The University of Sydney
School of Mathematics and Statistics

Solutions to Tutorial Week 3

MATH1905: Statistics (Advanced) Semester 2, 2017

Web Page: `http://sydney.edu.au/science/maths/MATH1905`
Lecturer: Michael Stewart

Please ask your tutor about any difficulties from week 2.

1. Evaluate the correlation coefficient for the following data set:

x_i:	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4
y_i:	10	15	30	35	25	30	50	45

Using your calculator, the value of r^2 is (2dp):

(a) 0.88 (b) 0.99 (c) 0.77 (d) 0.23 (e) none of the above

Solution: It is straightforward to compute the summary statistics

$$\begin{aligned} \sum_i x_i &= 10.8 & \sum_i y_i &= 240 \\ \sum_i x_i^2 &= 18.36 & \sum_i y_i^2 &= 8500 \\ \sum_i x_i y_i &= 385.5 & n &= 8 \end{aligned}$$

So we have

$$\begin{aligned} S_{xx} &= 18.36 - \frac{10.8^2}{8} = \frac{189}{50} = 3.78\,, \\ S_{yy} &= 8500 - \frac{240^2}{8} = \frac{189}{50} = 1300\,, \end{aligned}$$

and

$$S_{xy} = 385.5 - \frac{10.8 \times 240}{8} = 61.5$$

so the correlation coefficient is given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{61.5}{\sqrt{3.78 \times 1300}} \approx 0.8773\,.$$

But note that we are asked for r^2, which is

$$\frac{61.5^2}{3.78 \times 1300} \approx 0.7697\,.$$

So the correct answer is (c).

2. Evaluate the correlation coefficient for the following data set, both by hand and then check it with R:

x_i:	5	3	10	1
y_i:	2	1	5	0

Solution: Recall from last week that we have

$$
\begin{aligned}
\sum_i x_i &= 19 & \sum_i y_i &= 8 \\
\sum_i x_i^2 &= 135 & \sum_i y_i^2 &= 30 \\
\sum_i x_i y_i &= 63 & n &= 4
\end{aligned}
$$

and

$$
\begin{aligned}
S_{xx} &= 135 - \frac{19^2}{4} = 179/4 \\
S_{xy} &= 63 - \frac{19 \times 8}{4} = 25\,.
\end{aligned}
$$

In addition we have

$$S_{yy} = 30 - \frac{8^2}{4} = 14$$

and so the correlation coefficient is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{25}{\sqrt{14 \times 179/4}} \approx 0.9988\,.$$

```
x=c(5,3,10,1)
y=c(2,1,5,0)
cor(x,y)
```

```
[1] 0.9988022
```

Note also that we can recover the correlation coefficient from the least-squares slope and the sd's:

```
fit=lm(y~x)
coef(fit)
coef(fit)[2]
coef(fit)[2]/(sd(y)/sd(x))
```

```
(Intercept)           x
 -0.6536313   0.5586592
        x
0.5586592
        x
0.9988022
```

3. In R Type

- `data(swiss)` to obtain the `swiss` data set;
- `attach(swiss)` to obtain the 6 variables from the data frame;
- `help(swiss)` and read information about this data set.

(a) Type `cor(swiss)` to obtain the matrix of pairwise correlations. What are the 3 most correlated pairs?

Solution:

```
data(swiss)
attach(swiss)
```

```
cor(swiss)
```

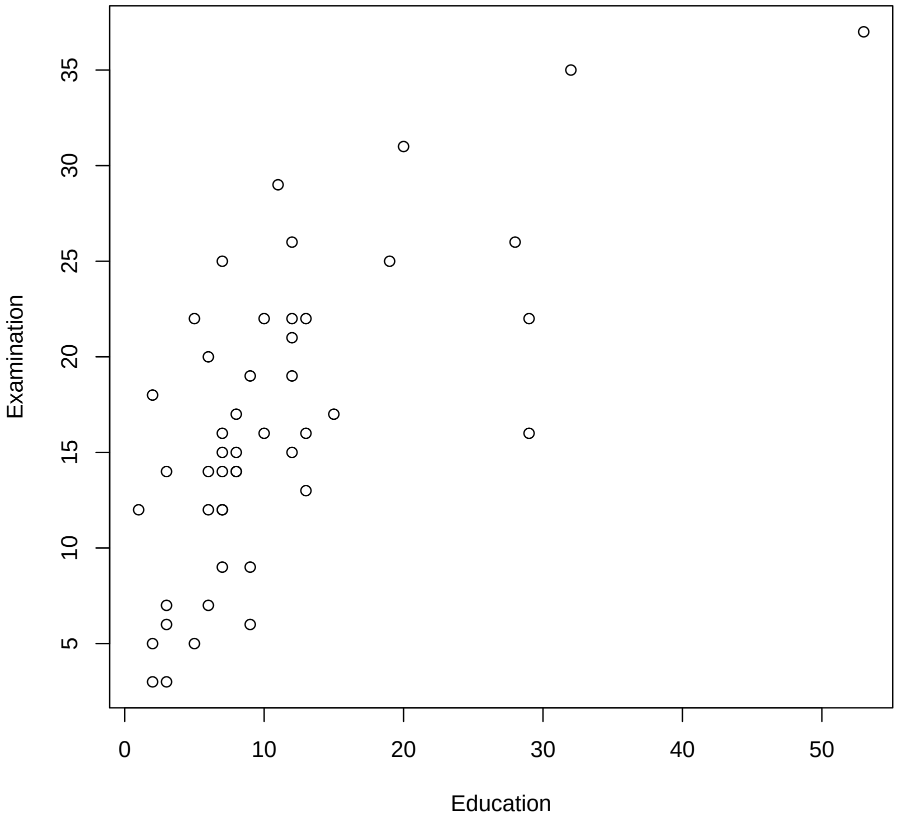
```
                  Fertility Agriculture Examination   Education   Catholic
Fertility         1.0000000  0.35307918  -0.6458827 -0.66378886  0.4636847
Agriculture       0.3530792  1.00000000  -0.6865422 -0.63952252  0.4010951
Examination      -0.6458827 -0.68654221   1.0000000  0.69841530 -0.5727418
Education        -0.6637889 -0.63952252   0.6984153  1.00000000 -0.1538589
Catholic          0.4636847  0.40109505  -0.5727418 -0.15385892  1.0000000
Infant.Mortality  0.4165560 -0.06085861  -0.1140216 -0.09932185  0.1754959
                 Infant.Mortality
Fertility              0.41655603
Agriculture           -0.06085861
Examination           -0.11402160
Education             -0.09932185
Catholic               0.17549591
Infant.Mortality       1.00000000
```

The 3 largest correlations are: (Education,Examination) with $r = 0.69$, (Agriculture,Examination) with $r = -0.68$, and (Education,Fertility) with $r = -0.66$.

(b) Produce (separate) scatter plots: `plot(Education,Examination)`, `plot(Education,Fertility)`, `plot(Agriculture,Examination)`, `plot(Catholic,Fertility)`. Do you see any pattern? If yes does it agree with the corresponding correlation. What do we learn from this data analysis?

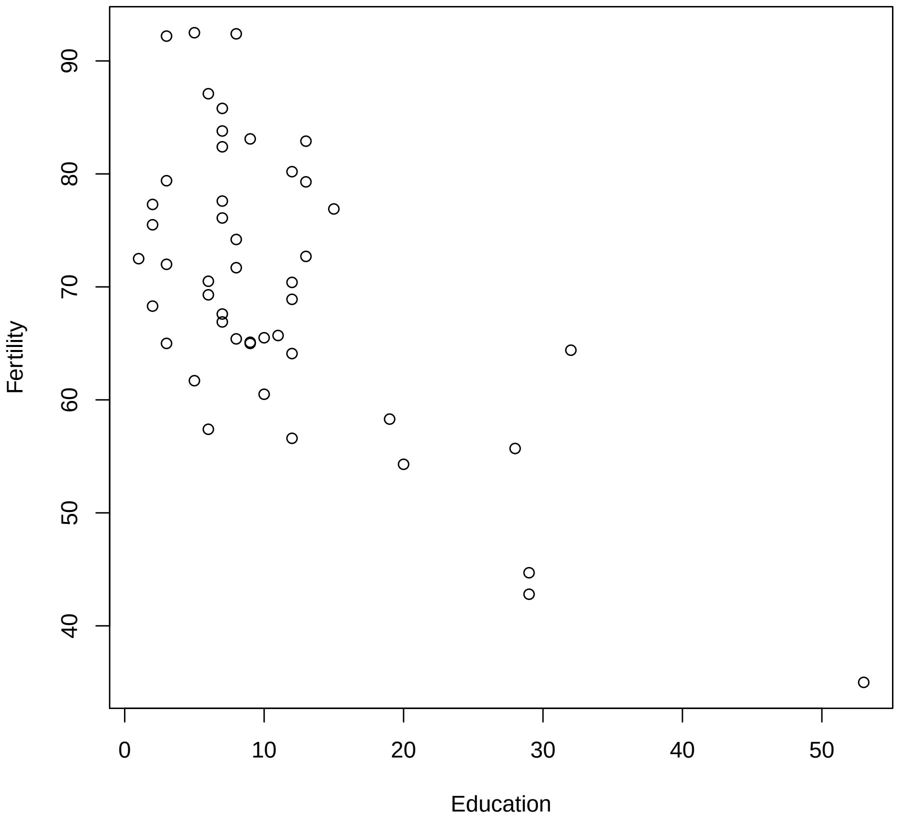
Solution:

```
plot(Education,Examination)
```



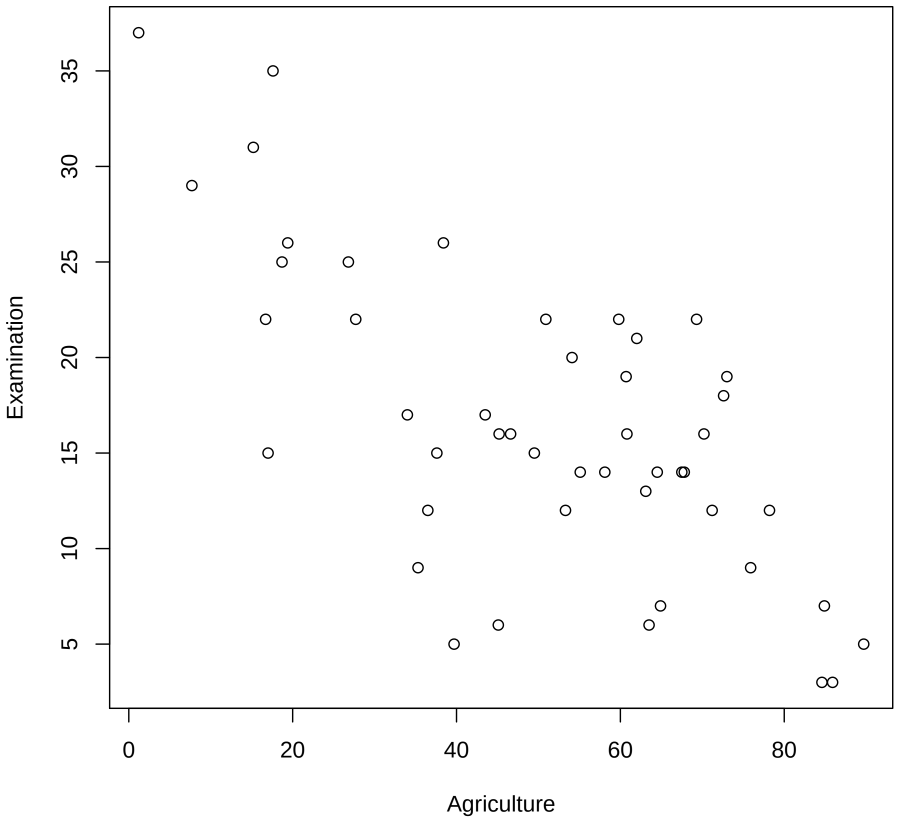
Overall (positive) linear pattern, medium strength association.

```
plot(Education,Fertility)
```



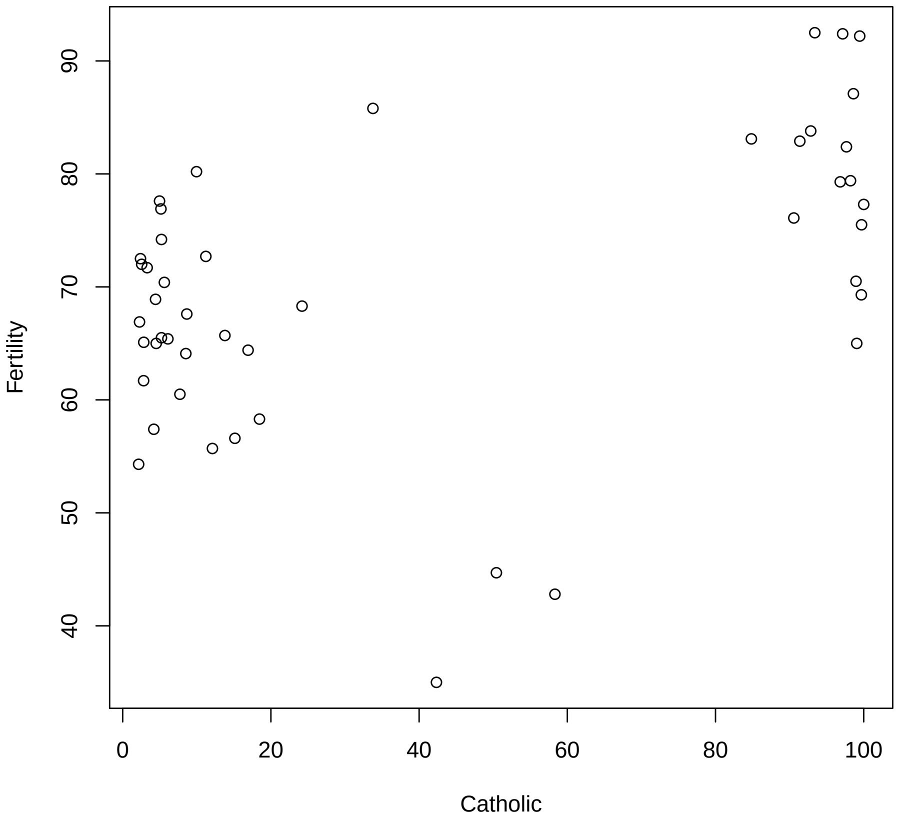
Overall (negative) linear pattern, medium strength association.

```
plot(Agriculture,Examination)
```



Overall (negative) linear pattern, medium strength association.

```
plot(Catholic,Fertility)
```



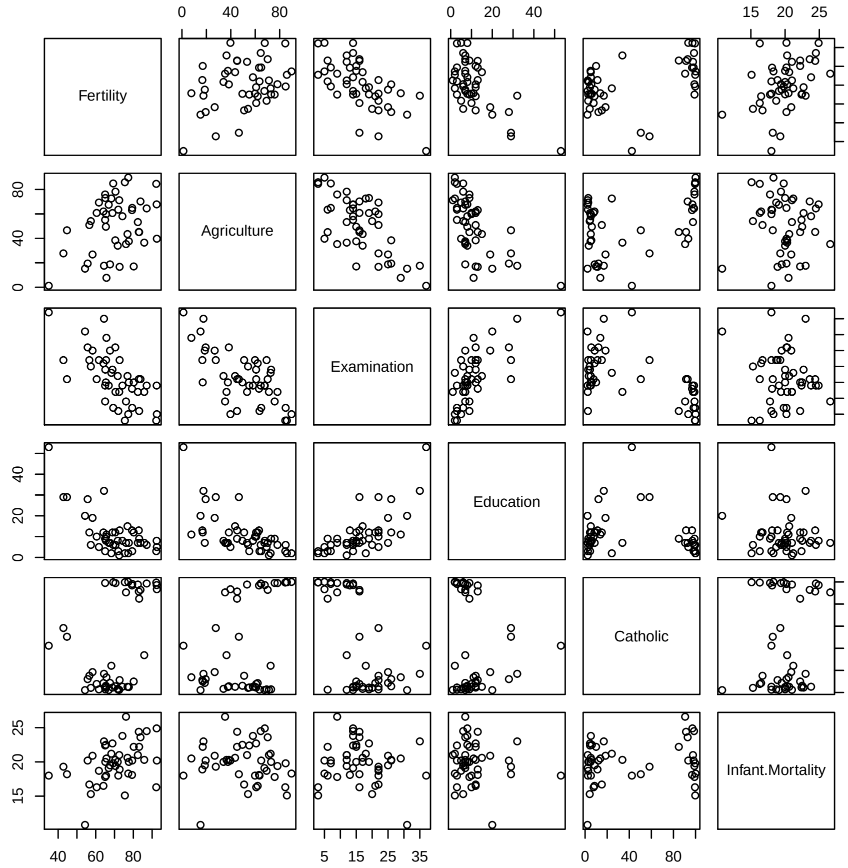
We see 2 clusters of data points, on the LHS provinces with low percentage of Catholic which tend to have lower Fertility index than provinces on the RHS where the percentage of catholic is higher. The general trend (between those 2 clusters) could be linear however there are 3 outliers (provinces) in the bottom of the plot where there are about 50percent of Catholic and a v.low fertility index. This could explain the relatively low value of $r = 0.45$ despite a visible linear trend. Remember that r is very sensitive to outliers. (In reading a scatter plot: we define an outlier as any pair (x_i, y_i) which (extremely) deviates from the overall pattern.)

The data analysis suggests that in French-speaking Switzerland at about 1888: an educated person was performing better at an exam (here the army test), people coming from rural areas (farming) had less education $r = -0.63$ and performed badly at exams ($r = -0.66$). Educated people had less children $r = -0.66$ and catholic regions tend to have more children (apart from 3 outliers, see above).

(c) Type `pairs(swiss)` to obtain all the paired scatter plots. Comment on the plots as well as on the pairwise correlations.

Solution:

```
pairs(swiss)
```



A look at the others scatter plots and correlations shows that the Infant mortality variable is only correlated to the Fertility variable (as expected since these regions have more births). The catholic variable is correlated to agriculture in a similar fashion as (Catholic,Fertility), see above; the other scatter plots for Catholic show some association which may not be linear and the results are harder to interpret.

4. Suppose $a = \bar{y} - b\bar{x}$ and $b = S_{xy}/S_{xx}$ (using the usual notation) are the least-squares intercept and slope associated with the points $(x_1, y_1), \dots, (x_n, y_n)$. Writing $\hat{y}_i = a + bx_i$ for the i-th fitted value and $\hat{\varepsilon}_i = y_i - \hat{y}_i$ for the i-th residual, show that $\sum_{i=1}^n (\hat{y}_i - \bar{y})\hat{\varepsilon}_i = 0$ and hence that $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2$.

Solution: Note firstly that we can write

$$\hat{y}_i = a + bx_i = (\bar{y} - b\bar{x}) + bx_i = \bar{y} + b(x_i - \bar{x}) \tag{1}$$

and thus

$$\hat{y}_i - \bar{y} = b(x_i - \bar{x})\,. \tag{2}$$

Using (1) we can also write

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - [\bar{y} + b(x_i - \bar{x})] = (y_i - \bar{y}) - b(x_i - \bar{x})\,. \tag{3}$$

Using both (2) and (3) we have

$$\begin{aligned}\sum_{i=1}^n (\hat{y}_i - \bar{y})\hat{\varepsilon}_i &= \sum_{i=1}^n b(x_i - \bar{x})\,[(y_i - \bar{y}) - b(x_i - \bar{x})] \\ &= b\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - b^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= bS_{xy} - b^2 S_{xx}\,.\end{aligned}$$

This is seen to equal zero when we substitute in $b = S_{xy}/S_{xx}$.

From here, note that since we can write

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y} = (\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i\,,$$

we have that

$$\begin{aligned}(S_{yy} =)\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}\left[(\hat{y}_i-\bar{y})+\hat{\varepsilon}_i\right]^2 &= \sum_{i=1}^{n}\left[(\hat{y}_i-\bar{y})^2+2(\hat{y}_i-\bar{y})\hat{\varepsilon}_i+\hat{\varepsilon}_i^2\right]\\ &= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})\hat{\varepsilon}_i+\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \qquad (4)\\ &= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2\end{aligned}$$

since as shown above the middle sum in (4) is zero.

This shows that we can decompose the “Total sum of squares” S_{yy} into the sum of two other sums of squares: the “Regression sum of squares” and the “Residual sum of squares”.

5. Using the fact that for events A, B and C, $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$, apply the general addition rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad (5)$$

repeatedly to prove that

$$P(A \cup B \cup C) = \ P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)\,.$$

Solution: Applying the general addition rule to the two events $(A \cup B)$ and C we get

$$P(A \cup B \cup C) = P\{(A \cup B) \cup C\} = P(A \cup B) + P(C) - P\{(A \cup B) \cap C\}\,. \qquad (6)$$

Next note that

$$P\{(A \cup B) \cap C\} = P\{(A \cap C) \cup (B \cap C)\} = P(A \cap C) + P(B \cap C) - P(A \cap B \cap C) \qquad (7)$$

after applying the general addition rule to the two events $(A \cap C)$ and $(B \cap C)$. The desired result follows after substituting (5) above and (7) into (6).

6. Two six-sided dice (of different colours) are rolled in such a way that all possible sequences of pairs of values are equally likely to show facing up when the dice come to rest. Let A be the event that a total of strictly less than 4 occurs. Its probability, $P(A)$ is:

(a) 1/6 (b) 3/6 (c) 9/36 (d) 1/36 (e) 3/36

Solution: There are 36 possible outcomes:

$$\begin{matrix} (1,1) & (1,2) & \cdots & (1,6)\\ (2,1) & (2,2) & \cdots & (2,6)\\ \vdots & \vdots & \ddots & \vdots\\ (6,1) & (6,2) & \cdots & (6,6)\end{matrix}$$

Of these, only $(1,1)$, $(1,2)$ and $(2,1)$ constitute the event A. If all 36 possible outcomes are equally likely then $P(A) = 3/36$ so (e) is the correct answer.

7. In the setting of the previous question let B be the event that the total showing is divisible by 3.

(a) Write down the event B.

Solution: The event B is the same as getting a total of 3, 6, 9 or 12:

$$\{\underbrace{(1,2),(2,1)}_{\text{total is 3}},\underbrace{(1,5),(2,4),(3,3),(4,2),(5,1)}_{\text{total is 6}},\underbrace{(3,6),(4,5),(5,4),(6,3)}_{\text{total is 9}},(6,6)\}\,.$$

(b) Determine the conditional probability $P(A|B)$.

Solution: With A as in the previous question we have that $A \cap B = \{(1,2),(2,1)\}$. If all 36 possible outcomes are equally likely then $P(B) = 12/36$ and $P(A \cap B) = 2/36$. By definition

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/36}{12/36} = \frac{1}{6}$$

since B has 12 outcomes.

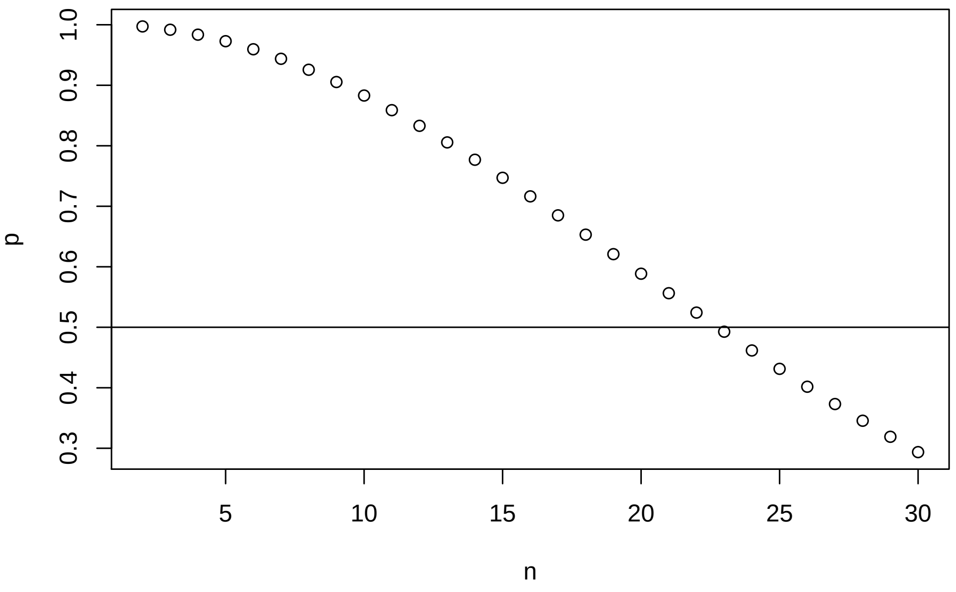
8. Suppose that for a group of n students it is known that none of them were born in a leap year. The students line up alphabetically and write their birth date (ignoring the year) in order on a whiteboard. Assuming each possible sequence of n birth dates is equally likely, write an expression (as a function of n) giving the probability that all birth dates are different. Plot this function using R for $n = 2, 3, \ldots, 30$ (**hint**: use the functions `choose()` and `factorial()`). What is the smallest n so that this probability is less than 0.5?

Solution: We may represent the sample space as all sequences of length n with each position occupied by a number from $\{1, 2, \ldots, 365\}$ and thus there are 365^n possible outcomes. The desired event consists of all those sequences with no repetitions, which has ${}^{365}P_n = 365 \times 364 \times \cdots \times (365 - n + 1) = 365!/(365-n)!$ outcomes in it. Thus the desired probability can be expressed as

$$\frac{365!}{(365-n)!365^n} = \frac{n!\binom{365}{n}}{365^n};$$

this last form is convenient for evaluation in R.

```
n=2:30
n
p=factorial(n)*choose(365,n)/(365^n)
plot(n,p)
abline(h=0.5)
```



The smallest `n` such that this is less that 0.5 is 23.