ÕPIOBJEKT

„*Andmeanalüüs MS Excelis (MS Excel 2010 baasil)*“

Tanel Kaart

http://ph.emu.ee/~ktanel/andmeanalyys_excelis/

Eesti Maaülikool
Estonian University of Life Sciences

Euroopa Liit
Euroopa Sotsiaalfond

Eesti tuleviku heaks

Õpiobjektid -> Andmeanalüüs MS Excelis (MS Excel 2010 baasil)

ANDMEANALÜÜS MS EXCELIS
(MS Excel 2010 baasil)

Õpiobjekti kirjeldus
Õpijuhis

Sissejuhatus

Peamised andmeanalüüsi teostamise vahendid *MS Excelis*
- Joonised
- Funktsioonid
- Protseduurid
- Risttabelid (*PivotTable*)

Sagedustabelid
- Pidev arvtunnus
- Diskreetne arvtunnus
- Mittearvuline tunnus

Arvkarakteristikud
- Valemid ja funktsioonid
- Protseduur *Descriptive statistics*
- Risttabel (*PivotTable*)
- Muud võimalused

Usalduspiirid
- Usalduspiirid keskmisele
- Usalduspiirid teistele parameetritele

Hüpoteeside kontrollimine (ühe ja kahe üldkogumi võrdlus)
- Üldskeem
- Hüpoteeside kontroll usalduspiiridega
- Z-test
- T-test
- F-test
- Mitteparameetrilised testid

Korrelatsioonanalüüs
- Pearsoni e lineaarne korrelatsioonikordaja
- Lineaarse korrelatsioonikordaja statistiline olulisus
- Spearmani e astakkorrelatsioonikordaja

Regressioonanalüüs
- Lineaarne regressioonanalüüs protseduuriga *Regression*
- Regressioonanalüüs graafiliselt
- Regressioonanalüüs funktsioonide abil
- Regressioonanalüüs *Solveri* abil

Kahemõõtmeline sagedustabel
- Kehamõõtmeline sagedustabel
- Hii^2-test
- Fisheri täpne test

Dispersioonanalüüs
- Ühefaktoriline dispersioonanalüüs
- Kahefaktoriline dispersioonanalüüs
- *Post-hoc* testid

Trikke ja nippe
- Kavalad funktsioonid ja valemid
- Excelile mitteomased joonised
- Andmeanalüüsil kasutatavad lisamoodulid

Lisa
- Kogu materjal ühe pdf-failina: stat_excelis.pdf

Õpiobjekti kirjeldus

Õppekava: Loomakasvatus (449)

Õppeaine: VL.0435 Katsetöö metoodika ja statistiline andmetöötlus

Maht: 0,5 EAP

Sihtrühm: Looma- ja kalakasvatuse õppekava magistrandid jt asjast huvitatud

Eesmärk: Õpiobjekti eesmärk on toetada õppeaine omandamist ning olla toeks edasisel õppe- ja teadustööl

Õpiobjekti läbinu:

- tunneb *Exceli* erinevaid andmeanalüüsi-alaseid võimalusi;
- oskab kasutada statistikafunktsioone ja -protseduure ning risttabelite konstrueerimise vahendit *PivotTable*;
- omab võimalust juhendites käsitletud näiteülesannete läbilahendamiseks.

Õppejõud ning sisuline ja tehniline teostus: Tanel Kaart

Eesti Maaülikool
sügissemester 2013

Järgmine >

Sisukord

Sissejuhatus 4
1. Peamised andmeanalüüsi teostamise vahendid MS Excelis 5
1.1. Joonised 5
1.2. Funktsioonid 7
1.3. Protseduurid 8
1.4. Risttabelid (*PivotTable*) 10
2. Sagedustabelid 13
2.1. Sagedustabel pidevale arvtunnusele 13
2.2. Sagedustabel diskreetsele arvtunnusele 18
2.3. Sagedustabel mittearvulisele tunnusele 19
3. Arvkarakteristikud 20
3.1. Valemid ja funktsioonid 20
3.2. Protseduur *Descriptive Statistics* 20
3.3. Risttabel (*PivotTable*) 23
3.4. Muud võimalused 24
4. Usalduspiirid 25
4.1. Usalduspiirid keskmisele 25
4.2. Usalduspiirid teistele parameetritele 27
5. Hüpoteeside kontrollimine (ühe ja kahe üldkogumi võrdlus) 30
5.1. Üldskeem 30
5.2. Hüpoteeside kontrollimine usalduspiiridega 31
5.3. z-test 32
5.4. t-test 35
5.5. F-test 39
5.6. Mitteparameetrilised testid 41
6. Korrelatsioonanalüüs 48
6.1. Pearsoni e lineaarne korrelatsioonikordaja 48
6.2. Lineaarse korrelatsioonikordaja statistiline olulisus 50
6.3. Spearmani e astakkorrelatsioonikordaja 52
7. Regressioonanalüüs 54
7.1. Lineaarne regressioonanalüüs protseduuriga *Regression* 54
7.2. Regressioonanalüüs graafiliselt 58
7.3. Regressioonanalüüs funktsioonide abil 61
7.4. Regressioonanalüüs *Solver*'i abil 70
8. Kahemõõtmeline sagedustabel 77
8.1. Kehamõõtmeline sagedustabel 77

8.2. χ^2-test 78
8.3. Fisheri täpne test 79
9. Dispersioonanalüüs 82
9.1. Ühefaktoriline dispersioonanalüüs 82
9.2. Kahefaktoriline dispersioonanalüüs 84
9.3. *Post-hoc* testid 88
10. Trikke ja nippe 90
10.1. Kavalad funktsioonid ja valemid 90
10.2. Excelile mitteomased joonised 99
10.3. Andmeanalüüsil kasutatavad lisamoodulid 100

Sissejuhatus

Järgnev materjal tutvustab andmete statistilise analüüsimise võimalusi MS Excelis.

Kuigi õpetuste aluseks on MS Excel 2010 inglisekeelne versioon, on tohiks eestikeelse või uuema Exceli versiooniga töötades suuri probleeme tekkida – funktsioonide ja statistikaprotseduuride nimed on samad ning menüükäsud paiknevad ka valdavalt samades kohtades.

NB! Joonistel esitatud valemites on järgnevalt kasutatud arvude kümnendkohtade eraldajana punkti ja valemi argumentide eraldajana koma – see on standard Inglise keeleruumi seadistustes MS Office puhul –, Eesti keeleruumi seadistustes arvuti puhul on arvude kümnendkohtade eraldajaks koma ja valemi argumentide eraldajana tuleb kasutada erinevalt joonistel näidatust semikoolonit.

Materjal ei ole mõeldud matemaatilise statistika teooria õpetamiseks, vastavad eelteadmised eeldatakse kasutajal enesel olemas olevat. Samas on mõnede kasutajalt enam tööd nõudvate punktide juures siiski pisut ka valemeid, samuti on näiteülesannete lahenduste lõpus sõnastatud tulemuste alusel tehtavad järeldused.

Tegu ei ole ka Exceli käsiraamatuga – baasteadmised selleski vallas eeldatakse kasutajal olemas olevaiks (või siis omandatavateks praktilise töö käigus). Seetõttu ei kirjeldata iga punkti juures üksikasjalikult, kuhu tuleks just vajutada ja mis operatsioon läbi viia, et soovitud tulemus saada, samuti ei ole juttu mõnikord 3-4 alternatiivsest variandist, kuidas mingi menüü avada või käsk valida.

Materjalis sisalduvate näidete aluseks on (enamasti) 155 esimese kursuse tudengi ankeedivastuseid sisaldav andmetabel, mille saate alla laadida aadressilt http://ph.emu.ee/~ktanel/andmeanalyys_excelis/ankeet.xlsx.

Selles Exceli failis on töölehel ’Kokku’ kõigi ning töölehtedel ’N’ ja ’M’ eraldi vastavalt neidude ja noormeeste andmed. Tudengite kohta on peale soo teada ka nende pikkus, kehamass, peaümbermõõt, jalanumber, gümnaasiumi matemaatika hinne ja vastused küsimustele: mida te hommikuti tavaliselt sööte, kas te sööte mannaputru, kas teil on lemmikloom, kas te olete olnud viimase 6 kuu jooksul haige, kas te tegelete regulaarselt (tervise)spordiga, kas teil on auto (kasutamise võimalus), kui mitu liitrit õlut te keskmiselt nädalas joote (sama ka jah/ei vastusevariandis), millal te viimati teatris käisite, millal te viimati kinos käisite.

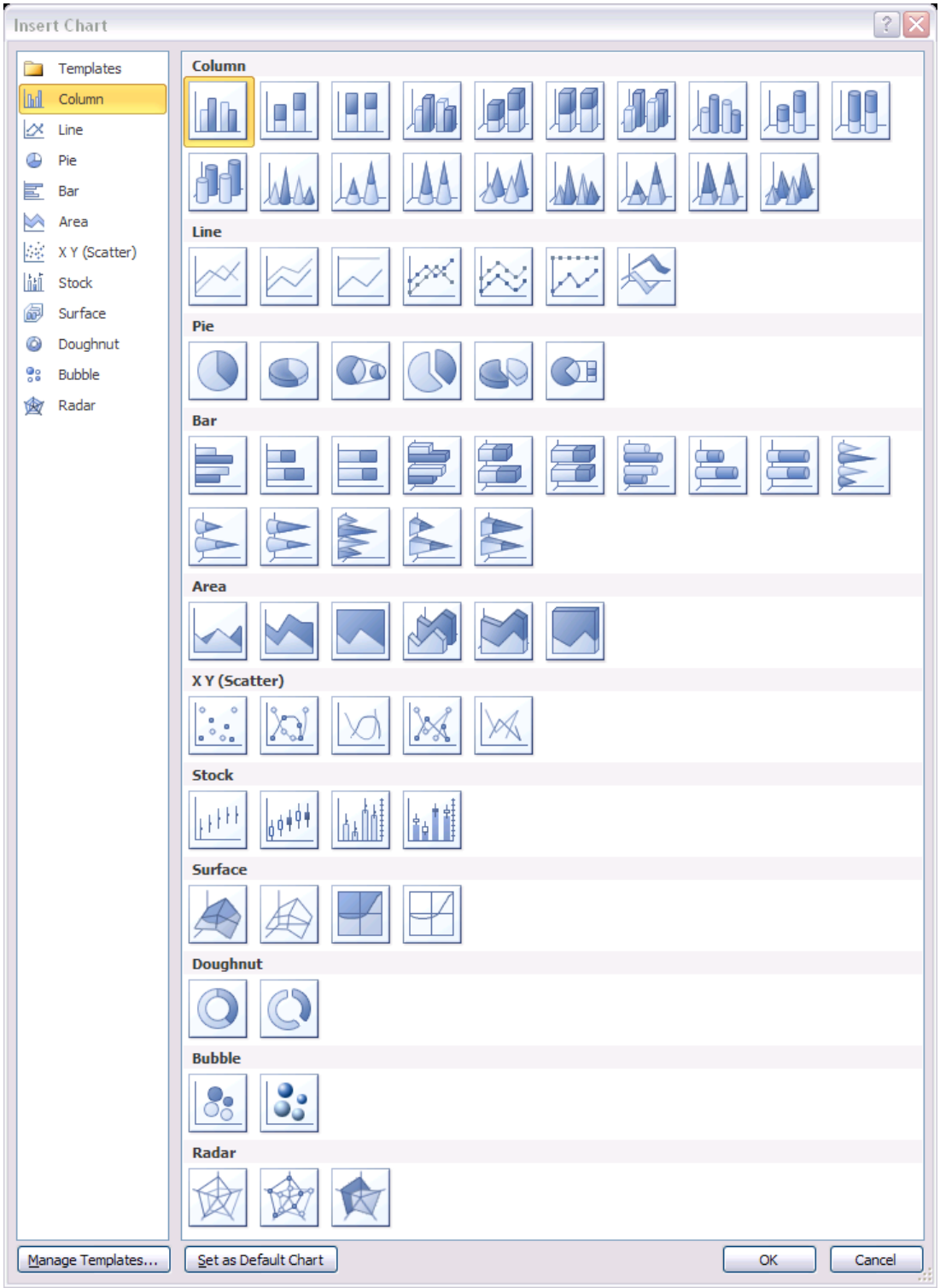
Ja sissejuhatuse lõpetuseks – MS Excel ei ole statistikapakett (miks, vt näiteks http://www.practicalstats.com/xlsstats/excelstats.html).

Samas sobib Excel hästi statistika baaskursuse omandamiseks ning standardsete ja lihtsatel katseplaanidel baseeruvate andmete analüüsiks (seega enamasti ka andmete analüüsiks tudengite bakalaureuse- ja mõnikord ka magistritööde raames). Asjatundliku kasutaja käes on Excel ka küllaltki võimas andmete haldamise ja jooniste tegemise tööriist.

1. Peamised andmeanalüüsi teostamise vahendid MS Excelis

1.1. Joonised

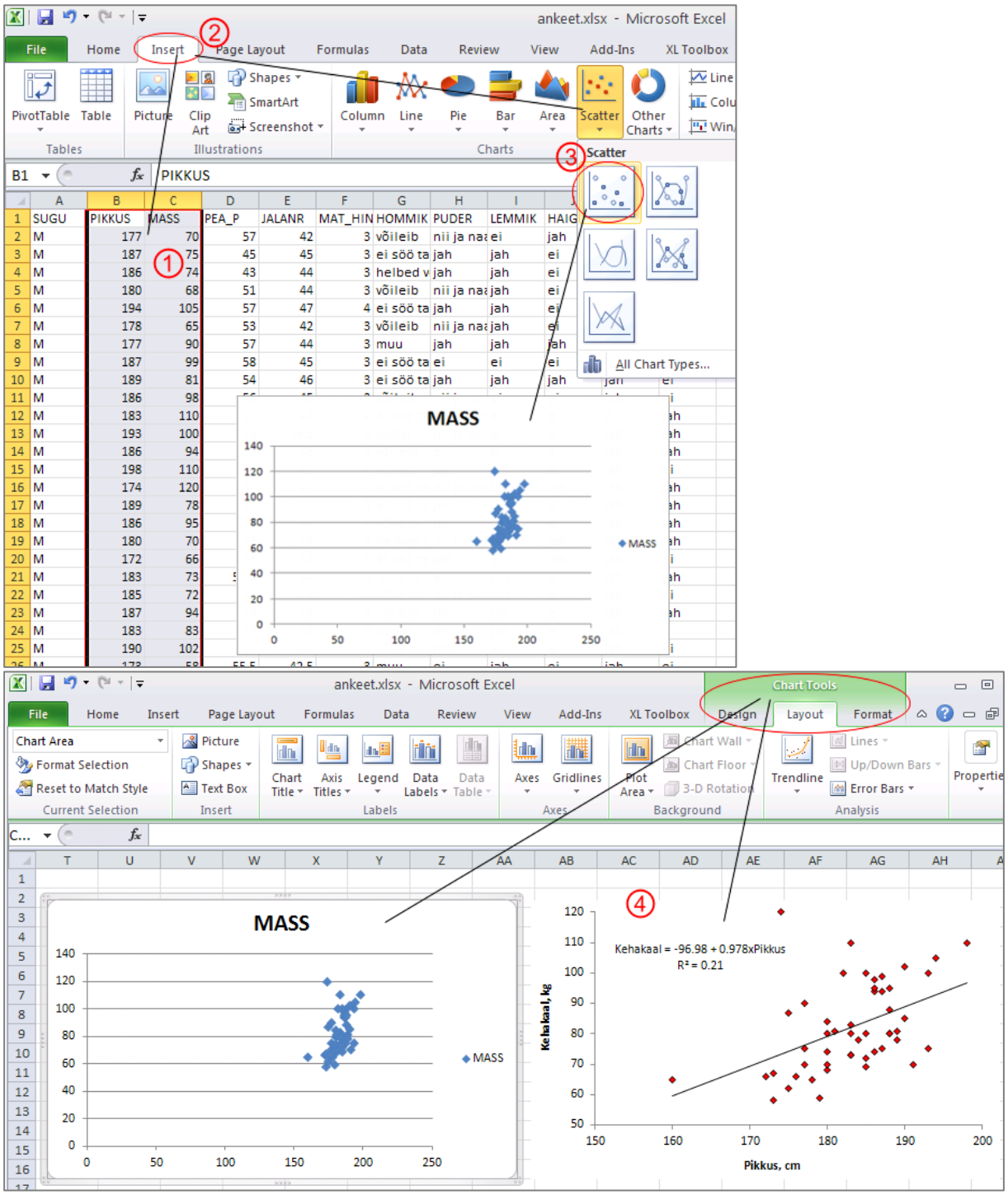
MS Excelis on hulk erinevaid sisse ehitatud joonise tüüpe ja alatüüpe, mis kõik on valitavad menüüsaki *Insert* alt (Joonis 1). Lisaks on erinevad joonised omavahel kombineeritavad ning väga mitmekesisel viisil korrigeeritavad, mistap on võimalik luua mistahes vajadustele vastavaid andmete graafilisi esitusi.



Joonis 1. MS Exceli jooniste tüübid ja alatüübid.

Joonise konstrueerimiseks MS Excelis tuleb (vt ka joonis 2)

1. võtta blokki joonise aluseks olevad andmed (vajadusel tuleb vastav tabel eelnevalt tekitada),
2. klikkida menüüsakil *Insert*,
3. valida Exceli joonisetüüpide hulgas sobiv ning
4. viia joonis telgede, tausta jm korrigeerimise-lisamise-eemaldamise teel sobivale kujule – a) paljud joonise osad on muudetavad neil lihtsalt klikkides, b) enamus täiendavaid muudatusi on tehtavad menüüsaki *Chart Tools* all avanevate valikute abil (**NB!** menüüsakk *Chart Tools* kuvatakse vaid joonise selekteerimise järel).



Joonis 2. Joonise konstrueerimise põhisammud MS Excelis.

NB! MS Excel 2013-s on jooniste modifitseerimise menüükäskude nupud toodud otse joonise kõrvale.

Märkus. Joonis on seotud selle aluseks olevate andmetega – kui andmed muutuvad, muutub automaatselt ka joonis.

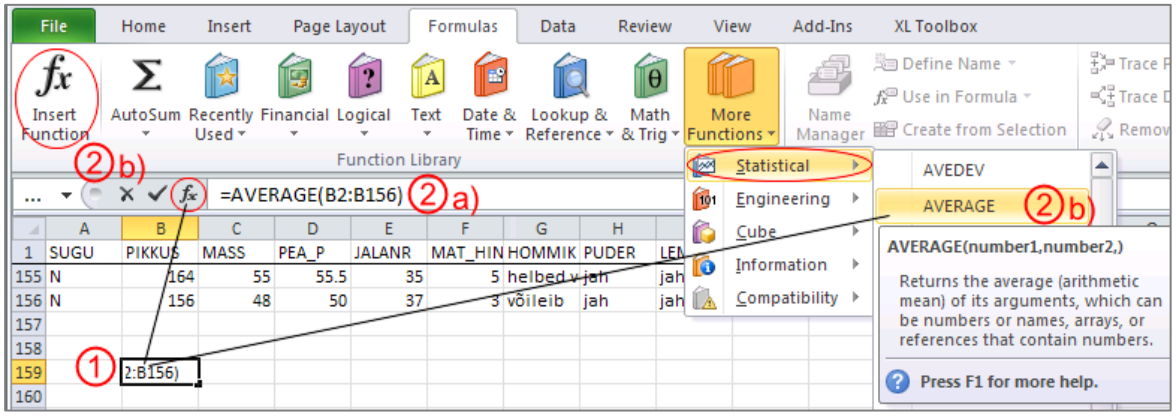
Lisa. MS Exceli poolt vaikimisi loodud jooniste korrigeeerimise, täiendamise ja kombineerimise näiteid võib uurida ka õpiobjektist „MS Excelile mitteomased andmeanalüüsil kasutatavad joonised“ (http://www.eau.ee/~ktanel/joonised_excelis/).

1.2. Funktsioonid

MS Excelis on sadu (valdkondade kaupa grupeeritud) valmis funktsioone, lisaks on erinevate funktsioonide ja tehete kombineerimise läbi võimalik defineerida just kasutaja vajadustele vastavad arvutuskäsud.

Funktsioonide rakendamiseks tuleb

1. **panna kursor lahtrisse, kuhu soovitakse saada funktsiooni tulemust** (massiivi-funktsiooni korral tuleb võtta blokki tulemustabeli jagu lahtreid) ning
2. anda ette arvutuskäsk. Viimaseks on mitmeid võimalusi:
   a) kiireim ja lihtsaim viis on trükkida **võrdusmärgiga alustatud** käsk otse selekteeritud lahtrisse või valemireale (eelduseks on vähemalt rakendatava funktsiooni nime alguse teadmine),
   b) alternatiiv on klikkida valemi rea alguses paikneval nupul f_x, samal nupul menüü-sakil *Formulas* või valida vastav funktsioon *Formulas*-sakil olevate funktsioonide kategooriate ja nende all avanevate funktsioonide hulgast (Joonis 3).

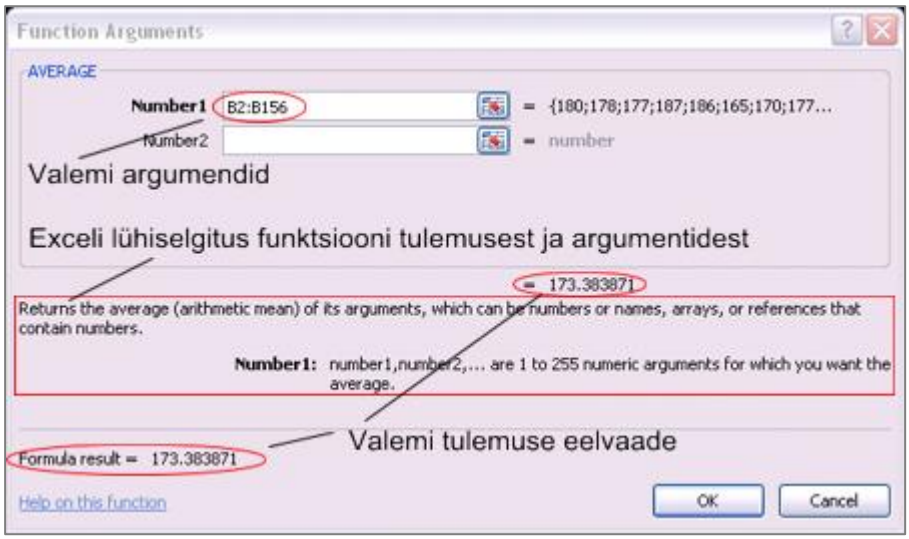


Joonis 3. Funktsioonide rakendamine MS Excelis.

Kui funktsiooni valik toimub menüürealt, avaneb andmete ette andmiseks vastav menüüaken (Joonis 4), mis sisaldab ka lühiselgitust funktsiooni tulemusest ja argumentidest ning kuvab andmete etteandmise järel funktsiooni tulemuse.

Exceli töölehel paiknevate andmete funktsioonile ette andmiseks võib selekteerida vastavad lahtrid hiirega (või klaviatuuri abil), trükkida lahtribloki aadressi ise (a’la B2:B156) või anda ette analüüsitava lahtribloki nime (viimase olemasolul).

NB! Tähele tuleb panna seda, et Exceli statistikafunktsioonid tahavad saada andmeid ette ilma andmetabeli esimeses reas paikneva tunnuse nimeta!



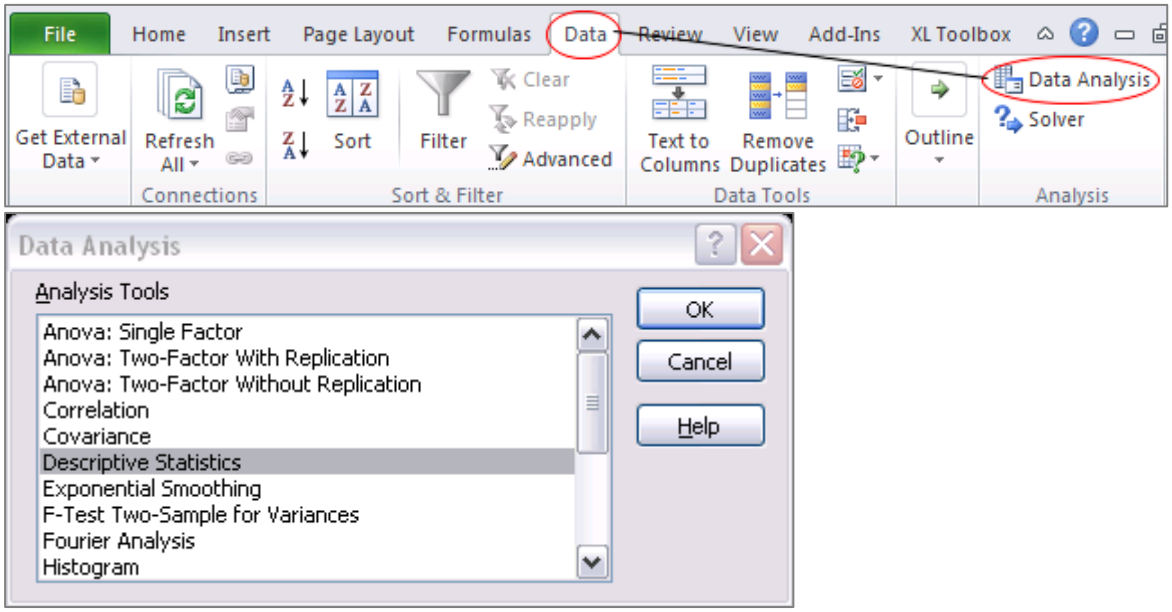
Joonis 4. Funktsiooni AVERAGE rakendamise aken.

Märkus 1. Funktsiooni tulemus on seotud selle aluseks olevate andmetega – kui andmed muutuvad, muutub automaatselt ka tulemus.

Märkus 2. Kõiksugu kokkuvõtlike karakteristikute ja andmetabeli vahele on alati kasulik jätta vähemalt üks tühi rida ja/või veerg, sest sellisel juhul ei loe Excel andmete sorteerimisel, filtreerimisel, *Pivot Table*'i rakendamisel jne arvutustulemusi andmetabeli osaks.

1.3. Protseduurid

MS Exceli statistikaprotseduuride loetelu avaneb menüü-sakilt *Data* valiku *Data Analysis* alt (Joonis 5).



Joonis 5. MS Exceli statistikaprotseduurid.

NB! Kui valik *Data Analysis* menüü-sakil *Data* puudub, tuleb järgida järgmist menüü-teekonda ning statistikaprotseduure sisaldav moodul nimega *Analysis ToolPak* sisse lülitada:

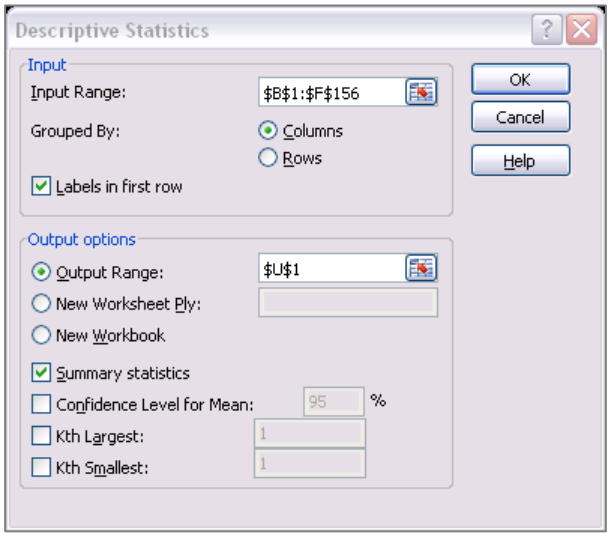
File -> Options -> Add-Ins -> Manage |Excel Add-ins| [*Go...*] -> *Analysis ToolPak*.

Sarnaselt funktsioonidega tuleb iga statistikaprotseduuride loetelust valitud analüüsi korral sisestada argumendid (andmeblokid). Kuid erinevalt funktsioonidest võib sisestatav andmeblokk sisaldada esimeses reas ka tunnuse nime, mida programm kasutab hiljem tulemuste väljatrükis. Sellisel juhul tuleb teha "linnuke" nimetuse *Labels* (*in First Row*) ees olevasse kasti (Joonis 6).

Samuti tuleb erinevalt funktsioonidest määrata tulemuste väljastamise asukoht:

- *Output Range* – tulemus väljastatakse olemasolevale lehele, määrata tuleb väljundi vasaku ülemise nurga aadress;
- *New Worksheet Ply* – vaikimisi valik, tulemus väljastatakse uuele loodavale töölehele (soovi korral saab viimasele anda ka nime, trükkides selle valiku taga asuvasse tekstikasti);
- *New Workbook* – tulemus väljastatakse uude loodavasse tööraamatusse (faili).

Ülejäänud valikud sõltuvad juba konkreetsest protseduurist ja saavad kirjeldatud selle õpetuse järgnevais osades.



Joonis 6. Statistikaprotseduuri *Descriptive Statistics* tellimisaken.

Võrreldes funktsioonidega sisaldab statistikaprotseduuride väljund märksa enam informatsiooni, koosnedes tavaliselt ühest või mitmest tabelist ja/või joonisest (Joonis 7).

PIKKUS		MASS		PEA_P		JALANR		MAT_HINNE	
Mean	173.384	Mean	68.8039	Mean	54.8774	Mean	40.5323	Mean	3.87097
Standard Error	0.75233	Standard	1.18499	Standard	0.30819	Standard	0.24078	Standard	0.0584
Median	172	Median	65	Median	56	Median	40	Median	4
Mode	180	Mode	55	Mode	56	Mode	39	Mode	4
Standard Deviation	9.36643	Standard	14.7531	Standard	3.83691	Standard	2.99766	Standard	0.72712
Sample Variance	87.7299	Sample \	217.653	Sample \	14.7219	Sample \	8.98597	Sample \	0.5287
Kurtosis	-0.55086	Kurtosis	0.74153	Kurtosis	3.95812	Kurtosis	-0.58018	Kurtosis	-1.08018
Skewness	0.24627	Skewnes	0.9816	Skewnes	-1.82998	Skewnes	0.52478	Skewnes	0.20278
Range	47	Range	73.5	Range	21	Range	13	Range	2
Minimum	151	Minimun	46.5	Minimun	41	Minimun	35	Minimun	3
Maximum	198	Maximun	120	Maximun	62	Maximun	48	Maximun	5
Sum	26874.5	Sum	10664.6	Sum	8506	Sum	6282.5	Sum	600
Count	155	Count	155	Count	155	Count	155	Count	155

Joonis 7. Statistikaprotseduuri *Descriptive Statistics* tulemus.

Märkus. Erinevalt joonistest ja funktsioonidest ei ole statistikaprotseduuride tulemused lingitud nende aluseks olevate andmetega – andmete hilisem muutmine ei muuda varem rakendatud statistikaprotseduuride tulemusi.

1.4. Risttabelid (*PivotTable*)

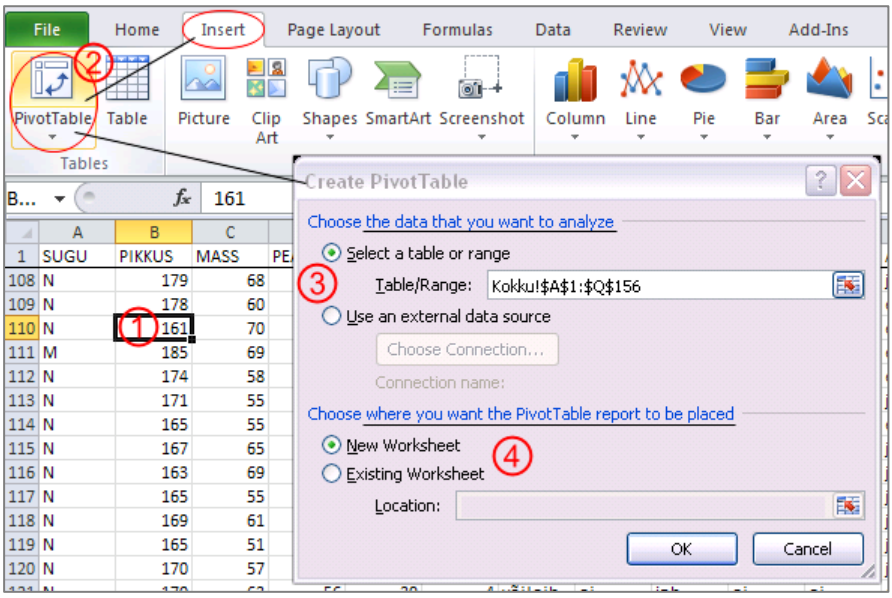
Vahendi *PivotTable* abil on võimalik kiirelt konstrueerida erinevaid kokkuvõtlikke tabeleid (ja vahendi *PivotChart* abil ka jooniseid).

PivotTable'i rakendamine eeldab, et andmetabelis on igal veerul nimi ning need nimed ei kordu.

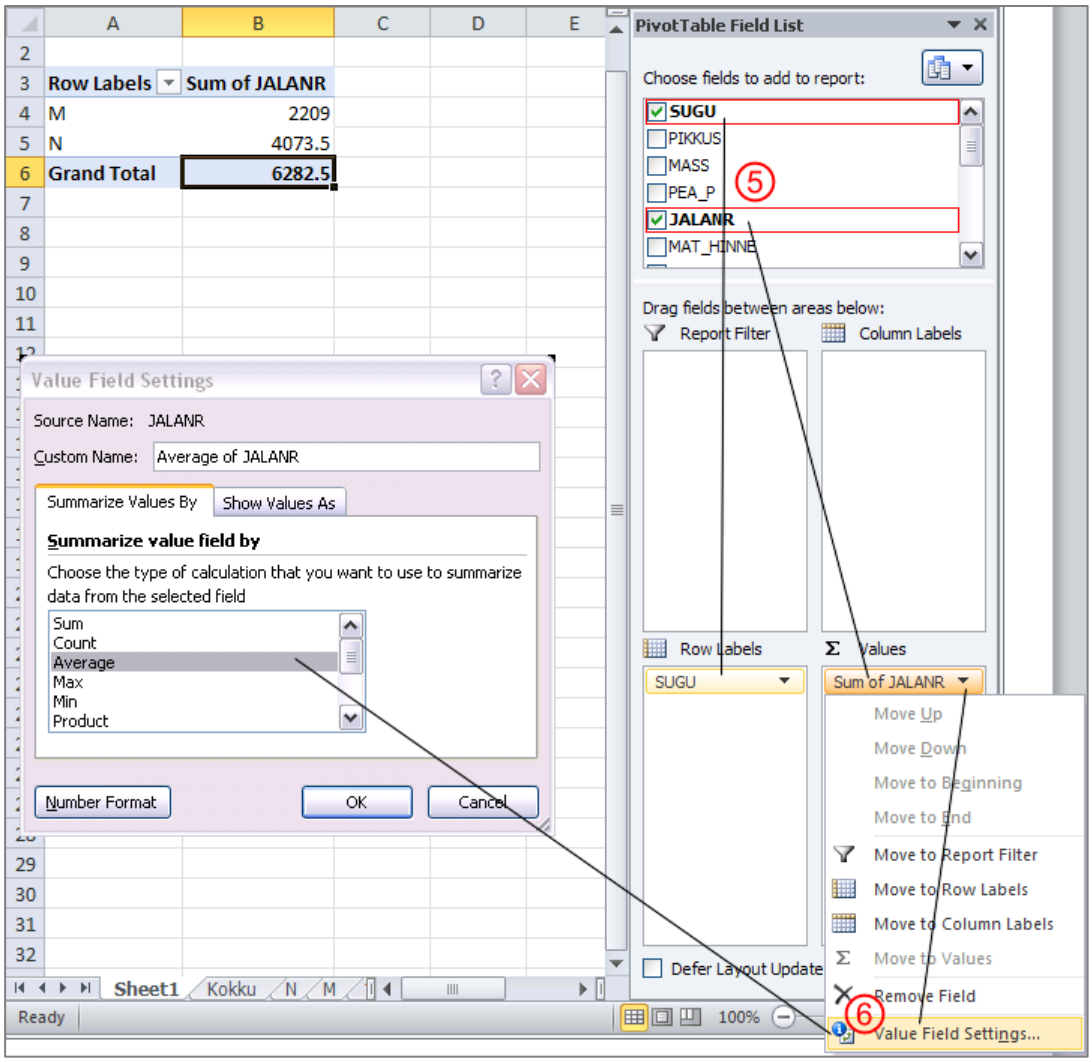
Risttabeli konstrueerimiseks (Joonised 8 ja 9) tuleb

1. esmalt panna kursor andmetabeli suvalisse lahtrisse (see on soovitatav, sest siis võtab Excel konstrueeritava tabeli aluseks automaatselt kõik andmed – jääb ära analüüsitavate andmete ette andmise lisatöö),
2. seejärel valida *Insert*-sakilt käsk *PivotTable*,
3. kahtluste korral kontrollida, kas Excel ikka valis analüüsitavaks kõik vajalikud andmed (või juhul, kui esmalt jäi kursor andmetabelisse panemata, anda ette analüüsitavate andmete asukoht),
4. anda ette loodava tabeli vasaku ülemise nurga asukoht (*Existing Worksheet*) või lasta Excelil teha loodava tabeli tarvis uus tööleht (*New Worksheet* – so vaikimisi variant),
5. lohistada tabeli konstrueerimise väljal tunnus, mille väärtuste alusel soovitakse tabelit ridadeks (või veergudeks) jagada, lahtrisse *Row Labels* (*Column Labels*), ning tunnus, mille kohta soovitakse midagi arvutada, lahtrisse *Values*,
6. muuta vajadusel arvutuskäsku (*Values*-lahtris paiknevale tunnusele rakendatud funktsiooni) valiku *Value Field Settings* või nupu *Summarize Values By* abil.

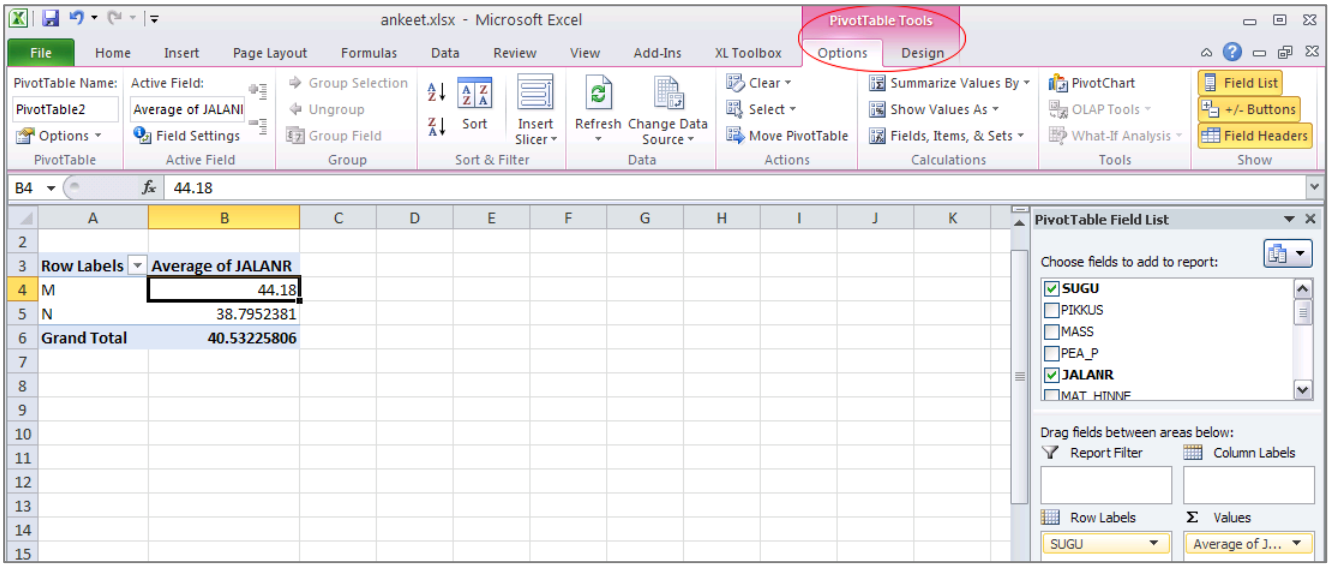
Nii valmis tabeli sisu kui ka väljanägemist on võimalik muuta ja täiendada spetsiaalse *PivotTable Tools* menüüsaki all leiduvate käskude abil (Joonis 10).



Joonis 8. *PivotTable* tellimine.



Joonis 9. *PivotTable* konstrueerimine.



Joonis 10. *PivotTable* muutmise käsud menüüsakil *PivotTable Tools*.

2. Sagedustabelid

2.1. Sagedustabel pidevale arvtunnusele

Sagedustabeli konstrueerimiseks pidevale arvtunnusele on MS Excelis kolm moodust: funktsioon FREQUENCY, statistikaprotseduur *Histogram* ja PivotTable.

Enne pideva arvutunnuse väärtuste grupeerimist tuleb otsustada, kui mitmesse ja millise suurusega klassi tunnuse väärtused jagada. Funktsioonile FREQUENCY ja (soovi korral) statistikaprotseduurile *Histogram* tuleb rühmitamiseeskiri ette anda rühmade ülemiste piiride bloki näol, st et klasside ülemised piirid tuleb sisestada Exceli töölehele.

Järgnevalt vaatame näitena tudengite pikkuse sagedustabeli konstrueerimist.

Sagedustabeli konstrueerimine pidevale arvtunnusele funktsiooni FREQUENCY abil

Kõige kiirem variant lasta Excelil kokku lugeda, kui palju vaatlusi mingisse ette antud klassi kuulub, on kasutada funktsiooni FREQUENCY.

Erinevalt enamustest MS Exceli funktsioonidest on funktsioon FREQUENCY massiivi-funktsioon, st et selle funktsiooni tulemuseks ei puugi olla üks väärtus eelnevalt valitud lahtris, vaid hulk väärtusi eelnevalt valitud lahtriteblokis.

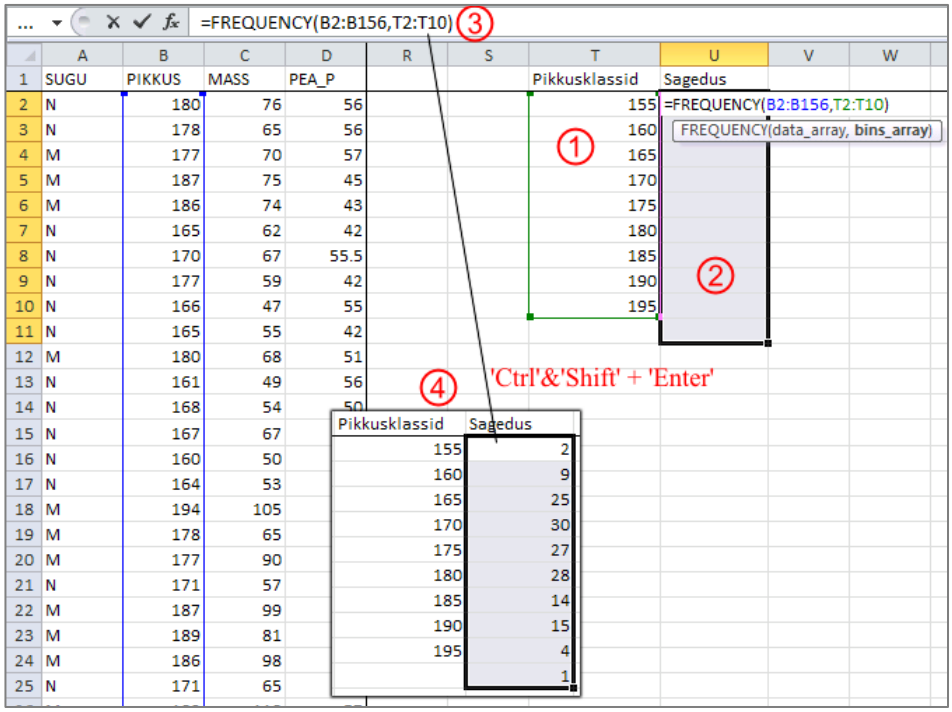
Sagedustabeli konstrueerimiseks funktsiooni FREQUENCY abil tuleb (Joonis 11a)

1. sisestada Exceli töölehele loodavate pikkusklasside ülemised piirid (näiteks soovides jagada tudengite pikkused kümnesse 5 cm pikkusesse klassi kujul 150-155, 155-160, …, 190-195 ja 195-200 cm, tuleb Exceli töölehele sisestada väärtused 155, 160, …, 195); **NB!**
   a) viimase klassi piiri ei ole vaja ette anda, sest Excel genereerib alati ise ühe lisaklassi rühmitamiseeskirjaga mittemääratud väärtuste tarvis (antud juhul siis tudengitele pikkusega üle 195 cm),
   b) klassi ülemise piiri loeb Excel kuuluvaks klassi sisse, st et väärtus 155 tähendab Exceli jaoks kõiki pikkuseid mis on väiksemad või võrdsed 155-st, väärtus 160 kõiki pikkuseid mis on väiksemad või võrdsed 160-st, aga ei kuulu eelmistesse klassidesse, jne;
2. võtta blokki lahtrid töölehel kohas, kuhu soovitakse sagedusi arvutada; arvutatavate sageduste ja seeläbi blokki võetavate lahtrite arv on määratud konstrueeritava sagedustabeli klasside arvuga (üks täiendav blokki võetud lahter vastab Exceli poolt täiendavalt moodustatavale klassile);
3. trükkida selekteeritud lahtriblokki (**NB!** koheselt sellesse lahtrisse, millest blokki võtmist alustasite, uuesti klikkida esimesel lahtril ei tohi!!) valem

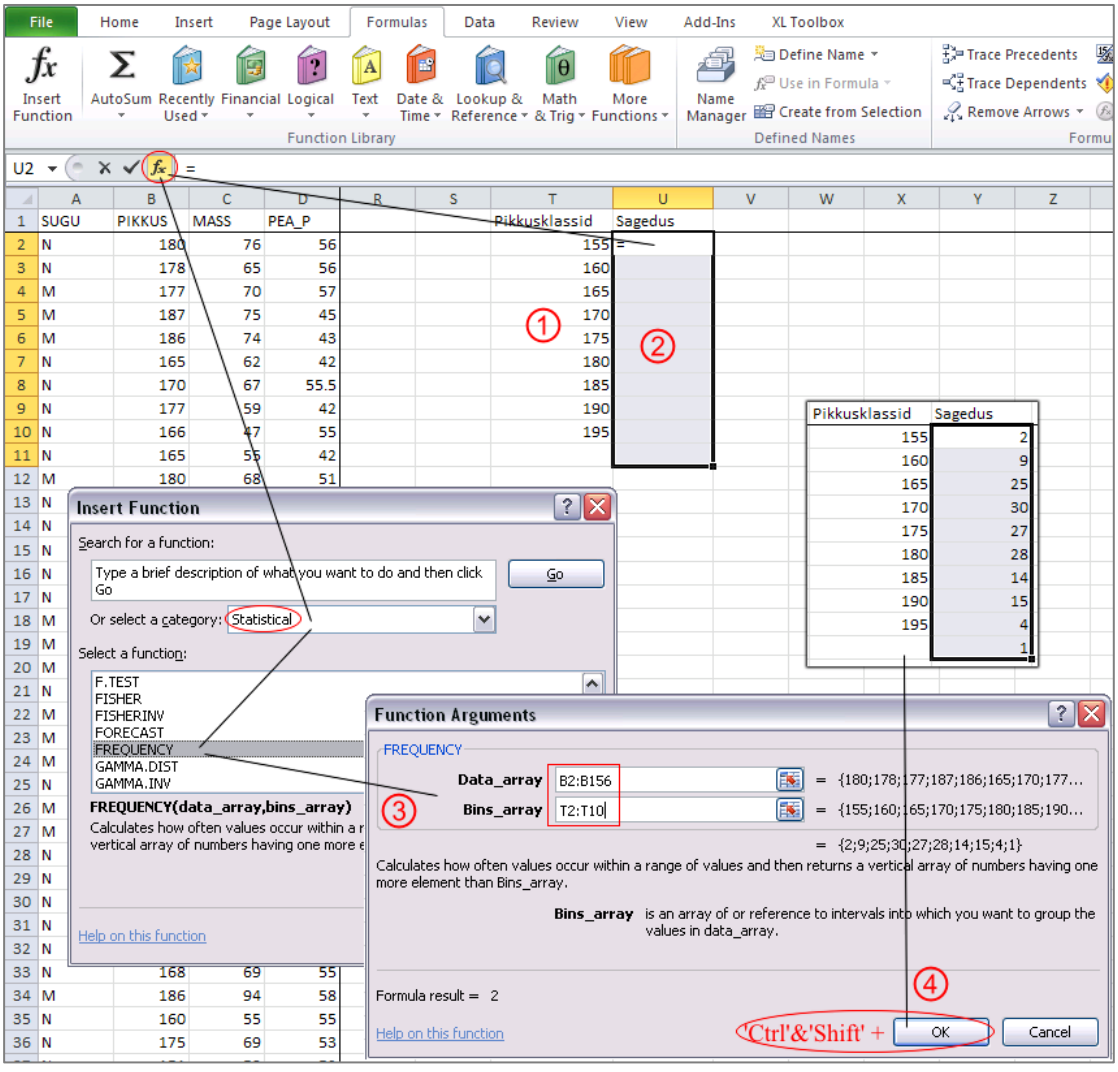
   =FREQUENCY(B2:B156;T2:T10)

   kus esimene argument annab ette uuritava tunnuse väärtused (antud juhul tudengite pikkused) ja teine argument klasside ülemised piirid;
4. vajutada alla klahvid **Ctrl** ja **Shift** ning seejärel **Enter** (st. 3 klahvi korraga).

Alternatiiv taolisele funktsiooni klaviatuurilt sisestamisele on lisada funktsioon menüüsid ja abiaknaid kasutades (Joonis 12).



Joonis 11. Sagedustabeli konstrueerimine pidevale arvtunnusele funktsiooniga FREQUENCY.

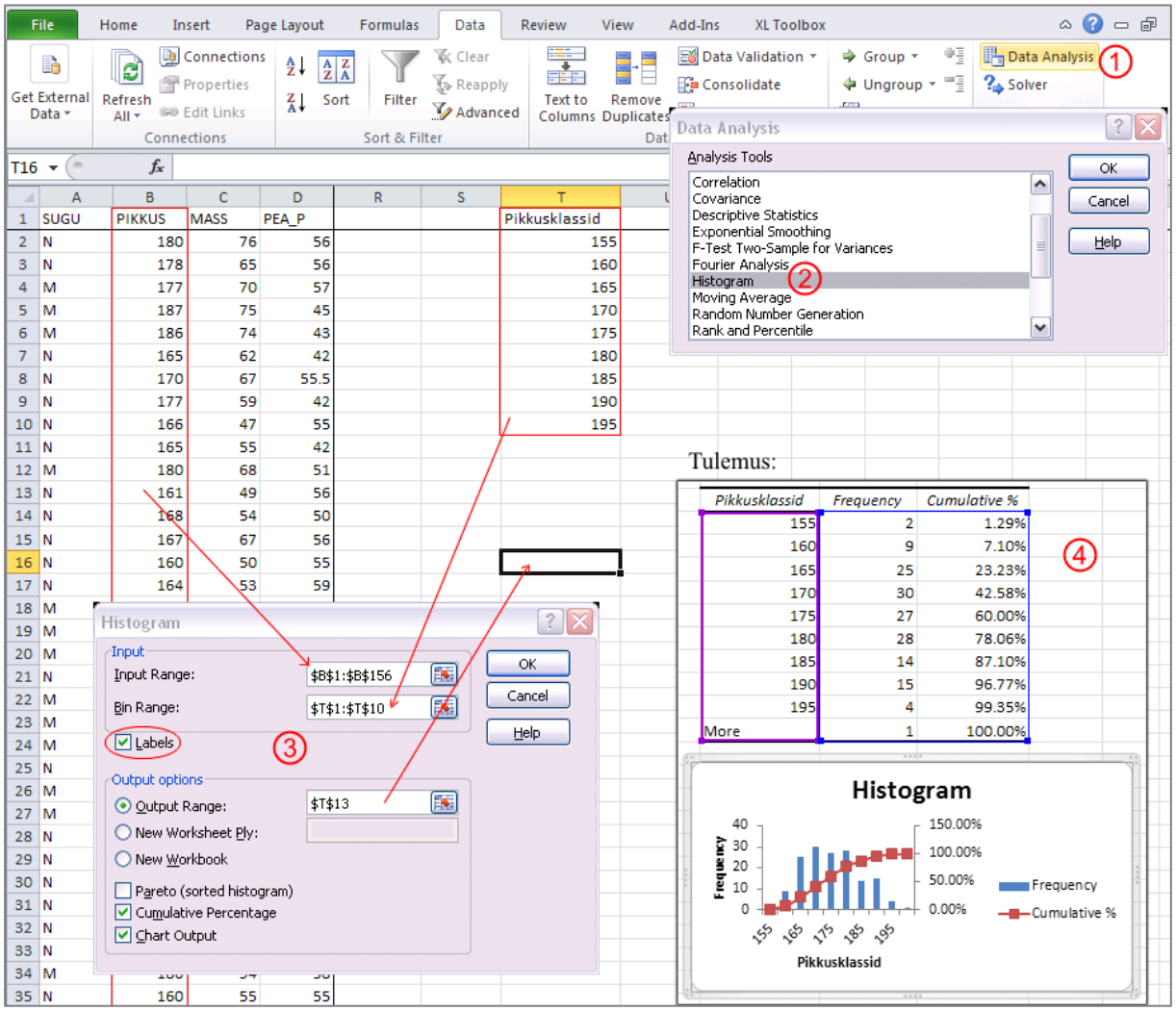


Joonis 12. Sagedustabeli konstrueerimine pidevale arvtunnusele funktsiooniga FREQUENCY menüüde ja abiakende abil.

Sagedustabeli konstrueerimine pidevale arvtunnusele protseduuri *Histogram* abil

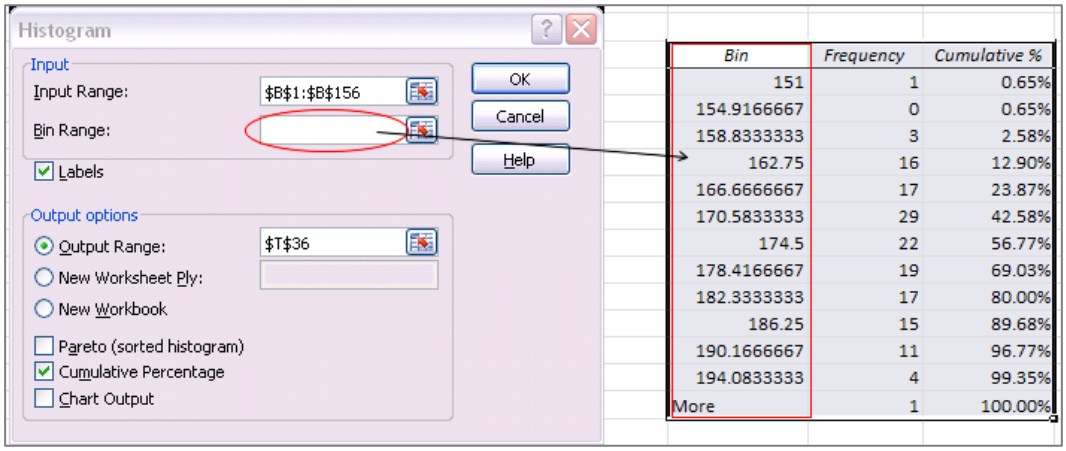
Sagedustabeli konstrueerimiseks protseduuri *Histogram* abil tuleb (Joonis 13)

1. valida *Data*-sakilt käsk *Data Analysis* ja
2. avanenud protseduuride loetelust *Histogram*;
3. avanenud aknas tuleb/võib täita järgmised väljad:
   - *Input Range* – algandmete blokk (tavaliselt üks tulp);
   - *Bin Range* – rühmade ülemiste piiride väärtuste blokk;
   - *Labels* – märgitakse tunnuse nime või tähise olemasolu korral andmebloki ülemises reas (**NB!** kui uuritava tunnuse väärtused on ette antud koos nimega, peab nimi olema ka klassipiiride blokil);
   - *Output options* – määratakse tulemuste väljastamise asukoht: samale töölehele (*Output Range*), uuele töölehele (*New Worksheet Ply*) või uude faili (*New Workbook*);
   - *Pareto* (*sorted histogramm*) – klassid järjestatkse nende sageduste alusel kahanevas järjekorras;
   - *Cumulative Percentage* – arvutatakse jaotusfunktsiooni väärtused;
   - *Chart Output* – tulemused väljastatakse lisaks tabelile ka graafikul (tulpdiagrammina);
4. tulemuseks saadud tabelis võiks selguse mõttes asendada klasside ülemised piirid klasside tegelike väärtustega ja joonis ei ole just kõige ilusam (kuigi esmase ettekujutuse saamiseks tudengite pikkuste jaotumisest kõlbab küll).



Joonis 13. Sagedustabeli konstrueerimine pidevale arvtunnusele protseduuriga *Histogram*.

NB! Lahtri *Bin Range* protseduuri *Histogram* tellimisaknas võib jätta ka tühjaks – siis moodustab Excel klassid ise. Reeglina ei klasside piirid siis küll „ümmargused“ arvud, aga esmase ettekujutuse sellest, kui mitu ja kui suure ulatusega klassi võiks antud andmetel konstrueerida, saab nii küll (Joonis 14).



Bin	Frequency	Cumulative %
151	1	0.65%
154.9166667	0	0.65%
158.8333333	3	2.58%
162.75	16	12.90%
166.6666667	17	23.87%
170.5833333	29	42.58%
174.5	22	56.77%
178.4166667	19	69.03%
182.3333333	17	80.00%
186.25	15	89.68%
190.1666667	11	96.77%
194.0833333	4	99.35%
More	1	100.00%

Joonis 14. Sagedustabeli konstrueerimine pidevale arvtunnusele protseduuriga *Histogram* Exceli moodustatud klasside korral.

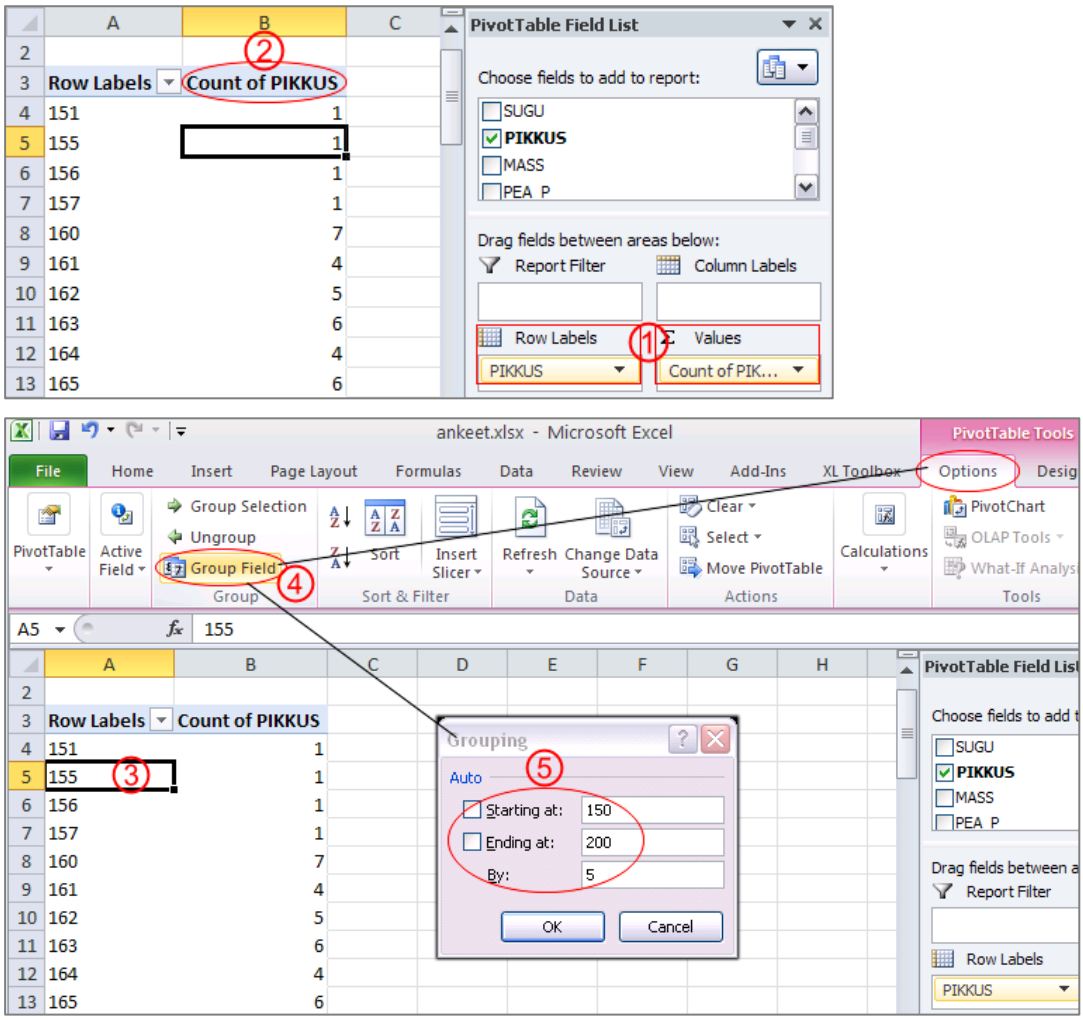
Sagedustabeli konstrueerimine pidevale arvtunnusele *PivotTable* abil

PivotTable abil saab pidevale arvtunnusele konstrueerida sagedustabeli vaid siis, kui tunnuse kõik väärtused on mõõdetud (pole puuduvaid väärtuseid).

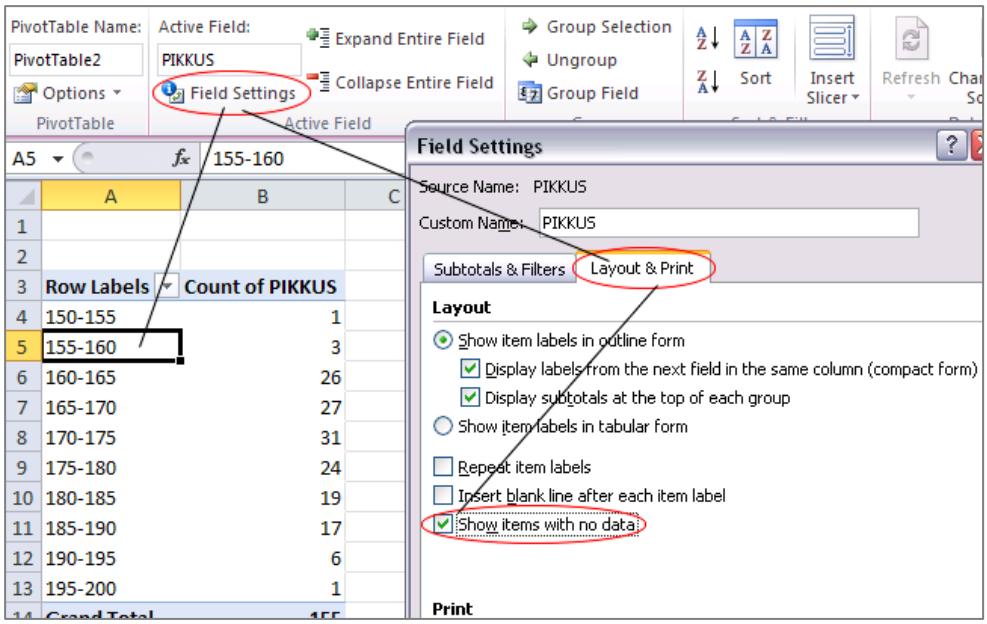
1. Esmalt tuleb konstrueerida tabel, kus nii grupeerivaks tunnuseks (*Row Label*) kui ka arvutuste aluseks olevaks tunnuseks (*Values*) on uuritav pidev arvtunnus (Joonis 15a),
2. seejärel peab arvutuskäsuks Exceli poolt arvtunnustele vaikimisi rakendatava summa asemel määrama loenduse (*Count*),
3. paigutama kursori mistahes lahtrisse veerus *Row Labels*,
4. valima kas hiire parempoolse nupu kliki järel avanenud rippmenüüst käsu *Group* või *PivotTable Tools -> Options -> Group Field* (Joonis 15b) ning
5. määrama avanenud aknas klasside algus- ja lõpp-punkti ning sammu (klassi suuruse).

NB! Vaikimisi ei kuva Excel klasse, kuhu ei kuulu ühtegi vaatlust! Olemaks kindel, et konstrueeritud sagedustabelis on kõik klassid – ka need, kuhu ühtegi vaatlust ei kuulu – kirjas,

- tuleks kursori paiknemisel veerus *Row Labels* klikkida käsul *Field Settings*,
- valida avanenud aknas lehekülg *Layout & Print*
- ning märkida ära valik *Show items with no data* (Joonis 16).



Joonis 15. Sagedustabeli konstrueerimine pidevale arvtunnusele *PivotTable* abil.

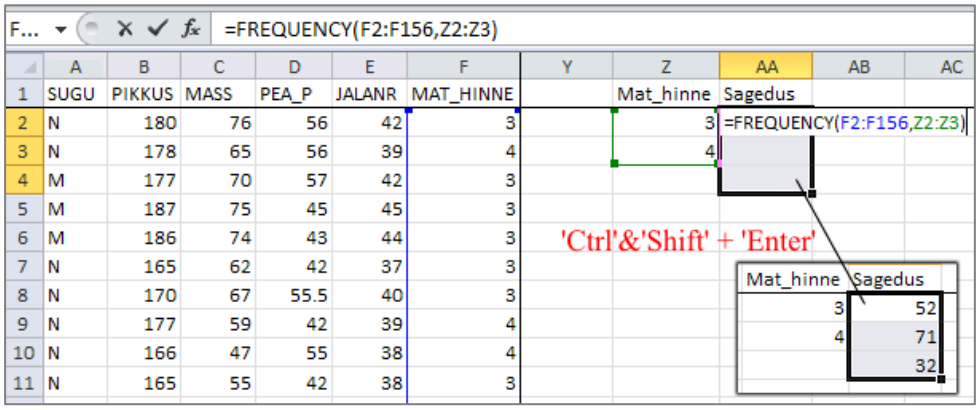


Joonis 16. Käsurada sundimaks *PivotTable*'t kuvama ka tühje klasse.

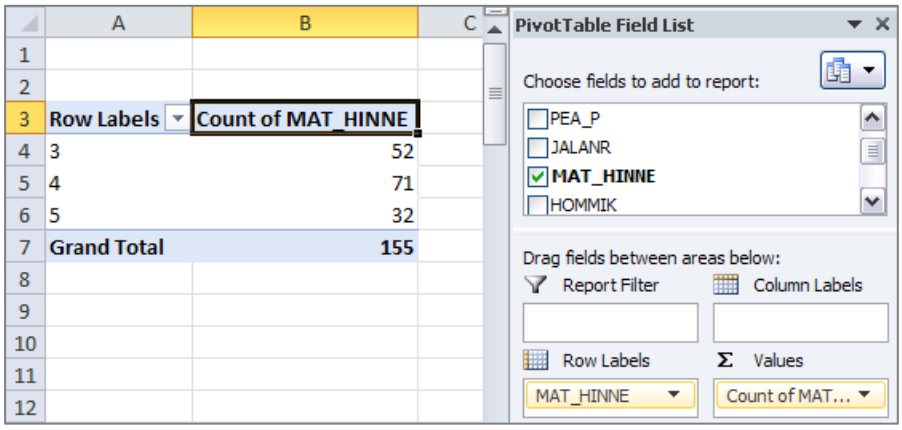
2.2. Sagedustabel diskreetsele arvtunnusele

Sagedustabeli konstrueerimiseks diskreetsele arvutunnusele on kasutatavad nii eelmises punktis käsitletud funktsioon FREQUENCY, protseduur *Histogram* kui ka *PivotTable*.

- Kahe esimese puhul peab lihtsalt klasside piiride näol tegema abitabeli kõigist diskreetse arvtunnuse väärtustest (va viimane väärtus; Joonis 17),
- *PivotTable* puhul aga jääb ära klasside moodustamine – grupeeriva tunnuse väärtused ongi ise klassid (Joonis 18).



Joonis 17. Tudengite matemaatika hinnete sagedustabeli konstrueerimine funktsiooniga FREQUENCY.

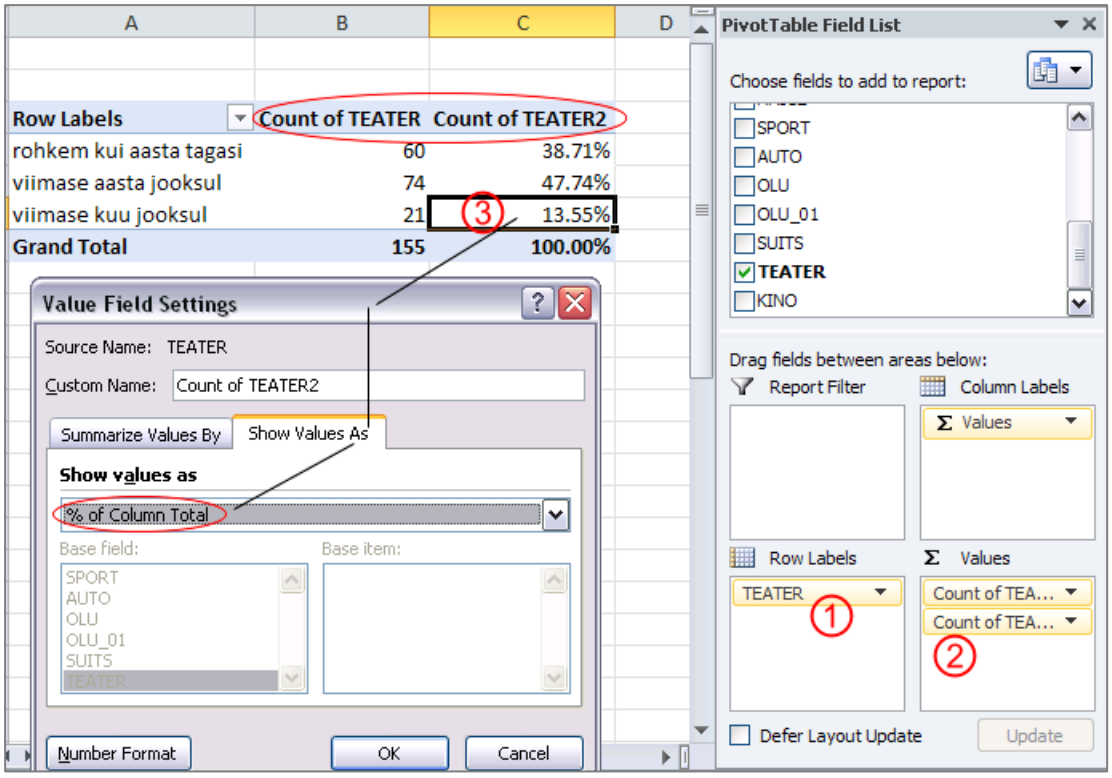


Joonis 18. Tudengite matemaatika hinnete sagedustabeli konstrueerimine *PivotTable* abil.

2.3. Sagedustabel mittearvulisele tunnusele

Mittearvulise tunnuse puhul on lihtsaim vahend sagedustabeli konstrueerimiseks *PivotTable*. Seejuures on põhimõte identne sagedustabeli konstrueerimisega diskreetsele arvutunnusele (Joonis 19):

1. loodav tabel tuleb jagada ridadeks uuritava mittearvulise tunnuse väärtuste järgi,
2. sama tunnus tuleb lohistada ka *Values*-lahtrisse ning määrata vajadusel arvutuskäsuks loendus (*Count*);
3. lohistades sama tunnuse *Values*-lahtrisse ka teine kord, on võimalik samas tabelis lisaks absoluutsetele sagedustele kuvada ka suhtelisi sagedusi (*Value Field Settings -> Show Values As -> % of Column Total*).



Joonis 19. Tudengite teatriskäimise sagedustabeli konstrueerimine *PivotTable* abil.

3. Arvkarakteristikud

3.1. Valemid ja funktsioonid

MS Exceli funktsioonidel on hulk positiivseid omadusi:

- argumentide ette andmine on enamasti intuitiivselt mõistetav,
- tänu võimalusele funktsioone kopeerida on kord juba sisestatud käsud lihtsalt rakendatavad uutele argumentidele (väärtustele, tunnustele),
- funktsioonide omavaheline kombineerimine võimaldab väljastada keeruliste avaldiste tulemusi.

Enamuste kirjeldavate karakteristikute väärtuste arvutamiseks on Excelis oma funktsioon, mille nimi tuleneb karakteristiku inglisekeelsest nimest. Näiteks arvutavad funktsioonid AVERAGE aritmeetilise keskmise ja MEDIAN mediaani väärtuse, MIN ja MAX minimaalse ja maksimaalse väärtuse.

Osade karakteristikute, mille arvutusvalem on valimi ja populatsiooni puhul erinev, väärtuste leidmiseks on Excelis eraldi funktsioonid, mida eristab funktsiooni nime lõpus olev laiend: .S valimi ja .P populatsiooni karakteristiku puhul. Näiteks arvutab funktsioon STDEV.S valimi standardhälbe valemist

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

ja STDEV.P populatsiooni standardhälbe valemist

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\ .$$

Joonisel 20 on kujutatud rea olulisemate arvkarakteristikute rakendamist tudengite pikkusele.

NB! Excelis puudub funktsioon (aritmeetilise keskmise) standardvea arvutamiseks. Aga teades standardvea arvutamise valemit:

$$se = s/\sqrt{n}$$

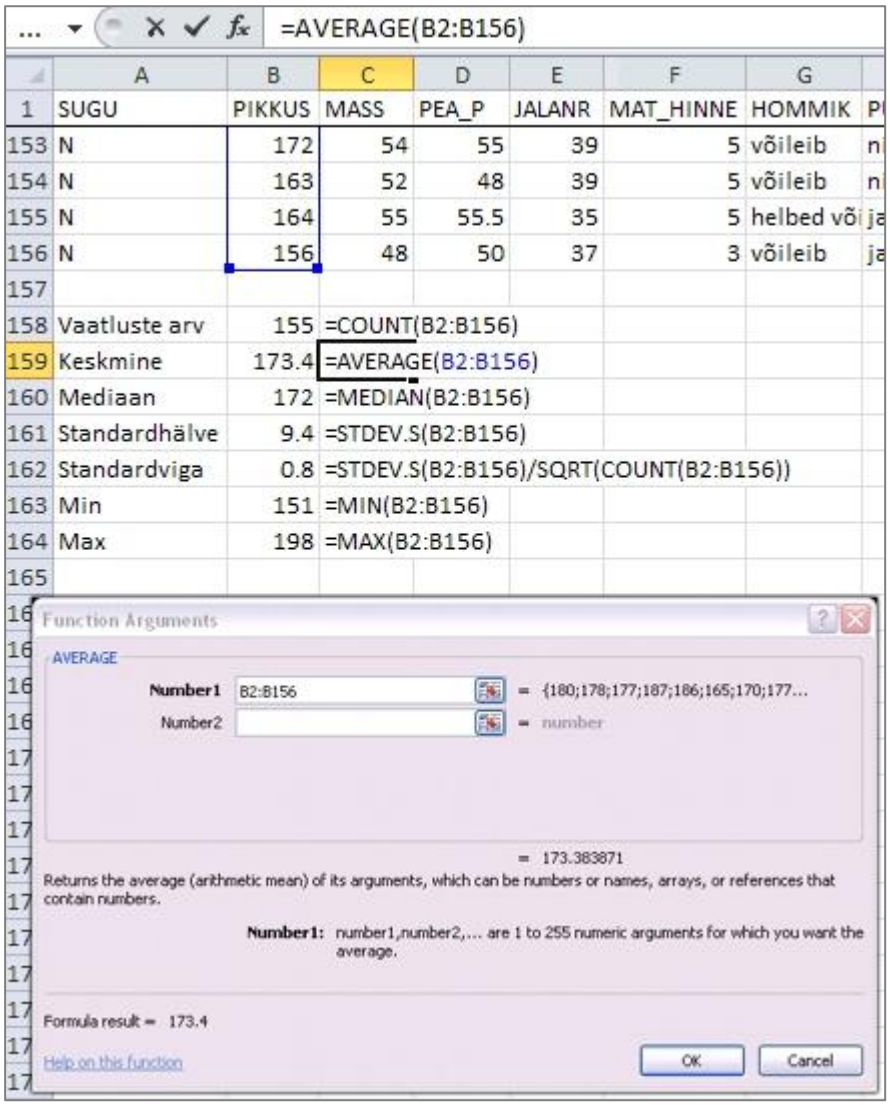
(standardviga võrdub standardhälve jagatud ruutjuurega vaatluste arvust), saab selle vajadusel ikkagi arvutada, andes Excelile ette vajaliku valemi (Joonis 20).

3.2. Protseduur *Descriptive Statistics*

Protseduur *Descriptive Statistics* väljastab ühekorraga 13 erinevat arvkarakteristikut, lisaks liidetava keskmise usalduspiiride arvutamiseks ja etteantud järjekorranumbriga väärtused. Arvutused võib teostada korraga ka enam kui ühe tunnusega, ainukesed tingimused on et tunnused peavad olema arvulised ja paiknema andmetabelis kõrvuti.

Protseduuri *Descriptive Statistics* rakendamiseks tuleb (Joonis 21)

1. valida *Data*-sakilt käsk *Data Analysis* ning seejärel avanenud aknast protseduur *Descriptive Statistics*,
2. anda ette analüüsitavad andmed ja määrata väljundi asukoht:
   - *Input Range* – algandmete blokk (võib sisaldada ka mitut veergu või rida, st et korraga võib analüüsida ka mitut tunnust eeldusel, et need on arvulised ja paiknevad andmetabelis kõrvuti);



Joonis 20. Mõningate arvkarakteristikute leidmine tudengite pikkusele Exceli funktsioonide abil.

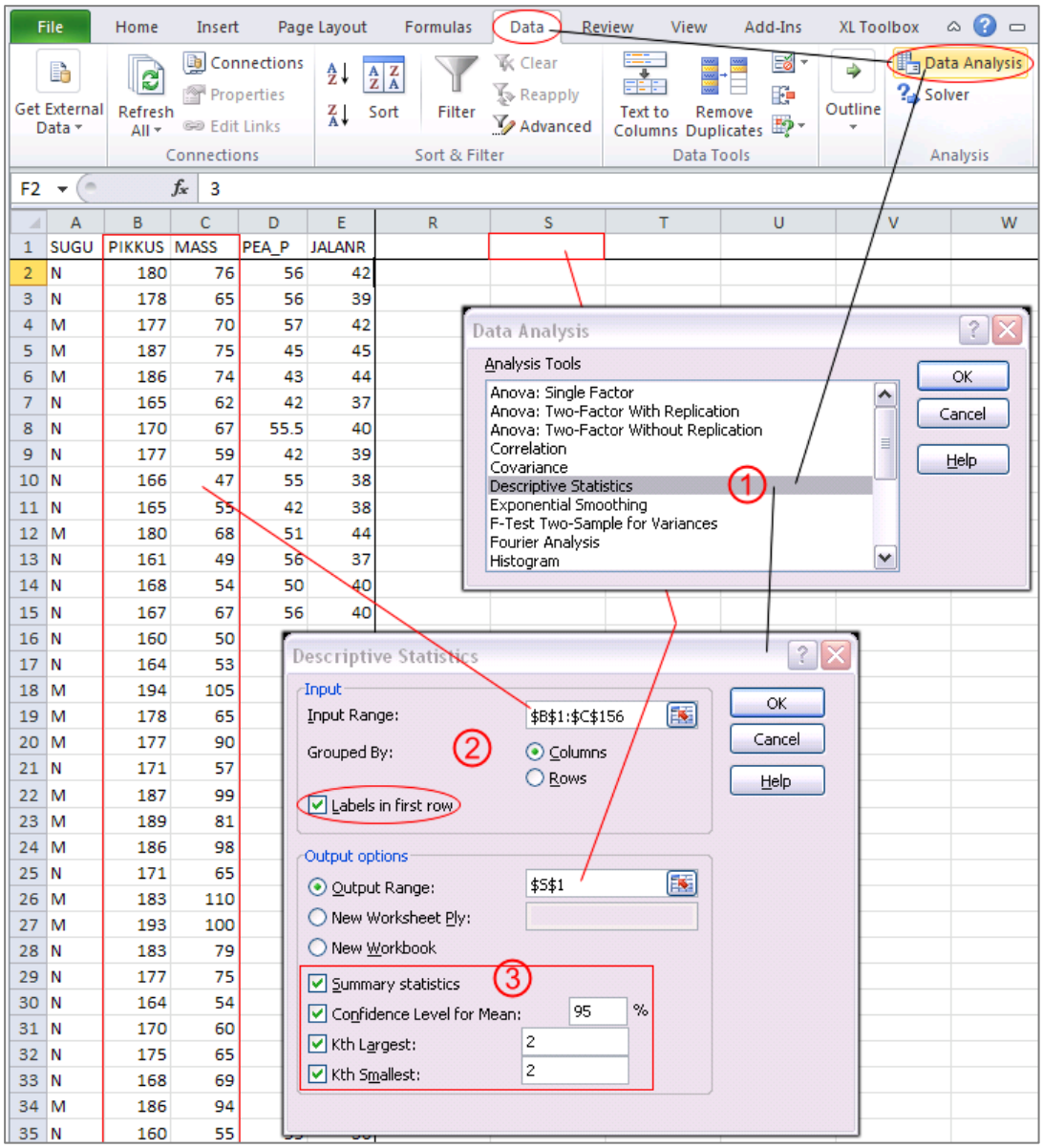
- *Grouped By* – määratakse andmete paigutus blokis, tavaliselt on erinevad tunnused paigutatud erinevatesse veergudesse (*Columns*), kuid võivad olla ka erinevates ridades (*Rows*);
- *Labels In First Column* – märgitakse tunnuse nime või tähise olemasolu korral andmebloki ülemises reas;
- *Output options* – määratakse tulemuste väljastamise asukoht: samale töölehele (*Output Range*), uuele töölehele (*New Worksheet Ply*) või uude faili (*New Workbook*);

3. täpsustada lisavalikute abil soovitud tulemuste hulka:
   - valiku *Summary statistics* tulemusena arvutab Excel kolmeteistkümne põhilise arvkarakteristiku väärtused;
   - valiku *Confidence Level for Mean* tulemusena arvutatakse pool usaldusintervalli suurusest (so suurus, mis tuleb keskmisele juurde liita ja lahutada, saamaks vastavalt ülemist ja alumist usalduspiiri); vaikimisi kasutatava 95% asemele võib ise trükkida mõne teise arvu (näiteks 90 või 99);
   - valikute *Kth Largest* ja *Kth Smallest* tulemusena väljastatakse järjekorranumbriga K väärtus vastavalt suuremate ja väiksemate väärtuste poolt lugedes; K = 1 korral on

tulemuseks maksimaalne ja minimaalne väärtus, et aga need suurused sisalduvad ka valiku *Summary statistics* väljundis, on enamasti mõistlik tellida näiteks suuruselt järgmised väärtused (siis K = 2).

NB! Kui teistel statistikaprotseduuridel on oma vaikimisi väljund ja lisavalikud ei ole kohustuslikud – need võimaldavad väljundit lihtsalt erinevate tabelite ja/või joonistega täiendada –, siis protseduur *Descriptive Statistics* nõuab vähemalt ühe lisavaliku määramist – vastasel juhul lõpeb arvutusprotsess veateatega.

Joonisel 22 esitatud protseduuri *Descriptive Statistics* tulemus koos leitud arvkarakteristikute eestikeelsete nimedega.



Joonis 21. Protseduuri *Descriptive Statistics* rakendamine.

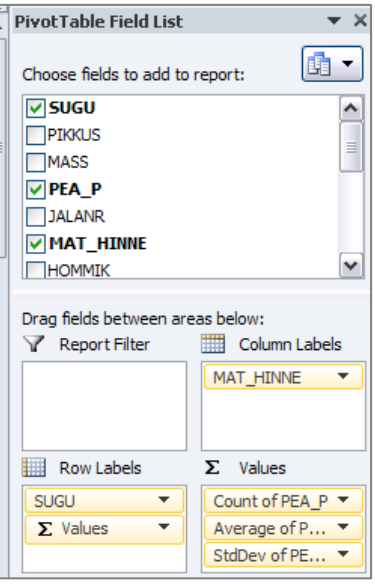
PIKKUS		*MASS*			
Mean	173.383871	Mean	68.803871	Keskmine	
Standard Error	0.75232943	Standard Erro	1.18499376	Standardviga	
Median	172	Median	65	Mediaan	
Mode	180	Mode	55	Mood	
Standard Deviation	9.36642584	Standard Devi	14.7530533	Standardhälve	
Sample Variance	87.729933	Sample Variar	217.652582	Dispersioon	
Kurtosis	-0.55086102	Kurtosis	0.74153233	Ekstsess e järsakuskordaja	
Skewness	0.24627128	Skewness	0.98159998	Asümmeetriakordaja	
Range	47	Range	73.5	Ulatus = Max - Min	
Minimum	151	Minimum	46.5	Miinimum	
Maximum	198	Maximum	120	Maksimum	
Sum	26874.5	Sum	10664.6	Summa	
Count	155	Count	155	Vaatluste arv	
Largest(2)	194	Largest(2)	110	Suuruselt teine väärtus	
Smallest(2)	155	Smallest(2)	47	Väiksuselt teine väärtus	
Confidence Level(95.0%)	1.4862178	Confidence Le	2.34094102	Liidetav usalduspiiride leidmiseks	

Joonis 22. Protseduuri *Descriptive Statistics* tulemus koos arvkarakteristikute eestikeelsete nimedega.

3.3. Risttabel (*PivotTable*)

PivotTable't on mugav kasutada leidmaks arvkarakteristikute väärtusi gruppides (näiteks keskmisi peaümbermõõte sõltuvalt tudengi soost ja matemaatika hindest – Joonis 23).

	A	B	C	D	E
3		Column Labels			
4	Row Labels	3	4	5	Grand Total
5	M				
6	Count of PEA_P	30	16	4	50
7	Average of PEA_P2	54.85	56.6875	56.25	55.55
8	StdDev of PEA_P3	4.47	2.02	1.26	3.74
9	N				
10	Count of PEA_P	22	55	28	105
11	Average of PEA_P2	53.45	54.68	55.18	54.56
12	StdDev of PEA_P3	4.98	3.93	2.41	3.86
13	**Total Count of PEA_P**	**52**	**71**	**32**	**155**
14	**Total Average of PEA_P2**	**54.26**	**55.13**	**55.31**	**54.88**
15	**Total StdDev of PEA_P3**	**4.70**	**3.68**	**2.31**	**3.84**



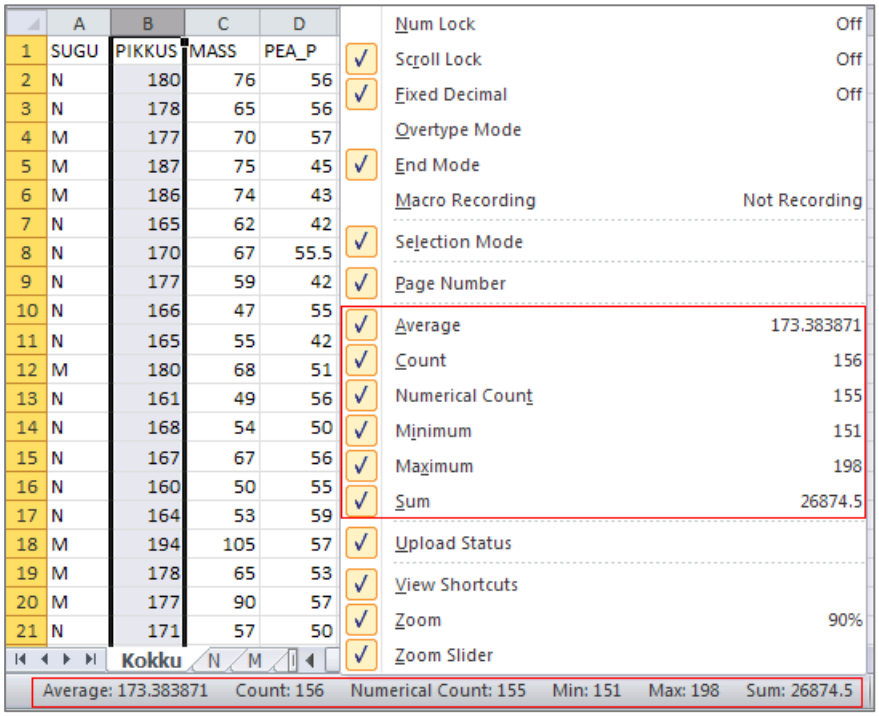
Joonis 23. Tudengite arv, keskmine peaümbermõõt ja peaümbermõõdu standardhälve sõltuvalt soost ja matemaatika hindest.

3.4. Muud võimalused

Saamaks kiirelt teada mõne arvkarakteristiku väärtust, ilma seda kuhugi töölehe lahtrisse arvutamata, võib kasutada töölehe allservas (olekuribal) kuvatavaid **selekteeritud lahtrite** sisu kirjeldavaid väärtusi (Joonis 24).

Arve sisaldavate lahtrite kohta kuvatakse keskmine, minimaalne ja maksimaalne väärtus ning väärtuste summa, lisaks ka veel arvuliste väärtustega lahtrite arv ning kõigi selekteeritud mittetühjade lahtrite arv.

Kui mõnda nimetatud karakteristikutest vaikimisi ei kuvata, saab selle tellida, klikkides hiire parempoolse nupuga olekuribal ja teostades vastava valiku avanenud rippmenüüs (Joonis 24).



Joonis 24. Olekuribal kuvatavad selekteeritud lahtrite sisu kirjeldavad väärtused.

Sarnaselt *PivotTable*'ga, ainult mitte eraldi tabelina, vaid analüüsitava andmetabeli sisse, on mitmesugused vahekokkuvõtted teostatavad ka *Data*-sakilt leitava valiku *Subtotal* abil.

4. Usalduspiirid

4.1. Usalduspiirid keskmisele

Keskmise usalduspiiride leidmine funktsioonide abil

Üldine valem mingi parameetri hinnangu usalduspiiride leidmiseks on kujul:

parameetri hinnang ± (tabeli väärtus * parameetri hinnangu standardviga).

Nn tabeli väärtus kujutab enesest mingi teoreetilise jaotuse protsendipunkti ning see sõltub nii ette antud protsendist kui ka hinnatava parameetri teoreetilisest jaotusest.

Mõnikord on viimase valikuks mitu võimalust. Näiteks juhul, kui uuritava tunnuse varieeruvus (dispersioon) on teada või on tegu suure valimiga, on keskmise usaldusintervall leitav standardse normaaljaotuse alusel valemist

$$\overline{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}.$$

Kui aga varieeruvust teada pole ja valim on väike, tuleb kasutada veidi „raskemate sabadega“ t-jaotust (sest tuleb arvestada ka dispersiooni hindamisel tekkinud võimaliku veaga, mis omakorda muudab keskmise hinnangu ebatäpsemaks ja seda eriti väikese valimi korral):

$$\overline{x} \pm t_{\alpha/2}(n-1) * \frac{s}{\sqrt{n}}.$$

Excelis ongi keskmise usalduspiiride arvutamiseks kaks eraldi funktsiooni: CONFIDENCE.NORM ja CONFIDENCE.T.

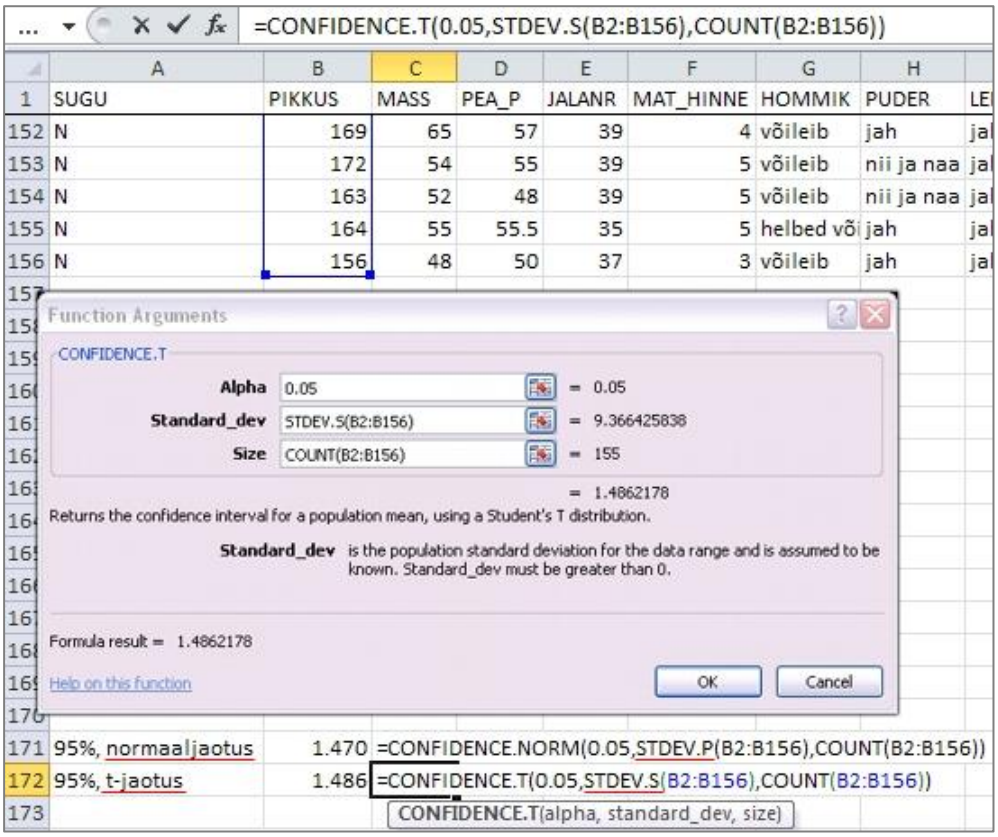
Mõlemad funktsioonid tahavad argumentidena ette (Joonis 25)

- olulisuse nivood *Alpha* (95%-lise usaldusintervalli korral on olulisuse nivoo 0,05),
- uuritava tunnuse standardhälvet või selle hinnangut *Standard_dev* (vastavalt funktsiooni CONFIDENCE.NORM või CONFIDENCE.T puhul, leituna funktsioonidega STDEV.P või STDEV.S),
- vaatluste arvu *Size*.

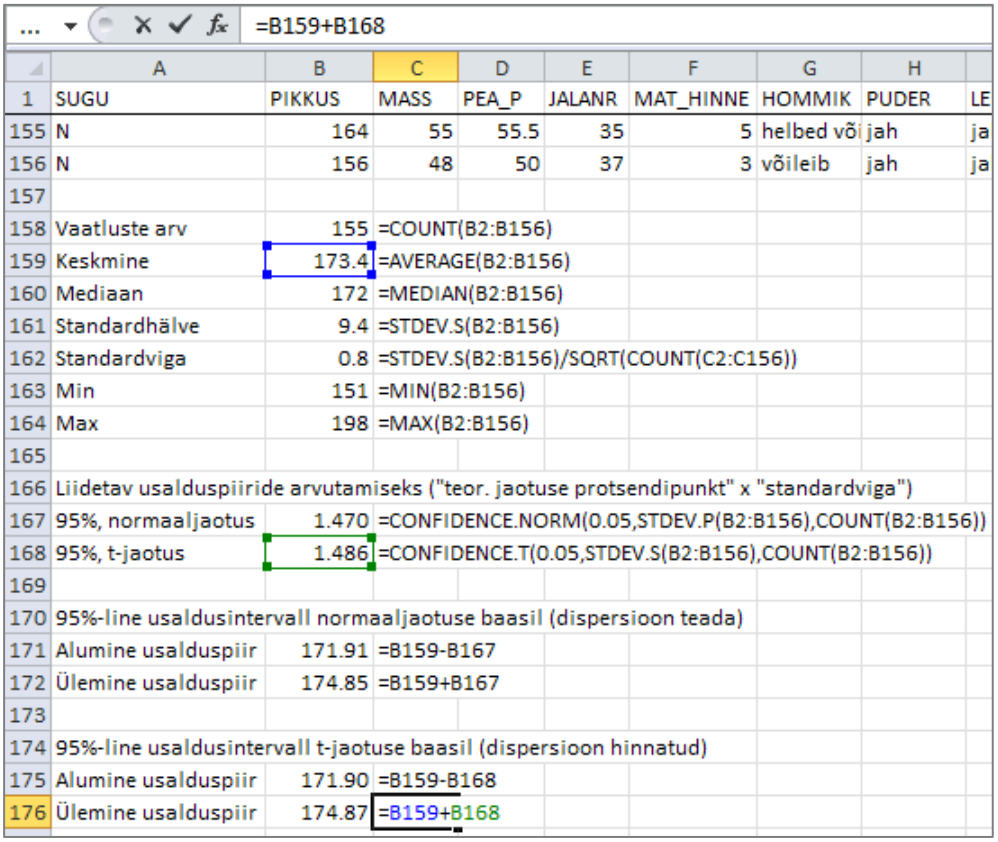
NB! Funktsioonide CONFIDENCE.NORM või CONFIDENCE.T tulemusena saadud arv näitab usalduspiiride kaugust keskväärtusest (poolt usaldusintervalli laiusest), usalduspiiride eneste leidmiseks tuleb see siis kas liita või lahutada aritmeetilisest keskmisest (Joonis 26).

Joonisel 26 toodud tulemustest nähtub, et tudengite keskmine pikkus jääb 95%-lise tõenäosusega vahemikku 171,9-174,9 cm. Seejuures on normaaljaotuse baasil hinnatud usaldusintervall vaid õige pisut kitsam, sest valim on piisavalt suur (n=155) garanteerimaks ka hinnangute täpsust.

Märkus. Ekslikult väidab Excel ka funktsiooni CONFIDENCE.T tellimisaknas ja abifailis, et funktsiooni argumendina ette antav standardhälve on populatsiooni teadaolev standardhälve. Tegelikult on see väide õige vaid funktsiooni CONFIDENCE.NORM puhul, funktsiooni CONFIDENCE.T rakendamisel eeldatakse ikka, et populatsiooni standardhälve ei ole teada ja on seetõttu andmetest hinnatud (funktsiooniga STDEV.S).



Joonis 25. Usalduspiiride arvutamine tudengite keskmisele pikkusele funktsioonidega CONFIDENCE.NORM ja CONFIDENCE.T.



Joonis 26. Usalduspiiride arvutamine tudengite keskmisele pikkusele funktsioonide CONFIDENCE.NORM ja CONFIDENCE.T tulemuste ning aritmeetilise keskmise alusel.

Keskmise usalduspiiride leidmine protseduuriga *Descriptive Statistics*

Kui uuritava tunnuse dispersioon ei ole teada (ja nii see tavaliselt on), on keskväärtuse usalduspiiride leidmiseks lisaks funktsioonile CONFIDENCE.T kasutatav ka protseduuri *Descriptive Statistics* valik *Confidence Level for Mean*.

Tellimusakna täitmine kulgeb analoogselt arvkarakteristikute leidmisel kirjeldatuga (Joonis 21), lisaks võib muuta usaldusnivood (vaikimisi on selleks 95%).

Tulemusena väljastatakse arvkarakteristikute tabelis suurus, mis näitab uuritava tunnuse keskmise väärtuse kaugust oma alumisest ja ülemisest usalduspiirist (poolt usaldusintervalli laiust). Usalduspiirid leitakse, liites ja lahutades saadud arvu tunnuse aritmeetilisele keskmisele (Joonis 27).

	S	T	U	V	W	X	Y	Z
1	*PIKKUS*		*MASS*					
2								
3	Mean	173.383871	Mean	68.803871		Keskmine		
4	Standard Error	0.75232943	Standard Erro	1.18499376		Standardviga		
5	Median	172	Median	65		Mediaan		
6	Mode	180	Mode	55		Mood		
7	Standard Deviation	9.36642584	Standard Devi	14.7530533		Standardhälve		
8	Sample Variance	87.729933	Sample Variar	217.652582		Dispersioon		
9	Kurtosis	-0.55086102	Kurtosis	0.74153233		Ekstsess e järsakuskordaja		
10	Skewness	0.24627128	Skewness	0.98159998		Asümmeetriakordaja		
11	Range	47	Range	73.5		Ulatus = Max - Min		
12	Minimum	151	Minimum	46.5		Miinimum		
13	Maximum	198	Maximum	120		Maksimum		
14	Sum	26874.5	Sum	10664.6		Summa		
15	Count	155	Count	155		Vaatluste arv		
16	Largest(2)	194	Largest(2)	110		Suuruselt teine väärtus		
17	Smallest(2)	155	Smallest(2)	47		Väiksuselt teine väärtus		
18	Confidence Level(95.0%)	1.4862178	Confidence Le	2.34094102		Liidetav usalduspiiride leidmiseks		
19								
20	Alumine 95% usalduspiir	171.897653	=T3-T18	66.4629299				
21	Ülemine 95% usalduspiir	174.870089	=T3+T18	71.144812				

Joonis 27. Usalduspiiride arvutamine tudengite keskmisele pikkusele ja kehamassile protseduuri *Descriptive Statistics* poolt väljastatud tulemuste alusel.

4.2. Usalduspiirid teistele parameetritele

Et Excelis on olemas funktsioonid mitmete erinevate teoreetliste jaotuste kvantiilide e protsendipunktide leidmiseks, on usalduspiirid leitavad ka otse vastavaid valemeid Excelis rakendades.

Usalduspiirid keskmisele

Eelmises punktis kirjeldatud usalduspiirid keskmisele võinuks leida ka otse lähtuvalt valemeist

$$\bar{x} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n}} \text{ või } \bar{x} \pm t_{1-\alpha/2}(n-1) * \frac{s}{\sqrt{n}}.$$

Suurused $z_{1-\alpha/2}$ ja $t_{1-\alpha/2}(n-1)$ neis valemeis on vastavalt standardse normaaljaotuse ja vaatluste arvule n vastava t-jaotuse 1-α/2-kvantiilid (väärtused, millest suuremaid väärtuseid

saab antud jaotuse korral olla vaid α/2*100%). 95%-lise usaldusintervalli korral on olulisuse nivoo α = 0,05 ja arvutusteks tuleb leida kas standardse normaaljaotuse või t-jaotuse 97,5%-punkt ehk (1 – 0,05/2) = 0,975-kvantiil. Leitavad on need kvantiilid vastavalt funktsiooniga NORM.S.INV ja funktsiooniga T.INV.

Tudengite keskmise pikkuse 95%-liste usalduspiiride arvutamine, lähtudes otse usalduspiiride valemeist, on esitatud joonisel 28.

Tulemustest nähtub, et tudengite keskmine pikkus jääb 95%-lise tõenäosusega vahemikku 171,9-174,9 cm. Seejuures on normaaljaotuse baasil hinnatud usaldusintervall vaid õige pisut kitsam, sest valimi on piisavalt suur (*n*=155) garanteerimaks ka ligikaudsete hinnangute täpsust.

	A	B	C	D	E	F	G	H	I	J	K	L
1	SUGU	PIKKUS	MASS	PEA_P	JALANR	MAT_HINNE	HOMMIK	PUDER	LEMMIK	HAIGE	SPORT	AUTO
154	N	163	52	48	39	5	võileib	nii ja naa	jah	jah	ei	ei
155	N	164	55	55.5	35	5	helbed või	jah	jah	jah	jah	ei
156	N	156	48	50	37	3	võileib	jah	jah	jah	jah	ei
157												
158	95%-line usaldusintervall normaaljaotuse baasil (dispersioon teada)											
159	Alumine usalduspiir	171.91	=AVERAGE(B2:B156) - NORM.S.INV(0.975) * STDEV.P(B2:B156)/SQRT(COUNT(B2:B156))									
160	Ülemine usalduspiir	174.85	=AVERAGE(B2:B156) + NORM.S.INV(0.975) * STDEV.P(B2:B156)/SQRT(COUNT(B2:B156))									
161				$\bar{x}$	±	$z_{0.975}$	*		$s/\sqrt{n}$			
162	95%-line usaldusintervall t-jaotuse baasil (dispersioon hinnatud)											
163	Alumine usalduspiir	171.90	=AVERAGE(B2:B156) - T.INV(0.975,COUNT(B2:B156)-1) * STDEV.S(B2:B156)/SQRT(COUNT(B2:B156))									
164	Ülemine usalduspiir	174.87	=AVERAGE(B2:B156) + T.INV(0.975,COUNT(B2:B156)-1) * STDEV.S(B2:B156)/SQRT(COUNT(B2:B156))									
165												
166				$\bar{x}$	±	$t_{0.975,155-1}$		*		$s/\sqrt{n}$		

Joonis 28. Usalduspiiride arvutamine tudengite keskmisele pikkusele Exceli funktsioonide abil lähtudes otse usalduspiiride valemeist.

Usalduspiirid protsendile

Suure valimi (enamasti tähendab see, et *n*>60) korral on protsendi usaldusintervalli arvutamiseks kasutatav juba keskväärtuse usaldusintervalli arvutamisest tuttav valem

$$\bar{x} \pm z_{1-\alpha/2} \frac{s}{\sqrt{n}},$$

kus aritmeetiline keskmine tähendab hinnangut uuritava sündmuse toimumise tõenäosusele, $\bar{x} = \hat{p}$, ja standardhälve avaldub vastavalt binoomjaotusele kujul $s = \sqrt{\hat{p}(1-\hat{p})}$.

Seega on uuritava sündmuse toimumise tõenäosuse usaldusintervall leitav valemist

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Usaldusintervalli protsendile saab, kui korrutada eelmine avaldis 100%-ga.

Joonisel 29 on näidatud 95%-lise usaldusintervalli arvutamist õlut joovate tudengite osakaalule. Tulemustest nähtub, et hinnanguliselt 36,1% tudengitest joob õlut, kusjuures 95%-lise tõenäosusega jääb õlut joovate tudengite protsent vahemikku 28,6-43,7%.

	M	N	O	P	
1	OLU	OLU_01	SUITS	TEATER	KINO
151	0	0	ei	rohkem kui aasta tagasi	viimase kuu j
152	0.5	1	ei	rohkem kui aasta tagasi	viimase kuu j
153	0	0	enam ei, a	rohkem kui aasta tagasi	viimase kuu j
154	0	0	ei	viimase aasta jooksul	viimase aasta
155	0	0	ei	viimase aasta jooksul	rohkem kui aa
156	0	0	ei	rohkem kui aasta tagasi	viimase kuu j
157					
158	Tudengite arv	155	=COUNT(N2:N156)		
159	Keskmine	0.36129	=AVERAGE(N2:N156)		
160	Standardhälve	0.480374	=SQRT(N159*(1-N159))		
161					
162	Alumine usalduspiir	0.285666	= N159 - NORM.S.INV(0.975) * N160/SQRT(N158)		
163	Ülemine usalduspiir	0.436915	= N159 + NORM.S.INV(0.975) * N160/SQRT(N158)		
164					
165			$\hat{p} \pm z_{0.975} * \sqrt{\hat{p}(1-\hat{p})/n}$		

Joonis 29. Usalduspiiride arvutamine õlut joovate tudengite osakaalule Exceli funktsioonide abil lähtudes otse usalduspiiride valemeist.

Usalduspiirid dispersioonile

Normaaljaotusega tunnuse dispersiooni alumine ja ülemine usalduspiir leitakse vastavalt valemeist

$$s^2 * \frac{n-1}{h_{1-\alpha/2}(n-1)} \text{ ja } s^2 * \frac{n-1}{h_{\alpha/2}(n-1)},$$

kus s^2 on valimi dispersioon, n on vaatluste arv ning $h_{1-\alpha/2}(n-1)$ ja $h_{\alpha/2}(n-1)$ on χ^2-jaotuse kvantiilid vabadusastmete arvu $n-1$ korral. Excelis on viimased leitavad funktsiooniga CHISQ.INV.

Standardhälbe alumine ja ülemine usalduspiir leitakse vastavalt valemeist

$$s * \sqrt{\frac{n-1}{h_{1-\alpha/2}(n-1)}} \text{ ja } s * \sqrt{\frac{n-1}{h_{\alpha/2}(n-1)}}.$$

Joonisel 30 on näitatud 95%-lise usaldusintervalli arvutamist tudengite pikkuse standard-hälbele. Tulemustest nähtub, et 95%-lise tõenäosusega jääb tudengite pikkuse standardhälve vahemikku 8,4-10,5 cm.

	A	B	C	D	E	F	G	
1	SUGU	PIKKUS	MASS	PEA_P	JALANR	MAT_HINNE	HOMMIK	P
2	N	180	76	56	42	3	muu	ja
3	N	178	65	56	39	4	ei söö tav	ja
4	M	177	70	57	42	3	võileib	n
188								
189	Vaatluste arv	155	=COUNT(B2:B156)					
190	Standardhälve	9.3664	=STDEV.S(B2:B156)					
191								
192	95%-line usaldusintervall standardhälbele							
193	Alumine usalduspiir	8.42696	= B190 * SQRT((B189-1)/CHISQ.INV(0.975,B189-1))					
194	Ülemine usalduspiir	10.5435	= B190 * SQRT((B189-1)/CHISQ.INV(0.025,B189-1))					
195								
196			$s * \sqrt{(n-1)/h_{0.025}(n-1)}$					

Joonis 30. Usalduspiiride arvutamine tudengite pikkuse standardhälbele.

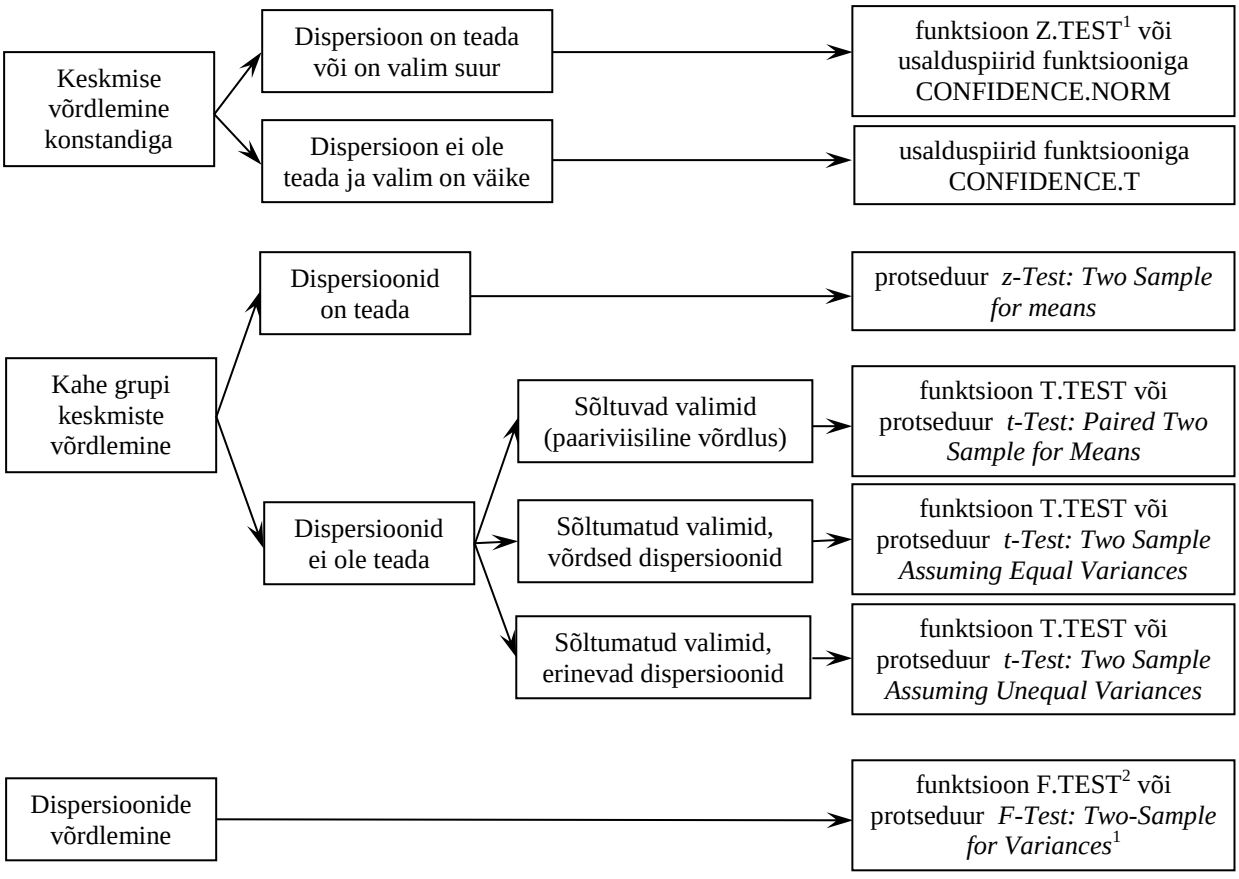
5. Hüpoteeside kontrollimine (ühe ja kahe üldkogumi võrdlus)

5.1. Üldskeem

Sõltuvalt andmete olemusest ja kontrollitava hüpoteesi tüübist on MS Excelis mitmeid erinevaid võimalusi ühe ja kahe üldkogumi võrdlemiseks. Skeem joonisel 31 annab lühiülevaate, millist funktsiooni või protseduuri millal kasutada.

NB! Kahe üldkogumi keskmiste võrdlemisel t-testiga sõltumatute valimite korral tuleb esmalt võrrelda F-testiga dispersioone, mille järel saab alles otsustada, kumba t-testi arvutuseeskirja rakendada – kas seda, mis eeldab võrdset varieeruvust võrreldavates üldkogumites (F-testi tulemusena saadud p-väärtus on suurem kui 0,05), või seda, mis arvestab võrreldavate üldkogumite erineva varieeruvusega (F-testi $p < 0,05$).

Lisaks tuleb arvestada, et kõik Excelis leiduvad ühe ja kahe üldkogumi võrdlemise vahendid eeldavad, et uuritav(ad) tunnus(ed) on normaaljaotusega või on andmeid palju. Mitteparameetriliste, normaaljaotust mitte-eeldavate testide teostamiseks Excelis sisseehitatud vahendid puuduvad. Siiski on mõnede mitteparameetriliste testide läbiviimine Excelis võimalik – kas siis testide aluseks olevate arvutuste samm-sammulise teostamise või spetsiaalsete lisamoodulite abil (vt peatükk 5.6).



Joonis 31. Ühe ja kahe grupi võrdluse üldskeem MS Excelis; [1] funktsioon Z.TEST ja protseduur *F-Test: Two-Sample for Variances* teostavad üksnes ühepoolse testi, [2] funktsioon F.TEST aga vaid kahepoolse testi.

5.2. Hüpoteeside kontrollimine usalduspiiridega

Juhul, kui kontrollitavaks hüpoteesiks on mingi andmete alusel hinnatud suuruse erinevus konstandist, tehakse otsus sageli 95%-lise usaldusintervalli alusel:

- kui konstant, millega andmeist arvutatud suurust võrreldakse, jääb usalduspiiride vahele, siis ei ole alust väita, et arvutatud suurus erineb antud konstandist;
- kui aga onstant jääb usaldusintervallist väljapoole, on arvutatud suurus konstandist statistiliselt oluliselt erinev ($p < 0,05$).

Näiteks soovides testida, kas esimese kursuse neidude keskmine pikkus erineb Eesti naiste keskmisest pikkusest 168 cm, piisab, kui võrrelda konstanti 168 neidude keskmise pikkuse 95%-lise usaldusintervalliga.

Kui oletada, et on täpselt teada, kui varieeruvad on esimese kursuse neidude pikkused – pikkuse standardhälve on 6,5 cm –, on keskmise pikkuse 95%-line usaldusintervall leitav funktsiooniga CONFIDENCE.NORM.

Kui pikkuse dispersiooni täpselt teada pole – ja enamasti see on nii –, tuleb dispersioon olemasolevatest andmetest hinnata ja keskmise pikkuse 95%-line usaldusintervall on leitav funktsiooniga CONFIDENCE.T (Joonis 32).

	A	B	C	D	E	F	G	H	I	J
1	SUGU	PIKKUS	MASS	PEA_P	JALANR	MAT_HIN	HOMMIK	PUDER	LEMMIK	HAIG
104	N	163	52	48	39	5	võileib	nii ja naa	jah	jah
105	N	164	55	55.5	35	5	helbed võ	jah	jah	jah
106	N	156	48	50	37	3	võileib	jah	jah	jah
107										
108										
109	**95%-line usaldusintervall normaaljaotuse baasil (standardhälve = 6,5 cm)**									
110	Alumine usalduspiir	167.542	=AVERAGE(B2:B106) - CONFIDENCE.NORM(0.05,6.5,COUNT(B2:B106))							
111	Ülemine usalduspiir	170.029	=AVERAGE(B2:B106) + CONFIDENCE.NORM(0.05,6.5,COUNT(B2:B106))							
112										
113	**95%-line usaldusintervall t-jaotuse baasil (standardhälve hinnatud funktsiooniga STDEV.S)**									
114	Alumine usalduspiir	167.54	=AVERAGE(B2:B106) - CONFIDENCE.T(0.05,STDEV.S(B2:B106),COUNT(B2:B106))							
115	Ülemine usalduspiir	170.04	=AVERAGE(B2:B106) - CONFIDENCE.T(0.05,STDEV.S(B2:B106),COUNT(B2:B106))							

Joonis 32. Neidude keskmise pikkuse usaldusintervalli hindamine funktsioonidega CONFIDENCE.NORM ja CONFIDENCE.T.

Et Eesti naiste keskmine pikkus 168 cm jääb esimese kursuse neidude keskmise pikkuse 95% usaldusintervalli sisse: 167,5<168<170,0, siis ei ole alust lugeda tõestatuks alternatiivset hüpoteesi keskmise pikkuse erinevusest 168 sentimeetrist ja tuleb jääda nullhüpoteesi juurde: esimese kursuse neidude keskmine pikkus ei erine 168 sentimeetrist.

Hea asi hüpoteeside kontrollimisel usalduspiiridega on see, et kui nüüd soovida testida, kas esimese kursuse neidude keskmine pikkus erineb maailma naiste keskmisest pikkusest 154 cm, ei pea midagi uuesti arvutama, piisab, kui võrrelda arvu 154 juba leitud usaldus-intervalliga – kuna 154 ei jää usalduspiiride vahele, võib lugeda tõestatuks, et esimese kursuse neidude keskmine pikkus erineb maailma naiste keskmisest pikkusest 154 cm (seejuures $p < 0,05$).

5.3. z-test

Keskväärtuse võrdlemine konstandiga

MS Exceli funktsioon Z.TEST testib normaaljaotuse ning teadaoleva dispersiooni (või suure valimi) eeldusel **ühepoolset** hüpoteesi kujul

$$H_0: \mu \leq \text{konstant}$$
$$H_1: \mu > \text{konstant}$$

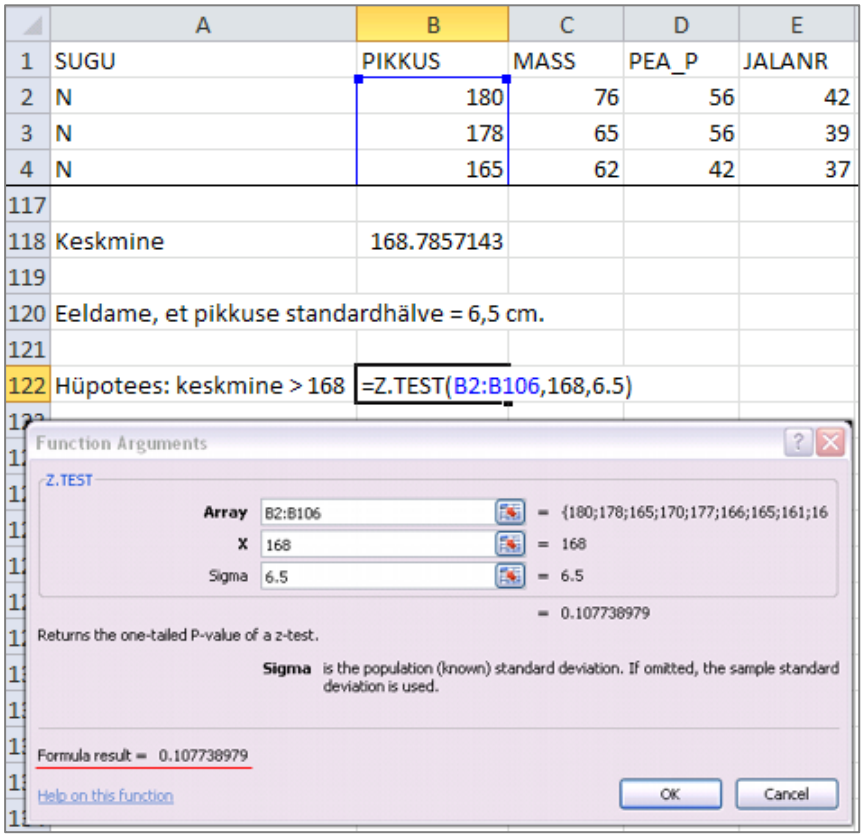
(μ on uuritava tunnuse keskväärtus).

Funktsioonile ZTEST tuleb ette anda (Joonis 32)

- *Array* – algandmete blokk (ilma tunnuse nimeta),
- *X* – konstant, millega võrdumist kontrollitakse,
- *Sigma* – populatsiooni teadaolev **standardhälve** (**NB!** võib ka puududa, siis arvutab Excel ise valimi standardhälbe valemiga STDEV.S ja kasutab seda).

Tulemusena väljastab Excel eelnevalt kursoriga määratud lahtrisse olulisuse tõenäosuse *p* väärtuse. Kui leitud $p < 0{,}05$, võib lugeda tõestatuks alternatiivse hüpoteesi H_1: uuritava tunnuse keskväärtus on võrreldavast konstandist suurem ja seda olulisuse nivool 0,05.

Näiteks kui eeldada, et esimese kursuse neidude pikkuse standardhälve on teadaolevalt 6,5 cm, siis testides hüpoteesi: esimese kursuse neidude keskmine pikkus on suurem, kui Eesti naiste keskmine pikkus 168 cm, on tulemuseks p-väärtus 0,11 (Joonis 33). Seega saab järeldada, et esimese kursuse neidude keskmine pikkus ei ole statistiliselt oluliselt suurem, kui Eesti naiste keskmine pikkus 168 cm ($p = 0{,}11$).



Joonis 33. Neidude keskmise pikkuse võrdlemine 168 sentimeetriga teadaoleva pikkuse standardhälbe 6,5 cm korral funktsiooniga Z.TEST.

Kuna funktsioon Z.TEST testib alati ühepoolset hüpoteesi: keskväärtus on konstandist suurem, on juhul, kui tegelik andmetest leitud keskmine on võrreldavast konstandist väiksem, tulemuseks kindlasti 0,05-st suurem p-väärtus.

Näiteks soovides võrrelda esimese kursuse neidude keskmist pikkust (mis antud andmete alusel on 168,8 cm) 170-ga, testib Z.TEST tegelikult seda, kas 168,8 > 170. Tulemuseks on p-väärtus 0,97 (Joonis 34) – esimese kursuse neidude keskmine pikkus ei ole statistiliselt oluliselt suurem 170-st. Mis on ka loomulik, sest tegelikult on neidude pikkus hoopis väiksem kui 170: 168,8 < 170.

Soovides testida ühepoolset hüpoteesi teistpidi: neidude keskmine pikkus on väiksem kui 170, piisab standardse normaaljaotuse sümmeetrilisuse tõttu nulli suhtes lihtsalt funktsiooni Z.TEST poolt väljastatud p-väärtuse ühest lahutamisest. Tulemuseks on $p = 0{,}028$ – neidude keskmine pikkus on statistiliselt oluliselt väiksem 170-st (Joonis 34).

Soovides testida kahepoolset hüpoteesi, piisab lihtsalt nö õiges suunas testitud ühepoolse hüpoteesi p-väärtuse kahega korrutamisest (Joonis 34). Ehk siis testides hüpoteesi: neidude keskmine pikkus erineb 170-st, on tulemuseks $p = 0{,}056$ – neidude keskmine pikkus ei ole statistiliselt oluliselt erinev 170-st.

Z... = 2 * MIN(Z.TEST(B2:B106,170,6.5), 1 - Z.TEST(B2:B106,170,6.5))

	A	B	C	D	E	F	G	H
1	SUGU	PIKKUS	MASS	PEA_P	JALANR	MAT_HIN	HOMMIK	PUDER
2	N	180	76	56	42	3	muu	jah
3	N	178	65	56	39	4	ei söö tav	jah
4	N	165	62	42	37	3	puder	jah
117								
118	Keskmine	168.7857143						
119								
120	Eeldame, et pikkuse standardhälve = 6,5 cm.							
121								
122	Hüpotees: keskmine > 168	0.107738979	= Z.TEST(B2:B106,168,6.5)					
123								
124	Hüpotees: keskmine > 170	0.972206882	= Z.TEST(B2:B106,170,6.5)					
125								
126	Hüpotees: keskmine < 170	0.027793118	= 1 - Z.TEST(B2:B106,170,6.5)					
127								
128	Hüpotees: keskmine ≠ 170	0.055586237	= 2 * MIN(Z.TEST(B2:B106,170,6.5), 1 - Z.TEST(B2:B106,170,6.5))					

Joonis 34. Ühe- ja kahepoolsete hüpoteeside testimine funktsiooniga Z.TEST.

Kahe üldkogumi keskväärtuste võrdlemine teadaolevate dispersioonide korral

Kahe üldkogumi keskväärtuse võrdlemine teadaolevate dispersioonide korral on teostatav protseduuriga *z-Test: Two Sample for means* (*Data*-sakk -> *Data Analysis*).

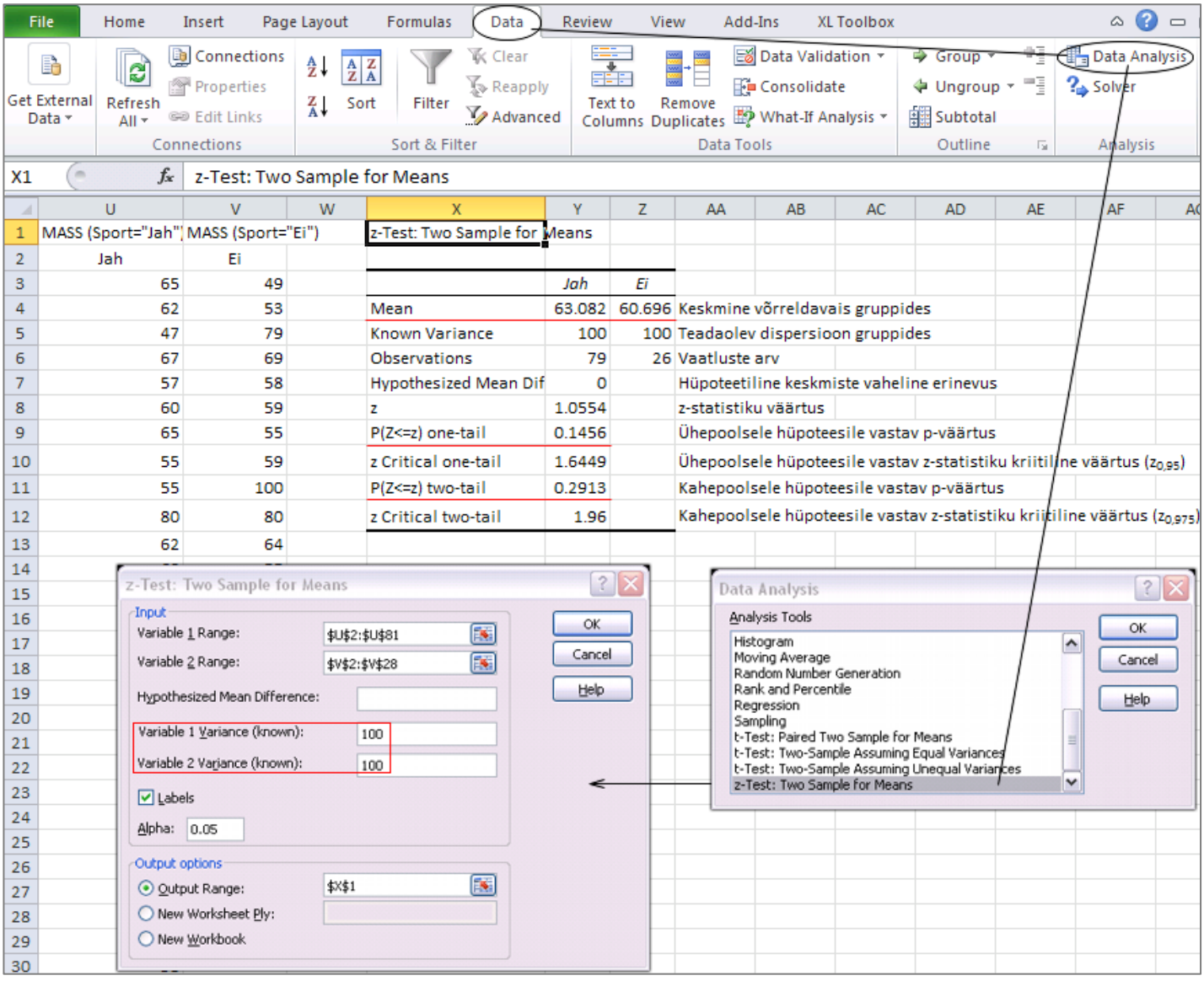
Protseduuril tuleb ette anda (Joonis 35)

- mõlema valimi andmete blokid – *Variable 1 Range* ja *Variable 2 Range* (seejuures võivad andmed paikneda nii veerus kui ka reas),
- oletatav keskväärtuste erinevus (vaikimisi null) – *Hypothesized Mean Difference*,
- mõlema populatsiooni teadaolevad **dispersioonid** – *Variable 1 Variance (known)* ja *Variable 2 Variance (known)*,

- kui andmete blokid sisaldavad esimeses reas/veerus nime, tuleb teha "linnuke" märgendi *Labels* ette,
- olulisuse nivoo (vaikimisi 0,05) – *Alpha*,
- tulemuste väljastamise asukoht (*Output options*): samale töölehele (*Output Range*), uuele töölehele (*New Worksheet Ply*) või uude faili (*New Workbook*).

NB! erinevalt funktsioonist Z.TEST, mille argumendiks oli populatsiooni standardhälve, tahab protseduur *z-Test: Two Sample for means* saada argumentidena ette populatsiooni dispersioone.

Joonisel 35 on näidatud spordiga tegelevate ja mittetegelevate esimese kursuse neidude keskmiste kehamasside võrdlemist protseduuriga *z-Test: Two Sample for means*. Kehamasside dispersiooniks sai mõlemas grupis võetud 100 kg^2 (oletame, et see on teada). Analüüsi tulemustest on näha, et spordiga tegelevate neidude kehamass on tegelikult hoopis kõrgem, kui spordiga mittetegelevatel neidudel – vastavalt 63,1 ja 60,7 kg – ju siis peab sportimiseks olema põhjus :) Siiski ei ole see erinevus statistiliselt oluline (p = 0,29).



Joonis 35. Sportivate ja mittesportivate neidude keskmiste kehamasside võrdlemine protseduuriga *z-Test: Two Sample for means*.

Ühepoolse hüpoteesi testimisel võtab protseduur *z-Test: Two Sample for means* alati aluseks tegelikud keskmised ning testib, kas suurem keskmine on ka statistiliselt oluliselt suurem kui väiksem keskmine. Antud näite puhul testitakse siis, kas spordiga tegelevate neidude kehamass on kõrgem, kui spordiga mittetegelevate neidude kehamass – vastus on, et on küll, ainult et mitte statistiliselt olulisel määral ($p = 0{,}15$).

Märkus. Et tunnuste varieeruvust üldkogumis tavaliselt ei teata, siis leiab vaadeldud protseduur ka vähest praktilist rakendust.

5.4. t-test

Kahe üldkogumi keskväärtuste võrdlemine sõltuvate valimite korral (paariviisiline võrdlus)

Sõltuvate vaatlustega/mõõtmistega on tegu, kui mõõdetud on samu või kõigi katsetulemust potentsiaalselt mõjutada võivate kriteeriumite poolest sarnaseid indiviide/objekte enne ja pärast teatavat „katset" (enne ja pärast ravimi manustamist, hommikul ja õhtul jne). Taolisel juhul moodustuvad „enne ja pärast sooritatud" mõõtmistest paarid – igal indiviidil/objektil on üks mõõtmine ühes ja teine mõõtmine teises grupis („enne ja pärast"). Gruppide keskmiste omavaheline võrdlemine on siis samaväärne keskmise muutuse nulliga võrdlemisega.

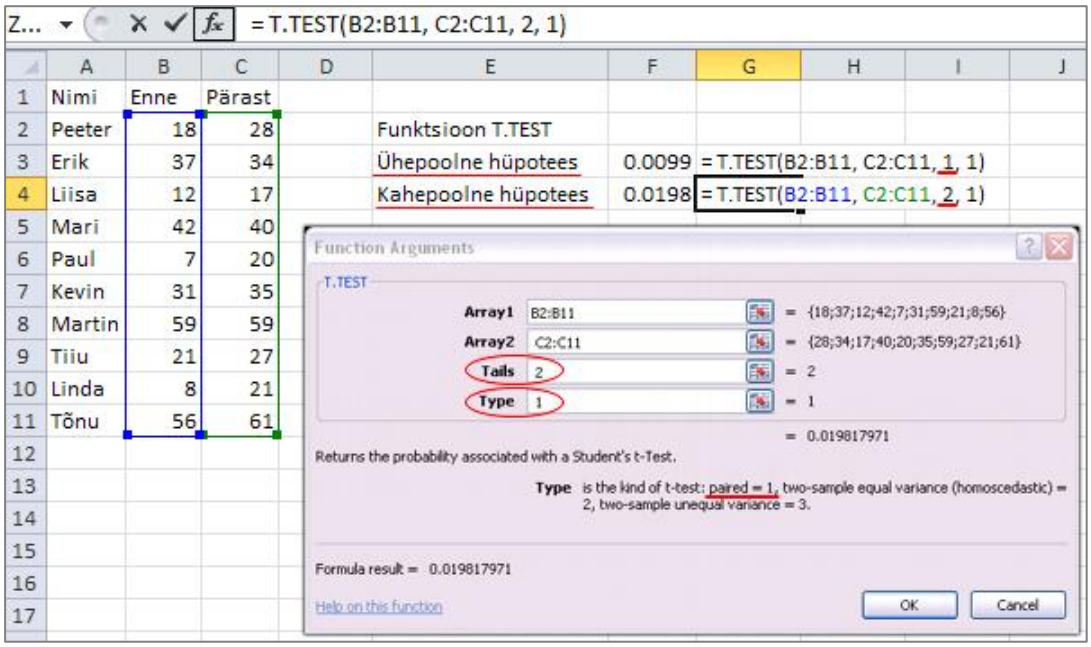
Excelis on kahe sõltuva (paaris) valimi keskmiste võrdlemiseks kasutatav funktsioon T.TEST ja protseduur *t-Test: Paired Two Sample for Means*.

Funktsiooni T.TEST, mis annab tulemuseks vaid olulisuse tõenäosuse *p* väärtuse, rakendamiseks tuleb panna kursor lahtrisse, kuhu tulemust soovite, valida Exceli funktsioonide hulgast või sisestada klaviatuurilt funktsioon T.TEST ja anda ette (vt ka Joonis 36)

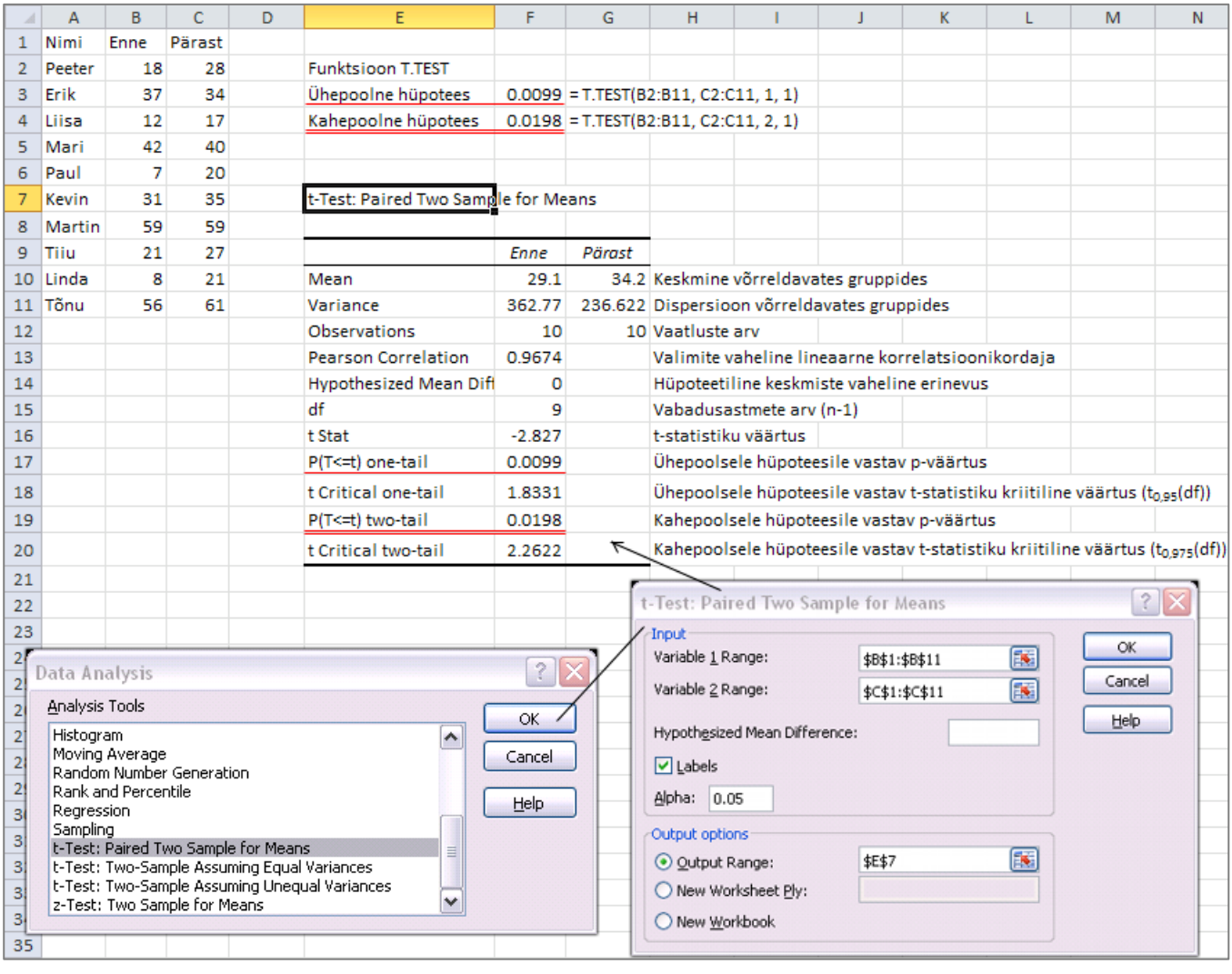
- mõlema valimi andmete blokid (*Array1* ja *Array2*),
- hüpoteesi tüüp (*Tails*): 1 – ühepoolne hüpotees (*one-tailed distribution*), 2 – kahepoolne hüpotees (*one-tailed distribution*),
- testi tüüp lähtuvalt andmete struktuurist ja varieeruvusest (*Type*): antud juhul 1 – sõltuvad valimid (*paired*); ülejäänud kaks tüüpi on: 2 – sõltumatud valimid ja võrdsed dispersioonid (*two-sample equal variance (hoscedastic)*) ning 3 – sõltumatud valimid ja erinevad dispersioonid (*two-sample unequal variance*).

Protseduur *t-Test: Paired Two Sample for Means* (*Data*-sakk -> *Data Analysis*) annab tulemuseks nii võrreldavaid gruppe kirjeldavad karakteristikud kui ka t-testi teostamisega kaasnevad arvutustulemused nii ühe- kui ka kahepoolse hüpoteesi kontrollimiseks ning selle rakendamiseks tuleb ette anda (vt Joonis 37)

- mõlema valimi andmete blokid – *Variable 1 Range* ja *Variable 2 Range* (seejuures võivad andmed paikneda nii veerus kui ka reas),
- oletatav keskväärtuste erinevus (vaikimisi null) – *Hypothesized Mean Difference*,
- kui andmete blokid sisaldavad esimeses reas/veerus nime, tuleb teha "linnuke" märgendi *Labels* ette,
- olulisuse nivoo (vaikimisi 0,05) – *Alpha*,
- tulemuste väljastamise asukoht (*Output options*): samale töölehele (*Output Range*), uuele töölehele (*New Worksheet Ply*) või uude faili (*New Workbook*).



Joonis 36. Funktsiooni T.TEST rakendamine tudengite enne ja pärast kursuse läbimist teostatud testi tulemuste võrdlemiseks sõltuvate valimite eeldusel.



Joonis 37. Protseduuri *t-Test: Paired Two Sample for Means* rakendamine tudengite enne ja pärast kursuse läbimist teostatud testi tulemuste võrdlemiseks sõltuvate valimite eeldusel.

Tulemustest (Joonis 37) võib järeldada, et tudengite testi tulemused enne ja pärast kursuse läbimist on statistiliselt oluliselt erinevad (p = 0,020). Ühepoolse hüpoteesina testivad nii funktsioon T.TEST kui ka protseduur *t-Test: Paired Two Sample for Means* alati seda, kas suurem valimi keskmine on statistiliselt oluliselt suurem, kui väiksem valimi keskmine. Kuna antud näite puhul on testi tulemus pärast kursuse läbimist keskmiselt kõrgem, näitabki ühepoolsele testile vastav p-väärtus, et tudengite testi tulemused pärast kursuse läbimist on statistiliselt oluliselt paremad, kui enne kursuse läbimist (p = 0,010).

Kahe üldkogumi keskväärtuste võrdlemine võrdsete dispersioonide korral

Kui

- võrreldavad valimid on sõltumatud ja
- on alust eeldada uuritava tunnuse võrdset varieeruvust gruppides (dispersioonide erinevuse test funktsiooniga F.TEST andis tulemuseks 0,5-st suurema p-väärtuse – vt peatükk 5.5),

on valimite keskmiste võrdlemiseks kasutatavad funktsioon T.TEST ja protseduur *t-Test: Paired Two Sample Assuming Equal Variances* abil.

Funktsioonile T.TEST, mis annab tulemuseks vaid olulisuse tõenäosuse *p* väärtuse, tuleb ette anda

- mõlema valimi andmete blokid (*Array1* ja *Array2*),
- hüpoteesi tüüp (*Tails*): 1 – ühepoolne hüpotees (*one-tailed distribution*), 2 – kahepoolne hüpotees (*one-tailed distribution*),
- testi tüüp lähtuvalt andmete struktuurist ja varieeruvusest (*Type*): antud juhul 2 – sõltumatud valimid ja võrdsed dispersioonid (*two-sample equal variance (hoscedastic)*); ülejäänud kaks tüüpi on: 1 – sõltuvad valimid (*paired*) ning 3 – sõltumatud valimid ja erinevad dispersioonid (*two-sample unequal variance*).

Protseduuri *t-Test: Paired Two Sample Assuming Equal Variances* (*Data*-sakk -> *Data Analysis*) tellimisaknas tuleb määrata:

- mõlema valimi andmete blokid – *Variable 1 Range* ja *Variable 2 Range* (seejuures võivad andmed paikneda nii veerus kui ka reas),
- oletatav keskväärtuste erinevus (vaikimisi null) – *Hypothesized Mean Difference*,
- kui andmete blokid sisaldavad esimeses reas/veerus nime, tuleb teha "linnuke" märgendi *Labels* ette,
- olulisuse nivoo (vaikimisi 0,05) – *Alpha*,
- tulemuste väljastamise asukoht (*Output options*): samale töölehele (*Output Range*), uuele töölehele (*New Worksheet Ply*) või uude faili (*New Workbook*).

Joonisel 38 on esitatud sportivate ja mittesportivate neidude keskmiste kehamasside võrdlemine nii funktsiooniga T.TEST kui ka protseduurga *t-Test: Paired Two Sample Assuming Equal Variances* (seejuures on enne keskmiste võrdlemist võrreldud dispersioone, et olla kindel, et võrdsete dispersioonide eeldamine võrreldavates gruppides on õige). Tulemustest nähtub, et kuigi kehamasside dispersioon sportimist mitte harrastavate neidude hulgas on märksa suurem (protseduuri *t-Test: Paired Two Sample Assuming Equal Variances* väljundtabelist nähtub, et 131,7 kg^2 sportivate neidude 86,1 kg^2 vastu), ei ole see erinevus

siiski statistiliselt oluline (F-test, $p = 0{,}16$). Samuti ei ole statistiliselt oluline keskmiste kehamasside vaheline erinevus (t-test, $p = 0{,}29$; siiski võib ära märkida, et sportivate neidude kehamass on pisut suurem).

	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	MASS (Sport="Jah"	MASS (Sport="Ei")											
2	Jah	Ei		**Dispersioonide võrdlus**									
3	65	49											
4	62	53		funktsioon F.TEST	0.1613	= F.TEST(U3:U81, V3:V28)							
5	47	79			p = 0,16 > 0,05 -> võib kasutada t-testi, mis eeldab kehamasside võrdset varieeruvust								
6	67	69											
7	57	58		**Keskmiste võrdlus**									
8	60	59											
9	65	55		funktsioon T.TEST									
10	55	59		Ühepoolne hüpotees	0.1434	= T.TEST(U3:U81, V3:V28 ,1,2)							
11	55	100		Kahepoolne hüpotees	0.2868	= T.TEST(U3:U81, V3:V28, 2,2)							
12	80	80											
13	62	64											
14	68	55		t-Test: Two-Sample Assuming Equal Variances									
15	85	55											
16	58	58			*Jah*	*Ei*							
17	54	70		Mean	63.082	60.696	Keskmine võrreldavais gruppides						
18	57	63		Variance	86.086	131.66	Teadaolev dispersioon gruppides						
19	61	52		Observations	79	26	Vaatluste arv						
20	63	59		Pooled Variance	97.147		Ühine dispersioon						
21	60	54		Hypothesized Mean Dif	0		Hüpoteetiline keskmiste vaheline erinevus						
22	75	50		df	103		Vabadusastmete arv (n-1)						
23	60	54		t Stat	1.0707		t-statistiku väärtus						
24	64	53		P(T<=t) one-tail	0.1434		Ühepoolsele hüpoteesile vastav p-väärtus						
25	74	63		t Critical one-tail	1.6598		Ühepoolsele hüpoteesile vastav z-statistiku kriitiline väärtus ($t_{0,95}(df)$)						
26	65	65		P(T<=t) two-tail	0.2868		Kahepoolsele hüpoteesile vastav p-väärtus						
27	61	50.1		t Critical two-tail	1.9833		Kahepoolsele hüpoteesile vastav z-statistiku kriitiline väärtus ($t_{0,975}(df)$)						

Joonis 38. Sportivate ja mittesportivate neidude keskmiste kehamasside võrdlemine funktsiooniga T.TEST ja protseduuriga *t-Test: Paired Two Sample Assuming Equal Variances*.

Kahe üldkogumi keskväärtuste võrdlemine erinevate dispersioonide korral

Kui

- võrreldavad valimid on sõltumatud ja
- uuritava tunnuse varieeruvus gruppides on erinev (dispersioonide erinevuse test funktsiooniga F.TEST andis tulemuseks 0,5-st väiksema p-väärtuse – vt peatükk 5.5),

on valimite keskmiste võrdlemiseks kasutatavad funktsioon T.TEST ja protseduur *t-Test: Paired Two Sample Assuming Unequal Variances*.

Mõlemal juhul on ette antavad argumendid ja väljund analoogsed samas peatükis eelnevalt kirjeldatuga. Vaid protseduuri *t-Test: Paired Two Sample Assuming Unequal Variances* väljundtabelis on üks rida vähem – et dispersioonid on erinevad, ei kasutata arvutamisel enam kahe valimi ühist dispersiooni (*Pooled Variance*).

5.5. F-test

Kahe üldkogumi dispersioonide võrdlemiseks on Excelis kasutatavad funktsioon F.TEST ja protseduur *F-Test: Two-Sample for Variances*.

Seejuures testib funktsioon F.TEST vaid kahepoolset hüpoteesi kujul

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_1: \sigma_1^2 \neq \sigma_2^2$$

kus σ_1^2 ja σ_2^2 on vastavalt esimese ja teise valimi dispersioon, ning protseduur *F-Test: Two-Sample for Variances* vaid ühepoolset hüpoteesi kujul

$$H_0: \sigma_1^2 \leq \sigma_2^2$$
$$H_1: \sigma_1^2 > \sigma_2^2$$

kus σ_1^2 ja σ_2^2 on vastavalt suurema ja väiksema varieeruvusega valimi dispersioonid (st, et protseduur *F-Test: Two-Sample for Variances* uurib alati, kas suurem dispersioon on ka statistiliselt oluliselt suurem).

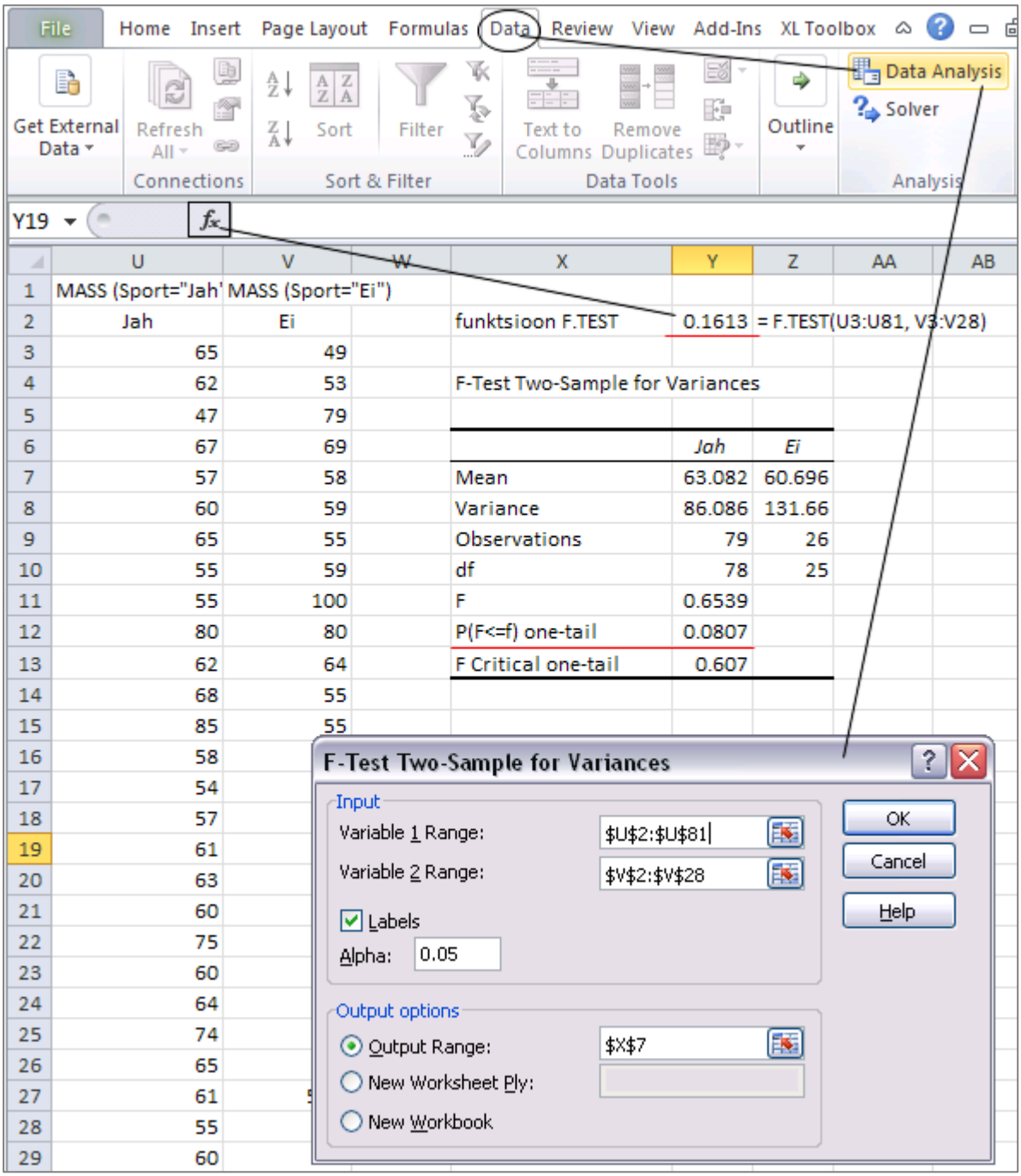
Funktsioonile F.TEST, mis annab tulemuseks vaid olulisuse tõenäosuse *p* väärtuse, tuleb ette anda mõlema valimi andmete blokid (*Array1* ja *Array2*).

Protseduur *F-Test: Two-Sample for Variances* (*Data*-sakk -> *Data Analysis*) annab tulemuseks nii võrreldavaid gruppe kirjeldavad karakteristikud kui ka F-testi teostamisega kaasnevad arvutustulemused ning selle rakendamiseks tuleb ette anda:

- mõlema valimi andmete blokid – *Variable 1 Range* ja *Variable 2 Range*,
- kui andmete blokid sisaldavad esimeses reas/veerus nime, tuleb teha "linnuke" märgendi *Labels* ette,
- olulisuse nivoo (vaikimisi 0,05) – *Alpha*,
- tulemuste väljastamise asukoht (*Output options*): samale töölehele (*Output Range*), uuele töölehele (*New Worksheet Ply*) või uude faili (*New Workbook*).

Joonisel 39 on näha nii funktsiooni F.TEST kui ka protseduuri *F-Test: Two-Sample for Variances* rakendamine ja tulemused võrdlemaks sportivate ja mittesportivate neidude kehamasside varieeruvust. Tulemustest nähtub, et kuigi kehamasside dispersioon sportimist mitte harrastavate neidude hulgas on märksa suurem (protseduuri *F-Test: Two-Sample for Variances* väljundtabelist nähtub, et 131,7 kg^2 sportivate neidude 86,1 kg^2 vastu), ei ole see erinevus siiski statistiliselt oluline (kahepoolsele hüpoteesile vastav p-väärtus funktsiooni F.TEST tulemusena on 0,16). Tõestatuks ei saa lugada ka ühepoolset hüpoteesi (protseduuri *F-Test: Two-Sample for Variances* poolt väljastatud ühepoolsele hüpoteesile vastav p-väärtus on 0,081).

NB! Et kahepoolsele hüpoteesile vastav olulisuse tõenäosus võrdub kahekordse ühepoolsele hüpoteesile vastava olulisuse tõenäosusega, saab mõlemat tüüpi hüpoteese kontrollida nii funktsiooniga F.TEST kui ka protseduuriga *F-Test: Two-Sample for Variances*, peab üksnes meeles pidama, mis tüüpi hüpoteesi kumbki neist kontrollib ja siis vajadusel tulemuse kahega jagama või korrutama.



Joonis 39. Sportivate ja mittesportivate neidude kehamasside varieeruvuse võrdlemine funktsiooniga F.TEST ja protseduuriga *F-Test: Two-Sample for Variances*.

5.6. Mitteparameetrilised testid

Juhul, kui uuritav tunnus ei ole normaaljaotusega ja valimi maht ei ole ka suur, ei ole Excelis olemas olevate z- ja t-testi rakendamine keskmiste võrdlemiseks korrektne (nagu ei ole korrektne ka dispersioonide võrdlemine F-testiga) ning kasutada tuleks mitteparameetrilisi normaaljaotust mitte-eeldavaid teste. Viimaste teostamiseks Excelis sisseehitatud vahendid puuduvad.

Siiski on mõnede mitteparameetriliste testide läbiviimine Excelis võimalik – kas siis testide aluseks olevate arvutuste samm-sammulise teostamise või spetsiaalsete lisamoodulite abil.

Märgitest funktsiooni BINOM.DIST abil

Näitena testi samm-sammulisest teostamisest on järgnevalt tutvustatud lihtsaima kahe sõltuva valimi võrdlemisel kasutatavat testi – märgitesti.

Kuna andmete näol on tegu sõltuvate (paaris) valimitega, saab iga indiviidi/objekti tarvis leida toimunud muutuse suuruse.

Märgitest

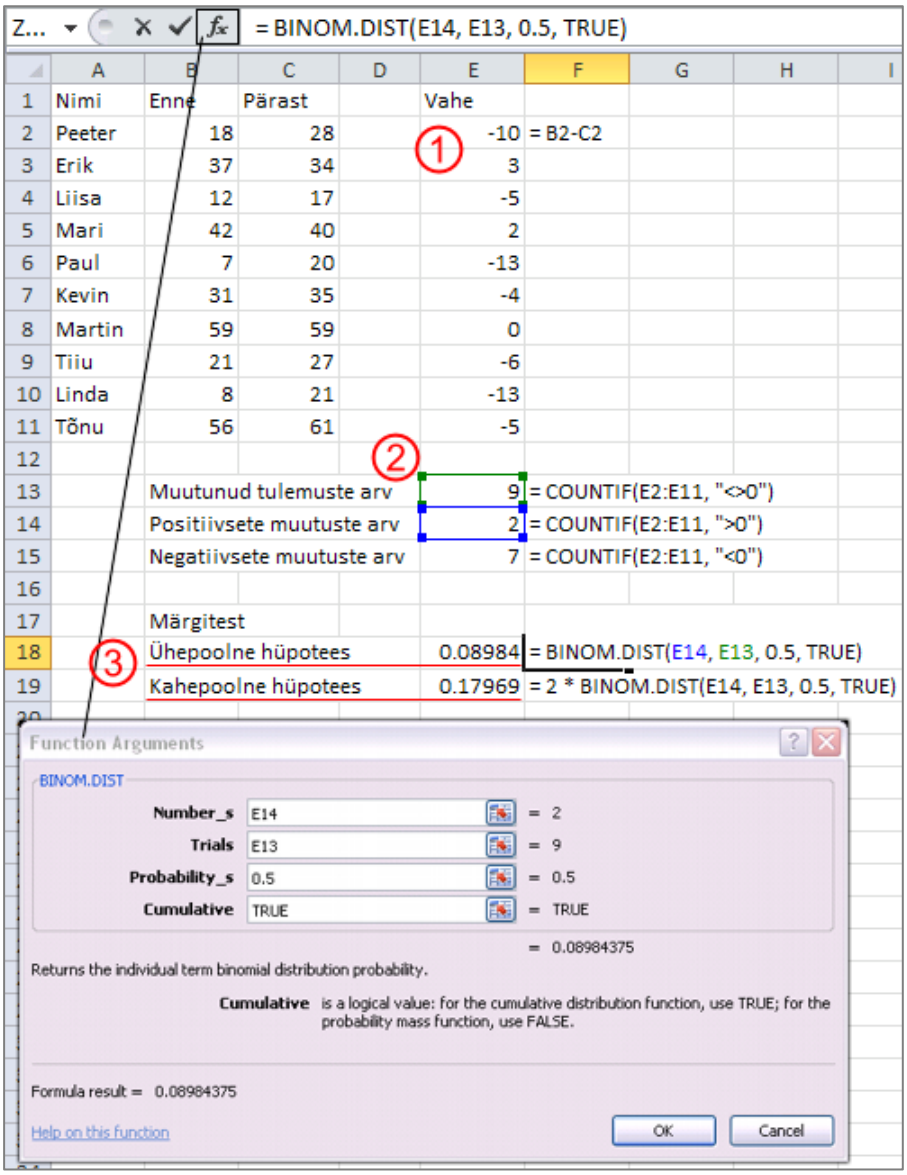
- loeb kokku, kui mitme indiviidi/objekti puhul üldse mingi muutus toimus (n_0) ja
- kui mitmel juhul oli muutus positiivne (miinusmärgiga, N_+) ja/või negatiivne (miinusmärgiga, N_-), ning
- leiab, tuginedes binoomjaotusele B(n_0, 0,5), kui suure tõenäosusega võinuks nii suur hulk samasuunalisi muutusi olla toimunud juhuslikult (juhusliku muutumise korral peaks iga indiviidi/objekti puhul olema nii positiivse kui ka negatiivse muutuse tõenäosus 0,5 – sellest ka binoomjaotuse teise parameetri väärtus).

Märgitesti teostamiseks (st olulisuse tõenäosuse *p* arvutamiseks) Excelis tuleb (Joonis 40)

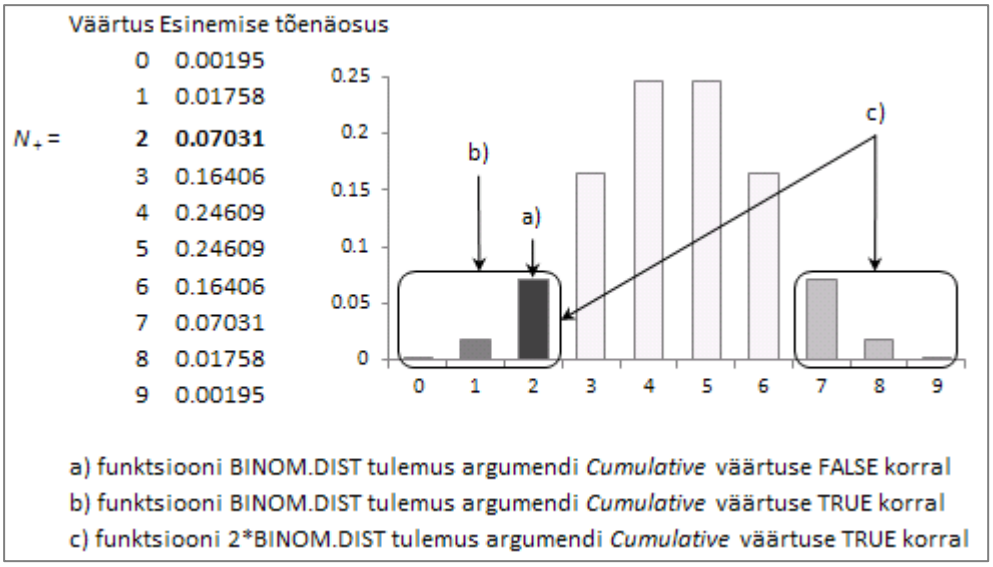
1. leida kõigi väärtustepaaride vahed,
2. lugeda kokku, kui mitmel juhul on uuritava tunnuse väärtus muutunud ning kui mitmel juhul oli muutus positiivne (ja/või negatiivne),
3. rakendada funktsiooni BINOM.DIST, millele tuleb ette anda
   - positiivsete või negatiivsete muutuste arv N_+ või N_- (*Number_s*),
   - kõigi toimunud muutuste arv n_0 (*Trials*),
   - positiivse muutuse toimumise tõenäosus nullhüpoteesi eeldusel, so 0,5 (*Probability_s*),
   - väärtus TRUE argumendile *Cumulative* (siis väljastab funktsioon BINOM.DIST nii antud muutuste arvu kui ka sellest vähemtõenäoliste muutuste arvu summaarse tõenäosuse; väärtuse FALSE puhul on tulemuseks vaid antud muutuste arvu tõenäosus – vt ka Joonis 41).

Et funktsiooni BINOM.TEST tulemuseks argumendi *Cumulative* väärtuse TRUE puhul on vaid ühepoolsele hüpoteesile vastav olulisuse tõenäosus *p*, tuleb standardse kahepoolsele hüpoteesile vastava p-väärtuse saamiseks funktsiooni BINOM.TEST tulemus korrutada kahega (Joonis 40).

Joonisel 40 esitatud märgitesti tulemusest nähtub, et tudengite testi tulemused enne ja pärast kursuse läbimist ei ole statistiliselt oluliselt erinevad (p = 0,18). Tulemus on erinev peatüki 5.5 alguses t-testiga leitust (Joonis 37), kus p-väärtus tuli 0,020. Põhjus on märgitesti robustsuses võrreldes t-testiga – t-test eeldab andmete normaaljaotuse-järgset jaotumist, märgitesti puhul on eelduseks vaid uuritava tunnuse väärtuste järjestatavus, lisaks ei arvesta märgitest toimunud muutuste suurusega.



Joonis 40. Märgitest Excelis funktsiooni BINOM.DIST abil.



Joonis 41. Funktsiooni BINOM.DIST tulemus sõltuvalt argumendi *Cumulative* väärtusest.

Lisamoodul „Kahe üldkogumi võrdlus“

Aastal 2005. kaitses Anu Iher Tartu Ülikooli matemaatilise statistika instituudis bakalaureuse töö „Olulisemad kahe üldkogumi võrdlemise testid ja MS Excel'i moodul nende läbiviimiseks“. Tööga, mis annab teoreetilise ja põhjaliku ülevaate erinevatest mitteparameetrilistest kahe üldkogumi keskväärtuste võrdlemisel kasutatavatest testidest, saab tutvuda siin: http://www.eau.ee/~ktanel/baca_AIher_2005.pdf.

Töö osana valminud Exceli lisamooduli ja selle abifaili saab alla laadida aadressilt http://ph.emu.ee/~ktanel/excel_addins/.

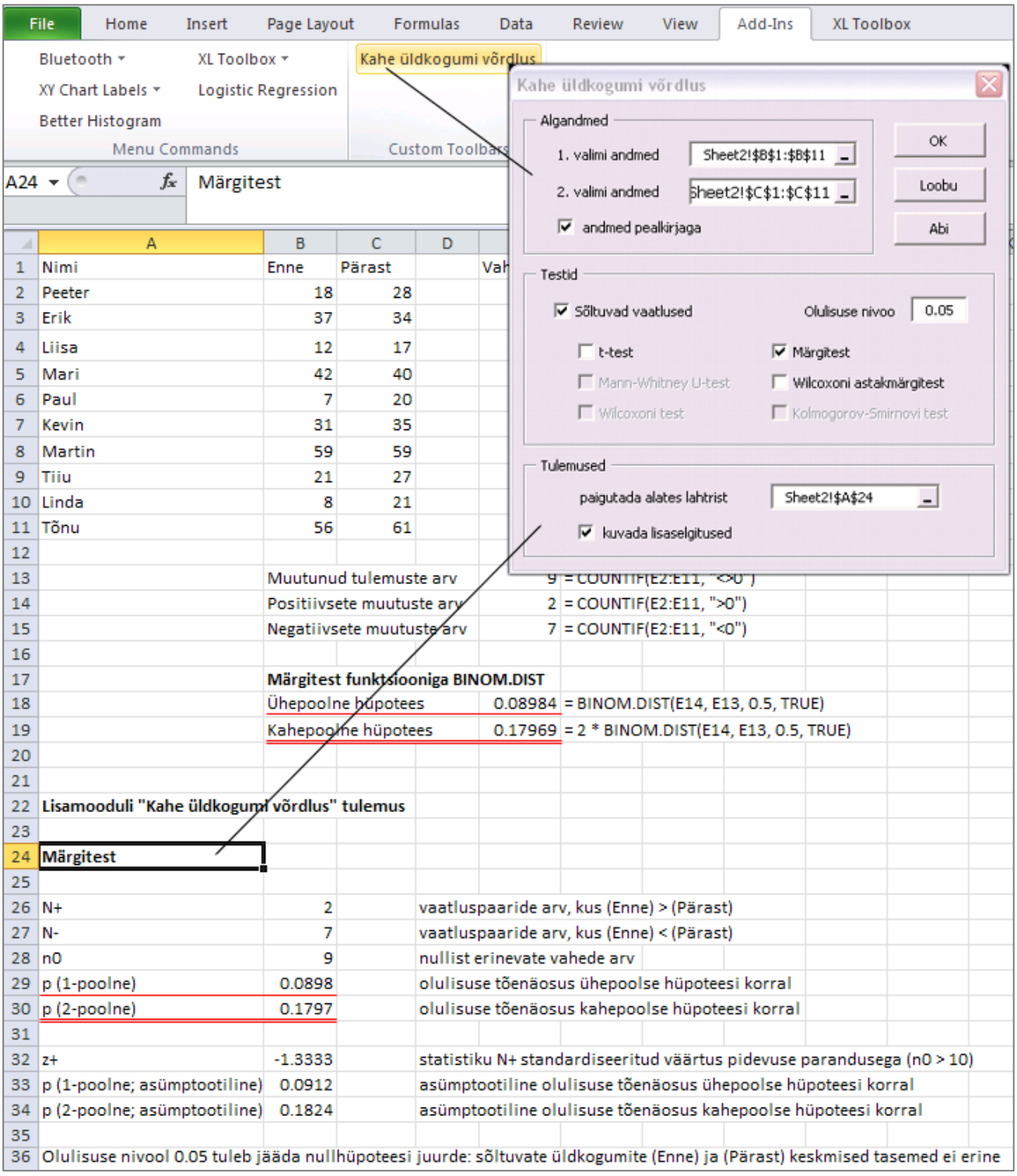
Lisamooduli rakendamiseks tekib peale selle installeerimist (analoogselt statistika-protseduuride paketi *Data Analysis* kasutuselevõtuga – vt pt 1.3) Exceli lisamoodulite saki (*Add-Ins*-sakk) alla valik |Kahe üldkogumi võrdlus|.

Lisamooduli tellimisaken on analoogne Exceli statistikaprotseduuride tellimisaknaga, määrata tuleb

- võrreldavate valimite andmed (võivad paikneda nii veergudes kui ka ridades),
- pealkirja olemasolu ette antud valimite esimeses reas/veerus,
- olulisuse nivoo (vaikimisi 0,05),
- võrreldavate valimite tüüp – sõltuvad või sõltumatud – ning soovitud test(id),

  NB!
  - korraga võib tellida mitu testi,
  - tellides sõltumatute valimite korral t-testi, teostatakse mõlemad, nii võrdseid kui ka erinevaid dispersioone eeldavad t-testid, ning lisaks ka F-test dispersioonide võrdlemiseks,
- väljundi asukoht,
- lisaselgituste soov (lisaks kõiksugu statistikute nimetustele/tähistustele ja arvutuste tulemustele kuvatakse väljundtabeleis ka vähe pikemad selgitused, sh lõppjäreldus).

Joonisel 42 on lisamoodulit „Kahe üldkogumi võrdlus“ rakendatud tudengite testitulemuste võrdlemiseks märgitestiga. Tulemused on identsed eelnevalt funktsiooni BINOM.TEST abil arvutatutega. Ainult lisaks täpsetele binoomjaotusel baseeruvatele p-väärtustele arvutab lisamoodul „Kahe üldkogumi võrdlus“ ka ligikaudsed normaaljaotusel baseeruvad p-väärtused – taolise tegevuse mõte on selles, et mitmete mitteparameetriliste testide arvutuseeskirjad on nende rakendamiseks suurte valimite puhul liiga töömahukad, samas on kasutatavate teststatistikute jaotus suurte valimite puhul lähendav standardse normaaljaotusega ja sestap saab sellisel juhul ka p-väärtuste arvutamisel lähtuda standardsest normaaljaotusest (z-statistikust).



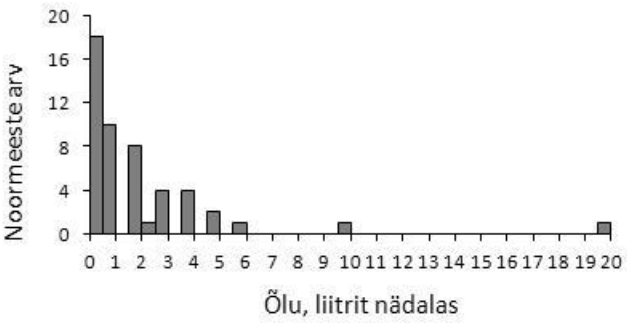
	A	B	C	D	
1	Nimi	Enne	Pärast		Vah
2	Peeter	18	28		
3	Erik	37	34		
4	Liisa	12	17		
5	Mari	42	40		
6	Paul	7	20		
7	Kevin	31	35		
8	Martin	59	59		
9	Tiiu	21	27		
10	Linda	8	21		
11	Tõnu	56	61		
12					
13		Muutunud tulemuste arv		9	= COUNTIF(E2:E11, "<>0")
14		Positiivsete muutuste arv		2	= COUNTIF(E2:E11, ">0")
15		Negatiivsete muutuste arv		7	= COUNTIF(E2:E11, "<0")
16					
17		**Märgitest funktsiooniga BINOM.DIST**			
18		Ühepoolne hüpotees		0.08984	= BINOM.DIST(E14, E13, 0.5, TRUE)
19		Kahepoolne hüpotees		0.17969	= 2 * BINOM.DIST(E14, E13, 0.5, TRUE)
20					
21					
22	**Lisamooduli "Kahe üldkogumi võrdlus" tulemus**				
23					
24	**Märgitest**				
25					
26	N+	2		vaatluspaaride arv, kus (Enne) > (Pärast)	
27	N-	7		vaatluspaaride arv, kus (Enne) < (Pärast)	
28	n0	9		nullist erinevate vahede arv	
29	p (1-poolne)	0.0898		olulisuse tõenäosus ühepoolse hüpoteesi korral	
30	p (2-poolne)	0.1797		olulisuse tõenäosus kahepoolse hüpoteesi korral	
31					
32	z+	-1.3333		statistiku N+ standardiseeritud väärtus pidevuse parandusega (n0 > 10)	
33	p (1-poolne; asümptootiline)	0.0912		asümptootiline olulisuse tõenäosus ühepoolse hüpoteesi korral	
34	p (2-poolne; asümptootiline)	0.1824		asümptootiline olulisuse tõenäosus kahepoolse hüpoteesi korral	
35					
36	Olulisuse nivool 0.05 tuleb jääda nullhüpoteesi juurde: sõltuvate üldkogumite (Enne) ja (Pärast) keskmised tasemed ei erine				

Joonis 42. Märgitesti rakendamine tudengite testitulemuste võrdlemiseks lisamooduliga „Kahe üldkogumi võrdlus“.

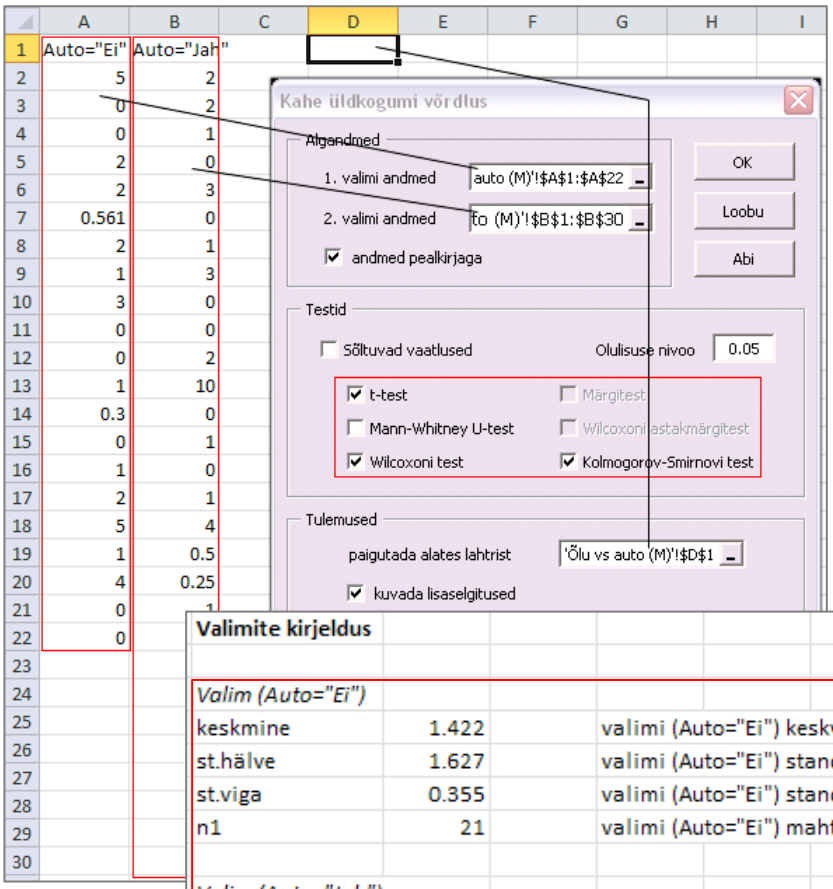
Joonisel 43 on korraga teostatud t-test, Wilcoxoni test ja Komogorov-Smirnovi test võrdlemaks autot omavate ja mitte omavate esimese kursuse noormeeste nädalas tarbitavaid õllekoguseid. Joonisel 43A on ära toodud analüüside tellimisaken ja osa lisaselgitustega varustatud väljundist – valimite kirjeldus ning t- ja F-testi tulemused. Joonisel 43B on ära toodud ülejäänud osa väljundist – Wilcoxoni ja Komogorov-Smirnovi testi tulemused.

Tulemustest nähtub, et autot omavad noormehed joovad nädalas keskmiselt 1,1 liitrit enam õlut kui autot mitte omavad noormehed (keskmised nädalas tarbitavad õllekogused on vastavalt 2,5 ja 1,4 liitrit), samas ei saa seda erinevust lugeda statistiliselt oluliseks (erinevatele dispersioonidele vastav t-test, $p = 0{,}19$). Küll võib varieeruvuse võrreldavates gruppides lugeda statistiliselt oluliselt erinevaks (F-test, $p < 0{,}001$) – seetõttu tuleb keskmiste võrdlemisel vaadata erinevatele dispersioonidele vastava t-testi tulemusi.

Iseküsimus on muidugi F- ja t-testi eelduste täidetus – noormeeste nädalas tarbitud õllekogused ei jaotu kohe kindlasti normaaljaotuse järgi (vt kõrvalolev joonis). Seetõttu on korrektsem kasutada autot omavate ja mitte omavate esimese kursuse noormeeste nädalas tarbitavate õllekoguste võrdlemiseks mitteparameetrilisi teste. Joonisel 43B ongi esitatud neist kahe tulemused.



Wilcoxoni testi täpse p-väärtuse arvutamiseks on andmeid liiga palju, mistõttu tuleb järeldused teha asümptootilise p-väärtuse alusel. Sarnaselt t-testile ei anna ka Wilcoxoni ja Komogorov-Smirnovi test alust lugeda autot omavate ja mitte omavate esimese kursuse noormeeste nädalas tarbitavaid õllekoguseid statistiliselt oluliselt erinevateks (vastavalt $p = 0{,}42$ ja $p = 0{,}88$). See, et p-väärtused suuremad, kui t-testi puhul, on loomulik, sest mõlemad mitteparameetrilised testid kontrollivad üldisemaid hüpoteese – Wilcoxoni test kahe valimi elementide mittejuhuslikku segunemist ja Kolmogorov-Smirnovi test jaotuste erinevust.



	A	B	C	D	E	F	G	H	I
1	Auto="Ei"	Auto="Jah"							
2	5	2							
3	0	2							
4	0	1							
5	2	0							
6	2	3							
7	0.561	0							
8	2	1							
9	1	3							
10	3	0							
11	0	0							
12	0	2							
13	1	10							
14	0.3	0							
15	0	1							
16	1	0							
17	2	1							
18	5	4							
19	1	0.5							
20	4	0.25							
21	0	1							
22	0								

Valimite kirjeldus

Valim (Auto="Ei")		
keskmine	1.422	valimi (Auto="Ei") keskväärtus
st.hälve	1.627	valimi (Auto="Ei") standardhälve
st.viga	0.355	valimi (Auto="Ei") standardviga
n1	21	valimi (Auto="Ei") maht
Valim (Auto="Jah")		
keskmine	2.5259	valimi (Auto="Jah") keskväärtus
st.hälve	4.0247	valimi (Auto="Jah") standardhälve
st.viga	0.7474	valimi (Auto="Jah") standardviga
n2	29	valimi (Auto="Jah") maht

Studenti t-test sõltumatute valimite korral

Võrdsed üldkogumite dispersioonid		
t	-1.186	t-statistiku väärtus
df	48	t-jaotuse vabadusastmete arv
p (1-poolne)	0.1207	olulisuse tõenäosus ühepoolse hüpoteesi korral
p (2-poolne)	0.2415	olulisuse tõenäosus kahepoolse hüpoteesi korral
Erinevad üldkogumite dispersioonid		
t	-1.3342	t-statistiku väärtus
df	39.2647	t-jaotuse vabadusastmete arv (SatterthWaite'i meetodil)
p (1-poolne)	0.0949	olulisuse tõenäosus ühepoolse hüpoteesi korral
p (2-poolne)	0.1898	olulisuse tõenäosus kahepoolse hüpoteesi korral
F-test üldkogumite dispersioonide võrdlemiseks		
F	0.1634	F-statistiku ehk valimite dispersioonide suhte (V1/V2) väärtus
f1	20	F-jaotuse 1. vabadusastmete arv
f2	28	F-jaotuse 2. vabadusastmete arv
p (2-poolne)	0.0001	olulisuse tõenäosus kahepoolse hüpoteesi korral

Olulisuse nivool 0.05 tuleb jääda nullhüpoteesi juurde: üldkogumite (Auto="Ei") ja (Auto="Jah") keskväärtused ei erine

Joonis 43A. Autot omavate ja mitte omavate esimese kursuse noormeeste nädalas tarbitavate õllekoguste võrdlus lisamooduliga „Kahe üldkogumi võrdlus“ – valimite kirjeldus ning t-test.

Wilcoxoni test		
W1	494	valimi (Auto="Ei") astakute summa ühises variatsioonreas
W2	781	valimi (Auto="Jah") astakute summa ühises variatsioonreas
W1/n1	23.5238	valimi (Auto="Ei") keskmine astak
W2/n2	26.931	valimi (Auto="Jah") keskmine astak
U-	263	U- avaldatud astakute summade W1 ja W2 kaudu
U+	346	U+ avaldatud astakute summade W1 ja W2 kaudu
U	263	U-statistik Min(U-, U+)
p (1-poolne)		valimid on täpse olulisuse tõenäosuse leidmiseks liiga suured
p (2-poolne)		valimid on täpse olulisuse tõenäosuse leidmiseks liiga suured
zW	-0.8059	standardiseeritud Wilcoxoni statistik pidevuse parandusega kasutades valimi (Auto="Ei") astakute summat (n1, n2 > 8)
p (1-poolne; asü	0.2102	asümptootiline olulisuse tõenäosus ühepoolse hüpoteesi korral
p (2-poolne; asü	0.4203	asümptootiline olulisuse tõenäosus kahepoolse hüpoteesi korral
Olulisuse nivool 0.05 tuleb jääda nullhüpoteesi juurde: üldkogumite (Auto="Ei") ja (Auto="Jah") keskmised tasemed ei erine		
Kolmogorov-Smirnovi test		
D+	0.1544	empiiriliste jaotusfunktsioonide vahe G(Auto="Ei")-G(Auto="Jah") maksimaalväärtus
D-	0	empiiriliste jaotusfunktsioonide vahe G(Auto="Jah")-G(Auto="Ei") maksimaalväärtus
D	0.1544	empiiriliste jaotusfunktsioonide G(Auto="Ei") ja G(Auto="Jah") maksimaalse erinevuse absoluutväärtus
p (2-poolne)	0.8797	olulisuse tõenäosus kahepoolse hüpoteesi G(Auto="Ei")<>G(Auto="Jah") korral
cD+	0.5387	statistik cD+
p (G1>G2; asüm	0.5021	asümptootiline olulisuse tõenäosus ühepoolse hüpoteesi G(Auto="Ei")>G(Auto="Jah") korral
cD-	0	statistik cD-
p (G1<G2; asüm	1	asümptootiline olulisuse tõenäosus ühepoolse hüpoteesi G(Auto="Ei")<G(Auto="Jah") korral
cD	0.5387	statistik cD
p (G1<>G2; asün	0.9337	asümptootiline olulisuse tõenäosus kahepoolse hüpoteesi G(Auto="Ei")<>G(Auto="Jah") korral
Olulisuse nivool 0.05 tuleb jääda nullhüpoteesi juurde: üldkogumite (Auto="Ei") ja (Auto="Jah") jaotused ei erine		

Joonis 43B. Autot omavate ja mitte omavate esimese kursuse noormeeste nädalas tarbitavate õllekoguste võrdlus lisamooduliga „Kahe üldkogumi võrdlus“ – Wilcoxoni ja Komogorov-Smirnovi test.

6. Korrelatsioonanalüüs

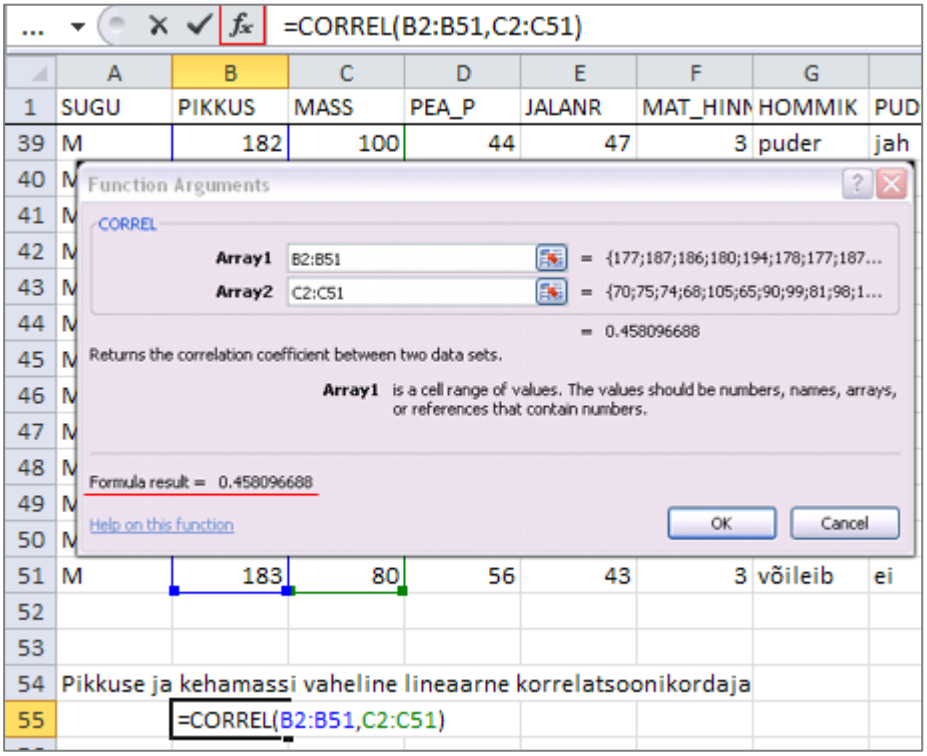
6.1. Pearsoni e lineaarne korrelatsioonikordaja

Funktsioonid CORREL ja PEARSON

Lihtsaim viis arvtunnuste vahelise lineaarse seose kirjeldamiseks on korrelatsioonanalüüs.

Kahe tunnuse vaheline lineaarne e Pearsoni korrelatsioonikordaja on Exceli keskkonnas leitav funktsioonidega CORREL ja PEARSON, millele tuleb ühte moodi ette anda kahe uuritava tunnuse väärtused (*Array1* ja *Array2*). Et ka mõlema funktsiooni tulemus on sama (sest arvutusvalem on sama ja minule on arusaamatu, miks neid funktsioone üldse kaks peab olema), on järgnevalt näidatud vaid funktsiooni CORREL rakendamist.

Joonisel 44 on kujutatud noormeeste pikkuse ja kehamassi vahelise korrelatsioonikordaja arvutamist Excelis. Tulemusena väljastatud lineaarse korrelatsioonikordaja väärtusest 0,46 järeldub, et noormeeste pikkuse ja kehamassi vahel on keskmise tugevusega positiivne seos – mida pikemad on noormehed, seda enam nad keskmiselt ka kaaluvad.



Joonis 44. Noormeeste pikkuse ja kehamassi vahelise lineaarse korrelatsioonikordaja arvutamine funktsiooniga CORREL.

Protseduur *Correlation*

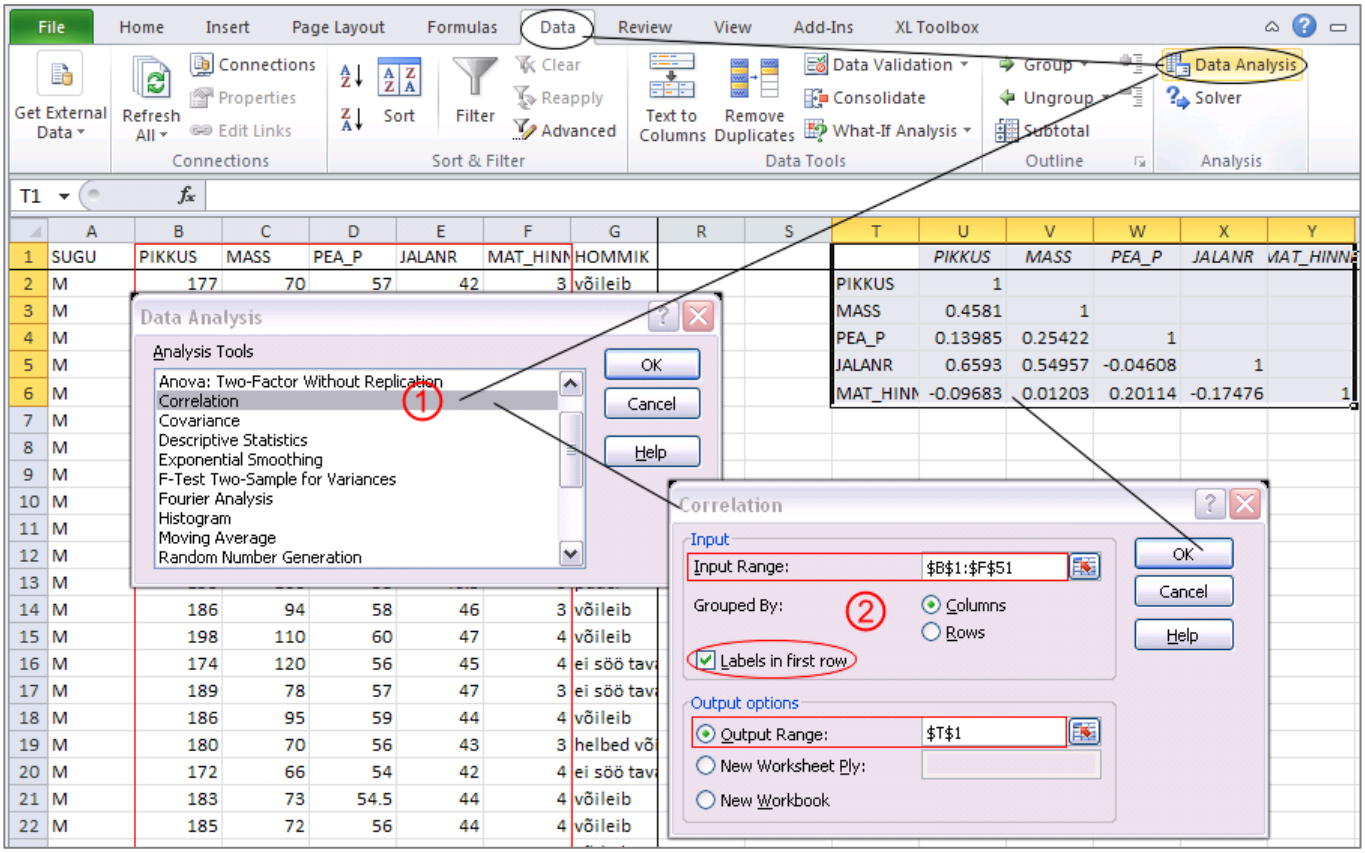
Mitme tunnuse korral on paarikaupa korrelatsioonikordajate tabeli (korrelatsioonimaatriksi) arvutamiseks kasutatav protseduur *Correlation* (Joonis 45):

1. *Data*-sakk -> *Data Analysis* -> *Correlation*,
2. avanevas sisestusaknas tuleb määrata:
   - *Input Range* – algandmete blokk (tunnused peavad paiknema järjestikustes veergudes ja olema arvulised);

- *Grouped by* – määratakse, kas tunnusvektorid on orienteeritud veerge (*Columns*, vaikimisi variant) või ridu pidi (*Rows*);
- *Labels in First Row* – märgitakse nimede või tähiste olemasolu korral tunnuste bloki esimeses reas;
- *Output options* – määratakse tulemuste väljastamise asukoht: samale töölehele (*Output Range*), uuele töölehele (*New Worksheet Ply*) või uude faili (*New Workbook*).

Tulemuseks on Exceli töölehele väljastatav kolmnurkse kujuga korrelatsioonikordajate maatriks.

Joonisel 45 esitatud korrelatsioonanalüüsi tulemustest ilmneb näiteks, et noormeeste pikkuse ja kehamassi vahel on keskmise tugevusega positiivne seos ($r = 0{,}46$) – mida pikemad on noormehed, seda enam nad keskmiselt ka kaaluvad, peaümbermõõdu ja matemaatika hinde vahel on aga nõrk positiivne ning jalanumbri ja matemaatika hinde vahel nõrk negatiivne seos – mida suurem on peaümbermõõt ja mida väiksem on jalanumber, seda kõrgem on keskmiselt matemaatika hinne.



Joonis 45. Noormeeste pikkuse, kehamassi, peaümbermõõdu, jalanumbri ja matemaatika hinde vaheliste lineaarsete korrelatsioonikordajate arvutamine protseduuriga *Correlation*.

6.2. Lineaarse korrelatsioonikordaja statistiline olulisus

Korrelatsioonikordaja *r* statistilise olulisuse kontrollimine seisneb hüpoteeside paari

$$H_0: r = 0$$
$$H_1: r \neq 0$$

testimises.

Kahjuks ei väljasta Excel korrelatsioonanalüüsi läbi viies automaatselt taoliste hüpoteeside kontrollimiseks vajalikke näitajaid (korrelatsioonikordaja või teststatistiku kriitilist väärtust või olulisuse tõenäosust *p*). Lahendusena tuleb kõne alla vähemalt kolm varianti.

Esiteks võib kasutada korrelatsioonikordajate kriitiliste väärtuste tabelit, mis on leitav enamuse statistikaõpikute lisades ja ka näiteks veebiaadressilt http://www.eau.ee/~ktanel/VL_0435/critical_values_of_Pearson_cor.pdf – kui leitud korrelatsioonikordaja väärtus on suurem vastavast kriitilisest väärtusest (viimane sõltub kordaja arvutamisel kasutatud väärtuste paaride arvust *n* ja olulisuse nivoost α), võib lugeda tõestatuks alternatiivse hüpoteesi H_1: korrelatsioonikordaja on nullist erinev ehk seos on statistiliselt oluline, vastasel juhul peab jääma nullhüpoteesi juurde.

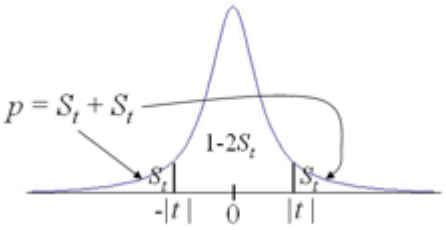
Teine võimalus uuritava lineaarse seose statistilise olulisuse kontrollimiseks on teostada kahe uuritava tunnusega tavaline lineaarne regressioon protseduuri *Regression* abil. Lineaarse regressiooniseose statistilist olulisust iseloomustav olulisuse tõenäosus *p* kehtib ka lineaarse korrelatsioonikordaja jaoks (täpsemalt vt peatükk 7.1).

Kolmas võimalus on leida olulisuse tõenäosuse *p* väärtus tuginedes teststatistikule

$$t = r\sqrt{n-2}\Big/\sqrt{1-r^2}\ ,$$

mis on nullhüpoteesi kehtides ligikaudu t-jaotusega parameetriga $n - 2$.

Otsuse, kumb hüpoteesidest on õige, vastu võtmiseks vajalik olulisuse tõenäosus *p* kujutab enesest leitud teststatistiku väärtuse poolt ära lõigatud *t*-jaotuse sabade osakaalu (kõrvaloleval joonisel pindalade S_t summa).

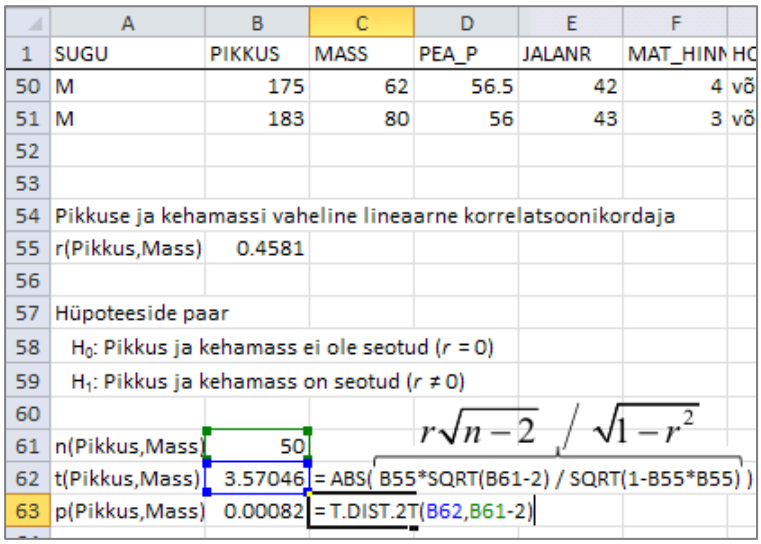


Excelis on p-väärtus leitav funktsiooniga T.DIST.2T, kus esimesena argumendina tuleb ette anda eelnevalt toodud teststatistiku absoluutväärtus ja teise argumendina korrelatsioonikordaja arvutamisel kasutatud väärtuste paaride arv $n - 2$.

Kui leitud olulisuse tõenäosus $p < 0{,}05$, võib lugeda kahe tunnuse vahelise seose statistiliselt oluliseks.

Joonisel 46 on esitatud noormeeste pikkuse ja kehamassi vahelise lineaarse korrelatsioonikordaja statistilise olulisuse testimine, kus vahetulemustena on välja kirjutatud ka vaatluste arv *n* ja teststatistiku absoluutväärtus $|t|$.

Tulemustest võib järeldada, et noormeeste pikkuse ja kehamassi vahel on keskmise tugevusega positiivne statistiliselt oluline seos ($r = 0{,}46$, $p < 0{,}001$).

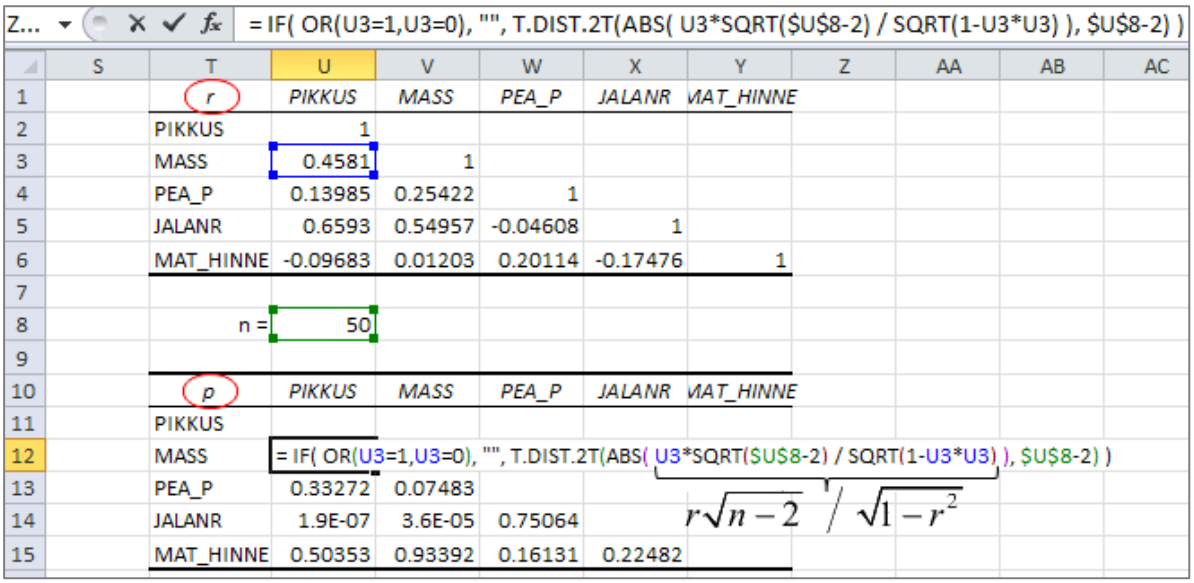


Joonis 46. Noormeeste pikkuse ja kehamassi vahelise lineaarse korrelatsioonikordaja statistilise olulisuse testimine.

Juhul, kui olulisuse tõenäosuseid soovitakse arvutada tervele korrelatsioonikordajate maatriksile (leituna protseduuriga *Correlation*), on mugav koondada arvutused analoogsesse tabelisse:

1. teha korrelatsioonikordajate tabelist koopia ja kustutada ära kopeeritud tabeli sisu (et ka hiljem oleks selge, mis arvud mis tabelis on, võib korrelatsioonikordajate ja loodava p-väärtuste tabeli ülemisse vasakusse nurka kirjutada vastava kordaja nime),
2. sisestada p-väärtuste tabeli lahtrisse valem olulisuse tõenäosuse arvutamiseks (Joonis 47),
   a) andes argumendina ette vastava korrelatsioonikordaja eelmises tabelis (lahtri aadressina) ja
   b) vaatluste arvu kas viitena seda sisaldavale lahtrile (NB! siis peab selle lahtri aadress olema fikseeritud) või lihtsalt arvuna ning
   c) lisades soovi korral valemi algusesse tingimuse funktsiooniga IF, mis juhul, kui korrelatsioonikordajate tabelis on arv 1 (peadiagonaalil) või mitte midagi (ülalpool peadiagonaali), jätab vastavad lahtrid p-väärtuste tabelis tühjaks,
3. kopeerida sisestatud valem kõigisse p-väärtuste tabeli lahtritesse.

NB! Kui arvutustes kasutatud vaatluste arv *n* on erinevate korrelatsioonikordajate puhul erinev (puuduvate väärtuste arv erinevatel tunnustel ja nende paaridel on erinev), tuleks enne p-väärtuste tabeli konstrueerimist teha analoogse struktuuriga tabel ka vaatluste arvude *n* tarvis ning kasutada p-väärtuste arvutamisel konktreetsele tunnuste paarile vastavat vaatluste arvu sellest tabelist.



Joonis 47. Olulisuse tõenäosuste maatriksi arvutamine korrelatsioonikordajate maatriksi alusel.

6.3. Spearmani e astakkorrelatsioonikordaja

Kui uuritavate arvtunnuste näol on tegu pidevate ja sümmeetrilise jaotusega tunnustega, on loomulikeim seosekordaja Pearsoni e lineaarne korrelatsioonikordaja, mis mõõdab kahe arvtunnuse vahelise lineaarse seose tugevust ja suunda.

Kui aga tunnused ei ole normaaljaotusega, leidub üksikuid erandlikke väärtuseid või ilmneb hajuvusdiagrammilt küll kasvav või kahanev, aga mittelineaarne seos, on mõttekas kasutada seose kirjeldamiseks lineaarse korrelatsioonikordaja asemel (või kõrval) Spearmani e astakkorrelatsioonikordajat.

Astakkorrelatsioonikordaja, mis mõõdab kahe arvtunnuse vahelise monotoonse seose tugevust ja suunda, arvutamiseks Excelis valmis vahendid puuduvad. Aga teades, et astakkorrelatsioonikordaja näol on tegu tavalise lineaarse korrelatsioonikordajaga väärtuste astakute ehk järjekorranumbrite vahel, saab seda kordajat arvutada ikkagi ka Excelis.

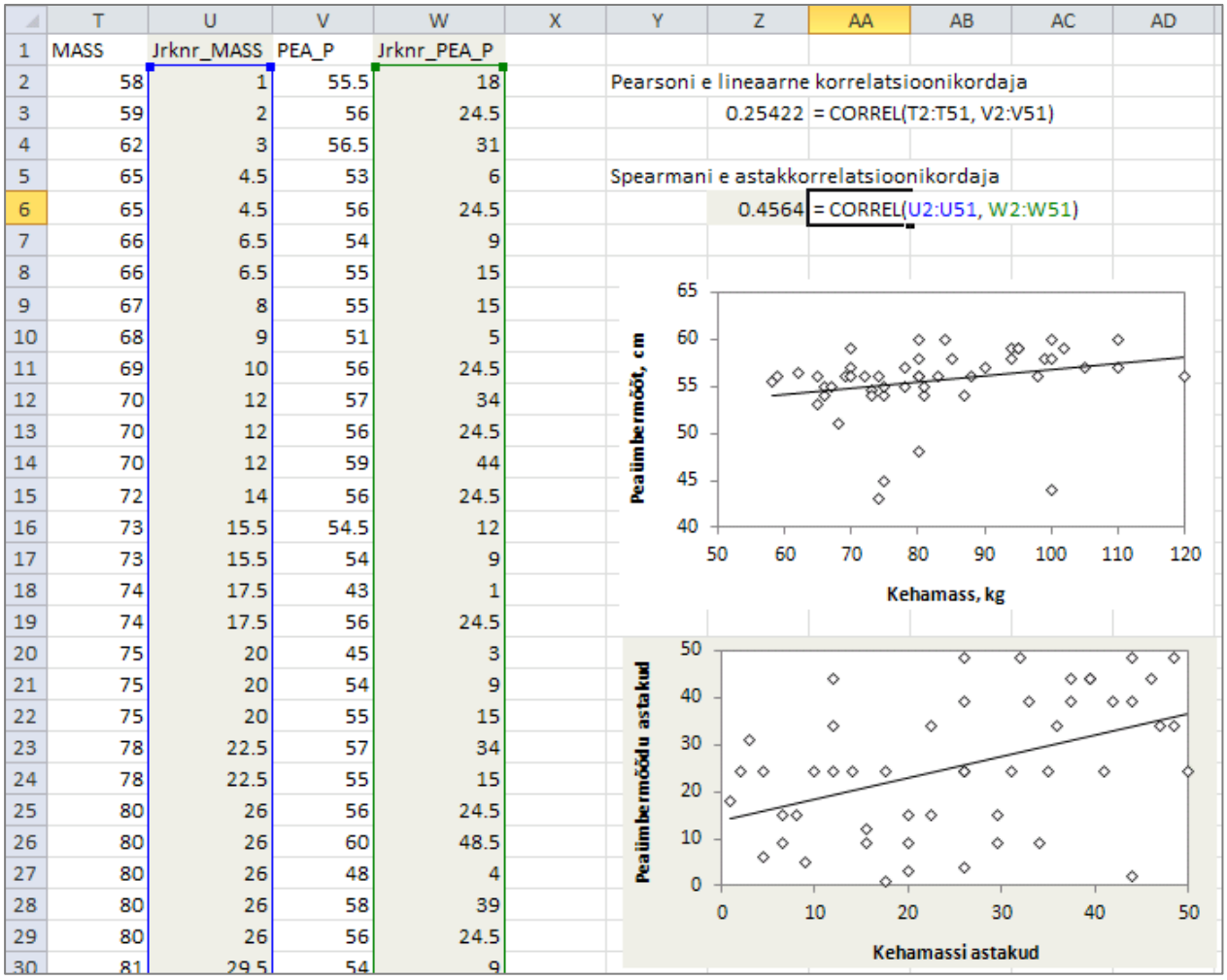
Astakkorrelatsioonikordaja arvutamiseks tuleb

1. arvutada uuritavate tunnuste kõrvale nende astakud funktsiooniga RANK.AVG (Joonis 48),
2. rakendada funktsiooni CORREL (või PEARSON) leitud astakutele (Joonis 49).

NB! Astakute arvutamiseks ei tohi kasutada funktsiooni RANK.EQ (või vanema Exceli funktsiooni RANK), sest erinevalt funktsioonist RANK.AVG, mis võrdsete väärtuste korral võtab nende astakuks järjekorranumbrite keskmise (näiteks kui 4. ja 5. väärtus on võrdsed, saab nende mõlema astakuks 4,5), võtab funktsioon RANK.EQ võrdsete väärtuste astakuks vähima järjekorranumbri (kui 4. ja 5. väärtus on võrdsed, saab nende mõlema astakuks 4). Astakkorrelatsiooni arvutusvalem eeldab, et võrdsete väärtuste astakuks on just nimelt nende järjekorranumbrite keskmine – so funktsiooniga RANK.AVG leitud suurus.

	S	T	U	V
1		MASS	Jrknr_MASS	
2		58	=RANK.AVG(T2,T$2:T$51,1)	
3		59	2	
4		62	3	
5		65	4.5	
6		65	4.5	
7		66	6.5	
8		66	6.5	
9		67	8	

Joonis 48. Noormeeste kehamassi astakute arvutamine funktsiooniga RANK.AVG.



	T	U	V	W
1	MASS	Jrknr_MASS	PEA_P	Jrknr_PEA_P
2	58	1	55.5	18
3	59	2	56	24.5
4	62	3	56.5	31
5	65	4.5	53	6
6	65	4.5	56	24.5
7	66	6.5	54	9
8	66	6.5	55	15
9	67	8	55	15
10	68	9	51	5
11	69	10	56	24.5
12	70	12	57	34
13	70	12	56	24.5
14	70	12	59	44
15	72	14	56	24.5
16	73	15.5	54.5	12
17	73	15.5	54	9
18	74	17.5	43	1
19	74	17.5	56	24.5
20	75	20	45	3
21	75	20	54	9
22	75	20	55	15
23	78	22.5	57	34
24	78	22.5	55	15
25	80	26	56	24.5
26	80	26	60	48.5
27	80	26	48	4
28	80	26	58	39
29	80	26	56	24.5
30	81	29.5	54	9

Joonis 49. Noormeeste kehamassi ja peaümbermõõdu vahelise astakkorrelatsioonikordaja arvutamine (võrdluseks on ära toodud ka lineaarne korrelatsioonikordaja).

Nagu jooniselt 49 näha, on noormeeste kehamassi ja peaümbermõõdu vahel nõrk positiivne lineaarne seos (lineaarne korrelatsioonikordaja $r = 0{,}254$) aga keskmise tugevusega positiivne monotoonne seos (astakkorrelatsioonikordaja $\rho = 0{,}456$). Seose tugevuse sedavõrd suure erinevuse põhjuseks on eelkõige mõned erandlikud väärtused (vt ülemist hajuvusdiagrammi joonisel 49), mis mõjutavad märgatavalt tunnuste vahelist lineaarset seost. Astakute vahelist seost kujutaval hajuvusdiagrammil (alumine diagramm joonisel 49) paiknevad punktid hajusamalt, aga puuduvad teistest hälbivad erandlikud väärtused – kokkuvõttes tähendab see tugevamat seost.

7. Regressioonanalüüs

7.1. Lineaarne regressioonanalüüs protseduuriga *Regression*

Kõige põhjalikuma väljundi lineaarse regressioonanalüüsi tulemustest annab protseduur *Regression* (*Data*-sakk -> *Data Analysis*).

Protseduuri *Regression* sisestusaknas tuleb määrata:

- *Input Y Range* – funktsioontunnuse andmete blokk,
- *Input X Range* – argumenttunnus(t)e andmete blokk (protseduur *Regression* võimaldab teostada ka mitme argumenttunnusega regressioonanalüüsi),
- *Labels* – märgitakse nimede või tähiste olemasolu korral tunnuste blokkide esimeses reas,
- *Constant is Zero* – märgitakse, kui tahetakse kontrollida tunnuste vahelist võrdelist sõltuvust (nõutakse, et $x = 0$ korral ka $y = 0$, st regressioonivõrrandi vabaliige $a = 0$),
- *Confidence Level* – usaldusnivoo parameetrite $(1-\alpha)$-usalduspiiride arvutamiseks, vaikimisi väärtus 95%,
- *Output options* – määratakse tulemuste väljastamise asukoht: samale töölehele (*Output Range*), uuele töölehele (*New Worksheet Ply*) või uude faili (*New Workbook*).

Nende argumentide alusel teostatud regressioonanalüüsi väljund koosneb kolmest tabelist: regressioonimudeli headuse karakteristikud, regressioonivõrrandi dispersioonanalüüs ja regressioonivõrrandi parameetrite hinnangud (Joonis 50).

Täiendavate valikutena võib tellida ka

- prognoosijäägid (*Residuals*),
- standardiseeritud prognoosijäägid (*Standardizised Residuals*),
- prognoosijääkide ja argumenttunnuse hajuvusdiagrammi (*Residuals Plot*),
- funktsioontunnuse ja prognooside graafiku argumenttunnuse suhtes (*Line Fit Plot*),
- funktsioontunnuse väärtuste ja empiiriliste kvantiilide punktdiagrammi punktdiagrammi e tõenäosuspaberi (*Normal Probability Plot*).

NB! Mitmese regressiooni korral peavad argumenttunnused paiknema üksteise kõrval, et neid saaks ette anda ühe pideva andmeblokina. Samuti eeldab protseduur *Regression*, et ette antud funktsioon- ja argumenttunnuste väärtuste blokid ei sisalda puuduvaid väärtuseid, vastasel korral lõpeb protseduuri rakendamine veateatega.

Joonisel 50 on kujutatud noormeeste kehamassi prognoosimine nende pikkuse alusel lineaarse regressioonivõrrandiga.

Esimeses väljundtabelis toodud tulemustest nähtub, et

- tegelike ja prognoositud kehamasside vahel on keskmise tugevusega seos – mitmene korrelatsioonikordaja (*Multiple R*), mis ongi uuritava tunnuse ja tema prognoositud väärtuste vaheline lineaarne korrelatsioonikordaja, $R = 0{,}458$,
- sobitatud mudel kirjeldab ära 21,0% noormeeste kehamasside tegelikust varieeruvusest – determinatsioonikordaja (*R Square*) $R^2 = 0{,}21$,
- keskmiselt osutub prognoositud kehamass valeks 13,1 kg võrra – mudeli standarviga (*Standard Error*), mis arvutatakse kui prognoosijääkide standardhälve, on 13,1.

Teises tabelis toodud tulemustest nähtub, et kuigi noormeeste kehamassi prognoos nende pikkuse järgi ei ole eriti täpne, on leitud regressioonivõrrand tervikuna siiski statistiliselt oluline ($p < 0{,}001$; veerg *Significance F*). St, et kasutades noormeeste kehamassi prognoosimiseks lineaarset funktsiooni nende pikkusest on tulemus statistiliselt oluliselt täpsem võrreldes tõdemusega, et kõik noormehed kaaluvad keskmiselt ühepalju.

Väljundi kolmandas tabelis on toodud regressioonivõrrandi kordajate hinnangud (veerus *Coefficients*), nende standardvead (*Standard Error*; näitavad, kui palju keskmiselt võib kordajate hinnang varieeruda), p-väärtused (*P-value*; testitakse hüpoteesi kordaja erinevusest nullist) ja 95%-lised usalduspiirid (*Lower 95%* ja *Upper 95%*).

Noormeeste kehamassi prognoosimisel kasutatav regressionivõrrand on vastavalt kordajate hinnangutele kirja pandav kujul

$$\text{Kehamass} = -97{,}0 + 0{,}978 * \text{Pikkus}.$$

- Sellest võrrandist tuleneb, et pikkuse suurenemisega 1 cm võrra suureneb keskmiselt ka noormeest kehamass 1 kg (täpsemelt 0,978 kg) võrra.
- Samuti saab arvutada, et 180 cm pikkune esimese kursuse noormees peaks hinnanguliselt kaaluma $-97{,}0 + 0{,}978 * 180 = 79{,}1$ kg.

Märkus. See, et võrrandi vabaliikmele vastav p-väärtus väljundi kolmandas tabelis on suurem kui 0,05 ($p = 0{,}059$), antud juhul ohu märk ei ole. Juhul, kui argumenttunnus ei saa reaalselt omandada väärtust 0 (ja tudeng ei saa kaaluda 0 kg), ei ole regressioonivõrrandi vabaliikmel sisulist tähendust, tegu on lihtsalt matemaatilise prognoosivõrrandi loomuliku osaga, mille kohta hüpoteeside testimine on mõttetu.

Joonistel 51 ja 52 on kujutatud protseduuri *Regression* lisavalikute tulemusena väljastatud tabelid ja joonised.

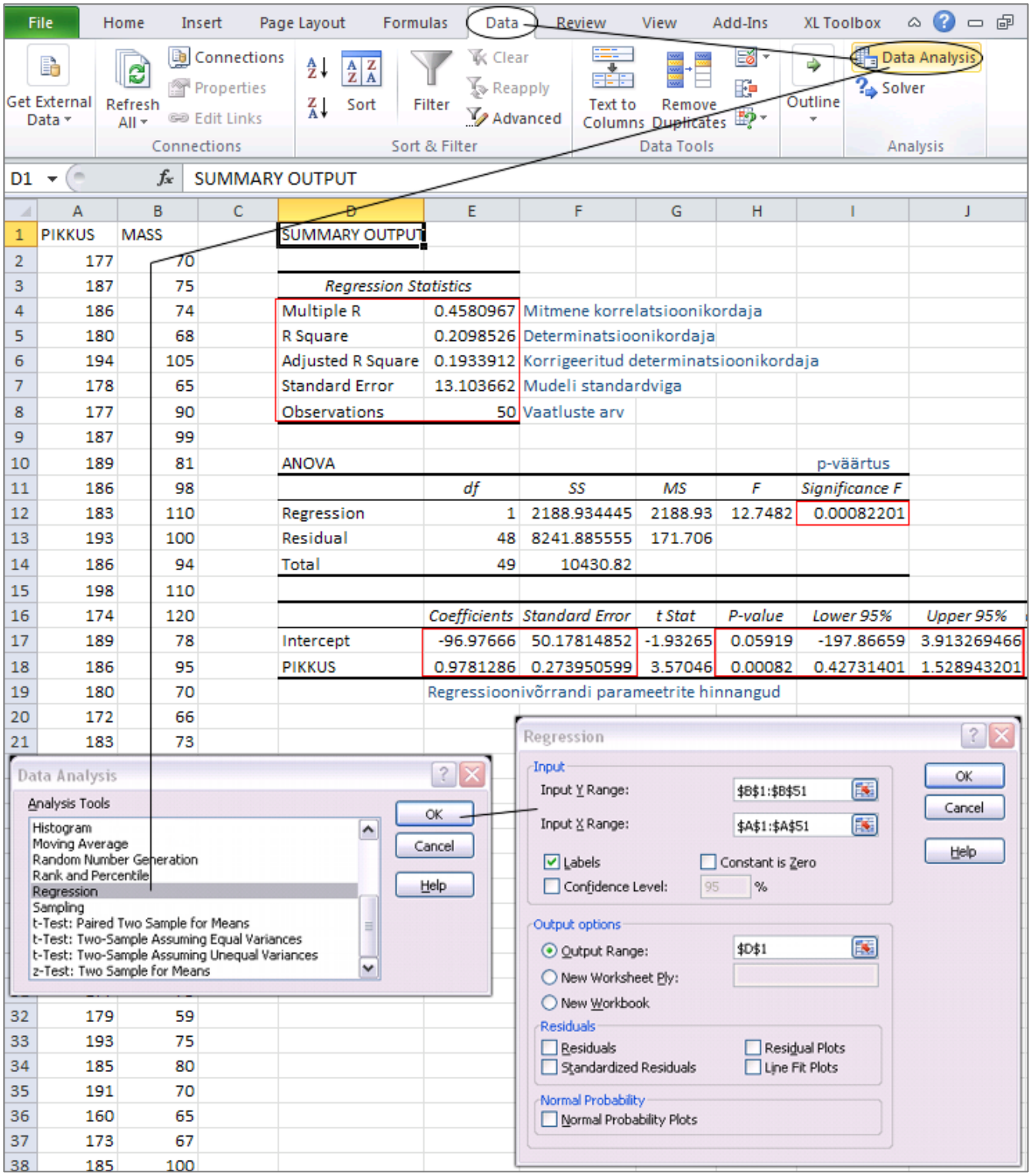
- Ükskõik milline lisavalikutest *Residuals*, *Residual Plots* või *Line Fit Plots* annab tulemuseks väljundtabeli, milles on üks rida iga andmetabeli rea tarvis ning selles on kirjas vaatluse järjekorranumber (*Observations*), prognoositud väärtus (*Predicted …*) ja prognoosijääk (*Residuals*). Prognoosijääk on seejuures arvutatud kui tegelik väärtus miinus prognoositud väärtus.
- Lisavalik *Standardizised Residuals* lisab prognoositud väärtuste ja prognoosijääkide tabelile täiendava, standardiseeritud jääkide veeru.
- Lisavalik *Normal Probability Plot* lisab väljundile tabeli, mis sisaldab kasvavalt sorteeritult kõiki uuritava tunnuse väärtuseid ja neile vastavaid protsendipunkte (emiirilisi kvantiile).

Joonistest annab

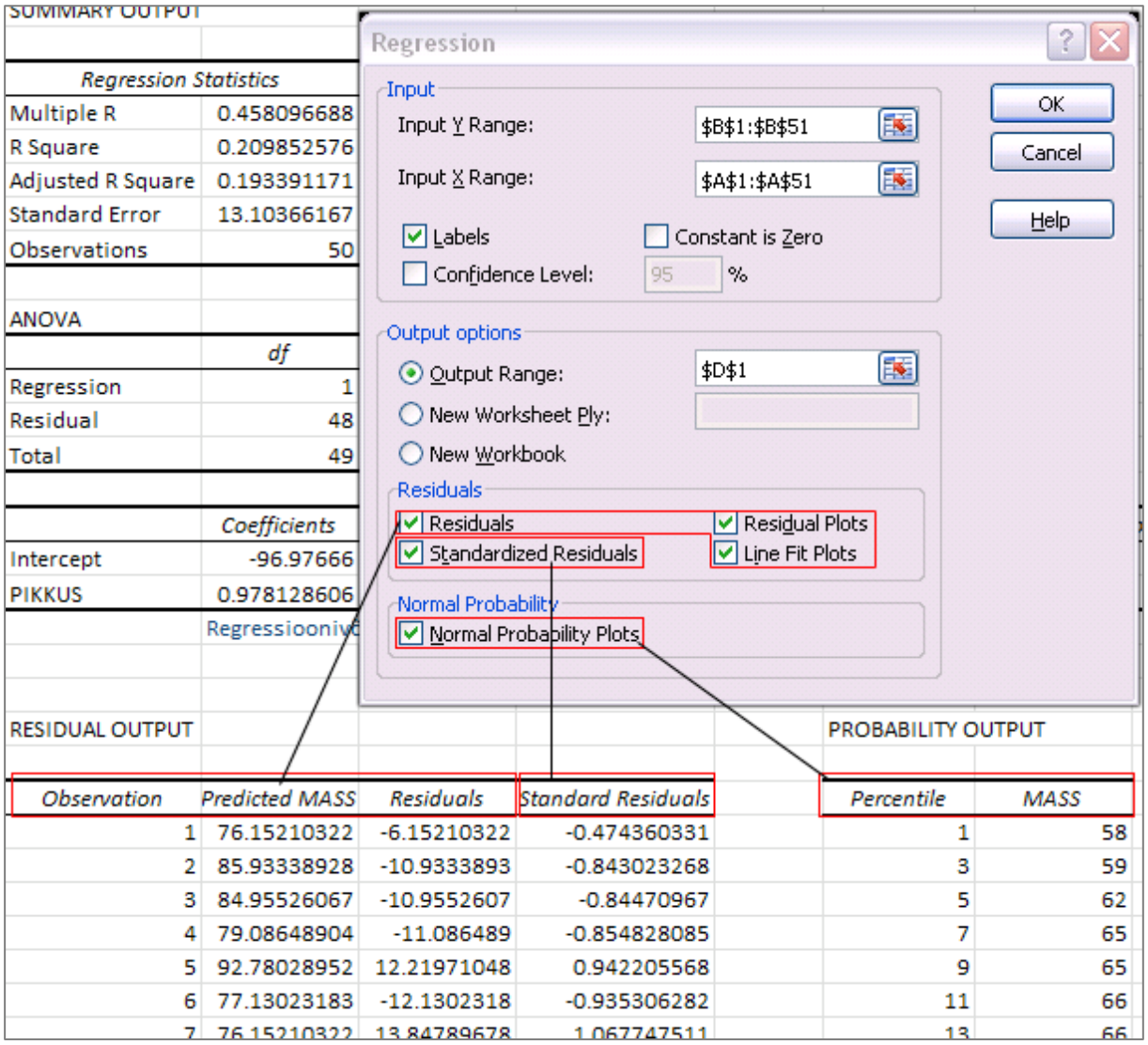
- lisavalik *Residual Plots* tulemuseks prognoosijääkide ja argumenttunnuse hajuvusdiagrammi – kui regressioonivõrrand on sobiv, peaksid punktid sellel graafikul paiknema juhuslikult, ühtlaselt hajutatud punktiparvena,
- *Line Fit Plots* funktsioon- ja argumenttunnuse hajuvusdiagrammi, kuhu on täiendava andmeseeriana kantud prognoositud väärtused,
- *Normal Probability Plot* funktsioontunnuse väärtuste ja empiiriliste kvantiilide punktdiagrammi e tõenäosuspaberi – kui funktsioontunnus on normaaljaotusega, peaksid punktid sellel diagrammil paiknema diagonaalsel sirgel.

NB! Mitmese regressioonanalüüsi puhul konstrueerib Excel valiku *Residual Plots* tulemusena eraldi iga argumenttunnuse ja prognoosijääkide hajuvusdiagrammid, valiku *Line Fit Plots* tulemusena aga eraldi iga argumenttunnuse ja prognoositud väärtuste hajuvusdiagrammid.

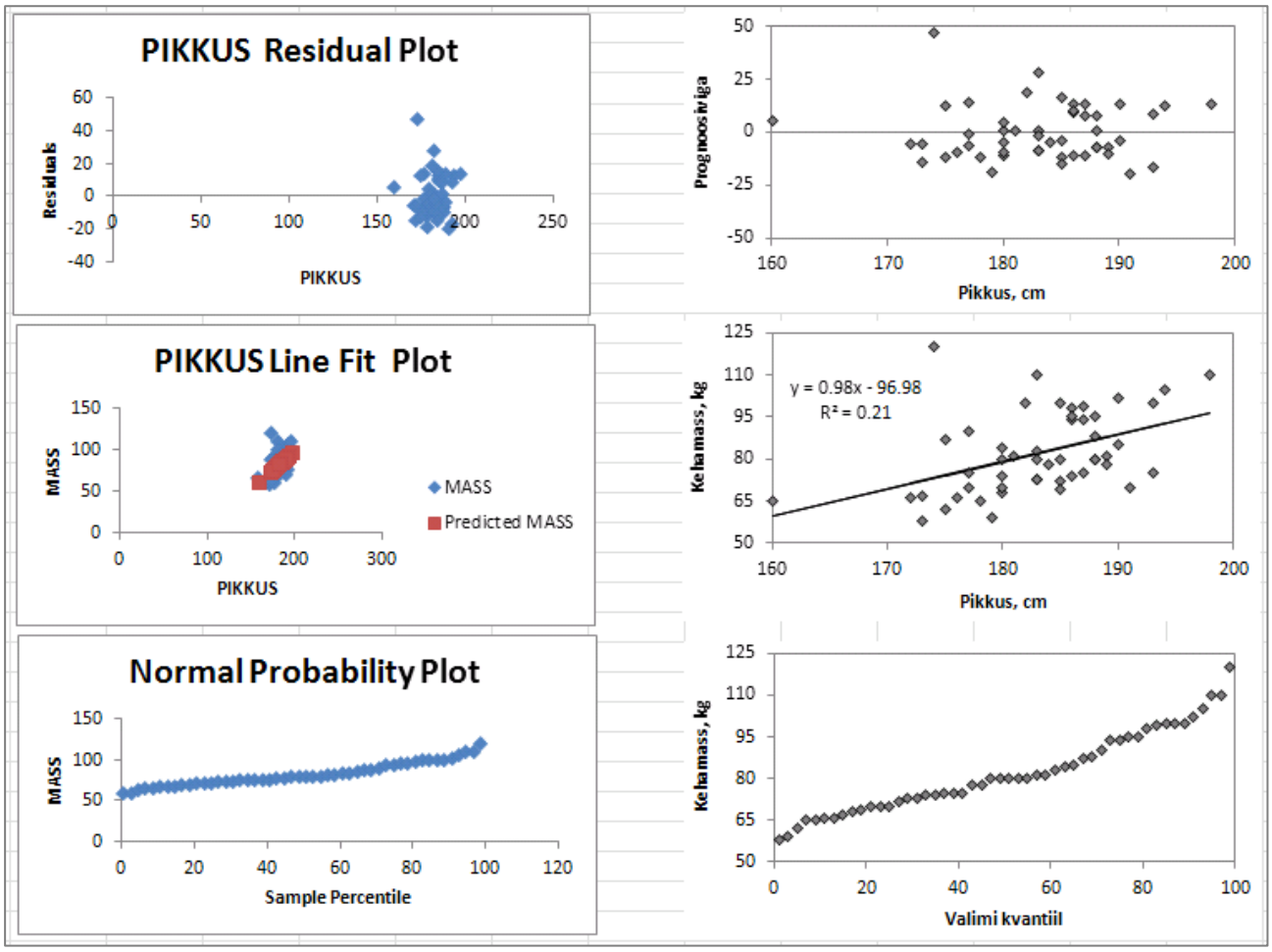
Et protseduuri *Regression* lisavalikute tulemusena saadud tabelid on üsna mahukad ja joonised vajavad üksjagu lisatööd viimaks neid kasutatavale kujule, on neid mõtet tellida üksnes siis, kui tõepoolest on plaanis regressioonivõrrandi põhjalikum diagnostika.



Joonis 50. Noormeeste kehamassi prognoosimine nende pikkuse alusel – protseduuri *Regression* tellimisaken ja vaikimisi väljastatavad tulemused.



Joonis 51. Noormeeste kehamassi prognoosimine nende pikkuse alusel – protseduuri *Regression* lisatulemused.



Joonis 52. Protseduuri *Regression* poolt väljastatavad diagrammid (vasakul pool) ning nende sobivamale kujule viidud variandid (paremal pool).

7.2. Regressioonanalüüs graafiliselt

Kui huvi pakub vaid kahe arvtunnuse vaheline seos, on seda mugav esitada punkt- e hajuvusdiagrammina, millele Excel võimaldab peale sobitada erinevate matemaatiliste funktsioonide abil hinnatud prognoosikõveraid – st teostada lineaarset ja mittelineaarset regressioonanalüüsi.

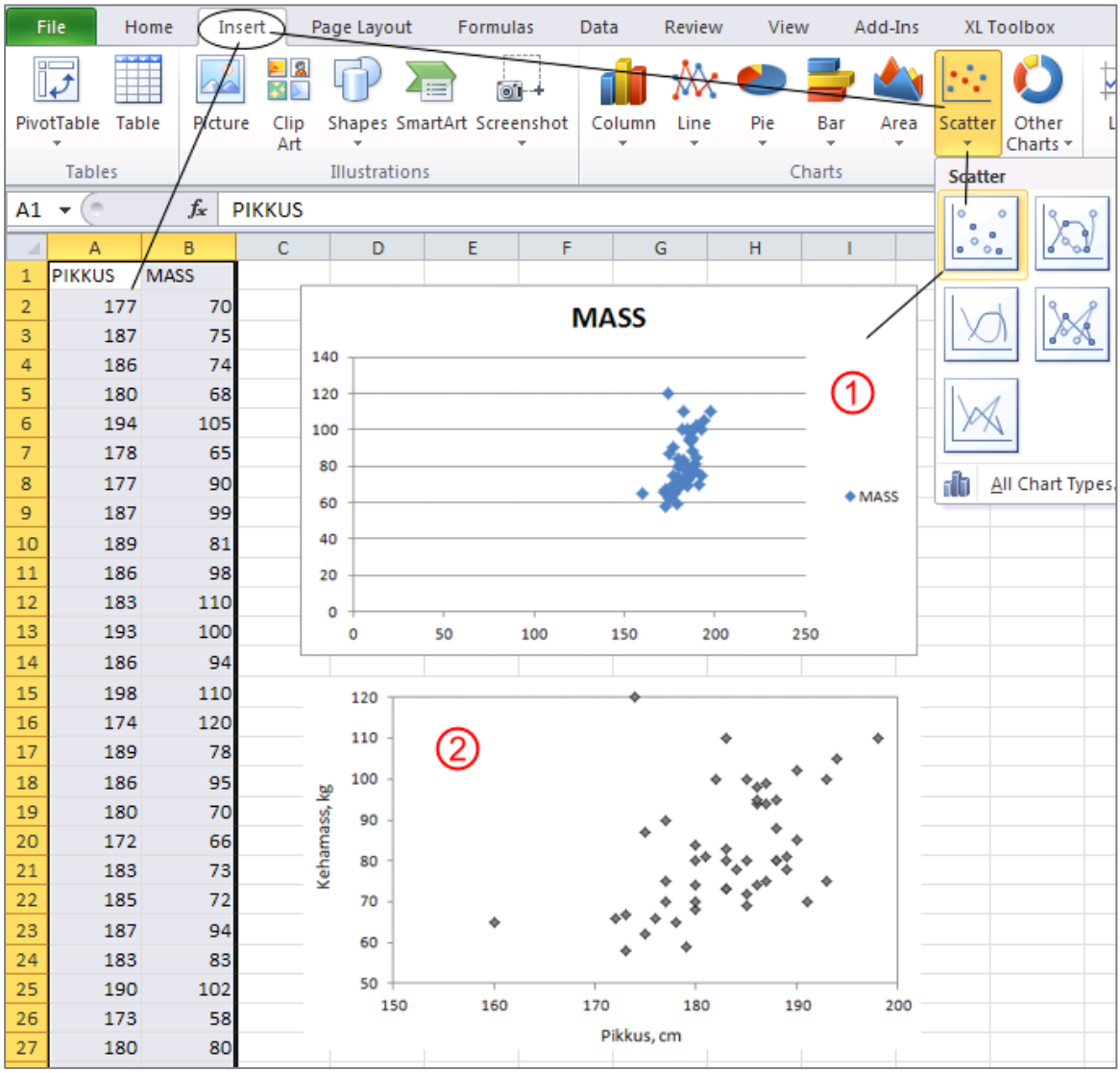
Regressioonanalüüsi teostamiseks graafiliselt tuleb (vt ka Joonised 53A ja 53B)

1. konstrueerida uuritavate tunnuste vaheline hajuvus- e punktdiagramm,
2. kujundada punktdiagramm sobivaks (enamasti tähendab see telgede ulatuse sobitamist andmetega, ruudujoonte ära kaotamist jmt),

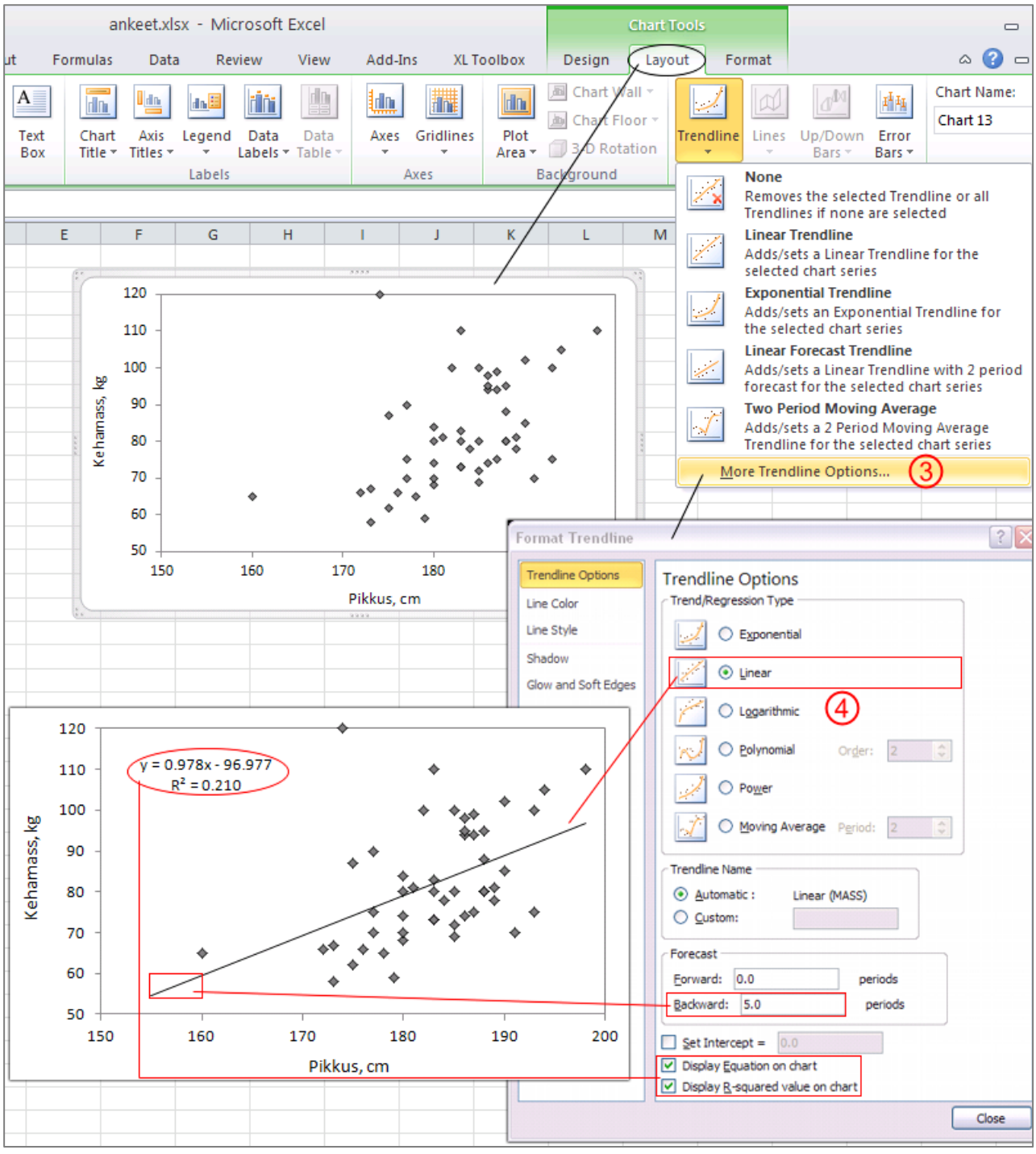
   NB! Excel paigutab joonise y-teljele alati andmetabelis vasakul pool paikneva tunnuse ja teostab ka regressioonanalüüsi, prognoosides y-teljel paikneva tunnuse väärtuseid – seega, vajadusel tuleb diagrammil *x*- ja *y*-teljel kuvatavad tunnused ära vahetada.
3. valida *Chart Tools*-sakilt alamenüü *Layout* ja sealt käsk *Trendline -> More Trendline Options…* (valik *More Trendline Options…* on Exceli vaikimisi pakutavate trendijoonte asemel mõttekam valik, kuna võimaldab koheselt määrata mitmeid trendijoone ja selle aluseks oleva funktsiooniga seotud lisaaspekte, mida vastasel juhul tuleks hiljem täiendavate käskudega muutma hakata),
4. määrata trendijoone tüüp (*Trend/Regression Type*) ja

- anda soovi korral trendijoonele nimi (*Trendline Name*; mõttekas, kui soovite näidata trendijoont joonise legendil),
- pikendada soovi korral trendijoont argumenttunnuse väärtuste piirkonnast väljapoole (*Forecast*; argumenttunnuse ühikutes),
- käskida trendijoonel läbida koordinaatide alguspunkti (vabaliige on siis null) või lõigata y-telge mõnes teises fikseeritud kohas (sisuliselt näitab see funktsioontunnuse *y* väärtust, kui argumenttunnus $x = 0$; *Set Intercept* =),
- tellida täiendavalt graafikule regressioonivõrrandi avaldis (*Display Equation on chart*) ja prognoosi headust kirjeldav determinatsioonikordaja (*Display R-squared value on chart*).

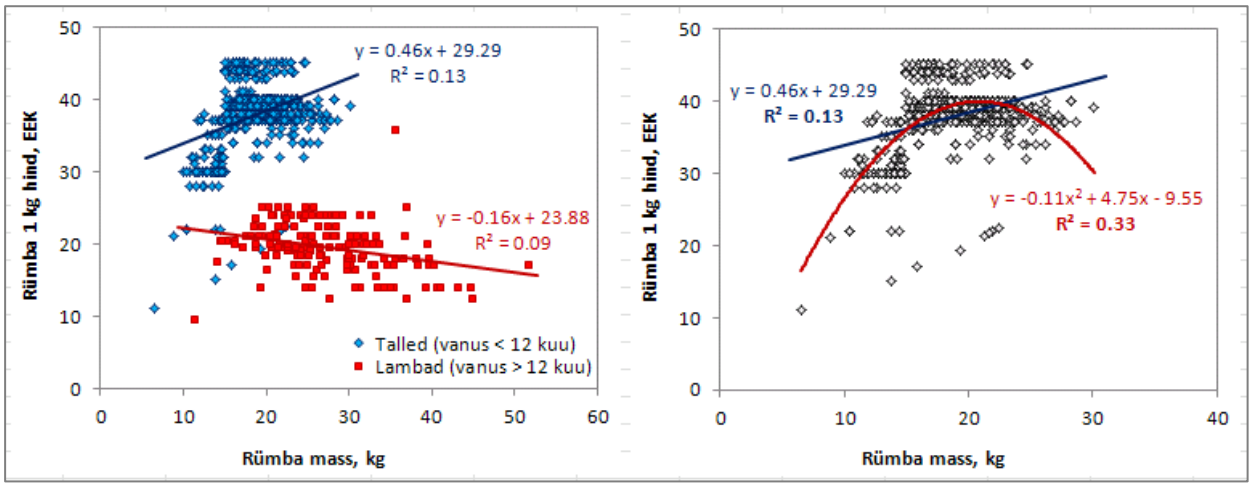
Regressioonanalüüsi graafilise teostamise peamine puudus statistiliste analüüside kontekstis on prognoosivõrrandi statistilist olulisust näitava p-väärtuse mitte arvutamine. Plussiks on aga võimalus kõrvutada erinevaid mudeleid nii visuaalselt kui ka võrrandi ja eriti determinatsioonikordaja R^2 alusel (Joonis 54).



Joonis 53A. Noormeeste kehamassi prognoosimine nende pikkuse alusel – punktdiagramm.



Joonis 53B. Noormeeste kehamassi prognoosimine nende pikkuse alusel – punktdiagrammile regressioonisirge ja regressioonivõrrandi lisamine.



Joonis 54. Näiteid regressioonanalüüsi graafilisest teostamisest – seoste võrdlemine.

7.3. Regressioonanalüüs funktsioonide abil

Funktsioonide abil on Excelis võimalik teostada nii lihtsat kui ka mitmest lineaarset regressioonanalüüsi ning sobitada andmetele ka eksponentfunktsiooni. Mõlemal juhul on võimalik nii funktsiooni parameetrite ja prognoosi headuse hindamine kui ka funktsioon-tunnuse väärtuste prognoosimine ette antud argumenttunnuste väärtuste korral.

Lihtne lineaarne regressioonanalüüs

Lihtsa lineaarse regressioonivõrrandi

$$y = a + bx,$$

kus y on uuritav ehk funktsioontunnus ja x on argumenttunnus, parameetrite a ja b hindamiseks Excelis on lihtsaim viis kasutada vastavalt funktsioone INTERCEPT ja SLOPE. Prognoosi headust kirjeldav determinatsioonikordaja R^2 on hinnatav funktsiooniga RSQ ja mudeli standardviga funktsiooniga STEYX. Kõigi nende funktsioonide puhul tuleb ühte moodi ette anda

- funktsioontunnuse *y* andmete blokk (*Known_y's*) ja
- argumenttunnuse *x* andmete blokk (*Known_x's*).

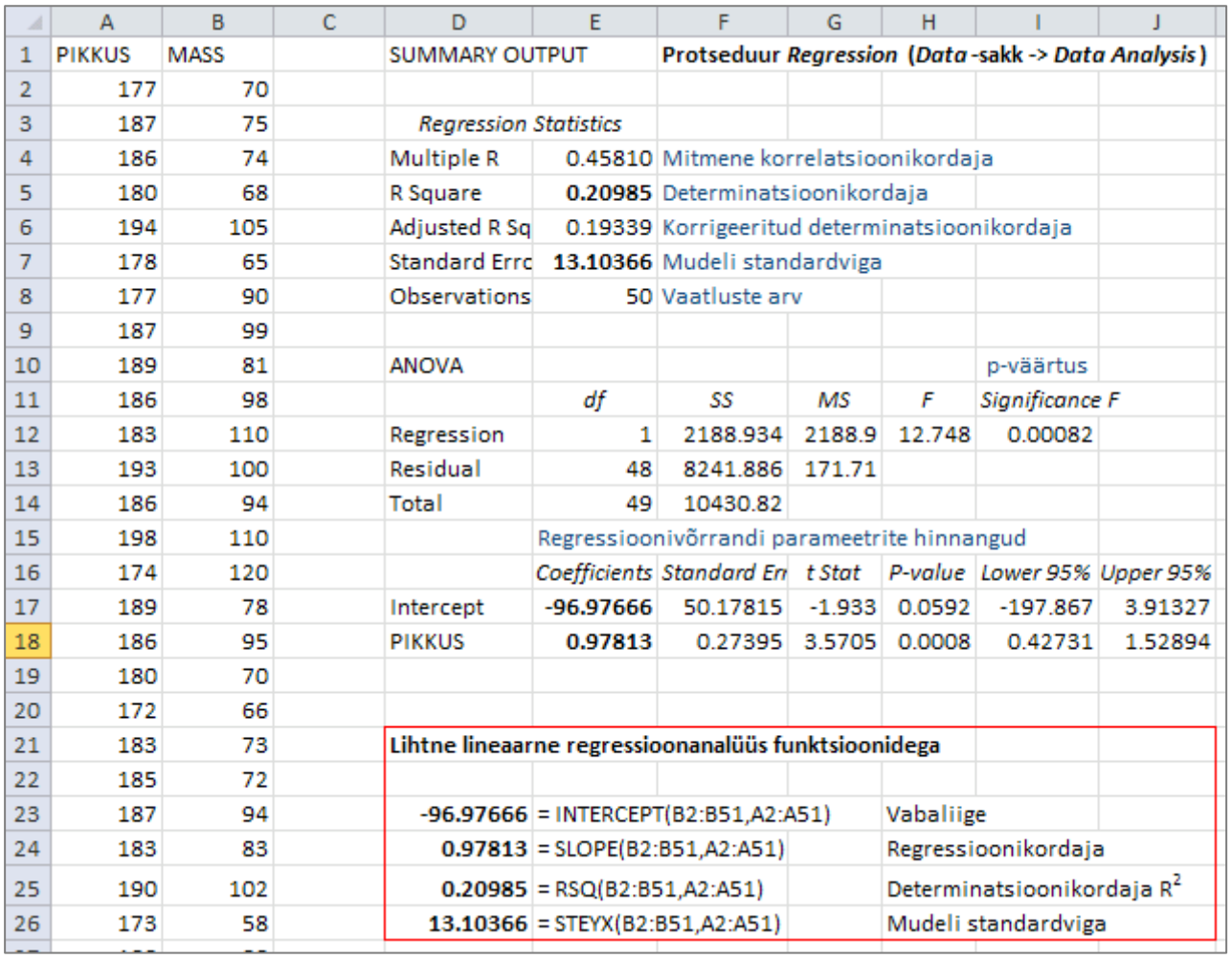
Joonisel 55 on näidatud lihtsa lineaarse regressioonanalüüsi teostamist funktsioonide INTERCEPT, SLOPE, RSQ ja STEYX abil prognoosimaks noormeeste kehamassi nende pikkuse alusel.

Tulemused on identsed protseduuriga *Regression* arvutatutele:

- noormeeste kehamass on nende pikkuse abil prognoostav valemist

  Kehamass = -97,0 + 0,978*Pikkus,

- kusjuures antud mudel kirjeldab ära 21,0% noormeeste kehamasside tegelikust varieeruvusest (R^2 = 0,21) ja keskmiselt osutub prognoositud kehamass valeks 13,1 kg võrra (*SEM* = 13,1).



	A	B	C	D	E	F	G	H	I	J
1	PIKKUS	MASS		SUMMARY OUTPUT		**Protseduur *Regression* (*Data*-sakk -> *Data Analysis*)**				
2	177	70								
3	187	75		*Regression Statistics*						
4	186	74		Multiple R	0.45810	Mitmene korrelatsioonikordaja				
5	180	68		R Square	**0.20985**	Determinatsioonikordaja				
6	194	105		Adjusted R Sq	0.19339	Korrigeeritud determinatsioonikordaja				
7	178	65		Standard Errc	**13.10366**	Mudeli standardviga				
8	177	90		Observations	50	Vaatluste arv				
9	187	99								
10	189	81		ANOVA					p-väärtus	
11	186	98			*df*	*SS*	*MS*	*F*	*Significance F*	
12	183	110		Regression	1	2188.934	2188.9	12.748	0.00082	
13	193	100		Residual	48	8241.886	171.71			
14	186	94		Total	49	10430.82				
15	198	110			Regressioonivõrrandi parameetrite hinnangud					
16	174	120			*Coefficients*	*Standard Er*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	189	78		Intercept	**-96.97666**	50.17815	-1.933	0.0592	-197.867	3.91327
18	186	95		PIKKUS	**0.97813**	0.27395	3.5705	0.0008	0.42731	1.52894
19	180	70								
20	172	66								
21	183	73		**Lihtne lineaarne regressioonanalüüs funktsioonidega**						
22	185	72								
23	187	94		**-96.97666**	= INTERCEPT(B2:B51,A2:A51)			Vabaliige		
24	183	83		**0.97813**	= SLOPE(B2:B51,A2:A51)			Regressioonikordaja		
25	190	102		**0.20985**	= RSQ(B2:B51,A2:A51)			Determinatsioonikordaja R^2		
26	173	58		**13.10366**	= STEYX(B2:B51,A2:A51)			Mudeli standardviga		

Joonis 55. Lihtne lineaarne regressioonanalüüs funktsioonidega INTERCEPT, SLOPE, RSQ ja STEYX prognoosimaks noormeeste kehamassi nende pikkuse alusel; võrdluseks on ära toodud ka protseduuriga *Regression* teostatud sama analüüsi tulemused, kusjuures nii funktsioonidega kui ka protseduuriga *Regression* arvutatavad väärtused on esitatud paksus kirjas.

Mitmene lineaarne regressioonanalüüs

Mahukaima väljundi annab tulemuseks funktsioon LINEST, mis võimaldab teostada nii lihtsat kui ka mitmest lineaarset regressioonanalüüsi ning sobitada andmetele ka argumenttunnuse funktsioone sisaldavaid mudeleid (näiteks kõrgema astme polünoome). Funktsioon LINEST on massiivifunktsioon (väljundiks on väärtuste tabel, mitte üksikväärtus), mille tulemuseks on sõltuvalt funktsiooni argumentidest kas vaid regressioonikordajad või regressioonikordajad pluss hulk teisi regressioonanalüüsiga kaasnevaid karakteristikuid.

Funktsioon LINEST rakendamiseks tuleb (Joonis 56)

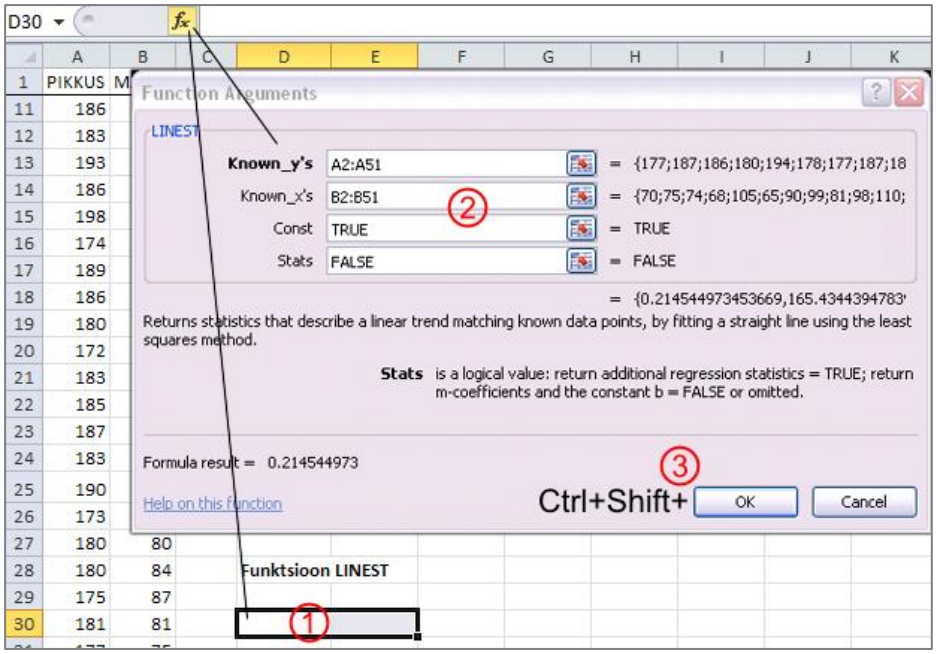
1. selekteerida Exceli töölehel väljundtabeli jagu lahterid (kui palju täpselt, sõltub funktsiooni LINEST kahest viimasest argumendist);
2. anda ette funktsiooni argumendid:
   - funktsioontunnuse *y* andmete blokk (*Known_y's*),
   - argumenttunnus(t)e *x* andmete blokk (*Known_x's*),
   - argument *Const* väärtustega TRUE (vaikimisi väärtus, hinnatav mudel sisaldab vabaliiget) või FALSE (hinnatav mudel ei sisalda vabaliiget),

- argument *Stats* väärtustega FALSE
  - vaikimisi väärtus, arvutatakse ja väljastatakse vaid mudeli parameetrite – so vabaliikme (argumendi *Const* = TRUE korral) ja regressioonikordja(te) – hinnagud,

  või TRUE
  - lisaks mudeli parameetrite hinnangutele väljastatakse ka
  - parameetrite hinnangute standardvead,
  - regressiooniseose headust kirjeldavad determinatsioonikordaja R^2 ja mudeli standardvea *SEM* väärtused,
  - F-statistiku (e F-suhte) väärtus, F-suhte nimetaja vabadusastmete arv (Df_2) ning nii mudelile kui ka mudeli jäägile vastavad ruutude summade väärtused (SS_{mudel} ja $SS_{jääk}$) mudeli dispersioonanalüüsi tabelis;

3. vajutada **Ctrl**+**Shift** ja **Enter** (või OK).

NB! Mitmese regressioonanalüüsi korral peavad argumenttunnused paiknema üksteise kõrval, et neid saaks ette anda ühe pideva andmeblokina. Samuti eeldab funktsioon LINEST, et ette antud funktsioon- ja argumenttunnuste väärtuste blokid ei sisalda puuduvaid väärtuseid, vastasel korral lõpeb funktsiooni rakendamine veateatega.



Joonis 56. Funktsiooni LINEST rakendamine prognoosimaks noormeeste kehamassi nende pikkuse alusel lineaarse regressioonanalüüsi abil.

Väljundtabeli, mille jagu lahtreid tuleb enne funktsiooni LINEST rakendamist blokki võtta,

- **ridade arv** on üks argumendi *Stats* = FALSE ja viis argumendi *Stats* = TRUE korral,
- **veergude arv** on võrdne mudeli parameetrite arvuga, st vabaliige pluss argumentide arv.

St, et näiteks mitmese regressioonivõrrandi

$$y = a + b_1*x_1 + b_2*x_2 + \ldots + b_m*x_m$$

korral on funktsiooni LINEST laiendatud väljundtabelis viis rida ja m+1 veergu (m on argumenttunnuste arv):

b_m	b_{m-1}	…	b_1	a
se(b_m)	se(b_{m-1})	…	se(b_1)	se(a)
R^2	SEM			
F	Df_2			
SS_{mudel}	$SS_{jääk}$			

NB! Regressioonivõrrandi parameetrid paiknevad funktsiooni LINEST väljundis nö tagurpidi – vasakult poolt esimesel kohal on viimane regressioonikordaja ning kõige viimasel kohal mudeli vabaliige (või selle puudumisel esimene regressioonikordaja).

Joonisel 57 on esitatud funktsiooni LINEST süntaks ja tulemused prognoosimaks noormeeste kehamassi nende pikkuse alusel lineaarse regressioonanalüüsi abil. Esitatud on nii funktsiooni LINEST vaikimisi väljund, mis sisaldab üksnes regressioonivõrrandi parameetreid, kui ka laiendatud väljund. Võrdluseks on esitatud ka sama ülesande lahendamisel protseduuriga *Regression* saadud tabelid ning vaid üksikväärtusi väljastavate funktsioonide INTERCEPT, SLOPE, RSQ ja STEYX tulemused.

Nii nagu varemalt kirjeldatud protseduuri *Regression* (pt. 7.1) ning antud punkti alguses kirjeldatud funktsioonide INTERCEPT, SLOPE, RSQ ja STEYX tulemuste alusel, saab nüüdki järeldada, et

- noormeeste kehamass on nende pikkuse abil prognoostav valemist

  Kehamass = -97,0 + 0,978*Pikkus,

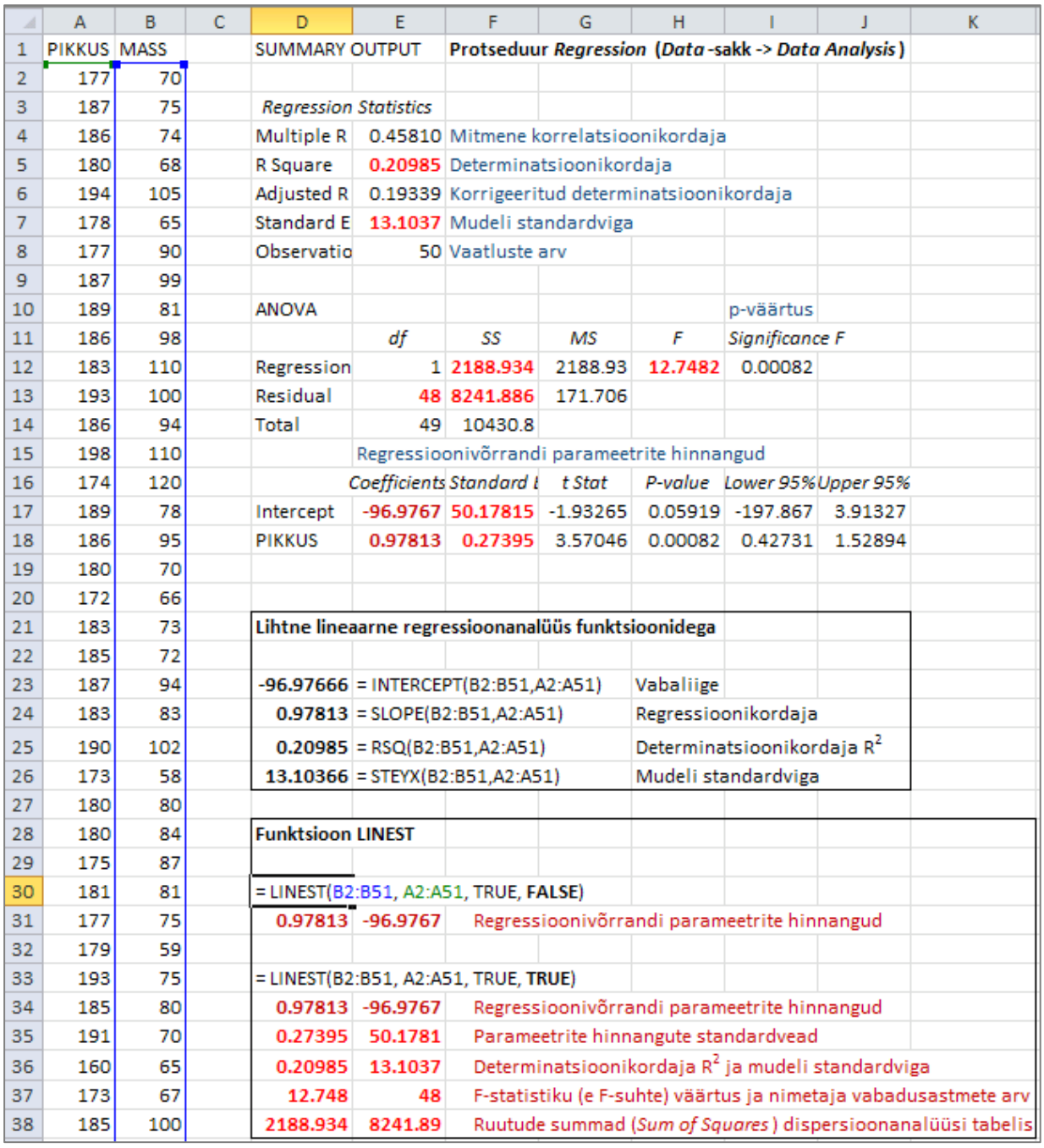
- kusjuures mudel kirjeldab ära 21,0% noormeeste kehamasside tegelikust varieeruvusest (R^2 = 0,21) ja keskmiselt osutub prognoositud kehamass valeks 13,1 kg võrra (SEM = 13,1).

Aga lisaks saab funktsiooni LINEST laiendatud väljundis toodud suuruste alusel testida ka hüpoteese nii mudeli kui terviku statistilise olulisuse kui ka üksikute liikmete statistilise olulisuse kohta.

Mudeli statistilise olulisuse testimine baseerub F-statistikul (ehk F-suhtel), mis on null-hüpoteesi kehtides F-jaotusega parameetritega Df_1 ja Df_2. F-suhte väärtus ja selle nimetajale vastav vabadusastmete arv Df_2 sisalduvad ka funktsiooni LINEST laiendatud väljundis (Joonis 58). F-suhte lugeja vabadusastmete arv Df_1 on juhul, kui mudel sisaldab vabaliiget, leitav valemist $Df_1 = n - Df_2 - 1$, vabaliikme puudumisel (funktsiooni LINEST argument *Const* = FALSE) aga valemist $Df_1 = n - Df_2$, suurus n tähistab valimi mahtu. Nn mudeli p-väärtus on arvutatav funktsiooniga F.DIST.RT, mille esimeseks argumendiks on F-statistiku väärtus ning teiseks ja kolmandaks argumendiks vastavalt F-jaotuse parameetrid Df_1 ja Df_2.

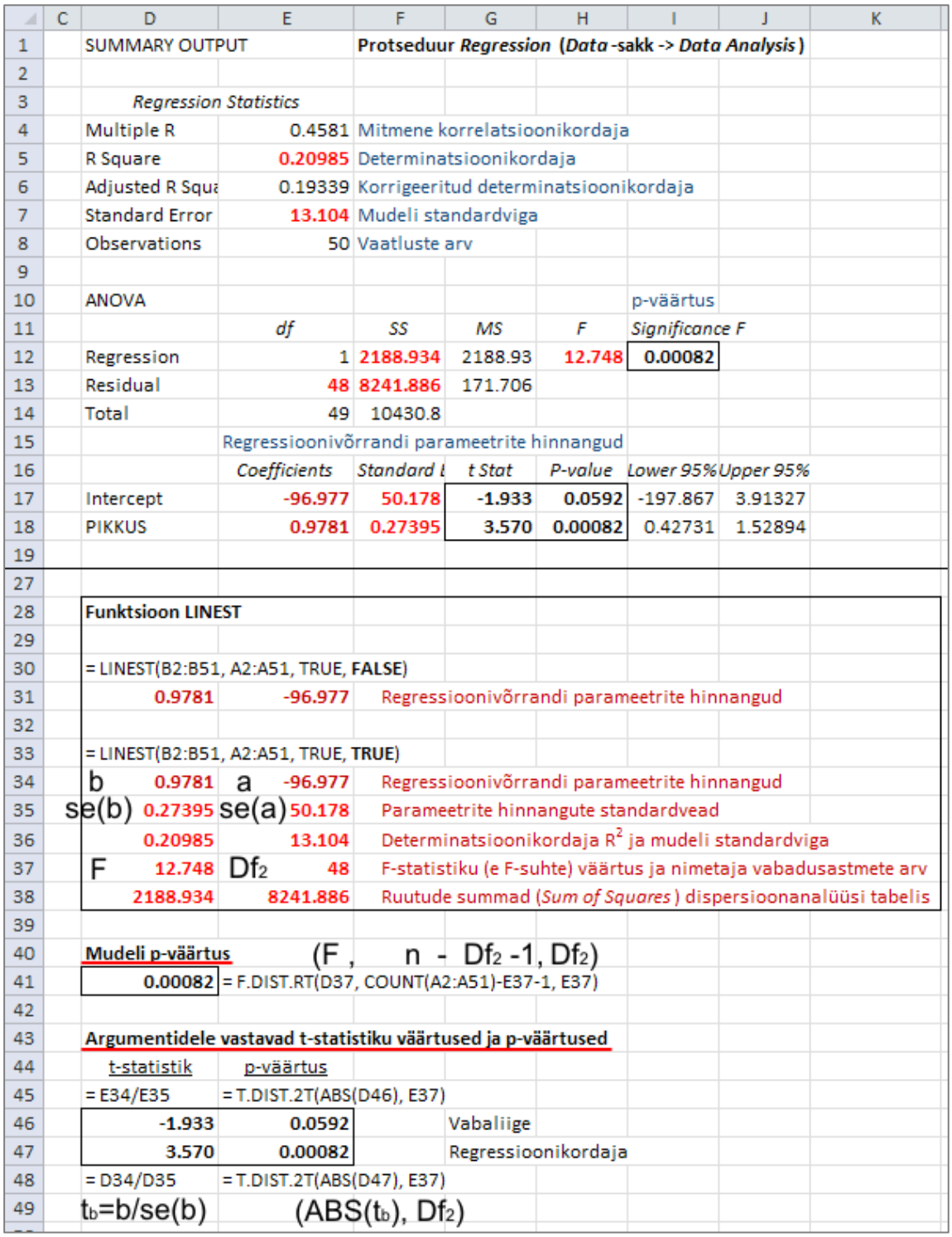
Iga üksiku regressioonivõrrandi parameetri statistilise olulisuse testimine baseerub t-statistikul, mis on arvutatav kui parameetri hinnangu suhe oma standardveasse (ja need suurused sisalduvad funktsiooni LINEST väljundis). Parameetri statistilist olulisust näitav p-väärtus on arvutatav funktsiooniga T.DIST.2T, mille esimeseks argumendiks on t-statistiku absoluutväärtus ning teiseks argumendiks ka terve mudeli statistilise olulisuse testimisel kasutatud funktsiooni LINEST väljundis sisalduv suurus Df_2.

Nagu näha jooniselt 58, on funktsiooni LINEST väljundi baasil arvutatud p-väärtused identsed protseduuri *Regression* poolt väljastatutega.



	A	B	C	D	E	F	G	H	I	J	K
1	PIKKUS	MASS		SUMMARY OUTPUT		Protseduur *Regression* (*Data*-sakk -> *Data Analysis*)					
2	177	70									
3	187	75		*Regression Statistics*							
4	186	74		Multiple R	0.45810	Mitmene korrelatsioonikordaja					
5	180	68		R Square	0.20985	Determinatsioonikordaja					
6	194	105		Adjusted R	0.19339	Korrigeeritud determinatsioonikordaja					
7	178	65		Standard E	13.1037	Mudeli standardviga					
8	177	90		Observatio	50	Vaatluste arv					
9	187	99									
10	189	81		ANOVA					p-väärtus		
11	186	98			*df*	*SS*	*MS*	*F*	*Significance F*		
12	183	110		Regression	1	2188.934	2188.93	12.7482	0.00082		
13	193	100		Residual	48	8241.886	171.706				
14	186	94		Total	49	10430.8					
15	198	110			Regressioonivõrrandi parameetrite hinnangud						
16	174	120			*Coefficients*	*Standard E*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	
17	189	78		Intercept	-96.9767	50.17815	-1.93265	0.05919	-197.867	3.91327	
18	186	95		PIKKUS	0.97813	0.27395	3.57046	0.00082	0.42731	1.52894	
19	180	70									
20	172	66									
21	183	73		**Lihtne lineaarne regressioonanalüüs funktsioonidega**							
22	185	72									
23	187	94		**-96.97666**	= INTERCEPT(B2:B51,A2:A51)			Vabaliige			
24	183	83		**0.97813**	= SLOPE(B2:B51,A2:A51)			Regressioonikordaja			
25	190	102		**0.20985**	= RSQ(B2:B51,A2:A51)			Determinatsioonikordaja R^2			
26	173	58		**13.10366**	= STEYX(B2:B51,A2:A51)			Mudeli standardviga			
27	180	80									
28	180	84		**Funktsioon LINEST**							
29	175	87									
30	181	81		= LINEST(B2:B51, A2:A51, TRUE, **FALSE**)							
31	177	75		**0.97813**	**-96.9767**	Regressioonivõrrandi parameetrite hinnangud					
32	179	59									
33	193	75		= LINEST(B2:B51, A2:A51, TRUE, **TRUE**)							
34	185	80		**0.97813**	**-96.9767**	Regressioonivõrrandi parameetrite hinnangud					
35	191	70		**0.27395**	**50.1781**	Parameetrite hinnangute standardvead					
36	160	65		**0.20985**	**13.1037**	Determinatsioonikordaja R^2 ja mudeli standardviga					
37	173	67		**12.748**	**48**	F-statistiku (e F-suhte) väärtus ja nimetaja vabadusastmete arv					
38	185	100		**2188.934**	**8241.89**	Ruutude summad (*Sum of Squares*) dispersioonanalüüsi tabelis					

Joonisel 57. Funktsiooni LINEST süntaks ning vaikimisi produtseeritav ja laiendatud väljund prognoosimaks noormeeste kehamassi nende pikkuse alusel lineaarse regressioonanalüüsi abil; võrdluseks on esitatud sama ülesande lahendamisel protseduuriga *Regression* saadud tabelid, milles ka funktsiooni LINEST poolt väljastatavad suurused on punases paksus kirjas, ning vaid üksikväärtusi väljastavate funktsioonide INTERCEPT, SLOPE, RSQ ja STEYX tulemused.



	C	D	E	F	G	H	I	J	K
1		SUMMARY OUTPUT		**Protseduur *Regression* (*Data* -sakk -> *Data Analysis*)**					
2									
3		*Regression Statistics*							
4		Multiple R	0.4581	Mitmene korrelatsioonikordaja					
5		R Square	0.20985	Determinatsioonikordaja					
6		Adjusted R Squa	0.19339	Korrigeeritud determinatsioonikordaja					
7		Standard Error	13.104	Mudeli standardviga					
8		Observations	50	Vaatluste arv					
9									
10		ANOVA					p-väärtus		
11			*df*	*SS*	*MS*	*F*	*Significance F*		
12		Regression	1	2188.934	2188.93	12.748	**0.00082**		
13		Residual	48	8241.886	171.706				
14		Total	49	10430.8					
15			Regressioonivõrrandi parameetrite hinnangud						
16			*Coefficients*	*Standard E*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	
17		Intercept	-96.977	50.178	**-1.933**	**0.0592**	-197.867	3.91327	
18		PIKKUS	0.9781	0.27395	**3.570**	**0.00082**	0.42731	1.52894	
19									
27									
28		**Funktsioon LINEST**							
29									
30		= LINEST(B2:B51, A2:A51, TRUE, **FALSE**)							
31		0.9781	-96.977	Regressioonivõrrandi parameetrite hinnangud					
32									
33		= LINEST(B2:B51, A2:A51, TRUE, **TRUE**)							
34		b 0.9781	a -96.977	Regressioonivõrrandi parameetrite hinnangud					
35		se(b) 0.27395	se(a) 50.178	Parameetrite hinnangute standardvead					
36		0.20985	13.104	Determinatsioonikordaja R^2 ja mudeli standardviga					
37		F 12.748	Df_2 48	F-statistiku (e F-suhte) väärtus ja nimetaja vabadusastmete arv					
38		2188.934	8241.886	Ruutude summad (*Sum of Squares*) dispersioonanalüüsi tabelis					
39									
40		**Mudeli p-väärtus**		(F , n - Df_2 -1, Df_2)					
41		**0.00082**	= F.DIST.RT(D37, COUNT(A2:A51)-E37-1, E37)						
42									
43		**Argumentidele vastavad t-statistiku väärtused ja p-väärtused**							
44		t-statistik	p-väärtus						
45		= E34/E35	= T.DIST.2T(ABS(D46), E37)						
46		**-1.933**	**0.0592**		Vabaliige				
47		**3.570**	**0.00082**		Regressioonikordaja				
48		= D34/D35	= T.DIST.2T(ABS(D47), E37)						
49		t_b=b/se(b)	(ABS(t_b), Df_2)						

Joonis 58. Hüpoteeside testimine funktsiooni LINEST tulemuste alusel; võrdluseks on esitatud sama ülesande lahendamisel protseduuriga *Regression* saadud tabelid.

Eksponentfunktsioon

Hindamaks uuritava tunnuse y ja argumenttunnuste $x_1, \ldots, x_m$ vahelist seost kujul

$$y = a * b_1^{x_1} * \ldots * b_m^{x_m}$$

ehk logaritmilisel skaalal kujul

$$\ln(y) = \ln(a) + x_1 * \ln(b_1) + \ldots + x_m * \ln(b_m)$$

on kasutatav funktsioon LOGEST.

Funktsiooni LOGEST argumendid ja ka väljund on identsed funktsiooniga LINEST.

Joonisel 59 on kujutatud noormeeste kehamassi prognoosimist nende pikkuse alusel eksponentfunktsiooniga kujul

$$\text{Pikkus} = a * b^{\text{Kehamass}}.$$

Tulemustest järeldub, et regressioonivõrrand on kujul

$$\text{Pikkus} = 8{,}21 * 1{,}013^{\text{Kehamass}},$$

kusjuures mudel kirjeldab ära 24,0% noormeeste kehamasside tegelikust varieeruvusest ($R^2 = 0{,}24$).

NB! Ülejäänud parameetrite – ja eelkõige standardvigade – tõlgendamisel peab aga silmas pidama asjaolu, et kõik funktsiooni LOGEST laiendatud väljundis sisalduvad tulemused (arvutatakse, kui argument *Stats* = TRUE) on leitud logaritmilisel skaalal, antud juhul siis mudeli

$$\ln(\text{Pikkus}) = \ln(8{,}21) + \text{Kehamass} * \ln(1{,}013)$$

baasil. St, et funktsioon LOGEST väljastab parameetrite a ja b hinnangute juurde standardvead kujul se[ln(b)] ja se[ln(a)].

Mudeli statistilise olulisuse testimine funktsiooni LOGEST poolt väljastatud F-statistiku väärtuse ja selle nimetaja vabadusastmete arvu *Df*2 alusel käib küll analoogselt funktsiooni LINEST puhul näidatule (vt Joonis 59) – tulemuseks on $p < 0{,}001$, st mudel on statistiliselt olulne –, aga hüpoteeside testimisel mudeli üksikute argumentide tarvis tuleb t-statistiku arvutamisel jagada parameetri logaritmitud väärtus funktsiooni LOGEST poolt väljastatud standardveaga (vt Joonis 59).

NB! Vaid ühe argumenttunnuse korral on eksponentfunktsiooni kujul avalduv seos andmetele sobitatav ka graafiliselt punktdiagrammi ja sellele lisatud trendijoonena. Ainult et graafilisel lahendamisel on vastav funktsioon Excelis defineeritud kujul

$$y = A * e^{B*x}$$

($e = 2{,}718\ldots$) ehk logaritmilisel skaalal kujul

$$\ln(y) = \ln(A) + B * x.$$

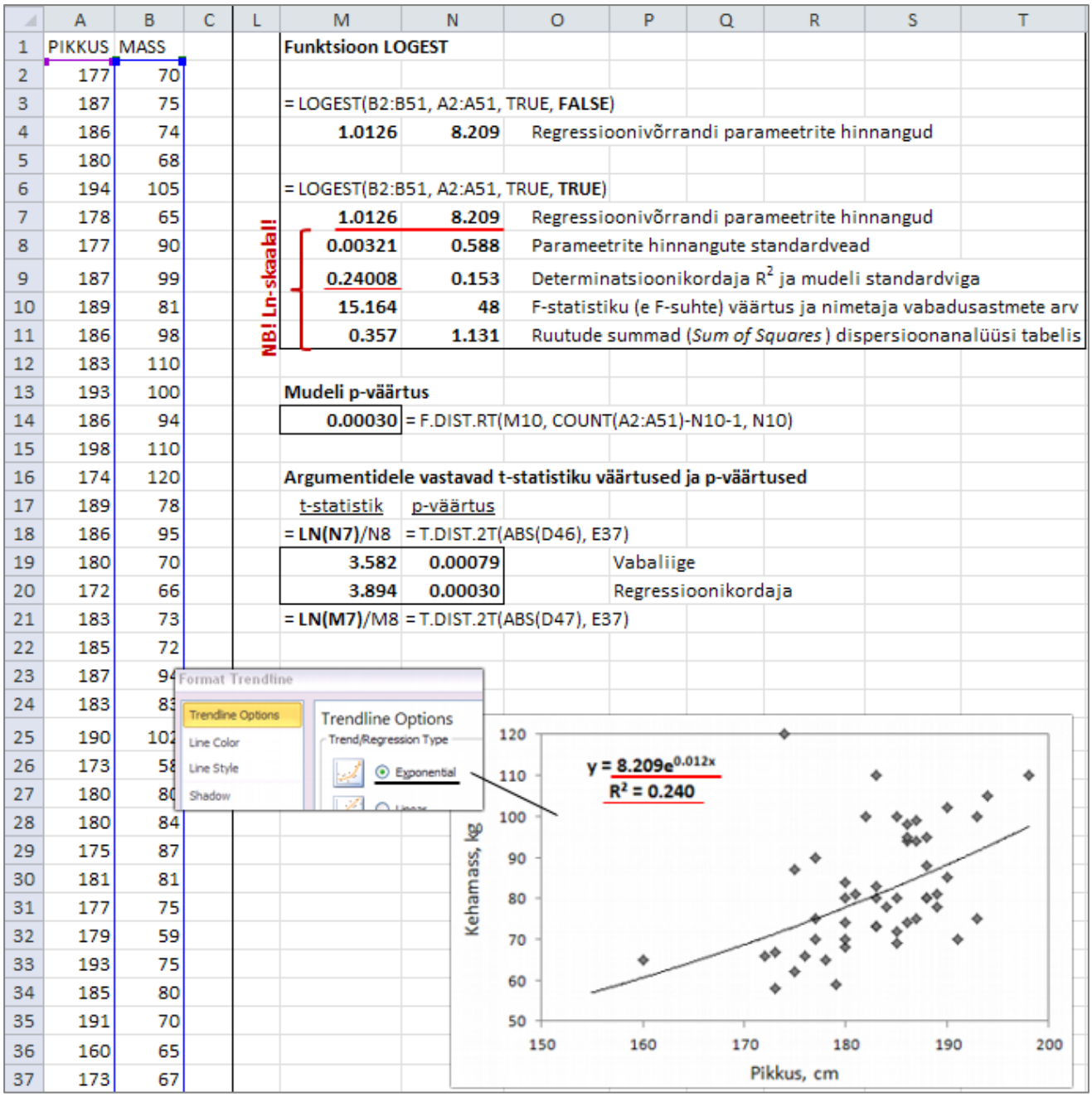
Kõrvutades neid võrrandeid funktsiooni LOGEST poolt hinnatavatega ilmneb, et mudeli vabaliige on nii funktsiooni LOGEST kui ka eksponentsiaalse trendijoone puhul sama: $a = A$ (kus a ja A on vastavalt funktsiooniga LOGEST ja eksponentsiaalse trendijoone abil hinnatavad mudeli vabaliikmed). Eksponentsiaalse trendijoone abil hinnatav regressioonikordaja B on aga võrdne naturaallogaritmiga funktsiooni LOGEST poolt hinnatud regressioonikordjast b: $B = \ln(b)$.

See seos funktsiooni LOGEST tulemuste ja eksponentsiaalse trendijoone vahel ilmneb ka jooniselt 59, kus on täiendavalt esitatud noormeeste kehamassi prognoosimine nende pikkuse alusel eksponentsiaalse trendijoonega – regressioonivõrrand joonisel on kujul

$$\text{Pikkus} = 8{,}21 * e^{0{,}012 * \text{Kehamass}}.$$

Võrrandi vabaliige 8,209 on sama, mis funktsiooni LOGEST poolt saadu, regressioonikordaja 0,012 on aga võrdne naturaallogaritmiga funktsiooni LOGEST väljastatud regressioonikordajast: 0,012 = ln(1,013).

Funktsiooni LOGEST ja eksponentsiaalse trendijoone abil hinnatud eksponentfunktsioonide samasust näitab ka see, et determinatsioonikordaja R^2 väärtus on mõlemal juhul sama: $R^2 = 0{,}24$.



Joonis 59. Noormeeste kehamassi prognoosimine pikkuse alusel eksponentfunktsiooni abil funktsiooniga LOGEST; võrdluseks on esitatud ka sama ülesande lahendus graafiliselt punktdiagrammi ja sellele lisatud eksponentsiaalse trendijoone abil.

Prognoosimine

Prognoosimiseks vastavalt lineaarsele regressioonivõrrandile on Excelis kasutatavad funktsioonid FORECAST ja TREND ning vastavalt eksponentvõrrandile funktsioon GROWTH.

Funktsioon FORECAST väljastab määratud lahtrisse vaid ühele argumenttunnuse väärtusele vastava prognoosi lineaarse regressioonivõrrandi alusel ning talle tuleb ette anda

- väärtus, millele vastavat prognoosi soovitakse leida (*X*),
- funktsioontunnuse *y* andmete blokk (*Known_y's*) ja
- argumenttunnuse *x* andmete blokk (*Known_x's*).

Funktsioonid TREND ja GROWTH on mõlemad massiivifunktsioonid ning nad võimaldavad arvutada prognoosid suurele hulgale ette antud argumenttunnuse väärtustele või mitmese regressiooni puhul argumenttunnuste väärtuste komplektidele. Mõlema funktsiooni rakendamiseks tuleb esmalt võtta blokki prognoositavate väärtuste jagu lahtreid (ükskõik kas veerus või reas), anda argumentidena ette

- funktsioontunnuse *y* andmete blokk (*Known_y's*),
- argumenttunnus(t)e *x* andmete blokk (*Known_x's*),
- väärtused, millele vastavaid prognoose soovitakse leida (*New_x's*), ja
- argumendi *Const* väärtus TRUE (vaikimisi väärtus, hinnatav mudel sisaldab vabaliiget) või FALSE (hinnatav mudel ei sisalda vabaliiget)

ning vajutada valemi rakendamiseks **Ctrl+Shift** ja **Enter** (või OK).

NB! Kui jätta funktsioonidele TREND ja GROWTH kolmas argument *New_x's* ette andmata, arvutavad mõlemad funktsioonid prognoosid kõigi andmeridade tarvis vastavalt neis paiknevate argumenttunnus(t)e väärtustele. Seejuures tuleks muidugi enne funktsioonide rakendamist selekteerida andmestiku suuruse jagu tühje lahtreid (näiteks andmestiku lõpus lisaveeruna), kuhu prognoosid arvutada.

Joonisel 60 on näidatud, kuidas arvutada noormeeste hinnangulised kehamassid etta antud pikkuste tarvis.

	A	B	C	V	W	X	Y	Z	AA
1	PIKKUS	MASS							
2	177	70		**Pikkused**	160	170	180	190	200
3	187	75							
4	186	74		**Prognoositud kehamassid**					
5	180	68			=TREND(B2:B51, A2:A51, W2:AA2, TRUE)				
6	194	105		Lineaarne võrrand	59.5239	69.3052	79.0865	88.8678	98.6491
7	178	65		Eksponentvõrrand	60.6131	68.6805	77.8217	88.1795	99.9159
8	177	90			=GROWTH(B2:B51, A2:A51, W2:AA2, TRUE)				

Joonis 60. Noormeeste kehamasside prognoosimine lineaarsest ja eksponentvõrrandist funktsioonide TREND ja GROWTH abil.

7.4. Regressioonanalüüs *Solver*'i abil

Solver on Exceliga kaasa tulev lisamoodul optimeerimisülesannete lahendamiseks. Menüüsakilt *Data* rakendatav *Solver* võimaldab leida, millised ühtedes lahtrites olevad parameetrite väärtused kas minimiseerivad või maksimeerivad teises lahtris paikneva, neist parameetritest sõltuva funktsiooni väärtuse.

Regressioonivõrrandi parameetrid hinnatakse klassikaliselt **vähimruutude meetodil**. St, et võrrandi parameetriteks valitakse sellised väärtused, mille korral summaarne ruuterinevus tegelike ja prognoositud väärtuste vahel on kõige väiksem. Seega on tegu optimeerimis-ülesandega, mistap saab regressioonivõrrandi parameetrite hindamiseks kasutada *Solver*'it.

Järgnevalt on näidatud, kuidas rakendada *Solver*'it lineaarse ja logistilise regressioonivõrrandi parameetrite hindamiseks. Analoogselt on hinnatavad ka eksponent- jm funktsioonide parameetrid.

NB! Kui valik *Solver* menüü-sakil *Data* puudub, tuleb järgida järgmist menüü-teekonda ning optimeerimisülesannete lahendamise moodul nimega *Solver* sisse lülitada:

File -> Options -> Add-Ins -> Manage |Excel Add-ins| [*Go...*] *-> Solver*.

Täpsemalt Exceli *Solver*'i olemusest ja kasutamisest vt
http://www.solver.com/content/basic-solver-overview-and-example.

Solver'i rakendamisest lineaarsete planeerimisülesannete lahendamisel vt näiteks
http://www.sauga.pri.ee/linplan/linplanfiles.html.

Regressioonanalüüsi teostamisest *Solver*'i abil vt näiteks
http://chemlab.truman.edu/chemlab_backup/DataAnalysis/Excel_Files/AdvancedRegression.htm.

Lineaarse regressioonivõrrandi parameetrid *Solver*'i abil

Uuritava tunnuse *y* väärtuste prognoosimiseks lineaarse regressioonivõrrandi

$$y = a + bx$$

abil vajalike parameetrite *a* ja *b* hindamiseks *Solver*'iga tuleb (Joonised 61A, 61B ja 61C)

1. määrata töölehel lahtrid regressioonivõrrandi parameetrite *a* ja *b* tarvis ning kirjutada sinna mingid **algväärtused**,
2. lisada andmetabeli kõrvale veerg uuritava tunnuse prognoositud väärtuste $\hat{y} = a + bx$ tarvis ning sisestada sinna valem arvutamaks välja prognoose iga andmetabeli rea kohta kasutades sammul 1 määratud lahtrites olevaid parameetrite algväärtuseid (**NB!** valem peab sisaldama viiteid neile lahtritele, mitte kordajate arvulisi väärtusi),
3. arvutada prognooside kõrvale välja prognoosivigade ruudud $(y - \hat{y})^2$ (nn **ruutvead**) ning
4. eraldi lahtrisse **summaarne ruutviga** – viimane kujutab enesest optimeerimisülesande **sihifunktsiooni**, mille väärtus on vaja vastavalt vähimruutude meetodile minimiseerida,
5. käivitada *Solver* (*Data*-sakk -> *Solver*) ning määrata avanenud aknas
   - lahter, milles paikneva valemi tulemuse suhtes soovite optimeerimist läbi viia (lahter, milles paikneb sihifunktsioon; *Set Objective*) – regressioonanalüüsi puhul on selleks summaarne ruutviga,

- mida soovitakse sihifunktsiooniga teha: maksimeerida (*Max*), minimiseerida (*Min*) või võtta võrdseks mingi väärtusega (*Value Of*) – regressioonanalüüsi puhul on eesmärgiks sihifunktsiooni, so summaarse ruutvea, minimiseerimine,
- lahtrid, milles paiknevate väärtuste muutmise teel soovitakse sihifunktsiooni väärtust optimeerida (*By Changing Variable Cells*) – regressioonanalüüsi puhul on neiks muudetavateks väärtusteks regressioonivõrrandi parameetrid *a* ja *b*,
- vajadusel määrata ka
  - lisakitsendused (*Subject to the Constraints*),
  - optimeerimisalgoritm (**NB!** regressioonanalüüsi puhul tuleks valida meetod *GRG Nonlinear*),
  - luba negatiivseteks lahenditeks – selleks tuleb ära võtta „linnuke“ valiku *Make Unconstrained Variables Non-Negative* eest (**NB!** see on oluline koht *Solver*'i rakendamisel regressioonivõrrandi parameetrite hindamisel, sest erinevalt mitmetest planeerimise ülesannetest regressioonivõrrandi parameetrite puhul väärtuste positiivsuse nõuet pole),
  - täpsustada optimeerimisalgoritmi tööd (valik *Options*) – kui sooviks on vaid summaarset ruutviga minimiseerivate regressioonivõrrandi parameetrite välja arvutamine ja optimeerimisprotsessi vahepealsed tulemused huvi ei paku, on mõttekas võtta ära „linnuke“ valiku *Show Iteration Results* eest,

6. käivitada optimeerimisalgoritm vajutades nupule *Solve* ning
7. lasta Excelil optimeerimisalgoritmi koondumise järel säilitada töölehel *Solver*'i poolt leitud parameetrite väärtused: *Keep Solver Solution*.

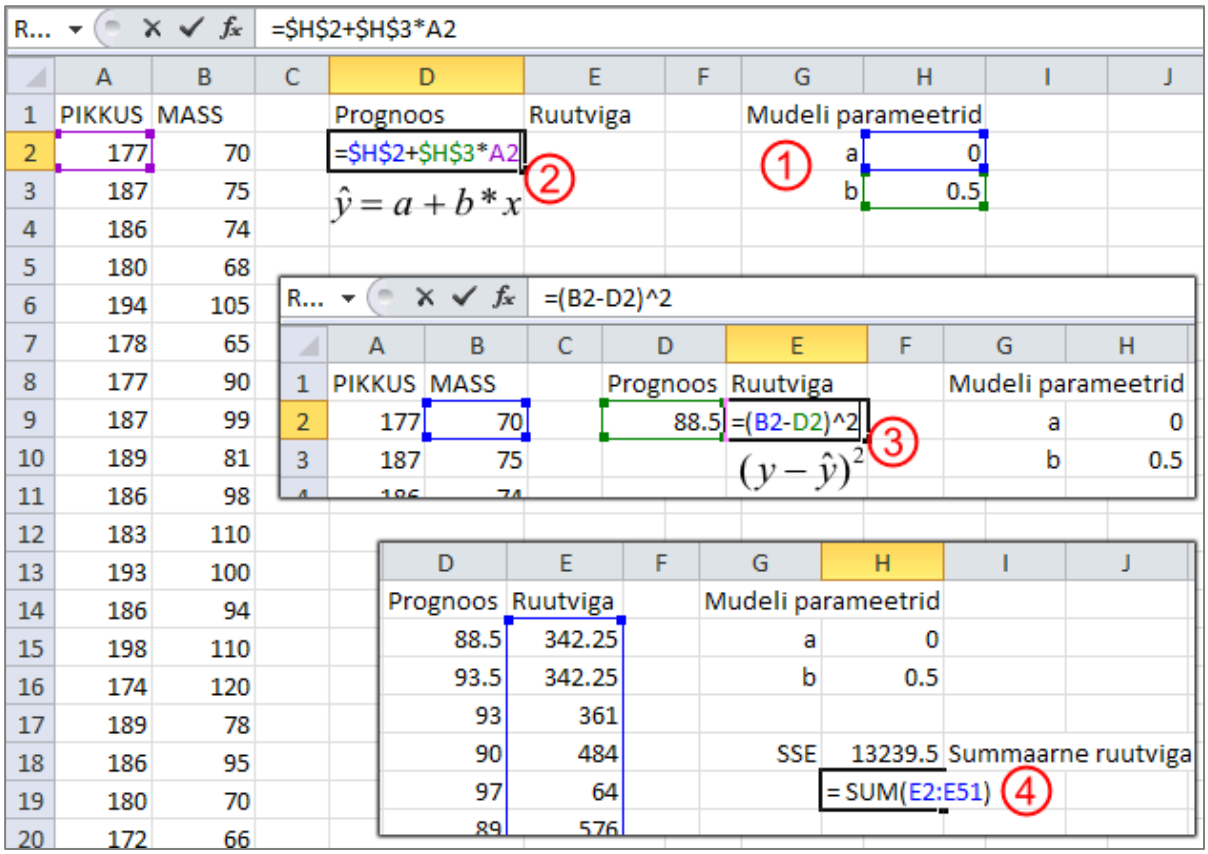
Joonistel 61A, 61B ja 61C on illustreeritud *Solver*'i rakendamist prognoosimaks noormeeste kehamassi nende pikkuse abil lineaarse regressioonivõrrandiga.

Võrrandi vabaliikme *a* ja regressioonikordaja *b* algväärtusteks sammul 1 on võetud vastavalt 0 ja 0,5. Miks need arvud? Lihtsalt niivõrd-kuivõrd loogiline arutelu – kui pikkus on null, siis peaks ka kehamass olema null (inimest pole), seega võiks vabaliige olla null; ja kuna pikkuse ja kehamassi vahel on ilmselt positiivne seos – mida pikem inimene, seda enam ta kaalub – ja vaevalt iga lisasentimeeter pikkuses kehamassile üle kilo lisab (aga mine tea), siis võiks regressioonikordaja alglähendina proovida mingit väikest positiivset arvu, näiteks 0,5 või 1. Hinnatavate parameetrite algväärtuste muutmine on ka üks võimalik tegutsemisvariant, kui *Solver* algselt paika pandud väärtustest alustades ei suuda lahendit leida.

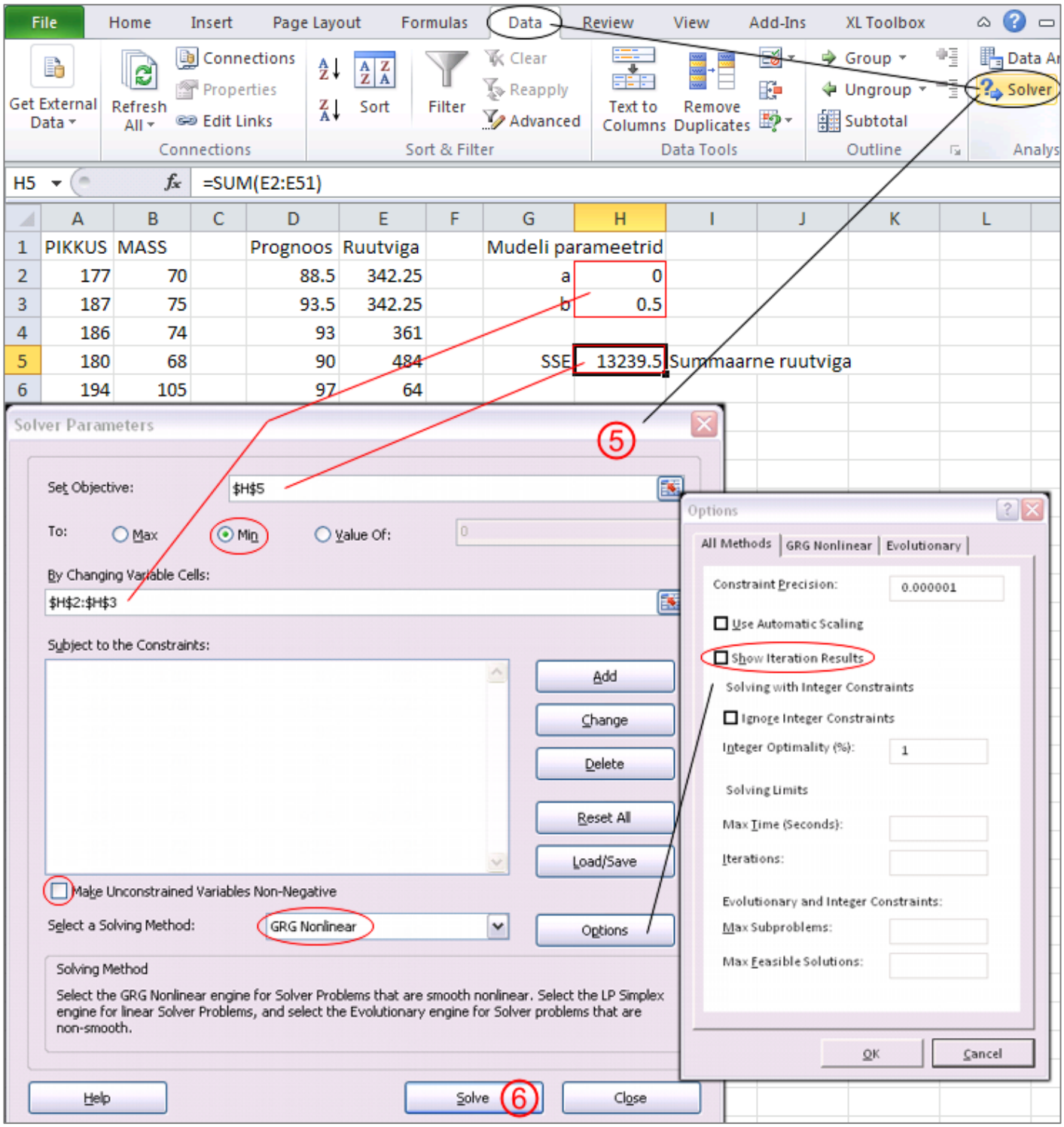
Joonisel 61C on Solver'i tulemused esitatud kõrvuti protseduuri *Regression* väljundiga. Tulemused – nii regressioonivõrrandi parameetrite hinnangud kui ka neile vastav vähim võimalik summaarne ruutviga – on võrdsed.

Regressioonivõrrandi parameetrite hinnangute usaldusväärsuse ja statistlise olulisuse üle otsustamiseks vajalikud arvutused on samuti võimalik läbi viia *Solver*'iga saadud hinnangute jaoks, ainult selleks peab teadma pisut põhjalikumalt vastavate analüüside matemaatilist tausta. Soovi korral saab abi näiteks artiklist:

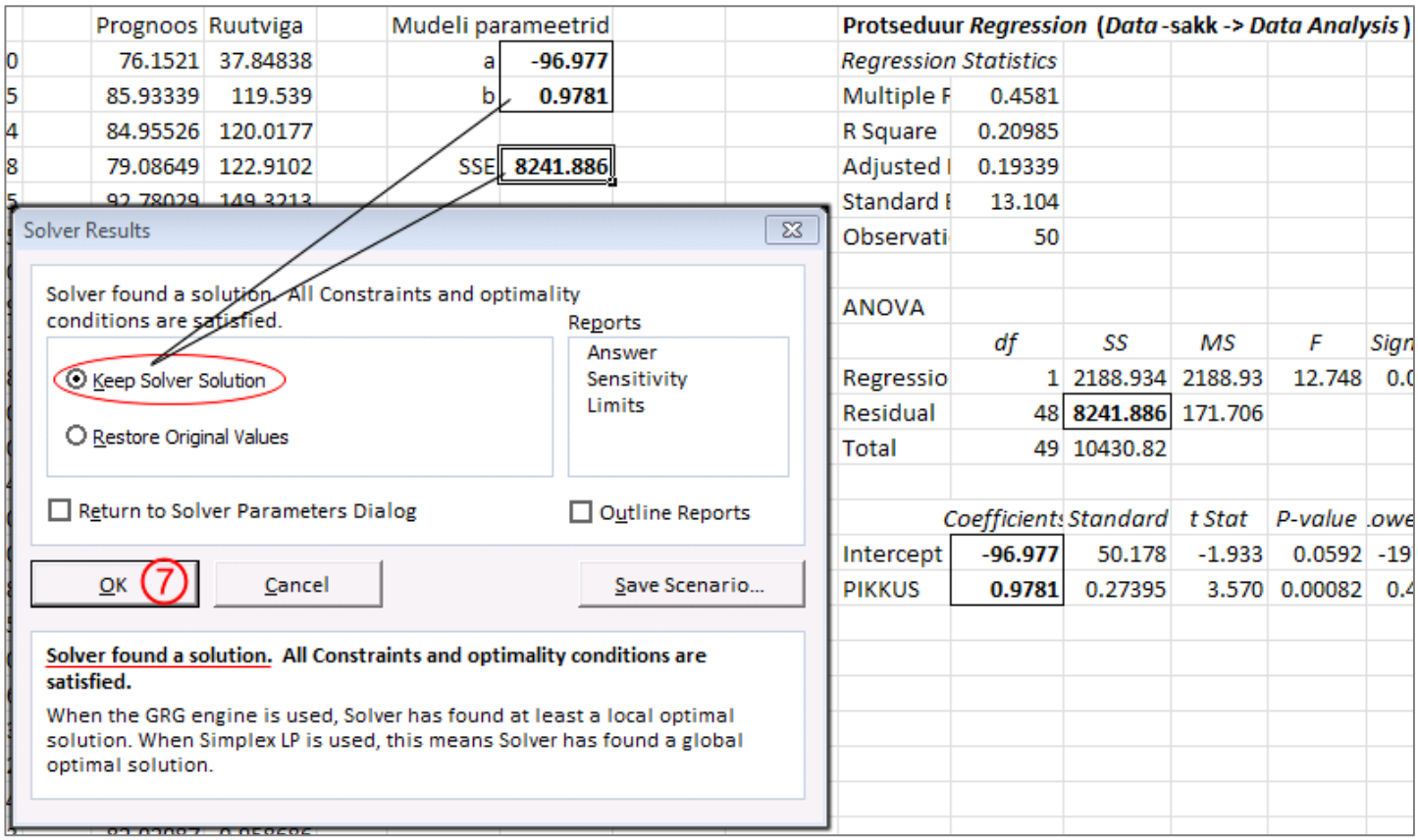
Harris, D. C. (1998). Nonlinear Least-Squares Curve Fitting with Microsoft Excel Solver. Journal of Chemical Education, 75, 119-121 (http://jchemed.chem.wisc.edu/Journal/Issues/1998/Jan/PlusSub/V75N01/p119.pdf).



Joonis 61A. Noormeeste kehamassi prognoosimine nende pikkuse alusel – mudeli parameetrite ja neist sõltuva, *Solver*'i abil minimiseeritava, summaarse ruutvea vahelise seose esitamine valemite abil.



Joonis 61B. Noormeeste kehamassi prognoosimine nende pikkuse alusel – mudeli parameetrite hindamine *Solver*'i abil.



Joonis 61C. Noormeeste kehamassi prognoosimine nende pikkuse alusel –*Solver*'i abil hinnatud parameetrite väärtused; võrdluseks on esitatud sama ülesande lahendamisel protseduuriga *Regression* saadud tabelid.

Logistilise regressioonivõrrandi parameetrid *Solver*'i abil

Eelnevalt lineaarse regressioonivõrrandi parameetrite hindamiseks kasutatud metoodika on rakendatav ka mittelineaarsete regressioonivõrrandite puhul, ainuke erinevus on sammul 2 kasutatavas prognoosivõrrandis.

Järgnevalt on lühidalt näidatud, kuidas hinnata logistilise regressioonivõrrandi

$$p = \mathrm{P}(y=1|x) = \exp(a + bx)/[1 + \exp(a + bx)] = 1 / [1 + \exp(-a - bx)]$$

ehk logit-funktsiooni

$$\ln[p/(1-p)] = \mathrm{logit}(p) = a + bx.$$

parameetreid *Solver*'i abil. Modelleeritavaks on siinkohal uuritava sündmuse y toimumise tõenäosus $p = \mathrm{P}(y=1)$.

Täpsemalt logistilise regressiooni olemusest vt vastavaid lehekülgi õpiobjektis „Binaarsete tunnuste analüüs" (http://ph.emu.ee/~ktanel/bin_tunnuste_analyys/).

Konkreetse näitena on vaatluse all tudengi meheks olemise tõenäosuse prognoosimine nädalas keskmiselt tarbitava õllekoguse alusel (**NB!** uuritav tunnus – antud näites sugu – peab olema kodeeritud arvuliseks väärtustega 0 ja 1). Ülesande lahendamiseks tuleb (vt ka Joonis 62)

1. määrata töölehel lahtrid regressioonivõrrandi parameetrite *a* ja *b* tarvis ning kirjutada sinna mingid **algväärtused** – viimaste väljamõtlemisel võib arvestada, et

- vabaliige *a* näitab meheks olemise tõenäosust juhul, kui tudeng üldse õlut ei joo, ja selleks võtta näiteks väärtuse 0 (või 0,25 või … mingi väärtuse nulli ja ühe vahel),
- regressioonikordaja *b* näitab meheksolemise logaritmilise šansi muutust kordades, kui tarbitav õllekogus suureneb ühe liitri võrra, ja ilmselt on selle väärtuse näol tegu mingi ühest pisut suurema arvuga, näiteks 1,1 – šanss olla mees suureneb $e^{1,1}$=3 korda, kui nädalas tarbitav õllekogus suureneb 1 liitri võrra (joonisel 62 esitatud näites on alglähendiks võetud küll 0,1, aga nagu näha, koondub hindamisprotsess ka sellise sisuliselt mitte kõige õigema, aga matemaatiliselt siiski suhteliselt sobiva alglähendi puhul),

2. lisada andmetabeli kõrvale veerg uuritava tunnuse prognoositud väärtuste

$$\hat{y} = 1 / \left(1 + \exp(-a - bx)\right)$$

   tarvis ning sisestada sinna valem arvutamaks välja prognoose iga andmetabeli rea kohta kasutades sammul 1 määratud lahtrites olevaid parameetrite algväärtuseid (*x* selles valemis on õllekogus),

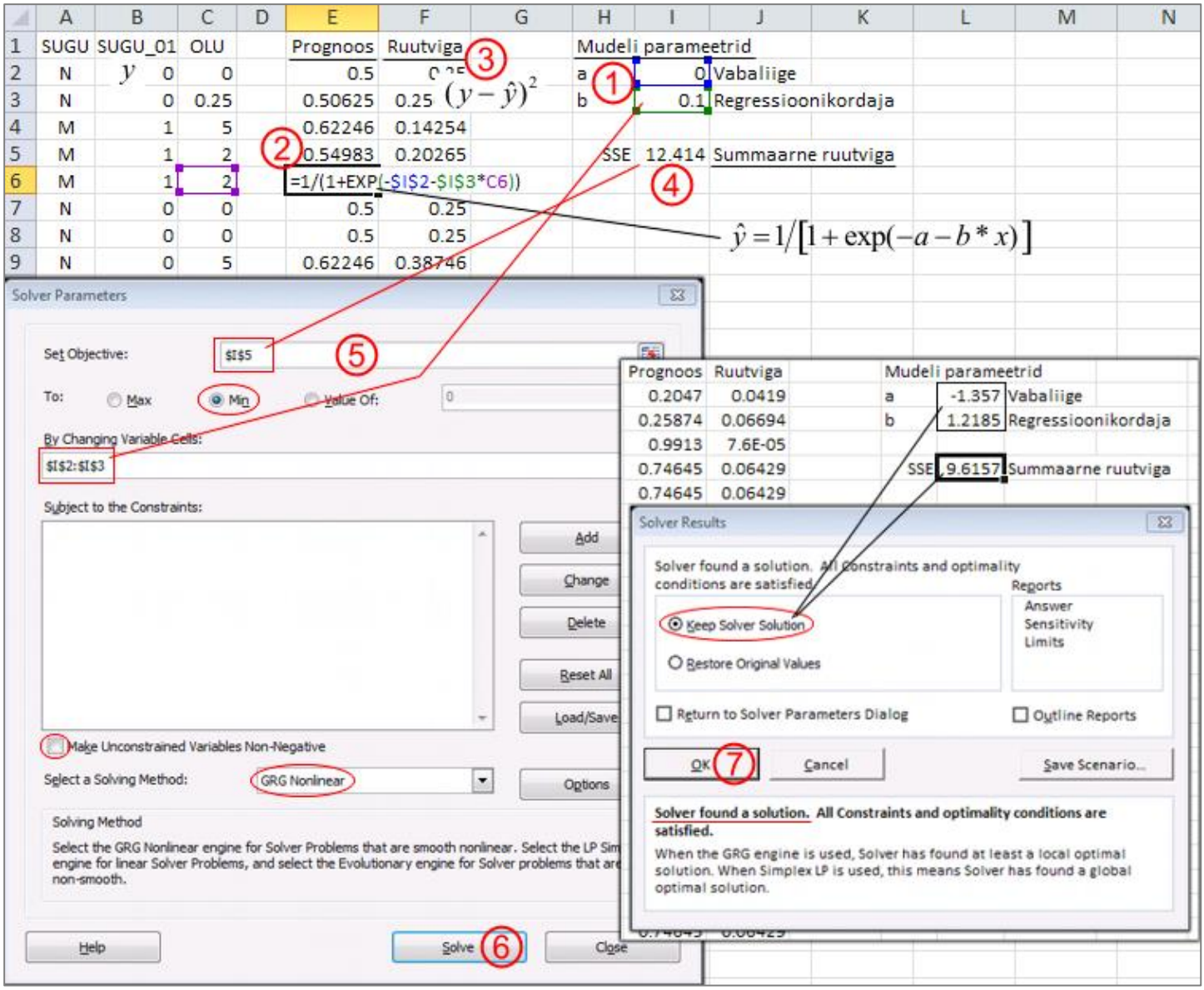
3. arvutada prognooside kõrvale välja prognoosivigade ruudud $(y - \hat{y})^2$ (nn **ruutvead**) ning
4. eraldi lahtrisse **summaarne ruutviga**,
5. käivitada *Solver* (*Data*-sakk -> *Solver*) ning
   - anda ette minimiseeritavat summaarset ruutviga sisaldav lahter (*Set Objective*),
   - määrata optimeerimise suunaks minimiseerimine (*Min*),
   - anda ette regressioonivõrrandi parameetrite algväärtuseid sisaldavad lahtrid (*By Changing Variable Cells*),
   - määrata optimeerimisalgoritmiks *GRG Nonlinear*,
   - võtta ära „linnuke“ valiku *Make Unconstrained Variables Non-Negative* eest (sest regressioonivõrrandi parameetrite puhul väärtuste positiivsuse nõuet pole),
   - võtta ära „linnuke“ valiku *Options -> Show Iteration Results* eest (kui ei ole just soovi igal iteratsiooni sammul leitud hinnanguid eraldi uurida),
6. käivitada optimeerimisalgoritm vajutades nupule *Solve* ning
7. lasta Excelil optimeerimisalgoritmi koondumise järel säilitada töölehel *Solver*'i poolt leitud parameetrite väärtused: *Keep Solver Solution*.

Joonisel 62 esitatud analüüsi tulemuste kohaselt on tudengi meheks olemise tõenäosus hinnatav valemist

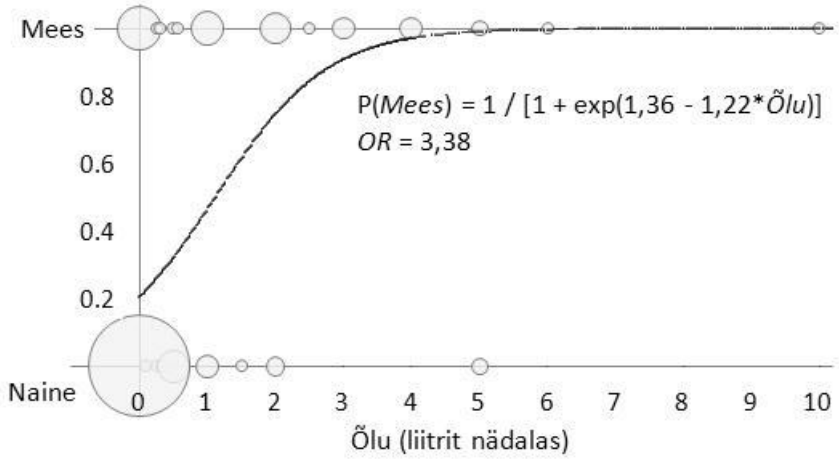
$$P(Mees) = 1 / [1 + \exp(1{,}357 - 1{,}22 * Õlu)$$

ja šanss olla mees suureneb $e^{1,22}$ = 3,38 korda, kui nädalas tarbitav õllekogus suureneb 1 liitri võrra.

Joonisel 63 on sama seos kujutatud graafiliselt (kuidas taolist joonist Excelis konstrueerida vt vastavat õpetust õpiobjektis „MS Excelile mitteomased andmeanalüüsil kasutatavad joonised“, http://ph.emu.ee/~ktanel/joonised_excelis/).



Joonis 62. Logistilise regressioonivõrrandi parameetrite hindamine *Solver*'iga prognoosimaks tudengi meheks olemise tõenäosust nädalas keskmiselt tarbitava õllekoguse alusel.



Joonis 63. Logistilise regressioonivõrrandi graafik prognoosimaks tudengi meheks olemise tõenäosust nädalas keskmiselt tarbitava õllekoguse alusel (ringid vastavad erinevatele õllekogustele ja ringi suurus tudengite arvule, pidev must joon on logistilise regressioonivõrrandi graafik).

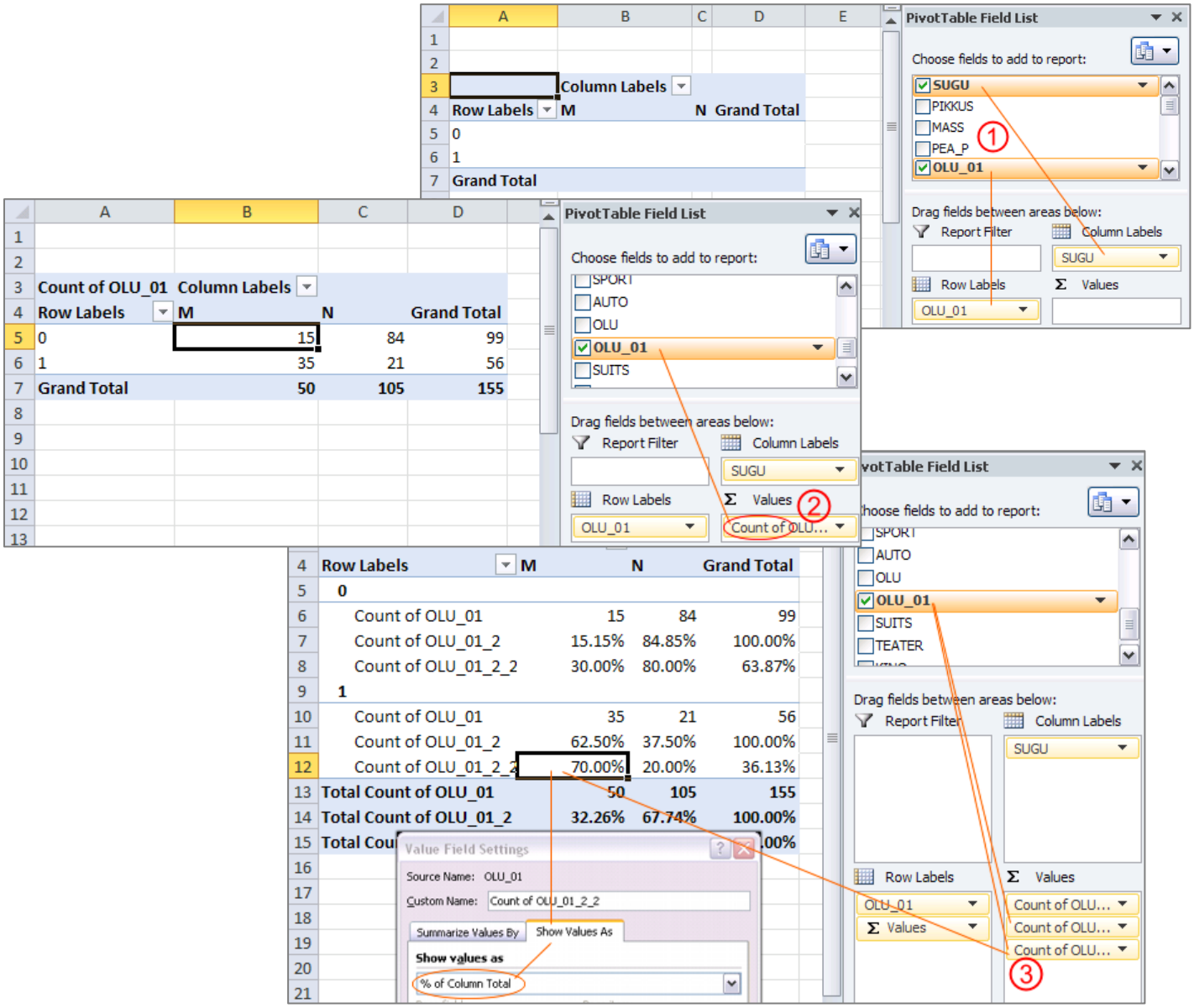
8. Kahemõõtmeline sagedustabel

8.1. Kehamõõtmeline sagedustabel

Mittearvuliste tunnuste ja/või diskreetsete arvtunnuste vahelise seose iseloomustamiseks kasutatav kahemõõtmeline sagedustabel on MS Excelis konstrueeritav *PivotTable* abil.

Selleks tuleb *PivotTable*'s (*Insert*-sakk -> *PivotTable*, vajadusel vt pt 1.4)

1. valida reafaktoriks üks ja veerufaktoriks teine uuritav tunnus (Joonis 64),
2. lohistada tabeli *Values*-lahtrisse ükskõik kumb uuritavatest tunnustest ning määrata leitavaks karakteristikuks vaatluste arv (*Value Field Settings -> Summarize Values By -> Count*),
3. rea ja/või veeru suhteliste sageduste leidmiseks tuleb korrata sammu 2 ning käskida Excelil kuvada absoluutsageduste asemel
   - reasagedusi (*Value Field Settings -> Show Values As -> % of Row Total*) ja/või
   - veerusagedusi (*Value Field Settings -> Show Values As -> % of Column Total*).



Joonis 64. Tudengite soo ja õllejoomise vahelise seose kirjeldamine kahemõõtmelise sagedus-tabeliga.

Tabelist joonisel 64 peale kolmanda sammu teostamist on näha, et kokku on valimis 155 tudengit, kellest 105 on neiud ja 50 noormehed, õllejoojaid on kokku 56 ja õlut mittejoovaid tudengeid 99. Õllejoomise ja soo vahelist seost näitab see, et 50-st noormehest 35 (so 70,0%) joovad õlut, samas kui 105-st neiust joob õlut vaid 21 (so 20,0%). Õlut joovatest tudengitest on 62,5% noormehed ja 37,5% neiud, õlut mittejoovatest tudengitest on aga vaid 15,2% noormehed ja tervelt 84,9% neiud.

8.2. χ^2-test

Mittearvuliste ja/või diskreetsete arvtunnuste vahelise kahemõõtmelise sagedustabeli kujul esitatud seose statistilise olulisuse testimiseks kasutatakse sagedaimini χ^2-testi, millele vastav hüpoteeside paar on kujul:

H_0: tunnused on sõltumatud ehk potentsiaalne riskifaktor ei mõjuta uuritava sündmuse toimumist,

H_1: tunnused on sõltuvad ehk potentsiaalne riskifaktor mõjutab uuritava sündmuse toimumist.

MS Excelis on χ^2-test teostatav funktsiooniga CHISQ.TEST, mis teostab andmetele vastava empiirilise ja nullhüpoteesile vastava teoreetilise sagedustabeli võrdluse ning annab tulemuseks olulisuse tõenäosuse *p* väärtuse.

Funktsiooni CHISQ.TEST rakendamiseks tuleb

1. konstrueerida uuritavate tunnuste vaid absoluutsagedusi sisaldav kahemõõtmeline sagedustabel (nö empiiriline, andmetele vastav sagedustabel),
2. arvutada teise tabelisse tunnuste sõltumatuse juhule (nullhüpoteesile) vastavad oodatavad teoreetilised sagedused valemist

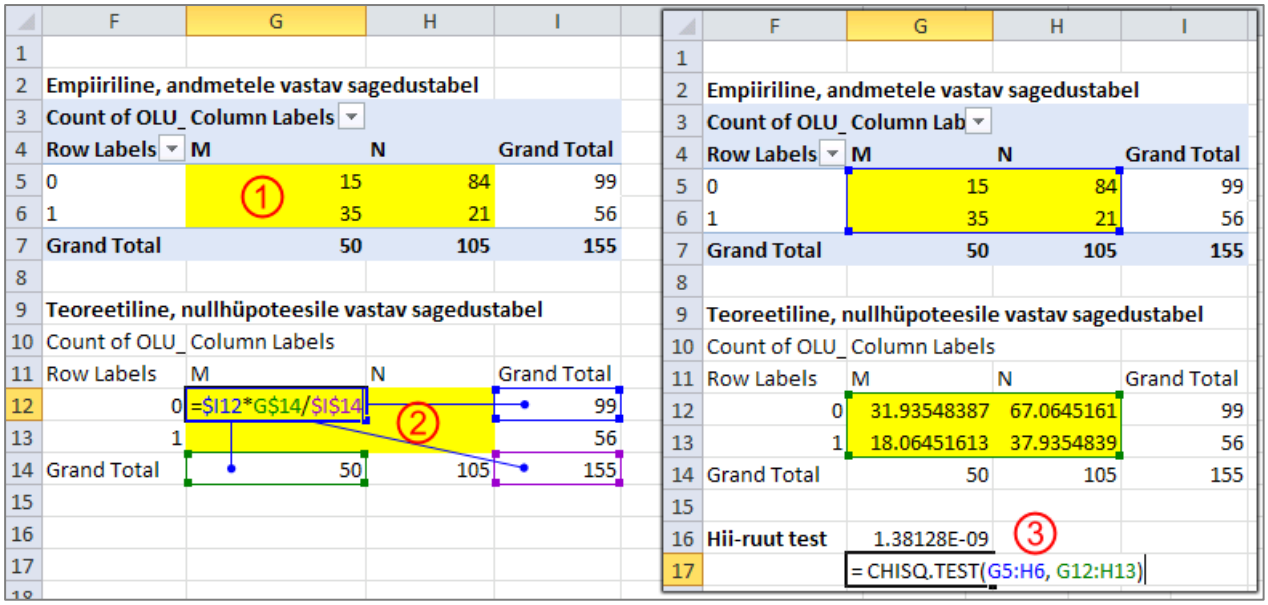
   $$n_{ij} = n_{i.} * n_{.j} / n,$$

   kus n_{ij} on oodatav sagedus tabeli *i*. reas ja *j*. veerus, $n_{i.}$ ja $n_{.j}$ on vastavalt sagedustabeli *i*. rea ja *j*. veeru summad ning *n* on vaatluste arv,
3. panna kursor lahtrisse, kuhu soovite saada χ^2-testi p-väärtust, ning rakendada funktsiooni CHISQ.TEST, andes argumentidena ette
   - empiirilised sagedused (*Actual_range*) ja
   - oodatavad teoreetilised sagedused (*Expected_range*).

NB! Funktsiooni CHISQ.TEST argumentidena antakse sagedustabelid ette ilma ääre-summadeta.

Joonisel 65 on näidatud χ^2-testi teostamist testimaks tudengite soo ja õllejoomise vahelise seose statistilist olulisust. Nagu eelmise punkti lõpus kirjutatud, on tudengite sugu ja õllejoomine seotud – noormeeste hulgas on õllejoojaid rohkem. χ^2-testi alusel saab väita, et see seos on ka statistiliselt oluline ($p < 0{,}001$).

Märkus. χ^2-test on asümptootiline test, st et arvutatav teststatistik on χ^2-jaotusega vaid ligikaudu ja testi tulemus on seda täpsem, mida rohkem on andmeid. Traditsiooniliselt loetakse χ^2-testi kasutamine õigustatuks, kui kõigis teoreetilise sagedustabeli lahtrites paiknevad sagedused on viiest suuremad ($n_{ij} > 5$).



Joonis 65. χ^2-testi läbiviimine Excelis funktsiooniga CHISQ.TEST testimaks tudengite soo ja õllejoomise vahelise seose statistilist olulisust kahemõõtmelise sagedustabeli alusel.

8.3. Fisheri täpne test

Fisheri täpne test on alternatiiv χ^2-testile. Fisheri täpne test annab, nagu nimigi ütleb, täpse olulisuse tõenäosuse *p* väärtuse ning on tänu oma töömahukusele rakendatav eelkõige väikeste valimite korral (suurte valimite korral annab piisavalt täpse tulemuse ka χ^2-test).

Kuigi Excelis puuduvad vahendid Fisheri täpse testi teostamiseks ja reaalseteks arvutusteks on mõttekam kasutada mõnda statistikapaketti või mõnda statistilisi analüüse teostavat internetilehekülge (vt http://www.eau.ee/~ktanel/bin_tunnuste_analyys/pt27.php), on järgnevalt nö pedagoogilistel kaalutlustel siiski näidatud ka Fisheri täpse testi teostamist Excelis (Joonis 66).

1. Fisheri täpse testi korral leitakse esmalt nagu χ^2-testi puhulgi andmetele vastav e empiiriline sagedustabel.
2. Seejärel pannakse kirja kõik sellised alternatiivsed sagedustabelid, mis erineva „sisu“ korral annavad tulemuseks ikkagi samad rea- ja veerusummad. Excelis on selliseks tegevuseks lihtsaim variant
   - otsida empiirilises sagedustabelis üles väikseimale rea- ja veerusummale vastav lahter ning avaldada ülejäänud sagedused valemina antud lahtriväärtusest ning rea- ja veerusummadest (Joonisel 66 on vastav lahter värvitud oranžiks),
   - teha taolisest empiirilisest sagedustabelist *k* koopiat, kus $k = \min(n_{i.},n_{.j})$, $n_{i.}$ ja $n_{.j}$ on vastavalt *i*. rea ja *j*. veeru summad (fikseeritud rea- ja veerusummade puhul on erinevaid sagedustabeleid $\min(n_{i.},n_{.j})+1$ tükki, koopiad on mõtet teha üks vähem, sest empiiriline sagedustabel on juba olemas), ning
   - muuta kopeeritud tabelites väikseimale rea- ja veerusummale vastavas lahteris paiknevat väärtust 0-st $\min(n_{i.},n_{.j})$-ni – eeldusel, et teiste sageduste arvutamiseks sisaldas tabel vastavaid valemeid, arvutatakse need kõik automaatselt.

3. Kõigi leitud sagedustabelite saamise tõenäosus fikseeritud rea- ja veerusummade puhul eeldusel, et kõik katseobjektid/indiviidid jaotuvad tabelisse juhuslikult (so tunnuste sõltumatuse korral), on arvutatav hüpergeomeetrilise jaotuse tõenäosusfunktsioonist kujul

$$p_{\text{tabel}} = \frac{\prod_{i=1}^{k} n_{i.}! \quad \prod_{j=1}^{m} n_{.j}!}{n! \prod_{i,j} n_{i,j}}, \quad n! = n \times (n-1) \times \ldots \times 2 \times 1,$$

kus n on vaatluste koguarv, n_{ij} sagedustabeli i. reas ja j. veerus paiknev sagedus ning $n_{i.}$ ja $n_{.j}$ vastavalt i. rea ja j. veeru summad ($n!$ on arvu n faktoriaal, mis Excelis on leitav funktsiooniga FACT). Juhul, kui mõlemad uuritavad tunnused on binaarsed, esitatakse kahemõõtmeline sagedustabel sageli kujul

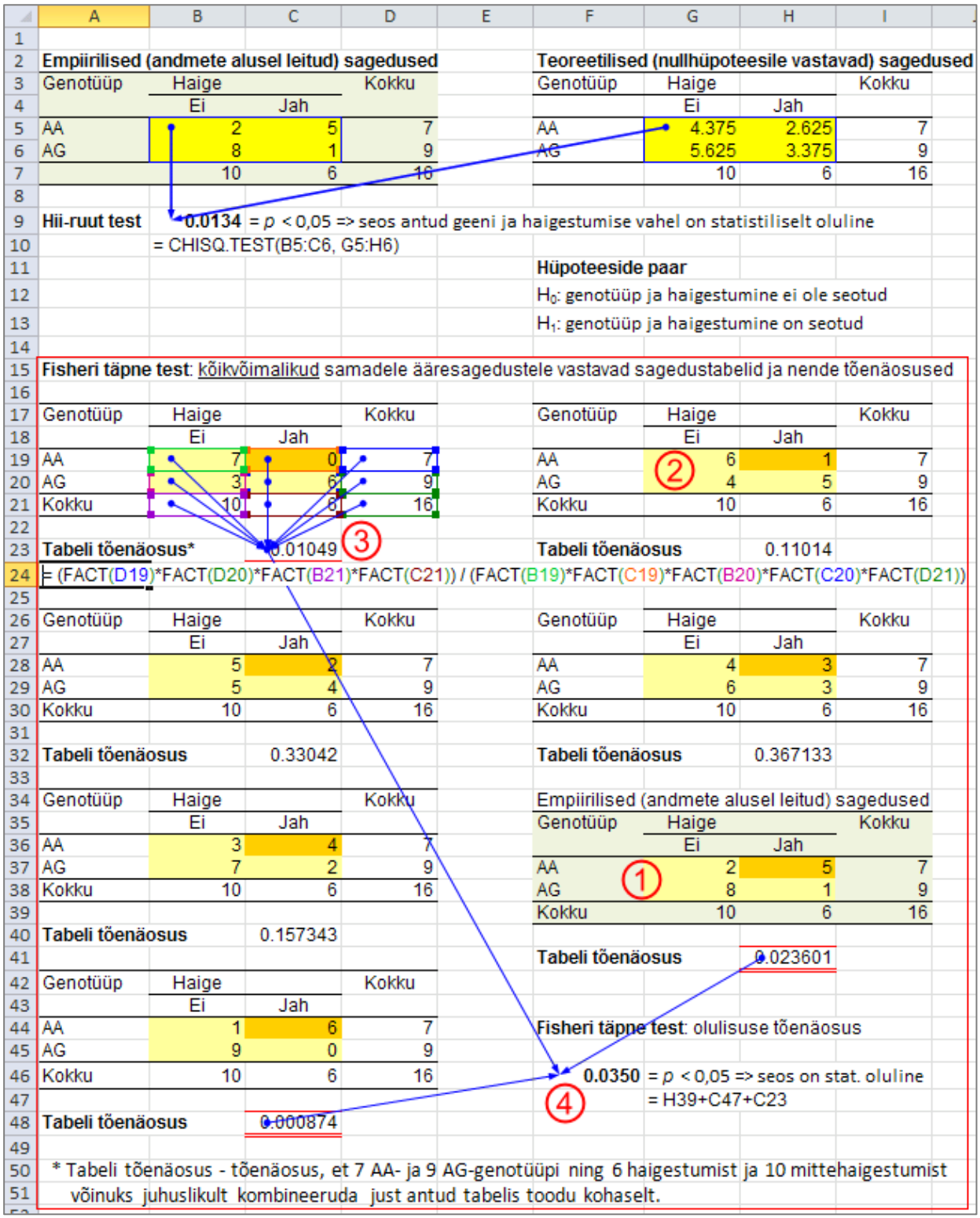
	Juhud	Kontrollid	Kokku
Eksponeeritud	a	b	$a+b$
Mitteeksponeeritud	c	d	$c+d$
Kokku	$a+c$	$b+d$	$a+b+c+d=n$

ning tõenäosus taolise tabeli saamiseks fikseeritud rea- ja veerusummade puhul juhuslikult avaldub valemiga

$$p_{\text{tabel}} = [(a+b)! \times (c+d)! \times (a+c)! \times (b+d)!] \,/\, [n! \times a! \times b! \times c! \times d!].$$

4. Viimaks liidetakse kokku empiirilise ning sellest ekstreemsemate (vähemtõenäolises suunas valitud) sagedustabelite esinemistõenäosused – tulemuseks on ühepoolsele hüpoteesile vastav olulisuse tõenäosus (p-väärtus);

   kahepoolsele hüpoteesile vastava olulisuse tõenäosuse saab, summeerides kõik empiirilise sagedustabeli esinemistõenäosusega võrdsed või sellest väiksemad tabeliste tõenäosused. Mõnikord leitakse kahepoolsele hüpoteesile vastav olulisuse tõenäosus ka korrutades ühepoolsele hüpoteesile vastava p-väärtuse lihtsalt kahega.



Joonis 66. Fisheri täpse testi teostamine Excelis. Võrdluseks on ära toodud ka χ^2-testi teostus ja tulemus. Empiiriline andmetabel on varjutatud taustaga, selle esinemistõenäosusega võrdsed ja väiksemad tõenäosused, mille summana kujuneb Fisheri täpse testi p-väärtus, on allajoonitud punase topeltjoonega.

9. Dispersioonanalüüs

9.1. Ühefaktoriline dispersioonanalüüs

Ühefaktoriline dispersioonanalüüs on Excelis teostatav protseduuriga *ANOVA: Single Factor* (*Data*-sakk -> *Data Analysis*).

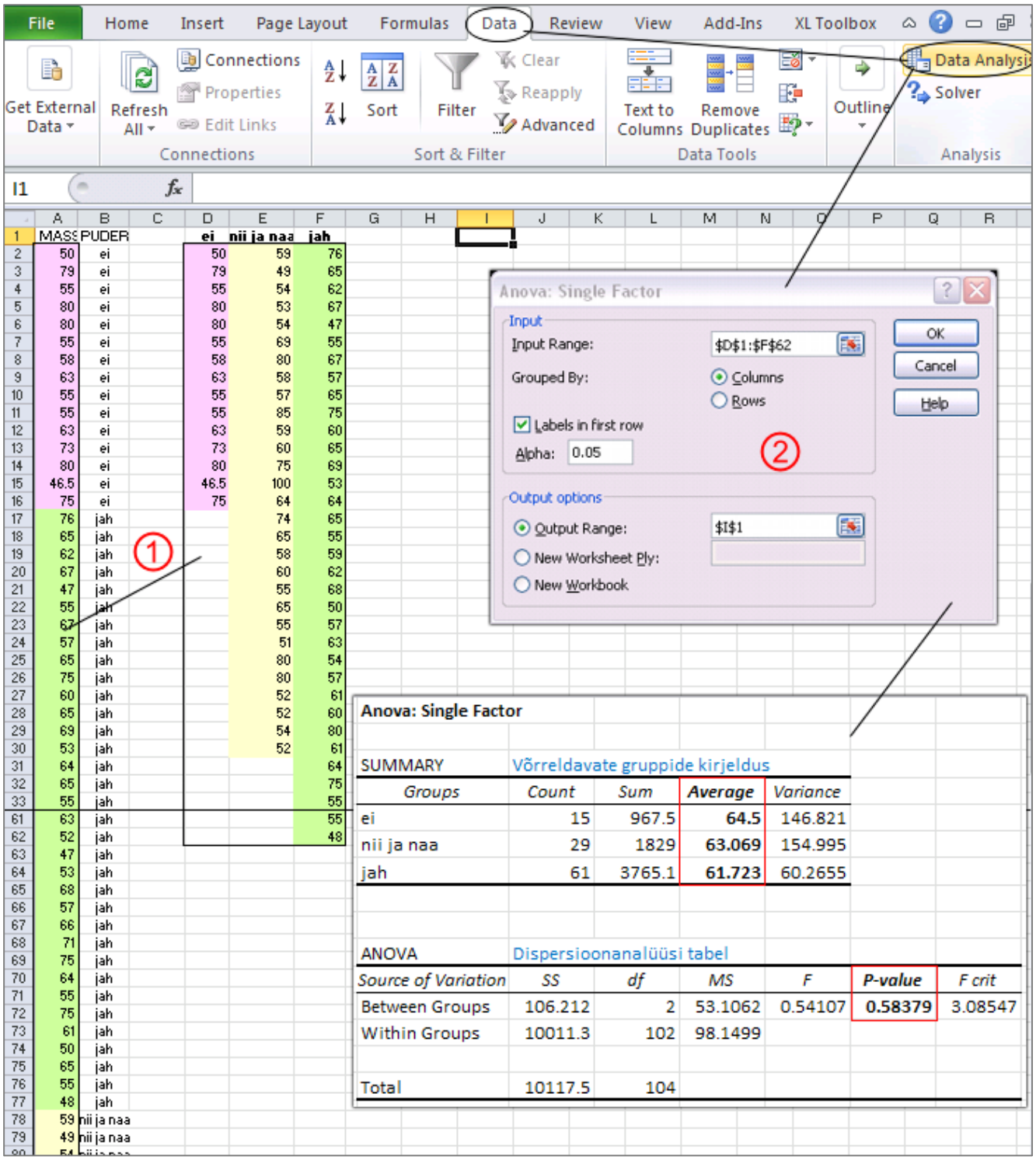
1. Esimese etapina dispersioonanalüüsi teostamisel tuleb esitada analüüsitavad andmed kujul, kus igale faktori tasemele vastaks üks veerg (või rida), milles paiknevad kõik sellel faktori tasemel sooritatud mõõtmised.

   NB! Erinevates gruppides võib olla sooritatud erinev arv mõõtmisi.

2. Teise etapina tuleb see abitabel anda ette protseduurile *ANOVA: Single Factor*, (argumendi *Input Range* väärtuseks), määrates täiendavalt,
   - kas võrreldavad grupid paiknevad kõrvuti veergudes (vaikimisi variant; *Grouped By = Columns*) või ridades (*Grouped By = Rows*),
   - kas andmetabeli esimeses reas (või veerus) paiknevad gruppide nimed (*Labels in first row*),
   - milline on olulisuse nivoo F-statistiku kriitilise väärtuse arvutamiseks (*Alpha*, vaikimisi 0,05),
   - kuhu paigutada tulemustabelid (*Output options*): samale töölehele (*Output Range*), uuele töölehele (*New Worksheet Ply*) või uude faili (*New Workbook*).

Joonisel 67 on näidatud ühefaktorilise dispersioonanalüüsi teostamist Excelis protseduuriga *ANOVA: Single Factor* testimaks mannapudru söömise mõju esimese kursuse neidude kehamassile.

Nagu dispersioonanalüüsi tulemustes sisalduvast kirjaldavate statistikute tabelist nähtub, on mannaputru söövaid neide kokku 61, mannaputru mõnikord (kui vaja, siis vaja) söövaid neide 29 ja mannaputru mittesöövaid neide 15. Keskmine kehamass on kõrgeim just nimelt viimastel – 64,5 kg, mis on peaaegu 2,8 kg enam võrreldes mannaputru söövate neidude keskmise kehamassiga 61,7 kg. Dispersioonanalüüsi tulemuste põhjal ei ole aga siiski teaduslikku alust rääkida mannapudru kaalu alandavast mõjust – erinevus mannapudrusse erinevalt suhtuvate neidude keskmiste kehamasside vahel ei ole statistiliselt oluline (p = 0,58).



Joonis 67. Mannapudru söömise mõju esimese kursuse neidude kehamassile – ühefaktoriline dispersioonanalüüs protseduuriga *ANOVA: Single Factor*.

9.2. Kahefaktoriline dispersioonanalüüs

Kahefaktoriline dispersioonanalüüs on Excelis teostatav üksnes tasakaaluliste andmete korral, st juhul, kui mõlema faktori kõigil tasemetel on sooritatud ühepalju mõõtmisi. Aga ka siis vaid ettemääratud mudeli kohaselt:

- protseduur *Anova: Two-Factor Without Replication* eeldab, et mõlema faktori kõigi tasemete kõigi kombinatsioonide puhul on teostatud üksnes üks mõõtmine, ja testib sedasi vaid mõlema faktori peamõju statistilist olulisust;
- protseduur *Anova: Two-Factor With Replications* eeldab, et mõlema faktori kõigi tasemete kõigil kombinatsioonidel on teostatud võrdne ja ühest suurem arv mõõtmiseid, ja testib mõlema faktori peamõju pluss nende koosmõju statistilist olulisust.

Seega sobivad Exceli kahefaktorilise dispersioonanalüüsi protseduurid vaid täpselt planeeritud ja läbi viidud väikesemahuliste katsete andmete analüüsimiseks.

Kahefaktoriline kordusteta dispersioonanalüüs

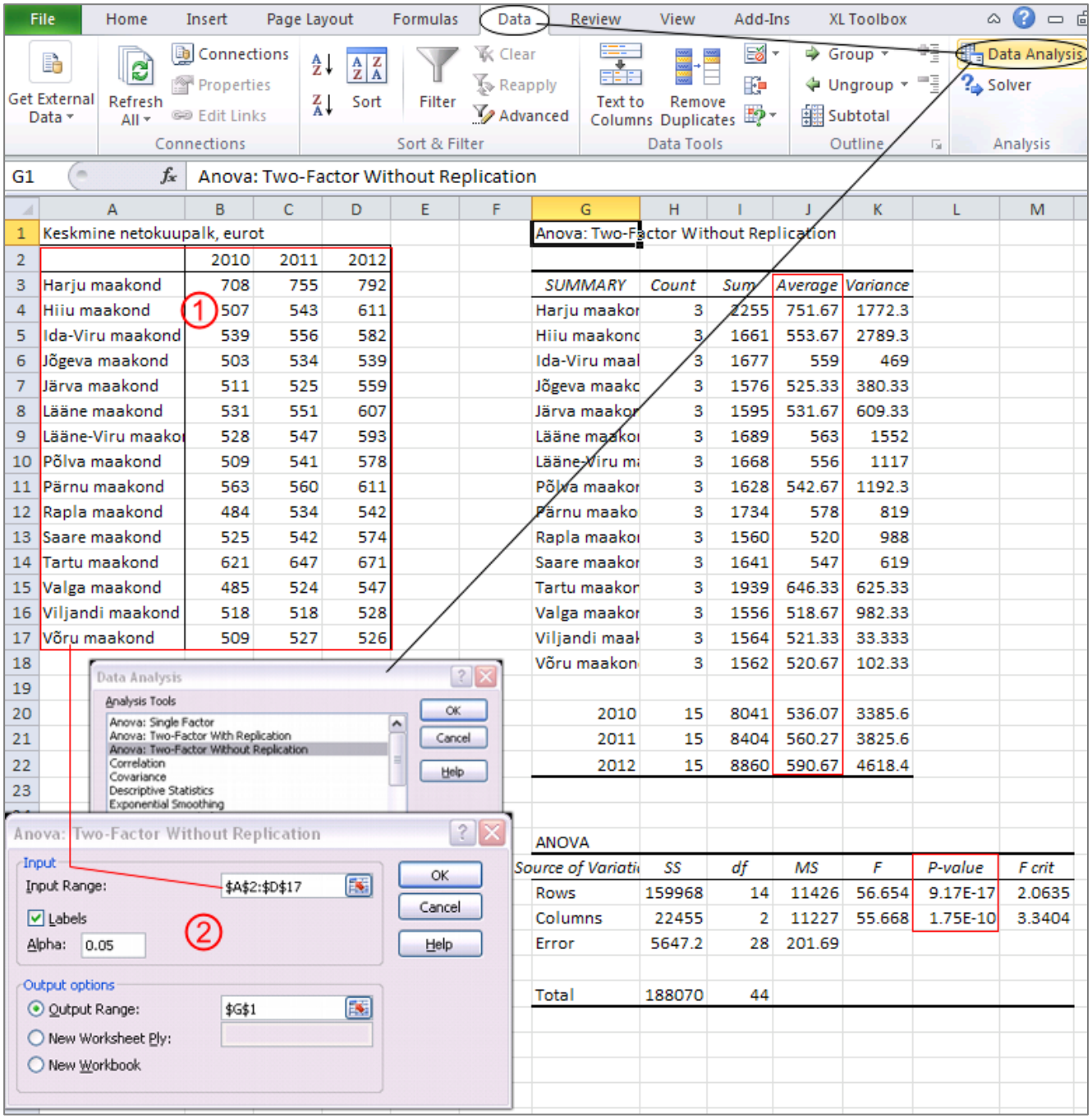
Kahefaktorilise kordusteta dispersioonanalüüsi teostamiseks Excelis tuleb (vt ka Joonis 68)

1. esitada analüüsitavad andmed risttabelina, mis on jagatud ridadeks ühe ja veergudeks teise faktori tasemete alusel ning kus igas lahtris paikneb täpselt üks uuritava tunnuse väärtus,
2. rakendada protseduuri *Anova: Two-Factor Without Replication* (*Data*-sakk -> *Data Analysis*), millele tuleb ette anda
   - sammul 1 loodud risttabel (*Input Range*),
   - infovõrreldavate gruppide nimede olemasolu kohta risttabeli esimeses reas ja veerus (st, et nimed peavad olemas olema kas mõlemat pidi või siis üldse puuduma; *Labels*),
   - olulisuse nivoo F-statistiku kriitilise väärtuse arvutamiseks (*Alpha*, vaikimisi 0,05),
   - tulemustabelite asukoht (*Output options*): samale töölehele (*Output Range*), uuele töölehele (*New Worksheet Ply*) või uude faili (*New Workbook*).

Joonisel 68 on näidatud kahefaktorilise kordusteta dispersioonanalüüsi teostamist protseduuriga *Anova: Two-Factor Without Replication*, testimaks keskmise netopalga erinevust maakondade ja aastate 2010-2012 vahel (andmed Statistikaameti kodulehelt http://www.stat.ee).

Tulemuseks on kaks tabelit, millest esimene sisaldab nii andmebaasi ridu (antud näites maakondi) kui ka veerge (antud näites aastaid) kirjeldavaid karakteristikuid. Neist oluliseimad on aritmeetilised keskmised, mille alusel saab vaadata, millised grupid (aastad ja maakonnad) omavahel enam ja mis suunas erinevad. Antud juhul on tulemused muidugi loomulikud – suurima keskmise netopalgaga aastatel 2010-2012 on olnud Harju maakonna töötajad (keskmine netopalk aastatel 2010-2012 751,67 eurot), millele järgneb Tartu maakond (646,33 eurot), madalaim keskmine netopalk on olnud Valga maakonnas (518,67 eurot). Uuritud aastate lõikes on palk suurenenud – kui aastal 2009 oli keskmine netopalk 536,07 eurot, siis aastal 2012 juba 590,67 eurot (mitte et see nüüd mingi eriti suur arv oleks …).

Dispersioonanalüüsi (ANOVA) tabelis vastab rida *Rows* maakonna mõjule ja rida *Columns* aasta mõjule (sest just niipidi oli algandmete tabel üles ehitatud). Mõlema faktori mõju on statistiliselt oluline ($p < 0{,}001$).



Joonis 68. Kahefaktorilise kordusteta dispersioonanalüüsi teostamine protseduuriga *Anova: Two-Factor Without Replication* ja selle tulemused, testimaks keskmise netopalga erinevust maakondade ja aastate 2010-2012 vahel (andmed Statistikaameti kodulehelt http://www.stat.ee).

Kahefaktoriline kordustega dispersioonanalüüs

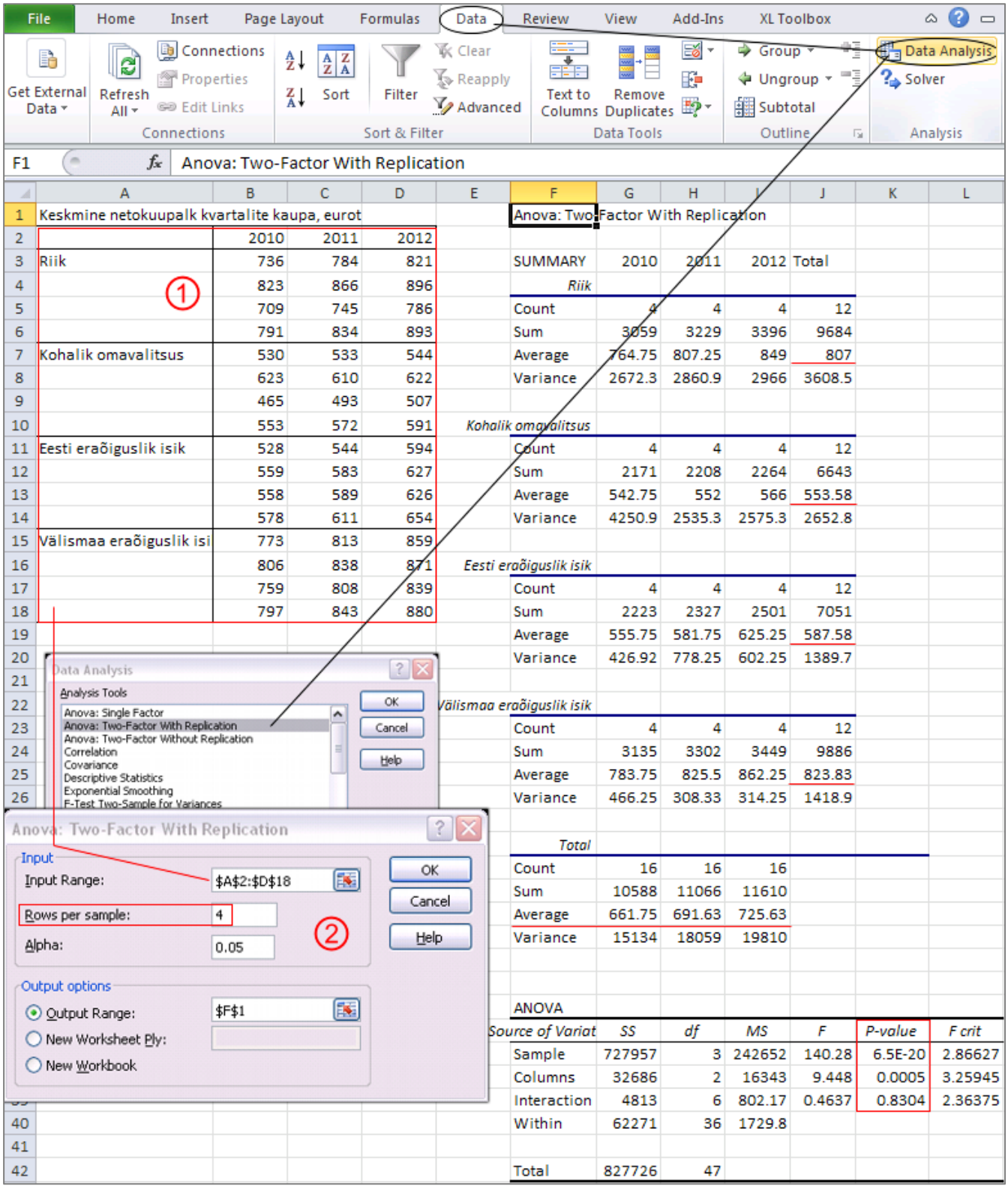
Kahefaktorilise kordustega dispersioonanalüüsi teostamiseks Excelis tuleb (vt ka Joonis 69)

1. esitada analüüsitavad andmed risttabelina, mis on jagatud ridadeks ühe ja veergudeks teise faktori tasemete alusel ning kus igale faktorite tasemete kombinatsionile vastab sama arv nö korduvaid mõõtmisi, mis paiknevad eri ridades faktorite samade tasemete sees,
2. rakendada protseduuri *Anova: Two-Factor With Replications* (*Data*-sakk -> *Data Analysis*), millele tuleb ette anda
   - sammul 1 loodud risttabel (*Input Range*),
   - mõõtmiste arv (kordsus) **reafaktori** taseme kohta (*Rows per sample*),
   - infovõrreldavate gruppide nimede olemasolu kohta risttabeli esimeses reas ja veerus (st, et nimed peavad olemas olema kas mõlemat pidi või siis üldse puuduma, seejuures võivad reafaktori tasemete nimed olla ka vaid iga taseme esimeses reas; *Labels*),
   - olulisuse nivoo F-statistiku kriitilise väärtuse arvutamiseks (*Alpha*, vaikimisi 0,05),
   - tulemustabelite asukoht (*Output options*): samale töölehele (*Output Range*), uuele töölehele (*New Worksheet Ply*) või uude faili (*New Workbook*).

Joonisel 69 on näidatud kahefaktorilise kordustega dispersioonanalüüsi teostamist protseduuriga *Anova: Two-Factor With Replications*, testimaks keskmise netopalga sõltuvust tööandja (omaniku) liigist ja aastast (andmed Statistikaameti kodulehelt http://www.stat.ee).

Tulemusena väljastab Excel kirjeldavate statistikute tabelid nii rea- kui ka veerufaktori kõigi tasemete kohta. Antud juhul on näha, et kõrgeim keskmine netopalk on olnud välismaa eraõiguslikust isikust omaniku puhul (823,83 eurot), palju ei jää maha keskmine netopalk ka riigi omandusega ettevõtetes – 807,00 eurot. Madalaim keskmine netopalk aastatel 2010-2012 oli kohalike omavalitsuste omanduses – 553,58 eurot. Aastate lõikes on palk järjest tõusnud.

Teisest Exceli poolt väljastatud tabelist on näha, et statistiliselt oluline erinevus keskmistes netopalkades on nii omanike (reafaktor, *Sample*) kui ka aastate (veerufaktor, *Columns*) vahel (mõlemal juhul $p < 0{,}001$). Küll ei ole statistiliselt oluline omaniku ja aasta koosmõju (*Interaction*, $p = 0{,}83$), st et omanike vaheline erinevus on samasugune sõltumata aastast nagu on ka aastate vaheline erinevus samasugune sõltumata omanikust.



Joonis 69. Kahefaktorilise kordustega dispersioonanalüüsi teostamine protseduuriga *Anova: Two-Factor With Replications* ja selle tulemused, testimaks keskmise netopalga sõltuvust tööandja (omaniku) liigist ja aastast (andmed Statistikaameti kodulehelt http://www.stat.ee).

9.3. *Post-hoc* testid

Juhul, kui dispersioonanalüüsi tulemus ütleb, et faktori mõju on statistiliselt oluline, kerkib sageli järgmine küsimus: millised võrreldud gruppidest on omavahel statistiliselt oluliselt erinevad? Taolisi, statistiliselt oluliseks osutunud dispersioonanalüüsi järgselt teostatud teste nimetatakse *post-hoc* testideks ning levinud on need eelkõige täpselt planeeritud ja läbi viidud katsete puhul (põldkatsed, söötmiseksperimendid jne).

Enam levinud *post-hoc* testid on Fisheri LSD (*Least Significant Difference*), Tukey, Scheffe ja Sidaki test. Mitte ühtki neist ei leidu Exceli statistilise analüüsi vahendite hulgas. Aga teades testide arvutuseeskirju ja vajadusel ka tabeleid kriitiliste väärtuste arvutamiseks, on nende testide teostamine Excelis võimalik.

Järgnevalt on näidatud, kuidas viia Excelis läbi kõige lihtsam *post-hoc* test – **Fisheri LSD test**.

Fisheri LSD testi puhul arvutatakse nn **vähim oluline vahe**, so vähim gruppide vaheline erinevus, mille võib veel lugeda statistiliselt oluliseks, valemist

$$LSD = t_{1-\alpha/2}(N-k)\sqrt{2MSE/n}\,,$$

kus k on võrreldavate gruppide arv (faktori tasemete arv), n on võrreldavate gruppide suurus, $N = nk$ on vaatluste arv, MSE on jääkidele vastav keskruut dispersioonanalüüsi tabelist ning $t_{1-\alpha/2}(N-k)$ on t-jaotuse kvantiil (protsendipunkt) kohal $1-\alpha/2$ vabadusastmete arvuga $N-k$, mis Excelis on leitav funktsiooniga T.INV.2T (argumentideks olulisuse nivoo α ja $N-k$).

Seega tuleb Fisheri LSD testi teostamiseks Excelis

1. viia läbi ühefaktoriline dispersioonanalüüs,
2. arvutada välja vähima olulise vahe *LSD* väärtus ja
3. võrrelda kõigi gruppide paarikaupa erinevusi *LSD*-ga:
   - kui kahe grupi keskmiste vaheline erinevus < *LSD*, siis ei ole võrreldud grupid statistiliselt oluliselt erinevad,
   - kui kahe grupi keskmiste vaheline erinevus ≥ *LSD*, siis on võrreldud grupid statistiliselt oluliselt erinevad.

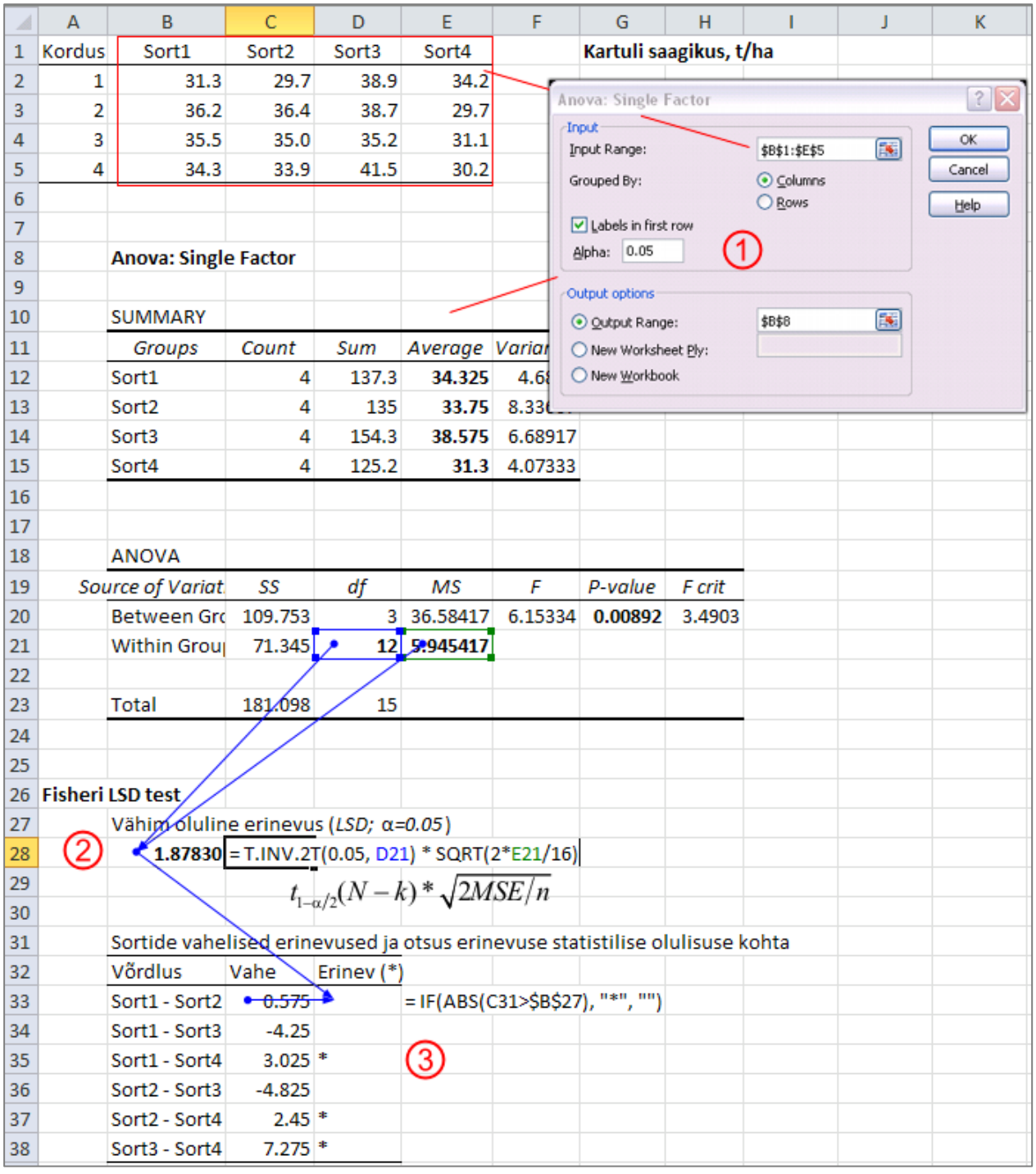
Joonisel 70 on näidatud nelja kartulisordi saagikuse võrdlemist ühefaktorilise dispersioon-analüüsiga ja selle järgselt Fisheri LSD testiga. Vähima olulise vahe *LSD* väärtuse arvutamiseks on kasutatud dispersioonanalüüsi tabelis sisalduvaid suuruseid, keskmiste paarikaupa võrdluste tarvis on tehtud abitabel, kuhu on esmalt välja arvutatud kõik keskmiste paarikaupa erinevused ning seejärel on neid erinevusi võrreldud vähima olulise vahe *LSD* väärtusega (kirjutades tärni statistiliselt oluliste erinevuste järele).

Märkus. Kui võrreldavad grupid on erineva suurusega, tuleb iga paariviisilise võrdluse tarvis arvutada oma *LSD* väärtus valemist

$$LSD = t_{1-\alpha/2}(N-k)\sqrt{MSE\Big/\left(\frac{1}{n_i}+\frac{1}{n_j}\right)},$$

kus n_i ja n_j on mõõtmiste arvud võrreldavates gruppides i ja j.

NB! Fisheri LSD test ei korrigeeri tulemusi mitmese testimise suhtes, st et kõigi paariviisiliste võrdluste peale kokku on tõenäosus, et mõni erinevus on ilmenud lihtsalt juhuslikult, suurem kui 0,05.



Joonis 70. Nelja kartulisordi saagikuse võrdlemist ühefaktorilise dispersioonanalüüsiga ja selle järgselt Fisheri LSD testiga.

10. Trikke ja nippe

10.1. Kavalad funktsioonid ja valemid

Tundes hästi MS Exceli funktsioone, on teinekord võimalik lahendada esmapilgul keeruka ja töömahukana näivaid ülesandeid vaid ühe arvutuskäsuga. Järgnevalt ongi esitatud mõningad kas Excelis juba olemasolevad või siis Exceli funktsioonide loogikat kasutavad valemid erinevate arvutuste teostamiseks.

NB! Kõik järgnevalt esitatud valemid on korrektsed inglise keele seadistustes Exceli puhul, eesti keele seadistustes Exceli puhul peab arvudes olema punkti asemel koma ja funktsioonide argumentide eraldajaks koma asemel semikoolon.

Tinglik keskmine, summa ja loendus

Exceli funktsioonid AVERAGEIF ja AVERAGEIFS, SUMIF ja SUMIFS ning COUNTIF ja COUNTIFS võimaldavad arvutada vastavalt keskmist, summat ning vaatluste arvu vaid teatud tingimustele vastavatest andmebaasi veergudest (või ridadest). Seejuures võib tingimus käia nii samades vaatlusalustes lahtrites paiknevate väärtuste kui ka teistes veergudes (või ridades) paiknevate väärtuste kohta. IF-lõpuga funktsioonid võimaldavad ette anda vaid ühe tingimuse, IFS-lõpuga funktsioonid aga kuni 127 tingimust.

Funktsioonide AVERAGEIF ja SUMIF süntaks on identne:

- esimese argumendina tuleb määrata lahtriblokk, mille kohta käib kontrollitav tingimus (võib olla sama, kui keskmise või summa arvutuste aluseks olev lahtriblokk; *Range*),
- teiseks argumendiks on kontrollitav tingimus (*Criteria*) – kas konkreetne väärtus, viide väärtust sisaldavale lahtrile või võrdlus mingi väärtusega (võrdlus väärtusega mingis lahtris ei ole võimalik, st, et kui näiteks lahtris A2 on väärtus 4, siis on mõeldavad tingimused kujul "=4", "=A2", ">4", aga mitte kujul ">A2"),
- kolmandaks argumendiks on arvutuste aluseks olev lahtriblokk (*Average_range* või *Sum_range*).

Funktsiooni COUNTIF puhul kolmandat argumenti, millega anda ette kokkuloetavaid väärtuseid sisaldav lahtriblokk, pole – kokku loetakse tingimusele vastavad väärtused samast, esimese argumendiga (*Range*) ette antud lahtriblokist.

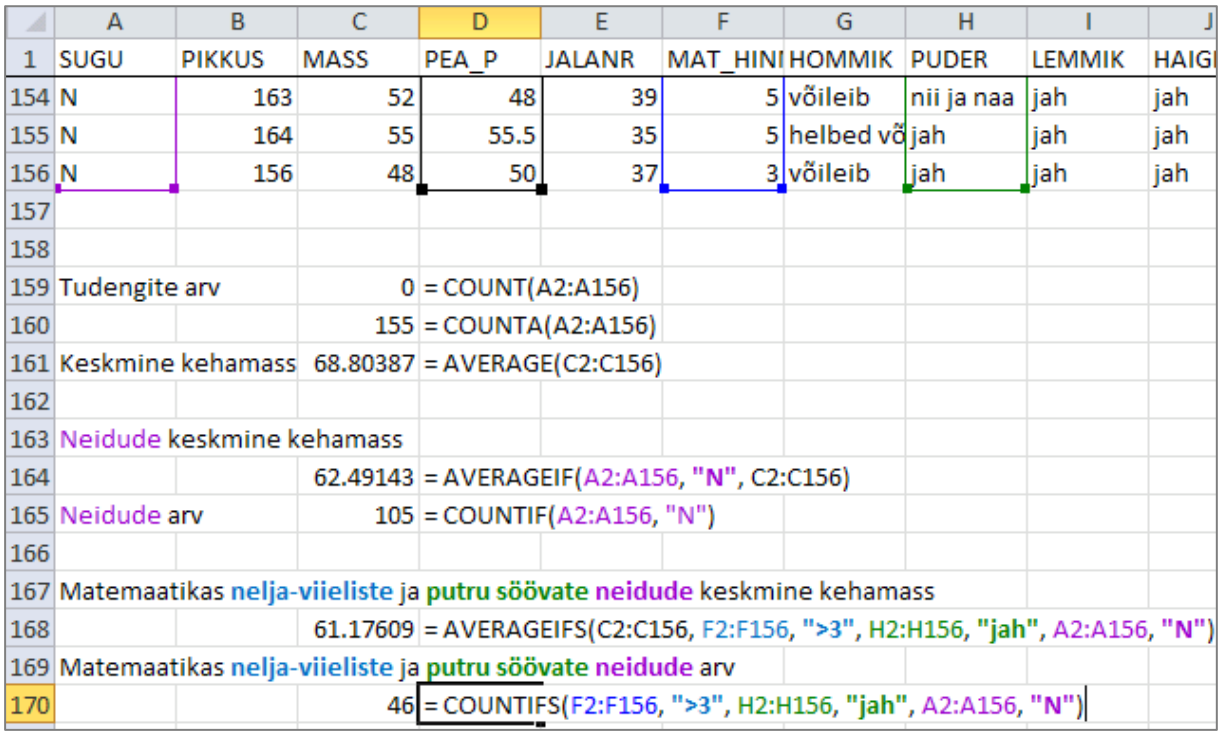
Funktsioonide AVERAGEIFS ja SUMIFS puhul on

- arvutuste aluseks olev lahtriblokk esimene argument (*Average_range* või *Sum_range*),
- millele järgnevad paarikaupa tingimuse aluseks olev lahtriblokk ja tingimus *Criteria_range1*, *Criteria1*, *Criteria_range2*, *Criteria2*, …).

Funktsiooni COUNTIFS puhul esimene arvutuste aluseks olev lahtriblokk puudub - kokku loetakse tingimus(t)ele vastavad väärtused esimese argumendiga (*Criteria_range1*) ette antud lahtriblokist.

NB! Erinevalt funktsioonist COUNT, mis loeb kokku vaid arvulised väärtused, toimivad funktsioonid COUNTIF ja COUNTIFS sama moodi nii arvuliste kui ka mittearvuliste väärtuste korral olles seega pigem funktsiooni COUNTA edasiarenduseks.

Jooonisel 71 on rakendatud tinglike arvkarakteristikute valemeid tudengite andmestiku näitel.



	A	B	C	D	E	F	G	H	I	J
1	SUGU	PIKKUS	MASS	PEA_P	JALANR	MAT_HINI	HOMMIK	PUDER	LEMMIK	HAIGI
154	N	163	52	48	39	5	võileib	nii ja naa	jah	jah
155	N	164	55	55.5	35	5	helbed võ	jah	jah	jah
156	N	156	48	50	37	3	võileib	jah	jah	jah
157										
158										
159	Tudengite arv		0	=COUNT(A2:A156)						
160			155	=COUNTA(A2:A156)						
161	Keskmine kehamass		68.80387	=AVERAGE(C2:C156)						
162										
163	Neidude keskmine kehamass									
164			62.49143	=AVERAGEIF(A2:A156, "N", C2:C156)						
165	Neidude arv		105	=COUNTIF(A2:A156, "N")						
166										
167	Matemaatikas nelja-viieliste ja putru söövate neidude keskmine kehamass									
168			61.17609	=AVERAGEIFS(C2:C156, F2:F156, ">3", H2:H156, "jah", A2:A156, "N")						
169	Matemaatikas nelja-viieliste ja putru söövate neidude arv									
170			46	=COUNTIFS(F2:F156, ">3", H2:H156, "jah", A2:A156, "N")						

Joonis 71. Tinglike arvkarakteristikute leidmine Excelis; tingimuse aluseks olevad lahtrid ja nendele järgnevad tingimused on esitatud sama värviga, arvutuste aluseks olevad lahtrid, kui need on eraldi argumendiks, on mustas kirjas.

Tinglikud arvkarakteristikud funktsioonide IF ja OR abil

Eelmises punktis käsitletud tinglike arvkarakteristikute funktsioonidel on mitmeid puuduseid:

- esiteks on olemas funktsioonid vaid keskmise, summa ja vaatluste arvu tarvis, aga mitte teiste, samuti sageli kasutatavate arvkarakteristikute jaoks,
- teiseks ei ole võimalik määrata tingimusi kujul „üks või teine“ – a'la leida keskmist pikkust tudengitel, kes kaaluvad alla 60 kg või üle 80 kg;
- kolmandaks ei saa tingimuste koostamisel kasutada teisi Exceli funktsioone.

Lahenduseks on funktsioonide IF ja OR kasutamine **massiivifunktsioonidena** statistika-funktsioonide siseselt. St, et esmalt valitakse välja vaid ette antud tingimustele vastavad andmebaasi read (või veerud) ja seejärel rakendatakse soovitud statistikafunktsiooni neile välja valitud ridades (või veergudes) paiknevatele väärtustele.

Näiteks tudengite andmebaasis on tudengite sugu määratud lahtrites A2:A156 paiknevate väärtustega M ja N ning tudengite kehamass on lahtrites C2:C156. Tütarlaste keskmine pikkus on siis leitav nii funktsioonidega

= AVERAGEIF(A2:A156, "N", C2:C156)

ja

= AVERAGEIFS(C2:C156, A2:A156, "N")

kui ka funktsiooniga

= AVERAGE(IF(A2:A156="N", C2:C156, "")).

NB! Et viimase funktsiooni näol on tegu massiivifunktsiooniga, tuleb selle rakendamiseks vajutada **Ctrl+Shift** ja **Enter**.

Viimast valemit modifitseerides on võimalik arvutada ka teisi tinglikke karakteristikuid ja testida hüpoteesegi (vt ka Joonis 72). Näiteks valem kujul

= **MEDIAN**(IF(A2:A156="N", C2:C156, "")).

annab tulemuseks neidude kehamassi mediaani, valem kujul

= **CORREL**(IF(A2:A156="N", C2:C156, ""), IF(A2:A156="N", B2:B156, ""))

annab tulemuseks neidude kehamassi ja pikkuse (veerus B) vahelise korrelatsioonikordaja, valem kujul

= **T.TEST**(IF(A2:A156="N", C2:C156, ""), IF(A2:A156="M", C2:C156, ""), 2, 2)

viib aga läbi teist tüüpi t-testi võrdlemaks neidude ja noormeeste keskmisi kehamasse.

Lisaks võib tingimus sisaldada funktsioone. Näiteks soovides arvutada keskmist kehamassi tudengitel, kes kaaluvad keskmisest enam, saab kasutada valemit kujul

=AVERAGE(IF(C2:C156>AVERAGE(C2:C156), C2:C156, "")),

keskmisest suurema pikkusega tudengite kehamassi mediaan on arvutatav aga valemist

=MEDIAN(IF(B2:B156>AVERAGE(B2:B156), C2:C156, "")).

Soovides aga rakendada mitut tingimust samaaegselt, näiteks arvutada matemaatikas nelja-viieliste ja putru söövate neidude keskmist kehamassi, saab seda teha eelmises alapunktis käsitletud valemiga

= AVERAGEIFS(C2:C156, F2:F156, ">3", H2:H156, "jah", A2:A156, "N"),

aga viimane ei ole üldistav teistele funktsioonidele ega võimalda kasutada tingimustes valemeid.

Alternatiiv on kasutada valemit, milles kõik samaaegselt kehtima pidavad tingimused on määratud üksteise sees paiknevate IF-lausetega:

= AVERAGE**(** IF**(**F2:F156>3, IF**(**H2:H156="jah", IF**(**A2:A156="N", C2:C156, ""**)**, ""**)**, ""**)** **)**

(tudengite matemaatika hinded paiknevad lahtrites F2:F156 ja info pudru söömise kohta lahtrites H2:H156).

Soovides arvutada keskmist pikkust tudengitel, kes kaaluvad alla 60 kg või üle 80 kg, st anda tingimust ette kujul „üks või teine“, saab seda teha OR-funktsiooniga IF-funktsiooni siseselt:

= AVERAGE(IF(OR(C2:C156<60, C2:C156>80), B2:B156, "")).

Ja viimaks, mitut IF-funktsiooni, OR-funktsioone nende siseselt ja lisaks ka valemitega ette antud tingimusi võib rakendada kõiki koos, saamaks suurest andmebaasist ilma mistahes sorteerimiste ja filtreerimisteta vaid ühe funktsiooniga kätte huvipakkuva arvkarakteristiku väärtust või hüpoteeside kontrolli tulemust. Näiteks alla 60 kg ja üle 80 kg kaaluvate matemaatikas keskmiselt parema hindega noormeeste pikkuse mediaan on leitav valemiga

= MEDIAN**(** IF**(** OR(C2:C156<60, C2:C156>80), IF**(**F2:F156>AVERAGE(F2:F156),
IF**(**A2:A156="M", B2:B156, ""**)**, ""**)**, ""**)** **)**.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	SUGU	PIKKUS	MASS	PEA_P	JALANR	MAT_HIN	HOMMIK	PUDER	LEMMIK	HAIGE	SPORT	AUTO	OLU
2	N	180	76	56	42	3	muu	jah	ei	ei	jah	jah	
3	N	178	65	56	39	4	ei söö tav	jah	jah	ei	jah	ei	0.2
4	M	177	70	57	42	3	võileib	nii ja naa	ei	jah	ei	ei	
171													
172	Neidude keskmine kehamass												
173			62.49143	=AVERAGEIF(A2:A156, "N", C2:C156)									
174			62.49143	=AVERAGEIFS(C2:C156, A2:A156, "N")									
175			62.49143	=AVERAGE(IF(A2:A156="N", C2:C156, ""))									
176													
177	Neidude kehamassi ja pikkuse vaheline korrelatsioonikordaja												
178			0.506151	=CORREL(IF(A2:A156="N", C2:C156, ""), IF(A2:A156="N", B2:B156, ""))									
179													
180	Neidude ja noormeeste keskmiste kehamasside võrdlus t-testiga												
181			5.67E-18	=T.TEST(IF(A2:A156="N", C2:C156, ""), IF(A2:A156="M", C2:C156, ""), 2, 2)									
182													
183	Keskmine kehamass keskmisest enam kaaluvatel tudengitel												
184			82.45455	=AVERAGE(IF(C2:C156>AVERAGE(C2:C156), C2:C156, ""))									
185	Keskmisest suurema pikkusega tudengite kehamassi mediaan												
186			75	=MEDIAN(IF(B2:B156>AVERAGE(B2:B156), C2:C156, ""))									
187													
188	Matemaatikas nelja-viieliste ja putru söövate neidude keskmine kehamass												
189		61.17609	=AVERAGEIFS(C2:C156, F2:F156, ">3", H2:H156, "jah", A2:A156, "N")										
190		61.17609	=AVERAGE(IF(F2:F156>3, IF(H2:H156="jah", IF(A2:A156="N", C2:C156, ""), ""), ""))										
191													
192	Alla 60 kg ja üle 80 kg kaaluvate tudengite keskmine pikkus												
193		173.3839	=AVERAGE(IF(OR(C2:C156<60, C2:C156>80), B2:B156, ""))										
194													
195	Alla 60 kg ja üle 80 kg kaaluvate matemaatikas keskmiselt parema hindega noormeeste pikkuse mediaan												
196		184.5	=MEDIAN(IF(OR(C2:C156<60, C2:C156>80), IF(F2:F156>AVERAGE(F2:F156), IF(A2:A156="M", B2:B156, ""), ""), ""))										

Joonis 72. Tinglike karakteristikute arvutamine Excelis funktsiooni IF abil.

Kaalutud keskmine

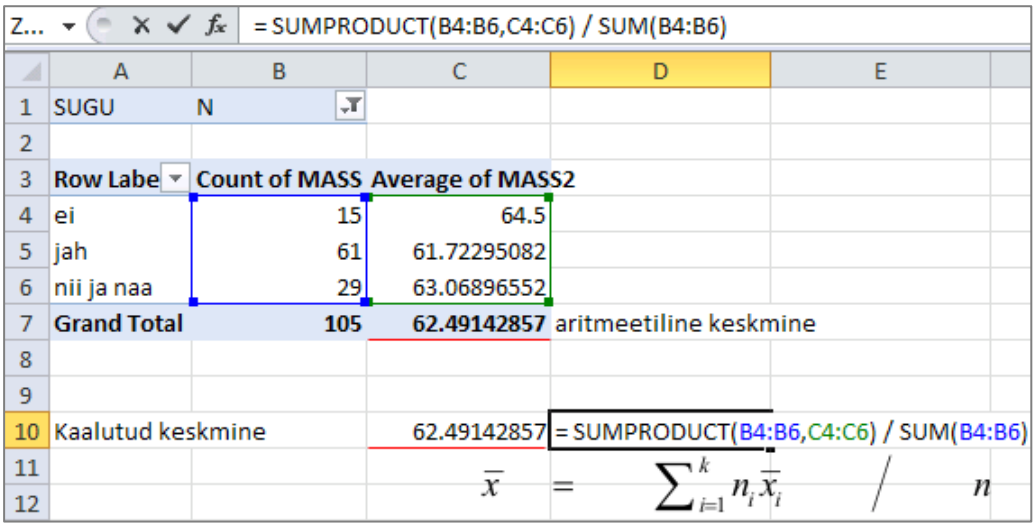
Mõnikord on vaja leida mingi näitaja keskmist väärtust kogu andmebaasis olukorras, kus seda andmebaasi ennast tegelikult kasutada pole, küll aga on olemas tabel keskmiste väärtustega mingites gruppides. Kui on teada ka gruppide suurused, on kogu andmebaasi keskmine arvutatav kaalutud keskmisena valemist

$$\bar{x} = \left(\sum_{i=1}^{k} n_i \bar{x}_i\right) \Big/ n ,$$

kus $\bar{x}$ ja $\bar{x}_i$ on vastavalt kogu andmebaasi keskmine ja *i*. grupi keskmine, n ja n_i on vastavalt kogu andmebaasi suurus ja *i*. grupi suurus ning k on gruppide arv.

Excelis on kaalutud keskmist mugav arvutada funktsiooni SUMPRODUCT abil.

Joonisel 73 on näidatud esimese kursuse neidude keskmise kehamassi arvutamist kaalutud keskmisena, võttes aluseks tudengite arvud ja keskmised kehamassid mannapudru söömise ja mittesöömise alusel moodustatud gruppides.



Joonis 73. Kaalutud keskmise arvutamine MS Excelis.

Erinevate väärtuste arvu leidmine

Korduvate väärtustega andmetabeli puhul võib sageli tekkida küsimus, kui palju on üldse erinevaid mõõdetud indiviide või kui palju on erinevaid mõõtmistulemusi. Üks võimalus on tekitada korduvate väärtusteta andmetabel (kas siis *PivotTable* või *Data*-sakil leiduva käsu *Remove Duplicates* abil) ja leida väärtuste arv selles. Teine võimalus on kasutada järgmist kavalat valemit.

Oletame, et väärtused, mille hulgast on vaja kokku lugeda unikaalsed, paiknevad lahtrites A2 kuni A20. Valem, mis arvutab (justnimelt arvutab, mitte ei loe kokku) erinevate väärtuste arvu, on siis kujul

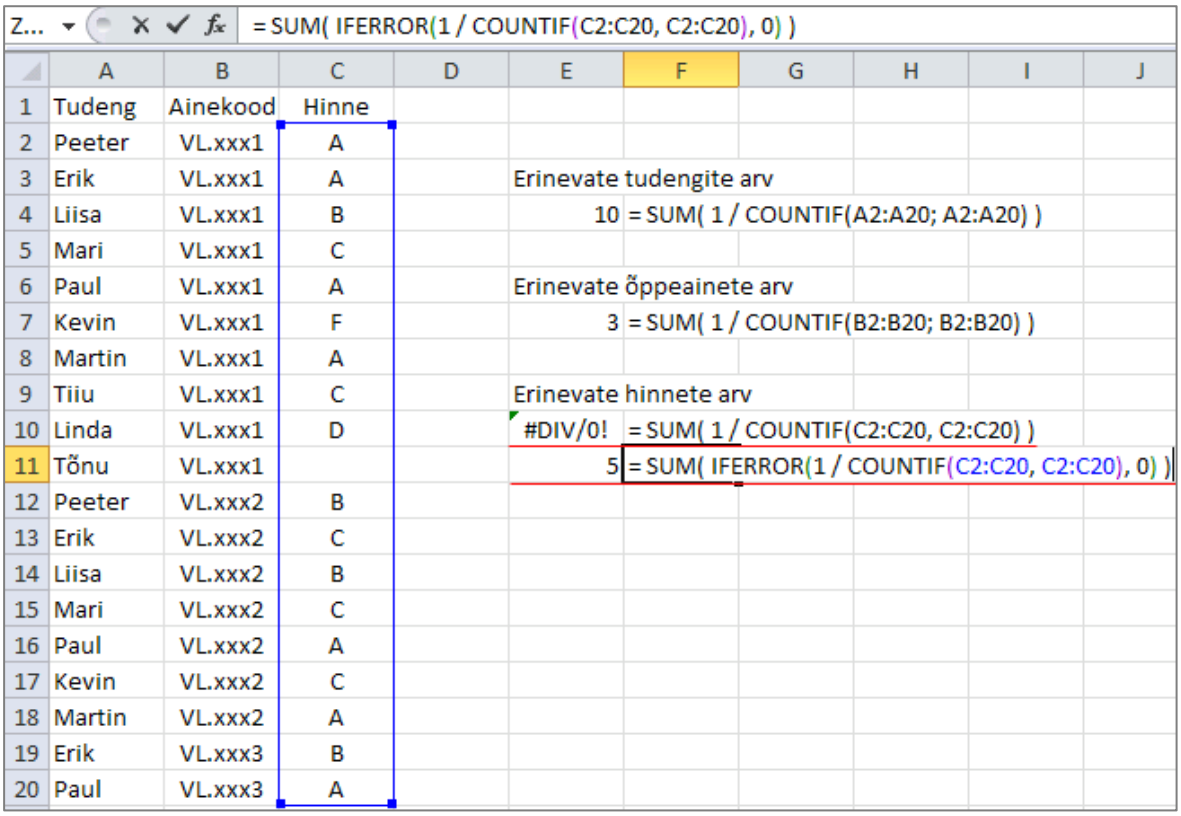
= SUM(1 / COUNTIF(A2:A20, A2:A20))

NB! Tegu on massiivifunktsiooniga, st, et valemi rakendamiseks tuleb vajutada korraga kolme klahvi: **Ctrl+Shift** ja **Enter**.

Kui lahtriblokk, milles paiknevate unikaalsete väärtuste arvu arvutatakse, sisaldab puuduvaid väärtuseid, lõpeb eelneva valemi rakendamine veateatega #DIV/0!. Lahenduseks on kasutada täendavalt funktsiooni IFERROR, mis puuduvate väärtuste korral võtab nende arvuks lihtsalt nulli (vt ka Joonis 74):

= SUM(IFERROR(1 / COUNTIF(A2:A20, A2:A20), 0))

ja siis Ctrl+Shift ja Enter.



	A	B	C
1	Tudeng	Ainekood	Hinne
2	Peeter	VL.xxx1	A
3	Erik	VL.xxx1	A
4	Liisa	VL.xxx1	B
5	Mari	VL.xxx1	C
6	Paul	VL.xxx1	A
7	Kevin	VL.xxx1	F
8	Martin	VL.xxx1	A
9	Tiiu	VL.xxx1	C
10	Linda	VL.xxx1	D
11	Tõnu	VL.xxx1	
12	Peeter	VL.xxx2	B
13	Erik	VL.xxx2	C
14	Liisa	VL.xxx2	B
15	Mari	VL.xxx2	C
16	Paul	VL.xxx2	A
17	Kevin	VL.xxx2	C
18	Martin	VL.xxx2	A
19	Erik	VL.xxx3	B
20	Paul	VL.xxx3	A

Joonis 74. Unikaalsete väärtuste kokku lugemine Excelis. **NB!** Valemi rakendamiseks tuleb korraga vajutada kolme klahvi: Ctrl+Shift jaEnter.

Mittelineaarne regressioonanalüüs funktsiooniga LINEST

Regressioonivõrrandeid, mis on argumentide suhtes lineaarsed, aga mille argumendid ise on argumenttunnuse mittelineaarsed funktsioonid – näiteks kolmandat järku polünoom

$$y = a + b_1*x + b_2*x^2 + b_3*x^3$$

või astmefunktsioon kujul

$$y = a + b * x^{1,5} = a + b\sqrt[3]{x^2}$$

või logaritmfunktsioon kujul

$$y = a + b*\ln(x)$$

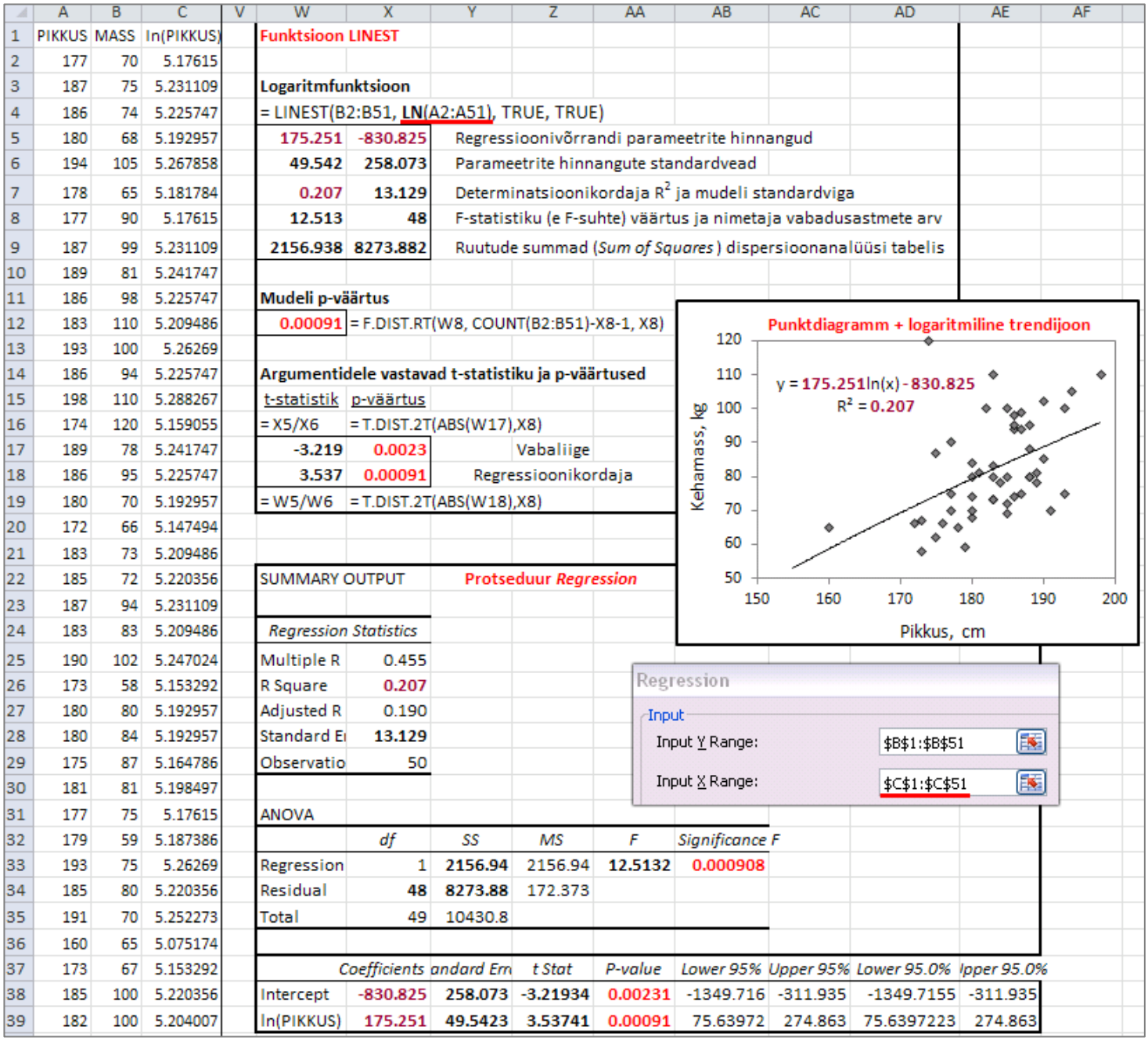
– on Excelis võimalik hinnata

- protseduuriga *Regression*, aga seda **eeldusel**, et kõik argumentide mittelineaarsed funktsioonid on eelnevalt töölehele üksteise kõrvale välja arvutatud, ja
- graafiliselt punktdiagrammi ja sellele lisatud sobiva trendijoone abil – aga seda vaid mõnede spesiifiliste funktsioonide puhul (ülaltooduist on punktdiagrammile lisatav kolmandat järku polünoom ja logaritmfunktsioon) ning ilma võimaluseta hinnata parameetrite hinnangute täpsust ja statistilist olulisust.

Joonisel 75 on näidatud noormeeste kehamassi prognoosimine pikkuse alusel logaritm-funktsiooniga kasutades nii funktsiooni LINEST, protseduuri *Regression* kui ka graafilist lahendamist – tulemused on identsed, ainult sarnaselt eelnevalt vaadatud logaritm-funktsioonile LINEST kujul

= LINEST(B2:B51, **LN(**A2:A51**)**, TRUE, TRUE)

- ei vaja erinevalt protseduurist *Regression* lahtrites A2:A51 paiknevate pikkuste logaritmi eelnevat välja arvutamist ning
- erinevalt graafilisest lahendusest on väljund rikkalikum, võimaldades täiendavalt testida ka hüpoteese regressioonivõrrandi statistilise olulisuse kohta (vastava teooria kohta vt pt 7.3 funktsiooni LINEST alapunkti).



Joonis 75. Noormeeste kehamassi prognoosimine pikkuse alusel logaritmfunktsiooniga kasutades funktsiooni LINEST, protseduuri *Regression* ja graafilist lahendamist. Tumepunases paksus kirjas on kõigil kolmel meetodil hinnatavad parameetrid, mustas paksus kirjas vaid funktsiooni LINEST ja protseduuri *Regression* väljundis sisalduvad parameetrid, helepunases paksus kirjas on p-väärtused, mis sisalduvad protseduuri *Regression* väljundis, aga on arvutatavad ka funktsiooni LINEST väljastatavate suuruste alusel. **NB!** Erinevalt protseduurist *Regression* ei vaja funktsioon LINEST ja graafiline lahendamine tudengite pikkuse logaritmi välja arvutamist andmetabelisse.

Joonisel 76 on näidatud noormeeste kehamassi prognoosimine pikkuse alusel kuup-polünoomiga kasutades nii funktsiooni LINEST, protseduuri *Regression* kui ka graafilist lahendamist – tulemused on identsed, ainult funktsioonile LINEST kujul

= LINEST(D2:D51, **A2:A51^{1,2,3}**, TRUE, TRUE)

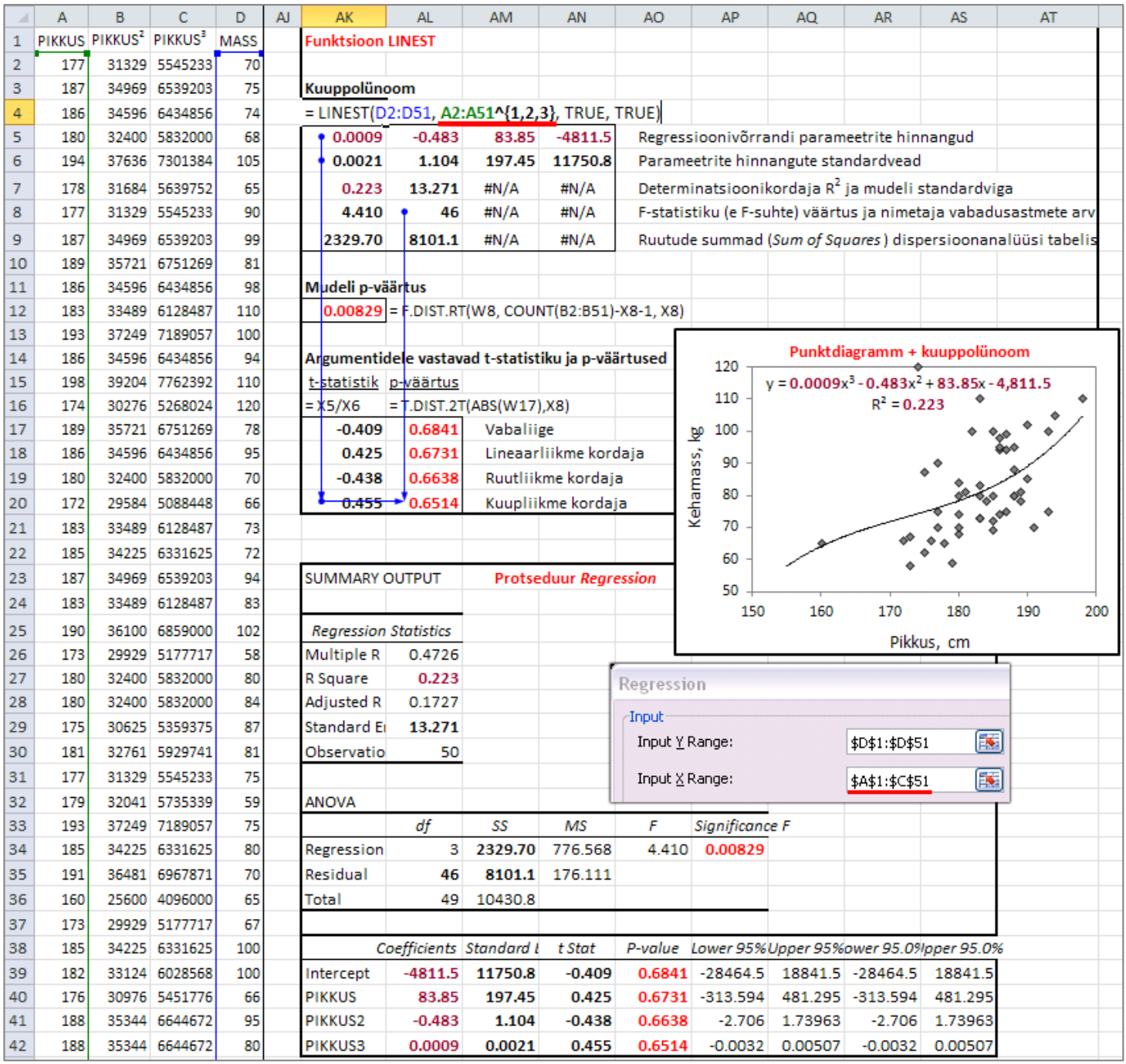
ei ole erinevalt protseduurist *Regression* vaja eraldi välja arvutada pikkuse ruutu ja kuupi ning erinevalt graafilisest lahendusest on väljund rikkalikum.

PS. Mingeid sisulisi järeldusi antud ülesande puhul teha ei maksa, sest mudel tervikuna on küll statistiliselt oluline (p = 0,008), aga ükski mudeli liige eraldi võetuna statistiliselt oluline pole (kõigi mudeli parameetrite puhul $p > 0{,}6$) ning ega varem vaadatud mudelitega (logaritmfunktsioon või peatükis 7.3 käsitletud lineaarne funktsioon) võrreldes prognoosi täpsus (R^2 ja mudeli standardviga) ka paremad pole – seega on antud mudel ilmselgelt liiga keerukas. Siinkohal on see ära toodud lihtsalt illustreerimaks funktsiooni LINEST võimalusi.

Astmefunktsiooni $y = a + b * x^{1,5} = a + b\sqrt[3]{x^2}$ parameetrid on funktsiooniga LINEST hinnatavad kujul

= LINEST(D2:D51, **A2:A51^1.5**, TRUE, TRUE)

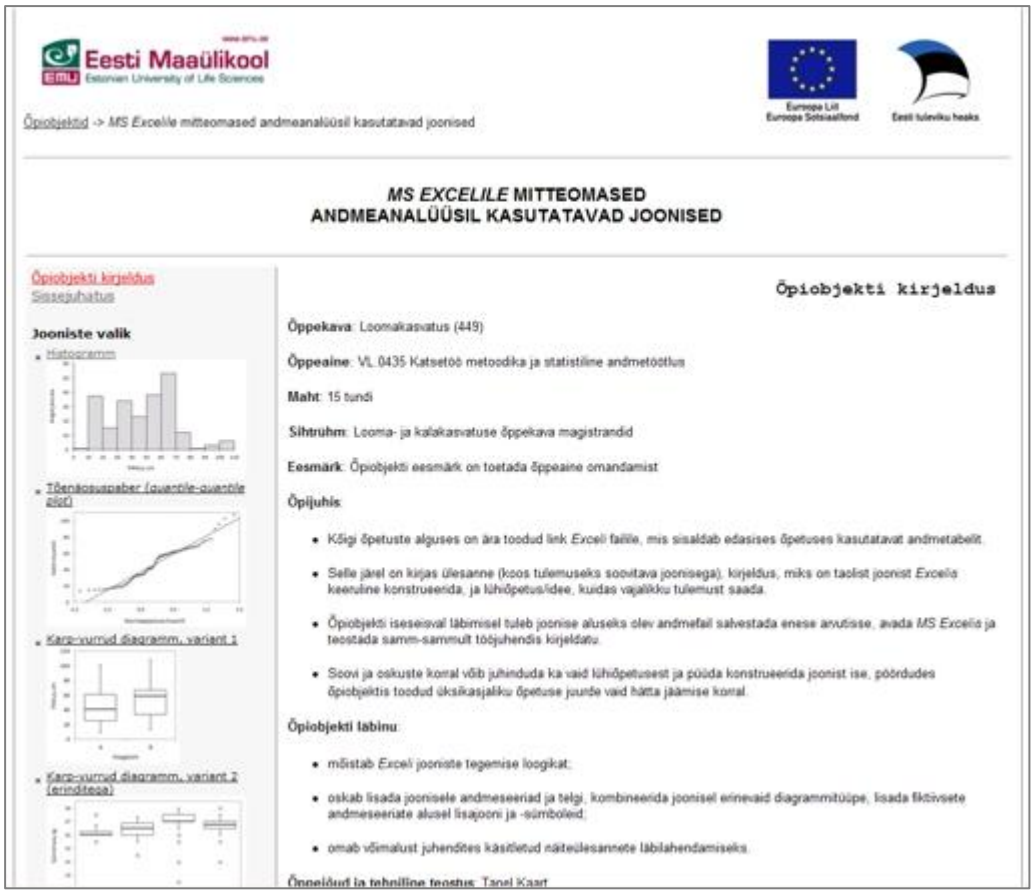
(funktsioontunnuse *y* väärtused paiknevad lahtrites D2:D51 ja argumenttunnuse *x* väärtused lahtrites A2:A59; **NB!** eesti keele seadistuses Exceli puhul peab arvus 1.5 punkti asemel olema koma ja funktsiooni argumentide eraldajaks koma asemel semikoolon).



Joonis 76. Noormeeste kehamassi prognoosimine pikkuse alusel kuuppolünoomiga kasutades funktsiooni LINEST, protseduuri *Regression* ja graafilist lahendamist. Tumepunases paksus kirjas on kõigil kolmel meetodil hinnatavad parameetrid, mustas paksus kirjas vaid funktsiooni LINEST ja protseduuri *Regression* väljundis sisalduvad parameetrid, helepunases paksus kirjas on p-väärtused, mis sisalduvad protseduuri *Regression* väljundis, aga on arvutatavad ka funktsiooni LINEST väljastatavate suuruste alusel. **NB!** Erinevalt protseduurist *Regression* ei vaja funktsioon LINEST ja graafiline lahendamine tudengite pikkuse ruudu ja kuubi välja arvutamist andmetabeli eraldi veergudesse.

10.2. Excelile mitteomased joonised

Valik näiteid ja õpetusi, kuidas konstrueerida MS Excelis mitte kõige igapäevasemaid, aga samas nii tudengite töödesse, projektide aruannetesse kui ka teadusartiklitesse sobivaid jooniseid, leidub õpiobjektis "MS Excelile mitteomased andmeanalüüsil kasutatavad joonised" (http://ph.emu.ee/~ktanel/joonised_excelis/).



10.3. Andmeanalüüsil kasutatavad lisamoodulid

Tänu MS Exceli ülilaialdasele kasutatavusele on sellele loodud hulk statistilisi analüüse teostavaid lisasid. Mahukamaid ja kasutajasõbralikumaid neist müüakse suure raha eest, vt näiteks

- XLSTAT (http://www.xlstat.com/),
- SigmaXL (http://www.sigmaxl.com/SigmaXL.shtml),
- statistiXL (http://www.statistixl.com/)

Tasuta lisamooduleid mõne spetsiifilise analüüsi tarvis võib otsida kas lihtsalt googeldades (a'la „*Tukey test in Excel*") võis siis mõnelt Exceli lisamooduleid koondavalt lehelt, näiteks

- http://www.dmoz.org/Science/Math/Statistics/Software/Excel_Add-In/,
- http://www.skilledup.com/learn/business-entrepreneurship/mostly-free-excel-add-ins/,
- http://www.mathtools.net/Excel/Statistics/.

Kuigi mina kasutan igapäevaseks andmeanalüüsiks statistikapakette SAS ja R, kulub ligikaudu 70% ajast siiski Excelis, sest sageli on andmete haldamiseks, samuti esmaste statistiliste analüüside teostamiseks või jooniste konstrueerimiseks lihtsaim ja kiireim variant kasutada Excelit. Sestap on mul töö hõlbustamiseks paigaldatud mitmeid tasuta lisamooduleid:

- „Real Statistics" (http://www.real-statistics.com/), mis sisaldab suurt hulka protseduure tava-Excelis puudu olevate andmeanalüüsimeetodite rakendamiseks – näiteks mitte-parameetrilised testid kahe grupi keskmiste võrdlemiseks, dispersioonanalüüsi järgsed *post-hoc* testid, logistiline regressioon jne –, aga ka mitmeid lisafunktsioone;
- „Daniel's XL Toolbox" (http://xltoolbox.sourceforge.net/), mis lisaks võimalusele teostada näiteks regressioonanalüüsi ilma puuduvate väärtustega ridu eemaldamata või dispersioonanalüüsi selleks spetsiifilist tabelit koostamata, võimaldab konstrueerida väga mitmesuguseid jooniseid ning eksportida neid teadusartiklitesse sobivatesse tiff-, png- ja emf-vormingutesse;
- „Kahe üldkogumi võrdlus" (http://ph.emu.ee/~ktanel/excel_addins/) – käesolevagi materjali peatükis 6.3 käsitletud Anu Iheri poolt 2005. aastal Tartu Ülikooli matemaatilise statistika instituudis bakalaureuse töö „Olulisemad kahe üldkogumi võrdlemise testid ja MS Excel'i moodul nende läbiviimiseks" osana valminud Exceli lisamoodul kahe grupi võrdlemisel kasutatavate parameetriliste ja mitteparameetriliste testidega;
- XY Chart Labeler (http://www.appspro.com/Utilities/ChartLabeler.htm), mis võimaldab lisada ja ümber paigutada joonistel kõikvõimalikke märgendeid (kasulik moodul atraktiivsete ja informatiivsete jooniste genereerimiseks);
- „Better Histogram" (http://www.treeplan.com/download-free-better-histogram-add-in.htm) – lihtne vahend teaduslikult korrektse histogrammi ja selle aluseks oleva sagedustabeli genereerimiseks.

Asjatundlikku Exceli kasutamist soovides,
Tanel Kaart