Statistics for geoscientists

Plotting data

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

– the mean of x is 9

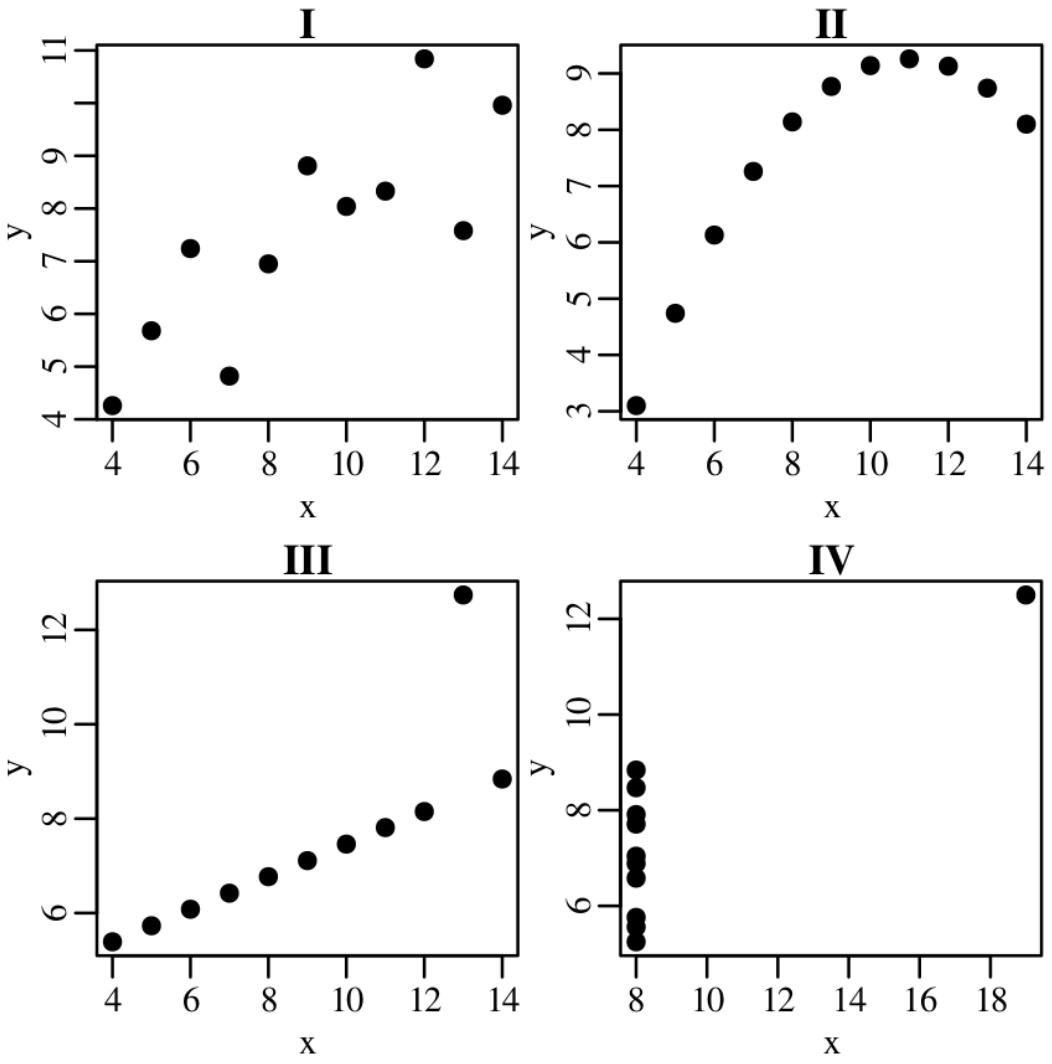
– the variance of x is 11

– the mean of y is 7.50

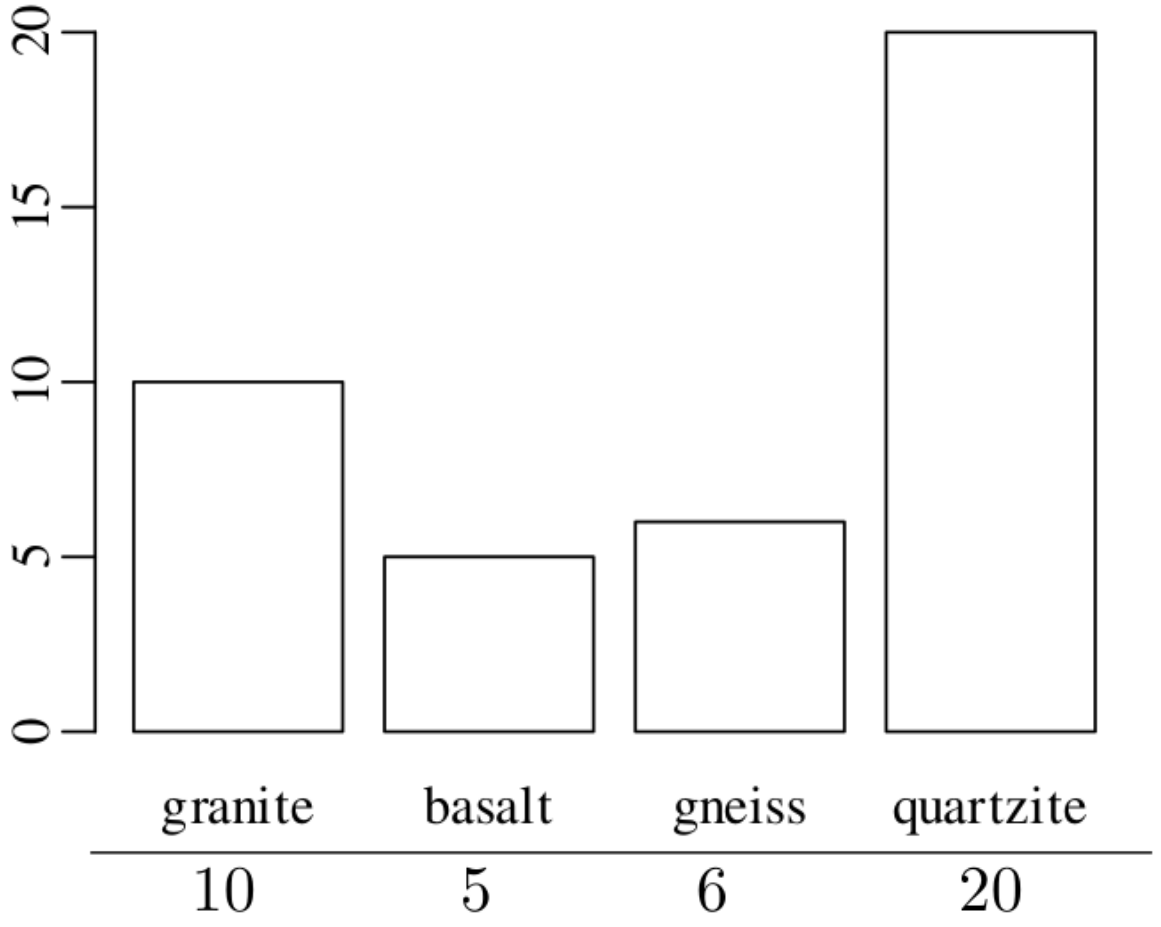
– the variance of y is 4.125

– the correlation coefficient of x and y is 0.816

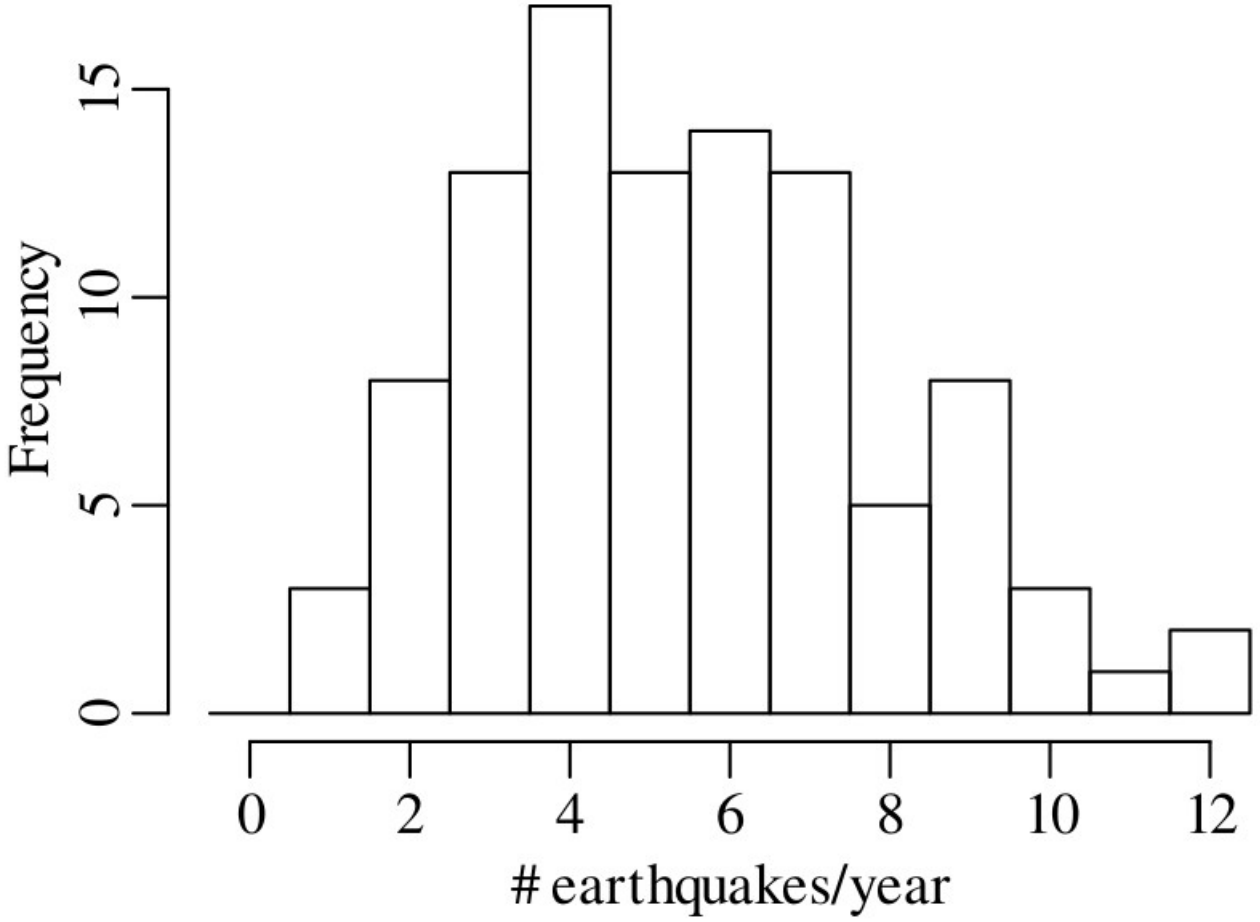
– the best fit line is given by $y = 3.00 + 0.500x$


I
II
III
IV
x
y
4
6
8
10
12
14
16
18
3
5
7
9
11

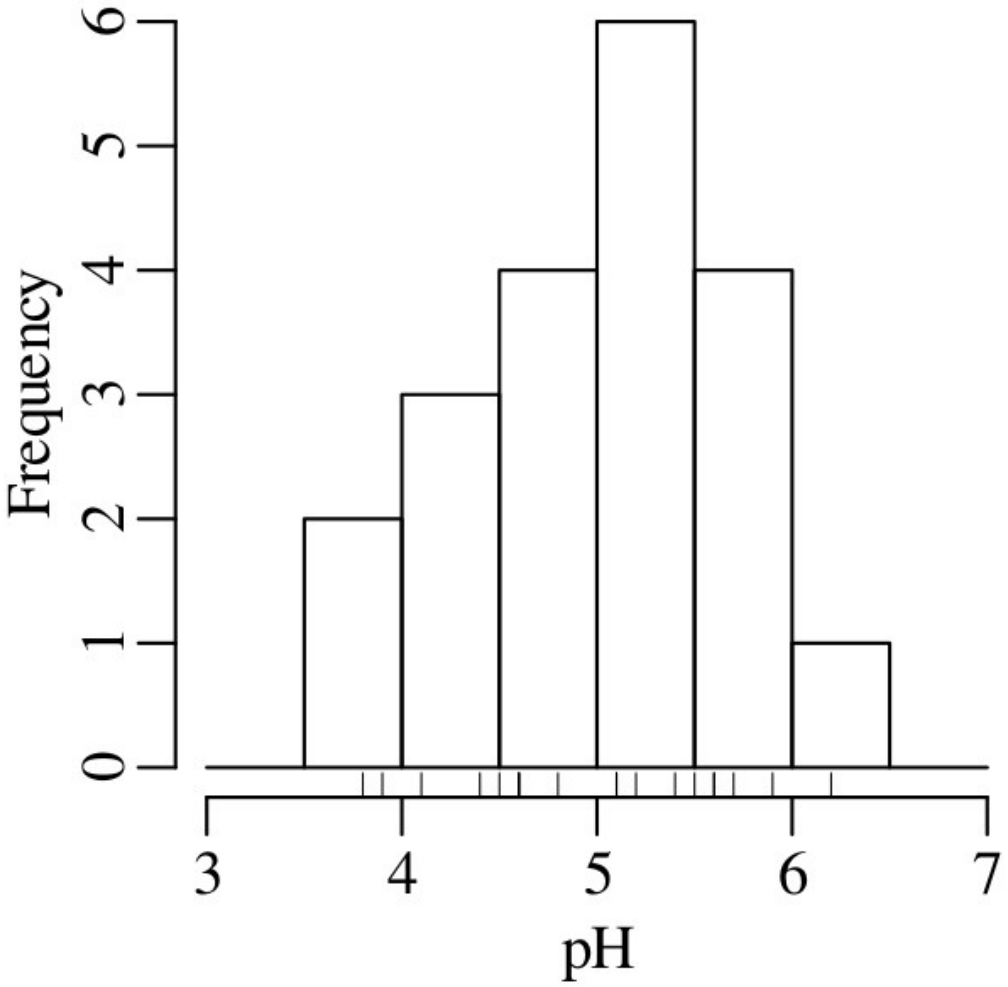
Categorical data



Count data

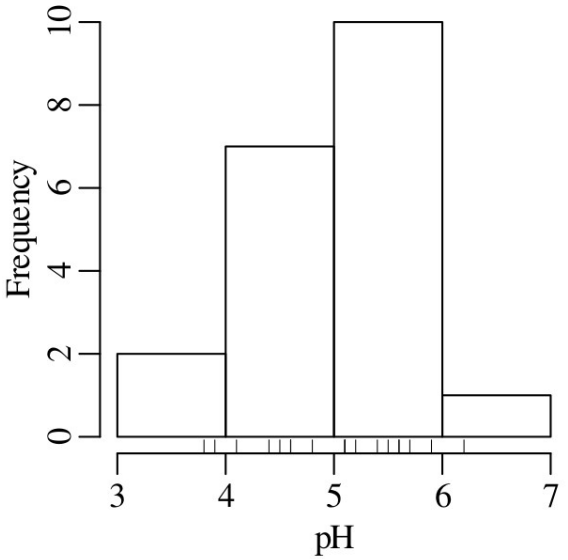


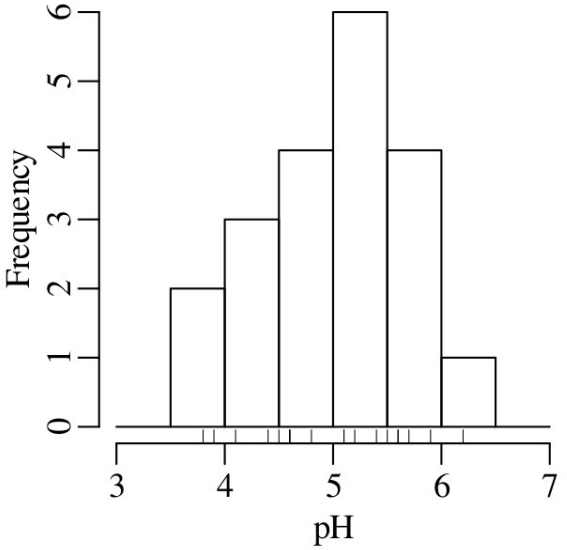
Continuous data

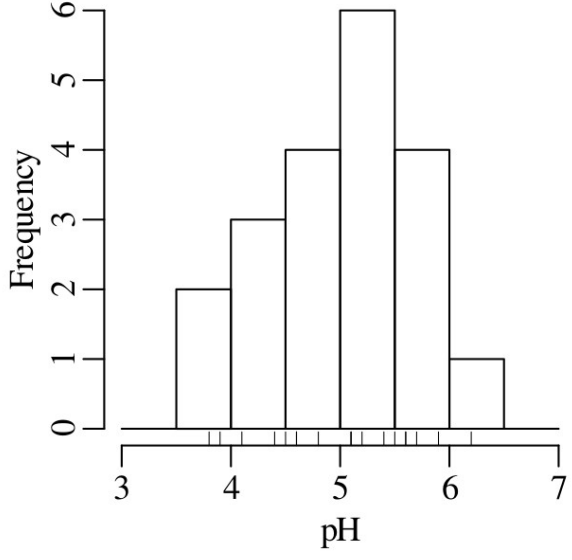


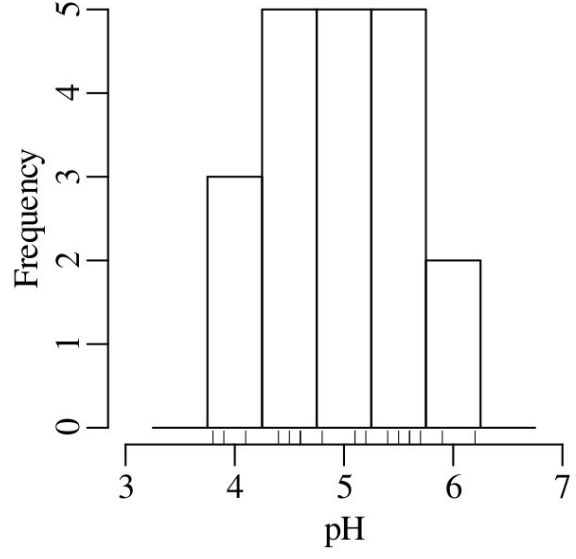
i. How many bins?

ii. Where to place the bins?



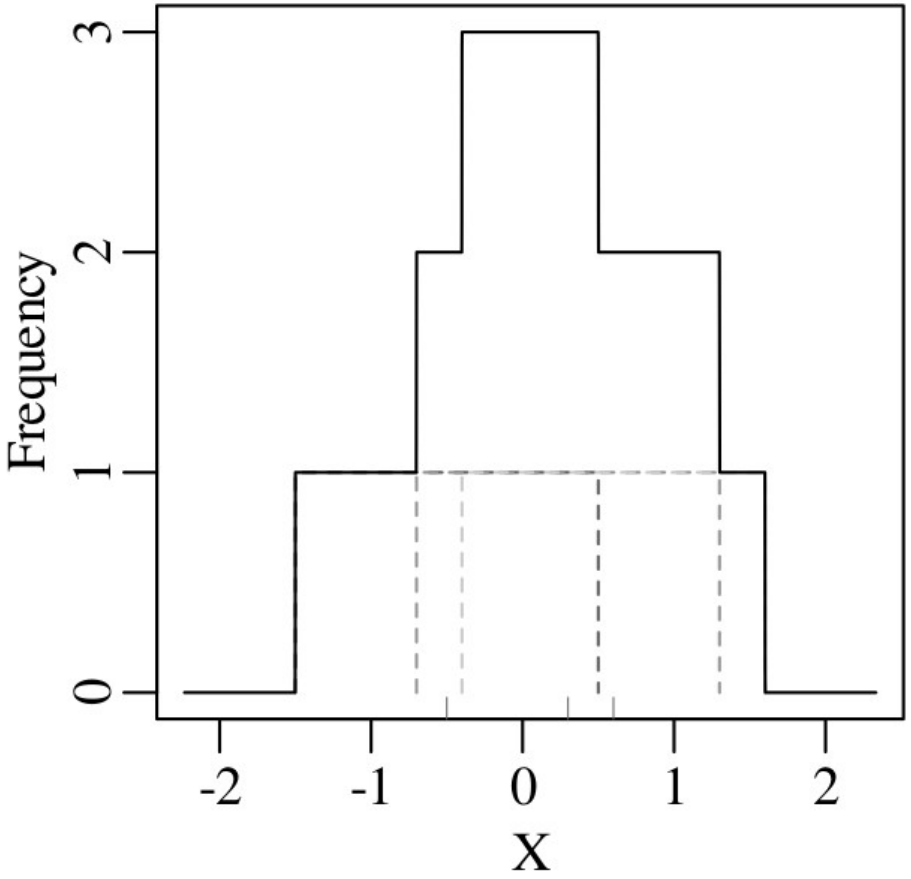


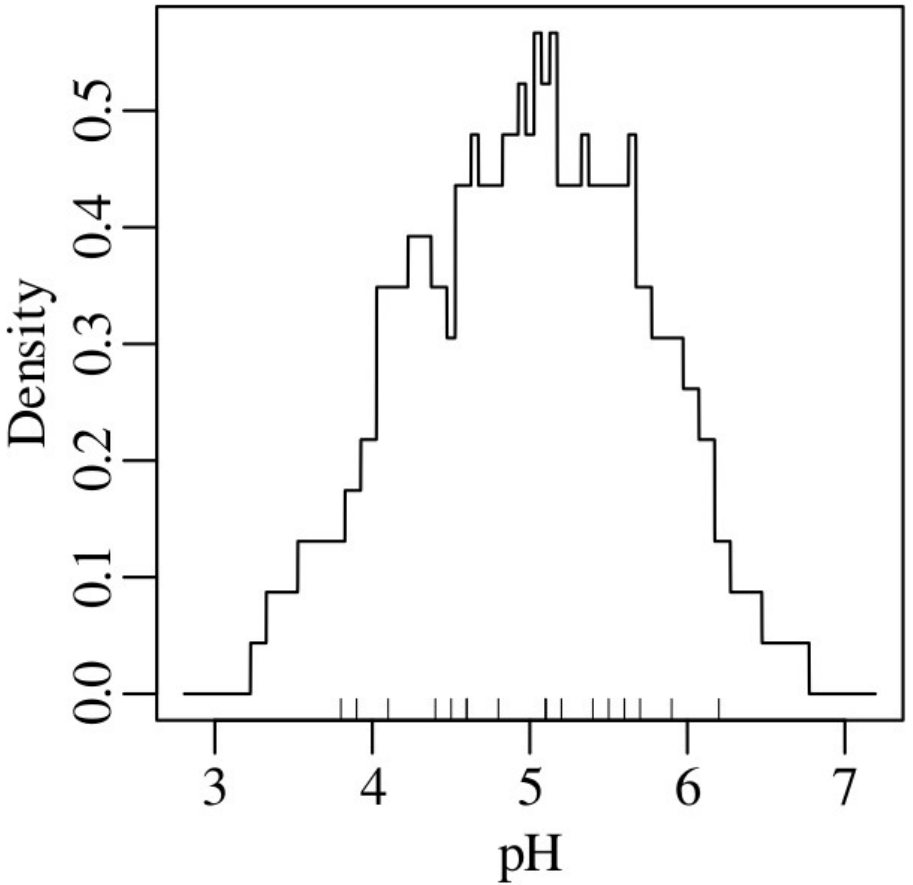




Kernel Density Estimate

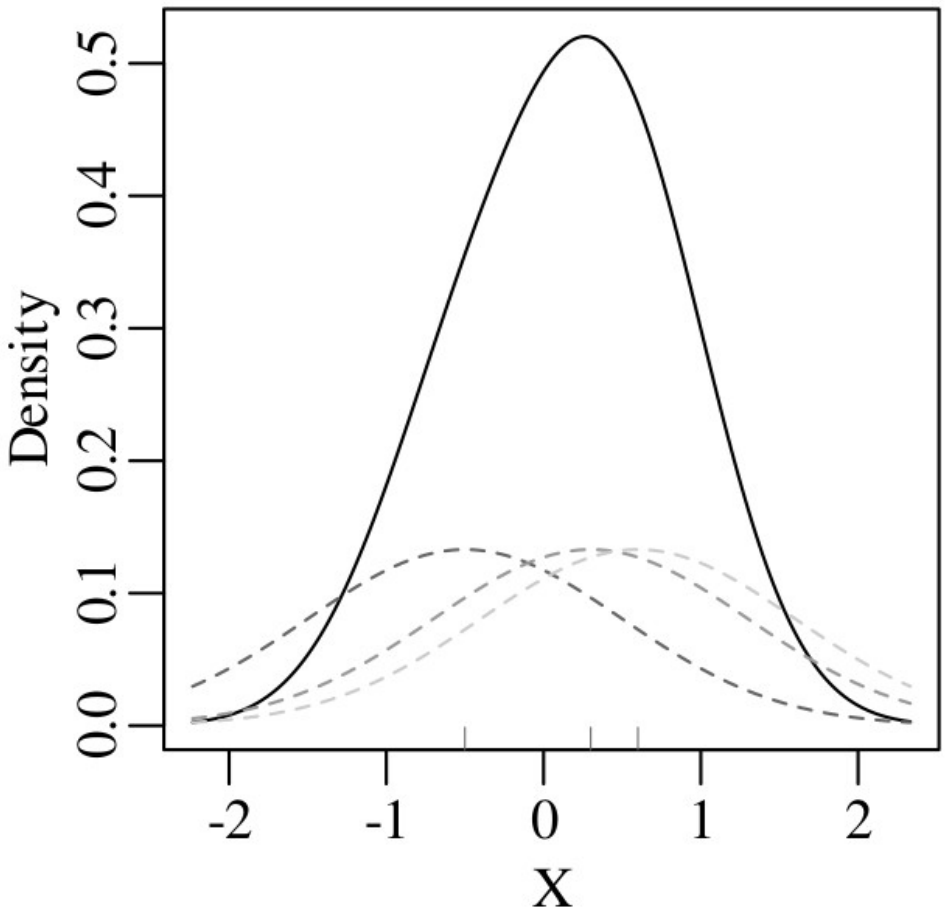
$$KDE(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

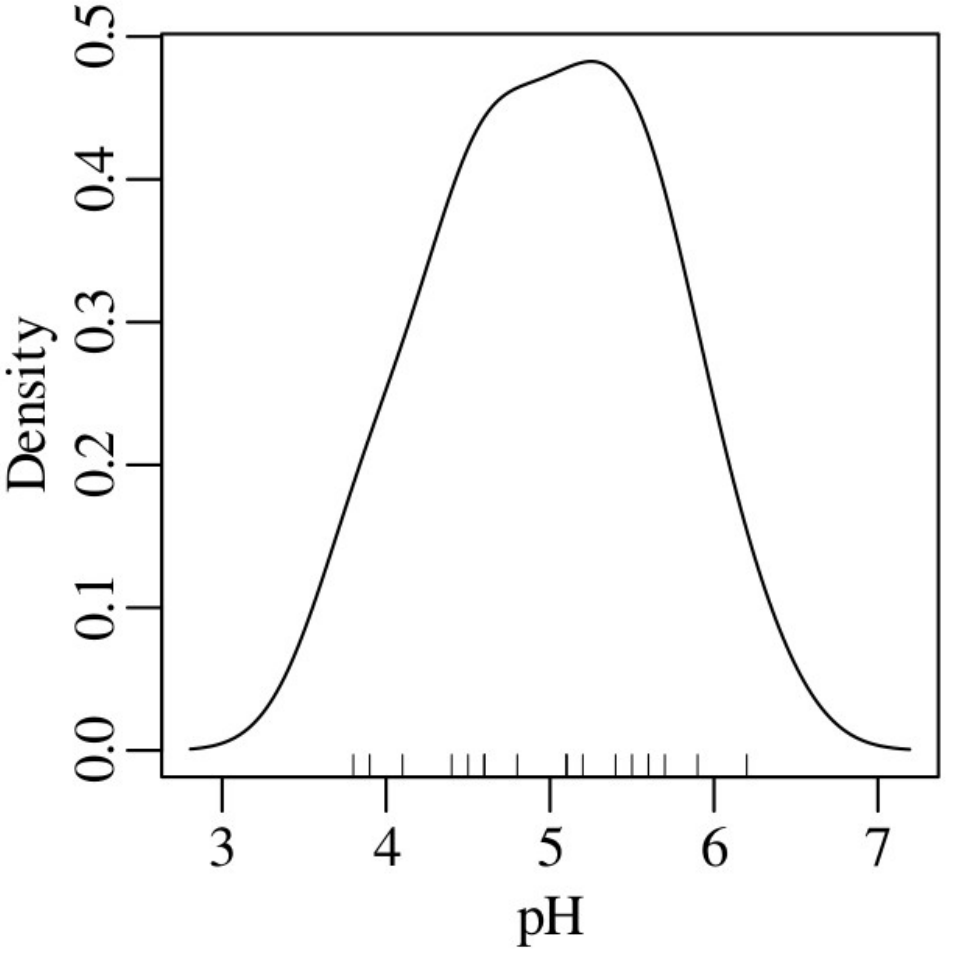




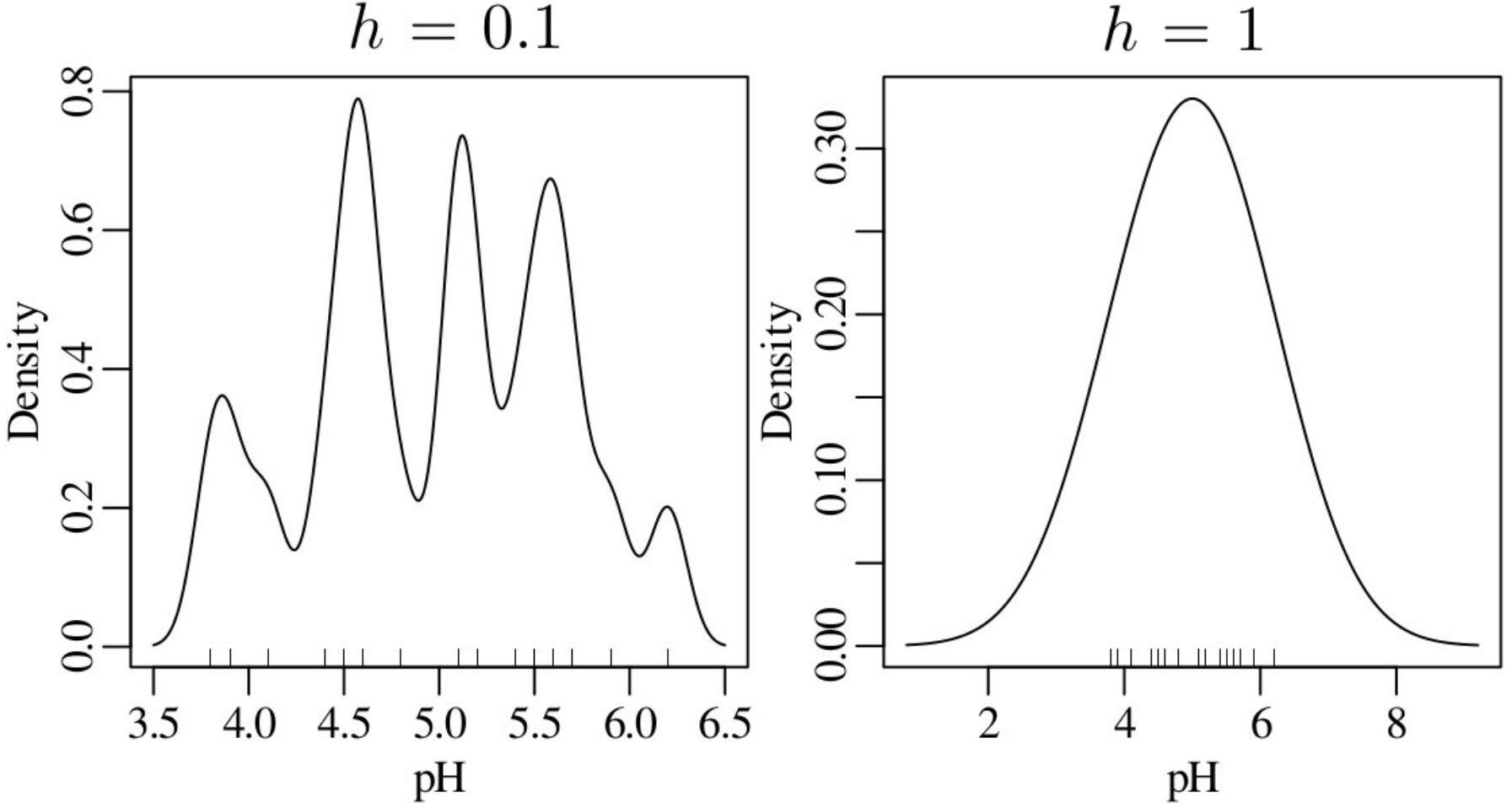
Gaussian kernel:

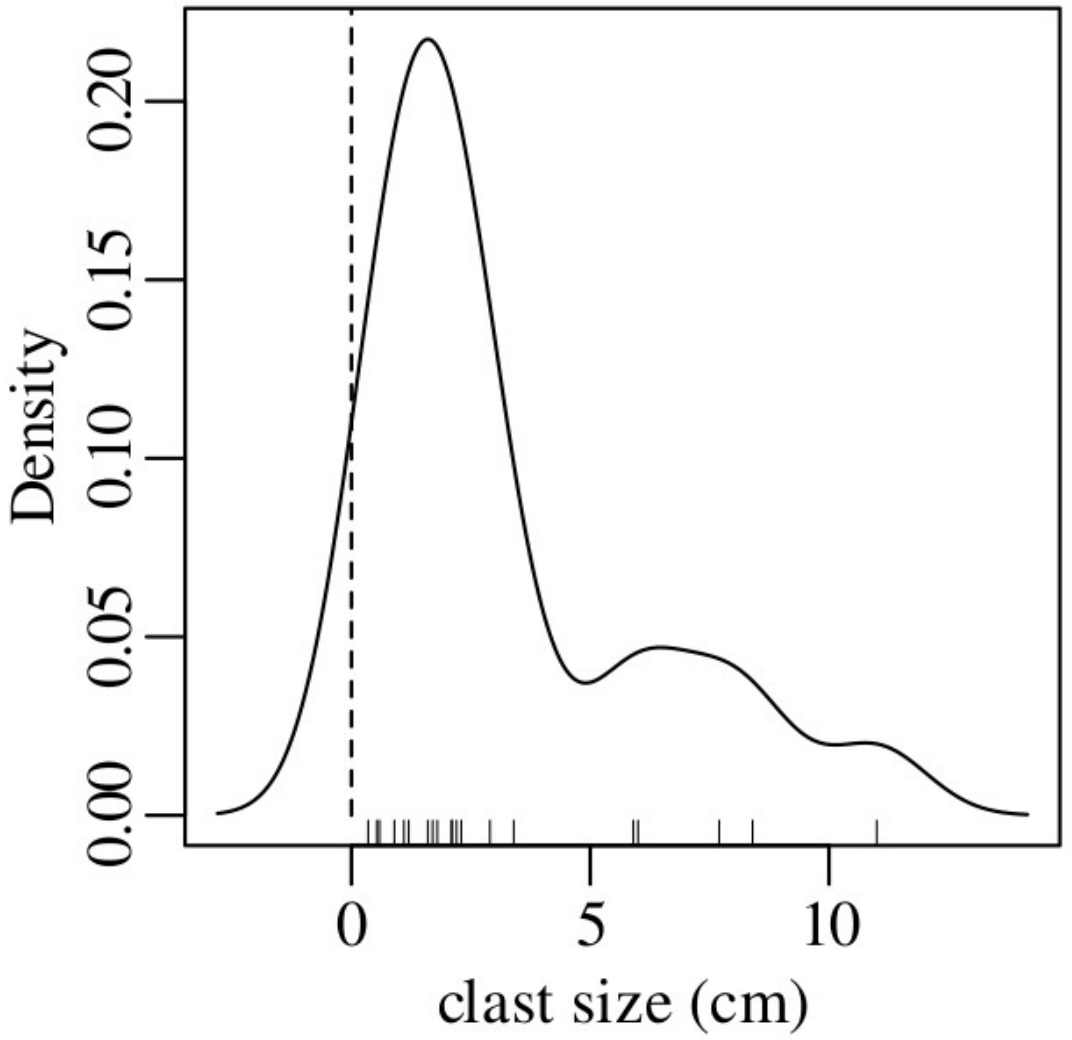
$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{u^2}{2}\right]$$



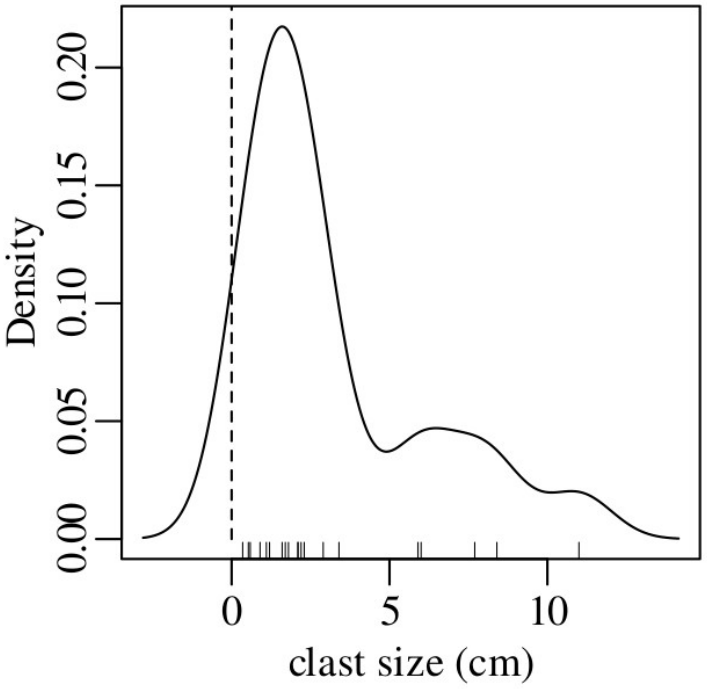


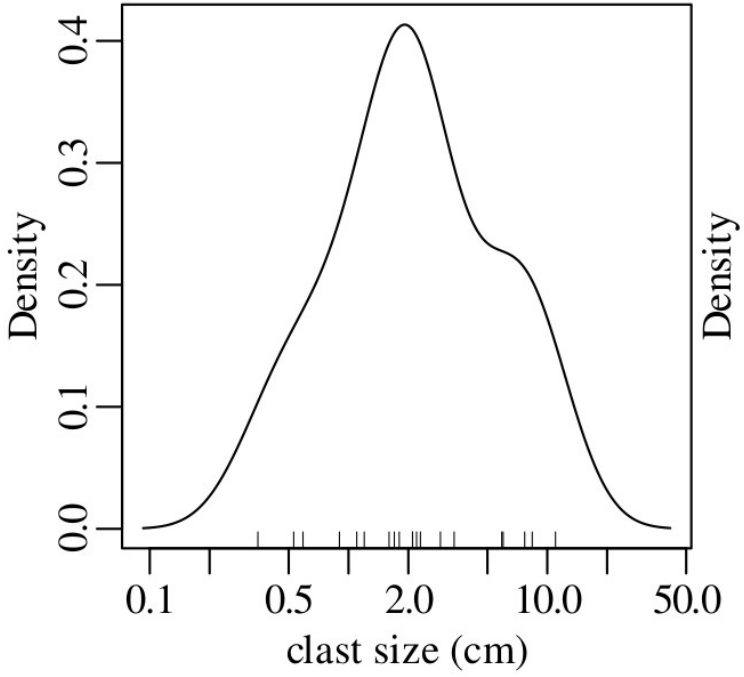
bandwidth

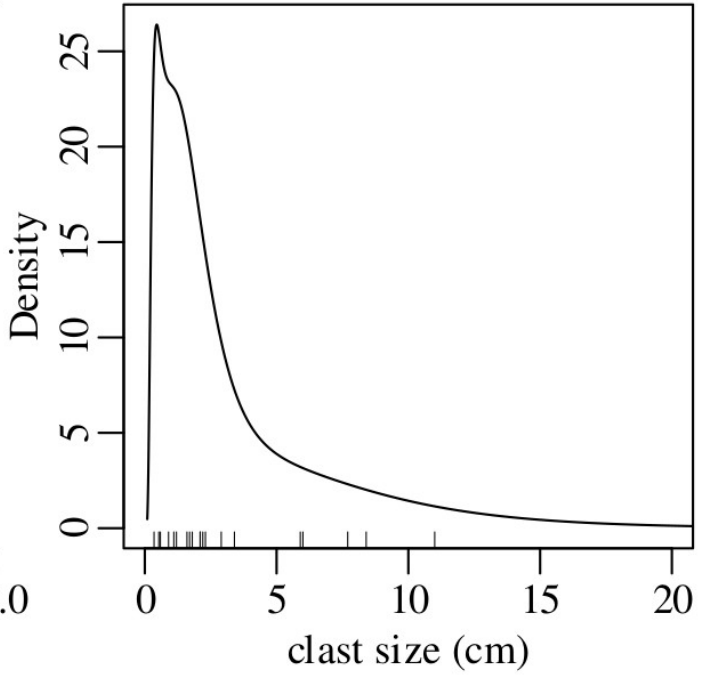

$h = 0.1$
Density
0.0
0.2
0.4
0.6
0.8
3.5
4.0
4.5
5.0
5.5
6.0
6.5
pH
$h = 1$
Density
0.00
0.10
0.20
0.30
2
4
6
8
pH

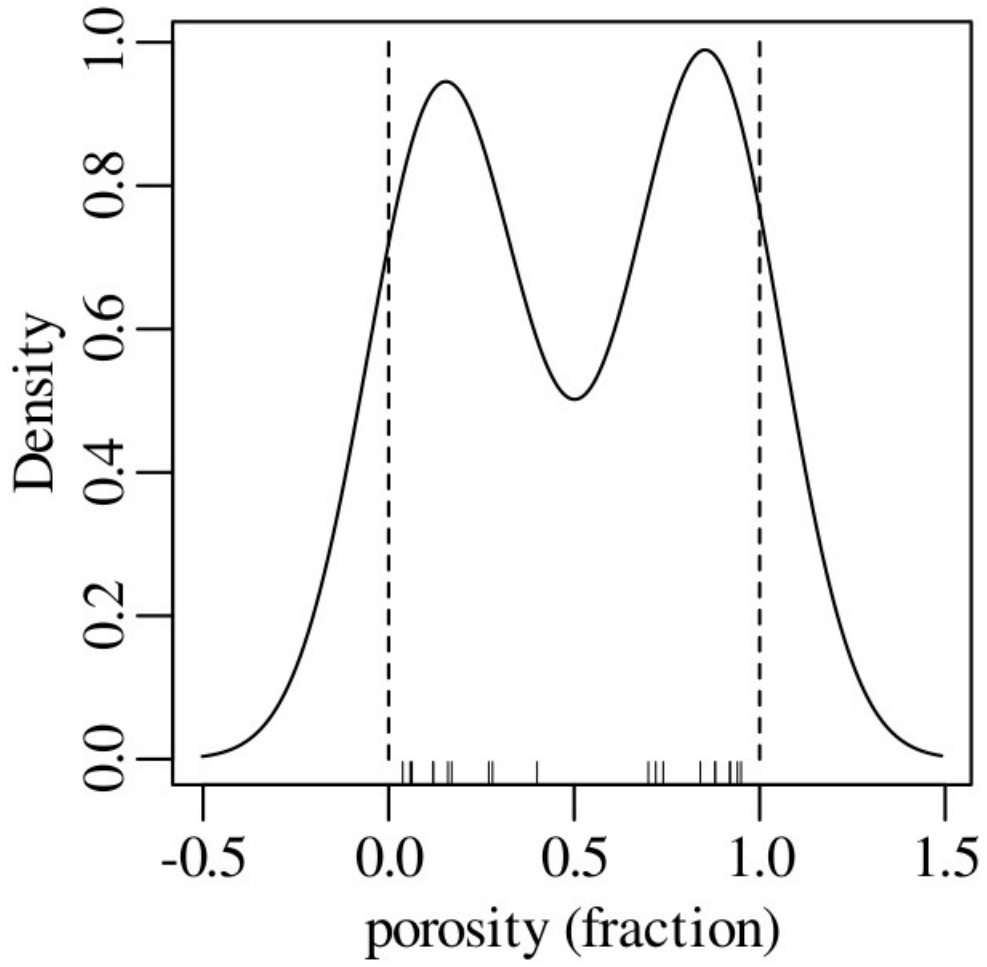

Density
0.00
0.05
0.10
0.15
0.20
0
5
10
clast size (cm)

logarithmic transformation





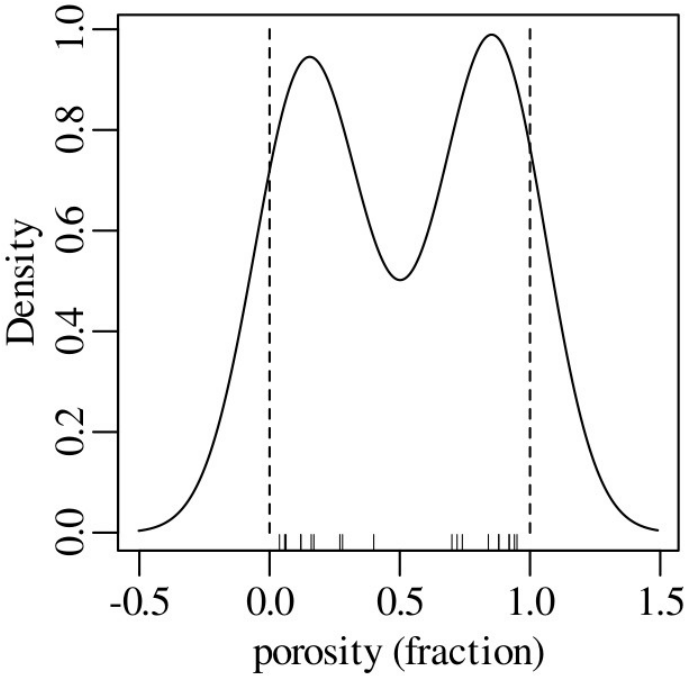


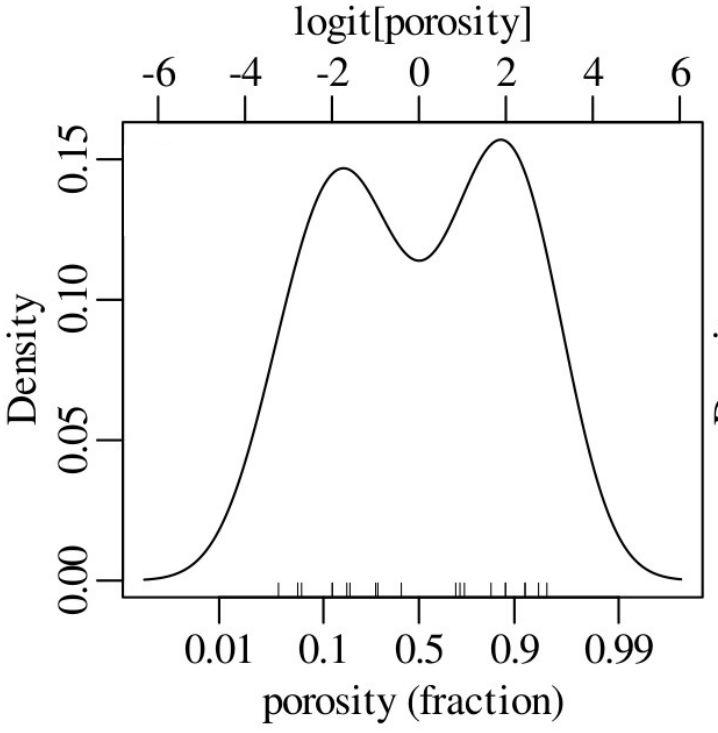

Density
0.0
0.2
0.4
0.6
0.8
1.0
-0.5
0.0
0.5
1.0
1.5
porosity (fraction)

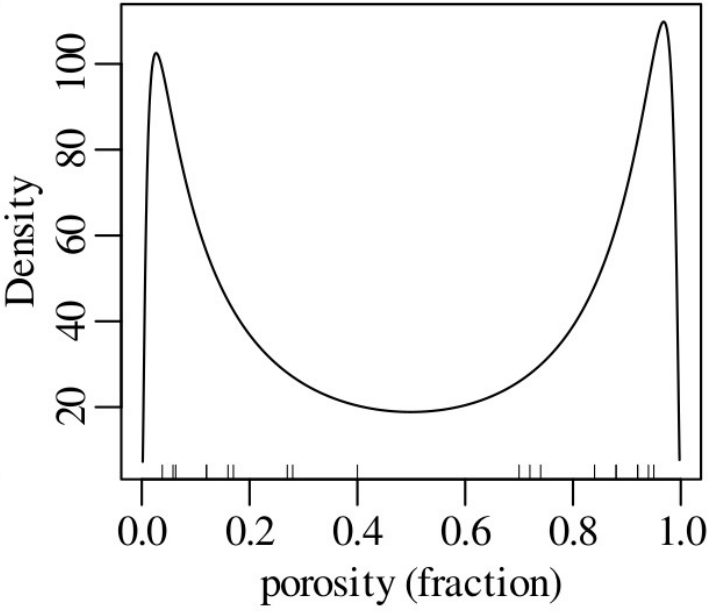
logistic transformation

$$u = \text{logit}(x) = \ln\left[\frac{x}{1-x}\right]$$

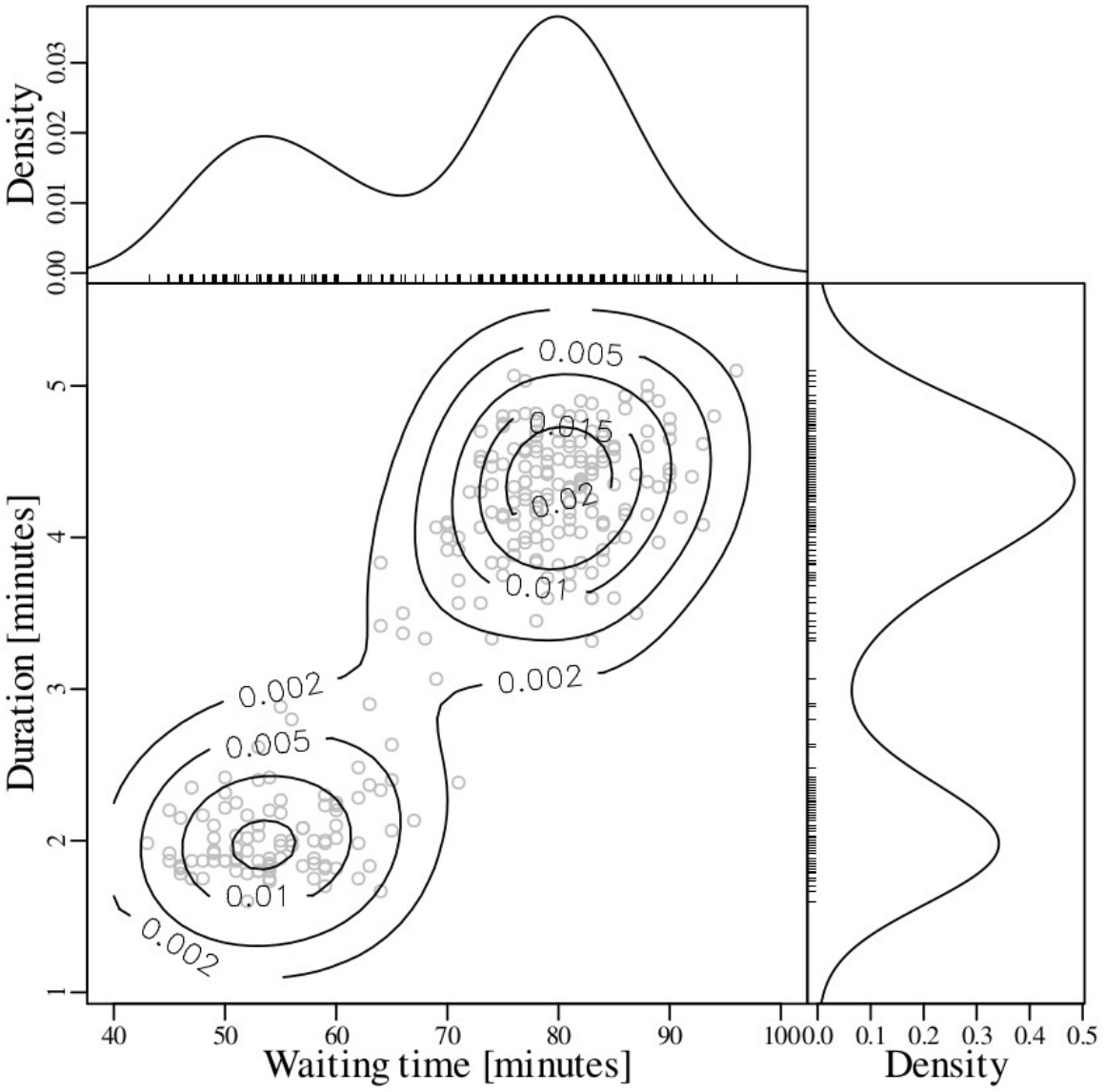
$$x = \text{logit}^{-1}(u) = \frac{\exp[u]}{\exp[u]+1}$$





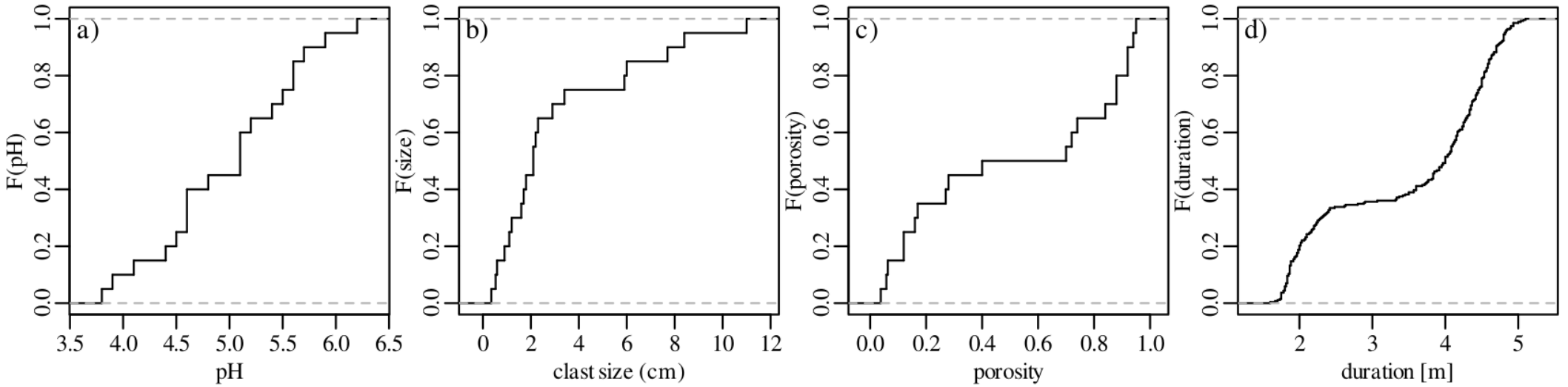


Multivariate distributions



Empirical cumulative distribution fuctions

$$F(x) = \sum_{i=1}^{n} 1(x_i < x)/n$$



Statistics for geoscientists

An introduction to R