PSP – Raport 1

Nela Tomaszewicz

Marzec 2020

1 Zadanie 1

W zadaniu 1 badamy próbkę danych 78 uczniów siódmej klasy pewnej szkoły w USA. Dla każdego ucznia została podana średnia ocen w skali F-A z odpowiadającymi liczbami 0-11, wynik standardowego testu IQ, płeć (F – kobieta, M – mężczyzna) oraz punktacja na teście psychologicznym Piers-Children's Self Concept Scale. Aby zrozumieć badane dane, należy dowiedzieć się, co oznaczają odpowiednie wyniki w odpowiednim kontekście. Ponieważ zadanie polega na narysowaniu histogramów zmiennych ilościowych oraz obliczeniu wartości statystyk takich jak minimum, maksimym, mediana, kwartle, średnia, odchylenie standardowe, wariancja czy współczynnik zmienności i wyciągnięciu wniosków, każdą z wielkości omówię przy okazji kolejnych histogramów, a następnie wyciągne wnioski ogólne.

Wszystkie histogramy zostały wygenerowane przy użyciu pakietu `ggplot2`, a optymalna szerokość binów została obliczona przy pomocy reguły Freedmana-Diaconisa, której wzór wygląda następująco:

$$d = 2 * \frac{IQR(x)}{\sqrt[3]{n}}$$

,gdzie d oznacza szerokość bina, x próbę, n liczność próby, natomiast IQR to rozstęp międzykwartylowy, czyli różnica pomiędzy pierwszym a trzecim kwartlem.

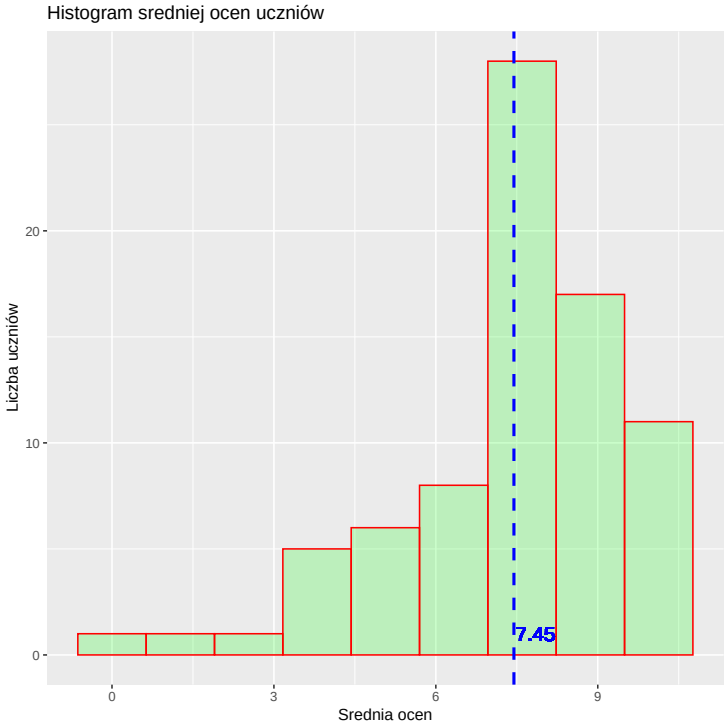
1.1 Analiza danych całej grupy uczniów

1.1.1 Średnia ocen uczniów

Pierwszą, i najprostszą do zinterpretowania zmienną ilościową jest średnia ocen uczniów. W USA oceny wystawiane są w skali F-A, która odpowiada liczbom od 0 do 11.

F	D-	D	D+	C-	C	C+	B-	B	B+	A-	A
0	1	2	3	4	5	6	7	8	9	10	11

Najgorszą oceną jest F, a najlepszą A.



Histogram jest jednomodalny z modą równą 9.167 (obliczone przy pomocy własnoręcznie zaimplementowanej funkcji w R), lewoskośny (sprawdzone przy pomocy funkcji `skewness` z pakietu `e1071`). Rozstęp międzykwartylowy wynosi 2.705, co w interpretacji na oceny nie jest aż tak ogromnym rozrzutem.Średnia ocena jest równa około 7.5, czyli znajduje się pomiędzy B- a B.

Jeśli chodzi o wartości statystyk podanych w zadaniu, są one następujące:

Min.	Max.	1st Qu.	Median	Mean	3rd Qu.	Sd.	Var.	CV
0.53	10.76	6.278	7.829	7.447	8.983	2.1	4.408	0.282

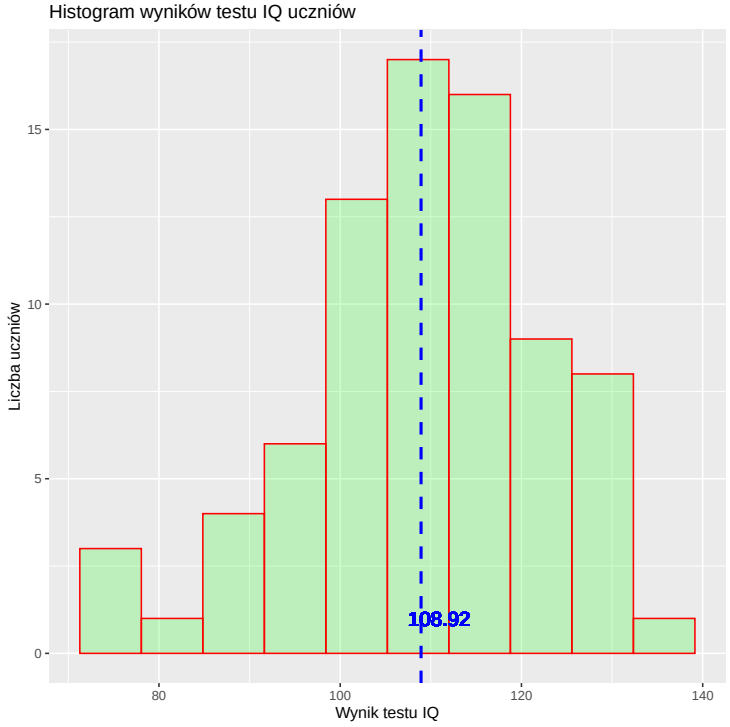
Korzystając z (2) dowiemy się, że dobrą Grade Point Average (GPA) dla przeciętnego ucznia 7 klasy jest 3.6, gdzie maksymalne GPA to 4.0. Dalej, biorąc pod uwagę informacje zawarte w (1) możemy stwierdzić, że GPA równe 3.6 leży pomiędzy B+ (GPA około 3.3) a A- (GPA około 3.7). Ważne jest jednak, aby pamiętać, że przelicznik podany w źródle nie jest uniwersalny i oceny zależą od szkół, do których chodzą uczniowie. Średnia ocen badanych uczniów wynosi około 7.5, więc zaokrąglając w górę daje nam to ocenę B-, czyli GPA równe około 2.67, zatem przy pomocy poprzednio otrzymanych informacji można powiedzieć, że mamy do czynienia z dość przeciętną grupą uczniów lub „trudną" szkołą. Oczywiście, analiza na razie dotyczy jedynie średniej bez uwzględnienia kolejnych danych ilościowych, które przeanalizujemy w pozostałych częściach raportu i które mogą znacząco wpłynąć na obraz badanej grupy.

1.1.2 IQ uczniów

IQ uczniów zostało wyznaczone poprzez standardowy test IQ. Oznacza to, że prawdopodobnie został przeprowadzony test *Wechsler Intelligence Scale for Children*((4)), w skrócie WISC. Jest to test przeprowadzany wśród dzieci w przedziale wiekowym 6-16. Najnowsza (piąta) wersja testu pochodzi z roku 2014 i wśród dzieci określono następującą klasyfikację ilorazu inteligencji (źródło: (5)):

Przedział IQ	Interpretacja
$\geqslant 130$	Inteligencja bardzo wysoka
120-129	Inteligencja wysoka
110-119	Inteligencja powyżej przeciętnej
90-109	Inteligencja przeciętna
70-89	Inteligencja niższa niż przeciętna
$\leqslant 69$	Upośledzenie umysłowe

Histogram dla ilorazu inteligencji badanych uczniów wygląda następująco:



Histogram jest jednomodalny, delikatnie lewoskośny. Moda wynosi 111, średnia 108.92, a mediana 110. Mamy zatem średnio do czynienia z uczniami o inteligencji przeciętnej i lekko powyżej przeciętnej. Rozstęp międzykwartylowy wynosi 14.5 i w kontekście współczynnika ilorazu inteligencji oznacza on, że dane są rozrzucone, ponieważ 14.5 to liczba punktów, która może sprawić, że dana

osoba zostanie zakwalifikowana do wyższej lub niższej grupy. Ponadto, wyliczona później wariancja również świadczy o dość dużym rozrzucie danych.

Wartości statystyk są następujące:

Min.	Max.	1st Qu.	Median	Mean	3rd Qu.	Sd.	Var.	CV
72	136	103	110	108.9	117.5	13.2	173.5	0.1

Zatem w badanej grupie znajdują się zarówno jednostki o bardzo niskiej inteligencji (IQ=72), jak i uczniowie wybitni o ilorazie inteligencji równym 136. Ponadto, 25% uczniów ma iloraz inteligencji większy niż 117.5 (trzeci kwartyl), co oznacza, że są to osoby z inteligencją powyżej przeciętnej, wysoką i bardzo wysoką.

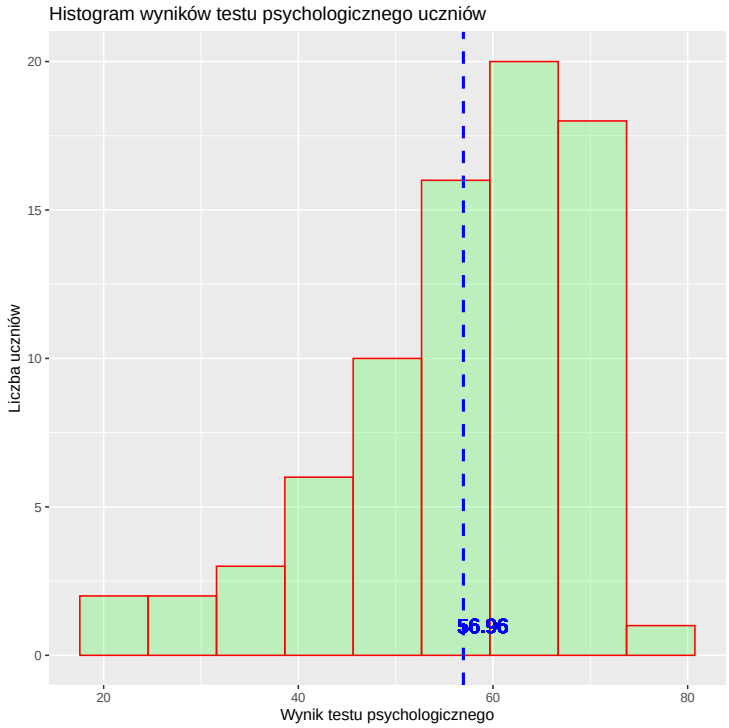
1.1.3 Wyniki testu psychologicznego przeprowadzonego na uczniach

Ostatnią omawianą zmienną ilościową jest punktacja na teście psychologicznym *Piers-Harris Children's Self–Concept Scale* (PHCSCS-2). Test jest przeprowadzany w grupie wiekowej 7-18 i polega na serii 60 pytań z odpowiedziami Tak/Nie. Co ważne, na pytania odpowiada uczestnik, a nie jego rodzic czy nauczyciel. Wynik końcowy oddaje ogólną samoocenę, ale test także na podstawie punktacji daje informacje z 6 kategorii: dostosowania behawioralnego (*Behavioral Adjustment*), wolności od lęków (*Freedom From Anxiety*), szczęścia i spełnienia (*Happiness and Satisfaction*), statusu szkolnego i intelektualnego (*Intellectual and School Status*), wyglądu fizycznego (*Physical Appearance and Attributes*) i akceptacji społecznej (*Social Acceptance*). Rozważa się wynik ogólny, a także wynik z każdej ze wspomnianych kategorii w celu dogłębnej analizy samooceny uczestnika. Niestety dla naszego raportu, test oraz instrukcja do niego są płatne, więc jedynym źródłem na temat interpretacji punktacji, jakie udało mi się znaleźć jest (6) (strona szkół w hrabstwie Montgomery w stanie Maryland).

Wynik PHCSCS-2	Interpretacja
$\geqslant 70$	Bardzo wysoka samoocena i duża pewność siebie
60-69	Wysoka samoocena
40-59	Badany w normie
56-59	Badany skłania się w stronę pozytywnej samooceny
40-44	Skłonności do negatywnej samooceny
$\leqslant 39$	Badany ma poważne zaburzenia samooceny
$\leqslant 29$	Istnieje duże prawdopodobieństwo, że badany jest zaburzony psychicznie

Interpretacja nie jest do końca oczywista i na pewno, aby była poprawna, musi zostać przeprowadzona przez wykwalifikowanego psychologa.

Histogram dla wyników testu psychologicznego wygląda następująco:



Histogram jest jednomodalny z modą równą 67, lewoskośny. Średnia wynosi 56.96, natomiast mediana 59.5, zatem możemy stwierdzić, że mamy do czynienia z uczniami z raczej pozytywną samooceną. Rozstęp międzykwartylowy wynosi 15. Tę wielkość nie jest łatwo ocenić, ale z pomocą przychodzi duża wariancja, dzięki której możemy stwierdzić, że dane są rozrzucone.

Wyniki statystyk są następujące:

Min.	Max.	1st Qu.	Median	Mean	3rd Qu.	Sd.	Var.	CV
20	80	51	59.5	56.96	66	12.4	154.1	0.2

W przypadku minimum i maksimum nie jesteśmy w stanie zinterpretować danych. Według źródła (6) przyczyna bardzo niskiego rezultatu jak i bardzi wysokiego, może nie być oczywista i należy brać pod uwagę takie czynniki jak losowość odpowiedzi czy zaznaczanie konsekwentnie jednej odpowiedzi niezależnie od zadanego pytania. Średnia wynosi prawie 57 punktów, zatem średnio badani uczniowie są w normie ze skłonnością do pozytywnej samooceny.

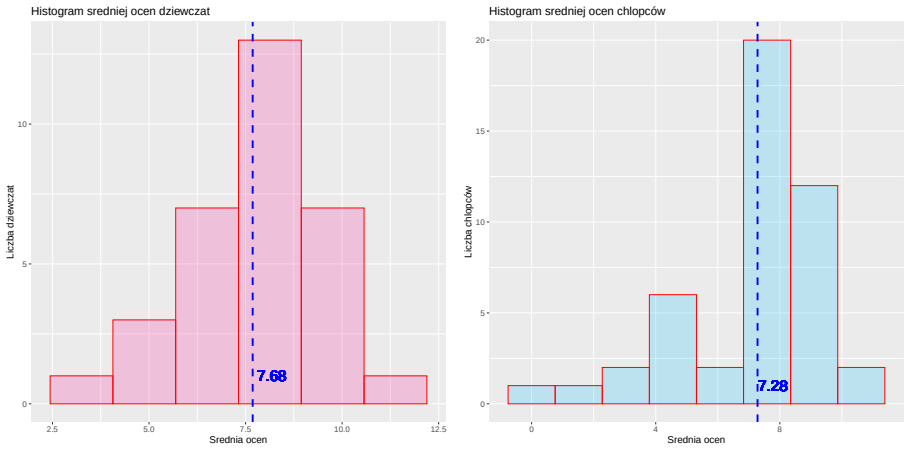
1.1.4 Podsumowanie i wnioski

Uzyskane informacje pozwalają na stwierdzenie, że badana grupa uczniów to osoby o raczej wysokim ilorazie inteligencji, oceniające siebie pozytywnie, jednakże mające średnio nie najlepsze oceny. Można wysnuć tezę, że szkoła, do

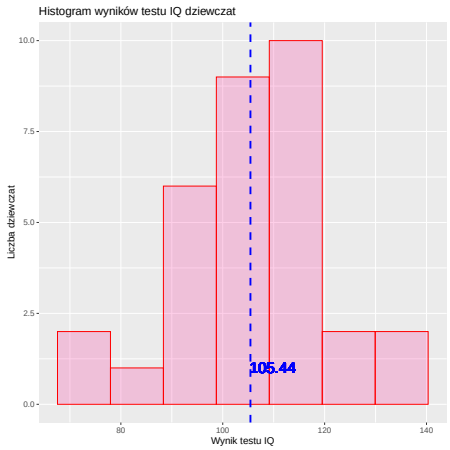
której uczęszczają uczniowie należy do szkół wymagających, w której nawet bardzo inteligentne jednostki nie osiągają bardzo wysokich rezultatów.

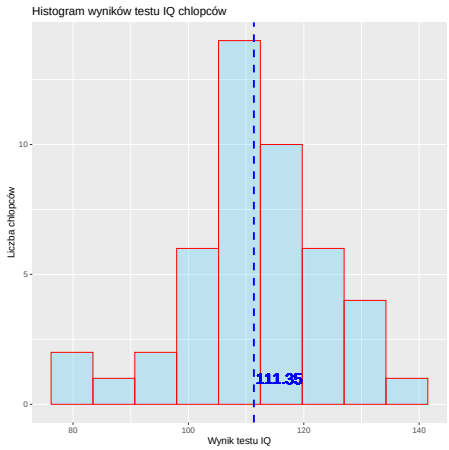
1.2 Analiza danych uczniów z uwzględnieniem płci

Istnieje wiele stereotypowych stwierdzeń dotyczących nastoletnich chłopców i dziewcząt. Pierwszą często powtarzaną tezą jest to, że dziewczynki na etapie nauki szkolnej osiągają lepsze wyniki od chłopców. Drugą, że dziewczynki są inteligentniejsze, a trzecią, że mają niższą samoocenę. Oczywiście, aby powiedzieć, że tego typu teza jest prawdziwa lub fałszywa musielibyśmy zbadać wszystkich uczniów i uczennice siódmych klas w Stanach Zjednoczonych, ponieważ próba 78 osób jest zdecydowanie za mała na wysnuwanie wniosków dotyczących ogółu. Jednakże, możemy sprawdzić, czy tego typu popularne stwierdzenia są prawdziwe dla badanej przez nas grupy uczniów. Zanim przystąpimy do badania, warto zauważyć, że próba dziewcząt to 32 osoby, a chłopców 46.

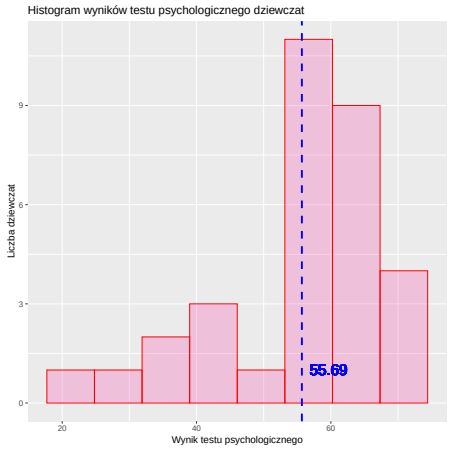


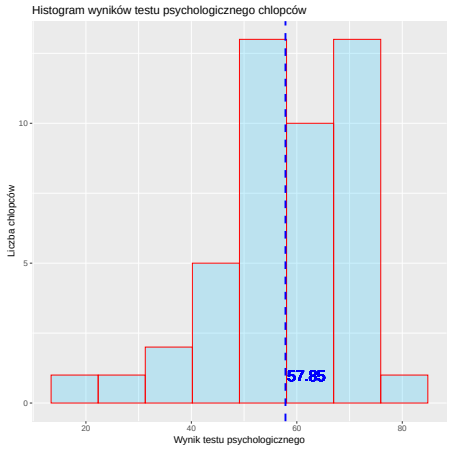
Jeśli chodzi o średnią, faktycznie dziewczynki mają ją lepszą niż chłopcy. Tak samo wynik minimalny, u chłopców wynosi on tylko 0.53, natomiast u dziewcząt 3.4. Jednakże, jeśli weźmiemy pod uwagę pozostałe statystyki, to mediana dla dziewcząt wynosi 7.83, a dla chłopców 7.88, czyli jest odrobinkę wyższa dla tej drugiej grupy. Różnice zwiększają się wraz z pozostałymi statystykami. Trzeci kwartyl dla średniej dziewcząt to 8.95, natomiast dla chłopców 9.08, maksimum dla chłoców również jest większe, ale nieznacznie. Tak więc, jeśli chodzi o średni poziom dziewcząt, jest on wyższy, ale więcej chłopców osiągnęło wyższe wyniki.





W przypadku testu IQ, dziewczynki osiągnęły odrobinę gorszy średni wynik niż chłopcy. Również mają mniejszy wynik minimalny od chłopców, bo równy 72, podczas gdy minimum w drugiej grupie wynosi 77. Dla każdej pozostałej statystyki, takiej jak maksimum, pierwszy kwartyl, mediana czy trzeci kwartyl sytuacja wygląda tak samo. Wariancja ma większą wartość u dziewcząt niż u chłopców, co wskazuje na większą różnicę w wynikach. Różnice mogą być spowodowane większą próbą wyników chłopców. Dodatkowo, warto zauważyć, że zazwyczaj niewielkie różnice w wynikach testu nie sprawiają, że dana osoba jest zakwalifikowana jako mniej lub bardziej inteligentna z powodu raczej szerokich przedziałów punktowych w teście.





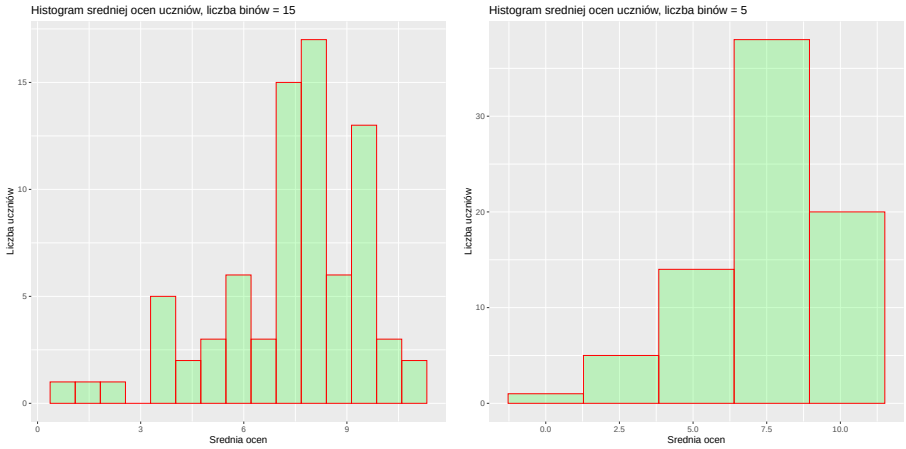
W kwestii samooceny, dziewczęta średnio wypadają gorzej niż chlopcy. Wartość minimalna, pierwszy kwartyl i mediana są u dziewcząt nieznacznie wyższe niż u chłopców – różnica wynosi zazwyczaj 1-1.5 punktu. Jednakże, już od średniej, przez trzeci kwartyl i do maksimum, chłopcy znaczącą przewyższają dziewczęta, osiągając w maksimum różnicę aż 8 punktów (72 u dziewcząt, 80 u chłopców), jednakże, jak juz wspomniałam w poprzednim punkcie ta wartość niekoniecznie może być uznana za prawidłową. Wyniki te, niestety, potwier-

dzają tezę o gorszej samoocenie wśród nastoletnich dziewcząt w porównaniu z chłopcami w tym samym wieku.

Reasumując, tezy postawione na początku badania nie zostały potwierdzone. Jednakże, różnice pomiędzy wynikami obydwu płci są raczej niewielkie.

2 Zadanie 2

Zmienną ilościową, dla której przeprowadzę badania jest średnia. W poprzednim zadaniu użyłam wzoru na „optymalną" szerokość binu pochodzącego z reguły Freedmana-Diaconisa. Istnieje jednak dużo innych wzorów zorientowanych na obliczenie optymalnej liczby binów (w R nie ma znaczenia czy używamy szerokości binu, czyli `binwidth` czy liczby binów `nbins`, ponieważ zgodnie z dokumentacją `ggplot2` jedna wielkość nadpisuje drugą). Początkowo, nie myśląc o regule Freedmana-Diaconisa, użyłam metody prób i błędów określając liczbę binów. Oczywiście, takie podejście jest dość fatalne i histogramy, które wtedy uzyskałam nie dają pewności na dobrą interpretację wyników.

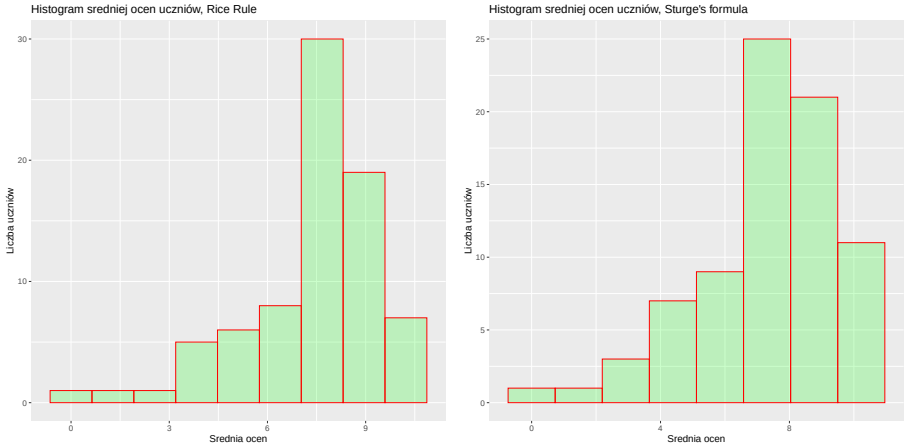


Liczba binów wyliczona regułą Freedmana-Diaconisa jest równa 9. Na powyższych histogramach widać liczbę większą i mniejszą od 9. Pierwszy histogram jest „postrzępiony", zbyt bliska odległość słupków utrudnia odczytanie informacji. Na drugim biny są za duże, tracimy część informacji.

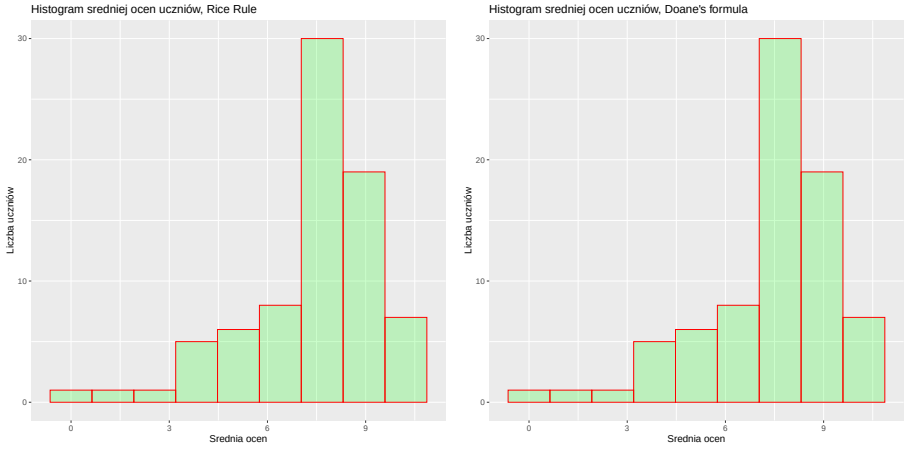
Rozważymy kilka wzorów z artykułu (7).

Nazwa	Wzór (k–liczba binów, n–liczba obserwacji)
Square-root choice	$k = \sqrt{n}$
Sturge's formula	$k = \log_2 n + 1$
Rice Rule	$k = 2\sqrt[3]{n}$
Doane's formula	$k = 1 + \log_2 n + \log_2 (1 + \frac{\lvert g_1 \rvert}{\sigma_{g_1}})$, gdzie g_1– skośność rozkładu, $\sigma_{g_1} = \sqrt{\frac{6(n-2)}{(n+1)(n+3))}}$

Dla pierwszych dwóch wzorów mamy następujące histogramy:



Pierwszy histogram jest w zasadzie taki sam jak ten wygenerowany przy pomocy reguły Freedmana-Diaconisa. Drugi natomiast ma mniej binów. Dla kolejnych dwóch wzorów mamy następujące histogramy.



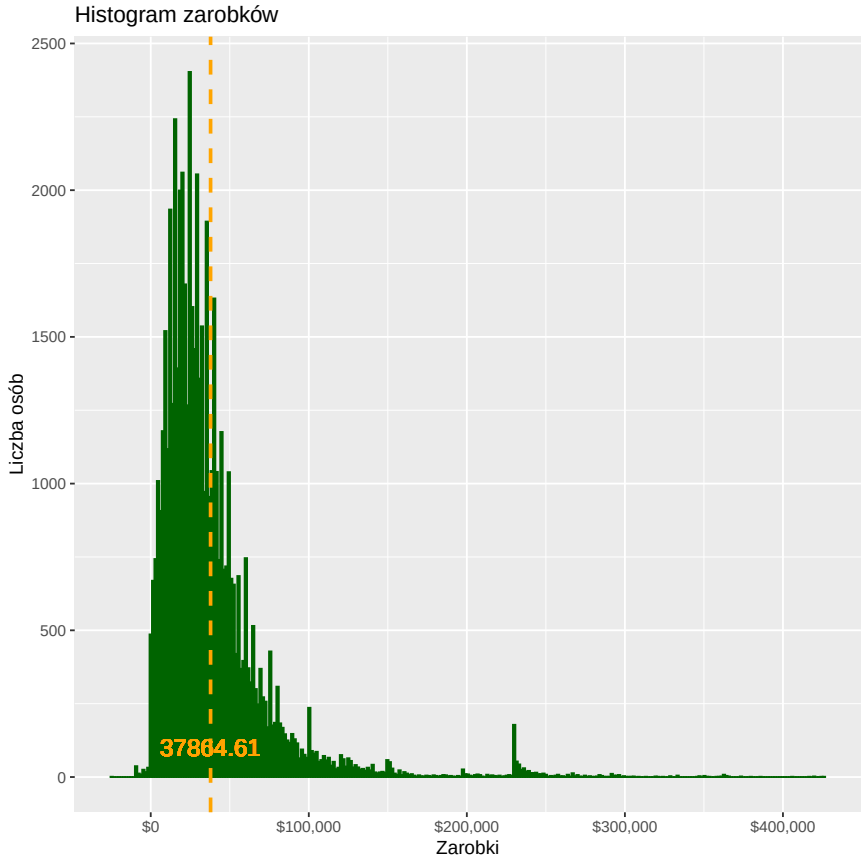
Obydwa histogramy są takie same jak ten wygenerowany jeszcze Zadaniu 1. Co ciekawe, wzór Doane (Doane's formula) jest ulepszeniem wzoru Sturge'sa, który nie działa dobrze na danych nie będących z rozkładu normalnego, co mogliśmy zauważyć na histogramie wygenerowanym przy jego pomocy. Reasumując, można zauważyć, że do poprzedniego zadania wybór wzoru dla liczby/szerokości binów nie miał znaczenia z wyłączeniem wzoru Sturgesa. Ważne było to, aby był to 'działający' wzór a niedopasowanie liczby binów metodą prób i błędów.

3 Zadanie 3

W zadaniu 3 zajmujemy się analizą zarobków grupy 55 899 osób z USA. Dostarczone dane to wyniki ankiet przeprowadzonych przez *Bureau of Labor Statistics*. Dla każdej osoby mamy podany wiek w latach, wykształcenie (1 – podstawowe, 2 – niepełne średnie, 3 – średnie, 4 – niepełne wyższe, 5 – wyższe (licencjat), 6 – wyższe(magisterium)), płeć (1 – mężczyzna, 2– kobieta), roczne zarobki w dolarach a także sektor zatrudnienia (5–prywatny, 6–publiczny, 7–samozatrudnienie).

3.1 Analiza zarobków

Szerokość binu w histogramie zostanie wyznaczona z reguły Freedmana-Diaconisa.



Histogram jest jednomodalny z modą równą \$20000. Jest wyraźnie prawoskośny, co również potwierdza funkcja `skewness`, zwracająca wartość powyżej 3. Rozstęp międzykwartylowy wynosi \$29503.5 i w kontekście danych oznacza to całkiem spory rozrzut, co potwierdzają wysokie odchylenie standardowe i

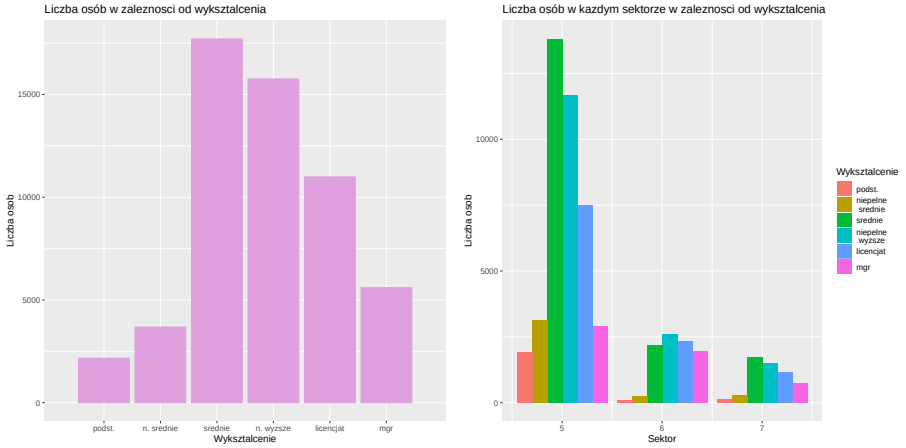
wariancja.

Wartości poszczególnych statystyk wynoszą:

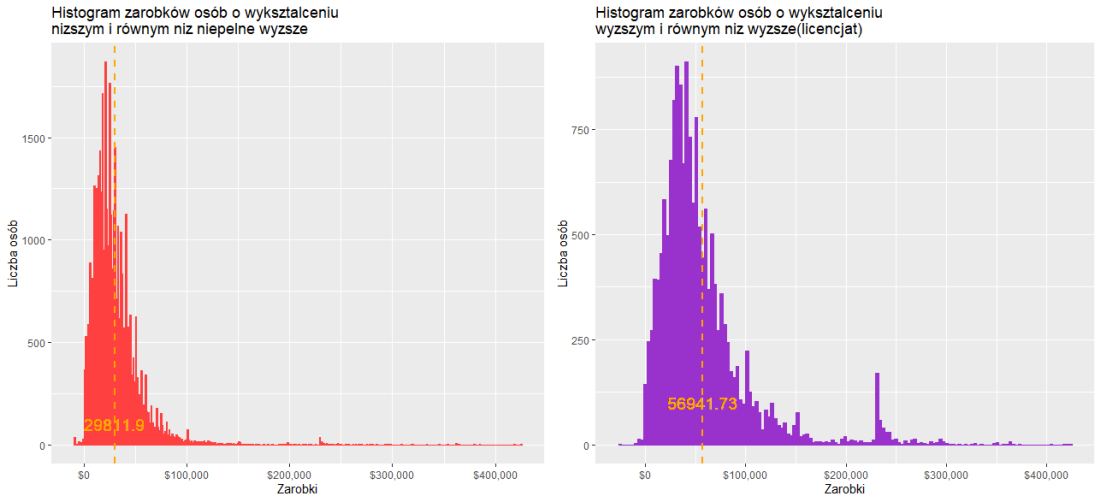
Min.	Max.	1st Qu.	Median	Mean	3rd Qu.	Sd.	Var.	CV
-24998	425510	17000	29717	37865	46504	36158	1307402975	0.955

Możemy zauważyć, że minimum zbioru obserwacji jest ujemne. Badając obserwacje odstające w kolejnym punkcie sprawdzimy, czy jest to błąd w zapisie czy być może prawdopodobne dane. Mediana to $29717, a dla porównania, według obecnych danych, mediana zarobków w USA wynosi około $60000.

Aby lepiej zrozumieć histogram, musimy przeanalizować dostarczone dane jakościowe. W tym celu użyję diagramów słupkowych.



Najwięcej osób w próbie ma wykształcenie średnie i pracuje w sektorze prywatnym. W sektorach: publicznym i samozatrudnieniu pracuje znacząco mniej osób, jednakże w sektorze publicznym zatrudnionych jest więcej osób z wykształceniem wyższym niż średnie. W sumie procentowy udział osób, które ukończyły studia (licencjackie lub magisterskie) wynosi około 29%. Zatem stosunkowo niska mediana zarobków spowodowana jest dużym udziałem pracowników, którzy mają jedynie średnie lub niepełne wyższe wykształcenie. Można sobie zadać pytanie, czy w takim razie osoby, które ukończyły studia zarabiają więcej? Spróbujemy znaleźć odpowiedź formułując dwa histogramy: jeden dla osób z wykształceniem niepełnym wyższym lub niższym a drugi dla osób z dyplomem studiów licencjackich lub magisterskich.

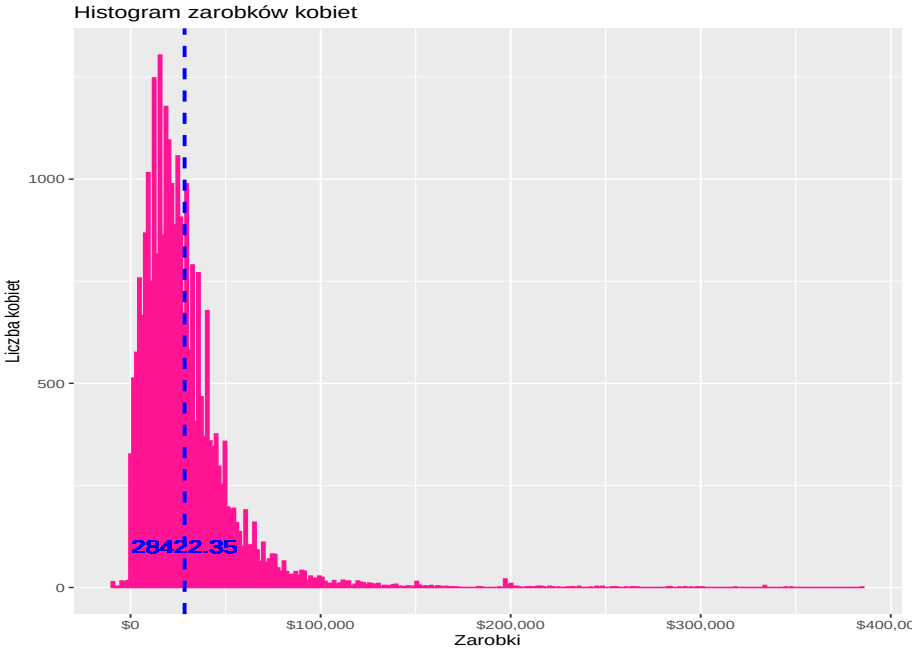


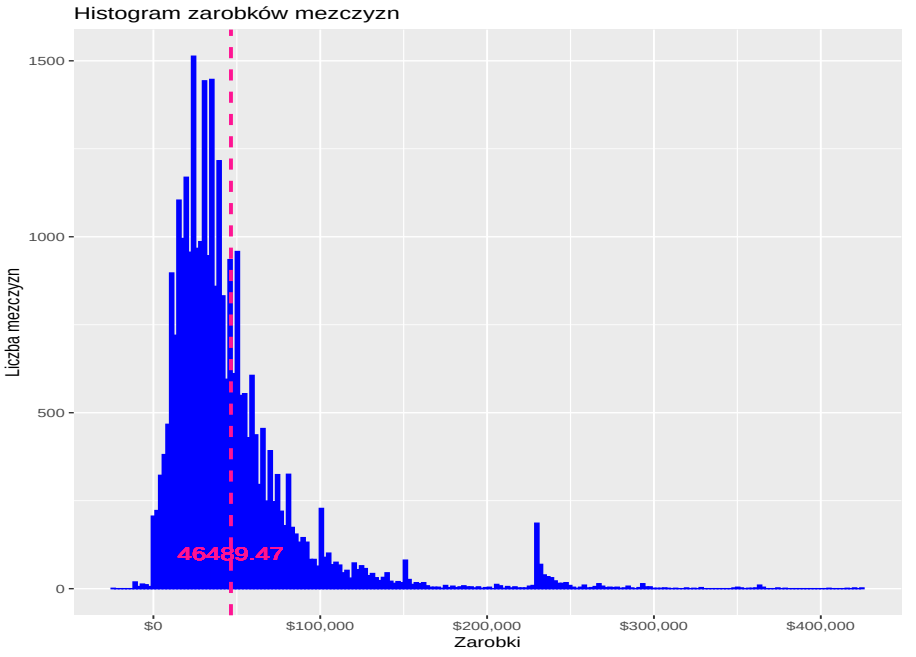
Faktycznie, osoby o wyższym wykształceniu zarabiają średnio o prawie $30 000 więcej od osób, które nie ukończyły studiów.

Jedyną zmienną, której dotychczas nie brałam pod uwagę jest wiek badanych osób. Po obliczeniach okazało się, że najliczniejszymi grupami są osoby w przedziale wiekowym 30-40 i 40-50 i ta wiedza niewiele daje, ponieważ dane mówią o ludziach pracujących. Gdybyśmy mieli badać, np. bezrobocie, to wiek byłby istotną informacją, jednakże w przypadku osób pracujących zdarzą się zarówno jednostki mające 25 lat i zarabiające bardzo dużo, jak i osoby w przedziale wiekowym 40-65 posiadające zarobki poniżej średniej.

3.2 Analiza zarobków w zależności od płci

Liczba badanych kobiet to 26685 osób, natomiast mężczyzn 29214.





Z histogramów wynika, że kobiety średnio zarabiają znacząco, bo aż o prawie

$20000, mniej niż mężczyźni. Może to byc potwierdzenie problemu szczególnie poważnego w Stanach Zjednoczonych, czyli *Gender Pay Gap*. Oznacza to, że kobiety zarabiają mniej pracując na tych samych stanowiskach co mężczyźni. Gender Pay Gap jest bardzo złożonym problemem, na który składają się (niestety) stereotypy, a także oczywiste i wspierane urlopy macierzyńskie i mało popularne urlopy tacierzyńskie. Po wyznaczeniu statystyk można zauważyć, że zarobki mężczyzn osiągają także bardziej „ekstremalne" wartości niż zarobki kobiet. Minimum u mężczyzn wynosi aż $-24998, podczas gdy u kobiet $-9999. Posiadając wiedzę z następnego punktu, możemy stwierdzić, że najbardziej stratny mężczyzna prawdopodobnie zainwestował zdecydowanie więcej i zarobił zdecydowanie mniej niż najbardziej stratna kobieta. Minimum to jedyna statystyka, której wartość kobiety w badanej próbie mają większą.

3.3 Obserwacje odstające

Obserwacje odstające to takie, które nie mieszczą się w przedziale

$$[Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR]$$

, gdzie IQR oznacza odstęp międzykwartylowy, a Q_1 i Q_3 oznaczają kolejno pierwszy i trzeci kwartyl. W przypadku zarobków ten przedział to $[-27255, 90759]$. Co ciekawe, zawiera się w nim ujemne minimum, które wyznaczyliśmy wcześniej. Oznacza to, że w próbie są osoby, które więcej zainwestowały w firmę niż na niej zarobiły. Natomiast maksimum, które początkowo można by uważać za prawdopodobną daną, jest obserwacją odstającą. Co więcej, po dokonaniu odpowiednich obliczeń, okazało się, że w zbiorze jest 3163 obserwacji odstających i są to wartości większe niż górne ograniczenie przedziału.

Literatura

[1] `https://en.wikipedia.org/wiki/Academic_grading_in_the_United_States`

[2] `https://calculatorgpa.com/middle-school-gpa-calculator`

[3] `https://en.wikipedia.org/wiki/Intelligence_quotient#Current_tests`

[4] `https://en.wikipedia.org/wiki/Wechsler_Intelligence_Scale_for_Children`

[5] `https://pl.wikipedia.org/wiki/Iloraz_inteligencji`

[6] *Źródło dotyczące interpretacji punktacji testu PHCSCS-2*, `https://www.montgomeryschoolsmd.org/uploadedFiles/community-engagement/linkages-to-learning/Piers-Harris-2-Interpretation-Overview.pdf`

[7] `https://en.wikipedia.org/wiki/Histogram`