Projektuppgift
Matematisk statistik för K, TMA074

Hur mycket fisk från Västra Götaland kan vi äta?

Bakgrund och målsättning

Projektet går ut på att studera halten av kvicksilver i fisk fångad i Västra Götaland, och se hur den förhåller sig till de gränsvärden som är bestämda för EU och Sverige. Uppgiften är delvis baserad på Miniprojekt 9.8 från boken Räkna Med Variation som används i kursen. Syftet med projektuppgiften är bland annat att ni ska träna på att

- använda Matlab för att tillämpa statistiska metoder på ett riktigt problem.
- konstruera och analysera statistiska modeller samt utföra kritisk granskning av modellernas förmåga att beskriva verkligheten.
- skriftligt redovisa modeller, antaganden och slutsatser från statistiska analyser.

Arbetet kommer genomföras i grupper om två till tre studenter. Projektet är indelat i fyra delar som motsvarar varsin datorövning. Arbetet kommer huvudsakligen utföras under datorövningarna men troligen kommer mer tid behövas för att slutföra och rapportera arbetet.

Projektet redovisas i form av en skriftlig rapport som lämnas in via PingPong senast **2018-10-17 klockan 23:59**. Lämna in rapporten som en PDF och bifoga koden som ni använde för att utföra beräkningarna. Organisera koden så att en fil utför analysen för varje del, låt `projektDel1.m` utföra analysen för del 1, `projektDel2.m` analysen för del 2 och så vidare. Återlämning och eventuell korrigering/komplettering görs vid datorövningen under den sista läsveckan. Vid större kompletteringar lämnas en korrigerad rapport in via PingPong senast **2018-10-31 klockan 23:59** I Appendix finns mer tips och instruktioner kring arbetet och rapporteringen av projektet.

Regler och gränsvärden för kvicksilver

Kvicksilver finns naturligt i miljön, men mängderna i naturen har ökat på grund av mänsklig påverkan. Detta gör till exempel att ohälsosamma mängder kvicksilver kan finnas i stora rovfiskar som gädda och abborre. Att få i sig kvicksilver kan vara skadligt för hjärnan och Livsmedelsverket rekommenderar därför att abborre, gädda, gös eller lake som du fiskat själv inte ska ätas oftare än en gång per vecka. Kvicksilver kan föras över till barnet hos gravida genom mammans moderkaka och bröstmjölk, och foster vars hjärna och nervsystem fortfarande är i utveckling är speciellt känsliga för kvicksilver. Livsmedelsverket rekommenderar därför också att kvinnor som är gravida, ammar eller planerar att skaffa barn, bör inte äta fisk som kan innehålla kvicksilver oftare än 2-3 gånger per år.

Den Europeiska myndigheten för livsmedelssäkerhet, EFSA, har tagit fram ett gränsvärde på $1.3 \cdot 10^{-3}$mg per kilo kroppsvikt som högsta veckointag av kvicksilver. Vidare får fisk som säljs inom EU inte innehålla mer än 0.5mg kvicksilver per kilo. Dock finns vissa undantag från denna regel, som vid försäljning av gädda i Sverige där gränsvärdet istället är 1.0mg/kg.

Datamaterialet

Svenska Miljöinstitutet har en stor databas med mätningar av bland annat kvicksilver, och mätningarna ni kommer arbeta med är hämtade från denna. Varje grupp får en bit av det fullständiga

datamaterialet att analysera. Datan ni ska analysera finns tillgänglig i filen `datagrxx` där xx är ert gruppnummer (grupp 4 hämtar alltså sin data från `datagr04`). Filerna finns tillgängliga på kurshemsidan och gruppens nummer anges i PingPong.

I filen finns en tabell där varje rad är en mätning. För varje mätning finns följande information:

Station	Var mätningen gjordes.
Datum	Vilket datum mätningen gjordes.
Art	Vilken art mätningen gjordes på.
Vikt	Vikten av fisken som mätningen gjordes på. Enhet gram.
HG	Vikten kvicksilver per kilo fisk. Enhet mg/kg.

Del 1: Undersökning av datan

Som alltid vid statistisk analys är det bra att börja med en översiktlig analys av datan för att se om det finns anledning till oro kring hur mätningarna utförts. När vi senare gör en fördjupad analys kommer vi till exempel anta att mätningarna är oberoende och likafördelade, vilket bör undersökas.

Vi börjar med att undersöka mätningarna hos gäddor. För att skapa en ny tabell med enbart mätningarna på gäddor kan följande kommando användas:

```
gadda = data(categorical(data.Art)=='Gädda',:);
```

- Plotta kvicksilvermätningarna som funktion av tiden när de gjordes. Tycks det finnas några trender eller andra konstigheter i datan? Om ni hittar några konstiga mätningar, undersök dessa i mer detalj (titta på vad de andra variablerna var för de mätningarna). Om ni hittar outliers fundera på om det kan vara värt att plocka bort dessa. Funktionen `datetime` är användbar för att konvertera datumen i tabellen till nummer som kan användas för plottning.
- Baserat på mätningarna, hur vanligt är det att kvicksilverhalten hos gäddor överskrider gränsvärdena 0.5 mg/kg och 1 mg/kg?
- Kan ni hitta någon statistisk modell som passar till datan? Kan en normalfördelning vara en rimlig modell, eller finns det någon annan fördelning som passar bättre? Funktionen `histfit` är bra att använda för att enkelt kunna utvärdera olika modeller.

Del 2: Modellanpassning och analys

- Använd funktionen `fitdist` för att skatta parametrarna i modellen ni valde i den första uppgiften. Förklara i rapporten hur funktionen tar fram skattningen och rapportera de skattade värdena.
- Vad är sannolikheten att överskrida de två gränsvärdena baserat på modellen? Hur skiljer sig dessa från de empiriska skattningarna? Finns det någon anledning att använda resultatet från modellen istället för att bara räkna andelen fiskar som har för hög halt?
- Verkar det säkert att sälja fisken i svenska affärer? Kan vi exportera den till resten av EU?
- Hur stor portion fisk kan en vuxen person som väger 67kg äta en gång i veckan, för att med 95% sannolikhet inte överskrida det rekommenderade högsta veckointaget av kvicksilver? Redovisa era beräkningar för hur ni tar fram portionsstorleken.
- Antag att personen äter en lika stor portion fisk varje vecka under ett år. Finns det någon lämplig (approximativ eller exakt) modell för årsintaget av kvicksilver baserat på modellen ni valde i första uppgiften? Hur stor portion kan personen äta varje vecka för att med 95% sannolikhet inte överskrida ett årsintag på $52 \cdot 1.3$mg kvicksilver per kilo kroppsvikt? Hur/varför skiljer sig svaret mot beräkningen som gjordes för en vecka?

Del 3: Jämförelse mellan abborre och gädda

Förutom mätningarna på gäddor har vi också mätningar på abborrar. Vi är nu intresserade av att undersöka om någon av de två arterna är säkrare att äta om man inte vill få i sig kvicksilver.

För att förenkla analysen, antag att den logaritmerade kvicksilverhalten hos både gäddor och abborrar är normalfördelad. Om ni inte gjort det tidigare, undersök gärna detta antagande med hjälp av `histfit`-funktionen.

Låt μ_a och μ_g beteckna väntevärdena för de två normalfördelningarna och låt σ_a^2 och σ_g^2 beteckna motsvarande varianser (a står får abborre och g för gädda). Börja med att testa om det är rimligt att anta att variansen hos de två arterna är samma:

- Utför hypotestestet

$$H_0 : \sigma_a = \sigma_g$$
$$H_1 : \sigma_a \neq \sigma_g$$

  med hjälp av ett F-test. Kan nollhypotesen förkastas?

Kontrollera era beräkningar genom att jämföra med resultatet som fås med `vartest2`. Testa nu om den förväntade halten hos de två arterna är samma:

- Utför hypotestestet

$$H_0 : \mu_a = \mu_g$$
$$H_1 : \mu_a \neq \mu_g$$

  Om ni inte kunde förkasta antagandet om lika varianser, utför testet med en poolad skattning av variansen. Utför annars testet när ni antar olika varianser. Kan nollhypotesen förkastas?

Kontrollera era beräkningar genom att jämföra med resultatet som fås med `ttest2`. Vad är era slutsatser? Om ni hittar en signifikant skillnad, kan ni komma på någon biologisk förklaring till denna?

Del 4: Har halten kvicksilver förändrats med tiden?

Om ni i Del 3 såg att det fanns skillnader mellan abborrar och gäddor skulle en förklaring till detta kunna vara att det verkligen finns biologiska skillnader kring hur gäddor och abborrar tar upp kvicksilver. En annan kunde vara att halten kvicksilver har förändrats över tiden, och att det finns skillnader i när den största delen gäddor och abborrar fångades. För att undersöka detta tittar vi nu i mer detalj på mätningarna av abborrar.

- Plotta mätningarna av kvicksilver hos abborrar mot tiden då de samlades in. Tycks det finnas någon trend i datan? Eftersom vi i Del 3 modellerade logaritmerade värden är det också här bäst att jobba med de logaritmerade halterna.
- Ansätt en regressionsmodell med tid som förklarande variabel och log-kvicksilver som responsvariabel, och skatta parametrarna. Funktionen `datenum` är användbar för att konvertera datumen i tabellen till tal som kan användas som förklarande variabel. Vad är parameterskattningarna?
- Undersök om regressionsmodellen passar bra genom att utföra lämplig residualanalys. Om en linjär regressionsmodell inte passar bra, verkar polynomregression fungera bättre?
- Verkar kvicksilverhalten ha förändras signifikant under tiden mätningar gjorts?

Som oftast vid analys av riktig data måste vi tänka på om det kan finnas några andra orsaksfaktorer som kan sammanblandas med regressionen. I detta fall har vi också samlat in vikterna hos fiskarna som mätningarna gjordes på, och ett rimligt antagande är att större fiskar har hunnit samla på sig mer kvicksilver.

- Anpassa en regressionsmodell då vi testar detta antagande. Skatta alltså en linjär regressionsmodell då fiskens vikt är förklarande variabel och log-kvicksilver är responsvariabel. Verkar det finnas ett signifikant samband?
- Rita ut mätningarna av abborrarnas vikt mot tiden då de samlades in. Tycks det finnas någon trend i datan? Kan det finnas fog för oron att vi har en sammanblandningseffekt för analysen vi gjorde kring om kvicksilverhalten har förändrats över tiden?
- Ansätt en multipel regressionsmodell för log-kvicksilverhalten med både tid och vikt som förklarande variabler. Skatta modellen och plotta det skattade regressionsplanet tillsammans med mätningarna.
- Är provdatumet forfarande en signifikant effekt i den multipla regressionsmodellen? Vad drar ni för slutsatser från den totala analysen som gjorts av datan?

Appendix - Riktlinjer för projektredovisning

I denna sektion följer anvisningar som är bra att tänka på när ni skriver rapporten för projektarbetet.

Ni ska med stöd av frågorna i projekthandledningen skriva en självständig rapport som täcker det väsentliga innehållet. Notera att det inte är tillåtet att sammarbeta med andra grupper varken under skrivandet av rapporten eller under arbetet i Matlab. Målgruppen för skriften i projektet ska vara en teknolog i samma årskurs som läst kursen men som inte är insatt i detaljerna kring projektet. Språket i rapporten ska vara anpassat för målgruppen och texten ska vara tillräckligt omfattande och tydlig så att en person i målgruppen utan större ansträngning skall kunna följa resonemang och motiveringar. Speciellt måste texten vara korrekturläst så att språk- och skrivfel är rättade.

Rapporten ska innehålla klara och tydliga formuleringar av frågeställningarna, modeller och antaganden. Texten ska vara välstrukturerad med tydliga avsnittsrubriker och ska kunna läsas utan tillgång till vare sig kod som använts för beräkningar eller denna projektbeskrivning. Figurer och tabeller, försedda med figurtexter och tydlig numrering, ska användas för att redovisa resultat när det är lämpligt. Rapporten skall innehålla titelsida, sammanfattning och referenslista. Källhänvisningar ska vara enligt gängse norm men hänvisningar till kursboken och föreläsningsanteckningar får göras i förenklad form.

Bedömning

Rapporten kommer i huvudsak bedömas enligt riktlinjerna ovan samt följande kriterier:

- att lämpligt och tydligt statistiskt språk används;
- att lämpliga tabeller, figurer och sammanfattande mått används för att beskriva datan och resultaten;
- att de statistiska beräkningarna är rätt utförda och verkar rimliga;
- att korrekta slutsatser dras med kommentarer om eventuella brister i analysen.

Nedan finns en checklista för granskning av rapporter. För att undvika onödiga misstag, gå igenom och bocka av listan med frågor innan rapporten lämnas in och se till att alla krav är uppfyllda.

Checklista för rapportering av projektet

		Ja	Nej
1.	Har alla moment i uppgiften blivit slutförda?	☐	☐
2.	Innehåller rapporten relevanta figurer och tabeller?	☐	☐
3.	Introduceras och förklaras all notation som används?	☐	☐
4.	Introduceras alla modeller som används?	☐	☐
5.	Presenteras och diskuteras resultaten tillräckligt?	☐	☐
6.	Är alla beräkningar kontrollräknade?	☐	☐
7.	Har rimligheten bedömts hos resultaten och har eventuella orimligheter kontrollerats och kommenterats?	☐	☐
8.	Har rapporten:		
	• Titel, författare och datum?	☐	☐
	• Inledning?	☐	☐
	• Resultat och slutsatser?	☐	☐
9.	Har rapporten korrekturlästs? Är stavningsfel rättade?	☐	☐
10.	Angående figurer och tabeller:		
	• Är de numrerade?	☐	☐
	• Har de lämpliga figurtexter?	☐	☐
	• Är de refererade i texten?	☐	☐
11.	Är texten välstrukturerad och indelad i stycken med lämpliga rubriker?	☐	☐
12.	Går det att läsa och förstå rapporten utan:		
	• tillgång till projektbeskrivningen?	☐	☐
	• tillgång till Matlabkod som använts?	☐	☐
13.	Om kursboken eller föreläsningarna refereras i texten, är referenserna tillräckligt specifika (anges kapitel/nummer på föreläsning)?	☐	☐