PEDAGOGISKA MÄTNINGAR

Skalor - Reliabilitet - Validitet

Sven-Eric Reuterberg

2

INNEHÅLLSFÖRTECKNING

I. INLEDNING

Kalle är 150 cm och väger 36 kg. Med detta uttryck har vi beskrivit två egenskaper hos Kalle: hans längd och hans vikt. Att förstå vad dessa egenskaper innebär vållar inga problem därför att det råder en fullkomlig enighet om vilka mätmetoder som skall användas för att få reda på dem.

Uppgiften om Kalle är också lätt att tolka därför att vi uttrycker den i kända och standardiserade måttenheter. Att måttenheterna är standardiserade betyder att 1 cm är en lika lång sträcka oavsett vilket måttband vi använder och på samma sätt har 1 kg samma tyngd oavsett vilken våg vi använder.

Uppgifterna om Kalle är också lätta att förstå av det skälet, att våra mätningar utgår ifrån en fast nollpunkt. Vi börjar ju att mäta Kalles längd från fotsulan och vågen skall stå på 0 när det inte finns någon tyngd på den.

Kalles längd och vikt har angivits i absoluta mått därigenom att vi har en fast nollpunkt att mäta utifrån och att vi mäter med standardiserade enheter, d.v.s. att skalstegen är lika stora över hela skalan. Mätningar med en fast nollpunkt och som dessutom har lika stora skalsteg sägs vara mätningar på KVOTSKALENIVÅ.

När vi utgår från mätningar som uttrycks på denna skalnivå kan vi uttala oss om kvoter. Eftersom Lisa är 120 cm kan jag t.ex. säga att Kalle är 1 1/4 gånger så lång som Lisa.

I stället för att ange Kalles och Lisas längd i absoluta värden, som vi gjort ovan, kan vi välja att ta Pers längd som utgångspunkt för våra mätningar. Per är 110 cm lång och följaktligen kan vi säga att Kalle är 40 cm längre och Lisa 10 cm längre än Per. Här har vi godtyckligt valt att ta Pers längd som utgångspunkt eller som nollpunkt för våra mätningar och därigenom är de inte längre absoluta utan relativa till Pers längd.

Genom att på detta sätt göra mätningen relativ har vi förlorat information. Vi vet ju inte längre hur långa Lisa och Kalle verkligen är utan bara att de är 10 resp. 40 cm. längre än Per. Vi kan nu inte heller ange någon kvot för hur lång Kalle är i förhållande till Lisa. Däremot kan vi fortfarande ange att Kalle är 30 cm längre än Lisa eftersom våra mätningar sker i den standardiserade måttenheten cm.

När vi mäter Kalles och Lisas längd med hjälp av en godtycklig nollpunkt (Per) men med lika skalsteg (cm) säger vi att mätningen sker på INTERVALLSKALENIVÅ. Uttrycket intervallskala anger att skalan har lika stora intervall eller mätenheter. Så länge våra mätningar är uttryckta i en skala med enhetliga skalsteg är det fortfarande meningsfullt att addera och subtrahera värden. Däremot är det inte möjligt att ange värden i kvoter. Antag nu att jag vill ange Kalles och Lisas längd men tyvärr har jag inte mätt Kalles, Lisas och Pers längd med måttband utan jag får nöja mig med att uttrycka deras längd i allmänna ordalag. Detta skulle kunna låta ungefär så här: "Kalle är mycket lång. Lisa är ganska kort medan Per är mycket kort."

Detta uttryck anger att Kalle är längst, därefter kommer Lisa och Per är kortast. Jag har dock ingen aning om deras verkliga längd och jag kan heller inte ange hur stora skillnaderna är mellan dem. I detta fall har jag uttryckt de tre individernas längd i ORDINALSKALA. Detta är en mycket vanlig skalnivå t.ex. i samband med frågeformulär. Ofta finner vi där svarsalternativ av typ:

mycket bra
ganska bra
varken bra eller dåligt
ganska dåligt
mycket dåligt

Dessa svarsalternativ går givetvis inte att addera. Inte heller har man någon grund för att påstå stegen mellan de olika alternativen är lika stora eller att "mycket dåligt" är uttryck för en absolut nollpunkt. Värden som är uttryckta i ordinalskala kan således ej adderas eller subtraheras. Det

enda vi kan göra med dessa värden är att rangordna dem. (I vissa fall kan det t.o.m. vara tveksamt om det är befogat med rangordning beroende på att svarsalternativen har något olika betydelse för olika människor. Sådana tolkningsproblem kan ofta undanröjas genom att man preciserar innebörden i alternativen:

ofta (varje dag)
ibland (någon gång/vecka)
osv.

Vi har ovan beskrivit vår lilla grupp om tre personer i variabeln längd och gjort denna beskrivning på tre olika nivåer kvotskale-, intervallskale- och ordinalskalenivå. Längd är ju en klart kvantitativ variabel - man kan ordna värdena i en skala från "mycket till lite" - och har därför varit lämplig som underlag för vår diskussion.

Genom att vi namngivit våra tre personer har gruppen också beskrivits i en annan variabel - kön. Denna variabel är kvalitativ och de olika värdena kan inte ordnas i en stigande eller fallande skala. Värden av denna typ hör mättekniskt hemma på NOMINALSKALENIVÅ, och de kan endast ligga till grund för en kategorisering. De anger däremot inte på något sätt hur kategorierna skall ordnas i förhållande till varandra. Som exempel på andra nominalskalevariabler kan nämnas nationalitet, blodgrupp, hårfärg osv.

Vi kan sammanfatta våra fyra skalnivåer och deras egenskaper med följande uppställning:

	Skalnivå			
	kvot	Intervall	Ordinal	Nominal
Absolut 0-punkt	x			
lika skalsteg	x	x		
rangordning	x	x	x	
likhet/olikhet	x	x	x	x

Vi beskrev tidigare vår lilla grupp med avseende på längd och kön d.v.s. med avseende på fysiska variabler. Vi går nu över

till att beskriva de tre personerna med avseende på en psykisk variabel. För "enkelhets skull" väljer vi resultatet på ett vanligt kunskapsprov.

Kalle går i årskurs 6. Lisa och Per går i årskurs 4 men i var sin parallellklass.

Alla tre har haft prov i samma ämne låt oss säga engelska. Deras resultat blev:

Kalle 9 poäng
Lisa 12 poäng
Per 10 poäng

Vad säger dessa resultat om elevernas kunskaper i engelska?
Platt intet!
Varför inte?

1. Vi saknar information om vilken typ av prov de tre eleverna genomgått. Antag att Kalles prov avsåg att översätta meningar till engelska. Lisas prov var ett glosprov med översättning från engelska till svenska medan Pers prov innebar översättning av svenska ord till engelska.

   Alla tre proven mäter kunskaper i engelska men det handlar om tre olika slags kunskaper. Man kan uttrycka det så att de mäter tre olika variabler. Att göra jämförelser mellan dessa tre prov är precis lika tokigt som om vi frågade oss om vem av Per och Lisa som var störst och vi försökte besvara frågan genom att ange Pers längd och Lisas vikt.

   För att tolka ett beteendevetenskapligt mätresultat måste vi först ta reda på vilken variabel som mätinstrumentet faktiskt utgör ett mått på. Eftersom psykiska variabler oftast är mycket svårdefinierade är frågan om vad mätinstrumenten faktiskt mäter mycket mer komplicerad än i samband med mätning av fysiska variabler. Frågan handlar om mätinstrumentets VALIDITET. Vi återkommer till denna fråga längre fram i kompendiet.

   För att komma ifrån validitetsproblemet ändrar vi förut-

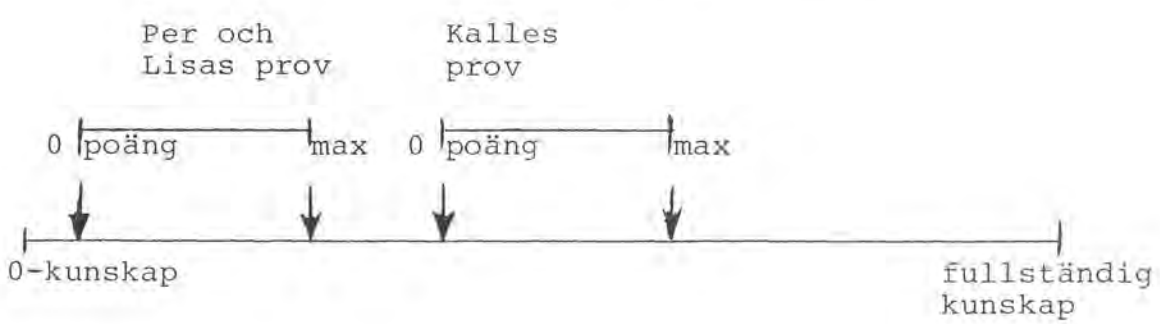
sättningarna så att alla tre eleverna har fått samma typ av prov - låt oss säga att alla proven bestod i översättning av svenska ord till engelska. Pers och Lisas prov innehåller dessutom exakt samma ord, medan Kalles prov innehåller betydligt svårare ord.

2. Trots att vi genom de ändrade förätsättningarna löst validitetsfrågan så att alla tre proven mäter samma typ av engelskkunskaper är resultaten omöjliga att tolka. Detta av det enkla skälet att resultaten inte är uttryckta i en absolut skala d.v.s. vi saknar absolut nollpunkt och vi saknar uppgift om vilken måttenhet som använts vid poängsättningen.

Om vi hade haft en absolut 0-punkt skulle detta innebära att 0 poäng på provet anger en total okunskap om engelska glosor.

I vårt exempel finns ingen som helst anledning att tro att så är fallet. En elev som får 0 poäng på Pers och Lisas prov kanske kan en hel del engelska ord även om eleven inte råkar kunna just de orden som ingår i provet. En elev som får 0-poäng på Kalles prov kan mycket väl ha ett betydligt bättre glosförråd än en elev som får full poäng på Pers och Lisas prov. (Förutsättningarna angav ju att Kalles prov innehöll betydligt svårare ord).

Enkelt kan vi åskådliggöra detta med nedanstående figur.



Att direkt avgöra hur två olika prov förhåller sig till varandra vad gäller svårighetsgrad är omöjligt. Ett sådant

avgörande kan aldrig bli annat än en subjektiv bedömning. Vi kan säga att Kalles prov är svårast men inte avgöra hur mycket svårare det är i jämförelse med det prov som Lisa och Per fick.

Genom att vi inte direkt kan avgöra hur de båda provens var för sig godtyckliga nollpunkter förhåller sig till varandra blir det självfallet omöjligt att jämföra resultat mellan de båda proven.

Eftersom Per och Lisa har fått exakt samma prov borde man ju utan betänkligheter kunna göra jämförelser mellan dem vad gäller provresultat. Ja, det kan man göra men bara under den förutsättningen att deras båda lärare har följt exakt samma bedömnings- och poängsättningsprinciper.

Låt oss antaga att de båda lärarna inte har gjort detta. Lisas lärare är mer petig i sin bedömning än vad Pers lärare är. Därför har han tillämpat principen att bedöma varje ord med en skala från 0-2 poäng, där 2 poäng endast ges för helt korrekta svar. 1 poäng har han givit då det förekommit något stavfel. Pers lärare har däremot använt skalan 0-1 poäng. I detta fall har följaktligen Lisas lärare använt sig av en poängskala som har mindre måttenhet eller skalsteg än den skala som Pers lärare använt.

Om måttenheterna varierar på detta sätt blir det självfallet omöjligt att göra jämförelser mellan Pers och Lisas klasser trots att de har samma prov och gemensam nollpunkt för sina mätningar.

En förutsättning för att direkt kunna jämföra två provresultat med varandra är således att resultaten härrör från samma prov, att mätningen sker utifrån en fast startpunkt och att måttenheten är enhetlig. Att justera för olikheter mellan två provresultat vad gäller mätningens utgångspunkt och för skalstegens storlek är möjligt med speciella statistiska metoder, vilka vi återkommer till i ett senare avsnitt.

Däremot torde det vara alldeles omöjligt att justera för det förhållandet att skalstegen faktiskt kan variera inom ett och samma prov.

Vid bedömningen av ett prov sätter vi en viss poäng för varje fråga. Dessa poäng summeras sedan till ett slutresultat på provet. Tanken bakom detta är naturligtvis att ju mer poäng en elev har fått på provet desto mer kunskaper har han visat.

Ofta belönas ett korrekt svar med samma poängantal på varje fråga, t.ex. 1 poäng. Det ser ut så här:

Uppg.	A	B	C
1	1	1	1
2	1	0	1
3	0	0	1
4	0	1	0
	2 p.	2 p.	3 p.

C har högst poäng på provet och därmed drar vi slutsatsen att C har de bästa kunskaperna inom det område, som provet mäter. A och B har 2 poäng var och är alltså jämbördiga. Är nu detta så säkert?
Ja, delvis.

C torde vara bättre än A eftersom båda har klarat uppgift nr 1 och 2 och C dessutom uppgift nr 3. Däremot är det inte helt säkert att C är bättre än B. Det kan nämligen vara så att uppgift nr 4 kräver så mycket kunskaper att de mer än väl uppväger de kunskaper som C visat genom att klara uppgifterna 2 och 3.

För att knyta an till vår inledande diskussion om skalnivåer skulle detta innebära, att man inte ens med en så enkel bedömningsprincip som 1 resp. 0 poäng på varje uppgift i ett prov kan vara säker på att mäta på intervallskalenivå. I många fall tas det hänsyn till detta och man försöker justera uppgifternas varierande svårig-

hetsgrad genom att ge olika poängtal för varje uppgift. Men problemet finns alltid där att avgöra hur en uppgift poängmässigt skall värderas i förhållande till andra uppgifter. Detta avgörande kan knappast bli annat än subjektivt.

Vi får naturligtvis inte överdriva denna problematik men den är ändå så viktig att den förtjänar ett noggrant övervägande innan man startar sitt bedömningsarbete.

Vi går nu tillbaka till resultaten av det engelska glosprovet. För att komma vidare i vårt resonemang måste vi införa en ny förutsättning:

> Vi måste antaga att Pers och Lisas prov, som ju var det samma, också har bedömts av lärarna på exakt samma sätt.

Denna nya förutsättning är ju nödvändig för att deras poäng på provet skall vara jämförbara.

Som vi minns hade Lisa 12 poäng och Per 10 poäng på samma prov som bedömts på samma sätt.

Kan vi nu säga att Lisa har ett bättre ordförråd i engelska än Per? (Kalle bortser vi ifrån eftersom han haft ett annat och svårare prov).
Nej!

3. Det kan mycket väl vara så att Per faktiskt har ett bättre ordförråd i engelska. På det här provet har emellertid Lisa haft en enorm tur medan Per hade "lite" otur. Provet råkade innehålla flera ord som Lisa repeterade dagen innan medan Per inte fick något av de ord som han repeterat. Lisa gissade dessutom rätt på ett par ord som hon var väldigt osäker på medan Per i sin nervositet råkade förväxla två ord som stavas nästan lika. Till råga på allt elände kände sig Per lite halvdålig den dag som provet gavs och hade svårt att koncentrera sig, och han gjorde också ett par ändringar i sitt prov så att det blev klad-

digt och läraren misstolkade hans svar. Han fick fel trots att han menade rätt.

I detta fall är det ju ingen måtta på den otur Per hade men allt som han råkade ut för är exempel på sådana tillfälligheter som kan inträffa i en provsituation och som gör att ett provresultat inte blir helt tillförlitligt. Frågan om provets tillförlitlighet kallas med en teknisk term provets RELIABILITET, och till detta skall vi också återkomma längre fram.

Vi har i denna inledning antytt tre olika frågeställningar, som man måste utreda för sig för att kunna tolka ett provresultat på ett adekvat sätt. Dessa tre frågeställningar är:

- Hur svårt är provet och hur stor är den måttenhet som provresultaten uttrycks i? (SKALNIVÅ?)
- I vilken utsträckning är ett provresultat påverkat av slumpfaktorer? (RELIABILITET?)
- Vad mäter egentligen provet? (VALIDITET?)

Det är dessa tre frågeställningar som vi kommer att diskutera i fortsättningen och vi kommer att behandla dem i den ordning som de här räknats upp.

2. SKALNIVÅ

När vi här diskuterar mätningarnas skalnivå bortser vi från de problem som hänger samman med brister i mätningarnas reliabilitet och validitet. Till dessa typer av problem återkommer vi som sagt i senare avsnitt.

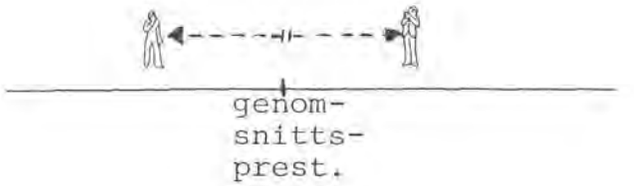
2.1 Absoluta — relativa mätningar

När vi inledningsvis angav Kalles längd till 150 cm är detta ett absolut värde. Alla som tar del av denna uppgift kan genast bilda sig en uppfattning om Kalles längd. Detta kan man göra därför att mätningen är gjord på kvotskalenivå d.v.s. den baseras på en fast nollpunkt och enhetlig måttenhet.

Uppgiften att Kalle hade 9 poäng på engelskprovet är däremot i sig omöjlig att tolka eftersom såväl nollpunkt som måttenhet är obekant.

Då det gäller beteendevetenskapliga mätningar har man i olika sammanhang försökt att definiera en fast punkt utifrån vilken mätningen görs. Ett exempel på detta är de s.k. målrelaterade mätningarna eller bedömningarna vilka mycket diskuteras i t.ex. betygssammanhang. Principen för denna typ av mätning är att man definierar en max-punkt - målet - för varje ämne i resp. årskurs. Mätningen sker sedan med avseende på avståndet till denna definierade maxpunkt.

Kursrelaterade mätningar är ett annat exempel på mätningar utifrån en definierad punkt. I detta fall skall man i stället för maxpunkten/målet definiera en genomsnittsprestation för varje ämne resp. årskurs. Mätningarna sker sedan i relation till denna genomsnittsprestation.



Tanken bakom denna princip med en fast definierad punkt utifrån vilken mätningen sker är att denna punkt skall vara riksgiltig för ett visst ämne och årskurs. Alla mätningar i ett visst ämne/årskurs skulle således kunna ske utifrån en fast punkt på samma sätt som att vi mäter en persons längd från fotsulan och därigenom ökas jämförbarheten mellan resultat från olika prov.

I praktiken medför både målrelateradeoch kursrelaterade mätningar stora svårigheter. För det första har det visat sig vara mycket svårt att definiera målet eller genomsnittsprestationen så entydigt att olika mätningar verkligen kan ske utifrån en gemensam punkt. För det andra är problemet med lika skalsteg inte löst även om det skulle vara möjligt att definiera målet/genomsnittsprestationen exakt. Man har också pekat på att denna mätningsprincip skulle förutsätta stor tillgång på centralt utarbetade prov vilket skulle medföra en hård styrning av undervisningen.

Vi saknar följaktligen referenspunkter, vilka kan definieras oberoende av det använda provet. För att tolka en individs resultat blir vi istället hänvisade till att jämföra denna individs värde med de värden som andra individer fått på samma prov. Denna princip för mätning kallas grupprelaterad eller relativ mätning. Vanligen sätter man en individs värde i relation till det medelvärde man får i en viss grupp.

Detta medelvärde får ej förväxlas med den genomsnittsprestation vi talade om i samband med kursrelaterade mätningar. Denna genomsnittsprestation anger vad en genomsnittselev i en viss årskurs omfattande hela landet verkligen kan göra oberoende av vilket prov vi använt för att mäta hans prestationer.

Det medelvärde vi använde som referenspunkt vid relativa mätningar är medelvärdet av den poäng som den grupp fått, vilken genomgått detta speciella prov. Eftersom de flesta prov bjuds till en mycket begränsad och icke riksrepresentativ grupp - vanligen den egna klassen - kan detta medelvärde variera mycket från en grupp till en annan.

Svårigheten att finna allmängiltiga referenspunkter (t.ex. giltiga för en hel årskurs elever i ett land) är en av de största nackdelarna då det gäller relativa mätningar. I brist på sådana allmängiltiga referenspunkter blir varje individs resultat, som vi sagt ovan, bedömt i relation till medelvärdet för den grupp han tillhör. Som vi vet kan det finnas stora olikheter mellan grupper och följaktligen kan två elever bli bedömda på helt olika grund.

Ett exempel:

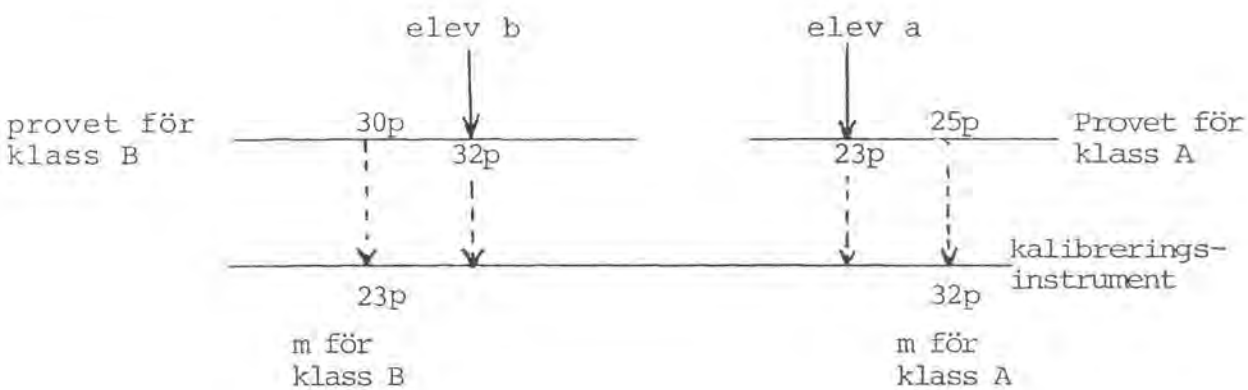
> Elev a går i klass A, vilken är "duktig" klass. Elev b går i klass B som är en medelmåttig klass. Båda klasserna har haft prov på samma avsnitt men läraren i klass A har anpassat sitt prov till den nivå som klassen normalt håller. Likadant har läraren i klass B gjort och följaktligen får eleverna i klass A ett svårare prov än eleverna i klass B.
>
> Medelvärdet i klass A är 25 poäng och i klass B 30 poäng. Elev a fick 23 poäng och elev b fick 32 poäng.

Om vi helt aningslöst ser till poängsumman verkar ju b ha lyckats bäst. Men så aningslösa är vi ju inte. Vi vet ju att mätning är relativ och följaktligen måste vi sätta de båda elevernas resultat i relation till medelvärdet på resp. prov. a ligger då 2 poäng under medelvärdet för sin grupp medan b ligger 2 poäng över medelvärdet för sin grupp. Fortfarande förefaller b att vara bäst.

Men i denna jämförelse har vi ju inte tagit hänsyn till att de två proven var olika svåra. För att få ett begrepp om hur skillnaden är mellan de två klasserna konstruerar vi ett nytt prov på samma avsnitt och ger detta till båda klasserna. Detta nya prov skall således fungera som ett "kalibrerings-instrument". För att förenkla vårt resonemang förutsätter vi att poängsättningen har följt samma principer på alla tre proven så att vi har lika stora skalsteg.

På vårt kalibreringsinstrument får klass A medelvärdet 32 medan klass B får 23 poäng.

Vi sammanfattar resultaten:



När vi nu justerat för provens olika svårighetsgrad och därigenom kan jämföra elev a och b på samma skala finner vi att a har bättre kunskaper än b!

Denna slutsats kan vi emellertid inte dra förrän vi fått tillgång till ett kalibreringsinstrument med vars hjälp vi kan justera för gruppernas olika prestationsnivå.

I Sverige är de centrala proven avsedda att utgöra kalibreringsinstrument med vars hjälp lärare kan få en uppfattning om hur den egna elevgruppen ligger i förhållande till en riksrepresentativ grupp. Med ledning av dessa uppgifter skall sedan läraren justera sin betygssättning så att den blir jämförbar över hela riket. Man kan emellertid ställa sig tveksam till om de centrala proven verkligen åstadkommer en för hela riket jämförbar betygs-

sättning. För det första ges centrala prov endast i vissa - låt vara de viktigaste - ämnena. Från dessa ämnen generaliseras sedan resultaten till andra ämnen som saknar centrala prov. En sådan generalisering från ett ämne till ett annat medför alltid en viss grad av osäkerhet.

För det andra ges centrala prov endast i vissa årskurser. Härigenom skall varje prov täcka in ett ganska brett kunskapsfält. Eftersom proven ändå är av begränsad omfattning kan de endast omfatta ett litet stickprov av alla de uppgifter, som skulle kunna ingå i provet. Risken finns således att provens innehåll kan bli bättre anpassat till vad som betonats i vissa klasser jämfört med andra. Dessa båda bristfälligheter med centrala proven torde knappast vara så stora att de allvarligt begränsar provens värde. Ett betydligt större problem är att det tycks råda tveksamhet om hur de centrala provens resultat skall användas. Denna tveksamhet visar sig i att de justeringar som görs i betygshänseende ofta blir för små. En bidragande orsak härtill torde vara att man känner sig alltför hårt bunden till de procenttal som anges för varje betygssteg för hela landets elever i en viss årskurs och med gemensam studiekurs. Dessa procenttal, som är hämtade från normalfördelningen och som anger att 7% av alla elever skall ha betyget 1, 24% betyget 2, 38% betyget 3, 24% betyget 4 och 7% betyget 5 gäller som vi sagt för hela landets elever i en viss årskurs och med gemensam studiekurs.[1] Däremot torde just dessa %-tal återfinnas inom en enskild klass endast i undantagsfall. Detta hänger samman bl.a. med att klasserna inte är slumpmässigt sammansatta utan sammansättningen är påverkad av olika selektiva mekanismer.

Sammantaget gör detta att elever i svaga klasser tenderar att bedömas lite för välvilligt medan elever i positivt gallrade klasser bedöms hårdare.

Totalt sett förefaller betygssättningen göras lite för generöst. I en undersökning utförd på ett riksrepresentativt urval normalåriga elever i årskurs 6 var medelbetyget i svenska resp. matematik inte 3.0 såsom normalkurvan föreskriver utan 3.2.

1) I Lgr 80 anges endast medelbetyget 3 för hela landet däremot inga %-tal. Genom att slopa %-talen finns nu inga direktiv för betygsskalans standardavvikelse för hela landet. Därigenom har man tagit bort alla möjligheterna att nå jämförbarhet mellan betyg i olika klasser.

Dessa brister i bedömningarna skulle emellertid undvikas om man modifierade sin betygssättning helt i enlighet med de värden som anges av de centrala proven.

Ett annat allvarligt problem med de relativa mätningarna är att vi saknar kalibreringsinstrument för att åstadkomma jämförbarhet mellan olika kurser t.ex. allmän och särskild kurs. Det torde inte vara en alltför våghalsig gissning om man antar att en och samma elev kommer att få ett bättre betyg om han väljer allmän i stället för särskild kurs i ett visst ämne.

Problemen med den ofta otillräckliga justeringen för skillnader mellan klasser med olika prestationsnivå och problemet med att åstadkomma jämförbarhet mellan olika kurser blir särskilt allvarligt i urvalssammanhang. Härvid jämställs ju ofta betyg som avser olika kurser eller betyg från olika linjer trots att dessa linjer skiljer sig åt en hel del vad gäller elevsammansättning.

Mot denna bakgrund är det knappast förvånande att valet av fortsatt utbildning i många fall avgörs mera med hänsyn till möjligheter att uppnå en hög betygsnivå än med hänsyn till hur väl utbildningen förbereder för nästa utbildningssteg. Flykten från gymnasieskolans Na-linje är bara ett exempel på detta. Sammanfattningsvis kan vi konstatera att de relativa mätningarna inte anger vad en individ verkligen kan utan endast vad en individ har presterat i jämförelse med andra individer. Det sätt på vilket en individs prestation bedöms blir härigenom åtminstone delvis en funktion av på vilken prestationsnivå jämförelsegruppen står.

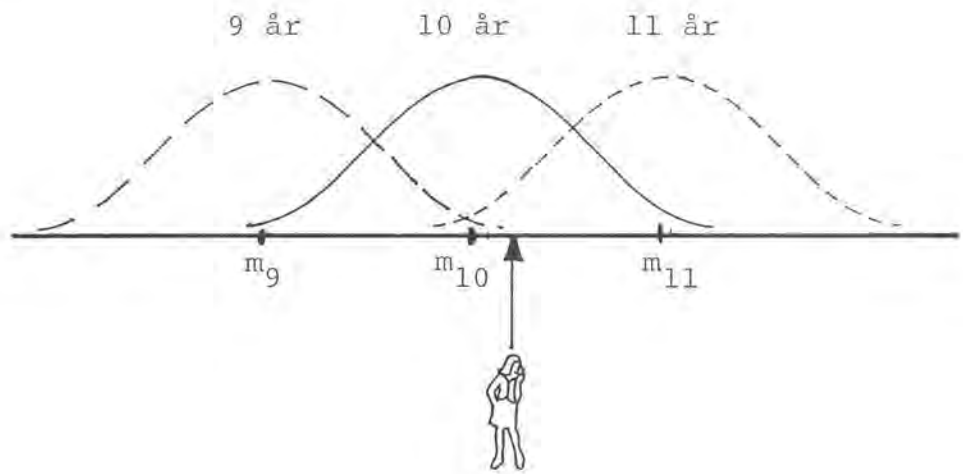
2.2 Normering

Av det vi hittills sagt om relativa mätningar framgår att en individs resultat på en mätning av denna typ blir meningsfull först efter det att man på empirisk väg skaffat jämförelsevärden. Att skaffa sig sådana jämförelsevärden kallas med en teknisk term normering. Normeringen kan ske på två olika sätt: antingen undersöker man hur olika grupper presterar på ett visst prov och individen jämförs med de olika gruppernas

genomsnittsprestation eller också jämförs individen med hur andra individer i hans egen grupp presterat på provet.

Åldersnormer innebär att man låter stora och för varje åldersgrupp representativa stickprov genomgå provet. Medelvärdet för varje grupp beräknas. När vi fått individens resultat jämförs detta med medelvärdena för de olika åldersgrupperna. Individens resultat anges sedan som den åldersgrupp vars medelvärde hans resultat ligger närmast. Ligger individens värde närmast medelvärdet för 10-åringar får han åldersnormen 10.

Figur 1.



Fördelen med åldersnormer är främst att de är relativt lätta att förstå även för en lekman. Det är kanske av det skälet som man åtminstone tidigare angav storleken på barnkläder i termer av åldersnormer. Detta görs nu allt mera sällan därför att denna typ av normering har stora nackdelar.

En sådan nackdel är att skalan har varierande storlek på skalstegen. Eftersom tillskottet i de flesta variabler tenderar att avta med stigande ålder blir avståndet mellan varje normvärde mindre och mindre ju högre upp på skalan vi kommer. Många variabler upphör helt att tillväxa efter en viss åldersgräns och då blir självfallet åldersnormerna meningslösa. - Vi köper t.ex. inte ett par byxor till en 50-åring.

Åldersnormerna är också ett mycket grovt mått. I många fall är steget mellan två åldersgrupper stort och man väljer ju det medelvärde som en individ kommer närmast. Följaktligen kan två individer ha ganska olika värden på provet men ändå få samma normvärde.

De här angivna nackdelarna tillsammans med att det är oerhört arbets- och kostsamt att fastställa åldersnormer, gör att de numera knappast använts. Dock var det denna typ av normer som låg till grund för den välbekanta intelligenskvoten (IK). I det kanske mest kända intelligenstestet grupperades uppgifterna efter åldersnivåer och den individ som testades tilldelades ett visst antal intelligensmånader för varje uppgift som löstes riktigt. På detta sätt erhölls en intelligensålder vilken med tiden kom att divideras med levnadsåldern. Den individ som fick samma intelligensålder som levnadsåldern erhöll då kvoten 1. Genom att multiplicera kvoten med 100 kom således en normalbegåvad individ att få IK=100.

Klassnormer fastställs på samma sätt som åldersnormer men i st.f. att beräkna medelvärden för olika åldersgrupper beräknas dessa på representativa urval elever ur varje årskurs. För- och nackdelar är i princip de samma som med åldersnormer och ilikhet med dessa är klassnormerna mycket sällsynta.

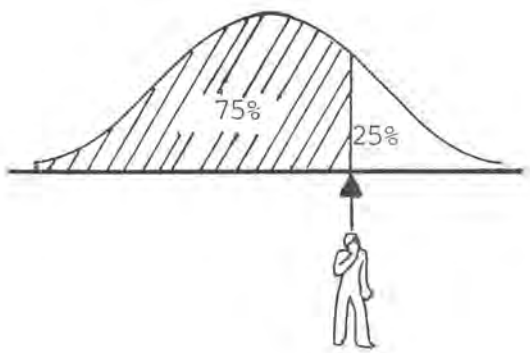
Percentiler anger hur stor procent av en viss grupp som hamnar på eller under en viss poäng på ett prov.

Antag att jag har givit ett visst prov till en grupp elever i en viss årskurs. Resultatet blev:

Poäng	N	%	Kumulativa %-tal = percentil
0	2	4	4
1	6	12	16
2	9	18	34
3	10	20	54
4	9	18	72
5	8	16	88
6	6	12	100
	Tot.50		

Percentilerna anger hur stor procent av en viss grupp (normeringsgruppen) som ligger på eller under en viss poäng. I vårt fall motsvarar poängen 2 percentil 34 därför att på värdet 2 eller därunder ligger 4 + 12 + 18% = 34% av gruppen. En individ som får percentilvärdet 75 har följaktligen presterat lika bra som eller bättre än 75% av den grupp jämförelsen avser.

Figur 2.



Percentilerna hjälper oss att bedöma hur väl en individ har lyckats på provet. Om jag anger att en individ hade 2 poäng på provet säger det inget om hans prestation. Om jag däremot uttrycker hans prestation i form av percentiler och säger att han fick percentil 34 ger jag betydligt mer information. Då anger jag ju att han är lika bra eller bättre än 34% av gruppen men samtidigt att han överträffats av 2/3 av gruppen. Alltså ingen lysande prestation i förhållande till gruppen.

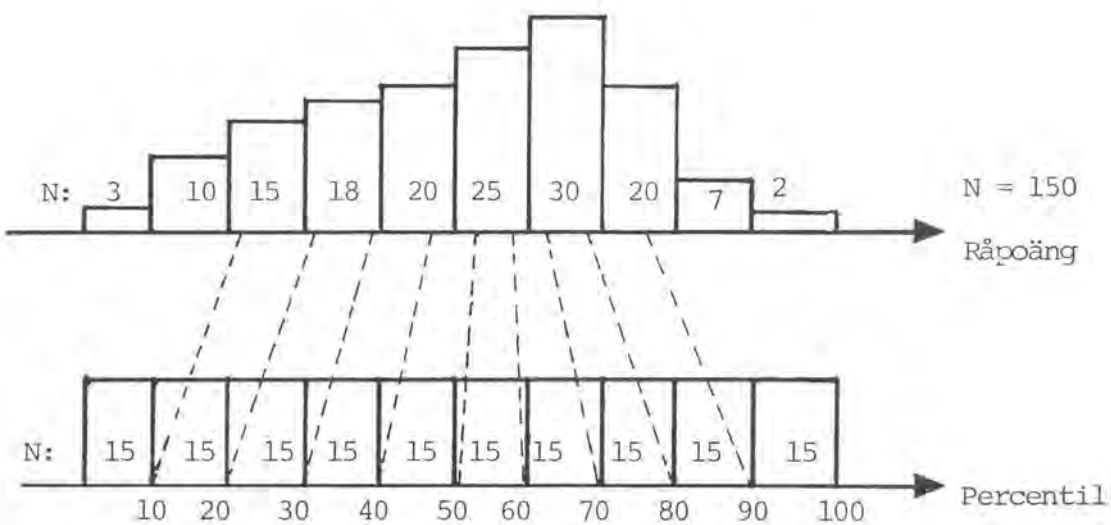
Vilket informationsvärde percentilerna får är naturligtvis beroende av vilken normgrupp jag har att relatera resultaten till. Om jag anger att en individ får percentilvärdet 34 för sin klass har jag givit mer information än om jag bara anger råpoängen. Kan jag säga att råpoängen motsvarar percentilvärdet 34 för elever i årskurs 6 i Göteborg har jag ökat informationsvärdet betydligt och ännu mer om jag kan säga att poängen motsvarar percentil 34 för årskurs 6 i Sverige.

Det största problemet med percentilskalan är att den inte ger

uttryck för enhetliga skalsteg. Som framgår av figuren nedan krävs det en större förändring mätt i råpoäng för att förändra sig från percentil 10 till 20 eller från percentil 90 till 100 jämfört med vad som krävs för att förändra sig från 40 - 50. Detta förklaras av att vi ofta konstruerar våra prov så att vi får en anhopning av individer i skalans centrala delar medan frekvenserna avtar ju längre ut mot skalans ändpunkter vi komner.

Figur 3.

TRANSFORMERING FRÅN RÅPOÄNG TILL PERCENTILER



De stora skalstegen i ytterkanterna leder naturligtvis till att skalan blir inexakt. Denna brist kan emellertid undvikas genom att byta jämförelsegrupp i de fall då det finns flera sådana grupper. Om en individ får percentilen 90 för 4-åringar kan denna uppgift kompletteras genom att ange vilket percentilvärde han får i jämförelse med 5-åringar.

Åldersnormer och klassnormer bygger således på den principen att en individs resultat jämförs med hur olika grupper presterat på samma prov. Detta gäller också för percentilerna om man skall uppnå en acceptabel precision i skalans ytterkanter. Följaktligen blir normeringsarbetet i dessa fall så omfattande och

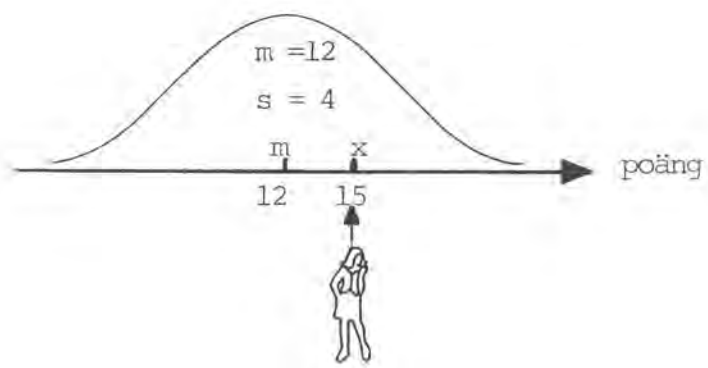
kostsamt att det normalt kan göras enbart för sådana prov som har en generell användning.

Som vi angivit har samtliga dessa normer också den nackdelen att skalstegen varierar vilket starkt begränsar deras användbarhet.

En typ av normer som undviker båda dessa problem är Z-skalan.

Z-skalan innebär att man jämför en individs resultat med normeringsgruppens medelvärde. Avvikelsen från detta medelvärde uttrycks med normeringsgruppens standardavvikelse som mätenhet.

Figur 4.



Om vi uttrycker detta i en statistisk formel får vi:

$$Z = \frac{x - m}{s}$$

där x anger individens poäng
m jämförelsegruppens medelvärde
s jämförelsegruppens standardavvikelse

Antag att en elev har fått 15 poäng på ett prov, där klassens medelvärde är 12 och standardavvikelsen i klassen är 4. Elevens z-värde blir då

$$Z = \frac{15 - 12}{4} = 0,75$$

Genom att uttrycka hans resultat i z-värde i st.f. i råpoäng tillför vi ytterligare information om hur väl eleven lyckats.

Vi vet nu att han måste ha presterat ett bättre resultat än genomsnittet i gruppen eftersom hans z-värde är positivt. Av formeln framgår ju att en individ som har samma poäng som medelvärdet får z-värdet 0 medan poäng under medelvärdet ger negativa z-värden.

M.a.o. anger tecknet på z-värdet hur en individ presterat i förhållande till jämförelsegruppens medelvärde.

Denna information hade vi i och för sig kunnat få genom att nöja oss med att ange att han ligger 3 poäng bättre än genomsnittsresultatet för jämförelsegruppen. Men med endast denna information kvarstår ju problemet att tolka vad 3 poäng innebär prestationsmässigt. Om mätningen är gjord med stora skalsteg innebär detta att vår elev har presterat betydligt bättre än genomsnittet. Är däremot skalstegen små är prestationen inte alls lika lysande.

Det är just för att få information om skalstegens storlek som vi har med standardavvikelsen i formeln för Z. Standardavvikelsen påverkas nämligen av skalstegens storlek så att ju större skalstegen är vid en mätning desto lägre blir standardavvikelsen.

Ett exempel:

Vi väger en grupp individer och får följande resultat:

a) uttryckt i kg:

kg	N
72	4
73	6
74	10
75	12
76	7
77	6
78	5
	50

m = 75
s = 1.46

b) uttryckt i hg:

hg	N
720	4
730	6
740	10
750	12
760	7
770	6
780	5

m = 750
s = 14.57

Som framgår av exemplet ovan blir standardavvikelsen 10 gånger så hög när mätningen sker i hg som i det fall då vi väljer kg som mätenhet.

En individ som väger 72 kg ligger i exempel a 3 enheter under medelvärdet och får $Z = \frac{72-75}{1.46} = -2.06$. När mätningen görs med hg som enhet ligger han 30 enheter under medelvärdet men även i detta fall blir $Z = \frac{720-750}{14.57} = -2.06$ eftersom vi nu dividerar med en 10 gånger så hög standardavvikelse. Vi kan uttrycka det så att vi med hjälp av standardavvikelsen justerar för olikheterna i skalsteg, och därigenom blir z-värdena jämförbara.

I vårt exempel är z-beräkningen självfallet onödig eftersom vi utgår från en standardiserad skala och därför känner såväl nollpunkt som mätenheternas storlekar (hg och kg).

Vid beteendevetenskapliga mätningar däremot saknar vi normalt kännedom om såväl fast nollpunkt som mätenheternas storlek, vilket vi ju diskuterat tidigare. Som ersättning för denna fasta nollpunkt har z-värdena jämförelsegruppens medeltal och under antagandet om att jämförelsegruppens medelvärde utgör en lika bra prestation på två olika mätningar kan vi göra direkta jämförelser mellan två olika beteendevetenskapliga mätresultat för en och samma individ då resultaten uttrycks i Z.

Kalle hade ju 9 poäng på provet i engelska. På ett matematikprov fick han 24 poäng. Vi ställer oss nu frågan: På vilket av de två proven har han lyckats bäst?

Att direkt jämföra de två råpoängen är lika tokigt som att fråga sig vem av två personer är störst: den som väger 100 kg eller den som är 2 m lång.

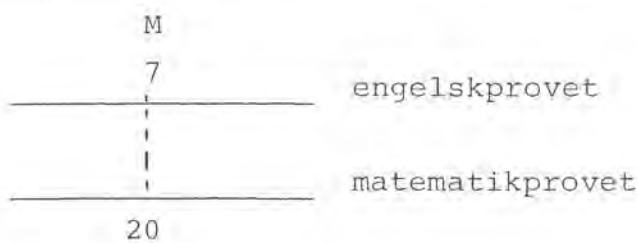
Råpoängen i sig säger ju inget om Kalles prestation därför att vi inget vet om vare sig skalornas nollpunkter eller om storleken på deras skalsteg.

För att kunna jämföra de två resultaten måste vi ha två informationsbitar:

1. klassens medelvärden på de två proven
2. klassens standardavvikelse på de två proven

Låt oss säga att klassens medelvärden är 7 poäng på engelskprovet och 20 poäng på matematikprovet.

Eftersom vi antar att klassens medelvärde utgör en lika bra prestation på de två proven kan vi således med hjälp av de två medelvärdena kalibrera skalorna:



7 poäng på engelskprovet motsvarar således 20 poäng på matematikprovet.

Kalle ligger ju 2 poäng över medelvärdet på engelskprovet och 4 poäng över medelvärdet på matematikprovet. Vilket resultat som är bäst blir nu beroende av hur mycket kunskaper som krävs för att höja sitt resultat 1 poäng på de två proven eller med andra ord hur stora skalstegen är.

Som vi angav tidigare ger standardavvikelserna besked om detta. Låt oss säga att klassens standardavvikelser var 2 på engelskprovet och 5 på matematikprovet. Vi justerar då för skalornas varierande skalsteg genom att dividera avvikelsen från medelvärdena med resp. standardavvikelse och vi får

för engelskprovet: $\frac{+ 2 \text{ poäng}}{2} = +1$

och för matematikprovet: $\frac{+ 4 \text{ poäng}}{5} = +0.8$

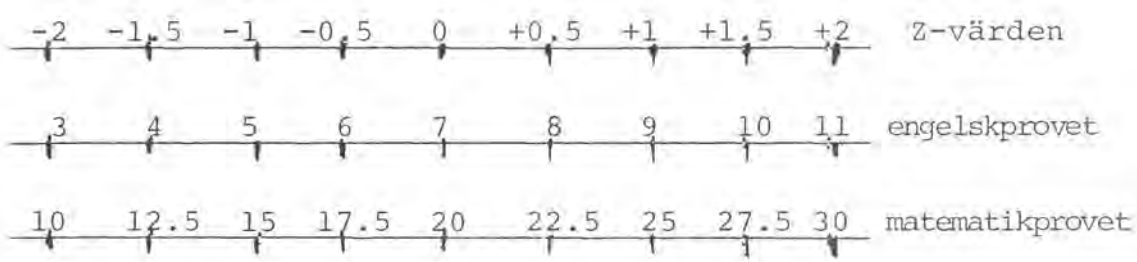
De två värdena +1 och +0.8 är ju Z-värden:

för engelskprovet $\frac{9 - 7}{2} = +1$

för matematikprovet $\frac{24 - 20}{5} = +0.8$

och eftersom de två råpoäng nu uttrycks i en och samma skala är de jämförbara. Kalle lyckades alltså bäst på engelskprovet. OBS att jämförbarheten gäller endast under förutsättning att m och s för de två proven är beräknade på samma jämförelsegrupp. I detta fall var det ju Kalles klass som utgjorde jämförelsegrupp på båda proven.

Genom att överföra värdena på de två proven till Z-värden kan vi göra jämförelser mellan vilka värden som helst på de två proven:



Här gäller den generella principen att två värden motsvarar varandra om de ligger lika många standardavvikelser från medelvärdet och på samma sida om medelvärdet.

Z-skalan kännetecknas av att dess medelvärde alltid är 0 och dess standardavvikelse är 1. Anledningen till att Z-skalan får m=0 är att dess värden utgör en avvikelse från ett medelvärde (x-m) och summan av alla avvikelser från medelvärdet är alltid 0. Standardavvikelsen =1 får vi därför att vi dividerar råpoängens avvikelse från medelvärdet med den standardavvikelse som gäller för råpoängen. Det är nämligen så att om alla värden multipliceras med en konstant så förändras standardavvikelsen så som konstanten anger. När vi beräknar Z-värdena dividerar vi ju alla värden med poängens standardavvikelse och därigenom blir standardavvikelsen för Z: $\frac{s}{s} = 1$.

Dessa två egenskaper, m=0 och s=1, medför ett par praktiska nackdelar med Z-skalan. Medelvärdet = 0 gör ju att alla värden som ligger under medelvärdet blir negativa och vi får således arbeta med positiva och negativa värden i en och samma grupp. Den låga standardavvikelsen (1) anger ju som vi sagt ovan att Z-skalan har stora skalsteg. För att få en tillräckligt hög precision i Z-värdena räcker det inte med att ange dessa i

heltal utan vi måste tillgripa decimaler - ibland två decimaler.

Nackdelen med decimalerna kan vi ju råda bot för genom att vi utgår från Z-värdena och multiplicerar varje sådant värde med en konstant t.ex. 10. Eftersom varje individs Z-värde behandlas på samma sätt åstadkommer vi ingen orättvisa mot någon. Varje individ bevarar sin position i gruppen men avståndet mellan individerna blir 10 gånger större efter denna operation jämfört med vad det var innan. Eftersom avståndet mellan individerna härigenom blivit 10 gånger så stort har vi också gjort standardavvikelsen 10 gånger så stor som den var för Z-värdena. D.v.s. att standardavvikelsen nu är 10 i st.f. 1.

Kvar står emellertid problemet med att värden under medelvärdet fortfarande är negativa. För att komma undan detta kan vi addera en konstant till alla värden vi fått efter att ha multiplicerat värdena med 10. Vi väljer då att till varje värde addera konstanten 50. Inte heller nu begår vi någon orättvisa. Varje individ får ju ett tillägg av 50 poäng, vilket innebär att vi flyttar hela gruppen 50 steg upp på skalan. Eftersom alla värden flyttats 50 steg upp följer naturligtvis medelvärdet med lika många steg. Z-skalans medelvärde var ju 0. Det nya medelvärdet blir därför 50. Förhoppningsvis hamnar nu ingen individ under 0 poäng utan alla värden är positiva. Vi har alltså genom dessa operationer ändrat medelvärdet från 0 till 50 och standardavvikelsen från 1 till 10. Följaktligen kan vi inte kalla de nya värdena Z-poäng. En skala med egenskaperna m=50 och s=10 kallas T-skalan.

En annan skala som främst används i samband med intelligensmätningar är IK-skalan, vars m = 100 och s = 15.

Överföringen av värden från en skala till en annan kallas lineär skaltransformation. En sådan transformation kan ske till en skala med vilka egenskaper som helst. Man är således inte bunden till att välja IK-skalan, T-skalan eller Z-skalan. Dock innebär den lineära skaltransformen alltid att man först

för över värdena till Z-poäng och därefter multiplicerar dem med en konstant och adderar en annan konstant. Eftersom Z-skalans standardavvikelse är 1 blir den konstant, med vilken man multiplicerar alla värden, standardavvikelsen i den nya skalan. Den konstant, som adderas till alla värden, blir medelvärde i den nya skalan eftersom Z-skalans medelvärde är 0.

Vi kan sammanfatta den liniära skaltransformationen med följande formel:

$$Y = S_y \cdot \frac{x - m_x}{S_x} + m_y$$

där

Y = individens värde i den nya skalan

S_y = standardavvikelsen i den nya skalan

x = individens värde i den gamla skalan

S_x = standardavvikelsen i den gamla skalan

m_x = medelvärde i den gamla skalan

m_y = medelvärde i den nya skalan

Observera att uttrycket $\frac{x - m_x}{S_x}$ är formeln för Z. Förenklat kan vi skriva formeln:

$$Y = S_y \cdot Z + m_y$$

Uttryckt på detta sätt beskriver formeln just de två operationer vi genomförde ovan då vi multiplicerade alla Z-värden med en konstant och sedan adderade en ny konstant. Kalles matematikprov gav m = 20 och S = 5. De olika poängen på detta prov kan nu med hjälp av formeln ovan översättas till värden i T-skalan eller IK-skalan.

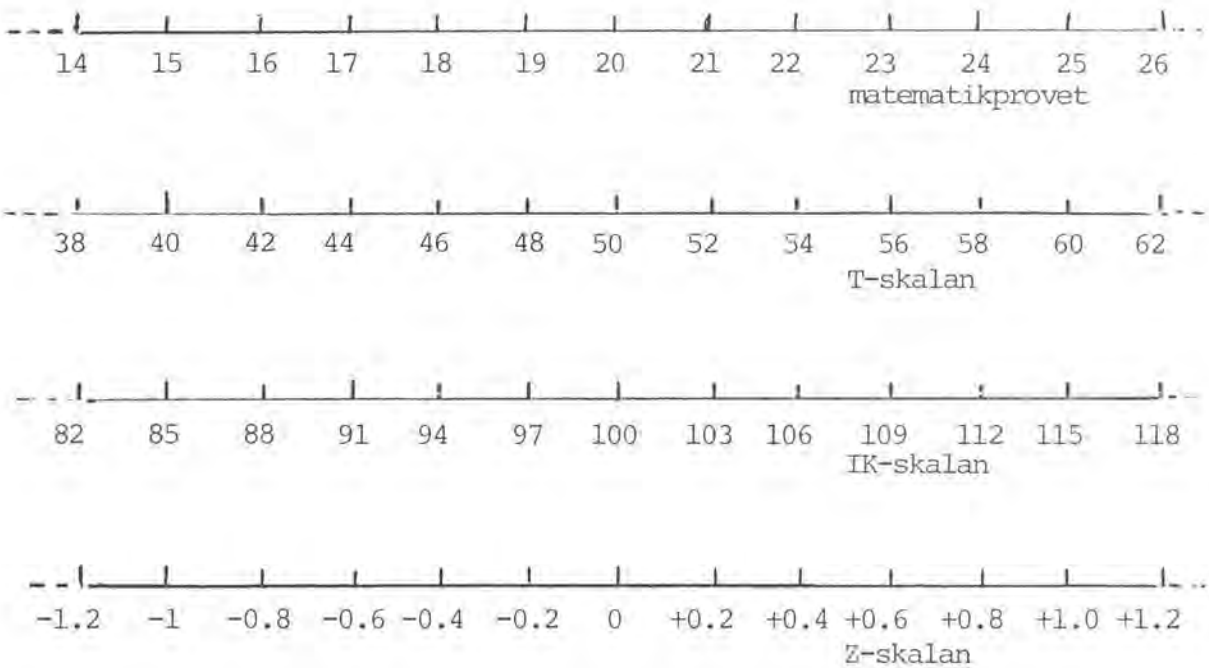
Kalle hade ju 24 p på provet. Vilken poäng skulle han få om resultatet uttrycktes i T-skalan?

$$y = 10 \cdot \frac{24 - 20}{5} + 50 \qquad y = 10 \cdot 0.8 + 50 = 58 \text{ poäng}$$

i IK-skalan?

$$y = 15 \cdot \frac{24 - 20}{5} + 100 \qquad y = 15 \cdot 0.8 + 100 = 112 \text{ poäng}$$

Genom att upprepa dessa beräkningar för vilket resultat som helst på matematikprovet kan dessa resultat översättas till T-poäng, IK-poäng, Z-poäng eller till en skala med vilka andra egenskaper som helst.



Som framgår av skalorna ovan motsvaras en förändring av 1 poäng i matematikprovet av 2 poängs ändring i T-skalan och av 3 poäng i IK-skalan. Förklaringen till att vi får just dessa relationer är att T-skalan har 2 gånger så hög standardavvikelse (10) som råpoängen (5) och IK-skalan har 3 gånger så hög standardavvikelse (15) som råpoängen.

Den lineära skaltransformationen innebär inte att vi ändrar variationen i den grupp som studeras. Vi mäter endast den befintliga variationen med hjälp av skalor som har olika stora skalsteg eller mätenheter. Jämför när vi uttryckte vikten i kg resp. hg.

Vad skall nu den lineära skaltransformationen tjäna för syften? En anledning till att göra en sådan transformation har vi redan nämnt:

> att föra över resultat, som uttrycks i olika skalor till en gemensam skala för att därigenom kunna göra jämförelsen mellan t.ex. en individs resultat på två eller flera olika prov.

Som vi angav tidigare förutsätter detta att medelvärden och standardavvikelser är beräknade på en och samma grupp.

Den lineära skaltransformationen är också till hjälp då vi vill slå samman flera olika mätvärden till en totalpoäng. Vid en sådan sammanslagning kommer nämligen de olika delresultaten att påverka skillnaderna i totalpoäng i enlighet med sina standardavvikelser.

Vi tar ett exempel för att förklara detta förhållande.

Jag har givit två prov till en klass. Med ledning av dessa två prov skall jag åstadkomma en rangordning av eleverna. De två provens standardavvikelser beräknade på hela klassen är:

prov I: $s = 4$
prov II: $s = 2$

Tre av eleverna i klassen får följande resultat på de två proven:

	I	II	Total poänst (I+II)
A	10	7	17
B	12	6	18
C	14	5	19

När jag på detta sätt slagit samman resultaten på de två proven får C den högsta totalpoängen och A den lägsta. Denna rangordning stämmer exakt med resultatet på prov I men den är samtidigt rakt motsatt den rangordning som prov II ger. Prov I har alltså helt bestämt rangordningen i totalpoängen. Förklaringen till detta är att prov I har en dubbelt så hög standardavvikelse (4) som prov II (2).

Kanske är den rangordning jag fått i totalpoängen helt korrekt. Om så är fallet har jag bedömt att prov I mäter ett område vilket är dubbelt så viktigt som det område prov II utgör mått på.

Antag i stället att jag anser båda områdena vara lika viktiga. Då skall rimligtvis också båda proven få lika stor inverkan på totalpoängen.

För att åstadkomma en lika vikt för båda proven måste jag se till att de två proven får samma standardavvikelse. En sådan viktförändring kan jag åstadkomma genom att multiplicera alla värden i prov I med konstanten $\frac{1}{2}$. Då förändras ju standardavvikelsen i enlighet med konstanten och den blir $4 \times \frac{1}{2} = 2$ d.v.s. lika stor som för prov II.

Resultatet blir:

	I	II	Total poäng
A	$10 \cdot \frac{1}{2} = 5$	7	12
B	$12 \cdot \frac{1}{2} = 6$	6	12
C	$14 \cdot \frac{1}{2} = 7$	5	12

När vi låter båda proven få samma vikt får alla tre eleverna samma totalpoäng och rangordningen blir helt olik den då vi bara slog samman råpoängen.

Ett alternativt sätt att ge båda proven samma vikt är att multiplicera resultaten på prov II med 2 varigenom prov II får s = 4. I detta fall skulle samtliga tre elever få 24 poäng i totalsumma.

Det resonemang som här förts med två olika prov gäller också för olika delar inom ett och samma prov. Ibland ser man prov där olika frågor bedöms med olika poängsättning. Vissa frågor bedöms med 0 resp. 1 poäng medan andra frågor bedöms 0 - 3 poäng. Tanken bakom en sådan differentierad bedömningsprincip är att de frågor som poängsätts från 0 - 3 skall få större inverkan på slutresultatet än frågor som bedöms med 0 resp. 1 poäng. Normalt blir effekten också i enlighet med intentionerna men man kan inte vara helt säker på att så är fallet. Det kan nämligen inträffa att alla elever får i stort sett samma poäng även på en fråga med en med mycket starkt differentierad skala. Om så är fallet kommer frågan att ha en liten inverkan på totalresultatet. Det är med andra ord inte den möjliga variationen som bestämmer en frågas vikt för resultatet utan vikten avgörs av den variation som frågan verkligen resulterar i.

Anledningen till att standardavvikelserna avgör det provs vikt i en totalpoäng är den att vi vid summeringen av resultaten från de olika proven tillmäter alla poäng samma värde oavsett hur stor mätenhet dessa poäng representerar. Som vi sagt upprepade gånger är det ju ett omvänt förhållande mellan mätenhetens storlek och standardavvikelsen. Genom att då enbart summera poäng från olika prov utan att ta hänsyn till hur stor mätenhet en poäng representerar kommer de poäng som erhålls på prov med hög standardavvikelse att övervärderas medan de poäng som härrör från prov med låg standardavvikelse undervärderas vid sammanslagningen.

Den lineära skaltransformationen medför inte att vi ändrar formen på den fördelning som råpoängen ger upphov till. Är fördelningen i råpoäng sned så kommer fördelningen i den nya skalan att vara precis lika sned. I vissa fall kan man vilja förändra en sned fördelning så att den blir mera symmetrisk och då väljer man vanligen att anpassa fördelningen till normalkurvan. Den process då man anpassar en viss fördelning av resultat till en normalfördelning kallas normalisering.

2.3 Normalisering

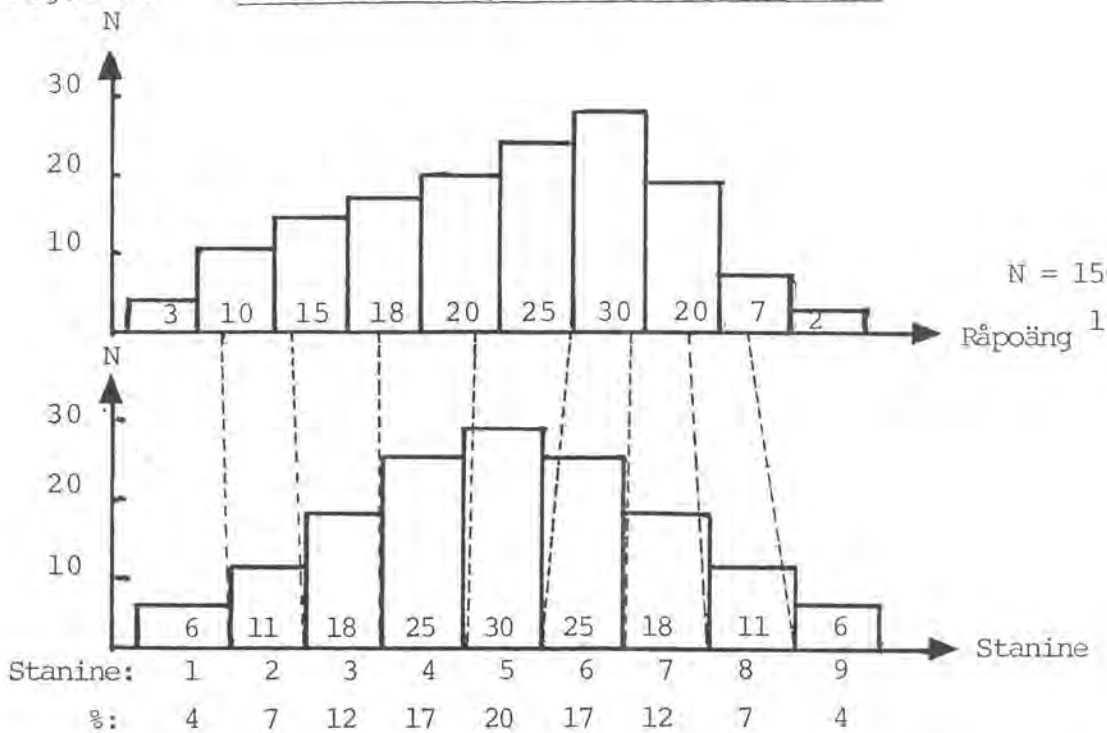
Normalisering innebär att man med hjälp av en normalfördelningstabell bestämmer hur stor andel av gruppen som skall hamna på varje skalvärde.

Ett exempel på en sådan skala är den niogradiga Stanineskalan. Enligt denna skala skall gruppen fördelas på de olika skalstegen enligt nedan.

Stanine-värde:	1	2	3	4	5	6	7	8	9
Andel av gruppen (%)	4	7	12	17	20	17	12	7	4

Dessa procenttal är bestämda i förväg och valda så att man får en fördelning som överensstämmer med normalfördelningen. Den råpoäng man fått tjänar i detta fall enbart det syftet att individerna rangordnas. Med ledning av denna rangordning fördelas individerna på stanine-värden så att de 4% med lägst råpoäng ges stanine-värde 1 nästa 7% får stanine-värde 2 o.s.v.

Figur 5. Transformering till stanineskalan:



1) Vi förutsätter här att råpoängen är klassindelad varigenom vi kan skilja mellan individer inom en och samma klass. Om värdena är exakta kan man givetvis inte skilja mellan individer som har samma poäng. Gränserna för staninevärden läggs då vid de råpoängsgränser som ger den bästa överensstämmelsen med den teoretiska stanineförde lningen.

Som framgår av figuren är fördelningen i råpoäng något negativt sned men genom att något "trycka ihop" individerna på lägre råpoäng och något "dra isär" de som ligger på höga råpoäng blir stanineförde lningen normaliserad. Uttrycket "trycka ihop" innebär att vi förminskar de skillnader som faktiskt finns och "dra isär" innebär att vi överdriver skillnaderna.

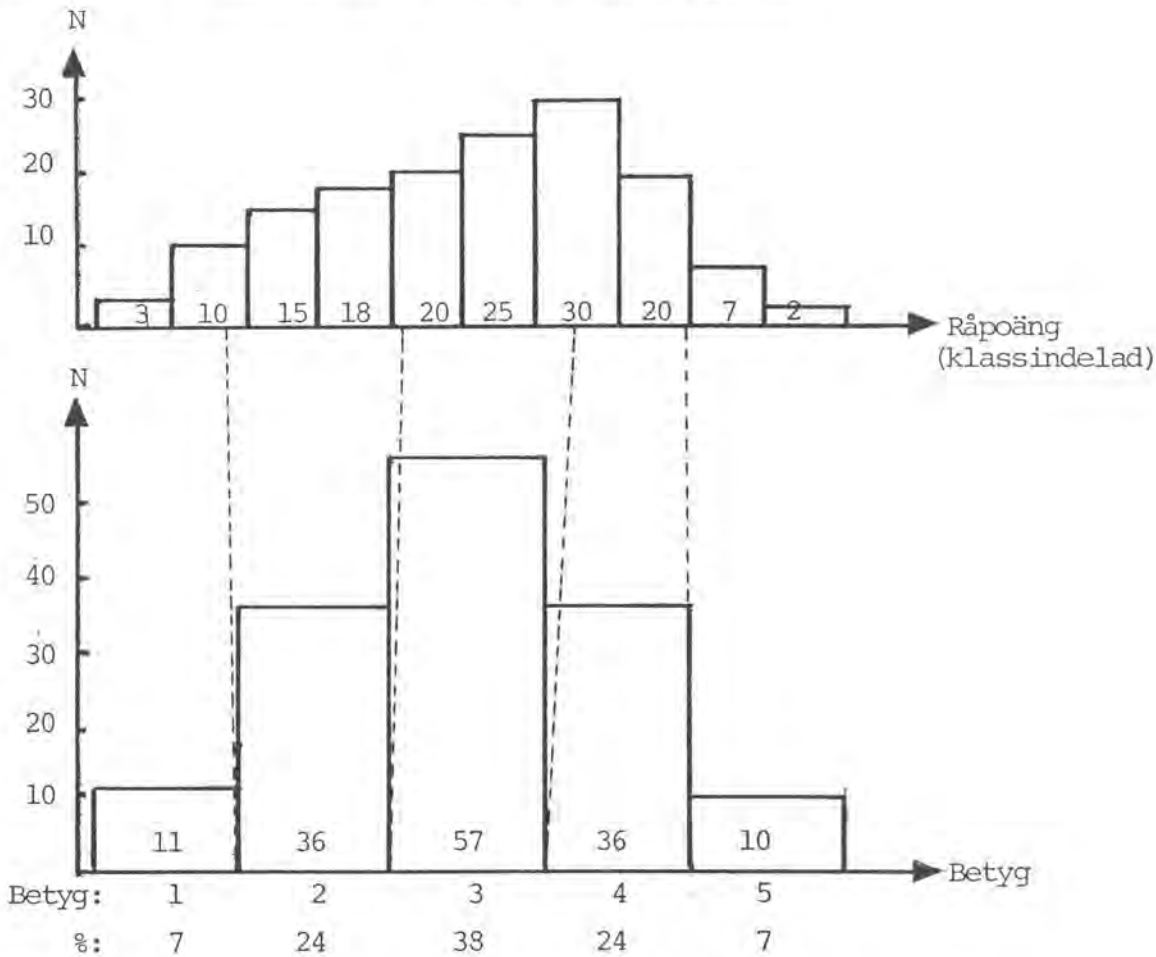
En annan skala med förutbestämda procenttal hämtade från normalfördelningen är den 5-gradiga betygsskalan. Procenttalen för de olika skalstegen är i denna skala

Betyg	1	2	3	4	5
%	7	24	38	24	7

Dessa procenttal har erhållits genom att man fört samman alla individer i en normalfördelning som ligger inom intervallet $Z = 0 \pm 0.5$. D.v.s. alla individer som finns inom en standard-

avvikelseenhet i mitten av fördelningen. Dessa individer ges betyget 3. Därefter har man tagit ut alla individer som ligger inom en standardavvikelse uppåt i fördelning resp. nedåt i fördelning, dessa har fått betyget 4 resp. 2. Återstående individer i skalans båda ytterkanter erhåller betyget 5 resp. 1.

Figur 6. Transformering till 5-gradig betygsskala



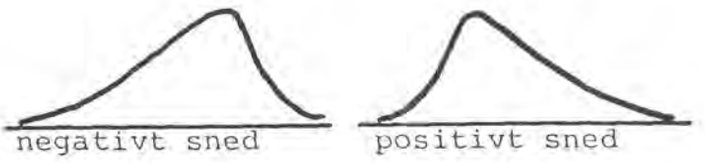
Man bör komma ihåg att såväl stanineskalan som den 5-gradiga betygsskalan är relativa skalor. De säger ingenting om vad en individ kan prestera utan anger endast vad en individ presterat i förhållande till andra individer.

Betyget 3 i ett visst ämne säger t.ex. endast att en individ är bättre än de 31% sämsta och sämre än de 31% bästa inom hela jämförelsegruppen. När det gäller betygen utgörs ju jämförelsegruppen av alla elever i Sverige i samma årskurs och som läst samma studiekurs.

Det faktum att beteendevetenskapliga mätningar ofta ger fördelningar av resultat som mer eller mindre liknar normalfördelningen kan inte tas som bevis för att de variabler vi mäter i sig är normalfördelade. För att få veta hur en viss variabel verkligen är fördelad i en grupp krävs ett mätinstrument som har exakt lika stora skalsteg över hela variationsområdet. Sådana beteendevetenskapliga mätinstrument finns inte.

I praktiken är det snarare så att vi konstruerar mätinstrument så att de ger en ungefärlig normalfördelning. Visar det sig vid en utprövning att fördelningen blir negativt sned tar man bort några lättare uppgifter och lägger till några svårare. Blir fördelningen positivt sned gör man tvärtom.

Figur 7. Negativt och positivt sned fördelning.



Genom utbytet av uppgifter rätas så fördelningen upp och närmar sig normalfördelningen.

Vid relativa mätningar, där vi främst är intresserade av skillnader mellan individer, får normalfördelningen en central plats. En sådan fördelning ger oss nämligen möjligheter att studera skillnader mellan individer på alla prestationsnivåer. Som framgår av figuren ovan medför en negativt sned fördelning svårigheter att skilja mellan elever på hög prestationsnivå och den positivt sneda differentierar dåligt mellan elever på låg prestationsnivå .

Frågan om huruvida en variabel i sig är normalfördelad eller inte kan således aldrig ges ett klart svar - dess fördelning är beroende av vilket mätinstrument som används för att mäta den.
Direkt tveksamt är det emellertid att förvänta sig normalfördelade mätresultat från relativt små grupper (t.ex. en skolklass) och i ännu högre grad gäller denna tveksamhet om grupperna är selegerade som t.ex. en klass i gymnasieskolan.

Om vi har anledning att förvänta en negativt sned fördelning, vilket bör vara fallet när vi arbetar med en positivt gallrad grupp, men ändå normaliserar fördelningen medför detta att vi tvingas till att överdriva de skillnader som faktiskt finns mellan individer på hög prestationsnivå. Skaltekniskt leder detta till att vi mäter skillnader mellan individer på hög-prestationsnivå med mindre skalsteg än de skalsteg med vilka vi mäter skillnader på låg nivå.

2.4 Slutkommentar

De beteendevetenskapliga mätningarna är inte jämförbara med fysiska mätningar. Normalt saknas såväl en absolut nollpunkt som kännedom om storleken på de mätenheter i vilka mätningar sker. Följaktligen säger ett mätningsresultat i sig ingenting innan vi genom olika tekniska åtgärder skaffat oss vissa referenspunkter. Dessa referenspunkter får vi från andra individer eller grupper. Mätningarna blir härigenom relativa - de säger något om en individs prestation i förhållande till andra individer men inget om individens prestation i sig.

En förutsättning för att kunna bedöma resultaten från en sådan relativ mätning och hantera dessa resultat adekvat är att man känner till åtminstone de mest elementära principerna för tillvägagångssättet då man skaffar referenspunkterna.

Det är just dessa elementära principer vi försökt beskriva i detta avsnitt.

Vid beteendevetenskapliga mätningar är det också betydligt svårare att uppnå en god precision än det är vid fysiska mätningar. Det är därför nödvändigt att man skaffar sig information om hur hög tillförlitlighet (reliabilitet) en viss mätning har och hur stora effekterna blir av de fel vi alltid måste räkna med vid beteendevetenskapliga mätningar. Denna typ av problem diskuterar vi i nästa avsnitt.

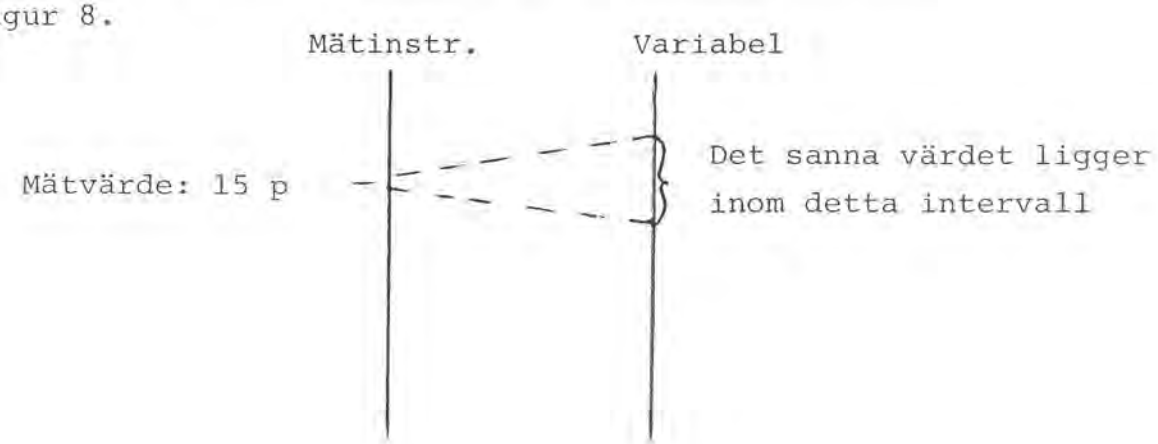
3. RELIABILITET

När vi mäter en individs längd eller vikt anger vi resultaten i exakta värden - 172 cm eller 74.3 kg. De mätinstrument vi då använder är så tillförlitliga - reliabla - att vi kan bortse från eventuella mätfel. Mätvärdet avviker knappast så mycket från den sanna vikten eller längden att det har någon praktisk betydelse.

Vid hastighetskontroller är man mer uppmärksam på att det värde radarn ger kan vara behäftat med fel. Man dömer inte en person för fortkörning om radarn registrerar 51 km/tim eftersom detta mätvärde kan vara påverkat av felfaktorer och det är fullt möjligt att bilens sanna hastighet ligger under hastighetsgränsen, d.v.s. 50 km/tim.

På samma sätt som man vid hastighetskontroller skiljer mellan det uppmätta värdet och det sanna värdet bör man också vid mätning av psykiska variabler hålla isär dessa två värden. De mätinstrument vi använder i beteendevetenskapliga sammanhang är normalt mer reliabla än om vi gör en subjektiv skattning men de har knappast lika hög precision som de mätinstrument vi använder för att mäta fysiska variabler. Detta innebär att även om mätningen resulterar i ett exakt värde t.ex. 15 poäng måste detta värde betraktas som en ungefärlig skattning av individens sanna värde. Det sanna värdet, vilket vi normalt inte känner, kan ligga högre eller lägre. Vi vet nämligen aldrig hur mätfelen inverkat på en enskild mätning.

Figur 8.



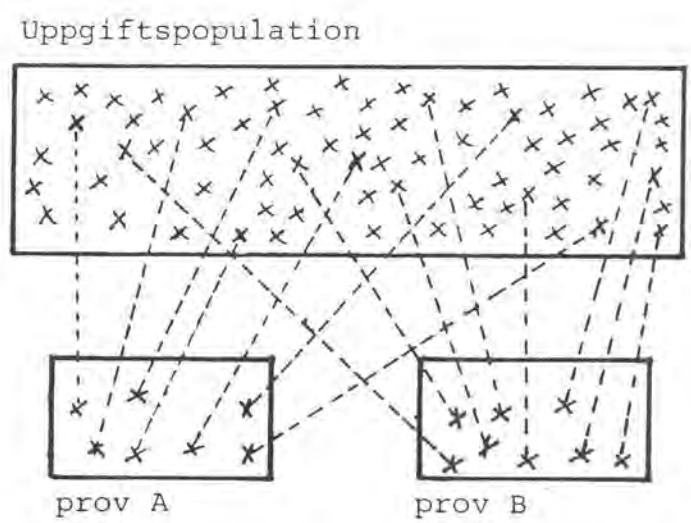
Det är självfallet det sanna värdet vi vill ange vid en mätning. Om man vill känna sig någorlunda säker på att göra detta räcket det följaktligen inte med att ange en exakt poäng. Man bör då istället ange det i form av ett intervall t.ex. 13-17 poäng. Ju lägre reliabilitet mätinstrumentet har desto större blir det intervall vi måste ta till för att med en viss säkerhet fånga in det sanna värdet.

3.1 Felkällor

De felkällor som orsakar reliabilitetsbristerna är av tre slag:

1. Variation i mätinstrumentet. Antalet uppgifter i ett mätinstrument måste av praktiska skäl begränsas. De uppgifter som finns med kan därför ses som ett stickprov draget ur en population av uppgifter, vilken omfattar alla tänkbara uppgifter för att mäta variabeln i fråga.

Figur 9.



Som alltid när vi arbetar med stickprov måste vi räkna med ett urvalsfel vilket gör att de två parallella instrumenten A och B kan ge något olika resultat. A kan innehålla någon eller några uppgifter som gynnar vissa individer medan någon eller några uppgifter i B kan gynna andra individer.

Man hör ibland personer som efter ett prov säger att de haft otur och fått frågor på sådant de varit osäkra på. Ett sådant uttryck kan vara en beskrivning av denna felkälla.

Omfattningen av denna felkälla beror delvis på hur många uppgifter mätinstrumentet innehåller. Ju flera uppgifter desto större är chansen att mätinstrumentet blir representativt för uppgiftspopulationen och desto mindre blir urvalsfelet.

Här tangerar vi validitetsproblematiken till vilken vi återkommer i nästa avsnitt.

Urvalsfelet är också beroende av hur mycket uppgifternas innehåll varierar i populationen. Ju mer uppgifterna i populationen liknar varandra d.v.s. ju mer homogen uppgiftspopulationen är desto mindre betydelse får ju själva urvalet av uppgifter. Av denna anledning brukar det fel som denna felkälla ger upphov till kallas homogenitetsfel.

2. Variation hos individen. Denna felkälla hänger samman med att en individs beteende fluktuerar något från ett tillfälle till ett annat. Vid vissa tillfällen är vi på bättre humör eller mera på allerten än vid andra tillfällen. Denna variation i "dagsformen" kan naturligtvis inverka på resultaten. Trots detta avgränsar vi inte mätresultatet till att gälla enbart för själva mättillfället utan vi gör en generalisering över tid och därmed riskerar vi att göra vissa fel.

3. Variation i mätsituationen. Till denna kategori hänförs alla de fel som kan uppstå vid själva mättillfället och i samband med bedömningen. Som exempel på denna felkategori kan nämnas

- otillbörlig hjälp eller vilseledande information från den som administrerar mätningen
- yttre störningar vid mättillfället
- feltolkning av frågor eller felmarkeringar av svar
- gissningseffekter
- bedömarens feltolkningar av svar eller felräkning av poäng

Listan på tänkbara felfaktorer i mätsituationen kan bli lång men vi nöjer oss med dessa olika exempel.

Om man avgränsar begreppet reliabilitet till att endast avse sådana felfaktorer som är direkt förknippade med mätinstrumentet är det givetvis fel att ta med "variation hos individen" och vissa felfaktorer som vi fört till kategorien "variation i mätinstrumentet". Eftersom mätresultatens tillförlitlighet även påverkas av sådana felfaktorer som ligger utanför själva mätinstrumentet har vi valt att definiera begreppet reliabilitet på detta något vidare sätt. Ett ytterligare skäl till att använda denna vidare definition är att vissa metoder för att bestämma reliabilitet innefattar också dessa typer av fel som inte direkt kan förknippas med själva mätinstrumentet.

3.2 Metoder för att bestämma reliabiliteten

Reliabilitetskoefficienten utgör ett mått på hur stor inverkan felfaktorerna får på mätresultaten. Det enklaste sättet att bestämma denna koefficient är att mäta en och samma variabel två gånger och se på överensstämmelsen mellan dessa båda mått. Ju bättre de två mätningarna stämmer med varandra desta större är tillförlitligheten eller reliabiliteten. I princip går vi alltså här tillväga på samma sätt som när vi kontrollerar om vår klocka visar rätt tid genom att jämföra med en annan klocka eller när vi räknar ut en summa två gånger för att kontrollera om vi räknat rätt.

När vi bestämmer reliabiliteten nöjer vi oss inte med ett enstaka par av mätningar utan de två mätningarna görs på en, inte alltför liten grupp av individer. Som mått på överensstämmelsen mellan de två mätningsresultaten beräknas en vanlig produktmomentkorrelationskoefficient (r). Ju bättre de två mätningsresultaten stämmer överens desto högre blir ju korrelationen. En hög korrelationskoefficient - i detta sammanhang kallar vi den för reliabilitetskoefficienten - visar att felfaktorerna har en liten inverkan på mätresultaten.

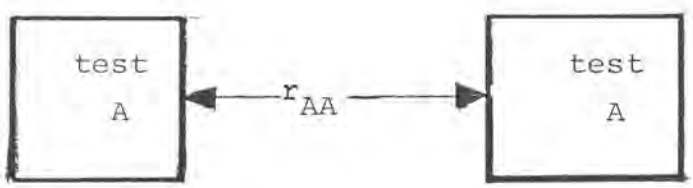
Denna slutsats kan vi dra därför att felfaktorernas inverkan är slumpmässig. D.v.s. den effekt de får på en individs resultat vid det ena tillfället är oberoende av den effekt de får

vid det andra. Reliabilitetskoefficienten ger följaktligen endast ett mått den inverkan som slumpmässiga fel har på mätresultaten. Om mätningen påverkas av systematiska fel ger detta däremot inte utslag i en reliabilitetskoefficient.

Av detta skäl definieras ibland reliabiliteten som mätmetodens förmåga att motstå slumpinflytande.

Test-retestmetoden. Det enklaste sättet att få två mått på en och samma variabel är att bjuda mätinstrumentet vid två olika tillfällen till en och samma grupp.

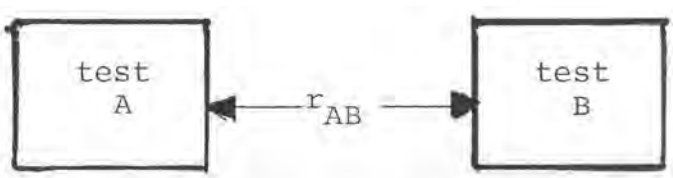
Figur 10.



Den reliabilitetskoefficient som beräknats enligt denna metod blir något svårtolkad eftersom vi inte kan bortse från att individerna minns sina svar från det första mättillfället eller från att det skett en viss inlärning vid det första mättillfället som sedan inverkar på den andra mätningen. Ytterligare ett problem med denna typ av reliabilitet är att vi inte får med den felkälla vi tidigare kallade "variation i mätinstrumentet". Uppgifterna är ju de samma vid båda tillfällena.

Parallelltestmetoden. De brister som vidlåder test-retest-reliabiliteten kan i stor utsträckning undvikas om man konstruerar två test som är så lika varandra som möjligt vad beträffar innehåll och svårighetsgrad, dock utan att innehålla identiska uppgifter. Tillvägagångssättet är i övrigt detsamma som vid test-retest-metoden.

Figur 11.

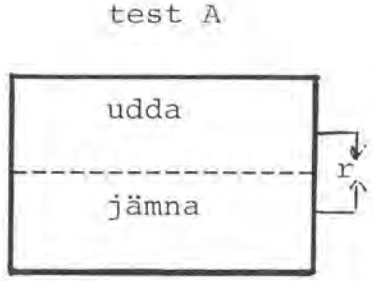


Split-half-metoden. Parallelltestmetoden är en arbetsam metod eftersom två olika mätinstrument måste konstrueras och bjudas vid två tillfällen. För att förenkla arbetet kan man istället ge ett enda mätinstrument till en grupp vid ett enda tillfälle. Efter det att mätningen genomförts delas testet upp i två halvor (split-half). Vardera halvan betraktas sedan som ett parallellt instrument vilket ger sitt eget resultat. Korrelationskoefficienten beräknas mellan resultaten på de två halvorna. Uppdelningen av mätinstrumentet på två halvor sker vanligen så att alla udda uppgifter förs till den ena halvan och alla jämna till den andra. Härigenom ökar ju chansen att de två halvorna kommer att likna varandra vad beträffar innehåll och svårighetsgrad.

Självfallet är även andra principer för uppdelning möjliga. Man skulle t.ex. kunna dela provet på mitten och beräkna korrelationen mellan resultaten på den första och den sista halvan. Med ett sådant förfarande är emellertid riskerna större att de två halvorna inte är parallella utan att de skiljer sig åt både vad gäller innehåll och svårighetsgrad. Korrelationen skulle därigenom sänkas.

Om man ser reliabiliteten som ett mått på överensstämmelsen mellan två parallella prov bör emellertid de två halvorna vara så lika varandra som möjligt både vad gäller innehåll och svårighetsgrad. För att uppnå denna likhet kan uppdelningen på udda och jämna uppgifter vara ett praktiskt tillvägagångssätt.

Figur 12.

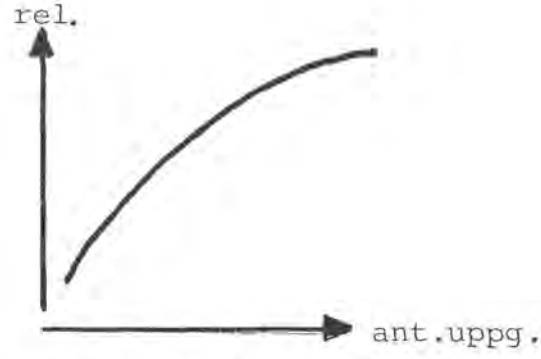


Den korrelation vi får mellan de två halvornas resultat kan inte direkt tas som mått på testets reliabilitet. Till grund för korrelationsberäkningen ligger ju resultaten från två h a l v a mätinstrument. Den erhållna korrelationen kommer härigenom att utgöra en underskattning av den reliabilitetskoefficient som gäller för hela provet sammanhållet.

Underskattningen av hela instrumentets reliabilitet orsakas av att det föreligger ett samband mellan provets längd och dess reliabilitet. En ökning av antalet uppgifter ökar reliabiliteten under förutsättning att de tillkommande uppgifterna är homogena med de redan befintliga och att provet inte blir så omfattande att resultatet påverkas av trötthetseffekter.

Förhållandet mellan antalet uppgifter och mätinstrumentets längd kan beskrivas med följande figur.

Figur 13.



Formelmässigt uttrycks förhållandet provlängd - reliabilitet:

$$r_{kk} = \frac{k \cdot r_{xx}}{1 + (k-1) \cdot r_{xx}}$$

där r_{kk} = reliabilitet efter förlängning

r_{xx} = reliabilitet innan förlängning

k = antal gånger som provet förlängs

Då vi skattar reliabiliteten för hela provet utifrån korrelationen mellan de två halvorna blir ju k=2 och därigenom förenklas formeln:

$$r_{22} = \frac{2 \cdot r_{xx}}{1 + r_{xx}}$$

Det faktum att reliabiliteten ökar med antalet uppgifter är ganska naturligt. Ju flera uppgifter vi har i provet desto större är möjligheterna att provet blir representativt för uppgiftspopulationen. Med ökat antal uppgifter ökar dessutom chansen att felfaktorerna tar ut varandra och vi får ett rättvisande slutresultat.

Relationen mellan mätningens längd och dess tillförlitlighet utnyttjas även i andra sammanhang. Inom idrotten är det t.ex. vanligt att man i viktigare tävlingar spelar längre matcher. Så sker bl.a. i de större bordtennis- och tennisturneringarna där matcherna går i bäst av 5 set mot normalt bäst av 3 set. På samma sätt avgörs SM-finalen i ishockey, där det lag vinner som först besegrat motståndarlaget 3 gånger.

I idrottssammanhang är det också vanligt att man använder sig av mer än en bedömare. Så sker i de flesta tävlingsgrenar där bedömningen är subjektiv t.ex. konståkning på skridskor, gymnastik och simhopp. Även detta är ett sätt att höja reliabiliteten - ett förfarande som bör utnyttjas också i sådana provsammanhang där bedömningen är svår t.ex. vid bedömningen av essäfrågor. Om man utnyttjar mer än en bedömare kan man på samma sätt som vi beskrivit ovan beräkna en korrelation mellan bedömarnas resultat och härigenom få ett mått på medbedömarreliabiliteten.

Som vi antydde i samband med split-half-reliabilitet är det inte lämpligt med alltför omfattande prov eftersom vi då får in trötthetseffekter som kan försämra reliabiliteten. Det kan också vara svårt att engagera medbedömare. För att undvika en låg reliabilitet orsakad av att mätningen innehåller för få uppgifter kan man ge flera var för sig homogena prov och sedan slå samman dessa prov till en totalpoäng enligt de principer vi diskuterade i föregående avsnitt. Ett sådant förfarande ger med säkerhet ett mer tillförligligt slutresultat än om man genomför en enda mätning som skall täcka ett större område.

På samma sätt som att flera prov för en och samman individ ökar tillförlitligheten i det resultat individen får kommer

sammanfattande gruppvärden såsom t.ex. medelvärdet att vara betydligt mer reliabelt än det resultat varje individ i gruppen får.

De metoder vi hittills diskuterat för bestämningen av reliabilitet bygger på den principen att vi beräknar korrelationen mellan två uppsättningar av mätvärden avseende en och samma variabel. Om vi använde oss av mer än två sådana prov skulle vi självfallet få en ännu bättre uppskattning av tillförlitligheten. Ett sådant förfarande är emellertid både praktiskt svårt att genomföra och kräver mer statistiskt kunnande varför vi inte behandlar det här. Det finns emellertid en metod för reliabilitetsbestämning där man utifrån ett enda prov kan göra en skattning av detta provs korrelation med ett parallellt prov nämligen.

Kuder-Richarsons metod. Denna metod ger ett mått på hur homogent ett prov är - eller med andra ord hur starkt uppgifterna i provet korrelerar inbördes.

Den grundläggande principen för Kuder Richardsons metod är att provets varians bestäms av två faktorer.

1) de enskilda uppgifternas egen varians x)

2) uppgifternas inbördes korrelationer

---

x)
På samma sätt som vi beräknar en varians för resultaten på ett helt prov kan vi beräkna varianser för den poäng individerna fått på en enstaka uppgift:
På en viss uppgift har 5 personer fått 0,1,2,3 resp 4 poäng. Medelvärdet för poängen på denna uppgift blir: $\frac{10}{5} = 2.0$ poäng.
Variansen blir då $\frac{(0-2)^2 + (1-2)^2 + (2-2)^2 + (3-2)^2 + (4-2)^2}{4} =$

$\frac{4 + 1 + 0 + 1 + 4}{4} = 2.5$ för just denna uppgift.

Ju högre varians de enskilda uppgifterna har och ju högre de korrelerar inbördes desto högre blir variansen i det totala provresultatet. Det första påståendet om uppgifternas varians är i det närmaste självklart. Om alla individer i en grupp får exakt samma poäng på varje enskild uppgift måste också deras slutresultat bli detsamma:

		Uppg. I	II	III	IV	Totalpoäng
Ind.	A	3	1	0	2	6
	B	3	1	0	2	6
	C	3	1	0	2	6
	D	3	1	0	2	6

Variansen i totalpoäng blir alltså noll.
Det andra påståendet om att korrelationen mellan uppgifterna påverkar variationen i totalresultat kan visas med följande två exempel.

Exempel A

	Uppgift I	II	III	IV	totalt
A	0	0	0	0	0
B	1	1	1	1	4
C	2	2	2	2	8
D	3	3	3	3	12
M	1.5	1.5	1.5	1.5	6

Exempel B

	Uppgift I	II	III	IV	totalt
A	3	1	3	2	9
B	2	3	2	0	7
C	1	0	0	1	2
D	0	2	1	3	6
M	1.5	1.5	1.5	1.5	6

I exempel A är korrelationen mellan de fyra uppgifterna 1.

Varje uppgift ger exakt samma poäng för en och samma individ. Detta medför att vi får en stor variation i slutresultatet.

I exempel B har vi en betydligt lägre korrelation mellan uppgifterna. Alla uppgifter har dock exakt samma varians i båda exemplen. Om vi beräknar variansen på vanligt sätt blir variansen för var och en av uppgifterna:

$$\frac{1.5^2 + 0.5^2 + (-0.5)^2 + (-1.5)^2}{3} = 1.67$$

Variansen i totalpoäng blir i exempel A: $\frac{(-6^2)+(-2^2)-2^2+6^2}{3} = 26.67$

Motsvarande värde i exempel B blir: 8.67

I formeln för Kuder-Richardsons reliabilitet ingår variansen för totalpoängen (S_x^2) och variansen för varje enskild uppgift (S_j^2). Dessutom en korrektionsfaktor för antalet uppgifter (k):

$$r_{xx} = \frac{k}{k-1} \cdot \left(\frac{S_x^2 - \Sigma S_j^2}{S_x^2} \right)$$

Om vi sätter in de värden vi fått i exempel A blir resultatet

$$r_{xx} = \frac{4}{3} \cdot \left(\frac{26.67 - 6.67}{26.67} \right) = \frac{4}{3} \cdot \frac{20}{26.67} = \frac{80}{80.01} = 1.00$$

I exempel B sjunker den totala variansen på grund av uppgifternas lägre inbördes korrelation till 8.67 och reliabiliteten sjunker då också till

$$\frac{4}{3} \cdot \left(\frac{8.67 - 6.67}{8.67} \right) = \frac{4}{3} \cdot \frac{2}{8.67} = \frac{8}{26.01} = 0.31$$

Vid våra exempel ovan poängsattes uppgifterna i en skala från 0 - 3 poäng. I många sammanhang bedöms uppgifterna enbart med rätt/fel eller 0/1 poäng. I ett sådant fall utbytes ΣS_j^2 mot Σpq där p anger den proportion individer som löst uppgiften riktigt och q den proportion som missat uppgiften. q blir följaktligen lika med 1-p.

Ett exempel:

	Uppgift				
Ind.	I	II	III	IV	totalt
A	1	0	0	0	1
B	1	1	1	0	3
C	1	1	0	1	3
D	1	1	1	1	4
p:	1.00	0.75	0.50	0.50	
q:	0.00	0.25	0.50	0.50	
p·q:	0.00	0.19	0.25	0.25	$\Sigma p \cdot q = 0.69$

S_x^2 beräknas på samma sätt som tidigare

$$r_{xx} = \frac{4}{3} \cdot \left(\frac{1.58 - 0.69}{1.58} \right) = \frac{4}{3} \cdot \frac{0.89}{1.58} = \frac{3.56}{4.74} = 0.75$$

Kuder-Richardsons reliabilitet är som vi visat ovan ett mått på uppgifternas inbördes korrelation. Dess reliabilitetskoefficient kallas därför ofta homogenitetskoefficienten.

Just därför att denna reliabilitetskoefficient är ett mått på provets homogenitet är den lämplig att använda som reliabilitetsmått då man eftersträvar ett homogent prov. I många fall t.ex. vid kunskapsprov kan det däremot vara önskvärt med en viss heterogenitet. Detta gäller t.ex. om provet avser att täcka ett område, som innefattar flera olika delmoment och därigenom i sig är heterogent. I ett sådant fall bör denna typ av reliabilitet ge högre värde för varje delområde än vad vi får för hela provet sammantaget. Ett problem är emellertid här att även de olika delområdena i sig kan vara något heterogena och det kan därför vara önskvärt att även för delområdena få en Kuder-Richardsonreliabilitet lägre än 1.00.

Det enklaste sättet att lösa detta problem är nog att välja någon av de andra reliabilitetsmetoderna i de fall man konstruerar prov för att mäta sådana områden som i sig inte är homogena.

3.3 Exempel

Som exempel på metoder för reliabilitetsbestämning skall vi här beräkna reliabiliteten på en statistiktentamen given på Al-kursen i pedagogik. Vi beräknar reliabiliteten enligt split-half-metoden och enligt Kuder-Richarsons metod.

Stud.	Uppgift nr 1	2	3	4	5	6	7	8	9	10	11	Tot.	Poängen beräknad på udda uppg.	jämna uppg.
A	2	2	2	2	2	0	2	2	0	0	1	15	9	6
B	2	2	2	2	0	2	2	2	2	0	1	17	9	8
C	2	2	2	2	0	2	0	0	2	1	2	15	8	7
D	2	2	2	2	2	2	2	2	1	0	2	19	11	8
E	2	2	2	2	2	1	2	1	2	2	2	20	12	8
F	2	0	0	2	0	0	2	1	1	2	1	11	6	5
G	2	2	2	2	2	2	2	1	2	2	2	21	12	9
H	2	2	0	2	0	0	1	1	2	0	2	12	7	5
I	2	0	0	2	0	1	2	2	0	1	2	12	6	6
J	2	2	2	2	0	0	1	2	2	0	0	13	7	6
K	2	2	2	2	0	0	2	2	0	0	1	13	7	6
L	2	2	2	2	2	0	2	2	0	2	2	18	10	8
M	2	0	2	2	1	2	0	1	1	0	2	13	8	5
N	2	2	2	2	1	0	2	1	2	2	1	17	10	7
O	2	2	2	2	0	0	1	1	2	2	0	14	7	7
P	2	0	2	2	0	0	0	0	0	2	1	9	5	4
Q	2	2	2	2	2	2	2	2	2	1	1	20	11	9
R	2	2	2	2	2	0	0	1	0	2	2	15	8	7
S	2	2	0	2	0	1	2	1	2	0	0	12	6	6
Uppgifternas varians (S_j^2)	0.00	0.70	0.70	0.00	0.92	0.84	0.71	0.45	0.84	0.89	0.56	ΣS_j^2 = 6.61		

För det totala provresultatet är variansen 11.7 och standardavvikelsen 3.4. Korrelationen mellan resultatet på udda resp. jämna uppgifter är r = 0.85.

Split-half-reliabiliteten får vi nu för hela provet genom att utgå från r=0.85 och sedan skatta reliabiliteten för det dubbla antalet uppgifter.

$$r_{kk} = \frac{2 \cdot 0.85}{1 + 0.85} = 0.92$$

Kuder-Richardsons reliabilitet blir:

$$r_{xx} = \frac{11}{10} \cdot \left(\frac{11.7 - 6.61}{11.7} \right) = \frac{11}{10} \cdot \frac{5.09}{11.7} = \frac{55.99}{117} = 0.48$$

I vårt fall blir Kuder-Richardson-reliabiliteten låg på grund av att provet är heterogent. Provet omfattade nämligen uppgifter som täcker en hel 4-poängskurs och innehåller såväl rena räkneuppgifter som diskussionsuppgifter av allmän undersökningsmetodisk karaktär. Hade provet enbart omfattat den ena typen av uppgifter skulle Kuder-Richardson-reliabiliteten blivit avsevärt högre. Dock hade detta medfört att provet hade haft en låg validitet.

Vid beräkningen av split-half-reliabiliteten finns båda typerna av uppgifter representerade i båda halvorna och därigenom blir denna typ av reliabilitet inte lika känslig för heterogenitet. Denna reliabilitet återspeglar mera hur väl detta prov korrelerar med ett annat lika heterogent prov. Eftersom just detta prov är avsett att mäta olika delar av en hel kurs och därför bör vara heterogent anser vi split-half-reliabiliteten vara ett mera relevant mått på provets tillförlitlighet.

Då reliabiliteten utgörs av en korrelationskoefficient kan den aldrig bli högre än 1.00. Vi kan således anse vår split-half-reliabilitet på 0.92 vara tillfredsställande.

Hur skall vi då tolka en reliabilitet på 0.92?
För att ge den en konkret innebörd måste vi föra in ett nytt begrepp:

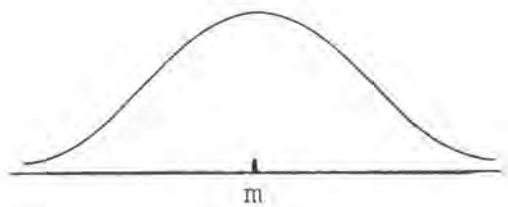
3.4 Enskilda mätningens medelfel

För att förklara detta begrepp tar vi ett exempel.

Antag att jag låter en relativt stor grupp människor skatta min vikt. På de värden jag får in beräknar jag medelvärdet och standardavvikelsen. Jag gör också upp en fördelning över värdena. Eftersom min kropp är någorlunda normal luras inte gruppen att göra systematiska felskattningar. Många skattningar kommer då att ligga på eller alldeles intill min sanna vikt. En del kommer dock att skatta för högt eller för lågt men ju längre bort från min sanna vikt vi kommer desto färre

blir skattningarna: Jag får med andra ord en ungefärlig normalfördelning vars medelvärde kommer att hamna på eller i varje fall ytterst nära min sanna vikt.

figur 14.



"sant värde"

Avvikelserna från medelvärdet beror på att vissa missbedömer vikten - de gör fel helt enkelt. Eftersom den variation som uppstår beror på fel ger vi standardavvikelsen ett litet "e"[x] som index: S_e. S_e är den enskilda mätningens medelfel.

Antag att jag byter ut de människor som skattat min vikt mot lika många vågar. Förmodligen skulle också vågarna ge en viss variation i resultatet men denna skulle bli avsevärt mycket mindre än då vikten skattades. Den enskilda mätningens medelfel skulle följaktligen bli betydligt mindre och detta beroende på att vågarna är mer reliabla än människor som skattar vikten.

Den enskilda mätningens medelfel är således beroende av reliabiliteten. Ju lägre reliabilitet en mätning har desto större inverkan får felen och desto större blir den enskilda mätningens medelfel.

I vårt exempel ovan var det lätt att fastställa enskilda mätningens medel eftersom vi kan upprepa mätningen många gånger utan att det sanna värdet påverkas. Vid beteendevetenskapliga mätningar är detta knappast möjligt. Därför måste vi skatta den enskilda mätningens medelfel och denna skattning kan göras med hjälp av provets totala standardavvikelse och dess reliabilitet enligt följande formel:

$$S_e = S_x \cdot \sqrt{1-r_{xx}}$$

x) e som i error

Om vi sätter in de värden vi fick på vår statistiktentamen blir den enskilda mätningens medelfel:

$$s_e = 3.4 \sqrt{1-0.92} = 3.4 \sqrt{0.08} = 3.4 \cdot 0.28 = 0.95$$

Detta innebär att, om vi gav en individ en mängd statistiktentor vilka alla var parallella med vår, skulle hans resultat variera så mycket att standardavvikelsen blev 0.95. Vi förutsätter här att det inte sker någon inlärning utan att det sanna värdet är konstant. Mätfelen ger upphov till en normalfördelning av mätvärden med det sanna värdet som medelvärde. Med detta antagande som grund kan vi utnyttja normalkurvans egenskaper för att skatta inom vilket intervall en individs värden skulle variera med en viss grad av säkerhet, om mätningen upprepades många gånger. I en normalfördelning faller ju 95% av alla värden inom intervallet $m \pm 1.96 \cdot S$ och 99% av alla värden inom intervallet $m \pm 2.58 \cdot S$.

För en individ vars rättvisande (sanna) värde på statistiktentamen är 15 poäng skulle 95% av alla mätvärden från parallella prov falla inom intervallet

$15 \pm 1.96 \cdot 0.95$ där 15 är individens sanna värde
1.96 en konstant hämtad från normalfördelningen
0.95 enskilda mätningens medelfel

Resultatet blir 15 ± 1.86 poäng. D.v.s. 13.14 - 16.86
99% av alla värden faller inom intervallet $15 \pm 2.58 \cdot 0.95 = 15 \pm 2.45$
12.55 poäng - 17.45 poäng.

Våra beräkningar är orealistiska därigenom att vi utgår ifrån ett sant värde och därifrån skattar mätvärdena. Normalt känner vi endast mätvärdet d.v.s. det värde som är behäftat med fel och utifrån detta får vi skatta var det sanna värdet ligger.

Beräkningsmetoden blir ändå exakt den samma som ovan.

Vi tar individ K-s resultat på statistiktentan som exempel. K. fick 13 poäng men vi har ingen aning om hur detta värde

ligger i förhållande till hans sanna värde. Det är nämligen omöjligt att säga hur mätfelen inverkat på ett visst mätresultat.

Genom samma beräkningsmetod som ovan skattar vi K-s sanna värde:

Med 95% säkerhet: $13 \pm 1.96 \cdot 0.95 \approx 11.14 - 14.86$ poäng
Med 99% säkerhet: $13 \pm 2.58 \cdot 0.95 \approx 10.55 - 15.45$ poäng

Om vi nöjer oss med att fånga in det sanna värdet med 95% säkerhet måste vi ta till ett intervall som är ungefär 3.7 poäng. Ökar vi kraven så att vi fångar in det sanna värdet med 99%-ig säkerhet blir vår osäkerhetszon nästan 5 poäng. Storleken på dessa osäkerhetsintervall blir de samma för varje individ i gruppen.

Vad händer med osäkerhetsintervallen om reliabiliteten ökas? Jo då blir den enskilda mätningens medelfel mindre och följaktligen minskar osäkerhetszonen. En reliabilitet som är 1.00 ger $S_e = 0.00$ (jfr formeln för S_e) och då blir osäkerhetszonen också 0. Vi mäter då enbart sanna värden.

Eftersom reliabiliteten gäller för hela provet och standardavvikelsen är beräknad på hela gruppen kommer den enskilda mätningens medelfel att vara lika stort för var och en av individerna.

Om vi ställer krav på oss att de resultat vi anger för en individ med 95% säkerhet skall innefatta hans sanna värde får vi således inte ange ett punktvärde i form av en poäng. Vi måste då ange ett intervall som för vårt prov innebär individens mätresultat $\pm$ 1.86 poäng t.ex. individ F-s kunskaper motsvarar ett värde mellan 9.14 och 12.86 poäng. Skärper vi kravet till 99%-ig säkerhet får vi för F ange ett värde mellan 8.55 och 13.45 poäng.

Reliabilitetsbristerna medför således att vi inte mäter exakta värden. Vi får med hjälp av mätinstrumentet endast en ungefärlig skattning av individernas verkliga prestationsförmåga. Hur

"ungefärlig" denna skattning är beror på hur hög reliabilitet vårt prov har.

När jag fick in resultaten från statistiktentan hade jag den obehagliga uppgiften att sätta ut gränser för hur många poäng som skulle krävas för godkänd resp. väl godkänd.

Låt oss säga att jag krävde 13 poäng för godkänd. Med dessa gränser skulle individerna H, I och S med 12 poäng, F med 11 poäng och P med 9 poäng underkännas. Däremot skulle t.ex. J, K och M med 13 poäng bli godkända.

Eftersom gränsen för godkänd egentligen blir 12.5 poäng kan vi beräkna ett osäkerhetsområde omkring denna gräns. Osäkerhetsområdet blir 12.5 ± 1.86 d.v.s. från 10.64 poäng upp till 14.36 poäng. De individer, vars poäng ligger inom dessa gränser, löper en risk som är större än 2.5%[1] att bli orättvist bedömda. D.v.s. att de oförtjänt blir godkända eller orättvist underkända. Ju närmare gränsvärdet deras poäng ligger desto större blir risken för en orättvis bedömning. På motsvarande sätt får jag ett osäkerhetsområde vid gränsen för väl godkänd.

Vad kan jag göra åt risken för felbedömningar, när provet väl är genomfört?

Ett sätt att i efterhand minska riskerna för en orättvis bedömning är att göra bedömningen extra noggrann för de individer som ligger i närheten av sådana gränser. Detta kan t.ex. ske genom att engagera en medbedömare eller att själv göra en ny bedömning. Detta leder dock till att man endast reducerar en felkälla nämligen bedömningsfelen. Övriga felkällor står kvar.

Naturligtvis skulle jag kunna genomföra en ny examination med "gränsfallen". Detta skulle kräva ett nytt prov och jag får en ny jämförelsegrupp. Därmed blir resultaten inte jämförda med de som erhölls vid den första mätningen.

Det säkraste sättet att undvika eller i varje fall reducera reliabilitetsbristerna är att inte bygga en slutgiltig bedöm-

1) Vi får här endast 2.5% därför att de godkända riskerar att bli underkända endast då felen verkar åt det ena hållet - när de sänker poängen.

ning på ett enda mätresultat utan att låta flera prov tillsammans ligga till grund för en slutbedömning. Som vi diskuterat tidigare ökar då möjligheterna till att mätfelen tar ut varandra för en individ och att det sammanslagna resultatet blir mer tillförlitligt.

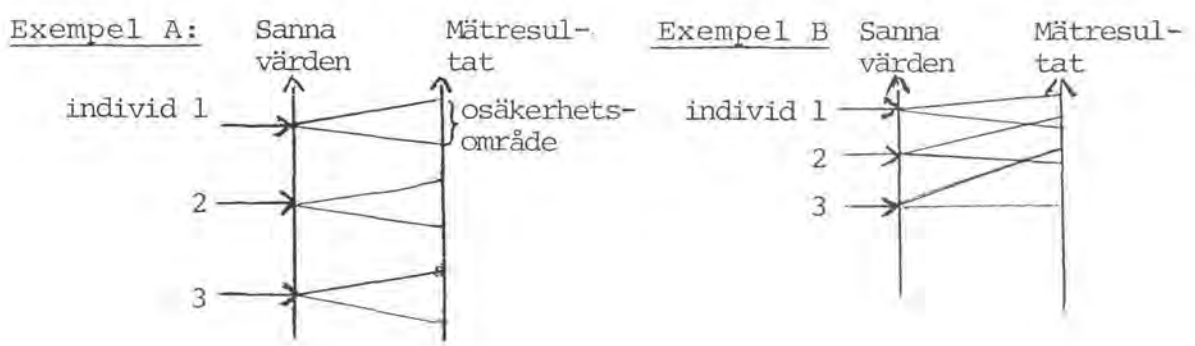
3.5 Reliabilitet och spridning.

Ett sätt att reducera antalet felbedömningar i statistiktentan hade varit att dra gränserna mellan olika betyg så att få individer hamnar nära gränsen. Idealet hade varit om det funnits luckor i fördelningen med ett antal poäng där inga individer låg. Betygsgränserna skulle då läggas mitt i dessa luckor.

I praktiken är detta naturligtvis ett orealistiskt förfarande. Betygsgränserna kan ju inte bestämmas av var eventuella luckor råkar uppstå i poängfördelningen. Resonemanget beskriver ändå ett viktigt förhållande mellan reliabiliteten och variationen i gruppen: Ju mer heterogen gruppen är desto högre reliabilitet får mätningarna.

Detta kan visas med följande enkla figur.

Figur 15.



I exempel A har vi en heterogen grupp, vilket visar sig i att de sanna värdena varierar mycket. Även om felfaktorerna också bidrar till en viss variation i mätresultaten är detta bidrag här litet jämfört med den variation som de sanna värdena står för. Felfaktorerna kan t.ex. knappast orsaka att rangordningen mellan individerna ändras.

I exempel B är däremot gruppen homogen och därför kan mätfelen, vilka här är lika stora som i exempel A, få betydligt större inverkan. Individ nr 2 kan t.ex. lika gärna få det bästa som det sämsta resultatet. Den variation som mätfelen står för är i detta fall stor i förhållande till variationen i sanna värden.

Vi har genom dessa två exempel antytt ett annat sätt att definiera reliabilitet nämligen som kvoten mellan sann varians och felvarians. Reliabilitetens beroende av den sanna variansen är egentligen ganska självklar. Det är t.ex. mycket lättare att åstakomma en rangordning mellan tre elever vars faktiska kunskaper är mycket olika jämfört med om jag skall rangordna tre elever, som kan i stort sett lika mycket. Jfr t.ex. betygsättning i en spärrad utbildning där eleverna är inbördes mycket lika varandra med att sätta betyg i en grundskoleklass.

Av detta resonemang kan vi dra den generella slutsatsen att reliabiliteten inte är någon konstant egenskap hos ett mätinstrument utan instrumentets tillförlitlighet är beroende av heterogeniteten i den grupp det används på.

OBS att det gäller heterogeniteten i gruppen. Hur stor denna heterogenitet är återspeglar sig normalt i standardavvikelsen för mätresultaten. Man får dock inte automatiskt sätta likhetstecken mellan gruppens heterogenitet och den standardavvikelse som mätningen resulterar i. Som vi diskuterade i föregående avsnitt är standardavvikelsen beroende av hur stora skalstegen är vid mätningen. Ju mindre skalsteg desto större standardavvikelse.

Det finns uppenbarligen en risk att man i bedömningen av ett prov vill åstadkomma en så stor variation i resultaten som möjligt. Detta underlättar ju rangordningen av eleverna, som

vi diskuterade ovan. För att uppnå denna grad av variation är det lätt att falla för frestelsen att göra bedömningen mycket fingraderad - d.v.s. att arbeta med små skalsteg. Blir dessa skalsteg tillräckligt små mister de dock sitt informationsvärde. D.v.s. det finns egentligen ingen relevant skillnad mellan två individer med en poängs skillnad i mätresultat.

Å andra sidan gör vi mätningen onödigt otillförlitlig om vi arbetar med alltför grova skalenheter. Detta kan ju medföra att två individer som är olika i faktisk prestationsförmåga hamnar på samma mätresultat.

Eftersom de beteendevetenskapliga mätningarna normalt är relativa - vi studerar skillnader mellan individer - är förmodligen den risken större, att mätenheterna görs för små än tvärtom.

Alltför små skalsteg medför således inte en förbättring av reliabiliteten. Däremot medför de att vi differentierar individerna på andra grunder än vad provet var avsett för. Detta blir då ett validitetsproblem och till det återkommer vi i nästa avsnitt.

Innan vi går över till att diskutera validitet skall vi ge ett exempel på risken med för små skalsteg. Jag såg för en tid sedan en provräkning i årskurs 5, där en elev fått en poängs avdrag för att vederbörande förväxlat två siffror när svaret skrevs ut. I själva uträkningarna stod det korrekta svaret! Samma elev hade på samma provräkning glömt att skriva ut kronor efter sitt svar på en fråga som löd: "Skriv i kronor 4 kr och 5 öre". Eleven skrev: 4.05. Läraren skrev: 0 p.!

Även om dessa principer för bedömning bidrog till att öka spridningen i gruppen - provräkningen föreföll vara lätt - så innebar inte dessa bedömningar att kunskaper i matematik premierades. Däremot utgjorde de ett mått på noggrannhet, vilket knappast provräkningen var avsedd att mäta.

3.6 Slutkommentar

Den reliabilitet vi får vid ett visst mättillfälle kan inte ses som en generell egenskap hos det prov som använts. Reliabilitetskoefficienten påverkas ju inte enbart av egenskaper hos själva provet utan också av faktorer som ligger utanför detta t.ex. individens variation, störningar i mätsituationen och bedömarens tillförlitlighet.

Som vi visat kan också olika metoder för reliabilitetsbestämning ge olika hög reliabilitetskoefficient beroende på att de olika metoderna varierar i känslighet för olika typer av fel.

Slutligen är reliabiliteten också en funktion av homogeniteten i den grupp, vars resultat ligger till grund för reliabilitetsbestämningen.

Av detta följer att reliabiliteten, även om man bestämmer den enligt samma metod, kan variera från mättillfälle till mättillfälle, från grupp till grupp och självfallet också från ett prov till ett annat även om proven är av samma typ. För att då få en uppfattning om vilken tillförlitlighet en uppsättning mätvärden har måste reliabiliteten bestämmas på just dessa data.

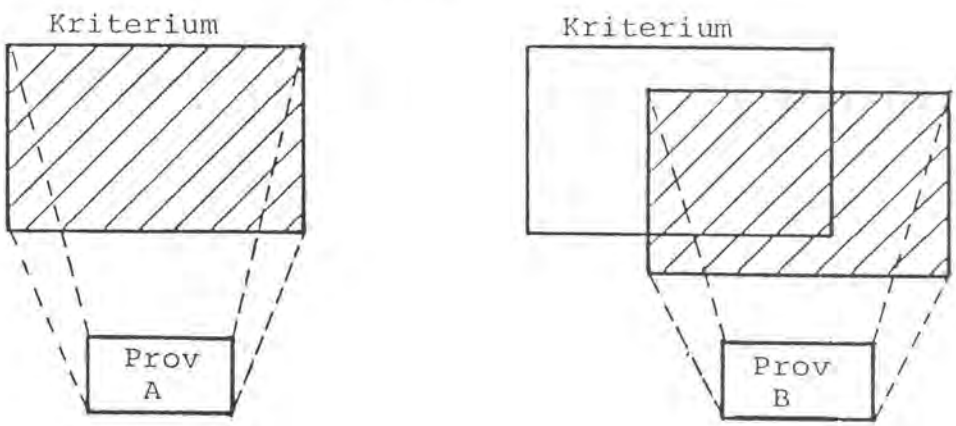
4. VALIDITET

Ordet validitet är en försvenskning av det engelska ordet validity, som betyder giltighet.

När vi konstruerar ett prov vill vi mäta en viss variabel. Eftersom psykiska variabler inte finns i materiell mening måste vi först definiera hur dessa variabler visar sig i beteenden. Vi definierar härigenom kriteriet för variabeln. Genom att sedan jämföra provets innehåll eller provresultaten med kriteriet får vi en uppfattning om provets validitet d.v.s. om provet har giltighet för den aktuella variabeln.

Figur 16.

VALIDITET



Av figuren ovan framgår att prov A har en mycket god, för att inte säga perfekt validitet. Detta prov återspeglar ju hela kriteriet och inget utöver kriteriet.

Prov B har däremot uppenbara validitetsbrister. Dels är det vissa delar av kriteriet, som inte alls täcks av provet, dels mäter provet också sådant, som inte hör till kriteriet.

För att konkretisera dessa två typer av validitetsbrister tänker vi oss att vi konstruerar ett kunskapsprov. Jag har bråttom och det visar sig vara väldigt svårt att hitta på bra frågor inom ett par delområden av kriteriet. För att få provet klart i tid måste jag hoppa över dessa områdena i provet. Na-

turligtvis leder detta till validitetsbrister. Ett annat vanligt exempel på denna typ av brist är att jag måste begränsa provets omfattning så mycket att vissa delar av kriteriet inte "får plats".

Provet är av essätyp och kräver ganska omfattande svar från eleverna. Naturligtvis kommer jag då vid rättningen att ställas inför många svårtolkade svar. I denna tolkningssituation är risken mycket stor att jag blir påverkad av elevernas förmåga att uttrycka sig. Då kan ett, ur innehållssynpunkt "luddigt" svar, ges en mera välvillig tolkning om det är språkligt elegant jämfört med om det är uttryckt med ett torftigt språk.

I detta fall kommer provresultatet inte att vara ett renodlat mått på elevernas kunskaper utan även ett mått på deras uttrycksförmåga, vilken inte ingår i kriteriet.

Med detta senaste exempel har vi antytt ett viktigt förhållande när det gäller validitet nämligen att validiteten inte helt avgörs av provets utformning i sig utan också av de bedömningsprinciper som tillämpas. Jag kan således ha ett prov som ur innehållssynpunkt är perfekt men vid bedömningen premieras inte enbart sådant som provet är avsett att mäta. Därmed kommer mätresultaten att vara behäftade med validitetsbrister.

En förutsättning för att man skall kunna bedöma ett provs validitet är naturligtvis att man kan exakt definiera kriteriet. Detta är relativt sett enkelt då det gäller kunskapsprov. I dessa sammanhang leder en noggrant genomförd målanalys till att också kriteriet definieras.

Då det gäller konstruktion av t.ex. intelligenstest och personlighetstest blir kriteriedefinitionen avsevärt mycket svårare. Hur skall vi kunna definiera kriteriet för en så mångfacetterad variabel som intelligens?

I praktiken är det omöjligt att definiera ett enda kriterium för denna variabel. Konsekvensen har blivit att olika testkonstruktörer haft olika kriterier - bl.a. beroende på deras

olika teoretiska utgångspunkter - och vi har därigenom fått en mängd olika typer av intelligenstest. En och samma individ kan härigenom få något olika resultat beroende på vilket intelligenstest som använts.

För att råda bot mot den förvirring, som detta kan medföra anger man ofta ett visst testresultat tillsammans med namnet på det test som använts. Att på detta sätt låta själva mätinstrumentet - själva operationen - definiera en variabel kallas för operationell definition.

Många av de variabler man arbetar med inom beteendevetenskapen är så komplexa att två test, avsedda att mäta en och samma variabel ändå kan ge upphov till något olika resultat beroende på att de fångar något olika sidor av denna variabel. Det kan därför vara klokt att gå tillbaka och se på innehållet i det test som använts, när man skall tolka ett visst testresultat. Därigenom får ju testresultatet en mer konkret innebörd.

Man kan skilja mellan olika typer av validitet:

4.1 Innehållsvaliditet.

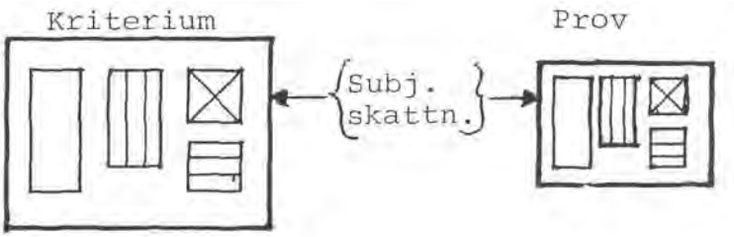
Denna typ av validitet är av central betydelse då det gäller kunskapsprov.

Låt oss säga att jag skall konstruera ett prov på ett visst huvudmoment i ett ämne i årskurs 6. Detta huvudmoment innefattar fyra olika delmoment och jag har innan undervisningen startade bestämt vilken vikt dessa delmoment skall ha i förhållande till varandra och vilken typ av kunskap som är viktig inom vart och ett av delmomenten.

När provet skall konstrueras måste jag se till
- att alla delmoment täcks av i provet.
- att delmomenten ges samma vikt i förhållande till varandra i provet som de har i kriteriet.
- att frågorna i provet mäter den typ av kunskap som är viktig inom vart och ett av delmomenten.

Figur 17.

INNEHÅLLSVALIDITET



De här angivna kraven måste vara uppfyllda för att mätresultaten skall ha en god validitet. Huruvida detta är fallet kan knappast avgöras på annat sätt än genom att låta någon som känner kriteriet väl göra en subjektiv jämförelse mellan kriterium och prov.

De hittills formulerade kraven garanterar emellertid inte en god innehållsvaliditet.

Som vi diskuterat ovan kommer också bedömningen att inverka på validiteten. Förutom risken att påverkas av ovidkommande faktorer såsom individens uttrycksförmåga löper vi också den risken att bedömningen sker med avseende på andra och ofta mer elementära former av kunskap än de som vi från början avsåg att mäta.

I stor utsträckning torde denna risk hänga samman med att man som bedömare känner ett starkt krav på sig att vara objektiv För att uppnå denna objektivitet konstrueras vanligen bedömningsmallar. Dessa bedömningsmallar innebär ofta en uppräkning av faktaenheter, vilka man bedömer vara relevanta i förhållande till frågan. Bedömningsarbetet består sedan i att räkna hur många faktaenheter svaret innefattar. Ju flera fakta desto högre poäng på frågan.

En bedömning enligt detta sätt blir naturligtvis relativt objektiv och därmed rättvis i en bemärkelse, nämligen den att alla svar bedöms på samma sätt. Det är däremot inte säkert

att bedömningen blir rättvis i den bemärkelsen att man premierar den typ av kunskap som från början eftersträvades. Speciellt stor blir risken om man med frågan avsåg att mäta elevens förståelse för ett sammanhang eller hans förmåga att tillämpa sina kunskaper. Dessa högre kunskapsnivåer låter sig knappast uttryckas i enkla, kvantitativa termer utan här krävs en kvalitativ och därmed lätt en mera subjektiv bedömningsprincip.

Man möter ofta schablonföreställningen att essäprov är bra prov därför att de mäter en mer förståelseinriktad kunskap medan objektiva prov med alternativfrågor, kompletteringsfrågor eller flervalsfrågor mäter mer elementära kunskaper och därför är sämre.

Mot denna föreställning kan man invända att det faktiskt går att konstruera objektiva frågor som mäter högre kunskapsnivåer men detta kräver att konstruktören både är medveten om vad frågan bör mäta och han/hon är en skicklig frågekonstruktör. För att mäta dessa högre kunskapsnivåer är det visserligen enklare att konstruera frågor av essätyp. Men för att utnyttja essäfrågans möjligheter ställs, som vi diskuterat ovan, speciella krav på bedömningsprinciperna.

Vi har hittills endast diskuterat mätning av högre kunskapsnivåer. Självfallet kan det också finnas anledning att pröva mera elementära former av kunskap. I sådana fall kan de objektiva provtyperna i den utformning de vanligen har vara utmärkta hjälpmedel.

Den viktigaste frågan vid kunskapsmätning är inte valet av provtyp i sig utan det gäller att vara medveten om vad man vill mäta, att konstruera ett prov som mäter just detta och att vid bedömningen premiera just de kunskaper, som man avsåg att mäta med provet.

4.2 Samtidig validitet

Samtidig syftar här på att kriterieresultaten är kända då jag konstruerar mitt mätinstrument.

Antag att jag skall göra en undersökning, där jag behöver få ett mått på "förmågan att ta egna initiativ i undervisningen". Av arbets- och kostnadsmässiga skäl måste jag konstruera ett eget mätinstrument.

Innan jag använder mitt mätinstrument i undersökningen måste jag kontrollera att det faktiskt mäter "förmågan till eget initiativ i undervisningen", d.v.s. att det har en god validitet. En sådan kontroll kan jag få genom att jag prövar mitt instrument på en elevgrupp som är lik den jag skall göra undelsökningen på men som inte skall vara med i undersökningen. Förutom resultatet på mätinstrumentet måste jag ha ett kriteriemått. Ett sådant mått skulle jag kunna få genom att jag låter läraren, som känner eleverna väl, göra en skattning av eleverna i den aktuella variabeln.

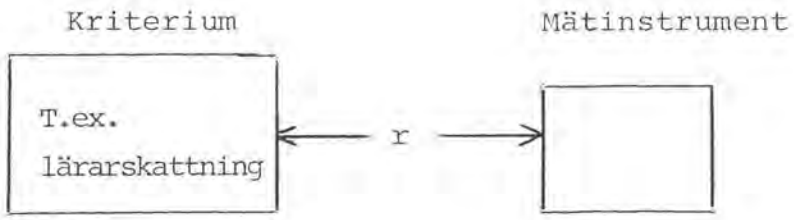
Som ett mått på mätinstrumentets validitet beräknas korrelationen mellan de resultat instrumentet ger och lärarens skattning. Ju högre korrelationen är desto bättre validitet anses då mätinstrumentet ha.

Bakom detta förfarande ligger naturligtvis ett antagande om att kriterieresultaten faktiskt återspeglar den variabel vi vill mäta. I vårt exempel skulle detta antagandes hållbarhet kunna prövas t.ex. genom att välja sådana klasser för utprövningen, vilka har två lärare. Dessa lärare får då göra var sin skattning oberoende av varandra. Ju bättre dessa parvisa skattningar stämmer överens desto säkrare kan jag vara att de uppfattat variabeln på samma sätt och desto mera kan jag lita på kriterieresultaten.

Den samtidiga validiteten fastställs således på empirisk väg och denna typ av validitet kan uttryckas som en korrelationskoefficient mellan kriterieresultaten (lärarnas skattning) och resultaten på mätinstrumentet.

Figur 18.

SAMTIDIG VALIDITET



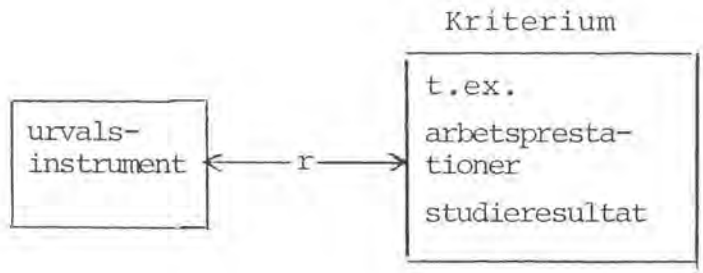
4.3 Prognostisk validitet

Ett annat exempel på empiriskt bestämd validitet är den prognostiska validiteten. Denna typ av validitet är av central betydelse då vi vill använda ett mätresultat som grund för en förutsägelse om vad en individ kan prestera senare. Exempel på sådana situationer är då vi gör urval till ett yrke/utbildning eller då vi bedriver studie- och yrkesvägledning.

När vi konstruerar ett mätinstrument för dessa syften känner vi ju inte kriterieresultaten. Dessa blir kända först efter det att individerna har varit verksamma i yrket under en tid eller har genomgått utbildningen.

Figur 19.

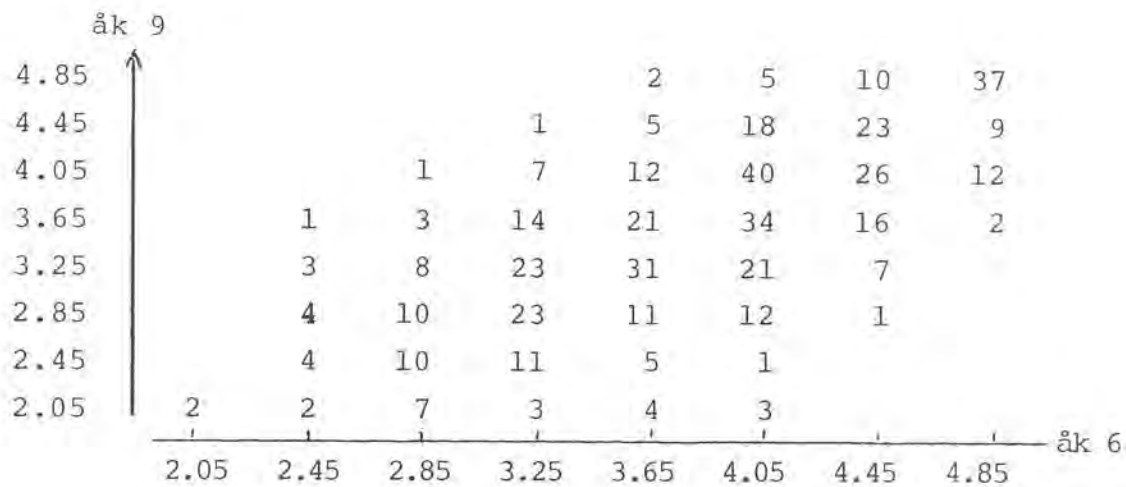
PROGNOSTISK VALIDITET



Som framgår av figuren ovan uttrycks också denna typ av validitet i form av en korrelationskoefficient. Ju högre korrelationen är mellan resultatet på urvalsinstrumentet och kriteriet desto säkrare kan detta urvalsinstrument förutsäga senare prestationer.

För att få ett konkret underlag för vår fortsatta diskussion skall vi här redovisa den prognostiska validiteten för betygen i årskurs 6 då det gäller att förutsäga betygen i årskurs 9 för en viss elevgrupp. Med betyg menar vi här medelbetyget i läroämnen i de två årskurserna. Det exempel som redovisas här är hämtat ur ett riksrepresentativt urval av normalåriga elever, som gick i grundskolans årskurs 6 år 1966. Resultaten nedan avser dock endast pojkar tillhörande en socialgrupp och som valt den mest teoretiska utbildningsvägen under högstadiet. Anledningen till att just denna grupp valts ut är att den uppvisar det högsta sambandet mellan betygen i årskurs 6 och årskurs 9.

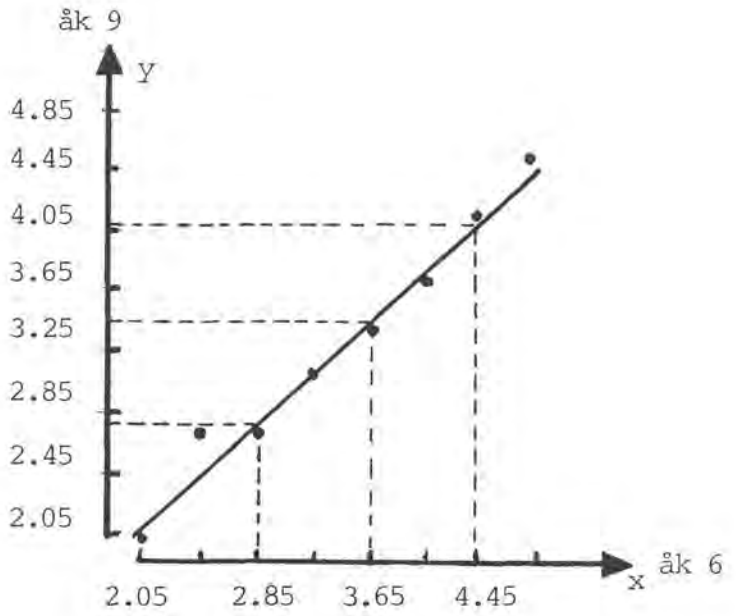
Figur 20.



$r_{xy} = 0.70$

Eftersom vi har ett klart positivt samband mellan de båda variablerna gäller den generella trenden att ju högre betyg man har i årskurs 6 desto högre betyg får man i allmänhet också i årskurs 9. Denna generella trend kan beskrivas med hjälp av en regressionslinje:

Figur 21.



Regressionslinjen utgör en sammanfattande beskrivning av hur de två variablerna förhåller sig till varandra. Den ligger så i diagrammet att individernas sammanlagda avvikelser, mätt på y-variabeln, blir de minsta möjliga. Regressionslinjen anger härigenom det värde i y-variabeln som utgör den bästa förutsägelsen utifrån ett visst värde i x-variabeln.

För en individ som fått betyget 2.85 i årskurs 6 är den bästa förutsägelsen av hans betyg i årskurs 9 c:a 2.85.(Se den streckade linjen). På motsvarande sätt är den bästa förutsägelsen av betyget i årskurs 9 c:a 4.05 för de elever, som har 4.45 i årskurs 6.

Om alla individer i gruppen låg nära regressionslinjen skulle vi med hjälp av denna linje ganska väl kunna förutsäga en enskild individs betyg i årskurs 9 utifrån kännedom om betyget i årskurs 6. En sådan ansamling i närheten av regressionslinjen förutsätter emellertid en mycket hög korrelation mellan prognosinstrument och kriterium. I vårt exempel var korrelationen 0.70, vilket måste betraktas som en jämförelsevis hög prognostisk validitet. Trots denna relativt höga korrelation kan två individer med samma betyg i årskurs 6 få mycket olika betyg i årskurs 9. Som framgår av fig. 20 varierar betygen i 9-an mellan 4.05 och 2.05 för de individer som i årskurs 6 hade betyget 2.85. På motsvarande sätt varierar de elever, vars betyg i årskurs 6 var 4.45, mellan 2.85 och 4.85 i årskurs 9-betyg.

Även om vår kännedom om betyget i årskurs 6 hjälper oss att förutsäga betygsnivån i årskurs 9 blir ändå våra förutsägelser på individnivå behäftade med ett ganska stort mått av osäkerhet.

Som ytterligare ett exempel på denna osäkerhet kan vi peka på att det finns individer i vår grupp med betyget 2.45 i årskurs 6 men som får ett högre betyg i årskurs 9 än vissa elever med så högt betyg i årskurs 6 som 4.45.

Vi kan sammanfatta vårt resonemang så här långt med att säga att när den prognostiska validiteten sjunker från 1.00 minskar möjligheterna till individuella förutsägelser mycket snabbt.

Om vi i stället inriktar oss på att göra förutsägelser på gruppnivå blir prognoserna betydligt säkrare.

De punkter, som finns i fig. 21 runt regressionslinjen anger medelbetyget i årskurs 9 för de elever som ligger på var och en av betygsnivåerna i årskurs 6. T.ex. de 14 elever, som har betyget 2.45 i årskurs 6 får ett medelbetyg i årskurs 9 på 2.76 och de 39 elever som hade 2.85 i årskurs 6 får medelbetyget 2.78 i årskurs 9.

Som framgår av figur 21 ligger dessa medelvärden ganska väl

samlade kring regressionslinjen och jag kan följaktligen göra en betydligt säkrare prognos av dessa gruppers medelvärden än vad som är möjligt att göra för enstaka individer.

Vi kan här dra parallellen med försäkringsbolagens sätt att arbeta. Med hjälp av statistik kan dessa bolag tämligen väl säga hur stor dödligheten är i en viss åldersgrupp. De har däremot inga möjligheter att förutsäga vilka individer i den aktuella åldersgruppen, som kommer att dö.

När det gäller urval av grupper utnyttjar vi ofta denna möjlighet att ställa prognos på gruppnivå utifrån resultatet på ett urvalsinstrument. Så sker t.ex. vid urvalet av sökande till vissa högre utbildningar, där urvalet till stor del sker utifrån betyg. Ett liknande förfarande används ibland också vid anställning av personal, även om man då ofta använder andra urvalsinstrument än betygen.

Oavsett i vilket sammanhang urvalet sker måste man ha tillgång till ett urvalsinstrument som uppvisar ett klart samband med kriteriet - d.v.s. ett urvalsinstrument med prognostisk validitet.

Låt oss gå tillbaka till vårt exempel med betygen i årskurs 6 och årskurs 9.

För att förenkla diskussionen betraktar vi de elever som får ett betyg på 3.25 och högre i årskurs 9 som framgångsrika elever medan de elever som får ett lägre betyg är elever utan framgång.

Med denna kategorisering kommer 392 av alla de 505 eleverna att tillhöra gruppen "framgångsrika". Dessa elever utgör 78% av samtliga.

Vi antar nu att samtliga 505 elever inte kan få plats på högstadiet utan att endast 450 elever kan tas emot. Då måste vi följaktligen göra ett urval. Eftersom vi från tidigare undersökningar vet att det finns ett samband mellan betyg i års-

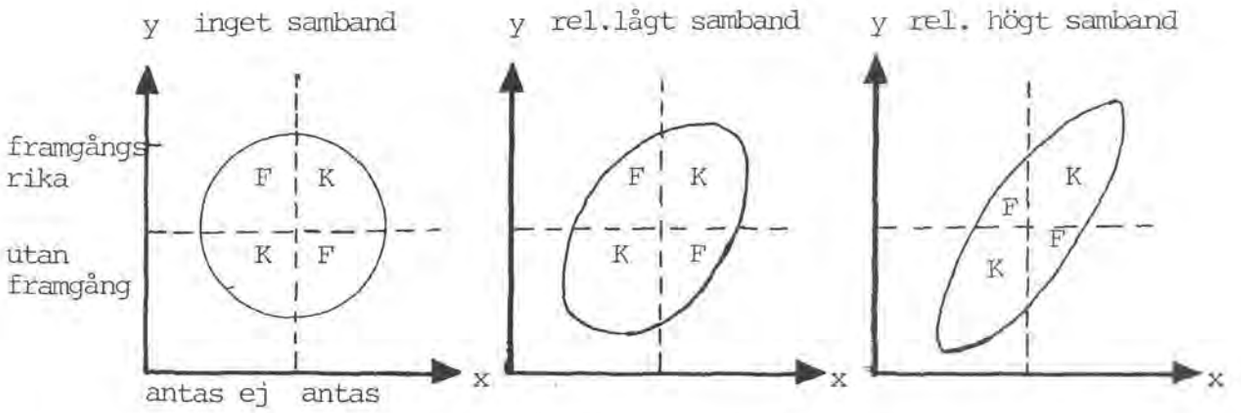
kurs 6 och 9 kan vi utnyttja betygen i årskurs 6 som urvalsinstrument. Med 450 tillgängliga platser kan alla elever i gruppen med betyget 3.25 och högre i årskurs 6 tas med.

Utav dessa 450 elever kommer 376 att tillhöra kategorien "framgångsrika" vilket motsvarar 84% av alla antagna elever. Om de tillgängliga platserna hade tilldelats genom lottning skulle andelen framgångsrika elever fortfarande varit c:a 78%. Urvalsinstrumentet har således hjälpt oss att göra ett bättre urval än vad vi kunnat göra utan detta. För denna vinst får vi dock betala det priset att vi utestänger 16 elever, vilka också skulle tillhört kategorien "framgångsrika" om de hade fått fortsätta utbildningen. (OBS. att vår resonemang förutsätter att betygskraven på högstadiet inte förändras p.gr.a. det selektiva urvalet).

Låt oss anta att antalet platser är ännu mer begränsade och uppgår till endast c:a 275. Då kommer antagningsgränsen att flyttas upp till lägst 4.05 i betyg i årskurs 6. Med bibehållen gräns för vad som betraktas som framgångsrika elever skulle 260 elever av de 277 som då antas höra till denna grupp. Detta motsvarar 94% av de antagna eleverna. Vi har således kunnat göra ett ännu bättre urval genom att ytterligare minska urvalskvoten. Precis som i förra exemplet sker denna vinst på bekostnad av att vi utestänger elever som skulle blivit framgångsrika om de antagits. I detta senare exempel kommer utestängningen att drabba inte mindre än 132 presumtivt framgångsrika elever.

Genom dessa två exempel har vi visat att urvalskvoten är en av de faktorer som avgör hur bra urval vi kan göra, när vi har tillgång till ett urvalsinstrument med prognostisk validitet. Den viktigaste frågan, när det gäller urval, är dock urvalsinstrumentets prognostiska validitet. Ju högre denna validitet är desto bättre urval kan vi göra. Detta framgår av nedanstående figurer.

Figur 22.



x = urvalinstrument
y = kriterium

F = felaktig prognos
K = korrekt prognos

När inget samband föreligger mellan urvalsinstrument och kriterium gör vi fel lika ofta som vi gör en korrekt prognos. När sambandet ökar ökar också andelen korrekta prognoser.

Den prognostiska validiteten är inte någon generell egenskap hos mätinstrumentet utan varierar med det kriterium, som urvalsinstrumentet skall förutsäga. Betyg i matematik har t.ex. säkerligen en högre prognostisk validitet för urval till en teknisk utbildning än för urval till humanistisk.

Den prognostiska validiteten påverkas också av hur stor variationen är i den variabel, som ligger till grund för urvalet.

Låt oss gå tillbaka till vårt exempel, där antagningsgränsen var betyget 4.05 i årskurs 6. Antag att denna gräns har gällt under några år. Då kommer knappast någon elev med betyg, som ligger betydligt under antagningsgränsen att söka. Låt oss för enkelhets skull säga att endast elever med betyget 3.65 och högre söker. Hur hög blir då den prognostiska validiteten hos betygen i årskurs 6 vid urvalet bland de sökande.

Beräknar vi korrelationen mellan betygen i årskurs 6 och betyg i årskurs 9 för endast de elever som ligger på betyget 3.65

eller högre i årskurs 6 blir denna korrelation 0.59.

Betygen i årskurs 6 har således en lägre prognostisk validitet bland de sökande än de har för hela gruppen. En del av betygens prognostiska förmåga kan sägas vara förbrukad genom dess betydelse för individernas beslut om att söka eller inte.

Om vi hade satt gränsen för antagning ännu högre på betygsskalan i årskurs 6 hade gruppen sökande förmodligen också koncentrerats ännu mer till höga betygsnivåer och därmed hade betygens prognostiska validitet sjunkit ännu mera.

Får man på detta sätt en grupp sökande som är mycket homogen i en urvalsvariabel kommer således denna variabels prognostiska validitet att sjunka och det är då fullt möjligt att andra variabler kan tjäna som ett bättre underlag för det slutgiltiga urvalet bland de sökande.

4.4. Begreppsvaliditet

Begreppsvalideringen avser att ge svar på frågan huruvida ett mätinstrument verkligen mäter den variabel, som det är avsett att mäta.

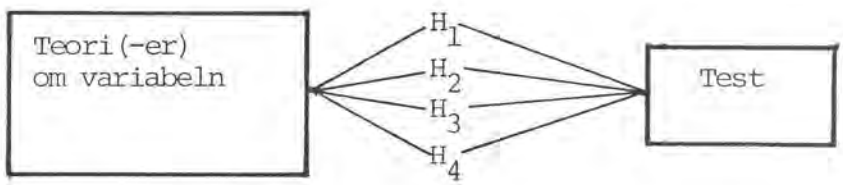
Psykologiska variabler är ofta svåra att definiera klart och entydigt. Därigenom uppstår frågan om vilket eller vilka kriterier man skall välja då man vill pröva ett mätinstruments validitet.

Om jag t.ex. vill konstruera ett mått på skolanpassning får jag utgå ifrån den eller de teorier, som finns om denna variabel.

Med ledning av dessa teorier formulerar jag ett antal hypoteser om hur skolanpassning visar sig i form av beteenden.

Figur 23.

BEGREPPSVALIDITET



Som exempel på sådana hypoteser kan nämnas:

"Om mitt mätinstrument mäter skolanpassning skall sådana elever som sällan bråkar i skolan få en högre poäng än de elever som ofta bråkar".
eller
"Om mitt mätinstrument mäter skolanpassning skall de elever somlärarna anser vara skolanpassade få en högre poäng på mitt mätinstrument än de elever som lärarna anser vara dåligt anpassade till skolan".
eller
"Om mitt mätinstrument mäter skolanpassning skall resultatet från detta mätinstrument uppvisa ett klart samband med andra mätinstrument som också avser att mäta skolanpassning".
o.s.v.

Ju fler hypoteser som jag empiriskt kan bekräfta med mitt instrument desto säkrare kan jag vara att instrumentet mäter den variabel det konstruerats för att mäta.

4.5 Förhållandet mellan reliabilitet och validitet.

En förutsättning för att få en hög validitet är att mätinstrumentet också har en hög reliabilitet men en hög reliabilitet garanterar inte en hög validitet.

Den första delen i påståendet är tämligen självklar. Eftersom reliabilitetsbristerna är orsakade av slumpmässiga fel leder dessa fel också till att sambandet mellan mätresultaten och kriteriet blir lågt. D.v.s. vi får en låg validitet.

Den andra delen i påståendet är kanske lättare att förbise i en praktisk situation. Detta hänger samman med att mätresultaten ofta ligger till grund för viktiga beslut rörande individen t.ex. betygsättning eller urval av individer till en anställning eller till en utbildning. I en sådan situation upplever helt naturligt den som skall fatta beslut en stark press på sig att vara objektiv eller rättvis. Resultatet av detta blir lätt att man väljer att grunda sitt beslut på den typ av mätning som har den högsta precisionen eller reliabiliteten. Det är emellertid långt ifrån alltid som just dessa mätningar är de mest relevanta måtten att basera beslutet på.

I praktiken visar detta sig ofta på det sättet att man t.ex. vid tillsättningen av en högre befattning grundar beslutet mera på en objektiv variabel t.ex. antal tjänsteår än på den mera svårbedömda variabeln lämplighet, trots att den sistnämnda torde vara mest relevant. På samma sätt väljer man ibland att använda sådana frågor i ett kunskapsprov, vilka låter sig rättas på ett kvantitativt och objektivt sätt i stället för att välja sådana frågor som kräver en mer subjektiv bedömning även om det senare sättet är det mest relevanta.

De två exempel vi tagit här innebär att vi tillgodoser kravet på reliabilitet på bekostnad av kravet på validitet. Om denna tendens blir stark blir mätresultaten meningslösa.

För att inte låta sig luras av mätningarnas synbart exakta resultat måste man först ställa sig frågan: Vad är mätningen mått på? Därefter: Hur tillförlitlig är mätningen? Det är svaret på dessa frågor som avgör vilken betydelse man kan tillmäta mätresultaten.