HÜPOTEESIDE STATISTILINE KONTROLLIMINE II

LOENGU TEEMAD

- Dispersioonide testimine, *F*-test
- Osakaalude testimine
- Märgitest
- χ^2 – test (hii-ruut test)
  - Jaotuse sobivuse testimine
  - Kahe tunnuse vaheline seos
- Dispersioonanalüüs

KESKVÄÄRTUSTE TESTIMINE, SÕLTUMATUD VALIMID

Ühesugune dispersioon

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2)$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Studenti test

Erinev dispersioon

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(\nu_{WS})$$

$$\nu_{WS} = \frac{\left(s_1^2 / n_1 + s_2^2 / n_2\right)^2}{\frac{(s_1^2 / n_1)^2}{n_1 - 1} + \frac{(s_2^2 / n_2)^2}{n_2 - 1}}$$

Welchi test

Kuidas testida dispersioone?

F-TEST DISPERSIOONIDE TESTIMISEKS

Nullhüpotees H_0 $\sigma_1^2 = \sigma_2^2$

Kahe kogumi dispersioonid on võrdsed.

Sisukas hüpotees H_1 $\sigma_1^2 \neq \sigma_2^2$

Teststatistiku empiiriline väärtus $F = \frac{s_1^2}{s_2^2}$

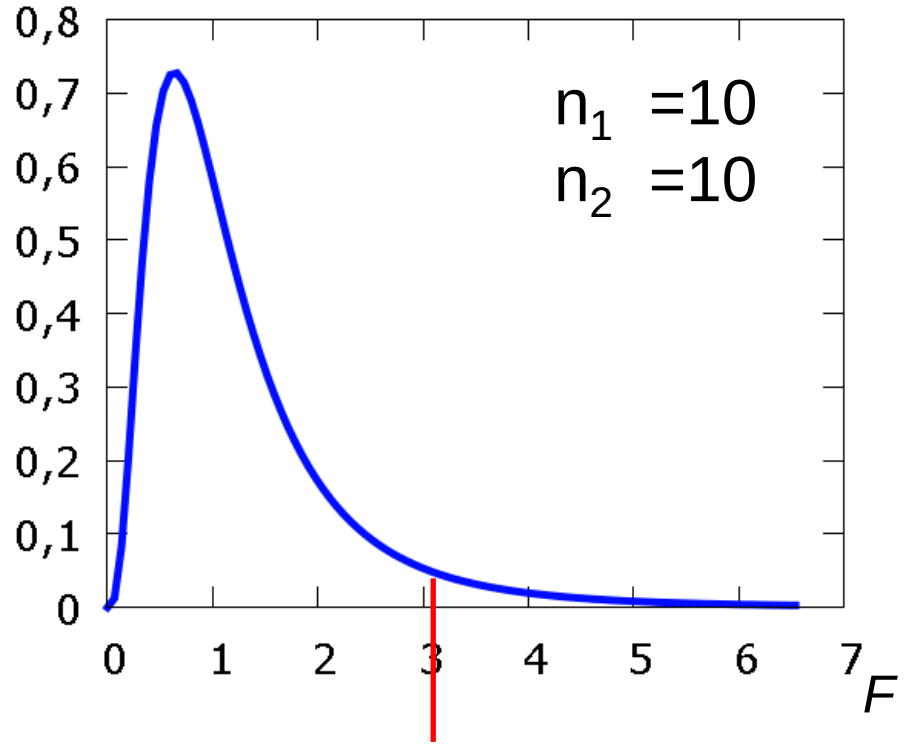
s_1^2 ja s_2^2 valimite dispersioonid

Nullhüpoteesi korral $F = 1$.

Nullhüpotees on ümber lükatud, kui empiiriline väärtus *F* erineb oluliselt **ühest.**

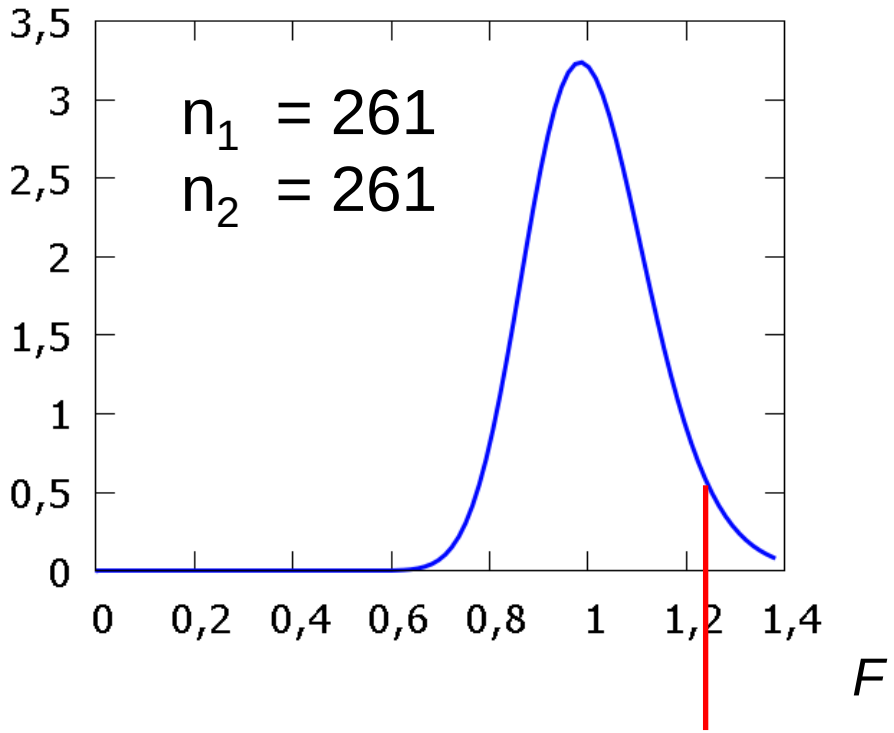
Kriitilised väärtused *F*-jaotusest (Fisheri jaotus) parameetritega $n_1 - 1$ ja $n_2 - 1$.

F - JAOTUS



F_{kr}= 3,18

Kriitiline väärtus nivool 0,05



F_{kr}= 1,23

Kriitilised leitud juhu jaoks, kui $s_1^2 > s_2^2$.

NÄIDE: MEESTE JA NAISTE SISSETULEKUTE VÕRDLUS, 1

Kas meeste sissetulek on naiste sissetulekust suurem?
Eesti Sotsiaaluuringu 2013 andmed.
Tegevusala: finantsvahendus, kinnisvara, rentimine ja äritegevus.
Ametiala: keskastme spetsialist, tehnik või ametnik.

Mehi 28, keskmine sissetulek aastas 9337,23 €, dispersioon s_1^2=43 292 030.
Naisi 71, keskmine sissetulek aastas 6757,37 €, dispersioon s_2^2=12 846 605.

Kasutada tuleb keskväärtuste võrdlemise *t*-testi, sõltumatud valimid.
Sobiva meetodi valikuks testime eelnevalt dispersioone, kasutades *F*-testi.

Nullhüpotees H_0 $\sigma_1^2 = \sigma_2^2$
Sisukas hüpotees H_1 $\sigma_1^2 \neq \sigma_2^2$

Teststatistik $F = \frac{s_1^2}{s_2^2} = \frac{43292030}{12846605} \approx 3{,}37$

Parempoolne kriitiline väärtus F_{krp}=1,81.

Järeldus: nullhüpotees on ümber lükatud, dispersioonid ei ole võrdsed.

NÄIDE: MEESTE JA NAISTE SISSETULEKUTE VÕRDLUS, 2

Keskväärtuste testimine. Küsimus: kas meeste (valim 1) keskmine sissetulek on suurem kui naistel (valim 2)? Ühepoolne hüpotees

$$H_0 : \mu_1 \leq \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

Kuna selgus, et dispersioonid ei ole võrdsed, tuleb kasutada *t*-testi erinevate dispersioonide korral.

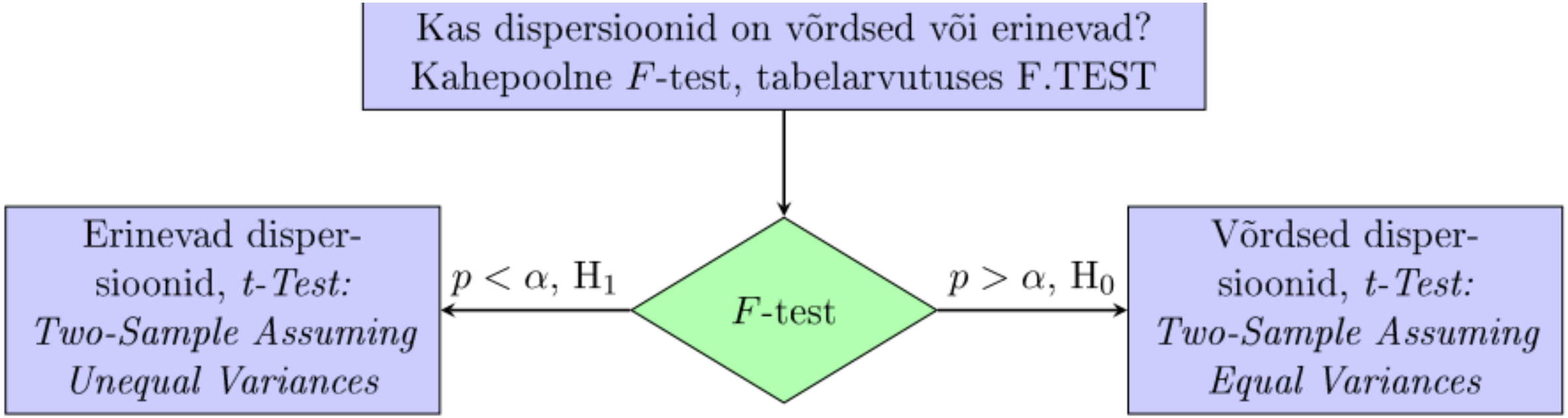
Excelis *t-Test: Two-Sample Assuming Unequal Variances*

	Mehed	Naised
Mean	9337,23	6757,73
Variance	43292030	12846605
Observations	28	71
Hypothesized Mean Difference	0	
df	34	
t Stat	1,963	
P(T<=t) one-tail	0,029	
t Critical one-tail	1,691	
P(T<=t) two-tail	0,058	
t Critical two-tail	2,032	

< 0,05

Võtame vastu sisuka hüpoteesi H_1: valitud tegevusalal ja ametialal on meeste sissetulek suurem kui naiste oma.

T-TEST SÕLTUMATUTE VALIMITE KORRAL



DISPERSIOONIDE TESTIMINE OMAETTE EESMÄRGINA

Mõningatel juhtudel võib dispersioonide testimine olla eesmärgiks omaette.

Näiteks aktsiahindade, pensionifondide volatiilsuse võrdlemine.

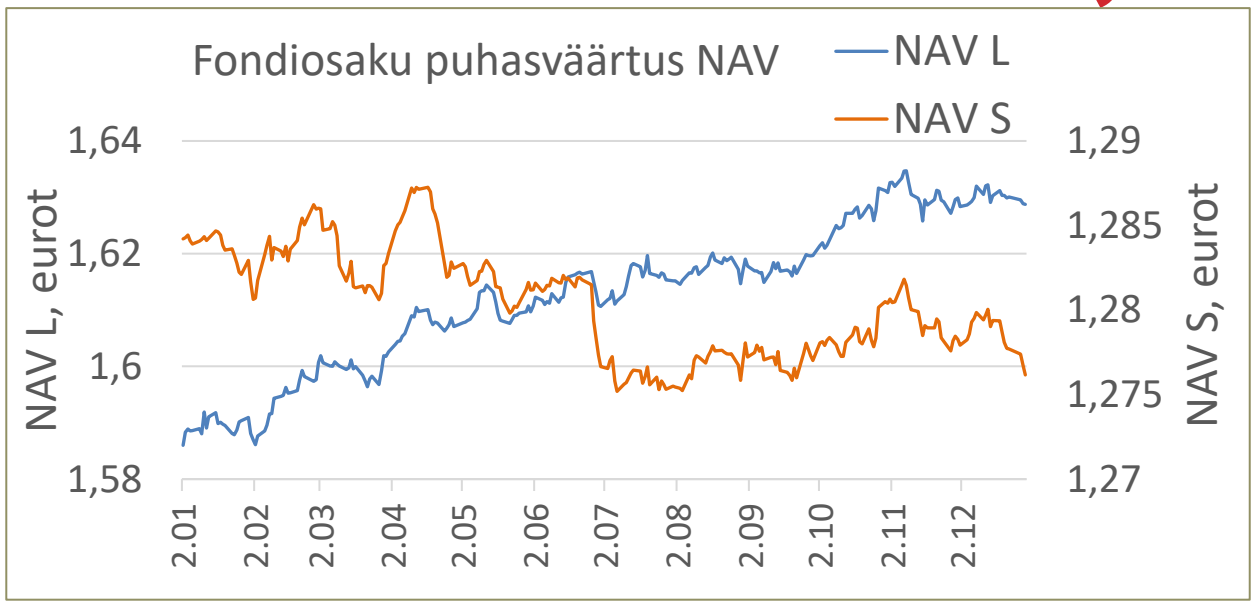
NÄIDE: PENSIONIFONDIDE VOLATIILSUS, 1

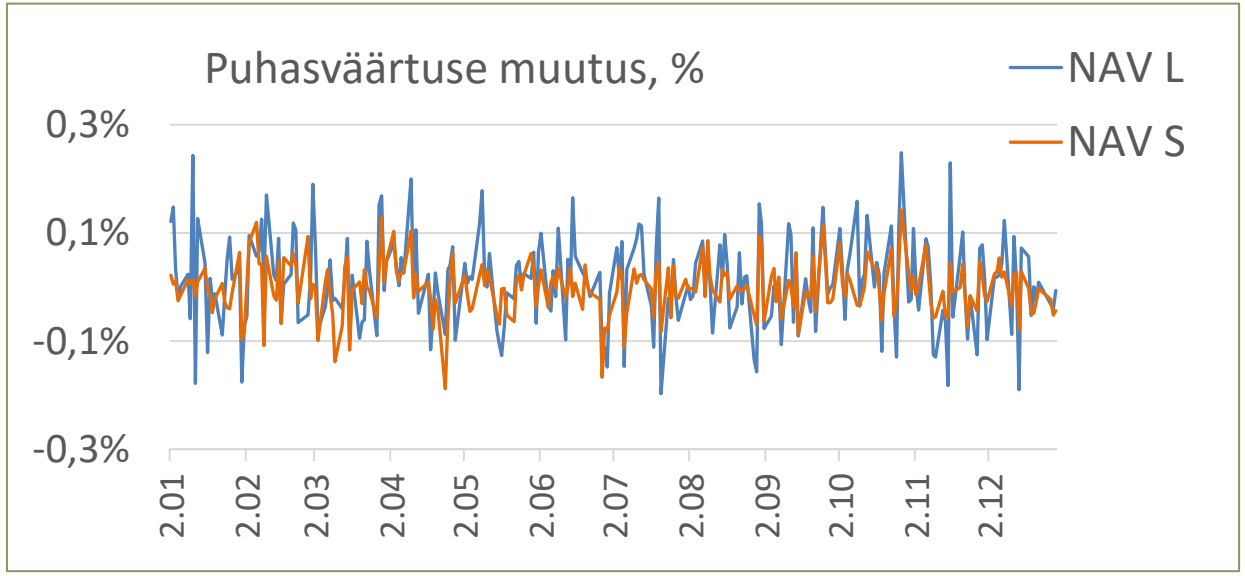
Kahe pensionifondi võrdlemine.
2.01.- 31.12.2017, 254 vaatlust

LHV pensionifond S on madala riskitasemega, konservatiivne.

LHV pensionifond L on keskmise riskitasemega, progressiivne.

Kas fondi L volatiilsus (riskitase) on oluliselt suurem?





NÄIDE: PENSIONIFONDIDE VOLATIILSUS, 2

Volatiilsuse kvantitatiivne näitaja on dispersioon.
Testime dispersioone, kasutades *F*-testi.
Kas fondi L dispersioon on oluliselt suurem?

$$H_0 : \sigma_L^2 \leq \sigma_S^2$$

$$H_1 : \sigma_L^2 > \sigma_S^2$$

LHV pensionifond L (progressiivne) $s_L^2 = 1{,}68 \cdot 10^{-4}$
LHV pensionifond S (konservatiivne) $s_S^2 = 9{,}86 \cdot 10^{-6}$

$$F = \frac{1{,}68 \cdot 10^{-4}}{9{,}86 \cdot 10^{-6}} = 17{,}1 > F_{kr} = 1{,}2$$

Võtta vastu sisukas hüpotees:
fondi L volatiilsus on oluliselt suurem kui fondi S oma.

OSAKAALU TESTIMINE

NÄIDE: VAESUSRISK EUROOPA LIIDUS JA EESTIS

Vaesusrisk tähendab seda, kui isiku sissetulek jääb allapoole suhtelist vaesuspiiri, mis on 60% sissetulekute mediaanist.

2013. aastal elas Euroopa Liidus 16,8% elanikkonnast vaesusriskis.

Eesti Sotsiaaluuring 2013. Valimi maht 15053 isikut, nendest 2770 elasid suhtelisest vaesuspiirist allpool. Vaesusriskis elavate inimeste osakaal valimis

$$\hat{p} = \frac{2770}{15053} = 0{,}184 = 18{,}4\%$$

Kas Eestis on selle uuringu järgi vaesusriskis elavate inimeste osakaal suurem kui Euroopa Liidus keskmiselt?

OSAKAALU TESTIMINE, ÜKS VALIM

Kahepoolne	Ühepoolne	Ühepoolne
$H_0 : p = p_0$	$H_0 : p \geq p_0$	$H_0 : p \leq p_0$
$H_1 : p \neq p_0$	$H_1 : p < p_0$	$H_1 : p > p_0.$

Osakaal valimis $\hat{p}$, maht n

Suurte valimite korral z-test:

$$z = \frac{\hat{p} - p_0}{se_p}$$

Standardvea leidmisel võetakse osakaaluks kogumis nullhüpoteesiga püstitatud väärtus p_0.

$$se_p = \sqrt{\frac{p_0(1-p_0)}{n}}$$

Kriitilised väärtused standardiseeritud normaaljaotusest.

NÄIDE: VAESUSRISK EUROOPA LIIDUS JA EESTIS

Kas Eestis on vaesusriskis elavate inimeste osakaal suurem kui Euroopa Liidus keskmiselt (16,8%)?

$$H_0 : p \le 0{,}168$$

$$H_1 : p > 0{,}168$$

Ühepoolne hüpotees

$$se_p = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0{,}168(1-0{,}168)}{15053}} \approx 0{,}003.$$

$$z = \frac{\hat{p} - p_0}{se_p} = \frac{0{,}184 - 0{,}168}{0{,}003} \approx 5{,}3$$

Kriitiline väärtus ühepoolse hüpoteesi korral olulisuse nivool 0,05 on 1,64.

Võtame vastu H_1.
Võime väita, et Eestis on vaesusriskis elavate inimeste osakaal suurem kui Euroopa Liidus keskmiselt.

NÄIDE: NAISTE OSAKAAL ÄRIÜHINGUTE JUHATUSTES

2013. aasta aprillis viidi Euroopa Liidus läbi uuring liidripositsioonide hõlvamise kohta naiste ja meeste poolt.

Naisjuhtide osakaal äriühingute juhatustes keskmiselt
Prantsusmaal 26,8% (valim 35 ettevõtet),
Inglismaal 18,5% (valim 46 ettevõtet).

Kas selle uuringu järgi võib järeldada, et Prantsusmaal kuulub keskmiselt rohkem naisi äriühingute juhatusse kui Inglismaal?

OSAKAALU TESTIMINE, KAKS VALIMIT

Kahepoolne	Ühepoolne	Ühepoolne
$H_0 : p_1 = p_2$	$H_0 : p_1 \geq p_2$	$H_0 : p_1 \leq p_2$
$H_1 : p_1 \neq p_2$	$H_1 : p_1 < p_2$	$H_1 : p_1 > p_2$

Valim 1 osakaal $\hat{p}_1$, maht n_1
Valim 2 osakaal $\hat{p}_2$, maht n_2

$$z = \frac{\hat{p}_1 - \hat{p}_2}{se_{p_1 - p_2}}$$

$$se_{p_1 - p_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Kriitilised väärtused standardiseeritud normaaljaotusest.

NÄIDE: NAISTE OSAKAAL ÄRIÜHINGUTE JUHATUSTES

Naisjuhtide osakaal äriühingute juhatustes keskmiselt Prantsusmaal 26,8% (valim 1, 35 ettevõtet), Inglismaal 18,5% (valim 2, 46 ettevõtet).

Kas Prantsusmaal kuulub keskmiselt rohkem naisi äriühingute juhatusse kui Inglismaal?

$$H_0 : p_1 \leq p_2$$
$$H_1 : p_1 > p_2$$

Ühepoolne hüpotees

$$z = \frac{\hat{p}_1 - \hat{p}_2}{se_{p_1 - p_2}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

$$z = \frac{0,268 - 0,185}{\sqrt{\frac{0,268 \cdot (1-0,268)}{35} + \frac{0,185 \cdot (1-0,185)}{46}}} = 0,881$$

Kriitiline väärtus ühepoolse hüpoteesi korral olulisuse nivool 0,05 on 1,64.

Võtta vastu H_0.
Ei saa väita, et Prantsusmaal on naisi äriühingute juhatustes rohkem kui Inglismaal.

NÄIDE: KODUTUD JA KODU OMAVAD PERED

Los Angeleses ca 35-50 tuhat kodutut, nendest ca 1/3 on peredena.
Uuring 1987-88: kodutud pered ja vaesed, kuid kodu omavad pered.

	Kodutud pered	Kodu omavad pered
Valimi maht	195	189
Ema kasvatas vallasema	40%	47%
Ema vanemad lahutatud	43%	36%
Ema vanemad tarvitasid alkoholi või narkootikume	49%	34%

Allikas: Wood, D.; Valdez, R. B.; Hayashi, T. & Shen, A. (1990), Homeless and housed families in Los Angeles: a study comparing demographic, economic, and family function characteristics. *Am J Public Health, American Journal of Public Health, American Public Health Association,* Vol. 80 , 1049-1052

NÄIDE: KODUTUD JA KODU OMAVAD PERED

Los Angeleses ca 35-50 tuhat kodutut, nendest ca 1/3 on peredena.
Uuring 1987-88: kodutud pered ja vaesed, kuid kodu omavad pered.

	Kodutud pered	Kodu omavad pered	z -statistik	Kriitiline	
Valimi maht	195	189			
Ema kasvatas vallasema	40%	47%	-1,39	-1,96	H_0
Ema vanemad lahutatud	43%	36%	1,41	1,96	H_0
Ema vanemad tarvitasid alkoholi või narkootikume	49%	34%	3,02	1,96	H_1

Allikas: Wood, D.; Valdez, R. B.; Hayashi, T. & Shen, A. (1990), Homeless and housed families in Los Angeles: a study comparing demographic, economic, and family function characteristics. *Am J Public Health, American Journal of Public Health, American Public Health Association,* Vol. 80 , 1049-1052

PARAMEETRILISED JA MITTEPARAMEETRILISED TESTID

Parameetrilised testid: teststatistik arvutatakse testitava suuruse väärtuse põhjal (aritmeetilised keskmised, dispersioonid). Testitav suurus peab olema intervallskaalas. Näiteks

- *t*-test
- *F*-test

Mitteparameetrilisi teste kasutatakse juhul, kui uuritava tunnuse mõõtmiseks ei saa kasutada intervallskaalat. Teststatistiku leidmisel ei kasutata tunnuste väärtusi. Kasutatakse näiteks

- märkide „+“ ja „–“ esinemissagedusi: märgitest;
- väärtuste esinemissagedusi: χ^2 - test (hii-ruut test).

MÄRGITEST

NÄIDE: MÄRGITEST, 1

Kas punast värvi autosid eelistatakse rohkem kui siniseid?

Küsitleme 10 inimest, juhuvalim.

Tähistame:
eelistab punast värvi +
eelistab sinist värvi -

Nullhüpotees
punase eelistajate arv = sinise eelistajate arv
"+" arv = "-" arv
punase eelistajate osakaal p = 0,5

EKVIVALENTSED FORMULEERINGUD

Sisukas hüpotees (ühepoolne)
punase eelistajate arv > sinise eelistajate arv
"+" arv > "-" arv
punase eelistajate osakaal p > 0,5

NÄIDE: MÄRGITEST, 2

Kriitilise väärtuse leidmisel aluseks binoomjaotus B(n,p).

n =10 katsete arv ehk valimi maht
p =0,5 nullhüpoteesile vastav osakaal

Kui kehtib nullhüpotees, st. punast ja sinist värvi eelistatakse võrdselt, siis tõenäosus et valimisse satub m sinise eelistajat on $P(X=m)$.

Kui 0 või 1, siis tõenäosus 0,011
Kui 0 või 1 või 2 siis tõenäosus 0,055

Kui olulisuse nivooks võtta 0,05, siis kriitiline väärtus on 2.

Kui 10 hulgast vähem kui 2 eelistab sinist, on nullhüpotees ümber lükatud: punase eelistajaid on rohkem.

Binoomjaotuse valemist

m	P(x= m)
0	0,001
1	0,010
2	0,044
3	0,117
4	0,205
5	0,246
6	0,205
7	0,117
8	0,044
9	0,010
10	0,001

NÄIDE: KORRIGEERITUD VALIMI MAHT

Kas punast värvi autosid eelistatakse rohkem kui siniseid?

Küsitleme 10 inimest, juhuvalim.

punast	+	vastasid 5 inimest
sinist	-	vastasid 2 inimest
ei ole eelistust		vastasid 3 inimest

KORRIGEERITUD VALIM

KORRIGEERITUD VALIMI MAHT 7

Kriitilise väärtuse leidmisel kasutame seda.

Kriitiline väärtus: 1 sinise pooldaja (ehk 6 punase pooldajat).

MÄRGITESTI KRIITILISED VÄÄRTUSED

Vasakpoolne kriitiline väärtus N_{krv} leitakse nii, et see on vähim arv, mille korral kehtib võrratus

$$P(N^+ \le N_{krv}) \ge \frac{\alpha}{2}$$

Tõenäosuse $P(N^+ \le N_{krv})$ saab leida binoomjaotusest B(n, 0,5)

Parempoolne kriitiline väärtus $N_{krp} = n - N_{krv}$

Näiteks, kui n=10, siis binoomjaotusest B(10, 0,5)

m	0	1	2	3
$P(N^+ = m)$	0,001	0,010	0,044	0,117
$P(N^+ \le m)$	0,001	0,011	0,055	0,172

$\alpha = 0,05, \quad P(N^+ \le N_{krv}) \ge 0,025 \qquad N_{krv} = 2 \quad N_{krp} = 10 - 2 = 8$

NÄIDE: MÄRGITEST INTERVALLSKAALA KORRAL

Kas veondus- ja laondusettevõtete müügitulu ühe töötaja kohta oli 2009. aastal väiksem kui 2008. aastal? Valimis 10 ettevõtet.

Ettevõtte ID	2008	2009	„+“ kui 2009. a on väiksem
1	112,6	115,3	-
2	147,2	145,5	+
3	60,9	52,6	+
4	111,7	89,1	+
5	186,5	180,8	+
6	164,2	130,7	+
7	73,3	55,6	+
8	28,7	37,4	-
9	60,9	38,7	+
10	405,3	310,9	+

Ühepoolne hüpotees, valimi maht n=10.

$$N_{krv} = 2 \quad N_{krp} = 10 - 2 = 8$$

Võtta vastu H_0 kui $N^+ \leq 8$

H_1 kui $N^+ > 8$

Tabelist N^+=8.

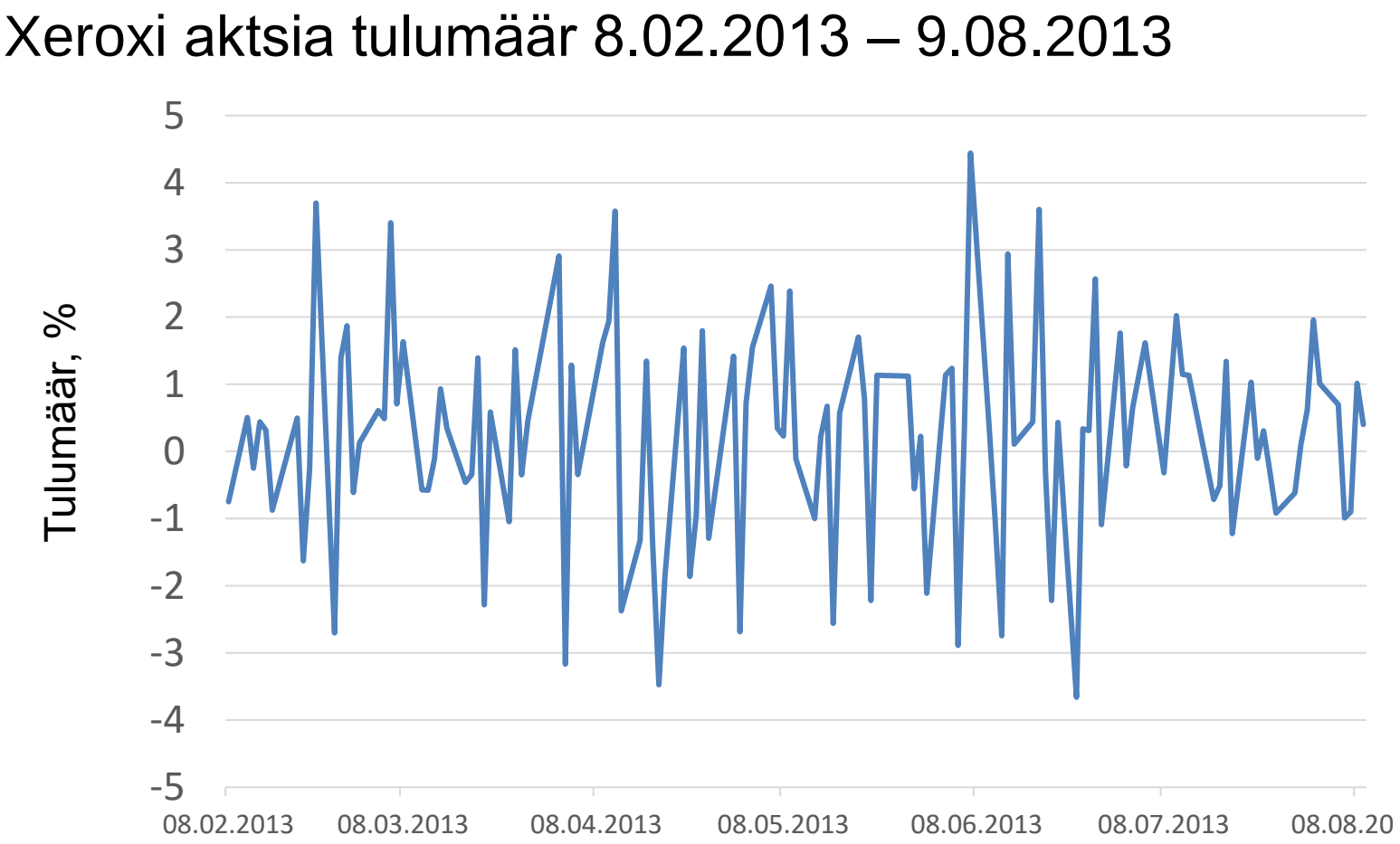
Märgitest ei tõesta, et veondus- ja laondusettevõtete müügitulu ühe töötaja kohta oli 2009. aastal väiksem kui 2008. aastal.

χ^2 TEST

MILLAL KASUTATAKSE χ^2 TESTI

- Jaotuse sobivuse testimine.
  - Kas valitud teoreetiline jaotus sobib empiirilise jaotuse kirjeldamiseks?
- Kahe kvalitatiivse tunnuse vahelise seose testimine.

NÄIDE: AKTSIA TULUMÄÄR JA NORMAALJAOTUS, I



Kas tulumäär allub normaaljaotusele?

NÄIDE: AKTSIA TULUMÄÄR JA NORMAALJAOTUS, II

Intervallime ja moodustame sagedustabeli

Klassi nr *i*	Ülemised piirid u_i	n^e
1	-2,4	8
2	-1,38	8
3	-0,36	22
4	0,66	39
5	1,68	28
6	2,7	10
7	3,72	6
8	4,74	1
KOKKU		122

Klasside sagedused

NÄIDE: AKTSIA TULUMÄÄR JA NORMAALJAOTUS, III

Võrdlemiseks kasutame normaaljaotusest leitud sagedusi n^o.

Normaaljaotusest

Klassi nr *i*	Ülemised piirid u_i	n^e	$F(u_i)$	p_i	n^o
1	-2,4	8	0,048	0,048	5,89
2	-1,38	8	0,156	0,108	13,12
3	-0,36	22	0,359	0,203	24,76
4	0,66	39	0,613	0,255	31,07
5	1,68	28	0,826	0,213	25,93
6	2,7	10	0,944	0,118	14,39
7	3,72	6	0,987	0,043	5,31
8	4,74	1		0,013	1,54
KOKKU		122		1	122

Võrdleme neid sagedusi

Vastava normaaljaotuse keskväärtus 0,208 ja standardhälve 1,57 on leitud empiiriliste andmete põhjal.

χ^2 –TEST JA TESTSTATISTIK

Nullhüpotees: empiiriline ja teoreetiline jaotus langevad kokku. St erinevus puudub, on null.
Sisukas hüpotees: empiiriline ja teoreetiline jaotus erinevad oluliselt.

Teststatistiku empiiriline väärtus

$$\chi^2 = \sum \frac{\left(n^e - n^o\right)^2}{n^o}$$

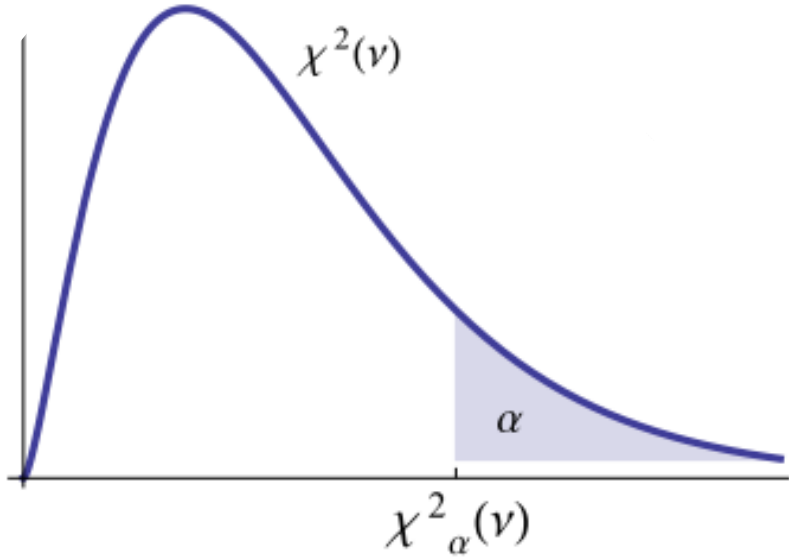
Nullhüpoteesile vastav väärtus $\chi^2 = 0$

Kui palju võib χ^2 empiiriline väärtus erineda arvust 0, kui kehtib nullhüpotees?

Kriitiline väärtus χ^2-jaotusest.

χ^2 TEST, KRIITILISTE VÄÄRTUSTE LEIDMINE

Kriitiline väärtus $\chi^2(\nu)$ jaotusest.
Excelis funktsioon CHIINV.



Määratud
- olulisuse nivooga;
- vabadusastmete arvuga.

Vabadusastmete arv $\nu = k - r - 1$

k on klasside arv ja *r* oodatava jaotuse parameetrite arv.

Parameetrite arv

ühtlane jaotus	0
binoomjaotus	1 (positiivse sündmuse esinemise tõenäosus)
Poissoni jaotus	1 (keskväärtus λ)
normaaljaotus	2 (keskväärtus ja standardhälve)

NÄIDE: AKTSIA TULUMÄÄR JA NORMAALJAOTUS, IV

Arvutame teststatistiku ja võrdleme kriitilisega

Klassi nr *i*	Ülemised piirid u_i	n^e	$F(u_i)$	p_i	n^o	$\frac{(n^e - n^o)^2}{n^o}$
1	-2,4	8	0,048	0,048	5,89	0,760
2	-1,38	8	0,156	0,108	13,12	1,997
3	-0,36	22	0,359	0,203	24,76	0,308
4	0,66	39	0,613	0,255	31,07	2,022
5	1,68	28	0,826	0,213	25,93	0,165
6	2,7	10	0,944	0,118	14,39	1,338
7	3,72	6	0,987	0,043	5,31	0,091
8	4,74	1		0,013	1,54	0,187
KOKKU		**122**		**1**	**122**	**6,867**

χ^2

Klasside arv k = 8. Normaaljaotuse parameetrite arv r = 2.

Vabadusastmete arv $v=k-r-1= 8 -2 -1 = 5$

Parameetri kriitiline väärtus olulisuse nivool 0,05 vabadusastmete arvuga 5: $\chi^2_{kr} = 11,07$

$6,867 < 11,07$

H_0: tulumäär allub normaaljaotusele

NÄIDE: KAS RIIKLIKU STATISTIKAGA MANIPULEERITAKSE?

German Economic Review 12(3): 243–255

Avaldatud 2011.a

Fact and Fiction in EU-Governmental Economic Data

Bernhard Rauch and Max Göttsche
University of Regensburg

Gernot Brähler
Ilmenau University of Technology

Stefan Engel
Catholic University of Eichstätt-Ingolstadt

Statistiliste aruannete analüüsimiseks kasutati Benfordi jaotust.
Jaotus iseloomustab numbrite 0,1,2,…,9 esinemist mitmesugustes reaalsest elust võetud arvudes (arved, hinnad, jõgede pikkused jne).

Näiteks mingi arvu esimesel kohal olevate numbrite tõenäosusjaotus:

d	1	2	3	4	5	6	7	8	9
P(d)	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046

NÄIDE: KAS RIIKLIKU STATISTIKAGA MANIPULEERITAKSE?

H_0: andmed alluvad Benfordi jaotusele, on tekkinud loomulikul teel.
H_1: ei allu, ei ole tekkinud loomulikul teel.

$X^2_{kr} = 15{,}51$

Euroala maad, aastad 1999 - 2009

	Keskmine χ^2	Astak
Kreeka	17,74	1
Belgia	17,21	2
Austria	15,25	3
…..	….	….
Portugal	10,19	15
Holland	7,83	16

Euroalast väljas olevad riigid, aastad 1999 - 2009

	Keskmine χ^2	Astak
Rumeenia	19,30	1
Läti	17,51	2
Eesti	15,86	3
…..	….	….
Ungari	10,07	10
Poola	9,74	11

Rauch, B., Göttsche, M., Engel, S., & Brähler, G. (2011). Fact and fiction in EU-governmental economic data. German Economic Review, 12(3), 243-255.

NÄIDE: KAS RIIKLIKU STATISTIKAGA MANIPULEERITAKSE?

Eesti lõpetas EL liitumisläbirääkimised

Eesti astus Euroopa Liitu

EESTI

1999	2000	2001	2002	2003	2004
16,46**	26,78**	10,53	3,79	15,09	26,73**

$X^2_{kr} = 15{,}51$

2005	2006	2007	2008	2009
18,12**	16,70**	10,76	9,71	6,82

2010. aastal sai Eesti eurotsoonilt nõusoleku liitumiseks.

** statistiliselt oluline nivool 0,05

Rauch, B., Göttsche, M., Engel, S., & Brähler, G. (2011). Fact and fiction in EU-governmental economic data. German Economic Review, 12(3), 243-255.

NÄIDE: χ^2 TEST JA KAHE TUNNUSE VÕRDLEMINE

Tootja soovib kontrollida hüpoteesi, et konkreetse kauba meeldivus sõltub ostja soost.

Nullhüpotees: ostja sugu ei mõjuta kauba meeldivust.
Sisukas hüpotees: ostja sugu mõjutab kauba meeldivust.

Hüpoteesi kontrollimiseks küsitleti 60 meest ja 90 naist ning saadi järgmised tulemused:

	mehed	naised	KOKKU
meeldib	17	14	31
neutraalne	29	45	74
ei meeldi	14	31	45
KOKKU	60	90	150

empiiriline jaotus

NÄIDE: OODATAVA JAOTUSE ARVUTAMINE

empiiriline jaotus

	mehed	naised	KOKKU
meeldib	17	14	31
neutraalne	29	45	74
ei meeldi	14	31	45
KOKKU	60	90	150

40% 60%

Meeste osakaal

$$\frac{60}{150} = 0,4$$

Nullhüpoteesi korral mehi, kellele meeldib

$$31 \times 0,4 = 12,4$$

Mehi, kes neutraalsed

$$74 \times 0,4 = 29,6$$

oodatav jaotus

	mehed	naised	KOKKU
meeldib	12,4	18,6	31
neutraalne	29,6	44,4	74
ei meeldi	18	27	45
KOKKU	60	90	150

NÄIDE: χ^2 TESTI EMPIIRILISE VÄÄRTUSE ARVUTAMINE

	Empiiriline sagedus n^e	Oodatav sagedus n^o	$\frac{(n^e - n^o)^2}{n^o}$
mehed, meeldib	17	12,4	1,706
mehed, neutraalne	29	29,6	0,012
mehed, ei meeldi	14	18	0,889
naised, meeldib	14	18,6	1,138
naised, neutraalne	45	44,4	0,008
naised, ei meeldi	31	27	0,593
KOKKU	150	150	4,346

χ^2

Kriitiline väärtus 5,99.

Vastu võtta H_0: kauba meeldivus ei sõltu ostja soost.

NÄIDE: SUHTUMINE PROGRESSEERUVASSE TULUMAKSU

Aastal 2000 üliõpilaste poolt läbiviidud uuring, vastajaid 90.
Vastati erinevatele küsimustele, vstusevariandid: Jah / Ei
Lisati hinnang oma majanduslikule olukorrale: väga halb, pigem halb, pigem hea, väga hea.
Kas vastus sõltub vastaja majanduslikust olukorrast?
Nullhüpotees: ei sõltu.

Küsimus	Olulisuse tõenäosus
3. Kas pooldate progresseeruva tulumaksu kehtestamist?	0,65
4. Kas progresseeruv tulumaks võib halvendada teie majanduslikku olukorda?	0,71
5. Kas progresseeruva tulumaksu kehtestamine Eestis võib suurendada maksude vältimist maksumaksjate poolt?	0,90
6. Kas olete nõus väitega, et progresseeruv tulumaksu süsteem piirab jõukamate inimeste vajadusi?	0,17
7. Kas olete nõus väitega, et progresseeruva tulumaksu korral kolivad rikkad riigist ära, sinna kus on suuremad maksusoodustused?	0,20

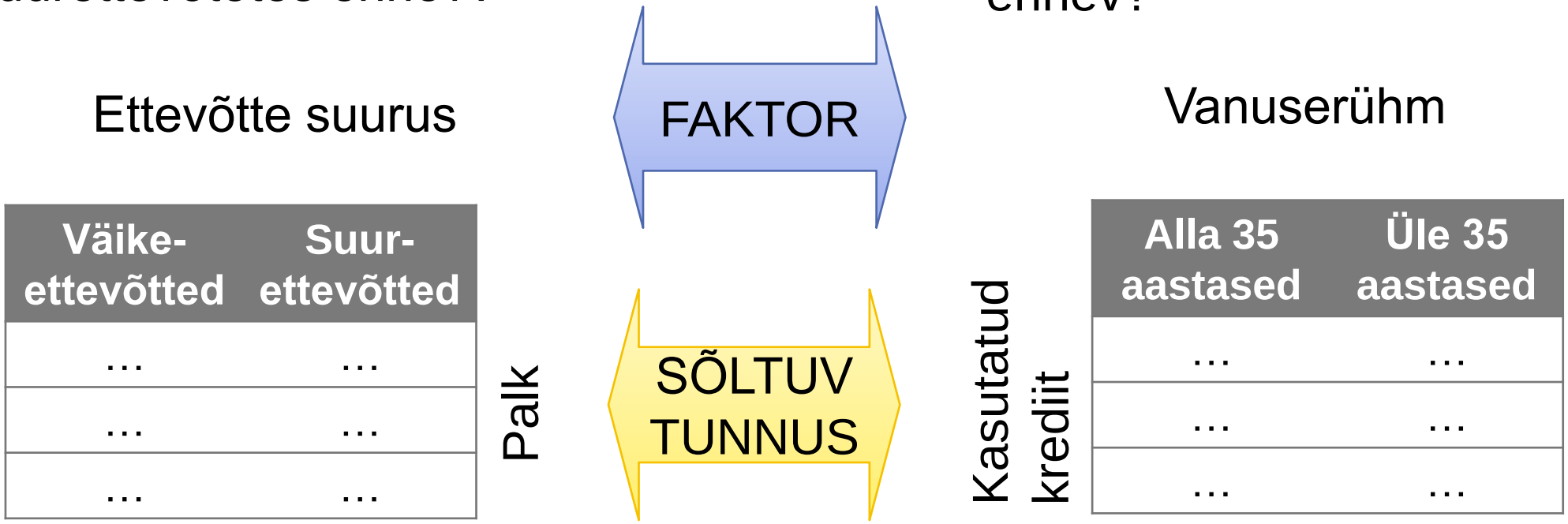
Kõigi küsimuste korral võtta vastu nullhüpotees: vastus ei sõltu vastaja majanduslikust olukorrast.

DISPERSIOON-ANALÜÜS

KAHE VALIMI VÕRDLEMINE: t-TEST

Kas naisjuhtide palgatase on väikeettevõtetes ja suurettevõtetes erinev?

Kas krediitkaardi kasutamine on kahes erinevas vanuserühmas erinev?



Faktor kas mõjutab või ei mõjuta sõltuvat tunnust.

ROHKEM KUI KAKS VALIMIT JA t-TEST

t-testi korral on faktoril maksimaalselt kaks taset.
Saame võrrelda maksimaalselt kahe valimi keskväärtust.

Kui aga faktoril on rohkem kui kaks taset?

Ettevõtted: väikesed, keskmised, suured. Kolm taset, st kolm valimit.

Mitu paarikaupa võrdlemist tuleb teha?

Kas palk sõltub ettevõtte tegevusalast? Tegevusalasid 21.

Mitu paarikaupa võrdlemist tuleb teha? 210

Kombinatsioonide arv 21-st kahekaupa $C_{21}^{2} = \frac{21!}{2!19!} = \frac{20 \cdot 21}{2} = 210$

ROHKEM KUI KAKS VALIMIT JA t-TEST

1. probleem: paarikaupa võrdlemisi väga palju.
Arvuti kasutamine lahendab selle probleemi.

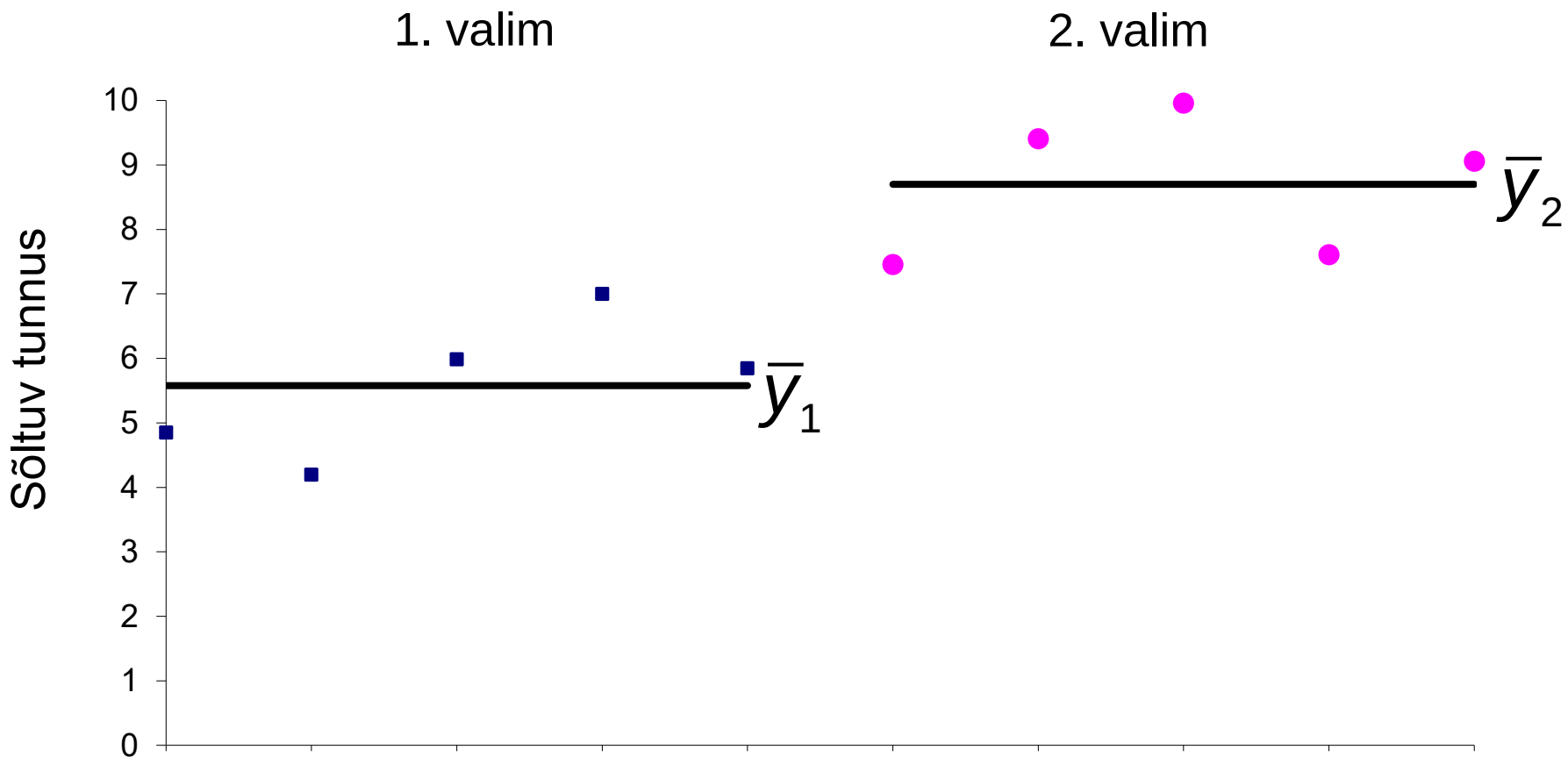
2. probleem: saame palju erinevaid tulemusi.
Iga paari kohta üks tulemus.

Näiteks 210 tulemust: võtta vastu H_0 või võtta vastu H_1.

Soovime üht tulemust: kas faktor mõjutab sõltuvat tunnust või mitte?

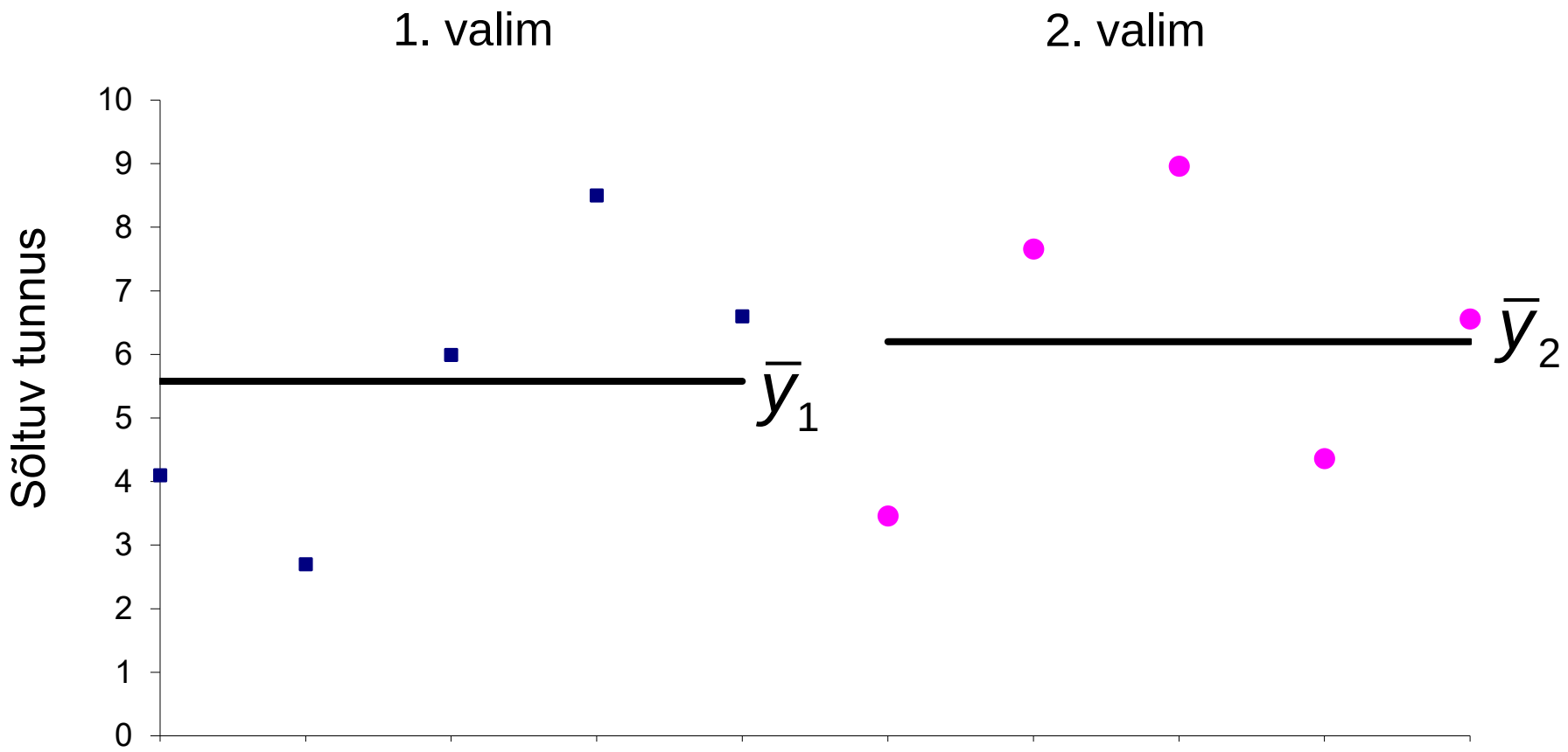
Lahendus: dispersioonanalüüs.

Punktid tähistavad üksikuid objekte
Kriipsud – valimite keskmisi



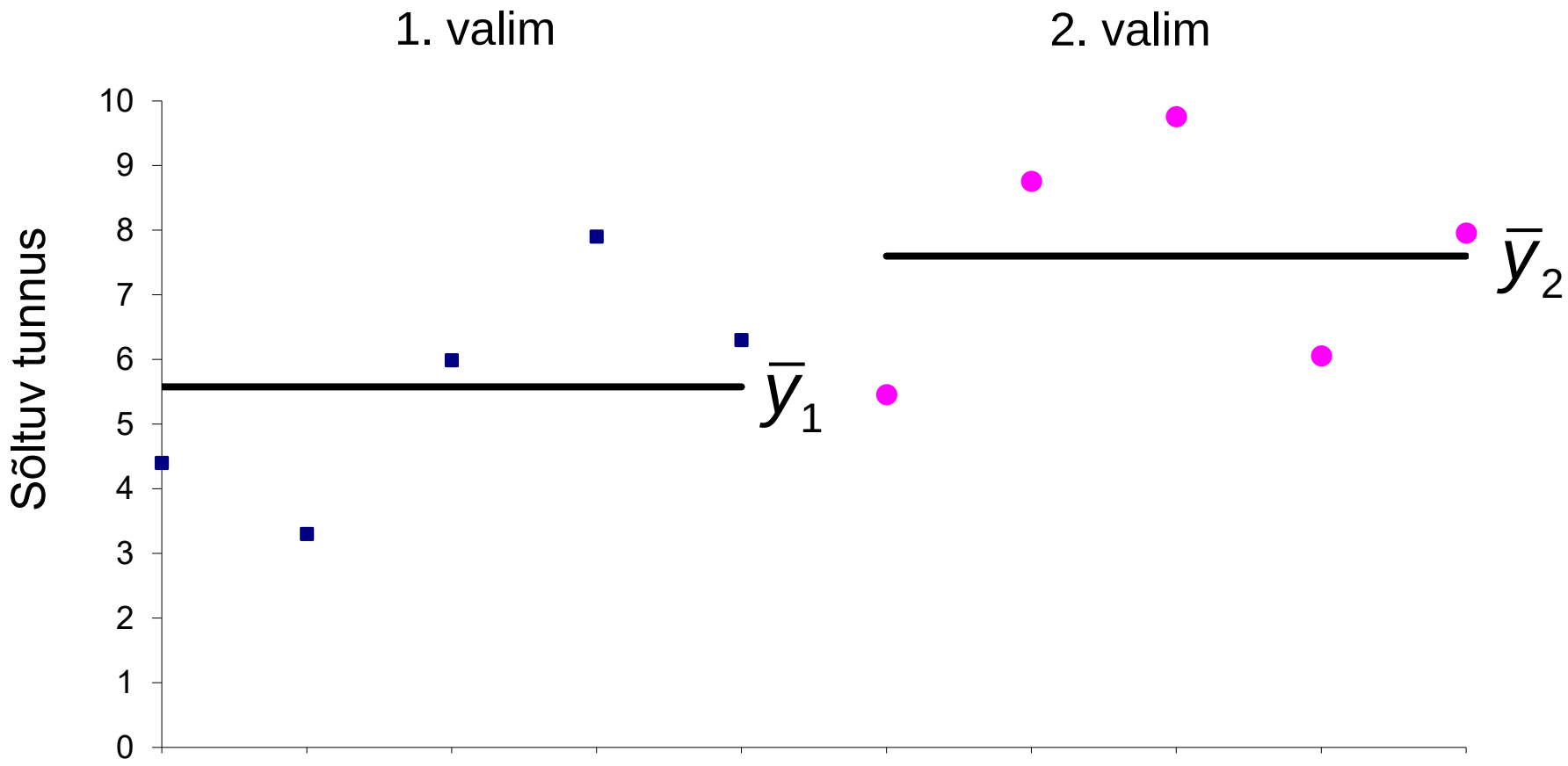
Kas valimid on ühest ja samast kogumist või kahest erineva keskväärtusega kogumist?

Punktid tähistavad üksikuid objekte
Kriipsud – valimite keskmisi



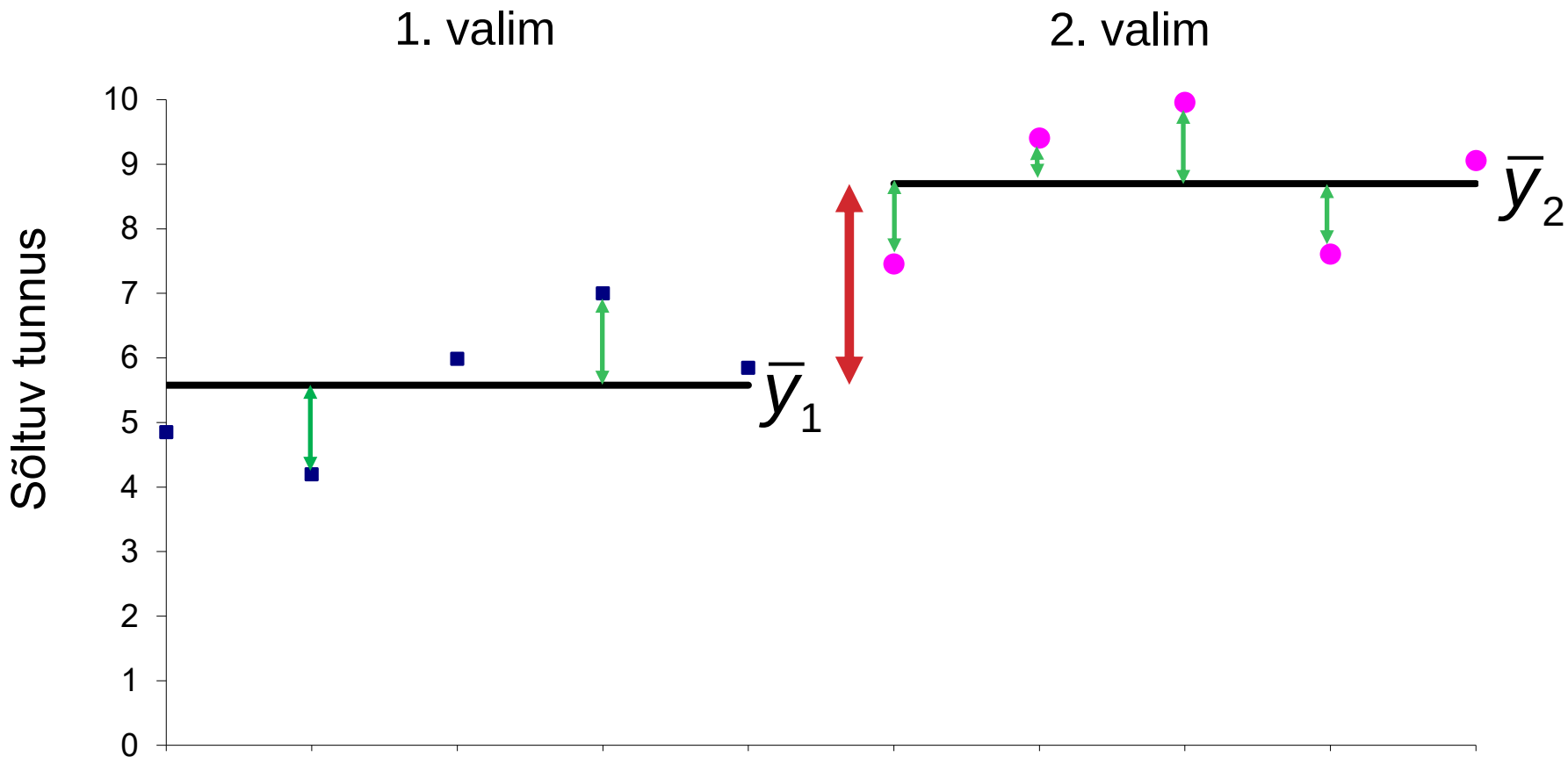
Kas valimid on ühest ja samast kogumist või kahest erineva keskväärtusega kogumist?

Punktid tähistavad üksikuid objekte
Kriipsud – valimite keskmisi



Kas valimid on ühest ja samast kogumist või kahest erineva keskväärtusega kogumist?

Punktid tähistavad üksikuid objekte
Kriipsud – valimite keskmisi

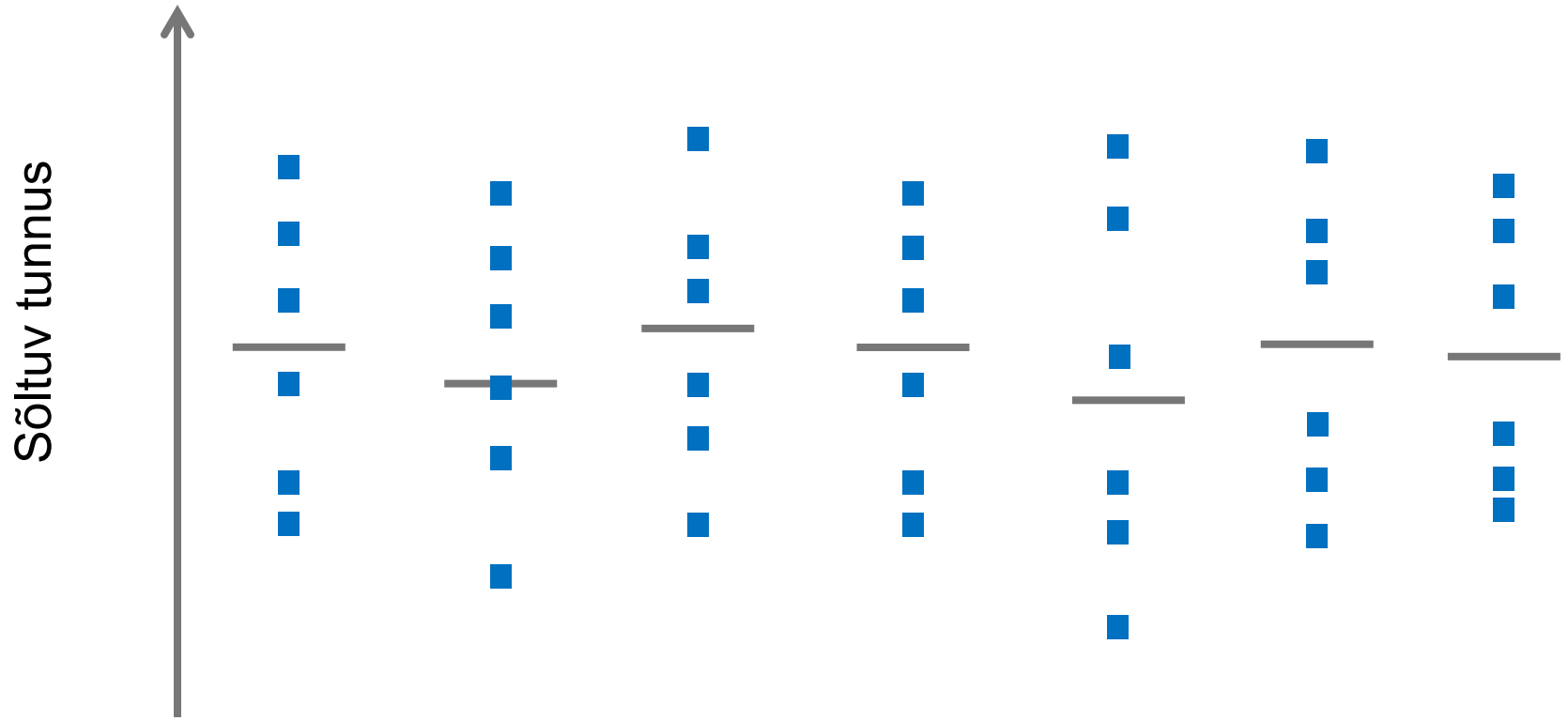


Üksikute punktide kaugust vastava valimi keskmisest

võrreldakse valimite keskmiste vahelise kaugusega

Punktid tähistavad üksikuid objekte
Kriipsud – valimite keskmisi

PALJU VALIMEID



Vaja objektiivset, arvkarakteristikut.

F-STATISTIK

$$F = \frac{MST}{MSE}$$

$$\frac{\text{rühmadevaheline, seletatud hajumine}}{\text{rühmasisene, seletamata hajumine}}$$

Demo: ANOVA

Kui suur peab *F* olema, et võiksime öelda: faktori põhjustatud seletatud hajumine on oluline?

Vaja kriitilist väärtust!

HÜPOTEESI KONTROLLIMINE

$$F = \frac{MST}{MSE}$$ allub *F*-jaotusele

Kriitilised väärtused *F*-jaotusest.

Nullhüpotees: funktsioontunnuse keskväärtused on kõikides rühmades võrdsed, **faktori mõju puudub**.
Sisukas hüpotees: leidub vähemalt kaks rühma, mille korral rühmade keskväärtused on oluliselt erinevad, **faktor mõjutab.**

Võtame vastu sisuka hüpoteesi, kui $F > F_{kr}$ ($p < \alpha$)

NÄIDE: TOOTLIKKUS ÜHE TÖÖTAJA KOHTA

Kas tootlikkus ühe töötaja kohta on erinevatel tegevusaladel erinev?
Valimis 340 ettevõtet viielt tegevusalalt.

Aruanne Excelis

SUMMARY

Groups	Count	Sum	Average	Variance
Tööstus ja energeetika	135	195015	1445	6311093
Jaekaubandus	56	168537	3001	13552978
Hulgikaubandus	62	1334470	21524	4,17E+09
Ehitus	70	115478	1650	2853717
Toiduainetetööstus	17	19153	1127	458631

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1,98E+10	4	4,96E+09	6,49	4,87E-05	2,40
Within Groups	2,56E+11	335	7,65E+08			
Total	2,76E+11	339				

Võtame vastu H_1, sest 6,49 > 2,40. Tootlikkus on erinevatel tegevusaladel erinev.

NÄIDE: TOOTLIKKUS ÜHE TÖÖTAJA KOHTA

Kas tootlikkus ühe töötaja kohta on erinevatel tegevusaladel erinev?
Valimis 340 ettevõtet viielt tegevusalalt.

Aruanne Excelis

SUMMARY

Groups	Count	Sum	Average	Variance
Tööstus ja energeetika	135	195015	1445	6311093
Jaekaubandus	56	168537	3001	13552978
Hulgikaubandus	62	1334470	21524	4,17E+09
Ehitus	70	115478	1650	2853717
Toiduainetetööstus	17	19153	1127	458631

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1,98E+10	4	4,96E+09	6,49	4,87E-05	2,40
Within Groups	2,56E+11	335	7,65E+08			
Total	2,76E+11	339				

Võtame vastu H_1, sest $4{,}87 \cdot 10^{-5} < 0{,}05$

DISPERSIOONANALÜÜS, KOKKUVÕTE

Dispersioonanalüüs on meetod, millega otsitakse vastust küsimusele, kas

- rühmakeskmiste erinevus on põhjustatud uuritava faktori mõjust

või

- valimite juhuslikkusest.

Kui erinevus on põhjustatud uuritava faktori mõjust (sisukas hüpotees), võib järgneda keskväärtuste mitmene võrdlemine.

SOBIVA TESTI VALIK

Tee kindlaks

1) mitu väärtust ehk taset on sõltumatul tunnusel (faktoril);
2) millist skaalat on kasutatud sõltuva tunnuse mõõtmisel.

???

Milliseid teste võib kasutada järgmiste probleemide analüüsimisel?
Võimalikud variandid:

t-test,
osakaalude testimine,
χ^2-test,
ANOVA.

Mõne probleemi korral on võimalik kasutada ka mitut erinevat testi.

1. Kas naiste ja meeste seas on töötus erinev? osakaalude testimine, χ^2-test

2. Kas neljas erinevas piirkonnas on keskmine palk erinev? ANOVA

3. Küsimus: kuidas tuled oma sissetulekuga toime? Vastusevariandid:

- enamasti jääb puudu,
- mõnikord jääb puudu,
- saan hakkama, jääb piisavalt üle ka investeerimiseks.

Kas vastus sõltub vastaja haridustasemest (3 taset)? χ^2-test

4. Kas avalikus sektoris ja erasektoris on keskmine tööjõukulu ühe töötaja kohta erinev? *t*-test