dr Izabela Miechowicz

Katedra i Zakład Informatyki i Statystyki

Zalecenia dotyczące badania zgodności i rzetelności skali

data przygotowania: 2024-05-31

Badanie rzetelności i zgodności kwestionariusza badawczego stosujemy, gdy:

- ✓ Istnieje kwestionariusz badawczy, który został zwalidowany, ale na innej populacji niż ta, którą chcemy przebadać,
- ✓ Istnieje już kwestionariusz badawczy, ale jest on długi, zawiera dużo pytań, co może powodować zmęczenie, zwłaszcza starszych lub mocno schorowanych respondentów, a chcielibyśmy zbudować narzędzie prostsze, ale równie dobrze badające dane zjawisko jak tradycyjne narzędzie badawcze,
- ✓ Nie ma narzędzia badawczego, które mogłoby zostać zastosowane do badania interesującego nas zjawiska i chcemy takie zbudować.

Podstawowe pojęcia?

Trafność

to stopień zgodności, z jaką narzędzie pomiarowe mierzy to, do mierzenia czego zostało skonstruowane. Odpowiada zatem na pytanie: czy przy użyciu tego narzędzia udało nam się zmierzyć to, co planowaliśmy zmierzyć?

Typy trafności:

- ✓ **Trafność treściowa** – odnosi się do zgodności treści pozycji testowych z definicją mierzonej cechy (pytania powinny dotyczyć tylko badanego zakresu)

Narzędzie będzie trafne treściowo, gdy:

1. Wszystkie pozycje narzędzia należą do zdefiniowanego uniwersum (opisuje cały zakres dziedziny, której narzędzie ma dotyczyć)
2. Narzędzie proporcjonalnie reprezentuje całe uniwersum

Trafność treściową można zmierzyć stosując współczynnik trafności treściowej Lawshego (CVR).

$$CVR = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

n_e - liczba sędziów (ekspertów), którzy uznali daną pozycje skali jako istotną (przyznali jej 2 punkty)

N – liczba wszystkich sędziów

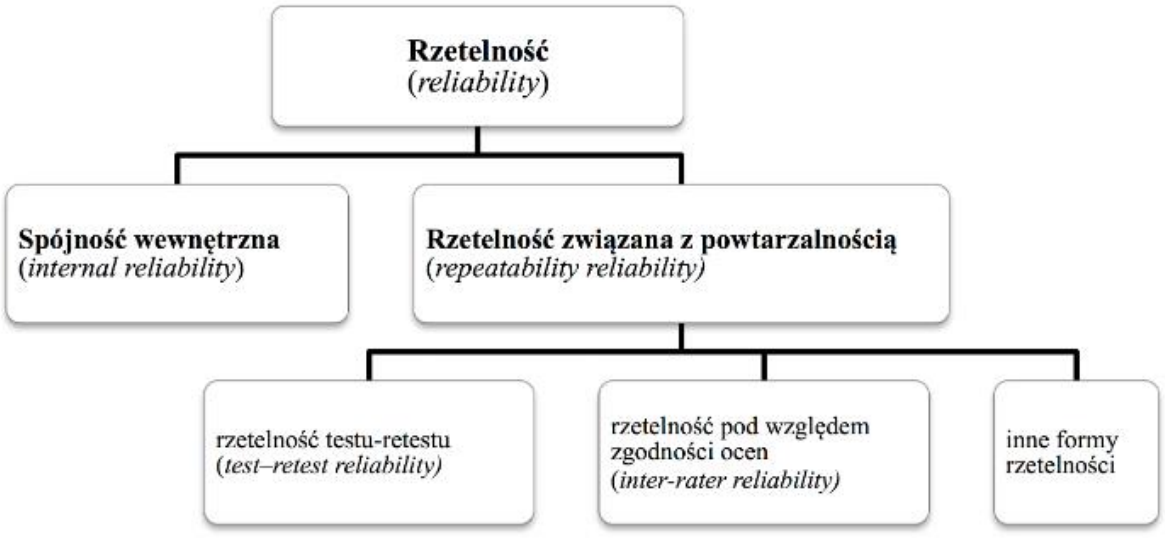
- ✓ **Trafność empiryczna** – to ocena prognostyczna narzędzia pomiarowego. Jeśli narzędzie pomiarowe jest trafne, to jeśli w rzeczywistości występuje związek pomiędzy mierzonym zmiennymi, to pomiędzy wynikami skali również taki związek powinniśmy wykazać. Trafność empiryczną możemy ocenić dokonując porównania wyników danego pomiaru z wynikami innych uznanych narzędzi pomiarowych. W praktyce do oszacowania siły związku pomiędzy wynikami narzędzia a zewnętrznym kryterium stosuje się współczynnik korelacji.

- ✓ **Trafność teoretyczna** – sprawdzenie, czy narzędzie pomiarowe jest powiązane z pojęciami i teoretycznymi założeniami obowiązującymi w analizowanej dziedzinie.

Rzetelność

określa stopień, w jakim stopniu wynik testu oddaje rzeczywistą wartość badanej cechy. Innymi słowy, jest to miara tego, jak bardzo **skala jest wolna od błędów**.

W sensie badawczym rzetelność oznacza: **zgodność, niesprzeczność i powtarzalność**.



https://przeglad.ump.edu.pl/uploads/2016/4/415_4_49_2016.pdf

Rzetelność związana ze spójnością wewnętrzną

W ramach badania rzetelności można oszacować **spójność wewnętrzną**. Służą do tego m. in. metoda połówkowa i zbadanie współczynnika alfa Cronbacha.

Aby wykorzystać **metodę połówkową** należy podzielić kwestionariusz na 2 części. Najczęściej dzieli się go na pytania parzyste i nieparzyste. Następnie oblicza się współczynnik korelacji Pearsona pomiędzy obiema połówkami kwestionariusza. Następnie oblicza się współczynnik r ze wzoru

$$r = \frac{2r_p}{1 + r_p}$$

Aby narzędzie uznać za spójne r powinien być większy od 0,6.

Współczynnik alfa Cronbacha – mierzy stosunek wariancji poszczególnych pozycji do wariancji całej skali. Przyjmuje się, że aby narzędzie uznać za spójne wewnętrznie alfa Cronbacha powinna osiągnąć wartość co najmniej 0,6.

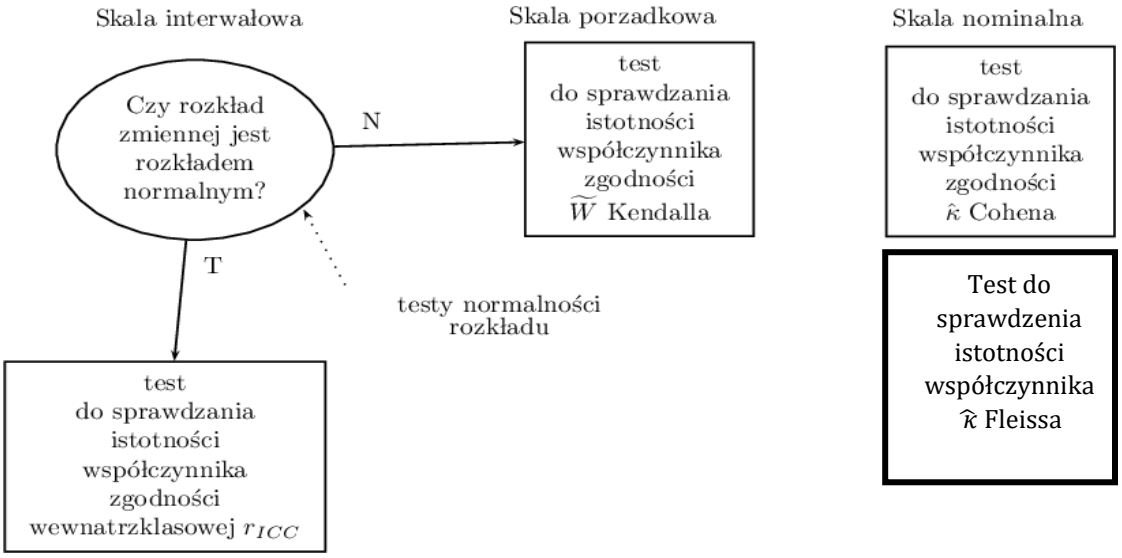
Rzetelność związana z powtarzalnością:

- ✓ **Test-retest** – polega na ponownym wykonaniu badania po pewnym czasie i zbadaniu czy otrzymujemy podobne wyniki. Problemem tutaj może byc czas pomiędzy badaniami – jeśli będzie zbyt krótki, wówczas musimy liczyć się z efektem uczenia (pamiętamy, jak oceniliśmy dany obiekt), jeśli będzie zbyt długi – w czasie pomiędzy badaniami może nastąpić istotna zmiana stanu ocenianego obiektu (np. pogorszenie stanu zdrowia). W tym celu wykonujemy testy służące do oceny zmian (t-Studenta dla prób powiązanych lub Wilcoxona) oraz badamy współczynnik korelacji pomiędzy pomiaramy (wówczas możemy sprawdzić czy zachowujemy tę samą tendencję w obu badaniach) – wysoka powtarzalność świadczy o dużej stabilności narzędzia

- ✓ **Rzetelność pod względem zgodności ocen –** zgodność obliczamy dla całej grupy sędziów, polega ona na ocenie zgodności ocen tego samego materiału przez grupę sędziów

Jaki współczynnik wybrać?

Wybór właściwego współczynnika zależy od tego na jakiej skali zmierzona jest cecha.



http://manuals.pqstat.pl/statpqpl:zgodnpl

- ✓ **<u>Współczynnik ICC</u>** obliczamy wtedy, gdy pomiar zmiennej odbywa się na skali interwałowej. Dodatkowo należy sprawdzić czy zmienne mają rozkład zgodny z normalnym. Dopuszcza się, by zamiast sprawdzenia normalności rozkładu każdej zmiennej zbadać czy róznice pomiędzy pomiarami mają rozkład zgodny z normalnym. ICC należy do przedziału (-1; 1), przy czym ujemne wartości tego współczynnika są traktowane tak samo, jak 0, czyli wskazuja na brak zgodności ocen sędziowskich.

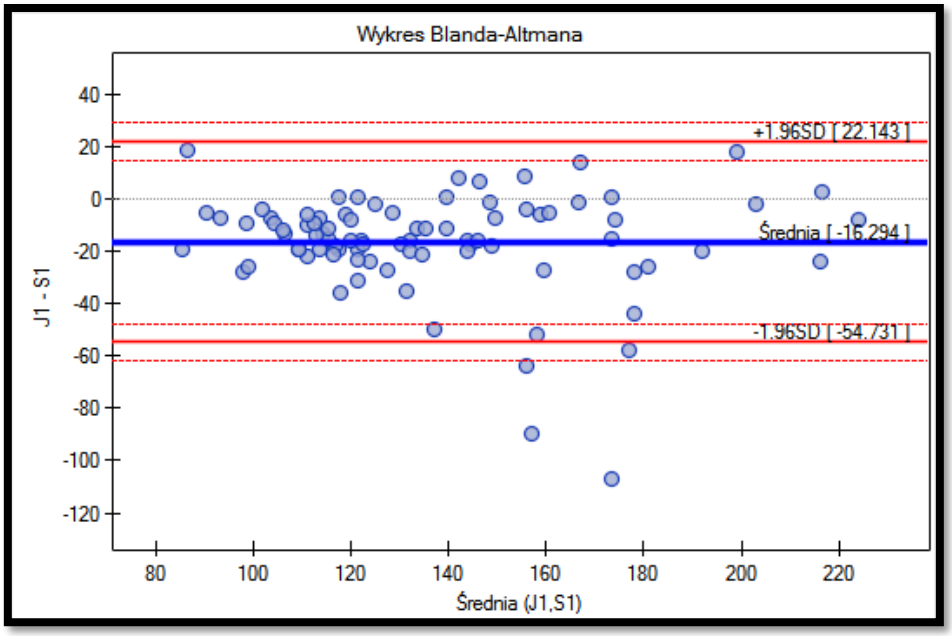
W celu przeprowadzenia tej analizy należy jeszcze określić jaki został zastosowany model:

- Czy wszystkie obiekty sa oceniame przez tych samych sędziów,
- Czy każdy z obiektów jest ocenainy przez innego sędziego.

Dodatkowo należy zdecydować co chcemy oszacować. Mamy dwie możliwości:

- Bezwzględna zgodność – czy oceny wystawione przez sędziów są identyczne (dzięki temu można okreslić dla każdego obiektu dokładną wartość jego wyniku)
- Bezwzględna spójność – sędziowie mogą wydawać różne wyniki (w różnym zakresie), ale poza pewnym przesunięciem ich oceny się nie różnią

Dla zmiennych zmierzonych na skali interwałowej można dodatkowo wykonać punktowy wykres Blanda-Altmana.



Co znajduje się na poszczególnych osiach?

- **Oś X** - średnia pomiarów dla porównywanych metod;
- **Oś Y** - różnica pomiędzy pomiarami dla porównywanych metod;
- **Średnia różnic** - jeśli wyniki uzyskane nową metodą są stale większe/mniejsze niż metodą starą, wówczas występuje przesunięcie, które nazywamy ang. bias, czyli linia obrazująca średnią różnic nie znajduje się na poziomie 0, ale jest przesunięta znacząco w górę lub w dół od tego poziomu (na powyższym wykresie jest to niebieska linia)
- **95% przedział zgodności** - jeśli różnice mają rozkład normalny, 95% różnic znajdzie się w przedziale (Średnia różnic ±1.96SD), gdzie SD, to odchylenie standardowe różnic.

Jak interpretować wartości współczynników?

Nie ma jasnych kryteriów, które określają kiedy możemy uznać zgodność za niewystarczającą, a kiedy za akceptowalną. Intuicyjnie czujemy, że im wartości są bliżej 1 tym zgodność ocen jest większa. Zaprezentowana tabela jest próbą ujednolicenia nazewnictwa i narzucenia pewnych zakrsów, natomiast nie należy jej traktować jako standard.

współczynnik	ocena zgodności
>0,80	bardzo dobra
0,60 – 0,80	dobra
0,40 – 0,59	zadowalająca
<0,40	niewystarczająca

https://media.statsoft.pl/pdf/czytelnia/wykorzystywanie_procedury_sedziow_kompetentnych.pdf

Jak przygotować dane?

Dane przygotuj w kolumnach

Ciśnieniomierz klasyczny	Ciśnieniomierz półautomatyczny
100	122
108	121
76	95
…	…

Zbadaj normalność obu zmiennych lub ich różnic (np. testem Shapiro-Wilka).

Wybierz odpowiedni model badania oraz to, co chcesz oszacować (bezwzględną zgodność czy spójność).

Jak interpretować wyniki?

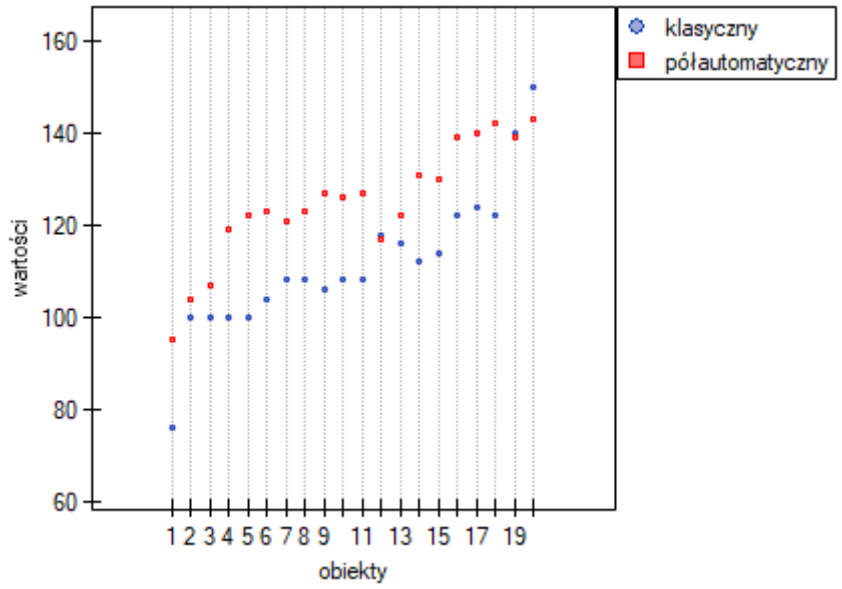
Pamiętaj, że **wartość α** najczęściej wynosi 0.05 i jest to poziom błędu którego nie chcemy przekroczyć prowadząc badanie.

Przykładowy wynik:

ICC(2,k) dla średniej k sędziów	0,901
-95% CI dla ICC(2,k)	0,751
+95% CI dla ICC(2,k)	0,961
Wartość p	<0,001

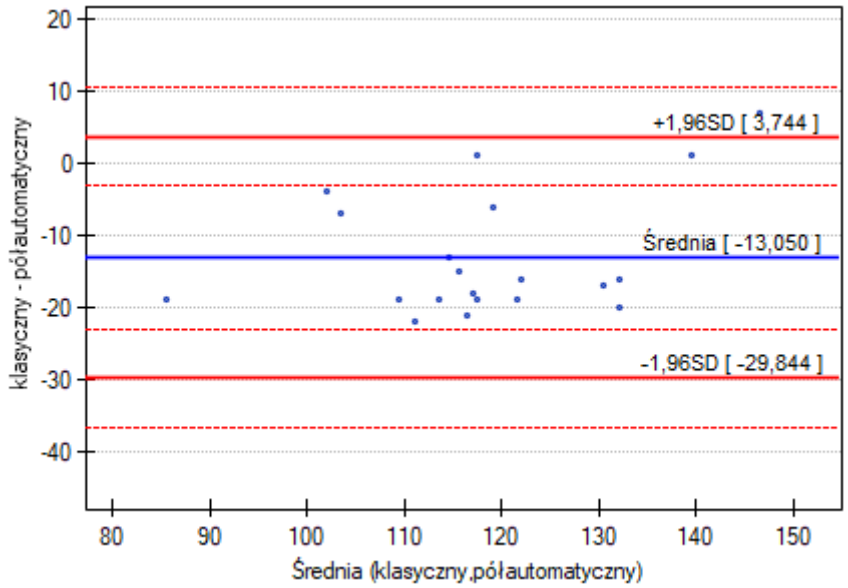
Wyniki otrzymane z pomiarów dwóch ciśnieniomierzy są zgodne ($p<0,05$). Współczynnik ICC=0,901, co świaczy o bardzo dobrej zgodności. 95% CI pokazuje, że dla populacji z prawdopodobieństwem 95% współczynnik ten będzie znajdował się w przedziale (0,751; 0,961).

Dodatkowo, w interpretacji może pomóc wykres punktowy pokazujący rozrzut wyników, po wczesniejszym uporządkowaniu pomiarów wg średniej:



Z wykresu można odczytać, że oba ciśnieniomierze mierzą podobnie (linie utworzone przez punkty sa prawie równiegłe), natomiast punkty czerwone są położone wyżej niż niebieskie, co świadczy o wyższych wynikach uzyskanych z ciśnieniomierza półautomatycznego.

Wykres Blanda-Altmana



Wykres punktowy w całości mieści się w przedziale średnia ±1,96SD (z wyjatkiem tylko jednego punktu). Niebieska linia obrazująca srednią z róznic pomiarów jest przesunięta w dół o 13,05 jednostki, co oznacza, że cisnieniomierz półautomatyczny średnio podaje wyższe wyniki o taką wartość.

- ✓ **<u>Współczynnik W Kendalla</u>** obliczamy wtedy, gdy pomiar zmiennej odbywa się na skali porządkowej. W Kendalla należy do przedziału (0; 1).

Jak przygotować dane?

W sytuacji, gdy chcemy zbadać zgodność z wykorzystaniem współczynnika W Kendalla większość pakietów statystycznych wymaga poprzecznego układu danych.

Dane przygotuj w tabeli:

	1 pomiar	**2 pomiar**	**3 pomiar**	**…**
metoda nowa I	1,76	1,71	1,84	…
metoda nowa II	1,87	2,19	2,11	…

Uwaga!

Niektóre pakiety statystyczne dają możliwośc wykorzystania danych przygotowanych w tradycyjny sposób (wzdłużny).

Jak interpretować wyniki?

<u>Przykładowy wynik:</u>

współczynnik zgodności Kendalla	0,993
średni współczynnik korelacji Spearmana	0,986
wartość p	<0,001

Wyniki otrzymane z pomiarów dwóch zastosowanych metod są zgodne ($p<0,05$). Współczynnik W Kendalla=0,901, co świaczy o bardzo dobrej zgodności. Średni współczynnik korelacji rangowej R Spearmana jest bardzo wysoki i wynosi 0,986, co świadczy o bardzo silnej dodatniej zależności pomiędzy pomiarami otrzymanymi w obydwu metodach.

- ✓ **<u>Współczynnik kappa Cohena (</u>**$\hat{\kappa}$ Cohena) obliczamy wtedy, gdy pomiar zmiennej odbywa się na skali nominalnej (ewentualnie porządkowej w kategoriach). Kappa Cohena należy do przedziału (-1; 1), przy czym ujemne wartości tego współczynnika są traktowane tak samo, jak 0, czyli wskazuja na brak zgodności ocen sędziowskich.

Jak przygotować dane?

Dane przygotuj w kolumnach

Lekarz 1	Lekarz 2
zapalenie płuc	zapalenie płuc
inne	zapalenie płuc
zapalenie oskrzeli	zapalenie oskrzeli
zapalenie oskrzeli	inne
…	…

Jeśli dane mają charakter nominalny, wybierając współczynnik kappa Cohena nie ustalamy wag. W przypadku, gdy dane są porządkowe (kategorie można uporządkować) możesz wybrać wagi (liniowe lub kwadratowe), by podkreślić, że istnieje relacja porządkująca pomiędzy poszczególnymi kategoriami.

Jak interpretować wyniki?

Przykładowy wynik:

Na podstawie zebranych danych tworzymy tabelę, w której przedstawiamy jakie diagnozy postawili poszczególni lekarze:

	Lekarz 2		
Lekarz 1	zapalenie płuc	zapalenie oskrzeli	inne
zapalenie płuc	28,182%	3,636%	3,636%
zapalenie oskrzeli	7,273%	35,455%	8,182%
inne	4,545%	6,364%	2,727%

Suma procentów na głównej przekątnej pokazuje % zgodnych diagnoz w próbie (66,364%). Otrzymany współczynnik kappa Cohena ma wartość niższą niż ten odsetek, ponieważ jest pomiejszony o przypadkową zgodność.

Współczynnik Kappa	0,446
-95% CI dla współczynnika Kappa	0,312
+95% CI dla współczynnika Kappa	0,579
Wartość p (asymptotyczne)	<0,001

Wyniki otrzymane z ocen dokonanych przez dwóch lekarzy są zgodne ($p<0,05$). Współczynnik kappa Cohena=0,449, co świaczy o zadowalającej zgodności.