Lars Andersen: Anvendelse af statistik. Notat om deskriptiv statistik, χ^2-test og Goodness of Fit test.

Anvendelser af statistik
Statistik er et levende og fascinerende emne, men at læse om det er alt for ofte dræbende kedeligt. Denne note vil forsøge at anskueliggøre det levende ved statistikken ved en anvendelse af virkelige eksempler indenfor deskriptiv statistik og χ^2 – test.

Statistik er en samling metoder til at træffe fornuftbestemte afgørelser, når man er stillet overfor uvished. Et problem kan behandles statistisk ved en procedure, der kan inddeles i fire stadier:

1) Indsamling og iagttagelse af data:
   indsamling af data og indledende behandling af data, det være sig beregning af få relevante tal (gennemsnit, spredning, median..) og fremstilling af relevante diagrammer (histogrammer, lagkagediagrammer, søjlediagrammer, kassediagrammer..)
2) Hypoteser:
   formulering af en formodning, en antagelse eller et mønster i data
3) Forudsigelser og fortolkning:
   hypoteserne af- eller bekræftes, forstået på den måde, at man foregriber, hvad man vil se i undersøgelser, der endnu ikke er foretaget. Hvor stor en tillid skal man have til konklusionerne og hvilke fejlmuligheder må der regnes med.
4) Verifikation:
   indsamling af nye data til at af – eller bekræfte forudsigelserne.

Vi vil i denne note se på hvorledes punkterne 1) – 3) "kommer i sving".

Essensen i (teoretisk) statistik er metoder til at konkludere om en helhed (populationen) ud fra en stikprøve. Det er derfor afgørende ved udtagelse af elementer i stikprøven, at stikprøven viser et miniaturebillede af populationen, man siger stikprøven skal findes ved simpel og tilfældig udvælgelse.

Hvis stikprøven ikke udtages simpel og tilfældig er det ikke muligt at fæstne nogen form for tillid til de hypoteser, der er blevet af – eller bekræftet. Eksemplerne herpå er utallige:

- kort før præsidentvalget i USA, 1936, foretog man en meningsmåling, hvor stikprøven blev udtaget blandt bilejere og husstande med telefon. Udfra stikprøven konkluderede man, at den republikanske guvernør Landon ville vinde i langt de fleste stater. Imidlertid blev demokraten Franklin D. Roosevelt genvalgt med den hidtil største valgsejr i USA's historie (hvad gik galt?)
- du kan godt bunde i åen. Gennemsnitsdybden er kun 50 cm (hvad er galt?)

Man skal også være opmærksom på de få relevante tal og diagrammer, man vælger at benytte til at beskrive data. Neden for ses nogle eksempler, der illustrerer dette.

Eksempel 1:

To hold er blevet bedømt ved en evaluering i matematik. De opnåede karakterer er:

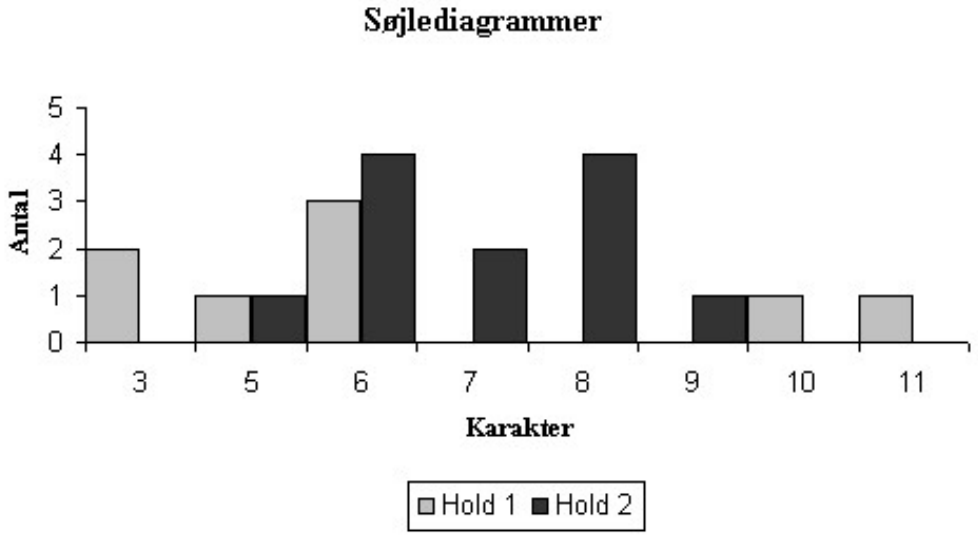
Hold 1	3	3	5	8	8	8	10	11				
Hold 2	5	6	6	6	6	7	7	8	8	8	8	9

Hvilket hold har klaret sig bedst?

De to hold har samme gennemsnit, så andre tal må "på bordet".
Nogle relevante tal kunne være minimum, maximum, variationsbredde, spredning og kvartilerne:

	Hold 1	Hold 2
Gennemsnit	7	7
Maximum	11	9
Minimum	3	5
Variationsbredde	11 – 3 = 8	9 – 5 = 4
Spredning	3,0	1,2
1. decil(10%)	3	6
1. kvartil(25%)	4,5	6
Median(50%)	8	7
3. kvartil(75%)	8,5	8
9. decil(90%)	10,3	8

Relevante diagrammer kunne være søjlediagrammer:



Eksempel 2:

Højden af eleverne på to nabo gymnasier vurderes udfra to stikprøver. Målingerne ses herunder

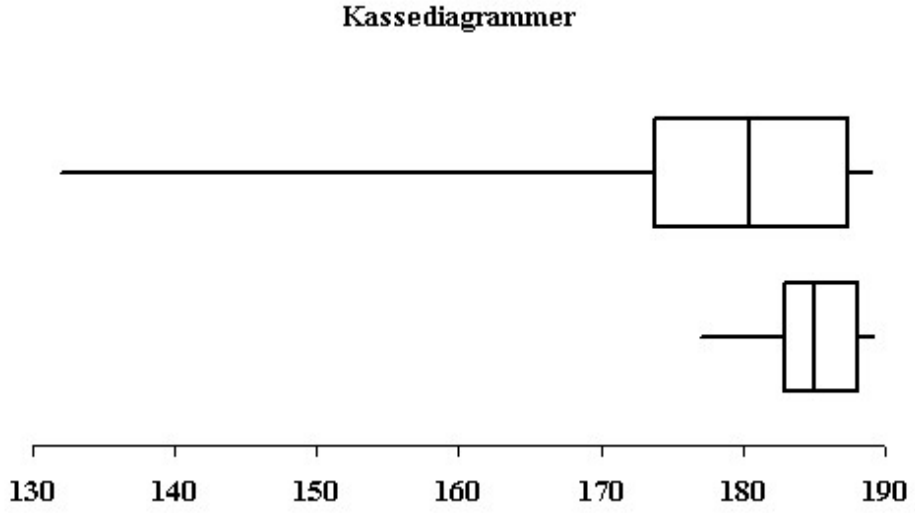
Gymnasium 1	173	183	185	188	190	
Gymnasium 2	91	173	176	185	188	190

Hvilke tal er ”bedst” til at beskrive data?

	Gymnasium 1	Gymnasium 2
Gennemsnit	183,8	168,3
Maximum	190	190
Minimum	173	91
Variationsbredde	190 – 173 = 17	190 – 91 = 99
Spredning	6,6	38,3
1. decil(10%)	177,0	132,0
1. kvartil(25%)	183,0	173,8
Median(50%)	185,0	180,5
3. kvartil(75%)	188,0	187,3
9. decil(90%)	189,2	189,0

Som det fremgår af tabellen er gennemsnittet følsomt overfor ekstreme værdier, værdier der så at sige falder uden for systemet. Dette er ikke tilfældet for medianen. I tilfælde som dette må det overvejes om ikke medianen er bedre end gennemsnittet til at beskrive et datasæt
At der i en af stikprøverne sandsynligvis findes ekstreme værdier, fremgår til dels af den høje spredning og den høje variationsbredde.

Relevante diagrammer kunne være kassediagrammer:

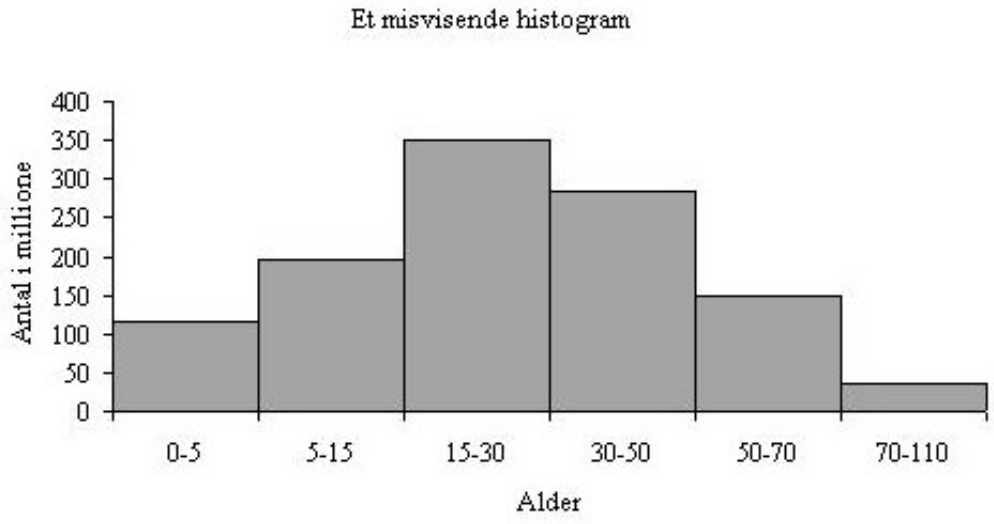


Eksempel 3:

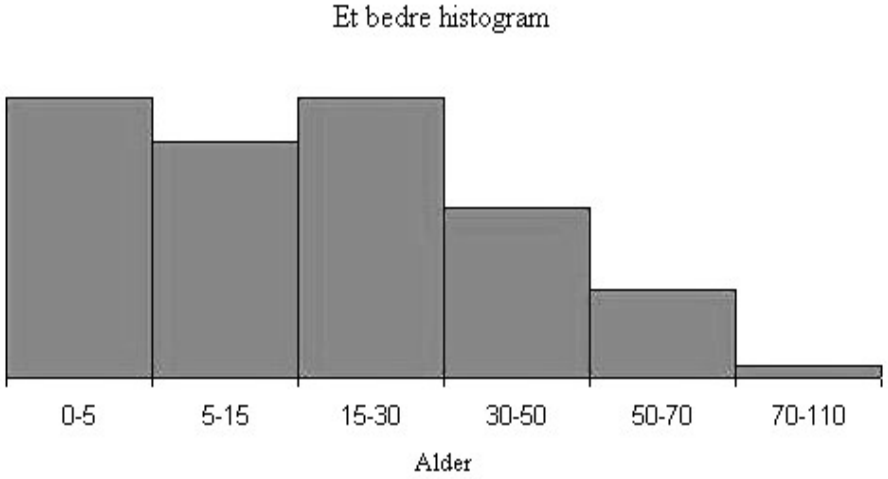
En befolkningstælling i Kina i 1990 viste følgende

Interval (alder)	0 - 4	5 - 14	15 - 29	30 - 49	50 - 69	70 - 110
Antal i millioner	116,6	196,9	350,5	283,1	147,9	36,8

Hvis man vælger at benytte disse data til at fremstille et histogram/søjlediagram uden at overveje problemerne med, at intervallerne ikke er lige bredde, vil dette se ud som på figur 1 neden for. Et bedre histogram, der tager hensyn til at intervallerne ikke er lige bredde, ser ud som på figur 2 neden for.



Figur 1



Figur 2

Bemærkning til figurerne 1 og 2: En mindre kritisk iagttager ville udfra figur 1 kunne konkludere, at børn og unge udgør en mindre andel af befolkningen i Kina. Dette må tilskrives Et-barns-

politikken i Kina. Den kritiske iagttager ville udtale, at figur 1 ikke kan benyttes som illustration heraf.

Krydstabeltest: χ^2 - test

Krydstabeller er en fællesbetegnelse for stikprøver, hvor data kan inddeles i forskellige kategorier. Den teori, der gennemgås i denne lektion, gør det da muligt at påvise, om der findes en signifikant sammenhæng mellem disse kategorier.
Bemærk, at testet ikke udtaler sig om, hvori en eventuel sammenhæng består.

Krydstabeltest omfatter to forskellige typer af test:

- test på uafhængighed
- test på homogenitet

De to nulhypoteser kan i ord formuleres som:

- H_0 : Der gælder uafhængighed mellem kategorierne
- H_0 : Der gælder homogenitet mellem kategorierne

hvor netop én af disse testes op mod den alternative hypotese:

- H_1 : Der gælder ikke uafhængighed mellem kategorierne
- H_0 : Der gælder ikke homogenitet mellem kategorierne

Om der er tale om et test på uafhængighed eller et test på homogenitet afhænger alene af den måde, hvorpå stikprøven udtages.
Der er tale om et *uafhængighedstest*, når man kun har oplysninger om stikprøvens størrelse, n.
Der er tale om et *homogenitetstest*, når man har yderligere oplysninger om stikprøvens fordeling på kategorier.

Forskellen kan illustreres ved følgende 2 små eksempler:

- udtag en stikprøve på 600 personer (uafhængighedstest)
- udtag en stikprøve på 600 personer fordelt på 400 mænd og 200 kvinder(homogenitetstest)

Den praktiske gennemførelse af et Krydstabeltest bygger på en sammenligning af to tabeller:

- en tabel over de observerede data, den såkaldte O – tabel
- en tabel over de data man ville forvente, hvis H_0 hypotesen er sand, den såkaldte E – tabel

Hvis H_0 hypotesen forkastes beregnes yderligere en tabel, den såkaldte Q – tabel, der så at sige viser, ”hvem den skyldige er”.

I noten betragtes endvidere et tredje test, det såkaldte Goodness of Fit test, hvor der her testes, om data kan anses som værende repræsentativ for populationen. Typisk vil man være interesseret i, om data er repræsentativ, hvad køn og alder angår.

Indledningsvis vises et eksempel på hvorledes opbygningen af et krydstabeltest gennemføres.

Eksempel 4:

Byrådet i en større sjællandsk provinsby har planer om at etablere et erhvervsbibliotek, hvis hovedformål er at formidle information til byens virksomheder. Da omkostningerne ved etablering af biblioteket er store gennemføres en stikprøveundersøgelse af 200 virksomheders holdning til erhvervsbiblioteket. Resultatet af undersøgelsen fremgår af nedenstående tabel:

Tabel over de observerede værdier(O – tabellen): (Antal kolonner: k = 3, antal rækker: r = 3)

Type\holdning	Interesseret	Ikke interesseret	Ved ikke	I alt
Industri	23	21	6	50
Bygge/anlæg	15	20	1	36
Service/liberalt	51	32	31	114
I alt	89	73	38	200

Hypoteser:

H_0: Der gælder uafhængighed mellem virksomhedstype og holdning
H_1: Der gælder ikke uafhængighed mellem virksomhedstype og holdning
Signifikansniveau = 0,05

Beslutningsregler :

Tabel over de forventede værdier (E – tabellen)

Type\holdning	Interesseret	Ikke interesseret	Ved ikke
Industri	22,25	18,25	9,50
Bygge/anlæg	16,02	13,14	6,84
Service/liberalt	50,73	41,61	21,66

Bemærk, at alle værdier i E – tabellen er over 5!

Teststørrelse: $T = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 16{,}61$; $T \sim \chi^2$ ($(k-1)\cdot(r-1) = 4$)

p – værdi: $P(T \geq 16{,}61) = 0{,}0023$

Da p – værdien er mindre end signifikansniveauet forkastes H_0, der gælder således ikke uafhængighed mellem virksomhedstype og holdning, der er altså forskel i virksomhedstypernes holdning til det nye bibliotek.

Da vi forkastede H_0 afsluttes med at opbygge Q – tabellen.

Tabel over de enkelte bidrag til teststørrelsen: (Q – tabellen)

Type\holdning	Interesseret	Ikke interesseret	Ved ikke
Industri	0,03	0,41	1,29
Bygge/anlæg	0,06	3,58	4,99
Service/liberalt	0,00	2,22	4,03

Bidragene til teststørrelsen opdelt efter kategorier.
Det væsentlige bidrag til teststørrelsen stammer fra Bygge/anlæg virksomheder samt virksomheder inden for Service/liberale !!

Eksempel 5:

På baggrund af en spørgeskema undersøgelse (se bilag 1) er der konstrueret krydstabeller over "Køn – Holdning (til indvandreres kontanthjælp)" ved brug af pivottabeller

Det skal nu undersøges, om der er signifikant forskel mellem mænds og kvinders holdning til indvandreres kontanthjælp samt om stikprøven er repræsentativ med hensyn til køn.

Tabel over de observerede værdier(O – tabellen): (Antal kolonner: k = 2, antal rækker: r = 2)

Køn\Holdning	Enig	Uenig	I alt
K	79	28	107
M	40	39	79
I alt	119	67	186

Hypoteser:

H_0: Der gælder uafhængighed mellem køn og holdning
H_1: Der gælder ikke uafhængighed mellem køn og holdning
Signifikansniveau = 0,05

Beslutningsregler:

Tabel over de forventede værdier (E – tabellen)

Køn\Holdning	Enig	Uenig	I alt
K	68,5	38,5	107
M	50,5	28,5	79
I alt	119	67	186

Bemærk, at alle værdier i E – tabellen er over 5!

Teststørrelse: $T = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 11{,}0$; $T \sim \chi^2$ ($(k-1)\cdot(r-1) = 1$)

p – værdi: $P(T \geq 11{,}0) = 0{,}0011$

Da p – værdien er mindre end signifikansniveauet forkastes H_0, der gælder altså ikke uafhængighed mellem køn og holdning. Der er signifikant forskel mellem mænds og kvinders holdning til indvandreres kontanthjælp.

Tabel over de enkelte bidrag til teststørrelsen: (Q – tabellen)

Køn\Holdning	Enig	Uenig	I alt
K	1,4	4,0	5,4
M	2,8	2,9	5,6
I alt	4,2	6,8	11,0

Bidragene til teststørrelsen opdelt efter kategorier. Mænd svarer generelt ikke som forventet, medens kvinderne er underrepræsenteret i kategorien "Uenig".

Goodness of Fit test

Undersøgelse af repræsentativitet med hensyn til køn:

Fordelingen af kvinder henholdsvis mænd i befolkningen antages her at være 51% - 49%. Denne fordeling skal også kunne findes i stikprøven, hvis undersøgelsen skal kunne betegnes som repræsentativ med hensyn til køn.

Hypoteser:

H_0: Andelen af kvinder er 51% og andelen af mænd er 49%
H_1: Andelen af kvinder er ikke som angivet ovenfor.

Signifikansniveau = 0,05

Beslutningsregler:

Tabel over de observerede værdier(O – tabellen):

Køn	Kvinder	Mænd	I alt
Antal	107	79	186

Tabel over de forventede værdier(E – tabellen):

Køn	Kvinder	Mænd	I alt
Antal	94,9	91,1	186

Teststørrelse: $T = \sum_i \frac{(O_{ij} - E_i)^2}{E_i} = 3{,}15$; $T \sim \chi^2$ (antal grupper – 1 = 2 – 1 = 1)

p – værdi: $P(T \geq 3{,}15) = 0{,}076$

Da p – værdien er større end signifikansniveauet accepteres H_0, stikprøven er altså repræsentativ med hensyn til køn.

Bilag:

Anvendte begreber i den deskriptive statistik:

Begreberne eksemplificeres ud fra data i eksempel 3, gymnasium 1:

	Gymnasium 1
Gennemsnit	183,8
Maximum	190
Minimum	173
Variationsbredde	190 – 173 = 17
Spredning	6,6
1. decil(10%)	177,0
1. kvartil(25%)	183,0
Median(50%)	185,0
3. kvartil(75%)	188,0
9. decil(90%)	189,2

Gennemsnit, $\bar{x} = 183{,}8$

Formlen til beregning af det almindelige gennemsnit er: $\bar{x} = \frac{\sum x_i}{n}$, hvor

x_i = den enkelte observation

n = stikprøvens størrelse

Spredning, s = 6,6:

Spredningen er den gennemsnitlige kvadratiske afvigelse fra gennemsnittet, altså et mål for, hvor meget observationerne spreder sig.

Formlen til beregning af spredning er her givet ved: $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

1.decil(10%) = 177,0:

10% af eleverne har en højde på 177 cm eller derunder

1.kvartil(25%) = 183,0:

25% af eleverne har en højde på 183 cm eller derunder

Median(50%) = 185,0:

50% af eleverne har en højde på 185 cm eller derunder

3.kvartil(75%) = 188,0:

75% af eleverne har en højde på 188 cm eller derunder

9.decil(90%) = 189,2:

00% af eleverne har en højde på 189,2 cm eller derunder

Fremstilling af histogram (ikke ens intervalbredde):

Udgangspunktet er data stammende fra en befolkningstælling i Kina i 1990.

Hvis histogrammet korrekt skal vise fordelingen af befolkningen, må der tages højde for, at intervallerne over alder ikke har samme bredde. Dette gøres ved at fremstille rektangler, hvor arealet er et mål for antal observationer i intervallet. Højden af rektanglet findes derfor ved følgende beregning:

$$\text{højde} = \frac{\text{antal}}{\text{bredde}}$$

Interval (alder)	0 - 4	5 - 14	15 - 29	30 - 49	50 - 69	70 – 109
Antal i millioner	116,6	196,9	350,5	283,1	147,9	36,8
Bredde af interval	5	10	15	20	20	40
Højde af rektangel	$\frac{116,6}{5} = 23,3$	$\frac{196,9}{10} = 19,7$	$\frac{350,5}{15} = 23,4$	$\frac{283,1}{20} = 14,2$	$\frac{147,9}{20} = 7,4$	$\frac{36,8}{40} = 0,9$

Det er nu højden af rektanglerne og ikke antallet af observationerne, der danner basis for histogrammet.

Tabellerne hørende til krydstabeltest og Goodness of Fit test, χ^2 - test

Opbygningen af tabellerne vises for data hørende til eksempel 4.

Beregning af de forventede værdier (E – tabellen):

Udgangspunktet er tabellen over de observerede værdier (O – tabellen)

O – tabel:

Type\holdning	Interesseret	Ikke interesseret	Ved ikke	I alt
Industri	23	21	6	50
Bygge/anlæg	15	20	1	36
Service/liberalt	51	32	31	114
I alt	89	73	38	200

De forventede værdier i E – tabellen beregnes ved at multiplicere række- og kolonnetotalerne og dele med stikprøvens størrelse:

$$E_{ij} = \frac{R_i}{n} \cdot K_j = \frac{R_i \cdot K_j}{n}$$

Type\holdning	Interesseret	Ikke interesseret	Ved ikke
Industri	$\frac{50 \cdot 89}{200} = 22{,}25$	$\frac{50 \cdot 73}{200} = 18{,}25$	$\frac{50 \cdot 38}{200} = 9{,}50$
Bygge/anlæg	$\frac{36 \cdot 89}{200} = 16{,}02$	$\frac{36 \cdot 73}{200} = 13{,}14$	$\frac{36 \cdot 38}{200} = 6{,}84$
Service/liberalt	$\frac{114 \cdot 89}{200} = 50{,}73$	$\frac{114 \cdot 73}{200} = 41{,}61$	$\frac{114 \cdot 38}{200} = 21{,}66$

Husk slutteligt at tjekke forudsætningen om, at alle indgangene i E – tabellen er over 5.

Beregning af bidragene til teststørrelsen (Q – tabellen):

Teststørrelsen er givet ved: $T = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$.

Bidragene til denne er derfor: $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$. En beregning involverer således O – og E – tabellen.

O – tabellen:

Type\holdning	Interesseret	Ikke interesseret	Ved ikke	I alt
Industri	23	21	6	50
Bygge/anlæg	15	20	1	36
Service/liberalt	51	32	31	114
I alt	89	73	38	200

E – tabellen:

Type\holdning	Interesseret	Ikke interesseret	Ved ikke
Industri	22,25	18,25	9,50
Bygge/anlæg	16,02	13,14	6,84
Service/liberalt	50,73	41,61	21,66

Q – tabellen:

Type\holdning	Interesseret	Ikke interesseret	Ved ikke
Industri	$\frac{(23-22{,}25)^2}{22{,}25} = 0{,}03$	$\frac{(21-18{,}25)^2}{18{,}25} = 0{,}41$	$\frac{(6-9{,}50)^2}{9{,}50} = 1{,}29$
Bygge/anlæg	$\frac{(15-16{,}02)^2}{16{,}02} = 0{,}06$	$\frac{(20-13{,}14)^2}{13{,}14} = 3{,}58$	$\frac{(1-6{,}84)^2}{6{,}84} = 4{,}99$
Service/liberalt	$\frac{(51-50{,}73)^2}{50{,}73} = 0{,}00$	$\frac{(32-41{,}61)^2}{41{,}61} = 2{,}22$	$\frac{(31-21{,}66)^2}{21{,}66} = 4{,}03$