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

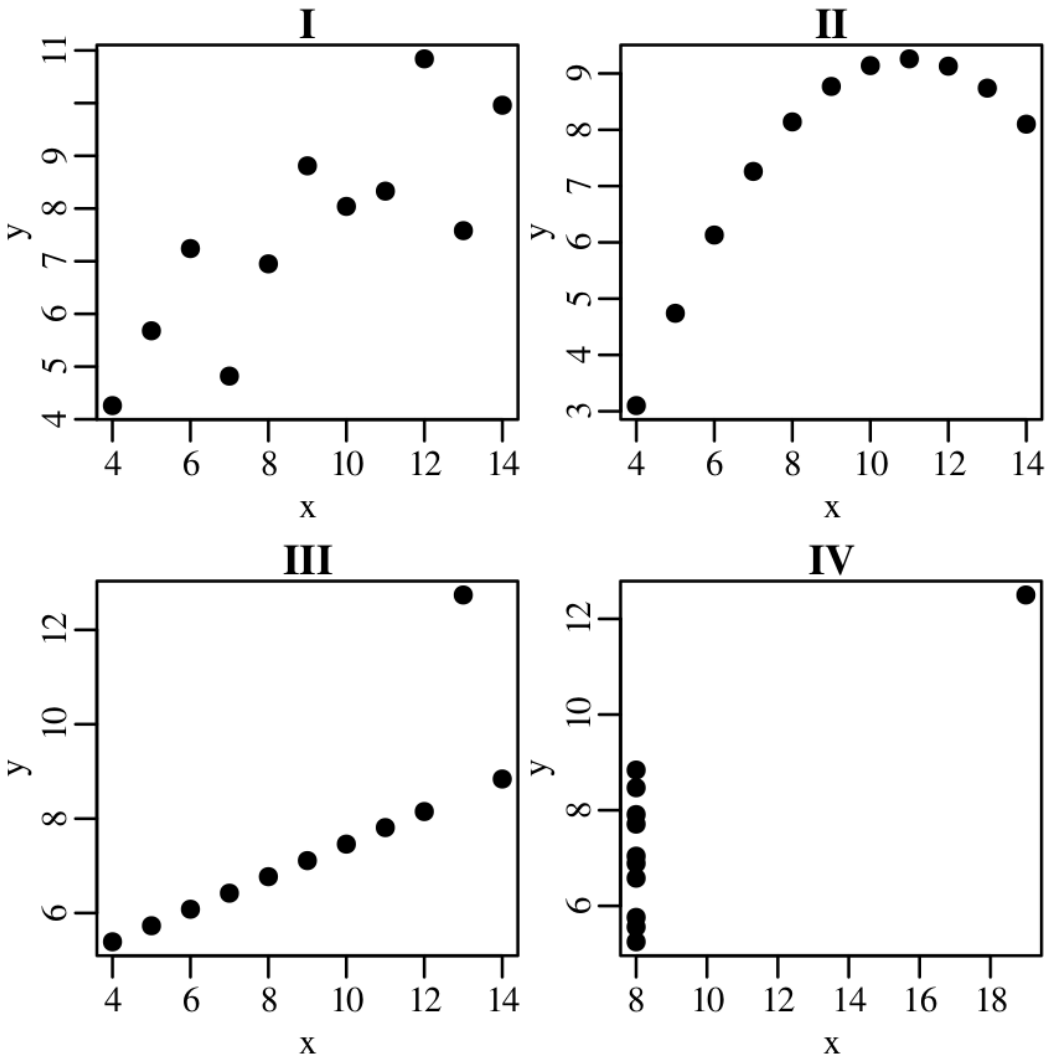
– the variance of x is 11

– the mean of y is 7.50

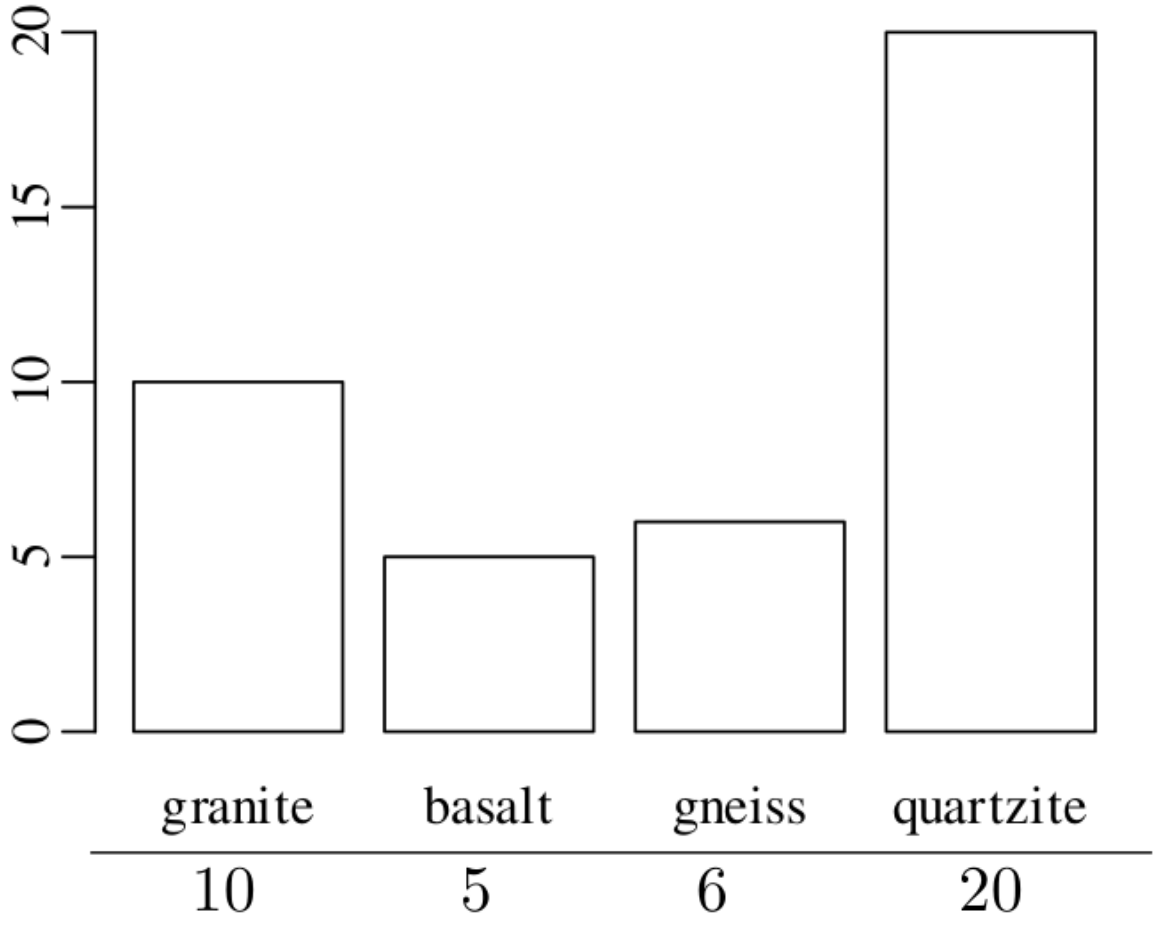
– the variance of y is 4.125

– the correlation coefficient of x and y is 0.816

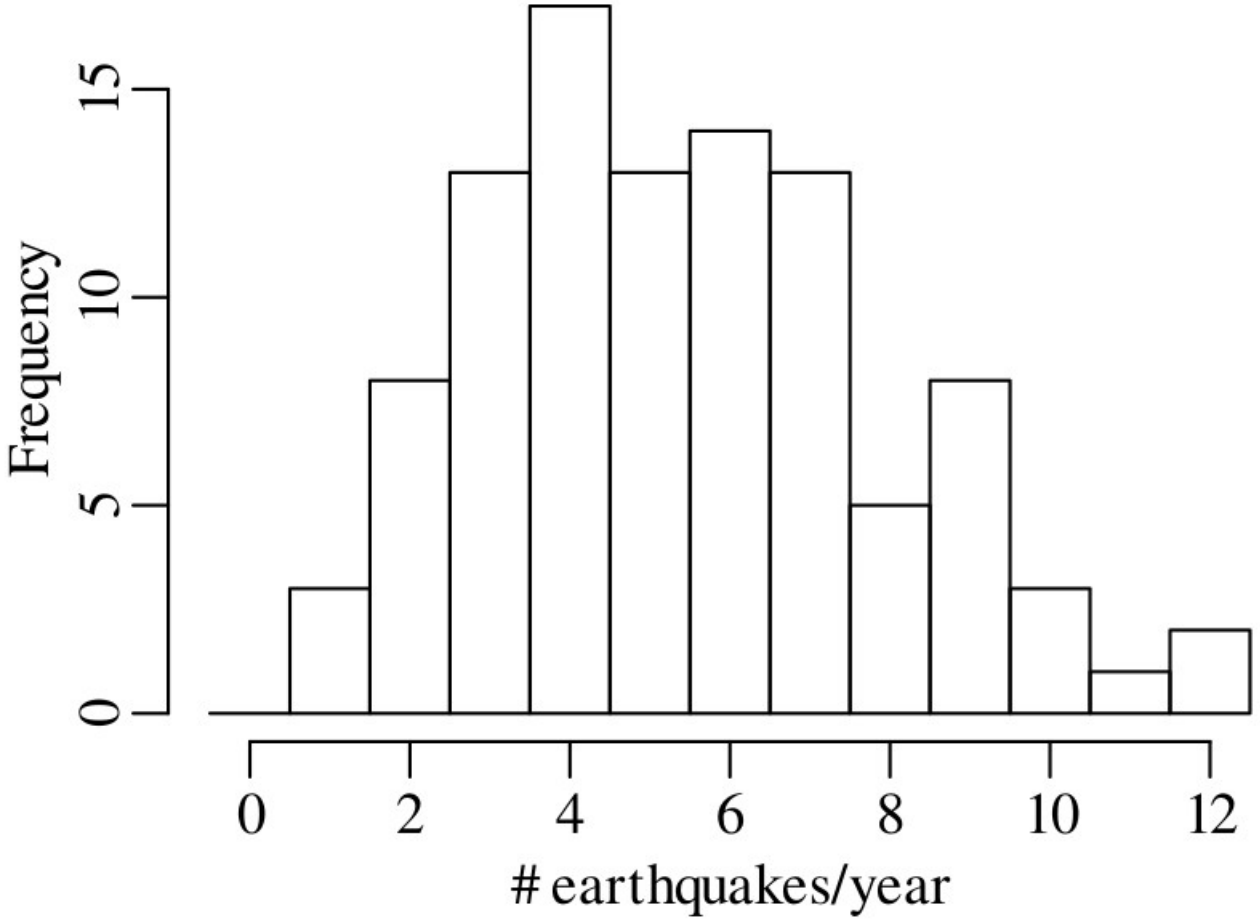
– the best fit line is given by $y = 3.00 + 0.500x$


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

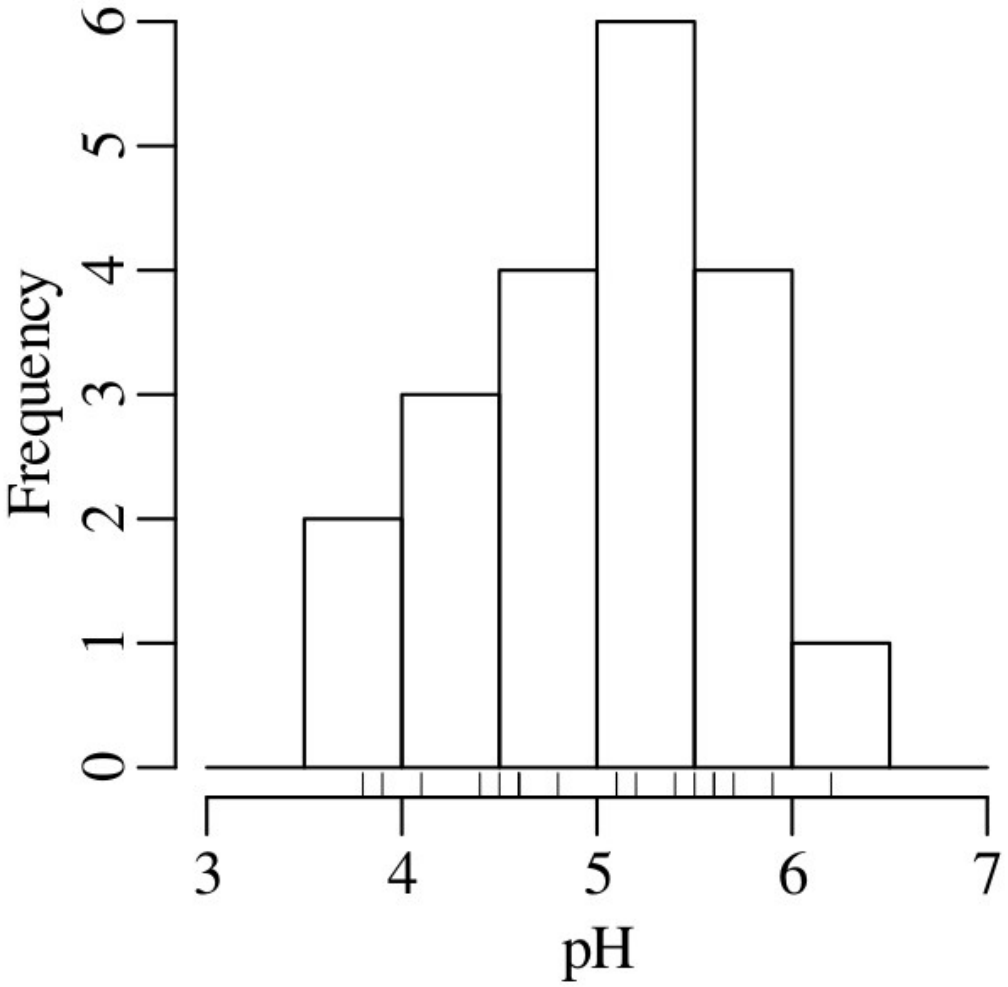
Categorical data



Count data

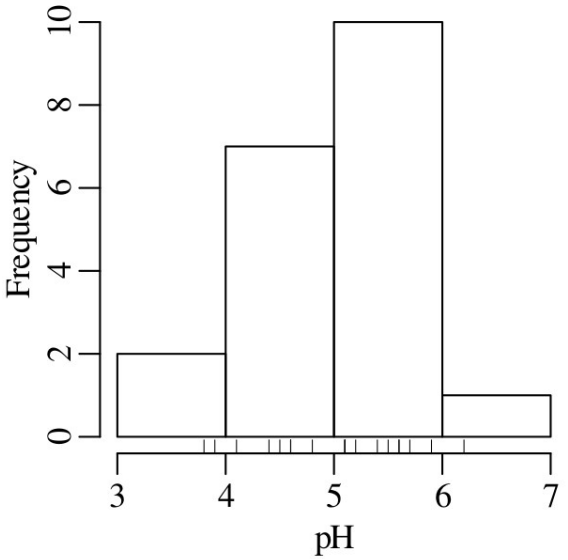


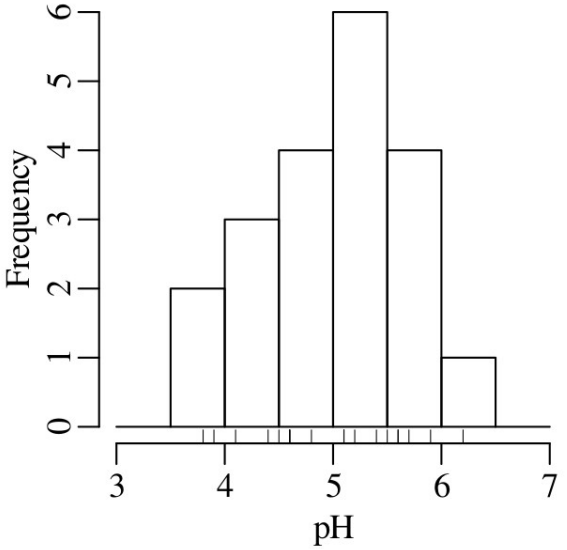
Continuous data

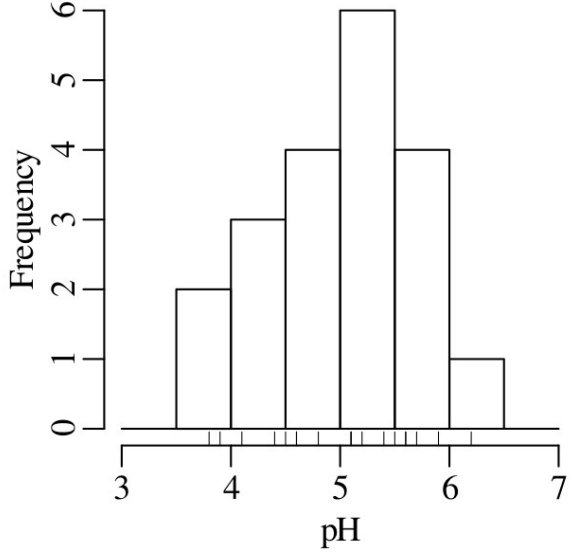


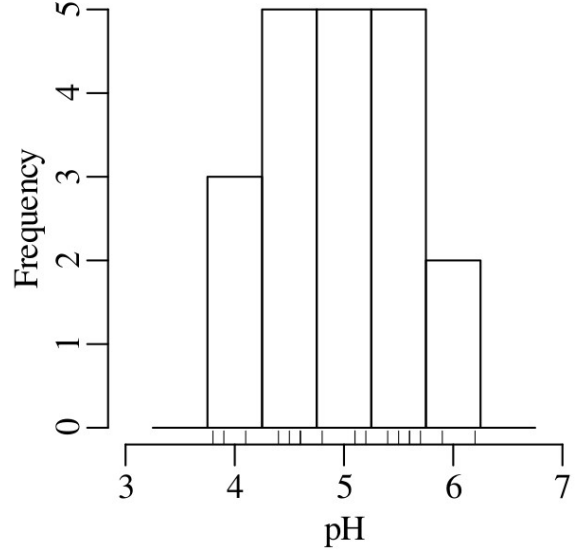
i. How many bins?

ii. Where to place the bins?



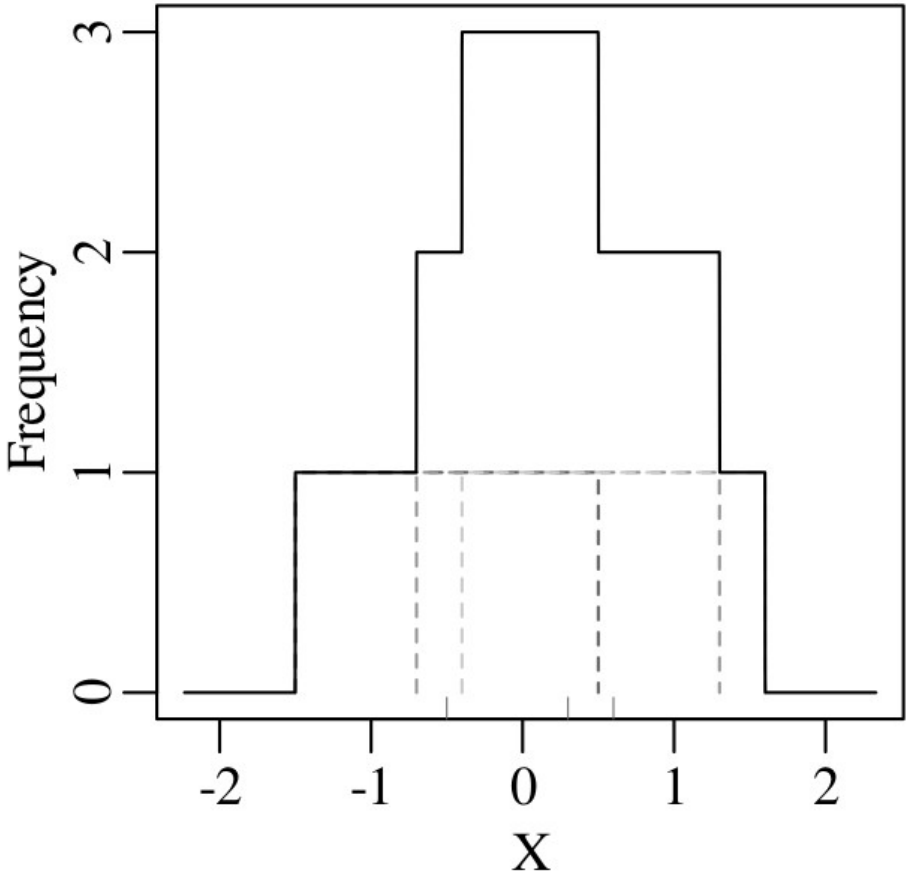


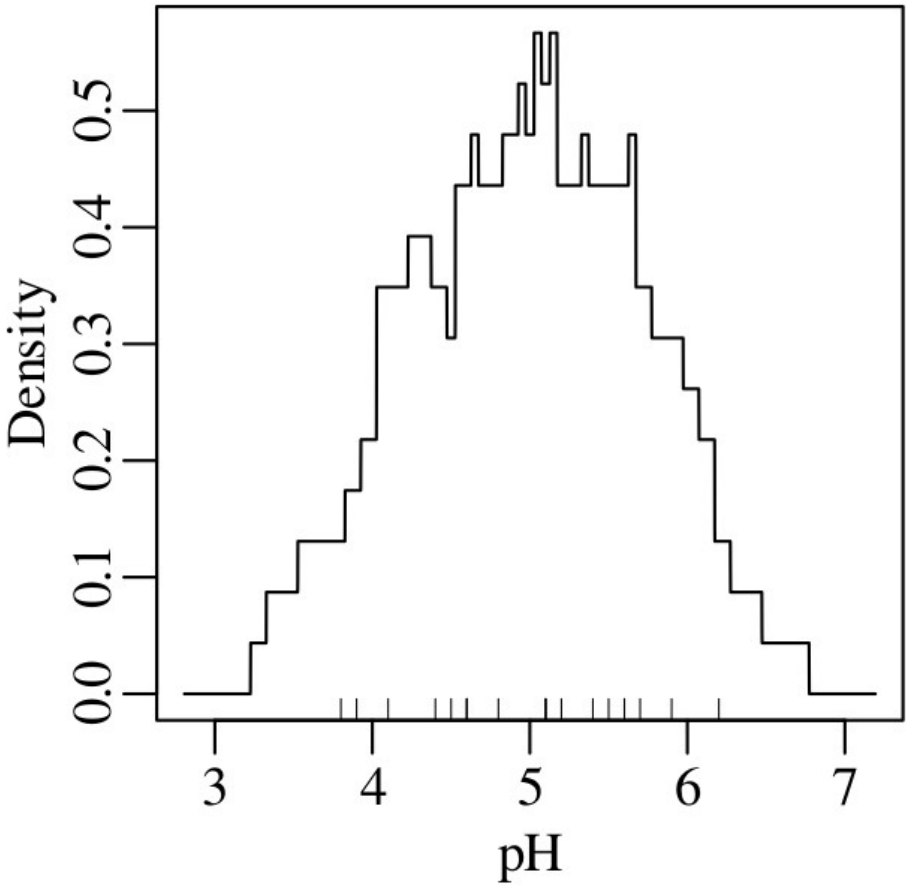




Kernel Density Estimate

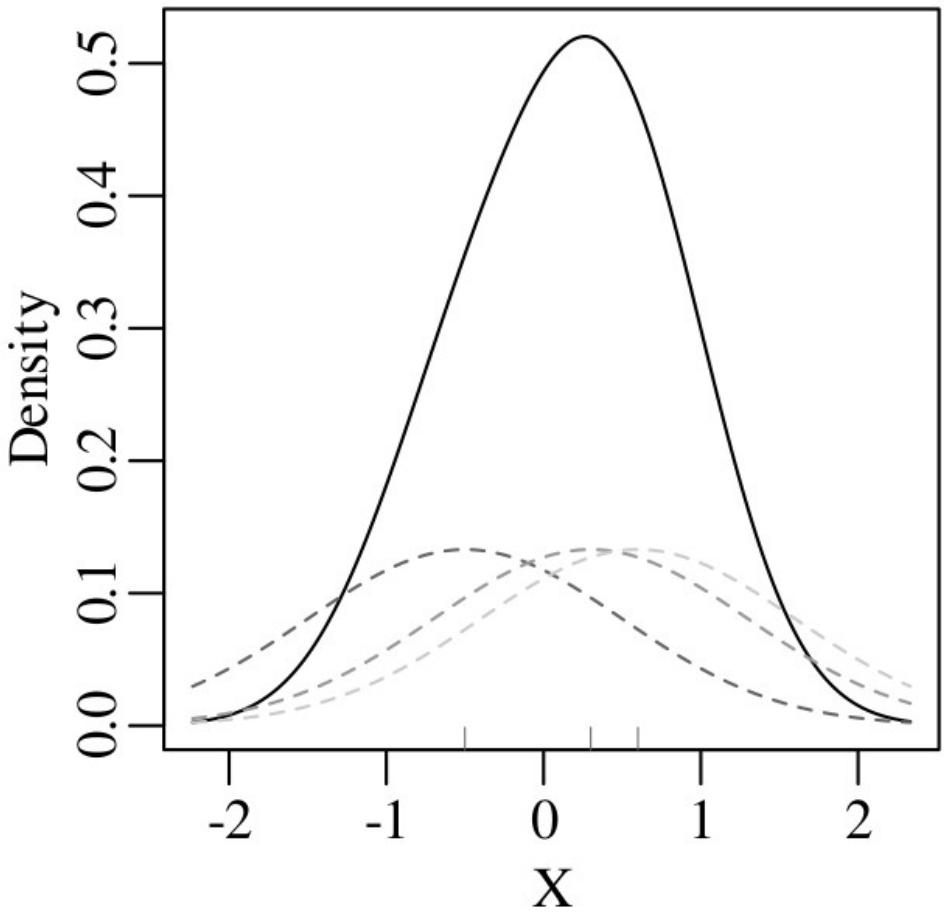
$$KDE(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

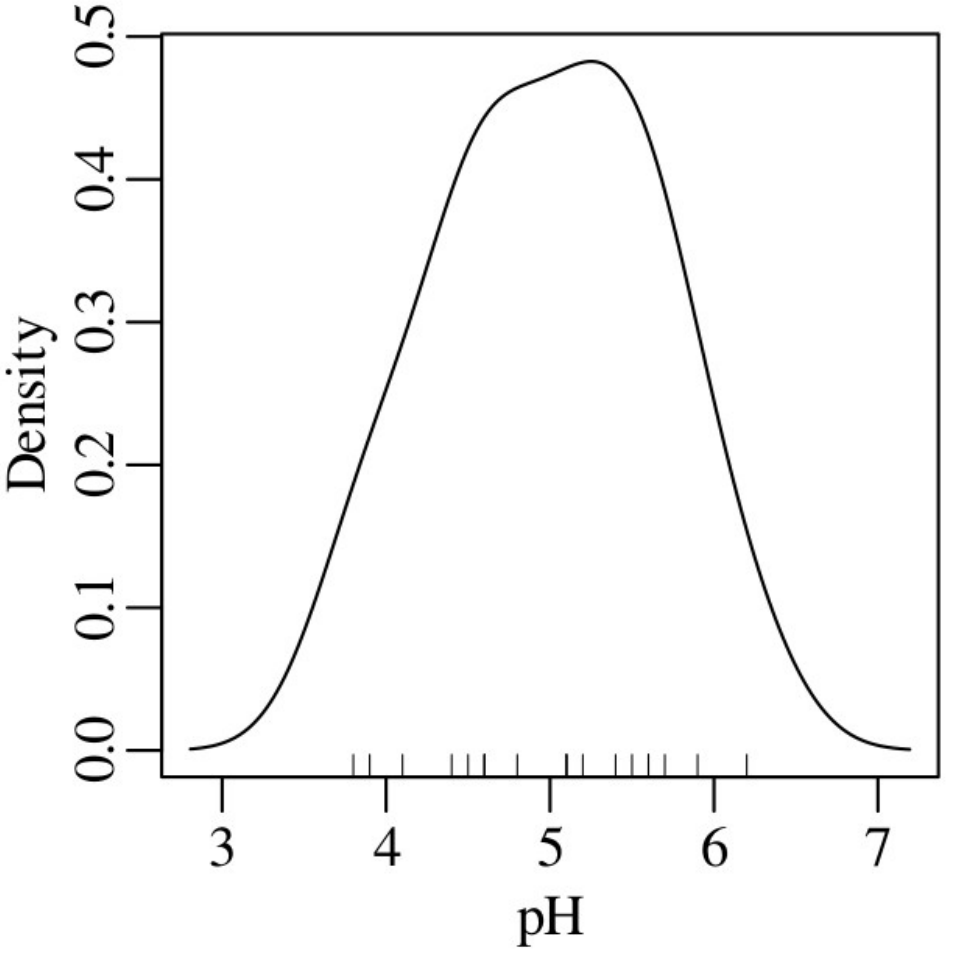




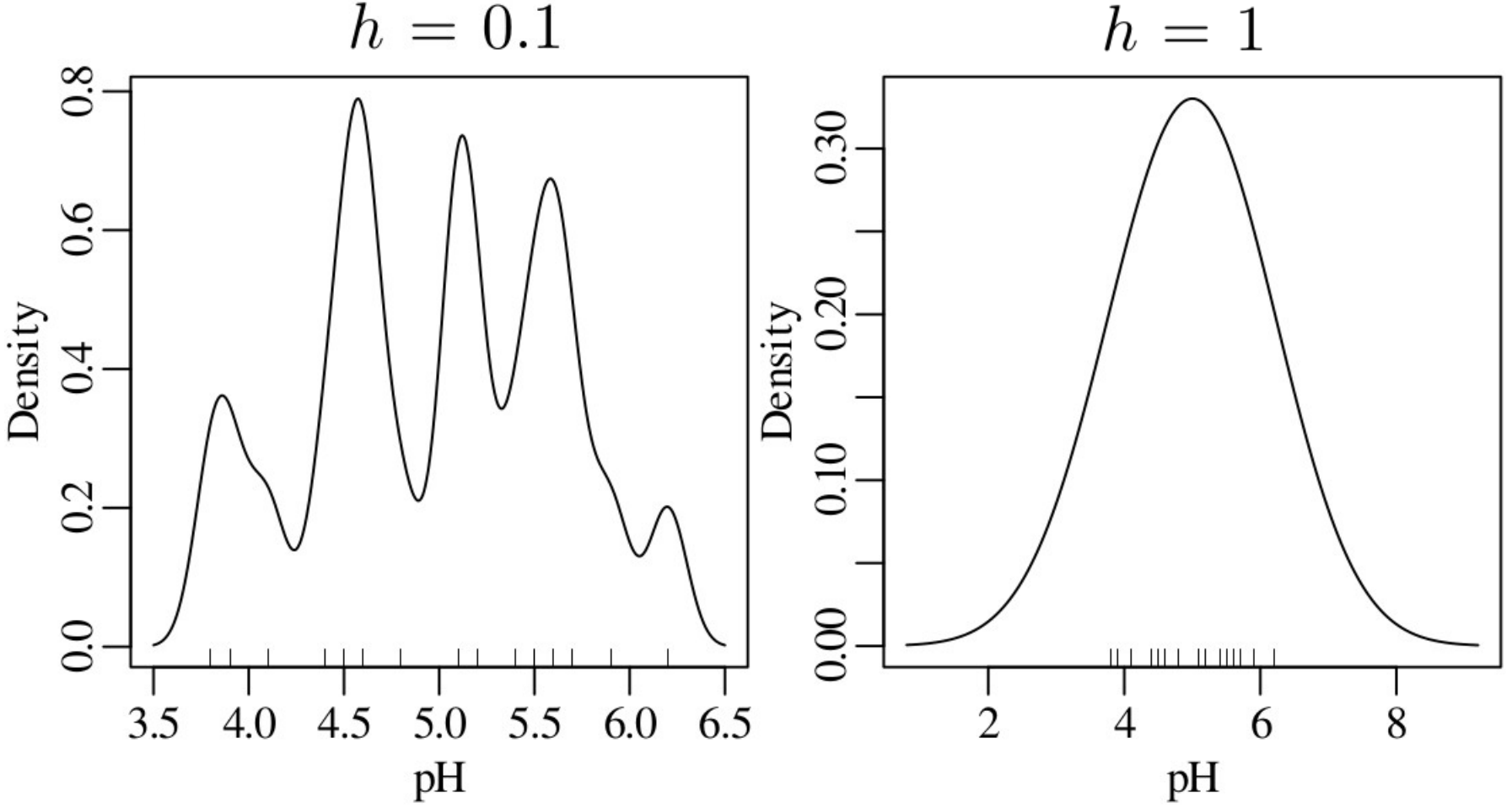
Gaussian kernel:

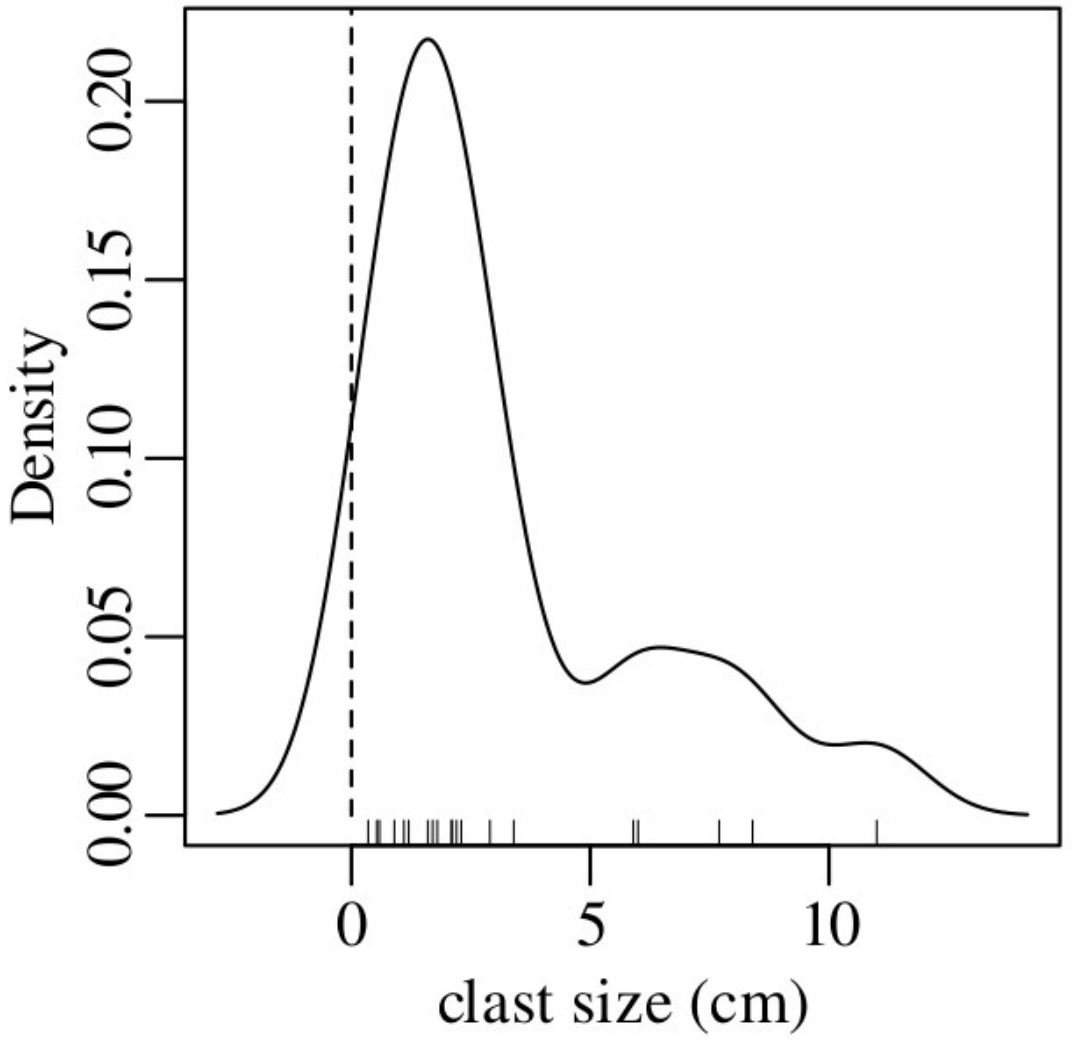
$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{u^2}{2}\right]$$



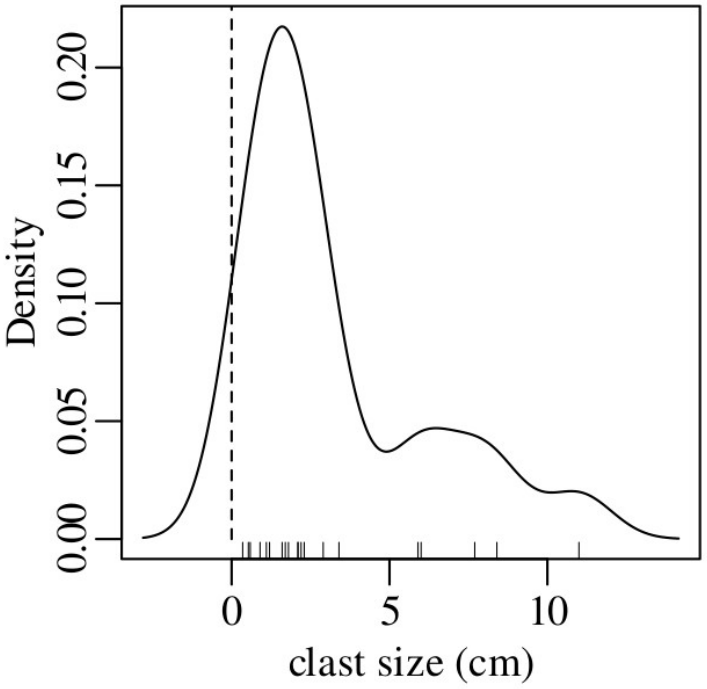


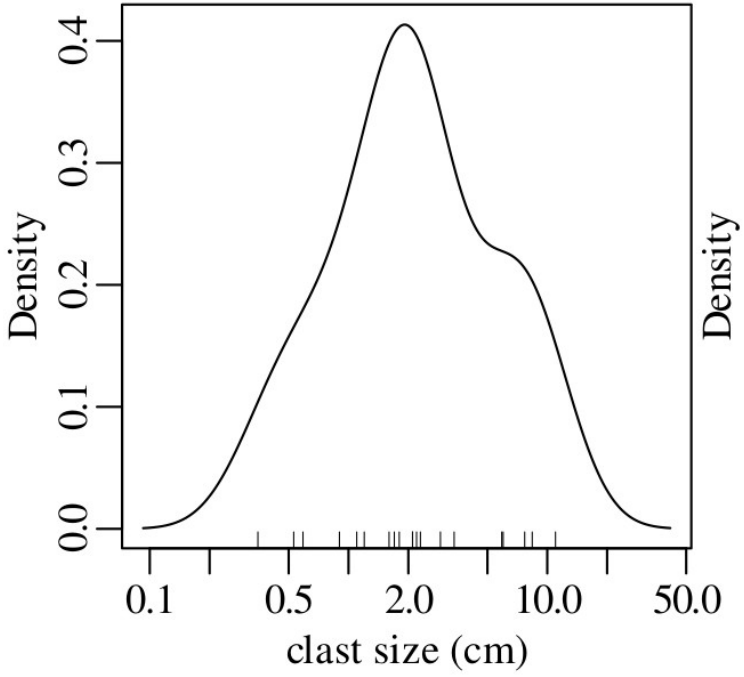
bandwidth

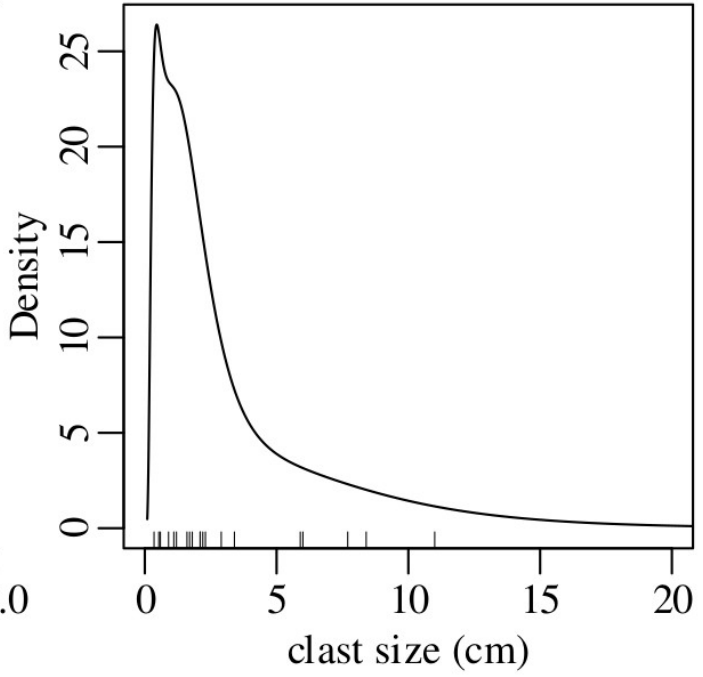

$h = 0.1$
Density
0.0
0.2
0.4
0.6
0.8
3.5
4.0
4.5
5.0
5.5
6.0
6.5
pH
$h = 1$
Density
0.00
0.10
0.20
0.30
2
4
6
8
pH

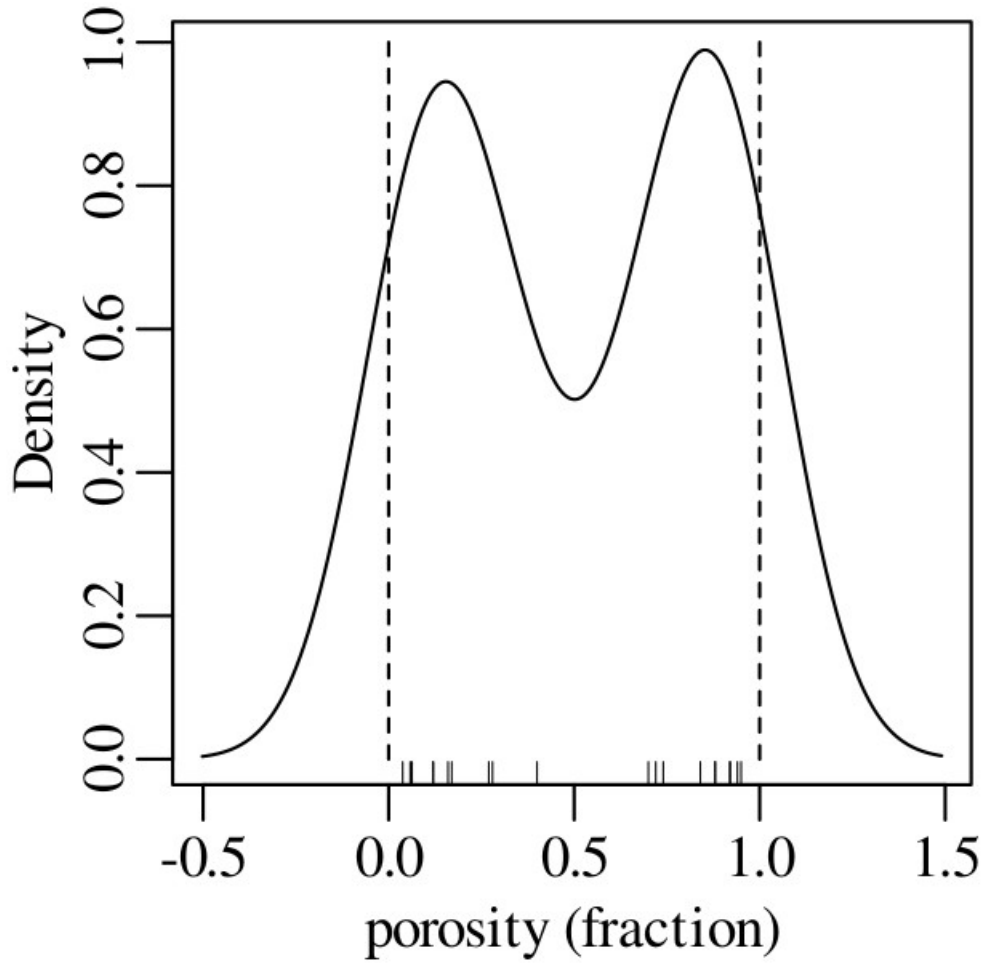

Density
0.00
0.05
0.10
0.15
0.20
0
5
10
clast size (cm)

logarithmic transformation





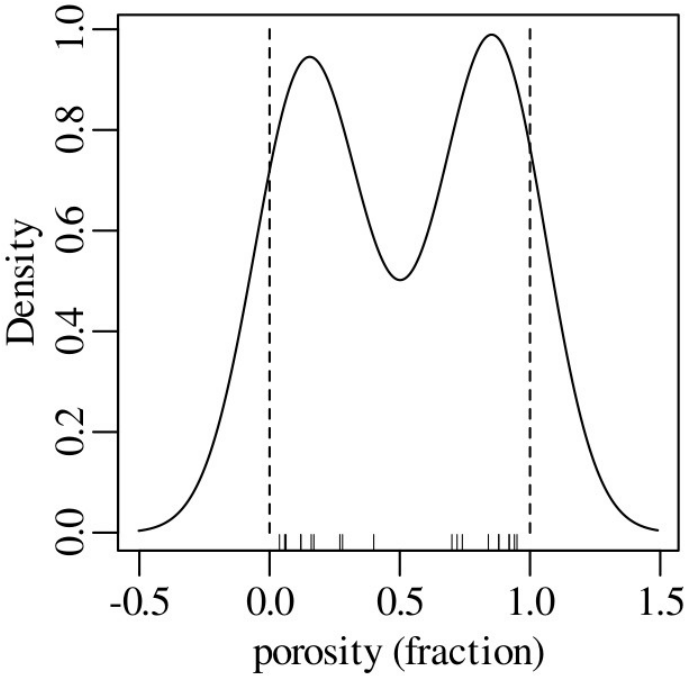


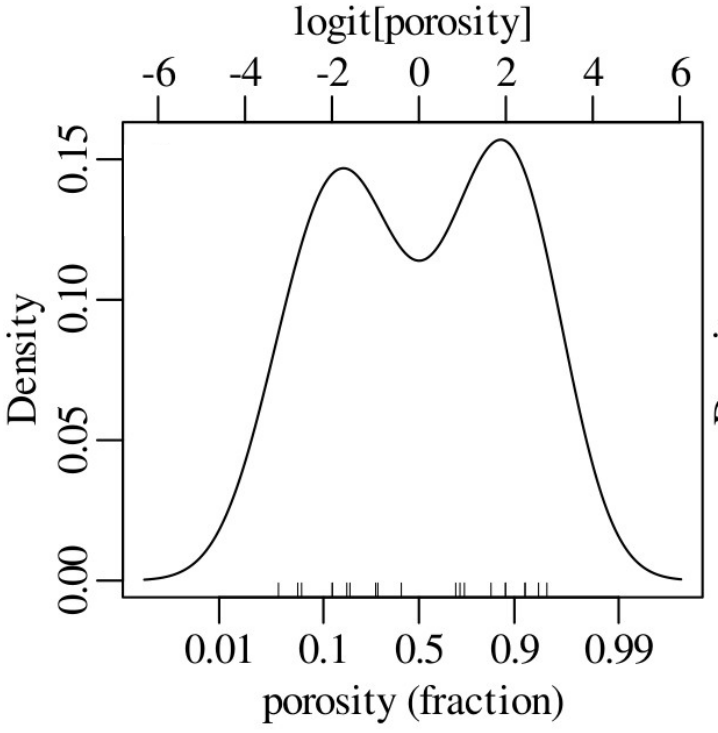

Density
0.0
0.2
0.4
0.6
0.8
1.0
-0.5
0.0
0.5
1.0
1.5
porosity (fraction)

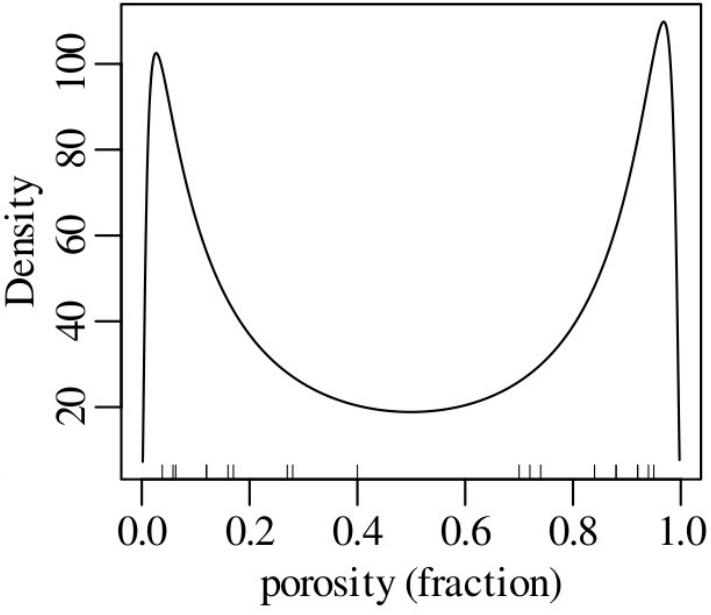
logistic transformation

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

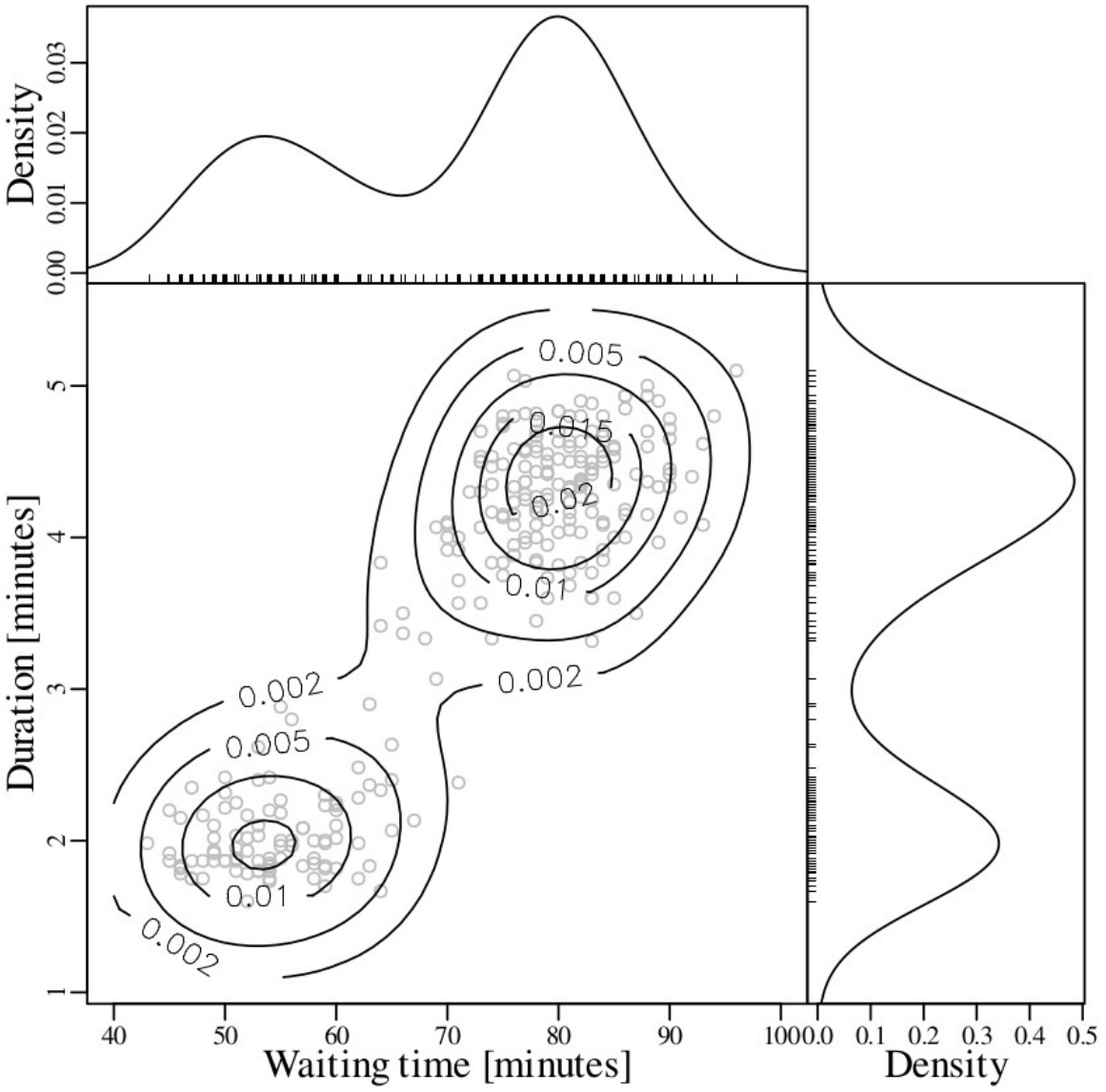
$$x = \text{logit}^{-1}(u) = \frac{\exp[u]}{\exp[u]+1}$$





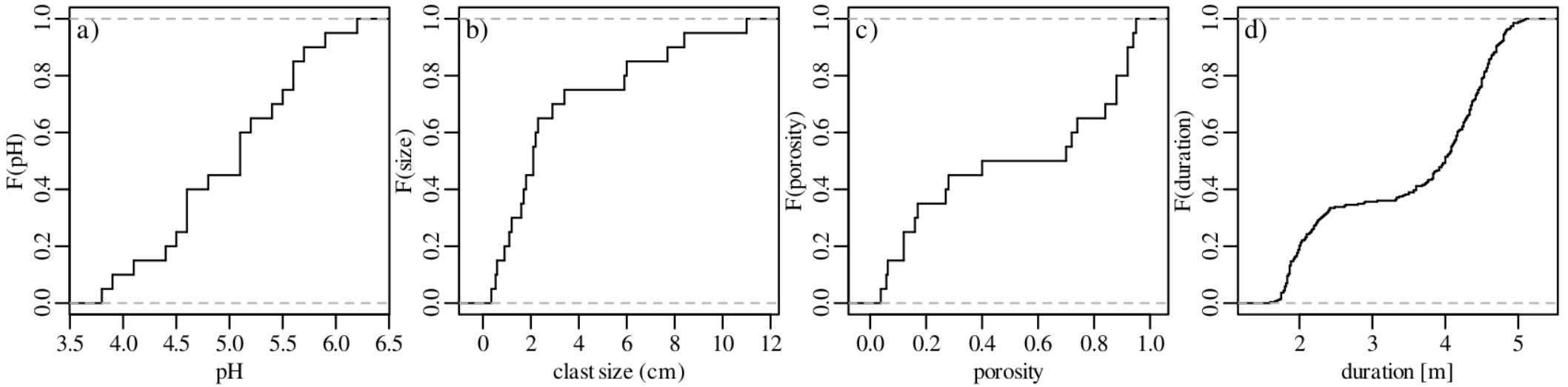


Multivariate distributions



Empirical cumulative distribution fuctions

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



Summary statistics

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

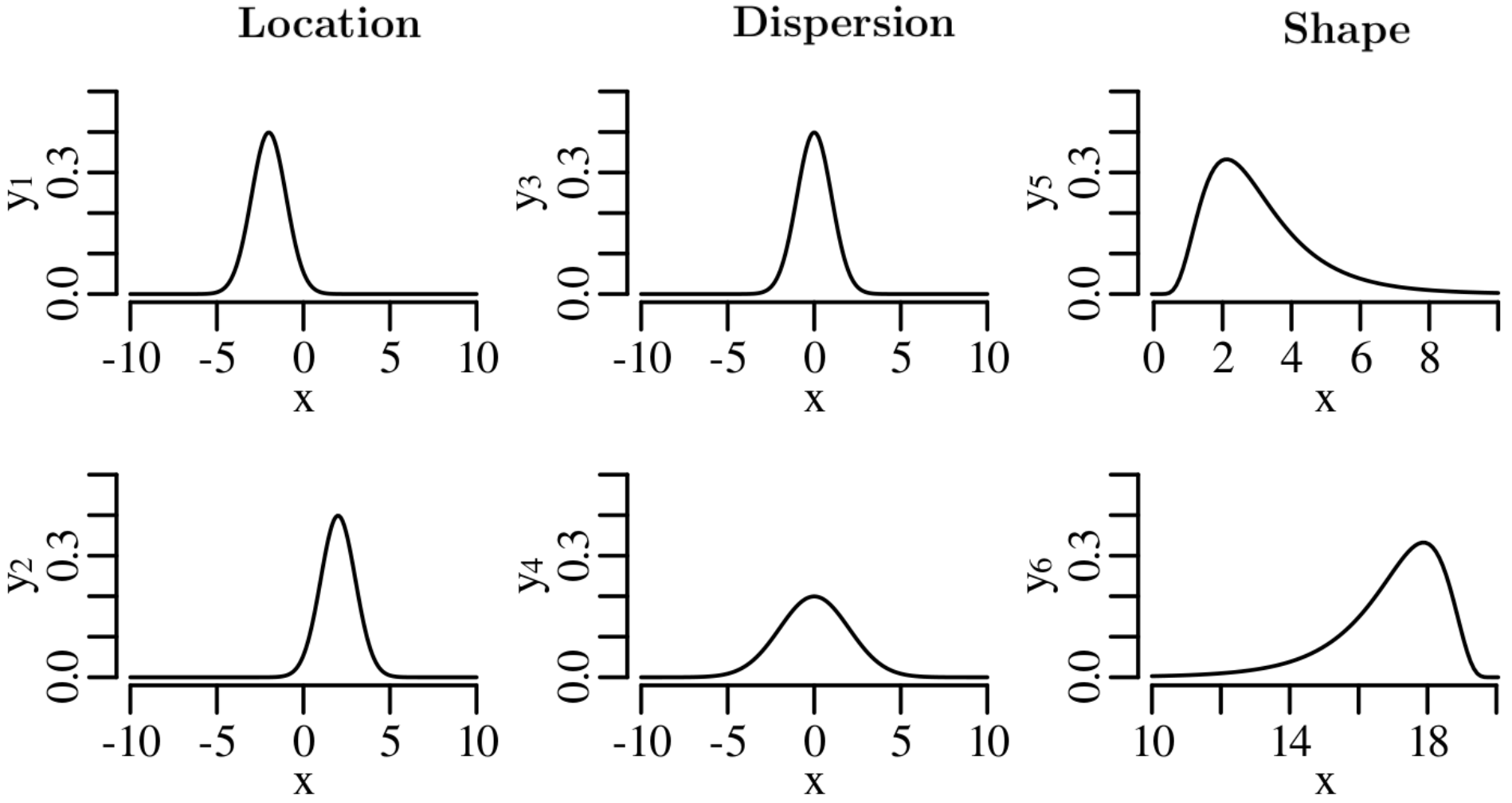
– the variance of x is 11

– the mean of y is 7.50

– the variance of y is 4.125

– the correlation coefficient of x and y is 0.816

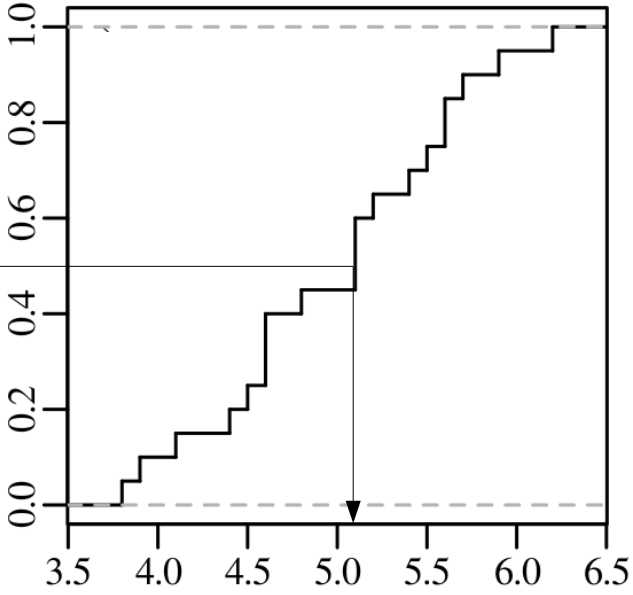
– the best fit line is given by $y = 3.00 + 0.500x$


Location
Dispersion
Shape
y1
y2
y3
y4
y5
y6
0.0
0.3
-10
-5
0
5
10
0
2
4
6
8
10
14
18
x

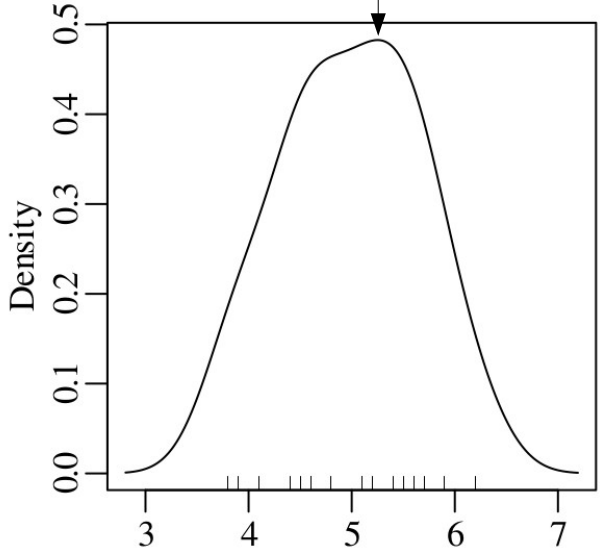
Location

Mean $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Median $\text{med}(x) = x_1 < \ldots < \mathbf{x_{n/2}} < \ldots < x_n$



Mode



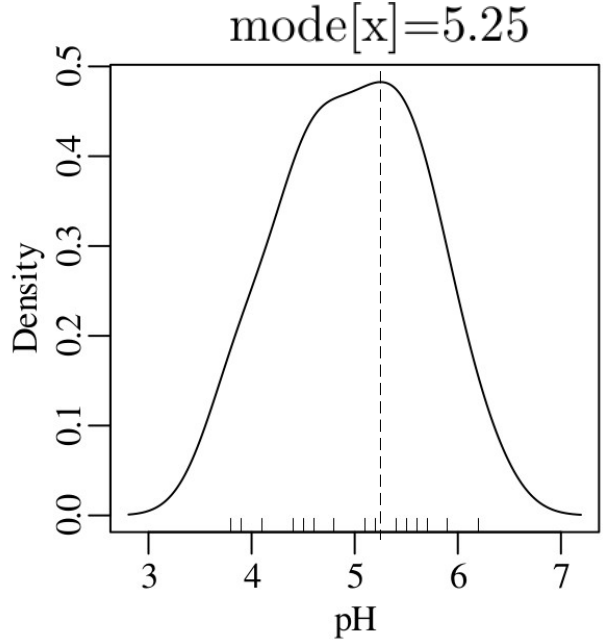
pH data 6.2, 4.4, 5.6, 5.2, 4.5, 5.4, 4.8, 5.9, 3.9, 3.8, 5.1, 4.1, 5.1, 5.5, 5.1, 4.6, 5.7, 4.6, 4.6, 5.6

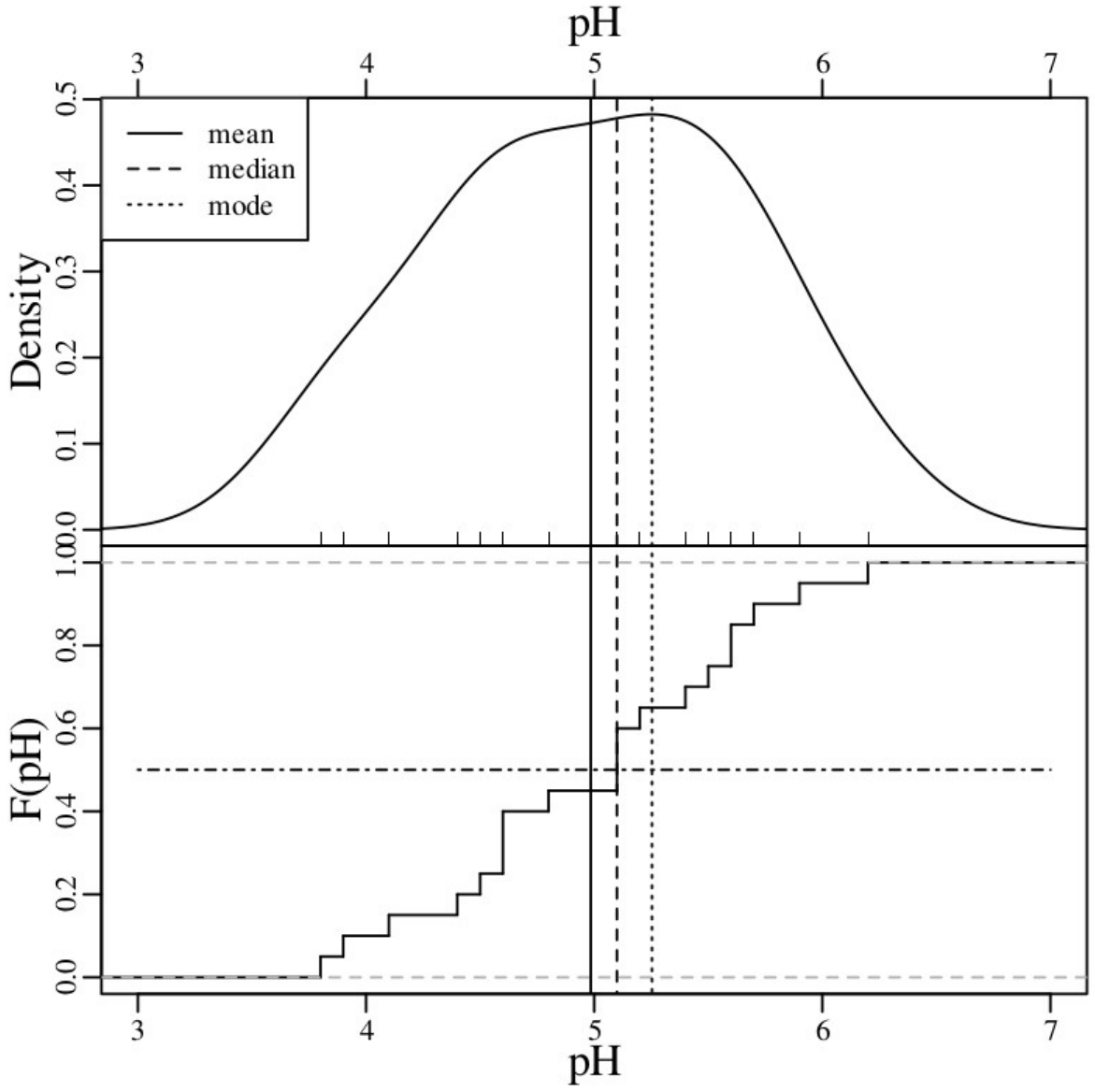
Mean

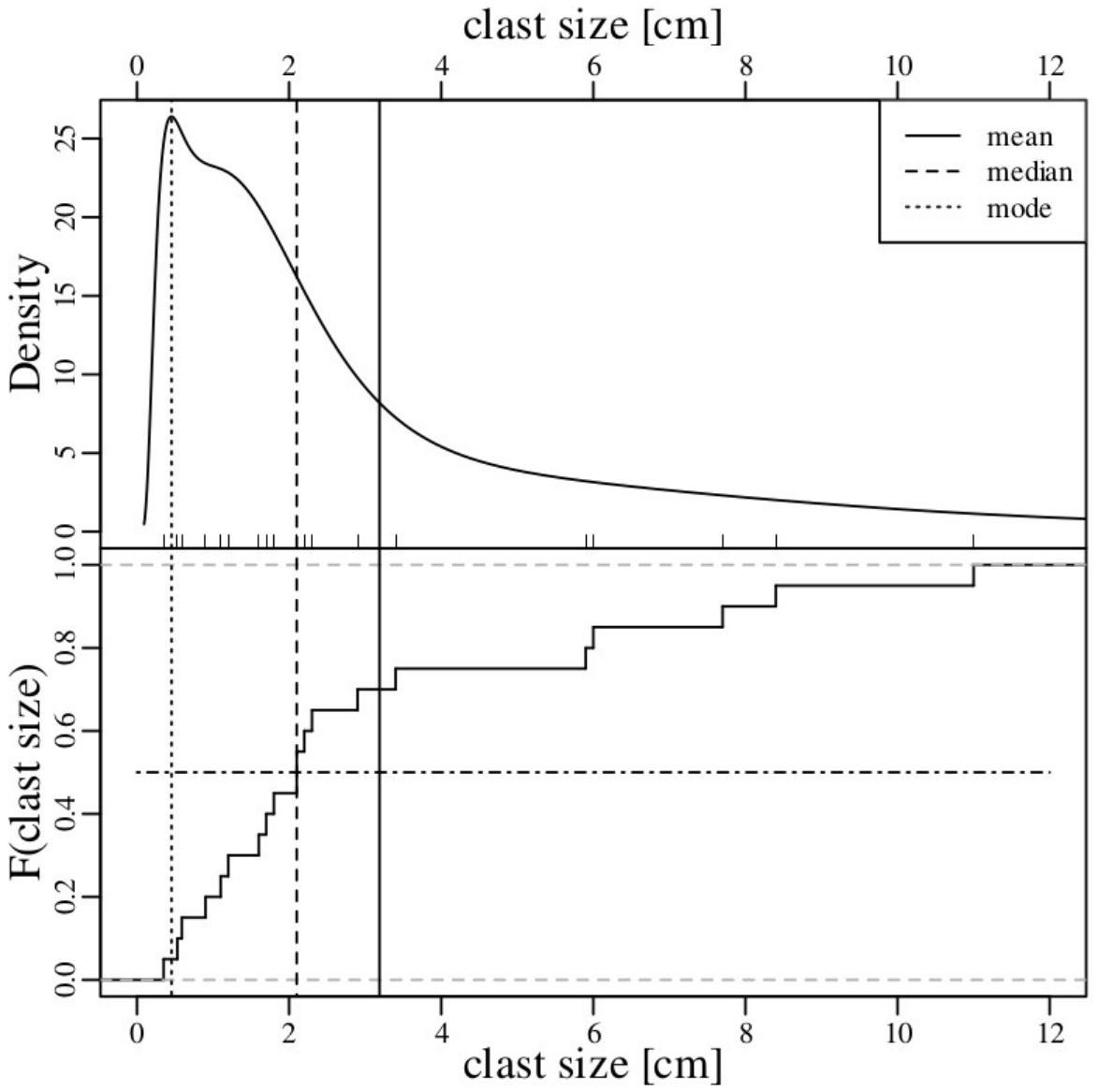
$$\bar{x} = \frac{\begin{array}{c}6.2 + 4.4 + 5.6 + 5.2 + 4.5 + 5.4 + 4.8 + 5.9 + 3.9 + 3.8 + \\ 5.1 + 4.1 + 5.1 + 5.5 + 5.1 + 4.6 + 5.7 + 4.6 + 4.6 + 5.6\end{array}}{20} = 5.00$$

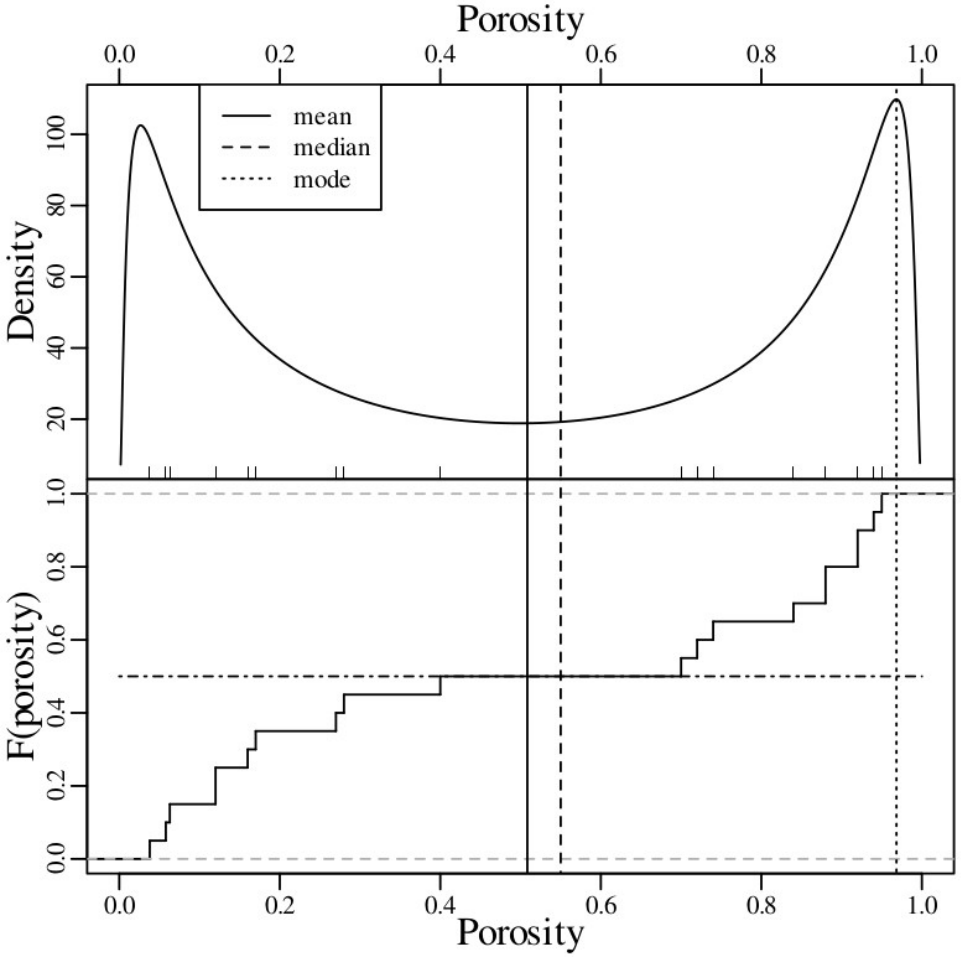
Median 3.8, 3.9, 4.1, 4.4, 4.5, 4.6, 4.6, 4.6, 4.8, **5.1**, **5.1**, 5.1, 5.2, 5.4, 5.5, 5.6, 5.6, 5.7, 5.9, 6.2

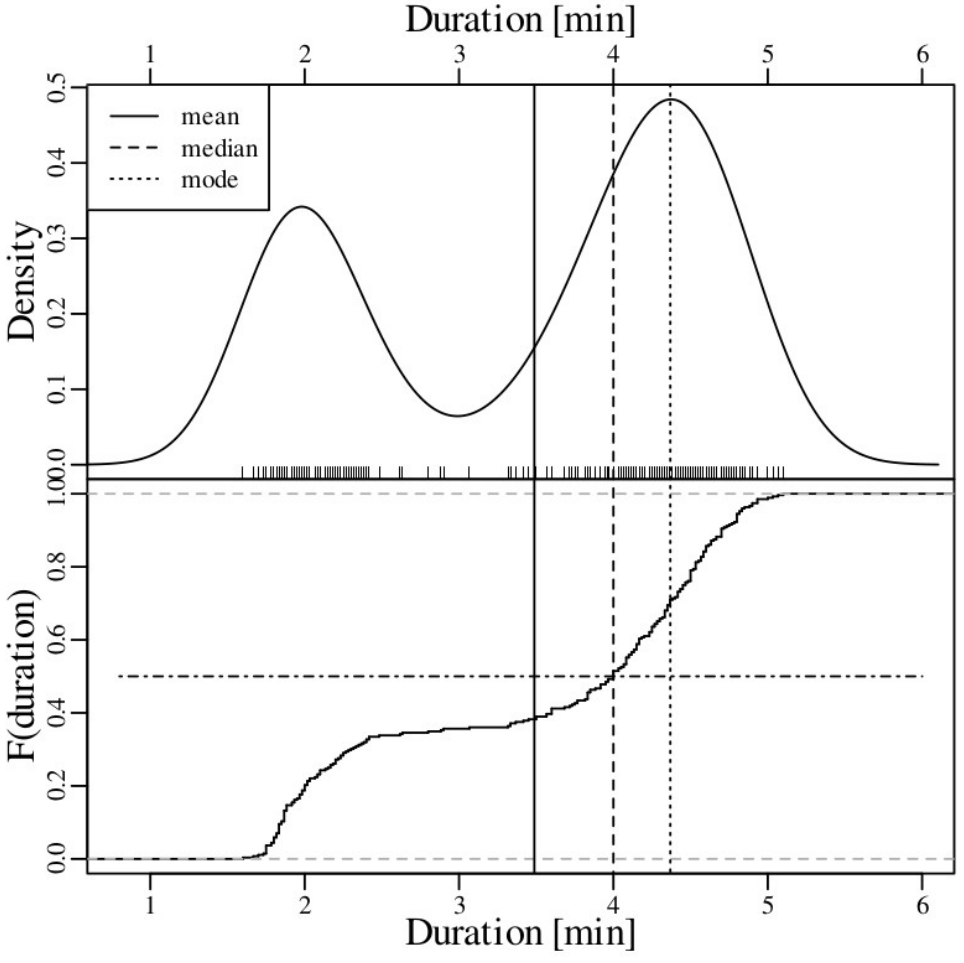
Mode




pH
3
4
5
6
7
mean
median
mode
Density
0.0
0.1
0.2
0.3
0.4
0.5
F(pH)
0.0
0.2
0.4
0.6
0.8
1.0
pH


clast size [cm]
0
2
4
6
8
10
12
Density
0
5
10
15
20
25
mean
median
mode
F(clast size)
0.0
0.2
0.4
0.6
0.8
1.0
clast size [cm]


Porosity
mean
median
mode
Density
F(porosity)
Porosity


Duration [min]
mean
median
mode
Density
F(duration)
Duration [min]

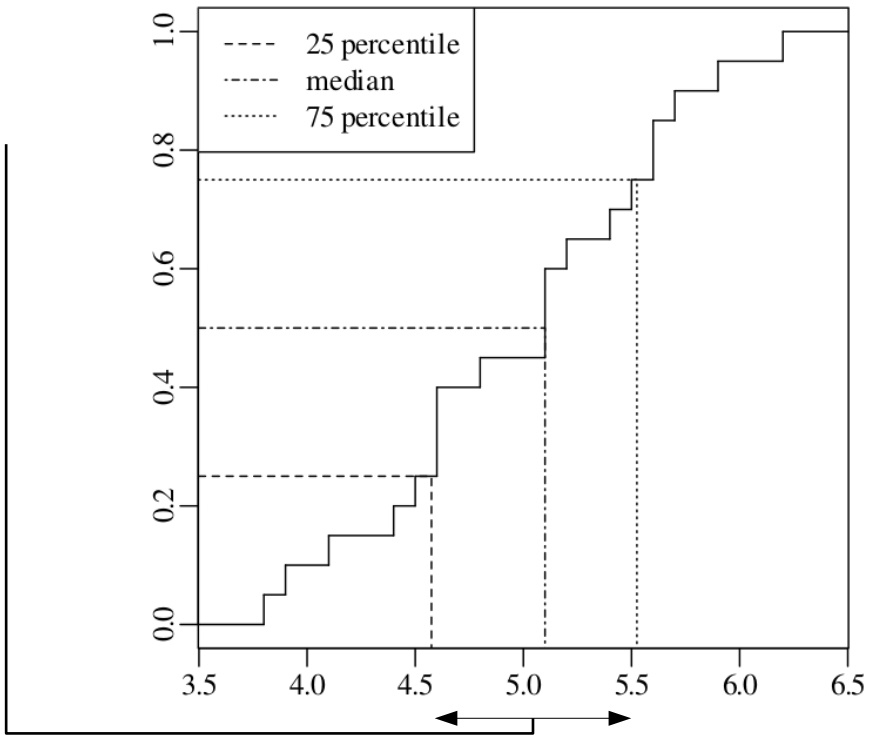
Dispersion

Standard deviation

$$s[x] = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Median Absolute Deviation $\mathrm{MAD} = \mathrm{median}|x_i - \mathrm{median}(x)|$

Interquartile range



Standard deviation

$$s[x] = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_i	6.2	4.4	5.6	5.2	4.5	5.4	4.8	5.9	3.9	3.8	5.1	4.1	5.1	5.5	5.1	4.6	5.7	4.6	4.6	5.6
$(x_i - \bar{x})$	1.20	-.58	.61	.21	-.49	.42	-.19	.92	-1.1	-1.2	.11	-.89	.11	.51	.11	-.39	.71	-.39	-.39	.61
$(x_i - \bar{x})^2$	1.5	.34	.38	.046	.24	.17	.034	.84	1.2	1.4	.013	.78	.013	.27	.013	.15	.51	.15	.15	.38

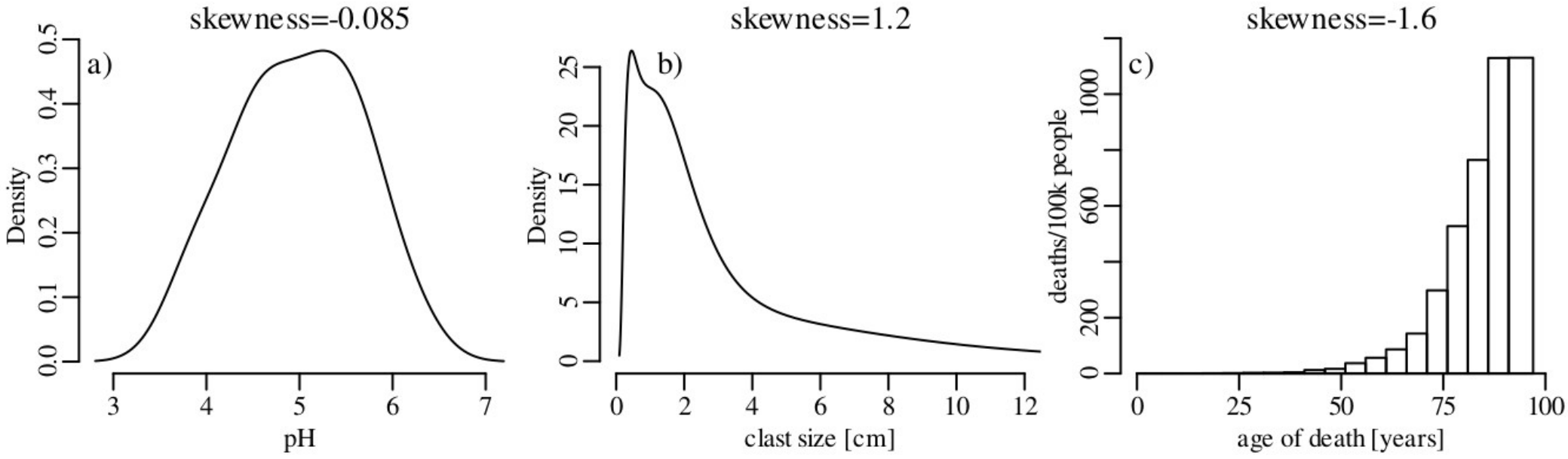
$$\sum_{i=1}^{20}(x_i - \bar{x})^2 = 8.52 \qquad s[x] = \sqrt{8.52/19} = 0.70$$

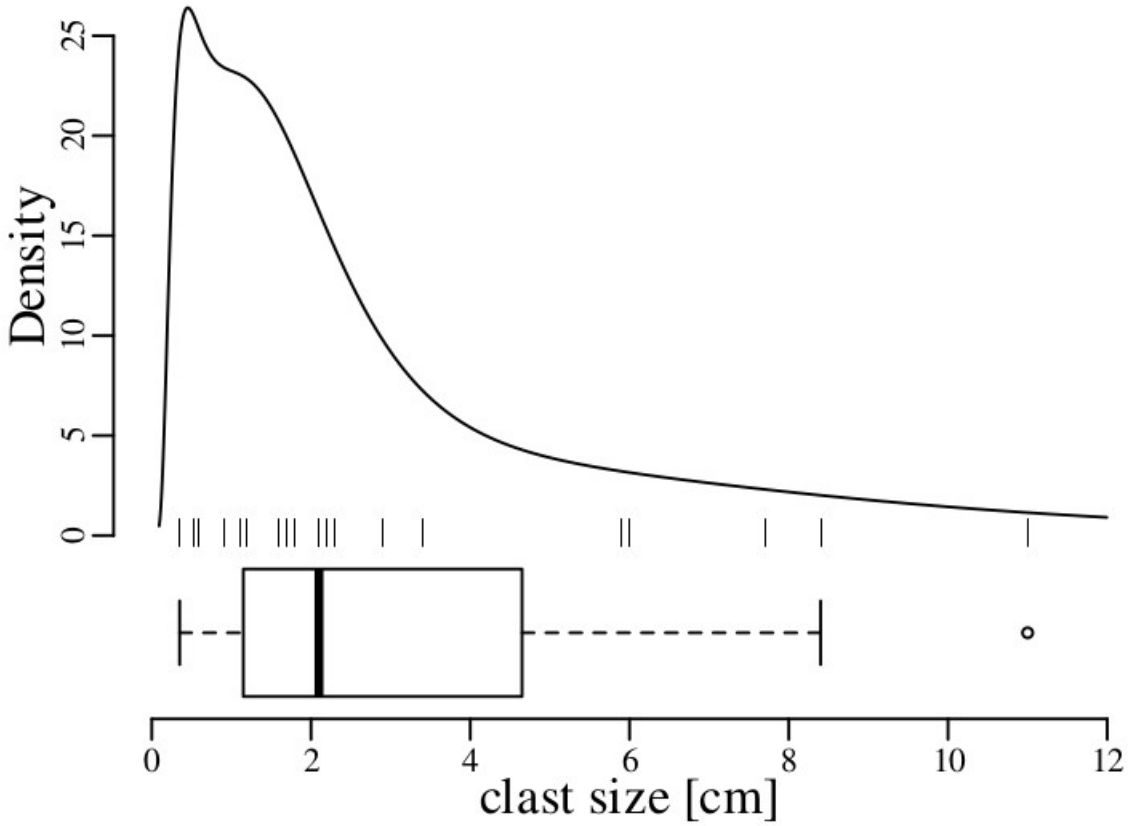
Interquartile range $\mathrm{IQR} = \boxed{5.55 - 4.55 = 1.00}$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_i	3.8	3.9	4.1	4.4	**4.5**	**4.6**	4.6	4.6	4.8	5.1	5.1	5.1	5.2	5.4	**5.5**	**5.6**	5.6	5.7	5.9	6.2
$x_i - \mathrm{med}(x)$	-1.3	-1.2	-1.0	-0.7	-0.6	-0.5	-0.5	-0.5	-0.3	0.0	0.0	0.0	0.1	0.3	0.4	0.5	0.5	0.6	0.8	1.1
$\lvert x_i - \mathrm{med}(x)\rvert$	1.3	1.2	1.0	0.7	0.6	0.5	0.5	0.5	0.3	0.0	0.0	0.0	0.1	0.3	0.4	0.5	0.5	0.6	0.8	1.1
sorted	0.0	0.0	0.0	0.1	0.3	0.3	0.4	0.5	0.5	**0.5**	**0.5**	0.5	0.6	0.6	0.7	0.8	1.0	1.1	1.2	1.3

Median Absolute Deviation $\mathrm{MAD} = \boxed{\mathrm{median}\lvert x_i - \mathrm{median}(x)\rvert}$

skewness $\text{skew}(x) = \frac{1}{n \cdot s[x]^3} \sum_{i=1}^{n} (x_i - \bar{x})^3$




Density
0
5
10
15
20
25
0
2
4
6
8
10
12
clast size [cm]

Statistics for geoscientists

Probability

$$P(A) = \frac{\text{the number of ways } A \text{ can occur}}{\text{the total number of outcomes}}$$

$$P(A) = \frac{\text{the number of ways } A \text{ can occur}}{\text{the total number of outcomes}}$$

$$\frac{\{H\}}{\{H\}\{T\}} = \frac{1}{2} = 0.5$$

$$P(A) = \frac{\text{the number of ways } A \text{ can occur}}{\text{the total number of outcomes}}$$

$$P(2 \times H \cap 1 \times T) = \frac{\{THH\}\{HTH\}\{HHT\}}{\{HHH\}\{THH\}\{HTH\}\{HHT\}\{TTH\}\{THT\}\{HTT\}\{TTT\}} = \frac{3}{8}$$

$$P(A) = \frac{\text{the number of ways } A \text{ can occur}}{\text{the total number of outcomes}}$$

$$P\left(1 \times \text{⚁} \cap 1 \times \text{⚅}\right) = \frac{\{\text{⚁ ⚅}\}\{\text{⚅ ⚁}\}}{\begin{array}{cccccc} \{\text{⚀ ⚀}\} & \{\text{⚁ ⚀}\} & \{\text{⚂ ⚀}\} & \{\text{⚃ ⚀}\} & \{\text{⚄ ⚀}\} & \{\text{⚅ ⚀}\} \\ \{\text{⚀ ⚁}\} & \{\text{⚁ ⚁}\} & \{\text{⚂ ⚁}\} & \{\text{⚃ ⚁}\} & \{\text{⚄ ⚁}\} & \{\text{⚅ ⚁}\} \\ \{\text{⚀ ⚂}\} & \{\text{⚁ ⚂}\} & \{\text{⚂ ⚂}\} & \{\text{⚃ ⚂}\} & \{\text{⚄ ⚂}\} & \{\text{⚅ ⚂}\} \\ \{\text{⚀ ⚃}\} & \{\text{⚁ ⚃}\} & \{\text{⚂ ⚃}\} & \{\text{⚃ ⚃}\} & \{\text{⚄ ⚃}\} & \{\text{⚅ ⚃}\} \\ \{\text{⚀ ⚄}\} & \{\text{⚁ ⚄}\} & \{\text{⚂ ⚄}\} & \{\text{⚃ ⚄}\} & \{\text{⚄ ⚄}\} & \{\text{⚅ ⚄}\} \\ \{\text{⚀ ⚅}\} & \{\text{⚁ ⚅}\} & \{\text{⚂ ⚅}\} & \{\text{⚃ ⚅}\} & \{\text{⚄ ⚅}\} & \{\text{⚅ ⚅}\} \end{array}} = \frac{2}{36} = \frac{1}{18}$$

multiplicative rule

$$P\left(\{2 \times H \cap 1 \times T\} \cap \{1 \times \square \cap 1 \times \square\}\right) = \frac{3}{8}\frac{1}{18} = \frac{3}{144} = 0.021$$

additive rule

$$P\left(\{2 \times H \cap 1 \times T\} \cup \{1 \times \square \cap 1 \times \square\}\right) = \frac{3}{8} + \frac{1}{18} = \frac{31}{72} = 0.43$$

Permutations

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
sampling with replacement							5	2	12	19	10	5	3	19	11	4			

$$\overbrace{n \times n \times \ldots \times n}^{k \text{ times}} = n^k$$

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

sampling with replacement $\boxed{5}$ 2 12 $\boxed{19}$ 10 $\boxed{5}$ 3 $\boxed{19}$ 11 4

$$\overbrace{n \times n \times \ldots \times n}^{k \text{ times}} = n^k$$

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

sampling without replacement 15 8 13 5 4 19 7 1 20 10

$$n \times (n-1) \times (n-2) \times \ldots \times (n-k+1) = \frac{n!}{(n-k)!}$$

what is the probability that two students in a classroom of k

celebrate their birthdays on the same day?

what is the probability that two students in a classroom of k

celebrate their birthdays on the same day?

365^k possible combinations of birthdays (sampling with replacement)

$365!/(365-k)!$ of these combinations do not overlap (sampling without replacement)

what is the probability that two students in a classroom of k

celebrate their birthdays on the same day?

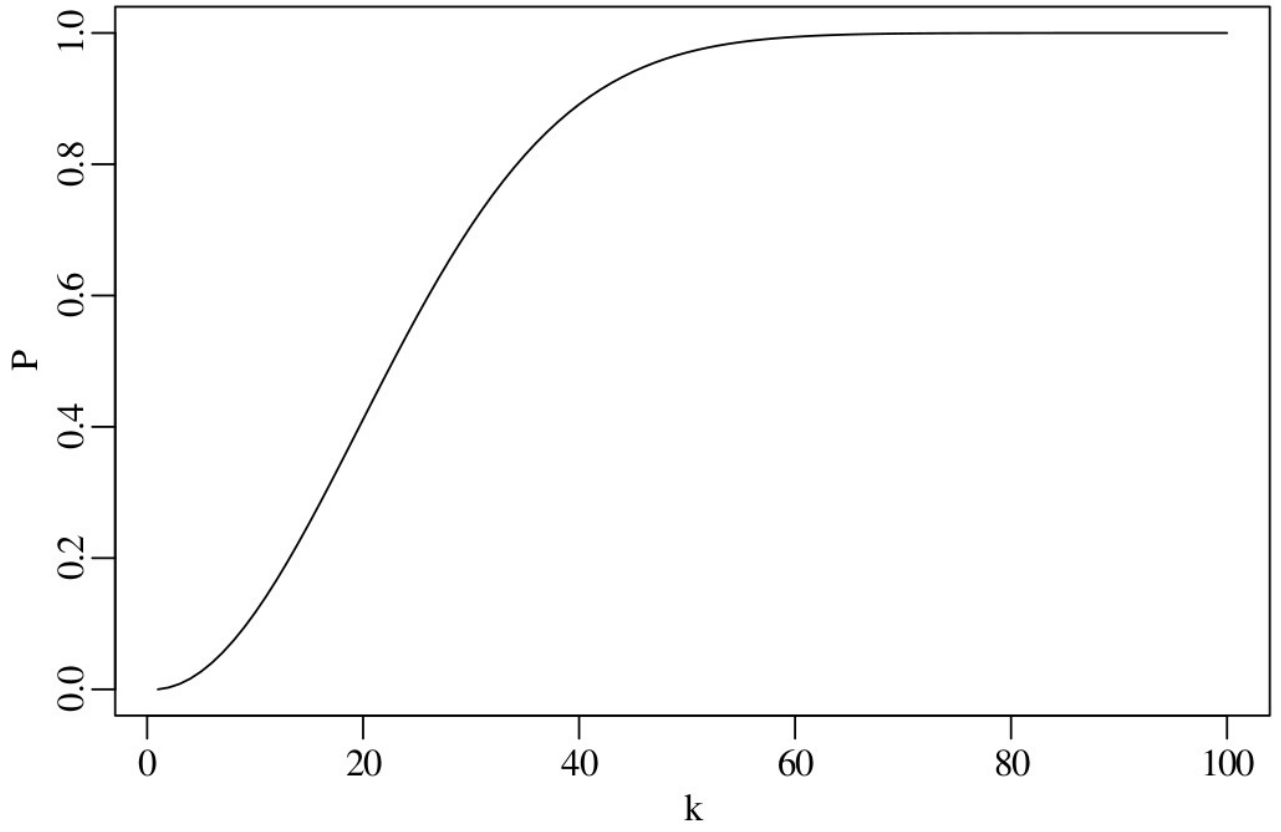
365^k possible combinations of birthdays (sampling with replacement)

$365!/(365-k)!$ of these combinations do not overlap (sampling without replacement)

$$P(\text{no overlapping birthdays}) = \frac{365!}{(365-k)!365^k}$$

$$P(>1 \text{ overlapping birthdays}) = 1 - \frac{365!}{(365-k)!365^k}$$

$$P(> 1 \text{ overlapping birthdays}) = 1 - \frac{365!}{(365-k)!365^k}$$

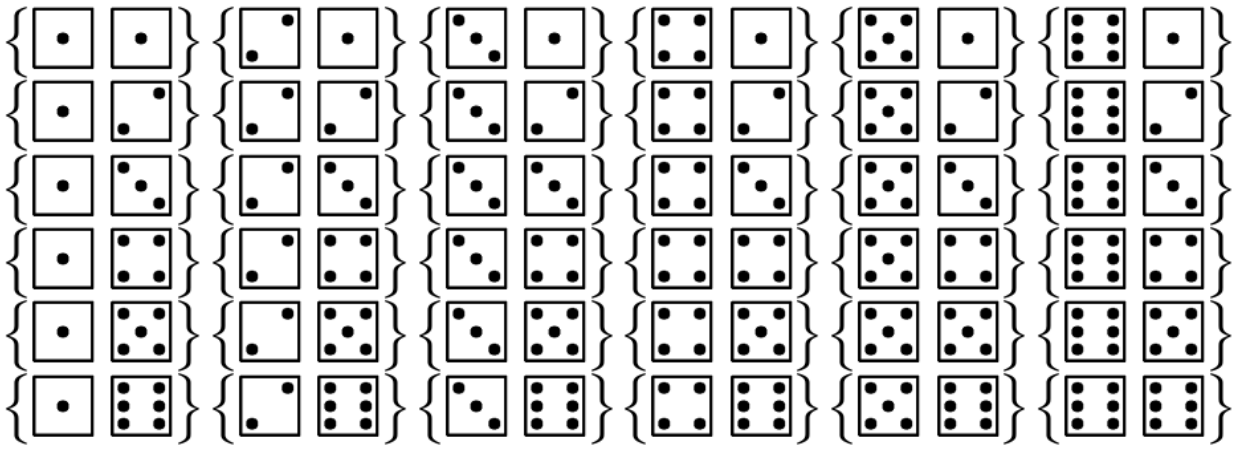


If $k = 23$, then $P(> 1 \text{ overlapping birthdays}) = 0.507$

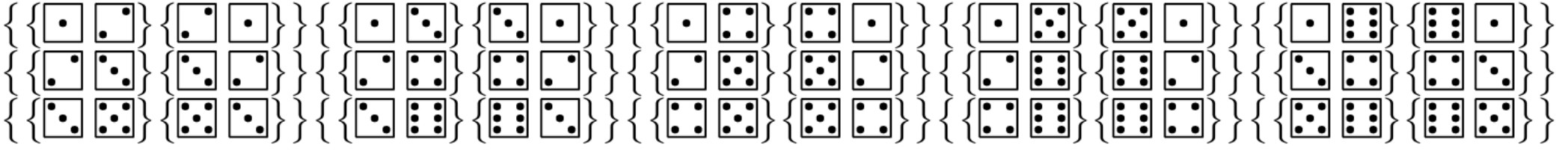
Combinations

$$\{HHH\}\{THH\}\{HTH\}\{HHT\}\{TTH\}\{THT\}\{HTT\}\{TTT\}$$

duplicates: $\{\{HTT\}\{THT\}\{HHT\}\}$ and $\{\{THH\}\{HTH\}\{TTH\}\}$



duplicates:



$$(\#\text{ ordered samples}) = (\#\text{ unordered samples}) \times (\#\text{ ways to order the samples})$$

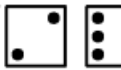
$$(\#\text{ unordered samples}) = \frac{(\#\text{ ordered samples})}{(\#\text{ ways to order the samples})}$$

$$= \frac{n!/(n-k)!\text{ ways to select } k \text{ objects from a collection of } n}{k!\text{ ways to order these } k \text{ objects}} = \frac{n!}{(n-k)!k!}$$

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

$\{THH\}\{HTH\}\{HHT\}$

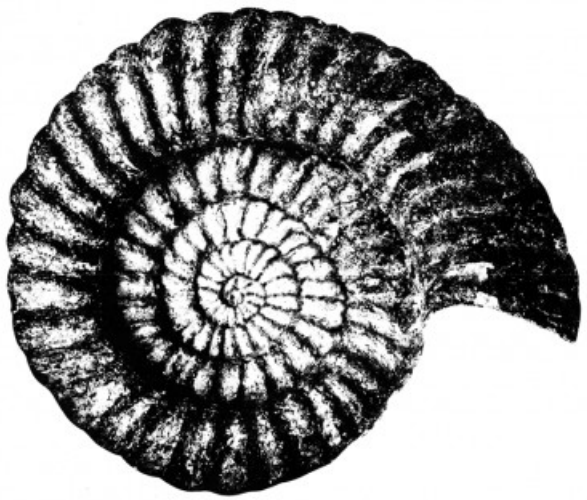
$\{$  $\}\{$  $\}$

$$\binom{3}{2} = \frac{3!}{1!2!} = \frac{6}{2} = 3$$

$$\binom{2}{1} = \frac{2!}{1!1!} = 2$$

Conditional probability

$P(A|B)$ = "The conditional probability of A given B"



JURASSIC			
	LATE	TITHONIAN	144
			151
		KIMMERIDGIAN	154
		OXFORDIAN	159
	MIDDLE	CALLOVIAN	164
		BATHONIAN	169
		BAJOCIAN	176
		AALENIAN	180
	EARLY	TOARCIAN	190
		PLIENSBACHIAN	195
		SINEMURIAN	202
		HETTANGIAN	206

multiplication law

$$P(A|B) = \text{“The conditional probability of } A \text{ given } B\text{”}$$

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) = P(B \cap A)$$

$$\left.\begin{aligned} P(B) &= 0.7 \\ P(A|B) &= 0.2 \end{aligned}\right\} \; 0.7 \times 0.2 = 14\%$$

law of total probability

$$P(A|B) = \text{“The conditional probability of } A \text{ given } B\text{”}$$

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

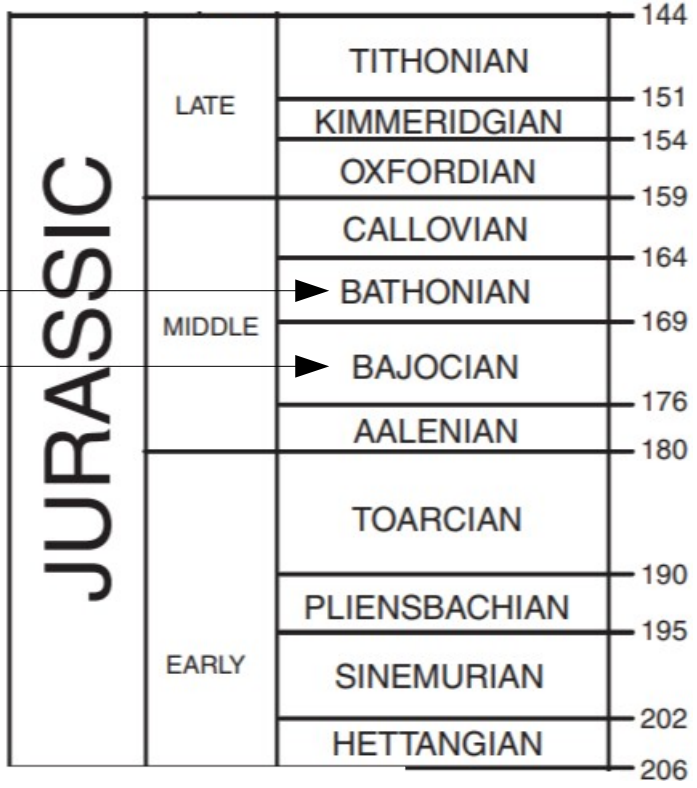
law of total probability

$P(A|B)$ = "The conditional probability of A given B"

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

$$P(B_2) = 0.3$$
$$P(B_1) = 0.7$$

$$P(A|B_2) = 0.5$$
$$P(A|B_1) = 0.2$$



$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) = 0.2 \times 0.7 + 0.5 \times 0.3 = 0.29$$

multiplication law $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) = P(B \cap A)$

Bayes' Rule $P(B|A) = \dfrac{P(A|B)P(B)}{P(A)}$

multiplication law $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) = P(B \cap A)$

law of total probability $P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$

Bayes' Rule $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Bayes' Rule

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Suppose that we have found an ammonite fossil in the river bed. What is its likely age?

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)}$$

$$= \frac{0.2 \times 0.7}{0.2 \times 0.7 + 0.5 \times 0.3} = 0.48$$

The binomial distribution

Bernoulli variable

1. a coin may land on its head (1) or tail (0);
2. a die may land on ⚅ (1) or not (0);
3. a 'wildcat' exploration well may find petroleum (1) or be dry (0).

five gold claims

$$p = 2/3$$

$$P(0 \times \text{gold}) = P(00000) = (1/3)^5 = 0.0041$$

five gold claims $\qquad p = 2/3$

$$P(0 \times \text{gold}) = P(00000) = (1/3)^5 = 0.0041$$

$$P(1 \times \text{gold}) = P(10000) + P(01000) + P(00100) + P(00010) + P(00001)$$

$$P(10000) = (2/3)(1/3)^4 = 0.0082$$
$$P(01000) = (1/3)(2/3)(1/3)^3 = 0.0082$$
$$P(00100) = (1/3)^2(2/3)(1/3)^2 = 0.0082$$
$$P(00010) = (1/3)^3(2/3)(1/3) = 0.0082$$
$$P(00001) = (1/3)^4(2/3) = 0.0082$$

$$P(1 \times \text{gold}) = \binom{5}{1}(2/3)(1/3)^4 = 5 \times 0.0082 = 0.041$$

five gold claims $\qquad p = 2/3$

$$P(0 \times \text{gold}) = P(00000) = (1/3)^5 = 0.0041$$

$$P(1 \times \text{gold}) = \binom{5}{1}(2/3)(1/3)^4 = 5 \times 0.0082 = 0.041$$

$$P(2 \times \text{gold}) = \binom{5}{2}(2/3)^2(1/3)^3 = 10 \times 0.016 = 0.16$$

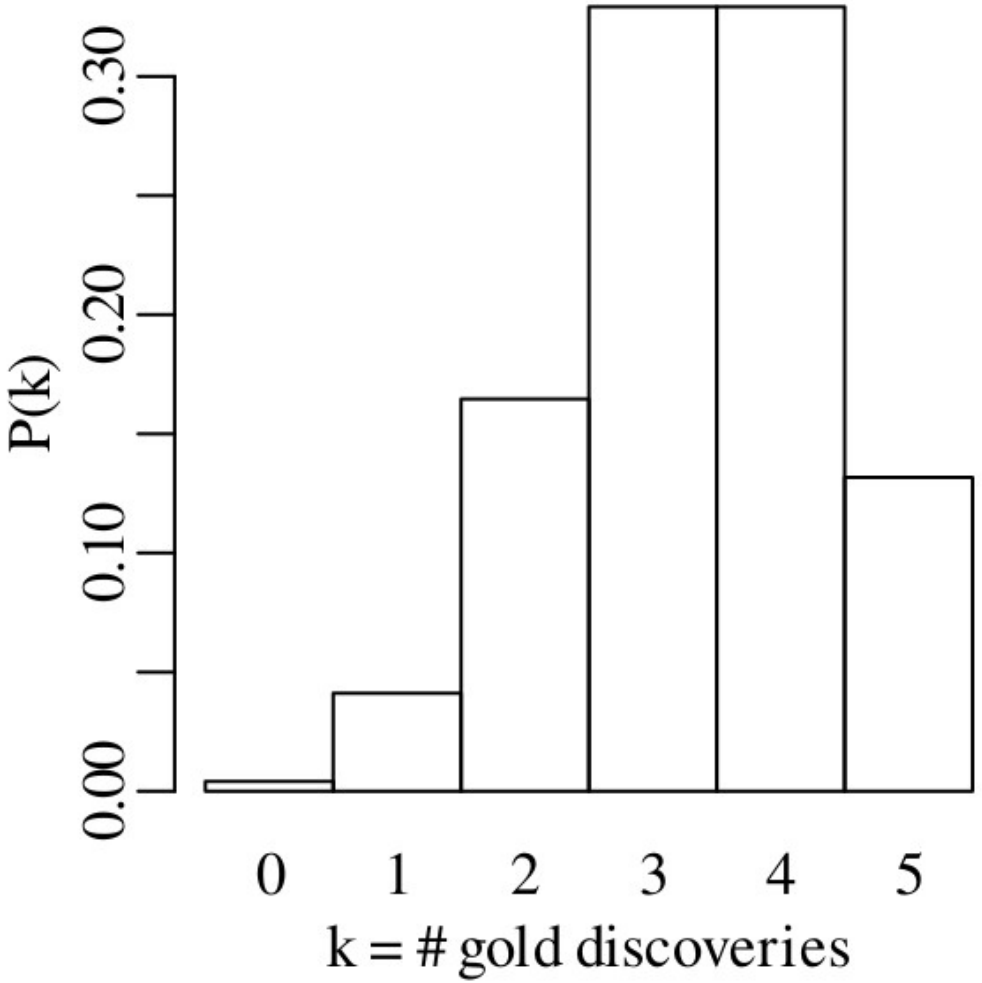
$$P(3 \times \text{gold}) = \binom{5}{3}(2/3)^3(1/3)^2 = 10 \times 0.033 = 0.33$$

$$P(4 \times \text{gold}) = \binom{5}{4}(2/3)^4(1/3) = 5 \times 0.066 = 0.33$$

$$P(5 \times \text{gold}) = (2/3)^5 = 0.13$$

probability mass function (PMF)

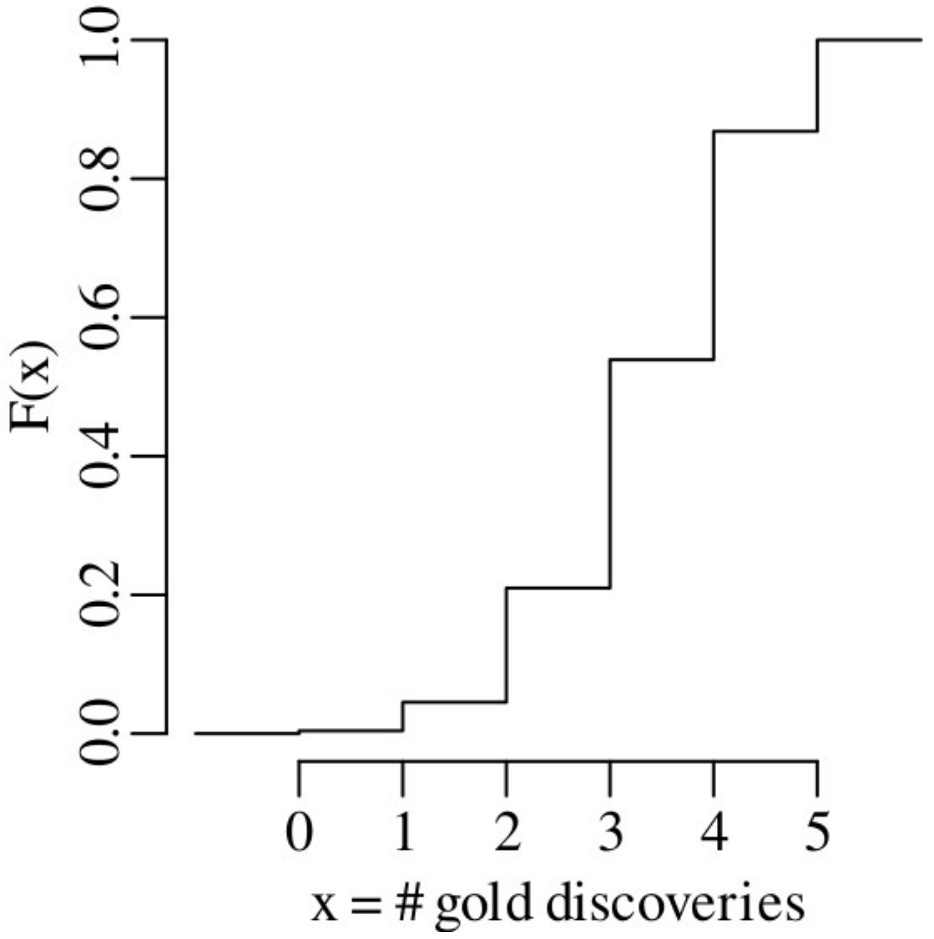
$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$



cumulative distribution function (CDF)

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$F(x) = P(X \leq x)$$



method of maximum likelihood

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathcal{L}(p|n,k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\hat{p} = \frac{k}{n} \qquad \text{e.g.} \qquad \hat{p} = \frac{2}{5} = 0.4$$

Hypothesis tests

$$\hat{p} = 2/5 \quad \longleftrightarrow \quad p = 2/3$$

1. Formulate two hypotheses:

   $H_\circ$ (**null hypothesis**) $\qquad p = 2/3$

   H_a (**alternative hypothesis**): $\qquad p < 2/3$

2. Calculate the **test statistic** T (k)

3. Determine the **null distribution** of T under $H_\circ$.

k	0	1	*2*	3	4	5
$P(T = k)$	0.0041	0.0411	0.1646	0.3292	0.3292	0.1317
$P(T \leq k)$	0.0041	0.0453	*0.2099*	0.5391	0.8683	1.0000

p-value

4. Choose a **significance level** ($\alpha = 0.05$)

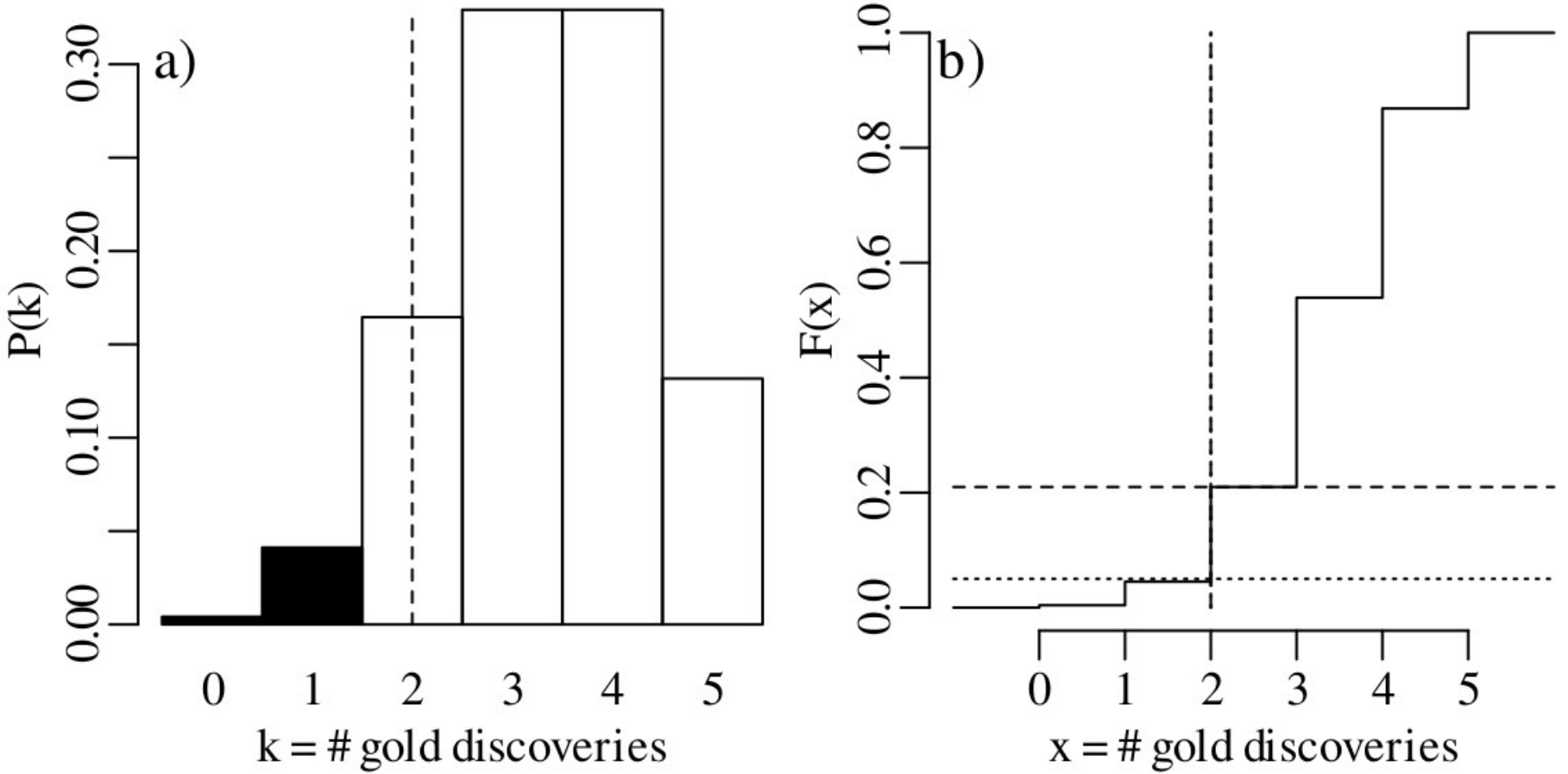
5. Mark the **rejection region**

k	**0**	**1**	*2*	3	4	5
$P(T = k)$	0.0041	0.0411	0.1646	0.3292	0.3292	0.1317
$P(T \leq k)$	**0.0041**	**0.0453**	*0.2099*	0.5391	0.8683	1.0000

rejection region $R = \{0, 1\}$

6. Reach a **decision** $k = 2 \notin R$

7. Alternatively, $0.2099 > \alpha$


a)
P(k)
0.00
0.10
0.20
0.30
0
1
2
3
4
5
k = # gold discoveries
b)
F(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
x = # gold discoveries

$$\hat{p} = 2/5 \quad \longleftrightarrow \quad p = 2/3$$

one-sided hypothesis test

$H_\circ$ (null hypothesis) $\qquad p = 2/3$

H_a (alternative hypothesis): $\qquad p < 2/3$

two-sided hypothesis test

$H_\circ$ (null hypothesis) $\qquad p = 2/3$

H_a (alternative hypothesis): $\qquad p \neq 2/3$

2. Calculate the **test statistic** $T \quad (k)$

3. Determine the **null distribution** of T under $H_\circ$.

k	0	1	*2*	3	4	5
$P(T = k)$	0.0041	0.0411	0.1646	0.3292	0.3292	0.1317
$P(T \leq k)$	0.0041	0.0453	*0.2099*	0.5391	0.8683	1.0000
$P(T \geq k)$	1.000	0.9959	*0.9547*	0.7901	0.4609	0.1317

4. Choose a **significance level** ($\alpha = 0.05$)

but evaluated twice at $\alpha/2$

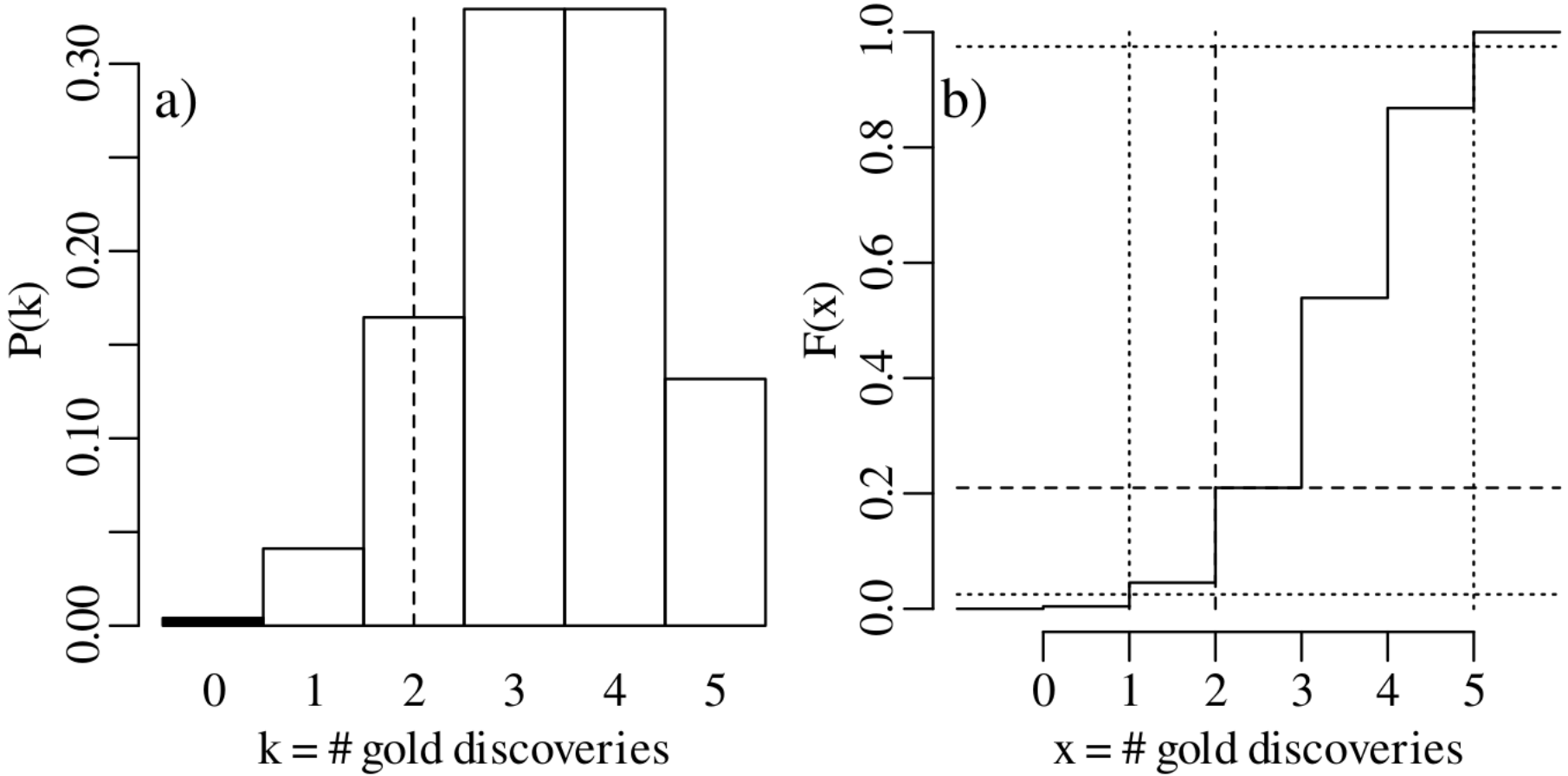
5. Mark the **rejection region**

k	**0**	1	*2*	3	4	5
$P(T = k)$	0.0041	0.0411	0.1646	0.3292	0.3292	0.1317
$P(T \leq k)$	**0.0041**	0.0453	*0.2099*	0.5391	0.8683	1.0000
$P(T \geq k)$	1.000	0.9959	*0.9547*	0.7901	0.4609	0.1317

$$R = \{0\}$$

6. Reach a **decision** $k = 2 \notin R$

7. Alternatively, $2 \times 0.2099 = 0.4198 > 0.05$



Statistical power

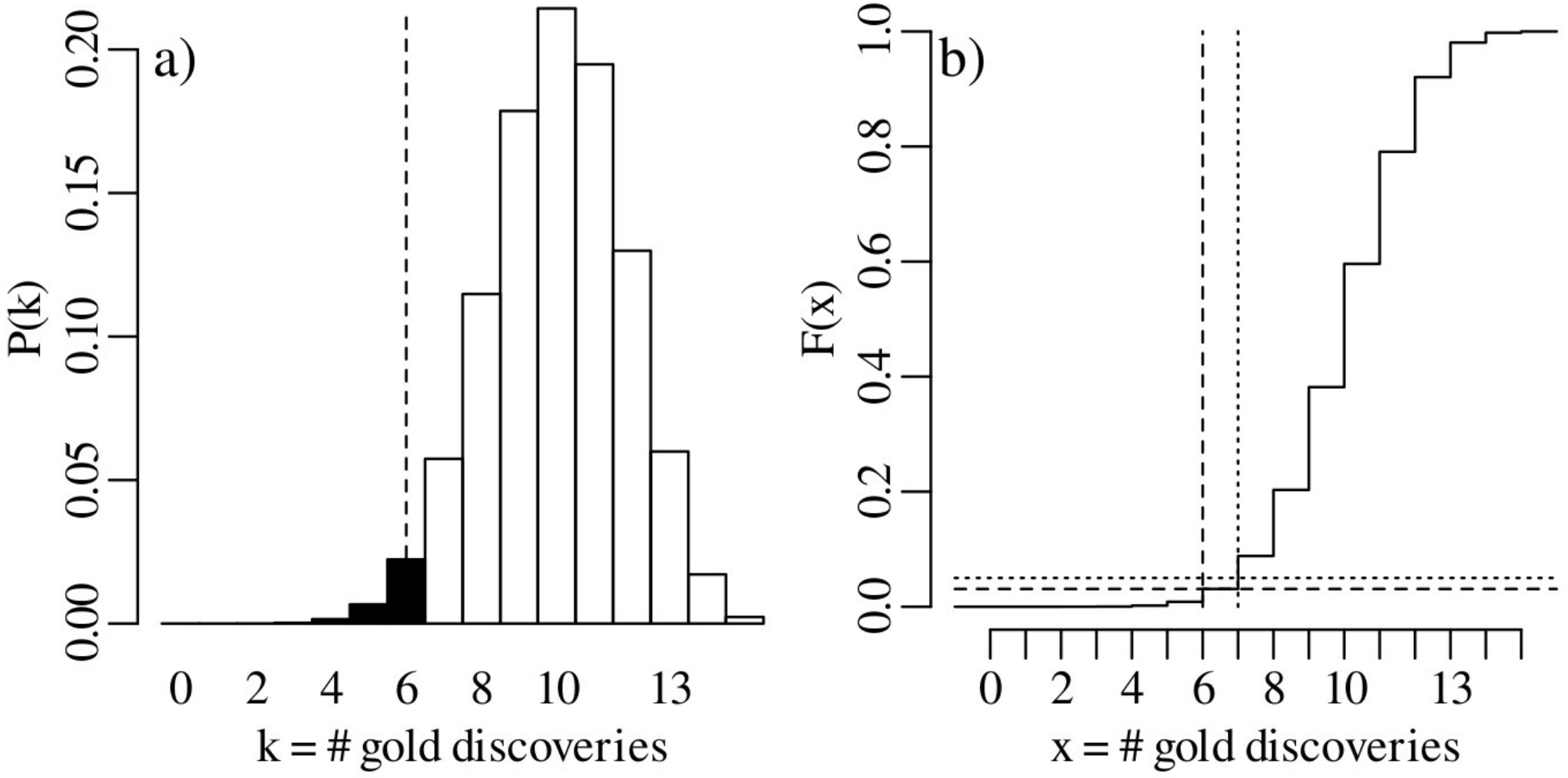
We failed to reject $H_\circ$ for $\hat{p} = \frac{2}{5} = 0.4$. How about $\hat{p} = \frac{6}{15} = 0.4$?

one-sided hypothesis test $\quad (H_\circ : p = 2/3 \text{ vs. } H_a : p < 2/3)$

k	**0**	**1**	**2**	**3**	**4**	**5**	***6***	7
$P(T = k)$	7.0×10^{-8}	2.1×10^{-6}	2.9×10^{-5}	2.5×10^{-4}	0.0015	0.0067	0.0223	0.0574
$P(T \leq k)$	$\mathbf{7.0\times10^{-8}}$	$\mathbf{2.2\times10^{-6}}$	$\mathbf{3.1\times10^{-5}}$	$\mathbf{2.8\times10^{-4}}$	**0.0018**	**0.0085**	***0.0308***	0.0882
k	8	9	10	11	12	13	14	15
$P(T = k)$	0.1148	0.1786	0.2143	0.1948	0.1299	0.0599	0.0171	0.0023
$P(T \leq k)$	0.2030	0.3816	0.5959	0.7908	0.9206	0.9806	0.9977	1.0000

$$R = \{0, 1, 2, 3, 4, 5, 6\}$$

$$k \in R \qquad \text{p-value} = 0.0308 < \alpha$$

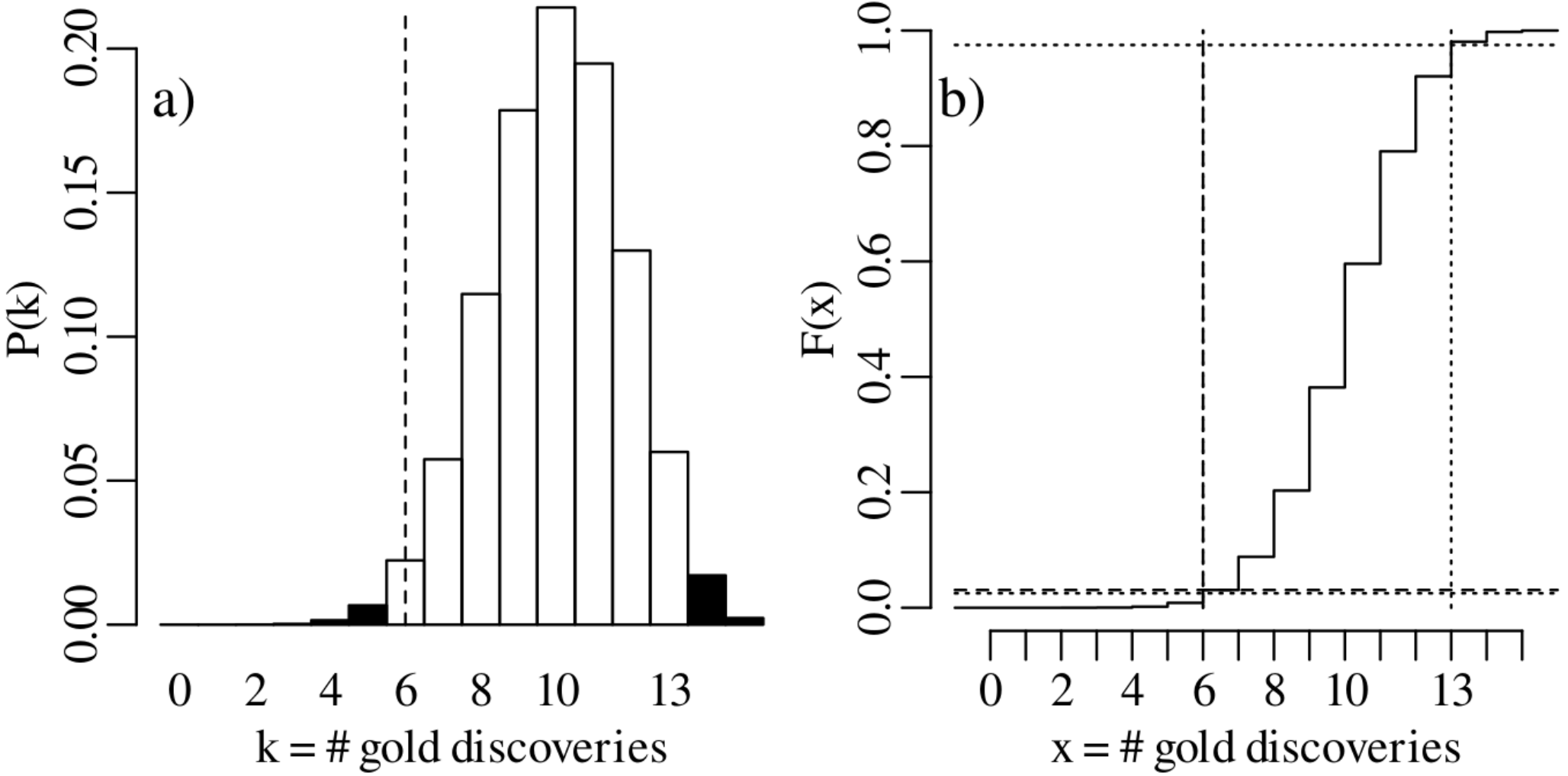

a)
P(k)
0.00
0.05
0.10
0.15
0.20
0
2
4
6
8
10
13
k = # gold discoveries
b)
F(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
13
x = # gold discoveries

two-sided hypothesis test ($H_\circ : p = 2/3$ vs. $H_a : p \neq 2/3$):

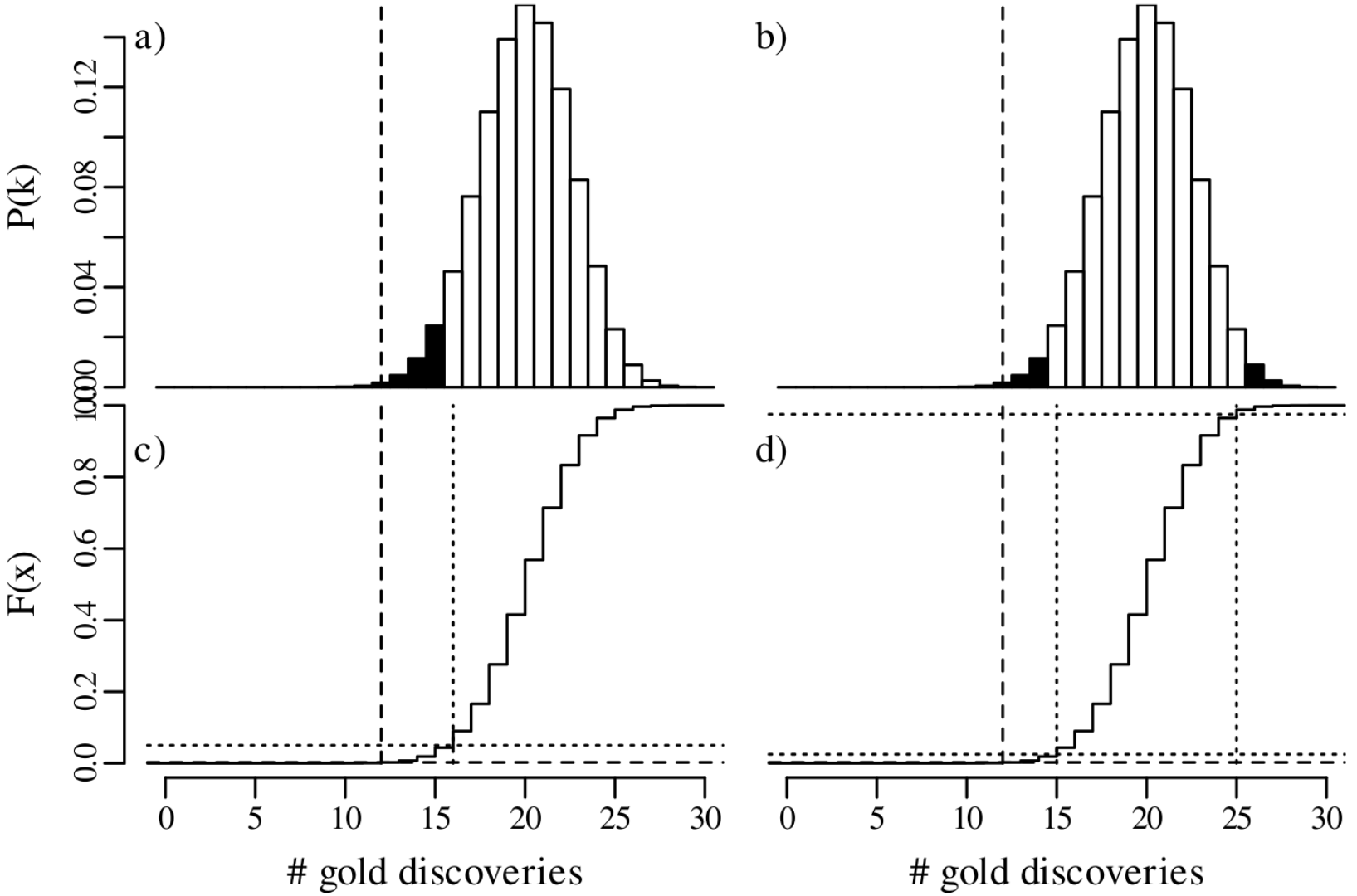
k	**0**	**1**	**2**	**3**	**4**	**5**	*6*	7
$P(T = k)$	7.0×10^{-8}	2.1×10^{-6}	2.9×10^{-5}	2.5×10^{-4}	0.0015	0.0067	0.0223	0.0574
$P(T \leq k)$	$\mathbf{7.0\times10^{-8}}$	$\mathbf{2.2\times10^{-6}}$	$\mathbf{3.1\times10^{-5}}$	$\mathbf{2.8\times10^{-4}}$	**0.0018**	**0.0085**	*0.0308*	0.0882
$P(T \geq k)$	1.0000	1-7.0×10^{-8}	1-2.2×10^{-6}	1-3.1×10^{-5}	1-2.8×10^{-4}	0.9982	0.9915	0.9692
k	8	9	10	11	12	13	**14**	**15**
$P(T = k)$	0.1148	0.1786	0.2143	0.1948	0.1299	0.0599	0.0171	0.0023
$P(T \leq k)$	0.2030	0.3816	0.5959	0.7908	0.9206	0.9806	0.9977	1.0000
$P(T > k)$	0.9118	0.7970	0.6184	0.4041	0.2092	0.0794	**0.0194**	**0.0023**

$$R = \{0, 1, 2, 3, 4, 5, 14, 15\}$$

$k \notin R$ p-value $= (2 \times 0.0308 =) 0.0616 > \alpha$


a)
P(k)
0.00
0.05
0.10
0.15
0.20
0
2
4
6
8
10
13
k = # gold discoveries
b)
F(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
13
x = # gold discoveries

Next, how about $\hat{p} = \frac{12}{30} = 0.4$?



The binomial distribution

$H_\circ$ is ...	false	true
rejected	correct decision	Type-I error
not rejected	Type-II error	correct decision

the accused is ...	guilty	innocent
sentenced	correct decision	Type-I error
acquitted	Type-II error	correct decision

depends on the **confidence level** α

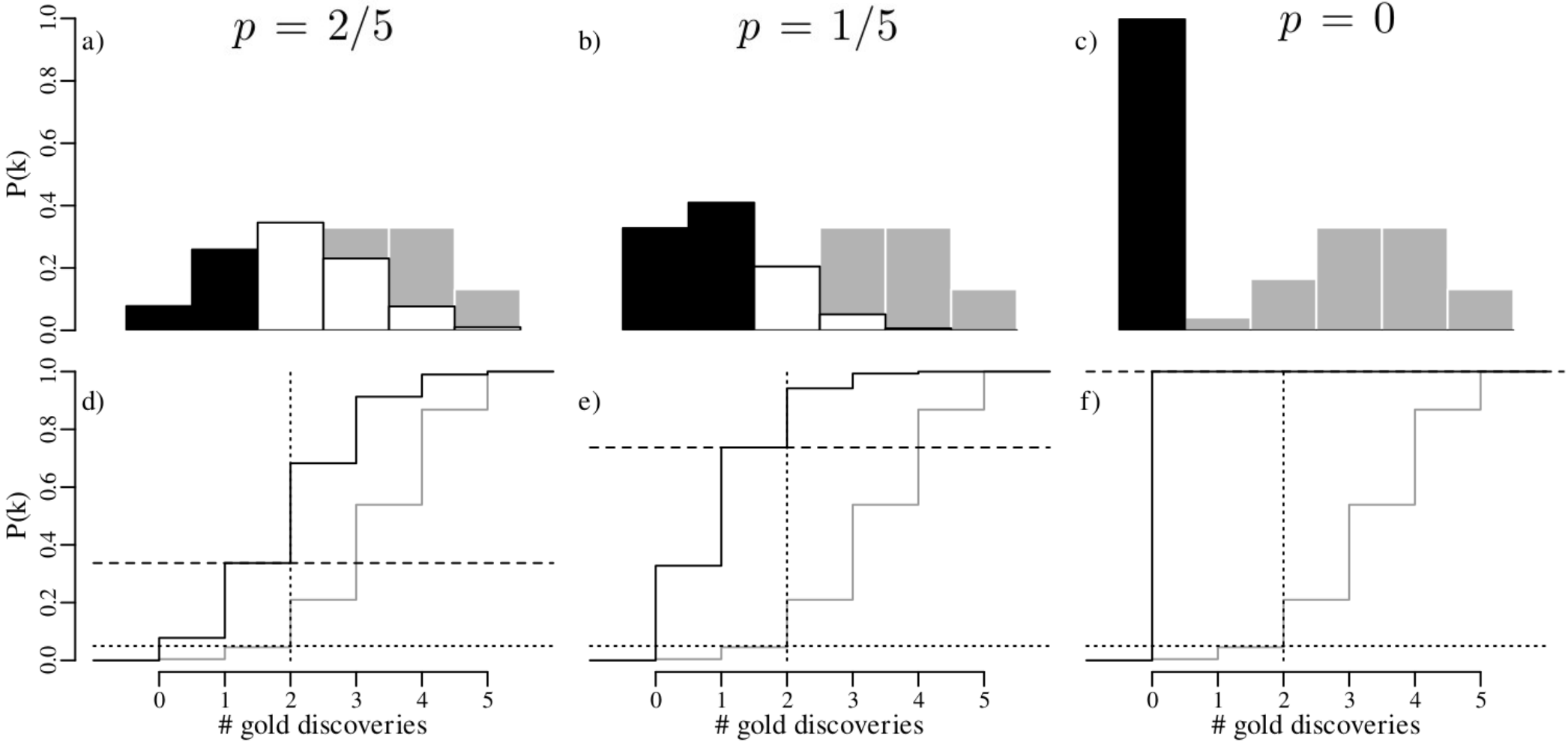
$H_\circ$ is ...	false	true
rejected	correct decision	Type-I error
not rejected	Type-II error	correct decision

$\beta = 1 - \text{power}$

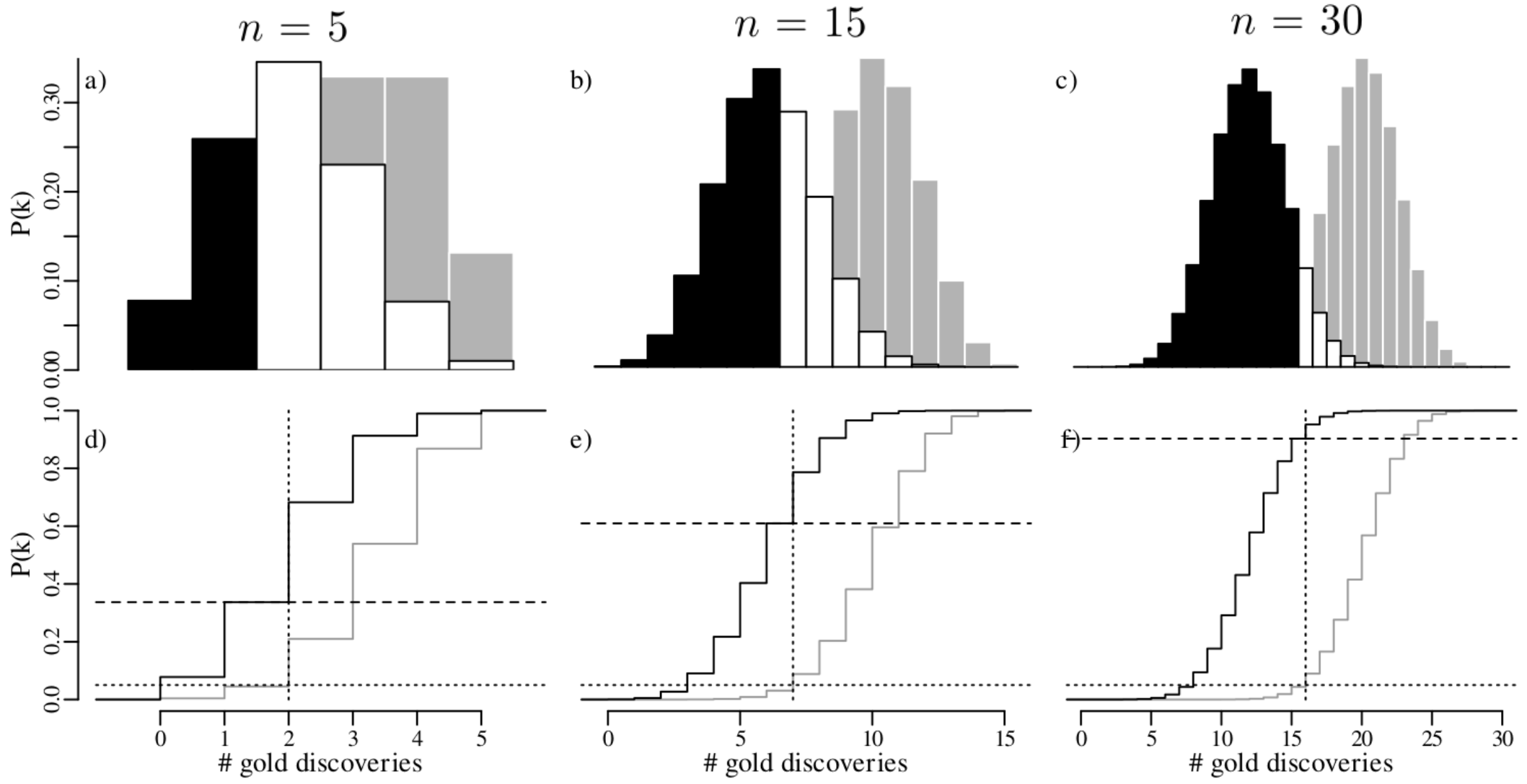
depends on 1. **The degree to which $H_\circ$ is false**

2. **Sample size**

1. The degree to which $H_\circ$ is false



2. Sample size



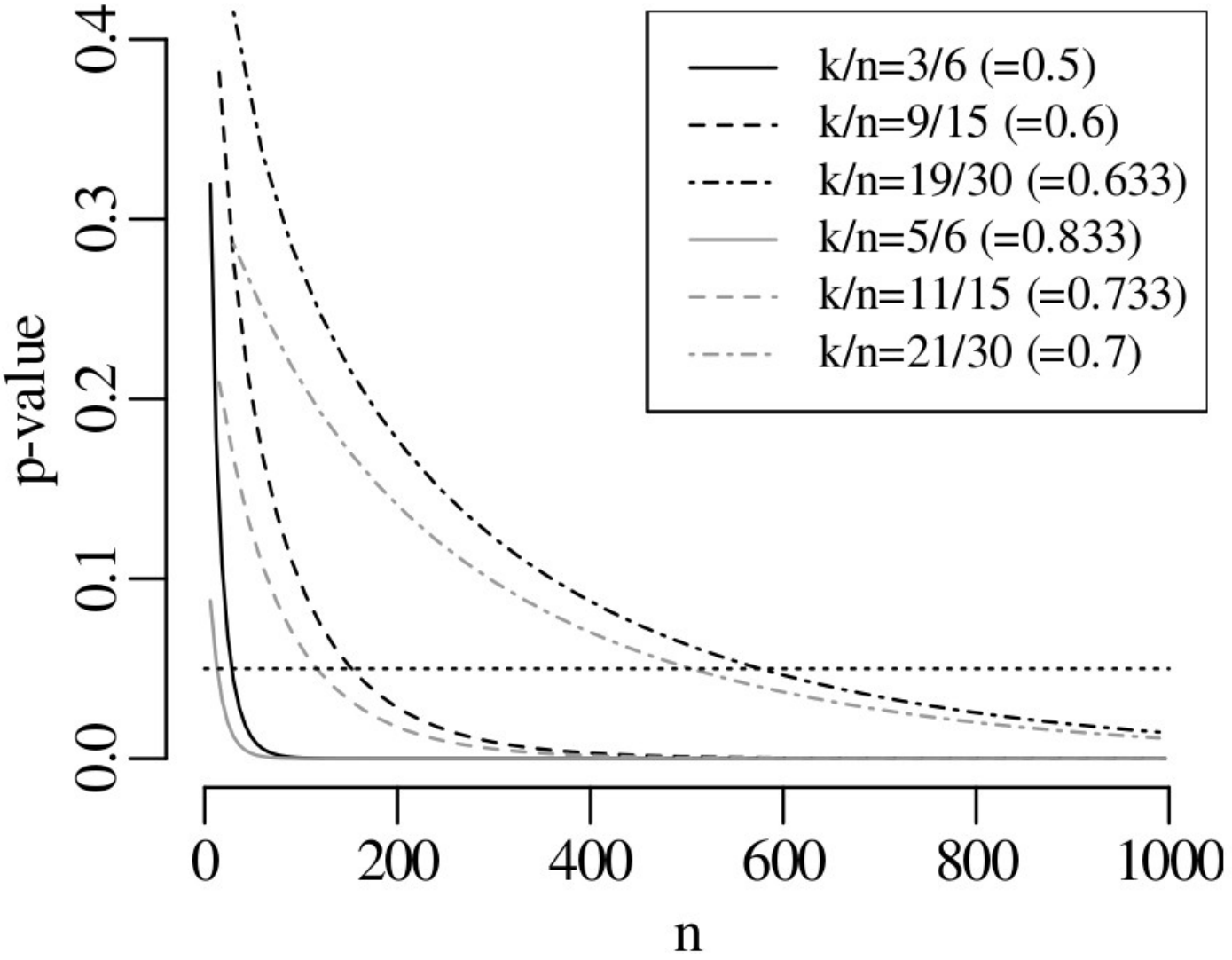
The scientific method

1. Formulate a hypothesis.
2. Design an experiment to test said hypothesis.
3. Carry out the experiment and check to see if it matches the prediction.

1. Hypothesis: Earth's lower mantle is made of olivine.
2. Test: Study the stability of olivine at lower mantle pressures (24-136 GPa).
3. Result: Olivine is not stable at lower mantle pressures.

1. Hypothesis: Earth's lower mantle is made of perovskite.
2. Test: Study the stability of perovskite at lower mantle pressures.
3. Result: Perovskite is stable at lower mantle pressures.

$p = 2/3$?


k/n=3/6 (=0.5)
k/n=9/15 (=0.6)
k/n=19/30 (=0.633)
k/n=5/6 (=0.833)
k/n=11/15 (=0.733)
k/n=21/30 (=0.7)
p-value
0.0
0.1
0.2
0.3
0.4
0
200
400
600
800
1000
n

Pitfalls

All hypotheses are wrong ... in some decimal place – John Tukey (paraphrased)

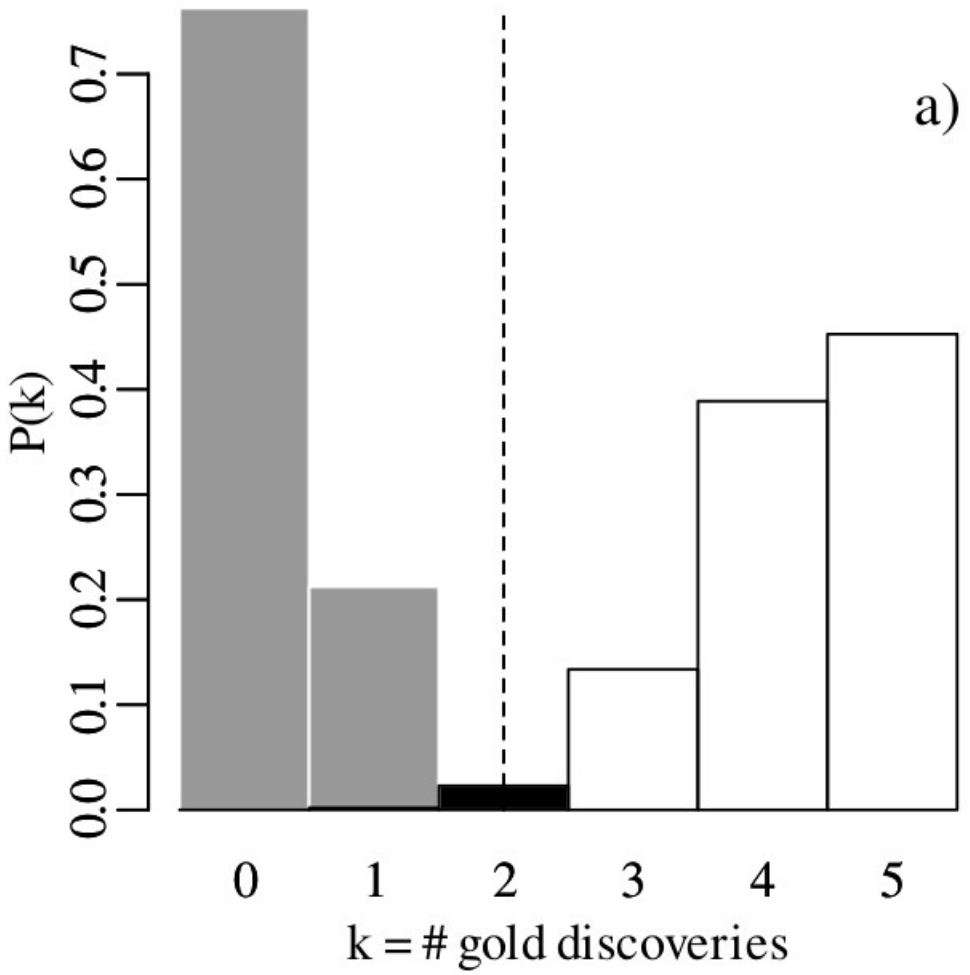
All models are wrong, but some are useful – George Box

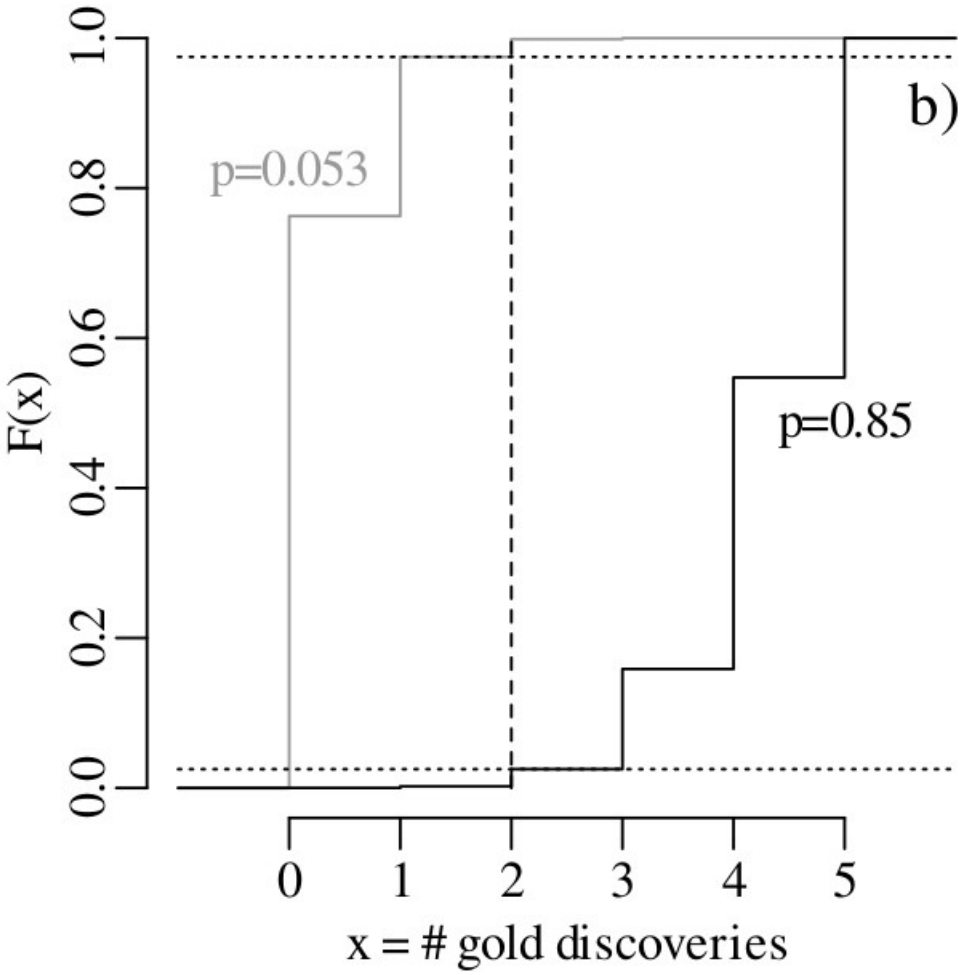
Confidence intervals

Which values of p are compatible with the observation of $k = 2$ successes out of $n = 5$ claims?

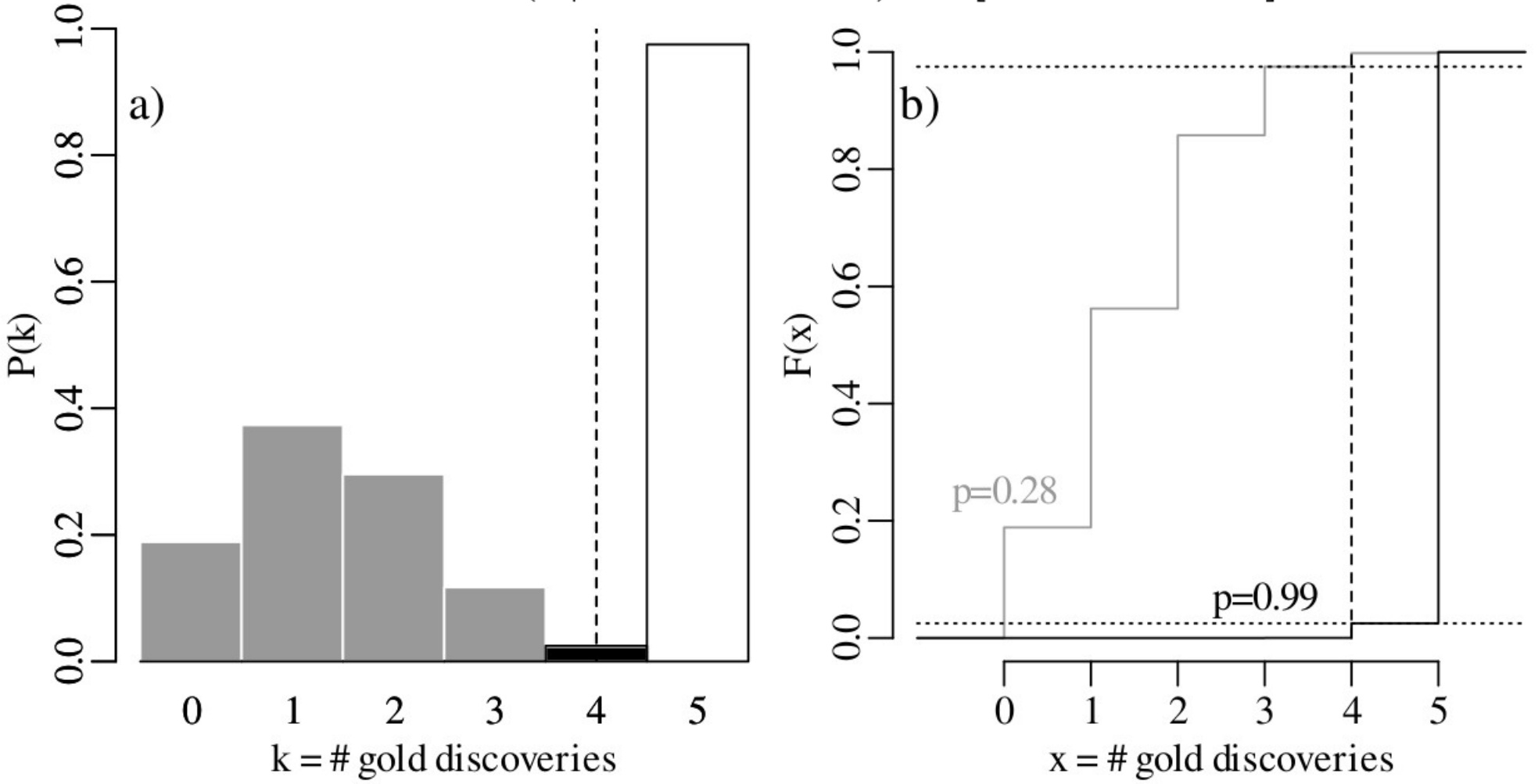
1. ~~**p=0?**~~
2. **p=0.1?** $P(k \geq 2|p = 0.1, n = 5) = \sum_{i=2}^{5} \binom{5}{i} (0.1)^i (0.9)^{n-i} = 0.081 \geq \alpha/2$
3. **p=2/5?** $= \hat{p}$
4. **p=2/3?** $P(k \leq 2|p = 2/3, n = 5) = \sum_{i=0}^{2} \binom{5}{i} (2/3)^i (1/3)^{n-i} = 0.21 \geq \alpha/2$
5. ~~**p=0.9?**~~ $P(k \leq 2|p = 0.9, n = 5) = \sum_{i=0}^{2} \binom{5}{i} (0.9)^i (0.1)^{n-i} = 0.0086 \leq \alpha/2$
6. ~~**p=1?**~~

$$\text{C.I.}(p|k=2,n=5) = [0.053, 0.85]$$

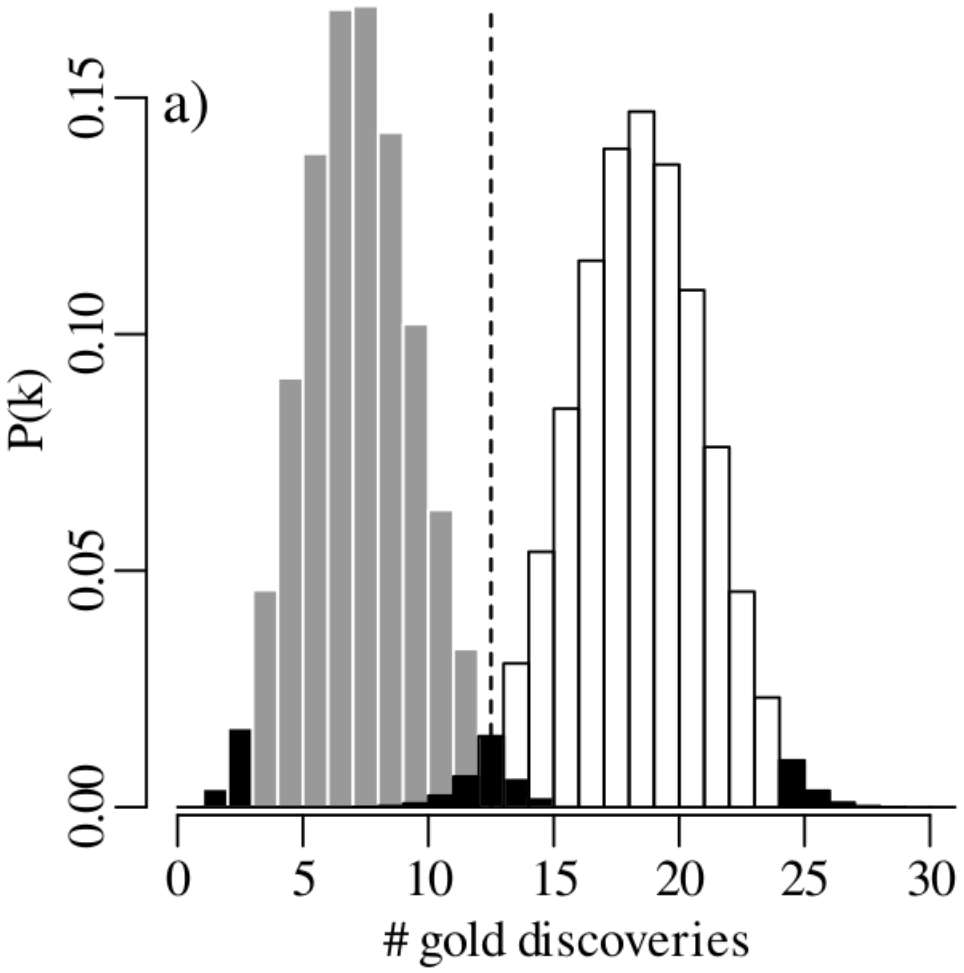


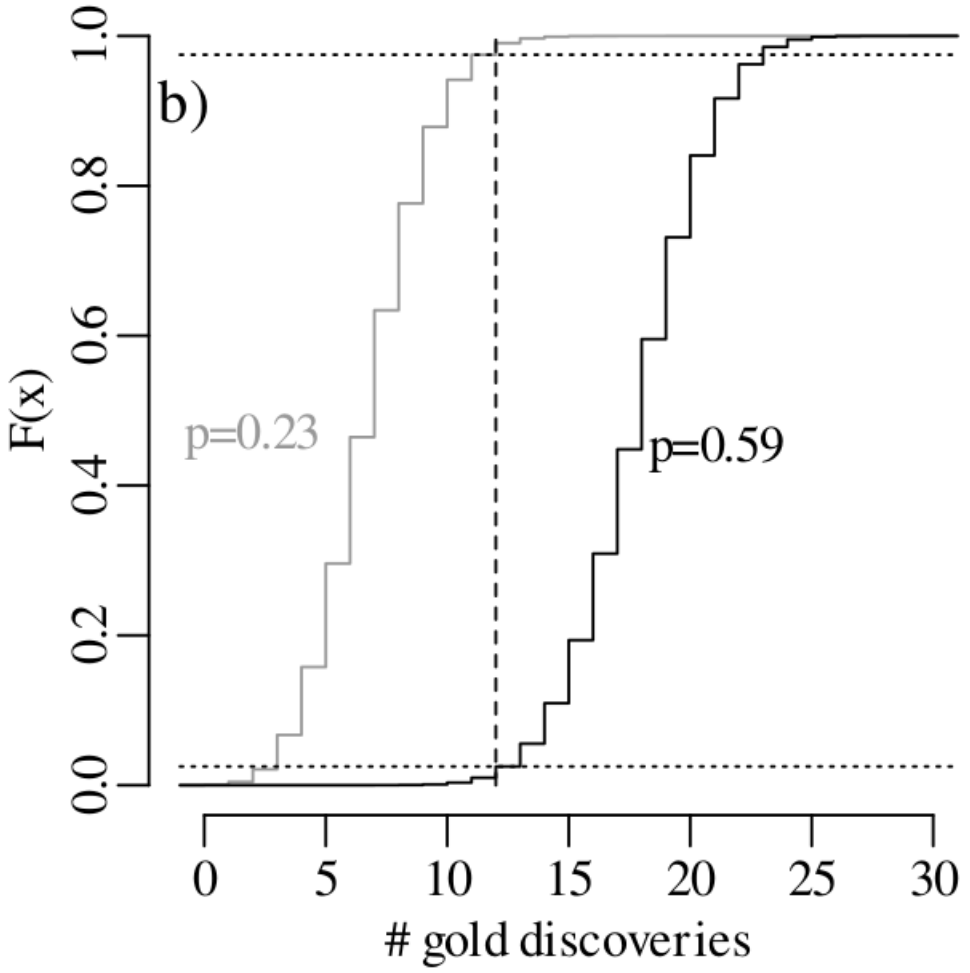


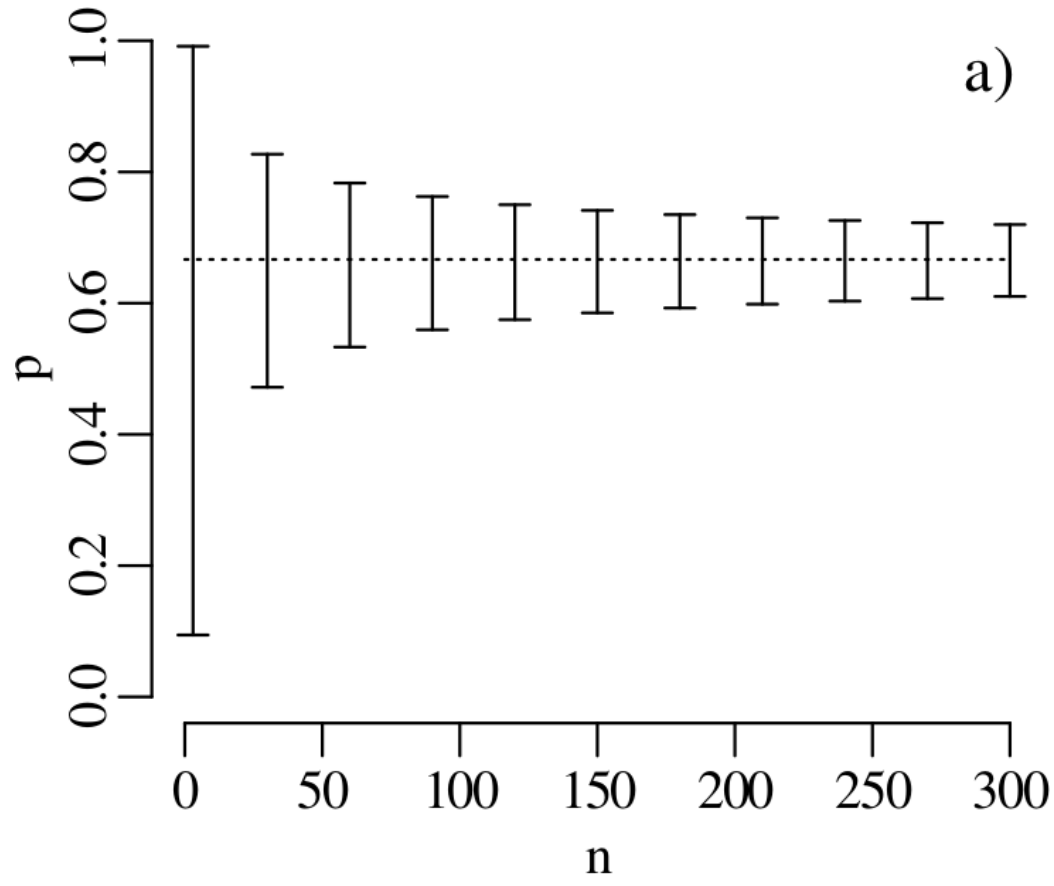
$$\text{C.I.}(p|k=4,n=5) = [0.284, 0.995]$$

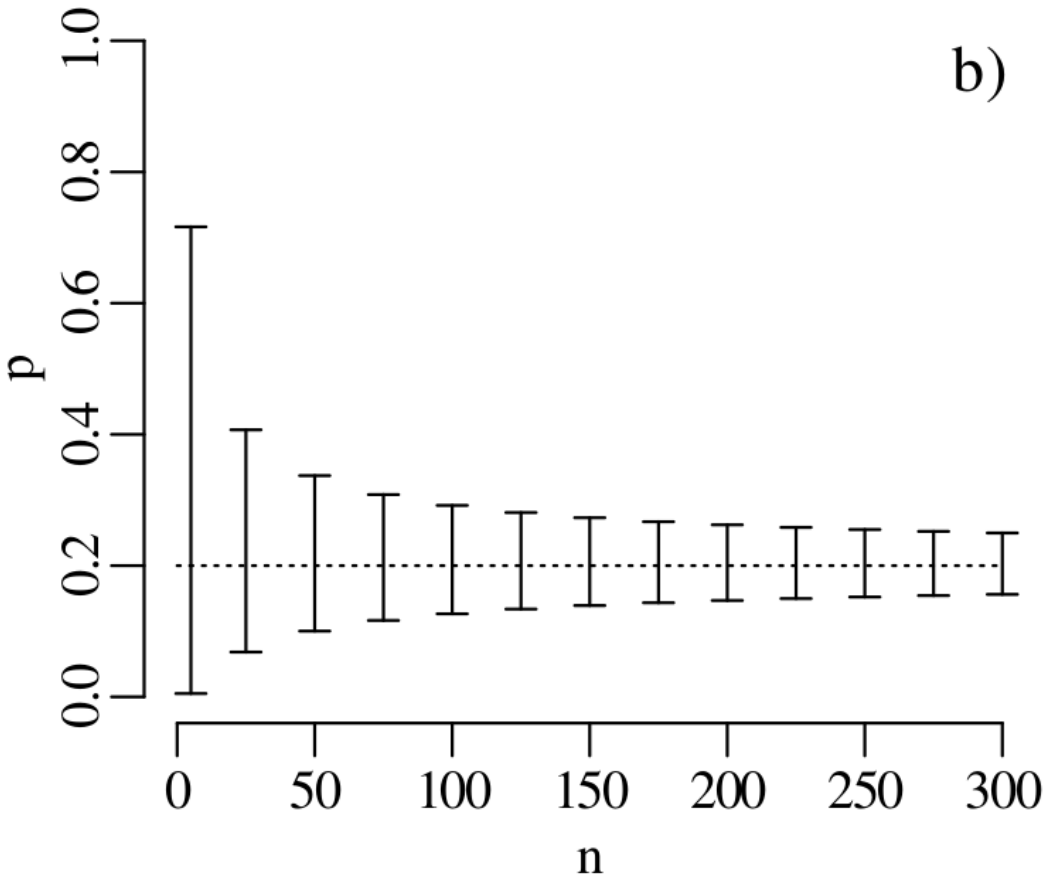


$$\text{C.I.}(p|k=12, n=30) = [0.23, 0.59]$$

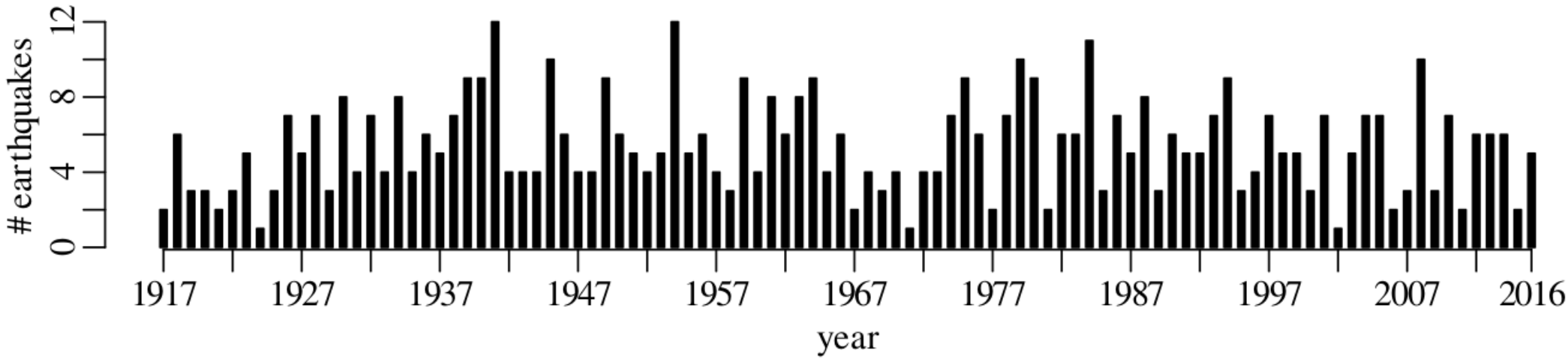


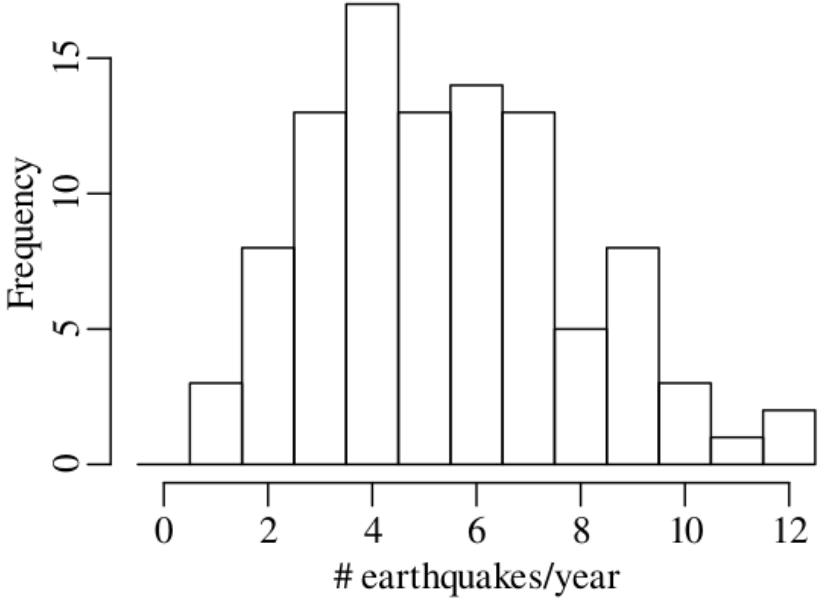



a)
p
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
n


b)
p
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
n

The Poisson distribution

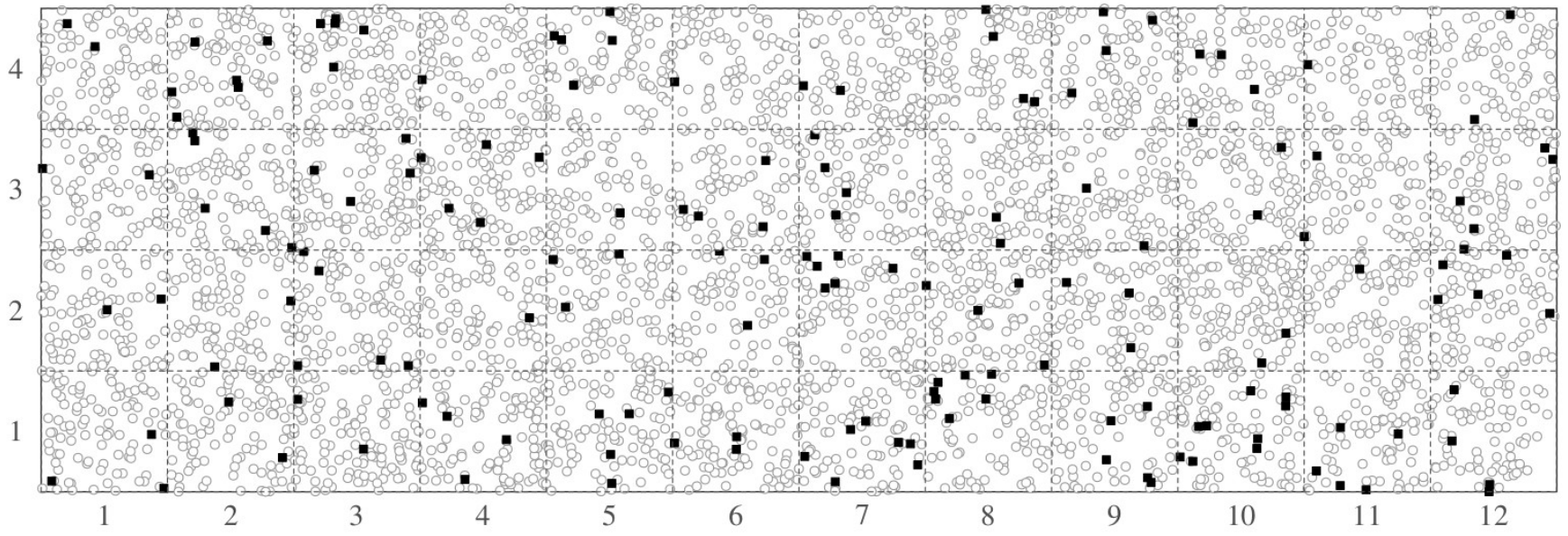




mean: 5.43

standard deviation: 2.50

variance: 6.24



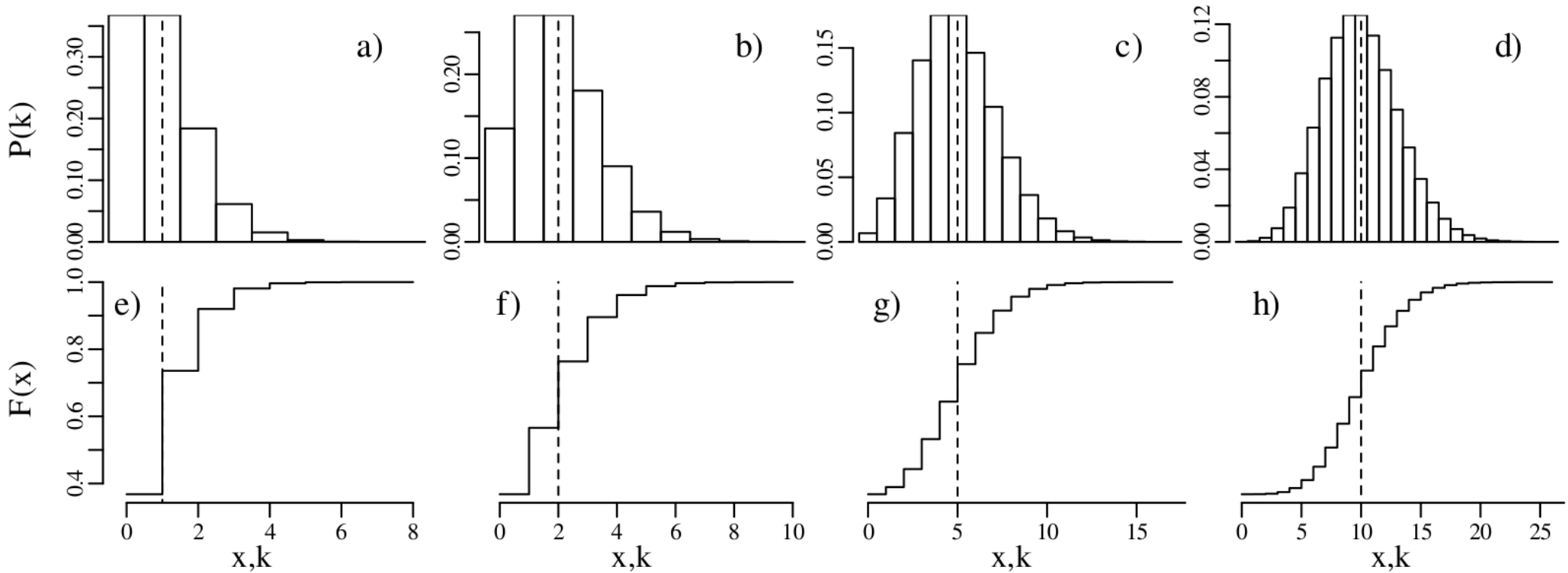
4	2	6	5	1	5	1	2	4	4	4	1	2
3	2	5	4	5	1	4	4	2	2	2	2	5
2	2	2	5	1	3	3	6	4	3	2	1	5
1	3	2	2	4	5	3	7	7	5	9	5	4
	1	2	3	4	5	6	7	8	9	10	11	12

mean: 3.50

standard deviation: 1.85

variance: 3.40

$$P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$



$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k} \quad \xrightarrow[p \to 0]{n \to \infty} \quad P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{where} \quad \lambda = np$$

n	10	20	50	100	200	500	1000	2000	5000	10000
p	0.5	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
$P(k \leq 2)$	0.0547	0.0913	0.112	0.118	0.121	0.1234	0.124	0.1243	0.1245	0.1246

Examples of Poisson variables

1. people killed by lightning per year;
2. mutations in DNA per generation;
3. radioactive disintegrations per unit time;
4. mass extinctions per 100 million years.

~~earthquakes including aftershocks~~

~~floods per year~~

Parameter estimation

$$P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\mathcal{L}(\lambda|k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\lambda = k$$

Hypothesis tests

$k = 9$ zircons given $\lambda = 3.5$

1. $H_\circ$ (**null hypothesis**) $\lambda = 3.5$

   H_a (**alternative hypothesis**): $\lambda > 3.5$

2. test statistic : $k = 9$ successes

3. The null distribution

k	0	1	2	3	4	5	6	7	8	*9*	10
$P(T = k)$	0.030	0.106	0.185	0.216	0.189	0.132	0.077	0.038	0.017	0.007	0.002
$P(T \geq k)$	1.000	0.970	0.864	0.679	0.463	0.275	0.142	0.065	0.027	*0.010*	0.003

4. significance level $\alpha = 0.05$

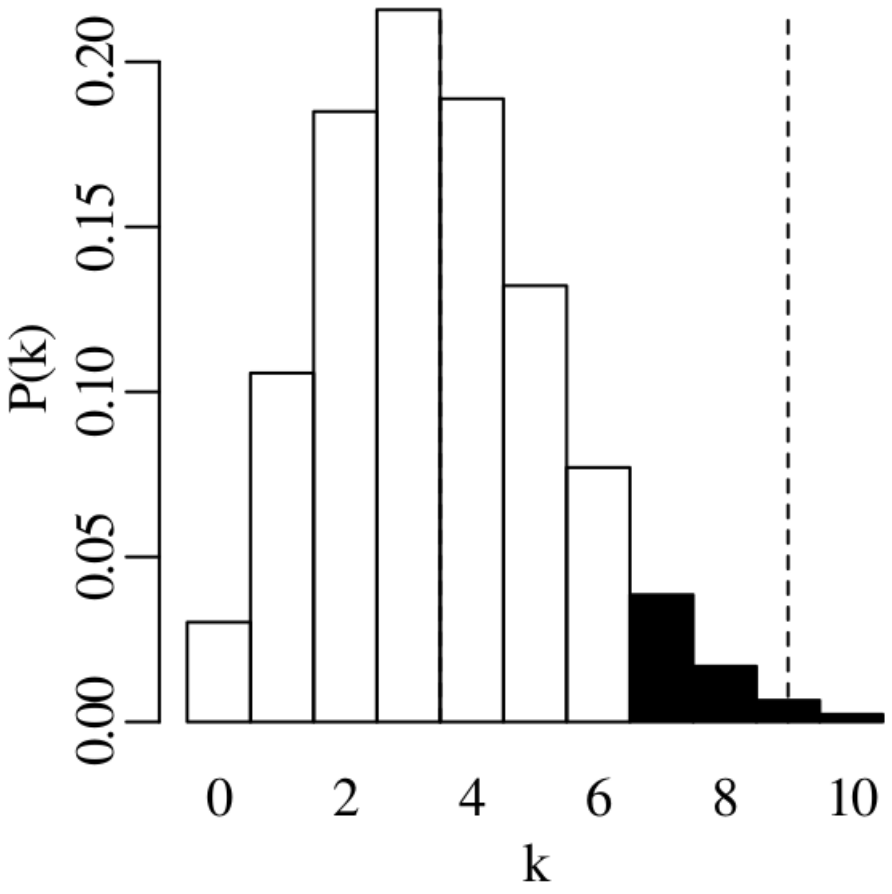
5. rejection region

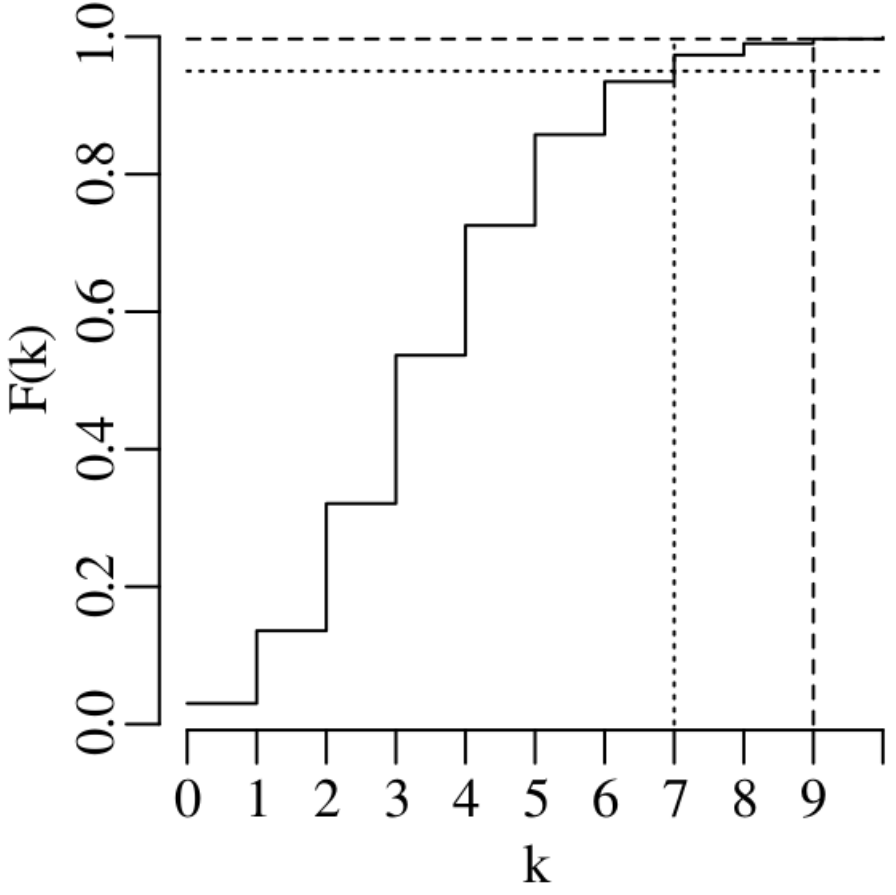
k	0	1	2	3	4	5	6	7	**8**	***9***	**10**
$P(T = k)$	0.030	0.106	0.185	0.216	0.189	0.132	0.077	0.038	0.017	0.007	0.002
$P(T \geq k)$	1.000	0.970	0.864	0.679	0.463	0.275	0.142	0.065	**0.027**	***0.010***	**0.003**

$R = \{8, 9, 10, \ldots, \infty\}$

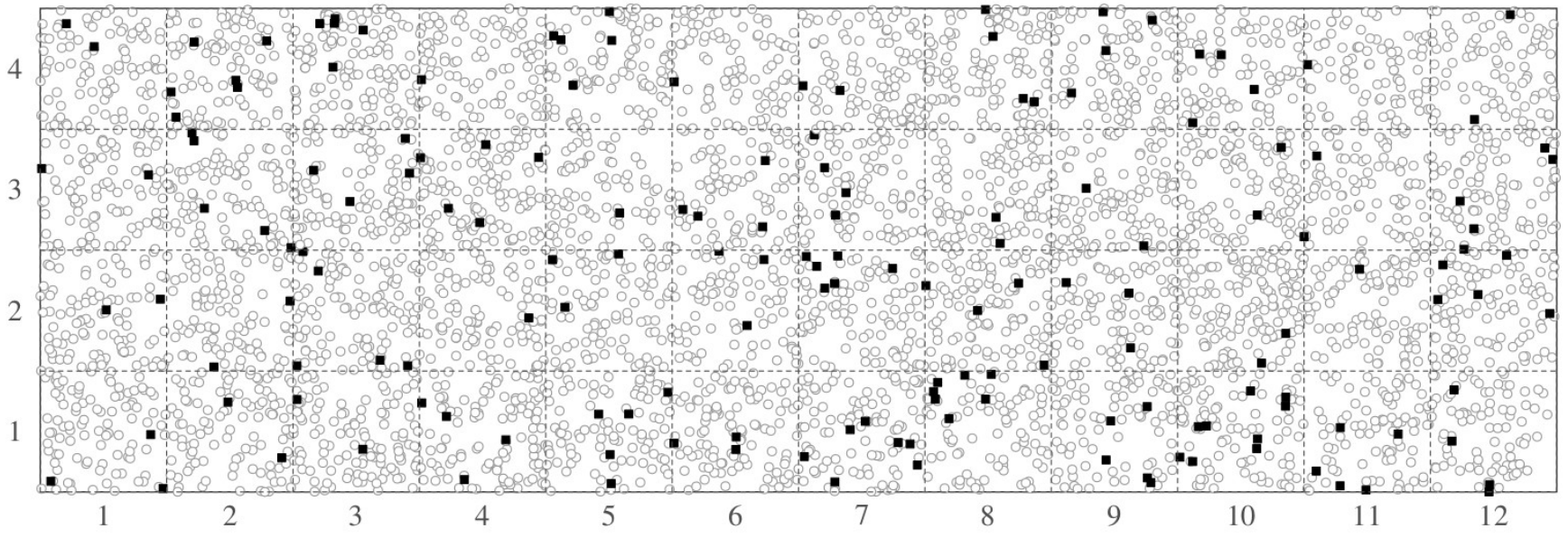
6. $k \in R$

7. p-value $= 0.010 < \alpha$


P(k)
0.00
0.05
0.10
0.15
0.20
0
2
4
6
8
10
k


F(k)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7
8
9
k

Multiple testing



4	2	6	5	1	5	1	2	4	4	4	1	2
3	2	5	4	5	1	4	4	2	2	2	2	5
2	2	2	5	1	3	3	6	4	3	2	1	5
1	3	2	2	4	5	3	7	7	5	9	5	4
	1	2	3	4	5	6	7	8	9	10	11	12

Assuming that $H_\circ$ is true, the probability of incurring a type-I error is:

$$1 \text{ experiment: } \alpha = 5\%$$

$$2 \text{ experiments: } 1 - (1-\alpha)^2 = 9.75\%$$

$$3 \text{ experiments: } 1 - (1-\alpha)^3 = 14\%$$

$$48 \text{ experiments: } 1 - (1-\alpha)^{48} = 91.5\%$$

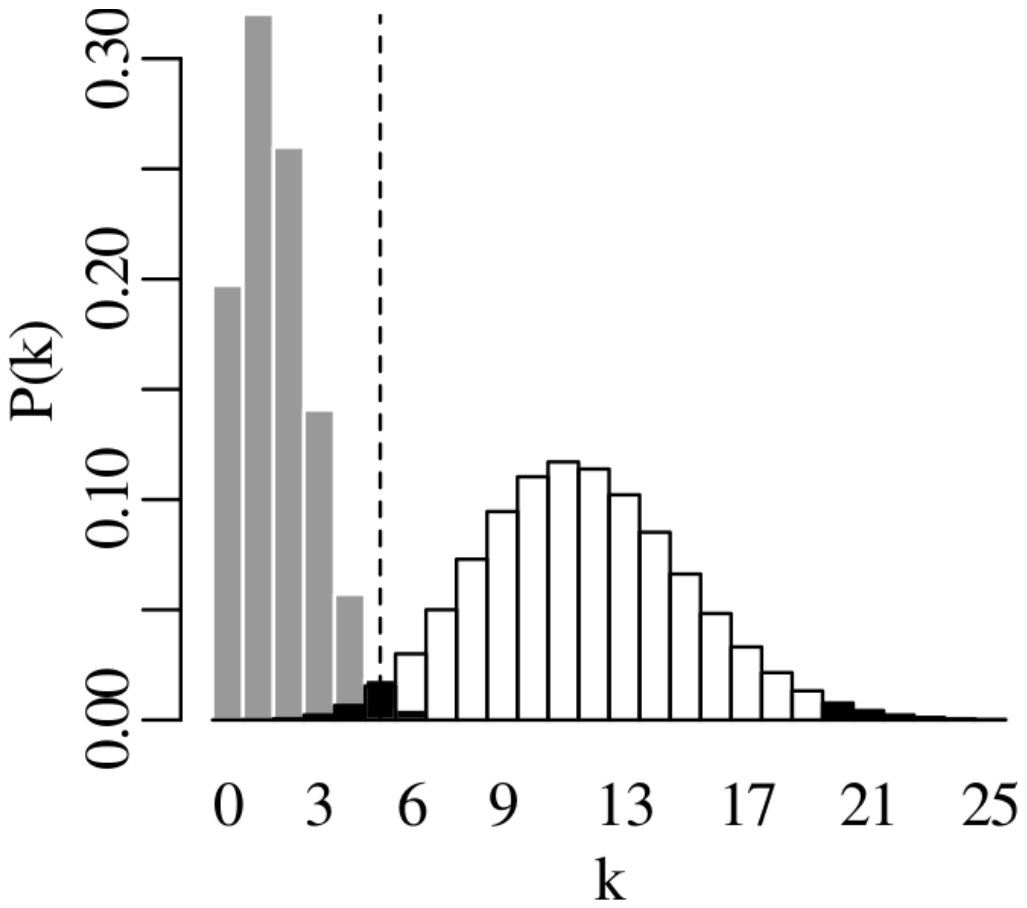
$$N \text{ experiments: } 1 - (1-\alpha)^N\%$$

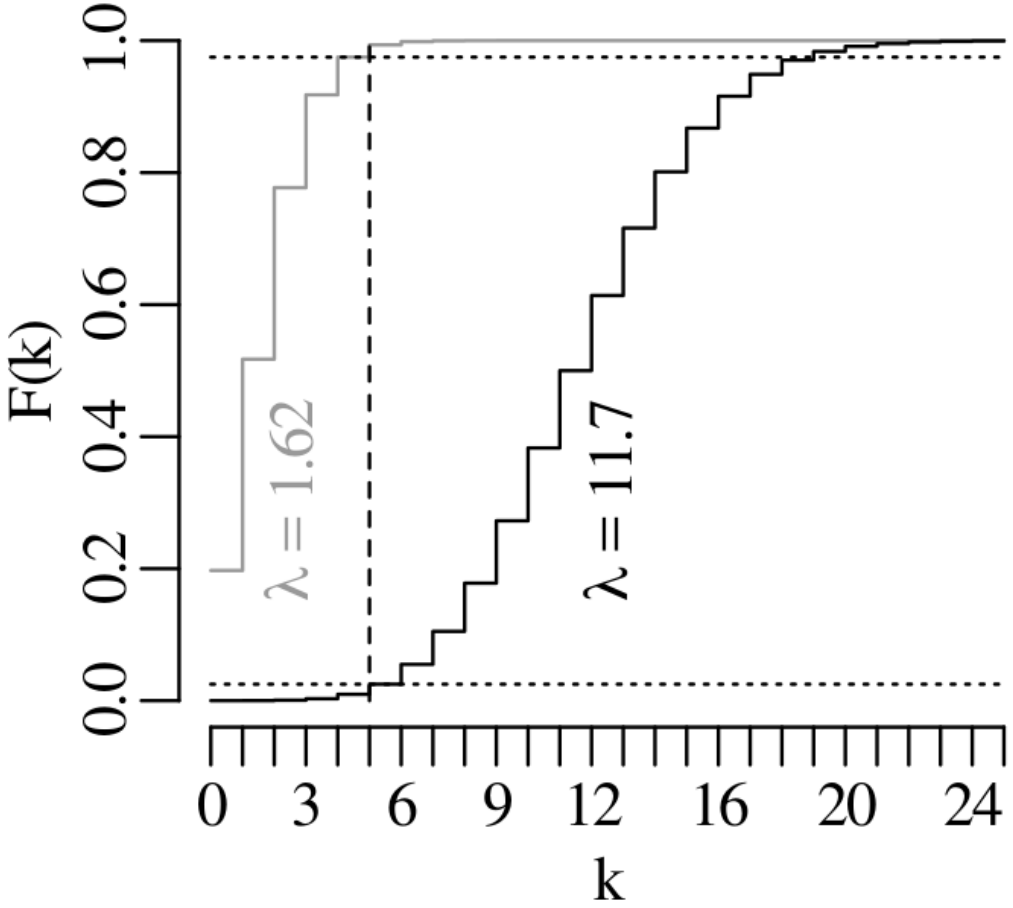
Bonferroni correction: α/N

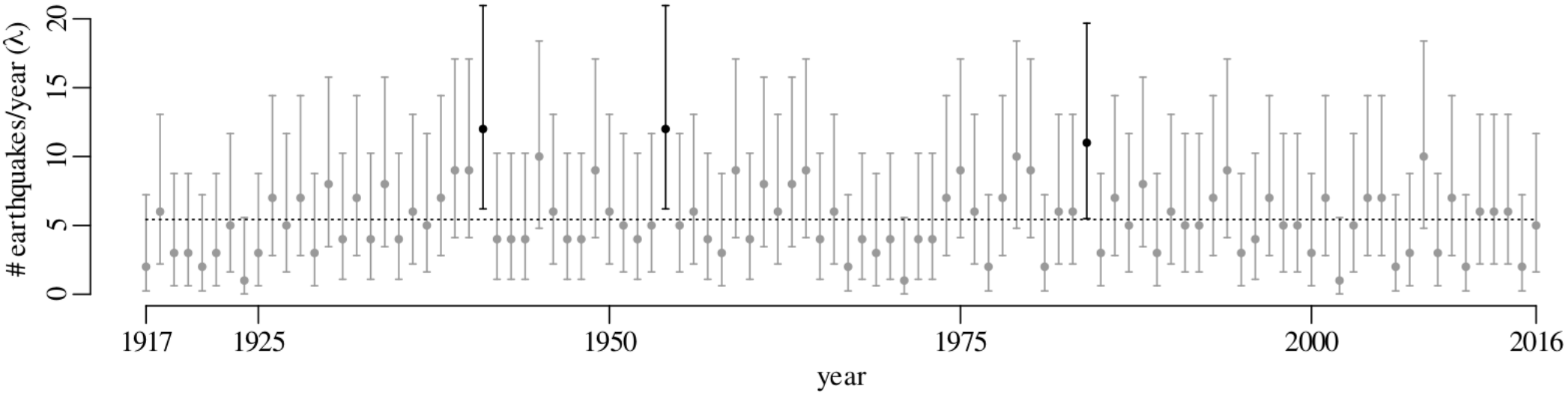
Confidence intervals

5 magnitude 5.0 or greater earthquakes occurred in the US in 2016

What is λ?






earthquakes/year (λ)
0
5
10
15
20
1917
1925
1950
1975
2000
2016
year

The normal distribution

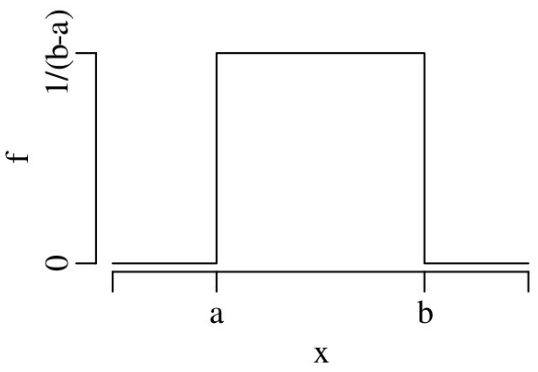
the binomial distribution $$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

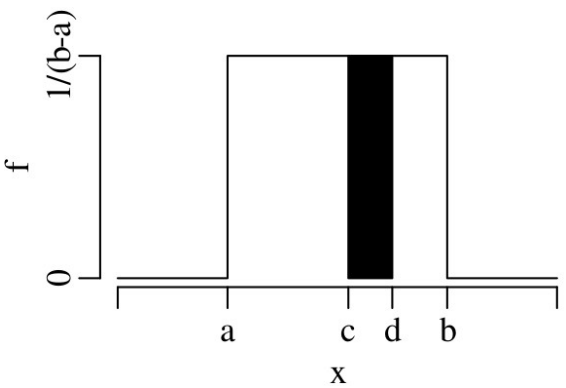
the Poisson distribution $$P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

the negative binomial distribution $$P(x|r,p) = \binom{r+x-1}{x} (1-p)^x p^r$$

the multinomial distribution $$P(k_1, k_2, \ldots, k_m | p_1, p_2, \ldots, p_m) = \frac{n!}{\prod_{i=1}^{m} k_i!} \prod_{i=1}^{m} p_i^{k_i}$$

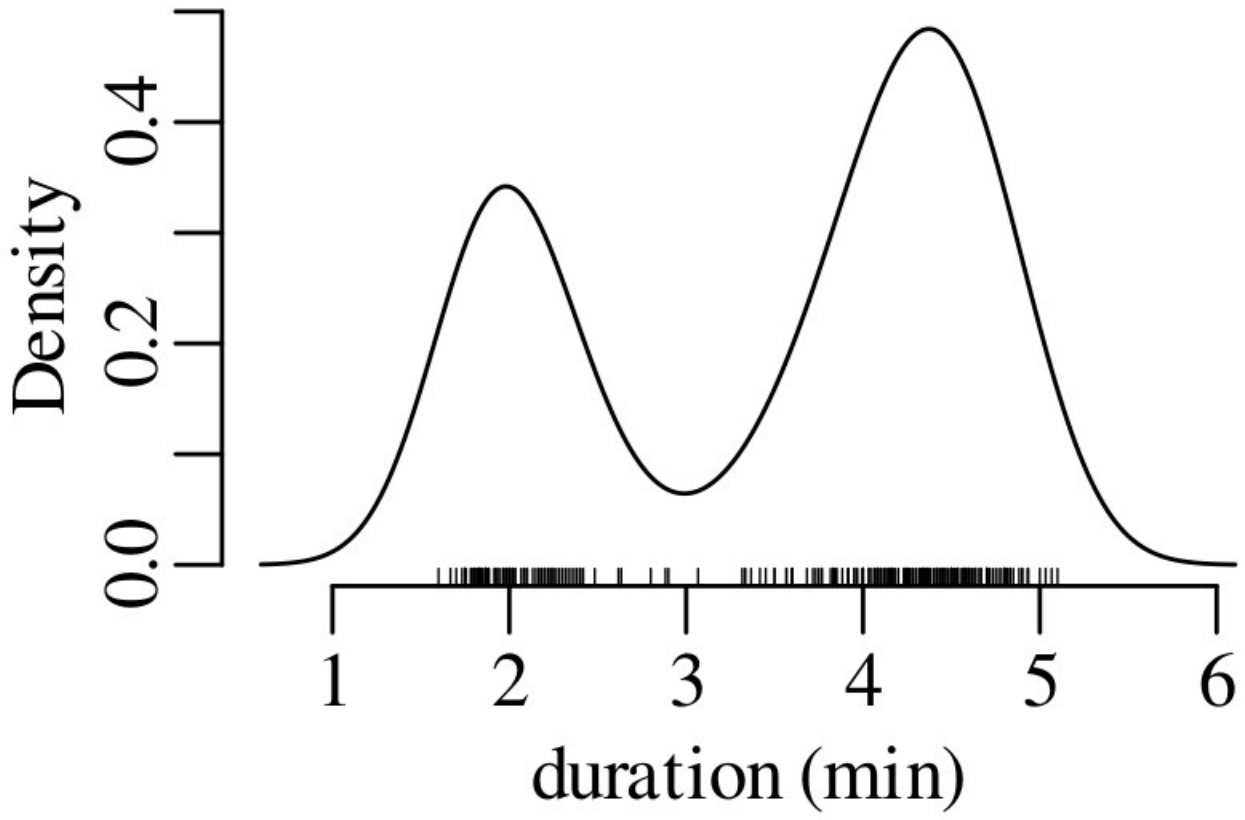
the uniform distribution $$f(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

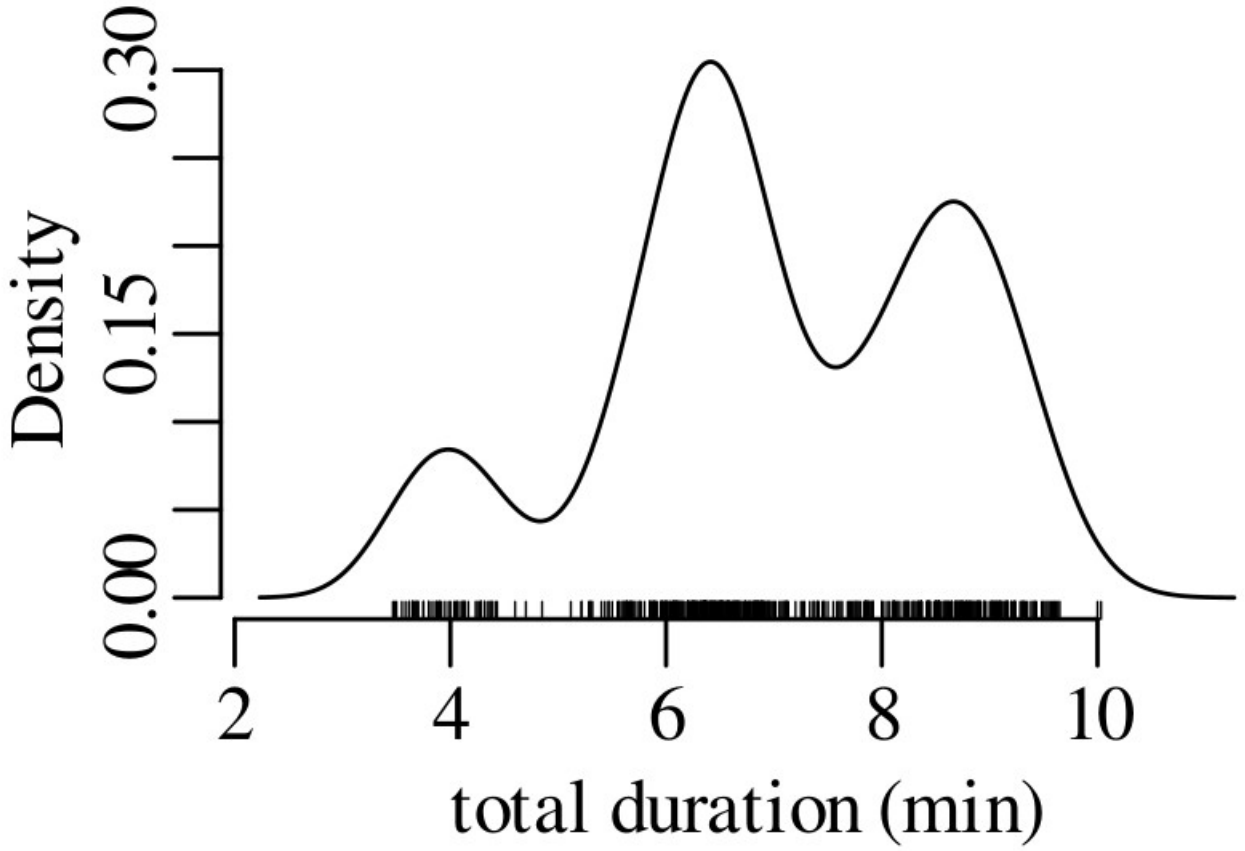


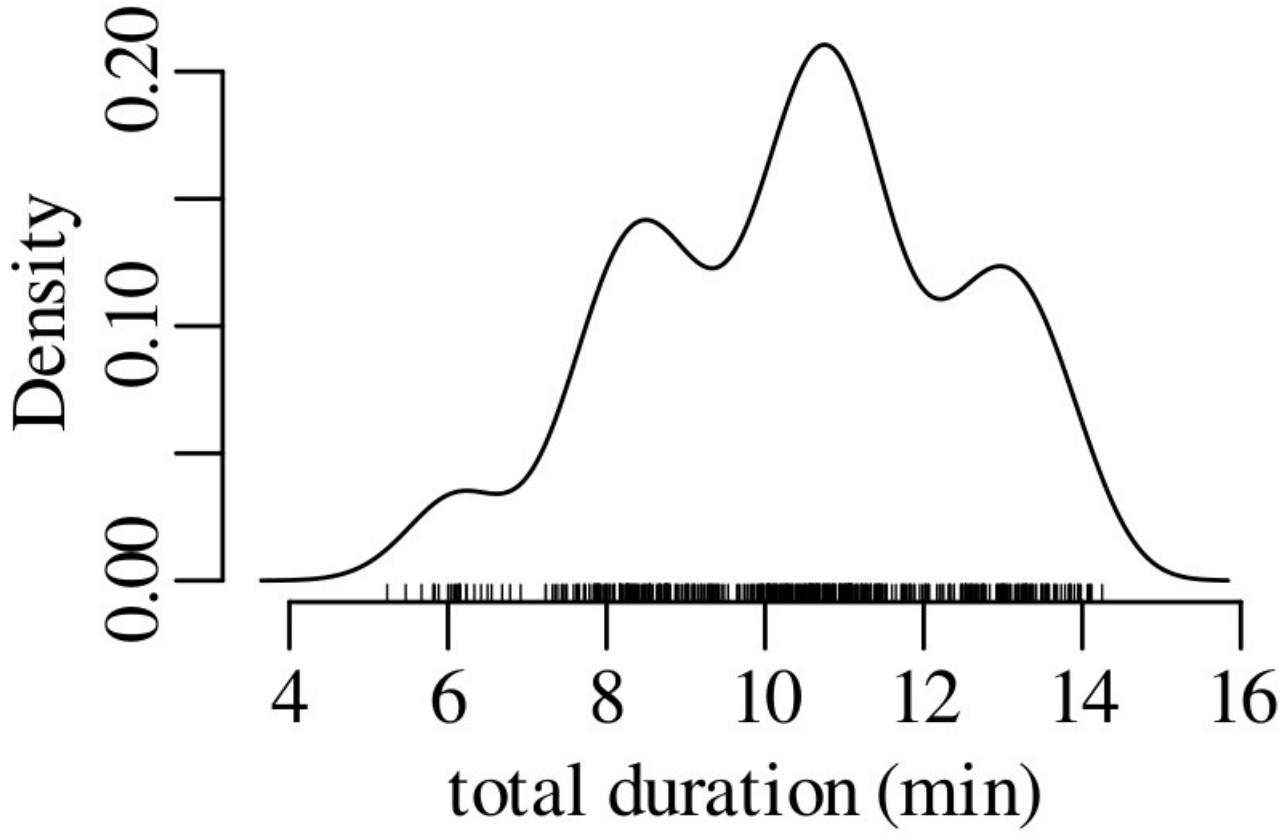


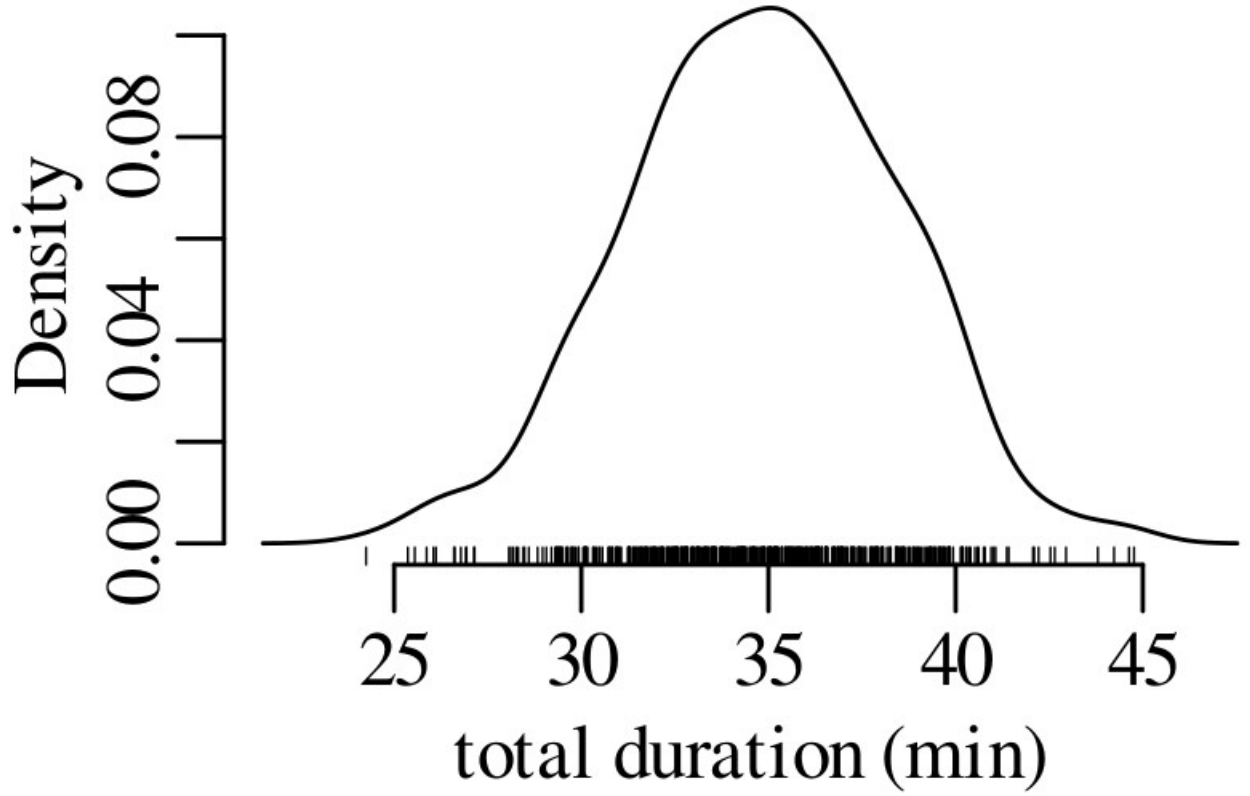
$$P(c \le x \le d) = \int_c^d f(x|a,b)\,dx$$

What is so normal about the Gaussian distribution?

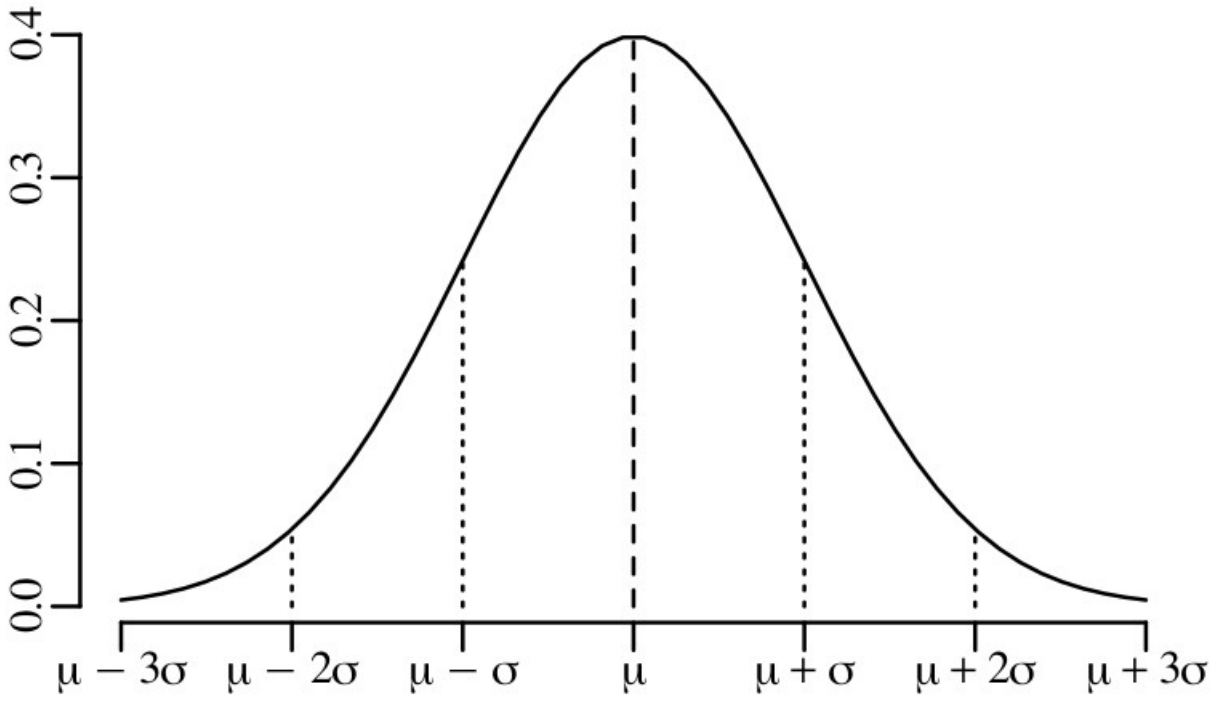


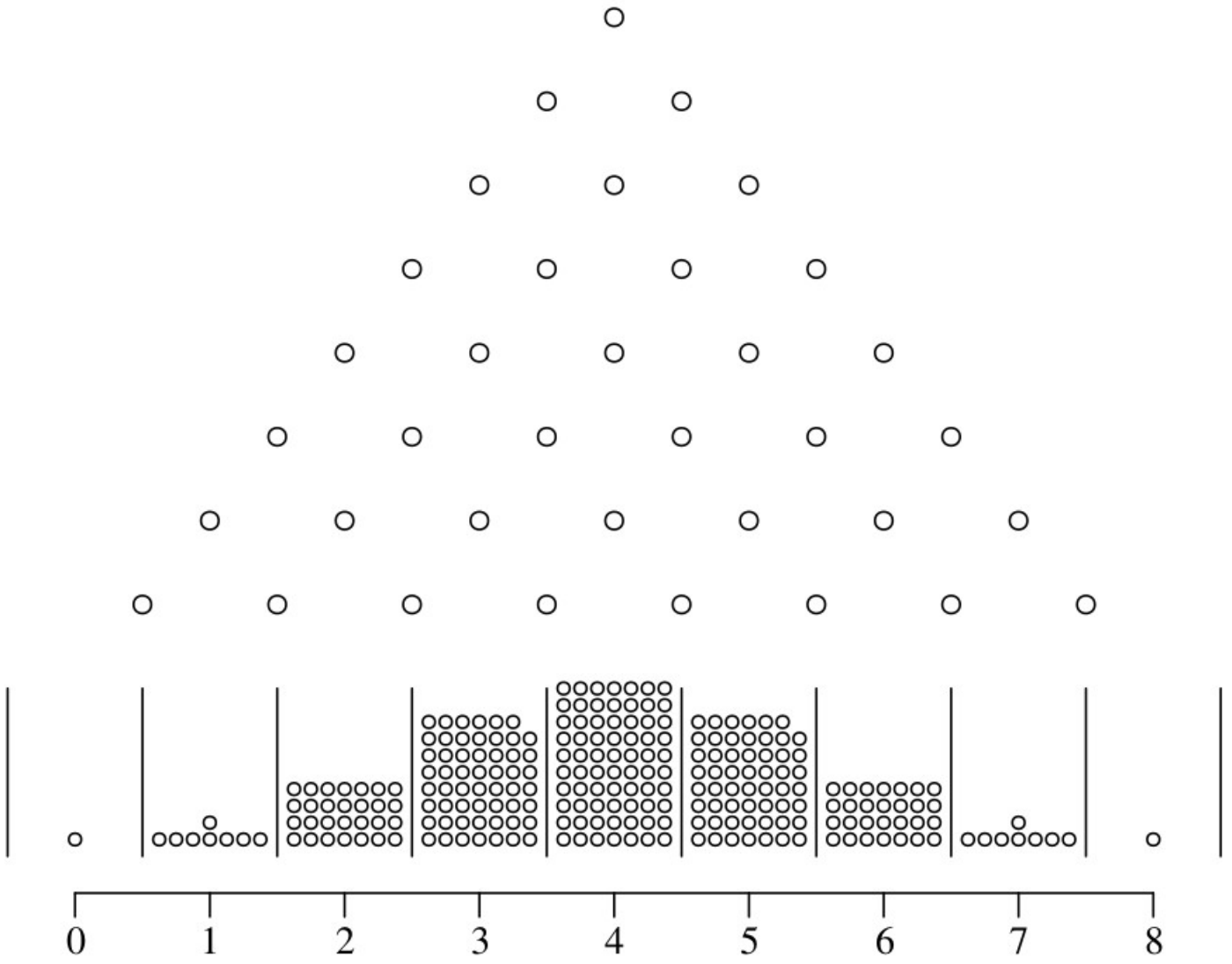

Density
0.00
0.15
0.30
2
4
6
8
10
total duration (min)

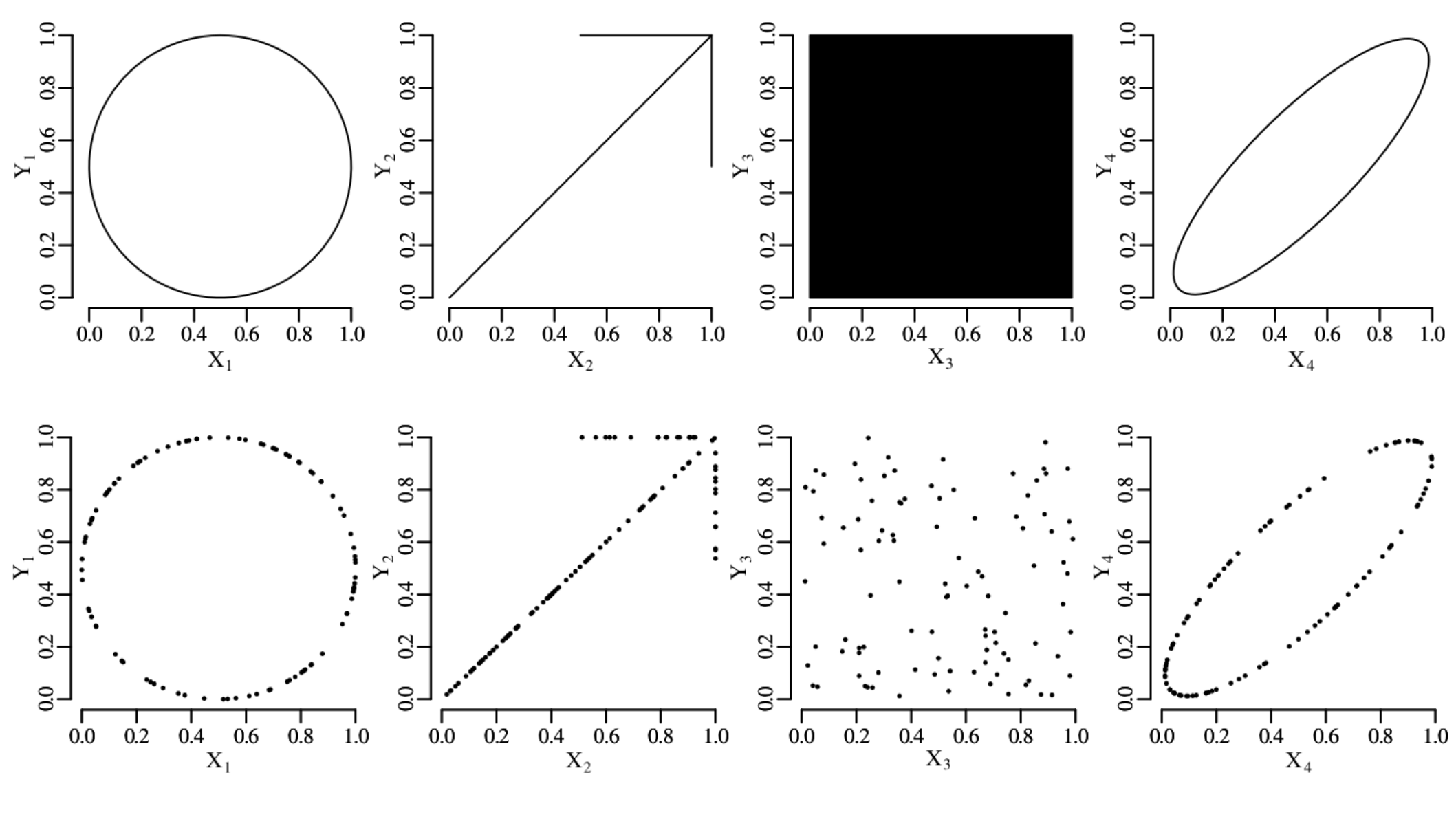

Density
0.00
0.10
0.20
4
6
8
10
12
14
16
total duration (min)

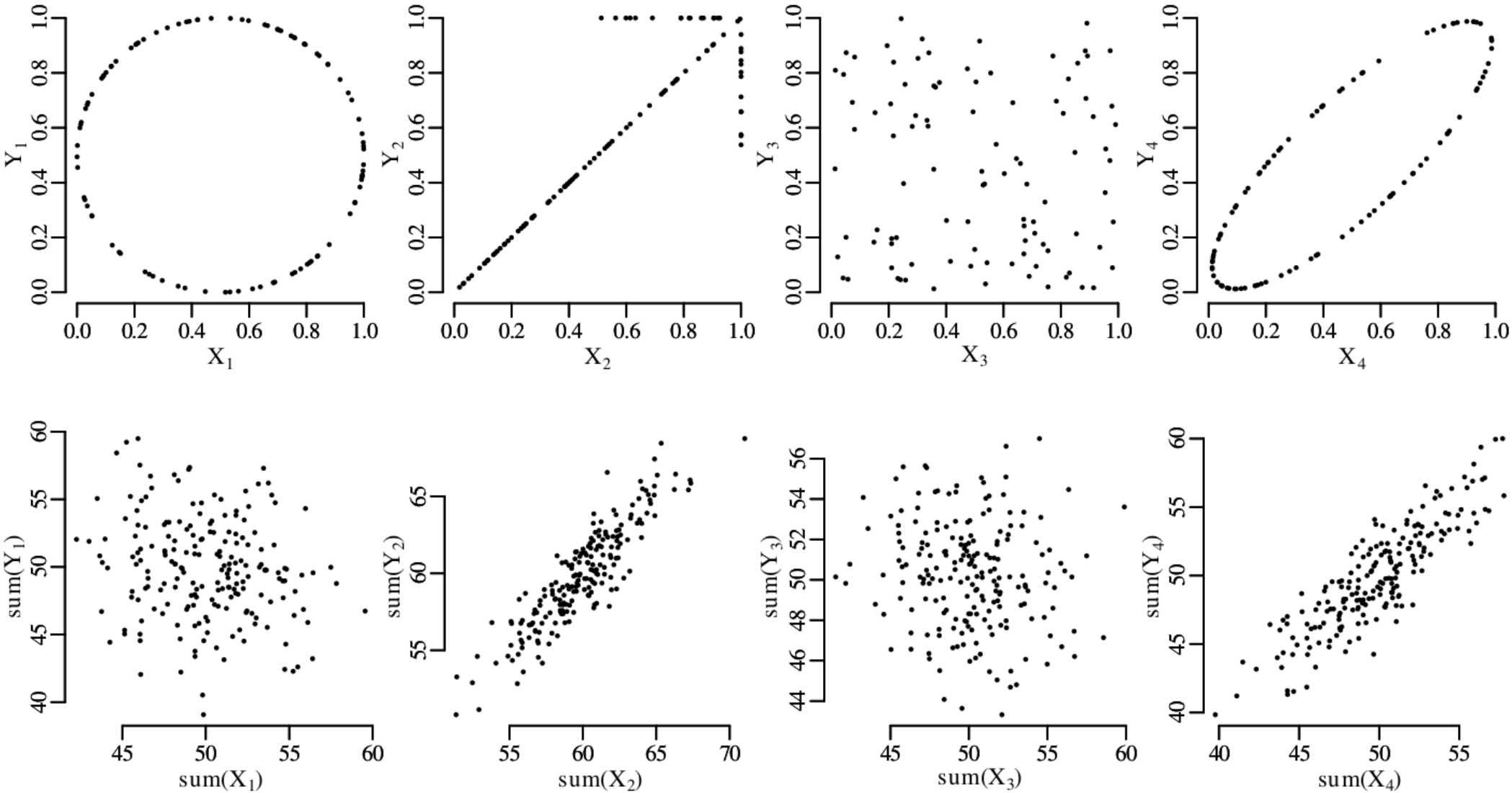

Density
0.00
0.04
0.08
25
30
35
40
45
total duration (min)

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

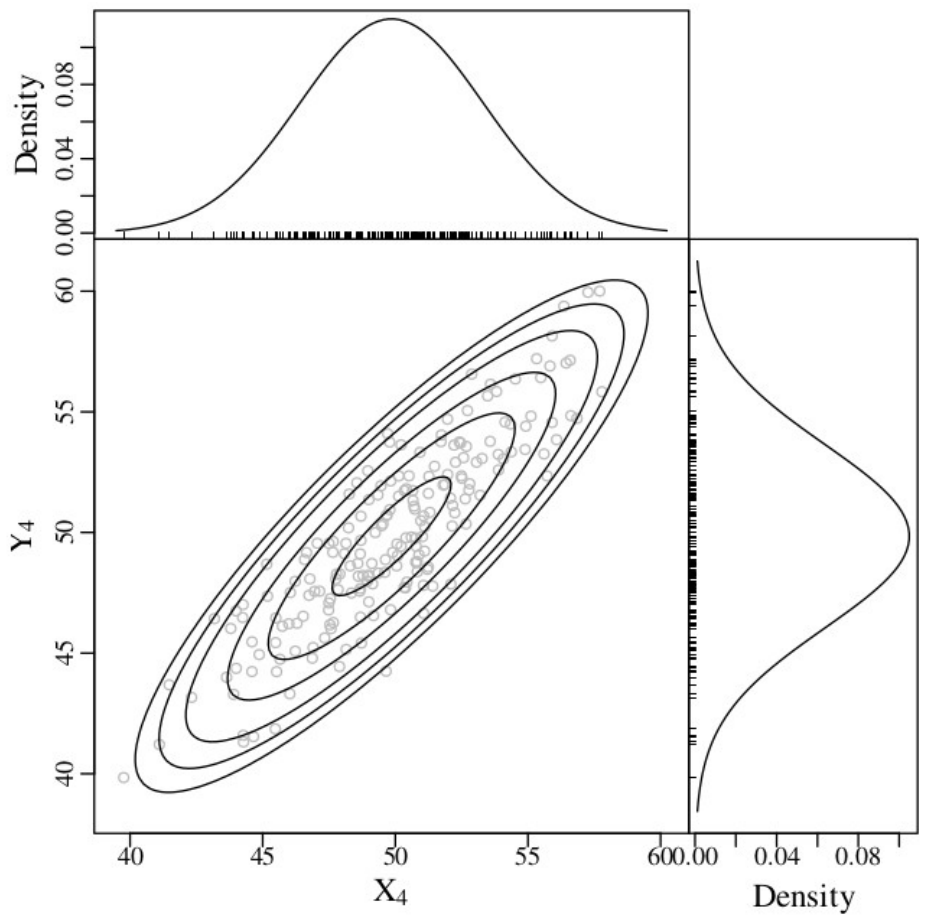



0
1
2
3
4
5
6
7
8

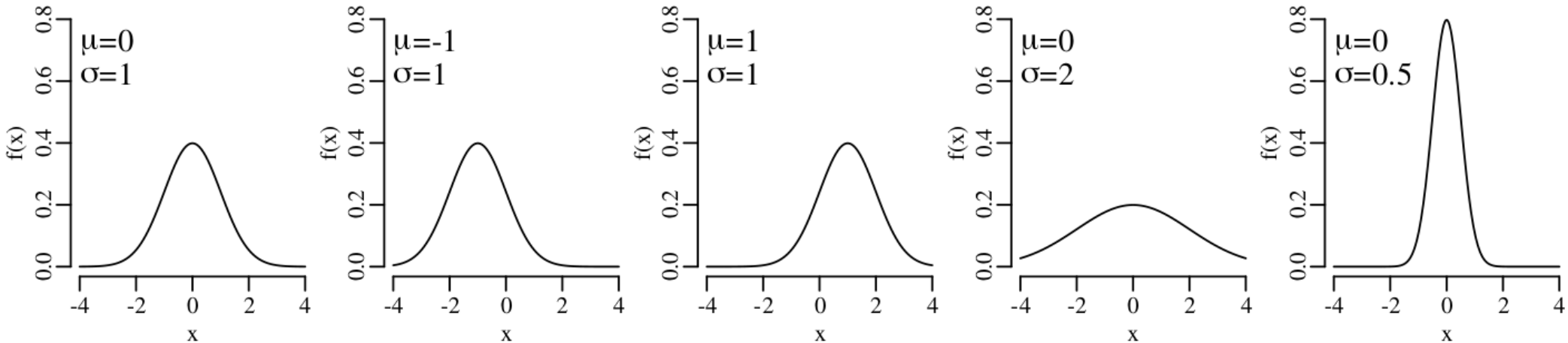

Y1
X1
Y2
X2
Y3
X3
Y4
X4
0.0
0.2
0.4
0.6
0.8
1.0


Y1
X1
Y2
X2
Y3
X3
Y4
X4
sum(Y1)
sum(X1)
sum(Y2)
sum(X2)
sum(Y3)
sum(X3)
sum(Y4)
sum(X4)

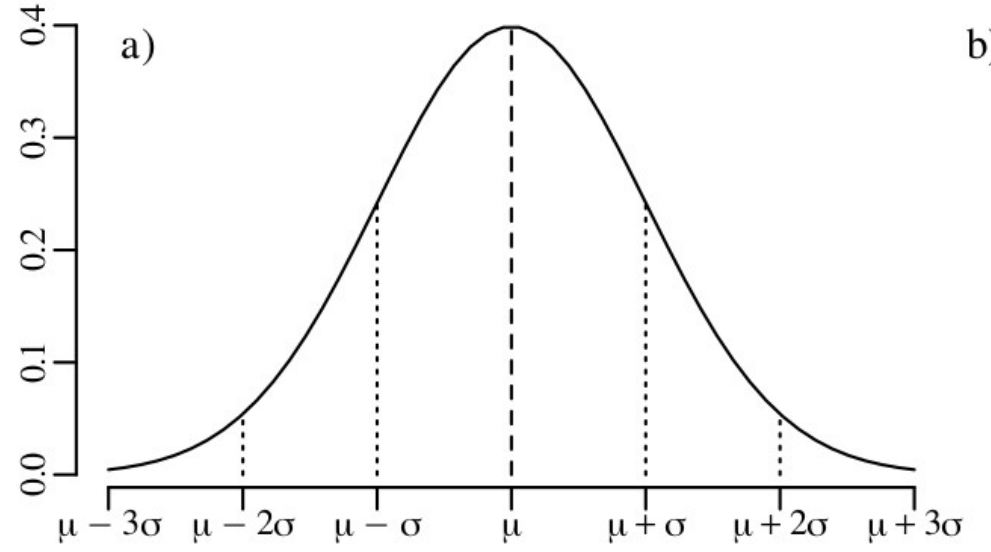
$$f(x, y|\mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{x,y}) = \frac{\exp\left(-\begin{bmatrix}(x-\mu_x) & (y-\mu_y)\end{bmatrix}\begin{bmatrix}\sigma_x^2 & \sigma_{x,y}\\ \sigma_{x,y} & \sigma_y^2\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_x\\ y-\mu_y\end{bmatrix}\Big/2\right)}{2\pi\sqrt{\begin{vmatrix}\sigma_x^2 & \sigma_{x,y}\\ \sigma_{x,y} & \sigma_y^2\end{vmatrix}}}$$

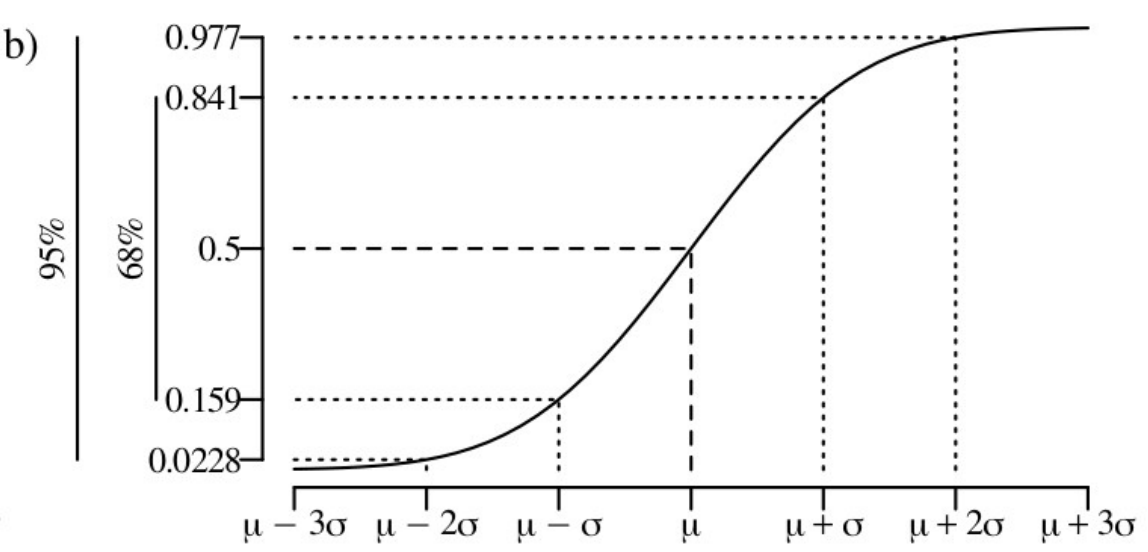


$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

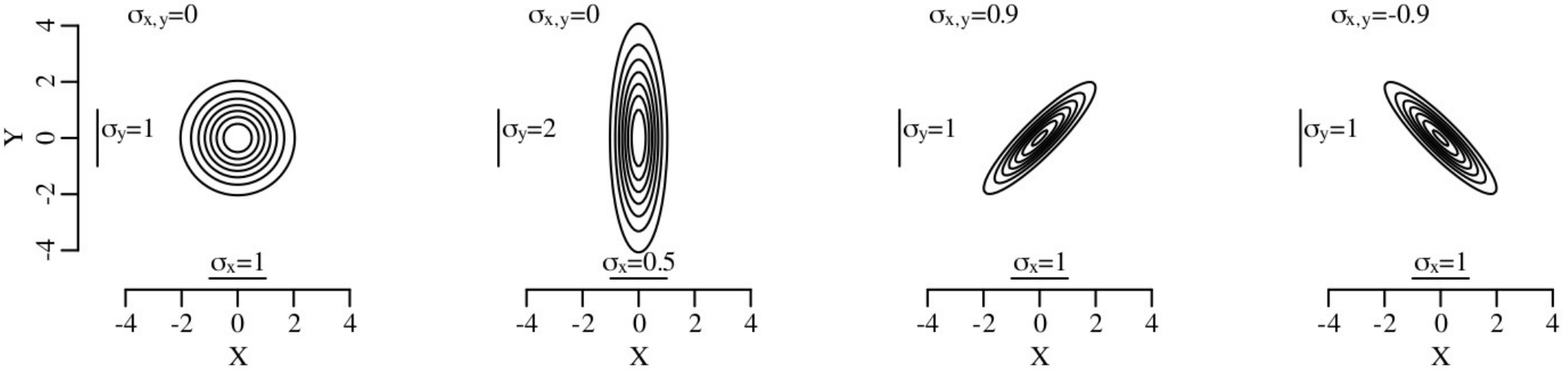


$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$





$$f(x,y|\mu_x,\mu_y,\sigma_x,\sigma_y,\sigma_{x,y}) = \frac{\exp\left(-\begin{bmatrix}(x-\mu_x) & (y-\mu_y)\end{bmatrix}\begin{bmatrix}\sigma_x^2 & \sigma_{x,y}\\ \sigma_{x,y} & \sigma_y^2\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_x\\ y-\mu_y\end{bmatrix}\Big/2\right)}{2\pi\sqrt{\begin{vmatrix}\sigma_x^2 & \sigma_{x,y}\\ \sigma_{x,y} & \sigma_y^2\end{vmatrix}}}$$



Parameter estimation

$$\mathcal{L}(\mu, \sigma | x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i | \mu, \sigma) \quad \text{with} \quad f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

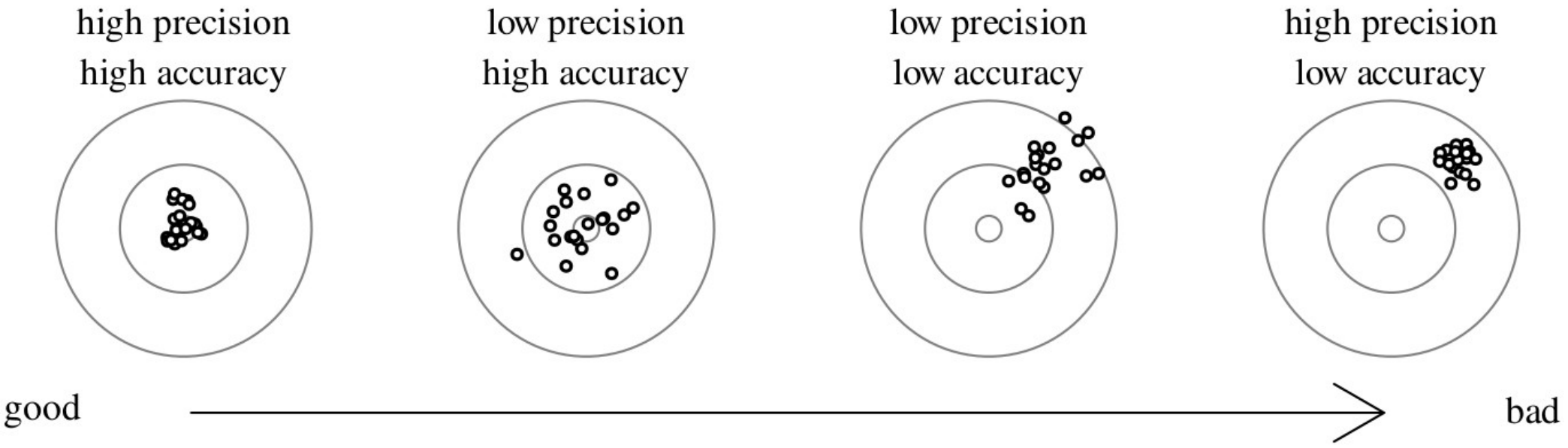
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \longleftrightarrow \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \quad \longleftrightarrow \quad s[x] = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\sigma}_{x,y} = \sum_{i=1}^{n} \frac{1}{n}(x_i - \mu_x)(y_i - \mu_y) \quad \longleftrightarrow \quad s[x,y] = \sum_{i=1}^{n} \frac{1}{n-1}(x_i - \bar{x})(y_i - \bar{y})$$

Statistics for geoscientists

Error propagation


high precision
high accuracy
low precision
high accuracy
low precision
low accuracy
high precision
low accuracy
good
bad

$z = g(x) \longrightarrow$ given x_i, for $i = 1...n$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s[x] = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$\bar{z} = g(\bar{x}) \longrightarrow$ What is $s[z]$?

Linear approximation

$$z = g(x)$$

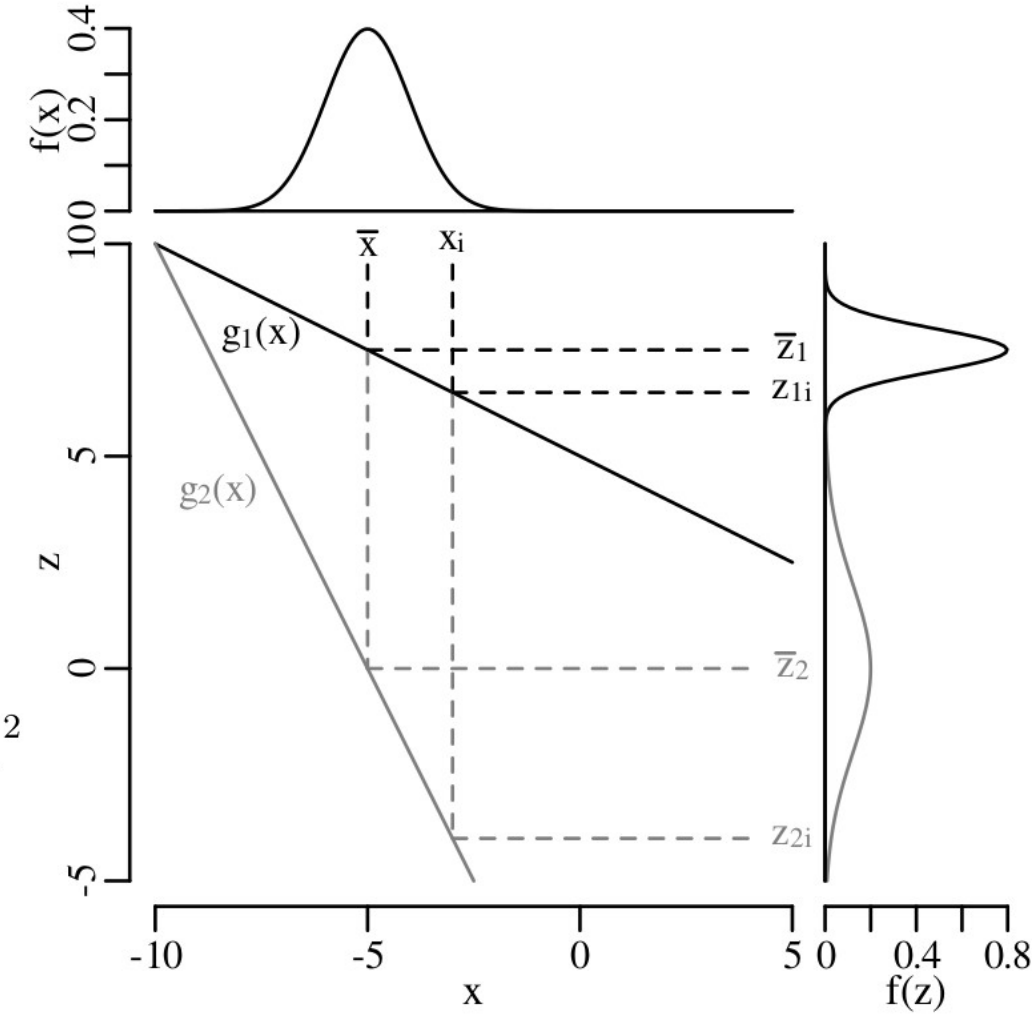
$$(z_i - \bar{z}) \approx \frac{\partial z}{\partial x}(x_i - \bar{x})$$

$$s[z]^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$$

$$s[z]^2 \approx \frac{1}{n-1}\sum_{i=1}^{n}\left[(x_i - \overline{x})\frac{\partial z}{\partial x}\right]^2$$

$$= \left[\frac{\partial z}{\partial x}\right]^2 \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right] = \left[\frac{\partial z}{\partial x}\right]^2 s[x]^2$$

$$\Rightarrow s[z] = \left|\frac{\partial z}{\partial x}\right| s[x]$$

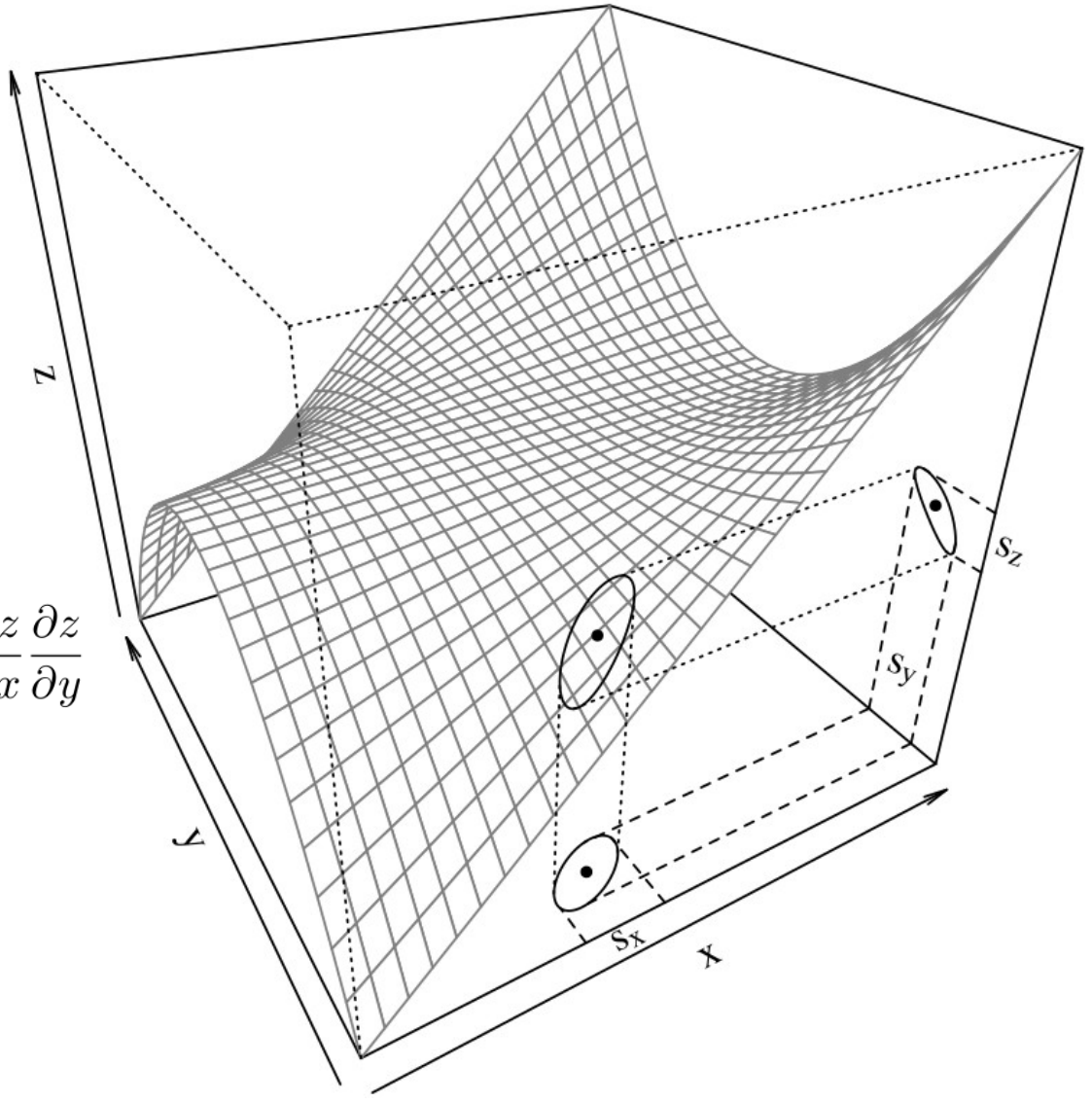


$$z = g(x, y)$$

$$s[z]^2 \approx \frac{1}{n-1} \sum_{i=1}^{n} \left[(x_i - \overline{x}) \frac{\partial z}{\partial x} + (y_i - \overline{y}) \frac{\partial z}{\partial y} \right]^2$$

$$s[z]^2 \approx s[x]^2 \left(\frac{\partial z}{\partial x} \right)^2 + s[y]^2 \left(\frac{\partial z}{\partial y} \right)^2 + 2\, s[x, y] \frac{\partial z}{\partial x} \frac{\partial z}{\partial y}$$

$$s[z]^2 \approx \begin{bmatrix} \frac{\partial z}{\partial x} \frac{\partial z}{\partial y} \end{bmatrix} \begin{bmatrix} s[x]^2 & s[x, y] \\ s[x, y] & s[y]^2 \end{bmatrix} \begin{bmatrix} \frac{\partial z}{\partial x} \\ \frac{\partial z}{\partial y} \end{bmatrix}$$



$$s[z]^2 \approx s[x]^2 \left(\frac{\partial z}{\partial x}\right)^2 + s[y]^2 \left(\frac{\partial z}{\partial y}\right)^2 + 2\ s[x,y]\frac{\partial z}{\partial x}\frac{\partial z}{\partial y}$$

addition

$$z = a + bx + cy$$

$$\frac{\partial z}{\partial x} = b \text{ and } \frac{\partial z}{\partial y} = c$$

$$s[z]^2 = b^2 s[x]^2 + c^2 s[y]^2 + 2bc\ s[x,y]$$

$$z = x + y \longrightarrow s[z]^2 = s[x]^2 + s[y]^2$$

$$s[z]^2 \approx s[x]^2 \left(\frac{\partial z}{\partial x}\right)^2 + s[y]^2 \left(\frac{\partial z}{\partial y}\right)^2 + 2\ s[x,y]\frac{\partial z}{\partial x}\frac{\partial z}{\partial y}$$

subtraction

$$z = ax - by$$

$$\frac{\partial z}{\partial x} = a \text{ and } \frac{\partial z}{\partial y} = -b$$

$$s[z]^2 = a^2 s[x]^2 + b^2 s[y]^2 - 2ab\ s[x,y]$$

$$z = x - y \longrightarrow s[z]^2 = s[x]^2 + s[y]^2$$

$$s[z]^2 \approx s[x]^2 \left(\frac{\partial z}{\partial x}\right)^2 + s[y]^2 \left(\frac{\partial z}{\partial y}\right)^2 + 2\ s[x,y]\frac{\partial z}{\partial x}\frac{\partial z}{\partial y}$$

multiplication

$$z = axy$$

$$\frac{\partial z}{\partial x} = ay \text{ and } \frac{\partial z}{\partial y} = ax$$

$$s[z]^2 = (ay)^2 s[x]^2 + (ax)^2 s[y]^2 + 2(ay)(ax)\ s[x,y]$$

$$\left(\frac{s[z]}{z}\right)^2 = \left(\frac{ay}{axy}s[x]\right)^2 + \left(\frac{ax}{axy}s[y]\right)^2 + 2\left(\frac{ay}{axy}\right)\left(\frac{ax}{axy}\right)s[x,y]$$

$$\left(\frac{s[z]}{z}\right)^2 = \left(\frac{s[x]}{x}\right)^2 + \left(\frac{s[y]}{y}\right)^2 + 2\frac{s[x,y]}{xy}$$

addition	$z = a + bx + cy$	$s[z]^2 = b^2 s[x]^2 + c^2 s[y]^2 + 2bc\ s[x,y]$
subtraction	$z = ax - by$	$s[z]^2 = a^2 s[x]^2 + b^2 s[y]^2 - 2ab\ s[x,y]$
multiplication	$z = axy$	$\left(\frac{s[z]}{z}\right)^2 = \left(\frac{s[x]}{x}\right)^2 + \left(\frac{s[y]}{y}\right)^2 + 2\frac{s[x,y]}{xy}$
division	$z = a\frac{x}{y}$	$\left(\frac{s[z]}{z}\right)^2 = \left(\frac{s[x]}{x}\right)^2 + \left(\frac{s[y]}{y}\right)^2 - 2\frac{s[x,y]}{xy}$
exponentiation	$z = ae^{bx}$	$\left(\frac{s[z]}{z}\right)^2 = b^2 s[x]^2$
logarithms	$z = a\ln[bx]$	$s[z]^2 = a^2 \left(\frac{s[x]}{x}\right)^2$
power	$z = ax^b$	$\left(\frac{s[z]}{z}\right)^2 = b^2 \left(\frac{s[x]}{x}\right)^2$

the **chain rule**

$$d = d_\circ + v_\circ t + gt^2$$

can be written as

$$d = x + y \qquad \text{where} \quad x \equiv d_\circ + v_\circ t \qquad \text{and} \qquad y \equiv gt^2$$

$$s[x]^2 = (v_\circ s[t])^2 \qquad s[y]^2 = g^2 \left(\frac{s[t]}{t}\right)^2$$

$$s[d] = \sqrt{s[x]^2 + s[y]^2}$$

$$s[d] = \sqrt{(v_\circ s[t])^2 + g^2 \left(\frac{s[t]}{t}\right)^2}$$

$$z = g(x)$$

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

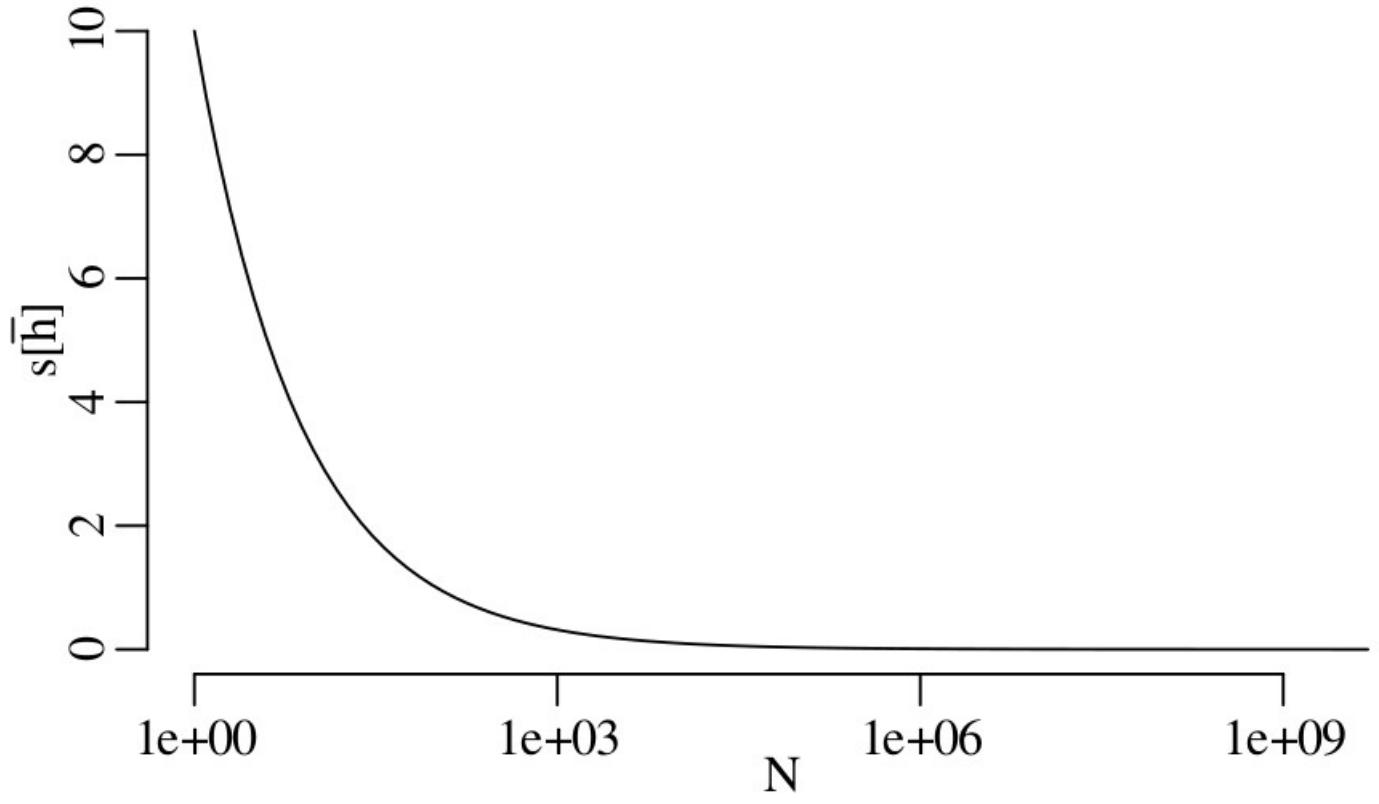
---

addition $z = a + bx + cy$ $s[z]^2 = b^2 s[x]^2 + c^2 s[y]^2$

$$s[\bar{x}]^2 = \sum_{i=1}^{n} \left(\frac{s[x_i]}{n} \right)^2 = \left(\frac{1}{n} \right)^2 \sum_{i=1}^{n} s[x_i]^2$$

$$s[\bar{x}]^2 = \left(\frac{1}{n} \right)^2 \sum_{i=1}^{n} s[x]^2 = n \left(\frac{1}{n} \right)^2 s[x]^2$$

$$s[\bar{x}] = \frac{s[x]}{\sqrt{n}}$$


0
2
4
6
8
10
s[h̄]
1e+00
1e+03
1e+06
1e+09
N

Statistics for geoscientists

An introduction to R