UNIWERSYTET GDAŃSKI

WYDZIAŁ MATEMATYKI, FIZYKI I INFORMATYKI

Ewa Bojke

MODELOWANIE MATEMATYCZNE
I ANALIZA DANYCH

Metody i modele bayesowskie w R Studio

Projekt zaliczeniowy
dr. Marta Frankowska

Gdańsk 2022

W tym projekcie zajmiemy się badaniem wpływu różnych czynników na to, czy dany pacjent przyjdzie na umówioną wizyte u lekarza, czy nie. Zastosujemy wnioskowanie bayesowskie aby się o tym przekonać.

Wyświetlmy zbiór danych, jaki posiadamy:

```
> head(show)
  show.AppointmentID show.Gender show.Age show.Diabetes show.Alcoholism show.SMS_received show.Show
1            5642903           F       62             0               0                0       Yes
2            5642503           M       56             0               0                0       Yes
3            5642549           F       62             0               0                0       Yes
4            5642828           F        8             0               0                0       Yes
5            5642494           F       56             1               0                0       Yes
6            5626772           F       76             0               0                0       Yes
> tail(show)
       show.AppointmentID show.Gender show.Age show.Diabetes show.Alcoholism show.SMS_received show.Show
110522            5651072           F       53             0               0                1       Yes
110523            5651768           F       56             0               0                1       Yes
110524            5650093           F       51             0               0                1       Yes
110525            5630692           F       21             0               0                1       Yes
110526            5630323           F       38             0               0                1       Yes
110527            5629448           F       54             0               0                1       Yes
>
```

Zmiennymi, które będą nas interesować to **Show**, **SMS.received**, **Gender**, **Diabetes**, **Alcoholism**, **Age**.

zmienna	opis
Gender	‘F‘ jeśli kobieta, ‘M‘ jeśli mężczyzna
Age	wiek pacjenta

zmienna	1	0
Show	pacjent pojawił się na wizycie	pacjent się nie pojawił
SMS.received	pacjent otrzymał wiadomość o wizycie	pacjent nie otrzymał
Diabetes	diabetyk	zdrowy
Alcoholism	alkoholik	zdrowy

Ogólnie, zastanówmy się nad tym, co tak naprawdę ma wpływ na przyjście pacjenta na wizytę. Czy to badanie doprowadzi nas do jakichś ciekawych wniosków? Przekonamy się o tym już za chwilę. Zacznijmy więc od tego, co już wiemy o danych.

Formalny opis wnioskowania bayesowskiego:

- x, jednostka obserwacji, może być to wektor,
- θ, parametr obserwacji tj. x $\sim$ p(x| θ)
- α , hiperparametr parametru, tj. $\theta \sim$ p($\theta \mid \alpha$). Może to być wektor hiperparametrów.
- **X** , zbiór n jednostkowych obserwacji, tj. x_1,... ,x_n.

Rozkład a priori to rozkład parametrów przyjęty przed zaobserwowaniem jakichkolwiek danych, tj. p($\theta \mid \alpha$). Reprezentuje wiedzę z jaką badacz rozpoczyna badanie.

Wyznaczmy prawdopodobieństwo a priori, jeśli chodzi o przyjście na wizytę pacjenta, ponieważ chcemy zobaczyć jak rozkładają się dane (ile osób ogólnie przyszło do lekarza).

```
> show %>%
+   tabyl(Show) %>%
+   adorn_totals("row")
  Show      n   percent
    No  22319 0.2019326
   Yes  88208 0.7980674
 Total 110527 1.0000000
> theta<-factor(c(rep("No", 22319), rep("Yes",88208)))
> priori.praw<-c(0.2019326,0.7980674)
```

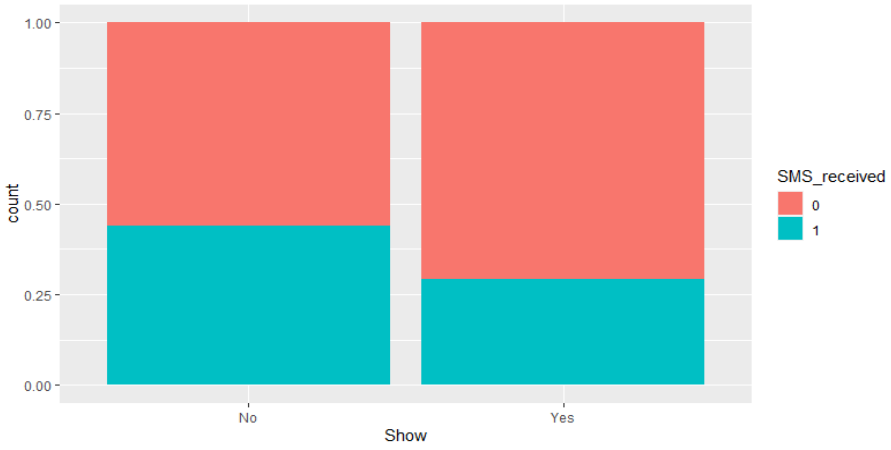
Wnioski: Widzimy, że większośc naszych danych wskazuje na to, że prawie 80% badanych przyszło do lekarza.

Rozkład z próby to rozkład obserwacji, zależnych od ich parametrów, tj. $p(\mathbf{X} \mid \theta)$. Nazywa się go również wiarygodnością, szczególnie gdy rozpatruje się ją jako funkcję parametrów, tj. L ($\theta \mid \mathbf{X}$) = p($\mathbf{X} \mid \theta$). Wyznaczmy rozkład wiarygodności zwracając uwagę, na to czy pacjent przyszedł do lekarza, ponieważ dostał SMSa o wizycie.

Zobaczmy najpierw ile osób dostało SMSa o wizycie oraz ile z tych osób pojawiło się faktycznie na wizycie.

```
> show %>%
+     tabyl(SMS_received,Show) %>%
+     adorn_totals("row")
 SMS_received    No   Yes
            0 12535 62510
            1  9784 25698
        Total 22319 88208
>
```

Obecność na wizycie a otrzymanie SMSa



Przyjrzyjmy się teraz wiarygodności:

```
> Tab<-table(SMS_received,Show)
> wiarygodnosc<- matrix(c(round(Tab[1,1]/sum(Tab[,1]),3),
round(Tab[2,1]/sum(Tab[,1]),3),round(Tab[1,2]/sum(Tab[,
2]),3),round(Tab[2,2]/sum(Tab[,2]),3)),2,2) #nie bylo go
 i nie otrzynal sms/byl ale mogl otrzymac i nie otrzymac
 sms
> rownames(wiarygodnosc)=c("no_received","received")
> colnames(wiarygodnosc)=c("No","Yes")
> wiarygodnosc
               No   Yes
no_received 0.562 0.709
received    0.438 0.291
```

Wnioski: Wiarygodność wskazuję nam na duży odsetek (prawie 71%) pacjentów, którzy nie otrzymali wiadomości o wizycie i przyszli na badanie. Ciekawe jest też to, że osób, które nie dostało SMSa i nieprzyszło na wizytę jest więcej od osób, które nieprzyszło i dostało SMSa o wizycie. Może więc, to czy ktoś dostałby z nieobecnych SMSa o wizycie zdecydowałoby, że jednak pojawili by się u lekarza. Trzeba się temu bardziej przyjrzeć.

Na potrzeby zadania obliczymy teraz rozkład łączny, który jest iloczynem obliczonej wcześniej wiarygodności i prawdopodobieństwa a priori. Rozkład łączny będzie nam potrzebny do obliczenia a posteriori, o którym zaraz wspomnimy.

```
> laczny<-matrix(c(wiarygodnosc[1,1]*priori.pra
w[1],wiarygodnosc[2,1]*priori.praw[1],
+                  wiarygodnosc[1,2]*priori.pra
w[2],wiarygodnosc[2,2]*priori.praw[2]),2,2)
>
> rownames(laczny)=c("no_received","received")
> colnames(laczny)=c("No","Yes")
> laczny
                    No       Yes
no_received 0.11348612 0.5658298
received    0.08844648 0.2322376
```

Rozkład a posteriori (in. wynikowy) to rozkład parametrów po uwzględnieniu zaobserwowanych danych. Jest określany przy pomocy twierdzenia Bayesa:

$$p(\theta \mid \mathbf{X}, \alpha) = \frac{p(\mathbf{X} \mid \theta)p(\theta \mid \alpha)}{p(\mathbf{X} \mid \alpha)}, gdzie$$

$p(\mathbf{X} \mid \alpha)$ to wiarygodność brzegowa dzięki, której możemy znormalizować nasze dane.

Jak wygląda rozkład brzegowy?

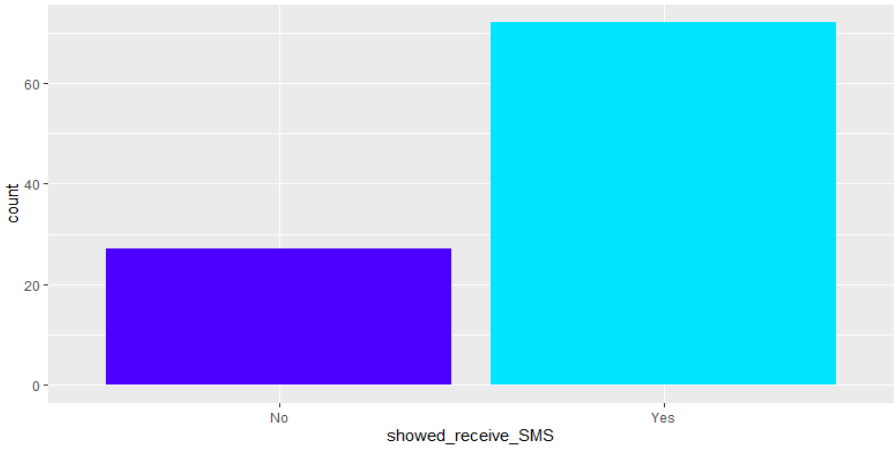
```
> X<-c("No","Yes")
> praw.X<-c(sum(laczny[1,]), sum(laczny[2,]))
> praw.X
[1] 0.6793159 0.3206841
```

Stąd, a posteriori wynosi:

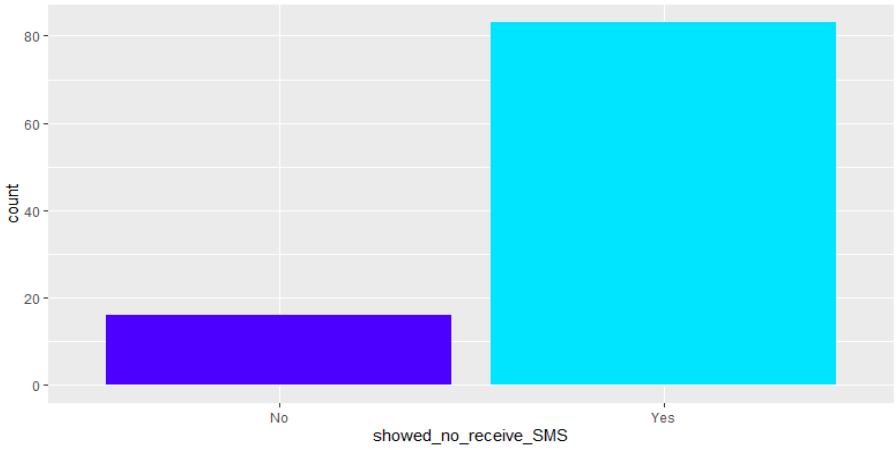
```
> posteriori<-matrix(c(laczny[1,1]/praw.X[1],laczny[2,
1]/praw.X[2],laczny[1,2]/praw.X[1],laczny[2,2]/praw.X
[2]),2,2)
> rownames(posteriori)=c("no_received","received")
> colnames(posteriori)=c("No","Yes")
>
> tab<-table(SMS_received,Show)
> tab<-as.data.frame(tab)
> show$SMS_received<-as.character(SMS_received)
> posteriori
                   No       Yes
no_received 0.1670594 0.8329406
received    0.2758056 0.7241944
>
```

Wnioski: O dziwo wyszedł nam bardzo ciekawy przypadek, bowiem tak jak myśleliśmy na poczatku, że otrzymanie SMSa o wizycie(patrząc na osoby, które były nieobecne) może mieć jakiś wpływ na to, czy dany pacjent przyjdzie na wizytę, czy nie, to teraz widzimy, że nie ma to żadnego wpływu. Otóz większość pacjentów, którzy otrzymali SMSa o wizycie nie przyszło porównując do tych osób, którzy nie otrzymali SMSa i również nie przyszli. Ciekawym wnioskiem jest fakt, że wzrosło bardzo prawdopodobieństwo osób, które przyszły na wizytę i otrzymały SMSa. Wcześniej takie prawdopodobieństwo wynosiło niecałe 30% a teraz prawie 73%.

Wykres a posteriori - SMS.received = 1



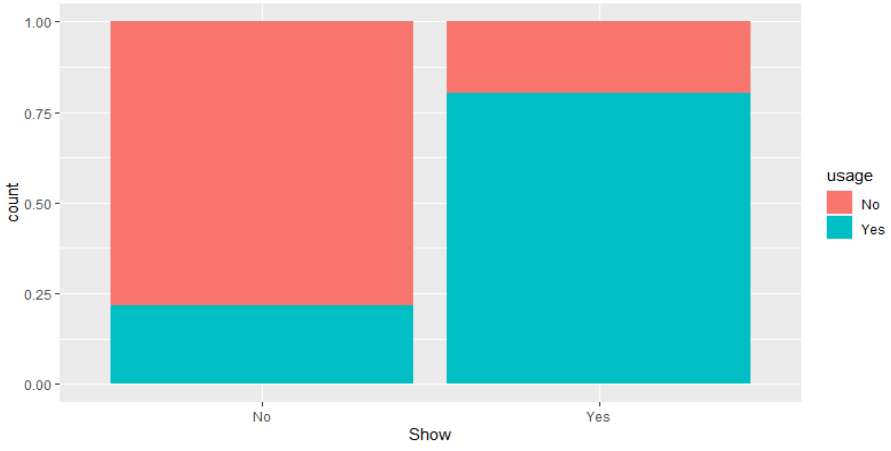
Wykres a posteriori - SMS.received = 0



Zróbmy test i wiedząc jakie jest a priori, zrobimy symulacje z próbki 1000 osób i porównamy wyniki z naszymi danymi:

```
> symulacje<-symulacje %>%
+     mutate(data_model=case_when(Show=="No"~0.205, Show=="Yes"~0.795))
> glimpse(symulacje) #dorzucenie jednej zmiennej data_model, ktora okresla prawdop tego
 ze SMS zostal otrztymany
Rows: 1,000
Columns: 2
$ Show       <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No", "No", "Yes", "Ye~
$ data_model <dbl> 0.795, 0.795, 0.795, 0.795, 0.795, 0.795, 0.205, 0.205, 0.795, 0~
>
> data<-c("No","Yes")
> symulacje<-symulacje %>%
+     group_by(1:n()) %>%
+     mutate(usage=sample(data, size=1, prob=c(1-data_model, data_model))) #usage dotyc
zy otrzyamnia SMS o wizycie
>
>
> symulacje %>%
+     tabyl(usage,Show) %>%
+     adorn_totals(c("col","row"))
 usage  No Yes Total
    No 143 160   303
   Yes  40 657   697
 Total 183 817  1000
>
```

Symulacje



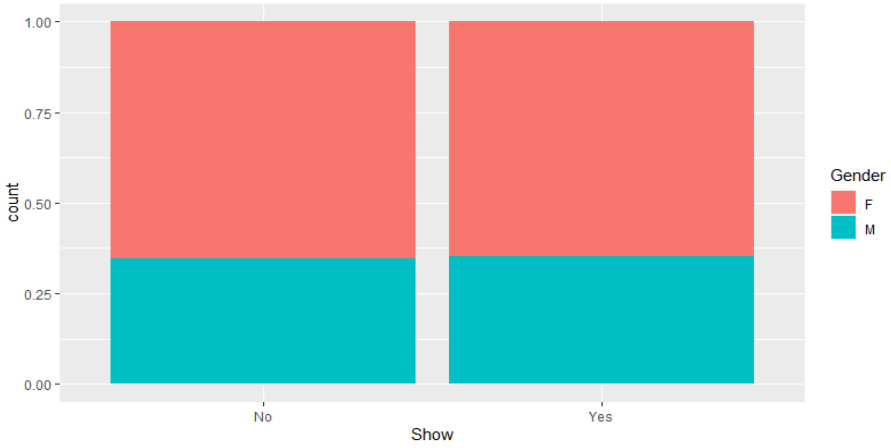
Wnioski: Możemy zauważyć, że symulacje wyszły bardzo podobnie do naszych danych, ponieważ większy odsetek jest ośób, które przyszły na badanie niż tych, które nie przyszły.

Zbadamy teraz wpływ zmiennej **Gender** na to, czy pacjent przyjdzie na wizytę, czy nie. Zobaczmy jak rozkłada się na wykresie a priori. Sprawdzimy jaka płeć częściej przychodzi na umówioną wizytę.

A priori ze względu na płeć:

```
> show %>%
+    tabyl(Gender,Show) %>%
+    adorn_totals("row")
 Gender     No   Yes
      F 14594 57246
      M  7725 30962
  Total 22319 88208
> 
```

Obecność na wizycie a płeć



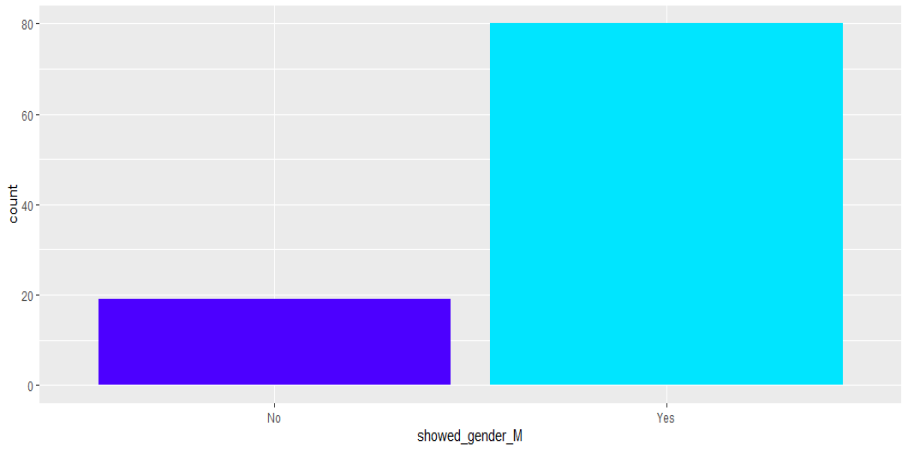
Wnioski: Widzimy, że łączna ilośc badanych kobiet jest przeważająca nad ilością mężczyzn po pierwszych danych, dlatego wykres wyglądą w ten sposób, że jest więcej kobiet, które przyszły niż mężczyzn i więcej kobiet, które nie przyszły na badanie niż mężczyzn. A priori daje na za mało informacji, dlatego musiy sprawdzić wiarygodność i wynik a posteriori.

Wiarygodność i rozkład a posteriori wynosi:

```
> wiarygodnosc
     No   Yes
F 0.654 0.649
M 0.346 0.351
> posteriori
          No       Yes
F 0.2031722 0.7968278
M 0.1996303 0.8003697
>
```

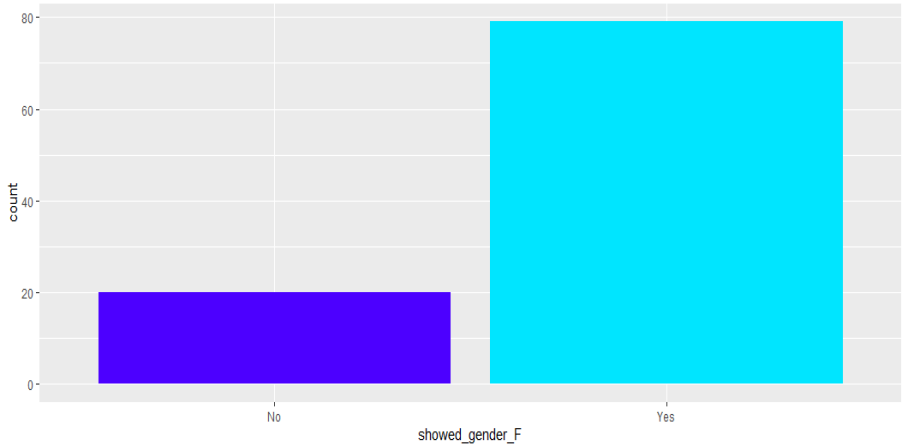
```
> showed<-c("No","Yes")
> showed_gender_M<-c(rep(showed[1],posteriori[2,1]*10
0),rep(showed[2],posteriori[2,2]*100))
> data_gender_M<-data.frame(showed_gender_M)
> ggplot(data_gender_M,aes(x=showed_gender_M))+geom_b
ar(fill=topo.colors(2))
```

Wykres a posteriori - mężczyźni



```
> showed<-c("No","Yes")
> showed_gender_F<-c(rep(showed[1],posteriori[1,1]*10
0),rep(showed[2],posteriori[1,2]*100))
> data_gender_F<-data.frame(showed_gender_F)
> ggplot(data_gender_F,aes(x=showed_gender_F))+geom_b
ar(fill=topo.colors(2))
`
```

Wykres a posteriori - kobiety



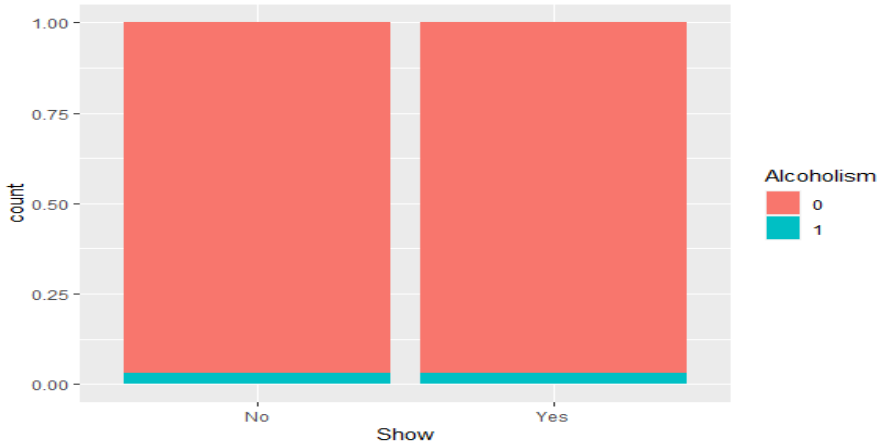
Wnioski: Po wstępnym pokazaniu a priori wydawało się, że to właśnie większość kobiet przychodzi na umówione wizyty, a jednak po większej analizie i obliczeniu a posteriori, widzimy, że te prawopodobieństwa pojawienia się na wizycie kobiet i mężczyzn są bardzo podobne.

Chcielibysmy zbadać jeszcze wpływ zmiennej **Alcoholism** na to, czy pacjent przyjdzie na wizytę, czy nie. Zobaczmy jak rozkłada się na wykresie a priori. Zobaczymy czy uzależnienie ma wpływ na przyjście na wizyte u lekarza.

A priori ze względu na zmienną alkohol:

```
> show %>%
+     tabyl(Alcoholism,Show) %>%
+     adorn_totals("row")
 Alcoholism    No   Yes
          0 21642 85525
          1   677  2683
      Total 22319 88208
```

Obecność na wizycie a alkoholizm



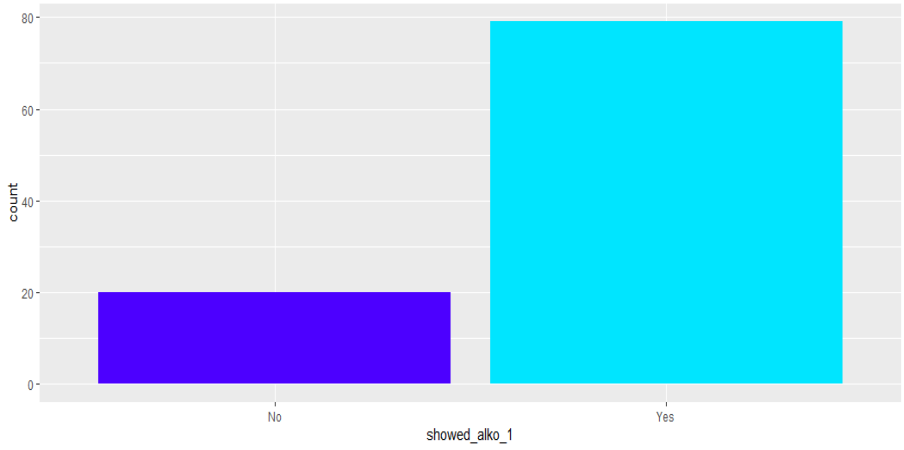
Wnioski: Można zauważyć, że bardzo mało ludzi pod wpływem alkoholu przychodzi do lekarza, większość pacjentów jest trzeźwych.

Wiarygodność i rozkład a posteriori wynosi:

```
> wiarygodnosc
    No  Yes
0 0.97 0.97
1 0.03 0.03
> posteriori
         No       Yes
0 0.2019326 0.7980674
1 0.2019326 0.7980674
```

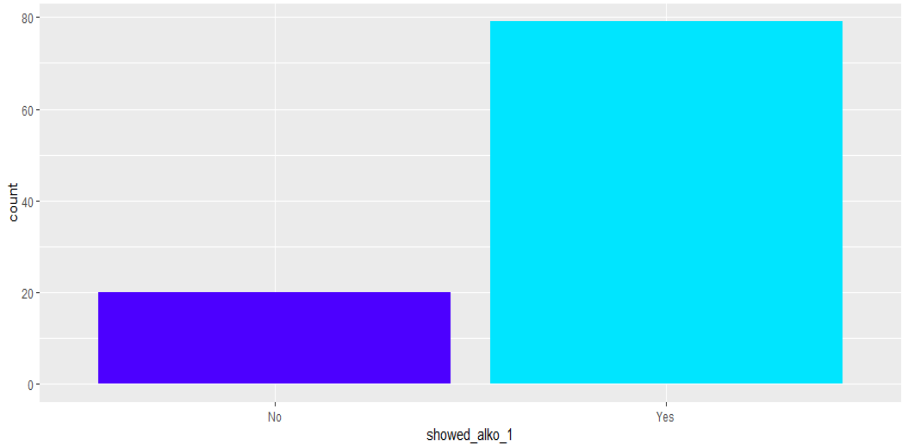
```
> showed<-c("No","Yes")
> showed_alko_1<-c(rep(showed[1],posteriori[2,1]*100),
rep(showed[2],posteriori[2,2]*100))
> data_alko_1<-data.frame(showed_alko_1)
> ggplot(data_alko_1,aes(x=showed_alko_1))+geom_bar(fi
ll=topo.colors(2))
```

Wykres a posteriori: alkohol = 1



```
> showed_alko_0<-c(rep(showed[1],posteriori[1,1]*10
0),rep(showed[2],posteriori[1,2]*100))
> data_alko_0<-data.frame(showed_alko_0)
> ggplot(data_alko_0,aes(x=showed_alko_0))+geom_bar
(fill=topo.colors(2))
```

Wykres a posteriori: alkohol= 0



Wnioski: Po wstępnym pokazaniu a priori wydawało się, że to właśnie większość osób pod wpływem alkoholu nie przyjdzie na wizytę. Trzeźwi, którzy przyszli na wizytę to aż 80% wszystkich trzeźwych, czyli około 20% osób trzeźwych nie przyszło. Natomiast nietrzeźwych, którzy przyszli na wizytę to 80% ze wszytskich nietrzeźwych, czyli 20% osób nietrzeźwych nie przyszło na wizytę.