6. Kalibrierung und Validierung von Simulationsmodellen

Zur Erinnerung an die Vorlesung MAO

- Modelle werden erstellt,
  - ▷ um Experimente mit realen Systemen zu vermeiden
  - ▷ um Aussagen über (potentielle) Objekt-Systeme zu erhalten, die (noch) nicht existieren
  - ▷ um Vorhersagen machen zu können
- Folgerungen aus Modellanalysen (hier: Simulationsexperimenten) sollten weitestgehend gleich sein zu Folgerungen, die aus entsprechenden (potentiell irgendwann möglichen) Objekt-Analysen/-Experimenten gewonnen würden
- Gleiche Beobachtungsschemata vorausgesetzt, sind Folgerungen (zumindest) dann identisch, wenn unmittelbare Beobachtungen/Resultate von Simulations- und Objekt-Experimenten identisch

Ist Identität der Resultate zu erwarten??

Sicher nicht immer, nicht allgemein, nicht in jedem Detail!!

Objekt- und Modell-System nicht identisch ⇒ Verhaltenunterschiede sind zu erwarten

Welche Verhaltensunterschiede sind tolerierbar?

6.1. Grundlegende Definitionen

Benötigt wird eine vernünftige Definition von

Realitätstreue, **Gültigkeit**, **validity**

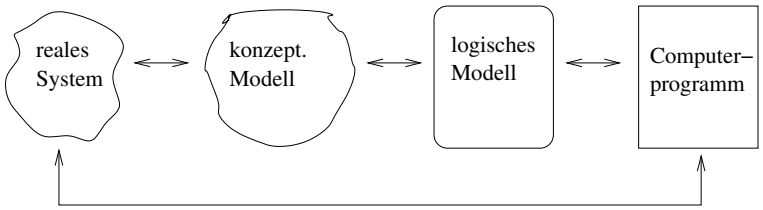
- müsste aufbauen auf Maß für Verhaltensunterschiede
  $D(V_R, V_S)$ (R: real, S: simuliert)
- müsste Realitätstreue bejahen,
  falls Abweichungsmaß unter bestimmtem, festzulegendem
  (problemabhängigem, Fragestellung-abhängigem)
  Grenzwert liegt
- müsste Grenzwert so festlegen, dass
  (unvermeidlich existierende) Verhaltensunterschiede die
  zu treffenden Folgerungen nicht beeinträchtigen
  (d.h. nicht zu unterschiedlichen Folgerungen führt)

Ursachen für Verhaltensunterschiede:

- strukturelle Unterschiede
  - ▷ Abstraktion, Aggregierung (beabsichtigt!)
  - ▷ Ungenauigkeiten (bugs) (unbeabsichtigt!)
    Fehler (faults) (unbeabsichtigt!)
- Parameterwerte
  - ▷ deterministische Werte + Abhängigkeiten
  - ▷ Verteilungsfunktionen

(• dazu stochastische Schwankungen)

Transformationsschritte im Rahmen der Modellbildung:



Transformationsprozess von vage-definierter Problemstellung
(schlecht-strukturiertem Realsystem)
zu wohl-definiertem Computerprogramm

Jeder angegebene Transformationsschritt kann zu neuen
Fehlern/Irrtümern/Verzerrungen führen

<u>Betrachtung der einzelnen Schritte:</u>

- reales System → konzeptuelles Modell
  Festlegung der Systemabgrenzung,
  Bestimmung der relevanten Systemelemente
- konzeptuelles Modell → logisches Modell
  Formulierung relevanter Zusammenhänge zwischen
  Systembestandteilen und Abbildung der Umwelteinflüsse
  Darstellung als Flussdiagramm, Petri-Netze etc.
- logisches Modell → Computerprogramm
  Codierung, Implementierung des Programms
- Computerprogramm → reales System
  Aussagen über Systemverhalten, Optimierung

Jeder Schritt stellt spezifische Anforderungen und
beinhaltet spezifische Fehlerursachen

Oftmals uneinheitliche Terminologie wir unterscheiden hier strikt:

- **Verifikation**:
  Bestätigung aller Modell-Eingabegrößen/-Annahmen
  inkl. struktureller Annahmen, Programmverifikation,
  Parameterwerte, . . .
- **Validierung**:
  Bestätigung der Modellresultate
- **Kalibrierung**:
  Reduktion von Verhaltensunterschieden
  (d.h. Reduktion von $D(V_R, V_S)$)
  falls Unterschiede als zu groß bewertet
  Anpassung i.a. durch Änderungen am Modell

Sicher (hoffentlich) große Mühe gelegt auf "gute" Wahl
der Eingangsgrößen (Hypothesen, . . . , Parameter)

"Hoffnung" auf gültiges Modell steigt damit

"Garantie" auf gültiges Modell aber nicht gegeben

Positivistischer Blick: Ergebnisse ok, alles ok reicht oft nicht

$\Rightarrow$ Kalibrierung ohne Verifikation/Validierung ist gefährlich

In allen Schritten Beachtung des Kosten/Nutzen Aspektes:

- Nutzen steigt mit dem Erkenntniswert
- Kosten steigen mit Aufwand der Erstellung

$\Rightarrow$ Kompromiss zwischen Resultat und Aufwand

Verifikation:

oft Einsatz (semi-)formaler/automatischer Techniken möglich

Wichtiger Aspekt: Verifikation des Simulationprogramms am logischen/konzept. Modell (also Programmverifikation)

Ansätze und Aspekte:

- Strukturierte Programmierung mit Test/Debugging von Modulen/Subprogrammen
- Code-review durch andere Mitarbeiter
- Bei Verwendung formaler Spezifikationstechniken (semi-)automatische Generierung von Programmcode
- Inkrementeller Entwurf und Test des Simulationsmodells
- Test des Simulators für unterschiedliche Eingabeparameter und Vergleich der Ausgaben mit erwarteten/berechneten Werten
- Erstellen und analysieren von traces
- Simulationsläufe unter vereinfachenden Annahmen (z.B. ohne ZVs), so dass Verhalten vorhersehbar
- "Durchspielen" gewisser typischer Abläufe
- Beobachtung der Animation
- Verwendung von zuverlässigen Simulationspaketen

weiterer Aspekt:

Verifikation der Eingabegrößen ($\Rightarrow$ statistische Verfahren/Tests)

Allgemeine Überlegungen zur Validierung

Simulationsmodell soll nützlich sein d.h. mit sinnvoller Genauigkeit Aussagen über das System erlauben

$\Rightarrow$ es gibt keine absolute Validität von Simulatoren

Zu beachten ist

- Validierung ist modell-individuell
  abhängig von System-/Modellstruktur und Aufgabenstellung
- Validierung ist graduell
  es gibt i.a. keine Abstufung in valide und invalide sondern einen Grad der Validität für ein Modell
- Validität ist oft Ergebnis eines (Verhandlungs-)Prozesses
  letztendlich auch eine Frage der Glaubwürdigkeit und Akzeptanz eines Modells
- Validierung ist ein Projekt-begleitender Prozess
  findet während der gesamten Projektlaufzeit unter unterschiedlichen Randbedingungen statt

Man unterscheidet:

- Funktionsbezogene Validierung (Test der Plausibilität)
- Ergebnisbezogene Validierung
  "Übereinstimmung" der Ergebnisse System/Modell
- Theoriebzogene Validierung
  "Übereinstimmung" der Ergebnisse anal./simul. Modell

Einsatz von Kalibrierung und Validierung setzt voraus,
dass Daten für Realsystem vorhanden sind
oder ermittelt werden können, damit V_R bestimmbar ist

Daten resultieren

- aus Messungen am Realsystem
  - ▷ Daten sind oft "gestört"
    oder müssen aufgearbeitet werden
  - ▷ Daten können nur in aggregierter Form oder
    für nicht relevante Situationen gewonnen werden

Probleme ähnlich zur Datenermittlung zur
Spezifikation quantitativer Modellgrößen

Falls Messungen am Realsystem nicht möglich sind
z.B. weil System (noch) nicht vorhanden

müssen Daten auf anderem Wege gewonnen werden

Möglichkeiten der Datengewinnung:

- aus Daten ähnlicher Systeme
  (u.U. zusätzliche Modifikation der Daten)
- aus anderen Modellen
- nur qualitative Informationen ermitteln

Wir gehen im folgenden davon aus, dass Daten vorhanden sind und betrachten Methoden zur Kalibrierung/Validierung

Kalibrierung schließt (zwangsläufig) ein:

- Identifikation der Ursachen für Verhaltensunterschiede
- entsprechende, gezielte Änderungen am Modell
  - ▷ Struktur: "Code"-Änderungen
  - ▷ Parameter/statische Attribute: "Werte"-Änderungen

Bei stochastischen Modellen resultieren zwei spezifische statistische Problem-Typen:

- Auswahl eines aus mehreren Modellen $(S_1, S_2, \ldots, S_k)$ auf Basis zugehöriger $D(V_R, V_{S_i})$ $(i = 1, \ldots, k)$
  - ▷ D (wie gewohnt) Zufallsvariable oder stochastischer Prozess
  - ▷ (folglich) müssen Unterschiede der D's "das normale Schwankungsmaß" übersteigen $\Rightarrow$ Unterschiede müssen **signifikant** sein

  Aufgabe: **Tests auf Signifikanz**
- "tuning" durch Parameter(wert)veränderungen

  Suche nach Parametervektor p_{opt} derart, dass

  $$\min_p D(V_R, V_S(p))$$

  erreicht wird

  Aufgabe: **Stochastische Optimierung**

Methodisch gesehen, ist

- kalibrieren eines stochastischen Simulators

identisches Problem wie

- Experimentieren mit stochastischem Simulator
  (bzw. stochastischem System)

Für gegebenes Realsystem mit Verhalten V_R wird
(über Experimente) versucht,
System mit Verhaltensgüte V_E zu finden derart, dass
$D^*(V_R, V_E)$ kleiner als vorgegebene Schranke

Möglichkeiten der Veränderung zur
"besten/akzeptablen" Alternative

- Ausprobieren struktureller Alternativen
- Tuning von System-Parametern

Bei der Suche nach Methoden, Hilfsmitteln für Kalibrieren, sollte man demnach fündig werden bei Methoden, Hilfsmitteln des qualitativen und quantitativen Experimentierens mit stochastischen Systemen/Modellen

Zunächst aber prinzipielle Klärung des Begriffs "Kalibrierung"

- angenommen, mittels Kalibrierung sei unmittelbares
  Ziel erreicht: D ist kleiner als (erträgliches) D_{ertr}
- ist damit auch inhaltliches Ziel erreicht: Folgerungen aus
  Objekt- und Modell-Experimenten sind gleich ??

Was genau, wurde getan?

Simulator-Verhalten wurde (mittels Änderungen)

einem Objektsystem-Verhalten "hinreichend" angepasst

- für einen Zustand der Umwelt
- für einen Zustand des Systems

d.h. für genau eine Situation

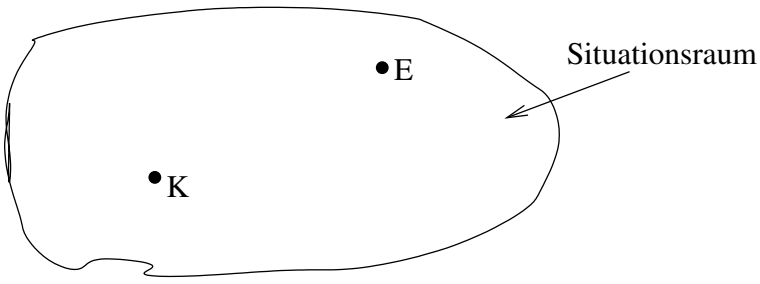
Als "Zweck" des Simulators letztlich angepeilt:

Experimentieren/Analysieren für andere Situationen;

über diese (im Kalibrierungsvorgang eingesetzte)

Situation besteht ja Klarheit (d.h. sind Ergebnisse bekannt!)

Prinzipienskizze:



- Kalibrierung für Situation K
- Einsatz des Simulators in Situation E

Frage damit:

- ist der Unterschied D bei E
- gleich/ähnlich Unterschied D bei K

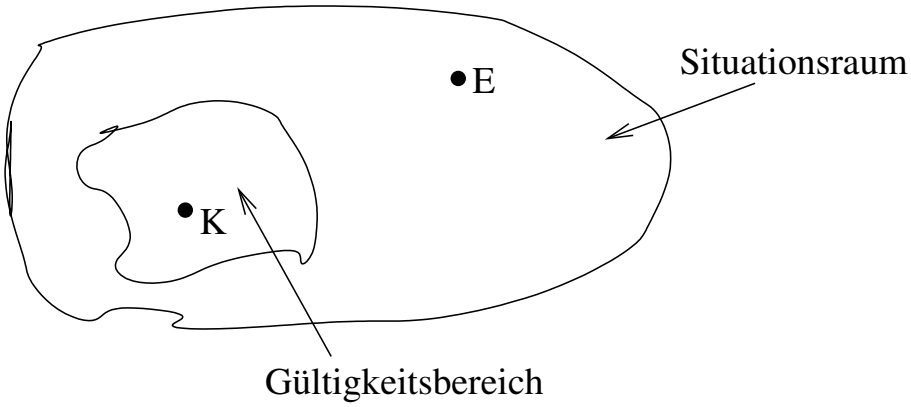
??? u.U. sehr fraglich

- wenn Verhaltensunterschiede existieren, dann ("höchstwahrscheinlich") unterschiedlich, je nach Situation
- zu erwarten
  - ▷ Bereich mit hinreichend kleinen Unterschieden

    Gültigkeitsbereich G
  - ▷ Bereich mit zu großen Unterschieden

    Bereich not G

Prinzipienskizze



und Frage damit:

gilt für (Experiment-Situation) E, dass

$E \in G$ oder $E \notin G$

nur beantwortbar (prinzipiell!), falls

- von Verhaltensunterschieden für gewisse Situation(en)
- auf Verhaltensunterschiede für andere Situationen geschlossen werden kann

also sicher **nicht allgemein beantwortbar**

Und damit (leider!) häufig Übergang notwendig

- von "Beweis der Gültigkeit"
- nach "Zuversicht in Gültigkeit"

bei jedem "vorhersagendem Modellieren"

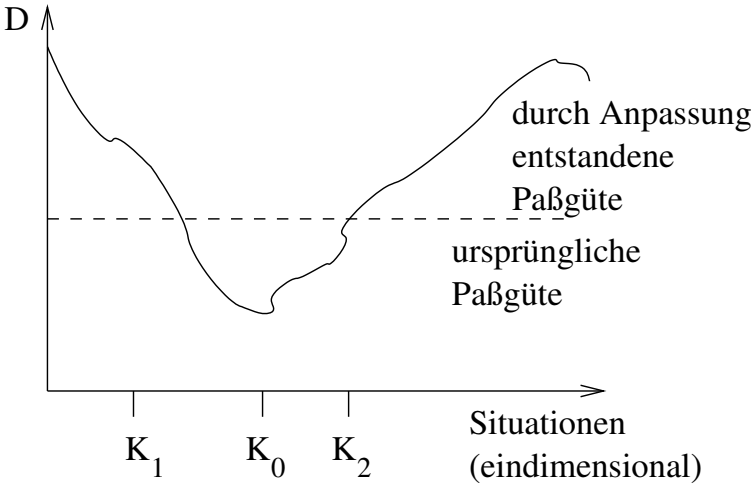
Schlimmer noch:

Kalibrierungsvorgang kann durchaus Problem

"Experimentsituation(en) gültig wiedergeben"

durch "Überanpassung" an spezielle Kalibriersituation verschärfen

Prinzipienskizze:



Reduktion der Gefahr einer Überanpassung

durch Verwendung mehrerer

Kalibriersituationen (Systemversionen, Umweltversionen)

zu erwarten

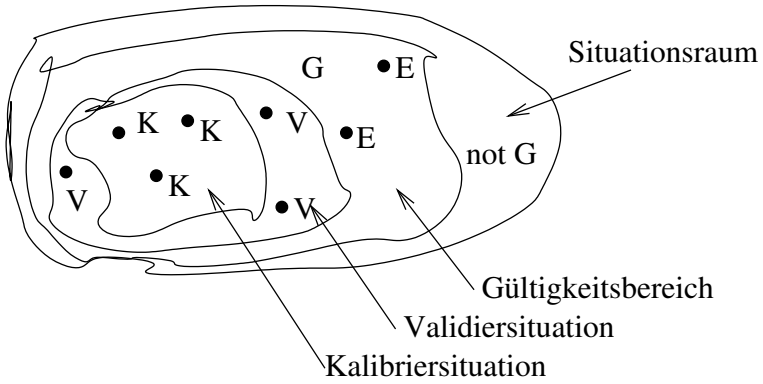
Dennoch sollte (formaler) gezeigt werden,

dass Kalibrierung nicht zu Überanpassung führt

Wie??

- Spezielle (unabhängige) Validierungsläufe für Situationen,
  die nicht zur Kalibrierung genutzt wurden
- Prüfung, ob Verhaltensunterschiede
  für diese (Validier-)Situationen
  "so etwa" wie für bekannte (Kalibrier-)Situationen

Prinzipienskizze:



Allerdings, ob eine Experiment-Situation E
im (unbekannten) Gültigkeitsbereich G liegt oder nicht
ist nach wie vor unbeweisbar (in irgendeinem strengen Sinne)
(bzw. nur retrospektiv beweisbar - dann aber "uninteressant" für "Vorhersagen")

Dennoch: "Vertrauen" wächst durch erfolgreiche Validierungen
(auch durch erfolgreiche retrospektive Validierungen)

Unterschiede auch
bzgl. "Interpolation/Extrapolation" von E.-Gültigkeit:
Vorsicht bei Extrapolation"weit weg" von K-/V-Bereichen

6.2 Messung von Verhaltensunterschieden

Bestimmung des Ausmaßes von
Verhaltensunterschieden Objekt/Modell war wesentlich bei

- **Kalibrierung**: Realitätstreue Modell verbessern durch Reduktion von Verhaltensunterschieden
- **Validierung**: Realitätstreue Modell bestätigen durch Überprüfung von Verhaltensunterschieden in (von Kalibrierung) unabhängigen Situationen

Kalibrierung/Validierung soll/muss Vertrauen in hinreichende Realitätstreue unterstützen, denn Modell erstellt für

- **Experimente**: System"güte" zu erhöhen durch Vergrößerung Verhaltensunterschiede "jetzt"$\rightarrow$ "später", bzw. "Plan 1"$\rightarrow$"Plan 2"$\rightarrow$ . . .

Bei Durchführung "Messen Verhaltensunterschiede" auftretende Fragen:

(a) mit Verhalten welchen Systems soll Verhalten Simulator verglichen werden?

(b) welche Verhaltens- Aspekte/-Beschreibungen sollen für den Vergleich gewählt werden?

(c) welche Vergleichs- Methoden/-Techniken sollen eingesetzt werden?

(a) zielt auf Festlegung "erwünschtes Simulatorverhalten"

(a1) Verhalten eines realen Objekt-Systems wäre ideale Basis

Idee realisierbar, falls

- System existiert und beobachtbar/meßbar ist
- Beobachtungen seines Verhaltens
  bei verschiedenen Umgebungssituationen,
  verschiedenen Struktursituationen (System-Versionen)

angestellt werden können

Dann Vorgehen: Man betreibe Realsystem und Modell

(in verschiedenen jeweils entsprechenden Versionen)

in identischer Umgebung ("Last")

(in verschiedenen, jeweils entsprechenden Umgebungen)

und vergleiche Verhalten

Da "Betreiben und Beobachten Realsystem

in verschiedenen Versionen/Umgebungen"

(zu) aufwendig (bis undenkbar) sein kann, ist Verfügbarkeit von ("historischen") Aufzeichnungen für verschiedene Versionen/Umgebungen erwünschte Fundgrube

In diesem Kontext liefert "trace driven simulation"

(Betreiben mit konkret aufgezeichneter, nicht stochastisch modellierter Last) gute Vergleichsbasis

(z.B. Bankschalter: Liste Ankunftszeiten der Aufträge)

Im Hinblick auf "retrospektive" (historische) Validierung: Nicht alle (Aufzeichnungen über) verfügbare Situationen für Kalibrierung "aufbrauchen" !

(a2) Verhalten analytischer Modelle: Ist dies ein Widerspruch?
(Simulation nur wenn's gar nicht anders geht!)

nicht unbedingt: Analytisches Modell stellt "Marginal"-Situation dar, für die analytisches Modell "spezifizierbar/lösbar", gleiche Situation sollte vom Simulator behandelbar sein und zu "ähnlichen/gleichen" Ergebnissen führen

(a3) weitere Möglichkeit deutlich "unterentwickelt",
sehr informell: "Turing's Test", "Delphi-Verfahren" benutzen Expertenmeinungen (z.B. Kann Experte Ausgabe Real-System/Simulator unterscheiden?) Befragung Experten sinnvoll, insbesondere zur Vertrauensbildung

(b) Verhaltens-Aspekte/-Beschreibung

Ziel ist es, System auf Basis bestimmten (ausgewählten) Leistungskriteriums zu beurteilen, zu verbessern

Daher offensichtlich Realitätstreue hinsichtlich dieses Leistungskriteriums (+ zugeordneten Maßes) am wesentlichsten

$\Rightarrow$ Einsatz dieses entscheidenden Leistungsmaßes auch für Kalibrierung/Validierung ("andere" nur zur Unterstützung)

(c) Vergleichs - Methoden / - Techniken

Fallunterscheidung:

(c1) reales Objektsystem / Simulator

Problemtyp: Vergleich zweier Stichproben

Entscheidung ob zwei Stichproben

"hinreichend ähnlich / unähnlich"

zunächst sicher:

Vergleich einfacher Charakteristik der Verteilungen

(z.B. Mittelwerte),

weiterhin:

Vergleich anderer Charakteristika

(immer auf Basis von Schätzern) denkbar, möglich

(c2) analytisches Modell / Simulator

Problemtyp: Vergleich analytische Verteilung

mit einer Stichprobe

Entscheidung ob Stichprobe

"hinreichend ähnlich/unähnlich"

wir kennen dazu bereits:

χ^2-Test, K.-S.-Test

Im folgenden also:

Vergleich von Stichproben

6.3 Testverfahren

Voraussetzungen:

Sei "Verhalten" beschrieben durch Zufallsvariable V

(z.B. Verweilzeit Kunden am Bankschalter

in der stationären Phase)

Zwei Stichproben liegen vor:

$v_R := (v_{R_1}, v_{R_2}, \ldots, v_{R_n})$

$v_S := (v_{S_1}, v_{S_2}, \ldots, v_{S_m})$

(z.B. "R" aus Beobachtung Realsystem, "S" aus Simulator)

Fragestellung: Sind beide Stichproben "hinreichend" ähnlich?

Man unterscheidet

- subjektive Verfahren (basierend im wesentlichen auf graphischer Darstellung)
- objektive Verfahren (basierend auf statistischen Tests)

Inspektionsansatz

Darstellung der Stichproben in graphischer Form

verwendete graphische Darstellungen:

- Histogramme
- Punktdiagramme
- Box-Plots

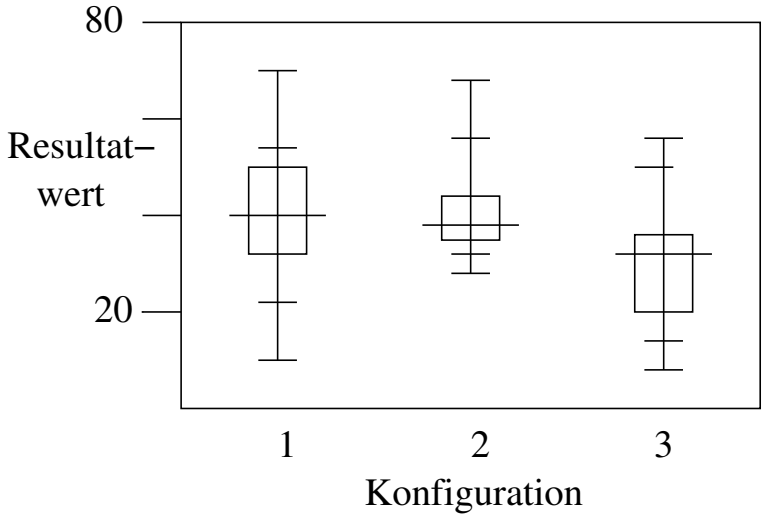
Box-Plots:

Verteilungen sind durch eine Vielzahl von Maßzahlen charakterisierbar

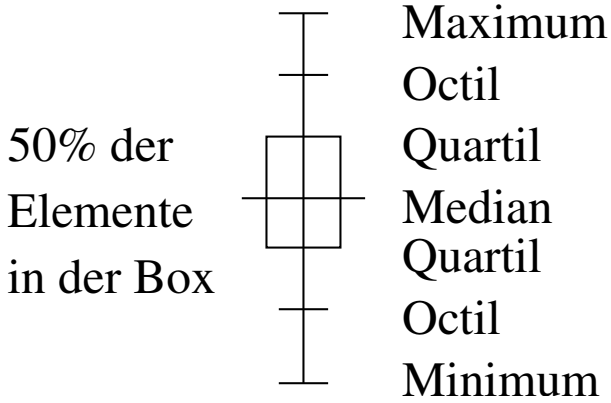
Wann sind zwei Stichproben ähnlich??

- Momente (niedriger Ordnung) beinhalten meist zu wenig Information
- Verteilungs- und Dichtefunktion sind nur schwer visuell vergleichbar

Box-Plots sind eine kompakte Darstellung wesentlicher Charakteristika, die einen einfachen Vergleich ermöglichen



Aufbau einer Box



Dargestellt werden

- alle Werte der Stichprobe
- Mittelwerte oder Varianzen über Teile der Stichprobe
- Korrelationsdiagramme

Anschließende Bewertung durch visuellen Vergleich

- Entwickler des Simulationmodells
- Experten im Anwendungsbereich
- Turing Test

Probleme bei diesem Vorgehen:

- graphische Darstellung visualisiert immer nur einen Teil der Information
- unterschiedliche Parametrisierung führt zu unterschiedlichen visuellen Eindrücken
- keine "objektivierbaren" Kriterien zu Entscheidungsfindung

aber Inspektionsmethode ist eine erster wichtiger Schritt und kann mit statistischen Auswahlverfahren kombiniert werden

wichtig ist insbesondere die Visualisierung von Korrelation in den Stichproben

Vergleich von Konfidenzintervallen

Ziel: Aussage ob Mittelwerte gleich oder unterschiedlich

Voraussetzungen

(deutlich restriktiver als bei der Inspektionsmethode):

- Werte jeder Stichprobe sind unabhängig identisch verteilt

<u>Ungepaarte Stichproben (Welch Verfahren)</u>

zusätzliche Voraussetzung:

- Werte zwischen den Stichproben sind unabhängig

keine Annahmen bzgl. identischer Varianz!

Berechne die Schätzer für Mittelwert und Varianz der Stichproben:

Schätze	durch
• μ_R	$\widetilde{\mu}_R = \frac{1}{n} \sum_{i=1}^{n} v_{R_i}$
• μ_S	$\widetilde{\mu}_S = \frac{1}{m} \sum_{i=1}^{m} v_{S_i}$
• σ_R^2	$\widetilde{\sigma}_R^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} v_{R_i}^2 - n \cdot \widetilde{\mu}_R^2 \right)$
• σ_S^2	$\widetilde{\sigma}_S^2 = \frac{1}{m-1} \left(\sum_{i=1}^{m} v_{S_i}^2 - m \cdot \widetilde{\mu}_S^2 \right)$

Differenz der Stichprobenmittelwerte:

$$\mu^*_{RS} = \mu^*_R - \mu^*_S$$

Standardabweichung der mittleren Differenz:

$$\sigma^*_{RS} = \sqrt{\frac{\sigma^{*2}_R}{n} + \frac{\sigma^{*2}_S}{m}}$$

Schätze die Zahl der Freiheitsgrade:

$$f^* = \frac{\left(\sigma^*_{RS}\right)^4}{\frac{1}{n-1}\left(\frac{\sigma^{*2}_R}{n}\right)^2 + \frac{1}{m-1}\left(\frac{\sigma^{*2}_S}{m}\right)^2}$$

Berechnung des Konfidenzintervalls

$$\mu^*_{RS} \pm t_{f^*,1-\alpha/2} \cdot \sigma^*_{RS}$$

f^* ist i.a. keine ganze Zahl $\Rightarrow$

runden oder

Werte der t-Verteilung durch Interpolation gewinnen

Auf Basis des Konfidenzintervalls kann entschieden werden,

- ob Stichprobenmittelwerte unterschiedlich (0 liegt innerhalb oder außerhalb des Konfidenzintervalls)
- wie groß die Abweichung der Mittelwerte (insbesondere im Vergleich zu den absoluten Werten) ist

Beispiel:

i	v_{R_i}	v_{S_i}	$v_{R_i} - v_{S_i}$
1	126.97	118.21	8.76
2	124.31	120.22	4.09
3	126.68	122.45	4.23
4	122.66	122.68	-0.02
5	127.23	119.40	7.83

Aus den Daten ergeben sich die Schätzer:

$\mu_R^* = 125.57$, $\mu_S^* = 120.59$, $\sigma_R^{*2} = 4.00$ und $\sigma_S^{*2} = 3.76$

sowie

$\mu_{RS}^* = 4.98$, $\sigma_{RS}^{*2} = 1.55$ und $f^* = 7.99$

Damit ergeben sich die Konfidenzintervalle:

- $4.98 \pm 1.24 \cdot 1.86 = [2.67, 7.29]$ zum Signifikanzniveau $\alpha = 0.1$ und
- $4.98 \pm 1.24 \cdot 3.36 = [0.81, 9.15]$ zum Signifikanzniveau $\alpha = 0.01$

$\Rightarrow$

0.0 ist in beiden Fällen nicht im Konfidenzintervall enthalten

damit kann die Hypothese gleicher Mittelwerte zum Niveau $\alpha = 0.01$ verworfen werden

Gepaarte Stichproben (Paired t-Konfidenzintervalle):

zusätzliche Voraussetzung:

- $n = m$ identische Anzahl von Beobachtungswerten

keine Annahmen bzgl. identischer Varianz oder Unabhängigkeit der Stichproben!

Berechne μ^*_{RS} wie vorher und

$$\overline{\sigma}^{*2}_{RS} = \frac{\sum_{i=1}^{n} \left((v_{R_i} - v_{S_i}) - \mu^*_{RS} \right)^2}{n \cdot (n-1)}$$

Berechnung des Konfidenzintervalls

$$\mu^*_{RS} \pm t_{n-1,1-\alpha/2} \cdot \overline{\sigma}^*_{RS}$$

Beispiel:

Daten wie im vorherigen Beispiel liefern

$$\mu^*_{RS} = 4.98 \text{ und } \overline{\sigma}^{*2}_{RS} = 2.44 \text{ bzw. } \overline{\sigma}^*_{RS} = 1.56$$

Damit ergeben sich die Konfidenzintervalle:

- $4.98 \pm 1.56 \cdot 2.13 = [1.66, 8.30]$ zum Signifikanzniveau $\alpha = 0.1$ und
- $4.98 \pm 1.56 \cdot 4.60 = [-2.20, 12.16]$ zum Signifikanzniveau $\alpha = 0.01$

$\Rightarrow$

Gleichheit der Stichprobenmittelwerte muss zum Niveau 0.1, nicht aber zum Niveau 0.01 abgelehnt werden

Vergleich der beiden Verfahren:

- Falls beide Verfahren anwendbar sind, ist nicht klar, welches von beiden schmalere Konfidenzintervalle liefert
- In der Regel wird nach "Datenlage" entschieden, welches Verfahren zur Anwendung kommt.

Statistische Testverfahren auf Basis von Hypothesen

Zahlreiche Verfahren zum Test der Hypothese $\mu_R = \mu_S$

Testverfahren

- verwerfen die Hypothese oder nehmen sie an, zu einem vorgegebenen Signifikanzniveau
- Entscheidung ist immer binär (annehmen, verwerfen)
- dadurch weniger differenzierte Information als bei der Berechnung von Konfidenzintervallen

allgemeine Beobachtung (für fast alle Testverfahren)

- für kleine Stichprobengrößen wird die Hypothese oft angenommen
- für große Stichprobengrößen wird die Hypothese fast immer abgelehnt
  (Simulation ist immer Approximation der Realität)

Zwei-Stichproben Test

Test zur Prüfung der Gleichheit zweier Stichproben

Oft verwendeter Test

(meist ohne die konkreten Voraussetzungen einzuhalten!)

Annahmen:

- v_R und v_S sind unabhängige Stichproben

  und wechselseitig unabhängig

  (alle Stichprobenvariablen V_{Ri} identisch unabh. vert.,

  alle Stichprobenvariablen V_{Si} identisch unabh. vert.,

  paarweise Unabhängigkeit aller V_{Ri}, V_{Si})
- beide Stichproben sind normalverteilt,

  mit identischer Streuung $\sigma_R = \sigma_S (=: \sigma)$
- beide Stichproben bestehen aus n Werten

Test-Hypothese

Stichproben besitzen identische Erwartungswerte: $\mu_R = \mu_S$

Alternativ-Hypothese(n)

entweder "zweiseitig": $\mu_R \neq \mu_S$

oder "einseitig": $\mu_R < \mu_S$ bzw. $\mu_R > \mu_S$

Test Algorithmus:

Schätze σ^2 durch $\widetilde{\sigma}^2 = \frac{1}{2} \cdot \left(\widetilde{\sigma}_R^2 + \widetilde{\sigma}_S^2\right)$

Bei zutreffenden Annahmen und zutreffender Hypothese ist

$D := \widetilde{\mu}_R - \widetilde{\mu}_S$

normalverteilt mit

Erwartungswert 0

Varianz $(\sigma_R^2 + \sigma_S^2)/n = 2\sigma^2/n$

und ist (demnach)

$D/\sqrt{2 \cdot \sigma^2/n}$

$N(0,1)$-verteilt sowie (Resultat aus der Statistik:)

$$T := D/\sqrt{2 \cdot \widetilde{\sigma}^2/n}$$

t-verteilt mit $2 \cdot n - 2$ Freiheitsgraden

falls (aus Stichproben errechneter) T-Wert

"zu groß" oder "zu klein" (zweiseitig),

"zu groß" bzw. "zu klein" (einseitig)

ist Testhypothese "gleiche Mittelwerte" zu verwerfen

zugunsten Alternativhypothese ($\neq$ bzw. $<$, $>$)

Also prüfe, ob

$$\frac{\mu_R^* - \mu_S^*}{\sqrt{2 \cdot {\sigma^*}^2/n}}$$

aus t_{2n-2}-Tafelwerten (bzgl. Niveau α) "herausfällt"

Anwendung auf das Beispiel liefert

$$D = 125.57 - 120.59 = 4.98$$

$$\sigma^{*2} = (4.00 + 3.76)/2 = 3.88$$

Für den Wert von T gilt:

$$T = \frac{4.98}{\sqrt{2 \cdot 3.88/5}} = \frac{4.98}{1.24} = 3.997$$

Für $\alpha = 0.01$ ist $t_{8,0.01} = 2.896 < 3.997$

damit ist die Hypothese gleicher Erwartungswerte
zum Niveau $\alpha = 0.01$ zu verwerfen

Insgesamt kritisch für Anwendung sind Annahmen:

- Normalverteilung
  ("parametrischer Test", d.h. Verteilungen als Annahme),
- Unabhängigkeit
  (ggf. höchstens Unkorreliertheit)
- identische Streuung
  (ähnlicher Test existiert für ungleiche Streuung)
- gleicher Stichprobenumfang
  (ähnlicher Test existiert für ungleiche Umfänge)

Mann-Whitney U-Test

Prüfung Gleichheit zweier unabhängiger Stichprobenverteilungen ohne Verteilungsvoraussetzung

(⇒ nicht parametrischer Test)

Werte innerhalb der Stichproben brauchen nicht unabhängig zu sein

Test-Hypothese:

Stichproben stammen aus identischer Verteilung

(hier formalisiert zu $P[V_R > V_S] = 0.5$)

Alternativ-Hypothese:

Stichproben stammen aus unterschiedlichen Verteilungen (hier formalisiert zu $P[V_R > V_S] \neq 0.5$ (zweiseitig) oder $P[V_R > V_S] > 0.5$ bzw. < 0.5 (einseitig))

Test-Algorithmus:

- vereinige Stichproben
- sortiere Werte der vereinigten Stichprobe merke dabei Herkunft der Werte
- bestimme für jeden Wert v_{S_i} der Stichprobe v_S die Anzahl der Werte v_{R_i} aus v_R, die kleiner als v_{S_i} sind
- addiere die Zahl der kleineren Werte für alle v_{S_i} diese liefert u^*

u^* ist eine Realisierung der ZV U

- für kleine Stichproben sind die kritischen Werte von U vertafelt
- für große Stichproben ist U normal-verteilt mit
  - Mittelwert $\mu_U = \frac{n \cdot m}{2}$
  - Varianz $\sigma_U^2 = \frac{n \cdot m \cdot (n+m+2)}{12}$
  - damit kann die Größe

    $$g^* = (u^* - \mu_U)/\sigma_U$$

    mit $N(0,1)$ Tafel einseitig/zweiseitig getestet werden

Beispiel:

i	v_{R_i}	v_{S_i}
1	9	6
2	11	8
3	15	10
4		13

v_R scheint größer zu sein $\Rightarrow$ teste $P[v_R > v_S] > 0.5$

Sortierte Stichprobe

6	8	9	10	11	13	15
S	S	R	S	R	S	R

$u^* = 3$

Tafelwert für kleine Stichproben liefert $P[U < 3] = 0.2$

$\Rightarrow$ Verwerfen würde einen Typ-1 Fehler von 0.2 bedeuten!

Wilcoxon Matched-Pairs Signed-Rank Test

Prüfung Gleichheit zweier "gepaarter" Stichprobenverteilungen

z.B. aus "trace-driven simulation" gleicher Kundenstrom

generiert die Stichproben

Test-Hypothese:

Stichproben stammen aus identischer Verteilung

(hier formalisiert zu $P[V_R > V_S] = 0.5$)

Alternativ-Hypothese:

Stichproben stammen aus unterschiedlichen Verteilungen

in der Regel einseitig $P[V_R > V_S] > 0.5$ bzw. < 0.5

Test-Algorithmus:

- bilde Differenzen $d_i = v_{R_i} - v_{S_i}$
- sortiere Differenzen
- ordne Differenzen Ränge zu
- ordne den Rangzahlen Vorzeichen der Differenzen zu
- summiere positive Rangzahlen zu s^+ negative zu s^-

Testgrößen S^+ und $|S^-|$ sollten bei zutreffender Test-Hypothese "annähernd gleiche" Werte aufweisen

kritische Werte für $T = \min(S^+, |S^-|)$ sind

- für kleine Stichproben vertafelt
- für große Stichproben ist T normalverteilt mit
  - Mittelwert $\mu_T = \frac{n \cdot (n+1)}{4}$
  - Varianz $\sigma_T^2 = \frac{n \cdot (n+1) \cdot (2n+1)}{24}$
  - damit kann die Größe

    $g^* = (t^* - \mu_T)/\sigma_T$

    mit $N(0,1)$ Tafel einseitig getestet werden

Beispiel:

v_R	v_S	d	Rang	mit Vorzeichen	
82	63	19	7	7	
69	42	27	8	8	
73	74	-1	1		-1
43	37	6	4	4	
58	51	7	5	5	
56	43	13	6	6	
76	80	-4	3		-3
65	62	3	2	2	
			Summe	32	-4

es ergibt sich $t = \min(32, |-4|) = 4$ bei $n = 8$

Resultate aus der Tafel

- Gleichheit bei $t \leq 4$ $(\alpha = 0.025)$ zu verwerfen zugunsten v_R größer

- Gleichheit bei $t = 4$ $(\alpha = 0.01)$ nicht verwerfen