Spotkanie Entuzjastów R: 24.04.2014

Psychometria w R: ukryte cechy, binarne wyniki, możliwości i ograniczenia

Tomasz Żółtak

Instytut Badań Edukacyjnych

Plan prezentacji

1. **Czym jest psychometria?**
    1. Modele psychometryczne – szybki przegląd i podstawowe założenia
    2. Bardzo powierzchownie o estymacji
2. **Problemy z dychotomicznymi (i porządkowymi) zmiennymi obserwowanymi**
    1. Rozkład cechy a rozkład sumy punktów
    2. Skala wyników
    3. Co mierzy zadanie?
3. **Psychometria w R**
    1. Szybki przegląd pakietów
    2. Wyskalujmy wzrost!

Czym jest psychometria?

Etymologicznie

- Psychometrię powstała w ramach psychologii...
- … w związku z pomiarem cech psychologicznych,
  - np. inteligencji (ale i mnóstwa innych).
- Jej elementy przyjęły się również w edukacji na potrzeby pomiaru umiejętności...
- … i w kontekście pomiaru umiejętności jest ona obecnie chyba najbardziej (co nie znaczy, że szeroko) rozpoznawana.
  - PISA, TIMSS, PIRLS, PIAAC.
- To, jak do pomiaru umiejętności podchodzą psychometrycy ma jednak niewiele wspólnego z tym, co na ten temat myślą *zwykli ludzie*.

Czym jest psychometria?

Współcześnie można powiedzieć, że jest to teoria pomiaru zakładająca, że:

- Mierzone cechy nie są obserwowalne bezpośrednio (są to cechy ukryte), lecz jedynie za pośrednictwem przejawów, które pozostają z nimi w zależnościach statystycznych.
  - Postać zależności musimy założyć w modelu, a parametry ją opisujące zwykle estymujemy na podstawie danych.
- Zmienne opisujące przejawy badanych cech nie są ciągłe, lecz mają charakter porządkowy (i to o niewielkiej liczbie różnych przyjmowanych wartości), w szczególności mogą to być zmienne dychotomiczne (0-1).

Czym jest psychometria? cd.

Różne modele psychometryczne – ogólna klasyfikacja:

<table>
<tr><th rowspan="2">zmienne ukryte</th><th colspan="2">zmienne opisujące przejawy</th><th rowspan="2">uwagi</th></tr>
<tr><th>dychotomiczne</th><th>porządkowe</th></tr>
<tr><td rowspan="2">ciągłe
(określone na ℝ;
na potrzeby estymacji bardzo często przyjmuje się, że posiadają one rozkład normalny)</td><td>Rasch, OPLM, 2PL, 3PL, 4PL i ich wersje probitowe</td><td rowspan="2">(S)GRM, RGRM, GPCM, RPCM, RSM i inne oraz ich wersje probitowe</td><td rowspan="2">jednowymiarowe lub wielowymiarowe</td></tr>
<tr><td>nieparametryczne modele IRT (Mokken)</td></tr>
<tr><td>porządkowe
(o niewielkiej liczbie przyjmowanych wartości)</td><td colspan="2" rowspan="2">Cognitive Diagnostic Models</td><td rowspan="2">co do zasady wielowymiarowe</td></tr>
<tr><td>dychotomiczne</td></tr>
</table>

Czym jest psychometria? cd.

Różne modele psychometryczne – ogólna klasyfikacja:

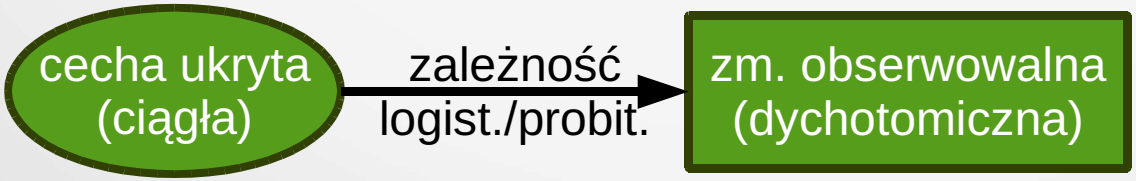
zmienne ukryte	zmienne opisujące przejawy		uwagi
	dychotomiczne	porządkowe	
ciągłe (określone na $\mathbb{R}$; na potrzeby estymacji bardzo często przyjmuje się, że posiadają one rozkład normalny)	**Rasch**, O… **2PL**, 3PL, 4… i ich wersje probitowe	RSM i inne oraz ich wersje probitowe	…we wielowymiarowe
	nieparametryczne modele IRT (Mokken)		
porządkowe (o niewielkiej liczbie przyjmowanych wartości)	*Cognitive Diagnostic Models*		co do zasady wielowymiarowe
dychotomiczne			

o tych modelach (jednowymiarowych) będziemy mówić dzisiaj

Dwie tradycje

- **Konfirmacyjna analiza czynnikowa** (CFA) i SEM, estymowane z macierzy korelacji między zmiennymi.
  - Dostosowanie założeń modelu do porządkowego/ dychotomicznego charakteru zmiennych – estymacja z macierzy korelacji polichorycznych.

- **Item Response Theory** – od początku konstruowana z myślą o tym, że przejawy zmiennych ukrytych mierzone są na skalach porządkowych Estymacja z pełnej macierzy danych.



Oba podejścia bardzo wiele łączy, a pod pewnymi warunkami są wręcz formalnie równoważne!

Dwie tradycje

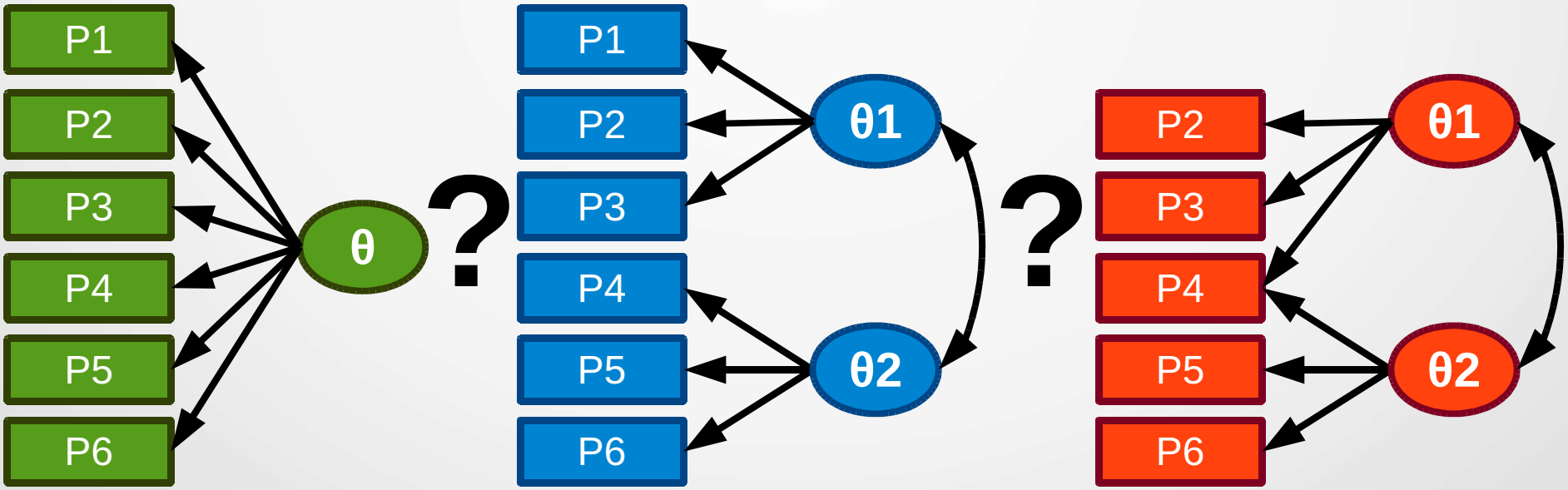
- **Konfirmacyjna analiza czynnikowa / SEM estymowane z macierzy korelacji:**
  - Podejście mniej złożone obliczeniowo.
  - Cała masa indeksów pozwalających oceniać jakość dopasowania modelu do danych.
  - Równoważna założeniu o probitowej funkcji łączącej.
  - Nie pozwala uwzględnić modeli 3PL i 4PL.
  - Nie da się zastosować do typowych schematów badawczych z planowymi brakami danych (musi dać się wyliczyć korelację między każdą parą zmiennych w modelu).

Dwie tradycje

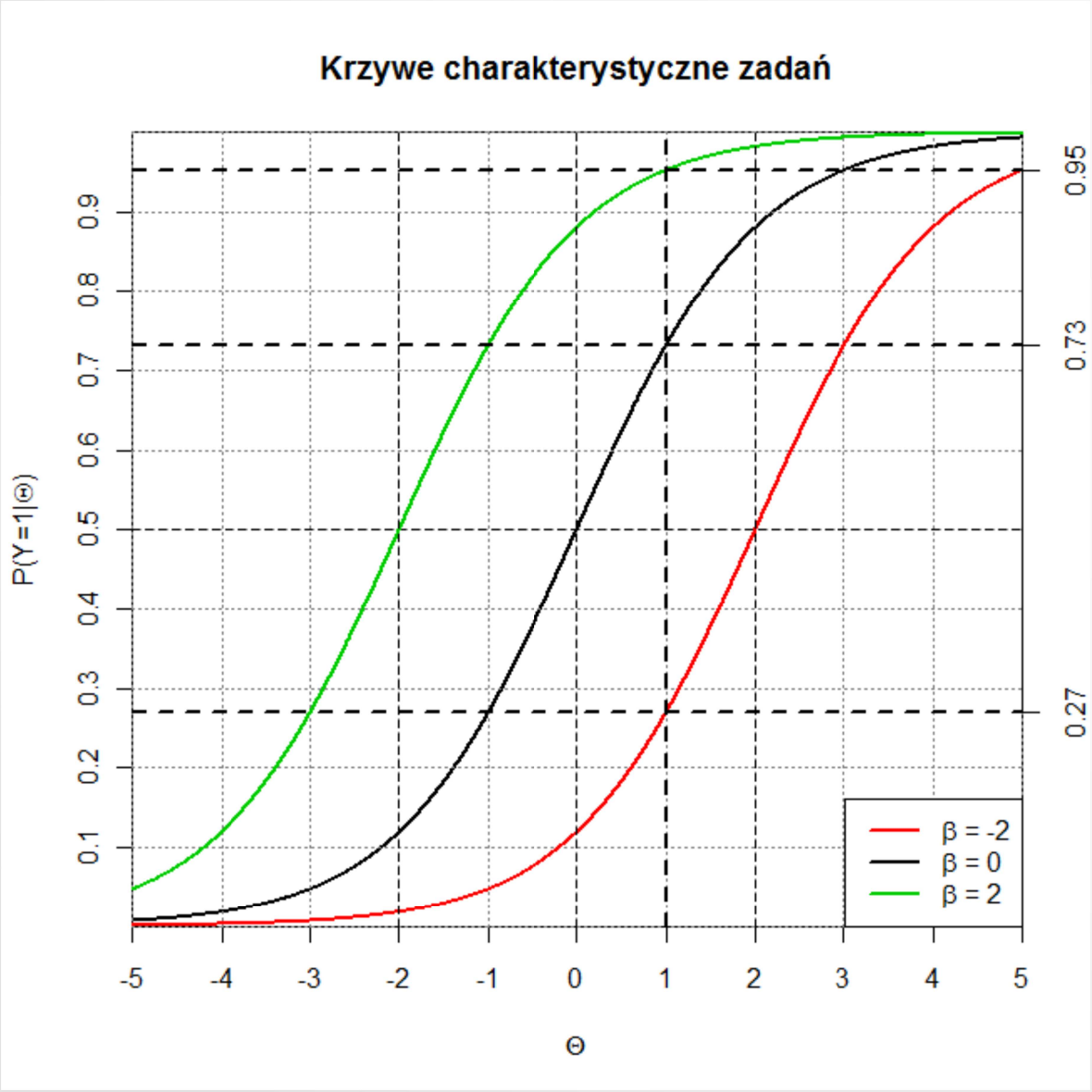
- **Item Response Theory:**
  - Większa swoboda wyboru modelu.
  - Da się zastosować do schematów badawczych z planowymi brakami danych (oczywiście musi występować pewna pula pytań wspólnych).
  - Ładnie ilustruje się wykresami.
  - Złożona obliczeniowo, zwłaszcza dla modeli wielowymiarowych (całkowanie numeryczne).
  - Problemy z oceną jakości dopasowania modelu do danych (najlepiej symulacyjnie, ale jest to możliwe tylko dla relatywnie prostych modeli).

Model psychometryczny - hipoteza

- Czy są podstawy by twierdzić, że dany zestaw pytań mierzy (w pewnym sensie) tę samą cechę?
- Jak mógłby wyglądać model, który lepiej opisywałbym (potencjalne) przyczyny obserwowanych zależności?



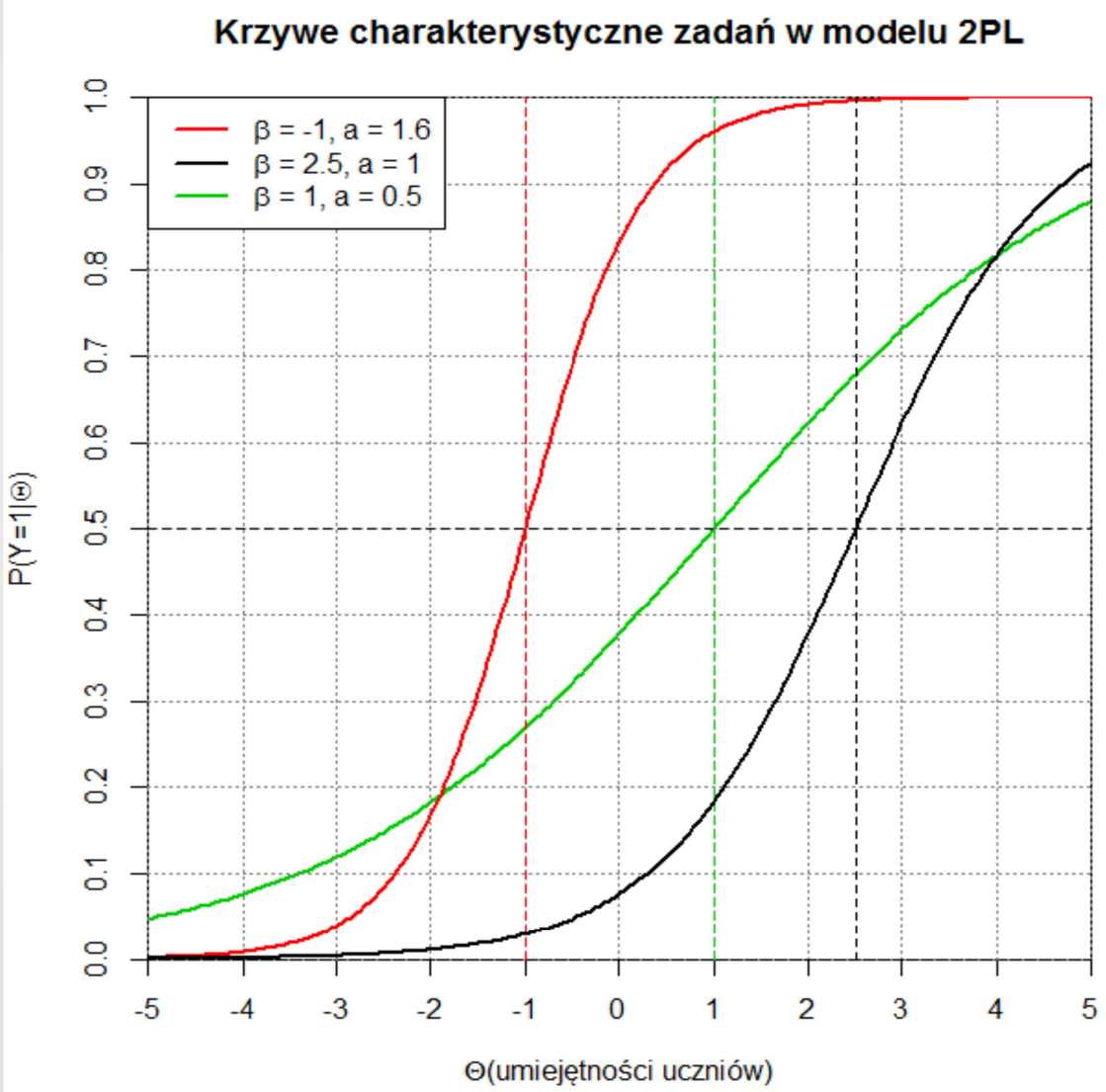
Model Rascha



$$P(X_i = 1|\Theta) = \frac{\exp(\Theta - b_i)}{1 + \exp(\Theta - b_i)}$$

- Prawdopodobieństwa udzielenia poprawnej odpowiedzi przez ucznia o poziomie umiejętności Θ=1 na pytania o **trudności**:
  - β_i=-2 jest równe 0,95;
  - β_i= 0 jest równe 0,73;
  - β_i= 2 jest równe 0,27.
- **Trudność** pytania to poziom umiejętności, dla którego prawd. poprawnej odpowiedzi jest równe 0,5

Model 2PL



$$P(X_i = 1|\Theta) = \frac{\exp[a_i(\Theta - b_i)]}{1 + \exp[a_i(\Theta - b_i)]}$$

- Wartość parametru **dyskryminacji** (a_i) wpływa na nachylenie krzywej charakterystycznej zadania.
- Im wyższa dyskryminacja, tym bardziej odpowiedź na dane pytanie związana z mierzoną cechą.
- **Trudność** pytania (β_i) przesuwa krzywą charakterystyczną w poziomie.
- Krzywe mogą się przecinać.

Przewidywanie poziomu cechy

- Poziom umiejętności przewidywany dla ucznia na podstawie punktacji, jaką uzyskał on z testu zależy od własności pomiarowych (jakości) zadań, które uczeń rozwiązał poprawnie:
  - Liczby zadań, które rozwiązał poprawnie – w modelu Rascha.
  - Parametrów dyskryminacji zadań, które rozwiązał poprawnie – w modelu 2PL.
- Przewidywany poziom umiejętności nie zależy od trudności zadań, które uczeń rozwiązał poprawnie (jeśli tylko wszyscy zdający rozwiązywali ten sam zestaw zadań).

Metody estymacji

Tradycja CFA (estymacja z macierzy korelacji):

1. Wyestymuj macierz korelacji polichorycznych.
2. Na jej podstawie wyestymuj parametry modelu.
   - Preferowana metoda Weighted Least Squares – musimy brać pod uwagę ten problem, że wariancja zmiennych obserwowalnych jest powiązana z ich średnią.
3. Ewentualnie wylicz oszacowania wartości cech ukrytych.
   - Wiele możliwych metod.

Metody estymacji

Tradycja IRT (estymacja z pełnej macierzy danych):

Modele Rascha i OPLM:

- Wiele metod: Joint ML, Conditional ML (obie nie nakładają założeń na rozkład badanej cechy), Marginal ML, metody bayesowskie.

Bardziej złożone modele:

- Marginal ML – zakładamy rozkład badanej cechy w populacji, z której pochodzi badana grupa; metody bayesowskie.

1. Wyestymuj parametry modelu.
2. Ewentualnie wylicz oszacowania wartości cech ukrytych.
   - Wiele możliwych metod.

Problemy z dychotomicznymi (i porządkowymi) zmiennymi obserwowanymi

Problemy: rozkład sumy punktów

- Jeśli odpowiedzi na pytania są przejawami tej samej cechy ukrytej, to najprostszym wskaźnikiem natężenia cechy może być suma punktów przypisanych do udzielonej odpowiedzi.
- Przyzwyczajenie wyniesione z modeli *stricte* liniowych każe nam oczekiwać, że rozkład takiej sumy powinien być (przy dużej liczbie badanych) zbliżony do rozkładu cechy ukrytej...
- … ale gdy cechę ukrytą ze zmiennymi obserwowalnymi łączy zależność logistyczna/probitowa będzie tak tylko pod warunkiem, że dobrze dobraliśmy trudność zadań.

Problemy: rozkład sumy punktów

Mała symulacja:

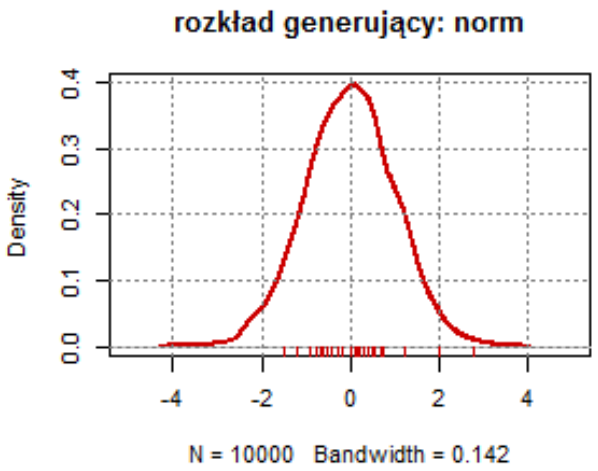
- Wygenerujmy trzy duże grupy *badanych*, losując im wartości cechy spod trzech różnych rozkładów: normalnego, log-normalnego i jednostajnego (a następnie wystandaryzujmy wartości cechy w każdej grupie do średniej 0 i odch. stand. 1).
- Dajmy im do *rozwiązania* pięć różnych *testów*, różniących się trudnością pytań. Dla uproszczenia załóżmy, że *odpowiadają* na nie zgodnie z założeniami modelu Rascha. Każdy *test* składa się z 30 pytań o trudnościach wylosowanych z:

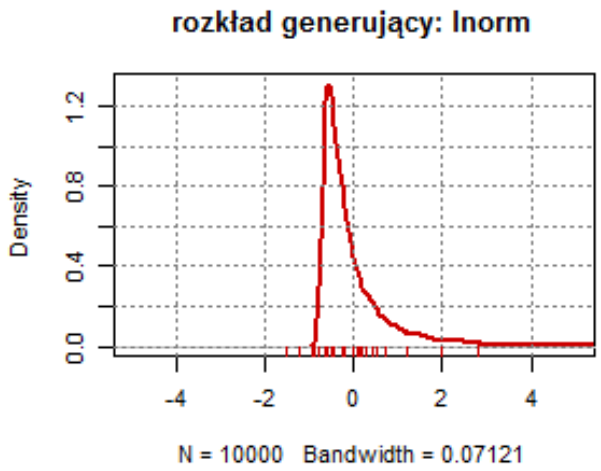
  1) N(0, 1) 2) mieszaniny 1:1 N(-1.5, 0.5) i N(1.5, 0.5),

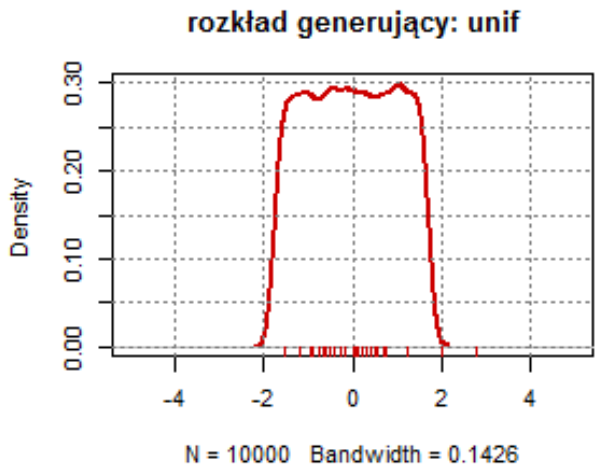
  3) N(0, 0.25) 4) N(1.5, 0.5) 5) N(-1.5, 0.5)
- Sprawdźmy, jak będą wyglądały rozkłady sum punktów.

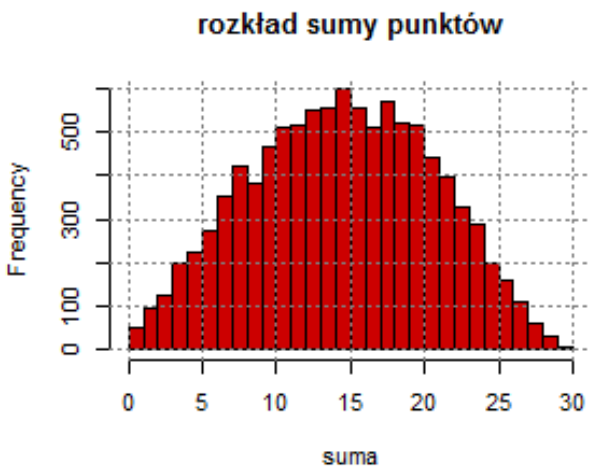
Dosyć dobrze dobrane trudności

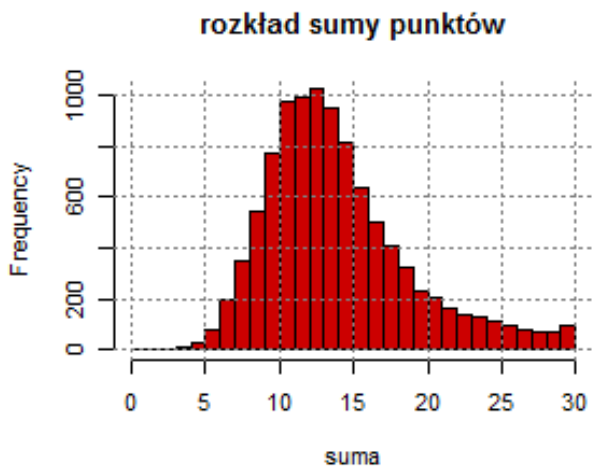
kształty rozkł. sumy punktów zbliżone do rozkł. generujących

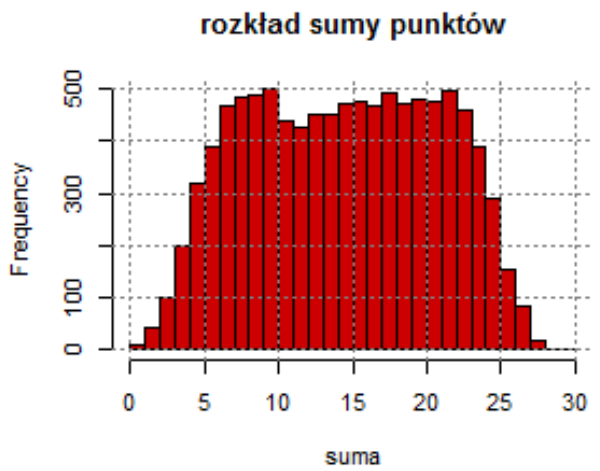






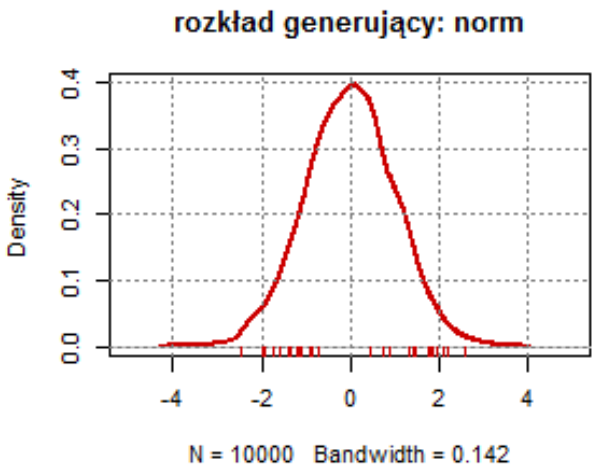


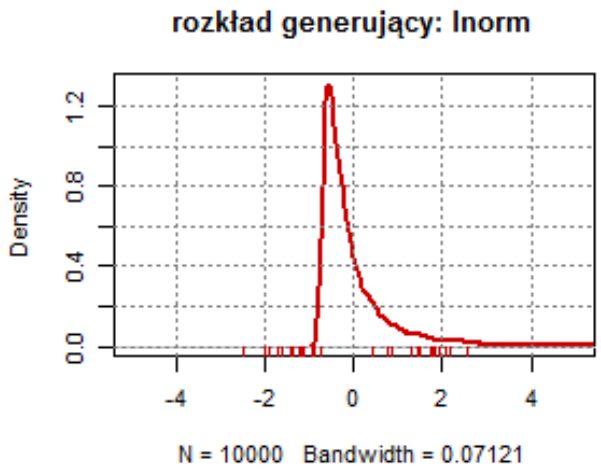


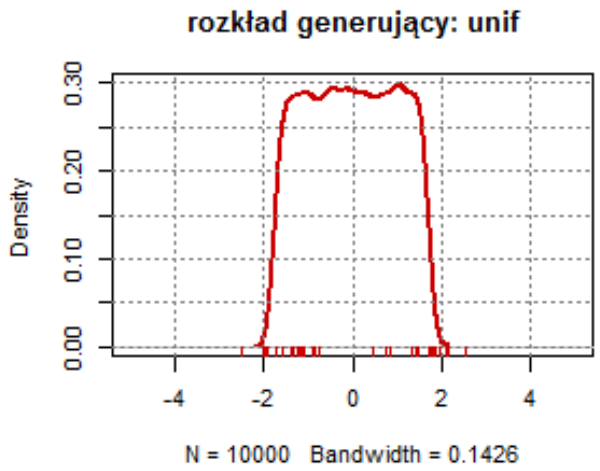


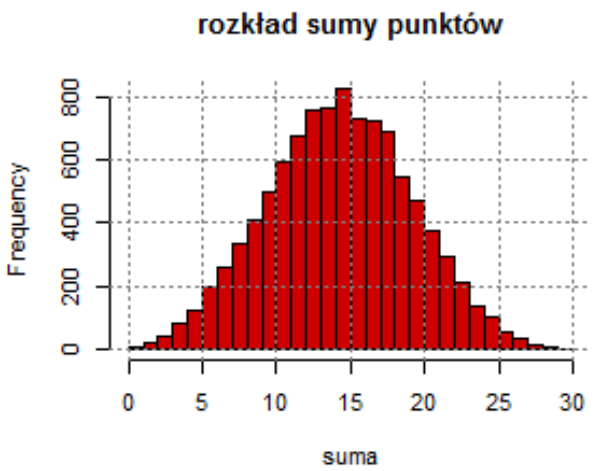
Brak zadań o średniej trudności

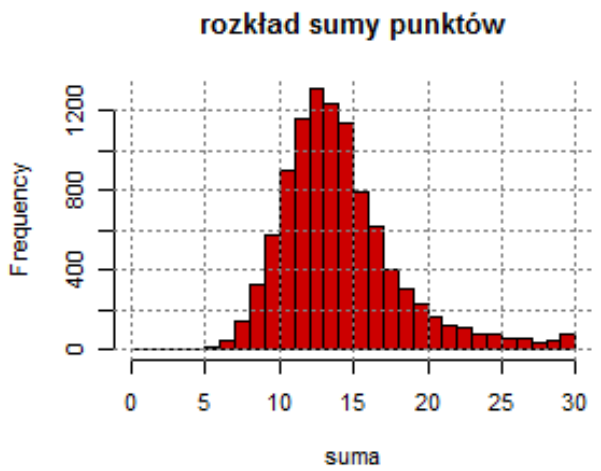
kształty rozkł. sumy punktów zbliżone do rozkł. generujących

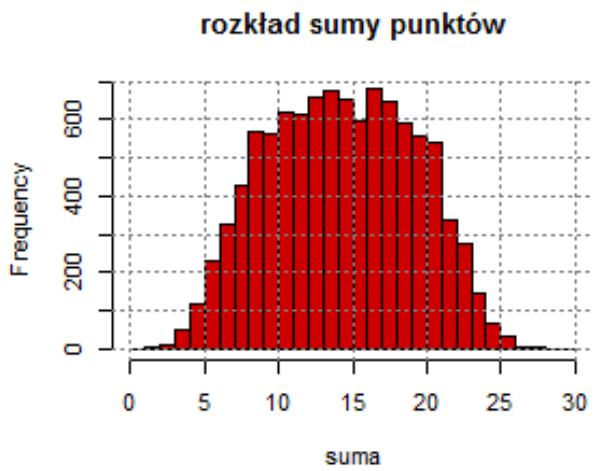






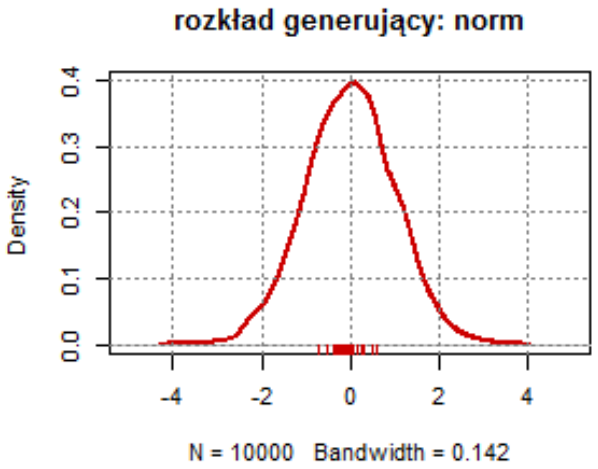


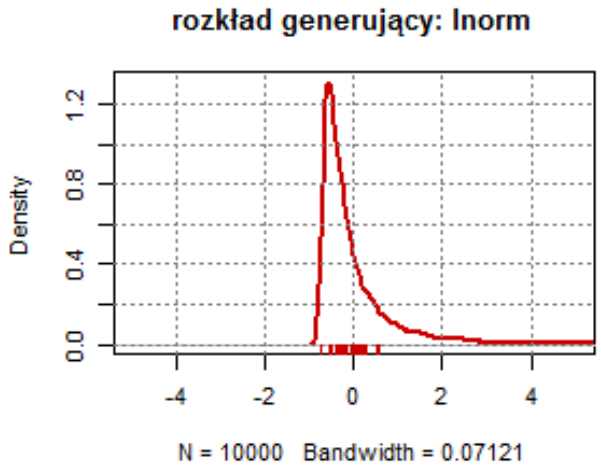


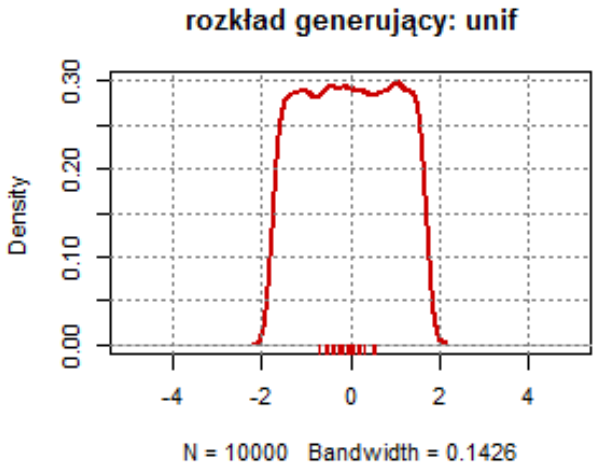


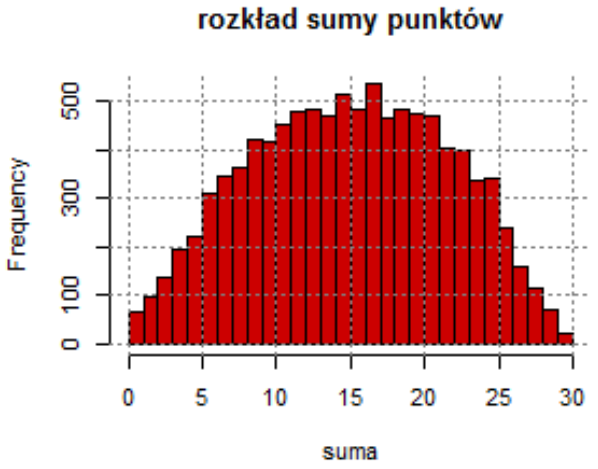
Zbyt małe zróżnicowanie trudności

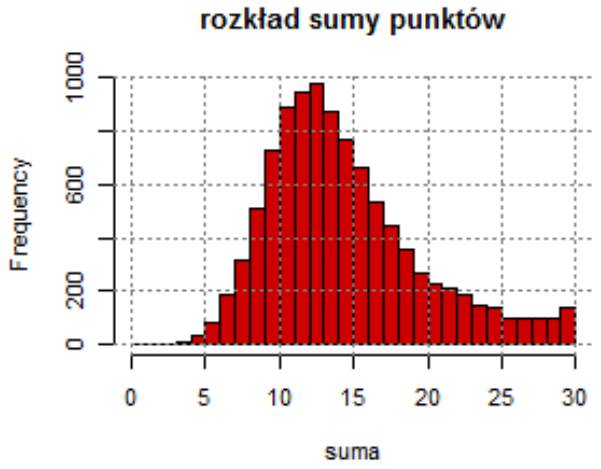
„rozszerzenie” rozkł. sumy punktów w jego środkowej części

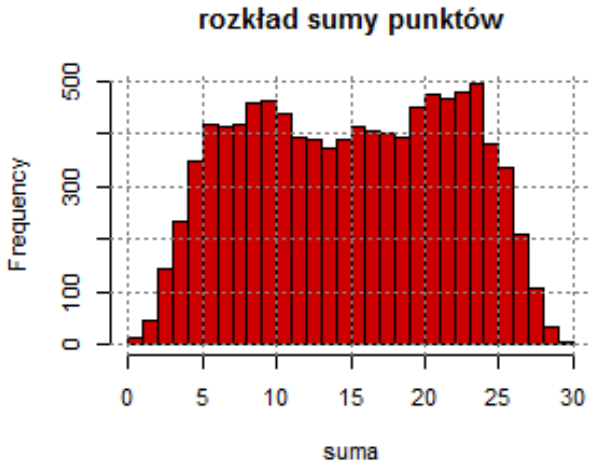






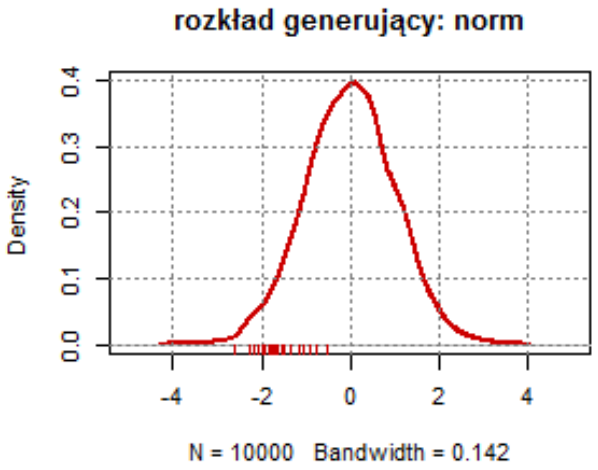


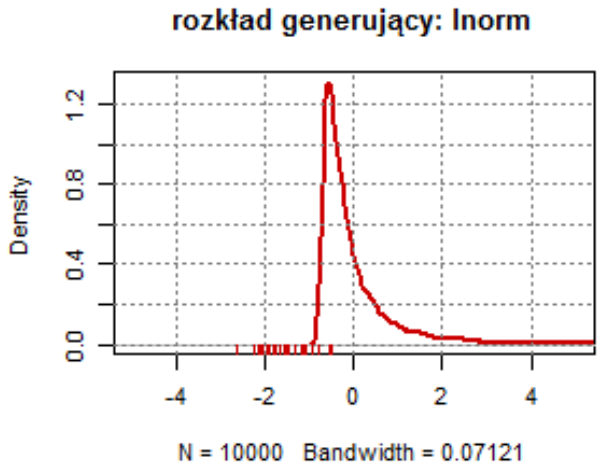


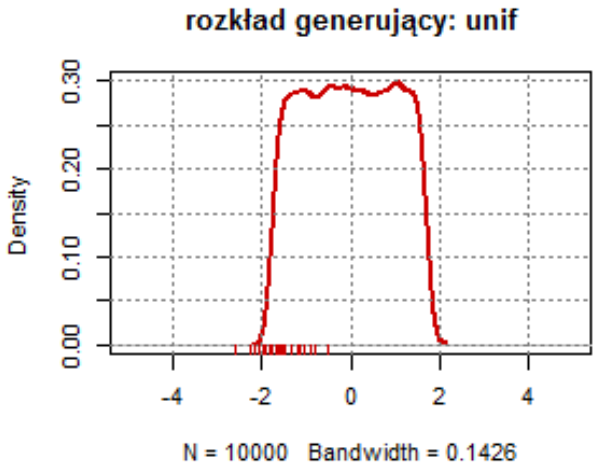


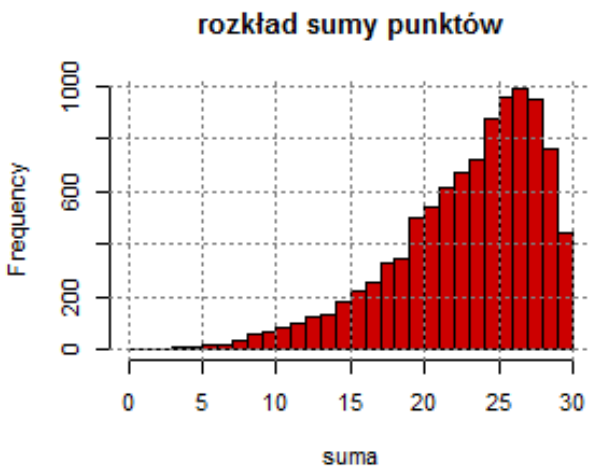
Brak trudnych zadań

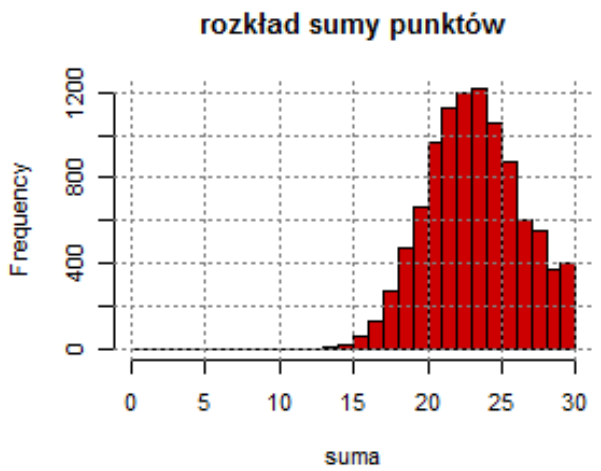
kształty rozkł. sumy punktów są do siebie bardzo podobne, bez względu na rozkł. generujący

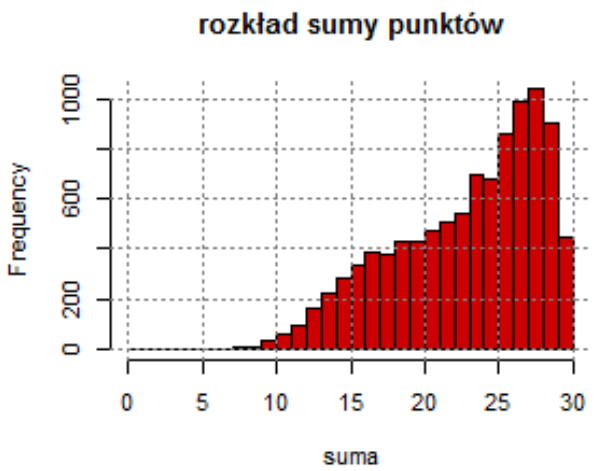






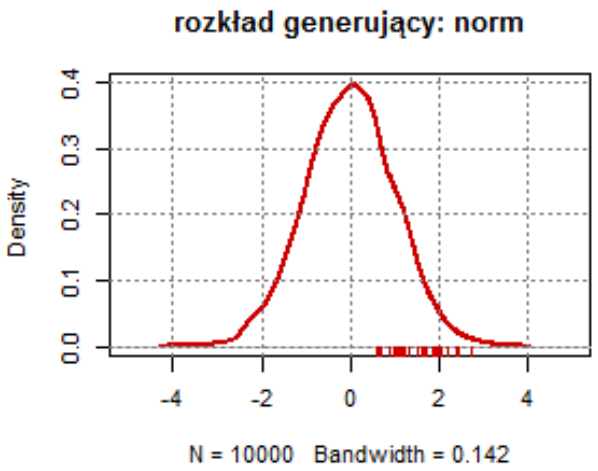


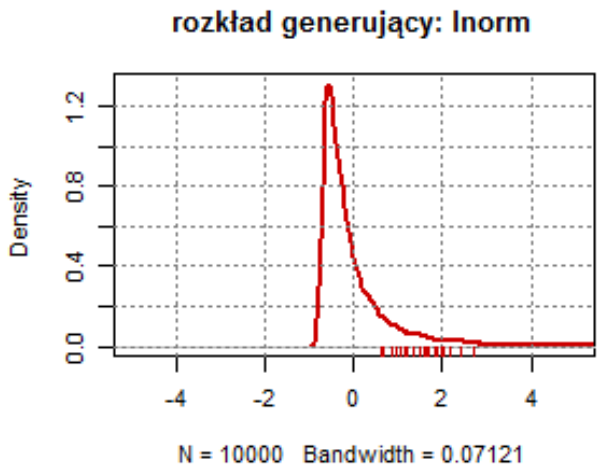


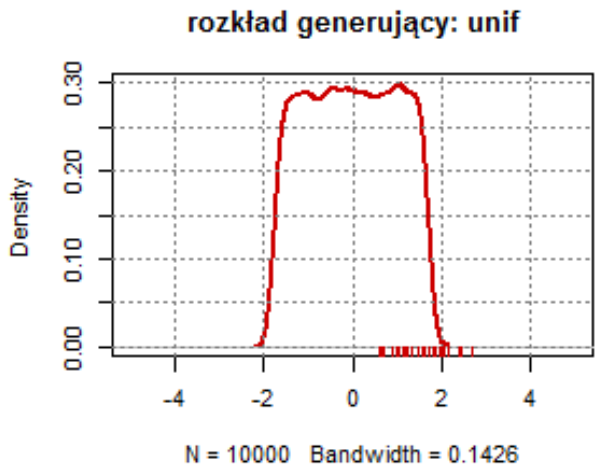


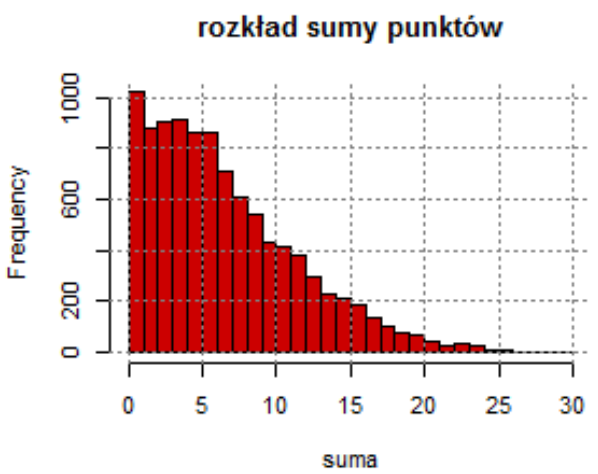
Brak łatwych zadań

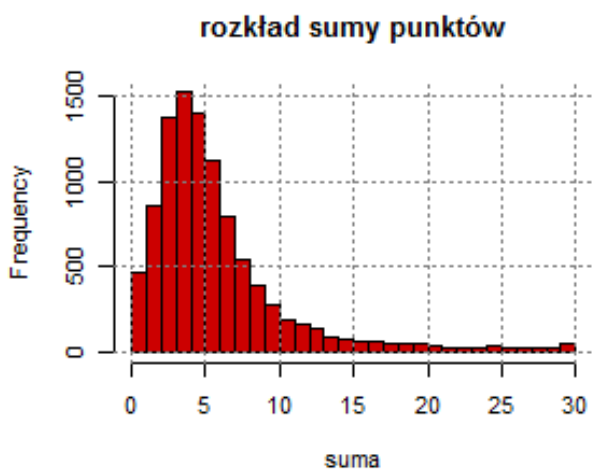
kształty rozkł. sumy punktów są do siebie bardzo podobne, bez względu na rozkł. generujący

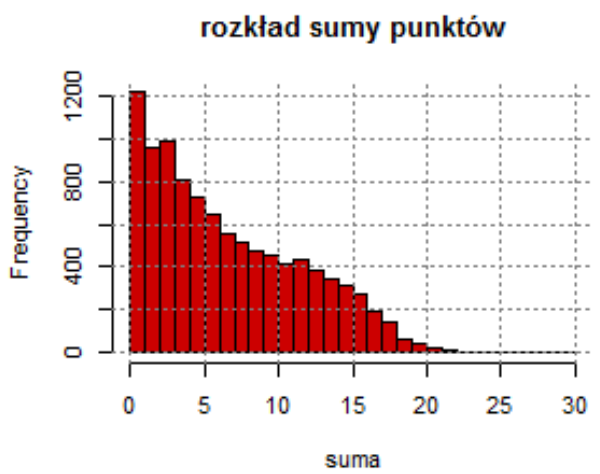












Problemy: skala wyników

- W modelach z ciągłą cechą ukrytą skalę zmiennej ukrytej musimy *zapożyczyć* z którejś zmiennej obserwowanej.
  - Jednak gdy cechę ukrytą ze zmiennymi obserwowalnymi łączy zależność logistyczna/ probitowa, taka skala jest bardzo kłopotliwa interpretacyjnie.
- Alternatywnie możemy ustalić skalę w odniesieniu do przewidywanych parametrów rozkładu zmiennej ukrytej w ramach badanej grupy (lub innej grupy, przebadanej już wcześniej tym samym testem).
  - Jest to rozwiązanie ułatwiające interpretację... ale często nie tak bardzo.

Problemy: skala wyników

Definicja skali PISA:

Wyniki testu PISA określone są na skali takiej, że w roku stanowiącym punkt odniesienia:

- Średnia wyników uczniów z krajów OECD biorących udział w badaniu, wyliczona tak, że każdy kraj ma równy wkład w wyliczaną średnią, jest równa 500.
- Odchylenie standardowe wyników uczniów z krajów OECD biorących udział w badaniu, wyliczona tak, że każdy kraj ma równy wkład w wyliczane odch. stand., jest równe 100.

I dlatego, aby odbiorcy mieli poczucie, że rozumieją, OECD woli *po prostu* mówić, że wyniki określone są na skali od 0 do 1000 punktów, która ma średnią 500...

Problemy: co mierzy zadanie?

Czy te trzy zadania mierzą to samo?

1. Jurek miał dwa żołnierzyki. Na urodziny dostał od Jacka jeszcze dwa. Ile żołnierzyków ma teraz?
2. 2 + 2 = ?
3. Ania miała dwie lalki. Na urodziny dostał od Zosi jeszcze dwie. Ile lalek ma teraz?

Co do zasady brak nam *obiektywnych* kryteriów oceny, czy zadanie (pytanie) mierzy to, co miało mierzyć. Możemy jednak sprawdzać:

- Czy zadanie mierzy coś podobnego do innych zadań?
- Czy mierzy w ten sam sposób w różnych grupach badanych?

Psychometria w R

Psychometryczne pakiety w R (wybór)

Tradycja CFA (estymacja z macierzy korelacji):

- **psych** – różne (również historyczne) odmiany CFA, elementy diagnostyki i wizualizacji danych; mało elastyczna specyfikacja modelu;
- **sem** – modele strukturalne – można dużo więcej, ale proste modele wymagają trochę więcej pisania;
- **lavaan** – modele strukturalne – jeszcze większe możliwości (np. efekty losowe); potrafi „naśladować” wyniki kilku różnych programów komercyjnych (Mplus, EQS);

Psychometryczne pakiety w R (wybór)

Tradycja IRT (estymacja z pełnej macierzy danych):

- **ltm** – pierwszy i najbardziej znany pakiet umożliwiający estymację wielu rodzajów modeli IRT metodą MML, ale: 1) tylko jedno- lub dwuwymiarowych, 2) problemy z działaniem na dużych danych;
- **mirt** – dość młody (od 2011 r.) i prężnie się rozwija; bardzo duże możliwości: 1) estymacja MML algorytmem EM lub bayesowsko, 2) dowolna liczba wymiarów, 3) obsługa regresji latentnej w tym z efektami losowymi (tylko przy estymacji bayesowskiej), 4) szybki i dobrze znosi duże dane;
- **lavaan** – wsparcie dla estymacji MML póki co na wstępnym etapie rozwoju;

Psychometryczne pakiety w R (wybór)

Tradycja IRT (estymacja z pełnej macierzy danych) cd:

- **difR, lordif** – diagnostyka zróżnicowanego funkcjonowania zadań;
- **mokken** – nieparametryczne modele IRT;
- **CDM** – *cognitive diagnostic models* – modele zakładające, że cechy ukryte mają charakter dychotomiczny (względnie porządkowy);
- sporo pakietów z wariacjami na temat modelu Rascha;

Wyskalujmy wzrost z pakietem mirt

- Choć zwykle nie mamy tego komfortu by móc sprawdzić, na ile dobrze nasze pytania mierzą to, co w założeniach miały mierzyć, tutaj posłużymy się testem, który bada cechę poddającą się łatwemu i dosyć precyzyjnemu pomiarowi w inny sposób.
- Oczywiście aby przeliczyć wyniki na centymetry musimy skądinąd znać średnią i wariancję wzrostu w badanej grupie...
- Dobrze określony konstrukt powinien nam pozwolić wyskalować model nawet na bardzo małych danych.

Wyskalujmy wzrost z pakietem mirt

Binarne **pytania o wzrost:**

P1. Zdarza mi się słyszeć, że mógłbym zostać koszykarzem.

P2. W autobusie mogę wygodnie trzymać się górnych (poziomych) uchwytów.

P3. Łóżka w hotelach są dla mnie zwykle za krótkie.

P4. Inni ludzie proszą mnie, żebym podał im rzeczy, które leżą wysoko.

P5. Wchodząc do pomieszczeń muszę nieraz uważać, aby nie uderzyć się w głowę.

P6. Sięganie po rzeczy z szafek wiszących w kuchni sprawia mi trudność.

P7. Siedząc w samochodzie (...) mam często zbyt mało miejsca na nogi.

P8. Wolał(a)bym być wyższy.

P9. Rozmawiając z innymi często muszę zadzierać głowę.

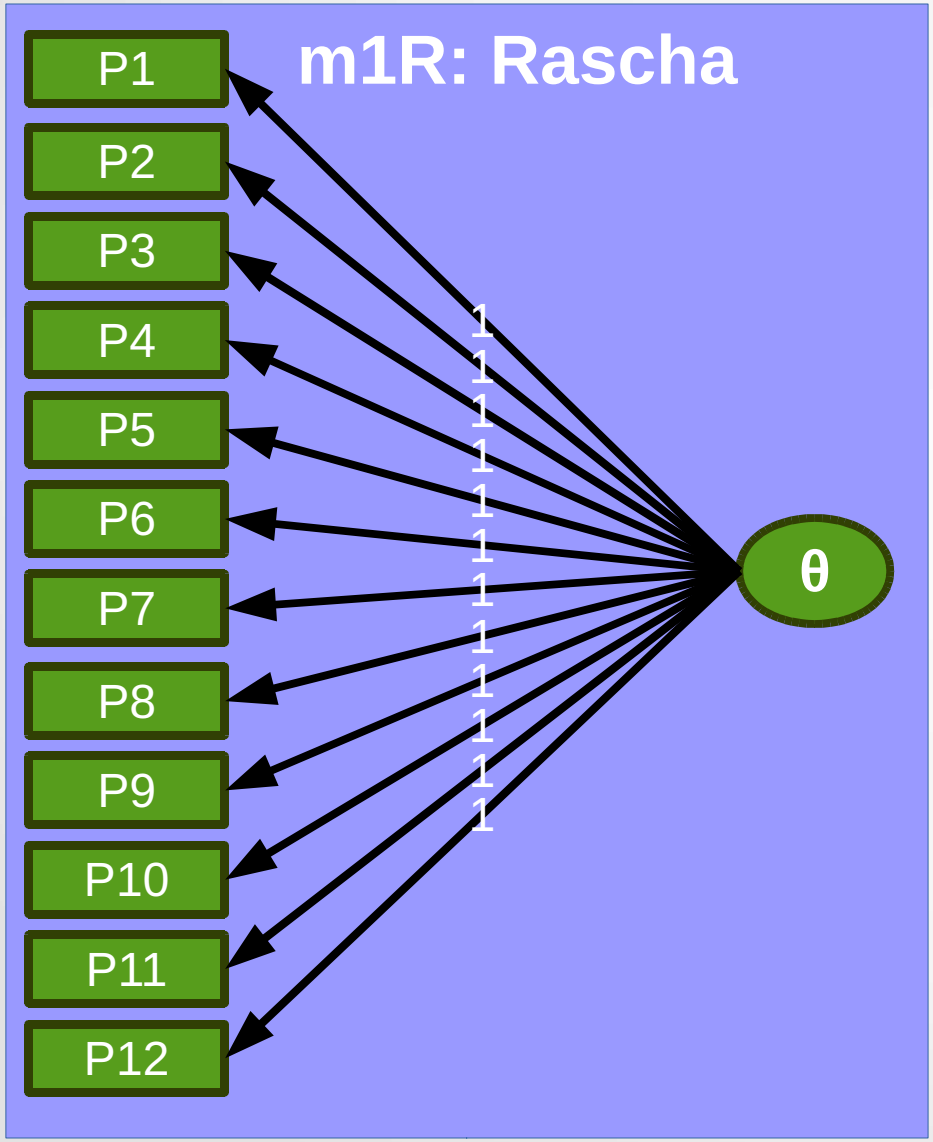
P10. Na większość ludzi mogę patrzeć z góry (dosłownie, nie w przenośni).

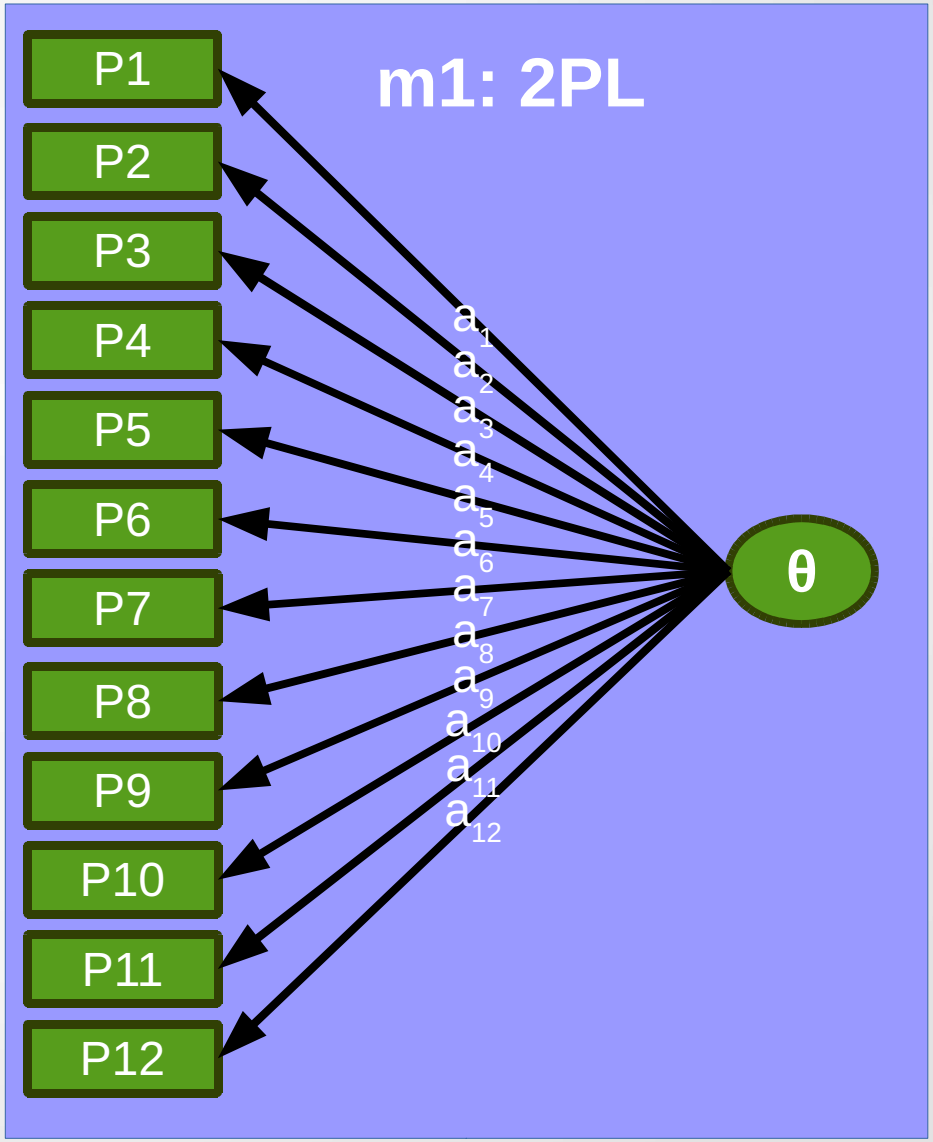
P11. Na fotografiach grupowych zwykle stoję w pierwszym rzędzie.

P12. Na koncertach muszę stać blisko sceny (ekranu), bo inaczej nic nie widzę.

Wyskalujmy wzrost z pakietem mirt

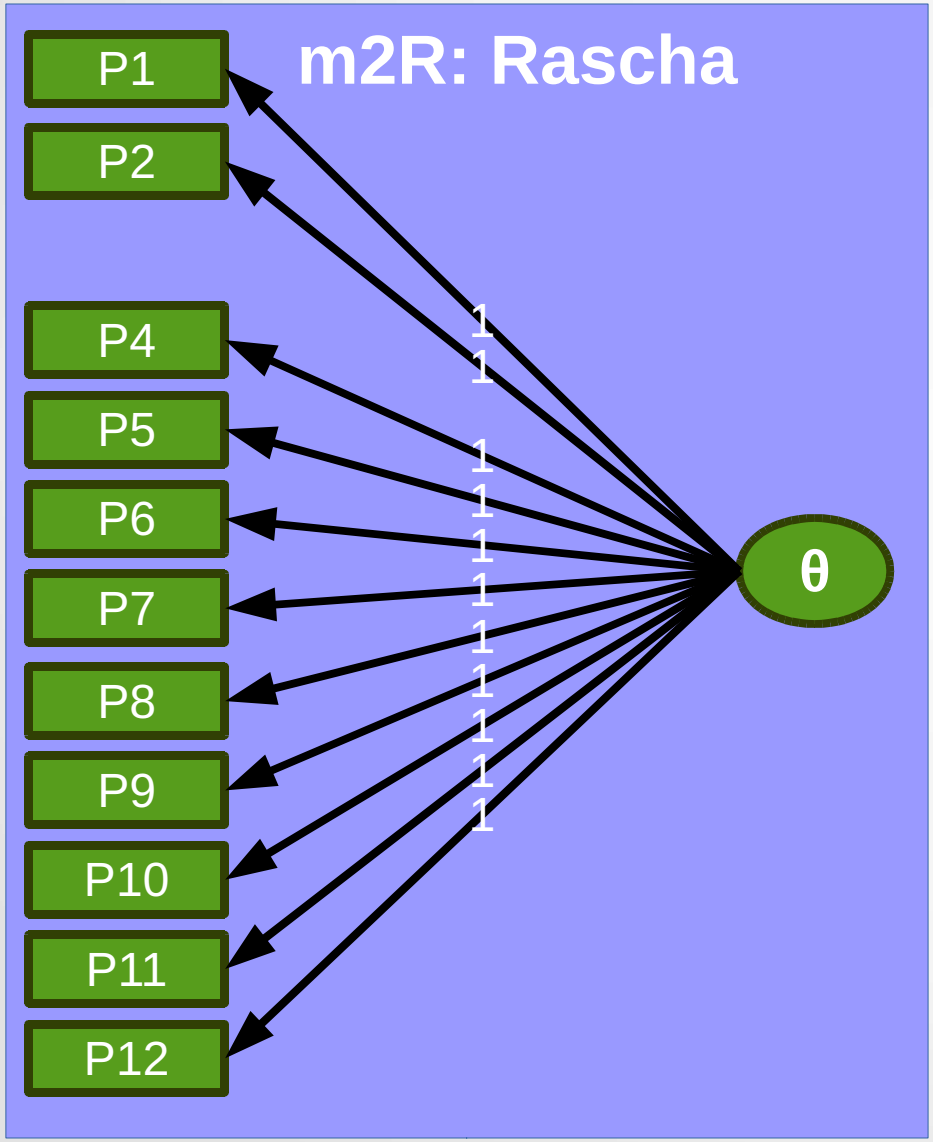
Nasze modele:

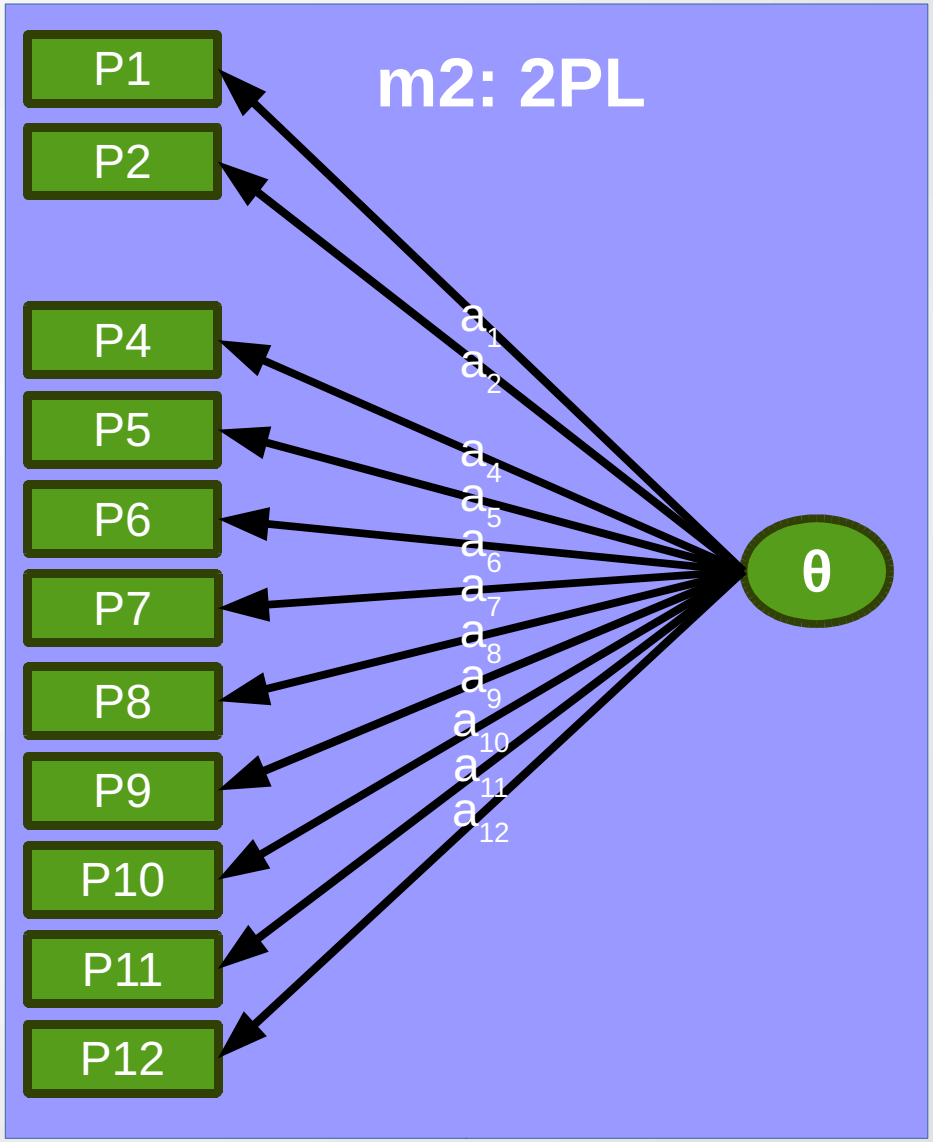




Wyskalujmy wzrost z pakietem mirt

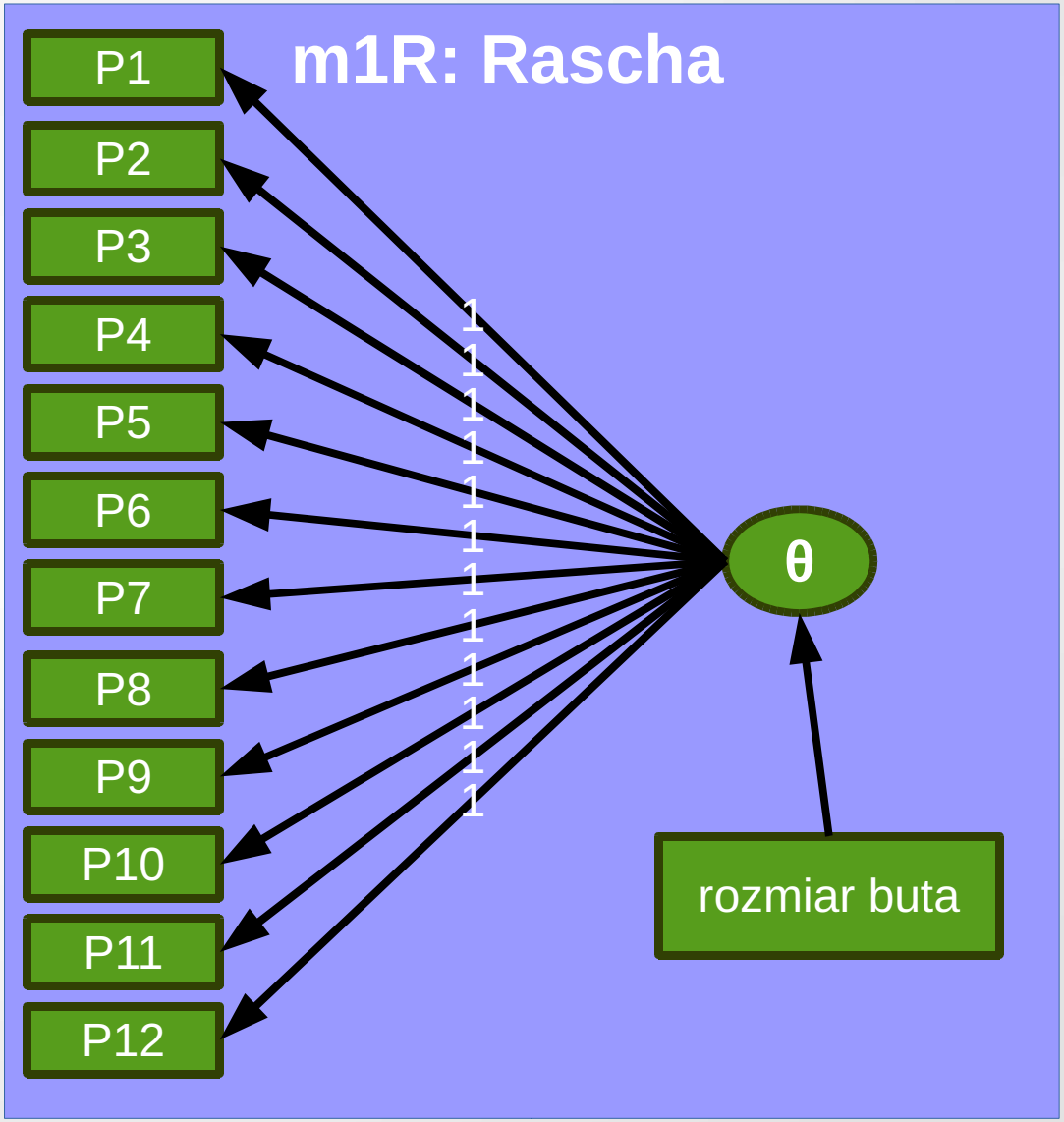
Nasze modele:





Wyskalujmy wzrost z pakietem mirt

Miała być jeszcze regresja latentna, ale mamy za mało danych, żeby wyszła sensownie:



Dziękuję za uwagę!

Tomasz Żółtak
t.zoltak@ibe.edu.pl