Tilastotieteen johdantokurssi

Apulaisprofessori Ville Hautamäki

School of Computing
University of Eastern Finland

October 9, 2023

Todennäköisyyslaskentaa (1)

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

Todennäköisyyslaskentaa (2)

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

Klassinen todennäköisyyden määritelmä

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

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

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

Subjektiivinen todennäköisyys

- Toinen yleisesti käytetty tulkinta todennäköisyydelle on **subjektiivinen todennäköisyys**, joka kuvaa uskomusta tapauksen todennäköisyydestä, mutta ei välttämättä sisällä mittalukua todennäköisyydestä.
- Esimerkiksi väite "Todennäköisesti huomenna sataa" kuvaa väitteen esittäjän subjektiivista uskomusta sateen mahdollisuudesta.
- Subjektiivinen todennäköisyys -pohjainen tilastotiede tunnetaan nimellä Bayesilainen tilastotiede. Siinä todennäköisyys on aina suhteessa (priori) alkuperäiseen uskomukseen.
- Kun data on havaittu, voidaan uskomus päivittää vastaamaan todellisuutta.
- Historiallisesti, Bayesiläinen todennäköisyyspäättely kehitettiin ensiksi (1700-luvulla) ja frekventistinen 1900-luvun alussa.
- Andrei Kolmogorov systematisoi ja formalisoi todennäköisyyslaskennan 1900-luvun alkupuolella.

Tilastollinen todennäköisyyden tulkinta

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

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

Todennäköisyysmalli

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

Todennäköisyysmallinnuksen perusperiaatteet

Edellä oleva kuvaus nopanheitosta sisältää kaksi osaa:

- Listan kaikista kokeen tulosvaihtoehdoista.
- Kunkin tulosvaihtoehdon todennäköisyydet

Nämä kaksi tekijää muodostavat perustan satunnaisilmiötä kuvaavalle todennäköisyysmallille.

Otosavaruus ja alkeistapaus

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

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

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

Esimerkki 5.2

Heitettäessä kolikkoa kahdesti voidaan saada neljä erilaista tulosta:

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

- Otosavaruus kahden kolikon heitossa on siis

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

  Tässä otosavaruudessa on siis neljä alkeistapausta.

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

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

Satunnaisotanta ja tapahtuma

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

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

Esimerkki 5.3

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

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

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

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

Esimerkki 5.4

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

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

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

Todennäköisyysmallista

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

A. Kolmogorovin todennäköisyyden perusominaisuudet

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

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

   Tämä on **yhteenlaskusääntö**. Yleisessä tapauksessa pätee

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

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

Esimerkki 5.5, veriryhmien todennäköisyyksistä

Satunnisesti valitun henkilön veriryhmä noudattelee oheista taulukkoa:

Veriryhmä	O	A	B	AB
Todennäköisyys	0.49	0.27	0.20	?

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

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

Esimerkki 5.5 (jatkuu)

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

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

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

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

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

Esimerkki 5.6

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

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

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

- Millä todennäköisyydellä 1-numeroinen satunnaisluku on pariton?

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

Esimerkki 5.6 (jatkuu)

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

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

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

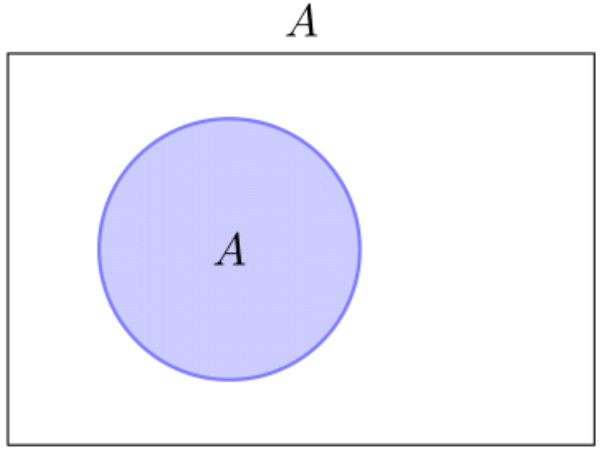
$$P(A \,\text{tai}\, B) = P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{13}$$

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

Visualisointi Venn-diagrammeilla

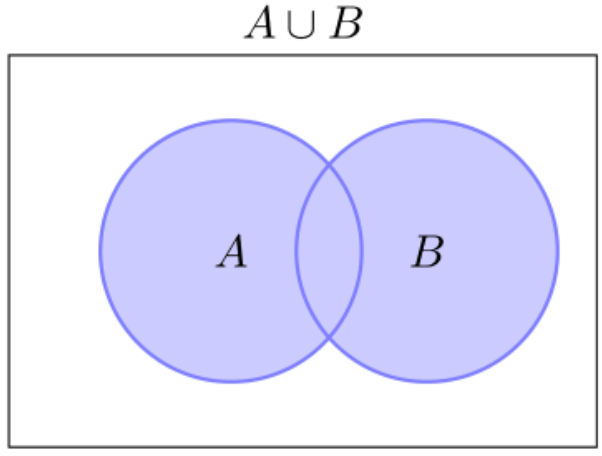
Ensiksi perusjoukko S

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

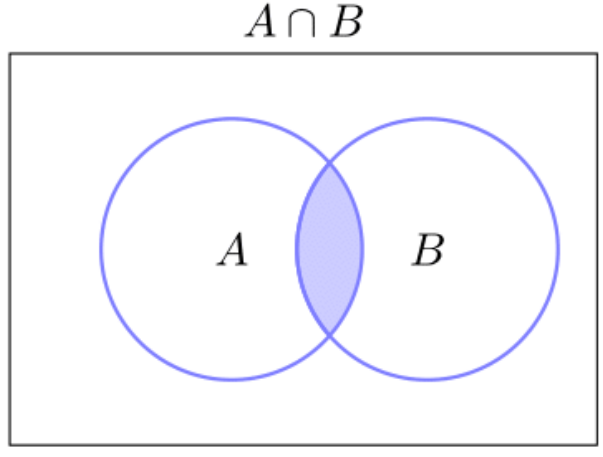


Visualisointi jatkuu

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

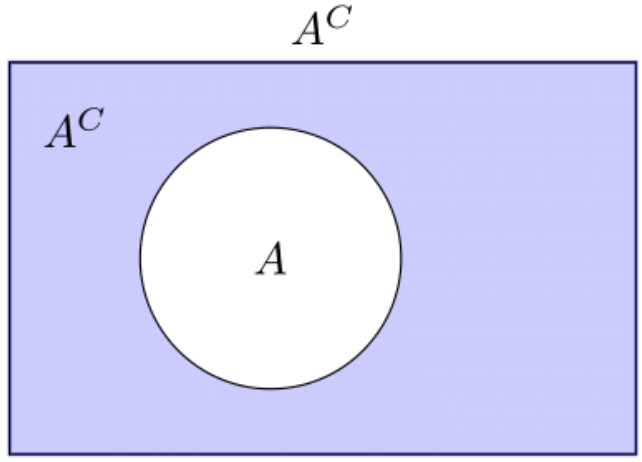


"A ja B" eli A ja B tapahtuu yhtäaikaa. Joukko-oppi: $A \cap B$

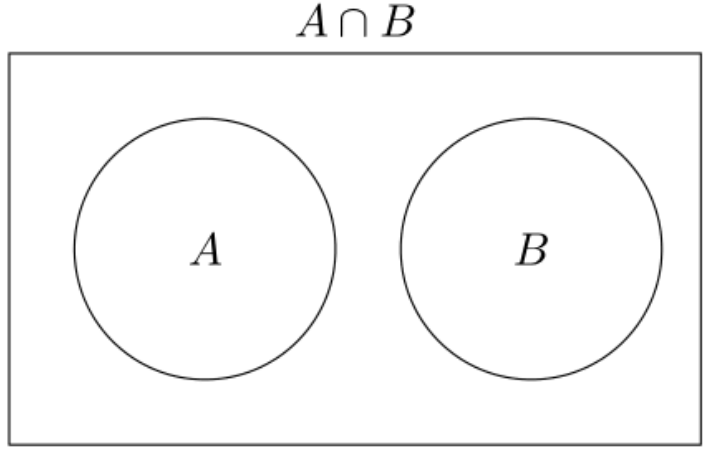


Visualisointi jatkuu

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



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



Esimerkki 5.8

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

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

Esimerkki 5.9

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

Ehdollinen todennäköisyys ja riippumattomuus

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

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

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

Esimerkki: vaalit Yhdysvalloissa

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

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

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

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

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

Riippumattomuus

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

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

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

Esimerkki 5.10

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

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

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

Esimerkki 5.10 jatkuu

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

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

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

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

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

Esimerkki 5.10 jatkuu

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

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

Esimerkki 5.11

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

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

Toisaalta voidaan laskea myös suoraan hyödyntäen riippumattomien tapahtumien kertolaskusääntöä

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

Esimerkki 5.12

Kuopion sosiaali- ja terveystointen yhdistämisen vaikutuksista tehdyssä tutkimuksessa kyselylomakkeessa kysyttiin henkilön sukupuoli ja mielipide siitä, takaavatko vastaanottotilat yksityisyyden.

	Mies	Nainen	Yhteensä
Hyvin	40	292	332
Huonosti	7	144	151
Yhteensä	47	436	483

Esimerkki 5.12 (jatkuu)

Poimitaan aineistosta yksi henkilö satunnaisesti ja tarkastellaan tapahtumia:
$A =$ "Henkilö kokee vastaanottotilojen takaavan yksityisyyden hyvin",
$B =$ "Henkilö on mies"

$$\begin{aligned} P(A) &= \frac{332}{483} \approx 0.687 \\ P(B) &= \frac{47}{483} \approx 0.0973 \\ P(A \cap B) &= \frac{40}{483} \approx 0.0828 \end{aligned} \tag{23}$$

$P(A \cap B)$ on todennäköisyys sille, että valittu henkilö kokee yksityisyyden vastaanottotiloissa hyväksi ja on mies.

Esimerkki 5.12 (jatkuu)

Jos tiedetään valitun henkilön olevan mies, niin hän kokee vastaanottotilojen takaavan yksityisyyden hyvin todennäköisyydellä

$$P(B|A) = \frac{P(A \cap B)}{P(B)} = \frac{40/483}{47/483} = \frac{40}{47} = 0.85 \tag{24}$$

Tapahtumat A ja B eivät ole riippumattomia, koska

$$P(A \cap B) \neq P(A)P(B) \tag{25}$$

eli

$$\frac{40}{483} \approx 0.0828 \neq \frac{332}{483} \times \frac{47}{483} \approx 0.067 \tag{26}$$

Lyhyesti kombinatoriikasta

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

Erilaisten lottorivien lukumäärä

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

$$\binom{40}{7} = \frac{40!}{7!(40-7)!} = \frac{40!}{7!33!} = 18\ 643\ 560 \qquad (27)$$

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

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad (28)$$

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

Kertomasta

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

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1 \tag{29}$$

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

Esimerkki 5.13

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

```
# Kertoma
factorial(3)
## [1] 6
# 3 yli 2
choose(3,2)
## [1] 3
```

Esimerkki 5.13

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

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

Luku 6. Satunnaismuuttujat

Johdanto: Satunnaismuuttujat

- **Satunnaismuuttuja** on mikä tahansa numeerinen muuttuja, jonka arvon määrää satunnaiskokeen tulos.
- Toisin sanoen se on numeerinen muuttuja, jonka havaittuun arvoon sattuma vaikuttaa.
- Satunnaismuuttujia merkitään isoilla kirjaimilla (esim. X, Y, Z, X_1, X_2)
- Satunnaismuuttujan saamia yksittäisiä arvoja taasen merkataan pienillä kirjaimilla, kuten x, y, z

Satunnaismuuttujien tyypeistä

- Satunnaismuuttujia on monenlaisia. Esimerkiksi nopan heiton tulos on usein käytetty esimerkki satunnaismuuttujasta.
- mutta samalla tavalla satunnaismuuttujina käsitellään tutkimuksissa mitattavia ominaisuuksia, esimerkiksi maanäytteen typpipitoisuutta.
- Nopan heiton tulos puolestaan on esimerkki diskreetistä satunnaismuuttujasta ja maanäytteen typpipitoisuus jatkuvasta satunnaismuuttujasta.

Satunnaismuuttuja voi saada vain numeerisia arvoja

- Erotuksena todennäköisyyslaskennan **tapahtumiin** satunnaismuuttuja voi saada vain numeerisia arvoja
- Esimerkiksi, kolikkoa heitettäessä pitää päättää millä numerolla merkitään kruunaa ja klaavaa (esim kruuna $= 0$ ja klaava $= 1$)
- Koodaustavan voi itse valita, mutta eri valinnat johtavat eri satunnaismuuttujiin.

Todennäköisyysjakauma ja odotusarvo

- Satunnaismuuttujan **todennäköisyysjakauma** kertoo, mitä arvoja satunnaismuuttuja voi saada ja millä todennäköisyyksillä se mitäkin arvoja saa.
- Satunnaismuuttujan **odotusarvo** (E[X], μ) ilmaisee minkä arvon ympärille satunnaismuuttujan arvot keskittyvät.
- Odotusarvo voidaan tulkita laskennallisena keskiarvona jos olemme keränneet äärettömän paljon dataa.

Varianssi ja keskihajonta

- Satunnaismuuttujan **varianssi** ($\mathrm{Var}(X)$, σ^2) ja sen neliöjuuri **keskihajonta** kuvaavat kuinka paljon satunnaismuuttujan yksittäiset arvot vaihtelevat odotusarvon ympärillä.
- Odotusarvo siis kertoo jakauman sijainnista.
- Varianssi kertoo jakauman "leveydestä".
- Odotusarvo ja varianssi eivät kuitenkaan kerro paljoakaan jakauman todellisesta muodosta $\rightarrow$ **erilaiset jakaumat voivat tuottaa saman odotusarvon ja varianssin!**

Diskreetti satunnaismuuttuja

- **Diskreetti satunnaismuuttuja** saa äärellisen tai numeroituvasti äärettömän määrän eri arvoja.
- Diskreetin satunnaismuuttujan todennäköisyysjakauma, eli **pistetodennäköisyysfunktio** kirjoitetaan muodossa

$$f(x) = \begin{cases} p_1 & \text{kun } x = x_1 \\ p_2 & \text{kun } x = x_2 \\ p_3 & \text{kun } x = x_3 \\ \vdots & \\ p_k & \text{kun } x = x_k \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

Pistetodennäköisyys taulukkomuodossa

- Pistetodennäköisyysfunktio voidaan esittää myös taulukkona:

X:n arvo	x_1	x_2	x_3	$\ldots$	x_k
Todennäköisyys	p_1	p_2	p_3	$\ldots$	p_k

- Taulukossa x_1, x_2, etc ovat arvoja mitä satunnaismuuttuja X voi saada, ne määrittelevät satunnaismuuttujan **arvojoukon**.
- Diskreetin satunnaismuuttujan kohdalla arvojoukko on siis mahdollisten (numeeristen) arvojen listaus.
- Jos satunnaismuuttujan mahdollisten arvojen määrä on numeroituvasti ääretön, niin silloin mahdollisten arvojen joukolla ei ole viimeistä arvoa, ja arvojoukko merkitään $x_1, x_2, \ldots$
- Toisella rivillä ovat pistetodennäköisyysfunktion arvot p_i , jotka kuvaavat todennäköisyyttä, että satunnaismuuttuja X saa arvon x_i, eli $p_i = P(X = x_i)$

Lisää pistetodennäköisyydestä

- Arvoilla, jotka eivät kuulu arvojoukkoon, pistetodennäköisyysfunktio saa arvon 0.
- Jokainen diskreetin satunnaismuuttujan saama todennäköisyys p_i saa arvon välillä $[0, 1]$.
- Satunnaismuuttujan todennäköisyyksien summa $\sum_{i=1} p_i = 1$.

Esimerkki 6.1

Kolikon heitto kaksi kertaa. Satunnaiskokeen otosavaruus on

$$S = \{\text{krkr}, \text{krkl}, \text{klkr}, \text{klkl}\} \tag{31}$$

Oletetaan kolikon olevan harhaton, eli

$$P(\text{"kr"}) = P(\text{"kl"}) = \frac{1}{2} \tag{32}$$

Merkitään satunnaismuuttujalla X kruunien lukumäärää kahdessa heitosssa. Nyt satunnaismuuttuja X voi siis saada arvon 0,1 tai 2. X on satunnaismuuttuja, jonka arvo määräytyy satunnaiskokeen tuloksen perusteella, ja sen en arvojoukko on $S = \{0, 1, 2\}$.

Esimerkki 6.1 (jatkuu)

Pistetodennäköisyysfunktion määrittämiseksi lasketaan kunkin arvon todennäköisyys:

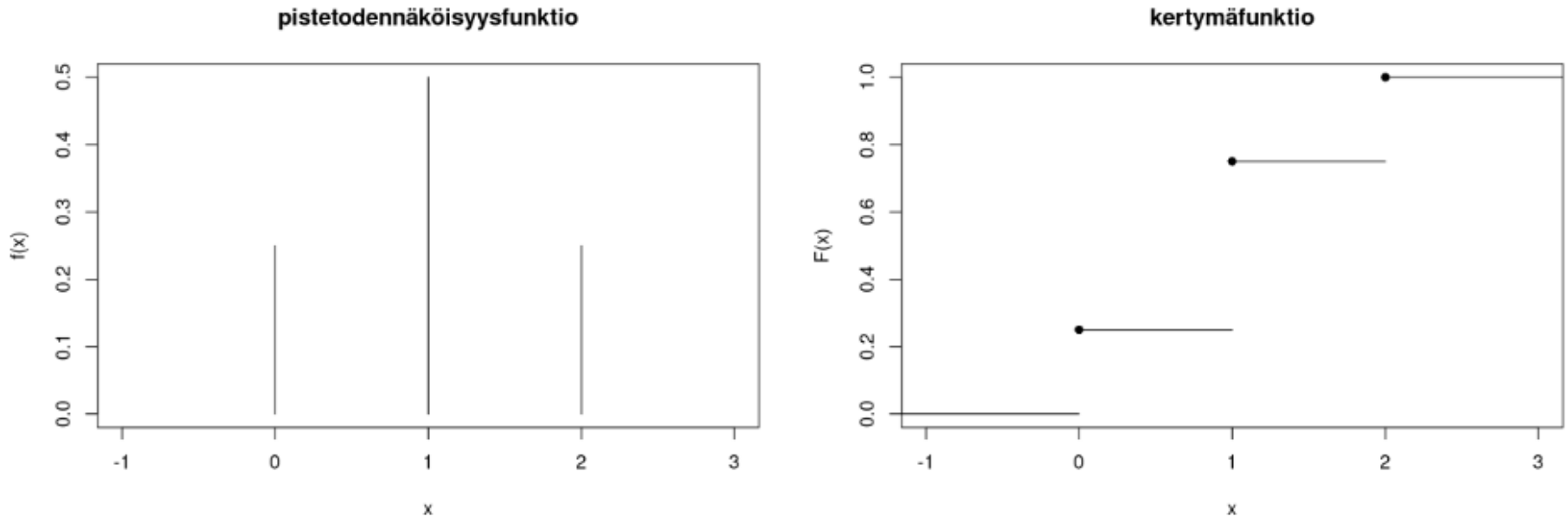
$$
\begin{aligned}
P(X=0) &= P(\text{klkl}) = \frac{1}{4} \\
P(X=1) &= P(\text{"krkl" tai "klkr"}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \\
P(X=2) &= P(\text{krkr}) = \frac{1}{4}
\end{aligned}
\tag{33}
$$

Satunnaismuuttujan X pistetodennäköisyysfunktio on siis

X:n arvo	0	1	2
Todennäköisyys	1/4	1/2	1/4

Esimerkki 6.1 (jatkuu)

Funktion kuvaaja on esitetty vasemmassa kuvassa



Voidaan laskea

$$P(X \leq 1) = P(X = 0) + P(X = 1) = \frac{3}{4} \qquad (34)$$

ja

$$P(X > 0) = P(X = 1) + P(X = 2) = \frac{3}{4} \qquad (35)$$

Kertymäfunktio diskreetille satunnaismuuttujalle

- **Kertymäfunktioksi** kutsutaan

$$F(x) = P(X \leq x) \tag{36}$$

- Kertymäfunktio (engl. cumulative distribution function, cdf, tai vain distribution function)
- Pistetodennäköisyysfunktio ilmaisee siis yksittäisiin tapahtumiin liittyvät todennäköisyydet ja kertymäfunktio ilmaisee todennäköisyyden, jolla satunnaismuuttuja saa enintään arvon x.
- Diskreetin satunnaismuuttujan kertymäfunktiolle on voimassa:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i) = \sum_{x_i \leq x} P(X = x_i) \tag{37}$$

Esimerkki 6.2

Satunnaismuuttujan X arvo on kruunujen lukumäärä kahden kolikon heitossa. Kertymäfunktion $F(x)$ arvot ovat:

$$\begin{aligned} F(0) &= P(X \leq 0) = P(X = 0) = 1/4 \\ F(1) &= P(X \leq 1) = P(X = 0) + P(X = 1) = 3/4 \\ F(2) &= P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 1 \end{aligned} \tag{38}$$

Joka voidaan kirjoittaa taulukkona

x_i	0	1	2
$f(x_i) = P(X = x_i)$	1/4	1/2	1/4
$F(x_i)$	1/4	3 / 4	1

Esimerkki 6.2 (jatkuu)

Matemaattisesti satunnaismuuttujan X kertymäfunktio voidaan ilmaista

$$F(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq x < 2 \\ 1 & 2x \end{cases} \tag{39}$$

Huomaa että oikean puoleinen raja ei kuuluu välille!

Jatkuva satunnaismuuttuja

- **Jatkuva satunnaismuuttuja** voi saada minkä tahansa arvon jollain välillä. Mahdollisia arvoja on siis ääretön (ylinumeroituva) määrä.
- **Tiheysfunktio** $f(x)$ on jatkuvan satunnaismuuttujan todennäköisyyden jakautumista kuvaava funktio.
- Tiheysfunktio on englanniksi **probability density function** (pdf)
- Tiheysfunktion arvot eivät kuitenkaan ole yksittäisen arvon esiintymisen todennäköisyyksiä; jatkuvalla satunnaismuuttujalla yksittäisen arvon todennäköisyys on aina nolla.

Tiheysfunktion ominaisuuksia

Tiheysfunktiolla on seuraavat ominaisuudet:

1. $f(x) \geq 0$, kaikilla x:n arvoilla. Eli tiheysfunktio ei voi saada negatiivisia arvoja.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$, eli tiheysfunktion ja x-akselin rajaaman alueen pinta-ala on 1.

Jatkuvalle satunnaismuuttujalle X tapahtuman $x_1 \leq X \leq x_2$ voidaan laskea integroimalla:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)dx. \tag{40}$$

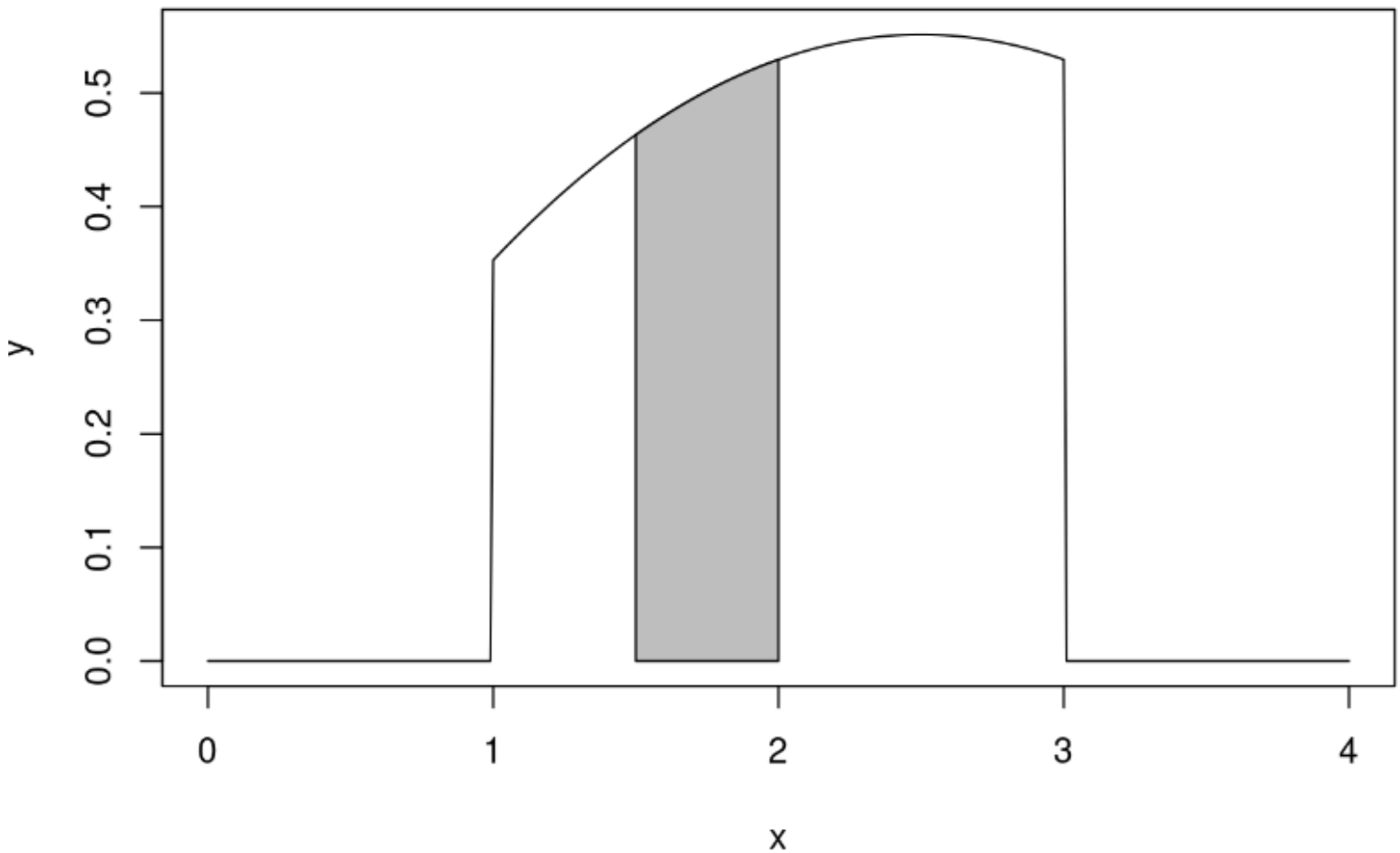
Todennäköisyys saadaan siis laskemalla tiheysfunktion ja x-akselin rajaaman alueen pinta-ala välillä $[x_1, x_2]$.

Esimerkki 6.3

Määritellään satunnaismuuttujan X tiheysfunktio kaavalla

$$f(x) = \begin{cases} -\frac{3}{34}x^2 + \frac{15}{34}x & \text{kun } 1 \leq x \leq 3 \\ 0 & \text{, muulloin} \end{cases} \tag{41}$$

a. Piirrä tiheysfunktion kuvaaja



Esimerkki 6.3 (jatkuu)

b. Osoita että funktio toteuttaa em. tiheysfunktion ominaisuudet. Funktio saa vain positiivisia arvoja (millä perusteella??), joten se toteuttaa em. ominaisuuksista ensimmäisen. Funktion ja x-akselin välille jäävä pinta-ala välillä $[1, 3]$ lasketaan (tarkkaan ottaen approksimoidaan, mutta approksimaatio on hyvin tarkka) ao. koodissa funktiolla

```
    # määritetään tiheysfunktio R-funktiona
fx <- function(x) -3/34*x^2+15/34*x
integrate(fx, 1, 3) # käyrän alle jäävä pinta-ala

## 1 with absolute error < 1.1e-14
```

Esimerkki 6.3 (jatkuu)

c. Laske todennäköisyys $P(1.5 \leq X < 2)$
Kysytty todennäköisyys voidaan laskea integraalina $\int_{1.5}^{2} f(x)dx$, jota approksimoidaan ao. koodissa funktiolla. Tulokseksi saadaan

$$P(1.5 \leq X < 2) = 0.25 \tag{42}$$

Kertymäfunktio jatkuvalle satunnaismuuttujalle

- Jatkuvan satunnaismuuttujan X kertymäfunktio määritellään samalla tavalla kuin diskreetille satunnaismuuttujalle:

$$F(x) = P(X \leq x) \tag{43}$$

- Kertymäfunktio on siis jatkuva ja yhtenäinen käyrä (vertaa diskreettiin tapaukseen!)
- Jatkuvan satunnaismuuttujan tiheysfunktio $f(x)$ on on kertymäfunktion derivaatta x: suhteen $f(x) = F'(x)$

Kertymäfunktion ominaisuuksia

- Jatkuvan satunnaismuuttujan kertymäfunktiolle on voimassa

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(s)ds \tag{44}$$

- Määritelmästä johtuen

$$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1) \tag{45}$$

- Tästä seuraa, että kun $x_2 \to x_1$, niin $F(x_2) \to F(x_1)$, joten jatkuvalle satunnaismuuttujalle $P(X = x_1) = 0$ aina!
- Eli yksittäisen arvon x todennäköisyys on aina nolla!

Jatkuvan ja diskreetin satunnaismuuttujan eroista

- Jatkuvaan satunnaismuuttujaan liittyvät todennäköisyydet lasketaan välien avulla - toisin kuin diskreettien satunnaismuuttujien tapauksessa!
- Diskreetti jakauma siis liittää todennäköisyydet yksittäisiin arvoihin ja jatkuva jakauma liittää todennäköisyydet väleihin.
- Jatkuvalla satunnaismuuttujalla ei ole merkitystä, ovatko välit avoimia, suljettuja vai puoliavoimia, esim:

$$P(X \leq x) = P(X < x). \tag{46}$$

Odotusarvo

- Satunnaismuuttujan odotusarvo $\mathrm{E}[X]$ kuvaa sitä, minkä arvon ympärille satunnaismuuttujan arvot keskittyvät.
- **Diskreetin satunnaismuuttujan** X odotusarvo määritellään lausekkeella

$$\mathrm{E}[X] = \sum_{i=1}^{k} x_i P(X = x_i) = \sum_{i=1}^{k} x_i p_i \tag{47}$$

- Odotusarvo saadaan siis kertomalla jokaisen satunnaismuuttujan arvo kyseiseen arvoon liittyvällä todennäköisyydellä ja laskemalla tulot yhteen.
- Kyseessä on siis painotettu keskiarvo.
- Odotusarvoa merkitään myös symbolilla μ.

Odostusarvosta lisää

- Satunnaismuuttujalla, jonka arvojoukko on numeroituvasti ääretön, summassa on äärettömän monta termiä eli odotusarvo on

$$E[X] = \sum_{i=1}^{\infty} x_i p_i \tag{48}$$

- Huomaa, että satunnaismuuttujan odotusarvo ei kuulu välttämättä satunnaismuuttujan arvojoukkoon; esimerkiksi harhattomalla nopalla heitettäessä nopan silmäluvun arvojoukko on $\{1, 2, 3, 4, 5, 6\}$, mutta odotusarvo on 3.5

Esimerkki 6.4

Jatkoa esimerkille 6.1, jossa määriteltiin pistetodennäköisyysfunktio

$$F(x) = \begin{cases} 1/4 & , \text{kun } x = 0 \\ 1/2 & , \text{kun } x = 1 \\ 1/4 & , \text{kun } x = 2 \\ 0 & , \text{muulloin} \end{cases} \tag{49}$$

Kruunujen lukumäärän odotusarvo kahta kolikkoa heitettäessä on

$$\mathrm{E}[X] = \sum_{i=1}^{3} x_i p_i = 0 \times 1/4 + 1 \times 1/2 + 2 \times 1/4 = 1 \tag{50}$$

Eli keskimäärin kahta kolikkoa heitettäessä saat yhden kruunun.

Jatkuvan satunnaismuuttujan odotusarvo

- Jatkuvan satunnaismuuttujan odotusarvo määritellään lausekkeella

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x) dx \tag{51}$$

- Mistä huomaamme että määritelmä on analoginen diskreetin satunnaismuuttujan tapauksen kanssa.

Odotusarvon laskusääntöjä

Lause 6.1

Odotusarvon laskusääntöjä

1. $\mathrm{E}[c] = c$ mille tahansa vakiolle c.
2. $\mathrm{E}[cX] = c\mathrm{E}[X]$ satunnaismuuttujalle X, kun c on vakio.
3. $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$, satunnaismuuttujille X ja Y.

Varianssi

- Satunnaismuuttujan **varianssi** määritellään lausekkeella

$$\mathrm{Var}(X) = \mathrm{E}[X - \mathrm{E}[X]]^2 = \mathrm{E}[X - \mu]^2 \tag{52}$$

- Satunnaismuuttujan varianssille käytetään myös merkintää σ^2
- Varianssin neliöjuurta kutsutaan **keskihajonnaksi**

$$\sigma = \sqrt{\sigma^2} = \sqrt{\mathrm{Var}(X)} \tag{53}$$

- Laskuissa varianssia käytetään sen matemaattisten ominaisuuksien vuoksi, mutta raportointia varten varianssit muunnetaan yleensä keskihajonnaksi, jolla on sama yksikkö kuin satunnaismuuttujalla ja sen odotusarvolla.

Varianssin laskusääntö

- Lauseen 6.1 laskusääntöjä soveltamalla voidaan varianssin määritelmä kirjoittaa muodossa

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2 \tag{54}$$

- Koittakaa vieruskaverin kanssa osoittaa tämä lause todeksi.
- Tämä kaava on alkuperäisiä määritelmiä helppokäyttöisempi laskennassa.
- Tarvitsemme kuitenkin tiedon siitä miten $g(X) = X^2$ voidaan laskea. Siihen tarvitsemme muunnoksen odotusarvon.

Diskreetin satunnaismuuttujan varianssi

- Diskreetille satunnaismuuttujalle varianssi saa muodon

$$\mathrm{Var}(X) = \sum_{i=1}^{k} (x_i - \mu)^2 p_i \tag{55}$$

- Numeroituvasti äärettömän arvojoukon tapauksessa varianssi lasketaan

$$\mathrm{Var}(X) = \sum_{i=1}^{\infty} (x_i - \mu)^2 p_i \tag{56}$$

Esimerkki 6.5

Jatkoa esimerkille 6.1. Kruunujen lukumäärän varianssi kahta kolikkoa heitettäessä on

$$\begin{aligned} \mathrm{Var}(X) &= (0-1)^2 \times \frac{1}{4} + (1-1)^2 \times \frac{1}{2} + (2-1)^2 \times \frac{1}{4} \\ &= \frac{1}{2} \end{aligned} \tag{57}$$

ja keskihajonta

$$\sigma = \sqrt{\frac{1}{2}} = 0.707 \tag{58}$$

Esimerkki 6.6

Tarkastellaan kahta noppaa heitettäessä saatua silmälukujen summaa X . Summan X pistetodennäköisyysfunktio on

2	3	4	5	6	7	8	9	10	11	12
1/36	2/36	3/36	4/36	5/36	6/35	5/36	4/36	3/36	2/36	1/36

Laske satunnaismuuttujan X odotusarvo ja varianssi.

Esimerkki 6.6 (jatkuu)

Diskreetin satunnaismuuttujan odotusarvon ja varianssin kaavoilla saadaan (kts. R-koodi alla)

$$\mathrm{E}[X] = 7, \quad \mathrm{Var}(X) = 5.83, \quad \sigma = 2.41 \tag{59}$$

```
# Diskreetti SM
x <- seq(2, 12)
f <- c(1,2,3,4,5,6,5,4,3,2,1)/36
sum(f) # tarkista että todennäköisyydet summautuu ykköseen
## [1] 1
```

Esimerkki 6.6 (jatkuu)

Ja sitten lopullinen tulos saadaan käyttäen jo laskettua odotusarvoa

```
mu1 <- sum(f*x)
sigma21 <- sum(f*(x-mu1)^2)
c(mu1, sigma21) # E(X) ja var(X)
## [1] 7.000000 5.833333
```

Jatkuvan satunnaismuuttujan varianssi

- Jatkuvan satunnaismuuttujan varianssi määritellään lausekkeella

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx \tag{60}$$

- Myös jatkuvan satunnaismuuttujan varianssille käytetään merkintää σ^2
- Samoin keskihajonta on edelleen $\sigma = \sqrt{\mathrm{Var}(X)}$

Esimerkki 6.7: tasajakauman odotusarvo ja varianssi

Laske satunnaismuuttujan X odotusarvo ja varianssi, kun satunnaismuuttujan tiheysfunktio on

$$f_X(x) = \begin{cases} \frac{1}{2} & 2 \leq x \leq 4 \\ 0 & \text{muulloin} \end{cases} \tag{61}$$

Tiheysfunktio saa siis arvon 1/2 välillä $2 \leq x \leq 4$, muutoin se saa arvon 0. Tätä jakaumaa kutsutaan myös nimellä tasajakauma (englanniksi **uniform distribution**).

Silloin yleisesti merkitään $X \sim \mathrm{Unif}(a, b)$ ja tämän esimerkin tapauksessa $X \sim \mathrm{Unif}(2, 4)$. Merkintä luetaan että "X on jakautunut tasaisesti välillä $[a, b]$".

Esimerkki 6.7 (jatkuu)

Approksimoidaan odotusarvoa ja varianssia numeerisella integroinnilla. Ao. R-koodilla saadaan $\mathrm{E}[X] = 3$ ja $\mathrm{Var}(X) = 0.333$. Huomaa, että integroinnin rajoina voidaan käyttää arvojen $-\infty$ ja ∞ sijasta satunnaismuuttujan vaihteluvälin päätepisteitä (2 ja 4), koska satunnaismuuttuja saa arvon nolla tämän välin ulkopuolella. Eikä se siten vaikuta integraalin arvoon.

Esimerkki 6.7 (jatkuu)

```
# Jatkuva satunnaismuuttuja
fx <- function(x) 1/2
xfx <- function(x) x*fx(x)
mu2 <- integrate(xfx,2,4)$value
mu2 # E(X)

## [1] 3
```

Esimerkki 6.7 (jatkuu)

Käyttäen jo laskettua odotusarvoa, saamme laskettua varianssin:

```
x2mufx <- function(x) (x-mu2)^2*fx(x)
sigma22 <- integrate(x2mufx,2,4)$value
sigma22 # var(X)
## [1] 0.3333333
```

Varianssin laskusääntöjä

1. $\mathrm{Var}(c) = 0$, missä c on vakio.
2. $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$, satunnaismuuttujalle X ja vakiolle c.
3. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \times \mathrm{Cov}(X, Y)$, satunnaismuuttujille X ja Y, kun $\mathrm{Cov}(X, Y)$ on muuttujien välinen kovarianssi. Kovarianssi on aiemmin määriteltyä otoskovarianssia vastaava populaation suure.

Lisää satunnaismuuttujien summan varianssista

- Jos satunnaismuuttujat X ja Y ovat **riippumattomia**, niin niiden $\mathrm{Cov}(X, Y) = 0$. Jolloin summan varianssi on varianssien summa (eli toimii samoin kuin odotusarvo).
- Edellisestä päättelystä seuraa että

$$\begin{aligned} \mathrm{Var}(X - Y) &= \mathrm{Var}(X) + \mathrm{Var}(-Y) + 2 \times \mathrm{Cov}(X, -Y) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) - 2 \times \mathrm{Cov}(X, Y) \quad (62) \end{aligned}$$

- Tästä aiheesta lisää tilastotieteen peruskurssilla.

Muunnoksen odotusarvo

- Usein haluamme tehdä satunnaismuuttujalle X muunnoksen ja laskea sitten muunnetun satunnaismuuttujan odotusarvon, kuten $\mathrm{E}[X^2]$.
- Muunnos voi olla muunmuuassa cX, X^2, $aX + b$ ja niin edelleen. Merkitään yleisesti tällaista muunnosta funktiolla $g(x)$

Jatkuvan satunnaismuuttujan muunnos

- Jatkuvan satunnaismuuttujan X muunnoksen $g(X)$ odotusarvo lasketaan

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx \tag{63}$$

- Esimerkiksi $g(X) = X^2$:

$$\mathrm{E}[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx \tag{64}$$

Diskreetin satunnaismuuttujan muunnos

- Diskreetin satunnaismuuttujan X muunnoksen $g(X)$ odotusarvo lasketaan

$$\mathrm{E}[g(X)] = \sum_{i=1}^{k} g(x) f(x) \tag{65}$$

- Esimerkiksi $g(X) = X^2$:

$$\mathrm{E}[X^2] = \sum_{i=1}^{k} x^2 f(x) \tag{66}$$

Esimerkki 6.8

Laske satunnaismuuttujan X^2 odotusarvo, kun X on "kahden nopan heitossa saatu silmälukujen summa". Diskreetille satunnaismuuttujalle käytetään suoraan määritelmää ja R-koodilla saatiin lasketuksi $\mathrm{E}[X^2] = 54.83$

Esimerkki 6.8 (jatkuu)

```
# Diskreetti satunnaismuuttuja
x <- seq(2, 12)
f <- c(1,2,3,4,5,6,5,4,3,2,1)/36
sum(f) # tarkista että todennäköisyydet summautuu ykköseen
## [1] 1
# kahden nopan heitto
EX21 <- sum(f*x^2) # E(X^2)
EX21 # E(X^2)
## [1] 54.83333
```

Esimerkki 6.8 (jatkuu)

satunnaismuuttujan X tiheysfunktio on

$$f_X(x) = \begin{cases} \frac{1}{2} & 2 \leq x \leq 4 \\ 0 & \text{muulloin} \end{cases} \tag{67}$$

Jatkuvalle satunnaismuuttujalle approksimoidaan pyydettyä integraalia numeerisesti R:n funktiolla **integrate**. Saadaan $\mathrm{E}[X^2] = 9.333$.

Esimerkki 6.8 (jatkuu)

```
# Jatkuva tasajakauma
x2fx <- function(x) x^2*fx(x)
EX22 <- integrate(x2fx,2,4)$value # E(X^2)
EX22
## [1] 9.333333
```

Esimerkki 6.9

Laske satunnaismuuttujan $(X^2 + 2X + 1)$ odotusarvo käyttäen lauseessa 6.1 annettuja odotusarvon laskusääntöjä, kun $X : n$ tiheysfunktio on

$$f_X(x) = \begin{cases} \frac{1}{2} & 2 \leq x \leq 4 \\ 0 & \text{muulloin} \end{cases} \tag{68}$$

Edellä laskettiin $\mathrm{E}[X^2]$ ja $\mathrm{E}[X]$, joten

$$\begin{aligned} \mathrm{E}[X^2 + 2X + 1)] &= \mathrm{E}[X^2] + 2 \times \mathrm{E}[X] + 1 \\ &= 9\frac{1}{3} + 2 \times 3 + 1 = 16\frac{1}{3} \end{aligned} \tag{69}$$

Esimerkki 6.10

Edellisen esimerkin satunnaismuuttujan X varianssi voidaan laskea kaavalla

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2 \tag{70}$$

joten

$$\mathrm{Var}(X) = 9\frac{1}{3} - 3^2 = \frac{1}{3} \tag{71}$$

Esimerkissä 6.6 saatiin sama tulos varianssin määritelmän avulla.

Esimerkki 6.11

Diskreetin satunnaismuuttujan odotusarvo, varianssi ja keskihajonta saadaan laskettua R:ssä seuraavasti. Määritellään ensin mahdolliset arvot vektorissa **x** ja niitä vastaavat pistetodennäköisyysfunktion arvot vektorissa **ptnf**.

```
x <- c(0, 1, 2)
ptnf <- c(1/4, 1/2, 1/4)
kert <- c(0, 1/4, 3/4, 1)
```

Esimerkki 6.11 (jatkuu)

Nyt, aiemmin esitetty kuva 6.1 saadaan tuotettua ao. koodilla.

```
plot(x, ptnf, type="h", ylim=c(0, 0.5), xlim=c(-1,3),
    xlab="x",ylab="f(x)",main="pistetodennäköisyysfunktio")
plot(stepfun(x,kert),ylim=c(0,1),xlab="x",ylab="F(x)",
    verticals=FALSE,pch=16,main="kertymäfunktio")
```

Esimerkki 6.11 (jatkuu)

Sitten lasketaan odotusarvo R-muuttujaan (objektiin) **EX**, varianssi **varX** :ään ja siitä keskihajonta **sdX** :ään.

```
EX <- sum(x*ptnf)
varX <- sum(ptnf*(x-EX)^2)
sdX <- sqrt(varX)
```

Vahva suurten lukujen laki

Lause 6.3

Vahva suurten lukujen laki. Olkoot $X_1, X_2, \ldots, X_n$ riippumattomia, samoin jakautuneita satunnaismuuttujia siten, että

$$\mathrm{E}[X_i] = \mu \tag{72}$$

ja

$$\mathrm{Var}(X_i) = \sigma^2 < \infty \tag{73}$$

kaikilla $i = 1, \ldots, n$, eli kaikilla X_i on sama odotusarvo ja varianssi. Tällöin

$$P(\bar{X}_n \underrightarrow{n \to \infty} \mu) = 1 \tag{74}$$

missä $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ n satunnaismuuttujien otoskeskiarvo. Tämä tarkoittaa sitä, että todennäköisyys on 1 sille, että suurella otoskoolla keskiarvo on hyvin lähellä odotusarvoa. Ts. kun otoskoko n kasvaa rajatta, niin otoskeskiarvo lähenee populaation odotusarvoa μ

Esimerkki 6.12

Pekan kodin vieressä on kasino. Hän on päättänyt käydä kasinolla joka päivä kahden vuoden ajan pelaamassa rulettia. Jokaisena päivänä Pekka aikoo pelata kymmenen kierrosta ja jokaisella kierroksella panostaa 10 euroa punaiselle. Yksittäisellä kierroksella punaisen todennäköisyys on $\frac{18}{37}$. Samoin mustan todennäköisyys on $\frac{18}{37}$ ja vihreän $\frac{1}{37}$. Yksittäisellä pelikierroksella Pekka voittaa 10 euroa todennäköisyydellä $\frac{18}{37}$ ja häviää $1 - \frac{18}{37}$. Näin ollen yhdellä kierroksella Pekan voiton odotusarvo on

$$10 \times \frac{18}{37} - 10 \times (1 - \frac{18}{37}) \approx -0.27\text{euroa} \qquad (75)$$

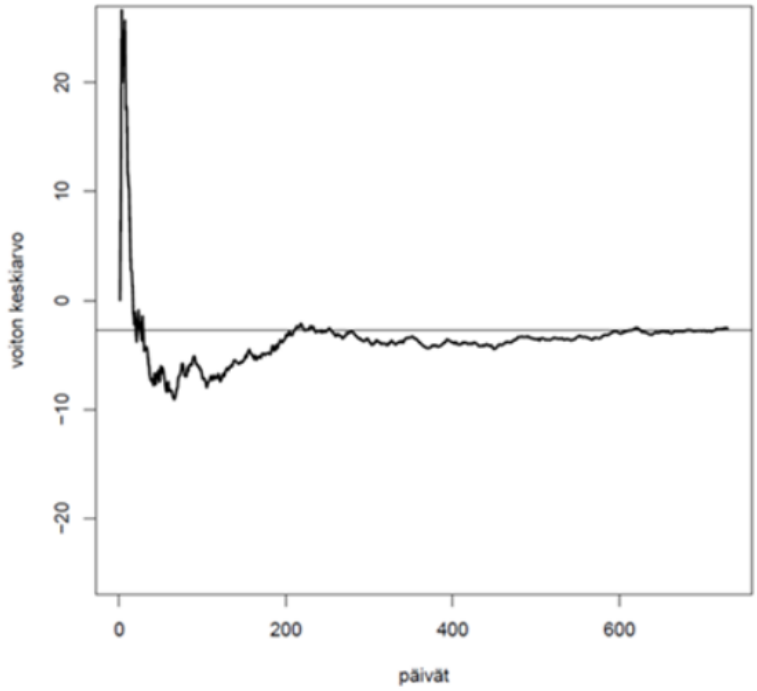
euroa, jolloin yhden päivän voiton odotusarvo on -2.7 euroa.

Esimerkki 6.12 (jatkuu)

Alla on päiväkohtaisia tunnuslukuja Pekan pelaamisesta kahden vuoden ajalta:

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
-100.000 -20.000 0.000 -2.575 20.000 100.000
```

Alla nähdään Pekan päiväkohtaisten voittojen keskiarvo muuttuu pelaamisen aikana päivästä 1 päivään 730.



Miksi normaalijakauma?

- Normaalijakauma on tilastotieteen tärkein jakauma, jonka tiheysfunktiota kutsutaan toisinaan Gaussin kellokäyräksi. Normaalijakauman keksijänä pidetään Karl Friedrich Gaussia (1777-1855).
- Normaalijakaumat sopivat kuvaamaan useita havaintoaineistojen jakaumia.
- Normaalijakautuneiden satunnaismuuttujien summan ja keskiarvon jakauma noudattaa normaalijakaumaa.
- Suurella otoskoolla riippumattomien, samoin jakautuneiden satunnaismuuttujien $X_1, X_2, \ldots, X_n$ summa $\sum_{i=1}^{n} X_i$ ja otoskeskiarvo ovat normaalijakautuneita **satunnaismuuttujan jakaumasta riippumatta!**
- Lineaarisessa regressiomallissa parametrit β_i ovat normaalijakautuneita.
- jne ...

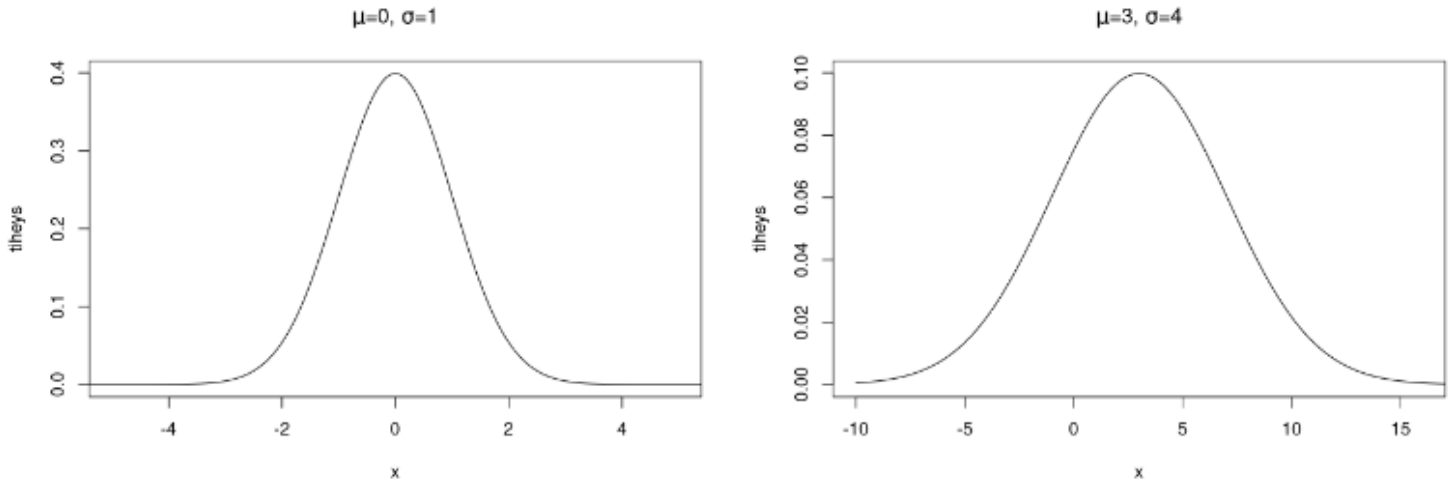
Normaalijakauman määrittely

- Normaalijakauman tiheysfunktio

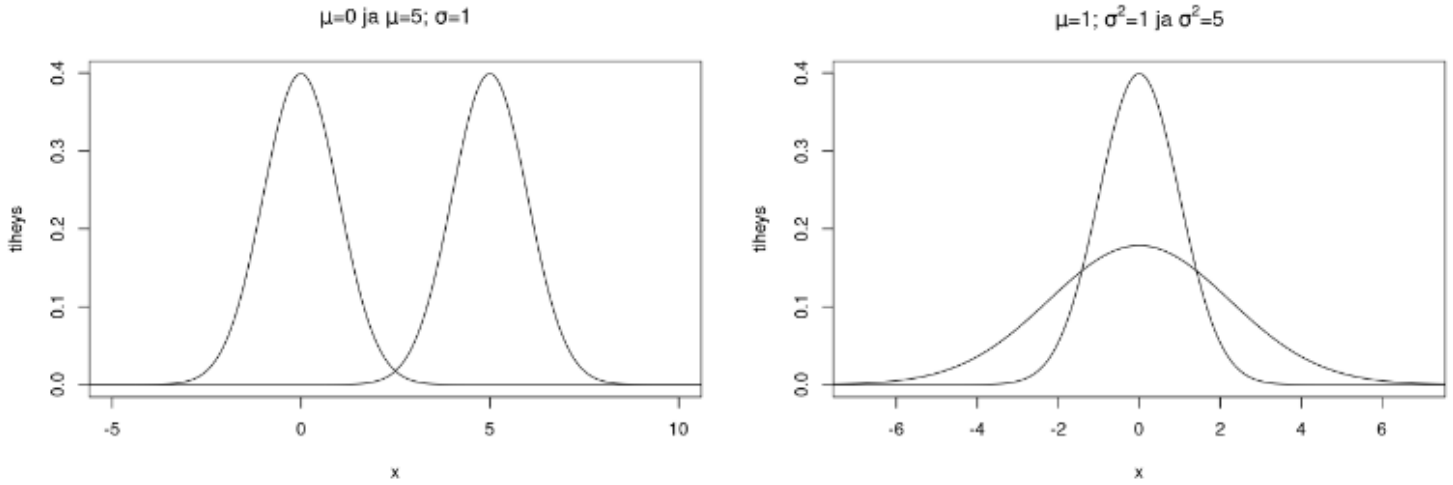
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{76}$$

- Missä e on Neperin luku (≈ 2.718), μ on jakauman odotusarvo ja σ^2 on sen varianssi. Normaalijakaumassa parametrit μ ja σ^2 määrittävät jakauman muodon täysin.
- Merkitäänkin, $X \sim N(\mu, \sigma^2)$ jos satunnaismuuttuja on jakautunut normaalijakauman mukaan parametreilla μ ja σ^2.
- Myös voidaan merkitä $\mathcal{N}(\mu, \sigma^2)$ ja vastaavasti tiheysfunktioon $f(x|\mu, \sigma^2)$.
- Englanniksi μ on yleensä **location** ja σ^2 on **scale**.

Esimerkkejä normaalijakaumasta eri parametrien arvoilla



Kuva 6.5: Normaalijakauman tiheysfunktion kuvaajia. Vasemmalla $\mu = 0$ ja $\sigma = 1$ (standardinormaalijakauma), oikealla $\mu = 3$ ja $\sigma = 4$.



Kuva 6.6: Normaalijakauman odotusarvon (vasen kuva) ja keskihajonnan (oikea kuva) vaikutus tiheysfunktion kuvaajaan.

Normaalijakauman kertymäfunktio

- Normaalijakauman kertymäfunktio, parametreilla μ ja σ^2

$$F(x|\mu,\sigma^2) = \int_{-\infty}^{x} f(u|\mu,\sigma^2)du \tag{77}$$

- Jos kyseessä on $N(0,1)$ normaalijakauma (eli ns. standardi normaalijakauma), niin kertymäfunktiota merkitään

$$\Phi(x) = \int_{-\infty}^{x} f(u|0,1)du \tag{78}$$

- Valitettavasti normaalijakauman kertymäfunktiolle ei olemassa suljetunmuodon ratkaisua. Tämän vuoksi laskuissa joudumme turvautumaan joko taulukkoon tai ohjelmistoon (**dnorm** ja **pnorm** funktiot R:ssä).

Standardinormaalijakaumasta

- Oletetaan $X \sim N(\mu, \sigma^2)$, tällöin muunnetulla satunnaismuuttujalla

$$Z = \frac{X - \mu}{\sigma} \tag{79}$$

  on jakautunut $N(0, 1)$ mukaan.
- Palauttaminen takaisin tapahtuu, kertomalla puolittain σ:lla ja lisäämällä μ.
- Lukua kutsutaan **standardoitu arvo** ja myös usein kutsutaan z -pistemääräksi (z-score).
- Kuvaa havainnon x poikkeamaa keskiarvosta μ käyttäen mittayksikkönä hajontaa σ.

Esimerkki 6.20

Opiskelija osallistui kahteen tenttiin, joista ensimmäisestä sai 30 pistettä ja toisesta 32 pistettä. Oletetaan ensimmäisen tentin pistemäärän olevan normaalijakautunut odotusarvolla 21 ja keskihajonnalla 6 ja toisen tentin pistemäärän olevan normaalijakautunut odotusarvolla 22 ja 8. Kummassa tentissä opiskelija pärjäsi suhteellisesti paremmin?
Tenttitulokset standardoituna ovat

- Tentti 1: $z_1 = \frac{30-21}{6} = 1.5$
- Tentti 2: $z_2 = \frac{32-22}{8} = 1.25$

Opiskelija menestyi suhteellisesti paremmin ensimmäisessä tentissä.

"68-95-99.7 -sääntö

Kun normaalijakauman parametrien μ ja σ^2 arvoja muutetaan, saadaan eri jakaumia. Näille kaikille pätee kuitenkin ns. "68-95-99.7" -sääntö, jonka mukaan $N(\mu, \sigma^2)$ jakaumasta peräisin olevista havainnoista

- noin 68% on välillä $(\mu - \sigma, \mu + \sigma)$ eli havainnoista noin 68% on enintään yhden keskihajonnan σ päässä odotusarvosta μ.
- noin 95% on välillä $(\mu - 2 \times \sigma, \mu + 2 \times \sigma)$ eli havainnoista noin 95% on enintään kahden keskihajonnan σ päässä odotusarvosta μ.
- ja noin 99.7% välillä $(\mu - 2 \times \sigma, \mu + 2 \times \sigma)$ eli havainnoista noin 99.7% on enintään kolmen keskihajonnan σ päässä odotusarvosta μ.

Huom! "68 - 95 - 99.7" -sääntö pätee yleisesti vain, jos muuttuja on normaalijakautunut.

Bernoullijakauma

- Kurssilla ollaan paljon puhuttu yksittäisen kolikon heitosta. Jos kyseessä on **harhaton** kolikko, niin silloin kruunan todennäköisyys $p = 0.5$.
- Samalla periaatteella voimme mallintaa mitä tahansa **binääristä** satunnaisilmiötä, esim. (pääsenkö läpi kurssista, sataako huomenna, jne). Tärkeintä on huomata että ilmiöllä täytyy olla ainoastaan kaksi tulosvaihtoehtoa.
- Tätä jakaumaa kutsumme Bernoulli jakaumaksi ja merkitään $X \sim \mathrm{Bern}(p)$, missä p on toisen (esim kruuna) tapahtuman todennäköisyys.
- Pistetodennäköisyysfunktio on

$$f(x|p) = p^x(1-p)^{1-x}, \tag{80}$$

missä x voi saada joko arvon 0 tai 1.

Binomijakauma

Binomijakauman avulla voidaan mallintaa satunnaisilmiötä, jos kaikki seuraavista neljästä ehdosta toteutuvat:

1. Satunnaiskoetta toistetaan n kertaa.
2. Satunnaiskokeet ovat toisistaan riippumattomia.
3. Jokaisessa satunnaiskokeessa on täsmälleen kaksi tulosmahdollisuutta (esim. kyllä/ei, oikein/väärin tai tapahtuu/ei tapahdu eli voidaan ajatella, että koe joko "onnistui" tai "epäonnistui").
4. Tulosvaihtoehtoihin liittyvät todennäköisyydet ovat jokaisessa kokeessa samat eli jokaisessa satunnaiskokeessa "onnistumisen" todennäköisyys on p ja "epäonnistumisen" todennäköisyys $1 - p$.

Huomaa yhteys Bernoullijakaumaan!

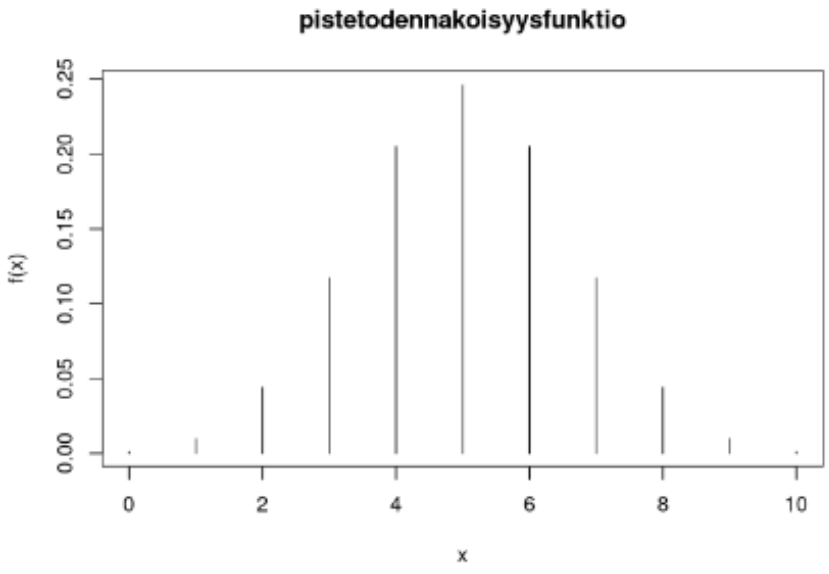
Binomijakauman pistetodennäköisyysfunktio

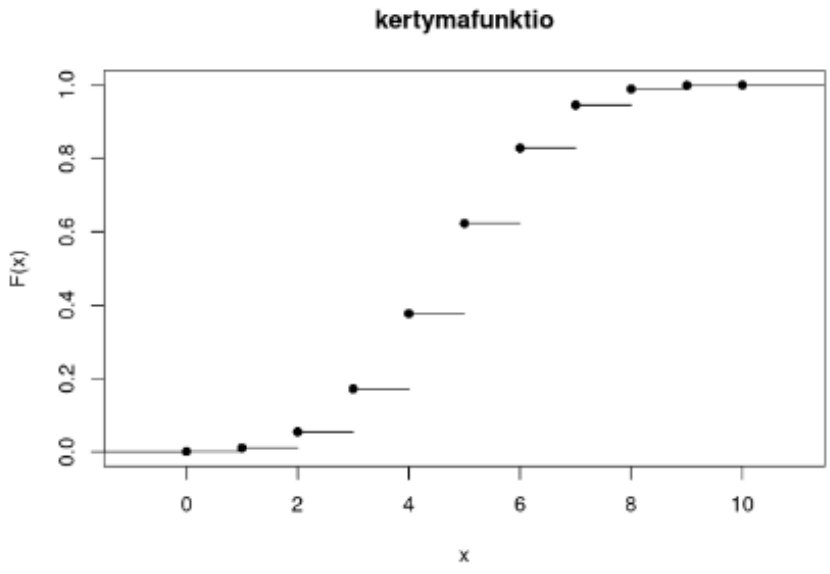
- Kun edelliset neljä ehtoa toteutuvat, onnistumisten lukumäärän X jakauma noudattaa binomijakaumaa parametrein n ja p
- n on satunnaiskokeen toistojen lukumäärä.
- p on onnistumisen todennäköisyys.
- Tälle käytetään merkintää $X \sim \mathrm{Bin}(n, p)$
- X saa siten arvoja (onnistumisten lukumääriä): $0, 1, 2, \ldots, n$
- Todennäköisyys (pistetodennäköisyysfunktio) on

$$P(X = k) = f(k|n, p) = \binom{n}{k} p^k (1 - p)^{n-k} \qquad (81)$$

Esimerkki 6.14

kuvassa on esitetty $\mathrm{Bin}(10, 0.05)$ jakauman pistetodennäköisyysfunktio ja sitä vastaava kertymäfunktio.





Esimerkki 6.14 (jatkuu)

Kuvassa esitetyt pistetodennäköisyysfunktio ja kertymäfunktio on toteutettu seuraavalla R:n koodilla:

```
x <- seq(0,10,1)
p <- dbinom(x,10,0.5)
kert <- pbinom(x,10,0.5)

plot(x,p,type="h", xlab="x",ylab="f(x)",
main="pistetodennakoisyysfunktio")
plot(stepfun(x,c(0,kert)),ylim=c(0,1),xlab="x",ylab="F(x)",
verticals=FALSE,pch=16,main="kertymafunktio")
```

Esimerkki 6.15

Heitetään kolikkoa 10 kertaa ja lasketaan kuinka monta kruunaa saadaan. Tämä vastaa binomikoetta, jossa toistojen lukumäärä $n = 10$. Tulosvaihtoehdot ovat kruuna ja klaava, joista tulos 'kruuna' tulkitaan onnistumiseksi ja tulos 'klaava', epäonnistumiseksi (kruuna valittiin vastaamaan kokeen onnistumista, koska ollaan ensisijassa kiinnostuneita näiden lukumäärästä). Oletetaan heittojen olevan toisistaan riippumattomia ja oletetaan kolikon olevan harhaton eli kruuna ja klaava ovat yhtä todennäköiset $P(\text{"kruuna"}) = P(\text{"klaava"}) = \frac{1}{2}$. Merkitään satunnaismuuttujalla X kruunujen lukumäärä kymmenessä heitossa. tällöin X noudattaa binomijakaumaa $\mathrm{Bin}(10, 1/2)$

Esimerkki 6.15 (jatkuu)

Nyt tapahtuman ‘saadaan 3 kruunaa’ todennäköisyys voidaan laskea lausekkeella

$$\begin{aligned} P(\text{"saadaan 3 kruunaa"}) &= P(X = 3) \\ &= \binom{10}{3}(1/2)^3(1-(1/2))^{10-3} \\ &= 0.117 \end{aligned} \tag{82}$$

Esimerkki 6.15 (jatkuu)

Vastaavasti tapahtuman ‘saadaan korkeintaan 1 kruuna’ todennäköisyys on

$$
\begin{aligned}
P(\text{"saadaan kork. 1 kruuna"}) &= P(X \leq 1) = P(X = 0) + P(X = 1) \\
&= \binom{10}{0} (1/2)^0 (1 - (1/2))^{10-0} \\
&+ \binom{10}{1} (1/2)^1 (1 - (1/2))^{10-1} \\
&= 0.011 \qquad (83)
\end{aligned}
$$

Esimerkki 6.15 (jatkuu)

Edellä esitetyt laskut voidaan helposti laskea R:llä

```
x <- seq(0,10,1)
p <- dbinom(x,10,0.5)
kert <- pbinom(x,10,0.5)
# todennakoisyys P(X=3) käsin ja funktiolla dbinom
choose(10,3)*0.5^3*(1-0.5)^(10-3)
## [1] 0.1171875
dbinom(3,10,0.5)
## [1] 0.1171875
# todennakoisyys P(X <= 1) käsin ja funktiolla pbinom
choose(10,0)*0.5^0*(1-0.5)^(10-0)+choose(10,1)*0.5^1*(1-0.5)^(10-1)
## [1] 0.01074219
pbinom(1,10,0.5)
## [1] 0.01074219
```

Riippumaton ja samoinjakautunut (iid)

- Usein oletetaan että meillä on satunnaismuuttujat $X_1, \ldots, X_n$ jotka ovat riippumattomia ja samoinjakautuneita.
- Satunnaismuuttujat $X_1, \ldots, X_n$ mallintavat aineiston yhden sarakkeen havaintoja.
- Oletetaan siis että aineiston muuttuja noudattaa ko. jakaumaa.
- Riippumattomuus tarkoittaa, että jos jakauman parametrit (esim. μ ja σ^2) tunnetaan niin X_i:n tunteminen ei anna lisätietoa X_j:n todennäköisyysjakaumasta, kunhan $i \neq j$

Luku 7. Otantajakauma

Parametri ja tunnusluku

- Käsitteellä **parametri** tarkoitetaan populaatiota kuvaavaa lukua, jonka arvoa ei yleensä tunneta. *Esimerkiksi populaation tai populaatiota kuvaavan mallin odotusarvo ja varianssi ovat parametreja.*
- Käsitteellä **tunnusluku** tarkoitetaan muuttujaa, jonka arvo voidaan laskea otoksesta. *Esimerkiksi otoskeskiarvo ja otosvarianssi ovat tunnuslukuja.*
- Parametrin arvoa estimoidaan usein tunnusluvulla. Esimerkiksi populaation odotusarvoa voidaan estimoida otoskeskiarvolla.
- On huomattava, että tunnusluku on satunnaismuuttuja, ja **tunnusluvun havaittu** arvo voidaan laskea otoksesta kun otos on poimittu.

Parametrit eivät ole satunnaislukuja

- Parametrit ovat kiinteitä lukuja, mutta tunnusluvut ovat satunnaismuuttujia.
- Tunnusluvut ovat satunnaismuuttujia, koska niiden arvot vaihtelevat otoksesta toiseen.
- Tunnuslukujen jakaumat on pystyttävä arvioimaan, kun otoksen avulla halutaan tehdä päätelmiä populaation parametrien arvoista.
- **Bayesilaisessa tilastotieteessä** myös parametrit ajatellaan satunnaismuuttujiksi, mutta tällä kurssilla ne ajatellaan kiinteiksi luvuiksi.

Tunnusluvun otantajakauma

- **Tunnusluvun otantajakauma** (sampling distribution) on tunnusluvun arvojen jakauma, kun tunnusluku lasketaan kaikista mahdollisista samasta populaatiosta samalla otantamenetelmällä poimituista samankokoisista otoksista.
- Jos aineistosta otetaan jollain satunnaismekanismilla, kuten yksinkertaisella satunnaisotoksella kaksi samansuuruista otosta, niin otoksiin ei yleensä saada täsmälleen samoja tilastoyksiköitä.
- Tätä otosten välillä esiintyvää vaihtelua kutsutaan **satunnaisvaihteluksi** tai **otosvaihteluksi**.
- Päättelyyn käytetään usein menetelmiä, jotka pohjautuvat tunnusluvun otantajakaumaan. Tunnuslukua sanotaan **harhattomaksi** (unbiased), jos sen otantajakauman odotusarvo on sama kuin estimoitava parametri.

Esimerkki 7.1

Laatikossa on maalattuja puupalloja yhteensä 260 kpl seuraavasti: punaisia 150 kpl, sinisiä 75 kpl, keltaisia 25 kpl, mustia 10 kpl. Ajatellaan että laatikko on populaatio, ja yksi sitä kuvaava parametri on sinisten pallojen osuus populaatiossa (merkitään p), jonka arvo on $p = 75/260 = 0.288$. Populaatiosta poimitaan 20 pallon otoksia niin, että aina uuden otoksen poimintaa varten aiemmin poimittu otos palautetaan laatikkoon. Kuvataan otoksia tunnusluvulla $\hat{p}$ "sinisten pallojen osuus". Ensimmäisessä otoksessa $\hat{p} = 7/20 = 0.35$, toisessa $\hat{p} = 9/20 = 0.45$, kolmannessa $\hat{p} = 1/20 = 0.05$ ja neljännessä $\hat{p} = 8/20 = 0.20$. TET-harjoittelija poimi kaiken kaikkiaan 30 otosta (Taulukko 7.1).

```
sin <- c(8,1,9,7,6,11,7,7,7,6,4,3,5,4,8,
         6,5,6,5,6,7,4,4,6,5,8,3,8,7,5)/20
hist(sin,xlab="Sinisten pallojen osuus")
```

Histogrammi kuvaa saatua sinisten pallojen osuuden jakaumaa. Se on sinisten pallojen osuuden empiirinen otantajakauma, kun otantamenetelmänä on 20 pallon satunnaisotanta palauttamatta.

Esimerkki 7.1 (jatkuu)

Otosten suhteellisten osuuksien keskiarvo on 0.297 ja otoskeskiarvojen keskihajonta on 0.10. Otosten keskiarvo poikkeaa populaation todellisesta sinisten pallojen osuudesta siksi, että otoksia on vain 30 kappaletta. Jos otoksia poimittaisiin lisää, otosjakauman keskiarvo lähestyisi arvoa $p = 75/260$, sillä voidaan osoittaa, että otoksen suhteellinen osuus on populaatio-osuuden harhaton estimaattori.

```
mean(sin)
## [1] 0.2966667
sd(sin)
## [1] 0.101653
```

Käytännössä otoksia on kuitenkin yleensä vain yksi

Käytännön tutkimuksessa käytettävissä on yleensä vain yksi otos kiinnostuksen kohteena olevasta populaatiosta. Siksi on ajateltava, että tunnusluvun arvot vaihtelisivat otosten välillä jos poimittaisiin useampia otoksia. Tämä ei muuta ilmiön luonnetta, mutta vaikeuttaa sen analysointia, sillä tunnusluvun otantajakauma pitää pystyä estimoimaan pelkästään yhden otoksen perusteella.

Esimerkki 7.2

Kannatuksen tutkimista varten haastatellaan 1000 satunnaisesti poimittua äänioikeutettua suomalaista. Heistä 370 ilmoitti äänestävänsä ehdokasta A. Otoksesta laskettu ehdokkaan A kannattajien suhteellinen osuus on

$$\hat{p} = \frac{370}{1000} = 0.37 \tag{84}$$

Esimerkissä ehdokkaan A kannattajien osuus populaatiossa on parametri p, jota ei tunneta. Otoksesta laskettu suhteellinen osuus $\hat{p}$ on tunnusluku, jonka avulla arvioidaan (estimoidaan) parametria p. Otokseen liittyy satunnaisvaihtelua ja jos poimittaisiin toinen 1000 hengen otos ja laskettaisiin siitä uusi estimaatti $\hat{p}$ suhteelliselle osuudelle p, niin saataisiin todennäköisesti eri arvo kuin 0.37.

Esimerkki 7.2 (jatkuu)

Jos 1000 hengen otoksia poimittaisiin suuri määrä ja laskettaisiin jokaisesta ehdokkaan A kannattajien suhteellinen osuus, niin voitaisiin muodostaa $\hat{p}$:n otantajakauma. Käytännössä emme kuitenkaan voi poimia suurta määrää otoksia Suomen äänioikeutetuista. Ilmiötä voidaan kuitenkin havainnollistaa simuloimalla, jolloin saadaan määritettyä $\hat{p}$:n otantajakauma. Alla olevalla R-ohjelmalla on simuloitu suhteellisen osuuden otantajakauma olettaen, että ehdokkaan A kannatus on 30% . Simuloinnissa poimittiin 1000 havainnon otoksia (gallupkyselyitä) 500 kappaletta. Simuloiduista otoksista laskettujen populaatio-osuuksien keskiarvo on 0.2999 ja keskihajonta 0.014. Populaatio-osuuksien jakauma on kuvan perusteella likimain normaalijakautunut. Jos emme tuntisi todellista populaatio-osuutta, voisimme normaalijakauman ominaisuuksien perusteella päätellä, että populaatio-osuus on hyvin todennäköisesti välillä 0.299 ± 20.014 eli välillä (0.271,0.327) . Päättely seuraa siitä, että normaalijakaumassa n. 95 % havainnoista on välillä $\mu \pm 2 \times \sigma$

Esimerkki 7.2 (jatkuu)

```
# Alustetaan satunnaislukugeneraattori, jolloin aina seuraavat
# koodit ajettaessa saadaan  samat otokset.
set.seed(1234)
p <- 0.3
n <- 1000
nrep <- 500
osuudet <- rep(NA,nrep)
for (i in 1:nrep) {
    osuudet[i] <- sum(runif(n)<p)/n
}
mean(osuudet)
## [1] 0.299932
sd(osuudet)
## [1] 0.0145822
hist(osuudet,xlab="Suhteellinen osuus",
main="G. otantajak. kun p=0.3",freq=FALSE,ylim=c(0,30))
x0 <- seq(0,1,0.001)
lines(x0,dnorm(x0,p,sqrt(p*(1-p)/n)))
```

Suhteellisen osuuden otantajakaumasta

- Täsmällisemmin ilmaistuna keskeisestä raja-arvolauseesta seuraa, että kokoa n olevasta otoksesta lasketun suhteellisen osuuden $\hat{p}$ otantajakauma on likimain normaalijakautunut parametrein p ja $p(1-p)/n$, missä p on populaatio-osuus

$$\hat{p} \sim N(p, p(1-p)/n) \tag{85}$$

- Otoksesta laskettu $\hat{p}$ on populaatio-osuuden harhaton estimaattori.
- Kun otoskoko n kasvaa, niin $\hat{p}$:n hajonta pienenee eli "tarkkuus" paranee. Sanotaan, että $\hat{p}$ on **tarkentuva** (konsistentti, consistent), koska hajonta lähenee nollaa, kun otoskoko kasvaa.
- Normaalijakauma on käyttökelpoinen, kun $np \geq 10$ ja $n(1-p) \geq 10$.