Limperg Instituut

Cijferanalyse op statistische grondslag

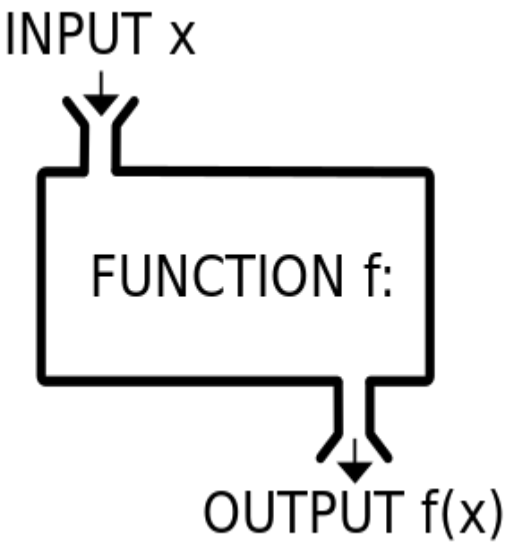
dr. ir. Rabin Neslo

Limperg Instituut

Inhoud presentatie

- Achtergrond
- Zin en onzin
- Context
- Gebruik cijferanalyse
  - Beschrijvende statistieken
  - Verkenning van data
  - Bevestigen van aannames of stellingen

Achtergrond



- Elke entiteit heeft een input (ingaande resources)
- Een functie (bewerking van de ingaande resources)
- En een output (uitgaande resources)
- De input, functie en output worden meestal vastgelegd
- (Externe) controleurs moeten een oordeel vellen op het geheel

Zin en onzin van cijferanalyse

- Zin
  - Volgens COS 520 helpt cijferanalyse:
    - bij het maken van financiële overzichten
    - bij het inschatten van risico's
    - bij het trekken van conclusies over een entiteit
  - Helpt cijferanalyse om de werking van een entiteit beter te leren kennen
  - Helpt bij het opsporen van fouten en verschillen
  - Uiteindelijk is de data een representatie van de werkelijkheid
- Onzin
  - Zonder goede checks and balances op het ontstaan van deze data bestaat de mogelijkheid om verkeerde conclusies te trekken

Limperg Instituut

Context cijferanalyse

- Wie is de leverancier van de cijfers
- Hoe zijn de cijfers samengesteld
- Wat voor informatie bevatten de cijfers
- Wanneer zijn de cijfers samengesteld
- Waarom en waarvoor zijn de cijfers samengesteld

Gebruik cijferanalyse

- Cijferanalyse kent verschillende fases
  - Beschrijvende statistieken (understanding the business)
  - Verkennen van data (initiële cijferbeoordeling)
  - Bevestigen van aannames of stellingen1
- Voor al die drie fases biedt de kansrekening en statistiek verschillende gereedschappen
- Met kansberekening veronderstel je de onzekerheid over het onderliggend proces en bepaal je wat de output zal moeten zijn
- Bij statistiek observeer je de uitkomsten en je gaat dan opzoek naar verdelingen die het best bij die uitkomsten passen

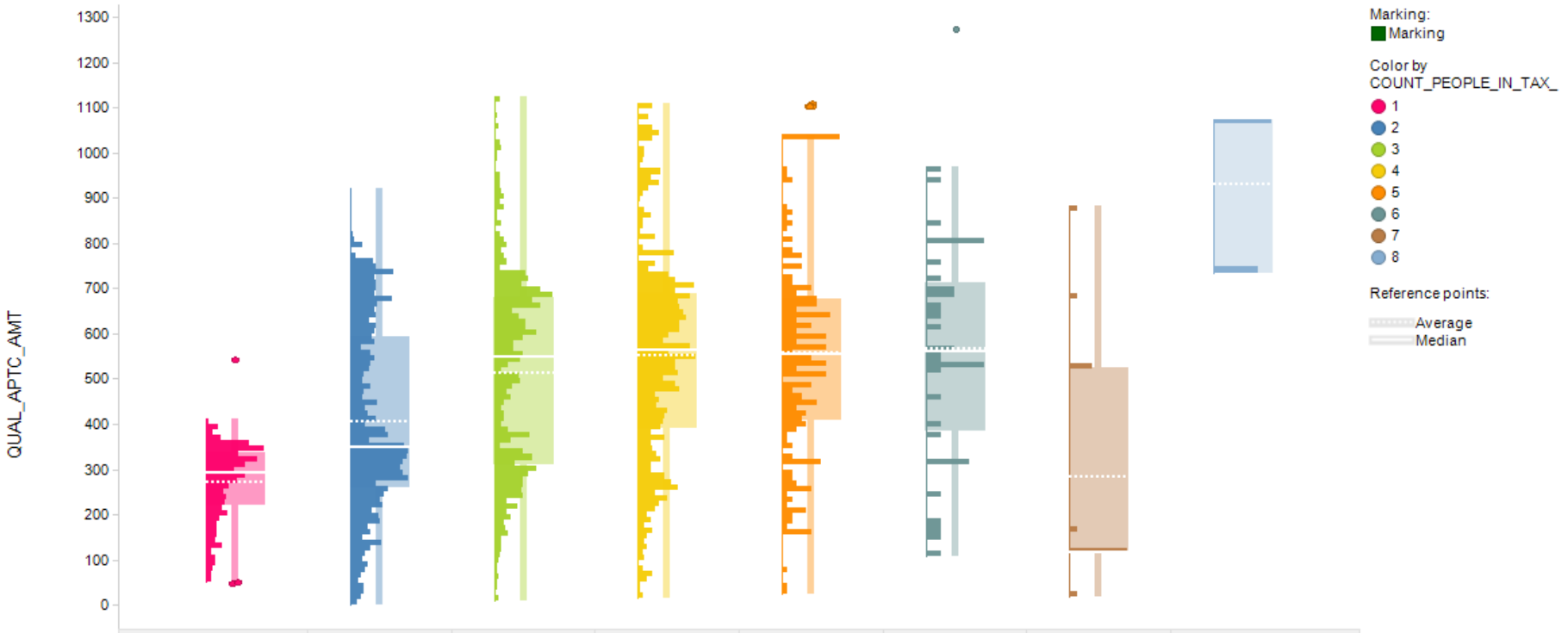
Beschrijvende statistieken

- Beschrijven, weergeven en samenvatten van cijfers
- Statistieken voor cijfers van afzonderlijke variabelen
  - Centrale tendens (rekenkundig en meetkundig gemiddelden, mediaan, modus)
  - Centrale spreiding (absolute of standaard afwijking, variantie, bereik, kwartielen, interkwartielbereik
- Voor verbanden tussen variabelen
  - Correlaties en covarianties
  - Kruistabellen
- Visualisaties helpen bij het duiden van deze kerngetallen
- Conclusies n.a.v. deze getallen geldt alleen voor de gebruikte cijfers

Limperg Instituut

Voorbeelden

Box Plot



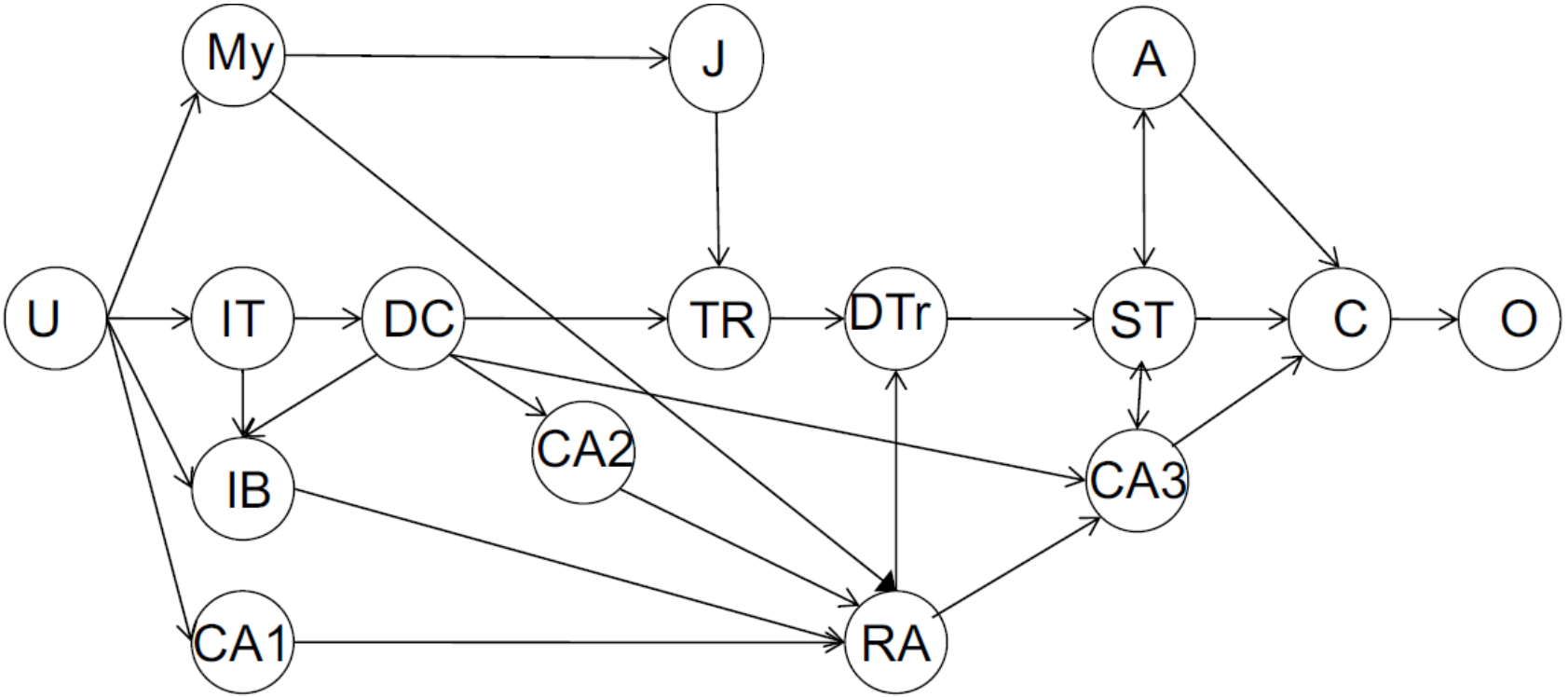
	1	2	3	4	5	6	7	8
Count	3865	2080	775	843	232	39	15	7
Avg	272.523	405.169	513.094	553.106	558.096	568.176	283.246	930.469
Median	291.99	350.58	547.94	563.39	556.645	562.25	116.06	1076.59
Outliers	3	0	0	0	6	1	0	0
Var	6492.56	41952.8	48283.6	53894.5	55912.6	65711.2	71602.4	33213.4
StdDev	80.5764	204.824	219.735	232.152	236.459	256.342	267.586	182.245

COUNT_PEOPLE_IN_TAX_HH

Beschrijvende statistieken

- Zijn ook kernwaarden voor complexere data als authorisatie data of netwerk data.
  - Degree centrality (knooppunt met de meeste connecties)
  - Betweenness centrality (maat die bijhoudt hoe vaak een knooppunt een lid is van het kortste pad tussen twee knooppunten).
  - Closeness centrality (maat voor het bijhouden van de gemiddelde afstand tussen alle knooppunten voor een gegeven knooppunt.
  - Katz centrality telt alle gewogen afstanden op naar alle andere knooppunten van een gegeven knooppunt.
- Netwerk data wordt nu vaak gebruikt om het verloop van interne processen te analyseren (DISCO processmining)

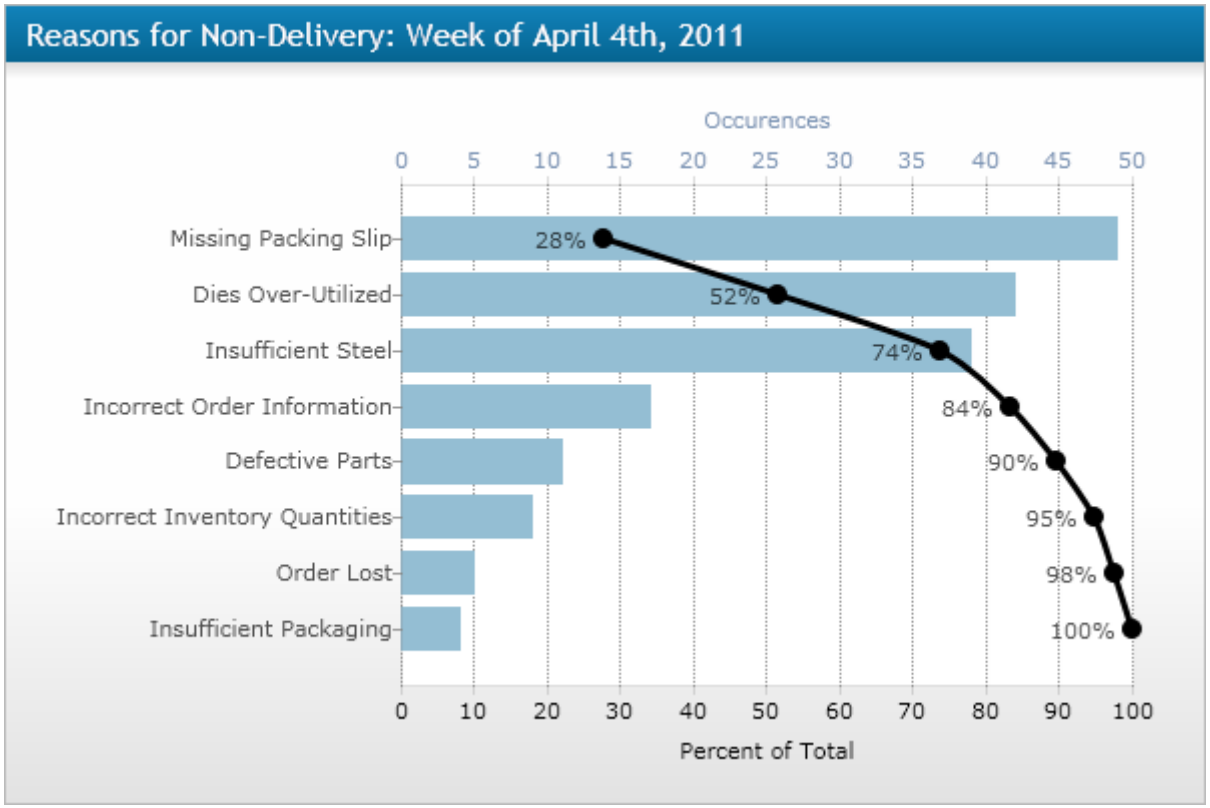
Voorbeelden



Verkenning van cijfers

- Testen van aannames van mogelijke oorzaken
- Selecteren van de juiste statistische methoden en technieken
- Formuleren van aannames voor statistische inferentie
- Mogelijke technieken zijn
  - Histogrammen
  - Pareto grafieken
  - Spreidingsdiagrammen
  - Dimensie reductie technieken

Voorbeelden



Bevestigen van aannames en stellingen

- Stringer & Stewart formaliseerden statistische analyse binnen het audit vakgebied (STAR)
- Meeste voorkomende statistische methode binnen STAR is regressie analyse
- Meerdere methodes om te bepalen als cijfers binnen een bepaald patroon passen
  - Logistic en Probit regression
  - Support vector machines (SVM)
  - Naïve Bayes classifier
  - Decision trees
- Deze methoden en technieken worden ook gebruikt binnen machine learning

Voorbeeld uit de praktijk