805305A JOHDATUS REGRESSIO- JA VARIANSSIANALYYSIIN, sl 2022

Harjoitus 2, viikko 36: mikroluokkatehtävät

Tässä harjoituksessa perehdymme todennäköisyysjakaumien ja erityisesti eräiden tilastotieteessä keskeisten tunnuslukujen otantajakaumien (kuten normaali- ja t-jakauma) numeeriseen käsittelyyn R:n avulla.

Lisäksi teemme simulaatioita, joiden avulla tutkimme, kuinka eräät keskeiset otostunnusluvut käyttäytyvät, kun samasta populaatiosta poimitaan toistuvasti samankokoisia satunnaisotoksia.

R:n tarjoamat jakaumafunktiot ovat neljää eri tyyppiä:
- `d`*jak* ; pistetodennäköisyys- tai tiheysfunktio,
- `p`*jak* ; kertymäfunktio,
- `q`*jak* ; kvantiili- eli fraktiilifunktio,
- `r`*jak* ; jakaumasta satunnaislukuja generoiva funktio,

jossa "*jak*" viittaa jakauman R-nimeen. Esimerkiksi normaalijakauman R-nimi on `norm`, jolloin sen tiheys- ym. funktiot ovat oikealta nimeltään `dnorm, pnorm, qnorm` ja `rnorm`. Binomijakauman R-nimi on puolestaan `binom`, jolloin sen pistetodennäköisyysfunktio on `dbinom` ja muut jakaumafunktiot `pbinom, qbinom` ja `rbinom`. Kullakin näistä on omanlaisensa parametrointi, joka on syytä selvittää tarpeen mukaan ao. jakauman `help`-sivulta (kysymysmerkki eteen, esim. `?rnorm`). Simuloinnissa tarvitaan erityisesti `r`*jak*-tyyppisiä funktioita.

1. Analysoidaan tämän viikon kotitehtävissä 2–4 käsiteltyä pituus-aineistoa. Mallioletuksemme on, että naisopiskelijan pituus (cm) noudattaa normaalijakaumaa. Jakauman odotusarvoa koskevana nollahypoteesina on $H_0 : \mu = 167 (= \mu_0)$ cm. Jakauman varianssista σ^2 ei tehdä tarkkaa oletusta. Pituuden Y havaintoarvot olivat:

165.0, 166.0, 171.0, 154.0, 166.0, 159.5, 166.5, 158.5

(a) Talleta havaintoarvot vektoriin `pituus`:

```
> pituus <- c(165.0, 166.0, 171.0, 154.0, 166.0, 159.5, 166.5, 158.5)
```

(b) Piirrä havainnoista vaakasuora pistekuvio funktiolla `beeswarm()` asettamalla funktiolle lisäargumentti `horizontal=TRUE`. Lataa sitä ennen paketti beeswarm:

```
library(beeswarm)
beeswarm(pituus, horizontal=TRUE)
```

(c) Kirjoita oma R-funktio `SEmean()`, joka laskee argumenttina annetun numeerisen muuttujan keskivirheen:

```
> SEmean <- function(x) sd(x)/sqrt( length(x) )
```

(d) Laske, tallenna omiin muuttujiinsa ja tulosta pituushavaintojen lukumäärä, keskiarvo, keskihajonta ja keskiarvon keskivirhe. `> n <- length(pituus)`

```
> mean.pit <- mean(pituus)
> sd.pit <- sd(pituus)
> se.pit <- SEmean(pituus)
```

Moodle 1 Edellä tehtyjen sijoitusten jälkeen saat tulostettua esimerkiksi pituushavaintojen lukumäärän komennolla `n` ja pituuden otoskeskiarvon komennolla `mean.pit`. Moodle 2 Moodle 3

(e) Laske nollahypoteesia $H_0 : \mu = 167$ vastaava testisuureen $T = \frac{\bar{Y}-\mu_0}{\mathrm{SE}(\bar{Y})}$ havaittu arvo ja talleta se nimellä `Thav`:

```
> mu0 <- 167
> Thav <- (mean.pit - mu0)/se.pit; Thav
```

Moodle 4

(f) Piirretään seuraavaksi kuva testisuureen otantajakaumasta, joka on t-jakauma vapausasteella 7. Lisätään piirrettyyn tiheysfunktiokuvaan edellä laskettu testisuureen havaittu arvo.

```
> curve(dt(x,7), from=-4, to=4)
> points(Thav, 0, pch=16)
```

(g) Asetelmaan liittyvä 2-suuntainen P-arvo on $p_{\text{hav}} = 2[1 - F_T(|t_{\text{hav}}|; n-1)]$, jossa $F_T(t; df)$ on Studentin t-jakaumaa vapausasteluvulla df noudattavan satunnaismuuttujan T kertymäfunktio. Tämän kertymäfunktion arvoja laskee R-funktio nimeltä `pt()`; ks. ao. `help`-sivua. Funktio `abs()` laskee argumenttinsa itseisarvon.

```
> Phav <- 2*( 1 - pt(abs(Thav), n-1) )
> c(Thav, Phav)
```

Moodle 5

Vertaa saatuja testisuureen havaittua arvoa ja P-arvoa kotitehtävässä 4 saatuihin arvoihin. Moodle 6

(h) Laske ja tulosta μ:lle 90% luottamusvälin ala- ja yläraja tavanomaisella kaavalla $\bar{Y} \ \pm \ t_{0.95}(n-1) \times \mathrm{SE}(\bar{Y})$. jossa luottamustasoa $100(1-\gamma)$ % vastaava fraktiili $t_{1-\gamma/2}(n-1)$ t-jakaumasta vapausasteluvulla $df = n-1$ löytyy R-funktiolla `qt()`.

```
> t.95 <- qt(0.95, n-1)
> ci <- mean.pit + c(-1,1) * t.95 * se.pit
> ci
```

Moodle 7

(i) Toteuta testiä ja luottamusväliä koskevat laskelmat yhdellä R-funktiolla `t.test()`, jonka oletusarvoja täytyy muuttaa vain argumenttien `mu` ja `conf.level` osalta.

```
> t.test(pituus, mu=167, conf.level=0.90)
```

Moodle 8

2. Jatkoa edelliseen tehtävään. Lähdemme nyt oletuksesta, että tarkasteltavassa populaatiossa naisopiskelijoiden pituus ($= Y$) noudattaa normaalijakaumaa odotusarvolla $\mu = 167$ cm ja lisäksi oletamme jakaumalle määrätyn varianssiarvon, joka on $\sigma^2 = 5^2$ cm^2, eli hajonta on $\sigma = 5$ cm. Huomaa, että R:n `norm`-funktioissa varianssin σ^2 asemesta hajontaparametrina käytetäänkin jakauman keskihajontaa σ, jonka R-nimi näissä funktioissa on `sd`.

(a) Piirrä jakauman $N(167, 5^2)$ tiheysfunktion kuvaaja vaihteluvälille [150, 185] cm:

```
> u <- seq(150, 185, by=0.1) ; u
> mu <- 167; sig <- 5
> plot( u, dnorm(u, mu, sig), type = "l", ylim=c(0,0.1) )
```

Vektori `u` sisältää hilan mahdollisia pituusarvoja 0.1 senttimetrin välein: $u_1 = 150.0, u_2 = 150.1, \ldots, u_{351} = 185$. Komento `plot()` piirtää murtoviivan (`type = 'l'` eli "*line*") pisteiden $(u_i,\ \frac{1}{4}\varphi[(u_i - 50)/4])$ kautta, jossa $\varphi(z)$ on $N(0,1)$-jakauman tiheysfunktio.

(b) R-kertymäfunktion `pnorm()` avulla voidaan laskea todennäköisyys tapahtumalle $Y \le a$. Tällöin ko. funktio tarvitsee kolme argumentin määritystä: määritellään laskentapiste a, normaalijakauman odotusarvo μ ja keskihajonta σ ja laskentaan tarvittava komento on puolestaan muotoa `pnorm(a,` μ, σ). Laske `pnorm()`-funktion avulla seuraavat todennäköisyydet

(*i*) $\mathbb{P}(Y \le 160)$ Moodle 9

(*ii*) $\mathbb{P}(Y \ge 175)$ Moodle 10

(c) R:n normaalijakaumaan liittyvällä kvantiilifunktiolla `qnorm()` voidaan puolestaan etsiä Y:n jakaumasta sellainen arvo, jolle pätee, että $P(Y \le a) = \gamma$. Edellä γ on Y:n jakauman haluttu fraktiilipiste. Esimerkiksi alakvartiilin laskennassa γ:n arvoksi asetetaan arvo 0.25. Hae `qnorm()`-funktion avulla naisten pituusjakauman ns. 95% viitevälin rajat eli 2.5 %:n ja 97.5 %:n fraktiilit.
Moodle 11 Moodle 12

3. Jatkoa kahteen edelliseen tehtävään. Simuloimme nyt satunnaisotantaa naisopiskelijoiden pituuden oletetusta populaatiojakaumasta.

(a) Poimi ja sijoita vektoriin `otos` seitsemän havainnon satunnaisotos jakaumasta $N(167, 5^2)$ funktiolla `rnorm()`, laske ja tulosta

```
> otos <- rnorm(7, mu, sig); otos
> summary(otos); sd(otos); SEmean(otos)
```

Mitä havaintoja teet otoskeskiarvon ja -hajonnan arvojen poikkeamista teoreettisiin arvoihin $\mu = 167$ ja $\sigma = 5$ verrattuna?

(b) Toista edellisen kohdan toimenpiteet ja vertaile tuloksia otosten välillä.

(c) Tee sama uudelleen kaksi kertaa, mutta nyt otoskoolla 1000. Enää ei kuitenkaan kannata tulostaa otosta kokonaisuudessaan.

```
> otos <- rnorm(1000, mu, sig); otos
> summary(otos); sd(otos); SEmean(otos)
```

Mitä nyt havaitset? Miten esim. otosarvojen vaihteluväli muuttuu pieniin otoksiin verrattuna? Moodle 13

(d) Piirrä viimeisimmän otoksen ($n = 1000$) arvoista histogrammi välille [150, 185] kahden senttimetrin luokkaleveyksin samaan kuvaan kuin tehtävän 2 (a) tiheysfunktiokäyrä:

```
> hist(otos, freq=FALSE, breaks=seq(130,210, by=2),
+           xlim=c(150,185), add=TRUE)
```

(Argumentilla `breaks` määrätään pylväiden leveydet ja varaudutaan laajaankin vaihteluväliin otosarvoissa, mutta itse kuvioon säädetään kapeampi väli argumentilla `xlim`.)

4. Jatkamme satunnaisotannan simulaatiotutkimuksia ja tarkastelemme nyt, miten keskeiset otostunnusluvut käyttäytyvät toistettaessa otantaa monta kertaa. Käytämme funktiota `normotos.sim()`, joka on FM Timo Knürrin alunperin laatima. Tällä funktiolla on seuraavat argumentit

- `n` = otoskoko n yksittäisessä otoksessa,
- `mu` = populaatiojakauman odotusarvo μ, `sig` = keskihajonta σ,
- `level` = luottamustaso $1 - \gamma$, oletusarvona `0.9` eli 90%,
- `nsim` = simuloitavien otosten lukumäärä, oletusarvona `nsim=20`,
- `kuva` = looginen muuttuja oletusarvona `TRUE`, jolloin piirretään otoksista graafinen esitys,
- `loc` = oletusarvona `TRUE`, jolloin käytetään graafisessa esityksessä interaktiivista `locator()`-funktiota, joka mahdollistaa kuvaan tulevien alkioiden piirtämisen otos kerrallaan.

Funktio tuottaa tuloksenaan datakehikon, joka sisältää kustakin otoksesta lasketut tunnuslukujen arvot. Kun `kuva = T`, se myös piirtää samaan kuvaan kaikkien otosten havainnot ja niistä lasketut 90% luottamusvälit odotusarvolle μ.

(a) Edellä kuvattu funktio on jaossa Moodlessa `Esanfunktiot.R`-nimisessä tiedostossa. Kopio tiedosto Moodlesta tietokoneesi tämän kurssin työhakemistoon, lataa ko. funktioiden kirjasto R-istuntoosi ja säädä tulostuksen desimaalitarkkuus neljään desimaaliin:

```
> source("Esanfunktiot.R")
> options(digits=4)
```

(b) Poimi kooltaan $n = 10$ suuruisia otoksia naisten pituusjakaumasta $N(167, 5^2)$ ja sijoita tulokset datakehikkoon

```
> otos10 <- normotos.sim(10, mu, sig)
```

(c) Siirry grafiikkaikkunaan ja klikkaa hiiren vasemmanpuoleista näppäintä, jolloin 1. otoksen havaintojen pistekuvio ilmestyy koordinaatistoon. Klikkaa toisen kerran, jolloin vasemmalle reunalle tulostuvat otoskeskiarvo ja -hajonta, ja lisäksi kuvioon ilmestyy μ:n luottamusväli.

Jatka klikkaamista rauhalliseen tahtiin ja seuraa, kuinka otosarvot, tunnusluvut ja luottamusväli vaihtelevat otoksesta toiseen, kunnes kaikkien 20 otoksen tulokset ovat näkyvillä. Mitä havaintoja teet? Kuinka moni luottamusväli ei peittänyt μ:tä? Moodle 14

(d) Listaa datakehikon `otos10` sisältö ja tulosta:

```
> summary(otos10)
```

Mitä havaintoja teet otostunnuslukujen vaihteluvälien suuruuksista?

(e) Toista kohdat(b)–(d) mutta nyt käyttäen otoskokoa $n = 100$ ilman `locator()`-funktiota ja ilman datakehikon tulostusta:

```
> otos100 <- normotos.sim(100, mu, sig, loc=F)
> summary(otos100)
```

Vertaile luottamusvälien ja muiden otostunnuslukujen vaihtelua kohdan (d) tuloksiin. Mitä havaitset? Moodle 15

5. Simuloidaan nyt peräti 10000 otosta naisopiskelijoiden pituuden mallista $N(167, 5^2)$, kukin kooltaan $n = 7$, ja tarkastellaan otostunnuslukujen jakautumista.

(a) Toteuta simulaatio, talleta datakehikkoon ja kiinnitä. Tulosta tunnuslukujen jakaumien tiivistykset.

```
> otos7.10k <- normotos.sim(7, mu, sig, nsim=10000, kuva=FALSE, loc=FALSE)
> attach(otos7.10k)
> summary(otos7.10k)
```

Mitä huomioita teet? Moodle 16 Tutki myös, kuinka lähellä otosvarianssien ja otoshajontojen keskiarvot ovat teoreettisia arvoja σ^2 ja σ. Mitä havaitset?

(b) Piirrä simuloitujen otosten keskiarvojen histogrammi 1 cm luokkavälein. Piirrä samaan kuvioon keskiarvon $\bar{Y}$ teoreettisen otantajakauman $N(\mu, \sigma^2/n)$ kuin myös alkuperäisen muuttujan Y jakauman $N(\mu, \sigma^2)$ tiheysfunktioiden kuvaajat: `> hist(keskiarvo, freq=F, br=150:185)`

```
> lines( u, dnorm(u, mu, sig/sqrt(7)) )
> lines( u, dnorm(u, mu, sig), lty=3 )
```

Moodle 17

Mitä havaintoja teet simuloitujen otoskeskiarvojen jakautumisesta suhteessa teoreettiseen otantajakaumaansa?

(c) Piirrä otoskeskihajontojen histogrammi:

```
> hist(hajonta, freq=FALSE)
```

Onko otantajakauma symmetrinen vai vino?

6. Jatkoa edelliseen tehtävään. Tarkastelemme seuraavaksi testaustunnuslukujen käyttäytymistä simuloidusta otoksesta toiseen. Datakehikon sarake `T.suure` sisältää otoksista lasketut arvot testisuureelle $T = (\bar{Y} - 167)/\mathrm{SE}(\bar{Y})$, ja sarake `P.arvo` vastaavat 2-suuntaiset P-arvot.

(a) Piirrä histogrammi simuloitujen otosten T-arvojen jakautumisesta välille [-6, 6] luokkavälein 0.2. Piirrä samaan kuvioon $N(0, 1)$-jakauman tiheysfunktion kuvaaja punaisella värillä:

```
> hist(T.suure, freq=F, br=seq(-20, 20, by=0.2), xlim=c(-6,6) )
> tval <- seq(-6,6, by=0.1)
> lines( tval, dnorm(tval), col="red" )
```

Kuinka hyvin standardinormaalijakauma kuvaa T:n otantajakaumaa tällä vapausasteluvulla? Edelleen piirrä vapausastein $n - 1$ Studentin jakauman tiheysfunktion kuvaaja sinisellä

```
> lines( tval, dt(tval, df=7-1), col="blue" )
```

Kuinka hyvin tämä otantajakauma kuvaa simuloitujen T-arvojen jakaumaa? Kumpi jakaumista on paremmin yhteensopiva simuloitujen T-arvojen jakauman kanssa? Moodle 18

(b) Piirrä simuloitujen P-arvojen histogrammi:

```
> hist(P.arvo, freq=FALSE)
```

Mitä päättelet P-arvojen otantajakaumasta H_0:n vallitessa? Moodle 19

(c) Simuloiduista otoksista laskettujen 90% luottamusvälien $\bar{Y} \pm t_{0.95}(5) \times \mathrm{SE}(\bar{Y})$ ala- ja ylärajat on talletettu datakehikon muuttujiin `mu.alar` ja `mu.ylar`. Laske, kuinka moni alempi luottamusraja on suurempi kuin odotusarvo μ:

```
> length(mu.alar[ mu.alar > 167])
```

Laske vastaavasti, kuinka moni yläraja on pienempi kuin μ. Kuinka suuri on siten μ:n "ohi osuneiden" luottamusvälien osuus kaikista simuloiduista otoksista? Moodle 20

Lisääkö tämä simulaatio sen väitteen uskottavuutta, että t-jakaumaan perustuvan luottamusvälin todellinen peittotodennäköisyys on sama kuin nimellinen luottamustaso, kunhan muuttujan vaihtelua koskevat oletukset ovat päteviä?