3.10.2025 – ćwiczenia 1

**Hasło do moodle:** aktuariat

**Warunki zaliczenia:**

* sprawdzian na ostatnich zajęciach
  + rozwiązywanie przykładów zadań, przeliczenia
  + forma quizu – miejsce na obliczenia i miejsce na wpisanie wyniku
  + mogą być pytania testowe
  + max 1 otwarte pytanie
  + tylko zagadnienia, które były na ćwiczeniach
* każde zgłoszenie to 1p.p., max 5p.p. wpisują się do oceny tylko jak ma się zaliczenie
  + interpretacje, dyskusje, zadania
* nieobecność na ponad 50% zajęć = nieklasyfikowanie
  + nieobecność usprawiedliwić można też przed zajęciami
  + można mieć 2 (?) nieobecności nieusprawiedliwione

**Tematyka:**

* zmienne kluczowe w ubezpieczeniach (liczba szkód – zm. licznikowa, stosujemy modele dyskretne)
* wysokość szkody (zm. ciągła – modele/rozkłady ciągłe)
* funkcje tworzące i charakterystyczne (operują na liczbach zespolonych), sploty funkcji, szybka transformata Fouriera – tematy bardziej teoretyczne
* model indywidualnego i kolektywnego ryzyka w ubezpieczeniach innych niż na życie
  + dział ubezpieczeń na życie jest specyficzny, bo oprócz komponentu statystycznego dochodzi komponent demograficzny - tablice trwania życia
* składki
* taryfikacja ryzyka – tworzenie taryf w zależności od różnych determinant

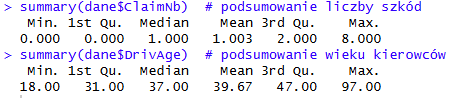
Będziemy pracować w R i głównie w Excelu

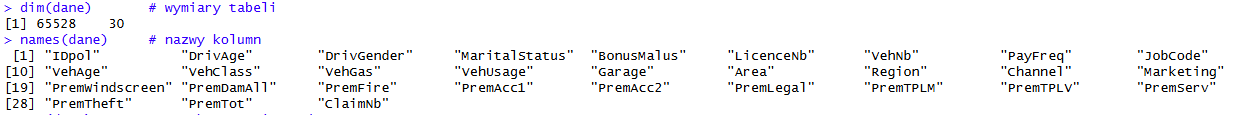
Dwa zajęcia będą online, w okolicach świąt – 19.12, 9.01 – przełożone na inne dni w tych tygodniach w które mamy zajęcia online

12.12 – przełożone na 10.12

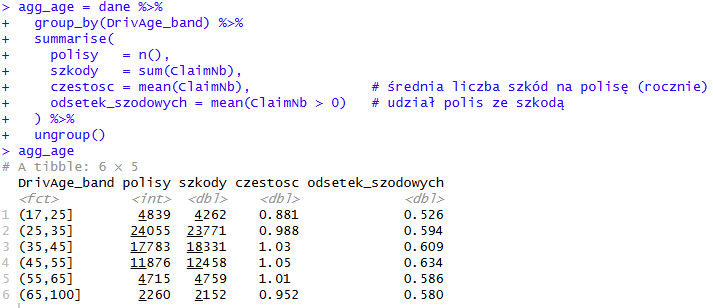
# ZAJĘCIA

Paczka danych – realne dane nt. ubezpieczeń kierowców



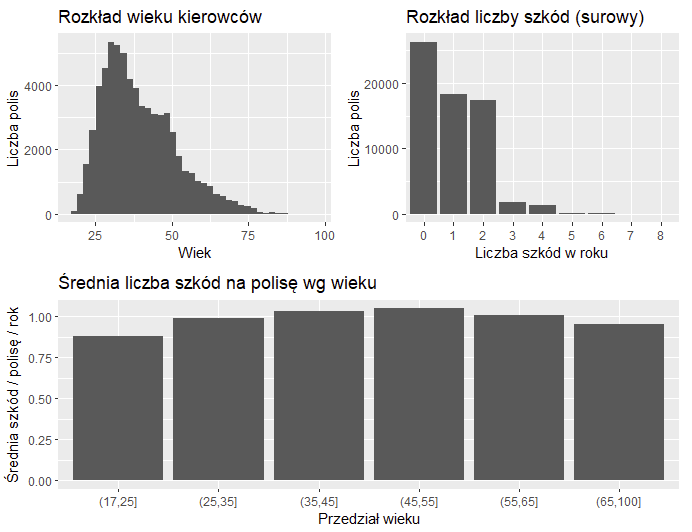


* 65528 polis, 30 zmiennych
* Id, wiek, płeć, stan cywilny, …, vehicle number – moc samochodu; rodzaj paliwa, czy garażowany, jaki obszar (gdzie ubezpieczony),..
* ClaimNb – liczba szkód w ciągu czasu wystawienia na ryzyko (exposure period, czas ubezpieczenia – dla nas roczny, czyli polisy są roczne, więc nie mamy takiej kolumny, bo dla wszystkich jest = 1) – najważniejsza dla nas zmienna



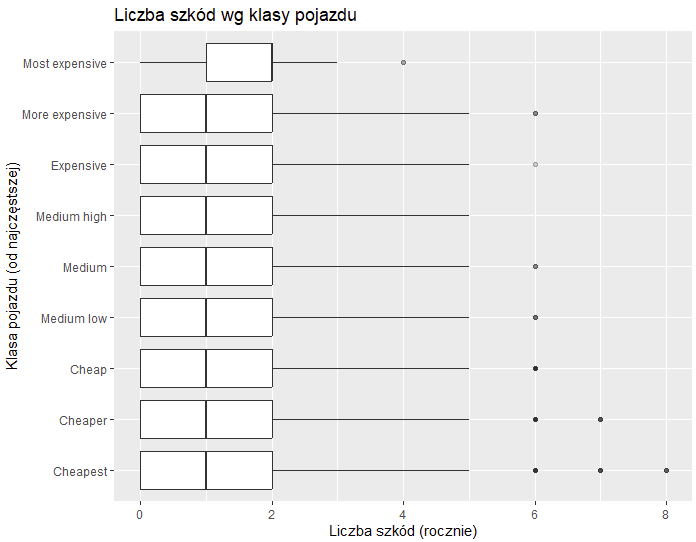
* Odsetek polis szkodowych – te polisy, gdzie jest przynajmniej 1 szkoda
* Najczęściej polisy ze szkodą są w przedziale wiekowym 46-55 (kto by pomyślał, że młodzi nie rozwalają aut najczęściej xd) – może w modelu wyjdzie inaczej

 🡨 zajebisty sposób na przedstawienie wykresów w innym układzie

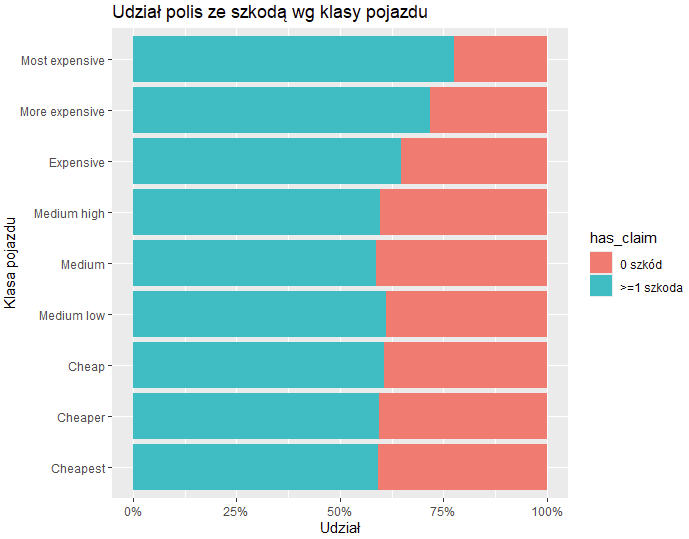


Rozkład liczby szkód – zmienna dyskretna; najwięcej polis jest bezszkodowych, a ogólnie najczęściej jest 0-2 szkód. Typowy rozkład liczby szkód ma płaski ogon, jest to problematyczne (szybki spadek, a potem płaski ogon)

**Boxplot**

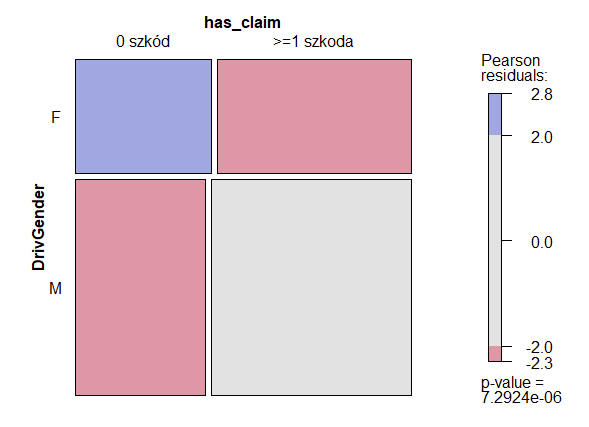


Most expensive jako jedyna kategoria, w której widać różnicę (ale ta kategoria jest najmniejsza) – 0 nie występuje w 75% w ogóle, a mediana to 2



W drogich samochodach odsetek polis szkodowych jest wyższy, bo:

* pokusa wykorzystania mocy silnika xD
* może bogatsi ludzie nie pilnują się aż tak, bo stać ich bardziej na wypadki
  + imo statystycznie bogaci chyba żyją bardziej ryzykownie z tego co pamiętam (bo żeby dużo zarobić trzeba mieć jaja i podejmować ryzyko), więc w teorii powinni uczestniczyć też w większej ilości wypadków niż biedniejsi którzy żyją statystycznie dużo bezpieczniej i ostrożniej
* może droższe samochody są firmowe więc nie jest ich aż tak szkoda
* jeśli samochód jest droższy to bardziej opłaca się zgłosić szkodę, bo w tańszym się nie opłaca

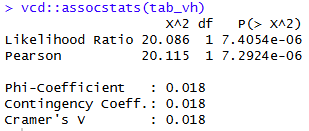


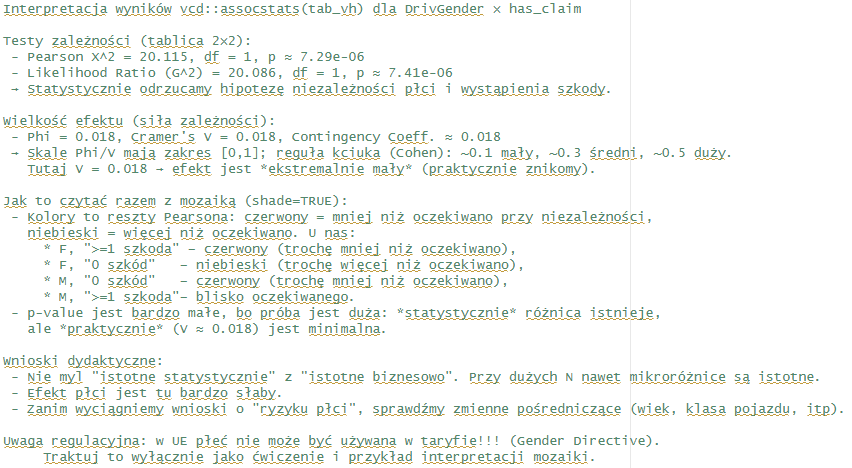
**Wykres Mozaiki**

* niebieski – wartości dodatnie, czyli więcej niż wynikałoby z równomiernego rozkładu
* szary – rozkład równomierny
* czerwony – wartości ujemne, czyli mniej niż wynikałoby z równomiernego rozkładu

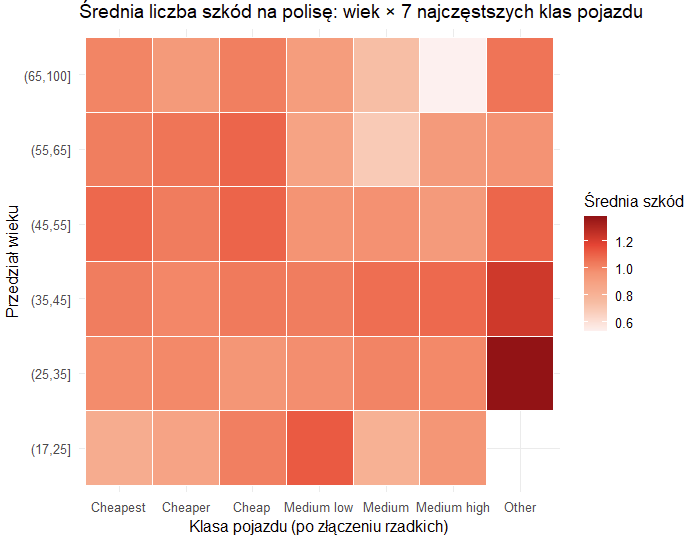
W zależności od płci – kobiet z 0 liczbą szkód jest więcej niż wynika z równomiernego rozkładu, a z >=1 jest mniej niż wynika z równomiernego rozkładu. To nie jest duże zróżnicowanie

Statystyki tej zależności:



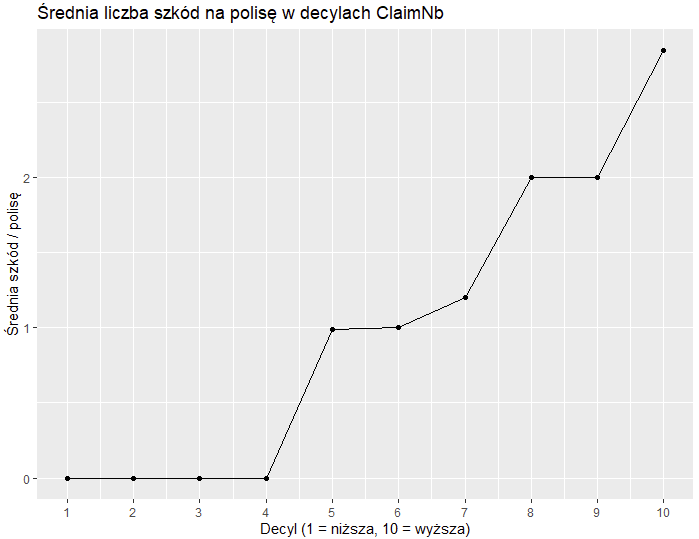


* Istotne statystycznie **=!** Istotne dla nas
* Duże dane = dużo danych istotnych statystycznie, ale niekoniecznie istotnych „życiowo”
* Funfact: prawnie ubezpieczyciel nie może uzależniać kwoty ubezpieczenia od płci

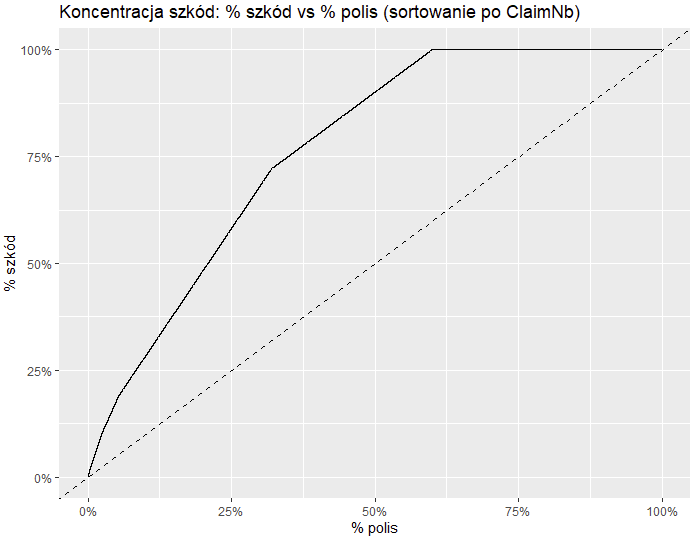


**Częstość szkód w danej klasie pojazdu**

* Im ciemniej tym więcej szkód
* Kategoria samochodów najdroższych wypada najgorzej pod względem liczby szkód
  + Im młodszy kierowca z drogim samochodem tym gorzej
* W najtańszych samochodach – im starsi kierowcy tym gorzej
* Klasa średnia samochodów – młodsi kierowcy najgorzej, najlepiej najstarsi kierowcy



10% najbardziej szkodowych polis średnio ma 3 szkody na polisę



Koncentracja szkód: czy szkody równomiernie się rozkładają pośród wszystkich polis

* Połowa polis generuje 50% szkód
* Nasza krzywa nad krzywą równomierną

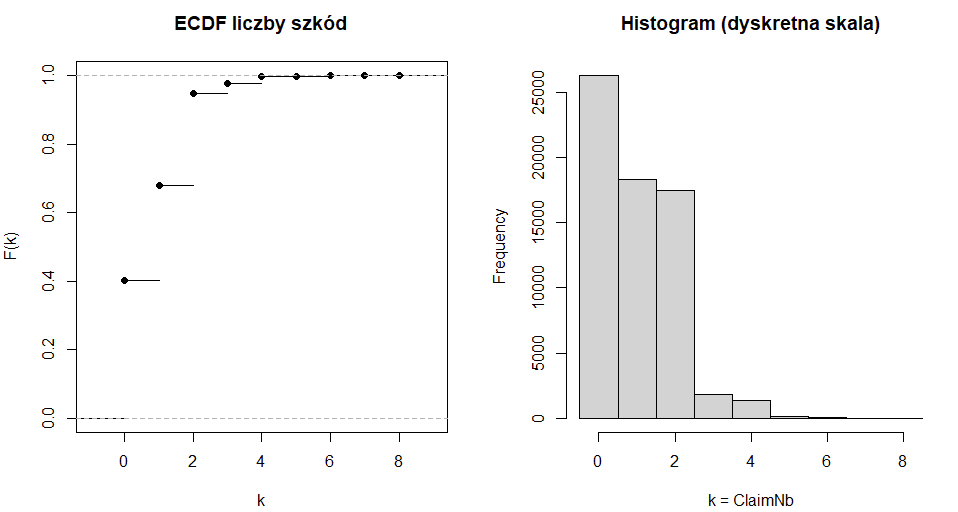
Wskaźnik pareto (80/20):



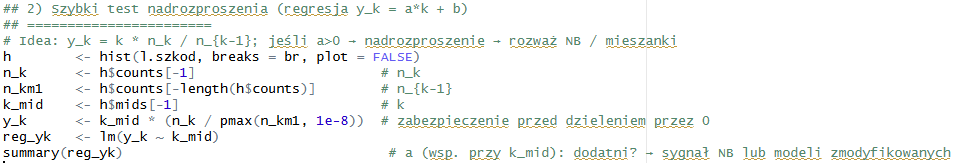
* 80% szkód generuje 40% polis
* Liczba szkód **nie jest** skoncentrowana w małej liczbie polis – bo w aż 40%, więc nie ma reguły pareto
* 100% szkód generuje ~60% polis

10.10 – ćwiczenia 2

* Zmienna licznikowa – jest określona na liczbach naturalnych
  + Mamy do tego specjalne modele:
    - Poissona
    - Bernoulliego
    - Ujemny dwumianowy
    - Geometryczny
* Dystrybuanta empiryczna (po lewej) i histogram (po prawej)
  + Najwięcej przypadków to 2 lub mniej szkód, najczęściej nie ma szkód
  + Ogon cienki i szybko spada – specyfika zmiennych aktuarialnych
  + Dystrybuanta – w 2 mamy jakieś 95% przypadków
  + Dobre oszacowanie ogona jest kluczowe dla ubezpieczyciela



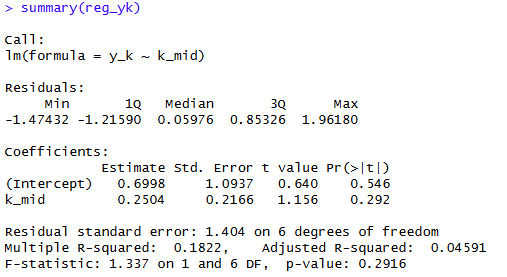
* Znak współczynnika powie nam czy mamy nadrozproszenie czy nie
* k: 1-8



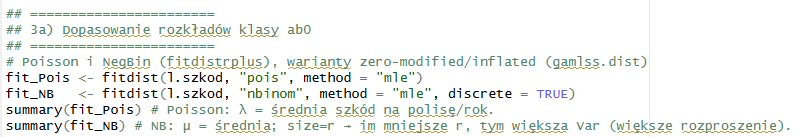
* Liczebność 1 szkody stanowi 70% liczebności zer
* Liczebność 2 szkód stanowi 95% liczebności jedynki
* Liczebność 3 szkód stanowi 32% liczebności dwójki



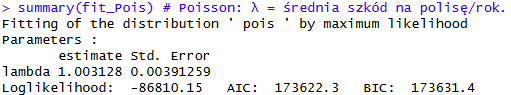
* Ocena parametru a – jak jest 0 to dobry będzie rozkład Poissona, jak > 0 to mamy nadrozproszenie, ale p-value > 0.05 (nieistotne statystycznie), więc i tak sprawdzi się Poisson
  + Jak > 0 to ujemny dwumianowy, lub geometryczny

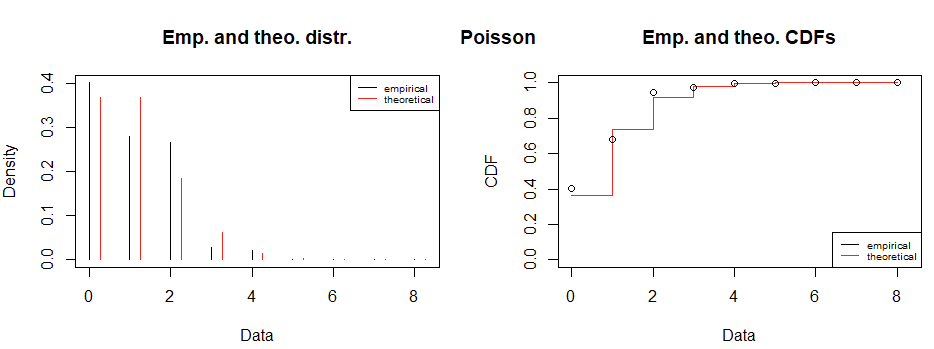


* Dystrybuanta – prawdopodobieństwo, że zmienna x będzie mniejsza od zmiennej a
* Rozkłady klasy ab0 (a, b, 0) – następujące po sobie prawdopodobieństwa jesteśmy w stanie uzależnić od dwóch stałych, rekurencyjnie:

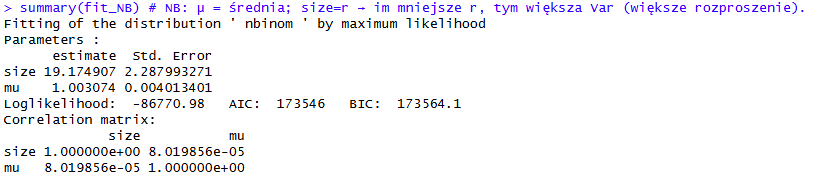


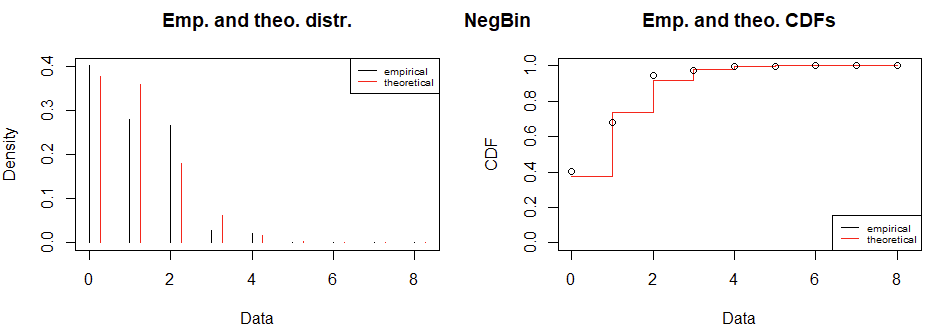
* Fitdist – do dopasowywania gołych rozkładów
* Poisson
  + Jeden parametr – lambda – średnia liczba szkód wynosi około 1 na polisę, +/- 0.004 szkody
  + Rozkład niedoszacowuje zerowych i 2 liczebności, przeszacowuje 1 i 3, dobry ogon



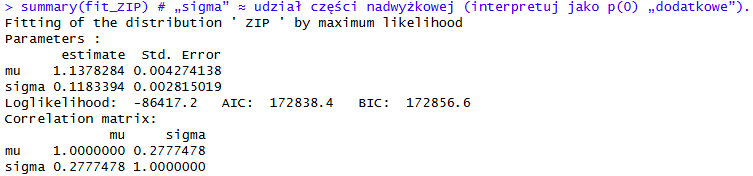


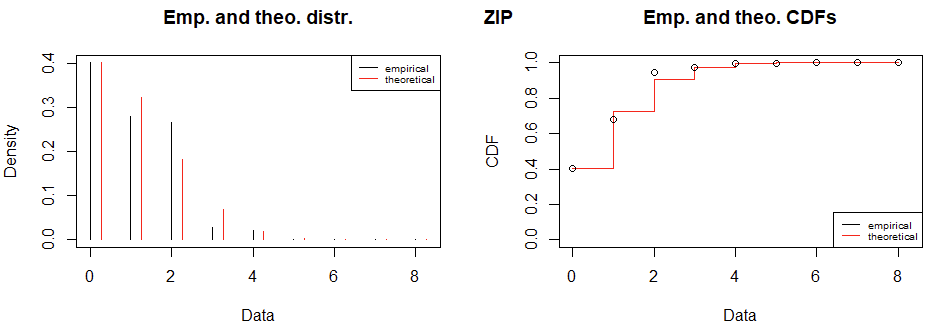
* Ujemny dwumianowy
  + Im mniejsze r tym większa wariancja (rozproszenie)
  + Średnia liczba szkód też około 1 na polisę, +/- 0.004 szkody
  + Size 19, więc dość niewielkie rozproszenie



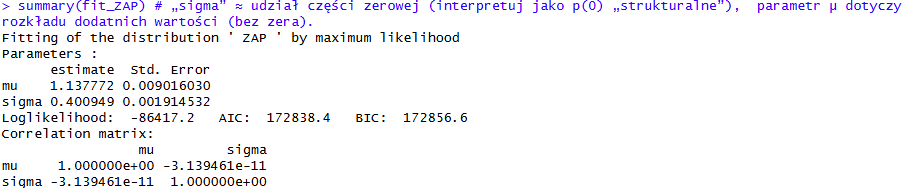


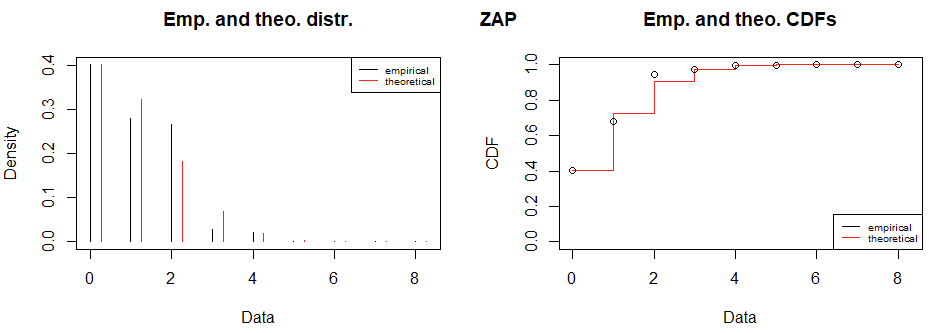
* Zero-Inflated Poisson (ZIP) – model Poissona z nadwyżką zer, czyli mieszanka 0 i Poissona
  + 0.12 – częstość występowania nadwyżkowych zer
  + W 0 jest idealny, właśnie po to dokładamy dodatkowy parametr na 0 żeby dobrze dopasować 0. Dzięki temu poprawiły się doszacowania w 1 i 2



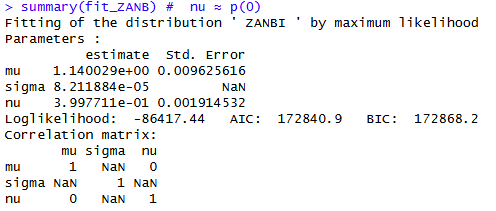


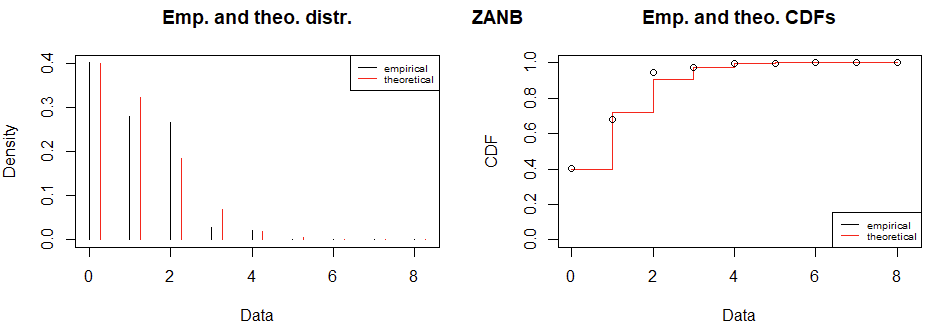
* Zero-Adjusted Poisson (ZAP) – zero jest osobno modelowane przez rozkład zero-jedynkowy, a wartości dodatnie rozkładem Poissona
  + Zero całe jest modelowane przez sigmę – prawdopodobieństwa zera w modelu to 0.4
  + Średnia około 1.13, jest większa, bo nie uwzględnia zer



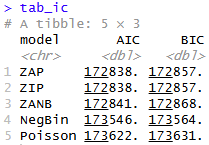


* Zero-Altered NegBin (ZANBI) – zero osobno modelowane, rozkład dwumianowy przesunięty
  + Sigma = 1 / size
  + nu – prawdopodobieństwo zera
  + średnia ~ 1.14
  + dodanie kolejnego parametru w sumie niewiele wniosło

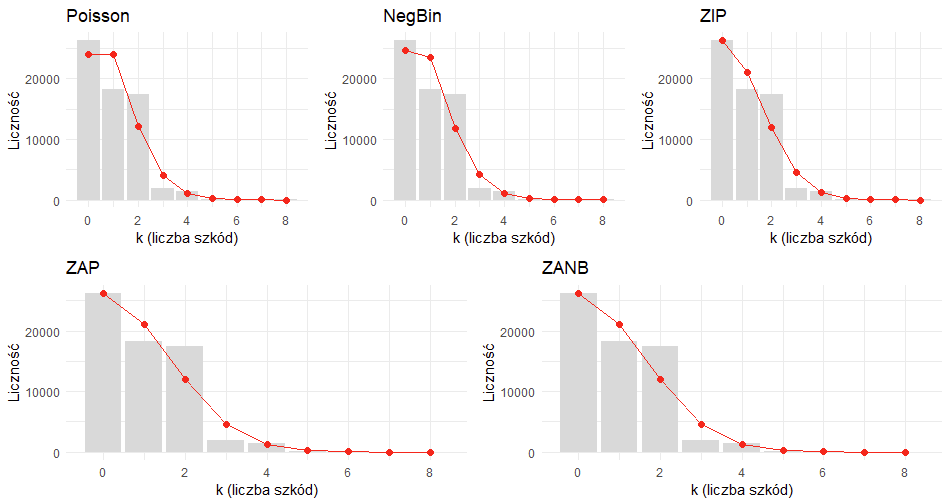




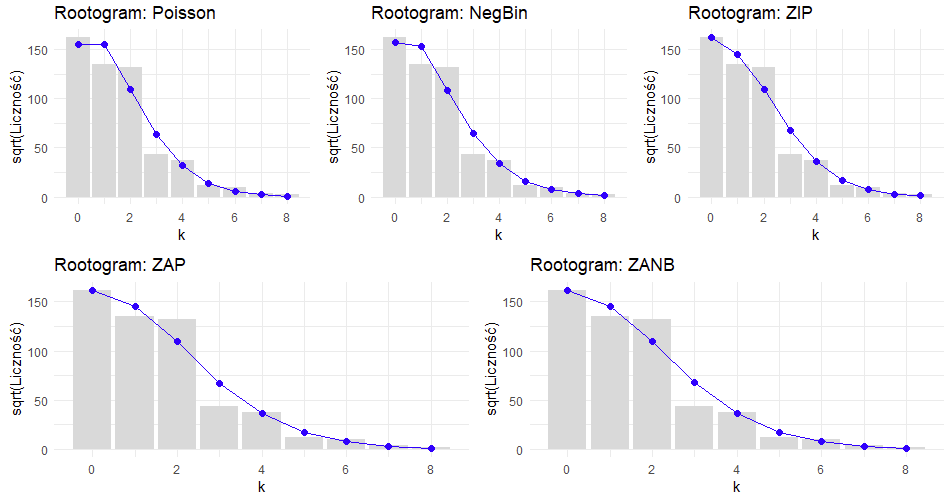
* Rozkłady z dodatkowym parametrem w 0 zdecydowanie dopasowują się lepiej
* Kryteria AIC, BIC
  + Najlepiej ZAP i ZIP (identycznie)
  + Więc ZAP czy ZIP?
    - ZAP – dla np. dzietności
    - ZIP – dla np. liczebności zawieralności małżeństw



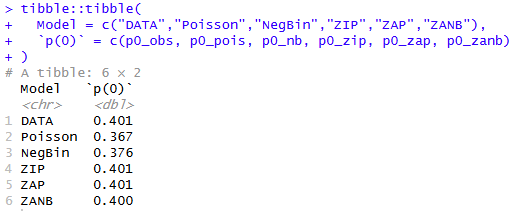
* Słupki – rozkład empiryczny
* Czerwone – rozkład teoretyczny
* Poisson meh, NegBin trochę lepiej ale nadal słabo, dopiero modele klasy ab1 (a, b, 1) - ZIP, ZAP i ZANB lepsze



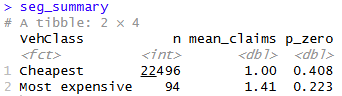
* Pierwiastek liczebności – wszystkie sobie w miarę dobrze poradziły



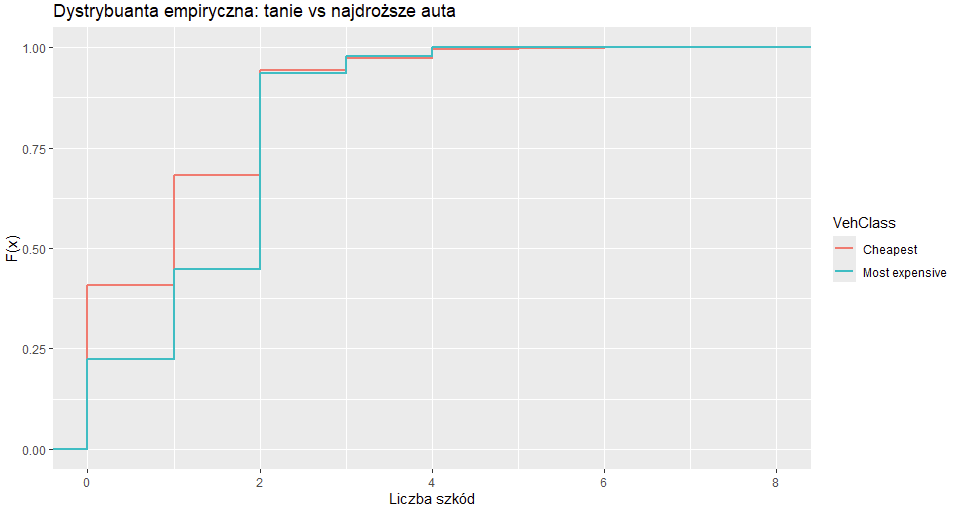
* Jak dokładnie różne modele dopasowują 0 lub 1?



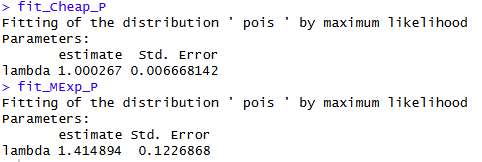
* Modele mogą być różne dla różnych podkategorii pojazdów:



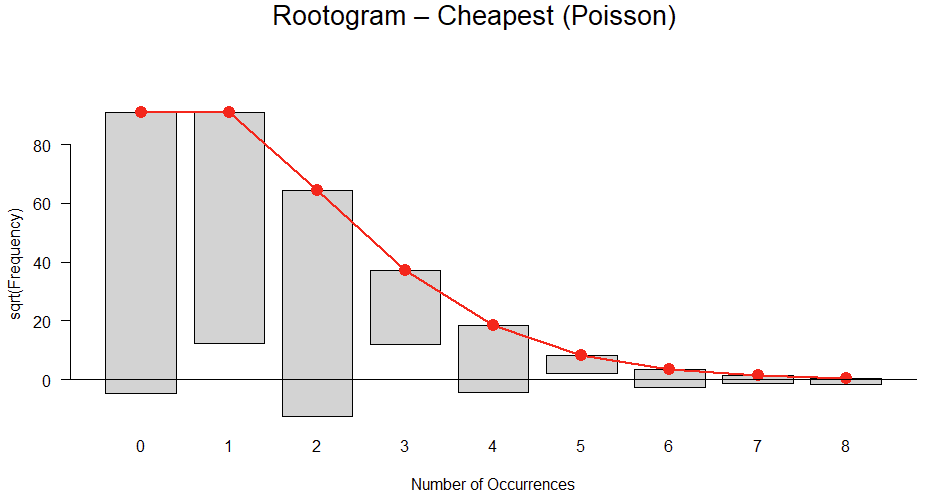
* Szybciej idzie w górę kategoria cheapest, bo jest więcej mniejszej ilości szkód niż w most expensive



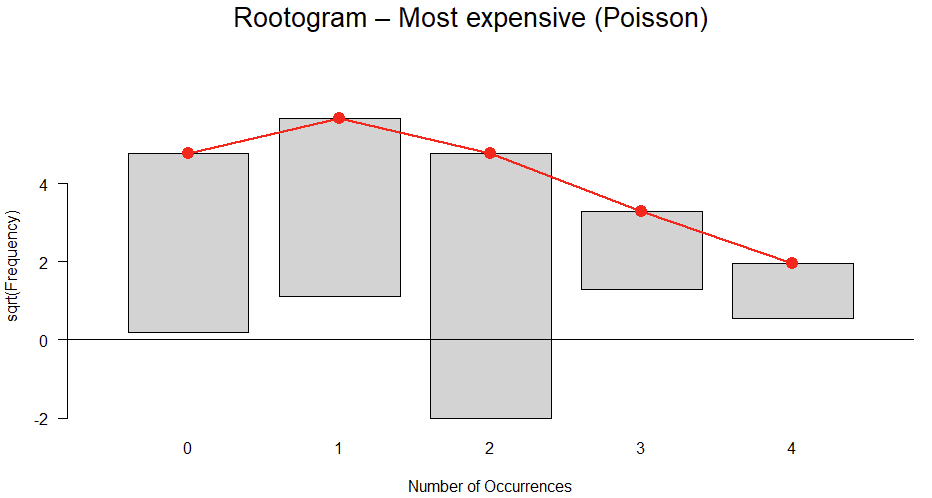
* Lambda większa dla most expensive, więc się zgadza bo częstsze szkody



* 0, 2 i 4 są niedoszacowane (poniżej osi), 1 i 3 przeszacowane



* Tylko 2 niedoszacowana, reszta przeszacowana



* Czyli powinno się cenę samochodu wziąć pod uwagę w taryfach, bo widać że ma wpływ na liczbę szkód

