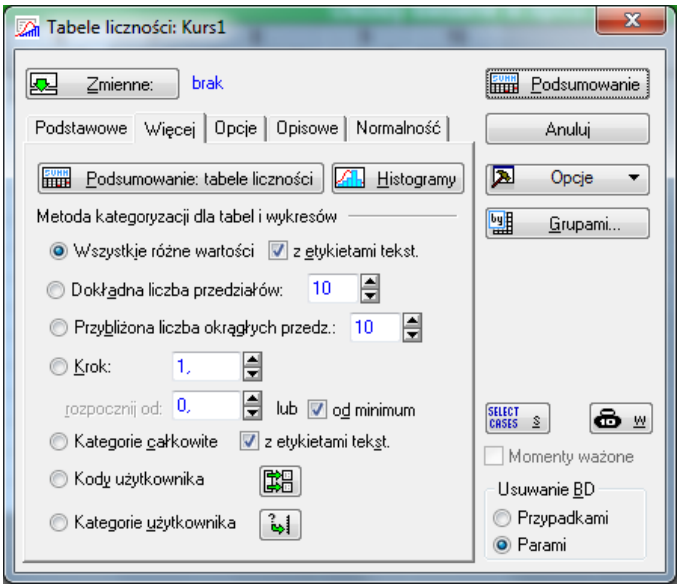
Grupowanie materiału statystycznego

Materiał liczbowy, otrzymany w wyniku przeprowadzonej obserwacji statystycznej lub pomiaru, należy odpowiednio usystematyzować i pogrupować. Doskonale nadają się do tego szeregi statystyczne i histogramy. Te ostatnie przemawiają do wyobraźni odbiorcy bardziej niż liczby. *STATISTICA* oferuje doskonałe narzędzia do tego typu opracowań. Moduł **Statystyki opisowe** daje nam już możliwość tworzenia szeregów statystycznych i histogramów. Jednak pełny zakres możliwości grupowania dostępny jest po wybraniu opcji **Tabele Liczności.** Obie opcje dostępne są w oknie **Statystyki Podstawowe i Tabele** wywołanej przy pomocy menu **Statystyka**. Warto tu zauważyć, że dla większość okien określania analizy dostępnych jest kilka kart zawierających opcje. Zazwyczaj dostępne są przynajmniej dwie grupy analiz. Pierwsza grupa znajdująca się na karcie **Podstawowe** zawiera najczęściej wykorzystywane opcje, umożliwiające szybkie określenie podstawowych analiz bez konieczności poszukiwania zbyt dużej liczby opcji.
Z kolei karta **Więcej** zawiera wszystkie opcje dostępne na karcie **Podstawowe** oraz wiele nieco rzadziej wykorzystywanych opcji. W niektórych bardziej złożonych analizach dostępne są również dodatkowe karty. Poniższy rysunek pokazuje opcje dostępne na karcie **Więcej** dla modułu **Tabele Liczności.**



Rys. 1. Opcje sposobu tabelaryzacji danych.

Tabele liczebności stanowią najprostsze i najczęściej używane narzędzie do wstępnej analizy danych ilościowych i jakościowych (danych w skali nominalnej). Umożliwiają one pogrupowanie danych według przyjętych kategorii dla ich uporządkowania i znalezienia interesujących różnic. Można też pogrupowane dane przedstawić graficznie w postaci histogramu. Tabele liczności informują o tym, jak często pojawiają się określone warianty analizowanej cechy w całym zbiorze danych. Kolejny przykład pokazuje jak utworzyć szereg rozdzielczy i powiązany z nim histogram dla wybranej zmiennej opisującej poziom sodu dla psów (Sód).

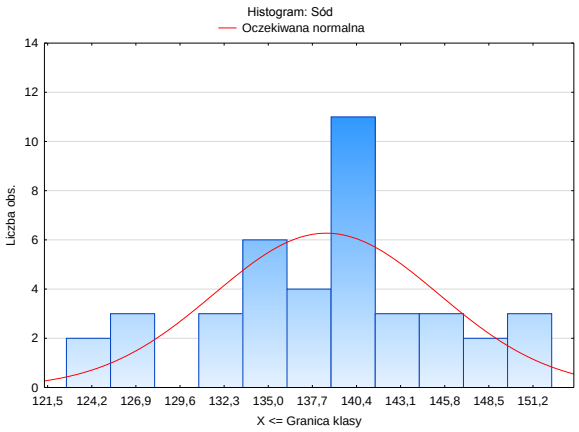
Przykład 1

1. Z menu **Statystyka** wybieramy opcję **Statystyki podstawowe i tabele**. Następnie w otwierającym się oknie wybieramy opcję **Tabele liczności.**
2. W oknie **Tabele liczności** klikamy przycisk **Zmienne** i wybieramy zmienną Sód. Następnie klikamy kartę **Więcej**, aby zobaczyć różne opcje dotyczące sposobu tabelaryzacji danych. Okno to zawiera wiele opcji służących do modyfikacji sposobu wyświetlania i kategoryzacji tabel częstości.

3. Dla potrzeb naszego przykładu wybieramy metodę kategoryzacji **Dokładna liczna przedziałów.** W takiej sytuacji rozstęp wartości każdej zmiennej zostanie podzielony na żądaną liczbę przedziałów. Przyjmujemy ustawienie 11 przedziałów i klikamy przycisk **Podsumowanie**: **Tabele liczności**, aby wyświetlić tabele liczności dla wybranej zmiennej. Otrzymana w ten sposób tabela liczności pokazana jest na poniższym rysunku.

Tabela liczności: Sód (Baza danych)

Od Do	Liczba	Skumulow. Liczba	Procent	Skumulow. Procent
122,6500<x<=125,3500	2	2	5,00000	5,0000
125,3500<x<=128,0500	3	5	7,50000	12,5000
128,0500<x<=130,7500	0	5	0,00000	12,5000
130,7500<x<=133,4500	3	8	7,50000	20,0000
133,4500<x<=136,1500	6	14	15,00000	35,0000
136,1500<x<=138,8500	4	18	10,00000	45,0000
138,8500<x<=141,5500	11	29	27,50000	72,5000
141,5500<x<=144,2500	3	32	7,50000	80,0000
144,2500<x<=146,9500	3	35	7,50000	87,5000
146,9500<x<=149,6500	2	37	5,00000	92,5000
149,6500<x<=152,3500	3	40	7,50000	100,0000
Braki	0	40	0,00000	100,0000

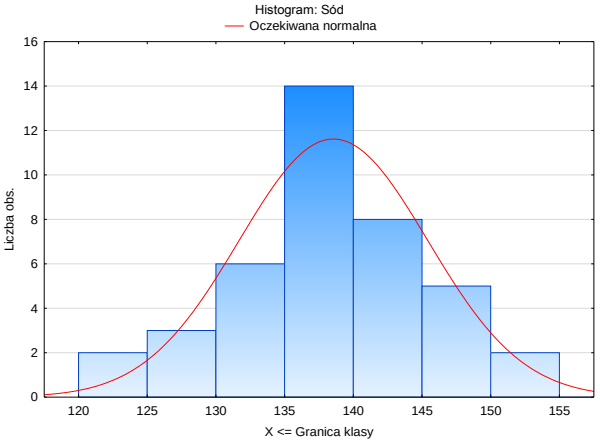


Rys. 2. Tabela liczności i histogram dla zmiennej Sód

4. Dla otrzymania powiązanego z tabelą histogramu wracamy do okna **Tabele liczności** i klikamy przycisk **Histogramy**. Otrzymany histogram pokazany jest na rysunku 2 po prawej stronie.
5. Prezentowana tabela zawiera niepotrzebnie przedziały o zerowej liczebności. Zmniejszamy więc liczbę przedziałów wybierając opcję **Przybliżona liczba okrągłych przedziałów.** Wówczas kategorie lub granice przedziałów klasowych i ich wielkości w tabelach liczności będą zaokrąglone. Można w ten sposób oczekiwać tablic łatwych do odczytania. Nowa tabela liczności oraz powiązany z nią histogram pokazuje rysunek 3.

Tabela liczności: Sód (Baza danych)

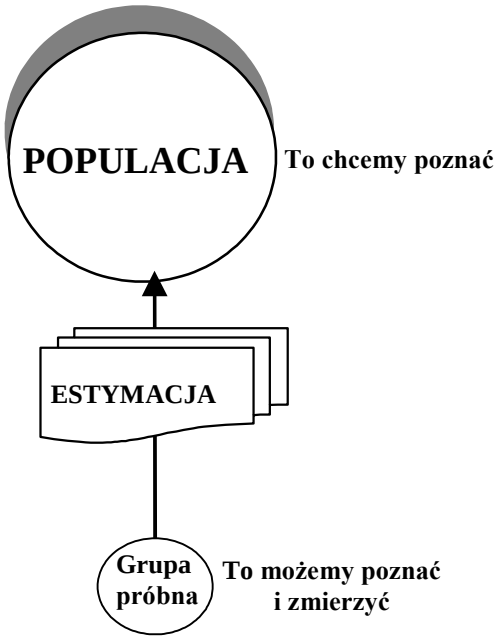
Od Do	Liczba	Skumulow. Liczba	Procent	**Skumul. Procent**
122,6500<x<=125,3500	2	2	5,00000	5,000
125,3500<x<=128,0500	3	5	7,50000	12,500
128,0500<x<=130,7500	0	5	0,00000	12,500
130,7500<x<=133,4500	3	8	7,50000	20,000
133,4500<x<=136,1500	6	14	15,00000	35,000
136,1500<x<=138,8500	4	18	10,00000	45,000
138,8500<x<=141,5500	11	29	27,50000	72,500
141,5500<x<=144,2500	3	32	7,50000	80,000
144,2500<x<=146,9500	3	35	7,50000	87,500
146,9500<x<=149,6500	2	37	5,00000	92,500
149,6500<x<=152,3500	3	40	7,50000	100,000
Braki	0	40	0,00000	100,000



Rys. 3. Druga wersja tabeli liczności i histogram dla zmiennej Sód

Wydaje się, że prezentowana tabela licznosci i histogram lepiej odzwierciedlają rozkład w badanej próbie. Rozkład liczebności informuje o liczbie jednostek w poszczególnych klasach natomiast procenty mówią o strukturze czyli o tym, jaką część całej zbiorowości stanowią jednostki przydzielone do poszczególnych klas. Prezentowany rozkład liczebności jest rozkładem empirycznym. Odzwierciedla rozkład wartości cechy wiek w badanej próbie. W praktyce intersuje nas rozkład badanej cechy w całej populacji. Stąd wykorzystanie rozkładów teoretycznych i testów zgodności badających dopasowanie rozkładu empirycznego do rozkładu teoretycznego.

Estymacja - przedziały ufności



Po wylosowaniu elementów do próby losowej i po ich obserwacji ze względu na interesujące nas cechy statystyczne, powstaje problem wnioskowania o populacji w oparciu o wyniki uzyskane z próby losowej. Na podstawie danych z próby możemy obliczyć średnią, medianę i odchylenie standardowe, ale tylko dla naszej próby. Otrzymane wnioski z tych wyników chcielibyśmy rozciągnąć na całą populację. Możliwość obliczenia średniej dla całej populacji przy pomocy średniej z próby to jest to, co jest nam potrzebne. Przyjrzyjmy się więc metodom wnioskowania statystycznego, które dotyczą sposobów oszacowań parametrów zmiennych losowych w całej populacji. Matematycy nazywają te metody **estymacją**. Podstawy teorii estymacji zostały sformułowane na przełomie XIX i XX wieku przez Karla Pearsona. Oczywiście estymacja może dotyczyć wyłącznie takich charakterystyk badanych cech, które przyjmują wartości liczbowe. Oszacowanymi parametrami są najczęściej średnia, frakcja, wariancja, współczynnik korelacji, ale estymować może też „obiekty" bardziej złożone jak linia regresji.

Na początek kilka słownikowych definicji

Estymacja, to proces, którego celem jest ocena nieznanej wartości parametru na podstawie obserwacji.

Estymator, to funkcja służąca do oceny nieznanej wartości parametru.

Wartość estymatora, to ocena wartości parametru wyliczona dla konkretnej próby.

Uwaga! Starajmy się dobrze rozróżniać estymator od wartości estymatora.

Punktem wyjściowym w estymacji jest wylosowanie z populacji n - elementowej próby i poznanie w niej interesującej nas zmiennej. **Estymacja** pozwala, w oparciu o wyniki z próby, wyznaczyć konkretną wartość (statystyka) będącą oszacowaniem nieznanego parametru populacji.

W zależności od sposobu, w jakim dokonujemy szacunku wartości parametrów estymację dzielimy na:

- **estymację punktową** - stosujemy ją, gdy nie znamy jednego lub kilku parametrów określających rozkład analizowanej zmiennej w populacji i chcemy ustalić ich wartości liczbowe na podstawie wyników próby, oczywiście przy zachowaniu odpowiednich reguł.
- **estymację przedziałowa** - tu dla oszacowania wyznaczamy pewien przedział liczbowy, który z pewnym prawdopodobieństwem zawiera wartość nieznanego parametru.

Podstawowym narzędziem szacowania nieznanego parametru jest estymator wyliczony na podstawie próby. Są to najczęściej statystyki tego samego typu, ale obliczone w próbie losowej. Przykładowo estymatorem wartości oczekiwanej jest średnia z próby losowej, a estymatorem wariancji dla całej populacji jest wariancja wyliczona na podstawie próby. Liczba możliwych estymatorów jest olbrzymia (ograniczona jedynie wyobraźnią statystyków), ale

użyteczne są jedynie te, które mają określone właściwości. Zaliczamy do nich przede wszystkim:

- **nieobciążoność**

Estymator nieobciążony to ten, którego przeciętna wartość jest dokładnie równa wartości szacowanego parametru. Innymi słowy przy wielokrotnym losowaniu próby średnia z wartości przyjmowanych przez estymator nieobciążony jest równa wartości szacowanego parametru. Obciążoność oznacza, że oszacowania dostarczone przez taki estymator są obarczone systematycznym błędem. Przykładowo średnia z próby jest nieobciążonym estymatorem średniej w całej populacji.

- **Efektywność**

Estymator jest tym efektywniejszy im mniejsza jest jego wariancja. Spośród dwóch estymatorów wybieramy ten, którego wariancja jest mniejsza.

- **Zgodność**

Zgodność oznacza, że jeśli rośnie liczebność próby, rośnie też prawdopodobieństwo, że oszacowanie przy pomocy estymatora będzie przyjmować wartości coraz bliższe wartości szacowanego parametru. Inaczej: zwiększając liczebność próby, zmniejszamy ryzyko popełnienia błędu większego niż pewna ustalona wielkość.

Estymatory o wszystkich tych własnościach są najbardziej użyteczne, zapewniają one otrzymanie wyników z próby zbliżonych do rzeczywistości. Jednak nawet bardzo wyrafinowane estymatory nie zapewniają oszacowania precyzji i wiarygodności uzyskanych wyników. Dlatego bardziej popularne są **przedziały ufności** pozbawione tych wad. Ich podstawy opracował w 1933 roku polski statystyk J. Spława-Neyman. Przedział ufności wyliczamy dla oszacowania wartości pewnej charakterystyki populacji na podstawie próby. Wartość tej charakterystyki dla próby będzie się nieco różnić od charakterystyki dla całej populacji. Wynika stąd, że dla różnych prób otrzymamy najczęściej różne wartości tej charakterystyki. Gdy próba jest reprezentatywna możemy oczekiwać niezbyt dużej różnicy między rzeczywistą wartością charakterystyki populacji a wyznaczoną przez nas wartością z próby. Przedział ufności określa nam prawdopodobny zasięg odchylenia naszych wyliczeń od wartości rzeczywistej. Wyznaczenie tego przedziału jest skomplikowane i wymaga zastosowania specjalnych wzorów, których postać zależy od liczebności próby oraz od pewnych założeń dotyczących rozkładu (najczęściej normalności) badanej cechy. Znajomość rozkładu to jak znajomość planu miasta, który pozwala zlokalizować każdy adres. Na pomoc przychodzi nam technika komputerowa. Większość bowiem programów statystycznych wylicza je precyzyjnie i bez problemu. Interpretacja przedziału ufności jest oczywista: im mniejszy przedział ufności, tym dokładniej obliczony przez nas estymator przybliża wartość rzeczywistą dla całej populacji. Odwrotnie szeroki przedział ufności oznacza możliwość dużych odchyleń wartości z próby od wartości z populacji - czyli małą wiarygodność naszych wyników.

Przykładowe okno z wyliczonym w pakiecie *STATISTICA* przedziałem ufności przeciętnej masy ciała przedstawione jest poniżej.

	Statystyki opisowe (Baza da		
Zmienna	**Średnia**	Ufność -95,000%	Ufność 95,000%
Waga	24,78250	18,08206	31,48294

Jak widać z każdym przedziałem związana jest liczba (oznaczana przez 1 - α) zwana **poziomem ufności**. Oznacza ona, że w średnio $\alpha \cdot 100\%$ przypadków jest źle tzn. otrzymamy przedziały niepokrywające estymowany parametr. Przykładowo przyjmijmy poziom ufności 0,95. Wówczas pobierając z populacji 100 prób i wyznaczając na ich podstawie przedziały ufności, to co najwyżej 5 przedziałów spośród 100 nie zawiera estymowanego parametru. Oczywiście w zastosowaniach praktycznych pobieramy tylko jedną próbę i wyznaczamy tylko jeden przedział ufności. W naszym konkretnym przypadku nie będziemy pewni, czy przedział zawiera wartość estymowanego parametru. Będziemy jednak „ufali", że tak jest o ile prawdopodobieństwo 1 - α jest dostatecznie duże. Powszechnie przyjmuje się wartość 1 - α = 0,95 jako tą najmniejszą. Musielibyśmy mieć wielkiego pecha (prawdopodobieństwo tego jest równe 0,05 lub mniejsze), aby nasz wyliczony z próby przedział ufności nie zawierał estymowanego parametru. Przyjmując z kolei poziom ufności 99% możemy się mylić raz na 100 razy. Aby mieć „pewność" możemy podnieść poziom ufności do 99,9%.
Przy interpretacji przedziałów ufności nie mówimy o prawdopodobieństwie, że nieznana wartość parametru P będzie zawarta w jakimś stałym przedziale. Przecież P **nie jest zmienną losową.**

Wydawać by się mogło, że przyjęcie wysokiego współczynnika ufności rozwiąże wszystkie nasze problemy. Zapewnimy sobie dowolnie dużą ufność wyliczonego przedziału. Niestety tak nie jest. Zwiększenie współczynnika ufności powoduje zwiększenie szerokości przedziału ufności, czyli zmniejszenie precyzji estymacji. Prowadzi to statystycznego paradoksu, że im chcemy być bardziej ufni, to jesteśmy mniej precyzyjni i odwrotnie. Poprawa precyzji jest możliwa pod warunkiem zwiększenia liczebności próby (istnieją na to specjalne wzory), a to w naukach medycznych nie zawsze jest możliwe. Taka sytuacja powoduje także zwiększenie kosztów eksperymentu. Musimy więc starać się wybrać złoty środek. A z tym wiadomo najtrudniej.

Reasumując estymacja pozwala nam przy ustalonym z góry prawdopodobieństwie (zwanym poziomem ufności) utworzyć przedział zawierający nieznaną wartość parametru populacji. Przedział ten nazywamy przedziałem ufności.

Starajmy się dla lepszej prezentacji wyników badań klinicznych podawać przedziały ufności. Granice przedziałów ufności prowadzą bowiem do lepszego zrozumienia zjawisk, a ich szerokość jest doskonałą wskazówką dokładności oszacowania badanych parametrów (czasów przeżycia, współczynników umieralności, metody leczenia itd.).