Communication of uncertainty in Official Statistics

Statistics Netherlands (CBS)

Edwin de Jonge

EMOS, webinar, April 7 2020

@edwindjonge/e.dejonge@cbs.nl

Edwin de Jonge

Who am I?

Methodologist/Data Scientist at CBS/Statistics Netherlands:

- Statistical consultancy (o.a. R, shiny, data vis)
- R-training
- Statistical R & D, *Complexity Science (e.g. Network analysis)*

What is statistics?

"***Statistics*** *is a branch of mathematics dealing with data collection, organization,* ***analysis***, ***interpretation***, *and* ***presentation***"

Wikipedia (April 2020)

…statistics is concerned with the use of data in the context of uncertainty and decision making in the face of uncertainty…

Wikipedia, April 2020

What is statistics?

Statistics is the science which uses mathematics to study and improve ways of drawing reliable inferences from incomplete, noisy, corrupt, irreproducible and otherwise imperfect data.

Cosmo Shalizi

Statistics and Uncertainty

Okay, uncertainty is part of making statistics.

- Is that also true for official Statistics?

Official Statistics and Uncertainty

What is not surrounded by uncertainty cannot be the truth,

Richard Feynman

For official statistics, at least two reasons:

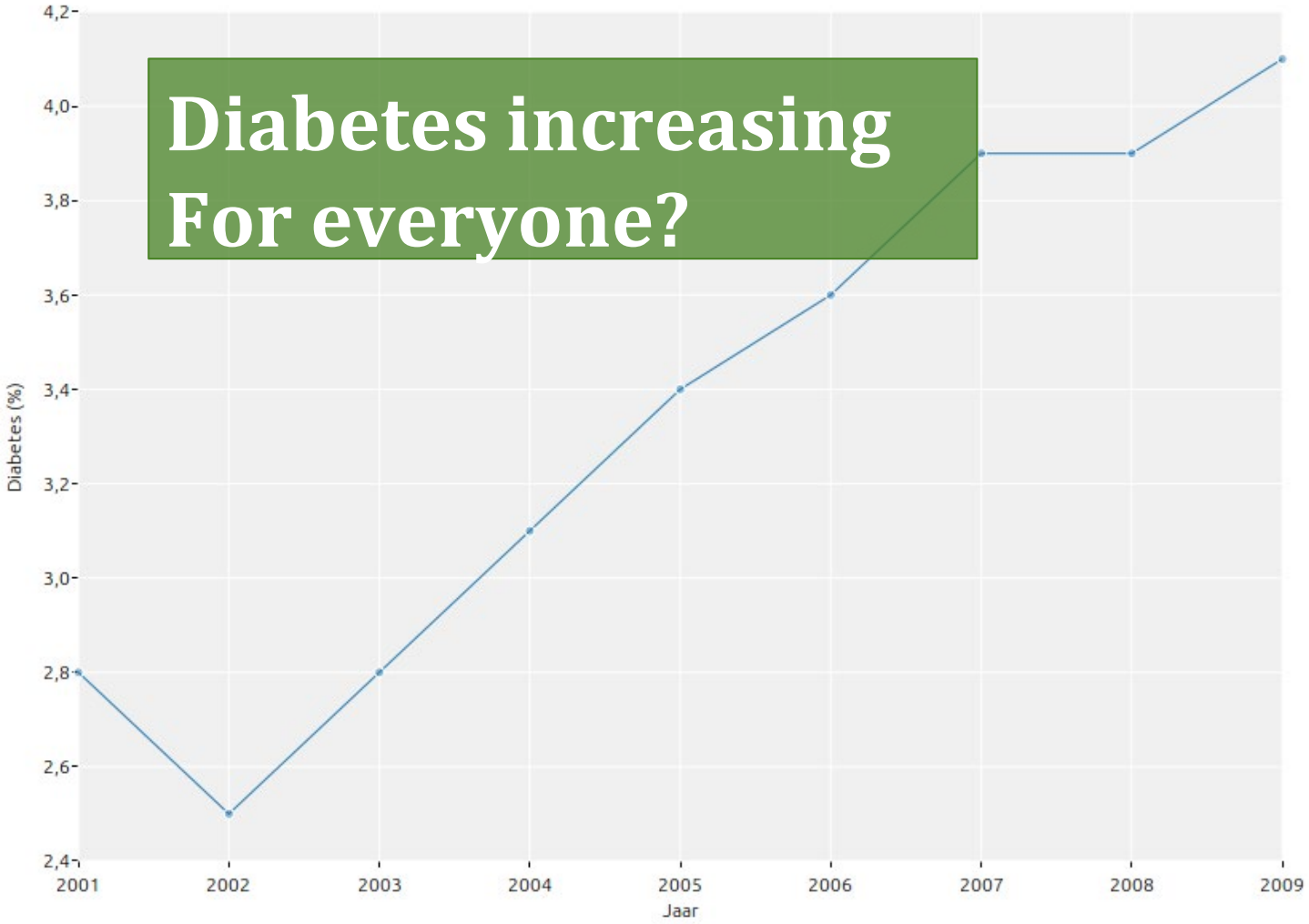
- Communicating accuracy
- Statistical/stochastic uncertainty

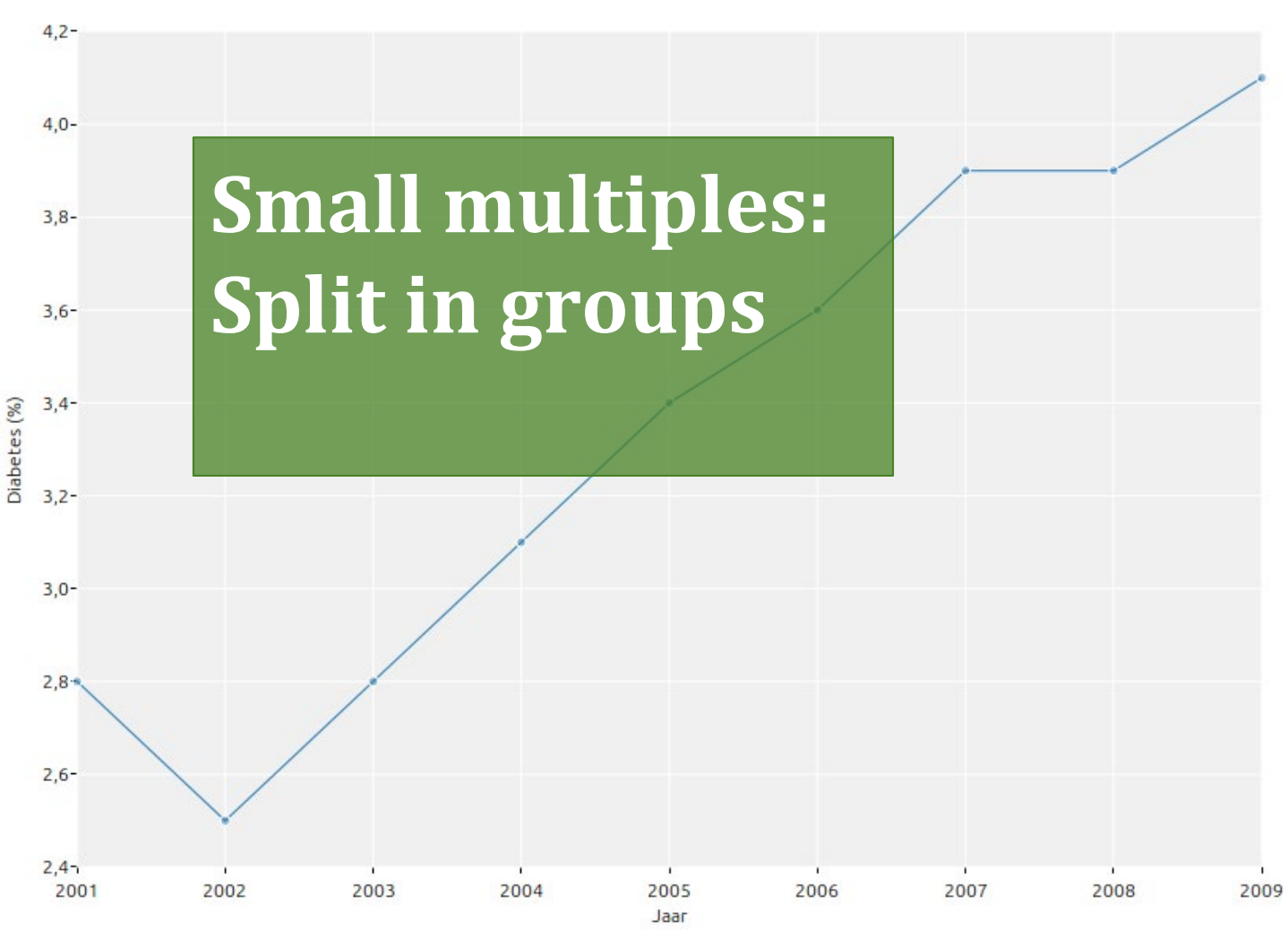
Let's view two cases of stats NL (CBS)

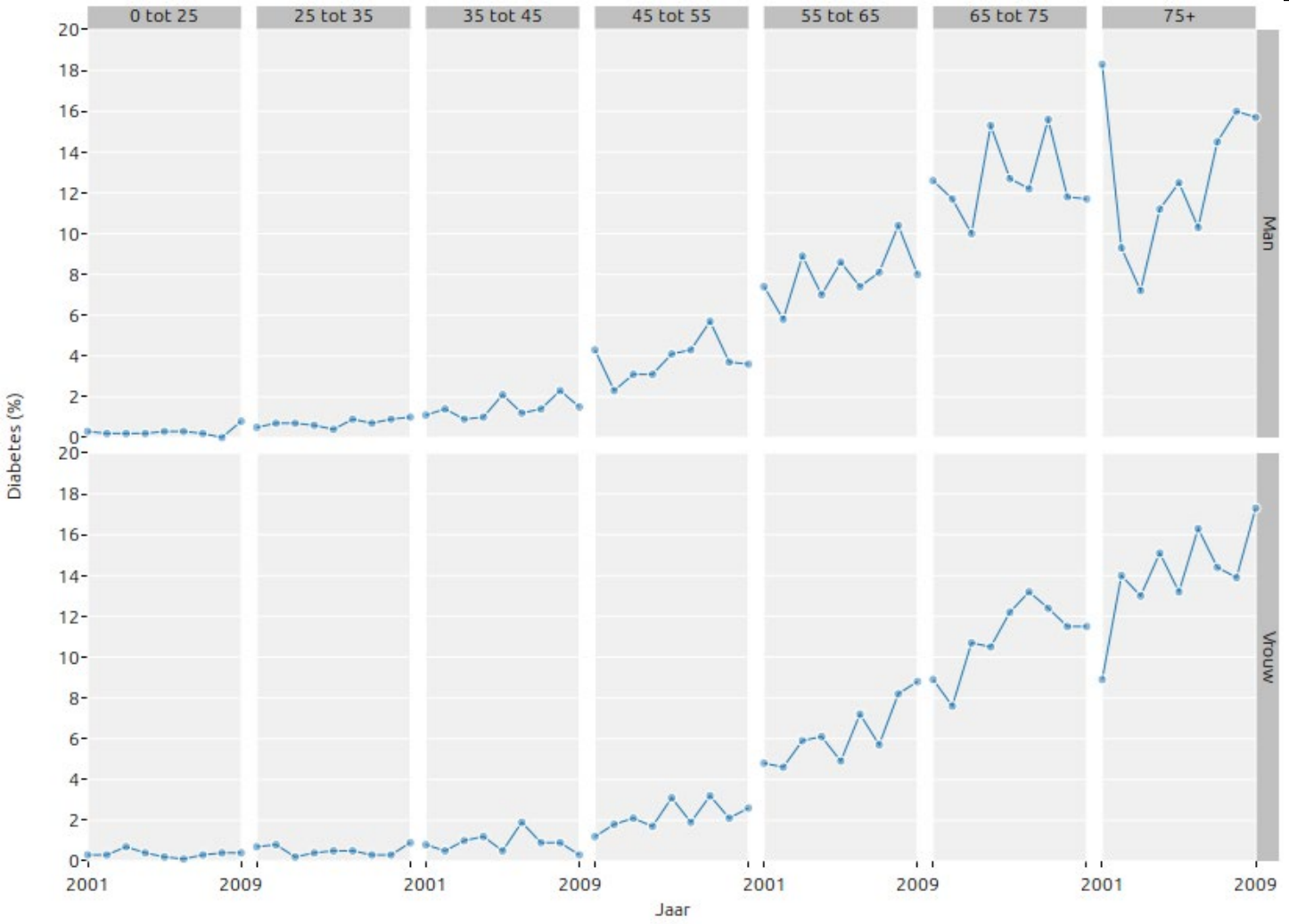
– Diabetes incidence

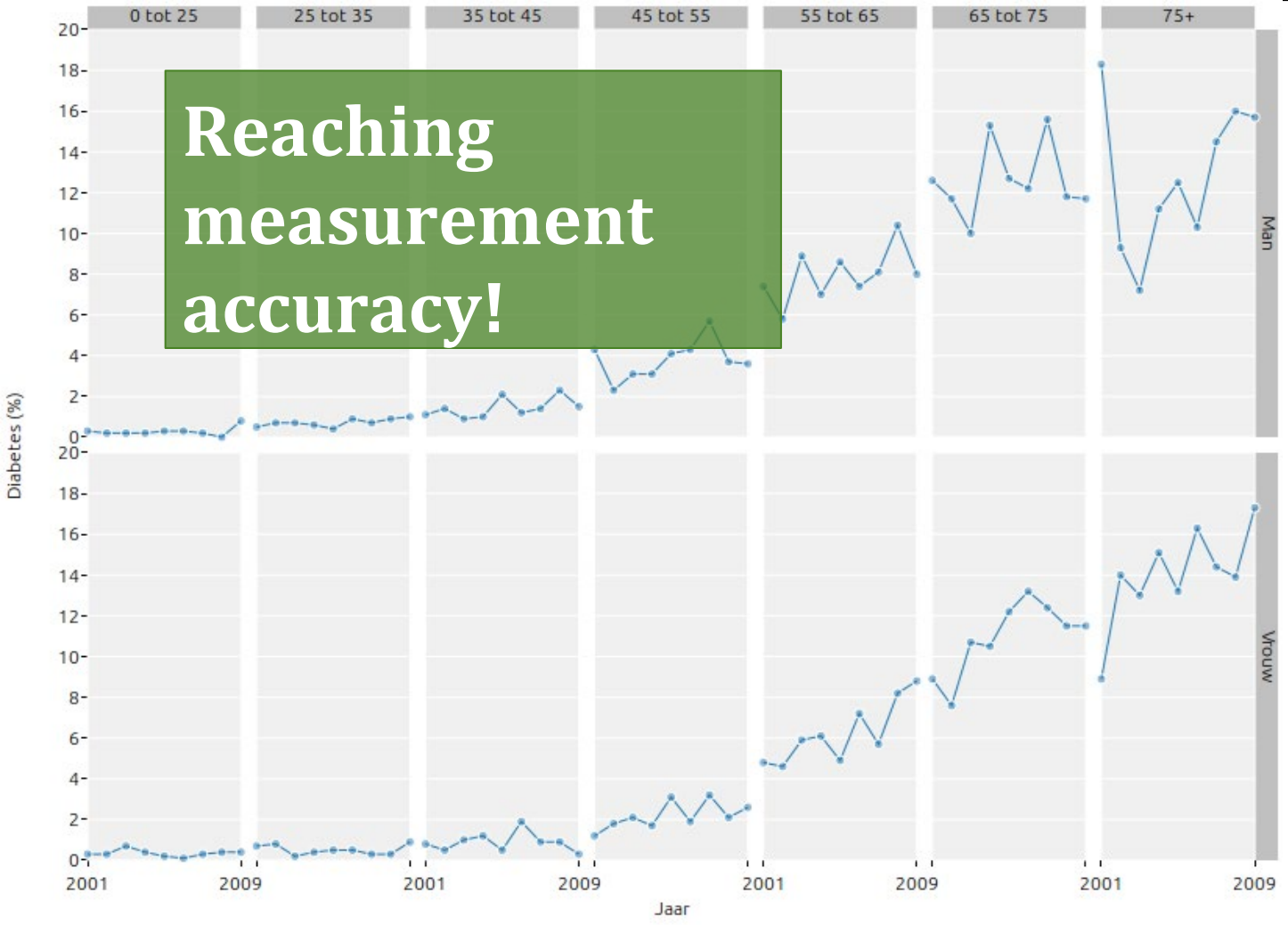
– Based on a (large) health survey of statistics netherlands (CBS)

Diabetes increasing For everyone?



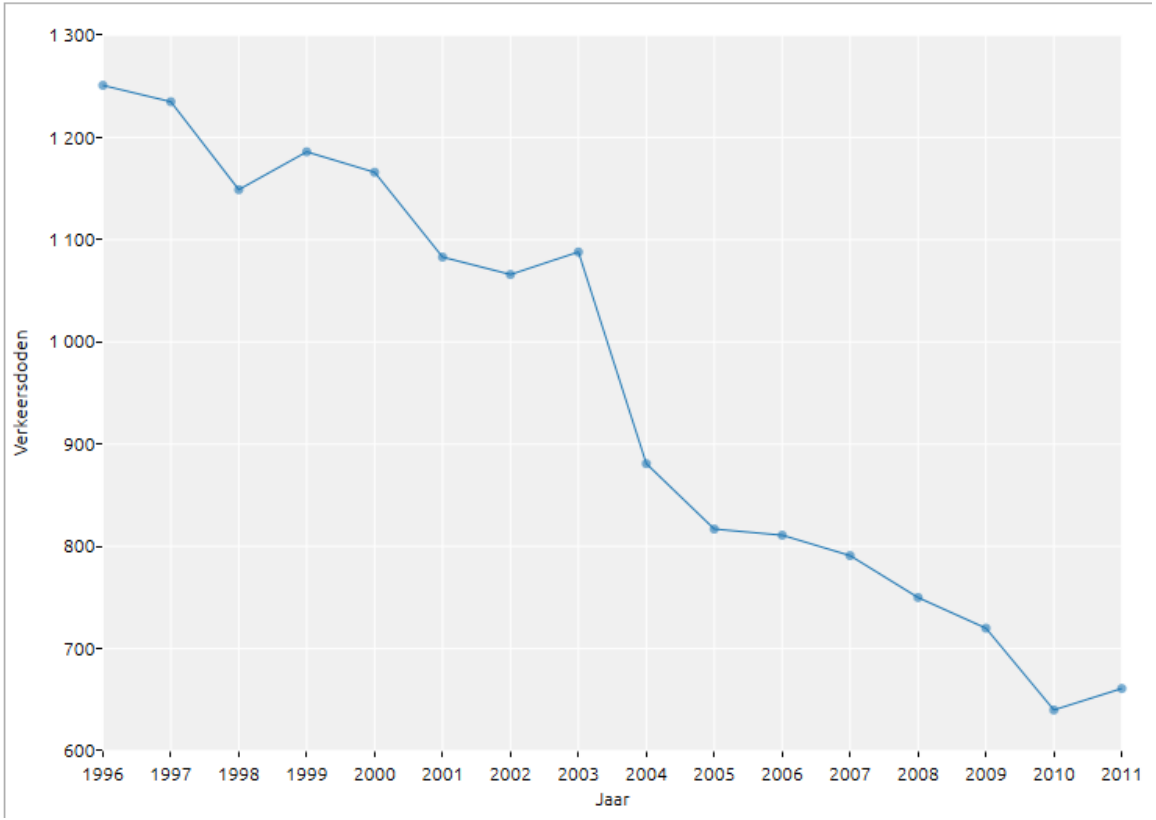

Small multiples:
Split in groups
Diabetes (%)
4,2
4,0
3,8
3,6
3,4
3,2
3,0
2,8
2,6
2,4
2001
2002
2003
2004
2005
2006
2007
2008
2009
Jaar


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

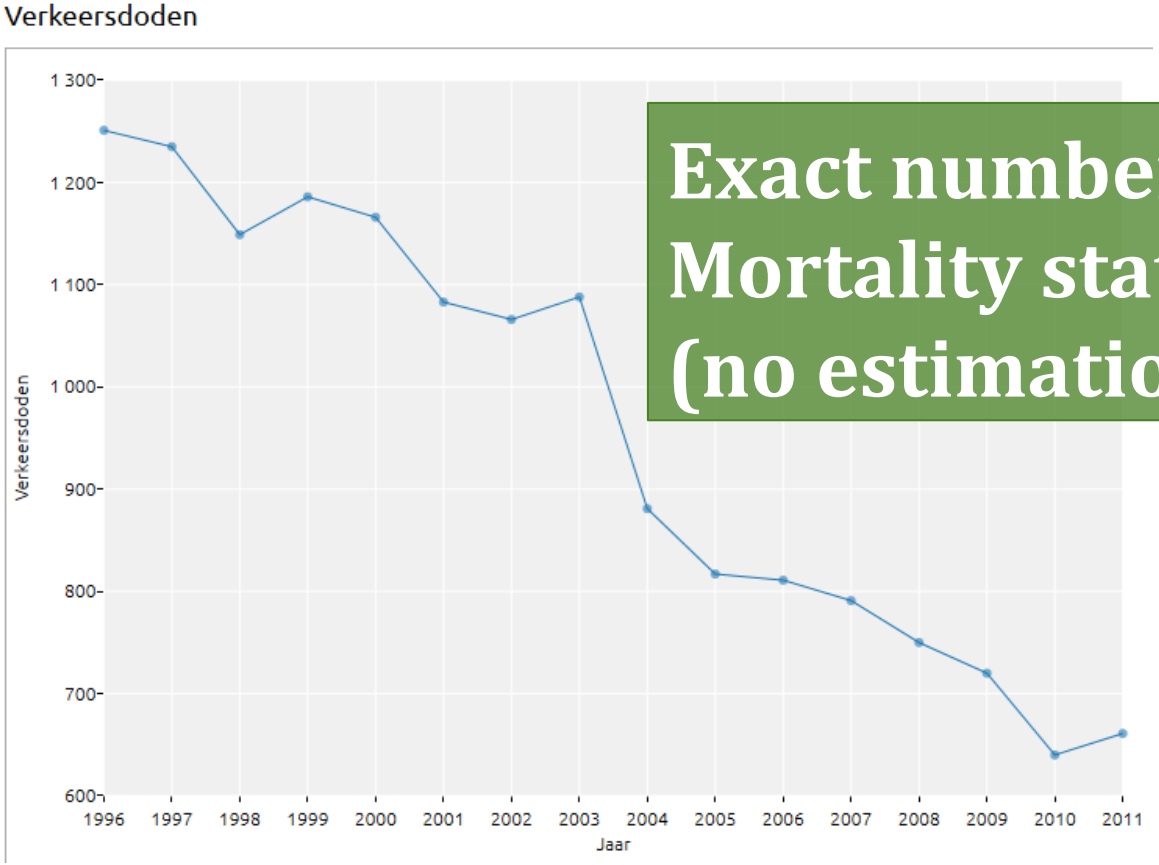

Reaching measurement accuracy!
0 tot 25
25 tot 35
35 tot 45
45 tot 55
55 tot 65
65 tot 75
75+
Man
Vrouw
Diabetes (%)
Jaar
2001
2009

Traffic fatalities

Verkeersdoden

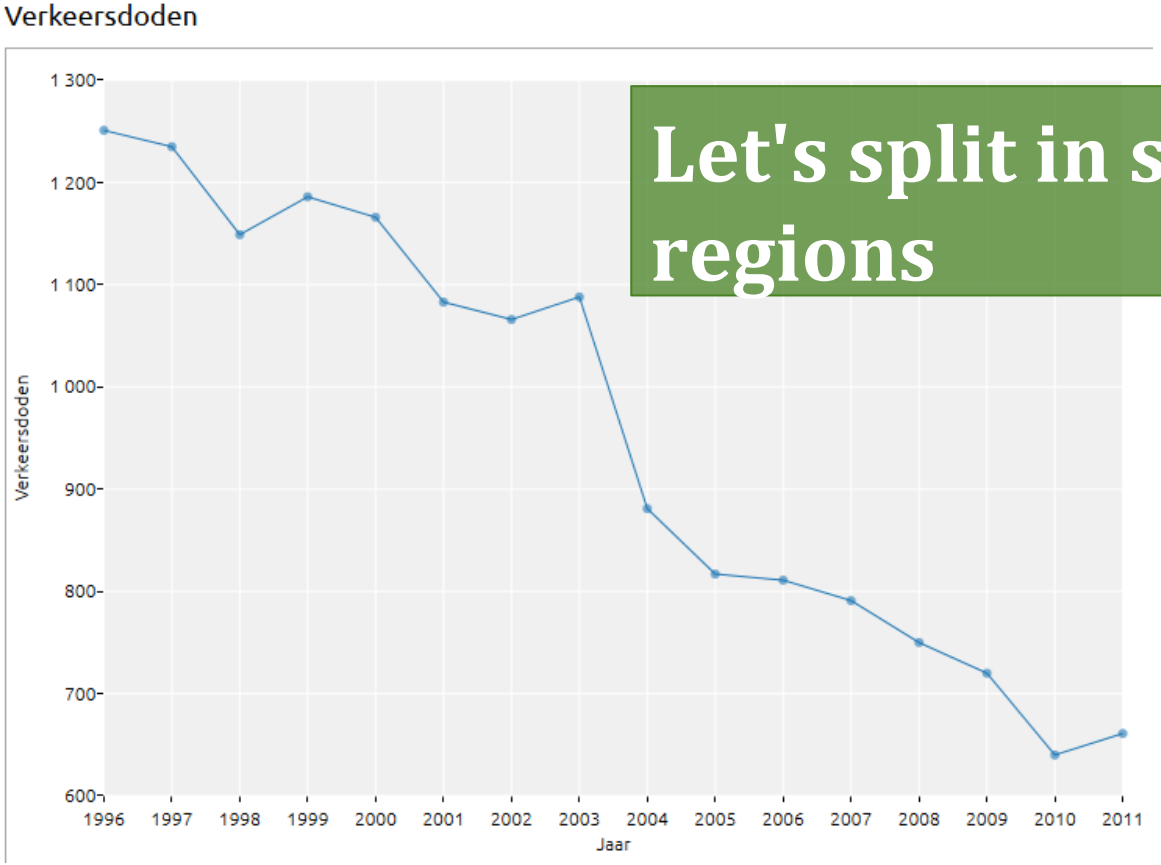


Traffic fatalities



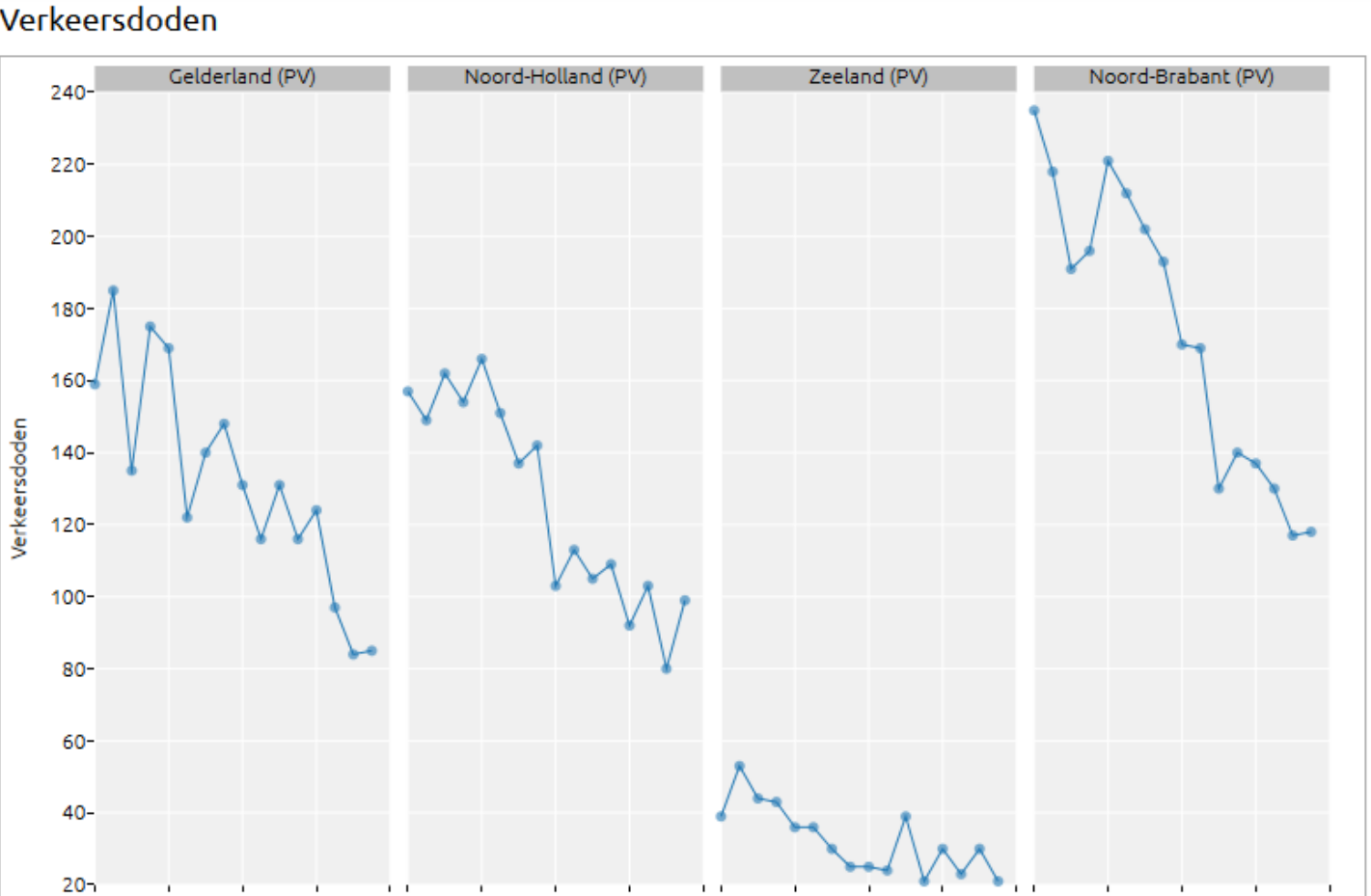
Exact numbers!
Mortality stats
(no estimation)

Traffic fatalities

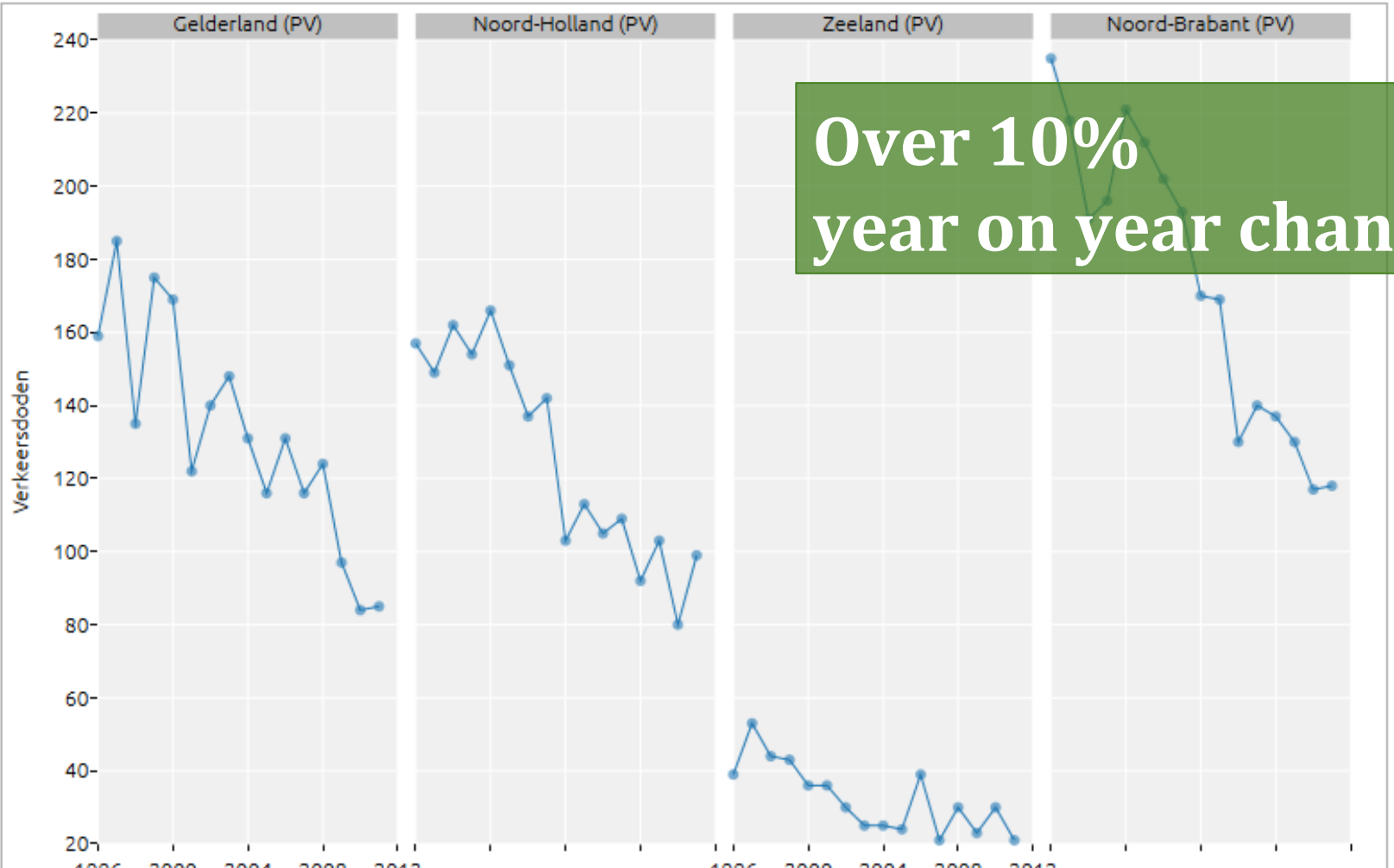


Let's split in smaller regions

Traffic fatalities



Verkeersdoden



Over 10%
year on year changes!

Case 2: Stochastic uncertainty

Uncertainty types

- Estimation uncertainty:
  - There is a **true** value, but lack of knowledge / measurement error / estimation error / processing error.
  - Official Stats: different versions of numbers will have improved accuracy / precision
  - (preliminary / final / revised versions).
- Prediction uncertainty:
  - No true value (yet!)

Intermezzo

Questions?

Uncertainty should be communicated

Communcation of uncertainty is of importance!

- Policy makers
- Decision makers
- Economists
- Scientists

All depend on the official numbers!

Official Stats Communcation Practices

Official statistics institutes are:

– very careful / prudent / reluctant on publishing uncertainty margins...

Why so prudent?

Possible reasons:

– “Users don't understand them”

– “Users dont need them”

– “Users may choose the number that best fits them”

– “We don't have an accurate estimation of the accuracy”.

– “Users might lose confidence in statistical institute”

Communication of Uncertainty

- Verbal
- Numerical
- Visual

Verbal Communication of Uncertainty

- Not common in official statistics
- Most verbal “terms” are on probability / prediction
- So no directly useful for official statitics.

Uncertainty terms (IPCC)

Virtually certain	> 99%
Very likely	90-99%
Likely	66%-90%
About as likely as not	33%-66%
Unlikely	10%-33%
Very unlikely	1%-10%
Exceptionally unlikely	<1%

Numerical Communication

Most often used in official stats.

Typically:

- standard error
- mostly sampling error (other measurement errors?)
- 95% confidence interval
- expressed in percentage of statistic...

Good practice Numerical communication

Express uncertainty as interval:

- Economic growth is 0.5% [0.3%-0.7%].

This is for most uses beter then:

- Economic growth is 0.5% +- 0.2%
- Economic growth is 0.5%

Uncertainty measures

- Frequentist confidence interval
- Bayesian Credible interval
- Prediction Interval

For statisticians really important which type!

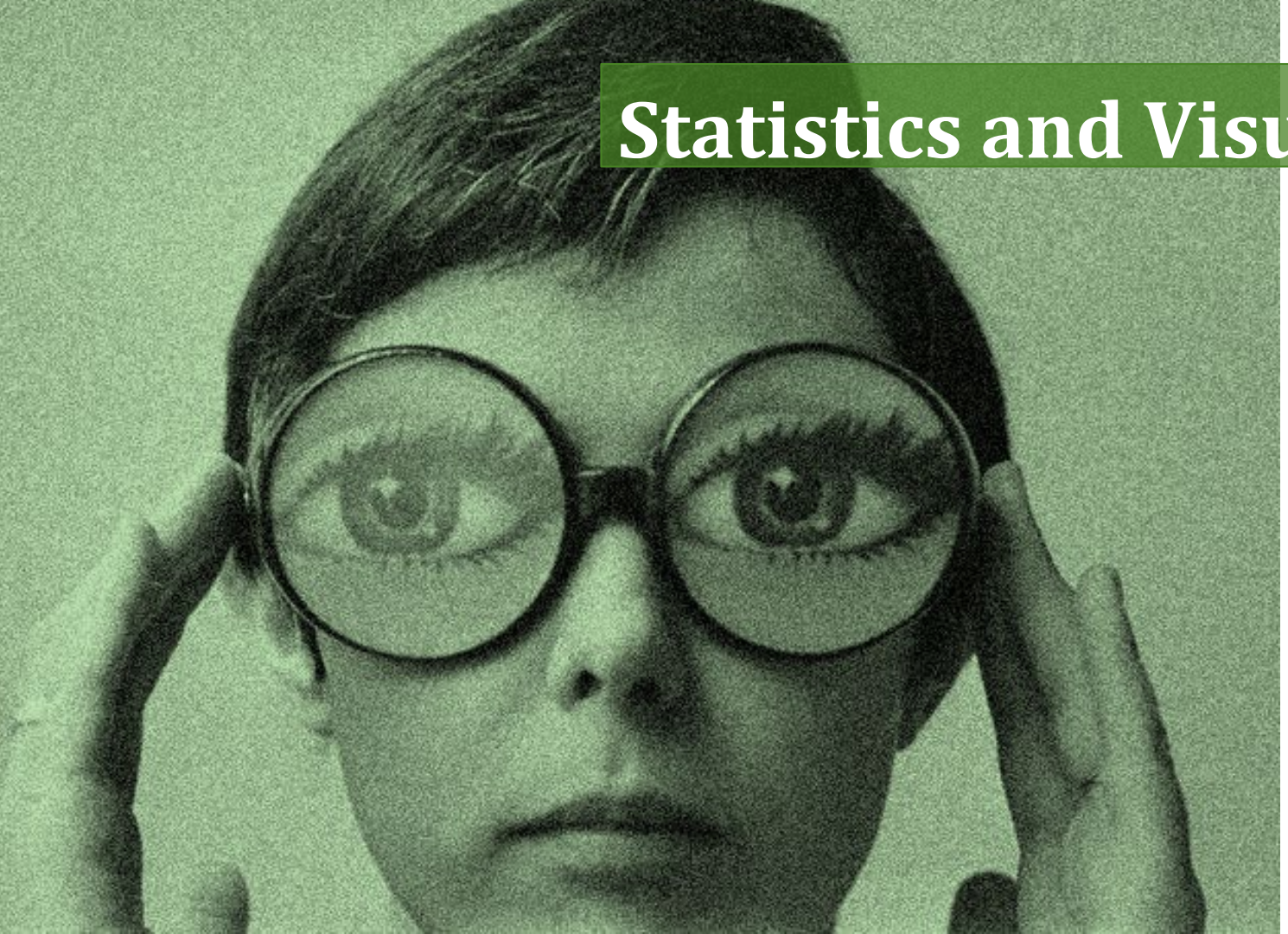
However for communication purposes in general not:

- statistical detail: the measure expresses confidence of institute in figure.

Intermezzo

Questions?

Statistics and Visualisation?



Visual communication

Data visualisation:

- important communication channel
- can summarize large quantity of information
- may encode uncertainty!

Why so prudent?

Possible reasons:

– **“Users don't understand them”**

– **“Users dont need them”**

– “Users may choose the number that best fits them”

– “We don't have an accurate estimation of the accuracy”.

– ?

- “Users might lose confidence in statistical institute”

User study 1:

- **The perception of visual uncertainty representation by non-experts**

  Tak, Toet, van Erp, *Transactions on Visualisation and computer Graphics*, 2014

- User experiment:

  140 users

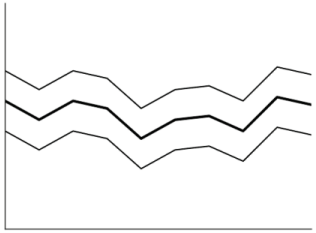
- Tests:

  Reading of certainty.

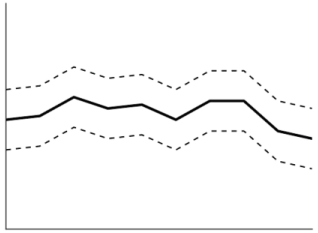
  Given a number, how certain is that value?


rtl XL
25
20
Normaal
15
10
5
DI
24.04
WO
DO
VR
ZA
ZO
29.04
MA
DI
WO

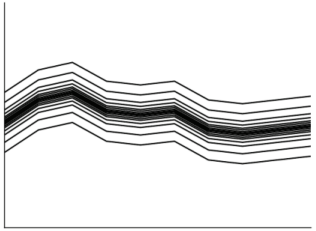
Chart types:



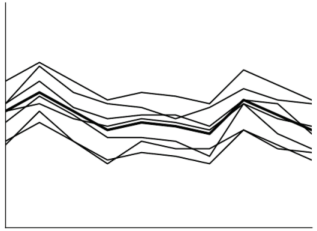
(a) solid border



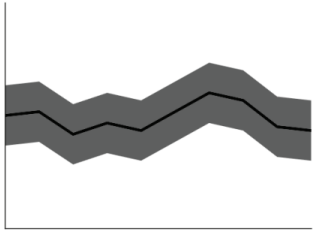
(b) dashed border



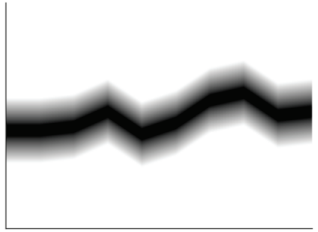
(e) thinning lines



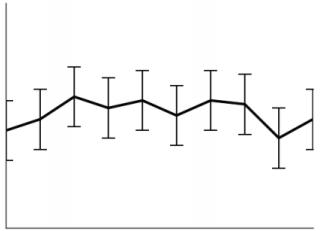
(f) random lines



(c) band



(d) gradient



(g) error bars

Results

- Non-expert can read probability intervals.
- However: users with high numeracy are better at it.
- No (significant) difference in response time.
- Random lines work well for stochastic numbers.

User study 2:

- **Effect of displaying uncertainty in Line and Bar charts**, Van der Laan, **de Jonge**, Solcer, IVAPP, 2015

  User study:

  110 persons

  Goal:

  Line: how does uncertainty effect (overall) trend? (main purpose line chart)

  Bar: how does uncertainty effect comparison? (main purpose bar chart)

Restricted to:

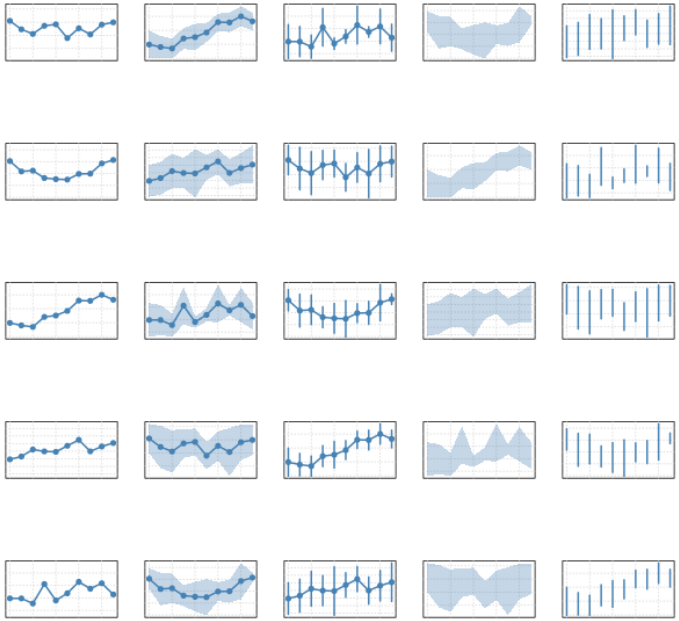
- How do users interpret CI's?
- And what does that affect the interpretation of facts?
- Do users need CI's?

Assumption:

- For test data set of point estimate with CI available

User test (100+) with synthetic data shows that:

- CI's improve validity of user statements (they are more correct)



Line Charts

– displaying uncertainty improve user statements (more correct)
– "*band* + *line*" works best for point estimate
, "*error bar*" works best for interval estimate

– Users do not "freak out" on uncertainty
– Appreciate it and ask for its definition.

StatMine | View | Table | Link | Info | Settings | Drugsgebruik

Line | Scatter | Bar | Mosaic

y
Onderwijsniveau (4/5)

size
Drugsgebruik (%)

color

top-to-bottom

side-by-side

variables

Jaar **['2009']**

Onderwerp **['Heeft ooit drugs gebruikt']**

Geslacht **['Totaal']**

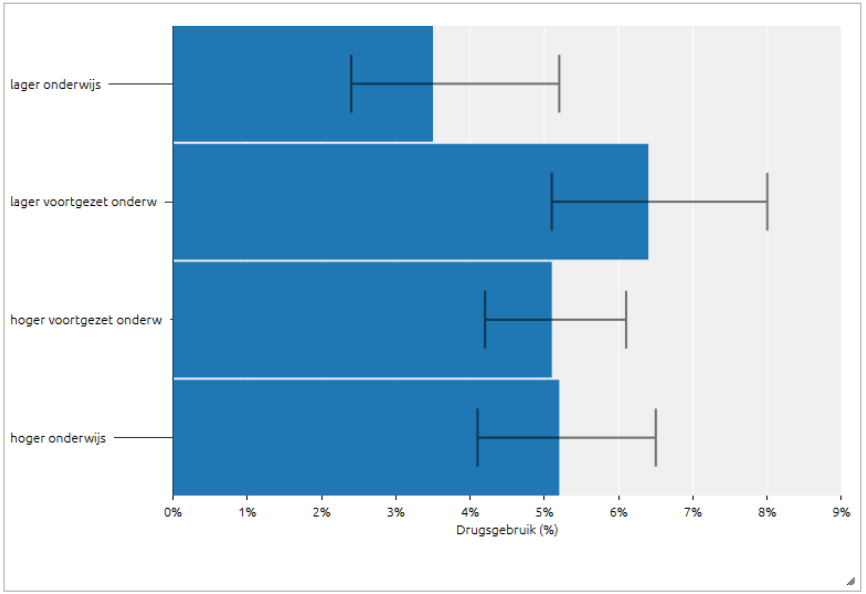
Stedelijkheid **['Totaal']**

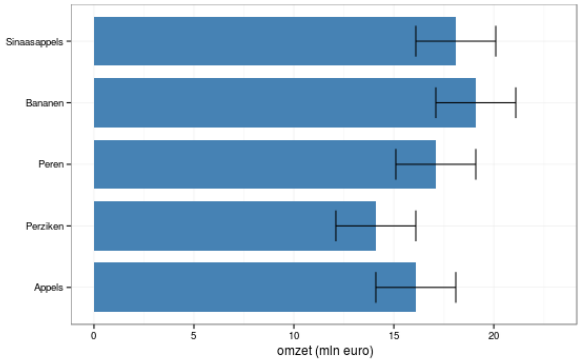
Leeftijd **['Totaal']**

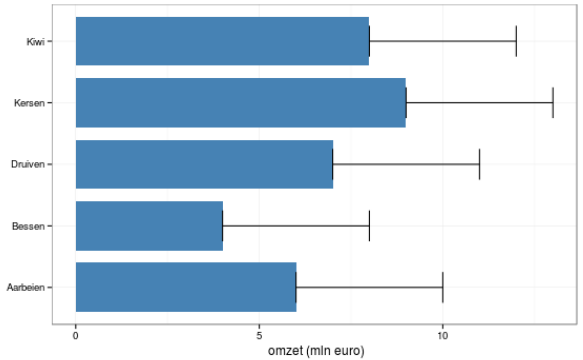
Soort **['Cocaine']**

- [ ] Amfetaminen
- [x] Cocaine
- [ ] Ecstasy
- [ ] Hasj of marihuana
- [ ] Heroine

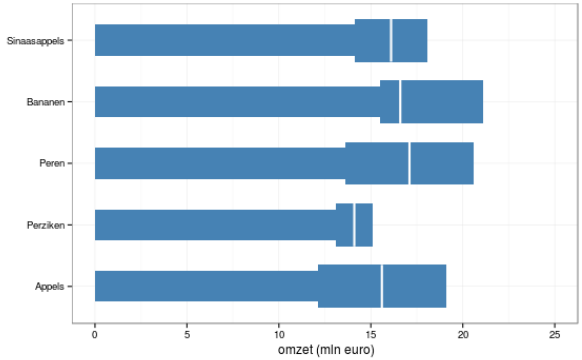
Drugsgebruik [2009, Heeft ooit drugs gebruikt, Cocaine]

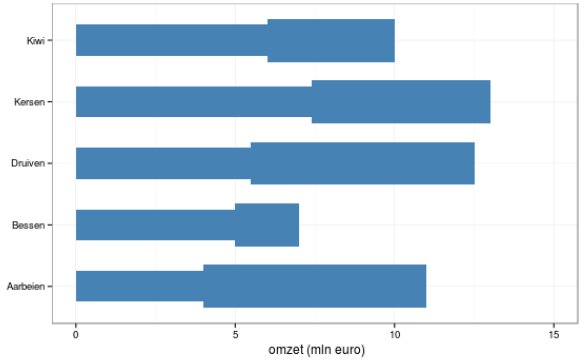


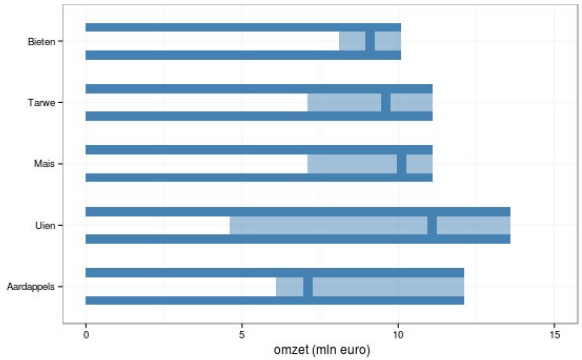

Sinaasappels
Bananen
Peren
Perziken
Appels
0
5
10
15
20
omzet (mln euro)

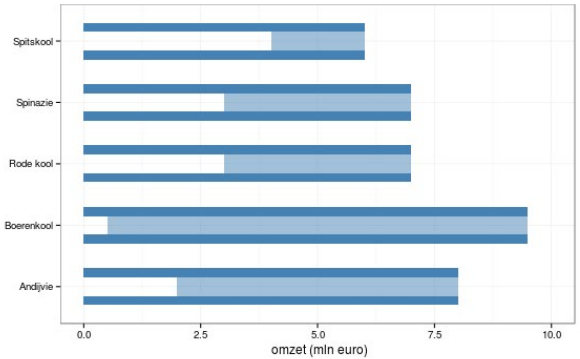

Kiwi
Kersen
Druiven
Bessen
Aarbeien
0
5
10
omzet (mln euro)

error
bars


Sinaasappels
Bananen
Peren
Perziken
Appels
0
5
10
15
20
25
omzet (mln euro)


Kiwi
Kersen
Druiven
Bessen
Aarbeien
0
5
10
15
omzet (mln euro)


Bieten
Tarwe
Mais
Uien
Aardappels
0
5
10
15
omzet (mln euro)


Spitskool
Spinazie
Rode kool
Boerenkool
Andijvie
0.0
2.5
5.0
7.5
10.0
omzet (mln euro)

Bar Charts:

– displaying uncertainty makes user less confident in comparison tasks. (which is good)
– No significant difference between methods
– PhDs prefer *error bar*, but *error bar* does do not perform better.

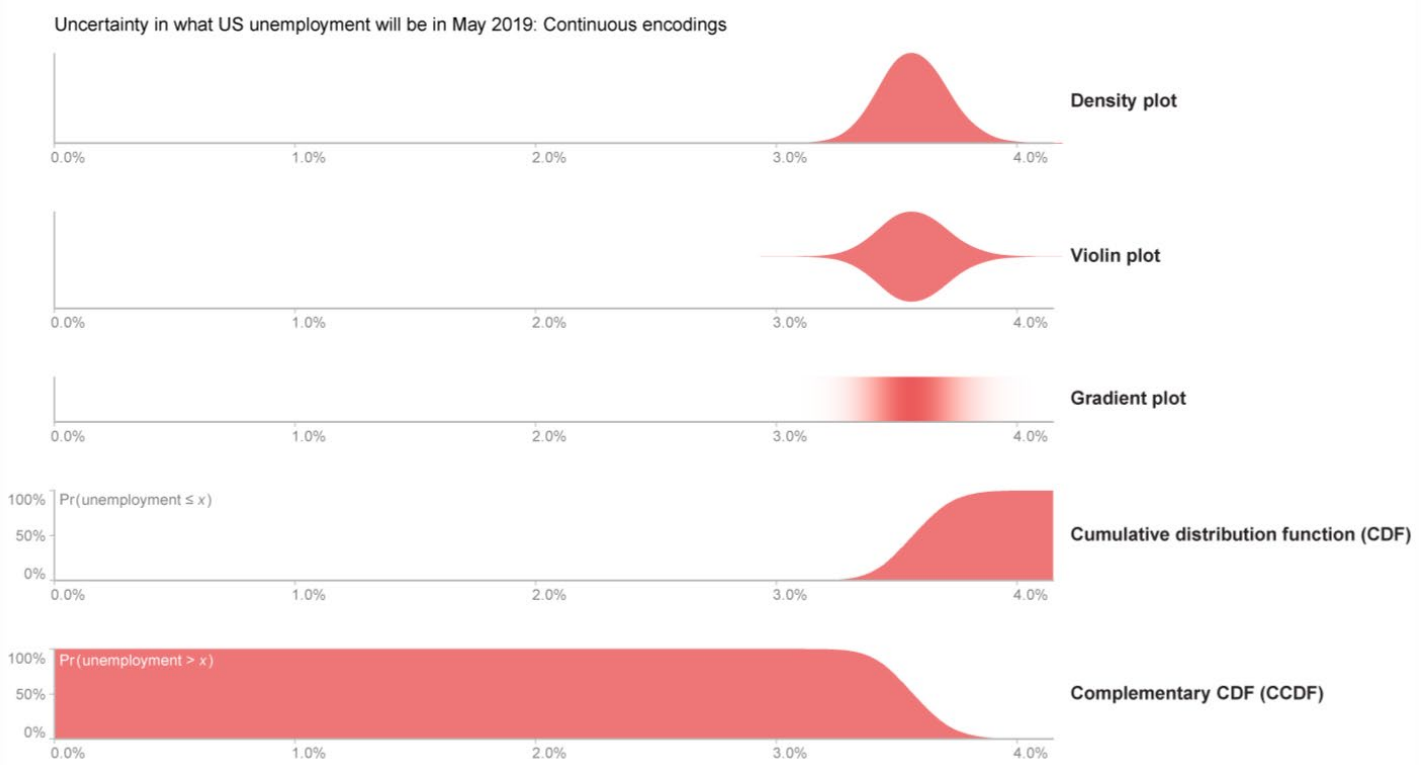
– When publishing intervals (without point):
    – cigarette is better

Uncertainty as distribution

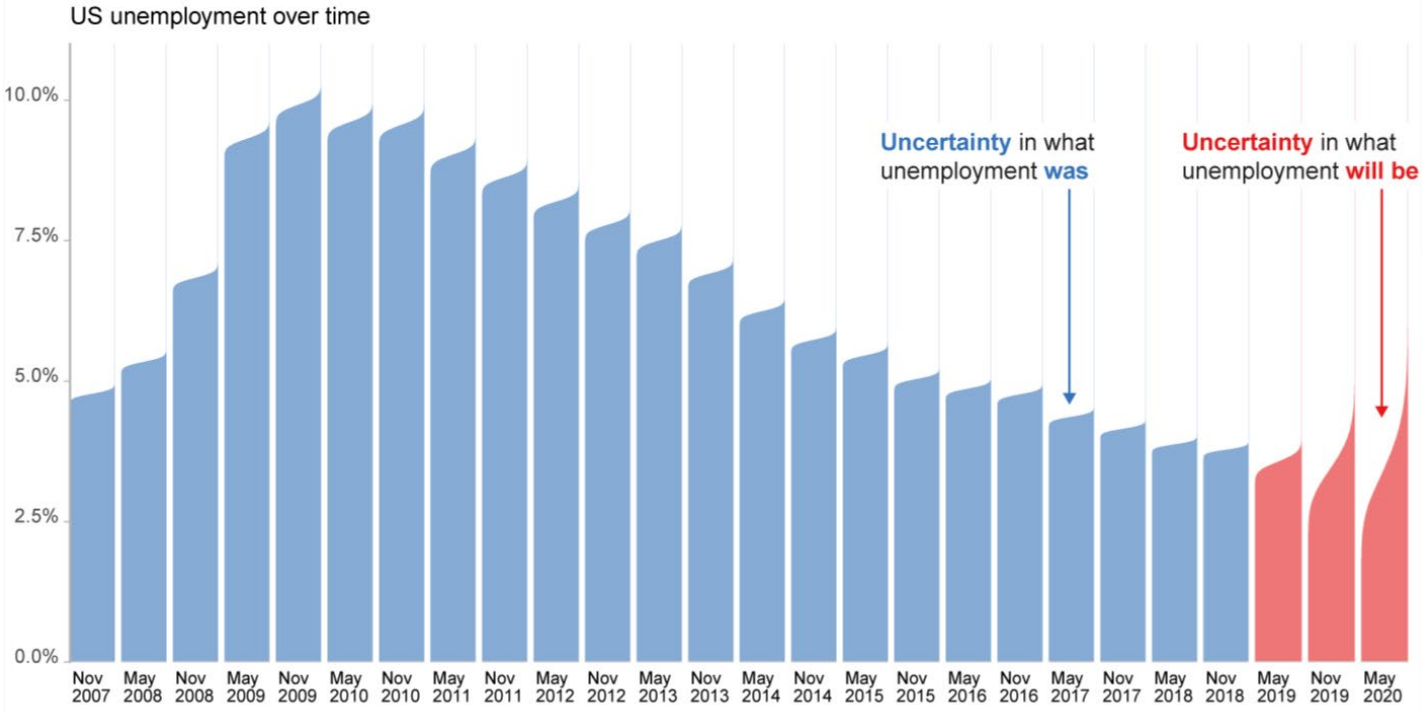
When a advanced statistical method is used:

- results in (probability) distribution (in stead of interval)

May use on of the following graphical methods:


Uncertainty in what US unemployment will be in May 2019: Continuous encodings
Density plot
Violin plot
Gradient plot
Cumulative distribution function (CDF)
Complementary CDF (CCDF)
Pr(unemployment ≤ x)
Pr(unemployment > x)
100%
50%
0%
0.0%
1.0%
2.0%
3.0%
4.0%

CCDF (Kay, 2019)



COMUNIKOS

COMUNIKOS: Eurostat project

- Goal: guidelines in COMmunicating Uncertain Knowledge in Official Statistics

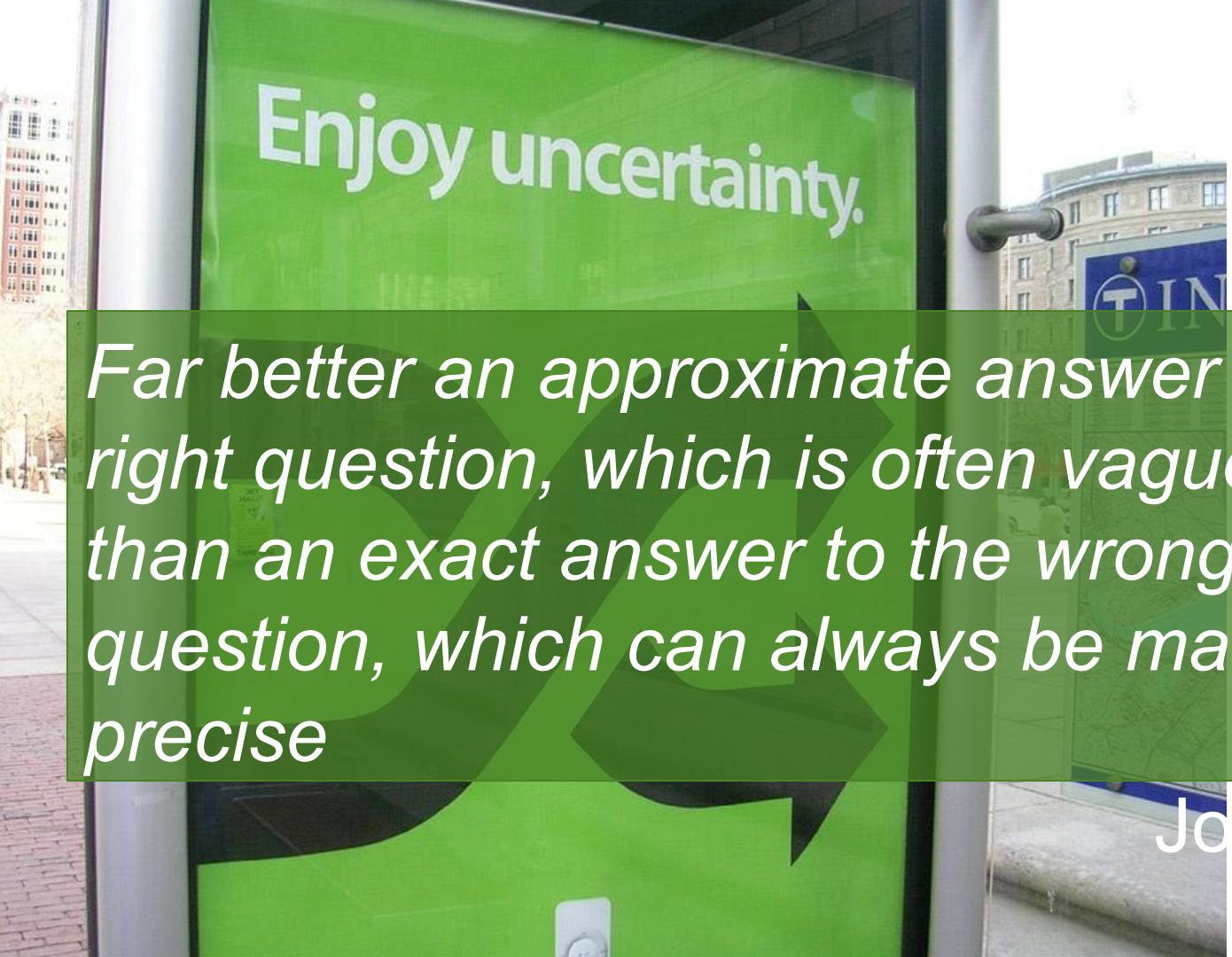
Tasks:
- Describe possible sources of uncertainty
- Visualisation Guidelines
- Methods for calculating uncertainty measures
- POC on Scanner Data

Advise:

Start publishing uncertainty measures

Plot them!

User appreciate it, and we are doing statistics aren't we?



Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise

Jo