ABCLakrids

Smilk

13 sep 2018

Introduktion til data og metode

Simon har samplet 4 poser ABC lakrids fra Haribo fra en Kvickly (her kan man overveje samplemetoden). Poserne er samplet ud af af en population vi kan kalde ABC lakridser produceret til Danmark, hvor populationen (N) formodes at være meget stor, hvorved vi må antage at 1/N og n/N er 0, hvilket også betyder at vores estimat for populationstotalen vil være uendelig. I stikprøven har vi oberveret hvilke bogstaver hver enhed har. Således er interessevariablen en indikatorfunktion som kan tage 29 værdier. Vi har dog måttet sande at vi ikke har observeret ø eller å og at det er umuligt at skeldne mellem M og W og Z og N, da det ikke fremgår klart hvordan de respektive bogstaver bør se ud. For at gøre dette så neutralt som mulig, og stadig gøre det muligt at lave beregninger, har vi valgt at antage, bogstaver som kan kategoriseres som to værdier fordeles til halv halv, og at de bogstaver som ikke er observeret ikke forekommer i totalpopulationen. Hver pose betragtes som en sample, af størrelse n (stokastisk værdi afhænger af vægten, da antallet af bogstaver varierer i hver pose). Det gennemsnitlige antal bogstaver pr pose er 80 (mere præcist 79.75 men vi kan ikke observere halve bogstaver)

##Overvejelser og problematikker med stikprøven. Det er en række forskellige eventuelle problemer med vores stikprøve, som vi er nød til at behandle inden vi laver beregninger på data. Som nævnt ovenfor kender vi ikke vores populationsstørrelse N, da vi ikke ved hvor mange ABC lakridser der blevet produceret og solgt i Danmark som vi ville have haft mulighed for at købe. Derfor har vi valgt at betragte en uendelig, men dog deterministisk, populationsstørrelse. En anden problematik er, at der er nogle bogstaver vi ikke kan kende forskel på, samt nogle bogstaver som vi slet ikke har samplet. Som nævnt ovenfor har vi derfor arbejdet med et alfabet med 27 bogstaver (uden ø og å). Problemstillingen med ikke at kunne se forskel på nogen bogstaver (f.eks N og Z) løste vi ved at tælle halvdelen af dem som N'er og halvdelen som Z'er. Vi mener at dette er den bedste løsning, selvom det muligvis kan introducere bias i vores stikprøve. Man kan også overveje om det er unbiased alle fire poser er samplet fra samme butik på samme tidspunkt. I stedet burde man have samplet ved SRS med tilbagelægning hvilke 4 butiker i Danmark hvor poserne skulle købes, da vi jo a priori ikke ved om det kunne skabe en bias i vores sample inden vi laver eksperimentet.

Overvejelser sammenlignet med alfabetet.

Hvis vi ønsker at skrive et bestemt ord, kræver det ofte (i alt fald på dansk) at vi har en rimmelig mængde af vokaler til rådighed. Lad derfor $y_v(i)$ betegne indikatorfunktionen for at individ i er en en vokal. Fra dette finder vi (jvf Corollary 2.13) at y bar inden for hver pose er den empiriske varians 0.0025 0.0024 0.0025 og 0.0025 for hhv. pose 1, 2, 3 og 4. med konfidens interval for hver af de givne poser som følgende.

```
id <- ifelse(rownames(Dat_new)%in% c("A","E","I",
                                     "O","U","Y",
                                     "AE"), TRUE,FALSE)
y_bar <- colSums(Dat_new[id,])/n
var_emp <- (1-0)*y_bar*(1-y_bar)/(n-1)
sqrt(var_emp) #Standard afvigelse
```

```
## [1] 0.04964741 0.04849737 0.04972873 0.05041362
```

```
uu <- qnorm(.975)
CI <- cbind(y_bar-uu*sqrt(var_emp),y_bar+uu*sqrt(var_emp))
round(CI,2)
```

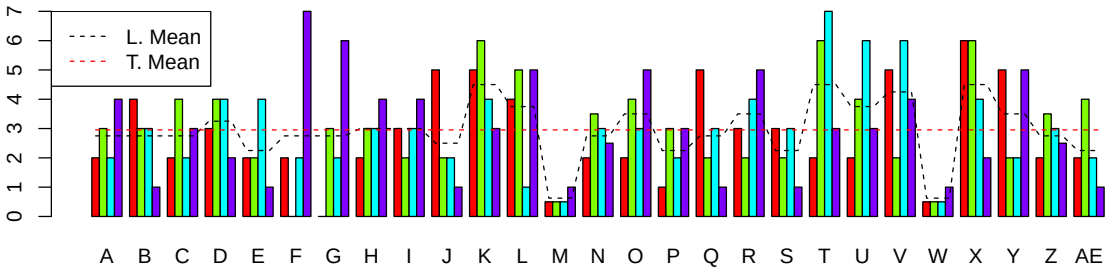


Figure 1: Bogstaver samplet fra de forskellige poser. Hver farvet bar indikerer en pose. Den sorte linje (L. Mean) viser gennemsnittet for hver bogstav over de 4 poser, mens den røde linje angiver gennemsnittet over alle bogstaver på tværs af poser

```
##      [,1] [,2]
## [1,] 0.14 0.34
## [2,] 0.16 0.35
## [3,] 0.17 0.37
## [4,] 0.19 0.38
```

Til sammenligning forventer vi at det danske alphabet med 27 bogstaver (ink. W ekls. ø og å) og 7 vokaler har en forekomst af vokaler på 0.26. vi forkaster altså ikke, at fordelingen af vokaler og konsonanter i vores stikprøver afviger fra afabetet.

Ifølge sproget.dk (https://sproget.dk/temaer/ord-og-bogstaver/hvad-er-de-mest-almindelige-bogstaver-pa-dansk#kilder-kjeld-kristensen-bogstavernes d. 13/9) forekommer bogstaverne e, n, d og r hyppigere end andre bogstaver i alphabetet. Lad os derfor gøre tilsvarende som ovenfor for at undersøge om disse forekommer i en frekvens der minder om alfabetet eller, hvis poserne er designet til at man kan stave ord, vil vi forvente disse forekommer oftere end andre bogstaver.

```
id2 <- ifelse(rownames(Dat_new)%in% c("E","N","D","R"),
                TRUE,FALSE)
y_endr <- colSums(Dat_new[id2,])/n

var_endr <- y_endr*(1-y_endr)/(n-1)
sqrt(var_endr) #Standard afvigelse
```

```
## [1] 0.03951660 0.03858221 0.04342978 0.03755426
```

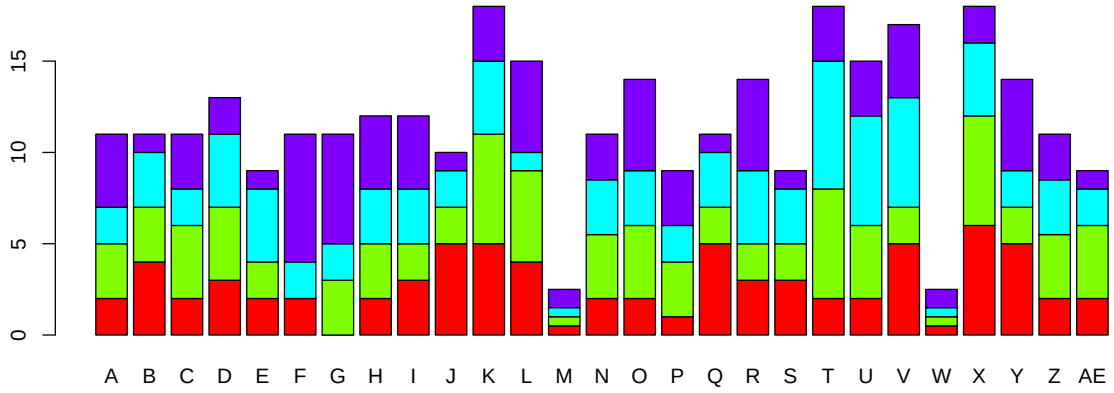
```
CI <- cbind(y_endr-uu*sqrt(var_endr),y_endr+uu*sqrt(var_endr))
round(CI,2)
```

```
##      [,1] [,2]
## [1,] 0.06 0.21
## [2,] 0.06 0.22
## [3,] 0.10 0.27
## [4,] 0.06 0.20
```

Med 4/27 aproksimativt lig 0.15 indeholdt i samtlige konfidensintervaller. Vi kan altså ikke forkaste at e, n, d og r ikke forekommer med samme hyppighed som i alfabetet.

Test for uniform fordeling af bogstaverne.

Vi vil nu teste hypotesen, at bogstaverne er uniformt fordelte, dvs. om vi kan antage, at der er nogenlunde det samme antal af hver bogstav i en pose. Vi vil hertil begynde med at betragte et histogram, som viser en fordeling af vores stikprøve. Et barplot summer over antallet af de forskellige bogstaver i `Dat_new` og viser således et histogram af de observerede bogstaver.



Ud fra histogrammet kan vi ikke udelukke en uniform fordeling af bogstaverne. Vi vil nu bruge Kolomogorov-Smirnov testet til at finde en sandsynlighed for, at bogstaverne er uniformt fordelte. Hertil finder vi en vektor, der indeholder tallene fra 1 til 27, som henholdsvis repræsenterer der observerede bogstaver fra `A` til `Æ`.

```
Dat_new2 <- Dat_new[-(28:29),]
let <- integer(sum(Dat_new2))
i = 1

for (k in 1:4){
  for (j in 1:27){
    while (Dat_new2[j,k] > 0){
      let[i] = j
      Dat_new2[j,k] = Dat_new2[j,k] - 1
      i = i+1
    }
  }
}

suppressWarnings(ks.test(let, "punif", 1, 27))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  let
## D = 0.061016, p-value = 0.1791
## alternative hypothesis: two-sided
```

Sandsynligheden $p = 0.179$ argumenterer dog ikke særlig stærk for en uniform fordeling af tallene.
Som nævnt, så er der ingen forskel mellem bogstaverne `M` og `W`. Vi vil derfor gentage Kolomogorov-Smirnov testen, hvor vi først lægger antallene af observerede `M`'er og `W`'er sammen. I den følgende kodestump bliver `W` transformeret til `M`. Derefter kører vi `ks.test`.

```r
for (i in 1:324){
  if (let[i] == 23){
    let[i] <- 13
  }
}

suppressWarnings(ks.test(let, "punif", 1,27))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  let
## D = 0.04867, p-value = 0.4266
## alternative hypothesis: two-sided
```

Sandsynligheden $p = 0.4266$ er allerede et meget stærkere argument for en uniform fordeling. Hvis `M` og `W` antages at være den samme bogstav, vil vi derfor ikke forkaste hypotesen om en uniform fordeling.

Konklusion