Stockholms universitet

Statistisk analys av kohort- och fall-kontrollstudier vid utredning av livsmedels-burna utbrott

A. Bjärenson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2016:24
Matematisk statistik
Augusti 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2016:24**
http://www.math.su.se/matstat

Statistisk analys av kohort- och fall-kontrollstudier vid utredning av livsmedelsburna utbrott

A. Bjärenson*

Augusti 2016

Sammanfattning

Kohort- och fall-kontrollstudier används vid utredningar där man försöker fastställa källan till livsmedelsburna sjukdomar. I Sverige upprättas dessa typ av studier vanligtvis av myndigheter såsom folkhälsomyndigheten. Datan får studierna samlas in på olika sätt vilket gör att den statistiska analysen kan skilja sig åt mellan dem. Både kohort- och fall-kontrollstudier kan innehålla ett stort antal livsmedel som ska undersökas. Det leder till multipla jämförelser och därmed behovet av att justera för det. Syftet med uppsatsen är att studera den statistiska teorin bakom analysen av kohort- och fall-kontrollstudier, samt att utforma en funktion, i programspråket R, som ska underlätta arbete och statistisk analys av studierna när man har ett stort antal livsmedel som ska undersökas. Syftet med uppsatsen är även att utföra en statistisk analys av en fall-kontrollstudie över ett utbrott av salmonella typhimurium i Danmark 1996. I analysen används enkel betingad logistisk regression för samtliga variabler som undersöks. Resultatet visade att två livsmedel, frukt och variabeln "Plant7", var signifikanta på 5% nivå. Efter justering för multipla jämförelser var endast variabeln "Plant7" signifikant.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: a.bjarenson@gmail.com. Handledare: Michael Höhle.

Abstract

Cohort studies and case-control studies are common in investigations that attempt to determine the source of foodborne diseases. The studies differ in the way the data is collected, leading to a bit different statistical analysis for the two of them. Both studies may contain a large number of food items which leads to multiple comparison and the need to adjust for it. The aim of this thesis is to study the statistical theory behind the analysis of cohort studies and case-control studies, and to implement a function in R that performs the batch analysis for a large number of food items. The aim of this thesis also includes performing an analysis of a salmonella outbreak in Denmark 1996. The analysis was conducted through fitting a simple conditional logistic regression model to each food item. The results showed that two food items was signifikant at 5% level, fruit and variable "Plant7". After adjusting for multiple comparison only "Plant7" remained significant.

Förord

Denna uppsats utgör ett examensarbete i matematisk statistik om 15 hp. Jag vill rikta ett tack till min handledare Michael Höhle, universitetslektor i matematisk statistik vid Stockholms universitet, för värdefulla råd och vägledning.

Innehåll

1 Introduktion

Livsmedelsburna sjukdomar är en sammanfattande benämning på en infektion eller förgiftning som uppstår till följd av förtäring av livsmedel innehållande skadliga mikroorganismer. En infektion orsakas vanligtvis när själva microorganimen, via födan, tränger sig in i mag- och tarmkanalen. Exempel på infektioner orsakade av bakterier är salmonella, E.Coli (EHEC) och Campylobacter. En förgiftning uppstår till följd av intag av toxiner som microorganismer, vanligtvis bakterier, bildat vid tillväxt. Dessa toxiner orsakar förgiftning även om själva bakterien inte längre är vid liv när födan intas. Några toxinbildande bakterier som orsakar förgiftning är Clostridium perfringens, Staphylococcus aureus och Bacillus cereus.

Enligt världshälsoorganisationen (WHO) definieras ett livsmedelsburet utbrott som två eller fler sjukdomsfall till följd av intag av samma livsmedel [6]. För att stoppa smittspridningen så snabbt så möjligt upprättas särskilda utredningar för att kunna identifiera smittkällan. I utredningar som bedrivs för att försöka fastställa källan, insamlas data över smittade individers matintag av specifika livsmedel. En fall-kontrollstudie eller kohortstudie kan sedan upprättas där svaren jämförs med data över friska individers intag av samma livsmedel. Datainsamling kan exempelvis ske genom intervju, enkäter eller genom webbaserade formulär som samlas in elektroniskt. Vilken av de ovannämnda studierna som väljs vid utredning beror bl.a. på datan. En statistisk analys kan exempelvis bestå av att studera observationerna med hjälp av 2×2 tabeller eller med en mer avancerad metod som exempelvis multipel logistisk regressions modell.

Vanligtvis upprättas dessa typ av studier för att utredare vill pröva hypotesen om att ett livsmedelsburet utbrott orsakas av någon specifik källa. När studiens fall och kontroller är insamlade kan utredarna dock väja att inkludera många fler misstänkta källor. Detta görs eftersom det alltid är svårt att på förhand veta vad som faktiskt har orsakat smittan. Det leder till att studien kan innehålla ett väldigt stort antal exponeringsvariabler. En sådan situation benämns ibland med engelskans “fishing expedition”.

1.1 Salmonella typhimurium

Danmark, Fyn county, drabbades under hösten 1996 av ett utbrott av salmonella typhimurium. Bakterien är en av de cirka 2300 serotyper som ingår i bakteriegruppen salmonella. Den är en av de vanligast serotyperna som hittas i livsmedel och orsakar en allvarlig maginfektion hos människor och även djur [7]. Vanliga symptom är feber, magkramper och kräkningar som uppträder cirka ett halvt till tre dygn efter smittotillfället.

För att försöka fastställa källan till smitta upprättade det danska zoonoscentret en fall-kontrollstudie. Varje fall matchades med två kontroller med avseende på kön, residens och ålder. Deltagarna telefonintervjuades om sina matintag under de senaste två veckorna. Av intresse var huruvida de hade ätit vissa livsmedel, bl.a. kött som genom återförsäljare kunde kopplas till produktionsanläggning "plant7". De tillfrågades även om de hade befunnit sig utomlands under samma tidsperiod. Datan för studien kan hämtas från R-paketet Epi. Den innehåller totalt 15 variabler med 136 observationer vardera. Bland dessa 136 observationer finns saknade värden, för vissa av exponeringsvariablerna, som har markerats med "NA". Datan innehåller även 5 set där fallet endast har matchats med en kontroll. Samtliga variabler i datasetet anges i nedanstående tabell.

Variabel	Tolkning	
case	sjukdomsstatus, fall eller kontroll	fall=1, kontroll=0
id	identifikationsnummer för varje deltagare	
set	index som anknyter varje deltagare till ett visst set som har matchats med avseende på ålder och kön	
age	ålder	
sex	kön	
abroad	om deltagare har befunnit sig utomlands	1=ja, 0=nej
beef	om deltagare ätit biff	1=ja, 0=nej
pork	om deltagare ätit fläsk	1=ja, 0=nej
veal	om deltagare ätit kalv	1=ja, 0=nej
poultry	om deltagare ätit fågel	1=ja, 0=nej
liverp	om deltagare ätit leverpastej	1=ja, 0=nej
veg	om deltagare ätit grönsaker	1=ja, 0=nej
fruit	om deltagare ätit frukt	1=ja, 0=nej
egg	om deltagare ätit ägg	1=ja, 0=nej
plant7	om deltagare ätit kött från plant7	1=ja, 0=nej

Vi lägger märke till att beskrivningen av variabeln "Plant7" inte klargör vad som menas med "kött". Det finns ingen riktig specifikation om vad "kött" syftar till. Vi skulle kunna göra tolkningen att det rör sig om ett annat livsmedel i köttväg som inte finns med i datan.

Det kan även tolkas som att variabeln syftar till några av de livsmedel vi har i datasetet. I köttväg har vi livsmedlen biff, fläsk, och kalv samt att även leverpastej innehåller kött. Även fågel, som är vitt kött, benämns ofta som kött. Men i sådant fall saknar vi dock information om det är samtliga av dessa livsmedel som "kött" står för, eller om det exempelvis endast är två av dem. Vi tittar närmare på datan. Om vi utgår från datasetet så skulle "kött" kunna syfta till antingen fläsk eller leverpastej. Det finns ett fall som har exponerats för både fläsk och "Plant7" men inte för någon annan exponeringsvariabler i köttväg. Det samma gäller leverpastej. Resterande rader i datasetet ger tyvärr ingen ytterligare information.

Då "Plant7" endast har angivits i datan tillsammans med de resterande exponeringsvariablerna, men det saknas information om hur den ska tolkas, väljer vi att avstå från att göra antagandet om att den syftar till leverpastej och fläsk. Vi analyserar den istället som en av de andra exponeringsvariablerna och låter definitionen av kött vara osagd.

Innan vi börjar vår statistiska analys, bekantar vi oss med den statistiska teorin som står bakom analysen för kohort- och fall-kontrollstudier.

2 Kohortstudie

I en kohortstudie studerar man en grupp individer, *kohorten*, för att undersöka sambandet mellan exponering, av en eller fler faktorer, och en sjukdomsstatus. Studiedesignen är *prospektiv*, man fastställer först om en individ i kohorten är exponerad eller inte och får därefter information om individens sjukdomsstatus.

Antag att vi inom kohorten delar in individer i två grupper. En grupp med exponerade individer E, och en grupp med ej exponerade individer $\overline{E}$. Vi fastställer sedan huruvida varje individ, i respektive grupp, är sjuk eller inte och inför beteckningarna D respektive $\overline{D}$. Vi sammanställer våra observationer i nedanstående 2×2 tabell

	E	$\overline{E}$	
D	n_{11}	n_{12}	n_{1+}
$\overline{D}$	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n

där n_{11} betecknar antalet exponerade individer som har utvecklat sjukdomen och n_{12}, n_{21} och n_{22} tolkas på motsvarande sätt. Vi har att $n_{+j} = \sum_{i=1}^{2} n_{ij}$ och $n_{i+} = \sum_{j=1}^{2} n_{ij}$ betecknar radsumman respektive kolumnsumman samt att n betecknar totala antalet individer i kohorten.
Vi är intresserade av de betingade sannolikheterna för sjukdom givet exponeringsstatus, dvs $\pi_1 = P(D|E)$ och $\pi_2 = P(D|\overline{E})$. Vi betraktar totala antalet exponerade individer n_{+1}, och ej exponerade individer n_{+2}, som fixerade i studien. Frekvenserna n_{11} och n_{12} är oberoende med fördelning $n_{11} \sim \text{bin}(n_{+1}, \pi_1)$ och $n_{12} \sim \text{bin}(n_{+2}, \pi_2)$. Sannolikheterna skattas som $\hat{\pi}_1 = n_{11}/n_{+1}$ och $\hat{\pi}_2 = n_{12}/n_{+2}$. Vi har enligt [1] att $\hat{\pi}_1$ och $\hat{\pi}_2$ är maximum likelihood-skattningar av π_1 och π_2 och att

$$\hat{\pi}_j \overset{d}{\approx} N\left(\pi_j, \pi_j(1-\pi_j)/n_{+j}\right).$$

2.1 Oddskvot

Nedanstående tabell sammanställer de betingade sannolikheterna för sjukdom givet exponeringsstatus i en 2×2 enligt

	E	$\overline{E}$	
D	π_1	π_2	
$\overline{D}$	$1-\pi_1$	$1-\pi_2$	
	1.00	1.00	

(1)

Oddset för grupp j definieras som $\psi_j = \pi_j/(1-\pi_j)$ där $j = 1, 2$. Oddset antar icke-negativta värden och tolkas som hur trolig sjukdomsstatus sjuk är i relation till sjukdomsstatus frisk i gruppen. Ett oddsvärde på $1/4$ tolkas som att 1 av 5 individer i gruppen kommer att vara drabbade av sjukdomen och att vara frisk är då 4 gånger så troligt som att vara sjuk. *Oddskvoten* definieras som

$$\text{OR} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} \qquad 0 \leq \text{OR} \leq \infty$$

och skattas som

$$\widehat{\text{OR}} = \frac{\hat{\pi}_1(1-\hat{\pi}_2)}{\hat{\pi}_2(1-\hat{\pi}_1)} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Oddskvoten är ett relativt mått på association mellan sjukdomsstatus och grupptillhörighet. När $OR = 1$ är oddset för sjukdomsstatus sjuk lika stor i gruppen med de exponerade individerna som i gruppen med de icke exponerade individerna. Vi har då en lika hög sannolikhet för sjukdom bland grupperna, $\pi_1 = \pi_2$, och alltså är är sjukdomsstatus oberoende av exponeringsstaus. För $0 \leq \text{OR} < 1$ är oddset för sjukdomsstatus sjuk mindre i gruppen med de exponerade individerna än i gruppen med de icke exponerade individerna och vi har att $\pi_1 \leq \pi_2$. När $\text{OR} > 1$ är oddset för sjukdomsstatus sjuk större i gruppen med de exponerade individerna än i den andra gruppen och vi har att $\pi_1 > \pi_2$.
Log oddskvoten definieras som

$$\log(\text{OR}) = \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right) = \log\left(\frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}\right)$$

där log syftar till den naturliga logaritmen och antar värden på intervallet $(-\infty, \infty)$. Vi ser att $\log(\text{OR}) = 0$ motsvarar $\text{OR} = 1$ för vilket $\pi_1 = \pi_2$. Det kan vara fördelaktigt att studera log(OR) istället för OR. Eftersom oddskvoten inte kan anta negativa värden så är den nedåt begränsad, men inte uppåt, dvs. den antar värden på intervallet $(0, \infty)$. Dens fördelning är skev. Log oddskvoten antar värden på hela reella linjen och är symmetrisk runt $\log(\text{OR}) = 0$. Eftersom $\hat{\pi}_j \overset{d}{\approx} N\left(\pi_j, \pi_j(1-\pi_j)/n_{+j}\right.$ så kan man, genom att använda deltametoden och Slutskys teorem, visa att log(OR) är approximativt asymptotiskt normalfördelad och beräkna väntevärde och varians [5]. Detta kommer till användning vid inferens, exepelvis vid konstruktion av konfidensintervall.

2.2 Konfidensintervall

Vi kan konstruera ett tvåsidigt konfidensintervall för oddskvoten genom att först bilda ett konfidensintervall för log oddskvoten. Genom att använda del-

tametoden så kan vi approximera variansen av log(OR). Den asymptotiska variansen för log oddskvoten ges av

$$\mathrm{Var}(\log \widehat{\mathrm{OR}}) = \mathrm{Var}\left(\log(\psi_1) - \log(\psi_2)\right) = \mathrm{Var}\left(\log(\psi_1)\right) + \mathrm{Var}\left(\log(\psi_2)\right) \quad (2)$$

där $\mathrm{Cov}\left(\log(\psi_1), \log(\psi_2)\right) = 0$ eftersom log oddsen, $\log(\psi_1)$ och $\log(\psi_2)$, är oberoende, vilket följer av att n_{11} och n_{12} är oberoende. Vi använder deltametoden och erhåller

$$\begin{aligned}
\widehat{\mathrm{Var}}(\log \widehat{\mathrm{OR}}) &= \left(\frac{\partial}{\partial \hat{\pi}_1} \log(\hat{\psi}_1)\right)^2 \mathrm{Var}(\hat{\pi}_1) + \left(\frac{\partial}{\partial \hat{\pi}_2} \log(\hat{\psi}_2)\right)^2 \mathrm{Var}(\hat{\pi}_2) \\
&= \left(\frac{1}{\hat{\pi}_1(1-\hat{\pi}_1)}\right)^2 \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_{+1}} + \left(\frac{1}{\hat{\pi}_1(1-\hat{\pi}_2)}\right)^2 \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_{+2}} \\
&= \frac{1}{n_{+1}\hat{\pi}_1(1-\hat{\pi}_1)} + \frac{1}{n_{+2}\hat{\pi}_2(1-\hat{\pi}_2)} = \frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}
\end{aligned}$$

Ett tvåsidigt asymptotiskt konfidensintervall för log oddsvoten ges därmed av

$$\mathrm{CI}_{\log(\mathrm{OR})} = \log \widehat{\mathrm{OR}} \pm z_{1-\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}} \quad (3)$$

på intervallet $(-\infty, \infty)$. Genom att ta exponentialfunktionen av (3) erhåller vi ett asymptotiskt konfidensintervall för oddskvoten

$$\mathrm{CI}_{\mathrm{OR}} = \exp\left\{\log \widehat{\mathrm{OR}} \pm z_{1-\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}}\right\} = (e^a, e^b) \quad (4)$$

på intervallet $[0, \infty)$.

2.3 Oberoendetest

Vi är intresserade av att testa om sjukdomsstatus är statistiskt oberoende av huruvida individen är exponerad eller inte. Vi låter π_{ij} beteckna sannolikheten att hamna i cellen belägen i rad i och kolumn j i tabell (1). Vi har att $\pi_{i+} = \pi_{i1} + \pi_{i2}$ och $\pi_{+j} = \pi_{1j} + \pi_{2j}$. Vi sätter upp nollhypotesen $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ som vi vill testa mot den tvåsidiga alternativ hypotesen $H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}$. Enligt nollhypotesen råder ett oberoende mellan sjukdomsstatus och exponering, medan H_1 säger att sjukdomsstatus beror på huruvida individen är exponerad eller inte. Två vanliga test som används för att testa oberoende är *Pearson chi-två test* och *likelihood-kvot test* [1]. Pearson chi-två statistika ges av

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

och likelihood-kvot statistikan av

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right).$$

Under nollhypotesen är *skattningen av de förväntade frekvenserna*

$$\begin{aligned}\hat{\mu}_{ij} &= n\hat{\pi}_{i+}\hat{\pi}_{+j} \\ &= n\frac{n_{i+}}{n}\frac{n_{+j}}{n} \\ &= \frac{n_{i+}n_{+j}}{n}.\end{aligned}$$

Under H_0 är både likelihood-kvot och Pearson statiskan symptotiskt χ^2-fördelad med 1 frihetsgrad [5]. När X^2 respektive G^2 är större än eller lika med $\chi^2_{1-\alpha}(1)$ förkastas H_0 och vi kan statistiskt säkerställa att sjukdomsstatus beror på status för exponering.

Vi går nu vidare med att att studera den statistiska teorin för en fallkontrollstudie. I kommande sektion använder vi oss av samma beteckningar som [5].

3 Fall-kontrollstudie

3.1 Fall-kontrollstudie omatchad

Fall-kontrollstudien undersöker sambandet mellan exponering, av en eller fler faktorer, och sjukdomsstatus. Vanligtvis så studeras huruvida en deltagare i studien är sjuk eller inte som responsvariabel, och exponeringsstatus som en förklarande variabel. Antag att vi har en grupp där samtliga individer har en viss sjukdom, dvs en grupp bestående av fall D och en grupp bestående av ett oberoende urvall av kontroller som inte har sjukdomen $\overline{D}$. Därefter fastställer vi huruvida varje individ, i respektive grupp, har exponerats för en faktor av intresse E eller inte $\overline{E}$. Vi kan sammanställa detta i följande 2×2 frekvenstabell

	D	$\overline{D}$	
E	n_{11}	n_{12}	n_{1+}
$\overline{E}$	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	N

där n_{11} betecknar antalet fall som har exponeringsstatus E, n_{12} antalet kontroller som har exponeringsstatus E och där n_{21} och n_{22} tolkas på motsvarande sätt. Vi har att N är totala antalet individer i studien och att $n_{+j} = \sum_{i=1}^{2} n_{ij}$ och $n_{i+} = \sum_{j=1}^{2} n_{ij}$ betecknar radsumman respektive kolumnsumman.

Datan är insamlad med en *retrospektiv metod*. Vi utgår från huruvida en deltagare är sjuk eller inte och fastställer därefter om denne är exponerad. Rad- och kolumnsumman n_{+1} respektive n_{+2} betraktas som fixerade i studien. Vi har att frekvenserna n_{11} och n_{12} är oberoende med fördelning $n_{11} \sim \mathrm{Bin}(n_{+1}, \phi_1)$ och $n_{12} \sim \mathrm{Bin}(n_{+2}, \phi_2)$. Vi kan skatta de betingade sannolikheterna för exponering givet sjukdomsstatus

$$P(E|D) = \phi_1, \quad P(E|\bar{D}) = \phi_2,$$

som $\hat{\phi}_1 = n_{11}/n_{+1}$ och $\hat{\phi}_2 = n_{12}/n_{+2}$. Vi sammanställer följande tabell.

	D	$\overline{D}$	
E	ϕ_1	ϕ_2	
$\overline{E}$	$1-\phi_1$	$1-\phi_2$	
	1.00	1.00	

Med hjälp av de betingade sannolikheterna kan vi beräkna den så kallade *retrospektiva oddskvoten*

$$\mathrm{OR}_{Reto} = \frac{P(E|D)/P(\overline{E}|D)}{P(E|\overline{D})/P(\overline{E}|\overline{D})} = \frac{\phi_1/(1-\phi_1)}{\phi_2/(1-\phi_2)}.$$

Eftersom vi vill studera sjukdomsstatus givet exponeringsstatus så är vi mer intresserade av den omvända betingningen, dvs sannolikheten för sjukdom givet exponering. Om vi antar att prevalensen för sjukdom i populationen som fallen och kontrollerna kommer från är känd, $P(D) = \delta$, så kan vi med hjälp av bayes sats beräkna de betingade sannolikheterna [5]. Vi får

$$\pi_1 = P(D|E) = \frac{P(D,E)}{P(E)} = \frac{P(E|D)P(D)}{P(E|D)P(D) + P(E|\overline{D})P(\overline{D})}$$

$$= \frac{\phi_1\delta}{\phi_1\delta + \phi_2(1-\delta)}$$

och

$$\pi_2 = P(D|\overline{E}) = \frac{P(D,\overline{E})}{P(\overline{E})} = \frac{P(\overline{E}|D)P(D)}{P(\overline{E}|D)P(D) + P(\overline{E}|\overline{D})P(\overline{D})}$$

$$= \frac{(1-\phi_1)\delta}{(1-\phi_1)\delta + (1-\phi_2)(1-\delta)}.$$

Vi kan nu beräkna den *prospektiva oddskvoten*

$$\mathrm{OR} = \frac{P(D|E)/P(\overline{D}|E)}{P(D|\overline{E})/P(\overline{D}|\overline{E})} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

Vi ser att den prospektiva oddskvoten sammanfaller med den retrospektiva oddskvoten

$$\text{OR} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{P(D|E)/P(\overline{D}|E)}{P(D|\overline{E})/P(\overline{D}|\overline{E})} = \frac{P(E|D)/P(\overline{E}|D)}{P(E|\overline{D})/P(\overline{E}|\overline{D})} = \frac{\phi_1/(1-\phi_1)}{\phi_2/(1-\phi_2)} \tag{5}$$

Vi vet i praktiken sällan värdet på δ men som vi ser så kancelleras den i ekvationerna ovan.

Vi kan konstatera att den retrospektiva oddskvoten sammanfaller med den prospektiva oddskvoten. För statistisk inferens av studien, när det gäller konfidensintervall och ett oberoendetest, så utförs dessa precis på samma sätt som för en kohortstudie. Vi får identiska konfidensintervall som de vi sammanställde i kohortstudien enligt (3) och (4). Även när det kommer till oberoendetest så kan vi både använda Pearson chi-två test och Likelihood-kvot test och de utförs på samma sätt som för kohortstudien.

3.2 Matchning

I den ovannämnda fall-kontrollstudien valdes kontrollgruppen genom ett oberoende urval. Det är dock vanligt att man önskar att öka jämförbarheten av deltagare mellan grupperna med avseende på en eller fler faktorer. Det kan exempelvis vara ålder, kön, sjukdomshistorik, med fler. Det görs för att undvika att andra faktorer, än de som studeras som exponeringar, påverkar studiens resultat. Ett sätt att åstadkomma detta är att matcha. Man väljer, till varje deltagare i fallgruppen, en eller fler kontroller som är matchade till denne med avseende på faktorerna.

3.3 Fall-kontrollstudie parmatchning

Antag att vi har en grupp, av storlek N, bestående av fall D. Till varje fall väljer vi ut en kontroll $\overline{D}$ som är matchad till fallet med avseende på en eller fler faktorer. Vi får därmed N stycken matchade par där en parmedlem är ett fall D och den andre är en kontroll $\overline{D}$. Observationerna inom varje par är korrelerade. Vi kan därför inte studera grupperna, D och $\overline{D}$, som oberoende. Vi betraktar istället varje par som en enhet och sammanställer följande

frekvenstabell och sannolikhetstabell.

		$\overline{D}$		
		E	$\overline{E}$	
D	E	a	b	$a+b$
	$\overline{E}$	c	d	$c+d$
		$a+c$	$b+d$	N

		$\overline{D}$		
		E	$\overline{E}$	
D	E	ϕ_{11}	ϕ_{12}	ϕ_{1+}
	$\overline{E}$	ϕ_{21}	ϕ_{22}	ϕ_{2+}
		ϕ_{+1}	ϕ_{+2}	1

Frekvensen a tolkas som antalet par där både fallet och kontrollen har exponeringsstatus E. Talet b betecknar antalet par där fallet har exponeringsstatus E och kontrollen $\overline{E}$. Frekvenserna c och d tolkas på motsvarande sätt. Vidare har vi ϕ_{11} och ϕ_{12} som är sannolikheten att både fallet och kontrollen har exponeringsstatus E, respektive sannolikheten att fallet har exponeringsstatus E och kontrollen $\overline{E}$. Resternade sannolikheter, ϕ_{21} och ϕ_{22}, tolkas på motsvarande sätt.

Par som ingår i a och d kallas *konkordanta par* för att både fallet och kontrollen, inom varje par, har samma exponeringsstatus. Par som ingår i b och c kallas *diskordanta par* för att fallet och kontrollen, i respektive par, har olika exponeringsstatus. Man kan anta att frekvenserna a, b, c och d är multinomialfördelade $(a,b,c,d)' \sim M_4(N,(\phi_{11},\phi_{12},\phi_{21},\phi_{22})')$ med sannolikhetsfunktionen

$$P(a,b,c,d|N) = \frac{N!}{a!b!c!d!}\phi_{11}^a\phi_{12}^b\phi_{21}^c\phi_{22}^d.$$

Då vi har parmatchad data så är inte oddskvoten, som beräknas med ϕ_{+j} och ϕ_{i+}, ekvivalent med oddskvoten för populationen. Vi väljer istället att studera sannolikheterna för ett fixt värde på den matchningsvariabeln Z. Då vi betingar på ett visst värde på Z, kan vi anta att dragningen av parmedlemmar, från respektive population, är oberoende. Medlemmarna inom respektive par är alltså sinsemellan betingat oberoende. Vi får att

$$\phi_{12|z} = \phi_{1+|z}\phi_{+2|z} = P(E|D,z)P(\overline{E}|\overline{D},z)$$

$$\phi_{21|z} = \phi_{2+|z}\phi_{+1|z} = P(\overline{E}|D,z)P(E|\overline{D},z)$$

och vi kan nu studera den *betingade retrospektiva oddskvoten*

$$\mathrm{OR}_z^* = \frac{P(E|D,z)/P(\overline{E}|D,z)}{P(E|\overline{D},z)/P(\overline{E}|\overline{D},z)} = \frac{P(E|D,z)P(\overline{E}|\overline{D},z)}{P(\overline{E}|D,z)P(E|\overline{D},z)} = \frac{\phi_{1+|z}\phi_{+2|z}}{\phi_{2+|z}\phi_{+1|z}}.$$

På samma sätt som för beräkningen av den retrospektiva och prospektiva oddskvoten i sektion 3.1, så sammanfaller även den betingade retrospektiva oddskvoten med den betingade prospektiva oddskvoten

$$\mathrm{OR}_z = \frac{P(D|E,z)/P(\overline{D}|E,z)}{P(D|\overline{E},z)/P(\overline{D}|\overline{E},z)} = \frac{P(D|E,z)P(\overline{D}|\overline{E},z)}{P(\overline{D}|E,z)P(D|\overline{E},z)}. \tag{6}$$

Om vi antar ett konstant värde på OR_z, även om $P(E|D,z)$ och $P(E|\overline{D},z)$ ändras med z, så kan vi även anta att $\mathrm{OR}_z = \mathrm{OR}_c \ \forall \ z$. Där vi låter OR_c beteckna den betingade konstanta oddskvoten. Det gäller då att ϕ_{21} och ϕ_{12} uppfyller

$$\phi_{21} = \int_z \phi_{21|z} f_D(z)dz = \int_z \phi_{2+|z}\phi_{+1|z} f_D(z)dz$$

och från OR_z får vi

$$\phi_{12} = \int_z \mathrm{OR}_z \, \phi_{21|z} f_D(z)dz = \mathrm{OR}_c \int_z \phi_{2+|z}\phi_{+1|z} f_D(z)dz = \mathrm{OR}_c \, \phi_{21}$$

$\Rightarrow \mathrm{OR}_C = \phi_{12}/\phi_{21}$. Vilket vi kan skatta som

$$\frac{\widehat{\phi}_{12}}{\widehat{\phi}_{21}} = \frac{b/N}{c/N} = \frac{b}{c},$$

där b och c är frekvenserna för de diskordanta paren. Vi lägger märke till att konkordanta par inte bidrar till skattningen av OR_C.

3.3.1 Test av association för parmatchad data

Vi kan jämföra de beroende grupperna, D och $\overline{D}$, genom att jämföra ϕ_{i+} och ϕ_{+j}. Vi ansätter

$H_0 : \phi_{1+} = \phi_{+1}$
$H_1 : \phi_{1+} \neq \phi_{+1}$.

Eftersom $\phi_{1+} - \phi_{+1} = (\phi_{11} + \phi_{12}) - (\phi_{11} + \phi_{21}) = \phi_{12} - \phi_{21}$ så är nollhypotesen ovan ekvivalent med *hypotesen om symmetri*

$H_0 : \phi_{12} = \phi_{21}$.

3.3.2 McNemar's test

För hypotesprövning kan vi använda Score test statistikan som vid parmatchad data endast använder sig av diskordanta par [5]. Statistikan ges av

$$Z = \frac{\phi_{12} - \phi_{21}}{\sqrt{\text{Var}(\phi_{12} - \phi_{21}|H_0)}}$$

och enligt [5] ges variansen för $\phi_{12} - \phi_{21}$ av

$$\text{Var}(\phi_{12} - \phi_{21}) = \frac{(\phi_{12} + \phi_{21}) - (\phi_{12} - \phi_{21})^2}{N}.$$

Under $H_0 : \phi_{12} = \phi_{21}$ får vi

$$\text{Var}(\phi_{12} - \phi_{21}|H_0) = \frac{\phi_{12} + \phi_{21}}{N}$$

vilket ger oss

$$Z_0 = \frac{\hat{\phi}_{12} - \hat{\phi}_{21}}{\sqrt{(\hat{\phi}_{12} - \hat{\phi}_{21})/N}} = \frac{b/N - c/N}{\sqrt{(b/N + c/N)/N}} = \frac{b - c}{\sqrt{b + c}}.$$

Statistikan Z_0 är asymptotiskt standard normalfördelad under H_0. Den kan användas för att testa den ensidiga mothypotesen $H_1 : \phi_{12} > \phi_{21}$ eller $H_1 : \phi_{12} < \phi_{21}$. För att testa den tvåsidiga mothypotesen $H_1 : \phi_{12} \neq \phi_{21}$ använder vi Z_0^2 som är asymptotiskt χ^2-fördelad med en frihetsgrad. Testet som använder test statistikan Z_0^2 är *McNemar's test* [5].

3.3.3 Konfidensintervall

Vi börjar med att sammanställa ett tvåsidigt konfidensintervall för log oddskvoten. Därefter kan vi använda det för att konstruera ett asymmetriskt konfidensintervall för oddskvoten. Man kan anta att frekvenserna $(a, b, c, d)'$ är multinomialfördelade. Eftersom multinomialfördelningen kan approximeras med normalfördelningen har vi att

$$(\phi_{11}, \phi_{12}, \phi_{21}, \phi_{22}) \overset{d}{\approx} N(\phi, \Sigma/N)$$

där

$$\Sigma = \begin{pmatrix} \phi_{11}(1-\phi_{11}) & -\phi_{11}\phi_{12} & -\phi_{11}\phi_{21} & -\phi_{11}\phi_{22} \\ -\phi_{11}\phi_{12} & \phi_{12}(1-\phi_{12}) & -\phi_{12}\phi_{21} & -\phi_{12}\phi_{22} \\ -\phi_{11}\phi_{21} & -\phi_{12}\phi_{21} & \phi_{21}(1-\phi_{21}) & -\phi_{21}\phi_{12} \\ -\phi_{11}\phi_{22} & -\phi_{12}\phi_{22} & -\phi_{21}\phi_{22} & \phi_{22}(1-\phi_{22}) \end{pmatrix}$$

Genom att använda deltametoden får vi enligt [5] att

$$\mathrm{Var}(\log(\widehat{\mathrm{OR}}_C)) = \frac{1}{N}\left(\frac{N}{b} + \frac{N}{c}\right)$$

och

$$\widehat{\mathrm{Var}}(\log(\widehat{\mathrm{OR}}_C)) = \left(\frac{1}{b} + \frac{1}{c}\right)$$

$$= \frac{b+c}{bc}.$$

Ett tvåsidigt asymptotiskt konfidensintervall för $\log(\mathrm{OR}_C)$ ges av

$$\mathrm{CI}_{\log(\mathrm{OR}_C)} = \log \widehat{\mathrm{OR}}_C \pm Z_{1-\alpha/2}\sqrt{\frac{b+c}{bc}}$$

på intervallet $(-\infty, \infty)$. Genom att använda exponentialfunktionen får vi ett asymptotiskt konfidensintervall för den betingade oddskvoten

$$\mathrm{CI}_{\mathrm{OR}_C} = \exp\left\{\log \widehat{\mathrm{OR}}_C \pm Z_{1-\alpha/2}\sqrt{\frac{b+c}{bc}}\right\}$$

på intervallet $[0, \infty)$.

4 Multipelt test

Vid ett hypotestest av $H_0 : \pi_1 = \pi_2$ mot $H_1 : \pi_1 \neq \pi_2$ uppstår *fel av första slaget* eller *typ 1 fel* då man förkastar nollhypotesen givet den är sann. Ett tests *signifikansnivå* eller *felrisk*, α, definieras som sannolikheten för fel av första slaget

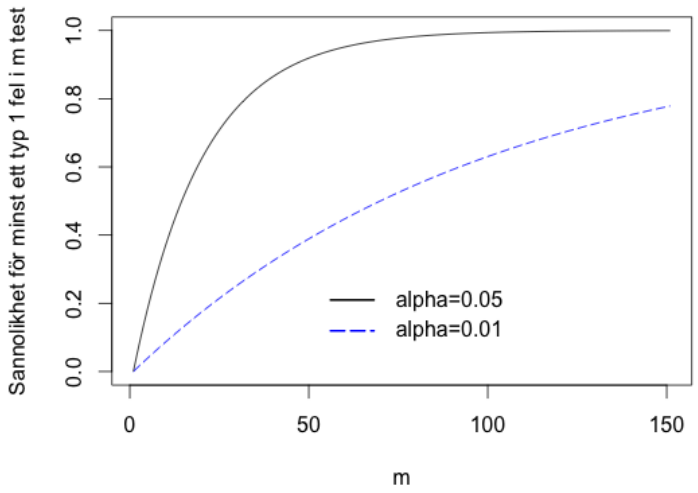
$$P(\text{att förkasta } H_0 | H_0) := \alpha.$$

Om vi väljer α=0.05 så innebär det att vi med 5% sannolikhet riskerar att förkasta H_0 av misstag. Antag istället att vi har m exponeringsvariabler och att vi, för varje variabel, vill utföra ett individuellt hypotestest. Om vi kan anta att hypotestesten är oberoende så har vi att X, som betecknar antalet typ 1 fel i m test, är binomialfördelad med parameter m och α. Om vi utför

m oberoende hypotesprövningar innebär det att vi, i minst ett av testen, riskerar att förkasta nollhypotesen av misstag med sannolikhet

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (1 - \alpha)^m$$

För $m = 20$ och $\alpha = 0.05$ blir ovanstående sannolikhet 64,15% och ökar ju fler förklarande variabler vi vill hypotespröva. Detta visas i figur 1 nedan.



Figur 1

Det är här som problemet med *multipla jämförelser* uppstår, vi får en inflation av fel av första slaget. Inom kohort- och fall-kontollstudier så kan vi vanligtvis inte anta att hypotesprövningarna är oberoende. Det beror på att datan vanligtvis består av ett set med samma personers exponeringsstatus för flera variabler. När hypotesprövningarna inte är oberoende så är ovanstående sannolikhet något mindre men en korrigering bör ändå göras [8].

Vi vill på något sätt korrigera för multipla jämförelser. Det finns flera metoder som tillämpas inom området. En av de enklaste metoder som ofta används är *Bonferronis metod* [1]. Antag att vi vill utföra m hypotesprövningar på signifikansnivå α. Enligt metoden så ska vi för varje enskilt test istället använda signifikansnivå α/m och förkasta nollhypotesen när $p_i \leq \alpha/m$, där $i = 1, 2, ..., m$. På samma vis kan vi korrigera konfidensintervall genom att sätta konfidengraden till $1-\alpha/m$. Vi kan även enkelt justera testens p-värden enligt $\tilde{p_i} = \min(p_i m, 1)$ och sedan jämföra dessa med den ursprungliga nivån α.

Bonferronis metod har fördelen att den är enkel att arbeta med då justeringen sker i ett steg. Men metoden blir väldigt konservativ när många hypotesprövningar ska utföras. Vid extremt stora värden på m blir α/m väldigt litet, vilket medför att det blir svårt att erhålla ett signifikant resultat. För $m = 100$ och $\alpha = 0.05$ krävs det att $p_i \leq 0,0005$ för att alternativ hypotesen ska vara signifikant.

En minde konservativ metod där justeringen sker stegvis är *Holms metod* [5]. Antag att vi utför m hypotesprövningar och erhåller motsvarande p-värden som vi ordnar efter stolek, enligt

$$p_1 \leq p_2 \leq \ldots \leq p_m.$$

Istället för att jämföra varje p-värde med α jämför vi

$$p_1 \leq \frac{\alpha}{m}\ , \quad p_2 \leq \frac{\alpha}{m-1}\ , \ldots ,\ p_m \leq \alpha,$$

och allmänt gäller

$$p_i \leq \frac{\alpha}{m-i+1}.$$

Vi kan även justera p-värdena och därefter jämföra dem med signifikansnivån α. De justerade p-värdena kan uttryckas som

$$\tilde{p}_i = \begin{cases} mp_1, & \text{för} \quad i = 1 \\ \max((m-i+1)p_i,\ \tilde{p}_{i-1}), & \text{för} \quad i = 2,3,\ldots,m. \end{cases}$$

där ett justerat p-värde > 1 sätts till 1. Vi får alltså att

$$\tilde{p}_1 = mp_1, \quad \tilde{p}_2 = \max((m-1)p_2,\ \tilde{p}_1) \quad \ldots \quad \tilde{p}_m = \max(p_m,\ \tilde{p}_{m-1}).$$

Med Holms metod utförs en stegvis korrigering vilket gör den något svårare

att applicera i jämförelse med Bonferronis metod. Men den är mer lämplig när studien innehåller ett stort antal hypotesprövningar eftersom den inte är lika konservativ som Bonferronis metod.

5 Enkel logistisk regression

Vi har sett hur en statistisk analys av kohortstudier och fall-kontrollstudier kan utföras med hjälp av 2×2 tabeller. Ett annat sätt att analysera dessa studier är att ansätta en logistisk regressions modell. Modellen tillhör klassen generaliserade linjära modeller och är en av de mest frekvent använda modeller för kategorisk data.

5.1 Enkel obetingad logistisk regressions modell

Antag att vi har en kohortstudie och vill skatta sannolikheten för sjukdom i gruppen med exponerade individer och för gruppen med individer som inte är exponerade. Vi utgår från samma beteckningar som i den sammanställda 2×2 tabellen med frekvenser för kohortstudie, enligt sektion 2.0. Vi har att n_{11} betecknar antal sjuka individer som är exponerade och n_{12} antal sjuka individer som inte är exponerade. Frekvensen n_{11} är binomialfördelad med parameter n_{+1} och sannolikhet π_1 och frekvensen n_{12} är binomialfördelad med parameter n_{+1} och sannolikhet π_1. Vi låter Y_i vara en binär responsvariabel som betecknar sjukdomsstatus för den i:te individen och antar värdet 1 eller 0 för sjuk respektive frisk. Den förklarande variabeln X_i betecknar exponeringsstatus för individ i och antar värdet 1 för exponerad och 0 för ej exponerad. Vi ansätter en enkel logistisk regressions modell för att skatta sannolikheten $P(y_i = 1|x_i = 1) = \pi_1$ och $P(y_i = 1|x_i = 0) = \pi_2$ enligt

$$\text{logit}(\pi_1) = \log\left(\frac{\pi_1}{1-\pi_1}\right) = \alpha + \beta, \quad \text{logit}(\pi_2) = \log\left(\frac{\pi_2}{1-\pi_2}\right) = \alpha \qquad (7)$$

Vilket vi, genom inversa transformationen, även kan skriva på formen

$$\pi_1 = \frac{\exp(\alpha+\beta)}{1+\exp(\alpha+\beta)}, \quad \pi_2 = \frac{\exp(\alpha)}{1+\exp(\alpha)}.$$

Med hjälp av (7) kan vi utrycka parametern β enligt

$$\begin{aligned}\beta &= \log\left(\frac{\pi_1}{1-\pi_1}\right) - \alpha = \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_2}{1-\pi_2}\right) \\ &= \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right) = \log\left(\frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}\right) \\ &= \log(\text{OR}).\end{aligned}$$

Vi ser därmed att modellparametern β är ekvivalent med log oddskvoten som vi räknade fram vid analys av 2×2 tabellen i sektion 2.1. Vi kan därför enkelt, direkt från modellen, beräkna oddskvoten enligt

$$e^{\beta} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} = \text{OR}$$

Vi kan skatta parametrarna α och β med maximum likelihood-metoden. Likelihooden av två binomialfördelade observationer ges av produkten av likelihood funktionerna enligt

$$\begin{aligned}L(\pi_1, \pi_2) &= P(n_{11}, n_{12} | n_{+1}, n_{+2}, \pi_1, \pi_2) \\ &= \binom{n_{+1}}{n_{11}} \pi_1^{n_{11}} (1-\pi_1)^{n_{21}} \binom{n_{+2}}{n_{12}} \pi_2^{n_{12}} (1-\pi_2)^{n_{22}} \\ &\propto \pi_1^{n_{11}} (1-\pi_1)^{n_{21}} \pi_2^{n_{12}} (1-\pi_2)^{n_{22}}\end{aligned}$$

Vi logaritmerar nu likelihoodfunktionen och får

$$\ell(\pi_1, \pi_2) = n_{11}\log(\pi_1) + n_{21}\log(1-\pi_1) + n_{12}\log(\pi_2) + n_{22}\log(1-\pi_{22})$$

$$= n_{11}\beta + (n_{11}+n_{12})\alpha - (n_{11}+n_{21})\log(1+e^{\alpha+\beta}) - (n_{12}+n_{22})\log(1+e^{\alpha})$$

$$= n_{11}\beta + n_{1+}\alpha - n_{+1}\log(1+e^{\alpha+\beta}) - n_{+2}\log(1+e^{\alpha})$$

Derivering med avseende på α och β respektive ger

$$\frac{\partial \ell}{\partial \alpha} = n_{1+} - \frac{n_{+1}e^{\alpha+\beta}}{1+e^{\alpha+\beta}} - \frac{n_{+2}e^{a}}{1+e^{\alpha}} = n_{1+} - n_{+1}\pi_1 - n_{+2}\pi_2 \tag{8}$$

$$\frac{\partial \ell}{\partial \beta} = n_{11} - \frac{n_{+1}e^{\alpha+\beta}}{1+e^{\alpha+\beta}} = n_{11} - n_{+1}\pi_1 \tag{9}$$

Vi kan nu sätta ekvationerna ovan till 0 för att därefter lösa ut parametrarna. Från ekvation (8) får vi att $\pi_1 = (n_{1+} - n_{+2}\pi_2)/n_{+1}$

Insättning av π_1 i den andra ekvationen ger

$$n_{11} - n_{+1}\left(\frac{n_{1+} - n_{+2}\pi_2}{n_{+1}}\right) = 0$$

$$\pi_2 = \frac{n_{1+} - n_{11}}{n_{+2}} = \frac{n_{12}}{n_{12}+n_{22}}$$

$$\frac{e^{\alpha}}{1+e^{\alpha}} = \frac{n_{12}}{n_{12}+n_{22}}$$

$$e^{\alpha} = \frac{n_{12}}{n_{22}}$$

och vi erhåller skattningen

$$\hat{\alpha} = \log\left(\frac{n_{12}}{n_{22}}\right).$$

För att lösa ut β använder vi $e^{\alpha} = n_{12}/n_{22}$ i ekvationen ovan och erhåller

$$\pi_1 = \frac{n_{11}}{n_{+1}}$$

$$\frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}} = \frac{n_{11}}{n_{+1}}$$

$$e^{\alpha+\beta} = \frac{n_{11}}{n_{+1}} \frac{n_{+1}}{(n_{+1}-n_{11})} = \frac{n_{11}}{n_{+1}-n_{11}}$$

$$\frac{n_{12}}{n_{22}} e^{\beta} = \frac{n_{11}}{n_{+1}-n_{11}}$$

$$e^{\beta} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

Vi får därmed, återigen, skattningen

$$\hat{\beta} = \log\left(\frac{n_{11}n_{22}}{n_{21}n_{12}}\right).$$

För $\beta = 0$ är oddskvoten 1 vilket, som tidigare nämnt, tolkas som att en risken för att drabbas av sjukdom är lika för fallgruppen och kontrollgruppen. Vi ser att då $\beta \to -\infty$ går OR $\to 0$ och att då $\beta \to \infty$ går OR $\to \infty$.

5.1.1 Konfidensintervall

Vi vill nu konstruera ett konfidensintervall för β och behöver därför variansen för modellparametern. Maximum likelihood-skattningarna är asymptotiskt normalfördelade med kovariansmatris $\text{Var}(\theta) = \boldsymbol{I}(\theta)^{-1}$ [1]. Här betecknar $\boldsymbol{I}(\theta)$ den förväntade fisherinformationsmatrisen som definieras som

$$\boldsymbol{I}(\theta) = -E\begin{bmatrix} \frac{\partial^2 \ell}{\partial \alpha^2} & \frac{\partial^2 \ell}{\partial \alpha \beta} \\ \frac{\partial^2 \ell}{\partial \alpha \beta} & \frac{\partial^2 \ell}{\partial \beta^2} \end{bmatrix}$$

där $\theta = (\alpha, \beta)'$. De partiella derivatorna ges av

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \alpha^2} &= -\frac{n_{+1} e^{\alpha+\beta}}{(1+e^{\alpha+\beta})^2} - \frac{n_{+2} e^{\alpha}}{(1+e^{\alpha})^2} \\ &= -n_{+1}\pi_1(1-\pi_1) - n_{+2}\pi_2(1-\pi_2), \end{aligned}$$

$$\frac{\partial^2 \ell}{\partial \beta^2} = -\frac{n_{+1} e^{\alpha+\beta}}{(1+e^{\alpha+\beta})^2} = -n_{+1}\pi_1(1-\pi_1),$$

$$\frac{\partial^2 \ell}{\partial \alpha \beta} = -\frac{n_{+1} e^{\alpha+\beta}}{(1+e^{\alpha+\beta})^2} = -n_{+1}\pi_1(1-\pi_1).$$

Allmänt så gäller att inversen av en inverterbar 2×2 matris är

$$C^{-1} = \frac{1}{\det(|C|)} \begin{pmatrix} c_{22} & -c_{21} \\ -c_{12} & c_{11} \end{pmatrix} = \frac{1}{c_{11}c_{22} - c_{12}c_{21}} \begin{pmatrix} c_{22} & -c_{21} \\ -c_{12} & c_{11} \end{pmatrix}$$

där $\det(|C|)$ är determinanten för matrisen C. Kovariansmatrisen för modellparametrarna ges därmed av

$$\widehat{\mathrm{Var}}(\hat{\theta}) = I(\hat{\theta})^{-1} = \frac{1}{n_1\pi_1(1-\pi_1)n_2\pi_2(1-\pi_2)} \begin{pmatrix} n_1\pi_1(1-\pi_1) & -n_1\pi_1(1-\pi_1) \\ -n_1\pi_1(1-\pi_1) & n_1\pi_1(1-\pi_1) + n_2\pi_2(1-\pi_2) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{n_2\pi_2(1-\pi_2)} & -\frac{1}{n_2\pi_2(1-\pi_2)} \\ -\frac{1}{n_2\pi_2(1-\pi_2)} & \frac{1}{n_1\pi_1(1-\pi_1)} + \frac{2}{n_2\pi_2(1-\pi_2)} \end{pmatrix}.$$

Vi förenklar variansen för β och får

$$\mathrm{Var}(\beta) = \frac{1}{n_1\pi_1(1-\pi_1)} + \frac{1}{n_2\pi_2(1-\pi_2)} = \frac{(1+e^{\alpha+\beta})^2}{n_1 e^{\alpha+\beta}} + \frac{(1+e^{\alpha})^2}{n_2 e^{\alpha}}$$

$$= \frac{(1+n_{11}/n_{21})^2}{(n_{11}+n_{21})(n_{11}/n_{21})} + \frac{(1+n_{12}/n_{22})^2}{(n_{12}+n_{22})(n_{12}/n_{22})} = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

Vi kan nu konstruera ett tvåsidigt konfidensintervall för modellparametern β enligt

$$\mathrm{CI}_\beta = \hat{\beta} \pm z_{1-\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

där $z_{1-\alpha/2}$ betecknar normalfördelningens kvantil. Vi lägger märke till att konfidensintervallet är ekvivalent med konfidensintervallet för oddskvoten vid analys av 2×2 tabellen enligt (4).

Som tidigare nämnt är oddskvoten 1 då modellparametern β är noll, vilket motsvarar att $\pi_1 = \pi_2$. Det är därför av intresse att testa nollhypotesen $H_0 : \beta = 0$ mot alternativ hypotesen $H_1 : \beta \neq 0$. Det finns flera test som används för detta. Ett bland de vanligaste är *Walds test* [1].

5.1.2 Wald test

Wald statistikan kan användas för att testa nollhypotesen $H_0 : \beta = 0$ mot alternativ hypotesen $H_1 : \beta \neq 0$, eller mot en ensidig alternativ hypotes

$H_1 : \beta > 0$. Statistikan ges av

$$Z_w = \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}} = \frac{\log(n_{11}n_{22}/n_{12}n_{21})}{\sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}}}$$

där β är maximum likelihood-skattningen och $\mathrm{Var}(\hat{\beta})$ är variansen under alternativhypotesen. Under H_0 är statistikan asymptotiskt standard normalfördelad.

5.2 Enkel betingad logistisk regression

En betingad logistisk regressions modell kan användas för att analysera en matchad fall-kontrollstudie. Modellen hanterar både parmatchning och matchade set med fler kontroller per fall. Antag att vi har N stycken matchade set av storlek n_i där $i = 1, ..., N$. Eftersom fall-kontrollstudien är retrospektiv, låter vi y_{ij} vara en binär responsvariabel som antar värdet 1 om j:te medlemmen i set i är exponerad och 0 om medlemmen inte är exponerad. Modellen kan uttryckas som

$$\phi_{ij} = P(y_{ij} = 1|x_{ij}) = \frac{\exp(\alpha_i + x_{ij}\beta)}{(1 + \exp(\alpha_i + x_{ij}\beta))}$$

eller

$$\mathrm{logit}(\phi_{ij}) = \log\left(\frac{\phi_{ij}}{1 - \phi_{ij}}\right) = \alpha_i + x_{ij}\beta$$

där x_{ij} är den förklarande variabeln som antar värdet 1 om den j:te medlemmen i set i är sjuk och 0 om denne är frisk. Varje matchat set i har ett tillhörande intercept α_i och en gemensam modellparameter β. För enkelhets skull antar vi att varje set i består av ett fall och en kontroll, dvs. parmatchning. Enligt [5] har vi att $\exp(\beta)$ är den betingade retrospektiva oddskvoten för i:te matchade par

$$\text{OR}_{\text{retro}} = \frac{\phi_{i1}/(1-\phi_{i1})}{\phi_{i2}/(1-\phi_{i2})} = \frac{\phi_{i1}(1-\phi_{i2})}{\phi_{i2}(1-\phi_{i1})} = \frac{P_i(E|D)P_i(\overline{E}|\overline{D})}{P_i(\overline{E}|D)P_i(E|\overline{D})} = e^{\beta}.$$

På samma sätt som i (6), har vi att den retrospektiva betingade oddskvoten sammanfaller med den betingade prospektiva oddskvoten.

Som vi ser så beror inte oddskvoten på α_i, $i = 1, ..., N$. För att skatta parametern β använder vi oss av den betingade maximum likelihood-metoden, på så sätt kan vi undvika att behöva skatta samtliga a_i. Vi låter S_i vara summan av antalet positiva utfall inom varje par, dvs då y_{ij} eller x_{ij} antar värdet 1. Vi har att alltså att $S_i = y_{ij} + x_{ij}$ kan anta värdet 0,1 eller 2, exempelvis så är $S_i = 1$ för diskordanta par, $(y_{i1} = 1, x_{i,2} = 0)$ eller $(y_{i1} = 0, x_{i,2} = 1)$. Den betingade likelihoodfunktionen kan uttryckas som

$$L_c(\beta)_{|S} = \prod_i^N \frac{P(y_{i1}, x_{i2})}{P(S_i)}$$

För $S_i = 1$ får vi

$$\begin{aligned} L_c(\beta)_{|S_i=1} &= \prod_{i=1}^N \left(\frac{P(y_{i1}=1, x_{i2}=0)}{P(S_i=1)}\right)^{y_{i1}(1-x_{i2})} \left(\frac{P(y_{i1}=0, x_{i2}=1)}{P(S_i=1)}\right)^{y_{i2}(1-x_{i1})} \\ &= \prod_{i=1}^N \left(\frac{\phi_{i1}(1-\phi_{i2})}{\phi_{i1}(1-\phi_{i2}) + \phi_{i2}(1-\phi_{i1})}\right)^{y_{i1}(1-x_{i2})} \left(\frac{\phi_{i2}(1-\phi_{i1})}{\phi_{i1}(1-\phi_{i2}) + \phi_{i2}(1-\phi_{i1})}\right)^{y_{i2}(1-x_{i1})} \\ &= \prod_{i=1}^N \left(\frac{e^{\beta}}{1-e^{\beta}}\right)^{y_{i1}(1-x_{i2})} \left(\frac{1}{1+e^{\beta}}\right)^{y_{i2}(1-x_{i1})} = \left(\frac{e^{\beta}}{1-e^{\beta}}\right)^{b} \left(\frac{1}{1+e^{\beta}}\right)^{c}. \end{aligned}$$

Vi får därmed ett uttryck som endast beror på β. Den betingade log likelihoodfunktionen är

$$\ell_c(\beta)_{|S_i=1} = b\beta - (b+c)\log(1+e^{\beta})$$

och derivering ger

$$\frac{\partial \ell_c(\beta)_{|S_i=1}}{\partial \beta} = b - \frac{(b+c)e^{\beta}}{1+e^{\beta}}. \tag{10}$$

Genom att sätta ekvation (10) till noll får vi skattningen $\hat{\beta}_{MLE} = \log(b/c)$. Vi ser att det endast är diskordanta par som används vid uträkning av $\hat{\beta}_{MLE}$. På motsvarande sätt kan vi även beräkna den betingade maximum likelihoodfunktionen, och skattningen av β, då vi har fler matchade kontroller per fall. Även då kommer endast diskordanta set att användas för skattning av parameter β [5].

5.2.1 Konfidensintervall

Variansen av parameter β ges av inversen av den förväntade fisherinformationen. Vi har att

$$\frac{\partial^2 \ell(\beta)_{|S_i=1}}{\partial \beta^2} = -\frac{(b+c)e^{\beta}}{(1+e^{\beta})^2}.$$

Den förväntade fisherinformationen är

$$I(\beta) = -E\left[-\frac{(b+c)e^{\beta}}{(1+e^{\beta})^2}\right] = \frac{bc}{b+c}$$

och slutligen har vi att

$$\mathrm{Var}(\beta) = I(\hat{\beta})^{-1} = \frac{b+c}{bc}.$$

Ett tvåsidigt konfidensintervall för $\beta = \log(\text{OR})$ ges av

$$\text{CI}_\beta = \hat{\beta} \pm z_{1-\alpha/2}\sqrt{\frac{b+c}{bc}}$$

där $z_{1-\alpha/2}$ betecknar normalfördelningens kvantil. Genom att använda exponentialfunktionen fås ett tvåsidigt konfidensintervall för $e^{\hat{\beta}} = \widehat{\text{OR}}$ enligt

$$\text{CI}_{e^\beta} = \exp\left\{\hat{\beta} \pm Z_{1-\alpha/2}\sqrt{\frac{b+c}{bc}}\right\}.$$

5.2.2 Wald test

Vi vill nu testa hypotesen om att modellparametern β är noll, $H_0 : \beta = 0$, mot alternativ hypotesen $H_1 : \beta \neq 0$. Vi kan använda Wald statistikan för att utföra testet, precis som vid obetingad logistisk regression. Statistikan ges av

$$Z_w = \frac{\hat{\beta}}{\sqrt{\widehat{\text{Var}}(\hat{\beta})}} = \frac{\log(b/c)}{\sqrt{(b+c)/bc}}$$

där β är maximum likelihood-skattningen, $\text{Var}(\hat{\beta})$ är variansen beräknad under alternativhypotesen och statistikan är asymptotiskt standard normalfördelad under nollhypotesen.

6 Statistisk analys av fall-kontrollstudie

Vi går nu vidare med att utföra en statistisk analys av fall-kontrollstudien över utbrottet av salmonella typhimurium i Danmark 1996. Vi utför analysen med hjälp av programvaran R. Eftersom datan innehåller två kontroller per fall väljer vi att använda oss av enkel betingad logistisk regression. Vi skapar en funktion som applicerar modellen för varje exponeringsvariabel

och sammanställer resultatet i en och samma tabell.

```
cctable<-function(caseColName, matchColName, exposureColName, data){}
```

För att använda funktionen anges, som "caseColName", namnet på responsvariabeln i datasetet, dvs. sjukdomsstatus. Därefter anges namn på matchningsvariabel som "matchColName". Funktionen är konstruerad på så sätt att användaren själv kan välja om samtliga exponeringsvariabler, i ett angivet dataset, ska användas eller endast några av dem. För att använda samtliga exponeringsvariabler i datasetet anges "all" som "exposureColName". Om man endast vill använda några exponeringsvariabler anges namnen på dessa i form av en vektor. Samtliga variabelnamn ska motsvara kolumnernas namn i datasetet.

Funktionen sammanställer antal exponerade och oexponerade fall och kontroller, oddskvoter för respektive exponeringsvariabel med 95% konfidensintervall, p-värden och justerade p-värden, där vi använder oss av Holms metod.

Fördelen med funktionen är att den kan använda samtliga exponeringsvariabler i ett dataset och sammanställa outputen, för varje variabel, i en och samma tabell. Detta ger en bättre överblick av resultat och är betydligt mer tidsbesparande än att utföra statistisk analys på en exponeringsvariabel i taget. Tanken med funktionen är även att den kan underlätta arbete och statistisk analys av framtida studier där man har ett stort antal exponeringsvariabler.

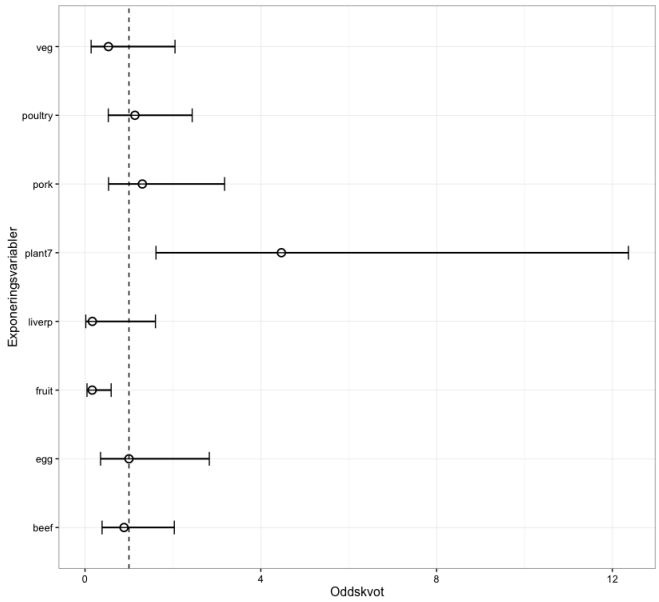
Vi skapar även en motsvarande funktion, som använder enkel obetingad logistisk regression, och kan användas vid analys av kohortstudier. Fullständig översikt av funktionerna presenteras i appendix.

6.1 Resultat

Vi använder funktionen till att analysera datan för fall-kontrollstudien över utbrottet av salmonella typhimurium och får nedanstående sammanställning.

	Exposed		Unexposed			CI 95%			
Exposure	Case	Total	Case	Total	OR	limit1	limit2	p	Adj.p
abroad	3	3	44	133	∞	0.000	∞	0.998	1.000
beef	28	88	15	42	0.8880	0.388	2.032	0.778	1.000
pork	33	98	8	31	1.3043	0.536	3.175	0.558	1.000
veal	5	7	37	124	∞	0.000	∞	0.998	1.000
poultry	14	41	26	88	1.1360	0.529	2.438	0.743	1.000
liverp	43	130	3	4	0.1667	0.017	1.602	0.121	0.966
veg	40	122	5	10	0.5326	0.139	2.048	0.359	1.000
fruit	32	113	12	19	0.1634	0.045	0.594	0.006	0.054
egg	31	98	9	29	1.0000	0.354	2.826	1.000	1.000
plant7	29	62	10	61	4.4686	1.615	12.37	0.004	0.039

Exponeringsvariablerna “fruit” och “plant7” är signifikanta på 5% nivån då vi utgår från de ojusterade p-värdena. Inga andra exponeringsvariabler är signifikanta på någon rimlig nivå. När vi tar hänsyn till att multipla jämförelser har gjorts, vi utgår från de justerade p-värdena, är endast exponeringsvariabeln “plant7” signifikant på 5% nivån. Vi har att $\widehat{\text{OR}}_{\text{plant7}} = 4.47$ vilket tolkas som att risken för att drabbas av salmonella typhimurium är 4.47 gånger större för dem som har ätit kött från “plant7” under de senaste två veckorna än för dem som inte har har gjort det. Som tidigare nämnt, använder vi oss av Holms metod för justering av p-värden. En plott med de skattade oddskvoterna och tillhörande konfidensintervall visas nedan, med undantag för variabeln “abroad” och “veal”.



Figur 2: oddskvot och konfidensintervall

Vi ser från sammanställningen att skattningarna av oddskvoten för exponeringsvariablerna abroad och veal är oändliga. Motsvarande 95% konfidensintervall för oddskvoterna visar en oändlig övre gräns. Detta kan bero på att det saknas diskordanta set där fallet är oexponerad. Vi undersöker exponeringsstatus för fallen och kontrollerna i varje set för "abroad" och "veal". För exponeringsvariabeln "abroad" får vi följande resultat

	Exposed controls	
	0	1
Unexposed case	5	0
Exposed case	0	0

	Exposed controls		
	0	1	2
Unexposed case	39	0	0
Exposed case	3	0	0

Vi ser att det finns fem matchade set, med endast en kontroll, där både fallet och kontrollen är oexponerade. Vi har 39 set med ett oexponerat fall och två oexponerade kontroller, samt tre diskordanta set med ett exponerat fall och två oexponerade kontroller. Vi saknar dock diskordanta set där fallet är oexponerad och kontrollen exponerad vilket förklarar skattningen.

För exponeringsvariabeln “veal” får vi

	Exposed controls	
	0	1
Unexposed case	3	0
Exposed case	0	1

	Exposed controls		
	0	1	2
Unexposed case	34	0	0
Exposed case	3	1	0

På samma sätt kan vi läsa av att veal endast har 3 diskordanta set där fallet är exponerad och de två kontrollerna är oexponerade. Alltså saknar vi även här diskordanta set där fallet är eoxponerad. Detta förklarar de skattade oddskovtarna och konfidensintervallens övre gränser för exponeringsvariablerna abroad och veal.

6.2 Slutsats och diskussion

Vi såg att kohort- och fall-kontrollstudier skiljer sig åt på sättet information om deltagarnas sjukdoms- och exponeringsstatus samlas in. Kohortstudien är prospektiv, vi får först information om huruvida en individ är exponerad, för att därefter få information om dennes sjukdomsstautus. I en fall-kontrollstudie har vi den omvända situationen. Vi erhåller först information om individens sjukdomsstatus och därefter huruvida individen är exponerad eller inte. Skillnaden i hur datan samlas in bör has i åtanke när man konstruerar oddskvoten i en fall-kontrollstudie då vi är intresserade av att studera sjudomsstatus givet exponering snarare än exponeringsstatus givet huruvida personen är sjuk eller inte. Men genom ett antagande om att prevalensen för sjukdom i populationen är känd, och med hjälp av Bayes sats, kan vi bestämma och skatta oddskvoten på ett korrekt sätt. Som vi såg, så behöver vi i slutändan inte känna till prevalensen numeriskt då denna kancelleras i beräkningen.

En annan viktig del i den statistiska teorin för studierna är multipla jämförelser. Vi såg att det uppstår en inflation av typ 1 fel när studien innehåller många exponeringsvariabler och när vi, för varje variabel, utför ett individuellt hypotestest av $H_0 : \pi_1 = \pi_2$ mot $H_1 : \pi_1 \neq \pi_2$. Vi gick igenom två metoder som ofta används för att korrigera för multipla jämförelser, Bonferronis metod och Holms metod. Vi såg dock att den förstnämnda metoden blir för konservativ när man har ett stort antal exponeringsvariabler, det blir svårt att erhålla ett signifikant resultat. Holms metod, där korrigeringen

sker stegvis till skillnad från Bonferronis metod, är inte lika konservativ och därmed lämpligare att använda.

Statistisk analys av kohort- och fall-kontrollstudier kan vara tidskrävande. För att underlätta arbetet skapade vi en funktion i programmet R som vi använde för att analysera fall-kontrollstudien över utbrottet av salmonella typhimurium i Danmark. Till grund för funktionen valde vi enkel betingad logistisk regression. Funktionen applicerar modellen för varje exponeringsvariabel och sammanställer resultatet, för varje variabel, i en och samma tabell. Funktionen är tidsbesparande, speciellt då man har många exponeringsvariabler som ska undersökas. Tanken är att den även kan underlätta statistisk analys av framtida fall-kontrollstudier.

I resultatet såg vi att två livsmedel, "fruit" och "plant7", var signifikanta på 5% nivå. När vi justerade för multipla jämförelser så var endast exponeringsvariabeln "plant7" signifikant. Vi fick att $\widehat{\mathrm{OR}}_{\mathrm{plant7}} = 4.47$, vilket tolkas som att risken för att drabbas av salmonella typhimurium är 4.47 gånger större för dem som har ätit kött från "plant7" under de senaste två veckorna än för dem som inte har har gjort det. Vi hade behövt mer information om hur "plant7" ska tolkas för att kunna dra en slutsats om resultatet, alternativt fortsätta analysen. Vi fann inget annat livsmedel som skulle kunna vara en tänkbar källa till utbrottet av salmonella typhimurium.

Referenser

[1] Agresti Alan, (2013), *Categorical Data Analysis*, Third edition, John Wiley & Sons

[2] Alm Sven Erick, Britton Tom, (2008), *Stokastik - Sannolikhethetsteori och statistikteori med tillämpningar*, Liber

[3] Grolemund Garrett, (2014), *Hands On Programming with R*, O'Reilly Media

[4] Jones Owen, Maillardet Robert, Robinson Andrew, (2009), *Introduction To Scientific Programming and Simulation Using R*, Taylor & Francis Group

[5] Lachin John, (2011), *Biostatistical Methods - The Assessment of Relative Risks*, Second edition, John Wiley & Sons

[6] **Webbplats** Folkhälsomyndigheten, (2015), *Sjukdomsinformation om livsmedelsburna utbrott inklusive matförgiftning*, hämtad från `https://www.folkhalsomyndigheten.se/amnesomraden/smittskydd-och-sjukdomar/ smittsamma-sjukdomar/livsmedelsburna-utbrott-inklusive-matforgiftning/`

[7] **Webbplats** Livsmedelsverket, (2015), *Sjukdomsframkallande mikroorganismer, salmonella*, hämtad från `http://www.livsmedelsverket.se/livsmedel-och-innehall/bakterier-virus-och-parasiter1/sjukdomsframkallande-mikroorganismer/salmonella/`

[8] **Webbplats** Glickman Mark, Rao Sowmya, Schultz Mark, Journal of Clinical Epidemiology, (2015), *False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies*, hämtad från `https://http://glicko.net/research/multiple-tests.pdf`

A Deltametoden

Låt T vara en stokastisk variabel med $E(T) = \mu$ och $\text{Var}(T) = \sigma^2$. Låt vidare $Y = g(t)$ där $g(\cdot)$ är en godtycklig funktion som är två gånger deriverbar. Taylorutveckling av $g(t)$ i en omgivning av μ ger

$$g(t) = g(\mu) + (t - \mu)g'(\mu) + O(\mu^*)$$

för något μ^* mellan t och μ, och där $O(\mu^*) = (t - \mu)^2 g''(\mu^*)/2$ är resttermen. För $t \xrightarrow{p} \mu$ är

$$g(t) \cong g(\mu) + (t - \mu)g'(\mu)$$

och väntevärde och varians för $Y = g(t)$ ges av

$$E(Y) = \mu_y = E[g(t)] \cong g(\mu) + g'(\mu)E(t - \mu) = g(\mu) =$$

$$\text{Var}(Y) = E(Y - \mu_y)^2 \cong E[g(t) - g(\mu)]^2 = E[g'(\mu)(t - \mu)]^2 = [g'(\mu)]^2 \,\text{Var}(T).$$

B Fördelningar

Tabell över fördelningar som används i arbetet

Fördelning		Sannolikhetsfunktion/Täthetsfunktion	E(X)	Var(X)
Binomial	$Bin(n,p)$	$\text{p}_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, .., n$	np	$np(1-p)$
Normal	$N(\mu, \sigma^2)$	$\text{f}_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty$	μ	σ^2
	$N(0,1)$	$\text{f}_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad -\infty < x < \infty$	0	1
Fördelning		Sannolikhetsfunktion	E(x_j)	Var(x_j)
Multinomial	$M_c(N, (\pi_1, \pi_2, .., \pi_c)')$	$P(x_1, ..., x_c) = \frac{N!}{\prod_{j=1}^c x_j!} \prod_{j=1}^c \pi_j^{x_j}$	$N\pi_j$	$N\pi_j(1-\pi_j)$

C Funktioner

```
#########################################################################################################
### Function:
###  The function applies the conditional logistic regression model using each exposure variables
###  as a single main effect.
###  It performs multiple testing using the Holm's method and summarizes the univariate findings in one table.
###
### Arguments:
###  caseColName - Case-control status, (response variable).
###  matchColName - A matching variable, optional.
###  exposureColName - The exposure variables.
###  The user can either choose some of the exposure variables given in a data frame or all of them.
###  Selected exposure variables should be entered as a vector and all exposure variables with the notation "all".
###  data - data frame
###
### Returns:
###  The function returns a table containing:
###  the number of exposed/unexposed cases and totals for each exposure
###  odds ratios for each exposure
###  confidence interval for the odds ratios
###  p-values and adjusted p-values using Holm procedure
###
#########################################################################################################


cctable <- function(caseColName, matchColName, exposureColName, data){
  if(exposureColName == "all" && length(exposureColName == 1)){
    l <- as.list(data[, !names(data) %in% c(caseColName, matchColName)])
  }
  else{
    l <- data[, exposureColName]
  }
  mod <- lapply(l, function(x) clogit(data[, caseColName]~x+strata(data[, matchColName])))
  tc <- sapply(l, function(f) table(data[, caseColName],f)) # number of exposed & unexposed cases and controls
  cc <- matrix(c(tc[4,],colSums(tc[c(3,4), ]), tc[2, ], colSums(tc[c(1,2), ])), nrow=length(l), ncol=4)
  # matrix with number of exposed/unexposed cases and total.
  conf <- sapply(mod, function(f) summary(f)$conf.int[3:4])      # matrix with  CI, upper & lower bounds
  OR <- sapply(mod, function(f) summary(f)$coef[2])               # vector with odds ratios
  p <- sapply(mod, function(f) summary(f)$waldtest[3])           # vector with p-values
  p.adj <- p.adjust(p, method = "holm", length(p))
  ccdata <- data.frame(cc, OR, t(conf), p, p.adj)
  colnames(ccdata) <- c("Exposed.case", "Exposed total", "Unexposed.case",
  "Unexposed.total", "Odds.ratio", "Conf.int.lower.bound", "Conf.int.upper.bound", "Pvalue", "Adj.pvalue")
  return(ccdata)
}



cctable(caseColName="case",matchColName="set",c("exposure1","exposure2","exposure2"),data=nameofdata)
# three exposures from the data frame

cctable(caseColName="case",matchColName="set",exposureColName="all",data=nameofdata)
# all exposures in the data frame




#########################################################################################################
### Function:
###  The function applies the  logistic regression model using each exposure variables as a single main effect.
###  It performs multiple testing using the Holm's method and summarizes the univariate findings in one table.
###
### Arguments:
###  caseColName - Case-control status, (response variable).
###  matchColName - A matching variable, optional.
###  exposureColName - The exposure variables.
###  The user can either choose some of the exposure variables given in a data frame or all of them.
###  Selected exposure variables should be entered as a vector and all exposure variables with the notation "all".
###  data - data frame
###
### Returns:
###  The function returns a table containing:
###  the number of exposed/unexposed cases and totals for each exposure
###  odds ratios for each exposure
###  confidence interval for the odds ratios
###  p-values and adjusted p-values using Holm procedure
###
#########################################################################################################

cstable <- function(caseColName, exposureColName, data){
  if(exposureColName == "all" && length(exposureColName == 1)){
   # check if the user wants to use all the exposure variables in the data frame
```

```
    l <- as.list(data[, !names(data) %in% caseColName])        # if true - create a list with all the exposure variables
  }
  else{                                     # if not true - create a list with exposure variables from the data frame
given by vector caseColName
    l <- data[, exposureColName]
  }
  model <- lapply(l, function(x) glm(data[, caseColName]~x, family=binomial))            # simple logistic regression
  or <- sapply(model, function(x) exp(summary(x)$coef[2,1]))                                    # vector with odds
ratios
  ci <- sapply(model, function(x) exp(coef(x)[2] + c(-1,1) * qnorm(1-0.01/2) * sqrt(vcov(x)[2,2])))
   # 95 % wald CI upper & lower bounds
  tab <- sapply(l, function(f) table(data[,caseColName], f))    # number of exposed & unexposed cases and controls
  status.cases.total <- matrix(c(tab[4,], colSums(tab[c(3,4),]), tab[2,], colSums(tab[c(1,2),])), nrow=length(l),
ncol=4)  # matrix with number of exposed & unexposed cases and total.
  p <- sapply(model, function(x) summary(x)$coef[2,4])                            # vector with p-values
  p.adj <- p.adjust(p, method = "holm", length(p))                  # vector with adjusted p-values, using Holm procedure
  csdata <- data.frame(status.cases.total, or, t(ci), p, p.adj)
  colnames(csdata) <- c("Exposed.case", "Exposed total", "Unexposed.case",
  "Unexposed.total", "Odds.ratio", "Conf.int.lower.bound", "Conf.int.upper.bound", "Pvalue", "Adj.pvalue")
  return(csdata)
}

cstable(caseColName="case",exposureColName="all",data=nameofdata) # all exposures in the data frame

cstable(caseColName="case",c("exposure1","exposure2","exposure3"),data=nameofdata)  #three exposures from the data
frame
```