Statystyka i Analiza Danych

Łukasz Jankowski
Wydział Informatyki i Telekomunikacji
Semestr IV, grupa I5.1
Nr albumu: 148081

Opis zbioru

Dane przedstawiają liczbę morderstw w latach 2014 i 2015 we miastach Stanów Zjednoczonych z liczbą mieszkańców powyżej 200 tysięcy, liczbę mieszkańców w każdym z miast oraz współczynnik morderstw na 100 tysięcy mieszkańców dla każdego z miast w poszczególnych latach. Dane pochodzą z FBI Uniform Crime Reports.

```
In [4]: (data <- read.csv("Dane.csv"))
```

A data.frame: 83 × 8

city	state	X2014_murders	X2015_murders	change	population	murder_rate_2014	murder_rate_2015
<chr>	<chr>	<int>	<int>	<int>	<int>	<dbl>	<dbl>
Baltimore	Maryland	211	344	133	620961	33.979590	55.398004
Chicago	Illinois	411	478	67	2763076	14.874727	17.299560
Houston	Texas	242	303	61	2160821	11.199447	14.022448
Cleveland	Ohio	63	120	57	372624	16.907124	32.204045
Washington	D.C.	105	162	57	705749	14.877811	22.954336
Milwaukee	Wisconsin	90	145	55	577222	15.591921	25.120318
Philadelphia	Pennsylvania	248	280	32	1569657	15.799630	17.838292
Kansas City	Missouri	78	109	31	495278	15.748731	22.007842
Nashville	Tennessee	41	72	31	689447	5.946795	10.443152
St. Louis	Missouri	159	188	29	300576	52.898435	62.546577
Oklahoma City	Oklahoma	45	73	28	681054	6.607406	10.718680
Louisville	Kentucky	56	81	25	596332	9.390742	13.583038
Denver	Colorado	31	53	22	715522	4.332501	7.407180
Los Angeles	California	260	282	22	3999759	6.500392	7.050425
Dallas	Texas	116	136	20	1304379	8.893121	10.426417
New York	New York	333	352	19	8336817	3.994330	4.222235
Orlando	Florida	15	32	17	287435	5.218571	11.132952
Minneapolis	Minnesota	31	47	16	382578	8.102923	12.285077
Omaha	Nebraska	32	48	16	408958	7.824764	11.737147
Sacramento	California	28	43	15	470956	5.945354	9.130365
Anchorage	Alaska	12	26	14	279671	4.290756	9.296638
Charlotte-Mecklenburg	North Carolina	47	61	14	731424	6.425821	8.339896
New Orleans	Louisiana	150	164	14	369250	40.622884	44.414353
Albuquerque	New Mexico	30	43	13	564599	5.313506	7.616025
Aurora	Colorado	11	24	13	386261	2.847815	6.213415
Fort Wayne	Indiana	12	25	13	223341	5.372950	11.193646
Long Beach	California	23	36	13	462257	4.975587	7.787876
Durham	North Carolina	21	34	13	217847	9.639793	15.607284
Indianapolis	Indiana	136	148	12	807584	16.840353	18.326267
Newark	New Jersey	93	104	11	279000	33.333333	37.275986
...	...	...	...	...	...	...	...
Chandler	Arizona	1	1	0	275987	0.3623359	0.3623359
Plano	Texas	4	4	0	285494	1.4010802	1.4010802
Stockton	California	49	49	0	291707	16.7976771	16.7976771
Toledo	Ohio	24	24	0	270871	8.8603062	8.8603062
Chula Vista	California	7	6	-1	275487	2.5409547	2.1779612
Phoenix	Arizona	114	112	-2	1608139	7.0889395	6.9645721
Riverside	California	12	10	-2	303871	3.9490442	3.2908701
San Jose	California	32	30	-2	1035317	3.0908408	2.8976632
Detroit	Michigan	298	295	-3	690074	43.1837745	42.7490385
Seattle	Washington	26	23	-3	713211	3.6454850	3.2248521
El Paso	Texas	21	17	-4	678815	3.0936264	2.5043642
Tucson	Arizona	35	31	-4	520116	6.7292681	5.9602089
Arlington	Texas	13	8	-5	394266	3.2972663	2.0290870
Lexington	Kentucky	20	15	-5	394266	5.0727174	3.8045381
Memphis	Tennessee	140	135	-5	633104	22.1132705	21.3235108
St. Petersburg	Florida	19	14	-5	265358	7.1601384	5.2758914
Columbus	Ohio	83	77	-6	902073	9.2010292	8.5358945
Honolulu	Hawaii	21	15	-6	343302	6.1170631	4.3693308
Laredo	Texas	14	8	-6	255205	5.4857859	3.1347348
Lincoln	Nebraska	7	1	-6	258379	2.7091985	0.3870284
Miami	Florida	81	75	-6	467963	17.3090608	16.0269081
Santa Ana	California	18	12	-6	334227	5.3855613	3.5903742
Mobile	Alabama	31	24	-7	188720	16.4264519	12.7172531
Fresno	California	47	39	-8	542107	8.6698751	7.1941517
Austin	Texas	32	23	-9	961855	3.3269048	2.3912128
San Antonio	Texas	103	94	-9	1434625	7.1795765	6.5522349
Corpus Christi	Texas	27	17	-10	317863	8.4942255	5.3482161
Pittsburgh	Pennsylvania	69	57	-12	302971	22.7744570	18.8136818
Boston	Massachusetts	53	38	-15	675647	7.8443329	5.6242387
Buffalo	New York	60	41	-19	278349	21.5556729	14.7297098

```
In [5]: city <- c(data$city)
        state <-c(data$state)
        murders_2014 <- c(data$X2014_murders)
        murders_2015 <- c(data$X2015_murders)
        diff <- c(data$change)
        population <- c(data$population)
        murders_rate_2014 <- c(data$murder_rate_2014)
        murders_rate_2015  <- c(data$murder_rate_2015)
        diff_rate <- murders_rate_2015 - murders_rate_2014
```

Analiza eksploracyjna

Miary położenia

```
In [6]: summary(murders_rate_2014)
        summary(murders_rate_2015)
        summary(diff_rate)

        (city[murders_rate_2014 == min(murders_rate_2014)])
        (city[murders_rate_2014 == max(murders_rate_2014)])

        (city[murders_rate_2015 == min(murders_rate_2015)])
        (city[murders_rate_2015 == max(murders_rate_2015)])


        (city[diff_rate == min(diff_rate)])
        (city[diff_rate == max(diff_rate)])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   4.170   6.729  10.248  14.876  52.898 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3623  4.7900  8.3399 11.7943 15.1685 62.5466 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-6.8260 -0.5120  0.6959  1.5467  2.8176 21.4184 
```

'Irvine'
'St. Louis'
'Chandler'
'St. Louis'
'Buffalo'
'Baltimore'

W roku 2014 średni współczynnik morderstw na 100 tys mieszkańców wynosi 10,248. Najbezpieczniejszym miastem było Irvine ze współczynnikiem 0, a najmniej bezpicznym St.Louis ze współczynnikiem 52.898 morderstw na 100 tyś mieszkańców.

W roku 2015 średni współczynnik morderstw na 100 tys mieszkańców wynosi 11.7943 . Najbezpieczniejszym miastem było Chandler ze współczynnikiem 0.3623, a najmniej bezpicznym St.Louis ze współczynnikiem 62.5466 morderstw na 100 tyś mieszkańców.

Średnia zmiana współczynnika morderstw ukształtowała się na poziomie 1,5467 co oznacza wzrost wskaźnika morderstw. Z kolei największy spadek liczby morderstw odnotowano w mieście Buffalo (-6,826), zaś najwiekszy wzrost charakteryzowało miasto Baltimore (21.4184).

Miary zmienności

```
In [7]: (var_murders_rate_2014 <- var(murders_rate_2014))
        (sd_murders_rate_2014 <- sd(murders_rate_2014))

        (var_murders_rate_2015 <- var(murders_rate_2015))
        (sd_murders_rate_2015 <- sd(murders_rate_2015))

        (var_diff_rate <- var(diff_rate))
        (sd_diff_rate <- sd(diff_rate))
```

94.6337292035006
9.727986903954
134.95518629078
11.6170214035604
15.2420465973317
3.90410637628276

Wariancja współczynnika morderstw w 2014 roku wynosi 94,63. Odchylenie standardowe - 9,728

Wariancja współczynnika morderstw w 2015 roku wynosi 134,95. Odchylenie standardowe - 11,617

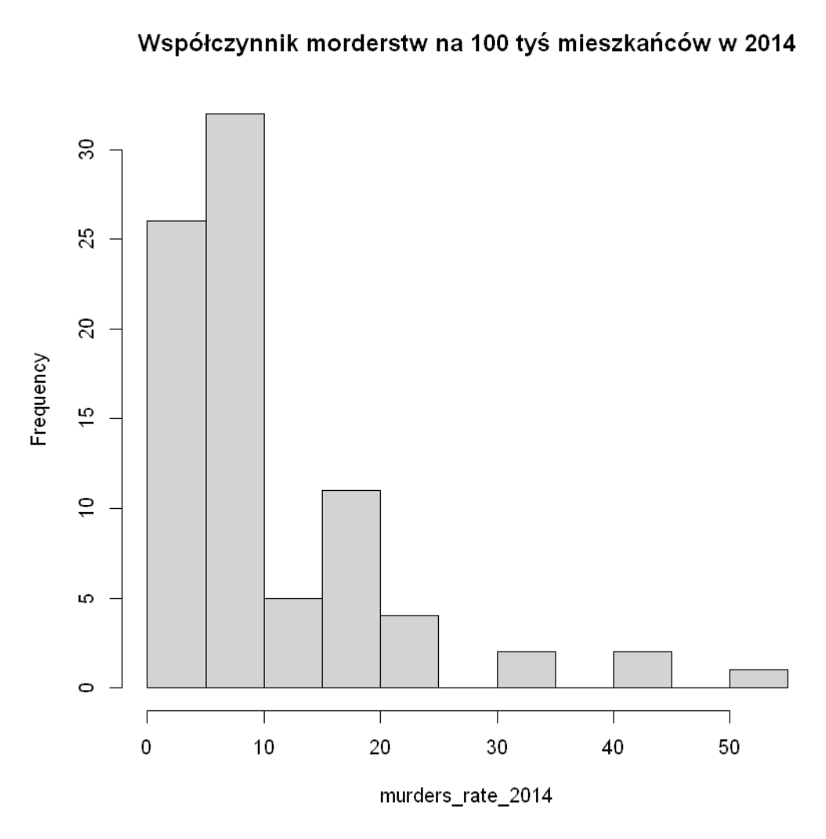
Wariancja różnicy współczynnika morderstw wynosi 15,242. Odchylenie standardowe - 3,904

Miary skośności

Liczbę przedziałów dla szerego rozdzielczego obliczam za pomocą wzoru $k = \sqrt{n}$, gdzie n to liczba badanych miast.

```
In [8]: k <- round(sqrt(nrow(data)))
        hist(murders_rate_2014, breaks=k, main="Współczynnik morderstw na 100 tyś mieszkańców w 2014")
        summary(murders_rate_2014)
        summary(murders_rate_2015)
        summary(diff_rate)
```

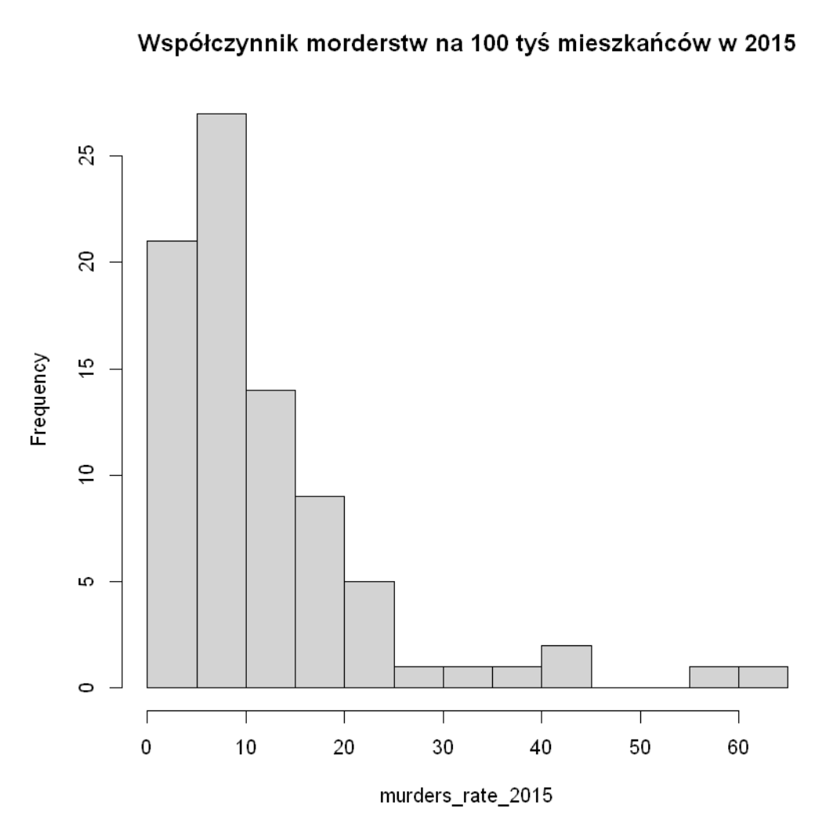
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   4.170   6.729  10.248  14.876  52.898 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3623  4.7900  8.3399 11.7943 15.1685 62.5466 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-6.8260 -0.5120  0.6959  1.5467  2.8176 21.4184 
```



Szereg o dodatniej asymetrii, mediana mniejsza od średniej, większość danych skupia się wokół wartości od 0 do 10. Mamy również kilka wartości oddstających, które charakteryzują się nawet kilkukrotnie wyższym współczynnikiem morderstw niż średni współczynnik

miasta, w których liczba morderstw jest na 100 tyś mieszkańców znacznie odbiega od wartości przeciętnych.

```
In [9]: hist(murders_rate_2015, breaks=k, main="Współczynnik morderstw na 100 tyś mieszkańców w 2015")
```



Również mamy do czynienia z szeregiem prawostronnie skośnym oraz mamy kilka wartości odstających.

```
In [10]: hist(diff_rate, breaks=k, main="Zmiana wskaźnika morderstw")
```

Większość danych jest w przedziale od -5 do 5. Wykres jest w dużym przybliżeniu symetryczny i zbliżony do rozkładu normalnego.

Test statystyczny

Test ma na celu zbadanie czy ilość morderstw w Stanach Zjednoczonych 2015 zmieniła się w stosunku do roku 2014. Przyjmuję poziom istotności na poziomie 5%.

Liczba morderstw w roku 2014 i 2015 w pojedynczym mieście jest od siebie zależna, natomiast ilość morderstw w różnych miastach nie jest - mamy do czynienia z danymi sparowanymi. Aby pozbyć się zależności wyznaczam różnicę między liczbą morderstw w danym mieście w roku 2015 i 2014.

```
In [11]: diffrence <- murders_2015 - murders_2014
```

Wielkość próby wynosi 83 więc zgodnie z Centralnym Twierdzeniem Granicznym rozkład będzie dążył do rozkładu normalnego. Również mamy nieznaną wariancję w populacji więc wyestymuję ją z próbki. Założenie testu T-studenta są spełnione więc mogę z niego skorzystać.

H_0: $\mu_{diff} = 0$

H_1: $\mu_{diff} \neq 0$

$T_{n-1} = \frac{\bar{X}-\mu_{diff}}{S}\sqrt{n}$

```
In [12]: n = length(diffrence)
         X = mean(diffrence)
         S = sd(diffrence)
         (t = X * sqrt(n) / S)

         alpha = 0.05
         # test dwustronny
         zb_kryt_l <- qt(p=alpha/2, df=n-1)
         zb_kryt_p <- qt(p=(1 - alpha/2), df=n-1)
```

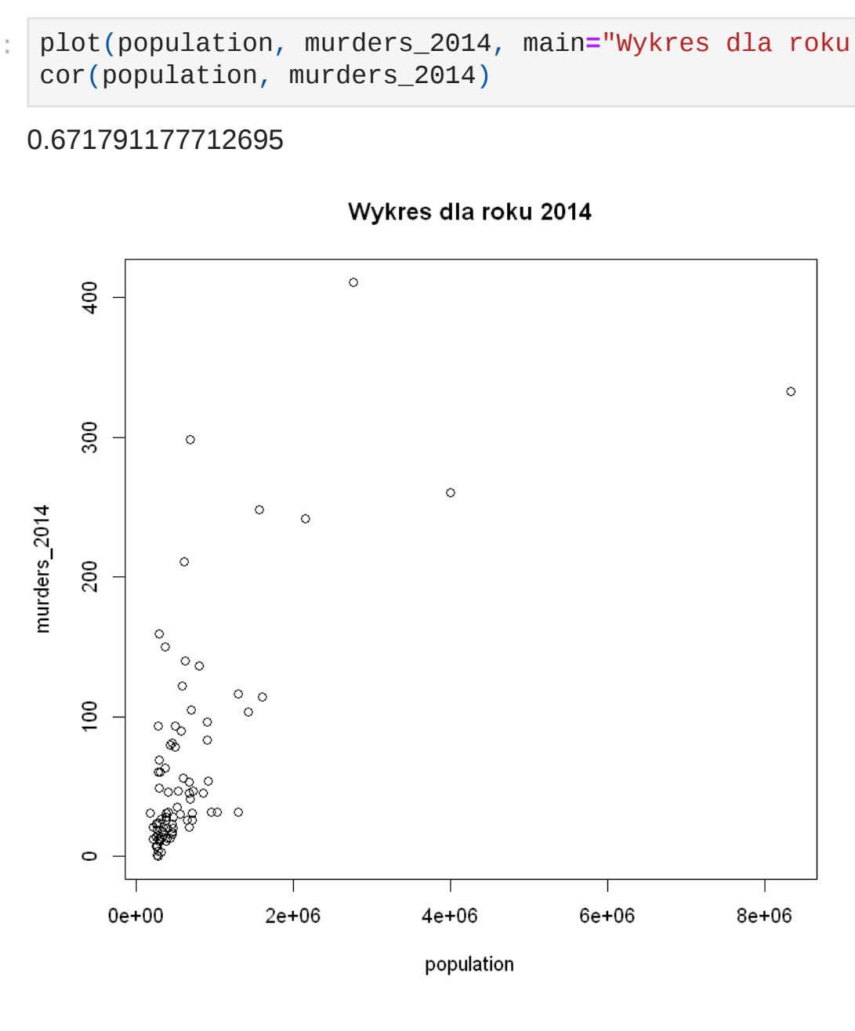
4.05738379300757

```
In [13]: x <- seq(-4, 4, length=100)
         y <- dt(x, df=n-1)

         #plot x and y as a scatterplot with connected lines (type = "l") and add
         #an x-axis with custom labels
         plot(x,y, type = "l", lwd = 2,, xlab = "x", ylab = "probability")
         abline(v=zb_kryt_l )
         abline(v=zb_kryt_p)

         x2 <- seq(-4,zb_kryt_l, 0.01)
         y2 <-  dt(x2, df=n-1)
         x2 = c(-4,x2,zb_kryt_l)
         y2 = c(0,y2,0)
         polygon(x2,y2, col="red", border=NA)

         x3 <- seq(zb_kryt_p, 4, 0.01)
         y3 <-  dt(x3, df=n-1)
         x3 = c(zb_kryt_p,x3, 4)
         y3 = c(0,y3,0)
         polygon(x3,y3, col="red", border=NA)
```



$T_{82} = 4.05738379300757$

Zbiór krytyczny - $(-\infty, -1,989) + (1,989, +\infty)$

Obszar krytyczny został oznaczony na wykresie kolorem czerwonym.

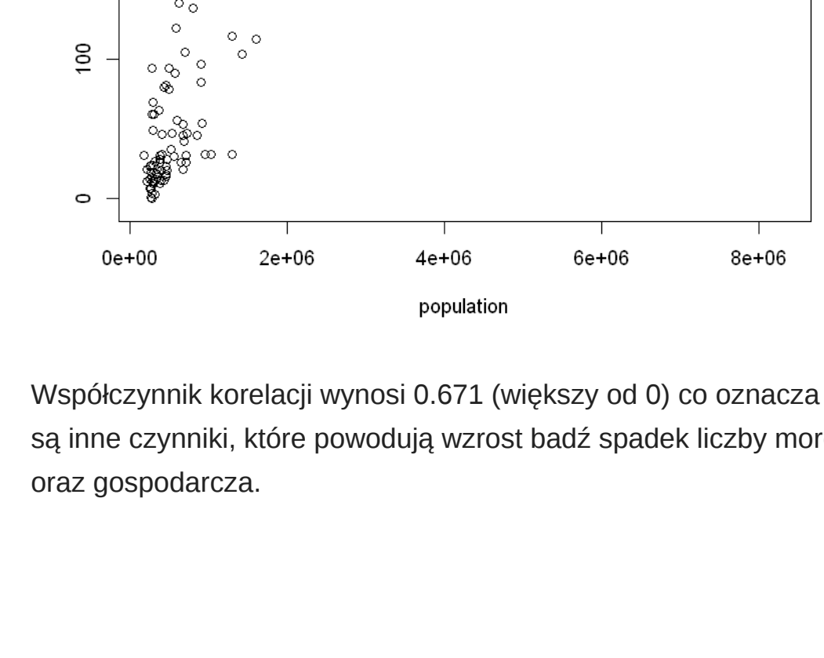
Statystyka T znajduje się w zbiorze krytycznym, zatem odrzucam hipotezę zerową i przyjmuję aternatywną. Liczba morderstw zmieniła się w 2015 roku w stosunku do roku 2014.

Analiza regresji

Celem badania będzie weryfikacja czy liczba morderstw jest zależna liniowo od liczby mieszkańców w danym mieście.

```
In [21]: plot(population, murders_2014, main="Wykres dla roku 2014")
         cor(population, murders_2014)
```

0.671791177712695



Współczynnik korelacji wynosi 0.671 (większy od 0) co oznacza że występuje pewna zależność, im więcej ludzi zamieszkuje dany teren tym więcej odnotowujemy tam zabójstw. Oczywiście nie jest to zależność liniowa, gdyż są inne czynniki, które powodują wzrost badź spadek liczby morderstw. Wpływ ma przede wszystkim infrastruktura ekonomiczna (bezrobocie, inflacja), społeczno-kulturowa(poziom życia, edukacja, styl życia, normy etyczne) oraz gospodarcza.