8. Kalibrierung und Validierung

Erinnerung an Abschn.1

- Modelle erstellt,
  - um Experimente
    mit realen (Objekt-) Systemen zu vermeiden
  - um Aussagen über (potentielle) Objekt-Systeme
    zu erhalten, die (noch) nicht existieren
  - um **Vorhersagen** machen zu können

- Folgerungen aus Modellanalysen
  (hier: Simulations-Experimenten)
  sollten (da als relevant für Realität verwendet)
  weitestgehend gleich sein zu Folgerungen, die aus
  entsprechenden (potentiell, irgendwann möglichen)
  Objekt-Analysen / Objekt-Experimenten
  gewonnen

- Gleiche Beurteilungs-, Folgerungs- Schemata
  vorausgesetzt,
  sind Folgerungen (zumindest) dann identisch,
  wenn unmittelbare Resultate / Beobachtungen
  von Simulations- und Objekt- Experimenten identisch

Identität Resultate zu erwarten ??

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

Objekt- und Modell-System nicht identisch,
⇒ Verhaltensunterschiede zu erwarten

vernünftige Definition von **Realitätstreue**
Gültigkeit
validity

- müßte aufbauen auf Maß 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
  (nicht zu unterschiedlichen Folgerungen führen)

Ursachen für Verhaltensunterschiede:

- strukturelle Unterschiede
  - Abstraktion, Aggregierung (beabsichtigt!)
  - Ungenauigkeiten (bugs), (unbeabsichtigt!)
    Fehler (faults) (unbeabsichtigt!)

- Parameterwerte
  - deterministische Werte + Abhängigkeiten
  - Verteilungsfunktionen
  - …

Uneinheitliche Terminologie, hier strikt unterschieden:

- **Verifikation**:
  Bestätigung aller Modell-Eingangsgrößen / -Annahmen
  incl. struktureller Annahmen,
  … ,
  Programmverifikation,
  Parameterwerten

- **Validierung**:
  Bestätigung der Modell-Resultate

(zu feineren Unterscheidungen vgl LaKe91, KnAr93
incl "confidence assessment methodology")

Sicher (hoffentlich) größte Mühe gelegt auf
"gute" Wahl Eingangsgrößen (Hypothesen, …, Parameter)

"Hoffnung" auf gültiges Modell damit zweifellos steigend,
"Garantie" für gültiges Modell aber nicht gegeben
("positivistischer" Blick sogar: Ergebnisse ok, alles ok)

Wenn Verhaltensunterschiede Modell / Realität entdeckt,
Versuch natürlich (und legitim), Modell zu ändern,
um Verhaltensunterschiede (Hoffnung:)
zu beseitigen, zu reduzieren:

Reduktion "Verhaltensunterschiede"
ist Reduktion eines $D(V_R,V_S)$

Vorgang "ausgelöst von (zu) großem D
Modell ändern mit Ziel D-Reduktion"
heißt **Kalibrierung**

Kalibrierung schließt also (zwangsläufig) ein:

- Identifikation Ursachen Verhaltensunterschiede

- entsprechende, gezielte Änderungen Modell
  - Struktur: "Code"-Änderung
  - Parameter / statische Attribute: "Werte"-Änderung

Bei stochastischen Modellen resultieren
zwei spezifische statistische Problem-Typen:

- Auswahl eines aus mehreren alternativen Modellen
  $(S_1, S_2, \ldots, S_k)$
  auf Basis zugehöriger
  $(D(V_R, V_{Si}) \; ; \; i=1,2,\ldots,k)$

  wo D (wie gewohnt) Zufallsvariable
  oder gar stochastischer Prozeß
  und Unterschiede der D's müssen (folglich)
  "das normale Maß der Schwankung" übersteigen,
  Unterschiede müssen **signifikant** sein

  Aufgabe: **Tests auf Signifikanz**

- "tuning" durch Parameter(wert)veränderungen,
  Suche nach Parametervektor $\underline{p}_{opt}$ derart, daß
  $$\min_{\underline{p}} D\left(V_R, V_S(\underline{p})\right)$$
  erreicht

  Aufgabe: **stochastische Optimierung**

Methodisch gesehen, ist
- Kalibrieren stochastischen Simulators

identisches Problem wie
- Experimentieren mit stochastischem Simulator
(bzw: stochastischem System)

Für gegebenes Realsystem mit Verhaltensgüte V_R
wird (über Experimente) versucht,
System mit Verhaltensgüte V_E zu finden derart, daß

$D^*(V_R,V_E)$

maximal (Veränderung hin zu "bester" Alternative);
als Möglichkeiten dafür erneut
- Ausprobieren struktureller Alternativen
- tuning von System-Parametern

Bei 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:
weitere prinzipielle Klärung "Kalibrieren"

- angenommen, mittels Kalibrierung sei
unmittelbares Ziel erreicht:
D ist unter (erträgliches) D_{ertr} gesunken

- ist damit auch inhaltliches Ziel erreicht:
Folgerungen aus Objekt- und Modell-Experimenten
sind gleich

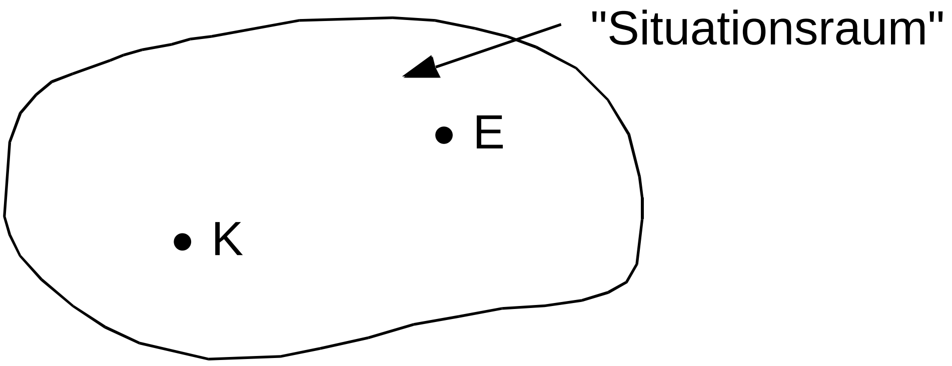
??

Was, genau, wurde getan?

Simulator-Verhalten wurde (mittels Änderungen)
einem Objektsystem-Verhalten "hinreichend" angepaßt
für **einen** Zustand der Umwelt (Last)
für **einen** Zustand des Systems (Konfiguration, Systemparameter)
dh für **eine Situation**

Als "Zweck" des Simulators letztlich angepeilt:
Experimentieren / Analysieren für **andere** Situationen;
über **diese** (beim Kalibrierungsvorgang eingesetzte) Situation besteht ja Klarheit !

Prinzipskizze:



und Frage damit:

ist Unterschied D bei E ("eigentlich" interessierend)
wie Unterschied D bei K (kraft Kalibrierung gesichert)

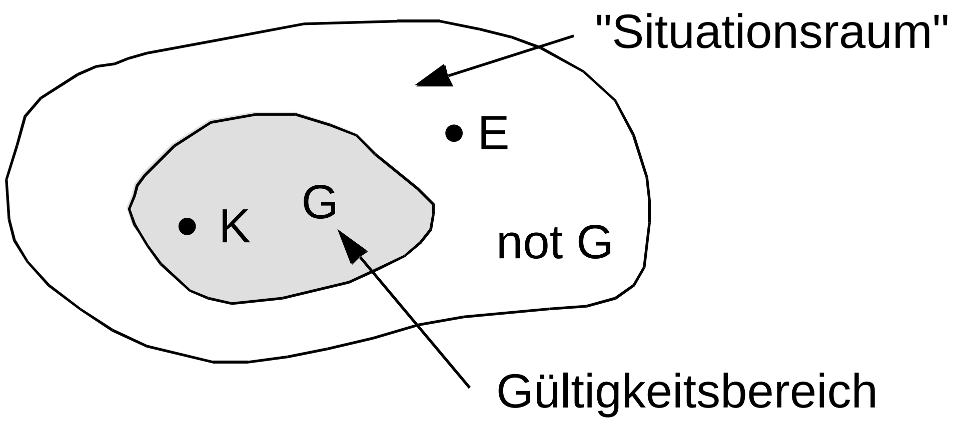
??

uU 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

Prinzipskizze:



und Frage damit:

gilt für (Experiment-Situation) E, daß
E G
oder E not G ??

nur beantwortbar (prinzipiell!), falls
- von Verhaltensunterschieden für gewisse Situation(en)
- auf Verhaltensunterschieden für andere Situation(en)

geschlossen werden könnte

also sicher **nicht allgemein beantwortbar**

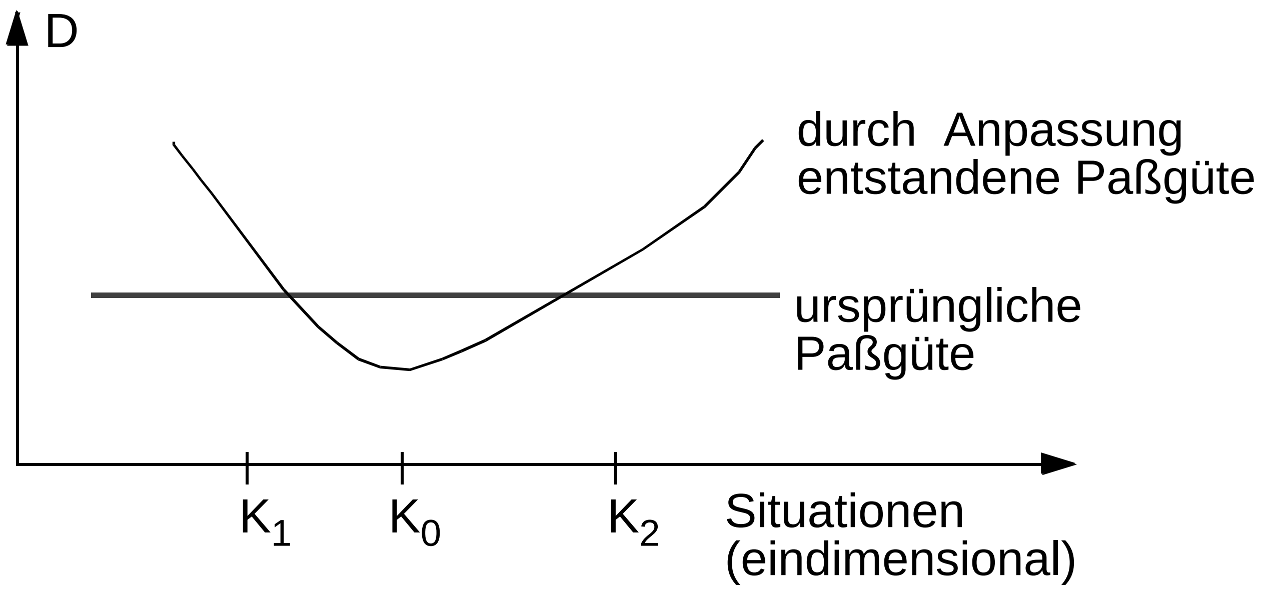
Und damit (leider !) häufig
Übergang notwendig
von "Beweis der Gültigkeit"
nach "Zuversicht in Gültigkeit"
bei jedem "vorhersagenden Modellieren" (LaKe91, KnAr93)

Schlimmer noch:
Kalibrierungsvorgang kann durchaus Problem

Experimentsituation(en) gültig wiedergegeben ?

durch "Überanpassung" an spezielle Kalibriersituation verschärfen

Prinzipskizze:

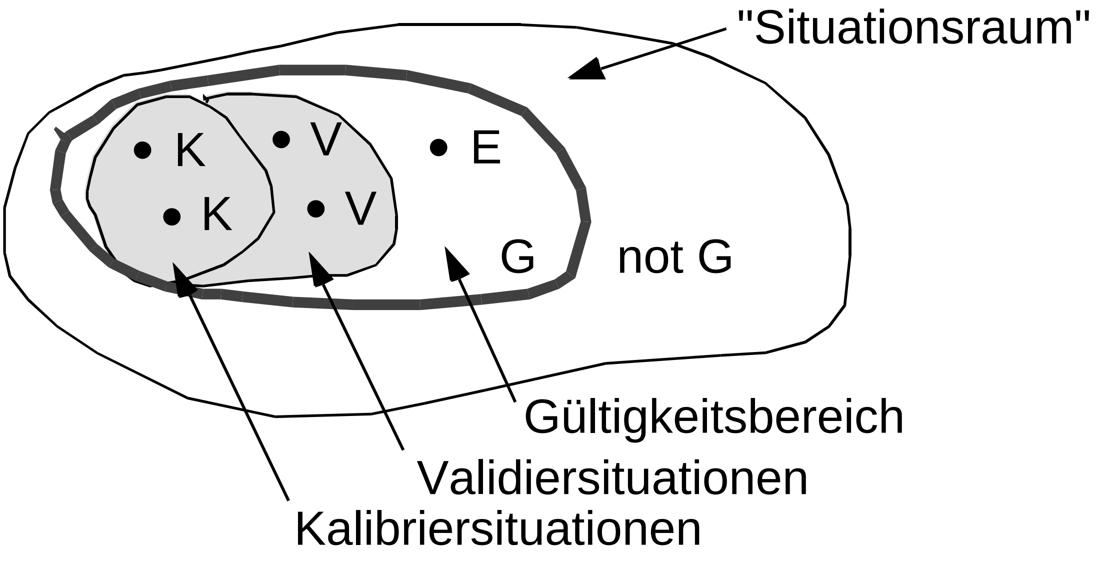


Reduktion der Gefahr einer Überanpassung
durch Verwendung mehrerer
Kalibriersituationen (Systemversion, Umweltversion)
zu erwarten

Dennoch sollte (formaler) gezeigt werden,
daß Kalibrierung nicht zu Überanpassung geführt hat

Wie ??
- Spezielle (unabhängige) **Validierungsläufe**
  für Situationen, die nicht zu Kalibrierung benutzt
- Prüfung, ob Verhaltensunterschiede
  für diese (Validier-)Situationen
  "so etwa" wie für bekannte (Kalibrier-)Situationen

Prinzipskizze:



Allerdings:
Ob eine Experiment-Situation E
im (unbekannten) Gültigkeitsbereich G liegt oder nicht
ist nach wie vor unbeweisbar
(in irgendeinem strengen Sinn)
(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 !

8.1 Zur Messung von Verhaltensunterschieden

Bestimmung des Ausmaßes von
Verhaltensunterschieden Objekt / Modell
war wesentlich bei

- **Kalibrierung**: Realitätstreue Modell zu verbessern durch Reduktion v. Verhaltensunterschieden

- **Validierung**: Realitätstreue Modell zu bestätigen durch Überprüfung v. Verhaltensunterschieden in (von Kalibrierung) unabh. Situationen

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

- **Experimentieren**: System"güte" zu erhöhen durch Vergrößerung Verhaltensunterschiede "jetzt" → "später", bzw Plan 0→ Plan 1 → …

Bei Durchführung "Messen Verhaltensunterschiede"
auftretende Fragen:

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

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

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

(a) zielt auf Festlegung "erwünschtes Simulatorverhalten"

- 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:
  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 /
  in verschiedenen Umgebungen"
  (zu) aufwendig (bis: undenkbar) sein kann,
  ist Verfügbarkeit von ("historischen")
  Aufzeichnungen für versch. Versionen/Umgeb'gen
  erwünschte "Fundgrube"

  In diesem Kontext liefert
  "trace-driven simulation"
  (Betreiben mit konkret aufgezeichneter,
  nicht stochastisch modellierter Last)
  gute Vergleichsbasis
  (zB Bankschalter: Liste Ankunftszeiten, Aufträge)

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

- Verhalten analytischer Modelle
  ist Widerspruch ?
  (Simulation nur wenn's gar nicht anders geht !)

  nicht notwendig:
  analytisches Modell stellt "Marginal"-Situation dar,
  für die analytisches Modell "lösbar",
  gleiche Situation sollte aber auch
  von Simulator behandelbar sein

- weitere Möglichkeiten deutlich "unterentwickelt",
  sehr informell:
  "Turing's Test", "Delphi-Verfahren", "face validity"
  benutzen Expertenmeinung
  (Befragung Experten aber sicherlich sehr sinnvoll:
  Zuversicht / Vertrauen !)

(b) Verhaltens -Aspekte / -Beschreibungen

Ziel ist, System auf Basis
bestimmter (bewußt gewählter) Leistungskriterien
zu beurteilen, zu verbessern

Daher offensichtlich Realitätstreue hinsichtlich
dieser Leistungskriterien (+ zugeordneter Maße)
am wesentlichsten

Demnach ratsam:
Einsatz des entscheidenden Leistungsmaße
auch für Kalibrierung / Validierung
("andere" nur zur Unterstützung)

(c) Vergleichs -Methoden / -Techniken

Fallunterscheidung

- reales Objektsystem / Simulator
  Problemtyp: Vergleich zweier Stichproben
  Entscheidung, ob zwei Stichproben
  "hinreichend ähnlich / unähnlich"

  zunächst sicher:
  Vergleich einfacher Charakteristika
  der Verteilungen (zB Mittelwerte)

  weiterhin: Vergleich anderer Charakteristika
  (immer auf Basis von Schätzern)
  denkbar, möglich

- analytisches Modell / Simulator
  Problemtyp: Vergleich analytische Verteilung
  mit einer Stichprobe
  Entscheidung, ob Stichprobe
  "hinreichend ähnlich / unähnlich"

  wir kennen dazu bereits (Abschn. 5.3):
  χ^2-Test,
  K-S-Test

Im Folgenden also:
Vergleich Stichproben

8.2 Der zwei-Stichproben-t-Test
(zur Prüfung der Gleichheit
zweier Stichprobenmittelwerte)

Viel verwendeter Test
(oft ohne die Voraussetzungen einzuhalten !)

Sei "Verhalten" beschrieben durch Zufallsvariable V
(zB Verweilzeit Kunden im Bankschalter-System,
in dessen stationärer Phase)

Mögen vorliegen zwei Stichproben

$$\underline{v}_R := (v_{R1}, v_{R2}, \ldots, v_{Rn})^T$$
$$\underline{v}_S := (v_{S1}, v_{S2}, \ldots, v_{Sn})^T$$

(zB "R" aus Beobachtung Realsystem, "S" aus Simulator)

Annahmen
- $\underline{v}_R$ und $\underline{v}_S$ sind unabhängige Stichproben
und wechselseitig unabhängig
(alle Stichprobenvariablen V_{Ri} identisch unabh. verteilt,
alle Stichprobenvariablen V_{Si} identisch unabh. verteilt,
paarweise Unabhängigkeit aller V_{Ri}, V_{Si})
- beide Stichproben sind normalverteilt,
mit identischer Streuung $\sigma_R = \sigma_S$ (=: σ)

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	durch
μ_R	$\tilde{\mu}_R = \frac{1}{n} \sum_i V_{Ri}$
μ_S	$\tilde{\mu}_S = \frac{1}{n} \sum_i V_{Si}$
σ_R^2	$\tilde{\sigma}_R^2 = \frac{1}{n-1} \left(\sum_i V_{Ri}^2 - n \tilde{\mu}_R^2 \right)$
σ_S^2	$\tilde{\sigma}_S^2 = \frac{1}{n-1} \left(\sum_i V_{Si}^2 - n \tilde{\mu}_S^2 \right)$
σ^2	$\tilde{\sigma}^2 = \frac{1}{2} \left(\tilde{\sigma}_R^2 + \tilde{\sigma}_S^2 \right)$

Bei zutreffenden Annahmen und zutreffender Hypothese ist

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

normalverteilt mit

Erwartungswert 0

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

und ist (demnach)

$$D \Big/ \sqrt{2\,\sigma^2/n}$$

N(0,1)-verteilt

sowie (Resultat aus der Statistik:)

$$T := D \Big/ \sqrt{2\,\tilde{\sigma}^2/n}$$

t-verteilt mit 2n-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 (≠ bzw >, <)

kurz:
prüfe, ob

$$\frac{\mu^*_R - \mu^*_S}{\sqrt{2\,\sigma^{*2}/n}}$$

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

Insgesamt kritisch für Anwendung sind die Annahmen:

- Normalverteilung
  ("parametrischer Test": Verteilungen als Annahme)

- Unabhängigkeit (ggf höchstens: Unkorreliertheit)

- identische Streuung
  (ähnlicher Test existiert für ungleiche Streuungen)

- gleicher Stichprobenumfang
  (ähnlicher Test existiert für ungleiche Umfänge)

vgl auch (breitere Diskussion): HaEK82

8.3 Der Mann-Whitney U-Test
(zur Prüfung der Gleichheit
zweier Stichprobenverteilungen)

"nichtparametrischer" Test
(ohne Verteilungsvoraussetzungen)
zu nichtparametrischen Tests
vgl Sieg56, BüTr78, HaEK82
in verschiedenen Varianten (hier i.w. "Siegel")

Sei "Verhalten" beschrieben durch Zufallsvariable V
(zB Verweilzeit Kunden im stationären Bankschalter)

Mögen vorliegen zwei Stichproben
$\underline{v}_I := (v_{I1}, v_{I2}, \ldots, v_{In})$
$\underline{v}_J := (v_{J1}, v_{J2}, \ldots, v_{Jm})$
(zB I=R aus Beobachtung Realsystem, J=S aus Simulator)

Annahmen
- Wertemengen der Stichproben unterliegen "zumindest" "Ordinalskala" (strenge Ordnungsrelation existiert)
- $\underline{v}_I$ und $\underline{v}_J$ voneinander unabhängige Stichproben, (nicht notwendig in sich unkorreliert !)
- im Vergleich: keine Annahmen über Verteilungstyp, Stichprobenumfänge

Test-Hypothese
Stichproben stammen aus identischer Verteilung;
formalisiert zu: $P[V_{I\bullet} > V_{J\bullet}] = 1/2$

Alternativ-Hypothese(n)
entweder "zweiseitig": $P[V_{I\bullet} > V_{J\bullet}] \neq 1/2$
oder "einseitig": $P[V_{I\bullet} > V_{J\bullet}] > 1/2$ (bzw $< 1/2$)
ist Normalfall, eine Stichprobe (zB I)
als uU "stochastisch größer" vermutet:
v "Wertemenge von V": $P[V_{I\bullet} \leq v] < P[V_{J\bullet} \leq v]$

Test-Algorithmus

- vereinige Werte aus beiden Stichproben,
  sortiere gesamte Menge von Werten aufsteigend,
  notiere aber Stichprobenherkunft der Werte

  Beispiel:

$\underline{v}_I$:	9	11	15				
$\underline{v}_J$:	6	8	10	13			

  Anschein: "I" ist größer
  daher Alternativhypothese: $P[V_{I\bullet} > V_{J\bullet}] > 1/2$

  sortierte gesamte Menge der Werte:

	6	8	9	10	11	13	15
Herkunft:	J	J	I	J	I	J	I

- zähle, für jeden J-Wert,
  Anzahl vorausgehender I-Werte
  und addiere Anzahlen zu Summe =: u*
  (Realisierung einer Test-ZV U)

  Beispiel:

Zählung:	0	0		1		2	

  Summe: u* = 3

- Testgröße U sollte kleinen Wert u* aufweisen,
  falls "I" wirklich größer ist
  Wahrscheinlichkeiten, mit denen
  (bei zutreffender Testhypothese)
  $U \leq$ "wert" (=u*)

  ausfällt, sind vertafelt
  (verschiedene Tafeln für
  "sehr kleine" / "kleine" / "verbleibende"
  Stichprobenumfänge)

Beispiel:

n_1 := Umfang kleinere Stichprobe = 3
n_2 := Umfang größere Stichprobe = 4

aus Tafel "n_2=4" für sehr kleine Stichproben:
$P[U \leq 3\ (=u^*)] = 0.2$
kein Anlaß, Testhypothese zu verwerfen
(falls nicht Typ1-Fehlerwahrsch. 0.2 toleriert)

- (Hinweis: falls Testwert u* nicht in Tafel zu finden, wurde Alternativhypothese falsch gestellt; Prüfung von $u' := n_1 n_2 - u^*$ testet entgegengesetzte -einseitige- Alternativhypothese)

- Tafeln sehr kleine Stichproben: $3 \leq n_2 \leq 8$

Tafeln kleine Stichproben: $9 \leq n_2 \leq 20$
(ein Beispiel: $n_1=6$, $n_2=13$, $u^*=19$ führt gemäß Tafel zu Verwerfen zum Niveau 0.05)

Für größere Stichproben ($n_2>20$) ist U in guter Näherung normalverteilt mit

$$\mu_u = \frac{n_1 n_2}{2}$$

$$\sigma_u^2 = \frac{n_1 n_2 \left(n_1+n_2+1\right)}{12}$$

und ("wie gehabt") Größe

$$g^* := (u^*-\mu_u)/\sigma_u$$

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

- rechnerisch effizientere Methode zur Berechnung von u* mittels Zuweisung von "Rängen" (ranking) zu den kombinierten Wertemengen beider Stichproben und nachfolgende Addition der Ränge je Stichprobe

im Beispiel:

					Stichproben-umfang	Rang-summen
$\underline{v}_I$:	9	11	15		$n_I = 3$	
Rang	3	5	7			$r_I = 15$
$\underline{v}_J$:	6	8	10	13	$n_J = 4$	
Rang	1	2	4	6		$r_J = 13$

zu errechnender Testgrößenwert ergibt sich zu

$$u^* = n_I\, n_J + \frac{n_I\left(n_I+1\right)}{2} - r_I \qquad (= 3)$$

bzw bei umgekehrter Alternativhypothese zu

$$u' = n_I\, n_J + \frac{n_J\left(n_J+1\right)}{2} - r_J \qquad (= 9)$$

(einfach: kleineres der Ergebnisse verwenden)

8.4 Der Wilcoxon Matched-Pairs Signed-Rank Test
(zur Prüfung der Gleichheit der Verteilungen
zweier "gepaarter" Stichproben)

Sei "Verhalten" beschrieben durch Zufallsvariable V
(zB Verweilzeit Kunden im stationären Bankschalter)

Mögen vorliegen zwei "gepaarte" Stichproben

$\underline{v}_I := (v_{I1}, v_{I2}, \ldots, v_{In})$

$\underline{v}_J := (v_{J1}, v_{J2}, \ldots, v_{Jn})$

(zB "trace-driven" Simulation,
gleicher Kundenstrom für Modellvarianten I und J;
natürliche Paarung: "i-ter Kunde bei Behandl'gen I, J")

Annahmen
- Wertemengen der Differenzen gepaarter Werte unterliegen zumindest Ordinalskala
- "je nach Literaturstelle" auch:
  ($\underline{v}_I$ und $\underline{v}_J$ unabhängige Stichproben,
  symmetrische Verteilungen,
  Verteilungsunterschiede nur in "Lokation")
- im Vergleich: keine Annahmen über Verteilungstyp

Test-Hypothese
Stichproben stammen aus identischer Verteilung;
formalisiert zu: $P[V_{I\bullet} > V_{J\bullet}] = 1/2$

Alternativ-Hypothese
generell "einseitig": $P[V_{I\bullet} > V_{J\bullet}] > 1/2$ (bzw $< 1/2$)
zweiseitig aber möglich!

Test-Algorithmus

- bilde Differenzen d_i der gepaarten Werte,
  sortiere Differenzbeträge,
  ordne den Differenzen Ränge zu,
  ordne den Rangzahlen Vorzeichen der Differenzen zu,
  summiere positive Rangwerte zu s^+, negative zu s^-

  Beispiel:

v_I	v_J	d	Rang Absolutwerte	samt 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	
			Summen	32	-4

- Testgrößen S^+, $|S^-|$ sollten
  bei zutreffender Testhypothese
  "annähernd gleiche" Werte aufweisen

  Wahrscheinlichkeiten, mit denen (unter Testhypothese)
  $T := \min(S^+,|S^-|) \leq$ "wert"
  ausfällt, sind vertafelt

  im Beispiel:
  t = 4, n = 8, einseitiger Test
  aus Tafel: Gleichheit bei $t \leq 4$ (α=0.025) zu verwerfen
  zugunsten "I" ist größer;
  Gleichheit bei t=4 (α=0.01) nicht verwerfen

- für große Stichproben (n>25):

  T in guter Approximation normalverteilt mit

$$\mu_T = \frac{n\,(n+1)}{4}$$
$$\sigma_T^2 = \frac{n\,(n+1)\,(2n+1)}{24}$$

  und ("wie gehabt") Größe

$$g := (t-\mu_T)/\sigma_T$$

  mittels N(0,1)-Tafel prüfbar