Huomautus merkinnöistä. Eri satunnaismuuttujien tiheysfunktiot erotetaan usein toisistaan alaindeksillä, esim. f_X ja f_Y. Alaindeksillä merkitään myös ehdollista tiheysfunktiota, esim. $f_{X|Y}$.

Bayes-päättelyssä käytetään usein lyhennysmerkintää, jossa kaikkia *eri* tiheysfunktioita merkitään vain f:llä ilman alaindeksiä. Tällöin sulkujen sisällä olevasta argumentista on pääteltävä, mistä funktiosta on kysymys: esim. $f(x)$ tarkoittaa samaa kuin $f_X(x)$. Lyhennysmerkintää ei pidä käyttää jos se aiheuttaa sekaannusta, esim. $f(5)$ ei kerro onko kyse X:n vai Θ:n tiheysfunktiosta. Vrt. luento 5A.

(Luentomonisteessa käytetään f:ää datan tiheysfunktioille ja p:tä parametrin tiheysfunktioille, mutta tämä merkintä ei ole kovin yleisesti käytössä.)

5B Bayesläinen päättely

Tuntitehtävät

5B1 (Kadonnut lentokone) Seuraavantyyppistä tilastollista hakumenetelmää on käytetty mm. yritettäessä paikantaa 8.3.2014 kadonnutta Malaysia Airlinesin lentoa MH370. Etsintöjen kohteena oleva alue on jaettu neljään suureen ruutuun. Lennon taustatietojen pohjalta arvellaan lentokoneen sijaitsevan ruudussa 1 tn:llä 50%, ruudussa 2 tn:llä 30%, ruudussa 3 tn:llä 10% ja ruudussa 4 tn:llä 10%. Olosuhteista johtuen arvellaan, että yksittäiseen ruutuun kohdistettu etsintäyritys epäonnistuu löytämään ruudussa sijaitsevan koneen tn:llä 70%, aiemmista etsintäyrityksistä riippumatta.

(a) Tulkitaan lentokoneen sijainti satunnaismuuttujaksi (tuntemattomaksi parametriksi) $\Theta \in \{1, 2, 3, 4\}$. Esitä sijainnin priorijakauma.

(b) Etsintäryhmä tutkii ensiksi ruudun 1. Merkitään $X = 1$ jos kone löytyy ja $X = 0$ muuten. Määritä Θ:n uskottavuusfunktio havainnolle $X = 0$.

(c) Konetta ei löydetä ensimmäisellä hakuyrityksellä. Määritä koneen sijainnin posteriorijakauma tämän havainnon suhteen. Kannattaako ruudusta 1 etsiä uudelleen?

(d) Etsintäryhmä päättää tutkia ruudun 1 uudelleen. Merkitään tämän toisen haun tulosta satunnaismuuttujalla Y. Tämänkin haun tulos on, että konetta ei löytynyt. Määritä lentokoneen sijainnin posteriorijakauma tämän datan valossa. Mistä etsintäryhmän kannattaa seuraavaksi etsiä?

Ratkaisu.

(a) Lentokoneen sijainnin Θ priorijakauma on

θ	1	2	3	4
$f(\theta)$	0.5	0.3	0.1	0.1

(b) Uskottavuusfunktio datapisteelle $x = 0$ on $f(x \,|\, \theta) = \mathrm{P}(X = 0 \,|\, \Theta = \theta)$. Jos lentokone sijaitsee ruudussa 1, epäonnistuu ruutuun 1 kohdistettu haku tn:llä 70%. Muussa tapauksessa ruutuun 1 kohdistettu haku epäonnistuu varmasti. Kysytty uskottavuus on siis

θ	1	2	3	4
$f_{X\mid\Theta}(0\mid\theta)$	0.7	1	1	1

(c) Posteriorijakauma $f(\theta\mid 0)$ on (pisteittäin) priorin ja uskottavuuden tulo, normitettuna sopivalla vakiolla $c>0$, eli

$$f_{\Theta\mid X}(\theta\mid 0) = c\cdot f_\Theta(\theta)\cdot f_{X\mid\Theta}(0\mid\theta)$$

tai lyhennysmerkinnöin

$$f(\theta\mid 0) = c\cdot f(\theta)\cdot f(0\mid\theta).$$

Kerroin c on itse asiassa $1/f(X) = 1/(\sum_\theta f(\theta)f(X|\theta)$, mutta helpointa on yleensä laskea vaiheittan seuraavasti.

Lasketaan ensin normittamaton posteriorijakauma $\theta\mapsto f(\theta)f(0\mid\theta)$.

θ	1	2	3	4
$f(\theta)f(0\mid\theta)$	0.35	0.30	0.10	0.10

Koska normittamattoman tiheyden lukuarvojen summa on 0.85, jakamalla ne tällä luvulla (eli normittamalla) saadaan aito posteriorijakauma

θ	1	2	3	4
$f(\theta\mid 0)$	0.412	0.353	0.118	0.118

Koska lentokoneen sijainnin posterijakauma maksimoituu ruudussa 1, kannattaa seuraavakin haku kohdistaa ruutuun 1.

1 on tuntemattoman parametrin Θ *posteriorimoodi* (engl. *posterior mode, maximim a posteriori estimate, MAP estimate* eli se kohta, jossa posterioritiheys on suurin.

(d) **Tapa 1.** Uskottavuusfunktio datapisteelle $(x,y)=(0,0)$ on

$$f(x,y\mid\theta) \;=\; \mathrm{P}(X=x, Y=y\mid\Theta=\theta)$$

Jos lentokone sijaitsee ruudussa 1, niin todennäköisyys havaita data $(0,0)$ eli epäonnistua kaksi kertaa on $0.7\cdot 0.7 = 0.49$. Jos lentokone sijaitsee jossain muualla, niin sitä ei varmasti löydetä ruudusta 1, joten todennäköisyys havaita data $(0,0)$ on $1\cdot 1 = 1$. Kysytty uskottavuus on siis

θ	1	2	3	4
$f(0,0\mid\theta)$	0.49	1	1	1

Posteriorijakauma $f(\theta\mid 0,0)$ datapisteen $(x,y)=(0,0)$ suhteen saadaan painottamalla ja normittamalla priorijakaumaa $f(\theta)$ uskottavuudella $f(0,0\mid\theta)$ samaan tapaan kuin (c)-kohdassa. Näin menetellen saadaan posteriorijakaumaksi

θ	1	2	3	4
$f(\theta \mid 0,0)$	0.329	0.403	0.134	0.134

Kahden epäonnistuneen ruutuun 1 kohdistetun hakuyrityksen jälkeen kannattaa seuraavaksi tutkia ruutu 2, koska sen posteriori-tn on suurin.

Tapa 2. Posteriorijakauma $f(\theta \mid 0,0)$ datapisteen $(x,y) = (0,0)$ suhteen voidaan laskea myös päivittämällä (c)-kohdan posteriorijakaumaa $f(\theta \mid 0)$ uskottavuusfunktiolla $f(y \mid \theta, x) = \mathrm{P}(Y = y \mid \Theta = \theta, X = x)$. Hakujen riippumattomuuden ja samoinjakautuneisuuden nojalla voidaan havaitaan, että

$$\begin{aligned} f_{Y|\Theta,X}(0 \mid \theta, 0) &= \mathrm{P}(Y = 0 \mid \Theta = \theta, X = 0) \\ &= \mathrm{P}(Y = 0 \mid \Theta = \theta) \\ &= \mathrm{P}(X = 0 \mid \Theta = \theta) \\ &= f(0 \mid \theta). \end{aligned}$$

Kysytty posteriorijakauma voidaan siis laskea päivittämällä priorijakaumaa 2 kertaa peräkkäin uskottavuusfunktiolla $f(0 \mid \theta)$.

5B2 (Metron vuoroväli.) Metro kulkee säännöllisesti θ minuutin välein. Tuomas saapuu metrolaiturille joka päivä satunnaiseen aikaan, joten i:ntenä päivänä metron odotusajalla X_i on tasajakauma välillä $[0, \theta]$, tiheydellä

$$f(x_i \mid \theta) = \begin{cases} \frac{1}{\theta}, & 0 \le x_i \le \theta, \\ 0, & \text{muuten.} \end{cases}$$

Koska Tuomas ei tunne θ:n arvoa, hän käsittelee sitä satunnaismuuttujana Θ. Yleistietonsa nojalla hän olettaa, että vuoroväli on ainakin 1 minuutti. Hän valitsee prioritiheydeksi

$$f_\Theta(\theta) = \begin{cases} 0.2\theta^{-1.2}, & \theta \ge 1, \\ 0, & \text{muuten.} \end{cases}$$

Viitenä päivänä Tuomas havaitsee odotusajat $\vec{x} = (7, 3, 2, 9, 6)$.

- Piirrä Tuomaksen prioritiheys välillä $[0, 20]$ käsin tai tietokoneella. Tarkista että kyseessä on todella tiheysfunktio, laskemalla sen integraali koko mahdollisella välillä $[1, \infty)$. Vihje: Vaikka eksponentti ei ole kokonaisluku, voit integroida funktion kuten polynomifunktion. Piirtämiseen riittää luultavasti funktion arvon laskeminen muutamassa pisteessä, jonka jälkeen näet funktion muodon suunnilleen.
- Laske posteriorijakauman tiheysfunktio (johda sen lauseke) ja piirrä se välillä $[0, 20]$. (Muista: priori kertaa uskottavuus, ja lopuksi normalisointi. Viiden havainnon uskottavuus on yksittäisten havaintojen uskottavuuksien tulo.)

- Tuomas päättää käyttää Θ:n *posteriorijakauman odotusarvoa* (huom. ei moodia) vuoroväli n Θ piste-estimaattina. Laske kyseinen odotusarvo. (Tunnet posteriorijakauman edellisen kohdan perusteella, laske odotusarvo normaaliin tapaan.)
- Käyttäen posteriorijakaumaa, laske todennäköisyys (havaintojen perusteella) että $\Theta < 15$. Ilmaise sanallisesti, mitä tämä tarkoittaa.

Ratkaisu.

(a) Kuva alempana. Integraali välillä $[1, \infty)$ on

$$\int_1^\infty 0.2\,\theta^{-1.2} d\theta \;=\; \Big/_1^\infty \left(\frac{0.2}{-0.2}\,\theta^{-0.2}\right) \;=\; 0 \;-\; \frac{0.2}{-0.2}\,1^{-0.2} \;=\; 1.$$

(b) Kun havainnot ovat $\vec{x} = (x_1, \ldots, x_5)$, niin uskottavuus on

$$f(\vec{x}\,|\,\theta) \;=\; \prod_{i=1}^{5} f(x_i\,|\,\theta) \;=\; \begin{cases} \theta^{-5}, & \theta \geq M, \\ 0, & \text{muuten}, \end{cases}$$

missä $M = \max(x_1, \ldots, x_5)$. Havainnoilla $\vec{x} = (7, 3, 2, 9, 6)$ on $M = 9$, joten Θ:n posterioritiheys on

$$f(\theta\,|\,\vec{x}) \;=\; c \cdot f(\theta) f(\vec{x}\,|\,\theta) \;=\; \begin{cases} c \cdot 0.2\,\theta^{-6.2}, & \theta \geq 9, \\ 0, & \text{muuten}, \end{cases}$$

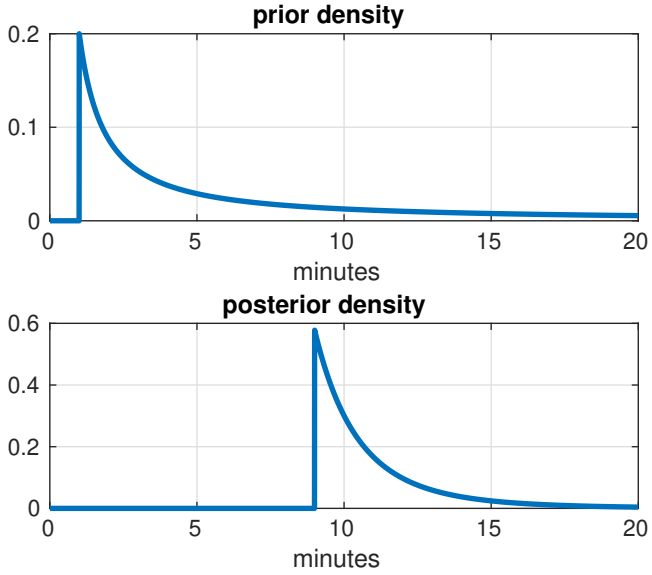
missä $c > 0$ on normalisointivakio. Voimme yksinkertaistaa laskuja merkitsemällä $c \cdot 0.2 = C$. Vakion arvo on valittava siten, että $\int f(\theta\,|\,\vec{x}) d\theta = 1$. Koska

$$\int_9^\infty \theta^{-6.2} d\theta \;=\; \Big/_9^\infty \left(\frac{1}{-5.2}\theta^{-5.2}\right) \;=\; \frac{1}{5.2 \cdot 9^{5.2}},$$

tulee valita $C = 5.2 \cdot 9^{5.2}$, joten posterioritiheys on

$$f(\theta\,|\,\vec{x}) = (5.2 \cdot 9^{5.2}) \cdot \theta^{-6.2}, \qquad \text{kun } \theta > 9$$

ja nolla muualla.



Huom. Sekä priori että posteriori ovat tässä *Pareto-jakaumia*, jollainen nähtiin jo harjoituksessa 2A1. Voidaan osoittaa yleisemminkin, että jos priori on Pareto-jakauma ja yksittäisen havainnon uskottavuus on tasajakauma, niin posteriorikin on Pareto-jakauma. Jos saadaan lisää havaintoja, posteriorin muoto pysyy edelleen Pareto-jakaumana, sen parametrit vain päivittyvät.

Yleisesti tällaista laskennallisesti mukavaa tilannetta kutsutaan *konjugaattiprioriksi*, ts. jos priorin ja uskottavuusfunktion muodot ovat tietyllä tavalla yhteensopivat (konjugaatit), niin posteriori on samaa muotoa (jakaumaperhettä) kuin priori.

(c) Käyttäen b-kohdassa laskettua posteriorijakaumaa, saadaan Θ:n (posteriori)odotusarvo tavalliseen tapaan integraalina

$$\begin{aligned}\int_{-\infty}^{\infty} \theta f(\theta \,|\, \vec{x})\, d\theta &= \int_{9}^{\infty} \theta\, C\theta^{-6.2}\, d\theta = C \int_{9}^{\infty} \theta^{-5.2}\, d\theta \\ &= C \cdot \frac{9^{-4.2}}{4.2} = \frac{5.2 \cdot 9^{5.2}}{4.2 \cdot 9^{4.2}} = \frac{5.2 \cdot 9}{4.2} \approx 11.143.\end{aligned}$$

(d) Todennäköisyys sille, että satunnaismuuttujan Θ arvo osuu eräälle välille, saadaan tavalliseen tapaan integroimalla sen tiheyttä ko. välin yli. Tässä käytetään siis Θ:n posteriorijakauman tiheyttä.

$$\begin{aligned}\mathrm{P}(\Theta < 15 \,|\, \vec{X} = \vec{x}) &= \int_{9}^{15} f(\theta \,|\, \vec{x})\, d\theta = \int_{9}^{15} (5.2 \cdot 9^{5.2}) \cdot \theta^{-6.2}\, d\theta \\ &= \frac{5.2 \cdot 9^{5.2}}{-5.2} \cdot \Big/_{9}^{15} \left(\theta^{-5.2}\right) = 9^{5.2} \cdot \left(9^{-5.2} - 15^{-5.2}\right) \approx 0.930.\end{aligned}$$

Sanallisesti: Viiden päivän havaintojensa perusteella Tuomas pitää melko todennäköisenä (tn = 93%), että metron vuoroväli on alle 15 minuuttia.

Kotitehtävät

5B3 (Lehtien pituudet) Botanisti arvelee, että tietyn kasvilajin lehtien pituus (cm) noudattaa normaalijakaumaa odotusarvona Θ (tuntematon) ja keskihajontana $\sigma = 2$. Hän olettaa lisäksi, että tuntematon odotusarvo Θ noudattaa normaalijakaumaa odotusarvona $\mu_0 = 10$ ja keskihajontana $\sigma_0 = 1$. Botanisti mittasi viiden kyseisen lajin kasvin lehtien pituudet ja sai tulokseksi $\vec{x} = (10, 14, 11, 17, 8)$ (lehtien pituudet voidaan olettaa toisistaan riippumattomiksi). Määritä tehtyjen havaintojen valossa:

(a) Satunnaismuuttujan Θ posteriorijakauman odotusarvo.

(b) Väli, joka sisältää tuntemattoman odotusarvon todellisen arvon 90% todennäköisyydellä.

Vihje. Ylläolevin oletuksin tuntemattoman odotusarvon posteriorijakauma on normaalijakauma, jonka odotusarvo ja keskihajonta löytyvät luentomonisteen luvusta 9 / luennosta 5A. Normaalijakauman kertymäfunktioon voi käyttää taulukoita tai esim. R:n funktioita `pnorm` ja `qnorm`.

Arviointiohje.
a)-kohta: +1p
b)-kohta: +1p
Pyöristysvirheistä tai pieniestä laskuvirheistä ei menetä pisteitä, jos saadut lukuarvot ovat järkeviä tehtävän kannalta.

Ratkaisu.

(a) Luentomonisteen lauseen 9.5 mukaan Θ:n posteriorijakauma on normaalijakauma odotusarvona

$$\mu_1 = \left(\frac{\frac{1}{\sigma_0^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}\right)\mu_0 + \left(\frac{\frac{n}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}\right) m(\vec{x}),$$

ja keskihajontana

$$\sigma_1 = \frac{1}{\sqrt{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}},$$

missä $n = 5$ on havaitun datajoukon $(x_1, \ldots, x_5)$ koko ja $m(\vec{x}) = \frac{1}{5}\sum_{i=1}^{5} x_i$ sen keskiarvo. Lehtien pituuksien keskiarvo on

$$m(\vec{x}) = \frac{1}{5}(10 + 14 + 11 + 17 + 8) = \frac{60}{5} = 12.$$

Sijoittamalla lukuarvot ylläolevaan kaavaan saadaan posteriorijakauman odotusarvoksi

$$\mu_1 = \left(\frac{\frac{1}{1^2}}{\frac{1}{1^2} + \frac{5}{2^2}}\right) 10 + \left(\frac{\frac{5}{2^2}}{\frac{1}{1^2} + \frac{5}{2^2}}\right) 12 = \frac{4}{9} \times 10 + \frac{5}{9} \times 12 = 11\frac{1}{9} \approx 11.11.$$

(b) Posteriorijakauman keskihajonnaksi saadaan

$$\sigma_1 \;=\; \frac{1}{\sqrt{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}} \;=\; \frac{1}{\sqrt{\frac{1}{1^2}+\frac{5}{2^2}}} \;=\; \frac{2}{3}.$$

Ehdolla havaittuun dataan $\vec{x}$, normitettu satunnaismuuttuja $Z = \frac{\Theta-\mu_1}{\sigma_1}$ noudattaa normitettua normaalijakaumaa. Määritetään ensin arvo $z > 0$, jolle pätee $\mathrm{P}(-z \le Z \le z \,|\, x) = 0.9$. Symmetrian perusteella normitetun normaalijakauman kertymäfunktio $\Phi(z)$ toteuttaa $\Phi(z) = 1 - \Phi(-z)$, joten

$$0.9 \;=\; \mathrm{P}(-z \le Z \le z \,|\, x) \;=\; \Phi(z) - \Phi(-z) \;=\; 1 - 2\Phi(-z).$$

Tästä ratkaistaan $\Phi(-z) = \frac{1-0.9}{2} = 0.05$. Mellinin taulukoista löydetään, että tämä yhtälö ratkeaa arvolla $z \approx 1.64$. Näin siis

$$\mathrm{P}\left(-1.64 \le \frac{\Theta-\mu_1}{\sigma_1} \le 1.64 \;\middle|\; \vec{x}\right) \;=\; 0.9,$$

eli

$$\mathrm{P}\left(\mu_1 - 1.64\sigma_1 \le \Theta \le \mu_1 + 1.64\sigma_1 \,|\, \vec{x}\right) \;=\; 0.9.$$

Halutunlainen väli voidaan siis esittää muodossa

$$\Theta \;=\; \mu_1 \pm 1.64\sigma_1 \;=\; 11.11 \pm 1.09.$$

5B4 (Vaarallinen tie) Juuri avatulla tiellä kuukauden aikana sattuvien kolarien määrän oletetaan noudattavan Poisson-jakaumaa,

$$f(x \mid \theta) \;=\; e^{-\theta}\frac{\theta^x}{x!}, \quad x = 0, 1, 2, \ldots,$$

tuntemattomalla odotusarvoparametrilla $\theta > 0$. Aiemman kokemuksensa perusteella bayesläinen insinööri arvelee, että tuntemattoman parametrin Θ arvo on 1 todennäköisyydellä 0.20, 2 todennäköisyydellä 0.60, ja 3 todennäköisyydellä 0.20. Koska parametrilla Θ on tässä vain kolme mahdollista arvoa, se on diskreetti satunnaismuuttuja.

(a) Ensimmäisenä kuukautena tiellä sattuu kaksi kolaria. Määritä odotusarvon $\Theta \in \{1, 2, 3\}$ posteriorijakauma tämän havainnon valossa. Havaintona on siis *yksi* luku, nimittäin 2, joka on peräisin Poisson-jakaumasta. Aloita selvittämällä uskottavuusfunktio.

(b) Toisena kuukautena tiellä ei satu yhtään kolaria. Määritä odotusarvon $\Theta \in \{1, 2, 3\}$ posteriorijakauma molempien kuukausien havaintojen valossa (oletetaan, että eri kuukausien kolarien lukumäärät ovat toisistaan riippumattomia). Havaintona on siis *kaksi* lukua, 2 ja 0, kumpikin peräisin samasta Poisson-jakaumasta.

Arviointiohje.
a)-kohta: +0.5 p jos uskottavuusfunktio oikein
a)-kohta: +0.5 p jos posteriorijakauma oikein
b)-kohta: +0.5 p jos uskottavuusfunktio oikein
b)-kohta: +0.5 p jos posteriorijakauma oikein
Pienet pyöristysvirheet sallitaan.

Ratkaisu. Merkitään X_i = kuukautena i tiellä sattuneiden kolarien lkm, $i = 1, 2$.

(a) Esitetään päättely taulukkona kuten luentomonisteen luvussa 9.1. Taulukkoon tulee yksi rivi kullekin mahdolliselle θ:n arvolle. Uskottavuus $f_{X_1|\Theta}(2 \,|\, \theta)$ lasketaan sijoittamalla Poisson-jakauman tiheysfunktion lausekkeeseen $x = 2$ ja (kullakin rivillä eri) θ.

θ	priori $f(\theta)$	uskottavuus $f(2 \,\|\, \theta)$	tulo	posteriori $f(\theta \,\|\, x_1)$
1	0.20	0.1834	0.0368	0.1508
2	0.60	0.2707	0.1624	0.6656
3	0.20	0.2240	0.0448	0.1836
$\sum$	1.00		0.2440	1.0000

Poisson-jakauman tiheysfunktion voi laskea käsin, laskimella, R:ssä komennolla `dpois`, tai Matlabissa/Octavessa komennolla `poisspdf`.

(b) Priori on sama kuin aiemmin. Uskottavuus eli tn havainnolle $(x_1, x_2) = (2, 0)$, kun parametrilla on arvo θ, on

$$f(2, 0 \mid \theta) = \left(\frac{\theta^2}{2!} e^{-\theta}\right) \left(\frac{\theta^0}{0!} e^{-\theta}\right) = \frac{\theta^2}{2} e^{-2\theta},$$

joka siis lasketaan kullakin rivillä kyseisellä θ:n arvolla.

θ	priori $f(\theta)$	uskottavuus $f(2, 0 \,\|\, \theta)$	tulo	posteriori $f(\theta \,\|\, 2, 0)$
1	0.20	0.0677	0.0135	0.3586
2	0.60	0.0366	0.0298	0.5823
3	0.20	0.0112	0.0022	0.0591
$\sum$	1.00		0.0378	1.0000