Računarska statistika

Snježana Lubura Strunjak

Zagreb, 15. travnja 2021.

Gamma(k,β) distribucija - podsjetnik

Neka osnovna svojstva $Gamma(k, \beta)$ razdiobe (k=0.5,β=1):

- mean (sredina)=$\mu = \frac{k}{\beta} (= 0.5)$
- standard deviation (standardna devijacija)=$\sigma = \frac{\sqrt{k}}{\beta} (= 0.707)$
- skewness (asimetrija)=$\gamma_1 = \frac{2}{\sqrt{k}} (= 2.83)$
- kurtosis (spljoštenost)=$\gamma_2 = \frac{6}{k} (= 12)$

Zadatak (Robusnost t statistike - nastavak)

Program *CHAPTER*1_2_*T*_*GAMMA*2.*SAS*, *CHAPTER*1_2_*T*_*GAMMA*2_*no*_*anim*.*SAS*, *CHAPTER*1_2_*T*_*GAMMA*4.*SAS*.

Ispitajte, pomoću Monte Carlo eksperimenta (i animacije) sampling distribuciju t statistike (za 1 populaciju) ako su podaci distribuirani po:

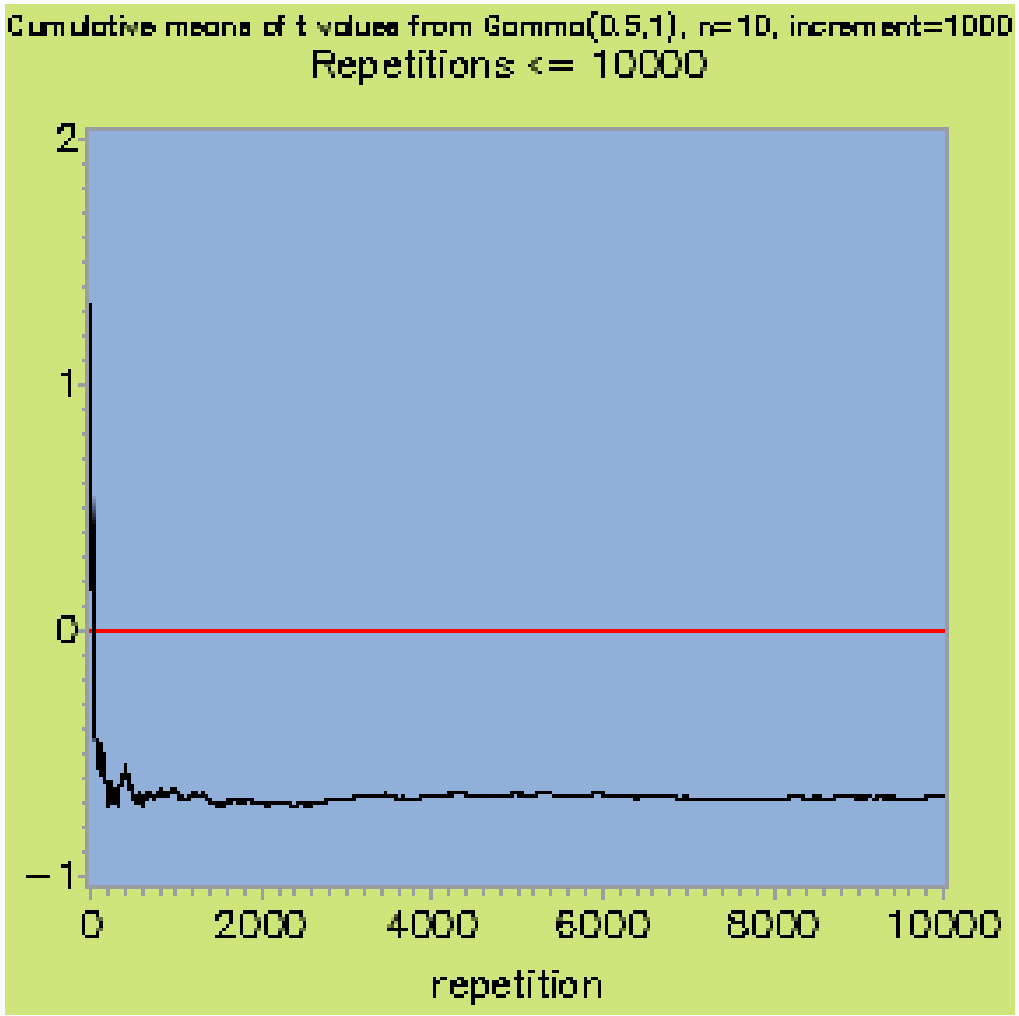
-Gamma distribuciji $Gamma(0.5, 1)$

Koristite slijedeće veličine uzoraka n=10,20,30,50,100,200.

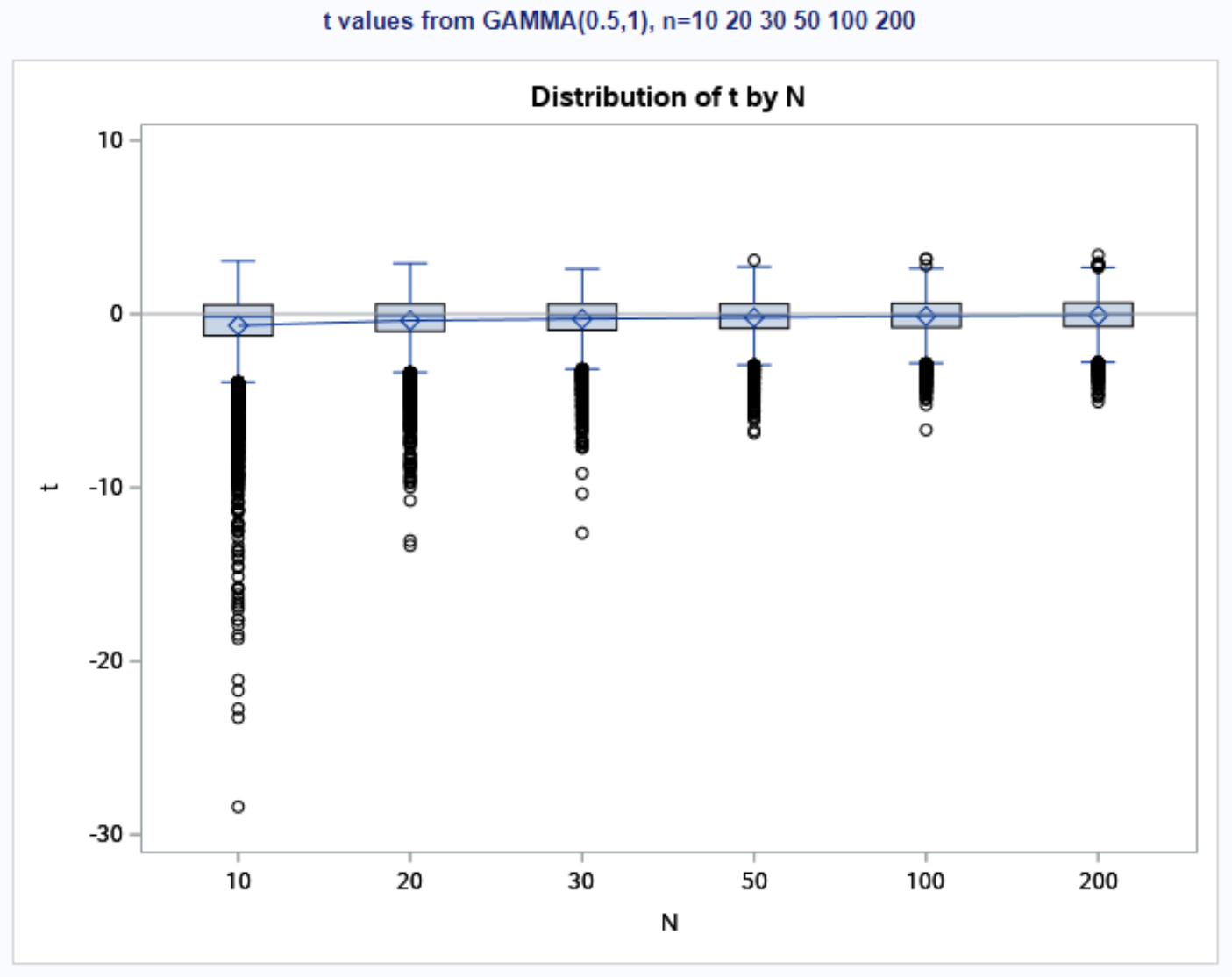
Usporedite momente dobivene simulacijom sa teoretskim momentima t distribucije.

Diskutirajte posljedice primjene t-testa u slučaju jako zakrivljene (asimetrične) originalne distribucije.

Slika: *CHAPTER1_2_T_GAMMA2.SAS*

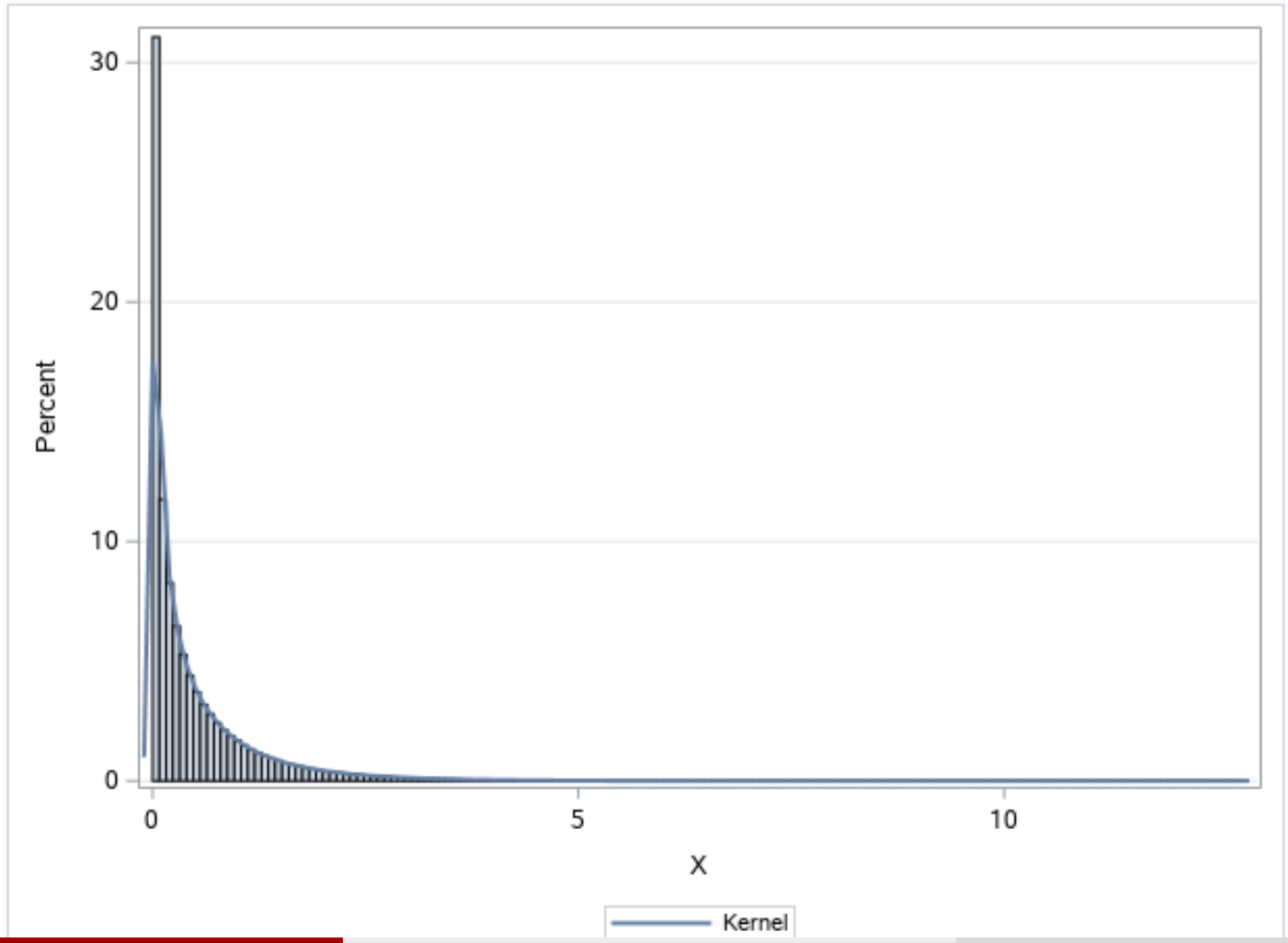


Slika: *CHAPTER1_2_T_GAMMA4.SAS*

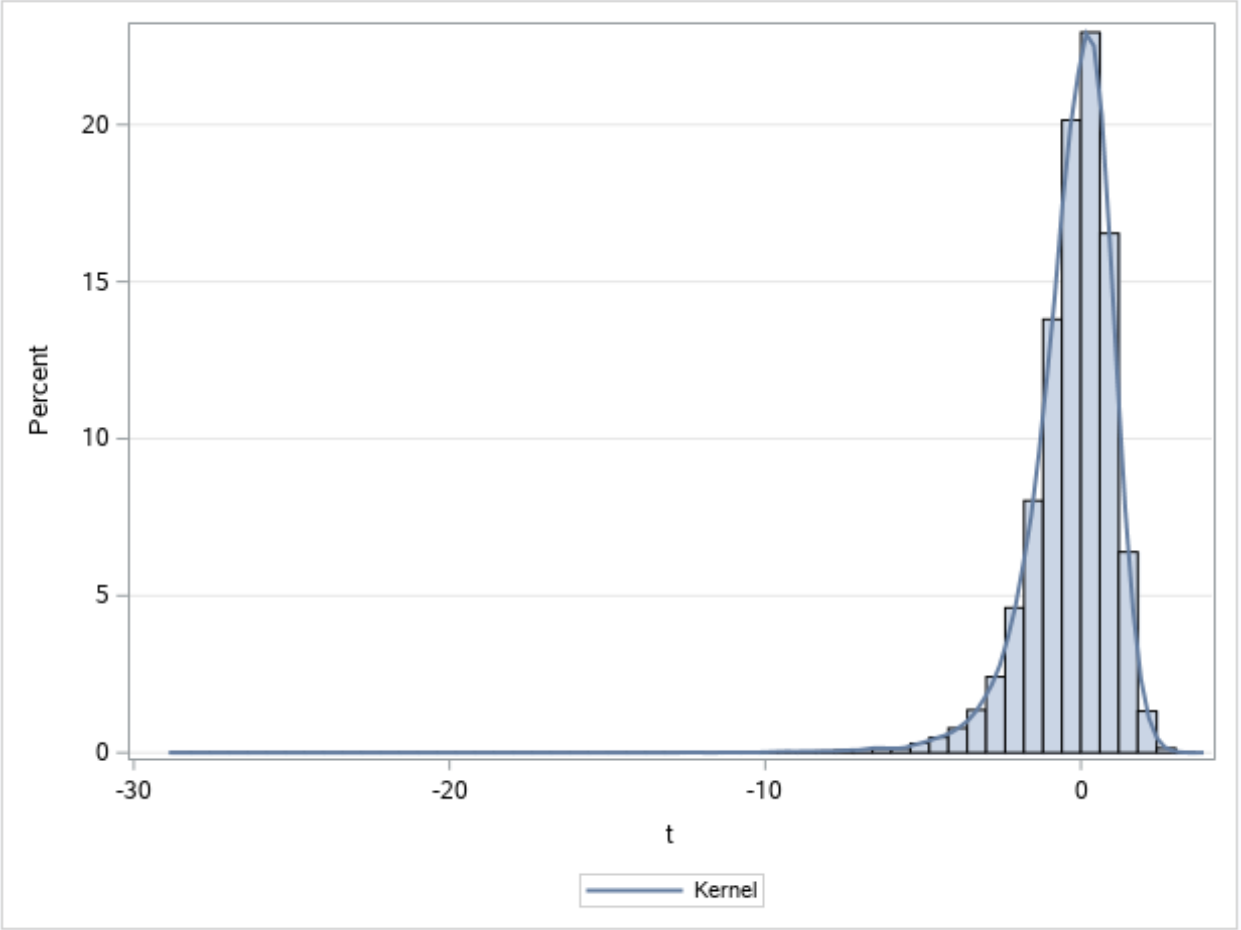


Pogledajmo histograme generiranih podataka po gamma razdiobi i t-statistika u tom slučaju:

Slika: *CHAPTER1_2_T_GAMMA4.SAS*-histogram generiranih podataka po Gamma distribuciji



Slika: *CHAPTER1_2_T_GAMMA4.SAS*-histogram t statistika



Uočimo: histogram podataka ima rep na desnoj strani, a histogram t statistika rep na lijevoj. To nije slučajno.

Razvoj u red potencija momenata t statistike

Ako je γ_1 skewness originalne populacije, a $\gamma_1(t)$ t-statistike, onda vrijedi:

- $\mathbb{E}(t) = -\frac{\gamma_1}{2\sqrt{n}} + O(n^{-\frac{3}{2}})$
- $Var(t) = 1 + n^{-\frac{1}{2}}(2 + \frac{7}{4}\gamma_1^2) + O(n^{-\frac{1}{2}})$
- $\gamma_1(t) = -2\frac{\gamma_1}{\sqrt{n}} + O(n^{-\frac{3}{2}})$

Uočimo: asimetrija t statistike je suprotnog smjera (predznaka) od asimetrije originalne populacije. Dakle, u našem primjeru, $\mathbb{E}(t) \approx -\frac{\gamma_1}{2\sqrt{n}} \approx -\frac{2.83}{2\sqrt{n}}$ i $\gamma_1(t) \approx -2\frac{\gamma_1}{\sqrt{n}} \approx -2\frac{2.83}{\sqrt{n}}$, što za različite n iznosi

n	10	20	30	50	100	200
$\mathbb{E}(t)$	-0.45	-0.32	-0.26	-0.20	-0.14	-0.10
$\gamma_1(t)$	-1.8	-1.3	-1.0	-0.8	-0.6	-0.4

Slika: *CHAPTER1_2_T_GAMMA4.SAS*

The MEANS Procedure

Analysis Variable : t						
N	N Obs	Mean	Std Dev	Std Error	Skewness	Kurtosis
10	10000	-0.661	1.998	0.020	-3.188	19.860
20	10000	-0.380	1.421	0.014	-1.677	5.696
30	10000	-0.290	1.282	0.013	-1.327	3.874
50	10000	-0.208	1.157	0.012	-0.882	1.507
100	10000	-0.125	1.063	0.011	-0.570	0.711
200	10000	-0.078	1.026	0.010	-0.365	0.363

Robusnost t statistike - gamma razdioba (Monte Carlo studija t- statistike)

Zadatak (Tablica proporcija generiranih t vrijednosti izvan kritičnih vrijednosti)
Koristite program *CHAPTER*1_2_*T*_*GAMMA*3.*SAS*.
Prilagodite program na slijedeći način:

- Postavite broj replikacija (nrep) na 10000 za sve vrijednosti

$$n = 10, 20, 30, 50, 100, 200,$$

- Izračunajte broj replikacija za koje t vrijednost pada izvan kritične vrijednosti t distribucije za $\alpha = 0.01, 0.025$ i 0.05, tj. procijenite $\mathbb{P}(t \leq t_{0.01}), \mathbb{P}(t \geq -t_{0.01}), \mathbb{P}(t \leq t_{0.025}), \mathbb{P}(t \geq -t_{0.025}), \mathbb{P}(t \leq t_{0.05}), \mathbb{P}(t \geq -t_{0.05})$. (Napomena: $-t_\alpha = t_{1-\alpha}$. U SAS-u $t_\alpha = TINV(0.01, n-1)$;)

Izvedite program. Što uočavate? Usporedite te procjene sa odgovarajućim α vrijednostima.
(Nakon što napišete sami program usporedite ga s rješenjem *CHAPTER*1_2_*T*_*GAMMA*3_*FRACTION*_*CRITICAL*_*VALUES*.*SAS* u folderu EXERCISES.)
Izvedite program *CHAPTER*1_2_*T*_*GAMMA*3_*FRACTION*_*CRITICAL*_*VALUES grafikon usporedbe gustoca vjerojatnosti*.*SAS* iz foldera EXERCISES i pogledajte slaganje teoretske *t* distribucije i empirijske dobivene pomoću gamma razdiobe.

Zadatak (Robusnost t statistike (eksploracija - gamma i uniformna distribucija))
Koristite program *CHAPTER1_2_T_GAMMA4.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?
- b)Izvedite istu analizu sa UNIFORMnom distribucijom (Uputa: u gornjem programu promijenite poziv GAMMA generatora u:
  $X = RAND(\text{"}UNIFORM\text{"});$
  $XT = X - \μ$
  Prilagodite naslove. Izvedite program, i koristeći dataset TALL pogledajte distribucije t statistike po veličinama uzoraka N.
  Što uočavate?

t test za male uzorke

Zadatak (t-test BMI (body-mass index) indeksa za 16 pacijenata)
Program *CHAPTER1_2_T_1SAMPLE.SAS*
Izvedite program. U datasetu *body_mass_index* nalaze se vrijednosti BMI za 16 pacijenata uključenih u studiju dijabetičara. Prije pocetka studije treba ispitati da li je BMI na tom uzorku konzistentan sa ranije nadjenom vrijednosti indeksa (BMI=31.9). Možemo li koristiti t-test?
Uputa: koristite Tasks and Utilities.

Monte Carlo testovi: simulacije podataka po hipotetskom modelu

- Za testiranje hipoteze da podaci čine slučajni uzorak iz specificiranog (hipotetskog) modela / populacije.
- Algoritam:
  1. Simuliraj uzorke iz specificirane populacije
  2. Usporedi vrijednosti test statistike za simulirane uzorke sa test statistikom sa stvarnog uzorka.
- Od posebne je vrijednosti u situacijama kada je (originalna) populacijska distribucija poznata, ali sampling distribucija test statistike nije poznata u analitičkoj formi.

Monte Carlo testovi: simulacije podataka po hipotetskom modelu

Primjer (Uniformno uzorkovanje bez nadomjeska/ponavljanja)
Program *CHAPTER1_2_HYPOTHESIS.SAS*
Scenarij: Proizvodna tvrtka primi pošiljku od novog dobavljača sa 1000 jedinica nekog proizvoda. Dobavljač je prethodno prihvatio ugovor po kojem 98% jedinica u svakoj pošiljci mora biti bez nesukladnosti (defekata). Slučajan uzorak od 100 jedinica je izvučen (bez nadomjeska/vraćanja) iz pošiljke. Svaka od 100 jedinica je testirana i na 4 su pronađene nesukladnosti (4%).
Da li je stopa nesukladnosti u stvari veća od 2%?
Izvedite Monte Carlo test (i nakon toga primijenite i točnu hipergeometrijsku distribuciju.)

Uniformno uzorkovanje bez nadomjeska (bez ponavljanja)

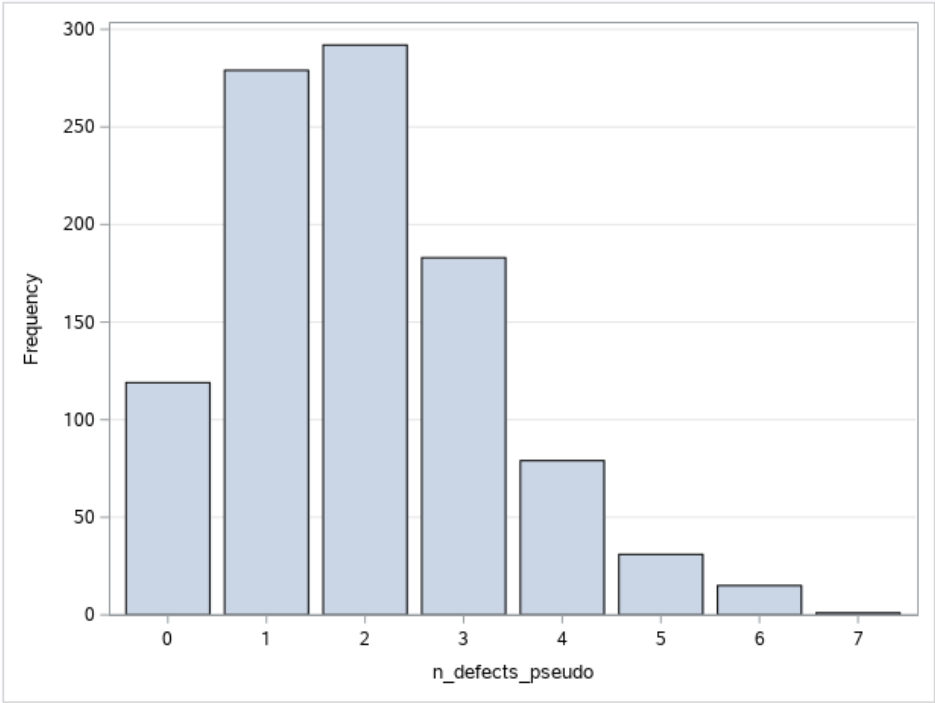
Monte Carlo uzorkovanje iz definirane (hipotetske) populacije:

- "Konstruirajte" reprezentaciju populacije (tj. generirajte 20 jedinica (nesukladnih jedinica) i 980 nula (sukladnih jedinica).
- Ponovite *NREP* puta:
  - Izvucite jednostavan slučajan uzorak (bez nadomjeska tj. bez ponavljanja, veličine 100) iz konstruirane populacije,
  - izračunajte test statistiku (broj nesukladnih jedinica),
  - usporedite je sa test statistikom na stvarnom uzorku.
- Izračunajte $p = \frac{N_{GRE}+1}{NREP+1}$, gdje je
  - N_{GRE} broj uzoraka na kojima je vrijednost test statistike $\geq$ od test statistike na stvarnom uzorku, a
  - NREP je broj izvučenih uzoraka.

Primjer (Uniformno uzorkovanje bez nadomjeska/ponavljanja)

Rezultati:

Slika: *CHAPTER1_2_HYPOTHESIS.SAS*



Primjer (Uniformno uzorkovanje bez nadomjeska/ponavljanja)

Slika: *CHAPTER1_2_HYPOTHESIS.SAS*

Obs	_TYPE_	_FREQ_	N_GRE	p	p_exact
1	0	999	126	0.127	0.13095

Zadaća

6. zadaća: rok za predaju 13.05.
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).