Karol Zioło 412360

Badanie rozkładu średniej ilości spożywanego alkoholu

1. Wprowadzenie

Celem przeprowadzonego badania było zdeterminowanie rozkładu, który najlepiej by odzwierciedlał rozkład średniej ilości spożywanego alkoholu wśród państw na świecie. Z ekonomicznego punktu widzenia analiza ta jest wyjątkowo istotna, ponieważ alkohol ma znaczący wpływ na wiele obszarów życia człowieka, a sama branża stanowi stosunkowo dużą część PKB, zwłaszcza wśród krajów zachodnich.

Tak jak powyżej wspomniano przemysł alkoholowy stanowi istotną część gospodarki wielu państw. Wpływ na to ma wiele czynników i dla każdego kraju są one inne. Bazując na polskim rynku można wytypować przede wszystkim wzrost poziomu życia, dążenie społeczeństwa do większej wygody i kompleksowości. Istotnym czynnikiem jest też rozwój kultury picia. W wielu krajach spożywanie alkoholu jest wpisywane w ich tradycję oraz stanowi znaczącą rolę w ich kulturze. Dodatkowo, ciągły rozwój branży przyczynia się do powstawania nowych rodzajów napoi alkoholowych. Skutkuje, to zwiększoną trafnością w gust ludzi, czego efektem jest wzrost popytu na alkohol.

Co więcej alkohol ma też pośredni wpływ na gospodarkę. Przede wszystkim jest jedną z najbardziej uzależniających używek na świecie. Mnóstwo ludzi poprzez jego nadmierną konsumpcję popada w alkoholizm niszcząc przy tym swój organizm. Ma to negatywne skutki na wielu płaszczyznach. Po pierwsze, silnie uzależnione osoby nie są w stanie pracować, co skutkuje obniżeniem siły roboczej w kraju. Efektem tego jest spadek PKB, ponieważ nie wszystkie zasoby są w pełni wykorzystane. Następnym skutkiem jest wzrost wydatków na służbę zdrowia. Wyniszczony przez alkohol organizm wymaga specjalistycznego leczenia zarówno pod względem zdrowia fizycznego, jak i psychicznego.

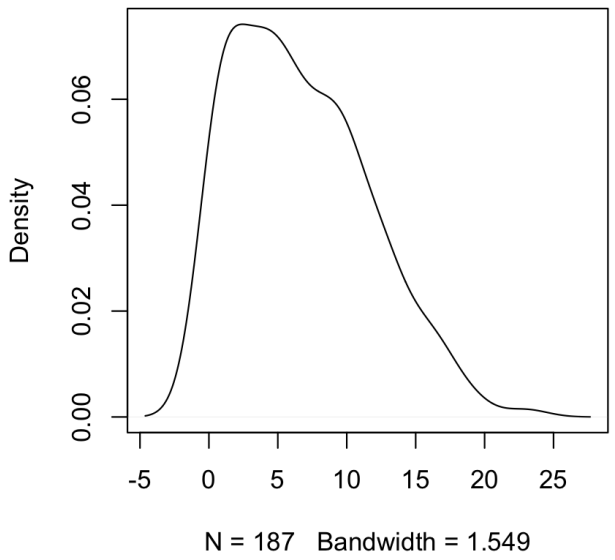
Przedstawione powyżej przykłady potwierdzają fakt, że badania konsumpcji alkoholu są istotne z ekonomicznej perspektywy. Przede wszystkim pomagają w analizie zachowań konsumenckich, poprzez dostarczanie producentom nowych informacji związanych z nowymi trendami i

preferencjami. Ponadto są one też istotne w analizie wydatków publicznych, ponieważ alkohol generuje tzw. koszty zewnętrzne. W niniejszym badaniu przeprowadzono analizę rozkładu średniej ilości spożywanego alkoholu wśród państw na świecie. Dostarczy ona więcej informacji dotyczących poziomu konsumpcji alkoholu oraz pozwoli zrozumieć tendencje do jego spożywania na świecie.

2. Propozycja rozkładów

W celu dobrania odpowiednich rozkładów stworzono wykres częstości (Wykres 1.).

Wykres 1. Wykres częstości średniej ilości spożywanego alkoholu wśród państw na świecie



Źródło: Opracowanie własne na podstawie danych WHO

Na podstawie analizy wykresu można było założyć, że zgromadzone dane mogą się cechować rozkładem normalnym lub rozkładem log-normalnym. W tym celu zebrano literaturę, opisującą te dwa rozkłady.

W badaniu statystycznym, przeprowadzonym przez Johna B.F. Fielda, zaproponowano rozkład log-normalny jako model opisujący rozkład konsumpcji alkoholu. Autor dokładnie opisał cechy jakimi się on charakteryzuje. Przedstawił też wykresy dla przykładowych parametrów. Na ich podstawie można dojść do wniosku, że zebrane dane mogą reprezentować rozkład log-normalny.

Drugim rozważanym przykładem był rozkład normalny, który stanowi podstawę do otrzymania rozkładu log-normalnego. Dlatego oba rozkłady cechują się tymi samymi parametrami, mu i sigma. Ponownie porównując wykresy można stwierdzić, że przedstawiony powyżej wykres zdecydowanie przypomina rozkład normalny.

3. Metodologia i wyniki

W celu wyestymowania parametrów powyższych rozkładów zastosowano Metodę Największej Wiarygodności. Dodatkowo uzyskano także parametry przy użyciu Metody Momentów, które posłużyły jako wartości startowe w procesie estymacji. Dodatkowo, w celu otrzymania wiarygodnych wyników dla każdego rozkładu zastosowano gradient i hessian. W poniższej tabeli przedstawiono wyniki estymacji:

	Rozkład normalny		Rozkład log-normalny	
	Mu	**Sigma**	**Mu**	**Sigma**
MM	6.69	4.90	1.77	0.49
MNK	6.69	4.88	1.42	1.24

Bazując na wynikach można zaobserwować, że w przypadku rozkładu normalnego wyniki otrzymane metodą momentów były bardzo zbliżone do wyników uzyskanych metodą największej wiarygodności. Przeciwnie było w przypadku rozkładu log-normalnego, gdzie wyniki znacząco się różniły.

3. Testowanie hipotez

W celu ustalenia hipotez posłużono się artykułem pt. *Determining the best population-level alcohol consumption model and its impact on estimates of alcohol-attributable harms*. Autorzy wyestymowali w nim parametry rozkładu log-normalnego dla zebranych danych dotyczących konsumpcji alkoholu wśród 43 krajów z podziałem na mężczyzn i kobiety. Jako wartości testowe wybrano średnią z otrzymanych parametrów dla mężczyzn, a wyniki zostały przedstawione w tabeli poniżej:

	Rozkład normalny		Rozkład log-normalny	
	Hipoteza	**Wynik**	**Hipoteza**	**Wynik**
z-test	mu = 7.03	p-value = 34% Brak podstaw do odrzucenia hipotezy zerowej.	mu = 1.95	p-value = 0% Odrzucamy hipotezę zerową.
LR-test	mu = 7.03 sigma = 4.71	p-value = 48% Brak podstaw do odrzucenia hipotezy zerowej.	mu = 1.95 sigma = 1.55	p-value = 0% Odrzucamy hipotezę zerową.

Ostatnim krokiem w przeprowadzonej analizie było zbadanie zgodności zebranych danych z założonymi rozkładami. W tym celu zastosowano test Kołmogorowa-Smirnowa, jego wyniki zostały zaprezentowane w tabeli poniżej:

	Rozkład normalny	Rozkład log-normalny
Test Kołmogorowa-Smirnowa	p-value = 12% Brak podstaw do odrzucenia hipotezy zerowej	p-value = 0% Odrzucamy hipotezę zerową

3. Wnioski

Na podstawie przeprowadzonych badań można wyciągnąć pare wniosków. Po pierwsze wyniki wykazały, że badane dane cechują się rozkładem normalnym. Wskazuje na to wynik testu Kołmogorowa-Smirnowa, w którym p-value wyniosło 12%, co oznacza, że nie ma podstaw do odrzucenia hipotezy zerowej o zgodności danych do badanego rozkładu. Zatem metoda największej wiarygodności pozwoliła wyestymować parametry opisujące rozkład normalny. Oznacza to, że średnia wynosi 6.69, a odchylenie standardowe 4.88. Dodatkowo w badaniu porównano otrzymane wyniki do parametrów uzyskanych w innym badaniu. Okazało się, że różnice pomiędzy parametrami są statystycznie nieistotne.

Jednakże, badanie dotyczące porównania zebranych danych do rozkładu log-normalnego wskazywało na przeciwne wnioski. Przede wszystkim test Kołmogorowa-Smirnowa wykazał niezgodność pomiędzy danymi, a badanym rozkładem, co oznacza, że sam rozkład i

wyestymowane parametry niepoprawnie opisują rozkład danych. Ponadto porównanie otrzymanych wyników do do parametrów uzyskanych w innym badaniu wykazało, że są one statystycznie różne.

Załącznik: Kod R

```
dane = read.csv("gapminder_alcohol.csv")
N = length(dane$country)
x = dane$alcconsumption

##### METODA MOMENTOW #####
# ROZKLAD NORMALNY

y = dane$alcconsumption
uklad.rownan = function(x){
  mu = x[1]
  sigma = x[2]
  r1 = mean(y)-mu
  r2 = var(y)-sigma^2
  return(c(r1,r2))
}

wynik_norm = multiroot(f=uklad.rownan,start=c(1,1))
wynik_norm$root

ks.test(y, "pnorm", wynik_norm$root[1], wynik_norm$root[2])

# ROZKLAD LOGNORMALNY

y = dane$alcconsumption
uklad.rownan = function(x){
  mu = x[1]
  sigma = x[2]
  r1 = mean(y)-exp(mu+(sigma^2)/2)
  r2 = median(y)-exp(mu)
  return(c(r1,r2))
}

wynik_lnorm = multiroot(f=uklad.rownan,start=c(1,1))
wynik_lnorm$root

ks.test(y, "plnorm", wynik_lnorm$root[1], wynik_lnorm$root[2])


##### METODA NAJWIEKSZEJ WIARYGODNOSCI #####
# ROZKLAD NORMALNY

mean(x)
sd(x)
N=length(x)

lnL = function(parametry){
  mu = parametry[1]
  sigma = parametry[2]
  val = -N/2*log(2*pi)-N*log(sigma)-1/(2*sigma^2)*sum((x-mu)^2)
}

gradient = function(parametry){
  mu = parametry[1]
  sigma = parametry[2]
```

```
  gr = rep(0,times=length(parametry))
  gr[1] = sum(x-mu)/sigma^2
  gr[2] = -N/sigma+sum((x-mu)^2)/sigma^3
  return(gr)
}

hessian = function(parametry){
  mu = parametry[1]
  sigma = parametry[2]
  h = matrix(0, nrow=2, ncol=2)
  h[1,1] = -N/sigma^2
  h[2,2] = N/sigma^2-3*sum((x-mu)^2)/sigma^4
  h[1,2] = -2*sum(x-mu)/sigma^3
  h[2,1] = h[1,2]
  return(h)
}

wynik_norm_MNW = maxNR(fn = lnL, grad=gradient, hess = hessian, start=c(wynik_norm$root[1], wynik_norm$root[2]))
summary(wynik_norm_MNW)

ks.test(x, "pnorm", wynik_norm_MNW$estimate[1], wynik_norm_MNW$estimate[2])

# ROZKLAD LOGNORMALNY

lnl = function(parametry){
  mu = parametry[1]
  sigma = parametry[2]
  val = -sum(log(x)) - N/2 * log(2*pi) - N*log(sigma) - 1/(2*sigma^2)*sum((log(x) - mu)^2)
}

gradient = function(parametry) {
  mu = parametry[1]
  sigma = parametry[2]
  gr = rep(0,times=length(parametry))
  gr[1] = sum(log(x))/sigma^2 - (2*N*mu) / (2*sigma^2)
  gr[2] = -N/(2*sigma^2) + sum((log(x) - mu)^2) / (2*sigma^4)
  return(gr)
}

hessian = function(parametry){
  mu = parametry[1]
  sigma = parametry[2]
  h = matrix(0, nrow=2, ncol=2)
  h[1,1] = -N / sigma^2
  h[2,2] = N / (2*sigma^4) - 2* sum((log(x) - mu)^2) / (2*sigma^6)
  h[1,2] = 0
  h[2,1] = 0
  return(h)
}

wynik_lnorm_MNW = maxNR(lnl, gradient, hessian, start=c(wynik_lnorm$root[1],wynik_lnorm$root[2]))
summary(wynik_lnorm_MNW)

ks.test(x, "plnorm", wynik_lnorm_MNW$estimate[1], wynik_lnorm_MNW$estimate[2])

##### TESTOWANIE HIPOTEZ #####
## TEST Z
# NORMALNY
mu_norm_z = exp(1.95)
```

```
vcov = -solve(wynik_norm_MNW$hessian)
std.err.k = sqrt(vcov[1,1])
z.test = (wynik_norm_MNW$estimate[1]- mu_norm_z)/std.err.k
print(z.test)

p.value = 2*(1-pnorm(abs(z.test),mean=0,sd=1))
print(p.value)

#LOG-NORMALNY
mu_lognorm_z = 1.95

vcov = -solve(wynik_lnorm_MNW$hessian)
std.err.k = sqrt(vcov[1,1])
z.test = (wynik_lnorm_MNW$estimate[1]- mu_lognorm_z)/std.err.k
print(z.test)

p.value = 2*(1-pnorm(abs(z.test),mean=0,sd=1))
print(p.value)



## TEST LR
# NORMALNY

mu_test_norm = exp(1.95)
sigma_test_norm = exp(1.55)

lnL_U = wynik_norm_MNW$maximum
lnL_R = lnL(c(mu_test_norm,sigma_test_norm))
LR.test.norm = 2*(lnL_U-lnL_R)
print(LR.test.norm)

alpha = 0.05
g = 2
qchisq(1-alpha, df=g)
pchisq()
p.value = 1-pchisq(q = LR.test.norm, df = g)
print(p.value)

# LOGNORMALNY
mu_test_lnorm = 1.95
sigma_test_lnorm = 1.55

lnL_U = wynik_lnorm_MNW$maximum
lnL_R = lnl(c(mu_test_lnorm,sigma_test_lnorm))
LR.test.lnorm = 2*(lnL_U-lnL_R)
print(LR.test.lnorm)

p.value = 1-pchisq(q = LR.test.lnorm, df = g)
print(p.value)
```

Bibliografia

Field, J. B. F. (1985). A statistical study of the distribution of alcohol consumption and consequent inferential problems/by John BF Field (Doctoral dissertation).

Kehoe, T., Gmel, G., Shield, K.D. *et al.* Determining the best population-level alcohol consumption model and its impact on estimates of alcohol-attributable harms. *Popul Health Metrics* **10,** 6 (2012). https://doi.org/10.1186/1478-7954-10-6

Zofia Mielecka-Kubien (2018) On the estimation of the distribution of alcohol consumption, Mathematical Population Studies, 25:1, 1-19, DOI: 10.1080/08898480.2017.1348749