Stochastiek 2

Inleiding in the Mathematische Statistiek

staff.fnwi.uva.nl/j.h.vanzanten

H.1 Introductie

Wat is statistiek? - 2

Statistiek is de kunst van het (wiskundig) modelleren van situaties waarin het toeval een rol speelt, en het trekken van conclusies op basis van data die zijn waargenomen in dergelijke situaties.

Bijma, Jonker, Van der Vaart, *Inleiding in de Statistiek*, 2013.

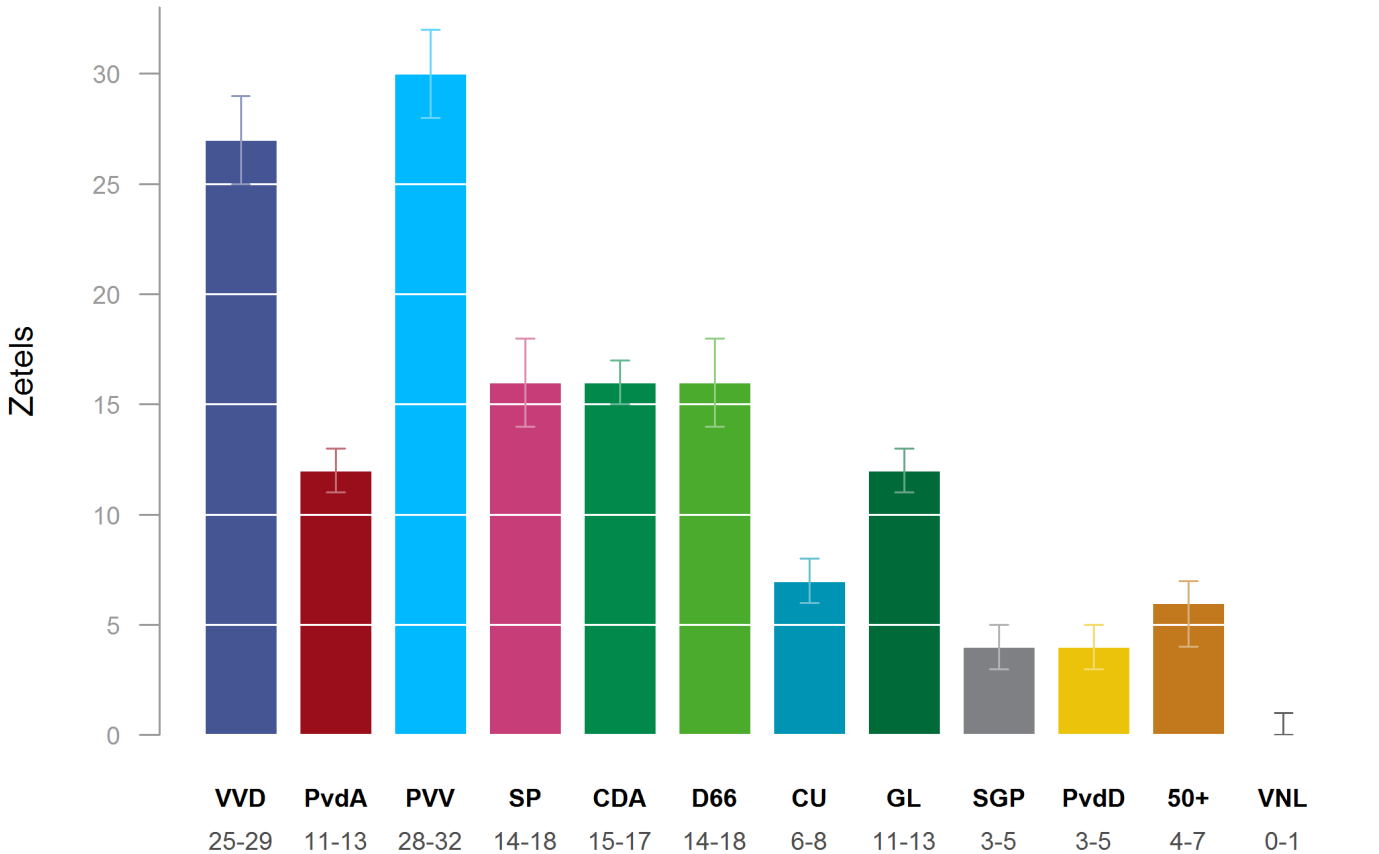
Wat is statistiek? - 2

Statistics concerns what can be learned from data. Applied statistics comprises a body of methods for data collection and analysis across the whole range of science, and in areas such as engineering, medicine, business, and law wherever variable data must be summarized, or used to test or confirm theories, or to inform decisions. Theoretical statistics underpins this by providing a framework for understanding the properties and scope of methods used in applications.

A.C. Davison, *Statistical Models*, Cambridge, 2003.

Typische statistische vragen

- Wat is de kans dat de Maas komend jaar buiten zijn oevers treed?
  $\rightarrow$ schatten (H.3)
- Is de nieuwe medische behandeling "significant" beter dan de oude?
  $\rightarrow$ hypothese toetsen (H.4)
- Wat is de onzekerheidsmarge in het voorspelde aantal zetels voor een politieke partij?
  $\rightarrow$ betrouwbaarheidsgebieden (H.5)



Peilingwijzer. Tom Louwerse, Universiteit Leiden. Bijgewerkt: 11-07-2016

Mathematische Statistiek

- We beschouwen de beschikbare data als realisatie(s) van een stochastische grootheid X. (Vaak een vector van de vorm $X = (X_1, \ldots, X_n)$.) Oftewel: we vatten de data op als *uitkomsten van een kansexperiment*.

Mathematische Statistiek

- We beschouwen de beschikbare data als realisatie(s) van een stochastische grootheid X. (Vaak een vector van de vorm $X = (X_1, \ldots, X_n)$.) Oftewel: we vatten de data op als *uitkomsten van een kansexperiment*.
- De variabiliteit in de data wordt beschreven met behulp van kansverdelingen.

Mathematische Statistiek

- We beschouwen de beschikbare data als realisatie(s) van een stochastische grootheid X. (Vaak een vector van de vorm $X = (X_1, \ldots, X_n)$.) Oftewel: we vatten de data op als *uitkomsten van een kansexperiment*.
- De variabiliteit in de data wordt beschreven met behulp van kansverdelingen.
- Statistisch model: collectie van kansverdelingen die de mogelijke variabiliteit in de data beschrijven.
- Doel van een statistische procedure: het identificeren van de verdeling die de data het "beste" beschrijft.

Voorbeeld 1.2

De opiniepeiler Maurice wil weten welke fractie van de Nederlandse kiezers op dit moment op de PVV zou stemmen. Daartoe vraagt hij 1000 willekeurig gekozen kiezers of ze dat zouden doen. Het aantal mensen dat bevestigend antwoordt is 143.

- Data:

Voorbeeld 1.2

De opiniepeiler Maurice wil weten welke fractie van de Nederlandse kiezers op dit moment op de PVV zou stemmen. Daartoe vraagt hij 1000 willekeurig gekozen kiezers of ze dat zouden doen. Het aantal mensen dat bevestigend antwoordt is 143.

- Data: (realisatie van) $X =$"aantal mensen dat zegt op de PVV te stemmen"
- Mogelijk model:

Voorbeeld 1.2

De opiniepeiler Maurice wil weten welke fractie van de Nederlandse kiezers op dit moment op de PVV zou stemmen. Daartoe vraagt hij 1000 willekeurig gekozen kiezers of ze dat zouden doen. Het aantal mensen dat bevestigend antwoordt is 143.

- Data: (realisatie van) $X =$"aantal mensen dat zegt op de PVV te stemmen"
- Mogelijk model: $\{\text{Bin}(1000, p) : p \in [0, 1]\}$. Interpretatie: p is de fractie van de Nederlandse kiezers die op dit moment op de PVV zou stemmen.
- Statistische probleem:

Voorbeeld 1.2

De opiniepeiler Maurice wil weten welke fractie van de Nederlandse kiezers op dit moment op de PVV zou stemmen. Daartoe vraagt hij 1000 willekeurig gekozen kiezers of ze dat zouden doen. Het aantal mensen dat bevestigend antwoordt is 143.

- Data: (realisatie van) $X =$"aantal mensen dat zegt op de PVV te stemmen"
- Mogelijk model: $\{\text{Bin}(1000, p) : p \in [0, 1]\}$. Interpretatie: p is de fractie van de Nederlandse kiezers die op dit moment op de PVV zou stemmen.
- Statistische probleem: we willen p schatten m.b.v. de beschikbare data. Voor de hand liggende keuze: $\hat{p} =$

Voorbeeld 1.2

De opiniepeiler Maurice wil weten welke fractie van de Nederlandse kiezers op dit moment op de PVV zou stemmen. Daartoe vraagt hij 1000 willekeurig gekozen kiezers of ze dat zouden doen. Het aantal mensen dat bevestigend antwoordt is 143.

- Data: (realisatie van) $X =$"aantal mensen dat zegt op de PVV te stemmen"
- Mogelijk model: $\{\text{Bin}(1000, p) : p \in [0, 1]\}$. Interpretatie: p is de fractie van de Nederlandse kiezers die op dit moment op de PVV zou stemmen.
- Statistische probleem: we willen p schatten m.b.v. de beschikbare data. Voor de hand liggende keuze: $\hat{p} = 143/1000 \approx 14,3\% \approx 21$ zetels.
- Belangrijk: wat is de betrouwbaarheid van deze schatting?

Voorbeeld 1.3

In 1879 deed de natuurkundige Michelson experimenten om de snelheid van het licht te bepalen. In één van de runs kreeg hij de volgende 20 waarden (in km/sec):

299850	299740	299900	300070	299930
299850	299950	299980	299980	299880
300000	299980	299930	299650	299760
299810	300000	300000	299960	299960

Wat is nu de snelheid van het licht?

- Data: vector van snelheden $X = (X_1, \ldots, X_{20})$.
- Model: ???

Voorbeeld 1.3

Poging 1

- Redelijke veronderstelling: $X_1, \ldots, X_{20}$ onafhankelijk.
- Idee: schrijf $X_i = \mu + e_i$, met

μ : snelheid van het licht

e_i : meetfout.

Gebruikelijke aanname: $\mathbb{E}e_i = 0$.

Voorbeeld 1.3

Poging 1

- Redelijke veronderstelling: $X_1, \ldots, X_{20}$ onafhankelijk.
- Idee: schrijf $X_i = \mu + e_i$, met

μ : snelheid van het licht

e_i : meetfout.

Gebruikelijke aanname: $\mathbb{E}e_i = 0$.

Model 1: Alle (simultane) verdelingen van $X = (X_1, \ldots X_{20})$ zodanig dat de $X_1, \ldots, X_{20}$ onafhankelijk zijn en $\mathbb{E}X_i = \mu$ voor $i = 1, \ldots, 20$.

Stat. probleem: schat μ, of geef een betrouwbaarheidsinterval.

Voorbeeld 1.3

Poging 2

- Redelijke veronderstelling: $X_1, \ldots, X_{20}$ onafhankelijk.
- Idee: schrijf weer $X_i = \mu + e_i$ en postuleer een aanname over de verdeling van e_i. Wat is redelijk?

Voorbeeld 1.3

Poging 2

- Redelijke veronderstelling: $X_1, \ldots, X_{20}$ onafhankelijk.
- Idee: schrijf weer $X_i = \mu + e_i$ en postuleer een aanname over de verdeling van e_i. Wat is redelijk?

Als de fout is samengesteld uit heel veel kleine, onafhankelijke bijdragen, dan suggereert de Centrale Limietstelling dat het redelijk is om aan te nemen dat e_i normaal verdeeld is. Vaak nemen we aan dat de fouten dezelfde variantie hebben.

Voorbeeld 1.3

Poging 2

- Redelijke veronderstelling: $X_1, \ldots, X_{20}$ onafhankelijk.
- Idee: schrijf weer $X_i = \mu + e_i$ en postuleer een aanname over de verdeling van e_i. Wat is redelijk?

Als de fout is samengesteld uit heel veel kleine, onafhankelijke bijdragen, dan suggereert de Centrale Limietstelling dat het redelijk is om aan te nemen dat e_i normaal verdeeld is. Vaak nemen we aan dat de fouten dezelfde variantie hebben.

Model 2: Alle (simultane) verdelingen van $X = (X_1, \ldots X_{20})$ zodanig dat de $X_1, \ldots, X_{20}$ onafhankelijk zijn en $X_i \sim N(\mu, \sigma^2)$ voor zekere $\mu \in \mathbb{R}$ en $\sigma^2 > 0$.

Algemene setup

- Data: stochastische grootheid X. Vaak $X = (X_1, \ldots, X_n)$.
- Model: collectie kansverdelingen $\{P_\theta : \theta \in \Theta\}$ die de mogelijke verdelingen van de data beschrijven. Terminologie: θ: parameter, Θ: parameter ruimte.

Algemene setup

- Data: stochastische grootheid X. Vaak $X = (X_1, \ldots, X_n)$.
- Model: collectie kansverdelingen $\{P_\theta : \theta \in \Theta\}$ die de mogelijke verdelingen van de data beschrijven. Terminologie: θ: parameter, Θ: parameter ruimte.

- Doel: schatten van θ, toetsen van hypotheses over θ, construëren van betrouwbaarheidsgebieden voor θ.

Meer voorbeelden: 1.4–1.8 in het boek.