Statystyki podstawowe

W tej części skupimy się na statystykach wymaganych przez każdego początkującego naukowca danych. Zbadamy sposoby próbkowania i uzyskiwania danych bez wpływu na uprzedzenia, a następnie użyjemy miar statystycznych do kwantyfikacji i wizualizacji naszych danych. Wykorzystując z-score i regułę empiryczną, zobaczymy, jak możemy standaryzować dane na potrzeby tworzenia wykresów i interpretacji. Przyjrzymy się następującym tematom:

• Jak pozyskiwać i próbkować dane?

• Miary środka, wariancji i względnej pozycji

• Normalizacja danych za pomocą z-score

• Reguła empiryczna

Czym są statystyki?

To może wydawać się dziwne pytanie, ale często dziwi mnie liczba osób, które nie potrafią odpowiedzieć na to proste, a jednocześnie mocne pytanie: czym są statystyki? Statystyki to liczby, które zawsze widzisz w wiadomościach i w gazecie. Statystyki są przydatne, gdy próbujesz udowodnić swoją rację lub próbujesz cię przestraszyć, ale czym one są? Aby odpowiedzieć na to pytanie, musimy zrobić kopię zapasową na minutę i porozmawiać o tym, dlaczego w ogóle je mierzymy. Celem tej dziedziny jest próba wyjaśnienia i modelowania otaczającego nas świata. Aby to zrobić, musimy przyjrzeć się populacji. Możemy zdefiniować populację jako całą pulę podmiotów eksperymentu lub modelu. Zasadniczo zależy ci na twojej populacji. O kim próbujesz porozmawiać? Jeśli próbujesz sprawdzić, czy palenie prowadzi do chorób serca, twoja populacja byłaby palaczami na całym świecie. Jeśli próbujesz badać problemy nastolatków związanych z piciem, twoją populacją będą wszyscy nastolatkowie. Teraz pomyśl, że chcesz zadać pytanie dotyczące twojej populacji, na przykład, jeśli twoja populacja to wszyscy twoi pracownicy (załóżmy, że masz ponad 1000 pracowników), być może chcesz wiedzieć, jaki procent z nich używa nielegalnych narkotyków. Pytanie nazywa się parametrem. Parametr możemy zdefiniować jako pomiar liczbowy opisujący charakterystykę populacji. Na przykład, jeśli zapytasz wszystkich 1000 pracowników i 100 z nich zażywa narkotyki, wskaźnik zażywania narkotyków wynosi 10%. Parametr tutaj wynosi 10%. Jednak bądźmy szczerzy, prawdopodobnie nie można zapytać każdego pracownika, czy zażywa narkotyki. A jeśli masz ponad 10 000 pracowników? Byłoby bardzo trudno wyśledzić wszystkich, aby uzyskać odpowiedź. Kiedy tak się dzieje, niemożliwe jest ustalenie tego parametru. W takim przypadku możemy oszacować parametr. Najpierw weźmiemy próbkę populacji. Możemy zdefiniować próbkę populacji jako podzbiór (losowy niewymagany) populacji. Więc być może zapytamy 200 z 1000 pracowników, których masz. Załóżmy, że z tych 200 osób 26 używa narkotyków, co oznacza, że wskaźnik zażywania narkotyków wynosi 13%. Tutaj 13% nie jest parametrem, ponieważ nie mieliśmy okazji zapytać wszystkich. To 13% to oszacowanie parametru. Czy wiesz, jak to się nazywa? Zgadza się, statystyka! Możemy zdefiniować statystykę jako pomiar liczbowy opisujący charakterystykę próbki populacji. Statystyka to tylko oszacowanie parametru. Jest to liczba, która próbuje opisać całą populację, opisując jej podzbiór. Jest to konieczne, ponieważ nigdy nie możesz mieć nadziei na przeprowadzenie ankiety każdemu nastolatkowi lub każdemu palaczowi na świecie. O to właśnie chodzi w dziedzinie statystyki - pobieranie próbek populacji i przeprowadzanie testów na tych próbkach. Zatem następnym razem, gdy otrzymasz statystykę, pamiętaj tylko, że ta liczba reprezentuje tylko próbkę tej populacji, a nie całą pulę badanych.

Jak pozyskujemy i przykładamy dane?

Jeśli statystyki dotyczą pobierania próbek populacji, bardzo ważne jest, aby wiedzieć, w jaki sposób uzyskujemy te próbki, i masz rację. Skoncentrujmy się tylko na kilku z wielu sposobów pozyskiwania i próbkowania danych.

Uzyskiwanie danych

Istnieją dwa główne sposoby zbierania danych do naszej analizy: obserwacyjne i eksperymentalne. Oba te sposoby mają oczywiście swoje plusy i minusy. Każdy z nich wytwarza różne rodzaje zachowań, a zatem uzasadnia różne rodzaje analizy.

Obserwacje

Możemy pozyskiwać dane za pomocą środków obserwacyjnych, które polegają na mierzeniu określonych cech, ale nie próbowaniu modyfikowania badanych osób. Na przykład masz na swojej stronie oprogramowanie śledzące, które obserwuje zachowanie użytkowników w witrynie, takie jak czas spędzony na określonych stronach i szybkość klikania reklam, a jednocześnie nie wpływa na wrażenia użytkownika, wtedy byłoby to badanie obserwacyjne. Jest to jeden z najczęstszych sposobów pozyskiwania danych, ponieważ jest to po prostu łatwe. Wszystko, co musisz zrobić, to obserwować i zbierać dane. Badania obserwacyjne są również ograniczone pod względem rodzajów danych, które możesz gromadzić. Dzieje się tak, ponieważ obserwator (ty) nie kontroluje środowiska. Możesz jedynie obserwować i zbierać naturalne zachowania. Jeśli chcesz wywołać określony rodzaj zachowania, badanie obserwacyjne nie będzie przydatne.

Eksperymentalnie

Eksperyment polega na leczeniu i obserwacji jego wpływu na badanych. Osoby biorące udział w eksperymencie nazywane są jednostkami eksperymentalnymi. W ten sposób zwykle zbiera dane większość laboratoriów naukowych. Umieszczą ludzi w dwóch lub więcej grupach (zwykle tylko dwóch) i nazywają ich grupą kontrolną i eksperymentalną. Grupa kontrolna jest wystawiona na określone środowisko, a następnie obserwowana. Grupa eksperymentalna jest następnie wystawiana na działanie innego środowiska, a następnie obserwowana. Eksperymentator następnie agreguje dane z obu grup i podejmuje decyzję o tym, które środowisko było korzystniejsze (korzystne to jakość, o której decyduje eksperymentator).

Weźmy pod uwagę przykład marketingowy, że pokazujemy połowę naszych użytkowników na określonej stronie docelowej z określonymi obrazami i określonym stylem (witryna A) i mierzymy, czy rejestrują się w usłudze. Następnie drugą połowę udostępniamy innej stronie docelowej, innym obrazom i różnym stylom (witryna B) i ponownie mierzymy, czy się zarejestrowali. Możemy wtedy zdecydować, która z dwóch stron wypadła lepiej i powinna być wykorzystywana dalej. Nazywa się to w szczególności testem A/B. Zobaczmy przykład w Pythonie! Załóżmy, że uruchamiamy poprzedni test i otrzymujemy następujące wyniki jako listę list:

results = [['A', 1], ['B', 1], ['A', 0], ['A', 0] …]

Tutaj każdy obiekt w wyniku listy reprezentuje podmiot (osobę). Każda osoba ma wtedy następujące dwa atrybuty:

• Na jakiej stronie internetowej byli obecni, reprezentowanej przez pojedynczy znak

• Czy dokonali konwersji (0 za nie i 1 za tak)

Następnie możemy dokonać agregacji i uzyskać następującą tabelę wyników:

users_exposed_to_A = []

users_exposed_to_B = []

utwórz dwie listy do przechowywania wyników każdej indywidualnej witryny

Po utworzeniu tych dwóch list, które ostatecznie będą zawierały wartość logiczną każdej indywidualnej konwersji (0 lub 1), powtórzymy wszystkie nasze wyniki testu i dodamy je do odpowiedniej listy, jak pokazano:

for website, converted in results: # iterate through the results

will look something like website == 'A' and converted == 0

if website == 'A':

users_exposed_to_A.append(converted)

elif website == 'B':

users_exposed_to_B.append(converted)

Teraz każda lista zawiera serię jedynek i zer. Pamiętaj, że 1 oznacza użytkownika, który faktycznie przechodzi na witrynę po obejrzeniu tej strony internetowej, a 0 oznacza, że użytkownik widzi tę stronę i opuszcza ją przed zarejestrowaniem się/konwersją. Aby uzyskać całkowitą liczbę osób narażonych na witrynę A, możemy użyć funkcji len() w Pythonie, jak pokazano:

len(users_exposed_to_A) == 188 #number of people exposed to website A

len(users_exposed_to_B) == 158 #number of people exposed to website B

Aby policzyć liczbę osób, które dokonały konwersji, możemy użyć sum() z listy, jak pokazano:

sum(users_exposed_to_A) == 54 # people converted from website A

sum(users_exposed_to_B) == 48 # people converted from website B

Jeśli odejmiemy długość list i sumę listy, otrzymamy liczbę osób, które nie dokonały konwersji dla każdej witryny, jak pokazano

len(users_exposed_to_A) - sum(users_exposed_to_A) == 134 # did not convert from website A

len(users_exposed_to_B) - sum(users_exposed_to_B) == 110 # did not convert from website B

Możemy zebrać i podsumować nasze wyniki w poniższej tabeli, która przedstawia nasz eksperyment testowania konwersji w witrynie:

	Did not sign up	Signed up
Website A	134	54
Website B	110	48

Możemy szybko zebrać statystyki opisowe. Można powiedzieć, że współczynniki konwersji witryny dla obu witryn są następujące:

• Konwersja dla witryny A: 54/134 + 54 = 0,288

• Konwersja dla witryny B: 48 / 110 + 48 = 0,3

Niewielka różnica, ale jednak inna. Mimo że B ma wyższy współczynnik konwersji, czy naprawdę możemy powiedzieć, że wersja B znacznie lepiej konwertuje? Jeszcze nie. Aby sprawdzić istotność statystyczną takiego wyniku, należy zastosować test hipotezy. Testy te zostaną szczegółowo omówione w następnym rozdziale, w którym ponownie przyjrzymy się dokładnie temu samemu przykładowi i zakończymy go odpowiednim testem statystycznym.

Próbkowanie danych

Pamiętaj, że statystyki są wynikiem pomiaru próbki populacji. Cóż, powinniśmy porozmawiać o dwóch bardzo powszechnych sposobach decydowania, kto dostąpi zaszczytu bycia w mierzonej przez nas próbce. Omówimy główny rodzaj doboru próby, zwany doborem losowym, który jest najczęstszym sposobem decydowania o wielkości naszej próby i naszych członków próby.

Próbkowanie prawdopodobieństwa

Próbkowanie prawdopodobieństwa to sposób próbkowania z populacji, w którym każda osoba ma znane prawdopodobieństwo wybrania, ale ta liczba może mieć inne prawdopodobieństwo niż inny użytkownik. Najprostszą (i prawdopodobnie najczęstszą) metodą próbkowania prawdopodobieństwa jest próbkowanie losowe.

Próbkowanie losowe

Załóżmy, że przeprowadzamy test A/B i musimy dowiedzieć się, kto będzie w grupie A, a kto w grupie B. Zespół ds. danych przedstawia następujące trzy sugestie:

• Oddzielni użytkownicy na podstawie lokalizacji: użytkownicy na zachodnim wybrzeżu są umieszczani w grupie A, podczas gdy użytkownicy na wschodnim wybrzeżu są umieszczani w grupie B

• Oddziel użytkowników na podstawie pory dnia, kiedy odwiedzają witrynę: Użytkownicy, którzy odwiedzają witrynę między 19:00. i o 4 rano dostaniesz miejsce A, podczas gdy reszta zostanie umieszczona w grupie B

• Zrób to całkowicie losowo: każdy nowy użytkownik ma 50/50 szans na umieszczenie w jednej z grup

Pierwsze dwie są prawidłowymi opcjami wyboru próbek i są dość proste do wdrożenia, ale obie mają jedną podstawową wadę: oba są narażone na wprowadzenie błędu próbkowania.

Błąd w doborze próby występuje, gdy sposób, w jaki pozyskiwana jest próbka, faworyzuje pewien wynik w porównaniu z wynikiem docelowym.

Nietrudno zrozumieć, dlaczego wybór opcji 1 lub opcji 2 może wprowadzać stronniczość. Jeśli wybieramy nasze grupy na podstawie tego, gdzie mieszkają lub o której godzinie się logują, to niepoprawnie przygotowujemy nasz eksperyment i teraz mamy znacznie mniejszą kontrolę nad wynikami. W szczególności istnieje ryzyko wprowadzenia do naszej analizy czynnika mylącego, co jest złą wiadomością.

Czynnikiem mylącym jest zmienna, której nie mierzymy bezpośrednio, ale łączymy zmienne, które są mierzone.

Zasadniczo czynnik mylący jest jak brakujący element w naszej analizie, który jest niewidoczny, ale wpływa na nasze wyniki. W tym przypadku wariant 1 nie uwzględnia potencjalnego mylącego czynnika smaku geograficznego. Na przykład, jeśli witryna A jest ogólnie nieatrakcyjna dla użytkowników z zachodniego wybrzeża, drastycznie wpłynie to na wyniki.

Podobnie opcja 2 może wprowadzić czasowy (oparty na czasie) czynnik zakłócający. Co jeśli witryna B jest lepiej oglądana w nocy (co było zarezerwowane dla A), a użytkownicy są wyłączani z tego stylu wyłącznie z powodu godziny. Są to oba czynniki, których chcemy uniknąć, dlatego powinniśmy wybrać opcję 3, która jest losowa próbka.

Chociaż stronniczość próbkowania może wprowadzać w błąd, jest to inne pojęcie niż mylące. Warianty 1 i 2 były obiema stronniczymi próbkami, ponieważ wybraliśmy próbki nieprawidłowo, a także były przykładami czynników mylących, ponieważ w każdym przypadku istniała trzecia zmienna, która wpływała na naszą decyzję.

Próba losowa jest wybierana w taki sposób, aby każdy członek populacji miał taką samą szansę na wybór jak każdy inny członek. Jest to prawdopodobnie jeden z najłatwiejszych i najwygodniejszych sposobów decydowania, kto będzie częścią twojej próbki. Każdy ma taką samą szansę na bycie w określonej grupie. Próbkowanie losowe to skuteczny sposób na zmniejszenie wpływu czynników zakłócających.

Próbkowanie z nierównym prawdopodobieństwem

Przypomnijmy, że wcześniej powiedziałem, że próbkowanie prawdopodobieństwa może mieć różne prawdopodobieństwa dla różnych potencjalnych członków próby. Ale co, jeśli faktycznie spowodowało to problemy? Załóżmy, że jesteśmy zainteresowani pomiarem poziomu szczęścia naszych pracowników. Wiemy już, że nie możemy zapytać wszystkich osób z personelu, ponieważ byłoby to głupie i wyczerpujące. Więc musimy pobrać próbkę. Nasz zespół danych sugeruje losowe pobieranie próbek i na początku każdemu przybija piątkę, ponieważ czują się bardzo inteligentni i statystycznie. Ale wtedy ktoś zadaje pozornie nieszkodliwe pytanie - czy ktoś zna procent mężczyzn/kobiet, którzy tu pracują? Przybijanie piątki ustaje i w pokoju zapada cisza. To pytanie jest niezwykle ważne, ponieważ seks może być czynnikiem mylącym. Zespół przygląda się temu i odkrywa, że w firmie jest 75% mężczyzn i 25% kobiet. Oznacza to, że jeśli wprowadzimy próbę losową, nasza próba prawdopodobnie będzie miała podobny podział, a tym samym będzie faworyzować wyniki dla mężczyzn, a nie kobiet. Aby temu zaradzić, możemy faworyzować uwzględnienie większej liczby kobiet niż mężczyzn w naszym badaniu, aby podział naszej próby był mniej korzystny dla mężczyzn. Na pierwszy rzut oka wprowadzenie systemu faworyzowania do losowego doboru próby wydaje się złym pomysłem, jednak złagodzenie nierównego doboru próby, a zatem praca nad usunięciem systematycznych uprzedzeń dotyczących płci, rasy, niepełnosprawności itd. jest znacznie bardziej adekwatna. Prosta próba losowa, w której każdy ma takie same szanse jak wszyscy, z dużym prawdopodobieństwem zagłuszy głosy i opinie członków populacji mniejszości. Dlatego wprowadzenie takiego systemu faworyzowania do technik pobierania próbek może być w porządku.

Jak mierzymy statystyki?

Gdy już mamy naszą próbkę, nadszedł czas, aby oszacować nasze wyniki. Załóżmy, że chcemy uogólnić szczęście naszych pracowników lub chcemy dowiedzieć się, czy wynagrodzenia w firmie bardzo różnią się w zależności od osoby. Oto kilka typowych sposobów mierzenia naszych wyników.

Miary środka

Miary środka to sposób, w jaki definiujemy środek lub środek zbioru danych. Robimy to, ponieważ czasami chcemy dokonać uogólnień na temat wartości danych. Na przykład, być może jesteśmy ciekawi, jakie są średnie opady w Seattle lub jaka jest mediana wzrostu dla europejskich mężczyzn. Jest to sposób na uogólnienie dużego zestawu danych, aby łatwiej było komuś przekazać. Miara środka to wartość w „środku" zbioru danych. Może to jednak oznaczać różne rzeczy dla różnych osób. Kto może

powiedzieć, gdzie jest środek zbioru danych? Jest tak wiele różnych sposobów definiowania centrum danych. Przyjrzyjmy się kilku. Średnią arytmetyczną zbioru danych można znaleźć, sumując wszystkie wartości, a następnie dzieląc ją przez liczbę wartości danych. Jest to prawdopodobnie najczęstszy sposób definiowania centrum danych, ale może być wadliwy! Załóżmy, że chcemy znaleźć średnią z następujących liczb:

import numpy as np.

np.mean([11, 15, 17, 14]) == 14.25

Dość prosto, nasza średnia wynosi 14,25 i wszystkie nasze wartości są do niej dość zbliżone. A co, jeśli wprowadzimy nową wartość: 31?

np.mean([11, 15, 17, 14, 31]) == 17.6

Ma to duży wpływ na średnią, ponieważ średnia arytmetyczna jest wrażliwa na wartości odstające. Nowa wartość, 31, jest prawie dwa razy większa niż reszta liczb, a zatem odchyla średnią. Inną, czasem lepszą miarą centrum jest mediana. Mediana to liczba znaleziona w środku zbioru danych, gdy jest on posortowany w zamów, jak pokazano:

np.median([11, 15, 17, 14]) == 14.5

np.median([11, 15, 17, 14, 31]) == 15

Zauważ, że wprowadzenie 31 przy użyciu mediany nie wpłynęło znacząco na medianę zbioru danych. Dzieje się tak, ponieważ mediana jest mniej wrażliwa na wartości odstające. Podczas pracy z zestawami danych z wieloma wartościami odstającymi czasami bardziej przydatne jest użycie mediany zbioru danych, natomiast jeśli dane nie zawierają wielu wartości odstających, a punkty danych są w większości blisko siebie, prawdopodobnie lepszym rozwiązaniem jest średnia. Ale jak możemy stwierdzić, czy dane są rozproszone? Cóż, będziemy musieli wprowadzić nowy rodzaj statystyki.

Miary zmienności

Miary środka są używane do ilościowego określenia środka danych, ale teraz zbadamy sposoby mierzenia, jak „rozprzestrzeniają się" gromadzone przez nas dane. Jest to przydatny sposób sprawdzenia, czy w naszych danych czai się wiele odstających elementów. Zacznijmy od przykładu. Weź pod uwagę, że bierzemy losową próbkę 24 naszych znajomych na Facebooku i piszemy ilu znajomych mieli na Facebooku. Oto lista:

friends = [109, 1017, 1127, 418, 625, 957, 89, 950, 946, 797, 981,

125, 455, 731, 1640, 485, 1309, 472, 1132, 1773, 906, 531, 742, 621]

np.mean(friends) == 789.1

Średnia tej listy to nieco ponad 789. Można więc powiedzieć, że według tej próbki przeciętny znajomy z Facebooka ma 789 znajomych. Ale co z osobą, która ma tylko 89 przyjaciół lub osobą, która ma ponad 1600 przyjaciół? W rzeczywistości niewiele z tych liczb jest naprawdę zbliżonych do 789. A może użyjemy mediany, jak pokazano, ponieważ na medianę zasadniczo nie wpływają wartości odstające:

np.median(friends) == 769.5

Mediana wynosi 769,5, co jest dość zbliżone do średniej. Hmm, dobra myśl, ale tak naprawdę nie wyjaśnia to, jak drastycznie wiele z tych punktów danych różni się od siebie. To właśnie statystycy

nazywają mierzeniem zmienności danych. Zacznijmy od wprowadzenia najbardziej podstawowej miary zmienności: zakresu. Zakres to po prostu wartość maksymalna minus wartość minimalna, jak pokazano

np.max(friends) – np.min(friends) == 1684

Zakres mówi nam, jak daleko znajdują się dwie najbardziej skrajne wartości. Obecnie zakres ten nie jest powszechnie używany, ale ma swoje zastosowanie w zastosowaniu. Czasami chcemy po prostu wiedzieć, jak rozłożone są wartości odstające. Jest to najbardziej przydatne w pomiarach naukowych lub pomiarach bezpieczeństwa. Załóżmy, że firma samochodowa chce zmierzyć, ile czasu zajmuje rozwinięcie się poduszki powietrznej. Znajomość średniej z tego czasu jest fajna, ale naprawdę chcą również wiedzieć, jak rozkłada się najwolniejszy czas w porównaniu z najszybszym. To dosłownie może być różnica między życiem a śmiercią. Wracając do przykładu z Facebooka, 1684 to nasz zakres, ale nie jestem pewien, czy mówi za dużo o naszych danych. Przyjrzyjmy się teraz najczęściej używanej mierze zmienności, czyli odchyleniu standardowemu. Jestem pewien, że wielu z was często słyszało ten termin i może nawet wzbudzać strach, ale co to tak naprawdę oznacza? Zasadniczo odchylenie standardowe, oznaczane przez s, gdy pracujemy z próbką populacji, mierzy, jak bardzo wartości danych odbiegają od średniej arytmetycznej. Jest to w zasadzie sposób, aby zobaczyć, jak rozłożone są dane. Istnieje ogólna formuła obliczania odchylenia standardowego, która wygląda następująco:

$$s = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

Tutaj:

• s jest naszym odchyleniem standardowym próbki

• x to każdy indywidualny punkt danych.

• $\overline{x}$ jest średnią danych

• n jest liczba punktów danych

Zanim wpadniesz, rozbijmy to. Dla każdej wartości w próbie weźmiemy tę wartość, odejmiemy od niej średnią arytmetyczną, podniesiemy do kwadratu różnicę, a kiedy dodamy w ten sposób każdy punkt, podzielimy całość przez n, liczbę punktów w próbie. Na koniec ze wszystkiego wyciągamy pierwiastek kwadratowy. Nie wchodząc w dogłębną analizę wzoru, pomyśl o tym w ten sposób: w zasadzie wywodzi się z wzoru na odległość. Zasadniczo to, co oblicza odchylenie standardowe, jest rodzajem średniej odległości odle,głości wartości danych od średniej arytmetycznej. Jeśli przyjrzysz się bliżej tej formule, zobaczysz, że ma ona sens:

• Biorąc , $x - \overline{x}$ znajdujesz dosłowną różnicę między wartością a średnią próbki.

• Podnosząc wynik do kwadratu, $(x - \overline{x})^2$ nakładamy większą karę na wartości odstające, ponieważ podniesienie do kwadratu dużego błędu powoduje, że jest on znacznie większy.

• Dzieląc przez liczbę pozycji w próbie, bierzemy (dosłownie) średnią kwadratową odległość między każdym punktem a średnią.

• Wyciągając pierwiastek kwadratowy z odpowiedzi, podajemy liczbę w zrozumiały dla nas sposób. Na przykład, podnosząc do kwadratu liczbę znajomych minus średnią, zmieniliśmy nasze jednostki na

kwadrat znajomych, co nie ma sensu. Wyciągnięcie pierwiastka kwadratowego sprawia, że nasze jednostki stają się z powrotem tylko „przyjaciółmi”.

Wróćmy do naszego przykładu na Facebooku, aby uzyskać wizualizację i dalsze wyjaśnienie tego. Zacznijmy obliczyć odchylenie standardowe. Więc zaczniemy obliczać kilka z nich. Przypomnijmy, że średnia arytmetyczna danych wynosiła około 789, więc użyjemy 789 jako średniej. Zaczynamy od różnicy między każdą wartością danych a średnią, podnosząc ją do kwadratu, dodając je wszystkie, dzieląc przez jeden mniej niż liczba wartości, a następnie wyciągając pierwiastek kwadratowy. Wyglądałoby to następująco:

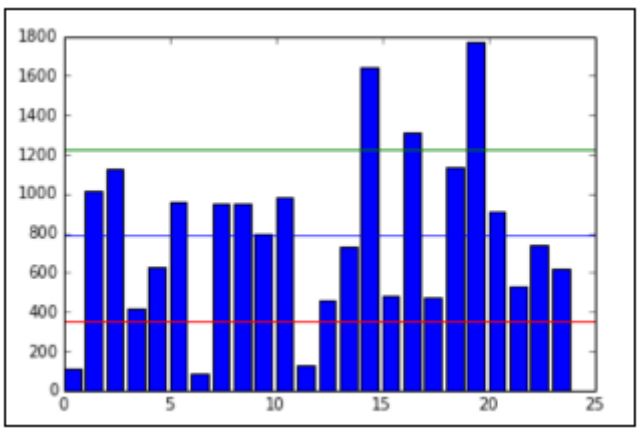
$$s = \sqrt{\frac{(109-789)^2 + (1017-789)^2 + \cdots + (621-789)^2}{24}}$$

Z drugiej strony możemy przyjąć podejście Pythona i zrobić to wszystko programowo (co jest zwykle preferowane).

np.std(friends) # == 425.2

Liczba 425 reprezentuje rozproszenie danych. Można powiedzieć, że 425 to rodzaj średniej odległości wartości danych od średniej. W prostych słowach oznacza to, że te dane są dość rozproszone. Tak więc nasze odchylenie standardowe wynosi około 425. Oznacza to, że liczba znajomych tych osób na Facebooku nie wydaje się być zbliżona do jednej liczby i jest to całkiem oczywiste, gdy wykreślimy dane na wykresie słupkowym, a także na wykresie średniej oraz wizualizacje odchylenia standardowego. Na poniższym wykresie każda osoba będzie reprezentowana przez pojedynczy słupek na wykresie słupkowym, a wysokość słupków reprezentuje liczbę znajomych, których mają osoby:

```
import matplotlib.pyplot as plt

%matplotlib inline

y_pos = range(len(friends))

plt.bar(y_pos, friends)

plt.plot((0, 25), (789, 789), 'b-')

plt.plot((0, 25), (789+425, 789+425), 'g-')

plt.plot((0, 25), (789-425, 789-425), 'r-')
```



Niebieska linia w środku jest narysowana na średniej (789), czerwona linia na dole jest narysowana na średniej minus odchylenie standardowe (789-425 = 364), a na koniec narysowana jest zielona linia w kierunku góry przy średniej plus odchylenie standardowe (789+425 = 1,214). Zwróć uwagę, że większość danych znajduje się między zieloną a czerwoną linią, podczas gdy osoby odstające żyją poza liniami. Mianowicie, są trzy osoby, które mają znajomego poniżej czerwonej linii i trzy osoby, które mają znajomego, powyżej zielonej linii. Należy wspomnieć, że jednostki odchylenia standardowego są w rzeczywistości tymi samymi jednostkami, co jednostki danych. W tym przykładzie powiedzielibyśmy, że odchylenie standardowe wynosi 425 znajomych na Facebooku. Inną miarą zmienności jest wariancja, jak opisano w poprzedniej części. Wariancja to po prostu odchylenie standardowe do kwadratu. Teraz wiemy, że odchylenie standardowe i wariancja są dobre do sprawdzenia, jak rozłożone są nasze dane i że możemy je wykorzystać wraz ze średnią do stworzenia pewnego rodzaju zakresu, w którym znajduje się wiele naszych danych. Ale co, jeśli chcemy porównać rozprzestrzenianie się dwóch różnych zbiorów danych, może nawet z zupełnie różnymi jednostkami? Tutaj w grę wchodzi współczynnik zmienności.

Definicja

Współczynnik zmienności definiuje się jako stosunek odchylenia standardowego danych do ich średniej. Ten współczynnik (który, nawiasem mówiąc, jest pomocny tylko wtedy, gdy pracujemy na poziomie pomiaru współczynnika, gdzie podział jest dozwolony i ma sens) jest sposobem na standaryzację odchylenia standardowego, co ułatwia porównywanie między zestawami danych. Często używamy tej miary, gdy próbujemy porównywać średnie, i rozprzestrzenia się ona na populacje, które istnieją w różnych skalach.

Przykład – pensje pracowników

Jeśli spojrzymy na średnią i odchylenie standardowe wynagrodzeń pracowników w tej samej firmie, ale w różnych działach, widzimy, że na pierwszy rzut oka porównanie różnic może być trudne.

Salaries of Company XYZ

Department	Mean Salary	SD	CoV
Mailroom	$25,000	$2,000	8.0%
Human Resources	$52,000	$7,000	13.5%
Executive	$124,000	$42,000	33.9%

Jest to szczególnie prawdziwe, gdy średnia pensja jednego działu wynosi 25 000 USD, podczas gdy w innym dziale średnia pensja jest sześciocyfrowa. Jeśli jednak spojrzymy na ostatnią kolumnę, która jest naszym współczynnikiem zmienności, staje się jasne, że ludzie w dziale wykonawczym mogą zarabiać więcej, ale pracownicy w dziale wykonawczym dostają szalenie różne pensje. Dzieje się tak prawdopodobnie dlatego, że dyrektor generalny zarabia znacznie więcej niż kierownik biura, który nadal jest w dziale wykonawczym, co sprawia, że dane są bardzo rozproszone. Z drugiej strony, wszyscy w kancelarii, chociaż nie zarabiają tyle pieniędzy, zarabiają mniej więcej tyle samo, co wszyscy inni w kancelarii, dlatego ich współczynnik zmienności wynosi tylko 8%. Dzięki miarom zmienności możemy zacząć odpowiadać na ważne pytania, takie jak rozłożenie tych danych lub ustalenie dobrego zakresu, w którym mieści się większość danych.

Miary względnej pozycji

Możemy łączyć zarówno miary centrów, jak i wariacji, aby tworzyć miary pozycji względnych. Miary zmienności miara, gdzie poszczególne wartości danych są pozycjonowane, względne do całego zbioru

danych. Zacznijmy od poznania bardzo ważnej wartości w statystyce, wskaźnika Z. Z-score to sposób na powiedzenie nam, jak daleko od średniej znajduje się pojedyncza wartość danych. Wynik z wartości danych x jest następujący:

$$z = \frac{x - \overline{x}}{s}$$

Gdzie:

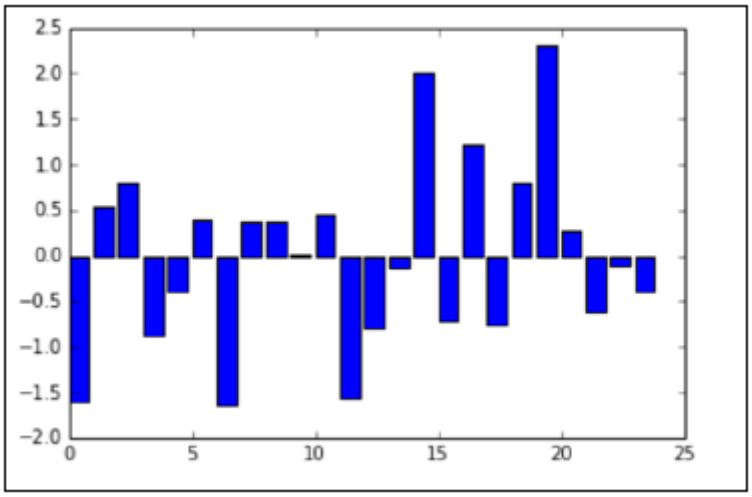
- x jest punktem danych
- $\overline{x}$ jest średnią
- s to odchylenie standardowe.

Pamiętaj, że odchylenie standardowe było (w pewnym sensie) średnią odległością danych od średniej, a teraz z-score jest indywidualną wartością dla każdego konkretnego punktu danych. Możemy znaleźć wynik Z wartości danych, odejmując ją od średniej i dzieląc przez odchylenie standardowe. Wynikiem będzie standaryzowana odległość, w jakiej wartość jest od średniej. Używamy wskaźnika Z w całej statystyce. Jest to bardzo skuteczny sposób normalizowania danych, które istnieją w bardzo różnych skalach, a także umieszczania danych w kontekście ich średniej. Weźmy nasze poprzednie dane o liczbie znajomych na Facebooku i ujednolicimy dane do z-score. Dla każdego punktu danych znajdziemy jego z-score, stosując poprzednią formułę. Weźmiemy każdą osobę, odejmiemy przeciętnych przyjaciół od wartości i podzielimy ją przez odchylenie standardowe, jak pokazano:

```
z_scores = []

m = np.mean(friends) # przeciętnych znajomych na Facebooku

s = np.std(friends) # znajomi z odchyleniem standardowym na Facebooku dla znajomego w znajomych:

z = (przyjaciel - m)/s # z-score

z_scores.append(z) # zrób listę wyników do wykreślenia
```

Teraz wykreślmy te wyniki Z na wykresie słupkowym. Poniższy wykres pokazuje te same osoby z naszego poprzedniego przykładu używającego znajomych na Facebooku, ale zamiast wysokości słupka pokazującej surową liczbę znajomych, teraz każdy słupek jest wynikiem z liczby znajomych, których mają na Facebooku. Jeśli wykreślimy z-scores, zauważymy parę rzeczy:

```
plt.bar(y_pos, z_scores)
```



• Mamy wartości ujemne (co oznacza, że punkt danych znajduje się poniżej średniej)

• Długości słupków nie reprezentują już surowej liczby znajomych, ale stopień, w jakim ta liczba znajomych różni się od średniej

Ten wykres bardzo ułatwia wytypowanie osób, które mają średnio dużo niższych i wyższych przyjaciół. Na przykład osoba z indeksem 0 ma średnio mniej znajomych . A co, jeśli chcemy narysować na wykresie odchylenia standardowe? Przypomnijmy, że wcześniej narysowaliśmy trzy poziome linie: jedną przy średniej, jedną przy średniej plus odchylenie standardowe (x + s) i jedną przy średniej minus odchylenie standardowe (x - s). Jeśli wstawimy te wartości do wzoru na z-score, otrzymamy:

$$\text{Z-score of } (\overline{x}\,\overline{x}) = \frac{x-\overline{x}}{s} = \frac{0}{s} = 0$$

$$\text{Z-score of } (x+s) = \frac{(\overline{x}+s)-\overline{x}}{s} = \frac{s}{s} = 1$$

$$\text{Z-score of } (x-s) \ \frac{(\overline{x}-s)-\overline{x}}{s} = \frac{-s}{s} = -1$$

To nie przypadek! Kiedy standaryzujemy dane za pomocą wskaźnika Z, nasze odchylenia standardowe stają się miarą z wyboru. Zobaczmy nasz nowy wykres z naniesionymi odchyleniami standardowymi:

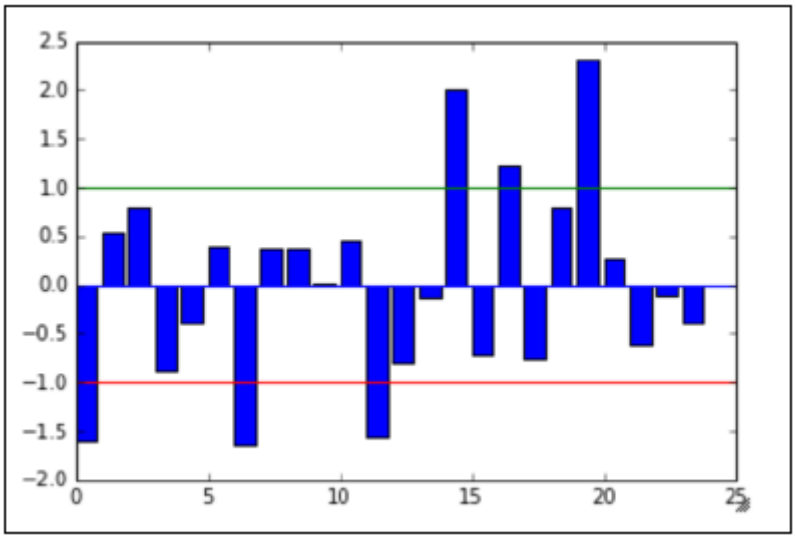
```
plt.bar(y_pos, z_scores)

plt.plot((0, 25), (1, 1), 'g-')

plt.plot((0, 25), (0, 0), 'b-')

plt.plot((0, 25), (-1, -1), 'r-')
```

Poprzedni kod dodaje w następujących trzech wierszach:

• Niebieska linia przy y = 0, która reprezentuje zero odchyleń standardowych od średniej (która znajduje się na osi x)

• Zielona linia, która reprezentuje jedno odchylenie standardowe powyżej średniej

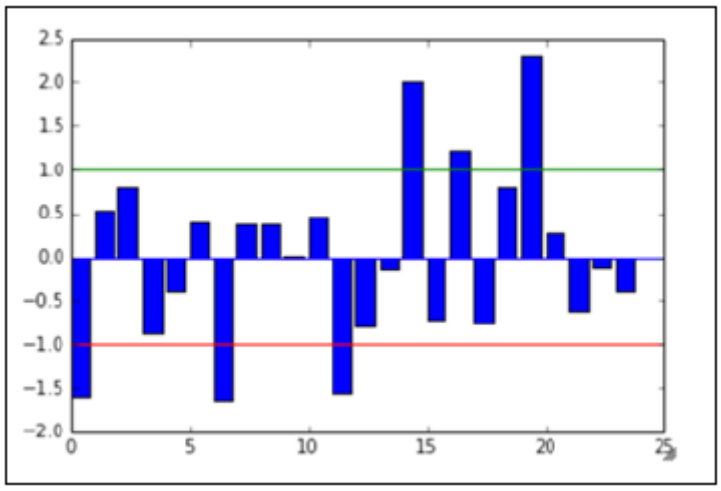
• Czerwona linia, która reprezentuje jedno odchylenie standardowe poniżej średniej



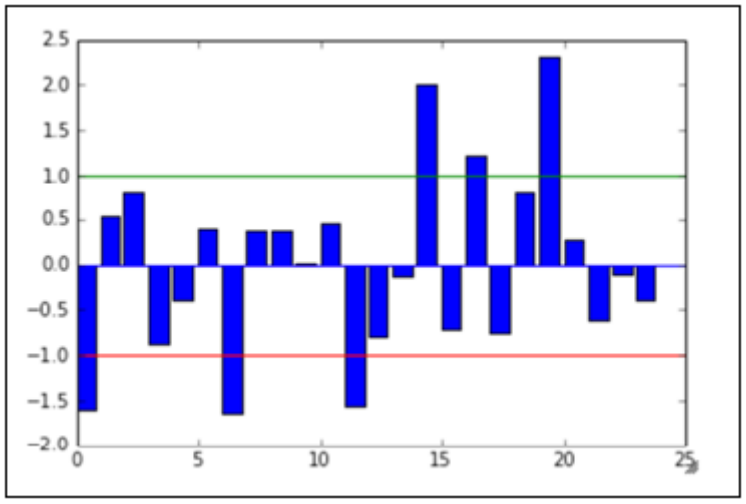
Kolory linii pasują do linii narysowanych na wcześniejszym wykresie surowej liczby znajomych. Jeśli przyjrzysz się uważnie, te same osoby nadal znajdują się poza zieloną i czerwoną linią. Mianowicie te same trzy osoby nadal spadają poniżej czerwonej (dolnej) linii, a te same trzy osoby spadają powyżej zielonej (górnej) linii.

W ramach tego skalowania możemy również użyć następujących stwierdzeń:

• Ten punkt danych jest o ponad jedno odchylenie standardowe od średniej:



• Ta osoba ma liczbę znajomych w granicach jednego odchylenia standardowego od średniej:



Wyniki Z to skuteczny sposób na standaryzację danych. Oznacza to, że cały zestaw możemy postawić na tej samej skali. Na przykład, jeśli zmierzymy również ogólną skalę szczęścia każdej osoby (która wynosi od 0 do 1), możemy mieć zestaw danych podobny do następującego zestawu danych:

```
friends = [109, 1017, 1127, 418, 625, 957, 89, 950, 946, 797, 981,
125, 455, 731, 1640, 485, 1309, 472, 1132, 1773, 906, 531, 742, 621]

happiness = [.8, .6, .3, .6, .6, .4, .8, .5, .4, .3, .3, .6, .2, .8,
1, .6, .2, .7, .5, .3, .1, 0, .3, 1]


import pandas as pd

df = pd.DataFrame({'friends':friends, 'happiness':happiness})
df.head()
```

	friends	happiness
0	109	0.8
1	1017	0.6
2	1127	0.3
3	418	0.6
4	625	0.6

Te punkty danych znajdują się w dwóch różnych wymiarach, z których każdy ma zupełnie inną skalę. Liczba znajomych może wynosić w tysiącach, podczas gdy nasz wynik zadowolenia utrzymuje się między 0 a 1. Aby temu zaradzić (a w przypadku niektórych modeli statystycznych/uczenia maszynowego ta koncepcja stanie się niezbędna), możemy po prostu ustandaryzować zbiór danych za pomocą gotowego pakietu standaryzacyjnego w scikit-learn, jak następuje:

from sklearn import preprocessing

df_scaled = pd.DataFrame(preprocessing.scale(df), columns = ['friends_

scaled', 'happiness_scaled'])

df_scaled.head()

Ten kod skaluje jednocześnie zarówno kolumny przyjaciół, jak i szczęście, ujawniając w ten sposób wynik Z dla każdej kolumny. Należy zauważyć, że w ten sposób moduł przetwarzania wstępnego w sklearn wykonuje następujące czynności osobno dla każdej kolumny:

• Znalezienie średniej kolumny

• Znalezienie odchylenia standardowego kolumny

• Zastosowanie funkcji z-score do każdego elementu w kolumnie

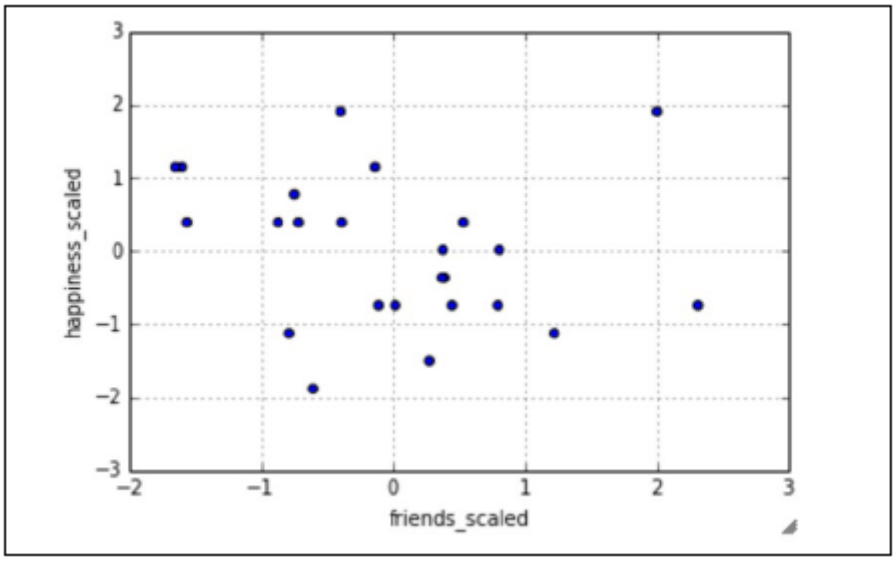
Wynikiem są dwie kolumny, jak pokazano, które istnieją w tej samej skali, nawet jeśli wcześniej nie były:

	friends_scaled	happiness_scaled
0	-1.599495	1.153223
1	0.536040	0.394939
2	0.794750	-0.742486
3	-0.872755	0.394939
4	-0.385909	0.394939

Teraz możemy wykreślić przyjaciół i szczęście w tej samej skali, a wykres będzie przynajmniej czytelny.

df_scaled.plot(kind='scatter', x = 'friends_scaled', y = 'happiness_

scaled')



Teraz nasze dane są standaryzowane do wskaźnika Z, a ten wykres punktowy jest dość łatwy do zinterpretowania! W dalszej części ta idea standaryzacji nie tylko sprawi, że nasze dane będą bardziej zrozumiałe, ale będzie również niezbędna w optymalizacji naszego modelu. Wiele algorytmów uczenia maszynowego będzie wymagało od nas standardowych kolumn, ponieważ są one zależne od pojęcia skali.

Część wnikliwa – korelacje w danych

W tym tekście omówimy różnicę między posiadaniem danych a posiadaniem praktycznych spostrzeżeń na temat danych. Posiadanie danych to tylko jeden krok do udanej operacji analizy danych. Możliwość uzyskania, oczyszczenia i wykreślenia danych pomaga opowiedzieć historię, którą dane mają do zaoferowania, ale nie może ujawnić morału. Aby posunąć cały ten przykład o krok dalej, przyjrzymy się relacji między posiadaniem znajomych na Facebooku a szczęściem. W kolejnych rozdziałach przyjrzymy się konkretnemu algorytmowi uczenia maszynowego, który próbuje znaleźć relacje między cechami ilościowymi, zwaną regresją liniową, ale nie musimy czekać do tego czasu, aby zacząć formułować hipotezy. Mamy próbkę ludzi, miarę ich obecności w sieci i zgłaszanego szczęścia. Pytanie dnia brzmi: czy możemy znaleźć związek między liczbą znajomych na Facebooku a ogólnym szczęściem? Oczywiście jest to duże pytanie i należy je traktować z szacunkiem. Eksperymenty mające odpowiedzieć na to pytanie powinny być prowadzone w warunkach laboratoryjnych,

ale możemy zacząć formułować hipotezę na temat tego pytania. Biorąc pod uwagę charakter naszych danych, tak naprawdę mamy tylko trzy możliwości postawienia hipotezy:

• Istnieje pozytywny związek między liczbą znajomych online a szczęściem (w miarę wzrostu jednego, rośnie drugi)

• Istnieje między nimi negatywny związek (w miarę wzrostu liczby przyjaciół twoje szczęście spada)

• Nie ma powiązania między zmiennymi (gdy jedna się zmienia, druga nie zmienia się tak bardzo)

Czy możemy użyć podstawowych statystyk do sformułowania hipotezy na to pytanie? Mówię, że możemy! Ale najpierw musimy wprowadzić pojęcie zwane korelacją. Współczynniki korelacji są miarą ilościową opisującą siłę powiązania/związek między dwiema zmiennymi. Korelacja między dwoma zestawami danych mówi nam o tym, jak poruszają się razem. Czy zmiana jednego pomoże nam przewidzieć drugie? Ta koncepcja jest nie tylko interesująca w tym przypadku, ale jest jednym z podstawowych założeń, które wiele modeli uczenia maszynowego przyjmuje na danych. Aby wiele algorytmów predykcyjnych działało, opierają się na fakcie, że istnieje jakiś związek między zmiennymi, na które patrzymy. Algorytmy uczące wykorzystują następnie tę zależność, aby dokonać dokładnych prognoz. Kilka rzeczy, o których należy pamiętać o współczynniku korelacji, jest następujące:

• Będzie leżeć między -1 a 1

• Im większa wartość bezwzględna (bliższa -1 lub 1), tym silniejszy związek między zmiennymi:

- Najsilniejsza korelacja to -1 lub 1

- Najsłabsza korelacja to 0

• Dodatnia korelacja oznacza, że gdy jedna zmienna wzrasta, druga również ma tendencję do wzrostu

• Ujemna korelacja oznacza, że gdy jedna zmienna wzrasta, druga ma tendencję do zmniejszania się

Możemy użyć Pand, aby szybko pokazać nam współczynniki korelacji między każdą funkcją a każdą inną funkcją w Dataframe, jak pokazano na rysunku:

```
# correlation between variables
df.corr()
```

	friends	happiness
friends	1.000000	-0.216199
happiness	-0.216199	1.000000

Ta tabela pokazuje korelację między przyjaciółmi a szczęściem. Zwróć uwagę na dwie pierwsze rzeczy, pokazane w następujący sposób:

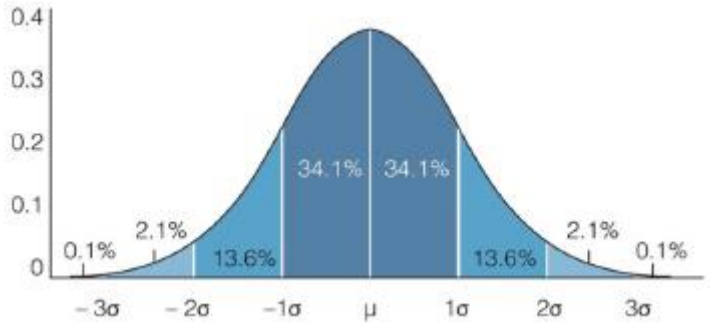
• Przekątna matrycy jest wypełniona pozytywem. Dzieje się tak, ponieważ reprezentują one korelację między zmienną a nią samą, która oczywiście tworzy idealną linię, dzięki czemu korelacja jest idealnie dodatnia!

• Matryca jest symetryczna na przekątnej. Dotyczy to każdej macierzy korelacji wykonanej w Pandas.

Istnieje kilka zastrzeżeń do zaufania do współczynnika korelacji. Po pierwsze, ogólnie rzecz biorąc, korelacja będzie próbowała zmierzyć liniową zależność między zmiennymi. Oznacza to, że jeśli nie ma widocznej korelacji ujawnionej przez tę miarę, nie oznacza to, że nie ma związku między zmiennymi, a jedynie, że nie ma linii najlepszego dopasowania, która łatwo przechodzi przez te linie. Może istnieć nieliniowa relacja, która definiuje te dwie zmienne. Ważne jest, aby zdać sobie sprawę, że związek przyczynowy nie jest implikowany przez korelację. Tylko dlatego, że istnieje słaba ujemna korelacja między tymi dwiema zmiennymi, niekoniecznie oznacza to, że ogólne szczęście spada wraz ze wzrostem liczby znajomych, których utrzymujesz na Facebooku. Ta przyczynowość musi być dalej testowana i w dalszej części postaramy się właśnie to zrobić. Podsumowując, możemy użyć korelacji, aby postawić hipotezy dotyczące związku między zmiennymi, ale będziemy musieli użyć bardziej wyrafinowanych metod statystycznych i algorytmów uczenia maszynowego, aby utrwalić te założenia i hipotezy.

Empiryczna reguła

Przypomnijmy, że rozkład normalny jest zdefiniowany jako mający określony rozkład prawdopodobieństwa, który przypomina krzywą dzwonową. W statystykach uwielbiamy, gdy nasze dane zachowują się normalnie. Na przykład, jeśli mamy dane przypominające rozkład normalny, tak:



Reguła empiryczna mówi, że możemy oczekiwać, że pewna ilość danych będzie istnieć między zestawami odchyleń standardowych. W szczególności reguła empiryczna określa dla danych, które są dystrybuowane normalnie:

• około 68% danych mieści się w 1 odchyleniu standardowym

• około 95% danych mieści się w 2 odchyleniach standardowych

• około 99,7% danych mieści się w 3 odchyleniach standardowych

Na przykład zobaczmy, czy dane naszych znajomych z Facebooka to wytrzymują. Użyjmy naszej ramki danych, aby znaleźć odsetek osób, które mieszczą się w zakresie 1, 2 i 3 odchyleń standardowych od średniej, jak pokazano:

```
# znalezienie odsetka osób w obrębie jednego odchylenia standardowego od
średnia
w obrębie_1_std = df_skalowane[(df_skalowane['znajomi_skalowane'] <= 1) & (df_
scaled['friends_scaled'] >= -1)].shape[0]
within_1_std / float(df_scaled.shape[0])
# 0,75
# znalezienie odsetka osób w obrębie dwóch odchyleń standardowych od
średnia
w ciągu_2_std = df_skalowane[(df_skalowane['znajomi_skalowane'] <= 2) & (df_
scaled['friends_scaled'] >= -2)].shape[0]
within_2_std / float(df_scaled.shape[0])
# 0,916
# znalezienie odsetka osób w obrębie trzech odchyleń standardowych od
średnia
w_3_std = df_skalowane[(df_skalowane['przyjaciele_skalowane'] <= 3) & (df_
scaled['friends_scaled'] >= -3)].shape[0]
within_3_std / float(df_scaled.shape[0])
# 1,0
```

Widzimy, że nasze dane wydają się być zgodne z regułą empiryczną. Około 75% ludzi mieści się w obrębie jednego odchylenia standardowego średniej. Około 92% ludzi mieści się w granicach dwóch odchyleń standardowych, a wszystkie mieszczą się w granicach trzech odchyleń standardowych.

Przykład - wyniki egzaminu

Załóżmy, że mierzymy wyniki egzaminu, a wyniki zazwyczaj mają rozkład normalny w kształcie dzwonu. Średnia z egzaminu wyniosła 84%, a odchylenie standardowe 6%. Z pewną dozą pewności możemy powiedzieć, że:

• Około 68% klasy uzyskało od 78% do 90%, ponieważ 78 to 6 jednostek poniżej 84, a 90 to 6 jednostek powyżej 84

• Gdybyśmy zostali zapytani, jaki procent klasy uzyskał wynik między 72 a 96%, zauważylibyśmy, że 72 to 2 odchylenia standardowe poniżej średniej, a 96 to 2 odchylenia standardowe powyżej średniej, więc reguła empiryczna mówi nam, że około 95 % klasy, która zdobyła punkty w tym zakresie.

Jednak nie wszystkie dane mają rozkład normalny, więc nie zawsze możemy zastosować regułę empiryczną. Mamy inne twierdzenie, które pomaga nam analizować każdy rodzaj dystrybucji. W następnym rozdziale omówimy szczegółowo, kiedy możemy przyjąć rozkład normalny. Dzieje się tak, ponieważ wiele testów statystycznych i hipotez wymaga, aby dane bazowe pochodziły z populacji o rozkładzie normalnym.

Wcześniej, gdy standaryzowaliśmy nasze dane do wyniku z, nie wymagaliśmy założenia rozkładu normalnego.

Podsumowanie

W tej części omówiliśmy wiele podstawowych statystyk wymaganych przez większość analityków danych. Omówiono wszystko, od tego, jak uzyskujemy/próbkujemy dane, po standaryzację danych zgodnie z z-score i zastosowaniami reguły empirycznej. W kolejnej części przyjrzymy się znacznie bardziej zaawansowanym aplikacjom statystyki. Jedną z rzeczy, które rozważymy, jest to, jak używać testów hipotez na danych, które możemy założyć, że są normalne. Korzystając z tych testów, będziemy również określać ilościowo nasze błędy i określać najlepsze praktyki w celu ich rozwiązania.