Analiza statystyczna *Iris flower data set*

407226, Igor Sitek, poniedziałek 14^{40}
AGH, Wydział Informatyki Elektroniki i Telekomunikacji
Rachunek prawdopodobieństwa i statystyka 2021/2022

Kraków, 26 stycznia 2022

Ja, niżej podpisany(na) własnoręcznym podpisem deklaruję, że przygotowałem(łam) przedstawiony do oceny projekt samodzielnie i żadna jego część nie jest kopią pracy innej osoby.

.....

1 Streszczenie raportu

Raport powstał w oparciu o analizę danych dotyczących szeroko znanego zestawu pomiarów kwiatów irysa dokonanych przez brytyjskiego statystyka i biologa Ronalda Fishera.

2 Opis danych

Dane pochodzą ze strony `https://gist.github.com/netj/8836201`. *Iris data set* jest zestawem bardzo małym - zawiera tylko 150 pozycji, z czego każda została opisana za pomocą 5 cech:

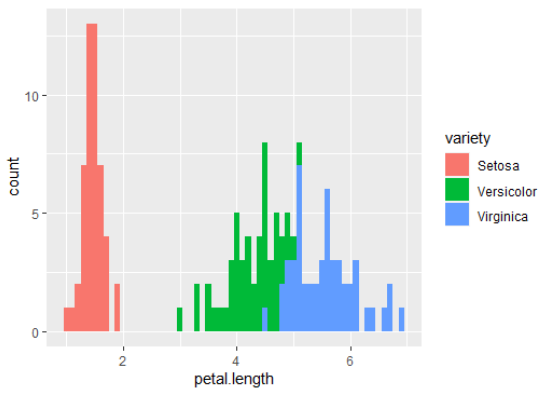
- **sepal.length** - długość kielicha kwiatu w cm,
- **sepal.width** - szerokość kielicha w cm,
- **petal.length** - długość płatka kwiatu w cm,
- **petal.width** - szerokość płatka w cm,
- **variety** - gatunek irysa (wartość typu string).

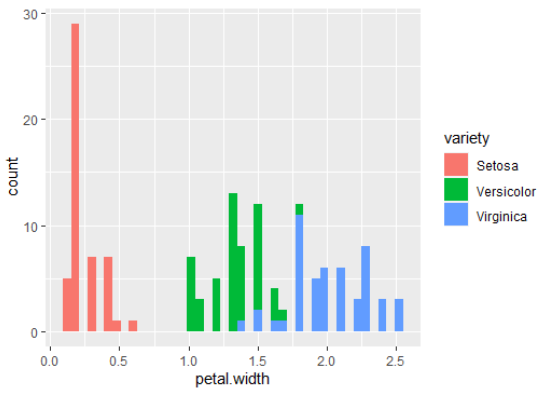
Powyższy zestaw danych jest wyczyszczony i nie wymaga żadnych modyfikacji cech (z tego powodu jest używany jako wstęp do Machine Learning'u). Można zatem od razu przejść do analizy statystycznej.

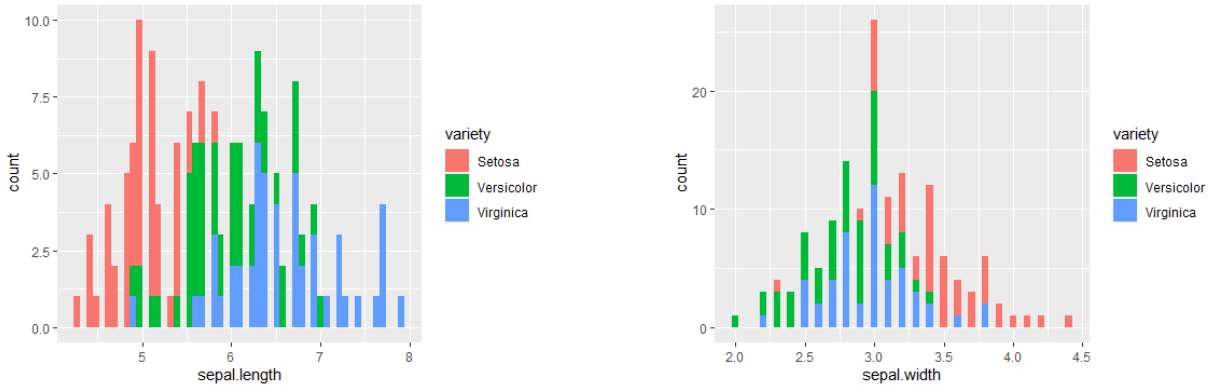
3 Analiza danych

3.1 Wydobywanie podstawowych informacji z danych

Na dobRy początek zwizualizujemy sobie rozkłady czterech zwymiarowanych cech. Dodatkowo zróżnicujemy sobie dane ze względu na gatunek irysa, dzięki czemu będziemy mieli dodatkowe informacje.







Dla każdej cechy numerycznej obliczymy też podstawowe statystyki, takie jak średnia, mediana itd.

```
 sepal.length    sepal.width     petal.length    petal.width          variety
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   Length:150
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   Class :character
Median :5.800   Median :3.000   Median :4.350   Median :1.300   Mode  :character
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

3.1.1 Parametr **petal.length** z podziałem na gatunki.

	Name	Min	Max	Mean	First_Quantile	Median	Third_Quantile	Variance	Std_Deviation	Skewness	Kurtosis
1	All	1.0	6.9	3.758	1.6	4.35	5.100	3.11627785	1.7652982	-0.2721277	1.604464
2	Setosa	1.0	1.9	1.462	1.4	1.50	1.575	0.03015918	0.1736640	0.1031751	3.804592
3	Versicolor	3.0	5.1	4.260	4.0	4.35	4.600	0.22081633	0.4699110	-0.5881587	2.925598
4	Virginica	4.5	6.9	5.552	5.1	5.55	5.875	0.30458776	0.5518947	0.5328219	2.743528

3.1.2 Parametr **petal.width** z podziałem na gatunki.

	Name	Min	Max	Mean	First_Quantile	Median	Third_Quantile	Variance	Std_Deviation	Skewness	Kurtosis
1	All	0.1	2.5	1.199333	0.3	1.3	1.8	0.58100626	0.7622377	-0.1019342	1.663933
2	Setosa	0.1	0.6	0.246000	0.2	0.2	0.3	0.01110612	0.1053856	1.2159276	4.434317
3	Versicolor	1.0	1.8	1.326000	1.2	1.3	1.5	0.03910612	0.1977527	-0.0302363	2.512167
4	Virginica	1.4	2.5	2.026000	1.8	2.0	2.3	0.07543265	0.2746501	-0.1255598	2.338652

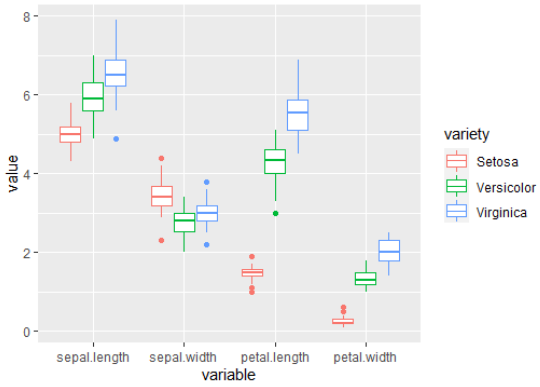
3.1.3 Parametr **sepal.length** z podziałem na gatunki.

	Name	Min	Max	Mean	First_Quantile	Median	Third_Quantile	Variance	Std_Deviation	Skewness	Kurtosis
1	All	4.3	7.9	5.843333	5.100	5.8	6.4	0.6856935	0.8280661	0.3117531	2.426432
2	Setosa	4.3	5.8	5.006000	4.800	5.0	5.2	0.1242490	0.3524897	0.1164539	2.654235
3	Versicolor	4.9	7.0	5.936000	5.600	5.9	6.3	0.2664327	0.5161711	0.1021896	2.401173
4	Virginica	4.9	7.9	6.588000	6.225	6.5	6.9	0.4043429	0.6358796	0.1144447	2.912058

3.1.4 Parametr **sepal.width** z podziałem na gatunki.

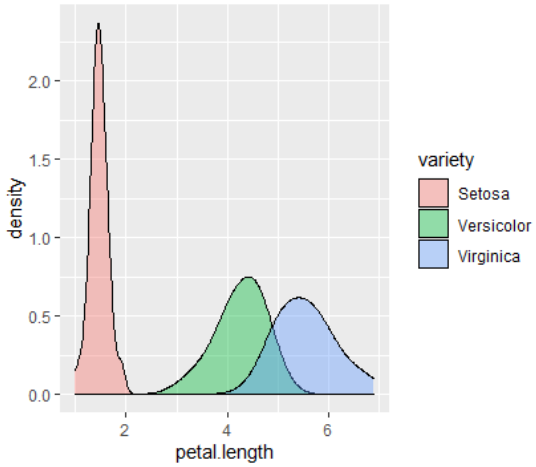
	Name	Min	Max	Mean	First_Quantile	Median	Third_Quantile	Variance	Std_Deviation	Skewness	Kurtosis
1	All	2.0	4.4	3.057333	2.800	3.0	3.300	0.18997942	0.4358663	0.31576711	3.180976
2	Setosa	2.3	4.4	3.428000	3.200	3.4	3.675	0.14368980	0.3790644	0.03992109	3.744222
3	Versicolor	2.0	3.4	2.770000	2.525	2.8	3.000	0.09846939	0.3137983	-0.35186750	2.551728
4	Virginica	2.2	3.8	2.974000	2.800	3.0	3.175	0.10400408	0.3224966	0.35487781	3.519766

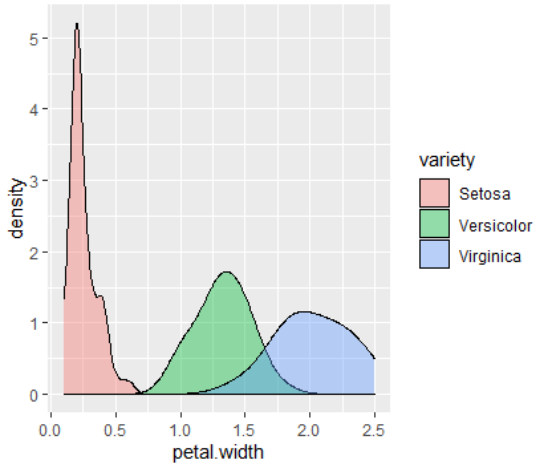
Nadszedł czas na skonstruowanie wykresów *box plot* dla każdej z badanych cech irysa, najpierw traktując wszystkie irysy łącznie, a następnie zrobimy to samo, grupując je według gatunków.

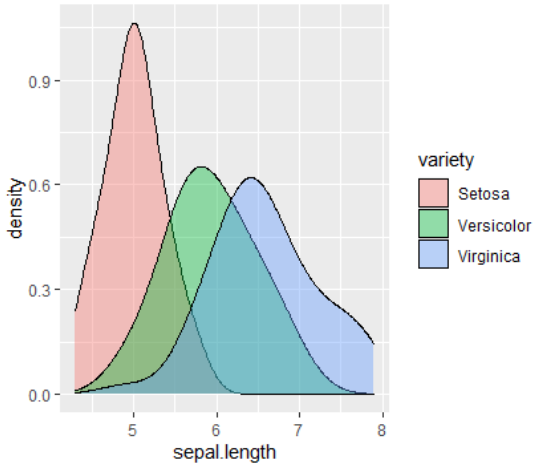


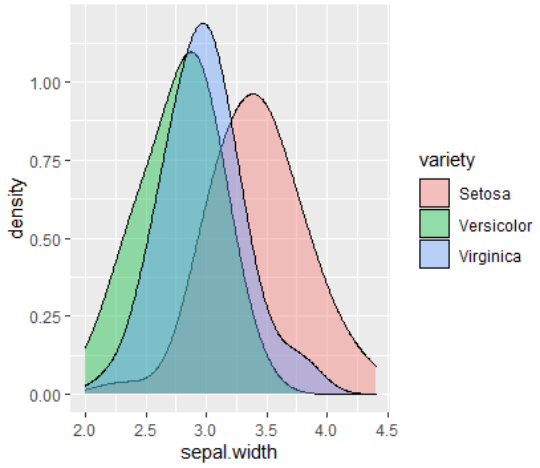
Na powyższych dwóch wykresach możemy zauważyć, że gatunek *Iris setosa* znacząco różni się od dwóch pozostałych, różnica ta jest najbardziej widoczna w długości płatka **(petal.length)**. Jest to więc argument za tym, żeby oprócz analizy całościowej przeprowadzać analizę gatunkową, gdyż pomimo równej reprezentacji (po 50 rekordów na gatunek) *Setosa* może wpływać na wyniki całościowe w większym stopniu, zaciemniając je.

Możemy też dla każdej cechy narysować gęstość jej rozkładu, tutaj również w celu uzyskania większej ilości informacji pogrupujemy dane.





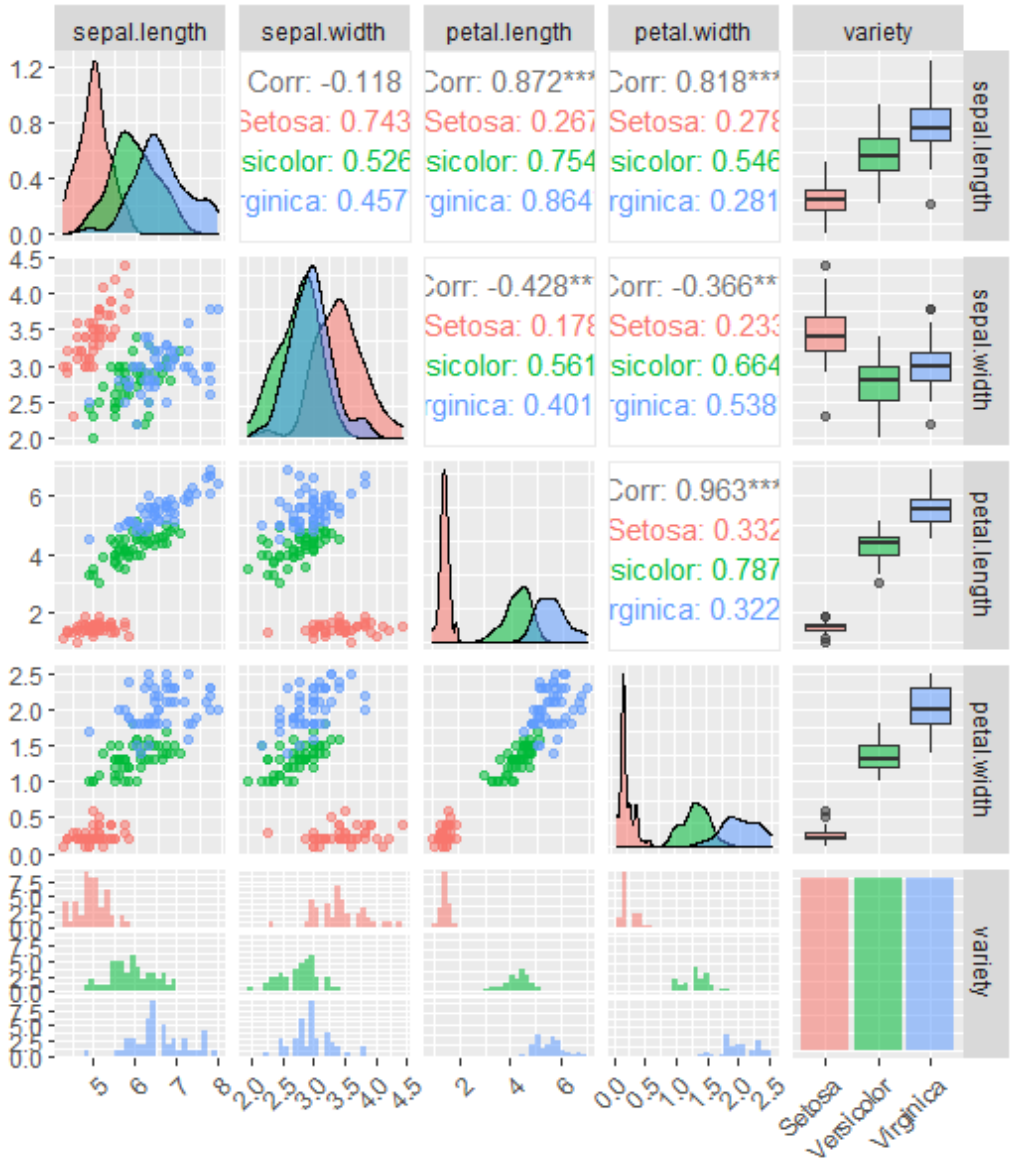




Gdyby tylko istniała funkcja, która generowała nam szybko wszystkie te wykresy, żeby nie trzeba było wszystkiego pisać ręcznie... Całe szczęście ktoś wcześniej też tak pomyślał, i dzięki prostemu wywołaniu:

```
> ggpairs(iris, aes(colour = variety, alpha = 0.4)) +
+     theme(axis.text.x = element_text(angle = 45, hjust=1))
```

Otrzymujemy taki multiwykres:



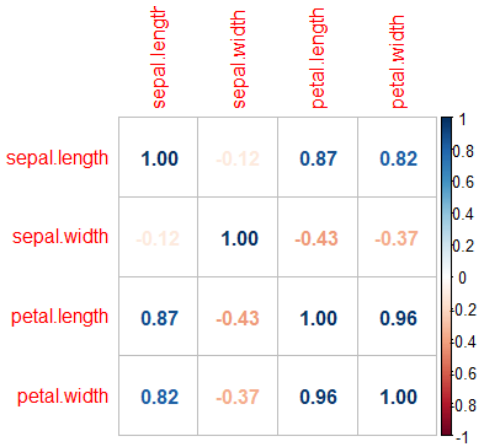
Dzięki temu uzyskujemy od razu takie wykresy jak:

- **histogram** każdej cechy z podziałem na gatunki - ostatni wiersz - co prawda grupy nie dla danej cechy nie znajdują się na jednym wykresie, tak jak my to zrobiliśmy, ale nie ma podstawy do uznania któregokolwiek podejścia jako jedynego prawidłowego
- zgrupowany **boxplot** każdej cechy - ostatnia kolumna
- wykresy **gęstości** każdej cechy, również z rozróżnieniem gatunków - przekątna główna macierzy
- wykresy **korelacji** par cech (wykresy rozrzutu) - dolny trójkąt macierzy - do tego zagadnienia przejdziemy w kolejnej sekcji

Uzyskane w ten sposób wykresy mają pewne brakujące elementy (nie w każdym wykresie mamy wartości na osi liczbowej), natomiast jako szybki sposób na zorientowanie się, z jakimi danymi mamy do czynienia, powyższa funkcja sprawdza się doskonale.

3.2 Korelacja zmiennych

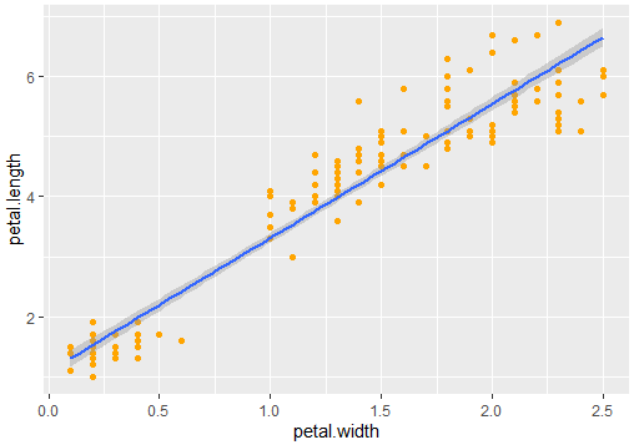
Zobaczmy, jak silnie korelują ze sobą nasze cechy, obliczając macierz korelacji - każda komórka macierzy zawiera wartość z przedziału $< -1, 1 >$, gdzie wartość bliska 1 oznacza korelację podobną do proporcjonalności wprost zmiennych, -1 - do odwrotnej proporcjonalności, a wartości około 0 sugerują brak powiązania.



Wartości 1.00 na głównej przekątnej są rzeczą oczywistą. Tak samo fakt symetryczności macierzy względem niej (przekątnej). To, co powinno wzbudzić nasz niepokój, to wartość współczynnika korelacji długości płatka od jego szerokości **(petal.width / petal.length)** - wartość 0.96 jest wartością bardzo dużą, dodatkowo - niepożądaną. Wysoki współczynnik korelacji posiada również **(petal.length / sepal.length)**, jednak nie jest ona aż tak alarmująco wysoka.

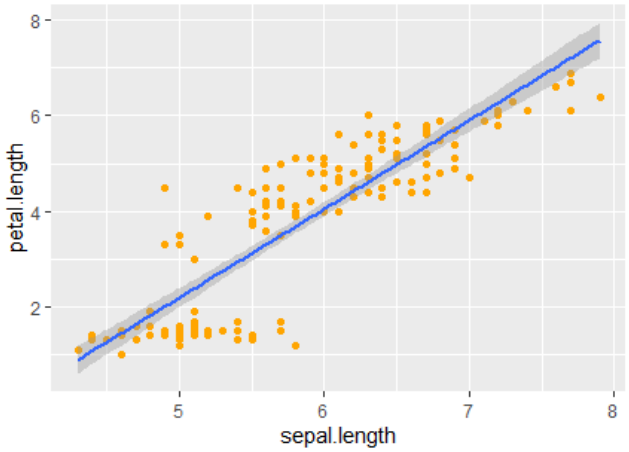
Taki wynik nie powinien nas jednocześnie dziwić - jesteśmy przywyczajeni, że rośliny w trakcie swojego wzrostu nie zmieniają kształtów swoich liści czy płatków, zatem w tym przypadku im większy irys, tym zarówno szerokość, jak i długość płatka zwiększają się w podobnym stopniu.

Tak wysoka korelacja jest jednak niepowiązana, zwłaszcza czy stosowaniu metod regresji liniowej, ponieważ zwiększamy wymiarowość problemu i czas koniecznych obliczeń (operacje komputerowe na macierzach są bardzo kosztowne), *de facto* nie wnosząc jednocześnie nowej informacji - znając jedną z tych wartości, jesteśmy z dużym prawdopodobieństwem oszacować wartość drugiej.



Jak widzimy, korzystając z regresji liniowej otrzymujemy zależność liniową, natomiast nie jest to zależność wprost proporcjonalna. Oddychamy zatem z ulgą, nie musząc przeprowadzać *feature engineering* na naszym zestawie danych, i możemy przejść do dalszej części naszej analizy.

Pro forma sprawdźmy też zależność **sepal.length / petal.length**, która również ma wysoki współczynnik korelacji.



Wykres daje nam taką samą informację jak w poprzednim przypadku - zależność liniowa, nie wprost proporcjonalna.

3.3 Estymatory przedziałowe

Przypomnijmy jeszcze raz wartości średniej oraz wariancji dla każdego parametru.

	Name	Petal.length_Mean	Petal.width_Mean	Sepal.length_Mean	Sepal.width_Mean	Petal.length_Var	Petal.width_Var	Sepal.length_Var	Sepal.width_Var
1	All	3.758	1.199333	5.843333	3.057333	3.11627785	0.58100626	0.6856935	0.18997942
2	Setosa	1.462	0.246000	5.006000	3.428000	0.03015918	0.01110612	0.1242490	0.14368980
3	Versicolor	4.260	1.326000	5.936000	2.770000	0.22081633	0.03910612	0.2664327	0.09846939
4	Virginica	5.552	2.026000	6.588000	2.974000	0.30458776	0.07543265	0.4043429	0.10400408

Nie znamy wariancji całego rozkładu, natomiast liczebność naszego rozkładu jest duża ($n > 30$), zatem przy obliczaniu przedziałów ufności możemy posłużyć się rozkładem normalnym, a nie t-Studenta. Dla każdego parametru obliczymy przedział ufności przy poziomie ufności 95%.

Przedział ufności dla wartości oczekiwanej
Skorzystamy ze wzoru:

$$P\left(\overline{X} - U_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} < m < \overline{X} + U_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Gdzie:

- $\overline{X}$ – średnia arytmetyczna zbioru
- $U_{1-\frac{\alpha}{2}}$ – wartość kwantyla $1 - \frac{\alpha}{2}$ rozkładu normalnego standardowego
- n – liczebność zbioru
- s – odchylenie standardowe próby

Aby rozwiązać ten układ nierówności, wywołamy w R poniższą komendę:

$$\overline{X} + c(-1, 1) * \frac{s}{\sqrt{n}} * qnorm(1 - \frac{\alpha}{2}),$$

wstawiając pod $\overline{X}$, s, n i α odpowiednie wartości.

Przykład wywołania dla *All* Petal.Length

$\overline{X} = 3.758$

$s = \sqrt{3.11627785} = 1.765298$

$\alpha = 0.05$

$n = 150$

```
> 3.758 + c(-1, 1)*1.765298/sqrt(150) * qnorm(0.975)
[1] 3.475499 4.040501
```

Przykład wywołania dla *Versicolor* Sepal.Length

$\overline{X} = 5.936$

$s = \sqrt{0.2664327} = 0.5161712$

$\alpha = 0.05$

$n = 50$

```
> 5.936 + c(-1, 1)*0.5161712/sqrt(50) * qnorm(0.975)
[1] 5.792927 6.079073
```

Przedział ufności dla wariancji
Skorzystamy ze wzoru:

$$P\left(\left(\frac{s}{1+\frac{U_{1-\frac{\alpha}{2}}}{\sqrt{2n}}}\right)^2 < \sigma^2 < \left(\frac{s}{1-\frac{U_{1-\frac{\alpha}{2}}}{\sqrt{2n}}}\right)^2\right) = 1-\alpha$$

Gdzie:

- $U_{1-\frac{\alpha}{2}}$ – wartość kwantyla $1-\frac{\alpha}{2}$rozkładu normalnego standardowego
- n – liczebność zbioru
- s – odchylenie standardowe próby

Aby rozwiązać ten układ nierówności, wywołamy w R poniższą komendę:

$$\frac{s^2}{\left(1+c(1,-1)*\frac{qnorm(1-\frac{\alpha}{2})}{\sqrt{2n}}\right)^2},$$

wstawiając pod s, n i α odpowiednie wartości.

Przykład wywołania dla *All* Petal.Length

$s^2 = 3.11627785$

$\alpha = 0.05$

$n = 150$

```
> 3.11627785 / (1 + c(1, -1) * qnorm(0.975) / sqrt(2 * 150)) ** 2
[1] 2.514908 3.962271
```

Przykład wywołania dla *Versicolor* Sepal.Length

$s^2 = 0.2664327$

$\alpha = 0.05$

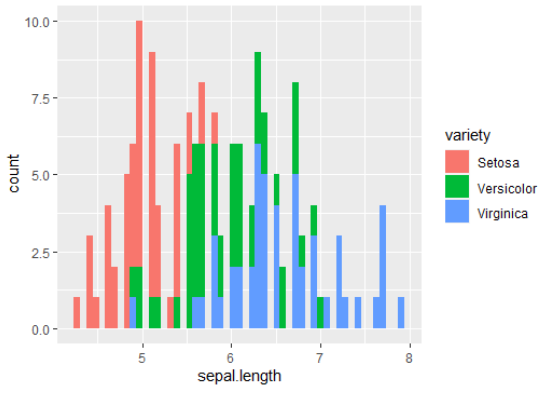
$n = 50$

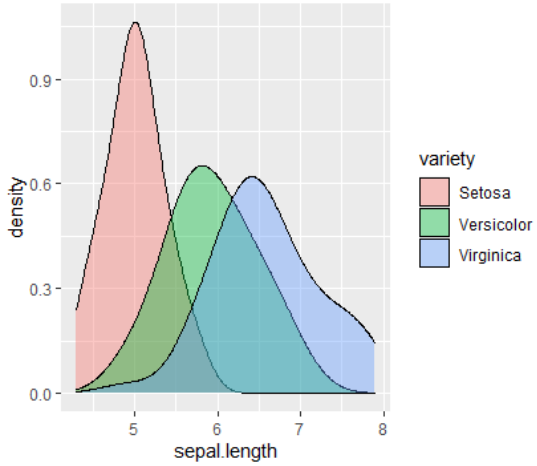
```
> 0.2664327 / (1 + c(1, -1) * qnorm(0.975) / sqrt(2 * 50)) ** 2
[1] 0.1862635 0.4121654
```

3.4 Testowanie hipotez

Ze względu na schematyczność operacji, nie będziemy testować każdego parametru wraz z grupowaniem na gatunki z osobna - zamiast tego na przykładzie **sepal.length** opiszemy procedurę działania.

Spójrzmy jeszcze raz na histogram i wykres gęstości tego parametru.

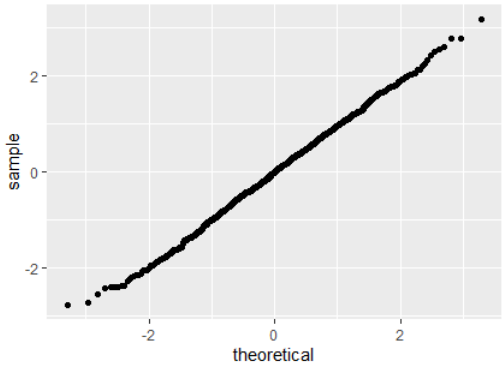




Wykresy gęstości ze względu na gatunki przypominają rozkład normalny. Żeby jednak to sprawdzić, dla wszystkich irysów, a potem także dla każdego fatunku narysujemy QQ-plot.

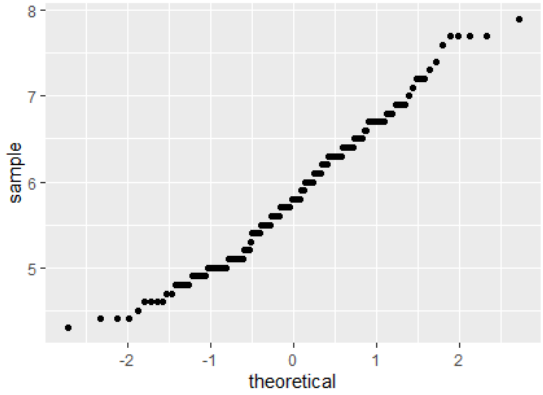
Wykres ten dla rozkładu normalnego jest linią prostą:

QQ-plot dla rozkładu normalnego

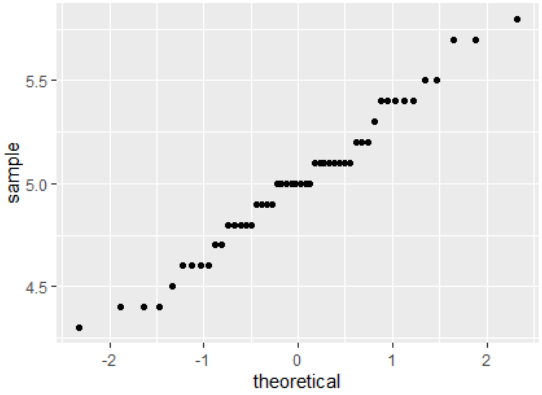


Zobaczmy teraz, jak wyglądają QQ-ploty dla naszych danych.

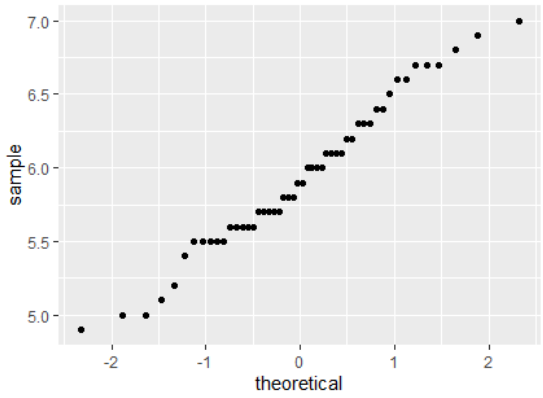
Wykres dla **sepal.length** całego zestawu *Iris*



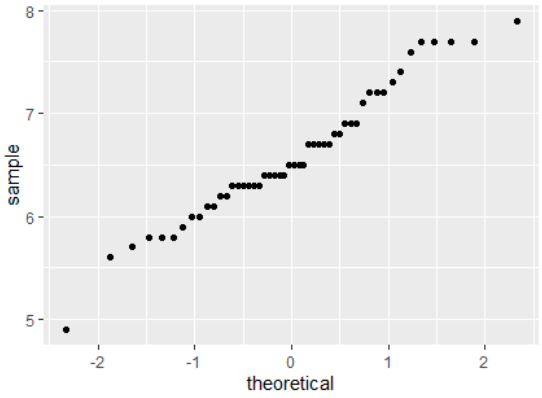
Wykres dla **sepal.length** *Setosa*



Wykres dla **sepal.length** *Versicolor*



Wykres dla **sepal.length** *Virginica*



Jak widzimy, dobrym kandydatem na rozkład normalny jest cały zestaw *Iris*. Podziały na gatunki tylko lekko przypominają linię prostą, i dla nich kończymy sprawdzanie normalności na tym etapie.

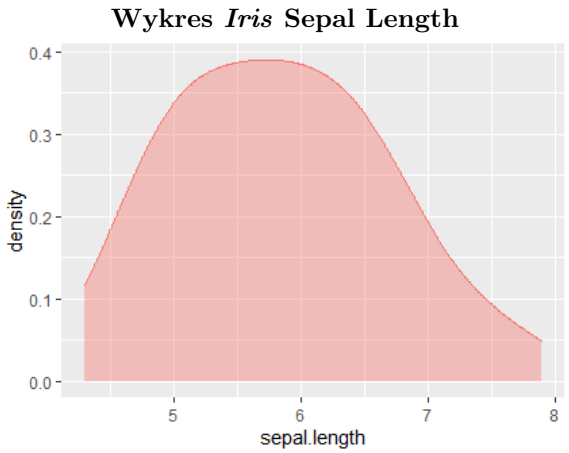
Z widocznych czterech wykresów najbliżej prostej jest rozkład całego zestawu *Iris*, więc test *Shapiro-Wilk* wykonamy właśnie dla niego. Ustalamy rozmiar próbki na 50 elementów.

```
> shapiro.test(sample(iris$sepal.length, 50))

        Shapiro-Wilk normality test

data:  sample(iris$sepal.length, 50)
W = 0.9686, p-value = 0.2028
```

Zgodnie z wytycznymi testu Shapiro-Wilka, aby rozkład był normalny, wartość p musi być większa od 0.05, a próg ten można zmieniać przy dużym W ($W > 0.98$). W naszym przypadku W nie jest aż tak duże, by próg zmienić, p jest większe od 0.05, zatem możemy uznawać rozkład **Sepal.Length** jako normalny.



Jak widzimy, wykres ten ma kształt krzywej dzwonowej, natomiast jest asymetryczny (ograniczony przez minimalne i maksymalne wartości długości kielicha).

4 Wnioski

Wnioski płynące z przeprowadzonej analizy, są następujące:

- gatunki irysa dość znacząco różnią się między soba, jeśli chodzi o wartości parametrów i cechy ich rozkładów,
- niektóre z cech irysa są ze sobą mocno skorelowane - przykładem tego może być zależność szerokości od długości płatka na poziomie korelacji 0.96,
- sprawdzono, że rozkład długości kielicha dla wszystkich irysów można traktować jako rozkład normalny, a także pokazano, jak przeprowadzić sprawdzenie dla pozostałych parametrów.