



lubuje się w budowaniu olbrzymich teleskopów w celu obserwacji nieba. Dla przykładu budowany właśnie w Chile *Large Synoptic Survey Telescope* (planowana data uruchomienia to 2019 rok) będzie teleskopem z ponad 8 metrowym lustrem, za którego pomocą naukowcy mają zamiar zmapować drogę mleczną oraz małe obiekty w układzie słonecznym czy wykrywać przemijające zdarzenia optyczne jak np. wybuchy supernowych. W tym celu teleskop będzie wykonywał zdjęcia całego dostępnego nieba raz na kilka nocy, a jedno takie zdjęcie będzie ważyło aż 30 TB. Dla porównania cały genom ludzki zajmuje mniej niż 1 GB, a 1 TB to odpowiednik 2 milionów książek (ciekawe ile czasu zajęłoby ci ich przeczytanie?). Jeżeli nadal dane astronomów nie zrobiły na tobie wrażenia, to spokojnie — już na 2020 rok planowane jest uruchomienie innego zestawu teleskopów, *Square Kilometre Array*, który będzie generował 1 EB (=1.048.576 TB) danych w każdej sekundzie swojego działania.

W środowiskach naukowych zaczyna się nawet mówić o nowym, stale rozwijającym się, czwartym paradygmacie nauki: nauki opartej na przetwarzaniu danych (ang. *science based on data-intensive computing*). W nauce opartej na tym paradygmacie nowe odkrycia będą/są dokonywane poprzez szeroką analizę danych pochodzącą z eksperymentów lub też po prostu danych o świecie (zdjęcia, nagrania, teksty). Z analizy danych będą wypływały także hipotezy badawcze np. poprzez automatyczną analizę artykułów naukowych (ang. *literature-based discovery*). Takiej analizie poddaje się zbiór artykułów np. dotyczących różnych reakcji i interakcji między substancjami. System sam potrafi „przeczytać” te artykuły, zbudować graf zależności pomiędzy substancjami, a następnie przeanalizować brakujące powiązania (nie sprawdzone dotychczas przez naukowców) w celu znalezienia tych o interesujących charakterystykach chemicznych. Takie powiązania są następnie sprawdzane przez naukowców i w ten sposób, o ile system się nie pomylił, dokonujemy nowego, interesującego odkrycia naukowego. Ciekawą i darmową książkę na temat tego rodzącego się paradygmatu w nauce udostępnił Microsoft Research [7].

Oczywiście astronomowie i inni naukowcy nie poradzą sobie sami z oglądaniem, analizowaniem i odkrywaniem wiedzy z tych olbrzymich danych. Podobnie analitycy bankowi nie będą w stanie dokonywać swoich analiz finansowych w arkuszach kalkulacyjnych, bo danych będzie zbyt dużo. Co więcej, wiele z danych gromadzonych przez bank będzie dla nich bezużyteczna, bo nie będą potrafili zastosować zaawansowanych technik ich analizy i przetworzenia. To rodzi na rynku pracy nową niszę na wysoko wykwalifikowanych specjalistów: inżynierów umiejących przetwarzać duże wolumeny danych wraz z wysokimi umiejętnościami analitycznymi i dużą wiedzą o statystyce i analizie danych. Takiego specjalistę nazywamy badaczem danych (ang. *data scientist*)<sup>1</sup>

### 1.1.2 Nowy zawód: Data Scientist

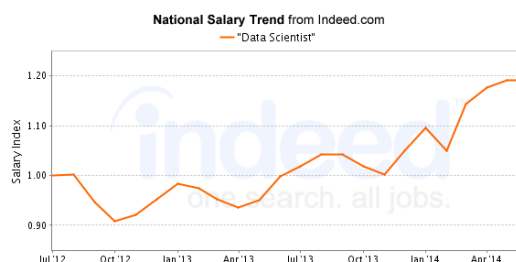
Najpierw zadajmy sobie pytanie: kim jest badacz danych? Ponieważ sama nauka o danych jak i skojarzony z nią zawód są terminami bardzo młodymi, nie istnieje żadna uniwersalna definicja tego terminu. Dlatego też pragnę przytoczyć w tym miejscu kilka z nich [1]:

- Badacz danych posiada unikalny zestaw umiejętności, które umożliwiają zrozumienie danych oraz opowiedzenie niesamowitej historii stojącej za nimi (ang. *A data scientist is that unique blend of skills that can both unlock the insights of data and tell a fantastic story via the data*)

<sup>1</sup> Polskie tłumaczenie zostało zaczerpnięte z książki [9], którą polecam zainteresowanym osobom w celu rozszerzenia wiadomości o eksploracji danych.

- Badacz danych jest rzadką hybrydą informatyka z umiejętnością zbudowania oprogramowania, które zbiera i zarządza danymi z wielu różnych źródeł, oraz statystyka, który wie jak wydobyć z nich interesujące wnioski. On/ona łączy te umiejętności w celu budowania nowych prototypów oraz odpowiadaniu na pytania dotyczące najgłębszych sekretów danych ze szczególną kreatywnością oraz dokładnością. (ang. *A data scientist is a rare hybrid, a computer scientist with the programming abilities to build software to scrape, combine, and manage data from a variety of sources and a statistician who knows how to derive insights from the information within. S/he combines the skills to create new prototypes with the creativity and thoroughness to ask and answer the deepest questions about the data and what secrets it holds*)
- Badacz danych posiada umiejętności inżynierskie do pobierania i zarządzania dużymi zbiorami danych, umiejętności statystyczne aby wyciągać wartościowe wnioski ze zbiorów danych oraz umiejętnością prezentowania danych szerokiej publiczności. (ang. *An ideal data scientist is “someone who has the both the engineering skills to acquire and manage large data sets, and also has the statistician’s skills to extract value from the large data sets and present that data to a large audience*)

Zapotrzebowanie na badaczy danych na rynku pracy stale rośnie, co wyraża się poprzez okrzyknięcie badacza danych mianem „the sexiest job of the 21st century”<sup>2</sup>[5]. Hal Varian, główny ekonomista w firmie Google powiedział: „The sexy job in the next 10 years will be statisticians. People think I’m joking, but who would’ve guessed that computer engineers would’ve been the sexy job of the 1990s?”. Szacuje się, że do roku 2020 rynek dużych danych będzie rósł aż sześć razy szybciej niż cały rynek IT (który i tak rośnie dość szybko). Według analityków spowoduje to utworzenie wielu nowych miejsc pracy, a na ponad 100 tysięcy z nich będzie brakowało wykształconej kadry (dane o USA). To rosnące zapotrzebowanie przekłada się na rosnące zarobki badaczy danych (patrz rysunek 1.2), a już dzisiaj przeciętny badacz danych zarabia w Stanach Zjednoczonych ok. 20% więcej niż inżynier oprogramowania.



Rysunek 1.2: Trend zarobków na stanowisku badacza danych wg. portalu Indeed [3]

Jeżeli do wysokich zarobków dołożymy możliwość robienia w pracy ciekawych rzeczy (patrz 1.1.3), zostanie badacz danych staje się interesującym sposobem na życie zawodowe. Rodzi się wtedy również pytanie: „jak zostać badaczem danych?”. Ponieważ pytanie to jest stawiane ostatnio coraz częściej została przygotowana specjalna grafika 1.3 przedstawiająca krok po kroku jakie umiejętności trzeba zdobyć. Studenci 2 roku kierunku „Informatyka” z pewnością opanowali już umiejętności z programowania oraz podstawowe umiejętności z baz danych. Na przedmiocie „Statystyka i analiza danych” zdobędziecie umiejętności oznaczone na rysunku na czerwono „2. Statistics”. Spojrzenie na kolejne punkty na czerwonej ścieżce, oznaczające umiejętności zaliczane do tej grupy, pokrywa się z programem tego przedmiotu. Jeśli więc jesteś ciekawy czego będziesz się uczył w

<sup>2</sup>Słowo „sexiest” jest tu użyte jako synonim pożądanego... niestety, na rynku pracy.

# MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

## MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

## PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

## DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

## COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



**MarketingDistillery.com** is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

*Marketing*  
**DISTILLERY**

Rysunek 1.1: Kim jest badacz danych? [11]





- automatyczne tłumaczenie tekstów (np. Google Translate)
- helikoptery, lub ogólniej roboty, same uczące się wykonywania zadań i poruszania się w zmieniającym się świecie
- określanie wieku i płci na podstawie zdjęcia <http://how-old.net>
- filtry SPAMu w poczcie elektronicznej (np. Gmail)
- rekomendacja nowych znajomych na portalach społecznościowych (np. Facebook)
- dostosowywanie reklam pod zainteresowania użytkownika (np. Google Adwords)
- automatyczna detekcja inwazji grypy na podstawie wyszukiwań użytkowników (Google Flu Trends, <https://www.google.org/flutrends/about/>)
- sztuczna inteligencja w grach
- automatyczne wykrywanie nowych zagrożeń (np. wirusów) w systemach komputerowych
- estymacja zatłoczenia ulic i czasu przejazdu (np. Google Maps, NaviExpert)
- wykrywanie oszustw np. podejrzanych transakcji bankowych
- analiza gry piłkarzy i podejmowanie decyzji o transferach lub strategii na mecz (np. Milan Lab, [http://www.acmilan.com/en/club/milan\\_lab](http://www.acmilan.com/en/club/milan_lab))
- automatyczna analiza nieba w celu detekcji nowych obiektów niebieskich

## 1.2 Organizacja zajęć

### 1.2.1 Zasady zaliczenia przedmiotu

Zasady obowiązujące studentów są następujące:

- każde laboratorium zaczyna się kartkówką (łatwe, zwykle obejmujące tylko ostatnie zajęcia, 40% oceny)
- przewiduje się jedno większe zadanie domowe na ocenę (dość proste, 15% oceny) oraz dodatkowo kilka mniejszych zadań domowych liczonych jako kartkówki.
- studenci piszą dwa kolokwia z wykładu (trudne), które równocześnie wpływają na ocenę z laboratoriów (45%)
- zaliczenie laboratoriów od 51% (3), próg każdej następnej oceny rośnie o 10% (np. 3.5 jest od 61%)
- dopuszcza się 2 nieusprawiedliwione nieobecności studenta na laboratoriach
- w przypadku nieobecności student otrzymuje 0 punktów z kartkówki
- 2 najgorsze wyniki z kartkówek są pomijane przy ocenie (podsumowując: korzystasz z prawa do 2 nieobecności = masz dwa „0”, liczę średnią z pozostałych kartkówek; chodzisz na wszystkie laboratoria = masz bonus w postaci pominięcia dwóch najgorszych wyników)
- w przypadku nieterminowego oddania zadania domowego odejmuje się 10% od ostatecznego wyniku za każdy rozpoczęty dzień spóźnienia
- w przypadku komunikacji mailowej z prowadzącym ([mateusz.lango@cs.put.poznan.pl](mailto:mateusz.lango@cs.put.poznan.pl)) uprzejmie proszę o rozpoczynanie tematu maila od skrótu przedmiotu: „[SiAD]”.
- materiały do przedmiotu są zamieszczane na ogólnodostępnej stronie internetowej <https://ophelia.cs.put.poznan.pl/webdav/ad/students/>, dodatkowe materiały dotyczące konkretnie prowadzonych przeze mnie grup i ew. ogłoszenia będą pojawiać się na mojej stronie internetowej [www.cs.put.poznan.pl/mlango](http://www.cs.put.poznan.pl/mlango) w zakładce „Teaching”.
- student przyłapany na ściąganiu lub plagiatowaniu zadań domowych otrzymuje



ocenę niedostateczną, niezależnie od innych ocen



Dodatkowo, zgodnie z Regulaminem Studiów Politechniki Poznańskiej:




- nieobecności studenta, w tym usprawiedliwione, przekraczające 1/3 zajęć są podstawą do niezaliczenia zajęć
- student zobowiązany jest do usprawiedliwienia u prowadzącego nieobecności na zajęciach w ciągu dwóch tygodni

### 1.2.2 Oprogramowanie statystyczne

Istnieje wiele dobrych pakietów statystycznych, oferujących zarówno podstawowe jak i zaawansowane metody analizy danych. Wśród nich pakiet Statistica (darmowy dla studentów PP poprzez platformę eProgramy), SPSS czy Stata.

W ostatnich latach dużą popularność w środowiskach analityków danych zyskał język/pakiet statystyczny , który w 2015 roku stał się 6 najbardziej popularnym językiem programowania na świecie, wyprzedzając PHP, JavaScript czy Ruby. Z kolei już w ankiecie branżowego portalu KDnuggets, badacze danych zapytani o oprogramowanie używane przez nich w ostatnim roku najczęściej wskazywali na  (46.9 %). Na drugi w rankingu RapidMiner wskazało 31,5% analityków, czyli o ponad 15% mniej.

Czemu  zawdzięcza tak dużą popularność? Niestety nie tym, że można się go szybko i łatwo nauczyć<sup>4</sup>. Dziwne maniere tego języka takie jak tablice indeksowane od 1 (a nie od 0) czy dziwny znak przypisania wartości do zmiennych długo daje się we znaki początkującemu programiście R<sup>5</sup>. Natomiast jest to język o wolnych źródłach, który każdy może pobrać ze strony projektu [r-project.org](http://r-project.org) i uruchomić na swoim ulubionym systemie operacyjnym (do wyboru Windows, OS X i Linux<sup>6</sup>). Dodatkowo  ma bardzo wiele pakietów, a więc praktycznie każda metoda analizy danych została już w nim zaimplementowana – wystarczy więc ściągnąć paczkę i używać bez znużonej implementacji. Powoduje to, że badacz danych nie jest ograniczony poprzez wąski zestaw metod zdefiniowany w umowie licencyjnej pakietu komercyjnego. Język ten staje się też powoli standardem w publikowaniu nowych osiągnięć w analizie danych – bardzo często do artykułów naukowych dołączany jest gotowy pakiet, aby każdy statystyk na świecie mógł zweryfikować opublikowane wyniki.

Z powyższych powodów większość zajęć laboratoryjnych z technik analizy danych będzie przeprowadzona w . Jednakże niektóre laboratoria, ze względu na możliwość lepszego ilustrowania przeprowadzanych przez studenta analiz, będą bazowały na arkuszach kalkulacyjnych. Studentów szczególnie zainteresowanym nauką języka  polecam stronę [datacamp.com](http://datacamp.com) gdzie poprzez oglądanie krótkich tutoriali oraz wykonywanie różnorodnych ćwiczeń można nauczyć się nie tylko programowania w , ale także zdobyć podstawowe doświadczenie w analizie i obróbce danych. Szczególnie polecam kurs „Data Analysis and Statistical Inference”<sup>7</sup>. Studenci podczas uczestnictwa w kursie ze „Statystyki i Analizy Danych” mają w pełni darmowy dostęp do wszystkich kursów na platformie DataCamp.

<sup>4</sup> „R will always be arcane to those who do not make a serious effort to learn it. It is \*\*\*\*not\*\*\*\* meant to be intuitive and easy for casual users to just plunge into. It is far too complex and powerful for that. But the rewards are great for serious data analysts who put in the effort.”, Berton Gunter

<sup>5</sup> Na portalu YouTube można nawet znaleźć filmiki zdecydowanych przeciwników tego języka na rzecz języka Python np. [youtu.be/4Iws2pv4kd8](https://youtu.be/4Iws2pv4kd8)

<sup>6</sup> Jest oczywiste, że tylko ten ostatni jest jedynym, prawdziwym i słusznym systemem operacyjnym.

<sup>7</sup> <https://www.datacamp.com/courses/statistical-inference-and-data-analysis>

### 1.3 Czym jest statystyka?

**Definicja 1.1 — Statystyka.** Statystyka (ang. *statistics*) to nauka zajmująca się zbieraniem, analizą i interpretacją danych.

Z jednej strony statystyka zakłada pewien brak idealności świata i dopuszcza istnienie szumu w danych (np. błędy pomiaru) oraz zjawisk losowych. Z drugiej zaś strony statystyka jest dziedziną matematyki, która jest nauką poruszającą się w idealnym (i abstrakcyjnym) świecie relacji, zbiorów i funkcji. To powiązanie z matematyką pozwala na formułowanie i udowadnianie pewnych twierdzeń oraz udzielania pewnych gwarancji osobom, które wykorzystują metody oferowane przez statystykę. Połączenie praktycznego podejścia wraz z ugruntowaną teorią matematyczną czyni ze statystyki potężne narzędzie wykorzystywane zarówno w naukach ścisłych jak i w naukach humanistycznych (np. psychologia, socjologia, a nawet literaturoznawstwo).

Tak powszechne użycie statystyki spowodowało, że wokół tej dyscypliny narosły, niestety, pewne mity. Popularne jest np. twierdzenie, że istnieją trzy rodzaje kłamstw: kłamstwa, okropne kłamstwa i statystyki. Powtarzane są one najczęściej przez osoby które statystyki nie znają, nie rozumieją i, co gorsza, czasami ją używają. Brak prawidłowego zrozumienia mechanizmów statystyki powoduje stosowanie metod nieodpowiednich do danego problemu (np. takiego w którym założenia metody nie są spełnione), pobieżną analizę na zasadzie „przecież to widać w danych” czy też wyciąganie błędnych wniosków z wyników analiz. Z tego powodu, z chwilą gdy poznajesz jakąś nową metodę statystyczną, oprócz dowiedzenia się jak to policzyć, niezwykle ważne jest znalezienie odpowiedzi na następujące pytania: „do jakich danych mogę tę metodę zastosować?”, „jaki są założenia tej metody?”, „co oznaczają wyniki uzyskane tą metodą?”. Warto powtórzyć ten sam zestaw pytań również przed każdym zastosowaniem dowolnej metody statystycznej.

#### 1.3.1 Populacja i próba

Z chwilą kiedy podejmujemy trud analizowania jakiś danych zwykle robimy to w celu znalezienia odpowiedzi na pewne pytanie badawcze. Na przykład, projektując nowe biurka do laboratoriów informatycznych, chcielibyśmy się dowiedzieć jaki jest średni wzrost studenta informatyki. Niestety, aby odpowiedzieć na tak ogólne pytanie musielibyśmy zmierzyć wszystkich studentów informatyki na świecie. Tego typu badanie byłoby straszliwie kosztowne, jeśli w ogóle wykonalne.

Problem ten można rozwiązać poprzez porzucenie pomysłu mierzenia wszystkich studentów informatyki i zdecydowaniu o wybraniu do badania jedynie kilku lub kilkunastu z nich, a otrzymaną w ten sposób średnią potraktować ze pewne przybliżenie. Nasuwają się od razu pytania: „ilu studentów powinniśmy zmierzyć?”, „w jaki sposób powinniśmy ich wybrać?” i w końcu „jak dokładna będzie uzyskana w ten sposób estymacja?”. Na te wszystkie pytania odpowiada statystyka.

Zbiór wszystkich elementów, których dotyczy pytanie badawcze nazywamy populacją – w naszym przykładzie są to wszyscy studenci informatyki na świecie. Z kolei zbiór elementów o których mamy dane (zwykle jest on dużo mniejszy od populacji) nazywamy próbą (studenci których wybraliśmy do zmierzenia). Statystyka nie zajmuje się odpowiadaniem na pytania badawcze mając do dyspozycji dane dotyczące całej populacji (jeśli masz dane dotyczące wszystkich studentów informatyki to po prostu policz średnią i zadanie wykonane) [8].



**Definicja 1.2 — Populacja.** Populacją (ang. *population*) nazywamy zbiór elementów podlegających badaniu lub analizie.

**Definicja 1.3 — Próba.** Próba (ang. *sample*) nazywamy podzbiór danych wybrany z populacji. Elementy próby nazywamy obserwacjami (ang. *observations*).

**Problem 1.1** Dla postawionych poniżej pytań badawczych wskaż populację.

1. Czy nowe lekarstwo na lekką krótkowzroczność rzeczywiście działa?
2. Czy jedzenie na śniadanie specjalnych płatek do mleka bez cukru powoduje spadek wagi u kobiet?
3. Ile procent studentów zdaje SiADy w pierwszym terminie?
4. Ile czasu zajmuje ukończenie studiów studentom Politechniki Poznańskiej w ostatnich 3 latach?

! Pomimo tego, że słowo „populacja” kojarzy nam się z np. populacją Polski (czyli po prostu liczbą ludności) to w statystyce populacja nie zawsze jest zbiorem ludzi. Na przykład jeżeli moim pytaniem badawczym będzie „Czy komputery z logiem jabłuszka są tanie?” to populacją będą tutaj wszystkie komputery z takim logiem.

### 1.3.2 Problem reprezentatywności próby

Trzeba zauważyć, że niezmiernie ważny jest też proces zbierania danych, które będziemy analizować. Już na tym etapie można popełnić kilka błędów. Na przykład pewna firma w Stanach Zjednoczonych chcąc obliczyć średnią wysokość Amerykanina pozyskała dane z bardzo popularnego portalu randkowego, gdzie użytkownicy uzupełniając dane o profilu również musieli podać swój wzrost. Dokonano starannego procesu czyszczenia danych usuwając dane użytkowników, którzy dla żartu wpisywali bardzo małe wartości (np. 16 cm) lub bardzo duże (np. 3256 cm). Pomimo tego przy porównaniu z wiarygodnymi danymi urzędu statystycznego okazało się, że rozkład wysokości ludzi, szczególnie wśród mężczyzn, był bardzo zawyżony. Nie jest to nic dziwnego, ponieważ powszechnie wiadomo (jak i potwierdziły to badania psychologiczne), że duży odsetek ludzi uzupełniając profile w tego typu portalach podają dane nieprawdziwe – korzystniejsze niż w rzeczywistości.

Ten przykład pokazuje, że zanim przystąpimy do analizy danych trzeba dobrze zrozumieć skąd one pochodzą i jak wygląda proces ich zbierania. Takie dane trzeba także oczyścić. Ponadto dowiedzieliśmy się, że niektóre próbki nie są tak samo dobre jak inne tj. wnioskowanie z nich o populacji jest obarczone znacznym błędem. Zwykle chcielibyśmy, aby próbka była reprezentatywna tj. częstość występowania badanej cechy w próbce nie powinna się znacząco różnić od jej częstotliwości w populacji. Co to w praktyce znaczy? Jeśli chcemy oszacować średnią wysokość człowieka to nie pobieramy próbki z koszykarzy grających w NBA, ale raczej pobieramy losowych ludzi w różnym wieku, różnej płci i z różnych krajów na świecie.

Aby ująć to trochę bardziej formalnie, większość metod statystycznych zakłada, że pracujemy z próbą prostą, która jest zdefiniowana poniżej.

**Definicja 1.4 — Próba prosta.** Próbką prostą  $n$ -wymiarową z rozkładu prawdopodobieństwa o dystrybuancie  $F$  nazywamy ciąg niezależnych zmiennych losowych

$X_1, X_2, \dots, X_n$  o jednakowym rozkładzie<sup>a</sup>,  $\forall_{i \in \{1, 2, \dots, n\}} P(X_i \leq x) = F(x)$ .

<sup>a</sup>W literaturze angielskojęzycznej często spotkasz skrót „i.i.d.” oznaczający „independent and identically distributed”

**Problem 1.2** Czy próba pobrana z populacji poprzez losowanie bez zwracania jest próbą prostą?

Kończąc dyskusję o reprezentatywności próby wymienimy kilka z powodów braku reprezentatywności prób:

- wygodne próbkowanie – bierzemy do próby przykłady, które są łatwo dostępne np. student AWFu bada ilu Polaków zdrowo się odżywia bazując na ankiecie przeprowadzanej wśród jego kolegów ze studiów.
- brak odpowiedzi – czasami respondent nie odpowiada na pytanie lub unika odpowiedzi. Powoduje to, że mamy dostępne dane o ludziach ze zdecydowanymi (często skrajnymi) poglądami na dane pytanie.
- odpowiedź wolontariuszy – jak wyżej, ankieta jest oparta na dobrowolnym wypełnieniu ankiety np. na stronie internetowej.

Najstławniejszym przykładem badania statystycznego w którym użyto niereprezentatywnej próby jest ankieta przeprowadzona w 1936 roku dla bardzo szanowanego magazynu „Literary Digest”, mająca przewidzieć wyniki wyborów prezydenckich w USA. Czasopismo wydało olbrzymią sumę pieniędzy przepytując ponad 2 400 000 wyborców i przewidziało zwycięstwo Alfreda Landon’a z 57% poparciem, podczas gdy wygrał Franklin D. Roosevelt z 62% poparciem<sup>8</sup>. Co było powodem tak dużego błędu? Ankieta została przeprowadzona drogą pocztową na 10 milionach ludzi z których odpowiedziało tylko 24%. Ponadto lista adresów została wzięta... z książki telefonicznej – posiadanie telefonu w 1936 roku nie było tak powszechne jak dzisiaj przez co ankieta trafiła tylko do bogatszej części społeczeństwa.

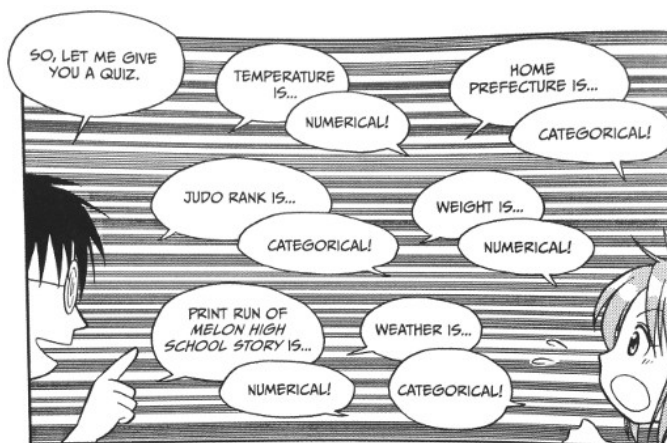
### 1.3.3 Obserwacja a eksperyment

Założmy, że chcemy przeprowadzić badanie, które wskaże czy jedzenie specjalnych płatków zbożowych na śniadanie sprzyja spadkowi wagi. W tym celu przeprowadziliśmy ankietę wśród konsumentów tych płatków i dowiedzieliśmy się, że ponad 90% nie ma żadnych problemów z nadwagą, a ich BMI jest w normie. Czy z tych danych możemy więc wnioskować o prawdziwości postawionej tezy?

Niestety, nie ponieważ nie mamy pewności czy nie istnieje jakaś inna zmienna która powoduje jednocześnie jedzenie płatków i spadek wagi. Takim czynnikiem może być „aktywność fizyczna”: ludzie którzy uprawiają sport interesują się zdrowymi produktami (a więc kupują nasze płatki) i jednocześnie w sposób oczywisty uprawianie sportu sprzyja spadkowi wagi. Jednak nasze dane nie uwzględniają tej zmiennej przez co wydaje nam się, że to jedzenie płatków wpływa na spadek wagi. Taką zmienną nazywamy zmienną zakłócającą (ang. *confounder*, *confounding variable*).

Co możemy zrobić, aby upewnić się że na wynik naszego badania nie będzie miała wpływu zmienna zakłócająca? Oczywiście, gdybyśmy ją znali wystarczyłoby włączyć ją do naszego badania i przeanalizować, ale co możemy zrobić w przypadku gdy jej nie znamy, albo na wynik badania ma wpływ jakiś niespodziewany czynnik?

<sup>8</sup>Więcej informacji znajdziesz na stronie: <https://www.math.upenn.edu/~deturck/m170/wk4/lecture/case1.html>



Rysunek 1.4: Rozpoznawanie różnych typów cech statystycznych.

Kluczem jest tutaj sposób pobierania danych. Dane w naszym przykładzie zostały zebrane poprzez obserwację, bez żadnego wpływu na otoczenie. Alternatywnym sposobem jest przeprowadzenie eksperymentu: zbierzmy losowo grupę ludzi, podzielmy ją (również losowo) na połowę i jednej z nich każmy jeść płatki, a drugiej nie. Po miesiącu mierzymy wagi wszystkim uczestnikom badania i wyciągamy wnioski. Zauważ, że nawet gdy istnieje jakaś zmienna zakłócająca, np. uprawianie sportu, to podczas losowego podziału w obu grupach powinno się znaleźć tyle samo reprezentantów ludzi uprawiających i nieuprawiających sport – co sprawia, że przy porównaniu tych dwóch grup zmienna ta nie będzie miała znaczenia (wpływie tak samo na wynik obu grup). Zasady projektowania eksperymentów możesz znaleźć w rozdziale 1.5.1 książki [6].

## 1.4 Typy cech statystycznych

Zanim zaczniemy pracę z danymi niezbędne jest zidentyfikowanie typu danych z którymi pracujemy. Typ danych określa nam operacje i metody które możemy na nich zastosować np. nie możemy pomnożyć koloru oczu. Najprostszy podział na typy zmiennych to podział na zmienne ilościowe (ang. *numerical*) i jakościowe (ang. *categorical*)<sup>9</sup>. Przykłady zmiennych obu typów możesz znaleźć na rysunku 1.4.

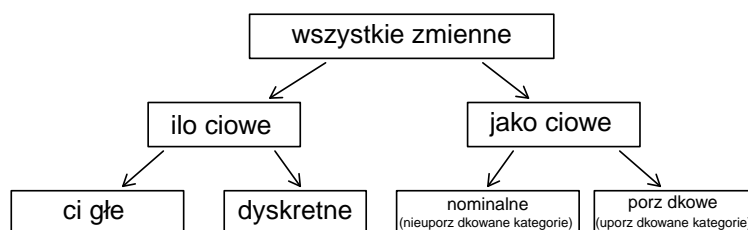
Dane ilościowe to takie które wyrażone są (zdziwienie) w postaci liczb i ma sens przeprowadzanie na nich operacji takich jak np. dodawanie, odejmowanie itd. Dodatkowo możemy uszczegółowić ich zaklasyfikowanie poprzez dodanie informacji o tym czy dane ilościowe są ciągłe (wysokość człowieka) czy dyskretne (liczba palców).

Pamiętaj, że nie wszystkie dane wyrażone w liczbach są danymi ilościowymi. Dobrym przykładem są tutaj numery na koszulkach reprezentacji Polski w piłce nożnej, ponieważ nie ma sensu ich dodawanie czy branie z nich średniej. Co więcej, nie zachodzi też relacja mniejszości/większości – czy zawodnik z numerem 1 jest gorszy od zawodnika z numerem 10? Tego typu zmienna jest przykładem zmiennej jakościowej.

Zmienne jakościowe są zwykle wyrażone w sposób opisowy. Do tej kategorii zaliczamy cechy takie jak kolor oczu, kraj urodzenia czy stan pogody („słonecznie”, „deszczowo” itd.). Jeżeli zbiór wartości zmiennej jakościowej jest uporządkowany (np. „niedosta-

<sup>9</sup>Czasami cechy ilościowe nazywamy mierzalnymi, a jakościowe niemierzalnymi

teczny”, „dostateczny”, „dobry” itd.) to zmienną tę nazywamy porządkową (ang. *ordinal*), w przeciwnym wypadku jest to zmienna nominalna (ang. *nominal*).



Rysunek 1.5: Typy cech statystycznych.

Innym bardzo często stosowanym podziałem jest podział zaproponowany przez Stanleya Stevensa na różne skale pomiarowe. Według tej kategoryzacji kolejne skale pomiarowe mają strukturę schodkową tj. każdy kolejny typ musi spełniać wszystkie wymagania swoich poprzedników.

**Skala nominalna** Zmienna ma jednoznacznie określony możliwy zbiór wartości i nic więcej – jedyną możliwą operacją jest porównywanie relacją równoważności (równa się/nie równa się). Przykłady: płeć, kolor oczu, typ komputera.



Rysunek 1.6: Podział cech wg. Stevensa [4].

**Skala porządkowa** Cecha wyrażona na skali nominalnej, która dodatkowo ma uporządkowany zbiór wartości, co powoduje że można porównywać wartości. Przykłady: poziom zadowolenia, wykształcenie („podstawowe”, „średnie”, „wyższe”). Zauważ, że możesz powiedzieć, że wykształcenie podstawowe jest niższe niż średnie, jednak nie możesz powiedzieć, że różnica pomiędzy wykształceniem podstawowym i średnim jest taka sama jak różnica pomiędzy wykształceniem średnim i wyższym.<sup>10</sup>

**Skala interwałowa** Cecha wyrażona na skali porządkowej, której różnice mają stały dystans. Przykłady: data urodzenia, temperatura w stopniach Celsjusza. Na przykład różnica pomiędzy 10°C a 15°C jest taka sama jak pomiędzy 155°C a 160°C. Zauważ, że obiekt mający 10°C nie jest wcale 2 razy cieplejszy od obiektu mającego 20°C (przelicz na Kelwiny!). W przypadku dat urodzenia dzielenie ich w ogóle nie ma sensu.

**Skala ilorazowa (proporcjonalna)** Cecha wyrażona na skali interwałowej której ilorazy mają interpretację. Na przykład możemy powiedzieć, że coś jest dwa razy cięższe (20 kg) niż coś innego (10 kg). Z tej własności wynikają kolejne: cecha ta nie przyjmuje wartości ujemnych oraz ma znaczące 0. Przykłady: masa, wysokość, temperatura w Kelwinach.

<sup>10</sup>Spójrz chociażby na długość kształcenia. Jeśli masz wykształcenie podstawowe to potrzebujesz 6 lat aby zdobyć wykształcenie średnie (gimnazjum+liceum), natomiast aby zdobyć wykształcenie wyższe mając średnie potrzeba 3(,5) lat kształcenia.

## 1.5 Szeregi rozdzielcze

W praktyce tabele z danymi są bardzo duże i nie jesteśmy w stanie przejrzeć całej tabelki, co znacznie utrudnia analizę. Rozwiązaniem tego problemu jest utworzenie z danych szeregu rozdzielczego (ang. *frequency table*).

Szereg rozdzielczy to ujęcie tabelaryczne danych w których poszczególne wartości zmiennej łączy się w przedziały (czasami używany jest też termin klasy), a następnie podaje się liczbę wartości należących do każdego z przedziałów. Jest to więc zwykle tabela posiadająca 2 kolumny: przedział wartości i liczbę wystąpień. W każdym wierszu takiej tabeli podajemy zakres wartości, a następnie liczbę wystąpień wartości zmiennej należących do danego zakresu. Zakresy wartości w szeregu są przedziałami, które nie nakładają się na siebie, tworząc jednocześnie jeden, ciągły przedział.

■ **Przykład 1.1** Studenci pewnej słynnej Politechniki, aby zaliczyć „Podstawy Programowania” musieli przygotować krótkie programy na zaliczenie. Po otrzymaniu kodów źródłowych, prowadzący postanowił sprawdzić czy przygotowane przez niego zadanie nie było zbyt skomplikowane. Aby to zrobić przeanalizował długość każdego z programów (liczbę linii kodu). Zebrane przez niego dane wyglądają następująco: 1123, 198, 473, 784, 305, 423, 397, 298, 698, 237.

Skonstruowany szereg rozdzielczy dla tych danych wygląda następująco:

przedział	liczność
(197,506]	7
(506,815]	2
(815,1124]	1

■

**Problem 1.3** Czy szereg rozdzielczy przedziałowy można zastosować do wszystkich typów danych?

Aby skonstruować szereg rozdzielczy dla pewnej próbki danych należy:

1. Obliczyć rozstęp badanej cechy/atributu.  $R = x_{\max} - x_{\min}$
2. Zdecydować o liczbie przedziałów  $k$ . Jest to bardzo ważny krok całego procesu, a jednocześnie nie ma jednoznacznej odpowiedzi jak powinno się to zrobić. Kilka popularnych heurystyk to:  $k = \sqrt{n}$ ,  $k \leq 5 \ln n$ ,  $k = 1 + 3.322 \ln n$ ,  $k = \frac{x_{\max} - x_{\min}}{2.64n^{-1/3}IQR}$  gdzie  $n$  to liczność badanej próbki, a  $IQR$  to rozstęp międzykwartyłowy<sup>11</sup>.
3. Obliczyć szerokość przedziału  $h = \frac{R}{k}$ , wartość tę można zaokrąglić w górę. Szczególnie powinno się to zrobić, aby dostosować szerokość przedziału do pewnych intuicji stojącymi za danymi np. jeżeli dane przedstawiają liczbę zjedzonych chipsów (liczba całkowita) przyjmowanie szerokości przedziału 6.7354 nie ma sensu, powinno przyjąć się wartość 7.
4. Zdefiniować przedziały szeregu  $(x'_{\min}, x'_{\min} + h]$ ,  $(x'_{\min} + h, x'_{\min} + 2h]$ ,  $(x'_{\min} + 2h, x'_{\min} + 3h]$  itd. Przedziały w szeregu zwykle są lewostronnie otwarte i prawostronnie zamknięte (choć to zależy od zastosowania, np. w epidemiologii popularniejsze są przedziały prawostronnie otwarte).
5. Zliczyć częstość występowania każdego z przedziałów i podsumować w tabelce.

Pewnie się zastanawiasz jaką wartość podstawić pod zmienną  $x'_{\min}$  i niestety nie ma prostej odpowiedzi na to pytanie. Po pierwsze gdyby przyjąć  $x'_{\min} = x_{\min}$  to do

<sup>11</sup>Zostanie on zdefiniowany na następnych zajęciach



pierwszego przedziału nie będzie należeć wartość  $x_{min}$ , gdyż przedział ten jest lewostronnie otwarty. Wartość ta z oczywistych względów nie będzie też należeć do żadnego innego przedziału – jest to więc sytuacja błędna. Wynika z tego, że  $x'_{min}$  powinno być mniejsze od  $x_{min}$ . Alternatywnym rozwiązaniem tego problemu byłoby oczywiście zdefiniowanie pierwszego przedziału jako przedziału obustronnie zamkniętego, jest to jednak rozwiązanie kontrowersyjne ponieważ łamiemy wtedy założenie o równej szerokości przedziałów. Inne rozwiązanie, również łamiące to założenie, to rozszerzenie wartości skrajnych przedziałów o pewną stałą wartość wynoszącą  $0,1\%h$ .

Kolejnym rozwiązaniem jest też obliczenie szerokości przedziału  $h$  dla  $k - 1$  i przyjęcie wtedy  $x'_{min} = x_{min} - \frac{1}{2}h$ , jednak jest wtedy oczywiste że pierwszy i ostatni przedział będzie tylko w połowie wykorzystany. Ten sposób konstrukcji szeregu również nie wydaje się optymalny.

Innym, i chyba najczęściej stosowanym ominięciem tego problemu, jest skorzystanie z możliwości zaokrąglenia szerokości przedziału  $h$  w górę (szczególnie, że najczęściej i tak to robimy dla zwiększenia intuicyjności przedziałów), a następnie ustalenia wartości  $x'_{min}$  na jakąś równie intuicyjną i „okrągłą” wartość.

W książce [8] autorzy proponują wykorzystanie wiedzy o dokładności raportowanych liczb  $\alpha$  poprzez odjęcie od  $x_{min}$  połowy tej wielkości. Dokładność raportowania danych używana jest też do zaokrąglenia w górę (!) szerokości przedziału. W naszym przykładzie z liczbą zjedzonych chipsów raportowane są tylko wartości całkowite, a więc dokładność wynosi  $\alpha = 1$ . W tej sytuacji ustalilibyśmy początek pierwszego przedziału na  $x'_{min} = x_{min} - 0.5$ . Takie ustalenie przedziałów ma dodatkowy plus: nie jest ważne czy przedziały są prawo- czy lewostronnie zamknięte, ponieważ szeregi dla obu rodzajów przedziałów wyglądają dokładnie tak samo. Wynika to z faktu, że żadna raportowana liczba nie będzie się znajdowała na granicy przedziału (skoro raportowane liczby są całkowite, żadna z nich nie będzie wynosiła np. 5,5).

Jak widzisz jest wiele sposobów na konstrukcję szeregu rozdzielczego, warto poeksperymentować z różnymi pomysłami i sprawdzić który dla Ciebie najlepiej się sprawdza.

■ **Przykład 1.2** W przykładzie 1.1 pokazaliśmy szereg rozdzielczy dla liczby linii kodu w programach studentów. Spróbujmy go teraz skonstruować krok po kroku. Przypomnijmy dane wyglądają następująco: 1123, 198, 473, 784, 305, 423, 397, 298, 698, 237.

1. Obliczmy rozstęp danych w naszej próbie. Znajdźmy najmniejszą i największą wartość:  $x_{min} = 198$  oraz  $x_{max} = 1123$ , a następnie podstawmy do wzoru  $R = x_{max} - x_{min} = 1123 - 198 = 925$ .
2. Zdecydujmy o liczbie klas  $k$  korzystając z heurystyki  $k = \sqrt{n} = \sqrt{10} \approx 3$ .
3. Wyznaczmy szerokość przedziału:  $h = \frac{R}{k} = \frac{925}{3} = 308\frac{1}{3}$ . Szerokość przedziału zaokrąglimy (zawsze w górę!) do  $h \approx 309$ .
4. Zaczniemy konstruować nasz szereg od  $x'_{min} = 197$  (jest to wartość mniejsza od  $x_{min} = 198$  i zarazem całkowita – ma więc sens jej zastosowanie w tym przypadku). Do definicji pierwszego przedziału  $(x'_{min}, x'_{min} + h]$  podstawiamy obliczone wartości  $(197, 197 + 309] = (197, 506]$ . Tak samo robimy dla pozostałych przedziałów otrzymując  $(x'_{min} + h, x'_{min} + 2h] = (506, 815]$  i  $(x'_{min} + 2h, x'_{min} + 3h] = (815, 1124]$ .
5. Sporządźmy tabelkę wypełniając pierwszą kolumnę nazwami przedziałów. Następnie możemy dodać do tabli dodatkową kolumnę w której będziemy stawiać kreski dla zliczenia każdej wartości należącej do przedziału (metoda kreskowa). Na koniec wystarczy policzyć finalną liczbę kresek w wierszu i wpisać jako ostateczną

liczebność przedziału. Ostateczny rezultat powinien wyglądać następująco:

przedział	liczność	miejsce na kreski
(197,506]	7	IIIII III
(506,815]	2	II
(815,1124]	1	I

Zauważ, że kreski stworzyły coś na kształt wykresu obrazującego nam gęstość danych w poszczególnych przedziałach. Wykres taki nazywamy histogramem i będziemy jeszcze o nim mówić na kolejnych laboratoriach.

! Przy konstruowaniu szeregu rozdzielczego zwróć uwagę na domknięcia przedziałów, gdyż jest to częste miejsce popełniania błędów. Zawsze sprawdź czy pomiędzy kolejnymi przedziałami nie ma „dziur”, a także czy najmniejsza i największa wartość należy odpowiednio do pierwszego i ostatniego przedziału.


**Ćwiczenie 1.1 — Konstrukcja szeregu rozdzielczego.** Otwórz arkusz kalkulacyjny dostępny pod następującym linkiem: <https://ophelia.cs.put.poznan.pl/webdav/ad/students/01/cw-1.xls>. Znajdziesz tam dane dotyczące wzrostu informatyków – wyniki należy pogrupować i przedstawić w postaci szeregu rozdzielczego.

**Ćwiczenie 1.2 — Funkcje tablicowe.** Otwórz arkusz kalkulacyjny dostępny pod następującym linkiem: <https://ophelia.cs.put.poznan.pl/webdav/ad/students/01/cw-2.xls>. To jest to samo ćwiczenie, które rozwiązywałeś wcześniej, ale rozwiąż je za pomocą funkcji tablicowych<sup>a</sup>

<sup>a</sup><http://pszyperski.republika.pl/Excel%202003/Funkcje%20Tablicowe.htm>

**Ćwiczenie 1.3 — Badanie liczby przedziałów w szeregu rozdzielczym.** Otwórz arkusz kalkulacyjny dostępny pod następującym linkiem: <https://ophelia.cs.put.poznan.pl/webdav/ad/students/01/cw-5.xls>. Dla danych w arkuszu poszukaj odpowiedniej liczby przedziałów dla szeregu rozdzielczego.

! Omawiane przez nas szeregi rozdzielcze to tak naprawdę szeregi rozdzielcze przedziałowe. Istnieją także szeregi rozdzielcze punktowe, w których nie tworzy się przedziałów tylko po prostu zlicza się wystąpienie każdej z wartości zmiennych. Z tego powodu nie są one stosowane dla danych ilościowych ciągłych, które mogą mieć bardzo duże zbiory wartości (potencjalnie każda wartość w próbie może być inna).

 Szeregi rozdzielcze można też łatwo skonstruować w pakiecie R. W tym celu najpierw wprowadźmy do systemu nasze dane. Możemy to zrobić np. poprzez ręczne utworzenie wektora z danymi wydając polecenie:

```
dane <- c(1123, 198, 473, 784, 305, 423, 397, 298, 698, 237)
```

Operator `<-` oznacza przypisanie<sup>a</sup>, a funkcja `c()` jest konstruktorem wektora. Wektor w R nie koniecznie musi zawierać liczb – może to być dowolny typ danych np. ciąg znaków, jednakże dane w całym wektorze zawsze muszą być jednakowego typu.

Aby utworzyć szereg rozdzielczy musimy podzielić nasze dane na przedziały. Możemy zrobić to za pomocą funkcji `cut(dane, breaks=liczba_przedziałów)` w której musimy ręcznie wyspecyfikować liczbę przedziałów poprzez podanie argumentu `breaks=`. Załóżmy, że dla naszego krótkiego wektora danych wystarczą 3 przedziały i wywołajmy polecenie:

```
dane_w_przedzialach <- cut(dane, breaks = 3)
```

Utworzona właśnie zmienna `dane_w_przedzialach` nie jest już wektorem, ale zmienna typu `factor`. Stało się tak dlatego, że funkcja `cut()` po określeniu zakresu przedziałów automatycznie przekonwertowała wszystkie wartości wektora `dane` na nazwy odpowiadających im przedziałów. Dane zmieniły więc swój typ i są teraz typu jakościowego. Do przechowywania tego typu danych wykorzystywany jest właśnie typ `factor`<sup>b</sup> (choć oczywiście, pomijając względy efektywności, mogłyby być one przechowane jako wektor ciągów znaków). Nie wierz temu co jest tu napisane na słowo – sprawdź jak wygląda zawartość zmiennej `dane_w_przedzialach` przez wpisanie jej nazwy do konsoli i wciśnięcia Enter.


Mając tak przygotowane dane aby dokończyć ćwiczenie wystarczy wyświetlić tabelkę ze zliczeniem wartości poszczególnych przedziałów - możemy to uzyskać funkcją `table()`.

```
table(dane_w_przedzialach)
```


Gratulacje! W ten sposób utworzyłeś swój pierwszy szereg rozdzielczy w R. Przypomnijmy: wymagało to obliczenia zakresu przedziałów i zastąpienia wartości liczbowych wektora odpowiadającymi im przedziałami (funkcja `cut()`), a następnie zliczenia nazw poszczególnych przedziałów i wyświetlenia tego w formie tabeli (funkcja `table()`). Przy okazji dowiedzieliśmy się jak utworzyć wektor z danymi oraz poznaliśmy typ `factor`.

<sup>a</sup>Zwróć uwagę, że postawienie spacji pomiędzy znakami `<-` zmienia operator przypisania na dwa operatory „mniejszy niż” oraz „minus”. Warto też wiedzieć, że w niektórych sytuacjach jest dozwolone korzystanie z „normalnego” operatora przypisania czyli znaku równości.

<sup>b</sup>Dla zainteresowanych: wewnętrznie `factor` jest reprezentowany jako wektor kolejnych liczb całkowitych (identyfikatorów) wraz ze słownikiem mapującym te identyfikatory na nazwy. Słownik ten można odczytać wydając polecenie `levels(dane_w_przedzialach)`. Pewną magią języka R jest to, że do tej funkcji możemy także przypisać wartości np. `levels(dane_w_przedzialach)<-c('a','b','c')` spowoduje zasępienie nazw przedziałów kolejnymi literami alfabetu.

**Ćwiczenie 1.4** Korzystając z dostępnej w pakiecie  pomocy (wpisz znak zapytania razem z nazwą komendy np. `?cut`) stwórz szereg rozdzielczy z przedziałami

prawostronnie otwartymi. ■

**Ćwiczenie 1.5** Typ wektorowy pozwala na przechowywanie danych tylko jednego typu. Co się stanie jeśli spróbujesz utworzyć wektor zawierający np. liczby i ciągi znaków? Sprawdź swoją hipotezę w . ■

## Literatura

### Literatura powtórkowa

Wprowadzenie do statystyki można znaleźć w rozdziałach 1.1—1.5 darmowej książki „OpenIntro Statistics” [6]. Książka jest dostępna do ściągnięcia na stronie [www.openintro.org/stat/textbook.php](http://www.openintro.org/stat/textbook.php). Natomiast informacje o szeregach rozdzielczych można znaleźć np. w rozdziale 1.2 książki [8].

### Literatura dla chętnych

Znanym przykładem na to jak bardzo ważne jest blokowanie innych, nieznanymi zmiennymi jest paradoks Simpsona. Pokazuje on jak nie wzięcie pod uwagę jednej ze zmiennych całkowicie zmienia wynik. Z tego powodu w tym tygodniu zachęcam do zapoznania się ze stroną <http://vudlab.com/simpsons/> na której znajdziecie wytłumaczenie tego paradoksu wraz z wizualizacją.

## Pytania sprawdzające zrozumienie

**Pytanie 1.2** Wyjaśnij różnicę między populacją i próbą. Jakie założenia spełnia próba prosta? Czy próba prosta jest zawsze reprezentatywna?

**Pytanie 1.3** Wyjaśnij różnice pomiędzy różnymi typami danych. Podaj typ danych dla cechy statystycznej.

**Pytanie 1.4** Skonstruuj szereg rozdzielczy dla podanych danych.

## Bibliografia

- [1] What is a data scientist? <http://bigdata-madesimple.com/what-is-a-data-scientist-14-definitions-of-a-data-scientist/>. Dostęp: 2016-02-10.
- [2] IBM. Bringing big data to the enterprise. <http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>. Dostęp: 2016-02-10.
- [3] Indeed.com: salary of data scientist. <http://www.indeed.com/salary?q1=%22Data+Scientist%22>. Dostęp: 2016-02-10.
- [4] Podział cech wg Stanleya Stevensa – czy płęć można mnożyć i dlaczego? <https://www.statystyczny.pl/podzial-cech-wg-stanleya-stevensa-czy-plec-mozna-mnozyc-i-dlaczego/>, 2016. Dostęp: 2016-02-10.

- [5] T.H. Davenport i D.J. Patil. Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, 2012. URL <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/>. Dostęp: 2016-02-10.
- [6] D.M. Diez, C.D. Barr, i M. Çetinkaya Rundel. *OpenIntro Statistics: Third Edition*. OpenIntro, Inc., 2015. ISBN 194345003X. URL [openintro.org](http://openintro.org).
- [7] T. Hey, S. Tansley, i K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009. URL [http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th\\_paradigm\\_book\\_complete\\_lr.pdf](http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf).
- [8] W. Krysicki, J. Bartos, W. Dyczka, K. Królikowska, i M. Wasilewski. *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach. Część II: Statystyka matematyczna*. Wydawnictwo Naukowe PWN, 2002. ISBN 8301113847.
- [9] C. O’Neil i R. Schutt. *Badanie danych. Raport z pierwszej linii działań*. Helion, 2014. ISBN 8324696261.
- [10] M. Walker. The professionalization of data science. <http://www.datasciencecentral.com/profiles/blogs/the-professionalization-of-data-science>, 2013. Dostęp: 2016-02-10.
- [11] K. Zawadzki. Data science skill-set explained. <http://www.marketingdistillery.com/2014/08/30/data-science-skill-set-explained/>, 2014. Dostęp: 2016-02-10.