Łagodne wprowadzenie do statystyki

Podręcznik dla studentów wydziałów nauk o zdrowiu

true

Spis treści

Rozdział 1

Przedmiot i metody badań statystycznych

1.1 Przedmiot statystyki

Wyraz statystyka ma wiele znaczeń: **statystyki zgonów** albo **statystyki alkoholizmu** czyli **dane** dotyczące zgonów lub alkoholizmu. Statystyka to też **dziedzina wiedzy**, upraszczając zbiór metod, które służą do tworzenia statystyk w pierwszym znaczeniu tego słowa. Wreszcie statystyka to **pojedyncza metoda** ze zbioru metod opracowanych w dziedzinie, np. średnia to statystyka. Trochę to niefortunne, ale świat nie jest doskonały jak wiemy…

Statystyka (obiegowo): dział matematyki, a w związku z tym wiedza absolutnie pewna i obiektywna. Nieprawda choćby z tego powodu, że nie jest działem matematyki. Korzysta z metod matematycznych jak wiele innych dziedzin.

Statystyka od strony czysto praktycznej to: **dane** + **procedury** (zbierania, przechowania, analizowania, prezentowania *danych*) + **programy**; Jeżeli statystyka kojarzy się komuś ze matematyką, wzorami i liczeniem, to jak widać jest to zaledwie podpunkt procedury→analizowanie.

1.2 Podstawowe pojęcia

Celem **badania statystycznego** jest uzyskanie informacji o interesującym zjawisku na podstawie danych. Zjawisko ma charakter masowy czyli dotyczy dużej liczby *obiektów*. Nie interesuje nas jeden zgon (obiekt) tylko zgony wielu ludzi.

Populacja (**zbiorowość statystyczna**) to zbiór obiektów będący przedmiotem badania statystycznego. Na przykład zgony w Polsce w roku 2022.

Każdy **obiekt** w populacji to **obserwacja** (zwana także **jednostką statystyczną** albo **pomiarem**) na jednej lub więcej **zmiennych**. Jeżeli interesującym zjawiskiem są zgony, obserwacją jest osoba zmarła a zmiennymi wiek, płeć, przyczyna zgonu oraz dzień tygodnia (w którym nastąpił zgon) zmarłej osoby.

Próba to część **populacji.** Na przykład część zgonów w Polsce w roku 2022.

Parametr: wielkość numeryczna obliczona na podstawie populacji.

Statystyka: wielkość numeryczna obliczona na podstawie próby.

Populacja powinna być zdefiniowana w taki sposób, aby nie było wątpliwości co tak naprawdę jest badane. Zgony to w oczywisty sposób za mało. *Zgony mieszkańców Kwidzyna w roku 2022*.

Zwróćmy uwagę, że *Zgony w mieście Kwidzyn w roku 2022* to nie to samo (ktoś może być mieszkańcem a umrzeć w Polsce i/lub ktoś może nie być mieszkańcem i umrzeć w Kwidzynie.)

Generalizacja: ocena całości na podstawie części. Badamy zjawisko wypalenia zawodowego pielęgniarek i pielęgniarzy w Polsce (populacja). Wobec zaporowych kosztów mierzenia wszystkich decydujemy się na przeprowadzenie ankiety wśród studentów pielęgniarstwa PSW (próba). Czy możemy twierdzić na podstawie próby, że wyniki badania dla całej Polski są identyczne? Raczej nie…

Próba, która pozwala na generalizację nazywa się próbą **reprezentatywną**. Najlepszym sposobem na uzyskanie próby reprezentatywnej jest losowanie.

W oczywisty sposób badanie na podstawie próby jest tańsze niż badanie całości, co nie oznacza że jest tanie. Kontynuując przykład: musielibyśmy mieć listę wszystkich pielęgniarek i pielęgniarzy w Polsce. Z tej listy wylosować próbę a następnie skontaktować się z wybranymi osobami (jak?). Dlatego też badania w oparciu o próbę nielosową są całkiem popularne (bo są tanie); należy jednakże mieć świadomość ich ograniczeń, w tym a zwłaszcza uogólnienia uzyskanych wyników.

Mądrość statystyczna nt liczebności próby i wnioskowania z próby niereprezentatywnej: badano czy nowy preparat podnosi nośność kur, w 33,3% przypadków podniósł w 33,3% przypadków nie podniósł, a na 33,3 nie wiadomo, bo kura uciekła.

1.3 Pomiar

Potocznie kojarzy się z linijką i wagą ale w statystyce używany jest w szerszym znaczeniu. Ustalenie płci albo przyczyny zgonu to też pomiar.

Pomiar to przyporządkowanie wariantom **zmiennej** liczb lub symboli z pewnej **skali pomiarowej**. Przykładowo jeżeli jednostką statystyczną jest zgon a zmiennymi wiek, płeć, przyczyna zgonu oraz dzień tygodnia to pomiar będzie polegał na ustaleniu (przyporządkowaniu) wieku w latach, płci ('K'/'M'), przyczyny (identyfikatora z katalogu ICD10 zapewne) oraz numeru dnia tygodnia (lub nazwy dnia tygodnia). Wiek oraz numer dnia są liczbami, płeć i przyczyna symbolem.

Wyróżnia się następujące **typy skal pomiarowych**:

- **nominalna** (*nominal scale*), klasyfikuje: płeć zmarłego;
- **porządkowa** (*ordinal scale*), klasyfikuje i porządkuje: dzień tygodnia w którym nastąpił zgon (po poniedziałku jest wtorek);
- **liczbowa**, mierzy w potocznym tego słowa znaczeniu: wiek zmarłego w latach

Mówimy **zmienna mierzalna** albo **zmienna ilościowa** dla zmiennych mierzonych za pomocą skali liczbowej. Mówimy **zmienna niemierzalna** albo **zmienna jakościowa** dla zmiennych mierzonych za pomocą skali nominalnej/porządkowej.

Zmienne mierzalne dzielą się na **skokowe** oraz **ciągłe**. Skokowe są to cechy, które przyjmują skończoną liczbę wartości, zwykle są to liczby całkowite; Ciągłe są to cechy, które przyjmują dowolne wartości liczbowe z pewnego przedziału liczbowego, np. ciśnienie krwi.

Rodzaje danych

- Przekrojowe (zmarli w Kwidzynie)
- Czasowe: każda obserwacja ma przypisany czas (liczba zmarłych w Polsce w latach 2000--20222)

- Przestrzenne : każda obserwacja ma przypisane miejsce na kuli ziemskiej (współrzędne geograficzne)

1.4 Rodzaje i sposoby analizy danych

Rodzaje **analizy statystycznej** zależą od rodzaju danych (jakie mamy dane takie możemy stosować metody):

- jedna zmienna/dane przekrojowe: analiza struktury
- jedna zmienna/dane czasowe: analiza dynamiki zjawiska
- co najmniej dwie zmienne: analiza współzależności (nadwaga powoduje cukrzycę)

Sposoby analizy danych zależą od sposobu pomiaru (populacja/próba/generalizacja):

Opis statystyczny -- (proste) przedstawienie badanych zbiorowości/zmiennych tabel, wykresów lub parametrów (np. średnia, mediana) ; Opis statystyczny może dotyczyć: -- struktury zbiorowości; -- współzależności; -- zmian zjawiska w czasie.

Wnioskowanie statystyczne: wnioskowanie na temat całości na podstawie próby; wykorzystuje metody analizy matematycznej

Opisujemy populację lub próbę. **Wnioskujemy** na podstawie próby o całości…

1.5 Sposoby pomiaru danych i organizacja badania

Sposób pomiaru/organizacja badania ma zasadnicze znaczenie dla interpretacji wyników. Są dwa fundamentalne rodzaje pomiaru (sposobu zebrania danych) **eksperyment** oraz **obserwacja**.

Mówimy w związku z tym **dane eksperymentalne** albo **dane obserwacyjne**.

Picie kawy a lepsza ocena
Chcemy ustalić czy spożywanie kawy w czasie sesji egzaminacyjnej skutkuje uzyskaniem lepszej oceny. W celu oceny prawdziwości takiej tezy przeprowadzono badanie wśród studentów pytając ich o to ile kawy pili w czasie sesji i zestawiając te dane z wynikami egzaminów. Średnie wyniki w grupie studentów pijących dużo kawy były wyższe w grupie pijącej mało kawy. Czy można powiedzieć, że udowodniono iż picie dużej ilości kawy poprawia wynik egzaminu?
Raczej nie: można sobie wyobrazić, że studenci którzy poświęcili więcej czasu na naukę pili w tym czasie kawę (na przykład żeby nie zasnąć). Prawdziwą przyczyną jest czas poświęcony na przygotowanie a nie to ile ktoś wypił lub nie wypił kawy. Inaczej mówiąc gdyby ktoś pił dużo kawy, bo uwierzył, że to poprawi mu wyniki i się nie uczył, to pewnie by się rozczarował.

Rodzaje badań: **eksperymentalne** vs **obserwacyjne**.

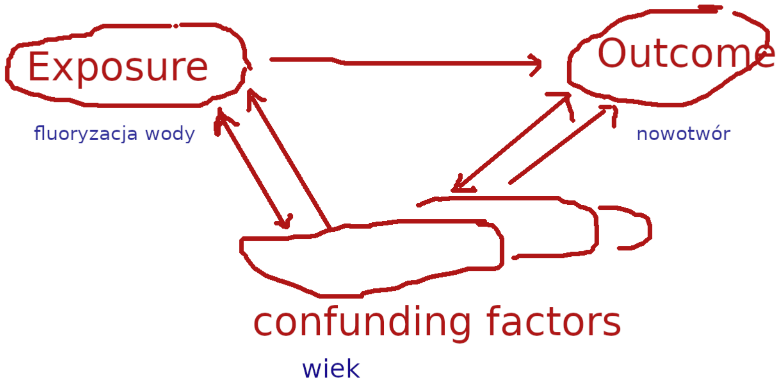
Eksperyment kontrolowany (zrandomizowany lub nie): służy do weryfikacja związku **przyczyna-skutek**. Skutek może być rezultatem działania wielu **czynników** (zmiennych). Eksperymentator manipuluje wielkością przyczyn (zmiennych **niezależnych**) oraz mierzy wielkość skutku (zmiennej **zależnej**); Wszystkie pozostałe czynniki (zmienne **ukryte**) są **kontrolowane** (w tym sensie, że ich wpływ na skutek jest ustalony.

Pomiarowi/manipulacji podlega zbiór jednostek podzielonych **losowo** na dwie grupy: grupa **eksperymentalna** (**experimental group**) oraz **grupa kontrolna** (**control group**)

W medycynie używa się terminu **badania kliniczne** czyli badania które dotyczą ludzi. Badania kliniczne także dzielą się na eksperymentalne oraz obserwacyjne. Eksperyment nazywa się RCT (*randomized clinical trial*/randomizowane kontrolowane badania kliniczne.) Manipulacja określana jest jako ekspozycja (**exposure**) albo leczenie (**treatment**) Zmienne ukryte określa się mianem **confunding factors** (**czynniki zakłócające**)

Rysunek 1.1 przedstawia zależność pomiędzy wynikiem (*outcome*), przyczyną oraz czynnikami zakłócającymi na przykładzie zależności dotyczącej domniemanego wpływu fluoryzowania wody na zwiększenie ryzyka zgonów z powodu nowotworów. W badaniu którego autor uważał, że udowodnił związek fluoryzowanie→nowotór porównał on współczynniki zgonów z miast fluoryzujących oraz nie fluoryzujących wodę. Okazało się, że przeciętnie współczynnik ten był wyższy w grupie miast fluoryzujących wodę. Czy to świadczy, że fluoryzowanie wody powoduje raka? Nie...

W innym badaniu tych samych miast okazało się, że w grupie miast fluoryzujących wodę przeciętnie mieszkają starsi ludzie. A ponieważ współczynniki zgonów rosną wraz ze wzrostem wieku, to nie można wykluczyć, że prawdziwą przyczyną obserwowanego zwiększenia wartości współczynników zgonów jest wiek a nie fluoryzacja wody.



Rysunek 1.1: Fluoryzacja, wiek a nowotwór

Efekt przyczynowy to **ilościowe** określenie wpływu ekspozycji na wynik poprzez porównanie wielkości wyniku dla różnych wielkości ekspozycji

Są dwa typy **efektu przyczynowego**: indywidualny efekt interwencji (*individual treatment effect*) oraz średni efekt interwencji (*average treatment effect*)

Individual Treatment Effect (ITE)

Indywidualny efekt interwencji (ITE) określa ilościowo wpływ interwencji dla konkretnej osoby, poprzez porównanie wyników dla różnych wartości interwencji.

Mogę pić kawę lub nie pić kawy a wynikiem będzie ocena. Oczywiście nie mogą zrobić tych dwóch rzeczy na raz...

Average Treatment Effect (ATE)

Średni efekt interwencji określa ilościowo wpływ interwencji dla grupy osób

W grupie studentów jedni pili kawę inni nie...

Jeżeli grupa (populacja) została uprzednio podzielona (losowo) na grupę **eksperymentalną** oraz **grupę kontrolną** możemy policzyć ATE oddzielnie dla obu grup. Wtedy efekt przyczynowy można zdefiniować jako:

ATT - ATC (albo ATT/ATC)

gdzie: ATT oznacza ATE w grupie eksperymentalnej a ATC oznacza ATE w grupie kontrolnej.

Kawa a ocena (kontynuuacja)
Można przypuszczać, że oprócz kawy na wynik egzaminu ma wpływ np. wrodzone predyspozycje w dziedzinie intelektualnej oraz czas poświęcony na naukę. Aby kontrolować ten czynnik można podzielić losowo grupę studentów; dzięki czemu średnia wielkość predyspozycji oraz czasu w obu grupach będzie podobna. Następnie zalecamy studentom w **grupie eksperymentalnej** picie 1l kawy dziennie a studentom w **grupie kontrolnej** podajemy 1l brązowej wody o smaku i zapachu kawy :-). Średnie wyniki w grupie studentów pijących 1l kawy okazały się wyższe niż w grupie pijącej kolorową wodę. Czy można powiedzieć że udowodniono iż picie dużej ilości kawy poprawia wynik egzaminu? Raczej tak...

Badania obserwacyjne można z kolei podzielić na **analityczne** i **opisowe**.

W badaniach **analitycznych** porównuje się grupę kontrolną z grupą poddaną ekspozycji/leczeniu; w badaniach przekrojowych nie ma grupy kontrolnej.

Badania analityczne dzielimy dalej na:

- **kohortowe**,
- **kliniczno-kontrolne**,
- **przekrojowe**.

Badanie **kohortowe** (**cohort study**): wieloletnie badania na dużej grupie jednostek. Pomiar zaczynamy od ekspozycji kończymy na wyniku/chorobie/zgonie (takie badanie nazywamy **prospektywnym**. Problem: koszty (np. choroby rzadkie wymagają ogromnych kohort).

Badanie **kliniczno-kontrolne** (**case-control study**): **restrospektywna** ocena ekspozycji dla jednostek, u których stwierdzono wynik (chorobę). Grupę kontrolną tworzą **dopasowane** jednostki u których wyniku nie stwierdzono (dopasowane w sensie, że są podobne podobne.) W praktyce badanie kliniczno-kontrolne to badanie chorych, którzy zgłosili się do przychodni; grupą kontrolną są podobni chorzy (wiek, płeć) z innej przychodni :-)

Problem1: błąd pamięci (**recall bias**) pacjenci -- zwłaszcza zdrowi -- słabo pamiętają fakty które miały miejsce lata temu. Problem2: trudności z **dopasowaniem** grupy kontrolnej (łatwiej powiedzieć niż zrobić.)

Badania **prospektywne**: od przyczyny do skutku (cohort); badanie **retrospektywne**: od skutku do przyczyny (case-control)

Badanie przekrojowe (**cross-sectional study**): badanie związku między wynikiem a ekspozycją bez podziału na grupę eksperymentalną i kontrolną.

Problem: nie da się określić związku przyczyna-skutek w taki sposób jaki się stosuje w badaniach analitycznych, ale można do tego celu zastosować **model** regresji liniowej.

Palenie a nowotwór
Badamy grupę pacjentów przychodni onkologicznej. Stwierdzamy że 90% z nich paliło papierosy. Czy z tego wynika że palenie powoduje raka? Niekoniecznie. Możemy **dopasować** pacjentów o podobnym profilu demograficzno-społecznym z innej przychodni (którzy nie chorują na raka) i stwierdzić że 20% z nich paliło. To już jest konkretny argument -- ale takie badanie nie jest już **przekrojowe** tylko **kliniczno-kontrolne**.

Kawa a ocena (kontynuuacja)
można oprócz pytania studentów o ilość kawy i wynik pytać ich jeszcze o czas poświęcony na naukę oraz o średnią ze studiów (wrodzone predyspozycje w dziedzinie intelektualnej). Za pomocą metody regresji możemy ustalić czy i jak bardzo kawa, czas i predyspozycje wpływają na ocenę. Teoretycznie zamiast eksperymentu można używać regresji, ale jest to w większości przypadków trudne--albo zmienne nie da się zmierzyć (czy średnia ze studiów jest dobrą miarą predyspozycji?) albo jakąś ważną zmienną pominiemy. Więcej na temat regresji w rozdziale 4.

Badanie **ekologiczne**: badanie (przekrojowe) zależności pomiędzy danymi **zagregowanymi** a nie indywidualnymi. Przykładowo zależność pomiędzy przeciętną wielkością dochodu narodowego, a przeciętną oczekiwaną długością życia np. na poziomie kraju.

Problem: błąd ekologizmu (**ecological fallacy**.) Zależności na poziomie indywidualnym oraz zagregowanym mogą być różne. Można oczekiwać że im większy dochód tym osoba dłużej żyje (poziom indywidualny.) Jeżeli w kraju występują duże różnice w dochodach (na przykład USA) to przeciętnie dochód jest wysoki, ale jest dużo osób o niskich dochodach, o ograniczonym dostępie do służby zdrowia, i krótszej oczekiwanej długości życia. Przeciętna oczekiwana długość życia na poziomie całego kraju jest niższa (bo jest sumą wysokiej dla bogatych + niskiej dla biednych); w rezultacie zależność na poziomie zagregowanym może się znacząco różnić od tej na poziomie indywidualnym.

1.5.1 Przykłady badań

Jest ustalony szablon artykułu naukowego, który powinien być podzielony na następujące części:

1. **Wprowadzenie**: określenie problemu badawczego, celu badania;
2. **Materiał i metoda**: Opis danych i zastosowanych metod statystycznych
3. **Wyniki**: Rezultaty analiz
4. **Dyskusja**: Znaczenie uzyskanych wyników, jeżeli we wstępie postawiono hipotezy to tutaj należy

Żeby się zorientować jakie dane (jakie zmienne i jak mierzone) oraz jakie metody statystyczne zostały wykorzystane w pracy wystarczy zapoznać się z treścią punktu **materiał i metoda**. W szczególności powinien tam być określony **rodzaj badania**: eksperyment, badanie kohortowe, kliniczno-kontrolne, przekrojowe lub inne…

Czy konsumpcja soli kuchennej szkodzi? (eksperyment)
Neal B. i inni zastosowali eksperyment kontrolowany do zbadania wpływu substytucji chlorku sodu chlorkiem potasu na choroby sercowo-naczyniowe (*Effect of Salt Substitution on Cardiovascular Events and Death, New England Journal of Medicine*, https://doi.org/10.1056/NEJMoa2105675). W badaniu przeprowadzonym w Chinach, uczestniczyli mieszkańcy 600 wsi, podzieleni losowo na dwie grupy. Uczestnik badania musiał mieć minimum 60 lat oraz nadciśnienie krwi. W badaniu uczestniczyło prawie 21 tysięcy osób. Przez pięć lat trwania eksperymentu grupa kontrolna używała soli zawierającej 75% chlorku potasu oraz 25% chlorku sodu; grupa badana zaś używała soli tradycyjnej czyli zawierającej wyłącznie chlorek sodu. Obserwowano w okresie pięcioletnim w obu grupach liczbę udarów, incydentów sercowo-naczyniowych oraz zgonów. Wpływ substytucji oceniono porównując współczynniki ryzyka w obu grupach.

Konflikt praca-dom w zawodzie pielęgniarki (przekrojowe)
Simon i inni badali konflikt Praca-Dom w zawodzie Pielęgniarki/Pielęgniarza (Work-Home Conflict in the European Nursing Profession Michael Simon 1, Angelika Kümmerling, Hans-Martin Hasselhorn; Next-Study Group Int J Occup Environ Health 2004 Oct-Dec;10(4):384-91.

doi: 10.1179/oeh.2004.10.4.384. https://pubmed.ncbi.nlm.nih.gov/15702752/)
Konflikt Praca-Dom (WHC) to sytuacja kiedy nie można zająć się zadaniami lub obowiązkami w jednej dziedzinie ze względu na obowiązki w drugiej domenie. Teoria zapożyczona z obszaru Nauk o Zarządzaniu zapewne. Ten konflikt jest mierzony odpowiednią skalą pomiarową składającą się z pięciu pytań. Czynnikami które WHC mają powodować są: czas pracy, grafik (w sensie rodzaj etatu/zmianowość), nacisk-na-nadgodziny (występuje lub nie), intensywność pracy, obciążenie emocjonalne oraz jakość zarządzania. (ostatnie trzy mierzone odpowiednimi skalami pomiarowymi, czytaj: serią pytań w ankiecie). Badano 27,603 osoby. Podstawowym narzędziem badawczym jak się łatwo domyśleć była ankieta, a przyczyny WHC ustalono za pomocą metody regresji wielorakiej.
Teraz porówajmy koszty badania #1, w którym jedynie starano się ustalić że sól szkodzi (lub nie) z badaniem #2, w którym starano się ustalić przyczyny stanów psychicznych badanych-:)

1.6 Miary częstości chorób

Populacja narażona (*population at risk*): grupa osób podatnych na zdarzenie (chorobę); rak szyjki macicy dotyczy kobiet a nie wszystkich.

Współczynnik chorobowości (*prevalence rate*): liczba chorych w określonym czasie (dzień, tydzień, rok) podzielona przez wielkość populacji narażonej. Ponieważ są to zwykle bardzo małe liczby, mnoży się wynik przez 10^n dla ułatwienia interpretacji. Czyli jeżeli chorych w populacji narażonej o wielkości 1mln jest 20 osób, to współczynnik wynosi 20/1mln = 0,000002 co trudno skomentować po polsku. Jeżeli pomnożymy owe 0,000002 przez 100 tys ($n = 5$), to współczynnik będzie równy 2, co interpretujemy jako dwa przypadki na 100 tys. (albo 0,2 na 10 tys, jeżeli $n = 4$, co już jednak brzmi trochę gorzej.)

Współczynnik zapadalności (*incidence rate*): liczba nowych chorych w określonym czasie (dzień, tydzień, rok) podzielona przez wielkość populacji narażonej. Też zwykle pomnożona przez 10^n

Współczynnik śmiertelności (*case fatality rate*): liczba zgonów z powodu X w określonym czasie (dzień, tydzień, rok) podzielona przez liczbę chorych na X w tym samym czasie. Śmiertelność jest miarą ciężkości choroby X.

Współczynnik zgonów (*death rate*): liczba zgonów w określonym czasie przez średnią liczbę ludności w tym czasie (pomnożone przez 10^n).

Jeżeli współczynnik zgonów nie uwzględnia wieku, nazywany jest surowym (*crude*); grupy różniące się strukturą wieku nie powinny być porównywane za pomocą współczynników surowych tylko standaryzowanych (*age-standardized* albo *age-adjusted*). Przykładowo jeżeli porównamy współczynnik zgonów USA i Nigerii to okaże się że w USA jest wyższy a to z tego powodu że społeczeństwo amerykańskie jest znacznie starsze (a umierają zwykle ludzie starzy)

Współczynnik zgonów standaryzowany według wieku to ważona średnia współczynników w poszczególnych grupach wiekowych, gdzie wagami są udziały tychże grup wiekowych w pewnej **standardowej populacji**

1.7 Oprogramowanie

Nie da się praktykować statystyki bez korzystania z programów komputerowych i mamy w tym zakresie trzy możliwości:

1. Arkusz kalkulacyjny. Przydatny na etapie zbierania danych i ich wstępnej analizy, później już niekoniecznie. Policzenie niektórych rzeczy jest niemożliwe (brak stosownych procedur) lub

czasochłonne (w porównaniu do 2--3)

2. Oprogramowanie specjalistyczne komercyjne takie jak programy STATA czy SPSS. Wady: cena i czas niezbędny na ich poznanie.

3. Oprogramowanie specjalistyczne darmowe: Jamovi oraz R Same zalety:-)

W większości podręczników opisuje się **procedury** oraz **program**, w którym te procedury można zastosować jednocześnie. My zdecydowaliśmy się oddzielnie przestawić teorię statystyki (rozdziały 1--5) a oddzielnie opis posługiwania się konkretnym programem (rozdział 6)

Rozdział 2

Analiza jednej zmiennej

Statystyka opisowa (opis statystyczny) to zbiór metod statystycznych służących do -- surprise, surprise -- opisu (w sensie przedstawienia sumarycznego) zbioru danych; w zależności od typu danych (przekrojowe, czasowe, przestrzenne) oraz sposobu pomiaru (dane nominalne, porządkowe liczbowe) należy używać różnych metod.

W przypadku **danych przekrojowych** opis statystyczny nazywany jest **analizą struktury** i sprowadza się do opisania danych z wykorzystaniem:

- tablic (statystycznych)
- wykresów
- parametrów (takich jak średnia czy mediana)

Rozkład cechy (zmiennej) to przyporządkowanie wartościom cechy zmiennej odpowiedniej **liczby wystąpień** (liczebności albo częstości (czyli popularnych procentów).

Analiza struktury (dla jednej zmiennej) obejmuje:

- **określenie tendencji centralnej** (**miary położenia**: wartość przeciętna, mediana, dominanta);
- **zróżnicowanie wartości** (rozproszenie: odchylenie standardowe, rozstęp ćwiartkowy);
- **asymetrię** (rozłożenie wartości zmiennej wokół średniej);

2.1 Tablice statystyczne

Tablica statystyczna to (w podstawowej formie) dwukolumnowa tabela zawierająca wartości cechy oraz odpowiadające tym wartościom liczebności.

Tablica dla cechy niemierzalnej (nominalnej albo porządkowej)

Absolwenci studiów pielęgniarskich w ośmiu największych krajach UE
Tablica: Absolwenci studiów pielęgniarskich w ośmiu największych krajach UE w roku 2018

kraj	liczba
Belgium	7203
Germany	35742
Spain	9936
France	25757
Italy	11207
Netherlands	9920
Poland	9070
Romania	18664

Źródło: Eurostat, tablica Health graduates (HLTH_RS_GRD)

W przykładzie **jednostką badania** jest absolwent studiów pielęgniarskich w roku 2018, **badaną cechą** zaś **kraj w którym ukończył studia** (cecha nominalna).

Tablica dla cechy mierzalnej liczbowej skokowej

Cecha skokowa to taka cecha, która może przyjąć skończoną liczbę wartości. Matematycznym odpowiednikiem cechy skokowej jest zbiór liczb całkowitych.

Jeżeli tych wartości jest mało tablica zawiera wyliczenie wartości cechy i odpowiadających im liczebności. Jeżeli liczba wariantów cechy jest duża tablica zawiera klasy wartości (przedziały wartości) oraz odpowiadające im liczebności.

Liczba przedziałów jest dobierana metodą prób i błędów, tak aby:

- przedziały wartości powinny być jednakowej rozpiętości.
- Na zasadzie wyjątku dopuszcza się aby pierwszy i ostatni przedział były **otwarte**, tj. nie miały dolnej (pierwszy) lub górnej (ostatni) **granicy**
- nie było przedziałów z zerową liczebnością
- przedziałów nie było za dużo ani za mało (typowo 8--15)
- większość populacji nie znajdowała się w jednej czy dwóch przedziałach

Gospodarstwa domowe wg liczby samochodów

Tablica: Gospodarstwa domowe we wsi X wg liczby samochodów w roku 2022

liczba samochodów	liczba gospodarstw	%
0	230	39.3162393
1	280	47.8632479
2	70	11.9658120
3 i więcej	5	0.8547009
razem	585	100.0000000

Źródło: obliczenia własne

Tablica dla cechy mierzalnej liczbowej ciągłej

Cecha ciągła to taka cecha, która może przyjąć nieskończoną/nieprzeliczalną liczbę wartości. Matematycznym odpowiednikiem cechy skokowej jest zbiór liczb rzeczywistych.

Tablica zawiera klasy wartości (przedziały wartości) oraz odpowiadające im liczebności.

Liczba przedziałów jest dobierana metodą prób i błędów, tak aby:

- przedziały wartości powinny być jednakowej rozpiętości.

- Na zasadzie wyjątku dopuszcza się aby pierwszy i ostatni przedział były **otwarte**, tj. nie miały dolnej (pierwszy) lub górnej (ostatni) **granicy**
- nie było przedziałów z zerową liczebnością
- przedziałów nie było za dużo ani za mało (typowo 8--15)
- większość populacji nie znajdowała się w jednej czy dwóch przedziałach
- zwykle przyjmuje się za końce przedziałów **okrągłe liczby** bo dziwnie by wyglądało gdyby koniec przedziału np. był równy 1,015 zamiast 1,0.

Dzietność kobiet na świecie
Współczynnik dzietności (*fertility ratio* albo FR) -- przeciętna liczba urodzonych dzieci przypadających na jedną kobietę w wieku rozrodczym (15--49 lat). Przyjmuje się, iż FR między 2,10--2,15 zapewnia zastępowalność pokoleń.
Dane dotyczące dzietności dla wszystkich krajów świata można znaleźć na stronie https://ourworldindata.org/grapher/fertility-rate-complete-gapminder) Zbudujmy tablicę przedstawiającą rozkład współczynników dzietności w roku 2018
Krajów jest 201. Wartość minimalna to 1.22 a wartość maksymalna to 7.13. Decydujemy się na rozpiętość przedziału równą 0,5; dolny koniec pierwszego przedziału przyjmujemy jako 1,0.
Tablica: Kraje świata według współczynnika dzietności (2018)

Wsp. dzietności	liczba krajów
(1,1.5]	24
(1.5,2]	61
(2,2.5]	40
(2.5,3]	17
(3,3.5]	8
(3.5,4]	15
(4,4.5]	11
(4.5,5]	12
(5,5.5]	6
(5.5,6]	5
(6,6.5]	1
(7,7.5]	1

Źródło: https://ourworldindata.org/grapher/fertility-rate-complete-gapminder

Każda tablica statystyczna **musi** mieć:

1. Część liczbową (kolumny i wiersze);
   - żadna rubryka w części liczbowej nie może być pusta (żelazna zasada); w szczególności brak danych należy explicite zaznaczyć umownym symbolem
2. Część opisową:
   - tytuł tablicy;
   - nazwy (opisy zawartości) wierszy;
   - nazwy (opisy zawartości) kolumn;
   - wskazanie źródła danych;
   - ewentualne uwagi odnoszące się do danych liczb.

Pominięcie czegokolwiek z powyższego jest **ciężkim błędem**. Jeżeli nie ma danych (a często nie ma--z różnych powodów -- należy to zaznaczyć a nie pozostawiać pustą rubrykę)

2.2 Wykresy

Wykresy statystyczne są graficzną formą prezentacji materiału statystycznego, są mniej precyzyjne i szczegółowe niż tablice, natomiast bardziej sugestywne.

Celem jest pokazanie rozkładu wartości cechy w populacji: jakie wartości występują często a jakie rzadko, jak bardzo wartości różnią się między sobą. Jak różnią się rozkłady dla różnych, ale logicznie powiązanych populacji (np rozkład czegoś-tam w kraju A i B albo w roku X, Y i Z).

Do powyższego celu celu stosuje się:

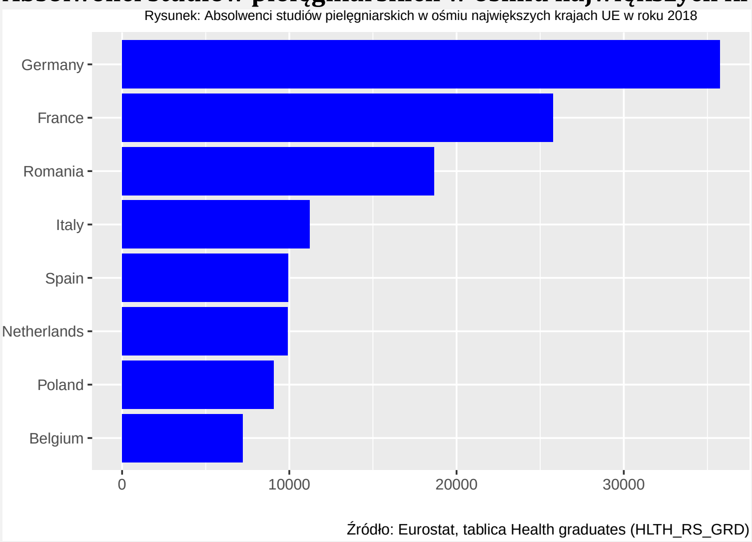
- **wykres słupkowy** (skala nominalna/porządkowa)
- **wykres kołowy** (skala nominalna/porządkowa)
- **histogram** (albo wykres słupkowy dla skal nominalnych)

Uwaga: **wykres kołowy** jest zdecydowanie gorszy od wykresu słupkowego i nie jest zalecany. **Każdy** wykres kołowy można wykreślić jako słupkowy i w takiej postaci będzie on bardziej zrozumiały i łatwiejszy w interpretacji.

2.2.1 Skala nominalna

Wykres słupkowy (*bar chart*)

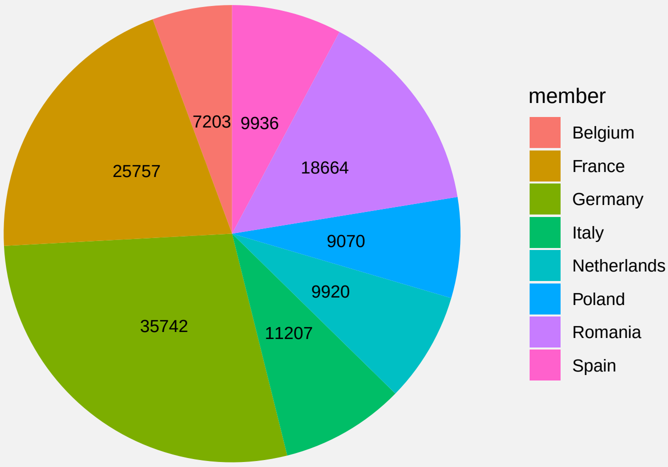
Absolwenci studiów pielęgniarskich w ośmiu największych krajach UE



Wykres kołowy (*pie chart*)

Absolwenci studiów pielęgniarskich w ośmiu największych krajach UE

Rysunek: Absolwenci studiów pielęgniarskich w ośmiu największych krajach UE w roku 2018



Źródło: Eurostat, tablica Health graduates (HLTH_RS_GRD)

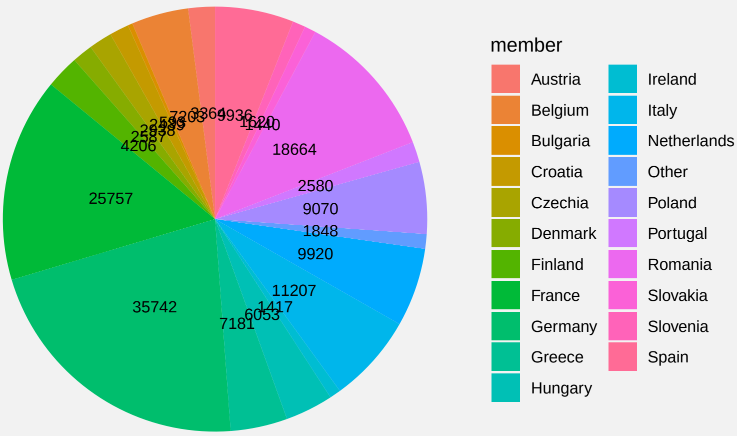
Wykres słupkowy i kołowy przedstawiają dokładnie to samo.

Wykres kołowy wygląda zapewne efektowniej (z uwagi na paletę kolorów) ale jest mniej efektywny. Wymaga legendy w szczególności, która utrudnia interpretację treści (nieustannie trzeba porównywać koło z legendą żeby ustalić który kolor to który kraj.)

Jeżeli zwiększymy liczbę krajów wykres kołowy staje się zupełnie nieczytelny (brakuje rozróżnialnych kolorów a wycinki koła są zbyt wąskie żeby cokolwiek wyróżniały):

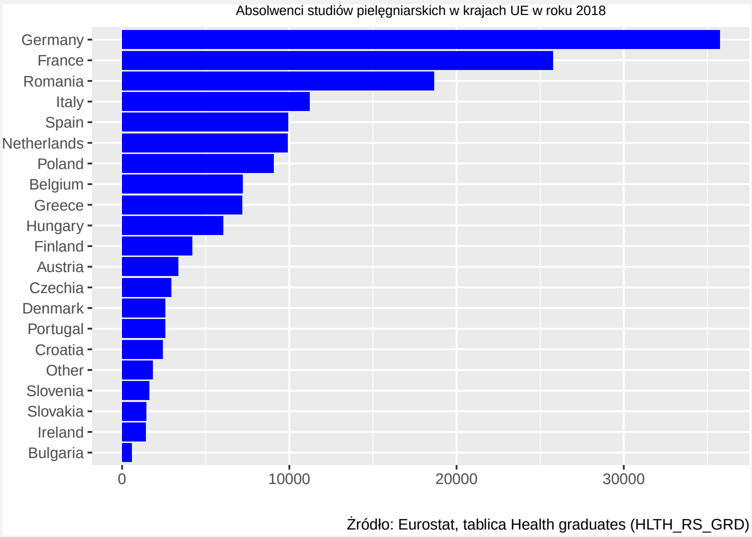
Absolwenci studiów pielęgniarskich w krajach UE

Rysunek: Absolwenci studiów pielęgniarskich w krajach UE w roku 2018



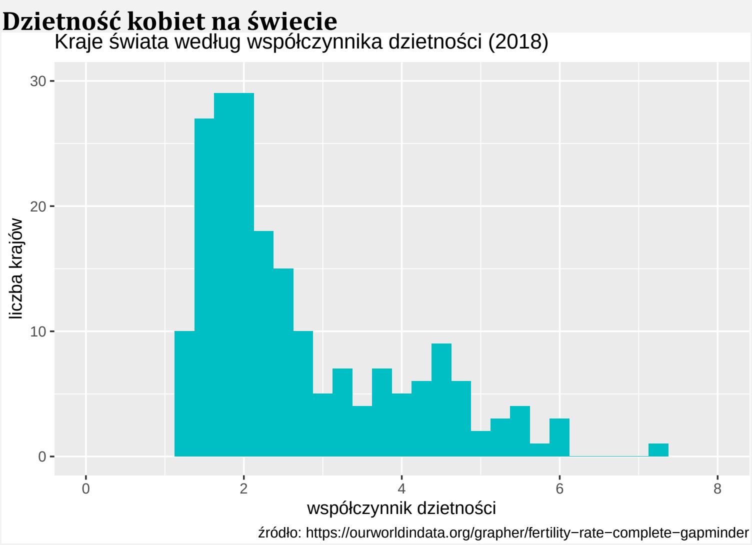
Źródło: Eurostat, tablica Health graduates (HLTH_RS_GRD)

Wykres słupkowy dalej jest natomiast OK:



2.2.2 Skala liczbowa

Histogram to coś w rodzaju wykresu słupkowego tylko na jednej osi zamiast wariantów cechy są przedziały wartości.

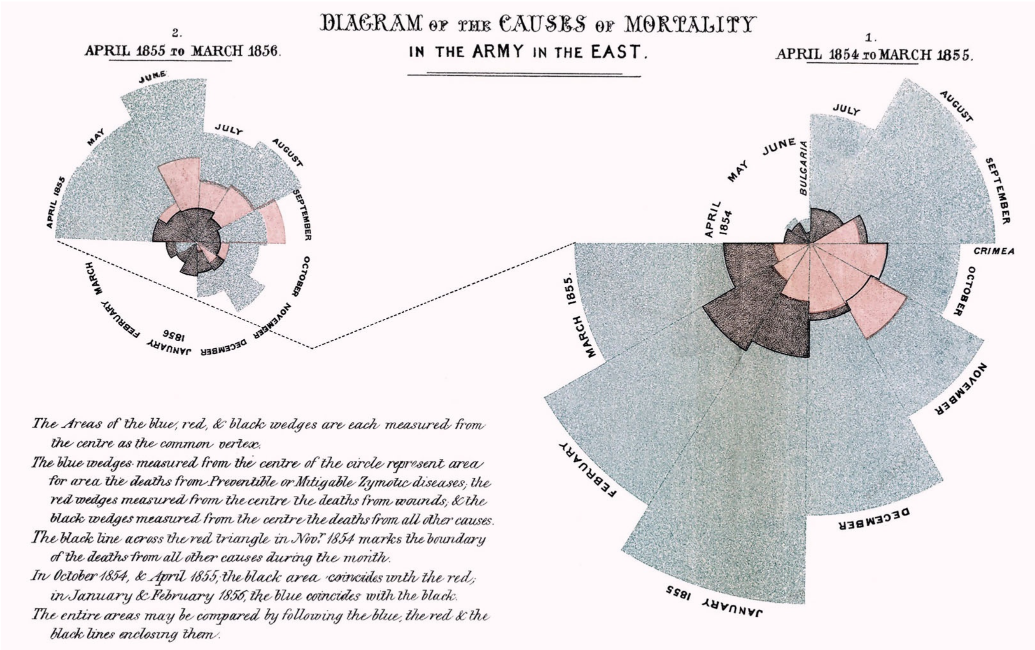


Podobnie jak tablice, rysunki powinny być opatrzone tytułem oraz zawierać źródło wskazujące na pochodzenie danych (zobacz przedstawione przykłady.)

2.3 Florence Nightingale

Nie każdy kto wie kim była Florence Nightingale, wie że była ona także statystykiem. W czasie wojny krymskiej nie tylko zorganizowała opiekę nad rannymi żołnierzami, ale również -- aby przekonać swoich przełożonych do zwiększenia nakładów na szpitale polowe -- prowadziła staranną ewidencję szpitalną oraz zgromadzone dane potrafiła analizować, używając wykresów własnego projektu.

W szczególności słynny jest diagram Nightingale zwane także różą Nightingale (rys. 2.1), które wprawdzie (podobno) nie okazały się szczególnie użyteczny, no ale nie każdy nowy pomysł jest od razu genialny:



Rysunek 2.1: Róża Nightingale

Jest to coś w rodzaju wykresu słupkowego tyle że zamiast słupków są wycinki koła. Wycinków jest dwanaście tyle ile miesięcy. Długość promienia a co za tym idzie wielkość pola wycinka zależy od wielkości zjawiska, który reprezentuje (przyczyna śmierci: rany/choroby/inne)

Wpisując Florence+Nightingale można znaleźć dużo informacji na temat, w tym: http://www.matematyka.wroc.pl/ciekawieomatematyce/pielegniarka-statystyczna

W 1859 roku Nightingale została wybrana jako pierwsza kobieta na członka Royal Statistical Society (Królewskie Stowarzyszenie Statystyczne) oraz została honorowym członkiem American Statistical Association (Amerykańskiego Stowarzyszenia Statystycznego).

Więc szanowi czytelnicy wnioski są oczywiste :-)

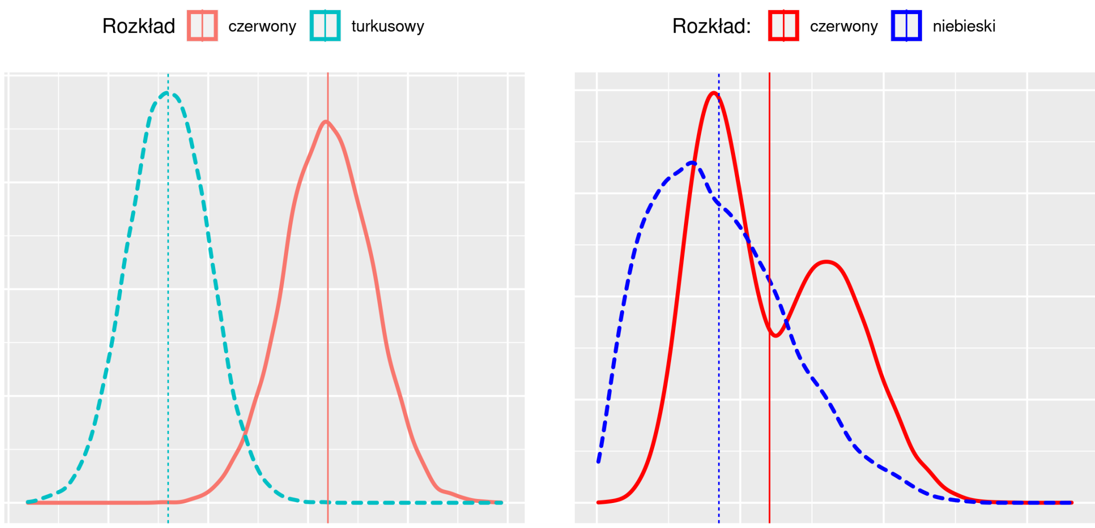
2.4 Analiza parametryczna

Analiza parametryczna z oczywistych względów dotyczy tylko zmiennych mierzonych na skali liczbowej.

2.4.1 Miary położenia

Miary przeciętne (**położenia**) charakteryzują średni lub typowy poziom wartości cechy. Są to więc takie wartości, wokół których skupiają się wszystkie pozostałe wartości analizowanej cechy.

Na rysunku 2.2 po lewej mamy dwa rozkłady różniące się poziomem przeciętnym. Rozkład czerwony ma przeciętnie większe wartości niż turkusowy. Są to rozkłady **jednomodalne**, czyli takie, w których rozkład



Rysunek 2.2: Rozkłady cechy a miary średnie

cechy skupia się wokół jednej wartości. Dla takich rozkładów ma sens obliczanie średniej arytmetycznej. Te średnie wartości są zaznaczone na rysunku linią pionową.

Na rysunku po prawej mamy rozkłady **nietypowe**: **wielomodalne** (czerwony) lub **niesymetryczne** (niebieski). W rozkładzie niesymetrycznym wartości skupiają się nie centralnie, ale po prawej/lewej od środka przedziału zmienności/wartości średniej).

W świecie rzeczywistym zdecydowana większość rozkładów jest jednomodalna. Rzadkie przypadki rozkładów wielomodalnych zwykle wynikają z łącznego analizowania dwóch różniących się wartością średnią zbiorów danych. Oczywistym zaleceniem w takiej sytuacji jest analiza każdego zbioru oddzielnie.

Rodzaje miar położenia

- klasyczne
  - **średnia arytmetyczna**
- pozycyjne
  - **mediana**
  - **dominanta**
  - **kwartyle**
  - ewentualnie kwantyle, decyle, centyle (rzadziej używane)

Średnia arytmetyczna (*Mean*, *Arithmetic mean*) to łączna suma wartości podzielona przez liczbę sumowanych jednostek. Jeżeli wartość jednostki i w N-elementowym zbiorze oznaczymy jako x_i (gdzie: $i = 1, \dots, N$) to średnią można zapisać jako $\bar{x} = (x_1 + \cdots + x_N)/N$

Uwaga: we wzorach statystycznych zmienne zwykle oznacza się małymi literami a średnią dla zmiennej przez umieszczenie nad nią kreski poziomej czyli $\bar{x}$ to średnia wartość zmiennej x.

Mediana (*Median*, kwartyl drugi) dzieli **uporządkowaną** zbiorowość na dwie równe części; połowa jednostek ma wartości cechy mniejsze lub równe medianie, a połowa wartości cechy równe lub większe od mediany. Stąd też mediana bywa nazywana wartością środkową.

Własności mediany: odporna na wartości nietypowe (w przeciwieństwie do średniej)

Kwartyle: coś jak mediana tylko bardziej szczegółowo. Kwartyli jest trzy i dzielą one zbiorowość na 4 równe części, każda zawierająca 25% całości.

Pierwszy kwartyl dzieli **uporządkowaną** zbiorowość w proporcji 25%--75%. Trzeci dzieli **uporządkowaną** zbiorowość w proporcji 75%--25%. Drugi kwartyl to mediana.

Kwantyle (D, wartości dziesiętne), podobnie jak kwartyle, tyle że dzielą na 10 części.

Centyle (P, wartości setne), podobnie jak kwantyle tyle że dzielą na 100 części. Przykładowo wartość 99 centyla i mniejszą ma 99% jednostek w populacji.

Współczynnik dzietności na świecie w roku 2018
Średnia: 2.68. Interpretacja: średnia wartość współczynnika dzietności wyniosła 2.68 dziecka.
Mediana: 2.2. Interpretacja mediany: współczynnik dzietności w połowie krajów na świecie wynosiła 2.2 dziecka i mniej.
Uwaga: średnia dzietność na świecie **nie wynosi** 2.68 dziecka (bo po pierwsze uśredniamy kraje a nie kobiety a po drugie kraje różnią się liczbą ludności). Podobnie dzietność połowy kobiet na świecie wyniosła 2.2 dziecka i mniej jest niepoprawną interpretacją mediany (z tych samych względów jak w przypadku średniej.)

Generalna uwaga: interpretacja średniej-średnich często jest nieoczywista i należy uważać. (a współczynnik dzietności jest średnią: średnia liczba dzieci urodzonych przez kobietę w wieku rozrodczym. Jeżeli liczymy średnią dla 202 krajów, to mamy *średnią-średnich*). Inny przykład: odsetek ludności w wieku poprodukcyjnym wg powiatów (średnia z czegoś takiego nie da nam odsetka ludności w wieku poprodukcyjnym w Polsce, bo powiaty różnią się liczbą ludności.)

Współczynnik dzietności (kontynuuacja):
Pierwszy kwartyl: 1.75; trzeci kwartyl 3.56 co oznacza że 25% krajów miało wartość współczynnika dzietności nie większą niż 1.75 dziecka a 75% krajów miało wartość współczynnika dzietności nie większą niż 3.56 dziecka.

2.4.2 Miary zmienności

Miary zmienności określają zmienność (dyspersję albo rozproszenie) w zbiorowości

Rodzaje miar zmienności:

- Klasyczne
  - Wariancja i odchylenie standardowe
- Pozycyjne
  - rozstęp
  - rozstęp ćwiartkowy

Wariancja (*variance*) jest to średnia arytmetyczna kwadratów odchyleń poszczególnych wartości cechy od średniej arytmetycznej zbiorowości. Co można zapisać

$$s^2 = \frac{1}{N}\left((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^N\right)$$

Przy czym często zamiast dzielenie przez N dzielimy przez $N - 1$.

Odchylenie standardowe (*standard deviation*, sd) jest pierwiastkiem kwadratowym z wariancji. Parametr ten określa przeciętną różnicą wartości cechy od średniej arytmetycznej.

Rozstęp ćwiartkowy (*interquartile range*, IQR) ma banalnie prostą definicję:

$$R_Q = Q_3 - Q_1$$

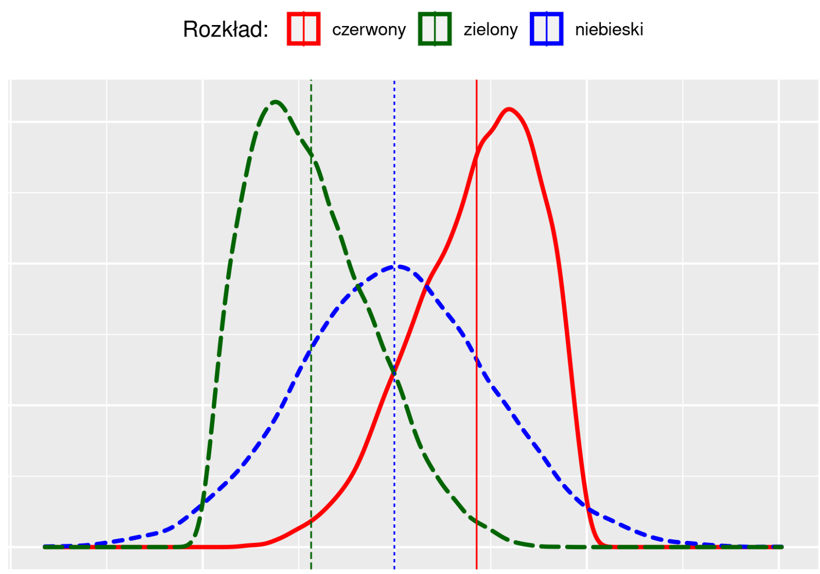
gdzie: Q_1, Q_3 oznaczają odpowiednio pierwszy oraz trzeci kwartyl.

Współczynnik dzietności (kontynuacja)
Średnie odchylenie od średniej wartości współczynnika wynosi 1.26 dziecka. Wartość rozstępu ćwiartkowego wynosi 1.81 dziecka.

Uwaga: odchylenie standardowe/ćwiartkowe są miarami mianowanymi. Zawsze należy podać jednostkę miary.

2.4.3 Miary asymetrii

Asymetria (*skewness*), to odwrotność symetrii. Szereg jest symetryczny jeżeli jednostki są rozłożone „równomiernie'' wokół wartości średniej. W szeregu symetrycznym wartości średniej i mediany są sobie równe. Skośność może być dodatnia (*positive skew*) lub ujemna (*negative skew*). Czym się różni jedna od drugiej widać na rysunku 2.3.



Rysunek 2.3: Rozkłady symetryczne i asymetryczne

Miary asymetrii:

- klasyczny współczynnik asymetrii (g)
  - przyjmuje wartości ujemne dla asymetrii lewostronnej; a dodatnie dla prawostronnej. Teoretycznie może przyjąć dowolnie dużą wartość ale w praktyce rzadko przekracza 3 do do wartości bezwzględnej.
  - wartości większe od 2 świadczą o dużej a większe od 3 o bardzo dużej asymetrii
- współczynniki asymetrii Pearsona (W_s)

  - wykorzystuje różnice między średnia Medianą: $W_s = (\bar{x} - Me)/s$

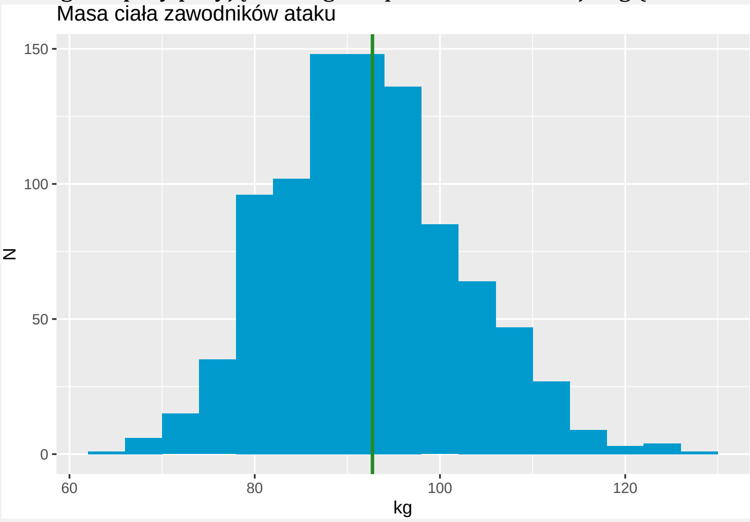
- Współczynnik asymetrii (skośności) oparty na odległościach między kwartylami lub decylami:

  - Obliczany jest według następującej formuły: $W_{sq} = \frac{(Q_3-Q_2)-(Q_2-Q_1)}{Q_3-Q_1}$

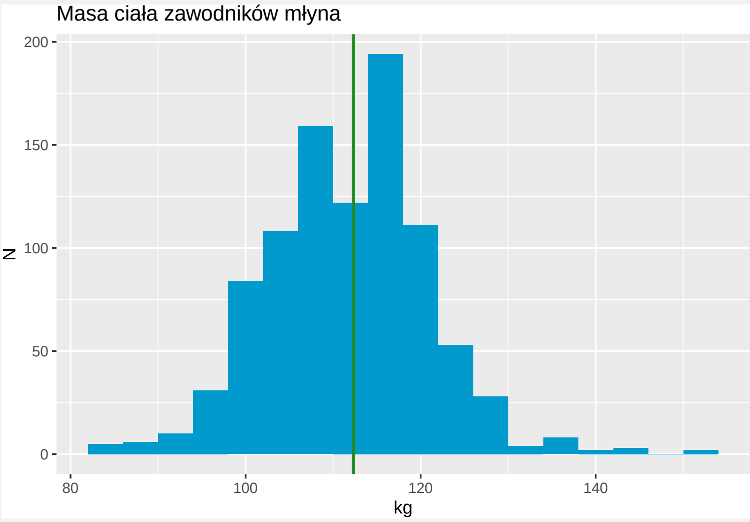
2.5 Porównanie wielu rozkładów

Często strukturę jednego rozkładu należy porównać z innym. Albo trzeba porównać strukturę wielu rozkładów. Pokażemy jak to zrobić na przykładzie.

Masa ciała uczestników Pucharu Świata w Rugby
W turniejach o puchar świata w Rugby w latach 2015, 2019 i 2023 uczestniczyło łącznie 1879 zawodników. W grze w rugby drużyna jest podzielona na dwie **formacje**: ataku i młyna. Należy scharakteryzować rozkład masy ciała zawodników obu formacji.
Zawodnicy ataku
Przeciętnie zawodnik ataku ważył 92.7 kg; mediana 92.0 kg (połowa zawodników ataku ważyła 92.0 kg i mniej); pierwszy/trzeci kwartyl 85.5/99 kg (1/4 zawodników ataku ważyła 85.5 kg i mniej; 1/4 zawodników ataku ważyła 99 kg i więcej;
Odchylenie standardowe 10.1 kg (przeciętnie odchylenie od średniej arytmetycznej wynosi 10.1 kg); rozstęp ćwiartkowy wynosi 13.5 kg (rozstęp 50% środkowych wartości wynosi 13.5 kg)
Histogram przy przyjęciu długości przedziału równej 4kg (linia zielona oznacza poziom średniej):



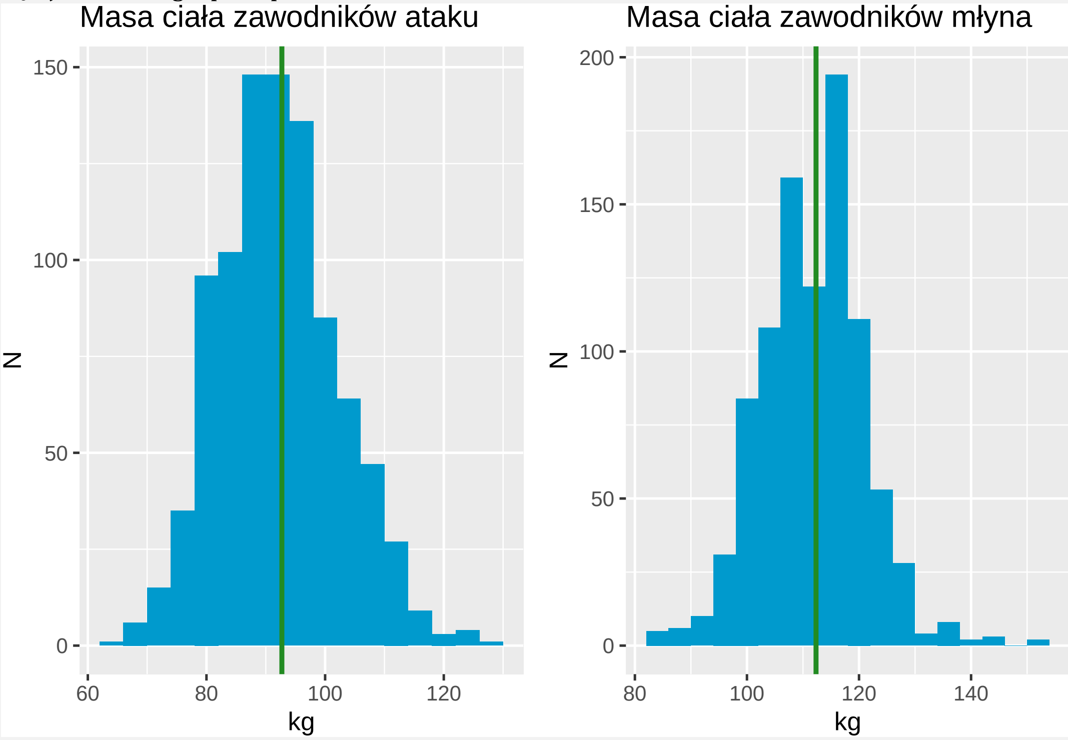
Zawodnicy młyna
Średnio zawodnik młyna ważył 112.3 kg; mediana 112.0 kg (połowa zawodników młyna ważyło 112 kg i mniej); pierwszy/trzeci kwartyl 106/118 kg (1/4 zawodników młyna ważyło 106 kg i mniej; 1/4 zawodników młyna ważyło 118 kg i więcej;
Odchylenie standardowe 9.2 kg (przeciętnie odchylenie od średniej arytmetycznej wynosi 9.2 kg); rozstęp ćwiartkowy wynosi 12 kg (rozstęp 50% środkowych wartości wynosi 12 kg)
Histogram przy przyjęciu długości przedziału równej 4kg (linia zielona oznacza poziom średniej):



Porównanie atak vs młyn

Miara	Atak	Młyn
średnia	92.7087379	112.327957
mediana	92	112
odchyl.st	10.0723816	9.2406513
iqr	13.5	12

średnio zawodnik młyna ważył prawie 20 kg więcej od zawodnika ataku (w przypadku mediany jest to dokładnie 20 kg więcej). Zmienność mierzona wielkością odchylenia standardowego oraz IQR jest w obu grupach podobna.



2.5.1 Wykres pudełkowy

Do porównania wielu rozkładów szczególnie użyteczny jest wykres zwany pudełkowym (**box-plot**).

Konstrukcja pudełka na wykresie (por rys. @(fig:boxplot)):

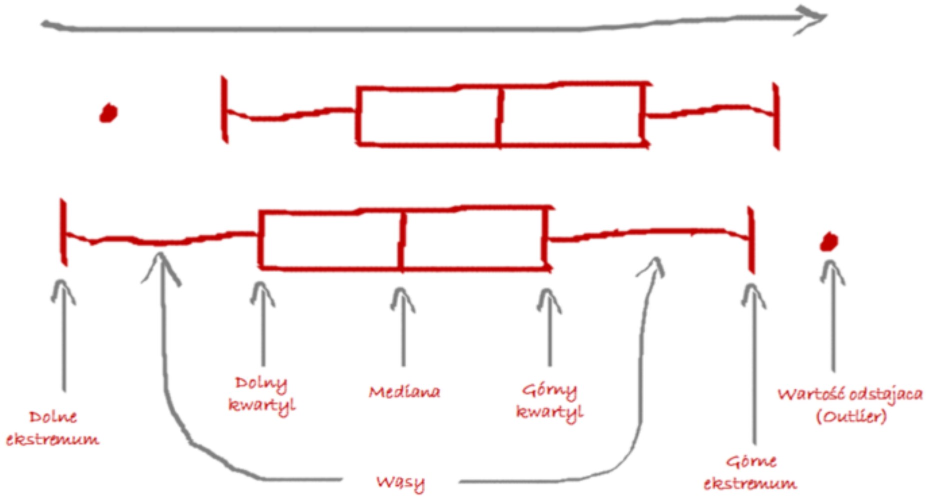
- górny dolny bok pudełka jest równy kwartylom;
- linia pionowa w środku pudełka jest równa medianie;
- linie poziome (zwane wąsami) mają długość równą $Q_1 - 1,5\mathrm{IQR}$ oraz $Q_3 + \mathrm{IQR}$ (dla przypomnienia: Q_1, Q_3 to kwartyle, zaś IQR to rozstęp ćwiartkowy);
- kropki nad/pod wąsami to wartości zmiennej większe od $Q_3 + 1,5\mathrm{IQR}$ lub mniejsze od $Q_1 - 1,5\mathrm{IQR}$.

Interpretacja pudełka:

- linia pozioma w środku pudełka określa przeciętny poziom zjawiska;
- wysokość pudełka oraz wąsów określa zmienność (im większe wąsy/wysokość tym większa zmienność);
- kropki nad/pod wąsami to **obserwacje nietypowe**.

Zwróć uwagę na sztuczkę: wartości nietypowe nie są definiowane jako na przykład górne/dolne 1% wszystkich wartości (bo wtedy **każdy rozkład** miałby wartości nietypowe), ale jako wartości mniejsze/większe od $Q_* \pm 1,5 \times \mathrm{IQR}$. Wszystkie wartości rozkładów o umiarkowanej zmienności mieszczą się wewnątrz czegoś takiego.

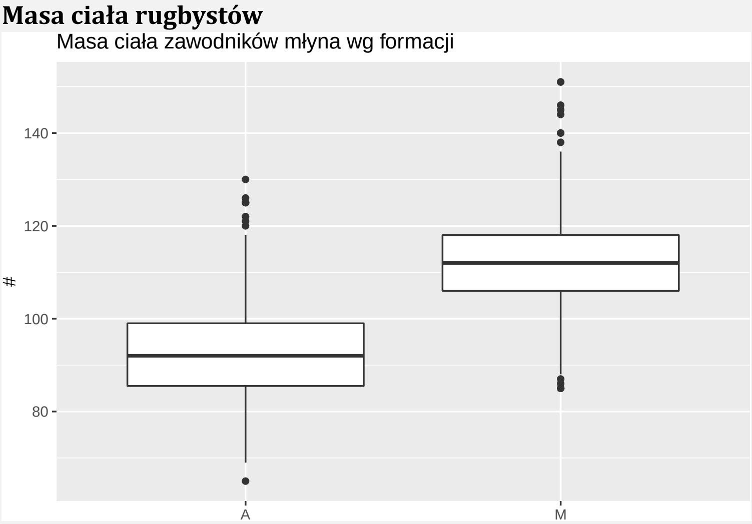
Typowo wykres zawiera wiele pudełek a każde pudełko wizualizuje jeden rozkład. Pudełka mogą być umieszczone jedno pod drugim, tak jak na rysunku @(fig:boxplot) lub jedno obok drugiego jak na



Rysunek 2.4: Wykres pudełkowy

przykładach poniżej.

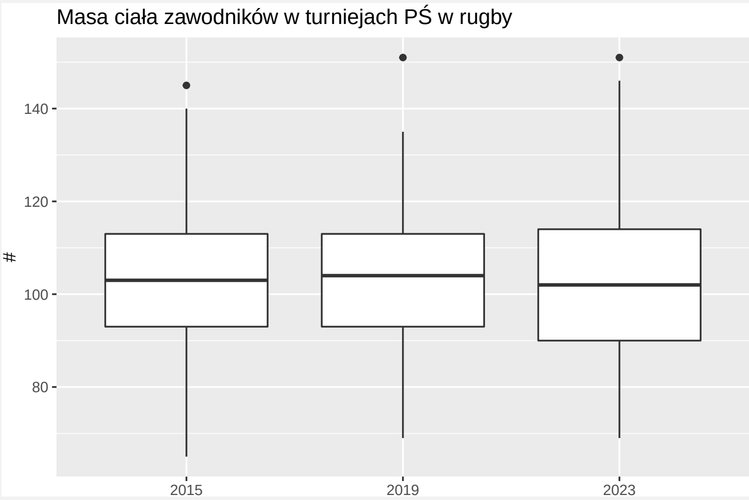
Masa ciała rugbystów



Z wykresu od razu widać, który rozkład ma wyższą średnią, który większe rozproszenie oraz w którym występują wartości nietypowe.

Pudełek może być więcej niż dwa oczywiście. Następny przykład pokazuje porównanie rozkładów masy ciała zawodników rugby na poszczególnych turniejach.

Masa ciała rugbystów



Od razu widać, że przeciętnie najcięższy zawodnicy byli na turnieju w roku 2019; największe zróżnicowanie masy ciała występowało na turnieju w roku 2023.

2.6 Zestawienie metod opisu statystycznego

W rozdziale przedstawiono osiem sposobów opisania rozkładu zmiennej:

1. Tablice statystyczne
2. Wykres słupkowy
3. Wykres kołowy (niezalecany)
4. Histogram
5. Wykres pudełkowy
6. Miary tendencji centralnej: średnia, mediana, kwartyle
7. Miary rozproszenia: odchylenie standardowe, rozstęp ćwiartkowy
8. Miary asymetrii

Rozdział 3

Łagodne wprowadzenie do wnioskowanie statystycznego

Chcemy się dowiedzieć czegoś na temat populacji (całości) na podstawie próby (części tej całości).

Przykładowo chcemy ocenić ile wynosi średnia waga główki kapusty na 100 h polu. Można ściąć wszystkie i zważyć, ale można też ściąć trochę (pobrać próbę się mówi uczenie) zważyć i poznać średnią na całym polu z dobrą dokładnością.

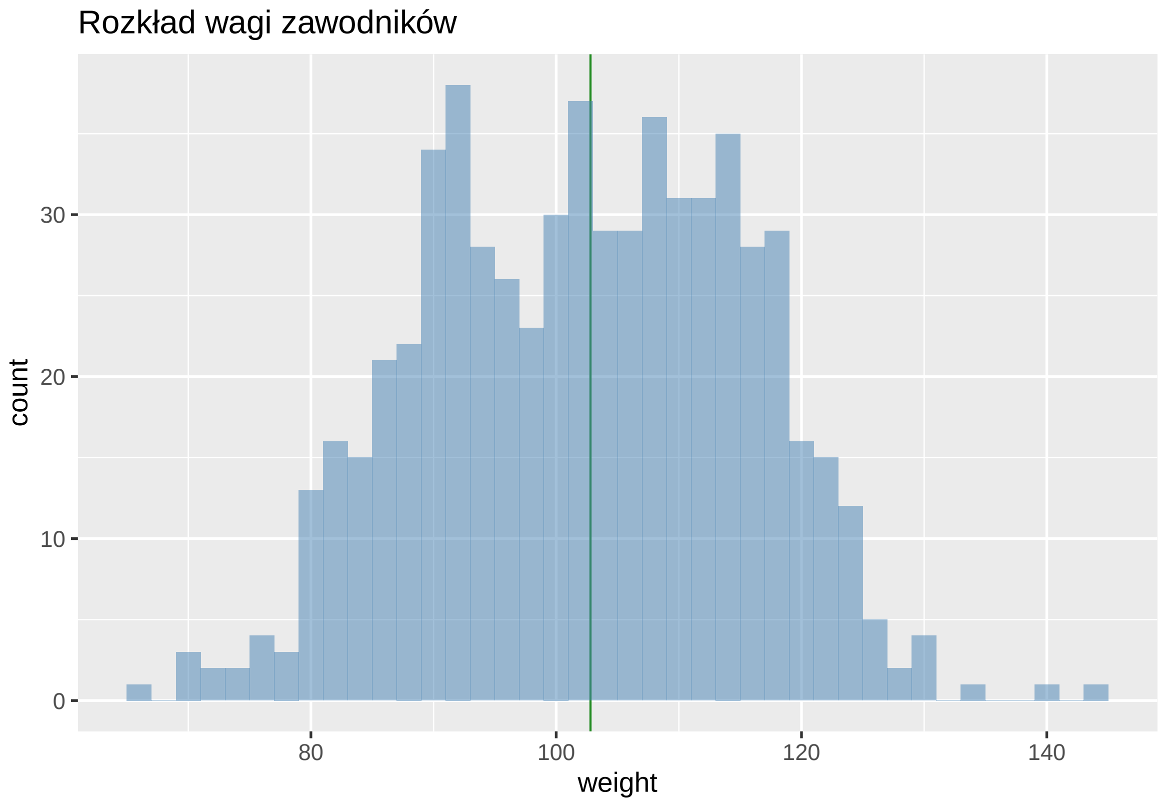
3.1 Masa ciała uczestników PŚ w rugby

W turnieju o Puchar Świata w rugby w 2015 roku uczestniczyło 623 rugbystów. Znamy szczegółowe dane odnośnie wzrostu i wagi każdego uczestnika turnieju. Obliczamy (prawdziwą) średnią, odchylenie standardowe i współczynnik zmienności masy ciała:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    65.0    93.0   103.0   102.8   113.0   145.0
```

Czyli średnio rugbysta na turnieju RWC'2015 ważył 102.80 kg (`Mean` na wydruku powyżej) a odchylenie standardowe (s) wyniosło 12.92 kg.

Wykres (rozkład jest dwumodalny; bo w rugby są dwie grupy zawodników, wcale nie wszyscy > 110 kg):



Szacujemy średnią na podstawie 2 zawodników pobranych losowo

Powtarzamy eksperyment 1000 razy (dwóch bo dla jednego nie obliczymy wariancji)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    74.0    95.5   102.0   102.1   109.0   130.0
```

średnia (średnich z próby) ma wartość 102.13 a odchylenie standardowe 9.28. Wartość $s/\sqrt{2}$ (odchylenie standardowe podzielone przez pierwiastek kwadratowy z liczebności próby) jest równa 9.14. Zauważmy że ta wartość jest zbliżona do odchylenia standardowego uzyskanego w eksperymencie (9.28 vs 9.14)

szacujemy średnią na podstawie 10 zawodników pobranych losowo

Powtarzamy eksperyment 1000 razy

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    89.3    99.7   102.7   102.5   105.4   116.5
```

średnia wyszła 102.47 a odchylenie standardowe 4.09. Wartość $s/\sqrt{10}$ jest równa 4.09.

szacujemy średnią na podstawie 40 zawodników pobranych losowo

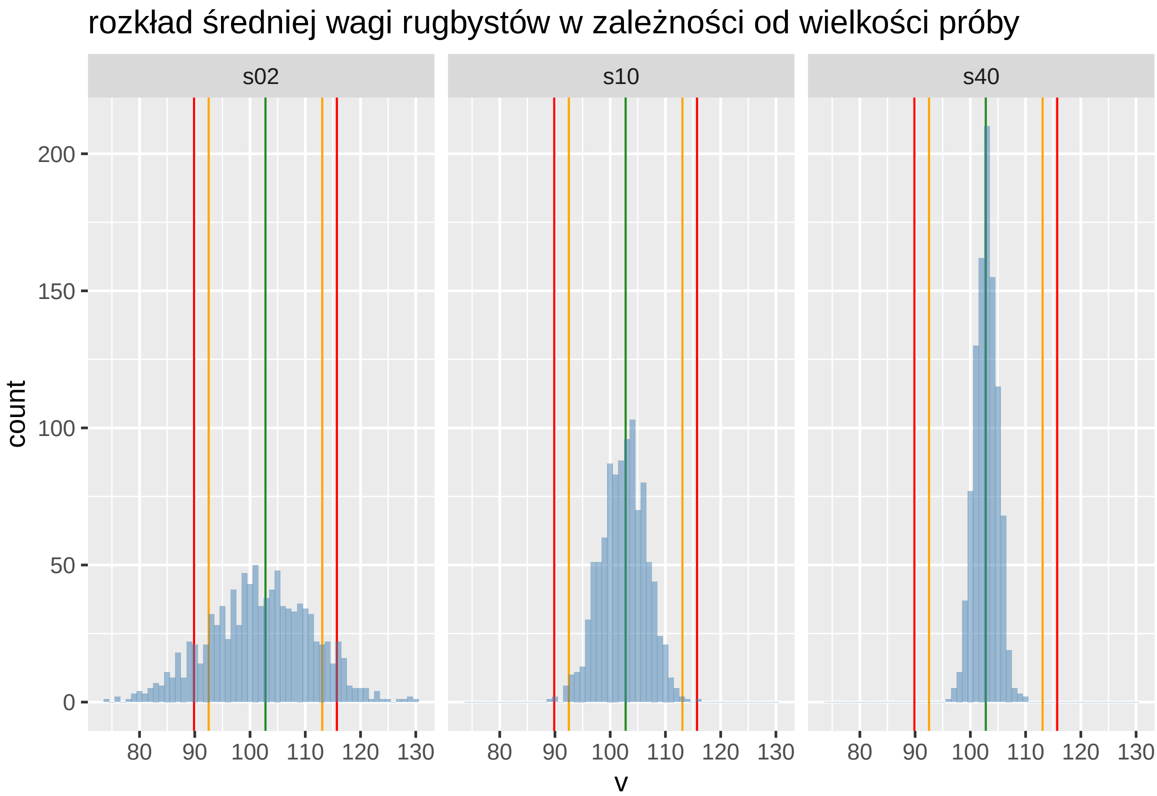
Uwaga: 40 zawodników to około 6.4% całego zbioru. Powtarzamy eksperyment 1000 razy

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   96.38  101.45  102.81  102.84  104.23  110.00
```

średnia jest równa 102.84 a odchylenie standardowe 2.04. Wartość $s/\sqrt{40}$ jest równa 2.04.

Wykres

Podsumujmy eksperyment wykresem rozkładu wartości średnich.



Wnioski z eksperymentu

Wartość średnią wyznaczamy na podstawie jakiejś konkretnej **metody**. Wydaje się na podstawie powyższych eksperymentów, że z dobrym skutkiem możemy jako metodę wykorzystać **średnią-z-próby**.

W ogólności taką metodą, która formalnie jest funkcją elementów z próby, nazywa się w statystyce **estymatorem**. Warto to pojęcie zapamiętać. Wnioskujemy o wartości parametru w populacji posługując się estymatorem.

Kontynuując wnioski z eksperymentu należy zauważyć, że wszystkie średnie-ze-średnich (bez względu na liczebność próby) są zbliżone do wartości prawdziwej (to się nazywa **nieobciążoność** estymatora); Mówiąc innymi słowy jeżeli będziemy oceniać wartość prawdziwej średniej na podstawie próby, a naszą ocenę powtórzymy wielokrotnie, to średnia będzie zbliżona do wartości prawdziwej (a nie np. niższa czy wyższa) Ta cecha jest niezależna od wielkości próby.

Jeżeli rośnie liczebność próby to zmienność wartości średniej-w-próbie maleje, co za tym idzie prawdopodobieństwo, że wartość oceniona na podstawie średniej z próby będzie zbliżona do wartości szacowanego parametru rośnie (to się nazywa **zgodność**). Co więcej dobrym przybliżeniem zmienności średniej-w-próbie jest prosta formuła $s/\sqrt{n}$ gdzie n jest liczebnością próby a s jest odchyleniem standardowym w populacji z której pobrano próbę.

Jeżeli mamy dwa rózne estymatory służące do oszacowania parametru, oba są **nieobciążone** oraz **zgodne**, to który wybrać? Ten która ma **mniejszą wariancję**. Taki estymator nazywa się **efektywny**.

Estymator zatem powinien być **nieobciążony**, **zgodny** oraz **efektywny** (czyli mieć małą wariancję). Można matematycznie udowodnić, że pewien estymator ma tak małą wariancję, że niemożliwe jest wynalezienie czegoś jeszcze bardziej efektywnego. Takim estymatorem średniej w populacji jest średnia z próby...

Konkretną wartość estymatora dla konkretnych wartości próby nazywamy **oceną** (parametru).

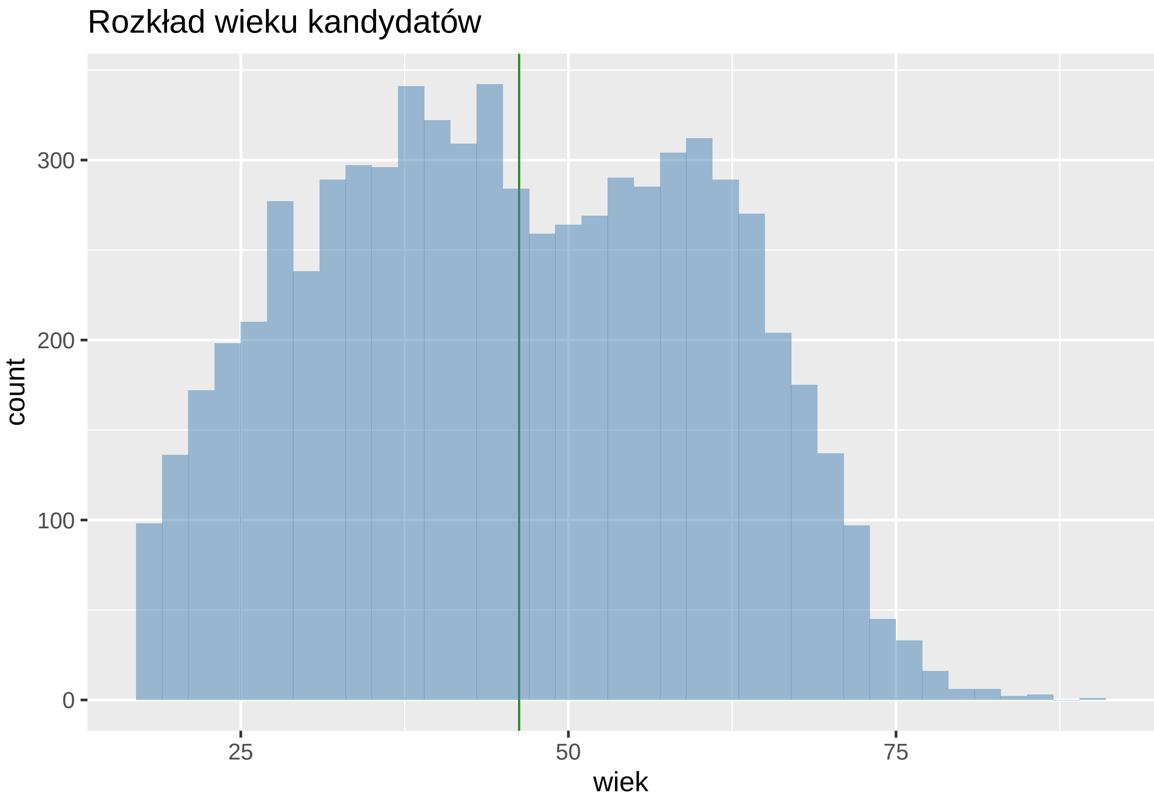
3.2 Wiek kandydatów na radnych

W wyborach samorządowychych w Polsce w roku 2018 o mandat radnego sejmików wojewódzkich ubiegało się 7076 kandydatów. Znamy szczegółowe dane odnośnie wieku każdego kandydata, bo to zostało publicznie podane przez Państwową Komisję Wyborczą. Obliczamy (prawdziwą) średnią, odchylenie standardowe i współczynnik zmienności wieku kandydatów:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   34.00   46.00   46.24   58.00   91.00
```

Czyli średnio kandydat miał 46.24 lat a odchylenie standardowe wieku wyniosło 14.61 lat.

Wykres (rozkład znowu jest dwumodalny z jakiś powodów):



Szacujemy średnią na podstawie 2 kandydatów pobranych losowo

Powtarzamy eksperyment 1000 razy

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   38.50   46.00   46.19   54.00   73.00
```

Średnia średnich z próby ma wartość 46.19 lat. Odchylenie standardowe wyniosło 10.62. Wartość $s/\sqrt{2}$ jest równa 10.33.

Szacujemy średnią na podstawie 10 kandydatów pobranych losowo

Powtarzamy eksperyment 1000 razy.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.30   43.10   46.20   46.28   49.40   62.70
```

Średnia średnich z próby ma wartość 46.28 lat. Odchylenie standardowe wyniosło 4.66. Wartość $s/\sqrt{10}$ jest równa 4.62.

Szacujemy średnią na podstawie 40 kandydatów pobranych losowo

Uwaga: 40 kandydatów to ok 0.6% całości. Powtarzamy eksperyment 1000 razy.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   39.00   44.74   46.27   46.23   47.76   53.27
```

Średnia średnich z próby ma wartość 46.23 lat. Odchylenie standardowe wyniosło 2.2729681. Wartość $s/\sqrt{40}$ jest równa 2.3105373.

Szacujemy średnią na podstawie 70 kandydatów pobranych losowo

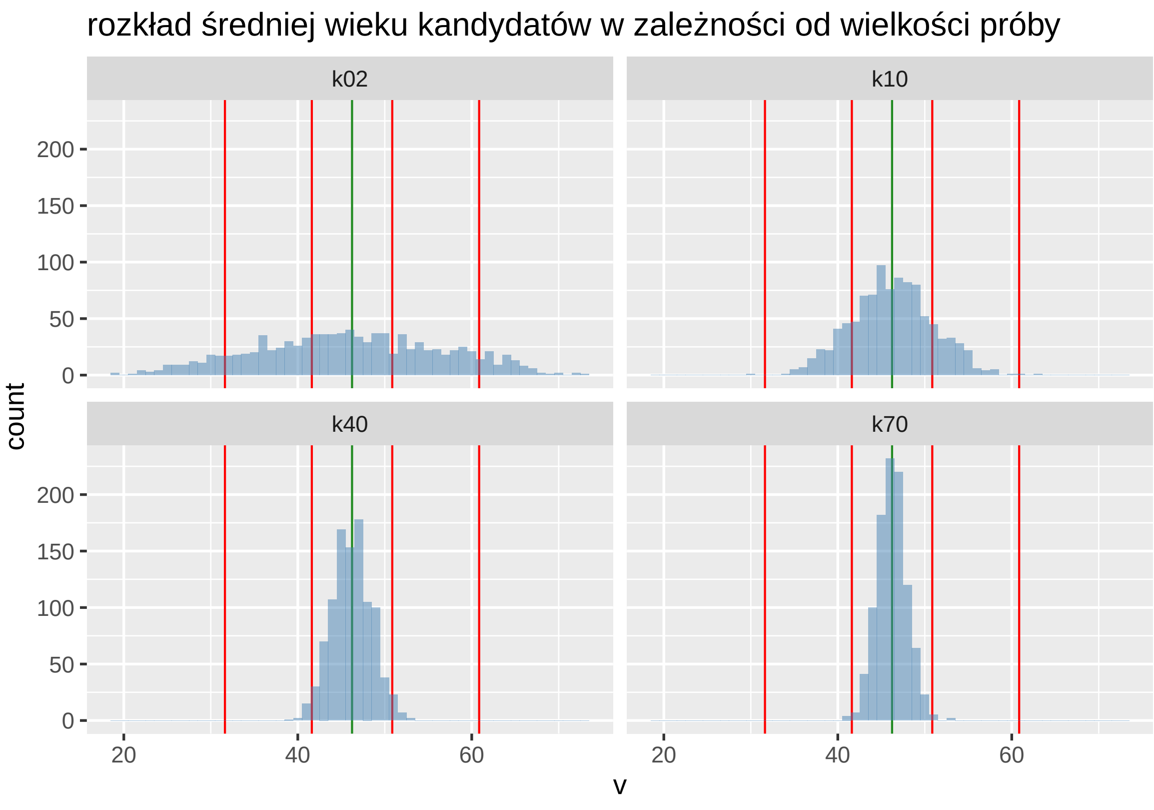
Uwaga: 70 kandydatów to około ok 1% całości (1000 powtórzeń)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   40.83   45.09   46.24   46.23   47.30   52.99
```

Średnia średnich z próby ma wartość 46.23 lat. Odchylenie standardowe wyniosło 1.6966764 Wartość $s/\sqrt{70}$ jest równa 1.746602.

Wykres

Podsumujmy eksperyment wykresem rozkładu wartości średnich.



Obserwujemy to samo co w przypadku wagi rugbystów: im większa próba tym dokładniejsza wartość średniej wieku. Bez względu na wielkość próby przeciętnie otrzymujemy prawdziwą wartość średniej.

Wniosek: precyzja wnioskowania zwiększa się wraz z liczebnością próby; tym szybciej im rozproszenie w populacji generalnej jest mniejsze. Żeby z dużą dokładnością wnioskować o średniej dla dużej populacji wcale nie trzeba pobierać dużej próby (w ostatnim przykładzie było to 1% całości).

3.3 Rozkład normalny

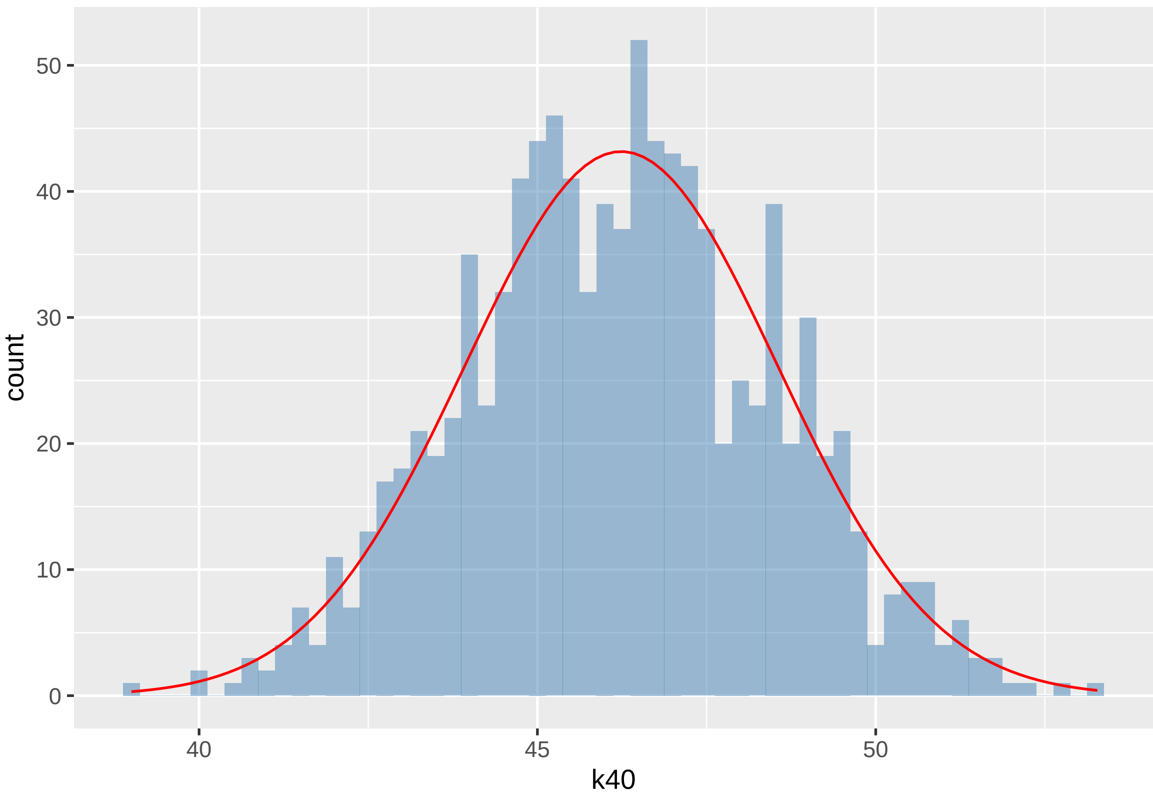
Rozkład empiryczny zmiennej to przyporządkowanie kolejnym wartościom zmiennej odpowiadających im liczebności.

Załóżmy że istnieje zapotrzebowanie społeczne na wiedzę na temat wieku kandydatów na radnych. Możemy to jak widać łatwo liczyć ale jednocześnie jest to kłopotliwe. Należy do tego mieć zbiór ponad 7 tys liczb. **Rozkład teoretyczny** to matematyczne uogólnienie **rozkładu empirycznego**. Jest to model matematyczny operujący pojęciem (ściśle sformalizowanym) **prawdopodobieństwa** (zamiast liczebności). **Rozkład teoretyczny** jest:

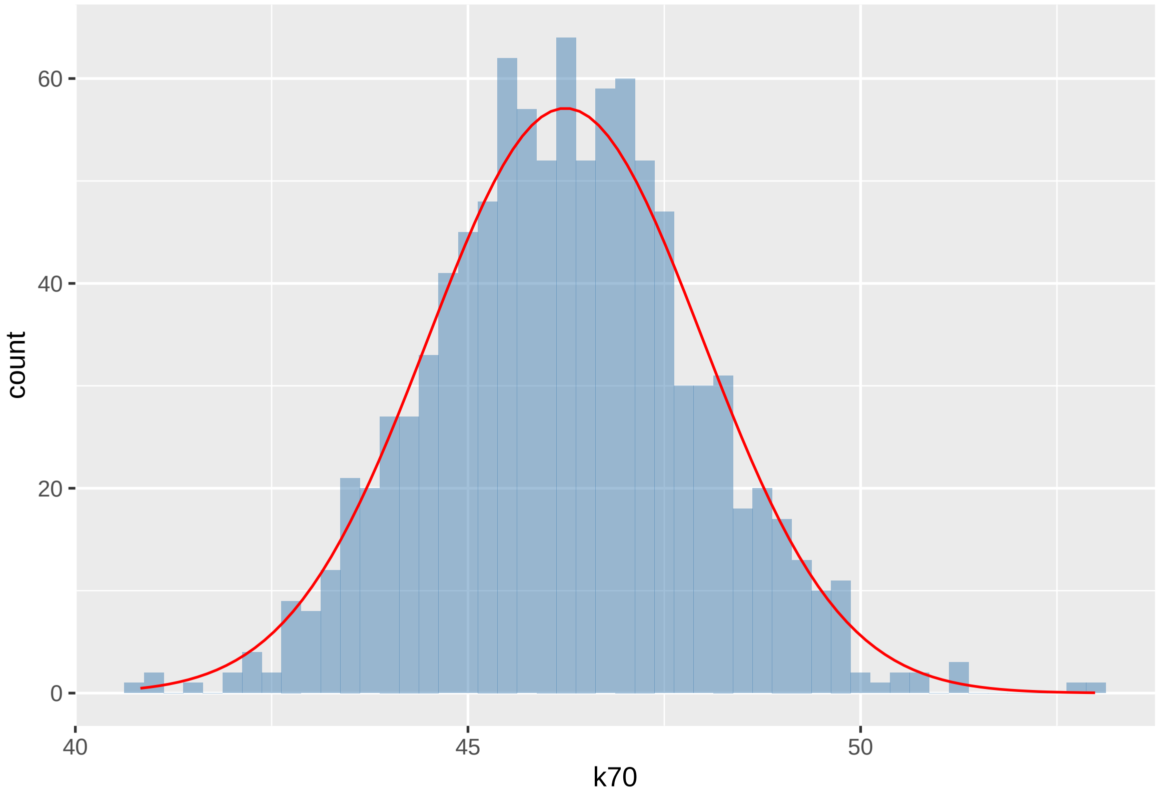
- zbliżony do empirycznego jeżeli chodzi o wyniki (jest przybliżeniem empirycznego)
- jest zdefiniowany za pomocą kilku liczb; nie ma potrzeby korzystania z liczebności

Żeby było ciekawiej istnieje dokładnie jeden **rozkład teoretyczny**, który z dobrą dokładnością opisuje rozkłady empiryczne będące wynikiem powyższej zabawy. Ten rozkład (zwany **normalnym**) zależy tylko od dwóch parametrów: średniej i odchylenia standardowego, gdzie średnia będzie równa (prawdziwej) średniej w populacji a odchylenie standardowe równe odchyleniu standardowemu w populacji podzielonemu przez pierwiastek z wielkości próby.

Dla próby 40-elementowej (wiek kandydatów) wygląda to tak:



dla próby 70-elementowej tak:



Prawda, że wynik jest całkiem dobry? Teoretyczność czerwonej krzywej polega na tym, że ona zawsze będzie identyczna, podczas gdy histogram będzie różny. Gdybyśmy powtórzyli nasz eksperyment (generowania 1000 losowych prób przypominam), to zapewne trochę by się różnił, bo byśmy wylosowali inne wartości do prób. Ta **teoretyczna abstrakcja** nazywa się **prawdopodobieństwem**. Rzucając monetą 1000 razy spodziewamy się po 500 orłów i reszek, co w modelu matematycznym będzie opisane jak: prawdopodobieństwo wyrzucenia orła wynosi 0,5. Rzucanie monetą to bardzo prosty eksperyment; nasz z liczeniem średniej wieku jest bardziej skomplikowany więc miło jest się dowiedzieć, że używając czerwonej krzywej można łatwo obliczyć jak bardzo prawdopodobne jest na przykład popełnienie błędu większego niż 10% średniej, albo większego niż 0,1 lat. Albo jak duża powinna być próba żeby ten błąd był nie większy niż 0,1 lat.

Interpretacja wartości rozkładu empirycznego zwykle jest w kategoriach ryzyka/szansy czy prawdopodobieństwa. Przykładowo interesuje nas prawdopodobieństwo, że kandydat ma mniej niż 30 lat. Takich kandydatów jest 1091 a wszystkich kandydatów dla przypomnienia jest 7076. Iloraz tych wartości będzie interpretowany jako ryzyko/szansa/prawdopodobieństwo (wynosi ono 15.42%.)

Podobnie można obliczyć prawdopodobieństwo, że wiek kandydata będzie się zawierał w przedziale 50--60 lat. Ponieważ kandydatów w wieku 50--60 lat jest 1570, to szukane prawdopodobieństwo jest równe: 22.19%.)

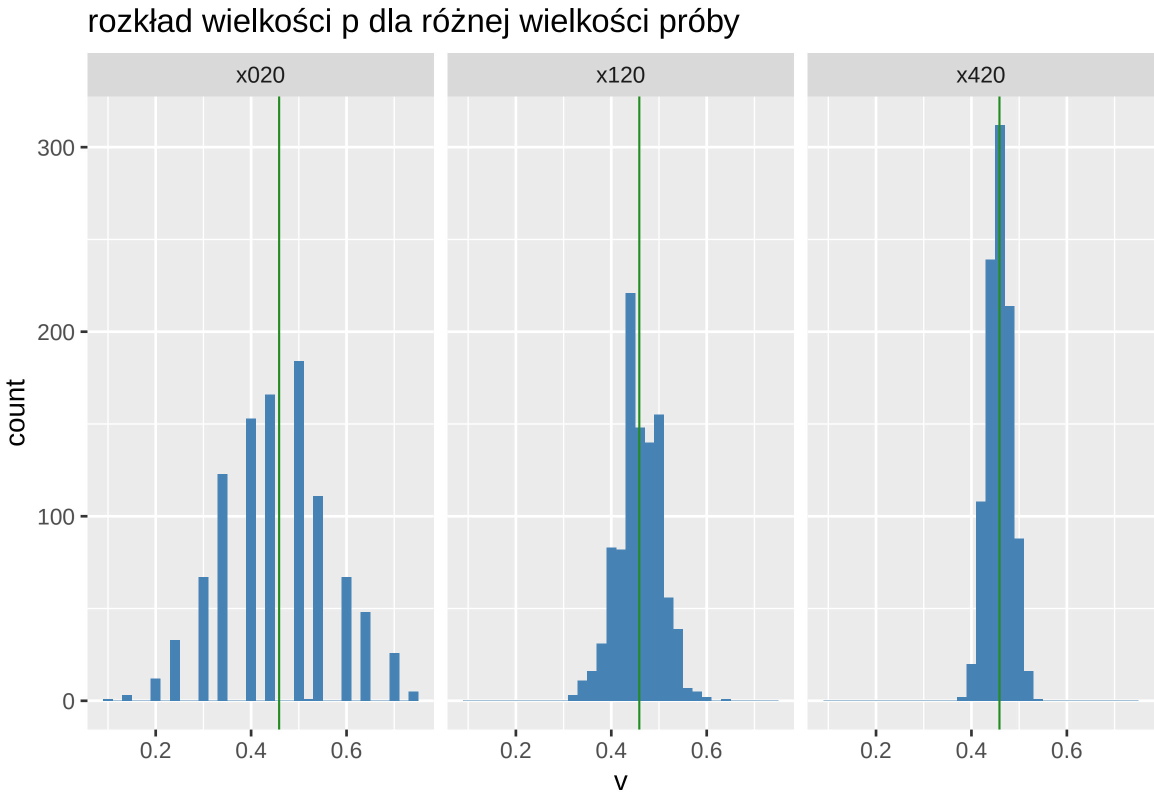
Jeżeli zamiast rozkładu empirycznego będziemy używać rozkład normalnego, który jak widzimy jest jego dobrym przybliżeniem, to nie musimy liczyć empirycznych liczebności. Wystarczy że znamy średnią i odchylenie standardowe a potrafimy obliczyć każde prawdopodobieństwo dla każdego przedziału wartości zmiennej.

W szczególności dla rozkładu normalnego prawdopodobieństwo $m \pm s$ (przyjęcie wartości z przedziału średnia plus/minus odchylenie standardowe) wynosi około 0,68 prawdopodobieństwo $m \pm 2 \times s$ wynosi około 0,95 a $m \pm 3 \times s$ około 0,997. Czyli w przedziale $[-3s < m, m + 3s]$ znajdują się praktycznie wszystkie wartości rozkładu. Albo innymi słowy przyjęcie wartości spoza przedziału średnia plus/minus trzykrotność odchylenia standardowego jest bardzo mało prawdopodobna.

Rozkład normalny będzie identyczny dla wagi rugbystów, wieku, wagi noworodków itd/itp. Uogólnieniem teoretycznym pojęcia **zmiennej statystycznej**, które do tej pory używaliśmy jest **zmienna losowa**, zmienna której wartości są liczbami a realizują się z określonym prawdopodobieństwem np. określonym przez rozkład normalny.

3.4 Odsetek kobiet wśród kandydatów na radnych

Dane dotyczące kandydatów na radnych do sejmików wojewódzkich zawierają także płeć kandydata. Ktoś może być ciekaw jaki był odsetek kobiet w tej grupie. Taki parametr nazywa się proporcją albo ryzykiem, a potocznie i niefachowo procentem. Matematycznym modelem jest **zmienna dwuwartościowa**, która z określonym prawdopodobieństwem przyjmuje wartość `kobieta`. Obliczmy empiryczną wartość tego prawdopodobieństwa jako liczbę kobiet do liczby wszystkich kandydatów. Wartość tego parametru wynosi 0.4587 (albo 45.87%). Potraktujmy to jako prawdziwą wartość prawdopodobieństwa (p), że kandydat jest kobietą i empirycznie sprawdźmy czy możemy szacować o prawdziwej wartości tego parametru używając (jako estymatora żeby się przyzwyczajać do nowych terminów) proporcji z próby. Tradycyjnie powtarzamy eksperyment 1000 razy dla trzech różnych wielkości próby. Rozkład otrzymanych wartości przedstawia rysunek.



Wnioski:

- Dla próby 20 elementowej rozkład nie przypomina rozkładu normalnego.
- Dla prób 120 i 420 elementowej rozkład jest podobny do normalnego.

- Zmienność estymatora maleje wraz ze wzrostem liczebności próby; każe nam to przypuszczać (i tak jest w istocie) że jest on zgodny.
- W każdym przypadku średnia z 1000 eksperymentów jest zbliżona do wartości prawdziwej; każe nam to przypuszczać (i tak jest w istocie) że estymator jest nieobciążony.

Rozkład normalny jest tak magiczny że nawet jeżeli zmienna, której parametr szacujemy nie ma rozkładu zbliżonego do normalnego (jak w przypadku zmiennej, która przyjmuje tylko dwie wartości) to i tak estymator tego parametru będzie normalny. Co najwyżej będziemy potrzebowali większej próby żeby „znormalniał'' (jak w opisywanym przykładzie).

3.5 Wnioskowanie statystyczne (*interferance*)

Analizując dane uzyskane z próby celem jest ich **uogólnienie** na całą populację. Przypominamy, że wnioskujemy o wartości parametru w populacji posługując się **estymatorem**. W przypadku wnioskowania o średniej estymatorem jest średnia-z-próby. Dobrze by było wiedzieć jak bardzo wiarygodna jest ta wartość (zwana oceną parametru) uzyskana na podstawie konkretnego estymatora, inaczej mówiąc jak dużo mogliśmy się pomylić.

Do oceny tej wiarygodności można użyć wariancji-średniej-z-próby (która nazywa się **wariancją błędu** albo **error variance**). Jeżeli wariancja błędu jest duża, to w pojedynczej próbie mogą wystąpić wartości znacznie różniące się od prawdziwej średniej; jeżeli jest mała to takie bardzo różniące się od prawdziwej średniej wartości mają małe szanse na zaistnienie. Do tego w przypadku rozkładu normalnego wiemy ze wariancja błędu jest równa s^2/n (gdzie s^2 jest wariancją w populacji a n wielkością próby.)

W ramach wnioskowania stosowane są trzy metody (podejścia):

- estymacja punktowa,
- estymacja przedziałowa,
- testowanie hipotez.

3.5.1 Estymacja punktowa

Szacujemy średnią (albo inny parametr) i tę wartość uznajemy za wartość prawdziwą; dokładność szacunku jest nieokreślona. Inaczej mówiąc wartość **estymatora** dla konkretnej próby przyjmujemy za ocenę parametru.

Estymatorem punktowym średniej jest średnia z próby a estymatorem punktowym proporcji/ryzyka jest proporcja/ryzyko z próby.

3.5.2 Estymacja przedziałowa

Nie można ustalić prawdopodobieństwa popełnienia błędu dla dokładnej wartości parametru (co wynika z właściwości matematycznych modelu), ale można dla dowolnego przedziału od--do.

Czyli nie można ustalić, że z prawdopodobieństwem 95% oszacujemy wartość średnią czegoś jako 5,000000, ale można z prawdopodobieństwem 95% oszacować **przedział**, w którym znajdzie się średnia (przykładowo, że będzie to przedział 4,9--5,1).

Estymacja przedziałowa to oszacowanie przedziału wartości od--do, który z zadanym z góry prawdopodobieństwem zawiera prawdziwą wartość parametru.

Z góry wyznaczone prawdopodobieństwo nazywa się **poziomem ufności** (określa jak często mamy się NIE rąbnąć).

3.5.3 Testowanie hipotez

Większość analiz statystycznych polega na porównaniu. W wyniku tego porównania otrzymujemy liczbę. Załóżmy, że mamy dwie próby dotyczące wieku kandydatów na radnych do sejmików wojewódzkich z roku 2018 (średnia 46,1) oraz z roku 2014 (47,2). Różnica wynosi 1,1 lat i może być spowodowana błędem przypadkowym (tj. gdybyśmy wylosowali jeszcze raz dwie próby, to wynik byłby zupełnie odmienny np 46,9 vs 46,5) i/lub wynikać z tego, że faktycznie w roku 2014 kandydaci byli starsi.

Formalnie stawiamy **hipotezę**, że różnica średnich wynosi zero. Jest to tzw. **hipoteza zerowa**. Niezbędne jest także postawienie **hipotezy alternatywnej**, którą może być proste zaprzeczenie zerowej. Zapisuje się to następująco (m_{14}/m_{18} oznacza odpowiednio średnie w latach 2014/2018):

H_0: różnica średnich wieku wynosi zero ($m_{14} = m_{18}$)

H_1: różnica średnich wieku jest różna od zera ($m_{14} \neq m_{18}$)

Hipotezy sprawdzamy wykorzystując **test statystyczny** czyli zmienną losową której rozkład prawdopodobieństwa zależy (jest funkcją powiedziałby matematyk) od wartości testowanych parametrów (w tym przypadku m_{14} oraz m_{18}). Tę zmienną losową nazywa się **statystyką testu**.

Nie jest chyba wielkim zaskoczeniem, że **statystyką testu** w teście różnicy średnich jest różnica średnich w próbie (poprawnie mówiąc różnica uwzględniająca liczebność próby oraz zmienność obu populacji). Całkiem **zdroworozsądkowo** możemy przyjąć, że duże różnice **statystyki testu** świadczą na rzecz hipotezy alternatywnej, natomiast małe na rzecz hipotezy zerowej.

Duża różnica pomiędzy **hipotezą** a wynikiem z próby może wynikać z tego, że

1. Pechowo trafiła nam się nietypowa próba, który zdarza się rzadko (rozkład normalny).
2. Hipoteza jest fałszywa, średnie mają inną wartość niż zakładamy w hipotezie zerowej.

Statystyk zawsze wybierze drugą wersję. Pozostaje tylko ustalić (dla statystyka) co to jest rzadko?

Rzadko to z prawdopodobieństwem mniejszym niż z góry ustalone prawdopodobieństwo otrzymania różnicy (zakładając, że hipoteza zerowa jest prawdziwa), którą otrzymaliśmy w próbie lub większej (coś jak założenie, że zrealizował się najlepszy z najgorszych scenariuszy).

Przyjmijmy przykładowo, że prawdopodobieństwo wystąpienia różnicy 1,1 lat (i większej) oszacowane na podstawie odpowiedniego modelu matematycznego (rozkład normalny) wynosi 0,3 co znaczy że coś takiego zdarza się względnie często -- trzy razy na 10 pobranych prób.

Załóżmy z kolei że, ta różnica wyniosła 3,2 lata. Prawdopodobieństwo wystąpienia takiej różnicy (i większej) wynosi 0,009 co znaczy że coś takiego zdarza się względnie rzadko -- 9 razy na tysiąc prób.

Przyjmując, że możemy się mylić 5 razy na 100 w pierwszym przypadku statystyk powie, że nie ma podstaw do odrzucenia hipotezy H_0. Różnica 1,1 lat wynika z przypadku. W drugim wypadku statystyk powie, że hipoteza jest fałszywa, bo zdarzyło się coś co nie powinno się zdarzyć.

Prawdopodobieństwo „graniczne'' ustalamy z góry i nazywa się ono **poziomem istotności**. Określa ono jak często możemy się rąbnąć **odrzucając hipotezę zerową, która jest prawdziwa**.

Ale jest jeszcze drugi przypadek popełnienia błędu: **przyjmujemy hipotezę zerową, która jest fałszywa**. W testach statystycznych nie określa się prawdopodobieństwa popełnienia tego błędu, a w związku z tym nie można **przyjąć hipotezy zerowej** (bo nie znamy ryzyka popełnienia błędu).

W konsekwencji hipotezę zerową albo się odrzuca albo **nie ma podstaw do odrzucenia**. Wniosek cokolwiek niekonkluzywny, ale tak jest.

Dlatego też często „opłaca się'' tak postawić hipotezę zerową aby ją następnie odrzucić, bo taki rezultat jest bardziej konkretny.

3.5.4 Testy nieparametryczne

Można testować hipotezy na temat wartości parametrów, ale można też testować przypuszczenia o charakterze mniej konkretnym. Na przykład, że dwie zmienne są niezależne (co to znaczy wyjaśniono w następnym rozdziale), albo że dwa rozkłady są podobne do siebie (rozkłady nie średnie). Takie hipotezy/testy określa się jako **nieparametryczne**. Przykładami są testy niezależności chi-kwadrat albo normalności Shapiro-Wilka (opisane w następnym rozdziale)

Oczywiste, ale powtórzmy: przypuszczenia o charakterze nieparametrycznym możemy tylko testować (sprawdzać hipotezy); nie obliczamy wtedy ani ocen ani nie wyznaczamy przedziałów ufności.

3.6 Słownik terminów które warto znać

Estymacja (punktowa, przedziałowa): szacowanie wartości parametru na podstawie próby.

Estymator (nieobciążony, zgodny, efektywny): funkcja na wartościach próby która służy do oszacowania parametru.

Hipoteza statystyczna: przypuszczenie dotyczące parametru lub rozkładu zmiennej.

Ocena (parametru): konkretna wartość estymatora dla pewnej próby.

Poziom istotności (testu; oznaczany jako α; zwykle 0,05): prawdopodobieństwo popełnienia błędu.

Poziom ufności = prawdopodobieństwo, że przedział ufności zawiera prawdziwą wartość parametru; oznaczany jako $1 - \alpha$; zwykle 0,95.

Rozkład (prawdopodobieństwa): przypisanie prawdopodobieństwa wartościom zmiennej losowej.

Test statystyczny: metoda weryfikacji hipotezy statystycznej.

Wnioskowanie statystyczne: wnioskowanie o całości na podstawie próby.

Rozdział 4

Analiza współzależności pomiędzy zmiennymi

Pomiędzy zjawiskami występują związki (zależności.) Nauki formułują te związki w postaci **praw**. Jak takie **prawo naukowe** powstaje? Typowo w dwu etapach, najpierw za pomocą **dedukcji** stawia się **hipotezę**, potem konfrontuje się hipotezę z danymi (podejście hipotetyczno-dedukcyjne). Na tym drugim etapie używa się statystyki (lub matematyki jeżeli prawo ma charakter deterministyczny)

Upraszczając *metoda hypodedukcji* sprowadza się do dedukcyjnego sformułowania hipotezy, która następnie jest empirycznie *falsyfikowana*, tj. próbuje się wykazać, że jest ona nieprawdziwa. Konsekwencje: nie można dowieść prawdziwości żadnej hipotezy, można natomiast wykazać, że hipoteza jest fałszywa.

Związki między cechami mogą być: **funkcyjne** (nauki przyrodnicze) -- wartościom jednej zmiennej odpowiada tylko jedna wartość drugiej zmiennej lub **stochastyczne** -- wartościom jednej zmiennej odpowiadają z pewnym przybliżeniem wartości innej zmiennej.

Problem: czy istnieje związek (zależność) pomiędzy cechami? Przykładowo czy istnieje związek pomiędzy paleniem (przyczyna) a chorobą nowotworową (skutek), wiekiem a prawdopodobieństwem zgonu z powodu COVID19 itd.

Jaki jest charakter zależności? Jaka jest siła zależności?

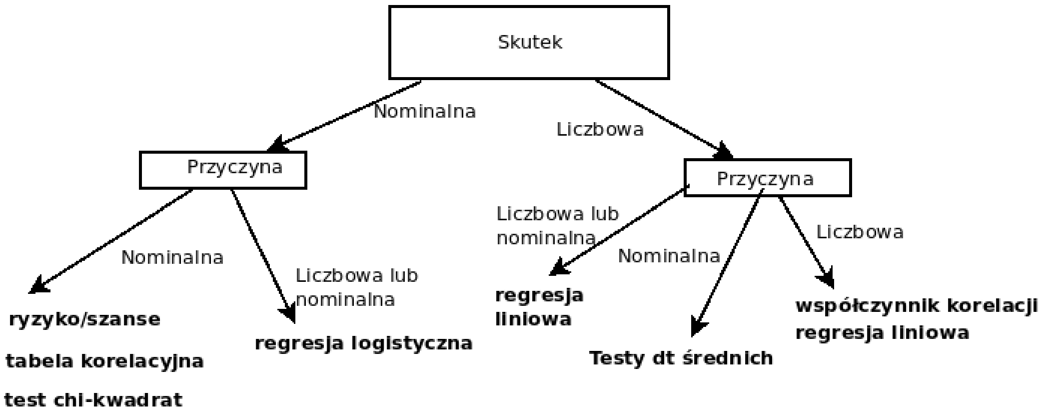
Rodzaj metod zastosowanej do empirycznej weryfikacji zależy w szczególności od sposobu pomiaru danych (nominalne, porządkowe, liczbowe.) co pokazano na rysunku 4.1.

Optymistyczną informacją jest że metod (oznaczonych krojem pogrubionym na diagramie), które omawiamy dalej w rodziale, jest raptem siedem czyli niedużo.

4.1 Dwie zmienne nominalne

4.1.1 Ryzyko względne oraz iloraz szans

Ryzyko to udział (iloraz) liczby sukcesów do liczby prób (zdarzeń pozytywnych/wyróżnionych do wszystkich). Zwykle podawany w procentach. Warto zauważyć że jest to empiryczny odpowiednik prawdopodobieństwa.



Rysunek 4.1: Metody statystycznej weryfikacji zależności pomiędzy zmiennymi

Podawanie witaminy C a przeziębienie/brak przeziębienia
Eksperyment przeprowadził Linus Pauling (laureat nagrody Nobla za odkrycie witaminy C). Eksperyment Paulinga polegał na tym, że podzielił 280 narciarzy na dwie grupy po 140 osób. Przez 5--7 dni podawał witaminę C jednej grupie oraz placebo drugiej grupie. Obserwował zachorowania na przeziębienie przez następne dwa tygodnie. Jeden narciarz nie dokończył eksperymentu. Historia milczy dlaczego :-)
W grupie 139 narciarzy, którym podano witaminę C (grupa C) zachorowało 17. W grupie 140 narciarzy, którym podano placebo (grupa P) zachorowało 31. Zatem:

- Ryzyko zachorowania w grupie C wyniosło 17/139 = 12,2%.
- Ryzyko zachorowania w grupie P wyniosło 31/140 = 22,14%

Na tzw. chłopski rozum jeżeli witamina C **nie działa**, to powinien zachorować ten sam odsetek narciarzy w obu grupach. A tak nie jest jak widać…

Prostymi miarami oceny siły zależności mogą być:

- różnica ryzyk (**risk difference**)
- ryzyko względne (**relative risk**), oraz
- iloraz szans (**odds ratio**).

Jeżeli r_e oznacza ryzyko w grupie eksperymentalnej (test group; grupa narażona/exposed group), a r_k w grupie kontrolnej (control group; grupa nienarażona/unexposed), to **różnica ryzyk** to po prostu $r_e - r_k$. W przykładzie będzie to $22,14 - 12,2 = -9,94\%$ Ta miara aczkolwiek prosta jest rzadko stosowana.

Znacznie częściej używa się **ryzyka względnego** definiowanego jako $RR = r_e/r_k$. W przykładzie będzie to $12,2/22,14 = 0,55$. Podanie witaminy C zmniejsza ryzyko zachorowania o prawie połowę. Oczywiste jest, że $RR < 1$ oznacza zmniejszenie ryzyka; $RR > 1$ oznacza zwiększenia; $RR = 1$ oznacza brak zależności.

Zamiast ryzyka (czyli ilorazu liczby sukcesów do liczby prób) można używać pojęcia szansa/szansy (**odds**) definiowanego jako iloraz sukcesów do porażek.

Jeżeli o_e oznacza szanse w grupie eksperymentalnej a o_k w grupie kontrolnej, to **iloraz szans** (*odds ratio*), jest definiowany jako stosunek $\mathrm{OR} = o_e/o_k$.

Przykładowo jeżeli w dwukrotnym rzucie monetą otrzymano orła i reszkę to ryzyko otrzymania orła wynosi 1/2 = 0,5 a szansa otrzymania orła wynosi 1/1 = 1.

Narciarze Paulinga (kontynuacja)
Ryzyko zachorowania w grupie C wynosi 12,2 (jak wiemy); natomiast szansa, że narciarz grupie C zachoruje wynosi 17/122 = 13,9%. (A w grupie P wynosi 28,44%).
Jak widać dla dużych ryzyk (rzut monetą) szansa różni się znacznie od prawdopodobieństwa, ale dla małych ryzyk obie miary mają zbliżoną wartość.
Zatem iloraz szans dla narciarzy wyniesie 13,9/28,44 = 0,48. Podanie witaminy C zmniejsza szansę na zachorowanie o ponad połowę. Albo 1/0,48 = 2,04, narciarz który nie brał witaminy C ma ponad dwukrotnie większą szansę na zachorowanie.

Właściwości ilorazu szans:

- jeżeli równe 1 to sukces/porażka równie prawdopodobne;
- jeżeli większe od 1 to sukces bardziej prawdopodobny;
- jeżeli jest mniejsze od 1 to porażka jest bardziej prawdopodobna.

Dane w badaniach wykorzystujących ryzyko/szanse mają często postać tabeli dwudzielnej o wymiarach 2×2, którą można przestawić następująco (a, b, c i d to liczebności):

Grupa	sukces	porażka
grupa kontrolna	a	b
grupa eksperymentalna	c	d

Dla danych w tej postaci:

- RR $= c(a+b)/a(c+d)$ oraz
- OR $= (ad)/(bc)$

Narciarze Paulinga (tabela dwudzielna)

Grupa	katar	zdrowy
grupa C	17	122
grupa P	31	109

4.1.2 Przedziały ufności dla ryzyka względnego oraz ilorazu szans

Ryzyko, ryzyko względne czy iloraz szans to parametry podobne do procentu kobiet wśród kandydatów na radnych z przykładu w poprzednim rozdziale. Wiemy, że estymatorem punktowym proporcji jest proporcja z próby. Nie będzie wielkim odkryciem, że estymatorem punktowym ryzyka jest ryzyko z próby, ryzyka względnego/ilorazu szans zaś ryzyko względne/iloraz szans z próby.

Standardem jest obliczanie dla ryzyka względnego oraz ilorazu szans oprócz ocen punktowych także przedziałów ufności czyli podawania dwóch wartości, pomiędzy którymi z zadanym prawdopodobieństem znajduje się nieznana wartość szacowanego parametru.

Narciarze Paulinga (przedziały ufności)
Końce przedziałów ufności dla ilorazu szans (ocena punktowa 0.4899524) wynoszą: [0.2569389; 0.934282] zaś dla ryzyka względnego (ocena punktowa 0.5523323) przedział ufności wynosi [0.3209146; 0.9506298].

Uwaga: nie jest specjalnie istotne jaka jest konkretna formuła obliczania przedziałów ufności, przecież obliczenia i tak koniec-końców wykona program komputerowy.

Przedział ufności dla ilorazu szans nie zawiera 1; zatem branie witaminy C zmniejsza szanse na zachorowanie; albo zwiększa na niezachorowanie od $1/25 = 4$ do $1/0,9 = 1,1$. Żeby to zabrzmiało ładnie i po polsku. Zwiększa na niezachorowanie od 300% do 10%.

Dlaczego taka znacząca rozpiętość? Bo próba jest względnie mała. Gdyby Pauling zwerbował nie 280 a 2800 narciarzy mógłby weryfikować działanie swojej witaminy z większą pewnością.

4.1.3 Tabele wielodzielcze

Łączny rozkład dwóch lub większej liczby zmiennych można przedstawić w tabeli. Taka tabela nazywa się dwudzielcza (dla dwóch zmiennych) lub wielodzielcza albo wielodzielna (dla więcej niż dwóch liczby zmiennych.) Inne nazwy tych tabel to krzyżowe albo kontyngencji (*cross-tabulation*, *contingency* albo *two-way tables*).

Ograniczmy się do analizy tabel dwudzielnych.

Narciarze Paulinga (kontynuacja)
Eksperyment Paulinga można przedstawić w postaci tablicy dwudzielczej (`P/C` oznacza czy narciarz zażywał witaminę czy placebo; `cold/nocold` czy zachorował czy nie zachorował na katar):

	nocold	cold	razem
C	122	17	139
P	109	31	140
Sum	231	48	279

Taka tabela składa się z wierszy i kolumn. Dolny wiersz (Sum czyli Razem po polsku) zawiera łączną liczebność dla wszystkich wierszy w danej kolumnie. Podobnie prawa skrajna kolumna zawiera łączną liczebność dla wszystkich kolumn dla danego wiersza. Dolny wiersz/Prawą kolumnę nazywamy **rozkładami brzegowymi**. Pozostałe kolumny oraz wiersze nazywane są **rozkładami warunkowymi**. Rozkładów warunkowych jest tyle ile wynosi suma $r + c$ gdzie r to liczba wariantów jednej cechy a c to liczba wariantów drugiej cechy.
Przy warunku że narciarz brał witaminę C, 122 takich osób nie zachorowało (**nocold**) a 17 zachorowało (**cold**). Drugi rozkład warunkowy: 109 narciarzy, którzy brali placebo nie zachorowało, a 31 zachorowało. Są także rozkłady warunkowe dla drugiej cechy. W grupie narciarzy, którzy zachorowali 122 brało witaminę C, a 109 brało placebo. Wreszcie w grupie narciarzy, którzy nie zachorowali 109 brało witaminę C, a 31 brało placebo. Rozkładów warunkowych jest 4 bo obie cechy mają po dwa warianty. Jest to najmniejsza możliwa tabela wielodzielcza.
Zamiast liczebności można posługiwać się odsetkami (procentami):

	N	Y	Sum
C	43.73	6.09	49.82
P	39.07	11.11	50.18
Sum	82.80	17.20	100.00

Narciarze, którzy brali witaminę C oraz nie zachorowali stanowią 43.73% wszystkich narciarzy. Mało przydatne…
Ciekawsze jest obliczenie procentów każdego wiersza osobno, tj. dzielimy liczebności w każdej kolumnie przez liczebności rozkładu brzegowego (wartości ostatniej kolumny):

	N	Y	
C	87.77	12.23	100
P	77.86	22.14	100
n.m	82.80	17.20	100

Otrzymaliśmy ryzyka zachorowania na katar (lub nie zachorowania). Ryzyko zachorowania dla całej grupy wynosi 17.20% a nie zachorowania 82.80%. Jest przyznajmy całkiem **zdroworozsądkowym założeniem** (uczenie hipotezą statystyczną), że jeżeli przyjmowanie witaminy nie ma związku z zachorowaniem lub nie na katar, to w grupie tych co brali i tych co nie brali powinniśmy mieć identyczne rozkłady warunkowe równe rozkładowi brzegowemu. Czyli powinno przykładowo zachorować 17.20% narciarzy, którzy brali witaminę C a widzimy , że zachorowało jedynie 12.23%.
Na oko księgowego witamina C działa (bo są różnice), ale dla statystyka liczy się czy ta różnica jest na tyle duża, że (z założonym prawdopodobieństwem) można wykluczyć działanie przypadku. Rozumowanie jest następujące: jeżeli prawdopodobieństwo wystąpienia tak dużej różnicy jest małe, to cechy nie są niezależne. Jest to istota i jedyny wniosek z czegoś co się nazywa testem istotności-chi-kwadrat. Test chi-kwadrat porównuje liczebności tablicy wielodzielnej z idealną-tablicą-wielodzielną, która zakłada niezależność jednej zmiennej od drugiej.
Można udowodnić, że taka idealna tablica powstanie przez przemnożenie dla każdego elementu tablicy odpowiadających mu wartości brzegowych a następnie podzieleniu tego przez łączną liczebność (czyli przykładowo pierwszy element poniższej „idealnej'' tablicy to 231 pomnożone przez 139 i podzielone przez 279; proszę sprawdzić, że jest to 115.086):

	N	Y	Sum
C	115.086	23.914	139
P	115.914	24.086	140
Sum	231.000	48.000	279

Proszę zwrócić uwagę że **rozkłady brzegowe** są identyczne, identyczna jest też łączna liczebność. Różnią się tylko rozkłady warunkowe (które nie są liczbami całkowitami ale tak ma być--nie jest to błąd)
Za pomocą testu chi-kwadrat obliczamy jakie jest prawdopodobieństwo wystąpienia tak dużych lub większych różnic. Wynosi ono 0.041864. Czyli wystąpienie tak dużych różnic pomiędzy **oczekiwanymi** (przy założeniu o niezależności zmiennych) liczebnościami a obserwowanymi liczebnościami zdarza się około 4 razy na 100.

Jeszcze raz przypominamy ideę testu: jeżeli prawdopodobieństwo zaobserwowanych różnic jest małe to zakładamy że

- albo mamy pecha i pięć razy podrzucając monetą zawsze nam spadła reszka (prawdopodobieństwo około 0,03), albo
- że założenie co do niezależności jest fałszywe.

Statystyk zawsze wybierze drugie. Pozostaje tylko ustalenie co to znaczy **małe**.

Małe to takie które jest mniejsze od arbitralnie przyjętego przez statystyka. Zwykle jest to 0,05 lub 0,01 (czasami 0,1) co oznacza że odrzucając założenie o braku związku pomiędzy katarem a braniem witaminy C pomylimy się pięć lub raz na 100.

Uwaga: proszę zwrócić uwagę że wniosek z testu niezależności jest słabszy niż z porówania ryzyk. Tam mamy informację że zależność istnieje i oszacowaną jej wielkość (np. za pomocą ryzyka względnego) tutaj tylko zweryfikowaliśmy fakt czy obie zmienne są niezależne czy też nie.

Palenie a status społeczno-ekonomiczny
Dla pewnej grupy osób odnotowujemy ich status-społeczno-ekonomiczny (wysoki/**high**, średni/**middle**, niski/**low**) oraz status-względem-palenia (wartości: pali/**current**, palił-nie-pali/**former**, nigdy-nie-palił/**never**). Obie zmienne są nominalne, obie mają po trzy wartości. Można poklasyfikować wszystkich badanych w następujący sposób:

	High	Low	Middle	Sum
current	51	43	22	116
former	92	28	21	141
never	68	22	9	99
Sum	211	93	52	356

Uwaga: status-społeczno-ekonomiczny to powiedzmy miara prestiżu używana w socjologii (można na Wikipedii doczytać co to dokładnie jest).
Tym razem tabela składa się z 3 wierszy i 3 kolumn (ostatni wiersz/kolumna się nie liczą bo to sumy--rozkłady brzegowe)
Przedstawmy tą tabelę w postaci udziałów procentowych sumujących się dla każdego wiersza osobno do 100% (tj. dzielimy liczebności w każdej kolumnie przez liczebności rozkładu brzegowego (wartości ostatniej kolumny):

	High	Low	Middle	
current	43.96552	37.06897	18.965517	100
former	65.24823	19.85816	14.893617	100
never	68.68687	22.22222	9.090909	100
n.m	59.26966	26.12360	14.606742	100

Rozumowanie jest identyczne jak dla narciarzy Pauliga. Jeżeli nie ma zależności pomiędzy paleniem a statusem to procenty w ostatnim wierszu powinny być identyczne jak w wierszach 1--3 (nagłówka nie liczymy). Tym idealnym procentom odpowiadają następujące liczebności:

	High	Low	Middle	Sum
current	68.75281	30.30337	16.94382	116
former	83.57022	36.83427	20.59551	141
never	58.67697	25.86236	14.46067	99
Sum	211.00000	93.00000	52.00000	356

Wartość prawdopodobieństwa dla testu chi-kwadrat określająca, że przy założeniu niezależności obu zmiennych tak duża różnica między liczebnościami rzeczywistymi a idealnymi (porównaj stosowne tabele wyżej) jest dziełem przypadku wynosi 0.000981. Jest to prawdopodobieństwo tak małe, że statystyk odrzuca założenie o niezależności statusu i palenia (myląc się w przybliżeniu 0.000981 ≈ raz na tysiąc)

4.2 Zmienna liczbowa i zmienna nominalna

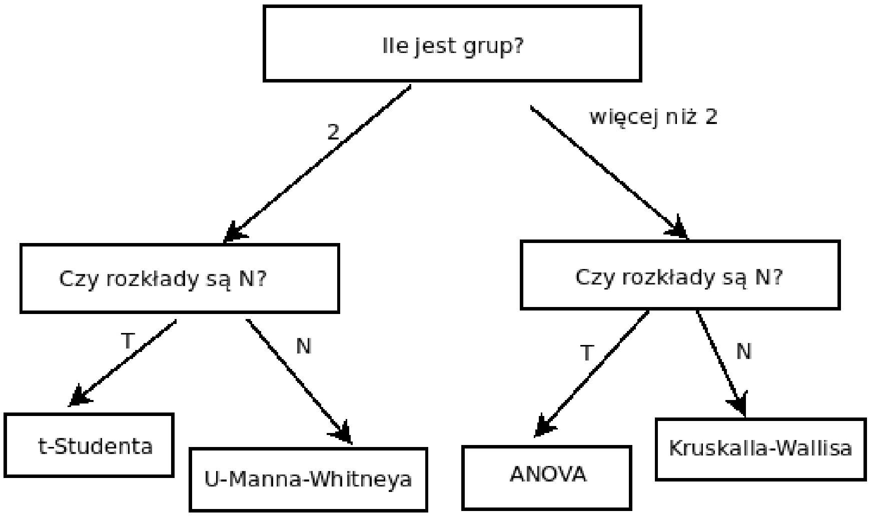
Obliczamy średnie wartości zmiennej liczbowej **w grupach** określonych przez wartości zmiennej nominalnej, np. wypalenie zawodowe w podziale na miejsce pracy. Grup może być dwie lub więcej.

Stawiamy hipotezę, że wartości średnie w każdej grupie są równe, wobec hipotezy alternatywnej że tak nie jest (że są różne jeżeli grup jest dwie; co najmniej jedna jest różna jeżeli grup jest więcej niż dwie). Stosujemy odpowiedni test statystyczny:

- jeżeli liczba grup wynosi 2 oraz można przyjąć założenie o przybliżonej normalności rozkładów, to stosujemy test t-Studenta (dla prób niezależnych);
- jeżeli liczba grup wynosi 2, ale nie można założyć normalności rozkładów to stosujemy test U-Manna-Whitneya;

- jeżeli liczba grup jest większa niż dwie oraz można przyjąć założenie o normalności rozkładów to stosujemy test pn. ANOVA;
- jeżeli liczba grup jest większa od dwóch oraz nie można przyjąć założenia o normalności rozkładów, to stosujemy test Kruskal-Wallisa.

Powyższe w postaci diagramu ze strzałkami przedstawiono na rysunku 4.2.



Rysunek 4.2: Testowanie istotości różnicy pomiędzy średnimi

Każdy z testów jest interpretowany identycznie:

1. Obliczana jest wartość statystyki testu t_k.
2. Obliczane jest prawdopodobieństwo $t \geq t_k$ czyli przyjęcia przez statystykę testu t równej lub większej od t_k (co do wartości bezwzględnej). To prawdopodobieństwo zwyczajowo oznacza się literą p albo p-value (czyli wartość p).
3. Jeżeli p jest mniejsze/równe od przyjętego poziomu istotności to hipotezę zerową odrzucamy; jezeli p jest większe od przyjętego poziomu istotności to nie ma podstaw do odrzucenia hipotezy zerowej.

Odrzucenie hipotezy zerowej oznacza, że istnieje związek pomiędzy jedną a drugą zmienną. Jeżeli nie ma podstaw do odrzucenia hipotezy zerowej to oznacza to że takiej zależności nie udało nam się wykazać.

Omawiając wynik należy podać się wartość t_k oraz p. Statystyka testu może się różnie nazywać i być oznaczana różnym symbolem, np.: t (test t-Studenta), U (test U Manna-Whitneya).

4.2.1 Test t-Studenta

Test stosujemy jeżeli porównujemy dwie średnie oraz można przyjąć założenie że rozkład wartości w obu grupach jest normalny.

Poziom depresji a miejsce pracy
Studenci pielęgniarstwa i ratownictwa PSW w 2023 roku wypełnili ankietę zawierającą test depresji Becka, mierzący **poziom depresji** (wartość liczbowa) oraz pytanie o rodzaj miejsca pracy (skala nominalna). Poniżej zestawiono średnie wartości **poziomu depresji** w podziale na

rodzaj miejsca pracy (szpital/przychodnia).

m-pracy	średnia	n
Przychodnia	7.833333	12
Szpital	8.450549	91

Kolumna n zawiera liczbność.

Średnie różnią się o 0.62. Pytanie czy to dużo czy mało?

Przyjmijmy (na razie bez sprawdzania), że rozkłady wartości poziomu depresji w obu grupach są (w przybliżeniu) normalne. Można zatem zastosować test *t*-Studenta

Grupa1	Grupa2	n1	n2	t	p
Przychodnia	Szpital	12	91	-0.3241142	0.749

Kolumna t zawiera wartość statystyki testu t_k. Kolumna p zawiera oczywiście wartość prawdopodobieństwa p.

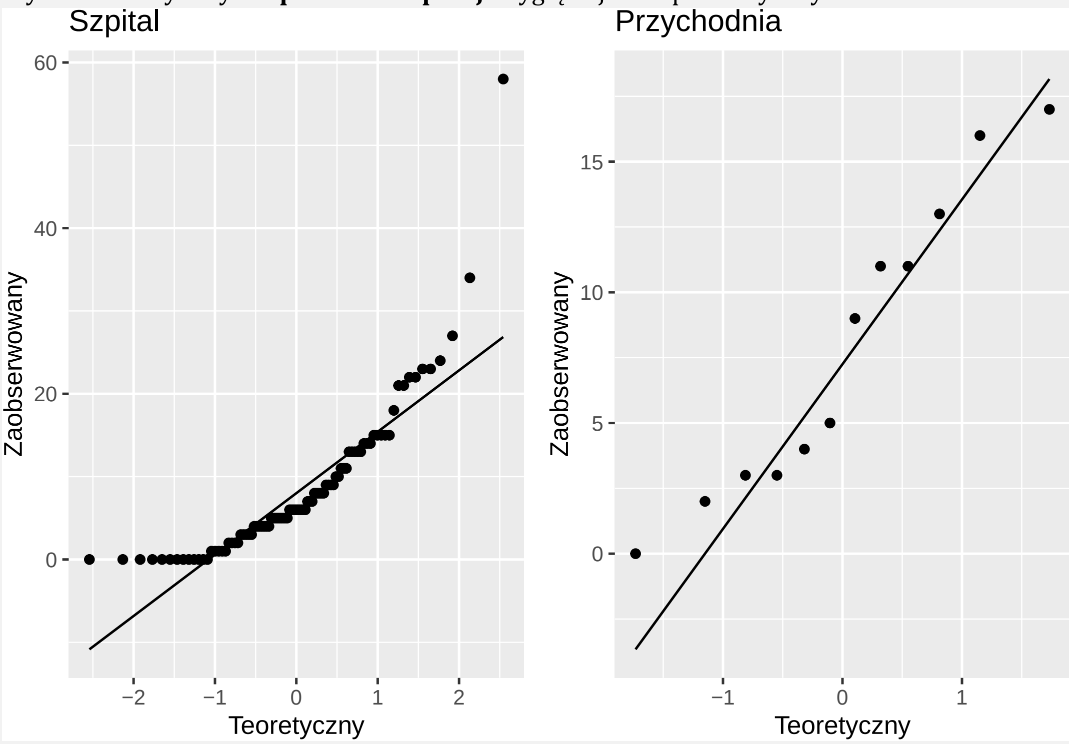
Ponieważ wartość p równa 0.749‘ jest większa od każdego zwyczajowo przyjmowanego poziomu istotności (0,05 na przykład, albo 0,1) nie ma podstaw do odrzucenia hipotezy, że średnie w obu grupach są równe. Skoro tak, to w konsekwencji stwierdzamy że pomiędzy poziomem depresji a miejscem pracy nie ma zależności.

4.2.2 Testowanie normalności

Statystyk nie przyjmuje założeń na słowo honoru. Kiedy zatem można przyjąć założenie o normalności a kiedy nie? Można to ocenić na podstawie wykresu kwantylowego. Oraz posługując się testem Shapiro-Wilka (bo statystycy na każde pytanie mają zawsze **jakiś** stosowny test).

Poziom depresji a miejsce pracy

Wykres kwantylowy dla **poziomu depresji** wygląda jak na poniższym rysunku



Prosta odpowiada teoretycznym wartościom kwantyli rozkładu poziomu depresji przy założeniu że mają one rozkład normalny. Punkty odpowiadają zaobserwowanym wartościom kwantyli. Im bardziej punkty nie pokrywają się z prostą (zwłaszcza na skrajach rozkładu) tym mniej wierzymy, że rozkład jest normalny.
W tym przypadku wygląda, że rozkład w grupie Szpital **nie jest** normalny. W grupie Przychodnia jest lepiej ale jednocześnie to lepiej jest mało wiarygodne z uwagi na małą liczebność grupy (zaledwie 12).
Wizualne obserwacja można potwierdzić stosując test Shapiro-Wilka (S-W). Interpretacja tego testu jest „standardowa", mianowicie małe wartości p świadczą przeciwko hipotezie zerowej (że rozkład jest normalny).

m-pracy	S-W	p
Przychodnia	0.9256178	0.3359655
Szpital	0.7865090	0.0000000

Rozkład w grupie `szpital` nie jest normalny (o czym świadczy niska wartość p). Nasze założenie co do normalności było niepoprawne i należy do weryfikacji hipotezy o równości średniej zamiast testu t-Studenta zastosować test U Manna-Whitneya.
Kolumna S-W zawiera wartości statystki testu S-W oczywiście.

4.2.3 Test U Manna-Whitneya

Poziom depresji a miejsce pracy
Ponieważ grup jest dokładnie 2 a rozkład nie jest normalny, stosujemy test U Manna-Whitneya.

Grupa1	Grupa2	n1	n2	U	p
Przychodnia	Szpital	12	91	564.5	0.853

Prawdopodobieństwo wystąpienia tak dużej różnicy przy założeniu, że średnie w obu grupach są identyczne wynosi 0.853 (różnica jest zatem nieistotna; obie średnie są identyczne--nie ma zależności). Kolumna U zawiera wartość statystyki testu U. Przypominamy, że dobry zwyczaj nakazuje podawać tę wartość omawiając wynik testu (więc ją podajemy).

4.2.4 Test ANOVA

Jeżeli liczba grup jest większa niż dwie ale można przyjąć założenie o normalności rozkładów to stosujemy test ANOVA.

Poziom depresji a staż pracy
W ankiecie, którą wypełnili Studenci pielęgniarstwa i ratownictwa PSW w 2023 roku było też pytanie o staż pracy. Oryginalną liczbową wartość zmiennej staż zamieniono na zmienną w skali nominalnej o następujących czterech wartościach: `<6` (oznacza od 0 do 6 lat stażu pracy), `07-12` (7--12 lat), `13-18` (13--18 lat) oraz `>19` (19 i więcej lat.)

staż (kategoria)	średnia	n
07-12	7.857143	7
13-18	7.666667	12
<06	8.512821	39
>19	8.533333	45

Zakładając, że rozkłady w grupach są normalne, do weryfikacji hipotezy o równości wszystkich średnich możemy zastosować test ANOVA. Na poniższym wydruku kolumna F zawiera wartość statystki test ANOVA a kolumna p jak zwykle wartość prawdopodobieństwa p:

```
## ANOVA Table (type II tests)
##
##   Effect DFn DFd     F     p p<.05   ges
## 1   staz   3  99 0.043 0.988       0.001
```

Wartość p równa 0.988 świadczy że nie istotnych różnic pomiędzy średnimi, co oznacza że pomiędzy poziomem depresji a kategoriami stażu pracy nie ma zależności.
Czy zastosowanie testu ANOVA było poprawne? Żeby się o tym przekonać trzeba zastosować (znowu) test Shapiro-Wilka:

m-pracy	S-W	p
07-12	0.8565271	0.1408865
13-18	0.7596157	0.0033736
<06	0.9008198	0.0023292
>19	0.6780397	0.0000000

Wobec takiego wyniku testu do oceny istotności różnic należy zastosować bardziej ogólny test Kruskala-Wallisa

4.2.5 Test Kruskala-Wallisa

Poziom depresji a staż pracy
Na poniższym wydruku wartość statystyki testu jest oznaczona jako `Kruskal-Wallis chi-squared` a wartość p symbolem `p-value`:

```
##
##  Kruskal-Wallis rank sum test
##
## data:  P by staz
## Kruskal-Wallis chi-squared = 1.5145, df = 3, p-value = 0.6789
```

Prawdopodobieństwo tak dużych różnic w wartościach średnich przy założeniu, że średnie we wszystkich grupach są identyczne wynosi 0.678922994359164 (różnice są zatem nieistotne; wszystkie średnie są identyczne--nie ma zależności)

4.3 Dwie zmienne liczbowe

4.3.1 Korelacyjny wykres rozrzutu

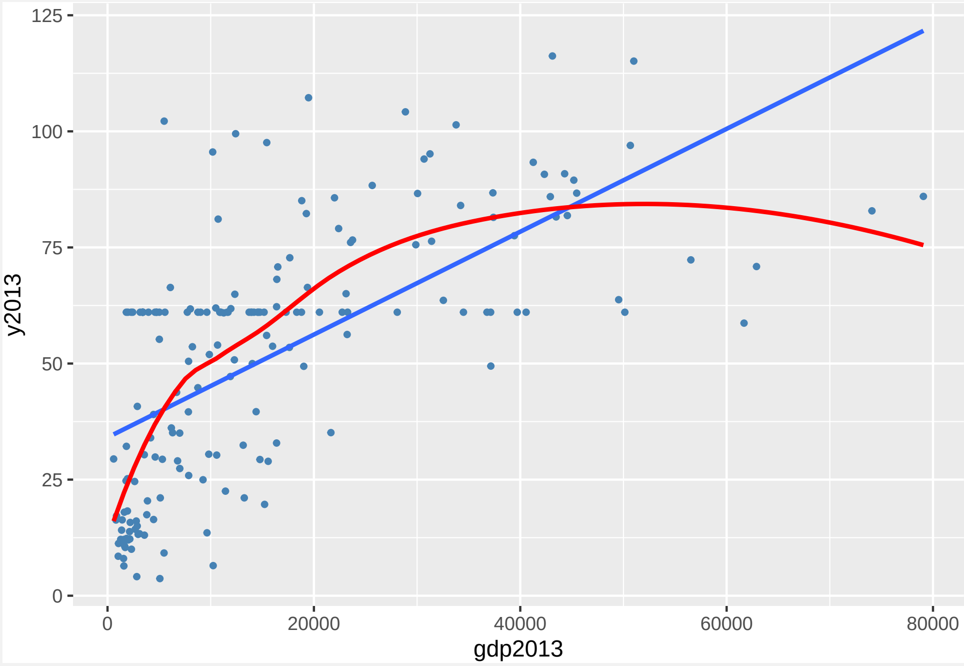
Wykres rozrzutu (*scatter plot*) znany także jako korelogram, albo wykres XY, to prosty wykres kreślony w układzie kartezjańskim, w którym każdej obserwacji (składającej się z dwóch liczb) odpowiada kropka o współrzędnych XY.

O występowaniu związku świadczy układanie się kropek według jakiegoś kształtu (krzywej). O braku związku świadczy chmura punktów niepodobna do żadnej krzywej.

Punkty układające się według prostej świadczą o zależności liniowej (wyjątek: linia pozioma lub pionowa o czym dalej) zaś punkty układające się według krzywej świadczą o zależności nieliniowej.

Zamożność a konsumpcja mięsa
Organizacja Narodów Zjednoczonych do spraw Wyżywienia i Rolnictwa znana jako FAO udostępnia dane dotyczące konsumpcji żywności na świecie (https://www.fao.org/faostat/en/#home). Bank światowy udostępnia dane dotyczące dochodu narodowego (https://data.worldbank.org/).

Konsumpcja mięsa jest mierzona jako średnia konsumpcja w kilogramach w każdym kraju (*per capita* się mówi); Dochód podobnie jako średnia wielkość dochodu narodowego *per capita*. Dane dotyczą roku 2013.



Przy dużej dozie wyobraźni można dostrzec relację liniową pomiędzy konsumpcją mięsa a GDP co oznaczono na wykresie linią prostą. Można też założyć, że relacja pomiędzy konsumpcją mięsa a GDP ma charakter nieliniowy (linia krzywa). Liniowa czy nieliniowa, relacja jest na pewno mocno przybliżona co jest najbardziej pewnym wnioskiem, który można wysnuć z wykresu rozrzutu.

4.3.2 Pomiar siły zależności: współczynnik korelacji liniowej Pearsona

Kowariancja to średnia arytemtyczna iloczynów odchyleń wartości zmiennych X, Y od ich wartości średnich. Dla n obserwacji na zmiennych X oraz Y można powyższe zapisać w postaci następującej formuły:

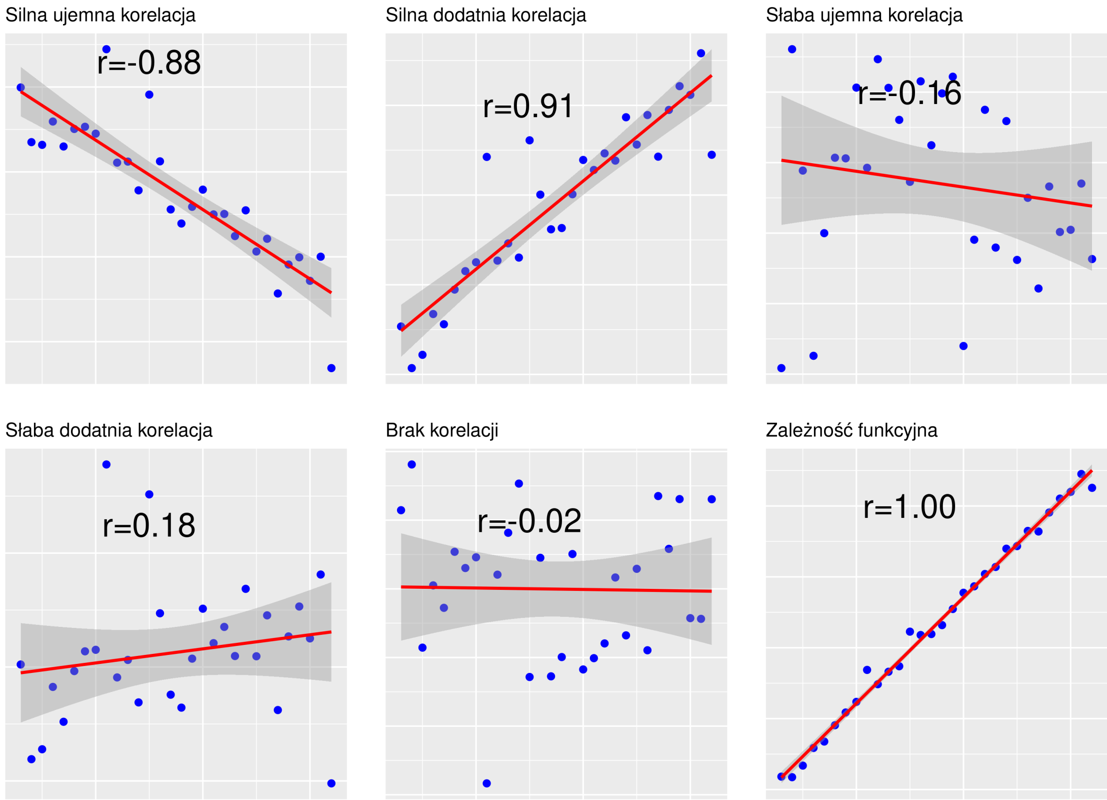
$$\mathrm{cov}(xy) = \frac{1}{n}((x_1 - \bar{x})(y_1 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y}))$$

Kowariancja zależy od rozproszenia (im większe tym większa), ma też dziwną jednostkę (jednostkaX · jednostkaY) oraz zależy od wybranych skal (tony vs gramy na przykład.)

Z powyższych powodów do pomiaru związku pomiędzy cechami używa się standaryzowanego współczynnika kowariancji, zwanego **współczynnikiem korelacji liniowej**, (*Pearson linear correlation coefficient*). Standaryzacja polega na podzieleniu wartości kowariacji przez iloczyn odchyleń standardowych s_x oraz s_y.

$$r_{xy} = \frac{\mathrm{cov}(xy)}{s_x \cdot s_y}$$

Współczynnik jest miarą niemianowaną, przyjmującą wartości ze zbioru $[-1; 1]$; Skrajne wartości ± 1 świadczą o związku funkcyjnym (wszystkie punkty układają się na linii prostej); wartość zero świadczy o braku związku co odpowiada linii poziomej lub pionowej (por. rysunek 4.3).



Rysunek 4.3: Wykresy rozrzutu dla korelacji o różnej sile

Interpretacja opisowa: wartości powyżej 0,9 świadczą o silnej zależności.

Zamożność a konsumpcja mięsa (kontynuacja)
Współczynnik korelacji liniowej wynosi 0.6823158 (umiarkowana korelacja).
Czy ta wartość jest istotnie różna od zera? Jest na to stosowny test statystyczny, który sprowadza się do określenia jakie jest prawdopodobieństwo otrzymania r = 0.6823158 przy założeniu że prawdziwa wartość r wynosi zero. Otóż w naszym przykładzie to prawdopodobieństwo wynosi 3.850676e-26 (czyli jest ekstremalnie małe -- r jest istotnie różne od zera).

4.3.3 Macierz korelacji

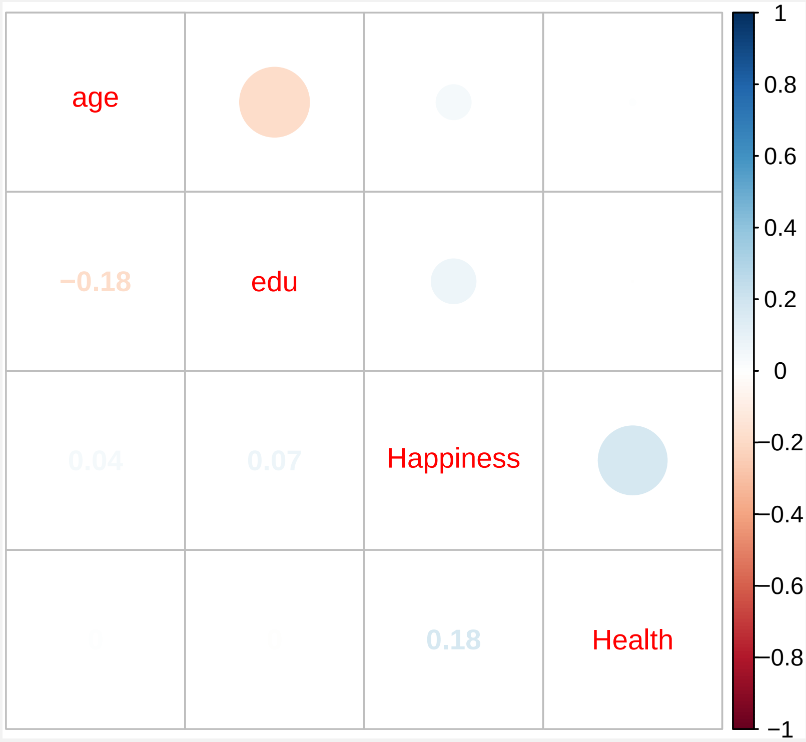
Wstępnym etapem analizy zależności między zmiennymi jest często hurtowa ocena współczynników korelacji w postaci kwadratowej **macierzy korelacji**.

Korelacja pomiędzy wiekiem, edukacją, szczęściem a stanem zdrowia
Mohammadi S. i inni badali zależność pomiędzy wiekiem, poziomem edukacji, szczęściem a stanem zdrowia. (The relationship between happiness and self-rated health: A population-based

study of 19499 Iranian adults; https://doi.org/10.1371/journal.pone.0265914)

```
##                  age          edu  Happiness         Health
## age        1.00000000 -0.18341325501 0.04491863  0.00125622963
## edu       -0.18341326  1.00000000000 0.07418519 -0.00003728405
## Happiness  0.04491863  0.07418519038 1.00000000  0.17863069296
## Health     0.00125623 -0.00003728405 0.17863069  1.00000000000
```

Albo w bardziej efektownej postaci tekstowo-graficznej:



Ze wszystkich zmiennych analizowanych w badaniu Mohammadiego i innych jedynie zależność pomiędzy wiekiem a wykształceniem (raczej trywialna) oraz szczęściem i zdrowiem (raczej oczywista) okazały się znacząco różne od zera.

4.3.4 Pomiar siły zależności: regresja liniowa

Regresja liniowa zakłada, że istnieje związek przyczyna-skutek i ten związek można opisać linią prostą (stąd liniowa). Skutek jest jeden i nazywa się go **zmienną zależną** a przyczyn może być wiele i noszą nazwę **zmiennych niezależnych** (albo **predyktorów**). W przypadku gdy związek dotyczy dwóch zmiennych mówi się o **regresji prostej**. Przykładowo zależność pomiędzy spożywaniem kawy w czasie sesji egzaminacyjnej a wynikiem egzaminu można formalnie zapisać jako:

$$\text{wynik} = b_0 + b_1 \cdot \text{kawa}$$

Współczynnik b_1 określa wpływ spożycia kawy na wynik egzaminu. W szczególności jeżeli $b_1 = 0$ to nie ma związku między spożywaniem kawy a wynikiem egzaminu.

4.3.5 Regresja prosta

Równanie regresji dla zmiennych Y (skutek) oraz X (przyczyna) można zapisać następująco:

$$Y = b_0 + b_1 \cdot X + e$$

$Y = b_0 + b_1 \cdot X$ to **część deterministyczna**, a e oznacza **składnik losowy**. O tym składniku zakładamy, że średnia jego wartość wynosi zero. Można to sobie wyobrazić, że w populacji jest jakaś prawdziwa zależność $Y = b_0 + b_1 \cdot X$ pomiędzy X a Y, która w próbie ujawnia się z błędem o charakterze losowym. Ten błąd może wynikać z pominięcia jakiejś ważnej zmiennej (model to zawsze uproszczenie rzeczywistości), przybliżonego charakteru linii prostej jako zależności pomiędzy X a Y (prosta ale nie do końca prosta) albo błędu pomiaru.

Współczynnik b_1 (nachylenia prostej) określa wielkość efektu w przypadku regresji, tj. siły zależności pomiędzy zmiennymi.

Współczynnik b_1 ma prostą interpretację: jeżeli wartość zmiennej X rośnie o jednostkę to wartość zmiennej Y zmienia się przeciętnie o b_1 jednostek zmiennej Y. Wyraz wolny zwykle nie ma sensownej interpretacji (formalnie jest to wartość zmiennej Y dla $X = 0$)

Oznaczmy przez y_i wartości obserwowane (zwane też empirycznymi) a przez $\hat{y}_i$ *wartości teoretyczne* (leżące na prostej linii regresji).

Wartości b_0 oraz b_1 wyznacza się minimalizując sumę kwadratów odchyleń wartości teoretycznych od wartości empirycznych, tj.:

$$(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + ... + (\hat{y}_n - y_n)^2$$

Rozwiązując powyższy **problem minimalizacyjny** otrzymujemy wzory definiujące parametry b_0 oraz b_1. Metoda wyznaczania parametrów linii prostej w oparciu o minimalizację sumy kwadratów odchyleń nosi nazwę **metoda najmniejszych kwadratów**.

Przypominamy, że **estymatorem** nazywamy metodę oszacowania parametru na podstawie próby. Ponieważ traktujemy b_0 oraz b_1 jako parametry jakieś populacji generalnej to wzory na b_0 oraz b_1 statystyk nazwie estymatorami parametrów b_0 oraz b_1. W konsekwencji tego b_0 oraz b_1 posiadają jakąś wartość średnią oraz wariancję.

Przypominamy dalej, że wartość średnia **dobrego estymatora** powinna wynosić zero (bo wtedy nie ma błędu systematycznego) oraz że wariancja estymatora powinna maleć wraz ze wzrostem liczebności próby. Można udowodnić że estymatory parametrów b_0 oraz b_1 uzyskane **metodą najmniejszych kwadratów** posiadają obie właściwości.

Graficznie **kryterium minimalizacyjne** przedstawia rysunek 4.4.

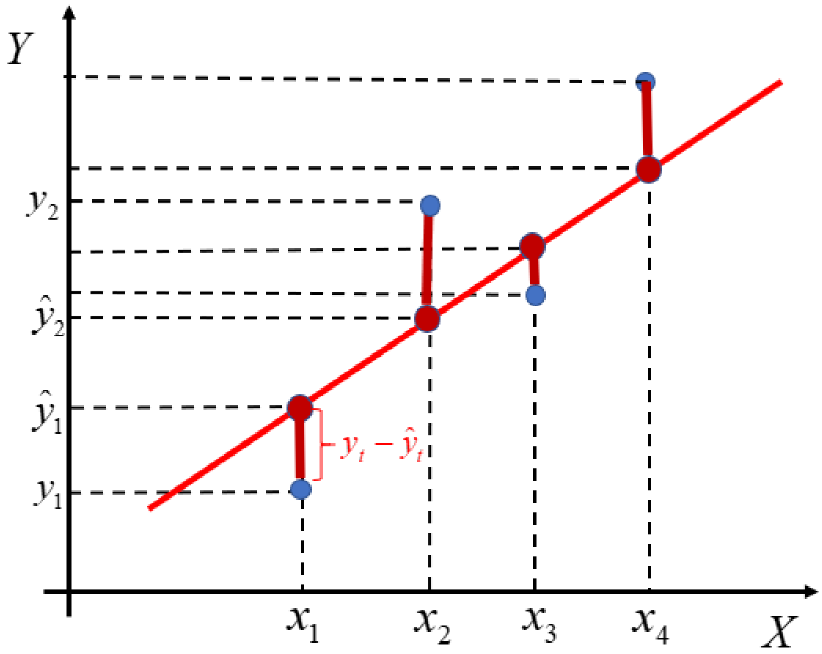
Suma podniesionych do kwadratu odległości pomiędzy czerwonymi (leżącymi na linii prostej w wersji czarno-białej) i niebieskimi kropkami ma być minimalna. Kropki niebieskie to wartości empiryczne; kropki czerwone to wartości teoretyczne. Zadanie wyznaczenie parametrów takiej prostej oczywiście realizuje program komputerowy.

Można udowodnić, że bez względu czy punkty na wykresie układają się w przybliżeniu wzdłuż prostej czy nie, zawsze **jakaś prosta** zostanie dopasowana (jeżeli tylko punktów jest więcej niż jeden.) Jak ocenić w sposób bardziej konkretny a nie tylko na oko jakość dopasowania prostej do wartości empirycznych?

Ocena dopasowania: wariancja resztowa oraz średni błąd szacunku

Oznaczając *resztę* jako: $e_i = y_i - \hat{y}_i$, definiujemy **wariancję resztową** jako:

$$s_e^2 = \frac{e_1^2 + e_2^2 + ...e_n^2}{n - k}$$



Rysunek 4.4: Metody statystycznej weryfikacji zależności pomiędzy zmiennymi

.

Gdzie n oznacza liczbę obserwacji (liczebność próby), a k liczbę szacowanych parametrów bez wyrazu wolnego czyli jeden w regresji prostej (a więcej niż jeden w regresji wielorakiej o czym dalej.)

Pierwiastek kwadratowy z **wariancji resztowej.** nazywamy **średnim błędem szacunku** (*mean square error*, MSE)

Ocena dopasowania: współczynniki zbieżności i determinacji

Suma kwadratów reszt (albo odchyleń wartości teoretycznych od wartości empirycznych, albo suma kwadratów błędów vel **resztowa suma kwadratów**):

$$\mathrm{RSK} = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + ... + (y_n - \hat{y}_n)^2$$

.

Suma kwadratów odchyleń **wartości empirycznych** od średniej (**ogólna suma kwadratów**):

$$\mathrm{OSK} = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + ... + (y_n - \bar{y})^2$$

Suma kwadratów odchyleń **wartości teoretycznych** od średniej (**wyjaśniona suma kwadratów**):

$$\mathrm{WSK} = (\hat{y}_1 - \bar{y})^2 + (\hat{y}_2 - \bar{y})^2 + ... + (\hat{y}_n - \bar{y})^2$$

Można wykazać, że OSK = WSK + RSK zatem (po podzieleniu obu stron równania przez OSK otrzymujemy:

$$1 = \text{WSK/OSK} + \text{RSK/OSK}$$

Współczynnik zbieżności oznaczany jako R^2 to WSK/OSK.

Współczynnik determinacji oznaczany jako Φ^2 (duża grecka litera Fi) to RSK/OSK.

Współczynniki przyjmują wartość z przedziału [0, 1] lub [0, 100]% jeżeli ich wartości zostaną pomnożone przez 100.

Interpretacja współczynnika zbieżności: udział (procent) zmienność wyjaśnianej przez linię regresji. Im R^2 jest bliższe jedności (lub 100% jeżeli jest współczynnik zbieżności jest wyrażony w procentach) tym lepiej.

Ocena dopasowania: istotność parametru b_1

Jeżeli: $Y = 0 \cdot X + b_0$, to $Y = b_0$ czyli nie ma zależności pomiędzy X oraz Y. Wartości b_1 bliskie zero wskazują na słabą zależność pomiędzy cechami.

Przypominamy, że **estymator** parametru b_1 ma średnią równą prawdziej wartości b_1. Dodatkowo zakładamy, że rozkład tego estymatora jest normalny. To założenie pozwala wiarygodnie oszacować wariancję; w konsekwencji znamy dokładny rozkład (bo przypominamy, że rozkład normalny jest określony przez dwa parametry: średnią oraz właśnie wariancję)

Można teraz zadać pytanie jeżeli faktycznie $b_1 = 0$, to jakie jest prawdopodobieństwo, że współczynnik $\hat{b}_1$ oszacowany na podstawie n obserwacji będzie (co do wartości bezwzględnej) większy niż b_e. Albo inaczej: otrzymaliśmy b_e, jakie jest prawdopodobieństwo otrzymania takiej wartości (lub większej co do wartości bezwzględnej) przy założeniu, że istotnie $b_1 = 0$.

Jeżeli takie prawdopodobieństwo jest duże, to uznajemy, że być może $b_1 = 0$, a jeżeli małe to będziemy skłonni uznać, że $b_1 \neq 0$. Duże/małe przyjmujemy arbitralnie, zwykle jest to $0,1$, $0,05$ lub $0,01$. Tak zgadza się, to prawdopodobieństwo to **poziom istotności**

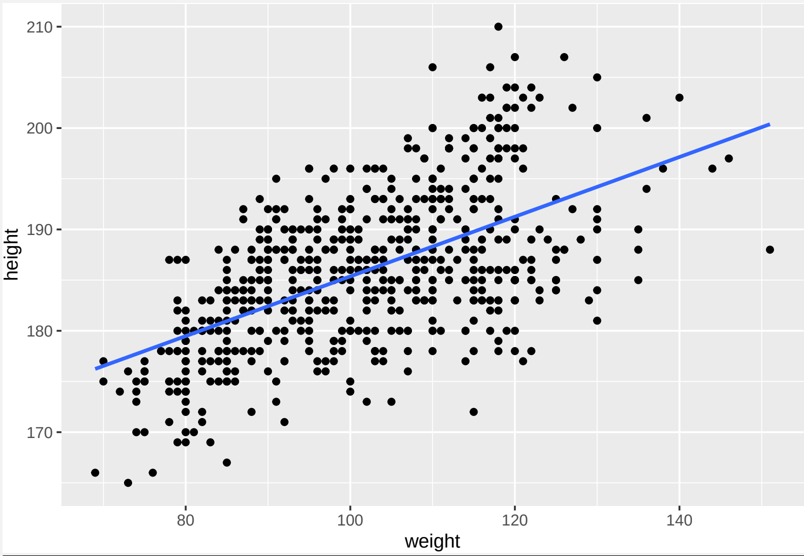
W każdym programie komputerowym na wydruku wyników linii regresji są podane wartości prawdopodobieństwa $b_1 > b_e$ (co do wartości bezwzględnej). Jeżeli jest ono mniejsze niż ustalony **poziom istotności** to b_1 ma wartość istotnie różną od zera.

Testowanie istotności współczynnika regresji jest ważnym kryterium oceny jakości dopasowania. Regresja z **nieistotnym** współczynnikiem nie może być podstawą do interpretowania zależności pomiędzy X oraz Y.

Waga a wzrost rugbystów
Zależność między wagą (`weight`) a wzrostem (`height`):

$$\text{height} = b_0 + b_1 \text{weight}$$

Oszacowanie tego równania na próbie 635 uczestników Pucharu Świata w rugby w 2023 roku daje następujące wyniki:



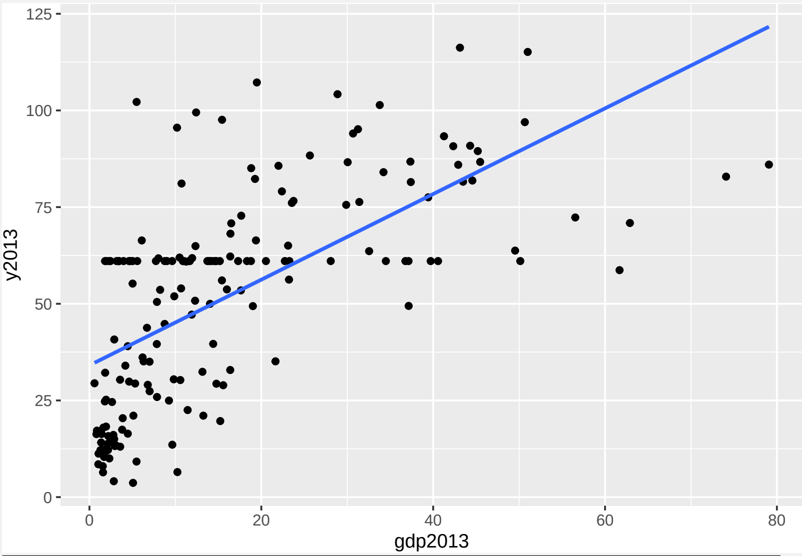
Zmienna	B	Błąd stand	z	p	CI95
(Intercept)	155.926	1.753	88.969	0	152.480 159.370
weight	0.294	0.017	17.305	0	0.260 0.330

Pierwsza kolumna `Zmienna` zawiera nazwy zmiennych (`(Intercept)` oznacza wyraz wolny). Druga kolumna oznaczona jako `B` zawiera oszacowane wartości (oceny) parametrów linii regresji. Kolumna `Błąd stand` zawiera oceny błędu standardowego estymatorów parametrów linii regresji. Kolumna `p` zawiera prawdopodobieństwo $b > b_e$.
Wzrost wagi zawodnika o 1kg skutkuje przeciętnie większym wzrostem o 0.294 cm. Współczynnik determinacji wynosi 32.86%. Współczynnik nachylenia prostej jest istotny ponieważ wartość p (tak mała, że w tabeli oznaczona jako 0) jest grubo poniżej zwyczajowego poziomu istotności (p < 0,05).
Kolumna `CI95` zawiera 95% przedziały ufności: z 95% prawdopodobieństwem wartość współczynnika nachylenia prostej znajduje się w przedziale 0,24--0,32.

Zamożność a konsumpcja mięsa
Następujący równanie opisuje zależność pomiędzy dochodem narodowym na głowę (tys USD *per capita*) a konsumpcją mięsa w kilogramach:

$$\text{konsumpcja} = b_0 + b_1 \text{gdp}$$

Model oszacowano dla krajów świata w roku 2013 na podstawie danych pobranych z bazy FAO Food Balance Sheet oraz Banku Światowego, otrzymując następujące wyniki:



Zmienna	B	Błąd stand	z	p	CI95
(Intercept)	34.085	2.232	15.268	0	29.680 38.490
gdp2013	1.108	0.100	11.124	0	0.910 1.300

Każdy tysiąc USD *per capita* więcej dochodu narodowego (GDP) oznacza przeciętny wzrost spożycia mięsa o 1.108 kg. Przeciętna różnica wartości teoretycznych od empirycznych wynosi 21,04 kg (średni błąd szacunku). Współczynnik zbieżności wynosi 40.88%. Współczynnik nachylenia prostej (którego wartość wynosi 1.108) jest statystycznie istotny.

Nie ma przykładów zastosowania regresji prostej w literaturze przedmiotu, bo jest ona zbyt dużym uproszczeniem rzeczywistości. Jest to jednak dobry punkt startu do bardziej skomplikowanego modelu **regresji wielorakiej**.

4.3.6 Zmienna liczbowa i zmienne liczbowe lub nominalne

Jeżeli zmiennych niezależnych jest więcej niż jedna, to mówimy o **regresji wielorakiej.** Przykładowo zależność pomiędzy wynikiem egzaminu, spożyciem kawy czasem nauki oraz predyspozycjami opisuje następujący model regresji:

$$\text{wynik} = b_0 + b_1 \cdot \text{kawa} + b_2 \cdot \text{czas} + b_3 \cdot \text{predyspozycje}$$

Współczynnik b_1 określa wpływ spożycia kawy, b_2 czasu poświęconego na naukę, a b_3 predyspozycji (intelektualnych, mierzonych np. średnią ocenę ze studiów). Ogólnie model regresji wielorakiej zapisać można jako:

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + ... + b_k \cdot X_k$$

Wpływ każdej zmiennej X_i na zmienną zależną Y jest określony przez odpowiedni współczynnik b_i. Zmienne X_i mogą być zmiennymi liczbowymi lub nominalnymi.

Podobnie jak w przypadku regresji prostej do oceny stopnia dopasowania modelu do danych wykorzystuje się: średni błąd szacunku, współczynnik zbieżności R^2 oraz weryfikuje się istotność współczynników b_i.

Standaryzacja współczynników regresji

Ponieważ współczynniki regresji $b_1, \ldots, b_k$ mogą być wyrażone w różnych jednostkach miary, bezpośrednie porównanie jest niemożliwe; mały współczynnik może w rzeczywistości być ważniejszy niż większy. Jeżeli chcemy porównywać wielkości współczynników to trzeba je **zestandaryzować**.

Standaryzowany współczynnik regresji dla i-tej zmiennej obliczony jest poprzez pomnożenie współczynnika regresji b_i przez s_{xi} i podzielenie przez s_y:

$$\beta_i = b_i \frac{s_{xi}}{s_y}$$

Dla przypomnienia s_{xi} to odchylenie standardowe zmiennej X_i, a s_y to odchylenie standardowe zmiennej Y. Interpretacja współczynnika standardyzowanego jest cokolwiek dziwaczna: zmiana zmiennej X_i o jedno odchylenie standardowe (s_{xi}) skutkuje zmianą zmiennej Y o b_i jej odchylenia standardowego s_y. Na szczęście współczynniki regresji standaryzuje się nie w celu lepszej interpretacji, tylko w celu umożliwienia porównania ich względnej wielkości (*wielkości efektu*). W publikacjach medycznych zwykle używa się litery b na oznaczenie współczynników niestandaryzowanych a litery β na oznaczenie współczynników standaryzowanych.

Wielkość efektu

Współczynniki regresji to miara wielkości efektu, która wskazuje na siłę zależności między zmiennymi. Standaryzacja pozwala na porównanie wielkości efektu zmiennych mierzonych w różnych jednostkach miary. Standaryzacja przydaje się także w przypadku posługiwania się skalami pomiarowymi mierzącymi przekonania i postawy, które z definicji są bezjednostkowe.

Wybór zmiennych objaśniających

Zwykle jest tak, że do objaśniania kształtowania się wartości zmiennej Y kandyduje wiele potencjalnych predyktorów X_k. Model zawierający wszystkie X_k predyktory niekoniecznie będzie najlepszy. Nie wdając się w omawianie szczegółowych zasad poprzestaniemy na dwóch kryteriach:

1. Model prostszy jest lepszy od modelu bardziej skomplikowanego jeżeli adekwatnie objaśnia zmienność Y (zasada brzytwy Ockhama, por. https://pl.wikipedia.org/wiki/Brzytwa_Ockhama).
2. Model powinien zawierać tylko zmienne o współczynnikach, których wartości są statystycznie różne od zera.

Regresja krokowa (*stepwise regression*) jest metodą wyboru najlepszych predyktorów spośród większego zbioru zmiennych. Występuje w dwóch wariantach **dołączania** i **eliminacji.** Ponieważ **eliminacja** wydaje się prostsza omówimy tylko ten wariant.

W metodzie eliminacji początkowym modelem jest model zawierający wszystkie potencjalne X_k predyktory. Następnie testujemy istotność wszystkich współczynników regresji i usuwamy ze zbioru predyktorów ten, który jest „najbardziej nieistotny" (ma największą wartość p) Procedurę powtarzamy dla modelu bez usuniętej zmiennej. Procedurę przerywamy gdy wszystkie współczynniki regresji są statystycznie istotne.

Zależność pomiędzy ciśnienie skurczowym, BMI oraz wiekiem

$$\text{ciśnienie} = b_0 + b_1 \text{BMI} + b_2 \text{wiek}$$

Dane pochodzą z badania: Zależność pomiędzy BMI i wiekiem a występowaniem cukrzycy wśród dorosłych osób w Chinach. Badanie kohortowe (Chen i inni, *Association of body mass index and age with incident diabetes in Chinese adults: a population-based cohort study.* BMJ Open.

2018 Sep 28;8(9):e021768. doi: 10.1136/bmjopen-2018-021768. PMID: 30269064; PMCID: PMC6169758.)
Oryginalny zbiór danych liczy 60 tysięcy obserwacji. Dla celów przykładu losowo wybrano 90, 490 oraz 4490 obserwacji. Zobaczymy jaki ma wpływ wielkość próby na wynik szacowania modelu.
Oszacowanie równania dla próby o wielkości 90 obserwacji daje następujące wyniki:

Zmienna	B	Błąd stand	z	p	Beta	CI95
(Intercept)	59.698	11.965	4.990	0.000	NA	35.920 83.480
BMI	1.742	0.486	3.583	0.001	0.330	0.780 2.710
age	0.484	0.124	3.906	0.000	0.360	0.240 0.730

Współczynnik zbieżności wynosi 26.24%. Kolumna `Beta` zawiera standaryzowane oceny parametrów regresji. Tej kolumny na poprzednich wydrukach (punkt 4.3.5) nie było, bo w przypadku regresji prostej standaryzacja jest zabiegiem raczej zbędnym. Dla wyrazu wolnego nie ma wartości standaryzowanej (co oznaczono jako `NA` czyli *not available*), ale to żadna strata -- oceny tego parametru nie są interpretowane. Wpływ `BMI` na wielkość ciśnienia jest nieco niższy niż `age`.
Oszacowanie równania dla próby o wielkości 490 obserwacji daje następujące wyniki:

Zmienna	B	Błąd stand	z	p	Beta	CI95
(Intercept)	79.061	4.378	18.057	0	NA	70.460 87.660
BMI	1.213	0.183	6.637	0	0.280	0.850 1.570
age	0.259	0.053	4.856	0	0.210	0.150 0.360

Współczynnik zbieżności wynosi 14.97%. Wpływ `BMI` na wielkość ciśnienia jest teraz wyższy niż `age`. Przedziały ufności są węższe co wynika z większej liczebności próby.
Oszacowanie równania dla próby o wielkości 4490 obserwacji daje następujące wyniki:

Zmienna	B	Błąd stand	z	p	Beta	CI
(Intercept)	74.011	1.530	48.358	0	NA	71.010 77.010
BMI	1.375	0.064	21.404	0	0.300	1.250 1.500
age	0.320	0.018	18.270	0	0.250	0.290 0.350

Współczynnik zbieżności wynosi 18.54%. Przedziały ufności są jeszcze węższe. Ocena `age` z 95% prawdopodobieństwem znajduje się w przedziale [0.290 0.350] a w pierwszym oszacowaniu dla znacznie mniejszej próby było to [0.240 0.730]. Przedział jest ponad 8 razy węższy...

4.3.7 Zmienne zero-jedynkowe

Zamiast (celem wykazania związku między zmienną licznową a nominalną) porównywać średnie w grupach możemy wykorzystać metodę regresji wielorakiej. Zmienna nominalna jest zamieniana na jedną lub więcej zmiennych binarnych, które przyjmują tylko dwie wartości 0 lub 1.

Przykładowo rodzaj miejsca pracy (skala nominalna; dwie wartości: szpital, przychodnia) można zamienić na zmienną binarną `praca` przypisując 1 = szpital, oraz 0 = przychodnia (lub odwrotnie). Załóżmy że poziom stresu zależy od stażu pracy, satysfakcji (obie mierzone na skali liczbowej) i rodzaju miejsca pracy. Możemy to zapisać jako następujące równanie regresji:

$$\text{stres} = b_0 + b_1\text{staż} + b_2\text{satysfakcja} + b_3\text{praca}$$

Jaka jest interpretacja współczynnika b_3? Zakładając że 0 = przychodnia, b_3 oznacza przeciętną zmianę wielkości stresu spowodowaną pracą w szpitalu w porównaniu do pracy w przychodni. Jeżeli ten współczynnik jest istotny statystycznie, to istnieje zależność pomiędzy stresem a miejscem pracy. Czyli zamiast stosować test t-Studenta i porównywać średnie w grupach, możemy oszacować model regresji z wykorzystaniem stosownej zmiennej zero-jedynkowej a następnie sprawdzić czy współczynnik stojący przy tej zmiennej jest istotny.

Jeżeli zmienna nominalna ma n wartości należy ją zamienić na $n-1$ zmiennych zero-jedynkowych. Załóżmy że stress zależy także od wykształcenia, mierzonego w skali nominalnej (średnie, licencjat, magisterskie.) Tworzymy dwie zmienne: magister (jeden jeżeli respondent ma wykształcenie magisterskie lub 0 jeżeli nie ma) oraz licencjat (jeden jeżeli respondent ma licencjat lub 0 jeżeli nie ma). Równanie regresji ma postać:

$$\text{stres} = b_0 + b_1\text{staż} + b_2\text{satysfakcja} + b_3\text{praca} + b_4\text{magister} + b_5\text{licencjat}$$

Jeżeli magister $=0$ oraz licencjat $=0$ to osoba ma wykształcenie średnie.

Interpretacja: b_4 (jeżeli istotne) oznacza przeciętną zmianę wielkości stresu osoby z wykształceniem magisterskim w porównaniu do osoby z wykształceniem średnim. Podobnie b_5 oznacza przeciętną zmianę wielkości stresu osoby z wykształceniem licencjackim w porównaniu do osoby z wykształceniem średnim.

Zależność pomiędzy ciśnienie skurczowym, BMI, wiekiem, płcią, paleniem i piciem
Poprzednio rozważany model zależności pomiędzy ciśnienie skurczowym, BMI oraz wiekiem rozszerzymy o trzy zmienne: płeć (kobieta/mężczyzna), status względem picia alkoholu (pije, pił, nigdy nie pił) oraz status względem palenia (palił, pali, nigdy nie palił). Zwróćmy uwagę że zmienne mierzące status względem palenia/picia mają nie dwie a trzy wartości. Należy każdą zamienić na dwie zmienne binarne, wg schematu:
`current.smoker` (pali) = 1 jeżeli pali, 0 w przeciwnym przypadku
`ever.smoker` (kiedyś palił) = 1 jeżeli palił ale nie pali, 0 w przeciwnym przypadku
Zmienna płeć `genderF` = 1 jeżeli kobieta, lub 0 jeżeli mężczyzna. Zauważmy, że nazwa zmiennej dwuwartościowej wskazuje która wartość jest zakodowana jako 1. Przykładowo `genderF` (*female* żeby się trzymać języka angielskiego) wskazuje że jedynką jest kobieta. Taka konwencja ułatwia interpretację. Gdybyśmy zamiast `genderF` nazwali zmienną `gender` to na pierwszy rzut oka nie było by wiadomo co zakodowano jako jeden. A tak wiadomo od razu jak interpretować parametr stojący przy tej zmiennej: zmiana wielkości ciśnienia u kobiet w porównaniu do mężczyzn.
Rozważany model ma postać:

$$SBP = b_0 + b_1\text{BMI} + b_2\text{age} + b_3\text{genderF} + b_4\text{current.smoker}+ \quad (4.1)$$
$$+ b_5\text{ever.smoker} + b_6\text{current.drinker} + b_7\text{ever.drinker} \quad (4.2)$$

Oszacowanie tego równania dla próby o wielkości 90 obserwacji daje następujące wyniki:

Zmienna	B	Błąd stand	z	p	Beta	CI
(Intercept)	90.332	15.745	5.737	0.000	NA	59.010 121.650
BMI	0.778	0.592	1.314	0.193	0.150	-0.400 1.960
age	0.441	0.121	3.658	0.000	0.330	0.200 0.680
genderF	-13.820	4.399	-3.141	0.002	-0.400	-22.570 -5.070
current.smoker	-6.890	3.972	-1.735	0.087	-0.180	-14.790 1.010
ever.smoker	7.626	6.882	1.108	0.271	0.110	-6.070 21.320
current.drinker	-3.959	8.523	-0.465	0.643	-0.040	-20.910 13.000
ever.drinker	-4.001	4.575	-0.875	0.384	-0.080	-13.100 5.100

Współczynnik zbieżności wynosi 38.11%. Tylko dwie na siedem zmiennych są istotne. Zwróćmy uwagę że nieistotnie zmienne mają przedziały ufności zawierające zero. W konsekwencji z 95% prawdopodobieństwem wartości tych współczynników mogą być raz ujemne raz dodatnie -- nie mamy nawet pewności co do kierunku zależności między zmienną objaśniającą a ciśnieniem. Zmienne, które okazały się istotne jednocześnie mają największą wielkość efektu (kolumna `Beta`) i nie jest to przypadek.

Oszacowanie tego samego równania dla próba o wielkości 4490 obserwacji daje następujące wyniki:

Zmienna	B	Błąd stand	z	p	Beta	CI
(Intercept)	80.089	1.623	49.334	0.000	NA	76.910 83.270
BMI	1.192	0.066	18.009	0.000	0.260	1.060 1.320
age	0.336	0.018	19.087	0.000	0.260	0.300 0.370
genderF	-5.329	0.508	-10.496	0.000	-0.160	-6.320 -4.330
current.smoker	-2.752	0.583	-4.717	0.000	-0.070	-3.900 -1.610
ever.smoker	-2.021	1.045	-1.933	0.053	-0.030	-4.070 0.030
current.drinker	3.621	1.535	2.360	0.018	0.030	0.610 6.630
ever.drinker	0.193	0.623	0.310	0.757	0.000	-1.030 1.410

Współczynnik zbieżności wynosi 20.72%. Zwiększenie liczebności próby z 90 do 4490 obserwacji spowodowało, że tylko dwie z siedmiu zmiennych mają nieistotne wartości. Analizując wartości standaryzowane możemy ustalić które zmienne mają największy wpływ na wielkość ciśnienia krwi.

Ktoś mógłby dojść do wniosku że wszystko da się **uistotnić** wystarczy zwiększyć wielkość próby. Teoretycznie tak, praktycznie nie. W praktyce nie interesuje nas niewielka wielkość efektu (znikomy wpływ czegoś na coś). Dodatkowo zebranie dużej próby może być kosztowne czyli w praktyce niemożliwe -- nie mamy dość dużo pieniędzy. Można teoretycznie określić jaka wielkość próby pozwoli nam na ocenę jakiej wielkości efektu. Sposób postępowania jest wtedy następujący: określamy jaka wielkość efektu ma **znaczenie praktyczne**, na tej podstawie określamy niezbędną minimalną liczebność próby. Takie zaawansowane podejście wykracza poza ramy tego podręcznika.

Regresja krokowa

W modelu zależność pomiędzy ciśnienie skurczowym, BMI, wiekiem, płcią, paleniem i piciem (próba 4490) zmienne `ever.drinker` oraz `ever.smoker` są nieistotne przy czym współczynnik przy zmiennej `ever.drinker` ma wartość p równą 0,309 zaś przy zmiennej `ever.smoker` ma wartość 0,05324. Usuwamy zmienną `ever.drinker` (bo wartość p jest większa) i szacujemy równanie regresji dla sześciu pozostałych zmiennych. Otrzymujemy:

Zmienna	B	Błąd stand	z	p	Beta	CI
(Intercept)	80.108	1.622	49.387	0.000	NA	76.930 83.290
BMI	1.193	0.066	18.056	0.000	0.260	1.060 1.320
age	0.335	0.018	19.097	0.000	0.260	0.300 0.370
genderF	-5.358	0.499	-10.740	0.000	-0.160	-6.340 -4.380
current.smoker	-2.740	0.582	-4.708	0.000	-0.070	-3.880 -1.600
ever.smoker	-1.980	1.037	-1.910	0.056	-0.030	-4.010 0.050
current.drinker	3.579	1.528	2.342	0.019	0.030	0.580 6.580

Współczynnik przy zmiennej `ever.smoker` dalej uparcie jest nieistotny. Usuwamy teraz tę zmienną. Otrzymujemy:

Zmienna	B	Błąd stand	z	p	Beta	CI
(Intercept)	79.865	1.618	49.375	0.00	NA	76.690 83.040
BMI	1.195	0.066	18.075	0.00	0.260	1.070 1.320
age	0.336	0.018	19.100	0.00	0.260	0.300 0.370
genderF	-5.155	0.488	-10.572	0.00	-0.160	-6.110 -4.200
current.smoker	-2.540	0.573	-4.435	0.00	-0.060	-3.660 -1.420
current.drinker	3.551	1.529	2.323	0.02	0.030	0.550 6.550

Wszystkie współczynniki mają istotnie różnie od zera wartości. Wartość współczynnika zbieżności ostatecznego modelu wynosi 20.66%. Usuwając nieistotne zmienne z modelu obniżyliśmy wartość współczynnika zmienności o 20.72% - 20.66% = 0.07%, czyli tyle co nic.

4.4 Przypadek specjalny: regresja logistyczna

Jeżeli zmienna Y jest zmienną **dwuwartościową**, czyli taką która przyjmuje tylko dwie wartości (np. chory/zdrowy), to metoda regresji nie może być zastosowana. Przykładowo jeżeli zakodujemy te wartości jako chory=0 i zdrowy=1, to zastosowanie regresji doprowadzi do obliczenia (teoretycznych) wartości Y różnych od 0 i 1. Taki wynik nie ma sensownej interpretacji…

Ale zamiast szacować regresję Y względem (X/X-ów) można szacować regresję względem ryzyka dla Y (czyli prawdopodobieństwa że Y przyjmnie wartość 1). Tutaj znowu pojawia się jednak trudność, bo ryzyko może przyjąć tylko wartości z przedziału $[0, 1]$. Nie wchodząc w matematyczne zawiłości model zapisuje się jako (ln oznacza logarytm naturalny):

$$\ln(\frac{p}{1-p}) = b_0 + b_1 \cdot x_1 + \ldots + b_k \cdot x_k$$

Zauważmy, że $o = \frac{p}{1-p}$ to nic innego jak szansa (*odds*, por. punkt 4.1.1). Parametr b_i jest miarą wpływu zmiennej X_i na zmienną Y. Jeżeli X_i wzrośnie o jednostkę, to logarytm ilorazu szans wzrośnie o $\ln(o)$ (przy założeniu, że pozostałem zmienne X mają pewne ustalone wartości a zmienia się tylko X_i). Jeżeli X_i jest zmienną **dwuwartościową** to interpretacja jest jeszcze prostsza: jest to logarytm ilorazu szans dla wartości $X_i = 1$ względem $X_i = 0$.

Zwykle zamiast **logarytmu ilorazu szans** wolimy interpretować zmianę w kategoriach **ilorazu szans**. Aby otrzymać ów iloraz należy wykonać następujące przekształcenie (exp oznacza podstawę logarytmu naturalnego):

$$o = \exp^{\ln(o)}$$

Dla przypomnienia: zwykle iloraz szans wyraża się w procentach, czyli mnoży przez 100. Jeżeli ta liczba jest większa od 100 oznacza to wzrost szansy, a jeżeli mniejsza od 100, spadek szansy.

Ocena dopasowania

Nie ma w przypadku regresji logistycznej możliwości obliczenia sumy kwadratów reszt (*residual sum of squares*) oraz współczynnika zbieżności. Model ocenia się używając jako kryterium dewiancję (*deviance*). Dewiancja to miara, której wielkość zależy od proporcji pomiędzy liczbą sukcesów obliczonych z modelu a liczbą sukcesów zaobserwowanych (jak dokładnie dewiancja jest liczona nie jest dla nas istotne).

Wyjaśnijmy to na przykładzie prostego modelu pomiędzy wystąpieniem osteoporozy a płcią. Model ma postać:

$$\ln(o) = b_0 + b_1 \text{płeć}$$

Po oszacowaniu b_0 oraz b_1 możemy łatwo obliczyć $\ln(o)$. Wiedząc że $\ln(o) = \frac{p}{1-p}$ możemy stąd obliczyć prawdopodobieństwo, które jak widać będzie różne dla kobiet i mężczyzn. Po pomnożeniu tych prawdopodobieństw przez liczebności dostajemy (teoretyczne) liczebności sukcesów (tj. wystąpienia osteoporozy). Dewiancja będzie tym większa im różnica między tymi teoretycznymi liczebnościami a liczebnościami empirycznymi będzie większa.

Jako minimum porównuje się wielkość dewiancji szacowanego modelu z modelem zerowym (*null model*), tj. modelem w którym po prawej stronie równania występuje tylko stała:

$$\ln(o) = b_0$$

W tym modelu prawdopodobieństwo osteoporozy jest identyczne dla kobiet i mężczyzn, zatem w oczywisty sposób dewiancja tego modelu będzie większa. Pytanie jest czy różnica jest istotna statystycznie. Jeżeli jest większa to przyjmuje się, że szacowany model jest lepszy od modelu trywialnego (warunek minimum przydatności.)

Jeżeli model zawiera wiele zmiennych w tym zmienne liczbowe, idea liczenia dewiancji jest podobna, ale oczywiście szczegóły są już bardziej skomplikowane. Szczegóły te nie są wszakże dla nas istotne bo zajmuje się tym program komputerowy.

Minimalne kryteria oceny przydatności modelu regresji logistycznej: istotnie mniejsza od modelu zerowego dewiancja oraz istotnie różne od zera parametry przy zmiennych niezależnych (predyktorach)

Ocena skuteczności klasyfikacji

Model regresji logistycznej nie oblicza wartości zmiennej prognozowanej, bo ta nie jest liczbą, tylko **klasyfikuje**, tj. ustala (albo prognozuje) wartość zmiennej nominalnej w kategoriach „sukces''/„porażka''. Ważnym kryterium oceny jakości modelu jest ocena jakości klasyfikacji, to jest ocena na ile model poprawnie przypisuje przypadkom kategorie zmiennej prognozowanej. Im mniejsza rozbieżność pomiędzy wartościami rzeczywistymi, a prognozowanymi tym oczywiście lepiej.

Tę jakość klasyfikacji ocenia się za pomocą dwóch wskaźników: czułość (*sensitivity*) oraz swoistość (*specifity*).

1. Odsetek sukcesów zaklasyfikowanych jako „sukces'' (**Czułość**); określany także jako TPR (*true-positive-rate*).
2. Odsetek porażek zaklasyfikowanych jako „porażka'' (**Swoistość**); określany także jako TNR (*true-negative-rate*).

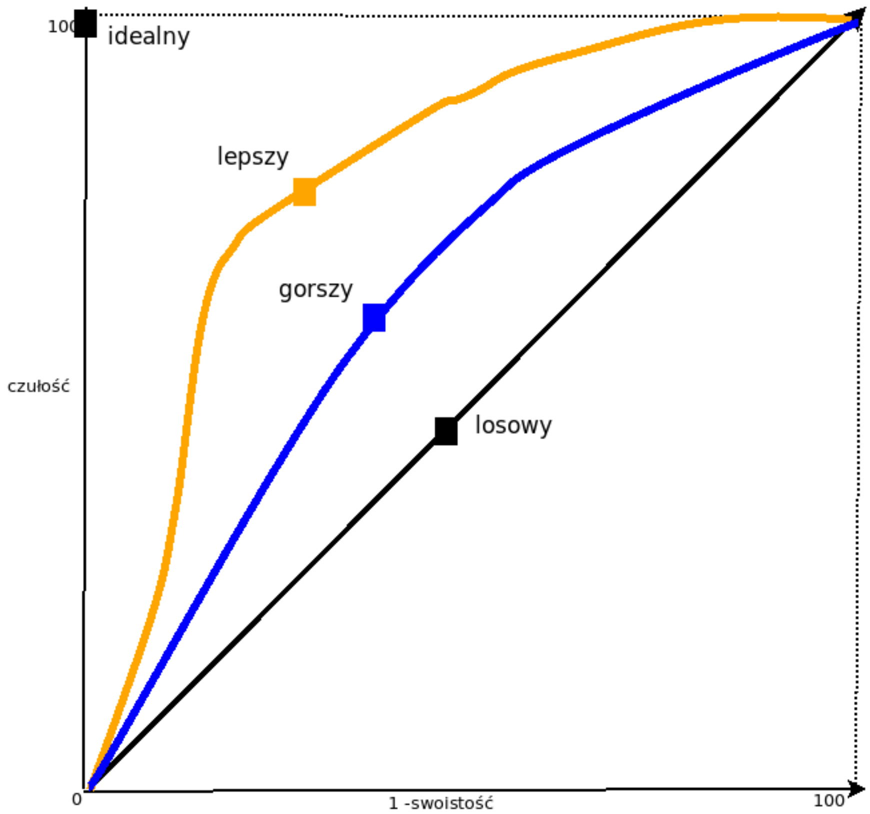
Klasyfikacja w modelu regresji logistycznej wygląda następująco. Jeżeli prawdopodobieństwo obliczone z modelu jest wyższe-lub-równe niż założona **wartość graniczna** (p_g), to zakładamy „sukces'', jeżeli tak nie jest, to zakładamy „porażkę''. Wartość graniczna jest ustala albo arbitralnie albo na podstawie jakieś dodatkowej (pozastatystycznej) informacji. Domyślnie za wartość graniczną przyjmuje się zwykle $p_g = 0,5$, co oznacza że wartości $p \geq 0,5$ zostaną zamienione na „sukces'' a wartości $p < 0,5$ zostaną zamienione na „porażkę''.

Ocena dopasowania: krzywa ROC

Czułości oraz swoistości zależą od prawdopodobieństwa granicznego. Im wyższa jest wartość prawdopodobieństwa granicznego tym mniej będzie „sukcesów".

Krzywa ROC przedstawia w układzie współrzędnych XY wartości czułości oraz swoistości dla różnych wartości granicznych. Współczynnik AUC (*area under curve*) to wielkość pola pod krzywą wyrażona w procentach pola kwadratu o boku 100%. AUC zawiera się w przedziale 50--100. Im większa wartość współczynnika tym lepiej. Model który klasyfikuje czysto losowo ma wartość AUC równą 50% (por. rysunek 4.5).

Osteoporoza i witamina D
Al Zarooni A.A.R i inni badali wpływ różnych czynników na ryzyka na wystąpienie osteoporozy (Risk factors for vitamin D deficiency in Abu Dhabi Emirati population; https://doi.org/10.1371/journal.pone.0264064), takich jak deficyt witaminy D, wiek oraz płeć w grupie 392 osób.



Rysunek 4.5: Krzywa ROC

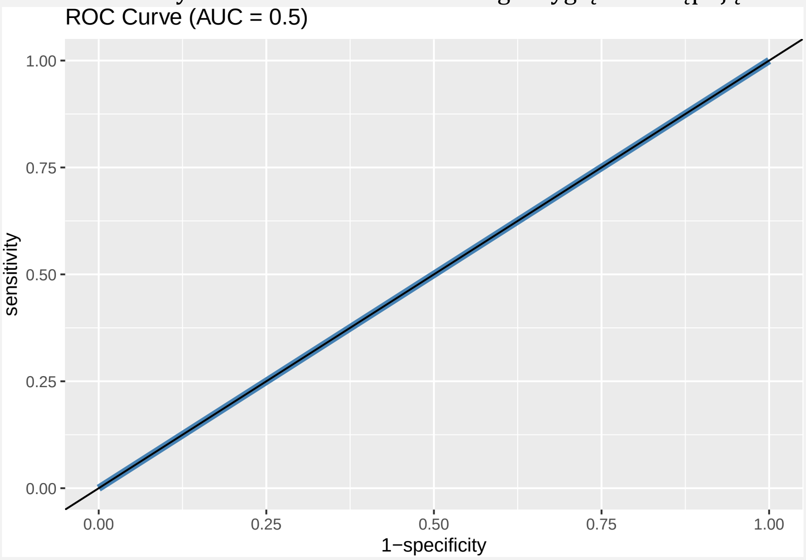
Zacznijmy od modelu zerowego tj. takiego w którym ryzyko/prawdopodobieństwo/szansa wystąpienia osteoporozy jest takie same bez względu na wielkości innych zmiennych. Odpowiada to następującemu równaniu:

$$\ln(o) = b_0$$

W tabeli zestawiono wartości parametrów oszacowanego modelu, ilorazy szans, przedziału ufności oraz prawdopodobieństwo

Parametr	Ocena	Błąd stand	z	p
(Intercept)	-2.644537	0.2029618	-13.02973	0

Można obliczyć że (teoretyczne) prawdopodobieństwo wystąpienia osteoporozy wyniosło 0.0663265. Krzywa ROC dla modelu zerowego wygląda następująco:



Model zerowy jak sama nazwa wskazuje może tylko służyć do porównania z bardziej skomplikowanymi modelami.
Takim bardziej skomplikowanym modelem będzie przykładowo zależność pomiędzy wystąpieniem osteoporozy a płcią, którą można opisać następującym równaniem regresji:

$$\ln(o) = b_0 + b_1\text{kobieta}$$

Zmienna `kobieta` przyjmuje wartość 1 jeżeli osoba była kobietą oraz zero w przypadku jeżeli była mężczyzną. Dla przypomnienia o jest szansą wystąpienia osteoporozy.
W tabeli zestawiono wartości parametrów oszacowanego modelu, ilorazy szans, przedziału ufności oraz prawdopodobieństwo

Parametr	Ocena	Błąd stand	z	p	OR	CI
(Intercept)	-3.367	0.455	-7.403	0.000	0.030	0.010 0.080
genderF	1.014	0.509	1.992	0.046	2.760	1.090 8.400

Znając wartości współczynników równania można obliczyć wartości $\ln(o)$
Dewiancja modelu jest istotnie mniejsza od modelu zerowego (wartość p wynosi bowiem 0.0303521)
Zależność pomiędzy wystąpieniem osteoporozy a płcią, wiekiem oraz poziomem witaminy D można opisać następującym równaniem regresji:

$$\ln(o) = b_0 + b_1\text{kobieta} + b_2\text{wiek} + b_3\text{poziomD}$$

W tabeli zestawiono wartości parametrów oszacowanego modelu, ilorazy szans, przedziału ufności oraz prawdopodobieństwo

Parametr	Ocena	Błąd stand	z	p	OR	CI
(Intercept)	-12.183	1.766	-6.898	0.000	0.000	0.000 0.000
d	0.005	0.009	0.536	0.592	1.000	0.990 1.020
age	0.156	0.026	5.930	0.000	1.170	1.120 1.240
genderF	2.463	0.662	3.722	0.000	11.740	3.540 48.760

Macierz pomyłek (*confussion matrix*)

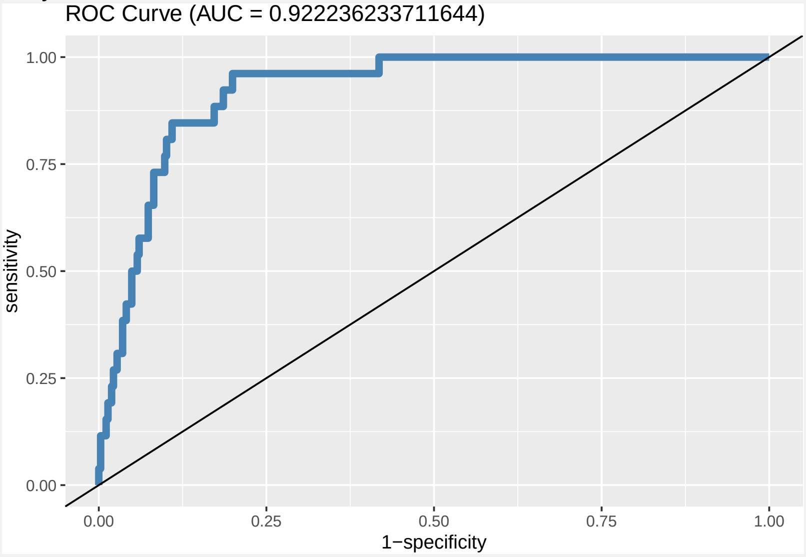
```
##           Osteoporoza
## Prognoza   0   1
##          0 362  22
##          1   4   4
```

Stąd: czułość 0.1538462; swoistość 0.989071

Istotność modelu

Dewiancja jest istotnie mniejsza od dewiancji modelu zerowego (p = 0)

Krzywa ROC



4.5 Przypadek specjalny: dwie zmienne co najmniej porządkowe

4.5.1 Pomiar siły zależności: współczynnik korelacji rang

Współczynnik korelacji rang Spearmana (*Spearman's Rank-Order Correlation*) może być stosowany w przypadku gdy cechy są mierzone w skali porządkowej (lub lepszej czyli liczbowej).

Obliczenie współczynnika Spearmana dla N obserwacji na zmiennych XY polega na zamianie wartości zmiennych X oraz Y na **rangi** (numery porządkowe od 1 do N). Następnie stosowana jest formuła współczynnika korelacji liniowej Pearsona (τ_x oraz τ_y oznaczają **rangi**):

$$\rho_{xy} = \frac{\mathrm{cov}(\tau_x, \tau_y)}{s_{\tau_x} s_{\tau_y}}$$

Współczynnik ρ_{xy} to – podobnie jak **oryginalny** współczynnik korelacji liniowej Pearsona -- miara niemianowana, o wartościach ze zbioru [-1;1];

Przykład: spożycie mięsa
Współczynnik Pearsona i Spearmana dla zależności między spożyciem mięsa w 1980 a spożyciem mięsa w 2013 roku (zmienna objaśniana):

```
## [1] "współczynnik Pearsona: 0.68"
## [1] "współczynnik Spearmana: 0.68"
```

Nie ma sensu liczenia współczynnika korelacji rang w przypadku kiedy obie cechy są liczbami, bo wtedy należy użyć normalnego współczynnika Pearsona. Ale nie jest to też błędem więc w powyższym przykładzie go liczymy :-)
Współczynnik korelacji liniowej Spearmana wynosi 0.68 (umiarkowana korelacja).
Czy ta wartość jest istotnie różna od zera? Jest na to stosowny test statystyczny, który sprowadza się do określenia jakie jest prawdopodobieństwo otrzymania $r_s = 0.68$ przy założeniu że prawdziwa wartość r_s wynosi zero. Otóż w naszym przykładzie to prawdopodobieństwo wynosi 2.302116e-26 (czyli jest ekstremalnie małe – r_s jest istotnie różne od zera).

4.6 Podsumowanie

Przedstawiono 7 następujących metod ustalania zależności między zmiennymi:

1. Wykres rozrzutu.
2. Tablica wielodzielcza i test chi-kwadrat.
3. Współczynnik korelacji liniowej Pearsona.
4. Współczynnik korelacji Spearmana.
5. Regresja liniowa.
6. Regresja logistyczna.
7. testy t-Studenta, U Manna-Whitneya, ANOVA albo test Kruskala-Wallisa.

Rozdział 5

Przykłady badań ankietowych

Uwaga: ankieta nie jest kolejną metodą statystyczną tylko techniką zbierania danych. Wszystkie metody już zostały przedstawione i żadna nowe nie będzie.

5.1 Jak zacząć badanie?

Każde w tym ankietowe.

Należy zastanowić się nad trzema sprawami:

1. Co chcemy ustalić?
2. Jakie dane są nam potrzebne żeby ustalić to co chcemy ustalić.
3. Jak te dane zebrać (czyli co i w jaki sposób zmierzyć)

Co chcemy ustalić?

Najlepiej jakąś zależność. Na przykład: Stress a wypalenie zawodowe; satysfakcja zawodowa a retencja; determinanty satysfakcji zawodowej

Może być od biedy opis czegoś lub porównanie czegoś z czymś. Przykłady: nadwaga wśród studentów wydziału zdrowia PSW; Analiza porównawcza wypalenia zawodowego pielęgniarek pracujących w różnych systemach opieki.

Co i jak mierzyć?

Jeżeli mamy zamiar badań nadwagę, to powinniśmy zmierzyć masę ciała. Jeżeli celem jest ustalenie zależności pomiędzy stresem a wypaleniem zawodowym to niewątpliwie powinniśmy zmierzyć stress i wypalenia. Jak dotąd banalnie prosto. Problem zaczyna się w momencie odpowiedzi na pytanie **jak**?

5.1.1 Mierzenie twardych faktów vs mierzenia przekonań

Możemy pytać w ankiecie o dwie rzeczy:

- **fakty** (wiek, staż, zawód, tętno, przebyte choroby);
- **przekonania**, **wartości**, **postawy**; **uczucia** (strach / radość) albo **zamiary** (w języku attitudes/emotions/intentions).

Mierzenie **faktów** nie wymaga dodatkowych objaśnień. Problem jest z mierzeniem **przekonań**.

Przekonanie to idea, którą jednostka uważa za prawdziwą. **Wartości** to trwałe przekonania o tym, co jest ważne dla jednostki. Stają się standardami, według których jednostki dokonują wyborów. **Postawy** to mentalne dyspozycje/nastawienie przed podjęciem decyzji, które skutkują określonym zachowaniem (zrobię to a nie tamto). Postawy kształtowane są wartościami i przekonaniami.

5.1.2 Pomiar przekonań, wartości i postaw

Postawy/uczucia/zamiary są to pojęcia abstrakcyjne. Często (albo zawsze) definiowane w obszarze psychologii, nauk o zarządzaniu itp.

Pomiar *przekonań* jest dokonywany w specyficzny sposób. **Definicja koncepcyjna** definiuje pojęcie (zaufanie do kogoś/czegoś to **przekonanie**, że *działania tego kogoś/czegoś okażą się zgodne z naszymi oczekiwaniami*; satysfakcja to **uczucie** *przyjemności, zadowolenia z czegoś*; samoskuteczność to **przekonanie**, iż *jest się w stanie zrealizować określone działanie lub osiągnąć wyznaczone cele*). **Definicja operacyjna** określa jak zmierzyć pojęcie (jak zmierzyć satysfakcję) Przejście od definicji koncepcyjnej do definicji operecyjnej bywa czasami mocno, hmm… arbitralne.

5.1.3 Skala Likerta

Przykładowo chcemy się dowiedzieć czy i jak bardzo respondenci boją się COVID19.

W najprostszej wersji się po prostu pytamy: **Czy pan/pani boi się COVID19?** i dajemy respondentowi trzy możliwe warianty odpowiedzi: Tak/Nie/Nie wiem.

Może też być pięć wariantów: bardzo się boję--boję się--ani/ani--nie boję się--zupełnie się nie boję.

Taką skalę pomiarową określamy jak wiemy jako **porządkową**. Pomiary nie są liczbami, ale są uporządkowane. Rangi wartości są już liczbami (np 1--5 w drugim przykładzie), można je np. uśredniać Tego typu skala pomiarowa, typowa dla ankiet, nosi nazwę skali **Likerta**. Można sobie wymyślać skalę Likerta 7-punktową i więcej.

Naszym zdaniem powyżej 7 wariantów normalny respondent będzie miał problem czy się bardziej-bardziej czy bardziej-bardziej-bardziej boi.

5.1.4 Skala pomiarowa czyli inwentarz albo kwestionariusz

Ponieważ skala Likerta jest zgrubna to uważa się powszechnie, że lepszy wynik da pomiar wielokrotny. W naukach podstawowych mierzymy (np. linijką) parę razy, a wynik uśredniamy co daje pomiar bardziej precyzyjny. Tutaj pytamy się parę razy o to samo co ma dać podobny efekt (mniejszy średni błąd pomiaru). Taka seria pytań nosi też nazwę skali albo **inwentarza**. Nie pytamy się zatem **Czy pan/pani boi się COVID19?** tylko zadajemy serię pytań o strach względem COVID19:

strach przed COVID19
The Fear of COVID-19 Scale: Development and Initial Validation. International Journal of Mental Health and Addiction, 1--9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7100496/
Lęk przed koronawirusem COVID-19 i lęk przed śmiercią -- polskie adaptacje narzędzi https://www.termedia.pl/Fear-of-COVID-19-and-death-anxiety-Polish-adaptations-of-scales,116,44937,1,1.html

1. I am most afraid of Corona
2. It makes me uncomfortable to think about Corona
3. My hands become clammy when I think about Corona
4. I am afraid of losing my life because of Corona
5. When I watch news and stories about Corona on social media, I become nervous or anxious.

6. I cannot sleep because I'm worrying about getting Corona.
7. My heart races or palpitates when I think about getting Corona

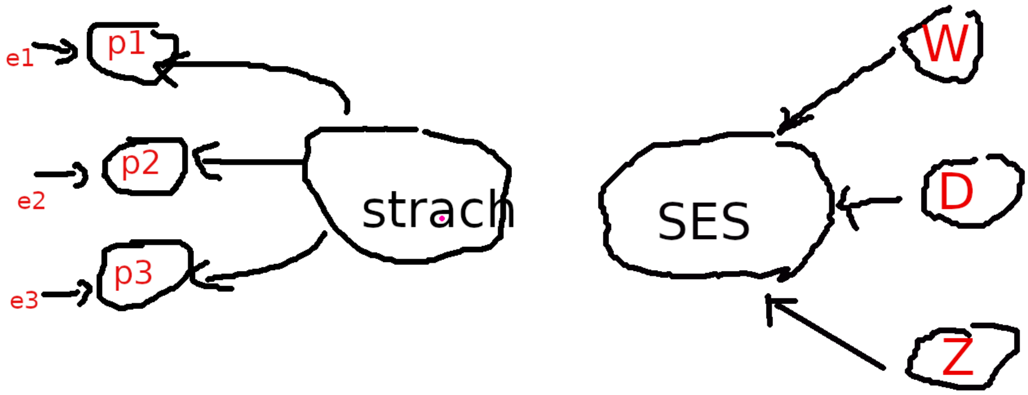
albo:

1. Boję się koronawirusa
2. Czuję dyskomfort, gdy myślę o koronawirusie
3. Pocą mi się dłonie, gdy myślę o koronawirusie
4. Boję się, że mogę stracić życie z powodu koronawirusa
5. Gdy oglądam wiadomości i czytam o koronawirusie w mediach społecznościowych, robię się nerwowy i niespokojny
6. Nie mogę spać, ponieważ martwię się, że ja lub moi bliscy zarażą się
7. Dostaję palpitacji serca, gdy myślę o tym, że mógłbym się zarazić.

Odpowiadający ma do wyboru pięć wariantów odpowiedzi: **zdecydowanie nie/nie/nie mam zdania/tak/zdecydowanie tak**

5.1.5 Model pomiaru

Ukryty czynnik (strach) kształtuje wartości indykatorów (odpowiedzi na pytania) Taki sposób pomiaru **ukrytego czynnika** (latent w języku) określa się mianem refleksyjnego (co jest kalką od *reflexive*). Na rysunku 5.1 kierunek strzałki obrazuje zależność (czynnik→indykator)



Rysunek 5.1: Modele pomiaru czynnika ukrytego

Alternatywny sposób definiowania ukrytego (w pewnym sensie, raczej złożonego) czynnika nosi nazwę **formatywnego** (albo indeksu): czynnik jest sumą indykatorów. Przykładem może być SES: status socjo-ekonomiczny będący agregatem wykształcenia (W), dochodu (D) oraz zawodu (Z).

W założeniu indykatory są jednakowo dobrymi miarami czynnika refleksyjnego i jako takie powinny być mocno skorelowane (mierzą to samo). Natomiast składniki czynnika formatywnego nie powinny być skorelowane, raczej każdy powinien mierzyć **inny aspekt** czynnika. Ktoś może być profesorem za przeproszeniem filozofii, nie mieć pracy i kiepskie dochody. Tylko jeden z trzech aspektów podwyższa mu SES; albo świetnie zarabiająca prostytutka bez matury.

Jeżeli w czynniku refleksyjnym pominiemy z trzech indykatorów to nic się nie stanie oprócz tego że pomiar będzie mniej precyzyjny. Jeżeli w czynniku formatywnym pominiemy indykator, to popełniamy gruby błąd, bo pomijamy istotny składnik całości.

Dobrą wiadomością jest, że najprostszy sposób pomiaru traktuje czynniki refleksyjne i formatywne jednakowo: wartością czynnika jest suma albo ewentualnie średnia wartości indykatorów. Jeżeli indykatory są mierzone za pomocą skali Likerta suma (albo średnia) wartości rang po prostu. W skali strachu przed COVID ten kto się najbardziej boi powinien odpowiedzieć 7 razy **zdecydowanie tak** co odpowiada sumie 35 rang (jeżeli rangujemy od 1 do 5). Ten który się wcale nie boi zaś będzie miał 7.

Małym utrudnieniem mogą być **pytania odwrócone**. Jeżeli pytamy o strach przed COVID i w każdym pytaniu jak bardzo ktoś się boi, albo jak bardzo mu serce bije, ale w jednym z pytań zapytamy **nie boję się COVID**, to ranga 5 odpowiada uczuciu **braku strachu**. Rangi w pytaniach odwróconych należy przeliczyć (odwrócić): 1 zamienić na 5, 2 na 4 itd… Jeżeli używamy cudzych skal to w opisie powinno być wskazane które pytania są odwrócone.

Zalecany schemat postępowania jeżeli w ankiecie mają być mierzone przekonania (strach, samoskuteczność, wypalenie zawodowe, stress czy satysfakcja):

- Dokształcamy się nieco z psychologii mimo wszystko.
- Robimy przegląd literatury i znajdujemy skalę, którą ktoś już wymyślił żeby mierzyć to co my chcemy zmierzyć, bo **raczej nie należy wymyślać własnych skal**.
- Robimy ankietę (w Internecie) i zbieramy dane.
- Wykonujemy analizę statystyczną.

Banalnie proste co udowodnimy na przykładach

5.2 Wiedza na temat szkodliwości palenia i jej uwarunkowania wśród studentów PSW

5.2.1 Cel

Celem jest ocena wielkości zjawiska palenia tytoniu oraz poziom wiedzy na temat szkodliwości palenia tytoniu wśród studentów RM/PO PSW oraz zweryfikowanie wpływu wybranych czynników warunkujących ten nałóg.

Postawiono następujące hipotezy badawcze:

1. Jaka jest wielkość zjawiska palenie tytoniu wśród studentów PSW?
2. Jaka jest wiedza na temat szkodliwości palenia tytoniu wśród studentów PSW?
3. Czy palenie jest skorelowane z płcią, stażem pracy i miejscem pracy?
4. Czy wiedza na temat szkodliwości palenie jest skorelowana z płcią, stażem pracy i miejscem pracy?
5. Czy palenie jest skorelowane z wiedzą na temat szkodliwości palenia?

5.2.2 Metoda

Badanie ankietowe wśród studentów RM oraz PO przeprowadzono w styczniu 2023. Ankieta zawierała pytania dotyczące palenia tytoniu (pali/nie pali/palił, jak długo pali itd), test wiedzy na temat szkodliwości palenia oraz pytania o rodzaj miejsca pracy, staż pracy i płeć itd.

Pięć następujących pytań oceniało wiedzę ankietowanego na temat szkodliwości palenia:

- Czy bardziej szkodliwe dla zdrowia jest czynne czy bierne palenie? (JW),
- Jakie według Ciebie choroby układu oddechowego mogą być spowodowane bezpośrednio przez palenie papierosów? (WW)
- Czy palenie papierosów powoduje choroby układu pokarmowego? (JW)

- Jakie według Ciebie choroby kardiologiczne mogą być spowodowane bezpośrednio przez palenie papierosów? (WW)
- Jaki według Ciebie ma wpływ palenie papierosów na narządy zmysłów? (WW)

W przypadku pytań jednokrotnego wyboru (JW), za wskazanie poprawnej odpowiedzi respondent otrzymywał 1 punkt. W przypadku pytań wielokrotnego wyboru (WW) za wskazanie prawidłowej odpowiedzi respondent otrzymywał 1 punkt, ale za wskazanie nieprawidłowej otrzymywał (minus) -1 punkt (aby nie opłacała się strategia zaznaczenia wszystkich odpowiedzi). Maksymalna możliwa do uzyskania liczba punktów wynosiła 19.

5.2.3 Zastosowane metody statystyczne

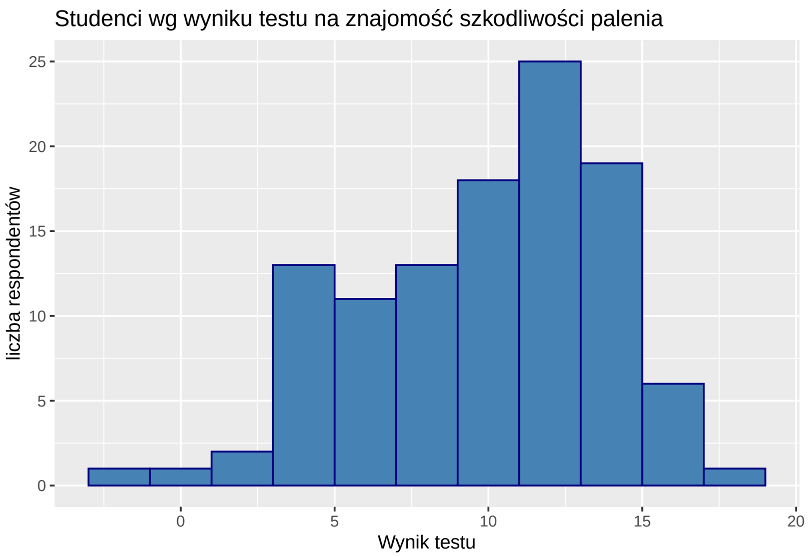
- Hipotezę 1 weryfikowano na podstawie wielkości odsetka respondentów palących.

- Hipotezę 2 weryfikowano na podstawie wielkości odsetka respondentów wykazujących się dobrą i bardzo dobrą wiedzą na temat palenia

- Hipotezy 3--5 zweryfikowano z wykorzystaniem tablic wielodzielnych/testu chi-kwadrat oraz porównania średniego poziomu depresji w grupach za pomocą testów Manna-Whitneya oraz Kruskala-Wallisa.

5.2.4 Metryczka (analiza respondentów)

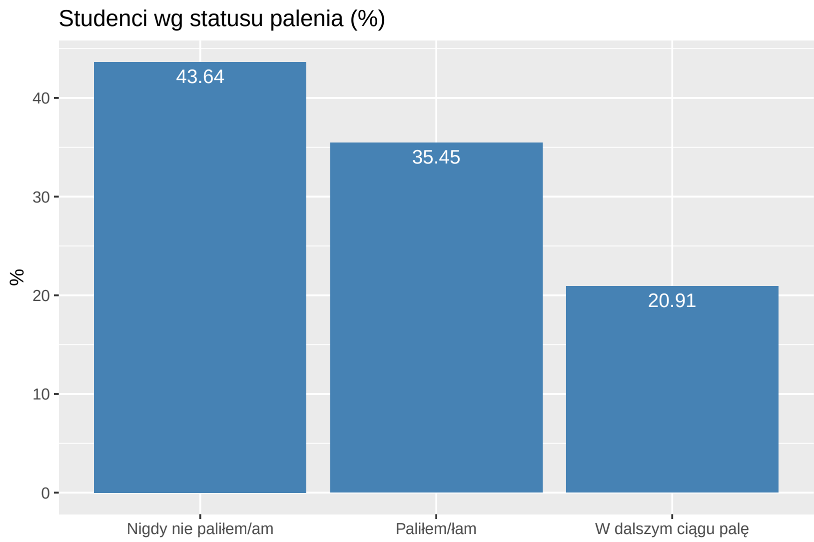
Rozkład ankietowanych wg wyniku testu na znajomość szkodliwości palenia przedstawiono na rysunku.



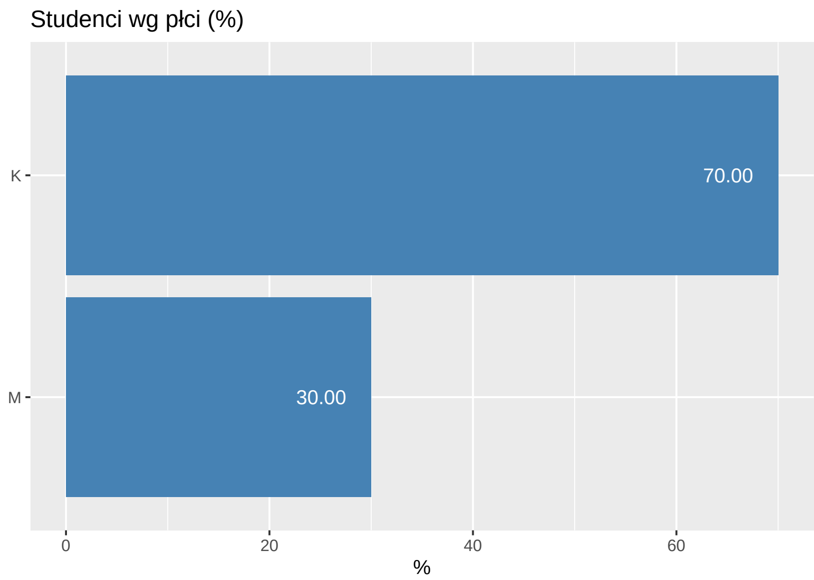
W badaniu wzięło udział 110 studentów. Otrzymano 110 poprawnie wypełnionych ankiet. Średnia wartość testu oceniającego wiedzę wyniosła 10.364 (odchylenie standardowe 3.997).

Rozkład ankietowanych ze względu na status względem palenia przedstawiono na rysunku.

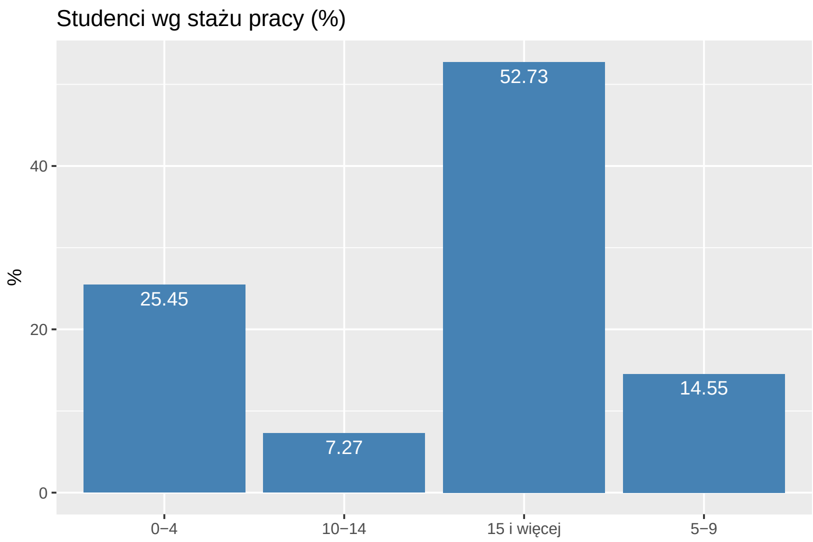
5.2. WIEDZA NA TEMAT SZKODLIWOŚCI PALENIA I JEJ UWARUNKOWANIA WŚRÓD STUDENTÓW PSW



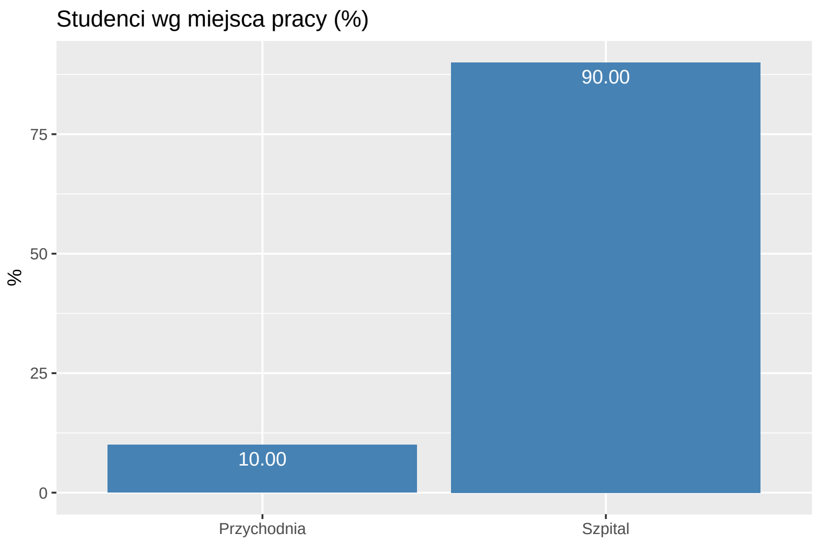
Rozkład ankietowanych ze względu na płeć przedstawiono na rysunku



Rozżad ankietowanych ze względu na staż pracy przedstawiono na rysunku



Rozkład ankietowanych ze względu na rodzaj miejsca pracy przedstawiono na rysunku



5.2.5 Weryfikacja hipotezy 1

Palą lub paliło 62 respondentów (56.36 %). Żeby stwierdzić czy to jest dużo czy mało to na przykład można by porównać z jakąś średnią ogólnopolską.

5.2.6 Weryfikacja hipotezy 2

Średnia wartość uzyskana w teście wyniosła 10.36 (mediana 11); 3/4 respondentów nie uzyskało więcej niż 13 (czyli 68.4 %).

5.2.7 Weryfikacja hipotez 3--5

Czy palenie jest skorelowane z płcią?

	K	M
Nigdy nie paliłem/am	31	17
Paliłem/łam	27	12
W dalszym ciągu palę	19	4

```
##
##  Pearson's Chi-squared test
##
## data:  t.sex.f
## X-squared = 2.4228, df = 2, p-value = 0.2978
```

Nie jest o czym świadczy wysoka wartość p (0.2978)

Czy palenie jest skorelowane ze stażem pracy?

	0-4	10-14	15 i więcej	5-9
Nigdy nie paliłem/am	14	4	24	6
Paliłem/łam	8	2	23	6
W dalszym ciągu palę	6	2	11	4

```
##
##  Pearson's Chi-squared test
##
## data:  t.staz.f
## X-squared = 1.7687, df = 6, p-value = 0.9397
```

Nie jest o czym świadczy wysoka wartość p (0.9397)

Czy palenie jest skorelowane z miejscem pracy?

	Przychodnia	Szpital
Nigdy nie paliłem/am	5	43
Paliłem/łam	3	36
W dalszym ciągu palę	3	20

```
##
##  Pearson's Chi-squared test
##
## data:  t.praca.f
## X-squared = 0.47674, df = 2, p-value = 0.7879
```

Nie jest o czym świadczy wysoka wartość p (0.7879)

Czy wiedza na temat palenia jest skorelowana z płcią:

płeć	średnia	n
K	10.831169	77
M	9.272727	33

```
## $p.value
## [1] 0.04883299
```

Porównujemy dwie grupy zatem stosujemy test Manna-Whitneya. Wartość p wynosi 0.04883299 zatem hipotezy o braku korelacji na poziomie 0,05 należy odrzucić. Istnieje zależność pomiędzy wiedzą na temat palenia a płcią.

Czy wiedza na temat palenia jest skorelowana z miejscem pracy:

m.pracy	średnia	n
Przychodnia	10.36364	11
Szpital	10.36364	99

```
## $p.value
## [1] 0.8026325
```

Porównujemy dwie grupy zatem stosujemy test Manna-Whitneya. Wartość p wynosi 0.80263254 -- nie ma podstaw od odrzucenia hipotezy o braku korelacji na poziomie 0,05.

Czy wiedza na temat palenia jest skorelowana ze stażem:

staż	średnia	n
0-4	9.928571	28
10-14	9.875000	8
15 i więcej	10.793103	58
5-9	9.812500	16

```
## $p.value
## [1] 0.6771844
```

Porównujemy więcej niż dwie grupy zatem stosujemy test Kruskala-Wallisa. Wartość p wynosi 0.67718441 -- nie ma podstaw od odrzucenia hipotezy o braku korelacji na poziomie 0,05.

Czy wiedza o szkodliwości palenia jest skorelowana ze statusem względem palenia? Chcemy zastosować tablicę wielodzielczą/test chi kwadrat. Musimy zatem zamienić skalę liczbową zmiennej mierzącej wiedzę nt szkodliwości palenia na nominalną, np tak: 0--5 mała; 6--10 średnia; 11--15 duża, 16--19 ogromna:

	Nigdy nie paliłem/am	Paliłem/łam	W dalszym ciągu palę
duża	22	16	14
mała	7	6	4
ogromna	2	3	2
średnia	17	14	3

```
##
##  Pearson's Chi-squared test
##
## data:  wiedza.status.t
## X-squared = 4.9954, df = 6, p-value = 0.5444
```

Widza i status wzg. palenia nie jest skorelowana o czym świadczy wysoka wartość p (0.5444)

Można to samo zweryfikować porównując średnie w grupach i stosując test Kruskala-Wallisa

status	średnia	n
Nigdy nie paliłem/am	10.31250	48
Paliłem/łam	10.07692	39
W dalszym ciągu palę	10.95652	23

```
## $p.value
## [1] 0.7787663
```

Wynik jest identyczny (wysoka wartość p 1).

5.2.8 Wnioski

- Ponad połowa studentów pali lub paliła.

- Istnieje zależność pomiędzy wiedzą o szkodliwości palenia a płcią.
- Nie ma związku pomiędzy statusem względem palenia a płcią, miejscem pracy i stażem.
- Nie ma związku pomiędzy wiedzą o szkodliwości palenia a miejscem pracy i stażem.

5.3 Depresja i jej uwarunkowania wśród studentów PSW

5.3.1 Cel

Celem jest ustalenie czy depresja jest istotnym problemem wśród studentów RM/PO PSW oraz zweryfikowanie wybranych czynników warunkujących depresję.

5.3.2 Metoda

Badanie ankietowe wśród studentów RM oraz PO przeprowadzono w styczniu 2023. Ankieta zawierała test samooceny depresji Becka oraz pytania o rodzaj miejsca pracy, staż pracy i płeć.

Test samooceny depresji Becka składa się z 21 pytań. W każdym pytaniu możliwe są 4 warianty odpowiedzi, odpowiadające zwiększonej intensywności objawów depresji, którym w związku z tym przypisuje się wartości od zera do 3 punktów. Maksymalna liczba punktów w teście wynosi 63 a minimalna 0.

Interpretacja wyników testu Becka: 0--19 brak/łagodna depresja; 20--25 umiarkowana; 26--63 cieżka depresja.

Postawiono następujące hipotezy badawcze:

1. Depresja stanowi duży problem wśród studentów PSW.
2. Problem depresji zależy od miejsca pracy.
3. Problem depresji zależy od stażu pracy.
4. Problem depresji zależy od płci.

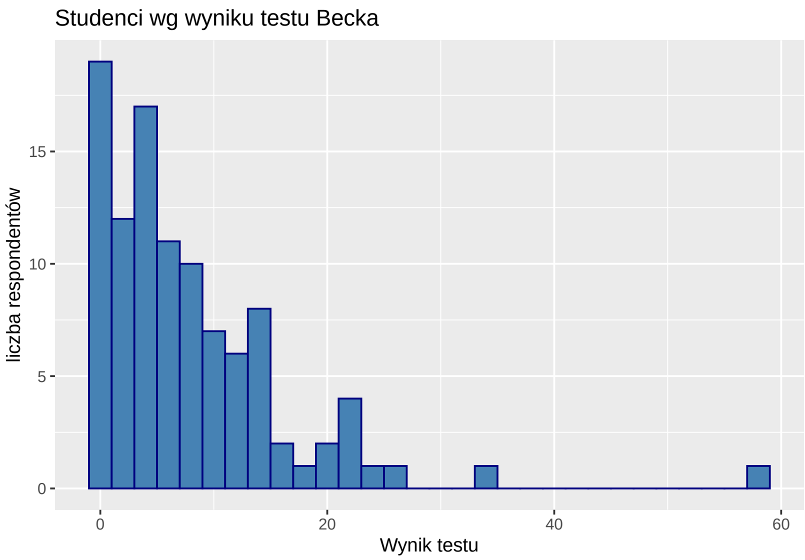
5.3.3 Zastosowane metody statystyczne

- Hipotezę 1 oceniono na podstawie odsetka respondentów wykazujących ciężką postać depresji.
- Hipotezy 2--4 zweryfikowano z wykorzystaniem tablic wielodzielczych/testu chi-kwadrat oraz porównania średniego poziomu depresji w grupach za pomocą testów Manna-Whitneya oraz Kruskala-Wallisa

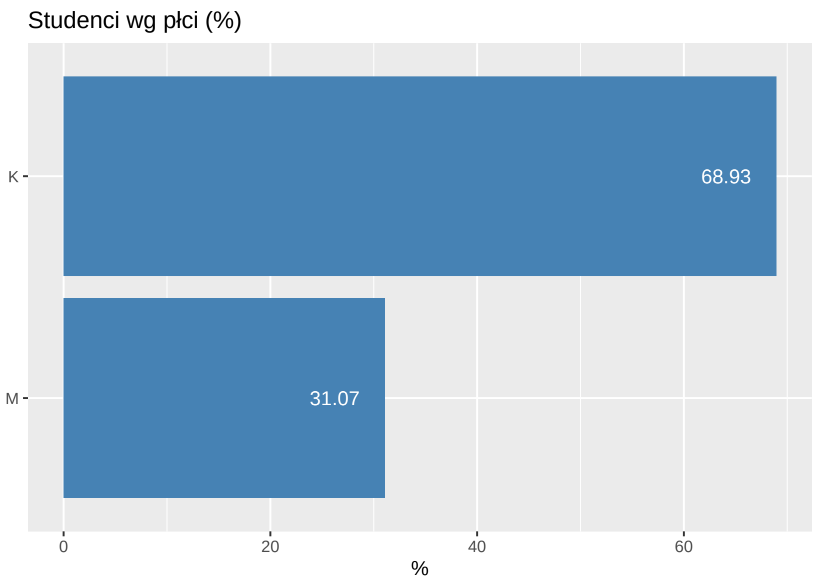
5.3.4 Metryczka

W badaniu wzięło udział 103 studentów. Otrzymano 103 poprawnie wypełnionych ankiet. Średnia wartość testu Becka wyniosła 8.379 (odchylenie standardowe 8.577)

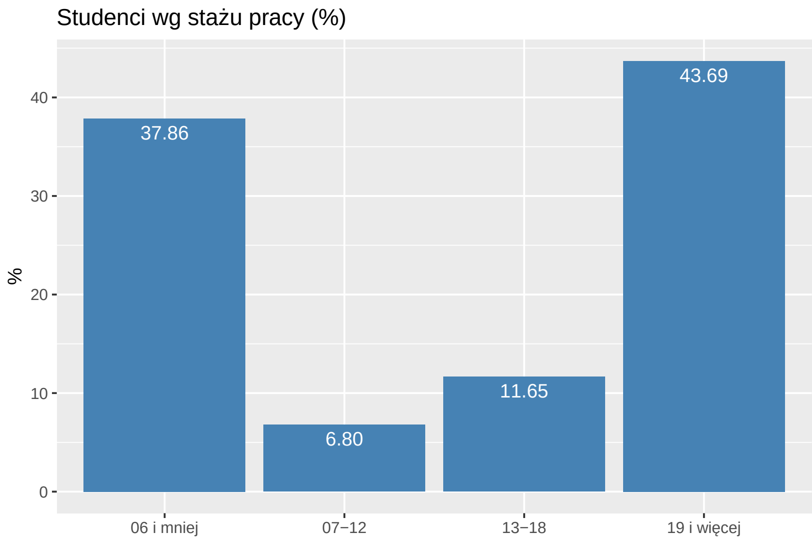
Rozkład ankietowanych wg wyniku testu Becka przedstawiono na rysunku.



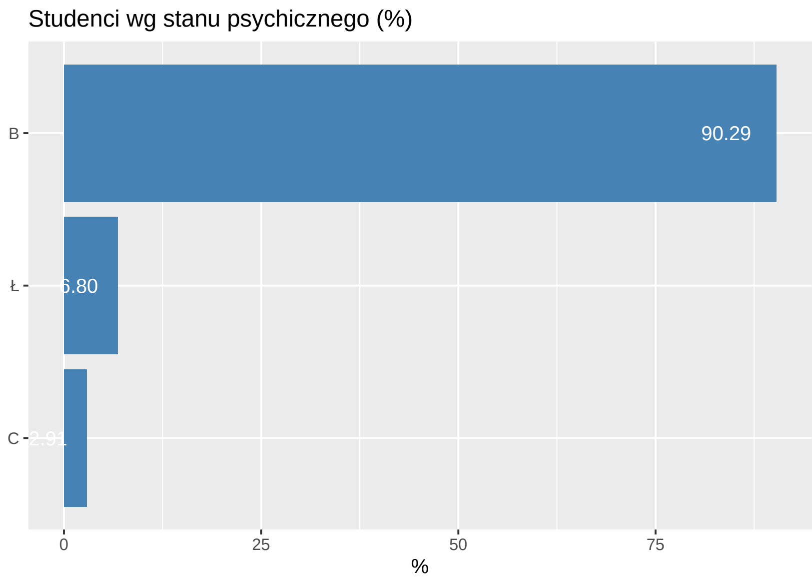
Rozkład ankietowanych ze względu na płeć przedstawiono na rysunku.



Rozkład ankietowanych ze względu na staż przedstawiono na rysunku.



5.3.5 Weryfikacja hipotezy 1



Ciężką postać depresji wykazuje zaledwie 3% studentów. Należy odrzucić hipotezę że depresja stanowi poważny problem wśród studentów RM/PO PSW.

5.3.6 Weryfikacja hipotez 2--4

Aby móc zastosować metody tablicy wielodzielczej i testu chi-kwadrat oryginalne wartości liczbowe depresji zamieniono na skalę porządkową: 0--19 brak/łagodna depresja (B); 20--25 umiarkowana (Ł); 26--63 ciężka depresja (C).

5.3.7 Depresja a płeć

Tablica wielodzielcza i test chi-kwadrat:

	K	M
B	64	29
C	2	1
Ł	5	2

```
##
##  Pearson's Chi-squared test
##
## data:  dep.sex.f
## X-squared = 0.028134, df = 2, p-value = 0.986
```

Nie stwierdzono zależności pomiędzy depresją a płcią (p = 0.986).

5.3.8 Depresja a staż

Tablica wielodzielcza i test chi-kwadrat:

	06 i mniej	07-12	13-18	19 i więcej
B	36	6	9	42
C	1	0	1	1
Ł	2	1	2	2

```
##
##  Pearson's Chi-squared test
##
## data:  dep.staz.f
## X-squared = 4.719, df = 6, p-value = 0.5803
```

Nie stwierdzono zależności pomiędzy depresją a stażem pracy (p = 0.5803).

Jeżeli depresję mierzymy na (oryginalnej) skali liczbowej można porównać wartości średnie i zastosować test Kruskala-Wallisa.

staż	średnia	n
06 i mniej	8.512821	39
07-12	7.857143	7
13-18	7.666667	12
19 i więcej	8.533333	45

```
## $p.value
## [1] 0.678923
```

Wynik jest ten sam (brak zależności)

5.3.9 Depresja a rodzaj miejsca pracy

Tablica wielodzielcza i test chi-kwadrat:

	Przychodnia	Szpital
B	12	81
C	0	3
Ł	0	7

```
##
##  Pearson's Chi-squared test
##
## data:  dep.praca.f
```

```
## X-squared = 1.4605, df = 2, p-value = 0.4818
```

Nie stwierdzono zależności pomiędzy depresją a miejscem pracy (p = 0.4818).

Jeżeli depresję mierzymy na skali liczbowej można porównać wartości średnie i zastosować test Manna-Whitneya

m-pracy	średnia	n
Przychodnia	7.833333	12
Szpital	8.450549	91

```
## $p.value
## [1] 0.8528214
```

Wynik jest ten sam (brak zależności)

5.4 Wnioski

- Depresja nie jest istotnym problemem wśród studentów RM/PO PSW.
- Nie ma związku pomiędzy depresją a stażem, płcią i miejscem pracy.

5.5 Satysfakcja, przywiązanie i zamiar odejścia

W przygotowaniu na podstawie ankiety: https://docs.google.com/forms/d/1bgNMhPEJMjiWeLfrmEAnPpns5U7Rooc-KOaeS7XlLV0/edit

5.6 Formularze ankiet

5.6.1 Skala Depresji Becka

Pytania

Pytanie 1

0. Nie jestem smutny ani przygnębiony.
1. Odczuwam często smutek, przygnębienie
2. Przeżywam stale smutek, przygnębienie i nie mogę uwolnić się od tych przeżyć.
3. Jestem stale tak smutny i nieszczęśliwy, że jest to nie do wytrzymania.

Pytanie 2

0. Nie przejmuję się zbytnio przyszłością.
1. Często martwię się o przyszłość.
2. Obawiam się, że w przyszłości nic dobrego mnie nie czeka.
3. Czuję, że przyszłość jest beznadziejna i nic tego nie zmieni.

Pytanie 3

0. Sądzę, że nie popełniam większych zaniedbań.
1. Sądzę, że czynię więcej zaniedbań niż inni.
2. Kiedy spoglądam na to, co robiłem, widzę mnóstwo błędów i zaniedbań.
3. Jestem zupełnie niewydolny i wszystko robię źle.

Pytanie 4

0. To, co robię, sprawia mi przyjemność.
1. Nie cieszy mnie to, co robię.
2. Nic mi teraz nie daje prawdziwego zadowolenia.
3. Nie potrafię przeżywać zadowolenia i przyjemności; wszystko mnie nuży.

Pytanie 5

0. Nie czuję się winnym ani wobec siebie, ani wobec innych.
1. Dość często miewam wyrzuty sumienia.
2. Często czuję, że zawiniłem.
3. Stale czuję się winny.

Pytanie 6

0. Sądzę, że nie zasługuję na karę
1. Sądzę, że zasługuję na karę
2. Spodziewam się ukarania
3. Wiem, że jestem karany (lub ukarany)

Pytanie 7

0. Jestem z siebie zadowolony
1. Nie jestem z siebie zadowolony
2. Czuję do siebie niechęć
3. Nienawidzę siebie

Pytanie 8

0. Nie czuję się gorszy od innych ludzi
1. Zarzucam sobie, że jestem nieudolny i popełniam błędy
2. Stale potępiam siebie za popełnione błędy
3. Winię siebie za wszelkie zło, które istnieje

Pytanie 9

0. Nie myślę o odebraniu sobie życia
1. Myślę o samobójstwie --- ale nie mógłbym tego dokonać
2. Pragnę odebrać sobie życie
3. Popełnię samobójstwo, jak będzie odpowiednia sposobność

Pytanie 10

0. Nie płaczę częściej niż zwykle
1. Płaczę częściej niż dawniej
2. Ciągle chce mi się płakać
3. Chciałbym płakać, lecz nie jestem w stanie

Pytanie 11

0. Nie jestem bardziej podenerwowany niż dawniej
1. Jestem bardziej nerwowy i przykry niż dawniej
2. Jestem stale zdenerwowany lub rozdrażniony
3. Wszystko, co dawniej mnie drażniło, stało się obojętne

Pytanie 12

0. Ludzie interesują mnie jak dawniej
1. Interesuję się ludźmi mniej niż dawniej
2. Utraciłem większość zainteresowań innymi ludźmi

3. Utraciłem wszelkie zainteresowanie innymi ludźmi

Pytanie 13

0. Decyzje podejmuję łatwo, tak jak dawniej
1. Częściej niż kiedyś odwlekam podjęcie decyzji
2. Mam dużo trudności z podjęciem decyzji
3. Nie jestem w stanie podjąć żadnej decyzji

Pytanie 14

0. Sądzę, że wyglądam nie gorzej niż dawniej
1. Martwię się tym, że wyglądam staro i nieatrakcyjnie
2. Czuję, że wyglądam coraz gorzej
3. Jestem przekonany, że wyglądam okropnie i odpychająco

Pytanie 15

0. Mogę pracować jak dawniej
1. Z trudem rozpoczynam każdą czynność
2. Z wielkim wysiłkiem zmuszam się do zrobienia czegokolwiek
3. Nie jestem w stanie nic zrobić

Pytanie 16

0. Sypiam dobrze, jak zwykle
1. Sypiam gorzej niż dawniej
2. Rano budzę się 1--2 godziny za wcześnie i trudno jest mi ponownie usnąć
3. Budzę się kilka godzin za wcześnie i nie mogę usnąć

Pytanie 17

0. Nie męczę się bardziej niż dawniej
1. Męczę się znacznie łatwiej niż poprzednio.
2. Męczę się wszystkim, co robię.
3. Jestem zbyt zmęczony, aby cokolwiek robić.

Pytanie 18

0. Mam apetyt nie gorszy niż dawniej
1. Mam trochę gorszy apetyt
2. Apetyt mam wyraźnie gorszy
3. Nie mam w ogóle apetytu

Pytanie 19

0. Nie tracę na wadze (w okresie ostatniego miesiąca)
1. Straciłem na wadze więcej niż 2 kg
2. Straciłem na wadze więcej niż 4 kg
3. Straciłem na wadze więcej niż 6 kg

Pytanie 20

0. Nie martwię się o swoje zdrowie bardziej niż zawsze
1. Martwię się swoimi dolegliwościami, mam rozstrój żołądka, zaparcie, bóle
2. Stan mojego zdrowia bardzo mnie martwi, często o tym myślę
3. Tak bardzo martwię się o swoje zdrowie, że nie mogę o niczym innym myśleć

Pytanie 21

0. Moje zainteresowania seksualne nie uległy zmianom
1. Jestem mniej zainteresowany sprawami płci (seksu)
2. Problemy płciowe wyraźnie mniej mnie interesują
3. Utraciłem wszelkie zainteresowanie sprawami seksu

**Treść pytań* (nie prezentowana ankietowanym): Odczuwanie smutku i przygnębienia (1), Martwienie się o przyszłość (2), Uważasz, że zaniedbujesz swoje obowiązki? (3), Jesteś zadowolony z siebie? (4), Czy często masz poczucie winy? (5), Czy zasługujesz na karę? (6), Zadowolenie z siebie (7), Czy czujesz się gorszy od innych? (8), Czy masz myśli samobójcze? (9), Często chce Ci się płakać? (10), Jesteś ostatnio bardziej nerwowy i rozdrażniony? (11), Czy zmieniło się coś w Twoim zainteresowaniu innymi ludźmi? (12), Czy ostatnio miewasz większe problemy z podejmowaniem różnych decyzji? (13), Czy uważasz, że wyglądasz gorzej i mniej atrakcyjnie niż kiedyś? (14), Czy masz większe trudności z wykonywaniem różnych prac i zadań? (15), Masz kłopoty ze snem? (16), Czy męczysz się bardziej niż zwykle? (17), Czy masz kłopoty z apetytem? (18), W ciągu ostatniego miesiąca nie stosowałem diety, aby schudnąć, lecz straciłem na wadze (19), Czy ostatnio bardziej martwisz się swoim stanem zdrowia? (20), Czy masz kłopoty z potencją? (21).

Interpretacja wyników

Punkty	Depresja
0--11	Brak
12--19	Łagodna
20--25	Umiarkowana
26--63	Ciężka

Źródło: https://psychiatra.bydgoszcz.eu/publikacje-dla-pacjenta/depresja/skala-depresji-becka/ oraz http://centrum-psychologiczne.com/files/files/Skala_Depresji_Becka_word.pdf

5.6.2 Ankieta Palenie

Poziom wiedzy personelu pielęgniarskiego na temat szkodliwości palenia tytoniu

Pytania

1. Płeć ☐ Kobieta ☐ Mężczyzna
2. Wiek
3. Pochodzenie ☐ wieś ☐ Miasto do 20 tys. mieszkańców ☐ Miasto powyżej 20 tyś mieszkańców
4. Wykształcenie ☐ Średnie medyczne ☐ Licencjat pielęgniarstwa ☐ Magister pielęgniarstwa ☐ Inne wyższe
5. Staż pracy ☐ Mniej niż rok ☐ 1-10 lat ☐ 10-15 lat ☐ Więcej niż 15 lat
6. Miejsce pracy ☐ Oddział zabiegowy ☐ Oddział zachowawczy ☐ Przychodnia/poradnia
7. Czy kiedykolwiek paliłeś/aś papierosy? ☐ Paliłem/łam ☐ W dalszym ciągu palę ☐ Nigdy nie paliłem/am
8. Od ilu lat palisz? ☐ Nie palę/Nie dotyczy ☐ Mniej niż rok ☐ 1-10 lat ☐ 11-15 lat ☐ Więcej niż 15 l
9. Czy zdarza Ci się palić w miejscu pracy? ☐ Tak ☐ Nie ☐ Nie dotyczy
10. Czy próbowałeś kiedykolwiek rzucić palenie? ☐ Tak, udało się ☐ Tak, ale wróciłem/am do nałogu ☐ Nie ☐ Nie dotyczy

11. Czy paląc przyznajesz się do uzależnienia ☐ Tak ☐ Nie ☐ Nie dotyczy
12. Uważasz, że bardziej szkodliwe dla zdrowia jest: ☐ Palenie czynne - dym tytoniowy wdychany bezpośrednio przez palacza ☐ Palenie bierne-boczny strumień dymu ☒ Każda forma kontaktu z dymem jest równie szkodliwa ☐ Nie wiem/Nie mam zdania
13. Czy uważasz,że przepisy prawa powinny zabraniać palenia w obecności dzieci poniżej 15 roku życia? ☐ Tak ☐ Nie ☐ Nie wiem/nie mam zdania
14. Czy przerwa na papierosa pomaga Ci w sytuacjach stresowych? ☐ Tak ☐ Nie ☐ Nie dotyczy
15. Czy palenie nikotyny sprawia Ci przyjemność? ☐ Tak ☐ Nie ☐ Nie dotyczy
16. Jakie według Ciebie choroby układu oddechowego mogą być spowodowane bezpośrednio przez palenie papierosów?(wielokrotnego) ☒ Przewlekła obturacyjna choroba płuc ☒ Astma oskrzelowa ☒ Alergie wziewne ☐ Gruźlica ☐ Zapalenie płuc ☒ Przewlekłe zapalenie oskrzeli ☒ Infekcje dróg oddechowych ☐ Palenie nie powoduje chorób układu oddechowego ☐ Nie wiem
17. Jakie według Ciebie choroby kardiologiczne mogą być spowodowane bezpośrednio przez palenie papierosów? (wielokrotnego) ☒ Nadciśnienie tętnicze krwi ☒ Zawał mięśnia sercowego ☒ Udar mózgu ☒ Choroba niedokrwienna serca ☒ Miażdżyca tętnic obwodowych ☒ Zaburzenie rytmu serca ☒ Choroba Buergera ☐ Hipercholesterolemia ☐ Tętniak aorty ☐ Palenie nie powoduje chorób kardiologicznych
18. Czy palenie papierosów powoduje choroby układu pokarmowego? ☒ Tak ☐ Nie ☐ Nie wiem
19. Jaki według Ciebie ma wpływ palenie papierosów na narządy zmysłów? (wielokrotnego) ☒ Upośledza węch i smak ☒ Powoduje podrażnienie spojówek ☒ Obniża apetyt ☒ Niszczy struny głosowe ☒ Zmniejsza ostrość wzroku ☐ Palenie nie ma negatywnego wpływu na narządy zmysłu.
20. Jak oceniasz swoją wiedzę na temat palenia papierosów i jego wpływu na zdrowie człowieka? ☐ Bardzo dobrze ☐ Dobrze ☐ Przeciętnie ☐ Źle
21. Czy wzorce sięgania po tytoń wyniosłaś/wyniosłeś z domu rodzinnego ☐ Tak ☐ Nie ☐ Nie dotyczy

(Prawidłowe odpowiedzi zaznaczono ☒)

Źródło: Na podstawie ankiety z pracy Izabeli Marendowskiej *Ocena poziomu wiedzy pielęgniarek na temat szkodliwości palenia papierosów* (promotor: dr Marzeny Barton); Kwidzyn 2022.

5.6.3 Ankieta: satysfakcja, przywiązanie i zamiar odejścia

Rozdział 6

Łagodne wprowadzenie z Jamovi

Jak wspomniano w rozdziale 1, statystkę można uprawiać (tj. liczyć statystyki w drugim tego słowa znaczeniu :-)) wykorzystując różne programy. My zdecydowaliśmy się promować **Jamovi**, program który naszym zdaniem jest najlepszym -- z punktu widzenia większości studentów Nauk o Zdrowiu -- połączeniem ceny, możliwości, prostoty i łatwości nauki.

6.1 Podstawy pracy z Jamovi

Jamovi jest oprogramowaniem rozpowszechnianym na licencji typu *Open Source*, a więc można go używać za darmo. Program jest dostępny ze strony https://www.jamovi.org/download.html. Klikamy, ściągamy, uruchamiamy instalator. Program jest dość duży, ale to nie jest aż tak wielki problem w czasach kiedy pojemności dysków w domowym komputerze zaczynają się od 250 gigabajtów. Jest też wersja Jamovi365, którą moża się posługiwać w przeglądarce i wtedy nic nie trzeba instalować na swoim komputerze.

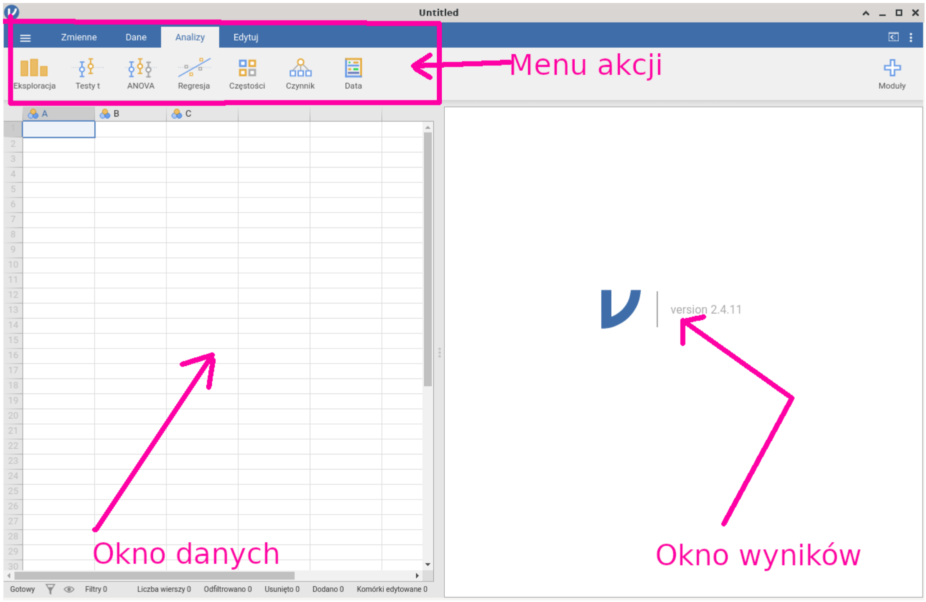
W naszym podręczniku pokazujemy jak uprawiać statystykę z tradycyjną wersją Jamovi. Po zainstalowaniu uruchamiamy program, którego ekrano startowy wygląda jak na rysunku 6.1.

Menu akcji umożliwia wykonanie podstawowych akcji:

- wczytanie danych i zapisanie danych (pierwsza pozycja menu oznaczona jako trzy poziome kreski)
- podgląd (w sensie skontrolowania wartości zmiennych) i modyfikację danych (pozycje **Zmienne** oraz **Dane**)
- wykonanie obliczeń (pozycja **Analizy**)
- modyfikowanie raportu (pozycja **Edit**)

Typowa sesja w **Jamovi**:

1. Wczytanie danych z pliku o praktycznie dowolnym formacie. Jeżeli przykładowo dane są wynikiem wykonania badania ankietowego z wykorzystaniem Formularzy Google to zalecamy posługiwanie się formatem CSV.
2. Transformacja danych. Przekodowanie wartości nominalnych na rangi. Przekodowanie wartości liczbowych na nominalne. Odwrócenie pytań odwróconych. Obliczenie sum/srednich rang dla wielu zmiennych.
3. Wykonanie obliczeń:
    1. Analiza struktury (**Eksploracja**),



Rysunek 6.1: Ekran startowy Jamovi

2. Analiza zależności między zmienną liczbową a nominalną (*testy t*/**ANOVA**),
3. Analiza zależności między zmiennymi liczbowymi: współczynnik korelacji liniowej/macierz korelacji (*Regresja*)
4. Analiza zależności między zmienną liczbową a zmiennymi liczbowymi/nominalnymi: regresja liniowa i logistyczna (**Regresja**)
5. Analiza zależności między zmiennymi nominalnymi: tablica wielodzielcza, test chi-kwadrat zgodności (*Częstości*)

Wykonanie obliczeń jest banalnie proste i sprowadza się do wybrania myszką odpowiednich zmiennych oraz procedury która ma być wykonana. Wynik obliczeń pojawia się natychmiast w **oknie wyników**. Jeżeli coś nam nie wyszło można procedurę poprawić a poprzedni wynik usunąć z okna wynikowego.

4. Zapisania danych (pozycja trzy poziome kreski). Po skończeniu pracy wynik można zapisać żeby np. wysłać wykładowcy lub nie zaczynać od zera jeżeli będziemy musieli pracę kontynuuować bo wykładowca chciał żebyśmy coś poprawili.

6.2 Analiza ankiety: satysfakcja--wiedza o paleniu--zamiar odejścia

Przykład nieco absurdalny, ale za to w zwartej postaci ilustrujący praktyczne sposoby transformacji danych oraz wykorzystania wszystkich procedur omawianych w podręczniku.

6.2.1 Wczytanie danych

W wyniku przeprowadzenia badania ankietowego zebrano za pomocą Formularza Google dane dotyczące satysfkacji/zamiaru odejścia oraz wiedzy nt. szkodliwości palenia tytoniu. Wyniki wyeksportowano do arkusza kalkulacyjnego, którego początek wygląda jak na rysunku 6.2:

9:9 fx 2023-12-02 11:56:39

	A	B	C	D	E	F	G	H	I	J	K
1	Sygnatura c	Ogólnie rzecz biorąc	Ogólnie rzecz biorąc j	Ogólnie rzecz biorąc, lu	Jakie według Ciebie choroby u bezpośrednio przez palenie pa	Często poważnie rozwa	Zamierzam rzucić obecr	Zacząłem szukać innej pracy	Płeć	Wiek (w	Staż pracy
2	2023-12-01	Zdecydowanie się nie	Nie zgadzam się	Zgadzam się	Alergie wziewne, Zapalenie płuc, Infekcje dróg oddechowych				Kobieta	5	14
3	2023-12-02	Zdecydowanie się nie	Zdecydowanie się zga	Zdecydowanie się zgac	Przewlekła obturacyjna chorob	Zdecydowanie się nie zg	Zdecydowanie się zgadz	Zdecydowanie się zgadzam	Kobieta	43	20
4	2023-12-02	Nie zgadzam się	Zgadzam się	Zgadzam się	Przewlekła obturacyjna chorob	Zdecydowanie się nie zg	Zdecydowanie się nie zg	Zdecydowanie się nie zgadzar	Kobieta	30	5
5	2023-12-02	Zdecydowanie się nie	Nie zgadzam się	Nie zgadzam się	Astma oskrzelowa	Nie zgadzam się	Nie zgadzam się	Zdecydowanie się zgadzam	Kobieta	14	4
6	2023-12-02	Zdecydowanie się nie	Zdecydowanie się nie	Zdecydowanie się nie z	Przewlekła obturacyjna chorob	Zdecydowanie się nie zg	Zgadzam się	Zdecydowanie się zgadzam	Kobieta	44	20
7	2023-12-02	Nie zgadzam się	Zgadzam się	Trudno powiedzieć	Astma oskrzelowa	Trudno powiedzieć	Trudno powiedzieć	Zgadzam się	Kobieta	40	20
8	2023-12-02	Zdecydowanie się nie	Zdecydowanie się zga	Zdecydowanie się zgac	Przewlekła obturacyjna chorob	Zdecydowanie się nie zg	Zdecydowanie się nie zg	Zdecydowanie się nie zgadzar	Kobieta	49	30
9	2023-12-02	Nie zgadzam się	Zgadzam się	Zgadzam się	Zapalenie płuc	Nie zgadzam się	Nie zgadzam się	Nie zgadzam się	Kobieta	48	29
10	2023-12-02	Trudno powiedzieć	Trudno powiedzieć	Trudno powiedzieć	Przewlekła obturacyjna chorob	Trudno powiedzieć	Trudno powiedzieć	Trudno powiedzieć	Kobieta	35	10
11	2023-12-02	Nie zgadzam się	Zgadzam się	Zgadzam się	Przewlekła obturacyjna chorob	Nie zgadzam się	Nie zgadzam się	Nie zgadzam się	Kobieta	50	30

Rysunek 6.2: Fragment przykładowej ankiety

Ankieta składa się z 10 następujących pytań:

`Ogólnie rzecz biorąc nie lubię swojej pracy` (kolumna `B`), `Ogólnie rzecz biorąc jestem zadowolony ze swojej pracy` (`C`), `Ogólnie rzecz biorąc, lubię tu pracować` (`D`), `Jakie według Ciebie choroby układu oddechowego mogą być spowodowane bezpośrednio przez palenie papierosów?` (`E`), `Często poważnie rozważam odejście z obecnej pracy` (`F`), `Zamierzam rzucić obecną pracę` (`G`), `Zacząłem szukać innej pracy` (`H`), `Płeć` (`I`), `Wiek (w latach)` (`J`), oraz `Staż pracy` (`K`).

Ponadto Formularz Googla dodał automatycznie sygnaturę czasową jako zawartość pierwszej kolumny (`A`).

Zmieniamy wartości w pierwszym wierszu, który powinien zawierać nazwy zmiennych. Nazwy zmiennych powinny być jednowyrazowe i w miarę krótkie żeby się później można nimi wygodnie posługiwać. Jednocześnie nie powinny być za krótkie żeby od razu było widać jakie dane zawiera zmienna.

Jak widać pytania z kolumn `B--C` mierzą to samo (satysfakcję) więc zmieniamy im nazwę na bardziej zwartą `s1`, `s2` oraz `s3` (`s` od satysfakcja). Podobnie ponieważ pytania z kolumn `F--H` też mierzą to samo (zmiar odejścia), to też zmieniamy nazwy na coś krótszego: `zo1`, `zo2`, `zo3`. Kolumnę `E` nazywamy `wiedza_nt_palenia` a kolumny `I`, `J` oraz `K` odpowiednio: `plec`, `wiek` oraz `staz`.

Teraz arkusz wygląda jak na rysunku 6.3.

	A	B	C	D	E	F	G	H	I	J	K
1	Sygnatura czasowa	s1	s2	s3	wiedza_nt_palenia	zo1	zo2	zo3	plec	wiek	staz
2	2023-12-01 15:45:5	Zdecydowanie się n	Nie zgadzam się	Zgadzam się	Alergie wziewne, Zapalenie płuc, Infekcje dróg oddechowych				Kobieta	5	14
3	2023-12-02 11:55:1	Zdecydowanie się n	Zdecydowanie się z	Zdecydowanie się z	Przewlekła obturacy	Zdecydowanie się n	Zdecydowanie się z	Zdecydowanie się z	Kobieta	43	20
4	2023-12-02 11:55:2	Nie zgadzam się	Zgadzam się	Zgadzam się	Przewlekła obturacy	Zdecydowanie się n	Zdecydowanie się n	Zdecydowanie się n	Kobieta	30	5
5	2023-12-02 11:56:0	Zdecydowanie się n	Nie zgadzam się	Nie zgadzam się	Astma oskrzelowa	Nie zgadzam się	Nie zgadzam się	Zdecydowanie się z	Kobieta	14	4
6	2023-12-02 11:56:1	Zdecydowanie się n	Zdecydowanie się n	Zdecydowanie się n	Przewlekła obturacy	Zdecydowanie się n	Zgadzam się	Zdecydowanie się z	Kobieta	44	20
7	2023-12-02 11:56:3	Nie zgadzam się	Zgadzam się	Trudno powiedzieć	Astma oskrzelowa	Trudno powiedzieć	Trudno powiedzieć	Zgadzam się	Kobieta	40	20
8	2023-12-02 11:56:3	Zdecydowanie się n	Zdecydowanie się z	Zdecydowanie się z	Przewlekła obturacy	Zdecydowanie się n	Zdecydowanie się n	Zdecydowanie się n	Kobieta	49	30
9	2023-12-02 11:56:3	Nie zgadzam się	Zgadzam się	Zgadzam się	Zapalenie płuc	Nie zgadzam się	Nie zgadzam się	Nie zgadzam się	Kobieta	48	29
10	2023-12-02 11:56:4	Trudno powiedzieć	Trudno powiedzieć	Trudno powiedzieć	Przewlekła obturacy	Trudno powiedzieć	Trudno powiedzieć	Trudno powiedzieć	Kobieta	35	10
11	2023-12-02 11:56:4	Nie zgadzam się	Zgadzam się	Zgadzam się	Przewlekła obturacy	Nie zgadzam się	Nie zgadzam się	Nie zgadzam się	Kobieta	50	30

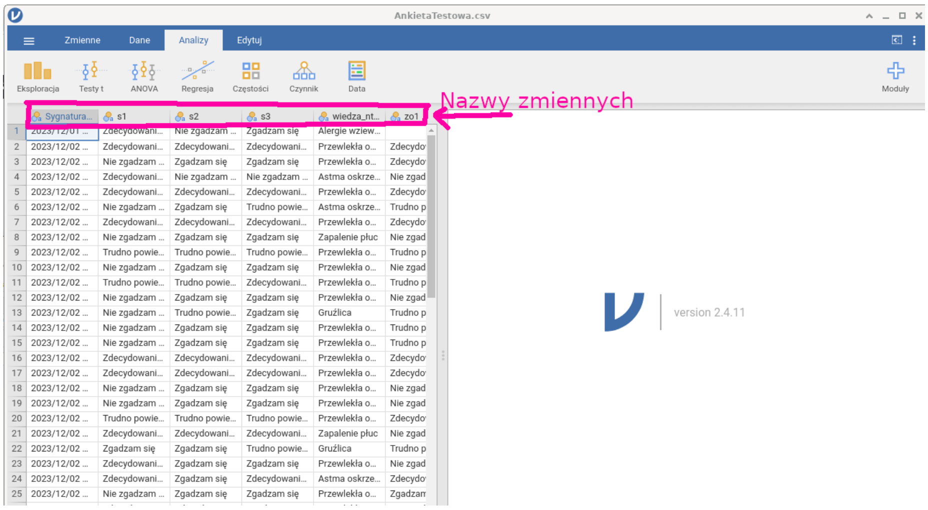
Rysunek 6.3: Fragment przykładowej ankiety

Arkusz eksportujemy wybierając format CSV. Bez problemu powinniśmy go wczytać do Jamovi (trzy poziome kreski →`Otwórz`). Jeżeli import się powiódł, to powinniśmy zobaczyć coś podobnego do tego co widać na rysunku 6.4.

Reasumując:

- Pytania oznaczone jako `s1`/`s2`/`s3` mierzą **satysfakcję z pracy**; pytania `zo1`/`zo2`/`zo3` mierzą **zamiar odejścia z pracy**. Pytania `s1--s3` oraz `zo1--zo3` są pytaniami jednokrotnego wyboru.
- Pytanie oznaczone jako `wiedza_nt_palenia` mierzy wiedzę na temat palenia tytoniu. Jest to przykład wykorzystania pytania z wielokrotnym wyborem.

- Pytania `plec`, `wiek`, `staz` mierzą płeć (`kobieta/mężczyna`), wiek (lata ukończone) oraz staż pracy (lata przepracowane)
- Pierwsza kolumna nie jest potrzebna ale jest dodawana przez aplikację Formularze Google.



Rysunek 6.4: Import danych

6.2.2 Przekodowanie danych

Zwykle zawartość arkusza zawierającego wyniki ankiety wymaga przekodowania. W naszym przykładzie należy wykonać:

- Zmienne `s1--s3` oraz `zo1--zo3` są mierzone w skali porządkowej. Wartości tych zmiennych chcemy zmienić (przekodować) na rangi wg schematu: `Zdecydowanie się nie zgadzam` = 1; `Nie zgadzam się` = 2; `Trudno powiedzieć` = 3 itd. Dodatkowo zauważmy że `s1` jest pytaniem odwróconym. W takich pytaniach należy przeliczyć rangi wg prostej formuły s1r = 6 - s1.
  - Miarą satysfakcji będzie suma rang s1r+s2+s3.
  - Miarą zamiaru odejścia będzie suma rang zo1+zo2+zo3
- Zmienna `plec` jest mierzona w skali nominalnej. Nie musimy jest przekodowywać
- Wartość zmiennej `wiedza_nt_palenia` należy przekodować na liczbę wg schematu: za wybranie poprawnej odpowiedzi plus jeden punkt; za wybranie błędnej odpowiedzi minus jeden punkt.
  - Miarą wiedzy nt. palenia będzy suma punktów uzyskanych za odpowiedzi prawidłowe minus suma punktów uzyskanych za odpowiedzi nieprawidłowe.

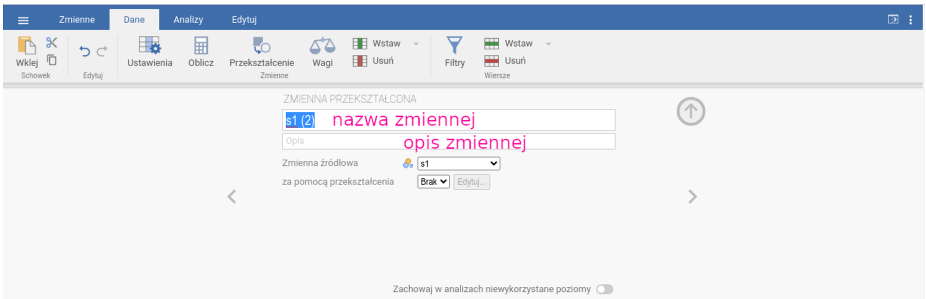
Uwaga: Sposób mierzenia wiedzy nt. palenia jest niepotrzebnie pokręcony; zamiast pytania z wielokrotnym wyborem spośród 8 możliwości/wariantów prościej jest zastosować 8 pytań Tak/Nie po czym pytania poprawne zsumować a pytania niepoprawne też dodać a wartość odjąć od sumy uzyskanej dla pytań poprawnych. My o tym wiemy, że tak jest bez sensu ale pokazujemy jako przykład przekodowania pytania z wielokrotnym wyborem.

- Wartości zmiennych `wiek` oraz `staz` są liczbami. Mogą być analizowane tak-jak-są (regresja/korelacja) ale można też je przekodować na wartości nominalne (mały-średni-duży staż) i zastosować metody z grupy zmienna-liczbowa/zmienna nominalna (takie jak test ANOVA czy Kruskala-Wallisa)

Przekodowanie wykonujemy wybierając **Dane** w menu głównym.

1. Klikamy w nazwę zmienną, którą zamierzamy przekodować. Niech to będzie `s1`. Kolumna po kliknięciu zmieni kolor.
2. Wybieramy ikonę `Przekształcenie`. Wypełniamy jak na rysunku 6.5.

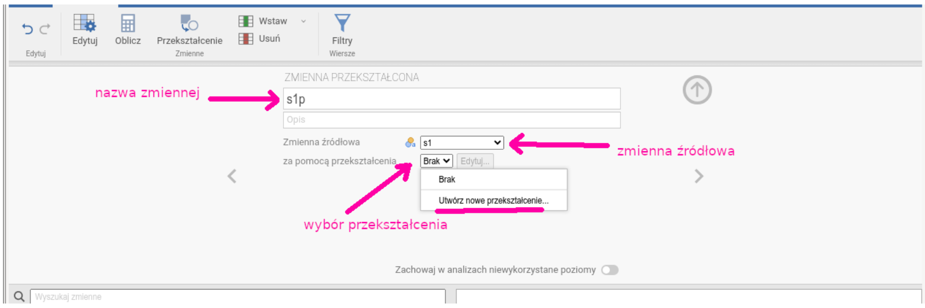
Uwaga: Jamovi nie zmieni wartości zmiennej `s1` tylko utworzy nową zmienną z przekodowanymi wartościami. Zmienna na podstawie której jest tworzona nowa zmienna nazywa się źródłową (`s1` w naszym przykładzie jest źródłową.)



Rysunek 6.5: Przekształcenie

Wpisujemy sensowną nazwę (na przykład `s1p` od przekodowana). Jak będziemy używać sensownych nazwa łatwiej będzie nam się pracowało. Dobrze jest też podać w opisie co zawiera zmienna.

Klikamy w pole wyboru na dole (obok napisu `za pomocą przekształcenia`) Powinniśmy zobaczyć coś podobnego do tego co widać na rysunku 6.6.



Rysunek 6.6: Przekształcenie

Wybieramy `Utwórz nowe przekształcenie`. Wpisujemy sensowną nazwę przekształcenia (na przykład `Likert2R5`) oraz formułę przekształcenia:

```
IF ($source=="Zdecydowanie się nie zgadzam", 1,
 IF ($source=="Nie zgadzam się", 2,
   IF ($source=="Trudno powiedzieć", 3,
      IF ($source=="Zgadzam się", 4, 5))))
```

Formuła może wydawać się przerażająca, ale jest koncepcyjnie bardzo prosta:

`IF (warunek, jeżeli-prawda, jeżeli-fałsz)`

Warunek to fragment `$source=="Zdecydowanie się nie zgadzam"`:

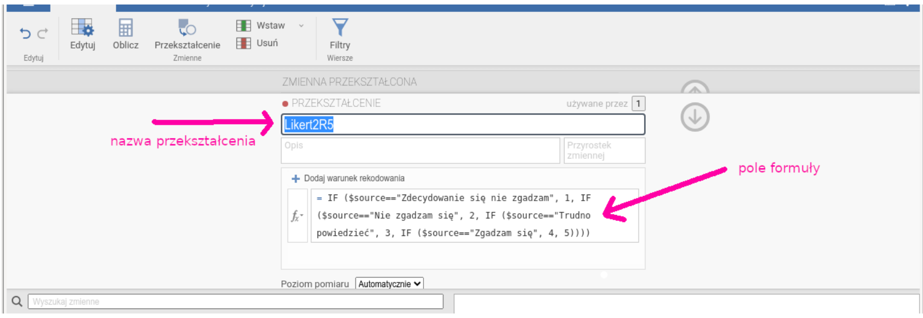
- `$source` oznacza bieżącą wartość zmiennej źródłowej
- == to **operator** równości; jest więcej operatorów, które można wybrać z menu

- `$source=="Zdecydowanie się nie zgadzam"` oznacza, że jeżeli bieżącą wartością w kolumnie źródłowej jest `Zdecydowanie się nie zgadzam` to wykonaj `jeżeli-prawda`; w wypadku przeciwnym wykonaj `jeżeli-fałsz`.

`jeżeli-prawda` to zwykle wstawienie nowej wartości; `jeżeli-fałsz` to często następna formuła `IF` albo wstawienie innej nowej wartości. Przykładowo jeżeli bieżącą wartością w kolumnie źródłowej jest `Zdecydowanie się nie zgadzam` to wstaw 1, jeżeli nie jest to wstaw 0:

```
IF ($source=="Zdecydowanie się nie zgadzam", 1,0)
```

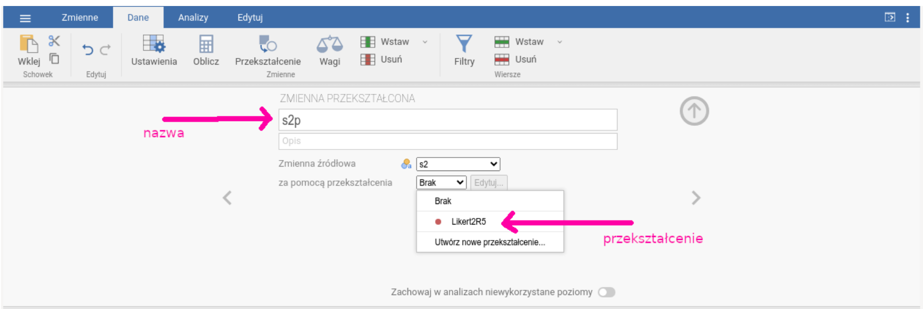
Ponieważ w naszym przykładzie mamy do przekodowania nie dwie a 5 wartości musimy użyć 4 warunków, które są zagnieżdżone jeden w drugim. Można powyższe przepisać, można też skopiować z podręcznika i wkleić do Jamovi (por. rys 6.7).



Rysunek 6.7: Przekształcenie

Naciskamy Enter i gotowe. Zostaje utworzona zmienna `s1p` zawierająca zamiast napisów rangi.

Jeżeli uporaliśmy się z przekodowaniem `s1` ustawiamy kursor na `s2` w okno danych. Naciskamy ikonę `Przekształcenie`. Upewniamy się że zmienną źródłową jest `s2`. Zmieniamy nazwę nowej zmiennej na `s2p`. Klikamy w pole wyboru przekształcenia. Poprzednio były tam tylko dwie pozycje `Brak` oraz `Utwórz nowe przekształcenie` teraz jest trzecia pozycja `Likert2R5` czyli przekształcenie które zdefiniowaliśmy dla zmiennej `s1p`. Wybieramy `Likert2R5` bo zmienną `s2` chcemy przekodować dokładnie w ten sam sposób jak `s1`. Po wybraniu przekształcenia w oknie danych pojawia się nowa zmienna `s2p` (por. rys 6.8).



Rysunek 6.8: Przekształcenie

W podobny łatwy sposób przekodowujemy `s3` oraz `zo1`, `zo2`, `zo3`

Uwaga: polecenie `IF` wpisujemy używając dużych liter. Słowo `$source` wpisujemy tak jak jest to zademonstrowane (`$Source` jest błędem.)

Przekodowanie pytanie z możliwością wielokrotnego wyboru jest równie proste tyle że pisania jest więcej. Zmienna `wiedza_na_temat_palenia` może zawierać do ośmiu następujących napisów oddzielonych średnikami: `Przewlekła obturacyjna choroba płuc`, `Astma oskrzelowa`, `Alergie wziewne`, `Gruźlica` (B), `Zapalenie płuc` (B), `Przewlekłe zapalenie oskrzeli`, `Infekcje dróg oddechowych`, `Palenie nie powoduje chorób układu oddechowego` (B).

Odpowiedzi błędne oznaczono jako (B).

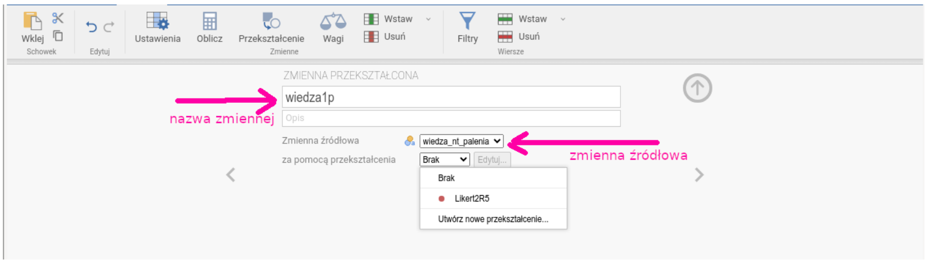
W arkuszu lub oknie danych Jamovi ta zmienna wygląda jakoś tak:

```
...,Przewlekła obturacyjna choroba płuc,...
...,Przewlekła obturacyjna choroba płuc;Astma oskrzelowa;Alergie wziewne;Gruźlica;Zapalenie płu
...,Astma oskrzelowa,
...,Astma oskrzelowa;Gruźlica;Przewlekłe zapalenie oskrzeli,...
...,Przewlekła obturacyjna choroba płuc;Astma oskrzelowa;Infekcje dróg oddechowych,...
```

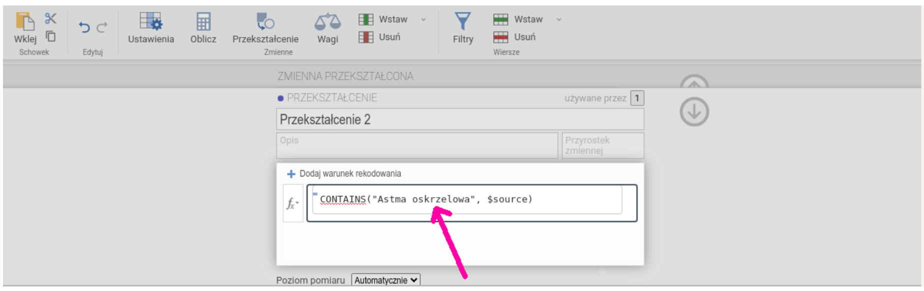
Należy zsumować wystąpienia poprawne i wystąpienia błędne. W tym celu trzeba utworzyć tyle nowych zmiennych ile jest wariantów odpowiedzi, czyli w naszym przykładzie osiem. Każda nowa zmienna jest przekodowywana za pomocą prostej formuły wykorzystującą funkcję `CONTAINS` (zawiera). Przykładowo pierwsza (nazwijmy ją `wiedz1p`) powinna być utworzona w oparciu o następujące przekształcenie:

```
CONTAINS("Przewlekła obturacyjna choroba płuc", $source)
```

Jak to wygląda w oknie programu Jamovi przedstawiono na rysunkach 6.9 oraz 6.10.



Rysunek 6.9: Przekształcenie



Rysunek 6.10: Przekształcenie

Funkcja CONTAINS wstawi 1 jeżeli $source zawiera `Przewlekła obturacyjna choroba płuc`. Oczywiście następna zmienna powinna zawierać `Astma oskrzelowa`:

```
CONTAINS("Astma oskrzelowa", $source)
```

I tak dalej aż do ostatniego wariantu odpowiedzi:

```
CONTAINS('Alergie wziewne', $source)
```

```
CONTAINS('Gruźlica', $source)
CONTAINS('Zapalenie płuc', $source)
CONTAINS('Przewlekłe zapalenie oskrzeli', $source)
CONTAINS('Infekcje dróg oddechowych', $source)
CONTAINS('Palenie nie powoduje chorób układu oddechowego', $source)
```

Każda zmienna `wiedza1`...`wiedza8` zawiera 1 jeżeli ankietowany wskazał dany wariant lub zero jeżeli nie wskazał.

Ostatnia sprawa to przekodowanie liczb na wartości nominalne. Przykładowo chcemy podzielić ankietowanych na grupy stażowe: mały (do pięciu lat), średni (5--15 lat), duży (16 i więcej) staż pracy.

Wartości liczbowe stażu pracy zawiera zmienna `staz`. Aby ją przekodować należy użyć następującego przekształcenia:

```
IF ($source < 5, "M",
  IF ($source < 16, "S", "D"))
```

Poleceń `IF` musi być o jedno mniej niż mamy klas. W naszym przykładzie zatem dwa. Jeżeli `staż` jest mniejszy od 5 wstawiony zostanie napis `M`, jeżeli `staż` jest mniejszy od 16 wstawiony zostanie napis `S` a w przeciwnym wypadku zostanie wstawiony napis `D`.

Gdyby ktoś się niepokoił że 3 spełnia jednocześnie `$source < 5` oraz `$source < 16` to dodamy, że pierwszy się liczy. Przekształcenie kończy działanie po spełnieniu pierwszego warunku i nie wykonuje dalszych porównań. Dlatego liczba 3 zostanie zamieniona na `M` a nie na `S`.

Podobnie przekodowujemy zmienną `wiek`.

6.2.3 Wyliczenie nowych zmiennych

Przekodowanie to była w zasadzie zamiana sposobu mierzenia. **Wyliczenie** to utworzenie nowej zmiennej, zwykle w oparciu o jakąś formułę matematyczną. Na przykład odwrócenie pytanie s1p realizuje `s1pr = 6 - s1p`. Satysfakcja to suma rang z trzech pytań: `satysfakcja = s1pr + s2p + s3p`.

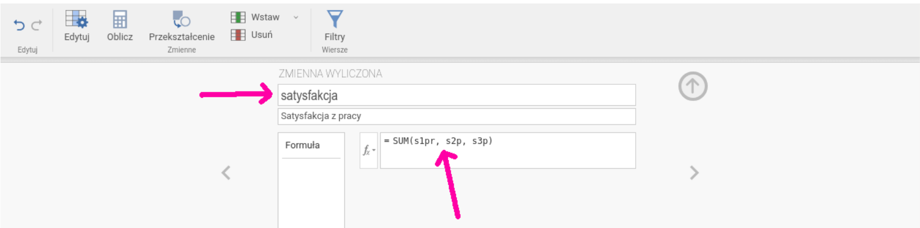
W celu wyliczanie nowych zmiennych należy wybrać Dane Oblicz. Pojawia sie okno zmiennej wyliczonej zatytułowane `ZMIENNA WYLICZONA`

Pierwszy pasek zawiera nazwę zmienną (domyślnie nazwę kolumny w konwencji arkusza kalkulacyjnego, w przykładzie jest to litera `H`) W polu definiowania zmiennej należy wpisać stosowną formułę matematyczną. W przypadku odwracania pytania `s1p` będzie to:

```
6 - s1p
```

W przypadku liczenia łącznej satysfakcji (por. rys. 6.11):

```
SUM(s1pr, s2p, s3p)
```



Rysunek 6.11: Obliczanie nowej zmiennej

Oczywiście wcześniej musimy utworzyliśmy zmienną `s1pr`, inaczej Jamovi zgłosi błąd.

Jeżeli nie chcemy sumy ale np. średnią powinniśmy użyć

```
MEAN(s1pr, s2p, s3p)
```

Inne funkcje matematyczne są dostępne po kliknięciu w pole wyboru znajdujące się po lewej stronie pola definiowania zmiennej.

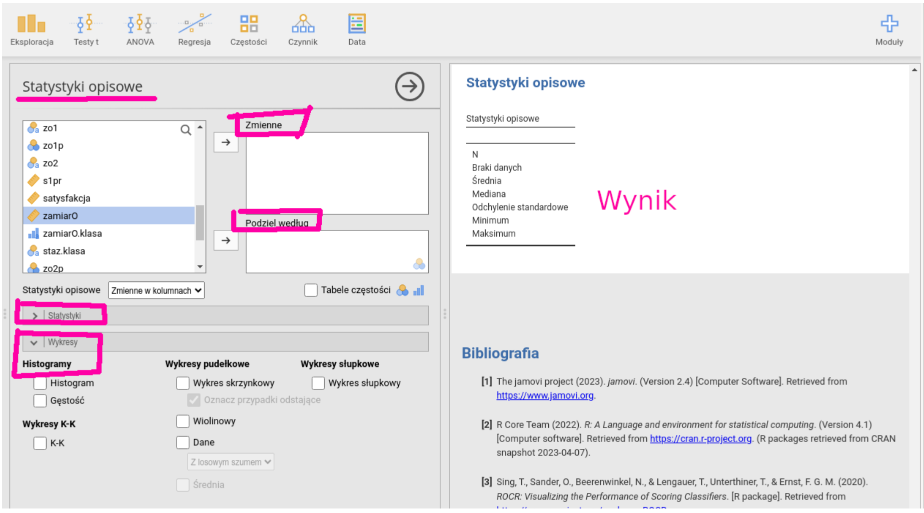
Powiedzieliśmy że miarą wiedzy nt. palenia będzy suma punktów uzyskanych za odpowiedzi prawidłowe minus suma punktów uzyskanych za odpowiedzi nieprawidłowe. Odpowiedzi prawidłowe to `w1p`, `w2p`, `w3p`, `w6` oraz `w7`. Odpowiedzi błędne to `w4p`, `w5p`, `w8p`. Zatem w polu definiowania zmiennej wpisujemy:

```
SUM(w1p, w2p, w3p, w6, w7) - SUM(w4p, w5p, w8p)
```

6.2.4 Analiza struktury

Wybieramy `Analizy`→`Eksploracja`→`Statystyki opisowe`.

W wyświetlonym oknie po lewej deklarujemy co ma być liczone. Wynik pojawi się po prawej (por rysunek 6.12).



Rysunek 6.12: Statystyki opisowe

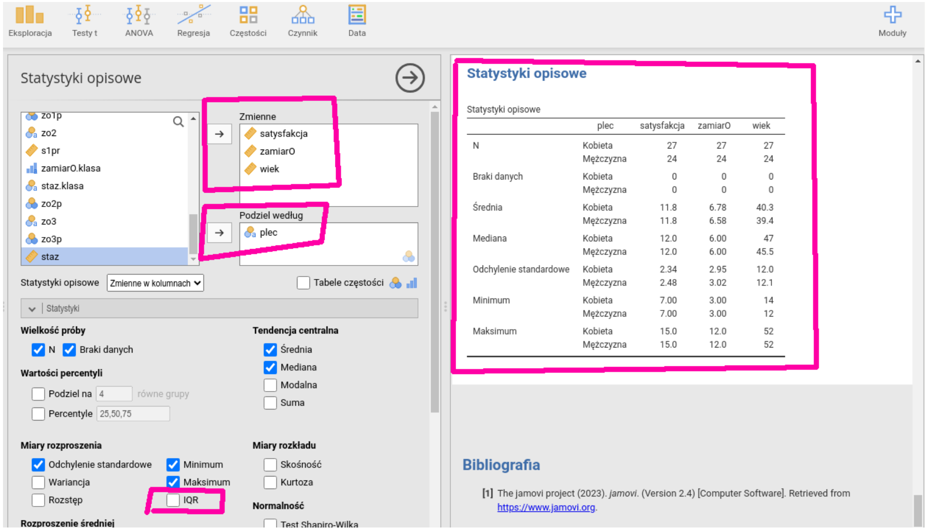
Ustawiamy kursor na zmiennej która nas interesuje i klikamy w strzałkę górną. Jeżeli chcemy podzielić wartości zmiennej na grupy według jakiejś zmiennej nominalnej, to ustawiamy kursor na tej zmiennej nominalnej (na przykład `plec`) i klikamy strzałkę dolną.

Można analizować wiele zmiennych na raz (por. rysunek ref(fig:statystykiSpisoweWynik)). Wystarczy w tym celu ustawić kursor na zmiennej i kliknąć w odpowiednią strzałkę. Zawartość okna wynikowego zostanie automagicznie uaktualniona.

Poniżej okien wyboru zmiennych są zakładki określające precyzyjnie co ma być obliczone oraz jakie wykresy mają zostać wyrysowane. Przykładowo domyślny wydruk nie zawiera rozstępu kwartylowego. Żeby go dodać do wyniku należy w zakładce `Statystyki` zaznaczyć przycisk `IQR`.

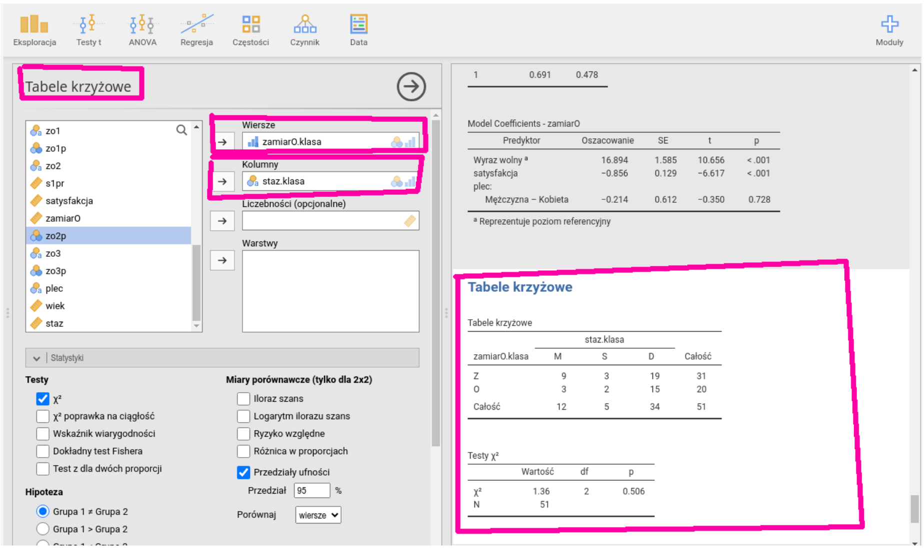
6.2.5 Analiza zależności: zmienne nominalne

Wybieramy `Analizy`→`Częstości`→`Próby niezależne`.



Rysunek 6.13: Statystyki opisowe

Podobnie jak w przypadku analizy struktury jest wyświetlana lista zmiennych oraz okna i strzałki pozwalające wygodnie wybrać to co ma być analizowane. Jest to tak proste że wystarczy przyjrzeć się przykładowemu rysunkowi żeby wiedzieć jak postępować. Przykładową analizę zależności pomiędzy zmiennymi nominalnymi `zamiarOklasa` oraz `staz.klasa` przedstawia rysunek @(fig:tabeleKrzyzowe).

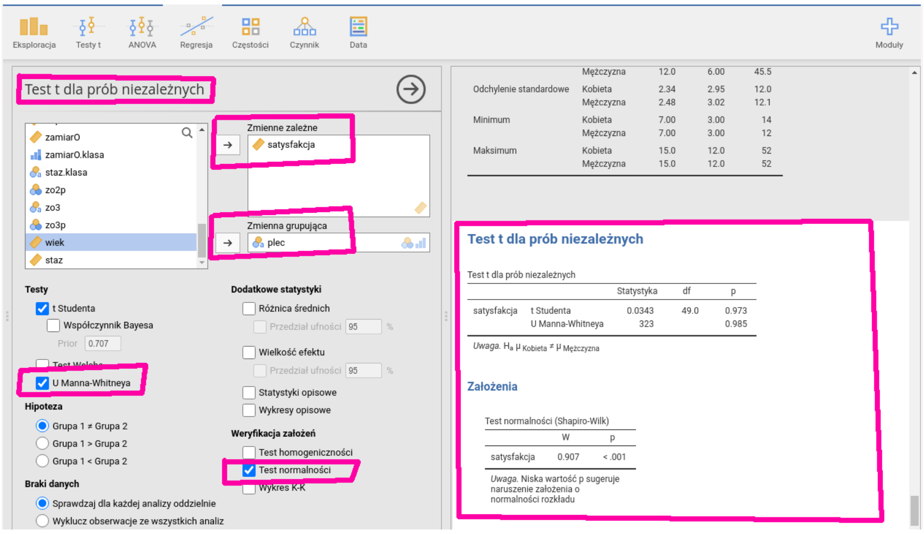


Rysunek 6.14: Tabele krzyżowe

6.2.6 Analiza zależności: zmienna liczbowa/zmienna nominalna

Wybieramy `Analizy→Testy t→Test t dla prób niezależnych` i/lub `Analizy→ANOVA→Jednoczynnikowa ANOVA`.

Zostanie wyświetlony znajomy interfejs. Wybieramy co trzeba. Wynik automagicznie pojawia się w lewym oknie (por. rysunek 6.15).

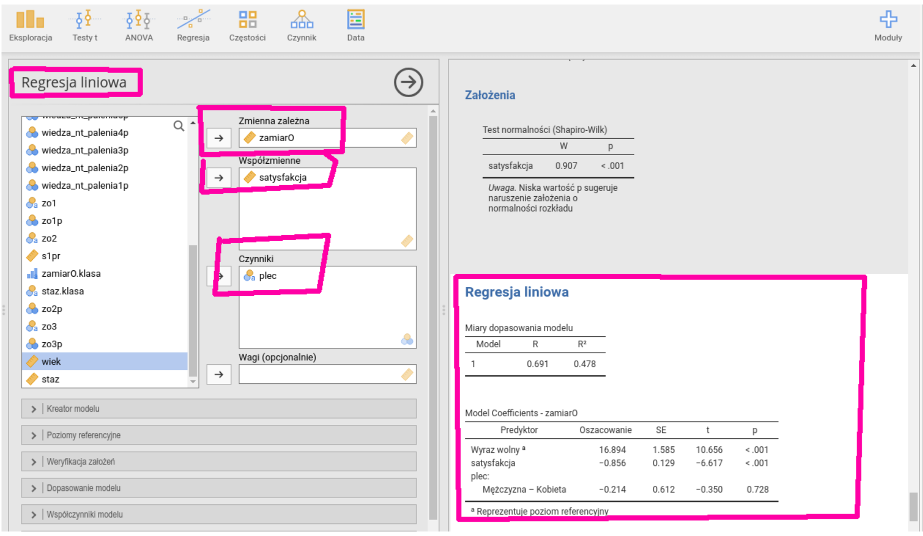


Rysunek 6.15: Test dla prób niezależnych

6.2.7 Analiza zależności: zmienna liczbowa/zmienna liczbowa lub nominalna

Wybieramy `Analizy`→`Regresja`→`Regresja liniowa`

Interfejs podobny do poprzednio opisywanych. Wybieramy zmienną zależną (musi oczywiście być liczbowa) klikając w górną strzłkę. Zmienne niezależne mierzone w skali liczbowej klikając w środkową strzałkę. Zmienne niezależne mierzone w skali nominalnej klikając w dolną strzałkę. Wynik automagicznie pojawia się w lewym oknie (por. rysunek 6.16).



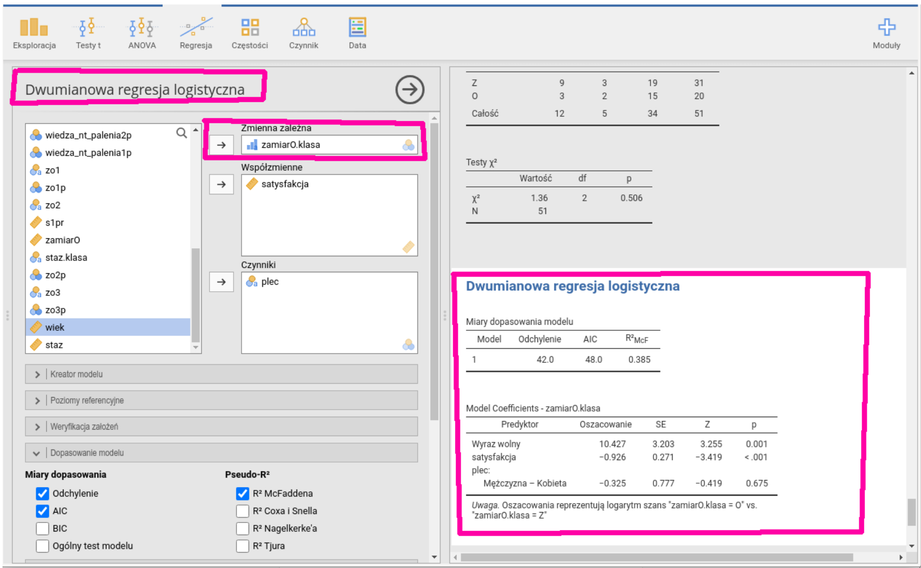
Rysunek 6.16: Regresja liniowa

6.2.8 Regresja logistyczna

Wybieramy `Analizy`→`Regresja`→`Regresja logistyczna`→`Dwie wartości`

Interfejs podobny do analizy regresji. Wybieramy zmienną zależną klikając w górną strzłkę. Zmienna ta **musi** być zmienną dwuwartościową.

Zmienne niezależne mierzone w skali liczbowej wybieramy klikając w środkową strzałkę a zmienne niezależne mierzone w skali nominalnej klikając w dolną strzałkę (por. rysunek 6.17).



Rysunek 6.17: Regresja logistyczna

6.2.9 Redagowanie raportu

Zwykle dobrze jest dodać jakieś dodatkowe objaśnienia do wyników obliczeń wygenerowanych przez program i Jamovi nam to umożliwia. Wybierając `Edytuj` przechodzimy do prostego edytora umożliwiającego redagowanie raportu w oknie wyników a obsługa tego menu jest tak banalnie prosta, że nie wymaga jakiś specjalnych objaśnień.

Rozdział 7

Literatura

Beaglehole, Robert, Ruth Bonita, and Tord Kjellstrom. 1996. *Podstawy Epidemiologii*. Instytut Medycyny Pracy.

Bland, Martin. 2015. *An Introduction to Medical Statistics*. Oxford University Press.

Machin, David, Michael J Campbell, and Stephen J Walters. 2007. *Medical Statistics*. John Wiley.