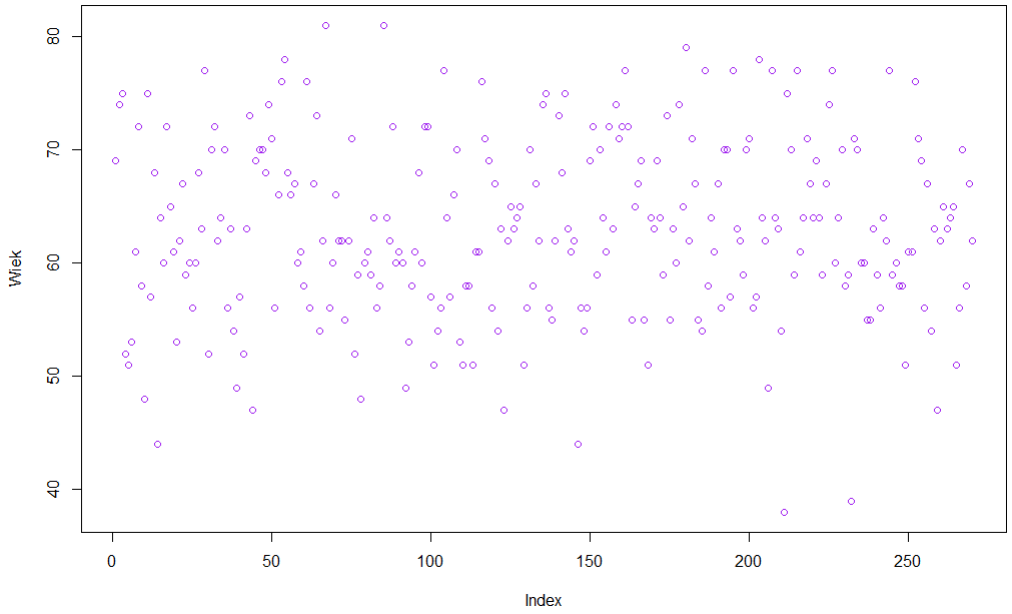
Statystyka Stosowana

Kierunek, nazwa wydziału	Grupa ćwiczeniowa, termin zajęć
Matematyka Stosowana, Wydział Matematyki	T00-64a Wtorek 7:30
Imię, nazwisko, numer albumu	**Data oddania**
Małgorzata Kowalczyk, 262295 Julia Mazur, 262296	15.05.2022 r.
Tytuł	**Prowadzący**
Analiza danych rzeczywistych przy pomocy metod statystyki opisowej	Dr inż. Aleksandra Grzesiek

Spis treści

1 Wstęp

Raport dotyczy analizy wieku ludzi, którzy zachorowali na raka płuc. Dane te pobrano ze strony https://www.kaggle.com/. Wśród nich, zawarte są osoby, które zachorowały, oraz te, które mimo czynników zwiększających ryzyko choroby, pozostały zdrowe. W związku z tym, że celem było badanie wieku chorych, usunięto z pliku osoby zdrowe. Zatem do analizy zostało poddane 270 osób. Zbadano podstawowe statystyki dla wieku, a następnie zwizualizowano dane. Sprawdzono także jak palenie papierosów oraz płeć wpływa na zachorowanie. Wszystkie wyniki zaprezentowane poniżej otrzymano przy pomocy języka R. Podczas przygotowania raportu korzystano z wykładów dr hab. inż. Krzysztofa Burneckiego.



Rysunek 1: Wykres przedstawiający dane dotyczące wieku zachorowania na raka płuc

2 Podstawowe statystyki

2.1 Miary położenia

2.1.1 Średnia arytmetyczna

Średnia arytmetyczna to suma wartości wszystkich badanych obiektów podzielona przez liczbę tych obiektów. Informuje o wartości średniej danego zbioru liczb. Jest ona najbardziej intuicyjna oraz powszechnie używana.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\bar{x} = 62.95185 \approx 62.95$$

Średni wiek zachorowania wynosi 62.95.

2.1.2 Mediana

Mediana to wartość środkowa. Dzieli dane na dwa podzbiory (powyżej i poniżej jej wartości) o równej liczbie obserwacji. Ustalana jest po uporządkowaniu danych w kolejności niemalejącej. Następnie, jeśli:

- n nie jest podzielne przez 2, to medianą jest wartość $x_{\left(\left[\frac{n}{2}\right]+1\right)}$;
- n jest podzielne przez 2, to medianą jest średnia arytmetyczna wartości $x_{\left(\frac{n}{2}\right)}$ i $x_{\left(\frac{n}{2}+1\right)}$.

$$Me = 62.5$$

Mediana wynosi 62.5. Oznacza to, że połowa chorych w badanym zbiorze danych jest w tym lub niższym wieku.

2.1.3 Dominanta

Dominanta inaczej nazywana jest modą. Jest to wartość, która w zbiorze danych występuje najczęściej. Jeżeli w zestawie występuje kilka danych z najwyższą częstością, to każda z nich jest dominantą. Jeżeli wszystkie dane występują z tą samą liczebnością, to zestaw danych nie ma dominanty.

$$M = 64$$

Dominanta wynosi 64, a więc jest to najczęściej występujący wiek zachorowania wśród ludzi w badanym zbiorze danych.

2.1.4 Kwartyle

Kwartyle są miarą położenia. Dzielą one zbiór danych na 4 równe, co do ilości elementów, grupy. Wyróżnia się kwartyl pierwszy, kwartyl drugi i kwartyl trzeci.

- Kwartyl pierwszy (Q_1) to mediana grupy obserwacji mniejszych od Q_2. Dzieli on uporządkowane niemalejąco dane na dwie części, w taki sposób, że 25% obserwacji ma wartości mniejsze lub równe kwartylowi pierwszemu a 75% równe lub większe od tego kwartyla.
- Kwartyl drugi (Q_2) dzieli uporządkowane niemalejąco dane na dwie równe części. Z tego powodu jest też nazywany medianą.
- Kwartyl trzeci (Q_3) to mediana grupy obserwacji większych od Q_2. Dzieli on uporządkowane niemalejąco dane na dwie części, w taki sposób, że 75% obserwacji ma wartości mniejsze lub równe kwartylowi trzeciemu a 25% równe lub większe od tego kwartyla.

$$Q_1 = 58 \qquad Q_2 = 62.5 \qquad Q_3 = 69$$

Zatem Q_1 informuje o tym, że 25% osób zachorowało przed lub w 58 roku życia, a 75% osób w późniejszych latach. Natomiast dzięki Q_3 wiadomo, że 75% osób zachorowało przed lub w 69 roku życia, a 25% później.

2.2 Miary rozproszenia

2.2.1 Rozstęp międzykwartylowy

Rozstęp międzykwartylowy to różnica między trzecim a pierwszym kwartylem.

$$IQR = Q_3 - Q_1 \tag{2}$$

Dla badanych danych $Q_3 = 69$ oraz $Q_1 = 58$.

$$IQR = 69 - 58 = 11$$

Różnica między medianą pierwszej i drugiej połowy uporządkowanych danych wynosi 11.

2.2.2 Rozstęp

Rozstęp to róznica między największą i najmniejszą wartością z danych.

$$R = x_{max} - x_{min} \tag{3}$$

Dla badanych danych $x_{max} = 81$ oraz $x_{min} = 38$.

$$R = 81 - 38 = 43$$

Zakres danych wynosi 43.

2.2.3 Wariancja

Wariancja jest średnią arytmetyczną kwadratów odchyleń od wartości średniej. Estymator nieobciążony wariancji jest postaci:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{4}$$

$$S^2 = 63.46607 \approx 63.47$$

Wariancja dla badanego zbioru danych wynosi 63.47, co świadczy o tym, że rozproszenie danych jest duże.

2.2.4 Odchylenie standardowe

Odchylenie standardowe jest pierwiastkiem z wariancji. Informuje o tym, jak szeroko wartości danej wielkości są rozrzucone wokół jej średniej.

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5}$$

$$S = 7.96656 \approx 7.97$$

Badany wiek zachorowania na raka płuc, różni się od przeciętnego średniego wieku osoby chorej o 7.97. Pozwala to stwierdzić, że badane dane różnią się między sobą.

2.2.5 Współczynnik zmienności

Współczynnik zmienności służy do badania stopnia zróżnicowania wartości zmiennej.

$$\nu = \frac{S}{\bar{x}} \cdot 100\% \tag{6}$$

$$\nu = 12.655\% \approx 13\%$$

Dla badanych danych współczynnik zmienności wynosi 13%. W związku z tym, można wnioskować, że dane charakteryzuje mała zmienność.

2.3 Miary asymetrii

2.3.1 Współczynnik skośności (współczynnik asymetrii)

Współczynnik skośności informuje o tym, jak wygląda rozkład i określa sposób, w jaki dane są rozłożone po obu stronach średniej.

$$\alpha = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S} \right)^3 \tag{7}$$

$$\alpha = -0.1047177 \approx -0.10$$

Dla badanych danych współczynnik skośności wynosi -0.10, co oznacza, że otrzymano niewielką lewostronną skośność. Pozwala to wysunąc wniosek, że w badanej grupie występowało więcej osób starszych.

2.4 Miary spłaszczenia

2.4.1 Współczynnik skupienia (kurtoza)

Kurtoza określa rozmieszczenie i koncentrację zbiorowości w pobliżu średniej.

$$K = \frac{\mu_4}{S^4} = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} \tag{8}$$

W ramach udogodnienia sprowadzono wzór do postaci, dla której kurtoza rozkładu normalnego przyjmuje wartość 0. Wzór ten nosi nazwę współczynnika ekscesu i pozwala porównać badane dane z danymi z rozkładu normalnego.

$$K = \frac{\mu_4}{S^4} - 3 \tag{9}$$

$$K = -0.1787868 \approx -0.18$$

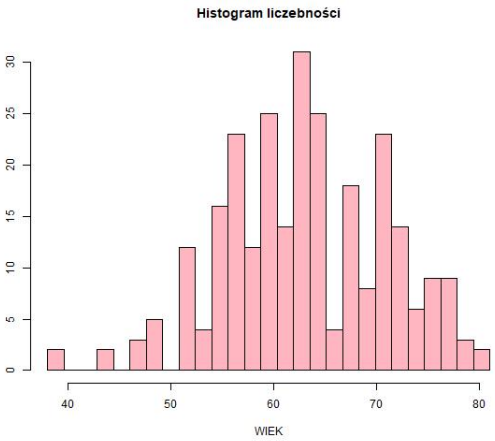
Kurtoza dla badanych danych osiągnęła wartość ujemną równą -0.18, co oznacza, że rozkład jest mniej wysmukły niż normalny oraz występuje w danych spora część wyników, które są znacznie oddalone od wartości średniej.

3 Wizualizacja danych

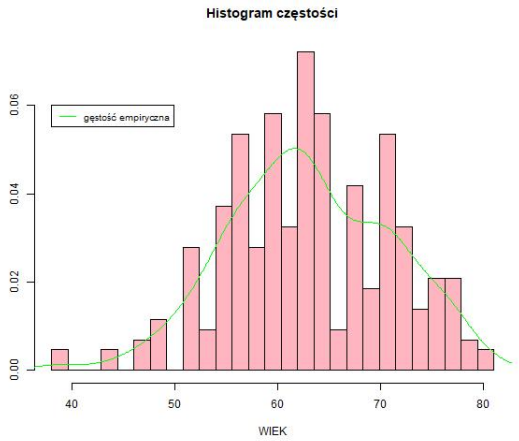
3.1 Histogram i gęstość empiryczna

Histogram to zestawienie danych, ilustrujące występowanie określonej cechy w badanych danych w zależności od podziałów klasowych. Wyróżniamy histogramy:

- liczebności, w których wartości na osi pionowej określają faktyczne liczby obiektów wykazujących daną cechę (Rysunek 2);
- częstości, w których wartości na osi Y określają ich częstości (Rysunek 3).



Rysunek 2: Histogram liczebności wieku zachorowań na raka płuc

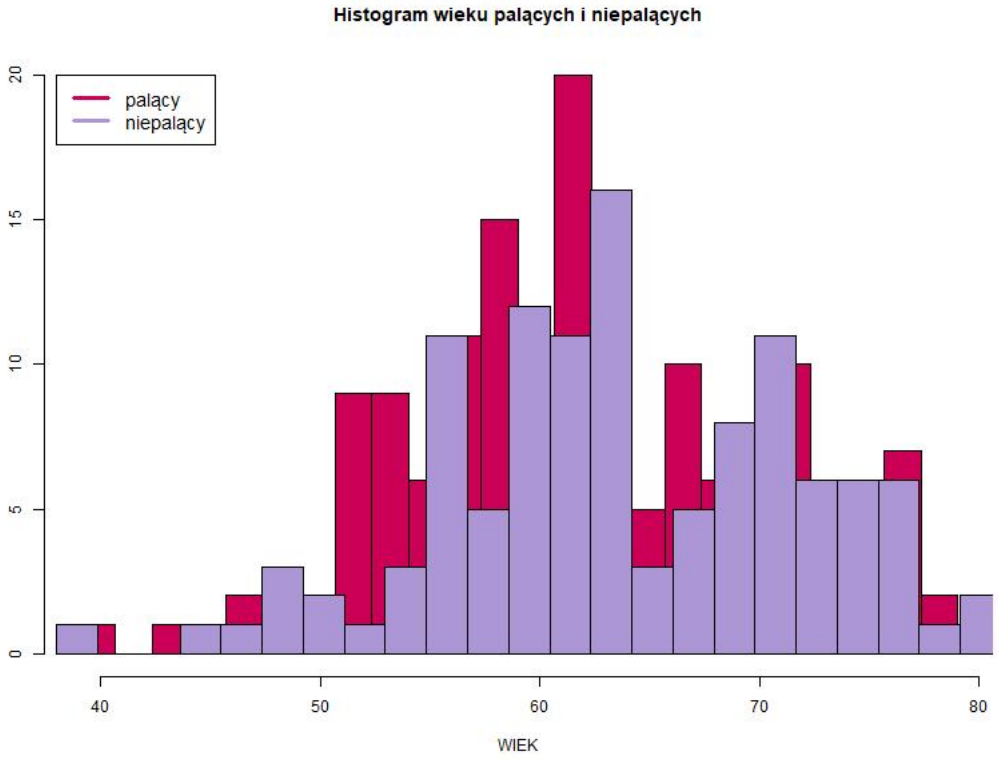


Rysunek 3: Histogram częstości wieku zachorowań na raka płuc wraz z gęstością empiryczną

Patrząc na powyższe histogramy, można zauważyć, że największa wartość jest osiągana dla wieku zbliżonego do wartości średniej arytmetycznej i mediany. Najliczniejszym przedziałem jest przedział od około 63 do 64.5 i to w nim znajduje się dominanta uzyskana w punkcie (2.1.3). Obserwacje nie są rozłożone równomiernie, co jest dość zrozumiałe w przypadku wieku zachorowania. Dodatkowo widać, że faktycznie większość danych różni się od wartości średniej o nie więcej niż wartość odchylenia standardowego wyliczonego w punkcie (2.2.4). Wykres jest delikatnie lewostronnie skośny i mniej wysmukły, co potwierdza wyniki otrzymane odpowiednio w punktach (2.3.1) i (2.4.1).

	Osoba paląca	Osoba niepaląca
Liczba danych	155	115
Średnia wieku	62.34	63.77

Tabela 1: Tabela wartości średniej wieku zachorowania na raka płuc z podziałem na osoby palące i niepalące

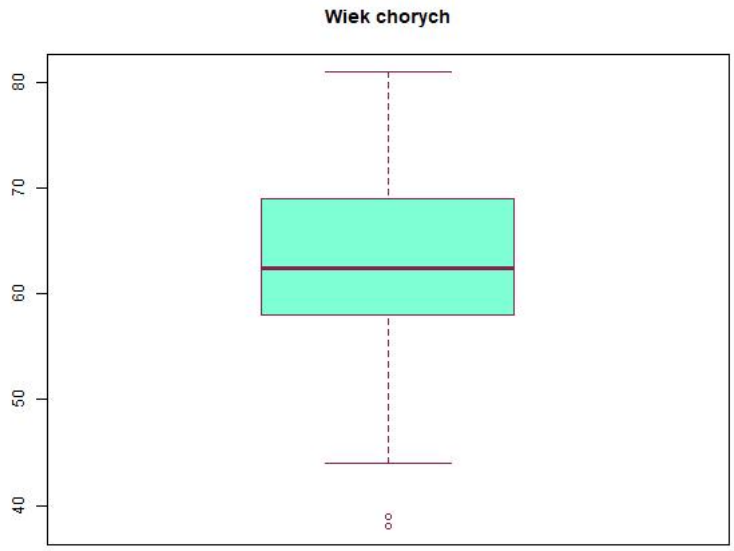


Rysunek 4: Histogram liczebności wieku zachorowania na raka płuc z podziałem na palących i niepalących

Po dokonaniu podziału na chorych palących i niepalących następuje niewielka zmiana w średnich dla badanych danych. Można zauważyć delikatne przesunięcie histogramu dla chorych palących, wskazujące na szybszy rozwój choroby. Główną różnicą jest jednak liczba zachorowań, która w przypadku osób palących znacznie wzrasta, sugerując negatywny wpływ palenia na zdrowie człowieka.

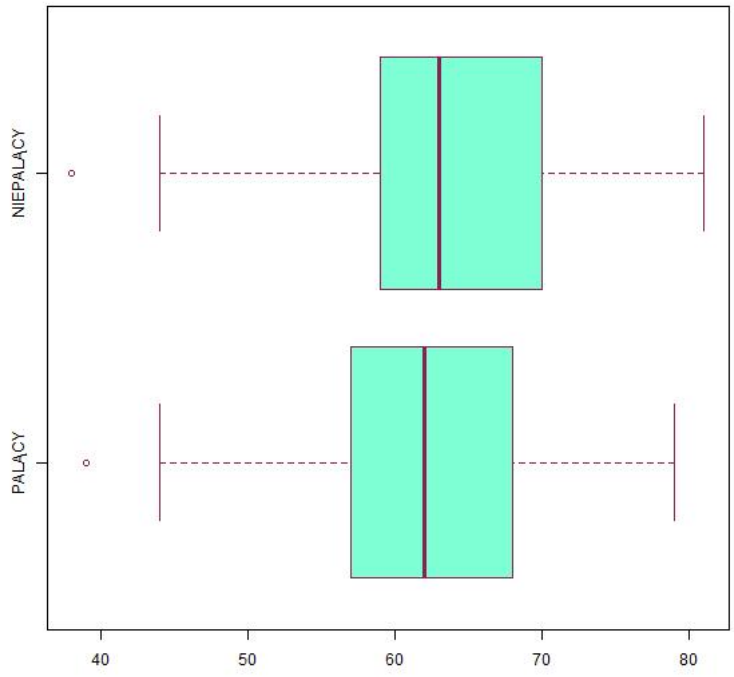
3.2 Wykres pudełkowy

Wykres pudełkowy jest graficzną interpretacją mediany, kwartyli pierwszego i trzeciego oraz minimum i maksimum z danych. Pozwala na przedstawienie statystyk opisowych, a także empirycznego rozkładu danej cechy na jednym rysunku. Można dzięki niemu określić m.in. rozproszenie i symetrię danych.



Rysunek 5: Wykres pudełkowy wieku badanych chorych

Pozioma linia, znajdująca się wewnątrz prostokąta oznacza medianę. Długość prostokąta reprezentuje rozstęp międzykwartylowy. Pudełko nie jest równo podzielone, co oznacza, że występuje asymetria danych. Obserwacje odstające znajdują się poniżej pudełka, zatem odległość mediany od minimum jest większa, niż odległość do maksimum. Potwierdza to otrzymane wyniki w punkcie (2.3.1) wskazujące na lewostronną skośność. Wąsy łączą pudełko z najwyższą i najniższą wartością nieodstającą. Z ich długości wynika, że dane są rozproszone. Górny wąs zaznaczający wartość maksymalną informuje o braku wartości odstających większych od $Q_3 + 1.5IQR$. Pojawiły się dwie wartości odstające mniejsze niż $Q_1 - 1.5IQR$, co wskazuje na to, że są nieliczne zachorowania osób znacznie młodszych od wartości średniej wieku.

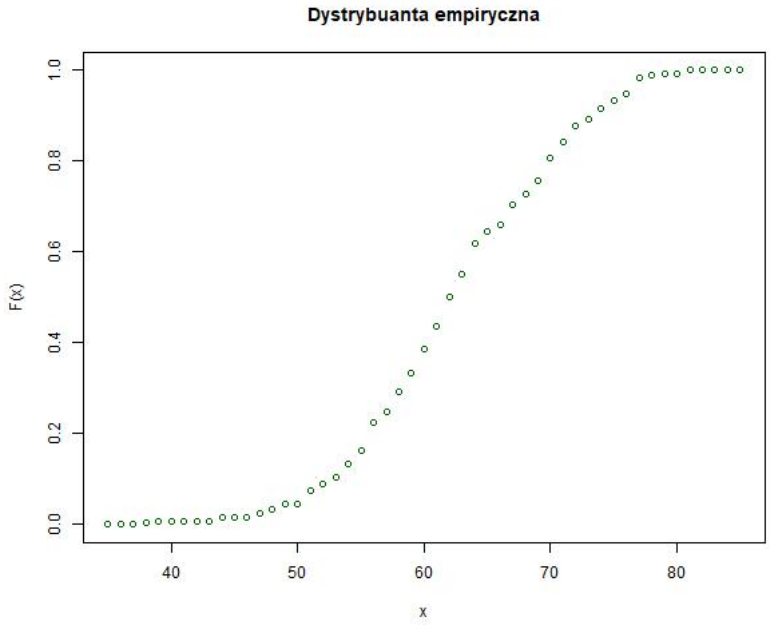


Rysunek 6: Wykres pudełkowy wieku z podziałem na palących i niepalących

W obu przypadkach występują wartości odstające, co oznacza, że zachorowania nie zależą jedynie od palenia. Można jednak zauważyć delikatne zmniejszenie mediany wieku dla chorych palących. Mniejsza maksymalna wartość wieku może świadczyć o krótszym okresie życia, związanym z późniejszym wykryciem nowotworu (np. w późniejszym stadium) ze względu na to, że niektóre skutki uboczne palenia pokrywają się z objawami raka płuc. To również tłumaczy podobny wiek badanych chorych niezależnie od stosowania rozważanej używki.

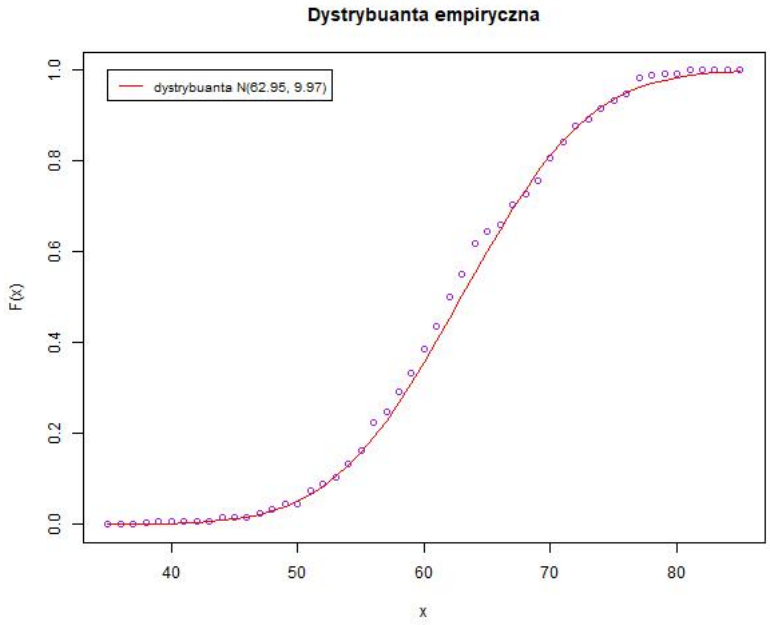
3.3 Wykres dystrybuanty empirycznej

Dystrybuanta empiryczna pozwala wyznaczyć częstość zdarzenia takiego, że obserwacje w próbce są mniejsze od wartości x.



Rysunek 7: Wykres dystrybuanty empirycznej wieku chorych

Można porównać otrzymaną dystrybuantę z teoretyczną pasującą do sytuacji. Jako że jest to mniej więcej symetryczny rozkład wieku, to można próbować dopasować dystrybuantę rozkładu normalnego o parametrach $\mu = \bar{x} = 62.95$ oraz $\sigma = S = 7.97$.



Rysunek 8: Wykres dystrybuanty empirycznej wieku chorych z dopasowaną teoretyczną dystrybuantą rozkładu $\mathcal{N}(\mu = 62.95, \sigma = 7.97)$

Powyższe dopasowanie jest sensowne, gdyż na postawie wykresu widać, że dystrybuanta teoretyczna rozkładu $\mathcal{N}(\mu = 62.95, \sigma = 7.97)$ bardzo dobrze pokrywa się z otrzymaną dystrybuantą empiryczną. Oznacza to, że dane mają rozkład zbliżony do rozkładu normalnego, czyli najwięcej danych znajduje się w okolicy średniej, co tłumaczy zbliżone wartości średniej arytmetycznej i mediany.

4 Porównanie statystyk ze względu na palenie i płeć

	Niepalący mężczyzna	Niepaląca kobieta	Palący mężczyzna	Paląca kobieta
Średnia	63.25	64.35	62.99	61.56
Mediana	63.00	64.00	62.00	61.00
Moda	63.00	56.00	58.00	61.00
Wariancja	50.36	86.42	59.80	60.34
Odchylenie standardowe	7.10	9.30	7.73	7.77
Kwartyl pierwszy	59.75	57.00	58.00	55.00
Kwartyl drugi	63.00	64.00	62.00	61.00
Kwartyl trzeci	69.00	71.50	69.00	67.00
Rozstęp międzykwartylowy	9.25	14.50	11.00	12.00
Rozstęp	34.00	43.00	40.00	34.00
Współczynnik zmienności	11%	14 %	12%	13%
Kurtoza	0.17	-0.20	0.06	-0.55
Współczynnik skośności	-0.35	-0.28	-0.08	0.07

Tabela 2: Podstawowe statystyki z podziałem na podgrupy

Na podstawie wyliczonych statystyk można stwierdzić, że palenie wpływa na zachorowanie na raka płuc. Różnice te nie są dość duże, jednakże należy pamiętać, że badanie zostało przeprowadzone na wąskiej grupie danych, więc nie były one zbyt zróżnicowane. Świadczy o tym niski współczynnik zmienności. Wpływ używki widać chociażby po średniej wieku, która zdecydowanie jest mniejsza dla osób palących, co oznacza, że zachorowali oni wcześniej. Co więcej, kwartyle, które dzielą zebrane obserwacje na cztery równe, co do ilości elementów, grupy również potwierdzają postawioną tezę. Widać, że palący otrzymali diagnozę wcześniej niż niepalący. Przykładowo, 75% niepalących kobiet zachorowało przed lub w trakcie 71 roku życia, natomiast dla kobiet palących czas ten skrócił się o 3 lata. Ciekawą obserwacją jest rozstęp, który dla niepalących kobiet jest dość wysoki, natomiast dla niepalących mężczyzn osiąga tę samą wartość co dla palących. Może mieć to swoje uzasadnienie w tym, iż mężczyźni statystycznie żyją krócej niż kobiety, dlatego nie wykrywa się u nich późno nowotworów. Warto zwrócić również uwagę na współczynnik skośności, który dla osób niepalących przyjmuje wartość ujemną, co świadczy o lewostronnej asymetrii. Oznacza to, że w badanych danych jest więcej wyników powyżej średniej, czyli osób starszych. Można stwierdzić zatem, że zachorowali później, ponieważ prowadzili zdrowszy tryb życia. Dla palących mężczyzn otrzymano już tylko delikatną lewostronną skośność, natomiast w przypadku kobiet palących rozkład w przybliżeniu jest rozkładem normalnym. Niestety, w przypadku tak szczególnej analizy niektóre statystyki nie są zbyt miarodajne przez zbyt małą liczbę obserwacji. Takim przykładem jest moda. Kolejnym wnioskiem jest wartość odchylenia standardowego, które odpowiada na pytanie o ile średnio, wartość wieku, odchyla się od średniej arytmetycznej. Z tabeli widać, że odchylenie

jest dość duże, co pozwala wysnuć fakt, iż jest wiele różnych czynników, takich jak predyspozycje genetyczne, stres czy spożywanie alkoholu, które wpływają na zachorowanie i nie da się tak naprawdę wyznaczyć jednego wieku zachorowania. Dane te zawsze będą od siebie odbiegać.

5 Podsumowanie

Dokonano analizy danych dotyczących wieku zachorowania na raka płuc. Podsumowując, można stwierdzić, że średni wiek zachorowania wynosi około 63 lata. Z histogramu widać, że przed 50 rokiem życia zdarzają się jedynie nieliczne zachorowania, być może wykryte dzięki często wykonywanym badaniom. Duża wartość wariancji oraz odchylenia standardowego świadczy o zróżnicowaniu w badanej populacji, co w przypadku analizowaniu wieku jest uzasadnione. Dane uwzględniały osoby palące i niepalące, jednakże brak w nich informacji o środowisku, w którym przebywali badani. Niepalący nadal mogą być biernymi palaczami, poprzez częste przebywanie z osobą palącą. Duży wpływ ma również zanieczyszczenie powietrza, które może wpływać na zdrowie badanej osoby nawet bez jej świadomości. Co więcej, wiele innych czynników przyczynia się do zachorowania, co potwierdza fakt, że ciężko analizuje się takie dane. Mimo ograniczonej liczby danych można jednak posunąć się o stwierdzenie, że palenie papierosów ma negatywny wpływ na zdrowie i przyśpiesza zachorowanie na raka płuc.

Bibliografia

[1] Encyklopedia Zarządzania

[2] Główny Urząd Statystyczny - słownik pojęć