Luku 1

Johdanto

1.1 Todennäköisyys ja tilastotiede

- Kurssi käsittelee todennäköisyyslaskentaa ja tilastotiedettä.
- Laaditaan satunnaisilmiöille todennäköisyysmalleja.
- Miten hyvin mallit kuvaavat todellisuutta? Tarvitaan havaintoja.

1.2 Havaitut frekvenssit ja empiiriset jakaumat

- *Satunnaiskoe,* esimerkiksi lantin heitto.
- Tapahtuman A lukumäärää eli *frekvenssi* n:n kokeen sarjassa $N_n(A)$, esimerkiksi klaavojen lkm 100:n ($n = 100$) heiton sarjassa.
- *Suhteellinen frekvenssi* $0 \leq \frac{N_n(A)}{n} \leq 1$.
- $\frac{N_n(A)}{n}$ lähenee lukua $P(A)$, kun toistojen lukumäärä n kasvaa.
- Todetaan, että $0 \leq P(A) \leq 1$.
- Luku $P(A)$ on tapahtuman A todennäköisyys.
- $\frac{N_n(A)}{n}$ on ominaisuuksiltaan todennäköisyyden kaltainen.
- Suurten lukujen laki

Empiirinen jakauma

- Luvut $x_1, x_2, \ldots, x_n$ ovat tavallisesti annetun *tilastollisen muuttujan* x, kuten esimerkiksi pituuden, painon jne., *mittalukuja* jossain havaintoaineistossa.

- $x_1, x_2, \ldots, x_n$ muodostavat muuttujan x *empiirisen jakauman.*
- Empiirisen jakauman $x_1, x_2, \ldots, x_n$ *empiirinen kertymäfunktio* (ekf) (reaalilukuakselilla) on

$$F_n(a) = \frac{1}{n}|\{\, i : 1 \leq i \leq n,\, x_i \leq a \,\}|,$$

  missä $-\infty < a < \infty$ ja $|\,.\,|$ on joukon alkioiden lukumäärä.
- Lukujen $x_1,\ x_2,\ \ldots,\ x_n$ *empiirinen jakaumafunktio* on

$$P_n(a, b) = F_n(b) - F_n(a).$$

- $P_n(a, b)$ on puoliavoimelle välille $(a, b]$ kuuluvien lukujen suhteellinen osuus lukujoukossa $\{x_1, x_2, \ldots, x_n\}$
- Histogrammi on empiirisen jakauman kuvaaja.
- Histogrammin piirtäminen: Valitaan ensin *jakopisteet* $b_1 < b_2 < \cdots < b_m$.

1.3 Todennäköisyysmallit

1.3.1 Satunnaiskoe

- Todennäköisyyslaskenta on *satunnaisilmiöiden* matemaattista teoriaa.
- *Satunnaiskoe,* esimerkiksi lantin heitto. Mahdolliset tulosvaihtoehdot tiedetään, yksittäisen kokeen tuloksesta ei varmuutta.
- Satunnaiskokeen kaikkien mahdollisten tulosten joukko Ω on *otosavaruudeksi.* Esimerkiksi

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\},$$

  missä $\omega_1,\ \omega_2,\ \ldots,\ \omega_n$ ovat alkeistapauksia eli satunnaiskokeen tulosvaihtoehtoja. Alkeistapausten lukumäärä $|\Omega| = n$. Otosavaruus voi olla myös ääretön.
- *Tapahtuma* on otosavaruuden Ω osajoukko. Esimerkiksi tapahtuma $A \subset \Omega$.
- Esimerkkejä satunnaiskokeista
  1. Heitetään tavallista noppaa, $\Omega = \{1, 2, 3, 4, 5, 6\}$.
  2. Valitaan kortti tavallisesta korttipakasta, Ω on 52:n kortin pakka. Esimerkiksi ”ruutu2”on alkeistapaus ja tapahtuma $A \equiv$ ”saadaan ässä”on $A = \{ruutu, hertta, risti, pata\} \subset \Omega$.

3. Tarkastellaan laitteen kestoa. $\omega > 0$ on laitteen kesto:

$$\Omega = \{\, \omega \in \mathbb{R} \mid \omega > 0 \,\}.$$

4. **Esimerkki 1.1** (a) Heitetään lanttia. Tulosvaihtoehdot ovat klaava (L) ja kruunu (R), joten otosavaruus $\Omega = \{\mathrm{L}, \mathrm{R}\}$ ja $|\Omega| = 2$.

   (b) Heitetään lanttia, kunnes saadaan ensimmäinen klaava. Silloin otosavaruus

   $$\Omega = \{\mathrm{L}, \mathrm{RL}, \mathrm{RRL}, \mathrm{RRRL}, \ldots\}$$

   ja $|\Omega| = \infty$. Jos tapahtuma A on 'enintään kaksi kruunua ennen 1. klaavaa', niin $A = \{\mathrm{L}, \mathrm{RL}, \mathrm{RRL}\}$.

   $\square$

5. Kliininen koe 20 potilaalle, kokeillaan lääkettä A. Olkoot yksinkertaistetut tulosvaihtoehdot: potilas paranee (1), ei parane (0). Silloin

   $$\Omega = \{(i_1, \ldots, i_{20}) | i_j = 0 \text{ tai } 1\},$$

   joten $|\Omega| = 2^{20}$.

1.3.2 Joukko-operaatiot

http://www.math.uah.edu/stat/

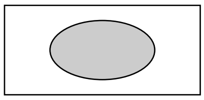
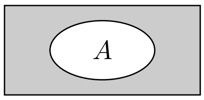
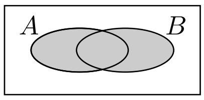
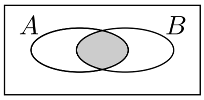
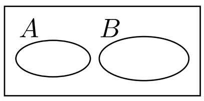
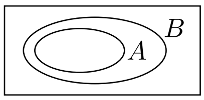
- Satunnaiskokeen $\mathcal{E}$ otosavaruus on Ω. Tapahtumat ovat Ω:n osajoukkoja.
- Jos A sattuu, niin kokeen $\mathcal{E}$ tulos $\omega \in A$.
- Vennin diagrammi
- Joukko-oppia esimerkiksi Diskreetin matematiikan kurssilla (Merikoski, Virtanen ja Koivisto, Diskreetti Matematiikka I, 3. luku). Verkossa esim. http://www.math.uah.edu/stat/
- Joukko-opin laskusäännöt, esimerkiksi *osittelulaki*

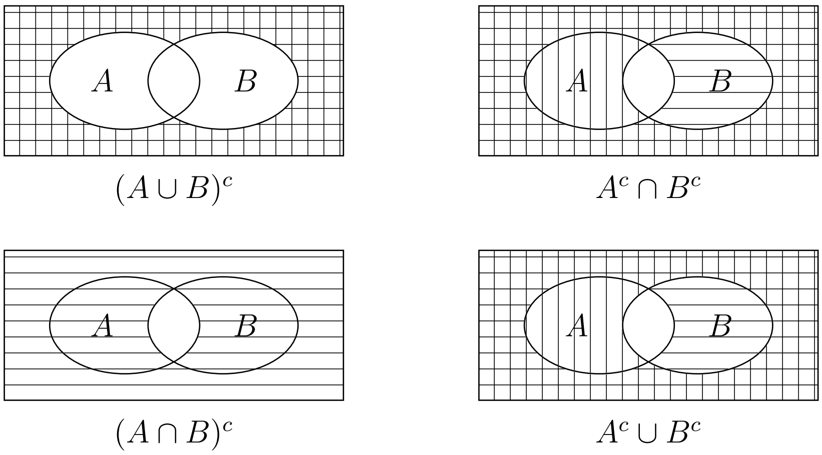
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

- Operaatiot *komplementti, yhdiste* ja *leikkaus* toteuttavat ns. *De Morganin lait:*

$$(A \cup B)^c = A^c \cap B^c,$$
$$(A \cap B)^c = A^c \cup B^c.$$

Taulukko 1.1. Joukko-opillisen ja todennäköisyyslaskennan terminologian vastaavuus.

Tapahtumat	Joukot	Joukkojen merkintä	Vennin diagrammi
otosavaruus	perusjoukko	Ω	
tapahtuma	Ω:n osajoukko	A, B, C jne.	
mahdoton tapahtuma	tyhjä joukko	$\emptyset$	
ei A, A ei satu	A:n komplementti	A^c	
joko A tai B tai molemmat	A:n ja B:n yhdiste	$A \cup B$	
sekä A että B	A:n ja B:n leikkaus	AB, $A \cap B$	
A ja B toisensa poissulkevat	A ja B pistevieraat	$A \cap B = \emptyset$	
jos A niin B	A on B:n osajoukko	$A \subset B$	



Kuvio 1.1. De Morganin lait.

- *Kaksinkertaisen komplementin sääntö*

$$(A^c)^c = A$$

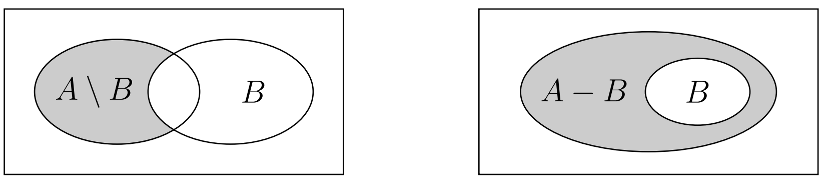
- Joukkojen A ja B *erotus* $A \setminus B$:

$$A \setminus B = A \cap B^c = \{\, \omega \mid \omega \in A \text{ ja } \omega \notin B \,\}.$$

- Jos $B \subset A$, merkinnän $A \setminus B$ sijasta voidaan käyttää myös merkintää $A - B$.

- Huomaa, että
$$A \setminus B = A - (A \cap B)$$
ja
$$A^c = \Omega - A.$$

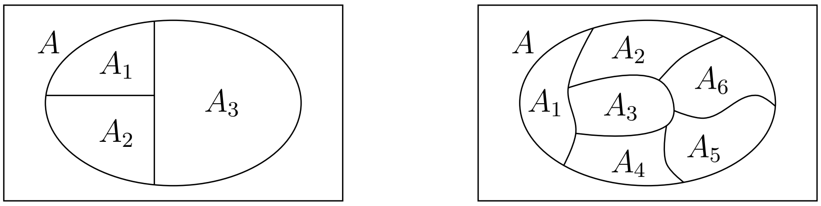
Ositus



Kuvio 1.2. Joukkojen erotus.

- Tapahtuman A *ositus* (tai jako) A_1, A_2, ..., A_m. Silloin $A = A_1 \cup A_2 \cup \cdots \cup A_m$ ja A_1, A_2, ..., A_m ovat toisensa poissulkevat.

- Esimerkiksi A, A^c on otosavaruuden Ω ositus ja $A \setminus B, A \cap B$ on A:n ositus.



Kuvio 1.3. Joukon A osituksia.

- Jos $A \cap B = \emptyset$, niin $A \cup B$:n sijasta voidaan kirjoittaa $A + B$. Silloin jo merkinnästä voidaan päätellä, että joukot ovat pistevieraat.

- Esimerkiksi
$$\Omega = A + A^c$$
ja
$$A = A_1 + A_2 + A_3,$$
joas A_1, A_2, A_3 on A:n jako.

1.3.3 Todennäköisyys

Kun otosavaruus on *äärellinen*, todennäköisyys voidaan määritellä alkeistapahtumien avulla.

Määritelmä 1.1 Olkoon $\mathcal{E}$ satunnaiskoe ja Ω sen äärellinen otosavaruus. Todennäköisyys on otosavaruudessa Ω määritelty reaaliarvoinen kuvaus

$$P\colon \Omega \to [0, 1],$$

jolla on seuraavat ominaisuudet:

1. $P(\omega) \geq 0$ kaikilla $\omega \in \Omega$, ja
2. $\sum\limits_{\omega \in \Omega} P(\omega) = 1$.

- $P(\omega)$ on *alkeistapahtuman* ω todennäköisyys.

- Tapahtuman $A \subset \Omega$ todennäköisyys on
$$P(A) = \sum_{\omega \in A} P(\omega).$$

- Näin jokaiseen tapahtumaan $A \subset \Omega$ voidaan liittää todennäköisyys $0 \leq P(A) \leq 1$.

- Todennäköisyys on joukkofunktio. Alkeistapahtuman todennäköisyyttä pitäisi oikeastaan merkitä $P(\{\omega\})$.

- Todennäköisyyttä kutsutaan yleisessä teoriassa todennäköisyysmitaksi.

- Jos $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, niin

$$\sum_{\omega_i \in \Omega} P(\omega_i) = \sum_{i=1}^{n} P(\omega_i) = 1.$$

- Jos esimerkiksi $A = \{\omega_1, \omega_3, \omega_5\}$, niin $P(A) = P(\omega_1) + P(\omega_3) + P(\omega_5)$.

- Satunnaiskokeen *todennäköisyysmalli* on pari (Ω, P). Se määritellään antamalla kokeen otosavaruus Ω ja siihen liittyvä funktio P, joka toteuttaa Määritelmän 1.1 ehdot.

- Jokainen alkeistapaus ω kuuluu joukkoon A tai sen komplementtiin, mutta ei molempiin samanaikaisesti. Siitä seuraa

$$\sum_{\omega \in A} P(\omega) + \sum_{\omega \in A^c} P(\omega) = \sum_{\omega \in \Omega} P(\omega) = 1.$$

- On siis voimassa sääntö $P(A) + P(A^c) = 1$. Sen kanssa yhtäpitävästi

$$P(A^c) = 1 - P(A).$$

- Koska $\Omega \cup \emptyset = \Omega$ ja $\Omega \cap \emptyset = \emptyset$, niin $\Omega^c = \emptyset$. Edellisen kohdan mukaan $P(\emptyset) = 1 - P(\Omega) = 0$.

- *Mahdoton tapahtuma* on tyhjä joukko $\emptyset$. Sen todennäköisyys on nolla, eli $P(\emptyset) = 0$.

- Määritelmän 1.1 mukainen funktio määrittelee *todennäköisyysjakauman* Ω:ssa. Jos $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, niin todennäköisyysjakauma on

$$\begin{array}{cccc} \omega_1 & \omega_2 & \ldots & \omega_n \\ p_1 & p_2 & \cdots & p_n, \end{array}$$

missä $p_i = P(\omega_i)$ ja $\sum_{i=1}^{n} p_i = 1$.

- **Esimerkki 1.2** Tavallisen nopan heitossa silmälukujen muodostama otosavaruus on $\Omega = \{1, 2, 3, 4, 5, 6\}$. Jos jokainen silmäluku on yhtä mahdollinen, todennäköisyysfunktio P on

$$P(i) = \frac{1}{6}, \qquad i = 1, \ldots, 6.$$

  - Nopanheiton todennäköisyysmalli on pari (Ω, P).

  - Tapahtuman 'silmäluku pariton' todennäköisyys on

$$P(\{1, 3, 5\}) = P(1) + P(3) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}.$$

$\square$

1.3.4 Äärettömät otosavaruudet

- Edellä on käsitelty vain *äärellisiä otosavaruuksia.*
- Esimerkissä 1.1 esitettiin myös ääretön otosavaruus.
- Jos Ω on numeroituvasti ääretön, niin
$$\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}.$$
- Silloin Määritelmän 1.1 2. ehdossa äärellinen summa korvataan äärettömällä summalla
$$\sum_{i=1}^{\infty} p_i = p_1 + p_2 + p_3 + \cdots = 1,$$
missä $P(\omega_i) = p_i$.
- Jos Ω ei ole numeroituva (eli on ylinumeroituva), niin Määritelmä 1.1 ei sovellu tapahtumien todennäköisyyden määrittelemiseen.

1.3.5 Todennäköisyyden tulkinnat

- Todennäköisyyslaskenta on aksiomaattinen matemaattinen teoria.
- Tapahtuman A *mahdollisuus* (*odds*) määritellään suhteena
$$\text{odds}(A) = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)}. \tag{1.3.1}$$
- Jos $\text{odds}(A)$ on annettu, niin A:n todennäköisyys on
$$P(A) = \frac{\text{odds}(A)}{1 + \text{odds}(A)}.$$
- **Esimerkki 1.3** Olkoon 1000 henkilön populaatiossa on 600 naista ja 400 miestä. Naisten suhteellinen osuus on
$$\frac{600}{600 + 400} = 0.6.$$
Jos $A = \{\text{nainen}\}$ ja $B = \{\text{mies}\}$, niin naisen mahdollisuus tulla valituksi on
$$\text{odds}(A) = \frac{P(A)}{1 - P(A)} = \frac{0.6}{0.4} = \frac{3}{2}.$$
□
- Uhkapelurit ovat kiinnostuneita *voiton mahdollisuudesta* (*payoff odds*). Pelikasinot ja vedonvälittäjät tarjoavat näitä mahdollisuuksia.

- Kasino tarjoaa esimerkiksi seuraavaa vetoa: Maksat osallistumisesta 1:n euron ja voitat 10 euroa, jos tapahtuma A sattuu. Jos A ei satu, häviät sen yhden euron.

- Voiton mahdollisuus on

$$\frac{\text{oma panos}}{\text{kasinon panos}} = \frac{1}{10}.$$

- Oma panos = 1, kasinon panos = 10 ja kokonasipanos = 11. A:n sattuessa saat kokonasipanoksen 11, joka sisältää maksamasi pelimaksun 1.

- Reilun pelin sääntö:

$$\text{voiton mahdollisuus } (VM) = \text{odds}(A).$$

- Jos veto (peli) toteuttaa reilun pelin säännön, niin pääset pitkässä sarjassa omillesi. Jos $VM < \text{odds}(A)$, häviät pitkässä sarjassa. Kasinoiden ja vedonvälittäjien toiminnan tuotto perustuu epäyhtälöön $VM < \text{odds}(A)$.

- *Mahdollisuuksien suhde* (*odds ratio*) $\theta(A, B)$ on

$$\theta(A, B) = \frac{\text{odds}(A)}{\text{odds}(B)} = \frac{P(A)/[1 - P(A)]}{P(B)/[1 - P(B)]}, \tag{1.3.2}$$

  vedonlyöntiterminologian mukaan θ on *vedonlyöntisuhde*.

- Vedonlyönnissä ja monissa sovelluksissa todennäköisyyksien arviointi perustuu usein henkilökohtaisiin käsityksiin ja kokemuksiin.

1.4 Ehdollinen todennäköisyys

- **Esimerkki 1.4** Heitetään harhatonta noppaa kuten Esimerkissä 1.2. Saadaan pariton silmäluku.

  - Mikä on todennäköisyys, että silmäluku on 5?
  - Olkoon $B = \{1, 3, 5\}$ ja $A = \{5\}$. B:n alkeistapaukset $1, 3$ ja 5 ovat yhtä todennäköisiä, joten silmäluvun 5 todennäköisyys B:ssä on 1/3.
  - Tapahtuman A *ehdollinen todennäköisyys* ehdolla B on 1/3:

$$P(A \mid B) = 1/3.$$

  - Tässä esimerkissä $P(A \mid B) \neq P(A) = 1/6$.

□

- Ehdollisen todennäköisyyden $P(A \mid B)$ kaava

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{1.4.1}$$

Todennäköisyys $P(A \mid B)$ on määritelty, kun $P(B) > 0$.

Esimerkki 1.5 Eloonjäämistaulukoissa esitetään eri ikäisenä elossa olevien odotettu lukumäärä 100000 elävänä syntynyttä kohti. Esimerkiksi seuraavassa taulukossa on annettu 20-, 45- ja 65-vuotiaana elossa olevien naisten lukumäärät eräässä väestössä 100000 elävänä syntynyttä tyttölasta kohti.

Ikä	20	45	65
Elossa	98040	95662	84483

Tässä voidaan ajatella, että alkuperäinen otosavaruus Ω on 100000 tyttölasta. Mikä on todennäköisyys, että 20-vuotias elää 45-vuotiaaksi (tarkoittaa itse asiassa, että elää ainakin 45-vuotiaaksi)? Olkoon $A =$ 'elää 45-vuotiaaksi' ja $B =$ 'elää 20-vuotiaaksi'. Koska 20-vuotiaaksi on elänyt 98040 naista ja näistä 45-vuotiaaksi 95662, niin kysytty todennäköisyys on $95662/98040 = 0.97574$. Laskettaessa ehdollista todennäköisyyttä valitaan perusjoukoksi B ja katsotaan kuinka moni näistä selviää 45-vuotiaaksi.

Nyt tapahtuma $A \cap B$ on 'elää 45-vuotiaaksi', koska 45-vuotiaksi eläneet ovat eläneet myös 20-vuotiaksi. Koska 20-vuotiaaksi elää 98040, niin $P(B) = 98040/100000 = 0.98040$. Vastaavasti $P(A \cap B) = 95662/100000 = 0.95662$. Ehdollinen todennäköisyys

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.95662}{0.98040} = 0.97574.$$

□

1.4.1 Ehdollisen todennäköisyyden frekvenssitulkinta

Ffrekvenssitulkinnan mukaan

$$P(A \mid B) \approx \frac{N_n(A \cap B)}{N_n(B)} = \frac{N_n(A \cap B)/n}{N_n(B)/n} \approx \frac{P(A \cap B)}{P(B)}, \tag{1.4.2}$$

kun toistojen lukumäärä n on suuri.

1.4.2 Kertolaskusääntö

Koska ehdollisen todennäköisyyden kaavassa (1.4.1) $P(B) > 0$, saadaan siitä kertolaskusääntö

$$P(A \cap B) = P(B)\, P(A \mid B) \tag{1.4.3}$$

tapahtuman $A \cap B$ todennäköisyyden laskemiseksi.

1.4.3 Riippumattomuus

Sanomme, että tapahtumat A ja B ovat *riippumattomat,* jos

$$P(A \cap B) = P(A)\, P(B). \tag{1.4.4}$$

Huomaa, että ehdollinen todennäköisyys (1.4.1) ei ole määritelty, jos $P(B) = 0$, mutta riippumattomuuden määritelmä (1.4.4) on silloinkin voimassa. Jos $P(B) \neq 0$ ja (1.4.4) pitää paikkansa, niin

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = P(A).$$

Jos A ja B ovat riippumattomat, niin tieto B:n sattumisesta ei vaikuta A:n todennäköisyyteen. Jos $P(A) > 0$, niin myös $P(B \mid A) = P(A \cap B)/P(A) = P(B)$, kun A ja B ovat riippumattomat.

Esimerkki 1.6 Tarkastellaan kaksilapsisia perheitä.

1. Valitaan satunnaisesti yksi perhe. Havaittiin, että perheessä on poika. Millä todennäköisyydellä hänellä on veli?

2. Valitaan kaksilapsisten perheiden lapsista satunnaisesti yksi. Jos valittlapsi on poika, millä todennäköisyydellä hänellä on veli?

□

1.5 Odotetut frekvenssit

- Kokeen $\mathcal{E}$ todennäköisyysmalli (Ω, P) on teoreettinen konstruktio. Mallin käytännössä tutkittava empiirisesti.

- Verrataan kokeen (empiirisen ilmiön) havaittuja tuloksia mallin perusteella odotettavissa oleviin tuloksiin.

- Koe toistetaan n kertaa. Jos tapahtuman A todennäköisyys on mallin mukaan p, niin silloin A:n *odotettu frekvenssi* eli *teoreettinen frekvenssi* on np.

- A sattui suoritetussa toistokokeessa n_A kertaa, joka on *havaittu frekvenssi.*

- Jos n_A poikkeaa ”liian paljon” odotetusta frekvenssistä np, niin havainnot eivät tue mallia (teoriaa).