Einführung in die Quantitative Datenanalyse

Sitzung 5: Lineare Regression III und Deskriptive Statistik

Proseminar an der Freien Universität Berlin
2.05.2017 - Marcus Spittler
Inhalt der 5. Sitzung

- Univariate Häufigkeitsverteilungen
  - Absolute Häufigkeit
  - Relative Häufigkeit
  - Kumulierte Häufigkeit

- Grafische Darstellung
  - Balkendiagramm
  - Histogramm
  - Dichteverteilung
  - Boxplot

- Lineare Regression III
  - Modellbildung
  - Interpretation
Univariate Häufigkeitsverteilungen

Annahme: nicht-häufbare Merkmale, das heißt: eine Untersuchungseinheit kann stets nur eine Ausprägung bei einem Merkmal haben. D.h. die Antwortmöglichkeiten sind diskjunkt, es gibt keine Schnittmenge zwischen ihnen.

Ordnet man den Merkmalsausprägungen Häufigkeiten zu, so erhält man eine Häufigkeitsverteilung.
Absolute Häufigkeit

Absolute Häufigkeit $h_{x_j}$ oder einfacher $h_j$: Die Anzahl der Merkmalsträger mit der Ausprägung $x_j$.

**Eigenschaften:**

$$0 \leq h_j \leq 1$$

$$\sum_{j=1}^{k} (h_j) = n$$
**AbsOLUTE HäUFIGKEIT**

**Beispiel:** In der Wahlstudie zur Europawahl 2014 wurde gefragt: *Bitte sagen Sie mir für jede der folgenden Aussagen, inwieweit diese Ihrer Ansicht oder Meinung entspricht bzw. nicht entspricht: Sie sind sehr an Politik interessiert.*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Absolute H.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ja, voll und ganz</td>
<td>529</td>
</tr>
<tr>
<td>Ja, teilweise</td>
<td>667</td>
</tr>
<tr>
<td>Nein, nicht wirklich</td>
<td>315</td>
</tr>
<tr>
<td>Nein, sicher nicht</td>
<td>137</td>
</tr>
<tr>
<td>Summe:</td>
<td>1648</td>
</tr>
</tbody>
</table>

Balkendiagramm (*Barchart*) mit absoluten Häufigkeiten
Relative Häufigkeit

Vergleich zweier absoluten Häufigkeitsverteilungen für ein Merkmal schwierig, v.a. wenn $n$ unterschiedlich, daher Verwendung der relativen Häufigkeit.

\[ f_{x_j} = \frac{h_{x_j}}{n} \]

$f_j \times 100$ ergibt die prozentualen Anteile

<table>
<thead>
<tr>
<th>Variable</th>
<th>Absolute H.</th>
<th>Relative H.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ja, voll und ganz</td>
<td>529</td>
<td>0.32</td>
</tr>
<tr>
<td>Ja, teilweise</td>
<td>667</td>
<td>0.40</td>
</tr>
<tr>
<td>Nein, nicht wirklich</td>
<td>315</td>
<td>0.19</td>
</tr>
<tr>
<td>Nein, sicher nicht</td>
<td>137</td>
<td>0.08</td>
</tr>
<tr>
<td><strong>Summe</strong></td>
<td><strong>1648</strong></td>
<td><strong>1</strong></td>
</tr>
</tbody>
</table>

Variable | Absolute H. | Relative H. |
Balkendiagramm (Barchart) mit relativen Häufigkeiten
Irreführende Grafiken: Balkendiagramm

Balkendiagramme sollten immer am Nullpunkt beginnen. Falls nicht, sollten sie das sehr deutlich dokumentieren.
Irreführende Grafiken: Balkendiagramm

High school graduation rates in the US

Die gleichen Daten in einer korrigierten Grafik
Kumulierte Häufigkeit

Mit den relativen Summenhäufigkeiten lässt sich die Summenhäufigkeitsfunktion $F_x$ definieren (empirische Verteilungsfunktion). Sie gibt zu jeder Merkmalsausprägung den Anteil der Untersuchungseinheiten an, die kleiner oder höchstens gleich einer Ausprägung sind. Die Summenhäufigkeitsfunktion hat (insbesondere bei nur wenigen Ausprägungen) das Bild einer Treppenfunktion.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Absolute H.</th>
<th>Relative H.</th>
<th>Kumulierte H.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ja, voll und ganz</td>
<td>529</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>Ja, teilweise</td>
<td>667</td>
<td>0.40</td>
<td>0.72</td>
</tr>
<tr>
<td>Nein, nicht wirklich</td>
<td>315</td>
<td>0.19</td>
<td>0.91</td>
</tr>
<tr>
<td>Nein, sicher nicht</td>
<td>137</td>
<td>0.08</td>
<td>0.99</td>
</tr>
<tr>
<td>Summe:</td>
<td>1648</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>
Balkendiagramm mit linearem Merkmal

*In der Politik spricht man von links und rechts. Welche Position haben Sie? Bitte geben Sie Ihren persönlichen Standpunkt auf einer Skala von 0 bis 10 an. 0 bedeutet □ links □ und 10 bedeutet □ rechts □. Welche Zahl gibt am besten Ihren Standpunkt wider?*

Quelle: EES Voter Study 2014
Balkendiagramm mit relativen Häufigkeiten
Treppenfunktion

Kumulierte Häufigkeit (F)

Links-/Rechts-Selbsteinstufung
Im Unterschied zum Balkendiagramm sind hier die Flächen interpretierbar. Metrische Merkmale werden in Klassen eingeteilt (engl. bins) mit konstanter oder variabler Klassenbreite.
Histogramm

Histogramm mit alternativer Klasseneinteilung
Hier ist ein Histogramm hilfreich, da die Klassen unterschiedlich breit sind.
Klassierte Daten

- Es gibt zwei Gründe, klassierte Daten zu betrachten:
  - Es gibt bei einer Befragung sehr viele unterschiedliche Merkmalswerte, so dass die empirische Verteilungsfunktion zu nahezu keiner Informationsverdichtung führt: Nachträgliche Klassenbildung.
  - Es sind in einer sekundärstatistischen Analyse nur klassierte Häufigkeitstabellen verfügbar: Rechnen mit vorgegebenen Klassengrenzen.

- Die Gestaltungsparameter der Klassierung sind Anzahl und Breite der Klassen.
Histogramm mit nachträglicher Klassenbildung
Histogramm mit nachträglicher Klassenbildung
**Klassierte Daten**

Beispiel: Dauer von Arbeitslosigkeit

<table>
<thead>
<tr>
<th>Klasse</th>
<th>Dauer (in Monaten)</th>
<th>Klassenbreite</th>
<th>Klassenmitte</th>
<th>Anzahl</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 bis 1</td>
<td>1</td>
<td>0.5</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>über 1 bis 2</td>
<td>1</td>
<td>1.5</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>über 2 bis 3</td>
<td>1</td>
<td>2.5</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>über 3 bis 6</td>
<td>3</td>
<td>4.5</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>über 6 bis 12</td>
<td>6</td>
<td>9.0</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>über 12 bis 24</td>
<td>12</td>
<td>18.0</td>
<td>6</td>
</tr>
</tbody>
</table>

Summe: 120
Häufigkeitsdichte

Die absolute Häufigkeit gibt an, wie viele der Beobachtungen in eine Klasse fallen. Wird mit ungleichen Klassenbreiten gearbeitet, so ist neben der absoluten bzw. relativen Häufigkeit auch die Häufigkeitsdichte interessant.

Zweck der Häufigkeitsdichte ist es bei ungleichen Klassenbreiten die tatsächlichen Häufigkeiten durch die jeweilige Klassenbreite zu relativieren.

Häufigkeitsdichte ist definiert als:

\[ f^*(x_j) = \frac{f_j}{\Delta x_j} = \frac{Relative\ H.}{Klassenbreite} \]
Häufigkeitsdichte

Beispiel: Dauer von Arbeitslosigkeit

<table>
<thead>
<tr>
<th><code>i</code></th>
<th><code>{x_{i-1}; x_i}</code></th>
<th><code>Δx_i</code></th>
<th><code>h_i</code></th>
<th><code>f_i</code></th>
<th><code>F_i</code></th>
<th><code>f_i^*</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0-1</td>
<td>1</td>
<td>19</td>
<td>0.16</td>
<td>0.16</td>
<td>0.160</td>
</tr>
<tr>
<td>2</td>
<td>1-2</td>
<td>1</td>
<td>12</td>
<td>0.10</td>
<td>0.26</td>
<td>0.100</td>
</tr>
<tr>
<td>3</td>
<td>2-3</td>
<td>1</td>
<td>24</td>
<td>0.20</td>
<td>0.46</td>
<td>0.200</td>
</tr>
<tr>
<td>4</td>
<td>3-6</td>
<td>3</td>
<td>28</td>
<td>0.23</td>
<td>0.69</td>
<td>0.077</td>
</tr>
<tr>
<td>5</td>
<td>6-12</td>
<td>6</td>
<td>31</td>
<td>0.26</td>
<td>0.95</td>
<td>0.043</td>
</tr>
<tr>
<td>6</td>
<td>12-24</td>
<td>12</td>
<td>6</td>
<td>0.05</td>
<td>1.00</td>
<td>0.004</td>
</tr>
<tr>
<td>Summe</td>
<td></td>
<td>120</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dichteverteilung
Dichteverteilung

![Histogramm der Links-/Rechts-Selbsteinstufung für Arbeiterklasse, Mittelschicht und Oberschicht](image-url)
Violin plot
Boxplot

- Outlier / Ausreißer
- Oberer Whisker
- $Q_3 + 1.5 \times IQR$
- Obere Quartilgrenze $Q_3$
- Arithmetisches Mittel
- Median
- Interquartilsrange $IQR = Q_3 - Q_1$
- Untere Quartilgrenze $Q_1$
- $Q_1 - 1.5 \times IQR$
- Unterer Whisker

Weitere Erläuterung
Boxplot

![Boxplot Image](image)

*Subjektive Klasse*

*Links-Rechts-Selbststeinstufung*

*Arbeiterklasse*  *Mittelschicht*  *Oberschicht*
Irreführende Grafiken: ABC News

Quelle: http://junkcharts.typepad.com/
Irreführende Grafiken: ABC News
Irreführende Grafiken: ABC News
Irreführende Grafiken: ABC News
Irreführende Grafiken: Tote durch Schusswaffen

**Gun deaths in Florida**

Number of murders committed using firearms

Source: Florida Department of Law Enforcement

Quelle: Ravi Parikh - Heap Analytics
Modellinterpretation

- Für das **Gesamtmodell** interessiert uns die **Modellgüte**. Diese lesen wir am **Bestimmtheitsmaß** $R^2$ ab.

- Für die **einzelnem Erklärungsfaktoren** des Modells prüfen wir:
  - Effektstärke (hier: *unstandardisierter Regressionskoeffizient*)
  - Richtung des Zusammenhangs
  - Signifikanz
lm(ptv.spd ~ europe.unification + left.right + gender, data=A) >>
summary()

## Call:
## lm(formula = ptv.spd ~ europe.unification + left.right + gender,
##    data = A)
##
## ## Residuals:
##    Min     1Q Median     3Q    Max
## -7.5280 -2.3478  0.0403  2.7652  6.1385
##
## ## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.22588    0.35116  20.577  < 2e-16 ***
## europe.unification  0.12773    0.03084   4.142   3.66e-05 ***
## left.right          -0.31746    0.04391  -7.229   8.08e-13 ***
## genderweiblich      0.21462    0.17726   1.211     0.226
##
## ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ## Residual standard error: 3.265 on 1355 degrees of freedom
## (289 observations deleted due to missingness)
## Multiple R-squared:  0.05526,  Adjusted R-squared:  0.05317
## F-statistic: 26.42 on 3 and 1355 DF,  p-value: < 2.2e-16
```
# Call:
# lm(formula = ptv.spd ~ europe.unification + left.right + gender, 
#     data = A)
# #
# # Residuals:    Min     1Q   Median     3Q    Max
# # -7.5280 -2.3478  0.0403  2.7652  6.1385
# #
# # Coefficients:   Estimate  Std. Error  t value  Pr(>|t|)
# # (Intercept)     7.22588   0.35116    20.577  < 2e-16 ***
# # europe.unification 0.12773   0.03084     4.142 3.66e-05 ***
# # left.right       -0.31746   0.04391    -7.229 8.08e-13 ***
# # genderweiblich   0.21462   0.17726     1.211   0.226
# #
# # Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# #
# # Residual standard error: 3.265 on 1355 degrees of freedom
# # (289 observations deleted due to missingness)
# # Multiple R-squared:  0.05526,  Adjusted R-squared:  0.05317
# # F-statistic: 26.42 on 3 and 1355 DF,  p-value: < 2.2e-16
```
Modellinterpretation

- **Beispiel:**
  - Das Modell erklärt die subj. Wahrscheinlichkeit SPD WählerIn zu sein. Die **Erklärungskraft** des Modells ist mit einem Bestimmtheitsmaß $R^2$ von 0.055 nur schwach, da nur ca. 5,5% der Varianz erklärt werden.
  - Die Zustimmung zu einer weiteren EU-Integration hängt **positiv** mit der Wahrscheinlichkeit SPD zu wählen zusammen. Mit jedem Punkt Zunahme auf der Skala EU-Integration nimmt die Wahrscheinlichkeit der SPD-Wahl um 0.12-Punkt zu. Dieser Zusammenhang ist **signifikant**.
  - Die Links/Rechts Skala hängt **negativ** mit der SPD-Wahl zusammen. Je linker eine Person eingestellt ist, desto wahrscheinlicher ist die SPD-Wahl. Mit jedem Skalenpunkt nach rechts nimmt die Wahrscheinlichkeit SPD zu wählen um 0.31-Punkte ab. Dieser Zusammenhang ist **signifikant**.
  - Der Effekt für Geschlecht ist **nicht signifikant**.
Modellbildung

![Graph showing the trend of time (in seconds) over time from 1900 to 2000, with data points for females and males.](image-url)
Erklärung durch den jährlichen Fortschritt

\texttt{lm(racetime \sim year, data = Swim)}
Erklärung durch das Geschlecht

```r
lm(racetime ~ sex, data = Swim)
```
Erklärung durch Jahr und Geschlecht

\texttt{lm(racetime ~ year + sex, data = Swim)}
Erklärung durch Jahr und Geschlecht, als *Interaktionsterm*

\[ \text{lm(racetime } \sim \text{ year + sex + year:sex, data = Swim)} \]
Erklärung durch Jahr Polynom

```r
lm(racetime ~ poly(year, 2), data = Swim)
```
Erklärung durch Jahr Polynom und Geschlecht, beide interagieren miteinander

```r
lm(racetime ~ sex:poly(year,2), data = Swim)
```