Datorlaboration 4

Statistik och dataanalys III, 15 hp

Mona sfaxi

I. Introduktion

I den här datorlabben kommer vi att använda oss av parametriska- och icke-parametriska metoder för att beräkna intervallskattningar såväl som för att göra hypotestest. Avslutningsvis kommer vi att ägna oss åt Bayesiansk inferens. Den första delen av datorlabben kommer förmodligen kännas väldigt bekant med innehållet från statistik 1 och kan ses som repetition. Medan de andra två delarna troligtvis kommer vara obekanta sedan tidigare.

II. Installera paket

Paketet `BSDA` kommer användas i den här datorlabben.

- ☐ Skapa en code-chunk där du laddar ner alla paket som du vill använda dig av. Skriv `#| output: false` längst upp i din code-chunk för att filtera bort störande meddelanden som dyker upp då man laddar paket. Dessa bör inte vara med i ditt färdiga dokument.
- ☐ På Athena finns några dataset som vi kommer använda oss av (`Salary_total`, `tacos`, `Salary`, `wc_women`, `wc_men` och `MWH`). Börja med att ladda ner dem och spara dem i din arbetsmapp, helst i samma mapp som din Quarto fil.
- ☐ Övningar som prioriteras i den här labben markeras med en *.

1. Parametriska metoder

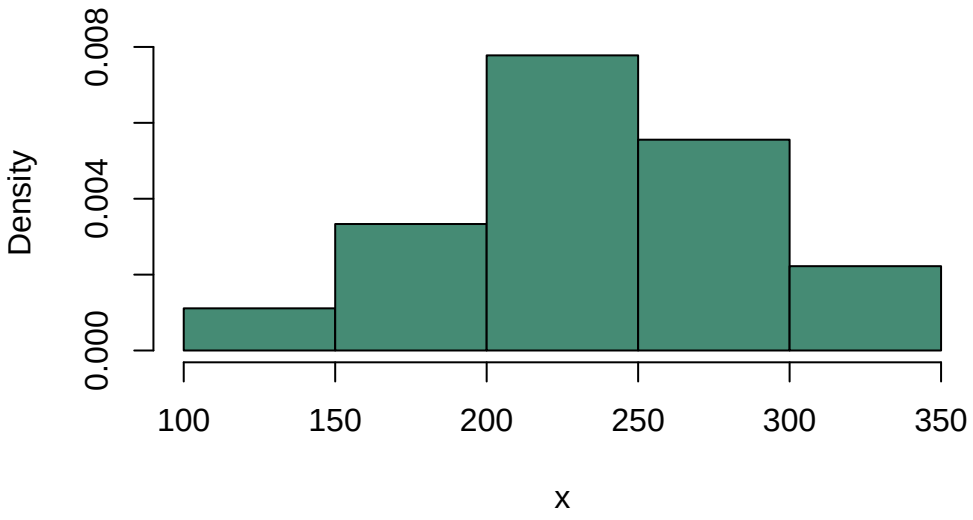
Du har hittills lyckats ducka t-fördelningen ganska rejält under datorlabbarna, förutom kanske den lilla avstickaren vid datorlabb 2. Men som du såg då så är det en väldigt trevlig fördelning och den används flitigt vid parametriska hypotestester och vis konstruktion av konfidensintervall som vi nu kommer syssla med.

I R finns inget "Z-test" istället används t-test för att testa hypoteser av numeriska variabler, oavsett hur stort urval man har. Innan vi börjar med uppgifterna kan vi först titta på ett litet fiktivt exempel på hur ett sånt här test skulle kunna se ut.

1.1 Ett fiktivt exempel för en numerisk variabel

Ett gamingcenter samlar in olika data om sina kunder. Något som är av intresse är hur länge varje användare uppehåller sig på centret. Låt X beteckan antal minuter en kund spenderar på centret. Nedan ser vi data på antal minuter som spenderats på centret bland 18 slumpmässigt utvalda användare. Antal minuter antas vara normalfördelat. Vi antar även att det är oberoende mellan individerna i stickprovet.

```
x <- c(213, 124, 192, 253, 216, 251, 243, 196, 268, 238, 229, 309, 181, 211,
       310, 256, 290, 214)
hist(x, col = "aquamarine4", main = "", freq = FALSE)
```



Under tidigare år har den genomsnittliga tiden legat på ca 205 minuter. Vi vill nu testa ifall den genomsnittliga tiden har förändrats. Vi låter μ beteckna väntevärdet för antal minuter som kunder spenderar på centret. Vi vet ej vad den sanna variansen är men eftersom fördelningen antas vara normalfördelad kan vi använda ett t-test. Vi börjar med att ställa upp våra hypoteser.

H_0: $\mu = 205$

H_A: $\mu \neq 205$

Detta är alltså ett dubbelsidigt test. Vår testvariabel ges av:

$$T = \frac{\overline{x} - \mu_o}{s/\sqrt{n}}$$

där s är stickprovsstandardavvikelsen och n är antal observationer i stickprovet. Vi vet att $T \sim t(n-1)$ under H_0.

Kritiskt värde ges av $t_{\alpha/2,n-1}$ och kan räknas ut i R med hjälp av funktionen `qt(p, df)` som ger oss kvantilvärdet på t-fördelningen givet ett visst percentilvärde och frihetsgrad. Låt oss använda en signifikansnivå på 5%, men här måste vi tänka på att det är ett dubbelsidigt test:

```
alpha <- 0.05
n <- length(x)
qt(p = 1-alpha/2, df = n - 1)
```

```
[1] 2.109816
```

Vår beslutsregel är alltså att förkasta H_0 ifall $|T_{obs}| > t_{crit}$, dvs ifall $|T_{obs}| > 2.11$. Låt oss utföra testet i R med hjälp av funktionen `t.test(x, mu, conf.int, alternative)`

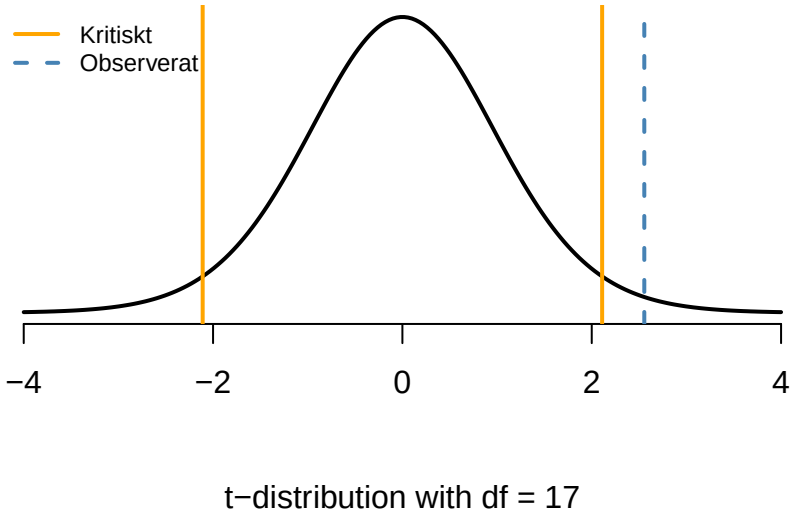
```
t_res <- t.test(x = x, mu = 205, conf.int = 0.95, alternative = "two.sided")
t_res
```

```
    One Sample t-test

data:  x
t = 2.5553, df = 17, p-value = 0.02049
alternative hypothesis: true mean is not equal to 205
95 percent confidence interval:
 209.8812 256.1188
sample estimates:
mean of x
      233
```

Här kan vi utläsa värdet på vår testvariabel som blev 2.55, och vi ser redan nu att det är större än det kritiska värdet och kan förkasta H_0. Men utskriften ger oss mer information, såsom antal frihetsgrader, p-värdet (sannolikheten att förkasta H_0 givet att H_0 är sann) samt ett konfidensintervall för μ och slutligen vad medelvärdet var i vårt stickprov.

Vi skulle även kunna illustrera t-fördelningen under H_0 tillsammans med det kritiska området och det observerade värdet (T_{obs}) med hjälp av vertikala linjer i en och samma graf. Men innan vi gör dett måste vi först skapa en grid av värden på x-axeln och sedan använda dessa värden i funktionen `dt()` för att få fram vilka funktionevärden i t-fördelningen som de motsvarar.

```
xGrid <- seq(-4, 4, length = 200)
fx <- dt(xGrid, df = n-1)
t_crit_low <- qt(p = alpha/2, df = n-1)
t_crit_high <- qt(p = 1-alpha/2, df = n-1)
t_obs <- t_res$statistic # extrahera värdet på det observerade t-värdet

plot(xGrid, fx, type = "l", xlab = "", bty = "n", yaxt = "n", ylab = "", lwd=2,
     sub = "t-distribution with df = 17")
abline(v = c(t_crit_low, t_crit_high), col = "orange", lwd = 2)
abline(v = t_obs, lwd = 2, col = "steelblue", lty = 2)
legend("topleft", legend = c("Kritiskt", "Observerat"), lwd = c(2,2),
       col = c("orange", "steelblue"), lty = c(1,2), bty = "n", cex = 0.75)
```



Varför finns det två orangea linjer i grafen och vad kan vi dra för slutsats av att titta på den?

(Argumentet bty = "n" tar bort "boxen" kring plotten och yaxt = "n" tar bort hela y-axeln, men var försiktiga med att använda dessa argument i en plot! oftast ska man aldrig ta bort y-axeln eftersom man då missar viktig information!)

1.2 Hypotestest och konfidensintervall för en variabel

Den nominella, genomsnittliga månadslönen år 2021 för samtliga yrken i Sverige var 37 100 kr. Ett slumpmässigt urval av löner år 2022 inom 41 yrken har valts ut och finns i filen `Salary_total`. Du ska nu jobba med det här datasetet och testa hypotesen ifall den genomsnittliga lönen har stigit år 2022.

Uppgift 1.1* Eftersom det här är lönedata är det lämpligt att använda sig av logaritmer. Börja med att skapa en ny variabel i ditt dataset; `ln_salary` som består av den naturliga logaritmen av `avg_salary`.

Uppgift 1.2* Testa ifall den logaritmerade månadslönen är större år 2022 jämfört med 2021 på en 5%-ig signifikansnivå. Börja med att ställa upp dina hypoteser, antal frihetsgrader och kritiskt värde för förkastelseområdet samt din teststatistika under H_0. Använd gärna `qt(sannolikhet, df = antal_frihetsgrader)` för att finna det kritiska värdet i R.

Uppgift 1.3* Använd funktionen `t.test(x = min_variabel, mu, alternative)` i R för att göra beräkningen, där `mu` syftar till vilken typ av hypotes du har om μ under H_0 (default är 0). Argumentet `alternative` anger din alternativhypotes och kan ta värdena "less", "two.sided" eller "greater". Vad kan du dra för slutsats? Försök att förklara med egna ord vad det innebär.

Uppgift 1.4 Rita upp den teoretiska t-fördelningen under H_0, utifrån antal frihetsgrader som du fann ovan, genom att först skapa en "grid" av *lämpliga* x-värden med hjälp av funktionen `seq(min, max, length)`. Kalla din vektor för något passande såsom exempelvis `xGrid`. Använd sedan `xGrid` i täthetsfunktionen för t-fördelningen; `dt(min_x_värden, df)` för att få fram teoretiska funktionsvärden från fördelningen.

Uppgift 1.5 Lägg till en vertikal linje i din graf som illustrerar kritiskt värde för förkastelseområdet och även en vertikal linje i en annan färg som illustrerar värdet på din teststatistika. (Tips: använd funktionen `abline(v = min_x_koordinat)`). Lägg gärna till en `legend` med etiketter så att man kan se vad varje linje representerar. Vad visar grafen? Beskriv med egna ord.

Uppgift 1.6* Vilket konfidensintervall (KI) för den genomsnittliga logaritmerade lönen μ skulle vara bredast och varför, ett 90%-igt eller ett 95%-igt? Fundera först över svaret och skapa därefter ett 90%-igt och ett 95%-igt KI för μ. Tips: använd dig av funktionen `t.test()` men skriv `alternative = "two.sided"` och ange konfidensgrad med hjälp av argumentet `conf.level`. Glöm inte att tolka intervallet!

1.3 Test av proportion i R med ett fiktivt exempel

Ovan tittade vi på hur man kunde utföra ett hypotestest för väntevärdet hos en numerisk variabel. I det här avsnittet ska vi istället titta på hur man kan utföra ett hypotestest för en proportion.

Funktionen `prop.test(x = antal_ettor, n = antal_obs, p = p0)` liknar funktionen `t.test()` och kan användas för att utföra ett hypotestest för en proportion. Men lägg märke till att variabeln man testar är kategorisk med två olika utfall, en så kallad dummy-variabel. Det första argumentet i funktionen bör vara antal "lyckade", dvs antal 1:or. Det andra argumentet i funktionen är `n` som representerar totalt antal observationer hos variabeln och sedan `p`, där `p` alltså är proportionen under nollhypotesen. Skriver man inte ut något värde här så kommer R anta att proportionen under nollhypotesen $= 0.5$. Annars så har funktionen samma argument som `t.test()`, dvs man kan sätta konfidensgrad och skriva ifall man vill ha ett enkelsidigt eller dubbelsidigt test osv.

Låt exempelvis säga att man vill undersöka ifall andelen som bär på ett paraply i Sverige när det regnar ute är större än 0.5. Man har sedan gjort mätningar på slumpmässigt utvalda personer under en regnig dag och alla som bar på ett paraply blev kodade som 1, och de som inte gjorde det blev kodade som 0. Man konstaterade i sina mätningar att $\frac{27}{50}$ av alla personer i stickprovet bar på ett paraply. Eftersom det här är ett hypotestest av en proportion så används inte t-fördelningen. I boken och på föreläsningarna använder man en normalapproximation ifall urvalet är tillräckligt stort (och $n \times p > 10$ samt $n \times (1-p) > 10$) men i funktionen `prop.test()` i R används istället en testvariabel som är χ^2-fördelad. Reslutaten kommer skilja sig åt väldigt lite så det går fortfarnde bra att använda sig av detta test. Det viktiga är då hur man tolkar utskriften. Nu kan vi testa:

$H_0 : p = 0.5$

$H_A : p > 0.5$

på en 10%-ig signifikansnivå

```
prop.test(x = 27, n = 50, p = 0.5, alternative = "greater", conf.level = 0.9)
```

```
    1-sample proportions test with continuity correction

data:  27 out of 50, null probability 0.5
X-squared = 0.18, df = 1, p-value = 0.3357
alternative hypothesis: true p is greater than 0.5
90 percent confidence interval:
 0.4400347 1.0000000
sample estimates:
```

```
   p
0.54
```

Vi ser här att det inte går att förkasta H_0 på en 10%-ig signifikansnivå eftersom p-värdet är 0.3357. Det finns alltså inget stöd för att andelen personer som har ett paraply en regnig dag överstiger 0.5.

1.4 Fredagsmys av proportionell betydelse

Ett företag som säljer tex-mex produkter hävdar efter en undersökning att 55% av alla svenskar äter tacos på fredagar. En student som läser statistik på SU ville testa detta påstående och gjorde en egen undersökning av 100 personer där respondenterna fick frågan ifall de äter tacos på fredagar (åtminstone 2 veckor per månad). Datasetet `tacos` innehåller resultatet från undersökningen där svaret "Ja" blivit kodat som 1 och svaret "Nej" blivit kodat 0.

Uppgift 1.7* Ladda datasetet i R med hjälp av funktionen:

```
tacos <- read.delim("tacos.txt", sep = " ")
```

Uppgift 1.8* Testa ifall påståendet hos företaget stämmer på en 5%-ig signifikansnivå genom att ställa upp noll- och alternativhypotes och använd sedan funktionen `prop.test()`. Vad drar du för slutsats? Tips: använd funktionen `table(min_variabel)` för att räkna ut antal 1:or i datasetet.

1.5 Lönedata för två variabler

Två andra vanliga t-test är för parvisa och icke-parvisa observationer. Viktigt att tänka på är att vid dubbelsidiga tester spelar det inte så stor roll hur du ställer upp din teststatistika *om* du endast är intresserad av resultatet. Vid enkelsidiga tester däremot är det väldigt viktigt att du tänker på hur frågan är formulerad när du ställer upp din testvariabel.

För att utföra ett hypotestest med **parvisa observationer** kan man använda sig av funktionen `t.test()`, principen är lik den innan med en variabel och argument för konfidensnivå, μ under H_0 . Nedan ser vi i princip vilka argument som används

```
t.test(x = obs_urval1, y = obs_urval2, mu, alternative, paired = TRUE)
```

Alternativt kan man istället skriva:

```
t.test(numerisk_variabel ~ kategorisk_variabel, mu, alternative, paired = TRUE)
```

ifall man har arrangerat datasetet på så sätt att alla värden i båda kategorier finns i en enskild kolumn ("numerisk_variabel" i koden ovan) och en annan variabel som anger grupptillhörighet finns i en annan kolumn ("kategorisk_variabel" i koden ovan). Men var då medveten om att R automatiskt kommer att beräkna medelvärdet för den grupp som förekommer allra först i datasetet. I praktiken så innebär inte detta ett problem ifall du formulerar dina hypoteser korrekt, men du får mindre valfrihet då när det gäller själva kodning. *Observera även att det inte går att använda ett t.test för parvisa observationer ifall man har olika många observationer i de två kategorierna.*

När det kommer till **icke-parvisa observationer** så är det nästan samma kod som ovan, dvs

```
t.test(x = obs_urval1, y = obs_urval2, mu, alternative, paired = FALSE)
```

Alternativt kan man istället skriva

```
t.test(numerisk_variabel ~ kategorisk_variabel, mu, alternative, paired = FALSE)
```

Enda skillnaden är alltså att man skriver `paired = FALSE`.

> **Uppgift 1.9*** Datasetet `Salary` består av ett urval av genomsnittslöner för män och kvinnor inom olika sektorer och yrken år 2022. Läs in datasetet i R och ta en titt på din dataframe. Lägg till en ny kolumn som består av den naturliga logaritmen av lönerna. Beräkna även gärna deskriptiv statistik för båda grupperna med hjälp av funktionen:

```
aggregate(Mitt_data$numerisk_variabel, list(Mitt_data$kategorisk_variabel), summary)
```

Alternativt med `dplyr` i `tidyverse`:

```
Mitt_data %>% group_by(kategorisk_variabel) %>%
  summarise_at(vars(numerisk_variabel), list(xbar = mean, sd = sd, n = length, ... osv))
```

Är detta parvisa, eller icke-parvisa observationer?

> **Uppgift 1.10** Rita fördelningarna av de logaritmerade lönerna hos de båda grupperna i samma graf i en boxplot och kommentera grafen. Tips: använd gärna koden:

```
boxplot(mitt_data$numerisk_variabel ~ mitt_data$kategorisk_variabel)
```

> **Uppgift 1.11*** Testa på en 10%-ig signifikansnivå ifall (den logaritmerade) genomsnittslönen skiljer sig åt mellan män och kvinnor. Är detta ett enkelsidigt, eller dubbelsidigt test?

2. Icke-parametriska metoder

Metoderna ovan används flitigt inom statistikens alla hörn, men hur går man tillväga ifall man har ett litet urval med okänd fördelning? Som du kan gissa av rubriken så kan man använda sig av icke-parametriska metoder.

Några vanliga sådana metoder är

- `Teckentest`
- `Wilcoxon's test`
- `Spearman's rank correlation`
- `Wann Whitney U test`

Dessa tester kräver dock att vi har data på åtminstone ordinalskala och de tre första testerna kräver att vi har parvisa observationer. Mann-Whitney U används alltså för icke-parvisa observationer.

Tänk också på att Tecken-, Wilcoxon- och Mann Whitney U används för att testa ifall två populationer har samma fördelning med samma läge.

> **Uppgift 2.1** Vad finns det för fördel, respektive nackdel med att använda sig av icke-parametriska metoder?

2.1 Teckentest

För att utföra Teckentest i R kan man använda sig av koden

```
SIGN.test(x = obs_urval1, y = obs_urval2, alternative, conf.level)
```

Där `alternative` kan ta värdena "two.sided", "less" eller "greater" precis som vid `t.test()` och `prop.test()`

2.2 Wilcoxon's test

För att utföra Wilcoxon's test i R används pseudo koden nedan:

```
wilcox.test(x = obs_urval1, y = obs_urval2, alternative, paired = TRUE)
```

2.3 Mann Whitney U test

Mann Whitney testet använder nästan exakt samma kod, men istället låter man argumentet `paired` vara lika med `FALSE`.

```
wilcox.test(x = obs_urval1, y = obs_urval2, alternative, paired = FALSE)
```

Koderna för Tecken-, Wilcoxon och Mann Whitney testerna är alltså väldigt lika koden för t-test i föregående avsnitt.

2.4 Spearman's rangkorrelation

Spearman's test kan utföras med R's inbyggda funktion `cor.test()` likt nedan

```
cor.test(x = obs_urval1, y = obs_urval2, alternative, method = "spearman")
```

Kom ihåg att korrelationen är ett tal mellan -1 och 1. Med argumentet `alternative` testar man här ifall den sanna korrelationen är större än 0 då man skriver `greater`, mindre än 0 då man skriver `less` och antingen större eller mindre än 0 då man skriver `two.sided`. Testet ger ett värde på teststatistikan, den skattade rangkorrelationen och ett p-värde.

2.5 Fotbolls-VM

Vi ska nu titta på fyra olika dataset som vi kommer göra lite olika tester på. Låt oss börja med fotbolls VM!

`wc_men` och `wc_women` består av 8 observationer och är ett urval av länder som spelade i fotbolls VM 2022 och 2023 i herr- respektive damfotboll. Läs in båda dataseten i R och ta en titt på dem. Den första variabeln är en lista av länder i urvalet. Den andra variabeln är antal mål de gjort under turneringen. Den tredje variabeln är antal gula kort varje land ådragit sig och den fjärde är antal mål som de släppt in.

> **Uppgift 2.2*** Är antal mål som görs i VM-sammanhang lika i herr- och damfotboll? Formulera dina antaganden och utför ett lämpligt icke-parametriskt test. Är detta ett test för parvisa eller icke-parvisa observationer?

En del av fotbollen är att ta risker, en annan är att försvara sig. Ofta kan gula kort uppstå vid sådana tillfällen. Två gula kort resulterar också automatiskt i ett rött kort och utvisning (oftast i alla fall, se Croatia vs Australia). En studie av Badiella et al. (2022) visar på att kort påverkar matchbilden. Röda kort har en särskilt stor negativ påverkan på det lag som ådragit sig det och enligt deras studie så tenderar det att leda till en större målskillnad mellan de båda lagen.

Uppgift 2.3* Testa med ett lämpligt test ifall antal insläppta mål (`Goals_conceded`) har ett positivt samband med antal gula kort (`YellowC`) inom herrfotboll. Vilken är den lägsta signifikansnivån du kan förkasta H_0 på utifrån ditt resultat?

2.6 Proteinpulver och muskelmassa

Företaget *IronMass* tillverkar proteinpulver och de vill veta ifall deras nya pulver som de håller på att lansera har en positiv påverkan på muskelmassan. De har gjort ett slumpmässigt urval av 7 individer och uppmätt dessas muskelmassa (som procentuell andel av kroppsvikten) före proteinintaget och sedan efter att de tagit proteinpulvret 2 gånger i veckan under en månad. De fick följande resultat i %

```
before <- c(36.3, 35.1, 32, 37.4, 41.7, 29.3, 31)
after <- c(37.9, 36.7, 29.9, 38.2, 42, 34.3, 31.5)
```

Uppgift 2.4* Testa ifall proteinpulvret har haft en positiv effekt på muskeltillväxten med ett lämpligt test på 10%-ig signifikansnivå. Vad drar du för slutsats?

2.7 VM-igen

En fotbollssupporter vill veta ifall antal reklaminslag som består av spelreklam under reklampauser skiljer sig åt före fotbolls-VM och under fotbolls-VM. Hon gjorde 10 olika mätningar på kanal 4 en månad före VM och sedan 10 olika mätningar under VM och fick följande siffror:

```
before <- c(3, 7, 2, 4, 3, 1, 5, 4, 3, 2)
during <- c(7, 7, 6, 6, 9, 6, 4, 6, 5, 5)
```

Uppgift 2.5* Testa ifall det finns en skillnad i antal reklaminslag före och under VM med ett lämpligt test. Börja med att ställa upp hypoteser såväl som dina antaganden. Vad säger resultatet?

3. Bayesiansk inferens

Hittills har vi i (klassisk) frekventistisk inferens behandlat våra parametrar som fixa, om än okända, och antagit att våra variabler X_i följer en viss fördelnig $f(x_i)$. Men du kanske minns Bayes sats från statistik 1 där vi har

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Men tänk om vår parameter också har en fördelning? Då kan vi äntligen skriva

$$P(\theta|x_1,...,x_n) = \frac{P(x_1,...,x_n|\theta)P(\theta)}{P(x)} = \frac{P(x_1,...,x_n|\theta)P(\theta)}{\int_\theta P(x_1,...,x_n|\theta)P(\theta)d\theta} \quad (1)$$

Det vi ser ovan är en Posteriorfördelning. Där $P(x_1,...,x_n|\theta)$ är vår likelihoodfunktion och $P(\theta)$ är vår prior (Priorfördelning). Nämnaren i det nedersta uttrycket är ett bekant resultat som kommer från lagen om total sannolikhet. Nämnaren brukar även kallas för normaliseringskonstanten (normalizing constant) och behövs för att hela bråket ska kunna integreras till 1.

> **Uppgift 3.1** Vad tror du är fördelen med att en parameter θ har en betingad fördelning?

Uttrycket må se vackert ut i Ekvation 1 men vi ser också att nämnaren behöver integreras. Detta kan lätt bli krångligt och det är inte säkert att man alltid kan integrera det analytiskt. Men här kommer vi endast att fokusera på fördelningar som man kan härleda analytiskt. Innan vi gör det så bör vi nämna begreppet konjugat prior

3.1 Konjugat prior

Kort sagt kan man säga att om man har en likelihoodfunktion från en viss fördelning och en konjugat prior så kan man härleda deras avkomma - posteriorfördelningen analytiskt. posteriorfördelningen kommer då att ha samma fördelning, om än med lite annorlunda parametrar, som priorfördelningen hade.

3.2 Likelihoodfunktionen

Dataset `MWh` innefattar den slutliga totala energianvändningen i megawatt timmar i 10-tals miljoner för år 2021 bland Sveriges kommuner (där saknade observationer och tre möjliga outliers tagits bort). Nedan finns även två mindre dataset definierade, `MWH_small` och `MWH_medium` som kommer från samma källa men som endast innehåller 5 respektive 15 slumpmässigt valda observationer.

```
MWH_small <- c(0.0504906, 0.2238188, 0.0555522, 0.0631950, 0.0563334)

MWH_medium <- c(0.0348163, 0.0300705, 0.0564024, 0.0339324, 0.0299269,
                0.0752713, 0.0751640, 0.0328736, 0.0336011, 0.0505392,
                0.5013083, 0.0555522, 0.0215687, 0.0101427, 0.0189661)
```

Börja med att läsa in alla tre dataset i R.

Vi kommer senare använda dessa dataset för att grafiskt illustrera olika Posteriorfördelningar där vi antar att vårt data (`MWh`) följer en exponentialfördelning med parameter λ.

Utgå alltså från att vi använder den alternativa parameteriseringen av exponentialfördelningen där $E(X) = \frac{1}{\lambda}$, dvs:

$$f(x) = \lambda e^{-\lambda x} \tag{2}$$

Det kan tyckas onödigt att byta definition nu, men som Bayesianer så kommer denna definition att underlätta våra eventuella analytiska beräkningar väldigt mycket i det här fallet.

> **Uppgift 3.2*** Finn uttrycket för likelihoodfunktionen $P(x_1, ..., x_n|\lambda)$ utifrån Ekvatin 2 ovan. Observera att du endast ska finna likelihoodfunktionen här, inte någon maximumlikelihoodskattning eller log-likelihoodfunktionen.

> **Uppgift 3.3*** Skriv en funktion i R vid namn `likelihood`, som tar argumentet `lambda` och `x`. Använd dig av uttrycket som du fick fram ovan för likelihooden i din funktion. Koden nedan är ett exempel som visar vilken output funktionen bör ge då `lambda` $= 1$ och `x` $= 0.23$ och 0.12.

```
likelihood(lambda = 1, x = c(0.23, 0.12))
```

```
[1] 0.7046881
```

3.3 Priorfördelningen

Antag vidare att du har en konjugat prior-fördelning för λ, där

$$\lambda \sim Gamma(\alpha, \beta)$$

Men använd inte kursbokens definition av Gammafördelningen där β är i nämnaren av funktionen. Använd istället definitionen där β är i täljaren, se Wikipedia

> **Uppgift 3.4*** Skriv ett uttryck för Priorn: $P(\lambda)$.

> **Uppgift 3.5*** Använd dig av uttrycket i Uppgift 3.4 för att skriva en funktion i R vid namn `prior`. Låt `prior` ta argumenten `a, b` och `lambda`. Ett exempel på vilken output du bör få ges nedan för `a = 5`, `b = 2` och `lambda = 1`.

```
prior(a = 5, b = 2, lambda = 1)
```

```
[1] 0.180447
```

Överkurs

Eftersom du har en konjugat prior så går det att härleda Posteriorfördelningen analytiskt.

> **Extaruppgift 1:** Härled Posteriorfördelningen $P(\lambda|x_1, ..., x_n)$ analytiskt genom att ställa upp bråket:
>
> $$P(\lambda|x_1, ..., x_n) = \frac{P(x_1, ..., x_n|\lambda)P(\lambda)}{\int_\lambda P(x_1, ..., x_n|\lambda)P(\lambda)d\lambda}$$
>
> Tips: tänk först på vilka termer du kan flytta utanför integralen i nämnaren innan du eventuellt skulle integrera och förkorta bort sådant som förekommer i täljare och nämnare. Snygga till uttrycket i bråket så mycket det går. Tänk sedan på vad uttrycket i nämnaren skulle behöva för att integrera till 1 och ersätt sedan integralen i nämnaren med detta nya uttryck. Kan du se vilken fördelning samt vilka parametrar $P(\lambda|x_1, ..., x_n)$ består av?

3.4 Den icke-normaliserade Posteriorfördelningen

Även om man inte får fram Posteriorfördelningen analytiskt så vet vi att den är proportionell mot $P(x_1, ..., x_n|\lambda)P(\lambda)$ eftersom integralen som vi hade i nämnaren endast var en konstant. Produkten av priorn och likelihooden kommer alltså att ha samma form rent grafiskt som Posteriorfördelningen.

> **Uppgift 3.6*** Skapa en funktion som består av produkten av din prior och din likelihoodfunktion. Kalla din funktion för `Posterior_unNormalized` och låt den ta argumenten `a, b, lambda` och `x`. Tips: Här kan du med fördel skriva en funktion som består av de andra två funktionerna `likelihood` och `prior`. Nedan ser vi ett exempel på vilken output funktionen bör ge ifall `a` = 5, `b` = 18, `x` består av värdena 0.1, 0.08 och 0.04 och `lambda` = 1.

```
Posterior_unNormalized(a = 5, b = 18, x = c(0.1, 0.08, 0.04), lambda = 1)
```

```
[1] 0.0009622897
```

3.3 En usel prior ¯_('J')_/¯

Du kommer nu att använda usla värden till din prior. Låt $\alpha = 2$ och $\beta = 2$.

Uppgift 3.7* Skapa en så kallad grid av lambda-värden bestående av minst 200 observationer och som som går mellan 0 till 15. Kalla den för något passande såsom `lambaGrid`. Rita sedan din likelihood, Prior och icke-normaliserade Posterior i tre separata grafer. Tips: Låt alltså `a = 2` och `b = 2`, `x` = värdena i `MWH_small` och lambda vara lika med `lambdaGrid` för att kunna rita linjediagrammen. Använd också din grid av lambda-värden på x-axeln i plottarna.

Uppgift 3.8* Gör samma sak som i Uppgift 3.7 ovan (för likelihoodfunktionen och den icke-normaliserade Posteriorfördelningen) men för datasetet `MWH_medium`.

Uppgift 3.9* Upprepa det du gjorde i 3.8, men för datasetet `MWH`, dvs med samtliga observationer på din x-variabel den här gången. Varför behöver man inte skapa tre olika linjer för Priorfördelningen (på samma sätt som för likelihoodfunktionerna och Posteriorfördelningarna)?

Uppgift 3.10* Vad tror du väntevärdet för din Posteriorfördelning $P(\lambda|x_1, ..., x_n)$ skulle vara? Vad kan du dra för slutsats av samtliga grafer?

4. Sammanfattning

I den här labben hav vi först ägnat oss åt parametriska tester, såsom test för proportioner och t-tester, både för en variabel och för grupper. Sedan har vi tittat på olika icke-parametriska tester som kan användas då man inte vet vilka antaganden som är uppfyllda. Avslutningsvis har vi ägnat oss åt Bayesiansk inferens och hur en Posteriorfördelning förhåller sig till likelihoodfunktioner och Priorfördelningen.

Referenser

Badiella, L.,Puig, P. & Peñas, C. & Casals, M. (2022). Influence of Red and Yellow cards on team performance in elite soccer. *Annals of Operations Research*. 325. DOI: 10.1007/s10479-022-04733-0.