UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2017:5

Cirkulära data och dess statistiska tillämpningar

Erik Persson

Examensarbete i matematik, 15 hp
Handledare: Jesper Rydén
Examinator: Jörgen Östensson
April 2017

Department of Mathematics
Uppsala University

Sammanfattning

I denna uppsats ges en introduktion till cirkulära data och dess statistiska tillämpningar. De mest grundläggande verktygen nödvändiga för tolkning av cirkulära data redovisas, såsom beräkning av medelvärde och spridning. Även metoder för att utföra tester presenteras. Avslutningsvis exemplifieras dessa statistiska redskap genom ett eget exempel med data från Transportstyrelsen.

Innehållsförteckning

1. Inledning

Vi använder oss utav cirkulära data dagligen utan att ägna någon större tanke åt saken. Exempelvis stöter vi på det i och med att vi avläser aktuell tid på ett ur, läser av en kompass för att undvika att gå vilse eller för den ornitologiintresserade som vill undersöka flyttfåglars migrationsmönster. Det dyker även upp i flera vetenskapliga fält såsom biologi, geografi, geofysik, medicin, meteorologi och oceanografi.[1] Hur vi än använder oss utav cirkulära data har de några gemensamma egenskaper som definierar dem.

Cirkulära data saknar tydlig nollpunkt och höga och låga värden är godtyckliga, i exemplet med kompass finns det inget som fysiskt berättigar att norr ska bli tilldelad grad 0 (eller 360) och inget som tyder på att en riktning av 180° är större än en på 90°. Även tiden på dygnet är utan nollpunkt och har därför tilldelats en nollpunkt vid midnatt. En timme motsvarar 15° på en cirkel och följaktligen motsvaras en grad av fyra minuter. All cirkulära data kan översättas till grader vilket ofta är nödvändigt då det representeras grafiskt med fördel utav just en cirkel och med antingen punkter eller staplar. Ytterligare en egenskap som följd av godtycklig nollpunkt är att, som i exemplet med årets månader, januari (första månaden på året) ligger lika nära februari (nr två) som december (nr tolv).

Huvudkällan i detta arbete har varit Zar (2010), då framför allt kapitel 26 och 27 om cirkulära data. I slutet av arbetet redovisas ett exempel som baseras på data från Transportstyrelsen.

2. Ett stickprovs medelvinkel $\bar{a}$ och spridning

När man behandlar cirkulära data finns det ett antal viktiga verktyg man behöver ha till sitt förfogande. Att kunna beräkna medelvinkeln på ett stickprov är nödvändigt för att kunna tolka materialet. Eftersom den cirkulära skalan har en godtycklig nollpunkt är vissa grundläggande metoder ej applicerbara. Ta, till exempel, tre riktningar på en kompass: 5°, 15° och 355°. Då hade medelvinkeln, beräknad som aritmetiskt medelvärde, blivit (5°+15°+355°)/3 = 125°. Detta ger mycket dålig förklaring för datamaterialet ty de tre vinklarna pekar norr på en kompass medan medelvärdet har en sydöstlig riktning. Det man istället bör använda sig av är ett mått som tar hänsyn till den cirkulära skalans egenskaper och ger en bättre förklaring av stickprovet.

Säg att man har ett stickprov med n stycken vinklar, a_1, a_2, a_3,...,a_n. För att beräkna medelvinkeln $\bar{a}$ behöver man först beräkna medelvinkelns rektangulära koordinater, X och Y, enligt följande:

$$X = \frac{\sum_{i=1}^{n} \cos a_i}{n} \quad , \quad Y = \frac{\sum_{i=1}^{n} \sin a_i}{n} \qquad (1), (2)$$

[1] Fisher, s 1.

Ofta är cirkulära data grupperade, för att ta hänsyn till detta gör man en alternering till ekvation (1) och (2) genom att helt enkelt multiplicera med frekvensen (f_i) av varje vinkel:

$$\boldsymbol{X = \frac{\sum_{i=1}^{n} f_i \cos a_i}{n} \quad , \quad Y = \frac{\sum_{i=1}^{n} f_i \sin a_i}{n}} \qquad \textbf{(3), (4)}$$

Med dessa två komponenter (X och Y) kan man sedan beräkna *r*, längden av medelvektorn, som beskriver hur väl medelvinkeln tillika medelvektorn beskriver datamaterialet (*r* är alltså ett mått på stickprovets koncentration):

$$\boldsymbol{r = \sqrt{X^2 + Y^2}} \qquad \textbf{(5)}$$

Ytterligare en aspekt att begrunda när man behandlar grupperade data är att det därefter beräknade r-värdet blir en aning skevt. För att korrigera detta bör man, om fördelningen är unimodal, multiplicera sitt *r* med en korrektionskoefficient *c*:

$$\boldsymbol{r_c = cr} \qquad \textbf{(6)}$$

där r_c är det korrigerade r-värdet och *c* fås enligt:

$$\boldsymbol{c = \frac{\frac{d\pi}{360°}}{\sin\left(\frac{d}{2}\right)}} \qquad \textbf{(7)}$$

där *d* är intervallängden av datamaterialets grupper, till exempel 30° för månadsvis data. Om *d* <30° blir korrigeringen oväsentlig.

Därefter kan man få fram medelvinkeln $\bar{a}$ med hjälp av:

$$\boldsymbol{\cos \bar{a} = \frac{X}{r} \quad , \quad \sin \bar{a} = \frac{Y}{r}} \qquad \textbf{(8), (9)}$$

Det finns endast en vinkel som har ovanstående värden på cosinus och sinus. Den erhåller man enklast med hjälp utav arccos och arcsin. Ytterligare en trigonometrisk likhet mellan medelvinkeln och dess rektangulära koordinater är:

$$\tan \bar{a} = \frac{\sin \bar{a}}{\cos \bar{a}} = \frac{X}{Y} \quad (10)$$

Om r = 0 finns ej någon medelriktning ty medelvinkeln är odefinierad.

2.1 Konfidensintervall för medelvinkeln

Konfidensgränserna och konfidensintervallet för medelvinkeln kan uttryckas enligt följande:

$$\bar{a} \pm d \quad (11)$$

eller

$$[\bar{a} - d,\ \bar{a} + d] \quad (12)$$

där *d* beräknas enligt följande för n ≥ 8 och r ≤ 0,9:

$$d = \cos^{-1}\left(\frac{\sqrt{\frac{2n\left(2R^2 - n\chi^2_{\alpha,1}\right)}{4n - \chi^2_{\alpha,1}}}}{R}\right) \quad (13)$$

och för n ≥ 8 och r ≥ 0,9:

$$d = \cos^{-1}\left(\frac{\sqrt{n^2 - (n^2 - R^2)e^{\chi^2_{\alpha,1}/n}}}{R}\right) \quad (14)$$

där

$$\boldsymbol{R = nr} \quad (15)$$

och kallas ”Rayleighs *R*”, återkommer till detta senare. I ekvation (13) och (14) innebär $\chi^2_{\alpha,1}$ chitvåfördelning med en frihetsgrad och konfidensgrad α.

2.2 Spridning

Ett medelvärde (eller medelvinkel i vårt fall) säger inte mycket om ett stickprov utan ett mått på spridningen. Dels kan man definiera stickprovets spann som vinkeln på den minsta cirkelbågen som innehåller all data. Till exempel om vi har ett stickprov innehållandes följande riktningar 23°, 41° och 355° är spannet minsta avståndet mellan de yttersta riktningarna. I vårt exempel blir det alltså avståndet mellan 355° och 41° som är 46°. Ett annat sätt att mäta spridning på cirkulära data är genom att beräkna värdet på den ovan nämnda variabeln *r*. Värdet på *r* kan variera mellan 0, där spridningen på stickprovet är så pass stor att någon medelvinkel ej existerar, och 1. Har *r* värdet 1 är alla observationer koncentrerade i en punkt. En anmärkning vid fallet *r* = 0 är att de ej medför att det är en likformig distribution utan det kan vara så att hälften av observationerna är koncentrerade vid 180° och andra hälften vid 0° varvid man erhåller ett värde på *r* nära 0. Som nämnt ovan så benämns variabeln *r* ibland som längden av medelvektorn då den beskriver hur väl medelvektorn beskriver stickprovet genom att anta en enhetslös längd mellan 0 och 1. Man kan tolka änden av medelvektorn, alltså längden av *r*, i medelvinkelns riktning som mittpunkten för stickprovets tyngd. Om alla observationer har samma vikt och placeras i utkanten av en disk, enligt respektives angivna vinkel, kommer disken kunna balansera på positionen av änden av medelvektorn.

Eftersom *r* är ett mått av koncentration får man helt analogt ett mått av spridning genom följande ekvation, som är en definition av cirkulär varians:

$$\boldsymbol{S^2 = 1 - r} \quad (16)$$

där ett värde på S^2 nära 1 tyder på stor spridning och brist utav spridning beskrivs av ett värde runt 0. Ett annat mått på spridning är vinkelvarians och definieras enligt följande:

$$\boldsymbol{s^2 = 2(1 - r)} \quad (17)$$

Detta anses av vissa vara en bättre beskrivning av spridning och bättre motsvara ”vanlig” linjär varians. Den senare ekvationen kan anta värden mellan 0 och 2. En anmärkning är att, på liknande vis som koncentration, ett värde av $S^2 = 1$ eller $s^2 = 2$ leder nödvändigtvis inte till att man kan dra slutsatsen att stickprovet är likformigt fördelat trots att det är fullkomligt utspritt på den cirkulära skalan.

Ytterligare ett mått av spridning kan beräknas, denna gång med hjälp av naturlig logaritm:

$$s_0^2 = -2 \ln r \qquad (18)$$

Detta mått kan anta värden från 0 till ∞. Att det saknar en övre gräns skiljer det från de två andra spridningsmåtten. Eftersom det är lättare att tolka ett spridningsmått på ett begränsat intervall (till exempel från 0 till 2) kommer s^2 i fortsättningen att användes vid tillfällen där ett stickprovs spridning skall beräknas.

3. Medianvinkel

För att bestämma medianvinkeln på ett stickprov behöver man först hitta diametern som delar upp observationerna i två lika stora grupper. Medianvinkeln är den radie på diametern som är närmast majoriteten av observationerna. Om antalet observationer är udda kommer medianvinkeln oftast vara belägen vid en av datapunkterna eller mittemot (180°) en. Däremot om antalet observationer är jämt kommer medianvinkeln vara placerad halvvägs mellan två datapunkter, helt analogt med linjära data.

Det är möjligt, dock ovanligt, att det kan förekomma fler än en medianvinkel. Då bör man, enligt konvention, beräkna ett medelvärde av de befintliga medianvinklarna.

3.1 Testa för symmetri runt medianvinkeln

Symmetri kring medianvinkeln kan testas genom att använda Wilcoxons teckenrangtest. För varje observerad vinkel, a_i, beräknar vi differensen till medianvinkeln. Vi benämner differensen $d_i = a_i$ – median. Då kan vi förslagsvis ställa upp H_0: stickprovets fördelning är symmetrisk runt medianvinkeln mot H_1: ej symmetrisk. Därefter fortsätter man som ett vanligt Wilcoxon teckenrangtest med att ta fram dessa differenser (d_i) och rangordna absolutbeloppet av dem ($|d_i|$). Addera sedan ihop absolutbeloppen av de positiva respektive de negativa differenserna till två statistiska variabler, T_+ och T_-. Slutligen jämför man T_+ och T_- med kritiskt tabellvärde av T (som en funktion av α och n) för att se om man kan förkasta nollhypotesen.

4. Axiell data

Ibland stöter man på cirkulära data som är bimodal (”tvåvägsdata”), alltså data som är uppdelad i två grupper, ofta motsatta riktningar. Ett exempel på detta hittar vi inom biologin,

närmare bestämt i limnologin och ett experiment med vandrande fisk. Om man släpper fri fisk i en flod i sydöstlig – nordvästlig riktning kan det vara utav intresse att undersöka ifall fisken vandrar till grundare vatten, i ena riktningen, eller djupare vatten, andra riktningen. Då får man anpassa beräkning av medelvinkeln ty om man skulle använda tidigare nämnda formel får man en skev bild av stickprovsfördelningen. Det man istället gör är att dubbla alla observationers vinklar (så att a_i, där i= 1,2,...,n, blir $2a_i$) och beräknar de modulo 360. Om man dubblar en vinkel a_i >180° resulterar det i att man subtraherar 360 från den dubblade vinkeln. Exempelvis vinkeln 190° blir, efter dubblering och modulo 360, 20°. Därefter beräknar man medelvinkeln enligt konvention förutom det att man avslutningsvis dividerar vinkeln med två. Detta eftersom man egentligen får fram $2\bar{a}$. Medelvinkeln man nu fått fram beskriver ej stickprovet väl men en linje från den beräknade $\bar{a}$ till $\bar{a}$ + 180° kommer generera en cirkeldiameter som löper mellan de två datagrupperna och är den eftersökta axeln av den bimodala datan.

4.1 Medelvärdet av medelvinklar för axiella data

Om man beräknar en medelvinkel för varje grupp av data i en bi- eller multimodal distribution kan det vara utav intresse att ta fram ett medelvärde för denna uppsättning medelvinklar, en så kallad huvudmedelvinkel. Dock kan man inte betrakta varje grupps medelvinkel som en observationsvinkel och sedan fortsätta att beräkna en huvudmedelvinkel med den vanliga metoden. Då skulle man anta att varje medelvinkel har ett r-värde på 1,0 vilket är högst osannolikt. Det man bör göra är att ta fram huvudmedelvinkelns rektangulära koordinater enligt följande:

$$\bar{X} = \frac{\sum_{j=1}^{k} X_j}{k} \quad , \quad \bar{Y} = \frac{\sum_{j=1}^{k} Y_j}{k} \qquad \textbf{(19), (20)}$$

Med k stycken grupper av data och X_j respektive Y_j erhålles som tidigare. När man fått fram $\bar{X}$ och $\bar{Y}$ kan man beräkna huvudmedelvinkel med den vanliga formeln. Om man skulle sakna värden på X och Y för varje grupp men istället har $\bar{a}$ och r så kan man använda:

$$\bar{X} = \frac{\sum_{j=1}^{k} r_j \cos \bar{a_j}}{k} \quad , \quad \bar{Y} = \frac{\sum_{j=1}^{k} r_j \sin \bar{a_j}}{k} \qquad \textbf{(21), (22)}$$

När man ska beräkna huvudmedelvinkeln med denna metod är det rekommenderat att alla grupper har lika många observationer, $n_1 = n_2 = ... = n_j$, fastän olika storlekar på gruppernas stickprov ej påverkar resultatet allvarligt.

Denna metod att dubblera (eller tripplera etc.) vinkeln är lämplig att använda generellt vid statistiska tester och annan statistik involverande bi- eller multimodala data.

5. Testa för cirkulär likformighet

5.1 Rayleighs test

Ju högre r-värde man får desto bättre beskriver medelvinkeln stickprovet, ekvivalent gäller för s-värde (spridning) fast där ger lägre värde bättre beskrivande $\bar{a}$. Ett lämpligt test att genomföra för att avgöra om ens stickprov är likformigt fördelat, alltså saknar medelvinkel, är Rayleightestet. Då ställer vi upp hypoteserna H_0: Populationen är likformigt fördelad runt en cirkel mot H_1: Populationen är ej likformigt fördelad. Testet centreras runt hur stort r-värdet måste vara för att säkerställa en icke likformig distribution. Detta utförs med hjälp av det så kallade Rayleighs *R*, som nämnts tidigare får man det av produkten av antal observationer och r-värdet ($R = nr$). Rayleighs *R* kan sedan nyttjas för att räkna ut Rayleighs *z*:

$$\boldsymbol{z = \frac{R^2}{n} = nr^2} \qquad \textbf{(23)}$$

Därefter jämför man resultatet med kritiskt värde av $z_{\alpha,\, n}$, där α är konfidensgrad och *n* antal observationer, från tabell för att avgöra om det är signifikant. För att få fram ett p-värde på Rayleighs *R* kan man använda:

$$\boldsymbol{P = e^{\left(\sqrt{1+4n+4(n^2-R^2)}-(1+2n)\right)}} \qquad \textbf{(24)}$$

När man utför Rayleightestet antar man att den underliggande fördelningen är von Mises, även kallat cirkulär normalfördelning och som det låter är det analogt med linjär normalfördelning. Om testet resulterar i att vi förkastar H_0 innebär det att det finns en medelvinkel och om vi inte förkastar H_0 kan vi dra slutsatsen att stickprovet har likformig fördelning runt cirkeln. Det sistnämnda gäller dock endast om vi kan anta att stickprovet bara har en grupp med data (alltså unimodal).

5.2 V-testet, ett modifierat Rayleigh test

Ett modifierat Rayleightest, även kallat v-test, är helt enkelt ett vanligt Rayleigh test med enda skillnaden att man har en specifik medelvinkel som mothypotes. Lämpligt tillfälle att använda sig av v-testet är, ännu ett exempel från biologin, om man ska undersöka vart honungsbin skulle flyga om de blev frisläppta norr om sin bikupa. Det naturliga antagandet är då att bina ställer in siktet på deras hem och flyger rakt söder ut (180°). Då skulle man ställa upp följande hypoteser, H_0: Populationens riktning är likformigt fördelat runt cirkeln, mot H_1: Populationens riktning är ej likformigt fördelat och medelvinkeln är 180°. Eftersom vi gissar

en medelvinkel, och därmed adderar mer information, är v-testet något kraftfullare än Rayleighs test. När man sedan skall räkna på det använder man:

$$\boldsymbol{V = R\cos(\bar{a} - \bar{a}_0)} \quad \textbf{(25)}$$

där $\bar{a}_0$ är den förslagna medelvinkeln. Signifikansen för variabeln v erhålls från:

$$\boldsymbol{u = V\sqrt{\frac{2}{n}}} \quad \textbf{(26)}$$

Detta jämförs med kritiskt tabellvärde på $u_{\alpha,\,n}$.

5.3 Ett-stickprovstest för medelvinkeln

Om man är ute efter att testa ifall ett stickprovs medelvinkel ($\bar{a}$) är lika med ett givet värde bör man göra ett test som är analogt med ett ”one-sample t test”. Då ställer man upp H_0: $\bar{a} = \bar{a}_0$ mot H_1: $\bar{a} \neq \bar{a}_0$. Sedan undersöker man ifall $\bar{a}_0$ ligger inom ett konfidensintervall för $\bar{a}$. Ligger det utanför förkastar man H_0.

5.4 Hodges-Ajne testet för likformighet

Som ett alternativ till Rayleightestet finns det så kallade Hodges-Ajnetestet, vilket ej antar någon specifik fördelning för stickprovet. Det fungerar bra för såväl unimodala som bimodala samt multimodala fördelningar. Om den underliggande fördelningen är von Mises (cirkulär normalfördelning), som är förutsättningen för att göra Rayleightestet, är också Rayleightestet det starkare av de två.

Givet ett stickprov med cirkulära data dras en linje genom centrum (en diameter) så att differensen mellan antal observationer på båda sidorna av diametern blir så stor som möjligt. På ena sidan har vi så många observationer som möjligt medan på andra sidan har vi så få som möjligt. Just det antalet, det lägsta, blir viktigt sedan när vi skall göra beräkningar så vi kallar det antalet *m*. P-värdet för ett *m* minst så litet som det observerade, under nollhypotesen att stickprovet är cirkulärt likformigt, är:

$$\boldsymbol{P = \frac{(n-2m)\binom{n}{m}}{2^{n-1}} = \frac{(n-2m)\frac{n!}{m!(n-m)!}}{2^{n-1}}} \quad \textbf{(27)}$$

För $n > 50$ kan man göra följande approximation:

$$\boldsymbol{P \approx \frac{\sqrt{2\pi}}{A} e^{\frac{-\pi^2}{8A^2}}} \quad \textbf{(28)}$$

där

$$\boldsymbol{A = \frac{\pi\sqrt{n}}{2(n-2m)}} \quad \textbf{(29)}$$

Man kan även direkt jämföra ens observerade m med ett tabellvärde som ger kritiska värden på m som funktion av α och n. Detta gäller för $n \geq 9$.

5.5 Batschelettestet, ett modifierat Hodges-Ajne test

På samma sätt som det finns ett modifierat Rayleighs test finns där även ett modifierat Hodges-Ajne test. Det så kallade Batschelet testet fungerar på liknande sätt som v-testet, att man ställer upp en nollhypotes med föreslagen medelvinkel. Därefter räknar vi antalet observationer som ligger inom ± 90° från den föreslagna medelvinkeln, vi benämner denna variabel $m´$:

$$\boldsymbol{C = n - m´} \quad \textbf{(30)}$$

Där värdet på det observerade C är det vi sedan jämför med kritiskt tabellvärde, där C är en funktion av α och n.

6. Goodness of fit test för cirkulära data

6.1 Chi två test

Chi två används för att se hur väl en teoretisk cirkulär fördelning stämmer överens med en observerad. Tillvägagångssättet är, som för ett vanligt chi två test, att bestämma förväntad frekvens för varje observerad. Detta görs genom att dela in det observerade materialet i grupper, till exempel 0°-30°, 30°-60°, 60°-90° etc., därefter beräkna förväntad frekvens för varje grupp. Enligt konvention bör observationerna grupperas så att ingen förväntad frekvens understiger fyra. Gruppernas intervall behöver inte vara lika men om de är det (exempelvis 30° som ovan) råds följande kriterium vara uppnått, $n/k \geq 2$, där n är antal observationer och k är antal grupper. För att slutföra sitt chi två test ska man beräkna testvariabeln χ^2 enligt följande:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - \hat{f}_i)^2}{k} \quad (31)$$

Där f_i är observerad frekvens och $\hat{f}_i$ är förväntad frekvens. Slutligen jämför man sitt beräknade χ^2 med kritiskt värde $\chi^2_{\alpha,k-1}$ från tabell, där α är konfidensgrad och k är antal grupper.

Om man skulle använda sig utav icke-grupperade data bör man, istället för chi två, används antingen Kuipertestet eller Watsons U^2-test för ett stickprov.

6.2 Watsons U^2 test för ett stickprov

Då Watsontestet och Kuipertestet är av likvärdig styrka kommer endast det förstnämnda testet att redovisas.

Det första man gör är att omvandla sina observerade vinklar (a_i) genom att dividera respektive vinkel med 360°.

$$u_i = \frac{a_i}{360} \quad (32)$$

Sedan beräknar man testvariabeln Watsons U^2:

$$U^2 = \sum_{i=1}^{n} u_i^2 - \frac{(\sum_{i=1}^{n} u_i)^2}{n} - \frac{2}{n}\sum_{i=1}^{n} i u_i + (n+1)\bar{u} + \frac{n}{12} \quad (33)$$

Tills sist jämför man sitt U^2 med kritiskt värde $U^2_{\alpha,n}$ från tabell, där α är konfidensgrad och n är antal observationer.

7. Problemlösning

Nedan kommer några av metoderna att demonstreras genom ett exempel. När en ekvation från arbetet används kommer det att finnas en hänvisning till höger om uträkningen. Detta exempel är baserat på data från Transportstyrelsen.

7.1 Exempel: Svårt skadade i trafikolyckor åren 2006-2012

Månad	a_i	f_i	sin a_i	f_i sin a_i	cos a_i	f_i cos a_i
Jan	0°	1522	0	0	1	1 522
Feb	30°	1435	0,5	717,5	0,866	1 242,7
Mar	60°	1505	0,866	1 303,4	0.5	752,5
Apr	90°	1824	1	1 824	0	0
Maj	120°	2209	0,866	1 913,1	-0,5	-1 104,5
Jun	150°	2722	0,5	1 361	-0,866	-2 357,3
Jul	180°	2564	0	0	-1	-2 564
Aug	210°	2346	-0,5	-1 173	-0,866	-2 031,7
Sep	240°	2178	-0,866	-1 886,2	-0,5	-1 089
Okt	270°	1965	-1	-1 965	0	0
Nov	300°	1813	-0,866	-1 570,1	0,5	906,5
Dec	330°	1807	-0,5	-903,5	0,866	1 564,9

Beräkning av medelvinkel:

$$n = 23\,890$$

Eftersom det är grupperade data måste vi multiplicera med frekvensen f_i när vi beräknar de rektangulära koordinaterna:

$$\sum f_i \sin a_i = -378{,}889$$

$$\sum f_i \cos a_i = -3\,157{,}86$$

$$\boldsymbol{Y = \frac{\sum f_i \sin a_i}{n} = -0{,}01586} \qquad (4)$$

$$\boldsymbol{X = \frac{\sum f_i \cos a_i}{n} = -0{,}13218} \qquad (3)$$

$$\boldsymbol{r = \sqrt{X^2 + Y^2} \approx 0{,}1331} \qquad (5)$$

Beräknar även det korrigerade r-värdet ty grupperade data:

$$\boldsymbol{r_c = cr = \frac{\frac{30*\pi}{360}}{sin\left(\frac{30}{2}\right)} \approx 0,1345} \quad (6)$$

Det korrigerade r-värdet skiljer sig ej mycket från det ursprungliga ty intervallen på 30° är ej stort nog för att påverka avsevärt.

$$\boldsymbol{sin\,\bar{a} = \frac{Y}{r} = \frac{-0,01586}{0,1331} = -0,11913} \quad (9)$$

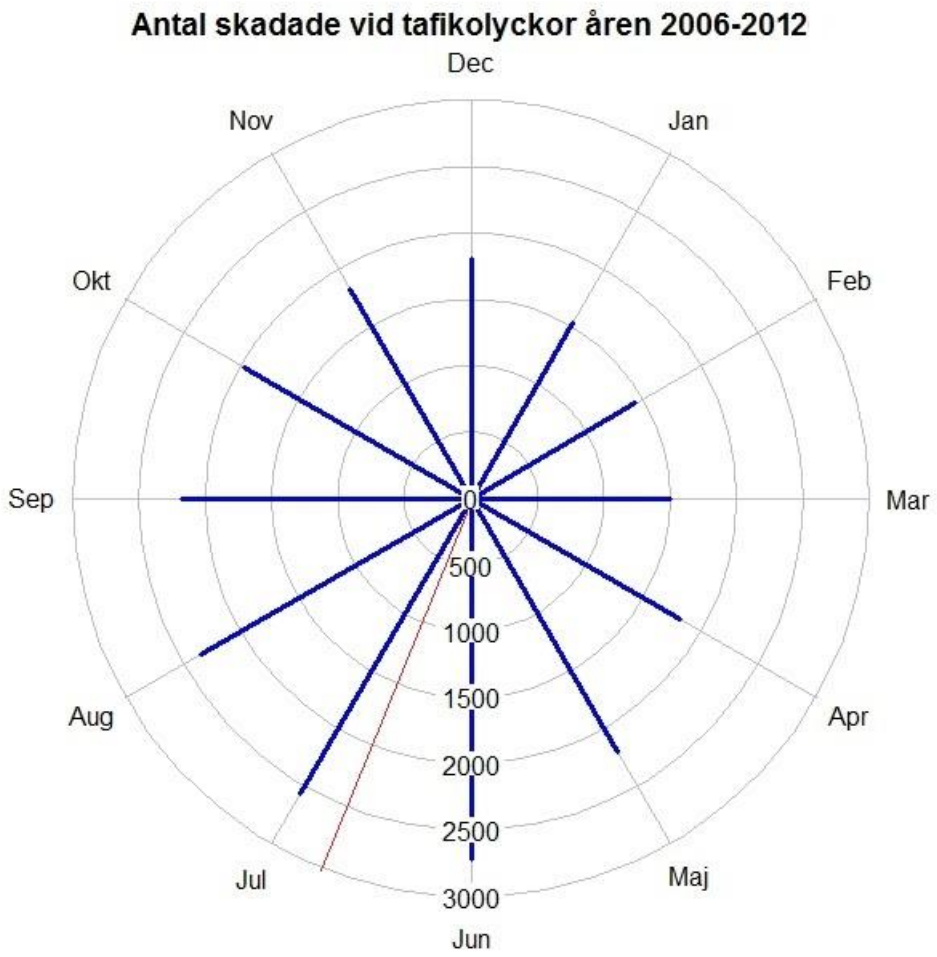
$$\boldsymbol{cos\,\bar{a} = \frac{X}{r} = \frac{-0,13218}{0,1331} = -0,99288} \quad (8)$$

Detta ger oss följande medelvinkel:

$$\rightarrow \bar{a} \approx 173°$$

Vi beräknar även spridningen:

$$\boldsymbol{s^2 = 2(1 - r) = 1,733738} \quad (17)$$



Figur 1: Schematisk bild över skador i trafiken år 2006-2012. Den röda linjen indikerar medelvinkeln. Bilden är gjord i R med paketet plotrix.

Rayleighs test:

H_0: Svårt skadade i trafiken är likformigt fördelat runt cirkeln (året).

H_1: $\neg H_0$.

$$\boldsymbol{z = nr^2 = 23\,890 * 0,1331^2 = 423,23} \quad \textbf{(23)}$$

Jämför sedan med tabellvärdet $z_{0,05,23\,890} = 2,9957$, vi kan förkasta H_0 på nivån 5 %. Eftersom vi har ett väldigt stort *n* får vi ett oerhört litet p-värde, $P < 0,0001$

Chi två test:

H_0: Svårt skadade i trafiken är likformigt fördelat runt cirkeln (året).

H_1: ¬H_0.

$k = 12$

Ta fram förväntade värden:

$$\hat{f}_i = \frac{n}{k} = \frac{23\,890}{12} \approx 1\,991$$

$$\boldsymbol{\chi^2 = \sum_{i=1}^{k} \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i} = \frac{(1\,522 - 1\,991)^2}{1991} + \cdots + \frac{(1\,807 - 1\,991)^2}{1991} \approx 969,7} \quad \textbf{(31)}$$

$$\chi^2_{0,05,11} = 19{,}675$$

Förkasta H_0 på nivån 5 %. Svårt skadade i trafiken är ej likformigt fördelat runt året.

8. Referenser

Böcker

Jerold H. Zar. *Biostatistical Analysis.* 5 uppl. Pearson Education, Inc. 2010.

N. I. Fisher. *Statistical analysis of circular data.* Cambridge University Press. 1993.

Webbsidor

Dödade och svårt skadade efter län, månad och år. (senast uppdaterad 2016-02-15) Transportstyrelsen. https://www.transportstyrelsen.se/sv/vagtrafik/statistik-och-register/Vag/Olycksstatistik/Polisrapporterad-statistik/Nationell-statistik/Manadsstatistik/ (Hämtad 2016-05-13)