Katseandmete analüüs

Õppematerjal esimeseks kodutööks: Excel

Käesolev õppematerjal tutvustab Exceli võimalusi esimese kodutöö ülesannete lahendamisel. Eeldatakse teatavat (vähemasti põgusat) kokkupuudet Exceli või muu tabelarvutusprogrammiga. Juhised on koostatud pidades silmas versiooni Excel 2013, aga ilmselt kehtib kõigile versioonidele alates 2010-st. Väga tugev ja soe soovitus on kasutada Google otsingut kõigi probleemide puhul nüüd ja edaspidi. Sealhulgas üks kasulik tulemus on näiteks Maaülikooli dotsent Tanel Kaarti koostatud õppematerjal aadressil <http://ph.emu.ee/~ktanel/andmeanalyys_excelis/index.php>.

Sisukord

[Lahtrid ja viited 3](#_Toc21042110)

[Keskmine, standardhälve ja teised funktsioonid 4](#_Toc21042111)

[Logaritmteisendus 5](#_Toc21042112)

[Tsentreerimine 5](#_Toc21042113)

[Analysis ToolPack 7](#_Toc21042114)

[Sagedusjaotus ja histogramm 9](#_Toc21042115)

[T-test 10](#_Toc21042116)

[Ühefaktoriline ANOVA 12](#_Toc21042117)

[Korrelatsioonanalüüs 13](#_Toc21042118)

[Hajusdiagramm 13](#_Toc21042119)

[Lineaarne regressioon 15](#_Toc21042120)

[*Analysis ToolPack*-iga 15](#_Toc21042121)

[Hajusdiagramm regressioonsirgega 16](#_Toc21042122)

[Kahefaktoriline ANOVA 18](#_Toc21042123)

[Graafik keskmiste võrdlemiseks 19](#_Toc21042124)

[Lisamaterjalid 21](#_Toc21042125)

[Sagedustabel ja histogramm funktsioonidega 21](#_Toc21042126)

[T-test funktsioonidega 22](#_Toc21042127)

[Sõltuvate valimite t-test lisapaketiga 23](#_Toc21042128)

[Lineaarne regressioon funktsioonidega 24](#_Toc21042129)

[Mitmene lineaarne regressioon lisapaketiga 24](#_Toc21042130)

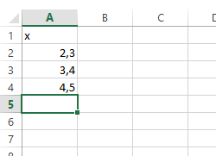
[Karpdiagramm 25](#_Toc21042131)

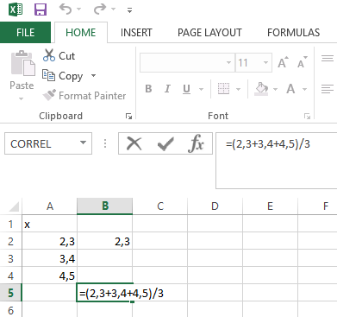
[Mitteparameetriline ANOVA 28](#_Toc21042132)

[Spearmani korrelatsioonikordaja 29](#_Toc21042133)

# Lahtrid ja viited

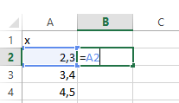
Usutavasti on lugeja kokku puutunud vähemalt mingisuguse tabelarvutusprogrammiga. Meeldetuletuseks mainime ära mõned vajalikud tehnilised nüansid, milleta järgnevas hakkama ei saa.

Enam-vähem kõik, mis Excelisse sisestada on võimalik, on organiseeritud tabeli lahtritesse. Tabeli read on nummerdatud (1, 2, 3, …) ja veerge tähistavad tähed (A, B, C, …). Näiteks tibatilluke andmestik:

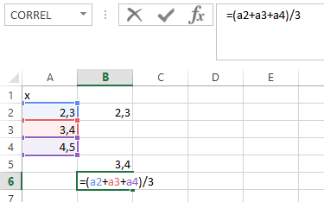
Tabelarvutusprogrammid oskavad arvutada. Selleks tuleb lahtrisse trükkimist alustada võrdusmärgiga =, pärast võrdusmärki tuleb aga trükkida arvutatav avaldis. Kui avaldis on trükitud, tuleb klaviatuuril vajutada Enter. Pane tähele, et pärast Enteri vajutamist kaob lahtrist avaldis ära ja selle asemele ilmub arvutuse tulemus. Näiteks eelneva andmestiku keskmise võib arvutada nii:

=(2,3+3,4+4,5)/3

Pane veel tähele, et parajasti märgitud lahtris sisalduvat avaldist on alati näha ka tabeli kohal.

Excelis on võimalik kirjutada avaldisi, mis sõltuvad teistest lahtritest. Selle sõltuvuse väljendamiseks tuleb vajaliku arvu või teksti asemel avaldisse kirjutada selle lahtri aadress (ehk viide), kust võtta see vajalik arv või tekst. Lahtri viide koosneb vastava veeru tähest ja rea numbrist, näiteks veeru A teises reas asuv lahter on A2 (suur- ja väiketähtedel vahet pole, võib trükkida ka a2). Kõige lihtsamal juhul saab lihtsalt näidata ühte asja kahes kohas, kui näiteks lahtrisse B2 kirjutada =A2 (ja ära unusta Enter vajutada, kui see lihtne avaldis on trükitud).

Kui nüüd lahtri A2 sisu muuta, siis muutub automaatselt ka lahtri B2 sisu.

Eeltoodud kolme arvu keskmise arvutamine, kasutades viiteid vastavatele lahtritele, näeb välja nõnda: =(a2+a3+a4)/3

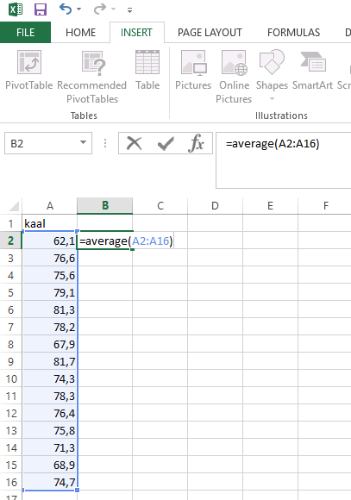
Kui nüüd on vaja andmeid parandada või muul põhjusel muuta, siis viiteid kasutavad keskmised (ja kõik muud arvutised) arvutavad ennast ise automaatselt ümber.

# Keskmine, standardhälve ja teised funktsioonid

Suur osa andmetöötlusest Excelis tugineb funktsioonidele. Exceli funktsiooni kasutamiseks tuleb hiirega klõpsata sellel lahtril, millesse tahetakse funktsiooni väärtust saada, ja sinna lahtrisse kirjutada kolm osa: alguses võrdusmärk, siis funktsiooni nimi, ja viimasena sulgude sees lähteandmed. Lõpuks tuleks vajutada Enter.

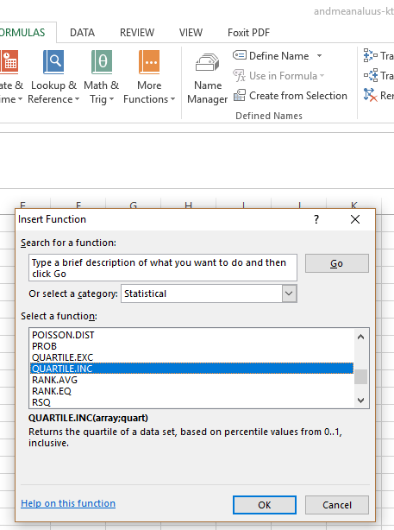
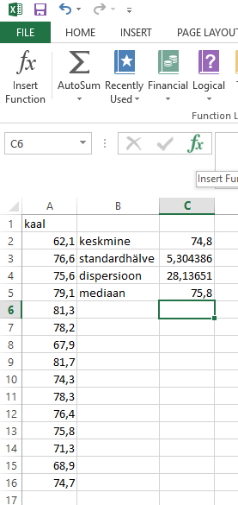
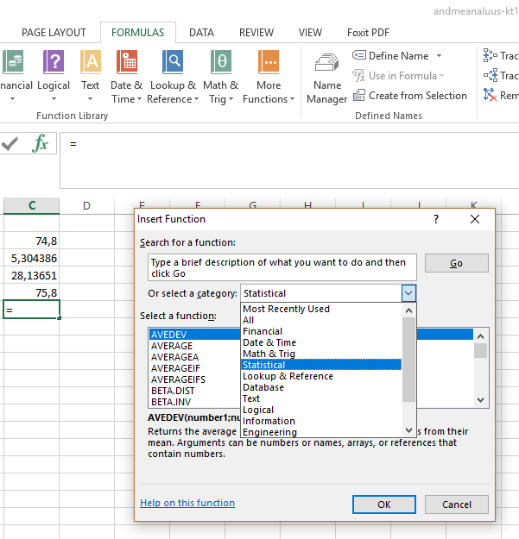
Lähteandmete endi asemel on mõistlik kasutada viidet neile. Õnneks ei ole siiski tarvis üksikute lahtrite viiteid ükshaaval kirja panna, vaid on võimalik viidata kõigile neile korraga. Ühes tulbas olevatele andmetele viitamiseks tuleb kirjutada ülemise lahtri viide, koolon, ja alumise lahtri viide. Sarnaselt saab viidata ka ühes reas või ristkülikus olevatele andmetele, mispuhul tuleb siis märkida ülemise vasakpoolse lahtri viide, koolon, ja alumise parempoolse lahtri viide. Näiteks mingite arvude aritmeetilise keskmise saab arvutada nii:

=AVERAGE(A2:A16)

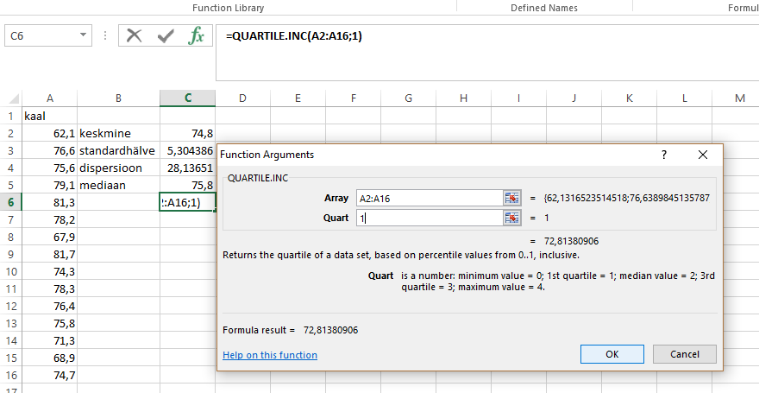


Lähteandmete vahemiku viite võib sulgudesse ise kirjutada või ka hiirega ära märkida, kui kirjutamine on sinnamaani jõudnud.

Standardhälve tuleks arvutada funktsiooniga STDEV.S, dispersioon funktsiooniga VAR.S. Kui on vaja kasutada mingit tundmatu nimega funktsiooni, võib uurida kõigi funktsioonide loendit (või Google-it!). Näiteks 25% kvartiili arvutava funktsiooni otsimine võib käia nõnda:



Seejärel avaneb dialoogiaken, mille kaudu saab funktsiooni lähteandmed kenasti sisestada, näha on ka nende lähteandmete nõutud vormi seletused. Näiteks 25% kvartiili saamiseks tuleb kirjutada 1.

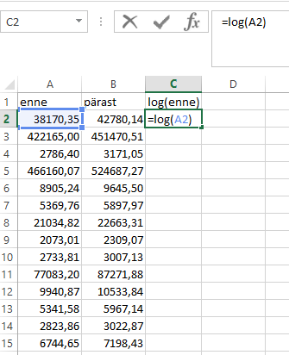


Excel töötab ka kalkulaatorina. Seda asjaolu saab kasutada ka siis, kui täpselt õiget funktsiooni leida pole õnnestunud. Näiteks variatsioonikordaja arvutamiseks võib kirjutada lahtrisse standardhälbe ja keskmise jagatise:

=STDEV.S(A2:A16)/AVERAGE(A2:A16)

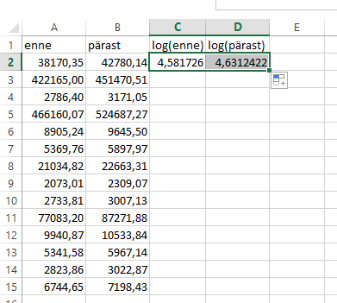
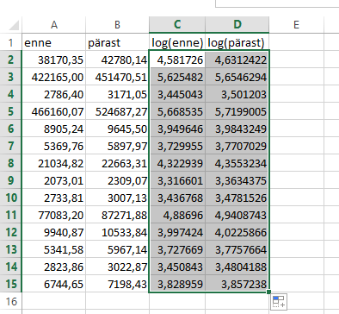
## Logaritmteisendus

Logaritmteisendus, nagu ka teised teisendused, seisnevad ühe ja sama funktsiooni rakendamises paljudele vaatlustele. Ühele vaatlusele me logaritmfunktsiooni rakendada ilmselt oskame:

=LOG(A2)

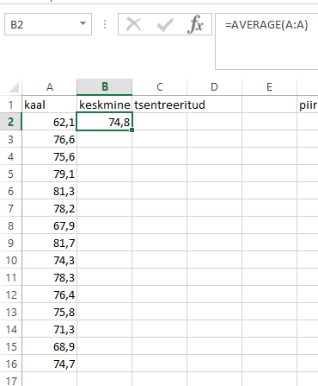
Kuidas aga logaritmida ära kõik need andmed kahes veerus? Muidugi võib logaritmfunktsiooni trükkida uuesti käsitsi 28 korda, aga õnneks on ka lihtsam lahendus. Nimelt saab funktsiooni juba sisaldava lahtri alumisest paremast nurgast hiirega kinni võtta ja lahtris leiduvat sisu kopeerida üle mitme lahtri, kusjuures funktsiooni argument tuleb kaasa. Selleks tuleb hiirega minna kohta, kus nooleke muutub väikeseks ristikeseks. Näiteks kui lohistada ühe lahtri võrra paremale, tekib ka lahtrisse D2 logaritmfunktsioon, kusjuures tema argumendiks võetakse B2.

Sarnaselt saab kopeerida ka mitme lahtri sisu korraga. Kui märgistatud on kaks külgnevat lahtrit ühes reas, võib alumisest paremast nurgast kinni võtta ja niimoodi mõlemat korraga allapoole lohistada.



## Tsentreerimine

Tsentreerimine (igast vaatlusest valimikesmise maha lahutamine) on nagu iga teinegi teisendus: tuleb ühte ja sama funktsiooni rakendada kõigile vaatlustele. Küll aga on tsentreerimisel võrreldes näiteks logaritmimisega Exceli silmis üks tähtis erinevus. Nimelt see valimi keskmine, mis tuleb igast vaatlusest maha lahutada, ei tohiks lohistamisel kaasa tulla.

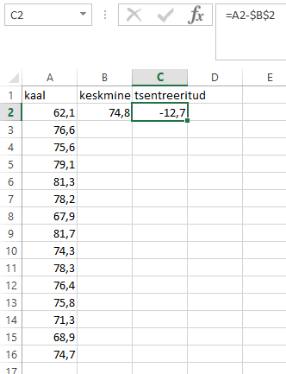
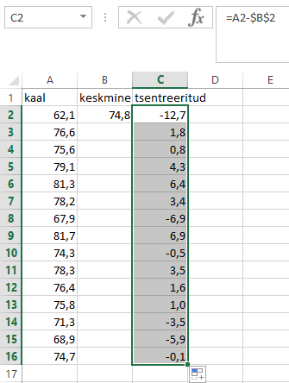
Olgu meil andmed nagu all näha ja olgu keskmine juba välja arvutatud. Pange tähele, et tervele veerule A saab viidata A:A, mistõttu veeru A kõigi vaatluste keskmise saab arvutada ka nii:

=AVERAGE(A:A)

Tsentreerimiseks tuleb lihtsalt vaatlusest maha lahutada keskmine, aga see keskmine lahtris B2 on kõigile vaatlustele ühine ja (erinevalt vaatlusest endist) ei tohi see lohistamisega muutuda. Selleks on Excelis olemas dollarimärgid: kui viite sümboli ees on $, siis see sümbol lohistamisega kaasa ei tule. Dollarimärke saab viitele tekitada ka klahviga F4 (vajuta seda klahvi, kui kursor parajasti viites istub). Praeguses näites tuleks esimene vaatlus tsentreerida nii:

=A2-$B$2

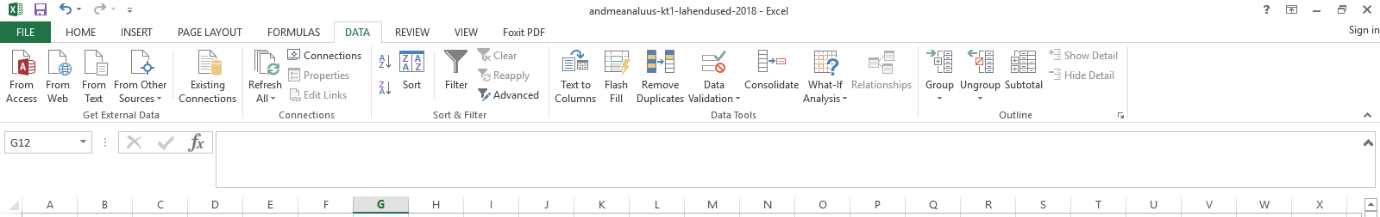
Seda sisu tuleks lohistada allapoole.



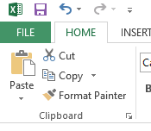
# Analysis ToolPack

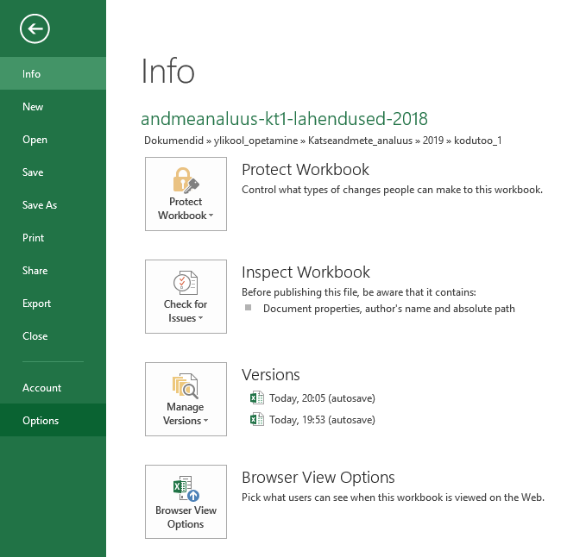
Mitmeid analüüse on Excelis lihtsam teha statistika-protseduure sisaldava lisapaketiga nimega *Analysis ToolPack*. See pakub rohkem abi analüüsi tellimisel ja annab reeglina rohkem paremini vormistatud tulemusi, kuid lisapaketiga tehtud analüüside tulemused ei uuene automaatselt, kui lähteandmed muutuvad. Kui on vaja, et analüüside tulemused uueneksid jooksvalt iga muudatusega lähteandmetes, siis tuleks eelistada funktsioone (vaata selle õppematerjali lisa).

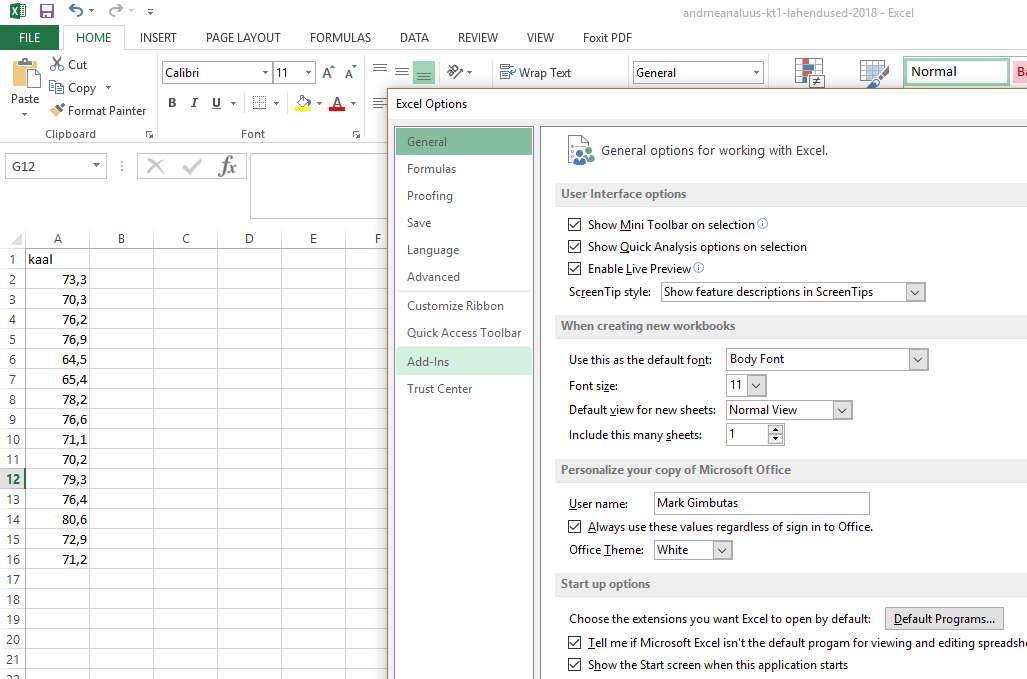
Pakett on Excelil küll põhimõtteliselt vaikimisi kaasas, kuid kasutamiseks tuleb see sõnaselgelt kasutusele võtta. Kui avaneb järgnev pilt, siis seda lisapaketti veel ei ole kasutusele võetud.

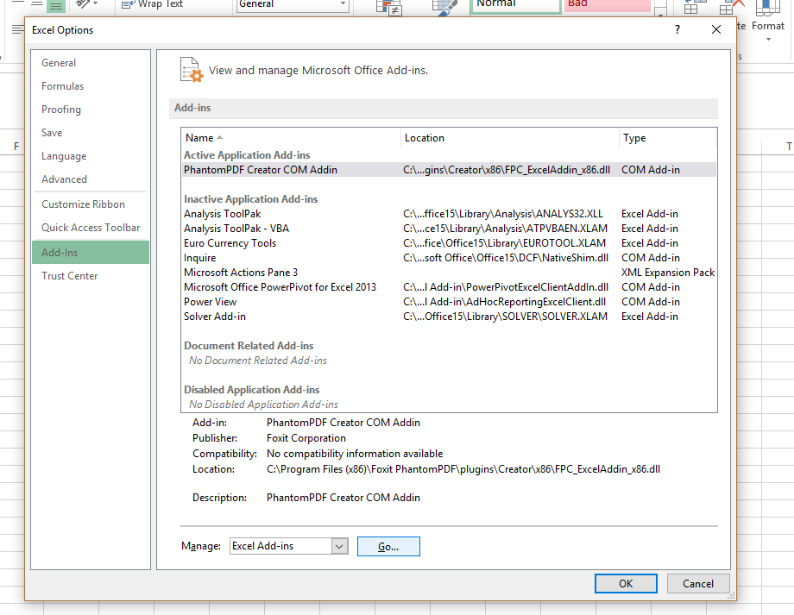


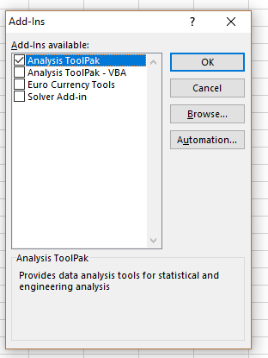
Andmeanalüüsi paketi (*Analysis ToolPack*) lisamiseks kliki nagu näitavad punased noolekesed all (sisuliselt järgi juhist <https://www.excel-easy.com/data-analysis/analysis-toolpak.html> ):



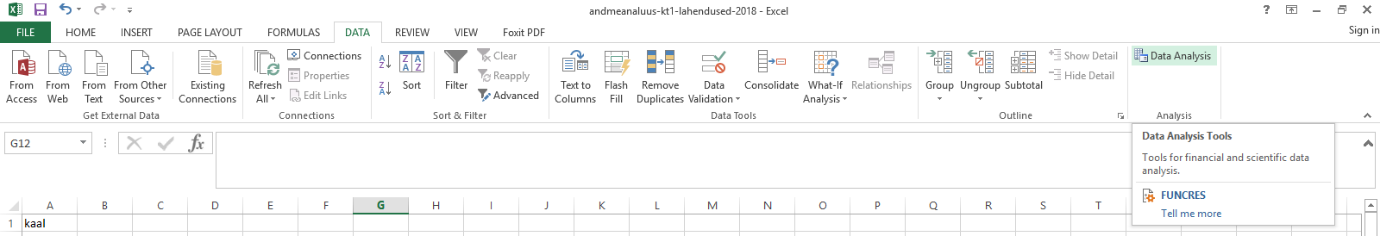






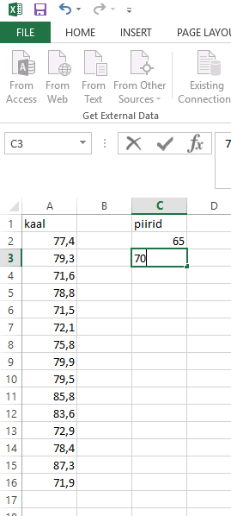
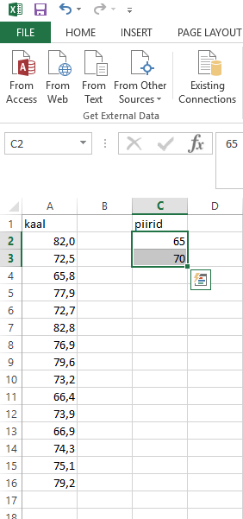
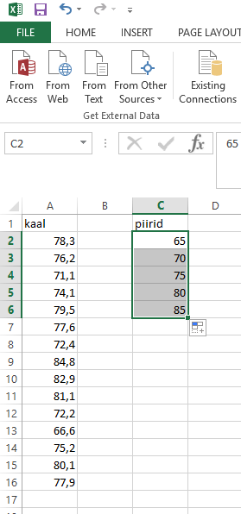


Tulemus peaks olema selline:

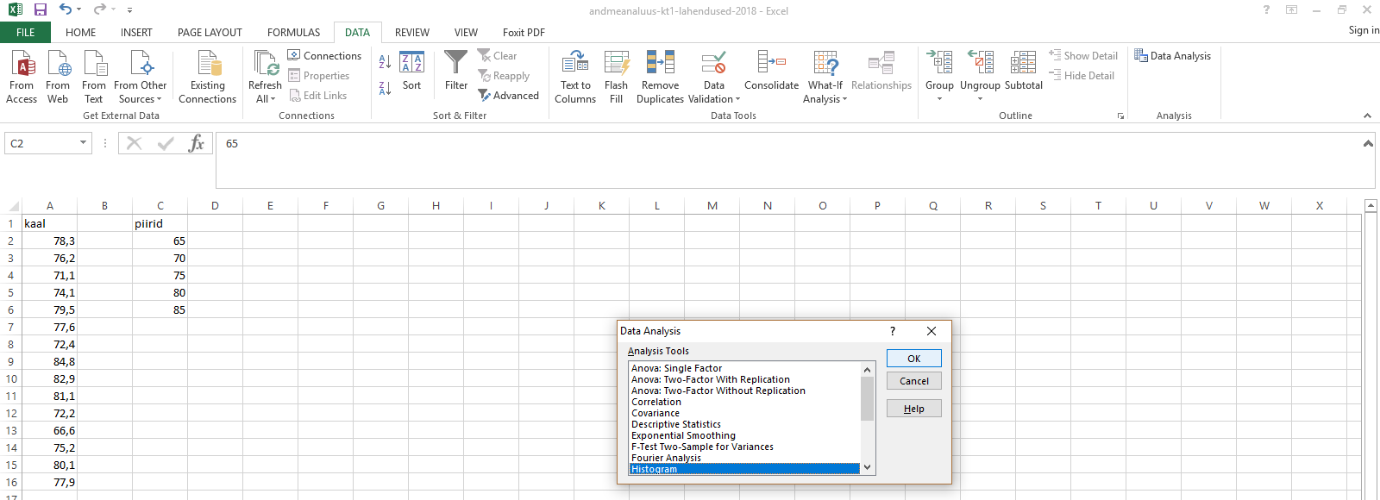


# Sagedusjaotus ja histogramm

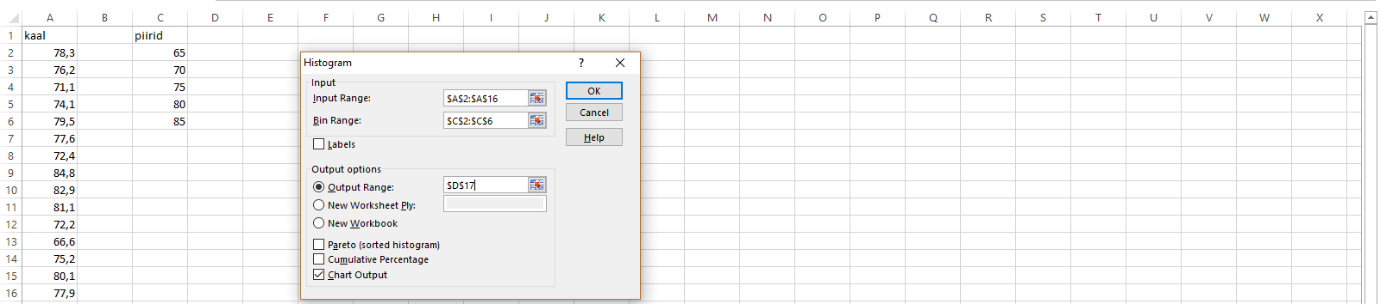
Nii sagedusjaotuse kui histogrammi jaoks on vaja määrata väärtuste vahemikud, mille lõikes me sagedusi näidata tahame. Selleks tuleks eraldi veergu kirjutada klasside piirid. Kui kõiki piire ei viitsi kirjutada, siis võib ära kasutada Exceli kavalust: kirjutada ainult kaks esimest, need ära märkida, ja alumisest parempoolsest nurgast allapoole lohistada (hiireklahvi all hoides) kuni vajalikud arvud käes. Excel saab aru soovitavast sammust ja kordab seda.



Kui andmeanalüüsi lisapakett on paigaldatud, pole vaja ise enne sagedustabelit teha. Lähenemise puuduseks on see, et graafik ei muutu koos andmetega: iga kord, kui andmeid muuta, tuleb järgnev protseduur uuesti läbida.

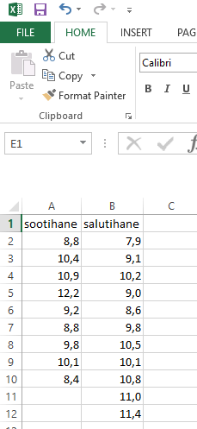


Avanevas aknas tuleb Excelile kätte näidata eelkõige andmete vahemik ja tulpade piirid, aga ka näiteks tulemuste tekitamise asukoht (uus leht või lahter (mis lahter?) käesoleval lehel) ja asjaolu kas tulemus peaks olema graafik või sagedustabel (ilmselt tahame me ikkagi graafilist tulemust).



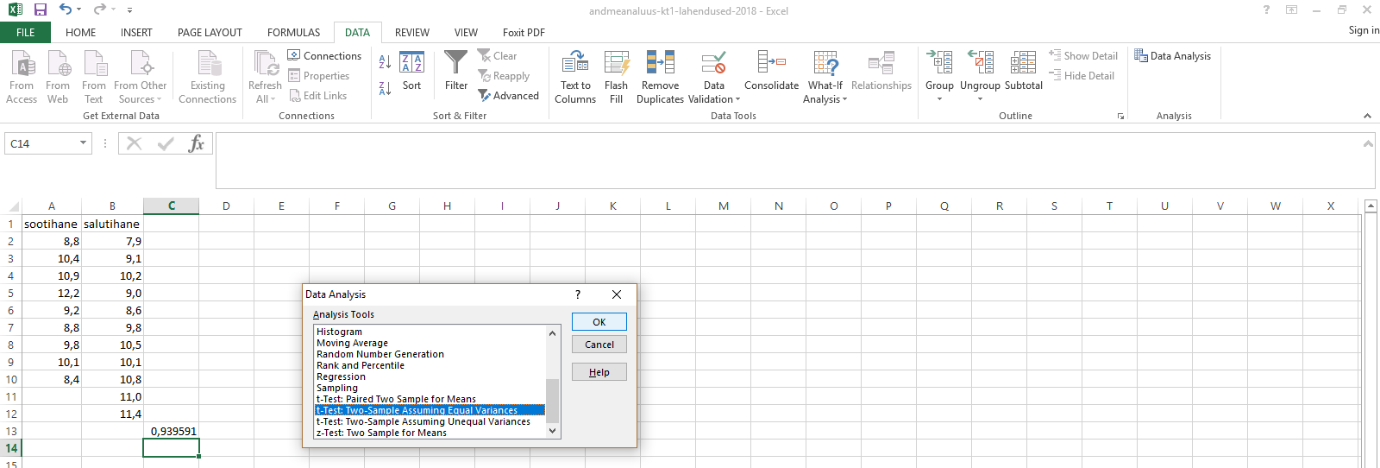
Eeltoodud dialoogiaknas vahemike määramiseks piisab hiirega vastavasse lahtrisse klõpsata ja siis lihtsalt vastav lahtrite vahemik ära märkida.

# T-test

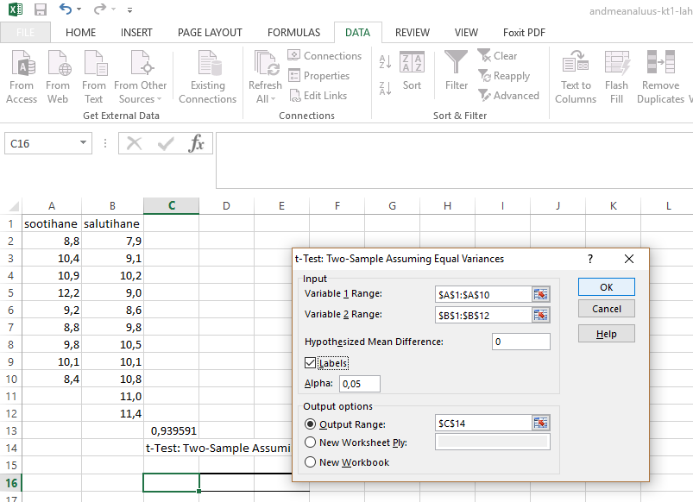
 t-testi tegemiseks on Excelis kaks võimalust: koos lisapaketiga ja ilma selleta. Nagu ennegi, seisneb peamine erinevus dünaamilisuses. Lisapaketiga on võib-olla ilusamini vormistatud tulemust lihtsam saada, aga algandmete muutumise korral ei arvutata tulemusi automaatselt ümber. Lisapaketti kasutamata, tuginedes ainult funktsioonidele, muutuvad aga tulemused koos andmetega. Järgnevalt on seletatud lisapaketi kasutamist, ilma lisapaketita protseduur on kirjeldatud lisas.

Andmed võiksid olla näiteks kahes eraldi veerus:

Järgi noolekesi piltidel. *Data Analysis* dialoogiakna lahtris tuleb ilmselt nimekirja *Analysis Tools* kerimisriba natuke allapoole kerida, sest t-testid on viimased seal. Kuna pildikesel kiri väike, siis hetkel teostame analüüsi nimega *t-Test: Two-Sample Assuming Equal Variances.*



Avanevas dialoogiaknas tuleb märkida andmevahemikud (nagu ennegi), *Hypothesized Mean Difference* null ja väljundi aadress (näiteks mingi lahter käesoleval töölehel). Kastikesse *Label* tuleks teha linnuke siis, kui märkisime andmevahemikesse kaasa ka veergude nimed (esimesel real). Lõpuks tuleks vajutada OK.



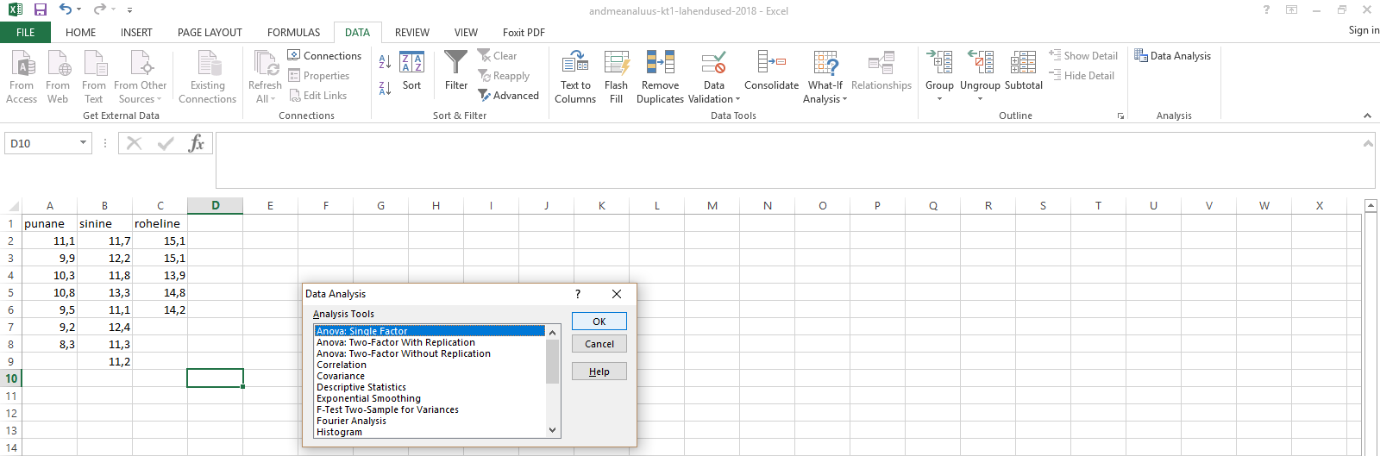
Tulemuseks võiks olla järgnev tabel (muidugi ilma seletusteta viimases veerus):

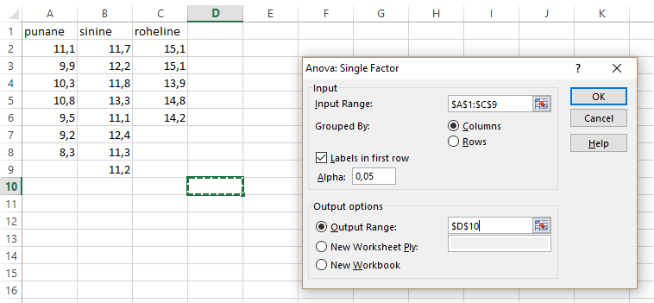
|  |  |  |  |
| --- | --- | --- | --- |
| t-Test: Two-Sample Assuming Equal Variances | | |  |
|  |  |  |  |
|  | *sootihane* | *salutihane* |  |
| Mean | 9,838856 | 9,878576 | valimite keskmised |
| Variance | 1,499378 | 1,180723 | valimite dispersioonid |
| Observations | 9 | 11 | valimite mahud |
| Pooled Variance | 1,322347 |  | valimite ühine dispersioon |
| Hypothesized Mean Difference | 0 |  | keskmiste erinevus nullhüpoteesi järgi |
| df | 18 |  | testi vabadusastmete arv |
| t Stat | -0,07685 |  | t-statistik |
| P(T<=t) one-tail | 0,469796 |  | ühepoolne p-väärtus |
| t Critical one-tail | 1,734064 |  | ühepoolse testi kriitiline väärtus (nivool 0,05) |
| P(T<=t) two-tail | 0,939591 |  | kahepoolne p-väärtus |
| t Critical two-tail | 2,100922 |  | kahepoolse testi kriitiline väärtus (nivool 0,05) |

# Ühefaktoriline ANOVA

Andmed peaksid olema esitatud nii, et ühele faktori tasemele vastavad vaatlused on kõik ühes veerus. Nii et näiteks kui meil on faktoril kolm taset (punane, sinine ja roheline), peavad vaatlused asuma kolmes veerus. ANOVA-t on võimalik teha ka ainult funktsioonidega, aga Analysis ToolPack-iga on see siiski oluliselt lihtsam. Vaatamegi ainult seda võimalust.

Algus on nagu ikka:



Seejärel tuleb avanevas dialoogiaknas märkida kogu andmevahemik (mitte igale faktori tasemele vastav veerg eraldi, vaid kõik koos), kas faktori tasemed on veergudes või ridades, kas esimene rida sisaldab faktori tasemete nimesid, ja kuhu läheb väljund.

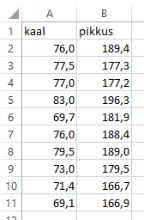
Tulemuseks võiks olla tabelid, millest esimene sisaldab lihtsalt rühmade kirjeldusi, ja teine on klassikaline ANOVA tabel.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Anova: Single Factor | | |  |  |  |  |
|  |  |  |  |  |  |  |
| SUMMARY | |  |  |  |  |  |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* |  |  |
| punane | 7 | 69,0362 | 9,862314 | 0,913037 |  |  |
| sinine | 8 | 95,00254 | 11,87532 | 0,544444 |  |  |
| roheline | 5 | 73,08512 | 14,61702 | 0,278586 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 65,94276 | 2 | 32,97138 | 53,87647 | 4,39E-08 | 3,591531 |
| Within Groups | 10,40368 | 17 | 0,611981 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 76,34643 | 19 |  |  |  |  |

P-väärtuse leiame muidugi veerust *P-value*. Tähistust 4,39E-08 tuleb lugeda kui 4,39 *korda kümme astmel* -8. Seega tegu on väga väikese p-väärtusega, mistõttu siniste, roheliste ja punaste üldkogumite keskmised usutavasti ei ole samad.

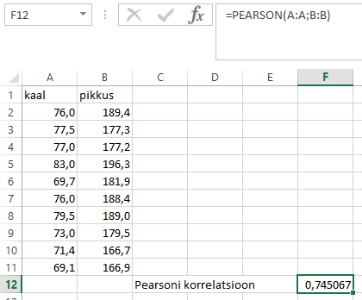
# Korrelatsioonanalüüs

Andmed korrelatsioonanalüüsiks tuleb esitada „tavalisel moel“:



Korrelatsioonanalüüsi saab teha nii funktsioonide abil kui ka lisapaketiga. Kuna korrelatsioon-analüüsi puhul aga lisapakett ei anna mitte mingit lisainfot võrreldes funktsiooniga, jätame selle praegu kõrvale. Tõsi küll, kui tunnuseid on vähemalt kolm, võimaldab *Analysis ToolPack*-i protseduur *Correlation* korraga arvutada kõigi tunnuste vahelised korrelatsioonid korraga.

Kahe tunnuse vahelise Pearsoni korrelatsioonikordaja saab arvutada funktsiooniga PEARSON:



Sellel funktsioonil on kaks argumenti: esimene argument on esimene tunnus ja teine argument on teine tunnus. Viidata saab ka kogu veerule korraga, näiteks A:A tähendab „kõik arvud veerus A“.

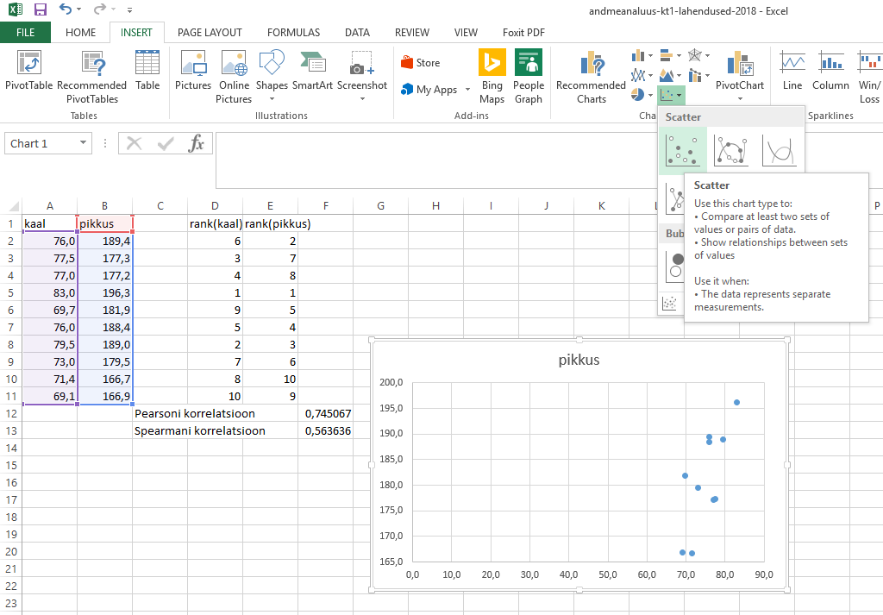
Vastava p-väärtuse saab arvutada järgneva valemiga:

=T.DIST.2T(F12\*SQRT(COUNT(A:A)-2)/SQRT(1-POWER(F12;2));COUNT(A:A)-2)

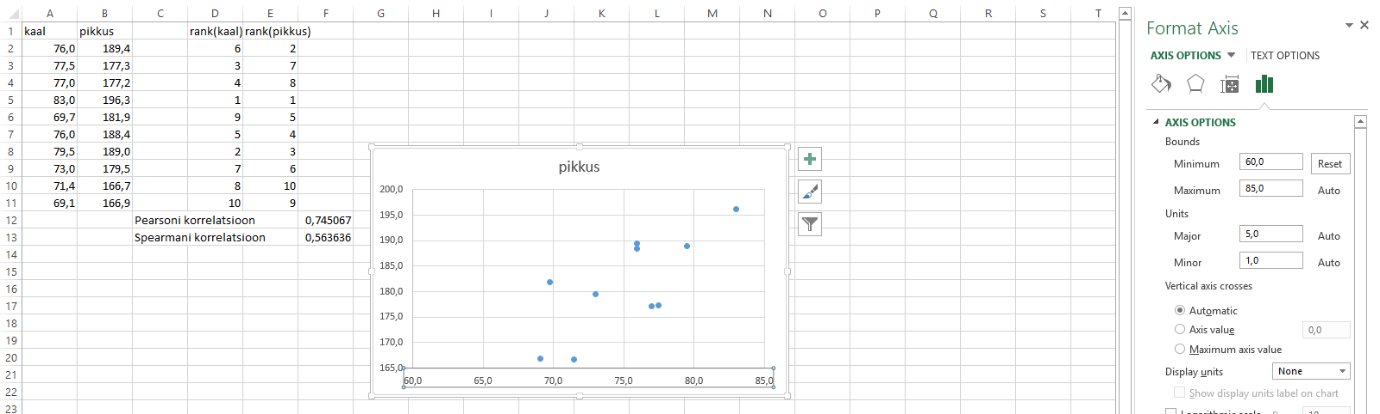
Valemi rakendamisel on kasulik teada, et lahtris F12 asus Pearsoni korrelatsioonikordaja ja funktsioon COUNT(A:A) loeb kokku kõik arvud veerus A (neid on kokku 10). P-väärtus tuleb praegu 0,013405.

### Hajusdiagramm

Korrelatsioonanalüüsi iseloomustavat graafikut on õnneks üsna lihtne joonistada. Et graafiku joonistamine läheks libedamalt, on mõistlik lähteandmed ära märkida enne graafiku joonistamist.



Tundub mõistlik x-telje skaalat natuke kitsamaks teha. Selleks tee klõps x-teljel. Akna paremas servas avaneb menüü, kus tuleb siis *Axis options > Bounds > Minimum* väärtus teha selliseks nagu vaja.



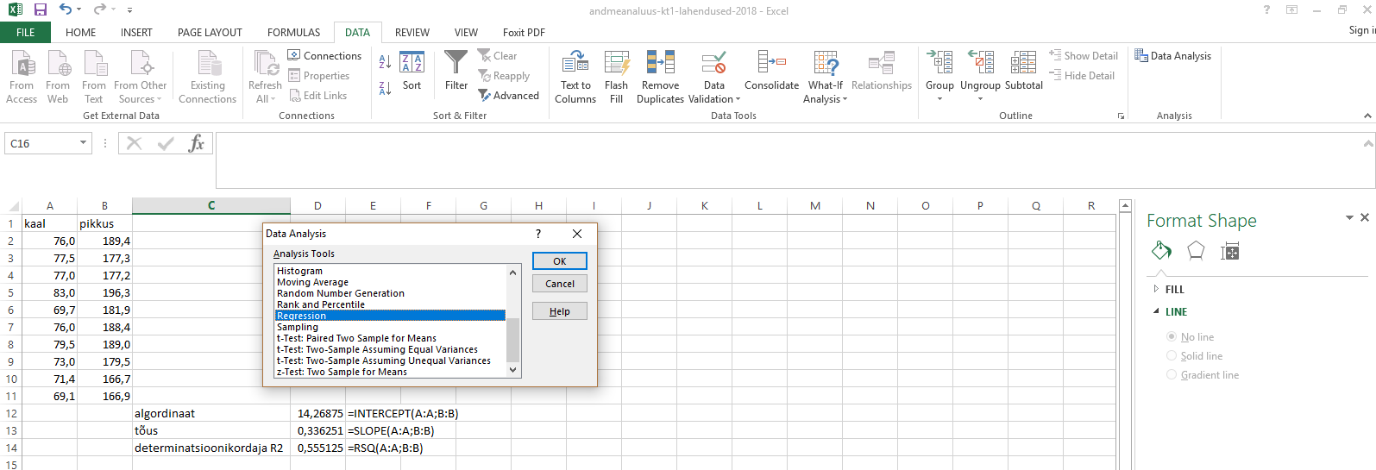
Üldiselt saab enam-vähem igat graafiku elementi natuke mudida, kui selle peal teha klõps ja uurida mis akna parempoolses servas avaneb.

# Lineaarne regressioon

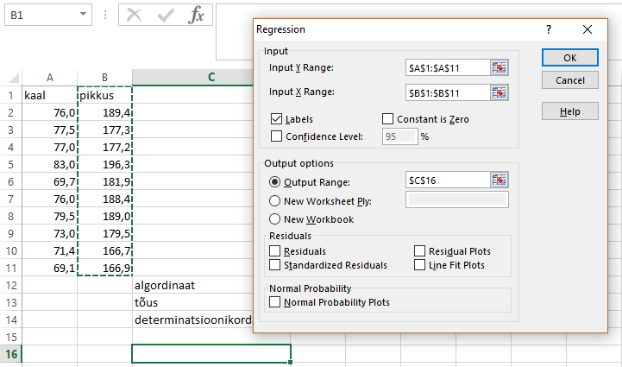
Regressioonanalüüsi, täpsemalt siis lineaarse regressiooni, andmed peaksid olema esitatud samamoodi nagu korrelatsioonanalüüsilgi. Samuti võib lineaarset regressiooni teha nii funktsioonide kui ka *Analysis ToolPack*-iga. Tõsi küll, funktsioonide abil on lihtne teha eelkõige lihtsat (ühe x-tunnusega) regressioonanalüüsi, paketi abil on lihtsa vaevaga võimalik teostada ka mitmest (mitme x-tunnusega) regressioonanalüüsi.

## *Analysis ToolPack*-iga

Tuleb valida protseduur *Regression*.



Avanevas dialoogiaknas tuleb määrata lähteandmed, asjaolu kas märgitud lähteandmete esimeses reas on tunnuse nimi, ja väljundi aadress. Erinevalt funktsioonide argumentidest ei tohi *Analysis ToolPack*-i andmevahemikuks anda tervet veergu (ehk siis asju stiilis et A:A), vaid tuleb ikkagi määrata konkreetne lõplik vahemik.



Tulemuseks on kolm tabelit (viimasest tabelist on siin kaks parempoolset veergu kustutatud, sest need sisaldasid täpselt sama infot mis kaks järgmist veergu): Olulisemad arvud on kollase taustaga.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0,745067 |  |  |  |  |  |
| R Square | 0,555125 |  |  |  |  |  |
| Adjusted R Square | 0,499515 |  |  |  |  |  |
| Standard Error | 3,114468 |  |  |  |  |  |
| Observations | 10 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 1 | 96,83005 | 96,83005 | 9,982568 | 0,013405 |  |
| Residual | 8 | 77,59931 | 9,699914 |  |  |  |
| Total | 9 | 174,4294 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 14,26875 | 19,3144 | 0,738762 | 0,481161 | -30,2703 | 58,80784 |
| pikkus | 0,336251 | 0,106425 | 3,15952 | 0,013405 | 0,090835 | 0,581667 |

Tabelist *Regression Statistics* on vast tähtsam näitaja determinatsioonikordaja R2 kohendatud variant *Adjusted R Square*. Tavaline ja kohendatud R2 mõõdavad mõlemad sama näitajat, aga kohendatud R2 mõõdab seda täpsemini, sest ta võtab arvesse ka mudeli keerukust. Näiteks ANOVA puhul, kui (äärmuslikul juhul) on faktoril nii palju tasemeid et peaaegu igal vaatlusel on oma tase, siis mis te arvate mis vastavate „gruppide“ keskmised tulevad ja kas jääke ka enam jääb. Olukord on mõneti analoogiline ja seotud ka sellega, et valimi dispersioon ja valimi põhjal üldkogumi dispersiooni hinnang on natukene erinevad. Meenutame loengust, et R2 peaks näitama mudeli poolt seletatud hajuvuse osakaalu kogu hajuvuses, ja „See SS on *sum of squares* ehk jääkide dispersiooni ümber oma rühma keskmise (jääkhajuvust) iseloomustav ruutude summa ehk erinevused keskmisest on ruutu tõstetud ja kokku liidetud […] SS erinevus dispersioonist on seega see, et see on N-ga (valimi suurusega) läbi jagamata.“ Nüüd see tehniline detail, kas jagada nii jääkide kui kogu SS mõlemad läbi N-ga või hoopis vastava vabadusastmete arvuga (mis võtab arvesse mudeli keerukust) määrabki, kas tuleb tavaline või kohendatud R2.

Tabel *ANOVA* võtab kokku mudeli kui terviku olulisuse, eelkõige siis p-väärtus viimases veerus.

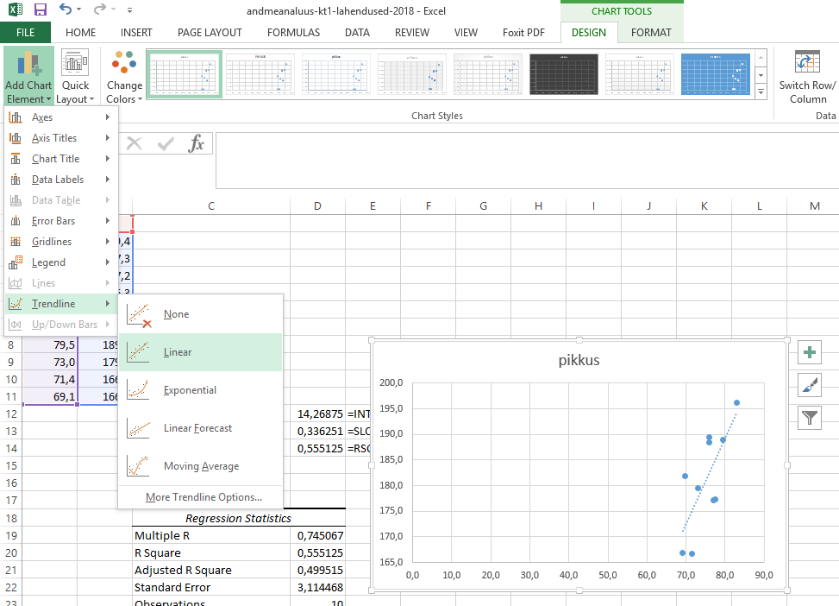
Kolmandas tabelis on üksikute parameetrite hinnangud koos vastavate p-väärtustega ja ka usaldusintervallidega.

Eelnevatel lehekülgedel toodud arve uurides võite märgata, et

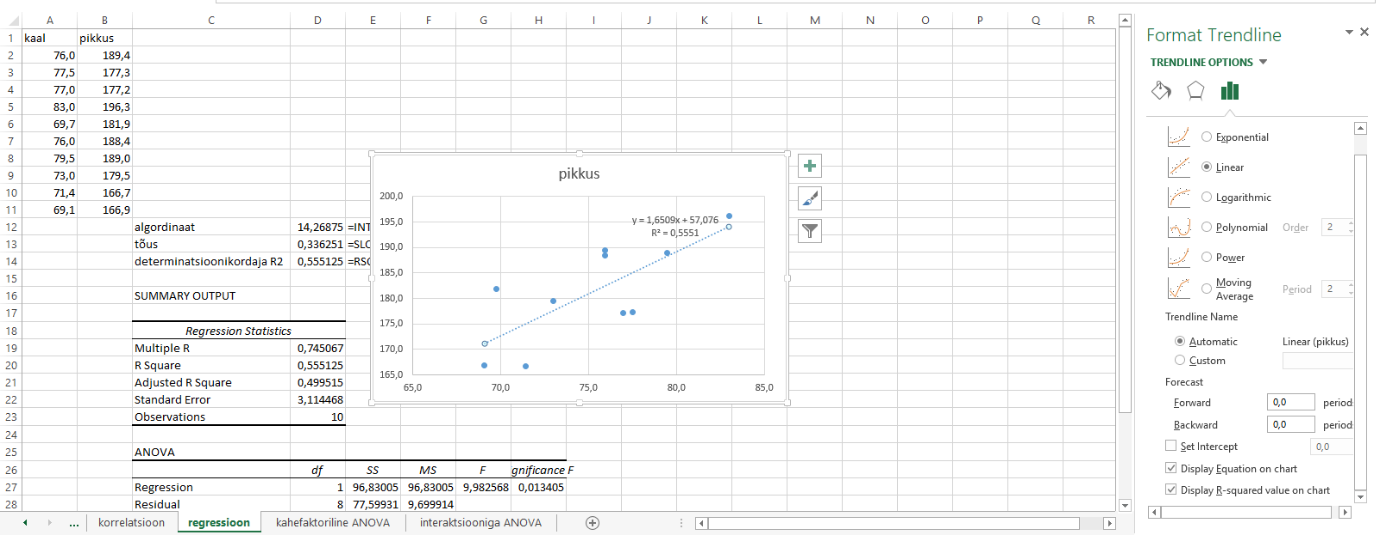
* *Multiple R* on Pearsoni korrelatsioonikordaja
* *Multiple R* ehk siis Pearsoni korrelatsioonikordaja ruut on *R Square*
* Lihtsa lineaarse regressioonimudel kui terviku p-väärtus on võrdne tema ainsa x-tunnuse p-väärtusega
* Pearsoni korrelatsioonikordaja p-väärtus on vastava lihtsa lineaarse regressioonimudeli x-tunnuse p-väärtus

## Hajusdiagramm regressioonsirgega

Alustame täpselt samasuguse hajusdiagrammi joonistamisest nagu korrelatsioonanalüüsi puhulgi. Kui hajusdiagramm joonistatud, lisame sellele regressioonijoone (*trendline*). Kõigepealt klikkige kuskil graafiku peal ja siis järgige noolekesi:



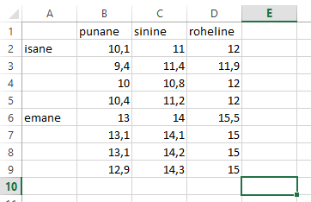
Nagu seletatud hajusdiagrammi juures, tuleks x-telje skaala teha paremaks. Samuti võime soovida regressioonisirge juures näidata sirge võrrandit ja R2 väärtust. Selleks tuleb klõpsata regressioonisirge peal ja teha paar linnukest paremal pool avanevas menüüs:



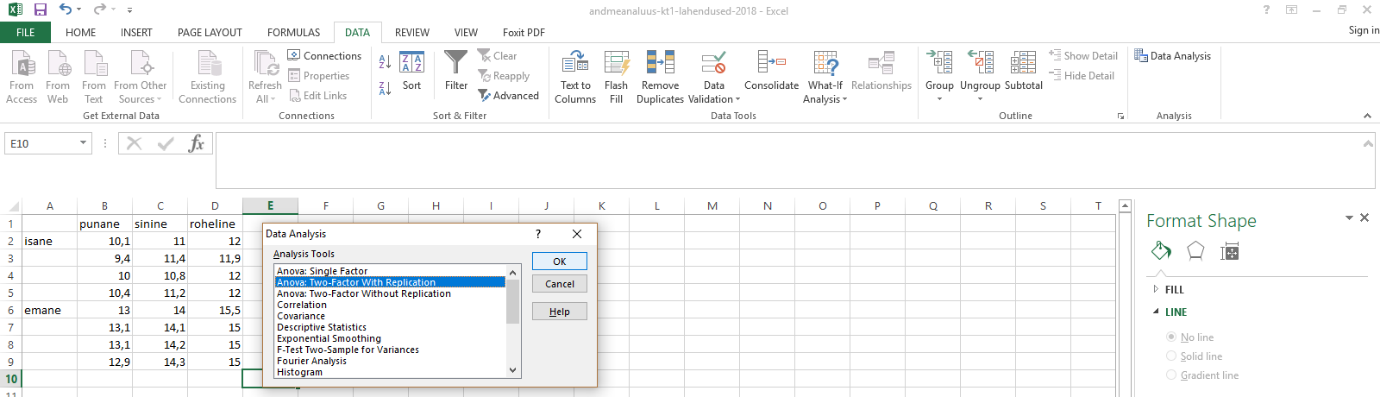
# Kahefaktoriline ANOVA

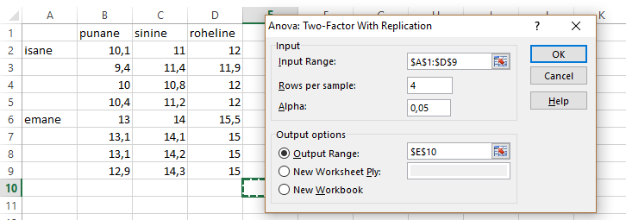
Kahefaktorilise ANOVA tegemisel Excelis on see oluline kitsendus, et analüüsida saab ainult tasakaalus andmestikku. Seega igal faktori tasemete kombinatsioonil peab olema täpselt sama palju vaatlusi.

Andmed peavad olema esitatud nii, et ühe faktori tasemed on veergudes ja teise faktori tasemed ridades. Igal faktori tasemete kombinatsioonil peab olema tehtud rohkem kui üks vaatlus, need vaatlused on eraldi ridades (vaata näidet):



Analüüsi teostamine käib sarnaselt eelmise näitegagi, analüüsimeetodiks tuleb ainult valida *Anova: Two-Factor With Replication* ja ainsa lisana tuleb täpsustada et mitu vaatlust siis on igal faktori tasemete kombinatsioonil tehtud (praegu on neid 4).





Tulemus kuvatakse jällegi tabelitena.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Anova: Two-Factor With Replication | | | |  |  |  |
|  |  |  |  |  |  |  |
| SUMMARY | punane | sinine | roheline | Total |  |  |
| *isane* |  |  |  |  |  |  |
| Count | 4 | 4 | 4 | 12 |  |  |
| Sum | 39,9 | 44,4 | 47,9 | 132,2 |  |  |
| Average | 9,975 | 11,1 | 11,975 | 11,01667 |  |  |
| Variance | 0,175833 | 0,066667 | 0,0025 | 0,797879 |  |  |
|  |  |  |  |  |  |  |
| *emane* |  |  |  |  |  |  |
| Count | 4 | 4 | 4 | 12 |  |  |
| Sum | 52,1 | 56,6 | 60,5 | 169,2 |  |  |
| Average | 13,025 | 14,15 | 15,125 | 14,1 |  |  |
| Variance | 0,009167 | 0,016667 | 0,0625 | 0,827273 |  |  |
|  |  |  |  |  |  |  |
| *Total* |  |  |  |  |  |  |
| Count | 8 | 8 | 8 |  |  |  |
| Sum | 92 | 101 | 108,4 |  |  |  |
| Average | 11,5 | 12,625 | 13,55 |  |  |  |
| Variance | 2,737143 | 2,693571 | 2,862857 |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Sample | 57,04167 | 1 | 57,04167 | 1026,75 | 2,5E-17 | 4,413873 |
| Columns | 16,86333 | 2 | 8,431667 | 151,77 | 5,4E-12 | 3,554557 |
| Interaction | 0,013333 | 2 | 0,006667 | 0,12 | 0,887624 | 3,554557 |
| Within | 1 | 18 | 0,055556 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 74,91833 | 23 |  |  |  |  |

Determinatsioonikordajat (R-ruut) ükski Exceli funktsioon ANOVA jaoks valmis kujul välja ei paku, küll aga on seda lihtne arvutada ANOVA tabeli väljundist. Nimelt tuleks mudeli poolt seletatud ruutude summa (SS) jagada kogu ruutude summaga. Mudeli poolt seletatud ruutude summa leidmiseks lihtsam viis on aga kogu ruutude summast (*Total SS*) lahutada maha seletamata jäänud ruutude summa (*Error SS*). Kokkuvõtteks siis tuleks kasutada valemit

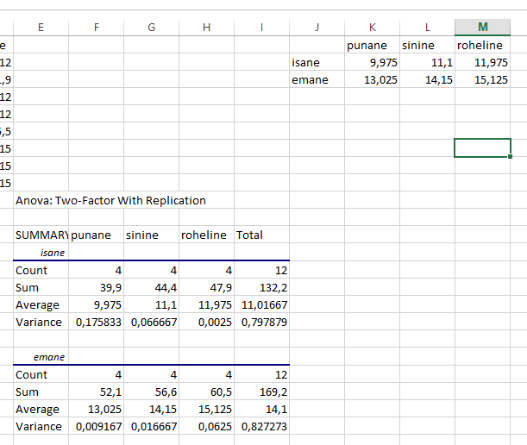
=(F39-F37)/F39

kus F39 viitab koguhajuvusele (*Total SS*) ja F37 viitab mudeli poolt seletamata hajuvusele (selles tabelis *Within SS*). Antud mudeli R2=0,986652 (nii palju tüvenumbreid ei ole üldiselt ilus esitada, tegin siin nii vaid kontrollimise mõttes).

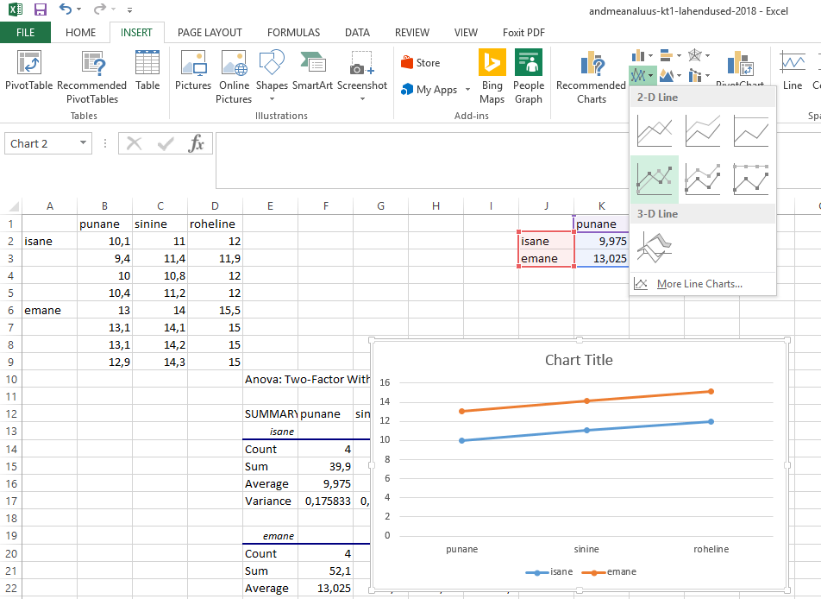
## Graafik keskmiste võrdlemiseks

Soovime graafikule kanda rühmade keskmisi, kus x-teljel on ühe faktori tasemed ja teine faktor on kajastatud erinevat värvi joontega.

Selle jaoks on kõigepealt vaja välja arvutada graafiku aluseks olevad keskmised: need võib kopeerida ANOVA väljundist.



Nüüd seletame graafiku lisamist. Et Excel oskaks ise pakkuda ilusaid eelvaateid, on mõistlik nüüd hiirega ära märkida graafiku andmed (J1:M3).

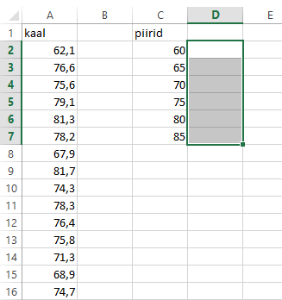
  
Jällegi on ilus lõpuks graafikule nimi panna.

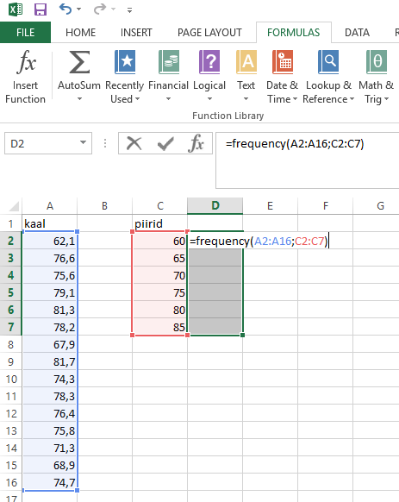
# Lisamaterjalid

## Sagedustabel ja histogramm funktsioonidega

Tutvustame histogrammi tegemist Exceli põhivahenditega. Alustada tuleb histogrammi aluseks oleva sagedustabeli koostamisest.

Selleks tuleb kõigepealt ära märkida kogu vahemik, kuhu tahame sagedused tekitada. Selguse huvides võiks see olla näiteks piiride vahemiku kõrval.

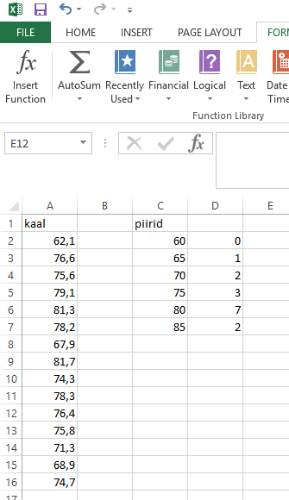


Alles siis kui kogu vahemik on märgitud, tuleks hakata klaviatuuril trükkima =frequency(

Kui esimene sulg on trükitud, võib märkida andmete vahemiku, seejärel trükkida klaviatuuril semikoolon (võimalik et teatavate keele-sätete puhul, eelkõige näiteks inglise keele puhul, on vaja hoopis koma), seejärel märkida piiride vahemik, ja seejärel trükkida lõpetav sulg. Muidugi võib ka vahemikud A2:A16 ja C2:C7 klaviatuuril trükkida. **Valemi sisestamiseks ei tohi vajutada Enter, vaid selle asemel Ctrl + Shift + Enter!**

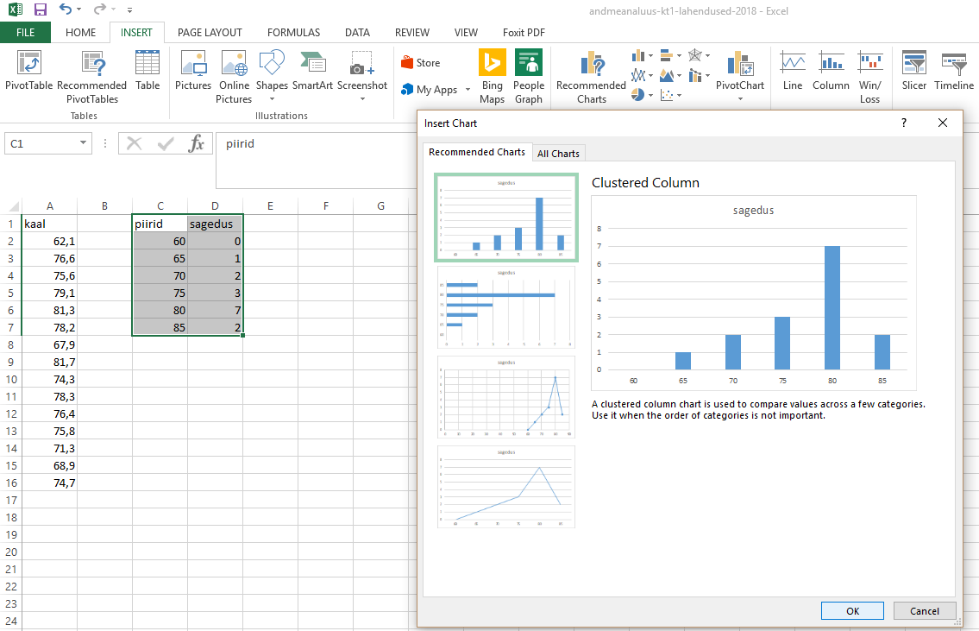
Valmis valem võiks välja näha näiteks nii:

=frequency(A2:A16;C2:C7)

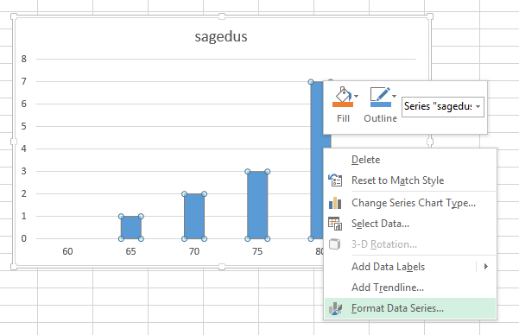
Tulemus võiks olla näiteks selline nagu näha parempoolsel joonisel.

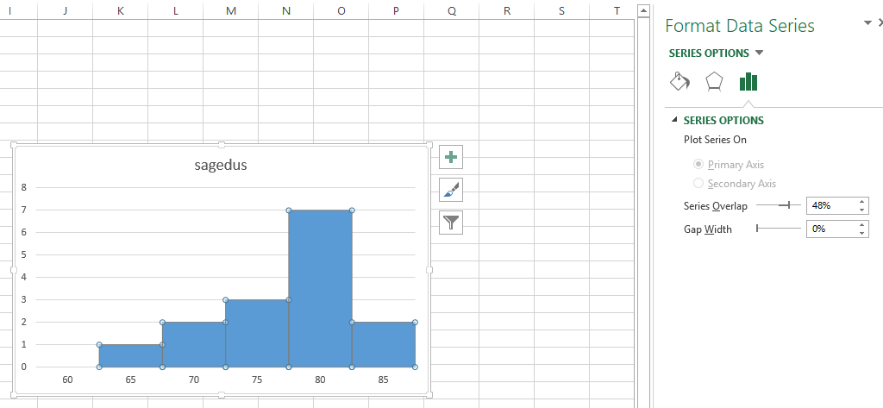
Iga number näitab et mitu väärtust langeb samal real asuva piiri ja sellest ühe rea võrra üleval pool asuva piiri vahele. Näiteks piiri 65 kõrval asuv 1 viitab sellele, et 60 ja 65 vahel asub 1 väärtus.

Nüüd on meil olemas sagedustabel ja sellest lähtuvalt saab nüüd teha histogrammi.



Histogramm on sisuliselt valmis. Seda saab natuke vormistada, näiteks kaotada ära vahed tulpade vahel. Selleks tee hiire parema klahviga klõps suvalisel histogrammi tulbal ja järgi juhist:

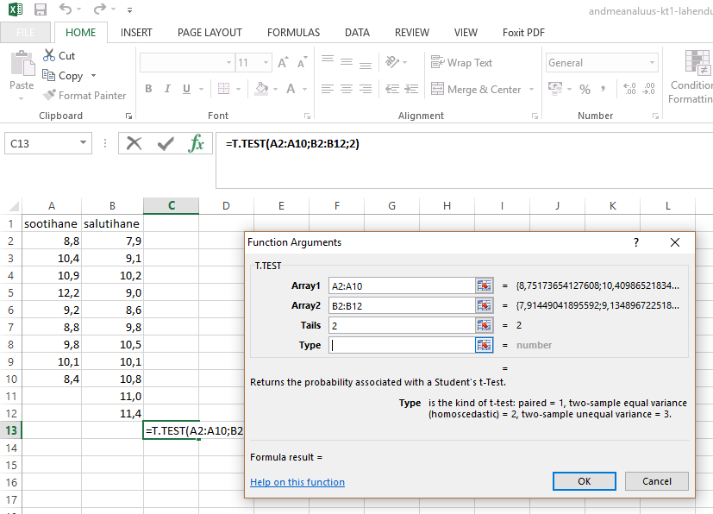




Tulpade all olevad arvud tähistavad sagedusklassi alumist piiri.

## T-test funktsioonidega

Sobiva funktsiooni nimi on, üllatus-üllatus, T.TEST. Kui selle funktsiooni poole pöörduda Exceli oma funktsiooniviisardi kaudu, on loota natuke kasutus-abi.



Funktsioonile tuleb ette anda võrreldavad valimid, märkida et teeme kahepoolse testi (lahtrisse *Tails* kirjuta 2) kahe sõltumatu valimi võrdlemiseks (lahtrisse *Type* kirjuta 2). Andmevahemikud A2:A10 ja B2:B12 (veeru nime pole vaja!) võib muidugi ka käsitsi trükkida, aga vast on mugavam need hiire abil märkida. Selleks tuleb kõigepealt klikkida vastavasse lahtrisse (s.t. **Array1** või **Array2** järele), nii et kursor sinna jääks, seejärel vahemik hiirega märkida nagu see ikka Excelis käib, ja siis võib näiteks Enter vajutada.

Kui kõik lähteandmed on antud, tuleks vajutada OK. Funktsioon annab tulemuseks ühe arvu: vastava t-testi p-väärtuse. T-statistiku enda välja arvutamiseks ainult funktsioone kasutades on Excelis mugavam mitte kasutada statistiku definitsiooni järgset valemit, vaid hoopis arvutada statistik p-väärtuse alusel. Täpsemalt tuleks kasutada funktsiooni T.INV.2T (miks 2T – sest et kasutasime kahepoolset testi ja p-väärtus on kahepoolne). Esimeseks argumendiks läheb p-väärtus ja teiseks argumendiks läheb vabadusastmete arv (antud juhul 9+11-2=18). Pange tähele, et see funktsioon annab teile alati plussmärgiga tulemuse – nii et kui esimese valimi keskmine on juhtumisi väiksem kui teise oma, siis tuleb tulemusele kindlasti ise miinusmärk ette kirjutada.

=T.INV.2T(C13;18)

Kasutada võib muidugi ka statistiku definitsiooni järgset valemit, aga see näeb üsna kole välja.

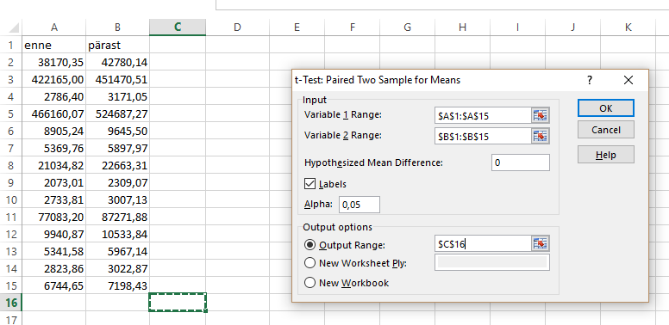
=(AVERAGE(A2:A10)-AVERAGE(B2:B12)) / SQRT((((9-1)\*VAR.S(A2:A10)+(11-1)\*VAR.S(B2:B12))/(9+11-2))\*(1/9+1/11))

## Sõltuvate valimite t-test lisapaketiga

Sõltuvate valimite (ehk siis tegelikult täpselt paariviisiliste vaatluste) t-testi tegemine on väga analoogiline eelnevalt kirjeldatud protseduuriga. Ainus (ja määrav!) vormiline erinevus seisneb selles, et vaatlusi peab olema täpselt sama palju. Ehk siis andmevahemikud peavad olema täpselt sama suured.

Sõltuvate valimite t-testi funktsiooni rakendamiseks tuleks vastavas sõltuvate valimite funktsioonis ära muuta ainult viimane argument: 2 või 3 asemel kirjutada number 1, seda lähenemist me hetkel üle ei tee.

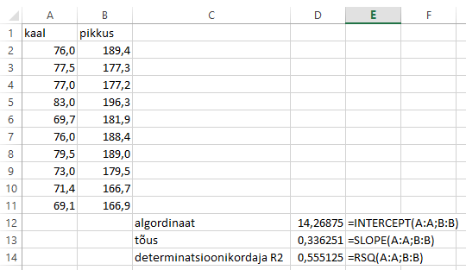
Ka lisapaketiga on sõltuvate valimite testi tegemine sarnane sõltumatute valimite testi tegemisega.

Dialoogiaknas tuleb ära märkida täpselt samad asjad mis sõltumatute valimite testi puhulgi. Tulemuseks peaks märgitud väljundaadressile tekkima järgnev tabel (muidugi ilma seletusteta viimases veerus):

|  |  |  |  |
| --- | --- | --- | --- |
| t-Test: Paired Two Sample for Means | | | |
|  |  |  |  |
|  | *enne* | *pärast* |  |
| Mean | 76523,758 | 84259,01 | valimite keskmised |
| Variance | 2,475E+10 | 3E+10 | valimite dispersioonid |
| Observations | 14 | 14 | valimite mahud |
| Pearson Correlation | 0,9995954 |  | Pearsoni korrelatsioonikoefitsient |
| Hypothesized Mean Difference | 0 |  | erinevus nullhüpoteesi järgi |
| df | 13 |  | testi vabadusastmete arv |
| t Stat | -1,7449726 |  | t-statistik |
| P(T<=t) one-tail | 0,0522832 |  | ühepoolse testi p-väärtus |
| t Critical one-tail | 1,7709334 |  | ühepoolse testi kriitiline väärtus (nivool 0,05) |
| P(T<=t) two-tail | 0,1045664 |  | kahepoolse testi p-väärtus |
| t Critical two-tail | 2,1603687 |  | kahepoolse testi kriitiline väärtus (nivool 0,05) |

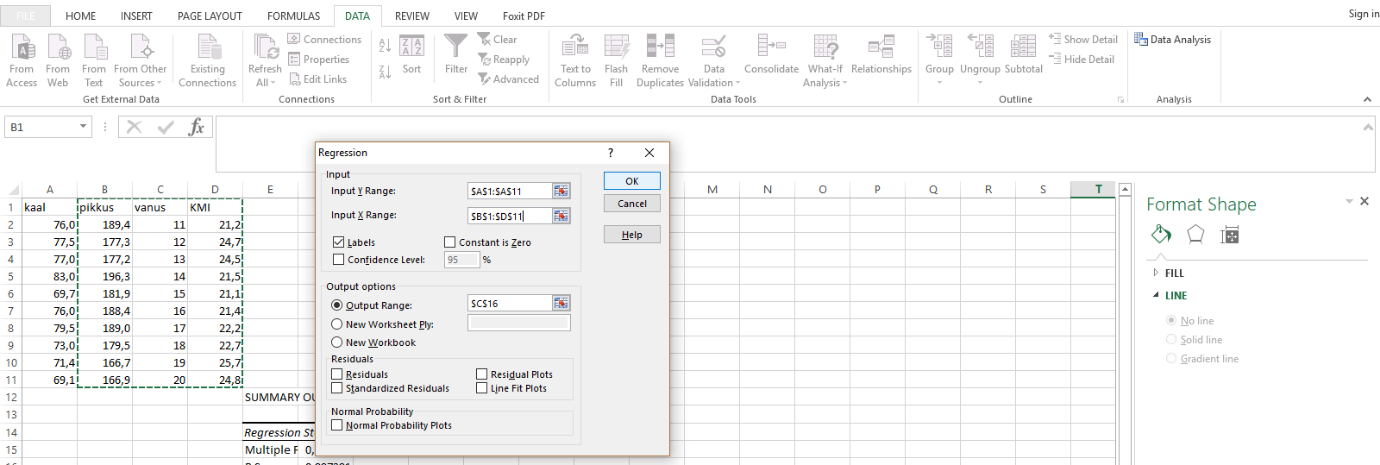
## Lineaarne regressioon funktsioonidega

Meenutame, et lihtne lineaarne regressioon üritab andmetele sobitada sirget võrrandiga y = a+bx. Vabaliikme a arvutab funktsioon INTERCEPT, tõusu b arvutab funktsioon SLOPE ja determinatsioonikordaja R2 annab RSQ. P-väärtust hetkel ei arvuta, sest selle annab väga lihtsasti kätte lisapakett.



## Mitmene lineaarne regressioon lisapaketiga

*Analysis Toolpack* võimaldab küllalt lihtsasti teha ka lineaarset mudelit, kui x-tunnuseid on mitu. Andmed peavad olema esitatud tavalise andmetabelina, kusjuures kõik x-tunnused peavad olema järjest (niimoodi ei saa, et y-tunnus on veerus B ja x-tunnused on veergudes A ning C). Ainus erinevus lihtsast lineaarsest regressioonist seisneb selles, et analüüsi tellimise dialoogiaknas *Input X Range* ei pea olema ainult üks veerg, vaid võib olla terve nelinurkne tabel.

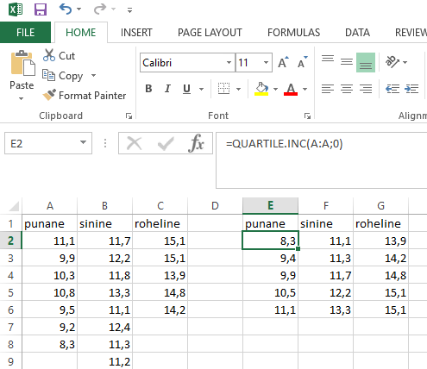


## Karpdiagramm

Karpdiagrammi tegemiseks eraldi nuppu ei ole ja selle saamiseks tuleb Excelit „loominguliselt kasutada“. Kõigepealt tuleb karpdiagrammil kujutatud kriipsukesed ise välja arvutada: miinimum, alumine kvartiil, mediaan, ülemine kvartiil ja maksimum. Õnneks on need kõik arvutatavad funktsiooniga QUARTILE.INC, muutes vaid viimast argumenti.

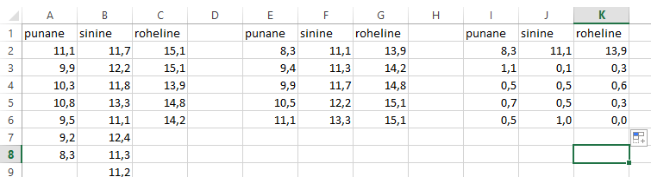
Lahtritesse trükitud funktsioonid näevad välja nii:

|  |  |  |  |
| --- | --- | --- | --- |
|  | punane | sinine | roheline |
| miinimum | =QUARTILE.INC(A:A;0) | =QUARTILE.INC(B:B;0) | =QUARTILE.INC(C:C;0) |
| alumine kvartiil | =QUARTILE.INC(A:A;1) | =QUARTILE.INC(B:B;1) | =QUARTILE.INC(C:C;1) |
| mediaan | =QUARTILE.INC(A:A;2) | =QUARTILE.INC(B:B;2) | =QUARTILE.INC(C:C;2) |
| ülemine kvartiil | =QUARTILE.INC(A:A;3) | =QUARTILE.INC(B:B;3) | =QUARTILE.INC(C:C;3) |
| maksimum | =QUARTILE.INC(A:A;4) | =QUARTILE.INC(B:B;4) | =QUARTILE.INC(C:C;4) |



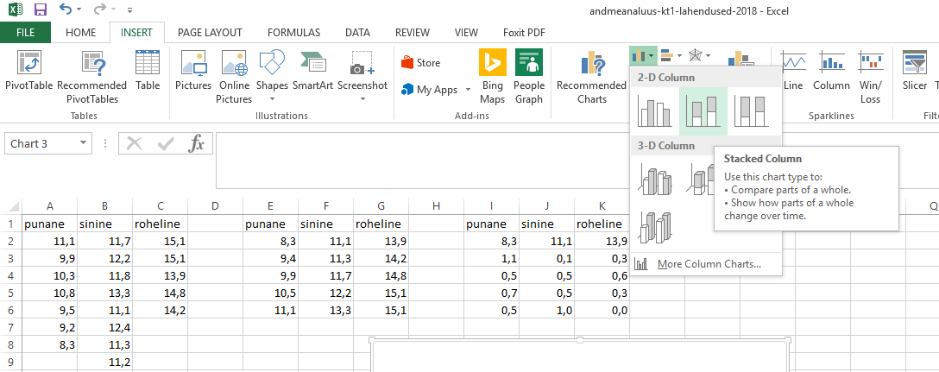
Seejärel tuleb arvutada karpdiagrammi osade kõrgused:

1. miinimum
2. alumine kvartiil miinus miinimum
3. mediaan miinus alumine kvartiil
4. ülemine kvartiil miinus mediaan
5. maksimum miinus ülemine kvartiil

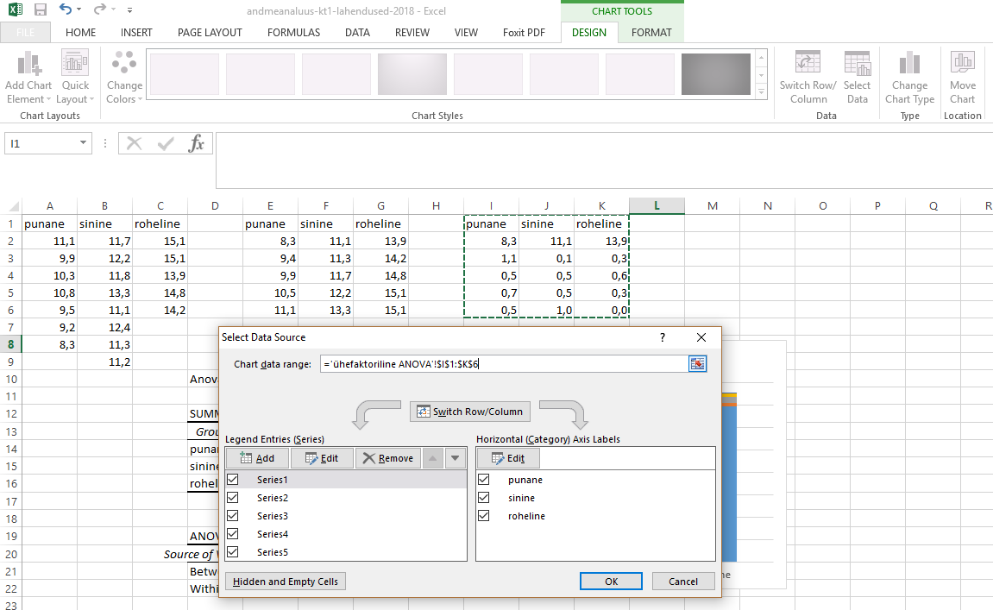
Kasutame selleks valemeid nagu näidatud allolevas tabelis. Pane tähele, et kui kollase taustaga valemid on ära trükitud, võib lahtrit I3 tõmmata allapoole, ja siis vahemikku I2:I6 paremale.

|  |  |  |
| --- | --- | --- |
| punane | sinine | roheline |
| =E2 | =F2 | =G2 |
| =E3-E2 | =F3-F2 | =G3-G2 |
| =E4-E3 | =F4-F3 | =G4-G3 |
| =E5-E4 | =F5-F4 | =G5-G4 |
| =E6-E5 | =F6-F5 | =G6-G5 |

Alles nüüd oleme valmis hakkama looma diagrammi. Valime menüüst välja *Stacked column* tüüpi graafiku.

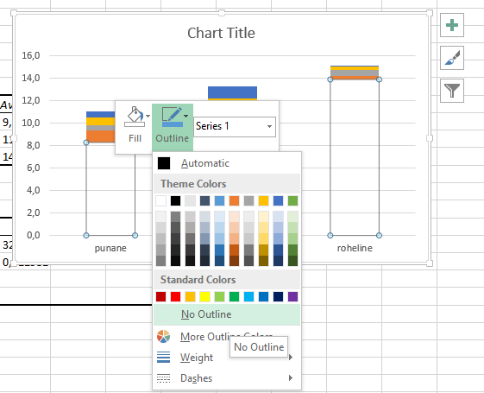


Seejärel valime välja andmed graafiku jaoks. Kui eelmise pildi järgi toimetamise ajal teie kursor juhtumisi oli kuskil andmestiku lähedal, võisite te juba näha mingit pilti, aga tõenäoliselt tuleb seda natukene muuta. Täpsemalt: tuleb jälgida, et veergude pealkirjad (punane, sinine ja roheline) oleksid parempoolses lahtris. Vajutage nuppu *Switch Row/Column*, kui see juba nii ei ole.

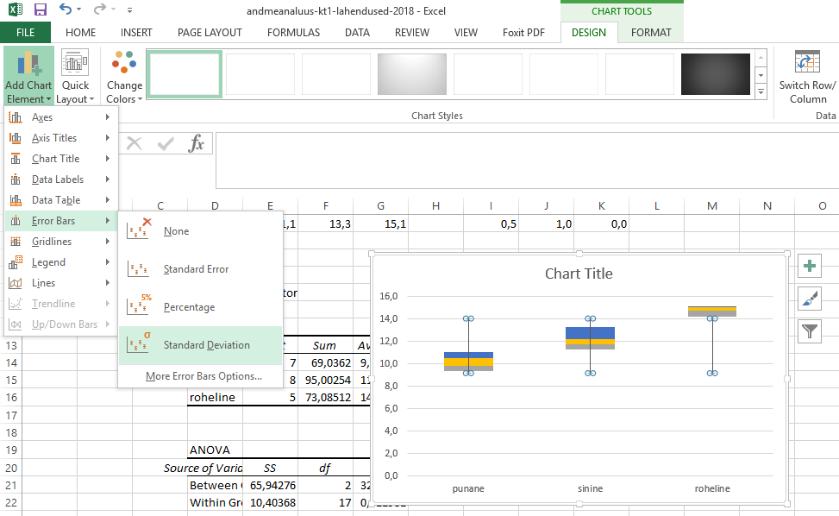


Nüüd tuleb tulemust vormistada: kõige alumine tulp tuleb ära peita, alt teine ja kõige ülemine tulp tuleb muuta hoopis vurrudeks, ülejäänud kaks tulpa tuleb muuta sama värvi.

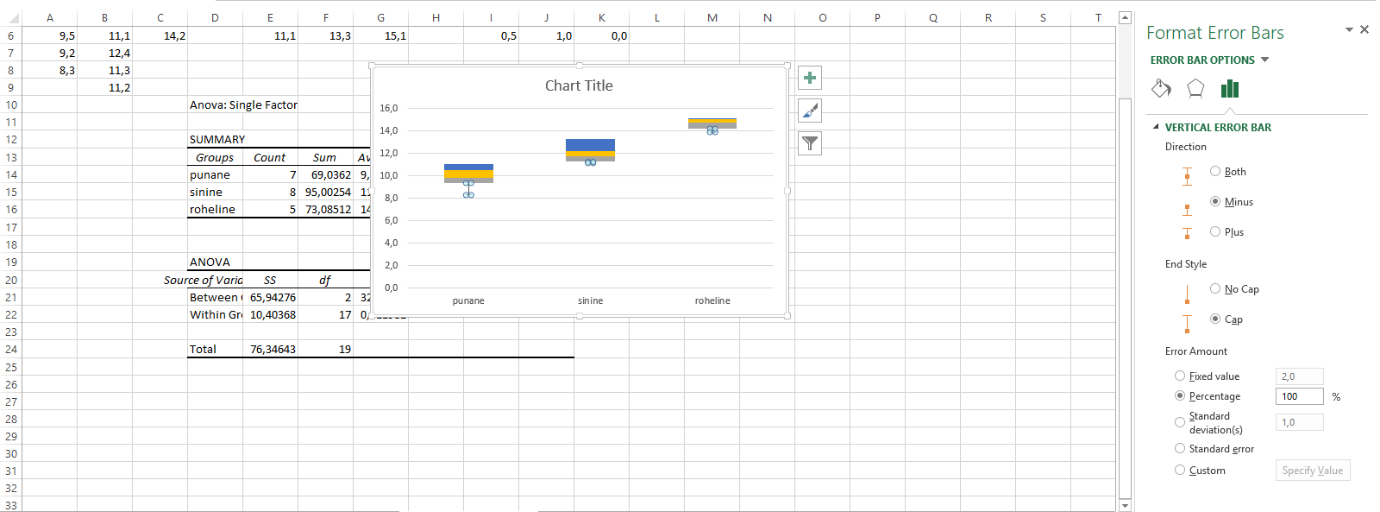
Alumise tulba peitmiseks tehke paremklõps suvalisel alumisel tulbal ja määrake sealt *No Fill* ja analoogiliselt *No Outline*.



Alt teise tulba vurruks muutmiseks tuleks ta kõigepealt ära peita täpselt nagu eelmisel pildil näidatud. Seejärel tuleks sisuliselt tekitada tema asemele *Error Bar*:

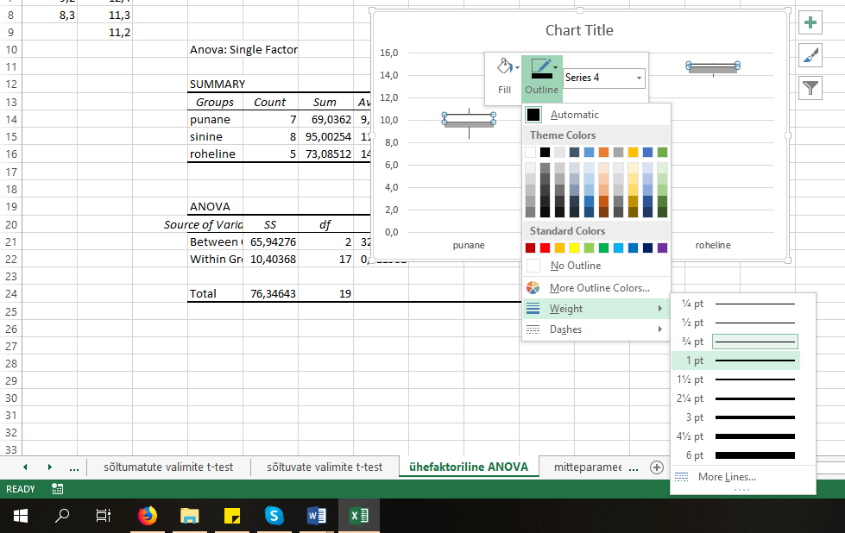


Kui vearibad on loodud, vaata akna paremasse serva, kas seal avaneb *Format Error Bars* menüü. Kui mitte, siis kliki ühe veariba peal. Seejärel toimeta nii:

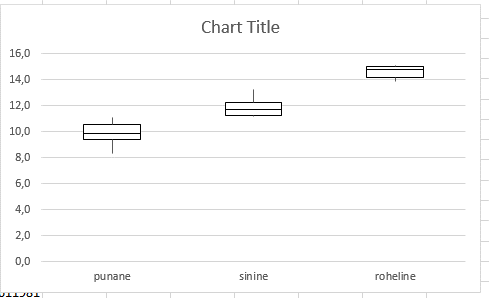


Toimeta sarnaselt ka kõige ülemise tulbaga.

Seejärel tuleks kahel ülejäänud tulbakesel määrata valge sisu (*Fill* valge värv) ja paks must ääris. Selleks klõpsa parema hiireklahviga ühel neist tulbakestest, järgi allolevat pilti, ja siis tee sama toiming veel viimase värvilise tulbakesega.



Tulemus on selline:



Otseloomulikult on hea veel graafikule pealkiri panna *Chart Title* asemel.

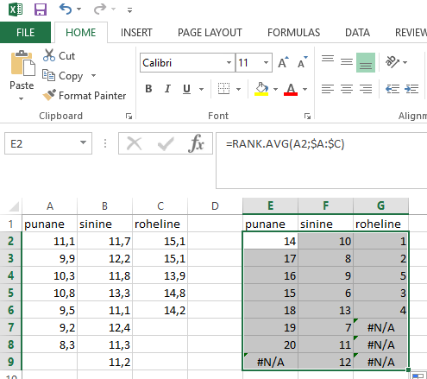
## Mitteparameetriline ANOVA

Excelis ei ole funktsiooni Kruskal-Wallise testi ehk mitteparameetrilise ANOVA tegemiseks. Vastav valem tuleb ise kirjutada. Proovimegi siis ise hakkama saada aadressil <https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance#Method> punktis 2 toodud valemi järgi arvutamisega Excelis.

Kõigepealt tuleb iga vaatluse jaoks välja arvutada tema järjekorranumber kõigi sorteeritud väärtuste seas (selle kohta öeldakse ka astak, inglise keeles *rank*). Selle jaoks on õnneks funktsioon olemas: RANK.AVG (see AVG seal lõpus tähendab, et võrdsetele väärtustele antakse keskmine astak: näiteks kui võitja ja teine koht on viigis, siis nad mõlemad saavad astaku 1,5). Ülemisse vasakpoolsesse lahtrisse kirjutame

=RANK.AVG(A2;$A:$C)

Esimene argument on see vaatlus, mille astakut arvutame. Teine argument on see vaatluste vahemik, mille seas me astakuid arvutame (A:C tähendab kõik veerud A kuni C). Pane tähele, et kuna vaatluste vahemik peaks jääma samaks, siis on A:C juures ka dollarimärgid. Lohistame selle lahtri sisu üle kõigi vaatluste.

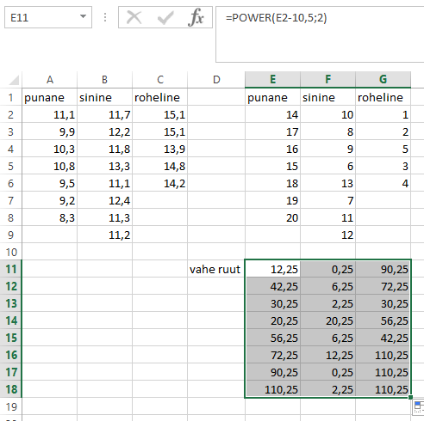


Kustutame puuduvatele väärtustele vastavad #N/A-d ära, need hakkavad segama.

Järgmisena tuleks arvutada iga astaku erinevuse ruut üld-keskmisest astakust. Üld-keskmine astak on õnneks lihtsasti arvutatav vaatluste koguarvust (vaatluste arv pluss üks, ja see jagada kahega). Antud näites on meil 20 vaatlust, seega keskmine astak on (20+1)/2=10,5. Tulemuseks on lihtne teisendus (POWER aitab meil ruutu tõsta):

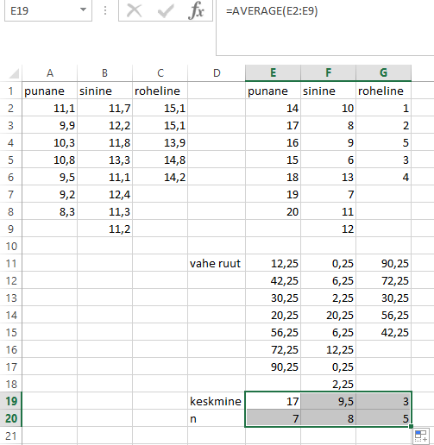
=POWER(E2-10,5;2)

ja seda tuleks lohistada üle kõigi vaatluste esmalt paremale ja siis alla:



Puuduvatele väärtustele vastavad arvud (110,25) tuleks kustutada (üks vasakpoolses ja kolm parempoolses veerus).

Nüüd tuleks arvutada keskmine astak igas kategoorias (rühma keskmine, kui soovite) ja iga rühma suurus (valimi maht rühmiti). Keskmist võite arvutada funktsiooniga AVERAGE, vaatlusi võite kokku lugeda funktsiooniga COUNT.



Lõpuks tuleks välja arvutada veel mõned summad vastavalt alltoodud tabelile (veergude F ja G sisu on täpselt sarnane veeru E sisuga, lihtsalt et E tähe asemel on vastavalt F ja G):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| D | E | F | G | H |
| lugejas | =E20\*POWER(E19-10,5;2) | 8 | 281,25 | =SUM(E22:G22) |
| nimetajas | =SUM(E11:E18) | 50 | 291,25 | =SUM(E23:G23) |
| Kruskal-Wallis H | |  |  | =(20-1)\*H22/H23 |

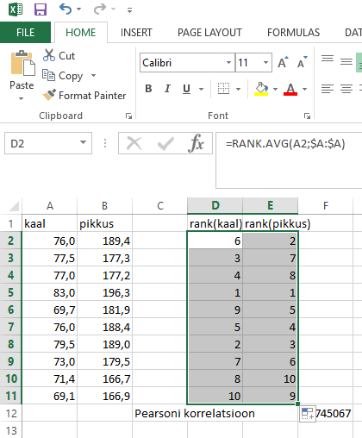
Lõpuks tuleks arvutada p-väärtus valemiga

=1-CHISQ.DIST(H24;2;TRUE)

kus esimene argument on vast-arvutatud statistik ja teine argument on võrreldavate gruppide arv miinus üks. Meie näites tuleb p=0,000235.

## Spearmani korrelatsioonikordaja

Kahe tunnuse vahelise Spearmani korrelatsioonikordaja saab arvutada, kasutades asjaolu et Spearmani korrelatsioonikordaja on võrdne Pearsoni korrelatsioonikordajaga astakute vahel (astak on vaatluse järjekorranumber, kui vaatlused järjestada kasvavas järjekorras, astakut saab arvutada funktsiooniga RANK.AVG). Seega esmalt tuleb arvutada astakud ja seejärel rakendada astakutele funktsiooni PEARSON.



Nüüd saab Spearmani korrelatsiooni arvutada kui korrelatsioonikordaja astakute vahel. Jälle on kasutatud viitamist kogu veerule korraga: D:D ja E:E. Siinkohal tuleb natuke jälgida, et te tulemust samasse veergu ei kirjuta (praegu on tulemus kirjutatud veergu F): vastasel korral tekib ringviide ja Excel kurjustab.

