

DRUGI KOLOKVIJ IZ STATISTIČKOG PRAKTIKUMAZadatak [40 bodova]

Baza podataka `boston.xls` sadrži dio rezultata istraživanja iz 1978. godine o kvaliteti stanovanja u 506 Bostonskih kvartova i okolnih mjesta koja gravitiraju Bostonu. Između ostalih, baza sadrži sljedeće varijable:

varijabla `CHAS` sadrži informaciju o tome je li naselje na obali rijeke Charles ili nije (1 - na obali je, 0 - nije na obali)

varijabla `RM` sadrži prosječan broj prostorija u stambenoj jedinici (kuća ili stan)

varijabla `RAD` sadrži indeks pristupačnosti auto ceste iz naselja

varijabla `MEDV` sadrži medijan cijena stambenih jedinica u privatnom vlasništvu 1978. godine (u tisućama dolara)

simulirana varijabla `MEDVnova` sadrži medijan cijena stambenih jedinica u privatnom vlasništvu 1990. godine (u tisućama dolara).

Nadalje, simulirana varijabla `MEDVEU` sadrži medijan cijena u tisućama dolara stambenih jedinica u privatnom vlasništvu u europskom gradu koji se po veličini, strukturi stanovnika i kvaliteti života podudara s Bostonom, za 1978. godinu.

Odgovorite na sljedeća pitanja i riješite sljedeće zadatke:

1. Kojim tipovima slučajnih varijabli biste modelirali varijable `CHAS` i `MEDV`? Obrazložite svoje odgovore. (2 boda)
2. Navedite procjenitelje za distribuciju i funkciju distribucije slučajne varijable. Procijenite vjerojatnost da je slučajno odabrana nekretnina iz promatrane populacije smještena na obali rijeke Charles (varijabla `CHAS`) te vjerojatnost da je za takve nekretnine medijan cijena veći od 20000 dolara (varijabla `MEDV`). (8 bodova)
3. Navedite procjenitelja za očekivanje slučajne varijable te navedite i interpretirajte njegova najvažnija asimptotska svojstva. Možete li na razini značajnosti $\alpha = 0.05$ tvrditi da je očekivani medijan cijena stambenih jedinica u privatnom vlasništvu (varijabla `MEDV`) manji od 25000 dolara? (10 bodova)
4. Interpretirajte kutijasti dijagram na bazi medijana za varijablu `MEDV` te definirajte mjere deskriptivne statistike koje on prikazuje. Možete li na razini značajnosti $\alpha = 0.05$ tvrditi da je medijan cijena stambenih jedinica u privatnom vlasništvu veći od 21000 dolara? (10 bodova)
5. Procijenite distribuciju slučajne varijable kojom modeliramo medijan cijena stambenih jedinica u privatnom vlasništvu (varijabla `MEDV`). Možete li na nivou značajnosti $\alpha = 0.05$ tvrditi da ta slučajna varijabla nije normalno distribuirana? Navedite koje ste statističke testove koristili. (10 bodova)
6. Možete li na nivou značajnosti $\alpha = 0.05$ tvrditi da je očekivani medijan cijena stambenih jedinica u privatnom vlasništvu 1978. godine manji od onog 1990. godine? Istu usporedbu napravite za očekivane medijane cijena 1978. godine u Bostonu i u spomenutom europskom gradu. Navedite koje ste statističke testove koristili i obrazložite zašto. (10 bodova)
7. Odredite razlike medijana cijena iz 1990. i medijana cijena iz 1978. godine i te vrijednosti kategorizirajte na sljedeći način:
 - (-1) - medijan cijena 1990. manji je od medijana cijena 1978. godine ($MEDVnova - MEDV < 0$)
 - 1 - medijan cijena 1990. veći je ili jednak medijanu cijena 1978. godine ($MEDVnova - MEDV \geq 0$).

Možete li na nivou značajnosti $\alpha = 0.05$ tvrditi da promjena medijana cijena ovisi o tome nalazi li se nekretnina na obali rijeke Charles ili ne (varijabla CHAS)? Koji ste statistički test koristili i zašto? (10 bodova)

8. Napravite jednostavan linearni regresijski model koji opisuje promjenu medijana cijene stambenih jedinica u privatnom vlasništvu za 1978. i 1990. godinu (varijable MEDV i MEDVnova). Koju ćete od ovih varijabli promatrati kao neovisnu, a koju kao ovisnu varijablu? Odredite procjenu regresijskog pravca te odgovorite na sljedeća pitanja:

- a) Što o linearnom regresijskom modelu možete reći na temelju analize reziduala? Ima li smisla vršiti daljnja zaključivanja o modelu i interpretirati model? Ako ima, odgovorite na pitanja pod b) i c).
- b) Što o linearnom regresijskom modelu možete reći na temelju koeficijenta smjera procijenjenog regresijskog pravca?
- c) Koliki je dio promjena u izmjerenim vrijednostima ovisne varijable objašnjen linearnim regresijskim modelom?

(20 bodova)