Kommentaarid
Iseseisvate tööde lahendustest võiksite tekitada tekstidokumendi (näit. *Wordis*), kuhu on kopeeritud küsimus ja selle järgi vastus (koos vajalike tabelite ja/või joonistega ning kommentaaridega).
Kuigi küsimused on koostatud praktikumides *R*-i abil uuritud andmestike ja teostatud analüüside baasil, võib analüüsimiseks kasutada ka mõnd teist tarkvara (miskit saab *Exceliski* ära teha, või siis *Statistica*-s või *SPSS*-s või …).

Iseseisev töö 1

Andmestik: http://www.eau.ee/~ktanel/DK_0007/studentsR.csv
või *Exceli* failina: http://www.eau.ee/~ktanel/DK_0007/studentsR.xls

Ülesanded

1.1. Leia keskmine ja mediaan tunnusele peaümbermõõt sõltuvalt matemaatika hindest.

1.2. Kas mannaputru söövate tudengite kehakaalude varieeruvus on suurem, kui mannaputru mitte söövatel tudengitel?

1.3. Kuidas tudengid jaotuvad matemaatika hinde lõikes – tee sagedustabel ja tulpdiagramm.

1.4. Kas erialati on matemaatika hinnete suhtelised sagedused erinevad?

1.5. Konstrueeri sektordiagramm tunnusele sugu ja kirjuta sektoritele juurde "naised" ja "mehed".

1.6. Joonista mannaputru söövate ja mittesöövate tudengite kehakaalude histogrammid ja tihedusfunktsioonid.

1.7. Konstrueeri sagedustabel tunnusele pikkus, jagades pikkused 6 klassi.

1.8. Konstrueeri karp-vurrud diagrammid tunnusele peaümbermõõt sõltuvalt matemaatika hindest.

1.9. Leidke naistudengite keskmise kaalu 95%-usalduspiirid. Kas naistudengite keskmine kaal erineb statistiliselt oluliselt 60 kg-st?

1.10. Kas nais- ja meestudengite keskmised kehamassiindeksid on statistiliselt oluliselt erinevad?

1.11. Kas tudengite kehamassiindeksite jaotus erineb normaaljaotusest?

Iseseisev töö 2

Ülesanded baseeruvad osal Mariann Nõlvaku poolt aastail 2004-2006 kogutud Eesti kalade andmebaasist.

Andmestik: http://www.eau.ee/~ktanel/DK_0007/kala.xls

Andmestik sisaldab järgmisi andmeid:
- kala number (lihtsalt identifitseerimiseks);
- liik (6 liiki: haug, särg, latikas, luts, ahven ja koha);
- rühm: röövkala või lepiskala;
- 5 püügikohta (Võrtsjärv, Kärevere, Kastre, Praaga ja Peipsi järv);
- püügisesoon (kevad-suvi või sügis-talv);
- kaal ja pikkus;
- sugu;
- laiussiga (*Diphyllobothrium latum*) nakatumine ('diphyl' = 0 või 1);
- laiussi leidude arv kalal (diph_arv).

Ülesanded

Võtame uurimise alla üksnes haugid (et ülesandeid lihtsam lahendada oleks, võite selleks teha kalade andmestiku alusel uue, üksnes haugide andmeid sisaldava, andmestiku).

2.1. Kas püügisesoon ja laiussiga nakatumine (on nakatunud / ei ole nakatunud) on seotud?

a) Konstrueerige 2-mõõtmeline sagedustabel nii absoluutsete kui ka suhteliste sagedustega. Kommentaarid?

b) Testige seose statistilist olulisust nii χ^2-testiga kui ka Fisheri täpse testiga.

c) Arvutage šansside suhe (OR) ja selle 95% usalduspiirid võrdlemaks nakatumist sügis-talvisel sesoonil püütud haugidel kevad-suvisel sesoonil püütutega. Järeldused?

2.2. Konstrueerige karp-vurrud diagramm illustreerimaks erinevatest kohtadest püütud haugide kaalude erinevust (või sarnasust). Kas suudate *R*-i *Help*'st välja lugeda, millal loetakse mingi kaal erindiks ja tähistatakse joonisel eraldi punktina?

2.3.

a) Leidke haugide pikkuse ja kaalu vaheline lineaarne (Pearsoni) korrelatsioonikordaja ning kaalu ja laiussi leidude arvu vahelised Spearmani and Kendalli korrelatsioonikordajad. Kas need seosed on statistiliselt olulised?

b) Leidke lineaarne regressioonivõrrand prognoosimaks haugide kaalu nende pikkuse järgi. Kas leiud võrrand on statistiliselt oluline? Illustreerige seost hajuvusdiagrammiga, kuhu kandke peale ka regressioonisirge ja viimase 95% usaldusintervall. Kui palju võiks leitud võrrandi alusel kaaluda 60 cm pikkune haug?

Iseseisev töö 3

Ülesanded

Esimesed ülesanded baseeruvad osal Mariann Nõlvaku poolt aastail 2004-2006 kogutud Eesti kalade andmebaasist.

Andmestik: http://www.eau.ee/~ktanel/DK_0007/kala.xls

Võtame uurimise alla üksnes haugid.

3.1. Uurige püügikoha, soo ja sesooni mõju haugide pikkusele. Kas kõigi nimetatud faktorite mõju on statistiliselt oluline?

3.2. Kas Peipsi ja Võrtsjärve haugid on erineva pikkusega? Kas see erinevus on statistiliselt oluline?

3.3. Kui pikad on keskmiselt Võrtsjärvest sügis-talvisel sesoonil püütud isased ja emased haugid? Leidke 95%-usalduspiirid sügis-talvisel sesoonil Võrtsjärvest püütud isaste ja emaste haugide mudeli abil prognoositud keskmistele pikkustele.

Järgneva ülesande tarvis on andmestik aadressil

http://www.eau.ee/~ktanel/DK_0007/lehm.xls

Andmestik sisaldab eesti holsteini tõugu lehmade esimese laktatsiooni summaarse piimatoodangu (kg) ning rasva- ja valguprotsendi andmeid. Samuti on teada loomade sünniaasta, omanik ja isa.

3.4. Käsitledes sünniaastat fikseeritud faktorina (NB! andmetes on sünniaasta arvulisena, mistap käsitleb *R* seda vaikimisi pideva argumendina …) ning omanikku ja isa juhuslike faktoritena hinnake, kui suur on isa ja farmi mõju osakaal piimatoodangu ning rasva- ja valguprotsendi koguvarieeruvusest.

Millise toodangunäitaja puhul on omaniku (so valdavalt söötmis-pidamistingimuste) mõju suurim ning millise toodangunäitaja puhul on isa mõju (so isalt pärandunud geenide efekt) suurim?

Iseseisev töö 4

Ülesanded

4.1. Andmestik *R*-i andmefailina: http://www.eau.ee/~ktanel/DK_0007/puud.rda

Puude andmestikus vastab üks rida ühele puule, veergudes on vastavalt:

veerus nimega 'A' puu vanused aastates; veerus 'D' puu diameeter sentimeetrites; veerus 'H' puu kõrgus meetrites; veerus 'ARENGUKL' puu arenguklass väärtustega: A – lage, N – noorendikud, L – latimets (noorendikust järgmine), K – keskealised, V – valmiv, Y – küps, S – selguseta, – puuduv väärtus;
veerus 'PE' puu liik: HB – haab, KS – kask, KU – kuusk, LH - lehis, LM – sanglepp, LV – hall lepp, MA – mänd, RE – remmelgas, SA – saar, TA – tamm;
…

Lisage andmebaasi uus binaarne tunnus (näiteks nimega 'KYPS') väärtustega 1 (puu on raieküps, 'ARENGUKL'=Y) ja 0 (puu ei ole raieküps, 'ARENGUKL'≠Y).

Linnamehest metsaomanik ei tea, kui vana on tema mets. Küll aga teab ta seda, et tal on metsas valdavalt kuused, ja samuti teab ta puude diameetreid.

a) Leidke logistilise regressiooni mudel, mis prognoosib, kui tõenäoliselt on mingi diameetriga kuusk raieküps.
Pange kirja logistilise regressiooni võrrand ning illustreerige tulemust joonisega.

b) Leidke optimaalne logistilise mudeliga hinnatud raieküpsuse tõenäosuse väärtus, millest alates maksab lugeda kuuske raieküpseks. Kui suur on sellise otsustusreegli korral testi tundlikkus ja spetsiifilisus?

c) Millise kuuse diameetri korral võib juba arvata, et kuusk on 90%-lise tõenäosusega raieküps?

4.2. Moodustage üksikute puude andmestiku baasil uus puuliikide andmestik, millesse pange kirja puu liik ja arvutage iga liigi kohta
- puude keskmine kõrgus vanuses 20 aastat,
- puude keskmine kõrgus vanuses 50 aastat,
- kõrguse juurdekasv vahemikus 60-80 aastat (modelleerige neljandat järku polünoomiga puude kõrguse sõltuvust vanusest, hinnake saadud mudelist puude kõrgus vanustes 80 ja 60 aastat ning arvutage viimaste vahe; nendele puuliikidele, mille maksimaalne vanus on alla 80 aasta, võtke kõrguse juurdekasvuks 0),
- maksimaalne kõrgus ja
- maksimaalne vanus.

Kõrvale jätke vähem kui 10 mõõtmisega liigid.

a) Teostage nende viie tunnuse alusel puuliikide klasteranalüüs (hierarhiline klasterdamine). Illustreerige ja kommenteerige tulemusi.

b) Teostage nende viie tunnuse peakomponentanalüüs. Kui mitu peakomponenti omavad suuremat kirjeldusvõimet kui üksikud tunnused eraldi võetuna ning kui suure osa algtunnuste varieeruvusest need peakomponendid ära kirjeldavad? Illustreerige peakomponentide ja algtunnuste vahelisi seoseid joonisega – milliseid erinevate puuliikide kasvu omadusi ühendab esimene ja milliseid teine peakomponent? Mida võite tehtud analüüside alusel järeldada erinevate liikide kohta?