5. Spezifikation quantitativer Modellgrößen

Modellbildung identifiziert
- Eigenschaften Ersatzsystem
  (auf Basis mentalen Modells, in gewählterAbstraktion)

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"

	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
	stochastisch:			
	Verteilung	stoch. Proz.	Verteilung	stoch. Proz.
endogen	bei festliegender Aufgabenstellung nicht existent		ergeben sich aus Analyse	
	deterministisch:			
			Resultat-	
			Wert	Fkt. der Zeit
	stochastisch:			
			Resultat-	
			Werte-menge (Stichprobe)	Wertverlauf-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 (vgl 3.3)

Notwendige Modellierungsschritte:
(i) Identifikation Verteilungstyp
(ii) (uU) 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: vgl Literatur

zu (d):
- sehr "ungemütliche" Situation !
- uU 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 $P[Y<u] \approx P[Y>o] \approx 0$
  → Vorschlag: **Gleichverteilung**

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

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

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

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

$$FY(y) = \begin{cases} 0 & y \le u \\ \dfrac{(y-u)^2}{(o-u)(w-u)} & u < y \le w \\ 1 - \dfrac{(o-y)^2}{(o-u)\,(o-w)} & w < y \le o \\ 1 & y > o \end{cases}$$

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

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

  → Vorschlag: **Beta-Verteilung** "beta(α_1,α_2)"

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

$$fY(y) = \begin{cases} y^{\alpha_1-1}(1-y)^{\alpha_2-1}/B(\alpha_1,\alpha_2) & 0<y\leq 1 \\ 0 & \text{sonst} \end{cases}$$

$$E[Y] = \frac{\alpha_1}{\alpha_1+\alpha_2} \quad (= m)$$

$$M[Y] = \frac{\alpha_1-1}{\alpha_1+\alpha_2-2} \quad (= w)$$

  $\alpha_1,\alpha_2 > 1$, aus m und w zu bestimmen,
  B(…) "Beta-Funktion":

$$B(\alpha_1,\alpha_2) := \int_0^1 t^{\alpha_1-1}\ (1-t)^{\alpha_2-1}\ dt$$

zu (a):
Einige Hinweise zu Verteilungsformen
auf Basis theoretischer Überlegungen
(vgl Mihr72,LaKe82)

- ZV Y, welche Summe
einer größeren Anzahl zufälliger Einflüsse darstellt,
könnte (zentraler Grenzwertsatz) **normalverteilt** sein
$N(\mu,\sigma^2)$-verteilt

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

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

$$F_Y(y) = \begin{cases} 1 - \exp\left(-(y/\beta)^{\alpha}\right) & y > 0 \\ 0 & y \le 0 \end{cases}$$

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 Δt
    (bei beliebiger Lage solchen Zeitintervalls)
    mit fester Wahrscheinlichkeit $\approx \lambda \cdot \Delta t$ auftreten
    (wo λ positive "charakterisierende" Konstante)
  - insgesamt mit konstanter "Rate" λ auftreten

  könnte (Satz über "seltene Ereignisse")
  (negativ) **exponentiell verteilt** sein

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

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,
  $LN(\mu,\sigma^2)$-verteilt

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

Insgesamt zu (a):
- Werte Verteilungsparameter?
- "wie gesagt", aus Stichprobe "schätzen"

vgl Abschnitt 5.2

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

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

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

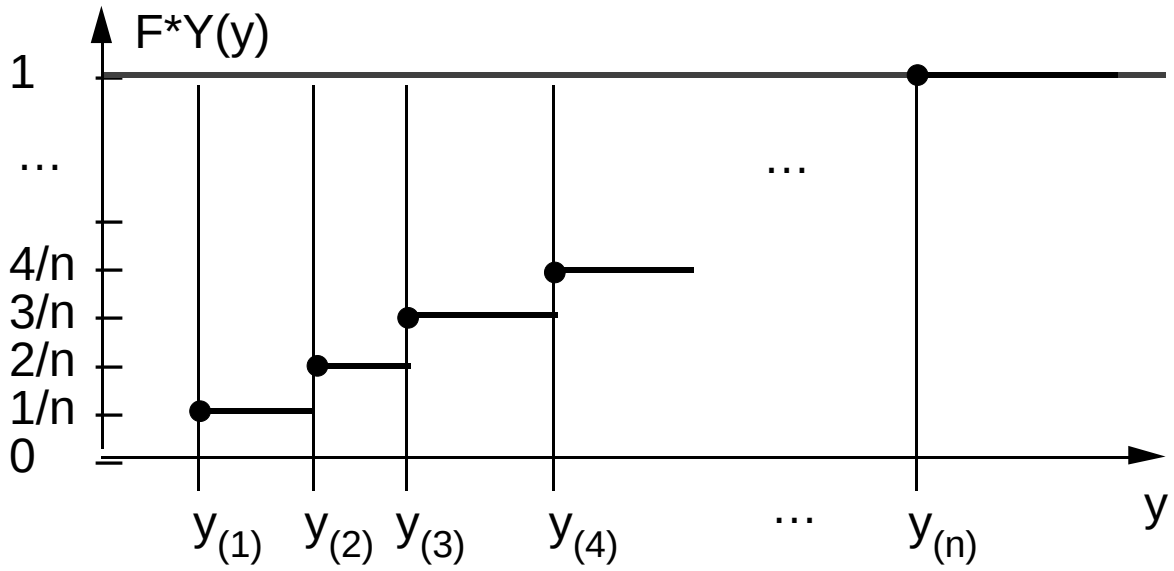
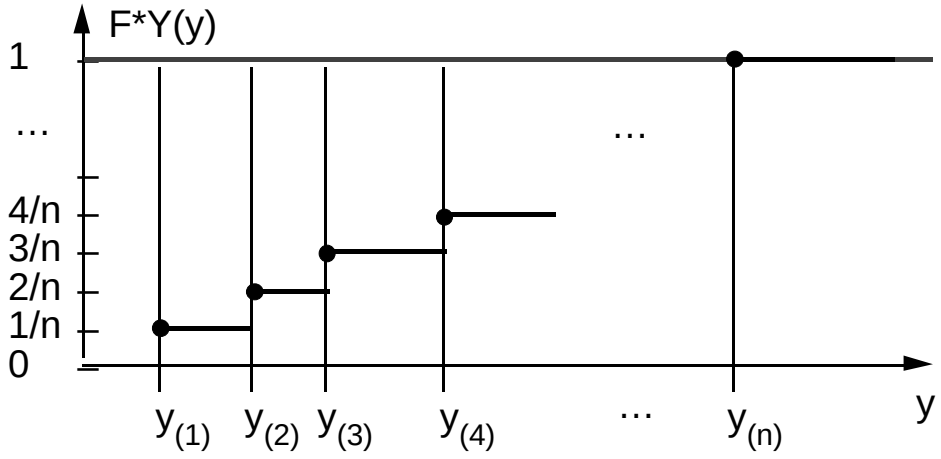


Abbildung 5.1.1: empirische Verteilungsfunktion

Empirische Verteilungsfunktion



schätzt Verteilungsfunktion erwartungstreu:

- aus Zeichnung:

$$\tilde{F}Y(y) = \frac{|\{\text{Realisierungen von } Y \le y \mid n\text{–Stichprobe}\}|}{n}$$

s. Zeichnung

$$= \sum_{1}^{n} I_{Y \le y} \Big/ n$$

wg. Unabhängigkeit

- wo "Indikatorfunktion":

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

- und damit:

$$E\left[\tilde{F}Y(y)\right] = \frac{1}{n} \; n \; E\left[I_{Y \le y}\right]$$
$$= P[Y \le y]$$
$$= FY(y)$$

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 $F_Y^{-1}(U)$ gemäß FY verteilte ZV

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

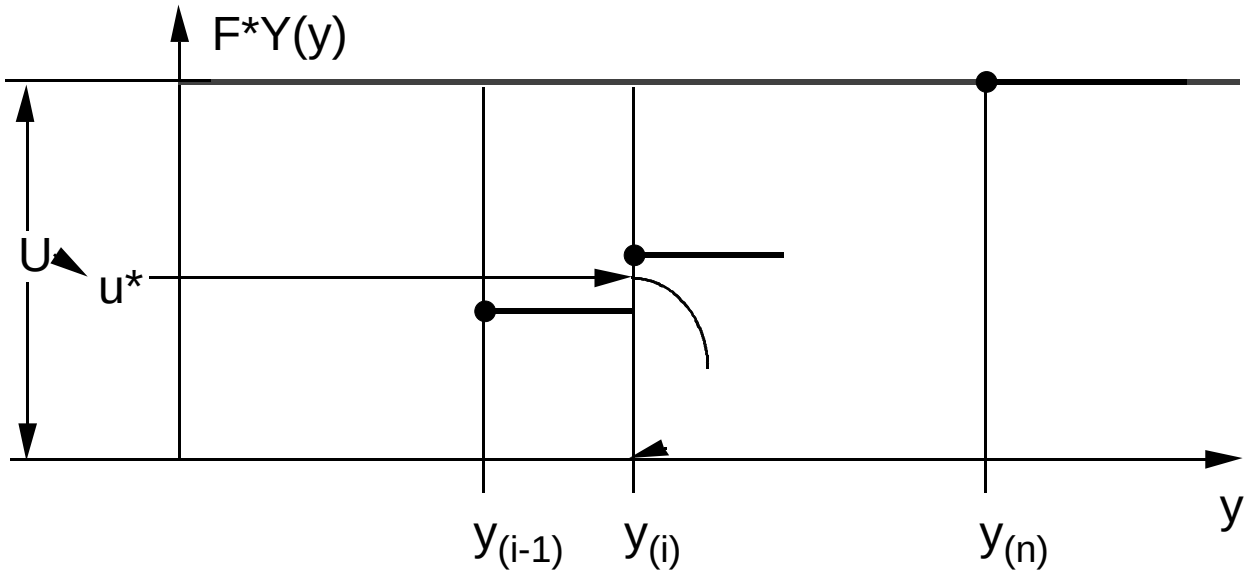


Abbildung 5.1.2: "Ziehen" aus
empirischer Verteilungsfunktion

bzw einfacher (identisches Ergebnis !):
alle Treppenstufen gleich hoch (1/n),
damit wegen gleichverteiltem U
- alle Werte $y_{(1)},y_{(2)},\ldots,y_{(n)}$
- bzw alle Werte $y_1,y_2,\ldots,y_n$

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 $y_{(1)},y_{(2)},\ldots,y_{(n)}$
bzw $y_1,y_2,\ldots,y_n$

gelegentlich empfohlen:

lineare Interpolationen zur Definition von F*Y
(auch Interpolationen höherer Ordnung)
klingt plausibel, zerstört aber Erwartungstreue!

numerisch uU

zwei oder mehr Werte der Stichprobe identisch:

- Treppenstufen entsprechend höher,
- bzw beide (alle) Werte ins n-Feld

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 (vgl Abschn. 4.1):

  $$\tilde{\mu}_1 = \sum_i Y_i / n$$

  $$\tilde{\sigma}^2 = \sum_i (Y_i - \tilde{\mu}_1)^2 / (n-1)$$

  Natürlicher (nicht erwartungstreuer) Schätzer für VK wäre:

  $$\tilde{\delta} := \sqrt{\tilde{\sigma}^2} / \tilde{\mu}_1$$

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

  "ähnliche" Ausschlußhinweise vgl zB LaKe00

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

- Weg:

  - vorhanden: Stichprobe $(y_1,y_2,\ldots,y_n)$

  - bilde "Klassen" von Wertebereichen
    $[b_0,b_1),[b_1,b_2), \ldots ,[b_{k-1},b_k)$
    gleicher Breite
    $\Delta b=b_j-b_{j-1}$ $\quad j=1,2,\ldots,k$

  - bestimme relative Häufigkeiten

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

    und zeichne "Histogramm":

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

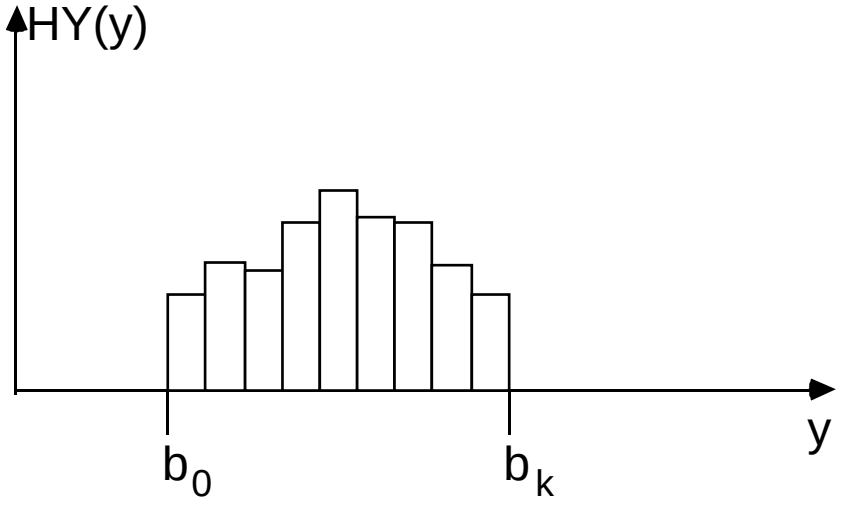


Abbildung 5.1.3: Histogramm

- Histogramm:
  - Aussehen sehr abhängig von Δb 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" uU vernachlässigen

- Histogramm HY(y) schätzt Dichte fY(y)
  (bis auf Konstante):
  - einerseits ist

$$P\left[b_{j-1} \le Y < b_j\right] = \int_{b_{j-1}}^{b_j} fY(y)\ dy$$

$$= \Delta b\ fY(y'_j)$$

    mit (Mittelwertsatz:) einem y'_j $[b_{j-1},b_j)$

  - andererseits ist für y' $[b_{j-1},b_j)$
    $HY(y') = r_j$
    mit
    $E[R_j] = P[b_{j-1} \le Y < b_j]$

  - so daß
    $r_j \approx \Delta b \cdot fY(y'_j)$
    und HY zu fY ungefähr proportional

  - und somit HY(y)/Δb Schätzer $\tilde{f}Y(y)$ für Dichte fY(y)

- Aus dem Bild (Histogramm) nun
  (subjektiv und erfahrungsabhängig)
  Typ einer uU zugrundeliegenden Verteilung identifizieren

Jede (analytisch charakterisierte) Verteilungsfamilie
weist gewisse Parameter auf,
hier erfaßt durch Parametervektor $\underline{Q}$
(zB: N: μ,σ EXP: λ)
Werte dieser Parameter (nach Identifizierung V-Familie)
noch zu bestimmen (schätzen) - analog Situation a) -
Diesbezügliche Verfahren vgl Abschn. 5.2

uU kann schon vor Parameterschätzung
Hypothese V-Typ verworfen werden;
so bei V-Familien, deren Mitglieder ausschließlich über
- Translation
- Dehnung / Stauchung

auseinander hervorgehen; zB: N

Dazu (wieder) "visuelle" Methode: probability plot

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

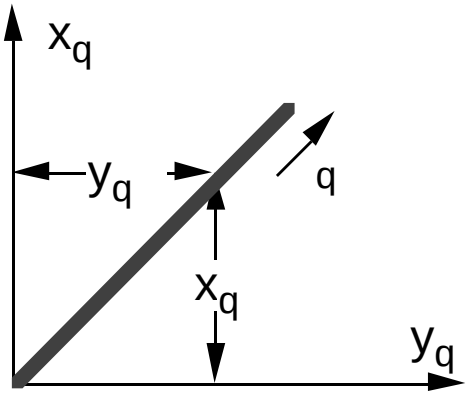


Abbildung 5.1.4

- hängen X und Y über lineare Transformation
  $X = a \cdot Y + b \qquad a>0$
  zusammen, dann

$$\begin{aligned} FX(x) &= P[X \le x] \\ &= P[a\,Y + b \le x] \\ &= P\left[Y \le \frac{x-b}{a}\right] \\ &= FY\left(\frac{x-b}{a}\right) \end{aligned}$$

  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

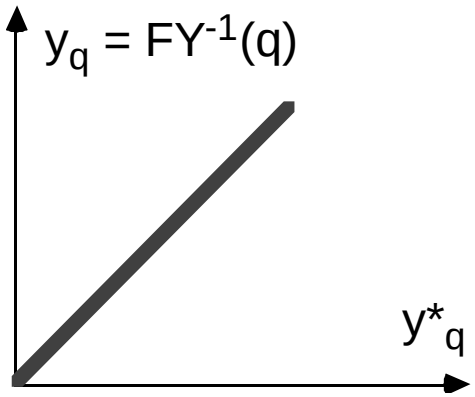


Abbildung 5.1.5: probability plot

- 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:

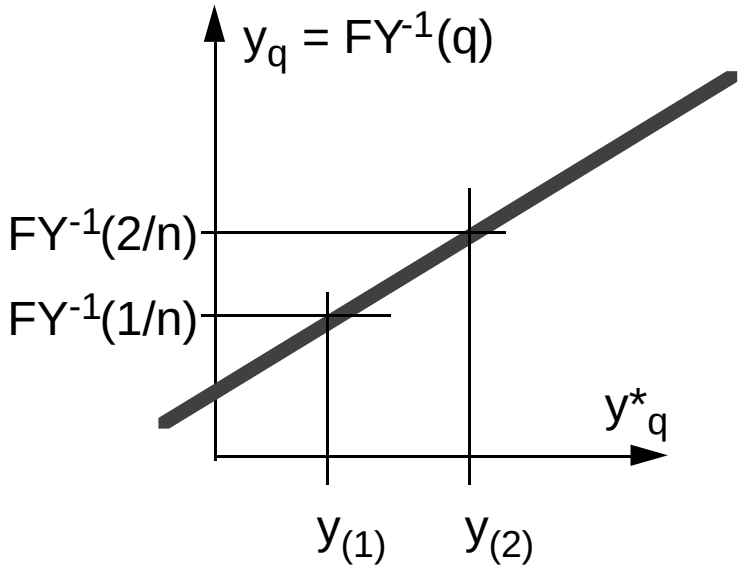


Abbildung 5.1.6: probability plot als Beurteilungsgraph

- 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: Hypothese ablehnen,
    "Ahnung FY(y) war falsch"

5.2 Schätzung von Verteilungsparametern

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

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

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

- deren Parameter

  $\underline{\Theta} = (\Theta_1,\Theta_2,\ldots,\Theta_p)^T$

  aber nicht bekannt,
  also zu bestimmen sind,

  zu schätzen sind aus vorliegender Y-Stichprobe

  $(y_1,y_2,\ldots,y_n)$

Für Aufgabe verfügbar diverse statistische Verfahren,
hier vorgestellt:

- Momentenmethode

- maximum likelihood Methode

Momentenmethode

- bedient sich der k-ten Momente von Y
  $\mu_k := E[Y^k]$

  bzw ihrer erwartungstreuen Schätzer (vgl Abschn. 4.1)

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

  bzw diesbezüglicher Schätzwerte μ_k^*

- Parameter Θ_j lassen sich oft ausdrücken
  als Funktionen der Momente
  $\Theta_j = \Phi_j(\mu_1,\mu_2,\ldots,\mu_p) \qquad j = 1,2,\ldots,p$

- Momentenmethode
  - substituiert (in diesen Funktionen)
    Momentenschätzer für Momente
  - gewinnt so **Parameterschätzer**
    $\tilde{\Theta}_j = \Phi_j(\tilde{\mu}_1,\tilde{\mu}_2,\ldots,\tilde{\mu}_p) \qquad j = 1,2,\ldots,p$

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

  (Erinnerung:
  Schätzer $\tilde{\Theta}$ für Größe Θ heißt
  - erwartungstreu, wenn $E[\tilde{\Theta}] = \Theta$
  - asymptotisch erwartungstreu, wenn $\lim_{n\to\infty} E[\tilde{\Theta}] = \Theta$
  - konsistent, wenn $\lim_{n\to\infty} P[\ |\tilde{\Theta}-\Theta| > \varepsilon\ ] = 0 \qquad \varepsilon>0$)

  oft dennoch keine "guten" Schätzer ("Form" Verteilung)

Beispiel: Exponentialverteilung

- Dichtefunktion:

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

- erstes Moment:

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

partielle Integration

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

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

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

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

- (folglich) Zusammenhang
  (erstes) Moment (μ_1) vs (einziger) Parameter (λ):
  $\lambda = 1/\mu_1$

- Parameterschätzer, mit Momentenschätzer (4.1.4) :

$$\tilde{\lambda} = n \Big/ \sum_1^n Y_i$$

maximum likelihood Methode

- Erinnerung:

  - Parameter $\underline{Q}$ der Verteilung einer ZV Y,
    deren Typ bekannt, zB als Dichte $f_Y(y;\underline{Q})$,
    zu schätzen aus Stichprobe $(y_1^*,y_2^*,\ldots,y_n^*)$
    (* für Stichprobenwerte zur Unterscheidung, wo nötig)

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

  - alle Stichprobenvariablen identisch verteilt:
    $f_{Y_i}(y_i;\underline{Q}) \equiv f_Y(y_i;\underline{Q}) \qquad i = 1,2,\ldots,n$

  - alle Stichprobenvariablen unabhängig verteilt:

    $$f_{\underline{Y}}(\underline{y};\underline{\Theta}) = \prod_{1}^{n} f_Y(y_i;\underline{\Theta})$$

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

- Idee der Methode:

  Parametervektor $\underline{Q}$ so wählen (bestimmen),
  daß Beobachtung $(y_1^*,y_2^*,\ldots,y_n^*)$
  in den Punkt maximaler Dichte,
  – Maximum von $f_{\underline{Y}}(\underline{y};\underline{Q})$ –
  zu liegen kommt

  Motivation: Umgebung dieses Punktes ist Bereich
  größter Beobachtungswahrscheinlichkeit

- Weg:

  - Maximierung der
    maximum-likelihood-Funktion

    $$L(\underline{\Theta}) := \prod_{1}^{n} fY(y_i;\underline{\Theta})$$

    bezüglich Komponenten des Vektors $\underline{Q}$

  - notwendige Bedingung für Maximum mehrdimensionaler, differenzierbarer Funktion (sei für gemeinsame Dichte $L(\underline{Q})$ vorausgesetzt) ist

    $$\left(\frac{\partial L}{\partial \Theta_j}\right) = (0)$$

    woraus p Bestimmungsgleichungen für die
    Θ_j $\quad j = 1,2,\ldots,p$
    folgen

  - uU muß explizit auf Vorliegen Maximum / Minimum / Sattelpunkt geprüft werden

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

- praktische Anwendung

  - Logarithmus log(L(Q)) der likelihood-Funktion,
    sog. **log-likelihood-Funktion**

$$\log(L(\underline{\Theta})) = \log\left(\prod_{1}^{n} fY(y_i;\underline{\Theta})\right)$$

$$= \sum_{1}^{n} \log(fY(y_i;\underline{\Theta}))$$

    hat wegen Monotonie der Logarithmus-Funktion
    Maximum an derselben Stelle wie likelihood-Funktion

  - log(L($\underline{Q}$)) wird zur Bestimmung
    des Maximums, daraus folgend der Schätzer $\tilde{\Theta}_j$
    wegen "leichterer" Differenzierbarkeit (Summe !)
    gern anstelle L($\underline{Q}$) verwendet

  - ist das Gleichungssystem
    zur Bestimmung der Schätzer $\tilde{\Theta}_j$
    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)

Beispiel: (wieder) Exponentialverteilung

- Dichtefunktion:

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

- log-likelihood Funktion:

$$\log L = \sum_{1}^{n} \log(\lambda e^{-\lambda y_i})$$

- partiell differenziert (einziger Parameter ist λ)

$$\frac{\partial \log L}{\partial \lambda} = \sum_{1}^{n} \frac{\lambda (-y_i) e^{-\lambda y_i} + e^{-\lambda y_i}}{\lambda e^{-\lambda y_i}}$$

$$= \sum_{1}^{n} \frac{1 - \lambda y_i}{\lambda}$$

- Maximumbedingung:

$$\frac{\partial \log L}{\partial \lambda} = 0 \qquad n - \lambda \sum y_i = 0$$

- daraus Parameterschätzer und Parameterschätzwert

$$\tilde{\lambda} = n / \sum Y_i \qquad \lambda^* = n / \sum y^*_i$$

- es ist "purer Zufall", daß in diesem Fall Momentenschätzer
  und **MLE** (maximum likelihood estimator)
  übereinstimmen

5.3 Überprüfung der Paßgüte angepaßter Verteilungen

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

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

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

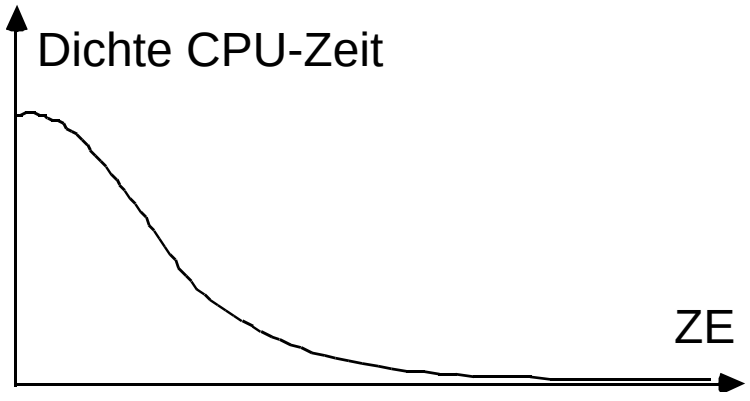


Abbildung 5.3.1: "erahnte" Dichte CPU-Zeit-Verteilung

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 ($y_1,y_2,\ldots,y_n$)
Beispiele: E[Y] > bestimmter Wert,
FY(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;
  dabei zwei "Typen" von "Fehlern" zu unterscheiden

Fehlertypen

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

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

Impliziert ein bestimmter Test
mit Wahrscheinlichkeit ≤ α 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 $S(Y_1,Y_2,\ldots,Y_n)$

dh Funktion der Stichproben -Variablen,

beim Einsatz: -Werte

festgelegt, deren Werte

$s(y_1,y_2,\ldots,y_n)$

umso größer sind, je unwahrscheinlicher H_0 ist

(und implizit: je wahrscheinlicher H_1 ist)

bei Vorliegen dieser Stichprobe

Zur Anwendung erforderlich:

- Bestimmung der Verteilung von S
  unter der Voraussetzung: H_0 zutreffend
- Ermittlung von "kritischen Werten" c_a
  (bzw c1−α: Vorsicht in Tafeln !)
  ab denen H_0 zum Niveau α zu verwerfen

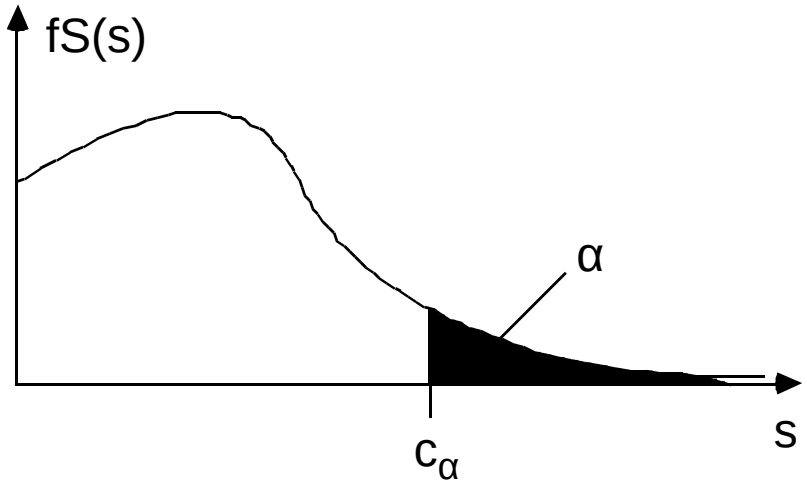


Abbildung 5.3.2: Prinzipskizze statistische Tests

nochmals:

Bestimmung c_a derart, daß $P[S>c_a \mid H_0] = \alpha$

oft spezielle Wahlen für α-Werte:

α = 0.05 "signifikant"

α = 0.01 "hochsignifikant"

5.3.1 Der Chi-Quadrat- (χ^2-) Test

Vorbereitung:

- Chi-Quadrat- (χ^2-) Verteilung
  ist in Statistik häufig verwendete Verteilungsfamilie

- Definition:
  Seien $Y_1, Y_2, \ldots, Y_k$
  k unabhängige, identisch N(0,1)-verteilte ZV, dh

$$fY_i(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \qquad i = 1,2,\ldots,k$$

  Dann ist

$$Y = \sum_{i=1}^{k} Y_i^2 \qquad \text{wieder ZV,}$$

  hat Verteilung ("so benannt")
  χ^2-Verteilung mit k Freiheitsgraden

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

Idee χ^2-Test:

- liege vor Stichprobe
  $\underline{y} = (y_1, y_2, \ldots, y_n)$
  dh n Beobachtungen einer ZV Y mit Dichte fY(y)

- werden Beobachtungen einsortiert
  in geschlossene Folge von Intervallen
  $[b_0, b_1), [b_1, b_2), \ldots, [b_{k-1}, b_k)$

  (analog Histogramm,
  aber gleiche Intervallbreiten nicht erforderlich)

- und werden Beobachtungen je Intervall gezählt

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

- dann sollten relative Häufigkeiten

$$r_i := z_i/n \qquad i = 1,2,\ldots,k$$

für hinreichend große Stichprobe (großes n) nahe theoretischen Wahrscheinlichkeiten

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

des Einnehmens dieser Intervalle liegen

- Differenzen

$$z_i - n \cdot p_i$$

liefern Maße der Abweichungen je Intervall, ihr gewogenes quadratisches Mittel

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

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 fY(y) gezogen

- jetzt die (standardmäßige) Überlegung:

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

  m.a.W.:
  Wie ist die Verteilung von D unter Hypothese FY(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)
    fD(d) also bekannt

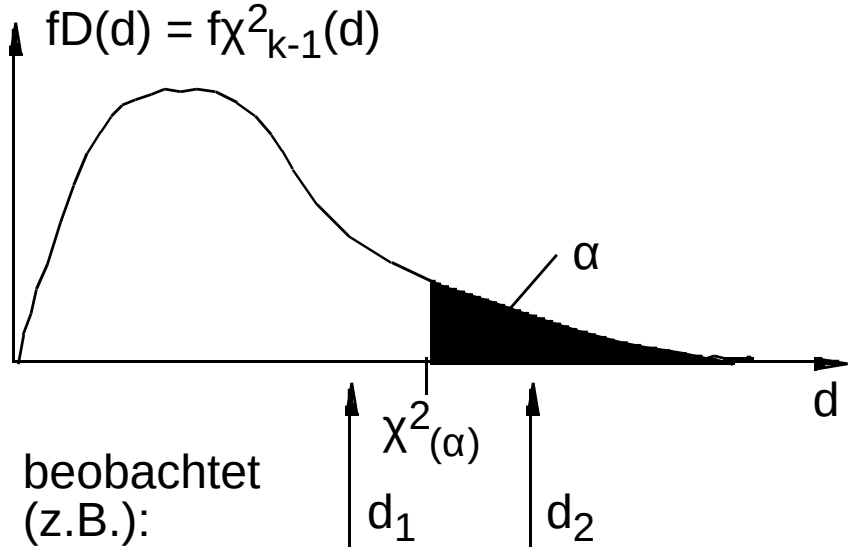


Abbildung 5.3.3: Skizze Entscheidungsverfahren

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

$$\alpha = \int_{\chi^2_{(\alpha)}}^{\infty} f\chi^2_{k-1}(x)\, dx$$

- d-Werte, die (bei zutreffender Hypothese)
  in lediglich wenig wahrscheinlichen Intervallen liegen
  zB $d > \chi^2_{(0.1)}$, $d > \chi^2_{(0.05)}$
  als Grund interpretiert, **Hypothese zu verwerfen**
  (in Bsp.Skizze: d_2)

- dabei in Kauf zu nehmen, daß mit gewisser W´keit
  zB in 10% 5%
  aller Fälle (aller Schätzvorgänge)
  Verwerfung fälschlicherweise vorgenommen (Typ 1 !)

- kleinere d-Werte:
  kein Anlaß, zu verwerfen,
  default: **Hypothese zu akzeptieren**
  (in Bsp.Skizze: d_1)

- dabei mit Typ 2 Fehlerwahrscheinlichkeit
  fälschlich akzeptiert

- kritische Werte aus Tabellen (Vorsicht: α vs 1-α)

- weiter bei Fallunterscheidung:
  obige Aussagen (Verteilung D) nicht zutreffend,
  wenn Hypothese aus Stichprobe gewonnen,
  also wenn (wie vorgeschlagen)
  Parameter aus $\underline{y}$ geschätzt
  (damit ist Teil der in $\underline{y}\_$ 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
    fD(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

    FD(d) also eingegrenzt

- Entscheidungsverfahren für letzteren Fall

  Beispiel: k=11, p=2, α=0.05

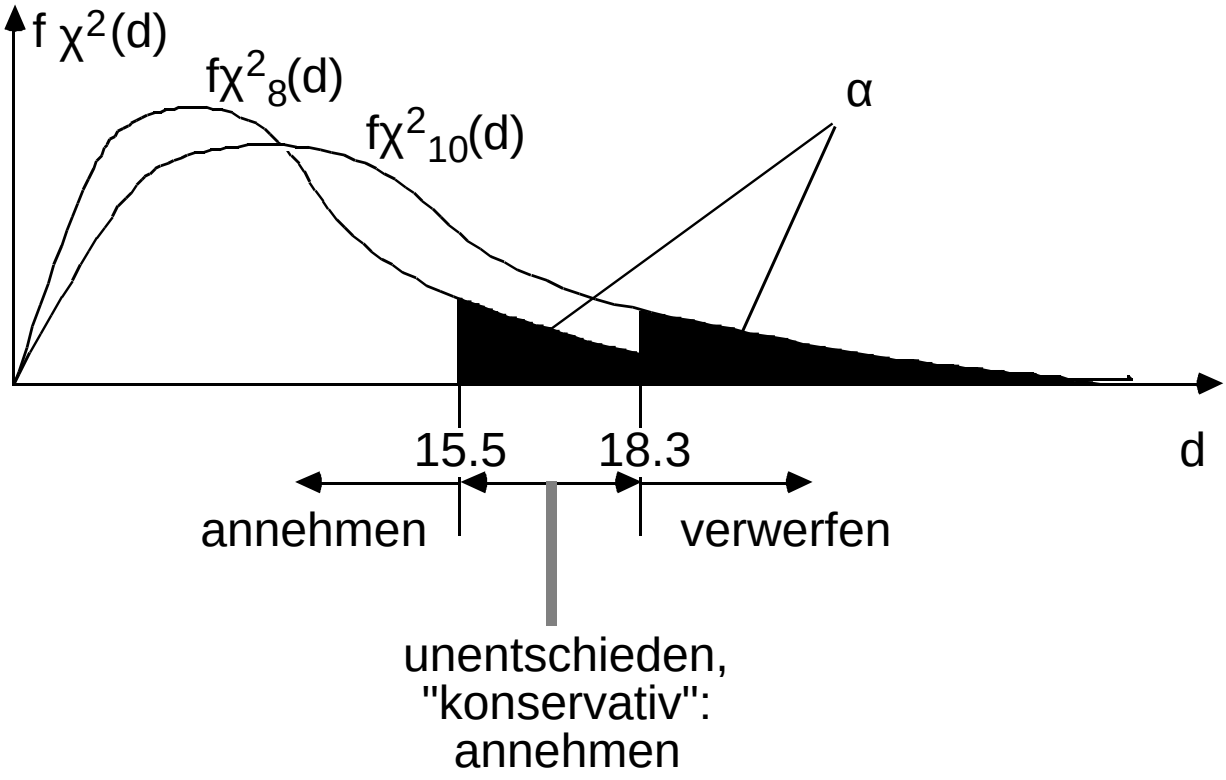


Abbildung 5.3.4: Skizze Entscheidungsverfahren

im "unentschiedenen" Fall
(irgendwo dort liegt der wahre kritische Wert):

ist sog "konservative" Entscheidung:
"zögern zu verwerfen" → annehmen
damit aber Typ 2 Fehler automatisch größer

Unterschied in praktisch häufigen Fällen
(p eher klein, k eher groß)
ohnehin gering

- 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, dh

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

    zB alle Intervalle gleichwahrscheinlich
    $p_i \equiv 1/k \qquad i = 1,2,\ldots,k$
    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 $F^*Y(y)$ aus Stichprobe
    und $FY(y)$ hypothetischerweise zugrundeliegende Verteilungsfkt.
    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(…), vgl unten)

  - Testhypothese
    H_0: $F^*Y(y) = FY(y)$ für alle y
    Alternativhypothese
    H_1: $F^*Y(y) \neq FY(y)$ 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 $\{c_\alpha\}$ vgl LaKe82

    wie üblich, H0 zu verwerfen falls
    $d_n > c_\alpha$ (wo zB $\alpha = 0.1, 0.05, 0.01$)

- 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

  * Normalverteilung (μ^*, σ^{*2} erwartungstreu geschätzt)
    mit approximativer Testgröße
    $d_n' := (\sqrt{n}-0.01+\mathbf{0.85}/\sqrt{n})\, D_n$

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

  * Exponentialverteilung (μ^* erwartungstreu geschätzt)
    mit approximativer Testgröße
    $d_n'' := (\sqrt{n}+0.26+\mathbf{0.5}/\sqrt{n})\,(D_n-0.2/n)$

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

  * Weibull-Verteilung (vgl Literatur)