Practice 12

Elena Tuzhilina

April 4, 2023

Question 1

You are given information about 15 Titanic passengers.

	PassengerId	Sex	Age	Class	Survived
773	773	female	57	2	no
698	698	female	NA	3	yes
652	652	female	18	2	yes
548	548	male	NA	2	yes
890	890	male	26	1	yes
875	875	female	28	2	yes
392	392	male	21	3	yes
788	788	male	8	3	no
330	330	female	16	1	yes
183	183	male	9	3	no
680	680	male	36	1	yes
560	560	female	36	3	yes
104	104	male	33	3	no
136	136	male	23	2	no
37	37	male	NA	3	yes

1. Compute the contingency table for Sex and Class variables.

female	male
1	2
3	2
2	5

2. Compute marginal frequencies for Sex and Class variables.

	female	male	Sum
1	1	2	3
2	3	2	5
3	2	5	7
Sum	6	9	15

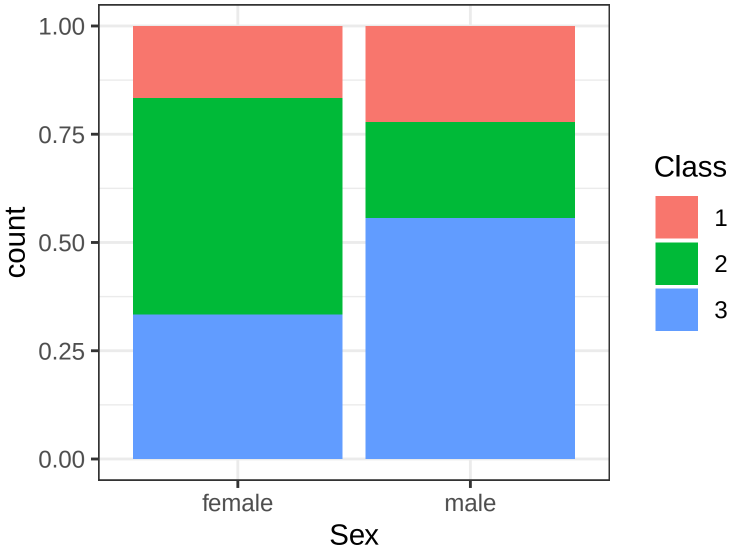
3. Compute joint probabilities for P(Sex = . . . , Class = . . .). Six probabilities in total.

Sex	Class	Frequency	Joint_Probability
female	1	1	0.0666667
female	2	3	0.2000000
female	3	2	0.1333333
male	1	2	0.1333333
male	2	2	0.1333333
male	3	5	0.3333333

4. Compute joint probabilities for P(Class = . . . |Sex = . . .). Six probabilities in total.

Sex	Class	Frequency	Joint_Probability	Conditional_Probability
female	1	1	0.0666667	0.1666667
female	2	3	0.2000000	0.5000000
female	3	2	0.1333333	0.3333333
male	1	2	0.1333333	0.2222222
male	2	2	0.1333333	0.2222222
male	3	5	0.3333333	0.5555556

5. Draw a stacked barplot. Do you think there is an association between these two variables?



Looks like there is an association.

Question 2

You are given Sex vs Class contingency table for all Titanic passengers, and you want to test these two variables for the independence.

```
tab = table(data$Class, data$Sex)
kable(tab)
```

female	male
94	122
76	108
144	347

1. State H_0 and H_a.

H_0 : Sex and Class are independent.

H_a : Sex and Class are dependent.

2. Find expected and observed counts for this table.

Sex	Class	Observed	Expected
female	1	94	76.12121
female	2	76	64.84400
female	3	144	173.03479
male	1	122	139.87879
male	2	108	119.15600
male	3	347	317.96521

3. Find the test statistic.

$\chi^2_{obs} = 16.971$

4. Find the p-value.

We use $df = (2-1)(3-1) = 2$. From the table, $p\text{-}value < 0.01$.

5. Draw the conclusion at significance level 0.01.

Since $p\text{-}value < 0.01$, we can reject null and conclude that Sex and Class variables are dependent.

```
chisq.test(x = data$Sex, y = data$Class, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data$Sex and data$Class
## X-squared = 16.971, df = 2, p-value = 0.0002064
```

Question 3

The iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris (150 flowers in total). The species are Iris setosa, versicolor, and virginica.

Here are summary statistics for Sepal and Petal lengths.

```
mean(Sepal.Length)
```

```
## [1] 5.843333
```

```
mean(Petal.Length)
```

```
## [1] 3.758
```

```
sd(Sepal.Length)
```

```
## [1] 0.8280661
```

```
sd(Petal.Length)
```

```
## [1] 1.765298
```

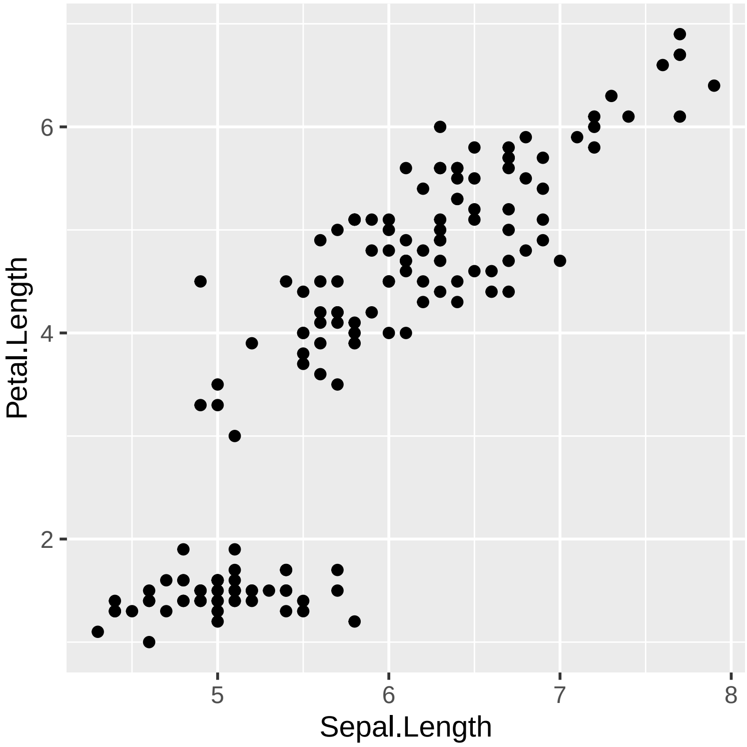
```
cor(Sepal.Length, Petal.Length)
```

```
## [1] 0.8717538
```

1. Do you think there is an association between Sepal and Petal lengths?

Probably, as the correlation is positive and close to 1.

2. You want to fit the regression line to the following scatterplot plot.



State the regression line equation.

$Petal.Length = a \cdot Sepal.Length + b$

3. Find regression coefficients.

$a = 1.858,\ b = -7.101$

4. What is the interpretation of the regression coefficients?

slope a: if Sepal length increases by one unit, Petal length will increase by 1.858

intercept b: Petal length is -7.101 for the flowers with zero Sepal length (does not make much sense in this context).

5. Check if point $Sepal.Length = 6$ and $Petal.Length = 3$ lies on the regression line.

The predicted value is $\hat{y}_i = 1.858 \cdot 6 - 7.101 = 4.047$.

It is not equal to 3, thus the point does not lie on the regression line.

6. Find the residual for point $Sepal.Length = 6$ and $Petal.Length = 3$. Does this point lie below or above the regression line?

$e_i = y_i - \hat{y}_i = 3 - 4.047 = -1.047$

The residual is negative, thus the point lies below the line.

7. Check that $Sepal.Length = \bar{x}$ and $Petal.Length = \bar{y}$ point lies on the regression line.

$1.858 \cdot 5.84 - 7.101 = 3.758$

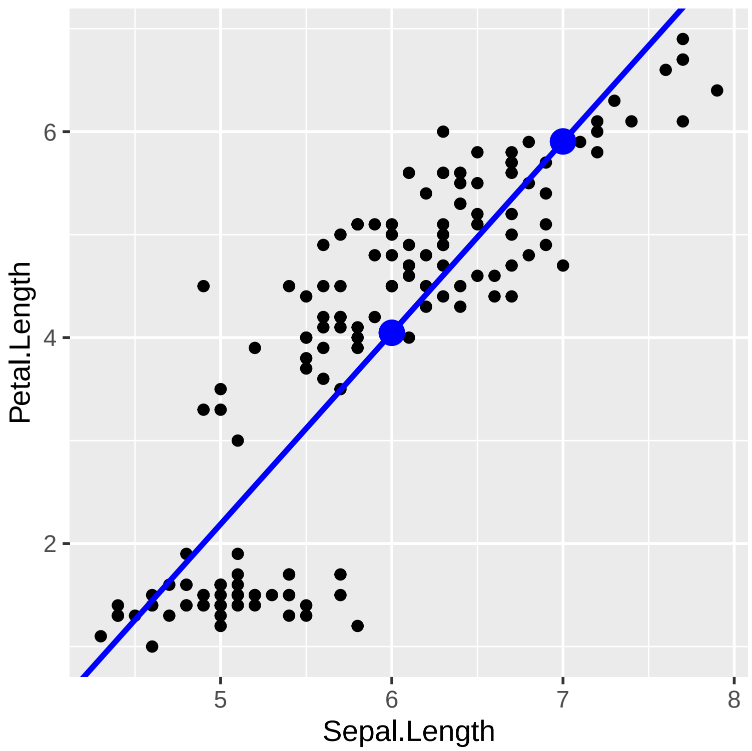
8. Use the regression line to predict the value of Petal length if Sepal length is 6 and 7.

$1.858 \cdot 6 - 7.101 = 4.047$

$1.858 \cdot 7 - 7.101 = 5.905$

9. Add the regression line to the scatterplot.

The regression line will pass through the points from the previous part.



10. Find TSS from the provided information.

From the sample standard deviation we find $TSS = (n-1) \cdot s_y^2 = 149 \cdot 1.76^2 = 461.5$

11. Find R^2 from the provided information. Do you think linear model fits the data well?

From the sample correlation we find $R^2 = r_{xy}^2 = 0.872^2 = 0.76$.
Yes, R^2 if close to 1.

12. Find RSS from the provided information.

From $R^2 = 1 - \frac{RSS}{TSS}$ we find $RSS = (1 - R^2) \cdot TSS = (1 - 0.76) \cdot 461.5 = 111$

13. Find ESS from the provided information.

$ESS = TSS - RSS = 461.5 - 111 = 350.5$

```
lm(Petal.Length~Sepal.Length)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length)
##
## Coefficients:
##  (Intercept)  Sepal.Length
##       -7.101         1.858
```