„Przeszłość Demograficzna Polski" 34, 2014

Radosław Poniat
Białystok

O wykorzystaniu wykresów pudełkowych do prezentacji danych demograficznych i o pożytku z użycia środowiska R z pakietem ggplot2

Historycy, z których wielu dopiero niedawno odkryło możliwość wykorzystania arkuszy kalkulacyjnych, a jeszcze większa grupa poprzestaje na edytorach tekstu traktowanych jako maszyny do pisania, nie są zazwyczaj świadomi bogactwa programów komputerowych stworzonych z myślą o prowadzeniu analiz statystycznych. Prezentowany tu artykuł ma przede wszystkim zwrócić uwagę na istnienie tego typu aplikacji oraz wskazać na ich możliwości, które zdecydowanie wykraczają poza ograniczony zestaw opcji oferowanych przez arkusze kalkulacyjne. Spośród licznych programów statystycznych i ogromnego zestawu zawartych w nich procedur uwaga została tu poświęcona cieszącemu się wzrastającą popularnością środowisku R oraz dostępnym w nim formom prezentacji graficznej, które powinny wzbudzić zainteresowanie wśród demografów historycznych. Opis jednej z rodzin takich wykresów został połączony z prezentacją sposobu ich interpretacji i omówieniem procedury ich generowania w programie R z dodatkiem ggplot2.

Pierwsza część artykułu zawiera omówienie samego programu/środowiska R. Część druga została poświęcona wykresowi pudełkowemu z równoczesnym wskazaniem na kilka jego wersji specjalnych, charakteryzujących się rozbudowanymi możliwościami przedstawiania rozkładu danych. Kolejna sekcja omawia procedury pozwalające na wygenerowanie wykresu pudełkowego za pomocą pakietu ggplot2. Część ostatnia zawiera podsumowanie artykułu.

1.

Konsekwencją trwającego od dziesięcioleci wzrostu znaczenia analiz statystycznych w biznesie, polityce i administracji oraz badaniach naukowych stał się rozwój wyspecjalizowanych przedsiębiorstw dostarczających swoim klientom szeroką paletę zaawansowanych programów statystycznych. Głównymi graczami na tym rynku są wielkie korporacje oferujące dobrze znane produkty, takie jak SPSS

(należący obecnie do IBM), SAS, Statistica, STATA. Firmy te zatrudniają tysiące pracowników i obracają co roku dziesiątkami milionów dolarów. Nawet jeśli niektóre z nich zachowują w swych nazwach terminy wskazujące na ich wywodzenie się ze świata akademickiego, to ich produkty są obecnie tworzone z myślą o przede wszystkim klientach biznesowych, gotowych zapłacić nawet tysiące dolarów za odnawiane co roku licencje. Odpowiadając na zapotrzebowanie zgłaszane przez przedsiębiorstwa, programy statystyczne stają się coraz lepiej dostosowane do badań marketingowych, operowanie nimi staje się prostsze, a rzadko używane funkcje zostają w nich ukryte. Zjawisko takie szczególnie dobrze widać na przykładzie tak popularnego wśród badaczy społecznych SPSS, ale inne korporacje też zmierzają w podobnym kierunku, o czym świadczy choćby JMP — program tworzony przez SAS Institute.

Opisany powyżej rozwój rynku programów statystycznych niesie też istotne konsekwencje dla badaczy pragnących wykorzystywać w swojej pracy naukowej metody ilościowe. Otrzymują oni do dyspozycji szeroką paletę produktów, oferujących możliwości daleko wykraczające poza to, na co pozwalają arkusze kalkulacyjne. Odbywa się to jednak kosztem wzrastających cen oprogramowania[1] oraz pociąga za sobą konieczność uczenia się procedur i opcji właściwych dla danej aplikacji. Badacz, który opanował prowadzenie analiz w SPSS, może mieć trudności w razie przejścia na SAS lub inny program, często też nie będzie mógł odczytać w nim danych zapisanych w formacie używanym przez inną aplikację. Zawsze też może się okazać, że wykorzystywany produkt jakiejś opcji nie oferuje lub wymaga w tym celu zakupu drogiego rozszerzenia. Wzrastające w środowisku naukowym niezadowolenie z takiej sytuacji doprowadziło do powstania nowych projektów mających na celu przełamanie dominacji rozwiązań oferowanych przez korporacje. Wydaje się, że spośród takich nowych niekomercyjnych propozycji[2] na szczególną uwagę zasługuje język R, który staje się obecnie groźnym konkurentem dla płatnych aplikacji[3].

Z technicznego punktu widzenia sam R jest nie tyle programem, ile raczej językiem programowania stworzonym z myślą o analizach statystycznych. Wywodzi się on z wcześniejszego rozwiązania o nazwie S, twórcami zaś jego pierwszej

[1] W wypadku osób zatrudnionych na uczelniach rozwiązaniem mogą być, oczywiście, różnego rodzaju licencje akademickie wykupywane przez jednostki. Wiążą się one jednak z pewnymi niebezpieczeństwami. Po pierwsze, zakupiony przez uczelnię program może nie do końca odpowiadać potrzebom badacza. Po drugie, studenci uczący się na nim statystyki mogą mieć w przyszłości trudności w używaniu innych aplikacji. Takie wychowywanie klienta stanowi zresztą jedną z najważniejszych przyczyn niższej ceny licencji akademickich.

[2] Obok opisywanego tu R, warto zwrócić uwagę na dwa inne projekty. PSPP stanowi darmową alternatywę dla SPSS i oferuje podobny sposób obsługi oraz zbliżony zestaw podstawowych funkcji. Z kolei Gretl został stworzony z myślą o zastosowaniach ekonometrycznych i może z powodzeniem służyć do dydaktyki oraz prowadzenia mniej skomplikowanych analiz.

[3] Wzrastająca rola R została już też dostrzeżona przez podmioty komercyjne. Możliwość wykorzystania komend pisanych w R oferują obecnie choćby SPSS oraz SAS.

wersji byli Ross Ihaka i Robert Gentleman, badacze związani z Uniwersytetem w Auckland[4]. Obecnie nadzór nad jądrem projektu sprawuje zespół ponad 20 autorów, wspomaganych przez liczne grono współpracowników. O popularności języka zadecydowały między innymi jego względna prostota oraz fakt, że jest on rozpowszechniany na licencji GNU, co oznacza pełną bezpłatność, prawo do komercyjnego wykorzystania oraz wprowadzania w nim dowolnych zmian. Co ważniejsze, wokół projektu powstało silne i dynamiczne środowisko twórców oraz użytkowników (kategorie te w wypadku R często trudno od siebie oddzielić), pojawiały się liczne publikacje, blogi i fora jemu poświęcone[5]. Aktywność taka zapewniła językowi szybki rozwój, a osobom z niego korzystającym wsparcie ze strony bardziej doświadczonych członków społeczności. Najważniejszym osiągnięciem osób skupionych wokół R jest ogromna liczba tak zwanych pakietów (*packages*), czyli dodatków do podstawowej wersji programu. Ich liczba jest zresztą trudna do oszacowania, obecnie w głównych repozytoriach jest ich około 6 tysięcy, ale należy tu doliczyć różnego rodzaju wersje testowe oraz rozwiązania prywatne, które nie są powszechnie udostępniane. Pakiety znacznie rozszerzają funkcjonalność R, pozwalają na dokonywanie w nim wielu dodatkowych operacji. Istnieją więc specjalne dodatki skierowane do genetyków, demografów, osób potrzebujących specjalnych rodzajów grafiki… Prowadzi to często do chaosu — nikt nie zna wszystkich pakietów, a wielu operacji można dokonać na kilka sposobów, w różnych dodatkach, za pomocą różniących się komend[6] — ale zapewnia też ogromną elastyczność. Dzięki pakietom nowe metody statystyczne najwcześniej pojawiają się właśnie w R. Podczas gdy użytkownicy innych programów mogą czasem latami czekać na wprowadzenie w nich jakiegoś rozwiązania, członkowie prężnego środowiska związanego z opisywanym tu językiem z reguły bardzo szybko opracowują dodatki implementujące nowe techniki obliczeniowe i prezentacyjne.

Fakt, że R jest przede wszystkim językiem programowania czy też środowiskiem programistycznym, powoduje, że w wersji podstawowej R nie przypomina programów komputerowych, do których od kilkunastu lat tak bardzo się przyzwyczailiśmy. Po jego zainstalowaniu użytkownik ma do czynienia z minimalistyczną aplikacją, przypominającą prosty edytor tekstu lub znaną starszym czytelnikom konsolę MS-DOS. W oknie takim można wpisywać komendy i funkcje, w nim

[4] Opis historii R można odnaleźć na stronie projektu oraz w licznych poświęconych mu publikacjach. W języku polskim warto tu polecić zwłaszcza prace Przemysława Biecka. Jest on między innymi autorem wyczerpującego wprowadzenia do całego środowiska, zob. Przemysław Biecek, *Przewodnik po pakiecie R*, wyd. 3, Wrocław 2014.

[5] Podstawową witryną internetową skierowaną do użytkowników R jest strona http://cran.r-project.org/. Można z niej pobrać sam program, zapoznać się z listą dodatków do niego oraz plikami pomocy.

[6] Dobry przykład mogą tu stanowić piramidy populacji. Piszący te słowa generował je w R na cztery sposoby, za pomocą trzech różnych pakietów. Jest też pewien, że istnieje jeszcze co najmniej kilka alternatywnych rozwiązań.

też pojawiają się wyniki. Ascetyczność podstawowej wersji R stanowi ogromne utrudnienie dla użytkowników, którzy są przyzwyczajeni do menu poleceń i klikania na wybrane opcje za pomocą myszki. Zamiast tego koniecznym staje się wpisywanie kodu, pilnowanie czy zostały w nim pozamykane wszystkie nawiasy i cudzysłowy, nazwy zaczynają się od małej czy wielkiej litery. Wszystko to sprawia, że pierwsze kroki w R mogą przypominać drogę przez mękę, a wielu potencjalnych odbiorców nawet nie próbuje do niego sięgać[7]. Na szczęście w ostatnich latach opracowano kilka aplikacji ułatwiających obsługę języka, ale odbywa się to kosztem znacznego zredukowania dostępnych w nim opcji i są one skierowane raczej do mniej zaawansowanych użytkowników[8]. W takiej sytuacji większość osób korzystających z R posługuje się w nim komendami tekstowymi. Nie musi to jednak oznaczać konieczności tworzenia własnych funkcji. Wiele z nich zostało już przygotowanych w formie gotowych poleceń. Jeśli na przykład użytkownik chce wyliczyć średnią arytmetyczną, minimum, maksimum i kwartale wraz z medianą, wystarczy, jeśli wpisze komendę summary, a następnie w nawiasie poda nazwę zmiennej, która ma być w ten sposób analizowana. Podstawowa wersja środowiska R oraz każdy pakiet pojawiają się wraz z instrukcjami, w których komendy takie zostały opisane. Jakość takich podręczników bywa, oczywiście, różna, obok bardzo wyczerpujących, pojawiają się też niezbyt pomocne i wymagające dużej wiedzy użytkownika, ale ich uważna lektura, pomoc udzielana na forach internetowych oraz metoda prób i błędów pozwalają zazwyczaj na dość szybkie opanowanie danej komendy. Dopiero gdy użytkownik jest niezadowolony z istniejących rozwiązań, pojawia się potrzeba pisania własnych funkcji, opierających się zresztą często na fragmentach już opracowanych komend.

Choć posługiwanie się komendami tekstowymi w miejsce graficznego interfejsu użytkownika (GUI) może się wydawać na pierwszy rzut oka rozwiązaniem niewygodnym, oferuje ono wiele zalet. Po pierwsze, pozwala na pełniejszą kontrolę nad prowadzonymi analizami, dokładniejsze śledzenia zachodzących procesów, określanie w nich dodatkowych parametrów. Zaawansowany użytkownik może dzięki temu wyjść daleko poza zakres opcji oferowanych mu przez autorów programów obsługiwanych przede wszystkim za pomocą klikania myszką. Po drugie, zachowanie raz wpisanych komend (co nakładki takiej jak RStudio robią nawet w odniesieniu do poprzednich sesji) pozwala na ich późniejsze powtarzanie. Daje to nie tylko szansę na szybkie odtwarzanie raz już przeprowadzonych analiz, ale

[7] Początkującym użytkownikom można polecić wiele podręczników ułatwiających zapoznanie się z podstawowymi funkcjami programu. Na uwagę zasługuje tu chociażby książka Johna M. Quicka, *Analiza statystyczna w środowisku R dla początkujących*, Gliwice 2011.

[8] Do rozwiązań takich należą między innymi R Commander, który przypomina nieco komercyjny SPSS, a także Deducer, Rattle oraz RExcel, integrujący się z arkuszami kalkulacyjnymi. Do nieco innej kategorii należy RStudio, będący upraszczającą wiele operacji nakładką na R, która jednak nadal wymaga posługiwania się komendami tekstowymi — korzysta z niego wielu bardziej zaawansowanych użytkowników, w tym i piszący te słowa.

też dzielenie się nimi ze współpracownikami (konieczne jest jedynie przesłanie danych i zapisanych komend), generowanie identycznie wyglądających wykresów, szybkie poprawianie błędów[9]. Po trzecie, użytkownik dysponujący już pewnym doświadczeniem w obsłudze środowiska R będzie przeprowadzał kolejne procedury szybciej niż ktoś korzystający z GUI. Posługując się znanymi sobie komendami (a w praktyce mało kto sięga w swych na co dzień prowadzonych analizach po więcej niż kilkadziesiąt zapamiętanych poleceń) zyskuje więc na szybkości, ma zapewnioną pełniejszą kontrolę nad procedurami i może je w elastyczny sposób modyfikować.

2.

Termin wykres pudełkowy (w literaturze można też spotkać określenia „skrzynkowy" lub „pudełkowy z wąsami") stanowi próbę przełożenia na język polski słowa *boxplot*[10]. Pomysłodawcą tej formy prezentacji danych jest znany statystyk John W. Tukey. Dzięki jego niezwykle wpływowej książce, dotyczącej technik analizy danych statystycznych[11], boxplot wszedł szybko do powszechnego użycia. Wykres pudełkowy miał zapewniać prosty w sporządzeniu i interpretacji (sam Tukey rysował je ręcznie) sposób na prezentowanie rozkładu danych oraz podstawowych miar go opisujących. Wykres ten nie oddaje więc wszystkich informacji związanych z dystrybucją zmiennej, a w to miejsce pozwala szybko dostrzec jej najważniejsze właściwości. Dobrego przykładu użyteczności takiego rozwiązania mogą dostarczyć dane pochodzące z badań Lidii A. Zyblikiewicz, dotyczących XIX-wiecznej populacji Krakowa[12]. Zaprezentowane na wykresie 1 histogramy rozkładu wieku służących płci żeńskiej zawierają wiele szczegółowych informacji, które do znacznej części analiz nie muszą być potrzebne, ale równocześnie wcale nie ułatwiają szybkiego wskazania na położenie podstawowych miar statystycznych. Tylko dominanta i rozstęp są na nich wyraźnie widoczne. Badacz pragnący przystępnie zaprezentować najważniejsze właściwości rozkładu może więc skorzystać na zastąpieniu histogramów pudełkami.

Widoczna na wykresie 2a podstawowa wersja boxplotu w wygodny sposób oddaje siedem miar statystycznych opisujących dystrybucję wieku krakowskich służących oraz pozwala na ich łatwe porównywanie między grupami. Pogrubione

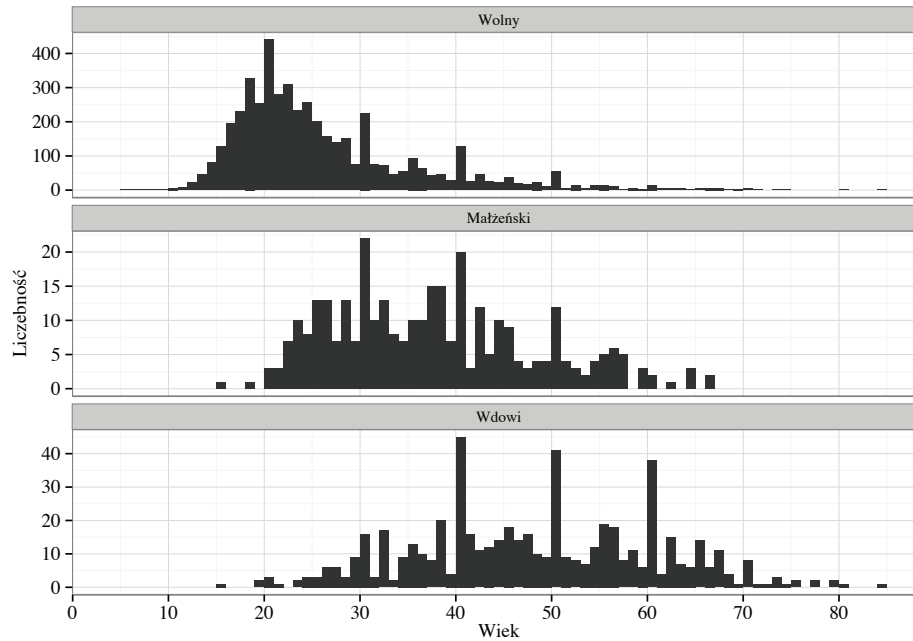
---

[9] Odkryta po wielu dniach literówka popełniona podczas przygotowywania zamieszczonego w tym artykule wykresu została poprawiona w ciągu minuty, a wymagało to jedynie ponownego uruchomienia pierwotnej komendy. Użytkownik Excela musiałby poświęcić znacznie więcej czasu na taką korektę, a uzyskany efekt wcale nie musiałby być identyczny. Ponownie opracowany wykres miałby zapewne nieco inne wymiary, legenda mogłaby się trochę przesunąć…

[10] Rodzimych historyków z wykresem skrzynkowym próbował zapoznać w swym podręczniku Michał Kopczyński, *Podstawy statystyki. Podręcznik dla humanistów*, Warszawa 2005, s. 42–45.

[11] John Tukey, *Exploratory data analysis*, Reading 1977.

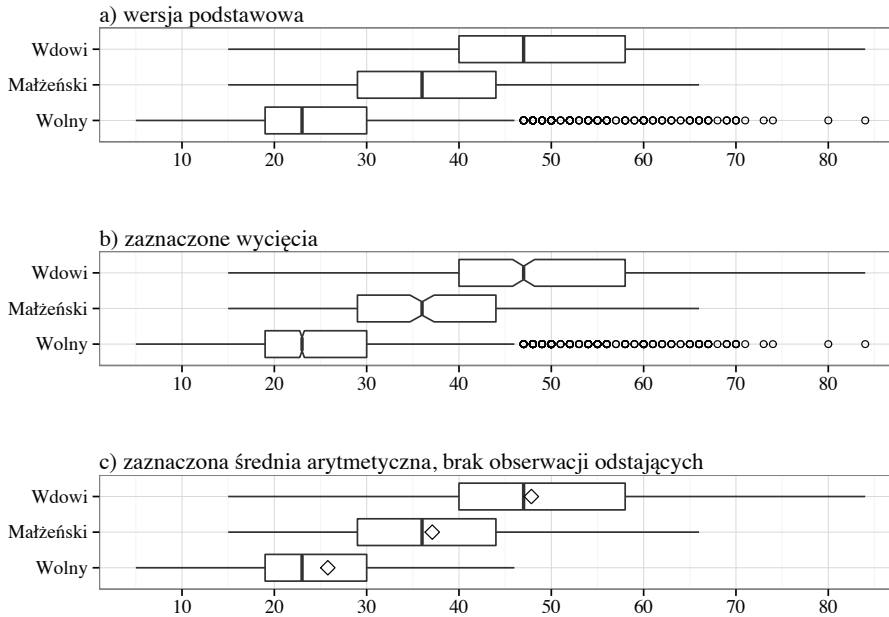
[12] Lidia A. Zyblikiewicz, *Kobieta w Krakowie w 1880 r. w świetle ankiet powszechnego spisu ludności. Studium demograficzne*, Kraków 1999.



Wykres 1. Histogramy rozkładu wieku kobiet służących w Krakowie w roku 1880 z podziałem na stan cywilny

Źródło: Lidia A. Zyblikiewicz, *Baza danych Spis ludności miasta Krakowa 1880*, Archiwum Państwowe w Krakowie, s. 108–134.

odcinki widoczne wewnątrz każdego pudełka wskazują na lokalizację mediany, czyli wartości określającej położenie środka rozkładu. Z kolei brzegi pudełek są wyznaczane przez pierwszy i trzeci kwartyl. Zobrazowanie położenia kwartyli i mediany, która zresztą jest w praktyce drugim kwartylem i tak też często bywa nazywana, pozwala odbiorcy na szybkie dostrzeżenie punktów dzielących badaną zbiorowość na ćwiartki. Poniżej pierwszego kwartyla znajduje się dolne 25% obserwacji, między pierwszym kwartylem a medianą kolejne 25% obserwacji charakteryzujących się większym nasileniem badanego zjawiska, ale wciąż niższym niż wartość środkowa, a w następnych przedziałach można odnaleźć obserwacje od mediany wyższe, także podzielone przez trzeci kwartyl na dwie ćwiartki. Kwartyle wskazują nie tylko na punkty dzielące daną grupę — za ich pomocą można też poznać wielkość rozstępu ćwiartkowego (zwanego też międzykwartylowym), czyli miary statystycznej wskazującej na zakres, w którym znajduje się środkowa połowa obserwacji. Ponieważ jest to odcinek zawierający się między pierwszym a trzecim kwartylem, na wykresie prezentuje go pudełko. Z kolei wychodzące z niego odcinki, nazywane „wąsami" lub „zawiasami", wskazują na zakres, w któ-



Wykres 2. Wykresy pudełkowe rozkładu wieku kobiet służących w Krakowie w roku 1880 z podziałem na stan cywilny

Źródło: jak wykres 1.

rym powinno się zawierać około 99% wszystkich obserwacji[13]. Punkty znajdujące się poza wąsami to tak zwane „obserwacje odstające", czyli w poważny sposób różniące się od reszty danej populacji. W wypadku grup niecharakteryzujących się znacznym zróżnicowaniem, punkty takie nie muszą się wcale pojawiać, co można dostrzec zresztą także na zaprezentowanym tu wykresie. Zakres między skrajnymi punktami lub wąsami przedstawia z kolei rozstęp — miarę zróżnicowania całej badanej grupy.

[13] W praktyce wyznaczony przez wąsy zakres może być jednak inny. Wynika to zazwyczaj z dwóch przyczyn. Po pierwsze, istnieje kilka konwencji określających sposób wyliczania zakresu wąsów: obok najczęstszego, który polega na nadaniu im długości wynoszącej 1,5 rozstępu ćwiartkowego z każdej strony pudełka, zaproponowano kilka innych, rzadziej spotykanych rozwiązań, dających nieco inne wyniki. Po drugie, w wypadku rozkładów zmiennej różniących się od rozkładu normalnego, opisany powyżej zakres wąsów może zawierać w sobie inny procent obserwacji. W związku z tym w praktyce lepiej nie traktować wąsów jako miar precyzyjnie informujących o położeniu 99% obserwacji, a w to miejsce interpretować je jako wskazujące na lokalizację prawie wszystkich badanych.

Ponieważ przedstawiony powyżej opis może się wydać czytelnikowi nadmiernie technicznym oraz przeładowanym nieznanymi mu miarami statystycznymi, warto więc przedstawić tu jego zastosowanie odnoszące się do codziennej praktyki badawczej. Posłużą temu zaprezentowane na wykresie 2a boxploty. Przede wszystkim, obserwując rozkład wieku służących stany wolnego można stwierdzić, że połowa spośród nich miała nie więcej niż 23 lata — wskazuje na to położenie mediany, zaznaczonej jako gruba linia w środku pudełka. Z kolei krańce pudełka wskazują, że najmłodsze 25% służących nie przekraczało 19. roku życia, aby zaś znaleźć się wśród 25% najstarszych służących, wystarczyło mieć więcej niż 30 lat. Wiek połowy służących stanu wolnego zawierał się między 19. a 30. rokiem życia. Najmłodsza z badanych kobiet miała zaledwie 5 lat, a najstarsza 84. Tak wysoki wiek był jednak rzadkością, długość wąsa wskazuje wyraźnie, że zdecydowana większość badanych służących nie przekraczała 46. roku życia. Łatwość interpretacji wykresu pudełkowego przekłada się też na szybkość i wygodę porównywania między sobą kilku boxplotów. Z zaprezentowanego tu wykresu bez trudu można wyczytać wyższy przeciętny wiek służących należących do stanów małżeńskiego i wdowiego, charakteryzujący te grupy większy rozstęp ćwiartkowy, częstsze i nie uznawane za wyjątkowe występowanie w nich osób starszych, przy równoczesnym braku kobiet mających mniej niż 15 lat. Prosty wykres dostarcza badaczowi zarówno łatwego wglądu w ogólne prawidłowości, jak i wielu szczegółowych miar statystycznych.

W świetle pokazanej powyżej łatwości interpretacji wykresu pudełkowego trudno się dziwić, że jest on jedną z najczęściej spotykanych i przez statystyków najbardziej cenionych form prezentacji graficznej. Szybko też pojawiły się jego alternatywne wersje, próby zastosowania go do danych wielowymiarowych lub dodania do niego kolejnych miar statystycznych. Z bogactwa takich propozycji z całą pewnością warto wskazać na dwie najbardziej popularne. Pierwsza z nich została zaprezentowana na wykresie 2b. Choć widoczne na nim pudełka nie różnią się znacznie od omawianych powyżej, dostrzec na nich można wcięcia (*notches*) okalające medianę. Wykorzystuje się je często w wypadku prezentacji wyników z próby, które pragnie się odnieść do całej populacji. W takiej sytuacji za pomocą wcięć oddaje się przedział ufności mediany, czyli zakres, w którym z określonym prawdopodobieństwem powinna się znajdować mediana całej populacji. Sposób wyliczania takiego przedziału jest wprawdzie w wypadku miar takich jak mediana nieco skomplikowany i nie powinien być traktowany jako odpowiednik właściwego testu istotności[14], ale wcięcia wciąż mogą pełnić ważną rolę informacyjną. Gdy wcięcia w pudełkach „zachodzą na siebie", badacz powinien być szczególnie

[14] Obustronny zakres wcięcia wylicza się, zazwyczaj dodając i odejmując od mediany rozstęp ćwiartkowy podzielony przez pierwiastek kwadratowy z liczby obserwacji i przemnożony przez stałą 1,57, zob. Robert McGill, John W. Tukey, Wayne A. Larsen, *Variations of box plots*, „The American Statistician" 32, 1978, z. 1, s. 12–16.

ostrożny i nie uznawać zaobserwowanych różnic między medianami za istotne. W sytuacji takiej istnieje bowiem możliwość, że otrzymane zróżnicowanie jest dziełem przypadku i nie oddaje rzeczywistych prawidłowości.

W prezentowanym tu przykładzie różnicę między medianą wyliczoną dla badanej grupy a przedziałem ufności, w którym powinna się zawierać mediana opisująca całą populację, widać dobrze w odniesieniu do służących będących wdowami. Wedle danych spisowych miały one przeciętnie 47 lat, ale przy założeniu, że zarejestrowane kobiety stanowiły jedynie frakcję całej populacji owdowiałych służących, należałoby stwierdzić, że rzeczywista mediana wieku z 95% prawdopodobieństwem powinna się zawierać w zakresie 46–48 lat. Podobny, dwuletni przedział ufności dostrzec też można w wypadku służących stanu małżeńskiego. Zdecydowanie mniejszy przedział widoczny w odniesieniu do kobiet stanu wolnego wynika przede wszystkim z dużej liczebności tej grupy, ale nawet tutaj można dostrzec nieznaczne wcięcie. Z punktu widzenia badacza kluczowe znaczenie ma tu jednak fakt, że wcięcia w żadnym wypadku się nie pokrywają, co stanowi mocny argument na rzecz tezy o zróżnicowaniu przeciętnego wieku służących ze względu na ich stan cywilny.

Innym stosunkowo częstym sposobem na rozszerzenie możliwości prezentacyjnych wykresu pudełkowego jest dodatkowe zaprezentowanie na nim średniej arytmetycznej. Na wykresie 2c została ona oddana za pomocą rombów. O ile w wypadku służących owdowiałych i zamężnych mediany i średnie są do siebie podobne, o tyle średnia wyliczona dla kobiet stanu wolnego jest wyższa od mediany o trzy lata. Tak wyraźna różnica stanowi konsekwencję występowania wartości odstających, znacznie odbiegających od przeciętnej i zawyżających przez to wynik. Ich istnienie widoczne było na wykresach 2a oraz 2b, tu zaś wskazuje na nie właśnie odstęp między średnią a medianą.

Zaprezentowane tu zalety wykresu pudełkowego nie oznaczają, że jego wykorzystanie nie wiąże się z pewnymi ograniczeniami. Podstawowa trudność w jego interpretacji wynika ze zróżnicowanej liczby obserwacji. Najczęściej wykorzystywane wersje tego wykresu nie pozwalają na wskazanie, że porównywane grupy mogą znacznie różnić się wielkością. Zjawisko takie wyraźnie widać w wypadku wykorzystywanych tu danych. Podczas gdy na histogramach na osi y bez trudu można dostrzec ogromną dysproporcję między służącymi stanu wolnego a pozostałymi kategoriami, wykresy pudełkowe fakt ten dość skutecznie ukrywają. Doświadczony badacz może wprawdzie zauważyć na wykresie 2b różnice w wielkości wycięć, co sugeruje większą liczebność niezamężnych służących, ale jest to jedynie dowód pośredni i nie dla każdego odbiorcy oczywisty. Graficznym rozwiązaniem problemu znacznej dysproporcji liczebności grup jest różnicowanie szerokości opisujących je pudełek. Zazwyczaj polega to na powiązaniu ich wymiarów z pierwiastkami kwadratowymi z liczby przynależących do danej grupy obserwacji. W praktyce jednak metoda taka wcale nie musi zapewniać wygodnego porównywania liczebności i na przykład do omawianego poniżej pakietu ggplot2

nie została ona wprowadzona. Wydaje się, że rozwiązaniem znacznie bardziej użytecznym może tu być raczej podawanie informacji dotyczących wielkości grup albo na wykresie (na przykład przy poszczególnych wartościach zmiennej grupującej), albo w sąsiadującym z grafiką tekście[15].

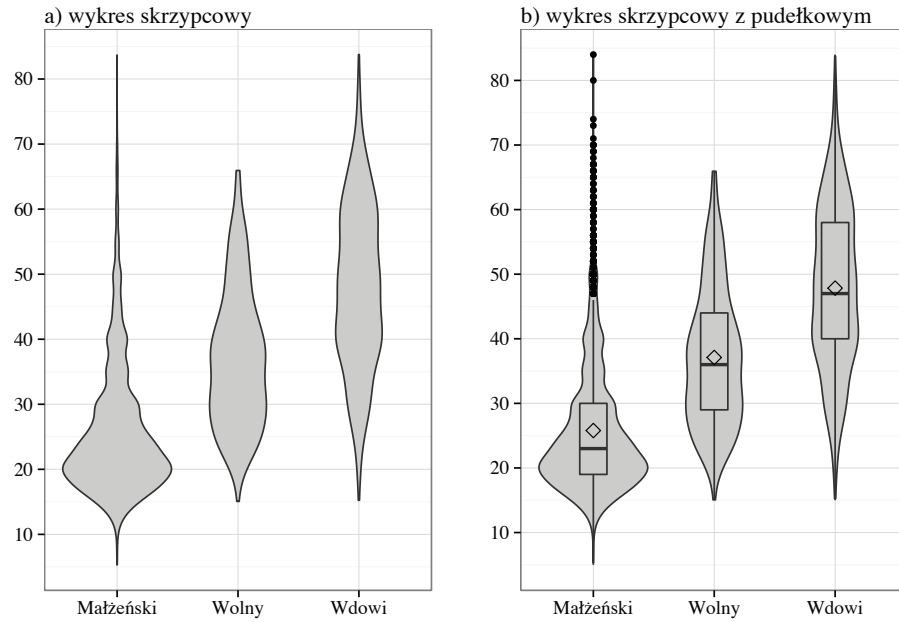
Drugie z podstawowych ograniczeń wykresu pudełkowego wynika z jego mocno uproszczonej, wręcz schematycznej formy graficznej. W wypadku rozkładów zmiennych charakteryzujących się kształtem mocno odbiegającym od rozkładu normalnego, w tym zwłaszcza rozkładów wielomodalnych[16], omawiany tu wykres nie wykaże ich istnienia. Konsekwencją takiego stanu może być poważny błąd badacza: wykorzystanie do analiz niewłaściwych miar statystycznych i sformułowanie nieadekwatnych wniosków. Jednym ze sposobów na uniknięcie takiego niebezpieczeństwa jest prezentacja danych na histogramie, ale odbywa się to zazwyczaj kosztem rezygnacji z łatwego podsumowania rozkładu oferowanego przez wykres pudełkowy. Na popularności zyskuje więc rozwiązanie alternatywne, polegające na wykorzystaniu wykresu skrzypcowego połączonego z pudełkowym.

Widoczna na wykresie 3a podstawowa wersja wykresu skrzypcowego nieco przypomina swym kształtem histogramy znane już z wykresu 1. Stanie się to szczególnie widoczne, jeśli weźmie się pod uwagę, że po dwóch stronach osi symetrii wykresu skrzypcowego zaprezentowane są lustrzane odbicia tego samego rozkładu. Spoglądając na tylko jedną jego połowę, czytelnik już bez trudu powinien dostrzec kształty zbliżone do tych zaprezentowanych wcześniej na histogramach. Jedyną różnicę stanowi tu ich wygładzenie, usunięcie z wykresu gwałtownych skoków (widocznych choćby w wypadku owdowiałych służących w 40., 50. oraz 60. roku życia) i widocznych spadków (przykładem takiego zjawiska mogą być zaskakująco nieliczne wdowy w wieku 49 lat). Opisane tu wygładzenie wynika z faktu, że wykres skrzypcowy w przeciwieństwie do histogramu nie oddaje prostego rozkładu zmiennej, a w to miejsce prezentuje tak zwany jądrowy estymator gęstości (*kernel density estimation*). Choć od strony matematycznej jego wyliczenie nie należy do najprostszych[17], wynik takiej estymacji może być interpretowany w sposób zbliżony do klasycznego histogramu. Nad tradycyjną formą estymator jądrowy ma

[15] W prezentowanym tu zbiorze znajdowało się 4837 służących stanu wolnego, 329 mężatek i 573 owdowiałe.

[16] Rozkład wielomodalny, to rozkład w którym zmienna ma więcej niż jedną wartość występującą najczęściej (dominantę, modę). W takiej sytuacji wiele klasycznych miar statystycznych charakteryzuje się ograniczoną użytecznością.

[17] Każda z obserwacji traktowana jest nie jako punkt, ale środek (jądro) rozkładu. W miejscach, gdzie wiele jąder występuje obok siebie, estymator przyjmuje większe wartości, podobnie do histogramu, który w takim punkcie charakteryzowałby się wyższym słupkiem. Jednak w przeciwieństwie do histogramu, w miejscach, gdzie żadna obserwacja nie wystąpiła, estymator może wciąż osiągać pewne wartości wynikające z istnienia rozkładów okalających jądra. W praktyce cała procedura jest nieco bardziej złożona, a jej wynik zależy od przyjętego kształtu rozkładu oraz parametru wygładzania. Stosunkowo przystępne wprowadzenie do zagadnienia można odnaleźć w pracy Randa Wilcoxa, *Intoduction to robust estimation and hypothesis testing*, wyd. 3, Amsterdam 2012, s. 46–54.



Wykres 3. Wykresy skrzypcowe rozkładu wieku kobiet służących w Krakowie w roku 1880 z podziałem na stan cywilny

Źródło: jak wykres 1.

w kilku punktach istotną przewagę: lepiej radzi sobie z prezentowaniem rozkładów wyliczanych na podstawie niedużych zbiorów danych i w mniejszym stopniu poddaje się zakrzywieniu przez pojedyncze skupienia obserwacji. Ta druga właściwość może być szczególnie użyteczna dla demografów historycznych, którzy nieustannie w swych badaniach muszą sobie radzić z trudnościami związanymi ze skupieniami wieku. Estymacja jądrowa pozwala na wygładzenie często obserwowanych na histogramach lub piramidach populacji wypadków nazbyt licznych deklaracji wieku zakończonego cyfrą 0.

W kontekście prezentacji graficznej za ważną zaletę wykresu skrzypcowego należy uznać fakt, że łatwo można go połączyć z wykresem pudełkowym. Przykłady takiego zestawienia zostały zaprezentowane na wykresie 3b. Pozwala on nie tylko na obserwację wcześniej już omówionych miar statystycznych — takich jak mediana czy rozstęp ćwiartkowy, które opisują wybrane aspekty zróżnicowania wieku krakowskich służących — ale też na śledzenie kształtu jego rozkładu. Ze skrzypiec można więc odczytać wyraźną koncentrację wieku służących stanu wolnego oraz brak takich skupień w wypadku pozostałych grup. Wśród owdowiałych służących dominował przedział 40–60 lat, a w jego ramach osoby w każdym roku

życia pojawiały się z podobną częstotliwością. W wypadku mężatek skupienie charakteryzujące się zbliżonym kształtem zawierało się między 27. a 44. rokiem życia. Wniosek taki może wydawać się sprzecznym z informacjami odczytanymi z histogramów zaprezentowanych na wykresie nr 1, ale w rzeczywistości jest on trafny — estymator jądrowy wygładził obserwowane wcześniej skupienia wieku i wskazał na rozkład bardziej zbliżony do rzeczywistości.

3.

Analizy statystyczne i generowanie wykresów w R musi być poprzedzone wprowadzeniem do programu danych, na których zostaną wykonane dalsze operacje. W wersji najprostszej sprowadza się to do ręcznego wpisania niezbędnych wartości. Dane takie zazwyczaj są wprowadzane w formie wektora, czyli zbioru jednorodnych informacji: liczbowych, tekstowo-kodowych, logicznych (true i false bez cudzysłowu). Dokonać tego można przy pomocy komendy:

nazwa wektora <- c(*ciąg liczb lub liter oddzielonych przecinkami*)

Nazwa wektora[18] może być dowolna i służy jedynie określeniu analizowanej zmiennej na czas trwania sesji. Ponieważ często będzie się pojawiać w kolejnych komendach, warto, aby była ona krótka, prosta, łatwa do zidentyfikowania i odróżnienia od innych wykorzystywanych w trakcie sesji zmiennych i zbiorów. Znak <– (można go zastąpić zwyczajnym znakiem równości) służy przypisaniu wartości do zmiennej. W praktyce oznacza to, że komendy lub dane podane po takim znaku będą w kolejnych procedurach przywoływane dzięki wpisaniu określenia znajdującego się przed znakiem. W powyższym wypadku ciąg liczb zostanie zapisany w pamięci programu pod nazwą nadaną mu przez użytkownika. Gdyby zamiast liczb wektor miał składać się ze słów lub liter, powinny one być wpisywane w cudzysłowach.

Obok ręcznego wprowadzania każdej wartości, co przy większych bazach nie jest zbyt wygodne, użytkownicy mogą korzystać z już istniejących plików zapisanych w bardzo wielu formatach. W części wypadków wymaga to wprawdzie sięgnięcia do bardziej wyspecjalizowanych pakietów, ale najczęściej spotykane rodzaje plików można wczytać za pomocą wersji podstawowej programu. Z praktycznego punktu widzenia najwygodniejszym rozwiązaniem jest tu zwłaszcza korzystanie z plików csv, które można zapisać i otworzyć w każdym arkuszu kalkulacyjnym (co pozwala na ewentualne modyfikacje danych w bardziej przyjaznym środowisku pakietu biurowego) oraz wielu edytorach tekstu. Podstawowa komenda pozwalająca na otwarcie takich plików w środowisku R wygląda następująco:

nazwa robocza zbioru <- read.csv("*nazwa pliku.csv*", header=**T**, sep=";", dec=",")

[18] Fragmenty komend wpisywane tu kursywą to nazwy plików, baz i zmiennych, nadawane przez samego użytkownika. Z kolei pogrubieniem zostały zaznaczone elementy, które mogą przyjmować wiele wartości, pochodzących z wykorzystywanego przez daną komendę zbioru określeń.

Określenie read.csv to już właściwa komenda, która uruchamia przygotowaną przez twórców programu funkcję wczytywania plików w formacie csv. W nawiasie znajdują się z kolei określane przez użytkownika parametry polecenia. Pierwszym z nich jest nazwa wykorzystywanego pliku (w wypadku, gdy znajduje się on w lokalizacji innej niż domyślna dla programu, trzeba podać jeszcze jego dokładną lokalizację), po niej następuje informacja o tym, że pierwszy wiersz pliku zawiera nazwy zmiennych, sep opisuje sposób, w jaki oddzielane są w nim kolumny (zamiast średnika mogą to być przecinki lub spacje), a dec określa format znaków dziesiętnych. Użytkownicy korzystający z arkuszy kalkulacyjnych, zapisujących dane według konwencji właściwych dla Europy, mogą skorzystać z uproszczonej wersji powyższej funkcji, w której wiele elementów zostało domyślnie określonych. Ma ona formę:

nazwa robocza zbioru <- read.csv2("*nazwa pliku.csv*")

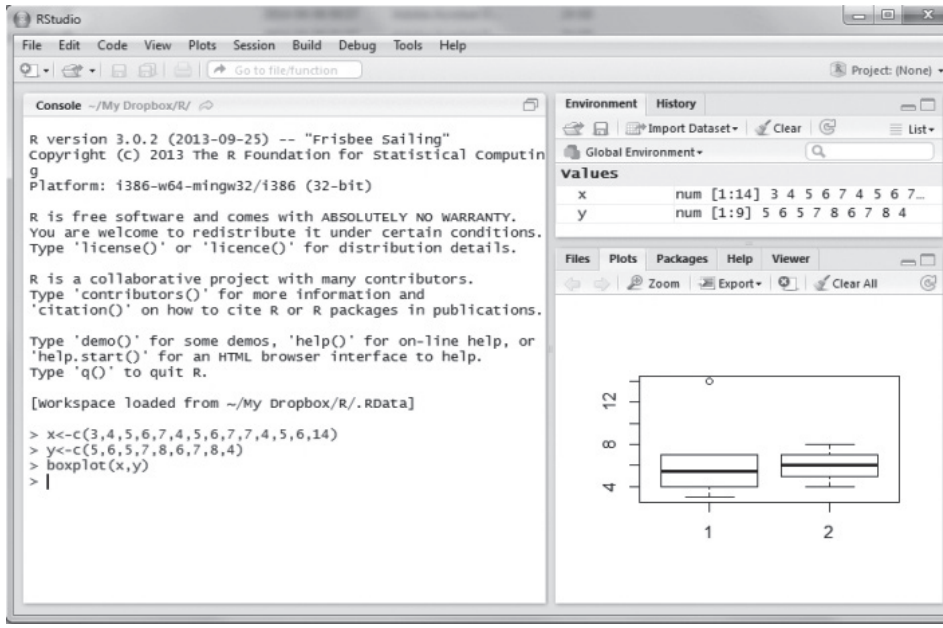
Tak wprowadzony zbiór danych może być odtąd poddawany analizom i przekształceniom w programie R. Wpisanie chociażby, wspomnianej już powyżej, komendy summary spowoduje więc wyliczenie podstawowych miar statystycznych opisujących rozkład zmiennej. Już sama podstawowa wersja programu oferuje dziesiątki takich procedur, zainstalowanie zaś dodatkowych pakietów może listę taką znacznie wydłużyć.

Po wprowadzeniu do programu wektorów lub bazy danych, wygenerowanie na ich podstawie wykresu skrzynkowego sprowadza się do wpisania komendy:

boxplot(*nazwy wektorów oddzielone przecinkami*)

Jeśli operacje wykonywane są na wprowadzonej do programu bazie danych, analizowane zmienne muszą zostać zapisane w sposób wskazujący na ich przynależność do bazy. Można tego dokonać na dwa sposoby: albo do komendy zostanie dodany zapis data="*nazwa robocza zbioru*", albo też nazwa każdej zmiennej musi być poprzedzona nazwą zbioru i znakiem $ (np. baza$zmienna). Wykres powstały w wyniku wpisania podanej powyżej komendy może być modyfikowany za pomocą dalszych poleceń. Tak więc wpisanie po nazwach wektorów terminu xlab="*nazwa*" pozwoli na nadanie nazwy osi X, a ylab uczyni to samo z osią Y. Podobnych komend jest więcej i pozwalają one na daleko posunięte modyfikacje, ale w opinii wielu osób wykresy generowane w podstawowej wersji R wciąż nie charakteryzują się szczególnie atrakcyjnym wyglądem. Zrzut ekranu programu R z nakładką RStudio, w którym na podstawie dwóch wektorów został wygenerowany wykres pudełkowy w jego podstawowej formie został zaprezentowany na ilustracji obok.

W ramach środowiska R znacznie atrakcyjniejsze wykresy, w tym i pudełkowe, pozwala łatwo generować wspomniany już na początku artykułu pakiet ggplot2. Dodatek ten, opracowany przez Hadley'a Wickhama, należy obecnie do najpopu-



larniejszych rozszerzeń wykorzystywanych w R oraz jeden z dwóch podstawowych pakietów służących do tworzenia zaawansowanych wykresów (drugim jest, oparty na nieco innej filozofii działania, dodatek o nazwie lattice). Podstawowe założenia określające sposób funkcjonowania ggplot2 wywodzą się ze znanej i wpływowej pracy Lelanda Wilkinsona[19]. Pakiet można określić wręcz mianem praktycznej realizacji teoretycznych założeń sformułowanych w tej książce. Zgodnie z propozycjami Wilkinsona tworzenie każdego wykresu rozpoczyna się więc od wskazania i ewentualnego przekształcenia wykorzystywanych danych, a następnie określa się formę ich graficznej prezentacji, wygląd osi i innych elementów. Zdefiniowania wykorzystywanych danych dokonuje się za pomocą komendy:

ggplot(*nazwa robocza zbioru*, aes(*zmienna z osi X, zmienna z osi Y*))

Nazwa zbioru odwołuje się do komend omówionych już powyżej, a zmienne prezentowane na osiach to wskazane przez użytkownika nazwy występujących w zbiorze kolumn. Obok zmiennych przedstawionych na osiach, mogą tu być zdefiniowane też kolejne zmienne, oddane na wykresie za pomocą kolorów, kształtów lub wielkości. Czyni się to poprzez wpisanie w nawiasie, rozpoczynającym się po aes, dodatkowych określeń. Na przykład polecenie fill=*zmienna grupująca* wskazuje na zmienną, ze względu na którą obserwacje będą się różnić kolorem wypełnienia. Zde-

[19] Leland Wilkinson, *The grammar of graphics*, New York 1999.

finiowane w ten sposób zmienne staną się podstawą dla generowanych wykresów, których kształt będzie określany w kolejnych częściach komendy, wprowadzanych po znaku dodawania. Dla wygody użytkownika podczas dalszych procedur, każda część tak rozbudowanego polecenia może zostać zapisana jako pojedynczy obiekt. Na przykład zamiast ponownie wpisywać definicję zbioru, wystarczy nadać mu nazwę, którą będzie się potem wpisywać podczas dalszej pracy:

nazwa obiektu <- ggplot(*nazwa robocza zbioru*, aes(*zmienna z osi X*, *zmienna z osi Y*))

Operacja taka może znacznie ułatwić dalsze procedury, skoro odtąd w miejsce długiego kodu będzie wystarczać wpisanie pojedynczego słowa czy wręcz znaku[20].

Uwaga poświęcona tu kwestii wstępnego zdefiniowania danych ma istotne uzasadnienie. Dalsze procedury związane z generowaniem grafiki w ggplot2 są uzależnione od tego, jak pakiet będzie odczytywać dane. Jeśli nie będą one spełniały określonych warunków, kolejne etapy nie zostaną zrealizowane, a użytkownik zobaczy jedynie długie linijki dziwnie brzmiących ostrzeżeń. Przede wszystkim należy pamiętać, że zmienne wykorzystywane w ggplot2 muszą zostać wprowadzone w formie tabeli danych, czyli kształcie przypominającym to, co generują zazwyczaj arkusze kalkulacyjne. Oddzielne ciągi zmiennych (zwane w środowisku R wektorami) nie zostaną przez pakiet przyjęte. Pakiet nie poradzi sobie też ze zmiennymi o różnej liczbie elementów, choć zazwyczaj bez trudności przetwarza odpowiednio zdefiniowane braki danych. W zależności od rodzaju wykresu konieczne jest też odpowiednie zdefiniowanie typów zmiennych znajdujących się na osiach. Wykres skrzynkowy wymaga, aby pierwsza zmienna, dla której będą wyliczane oddzielne boxploty, była zmienną nominalną lub porządkową (tutaj rolę tę pełni stan cywilny — cecha mierzona na poziomie nominalnym). Program sam zinterpretuje w taki sposób zmienne zapisane za pomocą liter (np. „kobieta", „mężczyzna"). W wypadku zmiennych zakodowanych za pomocą cyfr (co jest rozwiązaniem często stosowanym przy zapisie danych), użytkownik musi poprzedzić je określeniem factor, które wskaże programowi na ich dyskretny charakter. Zmienna grupująca zostanie więc zapisana jako: factor(*zmienna z osi X*). Druga z prezentowanych na wykresie zmiennych musi być ilościowa (czyli pozwalać na wyliczenie średniej arytmetycznej — w prezentowanych tu przykładach jest to wiek). W wyjściowej formie wykresu zmienna ilościowa zostanie przedstawiona na osi y, a grupująca na osi x, powinno się tak dziać nawet w sytuacji, gdy ostateczna forma wykresu przyjmie inny układ, co osiąga się poprzez komendę coord_flip(), która pozwala na przestawienie osi.

Po odpowiednim zdefiniowaniu zmiennych samo wygenerowanie wykresu pu-

[20] R dysponuje dodatkowo wygodną opcją przywoływania wcześniej wykorzystanych funkcji za pomocą klawiszy strzałek. Może to znacznie przyspieszać wpisywanie kodu.

dełkowego wymaga jedynie wpisania po znaku dodawania komendy geom_boxplot(). W pełnym zapisie przyjmie ona formę:

```
ggplot(nazwa robocza zbioru, aes(zmienna z osi X, zmienna z osi Y)) + geom_boxplot()
```

Modyfikacji podstawowej wersji pudełka dokonuje się poprzez wpisywanie dalszych komend w nawiasie znajdującym się po terminie boxplot. Wcięcia wymagają więc polecenia notch=TRUE, a ukrycie wartości odstających wymaga zapisu outlier.color=NA. Opcji przekształcających wygląd wykresu jest, oczywiście, więcej — są one omówione w załączonych do pakietu ggplot2 plikach pomocy.

Tylko nieznacznie bardziej skomplikowane jest dodanie do wykresu punktu wskazującego na położenie średniej arytmetycznej. Wystarczy w tym celu wpisać po znaku dodawania komendę:

```
stat_summary(fun.y="mean",geom="point",colour="black",shape=5, size=4)
```

Polecenie stat_summary nakazuje zaznaczenie na wykresie statystyki opisującej rozkład, a termin podany po fun.y określa, że ma to być średnia arytmetyczna. Kolejne, oddzielane przecinkami części komendy opisują sposób, w jaki średnia owa zostanie na wykresie zaprezentowana. W tym wypadku ma to być punkt, zaznaczony na czarno, w kształcie rombu (wskazuje na to cyfra 5 po słowie shape – listę dostępnych znaków i kodów je określających bez trudu można odnaleźć w pomocy pakietu) i o wielkości równej 4. Użytkownik może bez wielkiego trudu zmodyfikować takie ustawienia, wygenerować punkt w innym kształcie, kolorze, wielkości, a nawet oddający inną miarę statystyczną.

Dodanie do przygotowywanej grafiki wykresu skrzypcowego następuje za pomocą polecenia geom_violin(). W wypadku nakładania na siebie kilku wykresów należy pamiętać jednak o ich kolejności. Ponieważ skrzypce mają większą powierzchnię niż pudełko, powinny one być wyrysowane jako pierwsze, tak aby nie zasłaniały następującego po nich wykresu. Ciąg komend generujących podstawowe elementy wykresu 3b powinien więc wyglądać następująco:

```
ggplot(nazwa robocza zbioru, aes(zmienna z osi X, zmienna z osi Y))
+ geom_violin(fill="grey80") +geom_boxplot(fill="grey80", width=0.3)
+ stat_summary(fun.y="mean", geom="point", colour="black", shape=5, size=4)
```

Podane w dwóch miejscach określania fill="grey80" wskazują na kolor wypełnienia skrzypiec oraz pudełka i mogą być modyfikowane, w zależności od potrzeb oraz upodobań użytkownika. Z kolei polecenie width=0.3 określa szerokość pudełka[21]. W tym wypadku została ona tak zredukowana, aby krawędzie nie wystawały poza obręb skrzypiec.

[21] Ponieważ pierwszymi twórcami środowiska R byli Anglosasi, domyślnym separatorem dziesiętnym jest w nim kropka, a nie przecinek.

Podany powyżej zbiór komend, liczący około 200 znaków, pozwoli na wygenerowanie wykresu będącego w rzeczywistości kombinacją trzech różnych form prezentacji graficznej. Użytkownik pragnący poddać go dalszym modyfikacjom, mającym na celu jego lepszy opis lub uatrakcyjnienie, może sięgnąć do bogatego zestawu dodatkowych poleceń. W wypadku grafiki zaprezentowanej na wykresie 3b komendy takie dotyczyły przede wszystkim wyglądu osi. Sposób wyświetlenia opisu zmiennej grupującej został zdefiniowany za pomocą komendy:

```
+scale_x_discrete("", labels=c("Wolny","Małżeński»,»Wdowi»))
```

Ponieważ na osi x została przedstawiona zmienna nominalna, została tu wykorzystana komenda odnosząca się do skal dyskretnych. W następującym po niej nawiasie na pierwszym miejscu należy podać tytuł osi. W widocznym tu przykładzie pojawia się pusty cudzysłów, co oznacza, że żadna nazwa nie zostanie na wykresie wyświetlona. Z kolei polecenie labels definiuje wektor z nazwami grup. Niemal identyczne komendy pozwalają też na zdefiniowanie wyglądu osi y:

```
+scale_y_continuous("", breaks=seq(0,90,10))
```

Podstawowa różnica dotyczy tu ilościowego charakteru zmiennej. W związku z tym została tu wykorzystana komenda odnosząca się do skali ciągłej. Także w tym wypadku osi nie nadano nazwy. Polecenie breaks stanowi tu bezpośredni odpowiednik wcześniejszej komendy labels. Wskazuje ono, że widoczne na osi wartości powinny rozpoczynać się od cyfry 0, dochodzić do 90, a ich przyrost ma następować co 10 jednostek. Dodanie na końcu komendy theme_bw() decyduje o minimalistycznej formie wykresu, białym tle i jasnych liniach pomocniczych. Pełne polecenie generujące wykres 3b wygląda więc następująco:

```
ggplot(nazwa robocza zbioru, aes(zmienna z osi X, zmienna z osi Y))
+ geom_violin(fill="grey80") + geom_boxplot(fill="grey80", width=0.3)
+ stat_summary(fun.y="mean" ,geom="point" ,colour="black", shape=5, size=4)
+ scale_x_discrete("",labels=c("Wolny","Małżeński»,»Wdowi»))
+ scale_y_continuous("", breaks=seq(0,90,10)) + theme_bw()
```

Choć na pierwszy rzut oka wydaje się być ono rozbudowane i skomplikowane, w praktyce wymaga znacznie mniejszego nakładu pracy niż wygenerowanie podobnego wykresu w klasycznym programie statystycznym, w którym zresztą wszystkie wykorzystane tu opcje wcale nie muszą być dostępne. Co więcej, raz napisana komenda może zostać zachowana i po drobnych modyfikacjach wielokrotnie używana w przyszłości, podczas gdy użytkownik aplikacji wykorzystującej GUI w każdym wypadku będzie musiał rozpoczynać całą procedurę od zera, ponownie klikać i wybierać poszczególne opcje.

4.

Choć artykuł ten powstał przede wszystkim w celu zaprezentowania wykresu pudełkowego oraz sposobu jego generowania za pomocą pakietu ggplot2, warto też wskazać na kwestie wykraczające poza proste i przywodzące na myśl książkę kucharską zagadnienie przepisu do dobry wykres. Omówiony tu wykres pudełkowy stanowi wszak tylko jedną z dostępnych w ggplot2 form prezentacji graficznej, a samo środowisko R jest więcej niż tylko narzędziem do generowania wykresów. Istnienie darmowego programu, który nie tylko może konkurować z płatnymi aplikacjami, ale jest wręcz od nich lepszy, bardziej zaawansowany i w wielu dyscyplinach naukowych staje się standardowym narzędziem pracy badawczej, nie powinno umknąć uwadze historyków korzystających z metod statystycznych. Zamiast przeznaczać znaczne środki na zakup komercyjnych aplikacji, warto może sięgnąć po program darmowy i wszechstronniejszy. Jego wykorzystanie ograniczałoby się zapewne na początku do prowadzenia bardziej zaawansowanych analiz, których nie oferują rozwiązania płatne, ale można mieć nadzieję, że z czasem historycy zaczną traktować R jako podstawowe narzędzie służące analizom statystycznym, a następnie też dydaktyce.

Szersze wykorzystanie programów komputerowych może też odegrać istotną rolę w podniesieniu poziomu prowadzonych przez historyków badań kwantytatywnych. Uwalniają one użytkownika z obowiązku prowadzenia żmudnych obliczeń i pamięciowego opanowania wzorów, dają dostęp do metod statystycznych, z których nigdy nie będzie mógł skorzystać badacz posługujący się kalkulatorem czy nawet arkuszem kalkulacyjnym. Zamiast poświęcać czas stronie technicznej, użytkownik programów do analizy statystycznej może swoją uwagę skierować na zagadnienie interpretacji wyników oraz przystępnej ich prezentacji. Dobry przykład takiego zjawiska stanowi omówiony tu wykres skrzypcowy — choć pod względem matematycznym jądrowe estymatory gęstości są dość skomplikowane, ich zastosowanie we współczesnej aplikacji komputerowej sprowadza się do wpisania jednej komendy, a uzyskany wynik można za pomocą prostej formy graficznej zaprezentować przeciętnemu odbiorcy. Nawet zresztą do wygenerowania mniej skomplikowanego wykresu pudełkowego konieczne jest przeprowadzenie całego szeregu obliczeń statystycznych, w ten przecież sposób wyznaczane zostaje położenie mediany, kwartyli, obserwacji odstających… Biorąc pod uwagę bogactwo zaprezentowanych w ten sposób informacji trzeba też stwierdzić, że wykres taki przestaje być jedynie atrakcyjnym urozmaiceniem wywodu i zaczyna pełnić rolę ważnego jego uzupełnienia, a czasem wręcz najważniejszej części.