Seminarske naloge

SPLOŠNA NAVODILA

Navodila veljajo praviloma za vse naloge, seveda pa jih je od naloge do naloge potrebno malce prilagoditi.

1. Ugotovite, kaj je hipoteza, ki jo želimo preverjati, kakšen test bi bil zanjo smiseln in predstavite osnove teorije pri tem testu (kaj je testna statistika, porazdelitev le-te, v kakšnih primerih naj bi se uporabljal, kaj so predpostavke).

2. Izmislite si parametre, ki bodo generirali smiselne podatke, ki jih pričakujete v nekem konkretnem primeru.

3. Pri preverjanju lastnosti testov vedno najprej generirajte podatke tako, da ničelna domneva velja in tako preverite lastnosti testne statistike. Nato generirajte podatke tako, da velja alternativna domneva... in komentirajte rezultate.

4. Spreminjajte parametre, ki ste si jih izmislili v 2. točki, in poglejte, kako se spreminjajo rezultati testa.

5. Natanko opišite postopek generiranja podatkov in prikažite izsek iz tabele podatkov, ki ste jo generirali. Opišite, kateri parametri so vedno enaki, kateri se spreminjajo in zakaj.

Težavnost oz. obsežnost posameznih nalog je nakazana z zvezdicami. Naloge z eno zvezdico so nekoliko krajše in osnovnejše, v želji za najvišjo oceno je pri njih potrebno precej dodatne lastne kreativnosti. Naloge z dvema zvezdicama so kompleksnejše in primerne za vse študente, ne glede na želeno oceno. Naloge s tremi zvezdicami že v osnovni obliki zagotavljajo izziv in nekoliko več študijskega časa.

Kazalo

1 Izračun velikosti vzorca *

Raziskovalka se ukvarja z odpravljanjem motenj uriniranja pri otrocih v starosti 5-12 let. V preteklih letih so uvedli terapijo, ki jo otroci lahko opravljajo bodisi v bolnišnici bodisi v domačem okolju z rednim nadzorom svetovalca. Raziskovalki se zdi, da navkljub naporom terapija v domačem okolju ni enako uspešna, to želi dokazati z raziskavo. Predvideva, da bodo otroci v bolnišnici dosegali 90% uspešnost, medtem ko bo v domačem okolju uspešnih največ 75% otrok. Ker tovrstnih otrok letno ni veliko, jo zanima, kakšen je najmanjši možen vzorec, da bo moč njenega testa enaka 0,8.

Odločite se za ustrezni test, preverite njegovo obnašanje pod ničelno domnevo (torej, da res zavrača v 0,05 odstotkih) in s pomočjo simulacij izračunajte potrebno velikost vzorca. oglejte si tudi, kaj se zgodi, če se predvidevana odstotka nekoliko spremenita.

2 Izbira ustreznega testa *

Svetujete raziskovalki, ki želi primerjati stanje depresivnosti med bolnicami z rakom dojke in primerljivo splošno populacijo. Vse vključene posameznice so izpolnile vprašalnik, ki pomaga pri določitvi depresivnosti in ima naslednje vrednosti:

- 0 - 9 točk: depresivnost malo verjetna
- 10 - 17 točk: možnost blažje depresivnosti
- 18 - 21 točk: na robu depresivnosti
- 21 - 35 točk: blažja do srednja depresivnost
- 36 - 53 točk: srednja do huda depresivnost
- 54+ točk: huda depresivnost

Zbrala je 112 bolnic in 112 kontrolnih preiskovank. Porazdelitev rezultatov na testu je v obeh skupinah precej asimetrična, nad 21 točk doseže približno 35% posameznic, nad 36 pa le dober odstotek.

1. Raziskovalka želi posnemati neko raziskavo iz predhodno objavljene literature, zato bi rada uporabila test χ^2 in primerjala zgoraj omenjene skupine.
   Razložite, zakaj to v tem primeru ni mogoče. Skušajte generirati smiselne podatke (uporabite npr. gama porazdelitev, vrednosti naj bodo celoštevilske, vrednosti nad 60 spremenite v 60) in poglejte, kako reagira R.
2. Raziskovalka sedaj ni prepričana, kako naj analizira podatke, zato ji predlagate več možnosti
   - test t (enaki varianci)
   - Mann-Whitneyev test
   - test χ^2, pri čemer bolnike razdelite v dve skupini (do 21 in 21 ter več)
   - test χ^2, pri čemer bolnike razdelite v dve skupini (do 18 in 18 ter več)
   - test χ^2, pri čemer bolnike razdelite v tri skupine (do 10, 10-20, 21 ter več)

Za vsakega izmed teh testov uporabite ustrezno funkcijo v R in preverite velikost testa pod ničelno domnevo (v obeh skupinah generirate podatke z isto porazdelitvijo, preštejete v koliko simulacijah je vrednost p pod 0,05). Pri tem pri testu t preverite ali ne bi bilo bolj pravilno testno statistiko aproksimirati z normalno porazdelitvijo, pri testu χ^2 pa preverite ali je smiselno uporabljati Yatesov popravek, ki je prednastavljena možnost v R - `correct=T`.

3. Za neko smiselno alternativno domnevo izračunajte moči tistih testov, ki se dobro obnašajo pod ničelno domnevo. Nato si izmislite podatke, jih analizirajte z vsakim testom in razložite razliko v interpretaciji posameznih testov.

4. Raziskovalka se odloči, da bi najraje preverila rezultate vseh testov, nato pa se odločila za tistega, ki bo imel najnižjo vrednost p. Izračunajte, kakšna je verjetnost, da bo na ta način zavrnila ničelno domnevo, kadar v populaciji ni razlik.

3 Parni test za primerjavo deležev **

Fizioterapevti so skupini starostnikov predpisali individualno vadbo za moč 3x tedensko. Po treh mesecih preverijo stanje, za vsakega posameznika zabeležijo, ali vadi redno ali ne (DA in NE). Nato organizirajo skupinsko vadbo, ki poteka 1x na 14 dni. Po 3 mesecih skupinske vadbe znova preverijo, ali posamezniki redno izvajajo individualno vadbo. Zanima jih, ali je občasna skupinska vadba pripomogla k bolj redni individualni vadbi. V ta namen bodo uporabili McNemarjev test.

1. Kratko predstavite idejo testa in poiščite ustrezno funkcijo v R.
2. Generirajte smiselne podatke in preverite velikost testa pod ničelno domnevo (ali test res zavrača z deležem α). Preverite, kako se obnaša na majhnih vzorcih, preučite lastnosti popravka.
3. Z grafom ali tabelo predstavite, kako je velikost potrebnega vzorca (za moč 0,8) odvisna od predpostavk.
4. Generirajte ene podatke, jih analizirajte in interpretirajte.

4 Test t in Mann-Whitneyev test **

Test t je najpogosteje uporabljan test, čeprav marsikatera populacija ni normalno porazdeljena. V takem primeru ga zato pogosto nadomestimo z neparametričnim testom - kadar imamo dva neodvisna vzorca uporabimo test Mann-Whitney.

- Poiščite podatke, ki izhajajo iz asimetrične porazdelitve in kjer želimo pokazati, da je ena skupina premaknjena glede na drugo (razlika v povprečjih oz. mediani). Analizirajte z obema testoma in interpretirajte.

- Generirajte podatke iz več različnih porazdelitev: normalne, enakomerne, gama porazdelitve z različno stopnjo asimetrije. Dodatno generirajte tudi iz porazdelitve, ki je podobna tisti, ki jo opažate na vaših podatkih.
- Za vse porazdelitve preverite velikost obeh testov, torej v kakšnem deležu zavrnete ničelno domnevo, če so podatki generirani tako, da ničelni domnevi ustrezajo.
- V situacijah, kjer je velikost obeh testov ustrezna, primerjajte njuni moči.

5 Linearna regresija - variabilnost in interval zaupanja **

Imamo model linearne regresije z eno neodvisno spremenljivko

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

kjer je ε normalno porazdeljena spremenljivka s povprečjem 0.

1. Recimo, da odvisnost ocene na izpitu statistike od števila ur učenja dnevno sledi modelu linearne regresije z $\beta_0 = 40$, $\beta_1 = 15$ in $\varepsilon \sim N(0{,}20)$ (da bo model zares držal, privzemite, da je rezultat na izpitu lahko tudi negativen). Število ur učenja je enakomerno porazdeljena spremenljivka na intervalu [0,4] (ur na dan).

   - Generirajte vrednosti za 20 posameznikov ter narišite razsevni diagram števila točk na izpitu glede na število ur učenja (podatke shranite) (opomba: dovolite tudi negativne vrednosti na izpitu, če vam jih model generira).
   - Ocenite vektor koeficientov β ($\widehat{\beta} = (X^T X)^{-1} X^T Y$) ter oceno primerjajte z izpiskom iz R (Funkcija `lm`).
   - Na sliko narišite ocenjeni model (premico) ter ga primerjajte s populacijskim modelom.
   - Simulacijo ponovite 1000-krat (z istimi vrednostmi X) in izračunajte povprečje tako dobljenih ocen za regresijska koeficienta β_0 in β_1.

2. Variabilnost ocene koeficientov

   - Izračunajte kovariančno matriko za svoj konkretni primer z uporabo ustrezne formule in jo primerjaj z izpiskom funkcije `lm`.
   - Simulacijo pri enakih vrednostih neodvisne spremenljivke ponovite 1000-krat in izračunajte empirično variančno-kovariančno matriko. Primerjajte jo s povprečjem ocenjene matrike.
   - Ponovite prejšnjo točko tako, da na vsakem koraku generirate tudi neodvisne spremenljivke. Kaj se zgodi z variancami?

3. Interval zaupanja za premico

- Na podlagi podatkov o 20 študentih in izpiska funkcije `lm` izračunajte pričakovani rezultat na izpitu za študenta, ki se uči 2 uri na dan in dodajte 95% interval zaupanja (zanima nas torej $\beta_0 + 2\beta_1$, za populacijski vrednosti parametrov) Namig: reparametrizirajte model, tako da namesto X za neodvisno spremenljivko vzamete $X - 2$.
- Po enakem postopku izračunajte 95% interval za vsako izmed 20 točk. Nato 1000x ponovite simulacijo in preštejte, kolikokrat je premica znotraj vseh intervalov. Komentirajte pokritost intervala zaupanja (ali je enaka nominalni, zakaj ...).
- Kako izgleda interval zaupanja za premico, če nas zanima le naklon (torej če bi v simulaciji vse ocenjene premice risali skozi isto presečišče?).
- Kako bi napovedali dejanski rezultat na izpitu za študenta, ki se uči 2 uri na dan (ne zanima nas le ocena povprečja, temveč ocena dejanskega rezultata na izpitu).

6 Linearna regresija - velikost testa in moč **

Imamo model linearne regresije z eno neodvisno spremenljivko

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

kjer je ε normalno porazdeljena spremenljivka s povprečjem 0.

1. Recimo, da odvisnost ocene na izpitu statistike od števila ur učenja dnevno sledi modelu linearne regresije z $\beta_0 = 40$, $\beta_1 = 15$ in $\varepsilon \sim N(0{,}20)$ (možno je doseči tudi negativen rezultat). Število ur učenja je enakomerno porazdeljena spremenljivka na intervalu [0,4] (ur na dan).

   - Generirajte vrednosti za 20 posameznikov ter narišite razsevni diagram števila točk na izpitu glede na število ur učenja (podatke shranite).
   - Ocenite vektor koeficientov β ($\widehat{\beta} = (X^T X)^{-1} X^T Y$) ter oceno primerjajte z izpiskom iz R (Funkcija `lm`).
   - Na sliko narišite ocenjeni model (premico) ter ga primerjajte s populacijskim modelom.
   - Simulacijo ponovite 1000-krat (z istimi vrednostmi X) in izračunajte povprečje tako dobljenih ocen za regresijska koeficienta β_0 in β_1.

2. Velikost testa

   - Generirajte vrednosti pod ničelno domnevo ($H_0 : \beta_1 = 0$), velikost vzorca naj bo 20. Simulacijo ponovite 1000x (z istimi vrednostmi X) in narišite porazdelitev vrednosti p (Waldov test) za koeficient β_1. Ali je velikost testa ustrezna?
   - Izračunajte še posplošeni test razmerja verjetij za koeficient β_1 (vsakič prilagodite cel model in model, ki vsebuje le konstanto). Preštejte kolikokrat zavrne ničelno domnevo, čeprav ta drži. Je velikost testa ustrezna?

3. Moč

   - Generirajte vrednosti z $\beta_1 = 15$. Simulacijo ponovite 1000x in narišite porazdelitev vrednosti p (Waldov test) za koeficient β_1. Kolikšna je moč testa?

- Izračunajte še moč posplošenega testa razmerja verjetij za koeficient β_1
- Povečajte vzorec na 50, kaj se zgodi z močjo?
- Ostanite pri vzorcu velikosti 50, vendar vrednosti neodvisne spremenljivke generiraj le na intervalu [1,3]. Kaj se zgodi z močjo?

7 Preverjanje normalnosti in test t *

Test t za dva neodvisna vzorca predpostavlja normalno porazdelitev populacije. Seveda to v splošnem ne bo vedno res. V tej nalogi preverite tri možne strategije za primerjavo povprečij dveh neodvisnih vzorcev (z enako variabilnostjo):

- Vedno uporabimo test t
- Vedno uporabimo test Mann-Whitney
- Najprej preverimo normalnost (test Shapiro-Wilk), na podlagi tega testa se odločimo med testom t in testom Mann-Whitney

Preverite in primerjajte strategije, pri tem naj bodo vključeni naslednji koraki:

- Poiščite podatke, ki izhajajo iz asimetrične porazdelitve in kjer želimo pokazati, da je ena skupina premaknjena glede na drugo (razlika v povprečjih oz. mediani). Analizirajte z obema testoma in interpretirajte. Spreminjajte tudi velikost vzorca.

- Generirajte podatke iz različnih porazdelitev, pri tem poiščite porazdelitve, ki vam bodo dale različno stopnjo asimetrije. Vključite tudi Cauchyjevo porazdelitev in porazdelitev, ki bo s precejšnjo verjetnostjo enakomerna na intervalu $[a, b]$ in imela neko majhno verjetnost, da zavzame točko c, ki je izven intervala $[a, b]$.
- Podatke najprej simulirajte pod ničelno domnevo in primerjajte vse tri strategije (velikost testa).
- Nato podatke simulirajte še pod alternativno domnevo (različne vrednosti) in primerjajte moči.

Povzemite rezultate in podajte napotke za uporabo testov.

8 Primerjave opisnih spremenljivk, logistična regresija **

Zdravniki želijo primerjati dve vrsti operacije, imajo nek vzorec bolnikov (cca 300 bolnikov) z rakom, pri polovici bolnikov je bila narejena ena vrsta operacije, pri drugi polovici druga. Zanimata jih dve vprašanji:

- Ali je bila odločitev za vrsto operacije povezana s stadijem bolnika (vsak bolnik je v enem od štirih stadijev bolezni)
- Ali ena vrsta operacije povzroča manj zapletov kot druga.

Ker vsi zapleti niso enakovredni, primerjajo vsakega posebej, zanima jih 10 različnih zapletov. Naredijo 10 primerjav (za vsakega bolnika pogledajo ali se mu je nek zaplet zgodil ali ne). Pri preverjanju povezanosti s stadijem naredijo enako - 4 primerjave, za vsak stadij posebej torej primerjajo dve opisni spremenljivki (je oz. ni v stadiju in prva/druga vrsta operacije).

1. Oglejte si, kaj se zgodi z napako I. reda pri takem načinu preverjanja. Komentirajte.
2. Podajte kak predlog, kako bi tovrstne podatke lahko ustrezneje analizirali.

9 Parni test t in mešani modeli ***

Enostavni linearni model, ki ima le eno binarno neodvisno spremenljivko X, je ekvivalenten testu t za dva neodvisna vzorca. Zanima nas, kako delujejo mešani modeli (posplošitev linearnih modelov, ki dovoljujejo med seboj odvisne spremenljivke) v primerjavi s parnim testom t.

- Teoretično pokažite ekvivalentnost testa t in linearnega modela (predpostavite enako varianco Y v obeh skupinah glede na X)
- Kratko predstavite osnovne ideje mešanih modelov (teoretično in s konkretnim primerom, predlog literature: Wood, S., Pinheiro, J. C., Bates, D. M. (2001). Mixed-effects models in S and S-PLUS)
- Predstavite funkcijo v R-u, ki prilagodi mešani model podatkom (npr. funkcija `lmer`). Pokažite konkretni primer (generirajte podatke z nekimi parametri in pokažite, kje najdete ocene teh parametrov v izpisu)
- Primerjajte parni test t in mešani model (predpostavke in lastnosti, velikost, moč)

Dodatno - izberite eno od spodnjih dveh točk:

- Denimo, da bi radi primerjali meritve istih posameznikov ob dveh različnih časovnih trenutkih, vendar nam pri precejšnjem deležu posameznikov manjka prva oziroma druga meritev. Ponujajo se nam (vsaj) tri možnosti analize:
  - uporabimo le posameznike, ki imajo obe meritvi, naredimo parni test t
  - uporabimo vse meritve, ne upoštevamo povezanosti med meritvama, naredimo neodvisni test t
  - uporabimo vse meritve, podatek o povezanosti upoštevamo s pomočjo mešanega modela

  Primerjajte posamezne možnosti s simulacijami in komentirajte
- Multiplo linearno regresijo uporabljamo kot posplošitev testa t za dva neodvisna vzorca, kadar imamo poleg binarne spremenljivke X še neko drugo motečo neodvisno spremenljivko Z. Zanima nas, ali bi lahko mešani model z več neodvisnimi spremenljivkami na enak način uporabljali kot posplošitev parnega testa t. Proučite lastnosti tovrstne analize.

10 Test primerjave deležev in hi-kvadrat test *

Ta naloga ima dva dela.
Del 1: Pogosto primerjamo dve skupini glede na opisno spremenljivko z več kategorijami (npr. primerjamo skupini glede na nivo znanja jezika). Avtorje člankov zanima primerjava vsake kategorije posebej, zato naredijo teste za vsako kategorijo posebej (nizki nivo DA/NE, skupina 1/2).

Obrazložite, ali je tak način uporabe pravilen, s simulacijami si oglejte velikost te-

	Skupina 1	Skupina 2	p
Nizki nivo	15 (50%)	25 (92,6%)	<0,01
Srednji nivo	12 (40%)	2 (7,4%)	<0,01
Visoki nivo	3 (10%)	0 (0%)	0,09

Slika 1: *Primer uporabe χ^2 testa.*

stov. Na koncu poskušajte dodati razlago, ki bi bila primerna tudi za nestatistike.

Del 2: Zanima nas povezanost dveh binarnih neodvisnih slučajnih spremenljivk (X in Y). Ničelna domneva pravi, da sta deleža vrednosti $Y = 1$ enaka pri obeh skupinah glede na spremenljivko X. Za preverjanje te ničelne domneve lahko uporabimo test primerjave deležev ali test χ^2.

- Predstavite oba testa, poizkusite najti teoretično povezavo med testnima statistikama
- Primerjajte velikost in moč obeh testov na različnih velikostih vzorcev

11 Primerjava porazdelitve polimorfizmov *

Raziskovalec je primerjal porazdelitev nekega polimorfizma med bolniki (vzorec 31 bolnikov) in kontrolami (200 posameznikov). Njegova ničelna domneva je, da je porazdelitev med bolniki in kontrolami enaka. Na vzorcu je dobil naslednje številke:

	bolniki	kontrole
GG	10	32
GT	15	100
TT	6	68

Opazil je, da se alel G pojavlja precej bolj pogosto pri bolnikih kot pri kontrolah. Zato se je odločil, da bo preveril ali so te razlike statistično značilne. Preštel je vse alele G (pri bolnikih je dobil $35 = 10 + 10 + 15$) in T (pri bolnikih je dobil $27 = 15 + 6 + 6$) in nato primerjal pogostost s pomočjo testa χ^2.

Vaša naloga je, da ugotovite, kaj je narobe z njegovim pristopom ter problematiko tega pristopa ovrednotite s simulacijami. Pri načrtovanju simulacij se osredotočite predvsem na dejstvo, da pri njegovem pristopu enote niso neodvisne, vzorec je bil povečan na umeten način. Rezultate na koncu predstavite tako, da boste z njimi lahko prepričali raziskovalca, ki po svoji izobrazbi ni statistik.

12 Asimptotski z-test in aproksimacija s Studentovo porazdelitvijo **

Naj bodo $X_1, \ldots, X_n$ neodvisne, enako porazdeljene slučajne spremenljivke. Teorija pravi:

a) če so X_i porazdeljeni normalno $N(\mu, \sigma)$, velja

$$\frac{\bar{X} - \mu}{\sigma} \sim N(0,1) \text{ in } \frac{\bar{X} - \mu}{\widehat{\sigma}} \sim t_{n-1}$$

1. če X_i niso porazdeljeni normalno, nam centralni limitni izrek še vedno da aproksimacijo

$$\frac{\bar{X} - \mu}{\sigma} \sim N(0,1) \text{ in } \frac{\bar{X} - \mu}{\widehat{\sigma}} \sim N(0,1)$$

Centralni limitni izrek pove le porazdelitev, ko gre n proti neskončno, nas pa zanima kvaliteta aproksimacije na majhnih vzorcih. Zanima nas, ali ne bi morda porazdelitev t tudi v drugem primeru ponudila boljše aproksimacije.
V seminarski nalogi:

- Kratko povzemite relevantno teorijo.
- Generirajte vzorec X_i iz normalne porazdelitve. Ocenite σ z vzorca in izračunajte interval zaupanja s pomočjo porazdelitve z in t. Generirajte veliko število vzorcev in pokažite, da je pokritje s porazdelitvijo t bližje nominalni vrednosti (95%). Preizkusite nekaj različnih velikosti vzorcev.
- Nato generirajte vzorec iz neke druge porazdelitve ter ponovite gornji postopek. Zanima nas, ali je aproksimacija s t še vedno boljša izbira.

V zadnji točki izberite več različnih porazdelitev, npr. vse porazdelitve iz eksponentne družine. Poizkusite najti teoretične rezultate na to temo.
Dodatno: oglejte si, ali bi aproksimacija s t izboljšala lastnosti Waldovega testa v logistični regresiji.

13 Razmerje tveganj in razmerje obetov ***

Raziskovalci so zbrali podatke o vseh pacientih, ki so imeli med leti 2008-2012 neko bolezen, zanimale so jih neželene posledice te bolezni. Zbrali so 68 posameznikov in jih pregledali (podatki `pod1.r`). Zanimalo jih je, ali se pojavijo posledice in kako je to povezano s štirimi napovednimi spremenljivkami. Zanima jih, katero mero naj v ta namen ocenijo in kako naj se tega lotijo.

- Razložite, kaj je razmerje tveganj (RR) in kako ga ocenimo. Izpeljite varianco.
- Razložite, kaj je razmerje obetov (OR) in kako ga ocenimo. Izpeljite varianco. Razložite, kako sta povezana RR in OR.
- Razložite, kako testiramo ničelno domnevo v primeru OR in kako izračunamo intervale zaupanja
- Proučite metode, ki se uporabljajo za podajanje intervalov zaupanja in testiranje ničelne domneve v primeru RR. Različne možnosti primerjajte s simulacijami (nekaj možnosti npr. podaja R funkcija `epitab` iz knjižnice `epitool`).
- Ustrezno analizirajte podatke in interpretirajte rezultate.

14 Test t in test z za vzorce iz končnih populacij **

Pri vzorcih iz končne populacije vemo, da zaradi odvisnosti izračuni za varianco dobijo popravek. Zanima nas, kaj lahko v splošnem rečemo za test t in test z v primeru končne populacije. Za oba testa nas torej zanima:

- ali so predpostavke pri testu zadoščene in kakšne razmisleke ste pri tem naredili
- ali je velikost testa enaka $\alpha = 0{,}05$ in v katerih primerih

Preverite velikost testa za različne strategije in scenarije:

- Izmislite si nek konkreten primer populacije. Analizirajte učinek spremenljivke na celotni populaciji, nato vzemite vzorec, ponovite analizo na vzorcu in primerjajte.
  - Zgornji postopek ponovite večkrat in izračunajte velikost testa v tem primeru populacije.
- Generirajte podatke, tako da ustrezajo ničelni domnevi, predpostavite normalno porazdelitev. Naredite to tudi za različno velike populacije, da ugotovite, kakšen vpliv ima velikost populacije na končne rezultate.
  - Preverite, v kakšnem deležu simulacij zavrnete ničelno domnevo, če uporabite test z/brez popravka.
- Generirajte podatke tako, da vrednosti ne prihajajo iz normalne porazdelitve (še vedno drži ničelna domneva). Ugotavljajte vpliv velikosti populacije na končne rezultate.
  - Ponovite prejšnje simulacije, preverite delež simulacij, ki zavrnejo ničelno domnevo.

15 Logistična regresija ***

Kadar nas v regresiji zanima kot izid zanima le binarna spremenljivka, namesto linearne regresije uporabimo logistično. Recimo, da je odvisnost opravljenega izpita iz statistike (DA /NE) sledi modelu logistične regresije, v katerem sta neodvisni spremenljivki število ur učenja dnevno in doseženo število točk pri domačih nalogah.

- Razložite osnovno teorijo modela logistične regresije (le tisto, kar boste potrebovali pri naslednjih točkah).
- Simulirajte primer podatkov, ki sledijo modelu, in ga ilustrirajte. Zaradi poenostavitve naj bosta kovariati v modelu neodvisni med seboj
- Za vsako spremenljivko posebej bi radi preverili, ali je povezana z izidom. Zapišite ničelni domnevi in ju preverite z Waldovim testom ter testom posplošenega razmerja verjetij. Kratko predstavite oba testa in ju ilustrirajte na svojem primeru.
- S simulacijami primerjajte velikost in moč teh dveh testov.

16 Zbirateljstvo **

Cilj te naloge je konstruirati testno statistiko ter predlagati cenilke za parametre, ki nas zanimajo. Rešitve nalog lahko povsem temeljijo na simulacijah, če vam uspe del simulacij nadomestiti s teoritičnimi izpeljavami, to seveda naredite (vendar pa rezultate vseeno preverite s simulacijami).

Naloga:
Miha zbira sličice, v albumu je 200 različnih sličic. Kupil je že 300 sličic, trenutno ima 140 različnih.

- Predpostavite, da so vse sličice enako pogoste. Zanima nas pričakovano število različnih sličic glede na število kupljenih. S pomočjo simulacije izračunajte, kako število različnih sličic narašča s številom kupljenih. Prikazu dodajte še prikaz variabilnosti te količine.
- Miha bi rad preveril predpostavko, da so vse sličice enako pogoste. Zapišite ničelno domnevo, predlagajte testno statistiko in meje zavrnitve. Predlagani test uporabite na konkretnem primeru in interpretirajte rezultate .
- Proučite lastnosti testa predlaganega v prejšnji alineji - velikost testa, moč testa. Primerjajte z lastnostmi, ki bi jih dobili v situaciji, kjer bi bilo le 20 različnih sličic, Miha pa bi jih do sedaj kupil 30.
- Predpostavite, da je preostalih 60 sličic, ki Mihi še manjkajo, manj pogostih (manj kot 1/200, vendar vse enako). Predlagajte cenilko za to količino (verjetnost posamezne manj pogoste sličice) in utemeljite njene lastnosti.
- Na podlagi ocene iz prejšnje alineje ocenite, koliko novih sličic mora Miha kupiti, da bo verjetnost, da zbere vse, večja od 0,5.