ULF GRANDIN, INSTITUTIONEN FÖR MILJÖANALYS, SLU

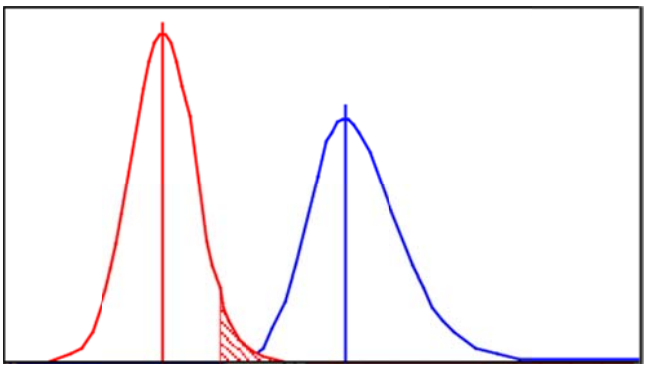
PÅ UPPDRAG AV

SWEDISH ENVIRONMENTAL PROTECTION AGENCY

03-01-13
REV. 12-01-30

Dataanalys och hypotesprövning för statistikanvändare

Ulf Grandin



Förord

För biologer, kemister och andra inom miljöövervakningen är statistik ett hjälpmedel för att undersöka om det t.ex. finns ett mönster eller några skillnader i en mängd siffror. Statistik är också en matematisk vetenskap, grundad på mer eller mindre avancerade matematiska resonemang. Som statistikanvändare behöver man inte kunna all teoretisk bakgrund, men man måste känna till olika statistiska testers förutsättningar och begränsningar. I och med datorernas intåg har detta blivit än mer aktuellt eftersom alla statistikprogram snällt räknar på de siffror man matar in, oavsett om data uppfyller förutsättningarna för ett visst test eller ej.

I denna bok har jag riktat mig till statistikanvändare som vill lära sig mer om grundläggande statistik och om vilka förutsättningar som gäller för olika statistiska tester. Några avsnitt är mer teoretiska än andra, men hela tiden har jag försökt att beskriva formler och föra resonemang på ett handfast och konkret sätt så att personer om inte är bevandrade i det matematiska språket ska kunna tillgodogöra sig innehållet.

Man skulle kunna likna användningen av statistik med ett husbygge. En lekman kan bygga ett hus som vid en första anblick ser gediget ut, men som vid en närmare granskning visar sig ha mängder av fel och brister som uppstått eftersom lekmannen inte haft tillräcklig kunskap om hur man bygger hus. Precis på samma sätt är det med statistiken. Om man inte känner till när man kan och inte kan tillämpa olika statistiska tester kommer man förr eller senare att få fuktskador i duschen och sättningar i husgrunden.

Det är således av stor vikt att kunna mer än hur man får sitt datorprogram att utföra ett visst test. Bara för att man kan spika vet man inte hur man bygger att hus. Min förhoppning är läsare även utan djupa kunskaper i matematik ska kunna använda den här boken för att bli säkrare på vilka förutsättningar som gäller för olika statistiska tester, och hur testerna fungerar.

Uppsala i november 2002

Innehåll

Sammanställning av olika statistiska tekniker och när de passar att användas.

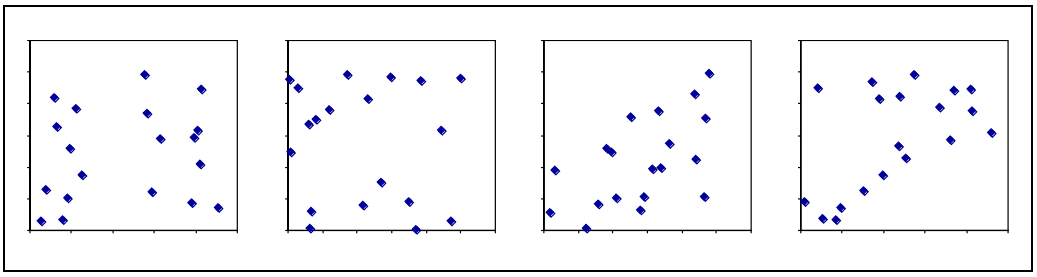
		Typ av data		
	Önskad analys	Mätdata (Normal-fördelade)	Rang, Poäng, Mätdata som inte är normalfördelade	Binomial (Endast två möjliga utfall)
Skillnader	Beskrivning av en grupp	Medelvärde, standardavvikelse	Median, kvanatiler, omfång	Proportion
	Jämföra en grupp med ett hypotetiskt värde	t-test för ett stickprov	Wilcoxons test	Chi-två eller binomialtest*
	Jämföra två oberoende grupper	Oparat t-test	Mann-Whitneys test	Fishers test*
	Jämföra två beroende grupper	Parat t-test	Wilcoxons test	McNemars test*
	Jämföra tre eller fler oberoende grupper	Envägs ANOVA	Kruskal-Wallis test	Chi-två-test
	Jämföra tre eller fler beroende grupper	ANOVA för upprepade mätningar	Friedmans test	Cochranes Q test*
Samband	Fastställa samband mellan två variabler	Pearsons korrelation	Spearmans korrelation	Kontingenskoefficient*
	Predicera ett värde från en annan uppmätt variabel	Enkel eller icke-linjär regression	Icke-parametrisk regression	Enkel logistisk regression
	Predicera ett värde från flera uppmätta variabler	Multipel linlär eller icke-linjär regression	-	Multipel logistisk regression*

* finns inte beskrivet i detta kompendium, se istället t.ex. Sokal & Rohlf (1995).

Introduktion

När behövs statistiska beräkningar?

Som framgått av andra avsnitt i Naturvårdsverkets handbok för miljöövervakning (t.ex. Planering och utformning av miljöövervakningsprogram) är det oftast omöjligt att göra totalundersökningar där man tar med alla objekt i en population. För att ändå kunna dra några slutsatser om den studerade populationen tar man stickprov och gör statistiska beräkningar. Målet med ett statistiskt test är att dra den starkast möjliga slutsatsen från en begränsad mängd data. Det finns emellertid flera anledningar till att en slutsats kan bli felaktig.
En anledning är att viktiga skillnader kan vara gömda i den stora variation som biologiska variabler ofta uppvisar. Detta gör att det kan vara svårt att skilja en verklig skillnad från slumpmässig variation.
En annan anledning är att människans hjärna alltid försöker hitta mönster. Detta leder till att vi hela tiden, både medvetet och omedvetet, söker efter och hittar mönster i alla typer av data. Det gäller såväl syntintryck som siffror. Ett exempel på vår medfödda förmåga att hitta mönster är hur ett spädbarn mycket tidigt lär sig känna igen det mönster av ögon, näsa och mun som är mammans ansikte, och kan skilja detta från andra ansiktsmönster. Ett annat exempel är alla stjärnbilder som är mönster konstruerade från en slumpmässig fördelning av stjärnor på himlen. Det är sålunda en naturlig och instinktiv handling att man konstruerar ett mönster eller ser en skillnad, oavsett om den existerar eller ej, och att man bortser från slumpvariation som stör det mönster hjärnan konstruerat. En illustration är figur 1 där 20 punkter har slumpats fyra gånger. I varje diagram kan man skönja ett samband eller mönster. I a och b tycker man sig se två tydliga två nivåer i x- respektive y-led och i c och d kan man ana linjära samband.



Figur 1a-d. Diagram med vardera tjugo slumpmässigt valda punkter.

För att undvika att hitta mönster som inte finns, eller att missa en verklig effekt som delvis döljs i naturlig variation är statistiska test nödvändiga. Det finns talesätt som säger att det man inte ser direkt i sina data är inget att fästa sig vid. Detta gäller mycket sällan och i synnerhet inte när man jobbar med miljödata eller biologiska data som båda har en stor naturlig variation.

Egenskaper hos mätdata

De data man samlar in har olika stora mängder information. Generellt gäller att man vill ha så mycket information som möjligt från varje mätning, men beroende på vilken frågeställning man har kommer data att variera i sitt informationsinnehåll.

Den enklaste typen av data ger information om **klasstillhörighet**. Exempel är kön, färg samt förekomst eller frånvaro av något man studerar (tabell 1a). Data på denna nivå ger däremot inte någon information om antalet av exempelvis honor och hanar.

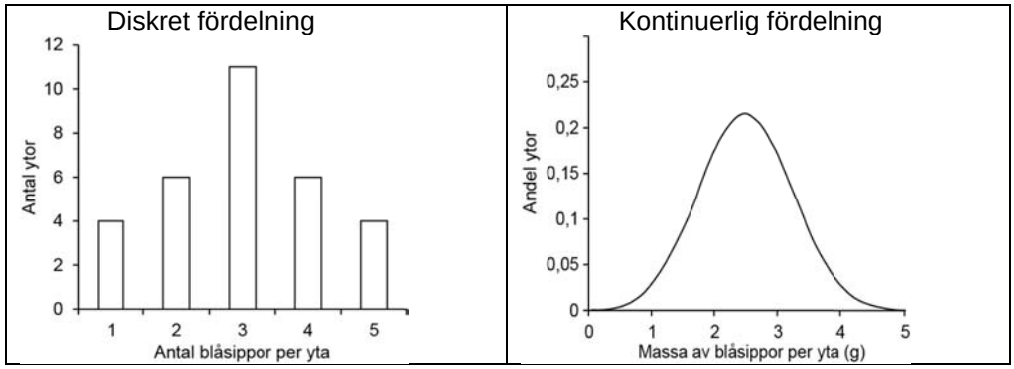
Mera information får man från data som är numeriska. Det kan vara **frekvenser**, dvs. att man räknat antalet av den variabel som studeras. Det kan till exempel vara antalet mygglarver i sedimentprov från olika hårt försurade sjöar. Om man istället mätt längden på mygglarverna hade man erhållit metriska värden, så kallade **skalvärden** (tabell 1b).

Tabell 1 a, b. Exempel på olika typer av data. I tabellen till vänster visar den första kolumnen **klasstillhörighet**. Kolumn två ger **frekvenser** för de olika klasserna. Tabellen till höger innehåller rådata till den vänstra tabellen där **klassvariabeln** färg och **mätvärdet** längd ges.

Färg	Antal
Röd	45
Blå	38
Grön	44

Prov Nr	Färg	Längd
1	Blå	25,3
2	Röd	26,1
3	Röd	31,4
.	.	.
.	.	.
.	.	.
126	Grön	21,8
127	Blå	22,0

Variabler som kan kvantifieras, dvs. anta olika värden, kan vara **diskreta** eller **kontinuerliga** (figur 2). En diskret variabel kan bara anta vissa värden, medan en kontinuerlig variabel kan anta vilket värde som helst inom ett intervall. Antalet blåsippor per provyta kan bara vara heltal, medan vikten av alla blåsippor i samma provytor i princip kan anta vilka värden som helst som inte är mindre än noll. Begränsande faktorer är ytans storlek och vågens noggrannhet.



Figur 2. Exempel på en diskret och en kontinuerlig variabel. Den diskreta variabeln kan bara anta vissa värden, medan den kontinuerliga variabeln kan anta vilket värde som helst. I detta exempel kan dock ingen av variablerna anta ett värde lägre än noll.

Skaltyper

Skaltyp är ett annat sätt att beskriva hur mycket information olika typer av data bär med sig. De olika typerna kan delas in i fyra klasser: **Nominalskala**, **Ordinalskala**, **Intervallskala** och **Kvotskala**. För var och en av dessa klasser gäller olika begränsningar vad gäller de sätt man kan behandla data statistiskt.

Data på **nominalskalan** är den enklaste typen av data. Detta är egentligen ingen skala utan en klassificering av olika objekt, eftersom man inte gjort någon mätning. Exempel är hona eller hane, norr eller söder om norrlandsgränsen samt krona eller klave.

Data på **ordinalskalan** bygger på en rangordning av vissa mätta egenskaper. Man ska kunna avgöra om ett värde är högre eller lägre än ett annat, men man kan inte säga hur mycket högre eller lägre. Ett exempel är Naturvårdsverkets bedömningsgrunder för miljökvalitet som använder en femgradig skala. Man kan säga att värdet 2 är bättre än värdet 3 ur miljösynpunkt, men inte hur mycket bättre eftersom det ligger en mängd miljökemidata bakom de fem klasserna.

Data på en **intervallskala** har alla egenskaper som data på ordinalskalan har, men dessutom har avståndet mellan värdena en innebörd. Avståndet mellan 2 och 3 är lika stort som avståndet mellan 3 och 4. Detta kallas ekvidistanta skalsteg. Man kan dock inte säga att 2 är dubbelt så mycket som 4 när man har data på en intervallskala.
Ett exempel är Celsiusgrader. Det går inte att säga att det idag är dubbelt så varmt som igår då det var 0 grader. På motsvarande sätt är inte +10° dubbelt så varmt som +5°.

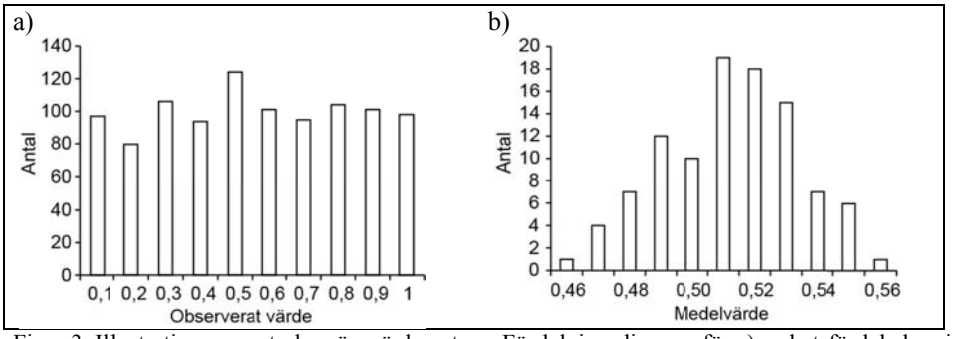
Anledningen till detta är att intervallskalan har en nollpunkt som är godtyckligt satt. Den saknar med andra ord en naturlig nollpunkt.

Slutligen, data på **kvotskalan** skiljer sig egentligen bara från data på intervallskalan genom att man kan bilda en kvot av två mätvärden, vilket kommer sig av att kvotskalan har en naturlig nollpunkt. Exempel på data på kvotskalnivå är vikt, fotosyntesaktivitet och avstånd.

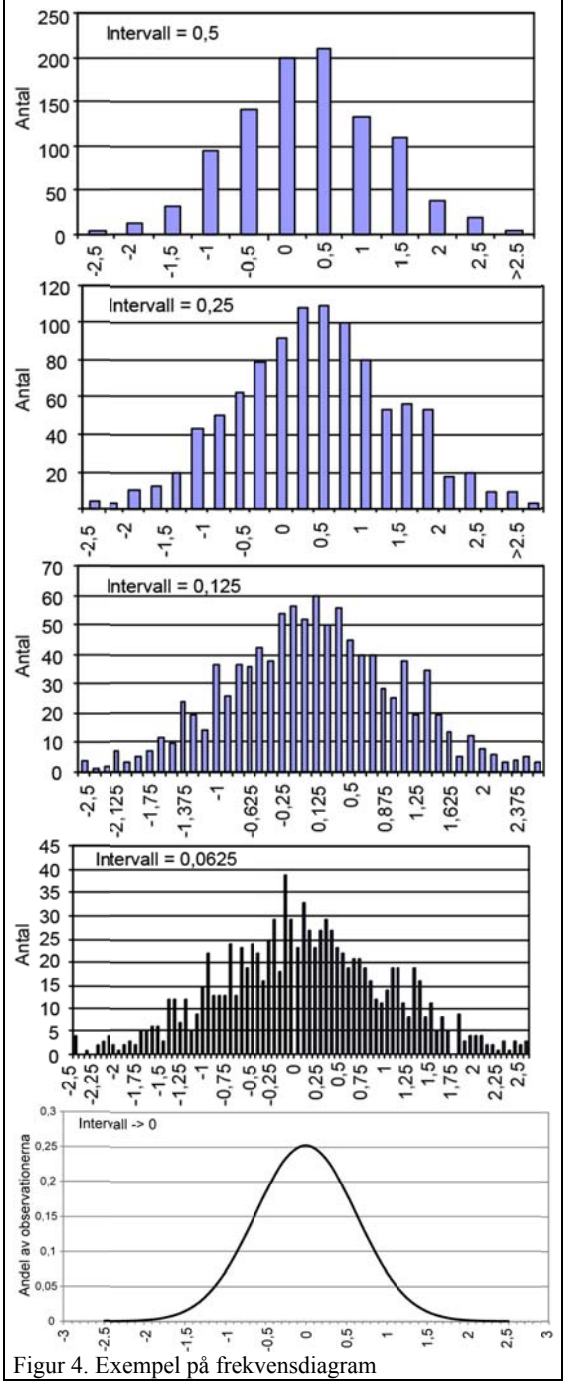
Normalfördelning och approximativ normalfördelning

Allt man observerar och mäter påverkas av en mängd olika slumpfaktorer. När en variabel påverkas av flera oberoende slumpvariabler kommer den att följa en klockformad kurva som kallas Gauss-fördelning eller normalfördelning (figur 4). Normalfördelningen har en central roll i statistiken på grund av sin relation till det vi kallar centrala gränsvärdessatsen. De matematiska reglerna för centrala gränsvärdessatsen gör att man om man har ca 20 eller fler observationer kan anta att data från de flesta fördelningar är s.k. approximativt normalfördelade. Detta gör att man kan använda de vanligaste statistiska testerna som t.ex. t-test och variansanalys även om populationerna inte är normalfördelade.

Ett tankeexempel får belysa vad som menas med approximativ normalfördelning. Grunden bygger på relationen mellan centrala gränsvärdessatsen och normalfördelningskurvan. Tänk dig en population med en känd fördelning (den behöver inte vara normalfördelad). Dra många stickprov från populationen, men lägg tillbaka alla prover efter varje dragning. Beräkna medelvärden för varje stickprov och gör ett histogram som visar fördelningen av medelvärdena. Enligt centrala gränsvärdessatsen kommer detta histogram visa en normalfördelning, även om urspringspopulationen inte var normalfördelad. Man säger att ursprungspopulationen är approximativt normalfördelad. I figur 3a visas ett frekvensdiagram för en betafördelad variabel med 1000 observationer. Figur 3b visar fördelningen av medelvärden från 100 stickprov med vardera 100 observationer från populationen i figur 3a, och precis som centrala gränsvärdessatsen förutsäger visar medelvärdena en normalfördelning.



Figur 3. Illustration av centrala gränsvärdessatsen. Fördelningsdiagram för a) en betafördelad variabel och b) medelvärden för 100 stickprov från a.



Figur 4. Exempel på frekvensdiagram

Frekvensdiagram

När man vill undersöka hur de data man samlat in är fördelade är det lämpligt att göra ett frekvensdiagram (figur 4). Dessa typer av diagram används också ofta slentrianmässigt när man vill förklara antaganden bakom olika statistiska tester.

Grunden i ett frekvensdiagram är att man delar in sina data i intervall och räknar hur många observationer som hamnar inom varje intervall. I det översta diagrammet i figur 3 är data indelade i intervall om 0,5 enheter och antalet observationer som hamnar inom varje intervall anges på y-axeln. I diagrammet närmast under är intervallen 0,25 enheter. Därunder visas två diagram där intervallstorleken ånyo halveras. Att intervallen blir minde och mindre framgår även av att staplarna blir smalare och smalare.

I den understa figuren är intervallen oändligt smala, vilket resulterar i en fördelningskurva istället för stolpdiagram. I en fördelningskurva är ytan under kurvan fram till en visst x-värde proportionell mot andelen observationer mindre än eller lika med det valda x-värdet. Hela ytan under kurvan har arean 1.

Alla diagrammen figur 4 visar fördelningen för samma 1000 mätvärden. Av kurvans utseende kan vi sluta oss till att dessa mätvärden är normalfördelade. I mera tveksamma fall finns det statistiska tester för att pröva om de data man samlat in är normalfördelade eller ej. Sådana tester ingår i de flesta statistikprogram för datorer, t.ex. Kolmogorov-Smirnovs test och Shapiro-Wilks test.

Parametriska eller icke-parametriska tester?

För att svara på frågan i rubriken måste vi först förstå vad som menas med parameter i detta avseende. Data som är normalfördelade eller approximativt normalfördelade kan beskrivas med medelvärde och något mått på variationen kring medelvärdet. Detta variationsmått kan vara standardavvikelse eller varians. Alla dessa beskrivningar kallas *parametrar* för ett stickprov eller en population. Data från fördelningar som inte kan inte beskrivas med medelvärden har sålunda inga parametrar av denna typ. Det är till exempel omöjligt att beräkna ett medelvärde av mätvärdena blå och röd. Om man inte kan beräkna parametrar kan man följaktligen inte använda statistiska tester som bygger på populationsparametrar utan är hänvisad till s.k. icke-parametriska tester.

Generellt gäller att icke-parametriska data innehåller mindre information än parametriska data. Ett exempel är Naturvårdsverkets bedömningsgrunder för miljökvalitet där avvikelseklass 1 är bättre än

klass 2, men man kan inte säga hur mycket bättre eftersom dessa siffror är på ordinalskalnivån. Om man istället ser till fosfathalt i sjövatten är 1 µg/l hälften av 2 µg/l, vilket är mer information än vad som ges av en etta och en tvåa på ordinalskalenivån. Om man vill använda avvikelseklasser i ett statistiskt test är man sålunda tvungen att använda icke-parametriska tester, medan tester som baseras på fosfathalter kan vara parameriska.

I de flesta icke-parametriska testerna räknar man på observationernas rangordningstal. Detta gör att alla extremer i mätdata försvinner och ersätts med rangtal. Extremernas rangtal hamnar visserligen i ytterkanten av fördelningen men de sticker inte ut som originaldata. I figur 5 illustreras hur fördelningen av ett antal mätvärden blir helt förändrad då man ersätter dem med deras rangordning. Medelvärdet av mätningarna är 4,25 medan medel för rangtalen är 14,5.

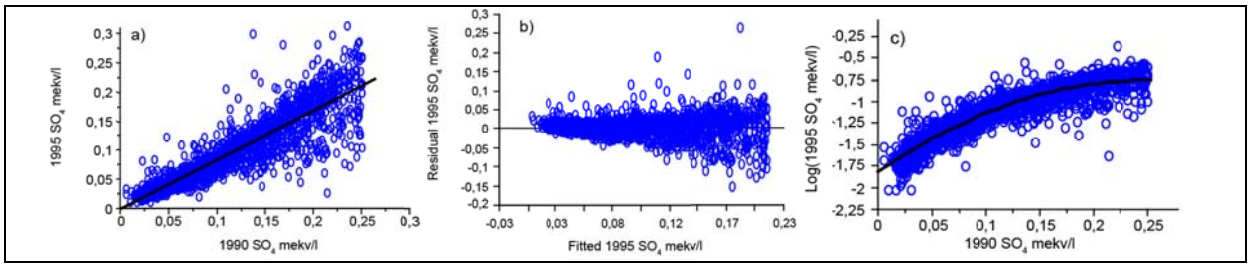
Figur 5. Jämförelse av fördelningen av några mätvärden och de rangtal samma mätserie ger.

Som synes kan en rankning helt förändra strukturen hos de data man analyserar. Man skall således vara medveten om att icke-parametriska tester inte har samma styrka som pararmetriska tester. Trots detta är det ibland är tvunget att ta till de icke-parametriska testerna eftersom data inte tillåter att man använder parametriska tester.

Datatransformation

Många tester bygger på att data är normalfördelade eller åtminstone approximativt normalfördelade. Om man vet att data kommer från en annan fördelning eller har vissa egenskaper bör man transformera sina data innan man börjar utföra några beräkningar (tabell 2).

Ibland är inte variationen jämt fördelad över ett helt mätintervall (figur 6). Detta kallas heteroskedasticitet och gör att parametriska tester inte kan användas. Om man har ojämnt fördelad varians kan datatransformation vara ett sätt att jämna ut variationen. I figur 6c visas hur en ojämn variation i 6a försvinner när y-variabeln logaritmeras. Den ojämnt fördelade variationen framgår tydligt i det s.k. residualdiagrammet i 6b där avstånden mellan regressionslinjen och mätvärdena ökar med ökande värde på x-axeln.



Figur 6. Jämförelse mellan SO_4 koncentrationer i Svenska sjöar provtagna 1990 och 1995. I b) visas avstånden mellan regressionslinjen i a) och mätpunkterna. I figur c) visas samma samband som i a) men med logaritmerade värden för 1995.

Några typer av data har inneboende egenskaper som leder till att statistiska tester blir felaktiga om man inte utför en datatransformation. Detta gäller t.ex. proportioner och angivelser av varaktighet. I tabell 2 visas exempel på transformationer som måste utföras, respektive kan utföras om variationen är ojämnt fördelad.

Tabell 2. Exempel på olika typer av datatransformation

Typ av data	Exempel	Obligat	Transformation
Proportioner eller %	Andel av olika funktionella grupper i prover	Ja	Arcsin√p
Tid eller varaktighet	Antal minuter per dag som ett gränsvärde överskrids	Ja	1/x
Cirkulära data	Olika kompassriktningar, α	Ja	Se t.ex. Zar (1999)
Antal (Poisson-fördelade)	Antal arter per prov	Nej	t.ex. √x
Mätvärde från en lognormal fördelning	Biomassa av olika arter i prover	Nej	t.ex. log(x + a)*

* om x = 0 går detta ej att beräkna, därför lägger man ofta till en konstant, oftast 0,1 eller 1

Hypotesprövning

Vid statistisk hypotesprövning arbetar man med två typer av hypoteser. Dels **forskningshypoteser** och dels **statistiska hypoteser**. En forskningshypotes är det antagande man vill testa med rigorösa vetenskapliga tester. Detta görs genom att formulera ett antal utsagor och konstruera experiment som ger tillräckligt med data för att kunna pröva om hypotesen stämmer. I miljöövervakningssamanhang kan man dock sällan göra experiment, istället får man använda insamlade data. En forskningshypotes kan vara formulerad hur som helst, t.ex. ”vi tror att kalkningen av Våtsjön har förändrat alkaliniteten jämfört med mätningen just innan kalkningen”.

Statistiska hypoteser följer till skillnad från forskningshypoteser ett strikt mönster. Först ställer man upp en s.k. **nollhypotes** som alltid formuleras så att den beskriver att det inte finns någon skillnad eller effekt. Sedan formuleras en eller flera s.k. **mothypoteser** eller alternativhypoteser. Nollhypotesen skrivs ofta H_0 (uttalas hå-noll) och mothypoteser skrivs H_{1a}, H_{1b}… o.s.v.

För att testa antagandet om alkalinitet i Våtsjön ställer man upp följande statistiska nollhypotes:

H_0: Alkaliniteten idag skiljer sig inte från alakaliniteten innan kalkningen.

En och tvåsidiga mothypoteser

Om man inte vet åt vilket håll en eventuell förändring har skett formulerar man en mothypotes som säger att det skett en förändring, en s.k. tvåsidig hypotes. Vet man att om det skett en förändring kan den bara innebära en ökning (eller en minskning) formulerar man istället en mothypotes som säger att det skett en ökning (minskning), en s.k. ensidig hypotes.

En tvåsidig mothypotes i exemplet ovan kan formuleras som:

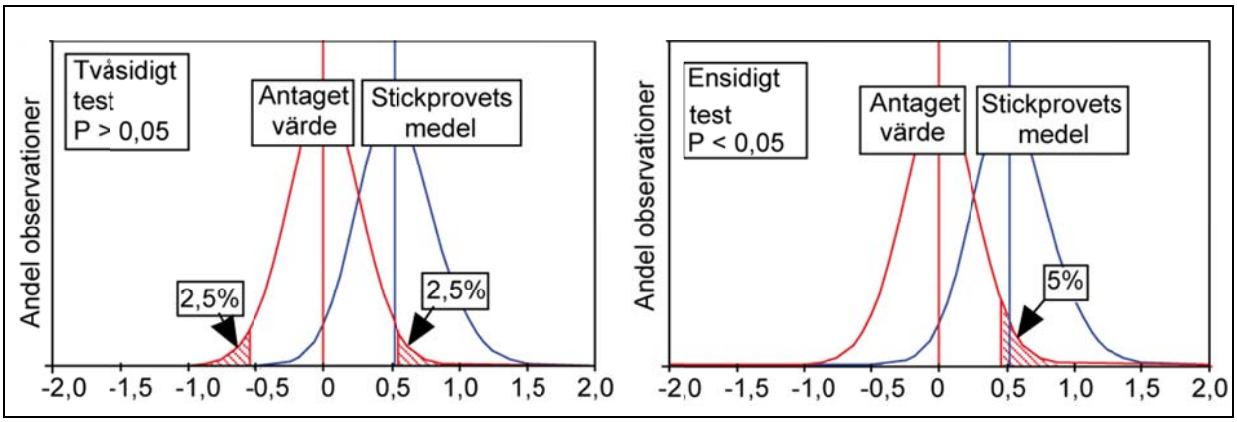
H_{1a}: Alkaliniteten idag skiljer sig signifikant från alkaliniteten för fem år sedan.

Om det skett kalkning i sjön kan man anta att alkaliniteten inte har minskat och man kan då ställa upp en ensidig mothypotes:

H_{1b}: Alkaliniteten har ökat sedan mätningen för fem år sedan.

Ovanstående exempel är dock inget annat än ett exempel för att illustrera skillnaden mellan en- och tvåsidiga tester. I verkigheten är det mer än en kalkning som påverkar alkaliniteten i en sjö och i princip skulle man kunna få lägre värden efter kalkningen.

I figur 7 illustreras hur utfallet av ett statistiskt test beror av hur mothypotesen är formulerad. I båda figurerna testas om ett stickprov med medelvärdet 0,52 skiljer sig från ett teoretiskt värde på 0. Om man ska anse stickprovet signifikant skiljt från 0 måste stickprovets medelvärdet hamna i svansen av fördelningen kring nollvärdet (den vänstra kurvan). Väljer man t.ex. 5%-nivån (se nästa avsnitt) innebär detta att stickprovets medelvärde måste hamna i den del av svansen som utgör 5% av hela ytan under kurvan runt nollvärdet. En fördelningskurva har dock två svansar därför är man tvungen att dela upp de 5 procenten på 2,5% i vardera svansen. I detta fall indikeras 5% av fördelningen för nollvärdet genom de streckade områdena. I den övre figuren har man en tvåsidig mothypotes och får resultatet att stickprovet inte skiljer sig från 0. I den undre figuren har man en ensidig mothypotes och alla 5% av fördelningen lagda i den högra svansen. I detta test får man resultatet att provet är skiljt från 0. Det är således lättare att upptäcka en skillnad om man i förväg kan förse testet med information om de data man testar. I det här fallet är informationen att förändringen bara kan gå åt ett håll.



Figur 7. Illustration av hur mothypotesen i statistisk hypotesprövning påverkar utfallet av ett statistiskt test. Vid ett tvåsidigt test krävs större skillnad i medelvärden för att kunna detektera en skillnad, än vid ett ensidigt test.

Teorin bakom hypotesprövning

Idag använder man nästa alltid datorer för statistiska beräkningar. Detta gör att man oftast inte följer teorin bakom den statistiska hypotesprövningen när man räknar. Om man inte känner till den teorietiska bakgrunden är det dock svårt att veta vad man testar och vad siffrorna ifrån datorn betyder. Oavsett om man räknar för hand eller utnyttjar dator finns det en strikt teori som man i princip ska följa för att få korrekta resultat. Teorin bygger på en hierarkisk ordning av beräkningssteg och dessa kan enklast presenteras i punktform:

1. *Urval*. Är de data man vill pröva ett *slumpmässigt urval* ur den bakomliggande populationen? Om inte gäller inte förutsättningarna för statistiska tester.
2. *Data*. Vilken skaltyp och vilken fördelning har de data man vill testa. Naturen hos de data man vill prova avgör vilka tester man kan använda. *Normalfördelade data* kan analyseras med *parametriska tester* medan data med andra fördelningar måste prövas med *icke-parametriska tester*.
3. *Hypoteser*. Vid all hypotesprövning måste (minst) två hypoteser ställas mot varandra. Den första hypotesen är den hypotes som ska prövas i det statistiska testet och kallas ofta *nollhypotes* eller *hypotesen om ingen skillnad*. Oftast ställs nollhypotesen upp enbart för att kunna förkastas. Den eller de andra hypoteserna är mothypoteser och formuleras efter den frågeställning man har. I testproceduren kan nollhypotesen antingen *förkastas* eller *inte förkastas*, men *aldrig* accepteras!
4. *Signifikansnivå*. Vilken risk är man villig att ta för att *förkasta en sann nollhypotes*. Signifikansnivån, eller α, är den risk man tar att det stickprov man samlat på grund av slumpen visar ett annat mönster än hela populationen och att man därigenom tar ett felaktigt beslut i frågan om att behålla eller förkasta nollhypotesen. Vid all hypotesprövning kan man hamna i fyra olika situationer (tabell 3).

Tabell 3. Illustration av α och β i statistisk hypotesprövning.

		Verkligt förhållande (okänt)	
		H_0 sann	H_0 falsk
Resultat av statistiskt test	H_0 förkastas	Typ I-fel, α	Korrekt beslut, Styrka (1-β)
	H_0 förkastas ej	Korrekt beslut, 1-α	Typ II-fel, β

Ett typ I-fel är att förkasta en sann nollhypotes och α är den risk man tar att detta ska inträffa. Ett typ II-fel är att behålla en falsk nollhypotes och denna risk ges av β. Typ II-fel är beskrivna under avsnittet statistisk styrka i *Planering av undersökningar* i Naturvårdsverkets handbok för miljöövervakning.

5. *Testfunktion*. Testfunktionen är en egenskap som beräknas utifrån stickprovet och fungerar som *beslutsunderlag* för att bestämma om nollhypotesen ska förkastas eller ej. För varje statistiskt test finns en specifik formel hur man beräknar testfunktionen. Beroende på vilket test man utför kommer testfunktionen att ha olika namn, t. ex. F, t eller χ^2 (chi-två). Dessa tester finns utförligt beskrivna längre fram.

6. *Beslutsregel*. Med utgångspunkt från testfunktionens fördelning formuleras en beslutsregel som ger det *kritiska värde* som anger om nollhypotesen ska förkastas eller ej. I tabeller finns olika testfunktioners fördelning för olika signifikansnivåer och för olika stora stickprov. Med hjälp av dessa kan man fastlägga vid vilket värde på testvariabeln nollhypotesen ska förkastas. En beslutsregel kan formuleras som ”Förkasta nollhypotesen om värdet på testvariabeln överstiger XX”.
7. *Beräkningar*.
8. *Statistiskt beslut*. Förkasta eller låt bli att förkasta nollhypotesen.

Detta är grunden i all statistisk hypotesprövning. Idag är det dock ovanligt att man följer alla dessa steg då man har datorprogram som direkt ger all information man behöver för att avgöra om man kan förkasta sin nollhypotes. De statistiska testerna i datorprogrammen bygger dock på ovanstående punkter. Av den anledningen är det inte korrekt att, som ofta sker, tala om "högsignifikanta resultat" eftersom svaret i den sista punkten endast kan bli förkasta eller inte förkasta nollhypotesen, varken mer eller mindre. Signifikansnivån (punkt 4 ovan) har man satt långt innan man tar sitt statistiska beslut (punkt 8). Däremot måste man ange den signifikansnivå man antagit då man tog sitt statistiska beslut. Det har dock blivit mer och med vanligt att man jobbar med den faktiska signifikansnivån som ges av P-värdet i ett test, i och med att datorerna gjort sitt intåg.

Upprepade tester och Bonferroni-korrektion

Om man t.ex. väljer $\alpha = 0{,}05$ kommer ett stickprov man samlat in i medeltal1 gång av 20 att vara så extremt att man begår ett typ I-fel (1/20 = 0,05). En analogi är ett lotteri med vinst på var tjugonde lott. I medeltal måste man ta tjugo lotter för att få en vinstlott, men man kan få extremen, d.v.s. vinstlotten, redan i första försöket. I ett statistiskt test motsvaras vinstlotten av ett stickprov där alla prover är extremer i förhållande till det verkliga, men okända, medelvärdet.

Om man utför upprepade tester på samma datamaterial och under samma nollhypotes (t.ex. flera t-tester eller flera korrelationer, se nedan) ökar risken att man begår ett typ I-fel för varje upprepat test. För upprepade t-tester på samma material (vid $\alpha = 0{,}05$) är sannolikheten för ett typ I-fel redan vi första upprepningen 0,13, vid en andra upprepning blir sannolikheten 0,21 och vid 20 upprepningar blir sannolikheten för ett typ I-fel så hög som 0,92! För att undvika detta brukar man justera nivån på α genom att dividera α med antalet upprepade tester. Detta kallar Bonferroni-korrektion. Så, vid t.ex. tre upprepade tester på samma datamaterial och under samma nollhypotes och $\alpha = 0{,}05$ måste P-värdet understiga 0,05/3 = 0,0167 för att testen ska anses signifikanta.

Avvikande värden ”Outliers”

Ett inledande steg vid all dataanalys bör vara att titta på en grafisk framställning alla data (se *Planering av undersökningar* i Naturvårdsverkets handbok för miljöövervakning). Om man upptäcker värden som avviker väldigt mycket från majoriteten av alla observationer finns det anledning att fundera över varför dessa är så avvikande. Ett avvikande värde kan uppstå av flera anledningar:

1. I vissa fall kan det röra sig om rena skrivfel, vid datainsamling eller vid inmatning i dator.
2. En provtagningsplats kan vara starkt avvikande p.g.a. tillfälliga lokala störningar.

3. Det avvikande värdet beror på att enstaka datum (= singular av data) i den studerade (statistiska) populationen skiljer sig så mycket från övriga datum att man kan anta att avvikande data inte tillhör samma population.

Om man med stor säkerhet vet att ingen av de tre påståendena ovan är anledningen till avvikelsen finns två möjligheter att gå vidare:

1. Det avvikande värdet representerar en del av den naturliga variationen. I detta fall bör det avvikande värdet tas med i analysen.
2. Den andra möjligheten är att det avvikande värdet beror av ett analys- eller observationsfel. Detta gör att värdet inte kommer från samma bakomliggande statistiska population som övriga analysvärden, vilket är en förutsättning för många statistiska tester. Detta gör att man bör ta bort det avvikande värdet.

Nu är det svårt att veta vilken av de två ovanstående punkterna som är den rätta. För att få hjälp att ta ett beslut i frågan kan man använda normalfördelningskurvan. Genom att utnyttja normalfördelningskurvans välkända egenskaper kan man beräkna hur stor sannolikhet det är att få ytterligare ett värde som avviker lika mycket från övriga värden som det värde man vill kontrollera. Ett sådant test är Grubbs test.

Det första man gör i Grubbs test är att beräkna ett standardiserat mått på hur långt det avvikande värdet är från övriga värden. Detta gör man genom att beräkna den s.k. z-kvoten:

$$z = \frac{|medel - värde|}{std.av.}.$$

För att sedan testa sannolikheten att det framräknade värdet på z indikerar ett avvikande värde kan man antingen använda en tabell på kritiska nivåer på z för Grubbs test, eller beräkna ett approximativt värde på P-nivån. Tabeller för Grubbs test finns i vissa statistikböcker (t.ex. Grubbs, F. E. & Beck, G. 1972. Technometrics, 14:847-854.) och på ett flertal webbsidor. Om man inte har tillgång till tabell får man beräkna signifikansnivån. Detta gör man genom att först beräkna ett s.k. t-värde m.h.a. följande formel:

$$t = \sqrt{\frac{N \times (N-2) \times z^2}{(N-1)^2 - N \times x^2}},$$

där N är antalet prover och z är det beräknade z-värdet.

Därefter tar man reda på det tvåsvansade P-värdet i t-fördelningen för det värde på t man beräknat och N-2 frihetsgrader. Detta kan man göra genom att titta i en tabell över t-fördelningen (tabell 4). Antag att man vill veta P-värdet för t = 3,25 vid 12 frihetsgrader. Detta gör man genom att i den rad som motsvarar aktuellt antal frihetsgrader leta rätt på det kritiska värde som är närmast lägre det värde man vill testa. I detta fall är det tabellvärdet 3,055 som är närmast lägre det beräknade värdet 3,25. Den signifikansnivå som motsvaras av 3,055 är 0,005. Vi kan således konstatera att det tvåsvansade P-värdet för t = 3,25 med 12 frihetsgrader är lägre än 0,005, men högre än 0,0025 som är P-värdet för t-värdet närmast högre än 3,25. En annan metod är att använda Microsoft Excel som kan ge det exakta P-värdet. Detta gör man genom att ge kommandot =TFÖRD(t;df;2) (TDIST på engelska). I exemplet ovan får man P = 0,0035.

Tabell 4. Utdrag ut t-tabell. Fullständiga t-tabeller finns i nästan alla statistikböcker.

df	Signifikansnivå							
	0,4	0,25	0,1	0,05	0,025	0,01	0,005	0,0025
…	…	…	…	…	…	…	…	…
11	0,260	0,967	1,363	1,796	2,201	2,718	3,106	3,497
12	0,259	0,695	1,356	1,782	2,179	2,681	3,055	3,428
13	0,259	0,694	1,350	1,771	2,160	2,65	3,012	3,372
…	…	…	…	…	…	…	…	…

Nästa steg är att multiplicera det erhållna P-värdet med N som är antalet prover. Den erhållna produkten är ett approximativt P-värde för Grubbs test. Om det slutgiltiga P-värdet är lägre än α kan man konstatera att det testade värdet är avvikande. Man får däremot inte något svar på om man ska ta bort den avvikande observationen eller ej eftersom detta beror på fler faktorer än bara det numeriska värdet.

Detta test ger ett approximativt P-värde. Om z är högt är dock P-värdet tillförlitligt. För låga värden på z kan P-värdet bli för högt. I exemplet ovan blir det slutgiltiga P-värdet $0{,}035 \times 14 = 0{,}049$. Vi kan således konstatera att det värde vi misstänker är avvikande är på gränsen till att betraktas som avvikare om man väljer $\alpha = 0{,}05$. Om man använt t-tabell får man istället att P-värdet är mindre än $(0{,}005 \times 14) = 0{,}07$ och större än $(0{,}0025 \times 14) = 0{,}035$.

Beskrivande statistik - ett stickprov

Centralmått och spridningsmått

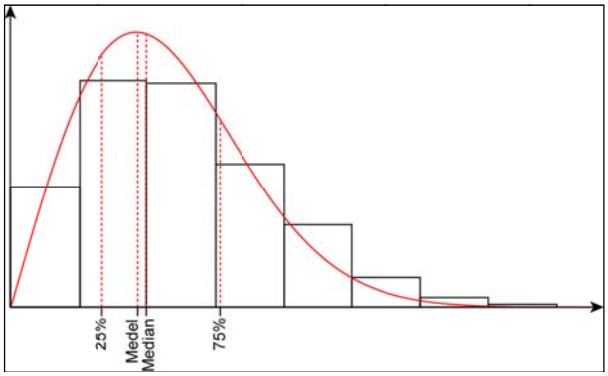
För att ge en bild av de data man samlat in kan man ge någon form av mått på det som kallas centraltendens, d.v.s. man anger det värde som är mest representativt för de data man samlat in. Ytterligare information om sina mätdata kan man ge genom att beskriva hur de sprider sig kring det centralmått man angivit.

Det finns tre mått på central tendens: *Typvärde*, *Medianvärde* och *Medelvärde*. Vilket mått som används beror på vilken skaltyp hos de data man vill beskriva. Eftersom de flesta datorprogram för statistik är på engelska ges här även den engelska termen.

Typvärde (eng. mode) är det mest förekommande värdet bland de data man samlat. För att beskriva spridningen använder man *variationsvidd* (eng. range), vilket är skillnaden mellan det högsta och det lägsta värdet man observerat. Dessa båda parametrar ger ganska lite information om de data man har vid handen. Typvärde är det enda korrekta mått man kan använda för att beskriva central tendens för data på nominalskalnivån.

Medianvärde (eng. median) är det värde som den mittersta observationen har. Om man har ett jämt antal observationer används medelvärdet av de två mittersta observationerna. Ett medianvärde delar en fördelning i två lika delar så att 50% av värdena är lägre än medianvärdet, och 50% är högre. Median används ofta som centralmått för data på ordinalskalnivå och vid så kallade sneda fördelningar av data på intervall- och kvotskalnivå.

Spridningen kring ett medianvärde anges med *kvantilavvikelse*. Oftast används percentil- eller kvartilavvilkese. Kvartilavvikelse innebär att man rangordnar observationerna och delar upp dem i fyra grupper så att de lägsta 25% av observationerna hamnar i den första kvartilen. Observationerna mellan 25% och medianvärdet hamnar i den andra kvartilen, och så vidare (figur 8).



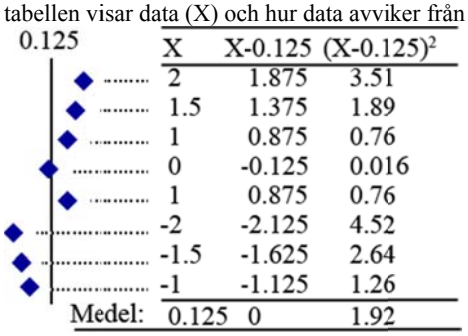
Figur 8. Median och kvartilavstånd i en skev fördelning.

Om en fördelning är normalfördelad eller approximativt normalfördelad används aritmetiskt *medelvärde* (eng. mean) för att beskriva central tendens.

Spridningen kring ett medelvärde ges av variansen (s^2) eller standardavvikelsen (s). Variansen är ett medelvärde av hur mycket de olika observationerna avviker från stickprovets medelvärde. Detta är illustrerat i tabell 5. Figuren till vänster visar åtta värden och deras avstånd till medelvärdet av de åtta värdena. Skillnaden mellan vart och ett av värdena och medelvärdet ges i den andra kolumnen. Som framgår av den sista raden är medelvärdet av avstånden i kolumn två lika med noll. Detta indikerar att alla observationer är lika med medelvärdet, vilket inte är fallet. Problemet är att negativa och positiva avvikelser från medelvärdet tar ut varandra då man beräknar ett medelvärde av avvikelserna. För att undkomma detta kvadrerar man värdet på varje avstånd, vilket är gjort i den tredje kolumnen. Medelvärdet av dessa är sedan ett mått på variationen, men det är inte den statistiska variansen. För

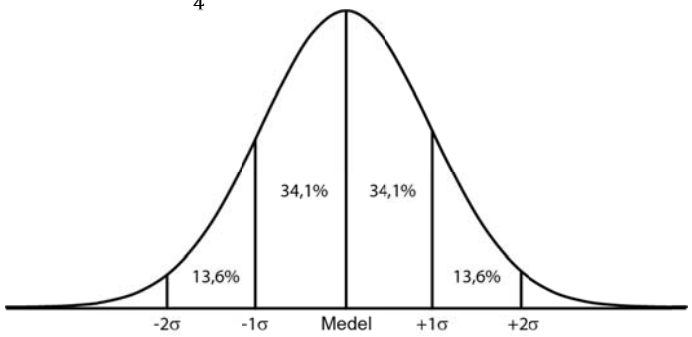
detta krävs en liten förändring i formeln för beräkning av medelvärde. Normalt beräknas ett medelvärde genom att summera alla värden och dela med antalet värden. För att beräkna variansen ska man dividera summa av alla värden med antalet värden minus 1. Formeln för detta blir:
$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$,
där s^2 är variansen, x_i de olika observationerna och $\bar{x}$ är medelvärdet.

Tabell 5. Illustration av det statistiska begreppet varians. Figuren visar hur data fördelar sig kring sitt medelvärde (0,125), och tabellen visar data (X) och hur data avviker från medelvärdet.



	X	X-0.125	$(X-0.125)^2$
	2	1.875	3.51
	1.5	1.375	1.89
	1	0.875	0.76
	0	-0.125	0.016
	1	0.875	0.76
	-2	-2.125	4.52
	-1.5	-1.625	2.64
	-1	-1.125	1.26
Medel:	0.125	0	1.92

Spridningen kring medelvärdet ges även av *standardavvikelsen* (s). I normalfördelningen ligger 68,3% av alla observationer inom intervallet medel ± 1 s och 95,4% av observationerna inom intervallet medel ± 2s (figur 9). En snabb skattning av standardavvikelsen får man genom att dividera skillnaden mellan det högsta och det lägsta värdet med 4. Stickprovets standardavvikelse beräknats enligt
$s = \sqrt{s^2} \approx \frac{x_{max} - x_{min}}{4}$.



Figur 9. Medelvärde och standardavvikelse i en normalfördelning. Området som utgörs av 1 standardavvikelse (σ) upp och ner från medelvärdet utgör 2 × 34,1 % av hela arean. Två standardavvikelser från medelvärdet inkluderar ytterligare 2 × 13,6 %, totalt 95,8 % av arean under normalfördelningskurvan.

När man ska välja vilket mått på central tendens man ska använda är fördelningens utseende av stor betydelse. Ofta är ett stickprov åtminstone approximativt normalfördelat och då kan det aritmetiska medelvärdet användas. Vid skeva fördelningar är det mera lämpligt att använda median (tabell 6). Observera att det engelska ordet *average* betyder centralmått och kan sålunda betyda flera olika saker.

Utöver det aritmetiska medelvärdet finns det andra sätt att beräkna medelvärdet. Vilket sätt man väljer beror på vilka data man har.

Trimmat medelvärde är det aritmetiska medelvärdet men där man tagit bort de högsta och de lägsta observationerna, t.ex. 10 % av alla observationer i vardera ändan. Detta gör att medvärdet blir mindre känsligt för starkt avvikande värden. Ett trimmat medelvärde av t.ex. 1, 2, 3, 4 och 5 blir (2+3+4)/3 = 3.

Geometriskt medelvärde är *n*:te roten ur produkten av *n* stycken värden. Används med fördel när man vill beskriva medelvärden av kvoter eller procentsatser. En förutsättning är att alla värden måste vara positiva och på kvotskalnivå. Ett geometriskt medelvärde av 1, 2, 3, 4 och 5 blir

$\sqrt[5]{1\times2\times3\times4\times5} = 2{,}6$.

Ett **harmoniskt medelvärde** är det inverterade värdet av det aritmetiska medelvärdet av de inverterade värdena av observationerna. Ett harmoniskt medelvärde av 1, 2, 3, 4 och 5 blir $\frac{1}{\left(\frac{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}}{5}\right)} = \frac{5}{\left(\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}\right)} = 2{,}2.$

Den andra varianten av beräkningen är en förenkling av den första varianten. Harmoniska medelvärden används t.ex. när man vill beräkna ett medelvärde av olika hastigheter eller varaktigheter.

Medelvärdet av **cirkulära data**, t.ex. gradtal, måste också beräknas på ett speciellt sätt. Numeriskt är det en väldigt stor skillnad mellan 359° och 1°, men i verkligheten är skillnaden bara 2°. Detta problem kommer man till rätta med genom att använd olika trigonometriska funktioner i beräkningarna. För ett antal, *n*, kompassriktningar, α_i, beräknas medelriktningen $\bar{\alpha}$ i följande steg:

$\bar{\alpha} = arccos\left(\frac{y}{r}\right)$,
där $y = \frac{\sum \cos \alpha_i}{n}$ och $r = \sqrt{x^2 + y^2}$, där $x = \frac{\sum \sin \alpha_i}{n}$.

Tabell 6. Sammanställning av olika sätt att beskriva central tendens och spridning, vid olika skaltyper. De olika sambandsmåtten gås igenom längre fram.

Skaltyp	Centralmått	Spridningsmått	Sambandsmått
Nominal	Typvärde	—	Phi, Cramérs V
Ordinal, och icke normal-fördelade data	Median	Kvantilavvikelse	Rangkorrelation
Intervall och kvotskala	Medelvärde	Standardavvikelse	Produktmoment-korrelation

Det är sällan som medelvärdet för ett stickprov är exakt detsamma som populationens medelvärde. Medelvärdet för ett stickprov beror av vilka värden som råkar hamna i stickprovet. Skillnaden mellan stickprovets och populationens medelvärden beror bl. a. på variationen i stickprovet och antalet prover. Genom att kombinera dessa två mått får man dels få ett mått **medelfelet** och dels ett **konfidensintervall** för stickprovet.

Medelfel, (eng. standard error of mean, sem) är ett mått på hur långt medelvärdet för ett stickprov är från populationens medelvärde. Eftersom man har större chans att komma i närheten av det populationsmedelvärdet ju större stickprov man tar, minskar medelfelet med ökande stickprovsstorlek. Medelfelet beräknas enligt: s/√n, och som synes minskar kvoten med större stickprov.

Konfidensintervallet (eng. confidence interval, CI) anger precisionen i en skattning av ett populationsmedelvärde från ett stickprov. För att kunna beräkna konfidensintervall måste man ange en konfidensnivå. Oftast väljs 95% eller 99%. Med en 95 procentig konfidensnivå kommer konfidensintervallet i minst 95 av 100 tänkta upprepade provtagningar att innefatta det sanna medelvärdet, och vid 99% kommer detta inträffa för 99 av 100 tänkta provtagningar. Om man har data från en (approximativ) normalfördelning beräknar man konfidensintervall med följande formel:

$\bar{x} \mp t_{\alpha/2,n-1} \times s/\sqrt{n}$,

där s är stickprovets standardavvikelse och $t_{\alpha/2,\, n-1}$ är det kritiska t-värdet för den konfidensnivå och det antal frihetsgrader man har i testet (vilket är antal prover - 1). Detta värde hämtas ur en tabell över t-fördelningen. Om man valt 95% nivån blir $\alpha = 0{,}05$.

Stickprovet i tabell 5 har en standaravvikelse på 1,48. Detta leder till att medelfelet blir 1,48/√8 = 0,524. För att beräkna ett konfidensintervall behövs $t_{\alpha/2,\, n-1}$. Om man väljer $\alpha = 0{,}05$ blir t = 2,36. Ett

95% konfidensintervall för medelvärdet blir då $0{,}125 \pm 2{,}36 \times (1{,}48/\sqrt{8}) = 0{,}125 \pm 1{,}24$. Detta betyder att om man upprepar provtagningen 100 gånger kommer i genomsnitt minst 95 av medelvärdena hamna i intervallet $0{,}125 \pm 1{,}24$, d.v.s. mellan $-1{,}11$ och $1{,}36$.

Vad skiljer standardavvikelse och medelfel, och när ska man använda det ena eller det andra? Standardavvikelse beskriver stickprovets spridning medan medelfelet är ett mått på hur väl stickprovet överensstämmer med populationen. Vilket variationsmått man väljer beror på vad man vill visa och vad det är som orsakat variationen. Om man har mycket naturlig biologisk variation är det bättre att välja standardavvikelse medan det i kontrollerade experiment är bättre att välja medelfelet.

Ofta är det dock så att medelvärden presenteras tillsammans med medelfelet eftersom det har ett lägre värde än standardavvikelsen. På detta sätt kan man, medvetet eller omedvetet, lura sig själv och läsare att variationen är lägre än vad den egentligen är.

Hypotestester - ett stickprov

Den hypotes man ofta ställer när man samlat ett enda stickprov är om stickprovet avviker från ett antaget värde. Det P-värde man erhåller från testet ger sannolikheten för att medelvärdet av slumpmässigt insamlade data från den population som ligger till grund för det antagna värdet är lika långt från det antagna populationsmedelvärdet som medelvärdet av stickprovet.

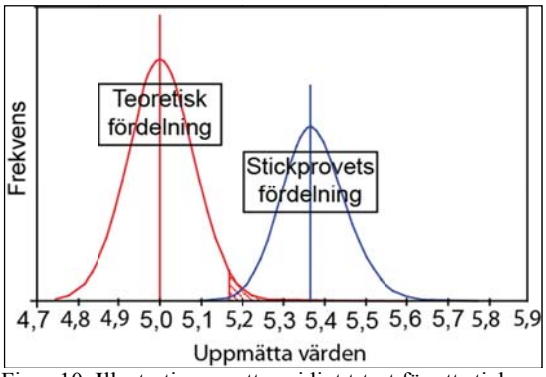
Om man har normalfördelade eller approximativt normalfördelade data kan man använda ett t-test. I de fall då stickprovet avviker grovt från normalfördelningen har man tre alternativ:

1. Utför någon form av datatransformation, t.ex. logaritmera, använd kvadratroten, invertera (= 1/värdet)
2. Om inte transformationen hjälper; använd Wilcoxons rangsummetest (se nästa avsnitt)
3. Om man har ett stort stickprov kan man använda t-test i alla fall, eftersom ett t-test vid stora stickprover är ganska robust mot avvikelser från normalitet.

t-test för ett stickprov

Ett exempel får belysa detta test. I ett försök är medelvärdet av 20 mätningar 5,36 (standardavvikelsen = 0,27). Detta är illustrerat av den högra kurvan i figur 10. Enligt biologisk litteraturen är det teoretiska medelvärdet 5,00. I ett t-test för ett stickprov utnyttjar man stickprovets variation och konstruerar med hjälp av denna variation en fördelningskurva kring det teoretiska medelvärdet, den vänstra kurvan i figur 10. Sedan testar man hur stor sannolikheten är att man drar ett lika extremt stickprov från den teoretiska fördelningen, som det stickprov man har i verkligheten. Som figuren visar är det ytterst otroligt att ett slumpmässigt urval från den vänstra kurvan skulle få den fördelning som visas av den högra kurvan. Det är dock inte omöjligt att detta skulle kunna inträffa.

Det skuggade området i den teoretiska fördelningen är 5% av hela arean. I figuren ser det ut som att kurvan slutar vid ca 5,25. I verkligheten forsätter kurvan oändligt långt åt båda sidor, men samtidigt oändligt nära noll. Om man väljer $\alpha = 0{,}05$ anser man att stickprovet är signifikant skiljt från 5,00 om stickprovets medelvärde hamnar inom den streckade delen. I detta fall ligger stickprovets medel och fördelning långt ut i ena svansen på den teoretiska fördelningen. Enligt P värdet för ett t-test på dessa data kommer stickprov från den vänstra kurvan i ett fall av ca 10000 upprepade dragningar ge ett medelvärde på 5,36, som i den högra kurvan. Vi kan således konstatera att det är ytterst osannolikt att stickprovet kommer en population som har medelvärdet 5,00.



Figur 10. Illustration av ett ensidigt t-test för ett stickprov.

Lite mer matematiskt beskrivet så beräknar man värdet på testvariabeln som

$$t = \frac{\bar{x} - \mu}{(s/\sqrt{n})},$$

där $\bar{x}$ = det medelvärde man vill prova, μ = det teoretiska medelvärdet, s = provets standardavvikelse och n är antalet prov. Antal frihetsgrader är n-1. Den kritiska nivån på t finns angiven i tabeller. Om det beräknade värdet på t är lägre än det kritiska värdet behålls nollhypotesen.

För exemplet ovan blir detta:
$t = \frac{5{,}36-5}{(0{,}27/\sqrt{20})} = \frac{0{,}36}{0{,}0604} = 5{,}96.$

Det kritiska t-värdet för $\alpha = 0{,}05$ och 19 frihetsgrader är 2,093. Eftersom detta är lägre än det beräknade värdet på t förkastar vi nollhypotesen.

Även om man får ett P-värde som är lägre än den konventionella 0,05-nivån är det inte säkert att man har ett resultat som är av vetenskapligt eller biologiskt intresse. Den skillnad man detekterat kan vara så liten att den inte har någon egentlig relevans. I vissa undersökningar kan dock avsaknaden av en skillnad vara det som är av intresse. Om så är fallet måste man istället fundera om ett P-värde straxt över 0,05 autoamtiskt innebär att det inte finns någon skillnad mellan de stickprov man testar. Innan man gör tolkningar av ett P-värde bör man således fråga sig hur stor skillnad man anser vara av vikt. För att svara på denna typ av fråga måste man använda sin biologiska expertkunskap. Det finns inga statistiska tester till hjälp eftersom frågorna är av helt annan karaktär än frågor som man besvarar med hjälp av statistik.

Viss vägledning får man dock genom att se till konfidensintervallet för stickprovet. I exemplet ovan är det 95-procentiga intervallet 5,23 - 5,48, d.v.s. om man tog 100 stickprov om vardera 20 prover skulle medelvärdet av 95 av stickproverna hamna inom intervallet. Om både den nedre och den övre gränsen är långt från det teoretiska värdet kan man anta att skillnaden är tillräckligt stor för att anses som biologiskt relevant. I fall där det motsatta gäller, att skillnaden är signifikant skild från det teoretiska värdet men att hela intervallet är nära gränsen, är skillnaden förmodligen ointressant trots att den är signifikant. I tabell 7 illustreras hur en liten skillnad kan bli signifikant om man tar tillräckligt många prover. Trots att det sista testet är signifikant skvallrar konfidensintervallet om att skillnaden förmodligen är så liten att den inte har någon egentlig betydelse.

Tabell 7. Resultat av fyra t-test på olika stora stickprover, alla med medelvärde 5,05 och standaravvikelse (s) 1,0. t-testerna provar om medelvärdena avviker från det teoretiska medelvärdet 5,00.

Antal prover	Medel	s	t-värde	P	95% Nedre	95% Övre
10	5,05	1,0	0,227	0,83	4,61	5,48
100	5,05	1,0	0,494	0,62	4,86	5,24
1000	5,05	1,0	1,682	0,0923	4,99	5,11
10000	5,05	1,0	4,991	< 0,0001	5,03	5,07

Wilcoxons rangsummetest

Ibland tillåter inte fördelningen hos de data man har att använda ett t-test för ett stickprov för att testa om medelvärdet av ett stickprov avviker från ett teoretiskt värde. I dessa fall får man istället använda Wilcoxons rangsummetest som är ett icke-parametriskt test. Förväxla inte detta test med Wilcocons teckenrangtest som beskrivs under tester för två beroende grupper.

Som alltid gäller att man måste ha ett någorlunda stort stickprov. För Wilcoxons rangsummetest är det till och med så att om man har fem eller färre prover kommer testet alltid att ge ett P-värde större än 0,05.

Eftersom Wilcoxons rangsummetest är ett test för data som inte kan beskrivas med medelvärde och standardavvikelse, utnyttjar man istället stickprovets avvikelse från ett hypotetisk medianvärde.
I Wilcoxons rangsummetest börjar man med att beräkna avståndet mellan varje observation och den hypotetiska medianen. Därefter rangordnar man alla avstånd som inte är noll. Det statistiska testet bygger på antagandet att om nollhypotesen är sann har summan av alla rangtal för positiva avstånd samma fördelning som summan av alla rangtal för negativa avstånd. Om det finns en skillnad kommer

rangsummorna däremot att skilja sig. Beroende på hur stor denna skillnad är kan man med hjälp av en tabell avgöra om stickprovet skiljer sig signifikant från den hypotetiska medianen.

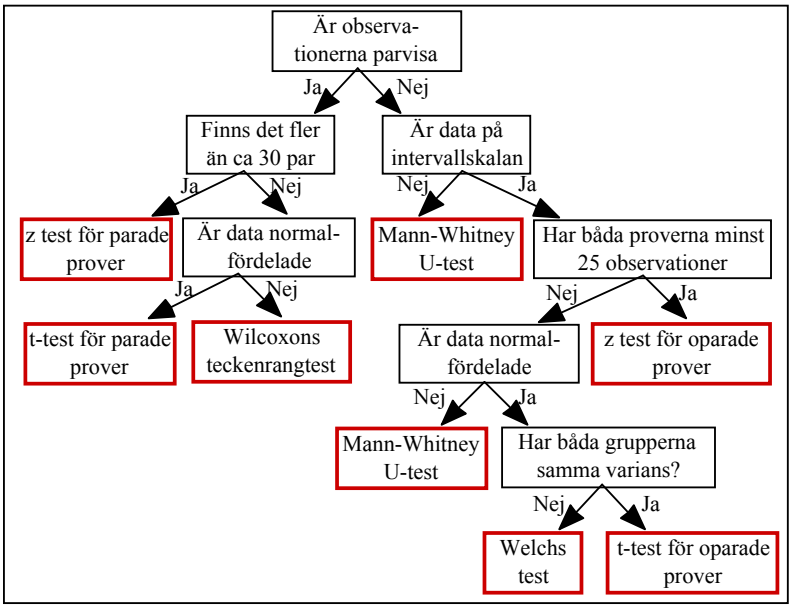
I tabell 8 visas samma data som i tabell 5, men här provas om stickprovet skiljer sig från den hypotetiska medianen 0. I kolumn 2 har skillnaden mellan varje värde och 0 beräknats och i den tredje kolumnen finns rangtalen för absolutvärdet av alla skillnader som inte är 0. Den fjärde kolumnen visar tecknet på rangtalet och därunder finns summan av alla positiva respektive negativa rangtal. Den lägsta av dessa två summor jämför man sedan med ett kritiskt värde från en tabell för Wilcoxons rangsummetest. Om α sätts till 0,05 skulle nollhypotesen i detta fall förkastas om den lägsta av rangsummorna var lika med eller mindre än 4. Vi kan således konstatera att medianen av stickprovet inte skiljer sig signifikant från det hypotetiska värdet 0, eftersom den lägsta rangsumman är 13.

Tabell 8. Exempel på Wilcoxons rangsummetest. Här testas om 8 mätvärden (X) skiljer sig från den hypotetiska medianen 0.

X	X - 0	Rang |X - 0|	Teckenrang
2	2	6,5	6,5
1,5	1,5	4,5	4,5
1	1	2	2
0	0	—	—
1	1	2	2
-2	-2	6,5	-6,5
-1,5	-1,5	4,5	-4,5
-1	-1	2	-2
Rangsumma; positiva avstånd: 15			
Rangsumma; negativa avstånd: 13			

Hypotesprövning - Två grupper

När man har två grupper av data kan de antingen bestå av par eller av två enskilda grupper. Om man kan ordna sin data parvis säger man att man har beroende data. Ofta är denna typ av data mätningar på samma objekt före och efter någon form av behandling. För den andra typen kan man inte säga att ett värde i de ena gruppen hör ihop med ett visst värde i den andra gruppen. I figur 11 finns ett schema som hjälper till att välja rätt statistiskt test när man vill testa hypoteser kring två grupper av mätdata. Om man har data som bygger på antal eller proportioner ska man istället använda s.k. kontingenstabeller som beskrivs senare.

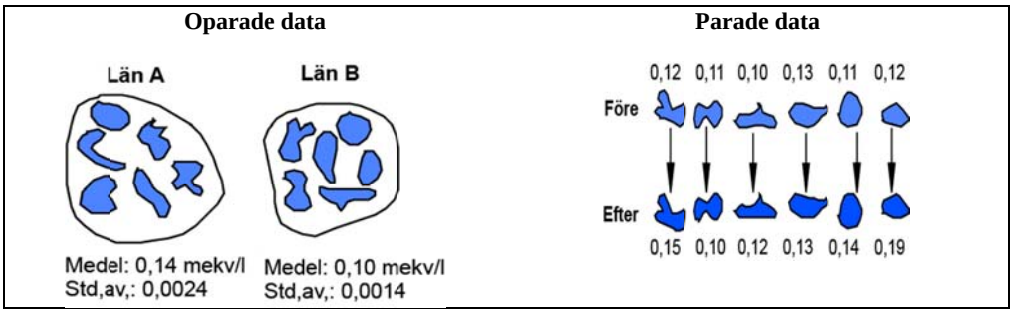


Figur 11. Flödesschema för att välja rätt statistiskt test för att prova om två grupper av mätvärden skiljer sig från varandra. Datorprogram skiljer sällan på z- och t-tester, utan kallar allt t-test.

Parat eller oparat test

Hela detta avsnitt behandlar metoder för att testa om två stickprov skiljer sig från varandra. När man drar två stickprov kan stickproven antingen vara oberoende eller beroende av varandra (figur 12). Oberoende, eller **oparade**, data får man t.ex. när man jämför exempelvis alkalinitet i 20 slumpvis valda sjöar i ett län med alkaliniteten i 20 slumpvis valda sjöar i ett annat län. Beroende, eller **parade**, data har man om man gör upprepade mätningar på samma objekt, ofta före och efter en behandling. Ett exempel kan vara alkaliniteten i 20 slumpvis valda sjöar före och efter kalknig. Man kan i detta fall para ihop mätningen före kalkningen i sjö 1 men mätningen efter kalkningen i sjö 1, och detsamma för de övriga 19 sjöarna.

Om man kan säga att ett värde i den ena gruppen hör ihop med ett visst värde i den andra gruppen har man beroende grupper och kan utföra ett parat test. Den faktor som gör att man kan para ihop värden måste dock vara en annan än den faktor man studerar. Generellt gäller att ett parat test har större styrka än ett oparat test eftersom man tillför mer information till testet genom att para ihop mätvärden två och två. Om man har möjlighet ska man därför redan vid planeringen av en studie sträva efter att lägga upp studien så att man får parade observationer.



Figur 12. Illustration av oparade och parade data. I den vänstra illustrationen kan inte en sjö i län A paras ihop med någon sjö i län B. I den högra illustrationen har man mätningar på samma sjö före och efter en behandling och har därmed par av data. Alla siffror anger alkalinitet i mekv/l.

Oparade tester

Förutsättningar

I ett oparat t-test förutsätts att båda de studerade grupperna kommer från samma bakomliggande population och därigenom har samma varians. Detta kan testas genom att först beräkna variansen för var och ett av stickproven. Därefter beräknar man en s.k. F-kvot genom att dividera den största av varianserna med den lägsta. Sedan kan man med hjälp av en tabell över F-fördelningen jämföra den erhållna kvoten med det kritiska värdet på F. Antalet frihetsgrader är lika med antalet prover minus 1, både för täljaren och nämnaren. Om den beräknade F-kvoten är högre än det kritiska värdet förkastas nollhypotesen att de två proven har samma varians. Om detta test visar att variansen är olika kan man använda Welchs test istället för ett t-test.

Det är dock mycket ovanligt att tillämpa Welchs test. Många förordar att man bör undvika det. Om man finner att man har olika varians kan detta vara ett resultat i sig, oavsett om medelvärdena skiljer sig eller ej. Om så är fallet kan man låta bli att göra ett efterföljande Welch-test. Ett annat sätt att undvika Welchs test är att transformera sina data och därefter testa varianserna igen. Ofta kan en transformering leda till att proverna får samma varians och då kan ett t-test tillämpas.

Oparat z- och t-test

De flesta datorprogram skiljer inte på de z- och t-tester som finns med i figur 11, utan kallar allt för t-test. I dessa fall använder man datorprogrammes t-test till alla stickprovsstorlekar.

Om man har fler än ca 25 till 30 observationer kan man genom centrala gränsvärdessatsen (se tidigare avsnitt) använda tester som bygger på normalfördelningen, s.k. z-tester. Har man färre observationer använder man t-tester. Det finns några skillnader mellan z-test och t-test. Den viktigaste är sättet att beräkna medelfelet.

I oparade z- och t-tester beräknas testvariabeln enligt:

$$z \text{ eller } t = \frac{\bar{x}_A - \bar{x}_B}{SE_{A-B}},$$

där $\bar{x}_A$ och $\bar{x}_B$ är medelvärdet av grupp A respektive grupp B och SE_{A-B} är medelfelet för skillnaden mellan medelvärdena av A och B. Om man har många prover är skillnaden mellan gruppernas medelvärden normalfördelad och då använder man z-testet. I annat fall gäller t-test.

I ett z-test beräknas medelfelet för skillnaden i mellan grupp A och grupp B enligt:

$$SE_{A-B} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}},$$

där s^2 är variansen och n är antal observationer för respektive grupp. Genom att kombinera de två formlerna ovan erhåller man ett värde på z. Detta värde jämförs med kritiska värden i en tabell över normalfördelningen.

I de fall man har färre än ca 30 observationer beräknar man standardavvikelsen för skillnaden i mellan grupp A och grupp B enligt:

$SE_{A-B} = \sqrt{\frac{s_{AB}^2}{n_A} + \frac{s_{AB}^2}{n_B}}$, där $s_{AB}^2 = \frac{(n_A-1)\times s_A^2 + (n_B-1)\times s_B^2}{(n_A-1)+(n_A-1)}$, där s_A^2 och s_B^2 är de olika provens varianser.

Värdet på t erhålls genom att sätta in SE_{A-B} i formeln ovan. Det kritiska värdet erhålls ur en t-tabell. Antalet frihetsgrader är n-2. Om det beräknade värdet på t är högre än det kritiska värdet i tabellen kan nollhypotesen förkastas.

Tolkning av resultat

Precis som för t-test för ett stickprov är det inte säkert att man har en biologisk intressant skillnad, även om skillnaden är signifikant. Precis som för t-test för ett stickprov kan man studera konfidensintervallen för att undersöka om den erhållna skillnaden är så pass stor att man ska betrakta den som relevant för den frågeställning man har. I tabell 7 illustrerades hur chansen att hitta en liten skillnad ökar med ökande antal prover, d.v.s. om man har tillräckligt många prover kan man hitta skillnader som är så små att de inte har någon praktisk betydelse. Om P-värdet däremot är litet och båda gränserna i konfidensintervallet indikerar en skillnad som har en reell betydelse kan man dra slutsatsen att de båda stickproven skiljer sig.

Mann-Whitneys test

I de fall man har data på ordinalskalnivå eller från fördelningar som inte ens efter transformering kan anses vara normalfördelade använder man Mann-Whitneys icke-parametriska test (kallas ibland även för Wilcoxons test, ej att förväxla med Wilcoxons teckenrangtest som beskrivs senare). I detta test jämför man de två gruppernas medianer istället för medelvärden som i t-tester. I detta test rankas alla observationer från lägst till högst, oavsett vilken grupp de tillhör (tabell 9). Om flera observationer har samma värde ges dessa sin medelrangordning. Efter att alla värden har rankats summeras rangtalen i varje grupp. Om skillnaden mellan dessa summor är stor indikerar detta att de båda grupperna skiljer sig åt.

Tabell 9. Illustration av Mann-Whitneys test. Rangsumman för grupp A är 28,5 och 49,5 för grupp B.

Grupp	A	A	A	A	A	A	B	B	B	B	B	B
Värde	1	2	1	2	1,5	1,3	2,1	2,3	1,5	1	2,1	3
Rang	2	7,5	2	7,5	5,5	4	9,5	11	5,5	2	9,5	12

Efter att man summerat rangtalen beräknas testvariabeln som kallas Mann-Whitneys U. Denna är den minsta av U_A och U_B, som beräknas enligt:

$U_A = n_A \times n_B + \frac{n_A(n_A+1)}{2} - \sum R_A$, respektive $U_B = n_B \times n_B + \frac{n_B(n_B+1)}{2} - \sum R_B$,

där n är antal observationer och R är rangsumman.

Via tabell får man reda på det kritiska värdet på U för de stickprovsstorlekar man har. I detta test blir det beräknade värdet på U lägre ju större skillnader det är mellan gruppernas medianer. Därför förkastas nollhypotesen om det beräknade värdet på U är lägre än det kritiska värdet.

I exemplet i tabell 9 är U_A = 28,5 och U_B = 7,5. Det kritiska värdet för Mann-Whitneys U för α = 0,05 och sex observationer i varje grupp är 5. Nollhypotesen kan således inte förkastas eftersom U_B är större än det kritiska värdet.

Ett lågt P-värde indikerar att de två populationerna som jämförs har olika medianer. Ett P-värde högre än den kritiska 0,05 gränsen säger att man inte kan dra slutsatsen att populationerna skiljer sig. Observera att detta inte är detsamma som att säga att de är lika! (detta gäller all statistisk hypotesprövning). Om man har små stickprov är Mann-Whitneys test olämpligt eftersom det i dessa fall har mycket låg styrka. Det är t.o.m. så att om man totalt har sju eller färre prover kommer testet alltid att ge P-värden högre än 0,05.

Parade tester

I de fall man har beroende, eller parade, observationer utför man det statistiska testet på differensen mellan paren. Som framgår av figur 11 finns även här en teoretisk skillnad mellan z- och t-test, men de flesta datorprogram tar inte hänsyn till denna skillnad utan kallar allt för t-test.

Parat t-test

I tabell 10 finns samma data som i tabell 9, men här ordnade som par och grupperna A och B har fått byta namn till ”Före” och ”Efter”. Den nollhypotes som prövas är att den behandling som utförts mellan mätningarna inte har givit någon effekt, d.v.s. att medelskillnaden mellan Före och Efter ska var noll.

H_0: Behandlingen har ingen effekt och medelskillnaden mellan ”Före” och ”Efter” är lika med 0.
H_1: Behandlingen har effekt och medelskillnaden mellan ”Före” och ”Efter” är skild från 0.

Tabell 10. Data för att illustrera ett parat t-test.

Prov nr.	Mätvärde		Skillnad
	Före	Efter	
1	1	2,1	-1,1
2	2	2,3	-0,3
3	1	1,5	-0,5
4	2	1	1
5	1,5	2,1	-0,6
6	1,3	3	-1,7
		Medel	-0,53
		Std.av.	0,91

Eftersom man testar om en observerad medelskillnad skiljer sig från en antagen medelskillnad är ett parat t-test detsamma som ett t-test för ett stickprov där den observerade medelskillnaden testas mot det hypotetiska värdet 0 (ingen skillnad). Från avsnittet om t-test för ett stickprov får vi:

$$t = \frac{\bar{x} - \mu}{\left(s/\sqrt{n}\right)} = \frac{-0{,}53 - 0}{\left(0{,}91/\sqrt{6}\right)} = \frac{-0{,}53}{0{,}37} = -1{,}4$$

Med $\alpha = 0{,}05$ och 5 frihetsgrader blir det kritiska värdet på t = 2,57. Då absolutvärdet av det beräknade t är lägre än 2,57 kan nollhypotesen inte förkastas, d.v.s. den behandling som utfördes mellan ”Före” och ”Efter” hade ingen effekt på den variabel som mättes.

Precis som tidigare gäller att även om nollhypotesen förkastas är det inte säkert att den upptäckta skillnaden har någon betydelse i praktiken. Ett hjälpmedel för att se hur stor skillnad man har är att beräkna konfidensintervallen. Hur man använder denna information framgår i avsnittet t-test för ett stickprov.

Wilcoxons teckenrangtest

I de fall man har data på ordinalskalnivå eller från fördelningar som inte ens efter transformering kan anses vara normalfördelade använder man Wilcoxons teckenrangtest som är ett icke-parametriskt test för två beroende, eller parade, grupper. Eftersom detta är ett ickeparametriskt test är det medianerna som jämförs. I detta test ger P-värdet svar på frågan om skillnaden mellan medianen före och efter behandling är en slumpvariabel eller ej. Testet tar hänsyn till både storlek och tecken på skillnaden mellan ”Före” och ”Efter”, och är således ett ganska starkt test trots att det är ett icke-parametriskt test.

Testet går till så att skillnaden mellan alla par beräknas. Därefter rangordnas absolutvärdet av alla skillnader. Skillnader som är 0 stryks, därefter får den lägsta skillnaden rangordning 1. Om flera skillnader är lika får dessa sitt medelrangtal. Därefter sätter man ut plus- och minustecken framför rangtalen och beräknar man summan. Testvariabeln beräknar man sedan enligt:

$$T = \frac{\Sigma R}{\sqrt{\Sigma R^2}},$$

där R är rangtalen med tecken. Det kritiska värdet för denna testvariabel får man ur en tabell över normalfördelningen.

För att illustrera detta använder vi samma data som i exemplet i tabell 10 och lägger till rangtal för skillnaden (tabell 11). Värdet på testvariabeln blir T = -15/√91 = -1,57. Från en tabell över normalfördelningen får man att det kritiska värdet för ett tvåsidigt test vid $\alpha = 0,05$ är -1,96. Då värdet på testvariabeln understiger det kritiska värdet behåller vi nollhypotesen, precis som i t-testet med samma data.

Tabell 11. Data för att illustrera Wilcoxons teckenrangtest.

Prov nr.	Mätvärde		Skillnad	Rang	Tecken-rang (R)	
	Före	Efter				
1	1	2,1	-1,1	2	-2	
2	2	2,3	-0,3	6	-6	
3	1	1,5	-0,5	5	-5	
4	2	1	1	3	3	
5	1,5	2,1	-0,6	4	-4	
6	1,3	3	-1,7	1	-1	
					Σ R	-15
					ΣR^2	91

Som för andra icke-parametriska tester har även detta test lägre styrka än motsvarande parametriska test. Vid små stickprovsstorlekar är det mycket svår att kunna upptäcka annat än mycket stora skillnader när man använder Wilcoxons teckenrangtest.

Hypotesprövning - Fler än två grupper

Hittills har vi gått igenom hur man gör statistiska tester på en och två grupper. I detta avsnitt kommer vi gå igenom tester för att prova hypoteser som involverar fler än två grupper. Om man har normalfördelade data går dessa typer av analyser ofta under samlingsnamnet variansanalys, eller ANOVA efter engelskans ANalysis Of VAriance. Namnet variansanalys är egentligen ganska missvisande eftersom man inte testar hypoteser som rör varianser. I själva verket är det hypoteser om medelvärden som testas, men av historiska skäl används namnet variansanalys.

Grupperna i en variansanalys kan vara ordnade på en mängd olika sätt beroende på vilket försök man utför och vad det är man studerar. I det enklaste fallet har man tre oberoende grupper av data där man testar om medelvärdet hos någon av grupperna skiljer sig från medelvärdet hos någon av de andra grupperna. Från detta enkla fall kan man bygga ut analyserna till ytterst komplicerade modeller. I denna handledning kommer några av de viktigaste aspekterna att tas upp. För mer avancerade modeller finns flera hundra statiskböcker om variansanalys, på mer eller mindre abstrakt nivå, att tillgå. En god start är Zar (1999).

ANOVA - grundläggande begrepp

I en variansanalys jobbar man med två grupper av variabler. Dessa grupper har många benämningar. En vanlig nomenklatur är beroende respektive oberoende variabler. Vanligt är också att de beroende variablerna kallas effektvariabler och de oberoende kallas faktorer. Olika nivåer i den oberoende variabeln kallas nivåer.

Ett exempel får belysa begreppen. Antag att man studerar effekten av olika mängder av ett gödningsmedel på växters tillväxt. Man planterar fem plantor vardera i tre miniväxthus och ger olika mängd av ett gödningsmedel till de olika växthusen. Efter två veckor mäter man tillväxten hos alla plantor. Därefter gör man om experimentet två gånger. Resultaten finns sammanställda i tabell 12.

I detta experiment är tillväxten den **beroende** variabeln, eller **effekt**variabeln. Näringsgiva och försöksomgång är **oberoende** variabler, eller **faktorer**. Näringsgiva har i detta exempel tre **nivåer**; 15 ml, 20 ml och 25 ml. De numeriska värdena på näringsgiva kommer dock inte att ingå i analysen utan är bara namn på de olika nivåerna.

Tabell 12 Tillväxt (mm) hos plantor efter 2 veckor vid odling i olika näringsgivor. Beskrivningar och formler för de olika beräkningarna finns nedan.

	Försök 1			Försök 2			Försök 3		
	Näringsgiva (ml)			Näringsgiva (ml)			Näringsgiva (ml)		
	15	**20**	**25**	**15**	**20**	**25**	**15**	**20**	**25**
	118	128	138	100	110	120	139	140	119
	119	129	139	110	120	130	122	142	118
	120	130	140	120	130	140	141	138	130
	121	131	141	130	140	150	120	128	132
	122	132	142	140	150	160	121	131	129
Summa	600	650	700	600	650	700	643	679	628
Summa av kvadraterna	72010	84510	98010	73000	85500	99000	83127	92353	79050
Kvadratsumma	10	10	10	1000	1000	1000	437,2	144,8	173,2
Antal obs.	5	5	5	5	5	5	5	5	5
Medel	120	130	140	120	130	140	129	136	126
Std.av.	1,58	1,58	1,58	15,8	15,8	15,8	10,5	6,02	6,58

I de två första försöksserierna i tabell 12 framgår det tydligt att tillväxten är högre vid högre näringsgiva och att variationen kring medelvärdena var densamma inom varje försök. Resultaten i den tredje upprepningen är svårare att tolka efter bara en snabb överblick.

Man kan tycka att ovanstående problem skulle kunna lösas med hjälp av tre t-tester, men detta är direkt olämpligt. Det olämpliga består i slumpens obehagliga nycker. Om man utför upprepade t-tester på samma dataset kommer risken att göra ett typ-I fel vara 1 på 20 för det första testet (vid $\alpha = 0{,}05$) och sedan öka för varje ytterligare test man utför på samma data, se även avsnittet om upprepade tester och Bonferroni-korrektion ovan. Om man i fallet med de tre näringsgivorna gör tre upprepade t-tester är risken för ett typ -I fel 13%. Hade man haft fem nivåer på näringsgivorna vore samma risk 29%. För att undkomma dessa ackumulerade risker för typ-I fel gör man istället en variansanalys.

Då man satte igång detta försök hade man den statistiska nollhypotesen att medeltillväxten är densamma vid alla näringsgivor. Mothypotesen är att medeltillväxten vid minst en av näringsgivorna skiljer sig från medeltillväxten vid minst en annan näringsgiva. Detta kan sammanfattas som:

H_0: $\bar{x}_{15} = \bar{x}_{20} = \bar{x}_{25}$

Mothypotesen omfattar många olika alternativ:

H_1: $\bar{x}_{15} < \bar{x}_{20} = \bar{x}_{25}$
$\bar{x}_{15} = \bar{x}_{20} < \bar{x}_{25}$
$\bar{x}_{15} < \bar{x}_{20} < \bar{x}_{25}$
$\bar{x}_{15} > \bar{x}_{20} = \bar{x}_{25}$
$\bar{x}_{15} > \bar{x}_{20} > \bar{x}_{25}$
och så vidare

Om nollhypotesen förkastas gäller minst en av alla mothypoteser. Vilken eller vilka som gäller får man svar på genom s.k. *post hoc* tester som vi kommer att beskriva längre fram.

För att testa nollhypotesen måste man ta reda på stickprovens medelvärde och spridningen kring medelvärdet. Vidare måste man ta reda på variationen i hela datamaterialet. Den totala variationen kan sedan delas upp i variation **mellan** grupper och variation **inom** grupper. I en variansanalys vill man ta reda på om variationen inom en grupp är mindre än variationen mellan grupperna. D.v.s. hur mycket av den totala variationen beror på olika behandlingar (näringsgivor i exemplet), och hur mycket av variationen är den naturliga variationen inom respektive grupp? Detta uttrycks genom kvoten mellan ”variationen mellan grupper” och ”den totala variationen”. Denna kvot kallas F-kvot och är ett centralt begrepp inom all variansanalys. Ju större F-kvoten är desto mer av variationen beror på olikheter mellan de olika behandlingarna. Benämningen F kommer från Sir Ronald Fisher.

Exempel på beräkningar - ANOVA-tabellen

En grundläggande del i all presentation av resultat från variansanalyser är den s.k. ANOVA-tabellen. I alla datorprogram för statistik presenteras resultaten i en tabell som kort och gott kallas ANOVA-tabell. I detta avsnitt kommer vi att gå igenom de olika komponenterna i en sådan tabell. Vi utgår från försöken i tabell 12. För att konstruera tabellen behöver vi de olika summorna som finns under själva datasammanställningen.

För varje nivå (näringsgivor i detta fall) beräknas följande:

Benämning	**Förklaring**	**Exempel, Försök 1 vid 15 ml.**
Summa	Summan av alla data i en grupp, Σx_i.	118+119+120+121+122 = 600
Summa av kvadraterna	Summan av varje mätvärde i kvadrat, i en grupp, $\sum x_i^2$.	$118^2 + 119^2 + 120^2 + 121^2 + 122^2 =$ 72010
Medel för grupp *j*	Medelvärde, $\bar{x}_j = \sum x_i / n_i$.	600/5 = 120
Kvadratsumma, SS	Summan av kvadraten på skillnaden mellan mätvärde och medelvärdet $\sum(x_i - \bar{x}_j)$.	$(118\text{-}120)^2 + (119\text{-}120)^2 +$ $(120\text{-}120)^2 + (121\text{-}120)^2 +$ $(122\text{-}120)^2 = 10$
Antal obs.	Antal observationer, n_i.	5
Std.av.	Standardavvikelse, s_i.	1,58

Till detta kommer ett antal termer som är gemensamma för hela försök 1:

Totalsumma	Summan av alla data i ett försök, ΣX.	1950
Summa av alla kvadrater	Summan av varje mätvärde i kvadrat, i ett försök, ΣX^2.	254530
Antal obs.	Antal observationer, N.	15
Antal grupper	De olika näringsgivorna, k	3

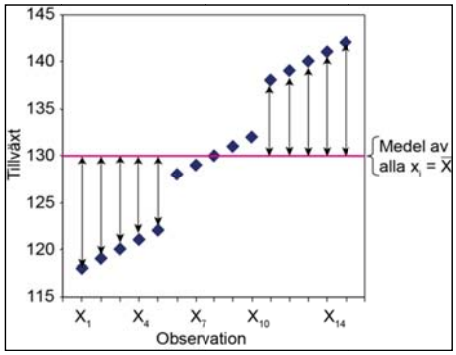
För att beräkna den s.k. F-kvoten behöver vi beräkna de olika varianskomponenterna. Dessa förkortas ofta SS vilket står för Engelskans ”Sums of Squares”. Nedan följer en beskrivning av de tre olika varianskomponenterna i den enklaste formen av variansanalys, en s.k. enfaktors-anova.

1) **Totala** kvadratsumman, SS_T. Detta är ett mått på avståndet mellan varje punkt och det gemensamma medelvärdet (figur 13). Detta beräknas med följande formel:

$SS_T = \Sigma(x_i - \bar{X})^2$.

Denna formel används inom många statistiska tester och vi har hittills stött på den vid beräkning av variansen för ett stickprov. För att göra beräkningarna lättare kan man förenkla formeln till:

$SS_T = \Sigma X^2 - (\Sigma X)^2/N = 254530 - (1950^2/15) = 1030$.



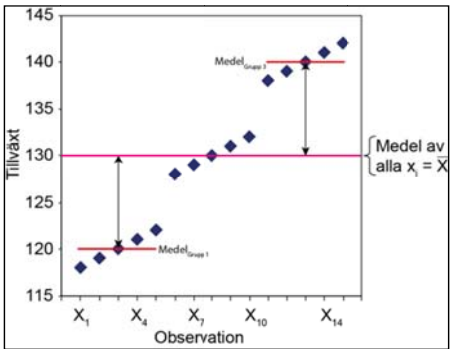
Figur 13. Illustration av de avstånd som ingår i beräkningen av **totala** kvadratsumman i en variansanalys. Det markerade avståndet motsvarar $x_i - \bar{x}$ i ekvationen ovan. Data kommer från försök 1 i tabell 12.

2) Kvadratsumman för variation **mellan** grupper, SS_M. Denna kvadratsumma ger ett mått på variationen mellan stickproven. Detta är ett mått på hur mycket stickprovsmedelvärdena ($\bar{x}_j$) varierar kring sitt medelvärde, d.v.s. ett mått på skillnaden mellan en grupps medelvärde och alla gruppernas gemensamma medelvärde (figur 14). Om alla grupper kommer från samma population kommer denna kvadratsumma, d.v.s. avstånden i figur 14, att vara liten. Om det däremot finns stora skillnader mellan grupperna antar summan ett stort värde. Detta beräknas enligt:

$$SS_M = \sum n_j(\bar{x}_j - \bar{X}_{medel})^2,$$

där n_j = antal observationer i grupp j, $\bar{x}_j$ = medelvärde av grupp j, och
$\bar{X}_{medel} = \frac{\Sigma n_j \times \bar{x}_j}{N}$,
d.v.s. ett viktat medelvärde av alla gruppers medelvärde. Viktningen består i att man tar hänsyn till antalet observationer i varje grupp. Genom detta får medelvärden från grupper med många observationer väga tyngre än medelvärden från grupper med få observationer.

Formeln för beräkning av SS_M kan förenklas till:
$SS_M = \Sigma(\Sigma x_i^2/n_i) - (\Sigma X)^2/N = (600^2/5 + 650^2/5 + 700^2/5) - (1950^2/15) = 1000.$



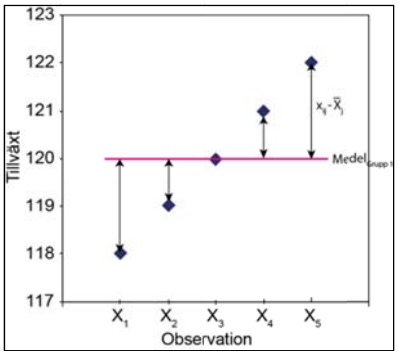
Figur 14. Illustration av de avstånd som ingår i beräkningen av kvadratsumman för variation **mellan** grupper i en variansanalys. De markerade avstånden motsvarar $\bar{x}_j - \bar{X}_{medel}$ i ekvationen ovan. Data kommer från försök 1 i tabell 12.

3) Kvadratsumma för variation **inom** grupper. Detta är ett mått på variationen inom de olika grupperna. Denna varianskomponent benämnes ofta ”Error” i statistikprogram. Formeln för detta är:

$$SS_i = \sum_i \sum_j (x_{ij} - \bar{x}_j)^2$$

där x_{ij} är mätvärdet för observation i i grupp j, och $\bar{x}_j$ är medelvärdet för grupp j. Med denna formel beräknas först kvadraten på summan av avstånden mellan varje observation i en grupp och gruppens medelvärde (figur 15). Detta görs för varje grupp och därefter summeras alla gruppsummor. Som framfår av figur 15 ger stora variationer (d.v.s. då x_{ij} - $\bar{x}_j$ är stort) ett högt värde, och liten variation ett lågt värde. En enklare variant av samma formel är:

$SS_I = \Sigma X^2 - \Sigma(\Sigma x_i^2/n_i) = 254530 - (600^2/5 + 650^2/5 + 700^2/5) = 30$



Figur 15. Del av figur 14 för illustration av de avstånd som ingår i beräkningen av kvadratsumman för variation **inom** grupper i en variansanalys. De markerade avstånden motsvarar x_{ij} - $\bar{x}_i$. i formeln ovan. Data kommer från försök 1 i tabell 12.

När de olika varianskomponenterna är beräknade presenteras de vanligtvis i en ANOVA-tabell (tabell 13). I tabellen ingår även andra komponenter än kvadratsummorna. Antal frihetsgrader är ett mått på

den mängd information som finns kvar i data efter att en parameter är beräknad. Vi går inte in närmare på detta här utan nöjer oss med att presentera formlerna i tabellen.

Tabell 13. De olika delarna i en ANOVA-tabell. I den övre tabellen visas formlerna för de olika komponenterna och i den undre visas värden från data från försök 1 i tabell 12.

Variansorsak	**Frihets-grader, f.g.**	**Kvadrat-summa, SS**	**Medelkvadrat-summa, MS**	**F**
mellan grupper	k-1	SS_M	SS_M/f.g.	MS_M/MS_I
inom grupper	N-k	SS_I	SS_I/f.g.	
Totalt	$(k-1)+(N-k) = N$-1	SS_T		

Variansorsak	**Frihets-grader, f.g.**	**Kvadrat-summa, SS**	**Medelkvadrat-summa, MS**	**F**	**p**
mellan grupper	3-1 = 2	1000	1000/2 = 500	500/2,5= 200	<0,001
inom grupper	15 - 3 = 12	30	30/12 = 2,5		
Totalt	15-1=14	1030			

Medelkvadratsumman är ett mått på den genomsnittliga variansen i de olika grupperna. Dessa värden används sedan för att beräkna den s.k. F-kvoten. En generell beskrivning av F-kvoten är:

$$F = \frac{\text{behandlingseffekt + individuella skillnader + mätfel}}{\text{individuella skillnader + mätfel}}.$$

Om behandlingseffekten (effekten av de olika näringsgivorna i detta fall) är liten blir kvoten nära 1 och man kommer inte att kunna hitta någon signifikant effekt av den behandling man utfört.

I praktiken beräknar man F-kvoten som $F = MS_{Mellan}/MS_{Inom}$. För att få reda på om behandlingen har någon signifikant effekt ser man i en tabell över F-fördelningen och läser av det kritiska värdet för de antal frihetsgrader som ingår i täljare och nämnare i F-kvoten, för den valda nivån på α. I exemplet ovan är det kritiska värdet vid $\alpha = 0{,}01$ och 2 respektive 12 frihetsgrader 5,39. Det beräknade värdet på F är 200 och vi kan därför på goda grunder förkasta nollhypotesen som sa att de olika medelvärdena inte skiljer sig åt.

Ett vanligt sätt att i löpande text presentera resultat av en variansanalys är ”det fanns en signifikant skillnad i tillväxt mellan de olika näringsgivorna (ANOVA, $F_{2, 12} = 200$; $P < 0{,}001$)”, eller varianter av detta.

I tabell 14 presenteras resultaten av variansanalyser på försök 2 och 3 i tabell 12, efter beräkning i ett datorprogram. Som framgår är ingen av dem signifikanta (d.v.s. P är större än 0,05). I försök 2 finns en stor variation både inom och mellan grupper som gör att man inte vet vad som är naturlig variation och vad som är effekt av de olika näringsgivorna. I försök 3 är den totala variationen densamma som i försök 1, men variationen inom grupperna är avsevärt mycke större. Därav går det inte att särskilja variation inom grupper från variation mellan grupper och testet ger stöd åt nollhypotesen.

Som du kanske märkt benämnes variation inom grupper ”Error” i det datorprogram som används för beräkningarna i tabell 14. Detta är en vanlig synonym till variation inom grupper. En annan vanlig synonym är ”Random”. Båda dessa synonymer indikerar att denna varianskomponent är ett mått på den slumpmässiga variationen inom en grupp.

Tabell 14a. Resultat av ANOVA på försök 2 i tabell 12.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Nutrient	2	1000	500	2,00	0,1780
Error	12	3000	250		
Total	14	4000			

Tabell 14b. Resultat av ANOVA på försök 3 i tabell 12.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Nutrient	2	274,8	137,4	2,183	0,1554
Error	12	755,2	62,93		
Total	14	1030			

Bakomliggande modeller

I detta avsnitt kommer vi att kort ta upp de bakomliggande modellerna i variansanalys. Den totala variationen består av summan av variationen i de olika grupperna som ingår i en modell, de s.k. varianskomponenterna. Variationen beskrivs av de olika kvadratsummorna. Vi får således:

$SS_{Total} = SS_{Inom} + SS_{Mellan}$.

Detta illustreras bl.a. i tabell 13 där totala kvadratsumman, 1030, är summan av gruppernas kvadratsummor som är 1000 och 30.

För tvåvägs variansanalyser (som beskrivs längre fram) består kvadratsumman för variation mellan grupper av flera delar:

1. en kvadratsumma som representerar variationen mellan prover, som orsakas av faktor A, SS_A,
2. en kvadratsumma som representerar variationen mellan prover, som orsakas av faktor B, SS_B och
3. en kvadratsumma som representerar variationen mellan prover, som orsakas av samverkan (interaktion) av faktor A och faktor B, SS_{AB}.

Vi får då:

$SS_{Total} = SS_{Inom} + SS_A + SS_B + SS_{AB}$.

Modellerna ovan är beskrivningar av hur variansen delas upp i olika komponenter. Det finns också modeller för hur varje enskild observation inkluderas i en variansanalys. I en enfaktors variansanalys har man följande komponenter (med exempel från försök 1 i tabell 12):

Y_{ij} =observation *i* hos nivå *j*, t.ex. 118 mm för planta 1 vid näringsgivan 15 ml,

μ = medeltillväxten hos hela populationen av den växt man använt, oavsett behandling,

α_j = avvikelsen från μ som en effekt av behandling *j*, kan vara både positiv och negativ; d.v.s. den tillväxt som erhålls som en effekt en näringsgiva,

ε_{ji} = ett mått på slumvariationen; d.v.s. ett mått på hur mycket planta *i* behandlad med näringsgiva *j* avviker från $\mu + \alpha_i$ på grund av slumpmässig variation.

Dessa komponenter bildar tillsammans följande ekvation:

$Y_{ji} = \mu + \alpha_j + \varepsilon_{ji}$.

I en variansanalys testar man om α_i har något inflytande i ekvationen. Om man inte har någon effekt av den behandling som utförts kommer α vara ungefär lika för alla *i*. D.v.s. att $\alpha_{15} \approx \alpha_{20} \approx \alpha_{25}$ i försök 1. Detta är i själva verket den egentliga nollhypotesen.

För en tvåfaktors ANOVA får man på motsvarande sätt:

$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$,

där β_j = effekten av nivå *j* i faktor β och $(\alpha\beta)_{ij}$ är samverkan mellan nivå *i* hos faktor α och nivå *j* hos faktor β.

Ovanstående ekvationer är exempel på den teoretiska grunden i all variansanalys och de beskriver sambandet mellan varje observation och effekten av de olika behandlingarna. I denna korta beskrivning av de bakomliggande modellerna i ANOVA tar vi bara upp de allra enklaste varianterna av dessa modeller. Ju mer komplicerad modell som testas desto mer komplicerad blir den ekvation som beskriver modellen.

Fixa, stokastiska och blandade modeller

De tre begreppen i rubriken är viktiga att känna till innan man sätter igång med en variansanalys eftersom sättet att beräkna F-kvoten skiljer sig beroende på vilken typ av modell man har. Om man inte specificerar vilken typ av modell man har kan man få felaktiga F-kvoter (och därmed felaktiga P-värden) när man använder datorprogram, eftersom datorn då inte vet hur den ska räkna utan använder den förvalda metoden.

Det finns två typer av faktorer, d.v.s. oberoende variabler; **fixa** och **stokastiska**. De flesta variansanalyser bygger på fixa faktorer. Denna typ av faktorer har fastställda nivåer i den oberoende variabeln, t.ex. de tre näringsgivorna i exemplet ovan. Nivåerna bestäms oftast innan ett experiment startar.

Stokastiska faktorer har nivåer som väljs slumpvis från en ”population” av tänkbara nivåer. Detta förfarande är inte ovanligt i miljöanalyssammanhang där man ibland inte kan designa experiment och i förväg bestämma nivåerna på sina faktorer. Ett exempel på en stokastisk faktor är trädslag förutsatt att man utför mätningar på slumpvis valda träd i ett område. Om man innan studien hade definierat vilka trädslag man skulle inventera vore trädslag däremot en fix effekt.

Beroende på kombination av fixa och stokastiska faktorer finns det olika sätt att beräkna varianskomponenterna och F-kvoten. Oftast är det nämnaren i F-kvoten som skiljer mellan de olika modellerna. Tabell 15 ger en bild av hur F-kvoten beräknas vid olika typer av data. För att få rätt F-kvot när man använder datorprogram för att beräkna en variansanalys måste man ange om man har stokastiska eller fixa faktorer. De olika modellerna benämnes Modell I, II och III. De engelska termerna för dessa tre modeller är antingen ”Model I, II, III” eller ”Fixed models”, ”Random models” och ”Mixed models”.

Modell I är den vanligaste typen och används när alla faktorer är fixa. De flesta datorprogram ger F-kvoter beräknade enligt modell I, om man inte specificerar annat.

Modell II används sällan och bygger på att alla faktorer har slumpmässigt valda nivåer.

Modell III används när man har en blandning av fixa och slumpmässiga faktorer och kallas ofta ”mixed model ANOVA”.

Tabell 15. Illustration av hur sättet att beräkna F-kvoten i en variansanalys skiljer sig mellan olika typer av studier, beroende på om faktorerna är fixa eller stokastiska. Error betecknar inomgruppsvariationen.

Variansorsak	**Modell I** (A och B är fixa)	**Modell II** (A och B är stokastiska)	**Modell III** (A är fix och B är stokastisk)
Faktor A	$\frac{\text{Faktor A MS}}{\text{Error MS}}$	$\frac{\text{Faktor A MS}}{\text{A} \times \text{B MS}}$	$\frac{\text{Faktor A MS}}{\text{A} \times \text{B MS}}$
Faktor B	$\frac{\text{Faktor B MS}}{\text{Error MS}}$	$\frac{\text{Faktor B MS}}{\text{A} \times \text{B MS}}$	$\frac{\text{Faktor B MS}}{\text{Error MS}}$
Interaktion mellan A och B	$\frac{\text{A} \times \text{B MS}}{\text{Error MS}}$	$\frac{\text{A} \times \text{B MS}}{\text{Error MS}}$	$\frac{\text{A} \times \text{B MS}}{\text{Error MS}}$

Envägs ANOVA

En envägs ANOVA använder man i de fall då man vill testa om det finns skillnader i medelvärdet hos en variabel t.ex. längd, koncentration, diversitet, etc., mellan olika nivåer i en enda faktor, t.ex. olika näringsgivor. Antalet nivåer ska vara minst tre. I exemplet ovan jämförde man hur medelvärdet av tillväxten varierade mellan tre olika näringsnivåer. Detta är ett typexempel på en envägs, eller enfaktors, variansanalys.

Innan man startar beräkningarna måste man fundera om data behöver transformeras. Avsnittet om datatransformation ovan beskriver olika möjligheter. Variansanalys bygger på att data är (approximativt) normalfördelade och om man inte kan uppnå en sådan fördelning med hjälp av tranformation ska man använda motsvarande ickeparametriska tester. En variansanalys är dock relativt okänsligt för avvikelser från normalfördelningen vilket gör att de flesta data som är på intervall- eller kvotskalan kan analyseras med ANOVA. Motsvarande ickeparametriska tester är Kruskal-Wallis test och Friedmans test. Mer om dessa tester följer.

Precis som t-test förutsätter nyttjandet av variansanalys att observationerna kommer från samma bakomliggande population och därmed har samma varians, s.k. homoscedasticitet. Om variansen är olika för olika nivåer har man s.k. heteroscedasticitet vilket leder till att det inte går att utföra en variansanalys. Ett vanligt test för att kontrollera om stickproven har samma varians är Barletts test och många statistikprogram beräknar detta test automatiskt när man gör en ANOVA. I Barletts test testas nollhypotesen att varianserna är lika mot alternativhypotesen att varianserna inte är lika. För att förutsättningarna för en ANOVA ska vara uppfyllda vill man i detta test således ha ett P-värde som är större än α, d.v.s. 0,05, så att nollhypotesen kan behållas. Om p-värdet blir mindre än 0,05 har minst en nivå hos den faktor man studerar en varians som skiljer sig från övriga nivåers varians.

Om Barletts test ger att varianserna skiljer sig kan man antingen konstatera att de studerade nivåerna skiljer sig, och nöja sig med detta. En skillnad i varians kan vara lika viktig som en skillnad i medelvärde. En annan åtgärd kan vara att utföra någon form av transformering för att reducera variansen och ett tredje sätt är att använda en variant av ANOVA som inte förutsätter jämn varians. Det senare alternativet är dock ovanligt. Ytterligare ett sätt är att använda en icke-parametrisk variant av ANOVA.

Om data inte är normalfördelade eller om de är på ordinalskalenivån finns flera ickeparametriska tester att tillgå. De vanligaste är Kruskal-Wallis test om man har oberoende grupper och Friedmans test då man har beroende grupper.

Kruskal-Wallis test

Detta är ett effektivt test för att pröva hypoteser av ANOVA-typ där data inte uppfyller kraven för en vanlig variansanalys. Testet kallas också för H-test eftersom testvariabeln benämnes H. Principen är att alla observationer rangordnas, oavsett grupp. Därefter jämför man summan av alla rangtal för de olika grupperna. Om nollhypotesen är korrekt, d.v.s. att alla stickprov kommer från samma population, kommer summan varje grupps rangtal att vara ungefär lika stora.

Testvariabeln H beräknas med följande formel:

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{N_i} - 3(N-1),$$

där N = det totala antalet observationer, N_i = antal observationer i grupp i och R_i = summan av rangtalen i grupp i.

Om det finns rangtal som är lika korrigerar man testvariabeln genom att dividera H med en korrektionsfaktor som kallas C. Man får då H_{korr} = H/C, där

$$C = 1 - \frac{\sum(t_i^3 - t_i)}{(N^3 - N)},$$

där t_i = antal lika rangtal i den i:te gruppen av lika rangtal.

Det kritiska värdet på testvariabeln får man i en chi-två-tabell. Antalet frihetsgrader är antalet grupper minus 1. Om det erhållna värdet på H är lägre än den kritiska nivån man erhåller från chi-två-tabellen behåller man nollhypotesen.

Om man utför Kruskal-Wallis test på försök 1 i tabell 12 får man de rangtal som visas i tabell 16. Underst i tabellen visas rangsummorna. Bara genom att titta på rangsummorna kan man gissa att de olika nivåerna (15, 20 resp. 25 ml) ger olika resultat.

Tabell 16. Rangtal för resultaten i försök 1 i tabell 12.

	Försök 1		
	Näringsgiva (ml)		
	15	**20**	**25**
	1	6	11
	2	7	12
	3	8	13
	4	9	14
	5	10	15
Antal obs. N_i	5	6	7
Rangsumma, R_i	15	40	65

Värdet på testvariabeln beräknas med hjälp av formeln ovan:

$$H = \frac{12}{15\times(15+1)} \times \left(\frac{15^2}{5} + \frac{40^2}{5} + \frac{65^2}{5}\right) - 3(15-1) = 18{,}5$$

Ur en chi-två tabell får man att det kritiska värdet på H för $\alpha = 0{,}001$ och 3-1 frihetsgrader är 13,8. Vi kan således konstatera att även den ickeparametriska varianten av variansanalys ger att tillväxten skiljer sig mellan behandlingarna.

Friedmans test

Detta test är den ickeparametriska varianten av en ”Repeated measures ANOVA”, som används när man har tre eller flera beroende grupper och man vill testa om det finns skillnader mellan grupperna. Som många andra ickeparametriska tester bygger även Friedmans test på rangtal. Principen är att rangsummorna blir ungefär lika för alla grupper om grupperna inte skiljer sig. En stor skillnad mellan rangsummorna tyder däremot på skillnader mellan grupper. Formeln för Friedmans test är besläktad med formeln för Kruskal-Wallis test men något mer komplex, därför presenteras inte formeln här. Testvariabeln är även i detta test chi-två-fördelad. Antalet frihetsgrader är antalet grupper minus 1.

Tvåvägs ANOVA

Hittills har alla tester som beskrivits i detta kapitel handlat om data med endast en faktor, med en, två eller fler nivåer. I en tvåvägs, eller tvåfaktors, ANOVA analyserar man effekter av två samverkande faktorer. Det kan t.ex. röra sig om en faktor med nivåerna ”Före” och ”Efter” en behandling samt en faktor med nivåerna ”Hane” och ”Hona” (tabell 17). I denna typ av försök kan man ha effekter från varje faktor för sig och genom en kombination av de båda faktorerna.

Tabell 17. Illustration av interaktioner i en tvåvägs ANOVA.

		Kön	
		Hona	Hane
Behand-ling	Före	Medelvärde_{11}	Medelvärde_{12}
	Efter	Medelvärde_{21}	Medelvärde_{22}

Eftersom vi kan förvänta oss effekter i tre nivåer kan vi konstruera tre nollhypoteser. Den första nivån rör den s.k. huvudeffekten som kommer av kön. Här försöker man besvara frågan om det finns några

könsanknutna skillnader utjämnat över alla nivåer i den andra faktorn, d.v.s. att summan av alla behandlingsgruppers medelvärden för honor ska vara lika med summa av alla behandlingsgruppers medelvärden för hanar. Något mer matematiskt uttryckt blir detta:

$H_{0,1}$: $(Medel_{11} + Medel_{21}) - (Medel_{12} + Medel_{22}) = 0$, d.v.s. medel för ”Honor” är lika med medel för ”Hanar”.

Nollhypotesen för behandlingseffekten testar om det finns någon skillnad mellan behandlingar, oavsett kön. På motsvarande sätt blir nollhypotesen att summan av honors och hanars medelvärden före behandling ska vara lika med summan av honors och hanars medelvärden efter behandling:

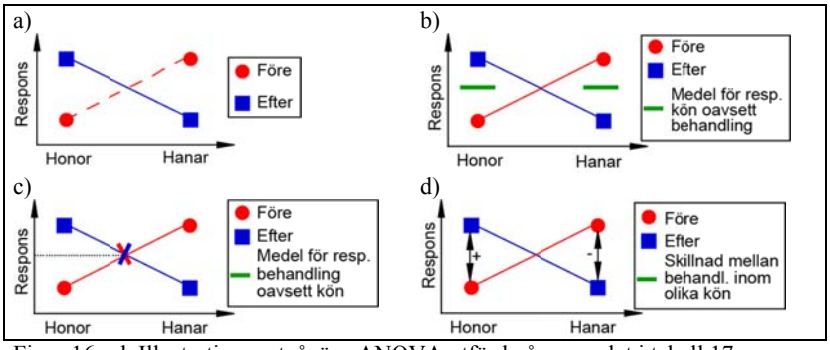
$H_{0,2}$: $(Medel_{11} + Medel_{12}) - (Medel_{21} + Medel_{22}) = 0$, d.v.s. medel för ”Före” är lika med medel för ”Efter”.

Slutligen ställer vi upp en nollhypotes om interaktionen, d.v.s. samverkan, mellan de båda faktorerna. Man vill då ha svar på frågan: finns det någon skillnad mellan könen i respons på behandlingen. Nollhypotesen blir matematiskt densamma som $H_{0,1}$.

$H_{0,3}$: $(Medel_{11} + Medel_{21}) - (Medel_{12} + Medel_{22}) = 0$, d.v.s. skillnaden mellan behandlingarna av honor ska vara lika stor som skillnaden mellan behandlingarna av hanar.

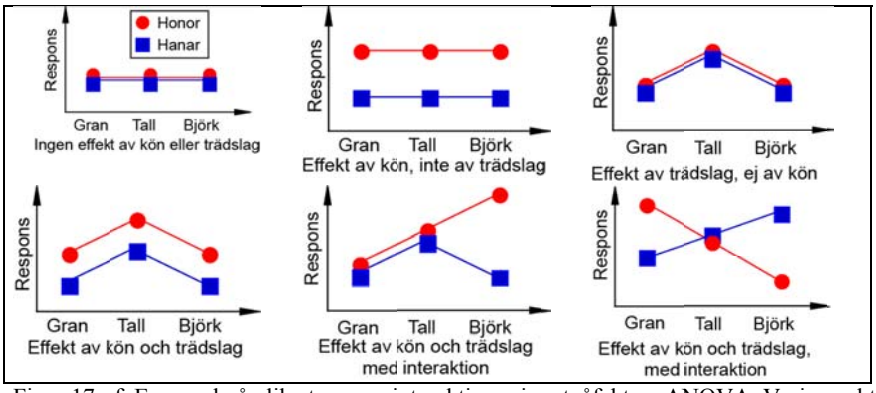
Resultaten från en tvåfaktorsanova kan illustreras med hjälp av bilder. I figur 16a visas medelvärdena av den uppmätta responsvariabeln i exemplet ovan. I 16b visas de olika könens medelrespons oavsett behandling. I detta exempel är de båda könens medelvärde lika om man inte tar hänsyn till behandlingen. Likaså är medelvärdena av de olika behandlingarna lika stora för båda könen (figur 16c). Däremot finns en tydlig effekt om man ser till interaktionen mellan kön och behandling (figur 16d). Skillnaden mellan medelvärdena för honor har motsatt tecken jämfört med skillnaden mellan medelvärdena för hanar.

Den grafiska beskrivningen visar således att det inte finns någon effekt av vare sig kön eller behandling om de testas var för sig, men att det finns en samverkanseffekt mellan de båda. Om det finns en interaktion är således linjerna i ett s.k. interaktionsdiagram, som i figur 16, inte parallella.



Figur 16a-d. Illustration av tvåvägs ANOVA utförd på exemplet i tabell 17.

Generellt kan man säga att en tvåfaktors ANOVA testar om skillnader mellan nivåer i en faktor varierar mellan nivåerna i den andra faktorn. I figur 17a-f visas exempel på interaktioner mellan två faktorer, kön (2 nivåer) och trädslag (3 nivåer). Responsen, d.v.s. det man mätt är vikten av en insektsart som man fångat på olika träd i ett skogsområde. Nollhypoteserna är 1) att det inte finns någon skillnad i vikt mellan honor och hanar, 2) att det inte finns någon skillnad i insekternas medelvikt oavsett kön mellan olika substratträdslag och 3) att det inte finns någon samverkan mellan kön och trädslag som påverkar insekternas medelvikt.



Figur 17a-f. Exempel på olika typer av interaktioner i en tvåfaktors ANOVA. Varje punkt visar medelvärdet av responsvariabeln för respektive kombination av kön och trädslag.

Trevägs ANOVA

Precis som man kan utöka en envägs ANOVA till en tvåvägs, kan man lägga till ytterligare en faktor till en tvåvägs ANOVA och på så sätt få en trevägs ANOVA. I en sådan finns effekter från varje faktor för sig, från interaktioner mellan två av de tre faktorerna och slutligen från interaktioner mellan alla tre faktorer. Dessa analyser kan dock bli så komplicerade att de faller utanför vad som är avsikten med detta kompendium och vi hänvisar till någon lärobok i ämnet, t.ex. Zar (1999).

Obalanserade modeller och saknade nivåer

För att få största möjliga styrka i en variansanalys bör man ha en s.k. **balanserad** studie. Detta innebär att man har lika många prover från varje kombination av faktorer (tabell 18a). Vidare bör man ha data från alla kombinationer av de studerade nivåerna. För att uppfylla dessa båda krav måste man ofta ha stora datamängder, speciellt vid två- och trevägs variansanalyser. Då man jobbar med miljöundersökningar kan det dock vara kostsamt och ibland även omöjligt att samla in så stora datamängder. I de fall man har olika stora stickprov från olika nivåer har man en s.k. **obalanserad** ANOVA. I vissa situationer kan det på grund av naturen hos de data man studerar komma sig att man helt saknar data för en nivå. Detta kan t.ex. bero på att vissa kombinationer av faktorer inte existerar (tabell 18b).

I de fall man har situationer som i tabell 18b måste man använda andra beräkningsrutiner än vid balanserade försök. Exemplen ovan gäller för balanserade modeller. Mer avancerade datorprogram kan lösa modeller med obalanserade data och med saknade nivåer med hjälp av en metod som kallas GLM, ”General Linear Models”, medan enklare statistikprogram saknar metoder för att lösa dessa typer av modeller. Teorin bakom detta och metoder för att lösa sådana modeller faller utanför intentionerna med detta kompendium. Metoderna finns beskrivna i många läroböcker, t ex. Rutherford (2001)

Tabell 18a. Exempel på en balanserad ANOVA. Siffrorna anger antal studerade objekt.

		Faktor A			
		Nivå 1	Nivå 2	Nivå 3	Nivå 4
Faktor B	Nivå 1	5	5	5	5
	Nivå 2	5	5	5	5
	Nivå 3	5	5	5	5

Tabell 18b. Exempel på en obalanserad ANOVA med en saknad nivå. Siffrorna anger antal studerade objekt.

		Faktor A			
		Nivå 1	Nivå 2	Nivå 3	Nivå 4
Faktor B	Nivå 1	•	5	6	5
	Nivå 2	4	5	3	5
	Nivå 3	2	4	2	6

“Repeated measures ANOVA”

Precis som för t-test kan man välja mellan **beroende** eller **oberoende** nivåer hos den faktor man studerar. Om man har beroende nivåer utför man en s.k. ”Repeated measures ANOVA”. Typiska situationer då man har beroende grupper är då man gjort upprepade mätningar på samma individer eller objekt, t.ex. före och efter en behandling. Behandlingen kan antingen vara en aktiv manipulation

eller att mätningarna upprepas med jämna intervall. I det senare fallet är tiden mellan mätningarna den behandling man utför. Vanligt är att man använder en av nivåerna som referensnivå, d.v.s. den nivå som gällde innan man satt igång någon form av manipulering.

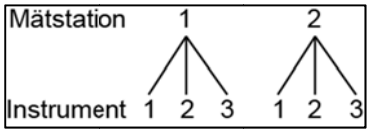
En förutsättning för att göra en ”Repeated measures ANOVA” är att man utför alla upprepade mätningar på alla objekt i studien. En annan förutsättning är att mätningarna på samma objekt är oberoende av varandra. Denna förutsättning är besläktad med kravet på homoskedasticitet (se avsnittet om datatransformation) som gäller för t-tester och variansanalyser där man jämför olika grupper. När man utför upprepade mätningar på samma grupp försvinner visserligen variationen mellan grupper, men vi måste istället ta hänsyn till s.k. kovarians hos de enskilda objekten. Eftersom man utför mätningar på samma objekt upprepade gånger finns det en risk att objekt som ger låga (eller höga) värden vid första mätningen även ger låga (höga) värden vid den andra mätningen, en s.k. kovariation. Detta risk är speciellt hög om man utför mätningarna så att den första mätningen påverkar resultatet i den andra mätningen. I beräkningarna i en ”repeated measure ANOVA” tar man hänsyn till denna kovarians, men en förutsättning för att detta ska fungera är att kovariansen är ungefär lika stor för varje objekt. På engelska kallas denna förutsättning för ”sphericity” eller ”circularity” och ovanstående beskrivning är en mycket grov förenkling för att ge en aning om vad som menas med ”sphericity”. Om data inte uppfyller detta krav kommer det framräknade P-värdet blir för lågt.

Matematiskt skiljer sig en ”Repeated measures ANOVA” från en vanlig ANVOA i princip genom sättet att beräkna F-kvoten. Om man studerar en faktor med endast två nivåer blir en ”repeated measures ANOVA” detsamma som ett t-test för beroende grupper.

Hierarkiska modeller

I vissa typer av variansanalyser är de olika faktorerna relaterade till varandra på ett hierarkiskt sätt. Med detta menas att olika nivåer inom en faktor kan delas upp i nya nivåer i en annan faktor. Man kan t.ex. ha flera mätstationer längs ett vattendrag och flera mätinstrument på varje mätstation. Här kan varje nivå i faktorn ”mätstation” delas upp i nya nivåer i faktorn ”instrument” (figur 18). Eftersom beroendet mellan de olika faktorerna är linjärt uppifrån och ner, eller tvärt om, har t.ex. mätinstrument nummer ett på station nummer ett ingenting att göra med instrument ett på station nummer två. Detta måste specificeras när man gör en variansanalys. Om man inte specificerar detta kommer beräkningarna att bli felaktiga. De flesta datorprogram har rutiner för att specificera om man har en hierarkisk utformning på sin studie.

De engelska uttrycken för denna typ av variansanalys är antingen ”Hierarchic ANOVA” eller ”Nested ANOVA”.



Figur 18. Illustration av hur faktorer är organiserade i en hierarkisk variansanalys. Instrument 1 vid mätstation 1 har inget samband med instrument 1 vid mätstation 2.

Post hoc tester

Om en variansanalys ger att man kan förkasta nollhypotesen vet man att medelvärdet hos minst en nivå i en faktor skiljer sig från minst en annan nivå i samma faktor. Däremot vet man inte vilken eller vilka nivåer som skiljer sig från varandra. I avsnittet om grundläggande begrepp inom ANOVA framgår att mothypotesen är en kombination av en mängd utsagor.

Om, och endast om, en variansanalys ger att man kan förkasta nollhypotesen kan man gå vidare med en grupp av analyser som gemensamt kallas ***post hoc*** tester. Dessa tester utför man för att ta reda på vilken eller vilka nivåer som skiljer sig signifikant. *Post hoc* tester fungerar i princip som upprepade t-tester där man jämför alla par av nivåer inom en faktor. Det finns dock en stor skillnad mot att utföra en massa upprepade t-tester. Om man gör flera t-tester på samma data kommer risken för ett typ-I fel

ackumulera för varje upprepat test. När man gör *post hoc* tester är denna risk betydligt lägre och de P-värden testet ger är betydligt mer tillförlitliga än de P-värden upprepade t-tester ger. Trots att *post hoc* tester är robusta finns det varianter som korrigerar α efter antalet upprepade tester, s.k. Bonferroni korrektion.

Vilket test som ska användas beror delvis på frågeställningen och delvis på hur konservativt test man vill använda. I dessa sammanhang innebär konservativ att testet kräver stora skillnader mellan två medelvärden innan det signalerar att det finns en signifikant skillnad mellan de nivåer man testar. Om man väljer ett test som är mycket konservativt finns en risk att man drar slutsatsen att det inte finns någon skillnad trots att det i själva finns en skillnad, ett s.k. typ-II fel. Väljer man däremot ett test som är mindre konservativt finns risken att man drar slutsatsen att det finns en skillnad när det egentligen inte finns det, d.v.s. ett typ-I fel.

Det finns minst ett femtiotal olika *post hoc* tester. Några av de vanligaste är Scheffes F-test, Tukeys test och Fishers PLSD och Bonferroni. Oftast används någon av dessa. Av de uppräknade testerna är Tukeys test det minst konservativa, följt av Fishers PLSD. Scheffes test är ganska konservativt och Bonferroni-testet blir mer och mer konservativt ju fler nivåer man jämför. Ju mer konservativt ett test är desto större måste skillnaden mellan medelvärdena hos de två nivåer man jämför vara för att testet ska ge att skillnaden är signifikant (tabell 19).

Tabell 19. Den lägsta skillnad i medeltillväxt mellan olika nivåer i försök 1 i tabell 12 som krävs för att två nivåer ska anses signifikant skilda, för fyra olika *post hoc* tester.

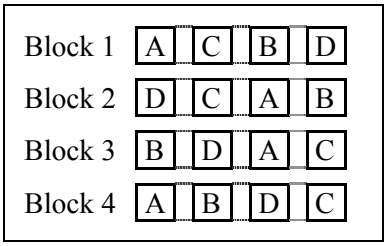
Signifikant* om skillnaden överstiger			
Tukeys	Fisher's PLSD	Bonferroni	Scheffe
3,56	4,32	4,96	5,09

* $P < 0,01$

Försöksdesign och ANOVA

Detta avsnitt bör läsas i kombination med avsnittet ”Bakomliggande modeller” ovan, då det bygger på de bakomliggande modellerna i variansanalys. I *Planering av undersökningar* i Naturvårdsverkets handbok för miljöövervakning nämns det att det finns tekniker som omvandlar störningar i ett experiment till kontrollerade slumpvariabler. Exempel på störande faktorer kan vara heterogenitet i markkemi när man utför ett gödslingsförsök. Vid sådana tillfällen kan man genom att välja rätt försöksdesign kontrollera slumvariationen i en tvåvägs ANOVA. Det finns flera olika typer av försöksdesign. Först beskrivs den design som på engelska kallas ”**Randomised Complete Block**”.

Antag att man vill se om det finns skillnader mellan fyra olika gödningsmedel i hur stor skörden blir. För att testa detta behövs minst fyra försöksytor, en yta per gödningsmedel. Men, med endast en yta per gödningsmedel får man inget mått på variationen vilket krävs i en ANOVA. Därför behöver man replikat. I detta fall väljer vi fyra replikat. Den totala area som krävs för detta försök kommer att bli ganska stor och man har då anledning att anta att det finns variationer i markkemi som kommer att påverka resultatet av studien. För att eliminera effekter från variation i markkemi i analysen låter man slumpen avgöra i vilken provyta man ska använda de olika gödningsmedlen. Man låter dock inte slumpen styra helt och hållet utan delar in ytorna i fyra s.k. block. Inom varje block slumpar man vilken yta som ska behandlas med vilket gödningsmedel (figur 19).



Figur 19. Exempel på randomiserad block-design. Fyra olika gödningsmedel (A - D) har använts i 4 × 4 ytor. Ytorna är indelade i block och ordningen inom varje block är slumpad.

Resultaten från studien kan analyseras med en tvåvägs ANOVA, där man testar nollhypotesen att de olika gödningsmedlen ska ge samma skörd. Som i alla tvåvägs variansanalyser kan man även ställa upp en nollhypotes om den andra faktorn, men i detta fall är vi inte intresserade av den nollhypotes som rör skillnader mellan de olika blocken. De olika blocken är dock en viktig varianskomponent och den måste tas med i beräkningarna. Vi är inte heller intresserade av interaktionen mellan block och gödningsmedel eftersom denna interaktion inte har någon biologisk relevans. Faktum är att om man tar med interaktionen som en varianskomponent i modellen blir resultatet felaktigt. Vi är bara intresserade av effekten av gödningsmedel och inget annat, men använder blockfaktorn för att resultatet inte ska vara påverkat av eventuell heterogenitet i marken.

Den ANOVA-tabell som beräkningarna ger kommer således att innehålla de två faktorerna och inomgruppsvariationen som ofta benämnes ”Error” eller ”Residual” (tabell 20a). Eftersom gödningsmedel är en fix faktor och block en stokastisk faktor kommer analysen att vara en s.k. ”mixed model ANOVA” och F-faktorn beräknas enligt modell III (tabell 15).

Ur tabell 20a kan vi utläsa att de olika gödningsmedlen gav en signifikant skillnad i tillväxt, och att de olika blocken påverkat resultatet. I tabell 20b visas det resultat man erhåller om man inte tar hänsyn till variation mellan blocken. Som synes är det i detta exempel avgörande för resultatet om vi tar med blockeffekten i ANOVA-modellen eller ej. I tabell 20b är den variation som beror på skillnader mellan blocken inkluderad i inomgruppsvariationen och därför blir denna väldigt hög (561). I 20a är denna variation uppdelad i två komponenter och inomgruppsvariationen blir då endast 15,1. Resterande 546 enheter av inomgruppsvariationen från tabell 20b är lagd i faktorn ”Block” i tabell 20a. Vi drar oss till minnes att F-kvoten är medelkvadratsumman för en effekt dividerat med medelkvadratsumman för variationen inom grupper. Detta leder till att F-kvoten för gödsel blir hög (och P-värdet därmed lågt) i tabell 20a eftersom man har en låg inomgruppsvariation.

Tabell 20a. ANOVA-tabell för exemplet i figur 19. Fiktiva data.

Source	DF	Sum of Squares	Mean Square	F-Value	P-Value
Gödsel	3	379	126	75,4	<0,0001
Block (Random)	3	546	182	109	<0,0001
Residual	9	15,1	1,67		

Tabell 20b. Samma ANOVA som i tabell 20a, men utan blockeffekten.

Source	DF	Sum of Squares	Mean Square	F-Value	P-Value
Gödsel	3	379	126	2,70	0,0926
Residual	12	561	46,8		

Utöver randomiserad block-design, finns flera andra vedertagna försöksuppställningar. En försöksuppställning är ”**Latin square**” (romersk kvadrat på svenska), som är en utökad variant av den randomiserade block-designen. Istället för att kontrollera en störande faktor tar man i ”Latin square”-designer hänsyn till två störande gradienter i undersökningsområdet, t.ex. samtidiga fuktighets- och ljusgradienter. En annan vanlig försöksuppställning är ”**Split plot design**” (försöksplan med delade rutor). I denna design har man till att börja med en randomiserad block-design som i figur 19, men för att kunna ta hänsyn till en ytterligare faktor, t.ex. behandling med olika herbicider (här kallade 1 och 2), delar man upp varje försöksyta i flera delar (figur 20). Inom varje yta är ordningen av 1 och 2 slumpad.

Block 1	A1	A2	C1	C2	B2	B1	D2	D1
Block 2	D2	D1	C1	C2	A2	A1	B1	B2
Block 3	B2	B1	D2	D1	A1	A2	C1	C2
Block 4	A1	A2	B1	B2	D2	D1	C2	C1

Figur 20. Exempel på en ”Split plot design” baserad på försöket i figur 19, men utökad med två sorters ogräsmedel (1 och 2).

De ANOVA-modeller som skall användas för de två ovanstående försöksuppställningarna är väl beskrivna i en mängd statistikböcker. Utöver de tre försöksuppställningar som presenterats här finns flera andra, mer eller mindre intrikata, försöksuppställningar. För dessa hänvisas till läroböcker i statistik. Några gemensamma nämnare är att man måste hålla reda på vilka faktorer som är fixa och vilka som är stokastiska, och om man ska ha med interaktionen i ANOVA-modellen.

Kovariansanalys - ANCOVA

Ovan beskrivs olika sätt att ta hänsyn till bakgrundsvariation genom att inkludera den som en faktor i en ANOVA-modell. Det finns dock ytterligare sätt att lösa problemet, som dessutom ger ett starkare test eftersom man tillför mer information. Istället för att låta variationen ingå som en faktor som i figur 19, kan man helt enkelt mäta markkväve innan experimentet startar. Sedan kan man inkludera denna information när man analyserar data. Det statistiska test man använder i detta fall kallas ANCOVA, vilket är en akronym för ”ANalysis of COVAriance”. Denna statistiska teknik är en kombination av variansanalys (ANOVA) och regression. Förklaringsvariablerna är dels faktorer som i en variansanalys och dels en kontinuerlig variabel på intervall- eller kvotskalnivån, som på engelska kallas ”covariable”. Utöver detta har man i vanlig ordning en responsvariabel, d.v.s. det objekt man gör sina mätningar på.

Om man utför en ANCOVA på exemplet i figur 19 kommer gödselmedel att vara faktor, markkväve att vara ”co-variabel” och skörd vara responsvariabel. I tabell 21 presenteras resultatet av en ANCOVA på samma data som i tabell 20, men med faktiska markkvävehalter medtagna som en co-faktor. Precis som i tabell 20a får man en signifikant effekt av de olika gödselmedlen när man tar hänsyn till variationen i markkväve.

Tabell 21. Resultat av en ANCOVA på samma data som i tabell 20.

Source	DF	Sum of Squares	Mean Square	F-Value	P-Value
Gödsel	3	403	134	21,8	<0,0001
Kväve	1	493	493	80,0	<0,0001
Residual	11	67,9	6,17		

De flesta statistikprogram har ingen speciell modul för ANCOVA, istället använder man ANOVA-funktionen och anger sin kontinuerliga co-faktor som en av faktorerna (de oberoende variablerna) i modellen.

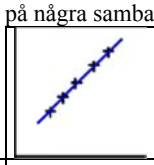
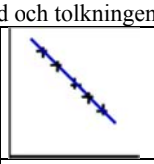
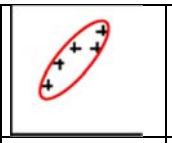
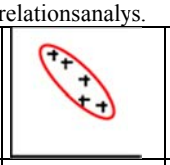
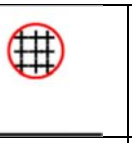
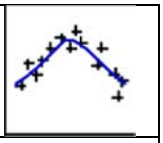
Korrelation och regression

Hittills har alla beskrivna tester handlat om olika sätt att beräkna om mätvärden skiljer sig från varandra. Nu ska vi istället beröra statistiska tekniker som man använder för att **beskriva samband** mellan variabler.

Korrelation

Om det finns ett samband mellan två kontinuerliga variabler kan man beskriva detta samband med hjälp av en korrelationsanalys. En förutsättning är dock att man kan ordna observationerna i par. Ytterligare en förutsättning är att sambandet är linjärt (tabell 22). I en korrelation beräknar man ett värde som säger hur starkt ett samband mellan två variabler är. Man tar inte hänsyn till vad som är X och Y, d.v.s. om det finns en beroende och en oberoende variabel. Oftast går det inte att säga vad som är den beroende och den oberoende variabeln, t.ex. som i exemplet i tabell 23. På grund av detta säger en korrelation inget om orsak och verkan, eller om det finns ett funktionellt samband mellan variablerna. Det kan mycket väl vara så att det finns orsakssamband, men en korrelation ger ingen information om detta. I tabell 22 finns några exempel på samband. Som synes finns det både positiva och negativa korrelationer.

Tabell 22. Exempel på några samband och tolkningen av dessa i en korrelationsanalys.

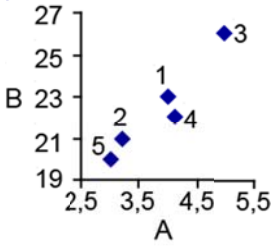
						
Beskrivning av relationen:	Perfekt positivt samband	Perfekt negativt samband	Starkt positivt samband	Starkt negativt samband	Inget samband	Icke-linjärt samband
Exempel på r:	1,0	-1,0	0,75	-0,75	0	Korrelation ej lämplig

Beroende på vilka data man har använder man olika varianter av korrelationstester. Vid (approximativt) normalfördelade data använder man **Pearsons produkt-momentkorrelation**. Om båda variablerna är av ordinalskaletyp eller om data inte är normalfördelade används istället Spearmans, eller Kendalls rangkorrelation.

Idéerna bakom korrelation bygger på att jämföra varje observation med medelvärdet. I tabell 23 visas data för fem par observationer och hur varje observation avviker från gruppmedelvärdet. Vi kallar observationerna A och B. Om en observation har höga värden i både A och B blir produkten av avvikelsen från respektive medelvärde hög och positiv, som t.ex. för observation 3. Detsamma gäller om både A och B i ett par är låga, som för observation 5. Generellt gäller att om man har data där höga värden i en variabel oftast hänger ihop med höga värden i den andra variabeln och låga värden hänger ihop med låga värden, kommer summan av produkterna vara hög och positiv.

Tabell 23. Exempel på beräkningar i en korrelation. Till höger en illustration av data i exemplet.

Obs. nr.	A	B	$A - \bar{A}$	$B - \bar{B}$	Produkt
1	4	23	0,14	0,6	0,084
2	3,2	21	-0,7	-1,4	0,92
3	5	26	1,14	3,6	4,10
4	4,1	22	0,24	-0,4	-0,096
5	3	20	-0,9	-2,4	2,06
Medel:	3,86	22,4		Summa:	7,08



Om höga värden i en variabel genomgående hänger ihop med låga värden i den andra variabeln blir summan av produkterna också hög, men i detta fall negativ. Om det inte finns någon genomgående trend får man ibland positiva och ibland negativa avvikelser från medelvärdet. Detta gör att summan av alla produkter kommer att bli nära noll.

Kovarians

I beskrivningen korrelation bryter vi här in med en beskrivning av ett närbesläktat begrepp; kovarians. Detta gör vi för att kovariansen ingår som en del i beräkningen av korrelation.

Summan av alla produkter i tabell 23 (7,08) beror av antalet observationer. Ju fler observationer desto högre blir summan. För att kunna jämföra olika undersökningar måste man därför dividera summan med antalet frihetsgrader vilket är n-1. Denna division ger en slags medelavvikelse, och resultatet är **kovariansen**. I exemplet i tabell 23 blir kovariansen 7,08/(5-1) = 1,77. Eftersom man inte tagit hänsyn till skillnader i variablernas varians är kovarians (som namnet antyder) ett mått på hur två variabler varierar tillsammans.

Genom beräkningen av kovarians har man kommit runt problemet med att A och B har olika medelvärden. Däremot har man inte tagit hänsyn till skillnader i de olika gruppernas varians. Detta gör att man inte kan använda kovarians om man vill beräkna samvariationen hos två variabler som är mätta i olika enheter. Ett exempel kan vara sambandet mellan pH och deposition. pH är enhetslöst medan deposition mäts i t.ex. kg/hektar/år. I dessa fall använder man korrelation för att beskriva sambandet.

Korrelation - fortsättning

För att komma från kovarians till korrelation dividerar man kovariansen med de båda gruppernas standardavvikelse. Genom denna beräkning erhåller man korrelationskoefficienten, r, för de båda grupperna. Vi får:

$$r_{AB} = \frac{\text{Kovarians}_{\text{AB}}}{\text{Standardavvikelse}_A \times \text{Standardavvikelse}_B}.$$

I exemplet ovan är standardavvikelsen 0,779 respektive 2,30 för grupp A och B. Om vi sätter in dessa siffror i formeln ovan få vi:

$$r_{AB} = \frac{1,77}{0,799 \times 2,30} = 0,96.$$

Genom division med standardavvikelsen blir korrelationskoefficienten standardiserad. Detta gör att en korrelationsanalys är oberoende av viken skala grupperna har. Det är matematiskt helt korrekt att t.ex. jämföra en sträcka mätt i millimeter med en koncentration mätt i mol/l, om man har sådana data. Om denna jämförelse även har någon funktionell betydelse får man dock inte svar på i en korrelationsanalys. En korrelation ger bara sambandet mellan två uppsättningar av siffror.

En naturlig fråga efter att man beräknat kovarians och korrelation är vad de två framräknade siffrorna betyder. Som redan visats i tabell 22 kan korrelationskoefficienten variera mellan +1 och -1. Vad som är ett starkt samband beror väldigt mycket på vad det är för data man undersöker. Biologiska data har ofta en stor naturlig varians. Detta gör att man får lägre r-värden än när man beräknar samband mellan variabler med en låg varians. De flesta datorprogram ger tillsammans med r-värdet även P-värden för korrelationer och detta är givetvis en stor hjälp när man ska tolka sitt r-värde.

Som beskrivits tidigare är ett P-värde ett mått på hur säker man kan vara att förkasta eller behålla en nollhypotes. I en korrelation säger nollhypotesen att det inte finns något samband mellan variablerna. Mothypotesen säger att det finns en så pass stark korrelation att den inte kan förklaras som ett resultat av slumpen.

I exemplet i tabell 23 är P-värdet 0,005. Detta indikerar att det är en ytterst låg sannolikhet att fem par slumpmässigt dragna värden skulle vara lika starkt korrelerade som de värden man observerat. Dock gäller som i tidigare avsnitt att om man har stora dataset kan man få signifikanta resultat som är så små att de inte har någon praktisk betydelse. Oftast är man ute efter ett P-värde lägre än 0,05. Om man får det kan man förkasta nollhypotesen. Men, om både r- och P-värdet är lägre än 0,05 (vilket kan

inträffa) bör man noga överväga om man har en korelation värd namnet. I tabell 24 finns en guide hur man ungefärligen kan tolka värdet på r (efter Fowler m. fl. (1998)).

Tabell 24. Ungefärlig tolkning av styrkan på en korrelationskoefficient.

Korrelationskoefficient, r	Styrka
0 till 0,19	Mycket svag
0,20 till 0,39	Svag
0,40 till 0,69	Måttlig
0,70 till 0,89	Stark
0,90 till 1	Mycket stark

Icke-parametriska korrelationstester

I de fall man har data som inte uppfyller kraven för parametriska tester finns flera icke-parametriska alternativ. Vanligast är **Spearmans rangkorrelation**, r_S, men **Kendalls τ** (tao) är också vanligt förekommande. Båda grundar sig i att man rangordnar de observerade värdena. I Spearmans test behandlas rangtalen som mätvärden på intervallskalan, medan Kendalls τ tar hänsyn till att rangtalen är just rangtal. Det finns teoretiska beskrivningar av när den ena eller andra ska användas, men i detta kompendium stannar vi med att säga att båda kan användas då man har icke-parametriska data.

Skillnader mellan olika r

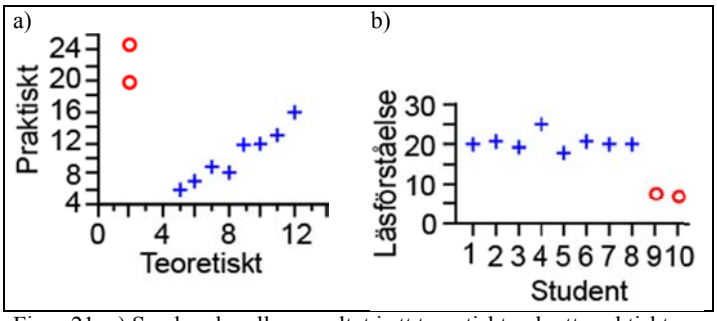
Slutligen är det dags att påpeka att de tre metoder för att beräkna korrelation som presenterats i detta avsnitt inte går att jämföra! I exemplet i tabell 23 är Spearmans r = 0,90 och Kendalls τ = 0,80, medan Pearsons r = 0,96. Följaktligen indikerar en korrelationskoefficient på t.ex. 0,8 olika starka samband beroende på vilket test som används. Därför måste man alltid ange vilket test man använt för att beräkna en korrelation.

Partiell korrelation

Partiell korrelation används när man vill studera sambandet mellan två variabler som i sin tur påverkas av andra variabler. Ett konstruerat och något orealistiskt men illustrativt exempel får belysa denna metod. Antag att man tränar fältpersonal att utföra standardiserade provtagningar. Mot slutet av utbildningen har man ett skriftligt prov och ett praktiskt prov. Man kan då tänka sig att det kommer att finnas ett positivt samband mellan resultaten i de båda proven. De som får höga resultat i det praktiska provet borde också ha höga resultat i det teoretiska, och tvärt om. Detta testar man genom en enkel korrelationsanalys enligt ovan.

Nu kan det dock vara så att fältpersonalen skiljer sig vad gäller läsförmåga. Några kan vara väldigt duktiga i fält men ha ett läshandikapp. Om detta är fallet skulle korrelationsanalysen ovan vara en direkt felaktig metod eftersom några personer kommer att få låga poäng på den teoretiska delen enbart p.g.a. sitt handikapp. För att testa sambandet mellan resultaten i de båda proven skulle man således behöva ta hänsyn till läsförmåga. Detta kan man göra med hjälp av en s.k. **partiell korrelation**. En förutsättning i detta exempel är dock att man har tillgång till resultaten från ett läsförståelseprov.

I en partiell korrelation testar man sambandet mellan två variabler medan man tar bort effekten av en eller flera andra variabler. I figur 21a illustreras sambandet mellan resultatet i ett praktiskt och ett teoretiskt prov för tio kursdeltagare. Som synes avviker två deltagares resultat kraftigt från den övriga gruppens (dessa två är markerade med cirklar). Korrelationen mellan resultaten från det teoretiska och det praktiska provet är låg och ickesignifikant (r = -0,37; P = 0,31). Om man istället utför en partiell korrelation med resultaten från ett läsförståelseprov (figur 21b) som s.k. kofaktor blir korrelationen mellan teoretiskt och praktiskt istället 0,73. Genom detta förfarande har man tagit bort effekten av skillnader i läsförmåga och på så sätt fått ett mer korrekt mått på korrelationen mellan resultaten i det praktiska och teoretiska provet.



Figur 21. a) Samband mellan resultat i ett teoretiskt och ett praktiskt prov i samma ämne.
b) resultat i ett läsförståelsetest. Två studenter har starkt avvikande resultat och dessa är markerade med cirklar.

Förutsättningar

För alla typer av korrelationsanalyser finns några gemensamma förutsättningar:

1) Sambandet måste vara linjärt.
2) Alla prover måste komma från samma population. Om man t.ex. jämför ett akvatisk bottenfaunaindex med pH i vattendrag kan man inte blanda både förorenade och rena vattendrag i samma korrelation.
2) Proverna måste vara oberoende i den bemärkelsen att A inte får vara en del av B. Om man t.ex. jämför antal arter per provyta som en otränad kartör hittar med antalet arter en erfaren kartör hittar ingår det antal arter som den oerfarna kartören hittade i antalet från den erfarna. I detta fall är A och B inte oberoende.
3) Man ska inte kunna avgöra vad som är beroende och oberoende variabel. Om man fixa nivåer på X-variabeln (t.ex. koncentration, tid eller dos) är denna en oberoende variabel. I detta fall ska man istället använda regression.
4) Båda variablerna måste vara normalfördelade för att man ska kunna göra en Pearson-korrelation. I annat fall gör man en Spearman- eller Kendall-korrelation.

Linjär Regression

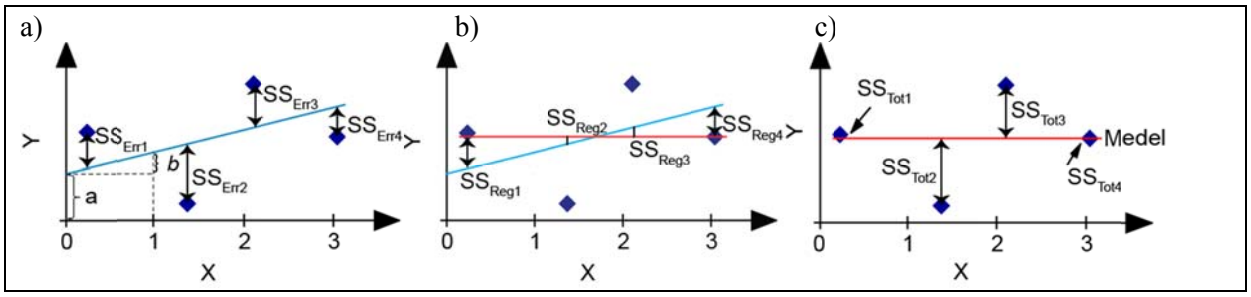
I en regression testar man sambandet mellan variabler som har ett **funktionellt samband**. Det gäller t.ex. för sambandet mellan ålder och vikt eller mellan grad av miljöpåverkan och bottenfaunaindex. I båda dessa exempel finns en beroende och en oberoende variabel. I det första exemplet är ålder oberoende eftersom vikt inte påverkar ålder men ålder kan påverka vikt. I det andra exemplet är miljöpåverkan den oberoende variabeln eftersom det är bottenfauna indexet som påverkas av miljön. Miljön kan däremot inte påverkas av indexet.

Regressionsanalys är väldigt nära släkt med variansanalys. Resultaten från en regression presenteras därför oftast i en s.k. ANOVA-tabell när datorprogram gör beräkningarna. I figur 22 illustreras de avstånd som används för beräkning av SS_{Err} och SS_{tot}, som båda ingår i ANOVA-tabellen för en regression. Om man studerar teorin bankom ANOVA (se ANOVA-avsnittet) märker man att ANOVA är en form av regressionsanalys.

Namnet ”linjär regression” antyder måste det finnas ett linjärt samband mellan de variabler som studeras. I detta sammanhang betyder linjär att de olika variablerna i regressionsekvationen är adderade och inte, som man kan tro, att det rör sig om en rät linje. I många fall beskriver en linjär regression en kurva. Om sambandet är inte är linjärt, dv.s. termerna är multiplicerade etc. finns flera andra metoder och dessa berörs senare i detta avsnitt och i nästa avsnitt.

Grunden i linjär regression är att man drar en linje genom den svärm av punkter man får när man prickar in sina data i ett diagram. Lutningen på denna linje ska vara sådan att linjen är så nära varje punkt som möjligt. På detta sätt visar linjen ett medelsamband mellan de båda variablerna. För att dra linjen på detta sätt använder man den s.k. minsta kvadratmetoden som vi träffade på i beskrivningen av ANOVA ovan. Men, i detta fall är det avståndet mellan de observerade värdena och regressionslinjen som ingår i beräkningarna, och inte avståndet mellan observerade värden och

medelvärdet som i ANOVA (figur 22a). Om lutningen på regressionslinjen är korrekt kommer summan av SS^2_{Err1}, SS^2_{Err2}, SS^2_{Err3} och SS^2_{Err4} i figur 22a vara den lägsta möjliga. Alla andra lutningar på regressionslinjen kommer att ge en högre summa.



Figur 22. Illustration av de olika kvadratsummorna i en enkel linjär regressionsanalys a) Avstånden mellan regressionslinjen och varje enskild observation (SS_{Err}), b) avstånden mellan medelvärdet av alla Y och regressionslinjen (SS_{Reg}), c) totala kvadratsumman (SS_{Tot}) som baseras på avståndet mellan medelvärdet av alla Y och de enskilda observationerna. I a) illustreras även regressionskoefficienterna *a* och *b*. OBS! För att göra de olika avstånden i a) och b) tydliga är regressionslinjen ej korrekt återgiven i detta exempel!

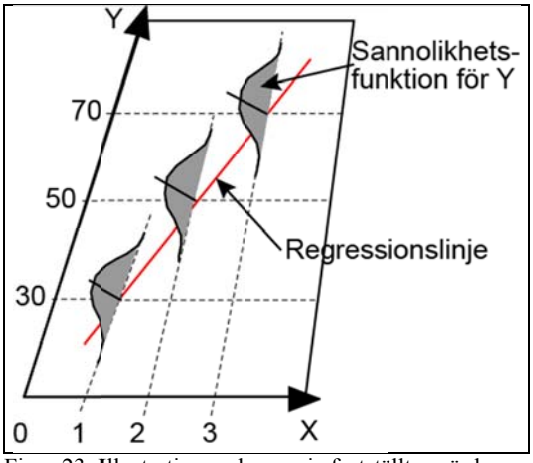
Matematiken bakom metoden för att beräkna lutningen på regressionslinjen är ganska omfattande. Därför nöjer vi oss med att presentera den färdiga regressionsekvationen. En regressionslinje beskrivs med följande ekvation:

$Y = a + b \times X,$

där *a* och *b* är s.k. **regressionskoefficienter**. *a* beskriver värdet på responsvariabeln, Y, när X = 0, och *b* beskriver linjens lutning, d.v.s. den förändring i y-led som motsvaras av 1 enhets förändring i x-led (figur 22a).

När man erhållit värden på *a* och *b* i ekvationen har man en ekvation som beskriver sambandet mellan X och Y i det stickprov man har till sitt förfogande.

Eftersom man utför regressionsanalysen på ett stickprov beskriver inte regressionsekvationen det samband som gäller för hela den studerade populationen, utan endast sambandet för stickprovet. Om man upprepar ett försök ett stort antal gånger kommer man för varje upprepning att få nya y-värden för de fastställda x-värdena. Efter oändligt många (tänkta) upprepningar kommer man att få en hel population av y-värden för varje x-värde (figur 23). Då en av förutsättningarna för regresion är att varje y-population måste vara normalfördelad är det störst sannolikhet att få ett y-värde nära populationens medelvärde.



Figur 23. Illustration av hur varje fastställt x-värde, motsvaras av en teoretisk population av y-värden.

En regressionsanalys är i grunden en hypotesprövning. Den nollhypotes som prövas är:

H_0: Lutningen på regressionslinjen är inte signifikant skild 0 (d.v.s. ett vågrätt streck).

Mothypotesen är följaktligen:
H_1: Lutningen på regressionslinjen är signifikant skild från 0.

När man prövar nollhypotesen utnyttjar man den teoretiska normalfördelningen kring varje y-värde (figur 23). Datorprogrammen ger direkt P-värdet för en regression utifrån detta värde kan man välja att behålla eller förkasta nollhypotesen.

Om man kan konstatera att lutningen, d.v.s. *b* i regressionsekvationen, är skild från 0 kan man genom den s.k. determinationskoefficienten, $\mathbf{r^2}$, få reda på hur mycket av variationen i y-variabeln som kan förklaras av variationen i x-variabeln. Ju högre värde r^2 har desto starkare är sambandet mellan X och Y. Determinationskoefficienten beräknas som 1 minus förhållandet mellan kvadratsumman för regressionslinjen (SS_{Err}) och totala kvadratsumman (SS_{Tot}) enligt:

$$r^2 = 1 - \frac{SS_{Err}}{SS_{Tot}}.$$

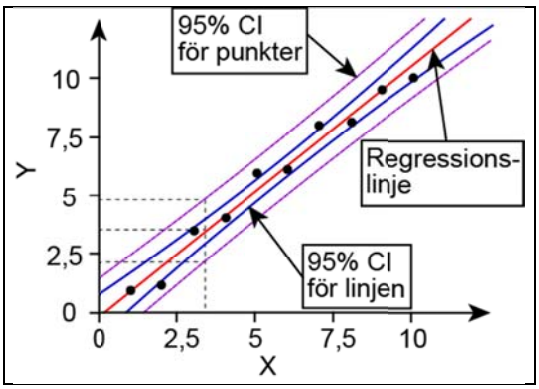
För att få ett högt värde på r^2 måste kvoten i ekvationen vara liten, d.v.s. avstånden i figur 22a måste vara mycket kortare än avstånden i figur 22c. Om man har ett svagt samband kommer avstånden i 22a och 22c vara nästa lika stora och därmed kommer kvoten bli nära 1, vilket leder till att r^2 blir lågt.

Konfidensintervall

Precis som för ett enda medelvärde kan man konstruera konfidensintervall för hela regressionen. Dessa intervall grundar sig på de teoretiska y-populationerna som illustreras i figur 23. Många datorprogram har rutiner för att presentera två olika konfidensintervall (figur 24). Det ena presenterar ett intervall som vid $\alpha = 0{,}05$ innefattar 95% av all lutningar som regressionslinjen kommer att få vid oändligt många (tänkta) upprepningar av experimentet.

Det andra intervallet är för de enskilda y-värdena och används vid prediktioner, (skattningar) av Y från X, och kallas därför ofta för prediktionsintervall. Vid $\alpha = 0{,}05$ kommer detta intervall att omfatta 95% av alla y-värden som motsvarar ett visst x-värde, d.v.s. större delen av den fördelning som illustreras i figur 22. Detta intervall används då man vill veta hur stor spridningen är kring den respons (d.v.s. y-värdet) som regressionen ger för ett visst x-värde.

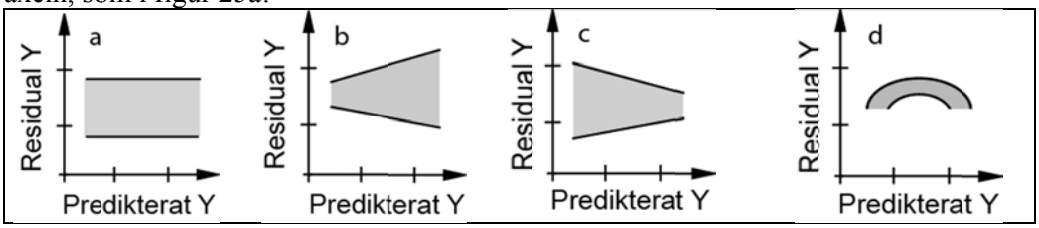
Det finns många synonyma namn på de konfidensintervall som beskrivits ovan. Konfidensintervall för regressionslinjen kallas bl.a.: ”confidence interval of the regression line”, ”confidence bands for slope” eller ”confidence curves fit”. Prediktionsintervall för enskilda skattningar kallas bl.a.: ”confidence of the prediction interval”, ”confidence bands for mean” eller ”confidence curves individual”



Figur 24. Illustration av konfidensintervall (CI) för regressionslinjen och för de enskilda datapunkterna. Medelrespons och 95%-igt konfidensintervall för x = 3,5 är markerat.

Residualanalys

En av förutsättningarna för att de ekvationer som används i regressioner ska ge korrekta värden är att variansen hos y-värdena är lika stor för alla x-värden. I avsnittet om datatransformation illustrerades hur variansen hos Y kan vara olika längs x-axeln (figur 6). Detta kallas heteroscedasticitet. Redan i 6a kan man se att spridningen ökar med ökande värden på X. För att illustrera detta tydligare brukar man göra ett s.k. residualdiagram (figur 6b). Residual är avståndet mellan regressionslinjen och de uppmätta värdena, d.v.s. d_i i figur 22a. I ett residualdiagram har man de Y-värden som regressionsekvationen ger på X-axeln och de residualerna (d_i i fig 22) på Y axeln. Om punkterna i detta diagram uppvisar ett mönster, oftast i form at en strut (figur 6b, 25b,c), kan man inte göra en linlär regression. Visserligen går det att låta en dator räkna, men resultatet kommer att vara felaktigt! För att en regression ska vara korrekt krävs att residualerna är slumpmässigt fördelade längs hela x-axeln, som i figur 25a!



Figur 25. Exempel på residualdiagram. a) homoscedastiska data, b och c) heteroscadastiska data, d) data som inte passar för linjär regression.

Det finns dock flera sätt att komma runt problemet med heteroscedasticitet. Det enklaste sättet är att transformera sina data som i exemplet i figur 6.

En titt på residualerna kan även ge annan information. Om residualerna visar en kurva, som i figur 25d, har man icke-linjära data. Om så är fallet är det felaktigt att försöka anpassa en rät linjes ekvation till ett samband som inte är rätlinjigt. Det finns dock flera sätt att anpassa kurvor till observerade samband. Antingen kan man använda polynomial regression eller icke-linjär regression. Båda dessa metoder gås igenom senare.

Förutsättningar

För att en linjär regression enligt minsta kvadratmetoden ska vara giltig måste ett antal krav vara uppfyllda:

1. Det måste finnas ett linjärt samband mellan en beroende y-variabel och en oberoende x-variabel. Detta samband måste vara ett s.k. orsakssamband, eller ett funktionellt samband.
2. För varje enskilt x-värde finns en teoretisk population av y-värden. Dessa populationer måste vara normalfördelade (figur 23).
3. Variansen måste vara lika stor hos alla y-populationer, s.k. homoscedasticitet.
4. X och Y måste vara oberoende i den meningen att X inte får ingå i Y.
5. X-variabeln får inte vara en stokastisk variabel, utan ska vara en kontrollerad variabel med fastställda nivåer. Detta krav är oftast inte uppfyllt, men för att få göra en regression räcker det med att veta att osäkerheten i mätningen av X-värdena är liten i förhållande till variationen hos Y-värdena.

Fixa, stokastiska och blandade modeller

Om inget av kraven i punkt 5 ovan är uppfyllda får man inte använda minsta kvadratmetoden för att beräkna värdet på regressionskoefficienterna! I detta avsnitt kommer vi att gå genom en metod för att ändå kunna utföra en regression.

Rubriken till detta avsnitt är densamma som rubriken till ett avsnitt i ANOVA-avdelningen. Denna upprepning kommer sig av den nära relationen mellan ANOVA och regression, och att man i båda teknikerna måste hålla reda på om man har fixa eller stokastiska variabler.

En av förutsättningarna för regression enligt minsta kvadratmetoden är att X-värdena är fixa, d.v.s. fastställda innan ett experiment startar (se punkt 5 under förutsättningen för regression). Denna förutsättning innebär bl.a. att man inte har någon slumpmässig variation hos de olika X-värdena i olika replikat eller i (tänkta) upprepningar av ett och samma experiment. Detta illustreras bl.a. i figur 23 där all variation finns i längs Y-axeln. X blir på detta sätt en fix variabel och Y en stokastisk. Detta är samma typ av modell som i en Modell III-ANOVA (jfr. tabell 15).

Inom många biologiska vetenskaper och speciellt inom miljöövervakningen är det dock ofta så att X-variabeln inte har fastställda nivåer. Ett exempel är sambandet mellan miljöpåverkan och antal taxa av bottenfauna i olika vattendrag. I detta exempel är miljöpåverkan utan tvekan en oberoende (X) och antal taxa en beroende (Y) variabel. Man har dock inte förutbestämda nivåer på X-variabeln som man skulle kunna ha i ett laboratorieexperiment. Istället får man de X-värden, liksom Y-värden, som observeras i naturen. Detta leder till att det finns slumpmässig variation hos både X och Y! Man har således två stokastiska variabler precis som i en Modell II-ANOVA.

Om man har denna typ av data och variationen är stor hos både X och Y, kan man inte använda minsta kvadratmetoden för att beräkna värdet på regressionskoefficienterna. Istället är man tvungen att använda ekvationer för s.k. Modell II-regression! Det finns flera metoder för Modell II-regression och här kommer vi att gå igenom den enklaste.

Lutningen hos regressionslinjen (här b’ för att indikera att det inte är samma lutning som b beräknad enligt minsta kvadratmetoden) beräknas enligt:

$$b' = \pm \frac{\text{standardavvikelsen för y}}{\text{standardavvikelsen för x}}.$$

Tecknet ± betyder att man själv måste avgöra om lutningen är positiv eller negativ. Kvoten mellan standardavvikelserna kommer alltid att vara positiv eftersom standardavvikelsen per definition alltid är positiv. Standardavvikelsen beräknas för alla värden i respektive stickprov, formeln för detta finns tidigare i detta kapitel.

Interceptet, a’ i regressionsekvationen, beräknas enligt:

$$a' = (\bar{y} - b' \times \bar{x}).$$

Efter att b’ och a’ är beräknade får man regressionsekvationen: Y = a’ + b’ × X, precis som vid minsta kvadratmetoden.

Det går inte att beräkna om b’, d.v.s. lutningen på regressionslinjen, är signifikant skild från 0. Man kan inte heller beräkna r^2 i denna typ av regression. Det är däremot korrekt att utföra en korrelation mellan X och Y och på så sätt ta reda på styrkan hos sambandet mellan X och Y.

Multipel regression

I de förra avsnittet behandlades regression mellan en oberoende och en beroende variabel. I detta avsnitt kommer vi att gå igenom hur man räknar om man har flera oberoende variabler som påverkar en responsvariabel. Detta är mer likt verkligheten än att bara en oberoende variabel påverkar den beroende variabeln.

Den modell (d.v.s. regressionsekvation) man använder är starkt relaterad till modellen för enkel linjär regression. Istället för en x-varibel lägger man till det antal x man vill pröva. Formeln blir då:

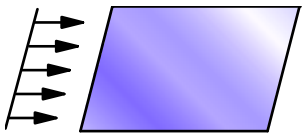
$$Y = a + b_1 \times X_1 + b_2 \times X_2 + b_3 \times X_3 + \ldots + b_n \times X_n,$$

där *a* och *b* är regressionskoefficienter.

Ett exempel är antalet taxa i bottenfauna i vattendrag som bl.a. beror av vattenkvaliteten vilken i sin tur är en kombination av en mängd variabler. I ett försök har man data på det totala antalet taxa bottenfauna från 30 olika vattendrag. Från varje provtagningsplats har man data på pH, och koncentrationerna av kalcium, magnesium och natrium. Man vill nu ta reda på hur vattenkvalitet, uttryckt i dessa kemivariabler, är relaterat till antalet taxa i bottenfauna. Detta löser man med hjälp av en multipel regression. Ekvationen som testas blir då:

Antal taxa = $a + b_1 \times pH + b_2 \times Ca + b_3 \times Mg + b_n \times Na$.

Eftersom detta är en ekvation med fyra termer är det omöjligt att rita ett diagram, men matematiskt sett konstruerar man ett diagram med fyra axlar och anpassar en linje genom den svärm av punkter som data ger upphov till, precis som i en enkel regression. Egentligen är det inte en linje utan ett s.k. plan som anpassas. Man kan tänka sig det hela som att linjen blivit utdragen till en duk som formas så att den ligger så nära varje punkt som möjligt.



När man utfört en multipel regression får man, precis som i en enkel regression, värden på r^2, P och regressionskoefficienterna (tabell 25). I exemplet blir $r^2 = 0,35$ och P =0,023. Detta gör att vi får regressionsekvationen:

Antal taxa = 1767 - 222 × pH - 85,1 × Ca + 5408 × Mg - 1732 × Na.

Men, även om r^2 och P indikerar att man har en modell som är signifikant (d.v.s. skild från ett mönster som skulle uppkomma av enbart slumpen) kan det vara så att några ingående X-variabler inte har någon inverkan på Y. Detta kontrollerar man genom att titta på P-värdena för regressionskoefficienterna (*b*) för de enskilda X-värdena (tabell 25).

Tabell 25. Regressionskoefficienter från en multipel regression mellan antal taxa av bottenfauna och vattenkemi. Urklipp från ett datorprogram.

Term	Estimate	Std Error	t Ratio	P
Intercept	1766,51	1347,89	1,31	0,2019
pH	-221,51	203,96	-1,09	0,2878
Ca mekv/l	-85,05	129,36	-0,66	0,5169
Mg mekv/l	5408,48	1494,28	3,62	0,0013
Na mekv/l	-1732,17	613,04	-2,83	0,0091

Dessa P-värden är resultaten av hypotesprövningar om huruvida varje enskild regressionskoefficient är skild från 0. Om regressionskoefficienten antar värdet 0 har den term som koefficienten gäller för ingen inverkan i regressionsekvationen, eftersom 0 gånger någonting blir 0. Av sista kolumnen i tabell 25 framgår att regressionskoefficienten för både pH och kalcium inte skiljer sig från 0, eftersom deras P-värden är högre än 0,05. pH och kalcium har därför inte någon inverkan på antalet taxa av bottenfauna i de vattendrag som undersökts. Därmed kan dessa variabler uteslutas i modellen.

Vi har nu konstaterat att pH och kalcium inte hade någon inverkan på antalet taxa, men vilken av de andra två variablerna har störst inverkan?

Detta får vi inte reda på genom att se på P-värdena. Dessa ger bara ett mått på sannolikheten att regressionskoefficienten är 0. Man kan inte heller se på värdet på regressionskoefficienterna, eftersom dessa värden hänger samman med vilka enheter man har i de olika variablerna. Det finns dock ett sätt att komma runt problemet med olika enheter: man standardiserar varje variabel så att alla får medel 0 och standardavvikelsen 1. Detta är en mycket vanlig standardisering och går till så att man subtraherar varje värde med medelvärdet och dividerar denna differens med standardavvikelsen, formeln för detta är:

$$X_{std} = \frac{x_i - \bar{X}}{std.av.(x)}.$$

Detta upprepar man för varje X-variabel. Efter standardiseringen kan man återigen utföra sin multipla regression (tabell 26). Trots transformationen av data är r^2 och P desamma som för originaldata. Likaså är P-värdena för regressionskoefficienterna desamma som tidigare. Däremot är det andra värden på själva regressionskoefficienterna. Dessa kan man nu jämföra och den variabel som har det högsta absolutvärdet på koefficienten har störst inverkan på Y-variabeln, vilket i detta exempel är magnesium.

Tabell 26. Regressionskoefficienter från samma multipla regression som i tabell 25, men här med standardiserade variabler.

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	498,57	79,37	6,28	<,0001
std_pH	-108,21	99,64	-1,09	0,2878
std_Ca	-62,77	95,47	-0,66	0,5169
std_Mg	622,32	171,93	3,62	0,0013
std_Na	-435,75	154,21	-2,83	0,0091

I detta exempel hade alla variabler utom pH samma enheter redan från början och därför får magnesium det högsta absolutvärdet på regressionskoefficienten i båda beräkningarna.

Stegvis multipel regression

I avsnittet ovan presenterades ett sätt att ta reda på vilka av flera X-variabler som har störst inverkan i en multipel regression. Ett mer elegant sätt är att använda **stegvis multipel regression**. I en sådan regression provar man varje X-variabel för sig och i den slutgiltiga modellen finns bara de signifikanta variablerna med. Det finns två huvudtyper av stegvis regression: ”forward selection” och ”backward elimination”.

I den första typen, ”**forward selection**”, startar man med en modell utan X-variabler. I det första steget inkluderas sedan den variabel som förklarar mest av variationen hos Y-variabeln. I steg två inkluderas den näst viktigaste variabeln o.s.v. Detta fortsätter tills det att alla variabler som uppfyller de signifikanskrav man ställt har tagits med i modellen.

I en stegvis regression på samma data som i exemplet med antal taxa hos bottenfauna och vattenkemi inkuderades magnesium in som första variabel i modellen. Därefter följde natrium och pH, medan kalcium inte togs med (tabell 27). I kolumnen RSquare visas hur r^2 ökar med antalet variabler i modellen, från 0,14 för bara magnesium till 0,34 med alla tre variablerna.

Tabell 27. Resultat av en stegvis regression med ”forward selection” på samma data som i tabell 25. Urklipp från datorprogram.

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Mallows Cp	p
1	Mg mekv/l	Entered	0,0443	998400	0,1367	7,3605	2
2	Na mekv/l	Entered	0,0207	1151689	0,2944	3,266	3
3	pH	Entered	0,1827	346514,4	0,3419	3,4323	4

Den andra metoden, ”**backward elimination**”, startar med en modell med alla X-variabler. Därefter tar man bort en variabel i taget av de variabler som förklarar minst av variationen hos Y, förutsatt att de bortplockade variablerna inte är signifikanta enligt en fördefinierad nivå.

I en stegvis regression med ”backward elimination” för samma exempel som ovan plockades kalcium bort i det första steget, och därefter pH (tabell 28). Den slutgiltiga modellen med denna teknik innehåller således bara två variabler och r^2 är 0,29.

Tabell 28. Resultat av en stegvis regression med ”backward elimination” på samma data som i tabell 25. Urklipp från datorprogram.

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Mallows Cp	p
1	Ca mekv/l	Removed	0,5169	81683,99	0,3419	3,4323	4
2	pH	Removed	0,1827	346514,4	0,2944	3,266	3

Som synes gav de olika metoderna olika modeller, en med tre variabler och en med två. Skillnaden ligger i att olika signifikansnivåer som valts för att plocka in variabler i ”forward selection” respektive för att ta bort variabler i ”backward elimination”. I detta exempel ser man att signifikansnivån för pH är så hög som 0,18, så den modell som erhölls m.h.a. ”backward elimination” är förmodligen mer korrekt.

Ytterligare sätt att undersöka hur många variabler som ska tas med i en stegvis multipel regression är att i ”forward selection” se till **Mallows Cp** (kolumn 7 i tabell 27). Detta värde ges automatiskt i många datorprogram. För att få en korrekt modell rekommenderas att Mallows Cp ska vara i ungefär lika med det antal parametrar (a och b_n) man tar med i modellen, vilket ges av p i tabell 27. I detta exempel är Cp-värdet för tre parametrar 3,2 vilket är den bästa överensstämmelsen mellan Cp och p. Enligt Mallows Cp bör man således välja modelen med två variabler (p är då 3 eftersom interceptet ingår i p).

Multikollinearitet

Om man tar med många oberoende variabler i en multipel regression finns det en stor risk att flera av dem är korrelerade och därigenom säger nästan samma sak. Om man inte är medveten om detta kommer det att ställa till problem när man försöker tolka resultaten från en multipel regression.

I exemplet ovan fanns pH och kalcium med som förklaringsvariabler. Normalt är dessa två variabler är starkt korrelerade, men detta var inte fallet i detta exempel. Däremot var magnesium och natrium starkt korrelerade (tabell 29).

Tabell 29. Korrelationer mellan de olika oberoende variablerna i exemplet i tabell 25.

	pH	Ca	Mg	Na
pH	1			
Ca	0,47	1		
Mg	0,44	0,41	1	
Na	0,22	0,24	0,83	1

Om man stoppar in korrelerade X-variabler (som alla har en signifikant inverkan på Y var för sig) i en multipel regression kommer resultaten visa att en eller några av de korrelerade X-variablerna är signifikant korrelerad med Y medan de övriga kommer att framstå som helt okorrelerade med Y.

För att göra en extra tydlig illustration av detta lägger vi till en konstruerad variabel till analysen som presenterades i tabell 26. Vi kallar den nya variabeln Mg_korr, där ”korr” står för att variabeln är mycket starkt korrelerad (r = 0,93) till magnesium (som var den viktigaste variabeln i modellen).

Vad som nu händer i en ny multipel regression är att natrium faller ut som den enda signifikanta variabeln (tabell 30). Regressionskoefficienterna för Mg och Mg_korr är nu inte signifikant skilda från 0. Om vi inte visste att Na, Mg och Mg_korr alla är starkt korrelerade skulle vi från tabell 30 tro att det endast var natrium som hade någon inverkan på antalet taxa av bottenfauna. Det är därför viktigt att göra korrelationsanalyser för alla par av oberoende variabler innan man gör en multipel regression. Om man hittar korrelerade X-variabler tar man bara med en av dem i regressionen. Vilken variabel som tas med beror av frågeställningen och korrelationen med Y. På detta sätt får man en starkare modell.

Ytterligare ett sätt att kontrollera om X-variablerna är korrelerade är att se till ”variance inflation factor, VIF”. Ju högre VIF-värden desto större sannolikhet att variabler är korrelerade och att de

därmed kan eller bör tas bort från analysen. I exemplet är VIF nästan en tiopotens högre för Mg och Mg_korr än för pH och Ca. Av detta bör man dra slutsatsen att magnesiumvariablerna är korrelerade. De flesta datorprogram har rutiner för att beräkna och presentera VIF.

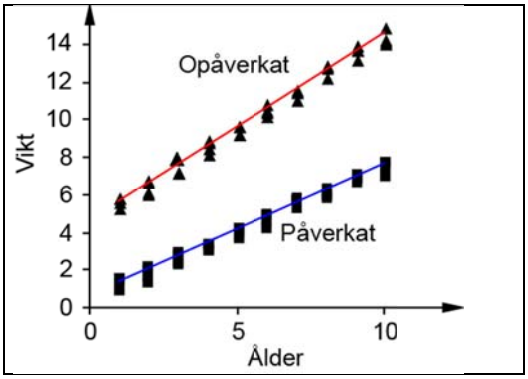
Tabell 30. Regressionskoefficienter från samma multipla regression som i tabell 26, men med ytterligare en variabel som är starkt korrelerad till magnesium.

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	498,57	80,22	6,22	<,0001	–
std_pH	-111,20	100,80	-1,10	0,2809	1,53
std_Ca	-44,16	100,23	-0,44	0,6635	1,51
std_Mg_korr	169,97	247,54	0,69	0,4989	9,21
std_Mg	477,38	273,43	1,75	0,0936	11,2
std_Na	-460,11	159,86	-2,88	0,0083	3,84

Att jämföra olika regressioner

Lutningen på regressionslinjen

Ibland vill man veta om lutningen på regressionslinjen i en regression skiljer sig från lutningen på regressionslinjen i en annan regression. Det kan t.ex. röra sig om samband mellan ålder och vikt hos en organism i påverkade och opåverkade sjöar (figur 26).



Figur 26. Samband mellan ålder och vikt hos en organism i påverkade respektive opåverkade sjöar.

Vid ett sådant test ställer man upp nollhypotesen:

H_0: Lutningen är densamma för de båda regressionslinjerna, $b_a = b_b$.

För att beräkna detta behövs följande:

	Förklaring	Opåverkat	Påverkat
$\sum x^2$	Summan av alla X-värden i kvadrat	1540	1540
$\sum xy$	Summan av alla x × y	2535,6	1173,3
$\sum y^2$	Summan av alla Y-värden i kvadrat	4340,6	897,28
Medel av x	Aritmetiskt medelvärde av alla x	9,99	4,25
r^2	Determinationskoefficienten	0,99	0,98
a	Interceptet	4,35	0,26
b	lutningen	1,02	0,72
SS_{resid}	$= \sum y^2 - \frac{(\sum xy)^2}{\sum x^2}$	165,7	3,32
Resid DF	Antal par - 2	38	38

Vidare behövs standardavvikelsen för skillnaden i lutning, som beräknas enligt:

$S_{b_{a-b_b}} = \sqrt{\frac{S^2_{Y\cdot X}}{(\sum x^2)_a} + \frac{S^2_{Y\cdot X}}{(\sum x^2)_b}}$, där $S^2_{Y\cdot X} = \frac{(SS_{resid})_a + (SS_{resid})_b}{(DF_{resid})_a + (DF_{resid})_b}$

När man erhållit värden på alla ovanstående komponenter kan man beräkna ett t-värde för nollhypotesen enligt:
$t = \frac{b_a - b_b}{S_{b_{a-b_b}}}$.
Därefter kan man gå in i en tabell och avläsa signifikansnivån vid α och ($n_1 + n_2$ - 4) frihetsgrader.

I exemplet får vi:

$$S^2_{Y\cdot X} = \frac{165{,}7 + 3{,}32}{38 + 38} = 2{,}22;\ S_{b_a - b_b} = \sqrt{\frac{2{,}22}{1540} + \frac{2{,}22}{1540}} = 0{,}054 \text{ och } t = \frac{1{,}02 - 0{,}72}{0{,}054} = 5{,}58$$

Via en t-tabell får fram vi att nollhypotesen ska förkastas om värdet på t överstiger 1,992 (α = 0,05, df = 76). Vi kan således förkasta nollhypotesen och konstatera att lutningen på de båda regressionslinjerna i figur 26 skiljer sig signifikant.

Interceptet

Om vi har förkastat nollhypotesen kan vi anta att vi har samlat två olika populationer. Om vi däremot inte hittade någon skillnad i lutning kan det vara idé att undersöka om interceptet, *a* i regressionsekvationen, skiljer sig mellan linjerna. D.v.s.:

H_0: interceptet för population a är detsamma som interceptet för population b, $a_a = a_b$.

Återigen kan man beräkna ett t-värde för att prova nollhypotesen. Detta gör man enligt:

$$t = \frac{(a_a - a_b) - b_{ab} \times (\bar{X}_a - \bar{X}_b)}{\sqrt{S^2_{Y\cdot X} \times \left(\frac{1}{n_a} + \frac{1}{n_b} + \frac{(\bar{X}_a - \bar{X}_b)^2}{(\sum x_a^2 + \sum x_b^2)}\right)}},$$

där b_{ab} är en gemensam lutningskoefficient för a och b som beräknas enligt:

$$b_{ab} = \frac{(\sum xy)_a + (\sum xy)_b}{(\sum x^2)_a + (\sum x^2)_b}.$$

Observera att denna beräkning endast är tillåten om man behållit nollhypotesen att lutningarna är lika, från förra avsnittet. För att illustrera det hela ändrar vi α i det förra avsnittet till 0,01. I och med detta är blir det kritiska värdet på t högre än det beräknade och vi kan behålla nollhypotesen. Vi kan nu testa om interceptet skiljer sig:

$$b_{ab} = \frac{2535{,}6 + 1173{,}3}{1540 + 1540} = 1{,}20, t = \frac{(4{,}35 - 0{,}261) - 1{,}20 \times (9{,}99 - 4{,}25)}{\sqrt{2{,}25 \times \left(\frac{1}{40} + \frac{1}{40} + \frac{(9{,}99 - 4{,}25)^2}{(1540 + 1540)}\right)}} = -7{,}57.$$

Absolutvärdet av detta t-värde jämförs sedan med t-värden i en t-tabell vid ($n_1 + n_2$ - 3) frihetsgrader. Vi kan konstatera att vårat t-värde på 7,57 är långt över det kritiska värdet för α = 0,0001. Intercepten skiljer sig således signifikant!

Det finns även tester för att jämföra fler är två lutningar eller intercept, men i detta kompendium går vi inte in närmare på detta utan hänvisar till en lärobok i ämnet., t.ex. Sokal & Rohlf (1995).

Icke-linjära samband

Hittills har vi behandlat metoder som kräver ett linjärt samband mellan X och Y, men hur gör man när sambandet inte är linjärt? Eftersom det inte är ovanligt med samband som inte är linjära har statistiker naturligtvis utarbetat flera sätt att lösa även denna typ av problem. De vanligaste är:

- polynomial regression,
- transformation av data så de blir linjära
- icke-linjär regression.

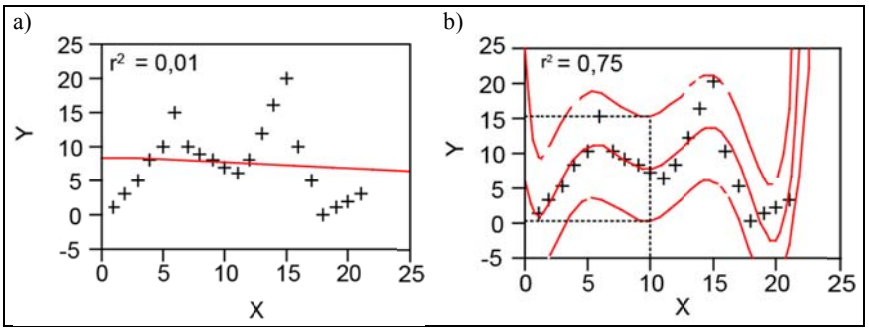
Polynomial regression

Grunden i en polynomial regression är samma ekvation som i multipel linjär regression, men de olika X-variablerna i den multipla regressionen är ersatta av polynom av den första X-variabeln, enligt:

$$Y = a + b_1 \times X + b_2 \times X^2 + b_3 \times X^3 + \ldots + b_n \times X^n.$$

Med polynom menas således den studerade X-variabeln upphöjd till någon faktor (2 och 3 upp till n i exemplet ovan). Trots att ekvationen ger en kurva är den linjär till sin natur och därför gäller samma förutsättningar som för alla annan linjär regression. Vi har bara en X-variabel, men den förekommer flera gånger fast som polynom av sig själv. Med en polynomial regression kan man anpassa vilken form av kurva som helst till de data man har. Det enda som begränsar är antalet observationer. Man bör dock undvika höga polynom eftersom de funktioner som dessa ger upphov till är osäkra. Inom biologin är det ytterst sällan man använder högre än tredjegradspolynom.

I figur 27 illustreras en enkel och en polynomial linjär regression på samma data. Den enkla regressionen hittar inte något samband medan den polynomiala uppnår ett r^2-värde på 0,75. Som synes är dock skattningar av Y från regressionskurvan i figur 27b mycket osäkra. Vid t.ex. X = 10 ligger det predikterade Y-värdet någonstans mellan 0,25 och 15,1, med $\alpha = 0,05$.



Figur 27. Linjär (a) och 6:e grads polynomial (b) regression på samma data. Den streckade linjen i b är 95%-igt prediktionsintervall.

Vägledning om hur många polynom som ska tas med får man genom att se till P-värdet för *b* i regressionsekvationen. När man använder denna teknik startar man med en kvadratisk regression och ser om b_1 och b_2 är signifikanta. Om b_2 är signifikant lägger man till X^3-termen och kontrollerar P-värdet för b_3. Detta upprepas tills det att *b* inte längre är signifikant. Det finns dock ett problem med denna metod; om man har komplicerade samband som i exemplet i figur 27. kan *b* vara osignifikant för alla polynom till och med tredje graden, men sedan bli signifikant för ett fjärdegradspolynom.

"Adjusted r^2"

I exemplet för multipel regression ansåg vi att den modell som hade två termer var den mest korrekta. Detta trots att modellen med tre termer hade ett högre r^2-värde, och modellen med fyra termer ännu högre r^2. Anledningen till att vi valde en modell med två termer berodde bl.a. på P-värdena för de olika X-variablerna. Som synes ökar r^2 med ökat antal termer, oavsett om termen är signifikant eller ej. Av denna anledning kan man inte använda r^2 för att bedöma vilken av olika modeller som är den bästa.

För att veta vilken modell som är den bästa i fråga om förklaringsgrad ser man istället till det justerade r^2-värdet, ”adjusted r^2” på engelska. Detta r^2-värde är justerat för antal termer i regressionsekvationen. Om man lägger till fler och fler termer i en regressionsekvation kommer, som sagt, r^2 att öka hela tiden medan det justerade r^2-värdet kommer att nå ett optimum och sedan minska (tabell 31). Det högsta antal termer man ska ta med är det antal som ger det högsta värdet på det justerade r^2-värdet. Detta gäller både multipel regression och polynomial regression!

För att illustrera effekten av antalet variabler på r^2 lägger vi till tio rena slumpvariabler till data i exemplet för multipel regression. I en ny multipel regression med alla tolv termer ökade r^2 till 0,6 istället för 0,29 som man fick med endast de två signifikanta kemivariablerna. Det justerade r^2-värdet är dock 0,22, vilket är lägre än 0,25 som var det justerade r^2-värdet för modellen med två variabler (tabell 31). P-värdet för modellen med tio slumpvariabler blev dock 0,18 vilket indikerar att modellen som helhet inte är signifikant. En stegvis regression med ”forward selection” gav tre variabler i den slutgiltiga modellen, varav en var en slumpvariabel. I denna regression blev r^2 = 0,34, ”adjusted r^2” = 0,32 och P = 0,0046 (tabell 31). Man kan således konstatera att modellen med två kemivariabler och en slumpvariabel matematiskt sett är den bästa att förklara variationen hos antal taxa av bottenfauna. (I verkligheten ska man givetvis inte lägga till slumpvariabler. I detta exempel har det bara råkat bli så att en av de tio slumpvariablerna var korrelerad till antalet taxa, vi har således blivit lurade av slumpen.)

Tabell 31. Illustration av hur r^2 ökar med antal termer, men att ”adjusted r^2” har ett optimum

Modell	r^2	Adjusted r^2	P
Original (2 termer)	0,29	0,25	0,0090
Original + 1 slumpvariabel	0,34	0,32	0,0046
Original + 10 slumpvariabler	0,60	0,22	0,18

Transformationer

Nästa grupp av metoder vi tar upp för att utföra regression på icke-linjära samband är olika transformationer som leder till att linjär regression kan användas.

I avsnittet om datatransformation presenterades några olika sätt att transformera data (tabell 2). Genom att tillämpa någon av de transformationer som nämndes där kan man åstadkomma ett linjärt samband av en icke-linjär funktion, och därigenom använda linjär regression. I vissa fall kan man genom logaritmlagarna omvandla ickelinjära samband till linjära. Detta gäller t. ex. modellen för exponentiell tillväxt, $Y = a \times e^{bX}$, som kan skivas om till:

$$\log(Y) = \log(a) + b \times \log(X).$$

I detta fall kan man använda linjär regression om modellen uppfyller alla förutsättningar för detta (se ovan). Om man inte kan utföra en linjär regression, t.ex. på grund av heteroscedastisitet efter transformation eller att modellen inte går att transformera till en linjär form, använder man istället icke-linjär regression (se nästa avsnitt).

Box-Cox-transformation

Det finns många metoder för transformation. I detta avsnitt kommer vi kort att beröra en familj av transformationer som kallas Box-Cox-transformation. Namnet kommer av att G.E.P. Box och D.R. Cox presenterade metoden i en publikation 1964. Box-Cox-transformation finns som standard hos de flesta datorprogrammen.

I Box-Cox-transformationer börjar man med att transformera Y-variabeln enligt:

$$y_\lambda = \begin{cases} \frac{y^\lambda - 1}{\lambda}, \text{ om } \lambda \neq 0 \\ \ln(y), \text{ om } \lambda = 0 \end{cases}$$

där λ oftast varierar mellan -2 och 2. Sedan gör man en regression för varje värde på λ inom det intervall man prövar. Oftast provar man λ = -2, -1,5 o.s.v., i steg om 0,5 upp till λ = 2.

För att ta reda på vilket av alla värden på λ som ger den bästa transformationen av de data man har använder man en metod som kallas ”maximum likelihood” eller maximimetoden. Vi går inte in på matematiken bakom detta utan nöjer oss med att presentera formeln:

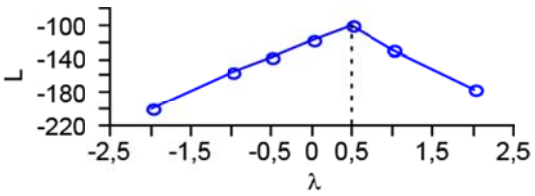
$$L_\lambda = -\frac{n}{2} \times \ln(MSE_\lambda) + (\lambda - 1) \times \sum \ln(y_i)$$

där MSE är ”Mean Square Error” som fås från en enkel linjär regression (ur ANOVA-tabellen), *n* är antalet par av observationer och y_i är de enskilda Y-värdena.

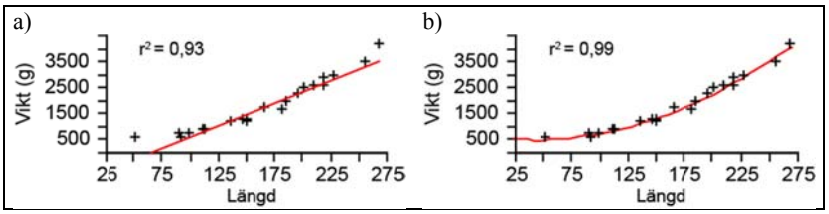
När man beräknat värdet på L för alla olika λ väljer man det λ-värde som ger det högsta värdet på L. Det enklaste sättet att avgöra vilket värde på λ som är det bästa är att göra ett diagram med L mot λ (tabell 32).

Tabell 32. ”Mean Square Error” från enkla linjära regressioner på samma data som i figur 28, men med Box-Cox-transformerade y-värden med olika värden på λ, och korresponderande värde på L enligt maximimetoden. Till höger en illustration av data i tabellen.

λ	MSE	L
-2	$6{,}27 \times 10^{-9}$	-200
-1	$2{,}59 \times 10^{-5}$	-157
-0,5	$2{,}05 \times 10^{-3}$	-137
0	0,157	-118
0,5	15,2	-101
1	$76{,}1 \times 10^{3}$	-129
2	$8{,}05 \times 10^{11}$	-175



Ett exempel får illustrera det hela. Sambandet mellan längd och vikt hos en organism ser ut som illustrerat i figur 28. Genom att titta på data ser man tydligt att de inte passar för en enkel linjär regression (figur 28a). En andragrads polynomial regression verkar däremot fungera utmärkt (figur 28b). I polynomiala regressioner transformerar man X-värdena så att data ska passa ekvationen för linjär regression. Med Box-Cox-transformationer ändrar man istället Y-värdena så att data ska passa en linjär modell.



Figur 28a, b. Samband mellan längd och vikt hos en organism. Enkel (a) och 2:a grads regression (b).

Om vi sätter in data på *n* och y_i blir formeln för $L_\lambda = -11.5 \times \ln(\mathrm{MSE}_\lambda) + (\lambda - 1) \times 139.2$. När vi väl har denna formel är det bara att prova med olika värden på λ. Det första man gör är att transformera alla *y* enligt formeln ovan till y_λ. För λ = -2 blir $y_\lambda = (y^{-2} - 1)/-2$. Därefter gör man en enkel linjär regression (tabell 33) för att få reda på värdet på ”Mean Square Error”. Detta upprepas sedan för alla värden på λ som man vill prova, och för varje λ beräknar man L (tabell 32).

Tabell 33. Resultat från en enkel linjär regression på data i figur 28, men med Box-Cox-transformerade y-värden (λ = -0,5).

ANOVA table	DF	Sum of Squares	Mean Square	F-Value	P-Value
Regression	1	19095,239	19095,239	1252,14	<0,0001
Residual	19	289,752	15,250		
Total	20	19384,990			

Coefficients	Coefficient	Std. Error	Std. Coeff.	t-Value	P-Value
Intercept	-25,639	2,624	-25,639	-9,771	<0,0001
X	0,527	0,015	0,992	35,386	<0,0001

Som framgår av tabell 32 och figuren bredvid tabellen är det λ = 0,5 som ger det högsta värdet på L. Vi kan nu kombinera den transformation vi gjort av y med λ = 0,5 med den regressionsekvation som vi fick för λ = 0,5 (tabell 33). Vi får då:

$\frac{y^{0,5}-1}{0,5} = -25,6 + 0,527 \times x$. Genom att lösa ut *y* ur denna ekvation* får vi:

$y = 139,24 - 6,219 \times x + 0,694 \times x^2$,

vilket är den linjära ekvation som beskriver sambandet mellan X och Y efter Box-Cox-transformering med λ = 0,5. Denna ekvation kan jämföras med ekvationen som andragradsregressionen (figur 28b) gav:

$y = 128,8 - 6,276 \times x + 0,072 \times x^2$.

Hur vet man då vilken av de två regressionerna som bör väljas?

Svaret på denna fråga får man genom att jämföra F-koterna för de två regressionerna, d.v.s. andragradsregressionen och regressionen med Box-Cox-transformerade y-data med λ = 0,5. Denna jämförelse låter sig göras eftersom båda F-kvoterna gäller samma data. I detta exempel är F-kvoten för andragradsregressionen 799 vilket ska jämföras med 1252 som är F-kvoten för Box-Cox-regressionen (med λ = 0,5). En högre F-kvot innebär en bättre modell (se utförligt resonemang i ANOVA-avsnittet) och vi kan konstatera att Box-Cox-transformationen gav en bättre regressionsekvation än andragradsregressionen.

*) y löses ut m.h.a. följande steg:

$\frac{y^{0,5}-1}{0,5} = -25,6 + 0,527 \times x \Rightarrow$

$2 \times (\sqrt{y} - 1) = -25,6 + 0,542 \times x \Rightarrow$

$y = \left(\left(\frac{-25,6 + 0,527 \times x}{2}\right) + 1\right)^2 = \left(\frac{-25,6}{2} + \frac{0,527 \times x}{2} + 1\right)^2 =$

$= \left(\frac{-25,6}{2} + 1\right)^2 + \left(\frac{0,527 \times x}{2}\right)^2 - 2 \times \left(\frac{-25,6}{2} + 1\right) \times \left(\frac{0,527 \times x}{2}\right) =$

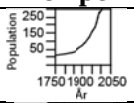
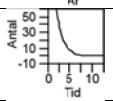
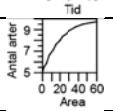
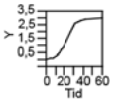
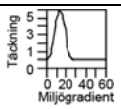
$= 139,24 + 0,694 \times x^2 - 6,219 \times x =$

$= 139,24 - 6,219 \times x + 0,694 \times x^2$

Icke-linjär regression

I detta sista avsnitt under huvudrubriken ”Icke-linjära samband” kommer vi att kort gå igenom en typ av regression som inte bygger på den linjära modell vi använt hittills. Om man har data som inte ens efter en transformation kan beskrivas med en rät linje, är man tvungen att använda andra typer av modeller för att beskriva sambandet mellan variablerna. Det finns många typer av icke-linjära samband och några är presenterade i tabell 34.

Tabell 34. Exempel på olika icke-linjära modeller.

Typ	Modell	Exempel	Beskrivning
Exponentiell tillväxt	$Y = a \times e^{bX}$.		Populationstillväxt
Exponentiellt avklingande	$Y = a \times e^{-bX}$.		Mortalitet
Asymptotiskt regression	$Y = a - b_1(e^{-b_2X})$		Kumulativa samband, t.ex. art-areakurvor
Logistisk tillväxt	$Y = \frac{a}{\left(1+b_1\left(e^{-b_2X}\right)\right)}$		Populationstillväxt vid predationstryck
Gaussformad logistisk	$Y = \frac{e^{(b_0+b_1\times X+b_2\times X^2)}}{1+e^{(b_0+b_1\times X+b_2\times X^2)}}$		En arts fördelning längs en miljögradient

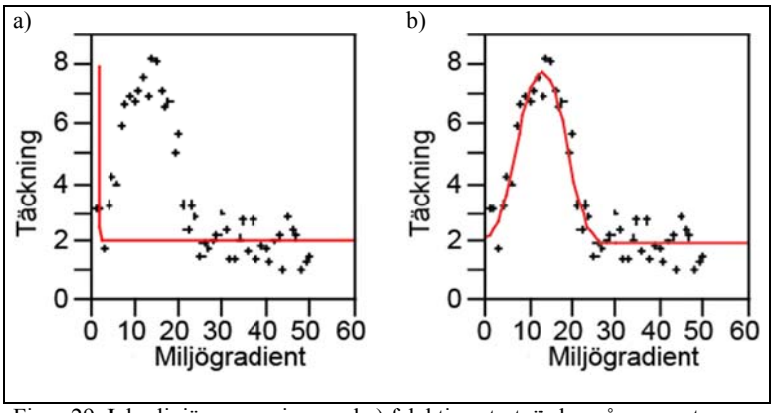
En viktig del av icke-linjär regression är att välja rätt modell för sina data. Detta görs nämligen inte automatiskt av det datorprogram man använder. Anledningen till detta är att val av modell inte är ett matematiskt utan ett vetenskapligt problem. Den modell som väljs måste vara grundad på teorin omkring de data man har. I de flesta datorprogram som klarar av icke-linjär regression skriver man in den modell man vill använda för hand.

Principen bakom både icke-linjär och linjär regression är att använda sina data för att hitta de bästa möjliga värdena på parametrarna i regressionsekvationen (a och b_n). När man hittat dessa får man den ekvation, eller modell, som på bästa sätt beskriver sambandet mellan X och Y. Det som skiljer de båda typerna av regression är dels att parametrarna inte adderas utan multipliceras i icke-linjär regression, och dels sättet att beräkna parametrarna.

I icke-linjär regression använder man inte minsta kvadratmetoden som i ANOVA och linjär regression för att komma fram till värden på parametrarna. Istället används en s.k. iterativ procedur som är så beräkningsintensiv att det är i stort sett omöjligt att utföra denna typ av regressioner för hand. Med iterativ procedur menas att programmet startar med godtyckliga värden på regressionsparametrarna (a och b_n i tabell 34) och från dem beräknas nya bättre värden, och från dessa beräknas ännu bättre värden, o.s.v. Detta fortsätter tills det att ytterligare steg inte leder till förbättrade värden på parametrarna, inom vissa gränser. De flesta datorprogrammen kräver att man själv ger startvärden på parametrarna. Här bör man välja värden som är någorlunda rimliga. Om startvärdena är helt fel kommer de upprepade beräkningarna inte leda till bättre värden på parametrarna. Detta kallas att iterationen inte konvergerar. Om detta skulle inträffa får man besked om detta av datorprogrammet.

Efter att man utfört beräkningarna har man i bästa fall anpassat en kurva till de data man har. Hur vet man då om anpassningen är bra eller ej. När man arbetar med icke-linjär regression kan man inte bara se till r^2-värdet. Det är till och med så att många program inte ger något värdet på r^2 eftersom det inte är ett bra mått på kurvanpassningen i icke-linjär regression. Således måste man se till andra mått.

En första kontroll är att förvissa sig om att iterationerna konvergerade. Om inte, får man prova en gång till, men med andra startvärden på parametrarna. Om detta lyckas är det dags för en visuell bedömning av resultatet. Återigen kan resultatet bli helt felaktigt om man valt fel parametrar (figur 29a). En annan anledning till att regressionslinjen hamnar långt från datapunkterna kan vara att man valt en felaktig modell.



Figur 29. Icke-linjär regression med a) felaktiga startvärden på parametrarna och b) korrekta startvärden.

Om man kan konstatera att regressionskurvan ser ut att vara korrekt, som i figur 29b, kan man gå vidare. Nästa steg blir att undersöka P-värden för de olika parametrarna. Om en parameter inte är signifikant skild från 0 är parametern överflödig och kan tas bort, eller så har man valt en felaktig modell. En del program ger inte P-värden för parametrarna utan istället får man ett konfidensintervall. I dessa fall måste hela konfidensintervallet vara högre eller lägre än 0 för att man ska kunna säga att parametern i fråga är signifikant skild från 0. Ytterligare sätt att kontrollera om en parameter är överflödig är att se till korrelationen mellan parametrarna. Om två parametrar är mycket starkt korrelerade finns anledning att fundera om någon av dem är överflödig. Om så är fallet kan man antingen behålla den modell man har eller så börjar man om med en ny modell. Man ska sträva efter att få en modell med så få parametrar som möjligt, men om man på teoretiska grunder vet att man valt rätt modell behåller man modellen även om anpassningen av de data man har indikerar att modellen har för många parametrar.

Logistisk regression

Alla regressioner vi gått igenom hittills har förutsatt att både X och Y är kontinuerliga variabler på lägst intervallskalnivån. Här kommer vi att kort gå igenom principen för regression när prediktorvariabeln, X, är kontinuerlig, men responsvariabeln, Y, är **nominal**, t.ex.:

• närvarade eller frånvarade
• hane eller hona

I dessa fall kan man inte använda vanlig regression. Istället får man använda s.k. **logistisk regression** som är en variant av den vanliga regressionen. Efter några matematiska operationer kan man anpassa en nominal Y-variabel till en linjär funktion av en eller flera X-variabler.

Eftersom X-variablerna fortfarande är kontinuerliga kan man uttrycka dem med den linjära modellen som vi känner från tidigare avsnitt. Vi få då:

$Y = a + b \times X$.

Responsvariabeln, Y, kan dock inte användas i den form som föreligger eftersom den inte är kontinuerlig. De olika klasserna hos Y-variabeln kodas ofta som 0 eller 1, men detta är bara namn på klasserna. Man skulle lika gärna kunna ge namn som ”Hane” och ”Hona”. Om man använt namnen ”0” och ”1” kan man inte använda dessa siffror som om de vore uppmätta responser även om man lätt kan lura ett datorprogram till att tro att det är mätvärden och inte namn!

För att kunna ha med en nominal Y variabel i en regression måste man således göra någon form av transformering. Det är här det logistiska i ”logistisk regression” kommer in. Y-variabeln görs om till en kvot mellan sannolikheten för 1 och sannolikheten för 0, enligt

$\ln\left(\frac{p}{1-p}\right)$,

där p är sannolikheten för ”1” och 1-p följaktligen sannolikheten för ”0”. Kvoten p/(1-p) kallas oddskvot (odds ratio, på engelska) och naturliga logaritmen för denna kvot kallas ”logit” vilket givit tekiken dess namn.

Genom att kombinera de två ovanstående ekvationerna får vi den logistiska regressionsmodellen:

$$\ln\left(\frac{p}{1-p}\right) = a + b \times X$$

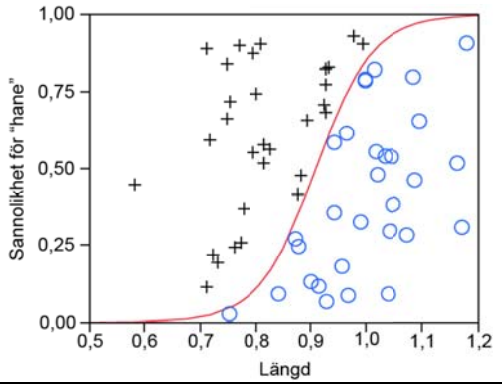
Denna modell kan skrivas om till en modell för sannolikheten för att Y-variabeln ska vara ”1” enligt:

$$\ln\left(\frac{p}{1-p}\right) = a + b \times X \Rightarrow \frac{p}{1-p} = e^{a+b\times X} \Rightarrow p = e^{a+b\times X} - p \times e^{a+b\times X} \Rightarrow \frac{p}{p} = \frac{e^{a+b\times X}}{p} - e^{a+b\times X} \Rightarrow 1 + e^{a+b\times X} = \frac{e^{a+b\times X}}{p} \Rightarrow$$
$$p = \frac{e^{a+b\times X}}{1 + e^{a+b\times X}}$$

vilket är en icke-linjär regressionsmodell (jfr. förra avsnittet).

Denna modell kommer att resultera i en S-formad kurva (=regressionslinjen) som är begränsad mellan 0 och 1 (figur 30). För varje X-värde ger avståndet under kurvan sannolikheten, p, för att Y är 0, och avståndet över kurvan sannolikheten att Y är 1 (eller tvärt om beroende på hur man kodat sina data). Om termen (a + b × X) i den logistiska regressionsmodellen har ett högt värde kommer p att närma sig 1 och vid ett lågt värde kommer p att närma sig 0.

Kön (Y)	Längd (X)
Hane	0,92
Hane	0,99
Hane	1,05
Hane	0,90
…	…
…	…
Hona	0,76
Hona	0,79
Hona	0,81
Hona	0,71



Figur 30. Exempel på en logistisk regression där man vill predicera kön från längd. Längst till höger ett utdrag ur datatabellen. Kors indikerar honor och ringar hanar. Regressionslinjen visar sannolikheten för ”Hane”, vilket är detsamma som sannolikheten för ”inte Hona”. I diagrammet är Y-värdet för varje observation är slumpat, medan X-värdet är den observerade längden.

Beräkningarna i en logistisk regression bygger precis som i icke-linjär regression på den s.k. maximimetoden (maximum likelihood method). Trots att beräkningarna skiljer sig från minsta kvadratmetoden som används vid vanlig regression presenteras resultaten i en tabell som påminner om ANOVA-tabellen.

I tabell 35 presenteras resultaten från regressionen i figur 30. Värdet på ”–LogLikelihood” motsvarar kvadratsumman i vanlig regression. Värdet som testas är skillnaden (*Difference* i tabellen) mellan den modell som erhålls om sannolikheten var lika stor över hela X-skalan (*Full*) och den modell som erhålls genom att anpassa data (*Reduced*). Testvärdet är chi-två-fördelat och P-värdet ger (något förenklat) sannolikheten att nollhypotesen är sann. Nollhypotesen i en logistisk regression är:

H_0: Sannolikheten för ”0” är 0,5 för alla X-värden.

Tabell 35a, b, c. Resultat från den logistiska regressionen presenterad i figur 30.

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	19,237180	1	38,47436	<0,0001
Full	22,351651			
Reduced	41,588831			

RSquare (U)	0,4626

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-17,577248	4,3405615	16,40	<,0001
Längd	19,3435094	4,736286	16,68	<,0001

For log odds of Hane/Hona

Determinationskoefficienten, R^2, ett mått på hur mycket bättre modellen blivit genom att ta med data, jämfört med att man har en modell som sätter sannolikheten för ”0” till 0,5. Det är således inte samma r^2 som i vanlig regression. Ofta skriver man r^2 när man använt minsta kvadratmetoden och R^2 när man använt maximimetoden.

Tabellen med värden för de olika parametrarna ger förutom parametervärdena även standardavvikelse och P-värden för parametrarna. Här gäller samma principer som för presentation av parametrar i icke-linjär regression i avsnittet innan.

Detta var en introduktion till logistisk regression. Exemplet var den enklaste tänkbara modell man kan testa m.h.a. denna teknik. Det är fullt möjligt att ha fler nivåer än två hos responsvariabeln. Dessutom kan man ha fler X-variabler som i en multipel regression.

Chi-två-tester och kontingenstabeller

Chi-två-tester

I det förra avsnittet gick vi igenom en metod som byggde på att en variabel är nominal och en kontinuerlig. I detta avsnitt kommer vi att ta upp olika tekniker som används då alla variabler är nominala. Dessa tekniker sammanfattas under det gemensamma namnet chi-två-tester och omfattar tester för homogenitet, slumpmässighet, association, oberoende och anpassning till en fördelning (goodness of fit). Trots mångfalden av tekniker bygger alla på samma princip: man jämför den observerade fördelningen med den fördelning man skulle förvänta om nollhypotesen är sann. Resultatet sammanfattas i en testvariabel, χ^2, som kan jämföras med tabellerade värden på χ^2 för att avgöra om nollhypotesen ska behållas eller förkastas.

Oftast gör man en sammanställning av resultaten i en s.k. kontingenstabell där antalen av de olika observerade kategorierna noteras (tabell 36). En kontingenstabell sammanfattar således data från studier där resultatet är olika kategoriska variabler, t.ex. hona eller hane eller olika tillståndsklasser eller avvikelseklasser i Naturvårdsverkets bedömningsgrunder.

Tabell 36. Antal sjöar i olika tillståndsklasser för totalfosfor enligt Naturvårdsverkets bedömningsgrunder för miljökvalitet, i norra och södra Sverige. Data från riksinventeringen av sjöar och vattendrag 2000.

	Tillståndsklass					Summa
Område	1	2	3	4	5	
Norra Sverige	1016	170	54	13	2	1255
Mellersta Sverige	1610	614	265	129	71	2689
Sydligaste Sverige	65	88	57	20	15	245
Summa	2691	872	376	162	88	4189

De frågor man kan besvara med hjälp av data från en kontingenstabell är av typen: är någon kombination av de klasser jag studerar överrepresenterad, t.ex. är det vanligare med tillståndsklass 1 i norr än i söder? Detta kommer att yttra sig som att data är koncentrerade till några av cellerna i tabellen. De statistiska hypoteserna blir:

H_0: Inom en kolumn är sannolikheten att en observation ska hamna i en viss rad är densamma för alla rader.

H_1: Inom en kolumn har minst två rader olika sannolikheter för att en observation ska hamna i de raderna.

Dessa hypoteser testas i ett s.k. chi-två-test (χ^2). Det första man gör i ett sådant test är att beräkna de frekvenser som skulle förekomma om nollhypotesen är sann, s.k. förväntade frekvenser (Expected frequency, E). För att beräkna de förväntade frekvenserna börjar man med att se till den fördelning sjöarna har mellan de olika tillståndsklasserna oavsett var i Sverige sjön ligger. Detta gör vi för att få reda på hur många sjöar i respektive klass vi har med i undersökningen. I exemplet är t.ex. andelen sjöar i klass 1 2691/4189 = 0,64, och andelen sjöar i klass 2 872/4189 = 0,21, o.s.v.

Nu vet vi hur vi samlat sjöar i olika tillståndsklasser. Nästa steg är att se hur antalet sjöar inom respektive klass skulle fördela sig mellan de tre områdena, om nollhypotesen är sann. Bland de 1255 sjöarna i norra Sverige borde 64% finnas i tillståndsklass 1, eftersom 64% av alla sjöar oavsett region befanns tillhöra klass 1. På motsvarande sätt borde 21% av de 1255 sjöarna i norr vara i klass 2 om nollhypotesen är sann. 64% av 1255 sjöar är detsamma som 0,64 × 1255 = 806,21 stycken sjöar. Detta är den förväntade frekvensen om nollhypotesen är sann. Detta värde och övriga förväntade frekvenser i exemplet är ifyllda i tabell 37. En allmän formel för att beräkna förväntade frekvenser blir:

$$E_{r,k} = \frac{n_k}{N} \times n_r,$$

där n_k = totala antalet observationer i kolumn *k*, n_r = totala antalet observationer i rad *r* och N = totala antalet observationer.

Tabell 37. Förväntade frekvenser från tabell 36, om nollhypotesen är sann.

	Tillståndsklass				
Område	1	2	3	4	5
Norra Sverige	806,21	261,25	112,65	48,53	26,36
Mellersta Sverige	1727,4	559,75	241,36	103,99	56,49
Sydligaste Sverige	157,39	51,00	21,99	9,47	5,15

Trots att de observerade frekvenserna bara kan var heltal ger man bråkdelar i förväntade frekvenserna eftesom detta är en del av den statistiska metoden och inte något man observerar.

För att testa nollhypotesen beräknar man som i övrig hypotesprövning värdet på en testvariabel. I detta fall kallas testvariabeln χ^2. Testvariabeln beräknar man enligt:

$$\chi^2 = \sum \frac{(O-E)^2}{E},$$

där O betyder observerad frekvens. Denna variabel är s.k. chi-två-fördelad med (r-1)×(k-1) frihetsgrader, där r = antalet rader och k = antalet kolumner. I detta fall har vi (3-1)×(5-1) = 8 frihetsgrader.

För att undersöka om vi ska behålla eller förkasta nollhypotesen jämför vi det erhållna värdet på χ^2 med det kritiska värdet man får ur en χ^2-tabell vid 8 frihetsgrader. I detta fall ska vi förkasta nollhypotesen om det beräknade värdet på χ^2 är högre än tabellvärdet som är 20,09 vid $\alpha = 0{,}01$.

Om vi stoppar in våra värden i formeln får vi:

$$\chi^2 = \frac{(1016-806{,}21)^2}{806{,}1} + \frac{(1610-1727{,}4)^2}{1727{,}4} + \cdots + \frac{(15-5{,}15)^2}{5{,}15} = 358{,}18.$$

Detta χ^2-värde för våra data är mycket högre än det kritiska värdet på χ^2 för 8 frihetsgrader och $\alpha = 0{,}01$. Vi kan således förkasta nollhypotesen och konstatera att åtminstone någon rad är överrepresenterad i någon kolumn. Detta kan uttrycka som: ”…fördelningen mellan olika tillståndsklasser var inte densamma i de olika delarna av landet…”.

Hur vet man då vilken eller vilka celler i datatabellen som avviker från det förväntade värdet, och som därmed orsaker ett högt värde på testvariabeln (χ^2)?

Förutsättningar

En av de viktigaste förutsättningarna för ett chi-två-test är att de förväntade frekvenserna blir tillräckligt höga. Ett generellt råd från statistiker är att ingen förväntad frekvens bör understiga 5. Ett krav är att ingen förväntad frekvens understiger 1, och att högst 20% av de förväntade frekvenserna understiger 5.

Som vid alla statistiska undersökningar måste proven vara oberoende av varandra, d.v.s. insamlandet av ett datum får inte påverka vilket värde nästa datum får (datum är singularformen av data).

Alla datum i en chi-två-analys måste vara kategoriska. Beräkningarna måste sedan ske på frekvenser. Om man gör en tabell med andelar eller procent och utför ett chi-två-test på en sådan tabell kommer resultat bli helt felaktigt!

Ovanstående är ett typexempel på chi-två analyser. Vi kommer nu att gå igenom några av de olika varianter som finns.

Test av frekvenser

Detta test innebär att man testar om antalet observationer i olika kategorier skiljer sig från en förväntad fördelning. Det kan antingen vara en jämn fördelning eller någon annan känd fördelning mellan kategorierna. Skillnaden mot ovanstående exempel är att man inte beräknar de förväntade frekvenserna från de observerade, utan använder den kända fördelning mellan klasser. Nollhypotesen blir:

H_0: Fördelningen hos på insamlade data skiljer sig inte från den teoretiska fördelningen.

Blommorna hos kungsängslilja *Fritillaria meleagris* har tre olika färger; purpur, vit och en mellanform. Enligt en tidigare undersökning förekommer de tre formerna i proportionen 80% purpur,15% vita och 5% i mellanformen. Antag att man vill testa detta genom en ny undersökning. Detta gör man genom att notera blomfärg hos 1000 slumpmässigt utvalda individer av kungsängslilja. Resultatet blev 900 purpurfärgade 95 vita och 5 i mellanformen. Om resultaten från den tidigare undersökningen är korrekta borde fördelningen av de olika färgerna i den nya undersökningen vara 800, 150 och 50. Vi kan direkt ana att de två undersökningarna inte givit samma resultat. För att vara på den säkra sidan utför vi ett chi-två-test för frekvenser.

Vi sammanställer resultaten i en tabell:

	Purpur	Vit	Mellanform
Observerad (O)	900	95	5
Förväntad (E)	800	150	50
$(O-E)^2/E$	12,5	20,2	40,5

Värdet på tesvariabeln blir 12,5 + 20,2 + 40,5 =73,2. Antalet frihetsgrader är i denna typ av test antalet kolumner -1. I detta exempel får vi df = 3-1 = 2. Från en tabell får vi att det kritiska värdet på χ^2 vid df = 2 och α= 0,01 är 9,21. Detta är lägre än det beräknade värdet på testvariabeln och vi kan precis som vi anade förkasta nollhypotesen att fördelningen skulle vara som i den tidigare undersökningen.

En frihetsgrad - Yates korrektion

Om man bara har två kategorier i ovanstående test kommer man att få 2-1 = 1 frihetsgrad. I detta fall kommer värdet på testvariabeln att bli för högt om man räknar på vanligt sätt. Detta innebär en förhöjd risk att man bedömer en fördelning signifikant skild från den teoretiska trots att så inte är fallet (ett typ I-fel).

För att undvika detta gör man en korrektion i beräkningen av testvariabeln. Istället för den allmänna formel som presenterades i början av detta avsnitt använder man följande variant:

$\chi^2 = \sum \frac{(|O-E|-0{,}5)^2}{E}$, där |O-E| är absolutbeloppet av skillnaden.

Test av associationer

En association är någon form av hopklumpning av data i ett insamlat material. Om data är insamlade på ett korrekt sätt kan man anta att samma hopklumpning gäller för hela den studerade populationen.

Det inledande exemplet visar hur man kan använda chi-två-tester för att hitta associationer. Vi ska här ge ett mer tydligt exempel, där det är lättare att förstå vad som menas med association i detta sammanhang.

Antag att vi vill testa om det finns någon association mellan torvmark och någon av revlummer eller lopplummer. Genom att utnyttja data från Ståndortskarteringen får vi antalet förekomster av de två arterna på torvmark respektive icke-torvmark (tabell 38). De förväntade frekvenserna beräknas enligt den allmänna formeln i inledningen av detta avsnitt.

Tabell 38. Förekomster av två lummerarter på olika substrat, samt förväntad förekomst och cell-chi-tvåt-värde.

Observerat Förväntat Cell-chi-två-värde	Lopplummer	Revlummer	Summa (observerat)
Icke-torvmark	172 219,12 9,919	3971 3923,9 0,5539	4143
Torvmark	88 40,88 53,172	685 732,12 2,989	773
Summa (observerat)	260	4656	4916

Eftersom antalet frihetsgrader är (r-1)×(k-1) = 1, är vi tvungna att tillämpa Yates korrektion när vi beräknar chi-två-värdet för respektive cell. Vi får då:

$$\chi^2 = \frac{(|172 - 219{,}12| - 0{,}5)^2}{219{,}12} + \frac{(|88 - 40{,}88| - 0{,}5)^2}{40{,}88} + \frac{(|3971 - 3923{,}9| - 0{,}5)^2}{3923{,}9} + \frac{(|685 - 732{,}12| - 0{,}5)^2}{732{,}12} =$$

$= 9{,}919 + 53{,}17 + 0{,}5534 + 2{,}969 = 66{,}61.$

Det kritiska värdet på χ^2 vid 1 frihetsgrad och $\alpha = 0{,}01$ är 6,63. Detta är långt lägre än värdet på vår framräknade testvariabel och vi kan förkasta nollhypotesen. Detta betyder att det finns en stark association mellan de variabler som ingick i testet. Genom att studera de förväntade frekvenserna och chi-två-värdena i varje cell (tabell 38) kan vi konstatera att lopplummer är associerad till torvmark.

G-tester

G-tester är ett alternativ till chi-två-tester och kan användas i alla situationer där ett chi-två-test kan användas. Vilken typ av test som används beror till stor del på tradition, även om statistiker menar att G-testerna har en starkare teoretisk grund.

För kontingenstabeller med en rad eller med 2 × 2 celler beräknas testvariabeln i ett G-test enligt:

$$G = \frac{2 \times \sum O \times \ln\left(\frac{O}{E}\right)}{\text{korrigeringsfaktor'}}$$

där korrigeringsfaktorn beräknas enligt:

$1 + (a^2 - 1)/6n\nu$, där

a är antalet kategorier, n är det totala antalet observationer och ν är antalet frihetsgrader vilket beräknas på samma sätt som för chi-två-tester.

För större kontingenstabeller beräknas G med samma formel men utan korrigeringsfaktorn.

Testvariabeln G är chi-två-fördelad och för att ta reda på signifikansnivån använder man en chi-två-tabell precis som i chi-två-tester.

Sokal & Rholf (1995) har en utförlig genomgång av G-tester och skillnaden mellan G- och chi-två-tester.

Multivariata metoder

I detta kompendium kommer vi inte att gå in på detaljer hos de multivariata analysteknikerna utan endast redogöra för huvuddragen hos några tekniker. Tanken med denna korta genomgång är att läsaren ska kunna förstå och tolka studier där man använt multivariata metoder, speciellt vad gäller s.k. ordinationstekniker. Klassifikation är något mer utförligt beskrivet.

Multivariata metoder är en grupp av statistiska analyser som skiljer sig från alla andra tekniker som tagits upp i detta kompendium. Den största skillnaden är att man oftast inte testar hypoteser, som i övriga tester. Multivariata metoder används istället för att hitta olika mönster och för att minska komplexiteten i stora dataset. Typexempel på dataset som lämpar sig för multivariat analys är vegetation i provytor eller kemidata i vattenprover (tabell 39). Multivariata metoder används bl.a. för att:

- Minska komplexitet i de data man har.
- Sortera prover efter likheter i t.ex. artsammansättning eller kemi.
- Hitta grupper eller klasser, som är så homogena som möjligt.
- Hitta (miljö-) gradienter som inte är uppenbara i stora dataset.

Tabell 39. Exempel på två dataset som lämpar sig för multivarait analys.

	Täckningsgrad (%)				
Provyta	Art 1	Art 2	Art 3	…	*Art j*
1	12	2	0	…	…
2	15	3	0	…	…
3	75	10	1	…	…
…	…	…	…	…	…
i	80	1	0	…	…

	Koncentration (olika enheter)				
Vattenprov	Tot-N	Alk.	Cl^-	…	*j*
a	720	,12	,066	…	…
b	352	,52	,050	…	…
c	509	,09	,024	…	…
…	…	…	…	…	…
nn	128	,13	,085	…	…

Det finns två huvudgrupper av multivariata analyser: klassifikation och ordination.

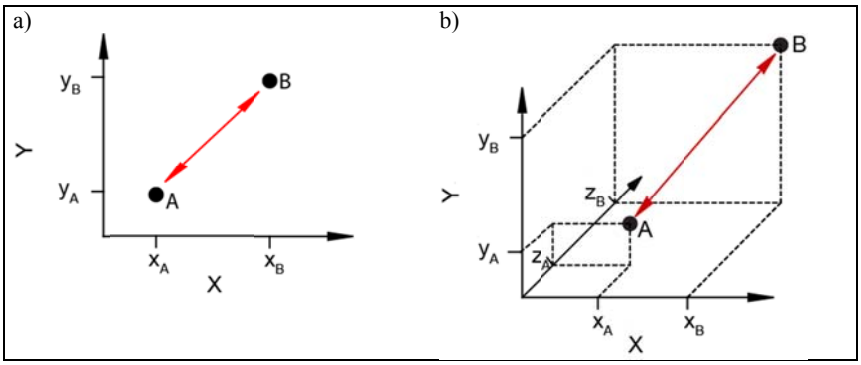
Klassifikation används för att sortera prover så att de som liknar varandra placeras i samma grupp, medan prover som är olika placeras i skilda grupper. Detta kan man göra för hand om man har en eller två variabler att ta hänsyn till för varje objekt, d.v.s. vattenprov etc. Har man däremot 30 variabler som beskriver varje objekt och kanske 300 objekt blir det en nära på omöjlig uppgift att sortera dessa för hand.

Ordination används för att minska komplexiteten i ett dataset och för att hitta mönster. Stora dataset är omöjliga att överblicka och det kan vara svårt att veta vilka variabler som är dominerande. Genom att först tillämpa en multivariat metod kan man reducera alla variabler som finns i hela datasetet till oftast två s.k. syntetiska variabler. T.ex. kan 30 kemivariabler från en stor mängd vattenprover sammanfattas i två eller tre syntetiska variabler, eller ordinationsaxlar som de också kallas.

Klassifikation

Klassifikation är som namnet indikerar ett sätt att dela in ett datamaterial i klasser där alla objekt liknar varandra. För att kunna dela in t.ex. vattenprover eller provytor i sådana klasser behöver man något mått på hur lika eller olika proverna är. Hur gör man då för att beräkna likheten mellan två objekt som är beskrivna av t.ex. 30 olika variabler? Svaret på denna fråga har många funderat över och det finns idag minst 50 olika sätt att beskriva likheter eller olikheter mellan objekt som är beskrivna med ett godtyckligt antal karaktärer.

Ett vanligt avståndsmått är Euklidiskt avstånd. Detta mått är relaterat till Pytagoras sats. Om två objekt (A och B) beskrivs av två karaktärer, vi kallar dessa X och Y, kan man illustrera dessa två objekt som i figur 31a.



Figur 31. Illustration av det Euklidiska avståndet mellan A och B, i två och tre dimensioner.

I figur 31a beräknas avståndet mellan A och B med hjälp av Pytagoras sats:

$$\sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}.$$

I figur 31b är A och B beskrivna med ytterligare en karaktär, Z, och avståndet därmed illustrerat i tre dimensioner. I detta fall beräknas avståndet som:

$$\sqrt{(x_B - x_A)^2 + (y_B - y_A)^2 + (z_B - z_A)^2}.$$

Som synes är den enda skillnaden mellan de båda formlerna antalet termer under rottecknet. I de fall man har ytterligare karaktärer lägger man till dem på motsvarande sätt.

I tabell 40 visas data på vattenkemi. Trots att varje objekt (= månad) är beskrivet av sex karaktärer kan man med lätthet beräkna det Euklidiska avståndet mellan objekten (tabell 41) med hjälp av formeln ovan. Vi kan konstatera att augusti och september är mest lika vad gäller vattenkemi och att augusti och mars skiljer mest. De framräknade siffrorna är avstånd och därför är två objekt mer lika varandra ju lägre värde på avståndet.

Tabell 40. Vattenkemidata från en provtagningsplats i en sjö, från olika månader.

Månad	pH	Syrgas mg/l	Ca mekv/l	Mg mekv/l	Na mekv/l	K mekv/l
Mars	6,66	11,18	1,117	0,253	0,263	0,045
Augusti	7,14	6,81	1,00	0,235	0,256	0,031
September	6,98	7,84	0,853	0,217	0,219	0,035
Oktober	7,36	9,44	0,985	0,218	0,242	0,033

Tabell 41. Euklidiskt avstånd mellan prover från de olika månaderna i tabell 40

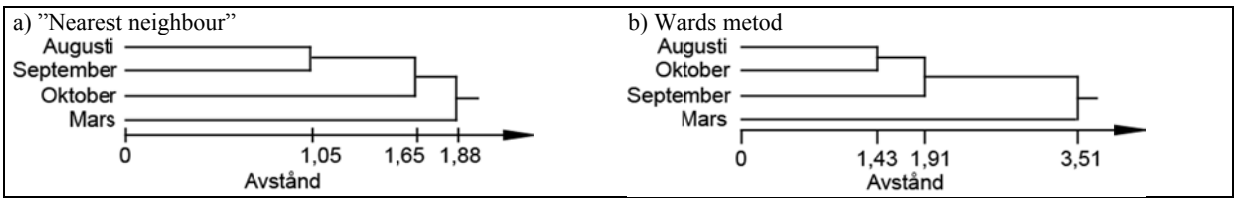
	Mars	Augusti	September	Oktober
Mars	0			
Augusti	4,40	0		
September	3,37	1,05	0	
Oktober	1,88	2,64	1,65	0

Efter att man sammanfattat avståndet mellan alla par av objekt till en enda siffra per par, är det dags att använda denna information för att gruppera objekten. De vanligaste metoderna är s.k. agglomerativa hierarkiska metoder. Dessa börjar med att se varje objekt som en enskild grupp. Därefter slår man samman grupper som liknar varandra. Detta upprepas med allt större grupper till det att alla objekt hamnar i samma grupp.

Det finns flera sätt att definiera vad som menas med ”lika” grupper. Ett vanligt sätt är det som kallas ”nearest neighbour” eller ”single linkage”. Med denna metod slås två grupper samman vid det kortaste

avståndet mellan ett objekt i den ena gruppen och ett objekt i den andra gruppen. Detta låter krångligt och därför får ett exempel belysa det hela.

För objekten tabell 41 kommer först alla objekt att vara egna grupper vid avståndet 0 (figur 32). Därefter kommer augusti och september att slås ihop vid avståndet 1,05 (markerat i tabell 41) eftersom de är de minst olika objekten. Efter sammanslagningen kommer vi vid detta avstånd få tre grupper: (augusti + september); mars; oktober. Det näst lägsta avståndet i tabell 41 finns mellan september och oktober. Därför slår man ihop gruppen (augusti + september) med oktober vid avståndet 1,65. Slutligen slår man ihop gruppen (augusti + september + oktober) med mars vid nivån 1,88 som är den tredje lägsta avståndet i tabell 41. Ett vanligt sätt att presentera denna typ av matematiska övningar är att illustrera sammanslagningarna i ett träddiagram eller dendrogram. I figur 32 illustreras resultaten av ovanstående klassifikation. Dessutom visas det dendrogram som erhålls då man istället använder Wards metod för att slå samman grupper.



Figur 32. Dendrogram som visar en klassificering av data i tabell 40 enligt avstånden i tabell 41, med två olika sätt att beräkna avstånd mellan grupper.

När man gör en klassificering finns det som antytts ovan flera olika sätt att beräkna avståndet mellan par av objekt, liksom det finns olika sätt att definiera vad som menas med att två objekt är lika. I exemplet ovan användes Euklidiskt avstånd och ”nearest neighbour”, fler sätt är beskrivna i tabellerna 42 och 44. Det går inte att rekommendera en viss kombination av avståndsmått och teknik för sammanslagning, utan detta måste avgöras från fall till fall. Som framfår av figur 32 kan resultatet bero på vilken metod man väljer.

De tekniker som anges i tabell 44 är alla olika varianter av s.k. hierarkisk sammanslagning av objekt till grupper. Om man har fler än ca 200 objekt förespråkar många statistiker att man istället använder en icke-hierarkisk metod för att skapa grupper. En sådan metod som förekommer i många statistikprogram är ”k-means”. Andra sätt att gruppera objekt går under samlingsnamnet ”Self-Organizing Feature Maps, SOM”, som bygger på neurala nätverk och artificiell intelligens. De senare är dock ovanliga än så länge.

För vegetation finns ytterligare en metod som heter TWINSPAN som är en förkortning av ”*Two Way INdicator SPecies ANalysis*”. Denna metod bygger på andra beräkningsmetoder än de som presenterats ovan och är ofta använd inom vegetationsekologin. Denna klassifikation gör man med hjälp av ett datorprogram med samma namn som metoden. OBS! Tidiga versioner av TWINSPAN har ett fel i algoritmerna. En korrekt version finns att ladda ner från Jari Oksanens hemsida (URL januari 2012: http://cc.oulu.fi/~jarioksa/).

Tabell 42. Några olika sätt att beräkna avstånd mellan två objekt.

Mått	**Formel**	**Kommentar**
Euklidiskt avstånd	$\sqrt{\sum (x_i - y_i)^2}$	Känsligt för olika skalor hos de karaktärer som beskriver objekten. Tyngdpunkt på dominanta karaktärer.
Manhattan (City block)	$\sum \lvert x_i - y_i \rvert$	Liknar Euklidiskt avstånd men ger mindre vikt åt avvikande värden.
Jaccards koefficient*	$S_j = \frac{a}{a+b+c}$	Vanligt för vegetation.
Sørensens index * (Community coefficient)	$S_j = \frac{2a}{2a+b+c}$	Besläktat med Jaccard, men ger större vikt åt arter som är närvarande i båda ytorna

*) förklaring av a, b, c och d ges i tabell 43

Tabell 43. Grunden för formlerna i tabell 42. Bokstäverna anger antalet karaktärer (t.ex. arter).

		Yta 1: Närvarande	Yta 1: Frånvarande	
Yta 2	Närvarande	a	c	a + c
	Frånvarande	b	d	b + d
		a + b	c + d	N

Tabell 44. Olika hierarkiska tekniker för sammanslagning av grupper

Metod	Beskrivning	Kommentar
Nearest neighbour (single linkage)	Avståndet mellan de närmaste objekten	Ger sällan distinkta grupper. Kan dock spåra diskontinuiteter i data.
Furthest neighbour (complete linkage)	Avståndet mellan de mest avlägsna objekten	Oftast bra metod om man har naturliga grupper, men kan ge starka falska grupper. Känsligt för avvikande värden.
Centroid	Avståndet mellan gruppernas viktade medelvärden	Relativt okänsligt för avvikande värden. Kan ge objekt i fel grupp.
Average	Medelvärdet av avstånden mellan alla par av karaktärer	Består av flera varianter. Vanligast är UPGMA (unweighted pair-group method using arithmetic averages). Tenderar att gruppera m.a.p. gruppers varians.
Ward	Använder ANOVA för att beräkna avstånd.	Elegant metod som tenderar att ge små grupper.

Ett träddiagram illustrerar hur olika objekt och grupper är relaterade till varandra. Ett problem som man inte får svar på i en klassifikation är var man ska avgränsa sammanslagningen. I exemplet ovan måste vi dra en gräns någonstans mellan 0 och 1,88 för att få ett antal relativt homogena grupper av de objekt vi hade från början. Om man sätter gränsen till 0 är varje objekt sin egen grupp och om man sätter gränsen över 1,88 hamnar alla objekt i samma grupp. Inget av dessa alternativ är önskvärda. En viss vägledning om var man ska dra gränsen kan man få om man kan se en praktisk eller naturlig gruppering. I exemplet kan det vara vettigt att dela så att mars bildar en ”vårgrupp” och de övriga månaderna bildar en ”höstgrupp”.

Ordination

Ordination är ett samlingsnamn för ett flertal besläktade statistiska tester och används bl.a. för att:
• minska komplexiteten i data,
• reducera antalet dimensioner,
• hitta och illustrera likheter och olikheter mellan objekt.

Med dimension menas i detta fall alla mätbara och omätbara faktorer som påverkar de objekt man mäter. Inom vegetationsekologin är det t.ex. alla biotiska och abiotiska faktorer som påverkar vilka arter som växer på en viss plats. Många av dessa är omöjliga att mäta men avspeglar sig indirekt i sammansättningen av växtarter. I en ordination reducerar man alla dessa dimensioner till oftast två gradienter som visar den största variationen i det dataset man analyserar. Dessa två gradienter, eller axlar, kan man plotta mot varandra i ett diagram. I detta s.k. ordinationsdiagram kommer objekt (provytor) som liknar varandra i artsammansättning att hamna nära varandra och objekt som är olika varandra att hamna långt ifrån varandra. Axel 1 visar alltid den störta variationen hos de data man analyserar och i ett ordinationsdiagram låter man alltid axel 1 vara den horisontella axeln.

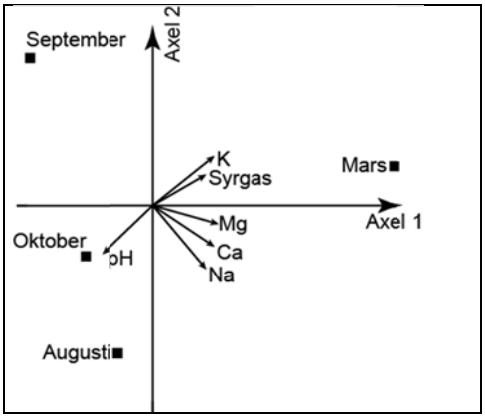
Ordinationsaxlar har inga enheter i traditionell mening. Man kan dock i efterhand kontrollera om objektens koordinater längs en axel korrelerar med någon variabel som inte ingått i själva beräkningarna. Det är t.ex. vanligt att man gör en ordination med artsammansättning som ingående data och att man efter det att provytorna fördelats längs ordinationsaxlar tolkar fördelningen m.h.a.

olika miljövariabler. Om man hittar en god korrelation mellan provytors position längs en axel och någon testad miljövariabel kan man anta att variationen längs den axeln till stor del beror av det testade variabeln.

En ordination (figur 33) av datasetet i tabell 40 visar att axel 1 skiljer mars månad från de tre höstmånaderna, precis som klassifikationen visade. De tre höstmånaderna är dock dåligt separerade längs den första ordinationsaxeln. Däremot får man en god separation av dessa tre månader längs axel 2. Ordinationen ger en liknande bild av likheter och olikheter i vattenkemi i de tre proverna som klassifikationen med Wards metod, men skiljer sig från ”nearest neighbour”-klassifikationen (jfr. figur 31).

En ordination ger även svar på frågan om vilka av de uppmätta variablerna som orsakar det mönster av objekt man erhållit. De olika variablernas inverkan framgår av deras placering i ordinationsdiagrammet (figur 33). Kontinuerliga variabler markeras ofta med en pil från origo för att indikera att variabeln ökar i den riktning som piken pekar. Ju längre från origo desto större inverkan, och ju närmare en axel desto mer relaterat till den axeln.

I figur 33 framgår att pH kontrasterar mot övriga variabler. Eftersom pH ligger till vänster om origo är det negativt korrelerat den första ordinationsaxeln, medan de övriga variablerna är positivt korrelerade. Mars-provet har sålunda höga värden hos alla variabler utom pH, medan de tre höstproverna har motsatta förhållanden hos de kemivariabler som använts i analysen.



Figur 33. Ordinationsaxel 1 och 2 från en principalkomponentanalys (PCA) av data i tabell 40.

En varning

Det finns två olika familjer av ordinationstekniker och beroende på vilka data man har väljer man den ena eller andra. Om man väljer fel typ av ordinationsteknik **blir resultatet felaktigt**! Principen för att välja rätt teknik är dock enkel, även om det kan förekomma svåra gränsfall. Det hela baseras på vilken typ av data man har. Kemidata och likande har s.k. en linjär respons. Data på organismers närvaro i provytor har däremot en Gaussformad fördelning. Några ordinationstekniker förutsätter att data har en linjär respons, medan andra bygger på en unimodal (Gaussformad) respons (tabell 45).

Tabell 45. Sammanställning av olika ordinationstekniker, uppdelade efter vilken typ av data som krävs för att tekniken ska ge korrekt resultat.

Linjär respons	Unimodal respons
PCA Principal Component Analysis	GO Gaussian ordination
RDA Redundancy analysis	CA Correspondence Analysis
PCoO Principal Co-ordinate Ordination	DCA Detrended Correspondence Analysis
PLS Partial Least Square Regression	CCA* Canonical Correspondence Analysis
	DCCA Detrended CCA
MDS Multi Dimensional Scaling	MDS Multi Dimensional Scaling

*Akronymen CCA förekommer även för ”canonical correlation analysis” som bygger på linjär respons.

Många allmänna statistikprogram har moduler för ordination, men dessa är nästan uteslutande metoder som förutsätter linjär respons. Om man har andra typer av data finns många specialprogram för de metoder som krävs i dessa fall. Inom växtekologin används ofta programmen CANOCO™ eller PC-ORD, men det finns många andra program att tillgå t.ex. gratisprogrammet PAST eller modulen ”vegan” för R.

Litteratur

För den som vill läsa vidare följer här en lista på dels citerade källor och dels på några böcker som rekommenderas för vidare studier i tillämpad statistik.

Conover, W. J. 1999: *Practical Nonparametric Statistics*. John Wiley & Sons.

Fowler, J., Cohen, L. & Jarvis, P. 1998: *Practical statistics for field biology*. 2 uppl. John Wiley & Sons Ltd.

Hammer, Ø., Harper, D.A.T. och Ryan, P.D. 2001. PAST: Paleontological Statistics software package for education and data analysis. *Palaeontologica Electronica* 4(1):9 pp. Finns att ladda ner från http://folk.uio.no/ohammer/past/index.html (januari 2012).

Helsel, D. R. & Hirsch, R. M. 1992: *Statistical Methods in Water Resources*. Elsevier.

Jongman, R. H., ter Braak, C. J. F. & Tongeren, O. F. R. 1995: *Data analysis in community and landscape ecology*. 2 uppl. Cambridge University Press.

Legendre, P. & Legendre, L. 1998: *Numerical ecology*. 2 uppl. Elsevier Scientific Publishing Company.

Motulsky, H. J. 1999: *Analyzing Data with GraphPad Prism*. GraphPad Software Inc. Finns att ladda ner gratis från http://www.graphpad.com/articles/analyzingdata.pdf (januari 2012).

Oksanen, J. 2006. *Vegan: R functions for vegetation ecologists*. Finns att ladda ner från: http://cc.oulu.fi/~jarioksa/softhelp/vegan.html (januari 2012).

Rutherford, A. 2001. *Introducing ANOVA and Ancova: A Glm Approach.* Sage Publications.

Sokal, R. R. & Rohlf, F. J. 1995: *Biometry: The Principles and Practice of Statistics in Biological Research*. 3 uppl. W. H. Freeman and Co.

Zar, J. H. 1999: *Biostatistical analysis*. 4 uppl. Prentice-Hall.