Tilastollinen todennäköisyys

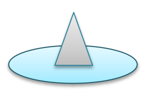
Klassisessa todennäköisyydessä oli ehdot: äärellisyys ja symmetrisyys. Tämä tilanne on usein mahdoton → ts. alkeistapauksia on usein ääretön määrä tai ne eivät ole symmetrisiä.

Määritelmä, *tilastollinen todennäköisyys:*
Kokeellisesti hankittua todennäköisyyden arvoa sanotaan *tilastolliseksi todennäköisyydeksi*. Eli tilastoaineistosta laskettua suhteellista frekvenssiä kutsutaan tilastolliseksi todennäköisyydeksi.

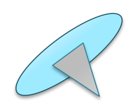
Määritelmä, *simulointi:*
Simuloinnilla tarkoitetaan käytännön tilanteen/tapahtuman jäljittelemistä esim. tietokonetta hyödyntäen.

Esimerkki: Nastaa heitetään kerran. Millä todennäköisyydellä se pudottuaan on kärki

a) ylöspäin



b) alaspäin



a) Todennäköisyys riippuu nastasta → heitetään nastaa riittävän monta kertaa ja tutkitaan mitä lukua alkeistapauksen ”kärki ylös” frekvenssi f_{ky} jaettuna kaikkien heittojen lukumäärällä n lähestyy. Kyseinen lukuhan on suhteellinen frekvenssi $\frac{f_{\text{ky}}}{n}$. Taulukointi antaa

$$\frac{f_{\text{ky}}}{n} \to 0{,}619 \quad \text{kun} \quad n \to \infty,$$

Eli

$$P(\text{"kärki ylös"}) = 0{,}619$$

n	f_{ky}	$\frac{f_{ky}}{n}$
200	126	0,630
400	244	0,610
600	375	0,625
800	485	0,606
1000	610	0,610
1200	743	0,619
1400	866	0,619

b) Alkeistapauksen ”kärki alas” frekvenssi f_{ka} on $1 - f_{\text{ky}}$, joten

$$\begin{aligned} P(\text{"kärki alas"}) &= 1 - P(\text{"kärki ylös"}) \\ &= 1 - 0{,}619 \\ &= 0{,}381 \end{aligned}$$

Määritelmä, *tapahtuman tilastollinen todennäköisyys:*
Tapahtuman A tilastollinen todennäköisyys on luku, jota suhteellinen frekvenssi lähestyy, kun koetta toistetaan yhä uudelleen, siis

$$P(A) \approx \frac{\text{tapahtuman } A \text{ esiintymiskertojen lukumäärä}}{\text{kokeen toistojen lukumäärä}}$$

Huomaa, että tämä ei ole tarkka todennäköisyyden arvo, vaan tilastollinen arvio tapahtuman todennäköisyydelle.

Jos tapahtuma A koostuu äärellisestä määrästä alkeistapauksia, niin sen todennäköisyys on suotuisien alkeistapauksien todennäköisyyksien summa.

Esimerkki: Suomalaisia kuuluu veriryhmiin seuraavasti: 44% A, 17% B, 8% AB ja 31% 0. Mikä on todennäköisyys, että satunnaisesti valittu suomalainen kuuluu veriryhmiin A tai B?
Nyt perusjoukko $E = \{\mathrm{A}, \mathrm{B}, \mathrm{AB}, 0\}$ ja tapahtuma $T = \{\mathrm{A}, \mathrm{B}\}$

$$P(T) = P(\mathrm{A}) + P(\mathrm{B}) = 0{,}44 + 0{,}17 = 0{,}61$$

Edellä käydyt tilastollisen todennäköisyyden ominaisuudet antavat aiheen todennäköisyyden laajempaan määrittelyyn:
Määritelmä, *todennäköisyys äärellisessä perusjoukossa*:
Äärellisessä perusjoukossa $E = \{e_1, e_2, e_3, \dots, e_n\}$ määritelty funktio P on *todennäköisyysfunktio*, jos seuraavat ehdot täyttyvät:

1) Funktion P arvot eli *alkeistapausten todennäköisyydet* ovat ei-negatiivisia reaalilukuja

$$P(e_1), P(e_2), P(e_3), \dots, P(e_n) \geq 0.$$

2) Kaikkien alkeistapausten todennäköisyyksien summa on 1. Toisin sanoen, varman tapahtuman todennäköisyys on 1

$$P(e_1) + P(e_2) + P(e_3) + \cdots + P(e_n) = 1$$

Tapahtuman $A = \{a_1, a_2, a_3, \dots, a_m\}$ todennäköisyys on suotuisien alkeistapausten $a_1, a_2, a_3, \dots, a_m$ todennäköisyyksien summa

$$P(A) = P(a_1) + P(a_2) + P(a_3) + \cdots + P(a_m).$$

Huomaa erityisesti 1) – kohdan ei-negatiivisuus, mikä tarkoittaa sitä, että jollekin alkeistapaukselle e_k voi olla $P(e_k) = 0$. Kuitenkin jollakin toisella tod.näk.funktiolla Q voi olla $Q(e_k) > 0$ (sama alkeistapaus).

Esimerkki 1: Olkoon $E = \{1,2,3,4\}$ ja $P(1) = 0{,}1$, $P(2) = 0{,}2$, $P(3) = 0{,}3$ ja $P(4) = 0{,}4$

a) Osoita, että P on todennäköisyysfunktio

b) Laske $P(A)$ kun $A = \{1,3\}$

a) Nyt $P(1)$, $P(2)$, $P(3)$, $P(4) \geq 0$ ja summa

$$P(1) + P(2) + P(3), + P(4) = 0{,}1 + 0{,}2 + 0{,}3 + 0{,}4 = 1$$

b)

$$P(A) = P(\{1,3\}) = P(1) + P(3) = 0{,}1 + 0{,}3 = 0{,}4$$

Esimerkki 2:

$P(e_9) = 0$, mutta $Q(e_9) = 0{,}19$

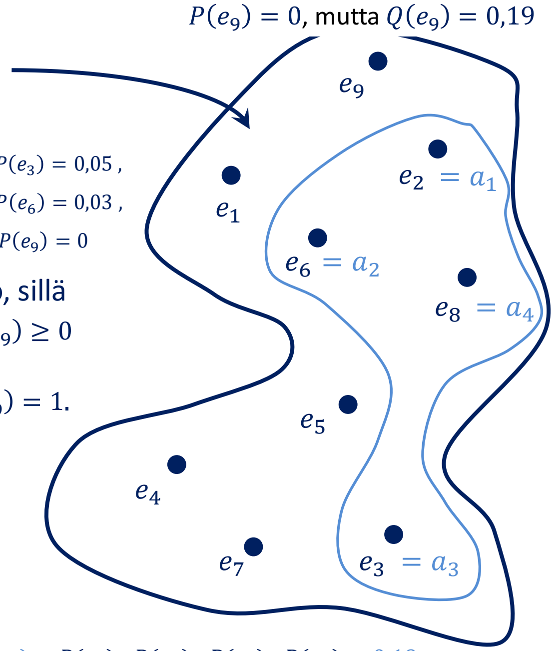
$E = \{e_1, e_2, e_3, \dots, e_9\}$

Tiedetään, että:

$P(e_1) = 0{,}13$, $P(e_2) = 0{,}05$, $P(e_3) = 0{,}05$,

$P(e_4) = 0{,}47$, $P(e_5) = 0{,}01$, $P(e_6) = 0{,}03$,

$P(e_7) = 0{,}2$, $P(e_8) = 0{,}06$, $P(e_9) = 0$



Tällöin P on tod.näk.funktio, sillä

$$P(e_1), P(e_2), P(e_3), \dots, P(e_9) \geq 0$$

ja

$$P(e_1) + P(e_2) + \dots + P(e_9) = 1.$$

$$A = \{a_1, a_2, a_3, a_4\}$$
$$= \{e_2, e_6, e_3, e_8\}$$

Tapahtumalle A pätee:

$$P(A) = P(a_1) + P(a_2) + P(a_3) + P(a_4) = P(e_2) + P(e_6) + P(e_3) + P(e_8) = 0{,}19$$

Tilastollinen todennäköisyys (jatkuu):
Kun tarkastellaan tietyn ilmiön alkeistapausta ja sen esiintymiskertoja, niin suoritetaan useita ilmiöön liittyviä satunnaiskokeita ja merkitään saadut tulokset muistiin. Pitkissä koesarjoissa kunkin alkeistapauksen esiintymiskertojen suhde tehtyihin satunnaiskoekertoihin lähestyvät alkeistapauksen todennäköisyyttä sitä paremmin mitä enemmän suoritetaan satunnaiskokeita.

Satunnaisilmiö	Alkeistapaus	Satunnaiskokeen tulos
Kolikon heitto	Saadaan kruuna	Kruuna
	Saadaan klaava	Klaava

1. heitto	2. heitto	3. heitto	4. heitto	5. heitto	6. heitto	7. heitto	8. heitto
kruuna	kruuna	kruuna	klaava	klaava	kruuna	kruuna	klaava
9. heitto	10. heitto	11. heitto	12. heitto	13. heitto	14. heitto	...	n. heitto
klaava	kruuna	klaava	klaava	klaava	klaava	...	kruuna

Tapahtuman, joka koostuu äärellisestä määrästä alkeistapauksia, todennäköisyys on suotuisien alkeistapausten todennäköisyyksien summa.

Alkeistapauksen ja tapahtuman todennäköisyyttä merkitään

$$P(\text{tapahtuma}) = \text{jotain}, \qquad \text{esim. } P(kr.) = 0{,}5.$$

Frekvenssillä tarkoitetaan alkeistapauksen esiintymiskertojen lukumäärää (esiintymistiheyttä), merkitään $f(\text{alkeistapaus})$, esim. $f(kr.) = 167$. Kaikkien satunnaiskokeiden lukumäärää merkitään kirjaimella n, esim. $\mathrm{n} = 346$. Alkeistapauksen "kruuna" *suhteellinen frekvenssi* on $\frac{f}{n} = \frac{167}{346} \approx 0{,}48 \approx P(kr.)$.

1. heitto	2. heitto	3. heitto	4. heitto
kruuna	kruuna	kruuna	klaava
5. heitto	6. heitto	7. heitto	8. heitto
klaava	kruuna	kruuna	klaava
9. heitto	10. heitto	11. heitto	12. heitto
klaava	kruuna	klaava	klaava
13. heitto	14. heitto	...	n. heitto
klaava	klaava	...	kruuna

n	$f(kr.)$	$\frac{f}{n}$
1	1	$\frac{1}{1} = 1$
2	2	$\frac{2}{2} = 1$
10	6	$\frac{6}{10} \approx 0{,}6$
13	6	$\frac{6}{13} \approx 0{,}46$
14	6	$\frac{6}{14} \approx 0{,}42$
n	167	$\frac{167}{364} \approx 0{,}48$

Geometrinen todennäköisyys

Kun perusjoukossa (otosavaruudessa) E on ääretön määrä alkeistapauksia, niin klassinen todennäköisyys ei toimi. Tällöin hyödynnetään tilastollista todennäköisyyttä.

Monesti tilastollisen todennäköisyyden selvittäminen on aikaa vievää ja ehkä haastavaakin. Tällöin voidaan hyödyntää satunnaisilmiöön (satunnaiskokeeseen) liittyvää geometriaa eli geometrisia mittoja, kuten pituuksia, pinta-aloja, tilavuuksia todennäköisyyksiä laskettaessa/määritettäessä.

Määritelmä, *geometrinen todennäköisyys:*
Jos perusjoukko E ja sen tapahtuma A voidaan tulkita tason alueiksi E ja A, niin tapahtuman A *geometrinen todennäköisyys* on näiden alueiden pinta-alojen $M(A)$ ja $M(E)$ suhde

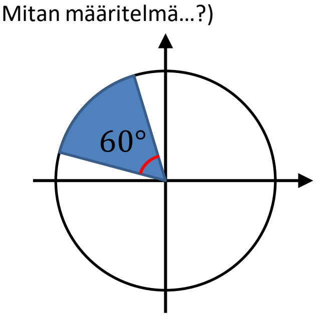
$$P(A) = \frac{M(A)}{M(E)}.$$

Määritelmä, *geometrinen todennäköisyys (jatkuu)*:
Jos perusjoukko E ja tapahtuma A voidaan tulkita janoiksi tai 3-ulot. kappaleiksi, niin tapahtuman A *geometrinen todennäköisyys* määritellään vastaavasti janojen pituuksien tai 3-ulot. kappaleiden tilavuuksien suhteena. Yleisesti

$$P(A) = \frac{M(A)}{M(E)} = \frac{\text{joukon } A \text{ geometrinen mitta}}{\text{joukon } E \text{ geometrinen mitta}}$$

Huom! Geometrista todennäköisyyttä hyödynnetään kun perusjoukko on ääretön, mutta jollakin tavoin mitattavissa. (Mitan määritelmä…?)

Esimerkki:
Tarkastellaan tason yksikköympyrän (säde on siis 1) pisteitä. Niitä on äärettömän paljon. Mikä on todennäköisyys, että umpimähkään ympyrältä valittu piste osuisi kuvassa näkyvään sektoriin?



VAST: Pinta-alojen suhteesta saadaan

$$\frac{\text{sektorin p.a.}}{\text{ympyrän p.a.}} = \frac{60}{360} = \frac{1}{6} \approx 0{,}1\overline{6}$$