---=== STATYSTYKA ===---

***---== Temat 3 ==---***

Populacja -> rzecz którą chcemy przebadać [ zbiór pewnych elementów ] np.: ludzie mieszkający w PL, zwierzęta w lesie lub populacja gwoździ w fabryce.

Zmienna/cecha -> określona rzecz którą badamy w populacji. Każda cecha ma swój rozkład prawdopodobieństwa. Bardzo często, nie jesteśmy w stanie zbadać całej populacji, dlatego też pracujemy na tak zwanej „próbie” – czyli podzbiorze populacji.

X = (y1, … , yn)T [ gdzie X: jest to cecha/zmienna || a Y: jest to pojedyncza obserwacja ]

Metoda analizy danych jest zależna od typu obserwacji (czy jest to obserwacja jakościowa czy ilościowa? Itd.)

MIARY TENDENCJI CENTRALNEJ:

* Średnia arytmetyczna jest dobrą miarą o ile mamy rozkład symetryczny. Jeśli natomiast mamy odstępstwa od tej symetrii (obserwacje odstające) wtedy średnia może być przez nie zniekształcona.
* Mediana (szczególny przypadek kwantylu rzędu 1/2) bazuje na próbie uporządkowanej (posortowanych wartościach od najmniejszego do największego) Połowa obserwacji jest większa od mediany i połowa obserwacji jest mniejsza od mediany. Mediana jest bardziej odporna na pojawienie się obserwacji odstających. (estymator odporny?)
* Przy rozkładzie symetrycznym średnia i mediana będą miały podobne wartości ale przy rozkładzie asymetrycznym – będę się już bardziej od siebie różnić.
* Kwantyle / kwartyle – można tworzyć je dla każdego rzędu. (1/4 np.: wskazuje nam podział, że 25% wartości jest mniejsze od K1/4 a 75% jest większe

MIARY DYSPERSJI (miary rozproszenia):

* Tutaj patrzymy czy dane obserwacji raczej są blisko wartości średniej czy raczej są bardziej rozrzucone. Obserwacje dalekie do wartości średniej. Im większa jest tam miara tym większy jest rozrzut.
* Odchylenie standardowe (mierzy rozrzut wartości obserwacji od średniej)
* Odchylenie standardowe jest o tyle dobre, bo praktycznie bardzo łatwo jest go zinterpretować, ponieważ ma jednostki zwykłe (jak coś mierzymy w cm. To wynik również będzie w cm.) W przeciwieństwie do wariancji s2 – która jest trudniejsza w interpretacji ale w teorii jest fajniejsza.
* Łatwiej się posługiwać wariancją, a interpretować łatwiej odchylenie standardowe.
* Współczynnik zmienności – podajemy w procentach. Jego zaletą jest brak jednostki i pozwala nam na porównywaniu zmienności tych samych cech tylko mierzonych różnymi sposobami.

MIARY ASYMETRII ROZKŁADU:

**Współczynnik asymetrii** pozwala nam liczbowe scharakteryzowanie odchylenia standardowego.

Pole które występuję pod krzywą wskazuje nam na prawdopodobieństwo otrzymania konkretnych wartości. Np.: więc w przypadku prawostronnej asymetrii, bardziej prawdopodobne jest wystąpienie wartości mniejszych w naszej czesze. Im większa wartość tym to pole znacznie się zmniejsza – a więc również zmniejsza się szansa na wystąpienie większych wartości w danej czesze.

Średnia w przypadku prawostronnej asymetrii jest większa od mediany. A W przypadku lewostronnej – mniejsza. (Mediana dlatego jest lepsza – bo średnia jest zawyżana/zaniżana)

Kolejną miarą jest **kurtoza (Współczynnik koncentracji rozkładu)** bazuje na 4potęgach odchylenia obserwacji. Ona mówi nam o występowaniu obserwacji skrajnych – o grubości ogonów obserwacji odstających.

Rozkład normalny ma cienkie ogony.

Rozkład jednostajny - ma zanik ogonów (wynik na minusie – bo ogony są węższe niż w rozkładzie normalnym i jest mniejsze ryzyko wystąpienia obserwacji odstających niż w rozkładzie normalnym)

Rozkład Cauchy’ego ma kurtoze około 200! Ma bardzo grube ogony – mają one spore pola. Dużo wartości odstających.

Wiele testów bazuje na kurtozie. |K| < 2 lub |K| < 3 jest to test który mówi nam że jest to taka bezpieczna wartość do stosowania testów parametrycznych które zakładają normalność.

LAB

Table() – funkcja table wyłuskuje nam wartości bez powtórzeń i liczy ile danych obserwacji zanotowano.

Prop.table(table() ) – liczy nam % udziału danej wartości w ogóle

Barplot(table() , xlab=”” , ylab=”” , main””) – wykres słupkowy

Przy zmiennej ilościowej ciągłem musimy pogrupować (stworzyć przedziały) nasze obserwacje aby móc stworzyć szereg rozdzielczy.

Cut() – funkcja ta dzieli obserwacje na grupy za pomocą breaks’ów. Te Break’sy to wektor który podaje krańce tych przedziałów.

R pomaga nam stworzyć dobrze wygenerowane breaks’y za pomocą i na podstawie wygenerowanego histogramu. Funkjca hist() rysuje wykres histogramu… ale jak dodamy argumnet plot=FALSE to nie narysuje. I pobieramy właśnie z tego te breaks’y.

Y <- Hist(x, plot = FALSE)$breaks

Możemy też podać ilość breaksów ile chcemuy aby było utworzonych… ale R może się nie posłuchać.

Wykres ramkowy – funkcja boxplot – raczej dla danych ilościowych ciągłych.

Czarna linia na tym wykresie to mediana. A ten prostokąt – im jest większa wysokość pudełka – tym jest większy rozrzut/zmienność danych.

Wąsy są to odcinki które w R są łączone z ramką za pomocą przerywanej linii. Ten górny wąs to obserwacja najwyższa, a dolny wąs to obserwacja najniższa/najmniejsza. Im dłuższe są te wąsy – tym większy jest rozrzut wartości.

Jak mediana powyżej środka pudełka i wąs górny jest krótszy niż dolny to występuje asymetria lewostronna.

Kółeczka oznaczają obserwacje odstające.