Statistik: Historier og eksempler

Helle Hvitved

Dette er et forsøg på at give en overordnet beskrivelse af statistik og statistiske begreber uden at gå for meget i matematiske detaljer. Derved vil der selvfølgelig mangle nogle ting, og det vil ikke blive helt præcist på alle punkter, men det er heller ikke meningen, at man skal blive ekspert i statistiske tests ud fra dette dokument. Dokumentet er ment som et supplement for lærere, der gerne vil have nogle historier at fortælle eleverne udover det, der står i lærebogen.

Introduktion til statistik generelt.

Vi mennesker har en indbygget trang til at kategorisere ting og udtale sig herom, f.eks. om økologisk mad er sundere end uøkologisk, om det var koldere i gamle dage, om kvinder er bedre til at multitaske end mænd, om man får større muskler, hvis man spiser et bestemt kosttilskud osv. I andre tilfælde kan det være vigtigt at finde f.eks. en type stof, der optræder i forskellig mængde hos syge og raske eller hos dopede og ikke-dopede mennesker eller at finde forskelle mellem folk med forskellig social/kulturel baggrund.

I dagligdagen er det de færreste ting, man er helt sikker på. *Deterministisk* betyder, at man er 100% sikker på, at noget sker. Det er karakteristisk for fysikkens love. F.eks. er vi 100% sikre på, at æblet falder ned fra træet og ikke op. Indtil videre er vi også 100% sikre på, at et fødende menneske er en kvinde, og vi er 100% sikre på, at hun ikke føder en elefant, men vi er ikke 100% sikre på, at babyen er et velskabt barn.

Når der nu er så mange ting i verden, som man ikke er 100% sikker på, så må vi jo nøjes med lidt mindre, og så sætte nogle retningslinjer for, hvornår vi bærer os ad med at vælge f.eks. om noget er bedre/anderledes end andet. Statistik er netop det redskab, der skal hjælpe os til at beslutte, om der er hold i et udsagn af ovennævnte typer eller hjælpe os til at afgøre, om man f.eks. kan adskille syge/raske eller dopede/ikke-dopede på baggrund af et bestemt stof.

Det første problem er at finde ud af, hvordan man måler det, man vil udtale sig om. Det vil vi ikke beskæftige os med lige her. Næste problem er at finde ud af, hvem eller hvad vi snakker om. Det kaldes *populationen*. Vi kan vælge økologiske gulerødder, temperaturer de sidste 200 år, kvinder og mænd mellem 20 og 60 år, 1. og 2. generations indvandrere fra bestemte lande, osv. Det, vi vælger at måle på kaldes en *variabel*. Vi kan se på køn (mand-kvinde), dyrkningsmetode (økologisk-traditionel), religion (protestant, muslim, andet), vælgere (blå-rød blok), stofmængde af et bestemt kemisk stof (målt i mg) eller højde (målt i meter), indkomst (kr. eller dollar). De første fire variable kaldes kategoriske og de tre sidste kaldes numeriske.

Indenfor denne population har vi så vores forestilling om en sammenhæng, og vi skal have denne sammenhæng formuleret som en hypotese eller rettere 2 hypoteser, idet vi skal bruge en alternativ hypotese, hvis vi ender med ikke at tro på vores udgangshypotese. Nu er det sådan, at det er lettere at regne med ligheder end med uligheder, for hvis to ting er ens, så er forskellen 0, men hvis de ikke er ens, hvad er forskellen så? 1? 1,5? 7? 133? Med kategoriske variable taler vi om uafhængighed i stedet for lighed, så her har vi med bestemte forholdstal, som vi så kan regne på. Statistik er sådan bygget op, at vores hovedhypotese, der kaldes nul-hypotesen, er den, hvor der er lighedstegn eller uafhængighed, og den alternative hypotese er den, hvor der er forskel eller afhængighed– og så er det i første omgang ligegyldigt, hvor stor den forskel er. Det er tit sådan, at man faktisk ønsker, at forkaste hypotesen om lighed og godtage den alternative hypotese om ulighed. Medicinalfirma A vil frygtelig gerne vise, at deres medicin er bedre en firma B's. Kvinderne vil gerne vise, at de er bedre til at multitaske en mænd, man vil gerne vise at arbejdsløsheden falder/stiger osv. Nul-hypotesen siger, at hvis man ser en forskel, så skyldes det rene tilfældigheder, mens den alternative hypotese påstår, at forskellen skyldes en underliggende årsag, f.eks. køn eller økologisk dyrkning eller den siddende regerings politik.

Nu skal vi så i gang med at måle på vores variable og lave nogle statistiske beregninger. Vi kan ikke måle alt. Hvis vi f.eks. måler vitaminindhold i alle verdens gulerødder, så er der jo ikke nogen tilbage at spise, og det vil tage for lang tid og være for dyrt. Derfor er vi nødt til at tage en *stikprøve*, som er et mindre og overkommeligt antal objekter. Hvordan disse udvælges, og hvor mange, der udvælges skal ikke behandles her.

Forskellige statistiske tests

For at få en overskuelig fornemmelse af, hvad der findes af statistiske tests, der er relevante for gymnasiet, laves følgende oversigt:

<table>
<tr><td colspan="3">Enkelt variabel</td></tr>
<tr><td>Kategorisk data</td><td colspan="2">Kategorisk data stammer fra undersøgelser, hvor man grupperer sine objekter i forskellige kategorier og tæller antallet i hver gruppe. Det kan f.eks. være en spørgeskemaundersøgelser med lukkede spørgsmål, hvor man kun kan sætte flueben, eller forsøg hvor man tæller objekter: 卌 |||. Hver variabel inddeles i et antal kategorier (svarmuligheder), og hvert individ eller objekt tæller 1 i hver kategori, f.eks. 1 kvinde, 1 rødhåret, 1 elev på matA, 1 ryger, 1 arbejdsløs, 1 grøn og rund ært. Når der kun er 1 variabel, kan man sammenligne antallet i kategorierne med en teoretisk eller tidligere funden fordeling. Det kan dreje sig om en bestemt genetisk fordeling, eller det kan være en opinionsundersøgelse, der sammenlignes med sidste valg.
Testen kaldes Goodness of fit (GOF) og er en variant af X^2-testen.</td></tr>
<tr><td>Numerisk data</td><td colspan="2">Numerisk data er f.eks. højde, vægt, indkomst. Man kan teste om pigernes menstruationscyklus nu også er 28 dage (uden P-piller), eller om fundne skeletter fra pestens tid har samme gennemsnitshøjde som moderne mennesker. Her bruges t-test. En variation af denne test er en parvis t-test, hvor man har de samme individer/objekter før og efter en hændelse, eller har individer/objekter, der er parrede på en eller anden måde. Her har man 1 variabel, der er forskellen på de målte værdier i parret. Det kunne være forskellen i kondi før/efter en bestemt træningsmetode eller forskellen i indtægt mellem ægtefæller.</td></tr>
<tr><td colspan="3">Dobbelt variabel</td></tr>
<tr><td>Kategorisk data</td><td>Kategorisk data</td><td>Man kan se på 2 variable samtidig og se, om der er en sammenhæng eller afhængighed imellem dem, f.eks. mellem køn og rygning, rygning og kræft, mellem forældres socialklasse og børns uddannelse, mellem regioner og socialklasser osv. Man sætter data op i en krydstabel og tester for uafhængighed af de to variable. Her anvendes en X^2-test.</td></tr>
<tr><td>Kategorisk data</td><td>Numerisk data</td><td>Her vil vi nøjes med at se på tilfælde med 2 kategorier, f.eks. mænd/kvinder, matematikere/samfundsfaglige, trænede/utrænede og så kan man sammenligne højde, vægt, BMI, fraværsprocent, karakterer, testscores, blodprocent. Hertil anvendes t-test.
Fodnote: Vær opmærksom på, at parret t-test hører til under 1 variabel, se ovenfor.</td></tr>
<tr><td>Numerisk data</td><td>Numerisk data</td><td>2 variable kan tegnes ind i et plot, og man kan lave en regression, f.eks. en lineær regression. Eleverne kender R^2, der giver et mål for, hvor meget punkterne spreder sig om linjen. Man kan så teste, om denne linje tyder på, at y-værdierne er uafhængige eller ukorrelerede med x-værdierne, så linjen er vandret, eller om man tværtimod kan se en stigende eller en faldende tendens.
Har anvendes enten en t-test eller en F-test, men de er stort set de samme, idet $F = t^2$.</td></tr>
</table>

Eksemplerne nedenfor er kodede i de forskellige farver.

For hver af disse situationer er der konstrueret en eller flere testmetoder, der resulterer i en *teststørrelse*, der er et tal. Hvordan denne teststørrelse beregnes, skal ikke behandles her. Fælles for alle teststørrelser er, at de er beregnet ud fra hypotesen om samme gennemsnit, ingen tidsmæssige ændringer op eller ned eller uafhængighed. Her har vi jo netop et tal at regne med (forskellen 0, hældningen 0 eller samme forholdsmæssige fordeling i kategorier). Vores stikprøve vil ikke have forskellen/ændringen præcis 0, for verden er ikke *deterministisk*. Der er variationer, og det vil der også være indenfor vores stikprøve. *Teststørrelsen* regner ud, hvor meget vores stikprøve varierer fra den ideelle verden, hvor vores hypotese gælder 100%. Alle teststørrelserne er sådan indrettet, at jo længere vi kommer fra en forudbestemt værdi, jo længere ligger vi fra hypotesens ideelle verden, og jo mindre troværdig bliver hypotesen. Opgaven er så at finde ud af, hvornår teststørrelsen passerer en grænse, hvor vi siger, at nu tror vi ikke længere på den hypotese, som vi har brugt til at udregne teststørrelsen.

Signifikansniveau

Man kan illustrere ideen bag denne grænse med følgende eksempel: Hvis man slår med en terning ved man, at der teoretisk er sandsynligheden 1/6 for en sekser og sandsynligheden 5/6 for en ikke-sekser, hvis man har en ægte terning. Det er så vores hypotese. Man kan så tælle, hvor længe der går, før man får en sekser (hvor man er slået hjem i Ludo, og de andre ræser rundt på banen mod mål). De første par slag uden en sekser accepterer man som uheld, men hvis der bliver ved med ikke at komme seksere, vil man begynde at undre sig og undersøge terningen. Der må være nogen, der har pillet ved terningen. Man kan beregne sandsynligheden for, at man stadig ikke har fået en sekser, men tallet når aldrig ned på 0%, for der er altid en lillebitte chance for, at det alligevel kan ske. Statistik går ud på at finde en grænse, hvor man siger, nu tror vi ikke på hypotesen længere. Vi skal vælge den sandsynlighed, hvor vi siger, at godt nok er hændelsen mulig, men den er så usandsynlig, at vi hellere vil tro på den alternative hypotese. Traditionelt vælger man 5%, og det betyder i eksemplet med terningen, at efter 17 slag uden en sekser stopper vi spillet og mistænker vores modspiller for at spille med falsk terning (udregnet side 26). Vi siger, at resultatet afviger *signifikant* fra det forventede, og vi anvender her et *signifikansniveau* på 5%. I f.eks. medicinalindustrien, hvor man vil være meget sikre på sine resultater, anvender man ofte 1% som signifikansniveau.

Ideen i eksemplet er, at man med statistik fastsætter den grænse, hvor man skifter fra at tro på hypotesen til at tro på alternativet. Antal slag uden en sekser er den simpleste teststørrelse, der findes, og den magiske grænse 17 kalder vi den *kritiske værdi* for denne teststørrelse.

De teststørrelser, der er nævnt ovenfor, behandles med sandsynlighedsregning, og man finder så nogle kritiske værdier, der angiver grænsen ved et hvilket som helst fastsat signifikansniveau. Vi vil kun interessere os for 5% signifikans og 1% signifikans, men alligevel bliver der hurtigt en masse forskellige teststørrelser, kritiske værdier og signifikansniveauer at holde styr på. Før computerne blev almindelige, havde man tabeller, som man måtte slå op i for at finde den kritiske værdi, der svarede præcis til den test og det antal ting/individer/hændelser, som man havde målt på.

P-værdien

Med de moderne computere vil man oftest bruge en enkelt værdi, der har den samme betydning, uanset hvilken undersøgelse man har udført. Dette kaldes *p-værdien*. P-værdien har den fordel, at når først den er regnet rigtigt ud, så er det bedøvende ligegyldigt for læseren, hvilken test der er udført, om det er en af de fem typer ovenfor eller en helt sjette type.

P-værdien måler nul-hypotesens troværdighed, og den angiver sandsynligheden for at få en teststørrelse, der er "længere ude", dvs. mindre sandsynlig end den man har fået – givet at nul-hypotesen er sand. Vi sammenligner nu bare p-værdien med signifikansniveauet 5% eller 1%, alt efter, hvilken vi har valgt. Hvis p-værdien er større end 5%, så er nul-hypotesen troværdig, og vi accepterer den, fordi vi er ikke ”så langt ude”. Omvendt, hvis p-værdien er mindre end 5%, så er vi ”langt ude”, og nul-hypotesen er ikke længere troværdig, og vi vælger den alternative hypotese. I praksis ser man f.eks. ofte P-værdier af størrelsen p< 0,0001, og så er vores test meget signifikant, og vi vælger alternativet. Andre gange kan p-værdien ligge og rode lige under 5%, og så er testen ikke så overbevisende.

Hvis vi f.eks. lader alle skolens elever slå med en ægte terning, indtil de får en sekser, og de hver især skal rapportere, hvor mange slag, det tager, så skal man regne med, at ca. hver tyvende elev (=5%) bruger 17 slag eller mere.

2 forskellige typer fejl

Vi accepterer altså en 5% fejl, hvor vi forkaster hypotesen, selv om den er sand. Det kaldes type I fejl. Samtidig har vi også mulighed for at lave en fejl, hvor vi ikke forkaster hypotesen, hvis den nu vitterlig er forkert. Dette kaldes type II fejl. Hvis man gør sandsynligheden for den første fejl lille, ved at mindske signifikansniveauet, så bliver sandsynligheden for den anden fejl større.

Man kan sammenligne med en helt anden situation fra det virkelige liv: Hvis man skal diagnosticere en slem sygdom, vil man selvfølgelig gerne have de raske bedømt raske og de syge bedømt syge, så de kan behandles. Men der er altid risiko for, at nogle af de syge bliver bedømt raske, så de ikke får behandling, og nogle af de raske bliver bedømt syge, så de bliver forskrækkede og unødigt bliver behandlet. Man skal her gøre sig klart, at mange undersøgelser har en gråzone, hvor det er svært at vurdere, hvorvidt prøven er positiv eller negativ.

Et andet eksempel er en anklaget i en retssag, der dømmes skyldig eller frikendes alt efter hvilken vægt man lægger på bevismaterialet:

	Patienten er Rask	Patienten er syg
Testen viser rask	OK	Katastrofal fejl
Testen viser syg	Ubehagelig fejl	OK (men trist)

	Den anklagede er uskyldig	Den anklagede er skyldig
Den anklagede frikendes	OK	..ærgerligt for ordensmagten
Den anklagede dømmes	justitsmord	OK

	0-hypotese er sand	0-hypotesen er falsk
Vi accepterer 0-hypotesen	OK	Type 2 fejl
Vi forkaster 0-hypotesen	Type 1 fejl	OK

Hvis man er meget bange for at lave en type 1 fejl, så tør man næsten ikke forkaste nul-hypotesen, men så bliver det ekstra svært nogensinde at ende med at vælge den alternative. Historisk har man brugt signifikansniveauet (=sandsynligheden for type 1 fejl) af størrelsen 5% eller 1%, og det er jo den, der definerer grænsen for troværdigheden af nul-hypotesen. Det er meget vigtigt, at man har besluttet sig for sit signifikansniveau, inden man begynder at lave sin undersøgelse, for man må ikke lade resultaterne påvirke sit valg.

Hollywood-setup: du bor i Japan, og der er lovet jordskælv og oversvømmelser sidst på dagen – og mobilnettet er brudt sammen. Der er busser, der skal evakuere området, og du vil gerne vente på din kæreste, for du tror på, at han/hun er i området endnu (din nul-hypotese). I får lov at lede til kl. 16:00, så kører bussen, for projektet er at redde så mange som muligt, og der er jo den mulighed, at kæresten har taget en anden bus (den alternative hypotese). Der er to udfald: I kører med kæresten eller I kører uden kæresten. Det giver 2 mulige katastrofer: I kører uden kæresten, og kæresten dør, fordi han/hun vitterlig var i området, eller 45 mennesker dør, fordi I blev for længe for at lede efter en, der allerede var kørt væk, og derfor blev I indhentet af vandmasserne. De to gode udfald er, at I finder kæresten og kører inden kl. 16:00, eller at I kører og finder kæresten i den anden ende, hvor han/hun forhåbentlig har været bekymret for dig. En vigtig pointe her er, at en eller anden har taget den beslutning, at kl. 16:00 er absolut sidste kørselstidspunkt. Sættes det tidligere, mindskes chancen for at finde kæresten, mens der er større chance for at redde bussen. Flyttes tidspunktet til senere, øges chancen for at finde kæresten, mens chancen for at redde bussen falder. Det svarer præcis til at vælge et signifikansniveau.

Statistik kan bruges, når vi har et projekt, der går ud på at kunne præsentere et resultat, en overskrift, en konklusion: ”Der gælder sådan og sådan”, eller ”vi skal gøre sådan og sådan”. Vi opstiller en hypotese og et alternativ, og vi beslutter os for et signifikansniveau, der fastsætter den grænse, hvor vi tipper fra hypotesen til alternativet. Så følger dataindsamling og udregninger, og til sidst står vi med en P-værdi, der skal sammenlignes med det valgte signifikansniveau: Hvis p er mindre end det valgte tal, forkaster vi hypotesen og vælger alternativet, hvis P er større end det valgte signifikansniveau, holder vi os til hypotesen.

Resume af begreber

Hypotese:

Nul-hypotesen og den alternative hypotese to skal udtrykke den sammenhæng, som man ønsker at undersøge. Nul-hypotesen skal udtrykke, at eventuelle forskelle i resultater kun skyldes tilfældigheder, og alternativet skal så udtrykke, at forskellene har en retning. Nul-hypotesen vil indeholde ord som ”uafhængig af”, ”lig med”, ”ændringen over tid er 0”, hvor alternativet lyder ”afhængig af”, ”forskellig fra”, ”stigende eller faldende over tid”.

Signifikansniveau

Signifikansniveauet er den maksimale sandsynligheden for at forkaste hypotesen, hvis den er sand. Man bruger ofte værdien 0,05 eller 0,01. Det betyder, at man prisgiver de 5% eller 1% af resultaterne, der er ”for langt ude” i forhold til nul-hypotesen, for i stedet at tro på den alternative hypotese.

P-værdi

Når man har fået nogle resultater, benytter man nul-hypotesen til at finde sandsynligheden for at få et resultat, der er længere væk fra uafhængighed/nul/vandret linje end den fundne stikprøve er. Denne sandsynlighed sammenligner man med signifikansniveauet og forkaster hypotesen, hvis P er mindre end signifikansniveauet og beholder hypotesen, hvis P er større.

EKSEMPLER

Her følger et antal eksempler, hvor udregningerne ikke er medtaget i første omgang. Her er kun P-værdier og konklusioner, og pointen med at læse dem er, at se en masse eksempler på tests, uden at skulle tage stilling til, hvordan de er lavet. Af praktiske grunde er eksemplerne delt i chi-i-anden, der er pensum, og de andre tests, der ikke er pensum, men som man måske alligevel kan få brug for.

X^2 –test (kategoriske data med 1 eller 2 variable)

Eksempel 1

En gruppe biologielever vil undersøge, hvor følsomme læberne er, ved at teste, om nogle forsøgspersoner kan føle forskel på 1 pril og 2 prik, de sidste med først 1 mm afstand og senere 2 mm afstand (10 af hver):

Hypotese: Svarene er helt tilfældige (personen kan ikke føle forskel og gætter bare på et tal).
Alternativ hypotese: personen kan føle forskel på 1 og 2 prik.

Sted: læbe		Stik med		Lis
Mm:1		1	2	sum
svar	et	6	4	10
	to	5	5	10
	sum	11	9	20

P= 0,6531 > 0,05
Hypotesen forkastes ikke, så forsøgspersonen har gættet

Sted: læbe		Stik med		Lis
Mm:2		1	2	sum
svar	et	7	3	10
	to	2	8	10
	sum	9	11	20

P=0,0246 < 0,05 (Udregnet side 26)
Hypotesen forkastes, og der er en overvægt af rigtige svar (7 og 8 mod 2 og 3).

Sted: læbe		Stik med		Bo
Mm:1		1	2	sum
svar	et	7	3	10
	to	4	6	10
	sum	11	9	20

P= 0,1775 > 0,05
Hypotesen forkastes ikke, så forsøgspersonen har gættet

Sted: læbe		Stik med		Bo
Mm:3		1	2	sum
svar	et	10	0	10
	to	0	10	10
	sum	10	10	20

P=0
Vi behøver slet ikke at teste dette klokkeklare resultat.

Her har vi 2 variable: stikket og svaret, og fire kategorier: 1stik – svar et, 1 stik – svar to, 2 stik – svar et og 2 stik – svar to.

Eksempel 2

En gruppe samfundsfagselever vil undersøge forældres holdning til 16 års valgret og til alkohol ved skolefester. (Udregnet side 28).

	16- års valgret		
Køn	For	Imod	I alt
Fædre	12	15	27
Mødre	21	6	27
I alt	33	21	54

	Alkohol til gymnasiefester		
Køn	For	Imod	I alt
Fædre	19	8	27
Mødre	21	6	27
I alt	40	14	54

Chi2-test	p-værdi	0,0120

Chi2-test	p-værdi	0,5346

Hypotese: Holdning til …… hos gymnasieelevers forældre er uafhængig af køn

Alternativ hypotese: Holdning til …… hos gymnasieelevers forældre er afhængig af køn

Konklusion: Holdningen til 16-års valgret er forskellig hos fædre og mødre til gymnasieelever, mens holdningen til alkohol er ens.

Vi har igen to variable, køn og standpunkt med fire kategorier: far –for, far –imod, mor –for og mor –imod.

Bemærk, at vi ikke tester noget om selve holdningen til spørgsmålet. Det er altså ligegyldigt, hvor mange der er for og imod. Det vi tester er, om fædre og mødre stemmer på samme måde.

Eksempel 3

En biologiklasse har sået byg med 2 genetiske varianter: grøn/hvid og høj/lav. De sår 90 frø, hvoraf de 87 spirer.

Hypotese: Plantevarianterne udspaltes i forholdet 9 : 3 : 3 : 1

Alternativ: Plantevarianterne udspaltes ikke i forholdet 9 : 3 : 3 : 1

Denne gang håber vi på, at hypotesen bliver bevaret, da det er vores teori. (Udregnet side 28)

	Høje og grønne	Høje og hvide	Små og grønne	Små og hvide	sum
Obs	54	14	12	7	87
Forventet	48,94	16,31	16,31	5,43	

P = 0,365. Konklusion: Hypotesen kan ikke forkaste, så ærterne følger den forventede fordeling.

Her er formålet ikke at teste uafhængighed mellem de to variable. I stedet kigger vi på de forskellige fænotyper, så vi får fire sideordnede kategorier, og vi tester for en bestemt teoretisk, forholdsmæssig fordeling i kategorierne. Denne type X^2 test kaldes Goodness of fit.

Eksempel 4

Bliver man forkølet af at være kold? Videnskab.dk, 9. december 2010

180 personer deles i to grupper, hvor af den ene gruppe sidder med fødderne i koldt vand og den anden sidder med fødderne i en tom balje. Det giver igen 2x2 kategorier.

Hypotese: forkølelse er uafhængig af føddernes temperatur

Alternativ: forkølelse er afhængig af føddernes temperatur

Tallene i parentes er de forventede værdier, hvis hypotesen er sand, og for overskuelighedens skyld er store afvigelser farvet enten røde eller blå. Dette er en rent subjektiv farvemetode, men den kan gøre det lettere at lave en konklusion:

	kolde fødder	kontrol
forkølet	13 (9)	5 (9)
Ikke forkølet	77 (81)	85 (81)
Sum	90	90

P = 0,047

Da p < 0,05, bliver hypotesen forkastet, så forkølelse er afhængig af føddernes temperatur. Det ses af tabellen, at kolde fødder giver mere forkølelse, men det er ikke et overbevisende resultat.

Eksempel 5

En klasse har AT om evolution. De laver forsøg med selektion i forhold til farve, og det gør de ved at samle bønner i 3 farver op fra et papir, der har samme farve som den ene slags bønner. Teorien er, at de bedst tilpassede overlever, og den bedst tilpassede her må være dem, der ligner baggrunden mest.

Hypotese: overlevelse er uafhængig af farve
Alternativ: overlevelse er afhængig af farven i forhold til baggrundsfarven.

Her ses klassens resultater for hvidt karton:

Hvidt karton						
	Bønne	hvid	brun	rød	I alt	
Observeret:	spist	21 (48)	51 (48)	72 (48)	144	
	ikke-spist	379 (352)	349 (352)	328 (352)	1056	P < 0,0001

Hypotesen forkastes. Der er altså en selektion, og det ses af resultaterne, at på hvidt karton er det de hvide, der bedst overlever.

Eksempel 6

Vold mod unge i Danmark, SFI 15.12.2010

Hypotese: vold mod børn er uafhængig af forældres og børnenes køn
Alternativ: vold mod børn er afhængig af forældres og børnenes køn

	Vold begået af		
	Mor	Far	
Piger	132	145	277
Drenge	134	146	280
	266	291	275

P = 0,96 > 0,05, så hypotesen forkastes ikke.

Konklusion: Mor og far er lige slemme til at slå, og de slår lige meget på drenge og piger.

Eksempel 7

I en sexlivsundersøgelse fra 2009 af mænd, der alle har været sammen med en mandlig partner, spørger man om HIV status, bopæl, og prævention. Nedenfor er vist 3 tabeller, hvor hypotesen i de to første tabeller er, at HIV-status er uafhængig af hhv. bopæl og beskyttelse, og i den tredje tabel er, at beskyttelse er uafhængig af alder.

Tabel 1	HIV +	HIV -	ukendt	
København	74 (58)	434 (449)	65 (66)	
Odense, Århus, Ålborg	10 (20)	168 (156)	21 (23)	
Udenfor store byer	17 (22)	174 (171)	28 (25)	P = 0,0134135

(Udregnet side 30)

Tabel 2	sikker sex	usikker sex
HIV +	38 (69)	70 (39)
HIV -	651 (569)	246 (328)
ukendt	39 (90)	103 (52)
	P =	5,03798E-33

Tabel 3	sikker sex	usikker sex
< 30 år	275 (198)	136 (213)
30 - 50	326 (264)	222 (284)
> 50	115 (254)	411 (272)
	P =	4,81337E-51

Konklusion: Første tabel viser en overrepræsentation af HIV+ mænd i København. Tabel 2 viser, at mænd, der er testet HIV – er mest tilbøjelige til at dyrke sikker sex, set i forhold til mænd, der er testet HIV+ og mænd, der ikke er testet. Tabel 3 viser, at det især er mænd over 50 år, der sløser med den sikre sex. Alle tre resultater er statistisk signifikante, da P < 0,05.

Eksempel 8

Mænd er så kloge - synes de selv *Søndag den 14. november 2010*

I en Gallupundersøgelse, Berlingske Tidende har fået lavet, svarer 54 procent af de danske mænd, at de er mere intelligente end de fleste andre. 35 procent af kvinderne svarer det samme.

	Klogere end de fleste	Ikke klogere end de fleste
Mænd	56 (45)	48 (59)
Kvinder	48 (59)	89 (78)

Hypotese: Mænd og kvinder har samme vurdering af sig selv

Her fås en p-værdi på 0,0035, så der er forskel på mænds og kvinders vurdering af sig selv.

Men er det nu mændene, der synes, at de er så kloge? Kunne det ikke lige så godt være kvinderne, der undervurderer sig selv? (Diskuter spørgsmålets formulering: ”klogere end de fleste” og evt. også medianens betydning her).

Eksempel 9

Mænd har mere ferie end kvinder, Newsdesk tic.travel, 17/01/2011 (hotels.com)

"Mens kun knap hver tredje kvinde (32 pct.) kan nyde 6 ugers ferie eller mere, gælder dette for næsten hver anden mand (44 pct.)."

Der er 705 kvinder og 448 mænd i undersøgelsen (oplyst af hotels.com).

	Kvinder	Mænd
≤ 5 ugers ferie	176 (165)	117 (128)
≥ 6 ugers ferie	226 (237)	197 (186)
P = 0,078		

Hypotese: mænd og kvinder har samme ferielængde. Denne hypotese kan ikke forkastes på 5% niveau, men den kan godt forkastes på 10% niveau, og 10% kan være godt nok til en markedsføringsstrategi.

"6 pct. af de danske mænd har således ferie mere end 12 uger om året. Dette gælder kun 3 pct. af kvinderne."

	Kvinder	Mænd
< 12 ugers ferie	378 (372)	285 (291)
≥ 12 uger	24 (30)	29 (23)
P = 0,098		

De små procenter og den store forskel på antallet af adspurgte mænd og kvinder betyder, at resultatet kun med nød og næppe er signifikant på 10% niveau.

Den interessante forskel ligger i ansættelser med og uden aftaler om ferie (tal oplyst af hotels.com):

	Kvinder	Mænd
job med ferie	402 (438)	314 (278)
udenfor feriesystemet	303 (267)	134 (170)
P < 0,001		

Hypotese: mænd og kvinder fordeler sig ligeligt på jobs med og uden ferie-aftaler. Den forkastes overbevisende. Det ses af tabellen at flest mænd har job med ferie-aftaler.

Eksempel 10

Unge muslimer råhygger uden alkohol videnskab.dk 19. november 2010

Man har undersøgt unge muslimers og unge danskeres alkoholvaner. Fra undersøgelsen har vi blandt andet:

Muslimer	Har aldrig drukket alkohol	Har drukket alkohol	
mænd	21 (27,7)	34 (27,3)	55
kvinder	41 (34,3)	27 (33,7)	68
	62	61	123

P = 0,0124 (Udregnet side 30)

Protestanter	Har aldrig drukket alkohol	Har drukket alkohol	
mænd	15 (14,1)	861 (861,9)	876
kvinder	16 (16,9)	1029 (1028,1)	1045
	31	1890	1921

P = 0,75

Hypotesen er, at der ikke er forskel på mænds og kvinders alkoholvaner blandt hhv. muslimer og protestsanter. Konklusionen er, at der er forskel på unge muslimske mænd og kvinders forsøg med alkohol, mens de unge danskere ikke adskiller sig på køn.

Eksempel 11

Placeboeffekt, BioNyt nr. 151, februar 2011

Placebo anvendes ofte i studier af medicin, hvor man lader halvdelen af forsøgsdeltagerne få medicinen og den anden halvdel få placebo, der ikke indeholder noget virksomt stof.

I det følgende studium har man derimod undersøgt selve effekten af placebo ved at lade halvdelen af forsøgsdeltagerne få placebo, som de ved er placebo, mens den anden halvdel intet får.

	Placebo	Intet	sum	
Bedring	22	15	37	
Uændret	15	28	43	
	37	43	80	P = 0,0279405

Hypotese: ingen effekt af placebo

Konklusion: P < 0,05, så der er påvist en effekt af placebo. Se dog artiklen, hvor der er opremset en del indvendinger mod forsøget.

Eksempel 12

Flere kvinder blev gravide efter klovnebesøg, Videnskab.dk, 20.01.2011

På fertilitetsklinikker har man det problem, at selv om man overfører befrugtede æg til kvinden, så er det ikke sikkert, at de sætter sig fast i livmoderen. En klinik prøvede at lade en klovn besøge kvinderne lige efter at ægget blev sat op. Det forbedrede chancen for graviditet:

	gravid	ikke-gravid	sum
klovn	40 (31)	70 (79)	110
ikke-klovn	22 (31)	87 (78)	109
sum	62	157	219

Hypotese: klovnene gør ingen forskel, Alternativ: Klovnene gør en forskel.

P = 0,007873, så hypotesen forkastes. Det ses, at klovnene forbedrer sandsynligheden for at bliver gravid. Effekten er formodentlig, at kvinderne slapper af. Der forlyder ikke noget om, om der er forskel på han-klovne og hun-klovne.

Eksempel 13

Sammenligning af to prognoser fra Altinget.dk. Megafon fra 01.04.2011 med 1022 adspurgte og Voxmeter fra 08.03.2011 med 2011 adspurgte.

	V	K	O	LA	S	SF	R	EL
Megafon	222	56	136	34	306	152	71	34
Voxmeter	225	68	158	72	361	187	70	48

Hypotese: Prognosen er uafhængigt af analyseinstituttet.

P = 0,080, så hypotesen forkastes ikke. De forskellige institutter får ikke forskellige resultater.

Eksempel 14

Skolebørnsundersøgelsen 2010. denne undersøgelse er en guldgrube af mulige tests, da man har været meget omhyggelig med at skrive, hvor mange personer, der indgår i hver undersøgelse. Undersøgelsen omfatter over 4000 elever i et repræsentativt udsnit af 11-, 13- og 15 elever i landet. Den gennemføres i 41 lande. De variable er køn, alder og socialgrupper, og af den sidste er følgende antal repræsenteret:

Socialgruppe	I	II	III	IV	V	VI
antal	292	1335	626	1126	584	209

Der er stillet mange spørgsmål og her er valgt et:

Socialgruppe	I	II	III	IV	V	VI
Er ude med venner mindst 3 aftener om ugen	53	280	131	293	164	63
ikke ovenstående	239	1054	494	833	420	146

P Hypotese: der er ikke forskel på 11-, 13- og 15-åriges adfærd i de 6 socialgrupper.

P = 0,000056, så hypotesen forkastes.

Andre typer tests: (☐ numerisk og ☐ kategorisk variabel)

Eksempel 15 (binomial test eller parvis t-test) ☐

Samfundsfagsklassen undersøger forældrene alder. De mente, at fædre er ældre end mødre, men hypotesen skal have lighed, så:

Hypotese: Fædre og mødre har samme alder
Alternativ: Fædre og mødre har ikke samme alder

Fars alder	41, 47, 47, 45, 42, 47, 45, 46, 53, 50, 50, 47, 53, 48, 48, 47, 48, 47, 53, 50, 49, 60, 47, 57, 48
Mors alder	40, 45, 40, 42, 51, 43, 45, 44, 47, 49, 46, 47, 57, 51, 45, 42, 50, 43, 52, 49, 45, 54, 43, 56, 48

Der er i alt 25 forældrepar. 20 hvor far er ældst, og 5 hvor mor er ældst (jeg har spurgt nærmere ind til de jævnaldrende). Hvis vi kun tager hensyn til ældre/yngre laves en binomialtest, hvor P = 0,002. Tager man også hensyn til aldersforskellens størrelse, laves parvis t-test, hvor P = 0,029, så hypotesen forkastes i begge tilfælde, og forældrene har ikke samme alder, og det ses af tabellen, at far er ældst. (Gennemregnet side 31).

Eksempel 16 (t-test) ☐☐

Eleverne i en biologiklasse laver farve-konflikttest. De skal læse farverne på ord, f.eks. læses GRØN som rød. Det siges, at pigerne er bedre til det end drenge, men vores hypotese er, at drenge og piger er lige gode, og alternativet er, at der er forskel.

	Antal	Tid, sec.	Gennemsnit
Piger	16	42, 60, 60, 56, 89, 59, 51, 70, 56, 68, 46, 60, 69, 59, 48, 64	60 sec
Drenge	8	45, 56, 75, 75, 70, 68, 65, 80	67 sec

Resultaterne er mildest talt modsat det, der siges, men vi kan jo godt teste hypotesen alligevel, for måske er drengene hurtigst. Vi laver t-test og får P = 0,18. Vi kan altså ikke forkaste hypotesen, så der er ingen forskel på drenges og pigers evner i denne opgave. (Udregnet side 33)

Eksempel 17 (t-test) ☐☐

Samf-klassen er træt af at høre, at lærerne roser mat-fys elevernes fremmøde og kritiserer samf-elevernes pjækkeri, så de får deres matematiklærer til at finde fraværsprocenterne i de to klasser. Matematiklæreren sorterer procenterne, så de bliver mere anonyme:

samf	0,45	1,61	5,6	6,48	6,72	8,71	9,11	9,66	11,5	12,19	12,4	13,09
	14,34	14,4	17,46	17,68	17,82	18,15	18,57	20,76	22,31	22,49	22,54	22,69
Ma-fy	0	0,22	4,71	4,87	5,33	7,69	7,69	7,76	7,93	8,62	10,02	10,49
	10,96	11,66	12	12,67	12,82	15,23	15,38	15,49	15,76	16,25	17,45	18,41

Samfernes hypotese er, at der ikke er forskel på klassernes fraværsprocent, og de får da også P = 0,0654, selv om middelfraværet er 13,6% i samf-klassen og 10,4% i mat-fys-klassen. Men hypotesen bliver altså ikke forkastet, så lærerne må holde op med at hakke på den stakkels samf-klasse, synes de.

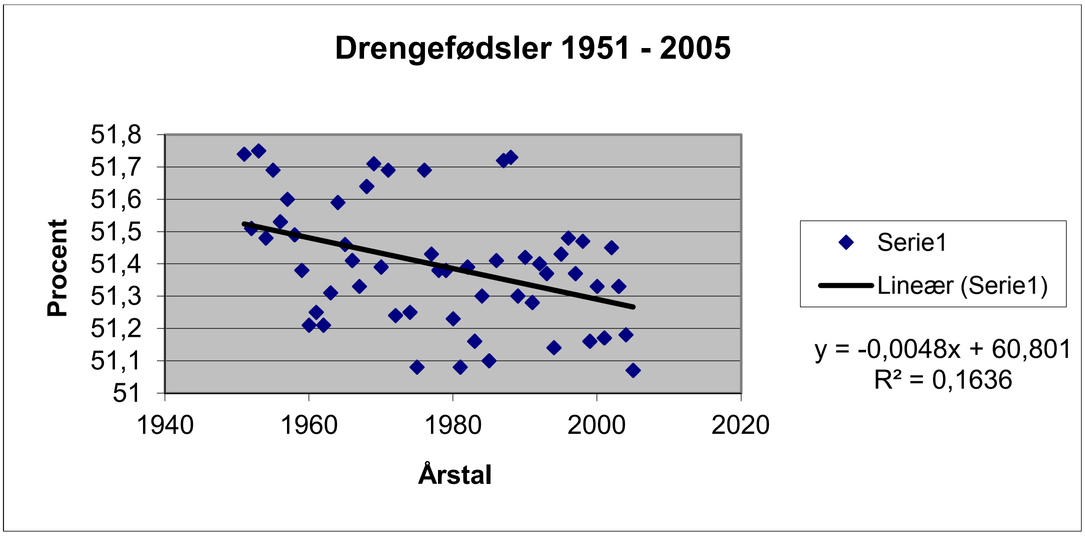
Mat-fys'erne påpeger, at data ikke er normalfordelte i de to grupper. Derfor laver de en simulering af data, hvor de blander tallene 2032 gange og finder 77 tilfælde, hvor samfklassen har højere fravær end den fundne. Det giver P = 0,0379. (Udregnet side 33)

"Hov", siger samf'erne, "I ser jo kun på det tilfælde, hvor vi har højere fravær end jer. Det kaldes en ensidig test. Vi ser også på det tilfælde, hvor jeres fravær er højest. Det er en to-sidig test. Så er det da ikke mærkeligt, at vores test giver en P-værdi, der er dobbelt så stor som jeres. (Se en gennemgang af en- og to-sidige tests under beregninger, side 34).

Eksempel 18 (test for lineær regression)

Vi har i mange år sagt, at der bliver født flere drengebørn end pigebørn. Gamle statistikbøger bruger tallene 52% drenge og 48% piger. Fra diverse statistiske årbøger kan vi finde procentdelen af drenge fra 1951 – 2005. Der laves lineær regression, og R^2 = 0,16. Det er ikke en imponerende R^2. Vi laver en test:

Hypotese: Procentdelen af drengebørn er uændret i perioden 1955 – 2005.
Alternativ: Procentdelen af drengebørn har ændret sig i perioden 1955 – 2005.

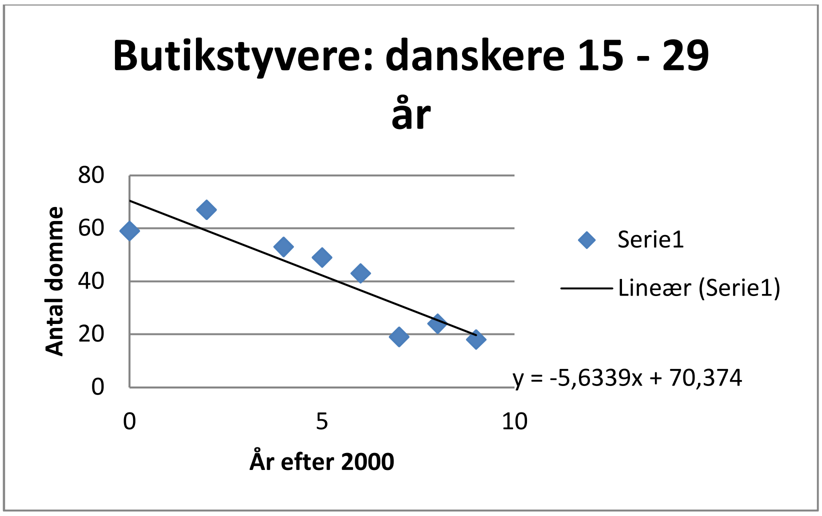


P = 0,0024, så vi forkaster altså hypotesen, og ser på figuren, at procentdelen er faldende. Så selv om punkterne er meget spredt, så er der er et signifikant fald.

Eksempel 19 (test for lineær regression)

En samfundsfagsklasse har undersøgt antallet af domme for butikstyveri blandt 15-29 årige i perioden 2000 – 2009 (fra Statistikbanken).

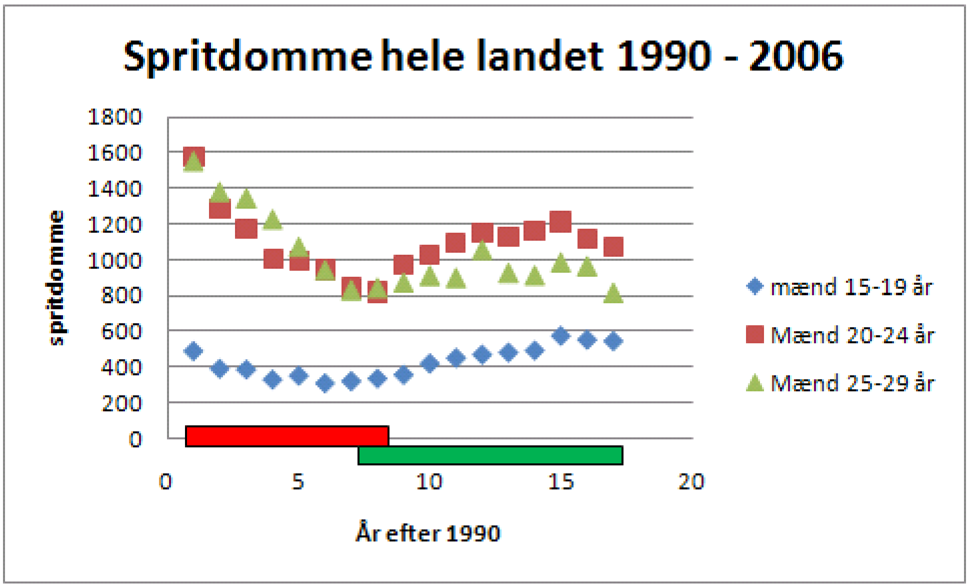
	2000	2002	2004	2005	2006	2007	2008	2009
15-29 år	59	67	53	49	43	19	24	18



Hypotese: antal domme er uændret. Alternativ hypotese: der er en systematisk ændring over årene.
P = 0,0019 < 0,05, så hypotesen forkastes, og der er en tendens, og den ses at være en faldende.

Eksempel 20 (test for lineær regression)

En samfundsklassen har undersøgt antal unge mænd sigtet for spritkørsel i Danmark i perioden 1990 – 2006 (fra Statistikbanken).



De beslutter at dele data i to perioder: 1990 – 1998 og 1997 – 1006, hvor de lader de to perioder overlappe med 2 år. Bagefter skal klassen undersøge, hvad der skete i 1998-99, der ændrede billedet.

Hypotese: tendenslinien i perioden er vandret, dvs. ændringen i antal domme er tilfældig spredning

Alternativ: tendenslinien er ikke vandret, så der er en stigende eller faldende tendens.

De 6 forskellige tests er skrevet op i en tabel:

	1990-1998	1997-2006
Mænd 15-19 år	P = 0,033	P < 0,001
Mænd 20-24 år	P = 0,003	P = 0,02
Mænd 25-29 år	P< 0,001	P = 0,62

(Udregnet side 36)

Konklusion: Der er et tydeligt fald i antal spritdomme for mænd mellem 20 og 29 år i årene 1990 – 1998. For unge mellem 15 og 19 er der kun et svagt fald i samme periode. I perioden 1997 – 2006 er antal spritdomme for mænd mellem 25 og 29 stabil, mens de helt unge ifølge grafen har en tydelig stigning. Det er ikke fordi stigningen er så stor, men den tilfældige variation er meget lille, så punkterne ligger tæt på en ret linie med en lille stigning. Mændene mellem 20 og 24 viser en stigning, men det havde nok set anderledes ud, hvis eleverne havde valg at starte den anden periode i 1999 i stedet for 1997. Eleverne går så i gang med at finde ud af, om der er sket noget specielt i 1997-98, der har påvirket spritdommen. Det havde her været mere korrekt at tilpasse en parabel i stedet for en ret linje.

Eksempel 21 (test for lineær regression)

Medlemmerne flygter fra de konservative, Jyllands-Posten 12.marts 2011.

Nul-hypotese: vælgersvingningerne er udtryk for tilfældigheder
Alternativ hypotese: der er en trend

År	Årefter2005	medlemmer
2005	0	19478
2006	1	18035
2007	2	16707
2008	3	15928
2009	4	15545
2010	5	13864

P = 0,000315 (Udregnet side 37)

Konklusion: P < 0,05 så der er en trend, der ses at være faldende medlemstal.

Eksempel 22 (test for lineær regression)

Milliardindsats på usikkert grundlag, Jyllands-Posten 14.april 2011.

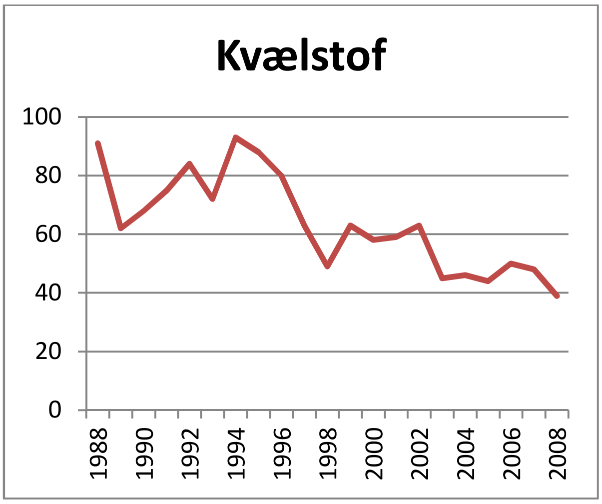
Vandmiljøplanen bygger på en formodning om, at en nedsættelse af kvælstofudledningen fra markerne vil give forbedrede tilstandene i vandmiljøet. Udbredelsen af ålegræs er valgt som succeskriterium, idet man ved, at denne plante vokser bedst og på størst dybde, hvis vandet er rent og klart og har et lavt indhold af kvælstof. Nedenstående kurver viser hhv. kvælstofindholdet (mmol/L) og maximale vanddybde med ålegræs (m) ved Skive. Det ser ud som om at kvælstofmængden falder, mens udbredelsen af ålegræs ikke ændrer sig – og den burde jo være steget.

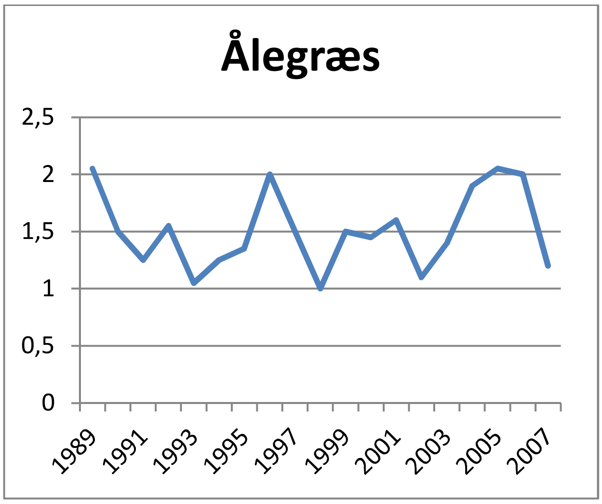
Hvis vores nul-hypotese er ingen ændring gennem årene, får vi p = 0,00002 for kvælstof og p = 0,57 for ålegræs. Så vores formodning holdt stik: kvælstofmængden er faldet signifikant, mens ålegræsdybden ikke har ændret sig signifikant. Fra 1988 til 2008 er faldet i kvælstof ret præcist 50%, hvis man bruger en lineær model ($R^2 = 0,62$).

Tallene bygger på aflæsning fra avisens kurve – det burde selvfølgelig være alle målte værdier:

	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
N-konc	91	62	68	75	84	72	93	88	80	63
Ål-dyb		2,05	1,5	1,25	1,55	1,05	1,25	1,35	2	1,5

1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
49	63	58	59	63	45	46	44	50	48	39
1	1,5	1,45	1,6	1,1	1,4	1,9	2,05	2	1,2	





Figurer fra analyser – specielt for samfundsfag

Eksempel 23

Strukturreformen udfordrer nærdemokratiet | Krevi 26. jun 2009

Man ser på hhv sammenlagte kommuner og på kommuner, der er blevet sammenlagt. Man har så spurgt til borgernes tillid til kommunen i 2001 og 2007. Hypotesen er, at tilliden er uafhængig af kommunesammenlægningen. Vi ved ikke, hvilken analysemetode, der er blevet anvendt, men P-værdien burde give mening:

Tabel 1. Udviklingen i lokalpolitisk tillid i sammenlægnings- og fortsætterkommuner (gennemsnitlige indeksværdier, matchet stikprøve)

År / Kommunegruppe	2001	2007	Ændring 2001-2007	Forskel i ændring
Sammenlægning	58,6 (n=100)	47,5 (n=602)	-11,1	-12,3***
Fortsætter	55,3 (n=392)	56,5 (n=594)	1,2	

****Signifikant på 0,01-niveau.*

Tabel 2. Udviklingen i lokalpolitisk selvtillid i sammenlægnings- og fortsætterkommuner (gennemsnitlige indeksværdier, matchet stikprøve)

År / Kommunegruppe	2001	2007	Ændring 2001-2007	Forskel i ændring
Sammenlægning	61,4 (n=139)	51,8 (n=590)	-9,6	-6,6 **
Fortsætter	57,8 (n=526)	54,9 (n=599)	-3,0	

***Signifikant på 0,05-niveau.*

Konklusionen er, at borgerne har mistet tillid til kommunerne, der er sammenlagt. Selvtilliden opfører sig på samme måde, men resultatet er knap så markant.

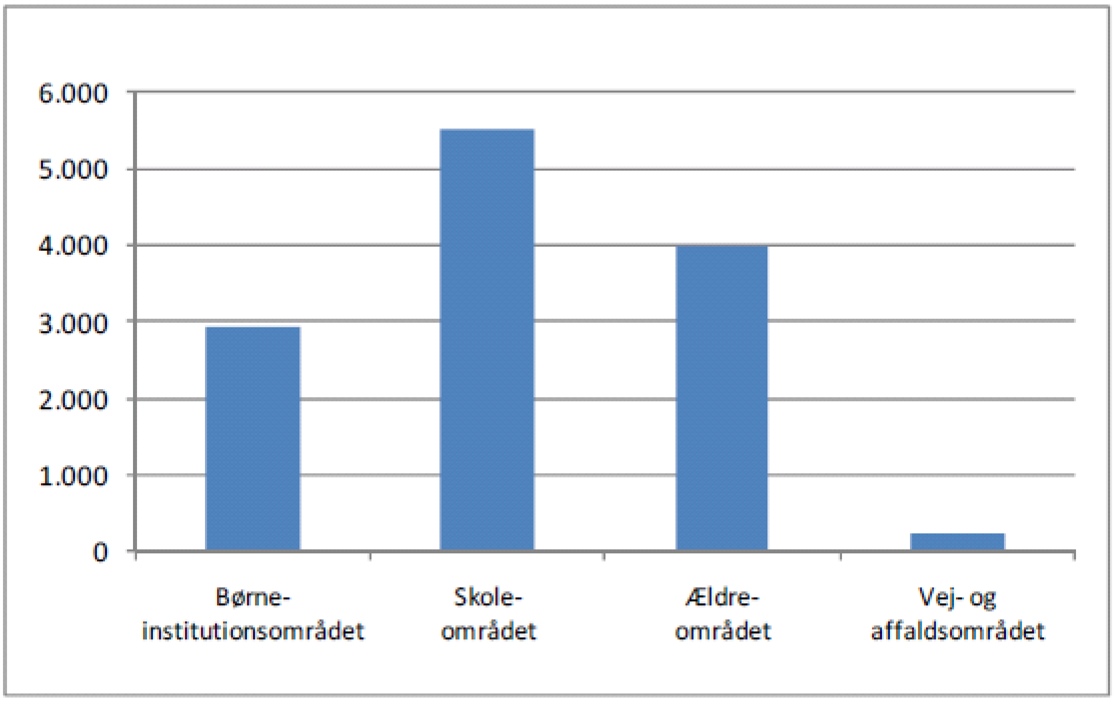
Eksempel 24

Brugeres og borgeres præferencer for kommunale serviceydelser| Krevi

01.10.10 | Danskerne vil ikke betale hvad som helst for at få forbedret en hvilken som helst kommunal serviceydelse. Og der er store forskelle i betalingsviljer mellem forskellige befolkningsgrupper. Det viser KREVIs undersøgelse af 5.824 danskeres præferencer for kommunale serviceydelser.

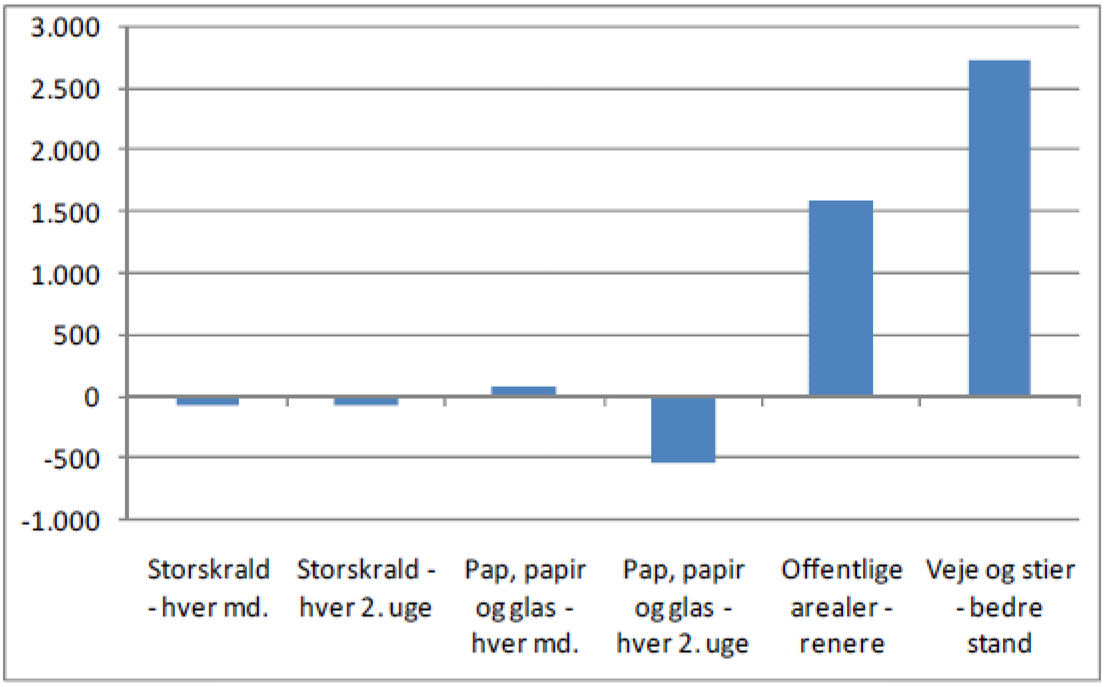
Deltagerne i undersøgelsen er blevet stillet forskellige spørgsmål, der skal vise, hvor mange kr. de er villige til at betale pr. husstand pr. år. for nogle ydelser. KREVI fortæller, at der er lavet nogle helt specielle analyser af spørgeskemaerne, og det kræver nogle meget avancerede statistiske metoder. Vi må nøjes med at se på P-værdierne. Hypoteserne er, at folk ikke vil betale for mere service. De tilfælde hvor P > 0, er specielt nævnt under figuren. Der er beløbet ikke signifikant forskel fra 0, så det er selvfølgelig de meget lave søjler. N er antal svarpersoner, og det kan variere fra spørgsmål til spørgsmål.

Figur 1.5: Gennemsnitlig betalingsvilje for udvalgte serviceforbedringer, tværgående prioritering, skattebetaling, kr./husstand/år



Note: Betalingsviljen for vej- og affaldsområdet er ikke signifikant forskellig fra 0. N = 5.824

Figur 1.4: Gennemsnitlig betalingsvilje for serviceændringer på vej- og affaldsområdet, skattebetaling, kr./husstand/år



Note: Betalingsviljerne for afhentning af storskrald hver måned eller hver anden uge, samt for afhentning af pap, papir og glas hver måned er ikke signifikant forskellige fra 0. N = 3.266

Desuden er der lavet diverse sammenligninger, hvoraf der her er et resume hentet på hjemmesiden:

Undersøgelsen viser **klare forskelle mellem befolkningsgrupper**:
• Venstreorienterede vil typisk betale mere end højreorienterede for serviceforbedringer.
• Kvinder vil typisk betale mere end mænd.
• Brugere vil betale mere end ikke-brugere.
• Jo tættere, vi kommer på pensionsalderen, des mere vil vi betale for praktisk hjælp til ældre.
• Sjællændere vil betale mere for økologi i frokostordninger og for at få hentet deres storskrald og haveaffald end jyder vil.

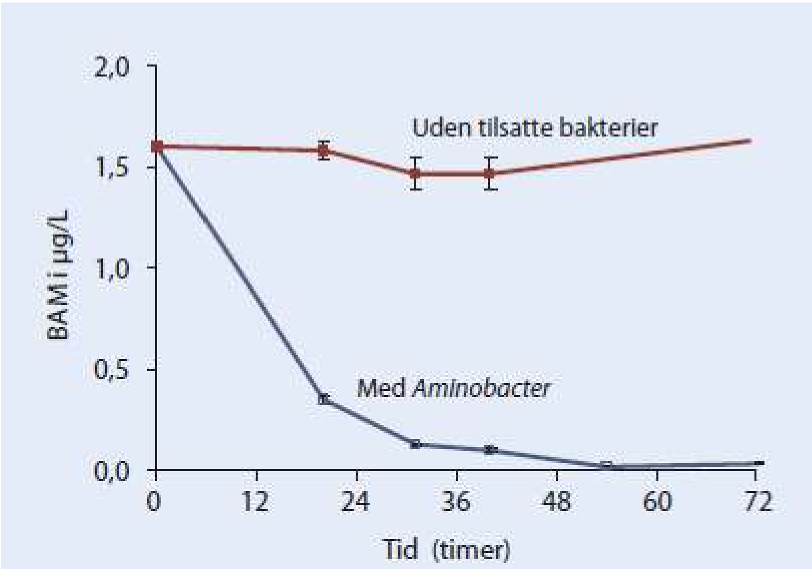
Figurer fra videnskabelige artikler – specielt for biologer

Her følger en kort gennemgang af, hvorledes data kan blive fremstillet grafisk i en popularvidenskabelig artikel, uden at man egentlig behøver at nævne de udførte tests. Der gives også en nem tommelfingerregel, der kan bruges, når man ser på de publicerede figurer, men som ikke må bruges som test, da den ikke er helt nøjagtig.

Eksempel 25

Bakterier renser drikkevand , A k t u e l N a t u r v i d e n s k a b | 6 | 2 0 0 6

Man undersøger, om bakterier (Aminobacter) fjerner stoffet BAM fra drikkevandet. Der bliver selvfølgelig lavet en masse små delforsøg, altså glas med bakterier og glas uden bakterier. Efter 20 timer, 30 timer og 40 timer måler man så mængden af BAM (μg/L) i hvert af de mange glas. Man udregner gennemsnittet af glassene uden bakterier og gennemsnittene af glassene med bakterier, og desuden beregner man spredningen. Spredningen divideres med kvadratrod n, og hedder nu *Standardfejl*. Der er her lavet 3 t-test til tre forskellige tidspunkter.
Gennemsnittene afsættes med en figur, her små kvadrater, og standardfejlen tegnes lodret op og ned, evt. afsluttet med en vandret streg. Den lodrette streg er altså 2 standardfejl lang. (2 er meget tæt på de 1,96, som du måske kender fra normalfordelingen).



De rå tal kan se sådan her ud for målingerne ved 20 timer:

Ved 20 timer	BAM (ug/L)	kommentar
uden bakterier	1,573; 1,552; 1,489; 1,633; ….osv	Gennemsnit af alle tallene er ca. 1,6
med bakterier	0,431; 0,397; 0,411; 0,405; ….osv	Gennemsnit af alle tallene er ca. 0,4

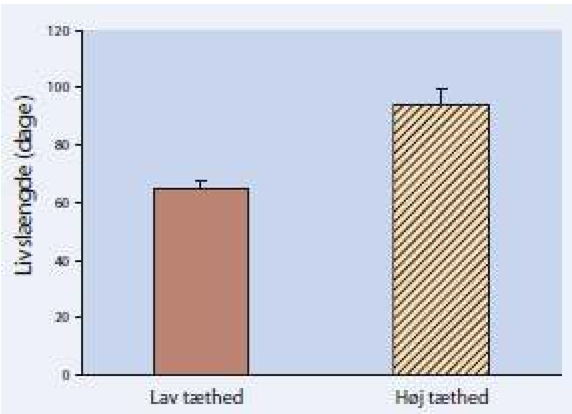
Man laver en t-test, hvor hypotesen er, at mængden af BAM er ens med og uden bakterier efter 20 timer. Her vil man få P<0,001, og man forkaster hypotesen. Det gør man for alle tre tidspunkter, og konklusionen er så, at bakterierne fjerner BAM fra drikkevandet.

Der er en øjemålsregel her, som siger, at hvis de lodrette streger fra standardfejlene når hinanden, så forkaster man IKKE hypotesen, og hvis de lodrette streger ikke når hinanden, så forkaster man, og der er forskel på gennemsnittene. I dette eksempel ligger gennemsnittene og stregerne langt fra hinanden, og det passer med artiklens konklusion.

Eksempel 26

Når stress er sundt, A k t u e l N a t u r v i d e n s k a b | 6 | 2 0 0 4

Når mange bananfluelarver vokser op på et lille område (høj tæthed), så får de signifikant højere levetid (livslængde):

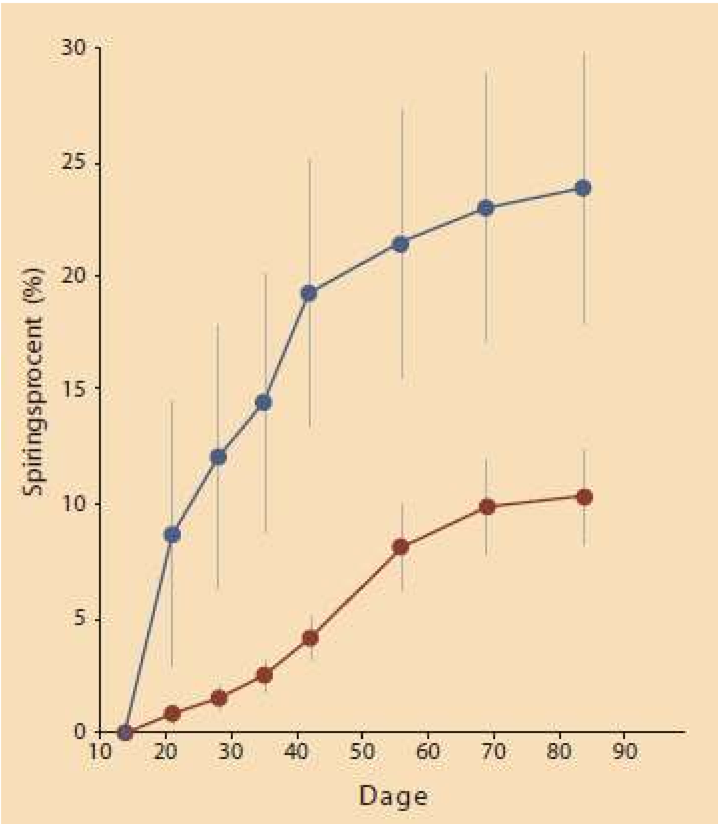


Vi ser to søjler med forskellig højde, og vi ser nogle meget små standardfejl, der dog kun er tegnet opad, men man kan selv forestille sig, at man tegner dem symmetrisk på søjlens overligger, og at man skubber søjlerne sammen, så de ligger oveni hinanden. De to standardfejl vil så aldrig kunne nå i nærheden af hinanden, så de to livslængder for hhv. lav og høj tæthed er altså signifikant forskellig. Der er lavet en t-test.

Eksempel 27

Indavl og undergang , AktuelNaturvidenskab | 5 | 2003

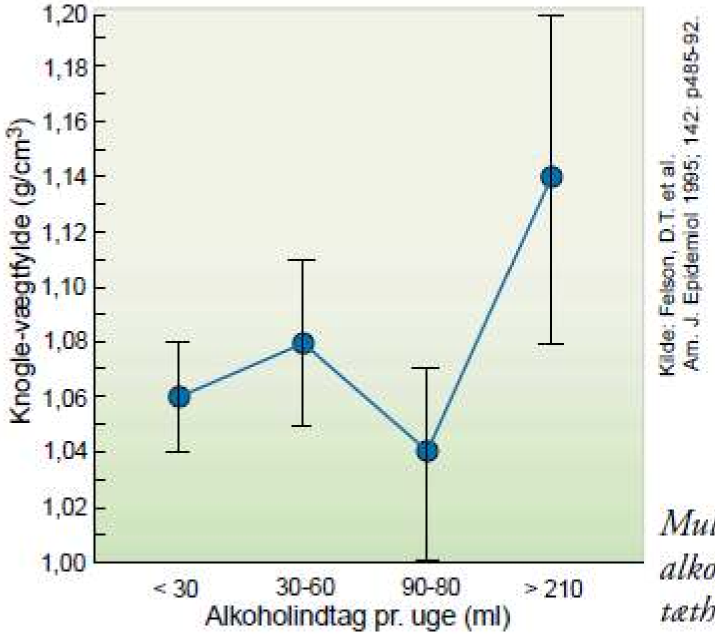
Planter, der er meget sjældne og kun findes i små spredte grupper, bliver let indavlede. Man har fundet to grupper af planten stenbræk, og så har man prøvet at krydsbestøve dem. Man undersøger frugtbarheden (spiringsprocenten) hos de indavlede (røde) og de krydsede (blå) planter. Der erlavet 7 t-tests, en for hver af de målte dage. Artiklen konkluderer, at de krydsede planter har signifikant større spiringsprocent end de oprindelige, og det passer med, at de lodrette standardfejl ikke overlapper.



HUSK: dette med at se på overlappende lodrette streger er <u>ikke</u> tilstrækkeligt, hvis man selv laver undersøgelsen og har ansvaret for resultaterne, men den er fortrinlig, hvis man bare er tilskuer, og ser på andre folks artikler og skal have et overblik.

EKSEMPEL 28

Vin og kvinder, Aktuel Naturvidenskab 2| 2000



Mulig sammenhæng mellem alkolforbrug og knoglernes tæthed.

Man mener, at knoglerne bliver mere skøre hos mennesker, der drikker meget alkohol, men denne her figur viser IKKE en klar sammenhæng, da standardfejlene overlapper eller næsten overlapper.. Artiklen foreslår da også, at man skal forske noget mere i denne sammenhæng. Der er lavet en test, hvor man kan sammenligne 4 variable.

Eksempel 29

Kvinder smækker røret på for at undgå indavl, Videnskab.dk, 7. december 2010

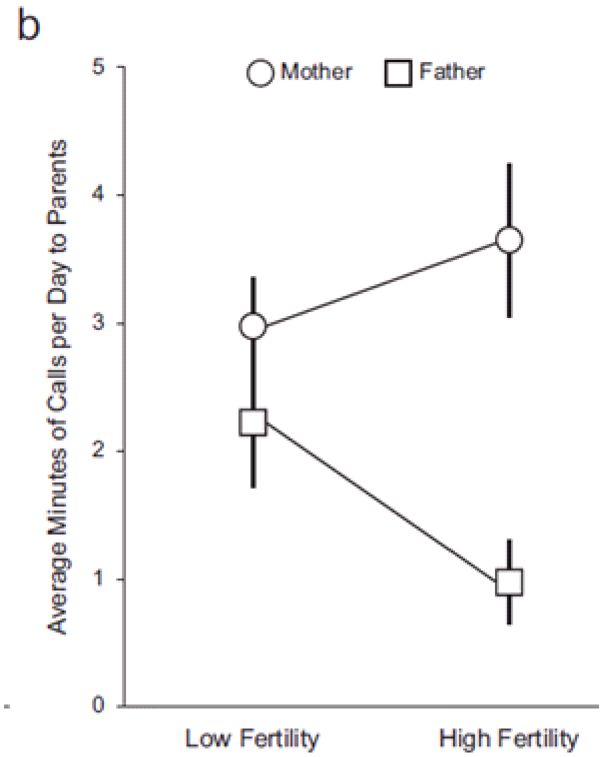
Far bliver frosset ud, og mor bliver forfremmet til personlig rådgiver, når kvinder er mest frugtbare. Det sker ubevidst, men det forhindrer indavl.

Studiet går ud på at se, om unge kvinder omkring ægløsningstidspunktet, hvor de er mest fertile, prøver at undgå deres far. Man har valgt at se på telefoni, og ser på længden af samtaler med henholdsvis fædre og mødre.

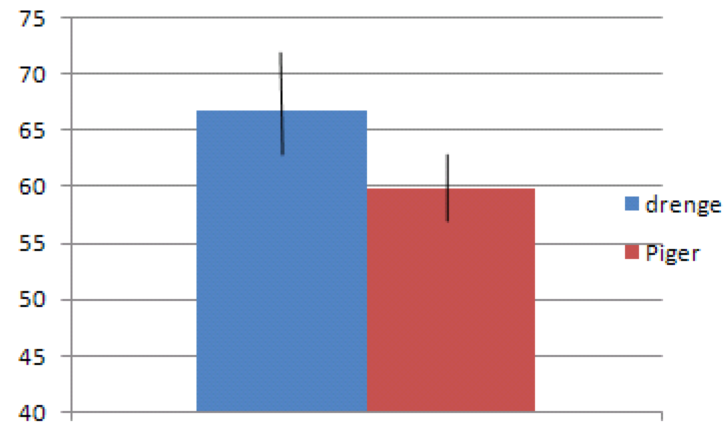
Fra hver kvinde ser man på, hvor længe hun snakker med sin mor, når hun ikke har ægløsning (low fertility), og hvor længe hun snakker med mor, når hun har ægløsning (high fertility), hvor længe hun snakker med far, når hun ikke har ægløsning (low fertility), og hvor længe hun snakker med far, når hun har ægløsning (high fertility). Man har spurgt 48 unge kvinder, og beregnet gennemsnittet og standardfejlene af de 4 grupper af tal. De hypoteser der undersøges, er, at længden af samtalerne er ens, ligegyldigt hvilke to grupper tal vi sammenligner.

På figuren ses til venstre, at kvinderne i gennemsnit snakker med mor i 3 minutter og med far i 2,1 minut ved lav fertilitet. Linjerne for standardfejlene støder sammen, så vi gætter på, at der ikke er nogen signifikant forskel på de to slags samtaler. Artiklen oplyser, at hvis der laves t-test, fås netop P > 0,05 (forkast ikke hypotesen). Kigger vi i stedet på samtaler omkring ægløsning, så ser vi, at kvinderne snakker 3,5 minut med mor og 1 minut med far, og de lodrette streger mødes ikke. Det passer med, at hvis vi laver en test her, får vi P < 0,001 (forkast). Så konklusionen er, at der i høj grad er forskel på samtaler med fædre og mødre omkring ægløsning, og den forskel ses ikke udenfor ægløsningstidspunktet.

Sammenligner vi i stedet samtalerne med mor under og udenfor ægløsning, så ser man, at de lodrette streger ville overlappe, hvis man skubbede dem hen til hinanden, hvorimod de tilsvarende streger for far ikke ville overlappe, hvis de blev skubbet sammen. Hvis man tester forskellen mellem 3 og 3,5 minutter for mor, får man P > 0,05, så forskellen på 3 og 3,5 er altså ikke signifikant. Tester man derimod taletiden med fædre (2,1 ved lav fertilitet mod 1,0 ved høj fertilitet), så får man P < 0,05. Så den samlede konklusion på alle testene er, at samtaletiden med far under ægløsning er signifikant forskellig fra de andre forældresamtaler, der stort set er ens.



Eksempel 30 (det er eksempel 16 i ny forklædning) Farvekonflikttest



Her er resultaterne tegnet med gennemsnit og standardfejlene. Bemærk at standardfejlene næsten kunne rører hinanden, hvis søjlerne blev skubbet sammen. Der blev heller ikke fundet nogen statistisk forskel mellem drenge og piger.

Gennemregninger af eksemplerne

Eksemplet med terningeslag:

Sandsynligheden for 16 ikke-seksere i træk er $(5/6)^{16}$ = 0,054
Sandsynligheden for 17 ikke-seksere i træk er $(5/6)^{17}$ = 0,045
Sandsynligheden for 25 ikke-seksere i træk er $(5/6)^{25}$ = 0,0105
Sandsynligheden for 26 ikke-seksere i træk er $(5/6)^{26}$ =0,0087
Så bruger man 5% er 17 den magiske grænse = den kritiske værdi, og bruger man 1%, er 26 grænsen. Men det er vigtigt at påpege, at man ALDRIG viser noget ved bare 1 forsøg.

Eksempel 1

Observeret værdi		Stik med 1	2	Lis sum
svar	1	**7**	**3**	10
	2	**2**	**8**	10
	sum	9	11	20

Forventet værdi		Stik med 1	2	sum
svar	1	**4,5**	**5,5**	10
	2	**4,5**	**5,5**	10
	sum	9	11	20

De forventede værdier beregnes som f.eks.

$$\frac{\text{Rækkesum*søjlesum}}{\text{total}} = \frac{10*9}{20} = 4{,}5$$

En 2x2 tabel siges at have 1 frihedsgrad (df = degrees of freedom). Det skyldes, at når man vil beholde rækkesummerne 10 og 10 samt søjlesummerne 9 og 11, så kan man kun vælge 1 af tallene i tabellen frit. Havde vi f.eks. fået 6 i stedet for 7 i kategorien 1 stik – svar 1, så ville vi automatisk få 4, 3 og 7 i de andre kategorier. X^2-testen bygger på, at række- og søjlesummer holdes konstant, når vi skal undersøge, om vi er ”langt ude”.

Ti-Interactive:

Insert Matrix:

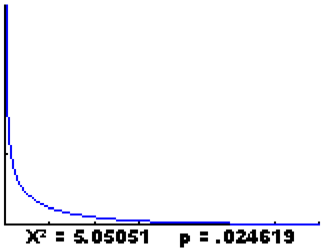
$obs := \begin{bmatrix} 7. & 3. \\ 2. & 8. \end{bmatrix}$

og

Observed Matrix: obs

Test or Interval Type: Chi-square test, Expected Matrix: forvent og ☑ Draw Result

Chi-square test
$p = .024619$
$X^2 = 5.05051$
$df = 1.$



Vil man se den forventede: $forvent = \begin{bmatrix} 4.5 & 5.5 \\ 4.5 & 5.5 \end{bmatrix}$

TI-nspire:

$obs := \begin{bmatrix} 7 & 3 \\ 2 & 8 \end{bmatrix}$

Vælg Statistik, Stat tests, χ^2- 2vejstest:

$\chi^2 2way\ obs: stat.results \quad \begin{bmatrix} \text{"Titel"} & \text{"}\chi^2\text{ 2-vejstest"} \\ \text{"}\chi^2\text{"} & 5.05051 \\ \text{"PVal"} & 0.024619 \\ \text{"df"} & 1. \\ \text{"ExpMatrix"} & \text{"[...]"} \\ \text{"CompMatrix"} & \text{"[...]"} \end{bmatrix}$

Under VAR finder man ExpMatrix:

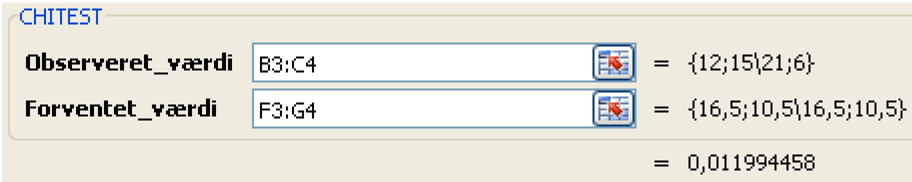
$stat.ExpMatrix \quad \begin{bmatrix} 4.5 & 5.5 \\ 4.5 & 5.5 \end{bmatrix}$

Eksempel 2

	A	B	C	D	E	F	G
1	Observeret	16- års valgret				Forventet	
2	Køn	For	Imod	I alt			
3	Fædre	12	15	27		16,5	10,5
4	Mødre	21	6	27		16,5	10,5
5	I alt	33	21	54			

$\frac{27*33}{54}$ = 16,5, og $\frac{27*21}{5,4}$ =10,5. I regnearket: F3=$D3*B$5/D5. Cellen kan nu trækkes til G3, og de to celler trækkes til F4:G4.

Excel



Eksempel 3 (togenskrydsning)

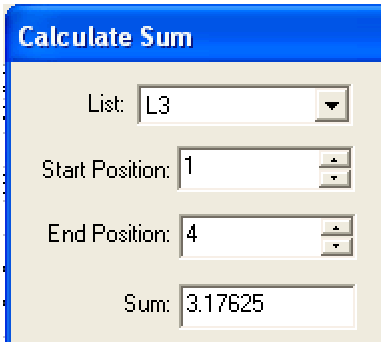
	Høje og grønne	Høje og hvide	Små og grønne	Små og hvide	sum
Obs	57	14	12	4	87
Forventet	48,94	16,31	16,31	5,43	

Her sammenligner vi en række observerede tal med en række teoretiske tal, som er 87*9/16, 87*3/16, 87*3/16 og 87/16. Der er 3 frihedsgrader, da det fjerde tal skal sørge for, at summen bliver 87.

Ti-Interactive:

L1 {...}	L2 {...}	L3 :=(L1-L2)^2/L2
57.	48.9375	I.328304597701149
14.	16.3125	.3278256704980843
12.	16.3125	1.140086206896552
4.	5.4375	0.3800287356321839

og



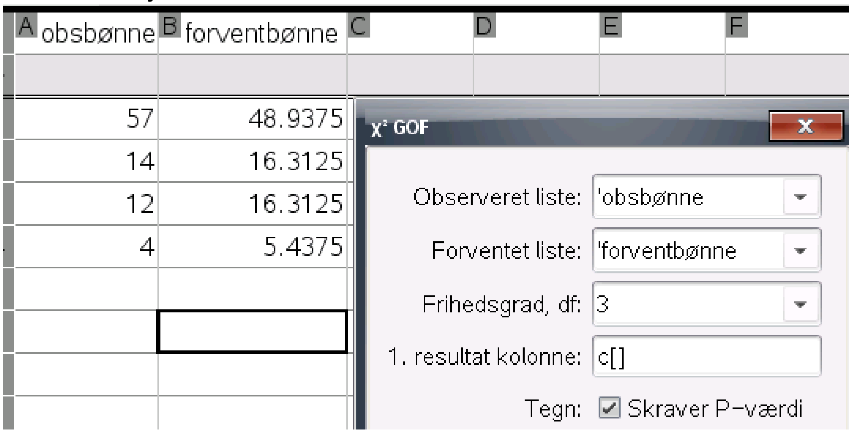
Vælg List, Calculate, Calculate sum. Programmet kan desværre ikkekopiere resultatet, så man må selv skrive det.

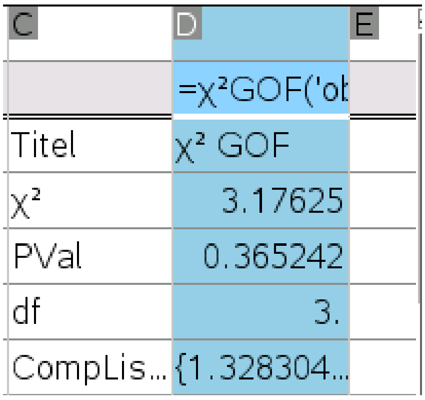
Chi^2 = 3,17625

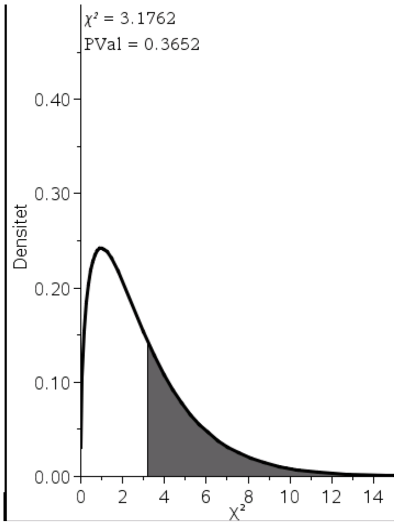
chisquarecdf$(3.17625, \infty, 3)$ = .365241

Ti-nspire:

Her skal arbejdes i lister:







Excel

Brug Indsæt funktion og vælger Chi-test. Marker både de observerede og de forventede, så får man P = 0,365.

Eksempel 7

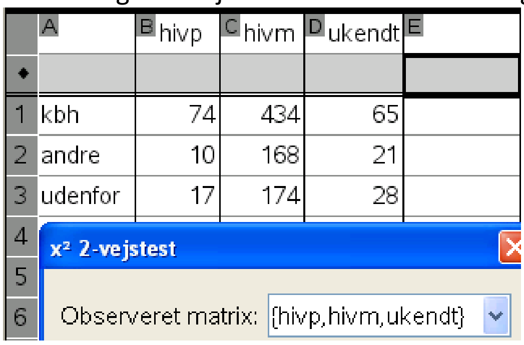
Observerede værdier:	HIV +	HIV -	ukendt	I alt
København	74	434	65	573
Odense, Århus, Ålborg	10	168	21	199
Udenfor store byer	17	174	28	219
I alt	101	776	114	991

Forventede værdier:				
	58	449	66	573
	20	156	23	199
	22	171	25	219
	101	776	114	991

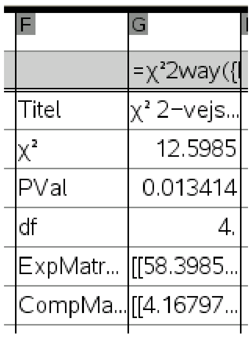
De forventede værdier beregnes således:

$\frac{573*101}{991} = 58$, $\frac{573*776}{991} = 449$, $\frac{219*776}{991} = 171$, $\frac{219*114}{991} = 25$

Ti-nspire: Man kan også arbejde i listerne i stedet for i grafregner:



OK



Eksempel 10

Mænd	Har aldrig drukket alkohol	Har drukket alkohol	
Ingen religiøs baggrund	3 (11)	369 (361)	372
Protestant	15 (26)	861 (850)	876
Muslim	21 (1,6)	34 (53)	55
	39	1264	1303

Vi kan ikke teste denne interessante tabel, fordi den ene forventede værdi (i parentes) er mindre end 5, og vi må kun lave en Chi-i-anden-test, hvis alle de forventede værdier er mindst 5.

Hvorfor må vi ikke lave chi-i-anden-test, når den forventede værdi er mindre end 5?

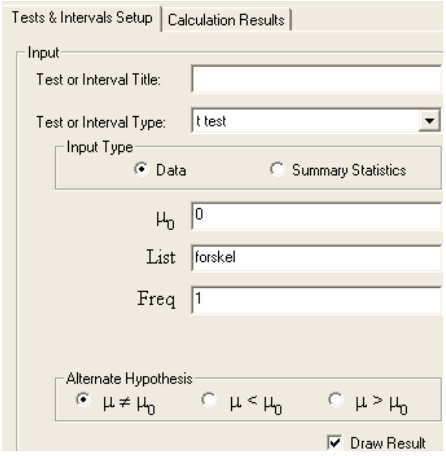
Chi-i-anden teststørrelsen fås ved at beregne (observeret – forventet)^2/forventet for hver eneste celle og så lægge dem sammen. Der er meget stor forskel på at dividere med 1 eller med 2, f.eks $(6-1)^2/1 = 25$ mens $(6-2)^2/2 = 8$. Hvis der derimod er 11 og 12 i nævneren (som er den forventede værid), så er $(16-11)^2/11 = 2{,}27$ mens $(16-12)^2/12 = 1{,}33$, så her er forskellen ikke så stor, selv om forskellen mellem de to nævnere stadig er 1. Ud fra erfaring er man så kommet til, at det giver brugelige resultater, bare de forventede værdier er større end 5.

Eksempel ***15*** (binomial test eller parvis t-test) ***(Er fædre ældre end mødre?)***

Man har fædrenes alder i en søjle og mødrenes i en anden søjle.

Ti-Interactive

Udregn L3 = L1 – L2, dvs man ser kun på aldersforskellen. Så vælger man Stat test and interval tool og t-test, og sætter μ = 0, dvs at man ser, om aldersforskellen er 0.



t test

$\mu \neq \mu_0$

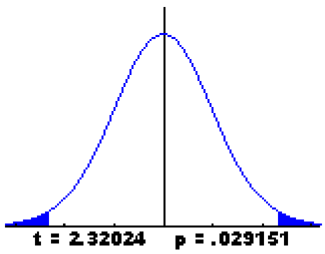
$t = 2.32024$

$p = .029151$

$\bar{x} = 1.64$

$Sx = 3.53412$

$n = 25.$



Her har man farvet den del af fordelingskurven, der har værdier numerisk større end det observerede. Det farvede areal er 3% af arealet under kurven, da p = 0,03.

TI-nspire

Vælg man Lister og regneark og udregner forskel = far – mor. Dvs at man kun ser på forskellen. Vælg t-test og $\mu_0 = 0$ samt $\mu \neq \mu_0$.

A far	B mor	C forskel	D	E
		=far−mor		=tTest(0,'forsk
41	40	1	Titel	t−test
47	45	2	Alternativ...	μ ≠ μ0
47	40	7	t	2.32024
45	42	3	PVal	0.029151
42	51	-9	df	24.
47	43	4	x̄	1.64
45	45	0	sx := sn-...	3.53412
46	44	2	n	25.

(Tabellen fortsætter i søjle far og mor).

Excel:

Her vælger man Data og derpå Dataanalyse og så t-test: Parvis dobbelt stikprøve for middelværdi, og sætter Hypotese for forskel i middelværdi = 0

	Far	*Mor*
Middelværdi	48,6	46,96
Varians	17,41667	21,95667
Observationer	25	25
Pearson-korrelation	0,687365	
Hypotese for forskel i middelværdi	0	
fg	24	
t-stat	2,320239	
P(T<=t) en-halet	0,014575	
t-kritisk en-halet	1,710882	
P(T<=t) to-halet	0,029151	
t-kritisk to-halet	2,063899	

Vi ser, at selv om aldersforskellen på fædre og mødre i gennemsnit kun er 1,6 år (48,6 – 46,96), så forkaster vi hypotesen, fordi vi jo ser på forældrene parvis.

Binomialtest:

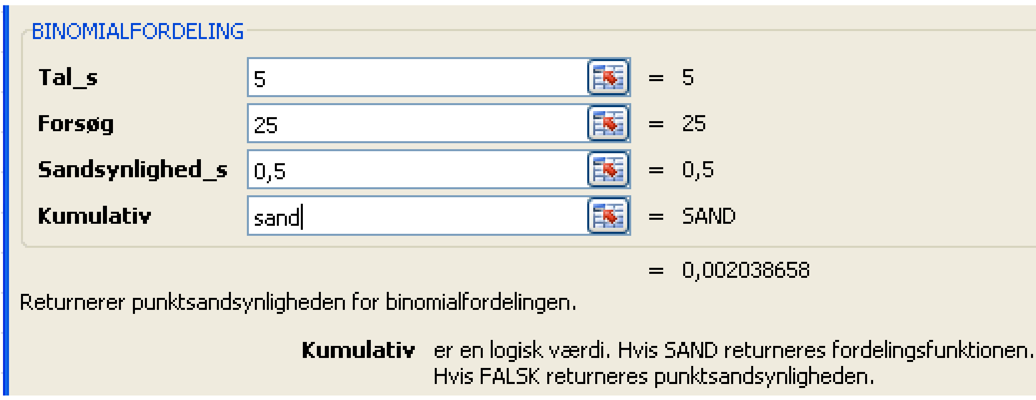
Ti-Interactive

n er antal adspurgte, sandsynlighed er 0,5 (50-50 for hvem der er ældst), og k = 5 er den mindste af de to observerede værdier

TI-nspire

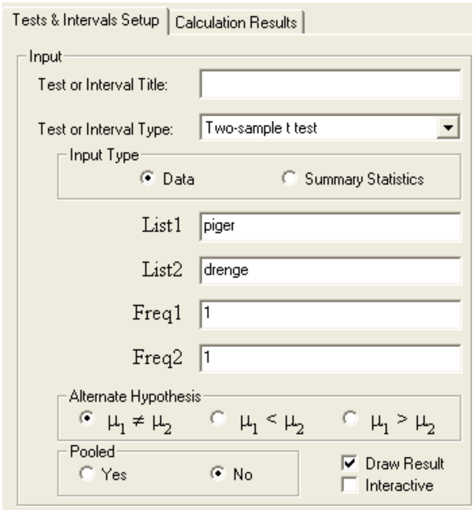
binomcdf(25, 0.5, 5) = .002039

Excel

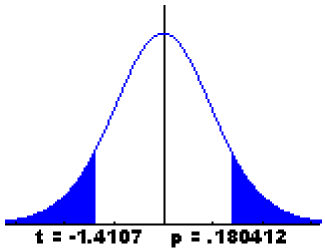


Eksempel 16 (farvekonflikttest)
Her har vi ikke parvis test, så man vælger som følger:

Ti-Interactive:



Two-sample t test
$\mu_1 \neq \mu_2$
$t = -1.41073$
$p = .180412$
$df = 13.8347$
$\bar{x}_1 = 59.8125$
$\bar{x}_2 = 66.75$
$Sx_1 = 11.1965$
$Sx_2 = 11.4362$
$n_1 = 16.$
$n_2 = 8.$



Her har man farvet den del af fordelingskurven, der har værdier numerisk større end det observerede. Det farvede areal er 18% af arealet under kurven, da p = 0,18.

TI-nspire

Vælg t-test med 2-prøver. Udfyldes som Ti-Interactive, men vælg NEJ ud for Samlet (fordi drenge og piger er i hver sin kolonne).

Excel

Vælg t-test: to stikprøver med ens varians. Resten er som parvis t-test.

Eksempel 17 (fravær i mat-fys klasse og samf klasse)
Simulering af fordelinger kan kun lade sig gøre i TI-nspire. Ideen bag simuleringen er, at hvis nulhypotesen er sand, så værdierne (her fraværsprocenterne) er tilfældigt fordelt, så kan man tage de fundne værdier (data) og blande dem og fordele dem tilfældigt på eleverne i klassen. Så er det fundne resultat bare et af mange, og man kan studere, hvordan resultatet ligger i forhold til alle de andre mulige udfald.

TI-nspire

Bjørn Feldsager har beskrevet metoden i sine t-test noter, og man kan se metoden animeret her i del 1 og del 2.
Data skrives i regneark i to søjler: søjle A er fag, så en matfys elev hedder mf, og en samfelev hedder sa. Søjle B er fravær og søjle C laver omrøring i fraværstallene.
Søjle D og E piller det observerede fravær for mat og samf ud i hver sin søjle, mens F og G gør det samme for de omrørte tal. Søjle C,F og G ændrer sig, hver gang man laver en ny omrøring.
Her ses søjlehovederne, hvor søjle C er skrevet ud for sig:

A fag	B fravær	C omrøri...	D mf_fravær	E sa_fravær	F mf_omrørt	G sa_omrørt
		=randsam	=b1:b24	=b25:b48	=c1:c24	=c25:c48

C	omrøring:=randsamp('fravær,dim('fravær),1)

H
obs_dif
-3.22167
sim_dif
0.7675
antal_mål
2032.
antal_sk...
77.
p_værdi
0.037894

Søjle H bruges til nogle enkeltstående udregninger. Først beregnes forskellen i gennemsnitsfravær hos matfys'ere og samf'ere, dels i de observerede tal, dels i den omrørte (simulerede) tal. Det andet tal ændrer sig altså for hver omrøring:

H2	obs_dif:=mean('mf_fravær)−mean('sa_fravær)

H4	sim_dif:=mean('mf_omrørt)−mean('sa_omrørt)

I H2 og H4 skal vi højreklikket og vælge variable -> gem var.
For at gemme en observation fra hver omrøring laves en ny variabel i søjle I:

I test
=capture('sim_dif,1)

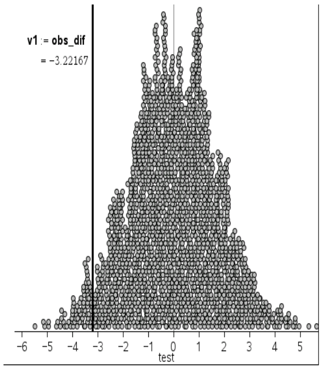
Vi skal også tælle antal omrøringer og antal tilfælde, hvor samf'erne har endnu større fravær i forhold til mat-fys'erne, end det observerede (forskellen bliver så et numerisk større negativt tal):

H6	dim_test:=dim(test)

H8	=countif(test,?≤obs_dif)

Og så udregnes den simulerede P fra de to ovenstående: $H10 \; = \frac{h8}{h6}$

Hold Ctrl + R nede, indtil dim_test har en passende størrelse.



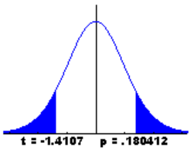
Lav en ny side eller gå under sidelauout og vælg 2-delt side. Klik og vælg data og statistik. Klik på x-aksen og vælg variablen Test.

Under Analyser  sælges Plot værdi. Vælg variablen obs_dif, der er den observerede forskel.

Bemærk, at man så kun tæller til venstre for den afsatte linie (ensidig), hvor man på figuren i eksemplet ovenfor, har farvet halerne i begge side af fordelingen (tosidig

Ensidig og tosidig test

I de numeriske tests, hvor man sammenligner målte værdier, har man hypoteser, der kan formuleres med lighedstegn og alternativer, der formuleres med ”forskellig fra”. Her er der så mulighed for at lave alternativ med formuleringen ”mindre end” eller ”større end”. F.eks. har en lineær regression en faldende tendens en linje med negativ hældning - altså hældning mindre end 0.



Mange fordelinger er symmetriske, som vi ser på figuren lånt fra eksempel 14 og som vi også kan se på den simulerede fordeling i eksempel 15. Så der er lige så mange procent, der er for langt ude til højre, som der er til venstre. Det vil sige, at hvis vi skifter fra en tosidig test til en ensidig test, så falder P-værdien til det halve. Det betyder, at hvis man har P < 0,05 i en tosidig test, så gælder det også for en ensidig test, og man forkaster hypotesen. Hvis P > 0,10, så er den ensidige p større end 0,05, og man vil ikke forkaste. Kun hvis P-værdien for den tosidige test ligger mellem 0,05 og 0,10, vil P for den ensidige test ligge under 0,05. Så der får man forskellig konklusion, og hvis man der har valgt en ensidig test, så vil man uvægerlig blive kritiseret for sit valg.

Situationen opstod i eksempel 17, hvor det eneste argument for ensidig test er, at ” der er da ingen der regner med, at mat-fys’erne pjækker mere end samf’erne”. Derfor laver man en nulhypotese, hvor for meget mat-fys fravær kommer sammen med lige meget fravær, og det tester man mod samf’ernes fravær i den alternative hypotese. Selvfølgelig bliver samf’erne vrede over det!

Man skal beslutte sig for, hvilken side man vil teste til, inden man overhovedet laver testen. Ved simuleringen kommer man nok let til at teste ensidigt i den side, hvor man nu lige er endt.

Eksempel 9 vil man også kunne teste ensidigt, hvis man ikke kunne drømme om, at kvinder kunne tænkes at have længst ferie.

Eksempel 18 (drengebørn)

Excel

Regressionsstatistik	
Multipel R	0,404425
R-kvadreret	0,163559
Justeret R-kvadreret	0,147474
Standardfejl	0,175414
Observationer	54

ANAVA

	fg	*SK*	*MK*	*F*	*Signifikans F*	
Regression	1	0,312876	0,312876	10,16819	0,002420619	= P
Residual	52	1,600044	0,03077			
I alt	53	1,91292				

	Koefficienter	*Standardfejl*	*t-stat*	*P-værdi*	*Nedre 95%*	*Øvre 95%*
Skæring	60,80124	2,950142	20,6096	1,11E-26	54,88135561	66,72113
X-variabel 1	-0,004756	0,001491	-3,18876	0,002421	-0,007748211	-0,00176

Se på de farvede tal: 0,016 kender vi som R^2 fra regressionen. Vores hypotese siger, at hældningen a = 0, og med en signifikans på P = 0,0024 forkaster vi denne hypotese. Vi ser også, at skæring med y-aksen = b = 70,4 og hældningen = a = -5,63 , og a ligger med 95% sikkerhed mellem -0,0077 og -0,00176.

Eksempel 20 (Spritdomme)

Ti-Interactive TI-nspire	Vælg Lineær reg t-test. Vælg x og y søjler.

Excel: Data, dataanalyse, regression. Vælg y- og x-værdier (i den rækkefølge)

RESUMEOUTPUT Mænd 15-19 1990 - 1998

Regressionsstatistik	
Multipel R	0,705902
R-kvadreret	0,498298
Justeret R-kvadreret	0,426627
Standardfejl	41,813
Observationer	9

ANAVA

	fg	*SK*	*MK*	*F*	*Signifikans F*
Regression	1	12155,27	12155,2	6,9525	0,033581
Residual	7	12238,29	1748,327		
I alt	8	24393,56			

	Koefficienter	*Standard-fejl*	*t-stat*	*P-værdi*	*Nedre 95%*	*Øvre 95%*
Skæring	421,7111	25,69979	16,4091	7,61E-07	360,94077	482,481
X-variabel 1	-14,2333	5,398035	-2,63676	0,03358	-26,99765	-1,4690

Se på de farvede tal: 0,49 kender vi som R^2 fra regressionen. Vores hypotese siger, at hældningen a = 0, og med en signifikans på P = 0,0336 forkaster vi denne hypotese. Vi ser også, at skæring med y-aksen = b = 422 og hældningen = a = -14,2 , og a ligger med 95% sikkerhed mellem -27 og -1,45.

Eksempel 21 (Medlemmer til det konservative folkeparti)

TI-nspire

Titel	Lineær R...
Alternativ...	β & ρ ≠ 0...
RegEqn	a+b*x
t	-11.6048
PVal	0.000315
df	4.
a	19187.
b	-1037.69
s	374.066
SESlope	89.4188
r²	0.971155

Vælg Lineær reg t-test. Vælg x og y søjler.

Excel bruger F-test og TI-nspire bruger t-test. Forskellen er bare, at $F = t^2$, og da det indgår i de videre beregninger, ender man med, at P-værdierne er præcis ens.

$t^2 = (-11{,}6048)^2 = 134{,}6$

P = 0,000315

Excel

Data, dataanalyse, regression. Vælg y- og x-værdier (i den rækkefølge)

RESUMEOUTPUT

Regressionsstatistik	
Multipel R	0,985472
R-kvadreret	0,971155
Justeret R-kvadreret	0,963943
Standardfejl	374,0659
Observationer	6

ANAVA

	fg	*SK*	*MK*	*F*	*Signifikans F*
Regression	1	18843854	18843854	134,6708	0,000315
Residual	4	559701,1	139925,3		
I alt	5	19403555			

Eksempel 25

Standardfejl eller spredningen på gennemsnittet:
Når vi har en stikprøve med *n* observationer, kan vi beregne gennemsnit, varians og spredning, som er kvadratroden af variansen. Gennemsnit fra forskellige stikprøver ligger tættere på hinanden end de oprindelige observationer, og man kan vise, at spredningen på gennemsnittene er stikprøvens spredning divideret med kvadratrod n, altså $\frac{s}{\sqrt{n}}$. Gennemsnit er tilnærmelsesvis normalfordelt, så man ved at ca. 95% af alle gennemsnit skal ligge indenfor gennemsnit ± 2 standardfejl. Så hvis man skal sammenligne to gennemsnit, kan man stå i den ene værdi og se, om den anden ligger indenfor 2 standardfejls afstand. Eller man kan se, om man mødes, hvis man starter i hvert sit gennemsnit og går en standardfejl mod hinanden. Det forudsætter selvfølgelig, at de to standardfejl er nogenlunde lige store. Hvis man ser lidt stort på nøjagtighederne her, så kan man lave en øjemålsregel, der kan bruges, når man læser en artikel eller hører et foredrag med Powerpoint. Her kan man gå ud fra, at forfatteren eller foredragsholderen har styr på den nøjagtige statistik, og man skal som læser eller tilskuer kun danne sig et overblik over resultaterne. Se eksempel 20 – 25.
Øjemålsregel: Hvis man skal sammenligne to værdier og har hypotesen, at gennemsnittene er ens, så sammenligner man bare de lodrette streger, der viser standardfejlene. Hvis stregerne når hinanden eller overlapper, så forkaster man IKKE hypotesen, og hvis streger IKKE når hinanden, så forkaster man hypotesen, og der er så forskel på gennemsnittene.