Databehandling och rensning

Outliers, transformationer och datakvalitet

Mathias Johansson Kristofer Söderström

2025-12-15

Innehållsförteckning

1 Introduktion

Innan statistisk analys måste data ofta rensas och bearbetas. Denna lektion behandlar identifiering och hantering av extremvärden (outliers) samt datatransformationer.

2 Outliers (Extremvärden)

2.1 Vad är en outlier?

En **outlier** är en observation som ligger onormalt långt från övriga värden i datasetet. Outliers kan vara:

1. **Genuina extremvärden**: Verkliga observationer som är ovanliga
2. **Mätfel**: Felaktiga värden från felinmatning eller mätfel
3. **Datainmatningsfel**: T.ex. "1000" istället för "100"

2.2 Identifiera outliers

2.2.1 1. Visuell inspektion

Boxplot: - Värden utanför 1.5 × IQR från kvartilerna markeras som outliers - IQR (Interquartile Range) = Q3 - Q1

Histogram: - Isolerade värden långt från huvuddistributionen

Scatterplot: - Punkter långt från huvudmönstret

2.2.2 2. Statistiska metoder

Z-score metod:

$$z = \frac{x - \bar{x}}{s}$$

- Om |z| > 3: Potentiell outlier (ligger mer än 3 standardavvikelser från medelvärdet)
- Om |z| > 2: Möjlig outlier

IQR-metod: - Nedre gräns: Q1 - 1.5 × IQR - Övre gräns: Q3 + 1.5 × IQR - Värden utanför dessa gränser är outliers

2.3 Exempel: Identifiera outliers

Data: Ålder på biblioteksanvändare

```
18, 19, 21, 22, 23, 24, 25, 26, 28, 95
```

Boxplot-metoden: - Q1 = 20, Q3 = 26 - IQR = 26 - 20 = 6 - Övre gräns: 26 + 1.5 × 6 = 35 - Värdet 95 är en outlier

Z-score-metoden: - Medelvärde 30.1 - Standardavvikelse 23.3 - Z-score för 95: (95 - 30.1) / 23.3 2.78

Värdet 95 är en möjlig outlier ($|z| > 2$).

3 Hantera outliers

3.1 1. Undersök orsaken

Frågor att ställa: - Är det ett mätfel eller inmatningsfel? - Är det ett genuint extremvärde? - Kan det förklaras av speciella omständigheter?

3.2 2. Olika strategier

3.2.1 A. Ta bort outliers

När: - Uppenbart fel (t.ex. ålder = 999) - Mätfel som inte kan korrigeras

Försiktighet: - Dokumentera alltid vad som tagits bort och varför - Rapportera antal borttagna observationer

3.2.2 B. Behåll outliers

När: - Genuina extremvärden som är viktiga för analysen - Studien handlar om att identifiera extremfall

3.2.3 C. Transformera data

När: - Outliers gör distributionen skev - Du vill behålla informationen men minska påverkan

3.2.4 D. Använd robusta metoder

När: - Outliers finns men du är osäker på orsaken - Använd median istället för medelvärde - Använd IQR istället för standardavvikelse

3.3 3. Exempel: Hantera outliers

Scenario: Löneddata med några extremt höga värden

Alternativ 1 - Ta bort:

```
Original: 30k, 32k, 35k, 38k, 40k, 42k, 500k
Renad: 30k, 32k, 35k, 38k, 40k, 42k
```

Alternativ 2 - Transformera: Använd logaritmtransformation (se nedan)

Alternativ 3 - Robust metod: - Använd median (38k) istället för medelvärde (102k)

4 Datatransformationer

4.1 Varför transformera?

1. **Hantera skevhet**: Göra fördelningen mer symmetrisk
2. **Stabilisera varians**: Göra spridningen mer konstant
3. **Linearisera samband**: Göra icke-linjära samband linjära
4. **Minska outliers påverkan**: Komprimera extremvärden

4.2 Vanliga transformationer

4.2.1 1. Logaritmtransformation

Formel:

$$y' = \log(y)$$

eller:

$$y' = \log(y + 1)$$

(om data innehåller nollor)

Användning: - Höger-skev data (lång svans åt höger) - Exponentiell tillväxt - Data som spänner över flera storleksordningar

Exempel:

```
Original: 10, 100, 1000, 10000
Log :    1,   2,    3,     4
```

Praktiska tillämpningar: - Inkomst/löner - Befolkningsstorlek - Husantika priser - Viruspartiklar

4.2.2 2. Kvadratrotstransformation

Formel:

$$y' = \sqrt{y}$$

Användning: - Måttligt skev data - Räknedata (count data) som följer Poisson-fördelning

Exempel:

```
Original: 1,  4,  9,  16,  25
Sqrt:     1,  2,  3,   4,   5
```

4.2.3 3. Kvadrattransformation

Formel:

$$y' = y^2$$

Användning: - Vänster-skev data - När man vill förstärka skillnader

4.2.4 4. Reciprok (invers) transformation

Formel:

$$y' = \frac{1}{y}$$

Användning: - Starkt höger-skev data - Tidsmätningar (konvertera tid till hastighet)

4.2.5 5. Box-Cox transformation

Formel:

$$y' = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{om } \lambda \neq 0 \\ \log(y) & \text{om } \lambda = 0 \end{cases}$$

Användning: - Automatisk val av bästa transformation - hittas genom att maximera normalitet

4.3 Exempel: Logaritmtransformation av inkomstdata

Original data (tusen kr/månad):

```
Data: 25, 28, 30, 32, 35, 38, 40, 45, 120
Medelvärde: 43.7
Standardavvikelse: 28.6
Skevhet: Höger-skev
```

Efter log-transformation:

```
Log-data: 3.22, 3.33, 3.40, 3.47, 3.56, 3.64, 3.69, 3.81, 4.79
Medelvärde: 3.66
Standardavvikelse: 0.45
Skevhet: Mindre skev, mer normalfördelad
```

Fördel: Extremvärdet (120) har mindre påverkan efter transformation.

5 Tolkning av transformerad data

5.1 Backtransformation

När du använt log-transformation:

```
log(y) = 2.5
y = 10^2.5   316  (om log )
y = e^2.5   12.2  (om ln)
```

5.2 Tolkning av resultat

Original skala: "Medellönen är 43,700 kr"

Log-skala: "Median log-lönen är 3.64, vilket motsvarar en lön på 10^3.64 43,650 kr"

Viktigt: Efter log-transformation representerar medelvärdet den geometriska medellönen, inte den aritmetiska.

6 Standardisering och normalisering

6.1 Z-score standardisering

Formel:

$$z = \frac{x - \bar{x}}{s}$$

Resultat: - Medelvärde = 0 - Standardavvikelse = 1

Användning: - Jämföra variabler med olika skalor - Före vissa statistiska analyser

6.2 Min-Max normalisering

Formel:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Resultat: - Värden mellan 0 och 1

Användning: - Maskininlärning - När du vill behålla relativa avstånd

7 Praktiska riktlinjer

7.1 Checklista för databehandling

1. **Inspektera data**
   - Visualisera med histogram, boxplot
   - Kontrollera för saknade värden
   - Identifiera potentiella outliers
2. **Validera outliers**
   - Kontrollera om det är mätfel
   - Undersök bakomliggande orsaker
   - Dokumentera alla beslut
3. **Välj strategi**
   - Ta bort, behåll eller transformera?
   - Basera på forskningsfråga och datatyp
4. **Dokumentera**
   - Antal borttagna observationer
   - Vilka transformationer använts
   - Motivering för beslut
5. **Rapportera**
   - Beskriv databehandling i metodavsnittet
   - Rapportera effekt på resultat
   - Var transparent

7.2 Exempel: Komplett arbetsflöde

1. Rådata: Antal besök på webbplats per dag

```
15, 18, 20, 22, 25, 28, 30, 35, 40, 500
```

2. Identifiera: - Boxplot visar 500 som outlier - Z-score för 500: 3.2 (stark outlier)

3. Undersök: - Kontrollera loggar: Visar sig vara botaktivitet, inte äkta besök

4. Beslut: - Ta bort värdet 500

5. Resultat:

```
Renad data: 15, 18, 20, 22, 25, 28, 30, 35, 40
Medelvärde: före = 73.3, efter = 25.9
```

6. Rapportering: "En observation (500 besök) identifierades som outlier och verifierades vara botaktivitet. Denna observation exkluderades från analysen, vilket resulterade i ett dataset med n=9."

8 Varningar och etiska överväganden

8.1 Undvik cherry-picking

- **FEL:** Ta bort outliers för att få "bättre" resultat
- **RÄTT:** Ta bort outliers baserat på objektiva kriterier definierade i förväg

8.2 Rapportera alltid

- Antal borttagna observationer
- Kriterier för borttagning
- Skillnad i resultat med/utan outliers

8.3 Var försiktig med transformationer

- Log-transformation kan dölja viktiga extremvärden
- Alltid visa både original och transformerad data
- Tolkning blir mer komplex

8.4 Sensitivitetsanalys

Kör analysen både: 1. Med outliers 2. Utan outliers 3. Med transformerad data

Om resultaten skiljer sig dramatiskt, var extra försiktig med slutsatser.

9 Viktiga begrepp

Svenska	Engelska
Extremvärde	Outlier
Transformation	Transformation
Logaritm	Logarithm
Kvadratrot	Square root
Standardisering	Standardization
Normalisering	Normalization
Skevhet	Skewness
Backtransformation	Back-transformation
Robust statistik	Robust statistics

10 Interaktiva verktyg

- **Datatyper** - Utforska effekten av log-transformation och outlier-borttagning
- **Statistikkalkylator** - Beräkna statistik och se effekten av outliers

11 Vidare läsning

- Beskrivande statistik - Grundläggande statistiska mått
- Presentera resultat - Visualisera data effektivt

12 Referenser

- Wikipedia: Outlier
- Wikipedia: Data transformation (statistics)
- Wikipedia: Logarithmic scale
- Wikipedia: Box-Cox transformation
- Wikipedia: Interquartile range
- Wikipedia: Standard score
- Wikipedia: Normalization (statistics)
- Wikipedia: Robust statistics
- Wikipedia: Skewness