B Kertaus ja yhteenveto

Emilia Blåsten

Matematiikan ja systeemianalyysin laitos
Perustieteiden korkeakoulu
Aalto-yliopisto

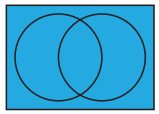
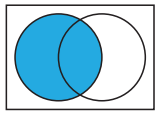
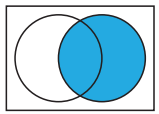
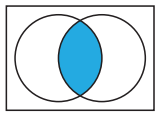
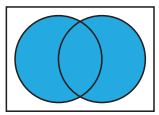
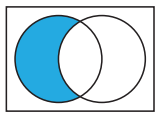
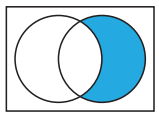
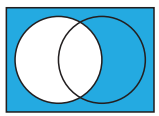
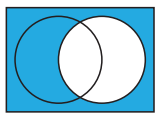
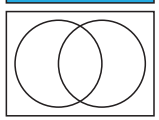
Lukuvuosi 2021–2022
Periodi II

Yleiset ohjeet tenttiin valmistautumiseen

Seuraavia kysymyksiä voi käyttää valmistautumisen pohjana (tai muita mieleenne tulevia kysymyksiä). Kannattaa käydä jokainen luento läpi kalvoineen ja pohtia seuraavia:

1. **Mitä ko. luennolla opitaan?** Luennon otsikko on noin 2–4 sanaa. Mitä aiheesta voisi kertoa 10–20 sanalla? Entä 50-100 sanalla, jos haluaa tiivistää ja selittää asian toiselle opiskelijalle?
2. Mikä on ko. luennon **tärkein kalvo**, joka kannattaa kerrata?
3. Miten luennon aihe liittyy kurssin **muihin luentoihin**? Käytetäänkö siinä aiempien luentojen menetelmiä, tai sen menetelmiä myöhemmin?
4. **Millaista matematiikkaa** ko. luennolla tarvitaan?
5. Millaisissa oman alasi tai muissa käytännön tilanteissa voit **käyttää** ko. luennon menetelmiä?
6. Mikä luennolla oli uutta/tuttua/yllättävää? Mikä oli helppoa/vaikeaa? Mitä lisäkysymyksiä tulee mieleesi?

L1A: Tn. peruslaskukaavat

Termi	Merkintä	Määritelmä	Venn-kaavio	Tulkinta
Perusjoukko	S	$\{x \in S : x \in S\}$		Varma tapahtuma
Osajoukko	A	$\{x \in S : x \in A\}$		A toteutuu
Osajoukko	B	$\{x \in S : x \in B\}$		B toteutuu
Leikkaus	$A \cap B$	$\{x \in S : x \in A \text{ ja } x \in B\}$		A ja B toteutuvat
Yhdiste	$A \cup B$	$\{x \in S : x \in A \text{ tai } x \in B\}$		A tai B toteutuu
Erotus	$A \setminus B$	$\{x \in S : x \in A \text{ ja } x \notin B\}$		A toteutuu mutta B ei
Erotus	$B \setminus A$	$\{x \in S : x \in B \text{ ja } x \notin A\}$		B toteutuu mutta A ei
Komplementti	A^c	$\{x \in S : x \notin A\}$		A ei toteudu
Komplementti	B^c	$\{x \in S : x \notin B\}$		B ei toteudu
Tyhjä joukko	$\emptyset$	$\{x \in S : x \notin S\}$		Mahdoton tapahtuma

L1A: Ehdollinen tn. ja Bayesin kaava

Ehdollinen todennäköisyys määritellään:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \text{kun } \mathbb{P}(B) \neq 0.$$

Riippumattomuus: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ tai $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Monen tapahtuman tulosääntö (kun $\mathbb{P}(A_{k-1} \cap \ldots \cap A_1) \neq 0$):

$$\begin{aligned}&\mathbb{P}(A_k \cap \cdots \cap A_1) \\ &\quad = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_2 \cap A_1) \cdots \mathbb{P}(A_k|A_{k-1} \cap \cdots \cap A_1).\end{aligned}$$

Osituskaava (kun $B_1, B_2, \ldots, B_n$ osittavat perusjoukon)

$$\mathbb{P}(A) = \sum_i \mathbb{P}(B_i)\mathbb{P}(A|B_i).$$

Bayesin kaava

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

L1A: Esim. harvinainen tauti

Erästä tautia esiintyy väestössä suhteessa 1/10000. Taudin toteamiseen on testi, joka tuottaa vääriä positiivisia ja vääriä negatiivisia tn:llä 1%. Millä tn satunnaisen henkilön testitulos on positiivinen?

H_- = "henkilö ei sairasta tautia" T_- = "testi on negatiivinen"
H_+ = "henkilö sairastaa tautia" T_+ = "testi on positiivinen"

$$\begin{aligned} \text{Osituskaava} \implies \mathbb{P}(T_+) &= \mathbb{P}(H_-)\mathbb{P}(T_+ \mid H_-) + \mathbb{P}(H_+)\mathbb{P}(T_+ \mid H_+) \\ &= 0.9999 \cdot 0.01 + 0.0001 \cdot 0.99 \\ &= 0.010098. \end{aligned}$$

Millä tn positiivisen testituloksen saanut henkilö sairastaa tautia? Aiemmin laskettiin $\mathbb{P}(T_+) = 0.010098$. Bayesin kaava $\implies$

$$\mathbb{P}(H_+ \mid T_+) = \frac{\mathbb{P}(H_+)\mathbb{P}(T_+ \mid H_+)}{\mathbb{P}(T_+)} = \frac{0.0001 \cdot 0.99}{0.010098} \approx 0.0098.$$

Onko tässä jotain outoa?

Esiintyvyysharha:

- Kaikista testituloksista 99% on oikeita
- Positiivisista testituloksista yli 99% on vääriä

L1A: Tn. kombinatorinen tulkinta

n alkion joukosta k alkion listoja

- toistojen kanssa n^k kappaletta,
- ilman toistoja $n(n-1)\cdots(n-k+1)$ kpl $(= n!/(n-k)!)$.

n alkiota voidaan järjestää listaan $n! = n(n-1)\cdots 2\cdot 1$ tavalla.

Järjestämättömiä k alkion osajoukkoja n alkion joukosta

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \text{ kpl.}$$

Esimerkit:

- puhelimen PIN-koodien lukumäärä,
- tavat jakaa SM-liigan mitalisijat olettaen ei jaettuja sijoja,
- tn. saada “kolmoset” pokerissa (ilman jokereita).

L1B: Sm. tiheys- ja kertymäfunktio

Diskreetti jakauma	**Jatkuva jakauma**
X:n arvojoukko äärellinen tai numeroituvasti ääretön	X:n arvojoukko ylinumeroituvasti ääretön
$\mathbb{P}(X = x) = f_X(x)$	$\mathbb{P}(X = x) = 0$ kaikilla x
Jakauma määräytyy kaavoilla	Jakauma määräytyy kaavoilla
$\mathbb{P}(X \in A) = \sum_{x \in A} f_X(x)$	$\mathbb{P}(X \in A) = \int_A f_X(x)\,dx$
$\mathbb{P}(X \leq t) = F_X(t)$	$\mathbb{P}(X \leq t) = F_X(t)$
Tiheysfunktion arvot ovat tarkkoja todennäköisyyksiä	Tiheysfunktion arvot ovat suhteellisia likiarvoisia todennäköisyyksiä
$f_X(x) = \mathbb{P}(X = x)$	$f_X(x) \approx h^{-1}\mathbb{P}(X = x \pm h/2)$
Esim. joukon $\{1, \ldots, 6\}$ tasajakauma	Esim. välin $[0, 10]$ tasajakauma

L1B: Kahden sm. yhteisjakauma ja reunajakaumat

$X_1 = 1.$ nopan silmäluku. $M = 1.$ ja $2.$ nopan maksimi.

	M						
X_1	1	2	3	4	5	6	Yht
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
2	0	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
3	0	0	$\frac{3}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
4	0	0	0	$\frac{4}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
5	0	0	0	0	$\frac{5}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
6	0	0	0	0	0	$\frac{6}{36}$	$\frac{1}{6}$
Yht	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	

Rivisummista saadaan X_1:n jakauma
Sarakesummista saadaan M:n jakauma

L1B: Sm. ehdollinen jakauma

Y:n ehdollinen tiheysfunktio X:n suhteen määritellään kaavalla

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

Tulkinta diskreetillä sm:llä $\mathbb{P}(Y = y|X = x) = f_{Y|X}(y|x)$.

Stokastinen riippumattomuus:

Palauttaen

	X_2		
X_1	0	1	Yht
0	$\frac{77}{80} \times \frac{77}{80}$	$\frac{77}{80} \times \frac{3}{80}$	$\frac{77}{80}$
1	$\frac{3}{80} \times \frac{77}{80}$	$\frac{3}{80} \times \frac{3}{80}$	$\frac{3}{80}$
Yht	$\frac{77}{80}$	$\frac{3}{80}$	

$$f_{X_1,X_2}(i,j) = f_{X_1}(i) f_{X_2}(j)$$

Palauttamatta

	X_2		
X_1	0	1	Yht
0	$\frac{77}{80} \times \frac{76}{79}$	$\frac{77}{80} \times \frac{3}{79}$	$\frac{77}{80}$
1	$\frac{3}{80} \times \frac{77}{79}$	$\frac{3}{80} \times \frac{2}{79}$	$\frac{3}{80}$
Yht	$\frac{77}{80}$	$\frac{3}{80}$	

$$f_{X_1,X_2}(i,j) \neq f_{X_1}(i) f_{X_2}(j)$$

L2A: Satunnaisluvun odotusarvo

Diskreetti:

$$\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x) = \sum_x xf(x)$$

Jatkuva:

$$\mathbb{E}(X) = \int xf(x)dx$$

Tulkinta: “monen riippumattoman toiston keskiarvo likimain odotusarvo”

Esim

Reilun nopanheiton silmäluvun X odotusarvo

$$\mathbb{E}(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

L2A: Suurten lukujen laki

Lause (Suurten lukujen laki)

Jos $X_1, X_2, X_3, \ldots$ ovat keskenään riippumattomia X:n tavoin jakautuneita satunnaislukuja, niin esim. tapahtuman

$$\frac{1}{n}\sum_{s=1}^{n} X_s \;=\; \mathbb{E}(X) \pm 0.001$$

todennäköisyys lähestyy ykköstä suurilla n arvoilla.

Esim (Suhteellinen esiintyvyys)

- Diskreetille tiheysfunktiolle $f(x) = \mathbb{P}(X = x)$:

$$\frac{\#\{s \;:\; X_s = x\}}{n} \;\approx\; f(x)$$

- Kertymäfunktiolle $F(t) = \mathbb{P}(X \leq t)$:

$$\frac{\#\{s \;:\; X_s \leq t\}}{n} \;\approx\; F(t)$$

L2A: Sm. muunnoksen odotusarvo

Fakta

- *Diskreetille satunnaismuuttujalle*

$$\mathbb{E}(g(X)) = \sum_x g(x)\, f(x).$$

- *Jatkuvalle satunnaismuuttujalle*

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)\, f(x) dx.$$

Sama pätee monen muuttujan muunnoksen odotusarvolle

- disk. $\mathbb{E}(g(X,Y)) = \sum_x \sum_y g(x,y)\, f(x,y)$,
- jva. $\mathbb{E}(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)\, f(x,y) dx dy$.

L2B: Varianssi, keskihajonta

Jos $\mu = \mathbb{E}(X)$, niin varianssi

$$\mathrm{Var}(X) = \mathbb{E}\big((X-\mu)^2\big) = \mathbb{E}(X^2) - \mu^2$$

ja keskihajonta

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

kuvaavat, paljonko X yleensä poikkeaa odotusarvostaan, eli jakauman f_X leveyttä.

Laskentakaavat diskreetille (muunnoksen odotusarvo)

$$\mathrm{SD}(X) = \sqrt{\sum_x (x-\mu)^2 f(x)},$$

$$\mathrm{SD}(X) = \sqrt{\mathbb{E}(X^2) - \mu^2} \quad \& \quad \mathbb{E}(X^2) = \sum_s x^2 f(x).$$

L2B: Chebyshevin epäyhtälö

Fakta (Tšebyšovin epäyhtälö, Chebyshev's inequality)

Jokaiselle satunnaisluvulle odotusarvona μ ja keskihajontana σ, tapahtuman $\{X = \mu \pm 2\sigma\} = \{\mu - 2\sigma \leq X \leq \mu + 2\sigma\}$ todennäköisyys on vähintään

$$\mathbb{P}(X = \mu \pm 2\sigma) \geq \frac{3}{4}.$$

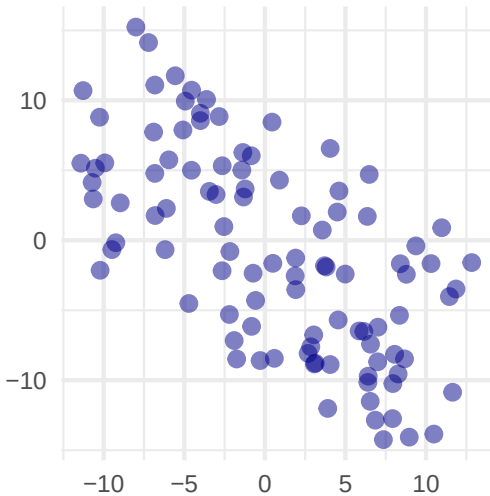
Yleisemmin $\mathbb{P}(X = \mu \pm r\sigma) \geq 1 - \frac{1}{r^2}$ kaikilla $r \geq 1$.

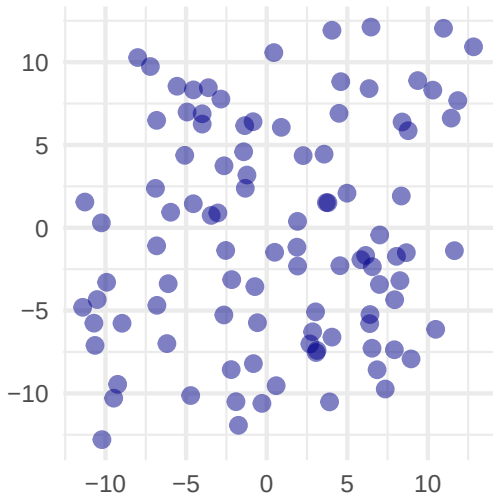
- X:n arvo sijaitsee melko todennäköisesti (tn $\geq 75\%$) kahden keskihajonnan sisällä odotusarvostaan
- X:n arvo sijaitsee hyvin todennäköisesti (tn $\geq 99\%$) kymmenen keskihajonnan sisällä odotusarvostaan

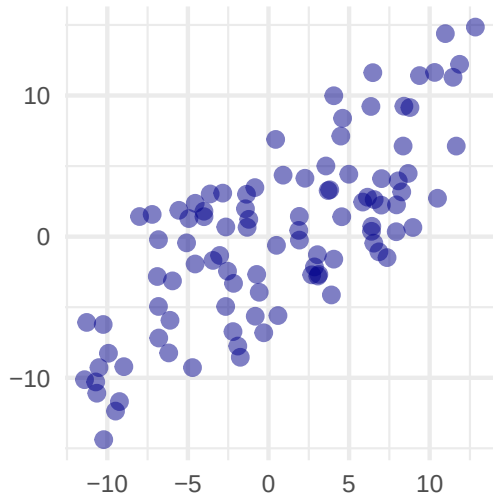
Tšebyšovin epäyhtälö antaa keskiosan tn:lle alarajan (ja häntätodennäköisyydelle ylärajan). Jos jakauman muoto tiedetään, voidaan saada tiukempiakin rajoja.

L2B: Kovarianssi, korrelaatio

Miten mitataan kahden satunnaismuuttujan X ja Y yhteisvaihtelua (suunta ja voimakkuus)?







$$\mathrm{Cov}(X, Y) = \mathbb{E}\big((X - \mu_X)(Y - \mu_Y)\big) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y),$$

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)}.$$

L3A: Satunnaislukujen summa

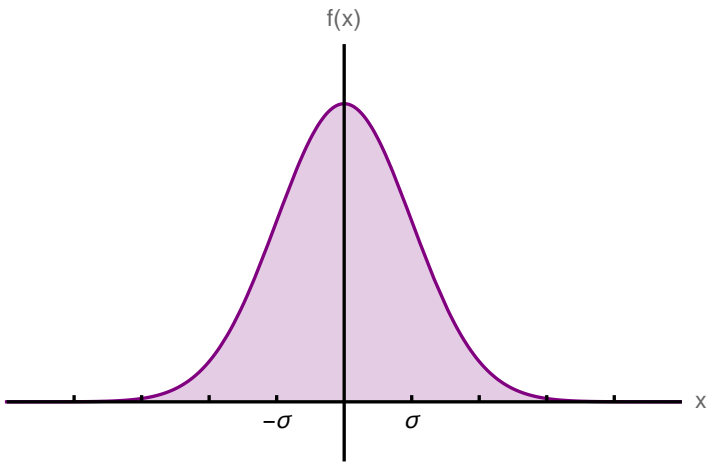
Kertaa, miten jakauma lasketaan yksinkertaisissa esimerkeissä.

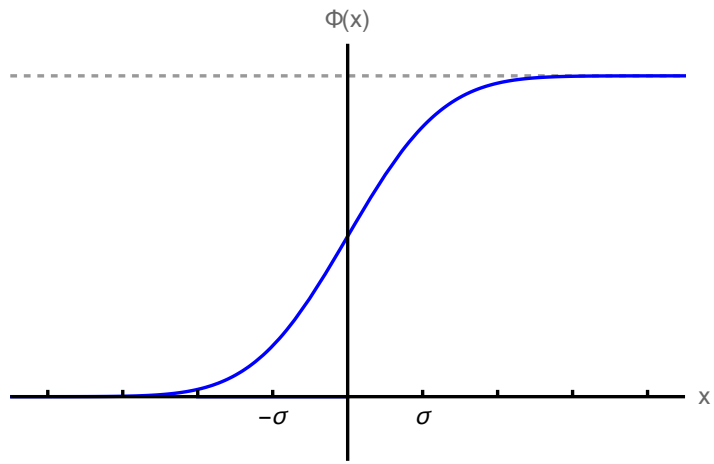
Satunnaislukujen $X_1, \ldots, X_n$ summan odotusarvo ja keskihajonta, kun $\mu_i = \mathbb{E}(X_i)$, $\sigma_i = \mathrm{SD}(X_i)$ ja $\rho_{i,j} = \mathrm{Cor}(X_i, X_j)$:

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	$\sigma \sqrt{n}$

L3A: Normaalijakauma, tapahtumien tn. laskusäännöt

Standardinormaalijakauma. Tiheysfunktio $f(x)$, kertymäfunktio $\Phi(x)$:





Jos X noudattaa ei-standardinormaalijakaumaa, niin normitettu satunnaisluku $Z = \frac{X-\mu_X}{\sigma_X}$ noudattaa standardinormaalijakaumaa ($\mu_Z = 0$ ja $\sigma_Z = 1$).

Välin todennäköisyys $\mathbb{P}(a \leq Z \leq b) = \Phi(b) - \Phi(a)$ kun $b \geq a$. Oikean hännän tn. $\mathbb{P}(Z \geq c) = 1 - \Phi(c)$.

L3A: Keskeinen raja-arvolause

Fakta (Keskeinen raja-arvolause)

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

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

suurilla n, missä Z:n tiheysfunktio on normitettu normaalijakauma.

$A \stackrel{\mathrm{d}}{\approx} B$ tarkoittaa, että A:n ja B:n jakaumat ovat likimain samat.

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

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.)

L3B: Datajoukon tunnusluvut

Lukuarvoinen yhden muuttujan datajoukko $\vec{x} = (x_1, \ldots, x_n)$

Keskiarvo $m(\vec{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i$

Varianssi $\mathrm{var}(\vec{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - m(\vec{x}))^2$

Keskihajonta $\mathrm{sd}(\vec{x}) = \sqrt{\mathrm{var}(\vec{x})}$

Esim. $\vec{y} = (0, 0, 1, 1, 2, 2)$

$$m(\vec{y}) \;=\; \frac{1}{6}\Big(0+0+1+1+2+2\Big) \;=\; 1$$

$$\mathrm{var}(\vec{y}) \;=\; \frac{1}{6}\Big((0-1)^2+(0-1)^2+(1-1)^2+(1-1)^2+(2-1)^2+(2-1)^2\Big) \;=\; \frac{2}{3}$$

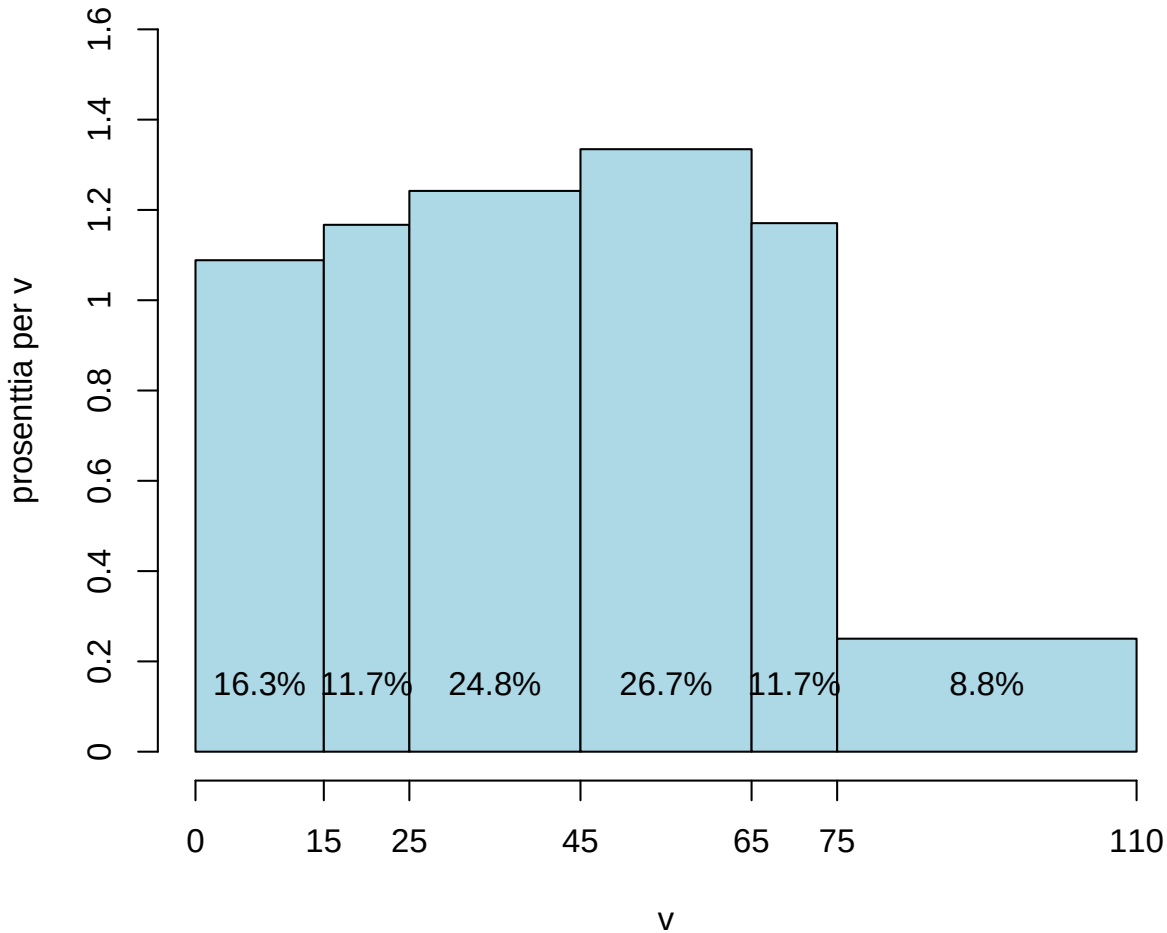
$$\mathrm{sd}(\vec{y}) \;=\; \sqrt{\frac{2}{3}} \;\approx\; 0.8165$$

Iso X satunnaismuuttuja. Pieni x sm:n arvo. $\vec{X}$ jono sm:iä. $\vec{x}$ jono havaintoja / dataa.

L3B: Datan havainnollistaminen

Histogrammi

Suomen väestörakenne ikäluokittain 31.12.2015 [Lähde: Tilastokeskus]



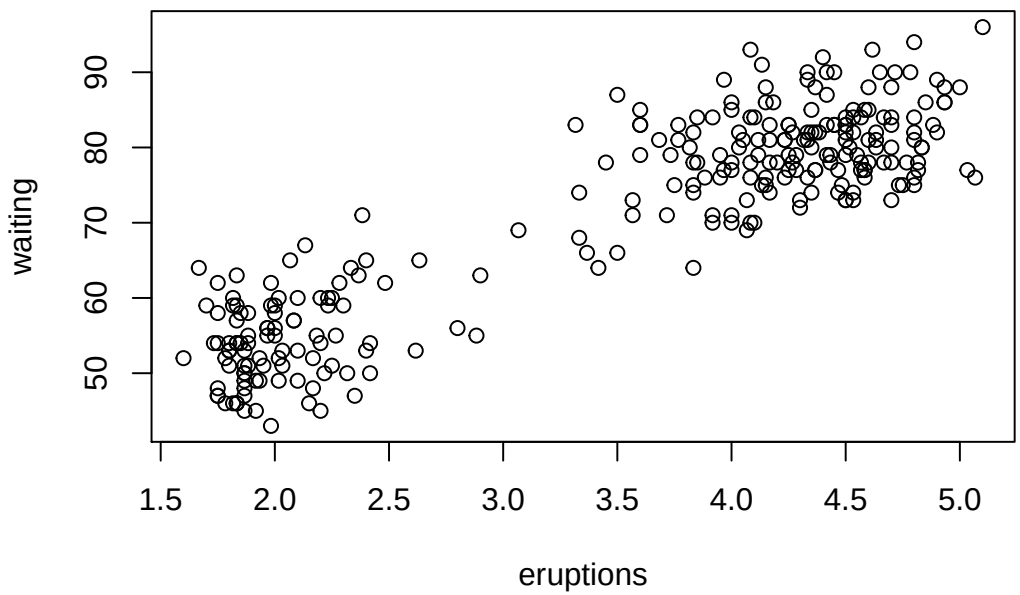
Ikä (v)	Lukumäärä
0–14	896 023
15–24	640 387
25–44	1 363 155
45–64	1 464 640
65–74	642 428
75–	480 675

Palkin leveys $\propto$ luokkavälin leveys.

L3B: Datan havainnollistaminen

Hajontakuvio

Hajontakuvio, 272 purkausta, *Old Faithful* (Yellowstone).
2 muuttujaa: Purkauksen kesto ja väliaika seuraavaan purkaukseen.



L3B: Datajoukko VS generoiva ilmiö

data	populaatio
Pearsonin 1000 isää ja poikaa	Kaikki isät ja pojat (missä? milloin?)
1000 gallup-vastausta	5 miljoonan suomalaisen mielipide (nyt)
272 geysirin purkausta	Kaikki Old F:n purkaukset (menneet/tulevat?)
Lääkkeen vaikutus 30 potilaalla	Vaikutus tulevilla potilailla
100 nopanheittotulosta	Potentiaalinen ääretön heittojono

Populaatio on tilastotieteen terminologiaa, ja tarkoittaa

- mistä/miten data on syntynyt (generoiva mekanismi; datalähde)
- se mitä datan perusteella yritetään ymmärtää

"Populaatio" ei välttämättä ole mikään konkreettinen kokoelma (esim. ihmisiä).

L4A: Tilastollisen päättelyn tarkoitus

Tavoitteena tehdä päätelmiä havaitun datan pohjalta.

1. Valitaan tilanteeseen sopiva stokastinen malli
   (jakaumaperhe, esim. “kaikki normaalijakaumat” tai “kaikki tasajakaumat $[0, m]$”)
2. Sovitetaan malli havaittuun dataan
   (estimoidaan mallin parametrit)
3. Lasketaan sovitetusta mallista tarvittavat tunnusluvut
4. Tehdään johtopäätökset

L4A: Parametrisoidut jakaumat

Monesti on hyödyllistä ajatella, että data tulee esim. *jostain* normaalijakaumasta tai *jostain* binomijakaumasta.

Esimerkkejä:

- Bernoulli: $f_p(1) = p$, $f_p(0) = 1 - p$, parametri $0 \leq p \leq 1$.
- Eksponentti: $f_\lambda(x) = \lambda e^{-\lambda x}$, $x > 0$, parametri $\lambda > 0$.
- Normaali: $f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, parametrit $\mu \in \mathbb{R}, \sigma^2 > 0$.
- Beta: $f_{\alpha,\beta}(x) = cx^{\alpha-1}(1-x)^{\beta-1}$, $0 \leq x \leq 1$, parametrit $\alpha, \beta \geq 0$. Huom! ei c.
- Binomi: $f_{n,p}(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, 2 \ldots, n$, parametrit $n \in \mathbb{N}, 0 \leq p \leq 1$.
- ...

L4A: Tilastolliset estimaattorit

1. Tuntematon datalähde. Tutkittavan suureen jakauma $f_\theta(x)$ ja parametri θ tuntematon.
2. Datalähteestä on saatu n riippumatonta ja samoin jakautunutta havaintoa $x_1, \ldots, x_n$.
3. Parametrin θ
   - estimaatti on datan $\vec{x} = (x_1, \ldots, x_n)$ pohjalta laskettu arvaus $\hat{\theta} = g(\vec{x})$,
   - estimaattori on funktio $(x_1, \ldots, x_n) \mapsto g(x_1, \ldots, x_n)$, joka kuvaa datan estimaatiksi.

Suurimman uskottavuuden estimaattori: Maksimoi

$$L(\theta) = f_\theta(x_1) \cdot f_\theta(x_2) \cdots f_\theta(x_n).$$

Harhattomuus (unbiased): $\mathbb{E}\big(\hat{\theta}(\vec{X})\big) = \theta$.

Osaa nämä binomijakauman p:lle ja normaalijakauman μ, σ^2:lle.

L4B: Stokastinen malli

Havaittu data

Datalähteestä on *havaittu* arvot $x_1, \ldots, x_n$. Halutaan päätellä (=arvata) tutkittavan suureen (tuntematon) jakauma $f(x)$.

Stokastinen malli

Tilastokokeen *mahdollista* tulosta mallinnetaan satunnaismuuttujilla $X_1, \ldots, X_n$, jotka ovat toisistaan riippumattomat ja noudattavat (tuntematonta tai oletettua) jakaumaa $f(x)$.

Stokastinen malli (generoiva malli, generoiva jakauma) kuvaa datalähteen toimintaa *yleensä* (millaisia lukuja se *voi* tuottaa ja millä tn:llä).

Stokastisen mallin mahdollisesta tuloksesta $X_1, \ldots, X_n$ voidaan laskea tunnuslukuja, niiden jakaumia, ja tunnuslukujen jakaumien tunnuslukuja. Esimerkiksi keskiarvon $m(\vec{X}) = \frac{1}{n} \sum_{i=1}^{n} X_i$ odotusarvo ja keskihajonta.

L4B: Luottamusväli

Stokastinen malli: $X_1, \ldots, X_n$ riippumattomia ja samoin jakautuneita satunnaislukuja odotusarvona μ ja keskihajontana σ. Meskiarvo $m(\vec{X})$ on parametrin μ estimaatti.

Normitetun virheen

$$\frac{m(\vec{X}) - \mu}{\sigma/\sqrt{n}}$$

odotusarvo on 0 ja keskihajonta 1.

99%:n luottamusväli estimaatille on $[m(\vec{X}) - \ell, m(\vec{X}) + \ell]$ jos pätee

$$\mathbb{P}(|m(\vec{X}) - \mu| \leq \ell) \approx 99\%.$$

Käytännössä mallin tuottamasta datasta lasketaan tuo väli. Mutta ℓ pitää laskea mallista, tai estimoida datasta myös.

Todennäköisyys johtuu datan generoinnin satunnaisuudesta, ei mallivirheestä, tai muusta epätarkkuudesta. Jos datasta $m(\vec{x}) = 5.1$ ja $\ell = 0.7$ luottamustasolla 99%, ei voida sanoa, että $\mu \in [4.4, 5.8]$ 99% todennäköisyydellä.

L4B: Normaalimallin μ-estimaatin luottamusvälin laskenta

Datalähteen normaalimalli: $X_1, X_2, \ldots$ riippumattomia ja normaalijakautuneita odotusarvona μ (tuntematon) ja keskihajontana σ (tunnettu)

Luottamusvälin määrittäminen:

1. Lasketaan havaitusta datasta keskiarvo $m(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$
2. Määritetään luku $z > 0$, jolle $\mathbb{P}(|Z| \leq z) = 1 - 2\Phi(-z) = 0.99$ $\implies$ $z = -\Phi^{-1}(\frac{1-0.99}{2}) \approx 2.58$
3. Asetetaan parametrin μ luottamusväliksi $m(\vec{x}) \pm z\frac{\sigma}{\sqrt{n}}$

Datalähteen tuottamalle satunnaisvektorille $\vec{X} = (X_1, \ldots, X_n)$

$$\mathbb{P}\left(|m(\vec{X}) - \mu| \leq z\frac{\sigma}{\sqrt{n}}\right) = \mathbb{P}\left(\left|\frac{m(\vec{X}) - \mu}{\sigma/\sqrt{n}}\right| \leq z\right) = \mathbb{P}(|Z| \leq z) = 99\%$$

Jos σ myös tuntematon, niin luottamusväliksi $m(\vec{x}) \pm z\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}$ ja

- Jos dataa on paljon (n iso), likimain normitettu normaalijakauma
- Jos dataa on vähän, korvataan $\mathrm{sd}(\vec{x})$ otoskeskihajonnalla $\mathrm{sd}_s(\vec{x})$ ja lasketaan $z = -F^{-1}_{t,n-1}(\frac{1-0.99}{2})$ t-jakaumasta

L4B: Binaarimallin p-estimaatin luottamusväli

Datalähteen binaarimalli: $X_1, X_2, \ldots$ riippumattomia ja $\{0,1\}$-arvoisia satunnaislukuja odotusarvona p (tuntematon)

Likiarvoisen luottamusvälin (n suuri) määrittäminen:

1. Lasketaan havaitusta datasta ykkösten suhteellinen osuus $\hat{p} = \hat{p}(\vec{x})$
2. Määritetään luku $z > 0$, jolle
   $\mathbb{P}(|Z| \leq z) = 1 - 2\Phi(-z) = 0.95$
   $\implies z = -\Phi^{-1}(\frac{1-0.95}{2}) \approx 1.96$
3. Asetetaan parametrin p luottamusväliksi $\hat{p} \pm z\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$.
   Konservatiivinen luottamusväli $\hat{p} \pm z\frac{0.5}{\sqrt{n}}$.

L5A: Bayesiläisen päättelyn filosofia

- Todennäköisyys kuvaa sekä satunnaisuutta, että myös epävarmuutta, epätietoutta.
- Stokastisen mallin (jakaumaperheen) tuntemattomalle parametrille muodostetaan piorijakauma kuvaamaan ennakkokäsitystämme siitä.
- Saatujen havaintojen nojalla priorijakauma päivitetään posteriorijakaumaksi Bayesin kaavan nojalla.
- Jos saadaan lisää havaintoja, edellinen posteriori päivitetään uudestaan.
- Posteriorista tehdään päätelmiä tuntemattomasta satunnaisprosessista tai ennustetaan tulevien havaintojen todennäköisyyksiä.

L5A: Posteriorijakauman laskeminen

Tuntematon parametri: Heitetyn kolikon tyyppi Θ

Priorijakauma $f(\theta) = \mathbb{P}(\Theta = \theta)$

Data $x = 0$ (havaittiin klaava)

Uskottavuus $f(x \mid \theta) = \mathbb{P}(X = x \mid \Theta = \theta)$

θ	Priori $f(\theta)$	Uskottavuus $f(0 \mid \theta)$	Tulo	Posteriori $f(\theta \mid 0)$
0	0.1	1.00	0.100	0.20
0.25	0.1	0.75	0.075	0.15
0.5	0.6	0.50	0.300	0.60
0.75	0.1	0.25	0.025	0.05
1	0.1	0.00	0.000	0.00
summa	1.0		0.500	1.00

L5A: Binomi-, beta-, ja normaalijakaumat Bayesiläisittäin

Betajakauma kun uskottavuus on Bernoulli- tai binomijakautunut

Kolikkoa n kertaa heitettäessä havaittiin x kruunaa. Kolikosta ei ole taustatietoja. Määritä parametrin Θ (kruunan tn) posteriorijakauma.

Priorijakauman tiheysfunktio: $f(\theta) = 1$, $\theta \in [0, 1]$
Uskottavuusfunktio datapisteelle x saadaan Bin(n, θ)-jakaumasta

$$f(x \mid \theta) \;=\; \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Posterioritiheys

$$f(\theta \mid x) \;=\; \frac{f(\theta) f(x \mid \theta)}{\int f(t) f(x \mid \theta')\, d\theta'} \;=\; c\, \theta^x (1-\theta)^y$$

on Beta$(x+1, y+1)$, missä $y = n - x$ on klaavojen lkm.

Priorijakauma $\Theta \sim \text{Nor}(\mu_0, \sigma_0)$
Uskottavuus: $(X_i \mid \theta) \sim \text{Nor}(\theta, \sigma)$

Fakta

Bayeslaisen normaalimallin posteriorijakauma havaitun datajoukon $\vec{x} = (x_1, \ldots, x_n)$ *suhteen on normaalijakauma* $\text{Nor}(\mu_1, \sigma_1)$, *missä*

$$\mu_1 = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2} m(\vec{x})}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \qquad \sigma_1 = \frac{1}{\sqrt{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}},$$

ja $m(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$ *on havaitun datajoukon keskiarvo.*

L5B: Posteriorijakauman käyttö

Mihin posteriorijakaumaa voi *käyttää*?

1. Mihin kysymykseen *halutaan* vastaus?
2. Mitä on (riittävän) helppo laskea?

Tyypillisiä tarkasteltavia asioita:

- posteriorin *moodi* = piste, jossa sen arvo on suurin
- posteriorin *odotusarvo* = tn-painotettu keskiarvo
- posteriorin *mediaani* = piste, jonka alla 50% tn:stä
- *uskottavuusväli*, joka sisältää esim. 95% posteriorista
- raportoidaan / piirretään *koko* posteriorijakauma
- tulevan datan *ennustaminen* posteriorin perusteella

L5B: Ennustejakauman laskeminen

Viisi havaintoa $\vec{x}$, ennustetaan kuudes havainto Y. Osituskaavan (additiivisuuden) perusteella

$$f_{Y|\vec{X}}(y \mid \vec{x}) = \int f(y \mid \theta) f(\theta \mid \vec{x}) d\theta.$$

Tulkinta: Eri θ-arvot antavat erilaisia ennusteita Y:lle. Osituskaava ottaa nämä kaikki mahdollisuudet huomioon ja *painottaa* niitä Θ:n posterioritiheyden mukaisesti.

Ennustetaan ovatko seuraavat kolme heittoa kruunia eli $\vec{Y} = (1, 1, 1)$.

- stokastinen malli $\mathbb{P}(Y_i = 1 \mid \Theta = \theta) = \theta$.
- Θ:n posteriorijakauma on Beta(5,2).

Joten lasketaan

$$f_{\vec{Y}|\vec{X}}(1, 1, 1 \mid \vec{x}) = \int f_{\vec{Y}|\Theta}(1, 1, 1 \mid \theta)\ f_{\Theta|\vec{X}}(\theta \mid \vec{x}) d\theta.$$

L5B: Multinomimalli

- n riippumatonta satunnaismuuttujaa $\vec{X} = (X_1, X_2, \ldots, X_n)$.
- Kullakin X_i sama diskreetti jakauma k:sta vaihtoehdosta
- Jakaumalla on k kpl todennäköisyysparametreja $\vec{p} = (p_1, p_2, \ldots, p_k)$
- Jos näitä pidetään satunnaismuuttujina, muodostuu *satunnaisvektori* $\vec{P} = (P_1, P_2, \ldots, P_k)$

Jo opitut menetelmät toimivat (kunhan yksityiskohdat selvitetään)

- Parametrivektorille oletetaan priorijakauma $f_{\vec{P}}(\vec{p})$
- Datalle oletetaan stokastinen malli $f(\vec{x} \,|\, \vec{p})$ (uskottavuus)
- Havaintojen perusteella lasketaan posteriori $f(\vec{p} \,|\, \vec{x})$

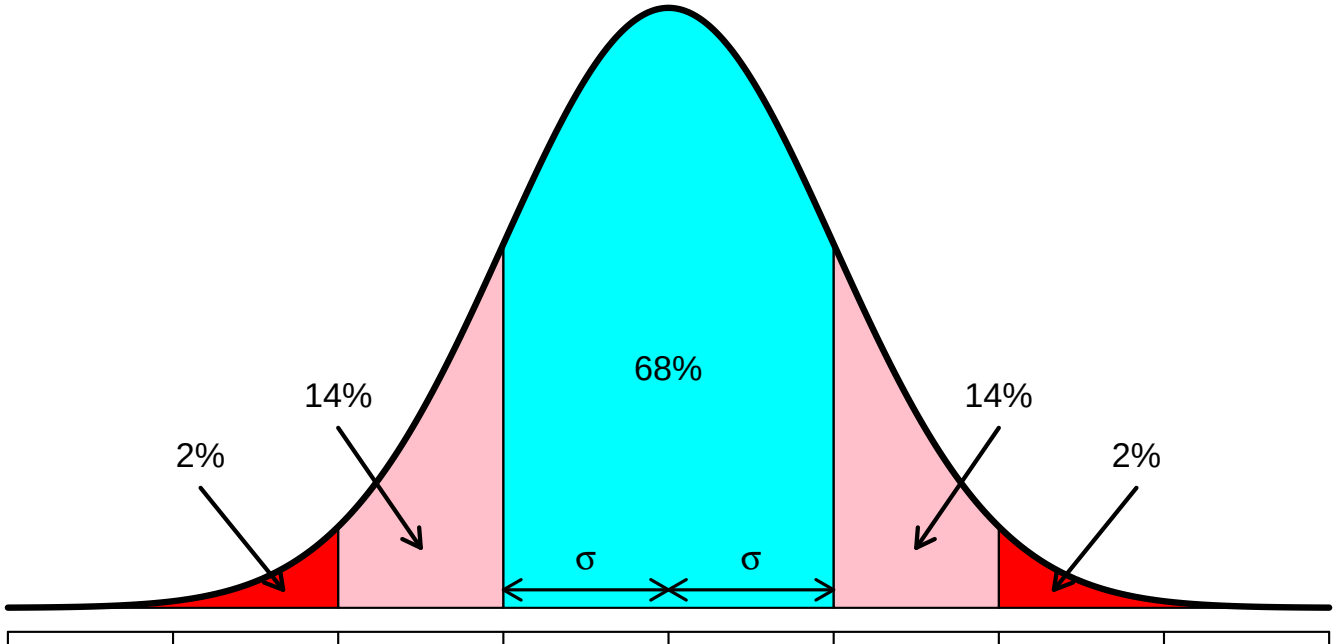
Saattaa tulla moniulotteinen jakauma, jos $k \geq 3$ ja on jatkuva parametri.

L6A: Hypoteesin testaus

Kuuluuko havaittu data hypoteesin perusteella liian kauas jakauman häntään?

1. Muotoile hypoteesi H_0 siitä, miten data syntyy.
2. Muotoile tunnusluku $t = t(\vec{X})$, joka lasketaan datasta.
3. Päättele t:n jakauma (jos H_0 on totta).
4. Laske arvoa $t(\vec{x})$ vastaavat häntä-tn:n molemmilla puolilla yhteensä. Tämä on p-arvo
5. Hylkää H_0 jos $p < \alpha$.

Jos $\alpha = 4\%$ niin hylätään H_0 jos $t(\vec{x})$ on punaisella alueella.



L6A: Hypoteesin testaus

Suuren datajoukon testi

11.05 9.65 10.93 9.46 10.27 10.02 10.07 10.74 11.15 10.40 10.12 11.20
10.07 10.27 9.99 9.80 10.83 10.21 11.26 10.11 10.49 10.10 10.15 11.02
10.00 11.68 10.51 11.20 11.29 10.15

Datajoukon keskiarvo $m(\vec{x}) = 10.473$, keskihajonta $\mathrm{sd}(\vec{x}) = 0.563$

H_0: $\mu = 10.0$
H_1: $\mu \neq 10.0$

Havaitun datajoukon testisuure:

$$t(\vec{x}) \;=\; \frac{m(\vec{x}) - \mu_0}{\mathrm{sd}(\vec{x})/\sqrt{n}} \;=\; \frac{10.473 - 10.0}{0.563/\sqrt{30}} \;=\; 4.60$$

Koska datajoukon koko $n = 30$ on melko suuri, käytetään suuren datajoukon testiä, ts. pidetään tunnuslukua $t(\vec{X})$ standardinormaalijakautuneena.

$$\text{p-arvo} \;\approx\; \mathbb{P}(|t(\vec{X})| \geq |t(\vec{x})| \,|\, H_0) \;\approx\; \mathbb{P}(|Z| \geq 4.60) \;\approx\; 4.2 \times 10^{-6}$$

Johtopäätös: Hyvin pieni p-arvo puoltaa vahvasti H_0:n hylkäämistä.

L6A: Epägaussinen data ja/tai tuntematon varianssi

Jos datalähteen tuottamat luvut noudattavat jotain muuta kuin normaalijakaumaa, niin keskeinen raja-arvolause! Testisuure aivan kuten normaalimallissa:

$$t(\vec{x}) \;=\; \frac{m(\vec{x}) - \mu_0}{\mathrm{sd}(\vec{x})/\sqrt{n}}.$$

Usein **datalähteen** keskihajonta σ ei ole etukäteen tiedossa, vaan estimoidaan otoksen keskihajonnasta $\mathrm{sd}(\vec{x})$.

Jos otos on suuri (esim. $n > 30$), estimaatti on melko tarkka.

Jos otos on pieni, testisuure

$$t(\vec{X}) = \frac{m(\vec{X}) - \mu_0}{\mathrm{sd}(\vec{X})\,\sqrt{n}}$$

on kahden satunnaismuuttujan osamäärä, eikä se olekaan normaalijakautunut, vaan t-jakautunut parametrilla $n - 1$.

L6A: Hyväksyminen ja hylkääminen

		Tehty johtopäätös	
		H_0 hyväksytään	H_0 hylätään
Totuus	H_0 tosi	Oikea päätös	Hylkäysvirhe
	H_0 epätosi	Hyväksymisvirhe	Oikea päätös

$p(\vec{x})$ = testisuureen p-arvo datajoukolle $\vec{x}$
$\vec{X} = (X_1, \ldots, X_n)$ mallintaa datalähteen tuottamia arvoja ennen niiden havaitsemista $\implies$ $p(\vec{X})$ on satunnaisluku
Hylkäysvirheen todennäköisyys on

$$\mathbb{P}(H_0 \text{ hylätään} \,|\, H_0) \;=\; \mathbb{P}(p(\vec{X}) < \alpha \,|\, H_0)$$

Hyväksymisvirheen todennäköisyys on

$$\mathbb{P}(H_0 \text{ hyväksytään} \,|\, H_1) \;=\; \mathbb{P}(p(\vec{X}) \geq \alpha \,|\, H_1).$$