Tilastotieteen johdantokurssi

Apulaisprofessori Ville Hautamäki

School of Computing
University of Eastern Finland

September 18, 2023

Todennäköisyyslaskentaa (1)

- Tapahtumia joissa sattuma vaikuttaa yksittäiseen tulokseen kutsutaan satunnaisilmiöiksi tai satunnaiskokeiksi.
- Satunnaisilmiön lopputulosta ei tiedetä etukäteen, mutta jos ilmiö toistuu useita kertoja, tuloksista muodostuu säännönmukainen jakauma.
- Satunnaisilmiöitä ja satunnaiskokeita voidaan mallintaa todennäköisyyslaskennan keinoin, jolloin saadun mallin avulla voidaan tehostaa päätöksentekoa epävarmuutta sisältävissä tilanteissa.

Todennäköisyyslaskentaa (2)

- Tilastotieteessä usein määritellään todennäköisyysmalli tutkittavalle ilmiölle.
- Tämä malli sisältää matemaattisessa muodossa sen mitä on perusteltua olettaa ilmiön ominaisuuksista, mutta numeerisia arvoja prosessia kuvaaville parametreille ei yleensä tunneta.
- Niiden arvot **estimoidaan** käyttämällä havaittuja arvoja satunnaisilmiön tuloksesta.

Klassinen todennäköisyyden määritelmä

- Klassinen todennäköisyys määritellmään suotuisten tapausten lukumäärän suhde kaikkien mahdollisten tapausten lukumäärään.

$$\text{Tapahtuman todennäköisyys} = \frac{\text{Suotuisten tapahtumien lkm}}{\text{Kaikkien tapahtumien lkm}} \tag{1}$$

- Klassisessa todennäköisyyden tulkinnassa oletetaan, että kaikki tapaukset ovat yhtä todennäköisiä. Esimerkiksi voidaan sanoa, että yhtä noppaa heitettäessä saadaan silmäluku 2 todennäköisyydellä $\frac{1}{6}$ (noppa oletetaan harhattomaksi eli jokaisen silmäluvun todennäköisyys on sama).
- Klassista eli frekventistä todennäköisyyden määritelmää voi myös kritisoida siitä että monesti toistaminen ei ole mielekästä (esim: "mikä on todennäköisyys sille että opettaja saapuu tiistain luennolle?"

Subjektiivinen todennäköisyys

- Toinen yleisesti käytetty tulkinta todennäköisyydelle on **subjektiivinen todennäköisyys**, joka kuvaa uskomusta tapauksen todennäköisyydestä, mutta ei välttämättä sisällä mittalukua todennäköisyydestä.
- Esimerkiksi väite "Todennäköisesti huomenna sataa" kuvaa väitteen esittäjän subjektiivista uskomusta sateen mahdollisuudesta.
- Subjektiivinen todennäköisyys -pohjainen tilastotiede tunnetaan nimellä Bayesilainen tilastotiede. Siinä todennäköisyys on aina suhteessa (priori) alkuperäiseen uskomukseen.
- Kun data on havaittu, voidaan uskomus päivittää vastaamaan todellisuutta.
- Historiallisesti, Bayesiläinen todennäköisyyspäättely kehitettiin ensiksi (1800-luvulla) ja frekventistinen 1900-luvun alussa.
- Andrei Kolmogorov systematisoi ja formalisoi todennäköisyyslaskennan 1900-luvun alkupuolella.

Tilastollinen todennäköisyyden tulkinta

- Tilastollisessa todennäköisyyden tulkinnassa tapahtuman todennäköisyys määritellään tapahtuman suhteelliseksi esiintymisfrekvenssiksi pitkässä koesarjassa, jossa tapahtumat ovat riippumattomia. Esimerkiksi jos seurataan suurta määrää mielivaltaisesti valittuja syntymiä voidaan laskea tilastollinen todennäköisyys sille, että syntyvä lapsi on poika.
- **Esimerkki 5.1**: 5.1 Aikavälillä 2000-2015 Suomessa syntyi kaikkiaan 929420 lasta. Syntyneistä lapsista 475550 oli poikia ja 453870 oli tyttöjä. Tämän perusteella voidaan tehdä arvio, että syntyvä lapsi on poika todennäköisyydellä

$$\frac{475550}{929420} = 0.512 \tag{2}$$

Todennäköisyysmalli

- Todennäköisyyslaskenta on satunnaisilmiöiden käsittelyä matematiikan keinoin. Kun tarkastellaan satunnaiskoetta (satunnaisilmiö), niin tulosvaihtoehdot tiedetään, mutta sattuman vuoksi ei voida varmuudella sanoa/ennustaa, mikä satunnaiskokeen tulos lopulta on.
- Esimerkiksi, jos noppaa heitetään yhden kerran, emme tiedä silmälukua etukäteen. Tiedämme vain, että heitosta saatava silmäluku on jokin arvo joukosta $\{1, 2, 3, 4, 5, 6\}$
- Voimme myös olettaa nopan olevan tasapainoinen (**harhaton**), jolloin kaikki silmäluvut ovat yhtä todennäköisiä.

Todennäköisyysmallinnuksen perusperiaatteet

Edellä oleva kuvaus nopanheitosta sisältää kaksi osaa:

- Listan kaikista kokeen tulosvaihtoehdoista.
- Kunkin tulosvaihtoehdon todennäköisyydet

Nämä kaksi tekijää muodostavat perustan satunnaisilmiötä kuvaavalle todennäköisyysmallille.

Otosavaruus ja alkeistapaus

- Satunnaiskokeen/satunnaisilmiön kaikkien tulosvaihtoehtojen muodostamaa joukkoa kutsutaan **otosavaruudeksi**. Esimerkiksi heitettäessä kolikkoa kerran tiedetään tuloksen olevan kruuna tai klaava eli otosavaruus on

$$S = \{\text{kruuna}, \text{klaava}\} = \{\text{kr}, \text{kl}\} \tag{3}$$

- Edellä otosavaruutta merkittiin symbolilla S, mutta otosavaruutta merkitään usein myös kreikkalaisella kirjaimella Ω
- Satunnaisilmiön tuottamaa tulosta kutsutaan myös **alkeistapaukseksi**. Esimerkiksi kolikonheitossa on kaksi alkeistapausta “kruuna” (kl) ja "klaava" (lk). Alkeistapaukset ovat otosavaruuden alkioita.

Esimerkki 5.2

Heitettäessä kolikkoa kahdesti voidaan saada neljä erilaista tulosta:

1. heitto	2. heitto
kruuna	kruuna
kruuna	klaava
klaava	kruuna
klaava	klaava

- Otosavaruus kahden kolikon heitossa on siis

$$S = \{\text{krkr}, \text{krkl}, \text{klkr}, \text{klkl}\} \tag{4}$$

  Tässä otosavaruudessa on siis neljä alkeistapausta.

- Jos kiinnostuksen kohteena olisikin kruunien lukumäärä kahden kolikon heitossa, niin otosavaruus olisi

$$S = \{0, 1, 2\} \tag{5}$$

Satunnaisotanta ja tapahtuma

Satunnaisotannassa jokainen otos on tulos, joten satunnaisotannan otosavaruus muodostuu kaikista mahdollisista otoksista.

- **Tapahtumaksi** kutsutaan satunnaiskokeen tulosta tai tulosten muodostamaa joukkoa. Toisin sanoen tapahtuma on otosavaruuden S osajoukko. Tapahtumia merkitään isoilla kirjaimilla $\{A, B, C, \ldots\}$
- Voimme sanoa esimerkiksi, että tapahtuma A sattuu (toteutuu), jos satunnaiskokeen tulos kuuluu joukkoon A.

Esimerkki 5.3

- Heitetään kolikkoa kunnes saadaan ensimmäinen kruuna tai kolikkoa on heitetty neljä kertaa. Tässä satunnaiskokeessa otosavaruus on

$$S = \{\text{kr}, \text{klkr}, \text{klklkr}, \text{klklklkr}, \text{klklklkl}\} \tag{6}$$

- Jos tapahtuma A on "Saadaan enintään kaksi klaavaa ennen kruunaa", niin tapahtuma A on joukko $A = \{\text{kr}, \text{klkr}, \text{klklkr}\}$

Merkintä, tapahtuman A todennäköisyyttä merkitään $P(A)$.

Esimerkki 5.4

Edellä todettiin, että yhtä noppaa heitettäessä saadaan silmäluku 2 todennäköisyydellä $\frac{1}{6}$. Harhattoman nopan tapauksessa todennäköisyys voidaan laskea suotuisten tapausten ja kaikkien tapausten lukumäärien suhteena $\Rightarrow$ otosavaruus sisältää nopanheitossa kuusi mahdollista alkeistapausta

$$S = \{1, 2, 3, 4, 5, 6\}, \tag{7}$$

joista yksi eli silmäluku 2 on suotuisa. Jos merkitään tapahtumaa A =“nopan silmäluku on 2”, niin tämän tapahtuman todennäköisyys on $P(A) = \frac{1}{6}$.

Todennäköisyysmallista

Todennäköisyysmalliksi kutsutaan satunnaisilmiön matemaattista mallia, jonka määrittelee otosavaruus sekä otosavaruuden tuloksiin liittyvät todennäköisyydet. Todennäköisyysmallissa todennäköisyys liittyy siis tapahtumiin.

A. Kolmogorovin todennäköisyyden perusominaisuudet

1. $0 \leq P(A) \leq 1$, Tapahtuman A todennäköisyys on vähintään 0 (mahdoton tapahtuma) ja enintään 1 (varma tapahtuma).
2. $P(A^C) = 1 - P(A)$, tapahtuman A **vastatapahtuman todennäköisyys**
3. $P(S) = 1$, otosavaruuden eli varman tapahtuman todennäköisyys on 1.
4. jos tapahtumat A ja B ovat **erilliset**, niin

$$P(A \text{ tai } B) = P(A \cup B) = P(A) + P(B) \tag{8}$$

   Tämä on **yhteenlaskusääntö**. Yleisessä tapauksessa pätee

$$P(A \text{ tai } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{9}$$

5. $P(A \text{ ja } B) = P(A \cap B) = P(A)P(B)$ kun tapahtumat ovat riippumattomia. Eli tieto toisesta tapahtumasta ei kerro mitään toisesta. Todennäköisyys kummallekin tapahtumalle on siten tulo. Tätä kutsutaan riippumattomien tapahtumien **tulosäännöksi**.

Esimerkki 5.5, veriryhmien todennäköisyyksistä

Satunnisesti valiten henkilön veriryhmä noudattelee oheista taulukkoa:

Veriryhmä	O	A	B	AB
Todennäköisyys	0.49	0.27	0.20	?

- Mikä on todennäköisyys, että henkilön veriryhmä on A tai B? Koska henkilön veriryhmä voi olla vain yhtä ryhmää, tapahtumat "veriryhmä on A" ja "veriryhmä on B" ovat erillisiä. Satunnaisesti valitun henkilön veriryhmä on A tai B todennäköisyydellä

$$\begin{aligned} P(\text{"Veriryhmä on A tai B"}) &= P(\text{"Veriryhmä on A"}) \\ &+ P(\text{"Veriryhmä on B"}) \\ &= 0.27 + 0.20 = 0.47 \quad (10) \end{aligned}$$

Esimerkki 5.5 (jatkuu)

- Mikä on veriryhmän AB todennäköisyys?
  Erillisten tapausten yhteenlaskusäännöllä saadaan:

$$\begin{aligned} P(\text{"Veriryhmä on A tai B tai O"}) &= P(\text{"O"}) + P(\text{"A"}) + P(\text{"B"}) \\ &= 0.49 + 0.27 + 0.20 = 0.96 \end{aligned}$$

  joten tapahtuman "veriryhmä on O tai A tai B"
  vastatapahtuman todennäköisyyden avulla saadaan

$$P(\text{"Veriryhmä on AB"}) = 1 - 0.96 = 0.04 \qquad (11)$$

  Voitaisiin ajatella myös näin: Todennäköisyyden perusominaisuuden 3 nojalla kaikkien otosavaruuteen kuuluvien alkeistapausten todennäköisyyksien summa on aina 1

Esimerkki 5.6

Valitaan 1-numeroinen satunnaisluku kokonaisluvuista 0-9 niin että kaikki luvut ovat yhtä todennäkoisiä. Todennäköisyyksiksi saadaan

Satunnaisluku	0	1	2	...	8	9
Todennäköisyys	0.1	0.1	0.1	...	0.1	0.1

Todennäköisyydet 0.1 saadaan ehdoista, että joku luku valitaan, ja että jokainen numero on yhtä todennäköinen ja lukujen todennäköisyyksien summa on 1.

- Millä todennäköisyydellä 1-numeroinen satunnaisluku on pariton?

$$\begin{aligned} P(\text{"Luku on pariton"}) &= P(\{1,3,5,7,9\}) \\ &= P(\{1\}) + P(\{3\}) + \ldots + P(\{9\}) \\ &= 0.1 + 0.1 + 0.1 + 0.1 + 0.1 = 0.5 \end{aligned}$$

Esimerkki 5.6 (jatkuu)

- Millä todennäköisyydellä luku on pariton tai pienempi tai yhtä suuri kuin 3?
  Määritellään tapahtumat A ja B seuraavasti:
  $A =$ "luku on pariton" ja $B =$ "luku ≤ 3" Nyt todennäköisyys, että luku on pariton on $P(A) = 0.5$ ja todennäköisyys, että luku on pienempi tai yhtä suuri kuin 3 on $P(B) = 0.4$

$$\begin{aligned} P(A \cup B) &= P(\text{"luku on pariton tai luku on} \leq 3\text{"}) \\ &= P(\{1,3,5,7,9\} \cup \{0,1,3\}) \\ &= P(\{0,1,2,3,5,7,9\}) = 0.7 \end{aligned} \tag{12}$$

  Huom! tämä on eri kuin $P(A) + P(B) = 0.9$. Meidän siis tulee käyttää yhteenlaskusääntöä:

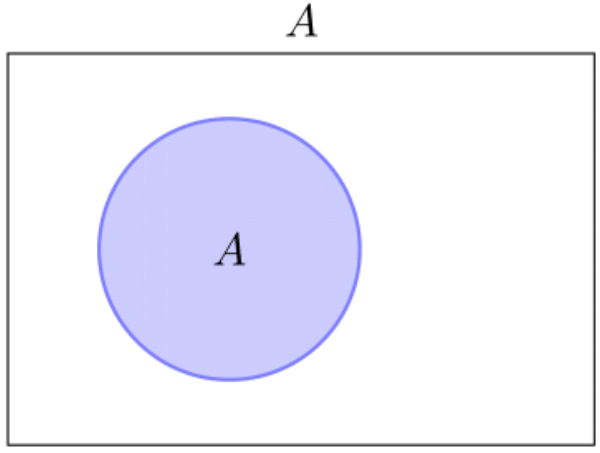
$$P(A \,\text{tai}\, B) = P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{13}$$

  eli pitää vähentää summasta $P(A \cap B) = 0.2$, koska tapahtumissa A ja B on **kaksi** yhteistä alkeistapausta, luvut 1 ja 3.

Visualisointi Venn-diagrammeilla

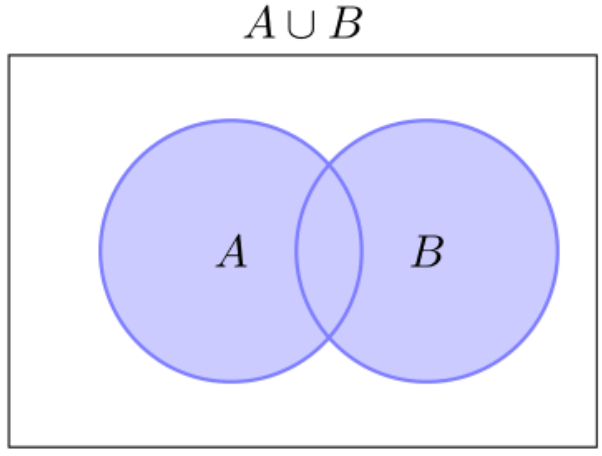
Ensiksi perusjoukko S

Ja sitten tapahtuma A, joukko-opin merkinnöillä $A \subseteq S$

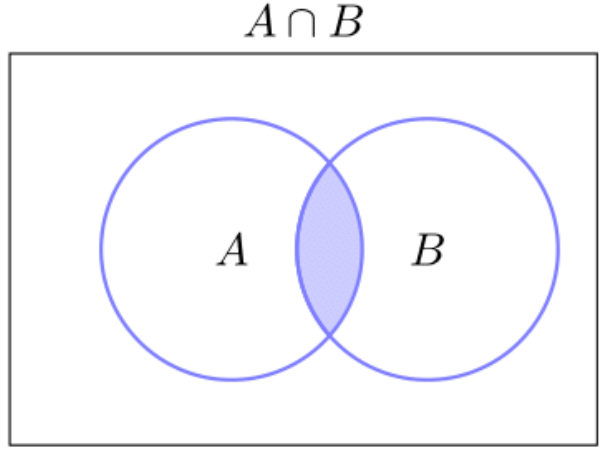


Visualisointi jatkuu

"A tai B" eli joko A tai B tai molemmat. Joukko-oppi: $A \cup B$

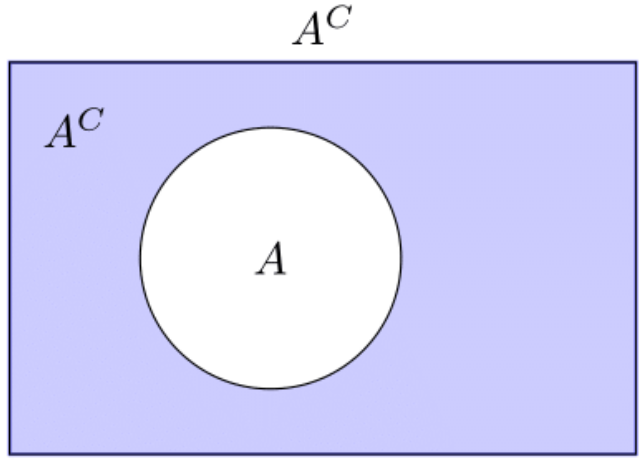


"A ja B" eli A ja B tapahtuu yhtäaikaa. Joukko-oppi: $A \cap B$

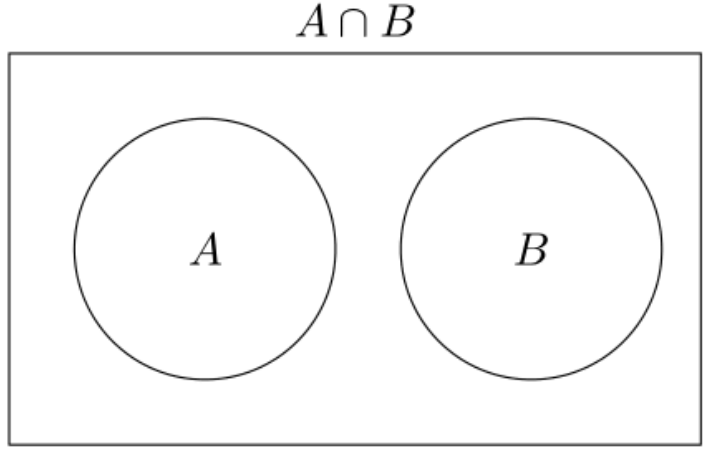


Visualisointi jatkuu

A ei esiinny. Joukko-oppi: A^C. (A:n komplementti)



A ja B ovat erillisiä. Joukko-oppi: $A \cap B = \emptyset$



Esimerkki 5.8

- Heitettäessä kolikkoa kaksi kertaa otosavaruus on $S = \{\text{krkr}, \text{krkl}, \text{klkr}, \text{klkl}\}$. Kaksi tapahtumaa, jotka ovat $A =$ "1. heitto on kruuna" ja $B =$ "2. heitto on kruuna". Tapahtumat A ja B eivät ole erillisiä, koska molemmat sattuvat jos molempien heittojen tulos on kruuna.
- Heitot oletetaan riippumattomiksi eli edellisen heiton tulos ei vaikuta seuraavan heiton tulokseen. Siksi voidaan käyttää kertolaskusääntöä laskettaessa todennäköisyys sille, että molemmilla heitoilla saadaan kruuna:

$$\begin{aligned} P(A \text{ ja } B) &= P(A \cap B) \\ &= P(A)P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \end{aligned} \qquad (14)$$

Esimerkki 5.9

Heitään vain yhtä noppaa. Tapahtumat ovat:
$A =$ "Silmäluku on vähintään 3" $= \{3, 4, 5, 6\}$
$B =$ "Silmäluku on 3" $= \{3\}$
Tapahtuma "A ja B" $= A \cap B = \{3, 4, 5, 6\} \cap \{3\} = \{3\} = B$, eli tapahtumat eivät ole erillisiä.
Jotta tapahtumat A ja B olisivat erillisiä, niillä ei saisi olla yhtään yhteistä alkeistapausta. Eli $A \cap B$ olisi silloin tyhjä joukko.

Ehdollinen todennäköisyys ja riippumattomuus

- Kun tapahtuma B on mahdollinen, eli $P(B) > 0$, niin **ehdollinen todennäköisyys** tapahtumalle A ehdolla B saadaan

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{15}$$

  Ehdollinen todennäköisyys $P(A|B)$ kertoo tapahtuman A todennäköisyyden, kun tiedetään, että tapahtuma B on tapahtunut.
- Ehdollinen todennäköisyys toimii kuin datasetin **filtteröinti**, esimerkiksi: todennäköisyys että saat nuhan on pieni, mutta jos ulkona on -10C pakkasta, niin todennäköisyys on suurempi.

Esimerkki: vaalit Yhdysvalloissa

Yhdysvaltojen 2012 vaalissa annettiin yhteensä 132 948 000 vaalilippua, joista 18-19-v. oli 20 539 000 ja 30-45-v. 30 756 000. Lasketaan ensiksi todennäköisyys, että äänestäjän ikä on alle 45:

$$P(\text{ikä} < 45) = \frac{20\ 539\ 000 + 30\ 756\ 000}{132\ 448\ 000} = \frac{51\ 295\ 000}{132\ 948\ 000} = 0.38,$$

Todennäköisyys on siis 0.38. Jos tiedämme että satunnaisen äänestäjän ikä tulee olla yli 29. Filtteröimme siis kaikkien äänestäjien joukosta kaikki alle 29 vuotiaat pois ja laskemme ehdollisen todennäköisyyden:

$$P(\text{ikä} < 45 | \text{ikä} > 29) = \frac{30\ 756\ 000}{112\ 409\ 000} = 0.27.$$

On tärkeää huomata että äänten kokonaislukumäärä tippuu 132 448 000:sta 112 409 000:han. Tämä on se filtteröinnin vaikutus.

Riippumattomuus

Tapahtumien riippumattomuus ehdollisen todennäköisyyden näkökulmasta. Kun tapahtumat A ja B ovat riippumattomia ja $P(B) > 0$ niin,

$$P(A|B) = P(A) \tag{16}$$

Edellinen tarkoittaa sitä, ettei tieto tapahtuman B tapahtumisesta vaikuta tapahtuman A todennäköisyyteen.

Esimerkki 5.10

Heitetään noppaa kaksi kertaa. Olkoon tapaukset sitten:
$A =$ "Saadaan ainakin kerran silmäluku 2" ja
$B =$ "Silmälukujen summa on pienempi kuin 6" Taulukoidaan kaikki heittojen mahdolliset tulokset:

	Ensimmäinen heitto					
Toinen heitto	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Todennäköisyydet voi ajatella suotuisten tapausten lukumäärien ja kaikkien tulosvaihtoehtojen lukumäärän (36 kpl) avulla, koska noppia heitettäessä jokainen silmäluku on yhtä todennäköinen (klassinen todennäköisyyden määritelmä).

Esimerkki 5.10 jatkuu

- Laske todennäköisyydet $P(A)$ ja $P(B)$
  Tapahtumalle A suotuisia tulosvaihtoehtoja on 11 kpl ja tapahtumalle B suotuisia tulosvaihtoehtoja on 10 kpl. Siten

$$P(A) = \frac{11}{36} \approx 0.305 \tag{17}$$

$$P(B) = \frac{10}{36} \approx 0.278 \tag{18}$$

- Laske todennäköisyys $P(A \cap B)$
  Molemmille tapahtumille suotuisia tulosvaihtoehtoja on 5 kpl, joten

$$P(A \cap B) = \frac{5}{36} \approx 0.134 \tag{19}$$

Esimerkki 5.10 jatkuu

- Laske ehdollinen todennäköisyys $P(A|B)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{5}{36}}{\frac{10}{36}} = \frac{1}{2} \tag{20}$$

Esimerkki 5.11

Kolikon heitto kolme kertaa. Tarkastellaan tapahtumaa
$A =$ "Saadaan kaikilla kolmella heitolla klaava"
Otosavaruudessa on 8 eri tulosmahdollisuutta (luettele ne!), joista jokainen on yhtä todennäköinen. Siten

$$P(A) = \frac{1}{8} \tag{21}$$

Toisaalta voidaan laskea myös suoraan hyödyntäen riippumattomien tapahtumien kertolaskusääntöä

$$P(A) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} \tag{22}$$

Lyhyesti kombinatoriikasta

Edellä todennäköisyydet voi ajatella siten, että todennäköisyys on tapahtumalle suotuisten tulosvaihtoehtojen lukumäärän suhde satunnaiskokeen kaikkien tulosvaihtoehtojen lukumäärään.
Erilaisten tulosten ja tapahtumalle suotuisten tulosvaihtoehtojen lukumäärän laskeminen voi kuitenkin olla hankalaa.
Kombinatoriikka käsittelee näiden lukumäärien laskemista.

Erilaisten lottorivien lukumäärä

- Kombinatoriikan avulla voidaan ratkaista esimerkiksi erilaisten lottorivien lukumäärä. Lotossa arvotaan 40:stä numerosta 7 numeroa sisältävä rivi (ei huomioida lisänumeroita). Erilaisten lottorivien lukumäärä (erilaisten 7 numeron kombinaatioiden lkm.) on

$$\binom{40}{7} = \frac{40!}{7!(40-7)!} = \frac{40!}{7!33!} = 18\ 643\ 560 \qquad (23)$$

- Edellisessä kaavassa $\binom{40}{7}$ on binomikerroin ja luetaan "40 yli 7". Binomikertoimen

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad (24)$$

avulla voidaan laskea kuinka monella tavalla n alkiosta voidaan valita k alkiota, kun tietty alkio voidaan poimia vain kerran ja valintajärjestyksellä ei ole merkitystä.

Kertomasta

- Merkintä $n!$ tarkoittaa **kertomaa**, eli

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1 \tag{25}$$

  joka kertoo kuinka monella eri tavalla n alkiota voidaan asettaa jonoon.
- On myös hyvä huomata että on sovittu että $0! = 1$.
- Kertoma on määritelty vain kokonaisluvuille, mutta sama voidaan tehdä myös reaaliluvuille ja sen function nimi on **Gamma-funktio** $\Gamma(n)$. Mikä saa kokonaislukujen tapauksessa saman arvon kuin kertoma. Tilastotieteessä Gamma-funktio esiintyy mm. Beta -jakaumassa.

Esimerkki 5.13

Kolme henkilöä voidaan asettaa jonoon 6:lla ($3! = 3 \times 2 \times 1 = 6$), kun järjestyksellä on väliä. Kolmen henkilön joukosta voidaan poimia kaksi henkilöä $\binom{3}{2} = 3$ kun poimintajärjestyksellä ei ole väliä.

Esimerkki 5.13

Ohessa yksinkertaiset esimerkit siitä kuinka kertoma ja binomikerroin voidaan laskea R:llä

- Kertoma eli 3! on R:ssä `factorial(3)`
- kolme yli kahden $\binom{3}{2}$ on R:ssä `choose(3,2)`