KOLOKVIJ 1A IZ RAČUNARSKE STATISTIKE

PMF-Matematika, Školska godina 2018/2019

Predavač: Doc.dr.sc. Vesna Lužar-Stiffler

Datum: _____

Ime, Prezime:_____

Potpis: _____

1)

a) Ako slučajna varijabla X slijedi lognormalnu distribuciju, a parametri μ i σ su sredina i standardna devijacija od ln(X), tada se X može izraziti kao

$X = \exp(\mu + \sigma Z)$,

gdje je Z standardna normalna varijabla ($Z \sim N(0,1)$).

Sredina, varijanca i koeficijent asimetrije lognormalne distribucije mogu se izraziti uz pomoč μ i σ na slijedeći način:

$E(X) = \exp(\mu + \sigma^2/2)$

$Var(X) = (\exp(\sigma^2) - 1) \exp(2\mu + \sigma^2)$

$Skewness(X) = (\exp(\sigma^2) + 2)(\exp(\sigma^2) - 1)^{1/2}$

Izračunajte E(X), Var(X) i Skewness (X) (i zaokružite na 3 decimalna mjesta),

ako je $\mu = 2$ i $\sigma = 1.5$

E(X)= _____

Var(X) = _____

Skewness (X) = _____

b) Generirajte niz od 25 pseudoslučajnih brojeva koji slijede lognormalnu distribuciju koristeći gore navedenu vezu između standardne normalne varijable Z i lognormalne varijable X, te uz $\mu = 2$ i $\sigma = 1.5$

Taj isti postupak ponovite 1000 puta. Uz pomoć procedure MEANS za svaku od 1000 replikacija odredite vrijednosti sredine (MEAN), mediana (MEDIAN), standardne devijacije (STD), varijance(VAR), koeficijenta asimetrije (SKEWNESS), t-vrijednosti za testiranje hipoteze da je sredina populacije =0 (H0: $\mu=0$) (T) i odgovarajuće p-vrijednosti (PROBT) za varijablu
X – E(X).

Na osnovu generiranih podataka odredite empirijsku procjenu snage t-testa da odbaci hipotezu da je sredina populacije =0 , na razini statističke značajnosti $\alpha=0.05$ (tj. odredite proporciju p-vrijednosti (PROBT) manjih ili jednakih 0.05)

Empirijska procjena snage testa = _____(zaokruženo na 3 decimalna mjesta)

95% interval pouzdanosti snage testa = _____
(zaokruženo na 3 decimalna mjesta)

c) Odredite prosječnu vrijednost i standardnu pogrešku sredine, mediana i varijance (tj. prosječnu vrijednost i standardnu devijaciju sredina (MEAN), mediana (MEDIAN) i varijanci (VAR) generiranih podataka):

MEAN(MEAN) = _____STD (MEAN) = _____
(zaokruženo na 3 decimalna mjesta)

MEAN(MEDIAN) = _____STD (MEDIAN) = _____ (zaokruženo na 3 decimalna mjesta)

MEAN(VAR) = _____STD (VAR) = _____
(zaokruženo na 3 decimalna mjesta)

d) Na osnovu simuliranih podataka pod nultom hipotezom (H0: $\mu=0$) u dijelu b), procijenite pogrešku I reda tj. izračunajte broj i udio replikacija za koje t vrijednost pada izvan kritične vrijednosti t distribucije za α =0.01, 0.025 i 0.05, tj. procijenite $Pr(t \leq t_{0.01})$, $Pr(t \geq -t_{0.01})$, $Pr(t \leq t_{0.025})$, $Pr(t \geq -t_{0.025})$, $Pr(t \leq t_{0.05})$, $Pr(t \geq -t_{0.05})$.

$Pr(t \leq t_{0.01})$ (fraction_crit_01_left): _____ (zaokruženo na 3 decimale)

$Pr(t \geq -t_{0.01})$ (fraction_crit_01_right): _____ (zaokruženo na 3 decimale)

$Pr(t \leq t_{0.025})$ (fraction_crit_025_left): _____ (zaokruženo na 3 decimale)

$Pr(t \geq -t_{0.025})$ (fraction_crit_025_right): _____ (zaokruženo na 3 decimale)

$Pr(t \leq t_{0.05})$ (fraction_crit_05_left): _____ (zaokruženo na 3 decimale)

$Pr(t \geq -t_{0.05})$ (fraction_crit_05_right): _____ (zaokruženo na 3 decimale)

NAPOMENA:

U 1 b) koristite početnu vrijednost **SEED**= 37749 (call streaminit(SEED) i RAND proceduru).

UPUTA:

Za određivanje vrijednosti sredine (MEAN), varijance (VAR), mediana (MEDIAN), STD, SKEWNESS, t-vrijednosti za testiranje hipoteze da je sredina populacije =0 (H0: $\mu=0$) (T) i odgovarajuće p-vrijednosti (PROBT) koristite proceduru MEANS:

```
proc means data= imeulaznedatoteke nway noprint  ;
 var imevarijable;
 by rep;
 output out= imeizlaznedatoteke mean=mean median=median var=var std=std
skewness=skewness t=t probt=probt;
 run;
```

Za određivanje 95% intervala pouzdanosti snage testa („EXACT CONF.LIMITS“) koristite proceduru FREQ:

```
proc freq data= imeulaznedatoteke;
      table imevarijable /binomial(level='1');
/*** za određivanje proporcije (i 95% interval pouzdanosti) vrijednosti “1”
kategorijske varijable “imevarijable” ***/
  run;
```

2. U data setu STUDIJA su zapisani rezultati mjerenja krvnog testa za dvije skupine pacijenata (1 i 2), nakon što su primili 2 različita tretmana. Eksperiment je proveden tako da su pacijenti nasumično pridruženi skupini 1 ili 2.

```
data  studija;
 input skupina   rezultat @@;
 datalines;
2 4.3 2 4.9 2 3.6 2 4.2 2 5.1 2 4.6 2 3.8
1 5.1 1 4.7 1 5.6 1 4.9 1 5.0 1 4.9 1 5.3 1 5.4 1 5.5
;
run;
```

a) Primjenom TTEST procedure testirajte hipotezu o jednakosti varijanci za varijablu REZULTAT po vrstama tijesta. Ako se hipoteza može odbaciti na razini značajnosti α=0.05, onda primijenite odgovarajuću transformaciju (tj. transformaciju nakon koje se hipoteza o jednakosti varijanci ne može odbaciti na razini α=0.05).

b) Aproksimativnim randomizacijskim testom (koristeći 300 permutacija) ispitajte neovisnost REZULTATA o SKUPINI (1, 2). Za test statistiku koristite apsolutnu vrijednost razlike srednjih vrijednosti grupa (tj. skupina).

   p-vrijednost=_____ (zaokruženo na 3 decimalna mjesta)

c) Aproksimativnim randomizacijskim testom (koristeći 300 permutacija) ispitajte neovisnost REZULTATA o skupini (1, 2). Za test statistiku koristite t vrijednost za testiranje hipoteze o razlici srednjih vrijednosti grupa („POOLED t-test za nezavisne uzorke“).

p-vrijednost=_____ (zaokruženo na 3 decimalna mjesta)

d) Usporedite rezultate randomizacijskog testa (p-vrijednost i 95% interval pouzdanosti za t-vrijednost) iz c) sa rezultatima t-testa (PROC TTEST) iz a). UPUTA: 95% interval pouzdanosti za randomizacijsku metodu odredite „percentilnom“ metodom (tj. odredite 2.5% i 97.5% percentil randomiziranih vrijednosti POOLED-t test statistike).

95% interval za POOLED-t (iz c):_____ (zaokruženo na 3 decimalna mjesta),

POOLED-t vrijednost za standardni t-test (iz PROC TTEST iz a):_____ (zaokruženo na 3 decimalna mjesta)

Da li obje metode (a i c) dovode do istog zaključka/odluke na razini značajnosti α=0.05 tj da li se vrijednost t-a iz a) nalazi izvan ili unutar 95% intervala pouzdanosti određenog randomizacijskom metodom? _____

UPUTE:

Prije proc ttest naredbe (u dijelu c) napišite i izvedite naredbu:

ODS graphics off; ili definirajte i pozovite macro ODSOFF.

Nakon ttest procedure pozovite macro ODSON.

U 2 b) i c) koristite početnu vrijednost **SEED= 12543.**

UPUTA:
Pod c) koristite formulu za „pooled“ t-test za nezavisne uzorke. Koristite ttest proceduru:

```
ods output ttests=ttests;
proc ttest data= ImeDataseta ;
var ImeVarijable;
class ImeGrupe;
run;
data means_t;
set ttests;
where upcase(method)="POOLED";
keep tvalue;
run;
```

odnosno (za randomizirane vrijednosti):

```
ods output ttests=ttests;
proc ttest data= approxperm ;
var ImeVarijable;
class ImeGrupe ;
by rep;
run;

data means_perm_t;
set ttests;
where upcase(method)="POOLED";
keep tvalue;
```

```
rename tvalue=tvalue_perm;
run;
```

Za d) koristite Jmp ili proc univariate

3) U datoteci FIT su navedeni podaci o težini i obimu struka (varijable WEIGHT i WAIST, u funtama i inčima) članova muškog spola fitness kluba.

```
data FIT;
   input Weight Waist ;
   datalines;
191  36
189  37
162  35
189  35
182  36
211  38
167  34
176  31
154  33
169  34
166  33
154  34
247  46
193  36
202  37
157  32
156  33
138  33
161  31
;
run;
```

Procijenite 90% interval pouzdanosti za median, 25. i 75. percentil za varijablu ***WAIST*** na slijedeći način:

i. Definirajte statističku populaciju primjenjujući karakteristike uzorka, tj. sredinu (MEAN), standardnu devijaciju (STD), koeficijent asimetrije (SKEWNESS) i spljoštenosti (KURTOSIS) kao populacijske parametre. Vrijednosti statistika (MEAN, STD, SKEWNESS i KURTOSIS) zaokružite na 2 decimalna mjesta.

ii. Generirajte ponavljano uzorke veličine 19 iz populacije definirane pod a). Generirajte nrep=300 uzoraka. U svakom ponavljanju izračunajte statistike od interesa (MEDIAN, 25. i 75. percentil).

iii. Procijenite percentilnom metodom 90% interval pouzdanosti za median, 25. I 75. percentil) tako da odredite 5. I 95. percentil za 300 vrijednosti mediana, 25. i 75. percentila iz koraka ii.

a) 90% interval pouzdanosti za median varijable ***WAIST*** (percentilnom metodom): _____,_____ (zaokruženo na 3 decimalna mjesta)

b) 90% interval pouzdanosti za 25.percentil varijable ***WAIST*** (percentilnom metodom): _____,_____ (zaokruženo na 3 decimalna mjesta)

c) 90% interval pouzdanosti za 75.percentil varijable ***WAIST*** (percentilnom metodom): _____,_____ (zaokruženo na 3 decimalna mjesta)

NAPOMENA: Za **SEED** koristite vrijednost 12456. UPUTA: Koristite proceduru MEANS.