6. Untersuchung von Messdaten

Zweck

- Modellierung von Eingangsdaten für eine Simulation
- Anpassen von Ergebnisdaten an theoretische Verteilungen

Vorgehensweise

- Gegeben: Messreihe $X_1, \ldots, X_n$.
- Gesucht: eine passende theoretische Verteilung.
- Gründe für eine theoretische Verteilung:
  - Messdaten selbst nur eine „historische" Realisierung des Prozesses
  - Empirische Verteilungen: unregelmäßig, feste Spannweite, mehr Parameter
  - Physikalische oder modellimmanente Gründe

6.1 Modellierung empirischer Daten durch theoretische Verteilungen

- Parameterbestimmung und Verteilungsanpassung setzen unkorrelierte Daten voraus
  - Bestimmung empirischer Autokorrelationskoeffizienten ρ_j für j=1, …, k mit k≥10
  - Beachten:
    - selbst bei statistisch unabhängigen Werten ist $\rho_j \neq 0$,
    - Annahme hinreichender Unkorreliertheit, falls alle ρ_j deutlich kleiner als 0.1
- Parameterschätzung verschiedener Verteilungstypen mittels
  - Maximum-Likelihood-Estimators (MLEs) oder
  - Least-Squares-Methoden oder
  - Verfahren basierend auf empirischen Momenten
- Passende Verteilung aus diesen Kandidaten auswählen mittels
  - optischer Vergleiche
  - Anpassungstests (χ^2-Test, Kolmogorov-Smirnov-Test)

6.2 Maximum Likelihood Estimator (MLE, LK 6.5)

Verfahren zum Schätzen von Parametern einer Verteilung auf Basis einer gegebenen Stichprobe und einer hypothetischen Verteilung. Es gibt aber auch andere Verfahren, die bei den meisten aber nicht notwendigerweise bei allen Verteilungstypen dieselben Ergebnisse liefern:

- Least Squares Estimator
- Unbiased Estimators
- Methods of Moments

Maximum Likelihood Estimator
Gegeben: Stichprobenmenge $\{X_i: 0 \le i < n\}$
Definition: Likelihood-Funktion

$$L(\theta) = \begin{cases} \prod_{0 \le i < n} p_\theta(X_i) & \text{für Verteilungen} \\ \prod_{0 \le i < n} f_\theta(X_i) & \text{für Verteilungsdichtefunktionen} \end{cases}$$

dabei ist θ der für die Verteilung charakteristische Parameter und möglicherweise mehrdimensional.
Definition: Maximum Likelihood Estimator (MLE): $\hat{\theta} : L(\hat{\theta}) = \max_{\theta}(L(\theta))$

- Beispiel Exponentialverteilung: $\theta = \lambda$

$$L(\lambda) = \prod_{0 \le i < n} \lambda \cdot \exp(-\lambda \cdot X_i) = \lambda^n \cdot \exp\left(-\lambda \cdot \sum_{0 \le i < n} X_i\right)$$

$$\ln(L(\lambda)) = \ln\left(\lambda^n\right) - \lambda \cdot \sum_{0 \le i < n} X_i$$

Maximum – Ermittlung durch Differentiation und Gleichsetzung mit 0

$$\frac{d\ln(L(\lambda))}{d\lambda} = \frac{n}{\lambda} - \sum_{0 \le i < n} X_i \overset{!}{=} 0$$

$$\frac{1}{\lambda} = \frac{1}{n} \cdot \sum_{0 \le i < n} X_i$$

$$\Rightarrow \hat{\lambda} = \left(\overline{X}(n)\right)^{-1}$$

- Beispiel Geometrische Verteilung: θ=p

$$L(p) = \prod_{0\le i<n} p\cdot(1-p)^{X_i} = p^n\cdot(1-p)^{\sum X_i}{}_{0\le i<n}$$

$$\ln(L(p)) = \ln\left(p^n\right) + \left(\sum_{0\le i<n} X_i\right)\cdot\ln(1-p)$$

Maximum – Ermittlung durch Differentiation und Gleichsetzung mit 0

$$\frac{d\ln(L(p))}{dp} = \frac{n}{p} + \left(\sum_{0\le i<n} X_i\right)\cdot\frac{-1}{(1-p)} \overset{!}{=} 0$$

$$\frac{n}{p} = \frac{1}{(1-p)}\cdot\left(\sum_{0\le i<n} X_i\right)$$

$$(1-p) = p\cdot\frac{1}{n}\cdot\left(\sum_{0\le i<n} X_i\right) = p\cdot\overline{X}(n)$$

$$1 = p\cdot\left(1+\overline{X}(n)\right)$$

$$\Rightarrow \hat{p} = \frac{1}{1+\overline{X}(n)}$$

Die Berechnung der MLE für mehrdimensionale Parameter ist deutlich schwieriger und stimmt nicht immer mit der Methods of Moments überein.

- MLE für unterschiedliche Verteilungen in LK Kapitel 6

- Beispiel
  - Messdaten: 1 (4x), 2 (4x), 3 (7x)
  - Aus Modellannahme wissen wir: Verteilung muss eine Gerade sein
  - Betrachtung von 2 Möglichkeiten
    - A) P(X=1)=3/15, P(X=2)=5/15, P(X=3)=7/15
    - B) P(X=1)=4/15, P(X=2)=5/15, P(X=3)=6/15
  - Bewertung der beiden Optionen mittels Least Squares
    - A) $(1/15)^2+(1/15)^2+0^2=1/450$
    - B) $0^2+(1/15)^2+(1/15)^2=1/450$
    - ⇒ A) und B) sind gleich gut geeignet
  - Bewertung der beiden Optionen mittels Likelihood
    - A) (3/15)^4*(5/15)^4*(7/15)^7=9,52*10^-8
    - B) (4/15)^4*(5/15)^4*(6/15)^7=1,02*10^-7
    - ⇒ B) ist besser geeignet als A)

6.3 Verfahren für optische Vergleiche

6.3.1 Messwerthistogramm

- Konstruktion der Histogrammfunktion h(x)
  - Aufteilung des von den Werten überdeckten Bereichs in Intervalle; für k Intervalle und $b_0 \le \min_i (X_i)$, $b_k > \max_i (X_i)$:

    $[b_0, b_1), [b_1, b_2), \ldots, [b_{k-1}, b_k)$
  - am einfachsten: konstante Intervallbreite $\Delta b = b_j - b_{j-1}$
  - bei VF mit unbeschr. Definitionsbereich wird b_0 auf $-\infty$ bzw. b_k auf ∞ gesetzt
  - h_j: relative Häufigkeit der Werte $X_i \in [b_{j-1}, b_j)$
  - $$h(x) = \begin{cases} 0 & \text{für } x < b_0 \\ h_j / \Delta b & \text{für } b_{j-1} \le x < b_j \quad \text{mit} \quad j = 1, \ldots, k \\ 0 & \text{für } b_k \le x \end{cases}$$
  - Fläche unter dem Histogramm = 1
  - Problem: Bestimmung einer geeigneten Zahl k von Histogrammintervallen
    - für große Werte von k ist das Histogramm unregelmäßig
    - für kleine Werte von k besteht Informationsverlust

      $\Rightarrow$ mehrere Werte testen
  - Definition: kumulatives Histogramm bzw. empirische Verteilungsfunktion: $H(x) = \sum_{y \le x} h(y)$
- $f_j = P(b_{j-1} \le X < b_j) = \int_{b_{j-1}}^{b_j} f(x)dx \approx h_j$ für passende Dichtefunktion f(x)
  - optischer Vergleich von f(x) und h(x) zeigt Eignung von f(x)
- Auswahl bei mehreren Kandidaten: Vergleich der quadratischen Abstände der f_j- und h_j-Werte $\sum_{j=1}^{k} (f_j - h_j)^2$

6.3.2 Naive Ansätze zum Vergleich der empirischen und analytischen Verteilungsfunktion

- Vergleich des Messwerthistogramms und der theoretischen Verteilungsdichtefunktion bzw. der theoretischen Verteilung in einer Graphik
  - Nachteil: Übereinstimmung hängt stark von Intervallbreite ab.

- Vergleich des kumulativen Histogramms und der theoretischen Verteilungsfunktion in einer Graphik
  - Nachteil: Unterschiede sind bei steigenden oder fallenden Linien schwer zu erkennen.
- Betrachtung der Differenz zwischen kumulativem Histogramm und der theoretischen Verteilungsfunktion in einer Graphik
  - Nachteil aller 3 Methoden: Differenz zeigt nur absolute Abweichung auf dem gesamten Wertebereich
- Betrachtung der relativen Differenz bezogen auf die Verteilungsdichtefunktion
  - Nachteil: bei sehr kleinen Dichten wird der Fehler schnell extrem groß.
  - Nachteil bei allen Verfahren: Messpunkte können sehr stark geclustert sein.

6.3.3 PP-Plots (LK 6.6.1)

- Gegeben: Messreihe $X_1,\ldots,X_n$
- Erstelle aufsteigend sortierte Messreihe: $X_{(1)},\ldots,X_{(n)}$ (Ordnungsstatistik ?)
- Punkte für Graphik: $\left(\frac{j-0.5}{n}, F\left(X_{(j)}\right)\right)$, wobei F die theoretische Verteilungsfunktion darstellt. Die Anpassung „-0.5“ wird vorgenommen, , damit $X_{(n)}=\propto$ nicht notwendig ist, um die theoretische Verteilungsfunktion genau erscheinen zu lassen.
- Falls der Plot eine 45° Gerade ergibt, stimmen die empirische und die theoretische Verteilungsfunktion überein.
- Exakt anwendbar für kontinuierliche und diskrete Verteilungen

6.3.4 QQ-Plots (Quantile-Quantile-Plots, LK 6.6.1)

- Gegeben: Messreihe $X_1,\ldots,X_n$
- Erstelle aufsteigend sortierte Messreihe: $X_{(1)},\ldots,X_{(n)}$ (Ordnungsstatistik ?)
- Punkte für Graphik: $\left(X_{(j)}, F^{-1}\left(\frac{j-0.5}{n}\right)\right)$, wobei F die theoretische Verteilungsfunktion ist. Die Anpassung „-0.5“ wird vorgenommen, damit $X_{(n)}=\propto$ nicht notwendig ist, um die theoretische Verteilungsfunktion genau erscheinen zu lassen.
- Falls der Plot eine 45° Gerade ergibt, stimmen die empirische und die theoretische Verteilungsfunktion überein.
- Exakt nur auf kontinuierliche Verteilungen anwendbar, da die Umkehrfunktion F^{-1} für beliebige Werte benötigt wird.

6.3.5 Vergleich von PP- und QQ-Plots

- PP-Plots orientieren sich an der Wahrscheinlichkeitsdomäne, sie machen den Unterschied für alle Messpunkte gleich stark sichtbar ⇒ Teile des

Wertebereichs mit überdurchschnittlicher Verteilungsdichte werden vergrößert

- QQ-Plots orientieren sich an der Wertedomäne ⇒ Teile des Wertebereichs mit unterdurchschnittlicher Verteilungsdichte bzw. Wahrscheinlichkeit werden vergrößert.
- Verteilungsfunktionen sind oft S-förmig, da die entsprechenden Verteilungsdichtefunktionen oder Verteilungen im Innern des Wertebereichs Maxima haben.
    - PP-Plots machen Unterschiede im Innern des Wertebereichs deutlicher sichtbar.
    - QQ-Plots machen Unterschiede am Rand des Wertebereichs deutlicher sichtbar.

6.4 Vorgehensweise in der Praxis

1. Berechnung von mehreren Histogrammen mit unterschiedlicher Intervallanzahl.
2. Bestimmung eines geeigneten Histogramms.
3. Bestimmung einer Auswahl von Verteilungen, die zu dieser Histogrammform passen.
4. Berechnung der Parameter dieser Verteilungen.
5. Optischer Vergleich und Berechnung des quadratischen Abstandes.
6. Durchführung der Anpassungstests
7. Auswahl der Verteilung die sowohl bei 5. als auch bei 6. gut abschneidet.

Bemerkung: In der Literatur findet man weitere optische und formale Verfahren zur Überprüfung der Anpassungsgüte einer theoretischen Verteilung an gegebene Daten (z.B. QQ-Plots, PP-Plots, Anderson-Darling-Test). Diese Tests werden durch Tools wie z.B. Expertfit unterstützt

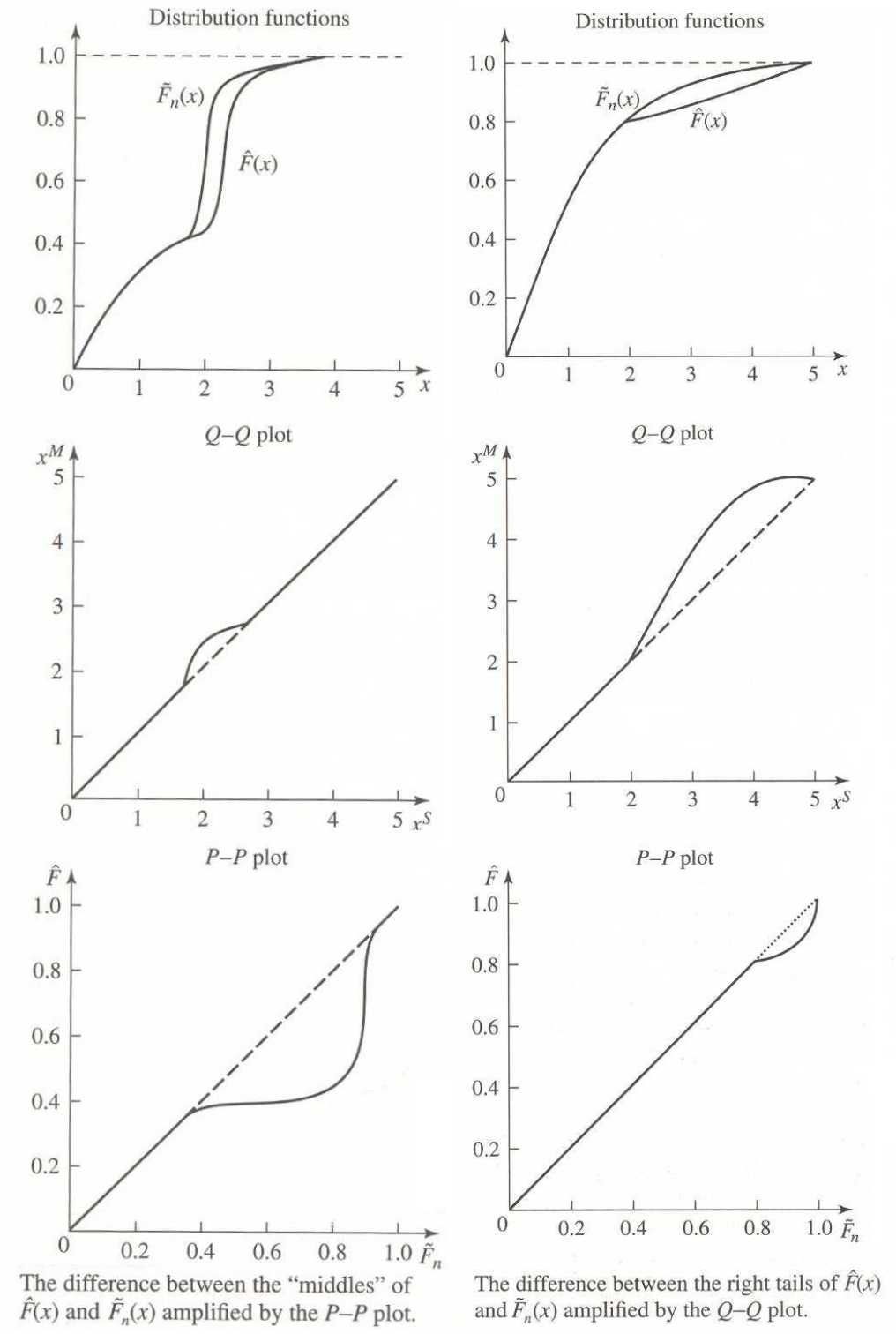

Abbildung 3.5: Vergleich von QQ- und PP-Plots. (aus Law/Kelton: "Simulation Modeling and Analysis", 3rd Edition, S. 353 und 354)