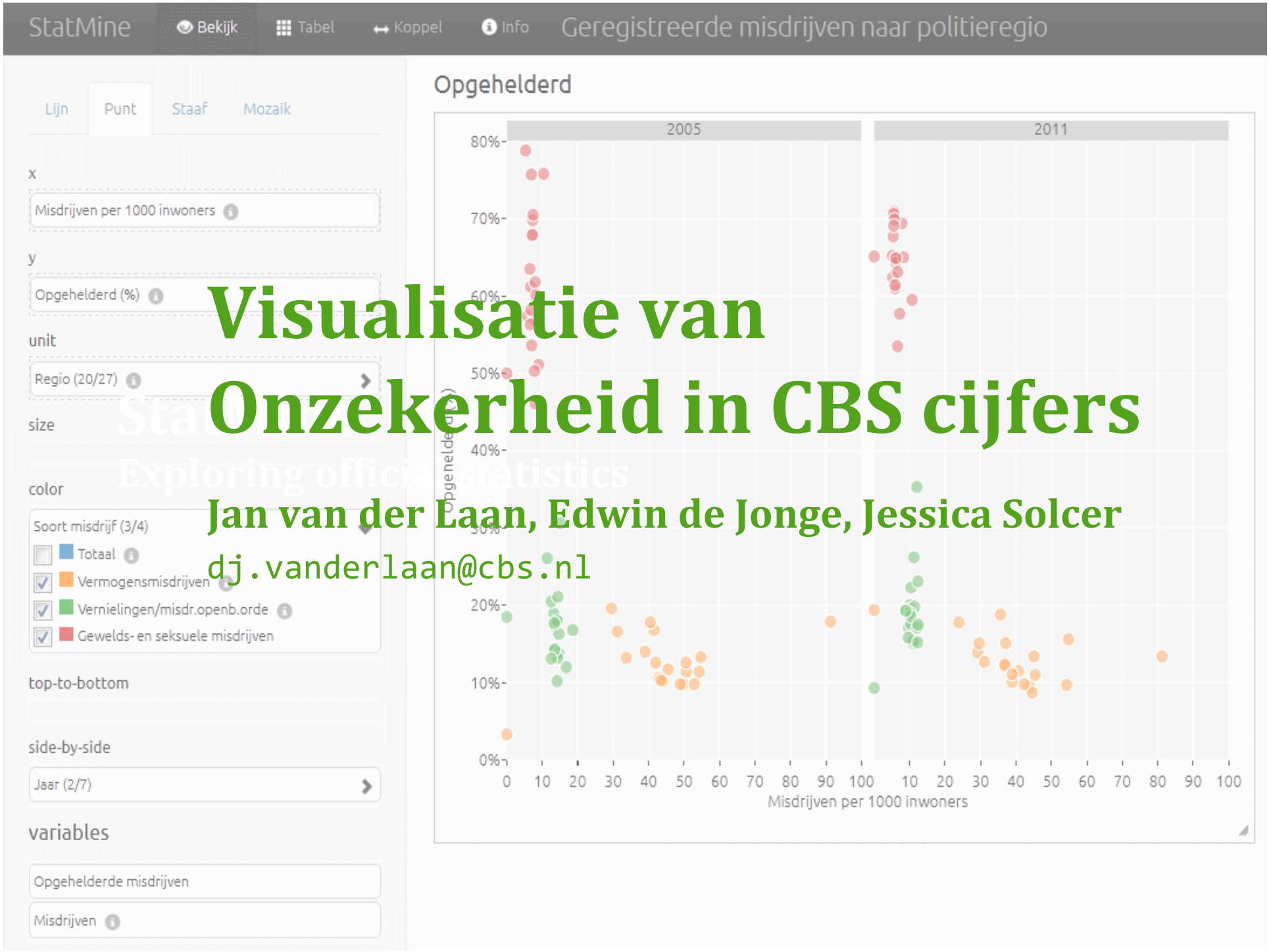


Visualisatie van Onzekerheid in CBS cijfers

Jan van der Laan, Edwin de Jonge, Jessica Solcer

dj.vanderlaan@cbs.nl

Context

Onderzoeksproject ***StatMine***:

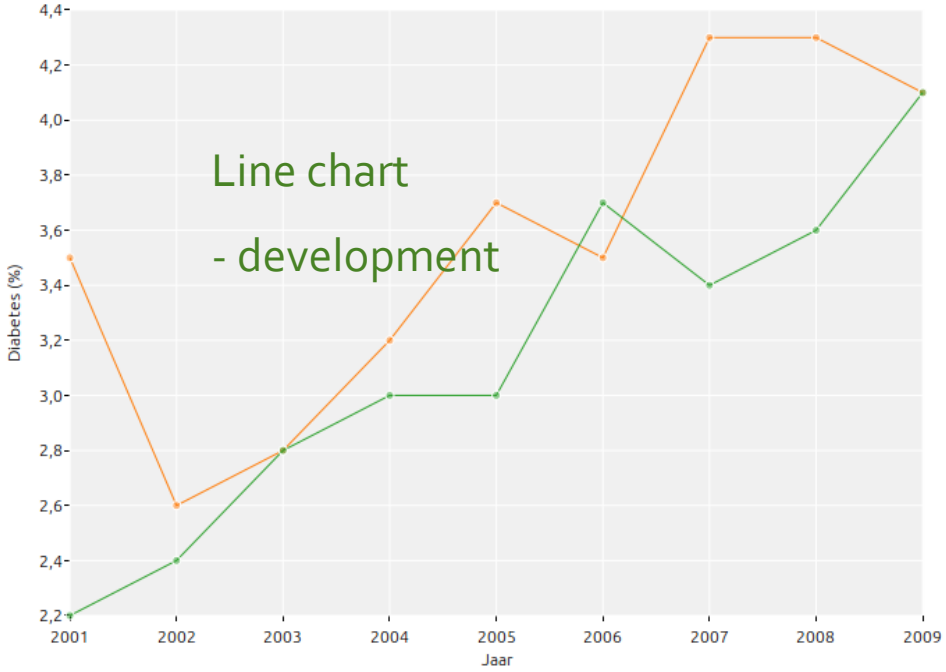
Doel: Toegankelijker maken van online CBS-database *StatLine*

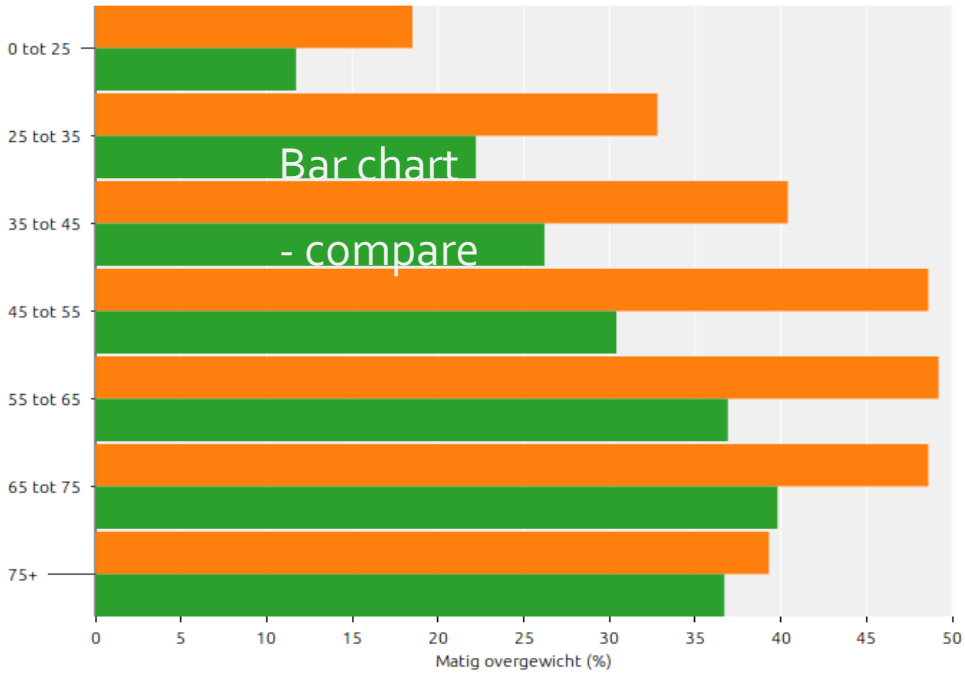
- Visuele analyse/exploratie
- Combineren StatLine tabellen
- Tonen van onzekerheidsmarges

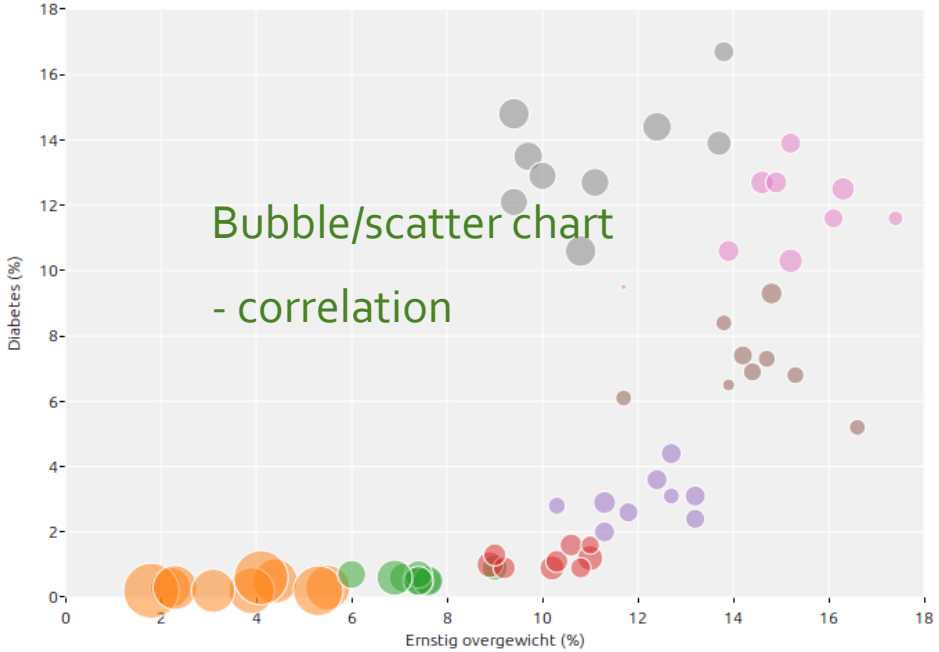
Waarom StatMine?

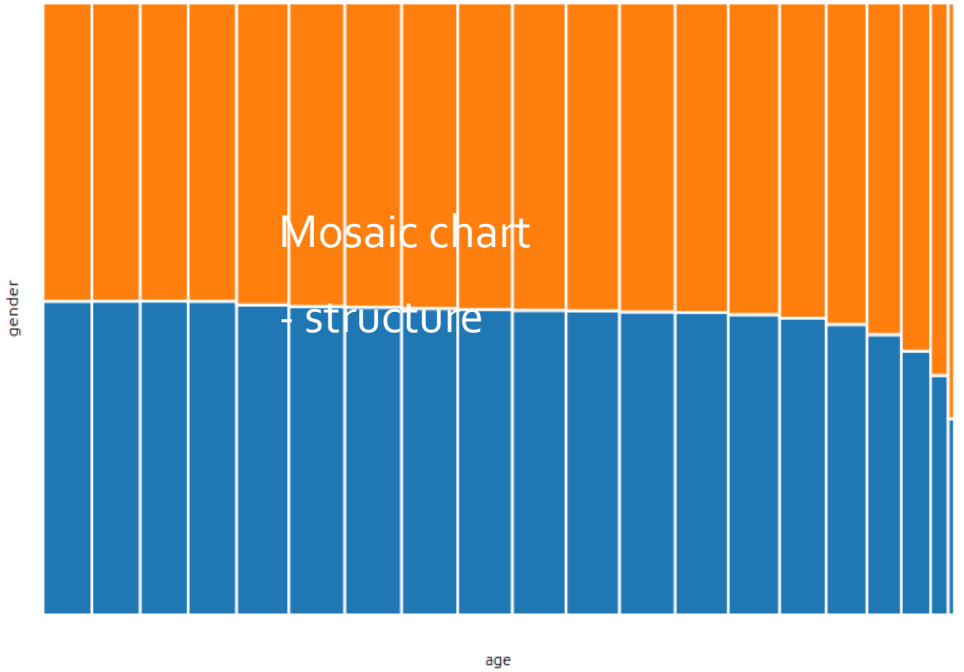
- StatLine bevat meer dan één miljard (10^9) **cijfers**
  - **Beleidsmedewerkers**
  - **Journalisten**
  - Burgers
  - Bedrijven
  - Economen
  - Sociologen etc.

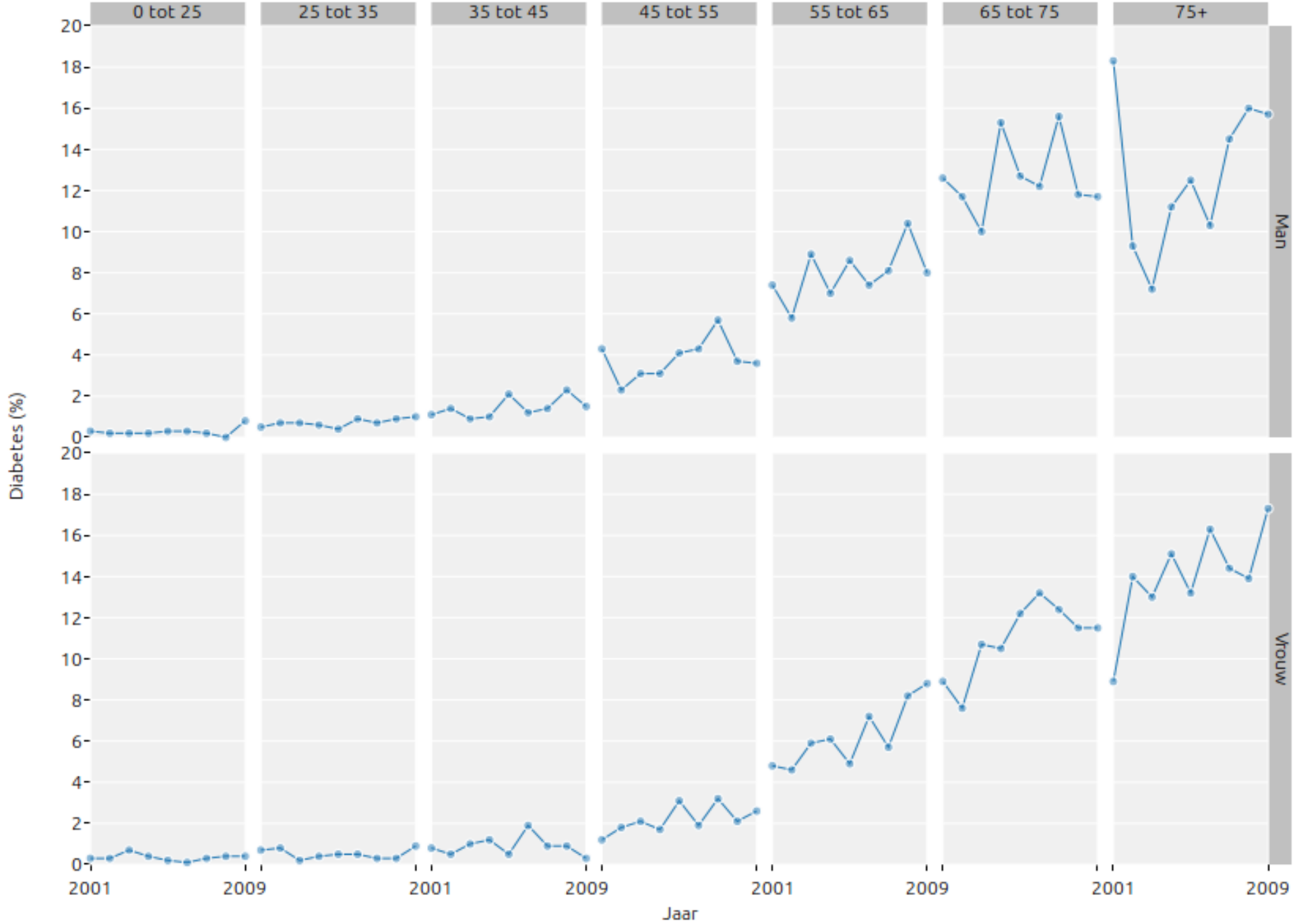

Data
analysis
=
Data
inzicht


Line chart
- development
Diabetes (%)
Jaar
4,4
4,2
4,0
3,8
3,6
3,4
3,2
3,0
2,8
2,6
2,4
2,2
2001
2002
2003
2004
2005
2006
2007
2008
2009


Bar chart
- compare
0 tot 25
25 tot 35
35 tot 45
45 tot 55
55 tot 65
65 tot 75
75+
0
5
10
15
20
25
30
35
40
45
50
Matig overgewicht (%)


Bubble/scatter chart
- correlation
Diabetes (%)
Ernstig overgewicht (%)
18
16
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16
18


Mosaic chart
- structure
gender
age


0 tot 25
25 tot 35
35 tot 45
45 tot 55
55 tot 65
65 tot 75
75+
Man
Vrouw
Diabetes (%)
0
2
4
6
8
10
12
14
16
18
20
2001
2009
Jaar

StatMine visuele analyse

Op meer dan 100 gebruikers getest:

Bevindingen:

- Test personen vinden dat data visualiseren waarde toevoegt (small multiples)
- Data makers kijken anders naar hun data.



Probleem:
Statistische cijfers zijn onzeker

Onderzoeksvraag:

Beperkt tot:

- Wat is het effect op de interpretatie van cijfers?
- Geven gebruikers aan dat ze marges waarderen?
- Kunnen gebruikers omgaan met enkel een intervalschatting (zonder puntschatting)?

Aanname:

- Voor iedere variabele is in ieder geval een intervalschatting beschikbaar (hoe die berekend wordt is buiten scope)

Gebruikerstest

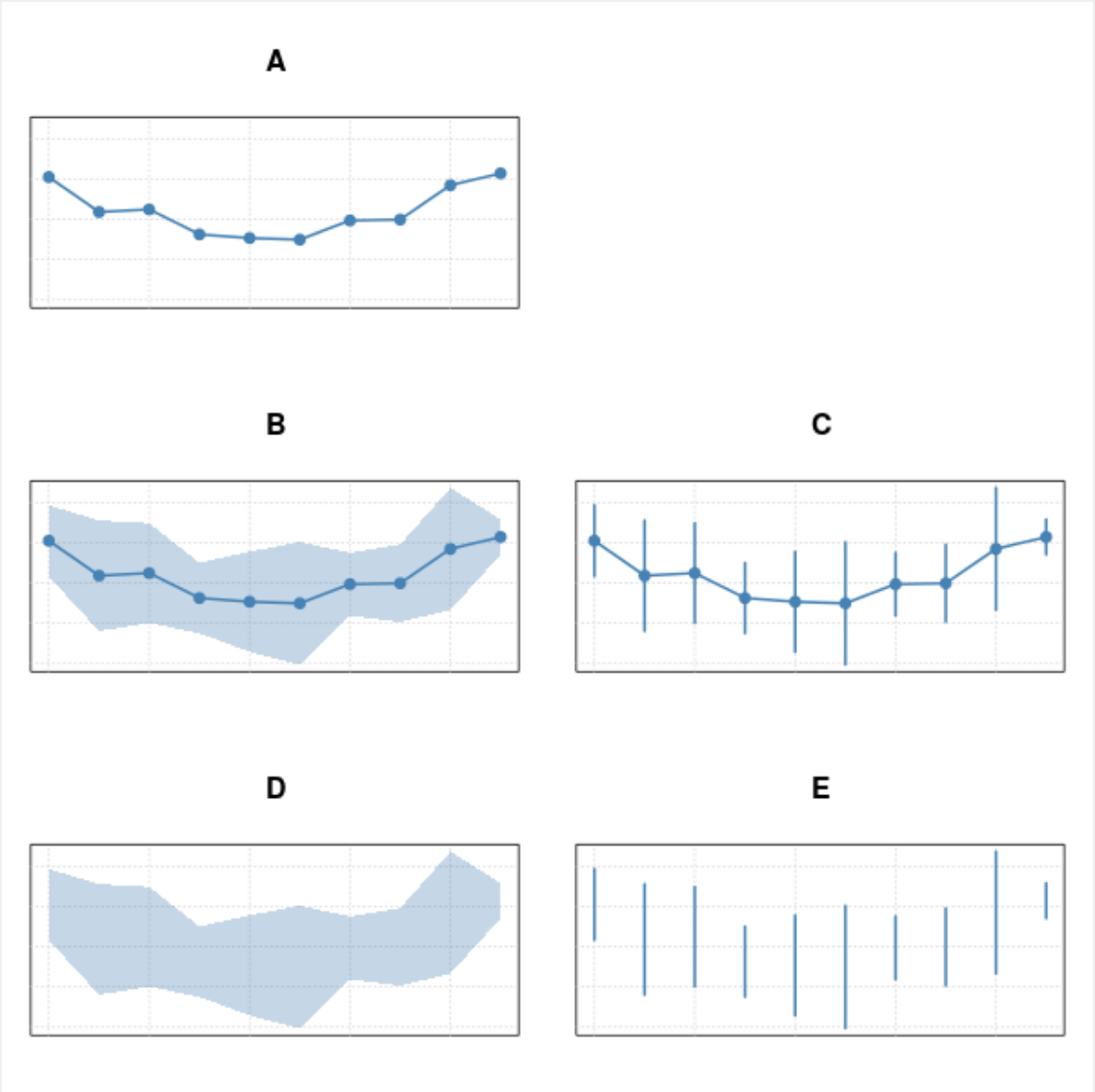
We testen met synthetische data op een groep van 108 testpersonen verschillende vragen met twee grafiektypen:

- Lijngrafiek
- Staafgrafiek

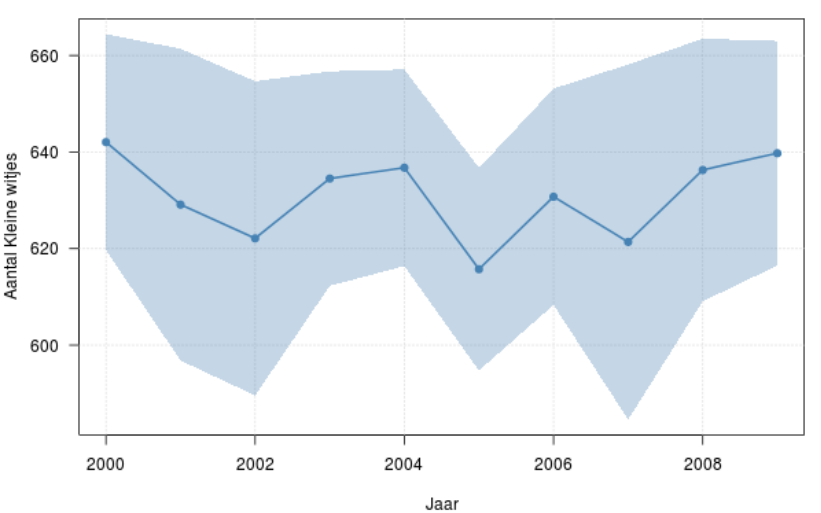
Websurvey:

- Contacten bij overheidsinstanties, data journalisten
- Binnen CBS

Lijngrafiek



Lijngrafiek: vragen



[Y] is afgelopen jaren?

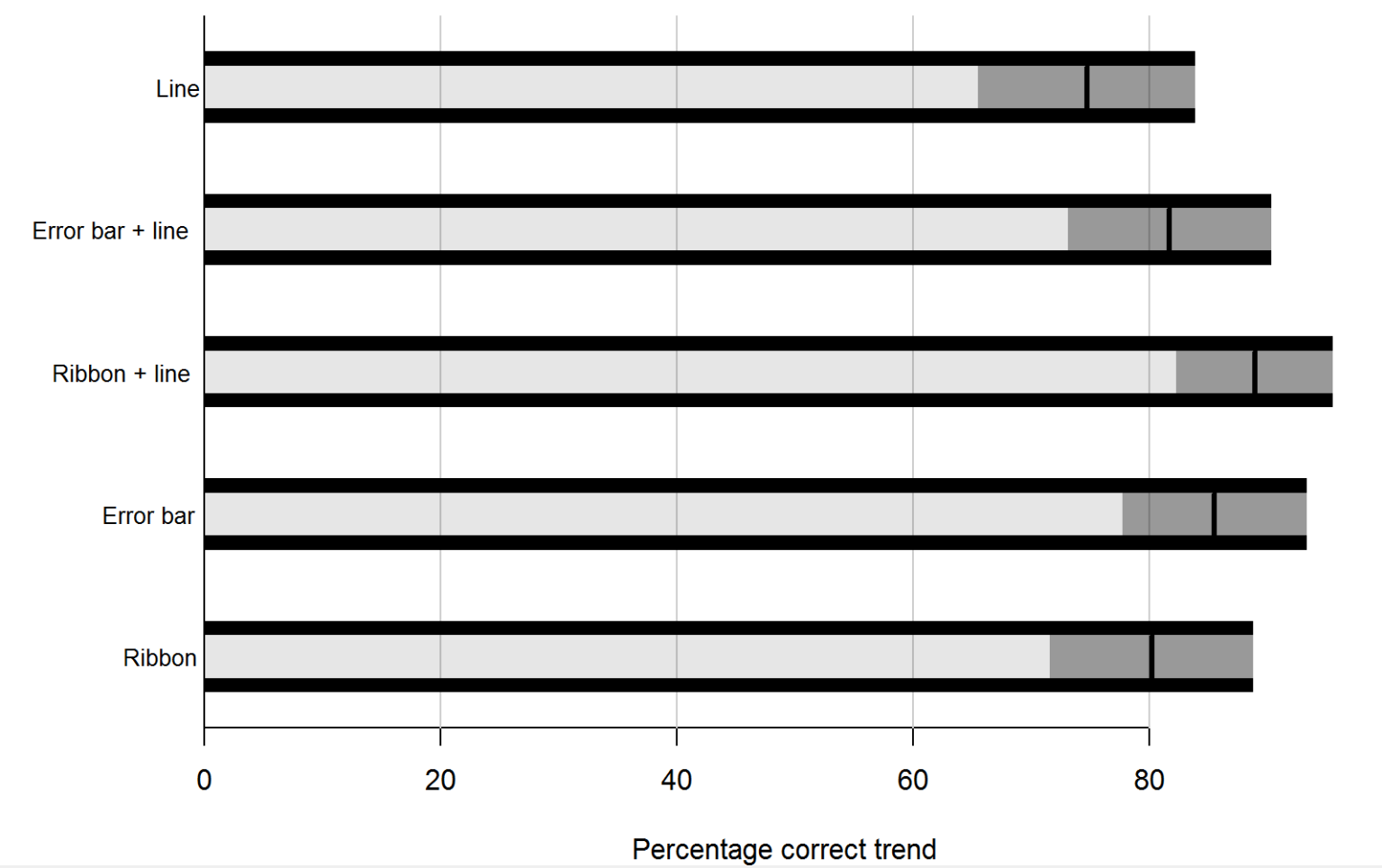
- Zeker gestegen
- Waarschijnlijk gestegen
- Waarschijnlijk gelijk gebleven
- Zeker gelijk gebleven

Is $[Y_t]$ waarschijnlijk hoger dan $[Y_{t-1}]$?

- Yes
- No

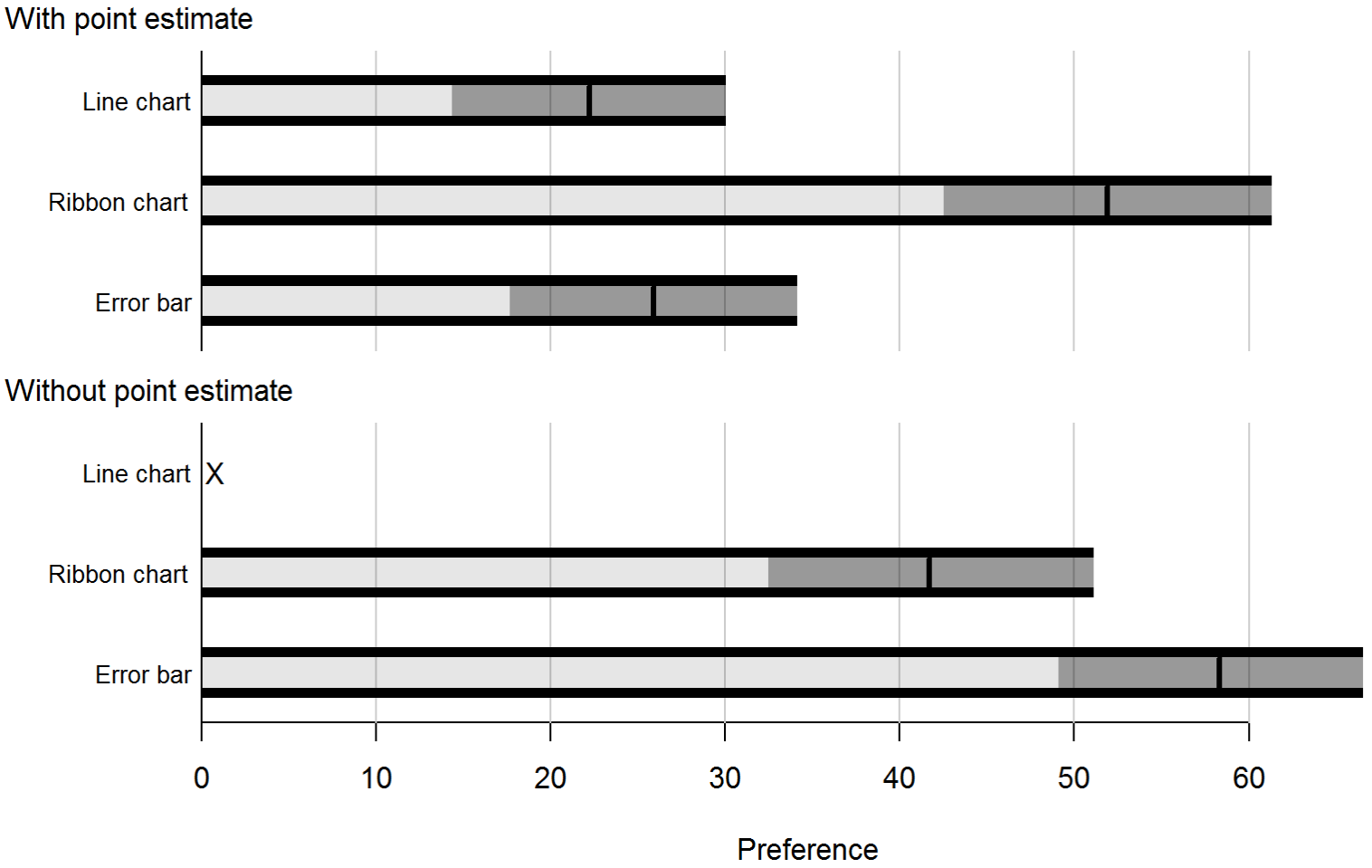
Hoeveel [Y] zijn er in [t]?

Correcte antwoorden uitlezen trend

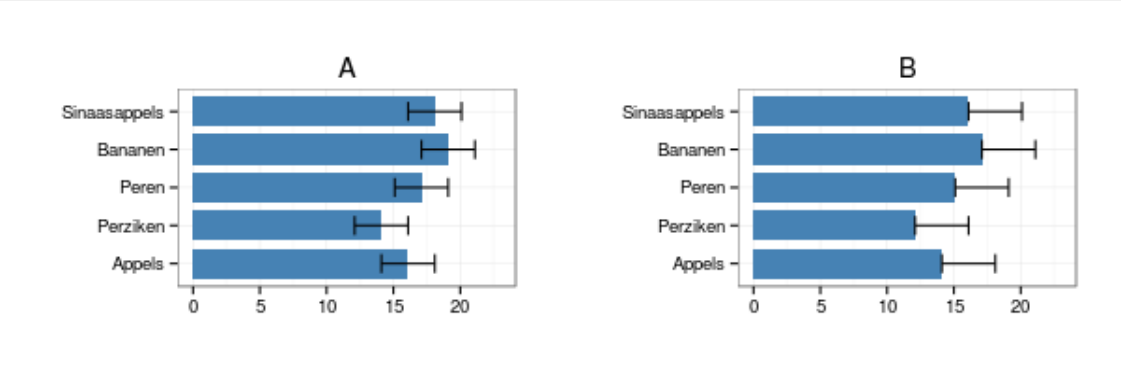


Opm: door het experimentele ontwerp zijn de antwoorden gecorreleerd en geven de betrouwbaarheidsintervallen geen goed beeld van de significantie.

Voorkeuren lijngrafiek



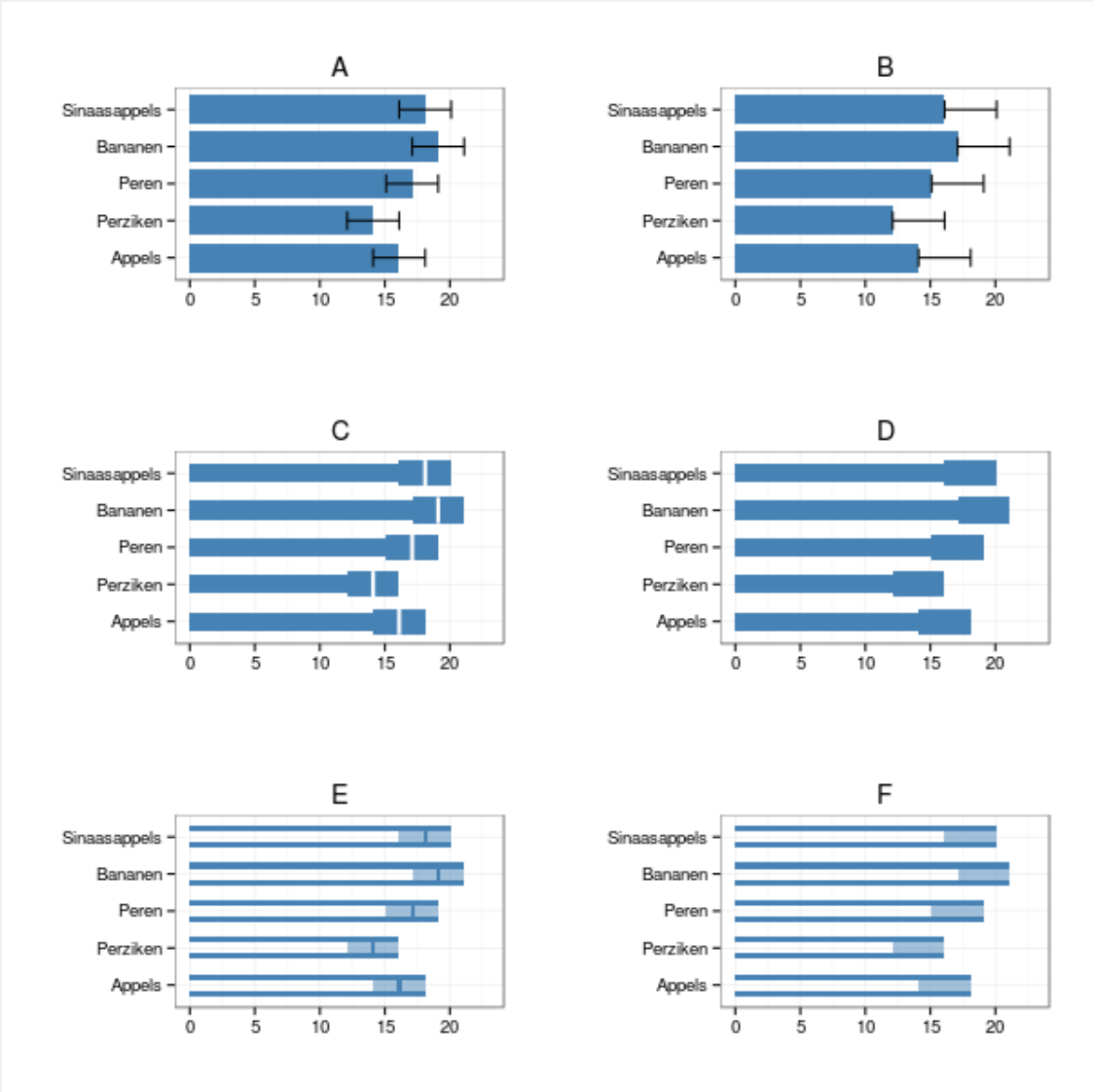
Staafgrafiek



Twee alternatieven

- Gebruik lengte om ***y*** te coderen
- Benadruk interval $[y_{lb}, y_{ub}]$
- Ondersteun plotten $[y_{lb}, y_{ub}]$ zonder y (interval zonder punt)
- y_{lb} en y_{ub} gelijkwaardig zichtbaar

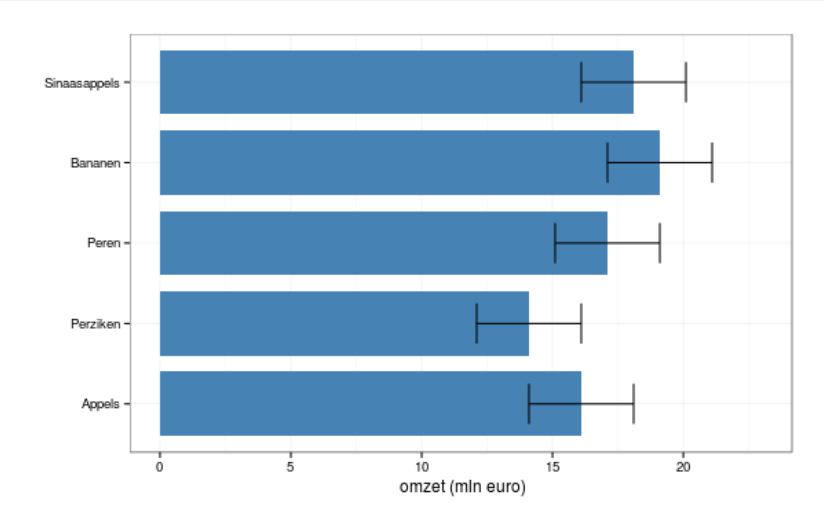
Staafgrafiek



Chisel chart

Cigarette chart

Staafgrafiek: vragen

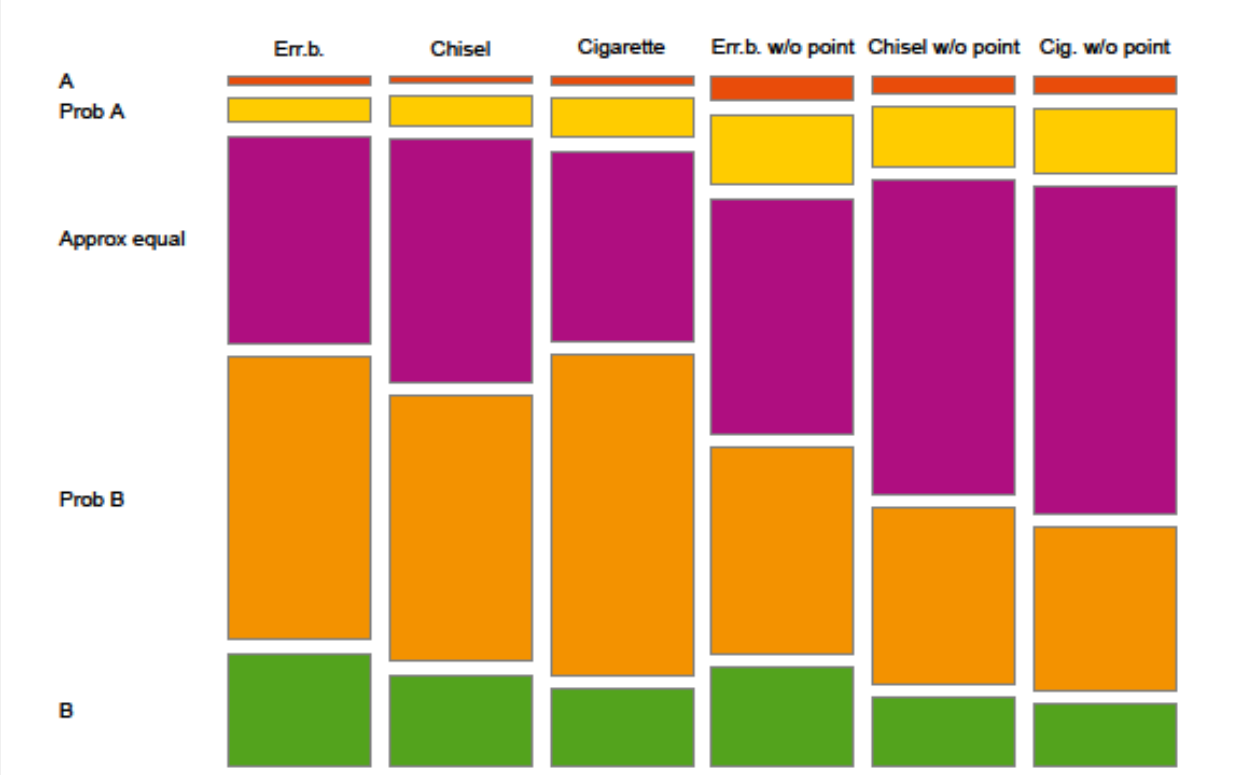


Is de omzet in [A] of [B] hoger?

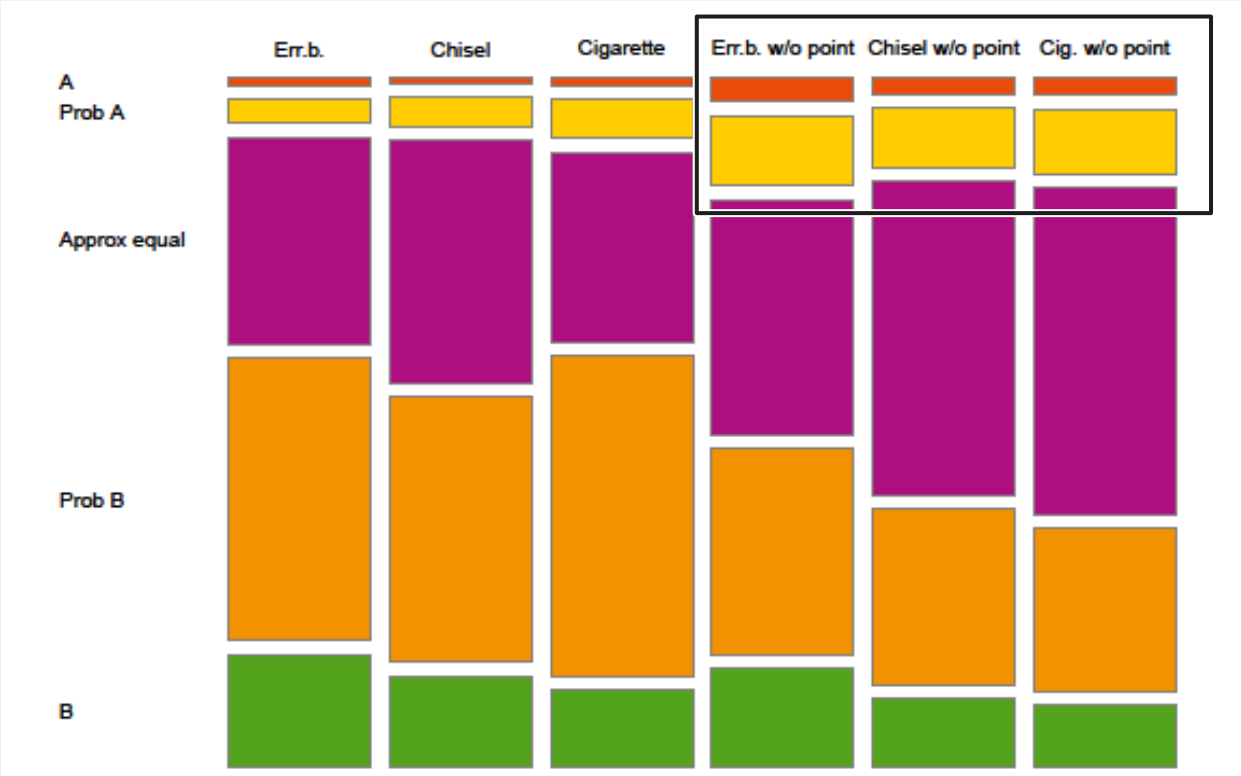
- [A]
- Waarschijnlijk [A]
- Ongeveer evenveel
- Waarschijnlijk[B]
- [B]

Wat is de omzet in [A]?

Verdeling antwoorden staafgrafiek



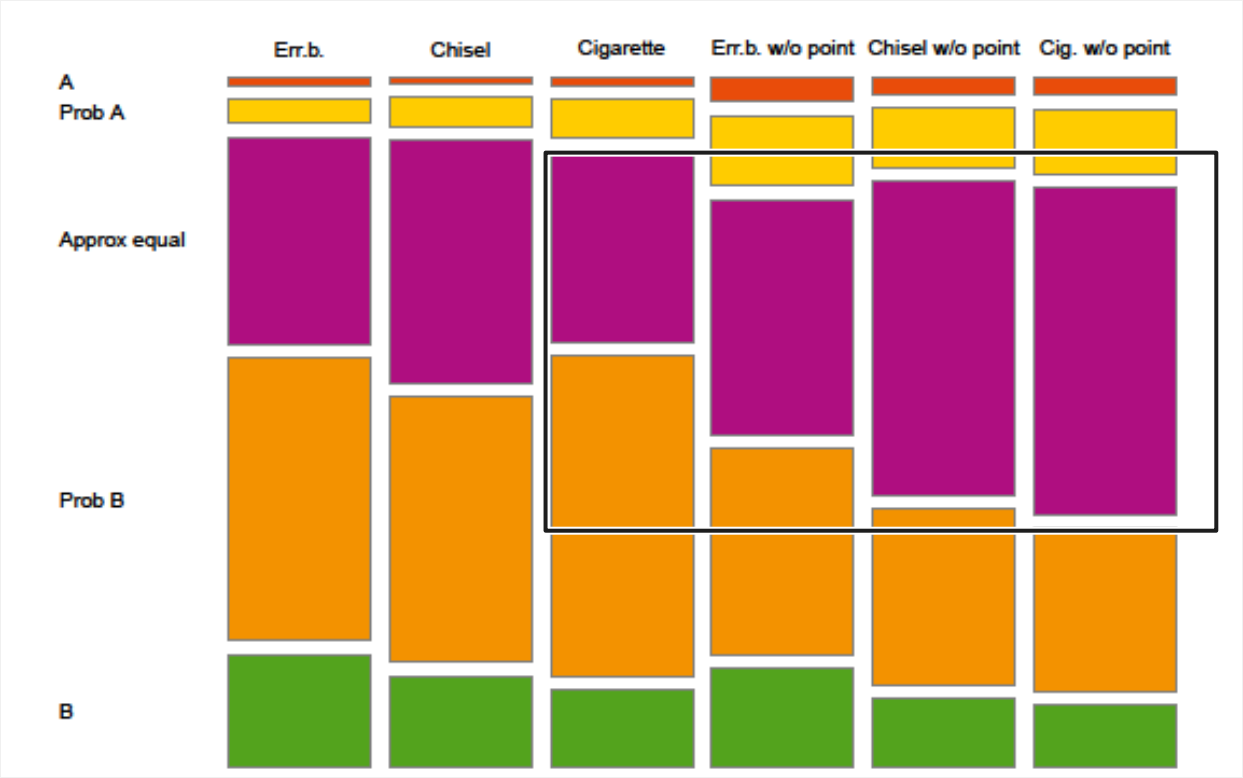
Verdeling antwoorden staafgrafiek



Meer incorrecte antwoorden wanneer

- geen puntschatting
- error bar zonder puntschatting

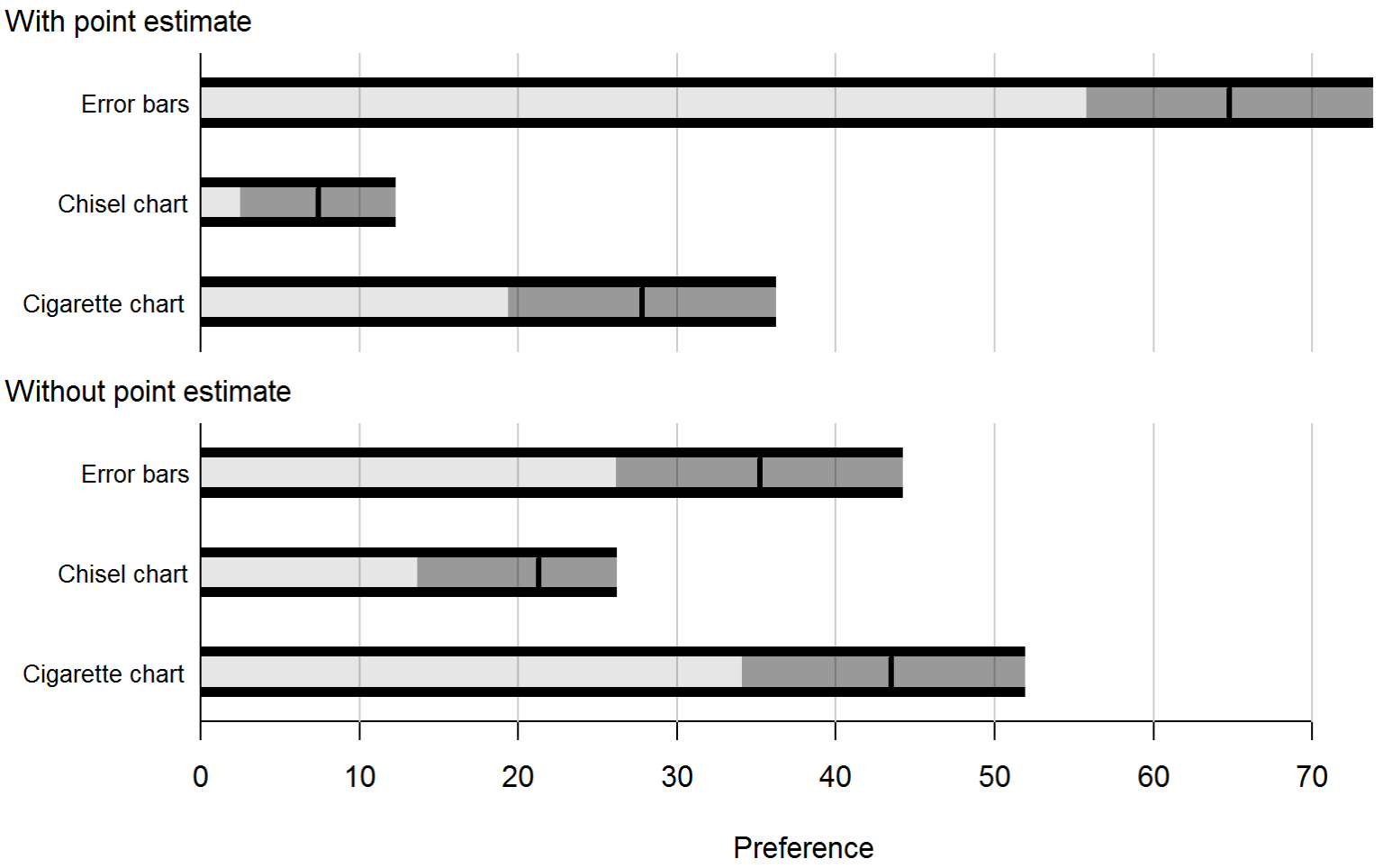
Distribution answers bar chart



‘Ongeveer gelijk’ vaker gekozen wanneer geen puntschatting

Relatief minder vaak voor cigarette plot.

Voorkeuren staafdiagram



Conclusie

Tonen onzekerheid verbetert uitspraken van gebruikers.

Lijngrafiek:
- Met puntschatting: ribbon
- Zonder puntschatting: error bars

Staafgrafiek:
- Met puntschatting: chisel/cigarette
  - Hoewel gebruikers de voorkeur geven aan error bars
- Zonder puntschatting: chisel/cigarette

Gebruikers waarderen weergave onzekerheid en zijn in staat om zulke grafieken correct te interpreteren.