Universität Bielefeld
Tanja Pasurek
Tutor: Ehsan Abedi WiSe 2023/24

Foundations of Statistics
Solutions to Homework 4

Random variables. Expectation and variance.

Exercise 1 (Indicator random variable).
Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Given an event $A \in \mathcal{A}$, define the indicator random variable

$$\mathbb{I}_A(\omega) := \begin{cases} 1, & \text{if } A \text{ occurs} \;\; (\text{i.e. } \omega \in A), \\ 0, & \text{if } A \text{ does not occur} \;\; (\text{i.e. } \omega \notin A). \end{cases}$$

(a) Prove that for any $A, B \in \mathcal{A}$

$$\mathbb{I}_A^2 = \mathbb{I}_A, \quad \mathbb{I}_{A\cap B} = \mathbb{I}_A \mathbb{I}_B, \quad \mathbb{I}_{A\cup B} = \mathbb{I}_A + \mathbb{I}_B - \mathbb{I}_A\mathbb{I}_B.$$

Solution: We have

$$\mathbb{I}_A^2(\omega) = \begin{cases} 1^2, & \omega \in A, \\ 0^2, & \omega \notin A. \end{cases} = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases} =: \mathbb{I}_A(\omega).$$

To prove the second statement, we start from the right-hand side and write

$$\begin{aligned}
\mathbb{I}_A(\omega)\,\mathbb{I}_B(\omega) &= \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases} \times \begin{cases} 1, & \omega \in B, \\ 0, & \omega \notin B. \end{cases} \\
&= \begin{cases} 1, & \omega \in A \text{ and } \omega \in B, \\ 0, & \omega \in A \text{ and } \omega \notin B, \\ 0, & \omega \notin A \text{ and } \omega \in B, \\ 0, & \omega \notin A \text{ and } \omega \notin B. \end{cases} \\
&= \begin{cases} 1, & \omega \in A \text{ and } \omega \in B, \\ 0, & \text{otherwise.} \end{cases} \\
&= \begin{cases} 1, & \omega \in A \cap B, \\ 0, & \omega \notin A \cap B. \end{cases} \\
&=: \mathbb{I}_{A\cap B}(\omega).
\end{aligned}$$

To prove the third statement, we spell out both sides of the equation:

$$
\begin{aligned}
\mathbb{I}_{A\cup B}(\omega) := &\begin{cases} 1, & \omega \in A \cup B, \\ 0, & \text{otherwise.} \end{cases} \\
= &\begin{cases} 1, & \omega \in (A \setminus B) \cup (B \setminus A) \cup (A \cap B), \\ 0, & \text{otherwise.} \end{cases} \\
= &\begin{cases} 1, & \omega \in (A \setminus B), \\ 1, & \omega \in (B \setminus A), \\ 1, & \omega \in (A \cap B), \\ 0, & \text{otherwise.} \end{cases} \\
= &\begin{cases} 1, & \omega \in A \text{ and } \omega \notin B, \\ 1, & \omega \notin A \text{ and } \omega \in B, \\ 1, & \omega \in A \text{ and } \omega \in B, \\ 0, & \omega \notin A \text{ and } \omega \notin B. \end{cases}
\end{aligned}
$$

Notice that the four conditions above split the sample space Ω into four disjoint sets. As a result, we can write

$$
\big(\mathbb{I}_A + \mathbb{I}_B - \mathbb{I}_A\mathbb{I}_B\big)(\omega) = \begin{cases} 1+0-1\times 0 = 1, & \omega \in A \text{ and } \omega \notin B, \\ 0+1-0\times 1 = 1, & \omega \notin A \text{ and } \omega \in B, \\ 1+1-1\times 1 = 1, & \omega \in A \text{ and } \omega \in B, \\ 0+0-0\times 0 = 0, & \omega \notin A \text{ and } \omega \notin B. \end{cases}
$$

Therefore, we conclude that the functions $\mathbb{I}_{A\cup B}$ are equal $\mathbb{I}_A + \mathbb{I}_B - \mathbb{I}_A\mathbb{I}_B$.

(b) Show that $\mathbb{I}_A \sim \text{Ber}(p)$ where $p = \mathbb{P}(A)$.

Solution: The random variable $\mathbb{I}_A : \Omega \to \mathbb{R}$ takes only two values:

$$
\mathbb{I}_A(\omega) := \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}
$$

Its PMF is given by

$$
\begin{aligned}
&\mathbb{P}(\mathbb{I}_A = 1) = \mathbb{P}\left(\{\omega \in \Omega : \mathbb{I}_A = 1\}\right) = \mathbb{P}\left(\{\omega \in \Omega : \omega \in A\}\right) = \mathbb{P}(A) =: p \\
&\mathbb{P}(\mathbb{I}_A = 0) = 1 - \mathbb{P}(\mathbb{I}_A = 1) = 1 - p,
\end{aligned}
$$

Therefore $\mathbb{I}_A \sim \text{Ber}(p)$.

(c) Check the fundamental relation $\mathbb{E}(\mathbb{I}_A) = \mathbb{P}(A)$.

Solution: We have

$$\begin{aligned}
\mathbb{E}[\mathbb{I}_A] &= \sum_x x\mathbb{P}(\mathbb{I}_A = x) \\
&= 0 \times \mathbb{P}(\mathbb{I}_A = 0) + 1 \times \mathbb{P}(\mathbb{I}_A = 1) \\
&= 0 \times (1-p) + 1 \times p \\
&= p = \mathbb{P}(A).
\end{aligned}$$

(d) Suppose that a random variable $U : \Omega \to [0,1]$ has a uniform distribution, i.e. $U \sim \text{Unif}(0,1)$. For some $0 < p < 1$ define a discrete random variable

$$X(\omega) := \begin{cases} 1, & \text{if } U(\omega) < p, \\ 0, & \text{if } U(\omega) \geq p. \end{cases}$$

Show that $X \sim \text{Ber}(p)$ and that it allows the representation $X = \mathbb{I}_A$.

Solution: We define the set

$$A := \{\omega \in \Omega : U(\omega) < p\},$$

thus, the random variable X can be written as

$$X = \mathbb{I}_A.$$

It follows from (b) that X has Bernoulli distribution. It only remains to find $\mathbb{P}(X = 1)$. We have

$$\mathbb{P}(X = 1) = \mathbb{P}(\mathbb{I}_A = 1) = \mathbb{P}(A) = \mathbb{P}(U < p) = \mathbb{P}(U \leq p) = \frac{p-0}{1-0} = p.$$

Accordingly we conclude that $X \sim \text{Ber}(p)$.

Exercise 2 (Properties of the variance).

(a) Let X be a random variable with $\mathbb{E}(X^2) < \infty$. Show that for any constants $a, b \in \mathbb{R}$, we have

$$\text{Var}(aX + b) = a^2\text{Var}(X).$$

Solution: We first need to check

$$\mathbb{E}[(aX + b)^2] < +\infty.$$

By the linearity of expectation we have

$$\begin{aligned}
\mathbb{E}[(aX+b)^2] &= \mathbb{E}[a^2X^2 + 2abX + b^2] \\
&= a^2\mathbb{E}[X^2] + 2ab\mathbb{E}[X] + b^2 < +\infty
\end{aligned}$$

which is indeed finite by assumption (note that $\mathbb{E}|X|^2 < \infty$ implies that $\mathbb{E}|X| < \infty$ and thus $\mathbb{E}[X]$ is well-defined and finite). Therefore, we can compute the variance

$$\begin{aligned}
\mathrm{Var}(aX+b) &:= \mathbb{E}\Big[(aX+b-\mathbb{E}[aX+b])^2\Big] \\
&= \mathbb{E}\Big[(aX+b-a\mathbb{E}[X]-b)^2\Big] \\
&= \mathbb{E}\Big[(aX-a\mathbb{E}[X])^2\Big] \\
&= a^2\mathbb{E}\Big[(X-\mathbb{E}[X])^2\Big] \\
&= a^2\mathrm{Var}(X)
\end{aligned}$$

(b) Let X, Y be independent random variables with $\mathbb{E}(X^2) < \infty$ and $\mathbb{E}(Y^2) < \infty$. Show that

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Solution: We again first need to check that

$$\mathbb{E}[(X+Y)^2] < +\infty.$$

By the linearity of expectation we have

$$\begin{aligned}
\mathbb{E}[(X+Y)^2] &= \mathbb{E}[X^2+Y^2+2XY] \\
&= \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\mathbb{E}[XY] \\
&\leq \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]} < \infty,
\end{aligned}$$

where we applied the Cauchy-Schwarz inequality (Eq.(15) in Ch. 1.4). Then we can calculate the variance:

$$\begin{aligned}
\mathrm{Var}(X+Y) &:= \mathbb{E}\Big[(X+Y-\mathbb{E}[X+Y])^2\Big] \\
&= \mathbb{E}\Big[(X-\mathbb{E}[X]+Y-\mathbb{E}[Y])^2\Big] \\
&= \mathbb{E}\Big[(X-\mathbb{E}[X])^2\Big] + \mathbb{E}\Big[(Y-\mathbb{E}[Y])^2\Big] + 2\mathbb{E}\Big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\Big] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y),
\end{aligned}$$

where by Theorem (1) in Ch. 1.5, we know that the covariance of two independent random variable is zero.

Exercise 3.
(a) Let $X : \Omega \to \mathbb{R}$ be a random variable with CDF F_X. Find the CDF of the random variable $Y := X^+ = \max\{0, X\}$.

Solution: We start from definition of CDF. Let us first consider $x \geq 0$

$$\begin{aligned}
F_{X^+}(x) :&= \mathbb{P}(X^+ \leq x) \\
&= \mathbb{P}(\max\{0, X\} \leq x) \\
&= \mathbb{P}\Big(\max\{0, X\} \leq x, X \geq 0\Big) + \mathbb{P}\Big(\max\{0, X\} \leq x, X < 0\Big) \\
&= \mathbb{P}(X \leq x, X \geq 0) + \mathbb{P}(0 \leq x, X < 0) \\
&= \mathbb{P}(0 \leq X \leq x) + \mathbb{P}(X < 0) \\
&= \mathbb{P}(X \leq x) \\
&= F_X(x).
\end{aligned}$$

Now consider $x < 0$

$$\begin{aligned}
F_{X^+}(x) :&= \mathbb{P}(X^+ \leq x) \\
&= \mathbb{P}(\max\{0, X\} \leq x) \\
&= \mathbb{P}(0 \leq \max\{0, X\} \leq x) \\
&= \mathbb{P}(\emptyset) \\
&= 0.
\end{aligned}$$

All in all, we obtain

$$F_{X^+}(x) = \begin{cases} F_X(x) & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

(b) Let $X : \Omega \to \mathbb{R}$ be a continuous random variable with CDF F_X and PDF f_X. Find the CDF and PDF of the the random variable $Y := -X$.

Solution: We again start from definition of CDF. Take any $x \in \mathbb{R}$ and observe that

$$\begin{aligned}
F_{-X}(x) :&= \mathbb{P}(-X \leq x) \\
&= 1 - \mathbb{P}\Big((-X \leq x)^c\Big) \\
&= 1 - \mathbb{P}(-X > x) \\
&= 1 - \mathbb{P}(X < -x) \\
&= 1 - \Big(\mathbb{P}(X \leq -x) - \mathbb{P}(X = x)\Big) \\
&= 1 - \mathbb{P}(X \leq -x) \\
&= 1 - F_X(-x),
\end{aligned}$$

where $\mathbb{P}(X = x) = 0$ becasue X is assumed to be a continuous random variable. Taking derivative w.r.t. x, we obtain

$$\begin{aligned} f_{-X}(x) &= \frac{d}{dx} F_{-X}(x) \\ &= \frac{d}{dx}(1 - F_X(-x)) \\ &= F'_X(-x) \\ &= f_X(-x). \end{aligned}$$

Exercise 4 (Linear change-of-units transformation).

(a) A random variable U is uniformly distributed over $[0, 1]$, i.e. $U \sim \text{Unif}(0, 1)$. Determine the distribution of $V := rU + s$ for any numbers $r > 0$ and $s \in \mathbb{R}$. What happens for $r = 0$ and $r < 0$?

Solution:
Case $r > 0$. First note that $V \in [s, s + r]$. Moreover, by (b) below, we find $f_V(y) = \frac{1}{r}$ over the range $[s, s + r]$. Thus, V remains uniform distributed.

$$V \sim \text{Unif}(s, s + r).$$

Case $r < 0$. Here we have $V \in [s + r, s]$. Similar as above, we find that $f_V(y) = \frac{1}{|r|}$ over the range $[r + s, s]$. Thus

$$V \sim \text{Unif}(s + r, s).$$

If $r = 0$

$$V = s \quad \Rightarrow \quad V \sim \delta_s.$$

(b) Let a continuous random variable $X : \Omega \to \mathbb{R}$ have CDF F_X and PDF f_X, and let us change units to $Y := rX + s$ for $r > 0$ and $s \in \mathbb{R}$. Show that

$$F_Y(y) = F_X\left(\tfrac{y-s}{r}\right) \quad \text{and} \quad f_Y(y) = \tfrac{1}{r} f_X\left(\tfrac{y-s}{r}\right).$$

Solution: We have

$$\begin{aligned} F_Y(y) :&= \mathbb{P}(Y \le y) \\ &= \mathbb{P}(rX + s \le y) \\ &= \mathbb{P}(X \le \frac{y - s}{r}) \qquad \text{(becasue } r > 0) \\ &= F_X\left(\tfrac{y-s}{r}\right). \end{aligned}$$

Taking derivative w.r.t. y, we obtain

$$f_Y(y) := F'_Y(y) = \frac{1}{r} F'_X\left(\tfrac{y-s}{r}\right) = \frac{1}{r} f_X\left(\tfrac{y-s}{r}\right)$$

(c) Determine the distribution of λX for $X \sim \text{Exp}(\lambda)$ using **(b)**, where $\lambda > 0$. What kind of distribution does λX have?

Solution: Recall first that $X \in [0, \infty)$. Thus $\lambda X \in [0, \infty)$. We apply (b) with $r = \lambda$ and $s = 0$. For $x \geq 0$, we have

$$\begin{aligned} f_{\lambda X}(x) &= \frac{1}{\lambda} f_X(\tfrac{x}{\lambda}) \\ &= \frac{1}{\lambda} \lambda e^{-\lambda \frac{x}{\lambda}} = e^{-x} \end{aligned}$$

and thus we conclude that $\lambda X \sim \text{Exp}(1)$.

Exercise 5. A hydrologist in Monsville models the maximum one-day rainfall X (inches) in a randomly selected year, following the density

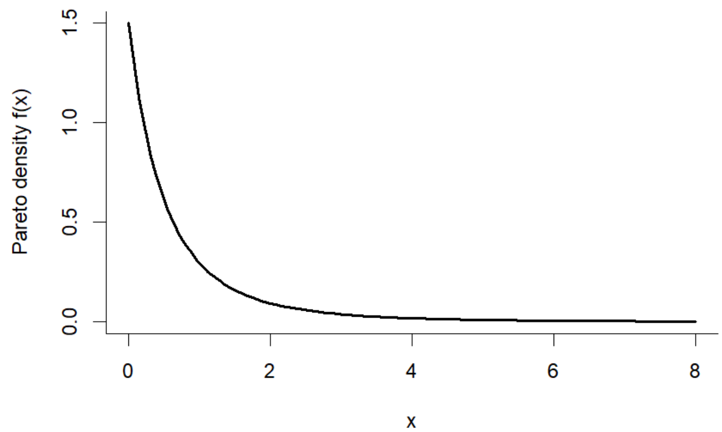
$$f(x) := \frac{3}{2}\left(\frac{x}{2} + 1\right)^{-4} \quad \text{for } x \geq 0.$$

This is a special case of *Pareto density*, which is used to model phenomena where the values taken tend to be mostly "small" but occasionally "big".

(a) Plot the density function. Prove its normalization first by direct computation and then using the numerical integration in `R`.

Solution: `R` code:

```
density_pareto <- function (x) {3/2*(x/2 + 1)^{-4}}
curve (density_pareto, from=0, to=8, bty="l", xlab ="x", ylab ="Pareto
density f(x)",lwd =2)
integrate(density_pareto, lower =0 , upper = Inf)
  1 with absolute error < 7e-08
```



We have

$$\begin{aligned}
\int_0^\infty f(x)dx &= \int_0^\infty \frac{3}{2}\left(\frac{x}{2}+1\right)^{-4} dx \\
&= \frac{3}{2}\int_1^\infty y^{-4}\cdot 2\cdot dy \qquad (y := \frac{x}{2}+1) \\
&= 3\frac{1}{-3}y^{-3}\Big|_1^\infty = 1
\end{aligned}$$

(b) The severve flooding in Monsville happens if there is more than 7 inches of rain in one day. What is the proportion of years with severve flooding?

Solution: We have

$$\mathbb{P}(X > 7) = \int_7^\infty f(x)dx = \frac{8}{729} = 1.097\%.$$

(c) Compute the average and variance of maximum-one day rainfall.

Solution: We have

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty xf(x)dx = 1, \\
\mathbb{E}[X^2] &= \int_0^\infty x^2 f(x)dx = 4, \\
\mathrm{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 3.
\end{aligned}$$

Exercise 6. The built-in-dataset `islands` in the package `stats` contains the size of the world's land masses that exceed 10,000 square miles. You can access that data for example by `require(stats);data(islands)`.

(a) How many are there? Print the 5 largest land masses using `sort()`.

(b) Find the smallest and the largest observations (i.e. its name and size) using the command `which`.

(c) Compute the 1st quantile and median based on their definitions. Check your result using the command `summary()`.

(d) Draw a histogram with a good choice of bin size. Try also
`dotchart(log(sort(islands), 10))`

Solution: see next page for the `R` code.

```r
require(stats);
head(islands);
```

```
##       Africa   Antarctica         Asia    Australia Axel Heiberg       Baffin
##        11506         5500        16988         2968           16          184
```

```r
# -- part (a)
# How many are there?
length(islands)
```

```
## [1] 48
```

```r
# Print the 5 largest land masses
sort(islands, decreasing = TRUE)[1:5]
```

```
##          Asia        Africa North America South America    Antarctica
##         16988         11506          9390          6795          5500
```

```r
# -- part (b)
# Find the largest
which.max(islands) # name
```

```
## Asia
##    3
```

```r
islands[which.max(islands)] # size (or simply use max(islands))
```

```
##  Asia
## 16988
```

```r
# Find the smallest
which.min(islands) # name
```

```
## Vancouver
##        47
```

```r
islands[which.min(islands)] # size (or simply use min(islands))
```

```
## Vancouver
##        12
```

```r
# -- part (c)
# median

median_function <- function(v){
  n <- length(v);
  if (n %% 2 == 1) {
    # length of vector is odd
    m <- sort(v)[(n + 1) / 2];
  } else {
    # length of vector is even
    m <- (sort(v)[n/2]+sort(v)[n/2+1])/2;
  }
  return(m);
}

median_function(islands)
```

```
## Iceland
```

```
##      41
```

```r
median(islands)
```

```
## [1] 41
```

```r
summary(islands)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    12.0    20.5    41.0  1252.7   183.2 16988.0
```

```r
# 1st quantile

i <- floor (25/100 * length(islands))
sort(islands)[i]
```

```
## Tierra del Fuego
##               19
```

```r
sort(islands)[i+1]
```

```
## Devon
##    21
```

```r
quantile(islands, p=0.25, type =1)
```

```
## 25%
##  19
```

```r
quantile(islands, p=0.25, type =2)
```

```
## 25%
##  20
```

```r
quantile(islands, p=0.25, type =3)
```

```
## 25%
##  19
```

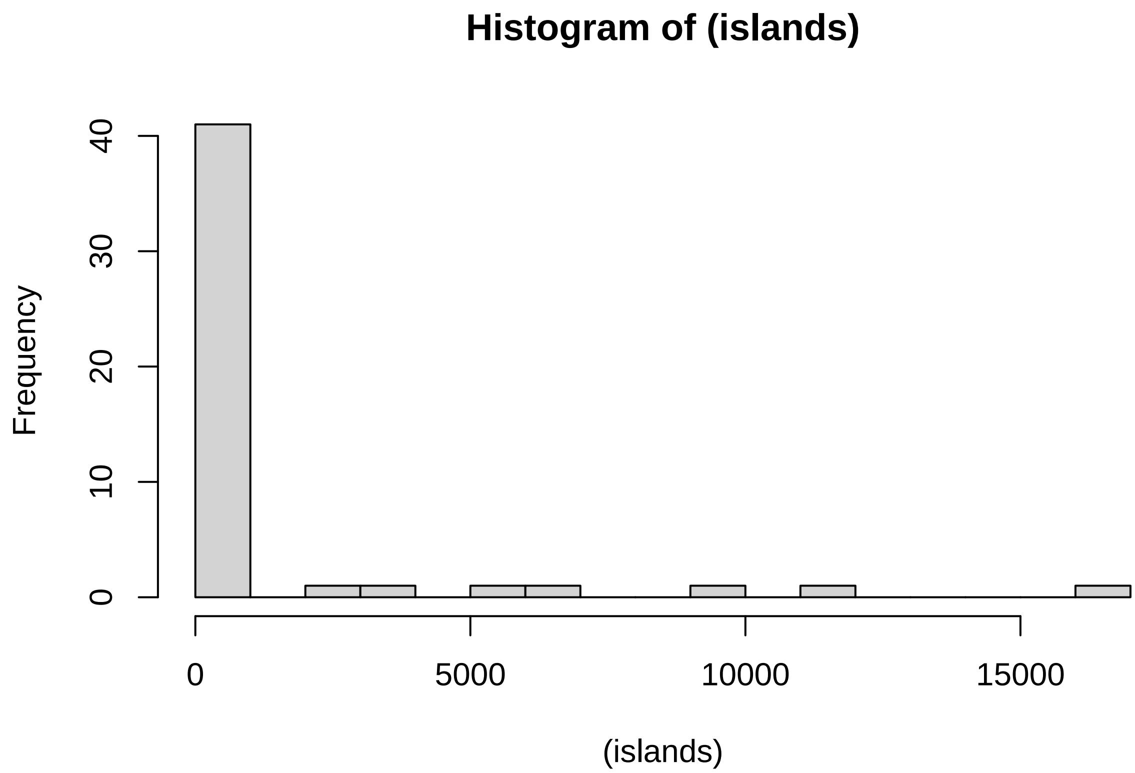
```r
quantile(islands, p=0.25, type =7) # this is default
```

```
##  25%
## 20.5
```

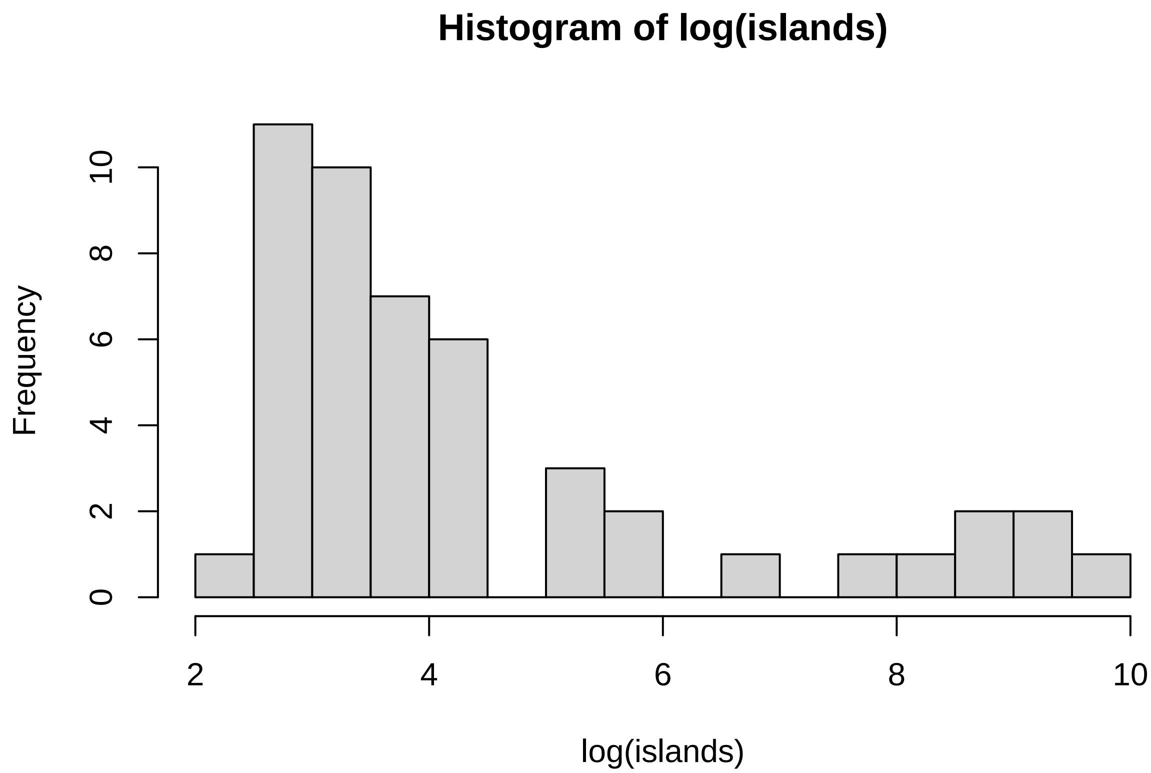
```r
quantile(islands, p=0.25)
```

```
##  25%
## 20.5
```

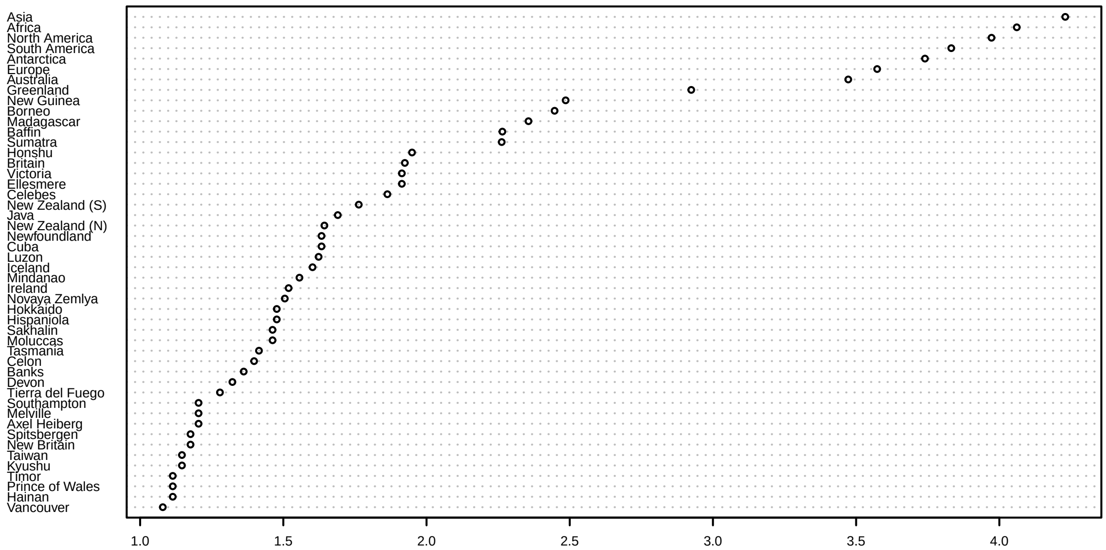
```r
# -- part (d)
hist((islands),breaks = 16)
```




```
hist(log(islands),breaks = 16)
```




```
dotchart(log(sort(islands), 10),cex = 0.4)
```


Asia
Africa
North America
South America
Antarctica
Europe
Australia
Greenland
New Guinea
Borneo
Madagascar
Baffin
Sumatra
Honshu
Britain
Victoria
Ellesmere
Celebes
New Zealand (S)
Java
New Zealand (N)
Newfoundland
Cuba
Luzon
Iceland
Mindanao
Ireland
Novaya Zemlya
Hokkaido
Hispaniola
Sakhalin
Moluccas
Tasmania
Celon
Banks
Devon
Tierra del Fuego
Southampton
Melville
Axel Heiberg
Spitsbergen
New Britain
Taiwan
Kyushu
Timor
Prince of Wales
Hainan
Vancouver
1.0
1.5
2.0
2.5
3.0
3.5
4.0