Računarska statistika

Snježana Lubura Strunjak

Zagreb, 08. travnja 2021.

Monte Carlo eksperimenti

Simulacija i MC simulacija

"Simulacija je reprezentacija ponašanja nekog fizikalnog ili apstraktnog sustava ponašanjem nekog drugog sustava" (Ralston, 1976).
Simulacija se primjenjuje kada je opserviranje/eksperimentiranje sa originalnim sustavom:

- Opasno (npr. epidemije, nuklearne reakcije),
- Nemoguće (npr. globalno zagrijavanje, udar meteora),
- Skupo (optimalni oblik novog vozila), ili
- Presloženo da bi se ispitalo egzaktnim analitičkim alatima (npr. multivarijatne statističke distribucije);
- Ako želimo ispitati utjecaj različitih čimbenika na sustav (utjecaj novih pravila/zakona na pad stope kriminaliteta), itd.

Simulacija i MC simulacija

Simulacija može biti

- Deterministička ili
- Stohastička ili MC simulacija.

U MC simulaciji se barem 1 varijabla ponaša slučajno. Generiraju se skupovi slučajnih brojeva po nekim apriornim distribucijama i istražuju rezultati modela.

Monte Carlo (MC) simulacija

Područja primjene Monte Carlo metoda uključuju biologiju, kemiju,računarstvo, ekonometriju i financije, inženjerstvo, ostala područja znanosti: fiziku, društvene znanosti, statistiku, itd.
MC simulacija znači koristiti dani mehanizam po kome se generiraju podaci kao model procesa koji želimo razumjeti, iz koga želimo generirati nove uzorke, i razmatrati rezultate dobivene na tim uzorcima.
U statistici taj se proces zove ponavljano uzorkovanje.
Ako se ponovno uzorkuje

- po hipotetskoj distribuciji - "klasični" Monte Carlo eksperimenti i testovi
- po empirijskoj distribuciji - Bootstrap, Jackknife i ostale metode ponovnog uzorkovanja (re-samplinga)

U folderu Primjeri\Simulating data with SAS, program *Ch*04_*sim.sas*

Primjene simulacija u statistici

- Osnovne tehnike
  - Aproksimativna sampling distribucija (utjecaj veličine uzorka, utjecaj distribucije podataka, procjene (pristranost, std.pogreška itd,))
  - Procjene vjerojatnosti
  - Evaluacije statističkih tehnika (int. pouzdanosti i vjerojatnost pokrivanja, robusnost testova na odstupanja od pretpostavki, snaga testa, procjene p-vrijednosti (MC testovi))
- Primjene u statističkom modeliranju (MC eksperimenti)
  - Linearni regresijski modeli (sa 1 prediktorom, sa više prediktora, interakcije, polinomijalni modeli, outlieri, povezane opservacije, nenormalnost, heteroscedastičnost)
  - Generalizirani linearni modeli (logistička regresija poissonova regresija)
  - ltd.

Aproksimativna sampling distribucija

- Simuliraj mnogo uzoraka iz iste populacijske distribucije
- Za svaki uzorak, izračunaj statistiku od interesa (npr. prosj. vrijednost, median, broj proporcija, ...). Skup tih statistika čini aproksimativnu sampling distribuciju (kraće ASD).
- Na osnovu analize ASD-a donose se zaključci – o procjenama, int. pouzdanosti, vjerojatnosti, i dr.

MC procjena je prosječna vrijednost statistika iz drugog koraka.
MC procjena je procjenitelj θ (očekivane vrijednosti statistike od interesa).

Primjer (ASD za spljoštenost (kurtosis) uzorka γ_2(za normalne, t, exponencijalne i lognormalne podatke))

Program *CHAPTER*1_1_*ASD primjer za koeficijent spljostenosti*.*sas*

Pročitajte Chapter 4 iz knjige „Simulating Data with SAS“ (Simulating Data with SAS Ch4.pdf) do str. 65.(potrebno i za Zadatak 4 iz zadaće 3)

1. Generaj pseudo-slučajne brojeve po zadanim distribucijama (n=50) iz tablice,

Distribucije	γ_2
normalna	0
t_5	6
eksponencijalna	6
lognormalna (0,0.503)	6

2. Izračunaj γ_2,
3. Ponovi 1-2 1000 x za svaku od 4 distribucije,
4. Sumariziraj

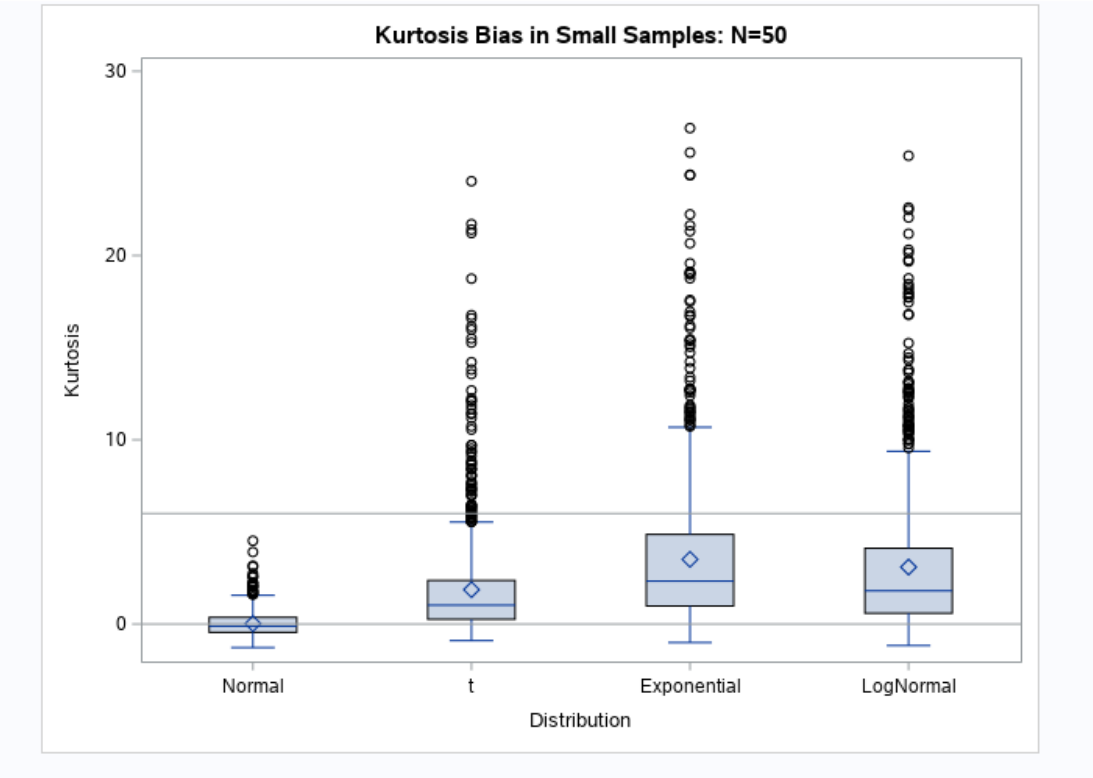
Napomena: γ_2 je „excess kurtosis” (tj. za normalnu distribuciju $\gamma_2 = 0$)

CHAPTER1_1_ASD primjer za koeficijent spljostenosti.sas

Program se sastoji od sljedećih dijelova:

- Definicija veličine uzorka i broja replikacija (macro varijable) (retci 3-4)
- Generiranje podataka (2 petlje, po replikacijama (&NumSamples) i broju opservacija u svakom uzorku (&N) (retci 5-16)
- Izračunavanje statistike od interesa (kurtosis) za svaku replikaciju (SampleID) i za svaku od 4 distribucije, spremanje u izlazni data set Moments (retci 18-22)
- Transpozicija (za svaki SampleID se vektor redak (1x4) transponira u vektor stupac (4x1)) tako da se data set Moments sa 1000 redaka i 4 stupaca transformira u data set Long sa 1000x4 redaka i 1 stupcem (+ dodatne ID varijable). Transpozicija je potrebna zbog grafičke procedure proc sgplot.
- Vertikalni paralelni boxplotovi za aproksimativne sampling distribucije statistike KURTOSIS (za 4 distribucije)

Slika: *CHAPTER1_1_ASD primjer za koeficijent spljostenosti.sas*



Uočite: Rezultati pokazuju da je kurtosis nepristran na uzorku od 50 samo kad uzorci dolaze iz normalne populacije (sredina sampling distribucije je otprilike jednaka populacijskoj vrijednosti γ_2 za normalnu distribuciju (0), dok su sredine sampling distribucija kod ostale 3 distribucije znatno niže od populacijske vrijednosti 6 za γ_2).

Procjene p-vrijednosti /Monte Carlo testovi

- Barnard (1963) je predložio metode "računarskog zaključivanja" (engl. computational inference): procjene kvantila test statistike, T pod nultom hipotezom primjenom MC metoda
- *nrep* slučajnih uzoraka iste veličine kao dani uzorak se generira pod nultom hipotezom. Test statistika se računa za svaki uzorak $(T_1^\star, T_2^\star, \ldots, T_{nrep}^\star)$, što čini uzorak test statistika.
- Empirijska CDF (funkcija kumulativne distribucije) uzorka test statistika se koristi kao procjena CDF test statistike.
- p-vrijednost opservirane test statistike se procjenjuje kao proporcija broja simuliranih vrijednosti $\geq$ i/ili $\leq$ od opservirane vrijednosti:

$$\text{p-vrijednost} = \frac{r}{nrep},$$

  gdje je r=broj simuliranih vrijednosti $\geq$ i/ili $\leq$ od opservirane vrijednosti.
- Ako je distribucija test statistike kontinuirana, tako definirana p-vrijednost je nepristrana.
- Obično se procjenjuje: $\text{p-vrijednost} = \frac{r+1}{nrep+1}$.

Procjene p-vrijednosti /Monte Carlo testovi

- Pretpostavke za primjenu: distribucija slučajne komponente pretpostavljenog modela mora biti poznata i mora biti moguće generirati slučajne uzorke iz te distribucije pod nultom hipotezom.
- Hope(1968), Marriott(1979): snaga MC testova može biti visoka čak i za razmjerno male vrijednosti nrep.

Simulacija po empirijskoj distribuciji

- Umjesto hipotetske vrijednosti, za parametar se uzima procjena iz uzorka.
- Uzorci se generiraju po procijenjenom modelu i na svakom uzorku se izračunavaju test statistike $(T_1^\star, T_2^\star, \ldots, T_{nrep}^\star)$, što čini uzorak test statistika.
- Empirijska CDF (funkcija kumulativne distribucije) uzorka test statistika se koristi kao procjena CDF test statistike.
- p-vrijednost opservirane test statistike se procjenjuje iz empirijske CDF.
- Distribucijska svojstva empirijske CDF u ovom slučaju NISU svojstva koja vrijede pod nekom nultom hipotezom, već su svojstva koja vrijede pod modelom sa parametrima koji odgovaraju procjenama iz podataka.
- Takav se tip pristupa zaključivanju zove parametarski bootstrap.

Monte Carlo eksperimenti

Model opisuje mehanizam po kome se generiraju podaci. Bolje razumijevanje modela može se postići na slijedeći način:

- Koristi model za simulaciju "umjetnih" podataka,
- Ispitaj podudaranje umjetnih podataka sa našim očekivanjima ili sa raspoloživim realnim podacima.
- Analiziraj podatke pomoću modela. Taj proces, koji je svojstvo računarske statistike, pomaže pri evaluaciji metoda analize. Pomaže nam u razumjevanju uloge pojedinih komponenti modela:
  - funkcionalne forme,
  - parametara,i
  - prirode stohastičke komponente. (J. Gentle, George Mason University)

Kada su MC studije neophodne?

- Kada teorijske pretpostavke statističke teorije nisu ispunjene - Npr. Ispitivanje posljedica neispunjenih teorijskih pretpostavki (robusnost) i
- Kada je statistička teorija nedovoljno razvijena ili je nema - Npr. Određivanje sampling distribucije statistike koja nema teorijsku distribuciju

Koraci u MC studiji

- Postavljanje pitanja koja se mogu ispitivati MC studijom
- Dizajn MC studije koja može dati odgovore na postavljena pitanja
- Generiranje podataka
- Implementacija kvantitativne/statističke tehnike od interesa (funkcija, procedura, macro, IML code)
- Izračunavanje i akumulacija statistike od interesa (u svakoj replikaciji)
- Analiza akumulirane statistike od interesa (vizualizacija i tablice)
- Donošenje zaključaka na osnovu empirijskih rezultata

Primjer: robustnost t-testa za 1 uzorak: Dizajn MC studije

Na osnovu postavljenih pitanja (npr. robusnost t statistike) identificiramo što utječe na sampling distribuciju t statistike:

- Veličina uzorka (1.faktor) - Npr. 10,20,50,100,200
- Distribucija (2. faktor) - Npr. Normalna, gamma, uniformna
- Broj replikacija (uzoraka) za svaku kombinaciju nivoa faktora - Npr. 10000

	Distribucije		
Veličina uzorka	normalna	gamma	uniformna
10	10000	10000	10000
20	10000	10000	10000
50	10000	10000	10000
100	10000	10000	10000
200	10000	10000	10000

Ukupno $3 \times 5 \times 10000 = 150000$ slučajnih brojeva.

Monte Carlo eksperimenti - treba ih koristiti samo kada analitičkim i numeričkim tehnikama ne možemo doći do odgovora.
Monte Carlo eksperiment (ima tri koraka, u našem primjeru gledamo robusnost t-statistike na odstupanja od normalnosti):

- Ulaz (dizajn, generiranje): Uzorkovanje iz normalne, gamma, itd., n= 10,20,50, 100, 200, br. replikacija= 10,000
- Model/modeli (statistička tehnika, izračunavanje, akumulacija): Računanje t statistike za svaku distribuciju, svaki n i svaku replikaciju, akumulacija vrijednosti
- Izlaz (Analiza rezultata i zaključci): Zaključci na osnovu sumarnih rezultata prikazanih u tablicama i grafikonima

Monte Carlo eksperimenti: Zašto ih koristiti

- Monte Carlo procjenjivanje
  - procjenjivanje odredjenog integrala
  - procjenjivanje varijance i pristranosti (bias)
- Monte Carlo testovi: simulacija podataka po hipotetskom modelu
- Bootstrap metode
  - parametarski bootstrap: simulacija podataka po procjenjenom (fitted) modelu
  - neparametarski bootstrap: ponovno uzorkovanje iz podataka

t test

t-statistika (t-test) se koristi za testiranja hipoteza o sredinama:

- Za 1 uzorak
- Za uparene uzorke
- Za 2 uzorka

Najjednostavnji slučaj: 1 uzorak.
MC eksperiment za ispitivanje robustnosti t-statistike (na odstupanje od uvjeta)

t test za 1 uzorak

Za testiranje nul hipoteze

$$H_0 : \mu = \mu_0$$

koristi se test statistika

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} \sim t(n-1),$$

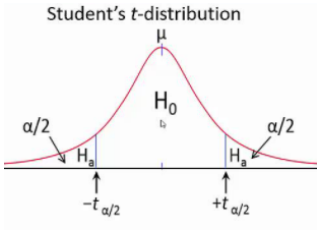
gdje je $\bar{x}$ aritmetička sredina uzorka, a $s_{\bar{x}}$ standardna pogreška ($s_{\bar{x}} = \frac{s_n}{\sqrt{n}}$).
Test može biti dvostrani ($H_1 : \mu \neq \mu_0$) ili jednostrani ($H_1 : \mu < \mu_0$ ili $H_1 : \mu > \mu_0$).
Pretpostavke za primjenu t testa:

- Normalnost podataka
- Ili normalnost sampling distribucije sredina (centralni granični teorem) - veliki uzorci
- Nezavisnost podataka

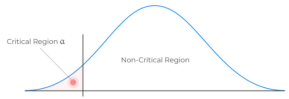
Pitanje: Koliko je t-test (odnosno distribucija t statistike) osjetljiv na odstupanja od normalnosti (robusnost)?

t-test: Kritično područje

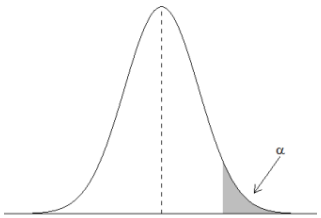
Slika: $H_1 : \mu \neq \mu_0$



Slika: $H_1 : \mu < \mu_0$



Slika: $H_1 : \mu > \mu_0$

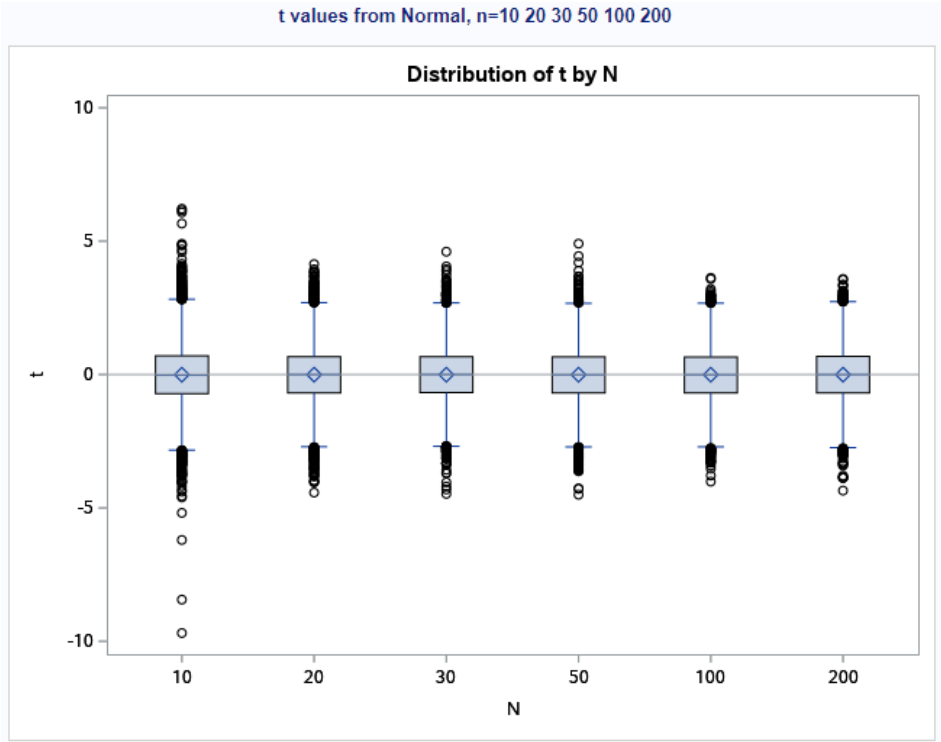


Procjenjivanje varijance i pristranosti (bias)

Primjer (Robusnost t statistike)
Program *CHAPTER*1_2_*T*_*NORMAL*1.*SAS*, *CHAPTER*1_2_*T*_*NORMAL*2.*SAS*, *CHAPTER*1_2_*T*_*NORMAL*4.*SAS*.
Ispitajte, pomoću Monte Carlo eksperimenta (i animacije) sampling distribuciju t statistike (za 1 populaciju) ako su podaci distribuirani po:
-Normalnoj distribuciji $N(0, 1)$
-Gamma distribuciji $Gamma(0.5, 1)$
Koristite slijedeće veličine uzoraka n=10,20,30,50,100,200.
Usporedite momente dobivene simulacijom sa teoretskim momentima t distribucije.
Diskutirajte posljedice primjene t-testa u slučaju jako zakrivljene (asimetrične) originalne distribucije.

Robusnost t statistike - normalna distribucija

Slika: *CHAPTER1_2_T_NORMAL4.SAS*



The MEANS Procedure

Analysis Variable : t						
N	N Obs	Mean	Std Dev	Std Error	Skewness	Kurtosis
10	10000	-0.006	1.135	0.011	-0.014	1.792
20	10000	-0.004	1.055	0.011	0.031	0.458
30	10000	-0.007	1.021	0.010	-0.032	0.286
50	10000	-0.008	1.016	0.010	-0.031	0.303
100	10000	-0.008	1.001	0.010	-0.027	-0.020
200	10000	-0.004	1.011	0.010	-0.024	0.053

Zadatak (Robusnost t statistike (eksploracija))

Koristite program *CHAPTER1_2_T_NORMAL4.SAS* Izvedite program, pa pogledajte dataset TALL.

Analizirajte distribucije sredina i t statistika (varijable MEAN I T) pojedinačno po veličinama uzoraka N. Dodajte qqplot.

Što uočavate? Kolike su sredine i standardne devijacije od MEAN, za pojedinačne N, a kolike su očekivane vrijednosti standardnih devijacija sredina (st.grešaka sredina) na osnovu centralnog graničnog teorema?

Zadaće

5. zadaća (7.zadatak napraviti nakon idućih predavanja): rok za predaju 22.04.
Zadaća se nalaze u folderu Zadaće na MERLINU.
UPUTE: Svaki zadatak iz zadaće mora biti u svom .sas programu. Sve .sas programe nazovite na način *prezime_ime_zad*1.*sas*, ako je npr. 1.zadatak u pitanju, itd. Sve što radite u zadaćama mora biti u obliku koda (možete koristiti sve dostupne materijale da dobijete tražene rezultate, ali sve mora biti napisano u obliku koda).