Szymon Szymankiewicz 151821

Zadanie Domowe

1. Opis zbioru

Dane, które wybrałem dotyczą ilości narodzin w USA w okresie od 2000 do 2014 roku. Dane pobrano z: https://github.com/fivethirtyeight/data/blob/master/births/US_births_2000-2014_SSA.csv. Są to dane obserwacyjne, które zostały zebrane przez amerykańską "Social Security Administration". Dane są dostępne jako pełna populacja i mają następującą strukturę:

Header	Definition
`year`	Year
`month`	Month
`date_of_month`	Day number of the month
`day_of_week`	Day of week, where 1 is Monday and 7 is Sunday
`births`	Number of births

Dysponując powyższą bazą danych postanowiłem wybrać dane z lat skrajnych (2000 oraz 2014) oraz wykonać analizę mającą na celu porównać średnią roczną ilość urodzeń w USA. Do wykonania testu zostały wykorzystane dane z roku 2000 oraz w ramach analizy została wykonana próba losowa z roku 2014 o rozmiarze 100 obserwacji. Próba losowa została wykonana poprzez interpreter R, w celu utrzymania powtarzalności wyników wygenerowanych liczb pseudolosowych została użyta funkcja set.seed().

```
set.seed(2014)
wielkosc_proby <- 100
proba_2014 <- sample(dane[dane$year == 2014, "births"], size = wielkosc_proby)
```

2. Analiza eksploracyjna

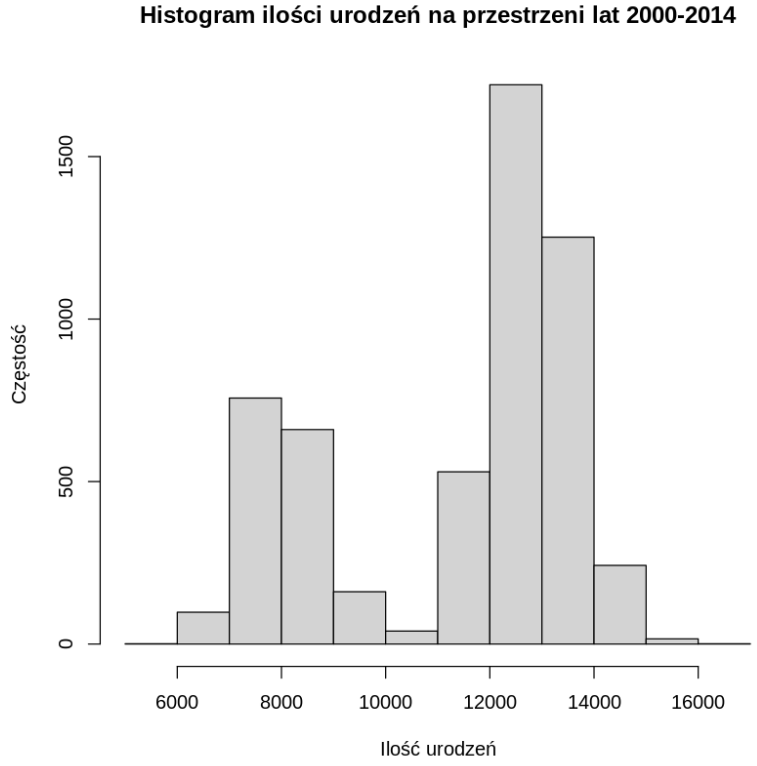
a. Podstawowe wartości statystyczne dla danych opisujących ilość urodzeń na przestrzeni lat 2000-2014:

```
summary(dane$births)
sd(dane$births)
hist(dane$births, main = "Histogram ilości urodzeń na przestrzeni lat 2000-2014", xlab = "Ilość urodzeń", ylab = "Częstość")
```

Wyniki:

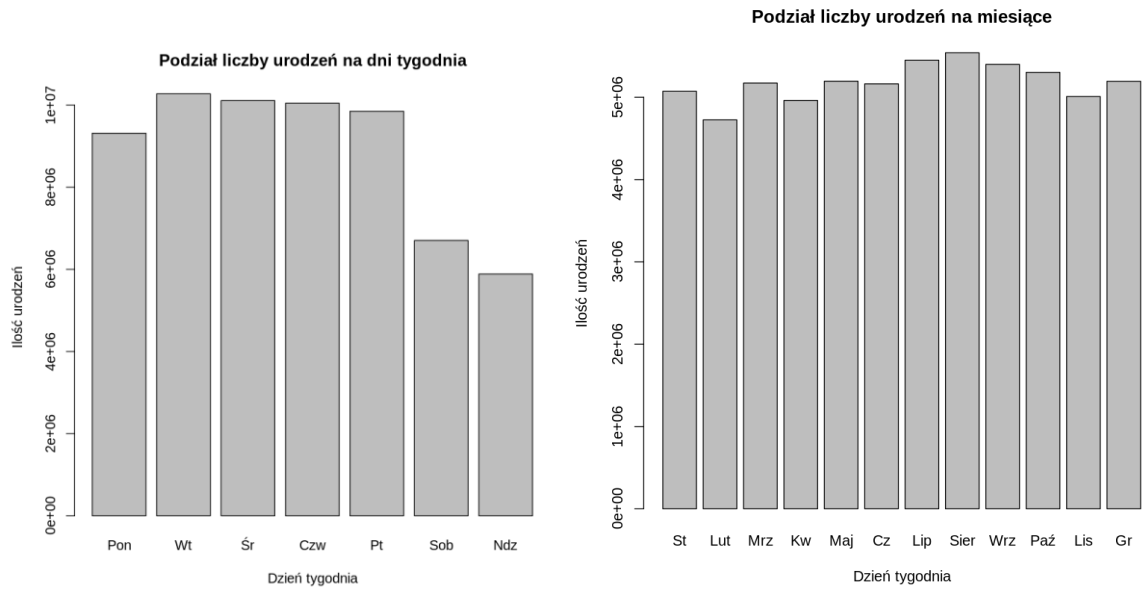
- Wartość minimalna: 5728
- Wartość maksymalna: 16081
- Wartość średnia: 11350
- Mediana: 12343
- Odchylenie standardowe: 2325,82
- Rozstęp międzykwartylowy: 4342

Aby łatwiej zobrazować, ile wynoszą i jak rozkładają się dane ilości urodzeń z lat 2000-2014 wykonany został histogram:



Na histogramie zauważyć można jeden mocno wybijający się szczyt, który przekracza nawet 16000 urodzeń w ciągu 1 dnia (po sprawdzeniu za pomocą interpretera R okazało się, że był to 3 września 2009 roku). Warte zaznaczenia jest również to, że mimo, iż wartość średnia oraz mediana nie różnią się znacznie jeżeli chodzi o wartość to między ich przedziałami występuje ogromna różnica w częstości występowania. Histogram pokazuje również, że występują bardzo duże wahania w liczbie urodzeń w ciągu jednego dnia.

W celu wykorzystania posiadanych danych oraz sprawdzenia, czy występuje zależność, która tłumaczy wspomniane wyżej wahania w interpreterze R wykonane zostały histogramy liczby urodzeń w zależności od dnia tygodnia oraz miesiąca:



Patrząc na wykres miesięcy nie można powiedzieć tutaj o żadnej występującej zależności. Wszystkie miesiące utrzymują się mniej więcej na tym samym poziomie z lekką górką w okolicach sierpnia. Jednak patrząc na wykres po lewej, gdzie ilość urodzeń została podzielona w zależności od dni tygodnia można zauważyć trend, w którym znacznie mniej dzieci jest rodzonych w weekend. Pozostałe dni tygodnia utrzymują się na podobnym poziomie. Między nimi, a sobotą i niedzielą występuje zauważalny spadek.

b. Podstawowe wartości statystyczne dla danych opisujących ilość urodzeń w roku 2000 zostały obliczone za pomocą interpretera R:

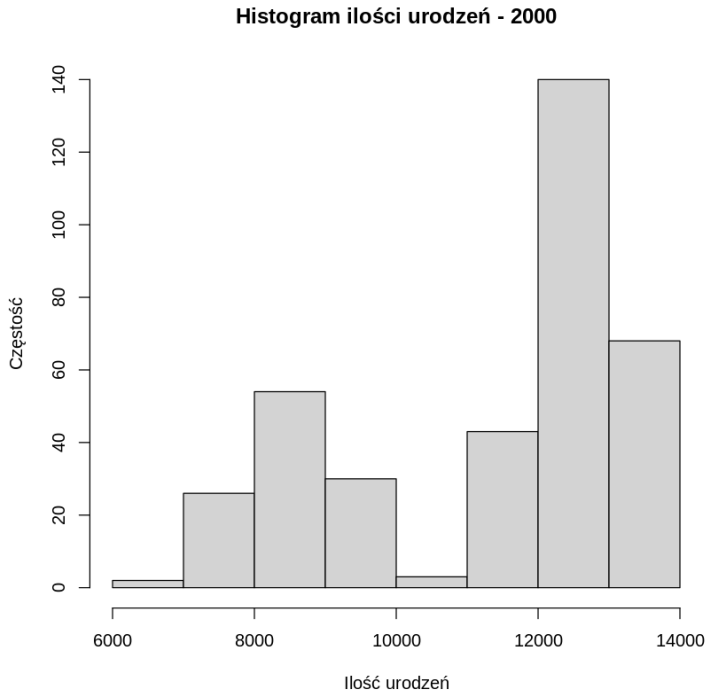
```
urodzeni_w_2000 <- subset(dane, year == 2000)

summary(urodzeni_w_2000$births)
sd(urodzeni_w_2000$births)
hist(urodzeni_w_2000$births, main = "Histogram ilości urodzeń - 2000", xlab = "Ilość urodzeń", ylab = "Częstość")
```

Wyniki:

- Wartość minimalna: 6719
- Wartość maksymalna: 13991
- Wartość średnia: 11337,7
- Mediana: 12240
- Odchylenie standardowe: 1978,11
- Rozstęp międzykwartylowy: 3665

Aby łatwiej zobrazować, ile wynoszą i jak rozkładają się dane ilości urodzeń w 2000 roku wykonany został histogram:



Jak widać powyżej histogram ma rozkład bipolarny. Charakteryzuje się on dwoma oddzielonymi skokami wartości. Warte zaznaczenia jest również to, że wartość średnia wynosząca 11337,7 oraz okoliczne wartości występują bardzo rzadko. Jest to spowodowane charakterystycznym rozkładem.

c. Podstawowe wartości statystyczne dla danych opisujących ilość urodzeń w próbie losowej pobranej z danych z 2014 roku:

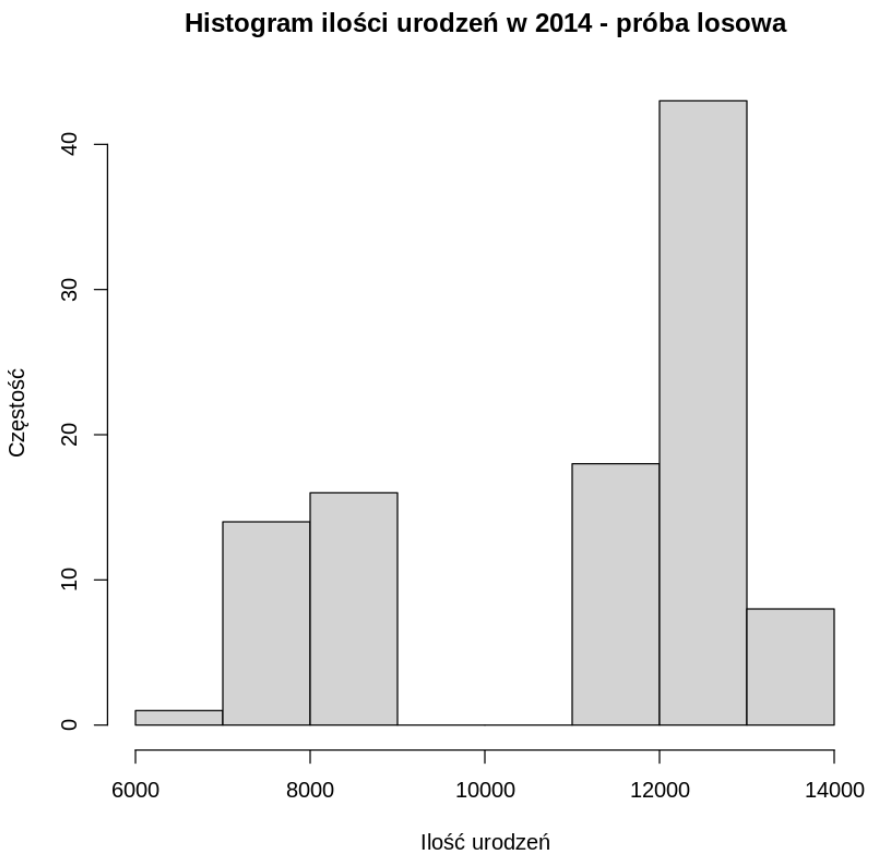
```
urodzeni_w_2014 <- subset(dane, year == 2014)

summary(urodzeni_w_2014$births)
sd(urodzeni_w_2014$births)
hist(urodzeni_w_2014$births, main = "Histogram ilości urodzeń - 2014", xlab = "Ilość urodzeń", ylab = "Częstość")
```

Wyniki:

- Wartość minimalna: 6973
- Wartość maksymalna: 13661
- Wartość średnia: 10966
- Mediana: 12010
- Odchylenie standardowe: 2138,68
- Rozstęp międzykwartylowy: 4054

Aby łatwiej zobrazować, ile wynoszą i jak rozkładają się dane ilości urodzeń z próbki w 2014 roku wykonany został histogram:



Jak można zauważyć, histogram w próbce wygląda na podobny rozkład jak w roku 2000, lecz tym razem w histogramie występuje 1 największy szczyt częstotliwości. Zakres ilości urodzeń nie zmienił się znacząco, a częstotliwość występowania wartości średniej spadła z bardzo niskiej do zerowej. Oznacza to, że żadna wartość w próbce nie jest równa wartości średniej.

3. Test statystyczny

Po wykonaniu analizy eksploracyjnej moim celem jest sprawdzenie prawdziwości stwierdzenia, że średnia ilość urodzeń w 2014 roku jest mniejsza niż średnia ilość urodzeń w roku 2000. Do testu wykorzystam dane z populacji z roku 2000 oraz przeanalizowaną wcześniej próbę losową z roku 2014. Do sprawdzenia tej hipotezy wykorzystam test Z dla średniej. Wynika to z tego, że wielkość próby jest wystarczająco duża (n>30), aby użyć testu Z zamiast testu T-Studenta. Poziom istotności α został ustalony na poziomie 5%.

Test Z

Hipoteza zerowa: Średnia liczba urodzeń w próbce z 2014 roku jest równa średniej liczbie urodzeń w populacji z 2000 roku.

$$H0: \mu_{2014} = \mu_{2000} = 11337,7$$

Hipoteza alternatywna: Średnia liczba urodzeń w próbce z 2014 roku jest mniejsza od średniej liczby urodzeń w populacji z 2000 roku.

Szymon Szymankiewicz 151821

H1: $\mu_{2014} < \mu_{2000}$

Przeprowadzenie testu:

Na podstawie danych populacji z 2000 roku oraz próbki losowej z 2014 roku przeprowadzono test t-Studenta dla średniej dla dwóch niezależnych prób:

Analiza wyników:

- Wartość statystyki: Z = -1,74

Obliczanie wartości statystyki zgodnie ze wzorem testu Z dla średnich:

$$Z = \frac{\overline{x} - \mu}{s} \cdot \sqrt{n},\ N{\sim}(0,1)$$

$$Z = \frac{10966 - 11337,7}{2138,68} \cdot \sqrt{100} \approx -1,74$$

- Liczba stopni swobody: df = 99

Obliczone ze wzoru df = n-1

- Obszar krytyczny dla przyjętej istotności 5%:

Dla podanego poziomu istotności dla rozkładu normalnego przedział krytyczny znajduje się w obszarze:

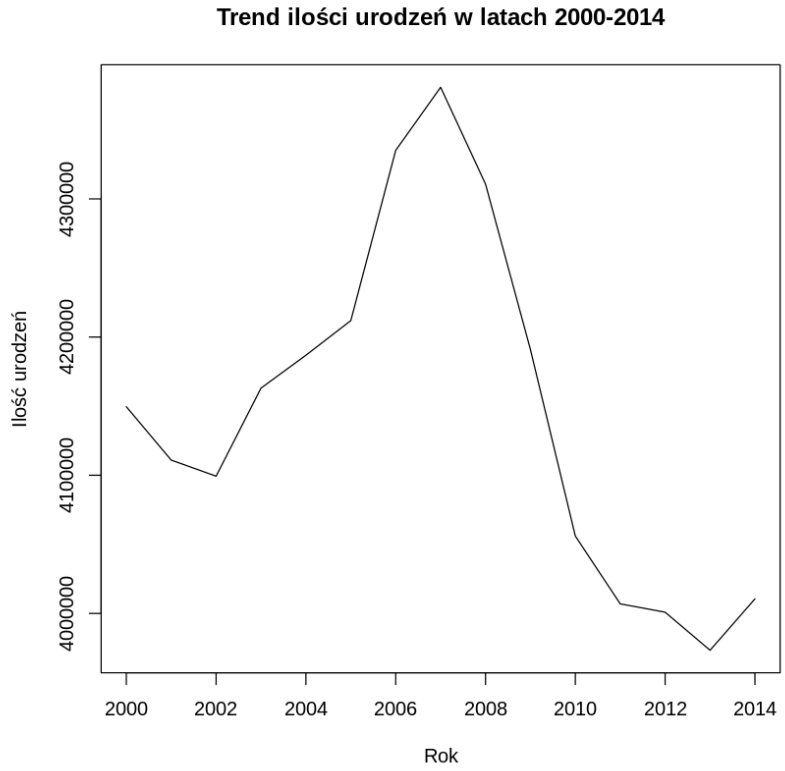
$$(-\infty\ ;\ -1,64)$$

Wnioski:

Po przeanalizowaniu wyników testu można stwierdzić, że odrzucamy hipotezę H0 mówiącą o równości średnich w obydwu latach na rzeczy hipotezy H1. Powodem jest to, że wartość statystyki znajduje się w obszarze krytycznym. Warto również wziąć pod uwagę to, że są to tylko dane pobrane z próby losowej oraz to, że sam test statystyczny nie daje pewności o prawdziwości lub fałszywości danej hipotezy. W tym przypadku oznacza to jedynie, że prawdopodobieństwo prawdziwości hipotezy zerowej jest bardzo niskie.

Jeżeli hipoteza H1 została przyjęta za prawdziwą, należy sprawdzić czy nie został popełniony błąd I rodzaju mówiący o tym, że odrzucona została prawdziwa hipoteza H0. Jako że dysponujemy danymi całej populacji z obydwu lat, łatwo jest sprawdzić czy utworzona próba losowa doprowadziła do

popełnienia błędu I rodzaju. Do sprawdzenia tego posłużyłem się wykresem wykonanym w interpreterze R:



Na wykresie dobrze widać jak na przestrzeni lat zmienia się sumaryczna ilość urodzeń. Maksymalna liczba urodzeń na rok wystąpiła w roku 2007, a jak można zauważyć różnica urodzeń między badanymi latami jest znacząca z przewagą w 2000 roku. Oznacza to, że błąd I rodzaju nie został popełniony.