Metody planowania i analizy eksperymentów
Zadanie domowe nr 2: Zastosowanie metod wnioskowania statystycznego

Kemal Erdem - 183705

28 maja 2020

Streszczenie

W zadaniu użyty został zbiór **'Wine Quality'** (1). Zbiór zawiera dane dotyczące cech różnych win z Portugalii oraz subiektywną jakość wina odnotowaną przez osoby testujące jego smak. W zbiorze danych przetestowane zostały dwa rodzaje wina (czerwone i białe). Jakość wina została opisana w 10-cio stopniowej skali (0-10) gdzie wartość 10 oznacza idealne wino.

Spis treści

1 Opis Danych

Zbiór danych przedstawia dane dotyczące win Portugalskich z roku 2009. Składa się on z 13 cech oraz 6497 obiektów (1599 prób win czerwonych oraz 4898 prób win białych). 11 cech jest cechami numerycznymi natomiast dwie są cechami kategorycznymi. Cechy numeryczne są to cechy ilościowe ciągłe odpowiadają za opis własności wina i są to między innymi: **kwasowość** (ang. *fixed acidity*), **zawartość cukru** (ang. *residual sugar*), **pH**. Dwie cechy kategoryczne to **jakość wina** (ang. *quality*, cecha jakościowa porządkowa) oraz **typ** (ang. *type*, cecha jakościowa nominalna). Jakość wina przyjmuje wartości od 0 do 10 (z czego faktycznie występujące wartości zawierają się w $\{3, 4, 5, 6, 7, 8, 9\}$), *typ* wina posiada dwie wartości: **0** - oznaczającą wino białe, **1** - oznaczającą wino czerwone.

2 Wnioskowanie statystyczne

Do przeprowadzenia wnioskowania statystycznego zostały wykorzystane następujące narzędzia: **python**, **pandas**, **matplolib**, **scipy**. Oryginalny zbiór danych był zapisany w dwóch oddzielnych plikach które musiały zostać połączone. Wszystkie operacje dokonane na danych są dostępne w dołączonym kodzie źródłowym.

2.1 Estymacja punktowa

Estymacja punktowa została wykonana dla dwóch cech: **jakość wina** i **kwasowość** (histogram cech dostępny jest w dodatku A). Punktowa ocena średniej jakości wina wyniosła $\bar{x} = 5.8184$, wariancja $s^2 = 0.7627$, odchylenie standardowe $s = 0.8733$. Punktowa ocena średniej kwasowości wina wyniosła $\bar{x} = 7.2153$, wariancja $s^2 = 1.6807$, odchylenie standardowe $s = 1.2964$.

2.2 Estymacja przedziałowa

W przypadku naszego zbioru nie posiadamy informacji o odchyleniu standardowym rozkładów dwóch testowanych cech więc aby obliczyć końce przedziału ufności konieczne jest skorzystanie ze wzoru:

$$\bar{x} \pm t_{\alpha/2} * \frac{s}{\sqrt{n}}$$

Ponieważ zbiór którego używamy składa się z 6497 prób, musimy odczytać wartość rozkładu t-Studenta dla 6496 stopni swobody (n-1) z kolumny odpowiadającej mierze pola pod jednym z ogonów krzywej gęstości równym 0.025 ($\alpha/2$ dla **95%** przedziału ufności). Używając już wyliczonych wartości s z poprzedniego poddziału (dla cechy **jakość wina**) otrzymujemy:

$$t_{\alpha/2} * \frac{s}{\sqrt{n}} = 1.9603 * \frac{0.8733}{\sqrt{6496}} = 0.0212$$

$$[5.8184 - 0.0212; 5.8184 + 0.0212] = [5.7972; 5.8396]$$

Następnie musimy obliczyć przedział ufności dla wariancji σ^2.

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]$$

W tym przypadku musimy użyć rozkładu chi-kwadrat o 6496 stopniach swobody i $\alpha/2 = 0.025$ $\chi^2_{\alpha/2} = 6721.29$, $\chi^2_{1-\alpha/2} = 6274.50$. Na tej podstawie możemy obliczyć 95% przedział ufności dla wariancji:

$$\left[\frac{(6496)0.7627}{6721.29}; \frac{(6496)0.7627}{6274.50}\right] = [0.7371, 0.7896]$$

W taki sam sposób możemy wyznaczyć przedział ufności dla **kwasowości**, tym razem obliczymy przedział **99%** przedział ufności.

$$t_{\alpha/2} = t_{0.005} = 2.5766$$

$$t_{\alpha/2} * \frac{s}{\sqrt{n}} = 2.5766 * \frac{1.2964}{\sqrt{6496}} = 0.0414$$

$$[7.2153 - 0.0414; 7.2153 + 0.0414] = [7.1739; 7.2567]$$

i dla wariancji:

$$\left[\frac{(6496)1.6807}{6793.35}; \frac{(6496)1.6807}{6206.16}\right] = [1.6071, 1.7592]$$

2.3 Testowanie Hipotezy

Ponieważ zbiór danych na którym operujemy składa się z dwóch rodzajów win (czerwone i białe), naszym zadaniem będzie weryfikacja hipotezy mówiącej że średnia wartość oceny jakości win w obu niezależnych populacjach jest taka sama (zakładamy że cecha **jakość wina** w obu populacjach pochodzi z rozkładów normalnych):

$$\text{czerwone } n_1 = 1599, \bar{x} = 5.6360, s = 0.8076, s^2 = 0.6522$$

$$\text{białe } n_2 = 4898, \bar{x} = 5.8779, s = 0.8856, s^2 = 0.7843$$

$$H_0 : \mu_1 = \mu_2$$

W przypadku tych populacji **wariancja rozkładu nie jest znana** więc będziemy musieli przeprowadzić **test t-Studenta**. Naszą hipoteza alternatywną jest hipoteza dwustronna w postaci:

$$H_1 : \mu_1 \neq \mu_2$$

Nasza statystyka testowa wygląda następująco:

$$t_{\text{emp}} = \frac{\bar{X_1} - \bar{X_2}}{S_r}$$

$$S_r = \sqrt{S_e^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, S_e^2 = \frac{varX_1 + varX_2}{n_1 + n_2 - 2}$$

Wartość krytyczna $t(\alpha; n_1 + n_2 - 2)$ przy poziomie istotności $\alpha = 0.05$ wynosi $t(0.05; 6495) = 1.9603$. Podstawiając wartości dla naszych prób otrzymujemy:

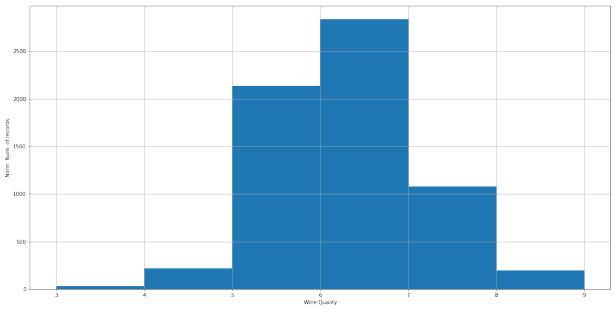
$$S_r = \sqrt{\frac{0.6522 + 0.7843}{1599 + 4898 - 2}\left(\frac{1}{1599} + \frac{1}{4898}\right)} = 0.4283 * 10^{-3}$$

$$t_{\text{emp}} = \frac{5.6360 - 5.8779}{0.4283 * 10^{-3}} = -564.7910$$

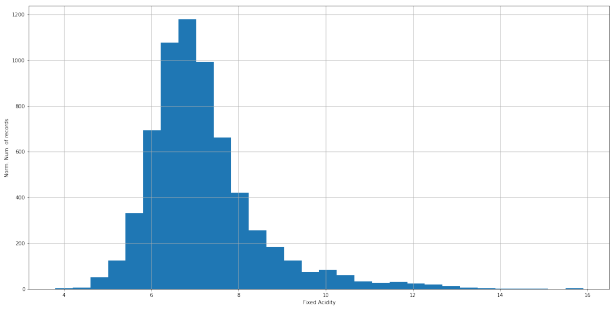
Ponieważ $|t_{emp}| > t(0.05; 6495)$ (wartość bezwzględna jest z powodu dwustronnego zbioru krytycznego C) odrzucamy hipotezę $H_0 : \mu_1 = \mu_2$ na rzecz hipotezy $H_1 : \mu_1 \neq \mu_2$. Można więc wysunąć wniosek że średnia jakość wina czerwonego nie jest taka sama jak średnia jakość wina białego. Mimo bardzo podobnych wartości $\bar{x_1}$ i $\bar{x_2}$ hipoteza jest odrzucona, jest to związane głównie z bardzo dużą próbą która zwiększa absolutną wartość t_{emp}.

A Dodatek - Histogramy cech

Histogramy cech: **jakość wina** (Rys. 1), **kwasowość** (Rys. 2)



Rysunek 1: Histogram jakości win



Rysunek 2: Histogram kwasowość

Literatura

[1] Guimarães Paulo Cortez, University of Minho. Wine quality, 2009. URL: `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`.