2. Statystyka dla analityki biznesowej

Witamy w świecie statystyki biznesowej, w którym dane przekształcają się w znaczące spostrzeżenia, kierując podejmowaniem decyzji i odkrywając ukryte prawdy. W tej kompleksowej eksploracji wyruszamy w podróż, aby zdemistyfikować podstawowe koncepcje i techniki statystyczne, które stanowią podstawę rygorystycznej analizy danych biznesowych. Od zrozumienia zawiłości rozkładów do stosowania testowania hipotez i konstruowania przedziałów ufności, każdy rozdział rozwija nowy aspekt umiejętności statystycznych.

W sercu analizy statystycznej leży **rozkład normalny**, krzywa w kształcie dzwonu, która przenika niezliczone zjawiska w przyrodzie i ludzkim zachowaniu. W tej części zagłębiamy się w istotę rozkładu normalnego, odkrywając jego właściwości i znaczenie we wnioskowaniu statystycznym. Za pomocą wizualizacji i rzeczywistych przykładów wyjaśniamy wszechobecność tego fundamentalnego rozkładu i jego rolę jako kamienia węgielnego teorii statystycznej.

Odchylenie standardowe służy jako kompas w krajobrazie statystycznym, prowadząc nas przez zmienność nieodłącznie związaną ze zbiorami danych. W tym rozdziale zostanie dokonana analiza pojęcia odchylenia standardowego, ujawniając jego znaczenie w ilościowym określaniu dyspersji i ocenie rozrzutu punktów danych. Uzbrojony w głębsze zrozumienie odchyleń standardowych, będziesz poruszać się po danych z pewnością siebie, precyzyjnie rozpoznając wzorce i wartości odstające.

Zmienne stanowią elementy składowe analizy statystycznej, a każda z nich posiada odrębne cechy i implikacje. Niniejszy rozdział wyjaśnia dychotomię między zmiennymi ciągłymi i dyskretnymi, rzucając światło na ich rolę w modelowaniu i interpretacji danych. Dzięki zrozumieniu niuansów związanych z typami zmiennych, będziesz mógł w pełni wykorzystać potencjał technik statystycznych dostosowanych do różnych struktur danych.

Rozkład statystyki próbkowania służy jako podstawa wnioskowania statystycznego, wypełniając lukę między obserwacjami z próby a parametrami populacji. W tym rozdziale rozwikłamy koncepcję rozkładu próbkowania, wyjaśniając jego znaczenie w formułowaniu probabilistycznych stwierdzeń dotyczących charakterystyki populacji. Dzięki konkretnym

przykładom rozwiniesz intuicyjne zrozumienie roli rozkładu próbkowania w solidnej analizie statystycznej.

Centralne Twierdzenie Graniczne jest jak latarnia morska statystycznego oświecenia, oświetlająca drogę do wiarygodnego wnioskowania w obliczu niepewności. Niniejszy rozdział demistyfikuje Centralne Twierdzenie Graniczne, ujawniając jego transformacyjną moc w stabilizowaniu średnich prób i ułatwianiu testowania hipotez. Uzbrojony w spostrzeżenia zebrane w tym rozdziale, będziesz mógł wykorzystać Centralne Twierdzenie Graniczne do wyciągnięcia znaczących wniosków z danych empirycznych.

Zrozumienie **testowania hipotez** jest niezbędne do podejmowania decyzji opartych na danych. Pozwala nam określić, czy zaobserwowane wzorce w danych są znaczące, czy po prostu wynikają z przypadku. Stosując testowanie hipotez, możemy oceniać założenia, porównywać grupy i oceniać istotność statystyczną wyników, dzięki czemu jest to niezbędne narzędzie w badaniach naukowych, analizach biznesowych i wielu innych dziedzinach

Wyniki *Z* i tabele *Z* służą jako pomoce nawigacyjne w morzu standardowego rozkładu normalnego, ułatwiając znormalizowane porównania i obliczenia prawdopodobieństwa. Niniejszy rozdział wyjaśnia zawiłości współczynników *Z*, umożliwiając interpretację znormalizowanych wyników i wykorzystanie tabel Z do analizy statystycznej. Dzięki biegłej znajomości współczynników Z będziesz poruszać się po rozległym obszarze rozkładu normalnego z pewnością i precyzją.

W sytuacjach, w których liczebność próby jest niewielka lub nieznane jest odchylenie standardowe populacji, **współczynniki *t*** i **tabele *t*** stają się niezbędnymi narzędziami analizy statystycznej. Niniejszy rozdział odkrywa tajemnice wyników *t*, prowadząc przez ich obliczanie i interpretację za pomocą tabel *t*. Uzbrojony w tę wiedzę, będziesz poruszać się po niuansach rozkładów *t* z finezją, zapewniając solidne wnioskowanie w różnych scenariuszach statystycznych.

Rozkład normalny i rozkład *t* są filarami teorii prawdopodobieństwa, a każdy z nich posiada unikalne cechy i zastosowania. W tym rozdziale wyjaśniamy różnice między tymi rozkładami, umożliwiając rozeznanie, kiedy zastosować każdy z nich w analizie statystycznej. Dzięki praktycznym przykładom i analizom porównawczym wybierzesz spośród rozkładów normalnego i *t*, wzbogacając swój zestaw narzędzi statystycznych.

Przedziały ufności zapewniają wgląd w niepewność otaczającą parametry populacji, umożliwiając nam ilościowe określenie precyzji naszych szacunków. W tym rozdziale zbadamy konstrukcję przedziałów ufności dla średnich i częstotliwości/odsetka/frakcji, odkrywając metodologię i interpretację tych podstawowych narzędzi statystycznych. Dzięki opanowaniu przedziałów ufności będziesz w stanie przekazać niepewność nieodłącznie związaną z wynikami w sposób przejrzysty i rygorystyczny.

Podczas gdy **wartości *p*** stanowią bramę do wnioskowania statystycznego, ich błędna interpretacja może prowadzić do błędnych wniosków i błędnych decyzji.
Niniejszy rozdział analizuje potencjalne pułapki związane z nadmiernym poleganiem na wartościach *p*, podkreślając znaczenie kontekstu i wielkości efektu w analizie statystycznej. Dzięki krytycznej analizie i praktycznym spostrzeżeniom będziesz poruszać się po złożoności wartości *p* z zachowaniem czujności, zapewniając integralność wniosków statystycznych.

Na tych stronach znajdują się klucze do odblokowania tajemnic analizy statystycznej, umożliwiając poruszanie się po złożoności danych z pewnością i precyzją. Wyruszając w tę podróż razem, niech ciekawość będzie naszym kompasem, a dociekanie naszym światłem przewodnim, oświetlającym drogę do głębszego zrozumienia i praktycznych spostrzeżeń.

2.1 Rozkład normalny

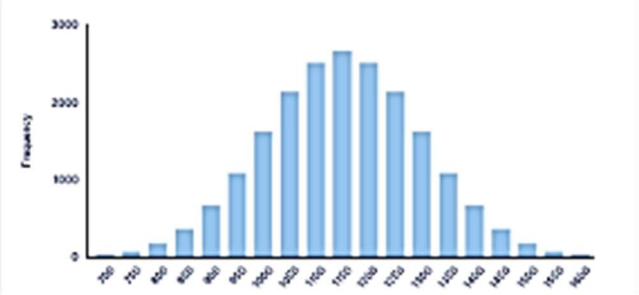
W sercu analizy statystycznej leży rozkład normalny, wszechobecny rozkład prawdopodobieństwa, który służy jako punkt odniesienia dla wielu technik statystycznych. Zagłębimy się w jego charakterystykę, symetryczną krzywą w kształcie dzwonu i jego znaczenie w zrozumieniu rozkładu danych.

Rozkład normalny znajduje zastosowanie w różnych dziedzinach, w tym w finansach, psychologii, inżynierii i biologii. Od modelowania cen akcji po zrozumienie rozkładu ludzkiego wzrostu, rozkład normalny służy jako wszechstronne narzędzie do analizy i interpretacji danych.

W tym rozdziale zagłębimy się w matematyczne właściwości rozkładu normalnego, badając, jak obliczać prawdopodobieństwa, percentyle i wskaźniki *z*-score. Ponadto omówimy praktyczne techniki wizualizacji i interpretacji rozkładów normalnych za pomocą histogramów, wykresów gęstości i funkcji rozkładu skumulowanego.

Pod koniec tego rozdziału będziesz kierował się głębokim uznaniem dla rozkładu normalnego i jego znaczenia w analizie statystycznej. Uzbrojony w tę wiedzę, będziesz dobrze przygotowany do radzenia sobie z bardziej zaawansowanymi koncepcjami statystycznymi i stosowaniem ich w rzeczywistych zestawach danych. Wyruszmy razem w tę podróż, aby rozwikłać tajemnice rozkładu normalnego.

Rozkład normalny, znany również jako rozkład Gaussa lub krzywa dzwonowa, wykazuje symetryczny rozkład danych bez asymetrii. Na wykresie dane tworzą krzywą w kształcie dzwonu, przy czym większość wartości gromadzi się wokół środka i maleje w miarę oddalania się od niego.



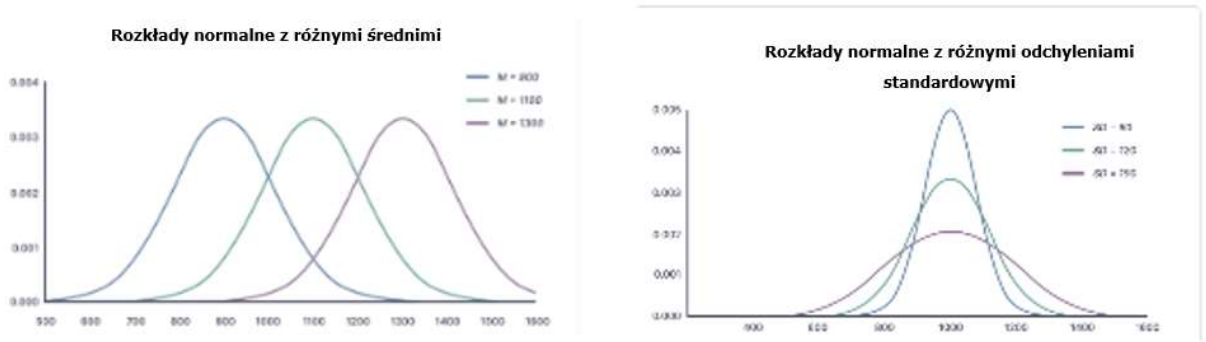
Rysunek 2.1 Przykład rozkładu gaussowskiego lub krzywej dzwonowej

Różne zmienne w naukach przyrodniczych i społecznych zazwyczaj wykazują rozkład normalny lub jego przybliżenie. Przykłady obejmują wzrost, wagę urodzeniową, umiejętność czytania, satysfakcję z pracy i wyniki SAT. Ze względu na powszechność zmiennych o rozkładzie normalnym, liczne testy statystyczne są dostosowane do takich populacji. Biegłość w zrozumieniu charakterystyki rozkładów normalnych upoważnia osoby do stosowania statystyk wnioskowania do porównywania grup i generowania szacunków populacji z próbek.

Rozkłady normalne mają kluczowe cechy, które łatwo zauważyć na wykresach:

- średnia, mediana i dominanta są dokładnie takie same,
- rozkład jest symetryczny względem średniej - połowa wartości spada poniżej średniej, a połowa powyżej średniej,
- rozkład można opisać za pomocą dwóch wartości: średniej i odchylenia standardowego.

Średnia służy jako parametr lokalizacji, dyktując środek szczytu krzywej. Dostosowanie średniej odpowiednio przesuwa krzywą: zwiększenie jej przesuwa krzywą w prawo, a zmniejszenie przesuwa krzywą w lewo. Tymczasem odchylenie standardowe działa jako parametr skali, wpływając na rozrzut lub szerokość krzywej.



Rysunek 2.2 Rozkład normalny z różnymi średnimi i różnymi odchyleniami

Odchylenie standardowe rozciąga lub ściska krzywą. Małe odchylenie standardowe skutkuje wąską krzywą, podczas gdy duże odchylenie standardowe prowadzi do szerokiej krzywej.

2.2 Reguła empiryczna

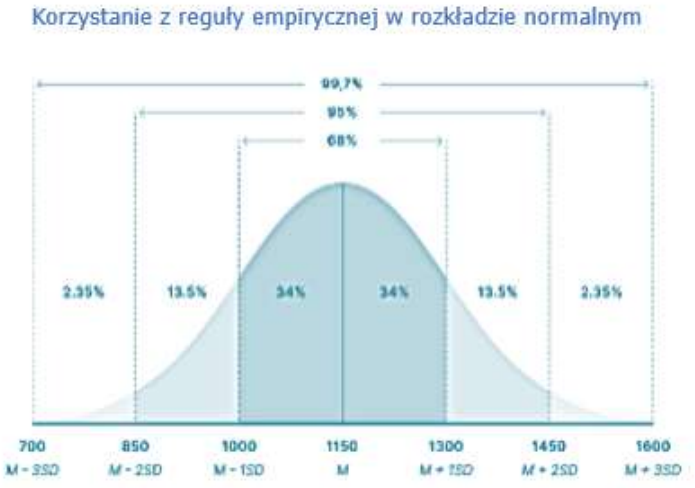
Reguła empiryczna, znana również jako reguła 68-95-99,7, zapewnia wgląd w rozkład wartości w rozkładzie normalnym:

- około 68% wartości mieści się w zakresie jednego odchylenia standardowego od średniej,
- około 95% wartości mieści się w zakresie dwóch odchyleń standardowych od średniej,
- około 99,7% wartości mieści się w zakresie trzech odchyleń standardowych od średniej.

Na przykład, rozważmy scenariusz, w którym zbierane są wyniki SAT od studentów na nowym kursie przygotowującym do testu, a dane są zgodne z rozkładem normalnym ze średnim wynikiem (M) 1150 i odchyleniem standardowym (SD) 150.

Zastosowanie reguły empirycznej daje następujące wnioski:

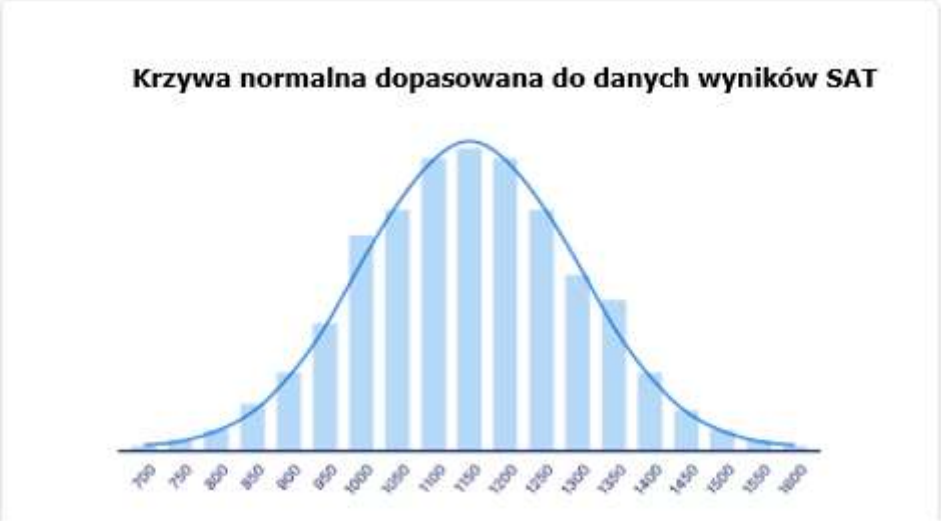
- około 68% wyników mieści się w zakresie od 1000 do 1300, co odpowiada 1 odchyleniu standardowemu powyżej i poniżej średniej,
- około 95% wyników mieści się w zakresie od 850 do 1450, co stanowi 2 odchylenia standardowe powyżej i poniżej średniej,
- prawie wszystkie wyniki, około 99,7%, mieszczą się w zakresie od 700 do 1600, obejmując 3 odchylenia standardowe powyżej i poniżej średniej.



Rysunek 2.3 Reguła empiryczna w rozkładzie normalnym

Reguła empiryczna oferuje szybką metodę oceny danych, umożliwiając wykrywanie wartości odstających lub wyjątkowych, które odbiegają od oczekiwanego wzorca. W przypadkach, w których dane z małych próbek znacznie odbiegają od tego wzorca, bardziej odpowiednie mogą być alternatywne rozkłady, takie jak rozkład *t*. Identyfikacja rozkładu zmiennej pozwala na zastosowanie odpowiednich testów statystycznych.

2.3 Wzór krzywej normalnej



Rysunek 2.4 Wyniki SAT wykazują rozkład zbliżony do normalnego

W funkcji gęstości prawdopodobieństwa obszar pod krzywą reprezentuje prawdopodobieństwo. Biorąc pod uwagę, że rozkład normalny służy jako rozkład prawdopodobieństwa, skumulowany obszar pod krzywą niezmiennie sumuje się do 1 lub 100%. Chociaż wzór na normalną funkcję gęstości prawdopodobieństwa może wydawać się skomplikowany, jego wykorzystanie wymaga jedynie znajomości średniej populacji i odchylenia standardowego. Podstawiając te parametry do wzoru, można określić gęstość prawdopodobieństwa związaną z dowolną wartością x.

- *f*(*x*) = prawdopodobieństwo,
- *x* = wartość zmiennej,
- μ = średnia arytmetyczna,
- σ = odchylenie standardowe,
- σ^2 = wariancja.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Przykład:

Korzystając z funkcji gęstości prawdopodobieństwa, chcesz poznać prawdopodobieństwo, że wyniki egzaminu SAT w próbie przekroczą 1380 punktów.

Na wykresie funkcji gęstości prawdopodobieństwa, prawdopodobieństwo jest zacieniowanym obszarem pod krzywą, który leży na prawo od miejsca, w którym wyniki SAT są równe 1380.

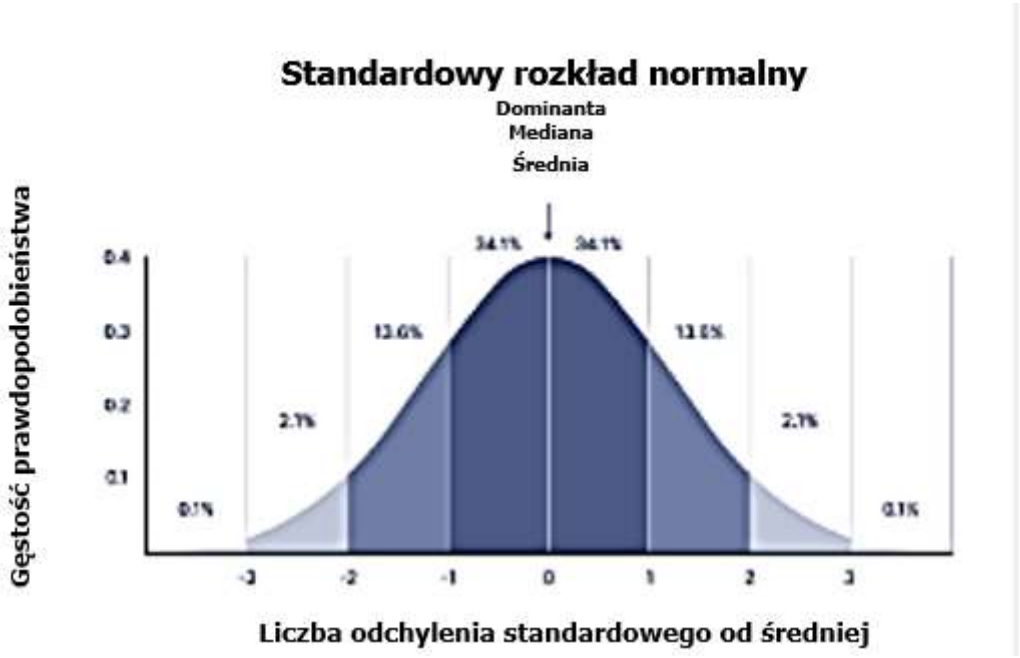
Rysunek 2.5 Wykres funkcji gęstości prawdopodobieństwa

Wartość prawdopodobieństwa tego wyniku można znaleźć przy użyciu standardowego rozkładu normalnego.

2.4 Standardowy rozkład normalny

Standardowy rozkład normalny, znany jako **rozkład *z*** wyróżnia się tym, że ma średnią równą 0 i odchylenie standardowe równe 1. Każdy rozkład normalny może być postrzegany jako transformacja standardowego rozkładu normalnego, podlegająca dostosowaniom skali, położenia lub obu tych elementów.

W kontekście rozkładu *z*, poszczególne obserwacje, które są zwykle oznaczane jako *x* w rozkładach normalnych, są określane jako wskaźnik *z*. Reprezentują one liczbę odchyleń standardowych, o jaką każda wartość odbiega od średniej. W związku z tym przekształcenie wartości z dowolnego rozkładu normalnego we wskaźniki *z* ułatwia porównanie i analizę w ramach standardowego rozkładu normalnego.



Rysunek 2.6 Wykres standardowego rozkładu normalnego

Wystarczy znać średnią i odchylenie standardowe rozkładu, aby znaleźć wskaźnik *z* dla danej wartości.

Objaśnienie wzoru na wskaźnik *z*:

- x = wartość indywidualna,
- μ = średnia arytmetyczna,
- σ = odchylenie standardowe.

$$z = \frac{x - \mu}{\sigma}$$

Rozkład normalny jest przekształcany w standardowy rozkład normalny z kilku powodów:

- aby znaleźć prawdopodobieństwo, że obserwacje w rozkładzie znajdą się powyżej lub poniżej danej wartości,
- aby znaleźć prawdopodobieństwo, że średnia z próby istotnie różni się od znanej średniej z populacji,
- aby porównać wyniki z rozkładów z różnymi średnimi i odchyleniami standardowymi.

2.5 Znajdowanie prawdopodobieństwa przy użyciu rozkładu *z*

Każdy wskaźnik *z* odpowiada prawdopodobieństwu, często określanemu jako wartość *p*, wskazując prawdopodobieństwo zaobserwowania wartości poniżej określonego wskaźnika *z*.

Przekształcając pojedynczą wartość we wskaźnik *z*, można określić prawdopodobieństwo wystąpienia wszystkich wartości do tego punktu w rozkładzie normalnym.

Rozważmy na przykład scenariusz, w którym chcemy ustalić prawdopodobieństwo, że wyniki SAT w rozważanej próbie przekroczą 1380. Początkowo oblicza się wynik z przy użyciu średniej i odchylenia standardowego rozkładu. Przy średniej 1150 i odchyleniu standardowym 150, wskaźnik *z* ujawnia liczbę odchyleń standardowych, o które 1380 odbiega od średniej.

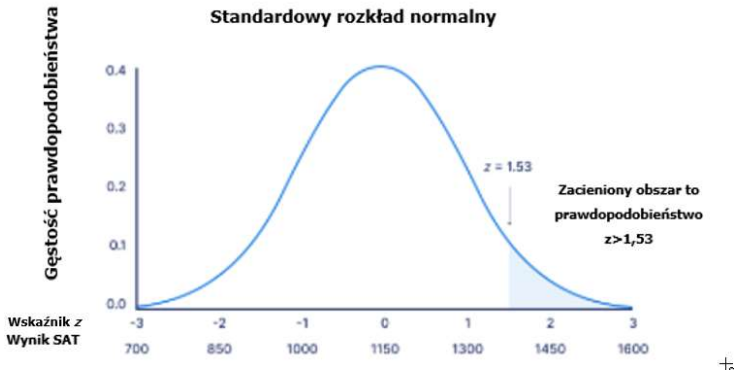
Wzór	**Obliczenie**
$z = \frac{x - \mu}{\sigma}$	$z = \frac{1380 - 1150}{150} = 1{,}53$

Dla wyniku z równego 1,53 wartość *p* wynosi 0,937. Jest to prawdopodobieństwo, że wynik SAT wynosi 1380 lub mniej (93,7%) i jest to obszar pod krzywą na lewo od zacieniowanego obszaru.

Aby znaleźć zacieniony obszar, należy odjąć 0,937 od 1, co stanowi całkowity obszar pod krzywą.

Prawdopodobieństwo x > 1380 = 1 – 0,937 = **0,063**

Oznacza to, że prawdopodobnie tylko 6,3% wyników SAT w próbie przekracza 1380.



Rysunek 2.7 Standardowy rozkład normalny ze wskazanym wynikiem

2.6 Rozkład próbkowania (rozkład próby)

Rozkłady próbkowania stanowią podstawę wnioskowania statystycznego, umożliwiając nam wyciąganie wniosków na temat populacji na podstawie danych z próby. Zagłębimy się w zawiłości rozkładów próbkowania, rozumiejąc, w jaki sposób odzwierciedlają one zmienność statystyk próby i ich kluczową rolę w testowaniu hipotez.

Rozkład próby odnosi się do rozkładu statystyki, takiej jak średnia próby lub odsetek z próby, uzyskanej z wielu prób o tej samej wielkości pobranych z populacji. Zapewnia wgląd w zachowanie statystyk próby i ich zmienność w różnych próbach.

2.7 Centralne Twierdzenie Graniczne i Rozkład Próbkowania

Centralne Twierdzenie Graniczne (CLT) to podstawowa koncepcja w statystyce, która leży u podstaw zachowania rozkładów próbkowania. Stwierdza ono, że rozkład średniej z próby zbliża się do rozkładu normalnego wraz ze wzrostem wielkości próby, niezależnie od kształtu rozkładu populacji. Twierdzenie to umożliwia nam wyciąganie wiarygodnych wniosków na temat parametrów populacji na podstawie danych z próby.

Centralne twierdzenie graniczne służy jako kamień węgielny zrozumienia rozkładów normalnych w statystyce. W warunkach badawczych uzyskanie dokładnego oszacowania średniej populacji często wiąże się z gromadzeniem danych z wielu losowych próbek w populacji. Te indywidualne średnie z próbek wspólnie tworzą tak zwany rozkład próbkowania średniej.

Centralne twierdzenie graniczne określa dwie kluczowe zasady:

1. **Prawo wielkich liczb:** Wraz ze wzrostem wielkości próby lub liczby prób, średnia z próby ma tendencję do zbliżania się do średniej z populacji.
2. **Normalność rozkładu próbkowania:** Pomimo oryginalnego rozkładu zmiennej, podczas pracy z wieloma dużymi próbkami, rozkład próbkowania średniej ma tendencję do zbliżania się do rozkładu normalnego.

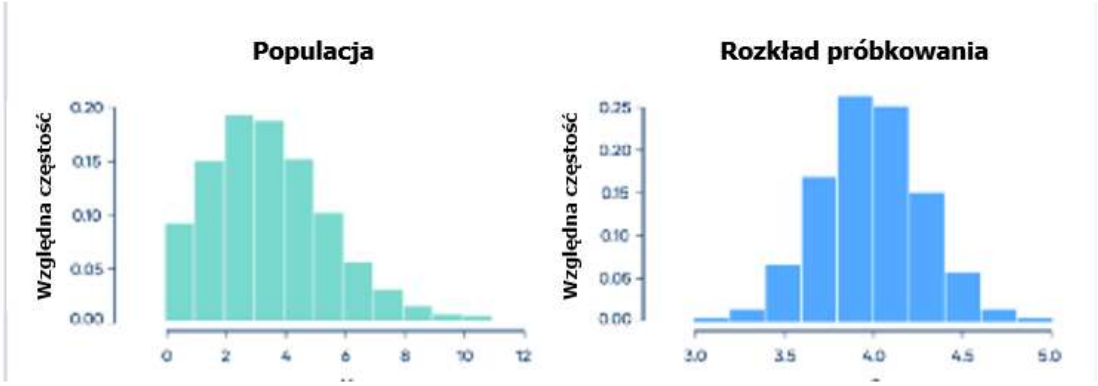
Parametryczne testy statystyczne konwencjonalnie zakładają, że próbki pochodzą z populacji o rozkładzie normalnym. Jednak centralne twierdzenie graniczne eliminuje konieczność tego założenia dla wystarczająco dużych prób. Przy dużych próbach testy parametryczne mogą być stosowane niezależnie od rozkładu populacji, pod warunkiem spełnienia innych istotnych

założeń. Wielkość próby wynosząca 30 lub więcej jest powszechnie uważana za wystarczająco dużą.

Z drugiej strony, w przypadku małych prób, zapewnienie założenia normalności ma kluczowe znaczenie ze względu na niepewność związaną z rozkładem próbkowania średniej. Dokładne wyniki wymagają potwierdzenia, że populacja jest zgodna z rozkładem normalnym przed zastosowaniem testów parametrycznych z małymi próbkami.

Przykładowo, centralne twierdzenie graniczne zakłada, że uzyskując wystarczająco duże próbki z populacji, średnie tych próbek będą wykazywać rozkład normalny, nawet jeśli podstawowy rozkład populacji odbiega od normalności.

Przykład: Rozważmy populację o rozkładzie Poissona (przedstawioną na lewym obrazku). Po pobraniu 10 000 próbek z tej populacji, z których każda składa się z 50 obserwacji, rozkład średnich próbek jest ściśle zgodny z rozkładem normalnym, zgodnie z centralnym twierdzeniem granicznym (jak pokazano na prawym obrazku).



Rysunek 2.8 Przykład populacji o rozkładzie Poissona i rozkładzie średnich z dużych próbek

Centralne twierdzenie graniczne opiera się na pojęciu rozkładu próbkowania, który reprezentuje rozkład prawdopodobieństwa statystyki obliczonej z wielu próbek pobranych z populacji.

Konceptualizacja eksperymentu może pomóc w zrozumieniu rozkładów próbkowania:

- Wyobraźmy sobie losowanie próbki z populacji i obliczanie statystyki, takiej jak średnia.
- Następnie losowana jest kolejna próba o identycznej wielkości, a średnia jest obliczana ponownie.

- Proces ten jest powtarzany wiele razy, w wyniku czego powstaje mnóstwo średnich, z których każda odpowiada próbce.

Agregacja tych średnich z próby stanowi przykład rozkładu próbkowania. Zgodnie z centralnym twierdzeniem granicznym, rozkład próbkowania średniej dąży do rozkładu normalnego, gdy wielkość próby jest wystarczająco duża. Co ciekawe, niezależnie od rozkładu populacji - normalnego, Poissona, dwumianowego lub innego - rozkład próbkowania średniej wykazuje normalność.

Na szczęście nie trzeba wielokrotnie próbkować populacji, aby określić kształt rozkładu próbkowania. Zamiast tego parametry rozkładu próbkowania średniej są zależne od parametrów samej populacji.

- Średnia rozkładu próbkowania jest średnią populacji.

$$\mu_{\bar{x}} = \mu$$

- Odchylenie standardowe rozkładu próbkowania to odchylenie standardowe populacji podzielone przez pierwiastek kwadratowy z wielkości próby.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Można opisać rozkład próbkowania średniej za pomocą tej notacji:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

gdzie:

- $\bar{X}$ to rozkład próbkowania średnich z próby,
- $\sim$ oznacza „podąża za rozkładem”,
- N jest rozkładem normalnym,
- μ jest średnią arytmetyczną populacji,
- σ jest odchyleniem standardowym populacji,
- n to wielkość próby.

Wielkość próby, oznaczana jako *n*, reprezentuje liczbę obserwacji pobranych z populacji dla każdej próby, zachowując jednorodność we wszystkich próbach. Wielkość próby znacząco wpływa na rozkład średniej w dwóch kluczowych aspektach.

1. Wielkość próby i normalność:
    - Większe rozmiary próbek zwykle dają rozkłady próbkowania, które ściśle przylegają do rozkładu normalnego.
    - I odwrotnie, przy małej liczebności próby rozkład próbkowania średniej może odbiegać od normalności. Ta rozbieżność powstaje, ponieważ ważność centralnego twierdzenia granicznego zależy od posiadania „wystarczająco dużej" wielkości próby.
    - Konwencjonalnie, próba o wielkości 30 lub więcej jest uważana za „wystarczająco dużą".
    - Gdy $n<30$, centralne twierdzenie graniczne nie ma zastosowania, a rozkład próbkowania odzwierciedla rozkład populacji. W związku z tym rozkład próbkowania jest normalny tylko wtedy, gdy rozkład populacji jest normalny.
    - I odwrotnie, gdy $n\geq30$, centralne twierdzenie graniczne jest prawdziwe, a rozkład próbkowania jest zbliżony do rozkładu normalnego.

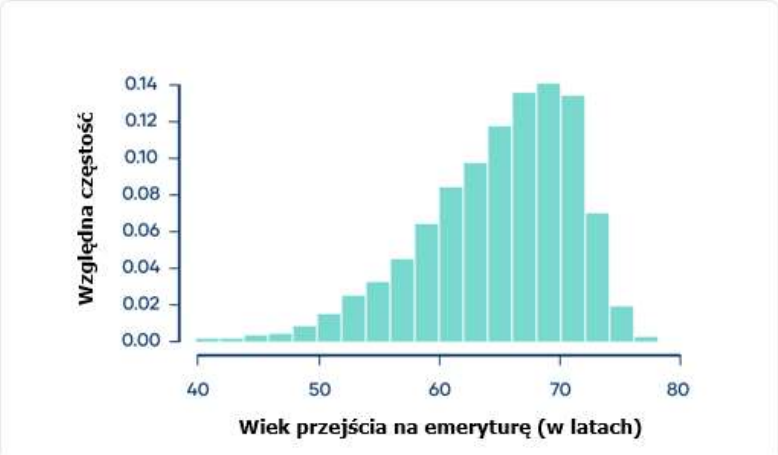
2. Wielkość próby i odchylenia standardowe:
    - Wielkość próby bezpośrednio wpływa na odchylenie standardowe rozkładu próbkowania, odzwierciedlając zmienność lub rozrzut rozkładu.
    - Przy mniejszej liczebności próby odchylenie standardowe jest zazwyczaj wyższe, co wskazuje na większą zmienność między średnimi z próby ze względu na ich niedokładne oszacowanie średniej populacji.
    - I odwrotnie, większe liczebności próby odpowiadają niższym odchyleniom standardowym, wskazując na mniejszą zmienność między średnimi z próby ze względu na ich dokładniejsze oszacowanie średniej populacji.

Znaczenie Centralnego Twierdzenia Granicznego:

Testy parametryczne, takie jak testy t, ANOVA i regresja liniowa, mają większą moc statystyczną w porównaniu z większością testów nieparametrycznych. Ta zwiększona moc statystyczna wynika z założeń dotyczących rozkładu populacji, które opierają się na centralnym twierdzeniu granicznym.

Rozkład ciągły

Rozważmy wiek emerytalny osób w Stanach Zjednoczonych. Populacja składa się ze wszystkich emerytowanych Amerykanów, a rozkład tej populacji można przedstawić w następujący sposób:



Rysunek 2.9 Wykres rozkładu ciągłego

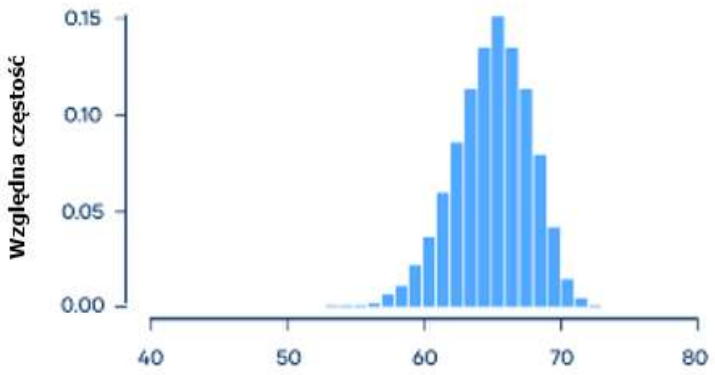
Rozkład wieku emerytalnego jest pochyla się w lewą stronę, a większość osób przechodzi na emeryturę w ciągu około pięciu lat od średniego wieku emerytalnego wynoszącego 65 lat. Istnieje jednak wydłużony ogon osób przechodzących na emeryturę znacznie wcześniej, na przykład w wieku 50 lub nawet 40 lat. Odchylenie standardowe w populacji wynosi 6 lat.

Wyobraźmy sobie, że przeprowadzamy próbę na małą skalę z tej populacji. Pięciu emerytów jest wybieranych losowo, a ich wiek emerytalny jest rejestrowany. Na przykład: 68, 73, 70, 62, 63.

Średnia z tej próby służy jako przybliżenie średniej populacji, aczkolwiek z ograniczoną precyzją ze względu na małą wielkość próby wynoszącą 5. Na przykład: Średnia = (68 + 73 + 70 + 62 + 63) / 5. Średnia = 67,2 lat.

Załóżmy teraz, że ten proces próbkowania jest powtarzany 10 razy, a każda próbka obejmuje pięciu emerytów. Obliczana jest średnia z każdej próby, co daje rozkład znany jako rozkład próbkowania średniej. Na przykład: 60,8; 57,8; 62,2; 68,6; 67,4; 67,8; 68,3; 65,6; 66,5; 62,1.

Ponieważ proces ten jest powtarzany wiele razy, histogram przedstawiający średnie tych próbek będzie zbliżony do rozkładu normalnego.



Rysunek 2.10 Normalny rozkład średnich

Pomimo tego, że rozkład próbkowania wykazuje nieco bardziej normalny kształt w porównaniu z populacją, nadal zachowuje lekkie odchylenie w lewo. Ponadto widać, że zmienność w rozkładzie próbkowania jest węższa niż w populacji.

Choć rozkład próbkowania wykazuje nieco bardziej normalny kształt w porównaniu z populacją, nadal zachowuje lekkie odchylenie w lewo. Ponadto widać, że zmienność w rozkładzie próbkowania jest węższa niż w populacji.

2.8 Statystyka testowa

Statystyka testowa reprezentuje wartość liczbową pochodzącą z testu hipotezy statystycznej, wskazującej stopień dopasowania obserwowanych danych do rozkładu oczekiwanego przy hipotezie zerowej tego testu.

Statystyka te odgrywa kluczową rolę w obliczaniu wartości *p* wyników, ułatwiając określenie, czy należy przyjąć, czy odrzucić hipotezę zerową?

Ale co dokładnie stanowi statystykę testową?

Statystyka testowa określa podobieństwo między rozkładem danych a rozkładem przewidywanym przy hipotezie zerowej zastosowanego testu statystycznego. Rozkład danych wyjaśnia częstotliwość każdej obserwacji, charakteryzując się tendencją centralną

i zmiennością wokół niej. Ponieważ różne testy statystyczne przewidują różne typy rozkładów, wybór odpowiedniego testu jest zgodny z postawioną hipotezą.

Statystyka testowa kondensuje obserwowane dane w pojedynczą liczbę, wykorzystując takie miary, jak tendencja centralna, zmienność, wielkość próby i liczba zmiennych predykcyjnych w modelu statystycznym.

Zazwyczaj statystykę testową wyłania się z dostrzegalnych wzorców w danych (np. korelacji między zmiennymi lub rozbieżności między grupami), podzielonych przez wariancję danych (tj. odchylenie standardowe).

Rozważmy następujący przykład:

Badany jest związek między temperaturą a datami kwitnienia określonego rodzaju jabłoni. Analizując kompleksowy zbiór danych obejmujący 25 lat, śledząc temperaturę i daty kwitnienia poprzez losowe pobieranie próbek 100 drzew rocznie z pola eksperymentalnego.

- Hipoteza zerowa (H_0): Nie istnieje korelacja między temperaturą a terminem kwitnienia.
- Hipoteza alternatywna (H_A lub H_1): Istnieje korelacja między temperaturą a terminem kwitnienia.

Aby sprawdzić tę hipotezę, należy przeprowadzić test regresji, uzyskując wartość *t* jako statystykę testową. Ta wartość *t* zestawia zaobserwowaną korelację między zmiennymi z hipotezą zerową o zerowej korelacji.

2.9 Rodzaje statystyk testowych

Poniżej, w tabeli 2.1, znajduje się podsumowanie najpopularniejszych statystyk testowych wraz z odpowiadającymi im hipotezami i kategoriami testów statystycznych, w których są one stosowane. Podczas gdy różne testy statystyczne mogą wykorzystywać różne metodologie do obliczania tych statystyk, podstawowe hipotezy i interpretacje statystyk testu pozostają spójne.

Tabela 2.1 Zestawienie najpopularniejszych statystyk testowych

Statystyka testowa	Hipotezy zerowe i alternatywne	Testy statystyczne, które ją wykorzystują
Wartość *t*	**Hipoteza zerowa:** Średnie dwóch grup są równe **Hipoteza alternatywna:** Średnie dwóch grup nie są równe.	• Test *T* • Testy regresji
Wartość *z*	**Hipoteza zerowa:** Średnie dwóch grup są równe. **Hipoteza alternatywna:** Średnie dwóch grup nie są równe.	• Test *Z*
Wartość *F*	**Hipoteza zerowa:** Zmienność między dwiema lub więcej grupami jest większa lub równa zmienności wewnątrz grup. **Hipoteza alternatywna:** Zmienność między dwiema lub więcej grupami jest mniejsza niż zmienność wewnątrz grup.	• ANOVA • ANCOVA • MANOVA
Wartość X^2	**Hipoteza zerowa:** Dwie próby są niezależne. **Hipoteza alternatywna:** Dwie próby nie są niezależne (tj. są skorelowane)	• Test chi-kwadrat • Nieparametryczne testy korelacji

W rzeczywistych scenariuszach zazwyczaj oblicza się statystykę testową za pomocą pakietu oprogramowania statystycznego, takiego jak R, SPSS lub Excel, który dostarczy również wartość *p* z nią powiązaną. Niemniej jednak, wzory do ręcznego obliczania tych statystyk można znaleźć w Internecie.

Na przykład, testując hipotezę dotyczącą temperatury i daty kwitnienia, przeprowadzasz analizę regresji. Test regresji daje następujące wyniki:

- współczynnik regresji 0,36,
- wartość t porównującą ten współczynnik z przewidywanym zakresem współczynników regresji przy hipotezie zerowej o braku związku.

Wynikowa wartość *t* z testu regresji 0,36, reprezentuje statystykę testową.

2.10 Błąd standardowy

Błąd standardowy średniej (SE lub SEM) służy jako wskaźnik prawdopodobnej rozbieżności między średnią populacji a średnią próby. Zapewnia wgląd w stopień zmienności, jaki można by przewidzieć w średniej próbie, gdyby badanie zostało powtórzone przy użyciu świeżych próbek pobranych z tej samej populacji.

Podczas gdy błąd standardowy średniej jest najczęściej wymienianą formą błędu standardowego, podobne miary istnieją dla innych parametrów statystycznych, takich jak mediany lub proporcje. Błąd standardowy funkcjonuje jako powszechny miernik błędu próbkowania, przedstawiając rozbieżność między parametrem populacji a statystyką próby.

Aby zmniejszyć błąd standardowy, zaleca się zwiększenie wielkości próby. Zatrudnienie dużej, randomizowanej próby służy jako najskuteczniejsza strategia minimalizowania stronniczości próbkowania i zwiększania wiarygodności ustaleń.

Błąd standardowy i **odchylenie standardowe** są miarami zmienności:

- **Odchylenie standardowe** opisuje zmienność **w obrębie pojedynczej próbki**.
- **Błąd standardowy** szacuje zmienność **w wielu próbkach** populacji.

Odchylenie standardowe służy jako statystyka opisowa pochodząca bezpośrednio z danych próbki, podczas gdy błąd standardowy reprezentuje statystykę wnioskowania, zwykle szacowaną, chyba że znany jest dokładny parametr populacji.

2.11 Wzór na błąd standardowy

Błąd standardowy średniej jest określany poprzez zastosowanie odchylenia standardowego wraz z wielkością próby. Ze wzoru wynika, że wielkość próby i błąd standardowy mają odwrotną zależność. Mówiąc prościej, wraz ze wzrostem wielkości próby, błąd standardowy maleje. Zjawisko to występuje, ponieważ większa próba zwykle daje statystykę próby bliższą parametrowi populacji.

Stosowane są różne wzory w zależności od tego, czy znane jest odchylenie standardowe populacji. Wzory te mają zastosowanie do próbek zawierających więcej niż 20 elementów (n > 20).

Gdy znane są parametry populacji

Gdy znane jest odchylenie standardowe populacji, można użyć go w poniższym wzorze, aby dokładnie obliczyć błąd standardowy.

Wzór	Objaśnienie
$SE = \frac{\sigma}{\sqrt{n}}$	• SE to błąd standardowy, • σ to odchylenie standardowe populacji, • n to liczba elementów w próbie.

Gdy parametry populacji są nieznane

Gdy odchylenie standardowe populacji jest nieznane, można użyć poniższego wzoru do oszacowania błędu standardowego. Wzór ten przyjmuje odchylenie standardowe z próby jako oszacowanie punktowe dla odchylenia standardowego populacji.

Wzór	Objaśnienie
$SE = \frac{s}{\sqrt{n}}$	• SE to błąd standardowy, • s to odchylenie standardowe próby, • n to liczba elementów w próbie.

Przykład: Korzystanie z formuły błędu standardowego. Aby oszacować błąd standardowy dla wyników SAT z matematyki, należy wykonać dwa kroki.
Najpierw należy znaleźć pierwiastek kwadratowy z wielkości próby (*n*).

Wzór	Objaśnienie
$n = 200$	$\sqrt{n} = \sqrt{200} = 14{,}1$

Następnie należy podzielić odchylenie standardowe próbki przez liczbę znalezioną w kroku pierwszym.

Wzór	Objaśnienie
$SE = \frac{s}{\sqrt{n}}$	$s = 180 \quad \sqrt{n} = 14{,}1 \quad \frac{s}{\sqrt{n}} = \frac{180}{14{,}1} = 12{,}8$

Błąd standardowy wyników SAT z matematyki wynosi 12,8.
Można przedstawić błąd standardowy obok średniej lub włączyć go do przedziału ufności, aby przekazać niepewność związaną ze średnią.

Przykład: Przedstawienie średniej i błędu standardowego. Średni wynik egzaminu SAT z matematyki dla losowej próby zdających wynosi 550 ± 12,8 (SE).

Podawanie błędu standardowego w przedziale ufności jest preferowane, ponieważ eliminuje potrzebę wykonywania przez czytelników dodatkowych obliczeń w celu uzyskania znaczącego zakresu.

Przedział ufności oznacza zakres wartości, w których nieznany parametr populacji będzie znajdował się najczęściej, jeśli badanie zostanie powtórzone z nowymi losowymi próbkami.

Przy poziomie ufności 95% oczekuje się, że 95% wszystkich średnich z próby będzie mieścić się w przedziale ufności obejmującym ± 1,96 błędu standardowego średniej z próby. Przedział ten służy jako oszacowanie, w którym prawdziwy parametr populacji znajduje się z 95% pewnością.

Przykład: Konstruujesz 95% przedział ufności (CI), aby oszacować średni wynik SAT z matematyki w populacji. Biorąc pod uwagę charakterystykę o rozkładzie normalnym, taką jak wyniki SAT, około 95% wszystkich średnich z próby mieści się w zakresie około 4 błędów standardowych średniej z próby.

Wzór na przedział ufności

$CI = \bar{x} \pm (1{,}96 \times SE)$

$\bar{x}$ = średnia z próby = 550
SE =błąd standardowy = 12,8

Dolna wartość graniczna	**Górna wartość graniczna**
$\bar{x} - (1{,}96 \times SE)$	$\bar{x} + (1{,}96 \times SE)$
$550 - (1{,}96 \times 12{,}8) =$ **525**	$550 + (1{,}96 \times 12{,}8) =$ **575**

W przypadku losowego doboru próby 95% przedział ufności (CI) [525 575] mówi, że istnieje prawdopodobieństwo 0,95, dla którego średni wynik SAT z matematyki w populacji wynosi od 525 do 575.

BIBLIOGRAFIA DO ROZDZIAŁU 2.

- Introductory Statistics. Bentham Science Publishers, Kahl, A. (Publish 2023). DOI:10.2174/9789815123135123010 1
- Introductory Statistics 2e, Openstax, Rice University, Houston, Texas 77005, Jun 23, senior contributing authors: Barbara Illowsky and Susan dean, De anza college, Publish Date: Dec 13, 2023, (https://openstax.org/details/books/introductory-statistics-2e);
- Introductory Statistics 4th Edition, Susan Dean and Barbara Illowsky, Adapted by Riyanti Boyd & Natalia Casper Published 2013 by OpenStax College) July 2021, (http://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf);
- Journal of the Royal Statistical Society 2024, A reputable journal publishing cutting-edge research and articles on various aspects of statistics, including theoretical advancements and practical applications. Recent issues have featured studies on sampling and hypothesis testing.
- Introductory Statistics 7th Edition, Prem S. Mann, eastern Connecticut state university with the help of Christopher Jay Lacke, Rowan university, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, 2011
- Introduction to statistics, made easy second edition, Prof. Dr. Hamid Al-Oklah Dr. Said Titi Mr. Tareq Alodat, March 2014
- Statistics for Business and Economics, Thirteenth Edition, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, 2017, 2015 Cengage Learning®

- Statistics for Business, First edition, Derek L Waller, 2008 Copyright © 2008, Derek L Waller, Published by Elsevier Inc. All rights reserved

Dodatkowe linki do literatury i filmów na Youtube do rozdziału 2

https://open.umn.edu/opentextbooks/textbooks/196

https://www.scribbr.com/category/statistics/

https://stats.libretexts.org/Bookshelves/Introductory_Statistics

https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf

https://saylordotorg.github.io/text_introductory-statistics/

https://drive.uqu.edu.sa/_/mskhayat/files/MySubjects/20178FS%20Elementary%20Statistics/Introductory%20Statistics%20(7th%20Ed).pdf

https://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf

https://www.geeksforgeeks.org/introduction-of-statistics-and-its-types/

https://onlinestatbook.com/Online_Statistics_Education.pdf

https://www.researchgate.net/profile/Tareq-Alodat-2/publication/340511098_INTRODUCTION_TO_STATISTICS_MADE_EASY/links/5e8de3dc4585150839c7b58a/INTRODUCTION-TO-STATISTICS-MADE-EASY.pdf

https://byjus.com/maths/statistics/

https://www.khanacademy.org/math/statistics-probability