MAB8: TILASTOT JA TODENNÄKÖISYYS II

1. Kombinatoriikkaa 2
2. Todennäköisyyslaskentaa 7
2.1. Kaikki vaihtoehdot yhtä todennäköisiä 7
2.2. Sanastoa ja merkintöjä 8
2.3. Jotkut vaihtoehdot ovat todennäköisempiä kuin toiset 9
2.4. Laskusääntöjä 10
3. Toistokoe ja binomitodennäköisyydet 13
4. LibreOffice Calcista 18
5. Keskiarvo ja -hajonta 27
6. Normaalijakauma 30
6.1. Normaalijakauma toisin päin 36
6.2. Satunnaismuuttujista ja todennäköisyysjakaumista yleisesti 37
7. Tilastotiedettä 40
7.1. Populaatio ja otos 40
7.2. Uskottavin arvo 41
7.3. Otoskeskihajonta 42
7.4. Luottamusväli suurella otoksella 43
7.5. Lisätietoa, vaikea ja vapaaehtoinen 46
8. Lisätehtäviä 48
9. Odotusarvo (kuuluu OPSiin, mutta ei ole pakollinen kurssilla) 50

8. Tilastot ja todennäköisyys II (MAB8)

Tavoitteet

Kurssin tavoitteena on, että opiskelija
- vahvistaa ja monipuolistaa tilastojen käsittelytaitojaan
- osaa määrittää tilastollisia tunnuslukuja ja todennäköisyyksiä jatkuvien jakaumien avulla hyödyntäen teknisiä apuvälineitä
- osaa käyttää teknisiä apuvälineitä digitaalisessa muodossa olevan datan hakemisessa, käsittelyssä ja tutkimisessa, todennäköisyysjakauman odotusarvon ja keskihajonnan määrittämisessä, todennäköisyyksien laskemisessa annetun jakauman ja parametrien avulla sekä luottamusvälin laskemisessa.

Keskeiset sisällöt
- normaalijakauma ja jakauman normittamisen käsitteet
- toistokoe
- binomijakauma
- luottamusvälin käsite

1. Kombinatoriikkaa

Kombinatoriikassa lasketaan, kuinka monella tavalla asioita voidaan järjestää, valita jostakin joukosta tai yhdistellä.

Esimerkki 1.1. Kuinka monessa järjestyksessä kirjaimet A,B,C voidaan kirjoittaa?

Kokeillaan:

ABC,
ACB,
BAC,
BCA,
CAB,
CBA.

Siis kuudella tavalla.

Huomaa, että kirjaimet kirjoitettiin systemaattisella tavalla: Alkukirjaimet käytiin läpi järjestyksessä ja sen jälkeen tulevat kaksi kirjainta aina ensin aakkosjärjestyksessä. Tällä tavalla voidaan olla jokseenkin varmoja siitä, ettei mikään mahdollisuus jäänyt huomoimatta.

Lukumäärän voi laskea myös kirjoittamatta yhdistelmiä paperille. Alkukirjaimia on kolme, ja jokaisen niistä jälkeen kirjoitetaan kaksi jäljelle jäävää kirjainta kummassakin järjestyksessä, yhteensä järjestyksiä on siis $3 \cdot 2 = 6$.

Esimerkki 1.2. Kuinka monessa järjestyksessä kirjaimet A,B,C,...,M voidaan kirjoittaa?

Nyt erilaisia kirjainjonoja on liikaa kirjoitettavaksi, mutta voimme silti laskea niiden määrän. Ensimmäiseksi kirjaimeksi meillä on 13 vaihtoehtoa. Kun se on kirjoitettu paperille, toinen kirjain voi olla jokin 12:sta jäljelle jääneestä. Kolmanneksi kirjaimeksi vaihtoehtoja on 11 jne. Lopulta erilaisten järjestysten lukumääräksi saadaan

$$13 \cdot 12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 6227020800.$$

Tätä numeroa voidaan merkitä myös lyhyemmin 13!, luetaan “kolmentoista kertoma”.

Yleisesti siis:

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

$\boxed{n \text{ kappaletta esineitä voidaan laittaa } n!\text{:ään eri järjestykseen.}}$

Tehtävä 1. Laske kynällä ja paperilla a) 4! b) 5! c) 6!.

LibreOfficen taulukkolaskentaohjelmalla esimerkiksi viiden kertoma saadaan kirjoittamalla `=fact(5)`. TI-Nspirellä se on `5!`.

Tehtävä 2. Laske laskimella tai LibreOfficella a) 10! b) 20!.

Tehtävä 3. Kuinka monella eri tavalla kirjaimet F,G,...,S,T voidaan järjestää? Vinkki: laske kirjainten lukumäärä sormilla.

Esimerkki 1.3. Kuinka monella tavalla kirjaimet A,B,C,D,E voidaan järjestää siten, että vokaalit tulevat ennen konsonantteja?

Vokaaleja ovat A ja E, ja ne voidaan järjestää kahdella tavalla. Konsonantteja on kolme kappaleta, ja ne voidaan järjestää 3!:lla tavalla. Kirjaimet voidaan siis järjestää vokaalit ensin $2 \cdot 3! = 2 \cdot 6 = 12$ tavalla.

Tehtävä 4. Kuinka monella tavalla kirjaimet E,F,G,...,N,O voidaan laittaa järjestykseen niin, että a) vokaalit ovat ennen konsonantteja b) konsonantit ovat ennen vokaaleja?

Järjestykseen laittamista nimitetään usein *valinnaksi ilman takaisinpanoa*. Ajatus tämän nimityksen takana on, että meillä on vaikkapa numeroituja palloja pussissa, josta niitä nostetaan sokkona. Valinnassa ilman takaisinpanoa palloja nostetaan yksitellen, eikä nostettuja palloja laiteta takaisin pussiin. *Valinnassa takaisinpanolla* nostetaan pallo, merkitään ylös sen numero, ja laitetaan se pussiin ennen seuraavaa nostoa. Eri tulosten määrän laskeminen valinnassa takaisinpanolla on helppoa: ensimmäisessä nostossa on n vaihtoehtoa, toisessa on n vaihtehtoa jne. Yhteensä vaihtoehtoja on siis $n \cdot n \cdots n$ kappaletta.

n:stä pallosta voidaan nostaa k kappaletta takaisinpanolla n^k eri tavalla.

Tehtävä 5. Montako viisikirjaimista sanaa voidaan muodostaa kirjaimista A,B,C,D,E? Sanojen ei tarvitse tarkoittaa mitään, eli esimerkiksi DCCDC kelpaa sanaksi.

Tehtävä 6. Montako nelinumeroista pin-koodia on olemassa?

Vaikeampi osio alkaa. Tämän täydellinen ymmärtäminen ei ole ensisijaisen tärkeää, mutta siitä on kyllä hyötyä.

Esimerkki 1.4. (Perustuu tositapahtumiin). Oppilas on unohtanut kaappinsa tunnusluvun. Kaksi tunnusluvun numeroista on samoja, ja loput eivät. Montako erilaista yhdistelmää numeroista saa?

Tätä esimerkkiä voi lähestyä monella tavalla, kuten useimpia todennäköisyyslaskennan kysymyksiä. Intuitiivisin tapa, ja myös käytännössä hyödyllisin on valita ensin kahden saman luvun paikat ja sitten laittaa loput kaksi numeroa paikoilleen. Merkitään numeroita A,A,B ja C. Tällöin A:n paikat olisivat:

AAXX
AXAX
AXXA
XAAX
XAXA
XXAA

Jokaisella rivillä B ja C voidaan laittaa paikoilleen kahdella eri tavalla, joten erilaisia yhdistelmiä tulisi 12 kappaletta. Tämä menetelmä auttaa meitä myös eri yhdistelmien läpikäynnissä.

Toista tapaa varten kuvitellaan, että A:t voitaisiin jotenkin erottaa toisistaan, ja että kaikki mahdolliset yhdistelmät olisi kirjoitettu paperille:

$$A_1A_2BC,$$
$$A_1A_2CB$$
$$A_1BA_2C$$
$$A_1CA_2B,$$
$$...$$

Nyt meillä on $4! = 24$ kappaletta neljän numeron/kirjaimen yhdistelmiä paperilla. Koska A_1 ja A_2 ovat kuitenkin oikeasti samoja, on jokainen yhdistelmä kirjoitettu kaksi kertaa, eli uniikkeja yhdistelmiä on $24/2 = 12$ kappaletta.

Jälkimmäinen tapa yllä mainituista voi olla monesti käytännöllisempi. Suurilla luvuilla on järkevää ensin laskea, onko ylipäänsä mahdollista käydä kaikkia vaihtoehtoja läpi. Toisaalta tällä tavalla voidaan varmistaa, ettei vaihtoehtoja ole unohtunut, ja lisäksi todennäköisyyslaskennassa usein ollaan kiinnostuneita vain vaihtoehtojen lukumäärästä.

Esimerkki 1.5. Montako anagrammia sanasta PAPPA voidaan muodostaa, jos anagrammien ei tarvitse tarkoittaa mitään?

Voitaisiin käydä kaikki vaihtoehdot läpi, mutta on nopeampikin tapa. Ajatellaan aluksi, että kirjaimet ovat kaikki erilaisia: $P_1A_1P_2P_3A_2$. Ne voidaan kirjoittaa paperille $5! = 120$ tavalla. Kuvitellaan, että olemme kirjoittaneet ne paperille. Jos alaindeksit pyyhitään pois, osa anagrammeista on samoja, esimerkiksi $P_1P_2A_1P_3A_2 = P_2P_1A_2P_3A_1$. Sanat ovat muuten samoja, mutta eri P:t ja eri A:t ovat niissä vain eri järjestyksessä.

Nyt ongelma on siis se, kuinka monta kertaa kukin sana esiintyy. Mietitään, kuinka monella tavalla alaindeksit voitaisiin kirjoittaa takaisin sanaan PPAPA. P-kirjaimien alaindekseille on $3! = 6$ eri järjestystä ja A-kirjaimien 2, joten jokainen sana esiintyy indeksittömässä listassa siis $6 \cdot 2 = 12$ kertaa.

Erilaisia sanoja on siis

$$\frac{120}{6 \cdot 2} = 10 \text{ kappaletta.}$$

Tätäkin esimerkkiä voidaan lähestyä monella tavalla. Voitaisiin vaikka kirjoittaa ensin paperille kolme P:tä ja sitten miettiä, monellako tavalla A:t voidaan laittaa niiden väliin ja sivuille. Ota kynä ja paperia esille ja kokeile. Jos A:t laitetaan sanaan vierekkäin, niille on neljä mahdollista paikkaa. Mieti sitten, monellako tavalla A:t voidaan laittaa, jos ne eivät ole vierekkäin.

Tehtävä 7. Kuinka monta anagrammia sanasta SAAPAS voidaan muodostaa, jos sanojen ei tarvitse tarkoittaa mitään?

VAIKEAMPI OSIO PÄÄTTYY.

Esimerkki 1.6. Nostetaan korttipakasta viisi korttia. Monellako tavalla kortit voidaan nostaa, jos nostojärjestyksellä on väliä?

Ensimmäinen kortti voi olla mikä tahansa pakan 52:sta eri kortista. Toinen kortti voi olla mikä tahansa 51:stä jäljellä olevasta, ja niin edelleen, kunnes viides kortti voi olla mikä tahansa 48 jäljellä olevasta. Eri vaihtoehtoja on siis

$$52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 = 311875200 \text{ kappaletta.}$$

Tässä esimerkissä on kaksi erityisen huomionarvoista asiaa: 1) viidennelle kortille ei ole 52-5 eri vaihtoehtoa, vaan 52-4, koska pakasta on poistettu sitä nostettaessa vasta neljä korttia. Tällaiset yhden kortin laskuvirheet ovat hyvin yleisiä. 2) Haluttu luku voidaan esittää muodossa

$$\frac{52!}{(52-5)!} = \frac{52!}{47!}.$$

k:n esineen valinta n:stä ilman takaisinpanoa voidaan tehdä $\frac{n!}{(n-k)!}$ tavalla.

Tehtävä 8. Pussissa on 29 palloa, joissa jokaisessa on eri kirjain. Niistä muodostetaan satunnainen sana nostamalla kuusi palloa yksi kerrallaan. Kuinka monta erilaista sanaa voidaan saada, jos a) pallot laitetaan noston jälkeen takaisin pussiin, b) palloja ei laiteta noston jälkeen takaisin pussiin?

Esimerkki 1.7. Kuinka monta erilaista viiden kortin kättä voidaan nostaa korttipakasta?

Tämä tehtävä poikkeaa aiemmasta sillä tavalla, että korttien nostojärjestyksellä ei ole väliä. Voimme lähestyä ongelmaa samalla tavalla kuin PAPPA-sanan anagrammeja. Ajatellaan, että meillä on luetelttuna kaikki edellisen esimerkin 311875200 eri tapaa nostaa viisi korttia pakasta. Tässä luettelossa jokainen käsi on mainittu monta kertaa, koska sama käsi on saatu nostamalla kortit eri järjestyksessä.

Jokainen viiden kortin käsi voidaan nostaa $5! = 120$ eri järjestyksessä, joten jokainen käsi on kirjoitettu luetteloon 120 kertaa. Erilaisia käsiä on siis

$$\frac{311875200}{120} = 2598960 \text{ kappaletta.}$$

Tästä esimerkistä saadaan yleisempikin sääntö. Kortteja voitiin nostaa

$$\frac{52!}{(52-5)!}$$

eri tavalla järjestyksessä, ja kun tämä lukumäärä jaettiin samanlaisten käsien lukumäärällä, saatiin

$$\frac{52!}{(52-5)!5!}.$$

Esimerkin logiikka ei tietenkään rajoitu korttipakkaan. Jos vaikka 20 henkilön luokasta pitäisi valita neljä henkilöä, voitaisiin heidät valita järjestykessä $20!/16!$ eri tavalla, ja kun järjestys jätettäisiin huomioimatta, olisi erilaisia valintatapoja $20!/(16!4!)$ kappaletta.

n:stä esineestä voidaan valita k kappaletta $\frac{n!}{k!(n-k)!}$:llä eri tavalla.

Tämä on tärkeä asia, ja siksi tälle luvulle on oma merkintä:

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

Vasemmalla oleva merkintä luetaan “än yli koon”.

Tehtävä 9. Laske kynällä ja paperilla a) $\binom{5}{2}$ b) $\binom{5}{3}$. Mitä nämä luvut merkitsevät?

LibreOfficen taulukkolaskentaohjelmalla esimerkiksi $\binom{5}{3}$ lasketaan `=combin(5,3)`, Nspirellä `nCr(5,3)`.

Tehtävä 10. Laske koneella a) $\binom{10}{4}$ b) $\binom{8}{3}$.

Tehtävä 11. Montako neljän kortin kättä voidaan nostaa korttipakan hertoista?

Tehtävä 12. (Vaikeahko). Kuinka monessa viiden kortin kädessä kaikki kortit ovat samaa maata?

Tehtävä 13. Heitetään kahta noppaa (kuten esim. Monopolissa tai backgammonissa). Jos tulos on vaikkapa 2 ja 6, on samantekevää, kumpi noppa on kakkonen ja kumpi kutonen. Montako erilaista tulosta voidaan saada? (Tätä varten kannattaa piirttää jonkinlainen taulukko silmäluvuista).

Tehtävä 14. Lotossa arvotaan 39:stä numerosta 7 voittonumeroa. Kuinka monta erilaista lottoriviä on olemassa? (Unohdetaan lisänumerot toistaiseksi)

Lopuksi vielä, koska n:stä esineestä voidaan valita n:n esineen joukko tasan yhdellä tavalla, ja toisaalta

$$\binom{n}{n} = \frac{n!}{0!n!},$$

haluamme määritellä, että $0! = 1$.

2. Todennäköisyyslaskentaa

2.1. Kaikki vaihtoehdot yhtä todennäköisiä.

Todennäköisyydestä puhutaan usein arkikielessä prosentteina. Esimerkiksi kolikonheitossa sanottaisiin kruunan todennäköisyyden olevan 50%. Matematiikassa käytetään todennäköisyyksinä välin $[0, 1]$ reaalilukuja siten, että 1 tarkoittaa varmaa tapahtumaa, 0 mahdotonta, ja kruunan todennäköisyys olisi 0.5. Laskuissa on parempi käyttää murtolukuja, jos desimaalit eivät lyhyesti ja tarkasti ilmoita lukua. Esimerkiksi luvulle $\frac{1}{10}$ on aivan ok käyttää desimaaliesitystä, mutta luvulle $\frac{1}{7}$ ei. Vastauksessa kannattaa kuitenkin kertoa myös desimaalilikiarvo, koska siitä näkee suoraan suuruusluokan.

Yllä käytännössä jo kerrottiin todennäköisyyksien ensimmäinen periaate: jos vaihtoehtoja on n kappaletta, ja kaikki ovat yhtä todennäköisiä, niin yhden vaihtoehdon todennäköisyys on $1/n$.

Esimerkki 2.1. Heitetään noppaa kerran. Todennäköisyys sille, että silmäluku on 3, on $1/6 \approx 0.167$.

Esimerkki 2.2. Nostetaan kortti hyvin sekoitetusta korttipakasta. Todennäköisyys sille, että kortti on patanelonen on $1/52 \approx 0.0192$.

Esimerkki 2.3. Nostetaan kortti hyvin sekoitetusta pakasta. Mikä on todennäköisyys sille, että kortti on pata?

Ensimmäinen tapa laskea tämä olisi, että maita on neljä kappaletta, joten todennäköisyys on $1/4 = 0.25$.

Toinen tapa on laskea kaikkien patojen määrä ja jakaa se kaikkien korttien määrällä: $13/52 = 1/4 = 0.25$. Tulos on tietenkin sama laskutavasta riippumatta.

Esimerkki 2.4. Pakasta on nostettu patanelonen, eikä sitä ole palautettu pakkaan. Millä todennäköisyydellä seuraava kortti on pata?

Nyt ei voida laskea maiden määrän perusteella, että todennäköisyys olisi $1/4$, sillä patoja on pakassa vähemmän kuin muita maita. Sen sijaan voidaan laskea, että patoja on jäljellä 12 kappaletta ja kortteja 51, joten todennäköisyys on $12/51 = 4/17 \approx 0.235$.

Tehtävä 15. Pakasta nostetaan satunnainen kortti. Millä todennäköisyydellä se on ässä?

Tehtävä 16. Millä todennäköisyydellä yksi rivi voittaa lotossa? (Käytä tehtävää 14 apuna). b) Matti on pelannut kymmenen eri riviä lottoa. Millä todennäköisyydellä hän saa täysosuman?

Tehtävä 17. Annelilla on pussissa 29 palloa, joissa jokaisessa on eri kirjain. Hän arpoo kaksikirjaimisen sanan. Millä todennäköisyydellä sana on "EU", jos a) hän palauttaa ensimmäisen pallon pussiin noston jälkeen b) ei palauta sitä?

2.2. **Sanastoa ja merkintöjä.**

Todennäköisyysksiä lasketaan TAPAHTUMILLE. Tapahtuma voi olla esimerkiksi se, että kolikko on kruuna, tai kortti on pata. Jos tapahtuma ei koostu "pienemmistä" tapahtumista, se on ALKEISTAPAUS. Esimerkiksi tapahtuma "kortti on pata" koostuu tapahtumista "kortti on pataässä, patakakkonen, ..., patarouva tai patakuningas". Kyseessä ei siis ole alkeistapaus. "Kortti on patanelonen" sen sijaan on alkeistapaus, koska sitä ei voi jakaa pienempiin osiin.

Periaatteessa todennäköisyyksiä varten pitäisi aina määrittää ensin mahdolliset tapahtumat jne., mutta yleensä ne ovat itsestäänselviä. Lisäksi niiden kunnollinen käsittely edellyttäisi aika pitkällekin menevää matematiikkaa, ja on omiaan sekottamaan oppilaita, joten ei tuhlata siihen nyt aikaa. Usein tapahtumat eivät kata kaikkia todellisia mahdollisuuksia, esimerkiksi kolikon kyljelleen jäämistä, mutta nämä mahdollisuudet voidaan ohittaa, jos ajatellaan, että kolikonheitto tällaisissa tapauksissa uusitaan niin kauan, kunnes saadaan kelvollinen tulos.

Tapahtuman todennäköisyyttä merkitään $P(\text{Kortti on pata})$.

Jos tarkastellaan esimerkiksi tapahtumaa "kortti on pata", sanotaan alkeistapauksia, jotka tyydyttävät tämän ehdon, SUOTUISIKSI. Siksi yllä käytetty todennäköisyyksien laskutapa ilmoitetaan usein

$$P(A) = \frac{\text{Suotuisten tapausten lukumäärä}}{\text{Kaikkien tapausten lukumäärä}}.$$

Huomaa, että sanalla suotuisa tarkoitetaan vain, että kysytty tapahtuma tapahtuu. Jos tehtävänä on laskea 6-oikein tuloksen todennäköisyys lotossa, ei tulos 7-oikein ole suotuisa, vaikka lottoaja sitä sellaisena varmaan pitäisikin.

Joskus kannattaa nimetä tapahtumia kirjaimilla.

Tehtävä 18. Sovitaan, että

$$A = \text{kortti on pata, ja}$$

$$B = \text{kortti on nelonen.}$$

Laske todennäköisyydet

a) $P(A)$

b) $P(A \text{ tai } B)$

c) $P(A \text{ ja } B)$?

2.3. **Jotkut vaihtoehdot ovat todennäköisempiä kuin toiset.** Todennäköisyys ei riipu aina vain vaihtoehtojen lukumäärästä. Jos juoksisin kilpaa Usain Boltin kanssa, mahdollisia lopputuloksia olisi (ainakin teoriassa) kaksi: minä voitan tai Bolt voittaa. Vaihtoehtojen todennäköisyydet eivät kuitenkaan ole samat.

Jos vaihtoehtoja on äärellinen määrä, ja niiden todennäköisyydet eivät ole samat, pitää seuraavien kuitenkin olla totta:

- Kaikkien vaihtoehtojen todennäköisyyksien summa on 1.
- Jos tapahtuma A koostuu tapahtumista $a_1, a_2, \ldots, a_n$, niin

$$P(A) = P(a_1) + P(a_2) + \ldots + P(a_n).$$

Esimerkki 2.5. Uhkapelurilla on huijausnoppa, jonka todennäköisyydet silmäluvuille ovat nämä:

Silmäluku	P
1	0
2	0.1
3	0.2
4	0.2
5	0.2
6	0.3.

Nämä todennäköisyydet ovat mahdolliset, koska

$P(0)+P(1)+P(2)+P(3)+P(4)+P(5)+P(6) = 0+0.1+0.2+0.2+0.2+0.3 = 1.$

Tässä $P(5)$ tarkoittaa tapahtumaa “silmäluku on vitonen” jne.

Todennäköisyys vaikkapa parittomalle silmäluvulle voidaan laskea näin:

$$P(\text{pariton}) = P(1) + P(3) + P(5) = 0 + 0.2 + 0.2 = 0.4$$

Tehtävä 19. Jos kolikonheitossa kruunan todennäköisyys on 0.54, mikä on klaavan todennäköisyys?

Tehtävä 20. Pelurin nopan todennäköisyydet silmäluvuille 1-5 ovat seuraavan taulukon mukaisia:

Silmäluku	P
1	0.15
2	0.03
3	0.11
4	0.21
5	0.09

Millä todennäköisyydellä hän heittää kutosen? Millä todennäköisyydellä heiton tulos on parillinen?

2.4. **Laskusääntöjä.**

$\boxed{\text{Peräkkäisten tapahtumien todennäköisyydet saadaan kertomalla.}}$

Esimerkki 2.6. Kolikko heitetään kolme kertaa. Millä todennäköisyydellä kaikki heitot ovat kruunia?

Vastaus:

$$P(\text{kruuna, kruuna, kruuna}) = 0.5 \cdot 0.5 \cdot 0.5 = 0.125.$$

Tämä vastaa myös arkijärkeä. Kuvitellaan, että kolikko heitettäisiin vaikka tuhat kolmen sarjaa. Puolessa niistä voisi olettaa ekan heiton olevan kruuna. Näistä sarjoista taas puolet olisi sellaisia, että toinenkin heitto olisi kruuna, ja niistä puolet sellaisia, joissa kolmas on kruuna.

Esimerkki 2.7. Pakasta nostetaan viiden kortin käsi. Millä todennäköisyydellä kaikki kortit ovat herttaa?

Tämän voi ajatella myös peräkkäisinä tapahtumina. Nostetaa kortti kerralla, ja kysytään todennäköisyyttä sille, että jokainen nostettu kortti on hertta. Nyt kuitenkin todennäköisyys hertan nostamiselle muuttuu jokaisen nostetun kortin myötä:

$$\begin{aligned} P(\text{kaikki herttaa}) &= P(H_1) \cdot P(H_2) \cdot P(H_3) \cdot P(H_4) \cdot P(H_5) \\ &= \frac{13}{52} \cdot \frac{12}{51} \cdot \frac{11}{50} \cdot \frac{10}{49} \cdot \frac{9}{48} \approx 0.000495 \end{aligned}$$

Tässä H_2 esimerkiksi tarkoittaa, että toinen nostettu kortti on hertta.

Huomionarvoista on se, että ylläoleva kertolasku voidaan kirjoittaa toisellakin tavalla:

$$\frac{13}{52}\cdot\frac{12}{51}\cdot\frac{11}{50}\cdot\frac{10}{49}\cdot\frac{9}{48}=\frac{13!}{8!}\frac{47!}{52!}.$$

Tästä on ensinäkin hyötyä joskus laskinta näppäiltäessä, mutta tämä vihjaa myös toiseen tapaan ajatella tehtävä: erilaisia viiden kortin käsiä on $\binom{52}{5}$ kappaletta. Erilaisia viiden kortin herttakäsiä taas on $\binom{13}{5}$ kappaletta. Niinpä herttakäden todennäköisyys on

$$\frac{\binom{13}{5}}{\binom{52}{5}}=\frac{13!}{8!\,5!}:\frac{52!}{5!47!}=\frac{13!\,5!\,47!}{8!\,5!\,52!}.$$

Koska 5! supistuu pois, tästä saadaan sama todennäköisyys kuin ensimmäisellä laskutavalla.

Tehtävä 21. Heitetään viittä noppaa. Millä todennäköisyydellä ei saada vitosta pienempiä silmälukuja?

Tehtävä 22. Korttipakasta nostetaan kuusi korttia. Millä todennäköisyydellä kaikissa on eri numero/kirjain?

Tehtävä 23. Uhkapeluri seuraa lottoarvontaa, jossa nostetaan 39:stä numerosta 7 voittonumeroa. Hän on saanut kuusi ensimmäistä numeroa oikein. a) Mikä on todennäköisyys, että seitsemäs numero on oikein? b) Mikä on todennäköisyys, että seitsemäs numero on väärin? c) Mikä on todennäköisyys ennen arvontaa, että kuusi ensimmäistä arvottua numeroa tulevat osumaan pelurin riviin, mutta seitsemäs ei?

Tehtävä 24. Opettaja lyö vetoa, että luokassa on oppilaita, joilla on sama syntymäpäivä (syntymävuosi voi olla eri). Oleta, että jokaisessa vuodessa on 365 päivää. Onko veto opettajan kannalta järkevä, jos oppilaita on a) 12 b) 24? c) Jos karkausvuodet huomioidaan, pieneneekö vai suureneeko opettajan voitonmahdollisuudet?

Seuraavat säännöt on aika helppo nähdä todeksi, ja osaa niistä on jo käytettykin:

$$\boxed{P(\text{ei-}A)=1-P(A)}$$

$$\boxed{P(A\text{ tai }B)=P(A)+P(B)-P(A\text{ ja }B).}$$

Tässä ei-A on nimeltään tapahtuman A KOMPLEMENTTI. Jos esimerkiksi $A=$ “kortti on pata”, niin ei-A on “kortti ei ole pata”, eli “kortti on hertta, ruutu tai risti”. Huomaa, että komplementti *ei* ole tapahtuman A vastakohta. Esimerkiksi lotossa tapahtuman “7 oikein” komplementti ei ole “0 oikein”, vaan “0-6 oikein”.

Toinen tärkeä asia on, että matematiikassa “A tai B” sisältää myös tapauksen “A ja B”. Arkikielessähän “tai” tarkoittaa usein, että pitää valita tasan yksi vaihtoehdoista. Esimerkiksi patanelonen sisältyy tapahtumaan "kortti on pata tai nelonen". Toinen ylläolevista laskusäännöistä tuleekin tästä, patanelonen on laskettu kaksi kertaa mukaan todennäköisyyteen $P(A) + P(B)$, joten toinen tapaus pitää miinustaa pois.

Esimerkki 2.8. Noppaa heitetään kaksi kertaa. Millä todennäköisyydellä ainakin toinen silmäluvuista on kutonen?

Tämän voi laskea kahdella tavalla. Merkitään aluksi $K_1=$ "Eka heitto on kutonen" ja $K_2=$ "toka heitto on kutonen". Nyt

$$\begin{aligned} P(\text{ainakin yksi kutonen}) &= P(K_1) + P(K_2) - P(K_1 \text{ ja } K_2) \\ &= \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36}. \end{aligned}$$

Toinen tapa on käyttökelpoinen hyvin monissa tilanteissa: Todennäköisyys sille, että kummallakaan nopalla ei saada kutosta on

$$\frac{5}{6} \cdot \frac{5}{6} = \frac{25}{36}.$$

Tällöin todennäköisyys sille, että toisella saadaan kutonen, on oltava $1 - \frac{25}{36} = \frac{11}{36}$. Tämän voi pukea fiinimpäänkin muotoon:

$$\begin{aligned} P(\text{ainakin yksi kutonen}) &= P(\text{ei-(molemmilla nopilla 1,2,3,4,5)}) \\ &= 1 - P(\text{molemmilla nopilla 1,2,3,4,5}) \\ &= 1 - \frac{5}{6} \cdot \frac{5}{6} \\ &= 1 - \frac{25}{36} = \frac{11}{36}. \end{aligned}$$

Tehtävä 25. Helsingissä oli 628 208 asukasta 1.1.2017. Heistä 105 240 oli alaikäisiä ja 36 197:n äidinkieli oli ruotsi. Oleta, että äidinkielellä ja henkilön iällä ei ole riippuvuutta (ts. ruotsinkielisten osuus kaikissa ikäluokissa on yhtä suuri). Millä todennäköisyydellä sattumanvaraisesti valittu helsinkiläinen on a) ruotsinkielinen ja alaikäinen b) ruotsinkielinen ja täysikäinen?[1]

Tehtävä 26. Peluri heittää kolikko kymmenen kertaa. Mikä on todennäköisyys, että hän saa ainakin yhden kruunan?

Tehtävä 27. Pelurin kolikko laskeutuu klaava ylöspäin todennäköisyydellä 0.53. Millä todennäköisyydellä hän saa kuudella heitolla ainakin yhden klaavan?

Tehtävä 28. Lotossa arvotaan 39:stä numerosta 7, ja sitä pelataan 7 numeron rivillä. Mikä on todennäköisyys, että rivissä on ainakin yksi numero oikein?

[1] `https://www.hel.fi/hel2/tietokeskus/julkaisut/pdf/17_06_28_Tilastoja_1_Maki_Vuori.pdf`

Tehtävä 29. Korttipakasta nostetaan viiden kortin käsi. Millä todennäköisyydellä ainakin kahdessa kortissa on sama numero/kirjain?

Tehtävä 30. Korttipakasta nostetaan kolme korttia. Millä todennäköisyydellä ainakin kaksi niistä on samaa maata?

Lopuksi vielä, kun nostettaan kortteja pakasta, ja aiemmat nostot vaikuttavat myöhempiin tapahtumiin, voidaan sanoa esimerkiksi, että $\frac{12}{51}$ on *ehdollinen todennäköisyys* sille, että toinen nostettu kortti on hertta ehdolla, että ensimmäinen oli. Tätä merkitään $P(H_2|H_1) = \frac{12}{51}$.

Ehdollisen todennäköisyyden ei tarvitse liittyä peräkkäisiin tapahtumiin tai aikaan. Voidaan esimerkiksi laskea ehdollinen todennäköisyys sille, että heitetään nopalla kutonen, ehdolla, että heitto on parillinen. Koska parillisia numeroita on kolme kappaletta, on

$$P(6|\text{parillinen}) = \frac{1}{3}.$$

Yleisesti ehdollinen todennäköisyys lasketaan tällä tavalla:

$$\boxed{P(A|B) = \frac{P(A \text{ ja } B)}{P(B)}, \text{ kun } P(B) > 0.}$$

Tätä kaavaa ei tarvitse kuitenkaan osata tai ymmärtää. Tällä kurssilla riittää, että ymmärrät perusajatuksen ehdollisesta todennäköisyydestä: patanelosen todennäköisyys on eri kuin patanelosen todennäköisyys ehdolla, että pakasta on jo nostettu kortti.

Tehtävä 31. Mikä on todennäköisyys sille, että toinen pakasta nostettu kortti on pata ehdolla, että ensimmäinen oli hertta?

Tehtävä 32. Laske todennäköisyys sille, että kahden nopan heitolla saadaan ainakin yksi kutonen ehdolla, että molemmat silmäluvut ovat parillisia.

3. Toistokoe ja binomitodennäköisyydet

Toistokokeessa tehdään sama asia monta kertaa, ja lasketaan tulosten lukumäärä. Esimerkkejä toistokoeista:

- Heitetään kolikkoa sata kertaa ja lasketaan kruunien määrä. Tämä voisi olla esimerkiksi uhkapeli, jossa peluria kiinnostaa vain, kuinka monesti hän voittaa.
- Pudotetaan voileipä tuhat kertaa pöydältä ja lasketaan kuinka monta kertaa se putoaa voipuoli ylöspäin. Tämä voisi olla tieteellinen koe, jossa selvitetään putoaako voileipä useammin voipuoli ylöspäin. Minkään yksittäisen pudotuksen tuloksella ei ole väliä, ainoastaan lukumäärällä kokeen lopussa.

- Arvataan kokeen monivalintatehtävässä vastaus jokaiseen kohtaan. Jos kaikkien tehtävien pisteytys on sama, tämä on toistokoe. Oppilasta ei kiinnosta, mihin tehtävään hän on valinnut oikein, vaan ainoastaan pisteet, joita hän saa.
- Kuljetaan junassa pummilla sama väli joka päivä kuukauden ajan. Tässä voitaisiin olla kiinnostuneita vain siitä, onko pummilla kulkeminen halvempaa kuin matkakortin ostaminen.
- Pelataan rivi lottoa joka viikko vuoden ajan. Pelaajaa voi esimerkiksi kiinnostaa, kuinka paljon hän voi odottaa tekevänsä tappiota tällä tavalla.

Seuraava esimerkki kertoo oleellisimman asian toistokokeista:

Esimerkki 3.1. Pelurilla on kolikko, joka antaa kruunan todennäköisyydellä 0.6 ja klaavan todennäköisyydellä 0.4. Hän heittää sitä kymmenen kertaa. Mikä on todennäköisyys, että hän saa neljä kruunaa?

Merkitään kruunia H = heads ja klaavoja T = tails. Eräs tapa saada neljä kruunaa olisi heittosarja

$$\text{H H H H T T T T T T.}$$

Sen todennäköisyys on

$$0.6 \cdot 0.6 \cdot 0.6 \cdot 0.6 \cdot 0.4 \cdot 0.4 \cdot 0.4 \cdot 0.4 \cdot 0.4 \cdot 0.4\cdot = 0.6^4 \cdot 0.4^6.$$

Eräs toinen tapa saada sama tulos olisi

$$\text{H T T H T T H H T T.}$$

Sen todennäköisyys on

$$0.6 \cdot 0.4 \cdot 0.4 \cdot 0.6 \cdot 0.4 \cdot 0.4 \cdot 0.6 \cdot 0.6 \cdot 0.4 \cdot 0.4\cdot = 0.6^4 \cdot 0.4^6.$$

Koska tulosten järjestyksellä ei ole väliä, on samantekevää, missä järjestyksessä kruunat ja klaavat tulevat, jokaisen heittosarjan todennäköisyys on sama, $0.4^6 \cdot 0.4^6$.

Nyt saamme tehtävän ratkaistua, jos osaamme laskea erilaisten nelikruunaisten heittosarjojen lukumäärän. Jos alamme kirjoittaa näitä heittosarjoja paperille, huomaamme pian miten niiden lukumäärä lasketaan.

$$\begin{array}{c}\text{H H H H T T T T T T}\\ \text{H H H T H T T T T T}\\ \text{H H H T T H T T T T}\\ \dots\end{array}$$

Paperille tulisi kaikki kymmenkirjaimiset rivit, joissa neljä kirjaimista on H ja kuusi T. Eli niitä olisi yhtä monta kuin on tapoja valita neljä kymmenestä, $\binom{10}{4}$. Huomaa, että tämä on sama luku kuin $\binom{10}{6}$, eli lopputulos ei riipu siitä, ajatellaanko sitä kruunien vai klaavojen kannalta.

Eri tapoja saada neljä kruunaa kymmenen heiton sarjassa on siis $\binom{10}{4} =$ 210 kappaletta, ja jokaisen niistä todennäköisyys on $0.6^4 \cdot 0.4^6$, joten neljän kruunan todennäköisyys on

$$\binom{10}{4} \cdot 0.6^4 \cdot 0.4^6 = 210 \cdot 0.6^4 \cdot 0.4^6 \approx 0.111.$$

Tästä saadaan yleisempi sääntö:

Binomitodennäköisyydet
Olkoon kolikonheitossa kruunan todennäköisyys p ja klaavan q. Tällöin n heiton sarjassa k:n kruunan saamisen todennäköisyys on $$\binom{n}{k} p^k q^{n-k}.$$

Huomaa: 1) $q = 1 - p$.
2) Tämä sääntö pätee tietenkin kaikkiin muihinkin toistokokeisiin, joissa on kaksi mahdollista tulosta. Sääntö on vain puettu konkreettisempaan muotoon yksinkertaisuuden takia.

Esimerkki 3.2. Kokeen monivalintatehtävässä on 15 kysymystä ja jokaisessa viisi vaihtoehtoa. Oppilas arvaa vastaukset lukematta edes kysymyksiä. Millä todennäköisyydellä hän saa 1) tasan viisi oikeaa vastausta 2) Vähintään viisi oikeaa vastausta?

Tämä on toistokoe, jossa onnistumisen todennäköisyys on $1/5 = 0.2$, joten a-kohdan vastaus on

$$\binom{15}{5} \cdot 0.2^5 \cdot 0.8^{10} \approx 0.103.$$

B-kohdassa voisi laskea todennäköisyydet 6,7,8,..,15-oikein tuloksille, mutta on helpompaa laskea todennäköisyydet 0,1,2,3,4-oikein tuloksille ja sitten vähentää nämäsaatu luku ykkösestä:

$$
\begin{aligned}
P(0 \text{ oikein}) &= \binom{15}{0} 0.2^0 \cdot 0.8^{15} = 0.8^15 \approx 0.035 \\
P(1 \text{ oikein}) &= \binom{15}{1} 0.2^1 \cdot 0.8^{14} \approx 0.132 \\
P(2 \text{ oikein}) &= \binom{15}{2} 0.2^2 \cdot 0.8^{13} \approx 0.231 \\
P(3 \text{ oikein}) &= \binom{15}{3} 0.2^3 \cdot 0.8^{12} \approx 0.250 \\
P(4 \text{ oikein}) &= \binom{15}{4} 0.2^4 \cdot 0.8^{11} \approx 0.188.
\end{aligned}
$$

Näistä saadaan $P(\text{enintään 4 oikein}) \approx 0.836$, joten $P(\text{vähintään 5 oikein}) \approx 0.164$.

Tällaisissa laskuissa pyöristykset voivat merkitä hyvin paljon. Pyöristettäessä kolmeen desimaaliin luku voi muuttua melkein 0.0005:n verran, ja jos pyöristettyjä lukuja lasketaan viisi yhteen, voi virhe olla jo 0.0025:n luokkaa. Siksi:

Älä käytä laskuissa pyöristettyjä lukuja, ellei ole pakko. Pyöristä vasta lopputulos.

Tehtävä 33. Heitetään viittä noppaa. Millä todennäköisyydellä saadaan a) tasan kolme kutosta b) Vähintään kolme kutosta?

Tehtävä 34. Henkilö A matkustaa pummilla saman reitin 40 kertaa kuussa. Sakkojen saamisen todennäköisyys jokaisella matkalla on 0.01. Millä todennäköisyydellä A saa useammat kuin kahdet sakot kuussa?

Binomitodennäköisyyksien järkevyyden ehto on, että aiemmat tapahtumat eivät vaikuta myöhempiin. Joissakin tilanteissa binomitodennäköisyyksiä ei siis voi käyttää

Esimerkki 3.3. Nostetaan pakasta viiden kortin käsi. Millä todennäköisyydellä siinä on tasan kolme herttaa?

Tätä ei voi laskea suoraviivaisesti binomitodennäköisyyksillä, koska jo nostetut kortit vaikuttavat tulevien korttien todennäköisyyksiin. Jos tehtävää mietii vähän aikaa kynän ja paperin kanssa, sille löytyy binomitodennäköisyyksiä muistuttava ratkaisu, mutta jätetään se nyt ylimääräiseksi harjoitustehtäväksi.

Joskus taas aiemmat tapahtumat vaikuttavat myöhempiin, mutta vaikutus on niin pieni, että se voidaan jättää huomioimatta. Esimerkiksi seuraavassa tehtävässä:

Tehtävä 35. Helsingissä on 628 208 asukasta, ja heistä 105 240 on alaikäisiä.
a) Jos valitaan satunnainen helsinkiläinen, millä todennäköisyydellä hän on alaikäinen?
b) Jos valitaan 20 satunnaista helsinkiläistä, millä todennäköisyydellä joukossa on enintään neljä alaikäistä? Oleta, että valinnan todennäköisyys on jokaisella valinnalla vakio (se, jonka laskit a-kohdassa). Nspirellä saat laskettua binomitodennäköisyyden suoraan komennolla `binomPdf(n,p,k)`, missä n on kokeiden lukumäärä, p onnistumisen todennäköisyys ja k onnistuneiden kokeiden todennäköisyys. LibreOfficen ohje on tiivistelmässä ja myöhemmin monisteessa, laatikossa ennen tehtävää 40.[2]
c) Tarkastellaan vielä b-kohdan oletuksen vakiotodennäköisyydestä järkevyyttä. Jos 19 ensimmäistä ryhmään valittua helsinkiläistä oli alaikäisiä, mikä on *oikea* todennäköisyys sille, että myös kahdeskymmenes on? Kuinka paljon se eroaa a-kohdan todennäköisyydestä?

Milloin voi olettaa, että todennäköisyys ei muutu, ja milloin ei? Tämä on suurelta osin makukysymys ja riippuu myös tuloksen odotetusta tarkkuudesta. Koulutehtävissä jotain voi päätellä siitä, kuinka paljon vaivaa tehtävän tekeminen ilman tätä oletusta vaatisi, tai olisiko se ylipäänsä mahdollista. Nyrkkisääntönä, korttipakassa aiemmilla valinnoilla on merkitystä, satunnaista helsinkiläistä valitessa ei.

TERMINOLOGIAA. Kokeessa, jossa heitetään kolikkoa sata kertaa, voidaan kruunien määrää sanoa kokeeseen liittyväksi SATUNNAISMUUTTUJAKSI. Kruunien lukumäärän lisäksi satunnaismuuttujia on näistä riippuvat asiat, esimerkiksi pelurin varallisuus kymmenen kolikonheiton jälkeen.

Esimerkki 3.4. Henkilö matkustaa pummilla 40 kertaa kuussa. Hänen rahanmenetys on satunnaismuuttuja, joka saa arvoja 0 €, 80 €, 160 €,...,3200 €.

Esimerkki 3.5. Galluptutkimuksessa valitaan satunnaisesti 4 000 äänioikeutettua suomalaista, joilta kysytään, ketä he äänestäisivät presidentinvaaleissa, jos se olisi nyt. Sauli Niinistön kannatus on otoksen valinnasta riippuva satunnaismuuttuja, joka voi saada arvoja nollasta prosentista sataan prosenttiin. Sekä nolla prosenttia, että sata prosenttia ovat mahdollisia tuloksia, koska Niinistön kilpailijoilla ja Niinistöllä oli kummallakin ainakin 4 000 äänestäjää. Ne ovat tietenkin hyvin epätodennäköisiä tuloksia.

[2] Jos tehtävässä kysytään vain yhtä todennäköisyyttä, käytä mieluummin binomitodennäköisyyden kaavaa. Jos laskeminen olisi muuten vaivalloista, voit käyttää `binomPdf`:ää tai vastaavaa.

Tehtävä 36. Oletetaan, että Niinistöä kannattaa presidentiksi 62.6% äänioikeutetuista.[3] Laiskassa vaaligallupissa valitaan satunnaisesti sata äänioikeutettua. Millä todennäköisyydellä Niinistön kannatus on gallupissa a) tasan 63% b) tasan 100% c) vähintään 97 % ?

4. LibreOffice Calcista

LibreOffice on ilmainen toimistotyökalu, ja Calc sen taulukkolaskentaohjelma. Se muistuttaa kaupallista Exceliä, ja myös Googlen Sheets on sille sukua. Jos näistä oppii käyttämään yhtä, muutkin ovat aika helppoja. Näitä ohjelmia käytetään erittäin paljon myös tosielämässä. Esimerkiksi tällä hetkellä (18.4.2018) työvoimatoimiston avoimissa työpaikoissa on 22 633 ilmoitusta, joista 480:ssä esiintyy sana “excel”.

LibreOfficen Calc on käytettävissä myös sähköisissä ylioppilaskirjoituksissa. Toinen käytössä oleva väline, TI-Nspire muistuttaa taulukko-ominaisuuksiltaan LibreOfficea, mutta on maksullinen, eikä sitä käytetä lukion jälkeen mihinkään. Jos Nspiren käyttö tuntuu sinusta mukavammalta, voit tehdä tehtävät silläkin, mutta suosittelen LibreOfficea.

Asennus. Googlaa libreoffice ja asenna sitten se koneellesi. Kannattaa valita englanninkielinen versio, koska siihen löytyy helpommin apua, joka on paremmin ajantasalla.

Tietokoneohjelmien käytössä tärkeintä on googlaus-taito.[4] Jos et käytä ohjelmaa pariin vuoteen, et muista miten mitäkin tehdään, eikä sinun tarvitsekaan. Ohjelmat myös muuttuvat harva se vuosi. On silti järkevää opetella takemään asiat ainakin jollakin ohjelmalla, koska kynnys tehdä ne toisella madaltuu, ja samalla oppii yleisiä periaatteita ohjelmien toiminnasta.

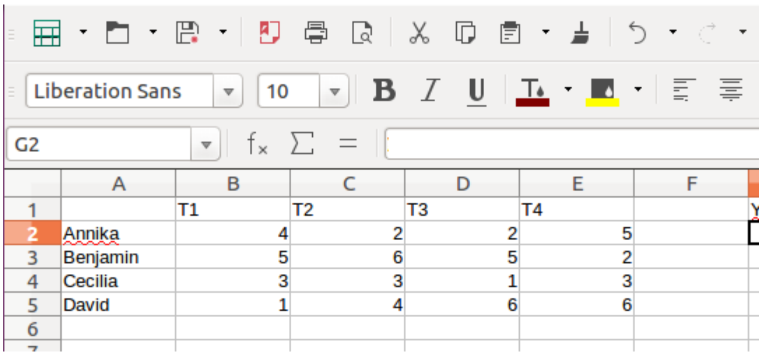
Googlaamisessa toimii yllättävän tarkatkin kysymykset. Kysymyksen lisäksi kannatta kirjoittaa hakuun jotain avainsanoja, kuten libreoffice calc. Jos haet tarkkaa lausetta tai sanajärjestyksellä on väliä, laita ne lainausmerkkeihin. Esimerkiksi haku `''Mauno Koivisto''` palauttaa tulokset, jossa nimet esiintyvät peräkkäin, haku `Mauno Koivisto` antaa myös tuloksia, joissa ei ole Mauno Koivistosta mitään, mutta muista Maunoista ja muista Koivistoista kylläkin.

Tehtävä 37. Mitkä ovat hyviä hakusanoja LibreOffice Calcin käytön aloittamiseen? Kokeile niitä hakukoneeseen ja laita bookmarkkeihin parhaat.

[3] Niinistö voitti tällä kannatuksella vaalit 2018, mutta äänestysprosentti oli vain 66,7%, joten tehtävän oletus on virheellinen. Se tekee laskemisesta kuitenkin paljon mukavampaa.

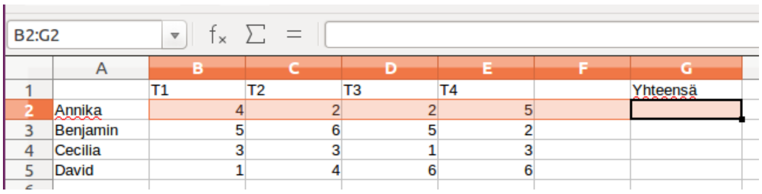
[4] Ajatellaan nyt, että googlaus tarkoittaa hakukoneen käyttä siitä riippumatta, mikä hakokone on kyseessä. Esimerkiksi `duckduckgo.com` on suosittu vaihtoehto niillä, jotka eivät pidä Googlen yksityisyyspolitiikasta.

Tehtävä 38. Katso tehtävän asiat läpi, jos ne ovat sinulle jo entuudestaan tuttuja, älä turhaan näe vaivaa, vaan hyppää seuraavaan tehtävään.
1)Tee seuraavanlainen tiedosto mielikuvitusoppilaiden pisteistä:

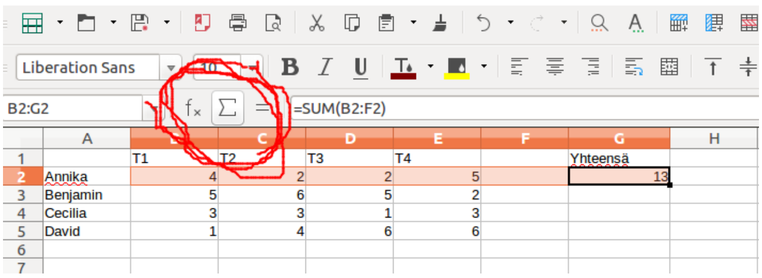


Syötettyäsi numeron pystyt siirtymään alaspäin enterillä ja oikealla sarkaimella (Caps Lockin yläpuolella).

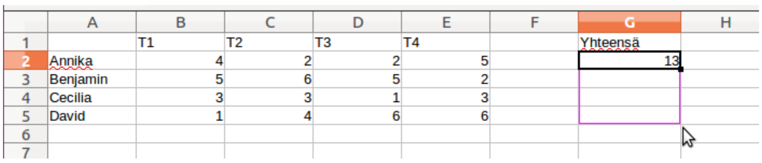
2) Klikkaa ruutua G2 ja maalaa vasemmalle ruutuun B2 asti:



3) Klikkaa ruutujen yläpuolella olevaa isoa sigmaa: Σ . Mitä LibreOffice laski ruutuun G2?



4) Klikkaa ruutua G2. Ruudun oikeassa alakulmassa on pieni musta neliö. Tartu siihen hiirellä ja vedä neljä riviä alaspäin:



5. Tallenna tiedosto .ods- ja .csv-muodoissa: Kun olet klikannut “save as”, aukeaa ikkuna, jossa voit valita tallennuspaikan, tiedoston nimen ja muodon. Tiedoston muoto on todennäköisesti ikkunassa alhaalla

save-napin yläpuolella, voit valita siitä “text csv”. Jos siinä kohdassa lukee “all formats”. Voit päättää tiedoston muodon myös kirjoittamalla nimen päätteeksi “.ods” tai “.csv”, esimerkiksi `mab8_tehtava37.csv`. Tätä tiedostoa tullaan käyttämään seuraavissa tehtävissä.

Tässä tehtävässä oli montakin huomionarvoista seikkaa

(1) Summaa laskettaessa on väliä sillä, miten alue maalataan. LibreOffice laskee summan siihen ruutuun, jota aluksi klikattiin.

(2) Klikkailun ja maalailun sijaan summan voi laskea myös pelkällä näppäimistöällä. Esimerkiksi G2-ruutuun oltaisiin voitu kirjoittaa `=SUM(B2:E2)` . Tässä `B2:F2` ovat alku- ja loppuruutu, josta summa lasketaan. Jos haluaa laskea lukuja yhteen sekä vaaka-, että pystysuunnassa, voi alkuruuduksi valita vasemman yläkulman ja loppuruuduksi oikean alakulman. Esimerkiksi kaikkien oppilaiden kaikkien pisteiden laskeminen yhteen olisi saatu kirjoittamalla johonkin ruutuun `=SUM(B2:E5)`. Kokeile vaikka.

(3) Summaan laskettiin tehtävässä mukaan ruutu F2, mutta koska se on tyhjä, sillä ei ollut vaikutusta.

(4) Hiiren käyttö on usein turhauttavaa, hidasta ja epäergonomista. Summa on aika helppo tehdä sillä, mutta useimmat muut asiat kannattaa tehdä pelkällä näppäimistöllä. Tietokoneen käytössä yleisestikin kannattaa suosia näppäimistöä.

(5) Jos avaat tallennetut tiedostot ja klikkaat ruutua G2, huomaat, että .ods-muodossa siihen on jäänyt =SUM(B2:F2), mutta .csv-muodossa siinä on luku 13. Vastaavasti, jos lisäät Annikalle pisteen vaikkapa tehtävään 2, .ods-muodossa tallennettu tiedosto lisää sen automaattisesti summaan, mutta .csv-muodossa tallennettu ei.

(6) .csv-muoto vaikuttaa edellisen huomion perusteella huonommalta, mutta se on erityisen hyvä raakadatan tallentamiseen, ja se on jopa jonkinlainen standardi siinä hommassa. Csv on lyhenne sanoista comma separated values, kyseessä on käytännössä tekstitiedosto, jossa eri sarakkeet on erotettu pilkuilla ja rivit rivinvaihdoilla. Koska joissakin maissa, esimerkiksi Suomessa, käytetään pilkkua desimaalierottimena, voi sarakkeiden erotin olla jokin muukin, esimerkiksi `&` tai `\t`. Jos käytät csv-tiedostoa, jossa on eri erottimet kuin ohjelmassasi, voit ohjelman säätöjen lisäksi muuttaa erottimia avaamalla sen tekstitiedostona ja käyttämällä find-replace all -toimintoa. Tämä kannattaa kuitenkin tehdä ajatuksen kanssa: jos muutat desimaalierottimet pisteistä

pilkuiksi ja arvojen erottimet pilkuista puolipisteiksi, pitää jälkimmäinen operaatio tehdä ensin. Miksi?

(7) Eri koneilla ja eri ohjelmistoilla desimaalierotin voi vaihdella. Jos desimaalierotin on piste ja kirjoitat LibreOfficessa ruutuun "5,4", se menee ruudun vasempaan reunaan, ja LibreOffice tulkitsee sen tekstiksi eikä numeroksi. Jos erotin on pilkku, saattaa "5.4" muuttua muotoon 5. maaliskuuta. Suosittelen käyttämään pistettä desimaalierottimena. Vaihto-ohjeet löytyy googlaamalla esimerkiksi `libreoffice decimal separator`.

(8) Tiedostoille kannattaa antaa kuvaavat nimet. Tärkeistä tiedostoista kannattaa ottaa varmuuskopiot, eikä kannata olettaa, että liitteenä lähetetty tiedosto päätyy vastaanottajalle. Toisinsanoen, jos lähetät vaikka koulutehtäviä opettajalle, tiedostoa ei kannata tuhota ennen kuin niistä on saatu pisteet, ja mieluiten koko kurssista arvosana.

(9) Kun ruutuun G2 oli laskettu summa ruuduista B2-F2, ja ruutu oli kopioitu alempiin raahaamalla, ei niihin tullut ruutujen B2-F2 summa, vaan vastaavien alempien rivien summa. Tätä sanotaan suhteelliseksi viittaukseksi, ja siitä on huomattavaa hyötyä taulukkolaskennassa.

(10) Syötetyn tekstin ei tarvitse olla isoilla kirjoitettua, esimerkiksi `=sum(B2:E2)` käy aivan hyvin. Isoja kirjaimia käytetään viestittämään, että tämä sana liittyy ohjelmaan itseensä, eikä kulloinkin käsiteltyyn tiedostoon tai tehtävään.

Tehtävä 39. Katso tehtävän asiat läpi, jos ne ovat sinulle jo entuudestaan tuttuja, älä turhaan näe vaivaa, vaan hyppää seuraavaan tehtävään.
Avaa tehtävässä 38 tehty .ods-muotoinen tiedosto.
a) Laske ruutuun G6 oppilaiden saamien kokonaispisteiden summa.
b) Laske ruutuun A6 oppilaiden määrä: `=COUNTA(A2:A5)`.
c) Laske ruutuun H6 oppilaiden kokonaispistemäärän keskiarvo: `=[jokin ruutu]/[jokin ruutu]`.
d) Laske ruutuun E6 kaikkien oppilaiden kaikkien tehtävien pistemäärä suoraan tehtäväruuduista: `=SUM(B2:E5)`.
e) Klikkaa ruutua E6, ota pienestä neliöstä sen vasemmassa alakulmassa kiinni ja vedä alaspäin. Mitä alapuolella oleviin ruutuihin tulee, ja miksi?
f) Poista e-kohdan laskut ja laske ruutuun B6 summa kaikkien oppilaiden tehtävästä 1 saamista pisteistä.
g) Ota ruudusta B6 kiinni ja maalaa oikealle niin, että saat jokaiseen tehtävään yhteispistemäärän.
h) Laske ruutuun B7 ykköstehtävstä saatujen pisteiden keskiarvo tällä tavalla `=[jokin ruutu]/[jokin ruutu]` .

i) Ota ruudusta B7 kiinni ja maalaa oikealle, että saisit jokaisesta tehtävästä keskimääräiset pisteet. Miksi tämä ei toimi?
j) Muuta ruudun B7 sisällöksi `=B6/$A$6`, ota ruudusta kiinni ja maalaa oikealle. Katso ruutujen C7-E7 sisältöä.

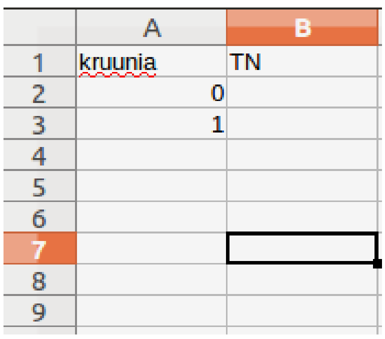
Tämän tehtävän ajatus oli näyttää suhteellisten viittauksien vaara: joskus viitatun ruudun halutaan pysyvän samana. Joskus samana pitäisi pysyä pelkän sarakkeen, joskus pelkän rivin, ja joskus molempien. Muuttumattomuus saadaan laittamalla dollarimerkki muuttumattoman pysyvän asian edelle. `$A6` tarkoittaa, että sarake ei muutu, `A$6` että rivi ei muutu, ja `$A$6` että kumpikaan ei muutu. Tehtävän j-kohdassa oltaisiin saatu sama tulos aikaiseksi myös kirjoittamalla B7-ruutuun `$A6`, koska ruutua maalattiin vain vaaksauorassa. Melkein aina kannattaa kuitenkin nähdä sen verran vaivaa, että kirjoittaa ajatellun sisällön mukaisesti, eikä pienimmällä riittävällä tavalla. Tällä tavoin dokumentista on helpompi ymmärtää, mitä siinä lasketaan, eikä mahdolliset tulevat muutokset tai lisäykset aiheuta ongelmia. Suhteellisista viittauksista saat enemmän tietoa googlettamalla esimerkiksi `libreoffice suhteellinen viittaus`.

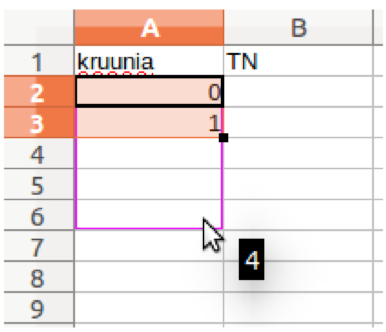
Seuraavassa tehtävässä käytetään funktiota, joka laskee binomitodennäköisyyksiä.

`=BINOM.DIST( 3, 8, 0.6, 0 )`
laskee LibreOfficessa todennäköisyyden saada kolme kruunaa kahdeksalla heitolla, kun kruunan todennäköisyys on 0.6.

Viimeisenä argumenttina olevasta nollasta ei tarvitse vielä välittää. Kirjoita siihen toistaiseksi nolla. Huomaa, että funktion nimeä tai argumenttien järjestystä ei tarvitse muistaa tarkasti, koska libreoffice ehdottaa funktiota ja kertoo argumentit, kun alat kirjoittaa sitä. Lisäksi sen saa selville googlettamalla.

Tehtävä 40. Avaa uusi dokumentti LibreOffice Calcissa ja kopioi siihen ruudut vasemmanpuoleisesta kuvasta. Maalaa sitten ruudut A2 ja A3, tartu A3:n oikeassa alakulmassa olevaan pieneen neliöön ja vedä alaspäin kunnes ruuduissa on numerot 0-10.





a) Ajatellaan koetta, jossa kolikkoa heitetään kymmenen kertaa. Kruunan todennäköisyys on 0.52. Laske ruutuun B2 binomitodennäköisyys nollalle kruunalle. Älä kuitenkaan käytä funktion ensimmäisenä argumenttina lukua 0, vaan viittaa ruutuun A2 ilman dollareita.
b) Ota ruudusta B2 kiinni ja raahaa binomitodennäköisyydet ruutuihin B3-B12.

c) Montako kruunaa on todennäköisin tulos? Entä epätodennäköisin?

d) Laske binomitodennäköisyyksien summa ruutuun B13. Mitä saatu luku kertoo tehtävän ratkaisusta?

Tehtävä 41. Oletetaan, että 62,6% kaikista äänioikeutetuista äänestäisi Sauli Niinistöä, jos vaalit järjestettäisiin nyt.
a) Laiska gallupin tekijä kysyy sadalta satunnaiselta äänioikeutetulta, ketä he äänestäisivät. Millä todennäköisyydellä hän saa Niinistön kannatukseksi 58% -68% (eli n. viiden prosenttiyksikön sisälle todellisesta kannatuksesta)?
b) Vielä laiskempi gallupin tekijä haastattelee vain kahtakymmentä satunnaista äänioikeutettua. Millä todennäköisyydellä hän saa Niinistön kannatukseksi 55% - 70%?
c) Iltapäivälehti kysyy nettisivuillaan lukijoilta, ketä he äänestäisivät presidentiksi. Tulosten mukaan Niinistö ei pääsisi edes toiselle kierrokselle. Onko tämä vain huonoa onnea, vai kenties salaliitto?

Kertymäfunktio. Heitettäessä kolikkoa tuhat kertaa jokainen kruunien lukumäärä on hyvin epätodennäköinen. Tällöin yleensä ollaan kiinnostuneempia todennäköisyydesta saada esimerkiksi 0-100 kruunaa kuin 434 kruunaa. Tehtävässä 41 nähtiin, kuinka näitä todennköisyyksiä voidaan laskea melko vaivattomasti pienille määrille tapauksia. Isommilla määrillä on hyödyllistä käyttä kertymäfunktiota, joka kertoo, mikä on todennäköisyys saada enintään n kappaletta kruunia.

Esimerkki 4.1. Oletetaan, että kolikonheitosta tulee kruuna todennäköisyydellä 0.57. Kolikkoa heitetään kymmenen kertaa. Taulukoidaan jokaisen kruunamäärän todennäköisyys ja kumulatiivinen todennäköisyys enintään n:lle kruunalle:

kruunia	TN	kumulatiivinen TN
0	0.0002161148	0.0002161148
1	0.0028647779	0.0030808927
2	0.0170887332	0.0201696259
3	0.0604066849	0.0805763108
4	0.1401294608	0.2207057716
5	0.2229036074	0.443609379
6	0.2462307291	0.6898401081
7	0.186513642	0.8763537501
8	0.0927146302	0.9690683803
9	0.0273112864	0.9963796667
10	0.0036203333	1

Taulukosta nähdään esimerkiksi, että todennäköisyys saada enintään neljä kruunaa on n. 0.22. Kumulatiivinen todennäköisyys kymmenelle kruunalle on ykkönen, koska on varmaa, että kymmenellä heitolla saadaan enintään kymmenen kruunaa.

Terminologiaan tutustuaksemme, keskimmäistä saraketta sanotaan ensimmäisen *pistetodennäköisyysfunktioksi.* Jos sitä merkitään f:llä, voidaan kirjoittaa esimerkiksi $f(4) \approx 0.140$. Oikeanpuoleinen sarake, eli kumulatiivinen todennäköisyys on taas f:n *kertymäfunktio.* Jos sitä merkitään F:llä, voidaan sanoa vaikkapa $F(7) \approx= 0.876$.

Tarkemmin määriteltynä pistetodennäköisyysfunktion f kertymäfunktiolle F on aina

$$F(n) = f(1) + f(2) + \cdots + f(n).$$

Toinen tapa ajatella kertymäfunktiota on

$$F(n) = P(\text{Enintään } n \text{ kruunaa})$$

Kertymäfunktio saadaan libreofficessa samalla tavalla kuin binomitodennäköisyys, mutta laittamalla viimeisen mystisen nollan paikalle ykkönen: `=binom.dist( 6, 10 , 0.57, 1)` antaa ruutuun 0.6898401081, kuten ylläolevan taulukon mukaan kuuluukin.

Kertymäfunktion suurin etu meille on se, että sen avulla on helppo laskea esimerkiksi

$$\begin{aligned} P(3 \leqq \text{kruunia} \leqq 9) &= P(\text{kruunia} \leqq 9) - P(\text{kruunia} \leqq 2) \\ &= F(9) - F(2) \\ &= 0.9963796667 - 0.0201696259 \\ &= 0.9762100408. \end{aligned}$$

Huomaa, että alarajan kolmonen muuttuu kakkoseksi. Jos miinustaisimme $F(9) - F(3)$, ei lopputulokseen jäisi tapausta kolme kruunaa.

Esimerkki 4.2. Niinistöä kannattaa presidentiksi 62,6% äänioikeutetuista. Ahkera gallupin tekijä haastattelee neljää tuhatta satunnaisesti valittua äänioikeutettua. Millä todennäköisyydellä hän saa Niinistön kannatukseksi 60.6% - 64.6%, eli kahden prosenttiyksikön sisälle todellisesta kannatuksesta?

Koska 4 000 äänioikeutettua on vielä pieni osa kaikista äänioikeutetuista, on tämä käytännössä binomijakautunut toistokoe. Merkitään haastateltujen Niinistön äänestäjien määrää n:llä. Prosenttiyksikköjä 60.6% - 64.6% vastaa $n = 2424 \ldots 2584$, joten olemme kiinnostuneita todennäköisyydestä $P(2424 \leqq n \leqq 2584)$.

Kertymäfunktiolle F pätee

$$F(2584) = P(N \leqq 2584)$$

ja

$$F(2423) = P(N \leqq 2423).$$

Näistä saadaan

$$P(2424 \leqq N \leqq 2584) = F(2584) - F(2423).$$

Laitetaan tämä LibreOfficeen

```
=binom.dist(2584,4000,0.626,1) - binom.dist(2423,4000,0.626,1)
```

ja tulokseksi saadaan n. 0.991.

Tehtävä 42. Alla on painotetun nopan todennäköisyydet. Laske oikeanpuoleiseen sarakkeeseen kertymäfunktio, eli todennäköisyydet, että silmäluku on enintään n.

Tulos	TN	Kertymäfunktio
1	0.2	
2	0.1	
3	0.1	
4	0.3	
5	0.1	
5	0.2	

Tehtävä 43. Kolikko on reilu, eli kruuna ja klaava ovat yhtä todennäköiset. Sitä heitetään kymmenen kertaa. Laske todennäköisyys saada a) enintään 7 kruunaa b) vähintään 2 kruunaa.

Tehtävä 44. Kolikonheitosta saadaan kruuna todennäköisyydellä 0.46. Millä todennäköisyydellä tuhannella heitolla saadaan a) enintään 400 kruunaa b) vähintään 500 kruunaa?

Tehtävä 45. Tutkijat pudottavat voileivän kaksisataa kertaa pöydältä lattialle, ja se putoaa 112 kertaa voipuoli alaspäin (merkitään $n = 112$). Kokeen merkittävyyttä tarkastellaan vertaamalla sitä todennäköisyyteen saada tällainen tai äärimmäisempi tulos, jos eri päin putoaminen olisi yhtä todennäköistä. Käytännössä siis tutkijoiden pitää laskea todennäköisyys

$$P(n \leqq 88) + P(112 \leqq n).$$

a) Laske tämä todennäköisyys
(Lisätietoa: tätä todennäköisyyttä sanotaan kokeen *p-arvoksi*, kun nollahypoteesi on, että molemmin päin putoaminen on yhtä todennäköistä. Lukua $p = 0.05$ pidetään yleensä tilastollisen merkittävyyden rajana. Lukua käytetään siksi, että ennen tietokoneaikaa nämä asiat laskettiin käsin ja taulukoiden avulla, ja yleensä taulkoidut arvot olivat 0.05, 0.01 ja 0.001. Tämä luku siis on jokseenkin mielivaltaisesti valittu).

Tehtävä 46. Helsingin sanomien 22.-24. 1. 2018 teettämään galluppiin[5] vastasi 500 äänioikeutettua. Niinistö kannatus oli 58% ja virhemarginaali 4,5 prosenttiyksikköä.
a) Kuinka montaa galluppiin vastannutta prosenttimäärät 53,5% ja 62,5% vastaavat? (Älä pyöristä vielä lukuja)
b) Oleta, että Niinistöä todellisuudessa kannatti 58% äänioikeutetuista. Millä todennäköisyydellä hänen kannatus on 500 henkilön gallupissa ilmoitettujen virherajojen sisälpuolella? (Että kannatus olisi virherajojen sisäpuolella, sen on oltava suurempi kuin alaraja ja pienempi kuin yläraja. Valitse a-kohdan pyöristykset tämän perusteella)

[5] `https://www.hs.fi/politiikka/art-2000005538817.html`

5. Keskiarvo ja -hajonta

Keskiarvo on varmaan kaikille tuttu: lasketaan luvut yhteen ja jaetaan niiden lukumäärällä.

> **Keskiarvo**
>
> ---
>
> Lukujen $x_1, x_2, \ldots, x_n$ keskiarvo on
> $$\bar{x} = (x_1 + x_2 + \cdots + x_n)/n.$$

Esimerkki 5.1. Oppilaiden A ja B matematiikan kurssien arvosanat ovat seuraavat:

	MAB1	MAB2	MAB3	MAB4	MAB5	MAB6	MAB7	MAB8
A	6	10	10	5	8	9	10	6
B	8	8	7	9	8	8	9	7

Matematiikan arvosana lopputodistukseen lasketaan näiden keskiarvona. A:n keskiarvo on

$$\frac{6+10+10+5+8+9+10+6}{8} = \frac{64}{8} = 8.$$

B:n keskiarvo on myös kahdeksan:

$$\frac{8+8+7+9+8+8+9+7}{8} = \frac{64}{8} = 8.$$

Merkinnällä $\bar{x}$ tarkoitetaan joskus muitakin asioita kuin keskiarvoa. Myöhemmin tulemme käyttämään keskiarvolle myös merkintää μ.

LibreOffice Calcissa keskiarvo esimerkiksi ruuduista A1-A8 saadaan näin: `=AVERAGE(A1:A8)`. TI-Nspiressä `=mean(a1:a5)`.

Tehtävä 47. Oppilaan matematiikan arvosanat kursseilta MAB1-MAB7 ovat

$$6,\ 7,\ 8,\ 6,\ 8,\ 9,\ 6.$$

a) Mikä on hänen matematiikan keskiarvo tällä hetkellä?
b) Mikä arvosana hänen on saatava kurssista MAB8, että keskiarvo olisi tasan 7?
c) Loppuarvosana on kurssien keskiarvo pyöristettynä lähimpään kokonaislukuun (tasan 0.5-loppuinen ylöspäin). Mitkä ovat oppilaan suurin ja pienin mahdollinen loppuarvosana MAB8:n jälkeen?

Esimerkissä 5.1 kummallakin oppilaalla oli sama keskiarvo, mutta toisen arvosanat vaihtelivat aika villistikin, kun toinen oli saanut tasaisesti kaseja muutaman seiskan ja ysin kanssa. *Keskihajonta* mittaa tätä asiaa: kuinka paljon luvut poikkeavat keskimääräisestä. Se on määritelty näin:

Keskihajonta

Lukujen $x_1, x_2, \ldots, x_n$ keskihajonta on

$$\sigma = \sqrt{\frac{1}{n}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)}$$

Esimerkki 5.2. Keskihajonnan laskeminen käsin on melko työlästä, mutta opettavaista, joten lasketaan esimerkistä 5.1 oppilaan A matematiikan arvosanojen keskihajonta taulukon avulla (muistetaan, että $\bar{x} = 8$):

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
6	-2	4
10	2	4
10	2	4
5	3	9
8	0	0
9	1	1
10	2	4
6	-2	4.

Sitten lasketaan yhteen oikeanpuoleinen sarake, summaksi saadaan 30. Jaetaan tämä lukujen lukumäärällä, saadaan $30/8 = 3\frac{3}{4}$, ja siitä otetaan lopuksi neliöjuuri: $\sigma = \sqrt{3\frac{3}{4}} \approx 1.94$.

Mitä keskihajonta kertoo meille? Ei välttämättä vielä paljoakaan. Jos toinen oppilas kertoisi, että hänen arvosanojensa keskihajonta on 1, tietäisimme, että hänen arvosananansa ovat paljon lähempänä keskiarvosa. Jos oppilaan keskihajonta olisi 2.5, ne olisivat kauempana keskiarvosta, ja jos se olisi 0, olisi hänen jokainen arvosana tasan 8.

Tehtävä 48. Esimerkissä 5.1 oppilas B sai arvosanat 8, 8, 7, 9, 8, 8, 9, 7. Laske hänen arvosanojensa keskihajonta.

LibreOfficelle keskihajonnan esimerkiksi ruuduista A1-A10 voi laskea näin: `=STDEVPA(A1:A10)`. TI-Nspiressä `=stdev(a1:a10)`.

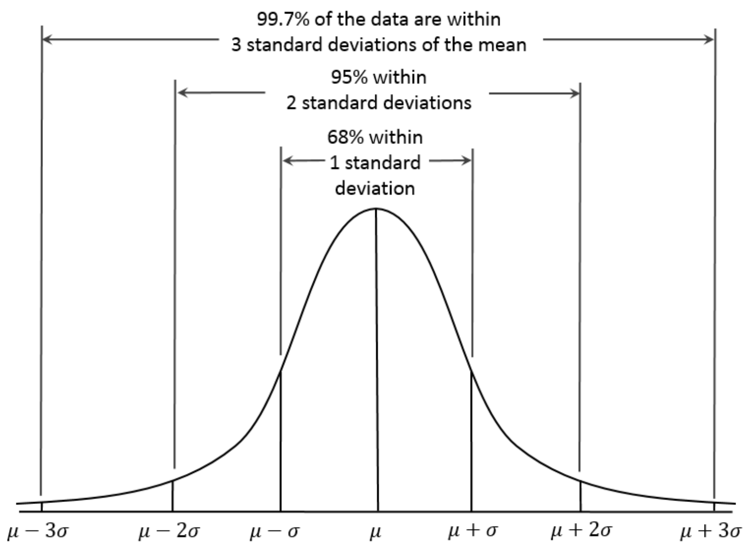
Tehtävä 49. Osoitteessa `https://www.cs.helsinki.fi/u/mokangas/mab8/data1.csv` on kolme sarjaa mittaustuloksia (A, B ja C). Lataa tiedosto ja laske jokaisesta sarjasta keskiarvo ja keskihajonta. (Mittaussarjat ovat eri esineiden terminaalinopeuksia putoamisen aikana).

Huomaa, että tiedosto ei välttämättä aukea koneellasi oikein. Siinä tapauksessa kannattaa kysyä apua. Tiedoston saa helposti näkymään oikein, mutta ohjeita lukemalla se on vaikeaa. Tämä on myös osa tehtävää, sillä tosimaailmassa väärin aukeavat tiedostot ovat yleinen hankaluus, jonka kanssa kannattaa opetella tulemaan toimeen.

Tehtävä 50. Osoitteessa `https://www.cs.helsinki.fi/u/mokangas/mab8/iris.csv` on Kurjenmiekkojen lehtien pituuksia ja leveyksiä. Lataa tiedosto ja laske keskiarvo ja -hajonta a) setosa-lajin sepal_widthille b) versicolorin petal_lengthille. (Mitat ovat senttimetrejä)

6. Normaalijakauma

Normaalijakauma on varmaan useimmille tuttu nimellä "kellokäyrä" tai "Gaussin käyrä". Käyrä näyttää tältä:



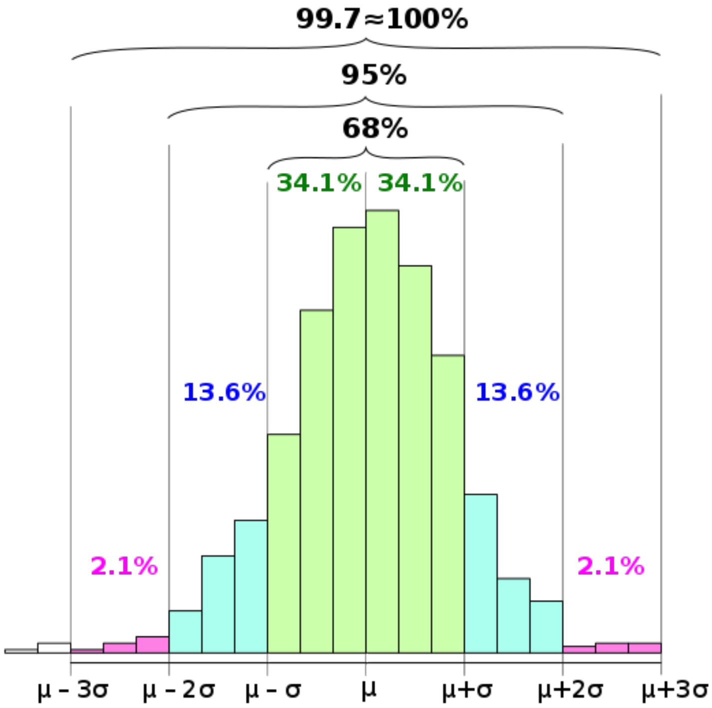
Kuva 6.1. Normaalijakauma. ©Dan Kernler

Suure on normaalijakautunut, jos siitä tehdyt mittaukset asettuvat käyrän mukaisesti: suurin osa mittauksista on lähellä keskiarvoa μ, ja siitä enemmän poikkeavat mittaukset ovat harvinaisempia. Keskihajonta σ (standard deviation) kertoo, kuinka leveä jakauma on. Pienellä keskihajonnalla suurin osa arvoista on lähellä keskiarvoa, kun suurilla keskihajonnoilla ne ovat kauempana siitä.

Tyypillinen esimerkki on ihmisen pituus: suomalaisten aikuisten miesten pituus esimerkiksi on keskimäärin n. 180 cm[6]. Tietoa keskihajonnasta en löytänyt, mutta oletetaan sen olevan vaikkapa 10 cm. Ylläolevasta kuvasta voidaan nyt katsoa, kuinka suuri osa suomalaisista aikuisista miehistä on 170-190 cm pitkiä, eli yhden keskihajonnan päässä keskiarvosta: 68 %. Vastaavasti kahden keskihajonnan sisällä keskipituudesta (160-200 cm) olisi 95% suomalaisista aikuisista miehistä ja kolmen keskihajonnan (150-210 cm) sisällä 99.7%.

Kun mitattava asia tiedetään normaalijakautuneeksi ja tunnetaan sen keskiarvo ja -hajonta, voidaan laskea, kuinka suurella todennäköisyydellä arvo asettuu millekin välille. Tässä toinen kuva, jossa sanotaan suunnilleen sama asia kuin edellisessä:

[6] `http://kasvukayrat.fi/tietoja/tietoa-terveydenhuollon-henkilokunnalle/`



KUVA 6.2. Normaalijakauma. Tässä kuvassa palkkien eroavat korkeudet kuvastavat sitä, että todellisuudessa harva muuttuja on täysin normaalijakautunut. ©Dan Kernler

Esimerkki 6.1. Suomalaisten aikuisten naisten pituuden keskiarvo on 167 cm ja keskihajonta 8 cm.[7] Millä todennäköisyydellä satunnaisesti valittu suomalainen aikuinen nainen on a) 151-183 cm pitkä, b) 159-183 cm pitkä c) 163-171 cm pitkä?

Käytetään apuna kuvaa 6.2 arvoilla $\mu = 167$ cm ja $\sigma = 8$ cm.
a) Koska yksi keskihajonta on 8 cm, on kaksi keskihajontaa 16 cm. 151 cm on siis kaksi keskihajontaa pienempi kuin 167 cm ja 183 cm on kaksi keskihajontaa suurempi kuin 167 cm. Tällä välillä on kuvan mukaan 95% havainnoista, joten kysytty todennäköisyys on 0.95.
b) Nyt välin alaraja on vain yhden keskihajonnan pienempi ja yläraja kaksi keskihajontaa suurempi kuin keskiarvo. Välillä on kuvan mukaan $34.1\% + 34.1\% + 13.6\% = 81.8\%$ aikuisista suomalaisita naisista, joten kysytty todennäköisyys on n. 0.82.
c) Tässä kysytään, millä todennäköisyydellä valittu aikuinen suomalainen nainen on alle puolen keskihajonnan päässä keskiarvosta. Kuva ei anna siihen vastausta.

[7] Keskipituuden lähde sama kuin aiemmin, keskihajonta vedetty hatusta.

Tehtävä 51. Kahvipaketteissa on keskimäärin 500 grammaa kahvia. Laitteiston epätarkkuuden vuoksi paketin massan keskihajonta on 3 grammaa. Millä todennäköisyydellä satunnaisessa kahvipaketissa on 500-506 grammaa kahvia?

Tehtävä 52. Ihmisten älykkyysosamäärän keskiarvo on 100 ja keskihajonta 15.[8]
a) Millä todennäköisyydellä satunnaisesti valitun henkilön ÄO on 85-130?
b) Joskus ÄO:n keskihajonnaksi on valittu 24. Jos henkilön ÄO on 130 keskihajonnalla 15, mikä se olisi keskihajonnalla 24?

Kuvasta katsominen on helppo tapa määrittää todennäköisyyksiä, mutta valitettavan rajoittunut. Siksi käytämme kahta edistyneempää tapaa. Toinen niistä on laskin, ja toinen on MAOLin tauluko. Voit valita, kumpaa käytät, mutta kannattaa opetella käyttämään taulukkoa, jos a) kirjoitat syksyllä matikan, eikä laskimessasi ole tähän soveltuvaa toimintoa, tai b) aiot hakea opiskelemaan alaa, jonka pääsykokeissa käytetään taulukoita. Tällaisia aloja ovat ainakin ennen olleet esimerkiksi psykologia ja lääketiede.

Seuraava esimerkki kertoo kaiken oleellisen normaalijakauman käytöstä.

Esimerkki 6.2. Palataan esimerkin 6.1 kohtaan c. Suomalaisten aikuisten naisten pituus on siis normaalijakautunut, sen keskiarvo on 167 cm ja keskihajonta 8 cm. Millä todennäköisyydellä satunnaisesti valittu suomalainen nainen on 163-171 cm pitkä?

Tehdään tämä kolmella tavalla: TI-Nspirella, LibreOffice Calcilla ja MAOLin taulukoilla.

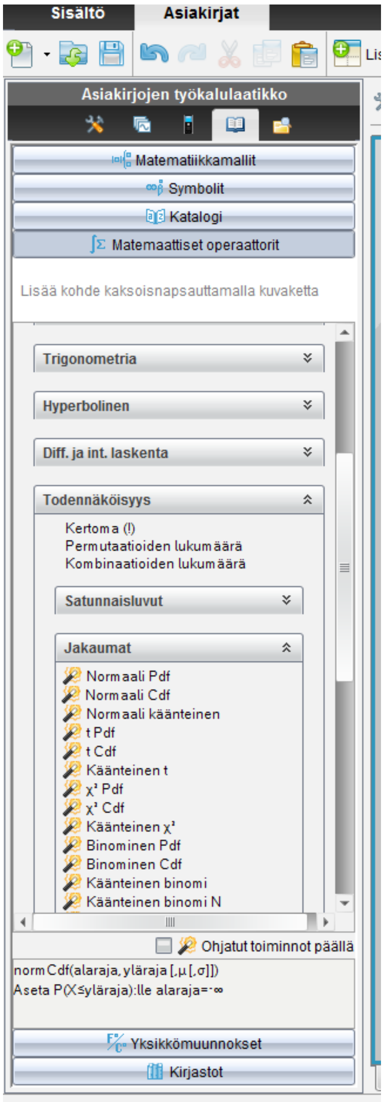
[8]ÄO on periaatteessa määritelty näin.

TI-NSPIREssä todennäköisyys saadaan näin:

`normCdf(163, 171, 167, 8).`

Luvut ovat siis ala- ja yläraja, keskiarvo ja -hajonta. Tätä ei tietenkään kannata opetella ulkoa, mutta kannattaa totutella etsimään se laskimesta (ks. kuva oikealla).

Huomaa myös, että laskin antaa ohjeet sulkuihin tuleville luvuille. Hakasuluissa on valinnaiset arvot, eli funktiota voi käyttää kaksi- tai kolmeparametrisenakin. Jos suluissa on vain kolme lukua, Nspire olettaa, että keskihajonta on 1. Jos lukuja on vain kaksi, Nspire olettaa, että keskiarvo on 0 ja -hajonta 1. Mikään ei tietenkään estä käyttämästä funktiota aina neljällä parametrilla, sellaisissakin tilanteissa, joissa kaksi tai kolme riittäisi. Se voi olla helpompaa, yksinkertaisempaa ja vähentää virheitä.



LIBREOFFICE. Käytetään apuna binomijakaumista tuttua kertymäfunktiota, eli funktiota, joka kertoo annetusta luvusta, millä todennäköisyydellä saadaan sitä pienempi tulos.

Selvitetään aluksi, millä todennäköisyydellä satunnaisesti valittu suomalainen aikuinen nainen on *alle* 171 cm pitkä:

`=NORM.DIST(171, 167, 8, 1).`

Argumentit ovat järjestyksessä: se pituus, jonka alle pitäisi olla, jakauman keskiarvo, jakauman keskihajonta ja ykkönen. Ykkönen tarkoittaa vain, että halutaan nimenomaan kertymäfunktio, samalla tavalla kuin binomijakaumassa.

LibreOffice antaa tulokseksi noin 0.691, joka on siis todennäköisyys valita alle 171-senttinen nainen. Lasketaan seuraavaksi todennäköisyys olla alle 163 senttiä pitkä:

```
=NORM.DIST(163, 167, 8, 1).
```

Tästä LibreOffice antaa vastaukseksi n. 0.309. Nyt todennäköisyys valita satunnainen suomalainen nainen väliltä 163-171 cm on

$$0.691 - 0.309 = 0.382.$$

Tämä lopputulos on altis pyöristysvirheille, joten paremmin sen saa kirjoittamalla suoraan ilman välivaiheita

```
=NORM.DIST(171, 167, 8, 1) - NORM.DIST(163, 167, 8, 1).
```

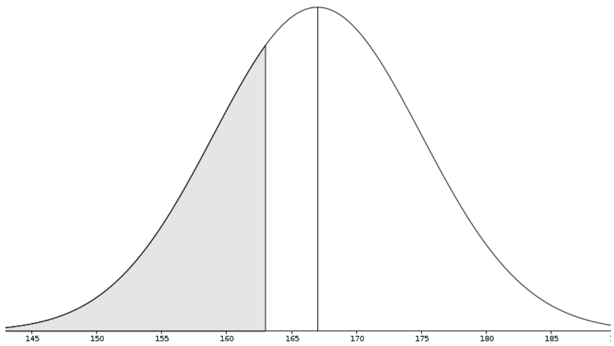
MAOLin taulukkokirjassa on kertymäfunktion arvoja normaalijakaumalle, jonka keskiarvo on 0 ja -hajonta 1. Sen käyttö on oikeasti helppoa, mutta käytön selittäminen tekstin ja kuvien avulla ei. Siksi asiasta kannattaa mieluummin kysyä tunnilla kuin lukea tätä selitystä. Voit myös hypätä tämän kohdan yli, jos et usko käyttäväsi taulukoita tulevaisuudessa.

Normaalijakauman kertymäfunktion taulukko

x	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0,1	5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0,2	5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0,3	6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0,4	6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0,5	6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0,6	7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0,7	7580	7611	7642	7673	7704	7734	7764	7794	7823	7852
0,8	7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0,9	8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1,0	8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1,1	8643	8665	8686	8708	8729	8749	8770	8790	8810	8830
1,2	8849	8869	8888	8907	8925	8944	8962	8980	8997	9015
1,3	9032	9049	9066	9082	9099	9115	9131	9147	9162	9177

Taulukkoa käytetään näin: Haluamme selvittää, millä todennäköisyydellä satunnaisen suomalaisen aikuisen naisen pituus on välillä 163-171 cm, kun pituus on normaalijakautunut, keskiarvo pituudelle on 167 cm ja keskihajonta 8 cm. Pituus 171 cm on puoli keskihajontaa suurempi kuin keskiarvo. Katsotaan taulukon vasemmasta laidasta rivi 0.5. Sarakkeissa on toiset desimaalit. Me etsimme kertymäfunktion arvoa 0.50:lle keskihajonnalle, joten valitsemme sarakkeesta 0 arvon 6915. Tämä luku tarkoittaa, että valitsemamme satunnainen nainen on 0.6915 todennäköisyydellä enintään 171 cm pitkä.

Seuraavaksi haluamme selviittää, millä todennäköisyydellä satunnainen nainen on alle 163 cm pitkä, minkä jälkeen saamme vähentämällä lopullisen todennäköisyyden, samalla tavalla kuin LibreOfficella. Tässä on kuitenkin pieni mutka 163 cm on puoli keskihajontaa *vähemmän* kuin 167 cm, eikä taulukossa ole negatiivisia arvoja. Negatiiviset arvot on jätetty pois, koska ne pystytään laskemaan positiivista. Allaolevassa kuvassa on varjostettu alle 163-senttisiä naisia vastaava alue. Koska normaalijakauma on symmetrinen, sen pinta-ala on yhtä suuri kuin sen peilikuvan pinta-alaa oikealla. Peilikuvan pinta-ala taas saadaan laskemalla $1 - 0.6915 = 0.3085$, missä 0.6915 on jo taulukosta katsomamme arvo. Alle 163-senttisiä naisia on siis 30.85% väestöstä, ja todennäköisyydeksi olla välillä 163-171 cm saadaan $0.6915 - 0.3085 = 0.383$.



Tiivistettynä siis: 1) Taulukosta saaadan todennäköisyyksiä, että tulos on *alle* jonkin tietyn rajan. 2) Jos kyseinen raja on pienempi kuin jakauman keskiarvo, eli se on keskihajonnoissa mitattuna "miinus jotakin", käytetään apuna jakauman symmetrisyyttä ja taulukosta löytyviä arvoja.

Toistaiseksi olemme sanoneet puhekielisesti esimerkiksi, että etsitty arvo on kaksi keskihajontaa suurempi tai puoli keskihajontaa pienempi kuin keskiarvo. Sillä tavalla asiaa on mielestäni helpointa ajatella. Taulukoidun arvon laskemiselle on kaavakin:

$$z = \frac{x - \mu}{\sigma}.$$

Tässä x on alkuperäinen arvo, z taulukoitu, *normitettu* arvo, μ jakauman keskiarvo ja σ sen keskihajonta. Arvon z laskemista sanotaan sen *normittamiseksi*. Jos haluaisimme tietää esimerkiksi, kuinka suuri osa naisista on alle 181-senttisiä, normittaisimme arvon 181 näin:

$$z = \frac{181 - 167}{8} = \frac{14}{8} = 1.75.$$

Sitten etsisimme taulukosta vastaavan arvon riviltä 1.7 ja sarakkeesta 5, jolloin saamme todennäköisyydeksi 0.9599. Tämän kaavan käyttäminen on juuri sitä, mitä olemme tehneet aiemmin epämuodollisesti: ensiksi lasketaan, kuinka kaukana ollaan keskiarvosta (14 cm), ja sitten muutetaan mittayksiköt senteistiä keskihajonnoiksi (1.75 keskihajontaa).

Tehtävä 53. Kahvipaketin massa on normaalijakautunut. Sen keskiarvo on 500 grammaa ja keskihajonta 3 grammaa.
a) Millä todennäköisyydellä satunnaisessa kahvipaketissa on alle 495 g kahvia?
b) Millä todennäköisyydellä satunnaisessa kahvipaketissa on yli 510 g kahvia?
c) Kuluttajan vaaka näyttää arvoa 500 g, jos paketin todellinen massa on 499.5-500.5 grammaa. Millä todennäköisyydellä näin käy?

Tehtävä 54. Kurjenmiekan terälehden pituus on normaalijakautunut. Sen keskiarvo on 3.8 cm ja keskihajonta 0.5 cm.
a) Millä todennäköisyydellä satunnaisen terälehden pituus on alle 2 cm?
b) Millä todennäköisyydellä satunnaisen terälehden pituus on 4-5 cm?

Tehtävä 55. Kurjenmiekan terälehden pituus on normaalijakautunut. Sen keskiarvo on 3.8 cm ja keskihajonta 0.5 cm.
a) Anni kertää kymmenen satunnaista kurjenmiekan terälehteä. Millä todennäköisyydellä ne ovat kaikki alle 4.5 cm pitkiä?
b) Benjamin kerää kaksikymmentä satunnaista terälehteä. Millä todennäköisyydellä hän saa ainakin yhden kuusisenttisen?

Merkintöjä. Normaalijakaumaa merkitään näin: $N(\mu, \sigma^2)$. Esimerkiksi $N(154, 4^2)$ tarkoittaa normaalijakaumaa, jonka keskiarvo on 154 ja keskihajonta 4.

Keskihajonnan toista potenssia nimitetään varianssiksi, ja vanhastaan on ollut tapana kertoa jakauman keskiarvo ja varianssi. Jotkut ilmoittavat mieluummin keskihajonnan, joka on omiaan aiheuttamaan sekaannusksia, periaatteessa merkinnästä $N(154, 16)$ ei tiedä, onko varianssi vai keskihajonta 16. Toisen potenssin kirjoittaminen näkyviin on hyvä tapa ilmoittaa, että tämä tarkoittaa nyt varianssia. Jos esimerkiksi ylioppilaskokeessa käytetään monitulkintaista merkintää, se on kokeen tekijöiden vika.

6.1. Normaalijakauma toisin päin.

Esimerkki 6.3. Tutkija tietää, että suomalaisten pituus noudattaa normaalijakaumaa keskiarvona 180 cm ja keskihajontana 10 cm. Hän haluaa koota snobistisen klubin, jonka jäsenet kuuluvat suomalaisten pisimpään prosenttiin. Kuinka pitkä pitää olla päästäkseen hänen klubiinsa?

LibreOffice Calc: Tutkija kirjoittaa ruutuun `=NORM.INV(0.99, 180, 10)`, ja LibreOffice antaa vastaukseksi 203.26 cm. TI-Nspirellä tutkija kirjoittaisi laskimeen `invnorm(0.99, 180, 10)`. Luvut molemmissa tapauksissa ovat 1) todennäköisyys, jota vastaavaa rajaa etsitään, 2) keskiarvo ja 3) keskihajonta.

Tehtävä 56. Maitotölkin massa on normaalijakautunut keskiarvolla 1000 grammaa ja keskihajonnalla 4 grammaa.
a) Kuinka painava tölkin on oltava ollakseen painavimmassa promillessa kaikista tölkeistä?
b) Millä todennäköisyydellä tällainen tölkki on tuhannen satunnaisen tölkin joukossa?
c) Minkä painon tölkittäjä voi ilmoittaa vähimmäispainoksi, jos halutaan, että alle prosentti tölkeistä on sitä kevyempiä?

Tehtävä 57. Ovien valmistaja haluaa, että alle miljoonasosa ihmisistä joutuu kumartumaan kulkiessaan heidän ovistaan. Jos ihmisten keskipituus on 175 cm ja keskihajonta 8.5 cm, kuinka korkeita ovien pitää olla?

6.2. **Satunnaismuuttujista ja todennäköisyysjakaumista yleisesti.** Tästä ei kysytä kokeessa, mutta kannattaa lukaista läpi.

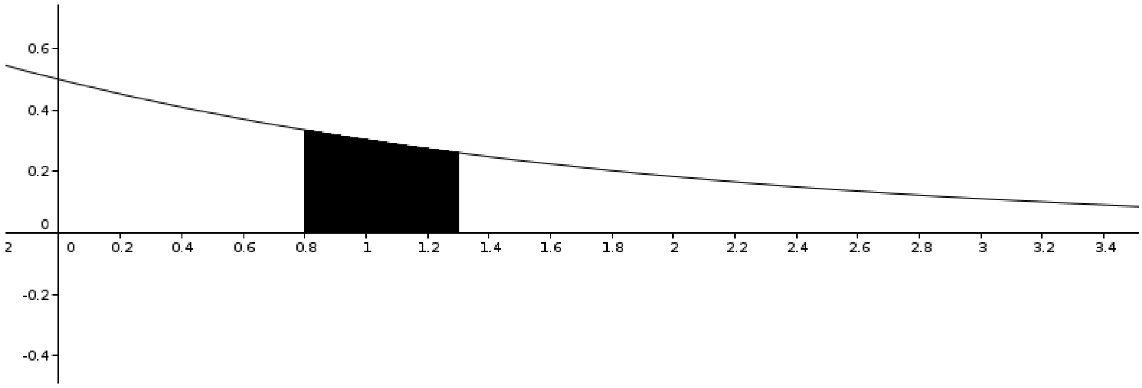
Satunnaismuuttujiksi sanotaan lukuja, jotka jollakin tavalla riippuvat jostakin satunnaisesta. Esimerkiksi heitettäessä kolikkoa sata kertaa, on kruunien lukumäärä satunnaismuuttuja. Valittaessa satunnainen ihminen, on hänen pituutensa satunnaismuuttuja. Valittaessa satunnainen otos ihmisiä on heidän keskipituutensa satunnaismuuttuja. Suomalaisten keskipituus ei ole satunnaismuuttuja, se on jokin tietty luku. Myöskään esimerkiksi Sauli Niinistön pituus ei ole satunnaismuuttuja.

Jos tarkempia ollaan, satunnaismuuttujan ei tarvitse olla luku, se voisi hyvin olla vaikkapa nimi, kansallisuus, päällä olevan paidan väri tms.

Jakauma on jonkinlainen sääntö satunnaismuuttujan ja sen arvojen todennäköisyyden välillä. Esimerkiksi kruunien määrä kolikonheitossa on noudattaa binomijakaumaa. Satunnaisesti valitun ihmisen pituus normaalijakaumaa. Binomijakauma on *diskreetti*, koska se voi saada arvoja vain tietystä äärellisestiä joukosta. Normaalijakauma on *jatkuva*, eli siinä satunnaismuuttujan arvo voi olla mikä tahansa reaaliluku.

Diskreettien jakaumien todennäköisyyksiä voidaan havainnollistaa parhaiten pylväsdiagrammeilla, mutta jatkuvissa jakaumissa parempi on käyrä, jota sanotaan jakauman *tiheysfunktioksi*. Käyrän pitäisi olla ei-negatiivinen ja sen alle jäävän pinta-alan pitäisi olla 1. Nämä asiat pitävät paikkansa normaalijakaumalle. Kun halutaan tietää, millä todennäköisyydellä henkilö on vaikkapa 184-187 cm pitkä, lasketaan käyrän alle näiden x-akselin pisteiden väliin jäävä pinta-ala. Muut jatkuvat jakaumat toimivat samalla periaatteella.

Esimerkki 6.4. *Eksponenttijakauma* kuvaa esineiden kestoikää. Sen tiheysfunktio on $f(x) = \lambda e^{-\lambda x}$, missä λ on esinekohtainen vakio, ja kestoikä x on postitiivinen. Mitä pienempi λ on, sitä pidempään esineen voi olettaa kestävän.



Tässä kuvassa on eksponenttijakauma λ:n arvolla 0.5. Käyrä jatkuu nollan vasemmalle puolelle vain laiskuuteni takia, oletan ymmärretyksi, että esineen kestoikä ei voi olla negatiivinen. Käyrän ja positiivisten akselien väliin jäävä pinta-ala on 1. Sovitaan, että x-akselin yksiköt olisivat vaikka vuosia. Tummennettun alueen pinta-ala on 0.15, ja se on todennäköisyys sille, että esine hajoaa 0.8-1.3 vuoden käytön jälkeen.

Käyrän alle jäävän pinta-alan voi laskea monella tavalla. Tässä esimerkikssä laskin sen GeoGebran komennolla `LowerSum(f, 0.8, 1.3, 10 000)`, suomenkielisessä GeoGebrassa `Alasumma` tjsp. Tämä komento piirtää käyrän f alle 10 000 suorakulmiota ja laskee yhteen niiden pinta-alan.

Toinen tapa on integraalilaskenta, jota opetetaan pitkässä matematiikassa. Integroimalla pinta-alalle on helppo laskea käsin tarkka arvo (toisin kuin normaalijakaumalle). Myös TI-Nspire osaa laskea pinta-alan komennolla

$$\int_{0.8}^{1.3} 0.5 \cdot e^{-0.5x}\, dx.$$

Tämä komento löytyy otsikon “matematiikkamallit” alta. Tämä merkintä voi näyttää pelottavalta, mutta se tarkoittaa yksinkertaisesti vain käyrän alle jäävää pinta-alaa sille funktiolle, joka on kirjoitettu merkkien $\int$ ja dx väliin. Todennäköisyyslaskennassa on parempi laskea likiarvo painamalla enteriä ctrl pohjassa. Likiarvo kertoo tässä tapauksessa jotain todennäköisyyden suuruusluokasta, tarkka arvo ei juurikaan.

Eräs tärkeä asia jatkuvissa jakaumissa on se, että ei ole järkeä sanoa, että henkilön pituus on tasan 189 cm tai esineen käyttöikä tasan puoli vuotta. Näiden tapahtumien todennäköisyys on tasan 0. Jatkuvilla jakaumilla pitää aina laskea todennäköisyys jollekin välille. Esimerkiksi

189 cm voisi tarkoittaa väliä 188.5-189.5 cm, ja puoli vuotta voisi tarkoittaa 182-183 päivää.

Laskinhommista pitää muistaa, että TI-Nspiren Pdf-loppuiset komennot tarkoittavat todennäköisyyttä juuri mainitulle arvolle (esim `binomPdf`) ja Cdf-loppuiset komennot arvolle ja sitä pienemmille arvoille (esim `binomPdf` ja `normalPdf`). LibreOfficessa tämä ero tehtiin jakaumakomennon viimeisellä parametrilla, joka on 0 tai 1. Jatkuvalle jakaumalle Pdf-loppuiset komennot ovat hyödyttömiä syystä, joka kerrottiin äsken: ei ole järkevää kysyä, millä todennäköisyydellä joku on tasan 189 cm pitkä.

7. Tilastotiedettä

Ajatellaan, että haluaisimme selvittää, mikä on Sauli Niinistön kannatus presidentiksi tällä hetkellä. Kannatus on (ainakin periaatteessa) jokin tietty prosenttiluku, joka kertoo, kuinka suuri osa äänestäjistä äänestäisi häntä, jos vaalit järjestettäisiin nyt. Tätä lukua ei todellisuudessa voi selvittää millään muulla tavalla kuin järjestämällä vaalit. Gallupit ovat yrityksiä arvioida tätä lukua, mutta ne ovat puutteellisia.

Gallupin puutteellisuus johtuu siitä, että kaikkia äänestäjiä ei voida haastatella, ja vaikka voitaisiinkin, ei totuudenmukaisista vastauksista olisi mitään takeita. Äänioikeutettuja vuoden 2018 presidentinvaaleissa oli n. 4,5 miljoonaa, ja vaikka heistä haastateltaisiin miljoonaa, ei voida varmasti sanoa, etteikö kaikki loput 3,5 miljoonaa äänestäisi esimerkiksi Nils Torvaldsia. Se on toki hyvin epätodennäköistä, mutta ei mahdotonta.

Varmaa tietoa ei siis voida saada, eikä sitä ole järkevää edes yrittää saada. Järkevämpää on etsiä uskottavinta tietoa ja arvioida sen paikkansapitävyyttä.

Vaaligallup on harvinainen esimerkki tilanteesta, jossa todellinen arvo myös selviää (tosin eri ajanhetkellä). Paljon yleisempiä ovat tutkimukset, joissa ei ole *mitään* muuta keinoa selvittää todellista arvoa kuin tilastollinen tutkimus, ja todellisen arvon olemassaolokin voi olla enemmän metafyysistä spekulaatiota. Varmuus on harvinainen ilmiö todellisessa maailmassa.

Lyhyesti gallupin tekeminen menee näin:

(1) Valitaan joukko henkilöitä, joilta kysytään heidän äänestämisestään (otos).

(2) Niinistön kannatuksen ennustetaan olevan yhtä suuri kuin hänen kannatus on otoksessa. Tätä sanotaan *uskottavimmaksi* arvoksi.

(3) Gallupin virhettä arvioidaan vastausten keskihajonnan avulla. Tarkemmin, siitä ja otoksen koosta lasketaan luottamusväli (esimerkiksi 58% - 68%), jolla Niinistön kannatus 95% todennäköisyydellä on. Luottamusvälejä voidaan laskea toki muillekin todennäköisyyksille kuin 95%.

Katsotaan seuraavaksi tarkemmin jokaista näistä kohdista.

7.1. **Populaatio ja otos.**

Populaatio ja otos
Populaatio on koko tutkittava joukko. Otos on jokin sitä pienempi joukko, jonka jäsenet on valittu satunnaisesti.

Esimerkki 7.1. a) Suomen äänoikeutetut on populaatio äänestystutkimuksessa. Väestörekisterin tietokoneen arpomat sata koehenkilöä on otos. Helsingin kuvataidelukion opettajakunta ei ole otos, koska se ei ole satunnaisesti valittu. Iltalehden lukijakunta ei ole otos, koska se ei ole satunnaisesti valittu.
b) EU:n kansalaiset on populaatio tutkittaessa EU:n kansalaisten keskipituutta. Suomalaiset eivät ole otos, koska suomalaiset eivät ole satunnaisesti valittu joukko.
c) Suomen lukiolaiset on populaatio tutkittaessa Suomen lukiolaisten varallisuutta. Kallion lukion opiskelijat eivät ole otos, koska eivät ole satunnaisesti valittu.
d) Tutkittaessa putoamiskiihtyvyyttä, voidaan mittaussarjaa ajatella otokseksi kaikien mahdollisten mittausten populaatiosta. Tämä on tietenkin täysin kuvitteellinen populaatio, mutta kuvitelma on yleensä mittaustuloksia käsiteltäessä käyttökelpoinen.

Populaatiota pienempää joukkoa, joka ei ole satunnaisesti valittu, sanotaan *näytteeksi*. Pedantti tilastotieteilijä voisi valittaa ylläolevasta otoksen määritelmästä, mutta se on MAB8-kurssille aivan riittävä.

Tehtävä 58. Kuuluuko Antti Tuisku populaatioon, kun tutkitaan
a) puoluekannatusta eduskuntavaaleissa 2019
b) ruotsalaisten keskipituutta
c) Euroopan nisäkkäiden elinikää
d) tilastotieteilijöiden työllisyysastetta.

7.2. **Uskottavin arvo.** Presidentinvaaligalluppia tehtäessä tutkimuksen lopputulos on ennuste, että Niinistön kannatus koko populaatiossa on sama kuin otoksessa. Jos tutkittaisiin suomalaisten keskipituutta, lopputulos olisi, että keskipituus on sama kuin keskipituus otoksessa. Näitä sanotaan *uskottavimmiksi* arvoiksi.

Tehtävä 59. Osoitteessa `https://www.cs.helsinki.fi/u/mokangas/mab8/vpjapituus.csv` on tiedosto, jossa on yläverenpaine ja pituus pieneltä otokselta aikuisia suomalaisia.
a) Määritä tiedostosta uskottavin arvo suomalaisten aikuisten yläverenpaineelle ja pituudelle.
b) Miksi nämä eivät ole uskottavimpia arvoja kaikkien suomalaisten yläverenpaineelle ja pituudelle?

Tehtävä 60. Osoitteessa `https://www.cs.helsinki.fi/u/mokangas/mab8/gallup.csv` on satunnaisesti generoitu galluptutkimus presidentinvaaleihin 2019.[9] Listassa 0 tarkoittaa, että vastaaja ei äänestäisi, ja muulloin numero on sen ehdokkaan numero, jota hän äänestäisi.
a) Laske uskottavin arvo äänestysprosentille. Nollat on helppo laskea näin: `=COUNTIF(A1:A45, ‘‘=0’’)`
b) Laske uskottavin arvo Niinistön (8) kannatukselle. Huomaa, että kannatus lasketaan prosenttina äänestäneistä, ei äänioikeutetuista.
c) Laske uskottavin arvo Haaviston (3) kannatukselle.

7.3. **Otoskeskihajonta.** Ajatellaan taas tapausta, jossa tutkitaan populaatiota otoksen avulla. Olkoon tutkittava asia vaikka suomalaisten aikuisten keskipituus, populaatio kaikki suomalaiset aikuiset ja otos satunnaisesti valittu sadan henkilön joukko. Aiemmin on jo kerrottu, että paras arvio koko populaation keskipituudelle on otoksen keskipituus.

Kun halutaan arvioida koko populaation keskihajontaa, ei kuitenkaan otoksen keskihajonta ole paras arvio. Pisimmät ja lyhimmät eivät välttämättä osu otokseen. Mitä pienempi otos, sitä epätodennäköisemmin siinä on mukana ääripäitä. Siksi otoksessa on yleensä vähemmän hajontaa kuin koko populaatiossa. **Paras arvio koko populaation keskihajonnalle on otoskeskihajonta:**

OTOSKESKIHAJONTA

Otoskeskihajonta on

$$\sigma_{n-1} = \sqrt{\frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n-1}},$$

missä $x_1, x_2, \ldots, x_n$ ovat havainnot otoksessa ja $\bar{x}$ niiden keskiarvo.
LibreOfficella (esimerkiksi) `=STDEV(A1:A13)` .
TI-Nspiressä `=stdevsamp(A1:A13)`.

HUOMAA:

- Otoskeskihajonta eroaa keskihajonnasta vain siinä, että jaetaan $n-1$:llä eikä n:llä, ja sen käyttötarkoitus on saada otoksesta ennuste koko populaation keskihajonnalle.
- Sinun ei tarvitse osata keskihajonnan tai otoskeskihajonnan kaavaa, eikä todennäköisesti edes käyttää niitä koskaan. Laske ne suosiolla koneella.
- Otoskeskihajonta on aina suurempi kuin keskihajonta.

[9] Vastaukset on arvottu vaalien todellisen äänten perusteella koneellisesti.

- Mitä suurempi otos, sitä pienempi on otoskeskihajonnan ja keskihajonnan ero. Otoskoolla 10 ero on n. 5%, otoskoolla 100 n. 0.5% .

Hyvin monessa ohjelmistossa tai laskimessa on kerrottu epäselvästi, antaako toiminto keskihajonnan vai otoskeskihajonnan. Ne saatetaan erottaa toisistaan esimerkiksi merkinnöillä σ_n ja σ_{n-1}. LibreOfficessa otoskeskihajonnan nimi on pelkkä `STDEV`, kun keskihajonnassa täsmennetään `STDEVPA`. Tämä johtuu siitä, että tarve laskea otoskeskihajonta on paljon yleisempi kuin tarve laskea koko populaation keskihajonta.

Esimerkki 7.2. Tutkija kerää kurjenmiekan terälehtiä ja mittaa niiden pituudet:

$$1.4, 1.4, 1.3, 1.5, 1.4, 1.7 \text{ cm}.$$

Hän laittaa luvut LibreOffice Calciin ja laskee niiden keskiarvoksi 1.45 cm, keskihajonnaksi 0.1258 cm ja otoskeskihajonnaksi 0.1378 cm.

Hän päättelee, että kurjenmiekkojen terälehdet yleisesti ovat 1.45 cm pitkiä, ja niiden pituuden keskihajonta on 0.1378 cm.

Luku 0.1258 cm on käytännössä hyödytön. Se kertoo vain, mikä on näiden kuuden nimenomaisen terälehden pituuksien keskihajonta. Todellisuudessa tutkija ei edes laskisi sitä.

Tehtävä 61. Tutkija yrittää selvittää norjalaisten aikuisten keskipituutta. Hän valitsee satunnaisen otoksen ja saa mitat

$$179, 174, 171, 169, 172, 148 \text{ ja } 171 \text{ cm}.$$

Mikä on hänen ennusteensa norjalaisten aikuisten pituudelle ja sen keskihajonnalle?

Tehtävä 62. Osoitteessa `https://www.cs.helsinki.fi/u/mokangas/mab8/iris.csv` on Kurjenmiekkojen mittoja. Laske uskottavin arvo versicolor-lajikkeen terälehden leveydelle (petal_width) ja sen otoskeskihajonta.

7.4. **Luottamusväli suurella otoksella.** Tutkimuksissa pelkkä uskottavin arvo on usein melko hyödytön, jos mukana ei ole jonkinlaista virheanalyysia.

Luonnontieteiden artikkeleissa voitaisiin ilmoittaa esimerkiksi putoamiskiihtyvyyden olevan mittausten perusteella 9.7 ± 0.5 m/s^2. Tämä tarkoittaisi, että tutkimuksessa otoskeskihajonta tulokselle on 0.5 m/s^2. Annetut rajat eivät siis ole suurimpia ja pienimpiä mahdollisia arvoja, eikä lopputulos kovin suurella todennäköisyydellä edes ole näissä rajoissa. Näiden artikkeleiden lukijat yleensä osaavat itse laskea tarpeidensa mukaan sopivan välin, jolla arvo milläkin todennäköisyydellä sijaitsee.

Vaaligallupeissa ja populaariteksteissä on usein tapana ilmoittaa 95%:n *luottamusväli*. Todellisen arvon voi ajatella olevan 95% todennäköisyydellä tällä välillä.[10] Galluppeja uutisoitaessa sanotaan usein esimerkiksi "todellinen kannatus voi vaihdella kaksi prosenttiyksikköä suuntaansa", joka on virheellinen väite. Todellinen kannatus *voi* olla melkein mitä tahansa. Viidessä prosentissa kaikista gallupeista todellinen kannatus ei ole annettujen rajojen sisällä.

Katsotaan ensin, miten luottamusväli lasketaan käytännössä mekaanisesti, ja sitten vasta, mistä se tulee.

Resepti 95% luottamusvälin laskemiseen:

(1) Tutkit jotakin asiaa koko populaatiossa, ja sitä varten valitset otoksen, jossa on n satunnaista populaation jäsentä.

(2) Laske otoksesta keskiarvo $\bar{x}$ ja otoskeskihajonta s (merkitään otoskeskihajontaa näin, että se olisi yhteneväinen MAOLin taulukoiden kanssa).

(3) Luottamusvälin keskipiste on $\bar{x}$. Sen pituus molempiin suuntiin on $1.96 \cdot \frac{s}{\sqrt{n}}$.

Esimerkki 7.3. Tutkitaan sadan henkilön otoksella suomalaisten aikuisten keskipituutta. Otoksessa keskipituus on 168.70 cm ja pituuden keskihajonta 9.5 cm.

95%:n luottamusväliä varten lasketaan ensin

$$1.96 \cdot \frac{s}{\sqrt{n}} = 1.96 \cdot \frac{9.5}{\sqrt{100}} = 1.862.$$

Luottamusvälin alaraja siis on $168.70 - 1.862 = 166.838 \approx 166.8$ cm ja yläraja $168.70 + 1.862 = 170.562 \approx 170.6$ cm.

Tutkimuksessa saatu 95%:n luottamusväli on siis $[166.8, 170.6]$.

Tehtävä 63. Suomalaisten aikuisten yläverenpainetta tutkitaan 50 henkilön otoksella, jossa keskiarvo yläpaineelle on 143.84 mmHg ja keskihajonta 25.77 mmHg. Mikä on 95%:n luottamusväli tutkimuksen mukaan?

Kun halutaan suurempia tai pienempiä luottamusvälejä, taikaluku 1.96 korvataan asiaankuuluvalla luvulla. Esimerkiksi 99%:n luottamusvälissä se on 2.58. Muuten laskut ovat täysin samat.

Tehtävä 64. Etsi MAOLin taulukoista luottamusvälit ja laske edelliseen tehtävään 99.9%:n luottamusväli.

[10] Pedantti tilastotieteilijä nuhtelisi tästä sanamuodosta. Yksittäisessä gallupissa todellinen kannatus joko on tai ei ole ilmoitetulla välillä. Kaikissa gallupeista n. 95:ssä prosentissa todellinen kannatus on ilmoitetulla 95%:n luottamusvälillä.

Tehtävä 65. Osoitteessa `https://www.cs.helsinki.fi/u/mokangas/mab8/pituuksia.csv` on otos amerikkalaisten aikuisten pituuksia. Laske sen perusteella
a) 95 %:n luottamusväli amerikkalaisten aikuisten keskipituudelle
b) 99%:n luottamusväli amerikkalaisten aikuisten keskipituudelle.

KESKIHAJONTA KOLIKONHEITTOKOKEESSA

Jos kolikkoa on heitetty n kertaa ja saatu k kruunaa, merkitään $p = \frac{k}{n}$ ja käytetään luottamusvälin laskemisessa otoskeskihajonnan paikalla lukua

$$s = \sqrt{p(1-p)}.$$

Luottamusväli lasketaan kruunien suhteelliselle osuudelle p.

Kolikonheiton keskihajonta saadaan siitä, että suurelle määrälle n toistoja kolikonheitto muistuttaa normaalijakaumaa keskiarvolla np ja keskihajonnalla $np(1-p)$, missä p on kruunan todennäköisyys. Tarkempi perustelu edellyttäisi taas vähän vaativampaa matematiikkaa, joten ei ryhdytä siihen nyt.

Huomaa, että tämä muotoilu poikkeaa vähän MAOLin taulukon muotoilusta (ainakin käytettävissäni olevassa versiossa): me laskemme ensin luvun s ja sitten sovellamme aiempaa kaavaa, jossa se jaetaan $\sqrt{n}$:llä, kerrotaan 1.96:lla jne. MAOLin taulukon kaavassa $\sqrt{n}$:llä jakaminen on sisällytetty keskihajonnan kaavaan. Meidän menetelmällä luottamusväleille ei tarvitse useampia eri kaavoja. (Tämä kannattaa katsoa MAOLin taulukosta läpi, niin tulet toimeen asian kanssa kokeissa).

Esimerkki 7.4. Vaaligallupissa haastateltiin sataa satunnaista äänioikeutettua, joista 57 sanoi, että äänestäisi Niinistöä, jos vaalit järjestettäisiin nyt. Mikä on 95%:n luottamusväli Niinistön kannatukselle?

Tämä on oleellisesti sama asia kuin kolikonheitto, joten $p = \frac{57}{100} = 0.57$ ja

$$s = \sqrt{p(1-p)} = \sqrt{0.57 \cdot 0.43} = \sqrt{0.2451}.$$

Ei lasketa vielä s:lle likiarvoa, ettei lopputulokseen tule liikaa pyöristysvirhettä. Luottamusvälin alarajaksi saadaan

$$p - 1.96 \cdot \frac{s}{\sqrt{n}} = p - 1.96 \cdot \frac{\sqrt{0.2451}}{\sqrt{100}} \approx 0.472$$

ja ylärajaksi

$$p + 1.96 \cdot \frac{s}{\sqrt{n}} = p + 1.96 \cdot \frac{\sqrt{0.2451}}{\sqrt{100}} \approx 0.667.$$

95%:n luottamusväli Niinistön kannatukselle on siis $[47.2\%, 66.7\%]$.

Järkevä laskija huomaa, että ala- ja ylärajojen kaavat eroavat toisistaan vain miinuksella ja plussalla, ja käyttää hyväkseen laskimen muistitoimintoja tai copy-pastea. Huomaa myös, että ala- ja ylärajat on tässä pyöristetty poispäin 0.57:stä, koska pyöristysvirheen ei haluta tekevän luottamusvälistä liian optimistista.

Tehtävä 66. Vaaligallupissa haastateltiin tuhatta satunnaisesti valittua äänoikeutettua. Heistä 712 sanoi aikovansa äänestää. Mikä on gallupin perusteella 95% luottamusväli vaalien äänestysprosentille?

Tehtävä 67. Osoitteessa `https://www.cs.helsinki.fi/u/mokangas/mab8/gallup2.csv` on koneellisesti arvottu vaaligallup presidentinvaaleista 2018. Luku 0 tarkoittaa hylättyä ääntä ja muut numerot ehdokkaiden numeroita. Laske 95%:n luottamusväli Haataisen (6) ja Vanhasen (4) kannatuksille. Oliko heidän todellinen kannatus luottamusvälillä? Käytä apuna sivua `https://fi.wikipedia.org/wiki/Suomen_presidentinvaali_2018` Laske hylätty ääni mukaan kokonaisäänimäärään.

Tehtävä 68. Osoitteessa `https://www.vauva.fi/keskustelu/3080040/eduskuntavaali-2019-gallup` kysytään, mitä puoluetta korkeatasoisesta keskustelusta tunnetun Vauva-lehden lukijat aikovat äänestää eduskuntavaaleissa 2019. Klikkaa “näytä vastaukset” vaihtoehtojen alapuolella nähdäksesi tulokset.
a) Mikä puolue olisi kyselyn perusteella vaalien jälkeen eduskunnan suurin?
b) Mikä on 95%:n luottamusväli puolueen kannatukselle?
c) Onko a- ja b-kohdan ennusteet päteviä? Miksi/miksi ei?

7.5. **Lisätietoa, vaikea ja vapaaehtoinen.** “Uskottavin arvo” tarkoittaa *sitä arvoa, joka tekee kerätyn aineiston todennäköisimmäksi.* Jos gallupissa 62% vastaajista sanoo äänestävänsä Niinistöä, tällaisen aineiston saaminen on todennäköisempää 62%:n kuin 61%:n tai 63%:n todellisella kannatuksella. Tämä on eri asia kuin sanoa, että 62%:n kannatus on todennäköisin. Asian tarkempi perustelu vaatisi enemmän matematiikkaa kuin tässä on tarkoituksenmukaista käydä läpi, joten tämä kannattaa ottaa vain uskon asiana. Olisi väärin sanoa, että 62%:n kannatus on todennäköisin, koska se yksinkertaisesti joko on todellinen kannatus, tai se ei ole, eikä galluppia tehtäessä ole mitään keinoa selvittää, kumpaa se on.

Mistä luottamusvälit tulevat? Ajatellaan, että tutkitaan jotain normaalijakautunutta suuretta, vaikka suomalaisten aikuisten pituutta. Kun arvio koko populaation keskipituudelle lasketaan otoksesta, on arvio satunnainen, se riippuu otoksen satunnaisesta valinnasta. Mitä suurempi otos, sitä lähempänä arvion voi olettaa olevan todellista arvoa. Jos koko populaatiossa keskipituus on μ ja pituuden keskihajonta σ, voidaan kohtalaisella yhtälönpyörittelyllä näyttää, että otoksesta laskettu keskiarvo on (melkein) normaalijakautunut keskiarvolla μ ja keskihajonnalla $\frac{\sigma}{\sqrt{n}}$.

Kun halutaan selvittää vaikkapa 95%:n luottamusväliä, halutaan että 47.5% on keskivälin kummallakin puolella. Luottamusvälin alaraja on siis se, jonka alapuolella on 2.5% kaikista otoksista lasketuista keskiarvoista ja yläraja se, jonka alapuolella on 97.5% niistä. Nämä sijoittuvat 1.96 keskihajonnan päähän keskiarvosta.

Yleiset luottamusvälit, esimerkiksi 80%:n tai 99.99%:n luottamusvälit saadaan nyt tällä reseptillä:

(1) Puolita luottamusväli ja laske se yhteen 50%:n kanssa. Sanotaan tätä luvuksi x. Esimerkiksi 80% luottamusvälitstä saat luvun $x = \frac{80\%}{2} + 50\% = 90\% = 0.9$.

(2) Korvaa 95%:n luottamusvälin kaavassa taikaluku 1.96 luvulla, jonka saat Nspirestä normaalijakauman käänteisfunktiolla näin: `invnorm(x,0,1)` tai LibreOffice Calcilla näin `=NORM.INV(x,0,1)`.

Pienillä otoksilla luottamusvälin laskeminen poikkeaa siten, että taikaluku 1.96 (95%:n luottamusvälillä) korvataan t-jakauman vastaavalla luvulla. Esimerkiksi 95%:n luottamusvälissä luku 1.96 vastaa normaalijakauman käänteisfunktion arvoa `invnorm(0.975, 0,1)`. Jos otoskoko olisi vaikka 8, korvattaisiin luku 1.96 sillä luvulla, jonka laskin antaa, kun siihen kirjoittaa `invt(0.975,7)`. Tässä luku 7 on ns vapausaste, ja se on otoksen koko miinus yksi. Käyttöohjeissa vapausaste lyhennetään usein esim df, eli degrees of freedom.

Suuren ja pienen otoksen raja riippuu halutusta tarkkuudesta. Kohtuullisen yleinen raja suuren ja pienen jakauman erolle on 30. Huomaa, että tämä raja ei siis kerro, onko gallupin otos riittävä järkevien tulosten saamiseen, vaan ainoastaan sen, pitäisikö sen luottamusväli laskea t-jakauman vai normaalijakauman mukaan. Periaatteessa luottamusväli pitäisi laskea aina t-jakaumasta, mutta suurilla otoksilla se muistuttaa normaalijakaumaa.

8. Lisätehtäviä

Kaikki nämä ovat vanhoja matematiikan yo-tehtäviä. Tehtävät eivät ole missään erityisessä järjestyksessä.

Tehtävä 69. (Syksy 17, lyhyt)

Pienestä lukiosta valmistui 22 ylioppilasta vuonna 2007. Kymmenen vuoden kuluttua valmistumisesta kaksi heistä päättää järjestää luokkakokouksen ja valitsee itselleen sopivan päivämäärän. Oletetaan, että jokaiselle muulle luokkakaverille tämä päivä sopii kuitenkin vain 85 % todennäköisyydellä.

a) Kuinka suurella todennäköisyydellä kaikki pääsevät paikalle?

b) Kuinka suurella todennäköisyydellä täsmälleen yksi ei pääse paikalle?

Tehtävä 70. (Kevät 16, lyhyt)

Hajamielinen professori muistaa ystäviensä ovikoodista vain, että se koostuu neljästä erisuuresta parittomasta numerosta.

a) Kuinka monta koodia hän joutuu huonoimmassa tapauksessa (enintään) kokeilemaan, jos hän käy systemaattisesti läpi kaikki vaihtoehdot?

b) Parin vuoden käyntien jälkeen professori huomaa koodissa seuraavan ominaisuuden: siinä ei ole numeroa 9 eikä peräkkäin "vierekkäisiä" parittomien numeroiden (1 ja 3, 3 ja 1, 3 ja 5, 5 ja 3, 5 ja 7, 7 ja 5, 7 ja 9 tai 9 ja 7) yhdistelmiä. Kuinka monta koodia pitää huonoimmassa tapauksessa kokeilla, kun otetaan huomioon myös nämä lisätiedot?

Tehtävä 71. (Kevät 16, lyhyt)

Uuteen 20-kerroksiseen tornitaloon asennettiin kolme hissiä. Todennäköisyys, että hissi tilataan johonkin kerroksista 2–20, on 0,025 kullekin. Todennäköisyys, että hissi tilataan kerrokseen 1, on 0,4 ja kellarikerroksessa sijaitsevaan parkkihalliin 0,125. Ruuhkattomina aikoina hissit palaavat seuraavanlaisille odotuspaikoilleen: yksi hissi on kerroksessa 1, yksi hissi on kerroksessa 8 ja yksi hissi on kerroksessa 16. Näistä hissiin haluava voi astua siihen suoraan. Jos tilaa hissin muualta, odotteluun kuluu 10 sekuntia ja lisäksi 5 sekuntia jokaista kerrosta kohden, jonka hissi joutuu kulkemaan. Kuinka suurella todennäköisyydellä tilattua hissiä joutuu odottamaan ruuhkattomana aikana yli 22 sekuntia?

Tehtävä 72. (Kevät 11, pitkä. Voi olla vaikea)

Lasten Lotossa rastitaan alle kuvatusta ruudukosta kolme ruutua ja arvonnassa muodostetaan kolmen numeron oikea rivi. Laske todennäköisyydet saada nolla, yksi, kaksi tai kolme oikein. Mikä on näiden todennäköisyyksien summa?

1	2	3	4	5	6	7	8	9	10

Tehtävä 73. (Syksy 10, pitkä)

Monivalintatestissä on 25 väitettä ja kussakin kaksi vastausvaihtoehtoa. Opiskelija tietää oikean vastauksen 10 väitteeseen, mutta joutuu arvaamaan loput. Millä todennäköisyydellä hän läpäisee testin, kun läpipääsyyn vaaditaan 15 oikeaa vastausta?

Tehtävä 74. (Kevät 08, pitkä)

CD-levyllä on viisi kappaletta, ja henkilö kuuntelee levyn päivittäin yhden viikon aikana siten, että hän asettaa soittimen toistamaan kappaleet satunnaisessa järjestyksessä. Millä todennäköisyydellä kappaleet tulevat ainakin kerran kuunnelluiksi siinä järjestyksessä, jossa ne ovat levyllä?

Tehtävä 75. (Kevät 03, pitkä. Ekassa kysymyksessä pitää ajatella ehdollisia todennäköisyyksiä.)

Tilastojen mukaan eräässä pääsykuulustelussa 25 % pyrkijöistä epäonnistuu matematiikan ja 17 % fysiikan kokeessa. Pyrkijöistä 10 % epäonnistuu kummassakin kokeessa. Laske todennäköisyys, että fysiikan kokeessa epäonnistunut pyrkijä epäonnistuu myös matematiikan kokeessa. Millä todennäköisyydellä pyrkijä epäonnistuu ainakin toisessa kokeessa?

Tehtävä 76. (Syksy 01, pitkä)

Eräällä paikkakunnalla sataa 60 prosentin todennäköisyydellä, jos edellisenä päivänä on satanut; poutasään todennäköisyys on tällöin 40 prosenttia. Jos taas edellisenä päivänä on ollut pouta, sateen todennäköisyys on vain 20 prosenttia ja poudan todennäköisyys vastaavasti 80 prosenttia. Millä todennäköisyydellä ylihuomenna sataa, kun tänään on pouta?

Tehtävä 77. (Kevät 01, pitkä)

Tutkimuksessa todettiin, että 200 gramman keksipakkausten massan keskiarvo oli 204 g ja keskihajonta 6 g. Oletetaan, että massa on normaalisti jakautunut. Kuinka monella prosentilla pakkauksista massa oli alle 200 g? Kuinka monella prosentilla pakkauksista massa oli välillä 200 g – 210 g?

Tehtävä 78. (Syksy 99, pitkä)

a) Kukansiemeniä sisältävän säkin kyljessä kerrotaan, että siementen itämistodennäköisyys on 95 % ja että 5 % säkin sisällöstä on samannäköisiä rikkaruohon siemeniä. Säkin siemenet jaetaan kahdenkymmenen siemenen pusseihin. Millä todennäköisyydellä puutarhuri, joka kylvää tällaisen pussillisen siemeniä, saa vähintään 19 haluamaansa kukantainta? Millä todennäköisyydellä hän kylvää vähintään yhden rikkaruohonsiemenen?

Tehtävä 79. (Syksy 12, lyhyt)

Eräässä tutkimuksessa mitattiin tiettyä lisäainepitoisuutta sadassa pullollisessa virvoitusjuomaa. Pitoisuuden keskiarvoksi saatiin $\bar{x} = 0,215$ % ja keskihajonnaksi $s = 0,005$ %. Lisäainepitoisuus noudattaa normaalijakaumaa. Millä todennäköisyydellä lisäaineen pitoisuus pullossa ylittää sallitun rajan $0,225$ %?

Tehtävä 80. (Kevät 12, lyhyt)

Vuorokauden keskilämpötila maaliskuussa on eräällä paikkakunnalla normaalijakautunut niin, että odotusarvo on $4,0$ °C ja 90 % vuorokautisista keskilämpötiloista on $2,0$ °C – $6,0$ °C. Laske keskilämpötilan keskihajonta.

9. Odotusarvo (kuuluu OPSiin, mutta ei ole pakollinen kurssilla)

Jos aiot heittää reilua kolikkoa sata kertaa ja laskea kruunien lukumäärän, odotat saavasi tulokseksi suunnilleen 50. Kokeen *odotusarvo* on siis 50.

> **Odotusarvo toistokokeessa**
>
> ---
>
> Jos koe toistetaan n kertaa ja yhden onnistumisen todennäköisyys on p, on onnistumisten lukumäärän odotusarvo np.

Tehtävä 81. Heität noppaa 600 kertaa. Mikä on kutosten lukumäärän odotusarvo?

Odotusarvoja voi laskea monelle muullekin asialle kuin onnistumisen lukumäärälle. Uhkapeleissä ja samanluonteisissa ilmiöissä (kuten talous, vakuutukset jne) saadun tai menetetyn rahan odotusarvoa voi pitää päätöksen tekemisen kriteerinä.

Esimerkki 9.1. Peluri A ehdottaa pelurille B, että he heittävät noppaa. Jos silmäluku on 1-5, A maksaa B:lle euron. Jos silmäluku on kuusi, B maksaa A:lle kuusi euroa. Kannattaako B:n suostua peliin?

Ajatellaan, että noppaa heitettäisiin yhteensä vaikka kuusi miljoonaa kertaa. Näistä luultavasti noin miljoonasta tulisi kutonen ja viidestä miljoonasta jokin muu. A siis maksaisi B:lle noin 5 000 000 euroa, ja B maksaisi A:lle noin 6 000 000 euroa. A jäisi siis pelissä voitolle, eikä B:n kannata suostua siihen.

Monen ensireaktio peliin olisi, että se on reilu, koska maksujen suhteet ovat 1:6 ja nopassa yksi kuudesta silmäluvusta on kutonen. Voittamisten suhde ei kuitenkaan ole 1:6, vaan 1:5, kutonen vastaan viisi muuta silmälukua.

Jos pysytään rahaesimerkeissä, odotusarvo lasketaan käymällä kaikki alkeistapahtumat läpi, kertomalla niiden todennäköisyydet niistä aiheutuvilla tuotoilla tai tappioilla, ja laskemalla näin saadut luvut yhteen. Seuraava esimerkki näyttää asian selkeämmin:

Esimerkki 9.2. Henkilö A matkustaa junalla 40 kertaa kuussa. Hän saa 80 euron sakot yhden matkan aikana todennäköisyydellä 0.01. Kuukausikortin ostaminen maksaa 50 euroa. A:ta ei kiinnosta pummilla matkustamisen moraaliset ongelmat, vaan ainoastaan raha. Kannattaako hänen ostaa lippu?

Helpo tapa: Kun matkoja on 40 ja sakkojen saamisen todennäköisyys on 0.01, on sakkojen odotusarvo kuukauden aikana 0.4 kappaletta. Sakkojen euromäärän odotusarvo on siis $0.4 \cdot 80 = 32$ euroa.

Vaikeampi tapa, jonka lopputulos on tietenkin sama, mutta joka on myös yleisempi. Käytetään taulukkolaskentaa:

C2 =80*A2*B2

	A	B	C	D
1	Matkojen lkm	Sakkojen TN	Odotusarvo	
2	0	0.6689717586	0	
3	1	0.2702916196	21.62332957	
4	2	0.0532392584	8.518281346	
5	3	0.006811757	1.634821672	
6	4	0.000636452	0.203664653	
7	5	4.62874211E-05	0.018514968	
8	6	2.7273733E-06	0.001309139	
9	7	1.33810522E-07	7.49339E-05	
10	8	5.57543844E-09	3.56828E-06	
11	9	2.00240213E-10	1.44173E-07	
12	10	6.27014809E-12	0.000000005	
13	11	1.72731352E-13	1.52004E-10	

Tässä ensimmäisessä sarakkeessa on kaikki mahdolliset saatujen sakkojen lukumäärät.
Toisessa sarakkeessa on todennäköisyydet saada näin monta sakkoa, B2:ssa `=BINOM.DIST(A2, 40, 0.01, 0)` ja muut kopioitu siitä.
Kolmannessa sarakkeessa on odotettu rahanmenetys kutakin tarkastusmäärää varten. Jos A saa sakot esimerkiksi kaksi kertaa, hän menettää rahaa 160 euroa, mutta C-sarakkeeseen tämä on kerrottu todennäköisyydellä saada sakot kaksi kertaa, eli n. 0.053:lla. C2:ssa siis on `=80*A2*B2`, ja alemmat ruudut ovat tästä kopioituja.

Lopullinen odotusarvo saadaan laskemalla yhteen C-sarakkeen luvut:

37	35	6.25759061E-65	1.75213E-61
38	36	8.77888694E-68	2.52832E-64
39	37	9.58655412E-71	2.83762E-67
40	38	7.64478E-74	2.32401E-70
41	39	3.96E-77	1.23552E-73
42	40	1E-80	3.2E-77
43			
44			32
45			

Pummilla matkustamisen odotusarvo olisi siis 32 euroa, ja kannattavampaa kuin lipun ostaminen.

Odotusarvo voidaan laskea myös monissa muissakin tapauksissa kuin toistokokeessa. Mietitään vain tapauksia, joissa vaihtoehtoja on äärellisen monta ja lasketaan odotusarvoa rahan saamiselle/menettämiselle. Tällöin:

(1) Käy jokainen vaihtoehto läpi ja laske

$$P(\text{vaihtoehto}) \cdot (\text{vaihtoehdon tuottama rahamäärä})$$

(2) Laske edellisessä kohdassa saadut luvut yhteen.

Esimerkki 9.3. Pelurin kolikko antaa kruunan todennäköisyydellä 0.54. Kannattaako hänen ehdottaa kollegallensa peliä, jossa hän voittaa euron kruunalla ja häviää 1.05 euroa klaavalla?

Pelin odotusarvo on

$$0.54 \cdot 1.00€ + 0.46 \cdot (-1.05€) = 0.057€.$$

Koska odotusarvo on positiivinen, hänen kannattaa ehdottaa peliä.

Esimerkki 9.4. Peluri pelaa peliä, jossa hän heittää noppaa ja voittaa tai häviää seuraavan taulukon mukaan.

Silmäluku	Voitto
1	4
2	3
3	-6
4	2
5	-1
6	-3

Jokaisen silmäluvun todennäköisyys on $\frac{1}{6}$, joten “todennäköisyys kertaa voitto” on $\frac{4}{6}$, $\frac{3}{6}$ jne., täytetään sarakkeeseen “E”

Silmäluku	Voitto	E
1	4	4/6
2	3	3/6
3	-6	-6/6
4	2	2/6
5	-1	-1/6
6	-3	-3/6

Kun sarake E lasketaan yhteen, saadaan $\frac{-1}{6} < 0$. Peluri siis voi odottaa häviävänsä, eikä peli ole hänelle kannattava.

Tehtävä 82. Pelaat peliä, jossa heitetään noppaa. Voitat parillisilla silmäluvuilla silmäluvun verran euroissa ja parittomilla häviät kolme euroa. Onko peli sinulle kannattava?

Tehtävä 83. Henkilö pelaa Monopolia. Hän heittää kahta noppaa. Silmäluvuilla 7 hän joutuu Mannerheimintielle ja joutuu maksamaan 24 000 pelirahaa. Silmäluvulla 9 hän joutuu Erottajalle ja maksaa 40 000 pelirahaa. Silmäluvulla 10 ja suuremmilla hän ylittää lähtöruudun ja saa 4000 pelirahaa. Muussa tapauksessa hän ei saa eikä menetä rahaa. Mikä on rahan saamisen odotusarvo?