

5. Spezifikation quantitativer Modellgrößen

Modellbildung identifiziert

- Eigenschaften Ersatzsystem (auf Basis mentalen Modells, in gewählter Abstraktion)

Eingeschlossen sind im Modell zu repräsentierende

- strukturelle und qualitative Eigenschaften (Objekte, Prozesse, Regeln, Relationen, ...)
- quantitative Eigenschaften (numerische Größen): **konkrete Werte / Wertverläufe** zu spezifizieren

Klassifikation quantitativer Eigenschaften (EmSi70):

- **exogene** (numerische) Faktoren; können (im Prinzip) verschiedene Werte haben, beeinflussen Modellverhalten, werden von Modellverhalten nicht beeinflusst (falls für eine Untersuchung/Analyse fest: "Konstante", "Parameter", "statische Größe" sonst: "Variable" (im PS-Sinn), "dynamische Größe")
- **endogene** (numerische) Faktoren; werden durch Modellverhalten (potentiell) beeinflusst

Erinnerung an Kap.1:

Unterscheidung

- kontrollierbare Größen: "willkürlich" einstellbar (für "was-wenn"-Fragen, zur Suche "optimale Güte")
- unkontrollierbare Größen: als "unbeeinflussbar" angesehen ("was-wenn"-Fragen)

Um Simulation ablaufen lassen zu können, müssen exogenen

(kontrollierbaren oder nicht kontrollierbaren)

Größen konkrete Werte / Werteverläufe zugewiesen sein

(Nachteil Simulation:

formale "Parameter" können nicht durchgetragen werden)

Woher Werte / Werteverläufe ??

3 prinzipielle Quellen:

- Theorie
- reale Welt
- Hypothesen (Annahmen)

Folgt Übersicht über

- exogene / endogene
- kontrollierbare / nicht kontrollierbare
- statische / dynamische
- deterministische / stochastische numerische Modellgrößen

und daraus resultierender "Bedarf an Daten"

Simulation

Spezifikation Modellgrößen

be/ja/2

5 - 3(4)

Simulation

Spezifikation Modellgrößen

be/ja/2

5 - 4(4)

	kontrollierbar		nicht kontrollierbar	
	statisch	dynamisch	statisch	dynamisch
exogen	Annahmen treffen		aus Theorie entnehmen oder beobachten: messen, schätzen	
	deterministisch:			
	Wert	Fkt. der Zeit	Wert	Fkt. der Zeit
endogen	bei festliegender Aufgabenstellung nicht existent		ergeben sich aus Analyse	
	deterministisch:			
	stochastisch:			
	Verteilung	stoch. Proz.	Verteilung	stoch. Proz.
			Resultat-	
			Wert	Fkt. der Zeit
			Resultat-	
			Wert-	Wertverlauf-
			menge (Stichprobe)	menge (Stichprob.-menge)

 Bedarf an Daten:
 exogen: für Simulation selbst
 endogen: potentiell zur "retrospektiven", "historischen" Validierung

Abbildung 5.0.1: Klassifikation von "Faktoren" und Bedarf an Daten

Für Modell insbesondere benötigt:

Bestimmung / Darstellung / Charakterisierung / "Modellierung"

nicht kontrollierbarer, exogener Faktoren

Fälle:

- (i) statisch / deterministisch: Wert
 - aus Theorie
 - aus Messung realer Welt (System / Umgebung)
- (ii) statisch / stochastisch: Verteilung
 - aus Theorie
 - aus Messung realer Welt (System / Umgebung) Realisierungen (Stichprobe) -> Verteilung
- (iii) dynamisch / deterministisch: Zeitfunktion
 - aus Theorie
 - aus Messung realer Welt (System / Umgebung)
- (iv) dynamisch / stochastisch: stochastischer Prozeß
 - aus Theorie
 - aus Messung realer Welt (System / Umgebung) Realisierungen (Zeitreihen) -> Generierungsgesetze

(i,iii) Vorgehen bekannt

(ii) im folgenden genauer betrachtet

(iv) schwieriger, meist auf (ii) zurückgeführt auch: mit "traces" behandelt

5.1 Modellierung von Zufallsvariablen

Statisch-stochastische, numerische (exogene) Faktoren
 - im Modell durch ZV beschrieben
 - (zugehörige) Verteilung benötigt, um (während Simulation) Realisierungen zu ziehen (s. 3.3)

Notwendige Modellierungsschritte:

- (i) Identifikation Verteilungstyp
- (ii) (u.U.) Schätzung Verteilungsparameter

Mögliche (typische) Situationen:

- (a) Verteilungstyp aus Theorie, Parameterschätzung aus real vorliegender Stichprobe
- (b) Empirische Verteilung aus Stichprobe, direkt zur Generierung Realisierungen verwendet
- (c) Verteilungstyp aus Stichprobe identifiziert, Parameterschätzung aus Stichprobe
- (d) Weder theoretische Hinweise noch reale Stichprobe vorhanden, lediglich "Charakter: ZV" feststehend

Literaturhinweis: LaKe82/...

Folgend alles für kontinuierliche Faktoren (als Bsp.), diskrete Faktoren: s. Literatur

- subjektiv Grenzen u, o Modalwert w Mittelwert m bekannt / ermittelbar

Vorschlag: **Beta-Verteilung** "beta(1, 2)"

Für u=0, o=1 (Translation, Dehnung/Stauchung möglich)

$$f_Y(y) = \begin{cases} y^{1-1} (1-y)^{2-1} / B(1, 2) & 0 < y < 1 \\ 0 & \text{sonst} \end{cases}$$

$$E[Y] = \frac{1}{1+2} \quad (= m)$$

$$M[Y] = \frac{1-1}{1+2-2} \quad (= w)$$

1, 2 > 1, aus m und w zu bestimmen, B(...) "Beta-Funktion":

$$B(1, 2) := \int_0^1 t^{1-1} (1-t)^{2-1} dt$$

zu (d):

- sehr "ungemütliche" Situation !
- u.U. Feld für (strittige!) "maximum entropy"-Verfahren
- im Anschluß: einige heuristische Tips zu Verteilung und Parametern einer zu modellierenden ZV Y

Alternative Situationen:

- subjektive Grenzen u, o bekannt / ermittelbar, derart, daß P[Y<u] P[Y>o] 0
 Vorschlag: **Gleichverteilung**

$$F_Y(y) = \begin{cases} 0 & y < u \\ (y-u)/(o-u) & u < y < o \\ 1 & y > o \end{cases}$$

$$E[Y] = \frac{u+o}{2}$$

M[Y] y-Wert maximaler Dichte (Modalwert, mode) nicht eindeutig

- subjektive Grenzen u, o und subjektiver Modalwert w bekannt / ermittelbar
 Vorschlag: **Dreiecksverteilung**

$$F_Y(y) = \begin{cases} 0 & y < u \\ \frac{(y-u)^2}{(o-u)(w-u)} & u < y < w \\ 1 - \frac{(o-y)^2}{(o-u)(o-w)} & w < y < o \\ 1 & y > o \end{cases}$$

$$E[Y] = \frac{u+w+o}{3} \quad M[Y] = w$$

zu (a):

Einige Hinweise zu Verteilungsformen auf Basis theoretischer Überlegungen (s. Mihr72, LaKe82)

- ZV Y, welche Summe einer größeren Anzahl zufälliger Einflüsse darstellt, könnte (zentraler Grenzwertsatz) **normalverteilt** sein N(μ, σ²)-verteilt

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

$$E[Y] = \mu$$

$$VAR[Y] = \sigma^2$$

Beispiele: "natürliche" Faktoren wie Größe Lebewesen, aber: auf y<0 achten ("truncation")

- ZV Y, welche Minimum einer größeren Anzahl zufälliger Einflüsse darstellt, könnte (Grenzwertsatz) **Weibull-verteilt** sein

$$F_Y(y) = \begin{cases} 1 - \exp\left(-\left(\frac{y}{\lambda}\right)^k\right) & y > 0 \\ 0 & y < 0 \end{cases}$$

Beispiele: "time between failures" komplexen technischen Systems

- ZV Y, welche zeitliche Abstände zwischen aufeinanderfolgenden Ereignissen darstellt, wo anzunehmen, daß Ereignisse
 - einzeln auftreten
 - in (kleinerem) Zeitintervall der Länge t (bei beliebiger Lage solchen Zeitintervalls) mit fester Wahrscheinlichkeit $\lambda \cdot t$ auftreten (wo positive "charakterisierende" Konstante)
 - insgesamt mit konstanter "Rate" auftreten könnte (Satz über "seltene Ereignisse") (negativ) **exponentiell verteilt** sein

$$FY(y) = \begin{cases} 1 - \exp(-\lambda y) & y > 0 \\ 0 & y \leq 0 \end{cases}$$

Damit (gleichwertig:) Zahl Ereignisse in beliebigem Zeitintervall **Poisson-verteilt**

Beispiele: Ereignisse, die zu "absolut zufälligem" Zeitpunkt von jeweils einem Mitglied großer Gesamtheit ausgelöst werden (etwa: an Vermittlung eintreffende Telefonanrufe)

- ZV Y, welche Produkt einer größeren Anzahl zufälliger Einflüsse darstellt, könnte (Grenzwertsatz) **log-normal verteilt** sein, LN(μ, σ^2)-verteilt

$$fY(y) = \begin{cases} \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right\} & y > 0 \\ 0 & y \leq 0 \end{cases}$$

Insgesamt zu (a):

- Werte Verteilungsparameter?
- "wie gesagt", aus Stichprobe "schätzen" vgl. Abschnitt 5.2

Simulation

Spezifikation Modellgrößen

be/ja/2

5 - 11(6)

zu (b): Explizite Bestimmung "Verteilungstyp" vermieden, stattdessen Verteilungsfunktion aus Stichprobe "direkt" geschätzt

Einschub:
 In "verteilungsfreier" / "nichtparametrischer" Statistik häufig Verwendung von "Ordnungsstatistiken" (order statistics):
 Ordnungst Statistik einer Stichprobe (Umfang n) ist Folge $(Y_{(1)}, Y_{(2)}, \dots, Y_{(n)})$ der größtmäßig aufsteigend geordneten Stichprobenvariablen Y_1, Y_2, \dots, Y_n
 Folge $(y_{(1)}, y_{(2)}, \dots, y_{(n)})$ der größtmäßig aufsteigend geordneten Stichprobenwerte y_1, y_2, \dots, y_n ist Realisierung Ordnungst Statistik

"Empirische Verteilungsfunktion" $F^*Y(y)$ wird aus Stichprobe (y_1, y_2, \dots, y_n) gewonnen:

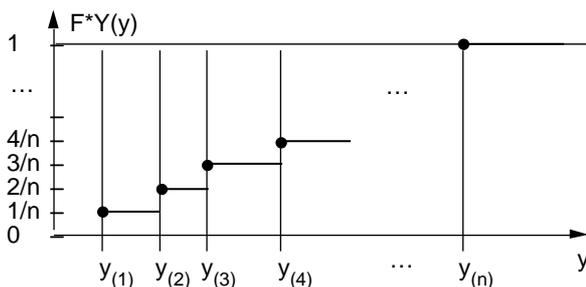


Abbildung 5.1.1: empirische Verteilungsfunktion

Simulation

Spezifikation Modellgrößen

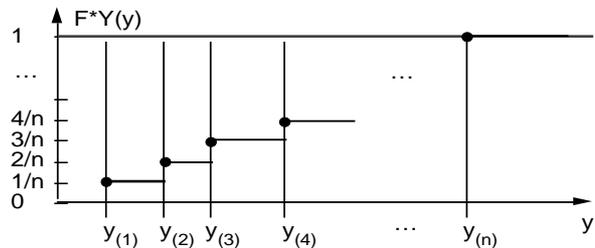
Simulation

Spezifikation Modellgrößen

be/ja/2

5 - 12(6)

Empirische Verteilungsfunktion



schätzt Verteilungsfunktion erwartungstreu:

- aus Zeichnung:

$$\tilde{F}Y(y) = \frac{|\{\text{Realisierungen von } Y \mid y \mid n\text{-Stichprobe}\}|}{n} = \sum_{y \leq y} I_{Y=y} / n \quad \text{s. Zeichnung}$$

wegen Unabhängigkeit

- wo "Indikatorfunktion":

$$I := \begin{cases} 1 & \text{wahr} \\ 0 & \text{falsch} \end{cases}$$

- und damit:

$$E[\tilde{F}Y(y)] = \frac{1}{n} \sum_{y \leq y} E[I_{Y=y}] = P[Y \leq y] = FY(y)$$

Simulation

Spezifikation Modellgrößen

Verwendung während Simulation:
 Generierung von Realisierungen von Y gemäß $F^*Y(y)$

Wieder Anleihe bei "inverse Transformation" (Abschn. 3.3):
 bei $[0,1)$ -gleichverteiltem U (+ Voraussetzungen)
 liefert $FY^{-1}(U)$ gemäß FY verteilte ZV

Als Rezept:
 Ziehe u, dann ist jenes $y_{(i)}$ die gesuchte Realisierung,
 für das gilt $F^*Y(y_{(i-1)}) < u < F^*Y(y_{(i)})$

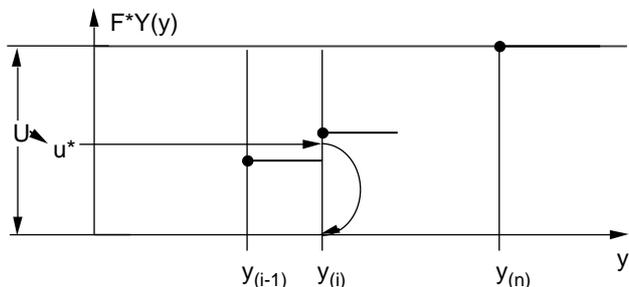


Abbildung 5.1.2: "Ziehen" aus empirischer Verteilungsfunktion

Bzw. einfacher (identisches Ergebnis !):
 alle Treppenstufen gleich hoch ($1/n$),
 damit wegen gleichverteiltem U
 - alle Werte $y_{(1)}, y_{(2)}, \dots, y_{(n)}$
 - bzw. alle Werte y_1, y_2, \dots, y_n
 mit gleicher Wahrscheinlichkeit $1/n$ zu wählen

praktisch:
 Ziehung aus diskret $[1,n]$ -gleichverteiltem U'
 liefert direkt Position des gesuchten Wertes
 in einem n-Feld mit Werten $y_{(1)}, y_{(2)}, \dots, y_{(n)}$ bzw. y_1, y_2, \dots, y_n

Gelegentlich empfohlen:
 lineare Interpolationen (auch Int. höherer Ordnung)
 zur Definition von F^*Y ;
 klingt plausibel, zerstört aber Erwartungstreue!

Numerisch u.U.
 zwei oder mehr Werte der Stichprobe identisch:
 - Treppenstufen entsprechend höher,
 - bzw. beide (alle) Werte ins n-Feld

Simulation

be/ja/2

Spezifikation Modellgrößen

5 - 15(6)

zu (c):

- Keine theoretisch fundierte Hypothese für Verteilungstyp, wohl aber Stichprobe verfügbar
- Versuch, Verteilungstyp aus Stichprobe zu "erahnen", zu "identifizieren"
- Auf diesem Weg erste Hinweise auf Ausschluß bestimmter Verteilungstypen auf Basis von (auf Momenten beruhenden) Verteilungscharakteristika

so etwa des **Variationskoeffizienten** $VK[Y]$

$$VK[Y] := \frac{\sqrt{VAR[Y]}}{E[Y]}$$

Bekannt sind Schätzer für Erwartungswert und Varianz (s. Abschn. 4.1):

$$\tilde{\mu}_1 = \frac{1}{n} \sum_{i=1}^n Y_i$$

$$\tilde{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \tilde{\mu}_1)^2$$

Natürlicher (nicht erwartungstreuer) Schätzer für VK wäre:

$$\tilde{VK} := \frac{\sqrt{\tilde{\sigma}^2}}{\tilde{\mu}_1}$$

Da gewisse Verteilungstypen VK's nur bestimmter Wertebereiche zulassen (z.B.: $<1, >1$), kann Schätzwert * zum Ausschluß ganzer Verteilungsfamilien ausreichen

- Subjektive ("visuell" geführte) Identifizierung Verteilungstyp mittels **Histogrammen** (Schätzer Dichtefunktion) (Verteilungsfunktion "visuell" schlecht einsetzbar)

Simulation

Spezifikation Modellgrößen

Simulation

be/ja/2

Spezifikation Modellgrößen

5 - 16(5)

Weg:

- vorhanden: Stichprobe (y_1, y_2, \dots, y_n)
- bilde "Klassen" von Wertebereichen $[b_0, b_1), [b_1, b_2), \dots, [b_{k-1}, b_k)$ gleicher Breite $b = b_j - b_{j-1} \quad j=1, 2, \dots, k$

bestimme relative Häufigkeiten

$$r_j := \frac{\#y_i\text{'s in } [b_{j-1}, b_j)}{n} \quad j = 1, 2, \dots, k$$

und zeichne "Histogramm":

$$HY(y) := \begin{cases} 0 & y < b_0 \\ r_j & b_{j-1} \leq y < b_j \\ 0 & y \geq b_k \end{cases}$$

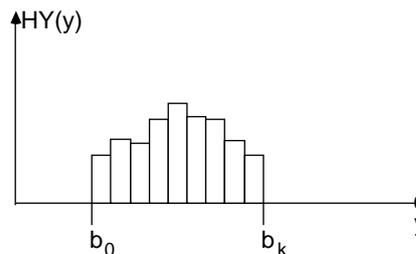


Abbildung 5.1.3: Histogramm

Simulation

Spezifikation Modellgrößen

- Histogramm:
 - Aussehen sehr abhängig von b und Lage b_0 , mehrere Alternativen "probieren"!
 - alle Intervalle sollten "hinreichende" Anzahl enthalten, Rat: > 5 ??
 - $[b_0, b_k)$ muß nicht alle y_i enthalten, "outlier" u.U. vernachlässigen
- Histogramm $HY(y)$ schätzt Dichte $fY(y)$ (bis auf Konstante):
 - einerseits ist

$$P[b_{j-1} < Y < b_j] = \int_{b_{j-1}}^{b_j} fY(y) dy = b fY(y'_j)$$
 mit (Mittelwertsatz:) einem $y'_j \in [b_{j-1}, b_j)$
 - andererseits ist für $y' \in [b_{j-1}, b_j)$

$$HY(y') = r_j$$
 mit

$$E[R_j] = P[b_{j-1} < Y < b_j]$$
 - so daß

$$r_j \approx b \cdot fY(y'_j)$$
 und HY zu fY ungefähr proportional
 - und somit $HY(y)/b$ Schätzer $\hat{f}Y(y)$ für Dichte $fY(y)$
- Aus dem Bild (Histogramm) nun (subjektiv und erfahrungsabhängig) Typ einer u.U. zugrundeliegenden Verteilung identifizieren

Jede (analytisch charakterisierte) Verteilungsfamilie weist gewisse Parameter auf, hier erfaßt durch Parametervektor θ (z.B.: $N: \mu, \sigma$, $EXP: \lambda$)
 Werte dieser Parameter (nach Identifizierung V-Familie) noch zu bestimmen (schätzen) - analog Situation a) - Diesbezügliche Verfahren s. Abschn. 5.2

u.U. kann schon vor Parameterschätzung Hypothese V-Typ verworfen werden; so bei V-Familien, deren Mitglieder ausschließlich über
 - Translation
 - Dehnung / Stauchung
 auseinander hervorgehen; z.B.: N

Dazu (wieder) "visuelle" Methode: **probability plot**

- Grundidee: Vergleich Quantile zweier Verteilungen
- dabei q -Quantil y_q ($0 < q < 1$) der Verteilung einer ZV Y definiert über

$$FY(y_q) = q$$
 d.h. bei kontinuierlichem, streng monotonem FY

$$y_q := FY^{-1}(q)$$
- zwei ZV X, Y genau dann identisch verteilt, wenn all ihre Quantile x_q, y_q übereinstimmen;

Graph x_q versus y_q ist dann (Ursprungs-) Gerade (mit Steigung 1)

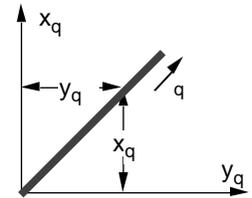


Abbildung 5.1.4:

Simulation

Spezifikation Modellgrößen

be/ja/2

5 - 19(6)

- hängen X und Y über lineare Transformation

$$X = a \cdot Y + b \quad a > 0$$
 zusammen, dann

$$FX(x) = P[X \leq x] = P[aY + b \leq x] = P\left[Y \leq \frac{x-b}{a}\right] = FY\left(\frac{x-b}{a}\right)$$
 und Graph x_q versus y_q ist Gerade (aus der sogar a und b abschätzbar)

- konkrete Aufgabe hier: visueller Vergleich
 - empirische Verteilungsfunktion $F^*Y(y)$ aus Stichprobe
 - hypothetische, analytische Verteilungsfunktion $FY(y)$ aus visueller Identifikation
 auf Basis jeweiliger Quantile
 - y^*_q für $F^*Y(y)$
 - y_q für $FY(y)$

Falls Hypothese: $F^*Y(y) = FY(y)$ zutreffend, sollte Graph y_q versus y^*_q (= "probability plot")

Ursprungsgerade der Steigung 1 ähneln

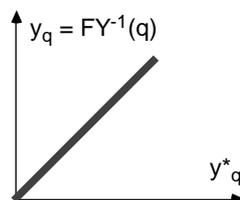


Abbildung 5.1.5: probability plot

Simulation

Spezifikation Modellgrößen

Simulation

Spezifikation Modellgrößen

be/ja/2

5 - 20(6)

- für konkrete Aufgabe ist (entsprechend Annahmen)
 - i/n -Quantil der Y^* -Vert'g gleich $y_{(i)}$ (Abszisse)
 - i/n -Quantil der Y -Vert'g gleich $FY^{-1}(i/n)$ (Ordinate)

Zu inspizierender Graph hat Aussehen:

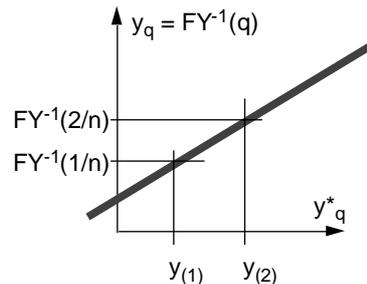


Abbildung 5.1.6: probability plot als Beurteilungsgraph

- danach
 - Hypothese $F^*Y(y) = FY(y)$ nicht abzulehnen, falls Punkte "so ungefähr" auf Winkelhalbierender
 - Hypothese $F^*Y(y)$ und $FY(y)$ nur durch Translationsparameter a und Dehnungsparameter b unterschieden nicht abzulehnen, falls Punkte "so ungefähr" auf Gerade
- sonst: Hypothesen ablehnen, "Ahnung $FY(y)$ war falsch"

Simulation

Spezifikation Modellgrößen

5.2 Schätzung von Verteilungsparametern

Sei Y kontinuierliche Zufallsvariable,
(für diskrete ZV: s. Literatur)

- deren Verteilungstyp bekannt / identifiziert sei (zumindest im Sinne "wohlbegründeter" Hypothese, s. Abschn. 5.1), etwa durch funktionale Form ihrer Dichte

$$f_Y(y; \underline{\theta})$$

- deren Parameter $\underline{\theta} = (\theta_1, \theta_2, \dots, \theta_p)$ aber nicht bekannt, also zu bestimmen sind, zu schätzen sind aus vorliegender Y-Stichprobe (Y_1, Y_2, \dots, Y_n)

Für Aufgabe verfügbar diverse statistische Verfahren, hier vorgestellt:

- Momentenmethode
- maximum likelihood Methode

Simulation

be/ja/2

Spezifikation Modellgrößen

5 - 23(6)

Beispiel: Exponentialverteilung

- Dichtefunktion:

$$f_Y(y; \lambda) = \begin{cases} \exp(-\lambda y) & y \geq 0 \\ 0 & y < 0 \end{cases}$$

- erstes Moment:

$$\mu_1 = \int_0^\infty y \cdot e^{-\lambda y} dy$$

partielle Integration

$$= \left[-y \frac{1}{\lambda} e^{-\lambda y} + \left(\frac{1}{\lambda} \right) \int e^{-\lambda y} dy \right]_0^\infty$$

$$= \left[-y e^{-\lambda y} + \left(\frac{1}{\lambda} \right) e^{-\lambda y} \right]_0^\infty$$

$$= \left[-e^{-\lambda y} \left(y + \frac{1}{\lambda} \right) \right]_0^\infty$$

$$\mu_1 = \frac{1}{\lambda}$$

- (folglich) Zusammenhang (erstes) Moment (μ_1) / (einziger) Parameter (λ): $\lambda = 1/\mu_1$
- Parameterschätzer, mit Momentenschätzer (4.1.4):

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n Y_i}$$

Simulation

Spezifikation Modellgrößen

Momentenmethode

- bedient sich der k-ten Momente von Y

$$\mu_k := E[Y^k]$$

bzw. ihrer erwartungstreuen Schätzer (s. Abschn. 4.1)

$$\tilde{\mu}_k := \frac{1}{n} \sum_{i=1}^n Y_i^k$$

bzw. diesbezüglicher Schätzwerte μ_k^*

- Parameter θ_j lassen sich oft ausdrücken als Funktionen der Momente

$$\theta_j = \theta_j(\mu_1, \mu_2, \dots, \mu_p) \quad j = 1, 2, \dots, p$$

- Momentenmethode
 - substituiert (in diesen Funktionen) Momentenschätzer für Momente
 - gewinnt so **Parameterschätzer**

$$\tilde{\theta}_j = \theta_j(\tilde{\mu}_1, \tilde{\mu}_2, \dots, \tilde{\mu}_p) \quad j = 1, 2, \dots, p$$

- gewonnene Schätzer nicht notwendig erwartungstreu, meist asymptotisch erwartungstreu und konsistent

(Erinnerung: $\tilde{\theta}$ Schätzer für Größe θ heißt
 - erwartungstreu, wenn $E[\tilde{\theta}] = \theta$
 - asymptotisch erwartungstreu, wenn $\lim_{n \rightarrow \infty} E[\tilde{\theta}] = \theta$
 - konsistent, wenn $\lim_{n \rightarrow \infty} P[|\tilde{\theta} - \theta| > \epsilon] = 0$ für jedes $\epsilon > 0$)
 oft dennoch keine "guten" Schätzer ("Form" Verteilung)

Simulation

be/ja/2

Spezifikation Modellgrößen

5 - 24(6)

maximum likelihood Methode

- Erinnerung:
 - Parameter $\underline{\theta}$ der Verteilung einer ZV Y, deren Typ bekannt, z.B. als Dichte $f_Y(y; \underline{\theta})$, zu schätzen aus Stichprobe $(y_1^*, y_2^*, \dots, y_n^*)$ (* für Stichprobenwerte zur Unterscheidung, wo nötig)

- Stichprobe zu sehen als Realisierung einer mehrdimensionalen ZV $\underline{Y} := (Y_1, Y_2, \dots, Y_n)$

- alle Stichprobenvariablen identisch verteilt:
 $f_{Y_i}(y_i; \underline{\theta}) = f_Y(y_i; \underline{\theta}) \quad i = 1, 2, \dots, n$

- alle Stichprobenvariablen unabhängig verteilt:

$$f_{\underline{Y}}(\underline{y}; \underline{\theta}) = \prod_{i=1}^n f_{Y_i}(y_i; \underline{\theta})$$

wo $\underline{y} := (y_1, y_2, \dots, y_n)$

- Idee der Methode:

Parametervektor $\underline{\theta}$ so wählen (bestimmen), daß Beobachtung $(y_1^*, y_2^*, \dots, y_n^*)$ in den Punkt maximaler Dichte,

zu liegen kommt
 - Maximum von $f_{\underline{Y}}(\underline{y}; \underline{\theta})$

Motivation: Umgebung dieses Punktes ist Bereich größter Beobachtungswahrscheinlichkeit

Simulation

Spezifikation Modellgrößen

- Weg:
 - Maximierung der **maximum-likelihood-Funktion**

$$L(\underline{\theta}) := \prod_{i=1}^n f_Y(y_i; \underline{\theta})$$
 bezüglich Komponenten des Vektors $\underline{\theta}$
 - notwendige Bedingung für Maximum mehrdimensionaler, differenzierbarer Funktion (sei für gemeinsame Dichte $L(\underline{\theta})$ vorausgesetzt) ist

$$\left(\frac{\partial L}{\partial \theta_j} \right) = (0)$$
 woraus p Bestimmungsgleichungen für die

$$j = 1, 2, \dots, p$$
 folgen
 - u.U. muß explizit auf Vorliegen Maximum / Minimum / Sattelpunkt geprüft werden
 - diese Beziehungen

$$\tilde{\theta}_j := l_j(\underline{y}) \quad j = 1, 2, \dots, p$$
 liefern, nach Substitution des Stichprobenvektors \underline{y}^* für den Variablenvektor \underline{y} die gesuchten **maximum-likelihood-Schätzwerte**

$$j^* := l_j(\underline{y}^*) \quad j = 1, 2, \dots, p$$

Beispiel: (wieder) Exponentialverteilung

- Dichtefunktion:

$$f_Y(y; \theta) = \begin{cases} \exp(-y/\theta) & y \geq 0 \\ 0 & y < 0 \end{cases}$$
- log-likelihood Funktion:

$$\log L = \sum_{i=1}^n \log \left(e^{-y_i/\theta} \right)$$
- partiell differenziert (einziger Parameter ist θ)

$$\frac{\partial \log L}{\partial \theta} = \sum_{i=1}^n \frac{(-y_i) e^{-y_i/\theta} + e^{-y_i/\theta}}{\theta^2} = \sum_{i=1}^n \frac{1 - y_i/\theta}{\theta^2}$$
- Maximumbedingung:

$$\frac{\partial \log L}{\partial \theta} = 0 \quad n - \sum_{i=1}^n y_i = 0$$
- daraus Parameterschätzer und Parameterschätzwert

$$\tilde{\theta} = n / \sum_{i=1}^n y_i \quad \theta^* = n / \sum_{i=1}^n y_i^*$$
- es ist "purer Zufall", daß in diesem Fall Momentenschätzer und **MLE** (maximum likelihood estimator) übereinstimmen

- praktische Anwendung
 - Logarithmus $\log(L(\underline{\theta}))$ der likelihood-Funktion, sog. **log-likelihood-Funktion**

$$\log(L(\underline{\theta})) = \log \left(\prod_{i=1}^n f_Y(y_i; \underline{\theta}) \right) = \sum_{i=1}^n \log(f_Y(y_i; \underline{\theta}))$$
 hat wegen Monotonie der Logarithmus-Funktion Maximum an derselben Stelle wie likelihood-Fkt.
 - $\log(L(\underline{\theta}))$ wird zur Bestimmung des Maximums, folgend der Schätzer $\tilde{\theta}$ wegen "leichterer" Differenzierbarkeit (Summe!) gern anstelle $L(\underline{\theta})$ verwendet
 - ist das Gleichungssystem zur Bestimmung der Schätzer $\tilde{\theta}$ nicht explizit lösbar, kann Maximum auch auf numerischem Weg ermittelt werden
- maximum-likelihood-Schätzer
 - sind (erneut) nicht notwendig erwartungstreu, aber meist konsistent
 - besitzen zusätzlich (unter gewissen Voraussetzungen) die (wünschenswerte) Eigenschaft minimal möglicher (asymptotischer) Varianz (vgl. Fish73, Mihr72, LaKe82)

5.3 Überprüfung der Paßgüte angepaßter Verteilungen

- Modellierung ZV umfaßte (falls analytische Form gefragt)
- Identifizieren Verteilungstyp (aus Theorie, durch intelligentes "Raten",...)
 - Schätzen Verteilungsparameter (aus Stichprobe, s. Abschn. 5.2)

- Man sollte
- sich vergewissern, ob (und wie "gut") gefundene Verteilungsform (samt Parameterwerten) mit Basisdaten übereinstimmt
 - entscheiden können, welche von u.U. mehreren Alternativen aus (i,ii) im Hinblick auf "Paßgüte" vorzuziehen

Beispielsweise aus vorliegender Stichprobe (z.B. über Histogramm) für "CPU-Anforderungen von jobs (in Zeiteinheiten)"

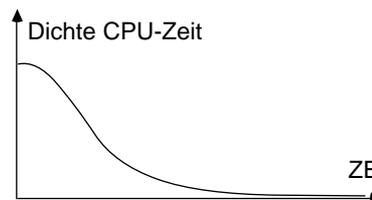


Abbildung 5.3.1: "erahnte" Dichte CPU-Zeit-Verteilung

Frage:
Exponentialverteilung oder Hyperexponentialverteilung oder Hypoexponentialverteilung oder COX-Verteilung oder Weibull-Verteilung wählen ??

Fragestellung in statistischer Form:

Liege vor: - Verteilungsdichte in funktionaler Form
(hier: identifizierte, angepaßte Dichtefunktion)
- Stichprobe (hier: Beobachtungsdaten)

Hypothese: Stichprobe ist aus dieser Verteilung "gezogen"

Frage: Muß Hypothese **verworfen** werden?
Oder könnte sie, alternativ dazu,
angenommen werden?

Beantwortung solcher Fragen durch statistische
Anpassungstests / goodness-of-fit Tests

zwei konkrete Tests in Abschn. 5.3.1, 5.3.2

Zuvor:

Hinweise zu "statistischen Tests"

- Ziel ist (hier), hypothetische Aussagen über Verteilung einer ZV zu überprüfen anhand vorliegender Stichprobe (y_1, y_2, \dots, y_n)
Beispiele: $E[Y] > m$,
 $F_Y(y)$ ist Verteilungsfunktion
(dies wäre "Anpassungstest")
- Zu überprüfende Aussage bezeichnet als **Nullhypothese** H_0
dazu alternative Aussage als **Alternativhypothese** H_1 (= not H_0)
- Überprüfung anhand Stichprobe kann, wegen deren "statistischen Schwankens", immer zu falschen Folgerungen führen; zwei Typen von "Fehlern" zu unterscheiden:

Simulation

Spezifikation Modellgrößen

be/ja/2

5 - 31(6)

Zur Anwendung erforderlich:

- Bestimmung der Verteilung von S unter der Voraussetzung: H_0 zutreffend
- Ermittlung von "kritischen Werten" c (bzw. c_{1-} : Vorsicht in Tafeln !)
ab denen H_0 zum Niveau α zu verwerfen

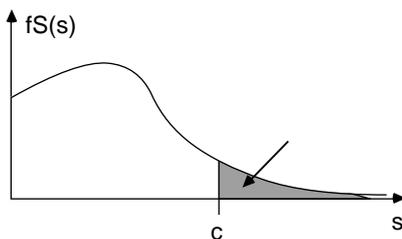


Abbildung 5.3.2: Prinzipskizze statistische Tests

nochmals:

Bestimmung c derart, daß $P[S > c | H_0] = \alpha$
oft spezielle Wahlen für α -Werte:
 $\alpha = 0.05$ "signifikant"
 $\alpha = 0.01$ "hochsignifikant"

Simulation

Spezifikation Modellgrößen

Fehlertypen

- (i) (statistischer) **Fehler 1.Art** (-Fehler)
wenn Entscheidung zugunsten H_1 getroffen, obwohl de facto H_0 gegeben
bedeutet: **fälschliches Verwerfen** (der Nullhypothese)
- (ii) (statistischer) **Fehler 2.Art** (-Fehler)
wenn Entscheidung zugunsten H_0 getroffen, obwohl de facto H_1 gegeben
bedeutet: **fälschliches Akzeptieren** (der Nullhypothese)

Impliziert ein bestimmter Test mit Wahrscheinlichkeit $1 - \alpha$ Fehler 1. Art, heißt er "Test zum Niveau α " (auch: "Niveau α -Test") (wo Niveau kurz für **Signifikanzniveau**), unabhängig von Wahrscheinlichkeit β eines Fehlers 2. Art

"Gütefunktion" / "Operationscharakteristika" / "power" von Tests zielen auf Aussagen über β (bei gegebenem α); häufig wenig darüber bekannt

Entscheidungsverfahren meist so, daß **Teststatistik** $S(Y_1, Y_2, \dots, Y_n)$
d.h. Funktion der Stichproben $-$ Variablen,
beim Einsatz: $-$ Werte festgelegt, deren Werte $S(y_1, y_2, \dots, y_n)$ umso größer sind, je unwahrscheinlicher H_0 ist (und implizit: je wahrscheinlicher H_1 ist)
bei Vorliegen dieser Stichprobe

Simulation

Spezifikation Modellgrößen

be/ja/2

5 - 32(6)

5.3.1 Der Chi-Quadrat- (χ^2 -) Test

Vorbereitung:

- Chi-Quadrat- (χ^2 -) Verteilung ist in Statistik häufig verwendete Verteilungsfamilie
- Definition:
Seien Y_1, Y_2, \dots, Y_k
k unabhängige, identisch $N(0,1)$ -verteilte ZV, d.h.

$$f_{Y_i}(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \quad i = 1, 2, \dots, k$$

Dann ist

$$Y = \sum_{i=1}^k Y_i^2$$

wieder ZV, hat Verteilung (so benannt)
 χ^2 -Verteilung mit k Freiheitsgraden

- Familie der χ^2 -Verteilung liegt tabelliert vor (keine explizite funktionale Form angebar)

Idee χ^2 -Test:

- liege vor Stichprobe $y = (y_1, y_2, \dots, y_n)$ (n Beobachtungen einer ZV Y mit Dichte $f_Y(y)$)
- werden Beobachtungen einsortiert in geschlossene Folge von Intervallen
 $[b_0, b_1), [b_1, b_2), \dots, [b_{k-1}, b_k)$
(analog Histogramm, gleiche Intervallbreiten nicht erforderlich)

Simulation

Spezifikation Modellgrößen

- und werden Beobachtungen je Intervall gezählt

$$z_i := \left| \left\{ y_j : b_{i-1} < y_j < b_i \right\} \right| \quad i = 1, 2, \dots, k$$

- dann sollten relative Häufigkeiten

$$r_i := z_i/n \quad i = 1, 2, \dots, k$$

für hinreichend große Stichprobe (großes n) nahe theoretischen Wahrscheinlichkeiten

$$p_i := \int_{b_{i-1}}^{b_i} f_Y(y) dy$$

des Einnehmens dieser Intervalle liegen

- Differenzen

$$z_i - n p_i$$

liefern Maße der Abweichungen je Intervall, ihr gewogenes quadratisches Mittel

$$d := \frac{1}{n} \sum_{i=1}^k \frac{(z_i - n p_i)^2}{p_i}$$

ist (ein) mögliches Maß der "Gesamtabweichung"

- Erwartung:
je kleiner d, desto geringer Abweichung analytische Verteilung / Beobachtungen und umgekehrt;
bzw.:
je kleiner d, desto wahrscheinlicher ist y tatsächlich aus $f_Y(y)$ gezogen

- Realisierungen d von D (errechnete Abw.maße), die $d > \chi^2_{(1)}$ erfüllen, treten bei zutreffender Hypothese auf mit Wahrscheinlichkeit

$$= \int_{\chi^2_{(1)}}^{\infty} f_{\chi^2_{(1)}}(x) dx$$

- d-Werte, die (bei zutreffender Hypothese) in lediglich wenig wahrscheinlichen Intervallen liegen
z.B. $d > \chi^2_{(0,1)}$, $d > \chi^2_{(0,05)}$
als Grund interpretiert, **Hypothese zu verwerfen** (in Bsp.Skizze: d_2)
- dabei in Kauf zu nehmen, daß mit gewisser W'keit
z.B. in 10% 5%
aller Fälle (aller Schätzvorgänge)
Verwerfung fälschlicherweise vorgenommen (Typ 1 !)
- kleinere d-Werte:
kein Anlaß, zu verwerfen,
default: **Hypothese zu akzeptieren**
(Erinnerung: mit Typ 2 Fehlerwahrscheinlichkeit fälschlich akzeptiert)
(in Bsp.Skizze: d_1)
- kritische Werte aus Tabellen (Vorsicht: $\chi^2_{1-\alpha}$ vs χ^2_{α})

- jetzt die (standardmäßige) Überlegung:

wenn y tatsächlich aus $f_Y(y)$ gezogen wird, welche Werteintervalle nimmt die ZV D (d ist deren Realisierung) mit welchen Wahrscheinlichkeiten ein?

i.a.W.:

Wie ist die Verteilung von D unter Hypothese $F_Y(y)$?

- Fallunterscheidung
 - sind Parameter der analytischen Y-Verteilung unabhängig von Stichprobe ermittelt (also **nicht** aus dieser geschätzt),
dann läßt sich zeigen, daß D asymptotisch χ^2_{k-1} -verteilt (für hinreichend große n approximativ χ^2_{k-1} -verteilt) $f_D(d)$ also bekannt

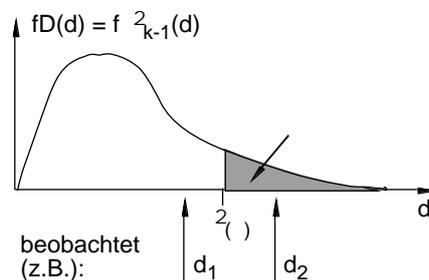


Abbildung 5.3.3: Skizze Entscheidungsverfahren

- weiter bei Fallunterscheidung:
obige Aussagen (Verteilung D) nicht zutreffend, wenn Hypothese aus Stichprobe gewonnen, also wenn (wie vorgeschlagen) Parameter aus y geschätzt (damit ist Teil der in y enthaltenen Information bereits ausgeschöpft)
- für bereits "benutzte" Stichprobe:
 - sind Parameter der analytischen Y-Verteilung derart aus Stichprobe ermittelt, daß Stichprobenwerte zunächst in k Intervalle sortiert, dann ML-Schätzer für p Parameter daraus ermittelt (**wir** hatten dies **nicht** getan)
dann läßt sich zeigen, daß D asymptotisch χ^2_{k-p-1} -verteilt $f_D(d)$ also erneut bekannt
 - sind Parameter der analytischen Y-Verteilung als ML-Schätzer aus **nicht** gruppierten Daten gewonnen,
dann läßt sich zeigen, daß D asymptotisch (im interessierenden Bereich) "zwischen" χ^2_{k-p-1} - und χ^2_{k-1} -verteilt $f_D(d)$ also eingegrenzt

- Entscheidungsverfahren für letzteren Fall

Beispiel: $k=11, p=2, \alpha=0.05$

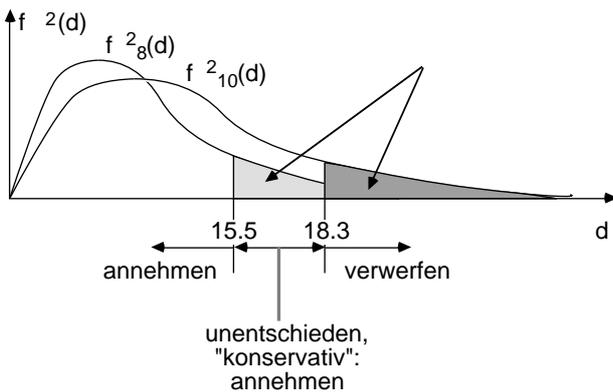


Abbildung 5.3.4: Skizze Entscheidungsverfahren

im "unentschiedenen" Fall
 (irgendwo dort liegt der wahre kritische Wert):
 konservative Entscheidung
 ("zögern zu verwerfen");
 damit aber Typ 2 Fehler automatisch größer

- Praktische Hinweise
 (χ^2_{k-1} -Verteilungen nur asymptotisch richtig)

- Intervalle nicht zu klein wählen,
 damit hinreichend viele Beobachtungen je Intervall

(oft kolportierte) Faustregel: $z_i > 5$ ($>10, >20$?)

bei (Voraus-) Intervallfestlegung also
 $n \cdot p_i > 5$
 wählen, d.h.

$$5 < n \int_{b_{i-1}}^{b_i} f_Y(y) dy = n (FY(b_i) - FY(b_{i-1}))$$

z.B. alle Intervalle gleichwahrscheinlich
 $p_i = 1/k \quad i = 1, 2, \dots, k$
 und damit
 $5 < n/k, \quad n > 5 \cdot k, \quad k < n/5$

- selbst bei vielen Daten Zahl Intervalle < 30
- i.allg.: für große Stichproben geeignet
 auch für diskrete Verteilungen anwendbar
 auch bei Parameterschätzung anwendbar

5.3.2 Der Kolmogoroff-Smirnoff-Test

- Grundidee:
 - empirische Verteilungsfunktion aus einer n -Stichprobe ist Treppenkurve (vgl. Abschn. 5.1):
 $F^*Y(y) = (\#y_i \leq y)/n$ ("#" für "Anzahl")
 - Abweichung zwischen $F^*Y(y)$ und $FY(y)$ aus Stichprobe hypothetischerweise zugrundeliegende V/Fkt. sollte als Maß der "Paßgüte" brauchbar sein; "Abweichung" noch zu definieren
- Abweichung im Kolmogoroff-Smirnoff- (KS-) Sinn ist maximaler Abstand zweier Verteilungsfunktionen
- Testgröße KS-Test (als Anpassungstest) bei n -Stichprobe ist entsprechendes

$$D_n := \max_y | F^*Y(y) - FY(y) |$$

(größter vertikaler Abstand der Funktionen, wo nötig, mit "sup" statt "max" definiert)

auch: $d_n := g(n, D_n)$

(mit speziellen Funktionen $g(\dots)$, s. unten)

- Testhypothese
 $H_0: F^*Y(y) = FY(y) \quad \text{für alle } y$
 Alternativhypothese
 $H_1: F^*Y(y) \neq FY(y) \quad \text{für wenigstens ein } y$

- Durchführung

Fallunterscheidungen:

- falls Parameter von $FY(y)$ **nicht** aus Stichprobe, ist Verteilung von D_n (unabhängig vom Typ der Verteilung von Y) bekannt und kritische Werte vertafelt
 approximativer Test durchführbar mit
 $d_n := (\sqrt{n} + 0.12 + 0.11/\sqrt{n}) D_n$
 und einer (von n unabhängigen) Tafel kritischer Werte $\{c\}$ vgl. LaKe82
 wie üblich, H_0 zu verwerfen falls $d_n > c$ (wo z.B. $c = 0.1, 0.05, 0.01$)
- falls Parameter von $FY(y)$ aus Stichprobe geschätzt, (und D_n sicher von Verteilungstyp abhängig) ist D_n -Verteilung nur bekannt für spezielle Y -Vert'gen so für: vgl. LaKe82

- Normalvert'g (μ^* , σ^2 erwartungstreu geschätzt) mit approx. Testgröße

$$d_n' := (\sqrt{n} - 0.01 + 0.85/\sqrt{n}) D_n$$

und zugehörigen kritischen Werten (Tafel) $\{c'\}$

- Exponentialvert'g (μ^* erwartungstreu geschätzt) mit approx. Testgröße

$$d_n'' := (\sqrt{n} + 0.26 + 0.5/\sqrt{n}) (D_n - 0.2/n)$$

und zugehörigen kritischen Werten (Tafel) $\{c''\}$

- Weibull-Verteilung (s. Literatur)