

MC meetodid, praktikum 10

Teema 1: Gibbsi meetod.

Eeldame, et suudame genereerida väärtuseid tinglikest jaotustest

$$X_i \mid X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_m, \quad i = 1, \dots, m$$

- Valime algpunkti $(x_{11}, x_{12}, \dots, x_{1m})$
- iga $i = 2, 3, \dots, n$ korral genereerime
 - x_{i1} jaotusest $X_1 \mid X_2 = x_{i-1,2}, \dots, X_m = x_{i-1,m}$,
 - x_{i2} jaotusest $X_2 \mid X_1 = x_{i,1}, X_3 = x_{i-1,3}, \dots, X_m = x_{i-1,m}$,
 - x_{i3} jaotusest $X_3 \mid X_1 = x_{i,1}, X_2 = x_{i,2}, X_4 = x_{i-1,4}, \dots, X_m = x_{i-1,m}$,
 - $\dots$
 - x_{im} jaotusest $X_m \mid X_1 = x_{i,1}, X_2 = x_{i,2}, \dots, X_{m-1} = x_{i,m-1}$,
- Järjestikused punktid ei vasta sõltumatutele juhuslikele vektoritele!

Harjutus 1

Vaatleme Gibbsi meetodi kasutamist selleks, et genereerida punkte funktsiooniga

$$f(x, y, z) = \begin{cases} e^{-x-2y-3z-xyz}, & x, y, z \geq 0, \\ 0, & \text{vastasel korral} \end{cases}$$

proportsionaalse tihedusfunktsiooniga jaotusest. Küllalt lihtne on veenduda, et kui fikseerida y, z , siis see funktsioon käitub x jargi nagu eksponentjaotuse $\text{Exp}(1 + yz)$ tihedusfunktsioon, kui fikseerida x, z , siis käitub see nagu eksponentjaotuse $\text{Exp}(2 + xz)$ tihedusfunktsioon jne. Genereeri Gibbsi meetodiga 100000 punkti ja uuri 'acf' käsuga iga koordinaadi puhul, kui kiiresti lähevad järjestikuste väärtuste autokorrelatsioonid nulli. Kui eesmärgiks on sõltumatute punktide genereerimine, siis millise sammuga võiks tekitatud jadast punkte valida?

Teema 2: Hüpoteesi püstitamine lähtuvalt andmetest

Idealis peaks olema otsustatud juba enne katsete teostamist, mida ja kuidas kontrollima hakatakse. Tegelikuses aga sageli kogutakse kõigepealt andmed ning kui neis paistab midagi imelikku, siis hakatakse testide teel tõestama, et nähtu ei ole puhtalt juhuslik. Sellisel juhul peab olema väga ettevaatlik, et nullhüpoteesi ja testi valik ei oleks otseselt katsetulumustest motiveeritud, sest muidu võib kõigest lihtsalt jõuda valedele järeldustele. Veendume selles järgmise harjutuse abil.

Harjutus 2

Kuuekümneme täringuviske tulemusena saadi allpool toodud tabeliga kirjeldatud tulemused (sagedustabelina).

```
## tulemused
## 1  2  3  4  5  6
## 10 17  5  9  9 10
```

Tulemustes esineb kahtesid kahtlaselt palju, nii et tekib kahtlus täringu aususe osas. Kui kindlalt aga saame öelda, et täring on ebaaus? Vaatleme erinevaid võimalusi, kuidas saaks nullhüpoteesi (H_0 - täring on aus) ümber lükata. Selleks tuleb fikseerida statistik, mida valimi põhjal saab leida ja fikseerida otsustuseeskiiri (millised väärtused loeme nullhüpoteesi korral ebaharilikeks). Tavaliselt statistik konstrueeritakse nii,

et otsustamiseks sobivad tingimused kujul: $\text{statsitik} > \text{arv}(\alpha) \Rightarrow$ lükkame nullhüpoteesi ümber, kusjuures $\text{arv}(\alpha)$ on valitud selline, et ümberlükkamise tõenäosus nullhüpoteesi kehtimisel on α . Sellise kriteeriumi korral valimi põhjal leitud statistiku väärtuse p -väärtus on aga tõenäosus, et statistik omandab vähemalt sama suure väärtuse kui valimis saadud väärtus.

Vaatleme järnevaid statistikuid

- $S_1 = n_2$, kus n_2 tähistab kahtede arvu valimis
- $S_2 = |n_2 - 10|$
- χ^2 -testi statistik $S_3 = \sum_{i=1}^6 \frac{(n_i - 10)^2}{10}$

Leia kõigi kolme statistiku p -väärtused. (Vihje: esimesel kahel juhul on abiks käsk `pbinom`, kolmandal juhul võib kasutada R käsku `chisq.test`). Kas meil on piisavalt infot, et olulisuse nivool 0,05 nullhüpoteesi ümber lükata? Mis tingimustel me võime eelnevate statistikute põhjal otsuseid teha?

Uurime simulatsioonide teel, mis juhtub, kui tegutseme järgmiselt:

1. Teeme katse (60 viiset)
2. Valime tulemuste põhjal kõige “kahtlasema” tahu ehk kõige rohkem esinenud silmade arvu
3. Rakendame selle tahu põhjal statistiku S_1 analoogi (st leiame p -väärtuse vastavalt valitud tahu esinemiste arvule) ja lükkame nullhüpoteesi ümber juhul, kui p -väärtus on väiksem kui 0.05

Leia 10000 katse põhjal hinnang, kui suure tõenäosusega tegelikult lükatakse nullhüpoteesi ümber eelmise stsenaariumi kohaselt tegutsedes, kui täring on tegelikult aus.

Teema 3: p -väärtuse leidmine MC simulatsioonide abil.

Statistika algusaegadel sai kasutada ainult selliseid statistilisi teste, mille korral teststatistiku jaotus (või vähemalt asümptootiline jaotus) on leitav analüütiliselt. Kaasajal ei ole see aga enam nii oluline, sest juhul, kui me oskame nullhüpoteesile vastavaid valimeid arvutil genereerida, saame järelikult genereerida ka teststatistiku väärtuseid ja nende põhjal hinnata statistiku jaotusega seotud arvnäitajaid (sh konkreetse valimi puhul saadud tulemuse p -väärtust, mille saame leida näiteks empiirilist jaotusfunktsiooni kasutades).

Harjutus 3

Täringu ebaausus väljendub selles, et tulemused ei ole kõik võrdse tõenäosusega. Kuna suurte arvude seaduse põhjal konkreetse tulemuse esinemista arv jagatuna katsete arvuga koondub selle tulemuse tõenäosuseks, siis statistik, mis leiab maksimaalse absoluutse erinevuse suhteliste sageduste ja $\frac{1}{6}$ vahel peaks piisavalt suure valimi korral käituma oluliselt erinevalt ausa täringu ja ükskõik millise ebaausa täringu korral. Seega võime kasutada täringu aususe testimisel ka statistikut

$$S = \max(|\frac{n_i}{n} - \frac{1}{6}|, i = 1, 2, \dots, 6),$$

kus n on valimi suurus ja n_i on i silma esinemiste arv valimis. Leia selle statistiku p -väärtus eelmises ülesandes toodud valimi korral, simuleerides selleks 10000 valimi moodustamist ausa täringu abil. Kasuta empiirilist jaotusfunktsiooni (R-s käsk `ecdf`) p -väärtuse arvutamiseks.

Märkus: Kui arvutatava statistiku jaotusfunktsioon on pidev (st tõenäosus saada täpselt sama väärtust erinevate valimite korral on 0), siis võib leida valimist leitud statistiku väärtusele $\hat{s}$ vastava p väärtuse kujul $1 - F_m(\hat{s})$, kus F_m on m valimi põhjal leitud statistiku väärtustele vastav empiiriline jaotusfunktsioon. Kui statistik võib omandada mingit fikseeritud väärtust positiivse tõenäosusega (mis kehtib alati, kui valim vastab diskreetsele juhuslikule suurusele), siis kehtib võrdus

$$P(S \geq \hat{s}) = 1 - F_S(\hat{s}) + P(S = \hat{s})$$

Kui oskame leida sellise arvu ε , et statistiku võimalike väärtuste vahe on suurem kui ε , siis saab eelnevat valemit lihtsustada kujule (mõtke järele, miks see nii on!)

$$P(S \geq \hat{s}) = 1 - F_S(\hat{s} - \varepsilon)$$

Teema 4: Testi võimsuse leidmine MC simulatsioonide abil.

Selleks, et valida test konkreetse uuringu jaoks, on mõistlik vaadata, milline testidest eristab kõige paremini meie jaoks nullhüpoteesist oluliselt erinevad jaotuseid. Oletame, et tüüpilised valed täringud on sellised, kus üks külg on veidi suurema tõenäosusega kui $\frac{1}{6}$ ja selle vastaskülg on samavõrra väiksema tõenäosusega ning ülejäänud käituvad samamoodi kui ausad täringud. Tee simulatsioonide abil kindlaks, kas eelmises ülesandes vaadeldud test on võimsam kui χ^2 -test juhul, kui neid kasutada $\alpha = 0.05$ korral 120 viske põhjal täringu aususe kindlakstegemiseks, kui visatava täringu visketulemuste tõenäosused on antud tabelina

tulemus	1	2	3	4	5	6
tõenäosus	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{4}$

Vihje Kõigepealt on mõistlik tekitada eelmise ülesande statistiku jaoks empiiriline jaotusfunktsioon nullhüpoteesi ja vaadeldava valimimahu jaoks, alles seejärel asuda alternatiivse jaotusega katseid tegema

Teema 5: Bootstrap meetodi kasutamine

Kui meil on lähtejaotuse (või üldkogumi) parameetri θ hindamiseks hinangufunktsioon (ehk statistik) $\hat{\theta}$, siis bootstrap hinnang nihkele avaldub kujul

$$\frac{1}{m} \sum_{i=1}^m \hat{\theta}_i - \theta_B,$$

Kus θ_B on vaadeldava parameetri väärtus bootstrap jaotuse (ehk valimile vastava empiirilise jaotuse) korral ning $\hat{\theta}_i$, $i = 1, 2, \dots, m$ on hinangufunktsiooni väärtused bootstrap jaotusest võetud sõltumatute valimite korral. Bootstrap usaldusintervall olulisusnivool α avaldub aga kujul

$$(\hat{\theta} + \theta_B - q_{1-\frac{\alpha}{2}}, \hat{\theta} + \theta_B - q_{\frac{\alpha}{2}}),$$

kus q_α tähistab hinnangu $\hat{\theta}$ bootstrap jaotuse korral leitud α -kvantiili ning $\hat{\theta}$ on esialgse valimi põhjal arvutatud parameetri hinnang.

Vaatleme juhtu, kus tahame hinnata lähtejaotuse puhul suurust $\lambda = \frac{1}{\sigma_X}$. Selle hindamiseks kasutame hinangufunktsiooni $\frac{1}{s}$, kus s on standardhälbe hinnang (arvutatud käsuga `sd()`). NB! Käsk `sd` on üldkogumi standardhälbe hinnang valimi põhja ning ei ole võrdne empiirilise jaotuse standardhällbega!

Harjutus 4

Olgu meil juhusliku suuruse X 15st väärtusest koosnev valim

```
## [1] 0.10 1.69 0.50 0.18 0.10 1.39 1.49 1.34 0.78 3.65 0.11 1.53 5.15 0.98 2.48
## [16] 0.50 2.43 1.35 4.54 0.18
```

Hinda bootstrap meetodil statistiku nihet ja leia selle abil parandatud hinnang λ väärtusele. Leia samuti olulisusnivoole $\alpha = 0.1$ vastav usaldusvahemik

Harjutus 5

Leia Monte-Carlo meetodil tegelik statistiku nihe 20-se valimi korral eeldusel, et lähtejaotus on $Exp(0.5)$

Kodune töö nr. 7 (1,5 punkti, tähtaeg 15.11.2020)

1. (0,5 punkti) Sageli me kasutame küllalt väikeste valimimahtude korral hinnanguid, mis kehtivad valimi mahu kasvamisel lõpmatusse. Näiteks tõenäosusega $1 - \alpha$ kehtiv keskväärtuse usaldusvahemik kujul

$$(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}),$$

kus $\bar{x}$ on keskväärtuse hinnang, s on standardhälbe hinnang ja z_q on standardse normaalkõrvale q -kvantiil, kehtib lõpliku dispersiooniga juhuslikule suurusele vastava sõltumatu valimi korral asümptootiliselt, st piisavalt suurte n väärtuste korral. Tee kindlaks, kui suure tõenäosusega paikneb õige keskväärtus selle usaldusvahemiku sees juhul, kui $\alpha = 0.05$ ja valim 30 vaatlusega on genereeritud jaotusfunktsioonile

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\sqrt{x}}, & x \geq 0, \end{cases}$$

vastavast jaotusest.

2. (0,5 punkti) Kasuta eelmises praktikumis kirjeldatud Metropolis-Hastingsi algoritmi versiooni (ehk Metropolis algoritmi) selleks, et genereerida vektoreid funktsiooniga

$$f(x, y) = \begin{cases} x^4 + 0.1y^2, & 0.1(x+1)^2 + (y-1)^2 \leq 1, \\ 0 & \text{mujal} \end{cases}$$

proportsionaalse tihedusfunktsiooniga jaotusest. Millise sammuga tuleks Sinu arvates väärtuseid võtta, et need oleks enam-vähem sõltumatud? Leia hinnang tõenäosusele, et saadud punktid satuvad ühikringi $x^2 + y^2 \leq 4$, kasutades selleks 1000 genereeritud sõltumatuse tingimusi rahuldavat punkti. (0,5 punkti) Kasuta Gibbsi meetodit selleks, et genereerida kolmnurgas tippudega $(0, 0)$, $(1, 0)$, $(2, 1)$ ühtlase jaotusega juhuslikke vektoreid. Tee kindlaks, millise sammuga peaks punkte võtma, et need võiksid vastata sõltumatutele vaatlustele sellest jaotusest, kujuta 100 sõltumatutele katsetulemustele vastavat punkti graafikul.