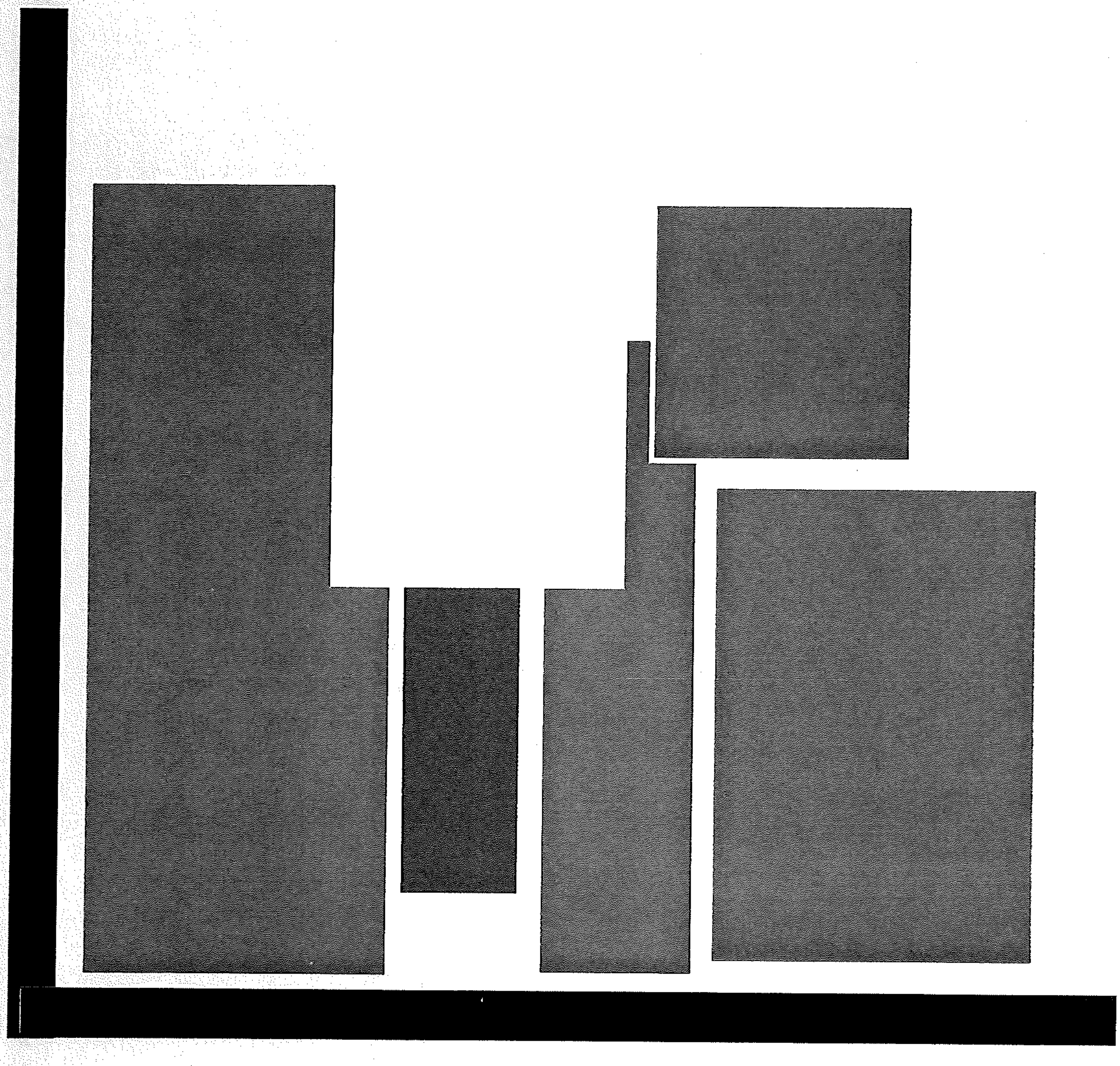
ALPO ÄIJÄLÄ



Tammertekniikka

TODENNÄKÖISYYSLASKENTA
JA
TILASTOTIEDE

ALPO ÄIJÄLÄ

Todennäköisyyslaskenta ja tilastotiede

Tammertekniikka

ISBN 951-9004-50-5

ALKUSANAT

Teknillisten oppilaitosten insinöörikoulutuksessa on jo pitkään kaivattu todennäköisyyslaskennan ja tilastotieteen oppikirjaa. Uudet ammattikorkeakoulut ovat osaltaan lisänneet tällaisen oppikirjan tarvetta. Toivon tämän kirjan osaltaan poistavan kyseisen puutteen teknillisten oppilaitosten sekä ammattikorkeakoulujen insinöörikoulutuksen osalta.

Matemaattisten taustatietojen osalta oppikirjassa oletetaan, että lukija tuntee yhden muuttujan differentiaali- ja integraalilaskennan perusteet. Tämän vuoksi kirja ajoittuu ensisijaisesti toiseen opiskeluvuoteen.

Todennäköisesti aikaresurssit eivät salli koko oppikirjan läpikäymistä. Opettajan harkinnan mukaan voidaan tällöin lähinnä jättää käsittelemättä osia neljästä viimeisestä luvusta.

Tampereella kesäkuussa 1993

Alpo Äijälä
matematiikan yliopettaja

SISÄLLYSLUETTELO

1 TILASTOLLISEN AINEISTON ESITTÄMINEN

1.1 MITTA-ASTEIKOT

Olkoon tarkastelun kohteena tietty **perusjoukko (populaatio)** E, jonka alkiot a_i ovat tutkimusobjekteja. Näitä sanotaan tilastotieteessä tavallisesti **tilastoyksiköiksi**. Tilastoyksiköihin a_i liittyviä tiettyjä ominaisuuksia x, y, z, ... sanotaan puolestaan **tilastollisiksi muuttujiksi.**

Tutkimuksen kannalta on tärkeää, että tutkittava ilmiö voidaan esittää numeerisessa muodossa eli on mahdollista suorittaa tilastoyksiköihin kohdistuvia mittauksia. Itse mittaamisella tarkoitetaan sääntöä, joka liittää tilastoyksikön a_i tarkasteltavaan ominaisuuteen x mittaluvun tai mittasymbolin. Matemaattisesti tulkittuna mittaaminen on tilastoyksiköiden joukossa määritelty reaaliarvoinen funktio f: E → ℝ. Luku $f(a_i) = x_i$ on tilastoyksikköön a_i liitetty **mittaluku.** Mittalukujen x_i muodostamaa joukkoa eli funktion f arvojoukkoa $Aj(f)$ sanotaan **mitta-asteikoksi**. Mittaaminen käsitetään usein laajassa mielessä siten, että mittaustulos voi luvun sijasta olla myös symboli (sukupuoli, väri, ...).

Alkeellisin mitta-asteikko on **luokitteluasteikko** (nominaaliasteikko). Tällöin tilastoyksiköt jaetaan ennalta määrättyihin luokkiin. Kahdesta tilastoyksiköstä voidaan todeta vain kuuluvatko ne samaan luokkaan vai eivät. Luokille voidaan antaa myös numeroarvot. Näillä suoritetuilla aritmeettisilla operaatioilla ei kuitenkaan ole mitään järkevää tulkintaa.

Järjestysasteikko (ordinaaliasteikko) saadaan, kun luokitteluasteikkoon lisätään järjestys tai paremmuus mitatun ominaisuuden suhteen (sotilasarvo, opintomenestys, sosiaaliryhmä, ...). Myöskään järjestysasteikon eri luokille annetuilla numeroarvoilla ei ole järkevää suorittaa aritmeettisia laskutoimituksia. Esimerkiksi luokiteltaessa viranhakijoita eri ominaisuuksien (tietopuoliset ansiot, työkokemus, ...) perusteella paremmuusjärjestykseen ei hakijoiden saamia järjestyslukujen summia saa käyttää lopullista valintaa suoritettaessa (vaikka joskus näkeekin näin meneteltävän).

Välimatka-asteikko (intervalliasteikko) voi luokittelun ja järjestyksen lisäksi ilmoittaa myös luokkien väliset erot numeerisesti. Tällaista asteikkoa varten tarvitaan asteikon mittayksikkö ja nollakohta. Esimerkkinä mainittakoon lämpötilan mittaaminen Celsius-asteita käyttäen.

Suhdeasteikko on muuten samanlainen kuin välimatka-asteikko, mutta siinä on absoluuttinen nollakohta eli kohta, jossa tarkasteltava ominaisuus "häviää". Suurin osa kvantitatiivisista suureista voidaan esittää suhdeasteikolla (pituus, pinta-ala, paine, lämpötila Kelvin-asteissa, asukasluku, ...). Tästä on syntynyt se yleinen harhakäsitys, että kaikki mittaaminen tapahtuu suhdeasteikolla.

1.2 FREKVENSSITAULUKKO JA HAVAINTOMATRIISI

Frekvenssitaulukossa havaintoarvot x_1, x_2, ..., x_n esitetään luokiteltuna. Muuttujan x arvojoukko jaetaan k:hon erilliseen (yhteispisteettömään) puoliavoimeen, usein yhtä pitkään, väliin $(c_1, c_2]$, $(c_2, c_3]$, ..., $(c_k, c_{k+1}]$ ja lasketaan montako havaintoa x_i joutuu kuhunkin väliin. Olkoon luokkiin (väleihin) kuuluvien havaintojen lukumäärät eli **luokkafrekvenssit** $f_1, f_2, \ldots, f_k$ sekä luokkien keskikohdat eli **luokkakeskukset** $m_1, m_2, \ldots, m_k$, missä

$m_i = \dfrac{c_i + c_{i+1}}{2}$. Luvut c_i ovat **todelliset luokkarajat**. Havaintoarvojen kokonaismäärä n on

luokkafrekvenssien summa: $n = \sum_{i=1}^{k} f_i$.

Frekvenssitaulukko näyttää periaatteessa seuraavalta:

luokkaväli	luokkakeskus	frekvenssi
$(c_1, c_2]$	m_1	f_1
$(c_2, c_3]$	m_2	f_2
.	.	.
.	.	.
.	.	.
$(c_k, c_{k+1}]$	m_k	f_k

Jos muuttuja x on diskreetti, niin luokkakeskukset ovat samalla yleensä muuttujan mahdolliset arvot. Jos x on jatkuva tai sen arvojoukossa on äärettömän monta erilaista mahdollista arvoa, niin hyvän yleisnäkemyksen vuoksi luokkien lukumäärän tulisi olla pieni. Toisaalta, mitä suurempi on luokkavälin pituus, sitä suurempi on virhe korvattaessa havaittu arvo vastaavalla luokkakeskuksella. Paljon sovelletun yleisohjeen mukaan luokkien lukumäärän (ilman nollaluokkia) tulisi symmetrisessä aineistossa olla välillä $\sqrt[3]{n} \cdots \sqrt{n}$, missä n on havaintojen lukumäärä. Epäsymmetrisessä (vinossa) aineistossa on syytä käyttää useampia luokkia.

Mitattaessa tilastoyksiköistä samalla kertaa useiden eri muuttujien x_1, x_2, ... , x_m arvoja esitetään mittaustulokset tavallisesti (ainakin eräissä tilastollisissa tietokoneohjelmissa) **havaintomatriisin** muodossa. Havaintomatriisilla tarkoitetaan taulukkoa, jossa eri muuttujien $x_1, x_2, \ldots, x_m$ arvot on taulukoitu tilastoyksiköittäin $a_1, a_2, \ldots, a_n$ seuraavasti:

	x_1	x_2	...	x_m
a_1	x_{11}	x_{21}	...	x_{m1}
a_2	x_{12}	x_{22}	...	x_{m2}
.	.	.		.
.	.	.		.
.	.	.		.
a_n	x_{1n}	x_{2n}	...	x_{mn}

Rivillä i olevat **soluarvot** x_{1i}, x_{2i}, ..., x_{mi} ilmoittavat tilastoyksikköön a_i liittyvät eri muuttujien saamat arvot.

1.3 OTOKSEN GRAAFINEN ESITTÄMINEN

Tarkastellaan esimerkin valossa välimatka- ja suhdeasteikolla käytettäviä esitystapoja.

Esim. Mitattiin 32 työntekijän erään työvaiheen suorittamiseen kuluva aika ja saatiin seuraavat tulokset (aika sekunneissa 0,1 s tarkkuudella).

35,6	36,6	36,3	35,3	35,4	34,5	36,0	34,9
34,6	35,3	34,6	34,6	35,0	33,9	34,9	34,7
35,0	33,9	37,3	34,3	35,0	36,3	33,8	35,5
35,3	35,5	34,8	35,4	33,8	36,0	36,5	35,9

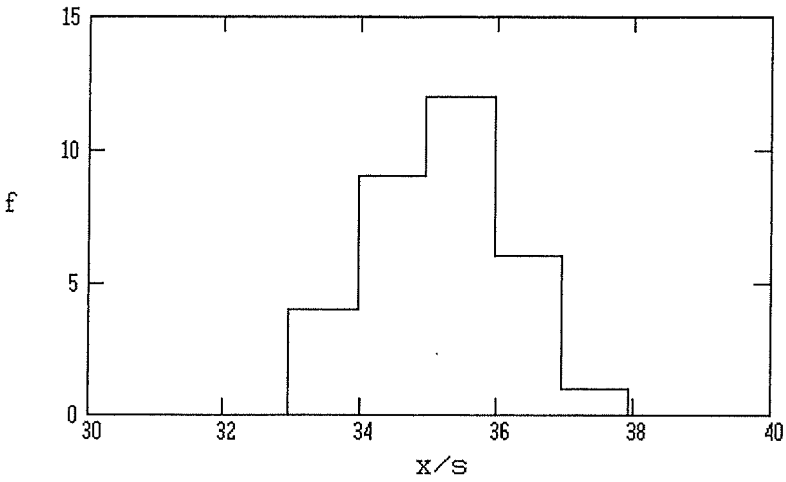
Aluksi aineisto on luokiteltu.

x/s	x_i/s	f_i	f_i/n
32,0-32,9	32,45	0	0
33,0-33,9	33,45	4	0,125
34,0-34,9	34,45	9	0,281
35,0-35,9	35,45	12	0,375
36,0-36,9	36,45	6	0,188
37,0-37,9	37,45	1	0,031
38,0-38,9	38,45	0	0
yhteensä		32	1,000

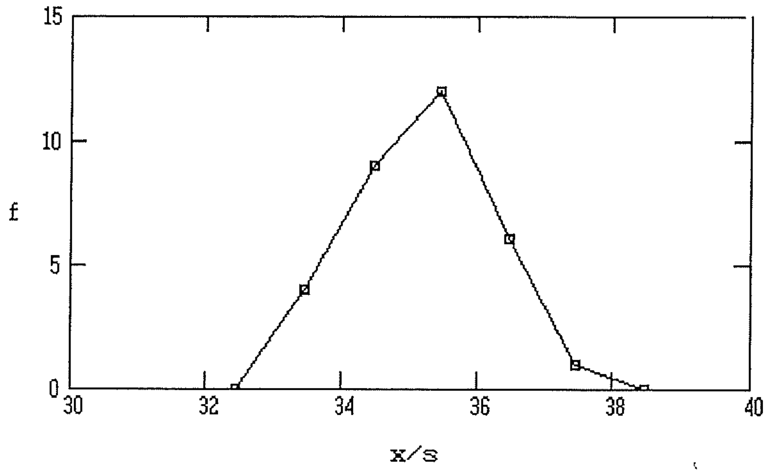
Taulukossa x_i on luokkakeskus ja f_i vastaava luokkafrekvenssi. Taulukkoon on myös laskettu luokkien **suhteelliset frekvenssit** f_i/n. Koska mittaustarkkuus on 0,1 s, niin todelliset luokkarajat ovat 31,95 s, 32,95 s, ..., 38,95 s.

Kuvassa 1.1 on esitetty tarkastellun jakauman **frekvenssihistogrammi** ja kuvassa 1.2 **frekvenssimonikulmio**. Havainnollisuuden lisäämiseksi histogrammiin piirretään usein myös nollaluokat histogrammin molemmille puolille.

Frekvenssimonikulmio saadaan histogrammista yhdistämällä vaakajanojen keskipisteet. Frekvenssimonikulmio alkaa nollaluokan keskipisteestä ja päättyy nollaluokan keskipisteeseen. Havaitaan, että histogrammi ja frekvenssimonikulmio rajoittavat x-akselin kanssa yhtä suuret pinta-alat.



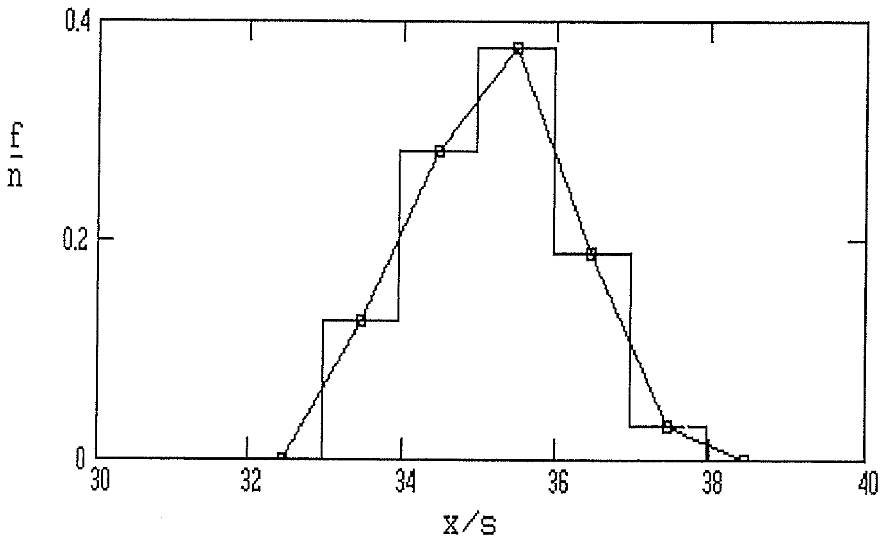
Kuva 1.1: Frekvenssihistogrammi



Kuva 1.2: Frekvenssimonikulmio

Suuressa aineistossa saadaan luokitusta tihentämällä frekvenssimonikulmion kulmat "häviämään", jolloin se lähestyy tasaisesti kaartuvaa käyrää. Tämän hypoteettisen rajakäyrän voidaan ajatella kuvaavan tutkittavan muuttujan teoreettista jakaumaa. Teoreettisia jakaumia tarkastellaan myöhemmissä luvuissa.

Korvaamalla frekvenssit f_i suhteellisilla frekvensseillä f_i/n saadaan **suhteellinen frekvenssihistogrammi** ja **suhteellinen frekvenssimonikulmio** (kuva 1.3). Jos suhteelliset frekvenssit ilmaistaan prosentteina, saadaan **prosenttinen jakauma.** Suhteellisten frekvenssijakaumien avulla on helppo vertailla eri kokoisia aineistoja toisiinsa.



Kuva 1.3: Suhteellinen frekvenssihistogrammi
ja frekvenssimonikulmio

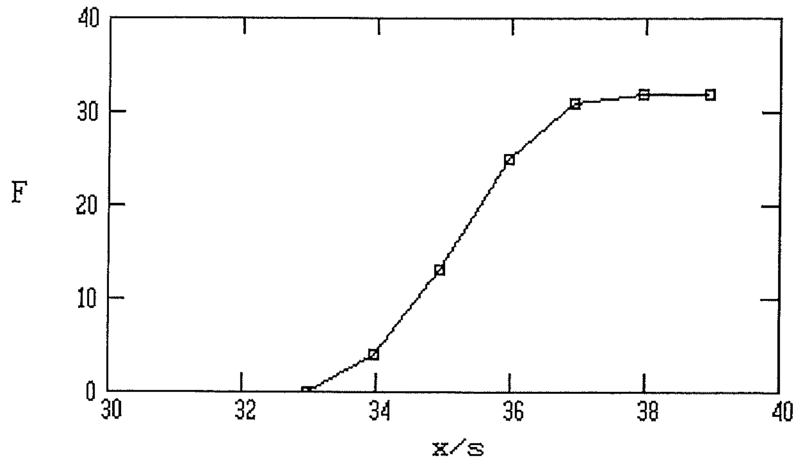
Frekvenssien avulla voidaan laskea **summafrekvenssit**

$$F_i = f_1 + f_2 + \ldots + f_i.$$

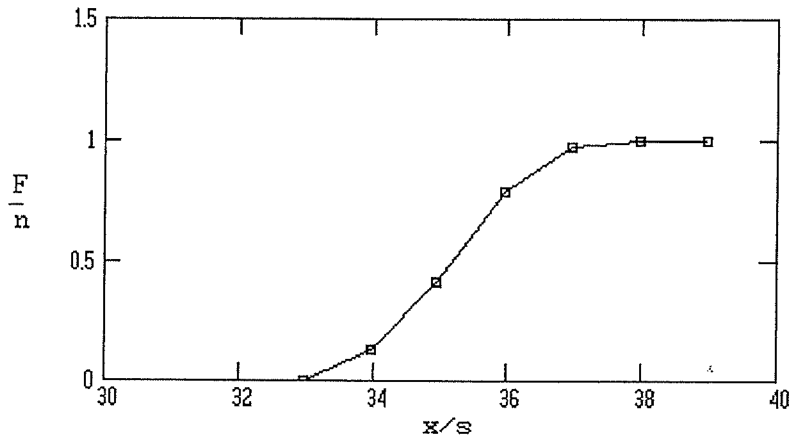
Tietyn luokan summafrekvenssi ilmaisee kaikkien niiden havaintojen lukumäärän, jotka ovat pienempiä tai yhtä suuria kuin ko. luokan todellinen yläraja. Vastaavaa taulukkoa sanotaan **summajakaumaksi** ja sen graafista esitystä **summakäyräksi** (kuva 1.4).

Huomaa, että ***kunkin luokan summafrekvenssiä vastaava piste on merkittävä luokan ylärajan kohdalle.***

x_i/s	f_i	F_i	F_i/n
32,45	0	0	0,000
33,45	4	4	0,125
34,45	9	13	0,406
35,45	12	25	0,781
36,45	6	31	0,969
37,45	1	32	1,000
38,45	0	32	1,000



Kuva 1.4: Summakäyrä



Kuva 1.5: Suhteellinen summakäyrä

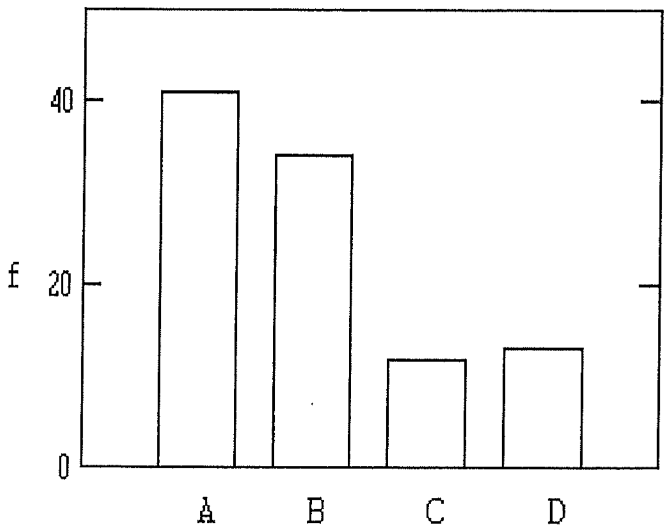
Jos summafrekvenssien F_i sijasta ilmoitetaan **suhteelliset summafrekvenssit** F_i/n saadaan **suhteellinen summajakauma**. Tämän graafinen esitys on **suhteellinen summakäyrä** eli **otoskertymäfunktion kuvaaja** (kuva 1.5). Otoskertymäfunktiota vastaava teoreettinen käsite **kertymäfunktio** näyttelee keskeistä osaa todennäköisyysjakaumien käsittelyssä.

Luokittelu- ja järjestysasteikolla esittämiseen soveltuu histogrammin sijasta **jana- ja pylväsdiagrammi.** Janadiagrammissa kunkin luokan frekvenssiä vastaa jana ja pylväsdiagrammissa pylväs. Joskus käytetään myös sektori- ja tilavuusdiagrammeja. Sektoridiagrammissa sektorin pinta-ala kuvaa tarkasteltavan luokan frekvenssiä. Tilavuusdiagrammissa frekvenssejä havainnollistetaan kappaleiden tilavuuksien avulla.

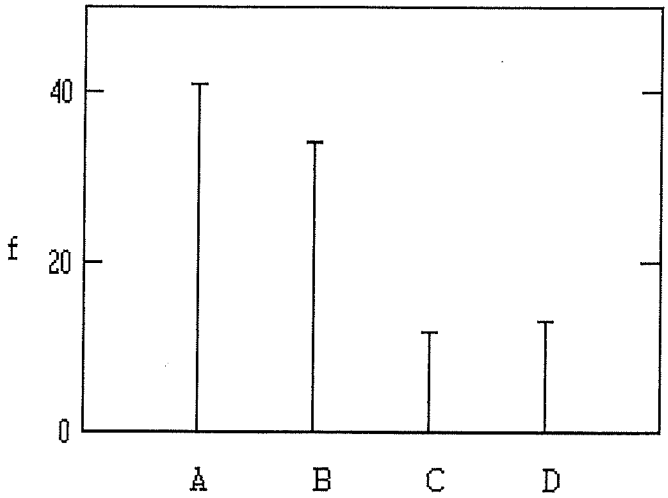
Esim. Mielipidetiedustelussa oli mahdollista vastata neljällä tavalla: (A) kannattaa ehdokasta a, (B) kannattaa ehdokasta b, (C) ei kannata kumpaakaan ehdokasta, (D) ei osaa sanoa. Vastauksista laadittiin frekvenssitaulukko:

mielipide	frekvenssi	suht. frekv.
A	41	41,0 %
B	34	34,0 %
C	12	12,0 %
D	13	13,0 %
yhteensä	100	100,0 %

ja piirrettiin pylväsdiagrammi (kuva 1.6) ja janadiagrammi (kuva 1.7). Käytännössä pylväsdiagrammia sanotaan myös histogrammiksi - varsinkin silloin, kun pylväät sivuavat toisiaan.



Kuva 1.6: Pylväsdiagrammi



Kuva 1.7: Janadiagrammi

HARJOITUSTEHTÄVIÄ

1.1 Luokittele seuraava tilastoaineisto

26,6	22,9	25,8	23,1	23,8	20,9	25,2	26,9	22,6	27,9
28,0	23,7	23,9	20,5	25,7	27,1	24,7	17,8	23,9	22,8
22,0	23,6	27,5	30,6	21,6	19,0	22,7	26,9	25,5	27,6
27,5	22,1	26,7	27,5	28,3	31,1	32,1	28,8	21,8	23,3

valiten todellisiksi luokkarajoiksi 17,25; 20,25; .... Piirrä frekvenssihistogrammi, frekvenssimonikulmio ja suhteellinen summakäyrä.

2 TILASTOLLISET TUNNUSLUVUT

2.1 YLEISTÄ

Jakauman sisältämä informaatio pyritään keskittämään muutamaan koko tilastoaineistoa kuvaavaan **tunnuslukuun.** Näistä tärkeimmät ovat keskiluvut ja hajontaluvut. Joskus käytetään myös vinouslukuja ja huipukkuuslukuja.

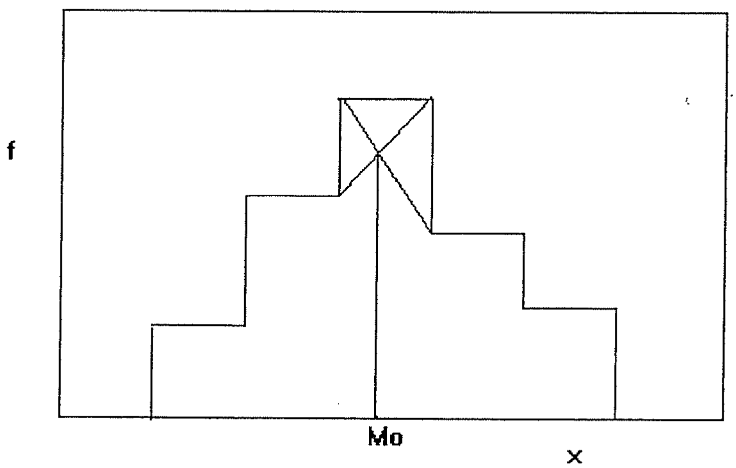
Keskiluku pyrkii esittämään muuttujan havaintoarvojen keskimääräistä suuruutta tai "keskikohtaa". Koska tunnusluvulla voidaan kuvata myös muuta kohtaa tilastoaineistossa, niin puhutaan usein keskilukujen sijasta **sijaintiluvuista.**

Hajontaluku kuvaa puolestaan jakauman "hajaantumista" keskiluvun ympärille. Jos hajontaa ei ole lainkaan, niin se osoittaa havaintoaineiston täydellistä yhdenmukaisuutta eli havaintoarvot ovat identtisiä. Pieni hajonta osoittaa havaintoaineiston keskittymistä. Jos taas hajonta on suuri, niin aineiston yhdenmukaisuus on vähäistä.

2.2 KESKILUVUT

2.2.1 Moodi

Moodi (Mo) eli tyyppiarvo on muuttujan tavallisin (yleisin) arvo. Se ei täten ole aina yksikäsitteinen. Se voidaan määrittää kaikilla asteikoilla. Luokittelu- ja järjestysasteikoilla moodi on se luokka, jolla on suurin frekvenssi. Välimatka- ja suhdeasteikoilla moodiluokan luokkakeskus on käytännössä riittävä moodin likiarvo. Jos halutaan tätä tarkempi arvio, niin voidaan menetellä histogrammissa kuvan 2.1 osoittamalla tavalla.

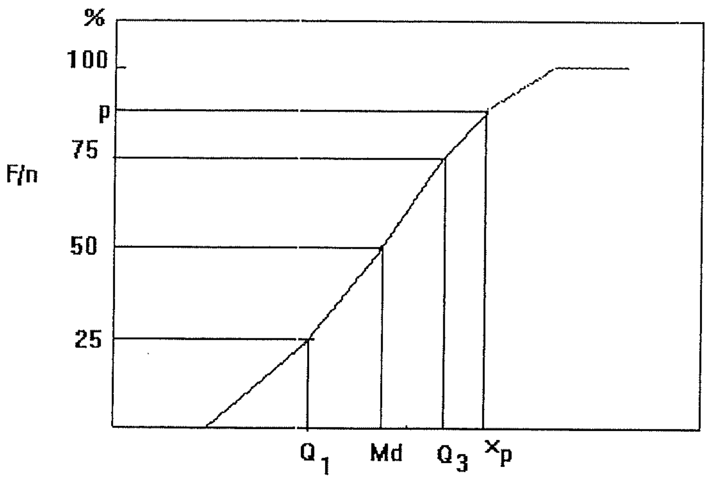


Kuva 2.1: Moodin määrittäminen

2.2.2 Mediaani ja fraktiilit

Mediaani (Md) on lähinnä järjestysasteikon tunnusluku, mutta sopii myös välimatka- ja suhdeasteikoille. Mediaani on muuttujan keskimmäinen arvo. Se jakaa siis havaintoaineiston kahtia.

Mediaani on erikoistapaus **fraktiileista**. p % fraktiili x_p on muuttujan arvo, jonka alapuolella on p % kaikista havaintoarvoista. Fraktiili x_p löydetään otoskertymäfunktion F kuvaajan avulla yhtälön $F(x_p) = p/100$ ratkaisuna (kuva 2.2).



Kuva 2.2: Fraktiilien määrittäminen

Fraktiileista mainittakoon erityisesti **kvartiilit:**

x_{25} = **alakvartiili**, merkitään usein Q_1
x_{50} = **mediaani** Md
x_{75} = **yläkvartiili**, merkitään usein Q_3

2.2.3 Aritmeettinen keskiarvo

Lukujen x_i, $i = 1, 2, \ldots, n$ **keskiarvo**

(2.1)
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Samojen lukujen **painotettu keskiarvo**, kun x_i:n painokerroin on h_i

$$\bar{x} = \frac{\sum_{i=1}^{n} h_i x_i}{\sum_{i=1}^{n} h_i}.$$

Luokitellulle aineistolle

(2.2)
$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n},$$

missä x_i on luokan i luokkakeskus, f_i vastaava luokkafrekvenssi ja $n = \sum_{i=1}^{k} f_i$. Tulos (2.2) voidaan kirjoittaa myös muotoon

(2.3)
$$\bar{x} = \sum_{i=1}^{k} p_i x_i,$$

missä $p_i = f_i/n$ = luokan i suhteellinen frekvenssi (analogia otoksen keskiarvon ja satunnaismuuttujan odotusarvon välillä).

Keskiarvon laskeminen edellyttää vähintään välimatka-asteikon mittauksia.

Keskiarvon ominaisuuksia:

a) Lukujen poikkeamien summa keskiarvosta = 0.
b) Lukujen poikkeamien neliöitten summa luvusta a on pienin, jos a on lukujen keskiarvo.

Ominaisuuksien todistus on sopiva harjoitustehtävä.

2.2.4 Geometrinen ja harmoninen keskiarvo

Positiivisten lukujen x_i, $i = 1, 2, ..., n$ **geometrinen keskiarvo (keskiverto)**

(2.4)
$$G = \sqrt[n]{x_1 x_2 \cdots x_n},$$

ja **harmoninen keskiarvo**

(2.5)
$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}.$$

H on siis lukujen x_i käänteisarvojen $1/x_i$ aritmeettisen keskiarvon käänteisluku. Yleistä käyttöä näillä keskiarvoilla on lähinnä indeksilukujen teoriassa; harmonisella keskiarvolla myös keskinopeutta kuvattaessa.

Esim. Matka s ajetaan k kertaa vakiovauhdeilla v_i, $i = 1, 2, \ldots, k$. Kokonaisaika on tällöin

$$t = \sum_{i=1}^{k} \frac{s}{v_i},$$

ja keskivauhti matkalla ks

$$v = \frac{ks}{t} = \frac{k}{\sum_{i=1}^{k} \frac{1}{v_i}}$$

eli vauhtien v_i harmoninen keskiarvo.

Esim. Suoritusaikojen jakaumalle saadaan:
Mo = 35,45 s (moodiluokan luokkakeskus)
Mo = 34,95 + 3/9 · 1,0 s = 35,28 s (interpoloitu kuvasta 1.1)
md = 35,2 s, Q_1 = 34,4 s, Q_3 = 35,9 s (kuvasta 1.5)
$\bar{x}$ = 35,2 s

2.3 HAJAANTUMISLUKUJA

2.3.1 Vaihteluväli

Vaihteluväli *R* on suurimman ja pienimmän havaintoarvon erotus. Luokitellussa aineistossa vaihteluväli on suurimman luokan todellisen ylärajan ja pienimmän luokan todellisen alarajan erotus. Vaihteluväli on siten yleensä suurempi luokitellulle aineistolle kuin ei-luokitellulle.

2.3.2 Kvartiilipoikkeama

Kvartiilipoikkeama

(2.6) $$Q = \frac{Q_3 - Q_1}{2}$$

voidaan muodostaa, mikäli mittaukset on suoritettu vähintään välimatka-asteikolla.

2.3.3 Keskihajonta ja varianssi

Lukujen x_i, $i = 1, 2, \ldots, n$ **keskihajonta**

(2.7)
$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Varianssi on keskihajonnan neliö σ^2.

Jos luvut x_i ovat vähintään välimatka-asteikkoon liittyviä havaintoarvoja (otos jostain populaatiosta) niin korvataan n luvulla n-1 ja merkitään keskihajontaa kirjaimella s.

(2.8)
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Luokitellulle aineistolle

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k}f_i(x_i - \bar{x})^2},$$

missä x_i on luokan i luokkakeskus ja f_i vastaava luokkafrekvenssi sekä n luokkafrekvenssien summa eli havaintojen kokonaismäärä.

Varianssin lauseke voidaan kirjoittaa muotoon, jossa ei esiinny keskiarvoa $\bar{x}$:

$$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{1}{n}\left(\sum x\right)^2\right].$$

Johto on sopiva harjoitustehtävä. Sama luokitellulle aineistolle

$$s^2 = \frac{1}{n-1}\left[\sum fx^2 - \frac{1}{n}\left(\sum fx\right)^2\right].$$

Esim. Suoritusaikojen jakaumalle saadaan

R = 37,3 s - 33,8 s = 3,5 s (luokittelematon aineisto)
R = 37,95 s - 32,95 s = 5,0 s (luokiteltu aineisto)
Q = (35,9 - 34,4) s/2 = 0,8 s
s = 1,02 s

2.3.4 Variaatiokerroin

Variaatiokerroin V määritellään vain suhdeasteikolla. Variaatiokerroin on keskihajonnan ja keskiarvon osamäärä

$$V = \frac{s}{\bar{x}}.$$

Variaatiokerroin on täten paljas luku. Käytännössä se ilmaistaan usein prosentteina. Koska V on riippumaton käytetystä mittayksiköstä, niin se on kätevä sellaisissa vertailuissa, joissa mittayksiköt ovat hyvin erilaiset, esimerkiksi vertailtaessa kahden huomattavasti eri painoisen tuotteen painojakaumia.

2.4 VINOUS

Tunnetuin vinoutta kuvaava tunnusluku on **Pearsonin vinousmitta** $vinous = \dfrac{\bar{x} - Mo}{s}$.

Empiirisesti on havaittu, että $(\bar{x} - Mo) \approx 3(\bar{x} - Md)$.

Tämän perusteella käytetään usein likiarvotulosta $vinous \approx \dfrac{3(\bar{x} - Md)}{s}$.

Jakauman sanotaan olevan **vino oikealle** (positiivisesti vino), jos vinous > 0 eli $\bar{x} > Mo$. Vastaavasti jakauma on **vino vasemmalle,** jos vinous < 0. Mitä symmetrisempi jakauma on sitä lähempänä nollaa on jakauman vinous.

HARJOITUSTEHTÄVIÄ

2.1 Tunnetaan seuraava luokiteltu tilastoaineisto

luokkakeskus	frekvenssi
12,5	3
17,5	7
22,5	16
27,5	12
32,5	9
37,5	5
42,5	2

a) Piirrä frekvenssihistogrammi ja suhteellinen summakäyrä. Lue kuvista moodi ja kvartiilit.
b) Laske keskiarvo, varianssi ja keskihajonta.

c) Laske geometrinen keskiarvo.
d) Laske harmoninen keskiarvo.

2.2 Lukujoukossa on 20 lukua = a, k lukua = a - 2 ja loput k lukua = a + 2. Kuinka suuri tulee luvun k olla, jotta lukujoukon keskihajonta on ≥ 1?

2.3 Määritä luvut x ja y siten, että lukujen x, y ja 6 keskiarvo on neljä ja lukujen keskihajonta mahdollisimman pieni.

2.4 Osoita, että
a) lukujen poikkeamien summa keskiarvosta = 0,
b) lukujen poikkeamien neliöitten summa luvusta a on pienin, jos a = lukujen keskiarvo.

2.5 Osoita, että lukujen x_i, $i = 1, 2, \ldots, n$ keskiarvo on
$\bar{x} = x_o + \frac{1}{n}\sum_{i=1}^{n}(x_i - x_o)$, missä x_o on ns. väliaikainen keskiarvo. Miten kuuluu vastaava tulos, jos aineisto on luokiteltu?

2.6 Osoita, että lukujen keskihajonta ei muutu, vaikka kaikista luvuista vähennetään sama luku.

3 KOMBINAATIO-OPPI

3.1 TULOPERIAATE JA SUMMAPERIAATE

Merkitään $m(A)$ tarkoittamaan joukon A alkioiden lukumäärää sekä $A \times B$ joukkojen A ja B tulojoukkoa:

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

Näitä merkintöjä käyttäen kuuluu tuloperiaate seuraavasti.

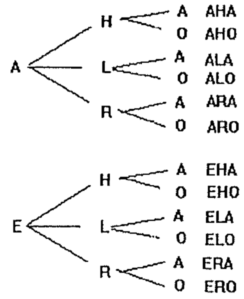
Tuloperiaate:

(3.1) $$m(A \times B) = m(A)m(B)$$

Esim. Muodostettava kolmen kirjaimen jonot siten, että

1.kirjain valitaan joukosta $\{A,E\} = A_1$
2.kirjain valitaan joukosta $\{H,L,R\} = A_2$
3.kirjain valitaan joukosta $\{A,O\} = A_3$

Tehtävänä on itse asiassa muodostaa tulojoukon $A_1 \times A_2 \times A_3$ kaikki alkiot. Näitä on tuloperiaatteen mukaisesti yhteensä $2 \cdot 3 \cdot 2 = 12$ kappaletta. Jonojen muodostamisessa voidaan käyttää apuna oheista "oksakaaviota".



Esim. Olkoon $E = \{1,x,2\}$. Silloin $E^{13} = E \times E \times \ldots \times E$ (13 tekijää) ja $m(E^{13}) = (m(E))^{13} \approx 1{,}6 \cdot 10^6$ (ilmaisee erilaisten veikkausrivien lukumäärän).

Summaperiaate: Jos joukot A ja B ovat erilliset eli $A \cap B = \emptyset$, niin

(3.2) $$m(A \cup B) = m(A) + m(B).$$

Esim. Montako korkeintaan 3-numeroista lukua voidaan muodostaa numeroista 1, 2, 3 ja 4, jos sama numero saa esiintyä luvussa a) vain kerran, b) kuinka monta kertaa tahansa?

a) Merkitään A_i = i-numeroisten lukujen muodostama joukko. Silloin tuloperiaatteen mukaan:

$$m(A_3) = 4 \cdot 3 \cdot 2 = 24$$
$$m(A_2) = 4 \cdot 3 = 12$$
$$m(A_1) = 4$$

Kysyttyjen lukujen määrä on $m(A_3 \cup A_2 \cup A_1) = 24 + 12 + 4 = 40.$

b) $m(A_3 \cup A_2 \cup A_1) = 4^3 + 4^2 + 4 = 84.$

3.2 VARIAATIOT, PERMUTAATIOT JA KOMBINAATIOT

Olkoon $B \subset A$ ja $m(B) = k$, $m(A) = n$ $(k \leq n)$. Jos B käsitetään A:n järjestettynä osajoukkona (jonona), niin sitä sanotaan A:n **variaatioksi** ja jos ei-järjestettynä osajoukkona, niin **kombinaatioksi**. Jos variaatioita muodostettaessa otetaan mukaan kaikki A:n alkiot eli valitaan $k = n$, niin saadaan A:n **permutaatiot**. Joukon permutaatio on siis mikä tahansa joukon kaikista alkioista muodostettu jono.

Esim. Olkoon $A = \{a,b,c\}$.

A:n kaksittain otetut variaatiot:	(a,b), (b,a), (a,c), (c,a), (b,c), (c,b)
ja kaksittain otetut kombinaatiot:	{a,b}, {a,c}, {b,c}.
A:n permutaatiot:	(a,b,c), (a,c,b), (b,a,c), (b,c,a), (c,a,b), (c,b,a).

Lause 3.1: n erilaisen alkion joukolla on

(3.3) $$(n)_k = n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

kappaletta k:ttain muodostettuja variaatioita.

Tod. 1. alkio voidaan valita n eri tavalla,
2. alkio voidaan valita n - 1 eri tavalla,
.
.
.
k. alkio voidaan valita n - k + 1 eri tavalla.

Tuloperiaatteen mukaan variaatioita on $n(n\text{-}1)\ldots(n\text{-}k\text{+}1)$ kappaletta.

Valitsemalla lauseessa 3.1 $k = n$ saadaan

Lause 3.2: n erilaisen alkion joukon permutaatioiden lukumäärä on $n!$.

Lause 3.3: n erilaisen alkion joukolla on

(3.4)
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

kappaletta k:ttain otettuja kombinaatioita.

Tod. Lauseessa 3.1 laskettiin variaatioiden lukumäärä $(n)_k$. Ne variaatiot, joissa on täsmälleen samat k alkiota muodostavat kuitenkin vain yhden kombinaation. Tällaisia samat k alkiota sisältäviä variaatioita on lauseen 3.2 mukaan $k!$ kappaletta (k:n erilaisen alkion erilaisten jonojen määrä). Siis kombinaatioiden lukumäärä on

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

Esim.1. Montako Suomen lipun kaltaista ristilippua saataisiin vain värejä vaihtamalla, jos käytettävissä on 6 erilaista väriä?

Kyseessä on kuuden alkion kaksittain otetut variaatiot, joita on $6 \cdot 5 = 30$ kappaletta.

Esim. 2. Kutsuille osallistui kaikkiaan 6 henkilöä, jotka kaikki kättelivät toisiaan. Kuinka monta kättelyä suoritettiin?

Erona edelliseen on nyt se, että kahden alkion osajoukot tulee käsittää ei järjestettyinä joukkoina. Kyseessä on siis kombinaatioiden lukumäärä $\binom{6}{2} = \frac{6!}{2!4!} = 15$.

3.3 BINOMIKAAVA JA PASCALIN KOLMIO

Potenssin $(a+b)^n$ kehitelmässä termin $a^k b^{n-k}$:n kerroin on $\binom{n}{k}$, sillä $(a+b)^n$ voidaan käsittää tulona:

$$(a+b)^n = (a+b)(a+b)\ldots(a+b) \quad (n \text{ tekijää})$$

ja kerrottaessa on k tekijästä otettava a ja lopuista n-k tekijästä b. Tällaisia valintoja on juuri n alkion k:ttain otettujen kombinaatioiden lukumäärä eli $\binom{n}{k}$. Näin saadaan binomikaava

(3.5)
$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

Erikseen on sovittava, että $0! = 1$ ja $1! = 1$.

Lause 3.4: Binomikertoimet $\binom{n}{k}$ toteuttavat ehdot

a) $\binom{n}{k} = \binom{n}{n-k}$

b) $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$

c) $\sum_{k=0}^{n} \binom{n}{k} = 2^n$.

Todistus on sopiva harjoitustehtävä.
Lauseen 3.4 kohtaan b) perustuu Pascalin kolmion käyttö binomikertoimia määritettäessä.

$$
\begin{array}{c}
1 \quad 1 \\
1 \quad 2 \quad 1 \\
1 \quad 3 \quad 3 \quad 1 \\
1 \quad 4 \quad 6 \quad 4 \quad 1 \\
1 \quad 5 \quad 10 \quad 10 \quad 5 \quad 1 \\
\text{------------------------}
\end{array}
$$

Esimerkiksi viidennen vaakarivin kolmas alkio $\binom{5}{2} = 10$ saadaan laskemalla yhteen neljännen rivin toinen ja kolmas alkio $\binom{4}{1} + \binom{4}{2} = 4 + 6 = 10$.

Lauseen 3.4 kohdan c) mukaan binomikertoimien summa kiinteällä n:llä on 2^n. Joukko-opillisesti tulkittuna tämä merkitsee sitä, että n alkion joukon kaikkien osajoukkojen lukumäärä on 2^n. Mukaan on tällöin luettu myös tyhjä joukko ja joukko itse.

Esim. Kuudesta eriarvoisesta lantista voidaan muodostaa $2^6 - 1 = 63$ erilaista rahasummaa.

HARJOITUSTEHTÄVIÄ

3.1 Montako erilaista lyöntijärjestystä voidaan laatia 9 pelaajan pesäpallojoukkueelle? Entä, jos paras lyöjä lyö neljäntenä?

3.2 Yhdistelmälukossa on kolme numerolevyä, joissa kussakin on 12 valintamahdollisuutta. Lukko avautuu, kun kuhunkin levyyn on valittu oikea vaihtoehto. Kuinka kauan enintään lukon avaaminen kokeilemalla kestäisi, jos kunkin vaihtoehdon kokeilu vie noin 3 sekuntia?

3.3 Kuinka monta sellaista rekisterilaattaa on olemassa, jossa on kolme kirjainta ja enintään kolmenumeroinen luku? Käytettävissä on 26 kirjainmerkkiä. Kaikki luvun numerot eivät saa olla nollia.

3.4 Laatikossa on 4 punaista ja 6 sinistä palloa. Kuinka monella tavalla voidaan laatikosta valita a) kolme palloa, b) kolme sinistä palloa, c) kolme palloa, joista yksi on punainen ja kaksi sinistä?

3.5 Tehtaassa on henkilöiden hakulaitteena valotaulu, jossa kukin viidestä lampusta voi toisistaan riippumatta palaa. Montako erilaista kutsua voi tällä laitteella saada, kun tapausta, jolloin mikään lamppu ei pala, ei katsota kutsuksi? Jos kullekin lampulle sallitaan kolmantena vaihtoehtona vilkkuminen, niin mikä on tällöin erilaisten kutsujen lukumäärä?

3.6 Kuinka monta sellaista veikkaussaraketta on olemassa, jossa on a) 7 ykköstä ja 6 kakkosta, b) 3 ykköstä, 4 ristiä ja 6 kakkosta, c) täsmälleen 10 ykköstä?

3.7 Viidestä naisesta ja kahdeksasta miehestä on valittava komitea, jossa on 2 naista ja 3 miestä. Monellako tavalla komitea voidaan valita, jos a) muita rajoituksia ei aseteta, b) erästä tiettyä naista ei voi valita ja eräs tietty mies on valittava?

3.8 Tarkastellaan jonoja, joissa on kahdeksan kappaletta merkkejä 0 ja 1 (esim. 10001100). a) Montako erilaista jonoa on olemassa? Montako jonoista on sellaisia, joissa on b) 5 nollaa ja kolme ykköstä, c) ainakin viisi nollaa?

HARJOITUSTEHTÄVIEN TULOKSIA

2.1 a) $Mo = 23{,}5$; $Q_1 = 21{,}1$; $Md = 25{,}4$; $Q_3 = 31{,}4$
b) $\bar{x} = 26{,}2$; $s = 7{,}34$ c) 25,2 d) 24,1

2.2 $k \geq 4$

2.3 $x = y = 3$

3.1 362880; 40320

3.2 1,44 h

3.3 $1{,}76 \cdot 10^7$

3.4 a) 120 b) 20 c) 60

3.5 31; 242

3.6 a) 1716 b) 60060 c) 2288

3.7 a) 560 b) 126

3.8 a) 256 b) 56 c) 93

4.1 2/9

4.2 a) 0,016 b) 0,984

4.3 a) 1/3 b) 2/5

4.4 0,416

4.5 0,575

4.6 0,024; 0,188; 0,452; 0,336

4.7 a) 10^{-4} b) $3{,}97 \cdot 10^{-6}$

4.8 0,94

4.9 a) 0,0356 b) 0,0440

4.11 0,491

4.12 0,90

5.1 $p(x) = (6 - |x - 7|)/36$, $x = 2, 3, \ldots, 12$; $\mu = 7$; $\sigma = 2{,}41$

5.2 $p(x) = (2x - 1)/36$, $x = 1, 2, \ldots, 6$; $\mu = 4{,}47$; $\sigma = 1{,}40$

5.3 $\mu = 0{,}21$

5.4 $\mu = \dfrac{2R}{3}$, $\sigma = \dfrac{R\sqrt{2}}{6}$

5.5 0,417

5.6 4/9

5.7 $c = 1/2$, $\mu = \pi/2$, $\rho^2 = \pi^2/4 - 2$

5.8 $E(X) = 4/3$, $Var(X) = 2/9$

5.9 a) 1 b) 2

5.10 a) $\sqrt[3]{\frac{1}{2}}$ b) 0

6.1 b) 0,568 c) $\mu = 5$; $\sigma = 1{,}58$

6.2 0,377

6.3 $P(A) = 0{,}603$

6.4 a) $\mu = 1{,}44$; $\sigma^2 = 1{,}267$ b) $\mu = 1{,}44$; $\sigma^2 = 1{,}12$

6.6 a) 0,978 b) 0,977

6.7 0,62

6.8 0,122

6.9 370

6.10 $2{,}5 \cdot 10^{-7}$

6.11 0,469

6.12 0,78

7.2 $\mu = 5$ min, $P = 0{,}2$

7.4 a) 0,050 b) 0,747

7.6 0,9544; 0,8493

7.7 a) 0,5328 b) 4,07

(Tehtävissä **7.8** ... **7.12** ja **8.3** ... **8.4** on huomioitu jatkuvuuskorjaus)

7.8 a) 0,62 b) 116 cm

7.9 a) 57,0 % b) 0,0 %

7.10 5,0 %

7.11 34,7 mm; 0,48 mm

7.12 0,62

8.1 $E(T) = 22{,}5$; $Var(T) = 22{,}9$; $E(\bar{X}) = 4{,}5$; $Var(\bar{X}) = 0{,}917$

8.2 a) 136,2 kg; 1,44 kg b) 136,2 kg; 0,29 kg

8.3 7

8.4 a) 82 % b) 97 %

8.5 0,94

10.1 59

10.2 [11,986 g;12,074 g]

10.3 [5,26;5,38]

10.4 [2,64;3,76]

10.5 a) 95,5 % b) [0,50;0,56]

11.1 H_0: $\mu = 10{,}0\ N/mm^2$ ei voi hylätä

11.2 a) ei ole mahdollista b) on mahdollista

11.3 ei ole mahdollista

11.6 kyllä

11.7 A on parempi pelaaja

11.8 Väite ei pidä paikkaansa

11.9 a) eivät eroa b) ei ole

11.10 a) on b) ei ole

12.1 [0,82;1,36], ei ole

12.2 $\chi^2 = 0{,}29$; on norm.jak.

12.3 $\chi^2 = 4{,}82$; on Poisson-jakautunut

12.4 $\chi^2 = 7{,}35$; ei ole sama alttius

12.5 Poikkeavat, $\chi^2 = 8{,}3$

12.6 $\chi^2 = 8{,}74$, frekvenssit yhtyvät, 1,4 %

12.7 ei ole eroa, $\chi^2 = 2{,}6$

13.1 a) $y = -75{,}2 + 0{,}845x$ b) $y = 118{,}4 + 0{,}748x$ c) 0,82

d) poikkeaa e) [0,24;0,96]

13.2 $y = 1{,}92e^{-0{,}98x}$

13.3 $y = -1{,}28x^2 + 7{,}42x + 0{,}09$

13.4 $y = 4{,}98 + 1{,}89x^2$

13.5 $a = 1{,}40$; $b = 16000$

TODENNÄKÖISYYSLASKENTA
JA
TILASTOTIEDE

helpottaa merkittävästi tämän alueen opiskelua ammattikorkeakoulujen insinöörikoulutuksessa. Aineen sovellutuksilla on laaja merkitys teknillisten, kaupallisten, yhteiskunnallisten ym. tieteiden aloilla.

Koska kurssia pidetään perinteisesti vaikeana, on kirjan sisältö valittu siten, että differentiaali- ja integraalilaskennassa tarvitaan vain yhden muuttujan funktioiden teoriaa.

Kirjan sisällön pääkohdat ovat:
- kuvaileva tilastotiede
- todennäköisyyslaskenta
- epäjatkuvia todennäköisyysjakaumia
- jatkuvia todennäköisyysjakaumia
- estimointi
- hypoteesin testaus
- regressio ja korrelaatio

Vaikka kirja on laadittu insinöörikoulutusta silmälläpitäen, se on käyttökelpoinen myös muilla ammattikorkeakoulujen opetusalueilla, kuten kaupallisten ja yhteiskuntatieteellisten aineiden opiskelussa.

Tammertekniikka
Hippoksenkatu 21
33530 Tampere
puh. (03) 261 1612
fax (03) 253 0306
e-mail tammerte@sci.fi

ISBN 951-9004-50-5

9 789519 004501