3B Normaaliapproksimaatio

Huom. Standardinormaalijakauman kertymäfunktion $F_Z(z) = \Phi(z)$ voi laskea esim. R:ssä komennolla `pnorm` tai Matlabissa/Octavessa komennolla `normcdf`. Jos tietokonetta tai tämän funktion osaavaa laskinta ei ole käytettävissä, voi käyttää esim. kurssisivulla kohdassa Materiaalit annettuja taulukoita. Mellinin taulukoissa on Φ annettu 0.01:n välein arvoille $-3.59 \ldots 3.59$.

Joissakin taulukoissa annetaan vain positiivinen puoli, jolloin negatiivisella puolella täytyy hyödyntää normaalijakauman symmetriaa $\Phi(-z) = 1 - \Phi(z)$.

Tuntitehtävät

3B1 (Teiden suolaus) Tieverkon sulana pitämiseen on varastoitu suolaa 200 cm lumimäärän varalle. Yksittäisen talvipäivän aikana lunta sataa keskimäärin 4.5 cm, keskihajonnan ollessa 2.5 cm.

(a) Laske normaaliapproksimaation avulla arvio todennäköisyydelle, että varastoitu suola riittää 50 talvipäiväksi.

(b) Mitä lisäoletuksia oletuksia sinun piti tehdä (a)-kohdan ratkaisemisen yhteydessä? Ovatko kyseiset oletukset mielestäsi perusteltuja?

Ratkaisu.

(a) Merkitään 50 talvipäivän lumikertymää (cm) satunnaismuuttujalla $S = X_1 + \cdots X_{50}$, missä X_i on i:nnen talvipäivän aikana satanut lumimäärä (cm). Oletusten mukaan $\mathrm{E}(X_i) = 4.5$ ja $\mathrm{SD}(X_i) = 2.5$. Tällöin 50 talvipäivän lumikertymän odotusarvo on

$$\mathrm{E}(S) = 50 \times 4.5 \;=\; 225.$$

Lumikertymän keskihajontaa *ei* voi määrittää ilman lisätietoja talvipäivien lumimäärien keskinäisistä riippuvuuksista. Näin ei myöskään normaaliapproksimaatiota voi käyttää ilman lisäoletuksia. Tehdään siis seuraava lisäoletus.

Oletus. Satunnaismuuttujat $X_1, X_2, \ldots$ ovat keskenään riippumattomat.

Lisäoletuksen vallitessa 50 talvipäivän lumikertymän keskihajonnaksi saadaan

$$\begin{aligned} \mathrm{SD}(S) &= \sqrt{\mathrm{SD}(X_1)^2 + \cdots \mathrm{SD}(X_{50})^2} \\ &= \sqrt{50 \times 2.5^2} \\ &= \sqrt{50} \times 2.5 \;\approx\; 17.68. \end{aligned}$$

Tällöin normaaliapproksimaation mukaan riipumattomien satunnaismuuttujien summa S noudattaa likimain normaalijakaumaa odotusarvona $\mathrm{E}(S) = 225$ ja keskihajontana $\mathrm{SD}(S) \approx 17.68$. Näin ollen suola riittää 50 talvipäiväksi likimain todennäköisyydellä

$$\mathrm{P}(S \le 200) \;=\; \mathrm{P}\left(\frac{S - \mathrm{E}(S)}{\mathrm{SD}(S)} \le \frac{200 - 225}{17.68} \right) \;\approx\; \mathrm{P}(Z \le -1.41),$$

missä Z noudattaa normitettua normaalijakaumaa. Normitetun normaalijakauman taulukoista (tai R-komennolla `pnorm(-1.41)`) todetaan, että kysytty todennäköisyys on likimain $\mathrm{P}(S \leq 200) \approx 7.9\%$.

(b) Lisäoletuksena vaadittiin, että yksittäisten talvipäivien lumimäärät ovat keskenään stokastisesti riippumattomat. Tämä lisäoletus ei välttämättä oikein tarkasti sovi Suomen sääolosuhteiden mallintamiseen.

3B2 (Voitolla kasinolla) Kasinon rulettipelissä arvotaan jokaisella kierroksella tasaisen satunnaisesti numero joukosta $\{0, 1, \ldots, 36\}$. Harry päättää panostaa joka kierroksella euron omalle onnennumerolleen. Mikäli rulettikuula osuu onnennumeroon, Harryn saa 36 euroa, eli hänen nettotuottonsa on 35 euroa. Muussa tapauksessa hän häviää panostamansa euron eli nettotuotto on -1 euro. Merkitään Harryn nettotuottoa n pelikierroksen jälkeen satunnaismuuttujalla $S_n = X_1 + \cdots + X_n$, missä X_i on pelikierroksen i tuotto.

(a) Määritä satunnaismuuttujan X_i odotusarvo ja keskihajonta.

(b) Määritä satunnaismuuttujan S_n odotusarvo ja keskihajonta.

Laske normaaliapproksimaatiota käyttämällä arvio todennäköisyydelle, että Harry on voitolla:

(c) 30 pelikierroksen jälkeen,

(d) 3 000 pelikierroksen jälkeen,

(e) 300 000 pelikierroksen jälkeen.

Ensimmäinen ylläolevista todennäköisyyksistä voidaan laskea myös tarkasti:

(f) Laske tarkka arvo (c)-kohdan todennäköisyydelle ja vertaa sitä normaaliapproksimaation tuottamaan arvioon.
Vihje: Mieti kaikki tapahtumaketjut, jotka johtavat siihen että Harry on tappiolla 30 pelikierroksen jälkeen.

Ratkaisu.

(a) Satunnaismuuttuja X_i saa arvon $r = 35$ tn:llä $p = 1/37$ ja arvon -1 tn:llä $1 - p$. Tästä voidaan laskea $\mathrm{E}(X_i) = (r+1)p - 1 \approx -0.027$ ja $\mathrm{SD}(X_i) = (r+1)\sqrt{p(1-p)} \approx 5.84$.

(b) $\mathrm{E}(S_n) = n\,\mathrm{E}(X_1) \approx -0.027 \times n$. $\mathrm{SD}(S_n) = \sqrt{n}\,\mathrm{SD}(X_1) \approx 5.84 \times \sqrt{n}$.

(c)–(e)

$$\begin{aligned}
\mathrm{P}(S_n > 0) &= \mathrm{P}\left(\frac{S_n - \mathrm{E}(S_n)}{\mathrm{SD}(S_n)} > \frac{n(1-(r+1)p)}{\sqrt{n}(r+1)\sqrt{p(1-p)}}\right) \\
&\approx \mathrm{P}\left(Z > \frac{n(1-(r+1)p)}{\sqrt{n}(r+1)\sqrt{p(1-p)}}\right) \\
&\approx \mathrm{P}\left(Z > \frac{0.027n}{5.84\sqrt{n}}\right) \\
&= \mathrm{P}(Z > 0.0046\sqrt{n}),
\end{aligned}$$

missä Z noudattaa normitettua normaalijakaumaa.

Tätä arviota vastaavat arviot on listattu alla.

Pelikierrosten lkm n	Voitolla olemisen tn
30	$\approx 49\%$
3 000	$\approx 40\%$
300 000	$\approx 0.6\%$

(f) Ainoa tapa olla tappiolla 30 kierroksen jälkeen on hävitä euro jokaisella jokaisella 30 pelikierroksella. Näin ollen

$$\mathrm{P}(S_{30} > 0) = 1 - \left(\frac{36}{37}\right)^{30} \approx 56\%.$$

Näin saatu tarkka arvo poikkeaa jonkun verran normaaliapproksimaation antamasta arvosta (49%), mutta on suuruusluokaltaan oikea. Molemmat todennäköisyydet kertovat, että Harryllä on karkeasti ottaen 50–50 mahdollisuudet olla voitolla 30 pelikierroksen jälkeen. Tarkka arvo tosin antaa mielikuvan, että 30 pelikierroksen rulettipeli olisi Harrylle kannattava. Tämä mielikuva on väärä, sillä 30 pelikierroksen nettotuoton odotusarvo on negatiivinen: $\mathrm{E}(S_{30}) = -0.81$ euroa.

Kotitehtävät

3B3 (Kyselyn otoskoko) Tutkijat ovat laskeneet, että jos vähintään 100 henkilöä vastaa heidän postitse jaettavaan maakunnan laajuiseen kyselyynsä, on otoskoko tarpeeksi iso jatkopäätelmien tekemiseen. Kokemuksesta tiedetään, että vain 70% kyselyn vastaanottaneista palauttaa vastauksensa tutkijoille. Tämän takia tutkijat ovat päättäneet lähettää kyselylomakkeen yhteensä 150 henkilölle. Tutkijat olettavat, että kukin henkilö vastaa tai jättää vastaamatta toisista riippumatta.

(a) Laske binomijakaumaa käyttämällä (ilman normaaliapproksimaatiota) todennäköisyys, että tutkijat saavat tasan x vastausta kyselyynsä, x:n arvoilla 100, 101 ja 130.
Katso binomijakauman tiheysfunktio luentomonisteen luvusta 5.1. Binomikertoimen $\binom{n}{x}$ laskemiseen kannattaa käyttää laskinta tai tietokonetta.

(b) Määritä normaaliapproksimaatiota käyttämällä arvio todennäköisyydelle, jolla tutkijat saavat vähintään 100 vastausta kyselyynsä.

Ratkaisu.

(a) Merkitään satunnaismuuttujalla S saatavien vastausten lukumäärää. Kyseinen satunnaismuuttuja voidaan esittää summana muodossa

$$S = X_1 + X_2 + \cdots + X_{150},$$

missä indikaattorimuuttuja X_i määritellään kaavalla

$$X_i = \begin{cases} 1, & \text{jos } i\text{:s kyselyn vastaanottaja palauttaa vastauksensa,} \\ 0, & \text{muuten.} \end{cases}$$

Koska indikaattorimuuttujat ovat riippumattomia, niiden summa summa S on binomijakautunut parametrein $n = 150$ ja $p = 0.7$. Kysytyt todennäköisyydet saadaan binomijakauman tiheysfunktiosta

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

sijoittamalla siihen jakauman parametrit n, p sekä (vuorotellen) arvot $x = 100$, $x = 101$ ja $x = 130$.

Lukuarvot ovat: $f(100) \approx 0.0467$, $f(101) \approx 0.0540$ ja $f(130) \approx 9.231 \cdot 10^{-7}$.

(b) Indikaattorimuuttujan X_i odotusarvo on

$$\mathrm{E}(X_i) = \mathrm{P}(X_i = 1) = 0.7.$$

Lyhyellä laskulla nähdään, että myös $\mathrm{E}(X_i^2) = 0.7$, joten

$$\mathrm{SD}(X_i) = \sqrt{\mathrm{E}(X_i^2) - \mathrm{E}(X_i)^2} \approx 0.4583.$$

Odotusarvon lineaarisuuden perusteella

$$\mathrm{E}(S) \;=\; 150 \times 0.7 \;=\; 105.$$

Koska vastaajien toiminta oletettiin riippumattomaksi, ovat indikaattorimuuttujat $X_1, X_2, \ldots$ riippumattomia. Siten S:n keskihajonnaksi saadaan

$$\begin{aligned}\mathrm{SD}(S) &= \sqrt{\mathrm{SD}(X_1)^2 + \cdots + \mathrm{SD}(X_{150})^2} \\ &= \sqrt{150 \times \mathrm{SD}(X_1)^2} \\ &= \sqrt{150} \times \mathrm{SD}(X_1) \;\approx\; 5.612.\end{aligned}$$

Käyttämällä normaaliapproksimaatiota riippumattomien satunnaismuuttujien summaan $S = X_1 + \cdots + X_{150}$, todetaan että S on likimain normaalijakautunut odotusarvona $\mathrm{E}(S) = 105$ ja keskihajontana $\mathrm{SD}(S) \approx 5.612$. Näin ollen kysytty todennäköisyys on arviolta

$$\mathrm{P}(S \geq 100) \;=\; \mathrm{P}\left(\frac{S - 105}{5.612} \geq \frac{100 - 105}{5.612}\right) \;\approx\; \mathrm{P}(Z \geq -0.891) \;=\; \mathrm{P}(Z > -0.891),$$

missä Z noudattaa normitettua normaalijakaumaa. Normaalijakauman taulukoista (tai R-komennolla `1-pnorm(-0.891)`) saadaan arvioksi

$$\mathrm{P}(S \geq 100) \;\approx\; \mathrm{P}(Z > -0.891) \;\approx\; 81.4\%.$$

3B4 (Osakeportfolio) Markkinoilla on kahden yrityksen, Xanadun ja Ypsilonin osakkeet. Kummankin arvo tällä hetkellä on 100 euroa. Xanadun osakkeen tuottoa seuraavan vuoden aikana mallinnetaan satunnaismuuttujalla X, jolla on normaalijakauma odotusarvolla $\mu_X = 12$ ja keskihajonnalla $\sigma_X = 10$. Ypsilonin osakkeen tuotto on satunnaismuuttuja Y, jolla on normaalijakauma odotusarvolla $\mu_Y = 8$ ja keskihajonnalla $\sigma_Y = 10$. Osakkeiden tuotot oletetaan riippumattomiksi.

Tämä on tietenkin yksinkertaistava malli. Todellisilla markkinoilla tuotot olisivat luultavasti riippuvia, eivätkä ne ehkä olisi normaalijakautuneitakaan.

(a) Abel ostaa 200 Xanadun osaketta, joten hänen tuottonsa tulee olemaan $A = 200X$. Määritä A:n jakauma. Mikä on todennäköisyys, että Abel menettää rahaa (saa negatiivisen tuoton)?

(b) Bertta ostaa 100 Xanadun ja 100 Ypsilonin osaketta, joten hänen tuottonsa tulee olemaan $B = 100X + 100Y$. Määritä B:n jakauma. Mikä on todennäköisyys, että Bertta menettää rahaa?

(c) Laske A:n ja B:n korrelaatio. Ovatko ne riippumattomat?

(d) Määritä satunnaismuuttujan $A - B$ jakauma. Mikä on todennäköisyys, että Abel saa enemmän tuottoa kuin Bertta?

Vihje. Kahden normaalijakautuneen muuttujan summa on myös normaalijakautunut. Selvitä sen parametrit tuntemillasi kaavoilla. Muista myös mitä vakiolla kertominen tekee odotusarvolle ja keskihajonnalle. Huomaa lisäksi, että $-Y = (-1) \cdot Y$.

Ratkaisu.

(a) A on vakio kertaa X, joten A on normaalijakautunut. Sen parametrit ovat

$$\mu_A = \mathrm{E}(A) = \mathrm{E}(200X) = 200\mu_X = 2400$$
$$\sigma_A = \mathrm{SD}(A) = \mathrm{SD}(200X) = 200\sigma_X = 2000$$

Nyt normitetulla muuttujalla $Z = (A - \mu_A)/\sigma_A$ on standardinormaalijakauma, joten

$$\mathrm{P}(A < 0) = \mathrm{P}\left(Z < \frac{0 - \mu_A}{\sigma_A}\right) = F_Z\left(\frac{0 - 2400}{2000}\right) = F_Z(-1.2) \approx \mathbf{11.51}\%$$

missä F_Z on standardinormaalijakauman kertymäfunktio.

(b) B on kahden riippumattoman normaalijakautuneen muuttujan $100X$ ja $100Y$ summa, joten myös B on normaalijakautunut. Sen parametrit ovat

$$\mu_B = \mathrm{E}(B) = \mathrm{E}(100X + 100Y) = 100\mu_X + 100\mu_Y = 2000$$
$$\sigma_B = \mathrm{SD}(B) = \mathrm{SD}(100X + 100Y) = \sqrt{\mathrm{Var}(100X + 100Y)}$$
$$= \sqrt{100^2 \,\mathrm{Var}(X) + 100^2 \,\mathrm{Var}(Y)} = \sqrt{100^2 \cdot 10^2 + 100^2 \cdot 10^2} \approx 1414.2$$

Nyt normitetulla muuttujalla $Z = (B - \mu_B)/\sigma_B$ on standardinormaalijakauma, joten

$$\mathrm{P}(B < 0) = \mathrm{P}\left(Z < \frac{0 - \mu_B}{\sigma_B}\right) \approx F_Z\left(\frac{0 - 2000}{1414.2}\right) \approx F_Z(-1.414) \approx \mathbf{7.87}\%$$

Huomataan, että Bertan tuoton odotusarvo on pienempi kuin Abelin, mutta myös sen keskihajonta on pienempi (hajautuksen vuoksi). Tässä tapauksessa Bertan pienempi keskihajonta johtaa myös pienempään todennäköisyyteen menettää rahaa.

(c) Koska $A = 200X$ ja $B = 100X + 100Y$, kovarianssin bilineaarisuuden nojalla

$$\begin{aligned}
\mathrm{Cov}(A, B) &= \mathrm{Cov}(200X,\ 100X + 100Y) \\
&= \mathrm{Cov}(200X, 100X) + \mathrm{Cov}(200X, 100Y) \\
&= \mathrm{Cov}(200X, 100X) \qquad \text{(note: } X \text{ and } Y \text{ independent)} \\
&= 200 \cdot 100 \cdot \mathrm{Cov}(X, X) \\
&= 20000 \cdot \mathrm{Var}(X) = 20000 \cdot 100 = 2000000.
\end{aligned}$$

Edelleen

$$\mathrm{Cor}(A, B) = \frac{\mathrm{Cov}(A, B)}{\mathrm{SD}(A)\,\mathrm{SD}(B)} = \frac{2000000}{2000 \cdot 1414.2} \approx \mathbf{0.71}.$$

Korrelaatio on positiivinen, joten Abelin ja Bertan tuotot ovat (positiivisesti) riippuvat. Riippuvuus johtuu siitä, että molemmat omistavat Xanadun osakkeita.

(d) Tuottojen erotus on satunnaismuuttuja

$$D = A - B = (200X) - (100X + 100Y) = 100X - 100Y = 100X + (-100) \cdot Y.$$

Huomataan, että D on normaalijakautunut, ja

$$\begin{aligned}
\mu_D &= \mathrm{E}(D) = 100\,\mathrm{E}(X) - 100\,\mathrm{E}(Y) = 2400 - 2000 = 400, \\
\sigma_D^2 &= \mathrm{Var}(D) = \mathrm{Var}(100X - 100Y) = 100^2\,\mathrm{Var}(X) + (-100)^2\,\mathrm{Var}(Y) = 2000000, \\
\sigma_D &= \mathrm{SD}(D) = \sqrt{\mathrm{Var}(D)} \approx 1414.2.
\end{aligned}$$

Normitetulla muuttujalla $Z = (D - \mu_D)/\sigma_D$ on standardinormaalijakauma, joten

$$\begin{aligned}
\mathrm{P}(D > 0) &= 1 - \mathrm{P}(D \le 0) = 1 - \mathrm{P}\left(Z < \frac{0 - \mu_D}{\sigma_D}\right) \\
&\approx 1 - F_Z\left(\frac{-400}{1414.2}\right) \approx 1 - F_Z(-0.2828) \approx 1 - 0.3887 = \mathbf{61.13}\%.
\end{aligned}$$

Tämä on todennäköisyys, että Abel saa enemmän tuottoa kuin Bertta.

Viimeiset desimaalit voivat vaihdella riippuen välitulosten pyöristämisestä.

Taulukoissa annetaan F_Z yleensä vain 0.01:n välein, esimerkiksi $F_Z(-0.28) \approx 0.3897$ ja $F_Z(-0.29) \approx 0.3859$. Tästä voidaan päätellä, että $F_Z(-0.2828)$ on sillä välillä. Tällä kurssilla tarkkuudeksi riittää katsoa arvo lähemmästä pisteestä -0.28. Tällöin d-kohdan tulos on $1 - 0.3897 = 61.03\%$.

Lisätietoa (yli kurssialueen). Jos tietokonetta ei ole käytettävissä, myös taulukoista saadaan huomattavasti tarkempia arvoja, kun käytetään ns. lineaarista interpolaatiota eli arvioidaan funktion F_Z muuttuvan taulukoitujen arvojen välillä lineaarisesti. Pisteiden -0.29 ja -0.28 välillä, kun z kasvaa 0.01 yksikköä, niin taulukon mukaan F_Z kasvaa 0.0038 yksikköä, joten sen derivaatta on noin 0.38. Joten pisteessä $z = -0.2828 = -0.29 + 0.0072$ sen arvon pitäisi olla noin

$$\begin{aligned}
F_Z(z) &\approx F_z(-0.29) + 0.0072 \times 0.38 \\
&\approx 0.3859 + 0.0027 = 0.3886,
\end{aligned}$$

joka on varsin lähellä oikeaa.