Analiza wybranych danych z wykorzystaniem metod statystyki opisowej

Statystyka Stosowana

Paweł Wojarnik (276027)

Adrian Stasiak (275991)

8.05.2024

1. Wstęp

1.1 Cel pracy

Celem raportu jest zbadanie zależności pomiędzy wysokością wynagrodzenia, a ilorazem inteligencji na podstawie wybranego przez nas zbioru danych. Dzięki analizie statystycznej będziemy mogli stwierdzić czy i w jakim stopniu IQ ma wpływ na zarobki.

1.2 Charakterystyka zbioru danych

Zbiór danych posiada 935 rekordów, z których każdy opisuje inną osobę, oraz przedstawia różne parametry.

Każdy rekord zawiera siedemnaście kolumn, reprezentujących różne cechy badanych:

- **wage:** miesięczne zarobki
- **hours:** średnia liczba przepracowanych godzin w tygodniu
- **IQ:** wynik IQ
- **KWW:** wynik testu znajomości świata pracy
- **educ:** lata edukacji
- **exper:** lata doświadczenia
- **tenure:** lata z aktualnym pracodawcą
- **age:** wiek
- **married:** =1 jeżeli po ślubie
- **black:** =1 jeżeli czarni

- **south:** =1 jeżeli z południa
- **urban:** =1 jeżeli z miasta
- **sibs:** liczba rodzeństwa
- **brthord:** kolejność urodzenia
- **meduc:** edukacja matki
- **feduc:** edukacja ojca
- **lwage:** logarytm naturalny płacy

Jednak do naszego raportu będziemy używać jedynie danych na temat miesięcznych zarobków oraz wyniku IQ. Obie te cechy można sklasyfikować jako dane ilościowe.

1.3 Źródło danych

Użyte dane o nazwie "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials" zostały zebrane i udostępnione przez Profesora D. Neumarka. Wykorzystane przez nas dane zostały jedynie z 1980 roku, dostępne na stronie nauczyciela akademickiego Vincenta Arel-Bundock' a: https://vincentarelbundock.github.io.

2. Podstawowe Statystyki

2.1 Średnia arytmetyczna

Średnia arytmetyczna $\bar{x}$ jest dobrą miarą położenia rozkładu i jednocześnie miarą tendencji centralnej. Jest to miara klasyczna rozkładu, czyli każda zmiana dowolnego elementu badanego zbioru pociąga za sobą zmianę wartości średniej. Średnia arytmetyczna to suma wszystkich wartości zmiennej x, dla badanej zbiorowości, podzielona przez liczbę jednostek w tej zbiorowości (n). Średnią arytmetyczną obliczamy ze wzoru:

$$\overline{\boldsymbol{x}} = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{x_i}}{\boldsymbol{n}}$$

gdzie:

$\bar{x}$ – średnia arytmetyczna,

x_i – wartość zmiennej dla i-tej jednostki w zbiorowości,

n – liczba jednostek w zbiorowości

Otrzymane wartości średniej arytmetycznej w badaniu:

płaca: **957,95**

wynik IQ: **101,28**

2.2 Mediana

Mediana – wartość cechy w szeregu uporządkowanym, powyżej i poniżej której znajduje się jednakowa liczba obserwacji. Mediana znalazła szerokie zastosowanie w statystyce jako średnia bardziej odporna na elementy odstające niż średnia arytmetyczna.

Aby obliczyć medianę ze zbioru n obserwacji, sortujemy je w kolejności od najmniejszej do największej i numerujemy od 1 do n. Następnie, jeśli n jest nieparzyste, medianą jest wartość obserwacji w środku (czyli obserwacji numer $\frac{n+1}{2}$). Jeśli natomiast n jest parzyste, wynikiem jest średnia arytmetyczna między dwiema środkowymi obserwacjami, czyli obserwacją numer $\frac{n}{2}$ i obserwacją numer $\frac{n}{2} + 1$.

Otrzymane wartości mediany:

płaca: **905**

wynik IQ: **102**

Mediana płacy jest wyraźnie mniejsza od średniej, co wskazuje na częste występowanie płac poniżej średniej, ale też na bardzo wysokie płace powyżej średniej.

Mediana wyniku IQ jest zbliżona do wartości średniej co wskazywałoby na równomierne rozłożenie wyników badanych.

2.3 Wariancja

Wariancja – miara zmienności zmiennej losowej będąca wartością oczekiwaną kwadratu różnicy wartości zmiennej losowej X i jej wartości oczekiwanej. W statystyce opisowej obliczana jest jako średnia arytmetyczna kwadratów odchyleń (różnic) poszczególnych wartości cechy od średniej. Obliczana ze wzoru:

$$Var(X) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

gdzie:

$Var(X)$ – wariancja,

$\bar{x}$ – średnia arytmetyczna,

x_i – wartość zmiennej dla i-tej jednostki w zbiorowości,

n – liczba jednostek w zbiorowości

Otrzymane wartości wariancji:

płaca: **163507,7**

wynik IQ: **226,58**

2.4 Odchylenie standardowe

Odchylenie standardowe jest pierwiastkiem kwadratowym z wariancji. Odchylenie standardowe informuje o tym, jak daleko wartości danej wielkości są rozrzucone wokół jej średniej. Im mniejsza wartość odchylenia tym obserwacje są bardziej skupione wokół średniej. Odchylenie standardowe obliczamy ze wzoru:

$$\boldsymbol{\sigma(x) = \sqrt{Var[x]}}$$

gdzie:

σ – odchylenie standardowe,

$Var(X)$ – wariancja

Otrzymane wartości odchylenia standardowego:

płaca: **404,36**

wynik IQ: **15,05**

Wysokie odchylenie standardowe płacy wskazuje na silny rozrzut wartości wokół średniej. Odchylenie to sugeruję, że wartości często mocno odbiegają od średniej.

Odchylenie standardowe wyników IQ jest umiarkowane. Sugeruje, że wartości skupiają się głównie wokół średniej.

2.5 Kowariancja

Kowariancja jest miarą, która określa, w jaki sposób dwie zmienne zmieniają się razem. Może być używana do określenia, czy zmienne są skorelowane (czy zmieniają się razem w podobny sposób), oraz w jakim stopniu. Ogólnie rzecz biorąc, kowariancja między dwiema zmiennymi jest dodatnia, gdy obie zmienne rosną lub maleją razem, ujemna, gdy jedna zmienna rośnie, a druga maleje, i bliska zeru, gdy brak jest wyraźnej zależności między zmiennymi.

Kowariancję między dwiema zmiennymi losowymi X i Y można obliczyć za pomocą następującego wzoru:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n}$$

gdzie:

cov – kowariancja,

$\bar{x}, \bar{y}$ – średnia arytmetyczna,

x_i, y_i – wartość zmiennej dla i-tej jednostki w zbiorowości,

n – liczba jednostek w zbiorowości

Otrzymany wynik kowariancji dla płacy oraz wyniku IQ: **1881**

Wynik jest dodatni co oznacza, że zmienne rosną lub maleją razem, otrzymana liczba jest duża co wskazywałoby na dużą zależność pomiędzy płacą, a IQ, jednak różnica w jednostkach może wpływać na wynik i prowadzić do fałszywej interpretacji.

2.6 Współczynnik korelacji

Współczynnik jest miarą liniowej zależności między dwiema zmiennymi. Ma wartość między -1 a 1. Wartość współczynnika korelacji wskazuje, jak silna jest liniowa zależność między dwiema zmiennymi. Im bliżej wartość współczynnika do 1 lub -1, tym silniejsza jest zależność, a im bliżej wartość do 0, tym słabsza jest zależność. Gdy współczynnik korelacji jest dodatni, oznacza to, że zmienne są dodatnio skorelowane. Oznacza to, że gdy jedna zmienna rośnie, druga zmienna również rośnie, a gdy jedna zmienna maleje, druga zmienna również maleje. Gdy współczynnik korelacji jest ujemny, oznacza to, że zmienne są ujemnie skorelowane. Oznacza to, że gdy jedna zmienna rośnie, druga zmienna maleje, a gdy jedna zmienna maleje, druga zmienna rośnie.

Współczynnik korelacji wyliczamy ze wzoru:

$$\boldsymbol{r_{xy} = \frac{cov(X,Y)}{\sigma(x)\sigma(y)}}$$

gdzie:

r_{xy} – współczynnik korelacji,

cov – kowariancja,

σ – odchylenie standardowe

Otrzymany współczynnik korelacji pomiędzy płacą, a wynikiem IQ wynosi: **0,309**

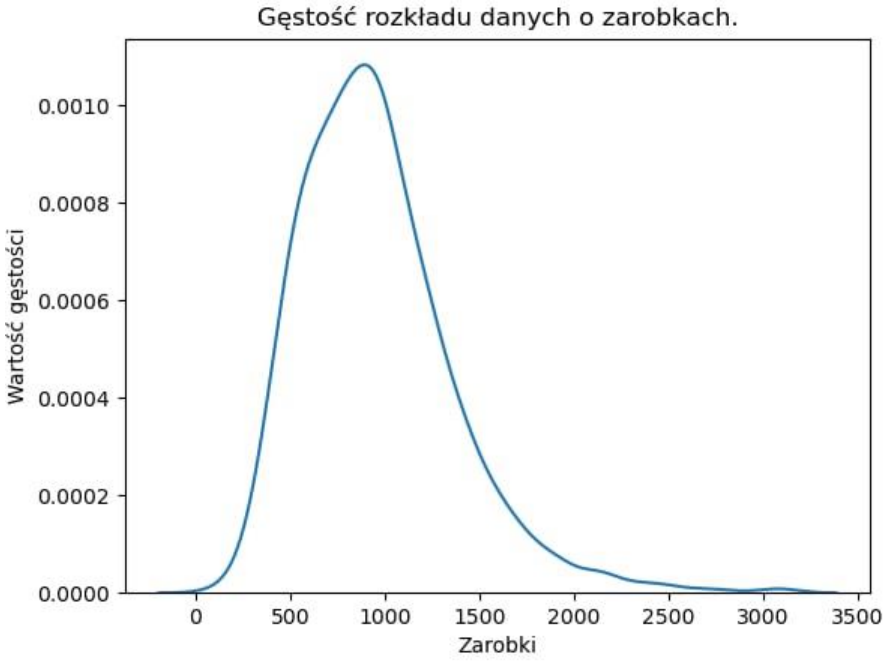
Współczynnik korelacji normalizuje kowariancję i zwraca wartości między -1 a 1 co ułatwia interpretację i zmniejsza rolę różnych jednostek zmiennych. Z otrzymanego współczynnika wynika, że istnieje umiarkowana i dodatnia zależność liniowa między płacą, a wynikiem IQ.

3. Wizualizacja danych

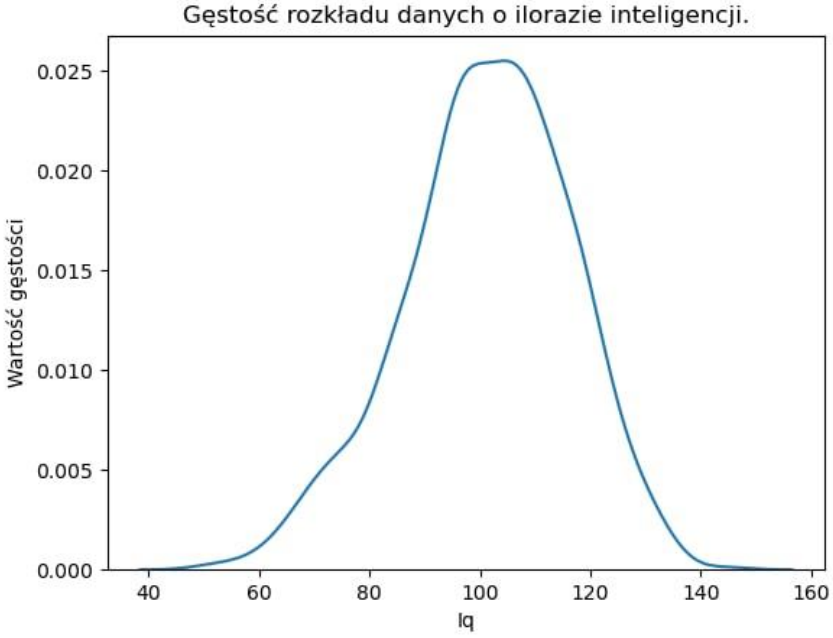
Wizualizacja danych to proces reprezentowania danych za pomocą grafik, wykresów, map, diagramów lub innych form wizualnych. Celem wizualizacji danych jest przekształcenie surowych danych w formę, która jest łatwiejsza do zrozumienia i analizy.

3.1 Gęstość

Gęstość reprezentowana graficznie pokazuje, jak wartości danych są rozmieszczone lub zagęszczone w określonej przestrzeni numerycznej. Analiza gęstości danych jest istotnym narzędziem w analizie danych i statystyce, ponieważ pomaga zrozumieć, jak dane są skoncentrowane wokół określonych wartości, jakie są ich rozkłady i relacje między nimi.



Przez widoczny ogon po prawej stronie gęstość rozkładu płacy jest asymetryczna, lewoskośna co oznacza że większość obserwacji skupia się po lewej stronie średniej w bliskim od niej położeniu, a wartości po prawej występują rzadziej, ale są bardziej ekstremalne.

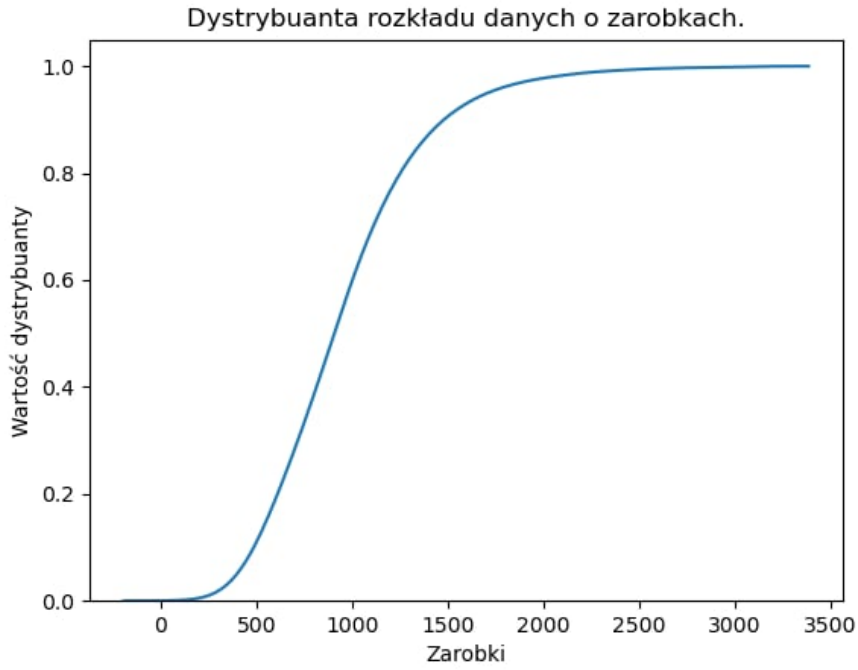


Gęstość rozkładu danych o wyniku IQ jest zbliżony do gęstości rozkładu normalnego, oznacza to skupienie wyników w środku badanego przedziału, w pobliżu średniej i mediany. Brak wyraźnej skośności na wykresie pokazuje, że wyniki wysokie i niskie były równie często uzyskiwane.

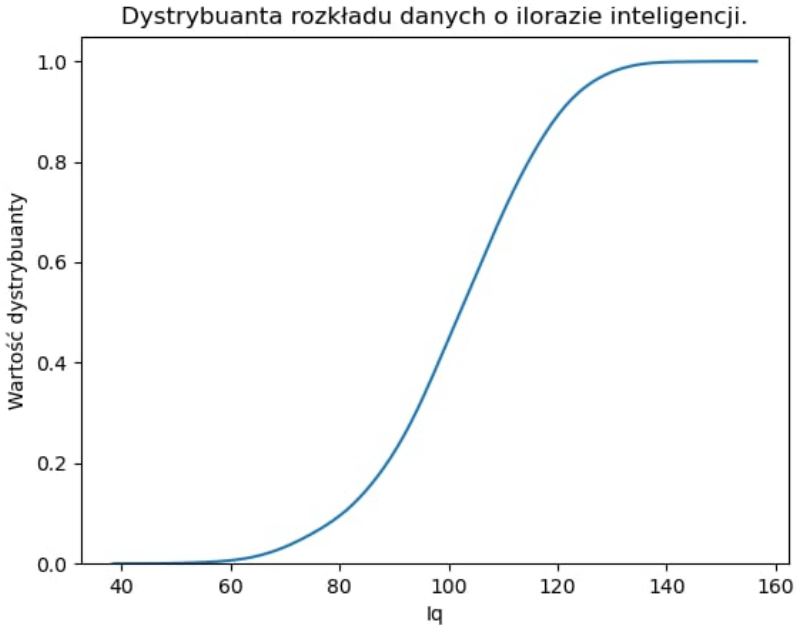
3.2 Dystrybuanta

Dystrybuanta w statystyce jest funkcją, która opisuje kumulatywny rozkład prawdopodobieństwa zmiennej losowej. Innymi słowy, jest to funkcja, która określa prawdopodobieństwo, że zmienna losowa będzie miała wartość mniejszą lub równą danej wartości. Dystrybuanta jest używana do analizy rozkładów prawdopodobieństwa i do obliczania kwantyli.

Dystrybuanta w kontekście analizy graficznej, dystrybuanta jest reprezentowana graficznie za pomocą wykresu kumulatywnego (lub krzywej dystrybuanty). Na osi poziomej wykresu znajdują się wartości zmiennej losowej, a na osi pionowej znajduje się prawdopodobieństwo, że zmienna losowa będzie miała wartość mniejszą lub równą danej wartości. Wykres dystrybuanty umożliwia szybką wizualizację kształtu i charakterystyk rozkładu danych, co pomaga w zrozumieniu ich rozkładu i właściwości.



Z wykresu dystrybuanty zarobków widać, kumulacje płac na przedziale od 500 do 1500 oraz charakterystyczny ogon na końcu rozkładu wynikający ze skośności rozkładu.



Wykres dystrybuanty wyników IQ jest zbliżony do rozkładu normalnego, z wykresu widać że prawdopodobieństwo jest zbliżone odo symetrycznego względem średniej.

4. Wnioski

Istnieje umiarkowana dodatnia zależność pomiędzy wynikiem IQ a wysokością wynagrodzenia. Współczynnik korelacji pomiędzy tymi dwoma zmiennymi wynosi 0,309, co wskazuje na istnienie pewnej liniowej zależności między nimi. Jednakże wartość tego współczynnika sugeruje, że wpływ IQ na zarobki występuję, ale jest niewielki, a inne czynniki również odgrywają rolę w określaniu poziomu wynagrodzenia.

Analiza gęstości rozkładu płac wykazała, że większość obserwacji skupia się w niższych przedziałach płac, z jednoczesnym występowaniem nielicznych, lecz ekstremalnych wartości na wyższych poziomach. Natomiast gęstość rozkładu danych dotyczących wyniku IQ była zbliżona do rozkładu normalnego, co oznacza, że wyniki IQ są równomiernie rozłożone wokół wartości średniej.

Choć istnieje pewna zależność między wynikiem IQ a wysokością zarobków, to jednak nie jest ona jednoznaczna i nie można jednoznacznie stwierdzić, że wyższe IQ automatycznie przekłada się na wyższe zarobki. Inne czynniki, takie jak doświadczenie zawodowe czy edukacja, mogą mieć równie lub bardziej istotny wpływ na wysokość wynagrodzenia.