FÖRETAGSEKONOMISKA INSTITUTIONEN

FE rapport 2004-406

Modeller för internprissättning: Re-analys av experimentella tester

Björn Lantz

Handelshögskolan
VID GÖTEBORGS UNIVERSITET
Företagsekonomiska institutionen

Modeller för internprissättning:
Re-analys av experimentella tester

Abstract: Dejong et.al. (1989) and Avila & Ronen (1999) conduct experimental tests of various transfer pricing mechanisms. The resulting quantity data is transformed to efficiency measures which the authors analyse primarily through t-tests and analyses of variance. The problem in both experiments is that the data cannot be seen as samples from approximate normal distributions. This means that non-parametric methods should be used to analyse the data in order to get valid conclusions. In this paper, the data sets from both experiments are re-analysed with non-parametric methods. Of course, there would be no story if the conclusions from these analyses did not differ with respect to the original sources. But they do. The overall picture is that there are significant differences between the tested mechanisms in more ways than the original authors found.

Keywords: Transfer pricing, truth-telling mechanism, experiment

JEL-code: C70, C91, D21, D82

Handelshögskolan vid Göteborgs universitet
School of Economics and Commercial Law at Göteborg University
Företagsekonomiska institutionen
Department of Business Administration
Box 610, 405 30 Göteborg
Björn Lantz, tel. 031-773 5245, e-mail: bjorn.lantz@handels.gu.se

SAMMANFATTNING

För att analysera mänskligt beteende under olika former av ekonomiska styrmodeller används ofta experimentella studier. På internprissättningsområdet märks två sådana studier, genomförda av Dejong m.fl. (1989) och Avila & Ronen (1999). Båda experimenten baserades på att försökspersoner parades ihop två och två, som köpare respektive säljare, varefter de under olika styrmodeller och asymmetrisk information fick förhandla om pris och kvantitet i en situation karakteriserad av bilateralt monopol. Varje par upprepade förhandlingen över ett antal perioder. Kvantiteterna räknades sedan om till effektivitetsmått som användes för att jämföra olika styrmodeller.

I båda studierna användes klassiska t-tester och variansanalyser för att analysera datamängderna som experimenten resulterade i. Parametriska tester av detta slag kräver att datan kommer från approximativt normalfördelade populationer. Framför allt t-testen är dock tämligen robust för måttliga avvikelser från normalfördelningsantagandet. I de båda aktuella studierna visar det sig emellertid vid närmare granskning att det är helt orimligt att anta normalfördelning för datamängderna, vilket innebär att de resultat som åstadkoms i båda fallen kännetecknas av dålig validitet.

I denna rapport re-analyseras därför experimentdatan från de båda studierna med framför allt icke-parametriska metoder. Det visar sig då att det finns betydligt fler signifikanta skillnader mellan de olika styrmekanismerna i experimenten än vad de ursprungliga författarna hittade. Exempelvis är marginalkostnadsmodellen signifikant effektivare än fri förhandling även när de inledande periodernas större andel helt misslyckade förhandlingar exkluderas. En annan skillnad var att subventionerad intern handel med straff-faktor visade sig leda till signifikant högre effektivitet är om straff-faktorn saknades, detta oavsett omständigheterna i övrigt.

INLEDNING

Experimentella tester av mänskligt beteende i syfte att verifiera eller förkasta olika slag av ekonomisk-teoretiska modeller är vanligt förekommande.[1] Sådana tester utförs vanligen i ”laboratoriemiljö”, vilket innebär att forskaren kontrollerar alla parametrar i den omgivning där försökspersonernas agerande observeras. Motpolen är ”fältstudier” där endast få aspekter rörande de observerade personernas agerande kan kontrolleras. Det faktum att alla parametrar bestäms av forskaren kan sägas vara både fördelen och nackdelen med experimentella studier. Under förutsättning att den huvudsakliga konfigurationen är densamma kan man förvänta sig samma typ av mänskligt beteende i olika situationer. Om någon viktig parameter i en viss situation däremot skiljer sig från den konfiguration som använts tidigare kan det mänskliga beteendet vara svårare att förutsäga med utgångspunkt från det experimentella resultatet.

Resultatet av ett experiment är vanligen en serie mätvärden för en eller flera variabler. Med hjälp av statistiska tester kan forskaren analysera hur utfallet för dessa variabler påverkas av olika sätt att konfigurera de parametrar som försökspersonernas omgivning utgörs av. Om forskare i olika experiment kan dra slutsatsen att närvaron av och/eller storheten hos en viss parameter, *ceteris paribus*, påverkar det mänskliga beteendet på ett visst sätt så kan denna kunskap utnyttjas i praktisk ekonomisk styrning när man vill uppnå en viss påverkan på det mänskliga beteendet.

Ett område inom företagsekonomiämnet där experimentella tester av styrmodeller har utförts är internprissättning (Dejong m.fl. 1989 och Avila & Ronen 1999). Det fundamentala styrproblemet med internprissättning när ingen extern marknad finns är att förmå köparen och säljaren att självständigt finna den lösning som är sammanlagt vinstmaximerande (se t.ex. Lantz, 2000). Svårigheten ligger i att båda parter har incitament att försöka utöva sin monopolmakt mot varandra vilket innebär att båda i utgångsläget strävar efter att transferera en kvantitet som är lägre än den optimala sett i det sammanlagda perspektivet. Det finns ingen teoretisk jämvikt mellan parterna i ett sådant bilateralt monopol och problemet är att bestämma en styrmodell som är incitamentskompatibel, d.v.s. leder till att parterna i en självständig förhandling kan hitta den sammanlagt optimala kvantiteten.

Dejong m.fl. (1989) och Avila & Ronen (1999) hade i stort sett samma experimentella upplägg i sina respektive studier. Båda använde t.ex. studenter som försökspersoner, vilka fick betalt för sin medverkan i förhållande till den ”vinst” de skapade åt sig själva under experimentet. Vidare var varje förhandling skriftlig och anonym och baserades på asymmetrisk fördelad information, varje förhandlingssituation upprepades ett antal gånger för att ge möjlighet till ”inlärning” och resultaten utvärderades med statistiska metoder. Det som primärt skilde studierna åt var de styrmodeller som testades.

Dejong m.fl. testade i sin experimentella studie tre olika styrmodellers påverkan på aktörerna i ett bilateralt monopol: Ronen & McKinneys (1970) modell med subventionerad intern handel, Hirshleifers (1956) modell med strikt marginalkostnadsprissättning samt fri förhandling utan egentlig styrning. Avila & Ronen testade Ronen & McKinneys modell samt Ronens (1992) modell med straff-faktor för att ge incitament till att uppge sann information. I båda fallen testades tre versioner: Grundformen, en variant med feedback på hur motparten har uppträtt

[1] Se t.ex. Roth (1988) för en exposé vad gäller såväl inriktningar som metodfrågor.

under förhandlingen samt en variant med möjlighet för parterna att koordinera sitt agerande. Totalt testades alltså sex olika slag av styrning i Avila & Ronens studie.[2]

I båda studierna observerades faktisk transfererad kvantitet i varje enskild förhandling, vilket sedan översattes till ett effektivitetsmått baserat på vilken sammanlagd vinst den transfererade kvantiteten motsvarade jämfört med den potentiellt maximala sammanlagda vinsten. Dejong m.fl. drog sammanfattningsvis slutsatsen att det inte fanns några signifikanta skillnader mellan de modeller som testades om man exkluderade de inledande perioderna, där fri förhandling annars var signifikant mindre effektiv p.g.a. ett större antal förhandlingar som inte ledde till avslut. Avila & Ronen konkluderade sin studie med att Ronens modell med straff-faktor kompletterad med antingen feedback eller koordinationsmöjlighet ledde till signifikant eller åtminstone marginellt signifikant högre effektivitet än de övriga fyra modellerna.

PROBLEM OCH SYFTE

Det åligger varje forskare att visa att den metod som används är den korrekta givet det forskningsproblem som forskaren arbetar med. Om forskaren tänker sig att med hjälp av stickprov beskriva hur någon parameter i verkligheten objektivt sett ser ut i något avseende passar t.ex. en statistisk metod ofta bäst. För att kunna resonera om varför ett visst fenomen råder kan en kvalitativ metod vara att föredra.

När väl den allmänna metodinriktningen för en studie har bestämts måste mer detaljerade beslut rörande det konkreta tillvägagångssättet fattas. Dejong m.fl. (1989) och Avila & Ronen (1999) använde sig där huvudsakligen av vanliga t-tester och variansanalyser för att utvärdera de datamängder som experimenten gav.

I båda artiklarna saknas emellertid helt diskussion om vilka modellantaganden som valda metoder baseras på. Därmed saknas även analys av huruvida dessa antaganden kan antas vara rimliga. Läsaren kan således inte veta om de båda studiernas slutsatser egentligen är valida. Denna rapport syftar därför till att re-analysera resultaten av de båda experimenten mot bakgrund av metodval baserat på en mer fullständig validitetsdiskussion.

Återstoden av rapporten utgörs av fyra huvudsakliga avsnitt. I det första beskrivs de datamängder som de båda experimenten genererade och sättet på vilket datamängderna åstadkoms och bearbetades i de båda artiklarna. I det andra avsnittet görs en kritisk analys av de antaganden som författarna gör i och med sina val av statistiska modeller. I avsnitt tre re-analyseras de befintliga datamängderna med statistiska modeller som diskussionen i det föregående avsnittet leder fram till. I den avslutande syntesen jämförs resultaten av re-analysen med de slutsatser som drogs i de ursprungliga artiklarna och betydelsen av eventuella skillnader diskuteras.

BESKRIVNING AV EXPERIMENTEN OCH DESS DATAMÄNGDER

De båda experimenten hade i stort samma upplägg. Ett antal studenter (54 för Dejong m.fl. och 128 för Avila & Ronen) parades ihop slumpvis där den ene fick vara ”köpare” och den andre

[2] Se respektive originalkälla för en mer detaljerad beskrivning och analys av respektive modell.

”säljare”. Varje köpare fick information om sin marginalnyttofunktion medan varje säljare fick information om sin marginalkostnadsfunktion. Båda funktionerna gällde för kvantitetsintervallet 1-10 enheter. Högre kvantiteter än 10 enheter var alltså inte tillåtet att transferera. På basis av denna information fick parterna i ett antal på varandra följande perioder via skriftlig kommunikation förhandla sig fram till ett ”avslut” i form av en kombination av pris och kvantitet. Kunde man inte komma överens om villkoren för ett ”avslut” i en viss period så var kvantiteten, och därmed även vinsten för var och en av parterna, 0 i den perioden. Den totala vinst som uppstod för varje enskild individ under förhandlingarna var den betalning som utgick för deltagandet i experimentet. Incitament till individuell vinstmaximering fanns således.

Den viktigaste mätvariabeln i båda experimenten är den kvantitet som respektive förhandling resulterar i. I samtliga förhandlingar oavsett styrmodell är både pris och kvantitet variabler, men eftersom den sammanlagda vinsten endast är en funktion av kvantiteten (priset visar endast hur den uppkomna vinsten ska fördelas mellan parterna) så baseras analyserna av styrmodellernas effektivitet såväl i artiklarna som här endast på variabeln kvantitet. Effektiviteten för en viss kvantitetsnivå utgörs av kvoten mellan den sammanlagda vinst som den aktuella kvantiteten gav parterna och den maximala sammanlagda vinst som hade varit möjlig att uppnå om ”rätt” kvantitet valdes. I tabell 1 framgår vilken effektivitet som olika kvantitetsutfallen motsvarade i de båda artiklarna. De små skillnaderna beror på att Avila & Ronen gjorde en liten justering av de värden som Dejong m.fl. använde, eftersom man ville ha ett unikt optimum. I Dejong m.fl. innebar både kvantiteten 6 och kvantiteten 7 att maximal effektivitet uppnåddes.

Kvantitet	Dejong m.fl.	Avila & Ronen
0	0	0
1	0,270833	0,270833
2	0,5	0,5
3	0,6875	0,6875
4	0,833333	0,833333
5	0,9375	0,9375
6	1	1
7	1	0,979167
8	0,9375	0,916667
9	0,8125	0,791667
10	0,604167	0,583333

Tabell 1: Effektivitet vid olika kvantiteter

Dejong m.fl. testade tre olika styrmodeller:

- Ronen & McKinneys (1970) modell med subventionerad intern handel (RM)
- Hirshleifers (1956) modell med strikt marginalkostnadsprissättning (H)
- Fri förhandling utan egentlig styrning (FF).

De 54 deltagarna utgjorde 27 förhandlingspar där 9 par fick förhandla under var och en av de tre modellerna i 10 på varandra följande perioder. Resultatet, i termer av kvantiteter, framgår av tabell 2.

Modell	Par	Period									
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
FF	*1*	0	5	4	5	4	6	5	6	2	7
	2	0	0	2	2	2	5	5	6	4	6
	3	10	0	1	0	0	0	5	6	5	5
	4	0	6	0	5	0	5	6	5	6	6
	5	6	7	7	6	6	6	6	6	6	4
	6	5	7	5	7	5	5	6	0	7	0
	7	0	2	1	4	3	2	5	5	3	4
	8	0	0	5	0	0	0	3	2	3	4
	9	0	0	1	1	0	5	6	6	5	5
H	*1*	3	6	3	5	5	7	6	6	6	6
	2	4	5	4	6	6	6	6	6	6	6
	3	0	0	0	4	4	3	4	4	5	6
	4	0	5	5	6	6	6	7	7	6	6
	5	4	4	4	5	5	5	5	5	6	6
	6	2	4	4	4	4	5	6	6	7	6
	7	2	1	2	0	6	6	6	6	5	6
	8	3	3	4	3	4	5	5	4	0	0
	9	3	2	3	3	3	5	5	5	5	0
RM	*1*	0	5	5	7	6	5	4	5	5	5
	2	0	0	5	4	4	6	5	6	4	3
	3	6	6	7	5	7	7	7	6	7	7
	4	1	6	2	5	5	5	5	5	5	5
	5	3	4	4	5	5	6	7	6	6	7
	6	0	6	5	6	5	6	6	6	6	6
	7	5	5	4	5	3	3	3	4	0	4
	8	6	5	7	5	6	6	6	6	7	6
	9	3	5	6	3	3	6	0	7	7	7

Tabell 2: Resultat av experiment – Dejong m.fl. (1989)

Avila & Ronen testade sex olika styrmodeller:

- Ronen & McKinneys (1970) modell med subventionerad intern handel utan någon feedback rörande motpartens agerande och utan möjlighet för parterna att koordinera sitt agerande (RM)
- Ronen & McKinneys modell med feedback men utan möjlighet till koordination (RMF)
- Ronen & McKinneys modell utan feedback men med möjlighet till koordination (RMC)
- Ronens (1992) modell med subventionerad intern handel och straff-faktor utan någon feedback rörande motpartens agerande och utan möjlighet för parterna att koordinera sitt agerande (R)
- Ronens modell med feedback men utan möjlighet till koordination (RF)
- Ronens modell utan feedback men med möjlighet till koordination (RC).

De 128 deltagarna utgjorde 64 förhandlingspar där 10 par fick förhandla under var och en av de sex modellerna i 10 på varandra följande perioder. Modellerna RM och RMF kompletterades dessutom med två extra förhandlande par vardera. Resultatet, i termer av kvantiteter, framgår av tabell 3.

Modell	Par	Period									
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
RM	*1*	0	0	7	7	0	0	0	0	7	0
	2	0	0	0	0	0	0	0	0	0	2
	3	0	0	0	0	0	0	0	0	8	7
	4	0	7	7	7	7	7	7	7	7	7
	5	0	0	0	0	0	0	6	6	6	6
	6	6	6	6	6	6	6	6	6	6	6
	7	0	0	5	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	0	0
	9	0	0	0	6	6	6	6	6	6	6
	10	0	0	0	6	6	6	6	6	6	6
	11	0	0	5	5	5	5	5	5	5	5
	12	5	5	5	5	5	0	5	5	5	5
RMF	*1*	0	0	0	0	5	5	5	5	0	0
	2	0	0	7	7	7	7	7	7	0	0
	3	6	6	6	6	6	6	6	6	0	6
	4	5	0	0	0	6	6	6	6	6	6
	5	0	6	0	0	6	6	6	0	6	6
	6	0	0	7	7	7	7	7	7	7	0
	7	0	0	7	7	7	7	7	7	7	7
	8	0	6	6	6	6	6	6	6	6	6
	9	0	5	0	5	0	5	5	0	0	5
	10	0	0	0	0	0	0	6	6	6	6
	11	0	0	0	7	0	5	5	5	5	5
	12	0	0	0	0	0	0	0	0	0	8
RMC	*1*	0	0	6	6	0	10	10	10	10	10
	2	3	3	0	0	3	4	10	10	10	10
	3	10	10	10	10	10	10	10	10	10	10
	4	0	6	7	7	7	7	7	7	7	7
	5	6	6	8	10	10	10	10	10	10	10
	6	0	0	6	7	9	8	0	8	8	8
	7	6	0	6	6	0	9	10	10	10	10
	8	0	0	0	6	0	6	6	6	6	6
	9	0	0	5	5	0	0	5	6	5	6
	10	0	0	0	4	5	7	6	10	10	10
R	*1*	6	6	0	6	0	6	6	6	6	6
	2	0	0	0	0	0	0	0	0	0	0
	3	6	6	6	6	6	6	6	6	6	6
	4	0	0	6	6	6	6	6	6	6	6
	5	5	0	5	5	6	6	0	6	6	6
	6	4	0	0	6	0	6	6	6	6	6
	7	0	8	0	0	0	6	6	6	0	6
	8	0	0	0	0	0	0	0	6	0	6
	9	0	6	0	0	0	0	0	0	0	0
	10	0	0	0	0	0	5	5	5	5	5
RF	*1*	0	0	0	0	0	0	6	6	6	6
	2	4	0	0	5	5	0	6	6	6	6
	3	0	5	0	0	6	6	6	6	6	6
	4	0	0	0	5	0	6	0	6	6	6
	5	6	6	6	6	6	6	6	6	6	6
	6	0	7	7	7	7	0	6	6	6	6
	7	0	0	0	0	0	0	0	4	4	4
	8	0	0	0	0	0	8	0	8	8	8
	9	4	4	4	4	0	0	6	6	6	6
	10	0	0	0	0	6	6	6	6	6	6

RC	*1*	0	5	0	9	8	6	6	6	6	6
	2	0	4	6	5	5	0	6	6	7	8
	3	0	0	0	0	0	4	6	4	4	4
	4	0	6	6	6	6	6	6	6	6	6
	5	0	0	0	3	5	5	5	6	6	6
	6	0	8	5	6	7	7	6	5	6	6
	7	5	5	6	10	6	6	6	6	6	6
	8	0	0	0	3	8	8	8	8	8	8
	9	5	6	6	6	5	5	4	5	4	5
	10	0	0	0	6	0	0	6	6	6	6

Tabell 3: Resultat av experiment – Avila & Ronen (1999)

I båda studierna räknades kvantiteterna sedan om till effektiviteter (enligt tabell 1) varefter den genomsnittliga effektiviteten de olika modellerna utvärderades och jämfördes, både för samtliga perioder och för de avslutande perioderna, med framför allt t-tester och variansanalyser. Dejong m.fl. kom fram till att det inte fanns några signifikanta skillnader mellan de modeller som testades om man exkluderade de inledande perioderna, där fri förhandling annars var signifikant mindre effektiv p.g.a. ett större antal förhandlingar som inte ledde till något avslut. Avila & Ronen konkluderade sin studie med att styrning med modellerna RF och RC båda ledde till signifikant ($\alpha = 0{,}05$) högre effektivitet än RM och RMF samt marginellt signifikant ($\alpha = 0{,}1$) högre effektivitet än R och RMC. Därmed drogs slutsatsen att straff-faktorn i Ronen (1992) i sig leder till högre effektivitet, om den kompletteras med antingen feedback eller möjlighet till koordination, jämfört med Ronen & McKinneys modell.

ANALYS AV MODELLANTAGANDEN

Om syftet med en studie är att kontrollera antalet gröna bilar som kör på en viss väg under en viss tidsperiod så hjälper det föga att använda sig av många oberoende personer som räknar bilar (för att åstadkomma högre reliabilitet, d.v.s. mindre känslighet för slumpfel), om det är så att personerna t.ex. fått instruktioner att räkna röda bilar istället för gröna eller att räkna mopeder istället för bilar (vilket förstör validiteten, d.v.s. resultatens giltighet). Att välja metod med god validitet är alltså ett nödvändigt villkor för att en vetenskaplig studie ska kunna leda fram till trovärdiga resultat. Att en viss metod har hög reliabilitet, eller hög statistisk power, är helt ointressant om metodens validitet är låg.[3] Det kan inte finnas någon mening med att mäta så exakt som möjligt om man ändå mäter något annat än det man faktiskt är intresserad av att mäta.

För forskning baserad på statistiska modeller så handlar validitet i hög grad om de antaganden som de utnyttjade modellerna grundas på. En validitetsanalys av resultat som härrör från statistisk bearbetning innebär alltså analys av rimligheten i de modellteoretiska antaganden om verkligheten som har gjorts. Variansanalyser och t-tester får generellt användas om två sådana villkor är uppfyllda (se t.ex. Bowerman m.fl., 2001):

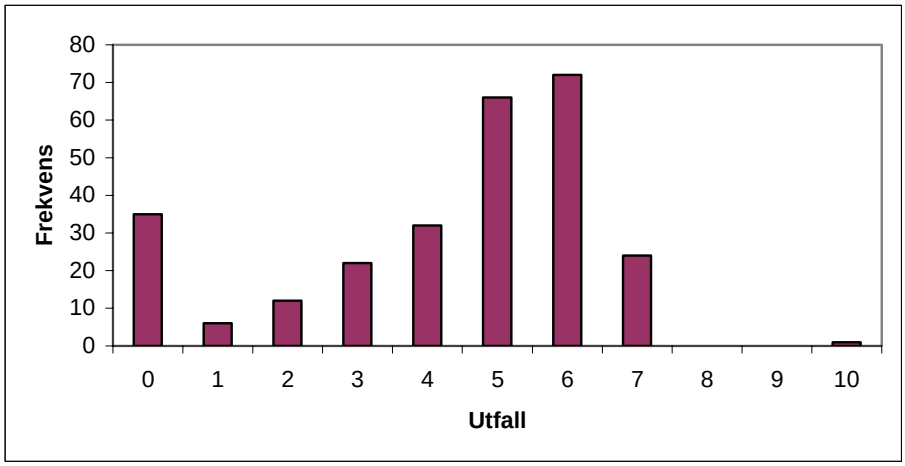
- Den data som används är av kvantitativ karaktär, d.v.s. den har en reell matematisk betydelse.
- Populationen som stickprovsdatan kommer ifrån kännetecknas av normalfördelning.

[3] En statistisk metods ”power” är sannolikheten att metoden leder till förkastande av nollhypotesen i en enkel hypotestest om det faktiskt är så att nollhypotesen verkligen är falsk.

Den data som experimenten genererade är uppenbarligen av kvantitativ karaktär. Om en förhandling resulterar i att kvantiteten 8 enheter ska transfereras så hade det varit dubbelt så mycket som om kvantiteten 4 enheter hade transfererats. Det första villkoret för att få använda variansanalys och t-test är alltså uppfyllt.

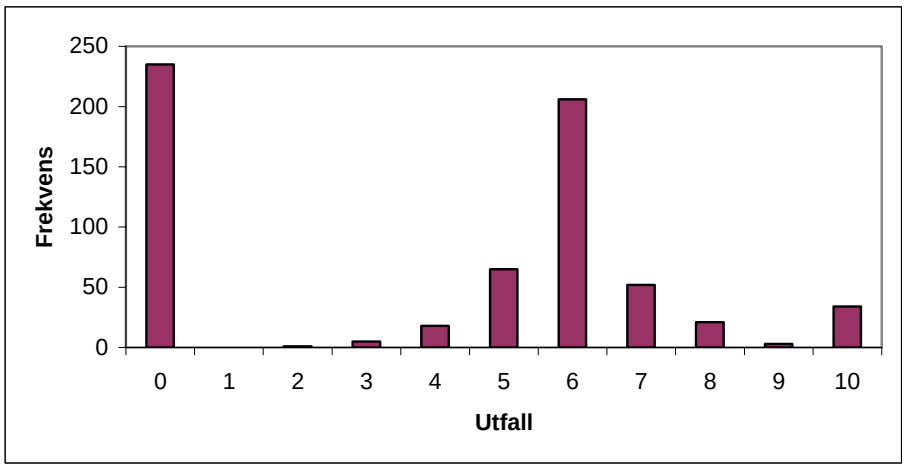
Det andra villkoret är att stickprovstagningen ska ha skett från en normalfördelning, och här uppstår direkt ett problem. Det finns exakt 11 olika tänkbara utfall för variabeln ”kvantitet” vid varje förhandling. Normalfördelningen är kontinuerlig, medan experimentdatan följer en diskret fördelning. För diskreta fördelningar med ett stort antal möjliga utfall kan ofta en kontinuerlig fördelning användas som approximation av den diskreta. I detta fall är antalet möjliga utfall emellertid lågt, vilket i sig är en faktor som talar emot användandet av parametriska metoder som t-tester och variansanalyser.

Betydligt mer allvarligt är dock det faktum att en normalfördelning ska vara ”klockformad” och symmetrisk kring ett medelvärde. Det går naturligtvis inte alltid att säga vilken slags fördelning som stickprovsdata kommer från, men ett sätt att i alla fall få en hyfsad bild av saken är att se hur stickprovsdatan i sig är fördelad. Vid stor stickprovsstorlek kommer stickprovsdata i sig att vara approximativt normalfördelad om det är så att ursprungspopulationen är normalfördelad. I figur 1 och 2 framgår de faktiska frekvenserna för de olika möjliga utfallen i de båda studierna på aggregerad nivå, och det framgår att normalfördelning är ett orimligt antagande i båda fallen, särskilt i Avila & Ronens studie där fördelningen är väldigt sned.[4]



Figur 1: Absoluta frekvenser för utfallen i studien av Dejong m.fl. (1989)

[4] I appendix 1 och 2 specificeras utfallens frekvenser för respektive styrmodell i respektive studie i samma typ av diagram. Det uppenbart icke-normalfördelade mönstret återfinns i alla situationer.



Figur 2: Absoluta frekvenser för utfallen i studien av Avila & Ronen (1999)

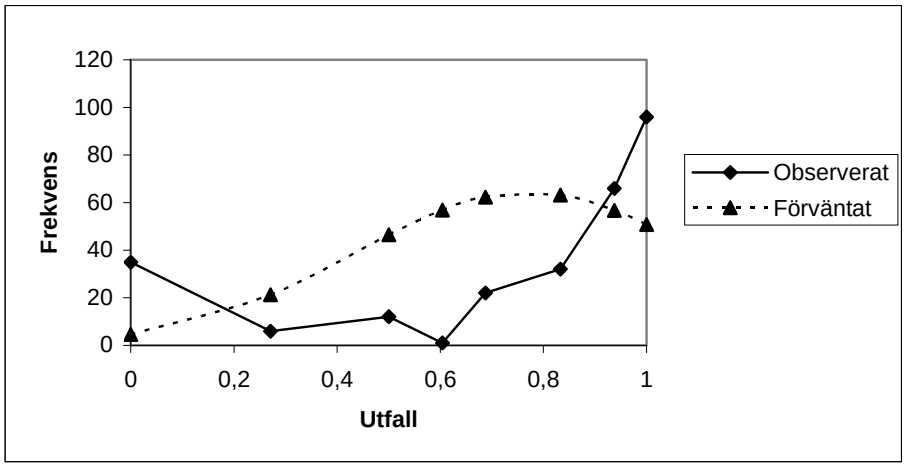
Studier har visat att t-testen är relativt robust för mindre avvikelser i modellantagandena förutsatt att stickprovsstorleken är stor.[5] När det gäller normalfördelningsantagandet så menar t.ex. Bowerman m.fl. (2001) att det räcker att ursprungspopulationen är någorlunda klockformad och symmetrisk kring medelvärdet. Det är dock minst sagt svårt att på ett trovärdigt sätt argumentera för att frekvenserna i figurerna 1 och 2 trots de stora stickprovsstorlekarna ger antydan om något i stil med normalfördelning. Problemet därvidlag är alla förhandlingar som inte ledde till något avslut, vilket motsvaras av utfallet 0 i figurerna 1 och 2. Hade det inte varit för dem hade faktiskt normalfördelningsantagandet sett ut att kunna stämma ganska bra i båda studierna. Som det ser ut nu, när antalet möjliga utfall dessutom är lågt, så måste t-testen anses vara ett orimligt analysverktyg. Variansanalysen är till och med mindre robust för avvikelser i modellantagandena än t-testen. Om inte t-testen rimligen kan användas så är variansanalysen därmed ännu mer orimlig.

Emellertid är det otillräckligt att studera kvantitetsdatans fördelning. Som vi noterade tidigare analyserar varken Dejong m.fl. eller Avila & Ronen sina resultat på basis av kvantitetsdata utan på effektivitetsmått som respektive kvantitetsutfall räknas om till (se tabell 1). I figurerna 3 och 4 illustreras hur fördelningen av stickprovsdatan ser ut efter denna operation i de båda fallen. I båda fallen jämförs den observerade fördelningen med en förväntad normalfördelningskurva för intervallet 0 till 1 som baseras på respektive datamängds medelvärde och standardavvikelse. Det framgår med all önskvärd tydlighet att ett antagande om normalfördelning är fullständigt orimligt i båda fallen.[6] Sammanfattningsvis måste validiteten i de slutsatser som Dejong m.fl. och Avila & Ronen presenterar därför ifrågasättas kraftigt.

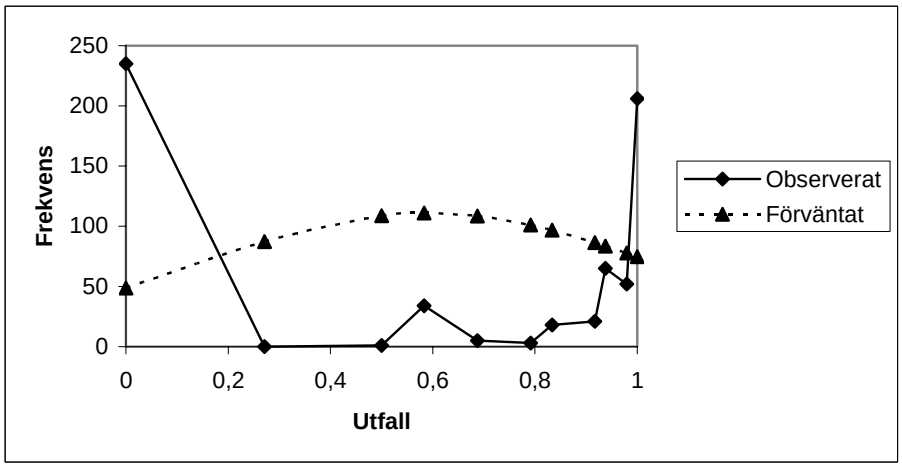
Eftersom den parametriska t-testen riskerar att ge invalida resultat p.g.a. kraftiga avvikelser i modellantagandena borde någon icke-parametrisk metod ha använts istället.[7] Närmast till hands ligger då Mann-Whitneys metod för jämförelse av två populationer, som inte förutsätter något annat om den data som ska analyseras mer än att den är rangordningsbar. Med denna metod kan de olika internprissättningsmodellernas effektivitet jämföras på ett sätt som ger valida slutsatser. Istället för variansanalys kan icke-parametrisk regression (Spearman's rangkorrelationskoefficient) användas för att analysera betydelsen av inlärning i förhandlingssituationen, d.v.s. om

[5] Se t.ex. Bartlett (1935), Bradley (1980), Geary (1947), Pearson & Please (1975), Pocock (1982) och Scheffe (1959).
[6] Chi-square-tester för goodness of fit har dessutom genomförts och dessa visar att datan inte i något av fallen kommer från normalfördelning ($p < 0{,}001$ i båda fallen).
[7] Se t.ex. Bowerman m.fl. (2001)

effektiviteten för en internprisättningsmodell påverkas av det ackumulerade antalet tidigare förhandlingar som genomförts. Inte heller denna metod förutsätter något mer än att datan är rangordningsbar, vilket innebär att den ger valida resultat i den aktuella situationen.



Figur 3: Observerad respektive förväntad fördelning för effektivitetsvärden i studien av Dejong m.fl.



Figur 4: Observerad respektive förväntad fördelning för effektivitetsvärden i studien av Avila & Ronen

När det gäller att jämföra de olika interprissättningsmodellernas andel faktiska avslut är en parametrisk test däremot både möjlig och lämplig. Eftersom antalet avslut per styrmodell är en binomialfördelad variabel kan en vanlig z-test användas i detta syfte, eftersom antalet observationer är stort. Någon sådan test gör emellertid varken Dejong m.fl. eller Avila & Ronen.

RE-ANALYS AV EXPERIMENTENS DATAMÄNGDER

Eftersom de analysmetoder som primärt ska användas här, Mann-Whitneys metod och Spearman's rangkorrelation, baseras på rangordning är det logiskt att komplettera de medelvärden för effektiviteterna som Dejong m.fl. och Avila & Ronen rapporterar med rangordningsbaserade beskrivande mått på datamängdernas centraltendens och spridning. I tabell 4 och 5 redovisas

dessa mått för de båda studierna, dels för samtliga perioder och dels för de 4 respektive 3 avslutande perioderna. Anledningen till att de avslutande 4 perioderna redovisas från Dejongs studie medan endast de 3 sista perioderna från Avila & Ronens studie tas med är att ”konvergensen” analyserades på just dessa sätt i respektive studie. Av jämförbarhetsskäl görs samma åtskillnad genomgående även här.

Redan här får vi en antydan om att resultaten, så som de redovisas i de aktuella studierna, kan vara tveksamma. En rangordning baserad på medelvärde mellan de sex modellerna i Avila & Ronens studie avviker från en rangordning baserad på median. Detta fenomen existerar oavsett om man jämför på basis av alla perioder eller endast de tre sista.

Dejong m.fl. drog slutsatsen att den enda signifikanta skillnad i effektivitet mellan de modeller som testades var att FF var signifikant mindre effektiv om alla perioder räknades med. För endast de 4 avslutande perioderna fanns inga signifikanta skillnader. I Avila & Ronens studie uppgavs att RC och RF i de tre avslutande perioderna båda var signifikant mer effektiva än RM och RMF, samt marginellt signifikant mer effektiva än R och RMC. Inga andra skillnader rapporterades vara signifikanta i de tre sista perioderna, och sett över alla perioder fanns inga signifikanta skillnader.

Modell	Sista 4 perioderna			Alla perioder		
	Medelvärde	Median	Interkvartila intervallet	Medelvärde	Median	Interkvartila intervallet
FF	0,8547	0,9375	0,1667	0,6495	0,9375	0,7292
H	0,8889	1	0,0625	0,8032	0,9375	0,2760
RM	0,8947	1	0,0625	0,8572	0,9375	0,1667

Tabell 4: Centraltendens och spridning för effektiviteten i studien av Dejong m.fl.

Modell	Sista 3 perioderna			Alla perioder		
	Medelvärde	Median	Interkvartila intervallet	Medelvärde	Median	Interkvartila intervallet
RM	0,6649	0,9375	1	0,4918	0,7083	0,9792
RMF	0,6800	0,9792	1	0,5877	0,9375	1
RMC	0,7375	0,5833	0,3854	0,6162	0,5833	0,3958
R	0,7271	1	0,7656	0,5325	0,9375	1
RF	0,9570	1	0	0,6000	0,9271	1
RC	0,9597	1	0,0781	0,7383	0,9375	0,2344

Tabell 5: Centraltendens och spridning för effektiviteten i studien av Avila & Ronen

Slutsatserna i skillnader i effektivitet var, som tidigare sagts, ett resultat av parvisa t-tester. P.g.a. den uppenbara icke-normalfördelningen i stickprovsdatan torde risken vara stor att dessa slutsatser kännetecknas av dålig validitet. Nya parvisa tester av modellernas effektivitet har därför gjorts med Mann-Whitneys metod (med justering för likanoteringar), dels för samtliga perioder och dels för de avslutande perioderna på samma sätt som i ursprungsstudierna. Resultatet redovisas i tabellerna 6 och 7 (Dejong m.fl.) samt 8 och 9 (Avila & Ronen). Tabellerna redovisar p-värden för parvisa tester med den genomgående nollhypotesen att ”kolumnmodellen” i mediantermer har högst lika hög sann effektivitet som ”radmodellen”.

I tabell 6 noterar vi att FF är signifikant mindre effektiv än H och RM när alla perioder räknas med, precis som Dejong m.fl. konkluderade. Det visar sig emellertid att RM i sig även är

signifikant effektivare än H. I tabell 7 kan vi konstatera att även när endast de 4 sista perioderna analyseras kvarstår att H är signifikant effektivare än FF. Övriga samband är dock inte signifikanta. Sammanfattningsvis skiljer sig resultaten här en aning jämfört med de slutsatser som Dejong m.fl. drog.

Från tabell 8 kan vi utläsa att det finns många skillnader i effektivitet som är signifikanta i Avila & Ronens studie när alla perioder räknas med. RC är signifikant effektivare än alla övriga modeller. RM är dessutom signifikant ineffektivare än RF samt marginellt signifikant ineffektivare än RMF och R, och RF är marginellt signifikant effektivare än RMC. Precis som för Dejong så ansåg emellertid Avila & Ronen att en stabilisering kunde skönjas i de avslutande perioderna, varför samma jämförelse görs även för de tre sista perioderna. Detta redovisas i tabell 9. De viktigaste slutsatserna här är att både RF och RC är signifikant effektivare än alla modeller som baseras på Ronen & McKinneys modell, d.v.s. RM, RMF och RMC. RF är även signifikant effektivare än R samt marginellt signifikant effektivare än RC. Även grundformen av Ronens modell, R, är effektivare än samtliga testade versioner av Ronen & McKinneys modell, även om skillnaden jämfört med RMF endast är marginellt signifikant. Den övergripande bilden är att resultaten avviker kraftigt jämfört med Avila & Ronens slutsatser.

Modell	FF	H	RM
FF	-	> 0,5	> 0,5
H	0,0224	-	> 0,5
RM	0,0003	0,0464	-

Tabell 6: p-värden för parvisa signifikanstester för effektivitet i studien av Dejong m.fl., alla perioder

Modell	FF	H	RM
FF	-	> 0,5	> 0,5
H	0,0475	-	0,3038
RM	0,1169	> 0,5	-

Tabell 7: p-värden för parvisa signifikanstester för effektivitet i studien av Dejong m.fl., 4 sista perioderna

Modell	RM	RMF	RMC	R	RF	RC
RM	-	> 0,5	> 0,5	> 0,5	> 0,5	> 0,5
RMF	0,0548	-	0,1779	> 0,5	> 0,5	> 0,5
RMC	0,1537	> 0,5	-	> 0,5	> 0,5	> 0,5
R	0,0533	0,3378	0,3138	-	> 0,5	> 0,5
RF	0,0189	0,2215	0,0750	0,3515	-	> 0,5
RC	0,0009	0,0651	0,0003	0,0924	0,01497	-

Tabell 8: p-värden för parvisa signifikanstester för effektivitet i studien av Avila & Ronen, alla perioder

Modell	RM	RMF	RMC	R	RF	RC
RM	-	> 0,5	0,2323	> 0,5	> 0,5	> 0,5
RMF	0,3915	-	0,1309	> 0,5	> 0,5	> 0,5
RMC	> 0,5	> 0,5	-	> 0,5	> 0,5	> 0,5
R	0,0467	0,0677	0,0170	-	> 0,5	> 0,5
RF	0,0002	0,0004	< 0,0001	0,0477	-	0,0782
RC	0,0116	0,0246	< 0,0001	0,3547	> 0,5	-

Tabell 9: p-värden för parvisa signifikanstester för effektivitet i studien av Avila & Ronen, 3 sista perioderna

Andelen misslyckade förhandlingar, d.v.s. andelen förhandlingar som inte ledde till något avslut och därmed effektiviteten 0, i respektive studie framgår av tabellerna 10 och 11. Den klart större andelen misslyckade förhandlingar under FF jämfört med H respektive RM i tabell 10 är signifikant (p=0,0042). Skillnaden H och RM emellan är däremot inte signifikant.

I tabell 12 redovisas p-värden för parvisa z-tester av datan i tabell 11. Den genomgående nollhypotesen är att ”kolumnmodellen” har minst lika hög sann andel misslyckade förhandlingar som ”radmodellen”. Ett antal samband är signifikanta. Framför allt gäller att både RMC och RC leder till signifikant lägre andel misslyckade förhandlingar än alla modeller som saknar möjlighet för aktörerna att koordinera sitt agerande. Möjligheten till koordination är alltså en faktor som i sig leder till större sannolikhet att en förhandling leder till avslut. Vidare gäller att RMF och RF leder till marginellt signifikant respektive signifikant lägre andel avslut än RM. Jämfört med grundmodellen med subventionerad intern handel leder således feedback rörande motpartens agerande under förhandlingen till större sannolikhet att fortsatta förhandlingar leder till avslut.

Modell	Antal förhandlingar	Antal misslyckanden	Andel misslyckanden
FF	90	21	0,2333
H	90	8	0,0889
RM	90	6	0,0667

Tabell 10: Andelen misslyckade förhandlingar per modell i studien av Dejong m.fl.

Modell	Antal förhandlingar	Antal misslyckanden	Andel misslyckanden
RM	120	59	0,4917
RMF	120	48	0,4
RMC	100	22	0,22
R	100	46	0,46
RF	100	38	0,38
RC	100	22	0,22

Tabell 11: Andelen misslyckade förhandlingar per modell i studien av Avila & Ronen

Modell	RM	RMF	RMC	R	RF	RC
RM	-	> 0,5	> 0,5	> 0,5	> 0,5	> 0,5
RMF	0,0766	-	> 0,5	0,3704	> 0,5	> 0,5
RMC	< 0,0001	0,0022	-	0,0002	0,0136	0,5
R	0,3198	> 0,5	> 0,5	-	> 0,5	> 0,5
RF	0,0483	0,3811	> 0,5	0,1259	-	> 0,5
RC	< 0,0001	0,0022	0,5	0,0002	0,0068	-

Tabell 12: p-värden för parvisa signifikanstester för andelen misslyckade förhandlingar i studien av Avila & Ronen

Avslutningsvis ska vi studera effektivitetens påverkan av antalet ackumulerade försök för de olika modellerna. Den fråga vi ställer oss är då om det finns en ”inlärningseffekt” i datan så att effektiviteten trendmässigt ökar med antalet genomförda försök. Detta testades i både Dejong m.fl. och Avila & Ronen med hjälp av variansanalyser. (Avila & Ronen gjorde även enkla linjära regressioner, men endast i beskrivande syfte.)

Variansanalys är dock inte en tillåten metod då populationerna i dessa fall, som vi har sett tidigare, inte kan anses vara åtminstone approximativt normalfördelade. Problemet är att det inte finns något icke-parametriskt alternativ till en flervägs variansanalys med mer än en observation per cell så som fallet är här. Istället kan icke-parametrisk regressionsanalys göras för att kartlägga om det finns en korrelation mellan modellernas effektivitet och den period som förhandlingen genomförs i. Genom rangordning av effektivitetsmåtten beräknades därför Spearman’s korrelationskoefficient för samtliga styrmodeller i båda studierna, och nollhypotesen att respektive modells sanna korrelationskoefficient är högst 0 testades. Korrelationskoefficienterna och p-värdena för respektive test finns i tabell 13 och 14.

Styrmodell	Spearman’s r	H0: $\rho \leq 0$
FF	0,36471	p < 0,01
H	0,56418	p < 0,01
RM	0,28337	p < 0,01

Tabell 13: Korrelationskoefficienter med p-värden i studien av Dejong m.fl.

Styrmodell	Spearman’s r	H0: $\rho \leq 0$
RM	0,31058	p < 0,01
RMF	0,30882	p < 0,01
RMC	0,18975	0,1 < p < 0,05
R	0,32017	p < 0,01
RF	0,59907	p < 0,01
RC	0,47808	p < 0,01

Tabell 14: Korrelationskoefficienter med p-värden i studien av Avila & Ronen.

Det framgår att alla modellerna i Dejongs studie kännetecknas av en signifikant positiv korrelation mellan antalet perioder och den effektivitet som uppnås. Samma bild erhålls från Avila & Ronens studie, med det lilla undantaget att korrelationen för RMC endast är marginellt

signifikant. Sammanfattningsvis betyder detta att försökspersonerna i experimenten ”lärde sig” av tidigare erfarenhet på så sätt att effektiviteten trendmässigt var ökande. Den effektivitet som modellerna låg på mot slutet av respektive experiment torde därför vara en god indikation på vilken faktiskt effektivitet som respektive modell bör tillskrivas.

SYNTES

I den här rapporten har resultaten av de två experiment som Dejong m.fl. (1989) och Avila & Ronen (1999) genomfört re-analyserats med valida metoder. Vad som framför allt är intressant är att de resultat som har erhållits vid den här genomförda re-analysen skiljer sig i olika avseenden från de slutsatser som drogs i de båda ursprungliga studierna. Dejong m.fl. (1989:57) skriver t.ex. att

> ”The surprising result of this laboratory experiment is how similar the transfer pricing mechanisms are with respect to performance, in particular the traditional [H] and the Ronen-McKinney [RM] mechanisms. When all periods of the experiment are considered, the direct negotiation [FF] mechanism is significantly less efficient than the other two mechanisms (due to a lower number of agreements). However, the efficiencies of the traditional [H] and Ronen-McKinney [RM] mechanisms are indistinguishable. In the last four periods, the efficiencies of all three mechanisms are indistinguishable.”

Detta är, som vi har sett i re-analysen, felaktigt. FF var mycket riktigt signifikant mindre effektiv än både H och RM när alla perioder räknades med, men H var i sig dessutom signifikant mindre effektiv än RM. Dessutom kvarstår att FF är signifikant mindre effektiv än H även när endast de avslutande 4 perioderna analyseras. Modellerna är alltså möjliga att skilja åt vad gäller faktisk styreffekt.

Avila & Ronen (1999:710) skriver att

> ”Regarding the mechanisms without the penalty factor [RM, RMF och RMC], there was no observable convergence towards the efficient level of transaction at the end of the experiment.”

Detta är uppenbarligen en felaktig slutsats eftersom samtliga tre modellers medianeffektivitet i verkligheten är signifikant eller marginellt signifikant positivt korrelerad med antalet perioder. Vidare konkluderar Avila & Ronen (1999:700) angående skillnaden mellan modellerna i de avslutande 3 perioderna att

> ”...the mechanisms with the penalty factor and some form of communication between the buyer and the seller (the RC and RF schemes) performed at least marginally better than all other four mechanisms.”

Tydligen är det istället så att de mekanismer som baserades på Ronens modell med straff-faktor genomgående var effektivare i de tre avslutande perioderna än de som inte hade någon straff-faktor. Det är således straff-faktorn i sig som är avgörande, inte möjligheten till feedback eller koordination som Avila & Ronen hävdade. Dessutom var RF effektivare än alla andra modeller i experimentet i det tre avslutande perioderna, alltså även effektivare än RC, vilket även det förbisågs av Avila & Ronen.

Sammanfattningsvis fanns i stora drag alltså fler samband som var signifikanta än vad Dejong m.fl. och Avila & Ronen hittade. I båda studierna uttrycktes viss förvåning över att skillnaderna inte var större, men en fundamental förklaring till detta är uppenbarligen att utnyttjade metoder kännetecknades av dålig validitet p.g.a. alltför restriktiva modellantaganden. Skillnaderna *var*

större, men felaktiga metodval gjorde att de inte upptäcktes. Det faktum att RMC i Avila & Ronens studie har högst medelvärde men lägst medianvärde av de tre modellerna utan straff-faktor, såväl för samtliga perioder som för de 3 sista perioderna, illustrerar väl det faktum att experimentdatans fördelning är alltför olik en normalfördelning för att signifikanstester baserade på medelvärden och parametriska metoder ska kunna användas. Ju snedare en fördelning är, desto större blir differensen mellan fördelningens median och medelvärde.

Vad är då implikationerna av de upptäckter som har gjorts här?[8] Ja, från re-analysen av Dejong m.fl. verkar det som om klara riktlinjer för internprissättning med uttalad målsättning att maximera den sammanlagda vinsten (Hirshleifers modell) kan vara en faktor som har en signifikant positiv styreffekt jämfört med att låta parterna förhandla fritt såväl i fallet med enstaka kontakter som när det gäller mer regelbundna förhandlingar. Signifikansen för denna effekt kan dock försvinna om styrmodellen i sig är mer komplicerad (Ronen & McKinneys modell). Om internprissättningen inte är regelbunden, så att parterna aldrig hinner lära känna varandra, så kan dock den positiva effekten hos en i övrigt tydligare beteendestyrande modell (Ronen & McKinney) överväga en lägre grad av komplexitet (Hirshleifer). Mer forskning behövs emellertid för att definitiva slutsatser av detta slag ska kunna dras.

Från re-analysen av Avila & Ronens studie visar det sig framför allt att aktörer, som drabbas av sanktioner när de i aktiv handling avviker från *ex ante* rapporterade marginalkostnads- eller marginalnyttofunktioner, tenderar att hitta sammanlagt effektivare lösningar än om inga sådana sanktioner finns. Ronens straff-faktor bör alltså vara en omständighet att räkna med i tillämpad ekonomisk styrning, och kanske även i andra sammanhang än just internprissättning. Denna faktors betydelse bör därför testas vidare i nya experimentella studier med andra styrmodeller.

Behov av två skilda slag av fortsatt forskning kan härledas från denna rapport. För det första är kunskapen om hur företagsekonomiska styrmodeller i stort påverkar mänskligt beteende tämligen dåligt utvecklad.[9] Specifikt inom internprissättning finns t.ex. många modeller som under olika antaganden är teoretiskt optimerande. Huruvida de faktiskt påverkar beslutsfattare på ”rätt” sätt har inte testats, vilket därför bör göras i form av t.ex. experimentella studier. Som tidigare nämndes är experiment för att bättre kunna förklara de skillnader som har hittats mellan olika styrmodeller också viktigt.

För det andra är det sannolikt att problemet med felaktiga metodval vid analys av experimentella studier förekommer i andra kontexter än just på interprissättningsområdet. Dessa kan, på liknande sätt som vi har sett här, ha lett till invalida slutsatser, vilka i sin tur kan ha påverkat den fortsatta vetenskapliga utvecklingen på ett felaktigt sätt. Det kan därför finnas stora behov av liknande re-analyser av data i andra sammanhang.

[8] Vi kan notera att varken Dejong m.fl. eller Avila & Ronen gjorde någon analys av den praktiska innebörden hos deras respektive resultat – de redogjorde i stora drag endast för vilka signifikanta skillnader som de hade upptäckt.

[9] Inom det nationalekonomiska ämnesområdet är det mer vanligt förekommande med experimentella studier av teoretiska modeller. Se t.ex. Roth (1988).

REFERENSER

Avila, M. & Ronen, J. (1999), ”Transfer Pricing Mechanisms: An Experimental Investigation”, International Journal of Industrial Organization, 17, 689-715.

Bartlett, M. S. (1935), "The Effect of Non-Normality on the t Distribution," Proceedings of the Cambridge Philosophical Society, 31, 223-231.

Bowerman, B. L., O’Connell, R. T. & Hand, M. L. (2001), ”Business Statistics in Practice”, McGraw-Hill Irwin, Boston.

Bradley, J. V. (1980), "Nonrobustness in Z, t, and F Tests at Large Sample Sizes," Bulletin of the Psychonomics Society, 16(5), 333-336.

Dejong, D. V., Forsythe, R., Kim, J-O. & Uecker, W. C. (1989), ”A Laboratory Investigation of Alternate Transfer Pricing Mechanisms”, Accounting, Organization and Society, 14, 41-64.

Geary, R. C. (1947), "Testing for Normality," Biometrika, 34, 209-242.

Hirshleifer, J. (1956), ”On the Economics of Transfer Pricing”, Journal of Business, 72-84.

Lantz, B. (2000), ”Internprissättning med effektiva incitament”, BAS, Göteborg.

Pearson, E. S., and Please, N. W. (1975), "Relationship Between the Shape of Population Distribution and Robustness or Four Simple Testing Statistics," Biometrika, 62, 223-241.

Pocock, S. J. (1982), "When Not to Rely on the Central Limit Theorem -- An Example from Absentee Data," Communications in Statistics, Part A -- Theory and Methods, 11(19), 2169-2179.

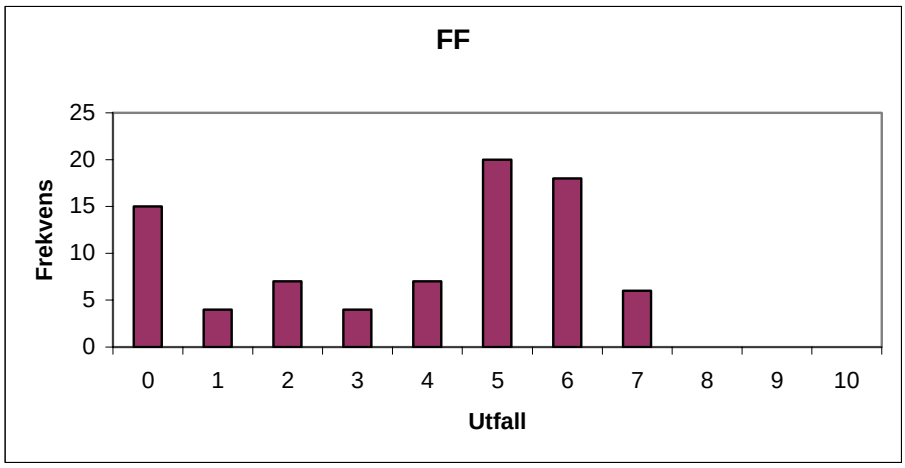
Ronen, J. (1992), ”Transfer Pricing Reconsidered”, Journal of Public Economics, 47, 125-136.

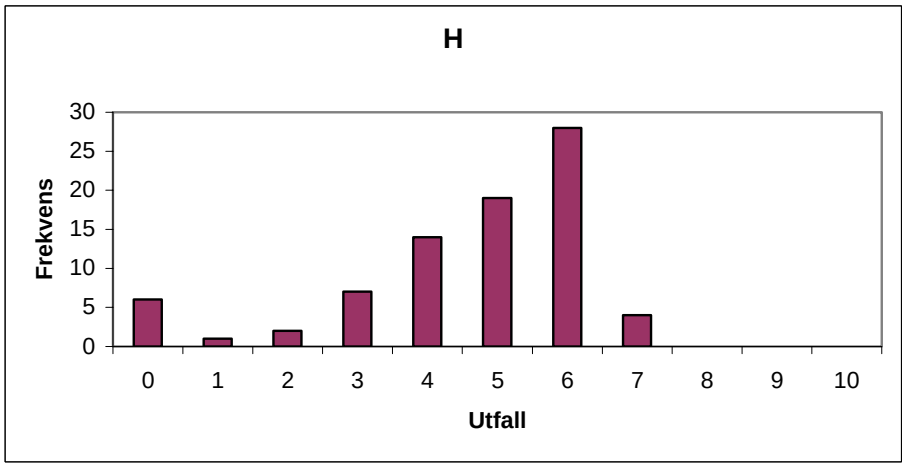
Ronen, J. & McKinney, G. (1970), ”Transfer Pricing for Divisional Autonomy”, Journal of Accounting Research, 99-112.

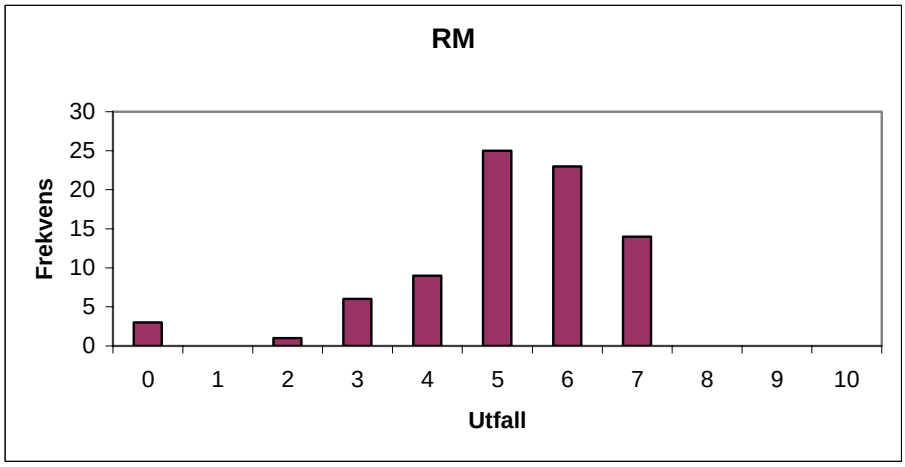
Roth, A. E. (1988), ”Laboratory Experimentation in Economics”, The Economic Journal, 98, 974-1031.

Scheffe, H. (1959), The Analysis of Variance, New York: Wiley.

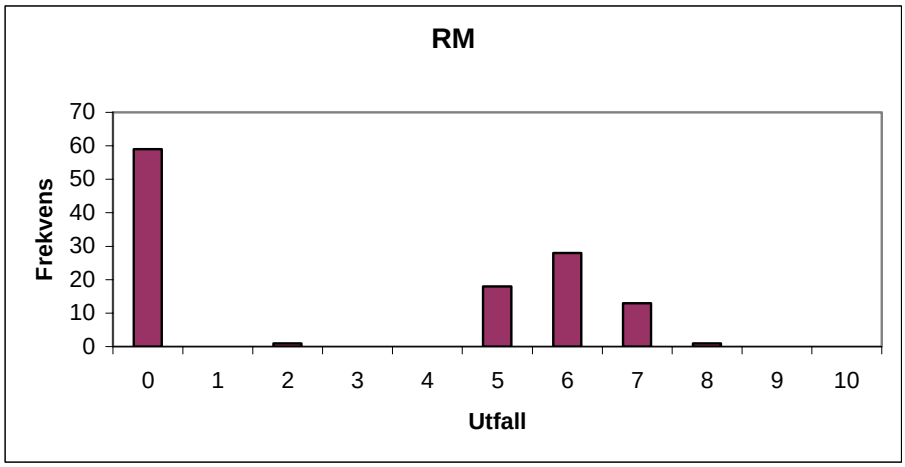
APPENDIX 1: UTFALL FÖRDELAT PÅ STYRMODELLER, DEJONG M. FL.

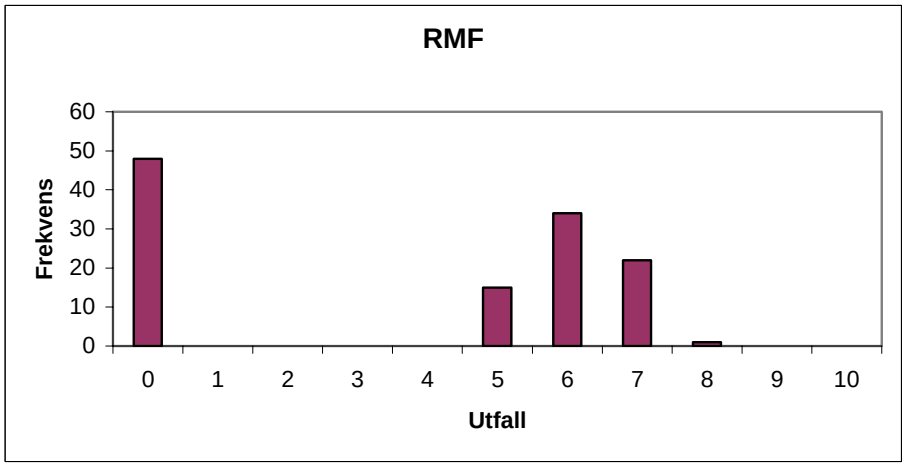


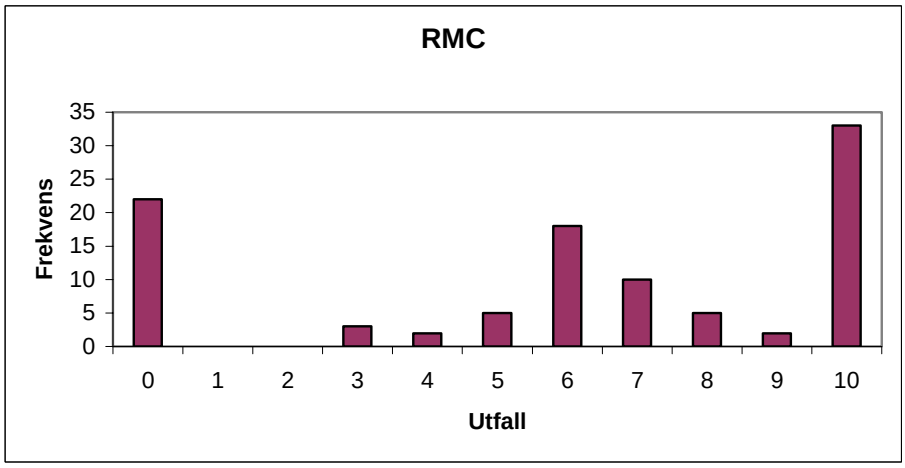


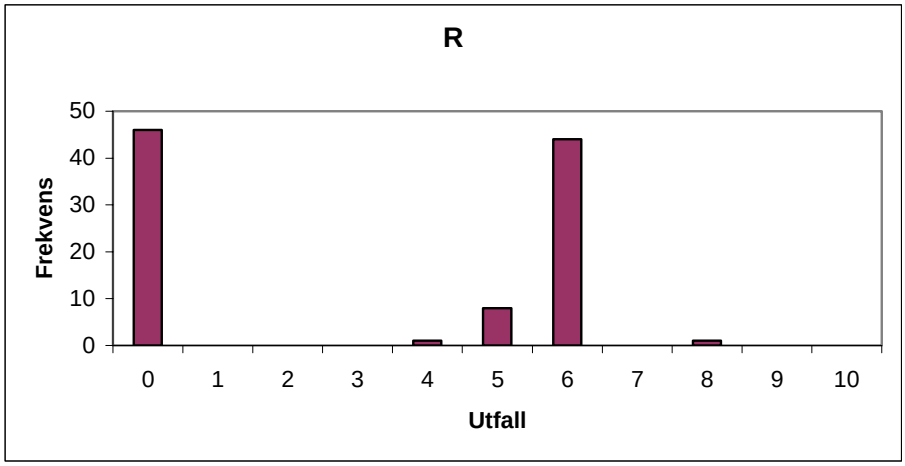


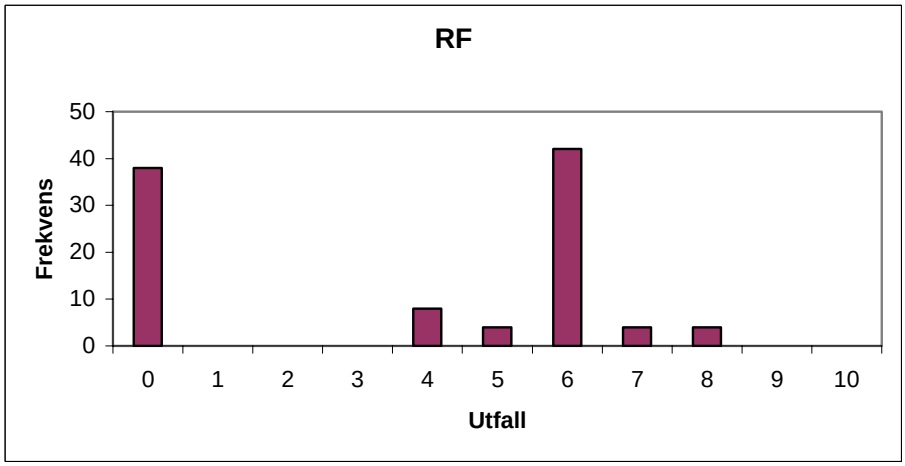
APPENDIX 2: UTFALL FÖRDELAT PÅ STYRMODELLER, AVILA & RONEN

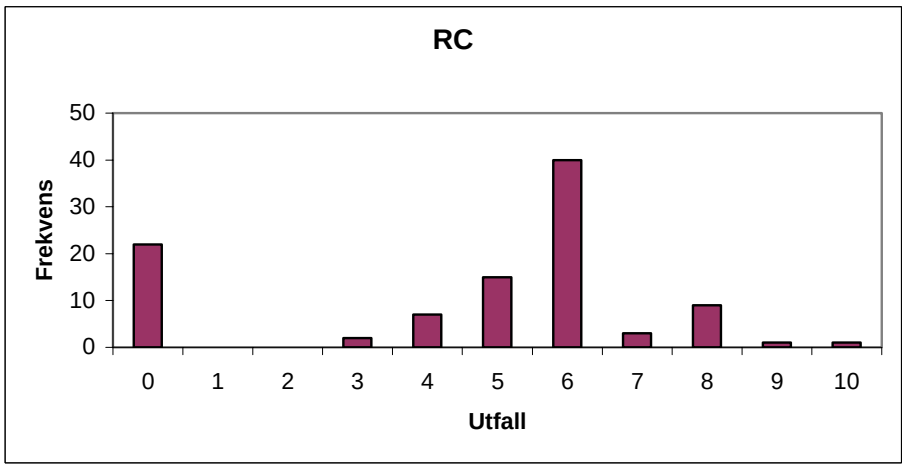







R
Frekvens
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10
Utfall


RF
Frekvens
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10
Utfall


RC
Frekvens
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10
Utfall