Kõik küsimused, mis puudutavad Excel'i kasutamist
(eelkõige Excel'i statistikat) võib saata aadressil tiit@zbi.ee

ANDMETE TEISENDAMINE

Korrutamine, liitmine, lahutamine, jagamine:

=A1*A2
=A1+A2+7
=(A1+A2)/A3
jne.

Ruutjuur **=SQRT(A7)**
Astendamine **=(A2)^2** Astendamise märgi saab Ctrl-Alt-Ä

IF-tüüpi lausete süntaks:

=IF(tingimus; väärtus kui tingimus täidetud; väärtus kui tingimus ei ole täidetud)

NT.

=IF(A2>7; B2; C2)
=IF(A2<48; B2/A2; B2)
=IF((A2/B2)<=2; 1; 0)
=IF(A2>B2; 1; 0)
=IF(AVERAGE(A1:A6)>20; 1; 0)

Kui tahame vastuseks teksti, siis tuleb see panna jutumärkidesse:
=IF(ABS(O3)>1,96;"p<0,05";"p>0,05")

VALIMI KIRJELDAMINE

Vajaliku statistiku leidmiseks sisestage ruutu **=statistikafunktsioon** ja sulgudesse **vaatlusandmete plokk**, mille kohta teavet soovitakse.

Keskväärtus `=AVERAGE(B2:B19)`
Mediaan `=MEDIAN(B2:B19)`
Mood `=MODE(B2:B19)`
Maksimumväärtus `=MAX(B2:B19)`
Miinimumväärtus `=MIN(B2:B19)`
Tinglik keskmine `=DAVERAGE(B1:B19; B1; A5:A6)`
`kus (DAVERAGE(andmeplokk; tunnus; filter)`
Mõõtmiste arv `=COUNT(B2:B19)`
Standardhälve `=STDEV(B2:B19)`
Dispersioon `=VAR(B2:B19)`
Standardviga (1) `=B27/SQRT(B26)`
Standardviga (2) `=STDEV(B2:B19)/SQRT(COUNT(B2:B19))`
`kui eelnevalt ei ole STDEV ja COUNT leitud`
Variatsioonikoefitsient
`=100*STDEV(B2:B19)/AVERAGE(B2:B19)`
`kui eelnevalt ei ole STDEV ja AVERAGE leitud`
Asümmeetrilisus `=SKEW(B2:B19)`
Järsakus `=KURT(B2:B19)`

Terve hulga erinevaid statistikuid korraga väljastab Excel järgmisel teel:

TOOLS > DATA ANALYSIS > DESCRIPTIVE STATISTICS

Samast, TOOLS > DATA ANALYSIS, võib leida ka suurema osa Excel'i statistilistest protseduuridest.

ANALÜÜSID DISKREETSETE MUUTUJATEGA
HII-RUUT-TEST

Lühidalt: Jutt sellistest muutujatest, mille väärtusi leiame loendamise, mitte mõõtmise teel. Kui sõltuv muutuja on sellist tüüpi, siis tuleb hoolega vältida algajatel sageli esinevat viga andmestikku jõuga allutada pidevate muutujate analüüsimiseks ette nähtud meetoditele. Sagedustabel tekib olukorras, kus jagame miskeid objekte klassidesse. (Toomase loengust nr. 6). Nii võime 100 püütud looma jagada emasteks ja isasteks, tulemuseks sagedustabel.

	emaseid	isaseid
vaadeldud jaotus	65	35
oodatav jaotus	50	50

Probleem: Kas vaadeldud emaste-isaste suhe erineb 50:50-st?

Sisestada:
=CHITEST(B2:C2; B3:C3), kus b2:c2 oleks empiiriline jaotus ja b3:c3 teoreetiline jaotus

Tulemused:

Tulemusena saame hii-ruut-testi *p* väärtuse. hii-ruut-statistiku arvutamine ei ole automatiseeritud. hii-ruut-jaotuse kriitiline väärtus arvutatakse järgmiselt: **=CHIINV(olulisuse nivoo; vabadusastmete arv)**

F-TEST ÜLDKOGUMITE DISPERSIOONIDE (ÜHEPOOLNE HÜPOTEES) VÕRDLEMISEKS (F-TEST TWO-SAMPLE FOR VARIANCES)

F-testi, t-testide ja ANOVA teoreetilise ülevaate saab Toomase II loengust

Lühidalt: vastust otsitakse küsimusele, kas üldkogumite varieeruvus on erinev.

Probleem: Kas erinevatel toidutaimedel kasvanud röövikute nukukaalude varieeruvus on erinev.

Sisestusaknas tuleb määrata:

- **Variable 1 range:** I algandmete plokk (koos rühma nimega, kui soovi on);
- **Variable 2 range:** II algandmete plokk (koos rühma nimega, kui soovi on);
- **Labels:** kasutada siis, kui soovitakse kasutada ka rühmade nimesid;
- **Alpha:** usaldusnivoo (vaikimisi 0,05). Ehk teisisõnu millise *p* väärtuse jaoks me kriitilist *F*-statistikut soovime.

Output options: määratakse, kuhu tulemused paigutatakse.

- ***Output range:*** sama töölehe piires;
- ***New Worksheet Ply:*** uuele töölehele; vaikimisi avatakse uus tööleht, mis saab nimeks Sheet #;
- ***New workbook:*** avatakse uus fail ja kirjutatakse tulemused sinna.

Soovitus: kasutage siin ja edaspidi ***New Worksheet Ply***. a) testi tulemused ei ole segi andmetega b) testi tulemuste vaatamiseks ei pea eraldi faili avama.

Tulemused:

F-Test Two-Sample for Variances

- **Mean:** uuritavate gruppide keskmised
- **Variance:** dispersioon
- **Observations:** valimi maht grupiti
- **df:** vabadusastmete arv
- **F:** F-statistik
- **P(F<=f) one tail:** olulisuse tõenäosus
- **F critical one-tail:** F-statistiku kriitiline väärtus

T-TEST KESKVÄÄRTUSTE VÕRDLEMISEKS SÕLTUVATE VALIMITE KORRAL (T-TEST: PAIRED TWO SAMPLE FOR MEANS)

Lühidalt: kasutatakse keskmiste võrdlemiseks, kui valimid ei ole sõltumatud, st. kui iga vaatlus I valimis on seotud vaatlusega teises valimis (näiteks: igat objekti on mõõdetud kahel korral, enne ja pärast teatud manipulatsiooni).

Probleem: Me soovime testida, kas kehatemperatuur on hommikuti ja õhtuti erineb.

Sisestusaknas tuleb määrata:

- **Variable 1 range:** I algandmete plokk (koos rühma nimega, kui soovi on);
- **Variable 2 range:** II algandmete plokk (koos rühma nimega, kui soovi on);
- **Hypothesized Mean Difference:** oletatav keskväärtuse erinevus, vaikimisi null (kui meil on põhjust eeldada, et keskmised erinevad);
- **Labels:** kasutada siis, kui soovitakse kasutada ka rühmade nimesid;
- **Alpha:** usaldusnivoo (vaikimisi 0,05). Ehk teisisõnu millise *p* väärtuse jaoks me kriitilist *t*-statistikut soovime.

Output options: määratakse, kuhu tulemused paigutatakse.

- ***Output range:*** sama töölehe piires;
- ***New Worksheet Ply:*** uuele töölehele; vaikimisi avatakse uus tööleht, mis saab nimeks Sheet #;
- ***New workbook:*** avatakse uus fail ja kirjutatakse tulemused sinna.

Tulemused:

t-Test: Paired Two Sample for Means

- **Mean:** uuritavate gruppide keskmised
- **Variance:** dispersioon
- **Observations:** valimi maht grupiti
- **Pearson correlation:** Pearson'i korrelatsioonikordaja
- **Hypothesized mean difference:** sai määratud sisestusaknas
- **df:** vabadusastmete arv
- **t stat:** t-statistik
- **P(T<=f) one-tail:** olulisuse tõenäosus ühepoolse hüpoteesi korral
- **t critical one-tail:** t-statistiku kriitiline väärtus ühepoolse hüpoteesi korral
- **P(T<=f) two-tail:** olulisuse tõenäosus kahepoolse hüpoteesi korral
- **t critical two-tail:** t-statistiku kriitiline väärtus kahepoolse hüpoteesi korral

T-TESTID KESKVÄÄRTUSTE VÕRDLEMISEKS (T-TEST: TWO-SAMPLE ASSUMING UNEQUAL VARIANCES ja T-TEST: TWO-SAMPLE ASSUMING EQUAL VARIANCES)

Probleem: Uurime, kas toomingal ja kasel kasvanud röövikute nukukaalud erinevad. Selle, kas kasutada t-testi, mis eeldab võrdseid dispersioone või mitte, teeme kindlaks F-testi abil.

Sisestusaknas tuleb määrata:

- **Variable 1 range:** I algandmete plokk (koos rühma nimega, kui soovi on);
- **Variable 2 range:** II algandmete plokk (koos rühma nimega, kui soovi on);
- **Hypothesized Mean Difference:** oletatav keskväärtuse erinevus, vaikimisi null (kui meil on põhjust eeldada, et keskmised erinevad);
- **Labels:** kasutada siis, kui soovitakse kasutada ka rühmade nimesid;
- **Alpha:** usaldusnivoo (vaikimisi 0,05). Ehk teisisõnu millise *p* väärtuse jaoks me kriitilist *t*-statistikut soovime.

Output options: määratakse, kuhu tulemused paigutatakse.

- ***Output range:*** sama töölehe piires;
- ***New Worksheet Ply:*** uuele töölehele; vaikimisi avatakse uus tööleht, mis saab nimeks Sheet #;
- ***New workbook:*** avatakse uus fail ja kirjutatakse tulemused sinna.

Tulemused:

t-Test: Paired Two Sample for Means

- **Mean:** uuritavate gruppide keskmised
- **Variance:** dispersioon
- **Observations:** valimi maht grupiti
- **Hypothesized mean difference:** sai määratud sisestusaknas
- **df:** vabadusastmete arv
- **t stat:** t-statistik
- **P(T<=f) one-tail:** olulisuse tõenäosus ühepoolse hüpoteesi korral
- **t critical one-tail:** t-statistiku kriitiline väärtus ühepoolse hüpoteesi korral
- **P(T<=f) two-tail:** olulisuse tõenäosus kahepoolse hüpoteesi korral
- **t critical two-tail:** t-statistiku kriitiline väärtus kahepoolse hüpoteesi korral

ÜHEFAKTORILINE DISPERSIOONANALÜÜS (ANOVA: SINGLE FACTOR)

Lühidalt: vastust otsitakse küsimusele, kas pideva tunnuse väärtused rühmiti erinevad. Küsitakse: kas valimi rühmakeskmiste erinevus on juhuslik või mitte.

Probleem: Liblikaröövikuid kasvatati 3 erineval toidutaimel. Meile pakub huvi, kas erinevatel toidutaimedel kasvanud liblikate nukukaal erineb. Seda, miks mitte antud probleemi lahendada mitme t-testiga, vt. Toomase loeng nr. 2.

Sisestusaknas tuleb määrata:

- **Input range:** algandmete plokk (koos rühmade nimedega, kui soovi on);
- **Grouped by:** määratakse, kas uuritavad rühmad on veerge või ridu pidi;
- **Labels in first row:** kasutada siis, kui soovitakse kasutada ka rühmade nimesid;
- **Alpha:** määratakse *F*-statistiku usaldusnivoo (0,05; 0,01). Ehk teisisõnu millise *p* väärtuse jaoks me kriitilist *F*-statistikut soovime. Suurt tähtsust ei ole, sest arvutab nagunii nii täpse *p* kui *F*-statistiku.

Output options: määratakse, kuhu tulemused paigutatakse.

- ***Output range:*** sama töölehe piires;
- ***New Worksheet Ply:*** uuele töölehele; vaikimisi avatakse uus tööleht, mis saab nimeks Sheet #;
- ***New workbook:*** avatakse uus fail ja kirjutatakse tulemused sinna.

Tulemused:

Summary

- **Groups:** uuritavate gruppide nimed (kui defineeritud)
- **Count:** rühmade suurused
- **Average:** valimi keskväärtused rühmade kaupa
- **Variance:** valimi dispersioon

ANOVA

- **SS:** hälvete ruutude summa (erinevused keskmisest on ruutu tõstetud ja kokku liidetud):
  - **Between groups** (SAS'is **MODEL)** - erinevused rühmade keskmiste vahel,
  - **Within groups** (SAS'is **ERROR**)- üksikväärtuste hajuvus ümber rühma keskmise.

- ***df*:** vabadusastmete arv
- **MS:** *MS=SS/df*, ehk siis vabaduastmega jagatud ruutude summa (Vt. Toomase loengut)
- **F**: *F*-statistiku väärtus
- **p-value:** usaldusnivoo
- **F crit:** *F*-jaotuse täiendkvantiil, ehk millise *F*-statistiku väärtuse korral on veel uuritav erinevus meie poolt etteantud *p* korral mittejuhuslik

KAHEFAKTORILINE DISPERSIOONANALÜÜS
(ANOVA: TWO-FACTOR WITHOUT REPLICATIONS)

Lühidalt: Meid võib huvitada mitme sõltumatu muutuja samaaegne mõju sõltuvale muutujale (või on see teine muutuja meie tahte vastaselt olemas ja mõjutamas ja asja segasemaks tegemas). Faktoreid on seega kaks (Toomase loengust nr. 4).

Probleem: Liblikaröövikuid kasvatati 3 erineval toidutaimel 3 erineval temperatuuril. Meile pakub huvi, kas toidutaim ja temperatuur mõjutavad liblikate nukukaalu.

Sisestusaknas tuleb määrata:

- **Input range:** algandmete plokk (koos rühmade nimedega, kui soovi on);
- **Labels:** kasutada siis, kui soovitakse kasutada ka rühmade nimesid;
- **Alpha:** määratakse *F*-statistiku usaldusnivoo (0,05; 0,01). Ehk teisisõnu millise *p* väärtuse jaoks me kriitilist *F*-statistikut soovime. Suurt tähtsust ei ole, sest arvutab nagunii nii täpse *p* kui *F*-statistiku.

Output options: määratakse, kuhu tulemused paigutatakse.

- ***Output range:*** sama töölehe piires;
- ***New Worksheet Ply:*** uuele töölehele; vaikimisi avatakse uus tööleht, mis saab nimeks Sheet #;
- ***New workbook:*** avatakse uus fail ja kirjutatakse tulemused sinna.

Tulemused

Anova: Two-Factor Without Replication

Summary

Grupiti esitatakse:

- **Count:** rühmade suurused
- **Sum:** mõõtmiste summa
- **Average:** valimi keskväärtused rühmade kaupa
- **Variance:** valimi dispersioon

ANOVA

- **Source of variation:** varieerumise allikas
  - **Rows**: faktori 1 tasemed
  - **Columns:** faktori 2 tasemed
  - **Error:** viga
- **SS:** hälvete ruutude summa
- ***df*:** vabadusastmete arv
- **MS:** *MS=SS/df*, ehk siis vabaduastmega jagatud ruutude summa (vt. Toomase loengut)
- **F**: *F*-statistiku väärtus
- **p-value:** usaldusnivoo
- **F crit:** *F* kriitiline väärtus

Lisaks sellele on Excel'is võimalik teha ka **korduvmõõtmistega ANOVA**'t:
Tools > Data analysis > ANOVA: Two-Factor with Replication

KORRELATSIOONIMAATRIKSI ARVUTAMINE (CORRELATION)

Lühidalt: Korrelatsioon kahe pideva muutuja vahel tähendab seda, et ühe muutuja suurematele väärtustele vastavad sagedamini teise muutuja suuremad (positiivne korrelatsioon) või väiksemad (negatiivne korrelatsioon) väärtused. Kui seost ei ole, on korrelatsioon null; maksimaalne on +1 ja minimaalne -1 (Toomase loengust).

Probleem: Leiame peremeesloomade ja neis arenenud parasitoidide karakteristikute (keha suurus, tiiva pikkus) vahelise korrelatsioonimaatriksi.

Sisestusaknas tuleb määrata:

- **Input range:** algandmete plokid (koos rühma nimedega, kui soovi on);
- **Grouped by:** määratakse, kas andmed on grupeeritud veerge või ridu pidi;
- **Labels in first row:** kasutada siis, kui soovitakse kasutada ka rühmade nimesid;

Output options: määratakse, kuhu tulemused paigutatakse.

- ***Output range:*** sama töölehe piires;
- ***New Worksheet Ply:*** uuele töölehele; vaikimisi avatakse uus tööleht, mis saab nimeks Sheet #;
- ***New workbook:*** avatakse uus fail ja kirjutatakse tulemused sinna.

Tulemused:

Tulemusena saadakse korrelatsioonimaatriks, kus on **Pearson**'i korrelatsioonikordajad.

REGRESSIOONANALÜÜS
(REGRESSION)

Lühidalt: Kahe korreleerunud muutuja puhul võime tahta seost kvantitatiivselt iseloomustada ehk siis vastata küsimusele, milline matemaatiline funktsioon seost kõige paremini kirjeldab. Sageli huvitab meid eelkõige ühe muutuja väärtuse ennustamine teise muutuja väärtuse järgi, ehk siis kui pikkus on teada, siis milline on tõenäone kaalu väärtus. Ennustamise otstarbel kasutame I tüüpi ehk nn. tavalist (lineaarset) regressiooni.

Probleem: Leiame seose peremeeslooma ja parasitoidi keha suuruse vahel.

Sisestusaknas tuleb määrata:

- **Input Y range:** sõltuva muutuja andmeplokk;
- **Input X range:** sõltumatu muutuja andmeplokk;
- **Labels:** kasutada siis, kui soovitakse kasutada ka rühmade nimesid;
- **Constant is zero:** kasutada, juhul kui eeldate, et tegemist on võrdelise sõltuvusega, ehk algordinaat $a = 0$ ($y = bx$);
- **Confidence level:** usaldusnivoo 1-á parameetrite usalduspiiride arvutamiseks.

Output options: määratakse, kuhu tulemused paigutatakse.

- ***Output range:*** sama töölehe piires;
- ***New Worksheet Ply:*** uuele töölehele; vaikimisi avatakse uus tööleht, mis saab nimeks Sheet #;
- ***New workbook:*** avatakse uus fail ja kirjutatakse tulemused sinna.

Tulemused:

Regressiooni statistikud (Summary output)

- **Multiple R:** korrelatsioonikordaja r
- **R Square:** determinatsioonikordaja r^2
- **Adjusted R square:**determinatsioonikordaja r^2 nihutamata hinnang
- **Standard error:** jääkstandardhälve
- **Observations:** vaatluste arv

Regressioonanalüüsi tulemuste dispersioonanalüüs (ANOVA)

- Esimeses reas on **varieeruvuse allikas**:
  - **Regression:** regressioonisirge
  - **Residual:** prognoosijäägid
  - **Total:** kokku
- ***df*:** vabadusastmete arv
- **SS:** hälvete ruutude summa
- **MS:** *MS=SS/df*, ehk siis vabaduastmega jagatud ruutude summa (Vt. Toomase loengut)
- ***F***: *F*-statistiku väärtus
- ***Significance F*:** mudeli olulisuse tõenäosus

Regressioonivõrrandi kordajate analüüs (.....)

- Esimeses reas on **regressioonivõrrandi parameeter:**
  - **Intercept:** vabaliige
  - **X variable:** sirge tõus
- **Coefficients:** parameetri hinnang
- **Standard error:** hinnangu viga
- **t stat:** t-statistik
- **P-value:** olulisuse tõenäosus
- **Lower 95%:** alumine 95%-line usalduspiir
- **Upper 95%:** ülemine 95%-line usalduspiir

Seega saame teada,
a) kui palju varieervusest seos kirjeldab ning
b) saame regressioonsirge valemi ja mudeli olulisuse tõenäosuse.