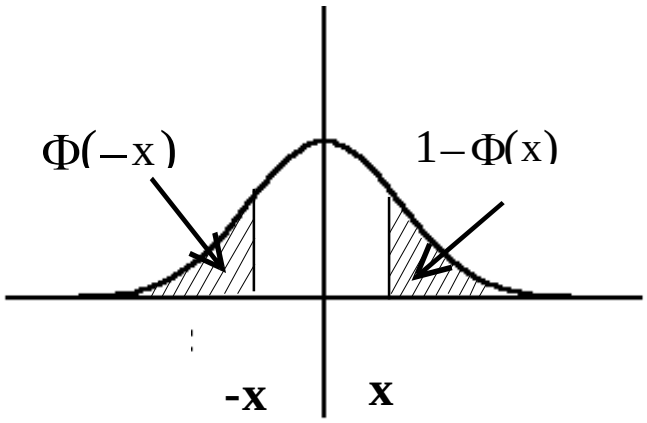
Antti Majaniemi

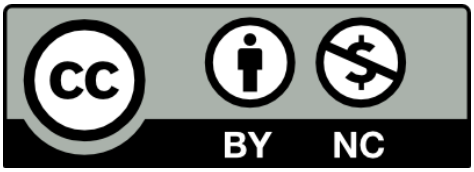
Matematiikka IV

Tilastot ja todennäköisyys



2016

ISBN 978-952-93-8171-5



Turussa 20.11.2016

Jari Majaniemi

jari @ anttimajaniemi.fi

Sisällys ja johdanto

Tämän monisteen alkupuoli on muokattu aikaisemmasta monisteesta mm. siten, että satunnaismuuttujia ja normaalijakaumaa koskeva luku on siirretty aikaisemmaksi. Näin siksi, että lukuvuoden lopussa aika yleensä loppuu kesken ja jos tämän monisteen aihepiiriä käsitellään viimeisenä, ainakin lukujen 1 – 2 asioita olisi hyvä ehtiä esitellä, jos ei muuten niin piirto- tai siirtoheittimen avulla. Nämä asiat voisi sijoittaa myös matematiikan kurssien johonkin aikaisempaan kohtaan, esim. differentiaaliyhtälöiden käsittelyn jälkeen. Koko monisteen läpikäyminen voisi sopia valinnaiselle tilastotieteen ja todennäköisyyslaskennan kurssille.

Monisteeseen loppupuolelle on lisätty aika paljon todennäköisyyslaskentaa (luvut 8 ja 9), jotta myös lisätylle tilastolaskennalle saadaan matemaattista pohjaa. Harjoitustehtävien laatimisessa olen välttänyt jollekin opintosuunnalle liittyviä sanontoja ja käyttänyt aika paljon yleisnimikkeitä "tuotteen", "jonkin ihmisjoukon" tms. "eräs ominaisuus". Ajatuksena on, että opettaja voi tarvittaessa muuttaa nimikkeitä (kirjallisuudesta löytyvien esimerkkien avulla ja opiskelijoiden kanssa pohtimalla) juuri opetettavalle opintosuunnalle soveltuviksi.

Kun tilastotieteessä tehdään matemaattisesta laskennasta (jonka pohjana on tämän monisteen tasolla aika usein normaali-, binomi-, Student- tai χ^2-jakauma) tilastollisia johtopäätöksiä, täytyy niiden tekemisessä olla erityisen varovainen, jotta vältyttäisiin virhepäätelmiltä. Oikeiden johtopäätösten teko voi vaatia paljon syvällisempää tilastollista näkemystä kuin tästä monisteesta saa. Joissakin oppikirjoissa on tällaisista asioista hyviä varoittavia esimerkkejä.

Monisteen laatimisessa olen käyttänyt apuna mm. seuraavia suomenkielisiä oppikirjoja ja monisteita: Holopainen: Tilastomatematiikan perusteet (Otava), Holopainen, Pulkkinen: Tilastolliset menetelmät (Weilin + Göös), Launonen, Sorvali, Toivonen: Teknisten ammattien matematiikka 3E (WSOY), Pelkonen: Tilastomatematiikka (Gummerus), Raija Tuohi: Tilastomatematiikan oppijakso (Moniste, Turun ammattikorkeakoulu), Äijälä: Todennäköisyyslaskenta ja tilastotiede (Tammertekniikka). Apuna on ollut myös käsin kirjoitettu moniste, jonka laadin aikoinaan Turun yliopistossa, kun luennoin siellä todennäköisyyslaskennan cl-kurssin. Mainittakoon myös seuraava aika uusi, laaja ja ainakin päällisin puolin katsottuna hyvän näköinen oppikirja: A. Hayter: Probability and Statistics for Engineers and Scientists sekä Shaumin sarjan monisteet Probability ja Statistics.

Kiitokset erityisesti Ritva Metsänkylälle, joka on oikolukenut monistetta sitä myöten kun olen saanut sitä kirjoitetuksi. Jäljelle jääneistä virheistä voin kuitenkin syyttää vain itseäni.

Turussa 11. 8. 1998

Antti Majaniemi

Olen päivittänyt Antti Majaniemen alkuperäiseen monisteeseen esimerkkejä ja harjoitustehtäviä tämän päivän tilanteeseen paremmin sopiviksi.

Turussa 20. 8. 2007 Jari Majaniemi

1 Tilasto-opin alkeita

1.1 Keskiarvot

Lukujen $x_1, x_2, \ldots, x_n$

- ***(aritmeettinen) keskiarvo*** $\bar{x} = \dfrac{x_1 + \ldots + x_n}{n} = \dfrac{1}{n}\sum x_i$

- ***geometrinen keskiarvo (keskiverto)*** $G = \sqrt[n]{x_1 x_2 \cdots x_n}$ $(x_i\text{:t} \geq 0)$

- ***harmoninen keskiarvo*** $H = \dfrac{n}{\dfrac{1}{x_1} + \ldots + \dfrac{1}{x_n}}$ $\therefore \dfrac{1}{H} = \dfrac{1}{n}\sum \dfrac{1}{x_i}$.

Täten H on lukujen x_i käänteislukujen $1/x_i$ aritmeettisen keskiarvon käänteisluku.

- Aritmeettinen eli "tavallinen" keskiarvo lasketaan usein ***painotettuna***:

$$\bar{x} = \frac{p_1 x_1 + \ldots + p_n x_n}{\sum p_i}, \quad p_i\text{:t painokertoimia}.$$

Esim. 1 Opintojaksosta 1 (2 op) opiskelija sai arvosanan 2 ja opintojaksosta 2 (6 op) arvosanan 4. Arvosanojen

– painottamaton keskiarvo $\bar{x} = \dfrac{2+4}{2} = 3$,

– painotettu keskiarvo, painoina opintopistemäärät

$$\bar{x} = \frac{2 \cdot 2 + 6 \cdot 4}{2 + 6} = 3\tfrac{1}{2}.$$

Esim. 2 Jos luvuista x_i on osa samoja (f_i kpl x_i:tä, yhteensä n kpl), niin keskiarvo voidaan laskea painotettuna, painoina lukujen x_i esiintymismäärät eli ***frekvenssit*** f_i:

$$\bar{x} = \frac{f_1 x_1 + \ldots + f_k x_k}{f_1 + \ldots + f_k} = \frac{1}{n}\sum f_i x_i \ .$$

Esim. 3 Erään luokan matematiikan arvosanojen ***jakauma*** oli seuraavanlainen

x_i	f_i	f_i / %
0	1	3,70
1	2	7,41
2	4	14,81
3	13	48,15
4	5	18,52
5	2	7,41
yht.	27	100,00

$$\bar{x} = \frac{1 \cdot 0 + 2 \cdot 1 + ... + 2 \cdot 5}{27} \approx 2{,}93 .$$

Keskiarvo voidaan laskea myös *frekvenssiprosenteilla*:

$$\bar{x} = \frac{3{,}7 \cdot 0 + 7{,}4 \cdot 1 + ... + 7{,}4 \cdot 5}{100} \approx 2{,}93 .$$

****Esim. 4*** Hyvinä aikoina alkupalkka = a ja

1. vuonna palkka nousi 10 % $\therefore$ palkka = 1,10 a

2. - " - nousu oli 6 % $\therefore$ palkka = 1,06·1,10 a

3. - " - - " - oli 2 % $\therefore$ palkka =1,02·1,06·1,10 a.

Samaan loppupalkkaan päästäisiin, jos kertojana olisi jokaisena vuonna luku $\sqrt[3]{1{,}02 \cdot 1{,}06 \cdot 1{,}10} \approx 1{,}0595$ $\therefore$ keskimääräinen nousu on 5,95 % (eikä 6 % kuten aritmeettinen keskiarvo antaisi). Keskimääräinen kasvu on siis *kasvukertoimien* geometrinen keskiarvo.

****Esim. 5*** Matka s ajettiin kolme kertaa ja nopeudet olivat 60, 80 ja 100 (km/h). Laske keskinopeus.

Kokonaismatka = 3s, kokonaisaika $t = \frac{s}{60} + \frac{s}{80} + \frac{s}{100}$.

$$v_k = \frac{3s}{t} = \frac{3}{\frac{1}{60} + \frac{1}{80} + \frac{1}{100}} \approx 76{,}6 \text{ (km/h)} \quad \text{(eikä 80 km/h).}$$

$\therefore$ v_k on nopeuksien harmoninen keskiarvo.

***Esim. 6** **Apukeskiarvon** x_0 käyttäminen: Arvioidaan, että lukujen $x_i\,(i=1,\ldots,k)$ aritm. keskiarvo $\approx x_0$. Lukujen x_i *poikkeamat* tästä apukeskiarvosta x_0 ovat

$$x_i' = x_i - x_0 \qquad \therefore\ x_i = x_0 + x_i'.$$

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{n}\sum (x_0 + x_i') = \frac{1}{n}\cdot nx_0 + \frac{1}{n}\sum x_i' = x_0 + \frac{1}{n}\sum x_i'.$$

$\therefore$ *Keskiarvo = apukeskiarvo + poikkeamien keskiarvo*. Esimerkki:

Luvut 12, 17, 14, 15, 18, 16, 13, 14, 16, 13 (10 kpl).

Valitaan $x_0 = 14$. Poikkeamat: –2, 3, 0, 1, 4, 2, –1, 0, 2, –1.

Poikkeamien summa = +8 $\quad\therefore\quad \bar{x} = 14 + \frac{8}{10} = 14{,}8$.

1.2 Tilastollisia peruskäsitteitä

Esimerkissä 3 esiintyi yksinkertainen esimerkki ***tilastosta***. Siihen oli koottu erään luokan matematiikan arvosanat frekvensseineen (viereinen kuva). Tilasto on siinä mielessä ***järjestetty***, että arvosanat on esitetty suuruusjärjestyksessä. Aineistoa on myös ***luokiteltu***, koska samat arvosanat on kerätty yhteen. Voidaan myös sanoa, että esim. arvosana 3 on arvosanavälin (***luokkavälin***) 2,5 … 3,5 ***luokkakeskus***.

x_i	f_i
0	1
1	2
2	4
3	13
4	5
5	2
yht.	27

Tässä tilastossa tutkimuskohteina (***tilastoyksikköinä***) ovat luokan oppilaat $a_1, a_2, \ldots, a_{27}$. Tilastoyksikköjen joukko $\{a_1, a_2, \ldots, a_{27}\}$ muodostaa ***populaation***. Tutkittavana on ominaisuus "menestyminen matematiikassa". Se on eräs tähän populaatioon liittyvä ***tilastollinen muuttuja x*** (*tilastollinen suure*).

Muuttujan *x* saamat arvot x_i: 0, 1, 2, 3, 4, 5
ja näiden frekvenssit f_i: 1, 2, 4, 13, 5, 2
muodostavat tilastollisen muuttujan *x* ***jakauman***.

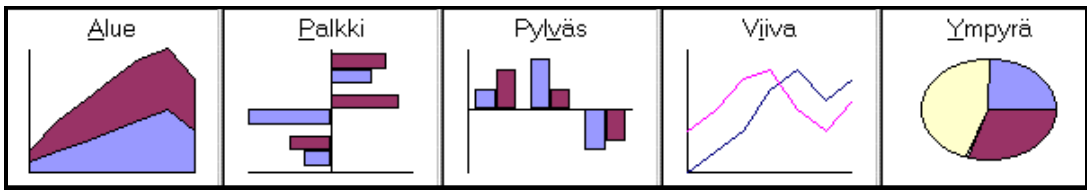
Tämäntyyppistä jakaumaa sanotaan myös ***frekvenssijakaumaksi***, koska se ilmoittaa, miten frekvenssit ovat jakautuneet *x*:n arvojen kesken (paljonko on nollia, ykkösiä jne ts. mikä on 0:n frekvenssi, mikä 1:n jne).

Edellinen muuttuja x on luonteeltaan ***diskreetti muuttuja***: sen arvot muuttuvat hyppäyksittäin. Diskreetin muuttujan "vastakohta" on ***jatkuva muuttuja*** (esim. ikä, pituus, nestemäärän paino jne). Toisentyyppinen perusjaottelu on seuraava:

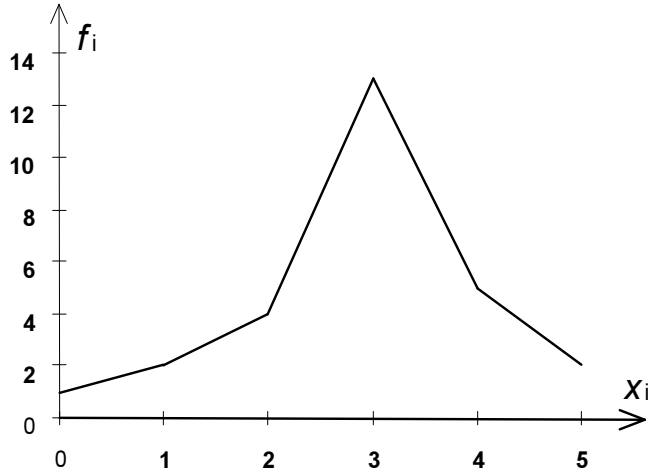
A. Edellinen muuttuja x on eräs ***kvantitatiivinen*** *(määrällinen)* ***muuttuja***. Sen arvot x_i ovat lukuja, joten ***x:n arvoista voidaan laskea esim. keskiarvo, keskihajonta*** ja ***vaihteluväli***.

B. ***Kvalitatiiviset*** *(laadulliset)* ***muuttujat*** kuten sukupuoli, puoluekanta, luonnetyyppi, ammatti jne ovat sellaisia, että *niistä ei voida laskea keskiarvoa tai keskihajontaa tms., vaan jakaumaa täytyy kuvata toisentyyppisillä* ***tunnusluvuilla*** (kuten *moodi*). Tunnuslukuja esitellään myöhemmin.

Frekvenssijakauman (x:n arvot x_i ja niiden frekvenssit f_i) erilaisia graafisia esitystapoja (kaavioita, *diagrammeja*) on suuri joukko mm. Excelissä. Seuraavassa kuvassa on näistä esitystavoista muutama.



Jatkon kannalta tärkeä esitystapa on ns. ***frekvenssikäyrä***, jossa on yhdistetty pisteet (x_i, f_i) murtoviivalla***:***



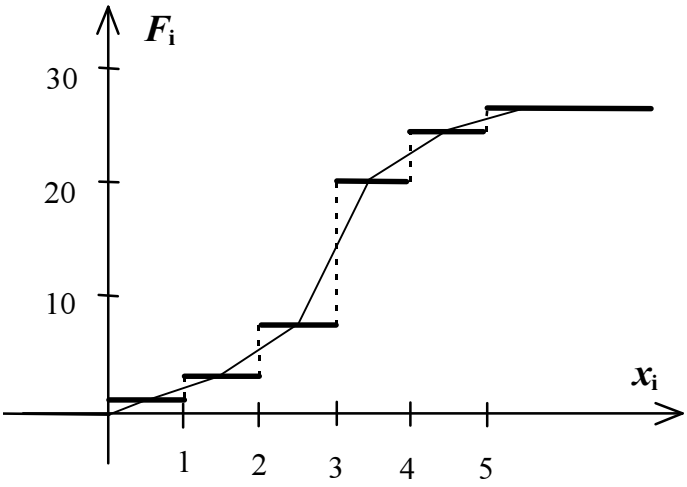
Käyrää voitaisiin sanoa myös jakauman (tai muuttujan x) ***tiheyskäyräksi***. Arvosanan 3 läheisyydessä käyrä on ylimmillään, ts. arvosanojen määrä yksikköä kohti on suurimmillaan eli *arvosanatiheys* on suurimmillaan.

Tiheys pienenee käyrän alku- tai loppupäähän päin mentäessä. Käyrästä nähdään mm., että se on aika hyvin *symmetrinen*.

Toinen tärkeä käyrä saadaan, kun tutkitaan miten frekvenssit kertyvät arvosanojen mukana. Tämä saadaan ***summafrekvenssien*** F_i avulla:

$$F_i = f_1 + \ldots + f_i .$$

x_i	f_i	$F_i = f_1 + \ldots + f_i$
0	**1**	**1**
1	**2**	**1 + 2 = 3**
2	**4**	**3 + 4 = 7**
3	**13**	**7 + 13 = 20**
4	**5**	**20 + 5 = 25**
5	**2**	**25 + 2 = 27**
yht.	**27**	



Summafrekvenssikäyrää sanotaan myös ***kertymäkäyräksi***. *Diskreetin muuttujan kertymäfunktio on porrasfunktio*, sillä esim. arvosanojen 2 ja 3 välillä ei kerry yhtään arvosanaa, joten tällä välillä funktion arvo on vakio. Kohdassa 3 arvosanoja kertyy 13 kpl, joten tässä kohdassa käyrässä on 13 yksikön suuruinen hyppäys.

Voidaan myös ajatella, että esim. arvosana 3 on luokan 2,5 … 3,5 luokkakeskus ja kaikki arvosanat 3 eivät ole samanarvoisia vaan niitä kertyy tasaisesti kohdasta 2,5 kohtaan 3,5. Täten 13 yksikön suuruinen hyppäys voidaan korvata janalla kohtaa 3 edeltävän portaan (askelman) keskikohdasta seuraavaan. Näin saadaan jatkuva viiva, joka myös on piirretty edelliseen kuvaan.

1.3 Jakauman tunnuslukuja

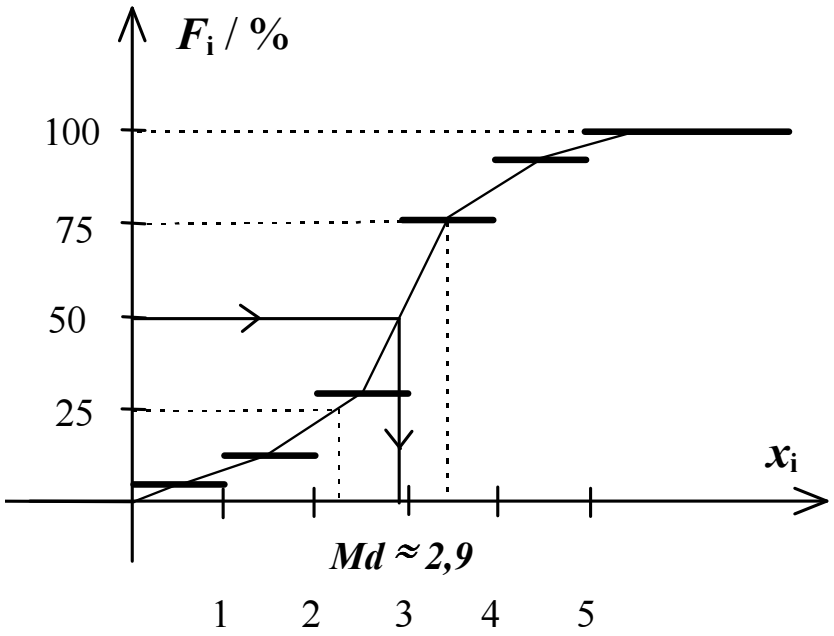
Jakaumaa kuvataan graafisten esitystapojen lisäksi ns. *tunnusluvuilla*. Nämä jaetaan kahteen tyyppiin:

A. KESKILUVUT (keskikohtaa ilmaisevat luvut):

- ***keskiarvo*** $\bar{x} = \frac{1}{n}\sum f_i x_i$,

- ***moodi*** eli tyyppiarvo (*Mo*). Tämä on se muuttujan arvo, jolla on suurin frekvenssi. Edellisessä esimerkissä *Mo* = 3 ja $\bar{x} \approx 2{,}93$. *Moodia voidaan käyttää myös kvalitatiivisille muuttujille*. Jos muuttuja *x* ilmaisee esim. puoluekannan (vaikkapa jossakin kunnassa), niin *x*:n jakaumasta ei voida laskea keskiarvoa, mutta moodi ilmaisee sen puolueen, jolla on suurin kannatus.

- ***mediaani*** (*Md*) on se kohta, johon mennessä on kertynyt puolet (eli 50 %) muuttujan arvoista. Tämä arvo nähdään jatkuvaksi viivaksi muutetusta kertymäkäyrästä "takaperin":



Mediaaniarvo on ns. *50 %* :n ***fraktiili***. 25 %:n fraktiilia eli sitä *x*:n arvoa, mihin mennessä on kertynyt 25 % *x*:n arvoista, sanotaan *ensimmäiseksi* ***kvartiiliksi*** tai ***alakvartiiliksi*** Q_1. Vastaavasti 75 %:n kohta on kolmas kvartiili eli ***yläkvartiili*** Q_3. Edellisessä kuvassa $Q_1 \approx 2{,}3$ ja $Q_3 \approx 3{,}4$.

B. HAJONTALUVUT (kuinka laajalle *x*:n arvot ovat hajaantuneet tai kuinka paljon ne keskimäärin poikkeavat jostakin keskikohtaa esittävästä *x*:n arvosta):

- ***vaihteluväli*** = suurimman ja pienimmän esiintyvän *x*:n arvon erotus. Esimerkiksi edellä käsitellyllä luokalla vaihteluväli on 5 – 0 = 5. Jollakin toisella luokalla keskiarvo voi olla sama, mutta vaihteluväli esim. 4 – 1 = 3.

- ***kvartiilipoikkeama*** $Q = \frac{Q_3 - Q_1}{2}$. Jos jakauma on symmetrinen, niin 50 % *x*:n arvoista on välillä $Md \pm Q$. Edellisessä kuvassa

$$Q \approx \frac{3{,}4 - 2{,}3}{2} = 0{,}55 .$$

- ***keskipoikkeama*** (*Mean deviation*) $d = \frac{1}{n}\sum f_i |x_i - \bar{x}|$ = poikkeamien itseisarvojen keskiarvo. Koska itseisarvolausekkeiden käsittely on hankalaa, käytetään yleensä neliöllisiä poikkeamia, jolloin saadaan seuraavat hajontaluvut:
- ***Keskihajonta*** (*Standard deviation*) $\sigma = \sqrt{\frac{1}{n}\sum f_i (x_i - \bar{x})^2}$ = poikkeamien neliöiden keskiarvon neliöjuuri.
- ***Varianssi = keskihajonnan neliö =*** σ^2 .

Keskihajonta σ on *koko aineiston hajonta*. Siitä käytetään myös merkintää σ_n ja sanontaa ***n*-hajonta**.

- Jos koko aineistosta otetaan n:n kappaleen ***otos***, niin suuretta

$$s = \sqrt{\frac{1}{n-1}\sum f_i (x_i - \bar{x})^2}$$

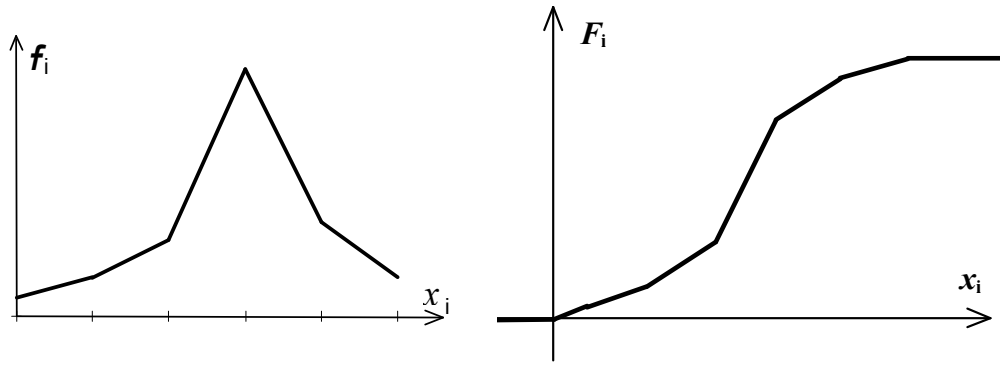
sanotaan ***otoshajonnaksi*** tai $(n-1)$-hajonnaksi ja merkitään myös s_{n-1}:llä. Vastaavasti puhutaan ***otosvarianssista*** s^2.

Jos esim. 10 kappaleen otoksesta lasketaan otoshajonta s, sitä voidaan käyttää koko aineistosta (esim. 10000 kpl) lasketun keskihajonnan σ likiarvona, jos jakauma on "lähes normaali". Tästä puhutaan lähemmin jatkossa.

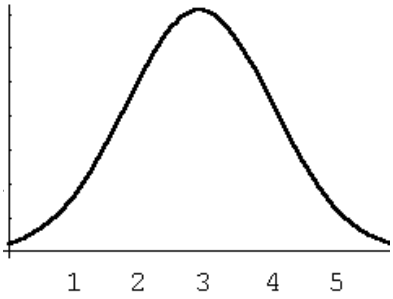
Mainitaan vielä jakauman vinoutta (*skewness*) mittaavia lukuja. ***Pearsonin vinousmitta*** $s_p = \frac{\bar{x} - Mo}{s} \approx \frac{3(\bar{x} - Md)}{s}$ (kokeellisesti saatu likiarvo). Positiivisesti eli oikealle vinolla jakaumalla ovat moodi ja mediaani keskiarvoa pienemmät ja siten $s_p > 0$. ***Vinouskerroin*** $g_1 = \frac{\sum (x_i - \bar{x})^4}{n \cdot s^4}$.

1.4 Likimain normaali jakauma. Luottamusvälit

Edellä käsitelty matematiikan arvosanojen jakauma on eräs tilastollinen jakauma. Sen frekvenssikäyrä (tiheysfunktion kuvaaja) ja summakäyrä (kertymäfunktion kuvaaja) olivat muodoltaan seuraavanlaiset:



Jos arvosanoja olisi paljon ja arvosteluväli olisi lyhempi (esim. ¼-numeroa), niin f_i-käyrä olisi muodoltaan lähellä ns. *Gaussin kellokäyrää* (viereinen kuva), joka on todennäköisyyslaskennan antaman erään teoreettisen (mutta käytännössä tärkeän) jakauman, ***normaalijakauman*** frekvenssifunktio.



Todennäköisyyslaskennassa frekvenssifunktioita sanotaan yleensä *tiheysfunktioiksi*.

Normaalijakaumaa käytetään useiden tilastollisten jakaumien *matemaattisena mallina.* Eräs todennäköisyyslaskennan avulla saatava "nyrkkisääntö" sille, milloin tilastollisen muuttujan *x* jakauma on likimain normaalijakauma on seuraava:

Jos tilastollisen muuttujan x arvojen vaihtelut johtuvat useista, toisistaan riippumattomista satunnaisista seikoista, niin x:n jakauma on likimain normaalijakauma. Tällöin *x*:n jakauman käsittelyssä voidaan käyttää normaalijakaumalle johdettuja matemaattisia tuloksia.

***Esim.** 7* Likimain normaaleja jakaumia voisivat olla esim. seuraavat:

- suuresta ihmisjoukosta muodostettu
  - painojakauma
  - pituusjakauma
  - arvosanajakauma,

- eri työntekijöillä saman työvaiheen tekemiseen käytetty aika, jos näitä työntekijöitä on aika paljon.
- jonkin autotyypin autonmoottorien kestoikä (esim. laskettuna km:issä ennen ensimmäistä moottoriremonttia).
- Jos suureen arvon määrittämiseksi tehdään useita toisistaan riippumattomia mittauksia samoissa olosuhteissa, niin mittausarvoissa esiintyy eroja. Tämä mittausarvojen jakauma on likimain normaali.

Normaalijakauman yhteydessä johdetaan myöhemmin tulos, joka antaa keskihajonnan σ suuruudesta jonkinlaisen kuvan. Likimain normaaleihin jakaumiin sovellettuna tulos on seuraava:

Lause 1 *Jos tilastollisen muuttujan x jakauma on likimain normaali, niin suunnilleen*

68 % x:n arvoista on välillä $\bar{x}-\sigma \ldots \bar{x}+\sigma$

95 % x:n arvoista on välillä $\bar{x}-1{,}96\sigma \ldots \bar{x}+1{,}96\sigma$

99 % x:n arvoista on välillä $\bar{x}-2{,}6\sigma \ldots \bar{x}+2{,}6\sigma$.

Esim. 8 Oletetaan, että suuresta määrästä matematiikan kokeita, joissa maksimipistemäärä on 20, on saatu tulokset

$$\bar{x}=12,\ \sigma=3{,}5\,.$$

Jos jakauman malliksi otetaan normaalijakauma, niin suurin piirtein

68 % (eli n. 2/3) pistemääristä on välillä 8,5 … 15,5 ja

95 % on välillä $12 \pm 1{,}96 \cdot 3{,}5$ eli suunnilleen välillä 5… 19.

Lauseessa 1 esim. väliä $\bar{x}-1{,}96\cdot\sigma \ldots \bar{x}+1{,}96\cdot\sigma$ sanotaan ***95 %:n luottamusväliksi***. Tämän välin päätearvoja $\bar{x}-1{,}96\sigma$ ja $\bar{x}+1{,}96\sigma$ ***sanotaan myös 95 %:n varmuusrajoiksi*** tai myös ***5 %:n riskirajoiksi***.

Esim. 9 Jos edellisen esimerkin mukaisesta koepistejoukosta valitaan umpimähkään yksi, niin 68 % varmuudella (ts. *todennäköisyydellä* 0,68) tämä arvosana on välillä 8,5 … 15,5. Koska vain 68 % arvosanoista on tällä välillä, niin otetaan aika suuri eli 32 %:n riski, jos väitetään että yhdessä valinnassa saadaan tällä välillä oleva arvosana.

1.5 Otos. Keskiarvon keskivirhe

Tilastollisen jakauman tutkimisessa joudutaan tyytymään usein otokseen (esim. laadunvalvonnassa tutkitaan vain osa tehdyistä tuotteista). Tällöin heräävät mm. seuraavat kysymykset:

◊ Miten pitkälle meneviä johtopäätöksiä otoskeskiarvosta $\bar{x}_{otos}$ ja otoshajonnasta eli $(n-1)$-hajonnasta s voidaan tehdä koko populaatiota ajatellen.

◊ Kuinka voimakkaasti otoskoko vaikuttaa. Esim. turhan suuri otos merkitsee usein lisäkustannuksia.

◊ Miten otos pitäisi tehdä.

◊ Miten kysely pitäisi tehdä, jotta se olisi luotettava (ei esim. puhelinkyselynä; lisäksi kaikkien kyselyyn mukaan joutuneiden pitäisi oikeastaan vastata, sillä muuten jakauma helpsti "vinoutuu").

Eräs aika itsestään selvältä tuntuva perustulos on seuraava:

Lause 2 *Otoskeskiarvo $\bar{x}_{otos}$ ja otoshajonta s ovat hyviä estimaatteja (arvioita) koko populaation keskiarvolle $\bar{x}$ ja keskihajonnalle σ (sitä parempia, mitä suurempi otos on ja mitä paremmin otos on onnistuttu suorittamaan).*

Esim. 10 Oletetaan, että tehtaan valmistamien tuotteiden kestoaikojen jakaumaa voidaan pitää likimain normaalina. Tehdas testasi kestoaikoja 100 kpl otoksella, ts. sataa tuotetta käytettiin, kunnes ne menivät rikki ja kestoajat luokiteltiin esimerkiksi $\sqrt{100}=10$ yhtä pitkään luokkaan. Ajatellaan, että tulokseksi saatiin esim. tunneissa laskettuina

$$\bar{x}_{otos}=990, \quad s=60 \quad ((n-1)\text{-hajonta}).$$

Tästä voidaan päätellä *Lauseen 2* mukaan, että kaikilla tuotteilla kestoaikojen keskiarvo ja keskihajonta ovat

$$\bar{x}\approx 990,\ \sigma\approx 60.$$

Lauseen 1 mukaan taas esim. 95 %:lla tuotteista kestoikä on suunnilleen välillä

$$990\pm 1{,}96\cdot 60 \text{ eli välillä } 870 \ldots 1110.$$

Lopuista 5 %:sta puolet eli 2,5 % kestävät alle 870 ja puolet yli 1110 tuntia. Täten 97,5 % *varmuudella* (tai 97,5 % todennäköisyydellä) umpimähkään valittu tuote kestää yli 870 tuntia.

Kovasti yksinkertaistaen voidaan sanoa myös, että jos tehdas antaa tuotteelleen 870 tunnin takuun, se ottaa 2,5 % *riskin*. Takuuehtojen määrittäminen ei tietenkään käytännössä ole näin yksinkertaista vaan riippuu mm. siitä, kuinka herkästi viallinen tuote palautetaan.

Otoshajonnan s avulla voidaan tutkia, paljonko otoskeskiarvo poikkeaa koko populaation keskiarvosta, ts. *paljonko otoskeskiarvossa on virhettä*. Kun käytetään todennäköisyyslaskentaa ja sovelletaan sen antamia tuloksia, saadaan tähän kysymykseen seuraava vastaus :

Lause 3 *Oletetaan, että tilastollisen muuttujan x jakauma on likimain normaali ja otoskoko* $n \geq 30$. *Silloin*

1) 95 %:n varmuudella koko populaation keskiarvo poikkeaa otoskeskiarvosta korkeintaan luvun $1{,}96 \cdot \frac{s}{\sqrt{n}}$ *verran, ts.*

$$\boxed{\bar{x}_{popul.} = \bar{x}_{otos} \pm 1{,}96 \cdot \frac{s}{\sqrt{n}}},$$

2) 99 %:n varmuudella

$$\boxed{\bar{x}_{popul.} = \bar{x}_{otos} \pm 2{,}6 \cdot \frac{s}{\sqrt{n}}}.$$

Esim 11 Edellisessä esimerkissä tuotteen kestoikää arvioitaessa otoksesta $n = 100$ saadut arvot olivat $\bar{x}_{otos} = 990$, $s = 60$. Täten

1) 95 % varmuudella kaikkien tuotteiden kestoiän keskiarvo on välillä

$$\bar{x} = 990 \pm 1{,}96 \cdot \frac{60}{\sqrt{100}} \approx 990 \pm 12 \quad \text{(h)},$$

2) 99 % varmuudella

$$\bar{x} = 990 \pm 2{,}6 \cdot \frac{60}{\sqrt{100}} \approx 990 \pm 16 .$$

Lauseessa 3 esiintyvää lukua $\boxed{\bar{s}_x = \frac{s}{\sqrt{n}}}$, joka kerrottuna luvulla 1,96 tai 2,6 antaa virherajat, sanotaan ***keskiarvon keskivirheeksi***. Jos otoskoko $n < 30$, virherajoja täytyy suurentaa eli ne täytyy kertoa eräällä 1:tä suuremmalla luvulla *t*. Tämä kerroin on peräisin eräästä toisesta jakaumasta, *Studentin t* -jakaumasta ja sen arvoja saadaan mm. seuraavasta taulukosta:

n	2	3	4	5	10	20	30
95%	2,2	1,6	1,4	1,3	1,1	1,0	1,0
99%	3,9	2,3	1,8	1,5	1,2	1,1	1,1

Esim. 12 Suureen arvon määrittämiseksi tehdyt mittaukset antavat tietyn mittaustulosten jakauman, joka on likipitäen normaali, jos mittauksia on suuri määrä ja ne ovat toisistaan riippumattomia. Kun mittauksia on vähäisempi määrä, kyseessä on itse asiassa otos tästä jakaumasta ja tähän otokseen voidaan soveltaa edellisiä tuloksia. Ajatellaan, että tehdään esim. 10 tällaista mittausta ja näiden keskiarvoksi saadaan $\bar{x}_{otos} = 28{,}5$ ja (*n*-1)-hajonnaksi $s = 0{,}68$. Tällöin mittausten keskiarvon keskivirhe on

$$\bar{s}_x = \frac{0{,}68}{\sqrt{10}} = 0{,}215...$$

Jos käytetään 95 % varmuutta, tämä luku on kerrottava luvulla 1,96 ja *t*-luvulla 1,1. Koska $1{,}1 \cdot 1{,}96 \cdot 0{,}215... = 0{,}463... < 0{,}5$, niin tulos on, että suureen todellinen arvo (äärettömän monen mittauksen keskiarvo) on

$$x = 28{,}5 \pm 0{,}5 \quad 95\ \%\ \text{varmuudella.}$$

Kun aikaisemmin monisteen I ja III osassa laskettiin suureiden $z = f(x, y)$ virheitä kokonaisdifferentiaalin avulla, muuttujien *x* ja *y* arvot virheineen annettiin yleensä esim. muodossa $x = 2{,}56 \pm 0{,}03$ ja $y = 32{,}1 \pm 0{,}5$, ts. ajateltiin, että esim. *x*:n arvoksi on mitattaessa saatu 2,56 ja arvioitiin, että mittausarvossa on virhettä korkeintaan 0,03 puoleen tai toiseen. Esimerkki 12 näyttää, miten mitattavien suureiden *x* ja *y* virheitä voidaan arvioinnin sijaan laskea tilastollisten menetelmien avulla.

Esim. 13 Tuotteen keskipainoksi ilmoitettiin 275 g. Punnittiin 32 tuotetta, joista keskiarvoksi saatiin 272 g ja otoshajonnaksi 11

g. Laske pitääkö ilmoitus tämän otoksen mukaan paikkansa, jos käytetään 95 % luotettavuustasoa (varmuutta).

Koska $1{,}96 \cdot \frac{11}{\sqrt{32}} \approx 3{,}8$, niin 95 % varmuudella kaikkien tuotteiden keskipaino on välillä $272 \pm 3{,}8$ eli välillä 268,2…275,8.

Täten kaikkien tuotteiden keskipaino voi olla jopa 275,8 g, ts. ilmoitettua keskipainoa 275 (g) ei voi tämän otoksen antaman tiedon perusteella pitää liian suurena. Siis ilmoitus pitää paikkansa.

Harjoituksia

A, B

Vastauksia monisteen lopussa.

1.1 Laske sopivaa apukeskiarvoa käyttäen lukujen 27, 24, 22, 26, 28, 26, 27 (aritmeettinen) keskiarvo. Laske myös samojen lukujen geometrinen ja harmoninen keskiarvo.

1.2 Tieosuudesta on 10 %:lla nopeusrajoitus 50, 30 %:lla 80 ja lopulla 100. Kuinka suuri on maksimikeskinopeus, johon tällä tieosuudella voidaan päästä nopeusrajoituksia noudattaen ?

1.3 Yrityksen liikevaihto laski kahtena peräkkäisenä vuotena 2 % kumpanakin. Millainen nousu tarvittaisiin kolmantena vuonna, jotta näiden kolmen vuoden keskimääräinen kasvu olisi 1 % ?

1.4 Diskreetin muuttujan x arvot ovat 4, 5, 6, 7, 8, 9 ja niiden frekvenssit 1, 2, 8, 13, 6, 2. Esitä x:n jakauma a) taulukkona, b) frekvenssikäyränä, c) *janadiagrammina*, jossa pisteistä (x_i, f_i) on piirretty x-akselille kohtisuoraan janat, d) *pylväsdiagrammina*, joissa em. janat on levennetty esim. ½ yksikön levyisiksi pylväiksi, e) *histogrammina*, jossa yhtä leveät pylväät ovat kiinni toisissaan ja y-akselina ovat frekvenssiprosentit. Tällöin pylväskuvion yhteispinta-ala on 1. f) Piirrä x:n kertymäkäyrä. g) Määritä jakauman keskiluvut ja hajontaluvut (opettele, miten laskimestasi saadaan keskiarvo, n-hajonta ja otoshajonta).

1.5 Pinossa olevat laudat luokiteltuina 30 cm pituisiin luokkiin jakautuivat seuraavan taulukon mukaisesti (lautojen pituuksien sijaan kyseessä voisi olla esim. työsuorituksiin käytettäviä aikoja tms.):

luokkaväli	*luokkakeskus* x_i	f_i %
315...345	330	5
345...375	360	8
375...405	390	17
405...435	420	35
435...465	450	20
465...495	480	9
495...525	510	6

Piirrä jakauman frekvenssi- ja summakäyrät ja määritä jälkimmäisestä jakauman mediaani, kvartiilipoikkeama sekä kuinka suuri osa laudoista on välillä 400…460. (Vihje: murtoviivaa ei tässä tehtävässä ole luonnollista vetää portaiden keskikohtien kautta.) Tämä antaa todennäköisyyden sille, että pinosta umpimähkään vedetty lauta on pituudeltaan kyseisellä välillä. Laske myös lautojen keskipituus ja keskihajonta.

1.6 Tilastollisen muuttujan jakauma on likimain normaali, keskiarvona 27,4 ja hajontana 2,86. Laske 95:n ja 99 %:n luottamusvälit.

1.7 Erään tuotteen keskipainon pitäisi olla vähintään 2,00 kg. Otostutkimus antaa seuraavat painot:

8 kpl 1,90 kg, 10 kpl 1,95 kg, 12 kpl 1,98 kg, 4 kpl 2,05 kg.

Kuinka paljon vähintään tuotteet ovat tämän otoksen perusteella alipainoisia, jos käytetään 95 % varmuutta?

1.8 Pakkauksen keskipainoksi ilmoitettiin 0,700 kg. Tutkituista 10 tuotteesta 6 painoi 692 g ja 4 painoi 701 g. Onko ilmoitettu keskipaino hyväksyttyjen rajojen välissä, jos käytetään 99 % varmuutta?

1.9 Auton erään osan tms. kulutuskestävyys arvioitiin 25 000 km:ksi ja keskihajonta 5000 km:ksi. Keskimääräisen kulutuskestävyyden tutkimiseksi päätettiin tehdä otos. Kuinka suureksi otoskoko on valittava, jotta otoskeskiarvo ei poikkeaisi todellisesta enempää kuin 5% 95%:n varmuudella?

2 Diskreetit ja jatkuvat todennäköisyysjakaumat

2.1 Satunnaismuuttuja ja pistetodennäköisyydet

Esim. 1 Viidessä kortissa ovat luvut 1, 2, 3, 4 ja 5 (yksi kussakin). Vedetään näistä korteista umpimähkään yksi. Tutkitaan mahdollisuutta saada tulokseksi esim. parillinen luku?

Kaikkien tulosten joukko on $E = \{1,2,3,4,5\}$. Tapausta "saadaan parillinen luku" vastaa osajoukko $A = \{2,4\}$. Koska kaikki viisi tulosta ovat yhtä mahdollisia ja A:ssa niistä on kaksi, niin A:n ***todennäköisyys*** (Probability) on

$$P(A) = \frac{2}{5} (= 0{,}4 = 40\%).$$

Esitetään tälle tehtävälle toisenlainenkin matemaattinen malli, joka sopii myös sellaisiin tapauksiin, missä tulokset eivät ole yhtä mahdollisia. Olkoon X ***satunnaismuuttuja***, joka ilmoittaa saadun pistemäärän. X:n arvot ja niiden todennäköisyydet (ns. ***pistetodennäköisyydet***) ovat

$$\begin{array}{ll} x_i: & 1,\ 2,\ 3,\ 4,\ 5 \\ p_i: & \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5} \end{array} \qquad (i = 1, 2, ..., 5).$$

Todennäköisyys, että X:n arvo on parillinen, on vastaavien pistetodennäköisyyksien summa:

$$P(X = \text{ parillinen}) = p_2 + p_4 = \frac{2}{5}.$$

Kaikkien pistetodennäköisyyksien summa on 1, ts. $\boxed{\sum p_i = 1}$.

Esim. 2 Neljästä kortista, joissa ovat luvut 1, 2, 3 ja 4, nostetaan umpimähkään yksi ja palautetaan se. Sitten nostetaan umpimähkään toinen kortti, joka voi siis olla myös sama kuin ensin nostettu. Olkoon X satunnaismuuttuja, joka ilmoittaa saadun kahden luvun summan. Viereisestä yhteenlaskutaulusta nähdään mahdolliset summan arvot ja kuinka monella eri tavalla kukin summa saadaan. Esim. 4 saadaan summina 3+1, 2+2 ja 1+3, ts.

+	1	2	3	4
1	2	3	4	5
2	3	4	5	6
3	4	5	6	7
4	5	6	7	8

kolmella korttiyhdelmällä 16:sta mahdollisesta korttiyhdelmästä. Siten todennäköisyys saada summa 4 on 3/16. Satunnaismuuttujan X ***jakauman*** muodostavat X:n arvot ja niiden todennäköisyydet:

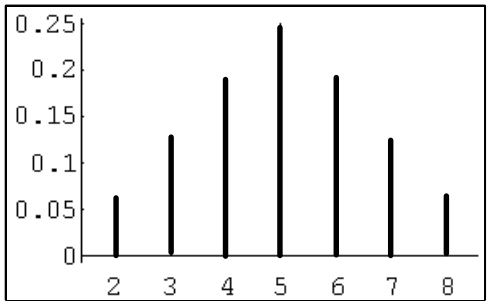
$$\begin{array}{ll} x_i: & 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8 \\ p_i: & \frac{1}{16}, \frac{2}{16}, \frac{3}{16}, \frac{4}{16}, \frac{3}{16}, \frac{2}{16}, \frac{1}{16} \end{array} .$$

Tässä esimerkissä kaikki pistesummat eivät enää ole yhtä mahdollisia. Todennäköisyys, että pistesumma on ainakin 7 eli että X saa arvon joka on 7 tai enemmän, on vastaavien pistetodennäköisyyksien summa:

$$P(X \geq 7) = \frac{2}{16} + \frac{1}{16} = \frac{3}{16}.$$

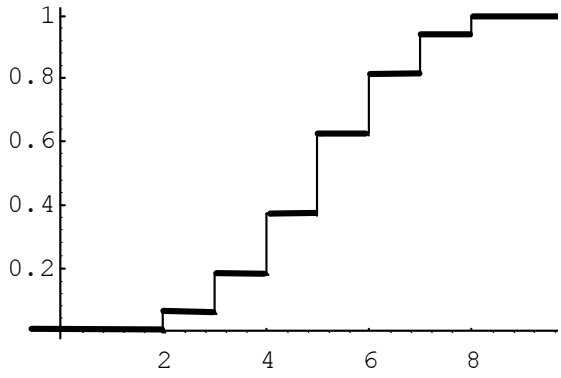
Koska pistetodennäköisyyksien summa on = 1, niin *komplementtitapauksen* "pistesumma on pienempi kuin 7 " todennäköisyys voidaan laskea seuraavasti:

$$P(X < 7) = 1 - \frac{3}{16} = \frac{13}{16}.$$



Koska X saa vain erillisiä arvoja, sen jakauma on ***diskreetti*** ja sitä kuvataan tavallisesti janadiagrammilla (viereinen kuva). Joskus on hyödyllistä ajatella todennäköisyys ***massaksi***, jota on kaikkiaan 1 yksikkö. Tässä esimerkissä todennäköisyysmassa on jakautunut lukujen 2, …, 8 kohdalle, symmetrisesti kohtaan 5 nähden.

Satunnaismuuttujan X ***kertymäfunktion*** F arvo kohdassa x saadaan, kun lasketaan yhteen kaikki kohtaan x mennessä kertynyt todennäköisyysmassa. Esim.



$$F(3) = \frac{1}{16} + \frac{2}{16} = \frac{3}{16}.$$

Teoreettisemmin sanottuna F voidaan määritellä yhtälöllä $F(x) = P(X \le x)$.

Kertymäfunktion kuvaaja on *porrasfunktio*, jonka pienin arvo on 0 (kun $x < 2$) ja suurin 1 (kun $x \ge 8$).

2.2 Diskreetin jakauman odotusarvo ja varianssi

Esim. 3 Lukujen 5, 6, 6, 6, 7, 7 keskiarvo

$$\bar{x} = \frac{5 + 3 \cdot 6 + 2 \cdot 7}{6} = \frac{1}{6} \cdot 5 + \frac{3}{6} \cdot 6 + \frac{2}{6} \cdot 7 \approx 6{,}17.$$

Tässä esiintyvät kertoimet ovat samat kuin pistetodennäköisyyksien arvot, jos ajatellaan luvuista valituiksi umpimähkään yksi ja satunnaismuuttuja X ilmoittaa saadun luvun. Siis

$$\bar{x} = \sum p_i x_i \approx 6{,}17.$$

Tätä summaa sanotaan todennäköisyyslaskennassa ***odotusarvoksi*** *(Expected value)*.

Määritelmä. *Jos satunnaismuuttuja saa arvot* $x_1, x_2, \ldots$ *ja näiden todennäköisyydet ovat* $p_1, p_2, \ldots$, *niin X:n* ***odotusarvo***

$$\boldsymbol{\mu = E(X) = \sum p_i x_i}$$

ja ***varianssi***

$$\boldsymbol{\sigma^2 = Var(X) = \sum p_i \cdot (x_i - \mu)^2}.$$

Varianssin neliöjuurta σ *sanotaan* ***keskihajonnaksi***.

Esim. 4 Jos X ilmoittaa yhdessä nopanheitossa saadun tuloksen, niin

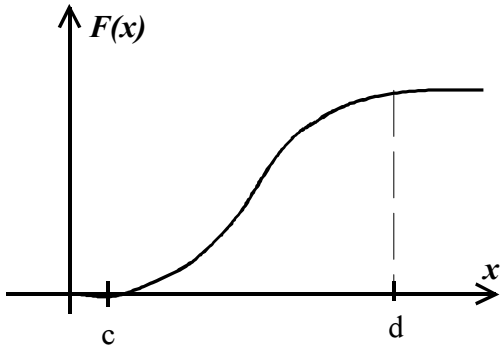
$$E(X) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3\frac{1}{2}$$

ja $\quad Var(x) = \frac{1}{6} \cdot (2{,}5^2 + 1{,}5^2 + 0{,}5^2 + 0{,}5^2 + 1{,}5^2 + 2{,}5^2) \approx 2{,}92.$

Jos suoritetaan useita heittoja peräkkäin, on odotettavissa että saatujen pistemäärien keskiarvo on $\approx 3{,}5$; sitä varmemmin, mitä enemmän heittoja on. Keskihajonta $\sigma = \sqrt{Var(X)} \approx 1{,}70$.

2.3 Jatkuvat jakaumat

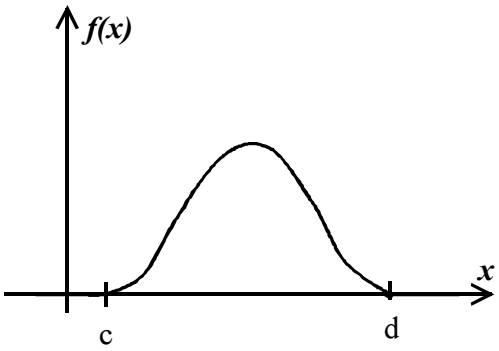
Painon, pituuden, ajan tms. mittauksissa kokeen tulos voi usein olla mikä tahansa reaaliluku jollakin välillä [*c*,*d*], mahdollisesti koko reaalilukualueella. Jos satunnaismuuttuja X ilmoittaa tällaisen kokeen tuloksen, X:n jakauma on ***jatkuva jakauma***. Tällöin X:n arvot x muuttuvat jatkuvasti, ilman hyppäyksiä ja siten kertymäfunktion F kuvaaja on nouseva, katkeamaton käyrä (sillä x:n kasvaessa todennäköisyysmassaa kertyy jatkuvasti).



Siinä kohdassa, missä kertymäfunktio $F(x)$ kasvaa jyrkimmin eli missä derivaatta $F'(x)$ on suurimmillaan, todennäköisyysmassan lisäys yksikköä kohti eli ns. *todennäköisyystiheys* on suurimmillaan. Samalla tavoin muissakin kohdissa kertymäfunktion derivaatta kuvaa todennäköisyystiheyttä tässä kohdassa. Siksi on luonnollista asettaa seuraava määritelmä:

Määritelmä. *Jatkuvan satunnaismuuttujan* X ***tiheysfunktio*** f täyttää ehdon

$$\boxed{f(x) = F'(x)}.$$



Todennäköisyys sille, että satunnaismuuttujan X arvo on jollakin välillä $a \ldots b$, saadaan kertymäfunktion avulla seuraavasti:

(1) $$\boxed{P(a < X \le b) = F(b) - F(a)},$$

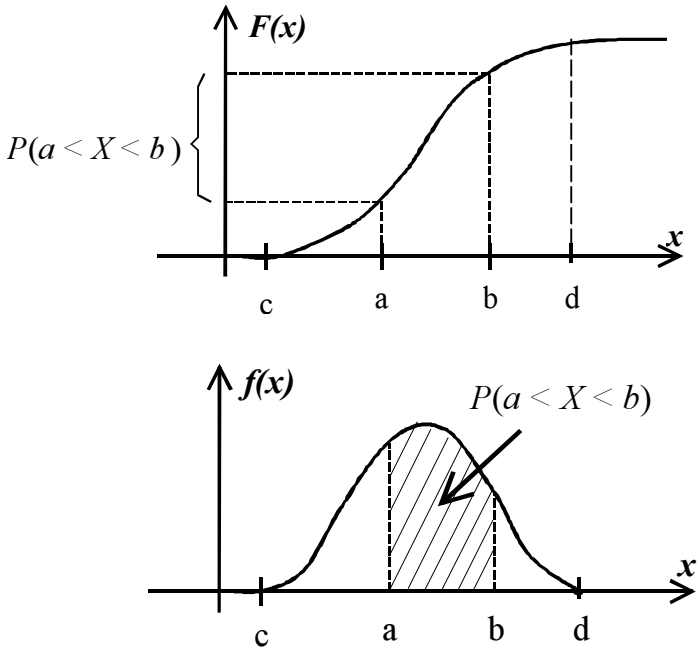
ts. laskemalla kohtaan b mennessä kertynyt todennäköisyysmassan määrä ja vähentämällä siitä kohtaan a mennessä kertynyt määrä. Sama todennäköisyys voidaan laskea myös tiheysfunktion avulla integraalina

(2) $$\boxed{P(a < X \le b) = \int_a^b f(x)\,dx},$$

sillä

$$\int_a^b f(x)\,dx = \Big|_a^b F(x) = F(b) - F(a) = P(a < X \le b).$$

Tulokset (1) ja (2) ovat graafisesti esitettyinä seuraavat:



Jatkuvalla jakaumalla esim. todennäköisyys, että X saa arvon b on $P(X = b) = 0$, sillä muuten kertymäkäyrässä olisi hyppäys kohdassa b. Siksi esim. välien $a < X < b$ ja $a < X \leq b$ todennäköisyydet ovat yhtä suuret. Yleisesti jatkuvalla jakaumalla yksittäisten pisteiden todennäköisyyksillä ei ole mitään käyttöä, koska ne ovat kaikki = 0. Niiden tilalle tuleekin "äärettömän lyhyelle välille dx sijoittuneen todennäköisyysmassan määrä $f(x)\,dx$.

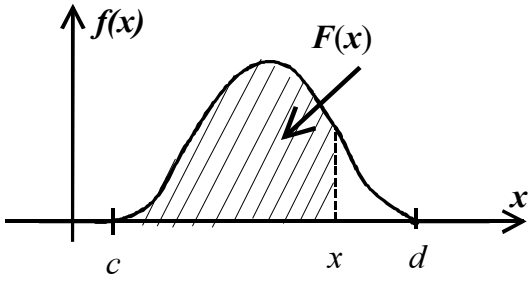
Jos X:n arvot muodostavat välin c … d (kuten edellisissä kuvissa), niin koko tälle välille sijoittuneen todennäköisyysmassan määrän tulee olla = 1 eli

$$\boxed{\int_c^d f(x)\,dx = 1}.$$

Diskreetillä jakaumalla vastaava tulos oli $\sum p_i = 1$.

Kertymäfunktion arvo kohdassa x saadaan "laskemalla yhteen" kohtaan x mennessä kertynyt todennäköisyysmassa:

$$\boxed{F(x) = \int_c^x f(x)\,dx}.$$



Esim. *Mathematica*-ohjelmassa tiheysfunktiolla on nimi *PDF = Probability Distribution Function* (= jakaumafunktio). Kertymäfunktiolla taas on nimi *CDF = Cumulative Distribution Function* (= kumulatiivinen jakaumafunktio).

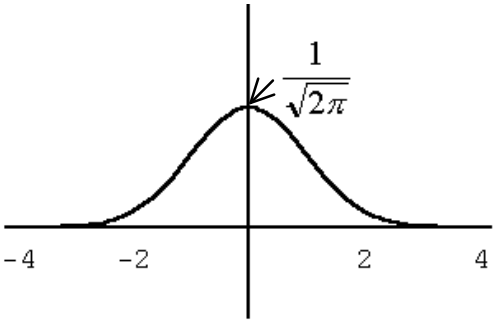
Jatkuvalla jakaumalla ***odotusarvo*** ja ***varianssi*** määritellään seuraavasti (jos X:n arvot sijoittuvat välille $c \ldots d$):

$$\mu = E(X) = \int_c^d x\, f(x)\, dx, \qquad \sigma^2 = \int_c^d (x-\mu)^2 f(x)\, dx.$$

Keskihajonta on varianssin neliöjuuri ja sitä merkitään σ:lla.

2.2 Normaalijakauma

1) Esitellään aluksi ***standardoitua normaalijakaumaa***. Sen tiheysfunktiota merkitään f:n sijasta φ:llä.



Tiheysfunktion φ arvot saadaan yhtälöstä

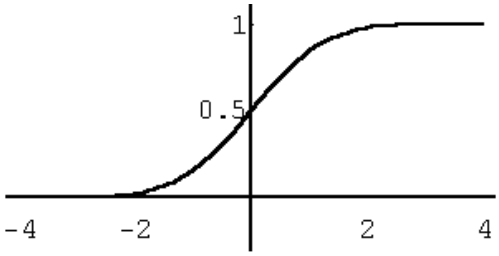
$$\boxed{\varphi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}},$$

missä x saa kaikki reaalilukuarvot. Kuvaaja on y-akselin suhteen symmetrinen, joten odotusarvo $\mu = 0$. Lisäksi keskihajonta $\sigma = 1$ kuten voidaan todistaa.

Tiheyskäyrä on tyyppiä $y = e^{-x^2}$, johon on eksponenttiin ja lausekkeen eteen lisätty sellaiset kertoimet, että kokonaispinta-ala eli todennäköisyysmassan määrä on = 1 ja keskihajonta on = 1.

Kertymäfunktiota merkitään vastaavalla isolla kirjaimella:

$$\Phi(x) = \int_{-\infty}^{x} \varphi(x)\, dx.$$



Kertymäfunktion arvot on taulukoitu,

koska integraalin laskemisessa jouduttaisiin käyttämään likiarvomenetelmiä. Eräs tällainen taulukko on kopioitu tämän monisteen loppuun.

Esim. 1 $\Phi(0) = \frac{1}{2}$ (symmetrian nojalla)

$\Phi(1) = 0{,}841$ (taulukko, laskin tms.)

$\Phi(2) = 0{,}977$ (- " -)

Edelleen, jos X on satunnaismuuttuja, jonka jakauma on standardinormaali, niin

$$P(1 \le X \le 2) = \Phi(2) - \Phi(1) = 0{,}136$$

$$P(X > 2) = 1 - P(X \le 2)$$

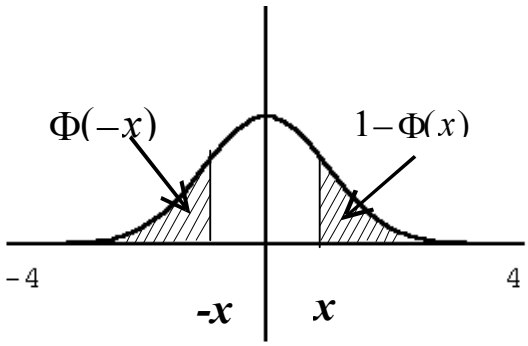
$$= 1 - \Phi(2) = 0{,}023$$

Symmetrian nojalla

$$\boxed{\Phi(-x) = 1 - \Phi(x)},$$

vrt. viereinen kuva. Esim.

$\Phi(-2) = 1 - \Phi(2) = 0{,}023$ ja



$$P(-1 \le X \le 1) = \Phi(1) - \Phi(-1)$$

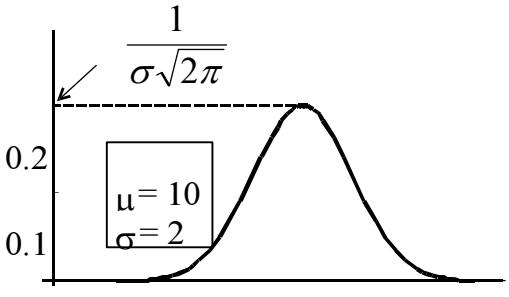
$$= \Phi(1) - [1 - \Phi(1)]$$

$$= 2 \cdot \Phi(1) - 1 = 0{,}682.$$

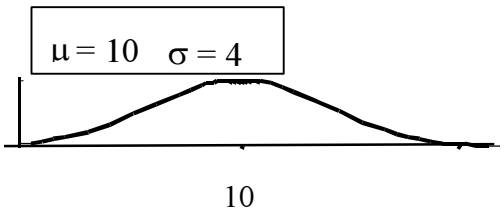
2) ***Yleinen normaalijakauma.*** Sen tiheysfunktio on

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

missä μ ja σ ovat jakauman parametrit (*odotusarvo* ja *keskihajonta*). Kuvaaja on kohtaan μ nähden symmetrinen.



Kuvaaja on sitä "laakeampi", mitä suurempi keskihajonta σ on.



Kertymäfunktion $F(x)$ arvojen

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

laskeminen palautetaan standardinormaalin jakauman kertymäfunktion arvojen laskemiseen seuraavan tuloksen avulla (tod. luvussa 8):

$$\boxed{F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)}.$$

Esim. 2 Oletetaan, että satunnaismuuttujan X jakauma on normaali, parametreina $\mu = 75$ ja $\sigma = 12$. Tämä merkitään lyhyesti $\boldsymbol{X \sim N(75,\ 12^2)}$. Vastaava yleinen merkintä on

$$\boldsymbol{X \sim N(\mu,\ \sigma^2)}$$

tai joskus $X \sim N(\mu,\ \sigma)$. Esim. todennäköisyys, että X:n arvo on korkeintaan 87 voidaan laskea seuraavasti:

$$P(X \le 87) = F(87) = \Phi\left(\frac{87-75}{12}\right) = \Phi(1) \approx 0{,}84.$$

Tämä jakauma voisi olla matemaattisena mallina suuren ihmisjoukon painojakaumalle, jos painojen keskiarvo on 75 kg ja hajonta 12 kg. Tuloksen mukaan korkeintaan 87 kg painavia on tässä joukossa n. 84 %.

Esim. 3 Laske todennäköisyys sille, että $(\mu,\ \sigma)$-normaalisen satunnaismuuttujan X arvo poikkeaa odotusarvostaan μ korkeintaan 1,96 kertaa keskihajonnan verran.

$$\begin{aligned}
&P(\mu - 1{,}96\cdot\sigma \le X \le \mu + 1{,}96\cdot\sigma)\\
&= F(\mu + 1{,}96\cdot\sigma) - F(\mu - 1{,}96\cdot\sigma)\\
&= \Phi\left(\frac{\mu + 1{,}96\cdot\sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - 1{,}96\cdot\sigma - \mu}{\sigma}\right)\\
&= \Phi(1{,}96) - \Phi(-1{,}96)\\
&= \Phi(1{,}96) - [1 - \Phi(1{,}96)]\\
&= 2\cdot\Phi(1{,}96) - 1\\
&= 2\cdot 0{,}9750 - 1 = 0{,}950.
\end{aligned}$$

Siis 95 % X:n arvoista poikkeaa odotusarvostaan korkeintaan 1,96 kertaa keskihajonnan verran. Toisin sanoen 95 % X:n arvoista on välillä $\mu \pm 1{,}96 \cdot \sigma$, jos X:n jakauma on normaali. Näin on perusteltu 1. luvun Lauseen1 toinen kohta.

Harjoituksia

A, B

2.1 Heität kerran noppaa ja markan rahaa. Rahan numeropuoli antaa pistemäärän 1 ja vastakkainen puoli pistemäärän 3. Satunnaismuuttuja X ilmoittaa saadun pistesumman. a) Määritä X:n saamat arvot ja niiden pistetodennäköisyydet. b) Piirrä jakauman graafinen esitys, c) Määritä X:n kertymäfunktio ja piirrä sen kuvaaja. d) Laske X:n odotusarvo μ ja keskihajonta σ.

2.2 Rahanheitossa numeropuoli on *klaava* ja toinen puoli *kruuna* (tai kruunu). Satunnaismuuttuja X ilmoittaa kruunien lukumäärän kolmessa heitossa. Määritä X:n jakauma ja kertymäfunktio.

2.3 Laske edellisen tehtävän mukaisen satunnaismuuttujan odotusarvo ja varianssi.

2.4 Sadasta arvasta yksi voittaa 200 euroa, 3 voittaa 50 euroa ja 8 voittaa 10 euroa. Laske voiton odotusarvo.

2.5 Satunnaismuuttuja X saa arvot 1,2,3,4,5 ja näiden todennäköisyydet ovat $5/k$, $4/k$, $3/k$, $2/k$, $1/k$. Laske k, μ ja σ.

2.6 a) Määritä sellainen k:n arvo, että $f(x) = \begin{cases} k \sin x, \text{ kun } 0 \le x \le \pi \\ 0 \text{ muulloin} \end{cases}$ käy jonkin jatkuvan satunnaismuuttujan X tiheysfunktioksi. Laske, millä todennäköisyydellä X:n arvo on b) korkeintaan 1, c) yli 1. d) Määritä vastaava kertymäfunktio. e) Laske kohdat b) ja c) kertymäfunktion avulla.

2.7 Olkoon $f(x) = \begin{cases} c / x^2, \text{ kun } x \ge 1 \\ 0, \text{ kun } x < 1 \end{cases}$. a) Laske, millä c:n arvolla $f(x)$ käy tiheysfunktioksi? b) Määritä sellainen luvun a arvo, että $P(x > a) = 1/3$.

2.8 Olkoon X:n jakauma normaali, parametreina $\mu = 1$ ja $\sigma = 2$. Laske $P(-3 < X < 4)$.

2.9 Oletetaan, että suomalaisten miesten pituusjakauma on (lähes) normaali, parametreina $\mu = 175$ cm ja $\sigma = 8$ cm. Kuinka monta % miehistä on yli 2 m pitkiä?

2.10 Olkoon $X \sim \mathrm{N}(0,1)$.

a) Todista, että $P(|X| > k) = 2 \cdot (1 - \Phi(k))$

b) Määritä sellainen k:n arvo, että $P(|X| > k) = 0{,}371$.

2.11 Tuotteen paino noudattaa normaalijakaumaa, jonka odotusarvo on 80 kg ja hajonta 4 kg.

a) Suuriko osa tuotteista on 77…80 kg painoisia?

b) Millä todennäköisyydellä umpimähkään valitun tuotteen paino on yli 88 kg ?

2.12 Olkoon $X \sim \mathrm{N}(120,\ 3)$ ($\therefore\ \sigma = 3$). Laske a) $P(X < 123)$, b) $P(X > 118)$, c) $P(119 < X < 122)$.

2.13 Sadan tuotteen otannassa saatiin keskimääräiseksi alkoholipitoisuudeksi 2,15 % ja hajonnaksi 0,05 %. Kuinka suuri osa tuotteista ylittää tälle tuotteelle sallitun pitoisuuden 2,25 % ?

2.14 Tehtaan valmistamien tuotteiden lujuuden tulisi olla (vähintään) 3000 N/cm^2. Tuotteista otettiin näytteitä aina silloin tällöin. Tämän perusteella arvioitiin lujuuden odotusarvon (keskiarvon) olevan 3700 N/cm^2 ja keskihajonnan 400 N/cm^2. Kuinka suuri osuus tuotteista täyttää lujuusvaatimuksen?

2.15 Levyistä hylätään alle 34 mm paksuiset ja yli 35 mm paksuiset. Mitattiin 500 levyä ja niistä 30 todettiin olevan liian ohuen ja 90 liian paksun. Laske levyjen paksuuden keskiarvo ja keskihajonta (olettaen, että paksuus on normaalijakautunut). Ohje: Yhtälöistä

$$P(X < 34) = \frac{30}{500} \quad \text{ja} \quad P(X > 35) = \frac{90}{500}$$

saat suureille μ ja σ yhtälöparin, kun esität ne Φ:n avulla. Monisteessa olevan taulukon käyttämisessä "takaperin" tarvitset sääntöä $\Phi(-x) = 1 - \Phi(x)$.

3 Todennäköisyyskäsitteistä

3.1 Joukko-opin käsitteistöä (kertaus)

Esimerkiksi joukkojen $A = \{2,3,4,5\}$ ja $B = \{4,5,6,7\}$

- ***yhdiste*** eli ***unioni*** muodostuu A:han tai B:hen (tai molempiin) kuuluvista alkioista: $A \cup B = \{2,3,4,5,6,7\}$.
- ***leikkaus*** muodostuu A:han ja B:hen kuuluvista alkioista eli näiden joukkojen yhteisistä alkioista: $A \cap B = \{4,5\}$.

Jos lisäksi on annettu jokin ***perusjoukko***, esim. $E = \{1,2,...,10\}$, voidaan puhua esim. joukon A ***komplementista*** (tämän perusjoukon suhteen). Se muodostuu niistä perusjoukon alkioista, jotka eivät kuulu A:han. A:n komplementtiä merkitään $\overline{A}$:lla. Tässä esimerkissä

- A:n ***komplementti*** $\overline{A} = \{6,7,8,9,10\}$.

Joukon A ***alkioiden lukumäärää*** eli ***kertalukua*** merkitään $n(A)$:lla. Tässä esimerkissä

$$n(A) = n(B) = 4.$$

Yhdisteen $A \cup B$ alkiomäärä saadaan kun lasketaan yhteen A:n ja B:n alkiomäärät ja summasta vähennetään A:n ja B:n yhteiset alkiot, jotka muuten tulisivat lasketuiksi mukaan kahdesti. Siis

(1) $$n(A \cup B) = n(A) + n(B) - n(A \cap B).$$

Tässä esimerkissä $n(A \cup B) = 4 + 4 - 2 = 6$. ***Tyhjää joukkoa*** merkitään $\varnothing$:llä.

3.2 Klassinen todennäköisyys

Lähtökohta: Kokeessa on vain *äärellinen määrä* tuloksia ja ne ovat *yhtä mahdollisia*. Joistakin tai kaikista tuloksista muodostuvia joukkoja sanotaan ***tapauksiksi***. Kaikkien tulosten joukko $E = \{e_1, e_2, ..., e_n\}$ on ns. ***varma*** tapaus.

Määritelmä: Jos kokeessa on n yhtä mahdollista tulosta, niin tapauksen A ***todennäköisyys*** on

$$\boxed{P(A) = \frac{n(A)}{n}}$$

Esim. 1 Nopanheitossa varma tapaus on $E = \{1,2,3,4,5,6\}$ ja eräs tapaus on se, että "tulos on *ainakin* 3 (eli *vähintään* 3)". Tämä tapaus on E:n osajoukko $A = \{3,4,5,6\}$ ja sen todennäköisyys on

$$P(A) = \frac{4}{6} = \frac{2}{3} \approx 66{,}7\ \%\,.$$

A:n *komplementtitapaus* (vastakohta) $\overline{A}$ on se, että tulos on *korkeintaan* 2, siis $\overline{A} = \{1,2\}$ ja sen todennäköisyys on

$$P(\overline{A}) = \frac{2}{6} = \frac{1}{3}, \quad \text{myös} \quad P(\overline{A}) = 1 - P(A) = \frac{1}{3} \approx 33{,}3\ \%\,.$$

3.3 Todennäköisyyden perusominaisuuksia

Lause 1 1) $\boldsymbol{P(A) \geq 0}$ eli jokaisen tapauksen todennäköisyys on ≥ 0.

2) $\boldsymbol{P(E) = 1}$ eli varman tapauksen todennäköisyys on 1.

3) $\boldsymbol{P(\varnothing) = 0}$ eli mahdottoman tapauksen todennäköisyys on 0.

4) $\boxed{\boldsymbol{P(\overline{A}) = 1 - P(A)}}$ eli komplementtitapauksen todennäköisyys on 1 miinus alkuperäisen tapauksen todennäköisyys.

5) ***yhteenlaskusääntö:*** $\boxed{\boldsymbol{P(A \cup B) = P(A) + P(B) - P(A \cap B)}}$

6) ***additiivisuus***: $\boxed{\boldsymbol{P(A \cup B) = P(A) + P(B),\ \text{jos}\ A \cap B = \varnothing}}$.

Kohtien 1) - 4) tulokset todistetaan klassisen todennäköisyyden määritelmän avulla seuraavasti:

$$P(A) = \frac{n(A)}{n} \geq 0, \quad P(E) = \frac{n}{n} = 1, \quad P(\varnothing) = \frac{0}{n} = 0,$$

Merkitään A:n alkiomäärää k:lla jolloin $\overline{A}$:ssa on $n - k$ alkiota ja

$$P(\overline{A}) = \frac{n-k}{n} = 1 - \frac{k}{n} = 1 - P(A)$$

Yhteenlaskusäännön mukaan todennäköisyys sille että ***A tai B*** tapahtuu (ts. että kokeen tulos kuuluu A:han tai B:hen tai molempiin) saadaan, kun lasketaan yhteen A:n ja B:n todennäköisyydet ja summasta vähennetään todennäköisyys sille, että ***A ja B*** tapahtuvat. Tämä tulos todistetaan

(klassisen todennäköisyyden tilanteessa) kohdan 3.1 lopussa mainitun tuloksen avulla seuraavasti:

$$P(A \cup B) = \frac{n(A \cup B)}{n} = \frac{n(A) + n(B) - n(A \cap B)}{n}$$
$$= \frac{n(A)}{n} + \frac{n(B)}{n} - \frac{n(A \cap B)}{n}$$
$$= P(A) + P(B) - P(A \cap B).$$

Additiivisuus taas seuraa välittömästi yhteenlaskusäännöstä ja kohdasta 3). Sen mukaan todennäköisyys, että *A tai B* tapahtuu on = *A*:n todennäköisyys + *B*:n todennäköisyys, mikäli tapaukset *A* ja *B* ovat ***erillisiä***, ts. ne eivät sisällä yhteisiä tuloksia.

Komplementtitapausta koskeva tulos $\boldsymbol{P(\overline{A}) = 1 - P(A)}$ voidaan esittää myös muodossa "todennäköisyys, että *A* ***ei tapahdu*** (ts. että tuloksena ei ole mikään *A*:han kuuluva tulos) on = 1 – *A*:n todennäköisyys".

Esim. 2 Nopanheitossa tapaus *"korkeintaan 2"* on $A = \{1,2\}$ ja *tapaus "ainakin 4"* on $B = \{4,5,6\}$. Nämä tapaukset ovat *erillisiä* ($A \cap B = \varnothing$). Täten additiivisuuden nojalla todennäköisyys, että saadaan *"korkeintaan 2 tai vähintään 4"* on

$$P(A \cup B) = \frac{2}{6} + \frac{3}{6} = \frac{5}{6}.$$

Tapaukset $C = \{2,4,6\}$ (*"saadaan parillinen tulos"*) ja $D = \{3,4,5,6\}$ (*"saadaan vähintään 3"*) taas eivät ole erillisiä vaan $C \cap D = \{4,6\}$. Todennäköisyyden, että saadaan *"parillinen tulos tai vähintään 3"* laskemiseen käy yhteenlaskusääntö:

$$P(C \cup D) = P(C) + P(D) - P(C \cap D) = \frac{3}{6} + \frac{4}{6} - \frac{2}{6} = \frac{5}{6}.$$

(Tarkistus: $C \cup D = \{2,3,4,5,6\}$, joten $P(C \cup D) = \frac{5}{6}$.)

Esim. 3 Todennäköisyys, että korttipakasta nostettu kortti on *pata* tai *ässä* on yhteenlaskusäännön nojalla

$$P(\textit{pata tai ässä}) = P(\textit{pata}) + P(\textit{ässä}) - P(\textit{pata ja ässä})$$
$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}.$$

3.4 Pistetodennäköisyydet

Kaikkien tulosten joukon $E = \{e_1, e_2, ..., e_n\}$ osajoukkoja sanottiin *tapauksiksi* ja esim. tapaus $A = \{e_1, e_2\}$ vastaa sitä, että kokeen tuloksena on e_1 tai e_2. Yksialkioisia osajoukkoja $\{e_1\}, \{e_2\}, ..., \{e_n\}$ sanotaan ***alkeistapauksiksi*** ja niiden todennäköisyyksiä $p_1, p_2, ..., p_n$ sanotaan ***pistetodennäköisyyksiksi***.

Klassista todennäköisyyskäsitettä voidaan käyttää vain, jos jokaisen tuloksen todennäköisyys eli jokainen pistetodennäköisyys p_i on yhtä suuri. *Pistetodennäköisyyksiä voidaan käyttää yleisemminkin*. Niiden ei tarvitse olla yhtä suuria, kunhan ne vain täyttävät seuraavat ehdot:

(1) *jokainen* $p_i \geq 0$,

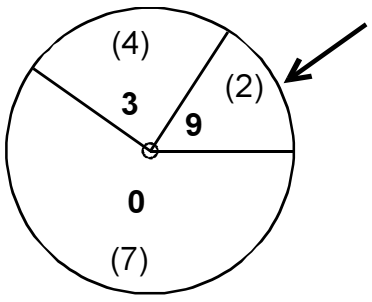
(2) $\sum p_i = 1$,

(3) *tapauksen A todennäköisyys on vastaavien pistetodennäköisyyksien summa*, ts. jos esimerkiksi

$$A = \{e_1, e_3, e_7\}, \text{ niin } P(A) = p_1 + p_3 + p_7 .$$

On helppo todistaa, että Lauseessa 1 mainitut 6 tulosta ovat edelleenkin voimassa ja siten käytettävissä.

Esim. 4 Onnenpyörää, joka on jaettu 3 sektoriin suhteessa 2 : 4 : 7, pyöräytetään kerran. Kokeessa on 3 mahdollista tulosta $x_1 = 9, x_2 = 3$ ja $x_3 = 0$ (kuva). Koska $2 + 4 + 7 = 13$, eri sektorien todennäköisyydet ovat



$$p_1 = \frac{2}{13}, p_2 = \frac{4}{13}, p_3 = \frac{7}{13} \qquad \left(\sum p_i = \frac{13}{13} = 1 \right)$$

(jos pyörittämisessä ei ole häiriöitä). Todennäköisyys, että yhdessä pyöräytyksessä saadaan tulos x_1 tai x_2 on $p_1 + p_2 = 6/13$. Todennäköisyys, että saadaan tulos $x_3 = 0$, on 7/13 ja todennäköisyys, että ei saada tulosta $x_3 = 0$, on $1 - 7/13 = 6/13$ (komplementtitapaus).

Tässä esimerkissä kokeelle saatiin *todennäköisyysmalli*, käyttämällä apuna onnenpyörän *geometriaa*: pistetodennäköisyydet valittiin sektorien koon perusteella.

Esim. 5 (*Tilastollinen todennäköisyys*). Jos onnenpyörän pysähtymisessä epäilee olevan "vilppiä", on turvauduttava tilastolliseen menetelmään, useaan pyöräytykseen. Jos näissä n:ssä pyöräytyksessä esim. tulos $x_3 = 0$ esiintyy f_n kertaa, niin suhteellisen frekvenssin f_n / n tulisi n:n kasvaessa "lähestyä" likiarvoltaan lukua $p_3 = 7/13 \approx 0{,}538$.

Lähestymiseen liittyy kuitenkin tietty tilastollinen epävarmuus (toisin kuin tavalliseen raja-arvoon). Voihan nimittäin käydä niin, että esim 100 pyöräytyksessä saadaan joka kerta sattumalta tulos 3, vaikka onnenpyörä olisi kunnossakin. Tällaisen poikkeuksellisen tuloksen mahdollisuuden pitäisi kuitenkin käydä hyvin pieneksi, jos toistomäärää kasvatetaan tai tehdään useita pienempiä koesarjoja.

Harjoituksia

A, B

3.1 Joukossa A on 27 alkiota ja joukossa B 34 alkiota. Yhteensä alkioita on 44. Laske $n(A \cap B)$.

3.2 Pussissa on 23 punaista, 15 mustaa ja 18 valkoista palloa. Näistä otetaan umpimähkään yksi. Laske todennäköisyys, että otettu pallo a) on punainen tai musta, b) ei ole musta. Käytä laskuissa Lauseen 1 merkintätapoja. (A = pun. pallojen joukko jne).

3.3 Noppaa heitetään kahdesti. Laske todennäköisyys, että saatu summa on a) 3 tai 9, b) korkeintaan 4, c) yli 4. (Tee yhteenlaskutaulu samaan tapaan kuin 2. luvun esimerkissä 2.)

3.4 Laske todennäköisyys, että kahden nopan heitossa pistesumma on a) alle 4, b) yli 10, c) vähintään 4, d) korkeintaan 10, e) alle 4 tai yli 10, f) välillä [4,10].

3.5 Opiskelijoista sai ala-arvoisen matematiikan kokeissa 15%, fysiikan kokeissa 12% ja molemmissa 7%. Millä todennäköisyydellä

satunnaisesti valittu opiskelija hylättiin ainakin toisessa näistä kokeista?

3.6 Luvuista 1, …, 100 valitaan umpimähkään yksi. Laske todennäköisyys, että tämä on jaollinen a) 3:lla, b) 5:llä c) 3:lla ja 5:llä (eli 15:llä), d) 3:lla tai 5:llä. Käytä apuna yhteenlaskusääntöä.

3.7 Tikka heitetään tauluun, joka muodostuu renkaista 1, …,10 ja jossa 10-ympyrän säde on sama kuin 1, …,9 -renkaiden leveydet. a) Laske pistetodennäköisyydet, olettaen että heitto osuu tauluun ja on täysin summittainen. b) Mikä on todennäköisyys, että heitossa saadaan pistemäärä 4, 5 tai 8? c) Entä todennäköisyys, että saadaan korkeintaan 8?

3.8 Valitaan yksi reaaliluku lukusuoran väliltä [0,3] ja toinen väliltä [−2,0] satunnaisesti. Laske todennäköisyys sille, että lukujen erotus on yli 3. (Ohje: Kaikki mahdolliset tapaukset voidaan esittää *xy*-tason alueena

$$E = \{(x, y) \,:\, 0 \le x \le 3, -2 \le y \le 0\}$$

ja lopputulokseltaan "suotuisat" tapaukset alueena

$$A = \{(x, y) \,:\, x - y > 3\} = \{(x, y) \,:\, y < x - 3\}.)$$

3.9 Lukusuoran väliltä [0,2] valitaan reaaliluku x satunnaisesti. Mikä on todennäköisyys, että sen 1. desimaali on 2 ja toinen on 3 (ts. että $0{,}23000... \le x \le 0{,}23999... = 0{,}24$ tai $1{,}2300... \le x \le 1{,}2399... = 1{,}24$).

4 Kokeiden yhdistäminen

2.1 Tuloperiaate

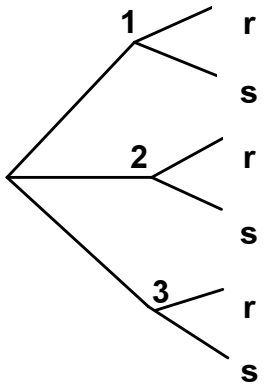
Joukkojen A ja B ***tulojoukko*** $\boldsymbol{A \times B}$ muodostuu kaikista sellaisista (järjestetyistä) pareista (a, b), missä $a \in A$ ja $b \in B$. Toisin sanoen

$$A \times B = \{(a,b) \,:\, a \in A, b \in B\}.$$

Esim. 1 Jos $A = \{1,2,3\}$ ja $B = \{r, s\}$, niin

$$A \times B = \{(1,r),(1,s),(2,r),(2,s),(3,r),(3,s)\}.$$

Tulojoukon alkiot (6 paria) voidaan havainnollisesti esittää ns. *puudiagrammin* avulla (viereinen kuva).



Tässä tulojoukossa on $3 \cdot 2 = 6$ paria, sillä jokaista A:n alkiota kohti löytyy 2 B:n alkiota, ja koska A:ssa on 3 alkiota, pareja on $3 \cdot 2$ kpl.

Samalla periaatteella päätellään yleisesti seuraava tulos:

Lause: Jos A:ssa on m alkiota ja B:ssä n alkiota, niin tulojoukossa $A \times B$ on mn alkiota (paria).

4.2 Kokeiden yhdistäminen

Tehdään kaksi koetta peräkkäin (esim. nopan ja rahanheitto tai kaksi nopanheittoa). Silloin ***todennäköisyys, että 1. kokeessa tapahtuu A ja 2. kokeessa B***, on

(1) $$\boxed{\boldsymbol{P(A \times B) = P(A) \cdot P(B)}}$$ (***kertosääntö***)

Perustellaan tätä klassisessa tapauksessa. Jos 1. kokeessa tapahtuu A ja 2. kokeessa B, niin tuloksena on jokin (a,b)-pari. Oletetaan, että 1. kokeessa on m tulosta ja niistä A:han kuuluu h kpl sekä 2. kokeessa on n tulosta ja niistä B:hen kuuluu k kpl. Tällöin (a,b)-pareja on hk kpl kaikista mn parista. Siten todennäköisyys, että $A \times B$ tapahtuu eli tuloksena on jokin (a,b)-pareista, on

$$P(A \times B) = \frac{hk}{mn} = \frac{h}{m} \cdot \frac{k}{n} = P(A) \cdot P(B).$$

Esim. 2 Heitetään noppaa kahdesti. Todennäköisyys, että 1. kerralla tulee 6 ja 2. kerralla 5 tai 6, on

$$\frac{1}{6} \cdot \frac{2}{6} = \frac{2}{36} = \frac{1}{18}.$$

Esim. 3 Pakasta vedetään 2 korttia. Laske todennäköisyys, että molemmat ovat patoja. Voidaan ajatella että kortit vedetään peräkkäin. Kun 1. kortti on pata, on toista vedettäessä jäljellä 12 pataa 51 kortista. Kertosäännön mukaan tulos on

$$P(\text{"1. on pata" ja " 2. on pata"}) = \frac{13}{52} \cdot \frac{12}{51} = \frac{1}{4} \cdot \frac{12}{51} = \frac{3}{51} = \frac{1}{17}.$$

Esim. 4 Suuressa tuote-erässä on 2 % virheellisiä.Tuotteista otetaan satunnaisesti kaksi. Laske todennäköisyys, että kumpikin on a) virheellinen, b) virheetön.

a) Todennäköisyys, että 1. on virheellinen, on 0,02. Koska tuote-erä on suuri, yhden virheellisen tuotteen poistuminen ei vaikuta sanottavasti tilanteeseen. Siten myös todennäköisyys, että 2. on virheellinen, on $0{,}02$. Kertosäännön mukaan todennäköisyys, että kumpikin on virheellinen, on

$$0{,}02 \cdot 0{,}02 = 0{,}0004 = 0{,}04\%.$$

b) Käyttämällä kertosääntöä edellisten tapausten komplementtitapauksiin, saadaan tulos

$$P(\text{"kumpikin virheetön"}) = 0{,}98 \cdot 0{,}98 \approx 96\%.$$

Esim. 5 Veikataan yksi rivi (13 ottelua) satunnaisesti. Todennäköisyys, että 1. ottelu on oikein, on 1/3. Todennäköisyys, että 1. ja 2. ovat oikein, on $1/3 \cdot 1/3$ jne. Todennäköisyys, että saadaan 13 oikein, on $(1/3)^{13} \approx 6{,}27 \cdot 10^{-7}$. Todennäköisyys, että ei saada yhtään oikein, on $(2/3)^{13} \approx 5{,}14 \cdot 10^{-3}$.

2.3 Kertosäännön ja additiivisuuden yhteiskäyttö

Yhteenlaskusääntö ja additiivisuus ovat "tai-sääntöjä", jotka koskevat yhden kokeen eri tapauksia. Esim. additiivisuus voidaan esittää seuraavassa muodossa:

$\boldsymbol{P(\text{"}A\text{ tai }B\text{ tapahtuu"}) = P(A) + P(B)}$, jos A ja B ovat *saman kokeen* kaksi erillistä tapausta (ts. ne eivät sisällä samoja tuloksia)

Kertosääntö ("ja-sääntö")taas koskee kahden kokeen suorittamista peräkkäin:

$\boldsymbol{P(\text{"}A\text{ ja }B\text{ tapahtuvat"}) = P(A) \cdot P(B)}$, jos A on 1. kokeen jokin tapaus ja B on 2. kokeen jokin tapaus

Näitä sääntöjä käytetään joskus yhdessä siihen tapaan kuin seuraavan esimerkin c)-kohta osoittaa.

Esim. 6 a) Yksi nopanheitto, $E = \{1,...,6\}$.

$$\begin{cases} A\text{: "saadaan 1 tai 2"} & A = \{1,2\} \\ B\text{: "saadaan 4, 5 tai 6"} & B = \{4,5,6\} \end{cases}$$

Todennäköisyys, että yhdessä kokeessa tapahtuu A tai B (ts. että tulos kuuluu joukkoon $A \cup B$) on

$$P(A \cup B) = P(A) + P(B), \text{ koska } A \cap B = \varnothing$$
$$= \frac{2}{6} + \frac{3}{6} = \frac{5}{6} .$$

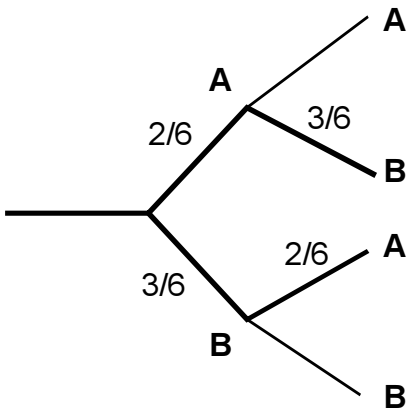
b) Kaksi nopanheittoa. A ja B kuten yllä. Todennäköisyys, että 1. kerralla tapahtuu A ja 2. kerralla B**,** on

$$P(A \times B) = \frac{2}{6} \cdot \frac{3}{6} = \frac{1}{6} .$$

c) Kaksi nopanheittoa yht'aikaa. Lasketaan todennäköisyys, että toisella heitolla tapahtuu A ja toisella B, ts. "1. kerralla A ja 2. kerralla B" <u>tai</u> "1. kerralla B ja 2. kerralla A" :

$$P(A \times B \cup B \times A) = \frac{2}{3} \cdot \frac{3}{6} + \frac{3}{6} \cdot \frac{2}{6} = 2 \cdot \frac{1}{6} = \frac{1}{3} .$$

Tilannetta voidaan havainnollistaa puudiagrammilla:



$$P(\text{"}A \text{ ja } B\text{" tai "}B \text{ ja } A\text{"}) = \frac{2}{6}\cdot\frac{3}{6}+\frac{3}{6}\cdot\frac{2}{6}.$$

Harjoituksia

A, B

4.1 Laske todennäköisyys, että

a) kahdessa nopanheitossa ei saada yhtään 6:ta,

b) vedettäessä kahdesta pakasta kummastakin yksi kortti, molemmat ovat ässiä.

4.2 Laske todennäköisyys, että kahdessa nopanheitossa saadaan

a) 1. kerralla 6 ja toisella ei,

b) tarkalleen yksi 6.

4.3 Mikä on todennäköisyys, että

a) lotottaessa yksi rivi saadaan kaikki 7 oikein (39 luvusta, ei lisänumeroita),

b) Vakioveikkauksessa veikattaessa 13 kohdetta umpimähkään saadaan yksi oikein (vaihtoehdot ovat 1, x, 2).

4.4 10 pallon joukossa on 3 mustaa. Palloista otetaan satunnaisesti kolme. Laske todennäköisyys, että

a) kaikki kolme ovat mustia,

b) mikään ei ole musta,

c) palloista kaksi on mustaa.

4.5 100 arvan joukossa on 10 voittoa. Ostat ensimmäisenä arpoja. Kaksi ensimmäistä arpaasi ei tuottanut voittoa. Laske todennäköisyys, että

a) kolmannella arvalla saat voiton,

b) kolmannella ja neljännellä saat voitot,

c) kolmannella tai neljännellä saat voiton, mutta et molemmilla.

4.6 Henkilöt A, B, ja C ampuvat maaliin. Heidän osumistodennäköisyytensä ovat vastaavasti 0,70, 0,90 ja 0,60. Kukin ampuu kerran. Laske seuraavat todennäköisyydet (vastaukset 2 numeron tarkkuudella):

a) kaikki saavat osuman,

b) A ja C osuvat, mutta B ei,

c) maaliin tulee tarkalleen kaksi osumaa,

d) maaliin tulee tarkalleen yksi osuma,

e) maaliin tulee korkeintaan yksi osuma.

4.7 a) Laatikossa on 3 valkoista ja 5 sinistä palloa. Otetaan kaksi palloa umpimähkään. Mikä on todennäköisyys, että ne ovat samanvärisiä?

b) "Pikakuljetuksen" kolme autoa ovat palvelukuntoisia todennäköisyydellä 0,85, 0,75 ja 0,55. Laske todennäköisyys, että kuljetustilauksen tullessa ainakin yksi autoista on kunnossa (käytä komplementtitapausta).

4.8 Kolminumeroisista luonnollisista luvuista valitaan satunnaisesti yksi. Mikä on todennäköisyys, että tässä luvussa esiintyvät numerot 5 ja 6 peräkkäin tässä järjestyksessä? (Ensimmäinen numero ei voi olla 0.)

5 Kombinaatio-oppia

5.1 Tuloperiaate

Kombinaatio-oppi eli *kombinatoriikka* käsittelee mm. äärellisiin joukkoihin liittyviä lukumääräkysymyksiä. Seuraavassa on muutama tällainen kysymys:

- *Moneenko eri järjestykseen 3 eri kirjainta a, b, c voidaan järjestää*: *abc*, *acb*, *bac*, *bca*, *cab*, *cba*, siis kuuteen järjestykseen.
- *Montako järjestettyä paria näistä kirjaimista saadaan*: (a, b), (b, a), (a, c), (c, a), (b, a), (a, b), siis 6 paria.
- *Montako 2-alkioista osajoukkoa eli yhdelmää näistä kirjaimista saadaan* (osajoukoissa ei alkioiden järjestys vaikuta mitään): $\{a, b\}$, $\{a, c\}$, $\{b, c\}$, siis 3 osajoukkoa.
- *Montako sellaista paria* (x, y) *saadaan, jossa x valitaan näistä kirjaimista ja y luvuista 2 ja 3*: $(a, 2)$, $(a, 3)$, $(b, 2)$, $(b, 3)$, $(c, 2)$, $(c, 3)$.

Tämäntapaisiin kysymyksiin etsitään jatkossa joitakin lainalaisuuksia ja niitä käytetään todennäköisyyksien laskemiseen.

Edellisessä luvussa oli itse asiassa eräs tällainen lukumääriä koskeva tulos, nimittäin se, että jos A:ssa on m alkiota ja B:ssä n alkiota, niin tulojoukossa $A \times B$ on mn alkiota (paria). Tämä tulos voidaan esittää myös seuraavassa muodossa:

Tuloperiaate: *Jos a voidaan valita m:stä ja b n:stä alkiosta, niin erilaisia* (a,b)*-pareja on mn kpl.*

Tuloperiaate yleistyy kolmikoihin (a,b,c) ja yleisesti (järjestettyihin) *k-jonoihin* $(a_1,a_2,\ldots,a_k)$.

Esim. 1 3 tytöstä ja 2 pojasta saadaan tyttö-poika pareja $3 \cdot 2 = 6$ kpl:

$$(t_1, p_1), (t_1, p_2), (t_2, p_1), (t_2, p_2), (t_3, p_1), (t_3, p_2).$$

(parit (p_i, t_j) eivät anna uusia tyttö-poika-yhdelmiä).

Esim. 2 Jos a:lla on 2 mahdollista arvoa a_1, a_2 ja b:llä samoin b_1, b_2 sekä c:llä 3 arvoa c_1, c_2, c_3, niin järjestettyjä (a,b,c)-kolmikkoja on $2 \cdot 2 \cdot 3 = 12$ kpl. Kirjoita ne näkyviin!

5.2 Permutaatiot ja kombinaatiot

Tuloperiaatteesta seuraa muutama kombinaatio-opin perustulos. Niitä esitellään seuraavassa numeroituna luettelona.

1) *n alkiota voidaan järjestää*

$$\boxed{n(n-1)(n-2)\cdots 2 \cdot 1 = n!}$$

eri järjestykseen. Toisin sanoen ***n:n alkion permutaatioiden (järjestysten) lukumäärä on n!*** (lue: "n kertoma")

Todistus: 1. alkioksi voidaan valita mikä tahansa n:stä alkiosta, 2. alkioksi mikä tahansa jäljellä olevasta $(n-1)$:stä alkiosta jne, ja viimeisen alkion kohdalla on vain yksi valintamahdollisuus.

Esim. 1 a) 6 henkilöä voidaan järjestää jonoon 6! = 720 eri järjestykseen (missä kaksi 6-jonoa ovat erilaisia, jos ne eroavat ainakin jonon yhdessä kohdassa toisistaan).

b) 4 kirjaimesta S, M, A, U saadaan 4! = 24 erilaista 4-kirjaimista "sanaa". Seuraavassa ne on kirjoitettu pystyriveittäin tiettyä systemaattista järjestystä noudattaen (yritä ymmärtää käytetty systematiikka):

SMAU	**MSAU**	**ASMU**	**USMA**
SMUA	**MSUA**	**ASUM**	**USAM**
SAMU	**MASU**	**AMSU**	**UMSA**
SAUM	**MAUS**	**AMUS**	**UMAS**
SUMA	**MUSA**	**AUSM**	**UASM**
SUAM	**MUAS**	**AUMS**	**UAMS**

2) *Jos n:stä alkiosta otetaan k alkiota kerrallaan, niin näistä n:stä alkiosta muodostettuja k-jonoja* $(a_1, a_2, \ldots, a_k)$ *saadaan*

$$\boxed{n(n-1)(n-2)\cdots(n-k+1)}$$

kappaletta. Toisin sanoen n:n alkion ***k-permutaatioiden*** *lukumäärä on* $\boldsymbol{n(n-1)\cdots(n-k+1)}$.

Todistus: 1. alkio voidaan valita *n*:stä, 2. alkio $(n-1)$:stä, jne, ja viimeinen eli *k*:s alkio $(n-(k-1))=(n-k+1)$:stä alkiosta.

Esim. 2 6 henkilöstä saadaan erilaisia 4 henkilön "soppajonoja"

$$6\cdot 5\cdot 4\cdot 3 = 360 \text{ kpl.}$$

3) *Jos n:stä alkiosta otetaan k alkiota, aina välillä palauttaen saatu alkio, niin erilaisia k-jonoja saadaan*

$$\boxed{n\cdot n\cdot \cdots\cdot n = n^k}$$

(sillä jokaisessa vaiheessa on *n* valintamahdollisuutta).

Esim. 3 a) Yksi veikkausrivi on 13-jono $(a_1,\ldots,a_{13})$, missä jokaisella ottelulla a_i on 3 tulosmahdollisuutta 1 × 2. Täten erilaisia veikkausrivejä on $3^{13}\approx 1{,}6$ milj. kappaletta.

b) Biteistä 0, 1 saadaan 8-bittisiä "tavuja" (ts. 8-jonoja) $2^8 = 256$ kpl. Vieressä on lueteltu kaikki 3-bittiset "tavut". Niitä on $2^3 = 8$ kappaletta.

0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

4) Seuraava esimerkki auttanee esimerkin jälkeisen tuloksen ymmärtämistä.

Esim. 4 Montako sellaista 5 numeroista lukua on, joissa on 3 ykköstä ja 2 nelosta (esim 11144, 11414 jne)?

5 numeroa voidaan järjestää 5! eri järjestykseen. Nyt kuitenkin ykkösten vaihto keskenään ei muuta lukua, joten ykkösten osalta aina 3! järjestyksistä antaa saman luvun ja vastaavasti nelosten osalta 2!. Siksi permutaatioiden määrä on jaettava 3!:lla ja 2!:lla:

$$\frac{5!}{3!\cdot 2!} = \frac{1\cdot 2\cdot 3\cdot 4\cdot 5}{1\cdot 2\cdot 3\cdot 1\cdot 2} = 10\,.$$

5) n:stä alkiosta saadaan k:n alkion <u>yhdelmiä</u> (ts. osajoukkoja, jolloin alkioiden järjestys ei vaikuta mitään)

$$\boxed{\binom{n}{k} = \frac{n!}{k!(n-k)!}} \qquad \textit{(lue: "n yli k")}$$

kappaletta. Toisin sanoen n:n alkion ***k-kombinaatioiden*** *lukumäärä on " n yli k".*

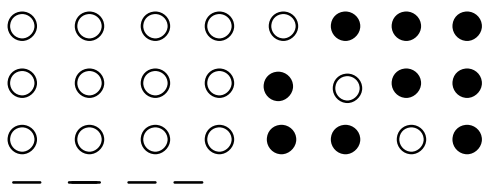
Esim. 5 a) (Lotto) 39 pallosta saadaan 7 pallon yhdelmiä

$$\binom{39}{7} = \frac{39!}{7!\cdot 32!} = \frac{39\cdot 38\cdot 37\cdot 36\cdot 35\cdot 34\cdot 33}{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7} \approx 15{,}4 \text{ milj.}$$

b) 52 kortista saadaan erilaisia 5 kortin yhdelmiä

$$\binom{52}{5} = \frac{52\cdot 51\cdot 50\cdot 49\cdot 48}{1\cdot 2\cdot 3\cdot 4\cdot 5} \approx 2{,}6 \text{ milj.}$$

Esim. 6 5 keskenään samanlaista valkoista ja 3 mustaa palloa asetetaan jonoon. Kuinka monta erinäköistä 8 pallon jonoa saadaan?



Kysymys voidaan muuttaa muotoon: kuinka monta 5 pallon yhdelmää 8 paikasta saadaan (näihin 5 paikkaan sijoitetaan valkoiset pallot ja loppuihin mustat). Vastaus on

$$\binom{8}{5} = \frac{8!}{5!3!} \overset{\text{(sup 5!:lla}}{=} \frac{8\cdot 7\cdot 6}{1\cdot 2\cdot 3} = 56\,.$$

Harjoituksia

A, B

5.1 a) Montako nollatonta 3-numeroista lukua on ? b) Entä sellaisia 3-numeroisia lukuja, joissa sama numero esiintyy vain kerran (numero 0 voi esiintyä muualla paitsi ei alussa)?

5.2 a) Kuinka monta istumajärjestystä 8 henkilöllä on 8-kulmaisessa pöydässä, jos jokainen paikka on eriarvoinen? b) Entä, jos paikat ovat samanarvoisia?

5.3 Kuudesta pojasta arvotaan $4 \times 100\,\text{m}$:n viestijoukkue. Montako erilaista joukkuetta voidaan valita, kun a) juoksujärjestyskin arvotaan, b) järjestykseen ei kiinnitetä huomiota?

5.4 Luokassa on 18 tyttöä ja 13 poikaa. Montako a) poika-tyttö, b) poika-poika, c) tyttö-tyttö -paria voidaan näistä muodostaa?

5.5 Montako erilaista istumajärjestystä luokan 32 oppilaasta saadaan?

5.6 3 poikaa ja 4 tyttöä asettuvat jonoon siten, että tytöt ovat alussa. Kuinka monta erilaista jonoa saadaan?

5.7 6 yhtäsuuresta pallosta on 3 samanväristä ja loput keskenään ja ensin mainittujen kanssa erivärisiä. Kuinka monta erinäköistä 6 pallon jonoa näistä saadaan? (vrt. Esim. 4)

5.8 a) 10 henkilöä kättelevät kaikki toisiaan. Montako kättelyä suoritetaan?
b) 8 joukkuetta pelaavat pareittain (yksi pari aina kerran ts. ei erikseen koti- ja vierasottelua). Montako ottelua pelataan?

5.9 Mielipidekyselyssä oli 6 kysymystä ja niistä jokaisessa oli 5 erilaista vastausmahdollisuutta. Paljonko vastausmahdollisuuksia oli kaikkiaan (kun jokaiseen kysymykseen pitää vastata)?

5.10 Oppilaan tulee vastata 10 kysymyksestä kahdeksaan. a) Montako erilaista vastauskombinaatiota (yhdelmää) hänellä on? b) Entä jos näiden 8 vastauksen joukossa tulee olla vastaukset 3 ensimmäiseen kysymykseen?

5.11 Montako erilaista komiteaa, jossa on 3 miestä ja 2 naista voidaan valita 7 miehestä ja 5 naisesta?

5.12 10 henkilöstä valitaan 6 henkilöinen lentopallojoukkue.Montako valintamahdollisuutta on ?

5.13 10 poikaa ryhmittyvät kahdeksi 5-jäseniseksi joukkueeksi. Montako erilaista ottelijayhdelmää voidaan muodostaa?

6 Diskreeteistä jakaumista

6.1 Binomijakauma

Toistetaan samaa koetta samoissa olosuhteissa ja olettaen, että edelliset toistot eivät millään tavoin vaikuta myöhempien tulokseen.

Lause 1 *Oletetaan, että*

1) koe toistetaan n kertaa,

2) tapauksen A todennäköisyys yhdessä kokeessa on p.

Merkitään komplementtitapauksen $\overline{A}$ *(= A ei tapahdu) todennäköisyyttä q:lla. Silloin* ***todennäköisyys, että A tapahtuu näissä n:ssä kokeessa tarkalleen k kertaa****, on*

$$\boxed{\binom{n}{k} p^k q^{n-k}}$$

Todistus: Eräs tulosjono, jossa A esiintyy tarkalleen k kertaa, on esim.

$$\underbrace{A, A, \ldots, A}_{k \text{ kpl}}, \underbrace{\overline{A}, \overline{A}, \ldots, \overline{A}}_{n-k \text{ kpl}}$$

Jokaisen tällaisen tulosjonon todennäköisyys on kertosäännön nojalla

$$p \cdot p \cdots p \cdot q \cdot q \cdots q = p^k q^{n-k}$$

Tällaisia tulosjonoja on yhtä monta kuin on mahdollisuuksia valita n:stä paikasta k:n paikan yhdelmiä (joihin A:t sijoitetaan). Yhdelmiä on $\binom{n}{k}$ kpl. Todennäköisyys, että saadaan jokin näistä tulosjonoista (ts. tuloksena on 1. tai 2. tai ... jono), on additiivisuuden nojalla

$$\binom{n}{k} \cdot p^k q^{n-k} .$$

Esim. 1 Heitetään noppaa 7 kertaa. Laske todennäköisyys, että saadaan 5 tarkalleen 3 kertaa. Tässä tehtävässä A on tapaus "saadaan 5 yhdessä heitossa", $p = P(A) = \frac{1}{6}$, $\therefore q = \frac{5}{6}$. Täten

todennäköisyys, että 7 heitossa saadaan 5 tarkalleen 3 kertaa, on

$$\binom{7}{3}\cdot\left(\frac{1}{6}\right)^3\cdot\left(\frac{5}{6}\right)^4 = \frac{7\cdot 6\cdot 5}{1\cdot 2\cdot 3}\cdot\frac{5^4}{6^7} = 7\cdot 5^5 / 6^7 \approx 0{,}078.$$

Muutetaan Lauseen 1 tulos satunnaismuuttujaa koskevaan muotoon:

Lause 2 *Oletetaan, että tapauksen A todennäköisyys yhdessä kokeessa on p ja tapauksen* $\overline{A}$ *on* $q\,(=1-p)$. *Koe toistetaan n kertaa. Jos X on satunnaismuuttuja, joka ilmoittaa, kuinka monta kertaa A tapahtuu näissä n:ssä kokeessa, niin X:n saamat arvot ovat* x_i: $0, 1, \ldots, n$ *ja näiden todennäköisyydet ovat*

$$\boldsymbol{p_k = P(X = k) = \binom{n}{k} p^k\, q^{n-k}} \quad (k = 0, 1, \ldots, n).$$

Satunnaismuuttujaa X, jonka arvot ovat jakautuneet tämän säännön mukaisesti, sanotaan ***binomijakautuneeksi, parametreina n ja p***. Tätä merkitään lyhyesti

$$\boldsymbol{X \sim Bin(n, p)}.$$

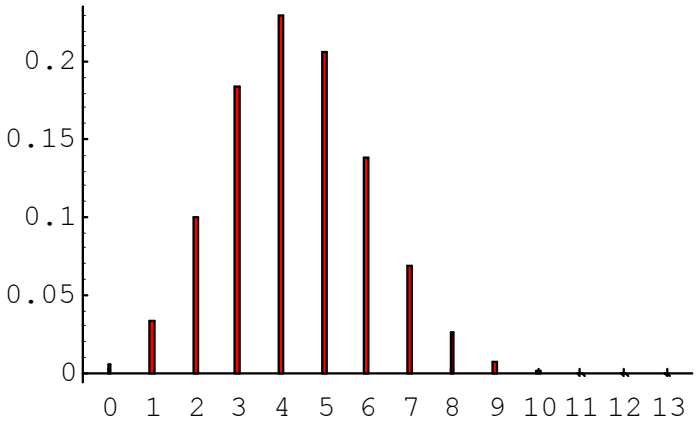
Binomijakauman yhteydessä lukua p sanotaan joskus *onnistumistodennäköisyydeksi* ja lukua q epäonnistumistodennäköisyydeksi. Lyhyesti sanottuna *todennäköisyys, että n:ssä toistossa onnistutaan tarkalleen k kertaa, on* $\binom{n}{k} p^k q^{n-k}$.

Esim. 2 Jos X ilmoittaa 13 ottelun veikkauksessa oikeiden määrän, niin X:n jakauma on binomijakauma, parametreina $n = 13$ ja $p = 1/3$ (= "onnistumistodennäköisyys" yhden yksittäisen ottelun veikkaamisessa). Vieressä ovat "oikeiden" määrät ja niiden todennäköisyydet, jotka on laskettu kaavasta

$$p_k = P(X = k) = \binom{13}{k}\left(\frac{1}{3}\right)^k\left(\frac{2}{3}\right)^{13-k}.$$

0	0,0051
1	0,0334
2	0,1002
3	0,1837
4	0,2296
5	0,2067
6	0,1378
7	0,0689
8	0,0258
9	0,0072
10	0,0014
11	0,0002
12	0,000016
13	$6{,}27\cdot 10^{-7}$

Seuraavassa kuvassa on tämä veikkausjakauma esitettynä janadiagrammina.



Edellisestä taulukosta voidaan laskea, mikä on todennäköisyys, että saadaan a) korkeintaan 4 b) vähintään 5 oikein:

a) $P(X \le 4) = p_0 + p_1 + p_2 + p_3 + p_4 \approx 0{,}552$, b) $1 - 0{,}552 = 0{,}448$.

Binomijakauma on eräs diskreetti jakauma. Diskreetin jakauman odotusarvo (*Expected Value*) ja varianssi määriteltiin jo luvussa 2. (kohta 2) seuraavasti:

$$\boxed{\boldsymbol{E(X) = \sum p_i x_i}}, \qquad \boxed{\boldsymbol{Var(X) = \sum p_i\,(x_i - E(X))}}.$$

Keskihajonta σ on varianssin neliöjuuri $\boxed{\boldsymbol{\sigma = \sqrt{Var(X)}}}$.

Voidaan todistaa seuraava tulos:

Lause 3 *Jos* $\boldsymbol{X \sim Bin(n, p)}$*, niin*

$$\boldsymbol{E(X) = n \cdot p} \quad \textit{ja} \quad \boldsymbol{Var(X) = n \cdot p \cdot q}.$$

Esim. 3 Veikkausjakaumalla on

$$E(X) = 13 \cdot \frac{1}{3} = 4\frac{1}{3}, \qquad Var(X) = 13 \cdot \frac{1}{3} \cdot \frac{2}{3} = 2\frac{8}{9}.$$

Esim. 4 Jos X ilmoittaa viiden rahan heitossa klaavojen määrän, niin $X \sim \text{Bin}(5, ½)$ ja $E(X) = 5 \cdot \frac{1}{2} = 2\frac{1}{2}$.

*6.2 Geometrinen jakauma

Kuten binomijakaumallakin, toistetaan samaa koetta kerta kerran jälkeen samoissa olosuhteissa ja olettaen, että edelliset toistot eivät millään tavoin vaikuta myöhempien tulokseen. Nyt kuitenkin tarkoituksena on toistaa koetta, kunnes "onnistutaan" (onnistuminen voi olla esim. tietyn vian ilmeneminen tai viallisen kappaleen löytyminen).

Olkoon yhdessä kokeessa "onnistumistodennäköisyys" p ja "epäonnistumistodennäköisyys" $q = 1 - p$

Merkitään X:llä satunnaismuuttujaa, joka ilmoittaa monennellako kerralla tapahtuu ensimmäinen onnistuminen. Kertosäännön mukaan todennäköisyys, että näin käy vasta k:nnella kerralla, on

$$p_k = P(X = k) = q \cdot q \cdot \ldots \cdot q \cdot p = q^{k-1} p$$

($k-1$ epäonnistumista ja sitten onnistuminen). Satunnaismuuttujaa X, jonka arvot ovat $x_i = 1, 2, 3 \ldots$ ja pistetodennäköisyydet $\boxed{p_k = q^{k-1} \cdot p}$, sanotaan *geometrisesti jakautuneeksi*.

Esim. 1 Heitetään noppaa, kunnes tulee 5 tai 6. Mikä on todennäköisyys, että tämä tapahtuu a) neljännellä kerralla, b) jollakin parittomalla kerralla?

Jos X ilmoittaa, monennellako kerralla onnistutaan, niin

$$P(X = k) = (\frac{2}{3})^{k-1} \cdot \frac{1}{3} \qquad (p = \frac{2}{6} = \frac{1}{3},\ q = 1 - \frac{1}{3} = \frac{2}{3})$$

Täten a) $P(X = 4) = \frac{2^3}{3^4} = \frac{8}{81}$. b) Additiivisuuden nojalla

$$P(X = 1 \text{ tai } 3 \text{ tai } 5 \text{ tai } \ldots) = \frac{1}{3} + (\frac{2}{3})^2 \cdot \frac{1}{3} + (\frac{2}{3})^4 \cdot \frac{1}{3} + \ldots.$$

Näin joudutaan suppenevaan *geometriseen sarjaan*

$$\frac{1}{3} \cdot (1 + q^2 + q^4 + \ldots) = \frac{1}{3} \cdot \frac{1}{1 - q^2} = \frac{1}{3 \cdot (1 - \frac{4}{9})} = \frac{3}{5}.$$

6.3 Hypergeometrinen jakauma

Hypergeometrista jakaumaa käytetään tilanteessa, jossa on kahdenlaisia alkioita. Esim. äärellisestä määrästä tuotteita osa on virheellisiä ja loput virheettömiä. Lotossa taas arvonnan jälkeen on 7 oikeaa numeroa (ilman lisänumeroita) ja 32 väärää.

Lause 4 *Joukossa on M alkiota. Näistä on N kpl lajia 1 ja loput $M - N$ kpl lajia 2. Otetaan joukosta r alkiota. Jos X ilmoittaa, montako otetuista alkioista on lajia 1, niin todennäköisyys, että näitä on k kappaletta, on*

$$p_k = P(X = k) = \frac{\binom{N}{k} \cdot \binom{M-N}{r-k}}{\binom{M}{r}} \qquad (k = 0,1,\ldots,N).$$

Esim. 1 100 tuotteen tavaraerässä on 5 virheellistä. Jos otetaan 10 tuotteen näyte, niin todennäköisyys, että näistä on 3 virheellistä, on

$$P(X = 3) = \frac{\binom{5}{3} \cdot \binom{95}{7}}{\binom{100}{10}} \approx 0{,}0064,$$

sillä kaikkiaan mahdollisuuksia ottaa 10 tuotteen otoksia 100 tuotteesta on $\binom{100}{10}$ kpl ja kysyttyjä otoksia, joissa on 3 virheellistä tuotetta 5:stä ja 7 virheetöntä 95:stä, on $\binom{5}{3} \cdot \binom{95}{7}$ kpl.

Samalla periaatteella voidaan todistaa Lause 4.

Huom. Laskimessasi on ehkä näppäin *nCr*, joka antaa $\binom{n}{r}$:n arvon.

Esim. 2 Todennäköisyys, että lotossa saadaan 5 oikein (ilman lisänumeroita), on seuraava:

$$P(X=5) \;=\; \frac{\binom{7}{5}\cdot\binom{32}{2}}{\binom{39}{7}},$$

sillä 7 oikeasta numerosta täytyy saada 5 ja 32 väärästä loput 2, kun taas yhteensä numeroita on 39 ja niistä 7 on oikeita.

Laske edellisen lausekkeen likiarvo. Todennäköisyydet "0 oikein" ja "7 oikein" voidaan laskea myös kertosäännöllä. Miten?

6.4 Poisson-jakauma

Olkoon $\lambda > 0$. Satunnaismuuttujan X jakauma on ***Poisson-jakauma, parametrina*** $\boldsymbol{\lambda}$, jos X saa arvot $x_i = 0, 1, 2, \ldots$ ja näiden todennäköisyydet ovat

$$\boxed{p_k = P(X=k) = e^{-\lambda}\cdot\frac{\lambda^k}{k!} \qquad (k=0,1,\ldots)}$$

Poisson-jakaumaa voidaan käyttää binomijakauman likiarvona, jos n on suuri ja p pieni (esim. $n=20,\ p=0{,}1$), sillä voidaan todistaa, että tällöin

$$\binom{n}{k}\cdot p^k q^{n-k} \approx e^{-\lambda}\cdot\frac{\lambda^k}{k!}, \text{ missä } \underline{\underline{\lambda = np.}}$$

Binomijakaumaa sovellettaessa p voi ilmoittaa, kuinka suuri osa kokeista keskimäärin onnistuu. (Jos esim. yksi koe 10:stä keskimäärin onnistuu, niin $p=0{,}1$). Täten $\lambda\,(=np)$ <u>ilmoittaa, kuinka monta n:stä kokeesta keskimäärin "onnistuu"</u>.

Esim. 1 Konekirjoittaja lyö keskimäärin 4 virhelyöntiä tunnissa. Laske todennäköisyys, että hän seuraavan puolen tunnin aikana lyö a) 3 virhelyöntiä, b) korkeintaan yhden virhelyönnin.

Tässä "kokeessa" satunnaismuuttuja X ilmoittaa "onnistumisten" eli virhelyöntien määrän kyseisen puolen tunnin aikana. Puolessa tunnissa tulee keskimäärin 2 virhelyöntiä, joten $\lambda = 2.$ Seuraavissa laskuissa tarvitaan tietoa $\boxed{0!=1}$ (sillä

eräs matemaattinen yleissopimus on, että tyhjän tulon eli tulon, jossa ei kerrota yhtään lukua keskenään, arvo on 0):

a) $P(X=3)=e^{-2}\cdot\frac{2^3}{3!}\approx 0{,}18$,

b) $P(X=0 \text{ tai } 1)=e^{-2}\cdot\frac{2^0}{0!}+e^{-2}\cdot\frac{2^1}{1!}=e^{-2}(1+2)\approx 0{,}41$

Esim. 2 Suuressa tuote-erässä on 1% virheellisiä. Laske todennäköisyys, että 200 kpl näytteessä on 3 virheellistä.

Koska virheellisiä on 1%, niin 200 kpl erässä on keskimäärin 2 virheellistä. Siten $\lambda=2$ ja $P(X=3)\approx e^{-2}\cdot\frac{2^3}{3!}\approx 0{,}180$. Sama voidaan laskea myös binomijakaumalla:

$$P(X=3)=\binom{200}{3}0{,}01^3\ 0{,}99^{197}\approx 0{,}181.$$

Poisson-jakauma on sovellettavissa moniin käytännön kokeisiin, joissa jokin poikkeuksellinen tapahtuma sattuu aika-ajoin (suhteellisen harvoin). Tällaisia ovat esim. virheellisen tuotteen esiintyminen tuotantoketjussa, asiakkaan saapuminen myymälään tai myöhästyminen tapaamisesta, väärän puhelinnumeron valitseminen, tietynlaiseen onnettomuuteen joutuminen, virheen tapahtuminen tiedonsiirrossa jne. Poisson-jakautunut satunnaismuuttuja X ilmoittaa tarkasteltavana aikavälinä sattuneiden poikkeuksellisten tapahtumien lukumäärän.

Esim. 3 Paikallistiellä ohittaa tietyn kohdan päiväsaikaan keskimäärin 48 autoa tunnissa (ilman ruuhkautumisia). Laske todennäköisyys, että aikavälinä 14.00 – 14.05 kyseisen paikan ohittaa vähintään 3 autoa.

5 minuutin aikana paikan ohittaa keskimäärin $\frac{5}{60}\cdot 48=\frac{48}{12}=4$ autoa, joten $\lambda=4$. Jos X ilmoittaa kyseisenä 5 min. aikavälinä ohittaneiden autojen lukumäärän, niin

$$P(X\geq 3)=1-P(X<3)=P(X=0,1 \text{ tai } 2)$$
$$=e^{-4}\cdot(\frac{4^0}{0!}+\frac{4^1}{1!}+\frac{4^2}{2!})=e^{-4}\cdot(1+4+8)\approx 0{,}24.$$

Harjoituksia

A

6.1 Eräästä lääketieteellisestä leikkauksesta on todettu toipuvan hyvin 95%. Millä todennäköisyydellä seuraavasta 8 potilaasta toipuu hyvin a) viisi, b) ainakin neljä?

6.2 Nasta jää sitä heitettäessä kärki ylöspäin todennäköisyydellä 0,4. Heitetään 8 nastaa. Laske todennäköisyys, että niistä on 3 kärki ylöspäin (ja 5 vinosti alaspäin).

6.3 Heitetään edellisen tehtävän nastaa, kunnes se jää kärki ylöspäin. Mikä on todennäköisyys, että tämä tapahtuu 3:nnella tai 4:nnellä kerralla?

6.4 Pussissa on 6 valkoista ja 4 mustaa palloa. Niistä otetaan viisi. Laske todennäköisyys, että palloista on 2 valkoista (ja 3 mustaa).

6.5 Kuten edellinen tehtävä, mutta pallot nostetaan peräkkäin ja aina jokaisen noston jälkeen pallo palautetaan.

6.6 Kesämökillä sattuu 3 kesäkuukauden aikana keskimäärin 5 sähkökatkosta. Mikä on todennäköisyys, että seuraavan heinäkuun aikana tulee vähintään 2 sähkökatkosta.

B

6.7 Pelatessaan B:tä vastaan A voittaa keskimäärin 2 peliä 5:stä. Laske todennäköisyys, että viidestä pelistä A voittaa a) kaksi, b) vähintään kolme.

6.8 Onko pitkän päälle edullista lyödä vetoa, sen puolesta, että 8 nopanheitossa saadaan ainakin 2 kuutosta?

6.9 Eräässä pelissä voitetaan 250 euroa todennäköisyydellä 0,3 ja hävitään 300 euroa todennäköisyydellä 0,2 ja 100 euroa todennäköisyydellä 0,4 (muulloin eli todennäköisyydellä 0,1 ei voiteta eikä hävitä). Laske voiton odotusarvo. Mitä saatu tulos merkitsee, jos pelataan usein tällaista peliä?

6.10 Olkoon $X \sim \mathrm{Bin}(4, 1/3)$. a) piirrä X:n jakauman graafinen esitys (janadiagrammi) ja kertymäfunktion kuvaaja yksikkönä y-akselilla 10 cm. b) Laske jakauman odotusarvo ja keskihajonta.

6.11 Särmiön muotoinen pieni kappale putoaa sitä heitettäessä lappeelleen todennäköisyydellä 2/3. Suoritetaan 4 heittoa. Ilmoittakoon X tulosten "putoaa lappeelleen" määrän.

a) Määritä X:n jakauma.

b) Piirrä jakauman graafinen esitys.

c) Piirrä kertymäfunktion kuvaaja.

d) Laske X:n odotusarvo ja varianssi.

6.12 A ja B heittävät kumpikin 4 tikkaa. A:n osumistarkkuus 7-renkaaseen tai sen sisälle on joka tikalla 60 %, kun taas B:n on ensimmäisellä tikalla 50% ja seuraavilla aina 5 prosenttiyksikköä suurempi. Laske todennäköisyys, että a) A, b) B saa vähintään 3 osumaa (laskentamenetelmät ovat erilaiset a- ja b-kohdissa).

6.13 Arpoja on 200 kpl, joista 5 voittoa. Ostat 3 arpaa. Laske hypergeometrisen jakauman avulla todennäköisyys, että saat a) 2 voittoa, b) ainakin yhden voiton. Miten muuten voit laskea esim. a)-kohdan? Mieti tilannetta, jos kaikki lukumäärät ovat 10-kertaiset (2000 arpaa, 50 voittoa, ostat 30 arpaa, saat 20 voittoa)?

6.14 Onnenpyörä on jaettu 8 yhtä suureen sektoriin, joista jokerisektori on yksi. Olkoon X satunnaismuuttuja, joka ilmoittaa jokerien lukumäärän 3 pyörityksessä. a) Määritä X:n jakauma, b) Laske odotusarvo.

6.15 Heitetään rahaa kunnes tulee klaava. Laske todennäköisyys, että tämä tapahtuu jollakin parittomalla heittokerralla.

6.16 Laatikossa on 50 oikeaa (mm-kokoista) mutteria ja 7 väärää (tuumakokoa). Laske todennäköisyys, että viiden otetun mutterrin joukkoon joutuu 2 väärää mutteria.

6.17 Koneen tekemistä tuotteista on 1 % virheellisiä. Laske todennäköisyys, että 100 tuotteen otoksessa on ainakin 3 virheellistä.

6.18 500-sivuisen kirjan sivuilla on 300 painovirhettä. Laske todennäköisyys, että valitulla sivulla on a) tasan 2 painovirhettä, b) 2 tai useampia painovirheitä.

6.19 Laske Poisson-jakauman odotusarvo.

7 Jatkuvia jakaumia

7.1 Yleistä. Tasainen jakauma

Edellä käsiteltiin muutamaa diskreettiä jakaumaa. Esim. *Mathematica 3*-ohjelma tuntee seuraavat diskreetit jakaumat, joista tavallisimmat on seuraavassa luettelossa lihavoitu:

Bernoulli Distribution

Binomial Distribution

Discrete Uniform Distribution

Geometric Distribution

Hypergeometric Distribution

LogSeries Distribution

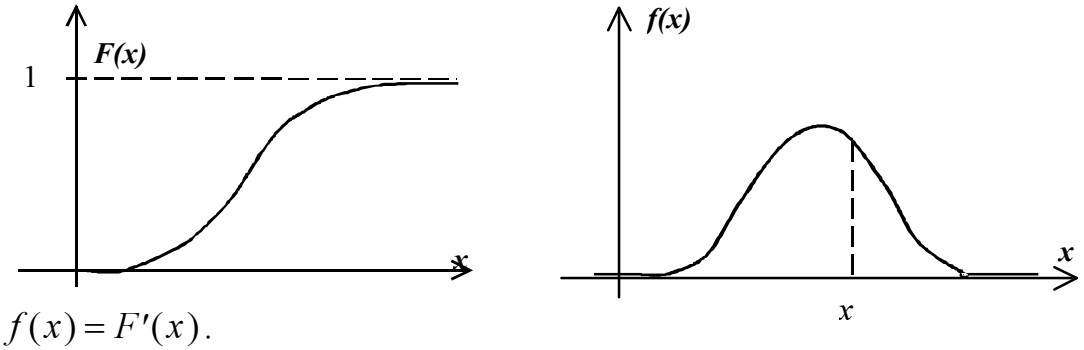
Negative Binomial Distribution

Poisson Distribution.

Kyseisen ohjelman *Help*issä on näiden ohjelmien kuvausta (*Help Browser, Add-ons, Standard Packages, Statistics, Discrete Distributions*). Jatkuvia jakaumia löytyy saman ohjelman luettelosta vielä suurempi joukko:

Beta Distribution
Cauchy Distribution
Chi Distribution
Chi Square Distribution
Exponential Distribution
Extreme Value Distribution
Fratio Distribution
Gamma Distribution
Half Normal Distribution
Hypergeometric Distribution
Laplace Distribution
Logistic Distribution
LogNormal Distribution
Noncentral Chi Square Distribution
Noncentral Fratio Distribution
Noncentral Student T Distribution
Normal Distribution
Pareto Distribution
Rayleigh Distribution
Student T Distribution
Uniform Distribution
Weibull Distribution

Luvussa 2 käsiteltiin jo jatkuvia jakaumia jonkin verran. Satunnaismuuttujan X jakauma on jatkuva jollakin välillä, jos *kertymäfunktio* F on jatkuva funktio (eli todennäköisyysmassa kertyy jatkuvasti tällä välillä). *Tiheysfunktio* f on kertymäfunktion derivaatta:



$f(x) = F'(x)$.

(*Tiheysfunktion kuvaajassa voi olla epäjatkuvuuskohtia (hyppäyksiä), sillä kertymäfunktiolla voi joissakin kohdissa olla vasemman- ja oikeanpuoliset derivaatat erisuuret.)

Todennäköisyysmassan kertymävälinä voi olla myös kaikkien reaalilukujen muodostama väli $-\infty .. +\infty$, kuten oli esim. normaalijakaumalla.

Kertymäfunktion arvo kohdassa x saadaan laskemalla kohtaan x mennessä kertyneen todennäköisyysmassan määrä, siis

$$\boxed{F(x) = \int_{-\infty}^{x} f(x)\,dx}$$

Välin $a \ldots b$ todennäköisyys voidaan laskea seuraavalla kahdella tavalla: 1) lasketaan kohtaan b mennessä kertynyt todennäköisyysmassan määrä ja vähennetään siitä kohtaan a mennessä kertyneen todennäköisyysmassan määrä tai 2) lasketaan tiheysfunktion avulla integroimalla välillä $a \ldots b$ olevan todennäköisyysmassan määrä:

1) $\boxed{P(a < X \le b) = F(b) - F(a)}$ 2) $\boxed{P(a < X \le b) = \int_a^b f(x)\,dx}$

Kokonaismassan määrän täytyy olla = 1, ts. $\boxed{\int_{-\infty}^{\infty} f(x)\,dx = 1}$.

Luvussa 2 mainittiin myös X:n ***odotusarvon*** ja ***varianssin*** lausekkeet:

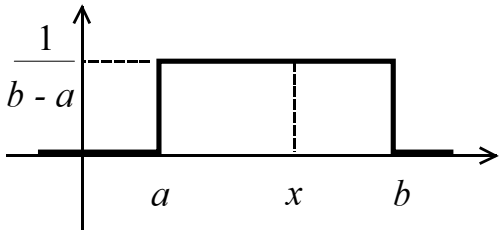
$\boxed{\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)\,dx}$, $\boxed{\sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx}$.

Varianssin neliöjuuri $\sigma = Sd(X)$ on nimeltään ***keskihajonta*** (*Standard deviation*).

*Tekniikan käsitteitä käyttäen odotusarvo on tiheysfunktion ja *x*-akselin rajoittaman alueen painopisteen *x*-koordinaatti (eli alueen staattinen momentti S_y jaettuna alueen alalla, joka on = 1), sillä yhden kohdassa *x* olevan pystyliuskan staattinen momentti on $dS_y = xdA = x \cdot f(x) \cdot dx$. Vastaavasti varianssi on saman alueen neliömomentti suoran $x = \mu$ suhteen.

Esim. 1 (*Uniform Distribution*) ***S***atunnaismuuttuja X on ***tasaisesti jakautunut*** välille *a ... b,* jos sen tiheysfunktio on

$$f(x) = \begin{cases} \dfrac{1}{b-a} \text{ välillä } a \ldots b \\ 0 \text{ muulloin} \end{cases}$$



Tämä käy tiheysfunktioksi, koska *x*-akselin ja viivan välisen alueen ala on

$$\frac{1}{b-a} \cdot (b-a) = 1 .$$

Kertymäfunktio saadaan laskemalla pinta-ala kohtaan *x* saakka (mikä ei tässä tapauksessa vaatisi integrointia):

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \begin{cases} 0, \text{ kun } x < a \\ \dfrac{1}{b-a}(x-a) \text{ välillä } a \le x < b \\ 1, \text{ kun } x \ge b \end{cases}.$$

Odotusarvo on

$$\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)\,dx = \int_{a}^{b} x \cdot \frac{1}{b-a}\,dx = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}$$

ja varianssi $\sigma^2 = Var(X) = \dfrac{(b-a)^2}{12}$ (harj.). Keskihajonta on täten $\sigma = \dfrac{b-a}{2\sqrt{3}}$.

7.2 Eksponenttijakauma

Eräillä tuotteilla (esim. joillakin elektroniikan komponenteilla) on se ominaisuus, että niiden jäljellä oleva toimivuusaika ei käytännöllisesti katsoen riipu siitä, miten pitkään ne ovat jo olleet käytössä.

Tällaisia tuotteita rikkoutuu aika-ajoin ja alussa eniten, koska tällöin tuotteita on vielä paljon jäljellä. Yleisesti jäljellä olevien, vielä toimivien tuotteiden *määrä vähenee ajan mukana eksponentiaalisesti*:

(1) $$m = m_o \cdot e^{-\lambda t},$$

missä m_o on hetkellä $t = 0$ toimivien tuotteiden määrä.

*Perustelu: tuotemäärän muutos $dm\,(<0)$ hetkellä t alkavalla lyhyellä aikavälillä dt on verrannollinen välin pituuteen ja siihen, paljonko tuotteita on vielä kyseisellä hetkellä jäljellä: $dm = -\lambda \cdot m \cdot dt$. Tämän differentiaaliyhtälön ratkaisu on (1).

Yhtälöstä (1) seuraa (kuten logaritmiopissa, monisteessa Algebra 2 on osoitettu), että vielä kunnossa olevien tuotteiden määrällä on *puoliintumisaika* $T = \ln 2 / \lambda$. Täten parametri $\boxed{\lambda = \ln 2 / T}$.

Jos satunnaismuuttuja X *ilmoittaa tällaisen tuotteen kestoiän*, niin X:n kertymäfunktion arvo hetkellä t saadaan, kun lasketaan, millä osalla tuotteista on kestoaika välillä $0 \ldots t$ eli mikä osa tuotteista on rikkoutunut hetkeen t mennessä:

$$F(t) = \begin{cases} \dfrac{m_o - m_o e^{-\lambda t}}{m_o} = \underline{\underline{1 - e^{-\lambda t}}}, & \text{kun } t \geq 0 \\ 0, & \text{kun } t < 0 \end{cases}.$$

Tästä saadaan derivoimalla tiheysfunktio

(2) $$\boxed{f(t) = \begin{cases} \lambda e^{-\lambda t}, & \text{kun } t \geq 0 \\ 0, & \text{kun } t < 0 \end{cases}} \qquad (\lambda > 0)$$

Satunnaismuuttujaa, jolla on tällainen jakauma (eli tällainen tiheysfunktio), sanotaan *eksponenttijakautuneeksi, parametrina λ.*

Eksponenttijakauman odotusarvoksi ja varianssiksi saadaan (osittais-integroimalla)

(3) $$E(X) = \int_{-\infty}^{\infty} t\, f(t)\, dt = \lambda \int_0^{\infty} t\, e^{-\lambda t}\, dt = \frac{1}{\lambda} \qquad Var(x) = \frac{1}{\lambda^2}.$$

Esim. 1 Eksponenttijakaumaa noudattavista tuotteista puolet on rikkoutunut viikon sisällä. Mihin mennessä jäljellä on vielä kolmasosa tuotteista?

Puoliintumisaika $T = 1$ viikko, joten $\lambda = -\ln 2 / (1\text{ viikko})$. Kysymys on, monenko viikon kuluessa rikkoutumisia on kertynyt 2/3:

$$\begin{aligned} 1 - e^{-\lambda t} &= 2/3 \\ e^{-\lambda t} &= 1/3 \\ -\lambda t &= \ln(1/3) = -\ln 3 \\ t &= \frac{\ln 3}{\lambda} = \frac{\ln 3}{\ln 2/(1\text{ viikko})} = \frac{\ln 3\,(1\text{ viikko})}{\ln 2} \\ &= \frac{\ln 3}{\ln 2}\text{ viikkoa} \approx 1{,}6\text{ viikkoa} \end{aligned}$$

Eksponenttijakaumaan joudutaan myös toisenlaisten probleemien yhteydessä. Oletetaan, että jokin satunnaismuuttuja Y ilmoittaa jonakin tiettynä aikavälinä tapahtuneiden poikkeuksellisten tapahtumien lukumäärän. Poikkeuksellisia tapahtumia ovat esim. asiakkaiden saapuminen palvelupisteeseen, valvontakohtaan tullut puhelu tai vikailmoitus, koneen rikkoutuminen, auton tuleminen huoltoon jne. *Y*:n jakauma on silloin diskreetti, *Poisson*-jakauma. Mutta jos tarkastellaankin satunnaismuuttujaa *X, joka ilmoittaa kahden poikkeuksellisen tapahtuman välisen ajan pituuden*, niin *X*:n jakauma on eksponentti-jakauma.

Jos esimerkiksi

$$f(t) = 4e^{-4t}, \text{ kun t} \geq 0 \text{ ja } [t] = 1\,s$$

niin tapahtumien välisen ajan odotusarvo on $E(X) = \frac{1}{\lambda} = \frac{1}{4}\,s$, ts. keskimäärin sattuu 4 tapahtumaa sekunnissa.

Esim. 2 Tietyn koneen vioittumisten välinen aika noudattaa eksponenttijakaumaa, jossa odotusaika on 17 vrk (ts. jos koneen toimintaa seurataan pitkän aikaa, niin keskimäärin koneeseen tulee vika n. 17 vuorokauden välein). Kuinka suuri

osa vioista tapahtuu korkeintaan 15 vuorokauden kuluessa edeltäneestä viasta.

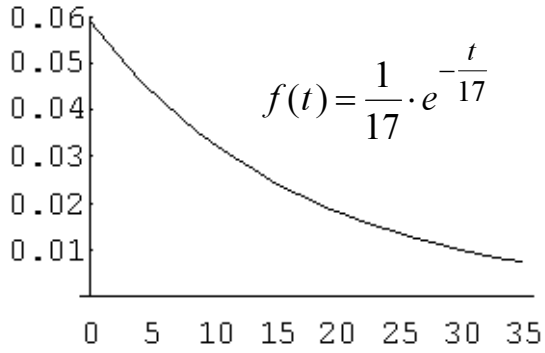
Odotusarvotuloksen (3) mukaan $\lambda = 1/E(X) = 1/17$ vrk. Täten

$$P(X \le 15 \text{ vrk}) = \mathrm{F}(15 \text{ vrk}) = 1 - e^{-\frac{1}{17 \text{ vrk}} \cdot 15 \text{ vrk}} \approx 0{,}59.$$

Täten n. 59 %:n todennäköisyydellä seuraava vika tulee 15 vrk kuluessa.

*Huomaa, että odotusarvo eli todennäköisyysmassan painopistekohta on 17, mutta edellisen tuloksen mukaan pinta-alasta on kohtaan 15 mennessä kertynyt jo 59 %.

Tässä esimerkissä tiheysfunktion kuvaaja on seuraavan näköinen:



Niiden tuotteiden osuus, jotka vioittuvat 15 vuorokauden kuluessa voidaan laskea myös tiheysfunktion avulla integroimalla:

$$\begin{aligned} P(0 \le X \le 15) &= \int_0^{15} \frac{1}{17} e^{-t/17}\, dt \\ &= -\int_0^{15} e^{-t/17} \left(-\frac{1}{17}\right) dt \\ &= -\Big|_0^{15} e^{-t/17} = -(e^{-15/17} - 1) \approx 0{,}59. \end{aligned}$$

Harjoituksia

A, B

7.1 Laske $P(0 < X \le 2)$, kun X:n jakauma on tasainen välillä $[-1, 5]$.

7.2 Satunnaismuuttujan jakauma on tasainen aikavälillä $[0, T]$. Laske todennäköisyys, että X saa arvon aikaväliltä

a) $-1 < t < 2$, kun $T = 10$, b) $|t| > 1$, kun $T = 1{,}5$.

7.3 Tehdas valmistaa tuotetta, jonka kestoaika noudattaa eksponenttijakaumaa. Määritä parametri λ, kun tiedetään, että tuotteista 80 % kestää vähintään 3 vuotta.

7.4 Huoltoinsinööri saa hätäpuhelun keskimäärin 3 tunnin välein. Jos oletetaan, että puhelujen välinen aika noudattaa eksponenttijakaumaa, laske todennäköisyys, että aika jostakin puhelusta seuraavaan on a) yli 3 tuntia, b) vähemmän kuin 4,5 tuntia.

7.5 Keskimääräinen aika, joka insinööriltä menee systeemissä olevan sähköisen vian korjaamiseen, on 2,7 tuntia. Laske todennäköisyys, että hän korjaa vian vähemmässä ajassa kuin tässä keskimääräisessä ajassa.

7.6 Tietyn konetyypin keskimääräinen vikaantumisväli on 400 tuntia. Laske todennäköisyys, että jonkin yksittäisen koneen vikaantumisväli on a) alle 350 tuntia, b) yli 450 tuntia.

7.7 Laske integroimalla eksponenttijakauman odotusarvo ja varianssi.

7.8 Todista, että a) diskreetin ja b) jatkuvan jakauman varianssin lausekkeet voidaan esittää myös seuraavissa muodoissa, jotka ovat laskennallisesti usein määritelmän mukaisia muotoja paremmat:

$$\boldsymbol{Var(X) = \sum x_i^2 p_i - E(X)^2}, \quad \boldsymbol{Var(X) = \int_{-\infty}^{\infty} x^2 f(x)\, dx - E(X)^2}.$$

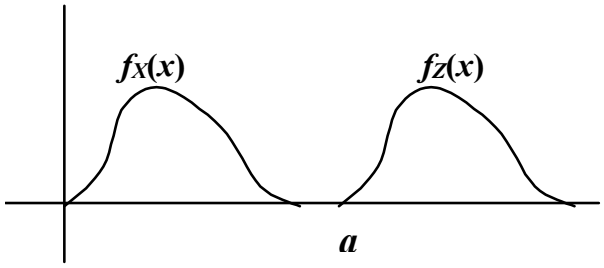
c) Sovellus: Laske $Var(X)$ ja $Sd(X)$, kun X:n tiheysfunktio on

$$f(x) = \begin{cases} 2x, \text{kun } 0 \le x \le 1 \\ 0 \text{ muulloin} \end{cases}.$$

8 Satunnaismuuttujien laskutoimituksia

8.1 Satunnaismuuttujan muunnokset

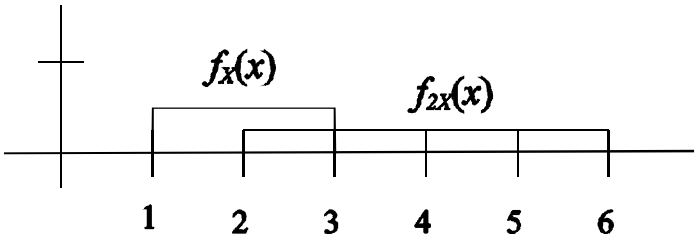
1) Jos satunnaismuuttujaan X lisätään vakio a eli tehdään siihen muunnos $Z = X + a$, merkitsee se, että X:n jokaista arvoa muutetaan a:n verran. Jakauman kuvaajaan (pistetodennäköisyyksiin tai tiheysfunktioon) tämä merkitsee a:n suuruista siirtoa x-suunnassa:



Täten odotusarvo muuttuu a:n verran ja keskihajonta ja varianssi eivät muutu miksikään:

(1) $\boxed{E(X+a) = E(X) + a, \;\; Sd(X+a) = Sd(X), \; Var(X+a) = Var(X)}$.

2) Jos satunnaismuuttuja X kerrotaan esim. 2:lla, merkitsee se, että X:n jokainen arvo kaksinkertaistetaan. Jos esimerkiksi X on tasan jakautunut välille 1 … 3, niin $2X$ jakautuu tasan välille 2 … 6 (vrt. seuraava kuva). Muunnoksessa odotusarvo (keskikohdan x-koordinaatti) kaksinkertaistuu, samoin hajonta. Täten varianssi muuttuu 4-kertaiseksi.



Lisäksi tiheysfunktion kuvaajan täytyy "madaltua", jotta kokonaispinta-ala pysyisi 1:nä.

Yleisesti voidaan odotusarvon ja varianssin määritelmien (tai harjoituksen 7. 8) avulla todistaa, että

(2) $\boxed{E(aX) = aE(X), \;\; Var(aX) = a^2 Var(X), \; Sd(aX) = |a| Sd(X)}$.

3) Tulokset (1) ja (2) esitetään yleensä yhdistettynä seuraavasti:

$$\boxed{E(aX+b) = aE(X) + b, \;\; Var(aX+b) = a^2 Var(X), \; Sd(aX+b) = |a| Sd(X)}$$

Esim. 1 (***Standardointimuunnos***) Jos satunnaismuuttujan X jakauman odotusarvo on μ ja keskihajonta σ, niin satunnaismuuttujan $\boxed{Z = \frac{X-\mu}{\sigma}}$ keskiarvo on 0 ja varianssi 1, sillä

$$E(Z) = E[\frac{1}{\sigma}(X-\mu)] = \frac{1}{\sigma}E(X-\mu) = \frac{1}{\sigma}(\mu-\mu) = 0$$

ja vastaavasti todistetaan (harj.), että $Var(Z) = 1$.

Jos erityisesti X $\sim N(\mu,\sigma^2)$ ***eli X:n jakauma on*** (μ,σ^2)***-normaali, niin Z:n jakauma on (0,1)-normaali***, sillä voidaan todistaa, että Z:n kertymäfunktio on luvussa 2 esitettyä muotoa (luvussa 2 tätä funktiota merkittiin Φ:llä).

*Todistus:

$$F_Z(t) = P(Z \le t) = P(\frac{X-\mu}{\sigma} \le t) = P(X \le \sigma t + \mu)$$

$$= F_X(\sigma t + \mu) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\sigma t+\mu} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \quad \left| \begin{array}{l} sij.\ \frac{x-\mu}{\sigma} = u \\ \therefore \frac{1}{\sigma}dx = du, \\ x: -\infty \ldots \sigma t + \mu \\ u: -\infty \ldots t \end{array} \right.$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-\frac{u^2}{2}}\, du.$$

*Tämän todistuksen alkuosasta takaperin saadaan myös seuraava, luvussa 2 todistamatta esitetty tulos:

$$F_X(\sigma t + \mu) = F_Z(t) \quad \left| \begin{array}{l} \text{merk. } \sigma t + \mu = x \\ \therefore t = \frac{x-\mu}{\sigma} \end{array} \right.$$

$\therefore F_X(x) = F_Z(\frac{x-\mu}{\sigma})$ eli luvun 2 merkinnöin $\underline{\underline{F(x) = \Phi(\frac{x-\mu}{\sigma})}}$.

****Esim. 2*** Määritä satunnaismuuttujan X^2 tiheysfunktio, jos X:n jakauma on standardinormaali (kertymäfunktiona $\Phi(x)$).

On selvää, että $f_{X^2}(t) = 0$, kun $t < 0$ (sillä X^2:n arvot ovat kaikki ≥ 0). Oletetaan siksi, että seuraavassa $t \geq 0$. Lasketaan aluksi kertymäfunktio (luvussa 2 esitetyillä laskutavoilla):

$$F_{X^2}(t) = P(X^2 \leq t) = P(-\sqrt{t} \leq X < \sqrt{t}) = \Phi(\sqrt{t}) - \Phi(-\sqrt{t})$$
$$= \Phi(\sqrt{t}) - [1 - \Phi(\sqrt{t})] = 2\Phi(\sqrt{t}) - 1$$

Tästä saadaan derivoimalla tiheysfunktio. Derivointi vaatii oletuksen $t > 0$. Kohdassa $t = 0$ voidaan asettaa lisämäärittely $f_{X^2}(0) = 0$ vaikuttamatta mitenkään minkään välin todennäköisyyteen. Näin saadaan tulokseksi

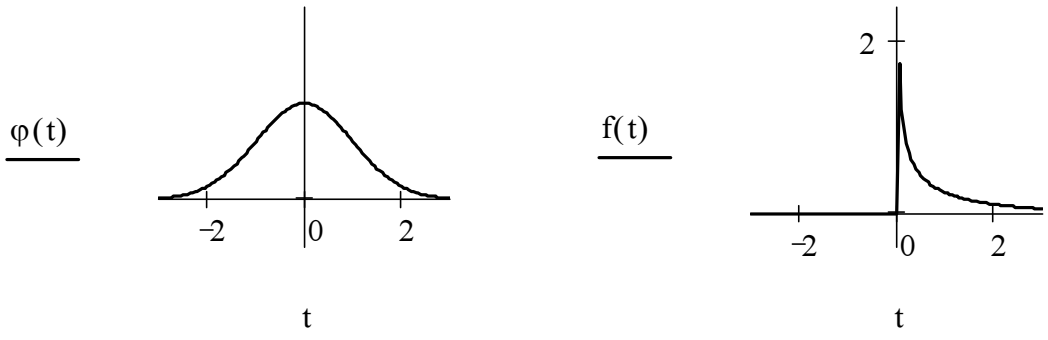
$$f_{X^2}(t) = \begin{cases} 2\varphi(\sqrt{t}) \cdot \dfrac{1}{2\sqrt{t}} = \dfrac{1}{\sqrt{2\pi}} \cdot \dfrac{1}{\sqrt{t}} e^{-\frac{(\sqrt{t})^2}{2}} = \dfrac{1}{\sqrt{2\pi}} \cdot \dfrac{1}{\sqrt{t}} e^{-\frac{t}{2}}, \text{ kun } t > 0 \\ 0, \text{ kun } t \leq 0. \end{cases}$$

Piirrä kuvaaja jollakin matematiikkaohjelmalla. Esim. Mathcadillä laskut ja tulos ovat seuraavan näköiset (dnorm = density function of normal distibution; on aika kummallisen tuntuista, että X^2:n tiheysfunktio menee origon kohdalla äärettömyyteen, vaikka X:n tiheysfunktiolla suurin arvo on vain $1/\sqrt{2\pi}$):

$t := -3, -2.95 .. 3$

$\varphi(t) := \text{dnorm}(t, 0, 1)$

$f(t) := \text{if}\left(t > 0, \dfrac{1}{\sqrt{2\,\pi\,t}} \cdot e^{-0.5\,t}, 0\right)$



8.2 Satunnaismuuttujien summa

Olkoot X ja Y samaan kokeeseen liittyviä satunnaismuuttujia. Kokeen tulosten joukkoa E sanotaan usein ***tapausavaruudeksi*** *(sample space)*.

Summa $X+Y$ saadaan laskemalla yhteen E:n jokaista alkiota vastaavat satunnaismuuttujan arvot:

$$(X+Y)(a) = X(a)+Y(a) \quad \text{(aina kun } a \in E\text{)}.$$

Tuloa XY muodostettaessa taas kerrotaan vastaavat satunnaismuuttujan arvot keskenään.

Esim. 3 Nopanheitossa nopan asennot $a_1, a_2, ..., a_6$ antavat pisteluvut 1, 2, …,6. Seuraavassa voidaan tapausavaruutena pitää joukkoa $E = \{a_1, a_2, ..., a_6\}$ tai myös tulosjoukkoa {1, 2, …, 6}.

Oletetaan, että X ilmoittaa yhdessä nopanheitossa saadun pisteluvun kaksinkertaisena ja Y taas saa arvon 1, jos pisteluku on pariton ja arvon 3, jos pisteluku on parillinen. Täten X:n ja Y:n jakaumat ovat seuraavat (jos noppa on säännöllinen):

x_i:	2	4	6	8	10	12
p_i:	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

y_j:	1	3
q_j:	$\frac{1}{2}$	$\frac{1}{2}$

a) Summalle $S = X + Y$ nopan eri asennot antavat seuraavat arvot:

$$S(a_1) = 2+1 = 3 \quad S(a_2) = 4+3 = 7 \quad S(a_3) = 6+1 = 7$$
$$S(a_4) = 8+3 = 11 \quad S(a_5) = 10+1 = 11 \quad S(a_6) = 12+3 = 15$$

Satunnaismuuttuja S saa siis arvot 3, 7, 11 ja 15. Näistä kaksi keskimmäistä arvoa saadaan kumpikin kahdella nopan asennolla ja reunimmaiset yhdellä. Täten summan S jakauma on

s_k:	3	7	11	15
r_k:	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

Jos ajattelet näiden jakaumien kuvaajia janadiagrammeina, huomaat että kahden "tasaisen diskreetin jakauman" summa ei suinkaan ole "tasainen" vaan "kulmamuotoinen".

b) Tulolle $U = XY$ nopan eri asennot antavat tässä esimerkissä kuusi erisuurta arvoa:

$2 \cdot 1 = 2, 4 \cdot 3 = 12, 6 \cdot 1 = 6,\ 8 \cdot 3 = 24, 10 \cdot 1 = 10, 12 \cdot 3 = 36$. Täten tulon jakauma on

$$u_l: \; 2 \quad 6 \quad 10 \quad 12 \quad 24 \quad 36$$
$$s_l: \; \tfrac{1}{6} \quad \tfrac{1}{6} \quad \tfrac{1}{6} \quad \tfrac{1}{6} \quad \tfrac{1}{6} \quad \tfrac{1}{6}$$

Kertaa miten lasket näiden neljän jakauman odotusarvot ja varianssit (Harj. 8.1).

8.3 2-ulotteinen jakauma

Saman tapausavaruuden E satunnaismuuttujista X ja Y voidaan muodostaa myös *yhdistetty* ***2-ulotteinen jakauma*** (*joint distribution*), jossa satunnaismuuttujana on 2-komponenttinen ***satunnaisvektori*** $[X,Y]$. Sen arvot ovat pareja (x, y), tarkemmin sanoen

$$[X,Y](a) = [X(a), Y(a)] \quad \text{(aina, kun } a \in E\text{)}.$$

Jos X ja Y ovat diskreettejä, niin samoin on niistä yhdistetty 2-ulotteinen jakauma. Siinä parin (x_i, y_j) pistetodennäköisyys on

$$p_{ij} = P(X = x_i, Y = y_j) = \text{todennäköisyys sille, että } X \text{ saa arvon } x_i \text{ ja } Y \text{ saa arvon } y_j$$

Esim. 4 Olkoot X ja Y edellisen esimerkin mukaiset satunnaismuuttujat. Pareja (x_i, y_j) on $6 \cdot 2 = 12$ kappaletta, koska X:llä on 6 arvoa $x_1 = 2, x_2 = 4, ..., x_6 = 12$ ja Y:llä kaksi $y_1 = 1, \quad y_2 = 3$. Satunnaisvektori [X, Y] saa arvoikseen seuraavat kuusi paria:

$$[X,Y](a_1) = (2,1) \quad [X,Y](a_2) = (4,3) \quad [X,Y](a_3) = (6,1)$$
$$[X,Y](a_4) = (8,3) \quad [X,Y](a_5) = (10,1) \quad [X,Y](a_6) = (12,3).$$

Jokaisen tällaisen parin todennäköisyys on = $\frac{1}{6}$ ja muiden parien = 0. Esimerkiksi juuri laskettujen arvojen perusteella

$$p_{22} = P(X = x_2, Y = y_2) = P(X = 4, Y = 3)$$
$$= \text{parin (4,3) todennäköisyys} = \tfrac{1}{6}$$

kun taas esim. $p_{32} = P(X = 6, Y = 3) = 0$.

Pistetodennäköisyyksistä saadaan seuraava taulukko (johon yleensä merkitään vain lukuarvot):

$x_i \setminus y_j$	1	3	
2	$p_{11}=\frac{1}{6}$	$p_{12}=0$	$\sum_j p_{1j}=\frac{1}{6}$
4	$p_{21}=0$	$p_{22}=\frac{1}{6}$	$\sum_j p_{2j}=\frac{1}{6}$
6	$p_{31}=\frac{1}{6}$	$p_{32}=0$	$\sum_j p_{3j}=\frac{1}{6}$
8	$p_{41}=0$	$p_{42}=\frac{1}{6}$	$\sum_j p_{4j}=\frac{1}{6}$
10	$p_{51}=\frac{1}{6}$	$p_{52}=0$	$\sum_j p_{5j}=\frac{1}{6}$
12	$p_{61}=0$	$p_{62}=\frac{1}{6}$	$\sum_j p_{6j}=\frac{1}{6}$
	$\sum_i p_{i1}=\frac{1}{2}$	$\sum_i p_{i2}=\frac{1}{2}$	

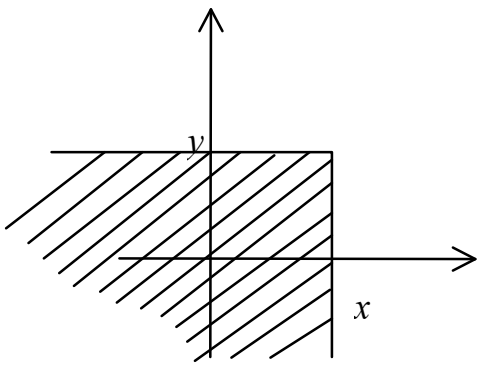
*Viimeisen sarakkeen/rivin luvut yhdessä ensimmäisen sarakkeen/rivin lukujen kanssa antavat muuttujien X ja Y jakaumat, joita sanotaan tämän 2-ulotteisen jakauman ***reunajakaumiksi***. Näiden pistetodennäköisyydet ovat

$$p_i=\sum_{j=1}^{2} p_{ij}\ \ (i=1,2,...,6), \qquad q_j=\sum_{i=1}^{6} p_{ij}\ \ (j=1,2).$$

Jatkuvalla 2-ulotteisella jakaumalla kertymäfunktio on kahden muuttujan funktio

$$F(x,y)=P(X\le x, Y\le y).$$

Geometrisesti tämä ilmaisee, kuinka suuri osa todennäköisyysmassasta on sijoittunut (kertynyt) viereiseen kulma-alueeseen.



Suorakulmioalueen todennäköisyys saadaan kertymäfunktion arvojen seuraavanlaisena lausekkeena (jonka voit perustella em. kuvan tapaan):

$$P(a < X \le b, c < Y \le d) = F(b,d) - F(a,d) - F(b,c) + F(a,c).$$

Tiheysfunktio saadaan kertymäfunktion toisena derivaattana:

$$f(x,y) = F_{xy}(x,y).$$

Kääntäen kertymäfunktion arvoja saadaan tiheysfunktiosta integroimalla:

$$F(x,y) = \int_{-\infty}^{x} \left[\int_{-\infty}^{y} f(u,v)\, dv \right] du$$

(sillä edellinen kulma-alue T on $\begin{cases} -\infty < u \le x \\ -\infty < v \le y \end{cases}$).

Suorakulmioalueen todennäköisyys on tiheysfunktion integraali

$$P(a < X \le b, c < Y \le d) = \int_{a}^{b} \left[\int_{c}^{d} f(u,v)\, dv \right] du.$$

*Reunajakaumien (satunnaismuuttujien X ja Y jakaumien) tiheysfunktioiden arvoja saadaan yhdistetyn jakauman tiheysfunktiosta "summaamalla" (vastaavasti kuin pistetodennäköisyyksiä edellä):

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx$$

8.4 Satunnaismuuttujien riippumattomuus

Saman tapausavaruuden E satunnaismuuttujia X ja Y sanotaan ***riippumattomiksi***, jos ne täyttävät kaikilla x:n ja y:n reaaliarvoilla ehdon

$$P(X \le x, Y \le y) = P(X \le x) \cdot P(Y \le y).$$

Tällöin 2-ulotteisen jakauman kertymä- ja tiheysfunktiot (tai diskreetissä tapauksessa pistetodennäköisyydet) saadaan X:n ja Y:n vastaavien suureiden tulona:

$$F(x,y) = F_X(x) \cdot F_Y(y), \qquad f(x,y) = f_X(x) \cdot f_Y(y).$$

Riippumattomuus merkitsee suurin piirtein sanottuna, että yhden satunnaismuuttujan arvon tunteminen ei vaikuta toiseen satunnaismuuttujaan liittyviin todennäköisyyksiin.

Esim. 5** Voidaan todistaa, että riippumattomien satunnaismuuttujien summan $S = X + Y$ tiheysfunktio saadaan X:n ja Y:n tiheysfunktioiden ns. ***konvoluutiona, ts. seuraavalla integraalilla:

$$f_S(u) = \int_{-\infty}^{\infty} f_X(x) \cdot f_Y(u-x)\, dx\,.$$

Tämän avulla on aika helppoa laskea esim. kahden riippumattoman eksponenttijakauman summan tiheysfunktio. Jos sekä X:llä että Y:llä on sama eksponenttijakauma, yhteisenä parametrinä λ, niin integraalin arvoksi ($2X$:n tiheysfunktioksi) tulee

$$\lambda^2 u \cdot e^{-\lambda u},\ \text{kun } u \geq 0\,.$$

Jos taas parametrit ovat $\lambda \neq \mu$, niin vastaava arvo on

$$\frac{\lambda\mu}{\mu - \lambda} \cdot (e^{-\lambda \cdot u} - e^{-\mu \cdot u})\,.$$

8.5 Odotusarvoa ja varianssia koskevia tuloksia

Kohdassa 8.1 olivat esillä mm. seuraavat tulokset:

(1) $\quad E(aX) = aE(X), \quad Var(aX) = a^2 Var(X),$

(2) $\quad E(aX + b) = aE(X) + b, \quad Var(aX + b) = a^2 Var(X)\,.$

Satunnaismuuttujien summaa ja tuloa koskevat seuraavat tulokset, joista *ensimmäinen (summan odotusarvotulos) on voimassa aina, kun taas kolme muuta vaativat, että X ja Y ovat riippumattomia:*

(3) $\quad \boxed{E(X+Y) = E(X) + E(Y)}\,, \quad E(XY) = E(X) \cdot E(Y),$

(4) $\quad \boxed{Var(X+Y) = Var(X) + Var(Y)}\,, \quad Var(XY) = Var(X) \cdot Var(Y)\,.$

*Todistetaan näytteenä tuloksista (3) edellinen jatkuvien jakaumien tapauksessa:

$$\begin{aligned} E(X+Y) &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} (x+y) f(x,y)\, dy \right] dx \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x f(x,y)\, dy \right] dx + \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y f(x,y)\, dy \right] dx \end{aligned}$$

Otetaan edellisessä integraalissa vakiotekijä sisemmästä integraalista ulos ja vaihdetaan jälkimmäisessä integraalissa summausjärjestystä (tämä voidaan tehdä, koska rajat ovat vakioita), jonka jälkeen y voidaan vakiona ottaa sisemmästä integraalista ulos. Näin saadaan seuraava tulos,

jossa sisemmät integraalit ovat reunajakaumien tiheysfunktioita (vrt. kohdan 8.3 loppu):

$$E(X+Y) = \int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f(x,y)\,dy\right]dx + \int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} f(x,y)\,dx\right]dy$$

$$= \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx + \int_{-\infty}^{\infty} y \cdot f_Y(y)\,dy = E(X) + E(Y).$$

Esim. 6 a) Tuloksissa (3) ja (4) yhteenlaskettavia tai kerrottavia satunnaismuuttujia voi olla useampiakin kuin kaksi:

$$E(2X_1 + 3X_2 - 5X_3) = 2E(X_1) + 3E(X_2) - 5E(X_3).$$

b) Oletetaan, että satunnaismuuttujien $X_1, ..., X_n$ odotusarvot ovat $\mu_1, ..., \mu_n$. Silloin (3):n ja (2):n nojalla

(5) $$E(\sum(X_i - \mu_i)) = \sum E(X_i - \mu_i) = \sum(\mu_i - \mu_i) = 0.$$

Riippumattomilla satunnaismuuttujilla vastaavasti

(6) $$Var(\sum(X_i - \mu_i)) = \sum Var(X_i - \mu_i) \overset{(2)}{=} \sum Var(X_i).$$

c) Jos satunnaismuuttujilla $X_1, ..., X_n$ on kaikilla sama odotusarvo μ ja sama varianssi σ^2, niin niiden aritmeettisen ***keskiarvon*** $\overline{X} = \frac{X_1 + ... + X_n}{n}$ vastaavat suureet ovat

$$\underline{\underline{E(\overline{X})}} = \frac{1}{n} \cdot n \cdot \mu \underline{\underline{= \mu}}$$

ja riippumattomilla satunnaismuuttujilla vastaavasti

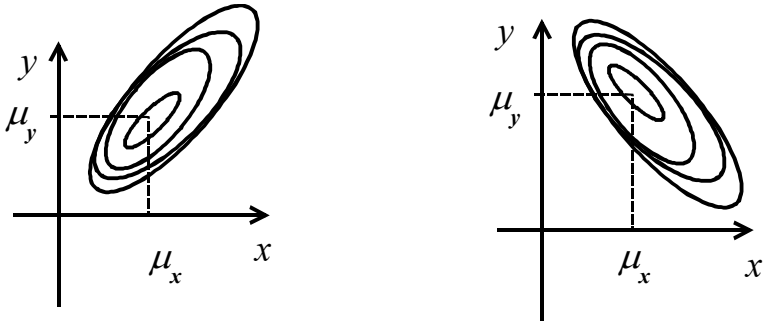
$$\underline{\underline{Var(\overline{X})}} = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \underline{\underline{\frac{\sigma^2}{n}}}.$$

8.6 Kovarianssi ja korrelaatio

Seuraavassa on tarkoituksena esitellä kaksi läheisessä yhteydessä toisiinsa olevaa tunnuslukua, ***kovarianssi*** ja ***korrelaatiokerroin***, jotka *mittaavat kahden satunnaismuuttujan välistä vuorovaikutusta, riippuvuutta toisistaan*.

Yleisesti satunnaismuuttujien X ja Y eri ominaisuuksista, siis myös niiden keskinäisestä vuorovaikutuksesta, antavat kuvan satunnais-

vektorin [X, Y] kertymäfunktio ja pistetodennäköisyys- tai tiheysfunktio. Geometrisesti tiheysfunktion kuvaaja on pinta xyz-koordinaatistossa ja tästä pinnasta saadaan käsitys piirtämällä pinnan korkeuskäyriä, siis käyttämällä topograafiesitystä kuten kartoissa.



Jos tiheysfunktion $f(x, y)$ kuvaaja on vasemmanpuoleisen kuvan tapainen, niin X:n ja Y:n arvojen välillä on (positiivista) riippuvuutta, sillä niillä on taipumusta poiketa odotusarvostaan samaan suuntaan: X:n kasvaessa myös Y keskimäärin kasvaa. Oikeanpuoleisessa kuvassa taas X:n ja Y:n arvoilla on taipumus poiketa odotusarvoistaan eri suuntiin: X:n kasvaessa Y keskimäärin vähenee.

Ääritapaus on se, että X:n arvot määräävät täysin Y:n arvot, ts. Y on X:n funktio. Tällaisesta *funktionaalisesta* riippuvuudesta on kyse esim. jos X:ää ja Y:tä sitoo yhtälö $Y = X^2$. Tällöin siis todennäköisyysmassa on kaikki sijoittunut vastaavalle käyrälle $y = x^2$. Jos käyrä on suora, kyseessä on *lineaarinen* riippuvuus.

Toinen ääritapaus on se, että X:llä ja Y:llä ei ole minkäänlaista vaikutusta toisiinsa, jolloin ne ovat riippumattomia myös todennäköisyyslaskennan mielessä ja siten

$$f(x, y) = f_X(x) f_Y(y).$$

Tässä tapauksessa satunnaisvektorin [X, Y] jakaumasta antavat kaiken tiedon (1-ulotteiset) X:n ja Y:n jakaumat.

Diskreetillä 2-ulotteisilla jakaumilla edellisten kuvien tilalla ovat (x_i, y_i)-pisteistä muodostuvat ***sirontakuviot***, joissa pisteet sijoittuvat enemmän tai vähemmän jonkin käyrän (esim. regressiosuoran) läheisyyteen.

Yhdellä diskreetillä satunnaismuuttujalla odotusarvo ja varianssi määriteltiin seuraavasti:

$$\mu = E(X) = \sum x_i p_i, \qquad \sigma^2 = Var(X) = \sum (x_i - \mu)^2 p_i$$

ja jatkuvalla satunnaismuuttujalla

$$\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx \qquad \sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx,$$

joten kummassakin tapauksessa varianssi antaa odotusarvosta laskettujen poikkeamien $X - \mu$ neliöiden odotusarvon, siis

$$Var(X) = E[(X-\mu)^2].$$

Vastaava suure, joka mittaa kahden satunnaismuuttujan välistä keskimääräistä poikkeamaa odotusarvosta, on ***kovarianssi***

(7) $$\boxed{Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]}.$$

Jos satunnaismuuttujat ovat diskreettejä/jatkuvia, kovarianssi on kaksinkertainen summa/integraali:

$$Cov(X,Y) = \sum_i \sum_j (x_i - \mu_X)(y_j - \mu_Y) p_{ij},$$

$$Cov(X,Y) = \int_{-\infty}^{\infty} [\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y)\, dy]\, dx.$$

Kovarianssin lauseke (7) voidaan muuntaa edellä olleiden tulosten (2) ja (1) avulla seuraavasti:

$$\begin{aligned} \underline{\underline{Cov(X,Y)}} &= E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y \\ &= E(XY) - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y = \underline{\underline{E(XY) - \mu_X \mu_Y}}. \end{aligned}$$

Jos erityisesti X ja Y ovat riippumattomia, niin kovarianssi on = 0, sillä

$$E(XY) - \mu_X \mu_Y = E(X)E(Y) - \mu_X \mu_Y = \mu_X \mu_Y - \mu_X \mu_Y = 0.$$

Tällöin sanotaan, että satunnaismuuttujat X ja Y eivät korreloi.

Määritelmä:

Satunnaismuuttujien X ja Y (lineaarinen) ***korrelaatiokerroin***

$$\boxed{r(X,Y)=\frac{Cov(X,Y)}{\sigma_X\cdot\sigma_Y}},$$

missä esim. $\sigma_X=\sqrt{Var(X)}$ (= X:n keskihajonta).

Edellä olevan mukaan, *jos X ja Y ovat riippumattomia, niin korrelaatiokerroin $r = 0$*. Esimerkillä voidaan osoittaa, että tämä tulos ei kuitenkaan ole käännettävissä. Edelleen voidaan todistaa, että *korrelaatiokertoimen arvo on aina välillä $-1\ldots1$ ja äärimmäiset arvot $+1$ tai -1 se saa silloin kun X:n ja Y :n välillä on lineaarinen riippuvuus $Y=aX+b$.*

Harjoituksia

A, B

8.1 Laske Esimerkin 3 mukaisille satunnaismuuttujille X, Y, $X+Y$ ja XY odotusarvot ja varianssit. Totea, että tässä esimerkissä lait

$$Var(X+Y)=Var(X)+Var(Y),\quad Var(XY)=Var(X)\cdot Var(Y)$$

$E(XY)=E(X)\cdot E(Y)$ eivät pidä paikkaansa, mutta sen sijaan laki $E(X+Y)=E(X)+E(Y)$ on voimassa tässä esimerkissä (kuten aina).

8.2 Laske Esimerkin 3 mukaisille satunnaismuuttujille X, Y ja $X+Y$ toisten potenssien odotusarvot $E(X^2), E(Y^2), E[(X+Y)^2]$.

8.3 Oletetaan, että satunnaismuuttujalla X on jakauma

x_i:	-2	-1	1	3
p_i	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

ja $Y=X^2$. Määritä Y:n jakauma.

8.4 Oletetaan, että X:n ja Y:n yhteisjakauma on seuraava:

$X \setminus Y$	–3	2	4	*summa*
1	0,1	0,2	0,2	0,5
3	0,3	0,1	0,1	0,5
summa	0,4	0,3	0,3	

a) Mitkä ovat X:n ja Y:n jakaumat?

b) Laske μ_X ja μ_Y sekä $E(XY)$, $E(X^2)$ ja $E(Y^2)$.

c) Osoita todennäköisyyden $P(X = 1, Y = -3)$ avulla, että X ja Y eivät ole riippumattomia.

d) Laske keskihajonnat σ_x ja σ_y.

e) Laske X:n ja Y:n välinen kovarianssi ja korrelaatiokerroin (joita käsitellään kohdassa 8.6).

8.5 Todista a) diskreettien, b) jatkuvien satunnaismuuttujien tapauksessa kohdan 8.1 tulokset (2).

8.6 Todista, että standardoidun satunnaismuuttujan $Z = \dfrac{X - \mu}{\sigma}$ varianssi on = 1.

C

8.7 Laske Esimerkissä 5 mainittujen riippumattomien eksponenttijakaumien summan tiheysfunktio kyseisessä esimerkissä esitettyä "konvoluutiointegraalia" käyttäen. Ohje: kun $x < 0$, niin $f_X(x) = 0$ ja kun $x > u$, niin $u - x < 0$ ja siten $f_Y(u - x) = 0$. Siten integroimisväliksi jää vain 0 … u.

8.8 a) Johdetaan Esimerkissä 5 mainittu ***konvoluutiointegraali***.

Satunnaismuuttujan X kertymäfunktio x:n arvolla u lasketaan X:n tiheysfunktion $f(x)$ avulla seuraavasti:

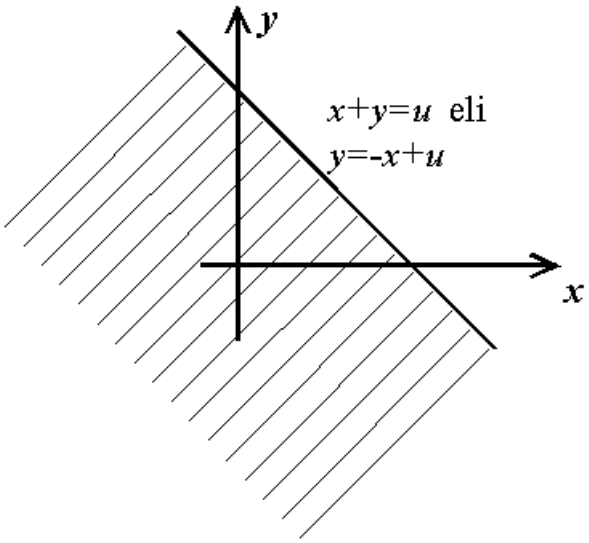
$$F(u) = P(X \le u) = \int_{-\infty}^{u} f(x)\,dx\,.$$

Sen tulkinta pienten differentiaalien menetelmällä on, että "summataan" dx:n levyisiä todennäköisyysmassaliuskoja $f(x)dx$ aina arvoon u saakka. Vastaavasti summan $X + Y$ kertymäfunktio

arvolla u saadaan vektorin $[X,Y]$ tiheysfunktion $f(x,y)$ avulla, kun "summataan" xy-tason neliöstä $dydx$ pinnalle $z = f(x,y)$ ulottuvia todennäköisyysmassapilareita $f(x,y)dydx$ yhteen:

$$F_{X+Y}(u) = P(X+Y \le u) = \iint_T f(x,y)dydx\,.$$

Integroimisalue T muodostuu kaikista niistä (x,y)-pareista, jotka täyttävät ehdon $x + y \le u$. Tämä ehto voidaan esittää siten, että x saa arvoja $-\infty$:stä $+\infty$:ään ja kullakin kiinteällä x:n arvolla y saa kaikki ehdon $y \le u - x$ täyttävät arvot eli kaikki arvot $-\infty$:stä suoralle $y = -x + u$. Siis



$$T: \begin{cases} -\infty < x < \infty \\ -\infty < y \le u - x \end{cases}. \quad \text{Täten} \quad F_{X+Y}(u) = \int_{-\infty}^{\infty} [\int_{-\infty}^{u-x} f(x,y)\,dy]\,dx\,.$$

Tiheysfunktio saadaan derivoimalla tämä integraali muuttujan u suhteen. Integraalissa on "summattavia" aina $-\infty$:stä $+\infty$:ään saakka. Summataan integraalin jokainen "termi" erikseen:

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} [\frac{d}{du} \int_{-\infty}^{u-x} f(x,y)dy]\,dx\,.$$

Sisempi integraali derivoidaan ylärajalla olevan parametrin u suhteen siten, yläraja $u - x$ sijoitetaan integroimismuuttujan y tilalle integroitavaan funktioon $f(x,y)$ ja tulos kerrotaan ylärajan (sisäfunktion) $u - x$ derivaatalla, joka on tässä esimerkissä = $1 - 0 = 1$. Siten

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f(x, u-x)\,dx\,.$$

Jos X ja Y ovat riippumattomia, niin tiheysfunktio $f(x,y)$ voidaan esittää X:n ja Y:n tiheysfunktioiden tulona, jolloin edellinen tulos muuttuu Esimerkin 5 mukaiseksi ***konvoluutiointegraaliksi***

$$f_{X+Y}(u) = \int_{-\infty}^{\infty} f_X(x) \cdot f_Y(u-x)\, dx\,.$$

b) Johda vastaavasti tulon *XY* ja osamäärän *Y/X* kertymäfunktion ja tiheysfunktion lausekkeet, olettaen, että *X*:n tiheysfunktio on = 0, kun $x < 0$ (Erona esim. tulolla on lähinnä vain se, että ylärajan *u*/*x* derivaatta ei ole = 1 vaan 1/*x*. Myös integroimisalue on erinäköinen.)

c) Laske väleille 0 …1 tasaisesti jakautuneiden, riippumattomien satunnaismuuttujien summan jakauma. Ohje: $f_X(x)$ ja $f_Y(u-x)$ saavat 0:sta eroavan arvon =1 vain kun $0 \le x \le 1$ ja $0 \le u - x \le 1$ eli $u - 1 \le x \le u$. Nämä ehdot merkitsevät *ux*-koordinaatistossa suorien $x = 0,\ x = 1,\ x = u - 1,\ x = u$ rajaamaa suunnikasaluetta (piirrä kuva). Tästä saadaan integroimisrajat *x*:n suhteen: Jos $0 \le u \le 1$, niin *x* muuttuu 0 … *u*. Jos $1 \le u \le 2$, niin *x* muuttuu $u - 1 \ldots 1$.

*d) Laske seuraavien jakaumien tulon jakauma:

$f_X(x) = xe^{-x^2/2}$, kun $x \ge 0$ (ja = 0 muulloin),

$f_Y(y) = 1/(\pi\sqrt{1-y^2})$, kun $-1 < y < 1$ (ja = 0 muulloin).

9 Normaalijakauman yhteys muihin jakaumiin

9.1 Keskeinen raja-arvolause

Monissa todennäköisyyslaskennan sovelluksissa lähtökohtana on usean satunnaismuuttujan X_i summa $X_1 + ... + X_n$. Sitä koskee seuraava tulos (jota ei todisteta):

Lause 1 *Jos satunnaismuuttujat X_i ovat riippumattomia ja niiden jakaumat ovat (μ_i, σ_i^2)-normaaleja, niin summan $X_1 + ... + X_n$ jakauma on tarkalleen $(\mu_1 + ... + \mu_n, \sigma_1^2 + ... + \sigma_n^2)$-normaali.*

T*odennäköisyyslaskennan keskeinen raja-arvolause* on edellisen lauseen yleistys ja se voidaan esittää esim. seuraavassa muodossa:

Lause 2 ***Kun n kpl <u>riippumattomia</u> satunnaismuuttujia lasketaan yhteen, niin summan jakauma on likimain normaali, kun n on riittävän suuri.***

Huomaa, että tämä lause (jonka todistus sivuutetaan) on voimassa *riippumatta siitä millaisia yksittäisten satunnaismuuttujien jakaumat ovat* (kunhan vain nämä jakaumat täyttävät tietyt, käytännössä lähes aina toteutuvat ehdot). Komponenttijakaumien ei siis tarvitse olla esim. normaalijakaumia.

Nämä lauseet näyttävät, miksi normaalijakauma on niin tärkeä sekä teoreettisesti että myös käytännön kannalta.

Esim. 1 ***Normaalijakauma binomijakauman likiarvona***. Toistetaan koetta n kertaa. Merkitään onnistumistodennäköisyyttä yhdessä toistossa p:llä ja epäonnistumistodennäköisyyttä q:lla ($q = 1 - p$). Jos X ilmoittaa onnistumisten määrän näissä n:ssä kokeessa, niin $X \sim Bin(n, p)$ ja

(1) $$p_k = P(X = k) = \binom{n}{k} p^k q^{n-k} \quad (k = 0, 1, ..., n).$$

Binomijakaumalla on $E(X) = n \cdot p$ ja $Var(X) = n \cdot p \cdot q$ (kuten aikaisemmin on mainittu).

a) X voidaan hajottaa n:n riippumattoman satunnaismuuttujan summaksi $X = X_1 + ... + X_n$, missä X_i ilmoittaa i:nnessä

kokeessa saadun tuloksen (onnistumisen = 1 tai epäonnistumisen = 0). Jokaisen X_i:n jakauma on $Bin(1, p)$ (yksi koe, jossa onnistumisen todennäköisyys on p). Täten $E(X_i) = 1 \cdot p = p$ ja $Var(X_i) = 1 \cdot p \cdot q = pq$. Keskeisen raja-arvolauseen mukaan X:n jakauma on likimain normaali, parametreinä

$$\mu = E(X_1 + ... + X_n) = E(X_1) + ... + E(X_n) = p + ... + p = np$$

ja riippumattomuuden nojalla

$$\begin{aligned} \sigma^2 &= Var(X_1 + ... + X_n) \\ &= Var(X_1) + ... + Var(X_n) = pq + ... + pq = npq \end{aligned}$$

Siis likimain $\boldsymbol{X \sim N(np, npq)}$.

Normaalijakauman kertymäfunktion arvoja voidaan laskea standardinormaalin jakauman kertymäfunktion avulla (vrt. kohdan 8.1 esimerkin 1 loppuosa):

$$F_X(t) = P(X \le t) \approx \Phi\left(\frac{t - \mu}{\sigma}\right) = \Phi\left(\frac{t - np}{\sqrt{npq}}\right).$$

Jos n on suuri, välien todennäköisyyksiä on helpompi laskea tämän tuloksen avulla kuin tuloksella (1). Seuraavassa näytetään, miten tätä tulosta käytetään ½ yksikön ns. *jatkuvuuskorjauksella*.

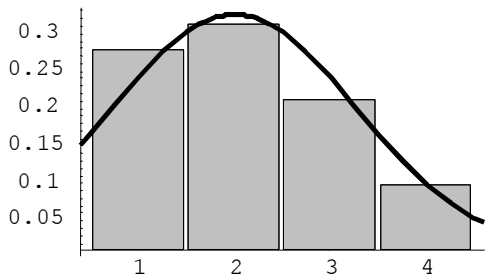
b) Jos X:n jakauma on binomijakauma, parametreinä esimerkiksi $n = 10$, $p = \frac{1}{5}$ (kokeiden määrä ja onnistumistodennäköisyys), niin

$$\mu = np = 10 \cdot \tfrac{1}{5} = 2 \text{ ja } \sigma = \sqrt{10 \cdot \tfrac{1}{5} \cdot \tfrac{4}{5}} = \sqrt{\tfrac{8}{5}}.$$

Esimerkiksi välin [1, 4] todennäköisyys on

$$P(1 \le X \le 4) = p_1 + p_2 + p_3 + p_4.$$

Summaa voidaan kuvata janadiagrammilla.
Jokainen jana voidaan korvata pylväällä, jonka korkeus = p_i ja kanta = 1. Kun tällaiset pylväät piirretään p_i-janojen



ympärille, saadaan porraskuvio, joka alkaa t:n arvosta ½ ja päättyy arvoon 4½. Tämä porraskuvio voidaan korvata likimain normaalijakauman tiheysfunktion ja t-akselin välisellä alueella välillä [½,4½]. Siis tässä esimerkissä

$$P(1 \le X \le 4) \approx \Phi\left(\frac{4.5-2}{\sqrt{\frac{8}{5}}}\right) - \Phi\left(\frac{0{,}5-2}{\sqrt{\frac{8}{5}}}\right)$$
$$\approx \Phi(1{,}976) - \Phi(-1{,}186)$$
$$= \Phi(1{,}976) - [1 - \Phi(1{,}186)]$$
$$= \Phi(1{,}976) + \Phi(1{,}186) - 1$$
$$\approx 0{,}9759 + 0{,}8822 - 1 \approx 0{,}858.$$

Binomijakaumaa käyttämällä saataisiin tulos 0,8598.

Vastaavasti yleisesti

$$\boxed{P(a \le X \le b) \approx \Phi\left(\frac{b+\frac{1}{2}-\mu}{\sigma}\right) - \Phi\left(\frac{a-\frac{1}{2}-\mu}{\sigma}\right)}$$

Huomaa, että normaalijakaumaa voidaan siis käyttää myös diskreettien jakaumien (kuten binomijakauman ja Poisson-jakauman) approksimaationa (likiarvona).

Seuraavissa esimerkeissä esitellään eräitä muita keskeisen raja-arvo-lauseen käyttöön liittyviä tuloksia. Nämä asian ymmärtämistä syventävät tulokset voidaan niin haluttaessa ohittaa ja voidaan siirtyä suoraan otoksen vastaaviin tuloksiin.

Esim. 2 Tarkastellaan riippumattomia satunnaismuuttujia X_i, joilla kullakin on tietty odotusarvo μ_i ja varianssi $\sigma_i{}^2$ $(i = 1, 2, \ldots)$.

a) Vähennetään ensin jokaisesta satunnaismuuttujasta oma odotusarvonsa, jolloin saatujen satunnaismuuttujien odotusarvot ovat = 0. Lasketaan sitten näin saadut satunnaismuuttujat yhteen, ts. muodostetaan uusi satunnaismuuttuja

$$S_n = (X_1 - \mu_1) + \ldots + (X_n - \mu_n).$$

Kohdan 8.5 esimerkin 6 mukaan

$$\underbrace{E(S_n)}_{\text{merk. } \mu} = 0 \quad \text{ja} \quad \underbrace{Var(S_n)}_{\text{merk. } \sigma^2} = \sigma_1{}^2 + \ldots + \sigma_n{}^2.$$

Keskeisen raja-arvolauseen mukaan (jos n on riittävän suuri) ***summan*** $\boldsymbol{S_n = (X_1 - \mu_1) + \ldots + (X_n - \mu_n)}$ ***jakauma on likimain (***$\boldsymbol{0,\ \sigma_1^2 + \ldots + \sigma_n^2}$***)-normaali***.

b) Suurilla n:n arvoilla summan S_n varianssi $\sigma_1^2 + \ldots + \sigma_n^2$ voi olla suuri vaikka jokaisen komponenttijakauman X_i varianssi σ_i^2 olisikin pieni. Mutta variansseiltaan suuret jakaumat eivät ole käytännön kannalta kovinkaan merkittäviä, koska jakauma on "laajalle hajaantunut" ja siksi "matala". Tämän vuoksi tarkastellaankin usein S_n:n sijasta vastaavaa standardoitua satunnaismuuttujaa

$$Z_n = \frac{S_n - \mu}{\sigma} = \frac{S_n - 0}{\sqrt{Var(S_n)}} = \frac{S_n}{\sqrt{Var(S_n)}}.$$

Kohdan 8.1 Esimerkin 1 mukaan ***standardoidun satunnaismuuttujan*** $\boldsymbol{Z_n}$ ***jakauma on likimain (0,1)-normaali***.

****Esim. 3*** a) Tarkastellaan <u>riippumattomia</u> satunnaismuuttujia X_i, joilla jokaisella on sama odotusarvo μ ja sama varianssi σ^2. Silloin summan

$$S_n = X_1 + \ldots + X_n$$

odotusarvo ja varianssi ovat (vrt. s. 64)

$$E(X_1 + \ldots + X_n) = E(X_1) + E(X_n) = n \cdot \mu,$$

$$Var(X_1 + \ldots + X_n) \overset{riippumattomuus}{=} Var(X_1) + \ldots + Var(X_n) = n\sigma^2.$$

Keskeisen raja-arvolauseen mukaan ***summan*** $\boldsymbol{X_1 + \ldots + X_n}$ ***jakauma on riittävän suurilla*** $\boldsymbol{n}$***:n arvoilla likimain*** $\boldsymbol{(n\mu, n\sigma^2)}$***-normaali***. Siten sen kertymäfunktiolle saadaan likiarvo standardinormaalin jakauman kertymäfunktion avulla (vrt. kohdan 8.1 esimerkin 1 loppuosa):

$$F_{X_1 + \ldots + X_n}(t) = P(\sum X_i \le t) \approx \Phi\left(\frac{t - n\mu}{\sqrt{n}\,\sigma}\right).$$

b) Jos riippumattomilla (mutta ei välttämättä normaalisti jakautuneilla) satunnaismuuttujilla X_i on kullakin oma

odotusarvonsa μ_i ja varianssinsa σ_i^2, niin riittävän suurilla n:n arvoilla summan $X_1 + ... + X_n$ jakauma on keskeisen raja-arvolauseen mukaan likimain normaali ja vastaavasti kuin a)-kohdassa voidaan todistaa että sen odotusarvo on $\mu_1 + ... + \mu_n$ ja varianssi on $\sigma_1^2 + ... + \sigma_n^2$.

c) *Jos riippumattomuuden lisäksi jokaisen X_i:n jakauma on normaali, niin lauseen 1 mukaan summan jakauma on tarkalleen normaali (n:n arvosta riippumatta).*

9.2 Otoskeskiarvo

Oletetaan, että otoskoko on *n*. Jos X_i antaa otoksen *i*:nnen havainnon arvon x_i, niin tämä arvo vaihtelee otoksesta riippuen. Siten jokainen X_i on satunnaismuuttuja. Muodostetaan niistä uusi satunnaismuuttuja

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

Sitä sanotaan ***otoskeskiarvoksi.*** Keskeisen raja-arvolauseen mukaan ***otoskeskiarvon jakauma on likimain normaali*** (koska kyseessä on satunnaismuuttujien $\frac{1}{n}X_1 + ... + \frac{1}{n}X_n$ summa).

Jos populaatio on suuri (tai otanta tapahtuu takaisinpanolla), niin satunnaismuuttujat X_i ovat riippumattomia ja jokaisella niistä on sama odotusarvo μ ja sama hajonta σ kuin koko populaatioon liittyvällä satunnaismuuttujalla X (joka ilmoittaa esim. kaikkien valmistettavien tuotteiden kestoiät). Otoskeskiarvon $\overline{X}$ odotusarvo ja varianssi ovat tällöin (kuten jo kohdassa 8.5 todettiin)

$$E(\overline{X}) = \frac{1}{n}(\mu + ... + \mu) = \mu, \quad Var(\overline{X}) = \frac{1}{n^2}(\sigma^2 + ... + \sigma^2) = \frac{\sigma^2}{n}.$$

Siis likimain (riittävän suurilla *n*:n arvoilla)

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}). \tag{2}$$

Otantateoriassa koko aineiston (populaation) keskiarvoa ja varianssia arvioidaan (estimoidaan) otoskeskiarvon $\overline{X}$ ja ns. otosvarianssin avulla. Kysymys on mm. siitä, kuinka hyvin otoskeskiarvo $\overline{X}$ tai otoksesta laskettu varianssi tai keskihajonta (= varianssin neliöjuuri) vastaavat

koko populaation vastaavia suureita. Tällöin koko populaation suureet (keskiarvo, varianssi jne.) ovat testattavia ***parametrejä*** ja otoksen vastaavat suureet näiden parametrien ***estimaattoreita.***

Koska tuloksen (2) mukaan *otoskeskiarvolla on sama odotusarvo* μ *kuin koko populaatioon liittyvällä satunnaismuuttujalla X*, sanotaan, että otoskeskiarvo $\overline{X}$ on *X*:n odotusarvon ***harhaton*** *(unbiased)* ***estimaattori***. Harhattomuuden vuoksi otoskeskiarvoa voidaan useissa tapauksissa käyttää koko populaation odotusarvon (keskiarvon, jota ei useinkaan tunneta) sijaan.

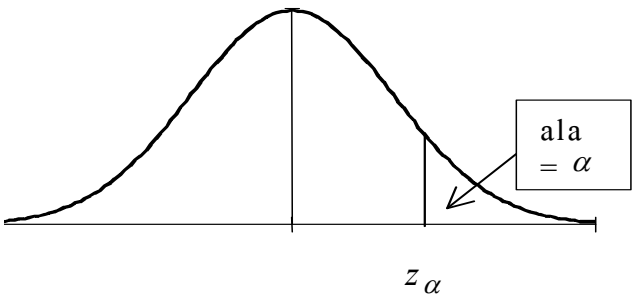
Otoskeskiarvon $\overline{X}$ keskihajonta on:

$$Sd(\overline{X}) = \sqrt{Var(\overline{X})} = \frac{\sigma}{\sqrt{n}}.$$

Lukua $\frac{\sigma}{\sqrt{n}}$ sanotaan ***keskiarvon keskivirheeksi.*** Sen voidaan ajatella ilmoittavan, paljonko otoskeskiarvossa on keskimäärin poikkeamaa koko populaation keskiarvoon verrattuna. Otoskoon *n* kasvaessa tämä luku $\to 0$. Tästä syystä sanotaan, että *otoskeskiarvo* $\overline{X}$ *on koko populaatioon liittyvän satunnaismuuttujan X odotusarvon* ***tarkentuva*** *(consistent)* ***estimaattori***. Suurten otosten keskiarvot poikkeavat siis keskimäärin hyvin vähän koko populaation keskiarvosta.

Tarkastellaan nyt yleisesti (0, 1)-normaalista satunnaismuuttujaa *Z*. Merkitään z_α:lla *Z*:n tiheysfunktion $\varphi(x)$ kuvaajan kohtaa, jonka jäljessä (jakauman loppupäässä) olevan todennäköisyysmassan määrä on α. Siis

$$P(Z \geq z_a) = \alpha, \qquad P(Z \leq z_a) = 1 - \alpha\,.$$



Lukua z_α sanotaan arvoa α vastaavaksi ***kriittiseksi arvoksi***. Luku α on nimeltään päättelyn tms. ***merkitsevyystaso***, ***riskitaso***, *p-arvo,* engl. significance level, *p*-value, tail probability (tail = häntä). Yleisimmin käytetyt α:n arvot (merkitsevyystasot, *p*-arvot) ovat 0,05 (5 %), 0,01 (1 %) ja 0,001 (0,1 %). Luku $1-\alpha$ on vastaava ***luottamustaso*** ja sen tavallisimmat arvot ovat 95 % (0,95), 99 % (0,99) ja 99,9 % (0,999).

Esim. 4 Eri luottamustasoja vastaavia kriittisiä arvoja voidaan laskea käyttämällä standardinormaalin jakauman kertymäfunktion taulukkoa "takaperin" tai esim. matematiikkaohjelmien avulla.

a) Lasketaan 80 %:n luottamustasoa (eli α:n arvoa 0,2) vastaava z_α:n arvo käyttämällä standardinormaalin jakauman kertymäfunktion taulukkoa "takaperin":

$$P(Z \le z_\alpha) = 0{,}8 \;\Leftrightarrow\; \Phi(z_\alpha) = 0{,}8 \;\Leftrightarrow\; z_\alpha \approx 0{,}842 .$$

b) Sama arvo voidaan laskea esim. Mathcadistä löytyvällä normaalijakauman kertymäfunktion käänteisfunktiolla:

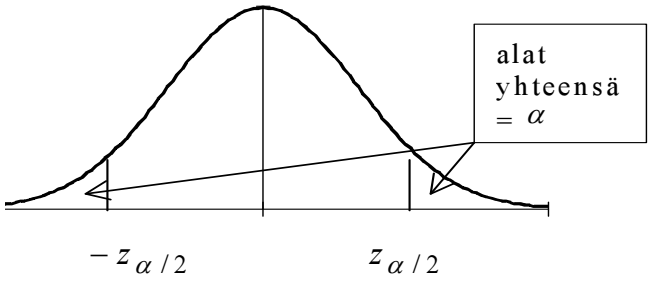
$$\text{qnorm}(0.8, 0, 1) = 0.84162 \;\blacksquare$$

Esimerkiksi ns. yksisuuntaisissa, yksitahoisissa (one-tailed) testeissä hylätään ne satunnaismuuttujan arvot, jotka ylittävät käytettyä merkitsevyystasoa α vastaavan kriittisen kohdan z_α. Havainnollisesti voidaan sanoa, että tiheysfunktiosta hylätään "loppupää" (tai vastaavasti alkupää).

Kaksisuuntaisissa testeissä hylätään kumpikin pää ja tarkastellaan väliä

$$[-z_{\alpha/2}, z_{\alpha/2}],$$

jonka ulkopuolelle jää α:n verran todennäköisyysmassaa. Tätä väliä, joka siis täyttää ehdon



(3) $$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha ,$$

sanotaan luottamustasoa $1-\alpha$ (tai merkitsevyystasoa α) vastaavaksi ***luottamusväliksi***.

Seuraavassa taulukossa on kolme tavallisinta luottamustasoa (eli $1-\alpha$:n arvoa), niitä vastaavat merkitsevyystasot (α-arvot, p-arvot) sekä vastaavat $z_{\alpha/2}$:n arvot.

luottamustaso $1-\alpha$	95 %	99 %	99,9 %
merkitsevyystaso α	0,05	0,01	0,001
$z_{\alpha/2}$	1,96	2,58	3,30

Esim. 5 a) Edellisen taulukon mukaan standardinormaalin satunnaismuuttujan Z 95 %:n luottamustasoa vastaava luottamusväli on

$$-1{,}96 \le Z \le 1{,}96 .$$

Tarkistetaan tulos:

$$\begin{aligned} P(-1{,}96 \le \overline{Z} \le 1{,}96) &= \Phi(1{,}96) - \Phi(-1{,}96) \\ &= \Phi(1{,}96) - [1 - \Phi(1{,}96)] = 2 \cdot \Phi(1{,}96) - 1 \\ &\approx 2 \cdot 0{,}9750 - 1 = 0{,}950. \end{aligned}$$

b) Laske 80 %:n luottamustasoa vastaava Z:n luottamusväli. Nyt $1-\alpha$:lla on arvo 0,8, joten α:n arvo on 0,2 eli $\alpha/2$:n arvo on 0,1. Lasketaan esimerkin 4 a) mukaisella tavalla kriittinen arvo $z_{\alpha/2}$, jonka yläpuolella on määrä $\alpha/2 = 0{,}1$ todennäköisyysmassaa eli alapuolella määrä 0,9:

$$\Phi(z_{\alpha/2}) = 0{,}9 \;\Leftrightarrow\; z_{\alpha/2} = 1{,}282 .$$

Siten kysytty luottamusväli on [–1,282; 1,282].

Palataan nyt otoskeskiarvon $\overline{X}$ jakaumaan. Jos otoskoko on riittävän suuri, niin tuloksen (2) mukaan otoskeskiarvon jakauma on likimain $(\mu, \sigma^2/n)$-normaali. Vastaavan standardoidun satunnaismuuttujan $\overline{Z}$ jakauma on siis seuraava:

(4) $$\boldsymbol{\overline{Z} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)} .$$

Ehdon (3) mukaan

(5) $$P(-z_{\alpha/2} \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}) = 1 - \alpha .$$

Tässä esiintyvistä epäyhtälöistä $-z_{\alpha/2} \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ ja $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}$ saadaan ratkaistua μ:lle rajat seuraavasti:

$\mu \le \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, $\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu$. Täten tulos (3) saa muodon

(6) $$P(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha .$$

Tässä (kuten edellisessäkin yhtälössä) $\overline{X}$ on käsitettävä satunnaismuuttujan $\overline{X}$ arvoksi (joka on luku), ts. otoksesta lasketuksi keskiarvoksi.

Tuloksen (6) nojalla luottamustasoa $1-\alpha$ vastaava luottamusväli on

$$[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

Esimerkiksi koko aineiston odotusarvo μ asettuu 95 % todennäköisyydellä välille $[\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}},\ \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}]$.

9.3 Otosvarianssi ja Student-jakauma

Voidaan todistaa, että satunnaismuuttuja

(7) $$S^2 = \frac{1}{n-1}[(X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2]$$

(eikä $\frac{1}{n}[(X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2]$, jonka odotusarvo on $\frac{n-1}{n}\sigma^2$) on koko populaatioon liittyvän satunnaismuuttujan X varianssin harhaton estimaatti, ts.

$$E[S^2] = \sigma^2 .$$

Satunnaismuuttujaa S^2 sanotaan ***otosvarianssiksi*** ja sen neliöjuurta ***otoskeskihajonnaksi***. Voidaan myös todistaa, että S ei ole X:n keskihajonnan harhaton estimaatti. Silti sitä voidaan joutua käyttämään käytännössä X:n keskihajonnan likiarvona.

Kohdan 9.2 tuloksen (2) mukaan otoskeskiarvon $\overline{X}$ jakauma on riittävän suurilla otoskoon n arvoilla likimain normaali, odotusarvona μ ja keskihajontana $\sigma/\sqrt{n}$. Täten (riittävän suurilla n:n arvoilla) vastaava standardoitu satunnaismuuttuja on likimain $(0,1)$-normaali:

$$\overline{Z} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Koko populaation keskihajontaa σ ei yleensä tunneta, vaan se joudutaan korvaamaan otoshajonnalla S. Mutta satunnaismuuttujan

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

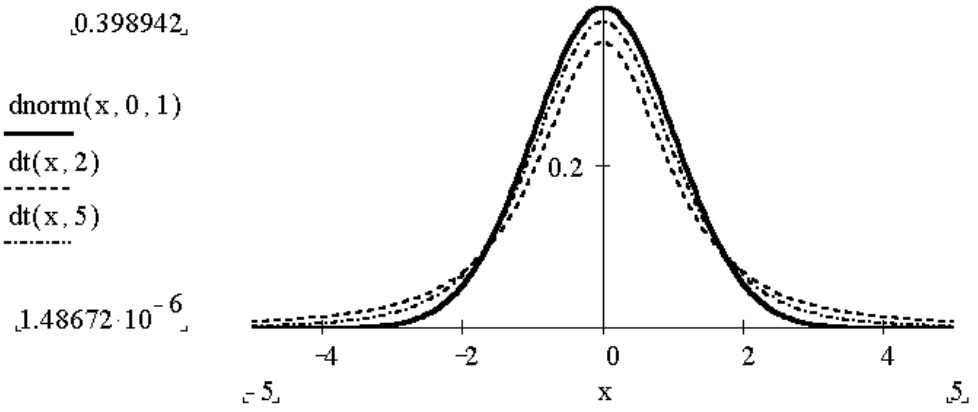
jakauma ei ole lähes normaali vaan likimain ns. ***Student-jakauma*** eli ***t-jakauma***.

Jos koko populaation jakauma on ***normaali***, niin voidaan todistaa, että T:n jakauma on tarkalleen t-jakauma – myös pienillä n:n arvoilla.

Lukua $f = n - 1$ (joka esiintyy S:n lausekkeen (7) nimittäjässä) sanotaan t-jakauman ***vapausasteeksi*** (degrees of freedom). Student-jakauman tiheysfunktio on muotoa

$$f(x) = k \cdot (1 + \frac{x^2}{n-1})^{-n/2},$$

missä k:lla on sellainen (n:stä riippuva) arvo, että ehto $\int_{-\infty}^{\infty} f(x)\,dx = 1$ tulee täytettyä (k on eräs ns. *gammafunktion* $\Gamma(n)$ lauseke). Seuraavaan kuvaan on piirretty Mathcad 7 -ohjelmalla (0, 1)-normaalisen jakauman sekä 2- ja 5-vapausasteisten t-jakaumien tiheysfunktioiden kuvaajat.



Kun vapausasteiden luku $f = n - 1$ kasvaa, niin t-jakauma lähenee (0,1)-normaalista jakaumaa. Edellisessä kuvassa f:n arvoa 5 vastaava kuvaaja on jo aika lähellä normaalijakauman kuvaajaa (yhtenäinen viiva).

Tuloksia (6) ja (5) vastaavat tulokset ovat yleisellä α:n arvolla

$$P(\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}) = 1 - \alpha$$

ja

$$P(-t_{\alpha/2} \le \frac{\overline{X} - \mu}{S/\sqrt{n}} \le t_{\alpha/2}) = 1 - \alpha\,.$$

Luku $t_{\alpha/2}$ riippuu paitsi α:n arvosta myös vapausasteesta $n - 1$.

Seuraavassa taulukossa on kolmelle tavallisimmalle luottamustasolle (tai niitä vastaaville α:n arvoille) esitetty eri vapausasteita vastaavat kriittiset arvot $t_{\alpha/2}$.

luottamustaso	95 %	99 %	99,9 %
α	0,05	0,01	0,001
vapausaste $f = n-1$	$t_{\alpha/2} = t_{0,025}$	$t_{\alpha/2} = t_{0,005}$	$t_{\alpha/2} = t_{0,0005}$
1	12,7	63,6	637
4	2,78	4,60	8,61
9	2,26	3,25	4,78
10	2,23	3,17	4,59
14	2,14	2,98	4,14
19	2,09	2,86	3,88
29	2,04	2,76	3,66

Huomaa, että Mathcadillä laskettaessa saat $qt(0.975,9) = 2.262$ ja $qt(0.025,9) = -2.262$, ts. jälkimmäinen arvo on negatiivinen, sillä se ilmoittaa kohdan, johon mennessä on kertynyt määrä 0,025 todennäköisyysmassaa. Tässä monisteessa taas $t_{0,025}$ tarkoittaa kohtaa, jonka jäljessä (vrt. kuva s. 77) on määrä 0,025 todennäköisyysmassaa.

Esim. 6 Vrt. *Esim. 12* sivulla 12. Samasta kohteesta tehtyjen 10 mittauksen keskiarvo on 28,5 ja otoshajonta 0,68. Lasketaan keskiarvolle 95 % luottamusväli. Tässä populaationa on äärettömän monen mittauksen tulos, joka on (ainakin likimain) normaali. Siten edellisiä tuloksia voidaan soveltaa, vaikka otoskoko on näin pieni (alle 30):

$$[\overline{X} - t_{0,025} \cdot S/\sqrt{n},\ \overline{X} + t_{0,025} \cdot S/\sqrt{n}]$$
$$= [28{,}5 - 2{,}26 \cdot 0{,}68/\sqrt{10},\ 28{,}5 + 2{,}26 \cdot 0{,}68/\sqrt{10}]$$
$$\approx [28{,}5 - 0{,}49,\ 28{,}5 + 0{,}49] \approx \underline{\underline{[28,\ 29]}}.$$

Siis suureen oikea arvo on $x = 28{,}5 \pm 0{,}5$ 95 % varmuudella.

9.4 Varianssin arviointi, χ^2-jakauma

Edellä ollut populaation odotusarvon μ arviointi perustui *t*-jakaumaan. Populaation varianssin σ^2 arviointi taas perustuu ***"khi toiseen"***-jakaumaan. Jos $Z_1, ..., Z_n$ ovat *riippumattomia (0,1)-normaaleja* satunnaismuuttujia, niin satunnaismuuttuja

$$\chi^2 = {Z_1}^2 + ... + {Z_n}^2$$

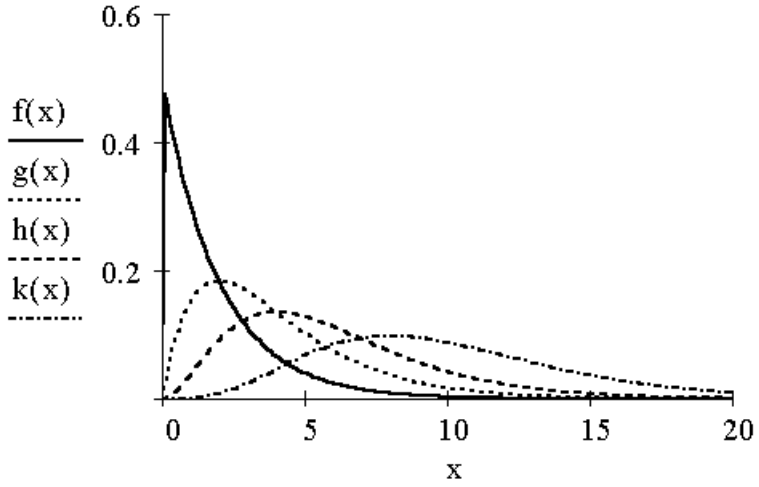
noudattaa ns. ***χ^2-jakaumaa***, *vapausasteena* $f = n - 1$. Sen tiheysfunktio on muotoa

$$f(x) = k \cdot x^{\frac{n}{2}-1} \cdot e^{-\frac{1}{2}x} \qquad (x \geq 0),$$

missä *k*:lla on sellainen (*n*:stä riippuva) arvo, että ehto $\int_{-\infty}^{\infty} f(x)\,dx = 1$ tulee täytettyä (tämäkin *k* on eräs gammafunktion lauseke). Tiheysfunktion kuvaaja ei enää ole symmetrinen. Vapausasteen $f = n - 1$ kasvaessa jakauma siirtyy oikealle ja sen muoto muuttuu oleellisesti. Seuraavassa kuvassa ovat vapausasteita 2, 4, 6 ja 10 vastaavat tiheysfunktioiden kuvaajat Mathcadillä piirrettyinä (ja siirrettyinä Wordiin *bittikartta*-muodossa).

$x := 0, 0.1 .. 20$ $\quad f(x) := \text{dchisq}(x, 2)$ $\quad g(x) := \text{dchisq}(x, 4)$

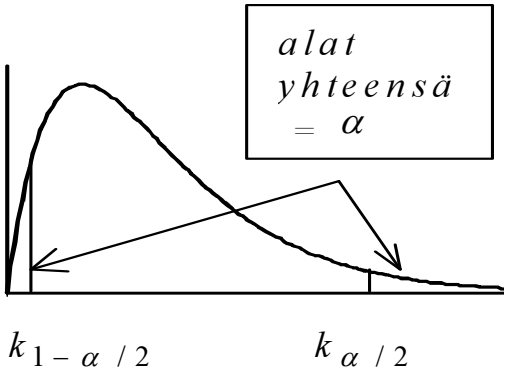
$h(x) := \text{dchisq}(x, 6)$ $\quad k(x) := \text{dchisq}(x, 10)$



Merkitään k_α:lla kohtaa (ns. ***kriittistä arvoa***), jonka jäljessä (jakauman loppupäässä) on α:n verran todennäköisyysmassaa, ts. $P(\chi^2 \geq k_\alpha) = \alpha$. Silloin

$$P(\chi^2 \geq k_{\alpha/2}) = \alpha/2, \quad P(\chi^2 \geq k_{1-\alpha/2}) = 1-\alpha/2,$$
$$P(\chi^2 \leq k_{1-\alpha/2}) = \alpha/2$$

ts. kohdan $k_{\alpha/2}$ yläpuolelle ja kohdan $k_{1-\alpha/2}$ alapuolelle (eli molempiin "häntiin") jää yhteensä α:n verran todennäköisyysmassaa (kumpaankin puolet).



Täten

$$P(k_{1-\alpha/2} \leq \chi^2 \leq k_{\alpha/2}) = 1-\alpha.$$

Oletetaan nyt, että koko populaation jakauma on (μ, σ^2)-normaali, jonka parametrejä μ ja σ^2 ei tunneta. Muodostetaan otosvarianssin $S^2 = \frac{1}{n-1}[(X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2]$ avulla uusi satunnaismuuttuja

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum (X_i - \overline{X})^2 \ .$$

Voidaan todistaa, että sen jakauma on χ^2. Täten

$$P(k_{1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq k_{\alpha/2}) = 1-\alpha.$$

Sulkeissa olevan kaksoisepäyhtälön oikea ja vasen puolikas voidaan kirjoittaa muotoihin $\frac{(n-1)S^2}{k_{\alpha/2}} \leq \sigma^2$ ja $\sigma^2 \leq \frac{(n-1)S^2}{k_{1-\alpha/2}}$ eli yhdistettynä

$$P(\frac{(n-1)S^2}{k_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{k_{1-\alpha/2}}) = 1-\alpha.$$

Näin saadaan varianssille σ^2 luottamustasoa $1-\alpha$ vastaava luottamusväli:

(8)
$$\left[\frac{(n-1)S^2}{k_{\alpha/2}}, \frac{(n-1)S^2}{k_{1-\alpha/2}}\right].$$

Seuraavassa on muutama 95 %:n ja 99 %:n luottamustasoja eli α:n arvoja 0,05 ja 0,01 vastaavien kertoimien $k_{\alpha/2}$ ja $k_{1-\alpha/2}$ arvo.

$luottamustaso\, 1-\alpha\, (prosenteissa)$	95 %	99 %
α	0,05	0,01
$f = n-1$	$k_{\alpha/2}$ ja $k_{1-\alpha/2}$	$k_{\alpha/2}$ ja $k_{1-\alpha/2}$
1	5,02 ja 0,00	7,88 ja 0,00
4	11,1 ja 0,48	14,9 ja 0,21
9	19,0 ja 2,70	23,6 ja 1,73
19	32,9 ja 8,91	38,6 ja 6,84

Tällaisia *k*-kertoimia voidaan laskea esim. Mathcadillä käyttämällä χ^2-jakauman kertymäfunktion käänteisfunktiota. Jos esim. vapausasteiden luku $f = 9$, niin 95 % tasoa vastaavat arvot $k_{\alpha/2}$ ja $k_{1-\alpha/2}$ saadaan, kun etsitään kohdat, joihin mennessä on kertynyt 97,5 % ja 2,5 % todennäköisyysmassasta. Tämä käy seuraavasti:

$\text{qchisq}(0.975, 9) = 19.0228$ $\quad$ $\text{qchisq}(0.025, 9) = 2.7004$

Esim. 7 Esimerkissä 6 laskettiin mitattavan suureen keskiarvon luottamusväli, kun 10 mittauksen keskiarvo oli 28,5 ja otoshajonta 0,68 ja käytettiin 95 % varmuutta (luottamustasoa).

Lasketaan nyt, millä välillä koko populaation (tässä tapauksessa äärettömän monen tai käytännössä hyvin monen mittaustuloksen, jotka luokiteltaisiin sopivasti) muodostaman normaalijakauman varianssi ja keskihajonta ovat 95 % varmuudella.

Käytetään edellä ollutta luottamustasoa $1-\alpha$ vastaavaa varianssin σ^2 luottamusväliä (8). Edellisestä taulukosta (tai sen jäljessä suoritetusta laskennasta Mathcadillä) saadaan ***kriittiset arvot*** $k_{\alpha/2} = 19{,}02$ ja $k_{1-\alpha/2} = 2{,}70$, joiden mukaan täksi väliksi tulee

$$\left[\frac{9 \cdot 0{,}68^2}{19{,}02},\ \frac{9 \cdot 0{,}68^2}{2{,}70}\right] \approx [0{,}22\,;1{,}54].$$

Vastaava keskihajonnan σ luottamusväli saadaan ottamalla edellisen välin päätepisteistä neliöjuuret: [0,47 ; 1,24]. Siis koko populaation keskihajonta on 95 % varmuudella tällä välillä.

Harjoituksia

A, B

9.1 Jos $X \sim Bin(50, 1/4)$, laske

a) binomijakaumaa käyttäen $P(X = 4 \text{ tai } 5)$,

b) normaalijakaumaa käyttäen $P(4 \le X \le 10)$

c) tarkista tulokset jollakin matematiikka-ohjelmalla.

9.2 Esitä esimerkki 3 b) yksityiskohtaisesti.

9.3 Oletetaan, että otoskoko on aika suuri. Laske todennäköisyys, että otoskeskiarvo poikkeaa koko aineiston odotusarvosta korkeintaan 2,5 kertaa keskiarvon keskivirheen verran.

9.4 Koko populaation varianssin arvioidaan (monien testausten perusteella) olevan 25,4. Tehdään 15 kpl suuruinen otos, josta keskiarvoksi saadaan 90,3. Laske näiden tietojen perusteella, millä välillä koko populaation odotusarvo on 99 % todennäköisyydellä.

9.5 15 kpl otoksesta saadaan otoskeskiarvoksi 90,3 ja otosvarianssiksi 25,4. Laske, millä välillä koko populaation odotusarvo on 99 % todennäköisyydellä.

9.6 Normaalijakaumasta otetaan 20 kpl otos, josta saadaan otoskeskiarvoksi 90,3 ja otoshajonnaksi 6,22. Laske, millä välillä koko populaation odotusarvo ja keskihajonta ovat 99 % todennäköisyydellä.

9.7 a) Laske Mathcadillä (tai muulla käytettävissä olevalla matematiikkaohjelmalla) $z_{\alpha/2}$, $t_{\alpha/2}$, $k_{\alpha/2}$ ja $k_{1-\alpha/2}$, kun otoskokona on 12 ja luottamustasona 82 %. b) Ratkaise sitten tehtävät 9.4, 9.5 ja 9.6 tätä otoskokoa ja luottamustasoa käyttäen.

10 Hypoteesin testaus

10.1 Normaalijakauman odotusarvon testaus

1) Oletetaan, että satunnaismuuttujan X jakauma on normaalijakauma, josta tunnetaan varianssi σ^2. Sen sijaan odotusarvoa μ ei tunneta, mutta otaksutaan sen olevan suuruudeltaan μ_o.

Tarkoituksena on tutkia tämän otaksuman (hypoteesin) paikkansapitävyyttä otoksen avulla. Sitä varten asetetaan ***nollahypoteesi*** $H_o: \mu = \mu_o$ ja ns. *kaksisuuntainen* ***vastahypoteesi*** $H_1: \mu \neq \mu_o$.

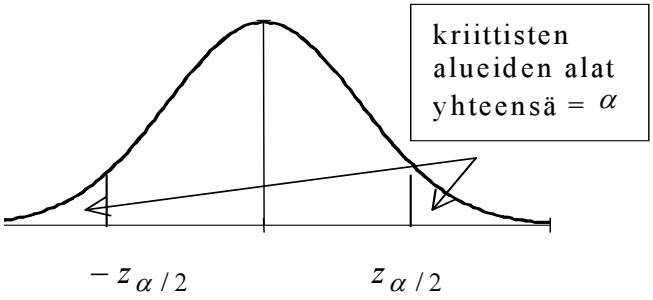
Tehdään n:n suuruinen otos. Mikäli hypoteesi pitää paikkansa, niin edellisen luvun tulosten mukaan otoskeskiarvon $\overline{X}$ jakauma on ainakin likimain $(\mu_o, \sigma^2 / n)$-normaali ja vastaava standardoitu satunnaismuuttuja (0, 1)-normaali:

$$\boxed{\overline{Z} = \frac{\overline{X} - \mu_o}{\sigma / \sqrt{n}}} \sim N(0,1).$$

Satunnaismuuttujaa $\overline{Z}$ voidaan käyttää populaation odotusarvoa testattaessa ns. ***testisuureena*** (kun koko populaation varianssi σ^2 tunnetaan). Lasketaan tästä otoksesta $\overline{Z}$:n arvo ja tutkitaan, jääkö se jotakin tavallista *merkitsevyystasoa* α vastaavan ***luottamusvälin eli hyväksymisalueen***

(1) $\boxed{-z_{\alpha/2} \leq \overline{Z} \leq z_{\alpha/2}}$

ulkopuolelle ns. ***kriittiseen alueeseen*** eli *hylkäämisalueeseen*.



Kaksoisepäyhtälö (1) voidaan kirjoittaa muotoon

$$-z_{\alpha/2} \leq \frac{\overline{X} - \mu_o}{\sigma / \sqrt{n}} \leq z_{\alpha/2}$$

ja edelleen muotoon $\mu_o - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu_o + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, joten voitaisiin yhtä hyvin tutkia, jääkö $\overline{X}$:n arvo tämän alueen ulkopuolelle.

Edellisessä luvussa esitettiin tavallisimpia α:n arvoja vastaavat kriittiset arvot $z_{a/2}$ seuraavanlaisena taulukkona:

luottamustaso $1-\alpha$	95 %	99 %	99,9 %
merkitsevyystaso α	0,05	0,01	0,001
kriittinen arvo $z_{\alpha/2}$	1,96	2,58	3,30

Jos otoksesta laskettu $\overline{Z}$:n tai $\overline{X}$:n arvo jää 99,9 % luottamustasoa vastaavan luottamusvälin ulkopuolelle (eli kriittistä arvoa 0,001 vastaavaan kriittiseen alueeseen), sanotaan, että otoskeskiarvo poikkeaa tilastollisesti ***erittäin merkitsevästi*** nollahypoteesista $H_o: \mu = \mu_o$.

Jos $\overline{Z}$:n arvo jää 99 % luottamusvälin ulkopuolelle mutta 99,9 % luottamusvälin sisään, sanotaan, että otoskeskiarvo poikkeaa tilastollisesti ***merkitsevästi*** nollahypoteesista. Vastaavasti 95 % luottamusvälin ulkopuolelle mutta 99 % luottamusvälin sisään jääminen merkitsee että otoskeskiarvo poikkeaa tilastollisesti ***melkein merkitsevästi*** nollahypoteesista. Jos halutaan käyttää esim. 99 % varmuutta ja otoksen antaman tuloksen poikkeama nollahypoteesista on vähintäänkin merkitsevä (ts. merkitsevä tai erittäin merkitsevä), nollahypoteesi on syytä hylätä.

Kaksisuuntaisen testauksen sijaan voidaan käyttää myös ***yksisuuntaista*** *testausta*. Jos vastahypoteesiksi tai paremminkin vaihtoehtoiseksi hypoteesiksi otetaan esim. $H_1: \mu > \mu_o$ puhutaan *oikeanpuoleisesta testauksesta*. Siinä hylkäämisalue on

$$\boxed{\overline{Z} > z_a} \text{ eli } \overline{X} > \mu_o + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Eri hylkäämisalueita vastaavat kriittiset arvot z_a voidaan laskea edellisen luvun esimerkin 4 mukaisesti kertymäfunktion Φ avulla (tai esim. matematiikkaohjelmilla):

$$\Phi(z_\alpha) = 0{,}95 \Leftrightarrow z_\alpha \approx 1{,}65 \text{ (melkein merkitsevä)}$$
$$\Phi(z_\alpha) = 0{,}99 \Leftrightarrow z_\alpha \approx 2{,}33 \text{ (merkitsevä)}$$
$$\Phi(z_\alpha) = 0{,}999 \Leftrightarrow z_\alpha \approx 3{,}08 \text{ (erittäin merkitsevä)}$$

Esim. 1 Valmistettavan tuotteen erästä suuretta (pituutta, painoa, paksuutta, käyttöikää, konsentraatiota tms.) on seurattu pitkän aikaa ja todettu sen noudattavan aika hyvin normaalijakaumaa, odotusarvona 32,5 ja keskihajontana 2,8. Tuotantoprosessin

jossakin vaiheessa näyttäisi siltä, että suureen keskimääräinen arvo (odotusarvo) olisi alkanut kasvaa (mutta sen sijaan valmistusprosessin luonteen perusteella voidaan olettaa, että keskihajonta ei ajan mukana oleellisesti muutu). Asian selvittämiseksi tutkitaan 10 tuotetta ja niistä saadaan kyseisen suureen keskiarvoksi 34,8.

Valitaan nollahypoteesiksi se, että ehkä suureen keskimääräinen arvo ei ole kuitenkaan muuttunut, vaan otoksen antama keskiarvo on sattumalta (ehkä otoksen pienuudesta johtuen) suurempi kuin pitkäaikainen keskiarvo. Siis valitaan $H_o: \mu = 32{,}5$ ja vastahypoteesiksi valitaan $H_1: \mu > 32{,}5$. Testisuureen $\overline{Z}$ arvoksi saadaan tässä tapauksessa

$$\overline{Z} = \frac{\overline{X} - \mu_o}{\sigma / \sqrt{n}} = \frac{34{,}8 - 32{,}5}{2{,}8 / \sqrt{10}} \approx 2{,}6.$$

Tämä arvo on suurempi kuin 99 %:n luottamustasoa (tasoa "merkitsevä") vastaava kriittinen arvo 2,33, joten otoksen perusteella suureen keskiarvo on kasvanut merkitsevästi ja siksi H_o on syytä hylätä, jos tyydytään tähän luottamustasoon. Ehkä kannattaa myös tutkia, mikä seikka tuotantoprosessissa on aiheuttanut muutoksen. Lasketaan vielä tarkemmin, mitä luottamustasoa arvo 2,6 vastaa:

$$\Phi(2{,}6) \approx 0{,}995 = 99{,}5\,\%.$$

2) Jos normaalisti jakautuneen satunnaismuuttujan X varianssia ei tunneta, testisuureena on käytettävä *Student-jakauman* vastaavaa testisuuretta

$$\boxed{T = \frac{\overline{X} - \mu_o}{S / \sqrt{n}}},$$

missä S on otoskeskihajonta [$S^2 = \frac{1}{n-1}(X_1{}^2 + ... + X_n{}^2)$]. Sitä sanotaan myös $(n-1)$-hajonnaksi ja sen arvo voidaan yleensä laskea mm. laskimella, jossa on tilastollisia toimintoja. Esimerkiksi kaksisuuntaisessa testissä luottamusvälinä (hyväksymisalueena) on $\boxed{-t_{\alpha/2} \le T \le t_{\alpha/2}}$ ja oikeanpuoleisessa testissä $T \ge t_\alpha$. Kriittisiä arvoja t_α löytyy taulukoista, jollainen on mm. kohdassa 9.3 (s.82) tai niitä voidaan laskea mm. matematiikkaohjelmilla. Jos esim. käytetään 80 % luottamustasoa, ts. $\alpha = 0{,}20$ ja vapausasteiden määrä $f = n - 1 = 14$ (eli otoskoko on 15), niin Mathcadillä saadaan

$t_\alpha = t_{0,80} = \text{qt}(0.80,\ 14)\ =\ 0.868,\ t_{\alpha/2} = t_{0,90} = \text{qt}(0.90,\ 14)\ = 1.345.$

Esim. 2 Oletetaan, että jonkin tutkittavan suureen jakauma on normaali ja odotusarvon arvellaan olevan 32,5. Arvelun varmistamiseksi tutkitaan 10 tuotetta. Otoskeskiarvoksi saadaan 34,6 ja otoshajonnaksi 2,8. Käytetään kaksisuuntaista testausta ja a) 99 % luottamustasoa (tasoa "merkitsevä"), b) 95 % luottamustasoa ("melkein merkitsevä").

Valitaan $H_o: \mu = 32{,}5$, $H_1: \mu \neq 32{,}5$ jolloin testisuureen arvo on $T = \dfrac{\overline{X} - \mu_o}{S/\sqrt{n}} = \dfrac{34{,}6 - 32{,}5}{2{,}8/\sqrt{10}} \approx 2{,}37$. Vapausasteiden määrä on $f = n - 1 = 9$. Kohdassa 9.3 (s. 82) olevasta taulukosta saadaan luottamusväleiksi a) $-3{,}25 \le T \le 3{,}25$, b) $-2{,}26 \le T \le 2{,}26$. Koska testimuuttujan arvo on edellisessä, mutta ei jälkimmäisessä välissä, otoskeskiarvo poikkeaa melkein merkitsevästi nollahypoteesin mukaisesta arvosta.

Lasketaan vielä arvoa 2,37 vastaava luottamustaso $1 - \alpha$ Mathcadillä kaksisuuntaisessa testauksessa, kun $f = 9$:

$P(-t_{\alpha/2} \le T \le t_{\alpha/2}) =$ $\text{pt}(2.37, 9)\ -\ \text{pt}(-2.37, 9)\ =\ 0.958$ ∎

Siis kyseessä on 95,8 % luottamustaso ja $\alpha = 0{,}042$ eli hylkäämisalueessa on todennäköisyysmassaa 0,042 yksikköä.

10.2 Normaalijakauman varianssin testaus

Oletetaan, että satunnaismuuttujan X jakauma on (ainakin likimain) normaali. Kun testataan, onko X:n varianssilla arvelujen mukainen arvo $\sigma_o{}^2$, tutkitaan n:n kappaleen otos, josta lasketaan otosvarianssi S^2, ja asetetaan hypoteesi $H_o: \sigma^2 = \sigma_o{}^2$ sekä joko kaksi- tai yksisuuntainen vastahypoteesi. Testisuureena on

$$\boxed{K = \frac{(n-1)S^2}{\sigma_o{}^2}},$$

joka on χ^2-jakautunut, vapausasteena $f = n - 1$ (vrt. kohta 9.4). Jos käytetään merkitsevyystasoa α ja kaksisuuntaista testausta, tutkitaan, onko K:n arvo luottamusvälin

$$k_{1-\alpha/2} \le K \le k_{\alpha/2}$$

ulkopuolella. Jos näin on, nollahypoteesi on syytä hylätä (mikäli tyydytään käytettyyn luottamustasoon). Kohdassa 9.4 (s. 85) oli pieni taulukko, josta saadaan kriittisiä arvoja $k_{1-\alpha/2}$ ja $k_{\alpha/2}$ 95 % ja 99 % luottamustasoille. Taulukon alla on esitetty, miten näitä arvoja lasketaan Mathcadillä. Oikeanpuoleisessa testauksessa tutkitaan, onko K:n arvo kriittistä arvoa k_α suurempi (ts. K:n tiheysfunktion loppupään muodostamassa hylkäysalueessa). Jos esim. käytetään 95 % luottamustasoa, niin $\alpha = 0{,}05$ $(= 5\,\%)$ ja vastaava kriittinen kohta on $q_\alpha = q_{0{,}05}$. Vasemmanpuoleisessa testauksessa taas tutkitaan, onko K:n arvo kriittistä arvoa $k_{1-\alpha}$ pienempi.

Esim. 3 Koneiden uusimisella pyrittiin vaikuttamaan siihen että tuote olisi tasalaatuisempaa, ts. tasalaatuisuutta kuvaavalla suureella (tuotteen paksuudella, painolla, happamuudella tms.) olisi aikaisempaa pienempi keskihajonta. Pitkäaikaisella seurannalla oli aikaisempia koneita käytettäessä saatu keskihajonnaksi 4,8.

Oletetaan, että satunnaismuuttuja X ilmoittaa testattavan suureen arvon ja X:n jakauma on ainakin likimain normaali. 10 kappaleen otoksesta lasketuksi otoshajonnaksi saatiin $S = 4{,}1$. Tutkitaan, onko hajonta pienentynyt merkittävästi.

$$H_o\colon \sigma^2 = 4{,}8^2 \approx 23 \qquad H_1\colon \sigma^2 < 23$$

$$K = \frac{(n-1)S^2}{{\sigma_o}^2} = \frac{9 \cdot 4{,}1^2}{23} = 6{,}57\ldots \approx 6{,}6$$

Luottamustasoa 99 % (eli α:n arvoa 0,01, "merkittävästi") ja vapausastetta $f = 9$ vastaava kriittinen arvo on $k_{1-\alpha} = 2{,}088$. Se saadaan taulukoista tai esim. Mathcadiä käyttämällä: $qchisq(0.01, 9) = 2.088$. Koska K:n arvo ei ole tätä arvoa pienempi, keskihajonnan ei tämän otoksen perusteella voida katsoa pienentyneen merkittävästi. Voidaan myös päätellä, että koska K:n arvo ei ole hylkäämisalueessa $K < k_{1-\alpha} = 2{,}088$, niin nollahypoteesia ei ole syytä hylätä.

10.3 Kahden jakauman odotusarvon vertailu

Tutkitaan, ovatko satunnaismuuttujien X_1 ja X_2 odotusarvot μ_1 ja μ_2 samat. Tehdään nollahypoteesi $H_o\colon\ \mu_1 = \mu_2$ ja joko kaksisuuntainen

vastahypoteesi H_1: $\mu_1 \neq \mu_2$ tai jompikumpi yksisuuntaisista hypoteeseista H_1: $\mu_1 > \mu_2$ tai H_1: $\mu_1 < \mu_2$.

a) Oletetaan, että *tunnetaan populaatioiden varianssit* ${\sigma_1}^2$ ja ${\sigma_2}^2$.

Otetaan edellisestä jakaumasta (populaatiosta) kokoa n_1 ja jälkimmäisestä kokoa n_2 oleva otos. Jos niiden otoskeskiarvot ovat $\overline{X}_1$ ja $\overline{X}_2$ ja otokset voidaan olettaa riippumattomiksi, niin otoskeskiarvojen erotuksen

$$X = \overline{X}_1 - \overline{X}_2$$

jakauma on keskeisen raja-arvolauseen mukaan likimain normaali (tai tarkalleen normaali, mikäli populaatioiden jakaumat ovat normaaleja). Odotusarvo on $\mu = \mu_1 - \mu_2$, joka on = 0, kun oletetaan että nollahypoteesi pitää paikkansa. Varianssille tulee lauseke

$$Var(X) = \frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2},$$

sillä yleisesti jos X ja Y ovat riippumattomia, niin

$$Var(X - Y) = Var(X) + (-1)^2 Var(Y) = Var(X) + Var(Y).$$

Täten vastaavan standardoidun satunnaismuuttujan (jonka yleinen muoto on $Z = \frac{X - \mu}{\sigma}$)

$$\boxed{Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}}$$

jakauma on ainakin likimain (0, 1)-normaali. Esim. kaksisuuntaisessa testauksessa luottamustasoa $1-\alpha$ vastaavana luottamusvälinä (hypoteesin hyväksymisalueena) on

$$-z_{\alpha/2} \leq Z \leq z_{\alpha/2}.$$

Jos erityisesti populaatioiden variansseilla ${\sigma_1}^2$ ja ${\sigma_2}^2$ on sama arvo, jota merkitään σ^2:lla, niin testisuure saa muodon

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

b) Oletetaan nyt, että *populaatioiden variansseja* σ_1^2 ja σ_2^2 *ei tunneta, mutta tiedetään niillä olevan sama arvo*. Lasketaan ensin *otosvarianssien painotettu keskiarvo*, painoina "otosten vapausasteet" $f_1 = n_1 - 1$ ja $f_2 = n_2 - 1$:

$$S^2 = \frac{f_1 \cdot S_1^2 + f_2 \cdot S_2^2}{f_1 + f_2}$$

(joka on populaatioiden varianssin tarkentuva ja harhaton estimaattori). Testisuureena käytetään nyt satunnaismuuttujaa

$$\boxed{T = \frac{\overline{X}_1 - \overline{X}_2}{S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}.$$

Kun oletetaan että nollahypoteesi pitää paikkansa, niin T on *Student*-jakautunut, vapausasteena $f = f_1 + f_2 = n_1 + n_2 - 2$.

c) Jos *koko populaatioiden variansseilla* σ_1^2 ja σ_2^2 *on erisuuret arvot, mutta niitä ei tunneta*, testisuureena on

$$\boxed{T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}}.$$

Tämä on likimain (0, 1)-normaalisti jakautunut, jos otoskoot ovat suuria (yli 50), mutta *Student*-jakautunut, jos otoskoot ovat pieniä. Vapausaste f on tällöin laskettava (kokonaisluvuksi pyöristettynä) yhtälöstä

$$\frac{1}{f} = \frac{c^2}{f_1} + \frac{(1-c)^2}{f_2}, \quad \text{missä} \quad c = \frac{S_1^2}{n_1} / \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right).$$

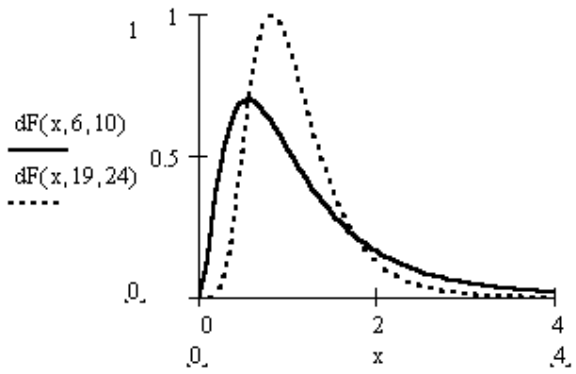
Kohtien a) – c) mukaisiin testauksiin joudutaan mm., kun tutkitaan kuinka hyvin kahden tuotantolinjan valmistamat tuotteet (esim. alkuperäisvaraosat/piraatit) ovat samanlaisia. Laskennassa käytetään kriittisiä arvoja z_α tai t_α ja luottamusvälejä tai hylkäämisalueita samaan tapaan kuin edellä.

10.4 Kahden jakauman varianssin vertailu

Tutkitaan, ovatko satunnaismuuttujien X_1 ja X_2 varianssit σ_1^2 ja σ_2^2 (joita ei tunneta) samat. Oletetaan, että *kummankin satunnaismuuttujan jakauma on ainakin likimain normaali ja niistä otetut otokset ovat riippumattomat*. Aikaisemman mukaan satunnaismuuttujat

$$K_1 = \frac{f_1 S_1^2}{\sigma_1^2} \quad \text{ja} \quad K_2 = \frac{f_2 S_2^2}{\sigma_2^2}$$

ovat χ^2-jakautuneita, vapausasteina $f_1 = n_1 - 1$ ja $f_2 = n_2 - 1$. Näiden suhde (*Fraction*) $F = K_1 / K_2$ on ns. ***F-jakautunut, vapausasteina*** $\boldsymbol{f_1, f_2}$ (ainakin likimain ja kun otoskoot ovat riittävän suuria). Seuraavassa kuvassa ovat F-jakauman tiheysfunktion kuvaajat, kahdella eri vapausasteparin arvolla:



Tehdään nollahypoteesi H_o: $\sigma_1^2 = \sigma_2^2$. Tällöin testisuure F yksinkertaistuu muotoon

$$\boxed{F = \frac{f_1 S_1^2}{f_2 S_2^2}}.$$

Esim. 4 Haluttiin tutkia merkitsevyystasolla 5 % (eli luottamustasolla 95 %) ja kaksisuuntaisella testauksella, antavatko uusi ja vanha menetelmä tuotteen tietylle, likimain normaalisti jakautuneelle suureelle saman keskihajonnan. Jos otokset ovat

$$n_1 = 20,\ S_1^2 = 37{,}2 \qquad n_2 = 25,\ S_2^2 = 43{,}2,$$

niin testisuureelle saadaan arvo $K = \frac{19 \cdot 37{,}2}{24 \cdot 43{,}2} \approx 0{,}682$. Kriittiset arvot $f_{\alpha/2} = f_{0.975}$ ja $f_{1-\alpha/2} = f_{0.025}$ voidaan laskea esim. Mathcadillä:

$$\mathrm{qF}(0.975, 19, 24) = 2.345 \qquad \mathrm{qF}(0.025, 19, 24) = 0.408$$

Koska K:n arvo on näiden välissä, nollahypoteesia ei ole syytä hylätä (ainakaan tällä luottamustasolla).

10.5 Suhteellisten osuuksien testaus

Jos halutaan tutkia esim. yhden tuotteen suhteellista osuutta, yhden puolueen kannatusprosenttia, joukkueen voittojen prosenttimäärää tms. tai jos halutaan verrata kahden tuotteen suhteellista osuutta keskenään, perustetaan testaus binomijakaumaan ja sitä approksimoivaan normaalijakaumaan.

a) *Yhden suhteellisen osuuden testaus*:

Esim. jonkin tuotteen suhteellisen osuuden p tutkimiseksi tehdään n kappaleen suuruinen otanta, josta oletetaan, että se on luonteeltaan toistokoe (ts. sellainen että yhden tutkittavan tuotteen ottaminen ei vaikuta seuraavan tuotteen ottamiseen). Jos satunnaismuuttuja X ilmoittaa "onnistumisten määrän", ts. montako näistä n:stä tuotteesta on kyseistä lajia, niin X:n jakauma on binomijakauma, *parametreinä n* ja "onnistumistodennäköisyys" p yhdessä kokeessa. Parametrin p arvo on tuntematon ja sille asetetaan jokin arvioitu, hypoteettinen arvo p_o ts. tehdään nollahypoteesi: H_o: $p = p_o$. Tällöin $E(X) = np_o$, $Var(X) = np_oq_o$, missä $q_o = 1 - p_o$ (arvioitu "epäonnistumisten" määrä). Onnistumisten suhteellisen osuuden ilmoittaa satunnaismuuttuja $P = X / n$. Sen odotusarvo ja varianssi ovat

$$E(P) = \frac{1}{n} \cdot E(X) = p_o,\ Var(P) = \frac{1}{n^2} \cdot Var(X) = p_oq_o / n.$$

Riittävän suurilla n:n arvoilla P:n jakauma on likimain normaali parametreinä $\mu = p_o$ ja $\sigma^2 = p_oq_o/n$. Otetaan testimuuttujaksi vastaava (0, 1)-normaalisti jakautunut (eli standardoitu) satunnaismuuttuja

$$Z = \frac{P - p_o}{\sqrt{p_oq_o / n}} = \frac{X / n - p_o}{\sqrt{p_oq_o / n}}.$$

Esim. 5 Tuotteen (esim. tietyn automerkin, pesuaineen tms.) markkinaosuus oli yhtenä vuonna 23 % . Kun kyseltiin 200 henkilöltä, jotka olivat käyttäneet tällaista tuotelajia (esim. jotain automerkkiä, jotain pesuainetta tms.), kuinka moni käytti juuri tätä tai vastaavaa tuotetta seuraavana vuonna, saatiin tulokseksi, että tällaisia oli 62. Laske, onko tämän tuotteen markkinaosuus kasvanut *merkittävästi* ts. 1 % merkitsevyystasolla (vrt. s. 88).

Tehdään hypoteesi ja oikeanpuoleinen vastahypoteesi

$$H_o\colon\ p = 23\,\% = \ 0{,}23\ ,\ \ H_v\colon\ p > 0{,}23\,.$$

Testisuureen arvoksi saadaan

$$Z = \frac{62/200 - 0{,}23}{\sqrt{0{,}23\cdot(1-0{,}23)/200}} \approx 2{,}69\,.$$

Koska kaksisuuntaisessa testissä merkitsevyystasoa 1 % eli luottamustasoa 99 % vastaava kriittinen arvo on 2,33 (s.88) ja arvo 2,69 ylittää sen, niin nollahypoteesi pitää hylätä ts. vastahypoteesi pitää paikkansa eli markkinaosuus on kasvanut.

Lasketaan vielä, mitä merkitsevyystasoa arvo 2,69 vastaa:

$$P(Z > 2{,}69) = 1 - \Phi(2{,}69) \approx 1 - 0{,}9964 = 0{,}64\,\%\,.$$

b) *Kahden suhteellisen osuuden vertailu:*

Jos satunnaismuuttujat $P_1 = X_1/n_1$ ja $P_2 = X_2/n_2$ ilmoittavat tietyn tapauksen A suhteelliset osuudet kahdessa riippumattomassa otoksessa, joiden avulla pyritään vertaamaan esim. kahden tuotteen suhteellisia osuuksia P_1 ja P_2 keskenään, niin lasketaan ensin suhteellisten osuuksien painotettu keskiarvo

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

ja asetetaan hypoteesi H_o: molempien populaatioiden todelliset, tuntemattomat osuudet p_1 ja p_2 ovat yhtä suuret eli $p_1 - p_2 = 0$. Sopiva testisuure on nyt

$$Z = \frac{P_1 - P_2 - (p_1 - p_2)}{\sqrt{P(1-P)(1/n_1 + 1/n_2)}} = \frac{X_1/n_1 - X_2/n_2}{\sqrt{P(1-P)(1/n_1 + 1/n_2)}}\,,$$

joka on riittävän suurilla otoskokojen arvoilla likimain (0, 1)-normaalisti jakautunut.

10.6 Käytetyn jakauman sopivuus tilastoaineiston malliksi

Edellä on monissa kohdissa käytetty tilastollisen aineiston (otoksen) mallina esim. normaalijakaumaa. Tutkitaan nyt, *kuinka merkittävästi jokin havaintoaineisto poikkeaa sen mallina käytetystä teoreettisesta jakaumasta.*

Ajatellaan, että n:n suuruisessa otoksessa on saatu erillisten tapahtumien (*luokkien*) $A_1, ..., A_r$ frekvensseiksi vastaavasti $f_1, ..., f_r$ ja mallina käytetyn teoreettisen jakauman mukaan näiden frekvenssien pitäisi olla $g_1, ..., g_r$. [Monissa oppikirjoissa frekvenssien f_i ja g_i tilalla käytetään merkintöjä o_i (= observed frequences) ja e_i (= expected frequences).]

Tehdään nollahypoteesi H_o : havainnot noudattavat käytettyä teoreettista jakaumaa, jolloin jokainen $f_i = g_i$ $(i = 1, ..., n)$. Käytetään testisuureena lukua

$$K = \sum \frac{(f_i - g_i)^2}{g_i}.$$

Voidaan osoittaa, että K on riittävän suurilla n:n arvoilla likimain χ^2-jakautunut, vapausasteena $f = r - 1 - k$, missä k on käytetyn teoreettisen jakauman sellaisten parametrien lukumäärä, jotka joudutaan arvioimaan otoksen perusteella (esim. normaalijakaumaa käytettäessä yleensä $k = 2$, sillä μ ja σ joudutaan arvioimaan otoksesta).

Esim. 6 Monisteen alussa sivuilla 2–5 käytettiin esimerkkinä arvosana-jakaumaa

x_i:	0	1	2	3	4	5
f_i:	1	2	4	13	5	2

jossa esim. arvosana 3 on 13:lla opiskelijalla, ts. välillä 2,5...3,5 olevien arvosanojen luokan frekvenssi on 13. Tämän jakauman keskiarvoksi ja otoshajonnaksi saadaan $\bar{x} = 2{,}925$ ja $s_{n-1} = 1{,}141$. Tutkitaan, *voidaanko tämän otoksen katsoa olevan peräisin normaalijakaumasta, jonka odotusarvo* $\mu = 2{,}925$ *ja keskihajonta* $\sigma = 1{,}141$. Tehdään nollahypoteesi, että näin on asian laita.

Lasketaan ensin normaalijakauma-mallin antamat (teoreettiset) todennäköisyydet. Jos $X \sim N(2{,}925;\ 1{,}141^2)$, niin arvosanoja 0, 1, ..., 5 vastaavien luokkien

$$A_1 = (-\infty, 0.5],\ A_2 = (0.5, 1.5],\ \ldots,\ A_5 = (3.5, 4.5],\ A_6 = (4.5, \infty)$$

todennäköisyydet ovat

$$P(A_1) = P(X \le 0.5) = F(0.5) = \Phi\left(\frac{0.5 - \mu}{\sigma}\right) = 0.017,$$

$$P(A_2) = F(1.5) - F(0.5) = \ldots = 0.089,$$
$$\vdots$$
$$P(A_5) = F(4.5) - F(3.5) = \ldots = 0.224,$$
$$P(A_6) = P(X \geq 4.5) = 1 - F(4.5) = 0.084$$

Kun nämä todennäköisyydet kerrotaan otoskoolla n = 27, saadaan eri luokkia A_i vastaavat teoreettiset frekvenssit:

$$g_1 = 0.452,\ g_2 = 2.402, \ldots, g_6 = 2.265\,.$$

Testisuureen K arvoksi saadaan K = 3,68. Vapausasteiden lukumäärä $f = 6 - 1 - 2 = 3$. Esimerkiksi 95 % luottamustasoa eli merkitsevyystasoa $\alpha = 0{,}05$ ja kaksisuuntaista testausta vastaava luottamusväli on $k_{1-\alpha/2} \leq K \leq k_{\alpha/2}$. Täksi väliksi saadaan [esim. laskemalla Mathcadillä arvot qchisq(0.025,3) ja qchisq(0.975,3)] $0.216 \leq K \leq 9.348$. Koska laskettu K:n arvo kuuluu tähän väliin, nollahypoteesi voidaan hyväksyä.

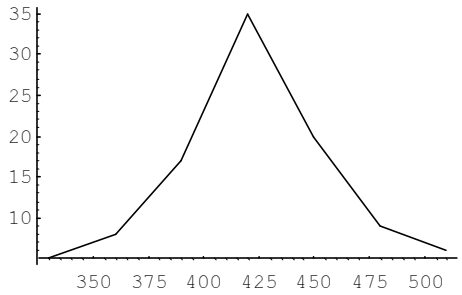
Huom. Oikeastaan edellisen esimerkin otoskoko on liian pieni, jotta yhteensopivuustestiä voitaisiin käyttää. Nyrkkisääntönä testin käyttämiseen on, että jokaisen g_i-luvun pitäisi olla vähintään 5.

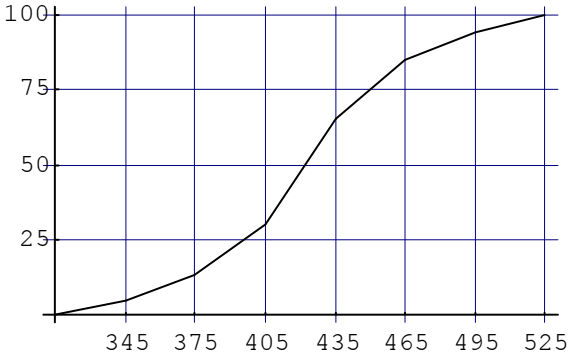
Harjoituksia

Suunnittele itse (tai etsi kirjallisuudesta) eri kohtiin liittyviä esimerkkejä ratkaisuineen. Vieraskielisessä ja samoin suomenkielisessäkin kirjallisuudessa merkinnät ja käytetyt sanonnat tai nimikkeet vaihtelevat. Tarkoitus on laatia esitykset käsillä olevan monisteen merkintöjen ja sanontojen mukaisiksi. Harjoitukset voitaisiin tehdä ryhmätöinä ja välittää muille seminaariesitys-tyyppisinä.

Vastauksia

1.1 $26-\frac{2}{7}=25\frac{5}{7}$, 25,64, 25,56 **1.2** 85 **1.3** 7,28 % **1.4** f) $\bar{x}\approx 6{,}84$, *Mo* = 7, *Md* ≈ 6,8, *vaihteluväli* = 5, *Q* ≈ 0,83, *keskipoikkeama* ≈ 0,830 , $\sigma \approx 1{,}09$, *s* ≈ 1,11 (jos kyseessä on otos), *varianssi* ≈ 1,19, *otosvarianssi* ≈ 1,23. **1.5**





Md ≈ 420, *Q* ≈ 27 (= (450-396)/2), välillä 400 …460 on ≈ 54 % (= 81,5 – 27,5), $\bar{x}\approx 422$, $\sigma \approx 42{,}8$ **1.6** 21,8…33,0 ja 20,0…34,8 **1.7** vähintään 24 g **1.8** on juuri ja juuri (rajat 691,0…700,2) **1.9** vähintään 62.

2.1 d) μ = 5,5 ja σ ≈ 1,98 **2.2** luettele ensin kaikki mahdolliset kolmikot (kr, kr, kr), (kr, kr, kl), (kr, kl, kr), (kl, kr, kr) jne.

$$x_i: 0,\ 1,\ 2,\ 3 \qquad p_i: \frac{1}{8},\frac{3}{8},\frac{3}{8},\frac{1}{8} \qquad F(x)=\begin{cases}0, \text{ kun } x<0\\ 1/8, \text{ kun } 0\le x<1\\ \vdots\\ 1, \text{ kun } x\ge 3\end{cases}$$

2.3 1,5 ja 0,75 **2.4** 4,3 eur (*X* ilmoittaa voiton suuruuden yhden arvan nostossa. *X*:n arvot 0, 10, 50, 200 jne.) **2.5** *k* = 15 (sillä $\sum p_i = 1$), $\mu = 2\frac{1}{3}$, σ ≈ 1,25 **2.6** a) *k* =1/2 b) 0,230 c) 0,770

d) $F(x)=0, \text{ kun } x\le 0$, $F(x)=-\frac{1}{2}(\cos x-1), \text{ kun } 0<x\le\pi$ ja $F(x)=1, \text{ kun } x>\pi$. **2.7** *c* = 1, *a* = 3 **2.8** 0,9104 **2.9** 0,09 % **2.10** a) vp $=1-P(|X|<k)=1-P(-k<X<k)=\ldots$ b) 0,895 **2.11** 27 % 0,0228 **2.12** 0,8413 0,7475 0,3781 **2.13** 2,3 % **2.14** 96% **2.15** σ= 0,405 mm, μ =34,63 mm.

3.1 17 **3.2** a) 19/28, b) 41/56 **3.3** a) 1/6, b) 1/6, c) 5/6 **3.4** a) 1/12, b) 1/12, c) 11/12, d) 11/12, e) 1/6, f) 5/6 **3.5** A: "hylätty matematiikassa", B: "hylätty fysiikassa" , $A \cup B$: " ...", $A \cap B$: " ..." , $P(A \cup B) = ... = \frac{1}{5}$ **3.6** a) 33/100, b) 1/5, c) 3/50, d) 47/100 **3.7** a) $(21-2i)/100$ $(i = 1,...,10)$, b) 0,29, c) 24/25 **3.8** 1/3 **3.9** 0,02

4.1 a) 25/36, b) 1/169 **4.2** a) 5/36, b) 5/18 (vrt. Esim. 6 c)) **4.3** a) $65 \cdot 10^{-9}$ b) $\approx 3{,}34$ % (vrt. Esim. 6 c)) **4.4** a) 1/120, b) 7/24, c) 7/40 (vrt. Esim. 6 c)) **4.5** a) 5/49, b) 45/4753 ≈ 0,00947 c) 880/4753 ≈ 0,185 **4.6** a) 0,378 ≈ 0,38, b) 0,042, c) 0,456 ≈ 0,46, d) 0,154 ≈ 0,15, e) 0,166 ≈ 0,17 **4.7** a) 13/28, b) ≈ 0,98 **4.8** 19/900.

5.1 a) 729, b) 648 **5.2** a) 40320, b) 5040 (ensimmäisen valitsema paikka on samantekevä, mutta sen jälkeiset paikat ovat eriarvoiset suhteessa ensimmäisen valitsemaan) **5.3** a) 360, b) 15 **5.4** a) 234, b) 78, c) 153 **5.5** $\approx 2{,}63 \cdot 10^{35}$ **5.6** 144 **5.7** 120 (= 6!/3!) **5.8** a) 45 (kättely-yhdelmien määrä, mutta yhdelmistä aina kaksi koskee samoja henkilöitä), b) 28 **5.9** 15625 **5.10** a) 45, b) 21 **5.11** 350 $\left(= \binom{7}{3} \cdot \binom{5}{2}\right)$ **5.12** 210 **5.13** 126 $\left(= \binom{10}{5} : 2\right)$

6.1 a) 0,54 %, b) 1 – 0,000015 = 0,999985 (ts. lähes varmasti) **6.2** 0,28 **6.3** 0,23 **6.4** 0,238 (hypergeom. jakauma, Toisin: tulosjonon v v m m m todennäköisyys on $\frac{6}{10} \cdot \frac{5}{9} \cdot \frac{4}{8} \cdot \frac{3}{7} \cdot \frac{2}{6}$, samoin muiden. Tulosjonoja on yhtä monta kuin mahdollisuuksia sijoittaa 2 valkoista palloa 5 paikalle **6.5** $144/625 \approx 0{,}230$ (binomijakauma) **6.6** noin 0,50 (Poisson-jakauma) **6.7** a) n. 0,346 b) n. 0,317 **6.8** Ei, sillä kyseinen todennäköisyys on alle 0,5 (= 0,395) **6.9** - 25 eur. Useassa pelissä keskimäärin hävitään 25 eur kerralla. **6.10** a) laske ensin pistetodennäköisyydet samaan tapaan kuin veikkausjakaumalla , b) $E(X) = 4/3$, $\sigma \approx 0{,}943$. **6.11** a) X:n arvot x_i: $0, 1, 2, 3, 4$ ja näiden todennäköisyydet p_i: $\frac{1}{81}, \frac{8}{81}, \frac{24}{81}, \frac{32}{81}, \frac{16}{81}$ d) 8/3 8/9 **6.12** a) n. 0,48 (binomitodennäköisyys) b) n. 0,43 (additiivisuus ja kertosääntö) **6.13** a) 0,00148 Toisin: additiivisuus ja kertosääntö b) 0,0735 **6.14** a) x:n arvot ovat 0, 1, 2, 3 ja näiden todennäköisyydet ovat n. 0,67, 0,29, 0,04, 0,002 (summa = 1) b)

0,38 **6.15** 2/3 **6.16** 0,0983 **6.17** 0,080 **6.18** a) 0,10 b) 0,122 **6.19** λ.

7.1 1/3 **7.2** a) 1/5, b) 0,33 **7.3** $-\frac{\ln(4/5)}{3} \approx 0{,}07$ **7.4** a) $1/e \approx 0{,}37$ b) $1-e^{-1{,}5} \approx 0{,}78$ **7.5** Olettaen, että jakauma on eksponenttijakauma, tulos on $(e-1)/e \approx 0{,}63$ **7.6** a) 0,58 b) 0,32 **7.7** Vastaukset löytyvät tekstistä **7.8** c) 1/18, $1/(3\sqrt{2})$.

8.1 Odotusarvot: 7, 2, 9, 15 ja varianssit: 11,7 1 14,7 134,3 **8.2** 60,7 (sillä $E(X^2) = \sum {x_i}^2 p_i$) 5 95,7 **8.3** y_j: 1 4 9, q_j: $\frac{1}{2}$ $\frac{1}{4}$ $\frac{1}{4}$ **8.4** a) x_i: 1 3 p_i: 0,5 0,5 y_j: -3 2 4 q_j: 0,4 0,3 0,3 b) 2 0,6 $E(XY) = \sum x_i y_j p_{ij} = \ldots = 0$ 5 9,6

c) $P(X=1, Y=-3) = 0{,}1$ mutta $P(X=1)\cdot P(Y=-3) = 0{,}5\cdot 0{,}4 = 0{,}2$ d) 1 ja $\sqrt{9{,}24} \approx 3{,}0$ e) $-1{,}2$ ja $\approx -0{,}4$ **8.7** Vastaus Esimerkissä 5 **8.8** b) Esim. tiheysfunktiot ovat riippumattomilla satunnaismuuttujilla $f_{XY}(u) = \int_0^\infty \frac{1}{x} f_X(x) f_Y(\frac{u}{x})\, dx$ ja $f_{Y/X}(u) = \int_0^\infty x f_X(x) f_Y(ux)\, dx$.

c) $f_{X+Y}(u) = \begin{cases} u, \text{ kun } 0 \le u \le 1 \\ 2-u, \text{ kun } 1 \le u \le 2 \\ 0 \text{ muulloin} \end{cases}$ d)

9.1 c) Mathcadillä:

$$\sum_{k=4}^{5} \text{dbinom}\left(k, 50, \frac{1}{4}\right) = 6.548 \cdot 10^{-3}$$

$$\text{pnorm}\left[(10.5), 50\cdot\frac{1}{4}, \sqrt{50\cdot\frac{1}{4}\cdot\frac{3}{4}}\right] - \text{pnorm}\left[(3.5), 50\cdot\frac{1}{4}, \sqrt{50\cdot\frac{1}{4}\cdot\frac{3}{4}}\right] = 0.255$$

9.2 Jakauma on likimain $N(\sum \mu_i, \sum {\sigma_i}^2)$ **9.3** 0,9876

9.4 $[\overline{X} - z_{\alpha/2}\cdot\sigma/\sqrt{n},\ \overline{X} + z_{\alpha/2}\cdot\sigma/\sqrt{n}] = [86{,}9\,;\,93{,}7]$ **9.5** [86,4 ; 94,2] (siis hiukan laajempi väli kuin edellisessä tehtävässä).

9.6 $86{,}3 \le \mu \le 94{,}3$ $4{,}36 \le \sigma \le 10{,}4$. **9.7** Mathcadillä:

qnorm(0.91, 0, 1) = 1.34076 qt(0.91, 11) = 1.432

qchisq(0.91, 11) = 17.653 qchisq(0.09, 11) = 5.405

Tehtävien 9.4 – 9.6 vastaukset näillä kriittisillä arvoilla (jos käytetään normaaleja katkaisusääntöjä; oikeastaan ehkä alaraja pitäisi pyöristää alaspäin ja yläraja ylöspäin, mutta joka tapauksessa luottamusvälit ovat likimääräisiä): [88,4; 92,3] [88,2; 92,4] [87,7; 92,9] [4,91; 8,87].

Liite: Standardinormaalin jakauman kertymäfunktion arvoja

Esim. $\Phi(1{,}23) = 0{,}8907$, $\Phi(1{,}237) \approx 0{,}8907 + \frac{7}{10} \cdot (0{,}8925 - 0{,}8907) \approx 0{,}8920$

t	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
0,3	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999