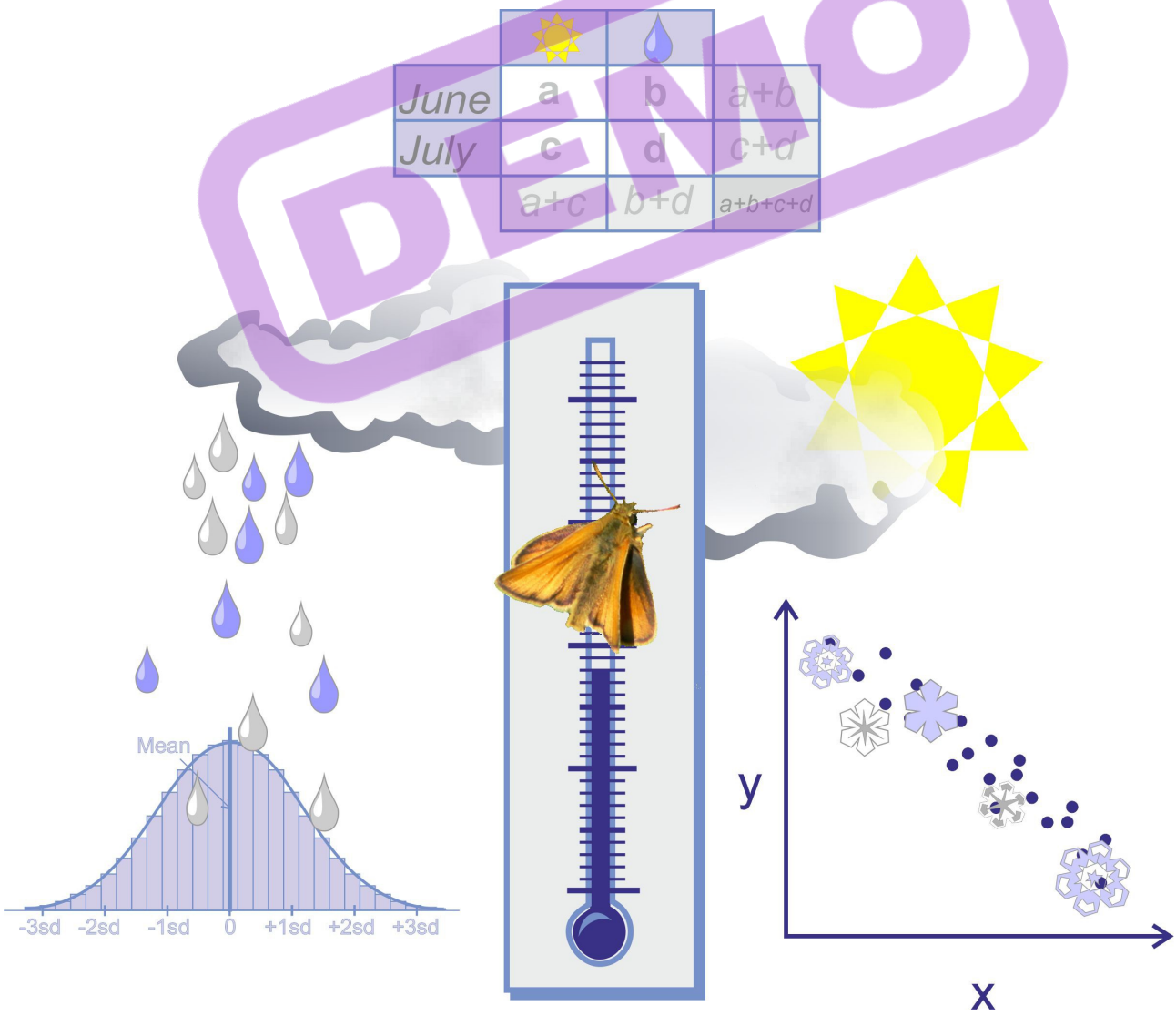
Utdrag ur utbildningsmaterial i grundläggande statistik, huvudsakligen baserat på väderobservationer från SMHI. Materialet finns även i engelsk översättning.

Grundläggande statistik

En kort introduktion



Carin Kullberg

2. Hypotestestning och Parametriska test

Hypotestestning

Man gör ett stickprov för att dra slutsatser om hela det studerade materialet/fenomenet/populationen. Stickprovet, urvalet, måste alltså vara *representativt* för hela studieobjektet. Tänk noga igenom inklusions- och exklusions-kriterier för att undvika urvalsbias, och val av mätmetod för att försäkra dig om att det du mäter verkligen speglar det du vill veta.

Man kan aldrig *bevisa* något om en studiepopulation med statistik, men man kan avfärda något som *mindre sannolikt.*

Man måste definiera exakt vad ett statistiskt test ska undersöka. Därför formuleras två motstående hypoteser.

Nollhypotes, H_0, H(0): Nollhypotesen är alltid utgångspunkten för alla statistiska test. Den säger vanligen att försökets resultat enbart beror på slumpvisa variationer. Det innebär så gott som alltid att den oberoende (=förklarande) variabeln inte har någon effekt på den beroende (=förklarade) variabeln, och att eventuella skillnader eller samband mellan grupper eller variabler beror på slumpen eller faktorer som inte testats.

Alternativhypotes: **H_1,** eller **H(a):** Alternativhypotesen är den idé som försöket utformades för att testa. Oftast innebär det att den oberoende variabeln påverkar värdet av den beroende variabeln, och att de uppkomna skillnaderna inte (enbart) på slumpen. Ibland ser man att en studie använt sig av flera alternativhypoteser (t ex H_1, H_2 och H_3) Man brukar ändå testa dem en i taget gentemot nollhypotesen.

		Verkligheten (alias Sanningen!)	
		H_0 is Sann	**H_0 is Falsk**
Statistiskt test	**Förkasta H_0**	**Typ I-fel** (Falskt positiv)	**Korrekt** (Sant positiv)
	Acceptera H_0	**Korrekt** (Sant negativ)	**Typ II-fel** (Falskt negativ)

Hypotesprövning innebär att med statistiska metoder undersöka om nollhypotesen kan förkastas, dvs. om sannolikheten att få samma resultat som i försöket av en slump är (tillräckligt) låg. Denna sannolikhet kan uttryckas på flera sätt, t ex som ett p-värde eller ett konfidensintervall. Konfidensintervall och p-värden kan beräknas även om data inte är normalfördelade, men beräkningsmetoderna varierar mellan olika testmetoder. Låt statistikprogrammet sköta beräkningarna, men kolla noga vilken metod som bör användas för varje analys.

Hypotestestning kan göras ensidigt eller tvåsidigt. **Tvåsidiga test** är absolut vanligast. De undersöker om t ex ett medelvärde skiljer sig signifikant från ett givet värde, vare sig det är större eller mindre än jämförelsevärdet. Referensvärdet kan vara ett tidigare känt värde, t ex vikten i en befolkning enligt offentliga databaser eller publicerade studier, eller medelvärdet i en annan grupp i samma studie, t.ex. jämföra mäns och kvinnors vikt.

Ensidiga test används inte lika ofta. Då undersöks bara om försökets medelvärde (eller vad som nu testas) skiljer sig signifikant från jämförelsevärdet *i en specifik riktning*. Kanske skulle man använda ett ensidigt test om ett nytt, dyrare läkemedel lanseras. Det skulle bara införlivas med standardsortimentet om det har tydligt bättre effekt än befintliga läkemedel, och alltså bara vara intressant om det är signifikant bättre än alternativen. Men å andra sidan vore det väl av intresse att vet även om det vore signifikant sämre, för då kan det vara viktigt att avråda patienter att använda det, trots påkostade reklamkampanjer, så varför inte göra ett två-sidigt test?

Sannolikhetsmått

p-värde: p-värdet visar sannolikheten att en skillnad som är *minst lika extrem* som den uppmätta uppkommit av en slump, trots att nollhypotesen är sann. Eller för att göra det enkelt, hur troligt det är att resultatet kan uppkomma av en slump. Ett lågt p-värde innebär låg sannolikhet att skillnaden uppstått av en slump, och då är troligen nollhypotesen felaktig.

Signifikansnivån är den accepterade risken att förkasta nollhypotesen trots att den är sann, t.ex. risken att felaktigt acceptera en slumpmässig variation som en verklig skillnad mellan två grupper. Signifikansnivån 5 % innebär 5 % risk att en uppmätt skillnad beror på slumpen. Om p-värdet är *högre* än signifikansnivån (t.ex. p=0,16 för en 5% signifikansnivå) innebär det att nollhypotesen *inte förkastas*. Vanligast är 5% signifikansnivå, men 1%, eller ännu lägre, används ganska ofta om man studerar många variabler samtidigt, eller farliga följdverkningar, t.ex. risken att en behandling av en lindrig åkomma kan orsaka svår skada eller död.

MEN: signifikansnivåer är inte naturlagar. Ett p-värde på 0,049 innebär inte att något är mer *sant* än om p-värdet är 0,051, bara lite, lite, lite mindre sannolikt att det beror på slumpen. Med få observationer i urvalet, eller få

mätvärden i en mätserie, kan det vara svårt att uppnå signifikans. För små grupper eller ett ovanliga fenomen kan det vara näst intill omöjligt att nå statistisk signifikans. Att jämföra längd hos män och kvinnor i en slumpmässigt vald grupp om 100 personer, men vill man *i samma urval* kontrollera om 4-barnsmammor från landsbygd är kortare än 4-barnsmammor från storstad kan det bli problem.

Å andra sidan innebär en 5% signifikansnivå att var 20:de test kommer att falla ut som signifikant trots att effekten bara beror på slumpvariation. Ju fler signifikanstest man gör desto större är risken att ett eller flera test visar på en effekt som inte finns i verkligheten. Detta problem brukar kallas **mass-signifikans**.

Det finns flera sätt att hantera detta. Det viktigaste, men allt för ofta ”bortglömt”, är att fundera på orsak och verkan. Som ett exempel: om man i WEATHER_1996_2016 MONTH tittar på hur olika variabler korrelerar (se del 3 om korrelationer) finns det en samvariation mellan medeltemperatur och antal observationer; ju fler observationer under månaden, desto högre medeltemperatur (Spearman correlation, p=0,00014). Är detta ”sant”? Ja, det finns en samvariation, men frågan är väl om det finns ett *kausalt samband* i meningen att temperaturen skulle påverka antal mätvärden eller, ännu mer befängt, att antal gånger under månaden man mäter påverkar temperaturen. Hur skulle den verkningsmekanismen se ut? Eller finns det ett systematiskt ”fel”, där mätapparaturen eller observatören tappar batterikraft respektive inte vill läsa av så ofta när det är kallt? Eller kan det hänga samman med att februari bara har 28 eller 29 dagar=mätvärden medan årets varmaste månader, juli och augusti har 31 dagar? **Ifrågasätt alltid dina resultat!**

Ett sätt att hantera problemet med massignifikans är att *begränsa antalet tester* man gör. Istället för att mäta allt koncentrerar man sig på ett fåtal variabler som är särskilt intressanta. Problemet är att det ofta blir ett laboratorieexperiment snarare än en bild av verkligheten. Bara ett fåtal fenomen påverkas av bara en eller ett fåtal faktorer. Genom att bara titta på enstaka förklarande variabler riskerar man att missa andra viktiga faktorer, som kanske förklarar fenomenet bättre, eller samverkan av flera faktorer.

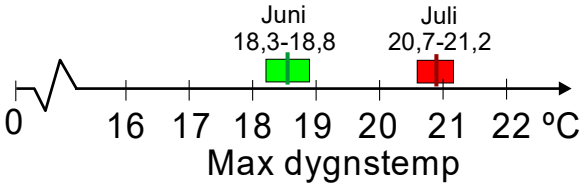
Ett mer ”mekaniskt” sätt att hantera problemet med massignifikans är att på något sätt korrigera statistik-testet för antalet test man gör. Det enklaste är **Bonferroni-korrektion**. Det innebär helt enkelt att man dividerar önskad signifikansnivå med antal test. Om man vill uppnå 5%-nivån för *ett* test ska man alltså ha ett $p \leq 0,05$. Kör man t.ex. t-test för 10 variabler kräver man ett p-värde som är 0,05/10 = 0,005 eller lägre, och kör man 100 t-test krävs $p \leq 0,0005$ för statistisk signifikans. Det finns olika ”skolor” om det är meningsfullt, eller ens korrekt, att göra den här sortens korrektioner. Många menar att det är bättre att överlåta värderingen av signifikansnivån till läsaren.

Hypotesprövning kan även göras med hjälp av **konfidensintervall**: Ett 95% konfidensintervall täcker med 95% sannolikhet det sanna värdet, t ex årsmedeltemperaturen, och motsvarar alltså en signifikansnivå på 5%.

Man kan dels använda konfidensintervall för att testa om ett uppmätt medelvärde skiljer sig signifikant från ett bestämt värde, dels för att kontrollera om två eller flera medelvärden från olika grupper eller mättillfällen skiljer sig från varandra.

Om konfidensintervallen för två grupper överlappar varandra är de inte signifikant skilda, åtminstone inte på den angivna nivån. Om konfidensintervallen *inte* överlappar är medelvärdena, eller vad man nu testar, signifikant åtskilda.

Exempel 2.1: Här visas 95% CI för dagens maxtemperatur under juni (grönt) och juli (rött). Intervallen överlappar inte varandra, alltså är de signifikant skilda med $p < 0,05$
(Data för Kilsbergen, juni och juli 1996-2016)



Förutsättningar för parametriska test

Uttrycket **parametriska test** syftar på att dessa test är beroende av *normalfördelningens parametrar*, dvs. medelvärde och varians (oftast uttryckt som standardavvikelse, roten ur variansen). Parametriska test bygger på visa förutsättningar, i princip samma grunder som för om man kan använda medelvärde och varians.

- variablerna är mätta på en **intervallskala**
- variablerna är **normalfördelade** och/eller mätfel är slumpmässigt fördelade

Dessutom behöver man förstås tillräckligt stort antal mätvärden. Har man ett stort material funkar faktiskt parametriska test även om variablerna inte är normalfördelade

Parametriska test

Alla slags t-test räknar ut ett p-värde för sannolikheten att skillnaden mellan två medelvärden, eller mellan ett medelvärde och ett annat givet värde uppkommit av en slump.

One-sample t-test

Den enklaste formen av t-test. Det jämför medelvärdet för *en* variabel mot ett bestämt värde. Det kan t.ex. vara ett referensvärde från offentlig statistik eller resultat i en publicerad artikel som rör samma slags data.

Exempel 2.2: Medeltemperaturen för juli månad 2006–2015 i Örebro enligt SMHI:s månadsstatistik visas i tabellen. Skiljer sig medeltemperaturen för denna period från referensperioden 1961–1990, då medeltemperaturen var 16,0ºC?

År	Medeltemp, °C
2006	19,7
2007	15,9
2008	17,5
2009	16,5
2010	19,1
2011	18,3
2012	16,8
2013	17,9
2014	20,0
2015	16,6

Ett one-sample t-test för dessa 10 månadsmedelvärden visas nedan. Medelvärdet för perioden är 17,8 ºC, SD =1,42 och 95% CI 16,9–18,8, dvs. konfidensintervallet omfattar inte jämförelsevärdet 16,0. Därmed var julitemperaturerna 2006–2015 signifikant högre än under referensperioden. P-värdet för differensen var 0,00281.

H0: MEAN = 16.00 VS. H1: MEAN <> 16.00

Variable	N	Mean	Standard Deviation	95% CI		t	df	p-Value
				Lower Limit	Upper Limit			
Meantemp Juli	10	17,8300	1,42287	16,8121	18,8479	4,0671	9,0000	0,00281

Oparat t-test

(eng: *two independent sample t-test*) är nog den vanligaste formen av t-test. Det används för att testa skillnaden i medelvärde mellan två oberoende variabler eller grupper är 0, dvs. *Nollhypotes*: skillnaden i medelvärde mellan grupperna = 0.

Oparat t-test förutsätter att värdena är normalfördelade, och att variansen inom grupperna är likartad, vilket testas med Levene's test. Är variationerna olika kan ett korrigerat t-test användas (betecknas ofta "separate variance t-test", eller "equal variances not assumed"). De flesta statistikprogram spottar ut resultat både för "pooled variance" och "separate variance" tillsammans med ett Levene´s test, och användaren kan välja vilket resultat man vill använda.

Levene's test

Det här testet används t.ex. med t-test eller ANOVA för att testa om två eller flera prover har samma varians. Nollhypotesen är att varianserna är lika stora (homogena). Om testets p-värde är signifikant (vanligen p<0,05) anses det osannolikt att varianserna är lika, och då brukar man använda modifierade testmetoder för att jämföra medelvärdena, t.ex. t-test för prov med olika varians.

Exempel 2.3: I exempel 2.1 om konfidensintervall konstaterade vi att CI95% för dygnsmedeltemperatur under juni respektive juli inte överlappade varandra. Medelvärdena är därmed signifikant skilda med 95% säkerhet. Om vi använder samma data till ett oparat t-test kan vi också få ett p-värde för skillnaden:

Variable	Month	N	Mean	Std Dev
Max temp, C	6	630	18,54540	3,54561
	7	651	20,92734	3,52845

Levene's Test

Variable		F-Ratio	df	p-Value
Max temp, C	Based on Mean	0,06863	1, 1279	0,79339
	Based on Median	0,10365	1, 1279	0,74755

Levene's test visar att varianserna är lika (p>0,05). Därför kan ett t-test med "pooled variance" (eller "equal variance assumed" i SPSS) användas.

Separate Variance

Variable	Month	Mean Difference	95% CI		t	df	p-Value
			Lower Limit	Upper Limit			
Max temp, C	6 7	-2,38195	-2,76977	-1,99412	-12,0493	1277,1879	0,00000

Pooled Variance

Variable	Month	Mean Difference	95% CI		t	df	p-Value
			Lower Limit	Upper Limit			
Max temp, C	6 7	-2,38195	-2,76974	-1,99416	-12,0502	1279,0000	0,00000

I det här fallet blir p-värdet detsamma för båda t-test-varianterna, p<0,00001, men konfidensintervallens gränser skiljer sig lite. Programmet Systat beräknar och visar skillnaden mellan grupperna, istället för medel och CI för varje grupp, men det statistiska testet och resultaten är desamma. Om medelvärdena är lika blir ju skillnaden = 0, och CI för skillnaden innefattar värdet 0, dvs. är inte signifikant.

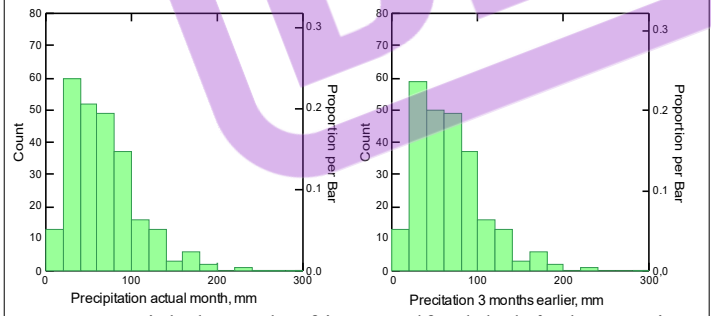
Parat t-test

Parat t-test kallas även matchat t-test eller beroende t-test. Det används vanligen när man mäter samma variabel på samma urval två gånger, t.ex. vid start och efter en viss händelse. Sådana matchade provtagningar minskar påverkan av variationen inom respektive grupp.

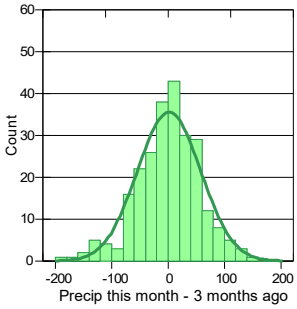
”Före-och-efter”-undersökningar är en vanlig studiedesign, som kontrollerar t.ex. vikt på samma individer före och efter en intervention, vilket ger en perfekt matchning. Det kan också användas för att jämföra tå mätserier av samma fenomen, t.ex. privata temperaturmätningar mot SMHI:s närmaste väderstation under samma tidsperiod. Eftersom man mäter samma fenomen under samma tidsperiod blir värdena matchade.

Även om de båda variablerna var för sig inte är normalfördelade kan man ganska ofta använda parat t-test. Det kräver att man har relativt många mätvärden (>50 åtminstone). Ett parat t-test kontrollerar egentligen om *differenserna* i medeltal avviker från 0, och är därför inte så beroende av de enskilda variablernas fördelning. Skillnaden mellan mättillfälle 1 och 2, eller vilka värden man nu matchar, kan mycket väl vara normalfördelade, även om värdena från respektive tillfälle inte är det. Gör ett histogram för differenserna för att kontrollera fördelningarna.

Exempel 2.4: För att demonstrera hur två skeva fördelningar ändå ger en hyfsat normalfördelad differens används data för nederbörd per månad, jämfört med nederbörden 3 månader tidigare. N=249. (För den misstänksamme kan meddelas att det inte fanns någon samvariation mellan de båda nederbördsvariablerna; regressionskoefficient -0,04, och p=0,516)



PRECIP-variabeln är kraftigt snedfördelad åt höger, inte normalfördelad, liksom nederbörden 3 månader tidigare (samma data, men förskjutet 3 mån i tid). Medelvärde 66 mm, SE±2,4, men kraftigt positivt skeva fördelningar



Differenserna mellan nederbörden en viss månad och 3 månader tidigare är nästan normalfördelade. Lite skeva åt vänster och lite toppiga, men båda p-värdena är nära 0,05; Medel -0,2 SE±3,5.

Innehåll

Absolut risk....41
Adjusted R2....26
Alternativhypotes....14
Andel....34
Anova....17
Antal inkluderade variabler....28f.
Attribuerbar risk....41
Beroende variabel....14, 23
Bias....13
Binomialfördelning....6
Bonferroni-korrektion....15
Centralvärde....4
Chi2 för linjär trend....38
Chi2-fördelning....7
Chi2-test....35
CI, konfidensintervall....9
Coefficient of Variation....11
Confounding....22
CV, Variationscoefficient....11
Deskriptiv statistik....3
Detektionsgräns....12
Diagnostiska test....41
Dikotom variabel....6
Dummy-variabel....27
Ensidiga test....14
Epidemiologi....35
Extrapolering....24
Faktor....17
Falsk negativ....12
Falskt positiv....12
Fisher's exact test....36
Frekvenstabell....4
Fördelningar....5
Förklarande variabel....23
Förväntat värde....35
Gauss-fördelning....5
Gruppstorlek....18
Histogram....4
Hypergeometrisk fördelning....7
Hypotestestning....14
Icke-parametriska test....31
Ickelinjärt samband....21
Incidens....40
Incidensrisk....41
Inter Quartile Range....8, 31
Interaktion....27
Intercept....23
Intervallskala....3
IQR, Inter Quartile Range....8
Kategori....3
Kategorisk variabel....3
Konfidensintervall....9, 24
Konstant....23
Kontinuerlig variabel....3
Korrelation....20
Korrelationskoefficient....20
Korstabell....38
Korstabeller....34
Kruskal-Wallis-test....32
Kumulativ frekvens....4
Kumulativ incidens....41
Kurtosis....10
Kvartiler....8, 31
Kvotskala....3
Levene's test....16
Likformig fördelning....6
Linjens ekvation....23
Linjär regression....23
Linjärt samband....21
Logaritmerad oddskvot....28
Logistisk regression....28
Lägesmått....4, 7
Mann-Whitney U-test....32
Mass-signifikans....15
Matchat t-test....17
Medelvärde....7
Median....7, 31
Modalvärde....7
Mode....31
Multinominal regression....30
Multipel linjär regression....26
Multipel regressionskoefficient....26
Negativt samband....21
Noggrannhet....12
Nollhypotes....14
Nominal skala....3
Normalfördelning....5
Oberoende variabel....14, 23
Odds....37
Odds ratio....37
Oddskvot....37, 41
Omfång....8
One-sample t-test....16
One-way anova....17
Oparat t-test....16
OR, Oddskvot....37
Ordinal skala....3
p-värde....4, 14
Parameter....4
Parametriska test....15
Parat t-test....17
Pearsons χ2....36
Percentiler....8, 31
Poissonfördelning....6
Positivt samband....21
Precision....12
Prediktionsintervall....24
Prediktivt värde....42
Prevalens....40
Proportion....34
Proportion av en variabel....34
Proportioner i två variabler....34
Pseudo R-square....29
Punktprevalens....40
r, Korrelationskoefficient....20
Range....8
Rankning....31
Ratioskala....3
Referensintervall....9
Regressionskoefficient....23
Rektangulär fördelning....6
Relativ frekvens....4
Relativ kumulativ frekvens....4
Relativ risk....36, 41
Reliabilitet....13
Residual....24
Riktighet....12
Risk....36
Risk ratio....36
Risk-differens....41
Riskkvot....36
Robust analys....24
ROC-kurva....29, 43
RR, Relativ risk....36
Rule of Ten....29
Räta linjens ekvation....23
Sannolikhetsmått....14
SD, Standardavvikelse....8
SE, Standardfel....9
Sensitivitet....42
Signifikansnivå....14
Simpson-paradoxen....18
Skalor....3
Skevhet....10
Skewness....10
Spearman-korrelation....33
Spearmans rho....33
Specificitet....42
Spridning....4
Spridningsmått....8
Standard error....9
Standardavvikelse....8
Standardfel....9
Stepwise regression....26
Stratifiering....39
t-fördelning....5
Toppighet....10
Tvåsidiga test....14
Two-way anova....18
Typ I-fel....12, 28
Typ II-fel....12
Typvärde....7
Upplösning....12
Validitet....13
Variabel....4
Varians....8
Variationskoefficient....11
Wilcoxon rank-sum test....32
Wilcoxon sign-ranked test....31
Wilcoxon two-sample test....32
Yates's korrigerat chi2....36
z-fördelning....5
Överanpassning....28
α-fel....12
β-fel....12
χ2-fördelning....7