Testy normalności rozkładu

Wiele testów parametrycznych wymaga, by dane pochodziły z rozkładu zbliżonego do normalnego. Dlatego testy badające normalność rozkładów są tak istotne. W testach tych zawsze przyjmuje się H0 - rozkład zmiennej jest normalny. Odrzucenie H0 jest wiec równoznaczne z przyjęciem hipotezy, że rozkład zmiennej nie jest normalny. Brak podstaw do odrzucenia nie oznacza przyjęcia hipotezy o normalności rozkładu. Musimy to jeszcze sprawdzić - w tym celu sporządzane są wykresy prawdopodobieństwo-prawdopodobieństwo.

Dla małych próbek wykonujemy dwa testy:

- *Test Kołmogorowa - Smirnowa (K-S) z poprawką Lilleforsa (K-S-L)*, która jest obliczana, gdy nie znamy średniej lub odchylenia standardowego całej populacji.
- *Test Shapiro - Wilka* (S-W) - najbardziej polecany, ale może dawać błędne wyniki dla próbek większych niż 2 tys.

Należy podkreślić, iż ostatnimi czasy test S-W stał się preferowanym testem normalności rozkładu prawdopodobieństwa ze względu na jego silną moc w porównaniu do innych dostępnych testów.

Przyjęcie hipotezy zerowej o normalności rozkładu powinno być potwierdzone graficzną reprezentacją danych np. funkcją qqplot !!!

Test Kolmogorova-Smirnova

Test K-S dla jednej próbki (K-S one-sample test) bazuje na maksymalnej różnicy pomiędzy empiryczną dystrybuantą (ECDF) a hipotetyczną dystrybuantą (CDF). Jeżeli statystyka D jest znacząca, wówczas hipoteza głosząca, że analizowany rozkład jest normalny powinna zostać odrzucona.
W wielu programach wykorzystywane wartości prawdopodobieństwa bazują na tych wyznaczonych przez Masseya (1951) - wartości te są prawidłowe w przypadku, gdy wartość średnia oraz odchylenie standardowe rozkładu normalnego są znane aproiri i nie są estymowane na podstawie zebranych danych.
Test K-S powstał w latach 30 XIX wieku. Pozwala on na realizację porównania pojedynczego rozkładu f(x) częstości z teoretycznym rozkładem g(x), lub dwóch zaobserwowanych rozkładów. W obu przypadkach realizacja testu wymaga

zdefiniowania dystrybuanty (CDF) F(x) jak i G(X) oraz wyznaczenia wielkości największej różnicy pomiędzy tymi funkcjami.

Założenia

Próbki są losowe (lub obie grupy próbek są losowe i niezależne).
Przestrzeń wartości powinna mieć przynajmniej zdefiniowany porządek (ordinal scale), a najlepiej gdyby była ciągła.
Znane są dokładne wartości (nie są estymowane z danych) średnie oraz odchylenia standardowe populacji z jakiej pochodzą próbki.

Hipoteza

H0: F(x) = G(x) (two-sided)
H1: F(x) != G(x) dla przynajmniej jednej wartości x

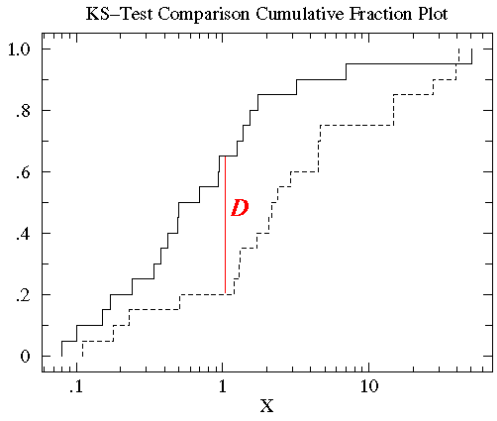
Test

Obliczana jest testowa statystyka:

D = sup[F(x) - G(x)]

Dla pojedynczej próby o rozmiarze n wartości D, Dk dla k=sqrt(n), dla zadanego poziomu istotności α wartości statystyki dane są tabelami lub obliczane programistycznie (np. z wykorzystaniem symulacji Monte Carlo).

Jedną z zalet testu K-S jest to, iż pozwala on na graficzną reprezentację danych, co pozwala użytkownikowi na swoistą weryfikację poprawności testu.



Dla dużych zbiorów danych (N>40) ze względu na CLT (centralne twierdzenie graniczne) t-test powinien pozwalać na poprawne wyniki nawet w sytuacji aberacji od rozkładu normalnego. Jednakże, w rzeczywistości rozkłady dalece nie-normalne mogą spowodować, iż t-test będzie generował omylne wyniki - nawet dla dużych zbiorów danych.

Wykres Dystrybuanty

Dystrybuanta / Empiryczna dystrybuanta pozwala na graficzną reprezentację charakteru rozproszoności zbioru danych. Przykładowo przyjmijmy zbiór controlB (posortowany od wartości najmniejszej do największej) :

> sorted controlB={0.08, 0.10, 0.15, 0.17, 0.24, 0.34, 0.38, 0.42, 0.49, 0.50, 0.70, 0.94, 0.95, 1.26, 1.37, 1.55, 1.75, 3.20, 6.98, 50.57}

Analizując dane widzimy, iż nie mamy wartości poniżej 0.08, 1/20 danych leży poniżej 0.10, 2/20 danych leżą poniżej 0.15, itd. Dla przypomnienia dla dowolnej wartości x, wartości dystrybuanty F(x) oznacza ułamek danych jakie są ściśle mniejsze od wartości x. Przykładowo F(3) dla danych controlB jest równe 17/20.

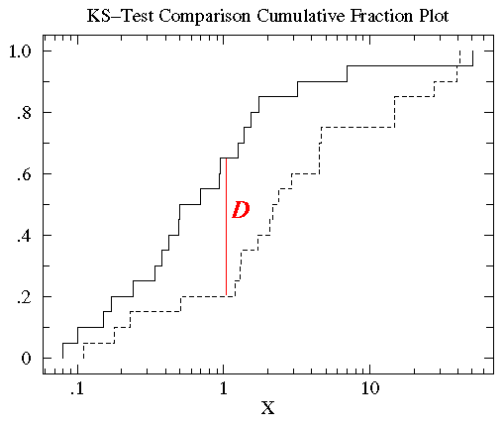
Ćwiczenie 1: Przygotuj wykres dystrybuanty dla zbioru danych controlB. Pamiętaj, że każdy punkt na wykresie odpowiada pojedynczemu punktowi w zbiorze danych.

Widać, iż większość punktów znajduje się w małym obszarze po lewej stronie wykresu, jest to swoisty znak nie-normalnego rozkładu danych. Jednakże aby lepiej zobaczyć charakterystykę danych ze zbioru controlB zmieńmy skalę osi X - tak aby więcej przestrzeni miały małe wartości.

Ćwiczenie 2: Przygotuj wykres dystrybuanty dla zbioru danych controlB z logarytmiczną skalą osi X.

Widać teraz, iż mediana przyjmuje wartość około 1.

Wracając do testu K-S. Wykorzystuje on największą wertykalną różnicę pomiędzy dwoma dystrybuantami jako statystykę D. W naszym przypadku różnica ta występuje w okolicach x=1 i wynosi około D=.45.



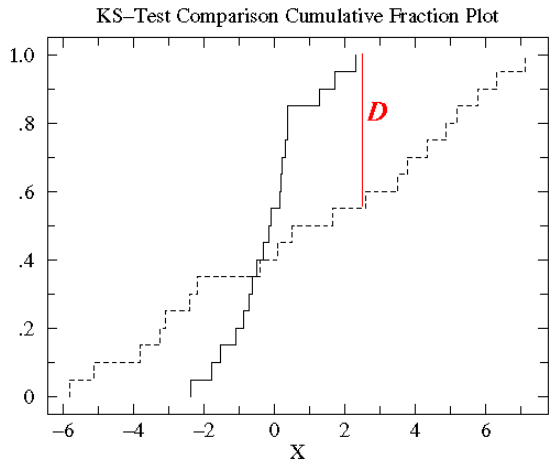
Warto zauważyć, iż w przeciwieństwie do statystyki T wartości statystyki D (stąd też wartość p) nie są podatne na zmianę skali, np. zastosowany logarytm.

Ćwiczenie 3: W analogiczny sposób przeprowadź rozumowanie dla grupy kontrolnej A oraz grupy poddanej terapii A:

controlA={0.22, -0.87, -2.39, -1.79, 0.37, -1.54, 1.28, -0.31, -0.74, 1.72, 0.38, -0.17, -0.62, -1.10, 0.30, 0.15, 2.30, 0.19, -0.50, -0.09}

treatmentA={-5.13, -2.19, -2.43, -3.83, 0.50, -3.25, 4.32, 1.63, 5.18, -0.43, 7.11, 4.87, -3.10, -5.81, 3.76, 6.31, 2.58, 0.07, 5.76, 3.50}

Sprawdź czy twój wynik zgadza się z:



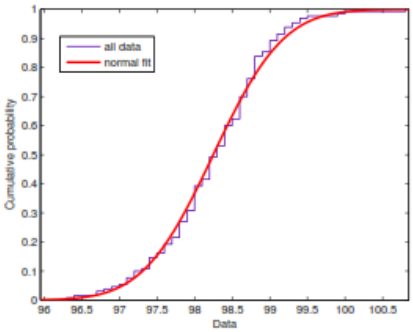
Przykład w Matlabie

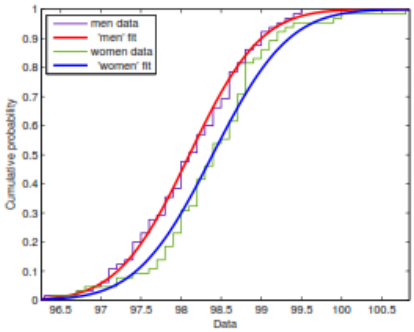
kstest: http://www.mathworks.com/help/stats/kstest.html

Załadujmy zbiory danych dotyczące temperatury ciała (*body_women* oraz *body_men*) - sumarycznie 130 obserwacji (zmienna all), każdy o 65 przykładach (zmienne odpowiednio women oraz men). Na początku spróbujmy ‘dopasować’ rozkład normalny do pełnego zbioru danych. Zacznijmy od obliczenia wartości średniej oraz odchylenia standardowego:

mean(all) = 98.2492, std(all,1) = 0.7304

Spróbujmy powtórzyć te kroki dla zbioru danych dotyczącego osobno kobiet i mężczyzn. Następnie narysujmy dystrybuantę rozkładu teoretycznego oraz empirycznego, analogicznie do poniższych wykresów:





Widzimy, iż pełny zbiór danych całkiem dobrze wpasowuje się dystrybuantę rozkładu normalnego. Sprawdźmy zatem czy podobne wnioski otrzymamy korzystając z analizy za pomocą testu K-S. W celu poprawnego wywołania testu musimy przygotować wektor wartości dystrybuanty rozkładu N(ˆμ, σˆ2):

CDFall = normcdf(all,mean(all),std(all,1))

Następnie możemy uruchomić test ‘kstest’:

[H,P,KSSTAT,CV] = kstest(all,[all,CDFall],0.05)

co da nam

H = 0, P = 0.6502, KSSTAT = 0.0639, CV = 0.1178

Zatem możemy zaakceptować H0 ponieważ p-value jest 0.6502. Dodatkowo wartość krytyczna ‘CV’ (dla odrzucenia H0 wymaga ‘KSSTAT’>’CV’) jest mniejsza od progu.

Test K-S może zostać wykorzystany do odpowiedzi na pytanie czy dane pochodzą z rozkładu normalnego, lecz także dla innych rozkładów log-normalnego, Weibulla, wykładniczego, czy logistycznego.

Zastanówmy się dalej czy oba rozkłady są takie same, zatem postawmy hipotezę H0: P1 = P2, gdzie P1 to odpowiednio rozkład dla mężczyzn, zaś P2 dla kobiet. W tym celu wykorzystamy funkcję ‘kstest2’:

[H,P,KSSTAT] = kstest2(men, women)

, która w wyniku podaje:

H = 0, P = 0.1954, KSSTAT = 0.1846.

Zatem powinniśmy zaakceptować hipotezę zerowa (p-value > 0.05 - domyślna wartość poziomu istotności). Zgodnie z przeprowadzonym testem, różnica pomiędzy obiema próbkami nie jest statystycznie znacząca aby móc stwierdzić, iż oba te rozkłady są różne.

Ćwiczenie 4: Na podstawie zbioru danych pacjenci.csv należy zbadać czy wzrost mężczyzn oraz wzrost kobiet mają takie same rozkłady? Czy obie próbki (wzrost mężczyzn i kobiet) traktowane osobno pochodzą z rozkładu normalnego?

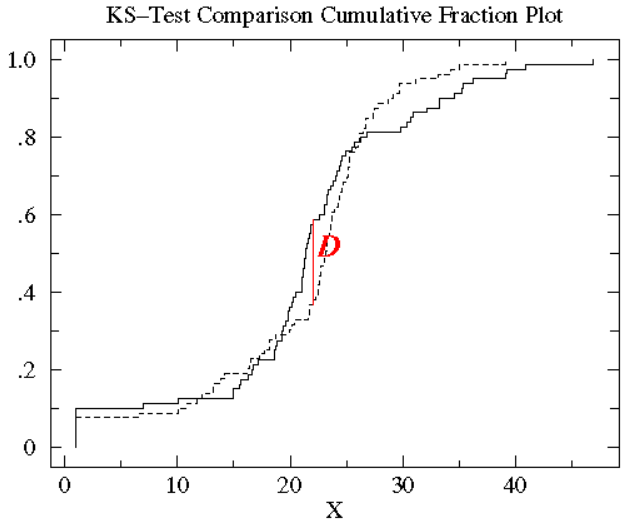
Ćwiczenie 5: Załóżmy, że pobliskie jabłonie właśnie owocują. Jedne jabłka to Delikates, zaś drugie to Renety. Zastanówmy się czy pobliskie pszczoły preferują tylko pewien gatunek jabłek. Wykorzystując stoper mierzymy czas jak długo pojedyncza pszczoła pozostaje na konkretnym drzewie - zaczynami, gdy dotyka ona drzewa, zaś kończymy, gdy odlatuje na odległość większą niż metr. Chcieliśmy zebrać równoliczne pomiary, jednakże zaczęło padać i udało nam się zebrać jedynie następujące dane:

delikates={23.4, 30.9, 18.8, 23.0, 21.4, 1, 24.6, 23.8, 24.1, 18.7, 16.3, 20.3, 14.9, 35.4, 21.6, 21.2, 21.0, 15.0, 15.6, 24.0, 34.6, 40.9, 30.7, 24.5, 16.6, 1, 21.7, 1, 23.6, 1, 25.7, 19.3, 46.9, 23.3, 21.8, 33.3, 24.9, 24.4, 1, 19.8, 17.2, 21.5, 25.5, 23.3, 18.6, 22.0, 29.8, 33.3, 1, 21.3, 18.6, 26.8, 19.4, 21.1, 21.2, 20.5, 19.8, 26.3, 39.3, 21.4, 22.6, 1, 35.3, 7.0, 19.3, 21.3, 10.1, 20.2, 1, 36.2, 16.7, 21.1, 39.1, 19.9, 32.1, 23.1, 21.8, 30.4, 19.62, 15.5}
renety={16.5, 1, 22.6, 25.3, 23.7, 1, 23.3, 23.9, 16.2, 23.0, 21.6, 10.8, 12.2, 23.6, 10.1, 24.4, 16.4, 11.7, 17.7, 34.3, 24.3, 18.7, 27.5, 25.8, 22.5, 14.2, 21.7, 1, 31.2, 13.8, 29.7, 23.1, 26.1, 25.1, 23.4, 21.7, 24.4, 13.2, 22.1, 26.7, 22.7, 1, 18.2, 28.7, 29.1, 27.4, 22.3, 13.2, 22.5, 25.0, 1, 6.6, 23.7, 23.5, 17.3, 24.6, 27.8, 29.7, 25.3, 19.9, 18.2, 26.2, 20.4, 23.3, 26.7, 26.0, 1, 25.1, 33.1, 35.0, 25.3, 23.6, 23.2, 20.2, 24.7, 22.6, 39.1, 26.5, 22.7}

Zacznijmy od sprawdzenia, czy zebrane dane pochodzą z rozkładu normalnego.

Na ostatnich zajęciach poznałeś funkcje t-test. Czy dla takich danych możemy wykorzystać t-test? Jak możemy interpretować wynik t-testu porównującego obie próbki ?

Następnie należy stworzyć wykres porównujący empiryczne dystrybuanty obu próbek. Wykorzystując test K-S proszę sprawdzić, czy dane te charakteryzują się różnymi rozkładami prawdopodobieństwa.



UWAGA Dodatkowa!

Jest kilka sytuacji w których błędem jest pełna ufność w wyniki t-testu:

1 W przypadku, gdy grupa kontrolna oraz poddana terapii nie różnią się wartością średnią, lecz różnią się w inny sposób, np:

controlA={0.22, -0.87, -2.39, -1.79, 0.37, -1.54, 1.28, -0.31, -0.74, 1.72, 0.38, -0.17, -0.62, -1.10, 0.30, 0.15, 2.30, 0.19, -0.50, -0.09}
treatmentA={-5.13, -2.19, -2.43, -3.83, 0.50, -3.25, 4.32, 1.63, 5.18, -0.43, 7.11, 4.87, -3.10, -5.81, 3.76, 6.31, 2.58, 0.07, 5.76, 3.50}

Oba zbiory danych są zbalansowane względem zera - zatem wartość średnia tych zbiorów oscyluje blisko zera. Jednakże widoczna jest znacząca wariancja w zbiorze w porównaniu do grupy kontrolnej (-6 - 6 do -2.5 - 2.5). Dane są różne jednakże t-test nie jest w stanie wychwycić różnicy

2 W przypadku, gdy grupy kontrolna oraz poddana terapii są małe (np. po 20 elementów) i różnią się wartością średnią, jednakże gdy znacząco nienormalny rozkład maskuje różnice, np:

controlB={1.26, 0.34, 0.70, 1.75, 50.57, 1.55, 0.08, 0.42, 0.50, 3.20, 0.15, 0.49, 0.95, 0.24, 1.37, 0.17, 6.98, 0.10, 0.94, 0.38}
reatmentB= {2.37, 2.16, 14.82, 1.73, 41.04, 0.23, 1.32, 2.91, 39.41, 0.11, 27.44, 4.51, 0.51, 4.50, 0.18, 14.68, 4.66, 1.30, 2.06, 1.19}

Oba te zbiory danych zostały wygenerowane z rozkładu lognormalnego i różnią się znacząco wartością średnią.

Test Lillieforsa

W teście K-S pojawia się dość restrykcyjne założenie, że znana jest dokładna wartość średnia oraz odchylenie standardowe analizowanej populacji - nie są one estymowane z danych. W przeciwnym przypadku wartości statystyki D wyznaczone przez Masseya nie są prawdziwe. Jednakże często nie mamy takiej wiedzy i wykorzystywane wartości średnie wyznaczana są na podstawie dostępnych danych wówczas test normalności przyjmuje dość skomplikowaną warunkowaną hipotezę - jak prawdopodobne jest osiągnięcie danej wartości statystyki D (lub większej), pod warunkiem, że wartość średnia oraz odchylenie standardowe zostały wyznaczone z danych. W takim przypadku należy wykorzystać tak zwane prawdopodobieństwa Lillieforsa (1967) - wartości statystyki - w celu wyznaczenia czy różnica obliczona zgodnie z KS jest statystycznie znacząca.

Założenia

Próbki są losowe.
Przestrzeń wartości powinna mieć przynajmniej zdefiniowany porządek (ordinal scale), a najlepiej gdyby była ciągła.
Nieznane są dokładne wartości (są estymowane z danych) średnie oraz odchylenia standardowe populacji z jakiej pochodzą próbki.

Hipoteza

H0: F(x) = G(x) (two-sided)
H1: F(x) != G(x) dla przynajmniej jednej wartości x

Test

Dla próbki składającej się z n obserwacji obliczana jest testowa statystyka D:

$$D = \sup[F(x) - G(x)]$$

,gdzie G(x) jest empiryczną dystrybuantą próby, zaś F(x) jest dystrybuantą rozkładu normalnego o wartości średniej równej wartości średniej próby i wariancji równej wariancji próby (nie obciążonej - z n-1 w mianowniku).
Jeżeli wartość statystyki D przekracza wartość krytyczną należy odrzucić hipotezę, iż obserwacje pochodzą z rozkładu normalnego. Wartości krytyczne wyznacza się stosując metodę symulacji Monte Carlo dla próby około 3-30 elementów i realizowane są automatycznie w nowoczesnym oprogramowaniu do analizy statystycznej.
Lillefors na bazie bogatego zbioru symulacji argumentował, iż test ten jest znaczącym udoskonaleniem testu chi^2 dla próby o małej liczbie elementów (<30) oraz, iż ma większą siłę niż test K-S.

Przykład w Matlabie

Lillietest: http://www.mathworks.com/help/stats/lillietest.html

Wykorzystujemy funkcje ‘lillietest’. Ponownie załadujmy zbiory danych dotyczące temperatury ciała (body_women oraz body_men) - sumarycznie 130 obserwacji (zmienna all), każdy o 65 przykładach (zmienne odpowiednio women oraz men).

```
[H,P,LSTAT,CV] = lillietest(all)
```

co powinno zwrócić:

H = 0, P = 0.1969, LSTAT = 0.0647, CV = 0.0777.

Zatem możemy zaakceptować normalność rozkładu danych ‘all’ z wartością p = 0.1969.

Ćwiczenie 6: Na podstawie zbioru danych pacjenci.csv należy zbadać czy wzrost mężczyzn oraz wzrost kobiet traktowane osobno pochodzą z rozkładu normalnego? Czy otrzymane wyniki różnią się od wyników dla testu K-S ?

Test Shapiro-Wilka

swtest: http://www.mathworks.com/matlabcentral/fileexchange/13964-shapiro-wilk-and-shapiro-francia-normality-tests/content/swtest.m

Test Shapiro-Wilka (W) (1968) wykorzystywany jest do testowania normalności rozkładu. Jeżeli statystyka W jest znacząca, to wówczas należy odrzucić hipotezę, iż analizowana próba pochodzi z rozkładu normalnego. Jest to preferowany test na normalność rozkładu ze względu na jego moc w porównaniu z innymi alternatywnymi testami. Niektóre programy statystyczne pozwalają na stosowanie testu W dla dużych danych (do 5000 obserwacji) przy wykorzystaniu rozszerzenia testu zaproponowanego przez Roystona (1982).
Test ten bazuje na spostrzeżeniu, iż analizując dopasowanie próbnego zbioru danych do rozkładu normalnego (jego wykresu w q-q plot) jest podobne do zadania liniowej regresji - linia diagonalna jest linią idealnego dopasowania, zaś wszystkie odchylenia od niej są podobne do residuów w zadaniu regresji. I właśnie analizując skalę tych odchyleń można określić jakość dopasowania.
Test ten, dobrze sprawdza się zarówno dla dużych, jak i małych zbiorów danych (<20).

Założenia

Próbki są losowe.
Przestrzeń wartości powinna mieć przynajmniej zdefiniowany porządek (ordinal scale), a najlepiej gdyby była ciągła.
Nieznane są dokładne wartości (są estymowane z danych) średnie oraz odchylenia standardowe populacji z jakiej pochodzą próbki.

Hipoteza

H0: F(x) ~ N(x)
H1: F(x)!~ N(x)

Test

W teście wykorzystywana jest następująca statystyka:

$$W = \frac{\sum_{i=1}^{n} a_i y_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Gdzie: yi to posortowane próbki, oraz ai to stałe wymagające weryfikacji.

Cała idea testu W polega na spostrzeżeniu, iż jeżeli faktycznie dane pochodzą z rozkładu normalnego o nieznanej wartości średniej μ oraz wariancji σ2, to wówczas powinniśmy być w stanie odwzorować te dane za pomocą prostego równania liniowego:

$$y_i = \mu + \sigma x_i,\ i = 1,2,\ldots n$$

Gdzie: xi to uporządkowany zbiór zmiennych losowych o rozkładzie N(0,1). Następnie metodą najmniejszych kwadratów możemy wyznaczyć nieznane współczynniki ai. Wektor tych wartości możemy wyznaczyć poprzez:

$$\mathbf{a}' = \frac{\mathbf{m}'\mathbf{V}^{-1}}{\sqrt{\mathbf{m}'\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m}}}$$

Gdzie: V jest macierzą co-wariancji elementów wektora x, zaś wektor m jest wektorem wartości oczekiwanych elementów x.
Statystyka W ma największą wartość równą 1, zaś najmniejszą wartość równą na12/(n-1) (stąd dla dużych wartości n>10, minimalna wartość jest w przybliżeniu równa kwadratowi najmniejszego współczynnika). Im większa wartość statystyki tym bliższe jest dopasowanie do rozkładu normalnego - jednakże nie jest to wystarczające, gdyż wysokie wartości mogą wynikać ze zbyt małych zbiorów danych, które w rzeczywistości nie są normalne. Natomiast przy odpowiednio małym W należy odrzucić hipotezę zerową.

Interpretacja

Hipotezę zerową - populacja charakteryzuje się rozkładem normalnym - należy odrzucić jeżeli wartość p jest mniejsza niż wyznaczony poziomu istotności. Wówczas możemy stwierdzić, iż dane nie są z rozkładu normalnego. W przypadku wartości p większych od wybranego poziomu istotności nie możemy odrzucić hipotezy zerowej, że dane pochodzą z rozkładu normalnego. Przykładowo dla poziomu istotności 0.05, otrzymana z danych wartość p 0.32 nie powoduje odrzucenia hipotezy, że dane pochodzą z rozkładu normalnego.

Przykład w Matlabie

W badaniu zebrano 11 losowo wybranych mężczyzn, których wzrost odnotowano. Po uporządkowaniu danych otrzymano następujące wyniki:

data = 148 154 158 160 161 162 166 170 182 195 236

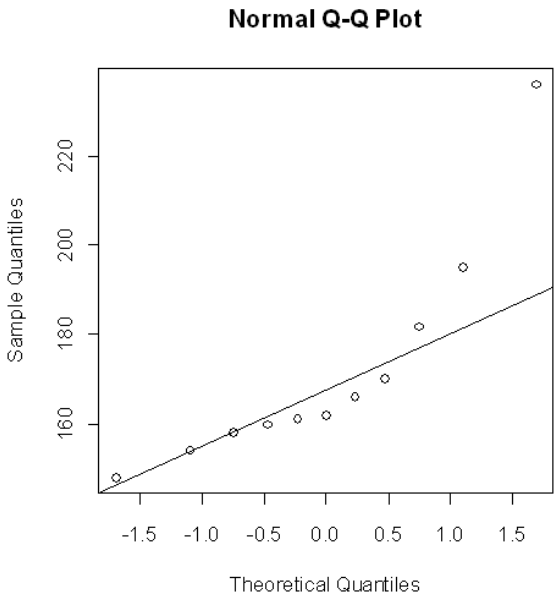
Analizując czy dane te pochodzą z rozkładu normalnego wykorzystamy test W:

[H, pValue, W] = swtest(data, alpha, 0)

Otrzymujemy jako wynik:

H=0, p=0.006704, W = 0.788840141398553

Zatem dane te raczej nie pochodzą z rozkładu normalnego. Fakt ten można również potwierdzić analizując wykres qq danych:



Ćwiczenie 7. Czy zmienna cukier z pliku *pacjenci.csv* ma rozkład normalny?

Ćwiczenie 8. Czas zużycia się 10 żarówek tego samego typu zawiera plik *zarowki.csv*. Na poziomie istotności 0,1 zweryfikuj hipotezę, że czas zużycia się żarówki ma rozkład normalny.

Ćwiczenie 9. Plik *kondensator.csv* zawiera maksymalne pojemności 40 kondensatorów (w pF). Przyjmując poziom istotności 0,05 zweryfikuj hipotezę, że rozkład pojemności kondensatorów jest normalny.

Ćwiczenie 10. Na podstawie danych z pliku *absolwenci.xsl* zbadaj, czy płace absolwentów rolnictwa i pedagogiki mają rozkłady normalne. Jakiego testu użyjesz do sprawdzenia rozkładu?