Eksamist

Eksami sooritamiseks tuleb hakkama saada kirjaliku eksamitööga ja esitada projekt. Tungivalt soovitatav (aga mitte kohustuslik) oleks ise eksamile järgneval tööpäeval (või kokkulepitud hilisemal ajal) tulla kohale eksamihinnet teada saama – eelkõige teie projektiga seoses võib olla õppejõul tekkinud arusaamatusi ja mittemõistmist – seega oleks võimalik isiklikult kohale tulles ise vajadusel täiendavaid selgitusi jagada. Vestluseks saab aega kinni panna samal päeval, mil kirjutate eksamitööd.

Kirjalik eksam koosneb kahest osast. Esimene osa testib mõisteid ja teadmisi, mida te peaksite peast teadma. Antud osa ülesanded/küsimused peaksid olema lihtsad, kui te antud mõistet/definitsiooni teate ja olete tema tähendust õieti mõistnud. Eksami esimene osa toimub väga tempokalt – aega vastamiseks antakse umbes 1-2 minutit iga küsimuse kohta. Eksami esimene osa tuleb sooritada ilma materiale kasutamata.

Eksami teises osas võib materiale kasutada. Teise osa põhiosa moodustavad ülesanded – tuleb otsustada küsimusega kaasa pandud materiali põhjal, kas tegemist on normaaljaotusega; arvutada ise usaldusintervall keskväärtusele vms. Ka eksami teine osa saab olema suhteliselt tempokas – aega on umbes 10 min iga ülesande kohta. Seega ei tasu lootma jääda, et te suudate eksami ajal loengumaterialiga tutvuda – loengumaterial peaks ikka eelnevalt olema läbi uuritud, muidu jääte eksamil lihtsalt ajahätta.

Kordamisküsimused

(*) - tähendab teemat, mille kohta esitatud küsimustele tuleb osata vastata ilma konspekti kasutamata. Vajadusel peate suutma ise, ilma arvuti ja abimateriali abita, ka vastava näitaja väärtust leida.

1. ANDMETE ESITAMINE; PÕHISTATISTIKUD

- keskmine, mediaan, mood (*)
- miinimum, maksimum, dispersioon ja standardhälve (*)
- sagedustabel (*)
- tulpdiagramm, karp-vurrud diagramm (*)
- tunnuste tüübid (*)

2. POPULATSIOON JA VALIM

- populatsioon ja valim (*)
- tunnuse jaotus populatsioonis, seonduvad mõisted (keskväärtus, populatsiooni dispersioon, populatsiooni standardhälve vms)
- keskväärtuse omadused (*)
- dispersiooni omadused (*)
- binoomjaotus. Kuna on juhusliku suuruse jaotuseks binoomjaotus. Peate oskama leida tõenäosust, et etteantud binoomjaotusega juhuslik suurus omandab mingi konkreetse väärtuse. Näiteks: Kui $X \sim B(4, 0.25)$, siis millise tõenäosusega näeme väärtust 2: $P(X=2)=\ldots.$.
- normaaljaotus: millal tekib, kuidas kontrollida, kas uuritava tunnuse jaotuseks on normaaljaotus.

3. VAHEMIKHINNANGUD

- prognoosiintervall (normaaljaotusega juhuslikule suurusele). Peate oskama arvutada või leida prognoosiintervalli. (*)
- usaldusintervall keskväärtusele. Peate oskama ka vaatluste või esitatud põhistatistikute pealt usaldusintervalli keskväärtusele arvutada.

4. HÜPOTEESIDE KONTROLLIMINE

- nullhüpotees/alternatiivne hüpotees (*)
- esimest liiki viga/teist liiki viga (*)
- olulisuse nivoo (*)
- olulisustõenäosus (*)
- testi võimsus
- t-test keskväärtuse kohta käivate hüpoteeside kontrollimiseks. Peate oskama vaatluste või esitatud statistiliste näitajate (näiteks: keskmine ja standardhälve) pealt t-testi teha: peate oskama arvutada t-statistiku väärtust ja oskama võtta vastu otsust (tõestan H1 / jään H0 juurde)
- hii-ruut test. Peate oskama esitatud vaatluste põhjal ise vajadusel arvutusi teha (teststatistiku väärtust leida ja võtta vastu otsust: H1 või H0).

5. STATISTILINE SEOS

- statistiline seos kahe tunnuse vahel (*)
- põhjuslik vs mittepõhjuslik seos – kuna on tarvis teada ühte, kuna piisab teisest

6. REGRESSIOONANALÜÜS

- lineaarse regressioonanalüüsi mudel, tema interpretatsioon
- regressioonanalüüsi juures kontrollitavad hüpoteesid
- determinatsioonikordaja, lineaarne (Pearsoni) korrelatsioonikordaja, tema tähendus
- regressioonanalüüsi eeldused

7. DISPERSIOONANALÜÜS

- mitmese võrdluse probleem, Bonferroni meetod (*)
- dispersioonanalüüsi mudel
- dispersioonanalüüsi eeldused

Hoiatus! Hüpoteeside statistilise kontrollimise (nullhüpotees; esimest liiki viga; p-väärtus; olulisuse nivoo) ja usaldusintervalli interpretatsiooni kohta võidaks esitada norivaid/trikiga küsimusi, mille vastust loetakse õigeks ainult täiesti korrektsete vastuste korral!

Eksami teise osa ülesannete lahendamisel on lubatud kasutada taskuarvuti abi.

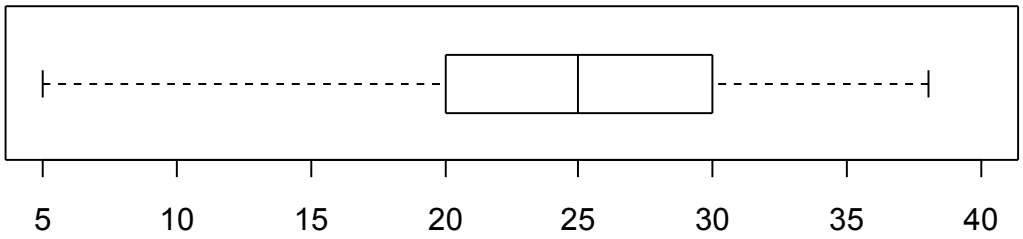
Näide eksamitöö A-osast

Biomeetria eksam. A osa.

(Sooritada ilma lisamaterialideta)

12. detsember 2007

1. Uuritava tunnuse väärtuseid iseloomustab järgmine karp-vurrud diagramm:



Millisesse mainitud vahemikest sattub 25% uuritava tunnuse väärtustest?

a) 5..25 b) 20..30 c) 5..30 d) 25..30 e) 25..38

2. Mis on olulisustõenäosus?

3. Viie linnu lennutee pikkused olid järgmised: 6 km; 4,5 km; 0,2 km; 4,4 km; 5,3 km. Leia antud valimit kasutades mediaan linnu lennutee pikkusele.

4. Mis on teist liiki viga?

5. Täiskasvanud kollanoka tiiva pikkuse (mõõdetud millimeetrites) dispersioon on 9. Kui mõõtmisi oleks tehtud millimeetrite asemel sentimeetrites, mis oleks siis saadud tiiva pikkuse dispersiooniks?

6. Tahetakse teada, milline/sed toidulisandid soodustavad pleektatsudel liimja lima tootmist. Katsetati kümmet eri toitu ja tehti 10 t-testi (lima hulk ilma toidulisandita vs toidulisandiga toitmisel). Saadi järgmised olulisustõenäosused:

1. 0,896 2. 0,042 3. 0,969 4. 0,002 5. 0.444
6. 0.044 7. 0.500 8. 0.051 9. 0.555 10. 0.712

Millistel juhtudel võib vastu võtta alternatiivse hüpoteesi ja öelda, et toidulisand mõjutab lima teket?

7. Suur valimi standardhälve näitab, et mõõtmistulemused on

a) suured
b) väikesed
c) üksteisest väga erinevad
d) statistiliselt mitteolulised

Näiteid eksami B-osa ülesannetest

1. Sooviti uurida kirjatuvide lennukiirust. 15 lindu viidi tuvilast 200 km kaugusele ja lasti vabaks. Kahe esimese päeva jooksul jõudis tuvilasse tagasi 10 lindu, neil kulus saabumiseks 18, 24, 28, 29, 31, 35, 36, 37, 39 ja 43 lennutundi. Kas kahe päeva jooksul saabunud tuvide andmestiku põhjal on võimalik iseloomustada kirjatuvide tüüpilist lennukiirust? Kas selleks on sobilikum kasutada valimikeskmist või valimimediaani? Miks?

Vastus: Valimikeskmist ei ole võimalik arvutada, sest meil pole teada viie linnu lennuaeg. Seevastu valimimediaani on võimalik arvutada (variatsioonrea 8. element on meil teada – 37 tundi – sest me teame, et ülejäänud viiel linnul kulus lennu peale rohkem aega kui teistel). Seega on antud näite puhul ainus võimalus iseloomustada kirjatuvide lennukiirust valimi mediaani abil (mis hindab üldiselt kirjatuvidel antud distantsi läbimiseks kuluva aja mediaani)

2. Kääbikupaar toob igal kevadel ilmale 5 pisikest ja karvast kääbikulast. Kui nii kääbikuema kui ka kääbikuisa kannavad retsessiivset pikakõrvalisuse geeni (st. nende laps omandab tõenäosusega 1/4 mõlemalt vanemalt pikakõrvalisuse geeni ja kasvatab seetõttu suureks saades jäneslikult kikkis kõrvad), kui suur on siis tõenäosus, et taolise vanematepaari viiest lapsest pole ükski pikkkõrvaline? Aga kui suure tõenäosusega on viiest lapsest just kaks pikkkõrvalised? (Vihje: ehk aitab binoomjaotus?)

Vastus:
Antud juhul saab tõepoolest kasutada binoomjaotust - võttes ühe "katse" "õnnestumise" tõenäosuseks 1/4=0,25 ja "katsete" arvuks 5, saame (võttes arvesse, et 0!=1 ja $x^0 = 1$):

a) leida tõenäosuse, et lastest pole keegi pikakõrvaline:

$$\begin{aligned} P(X=0) &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \frac{5!}{0!(5-0)!} 0{,}25^0 0{,}75^{5-0} \\ &= 0{,}75^5 \quad (= 0{,}237) \end{aligned}$$

b) tõenäosus, et lastest täpselt kaks on pikakõrvalised:

$$\begin{aligned} P(X=0) &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \frac{5!}{2!\,3!} 0{,}25^2 0{,}75^3 \\ &= 10\; 0{,}25^2\; 0{,}75^3 \quad (= 0{,}2636...) \end{aligned}$$

3. Kümnes erinevas Eestimaa paigas mõõdeti 4.aprillil õietolmu sisaldust õhus. Tulemuseks saadi (mõõtmistäpsuse piirides) järgmised arvud:

3, 4, 8, 5, 4, 2, 6, 4, 6, 5

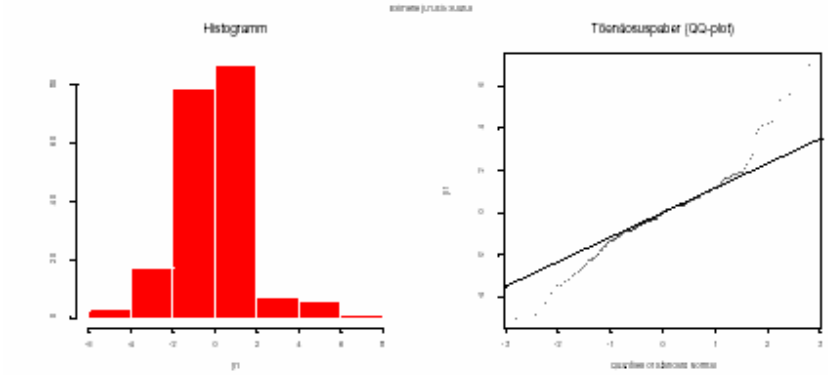
Leia antud valimi keskväärtus, dispersioon, standardhälve ja mediaan. Leia usaldusintervall 4.aprillil keskmisele õietolmu sisaldusele õhus.

Vastus:
keskmine on 4,7
dispersioon s^2=2,9
standardhälve s=1,7029...
mediaan=4,5
95% usaldusintervall oleks (3,48...5,92)
(t-jaotuse vastav täiendkvantiil oli tabeli andmetel 2,26)

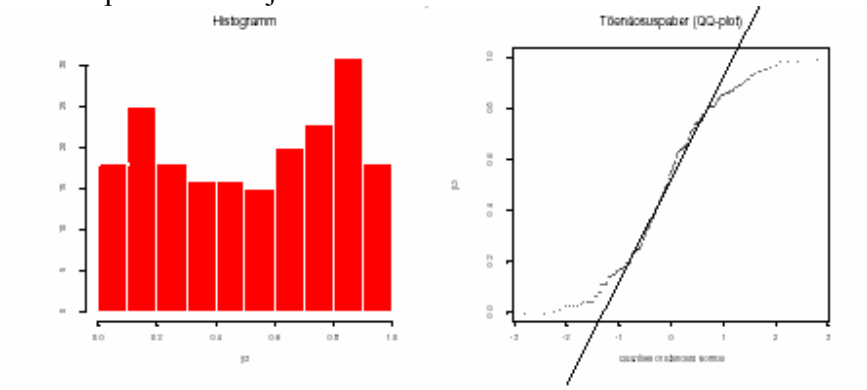
4. Loodusuurija Harry Hoolas on oma elu jooksul leidnud 2000 linnuliigi keskmisele sabapikkusele 97%-lised usaldusintervallid. Kui teil palutakse hinnata, mitme linnuliigi tegelik keskmine sabapikkus (sabapikkuse keskväärtus) ei asu prof. H.Hoolas poolt leitud usaldusvahemikes, siis millist vastust pakute?

Vastus:
Tegelik keskmine ei sattu 97%-lisse usaldusintervalli keskmiselt 3% juhtudest ehk tõenäosusega p=0,03. 2000 usaldusintervalli seas oleks seega ligikaudu 0,03*2000=60 tegelikku keskmist sabapikkust mittesisaldavat usaldusintervalli.

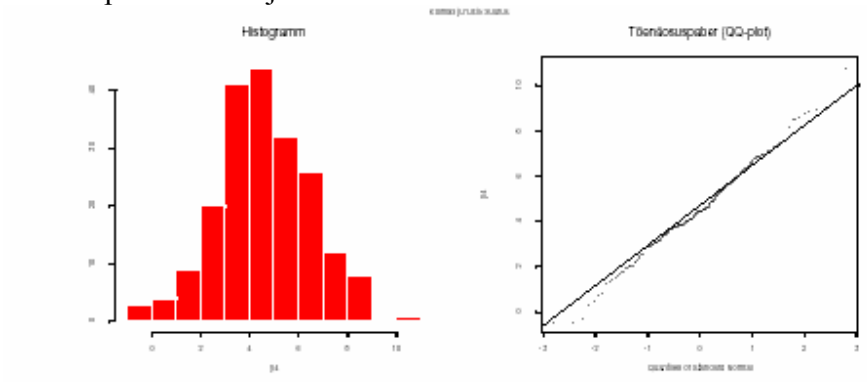
5. Ütle, milliste alljärgnevate juhuslike suuruste jaotus võiks olla normaaljaotus?



Vastus: pole normaaljaotus



Vastus: pole normaaljaotus



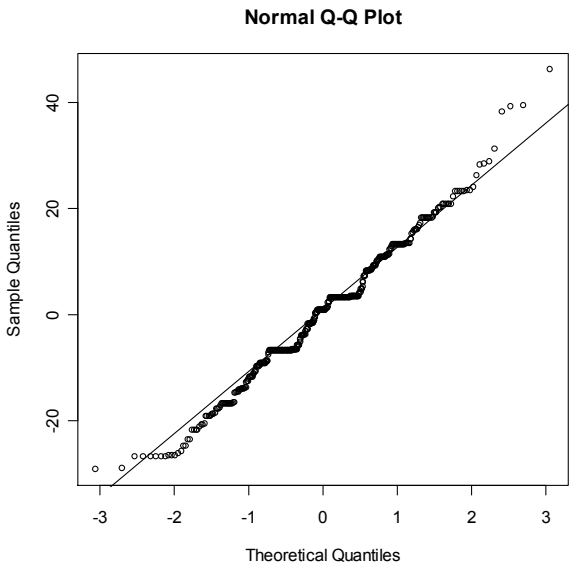
Vastus: võib olla normaaljaotus

6. Soovitakse teada, kas sportimise ja vererõhu vahel on mingit statistilist seost (näiteks: kas neil, kes rohkem spordivad, võiks olla kõrgem või madalam vererõhk kui neil, kes vähem teevad sporti). Tunnus sport (mitu korda nädalas teed sporti) on kodeeritud järgmiselt: :
1-„ei tee sporti“; 2-„1-2 korda“; 3- „3-4 korda “; 4-„5 või enam“.
Sellele küsimusele vastamiseks tehakse dispersioonanalüüs. Dispersioonanalüüsi tulemused olid alljärgnevad:

```
Call:
lm(formula = SVR ~ factor(sport))

Residuals:
     Min       1Q   Median       3Q      Max
-29.2078  -6.9721   0.7922   8.7922  46.0279

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     116.7326     1.3787  84.671   <2e-16 ***
factor(sport)2    0.2396     1.5975   0.150   0.8809
factor(sport)3    2.4752     2.0059   1.234   0.2179
factor(sport)4    7.3979     3.0013   2.465   0.0141 *
---

Residual standard error: 12.79 on 436 degrees of freedom
Multiple R-Squared: 0.0213,     Adjusted R-squared: 0.01232
F-statistic: 2.372 on 3 and 436 DF,  p-value: 0.05166
```

Vasta:

a) Kas eri intensiivsusega sportivatel inimestel on vererõhu keskväärtus erinev?
b) Kirjelda seost vererõhu ja sportimise vahel – kas sportimine suurendab või vähendab vererõhku? Kas tulemus on üllatav? Mis põhjusel võiks seos tulla selline, nagu ta tuli?
c) Kui suur on keskmine vererõhk neil, kes ei tee sporti (sport=1)? Kui suur on keskmine vererõhk iga päev sporti tegevatel tudengitel (sport=4)?
d) Kirjelda, kui hästi me saame inimese vererõhku prognoosida, kui teame tema sportimisharjumusi. Kas sportimisharjumuste teadmine võimaldab saada täpset prognoosi vererõhule?
e) Selleks, et dispersioonanalüüsi tulemusi saaks usaldada, peavad olema täidetud dispersioon-analüüsi eeldused. Üheks eelduseks on nõue, et mudeli jäägid peaksid olema normaaljaotusega. Kas see nõue on praegu rahuldatud? Vaata ka mudeli jääkide jaoks joonistatud tõenäosus-paberit!



7. Mis on laiem, kas 90%-usaldusintervall või 95%-usaldusintervall? Miks?

8. Biotoobi kirjeldamisel kasutati tunnust "sademete hulk" mõõdetuna ühikuis mm/kuus. Mis tüüpi tunnusega on tegemist?

9. Populatsioonis leidub väike grupp isendeid kellel uuritav tunnuse väärtused on järsult suuremad kui ülejäänutel. Kui arvutada välja nii populatsiooni mediaan kui ka keskväärtus, siis kumb neist võiks tulla suurem? Miks?

10. Väliseksperdid uurisid eesti vaenulindude sabalaiust X ja leidsid, et $EX=4$, $DX=0{,}25$. Mõõtmisühikuks oli neil toll. Millised oleksid tulnud keskväärtus ja dispersioon kui uurijad oleksid kasutanud sentimeetreid mõõtmisühikutena (üks toll on ligikaudu 2,5cm)?

11. Genotüübiga AA madagaskari kärnkonnadel kasvavad teistest sama liigi kärnkonnadest lühemad tagajalad. Genotüüpidega Aa ja aa isenditel on koibade pikkus normaalne. Loodusuurijad uurisid 1000 juhuslikult valitud konna, kellest 42 olid genotüübiga AA, 316 genotüübiga Aa ja 642 genotüübiga aa. Geenialleelide jaotus sarnaneb olukorrale, kus looduslikku valiku survet märgatavalt ei esine (populatsioon näib olevat Hardy-Weimbergi tasakaalus). Loodusuurijaid huvitab, mis võiks kompenseerida lühemate (ja seetõttu halvema hüppevõimega) jalgade omamist. Hüpoteesiks on, et lühemate tagajalgadega konnade keskmine reaktsioonikiirus on suurem. Kõigil 1000 konnal mõõdeti reaktsioonikiirused (ms). Millise testi abil saaks mainitud hüpoteesi kontrollida?

12. Eksperimendis kasutatakse lisaainet X, mida saadakse tablettide kujul (tablett lahustatakse ja manustatakse katseloomale). Ühes tabletis on toimeainet 2mg, standardhälve 0,03. Eeldades, et toimeaine kogus tabletis on normaaljaotusega juhuslik suurus, siis millisesse vahemikku jääb 95% tablettide puhul toimeaine kogus?

13. Kolm haruldase linnu muna sattus uurijate kätte. On teada, et antud linnuliigi puhul on emaslinnu koorumise tõenäosus (0,6) veidi suurem kui isaslinnu koorumise tõenäosus (0,4). Milline on tõenäosus, et nendest kolmest munast kooruvad linnulapsed selliselt, et tekiks vähemalt üks linnupaar (kooruks vähemalt üks emaslind ja vähemalt üks isaslind)?

14. Otsusta, kas eksisteerib seos tunnuste X ja Y vahel. Kas see seos on tugev? Milline võiks olla Y-i väärtus, kui $X=20$? Põhjenda oma arvamusi!

```
> summary(lm(y~x))
Residuals:
      Min        1Q    Median        3Q       Max
-14.75012  -3.03364   0.09066   3.00279  12.53587

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.1505     2.0831   4.393 2.84e-05 ***
x             0.5372     0.1164   4.616 1.19e-05 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.913 on 98 degrees of freedom
Multiple R-Squared: 0.1786,     Adjusted R-squared: 0.1702
F-statistic: 21.31 on 1 and 98 DF,  p-value: 1.186e-05
```