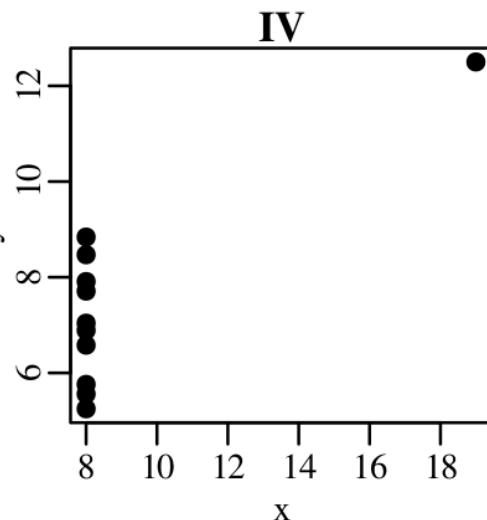
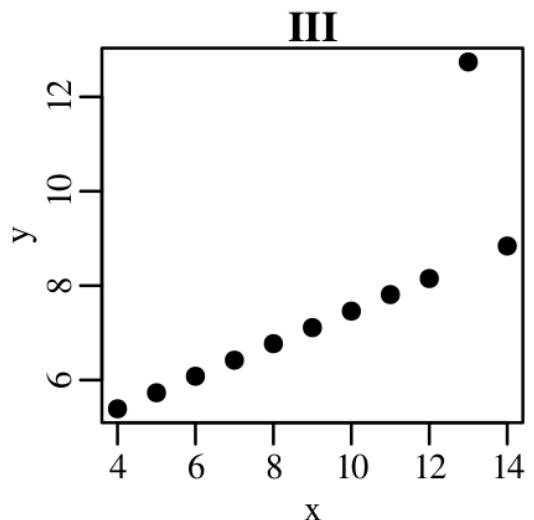
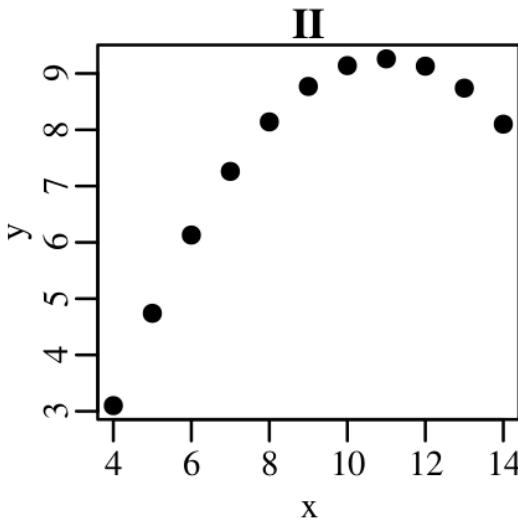
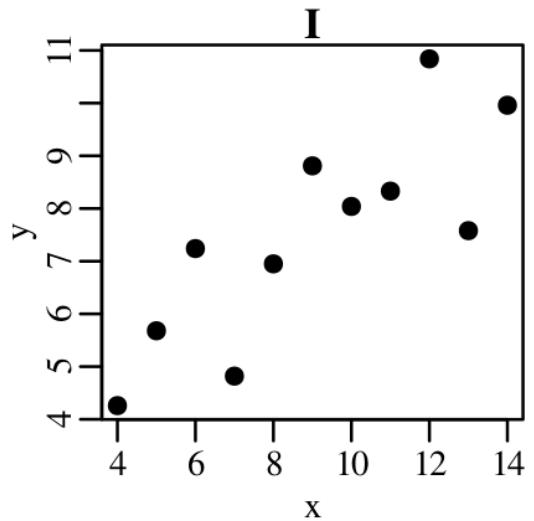


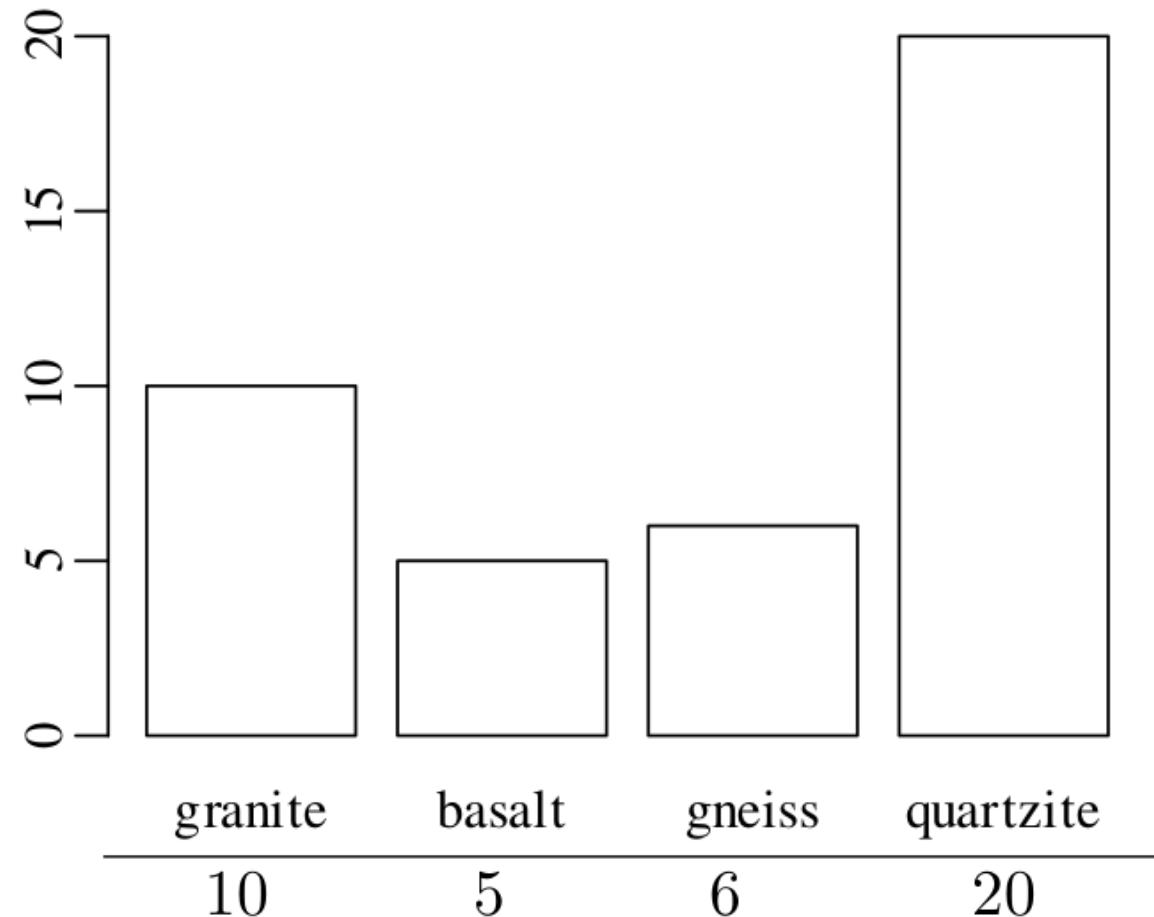
Statistics for geoscientists

Plotting data

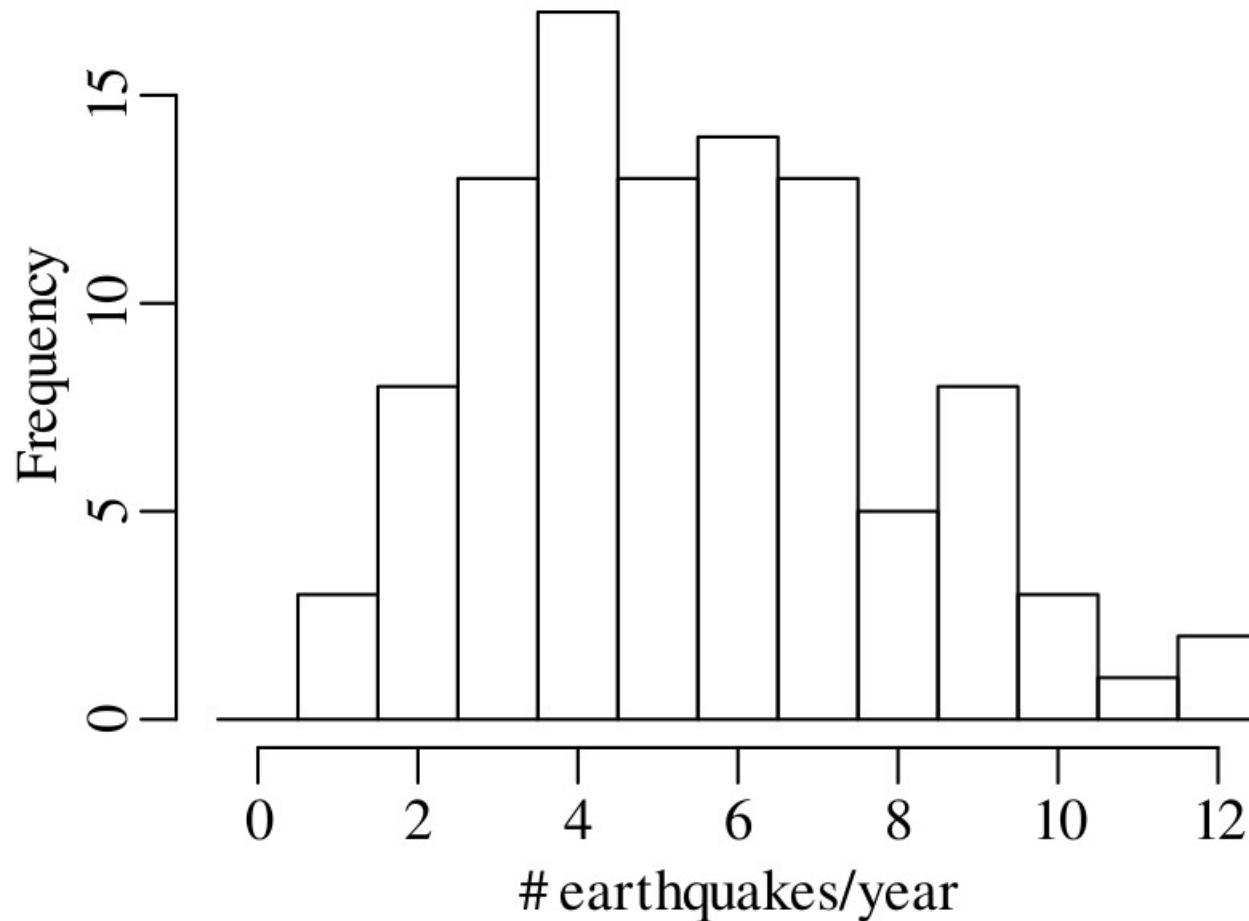
I	II		III		IV			
x	y	x	y	x	y	x	y	
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58	– the mean of x is 9
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76	
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71	– the variance of x is 11
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84	
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47	– the mean of y is 7.50
14.0	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	– the variance of y is 4.125
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	– the correlation coefficient of x and y is 0.816
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 best fit line is given by $y = 3.00 + 0.500x$



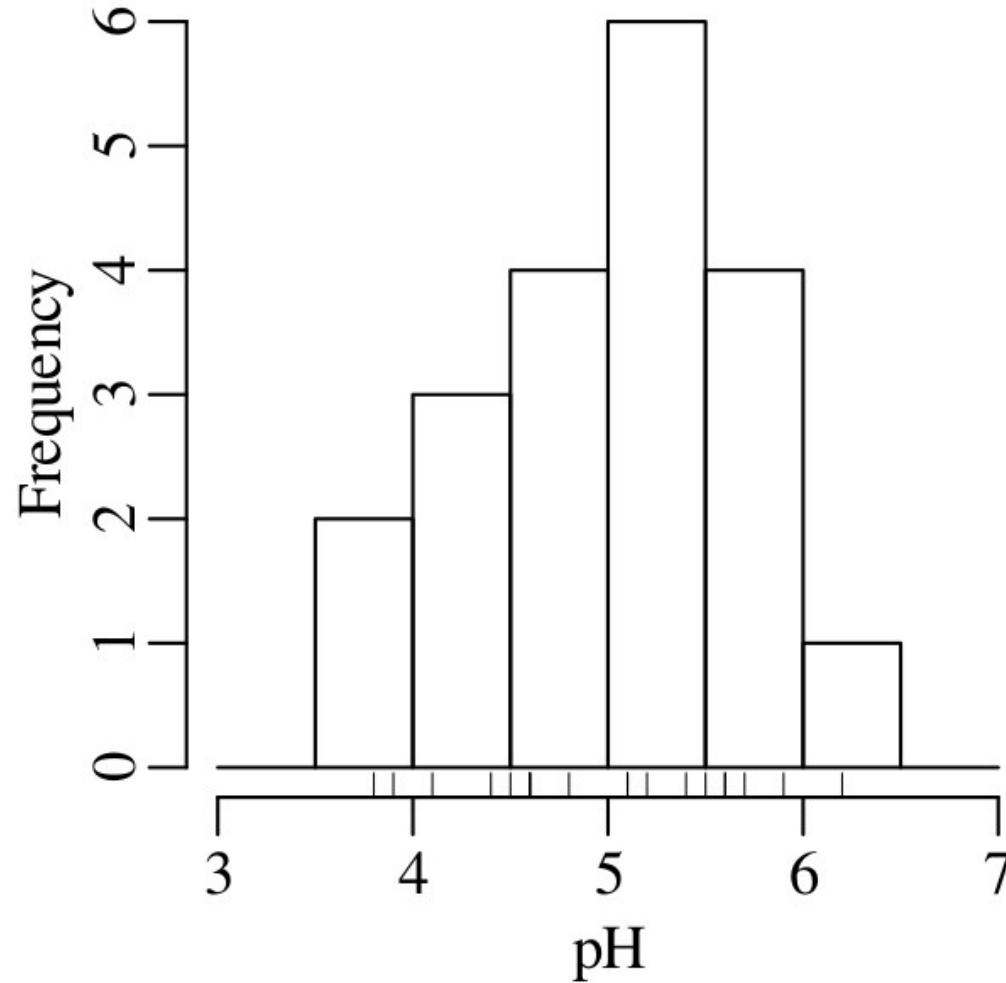
Categorical data



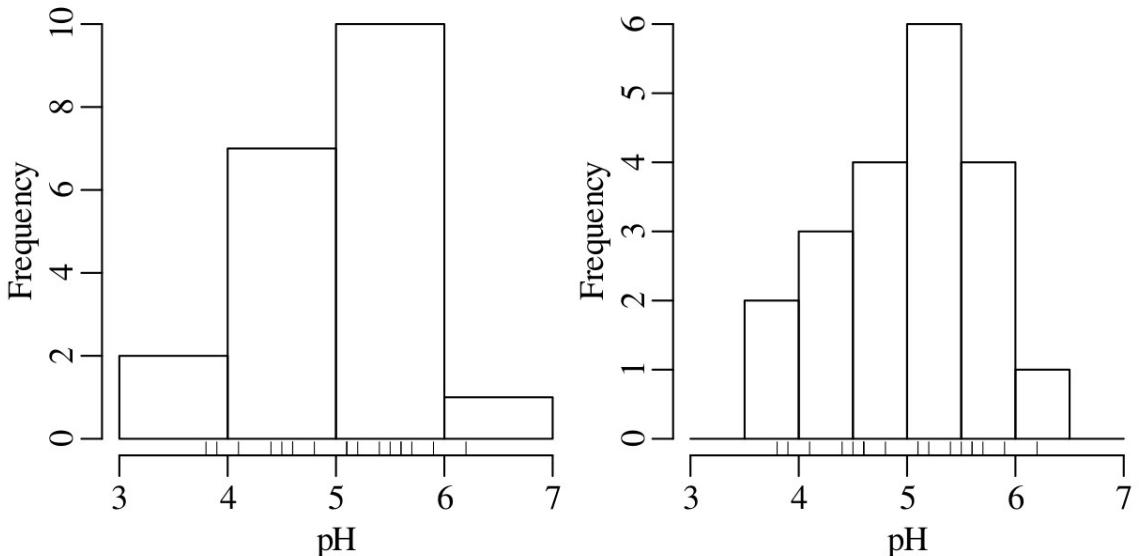
Count data



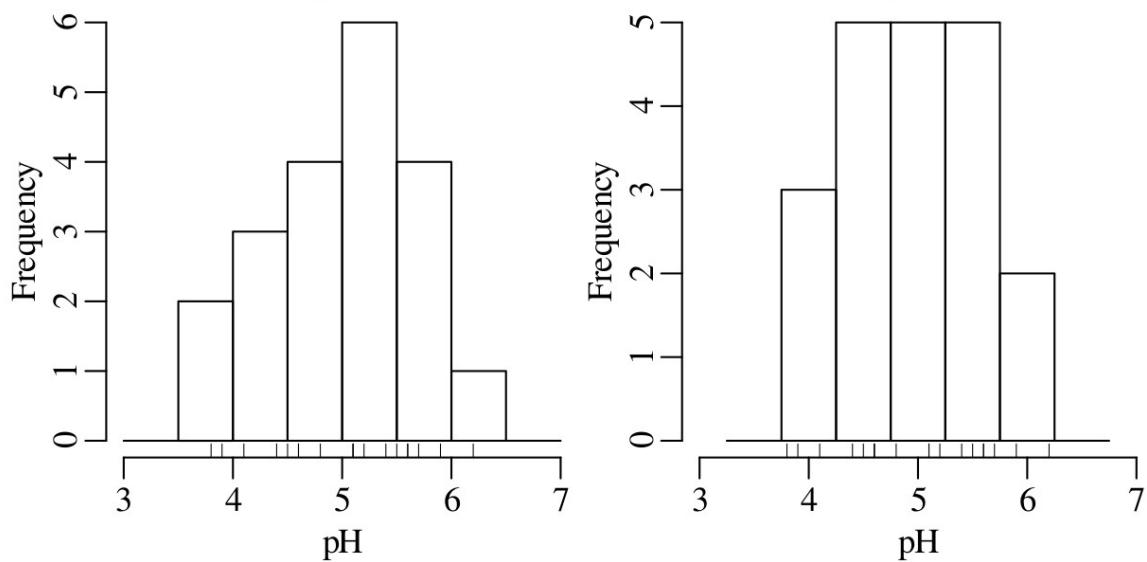
Continuous data



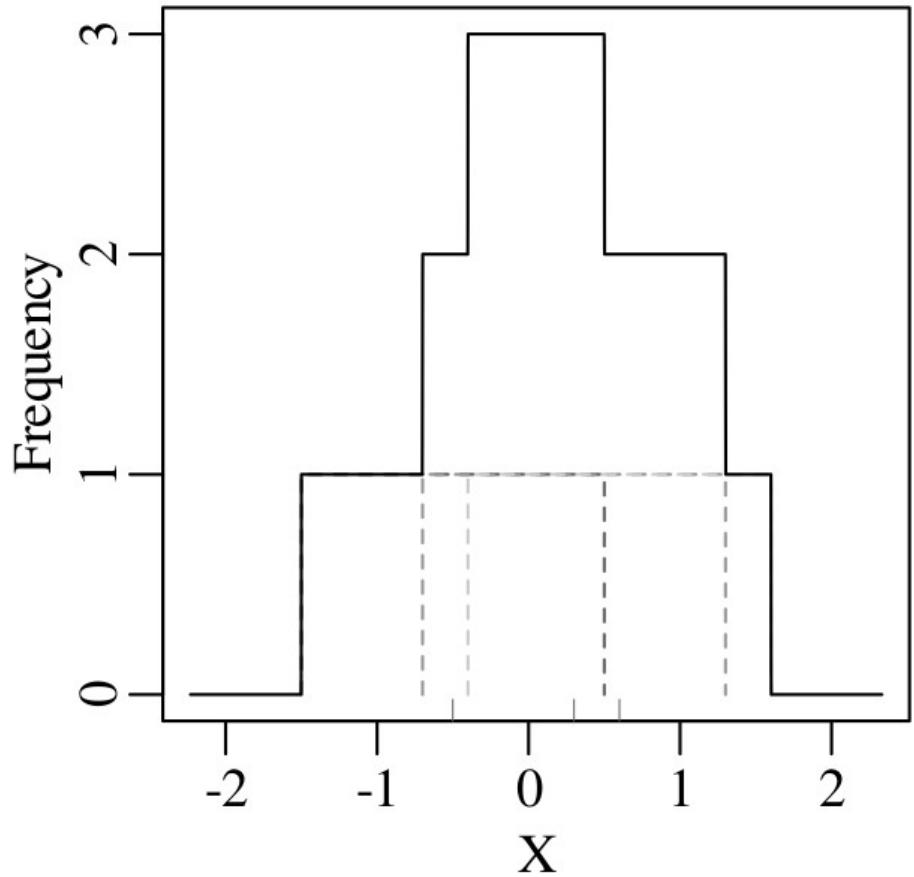
i. How many bins?



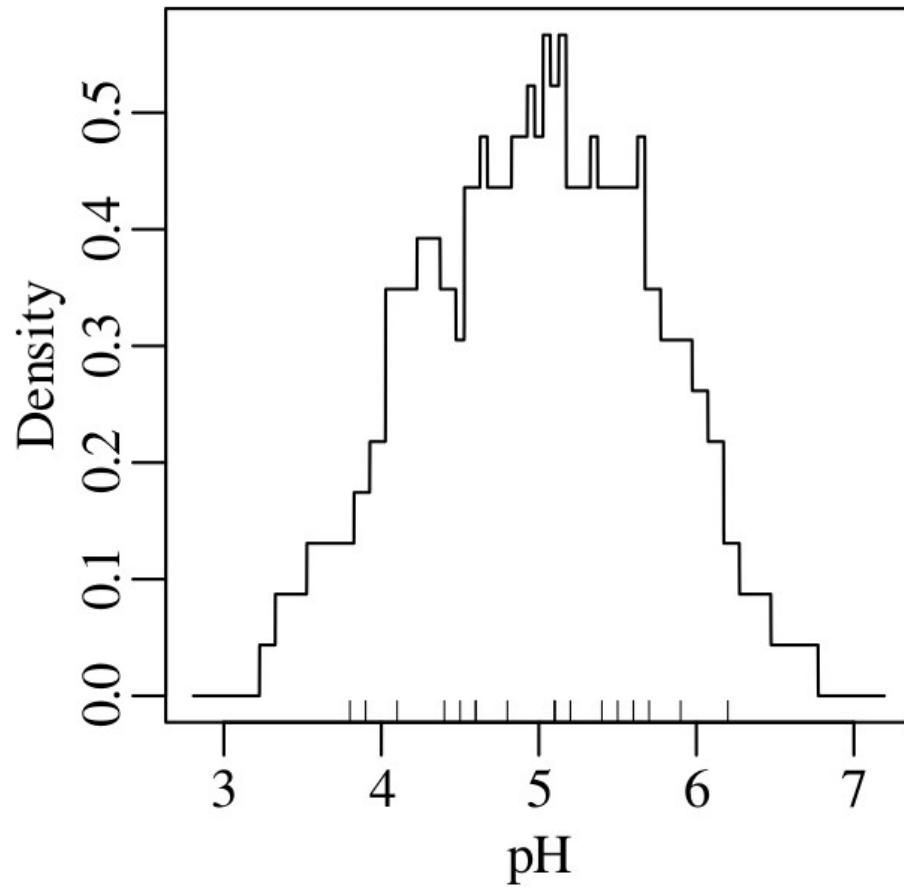
ii. Where to place the bins?



Kernel Density Estimate

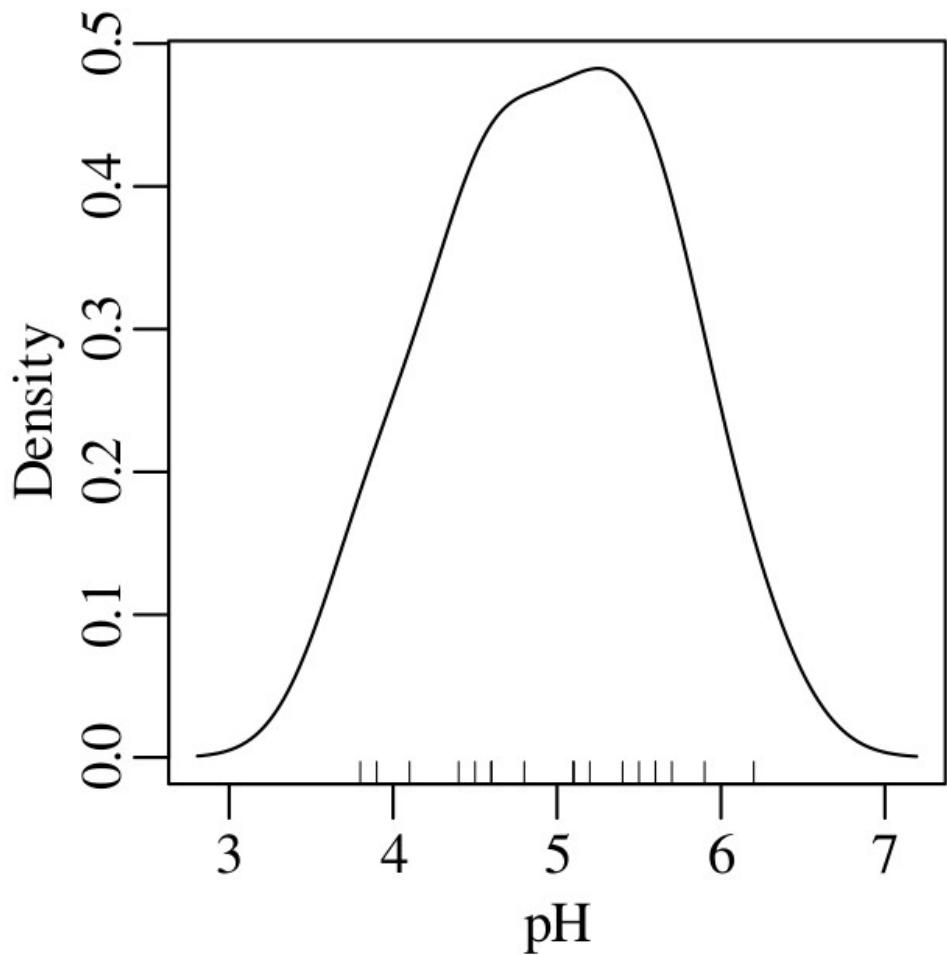
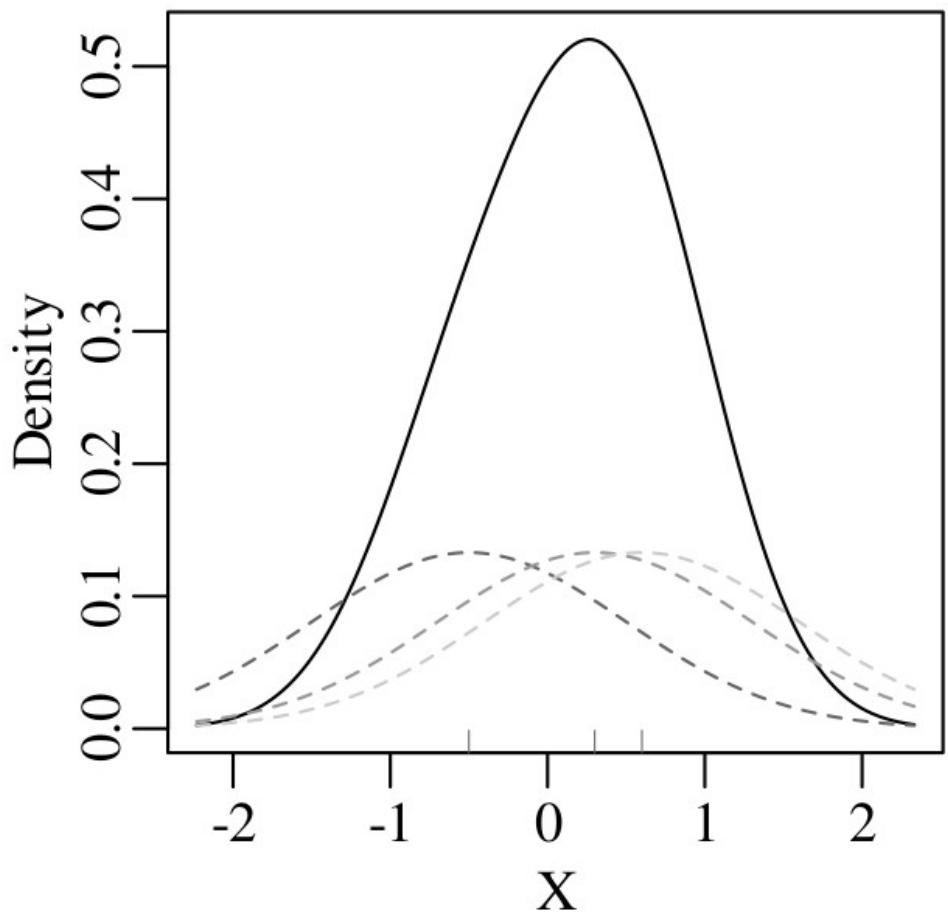


$$KDE(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

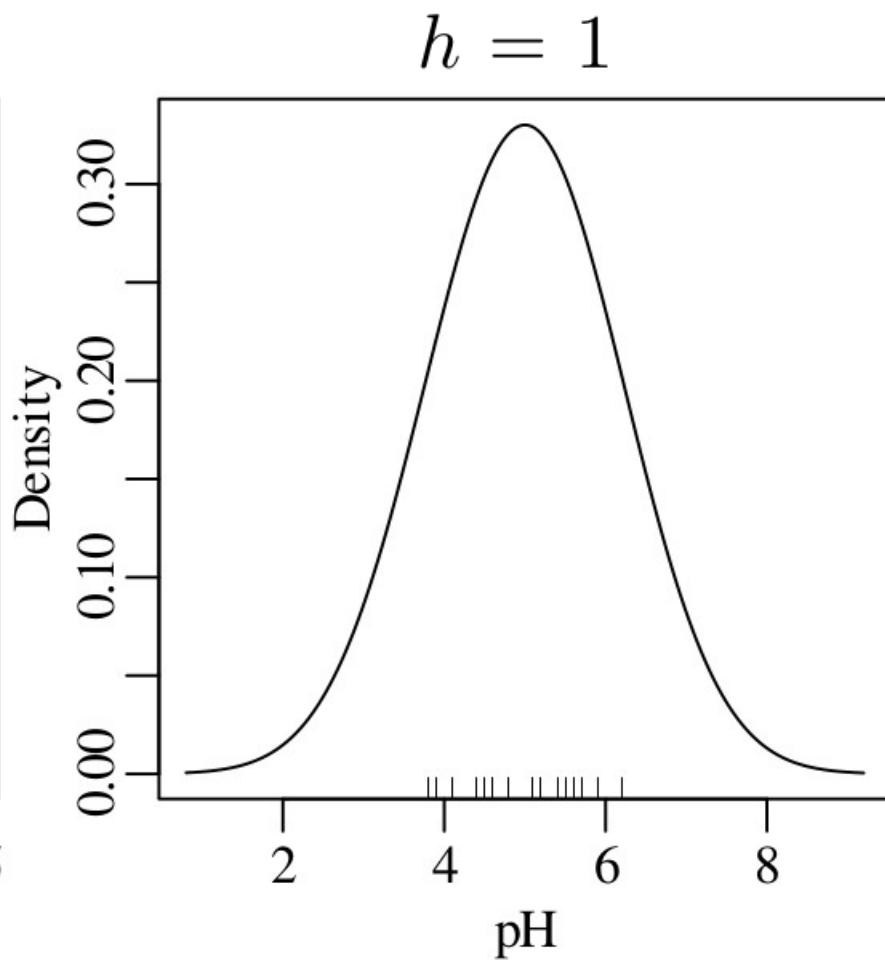
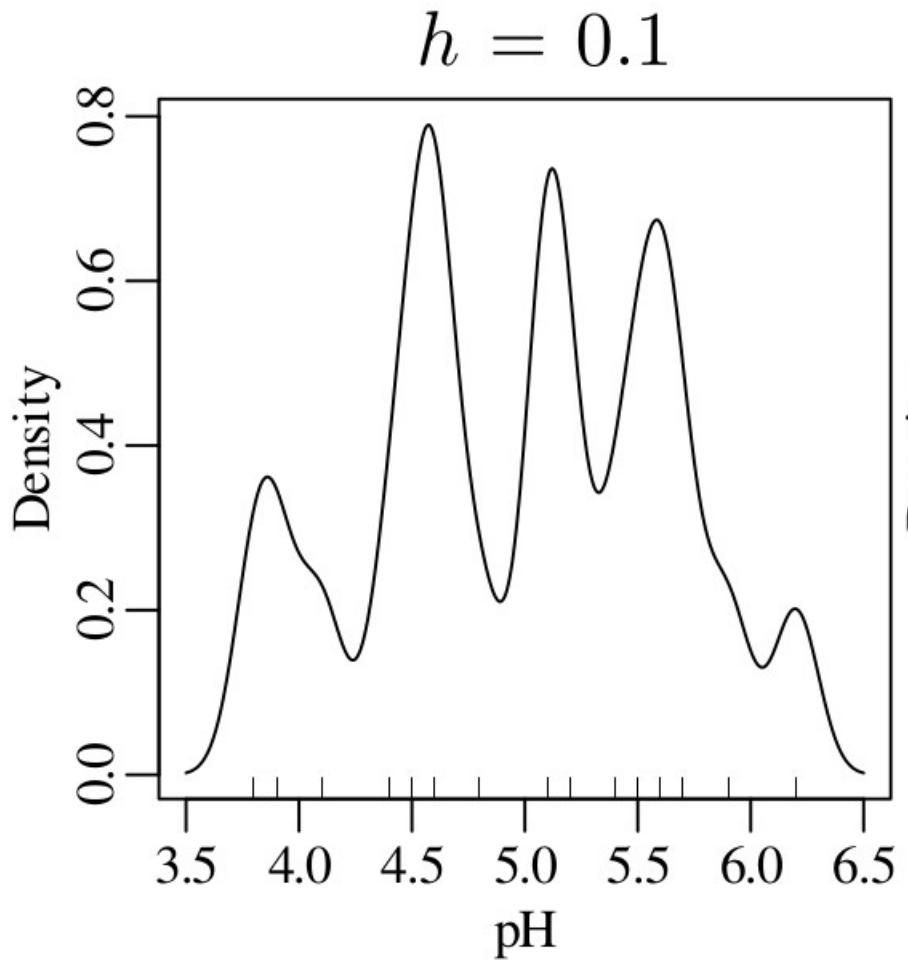


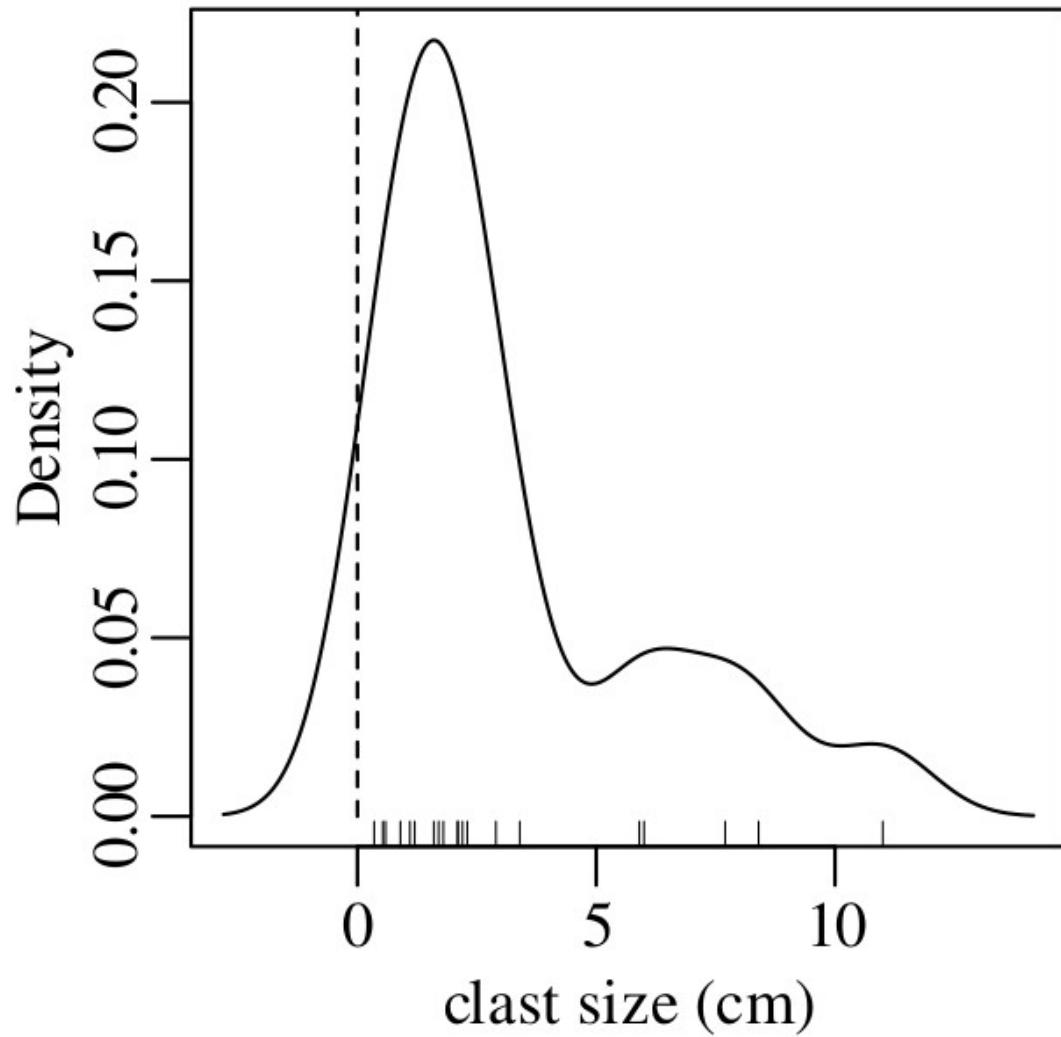
Gaussian kernel:

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

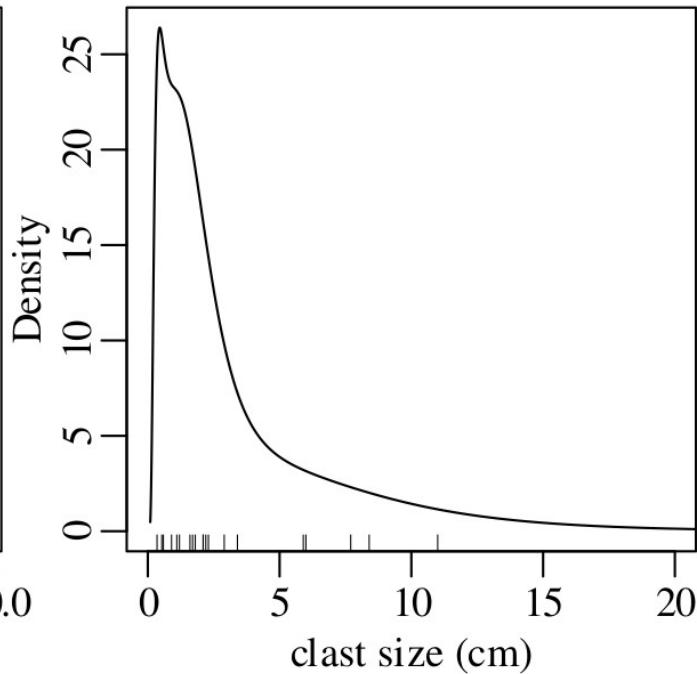
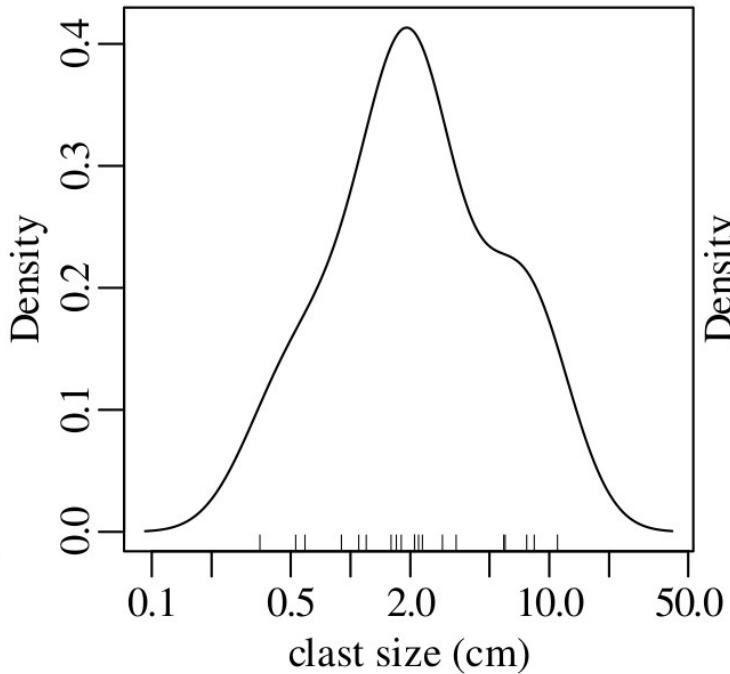
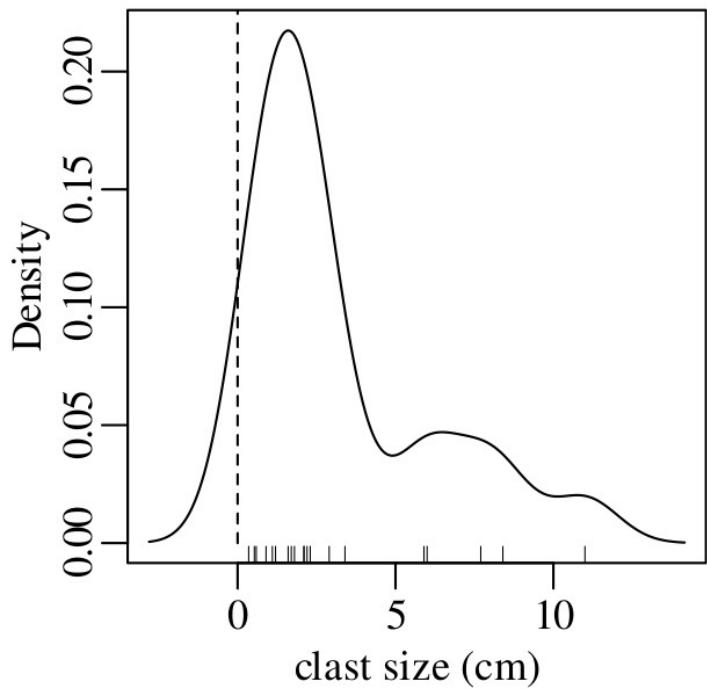


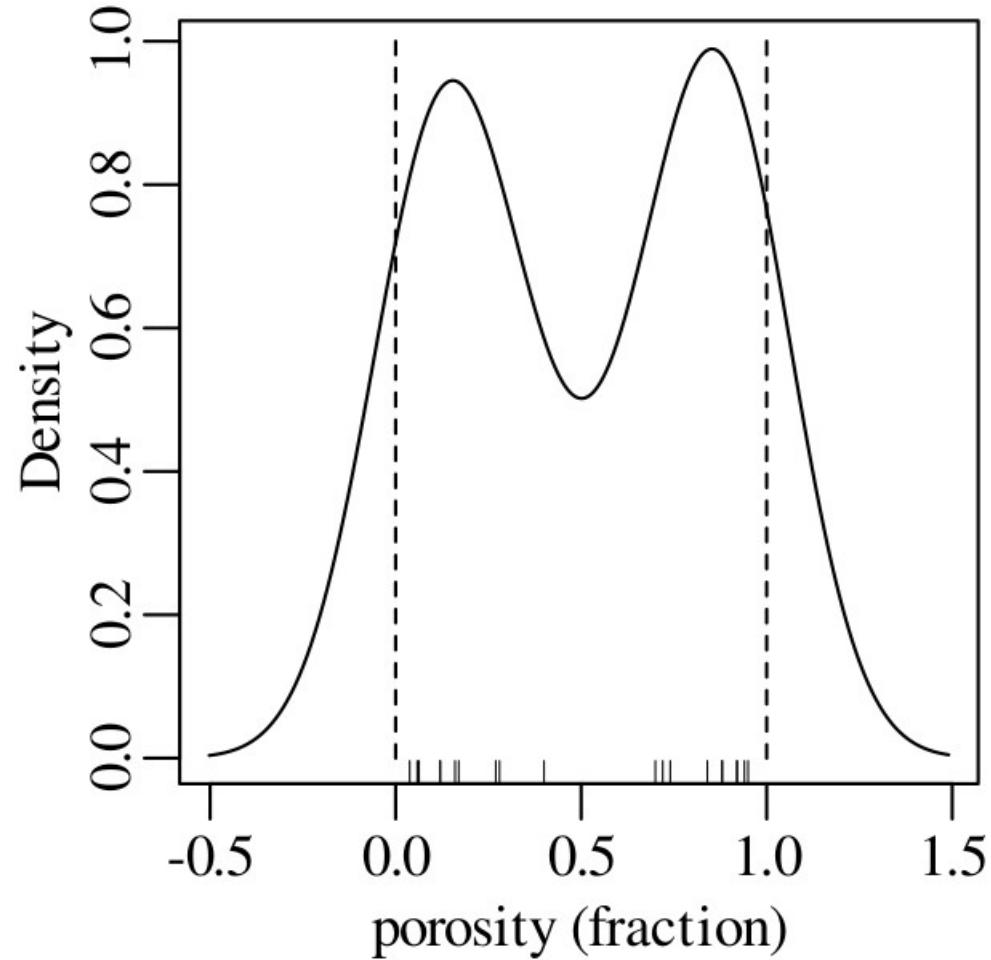
bandwidth





logarithmic transformation

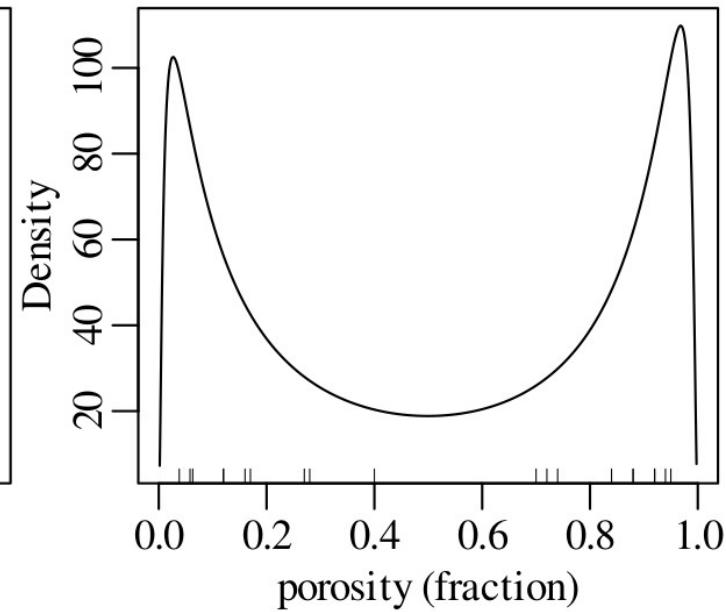
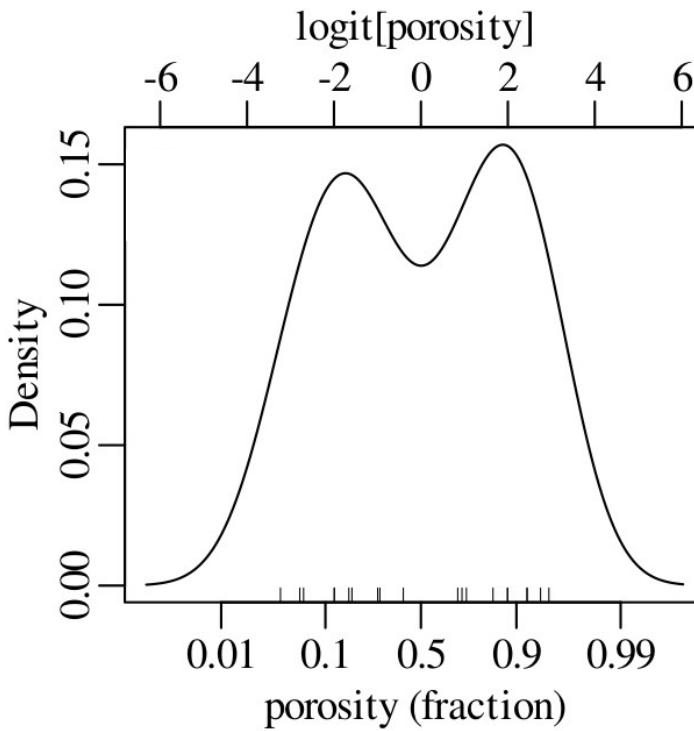
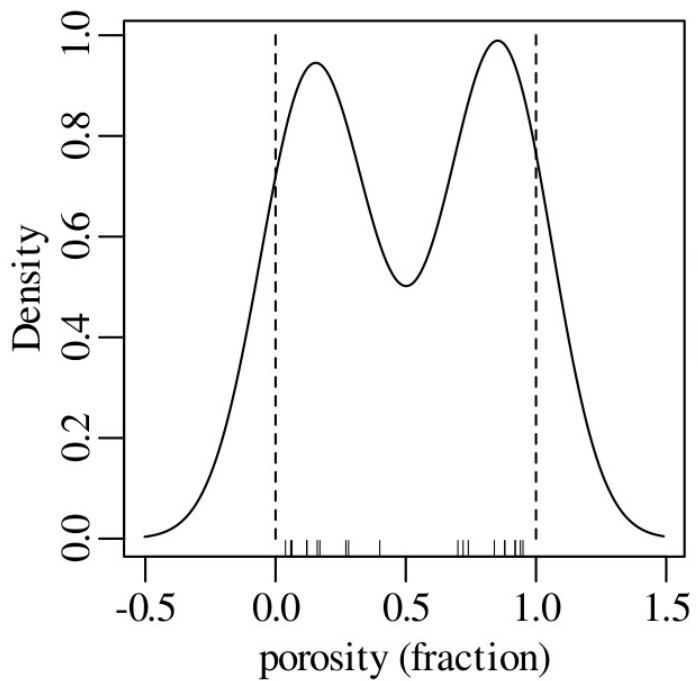




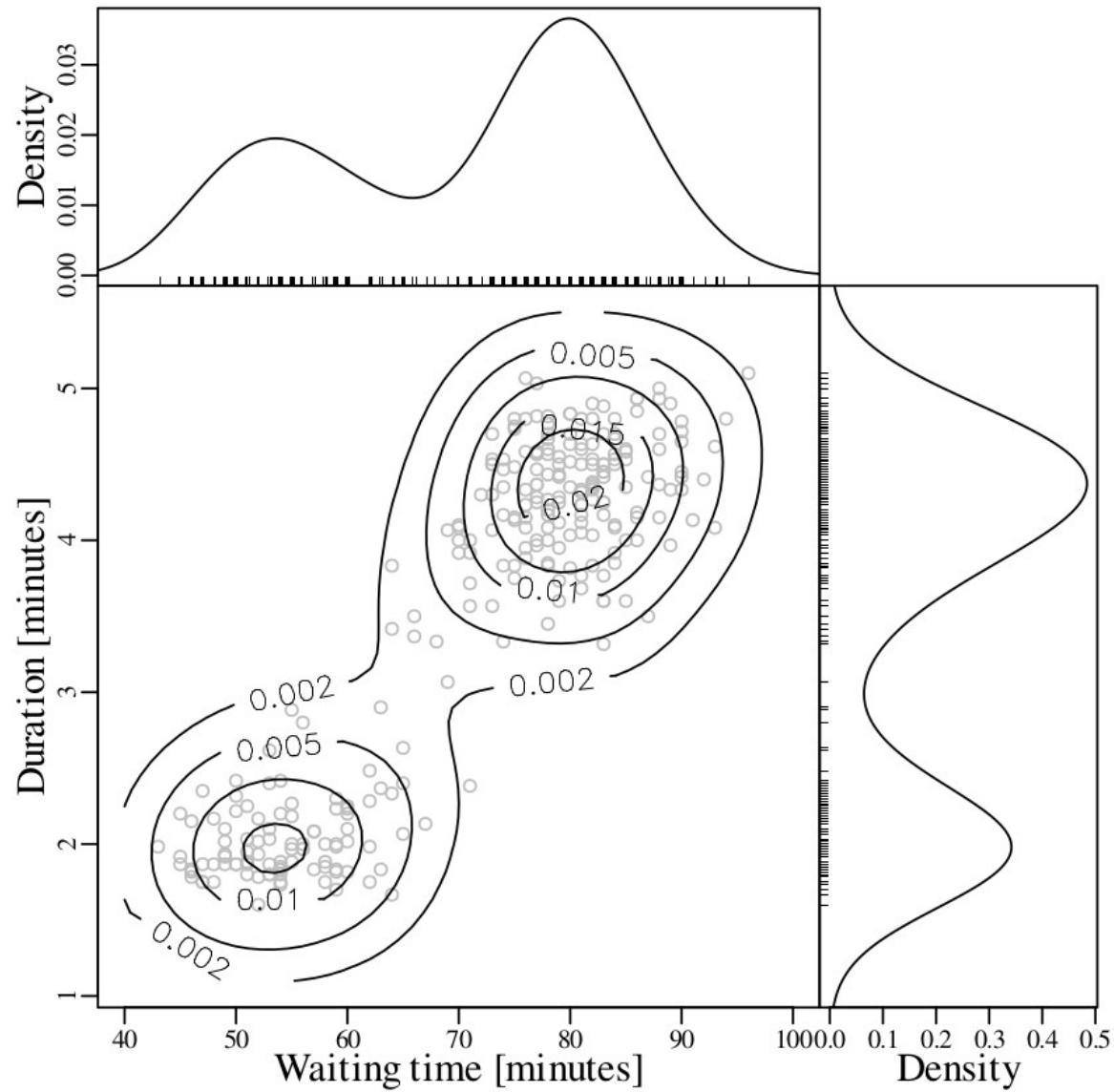
logistic transformation

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

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

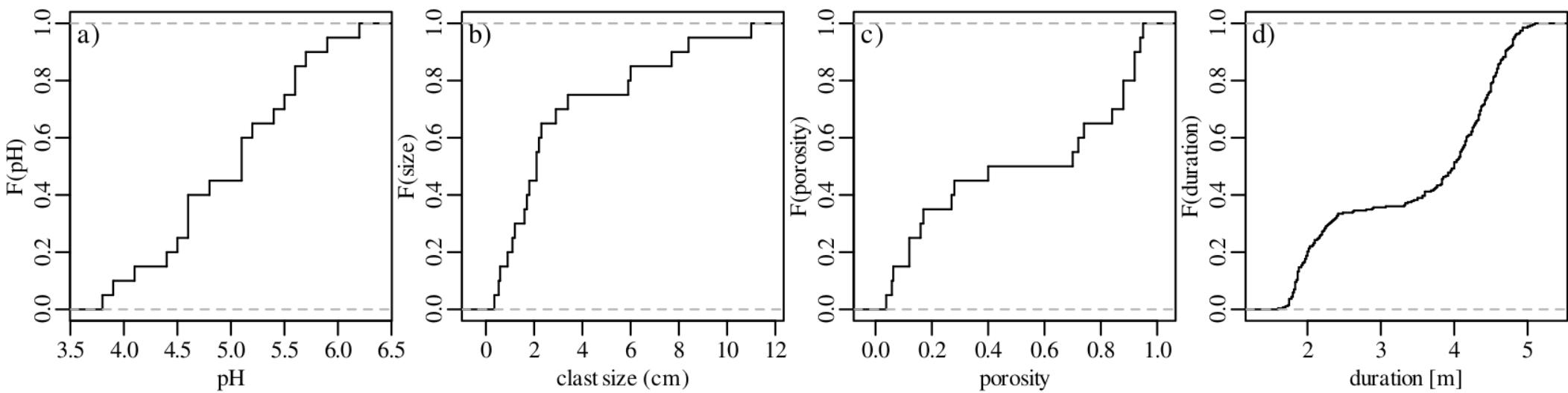


Multivariate distributions



Empirical cumulative distribution fuctions

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

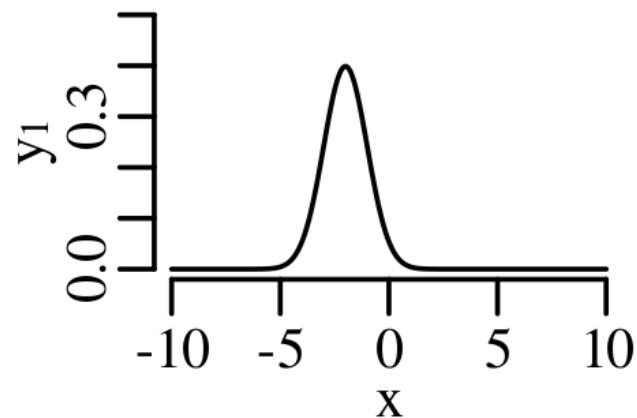


Statistics for geoscientists

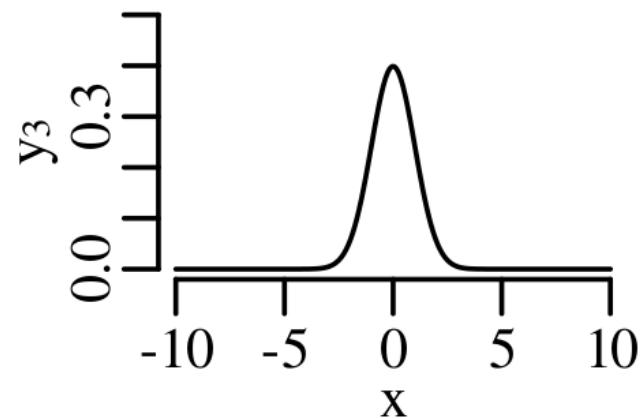
Summary statistics

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	– the mean of x is 9
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	– the variance of x is 11
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	– the mean of y is 7.50
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	– the variance of y is 4.125
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	– the correlation coefficient of x and y is 0.816
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 best fit line is given by $y = 3.00 + 0.500x$

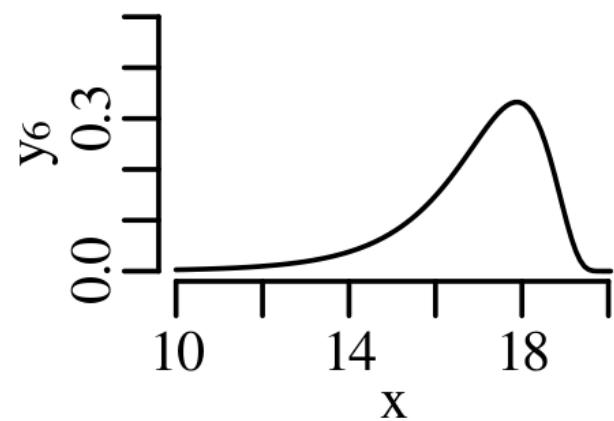
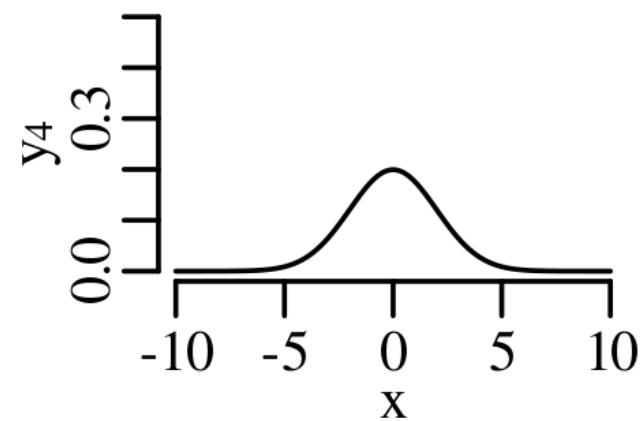
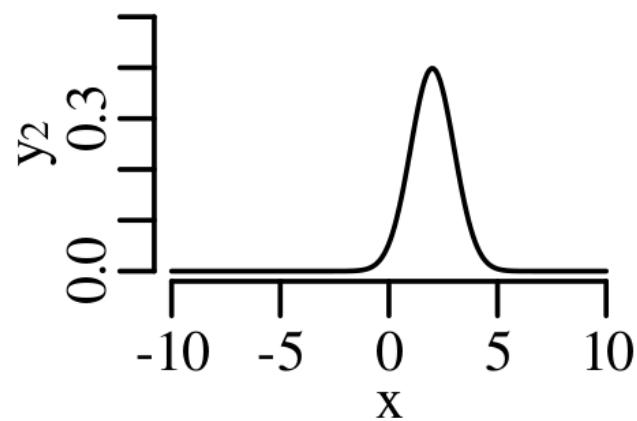
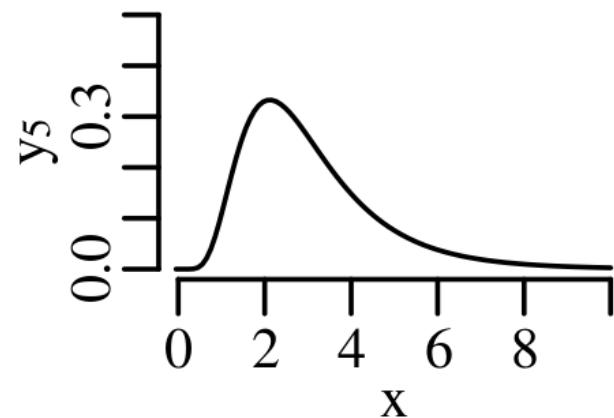
Location



Dispersion



Shape

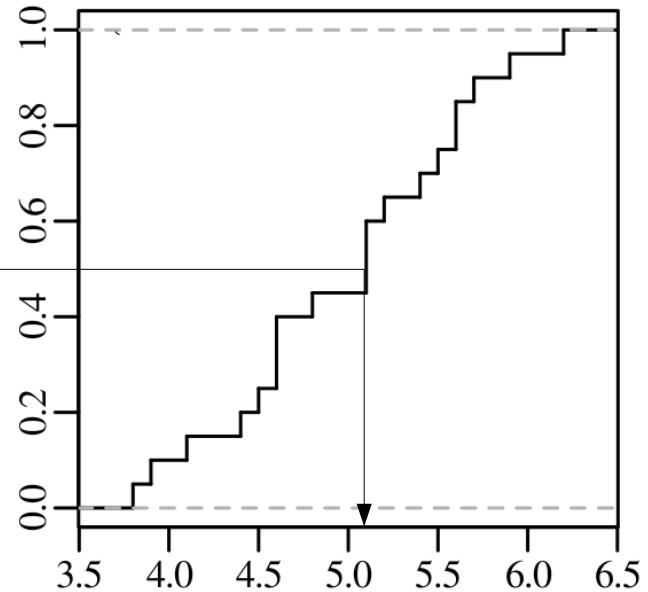
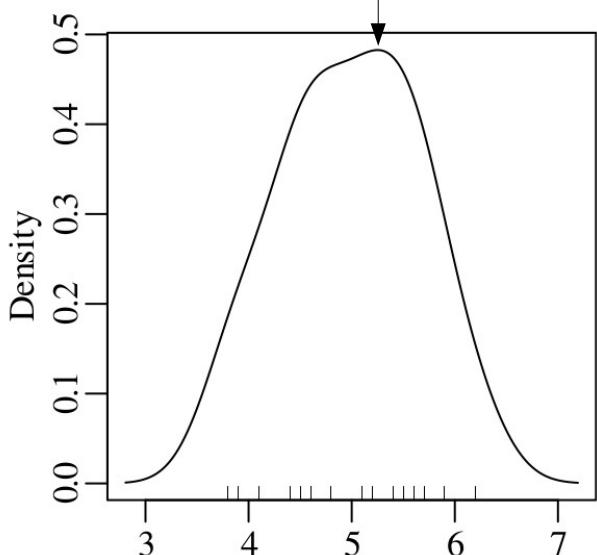


Location

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

Median $\text{med}(x) = x_1 < \dots < x_{\frac{n}{2}} < \dots < x_n$

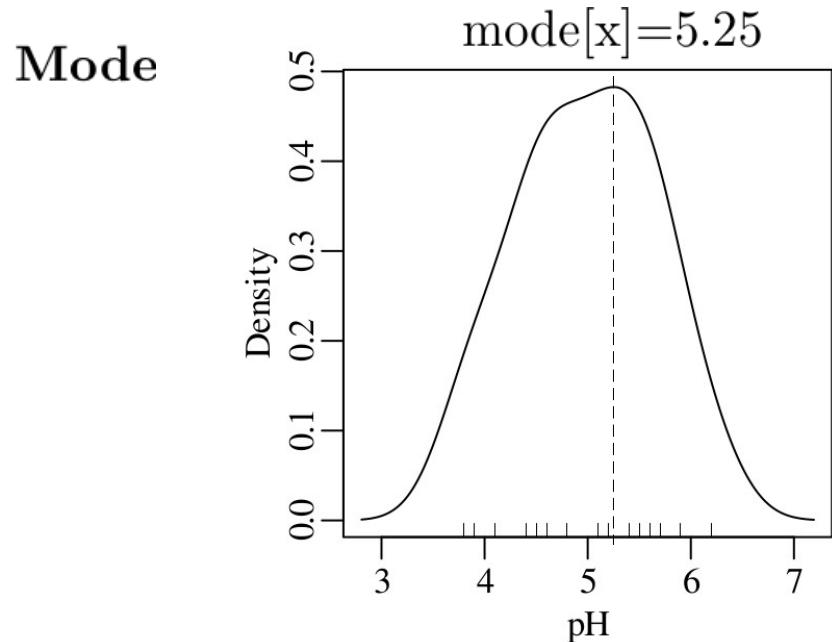
Mode

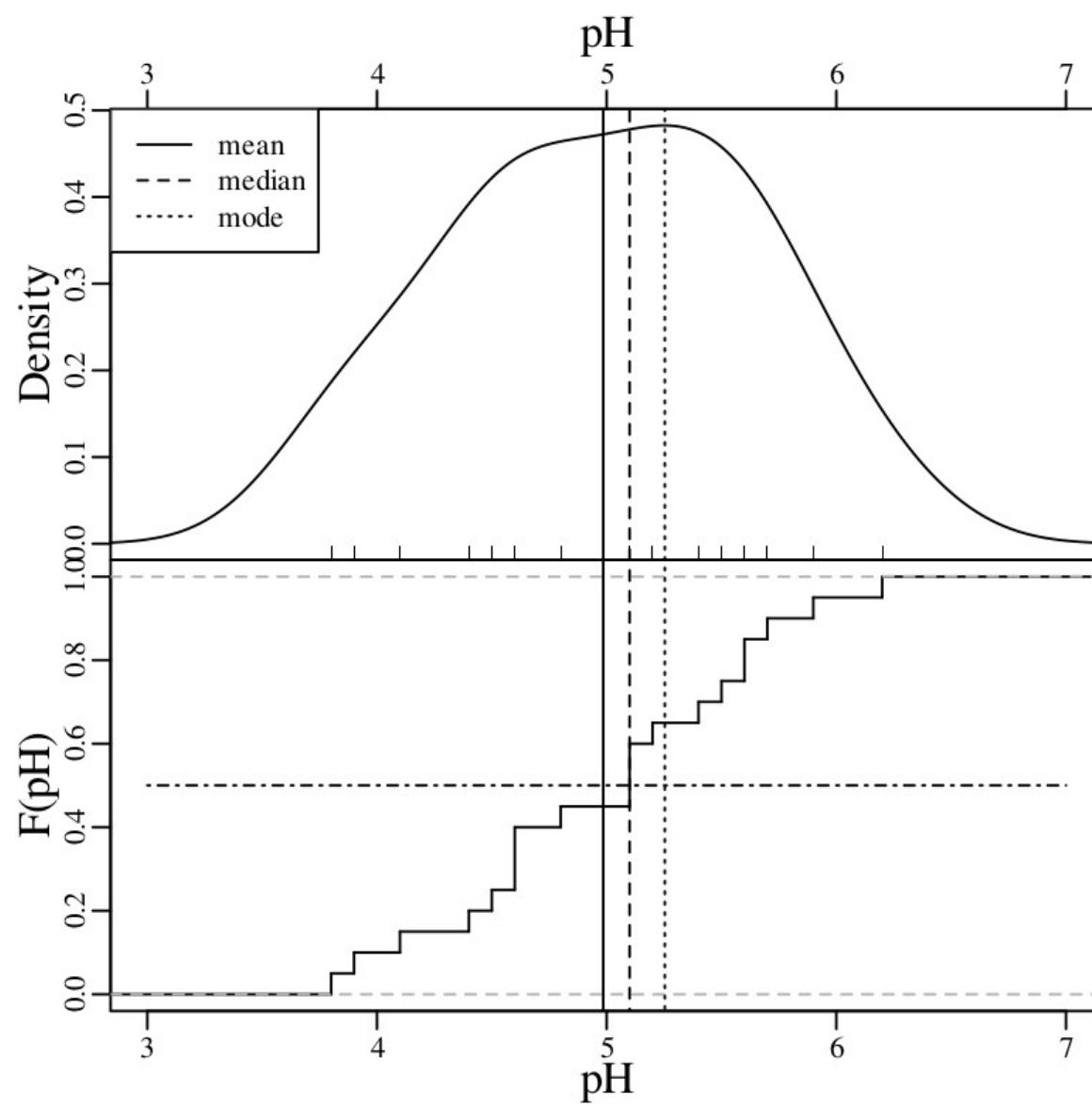


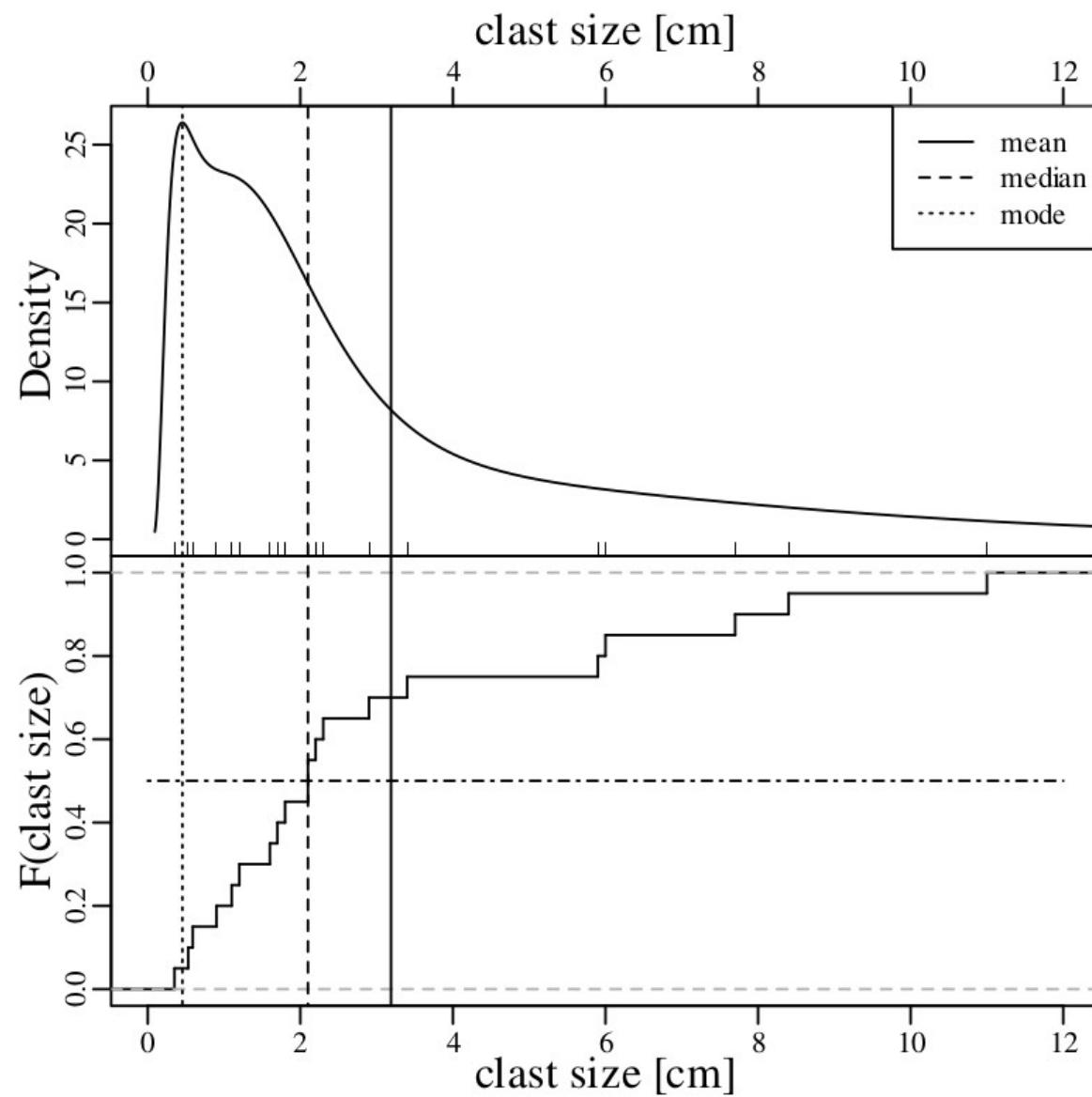
pH data 6.2, 4.4, 5.6, 5.2, 4.5, 5.4, 4.8, 5.9, 3.9, 3.8, 5.1, 4.1, 5.1, 5.5, 5.1, 4.6, 5.7, 4.6, 4.6, 5.6

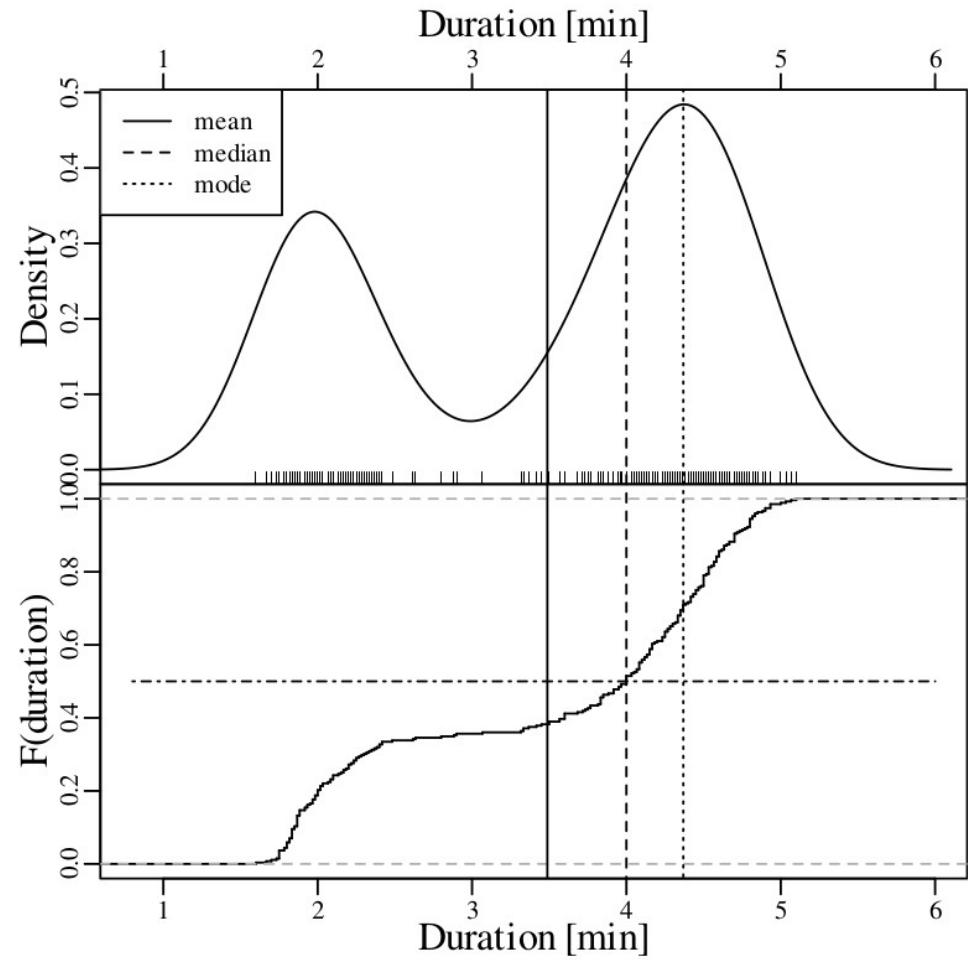
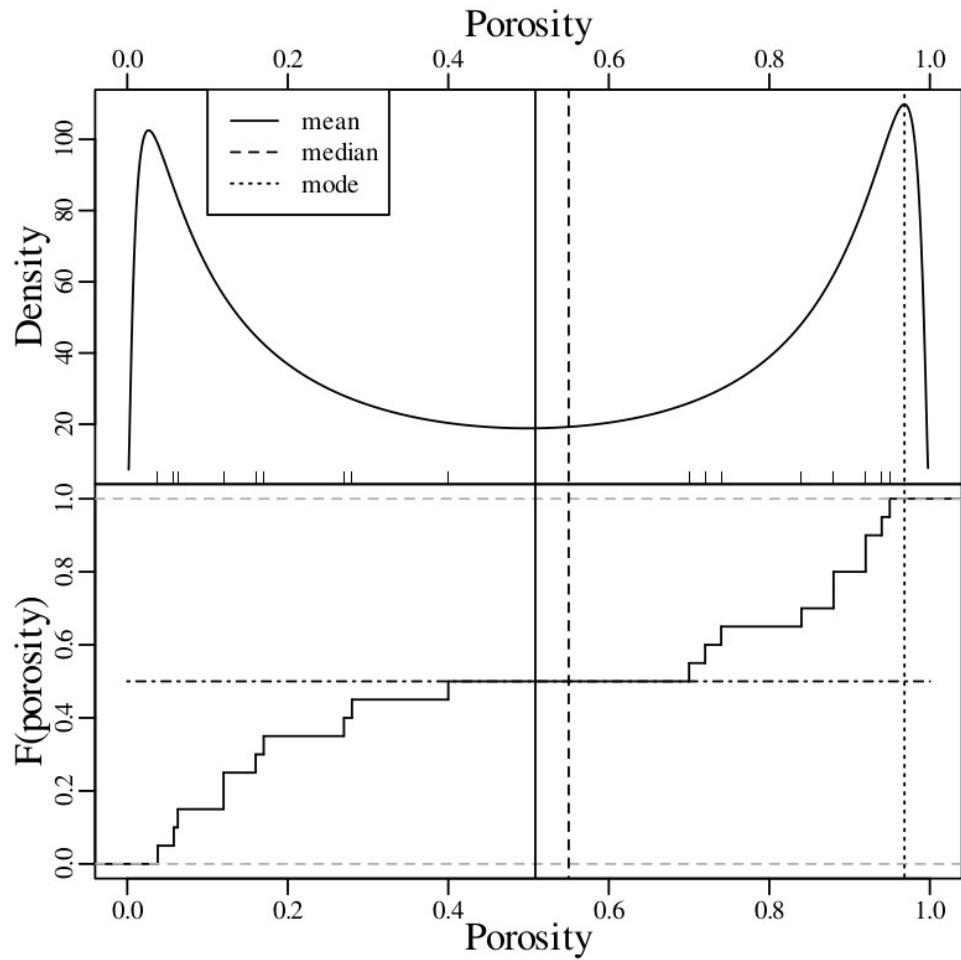
$$\text{Mean } \bar{x} = \frac{6.2 + 4.4 + 5.6 + 5.2 + 4.5 + 5.4 + 4.8 + 5.9 + 3.9 + 3.8 + 5.1 + 4.1 + 5.1 + 5.5 + 5.1 + 4.6 + 5.7 + 4.6 + 4.6 + 5.6}{20} = 5.00$$

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









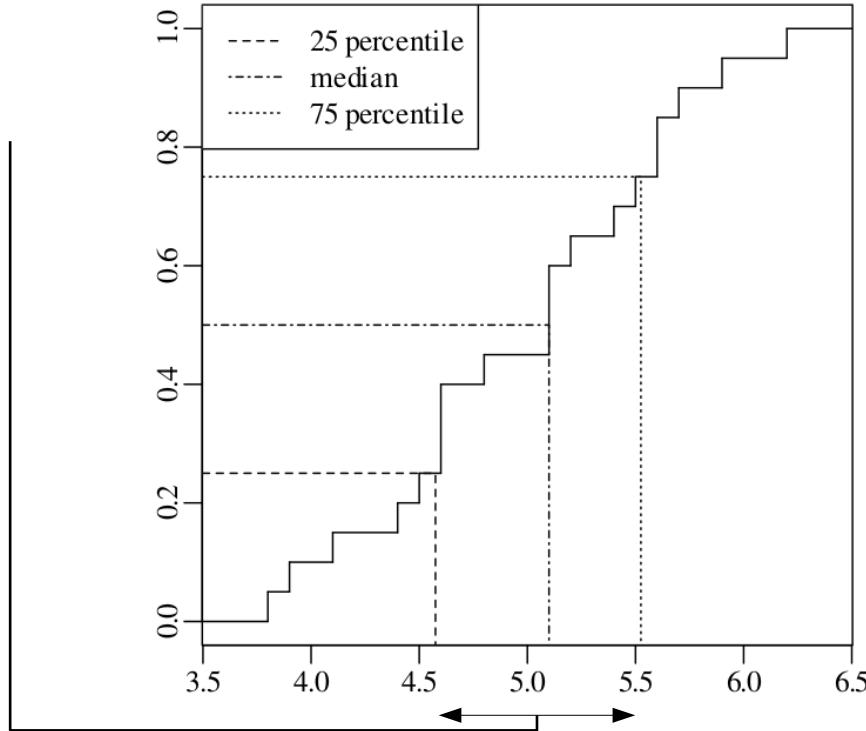
Dispersion

Standard deviation

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

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

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

Interquartile range IQR = $5.55 - 4.55 = 1.00$

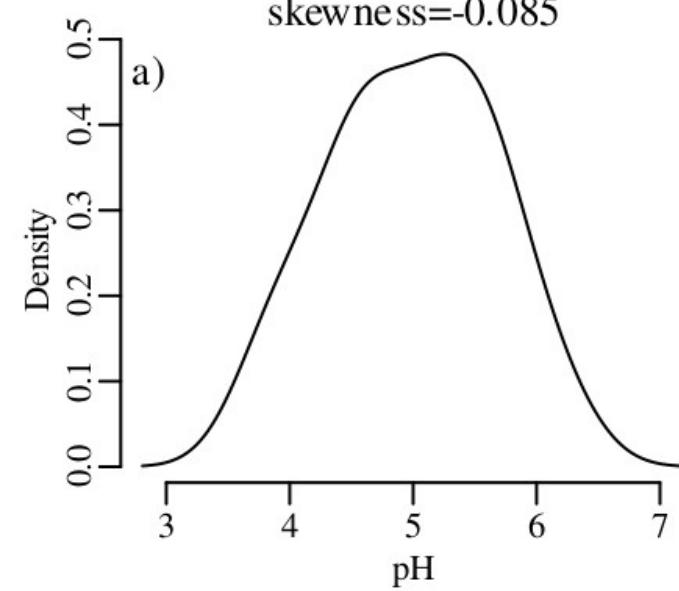
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_i	3.8	3.9	4.1	4.4	4.5	4.6	4.6	4.6	4.8	5.1	5.1	5.1	5.2	5.4	5.5	5.6	5.6	5.7	5.9	6.2
$x_i - \text{med}(x)$	-1.3	-1.2	-1.0	-0.7	-0.6	-0.5	-0.5	-0.5	-0.3	0.0	0.0	0.0	0.1	0.3	0.4	0.5	0.5	0.6	0.8	1.1
$ x_i - \text{med}(x) $	1.3	1.2	1.0	0.7	0.6	0.5	0.5	0.5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

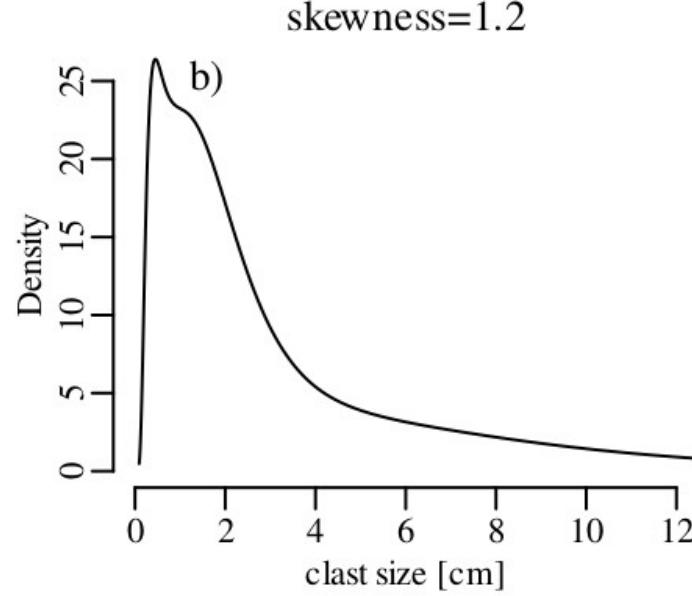
skewness

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

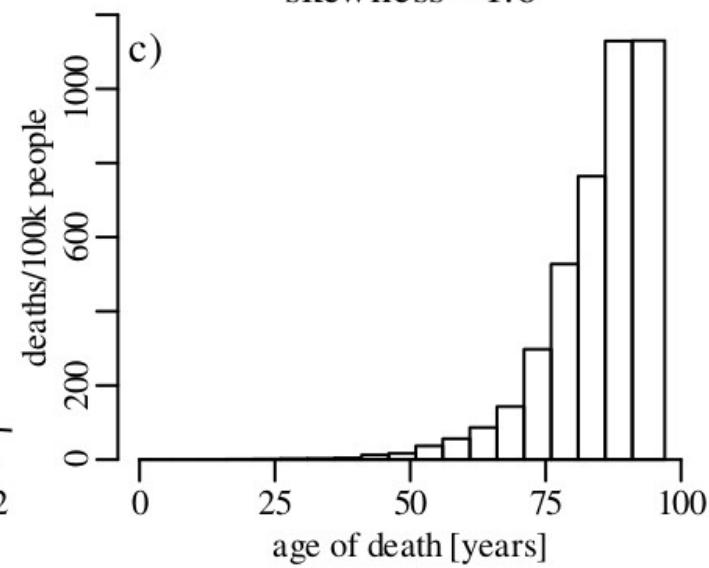
skewness=-0.085

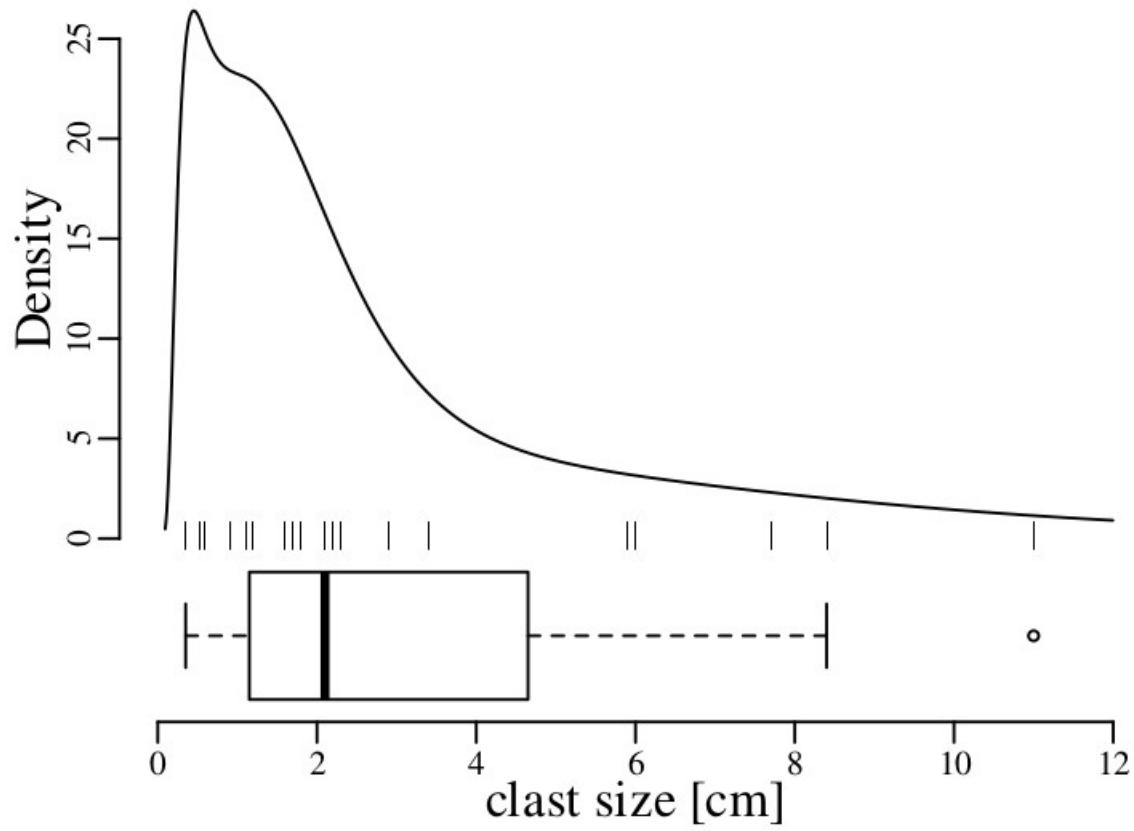


skewness=1.2



skewness=-1.6





Statistics for geoscientists

Probability

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

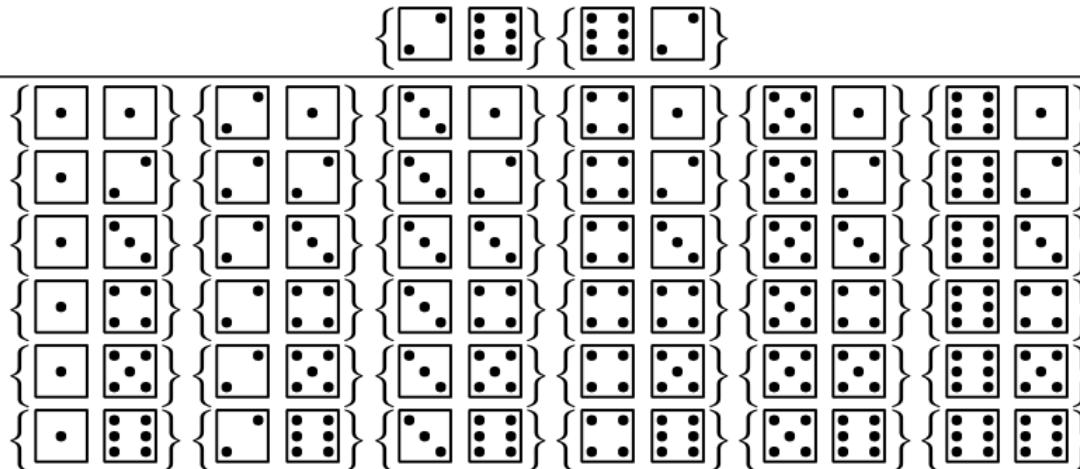
$$P(A) = \frac{\text{the number of ways } A \text{ can occur}}{\text{the total number of outcomes}}$$

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

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

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

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

$$P(1 \times \begin{array}{|c|} \hline \bullet \\ \hline \end{array} \cap 1 \times \begin{array}{|c|c|} \hline \bullet & \bullet \\ \hline \end{array}) = \frac{\left\{ \begin{array}{|c|c|} \hline \bullet & \bullet \\ \hline \bullet & \bullet \\ \hline \end{array} \right\} \left\{ \begin{array}{|c|c|} \hline \bullet & \bullet \\ \hline \bullet & \bullet \\ \hline \end{array} \right\}}{36} = \frac{2}{36} = \frac{1}{18}$$


multiplicative rule

$$P(\{2 \times H \cap 1 \times T\} \cap \{1 \times \begin{array}{|c|c|}\hline \bullet & \bullet \\ \hline \end{array} \cap 1 \times \begin{array}{|c|c|c|}\hline \bullet & \bullet & \bullet \\ \hline \end{array}\}) = \frac{3}{8} \frac{1}{18} = \frac{3}{144} = 0.021$$

additive rule

$$P(\{2 \times H \cap 1 \times T\} \cup \{1 \times \begin{array}{|c|c|}\hline \bullet & \bullet \\ \hline \end{array} \cap 1 \times \begin{array}{|c|c|c|}\hline \bullet & \bullet & \bullet \\ \hline \end{array}\}) = \frac{3}{8} + \frac{1}{18} = \frac{31}{72} = 0.43$$

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

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

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

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

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

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

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?

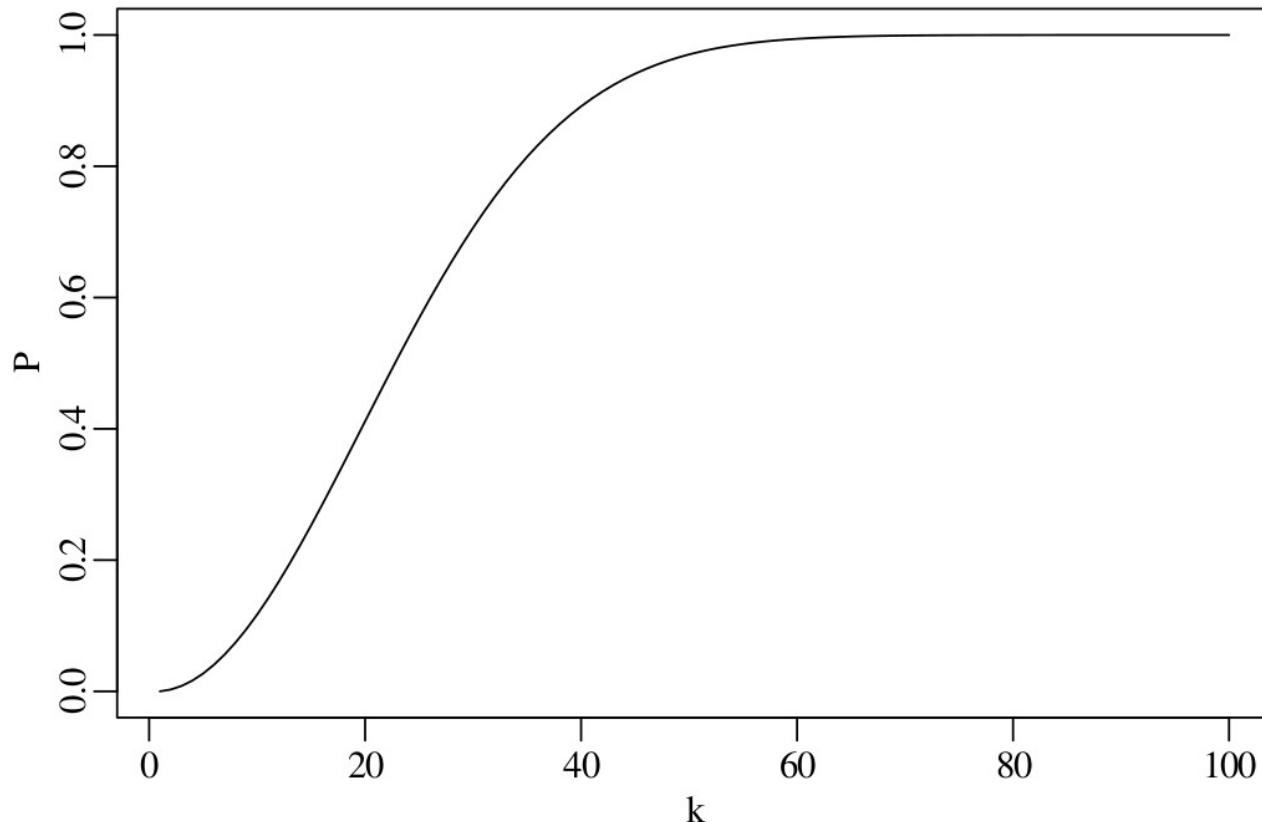
365^k possible combinations of birthdays (sampling with replacement)

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

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

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

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

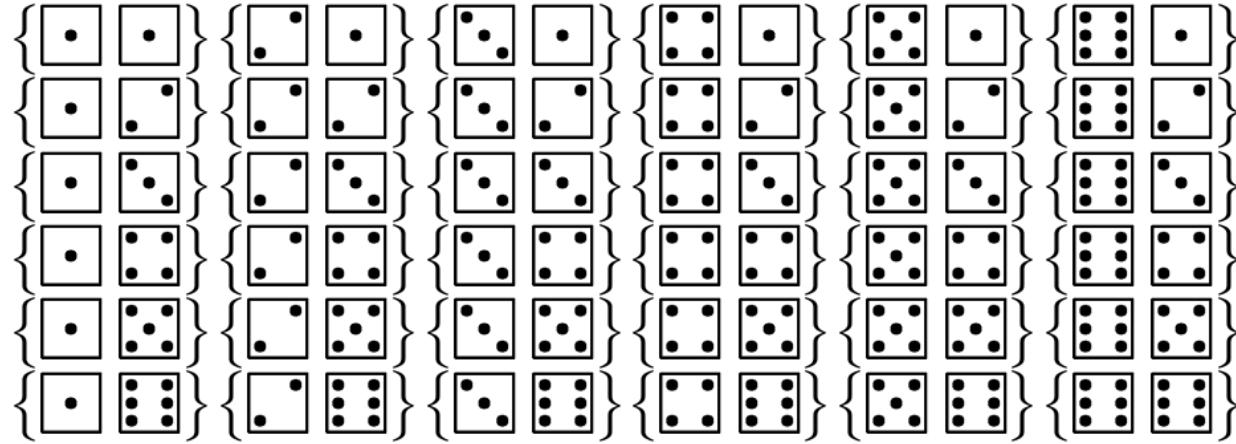


If $k = 23$, then $P(> 1 \text{ overlapping birthdays}) = 0.507$

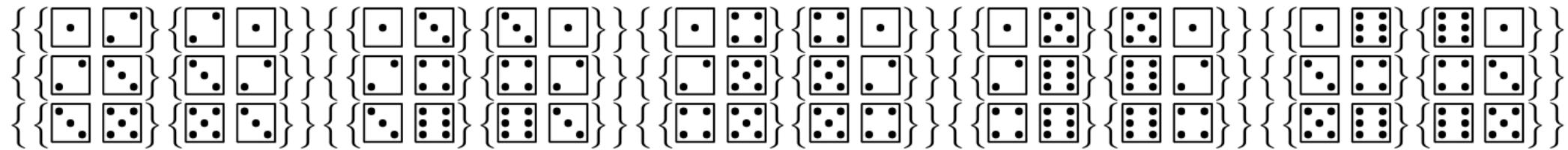
Combinations

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

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



duplicates:



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

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

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

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

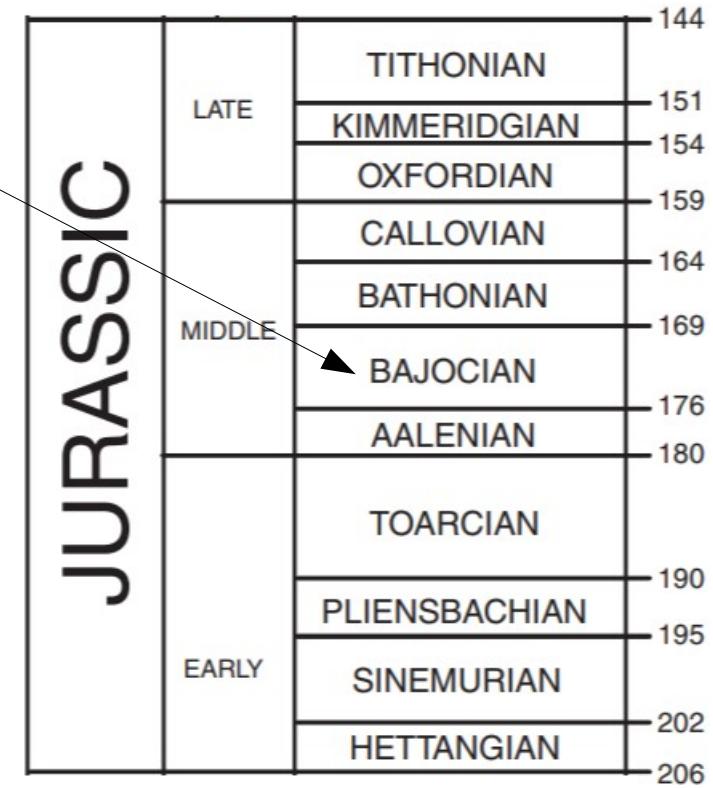
$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

$$\{THH\}\{HTH\}\{HHT\} \hspace{10cm} \left\{\begin{array}{|c|c|}\hline \bullet & \bullet \\ \hline \bullet & \end{array}\right\} \left\{\begin{array}{|c|c|}\hline \bullet & \bullet \\ \hline \bullet & \bullet \\ \hline \end{array}\right\} \left\{\begin{array}{|c|}\hline \bullet \\ \hline \end{array}\right\}$$

$$\binom{3}{2}=\frac{3!}{1!2!}=\frac{6}{2}=3 \hspace{10cm} \binom{2}{1}=\frac{2!}{1!1!}=2$$

Conditional probability

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



multiplication law

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

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

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

law of total probability

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

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

law of total probability

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

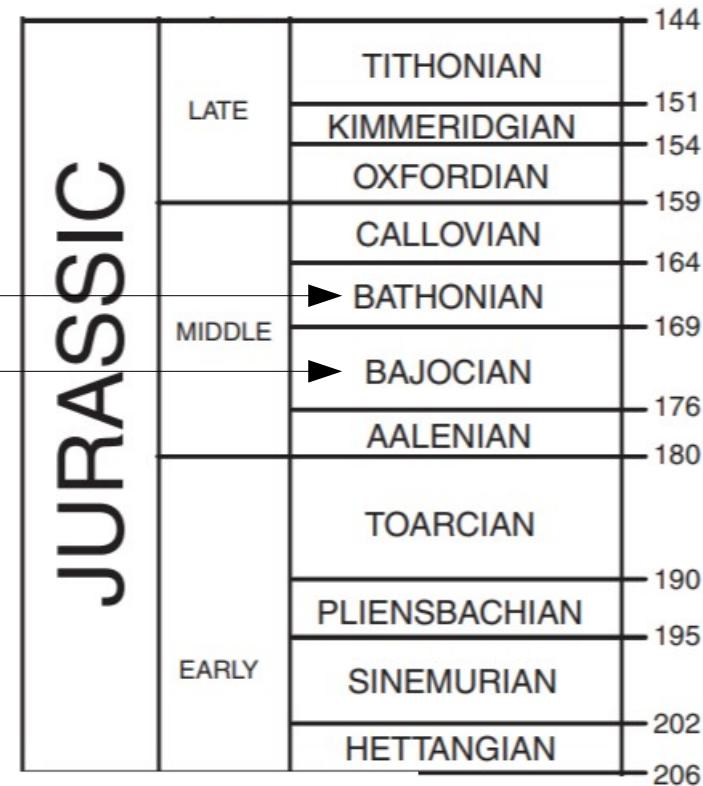
$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

$$P(B_2) = 0.3$$

$$P(B_1) = 0.7$$

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

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



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

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

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

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?

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

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

Statistics for geoscientists

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 $p = 2/3$

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

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

$$P(10000) = (2/3)(1/3)^4 = 0.0082$$

$$P(01000) = (1/3)(2/3)(1/3)^3 = 0.0082$$

$$P(00100) = (1/3)^2(2/3)(1/3)^2 = 0.0082$$

$$P(00010) = (1/3)^3(2/3)(1/3) = 0.0082$$

$$P(00001) = (1/3)^4(2/3) = 0.0082$$

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

$$\text{five gold claims} \quad p = 2/3$$

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

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

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

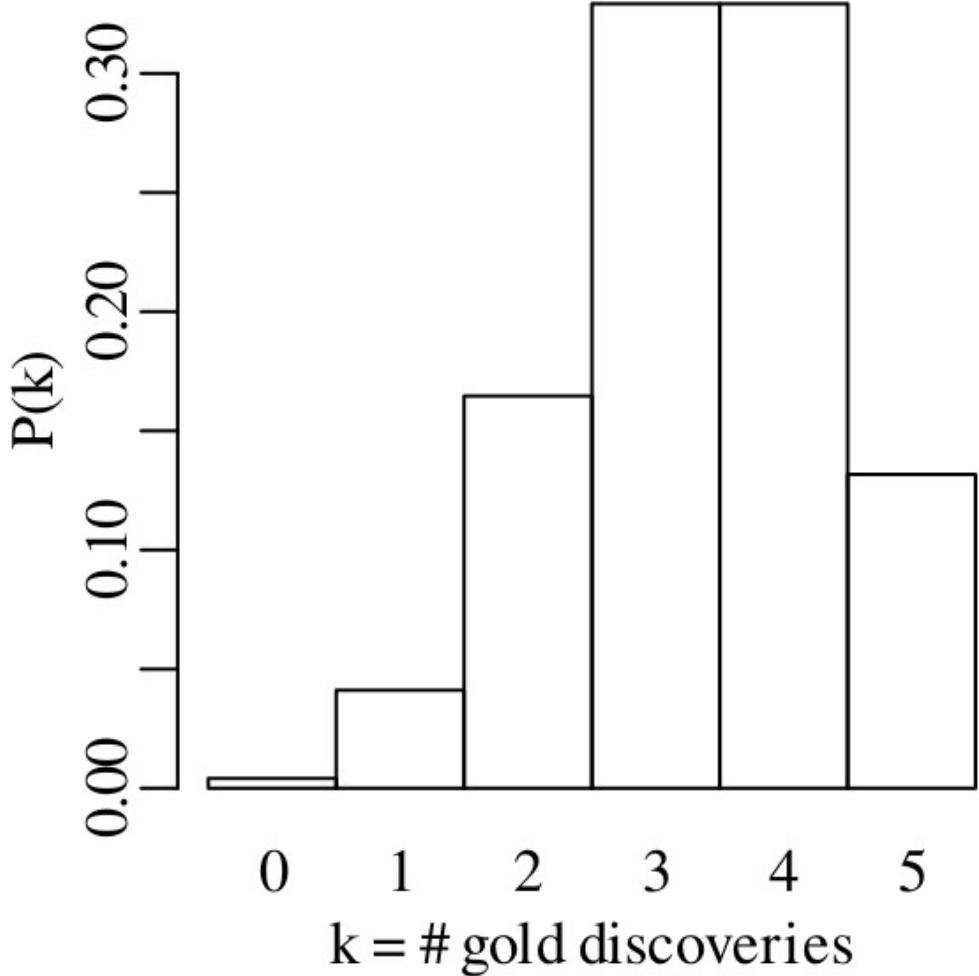
$$P(3 \times \text{gold}) = \binom{5}{3} (2/3)^3 (1/3)^2 = 10 \times 0.033 = 0.33$$

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

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

probability mass function (PMF)

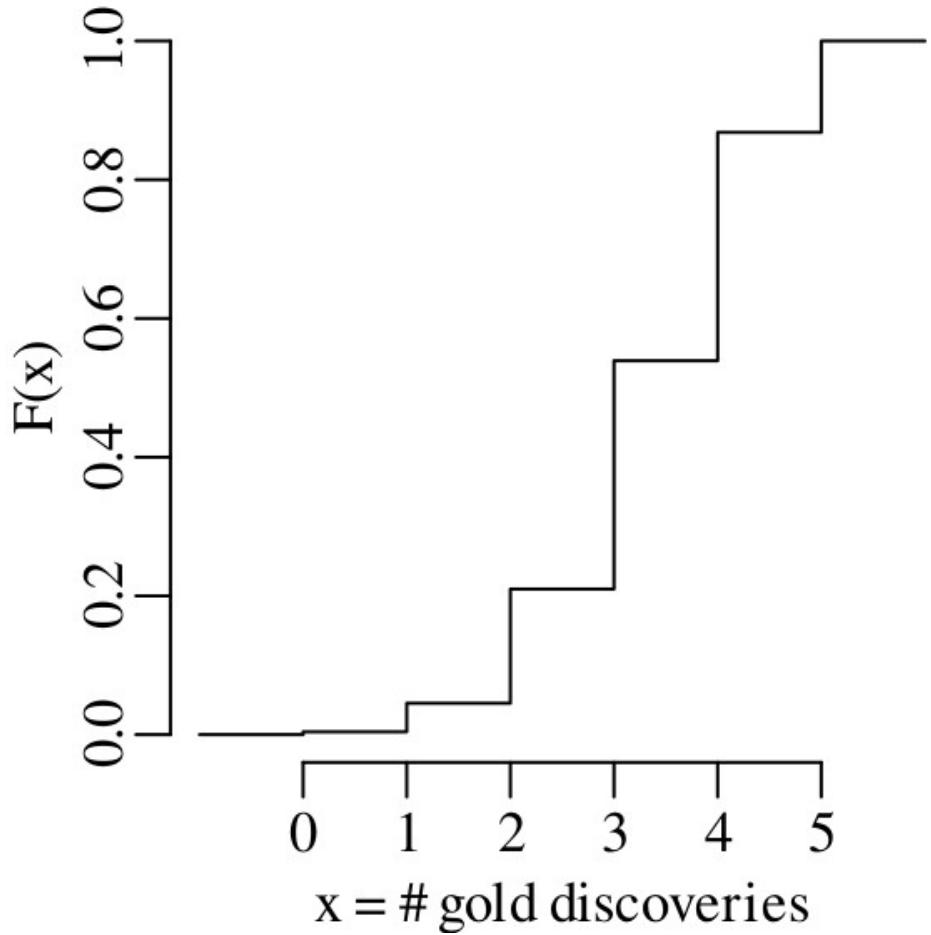
$$P(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$



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

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

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

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

Hypothesis tests

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

1. Formulate two hypotheses:

$$H_0 \text{ (null hypothesis)} \qquad \qquad p = 2/3$$

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

2. Calculate the **test statistic** T (k)
3. Determine the **null distribution** of T under H_0 .

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** ($\alpha = 0.05$)

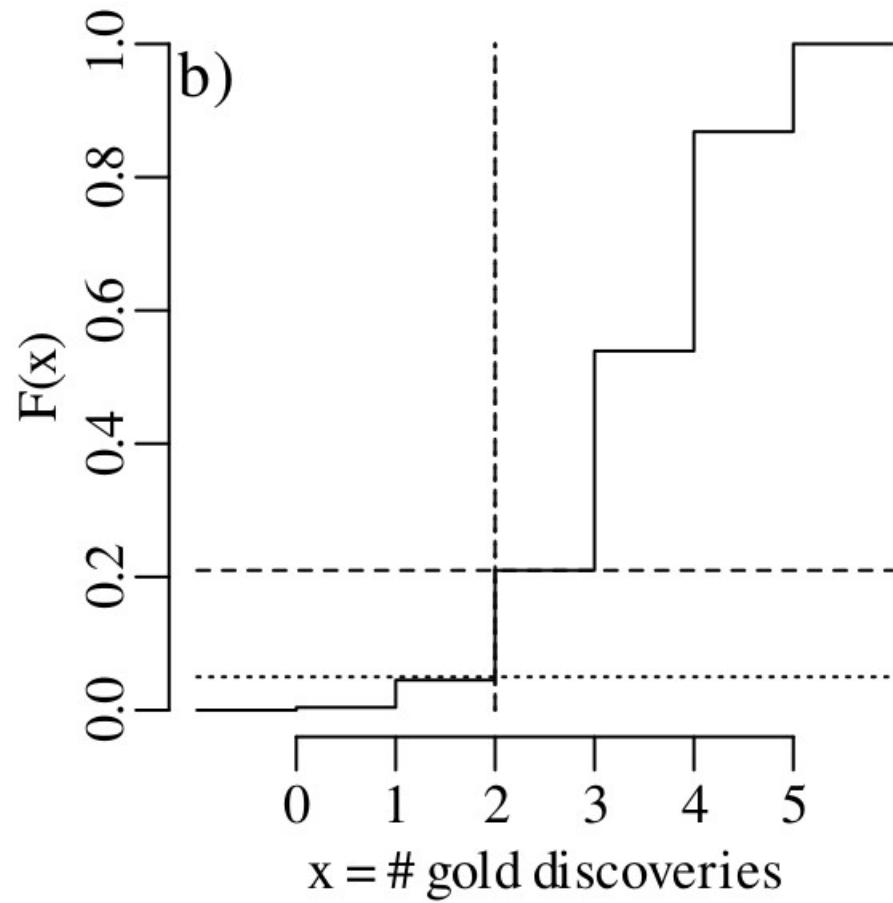
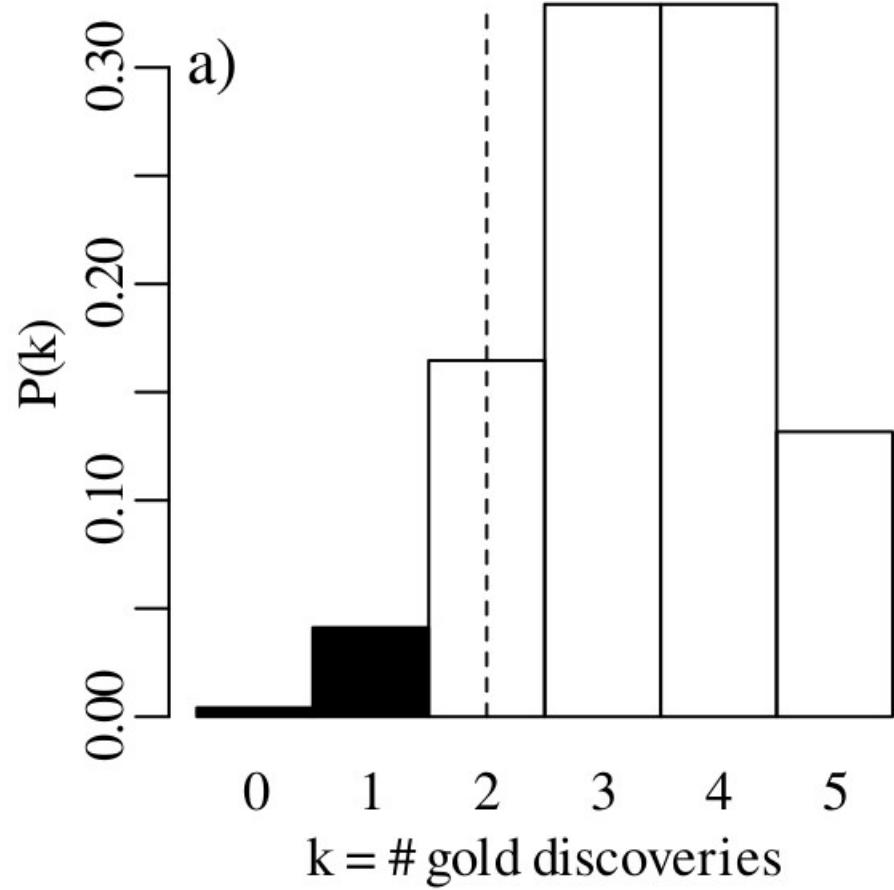
5. Mark the **rejection region**

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	<i>0.2099</i>	0.5391	0.8683	1.0000

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

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

7. Alternatively, $0.2099 > \alpha$



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

one-sided hypothesis test

$$H_0 \text{ (null hypothesis)} \qquad \qquad \qquad p = 2/3$$

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

two-sided hypothesis test

$$H_0 \text{ (null hypothesis)} \qquad \qquad \qquad p = 2/3$$

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

2. Calculate the **test statistic** $T^-(k)$
3. Determine the **null distribution** of T under H_0 .

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 $\alpha/2$

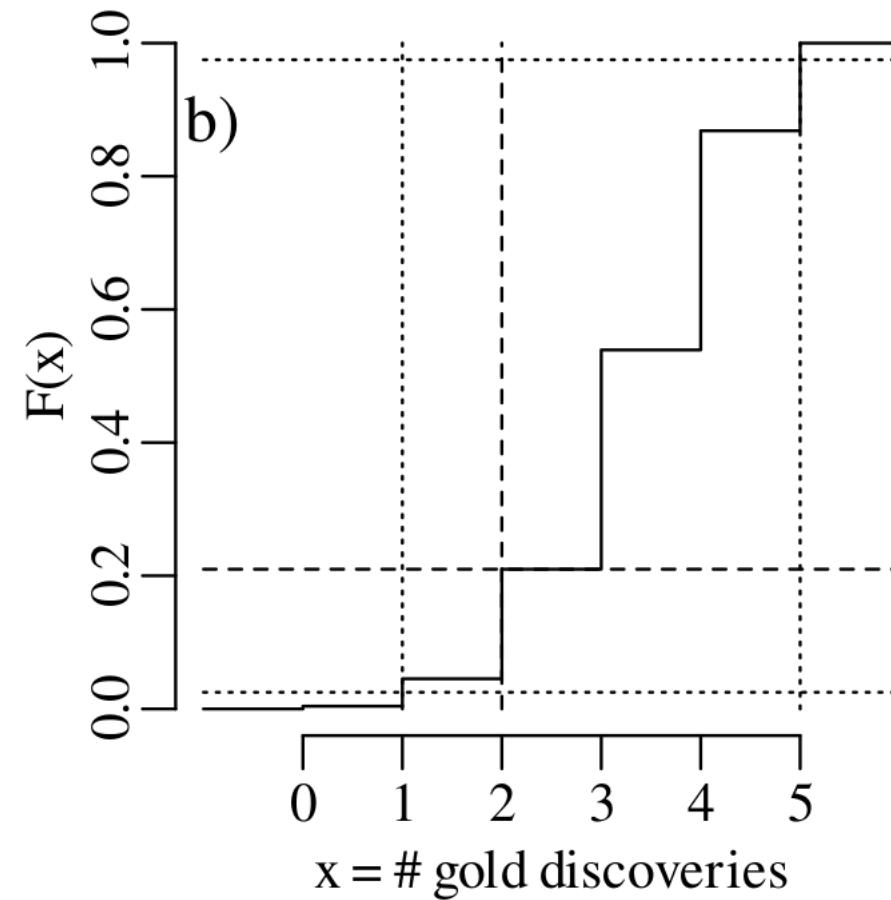
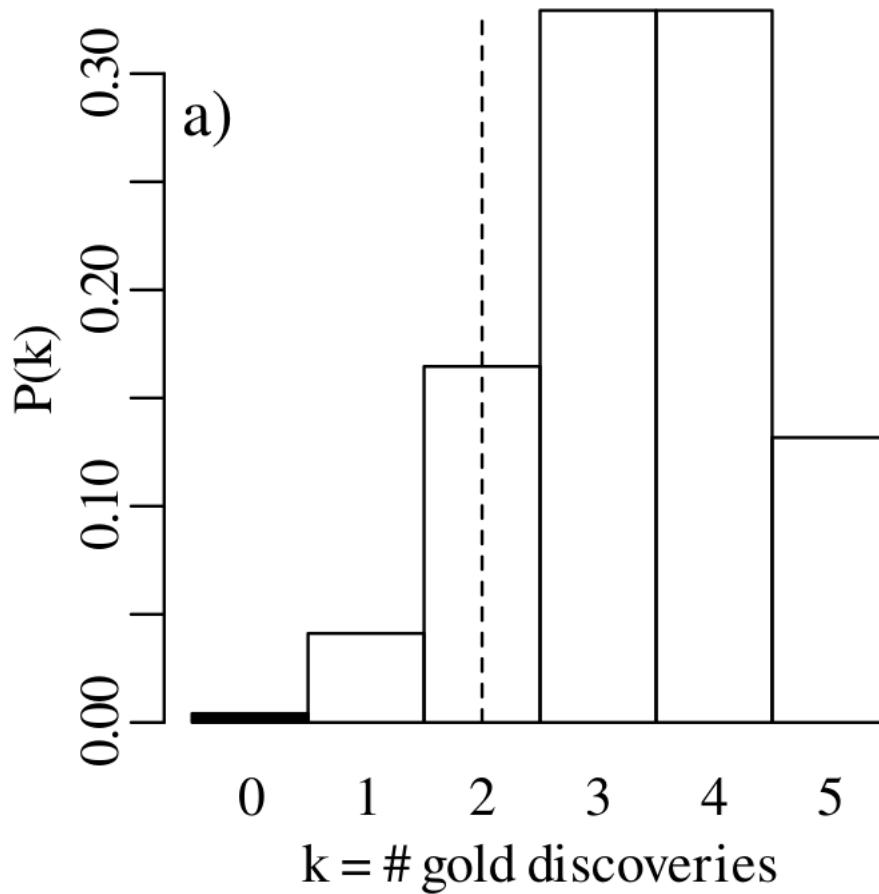
5. Mark the **rejection region**

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	<i>0.2099</i>	0.5391	0.8683	1.0000
$P(T \geq k)$	1.000	0.9959	<i>0.9547</i>	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_0 for $\hat{p} = \frac{2}{5} = 0.4$. How about $\hat{p} = \frac{6}{15} = 0.4$?

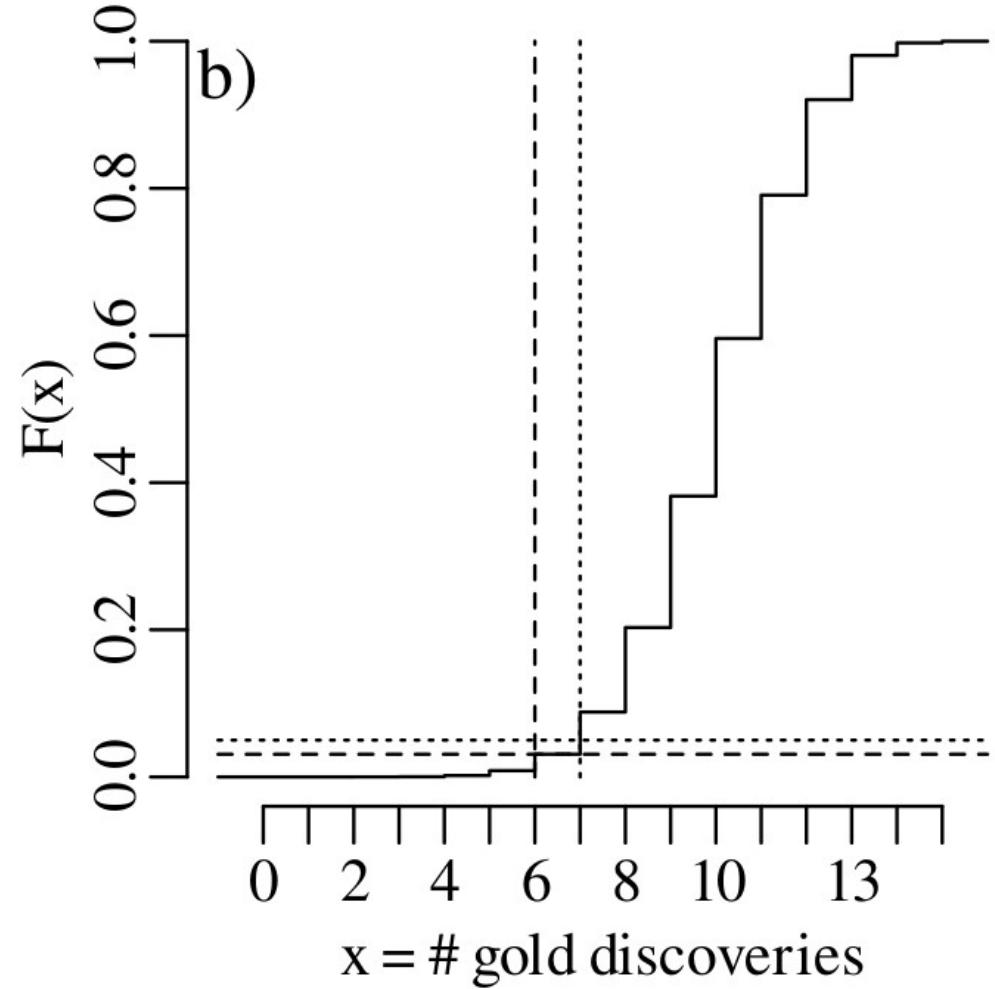
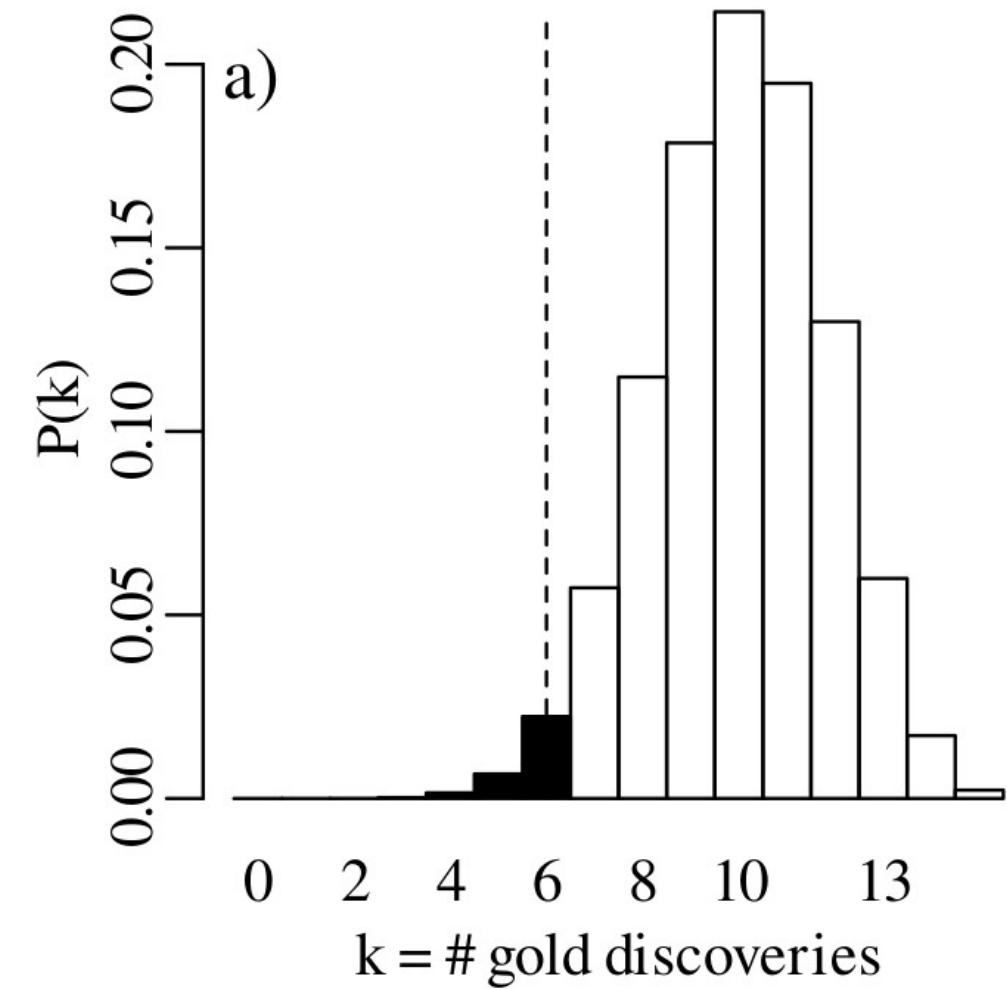
one-sided hypothesis test ($H_0 : p = 2/3$ vs. $H_a : p < 2/3$)

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
$P(T \leq k)$	7.0×10^{-8}	2.2×10^{-6}	3.1×10^{-5}	2.8×10^{-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

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

$$k \in R$$

$$\text{p-value} = 0.0308 < \alpha$$



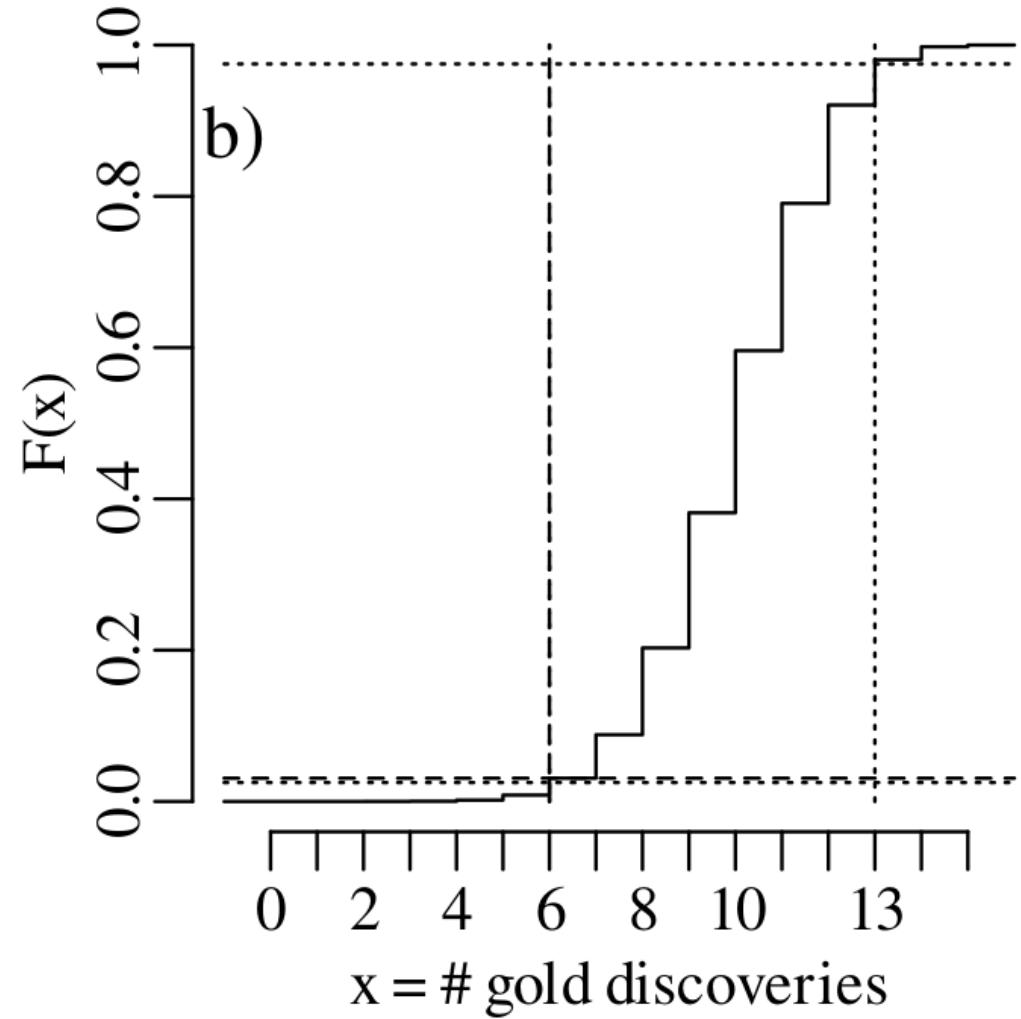
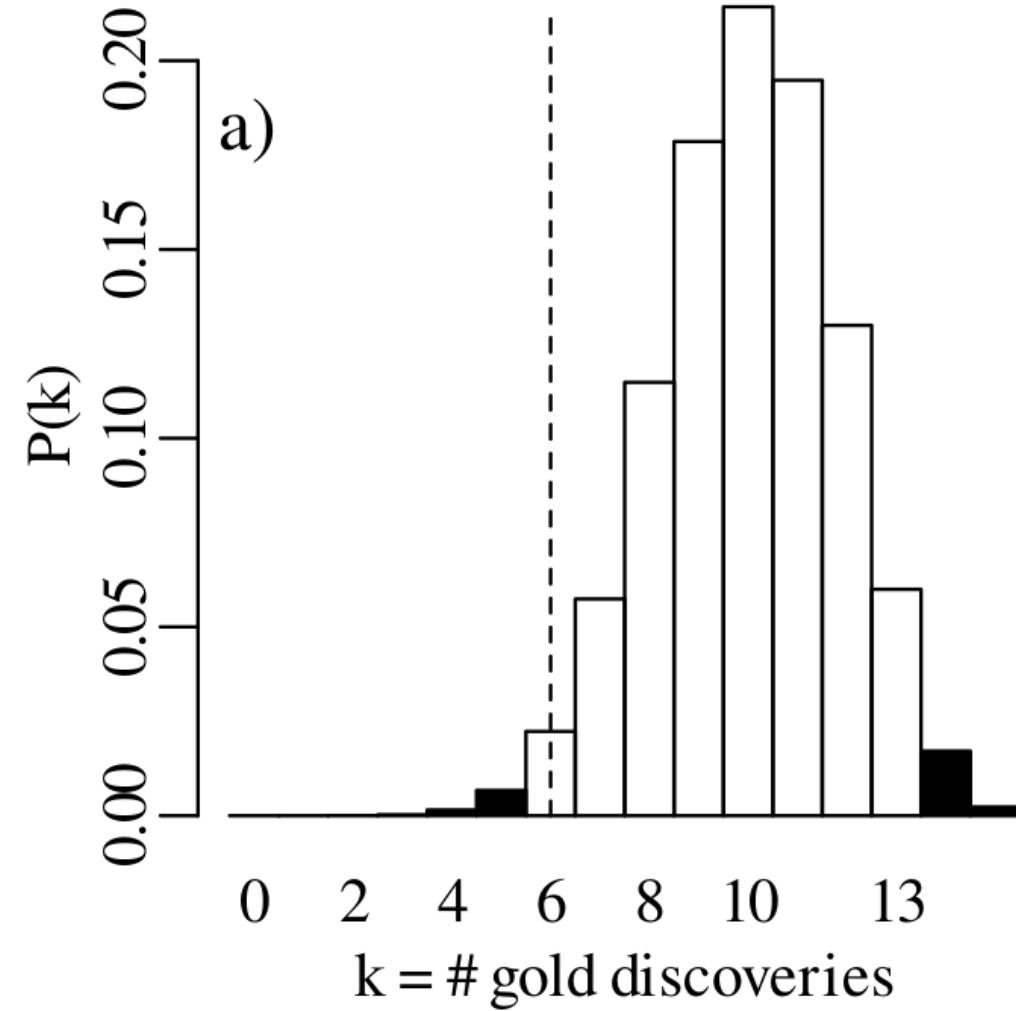
two-sided hypothesis test ($H_0 : p = 2/3$ vs. $H_a : p \neq 2/3$):

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
$P(T \leq k)$	7.0×10^{-8}	2.2×10^{-6}	3.1×10^{-5}	2.8×10^{-4}	0.0018	0.0085	0.0308	0.0882
$P(T \geq k)$	1.0000	$1 - 7.0 \times 10^{-8}$	$1 - 2.2 \times 10^{-6}$	$1 - 3.1 \times 10^{-5}$	$1 - 2.8 \times 10^{-4}$	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

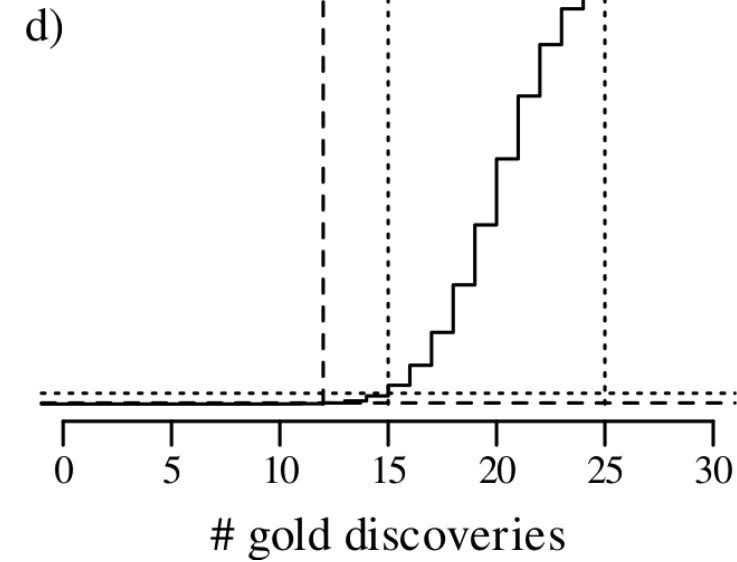
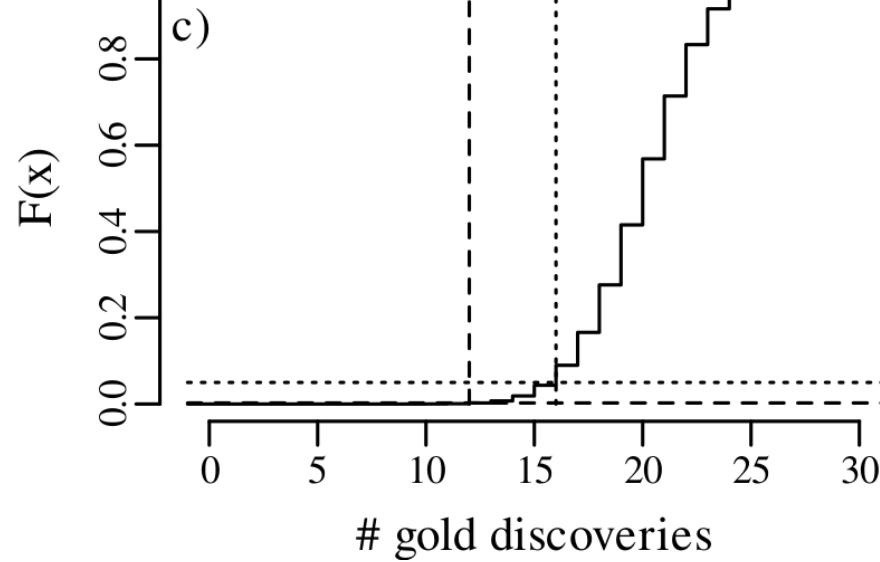
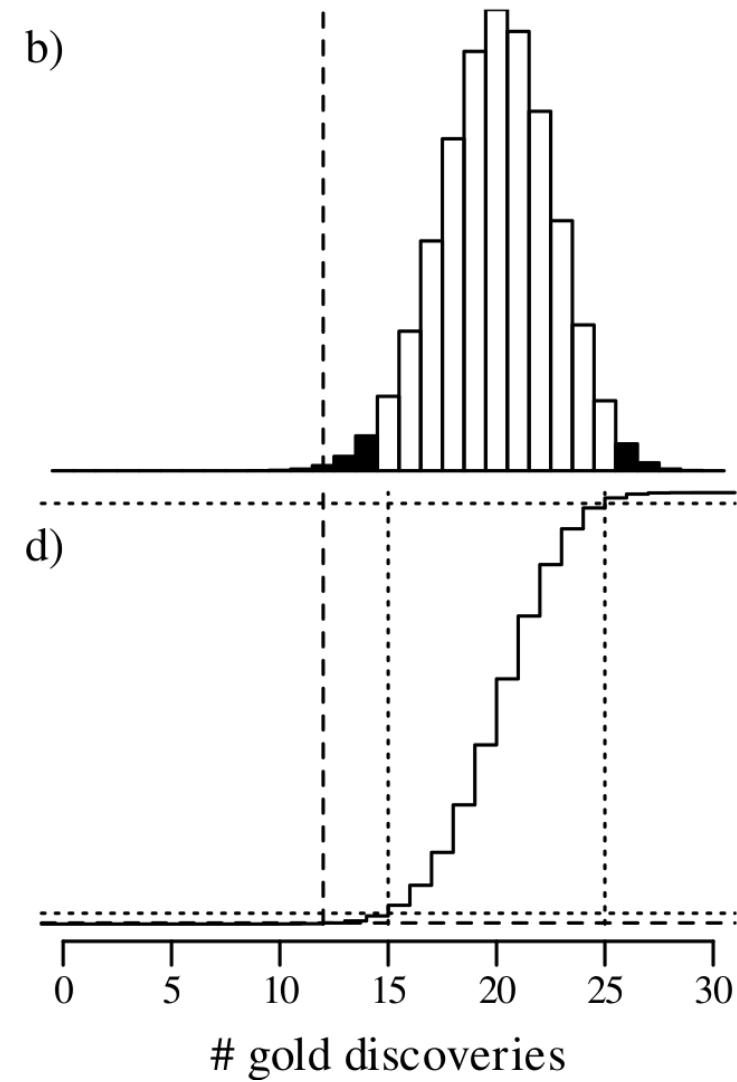
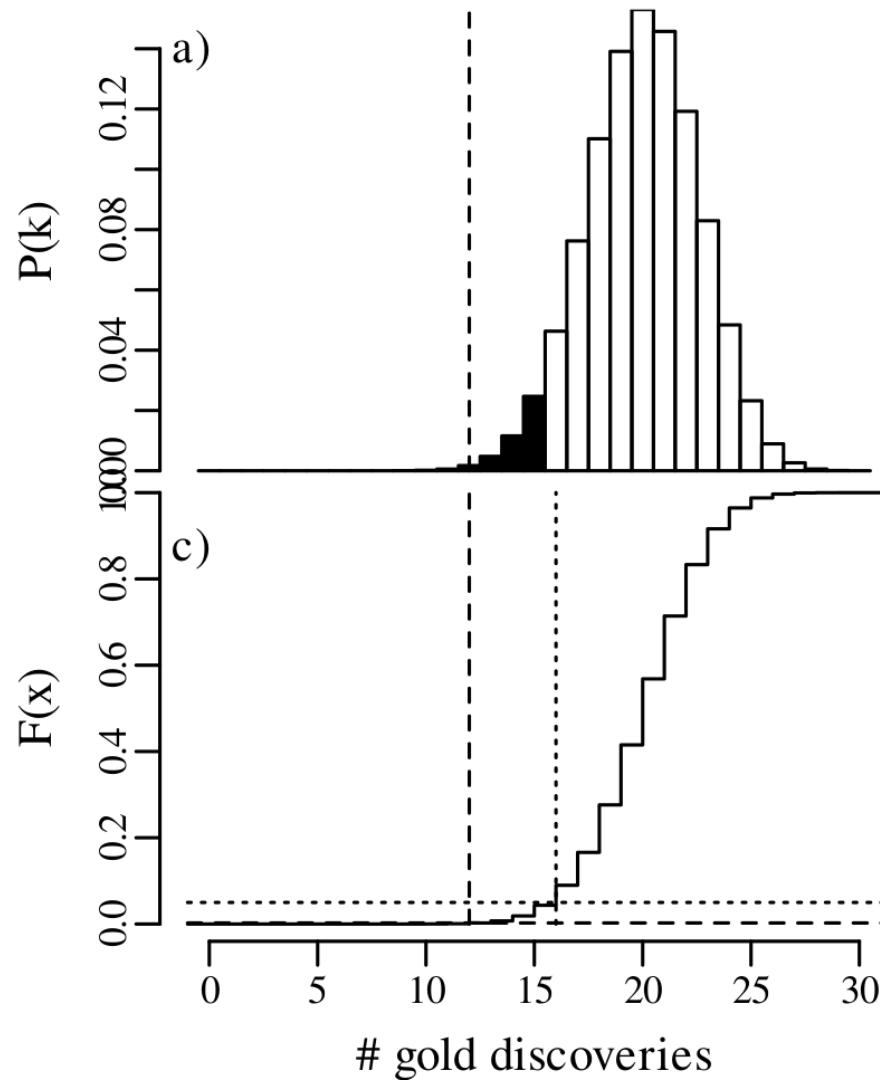
$$R = \{0, 1, 2, 3, 4, 5, 14, 15\}$$

$$k \notin R$$

$$\text{p-value} = (2 \times 0.0308 =) 0.0616 > \alpha$$



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



H_0 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

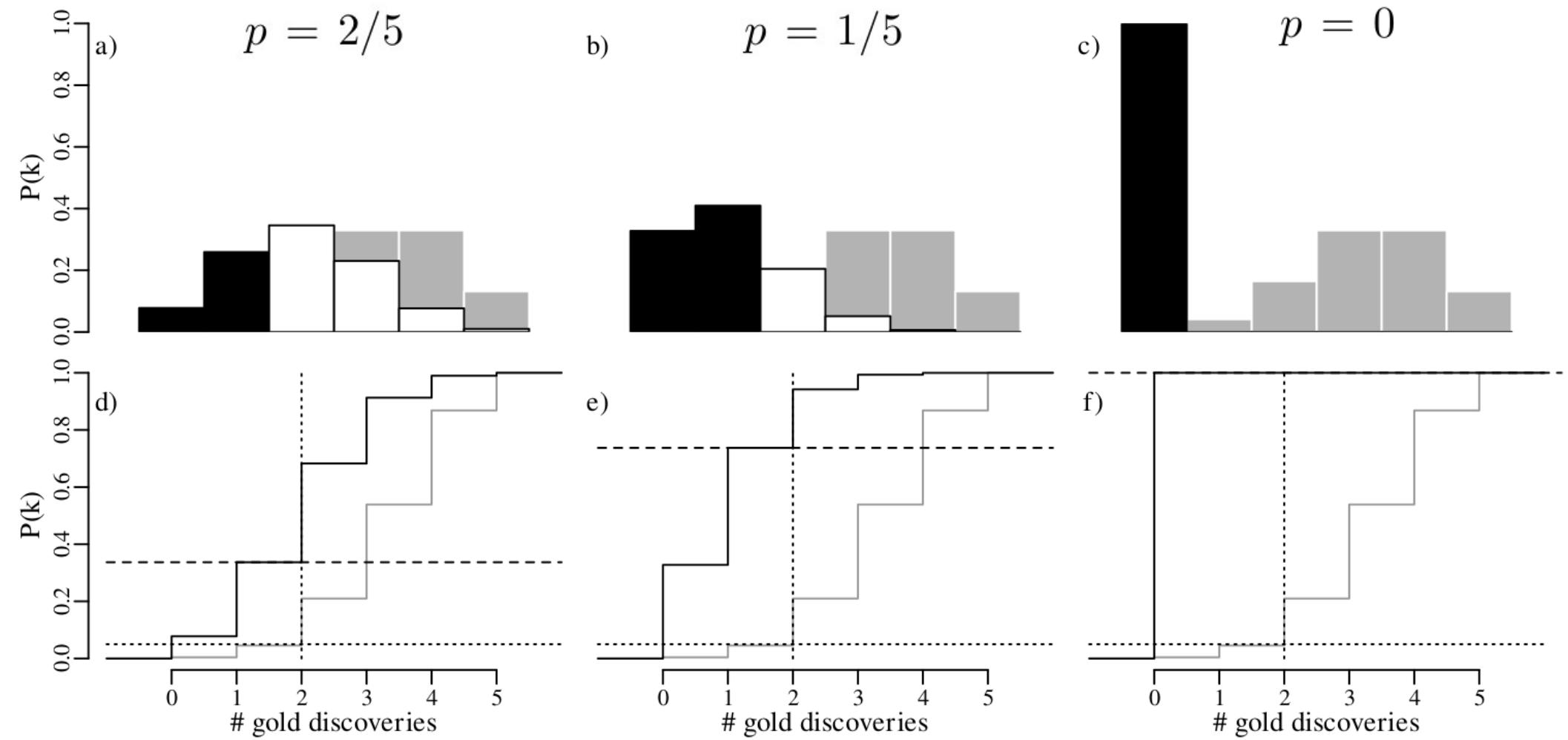
depends on the **confidence level** α

H_0 is ...	false	true
rejected	correct decision	Type-I error
not rejected	Type-II error	correct decision

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

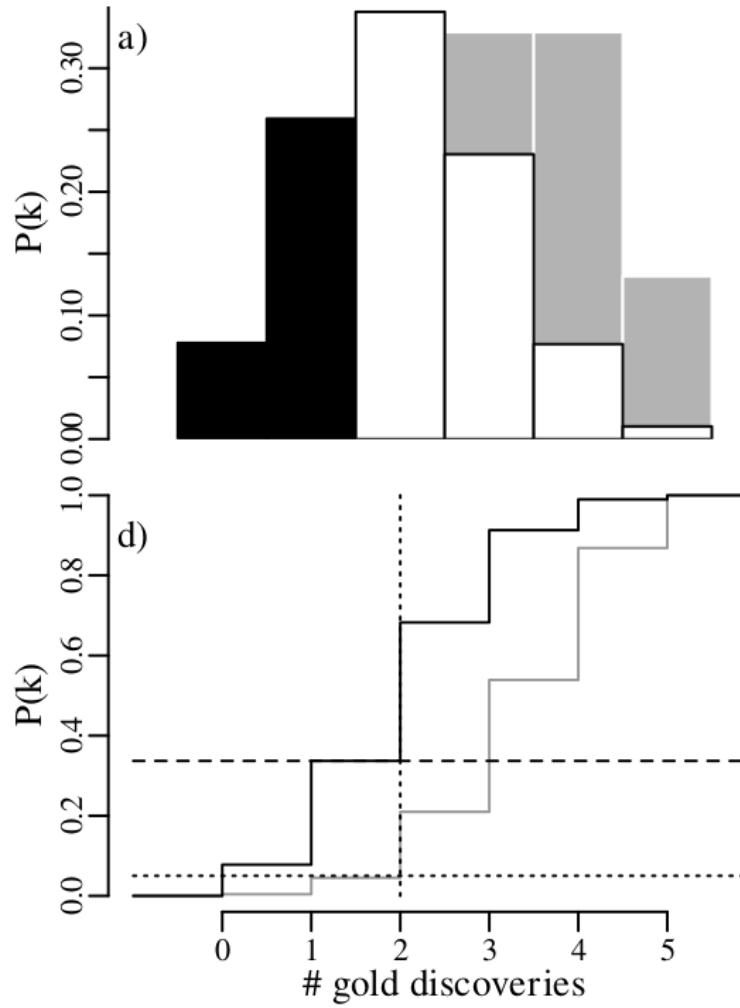
depends on 1. The **degree to which H_0 is false**
2. **Sample size**

1. The degree to which H_0 is false

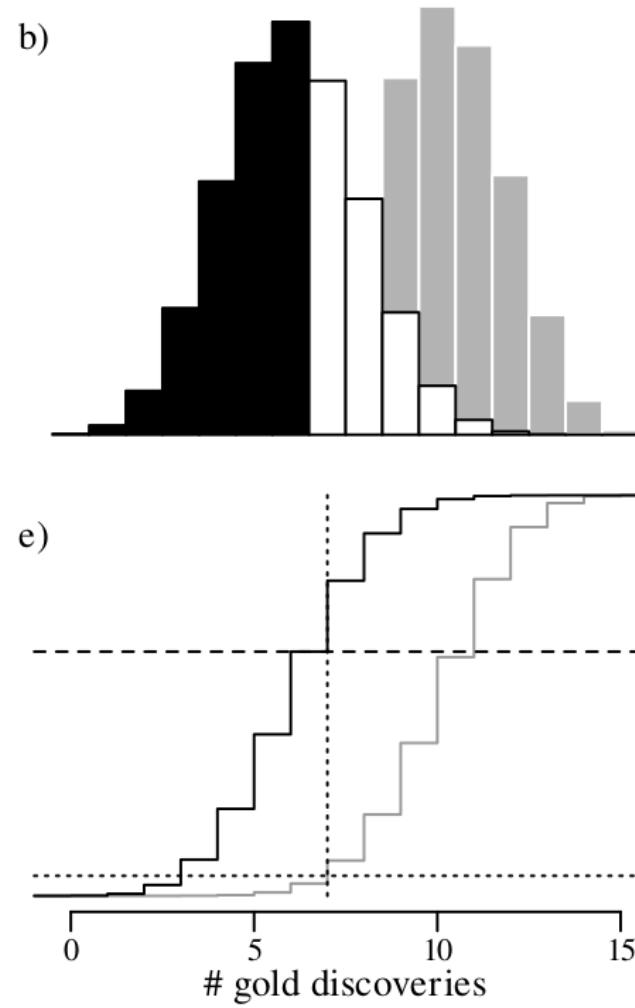


2. Sample size

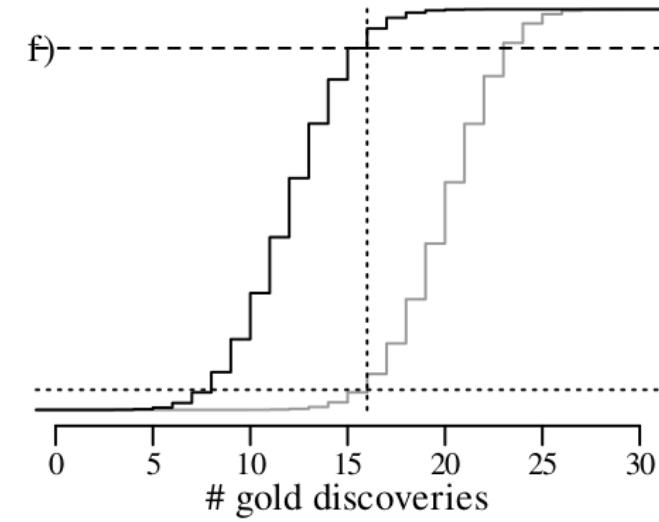
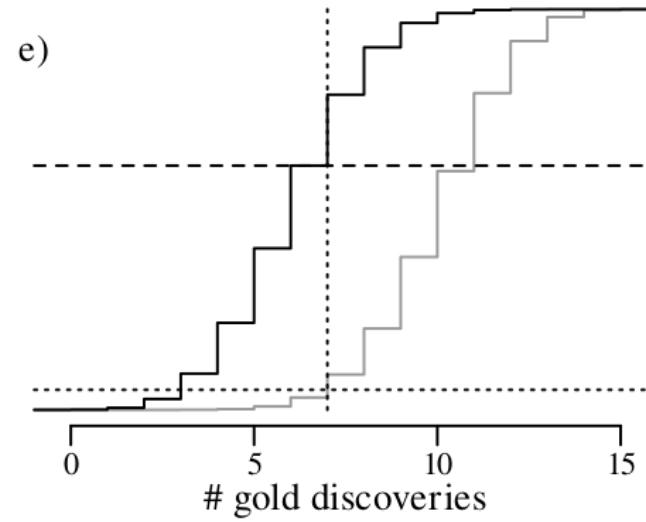
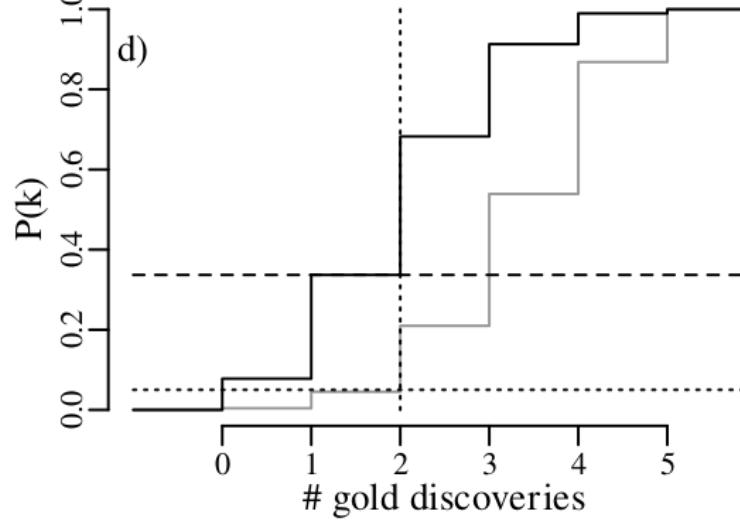
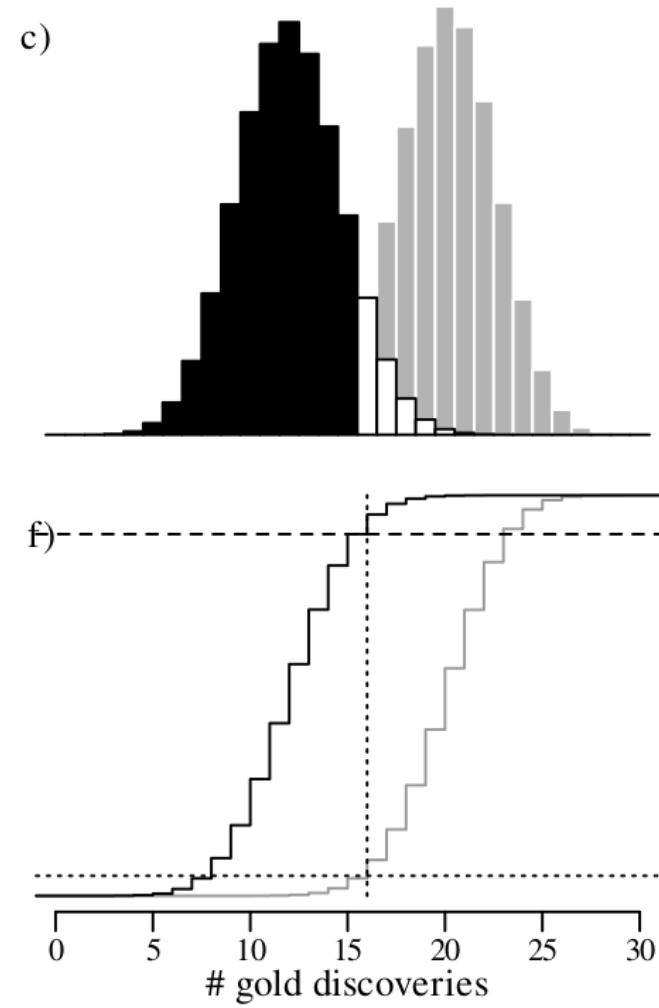
$n = 5$



$n = 15$



$n = 30$



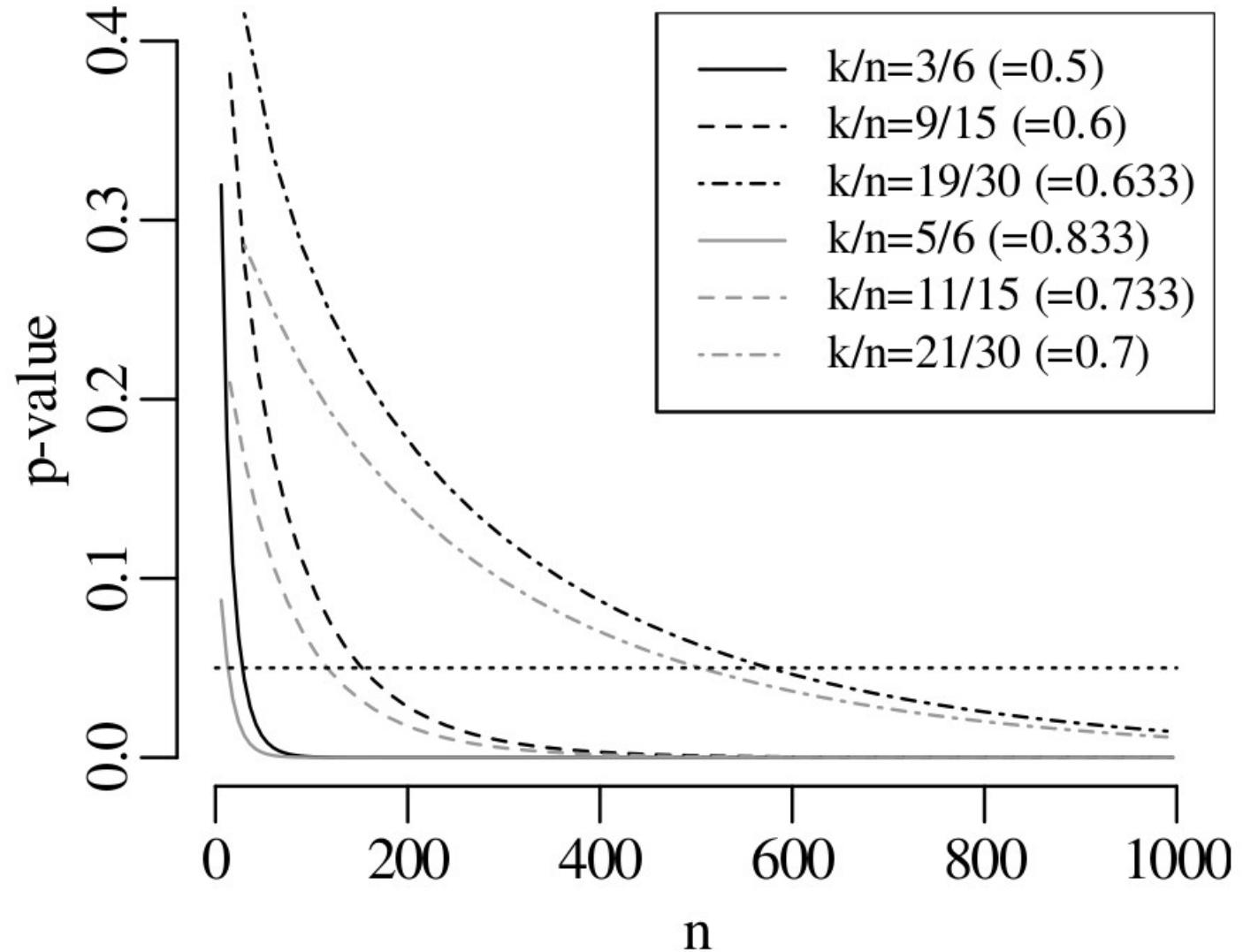
The scientific method

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.

$p = 2/3$?



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. $p=2/3?$

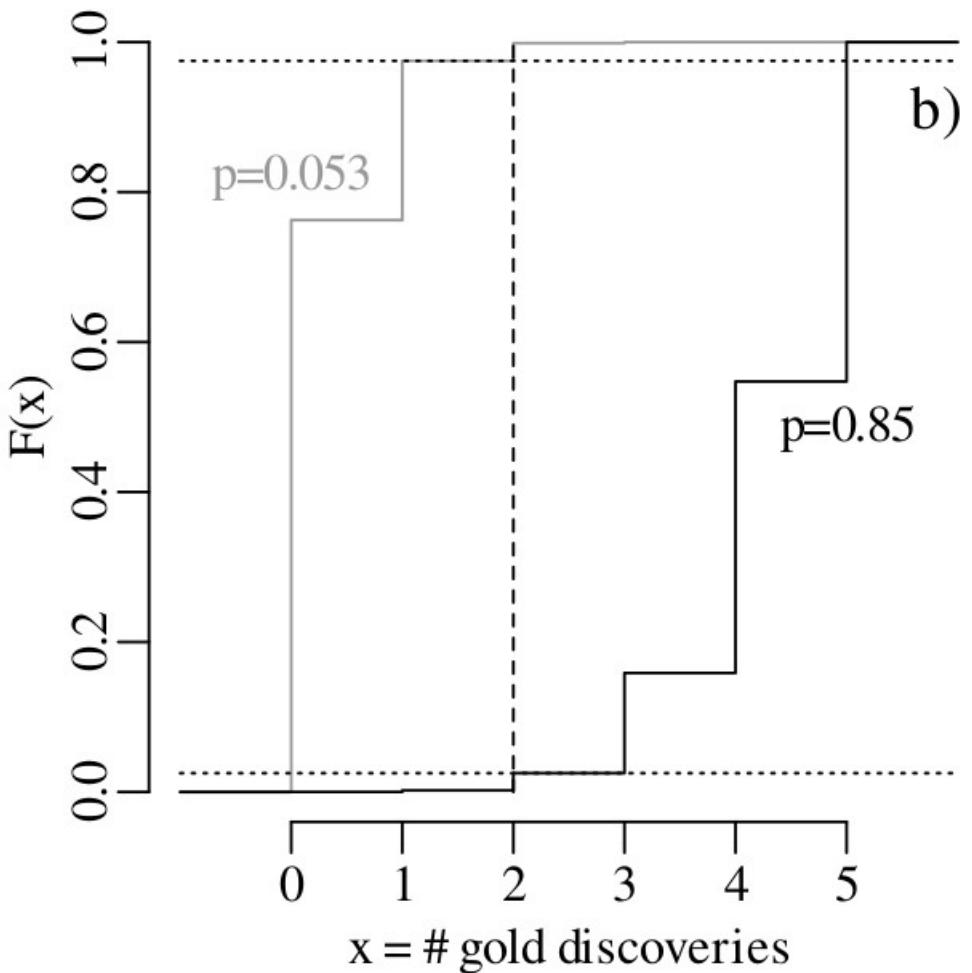
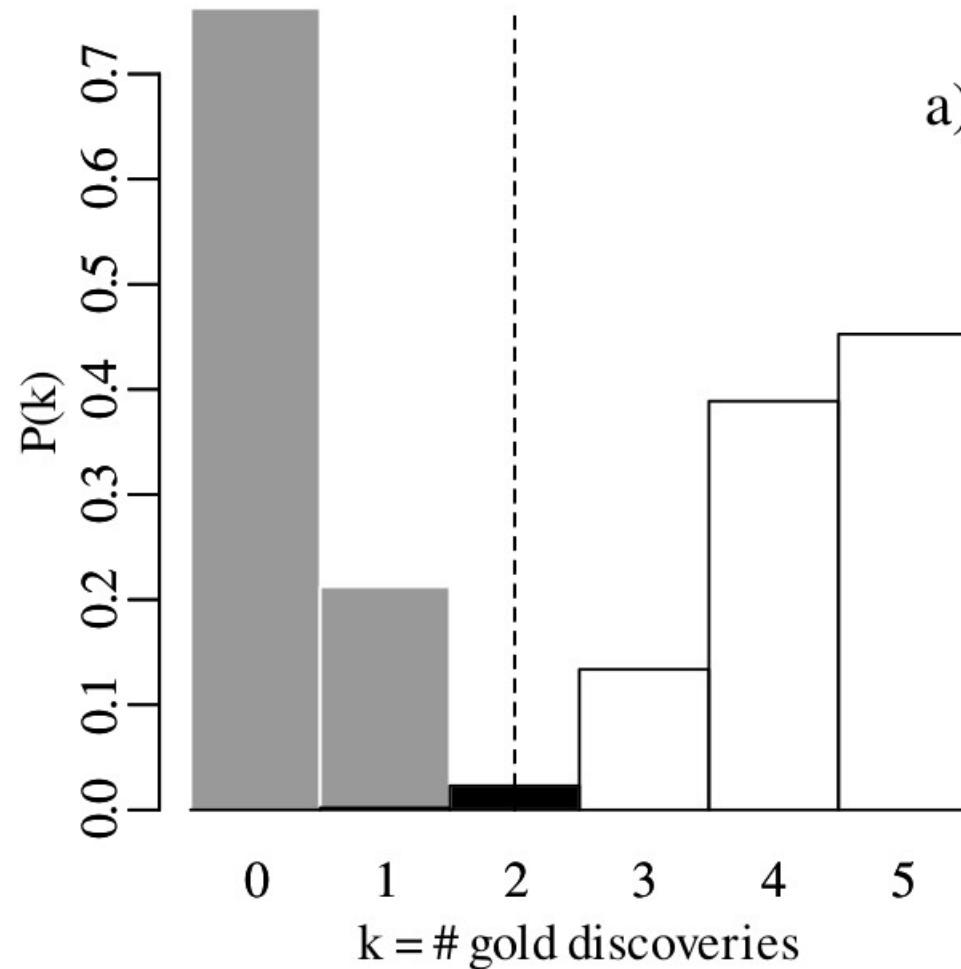
$$P(k \leq 2 | p = 2/3, n = 5) = \sum_{i=0}^2 \binom{5}{i} (2/3)^i (1/3)^{n-i} = 0.21 \geq \alpha/2$$

5. ~~$p=0.9?$~~

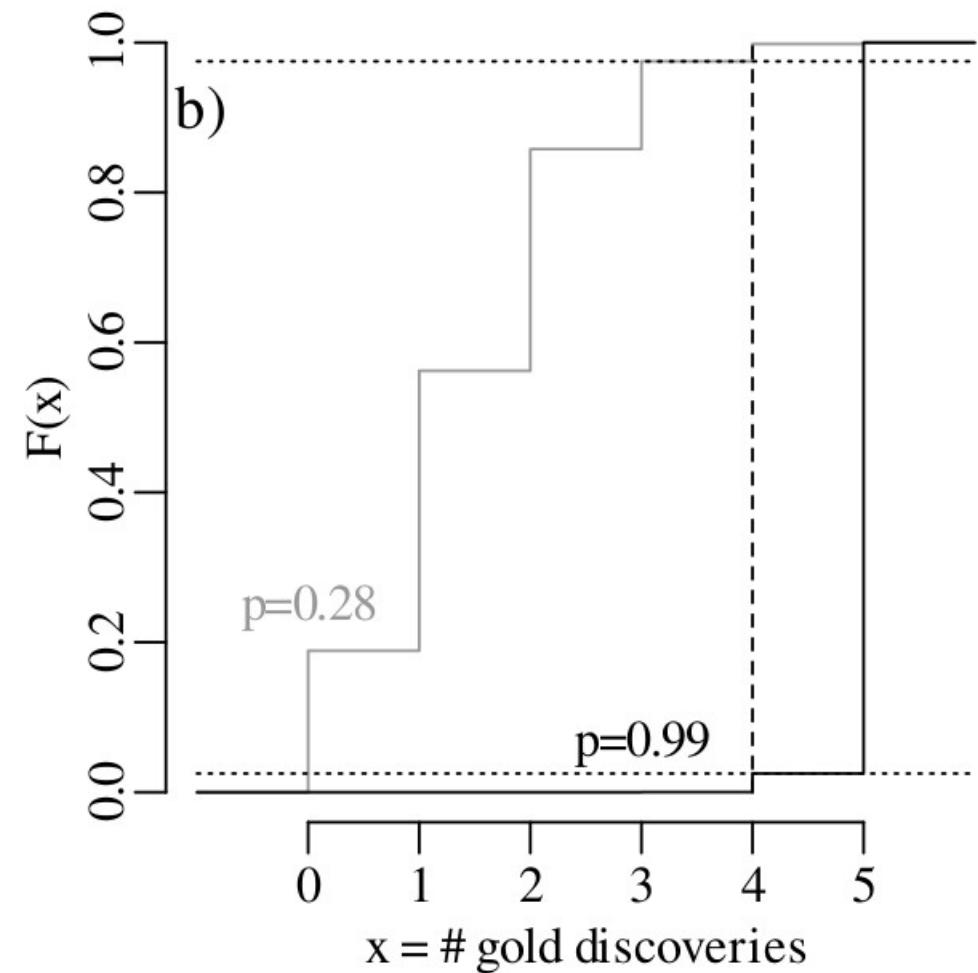
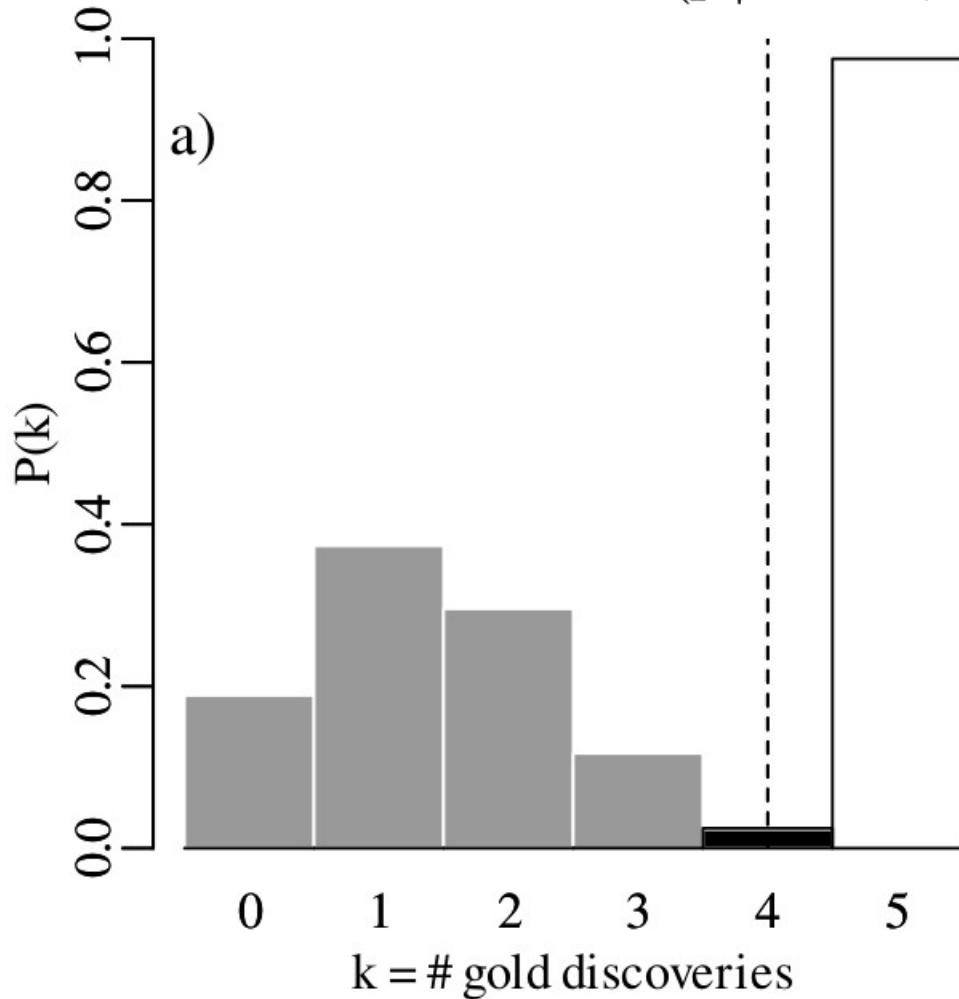
$$P(k \leq 2 | p = 0.9, n = 5) = \sum_{i=0}^2 \binom{5}{i} (0.9)^i (0.1)^{n-i} = 0.0086 \leq \alpha/2$$

6. ~~$p=1?$~~

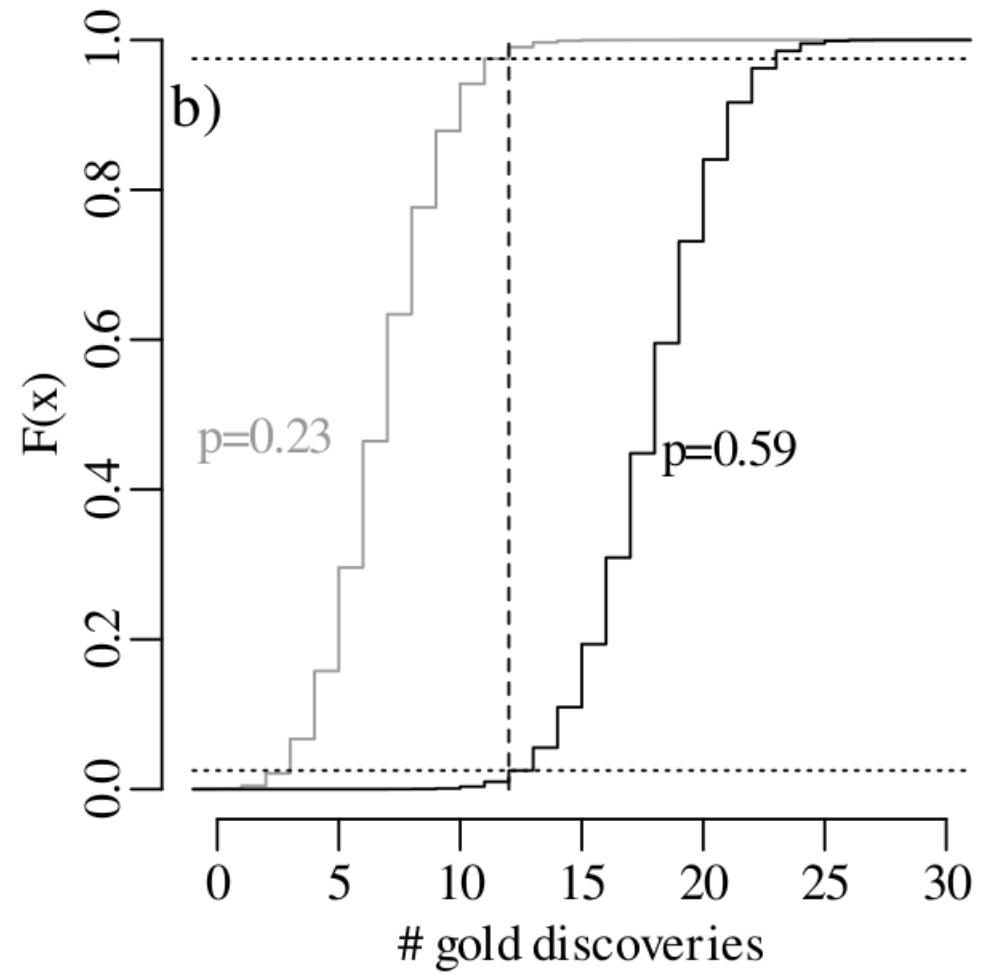
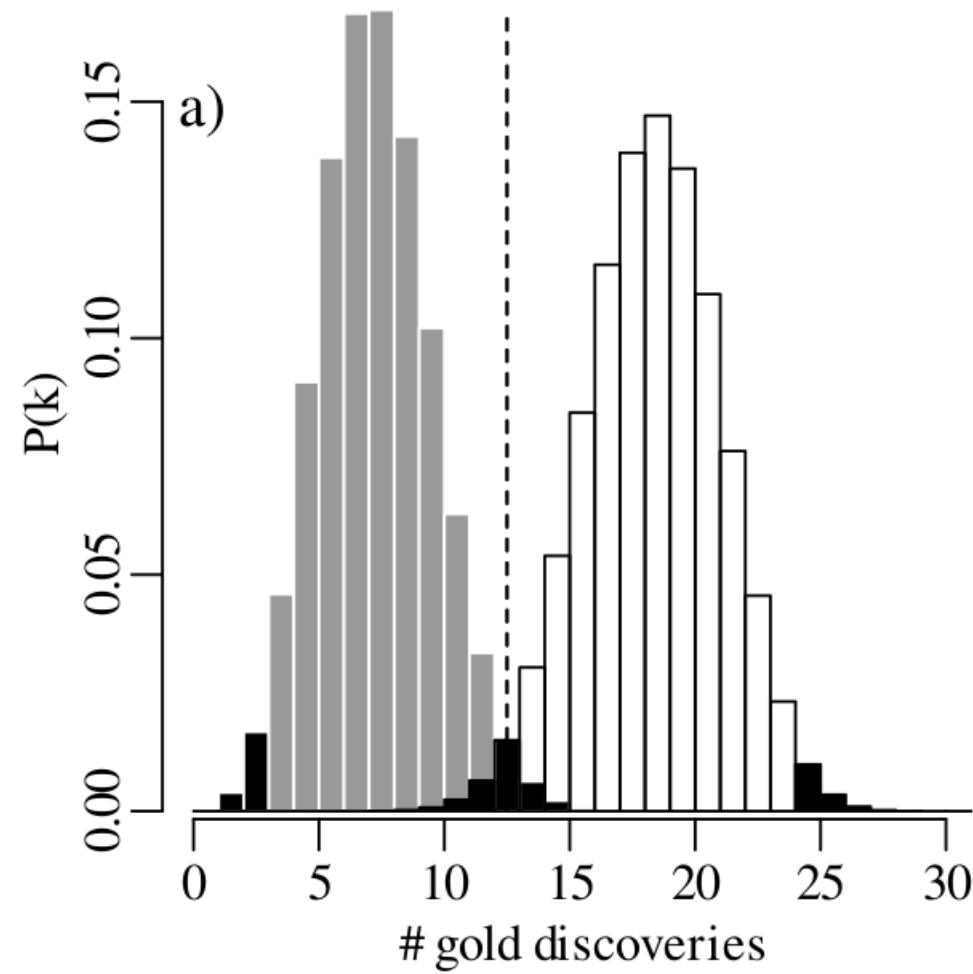
$$\text{C.I.}(p|k=2, n=5) = [0.053, 0.85]$$

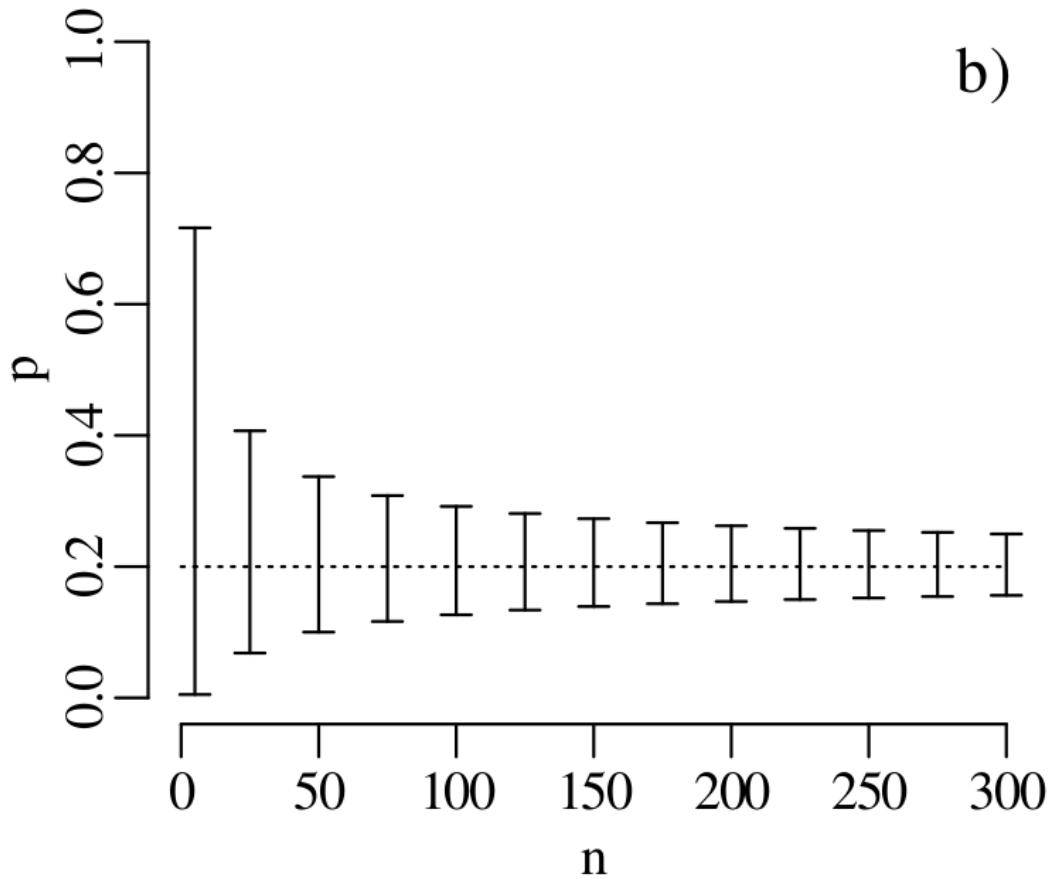
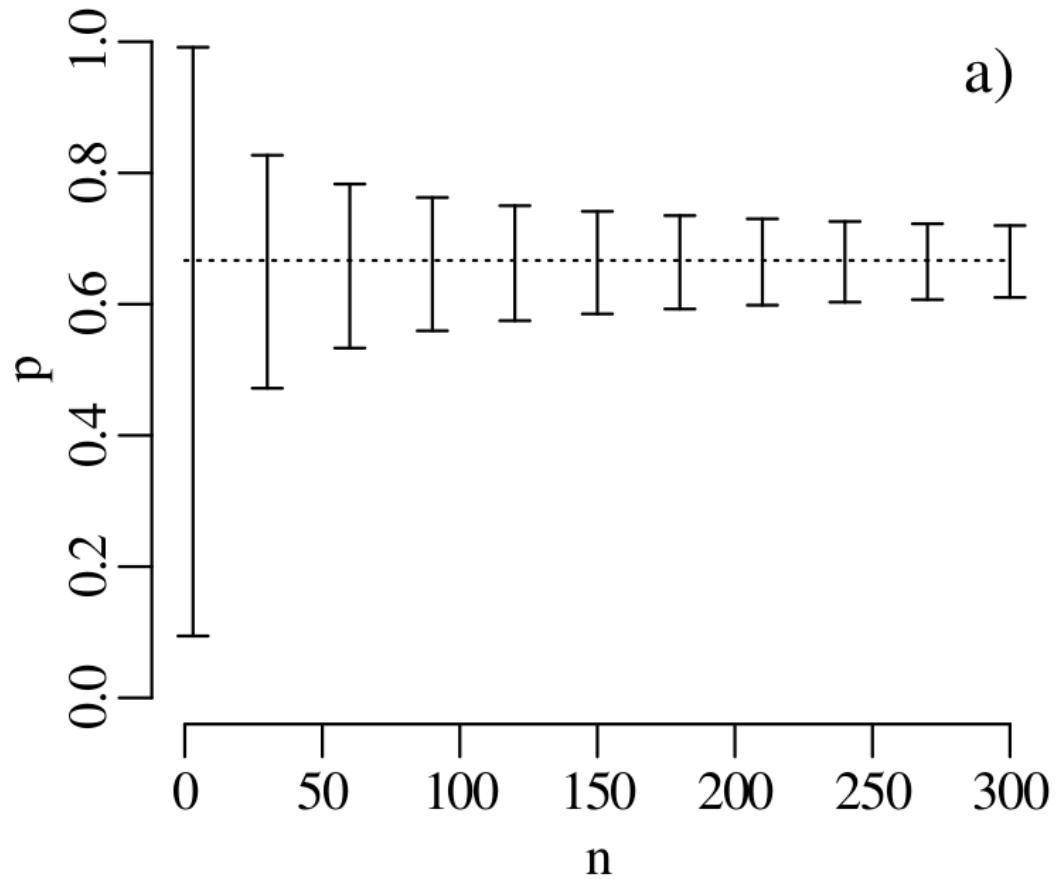


$$\text{C.I.}(p|k=4, n=5) = [0.284, 0.995]$$



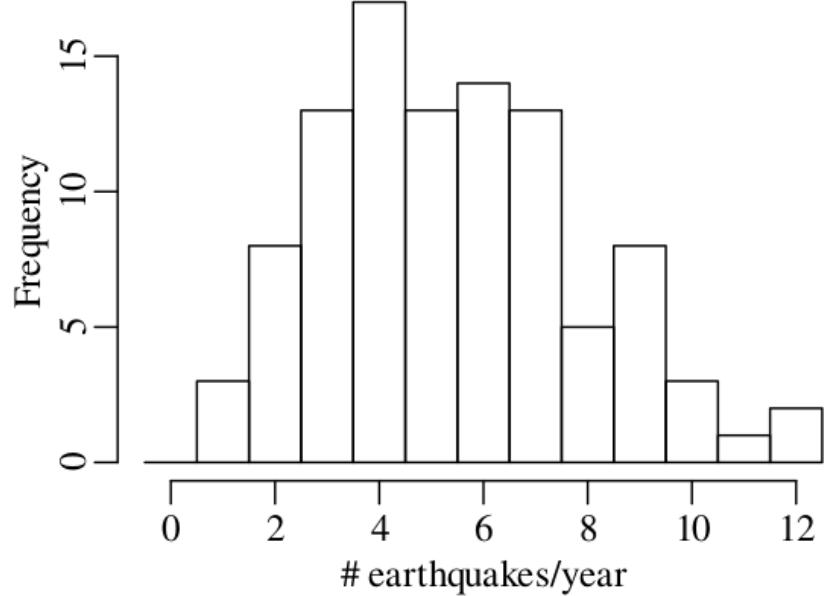
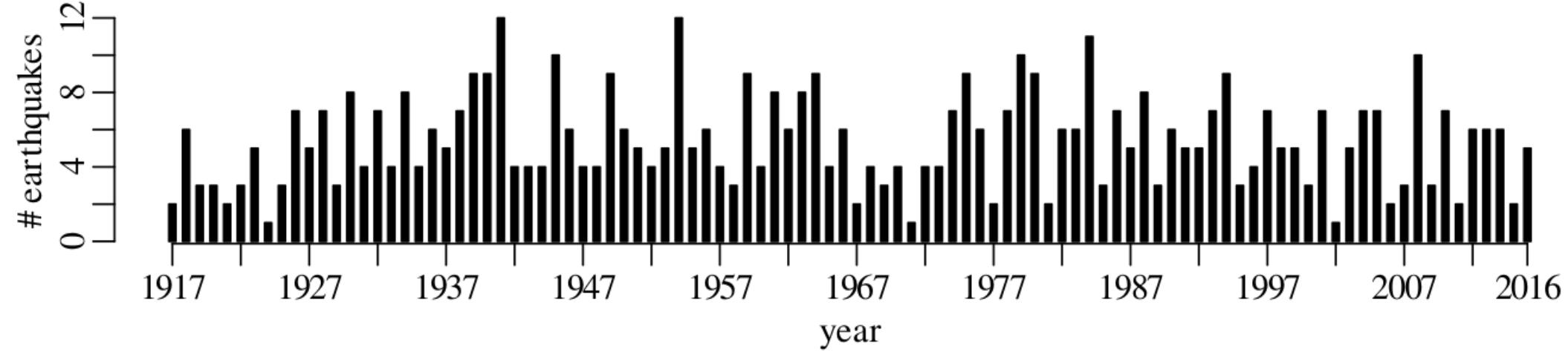
$$\text{C.I.}(p|k = 12, n = 30) = [0.23, 0.59]$$





Statistics for geoscientists

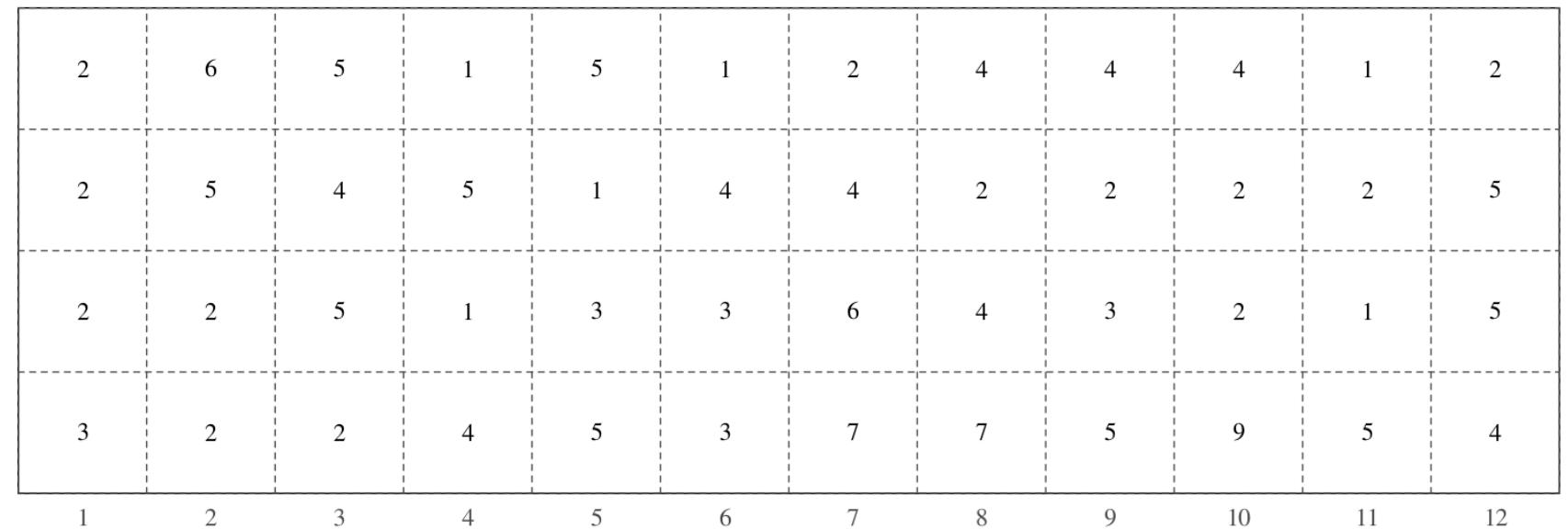
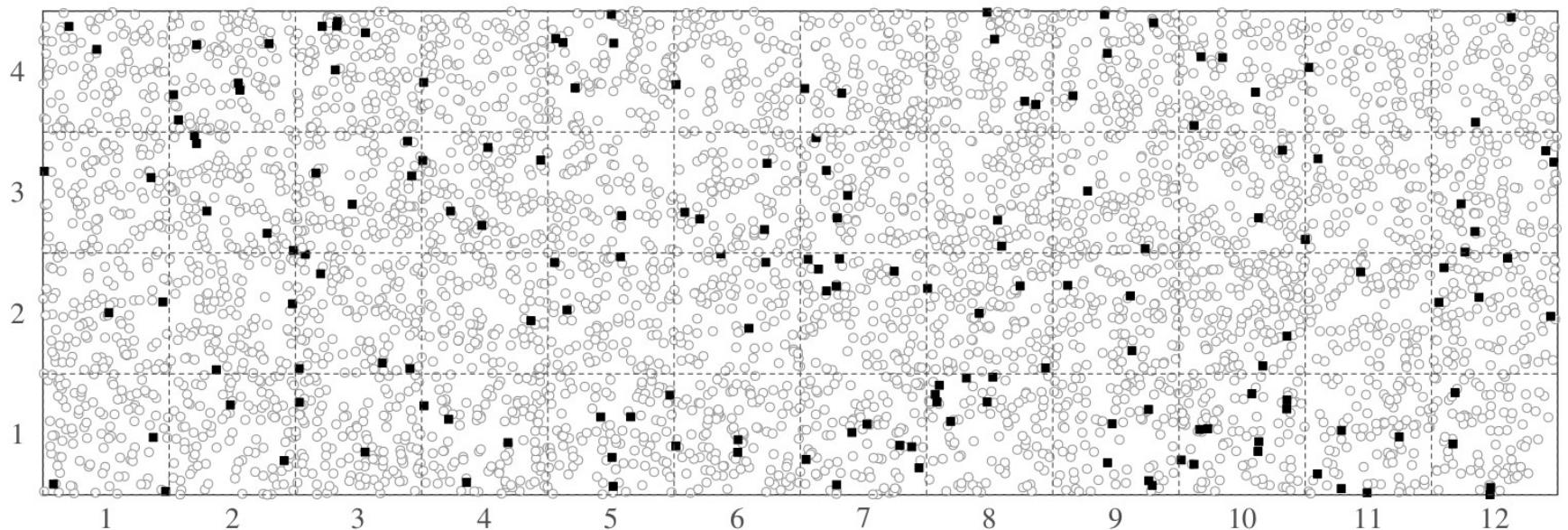
The Poisson distribution

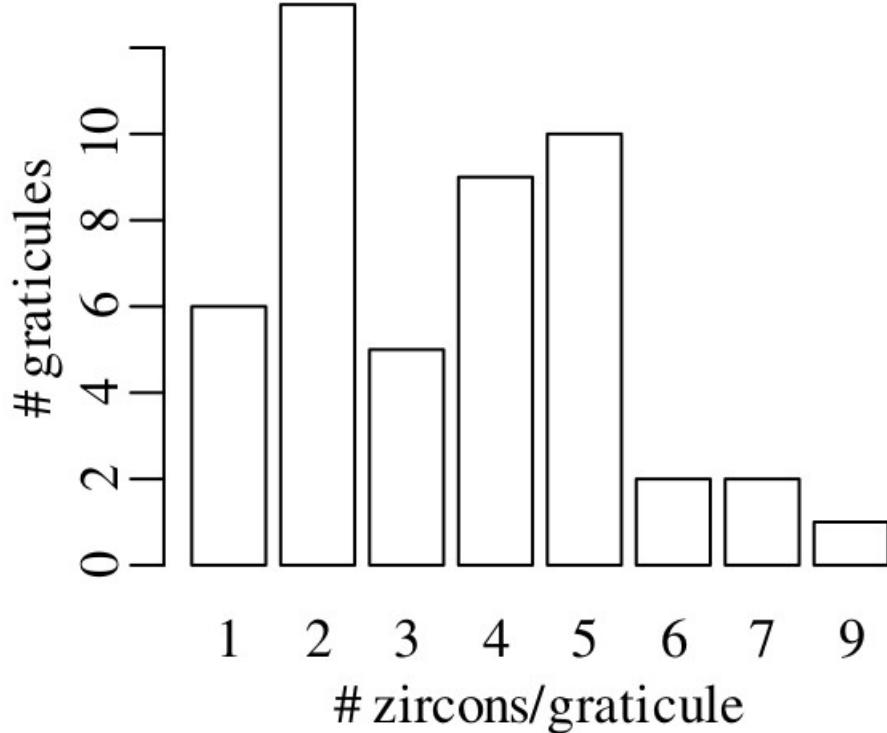


mean: 5.43

standard deviation: 2.50

variance: 6.24





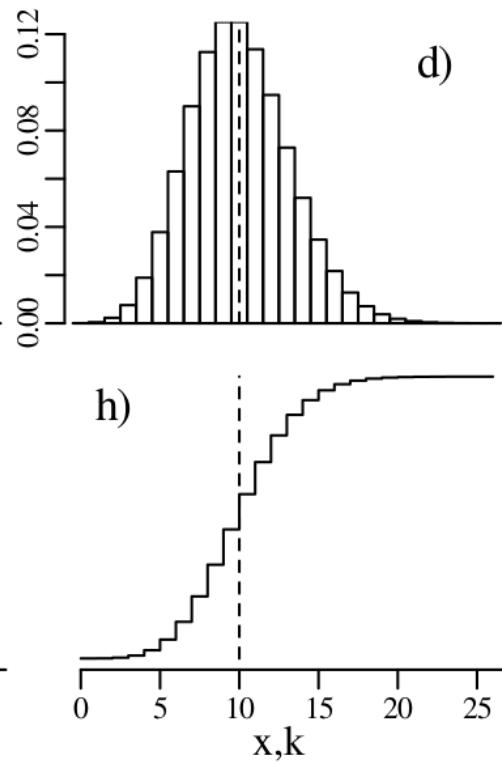
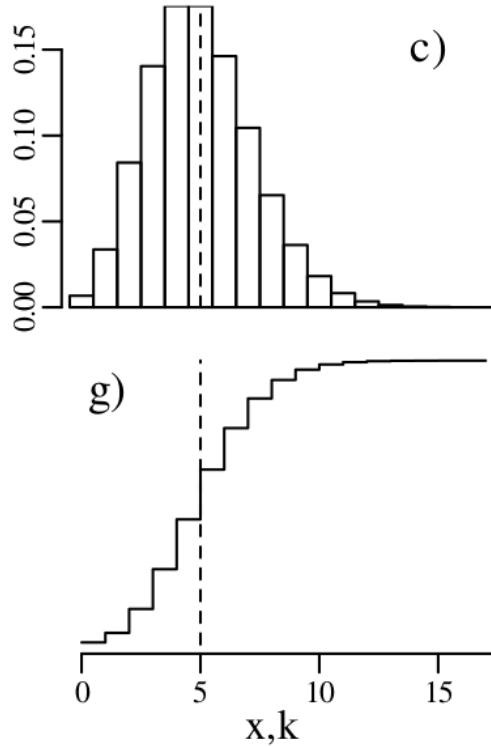
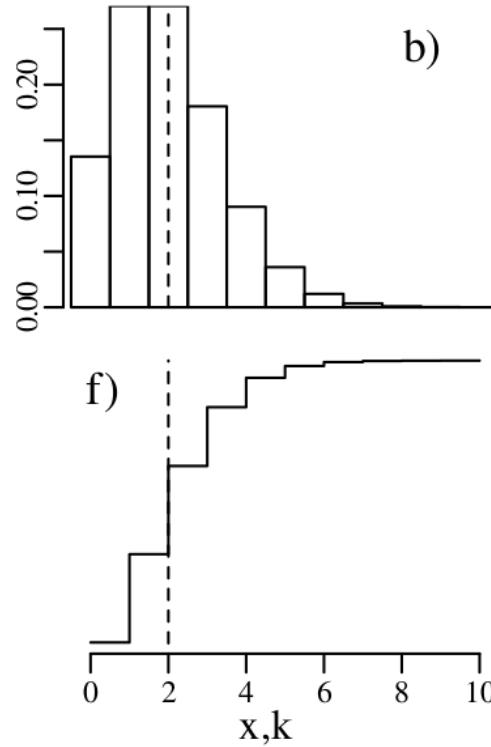
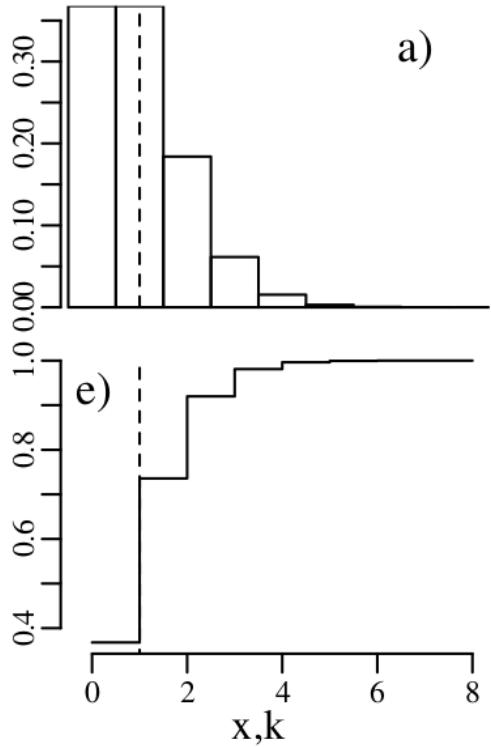
mean: 3.50

standard deviation: 1.85

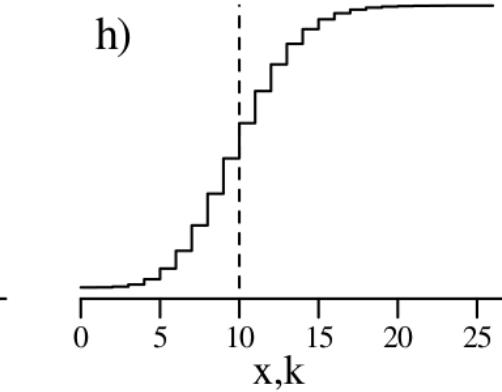
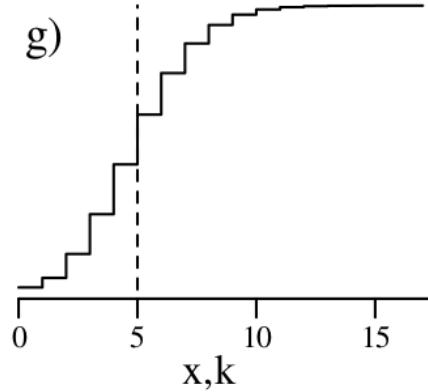
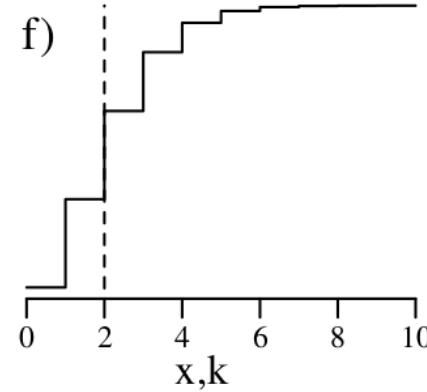
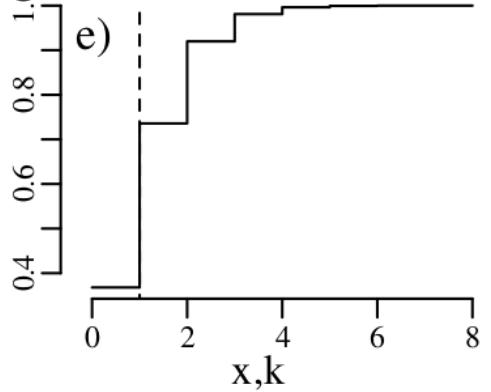
variance: 3.40

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

$P(k)$



$F(x)$



$$P(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k} \quad \xrightarrow{\begin{array}{l} n \rightarrow \infty \\ p \rightarrow 0 \end{array}} \quad P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{where} \quad \lambda = np$$

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_0 (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 $\alpha = 0.05$

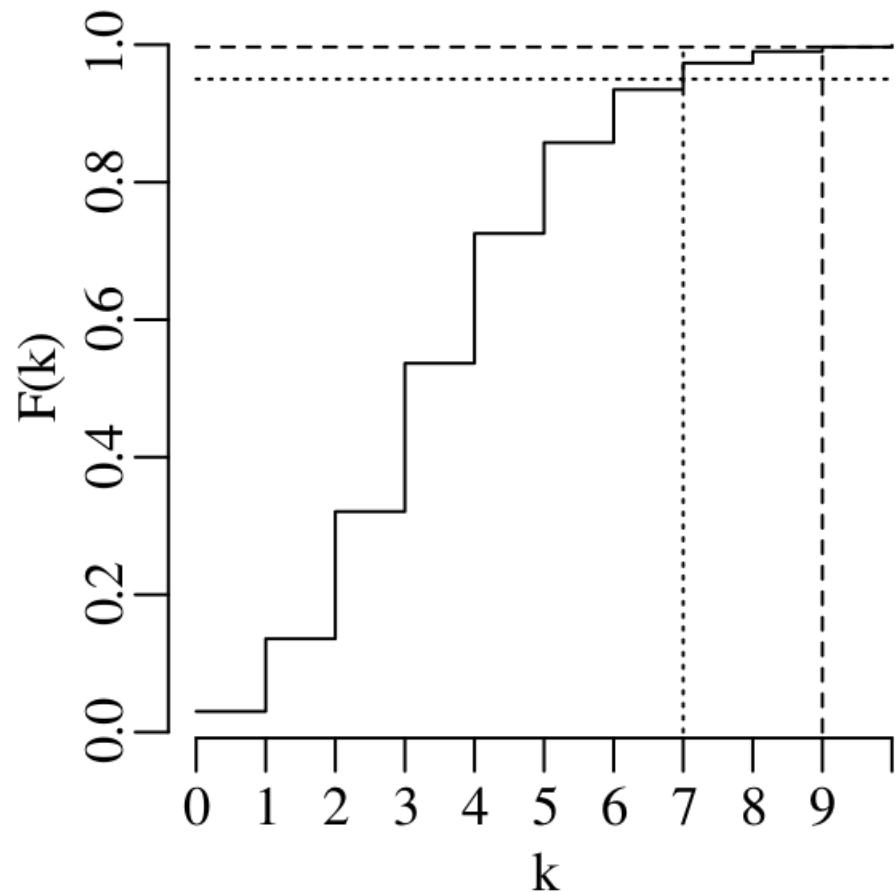
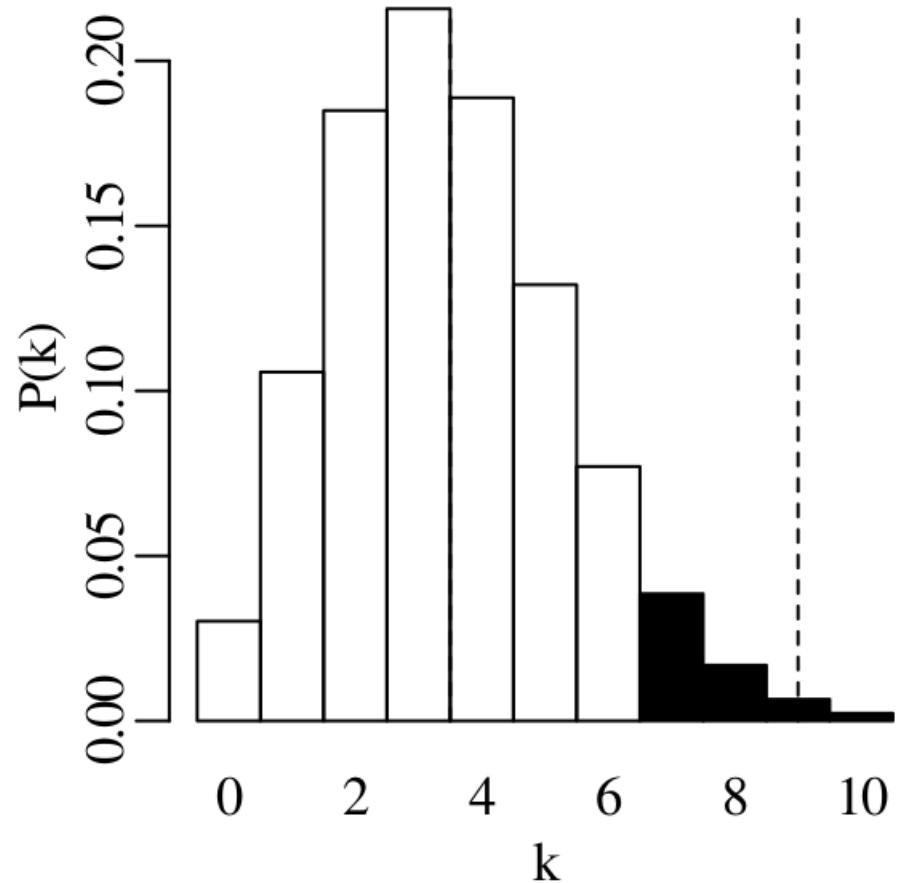
5. rejection region

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

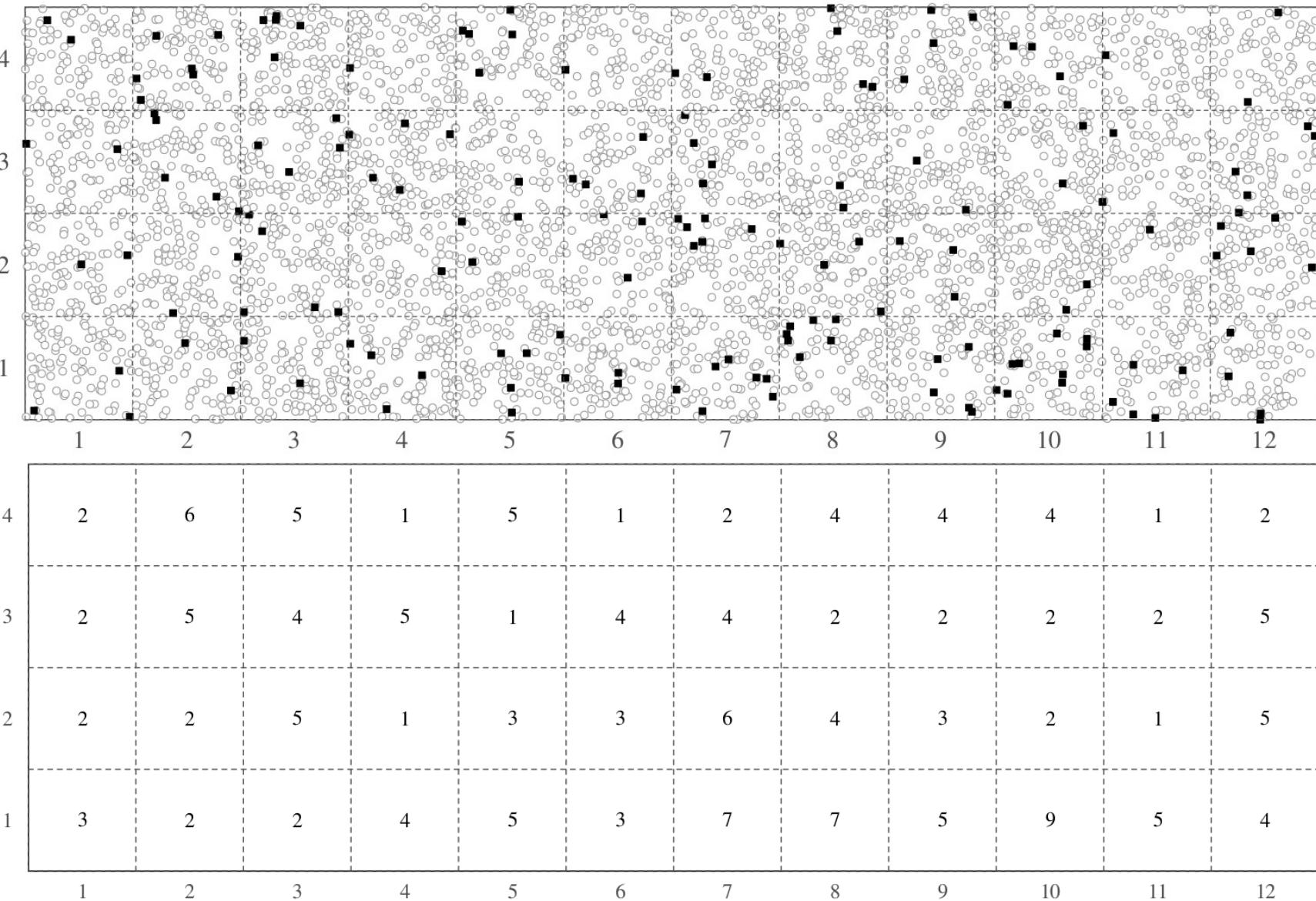
$$R = \{8, 9, 10, \dots, \infty\}$$

6. $k \in R$

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



Multiple testing



Assuming that H_0 is true, the probability of incurring a type-I error is:

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

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

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

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

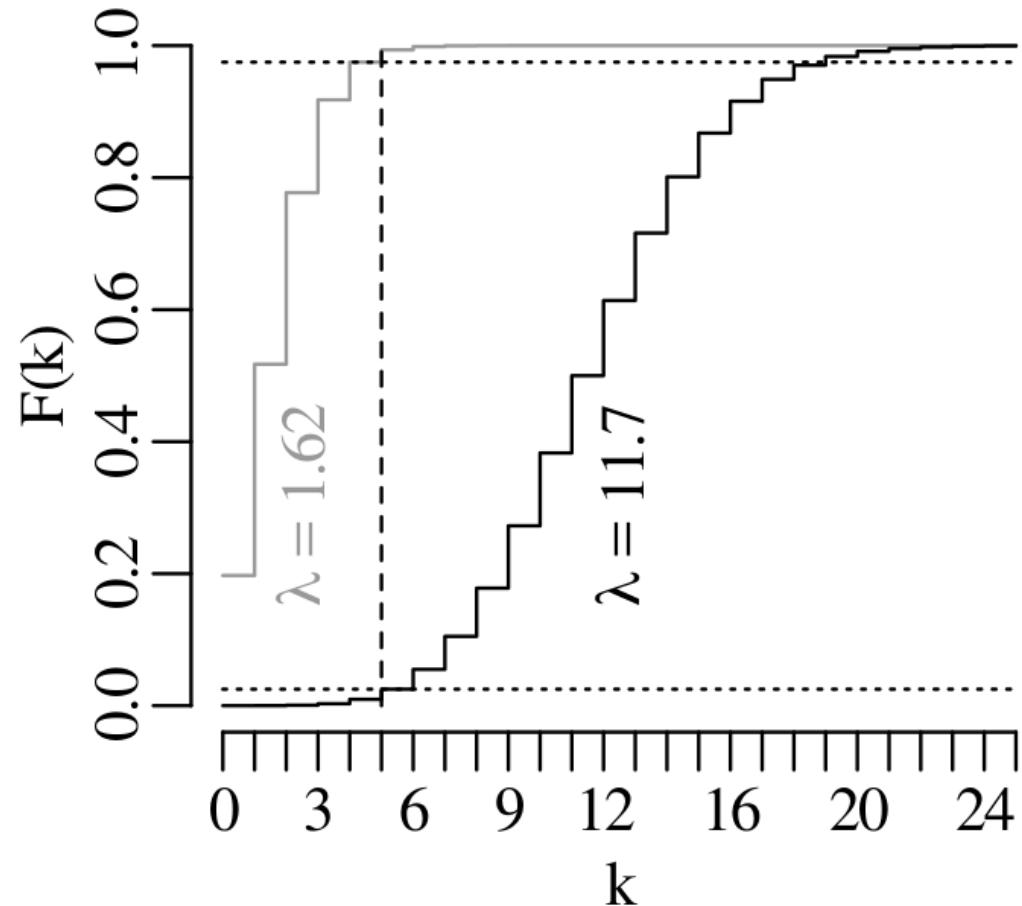
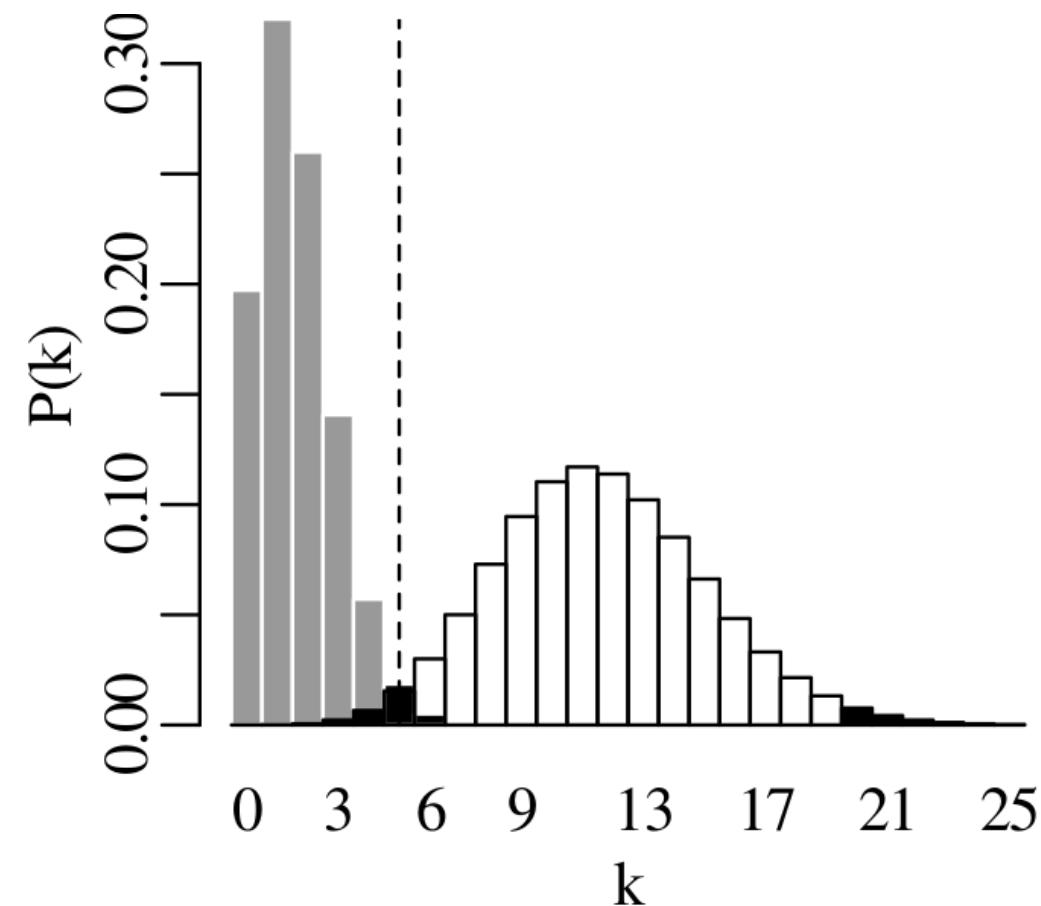
$$N \text{ experiments: } 1 - (1 - \alpha)^N\%$$

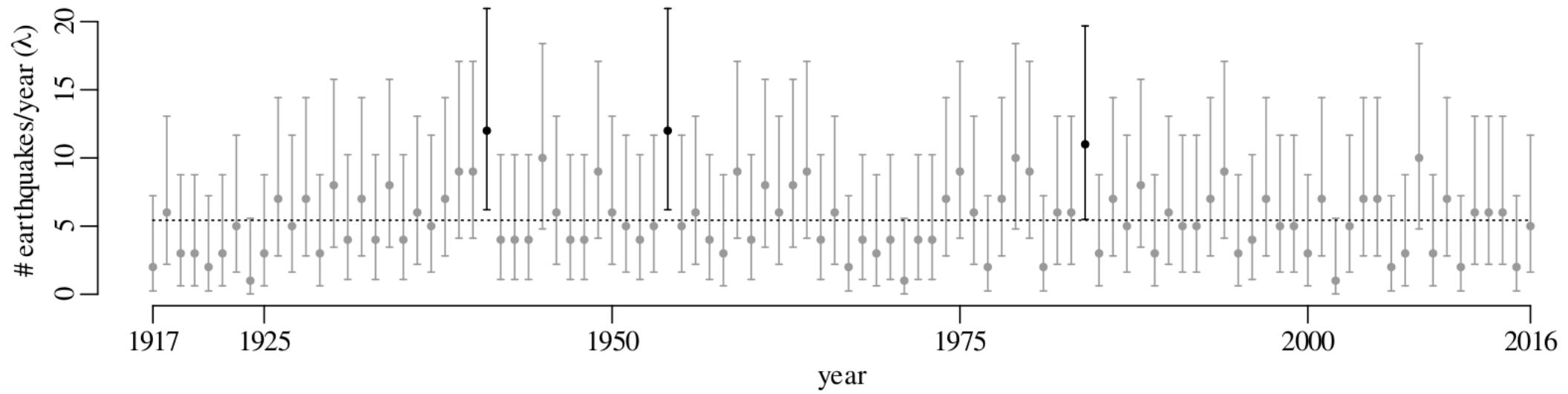
Bonferroni correction: α/N

Confidence intervals

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

What is λ ?





Statistics for geoscientists

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

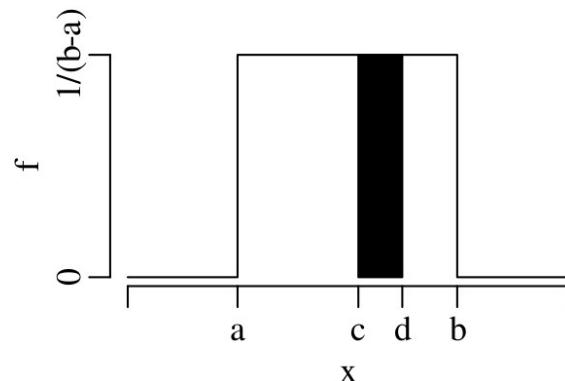
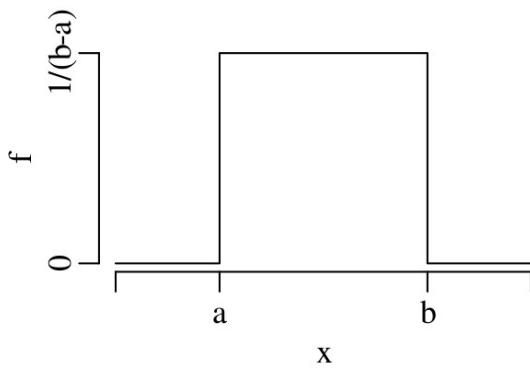
$$P(x|r,p) = \binom{r+x-1}{x} (1-p)^x p^r$$

the multinomial distribution

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

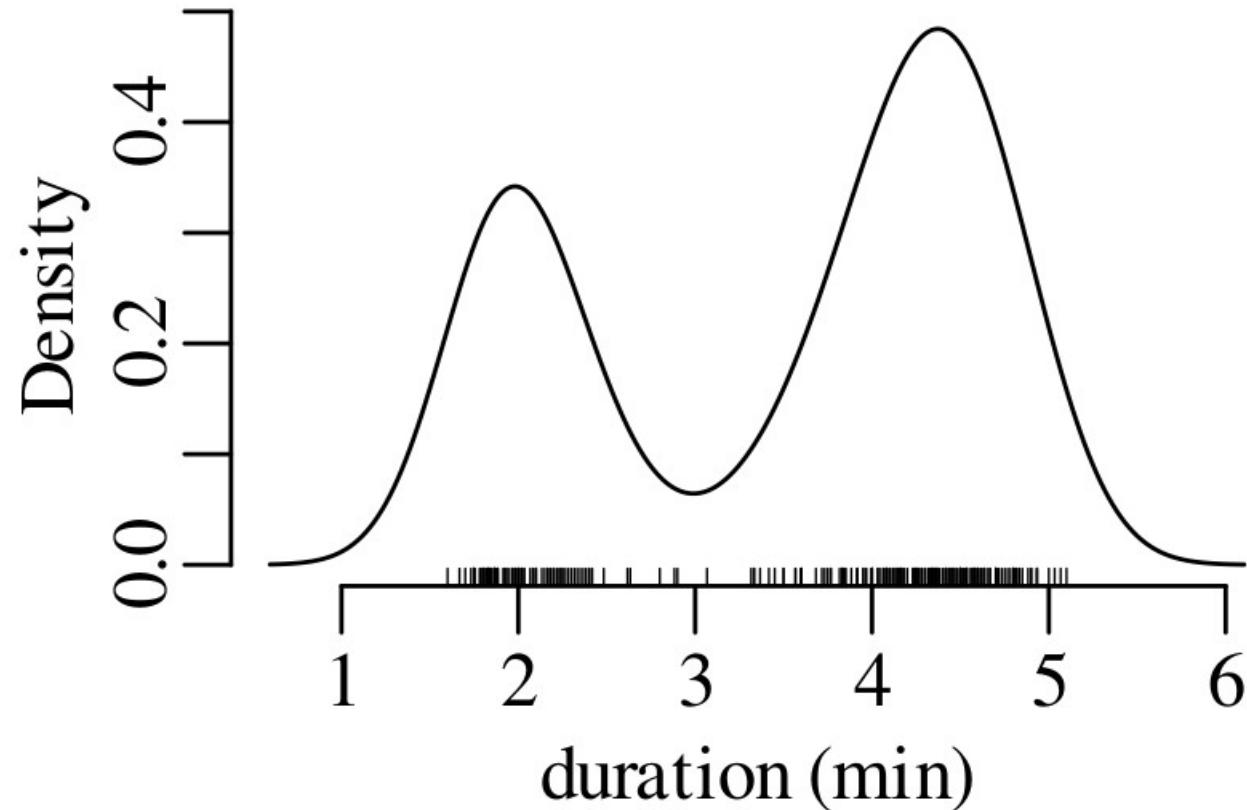
the uniform distribution

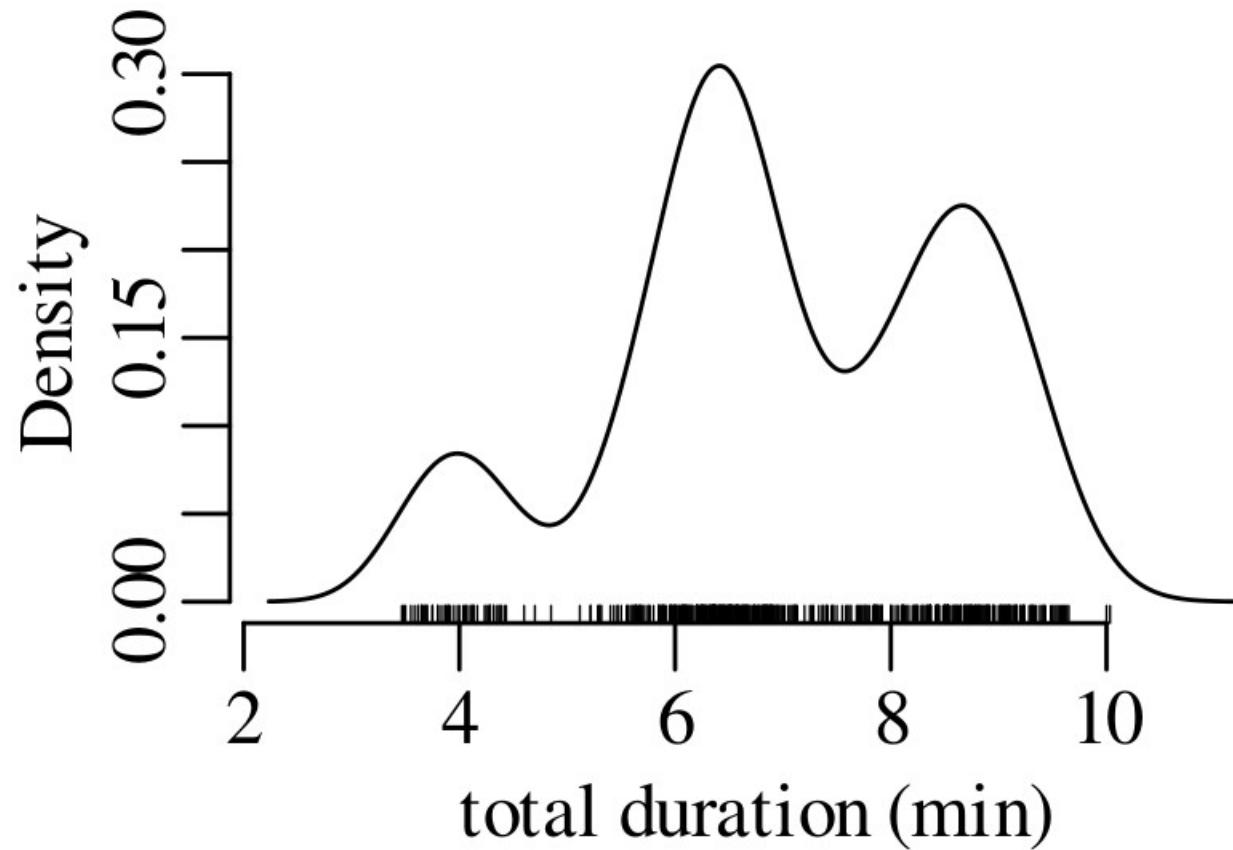
$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

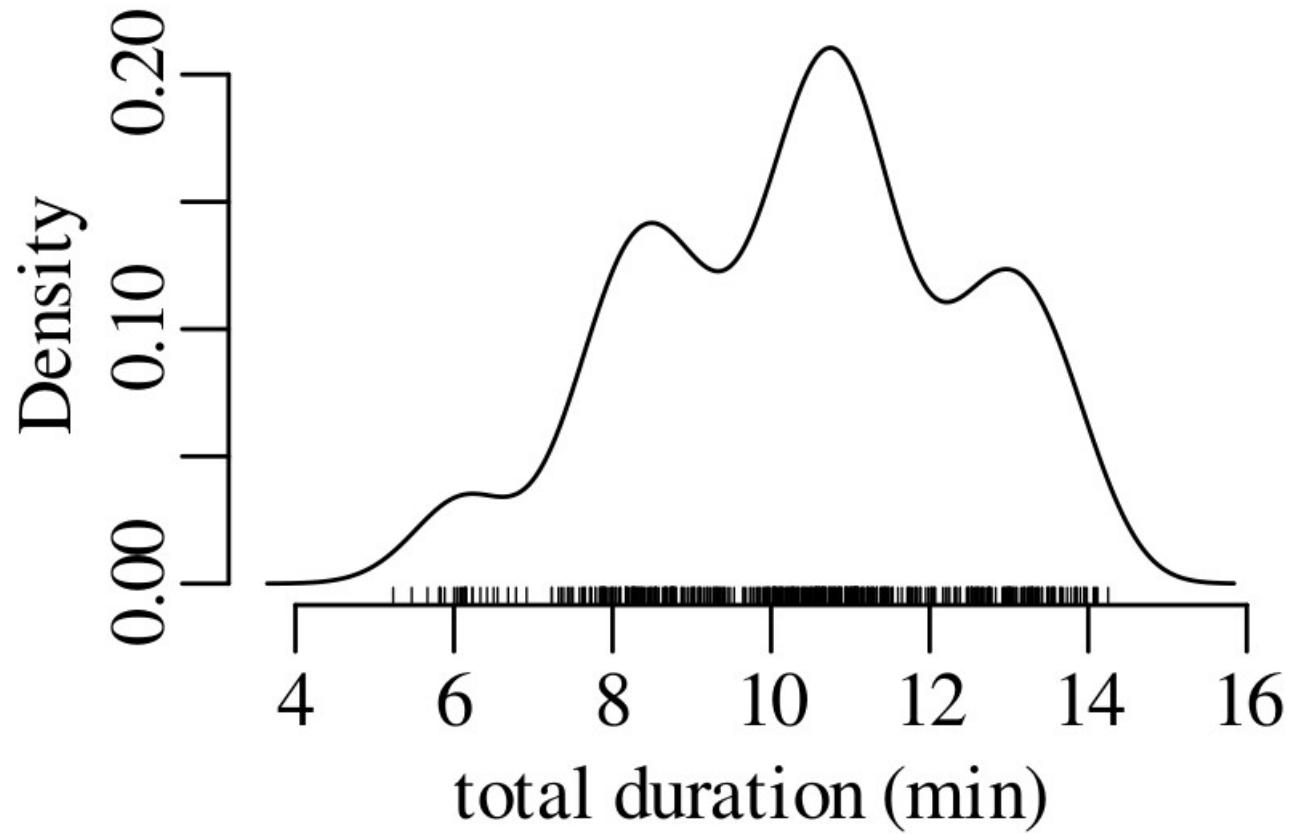


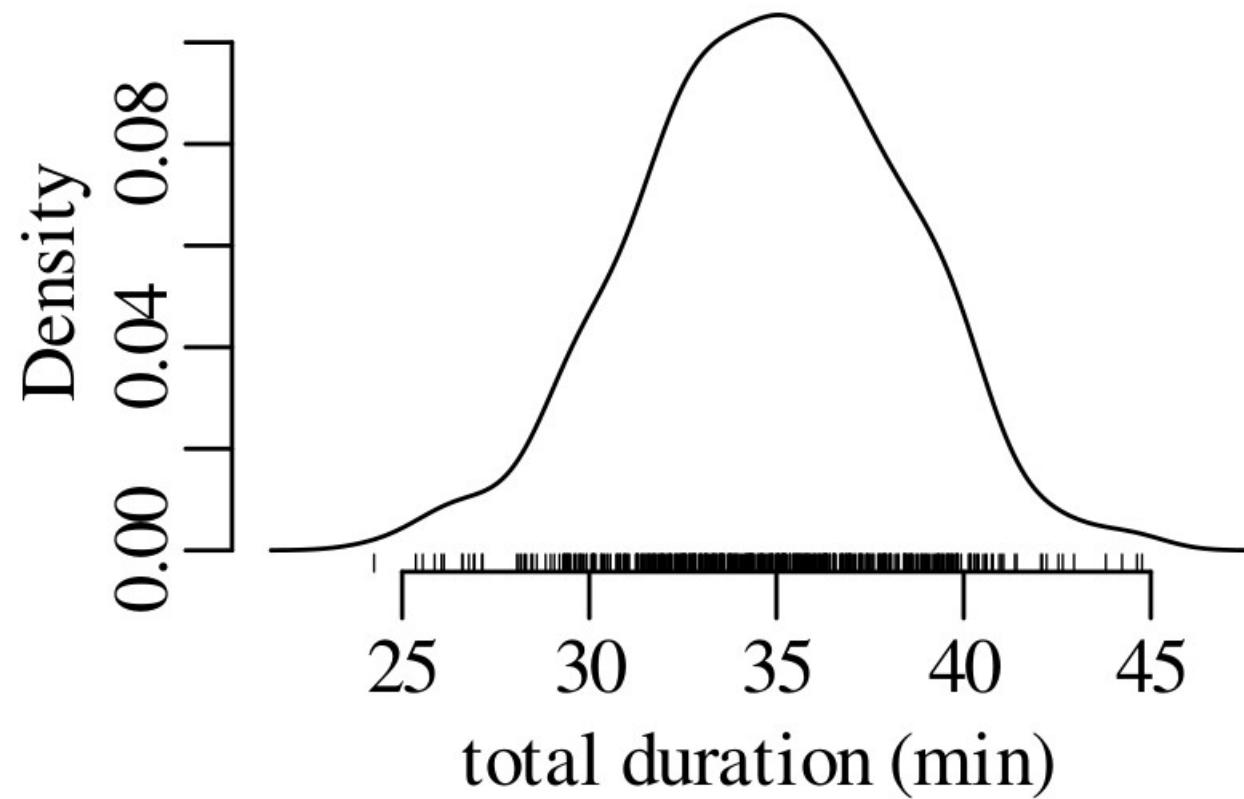
$$P(c \leq x \leq d) = \int_c^d f(x|a,b) dx$$

What is so normal about the Gaussian distribution?

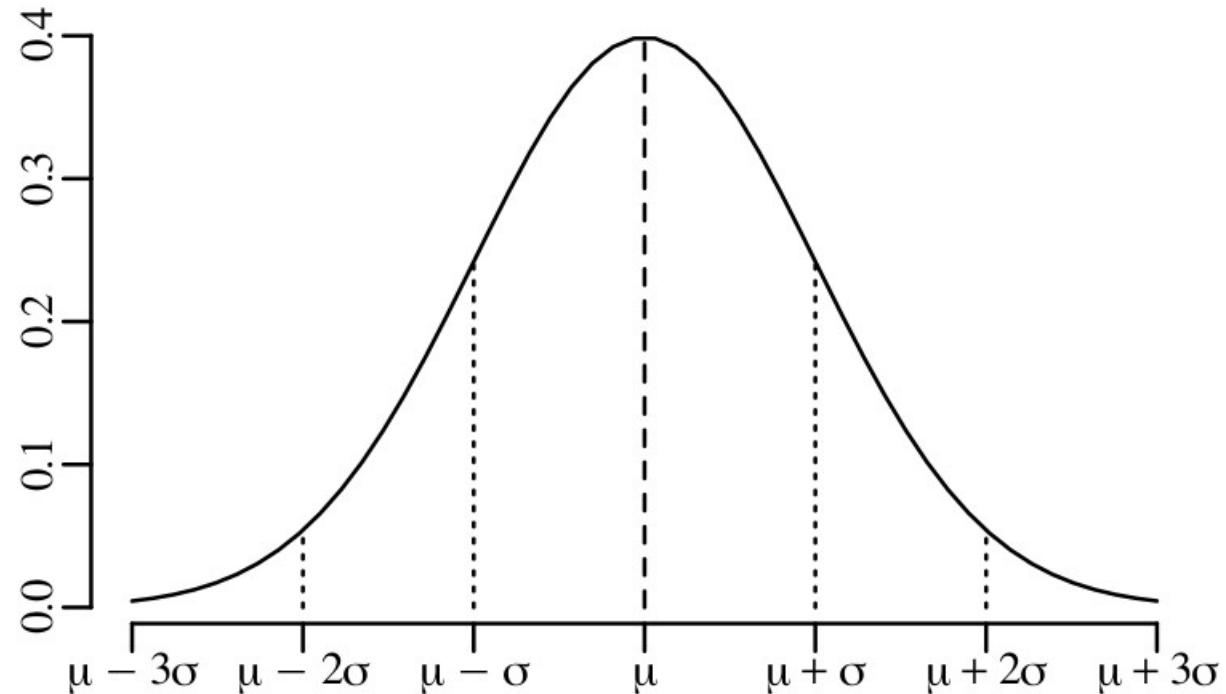


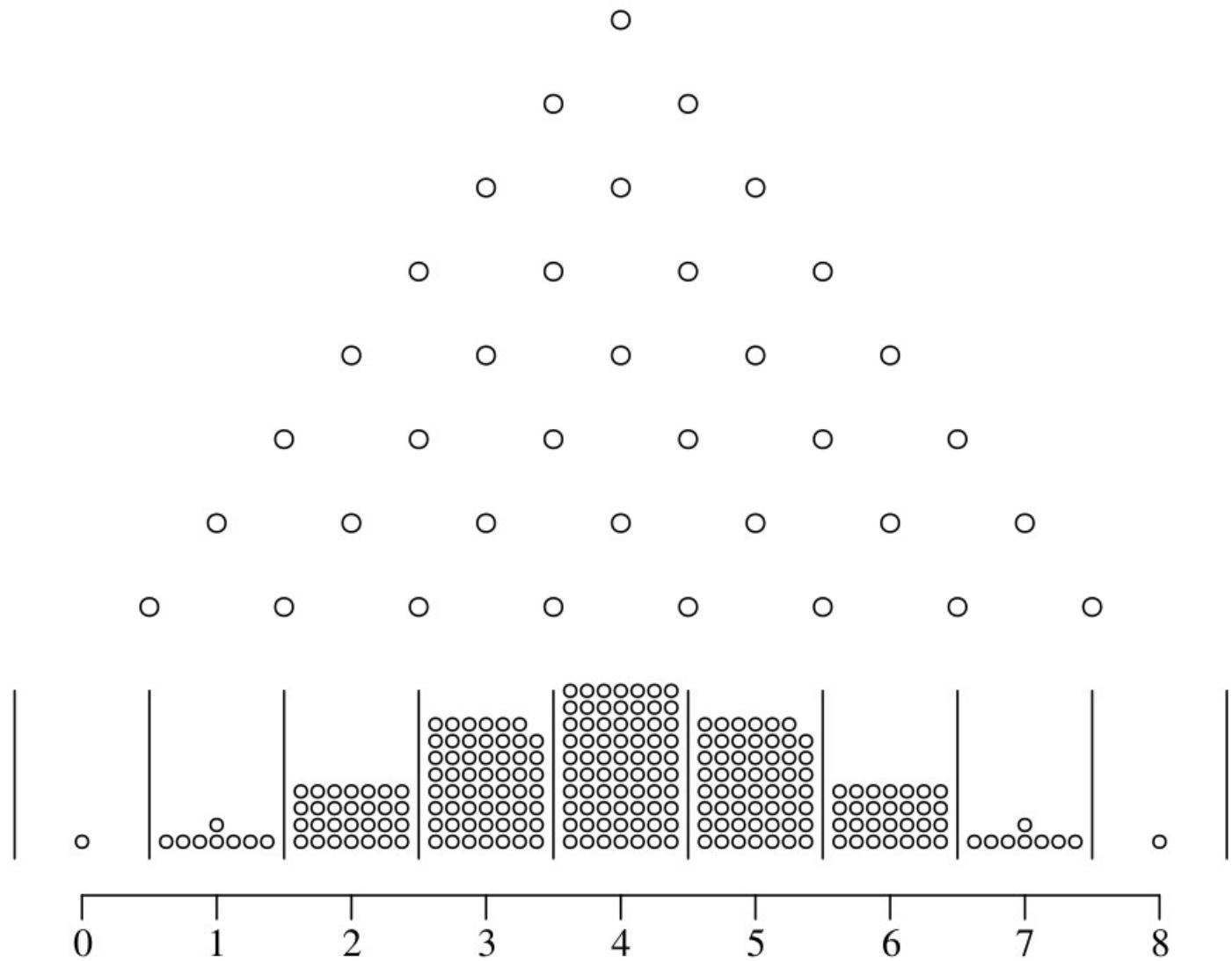


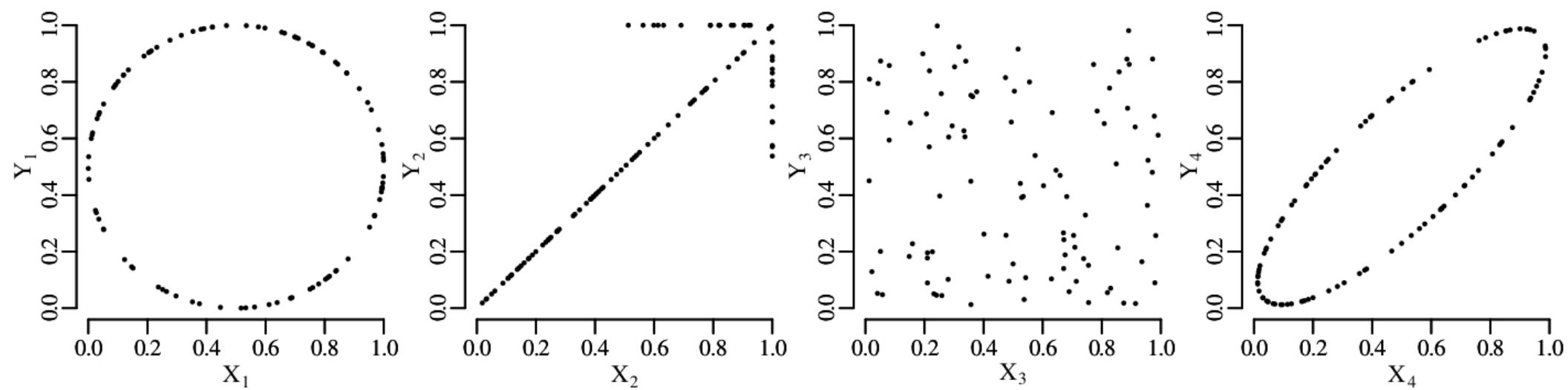
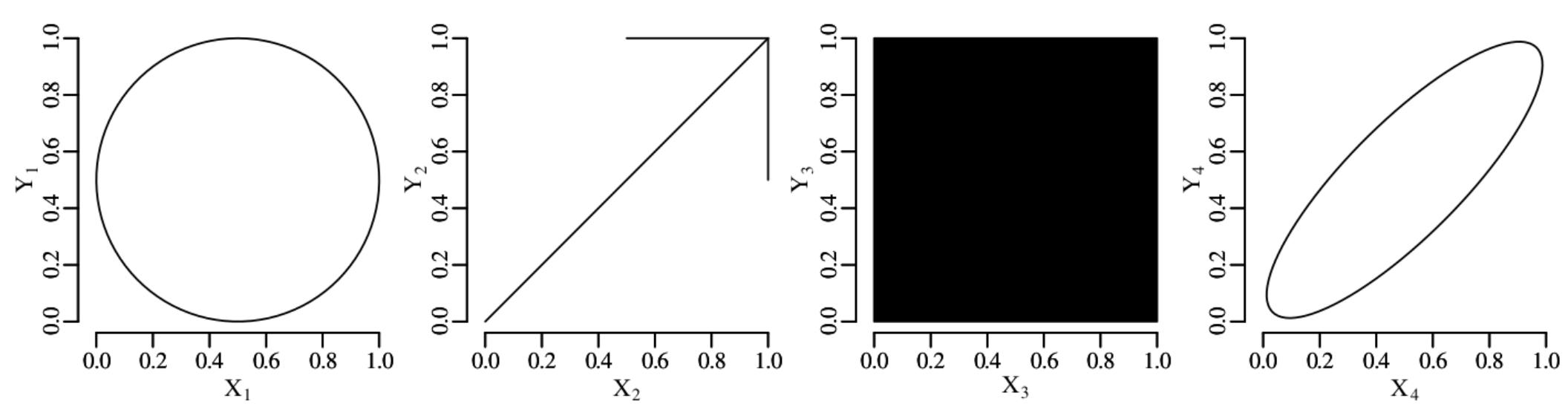


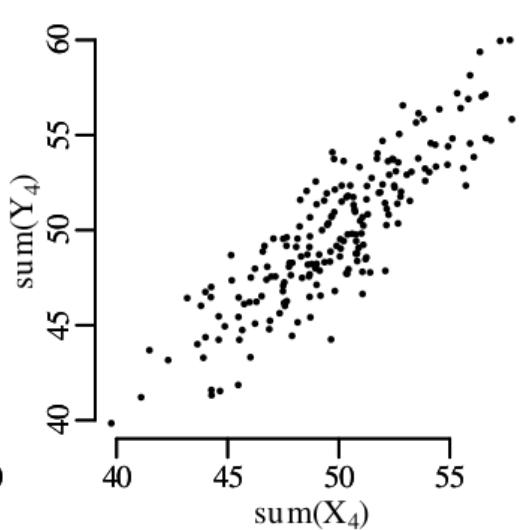
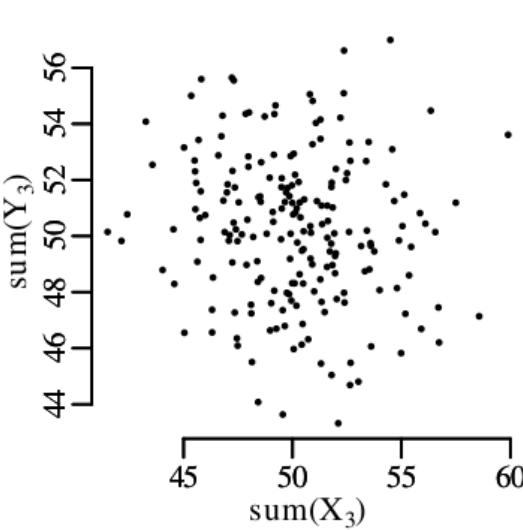
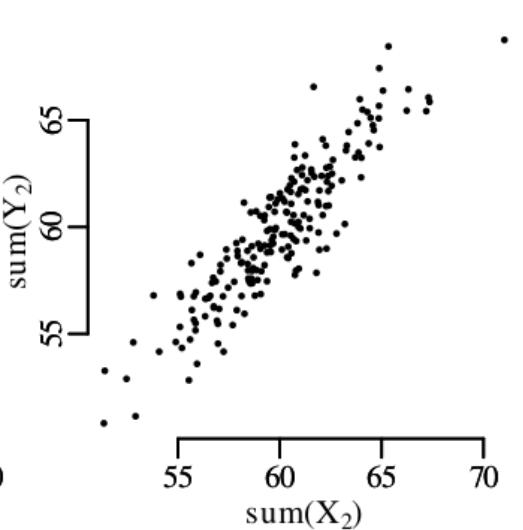
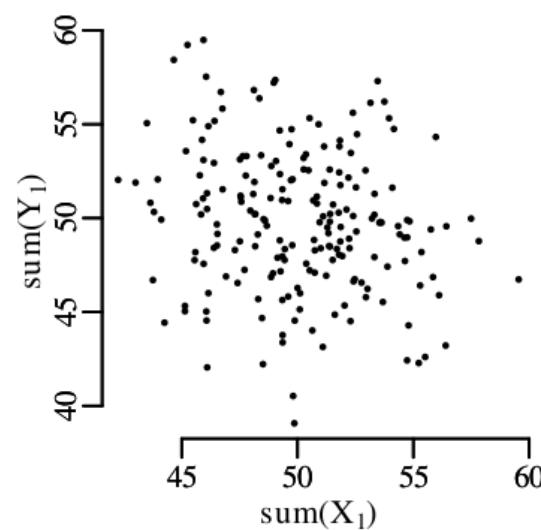
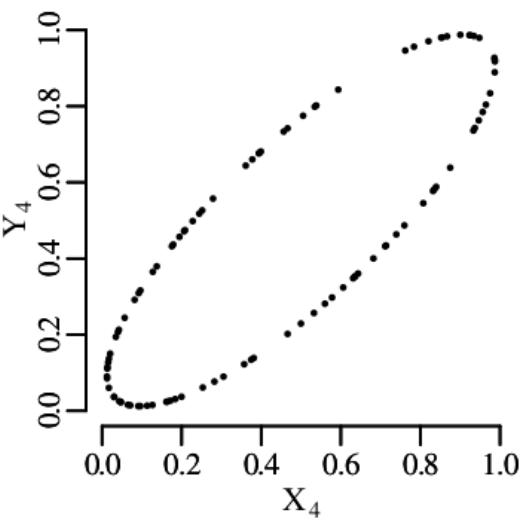
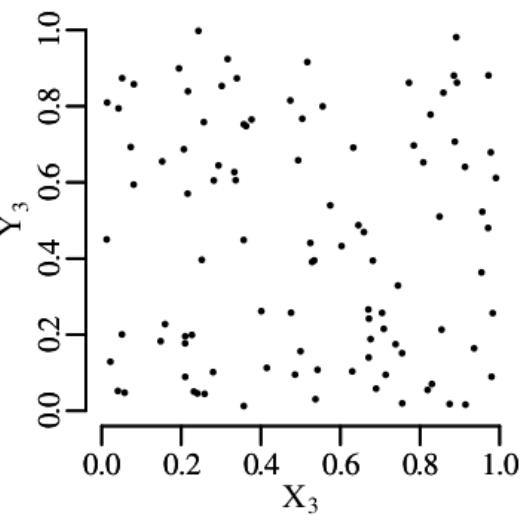
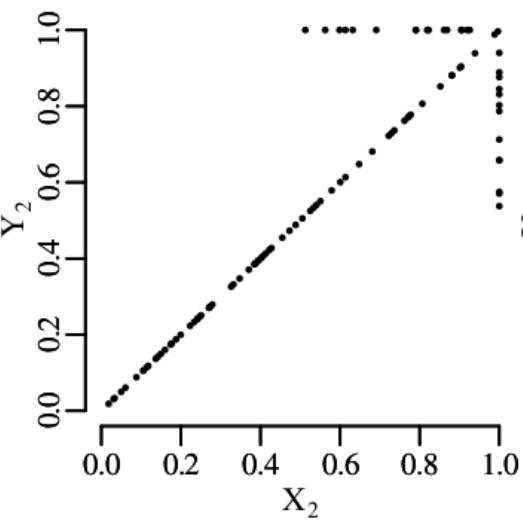
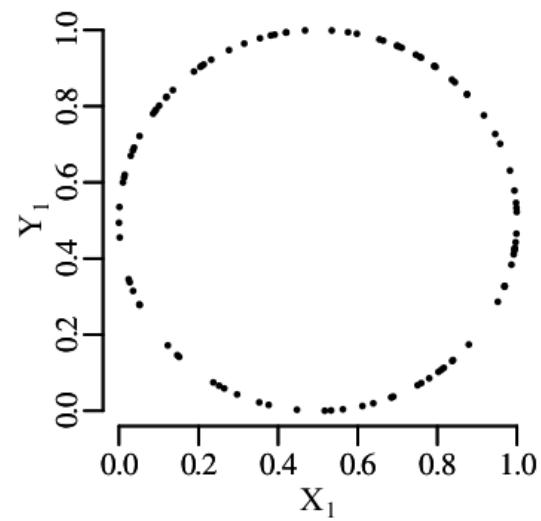


$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

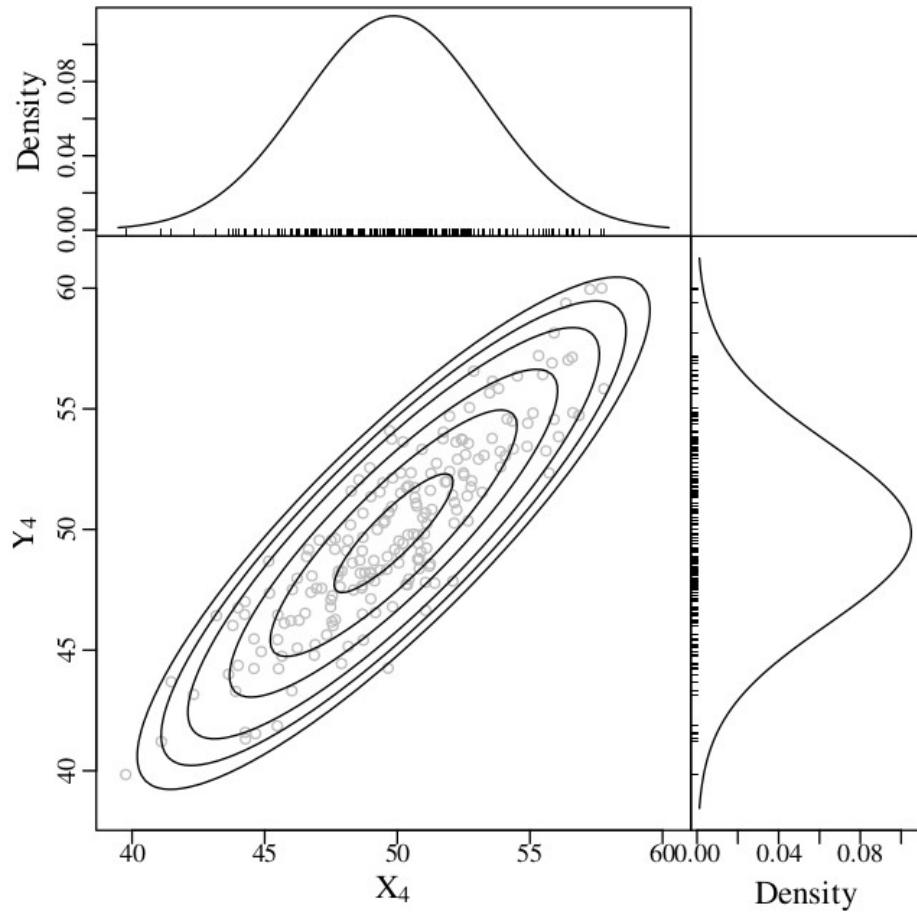




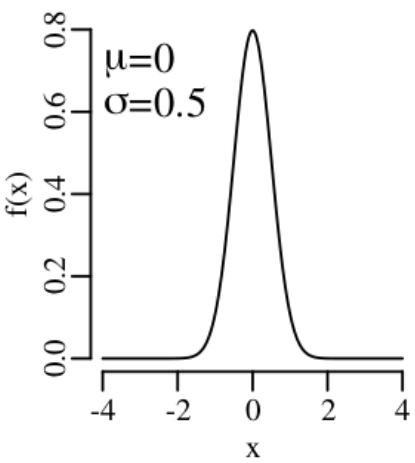
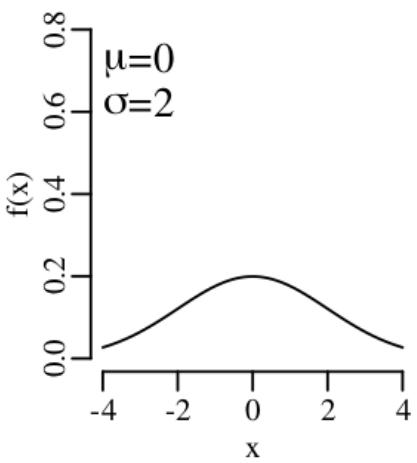
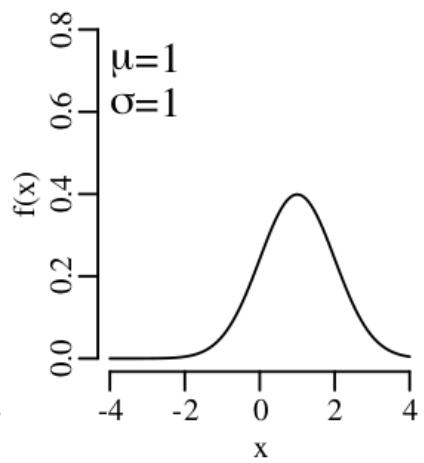
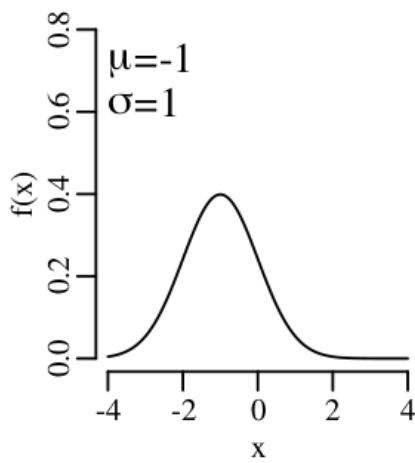
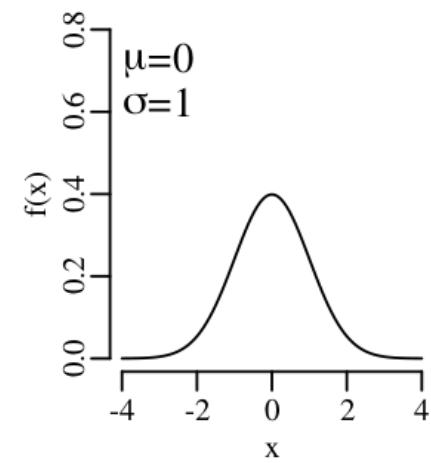




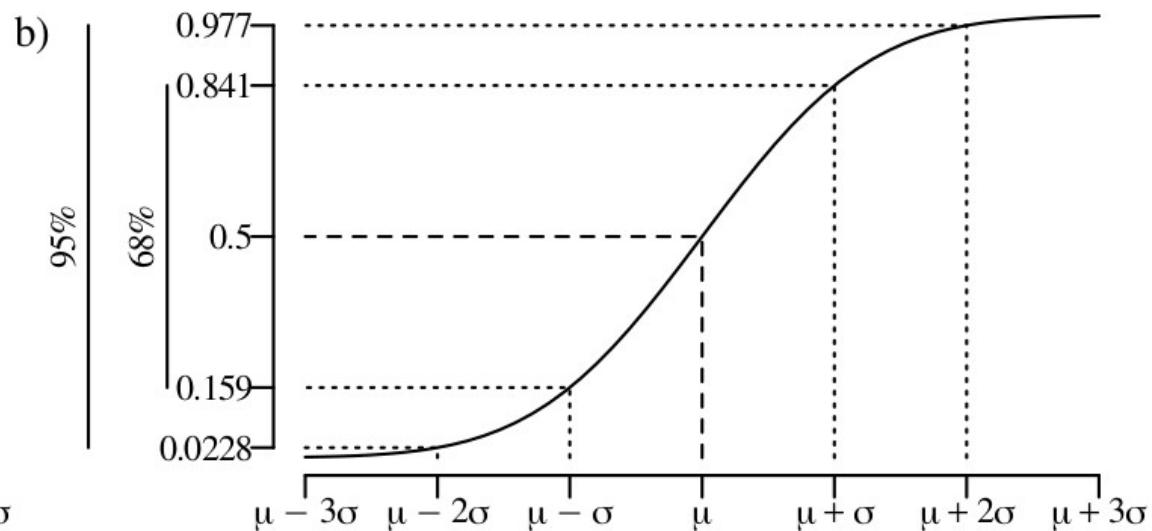
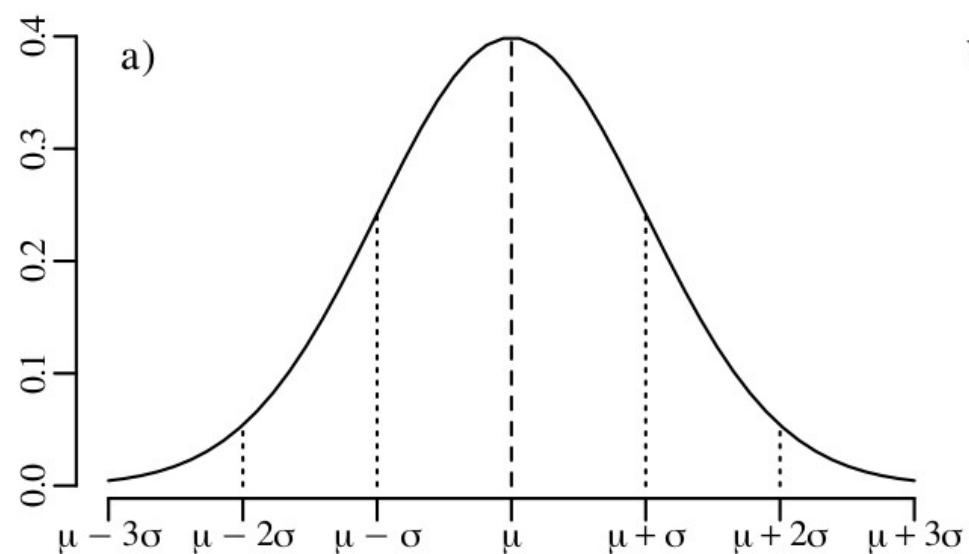
$$f(x, y | \mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{x,y}) = \frac{\exp\left(-\left[(x - \mu_x) \quad (y - \mu_y)\right] \begin{bmatrix} \sigma_x^2 & \sigma_{x,y} \\ \sigma_{x,y} & \sigma_y^2 \end{bmatrix}^{-1} \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix} / 2\right)}{2\pi \sqrt{\left|\begin{bmatrix}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



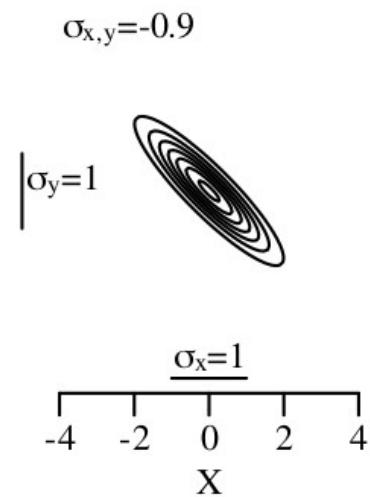
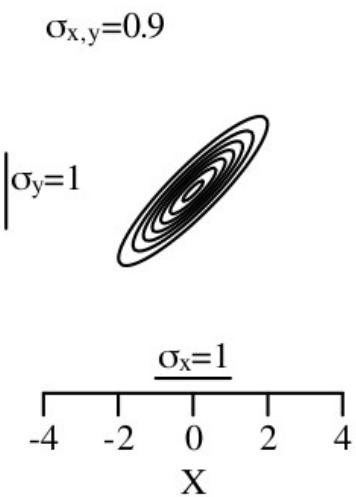
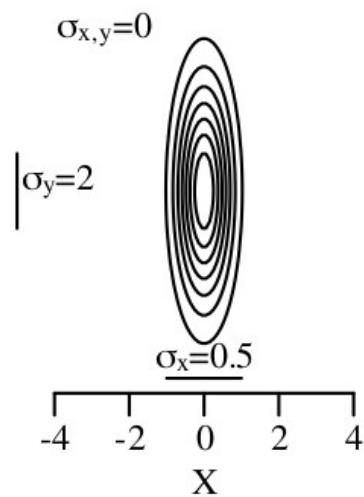
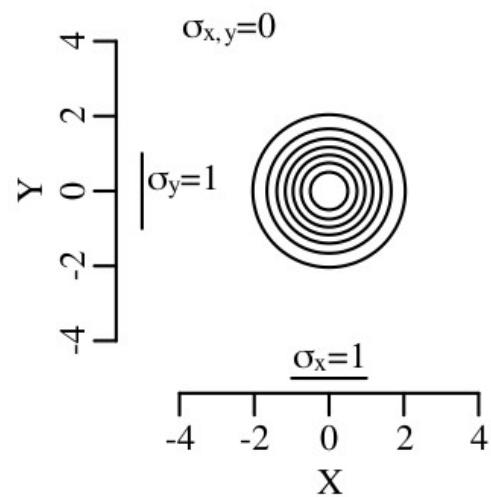
$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



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



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



Parameter estimation

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

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

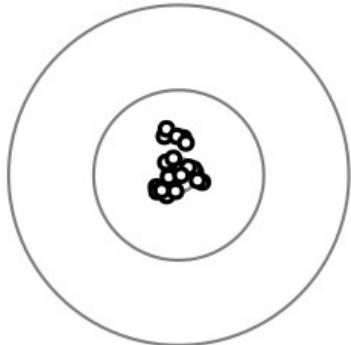
$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2} \qquad \longleftrightarrow \qquad s[x] = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

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

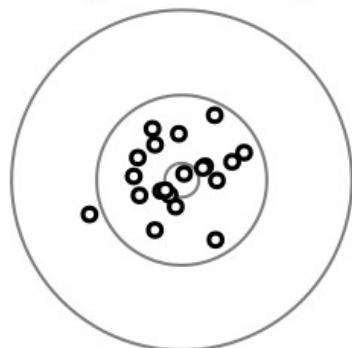
Statistics for geoscientists

Error propagation

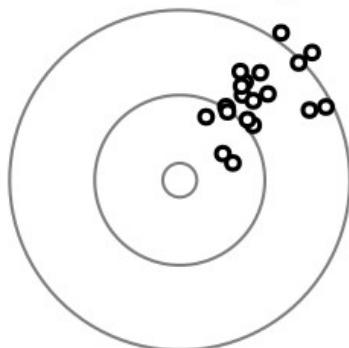
high precision
high accuracy



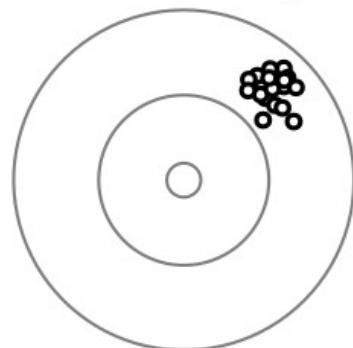
low precision
high accuracy



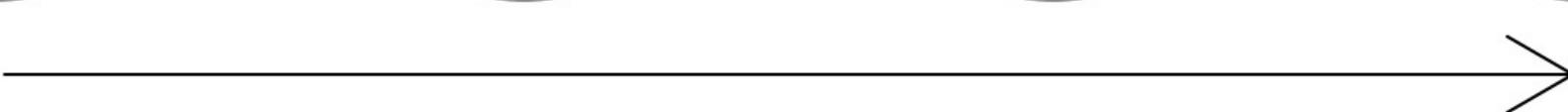
low precision
low accuracy



high precision
low accuracy



good



bad

$z = g(x) \rightarrow$ 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

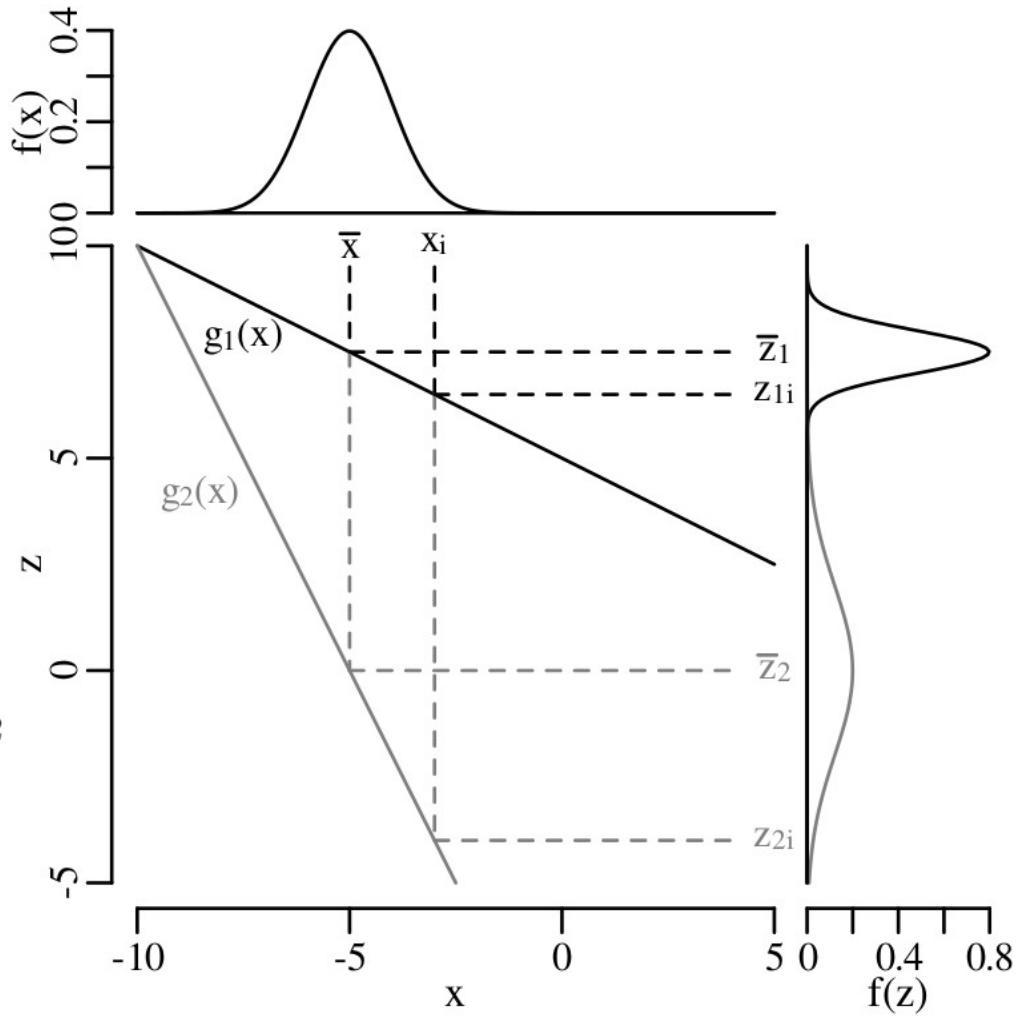
$$z = g(x)$$

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

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

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

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

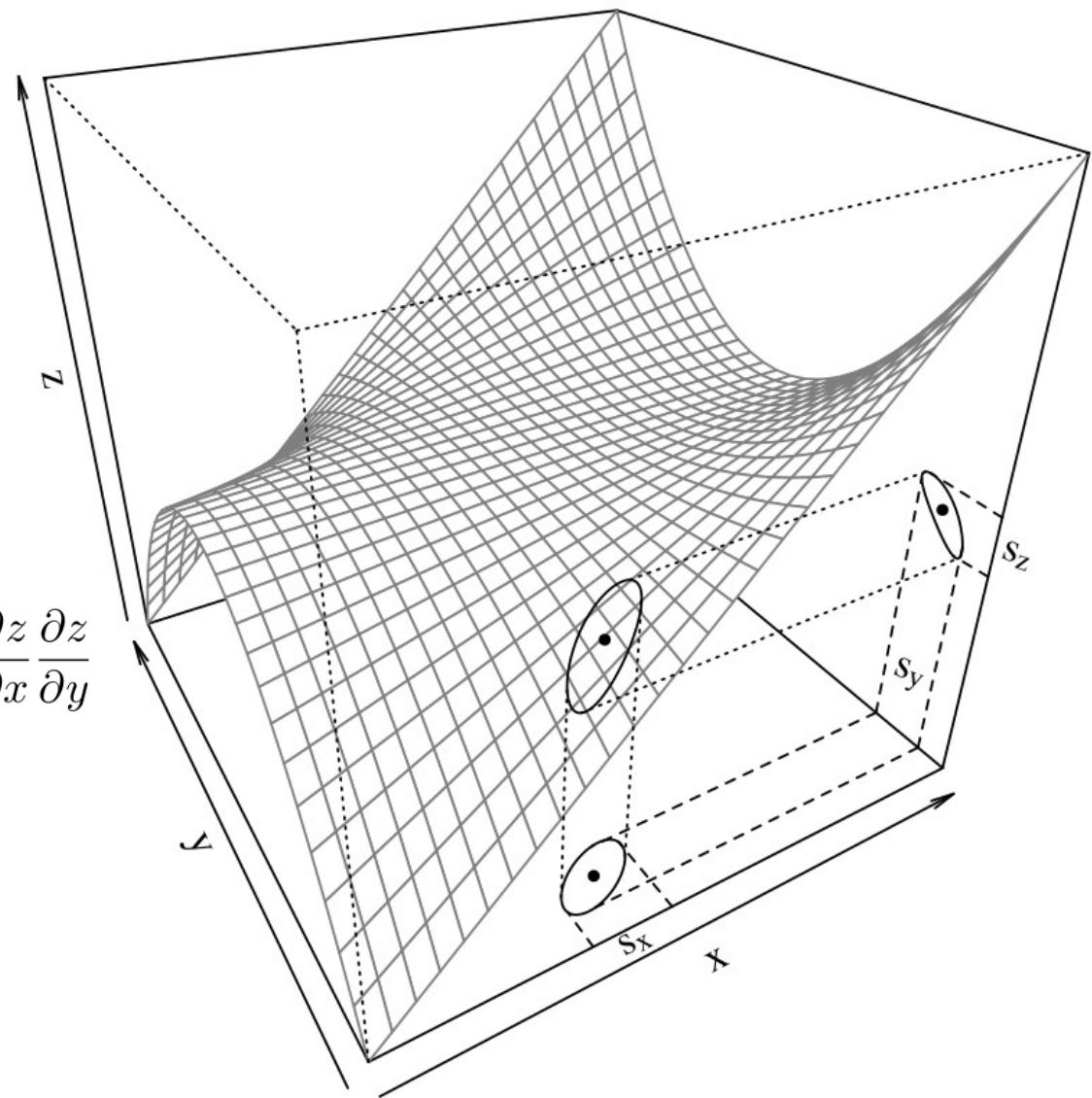


$$z = g(x, y)$$

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

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

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



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

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

$$d = d_o + v_o t + g t^2$$

can be written as

$$d = x + y \quad \text{where} \quad x \equiv d_o + v_o t \quad \text{and} \quad y \equiv g t^2$$
$$s[d] = \sqrt{s[x]^2 + s[y]^2}$$
$$s[x]^2 = (v_o s[t])^2$$
$$s[y]^2 = g^2 \left(\frac{s[t]}{t} \right)^2$$
$$s[d] = \sqrt{(v_o s[t])^2 + g^2 \left(\frac{s[t]}{t} \right)^2}$$

```
graph TD; A["d = x + y  
x ≡ d_o + v_o t  
y ≡ g t^2"] --> B["s[d] = sqrt(s[x]^2 + s[y]^2)"]; B --> C["s[x]^2 = (v_o s[t])^2"]; B --> D["s[y]^2 = g^2 (s[t]/t)^2"]; C --> E["s[d] = sqrt((v_o s[t])^2 + g^2 (s[t]/t)^2)"]; D --> E;
```

$$z = g(x)$$

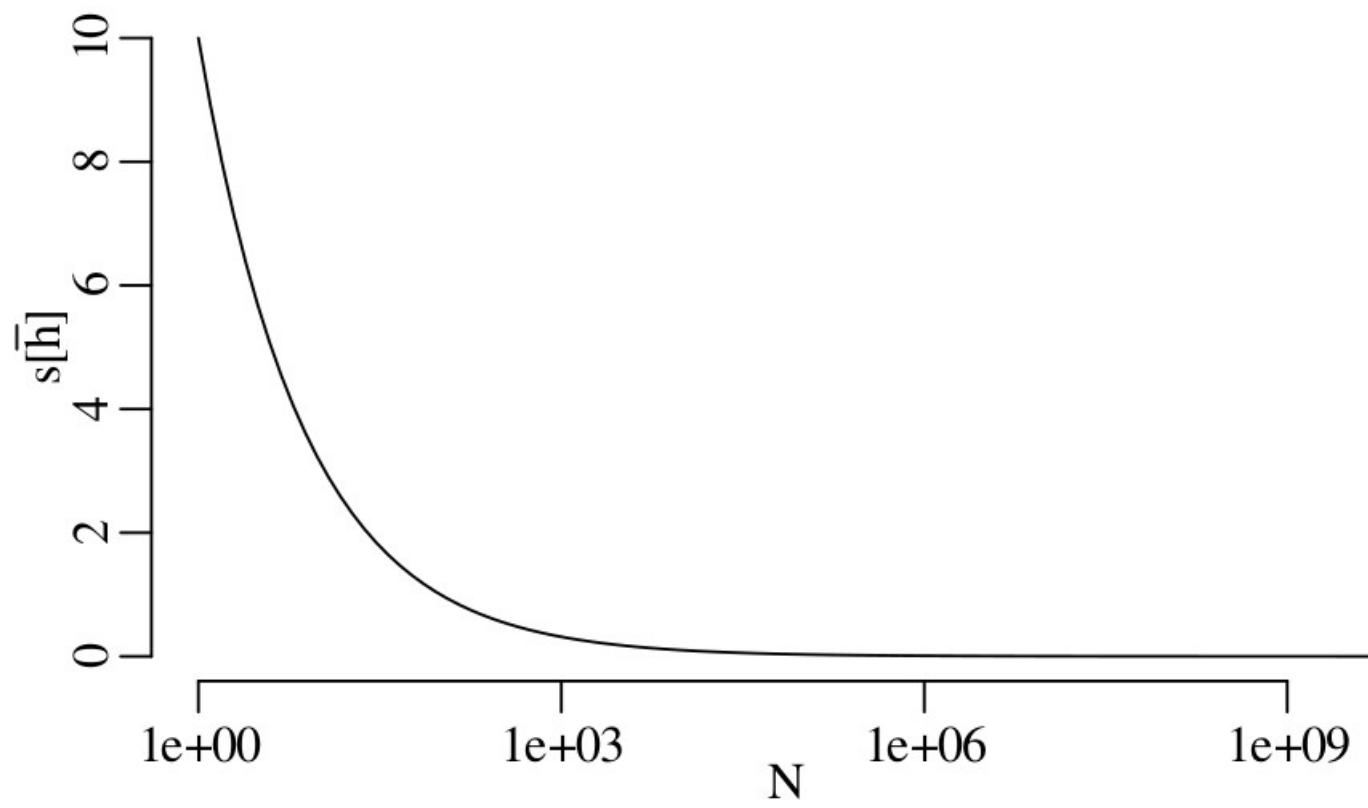
$$\bar{x} = \sum_{i=1}^n \frac{x_i}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

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

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

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

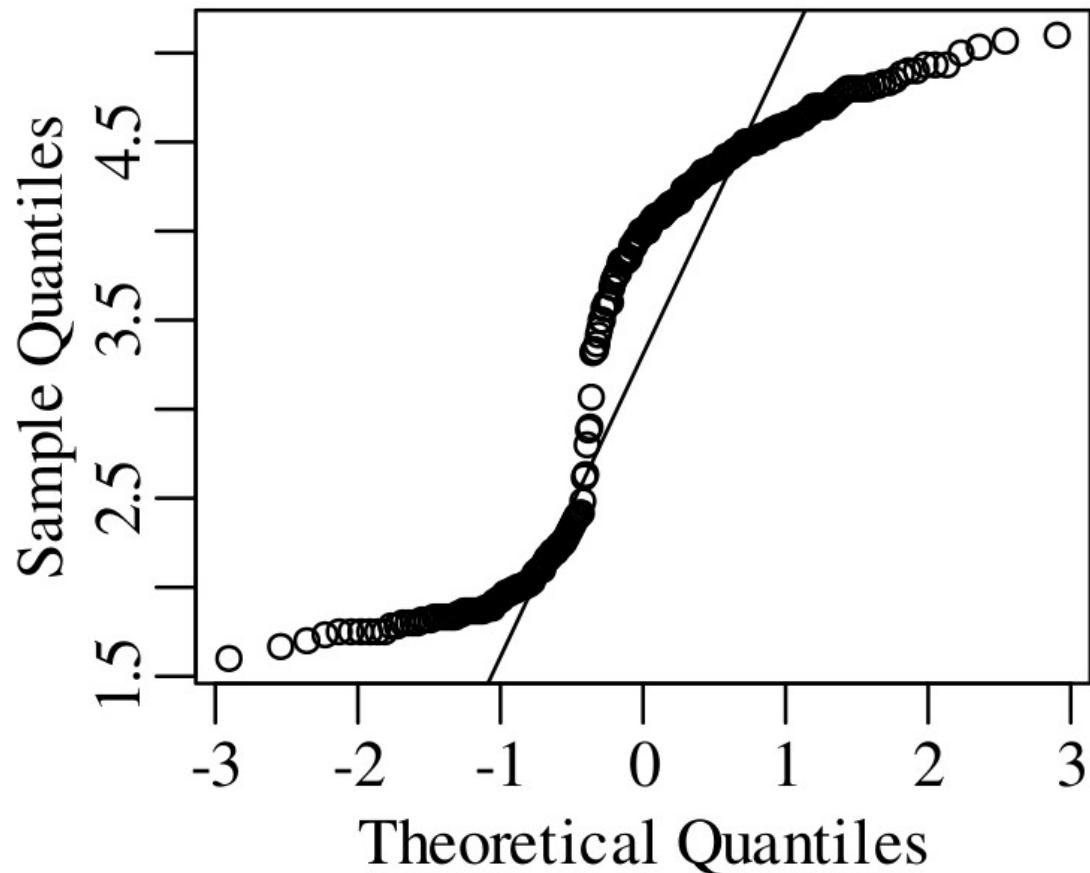
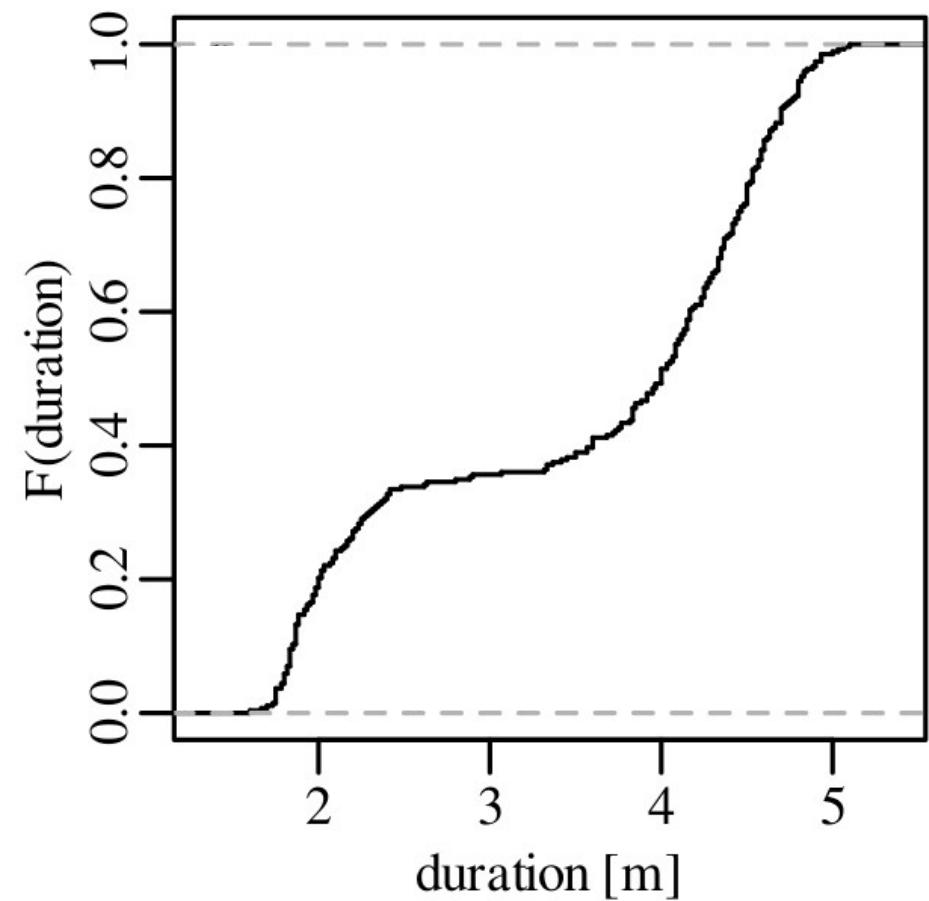
$$s[\bar{x}] = \frac{s[x]}{\sqrt{n}}$$

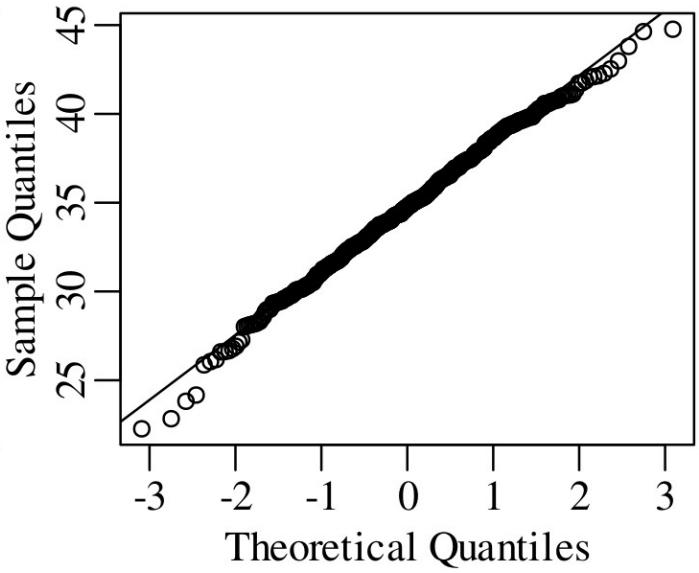
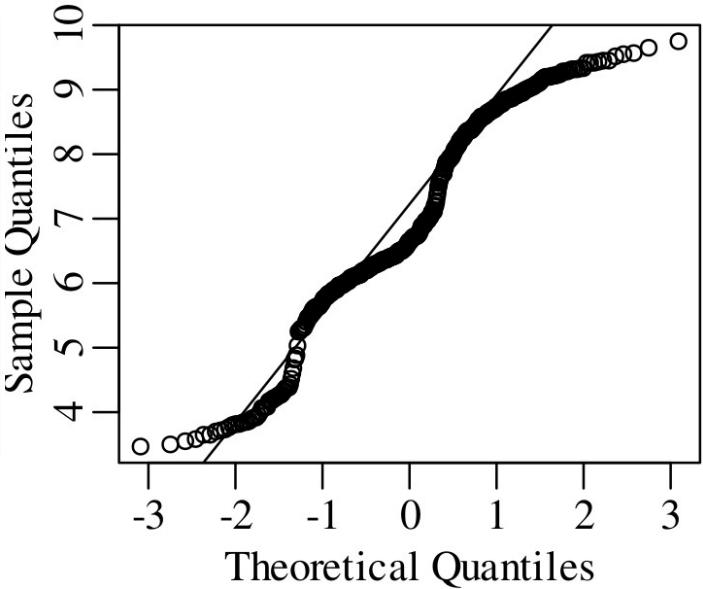
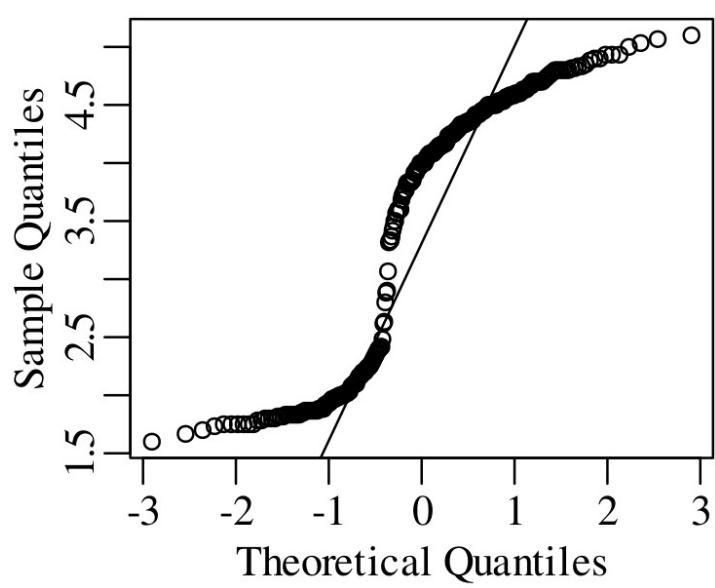


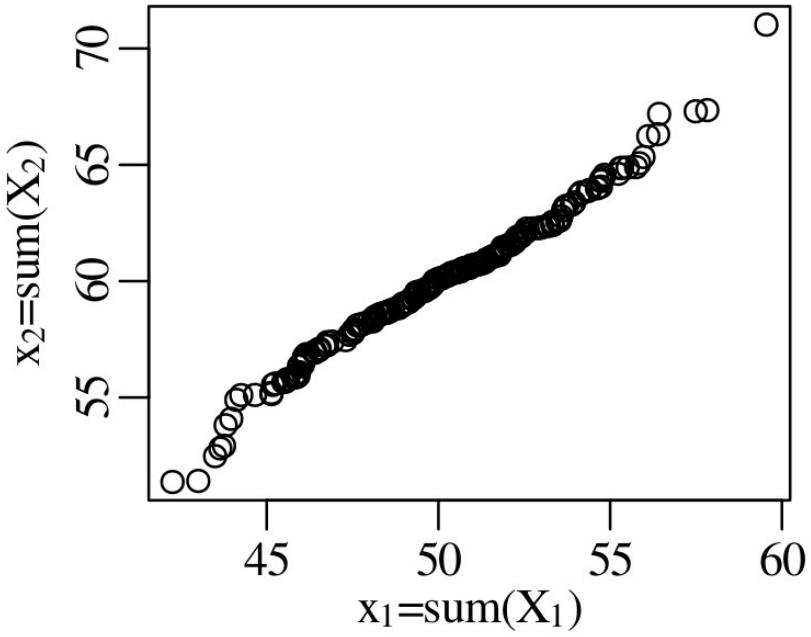
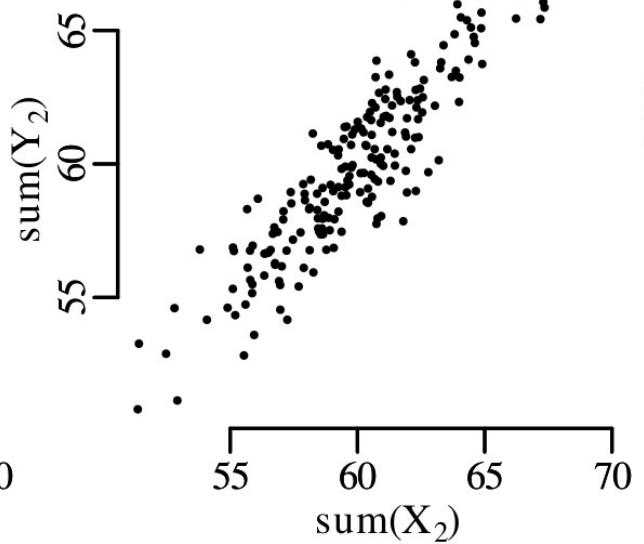
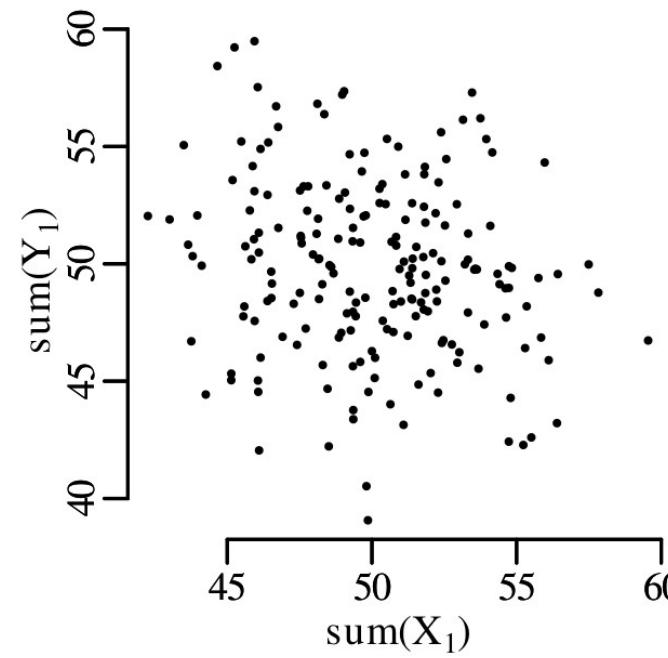
Statistics for geoscientists

Comparing distributions

Q-Q plots







The t-test

coin #	1	2	3	4	5
density (g/cm ³)	19.07	19.09	19.17	19.18	19.31

$$\bar{x} = 19.164$$

density of pure gold = 19.30 g/cm³

coin #	1	2	3	4	5
density (g/cm ³)	19.07	19.09	19.17	19.18	19.31

$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ follows a **standard normal distribution**

$t = \frac{\bar{x} - \mu}{s[\bar{x}]/\sqrt{n}}$ follows a **Student t-distribution**
with $(n - 1)$ **degrees of freedom**

1. H_0 (null hypothesis) $\mu = 19.30$

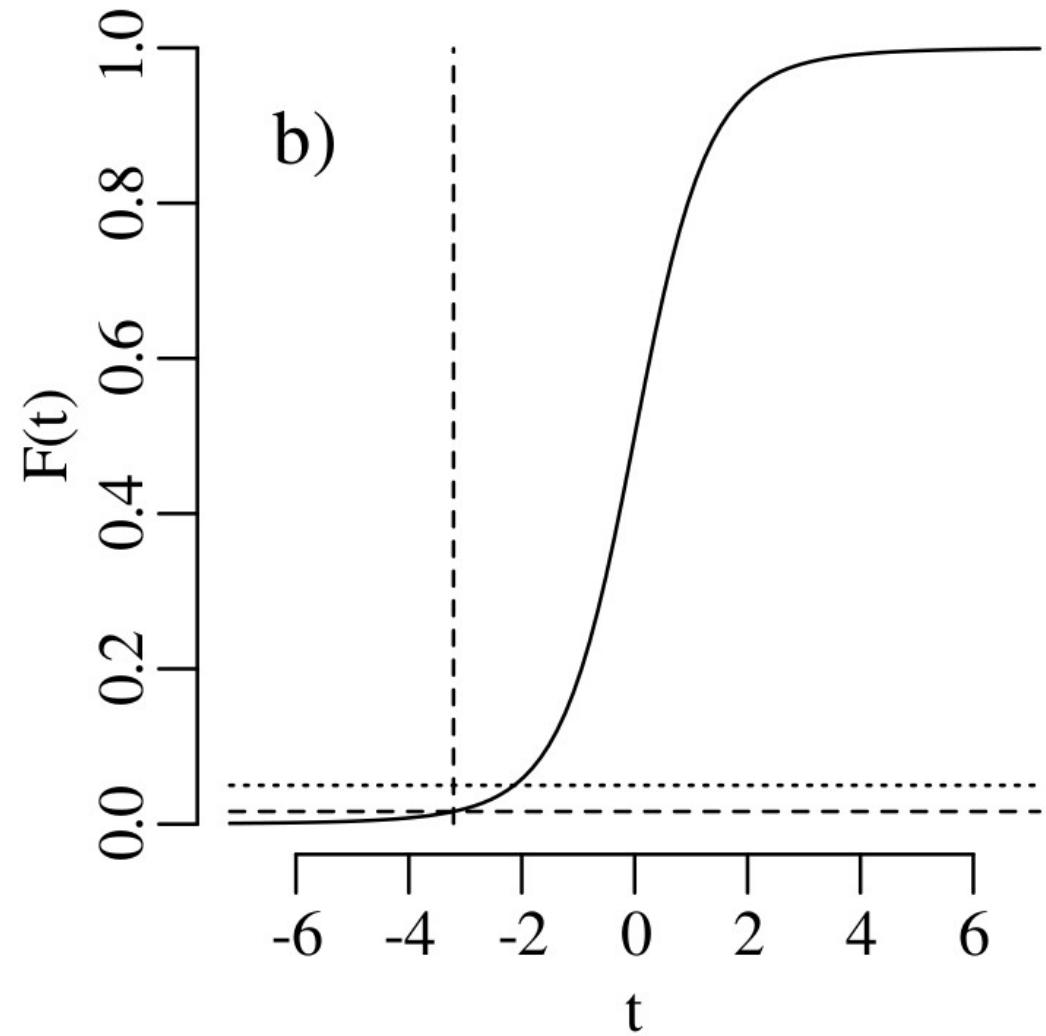
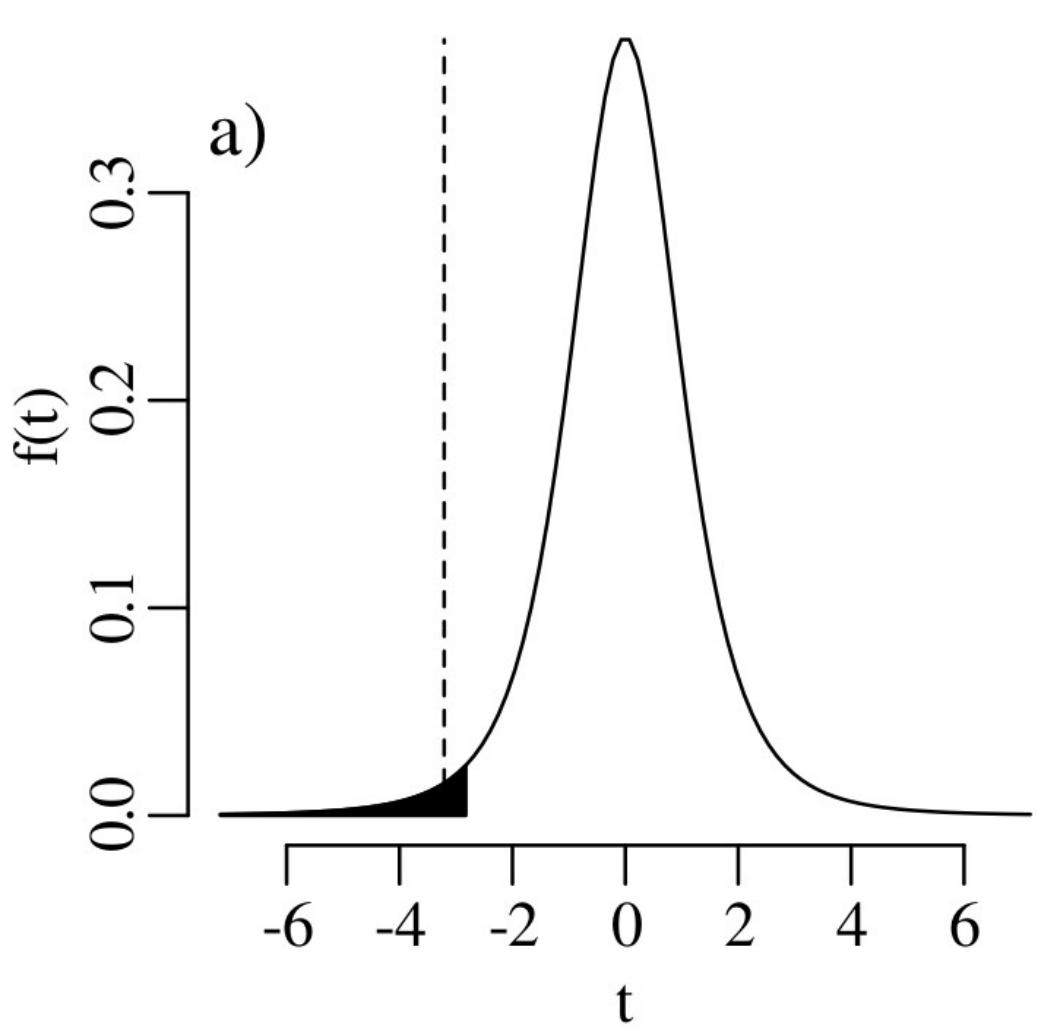
H_a (alternative hypothesis): $\mu < 19.30$

2. $t = \frac{\bar{x} - \mu_0}{s[x]/\sqrt{n}}$ where $\mu_0 = 19.30$ $\bar{x} = 19.164$
 $s[x] = 0.0948$ so that $t = -3.2091$
 $n = 5$

3.	t	-3.70 -3.2091 -2.80 -2.10 -1.50 -0.74 0.00 0.74 1.50 2.10 2.80 3.70
	$P(T \leq t)$	0.01 0.0163 0.025 0.05 0.1 0.25 0.5 0.75 0.9 0.95 0.975 0.99

4. $\alpha = 0.05$

5.	t	-3.70 -3.2091 -2.80 -2.10 -1.50 -0.74 0.00 0.74 1.50 2.10 2.80 3.70
	$P(T \leq t)$	0.01 0.0163 0.025 0.05 0.1 0.25 0.5 0.75 0.9 0.95 0.975 0.99



two sample t-test

coin #	1	2	3	4	5
density (1 st collection)	19.07	19.09	19.17	19.18	19.31
density (2 nd collection)	19.17	19.30	19.31	19.32	

$$\bar{x}_1 = 19.164 \longleftrightarrow \bar{x}_2 = 19.275$$

$$1. \quad H_0 \text{ (null hypothesis)} \quad \mu_1 = \mu_2$$

$$H_a \text{ (alternative hypothesis):} \quad \mu_1 \neq \mu_2$$

$$2. \quad t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{where} \quad s_p = \sqrt{\frac{(n_1 - 1)s[x_1]^2 + (n_2 - 1)s[x_2]^2}{n_1 + n_2 - 2}}$$
$$= -2.014$$

3. Student t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom

t	-3.00	-2.40	-2.014	-1.90	-1.40	-0.71	0.00	0.71	1.40	1.90	2.40	3.00
$P(T \leq t)$	0.01	0.025	0.042	0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.975	0.99
$P(T \geq t)$	0.99	0.975	0.958	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01

4. $\alpha = 0.05$

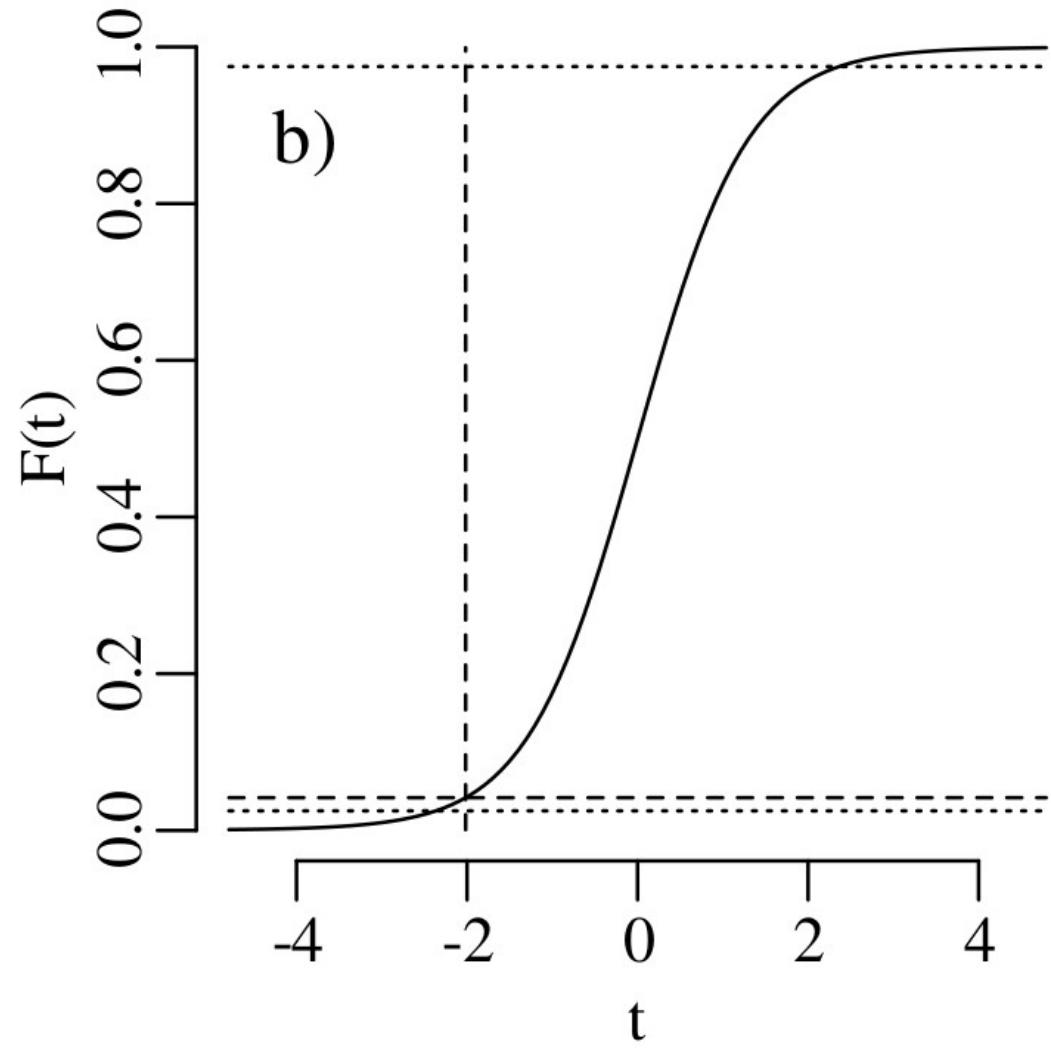
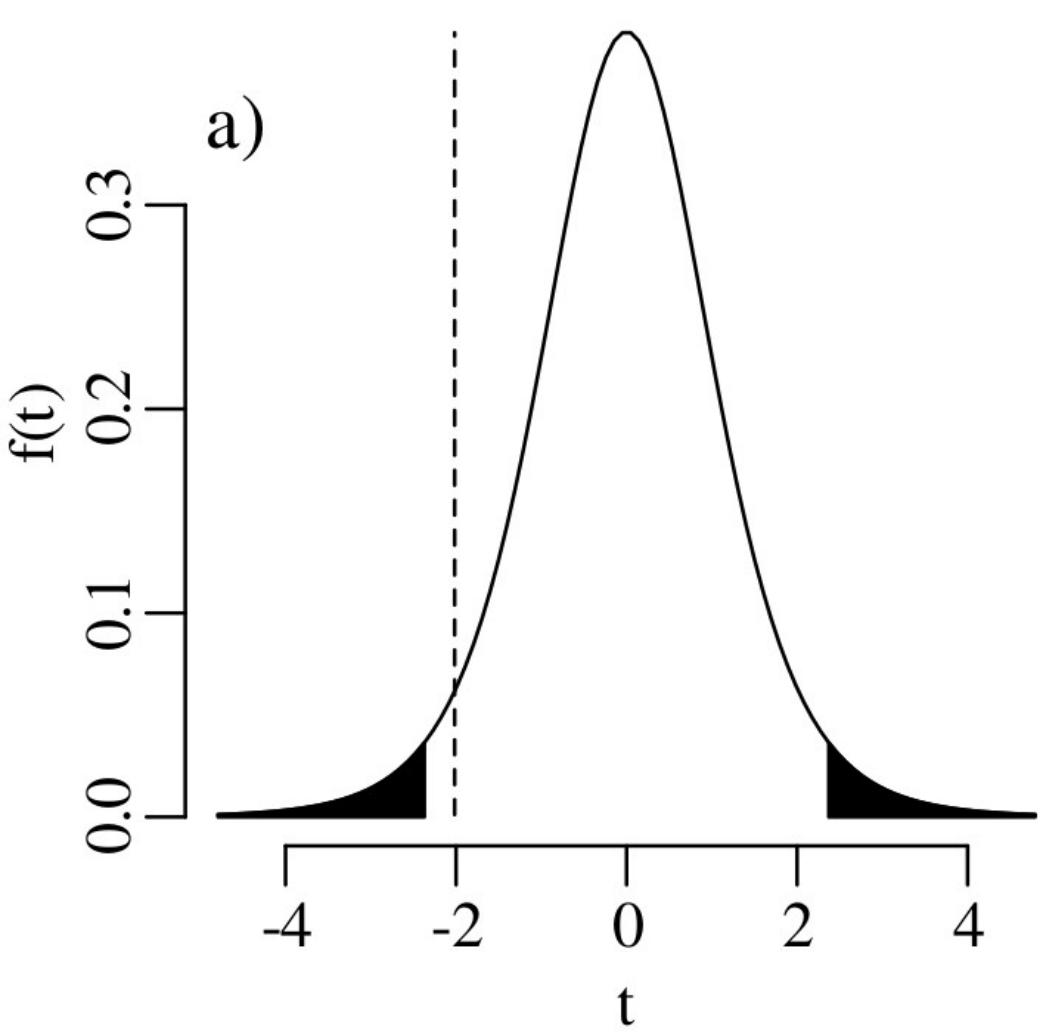
5.

t	-3.00	-2.40	-2.014	-1.90	-1.40	-0.71	0.00	0.71	1.40	1.90	2.40	3.00
$P(T \leq t)$	0.01	0.025	0.042	0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.975	0.99
$P(T \geq t)$	0.99	0.975	0.958	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01

$$R = \{t < -2.40 \text{ and } t > 2.40\}$$

6. $t = -2.014 \notin R$

7. p-value = 0.084 ($= 2 \times 0.042$) > 0.05



Confidence intervals

$$t = \frac{\bar{x} - \mu}{s[x]/\sqrt{n}}$$

$$t_{df,\alpha/2} \leq t \leq t_{df,1-\alpha/2}$$

$$t_{df,\alpha/2} \leq \frac{\bar{x} - \mu}{s[x]/\sqrt{n}} \leq t_{df,1-\alpha/2}$$

$$\bar{x} - t_{df,\alpha/2} \frac{s[x]}{\sqrt{n}} \geq \mu \geq \bar{x} - t_{df,1-\alpha/2} \frac{s[x]}{\sqrt{n}}$$

with $t_{df,1-\alpha/2} = -t_{df,\alpha/2}$

$$\bar{x} + t_{df,\alpha/2} \frac{s[x]}{\sqrt{n}} \leq \mu \leq \bar{x} - t_{df,\alpha/2} \frac{s[x]}{\sqrt{n}}$$

$\rightarrow \mu \in \left\{ \bar{x} \pm t_{df,\alpha/2} \frac{s[x]}{\sqrt{n}} \right\}$

$$\mu \in \left\{ \bar{x} \pm t_{df,\alpha/2} \frac{s[x]}{\sqrt{n}} \right\}$$

coin #	1	2	3	4	5
density (g/cm ³)	19.07	19.09	19.17	19.18	19.31

$$\bar{x} = 19.164$$

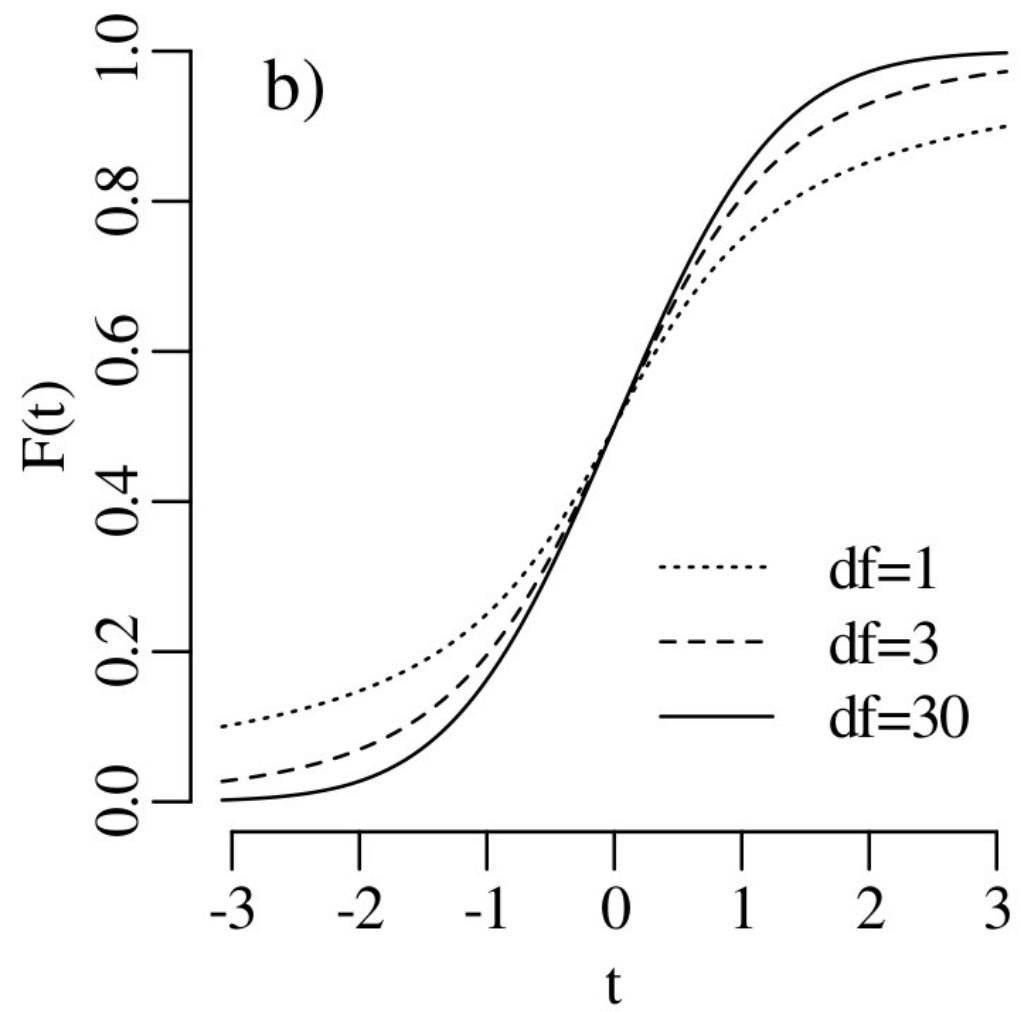
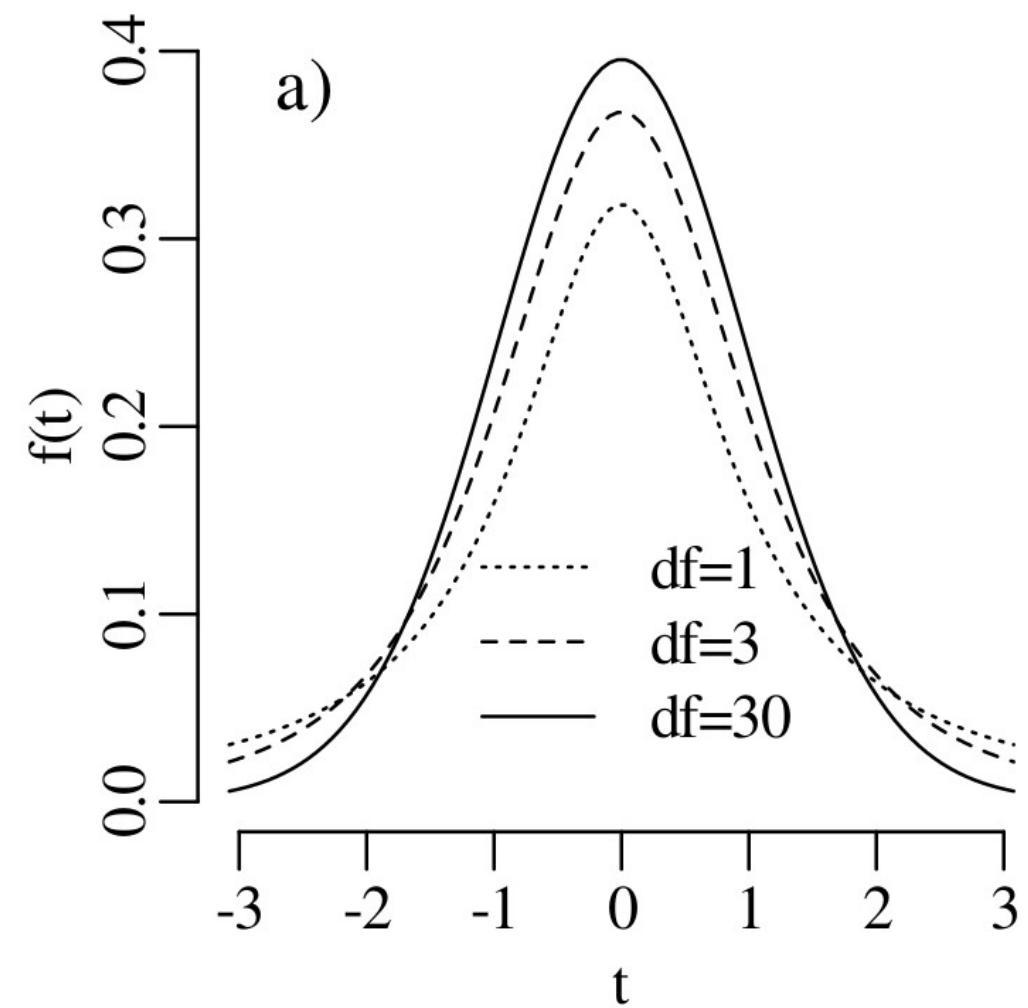
$$s[x] = 0.0948 \qquad \qquad \mu = 19.16 \pm 0.12 \text{ g/cm}^3$$

$$df = 4$$

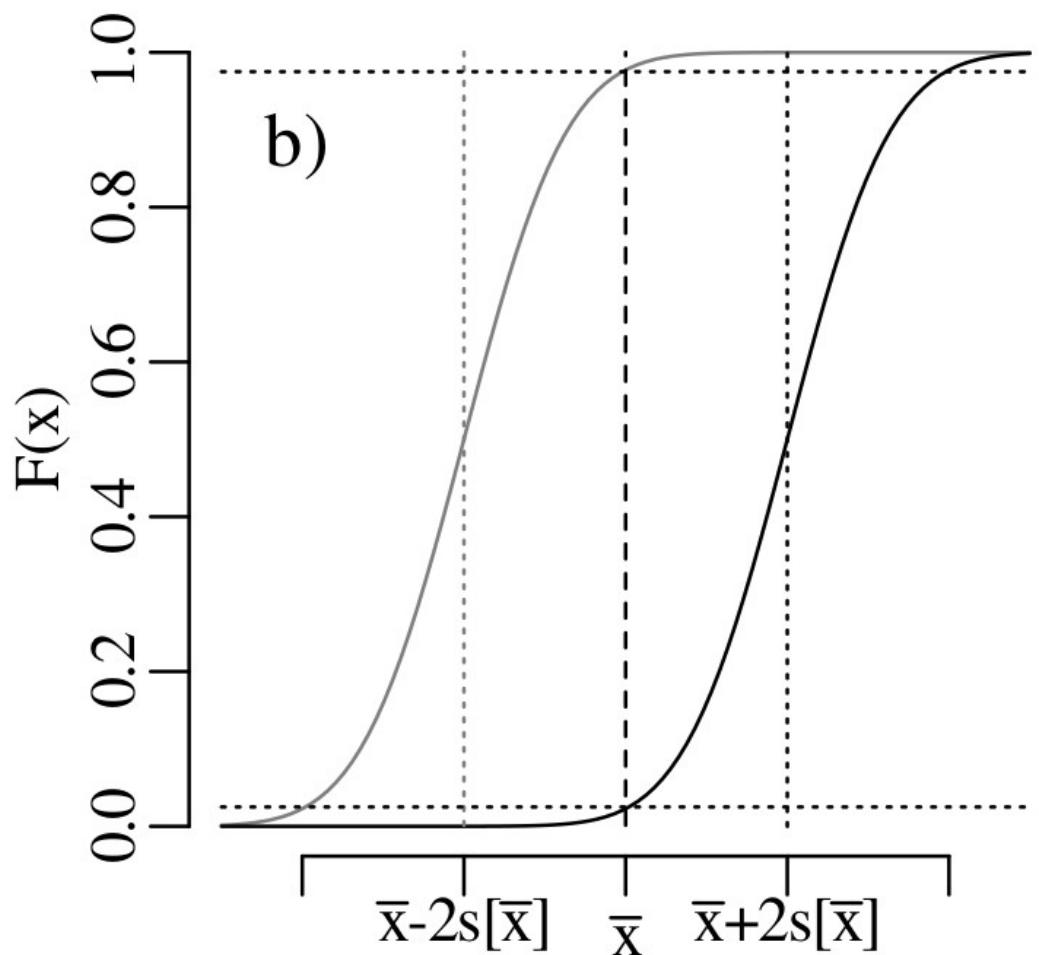
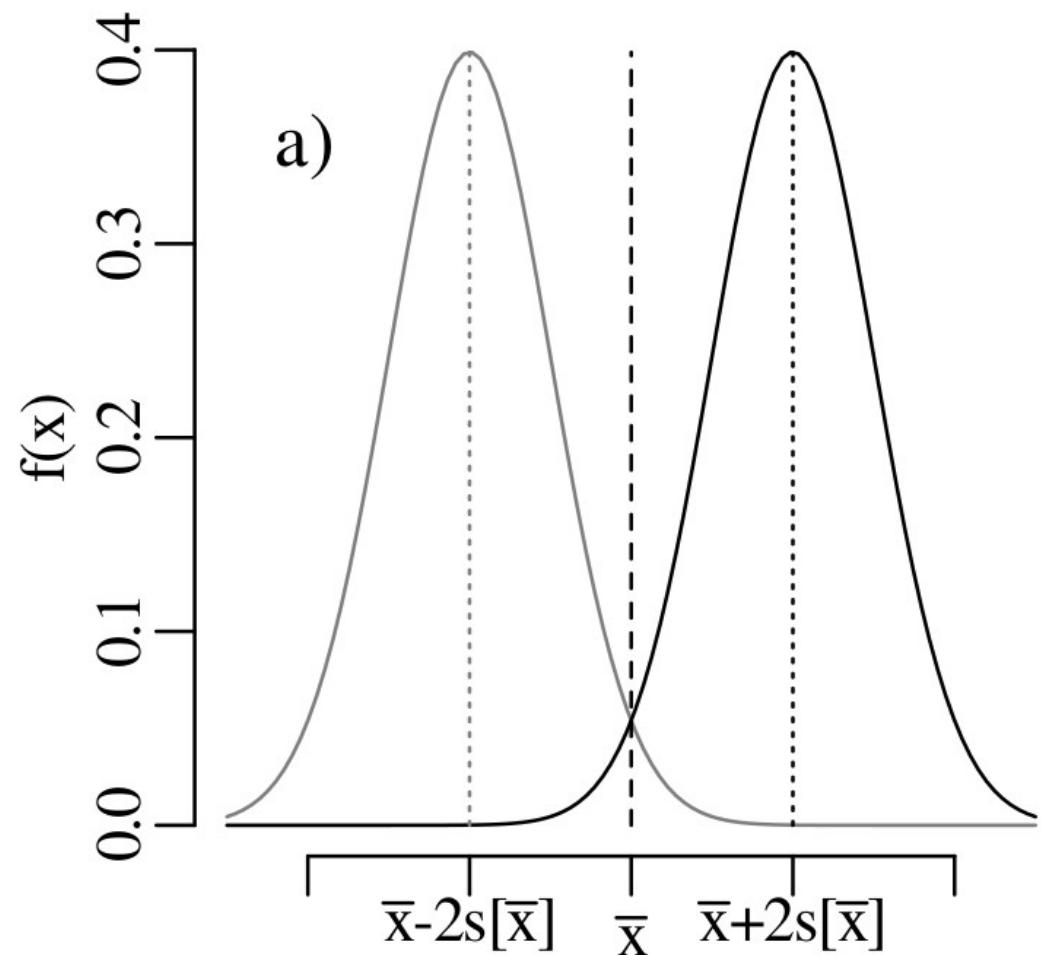
$$t_{4,0.025} = -2.776$$

df	1	2	3	4	5	6	7	8	9	10	30	100	1000
$t_{df,0.975}$	12.710	4.303	3.182	2.776	2.571	2.447	2.365	2.306	2.262	2.228	2.042	1.984	1.962

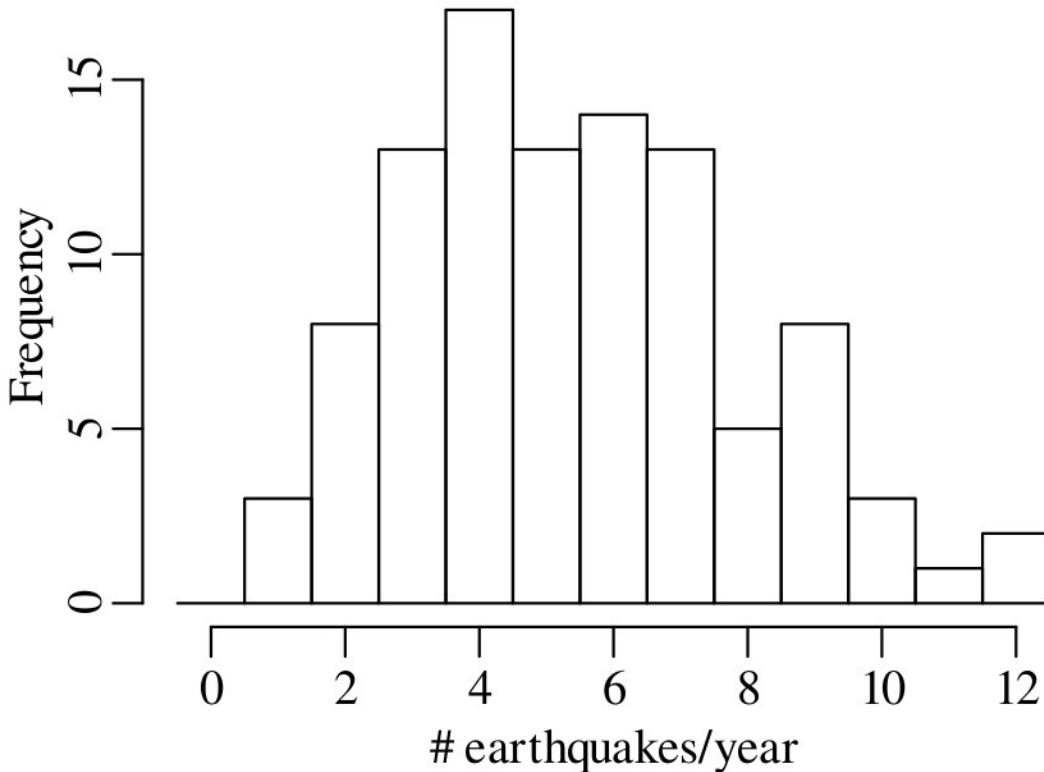
$$\mu \in \left\{ \bar{x} \pm t_{df,\alpha/2} \frac{s[\bar{x}]}{\sqrt{n}} \right\} \xrightarrow{} \mu \in \{ \bar{x} \pm 2s[\bar{x}] \}$$



$$\mu \in \left\{ \bar{x} \pm t_{df,\alpha/2} \frac{s[x]}{\sqrt{n}} \right\} \longrightarrow \mu \in \{ \bar{x} \pm 2s[\bar{x}] \}$$



The χ^2 -test



mean: 5.43

standard deviation: 2.50

variance: 6.24

number of earthquakes per year	0 1 2 3 4 5 6 7 8 9 10 11 12
number of years	0 3 8 13 17 13 14 13 5 8 3 1 2

observed

number of earthquakes per year	0	1	2	3	4	5	6	7	8	9	10	11	12
number of years	0	3	8	13	17	13	14	13	5	8	3	1	2

number of earthquakes per year	≤ 2	3	4	5	6	7	8	9	≥ 10
number of years	11	13	17	13	14	13	5	8	6

expected

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

number of earthquakes per year (k)	≤ 2	3	4	5	6	7	8	9	≥ 10
$N \times P(k \lambda = 5.43)$	9.28	11.7	15.9	17.2	15.6	12.1	8.22	4.96	5.02

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

1. H_0 (**null hypothesis**): the earthquake data follow a Poisson distribution

H_a (**alternative hypothesis**): the earthquake data do not follow a Poisson distribution

$$2. \chi^2 = \frac{(11 - 9.28)^2}{9.28} + \frac{(13 - 11.7)^2}{11.7} + \frac{(17 - 15.9)^2}{15.9} + \frac{(13 - 17.2)^2}{17.2} + \\ \frac{(14 - 15.6)^2}{15.6} + \frac{(13 - 12.1)^2}{12.1} + \frac{(5 - 8.22)^2}{8.22} + \frac{(8 - 4.96)^2}{4.96} + \frac{(6 - 5.02)^2}{5.02} = 5.14$$

3. χ^2 distribution with $n - 2$ degrees of freedom

χ^2	1.24	1.69	2.17	2.83	4.25	5.14	6.35	9.04	12.0	14.1	16.0	18.5
$P(X \leq \chi^2)$	0.01	0.025	0.05	0.1	0.25	0.36	0.5	0.75	0.9	0.95	0.975	0.99
$P(X \geq \chi^2)$	0.99	0.975	0.95	0.9	0.75	0.743	0.5	0.25	0.1	0.05	0.025	0.01

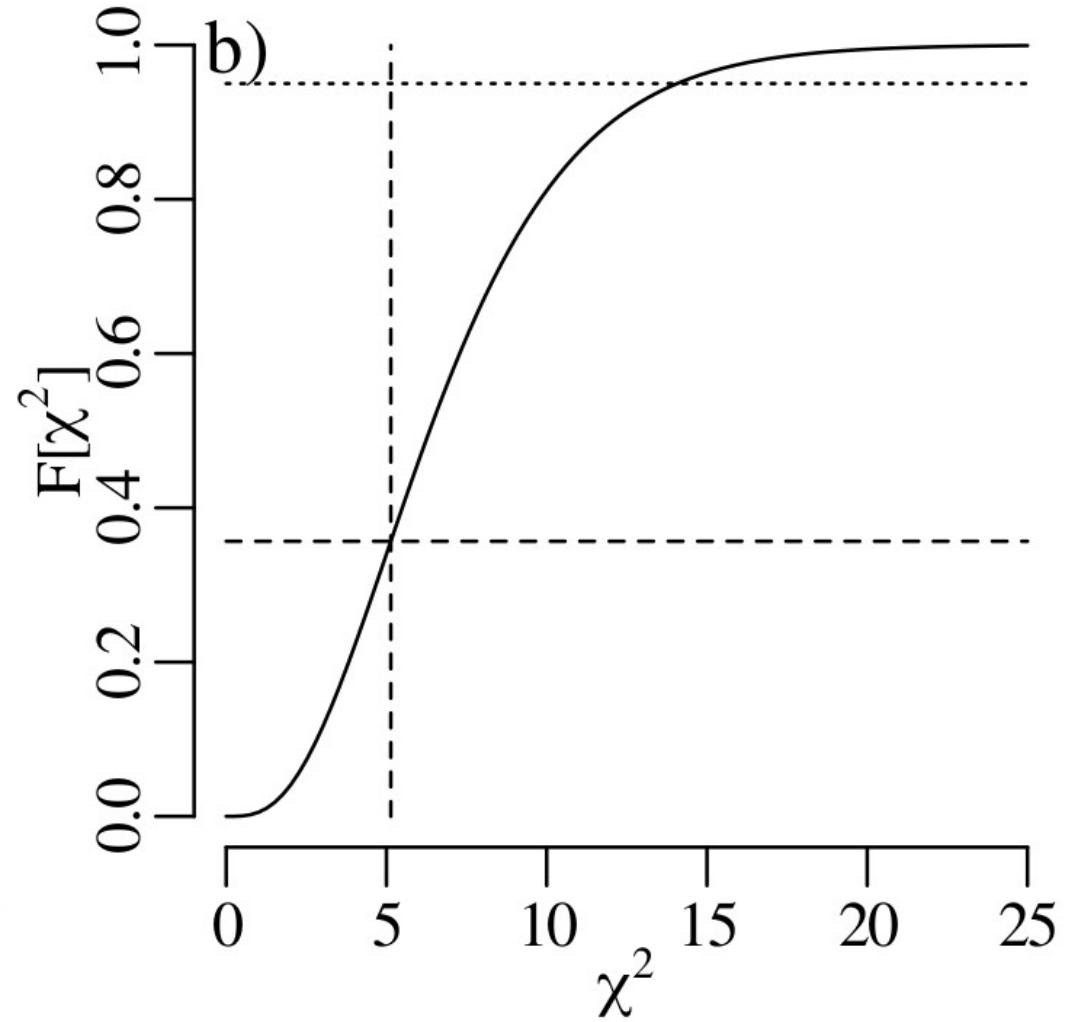
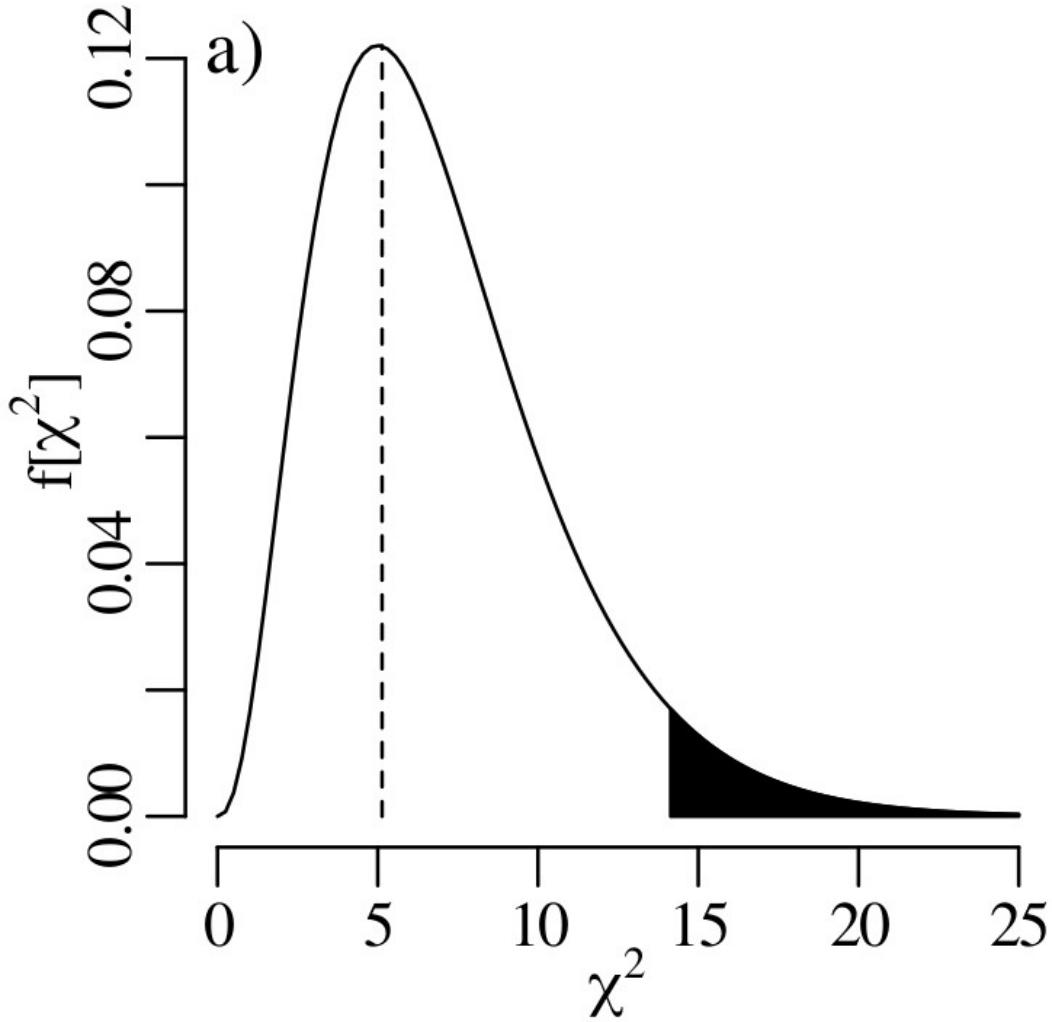
4. $\alpha = 0.05$

5.	χ^2	1.24 1.69 2.17 2.83 4.25 5.14 6.35 9.04 12.0	14.1 16.0 18.5
	$P(X \geq \chi^2)$	0.99 0.975 0.95 0.9 0.75 0.743 0.5 0.25 0.1	0.05 0.025 0.01

$$R = \{\chi^2 \geq 14.1\}$$

6. $\chi^2 = 5.14 \notin R$

7. p-value = 0.743 > 0.05



Comparing two or more samples

lithology	granite	basalt	gneiss	quartzite
sample A	10	5	6	20
sample B	25	12	10	35

lithology	granite	basalt	gneiss	quartzite	row sum
sample A	10	5	6	20	41
sample B	25	12	10	35	82
column sum	35	17	16	55	123

expected counts of bin $(i, j) = \frac{(\text{sum of row } i) \times (\text{sum of column } j)}{(\text{sum of all the cells})}$

lithology	granite	basalt	gneiss	quartzite
sample A	11.7	5.67	5.33	18.3
sample B	23.3	11.30	10.70	36.7

1. H_0 (**null hypothesis**): samples A and B have the same composition

H_a (**alternative hypothesis**): samples A and B do not have the same composition

$$2. \chi^2 = \frac{(10 - 11.7)^2}{11.7} + \frac{(5 - 5.67)^2}{5.67} + \frac{(6 - 5.33)^2}{5.33} + \frac{(20 - 18.3)^2}{18.3} + \\ \frac{(25 - 23.3)^2}{23.3} + \frac{(12 - 11.30)^2}{11.30} + \frac{(10 - 10.70)^2}{10.70} + \frac{(35 - 36.7)^2}{36.7} = 0.86$$

3. χ^2 -distribution with $(2-1) \times (4-1) = 3$ degrees of freedom

χ^2	0.115	0.216	0.352	0.584	0.86	1.21	2.37	4.11	6.25	7.81	9.35	11.3
$P(X \leq \chi^2)$	0.01	0.025	0.05	0.1	0.157	0.25	0.5	0.75	0.9	0.95	0.975	0.99
$P(X \geq \chi^2)$	0.99	0.975	0.95	0.9	0.843	0.75	0.5	0.25	0.1	0.05	0.025	0.01

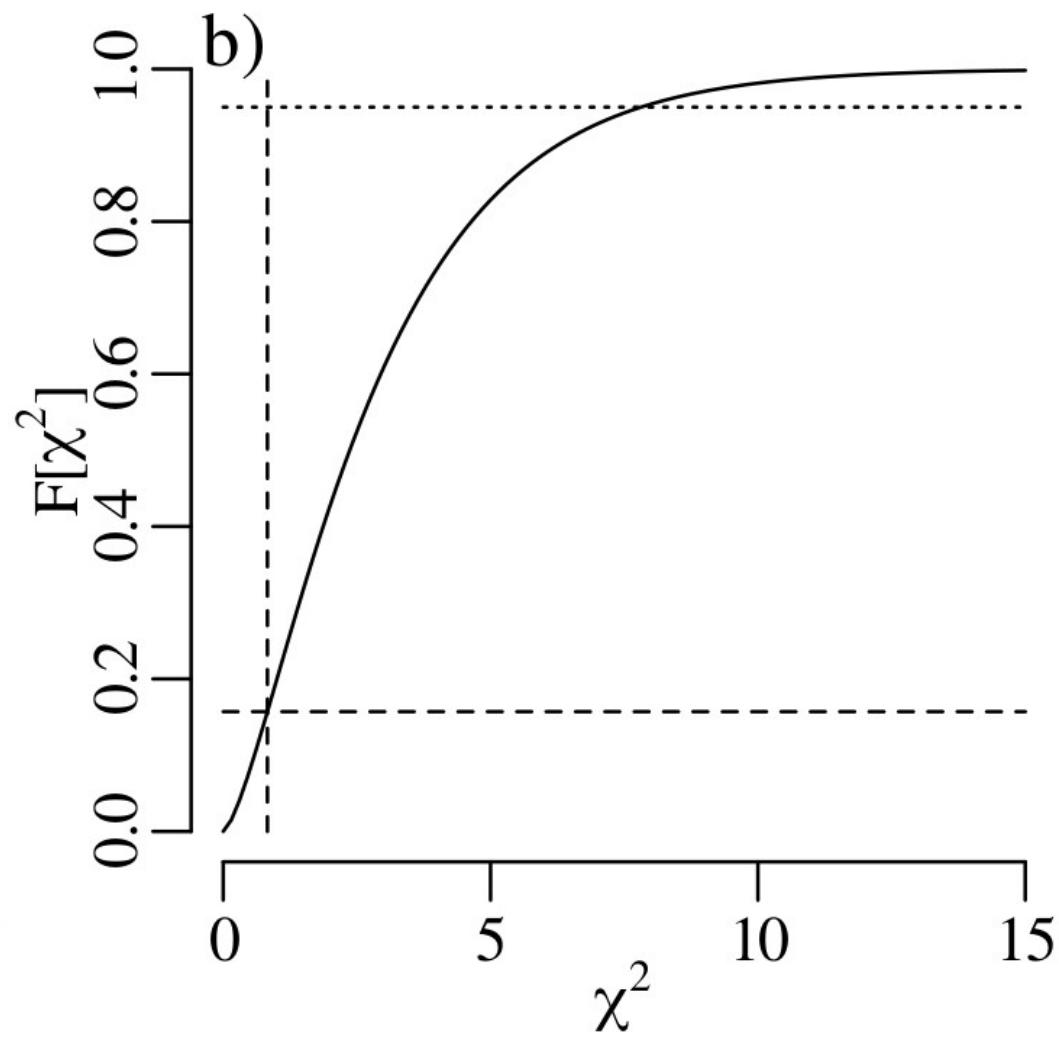
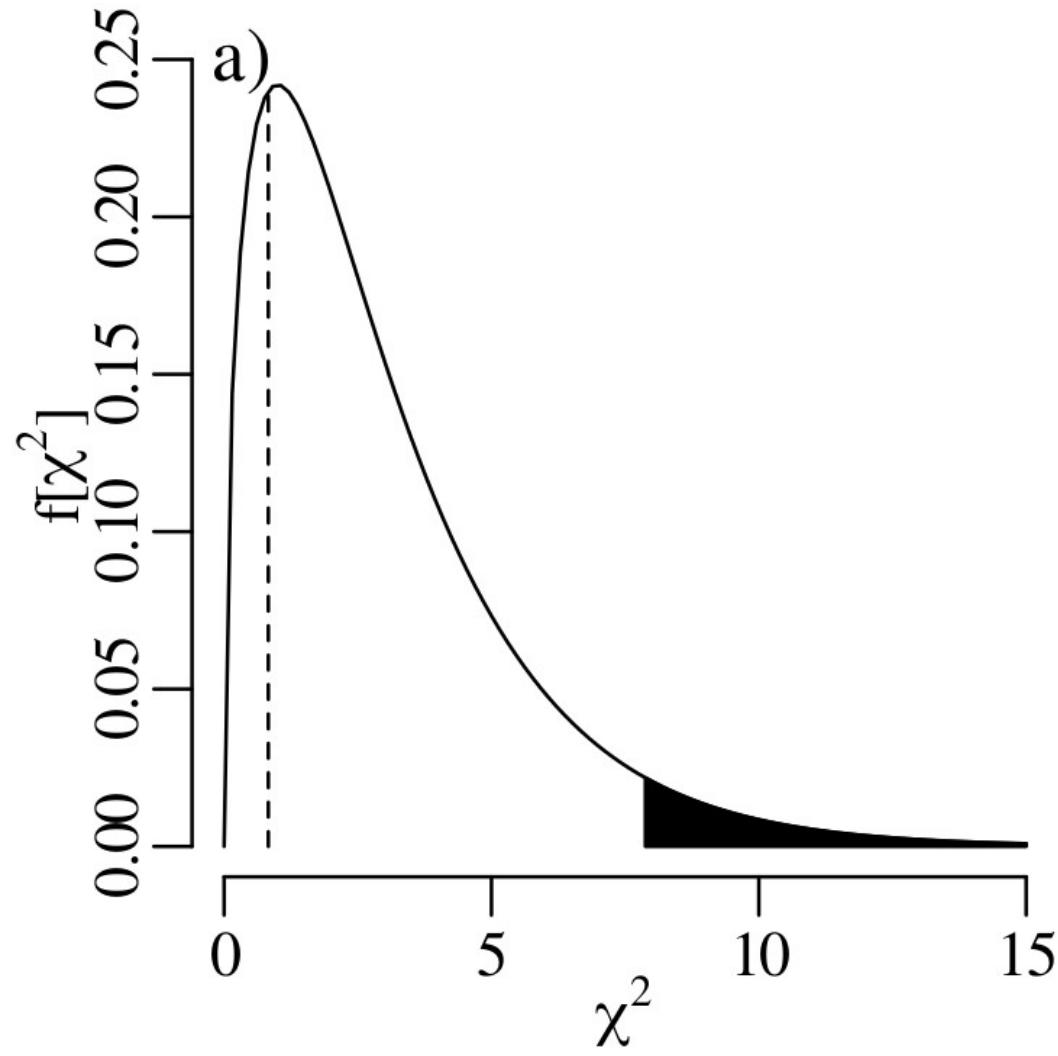
4. $\alpha = 0.05$

5.	χ^2	0.115 0.216 0.352 0.584 0.86 1.21 2.37 4.11 6.25 7.81 9.35 11.3
	$P(X \geq \chi^2)$	0.99 0.975 0.95 0.9 0.843 0.75 0.5 0.25 0.1 0.05 0.025 0.01

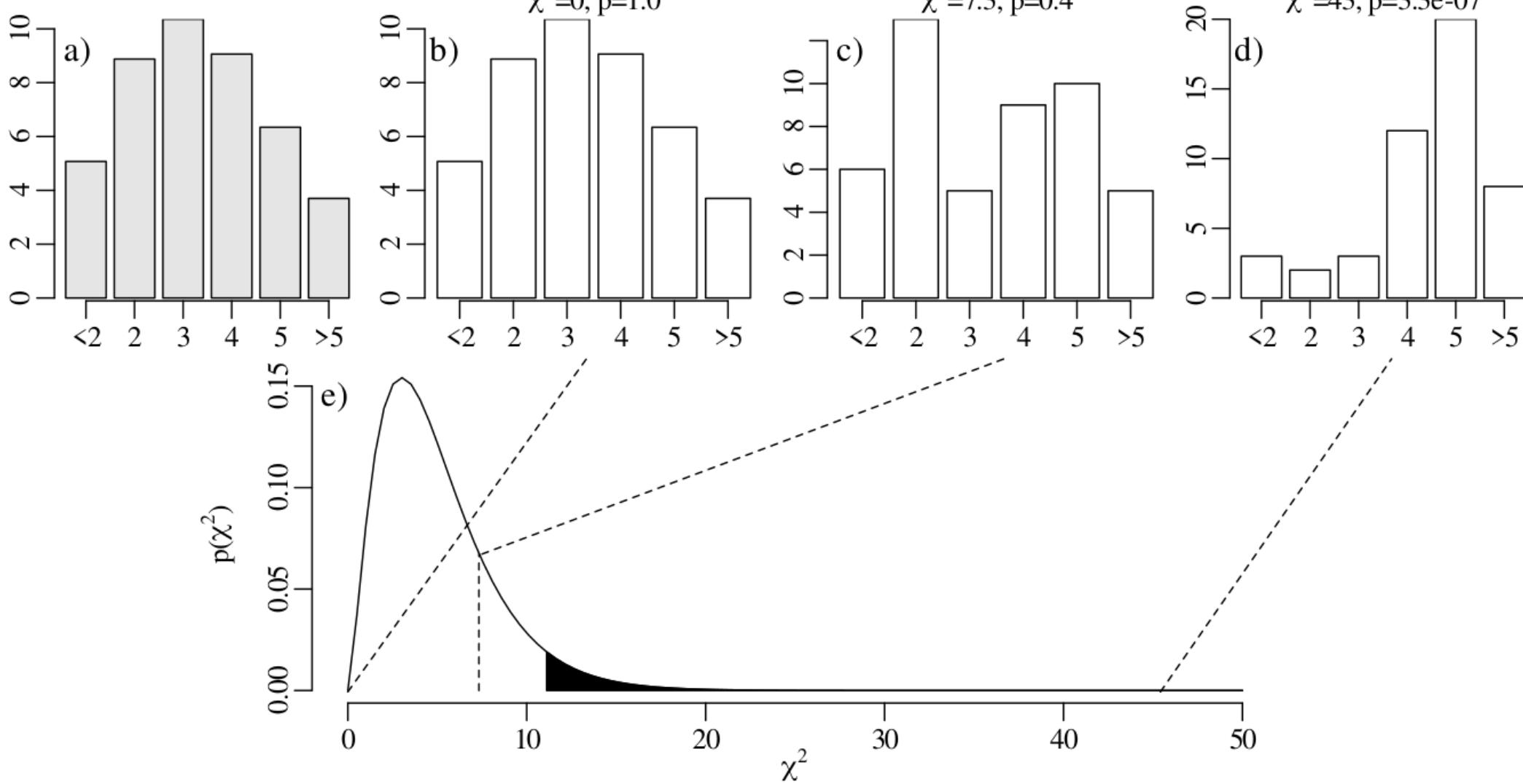
$$R = \{\chi^2 > 7.81\}$$

6. $\chi^2 = 0.86 \notin R$

7. p-value = 0.843 > 0.05



Cherry picking (Type-I errors revisited)



Effect size (Type-II errors revisited)

lithology	quartz	plagioclase	alkali feldspar	lithics
sample A	29544	14424	13706	47864
sample B	29454	14788	13948	47311

lithology	quartz	plagioclase	alkali feldspar	lithics	row sum
sample A	29544	14424	13706	47864	105538
sample B	29454	14788	13948	47311	105501
column sum	58998	29212	27654	95175	211039

1. H_0 (**null hypothesis**): samples A and B have identical compositions

H_a (**alternative hypothesis**): samples A and B have different compositions

lithology	quartz	plagioclase	alkali feldspar	lithics	
sample A	29504	14609	13829	47596	$\chi^2 = 10.0$
sample B	29494	14603	13825	47579	

3. χ^2 -distribution 3 degrees of freedom

χ^2	0.115	0.216	0.352	0.584	1.21	2.37	4.11	6.25	7.81	9.35	10.0	11.3
$P(X \leq \chi^2)$	0.01	0.025	0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.975	0.9814	0.99
$P(X \geq \chi^2)$	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.0186	0.010

4. $\alpha = 0.05$

5.

χ^2	0.115	0.216	0.352	0.584	1.21	2.37	4.11	6.25	7.81	9.35	10.0	11.3
$P(X \geq \chi^2)$	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.0186	0.010

$$R = \{\chi^2 > 7.81\}$$

6. $\chi^2 = 10.0 \in R$

7. p-value = 0.0186 < α

depends on the **confidence level** α

H_0 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_0 is false

2. Sample size

depends on the **confidence level** α

H_0 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_0 is false
 2. Sample size

effect size

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$w = \sqrt{\sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i}}$$

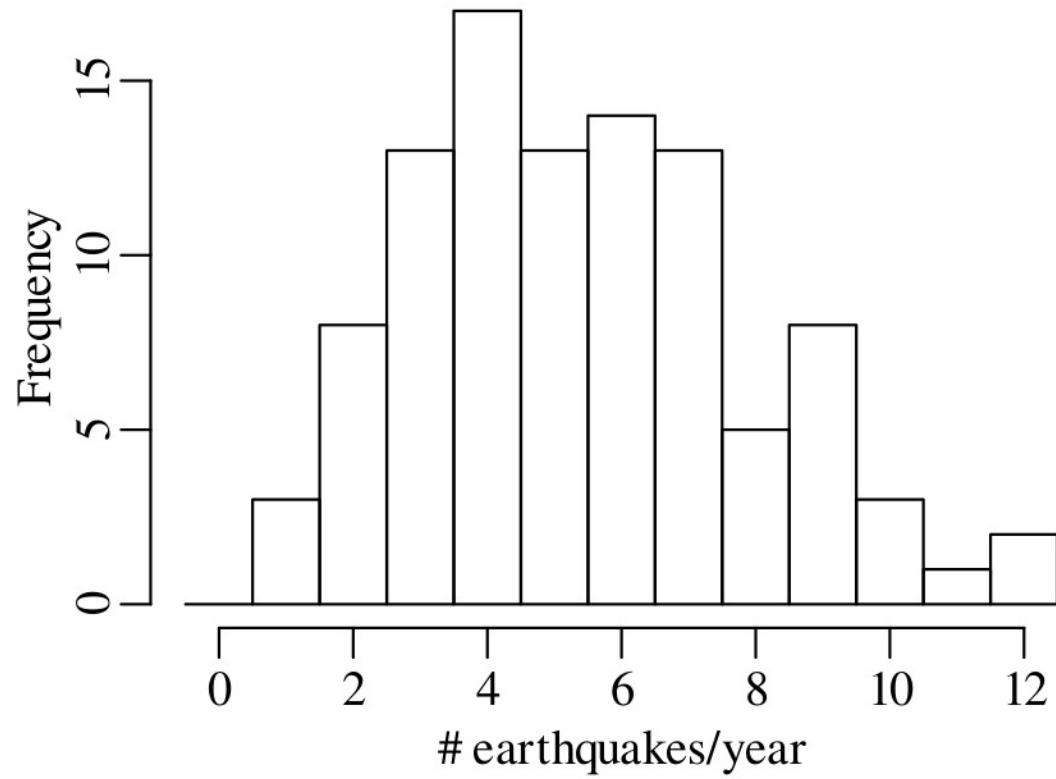
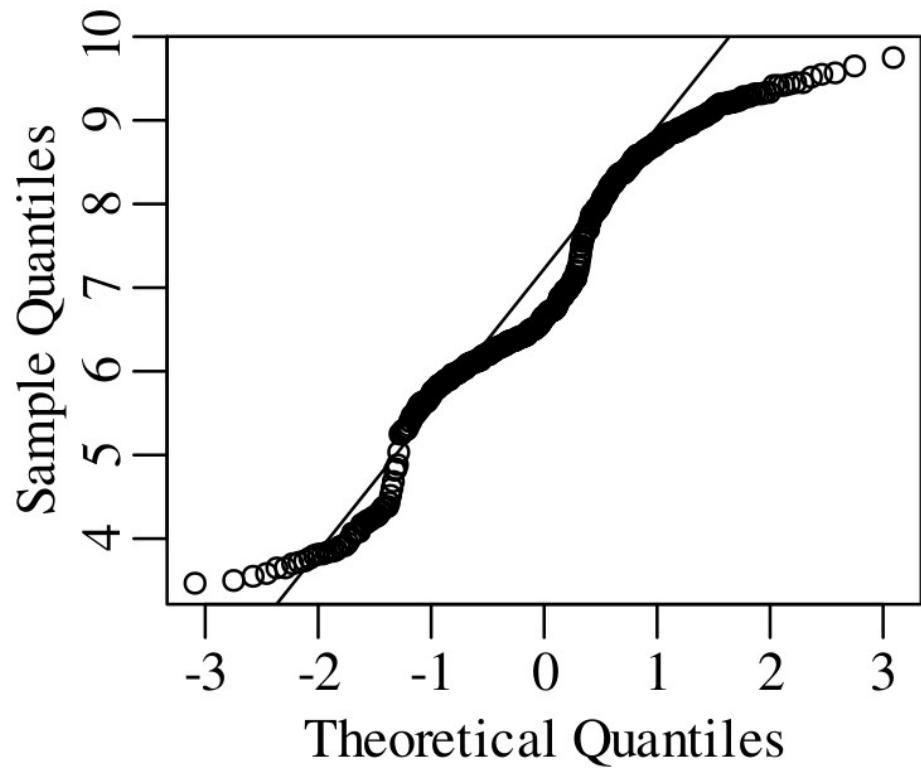
where $o_i = O_i/N$ and $e_i = E_i/N$

with $N = \sum_i^n O_i = \sum_i^n E_i$

effect size	small	medium	large
w	0.1	0.3	0.5

effect size		small	medium	large	
	w	0.1	0.3	0.5	
$o =$	lithology	quartz	plagioclase	alkali feldspar	lithics
	sample A	0.1400	0.06835	0.06495	0.2268
	sample B	0.1396	0.07007	0.06609	0.2242
$e =$	lithology	quartz	plagioclase	alkali feldspar	lithics
	sample A	0.1398	0.06922	0.06553	0.2255
	sample B	0.1398	0.06920	0.06551	0.2255

$$w = 0.00688$$



Non-parametric tests

1. Wilcoxon test

coin #	1	2	3	4	5
density (1 st collection)	19.07	19.09	19.17	19.18	19.31
density (2 nd collection)	19.17	19.30	19.31	19.32	

sample	1	2	3	4	5
A	A_1	A_2	A_3	A_4	A_5
B	B_1	B_2	B_3	B_4	

sample	1	2	3	4	5
A	A_1	A_2	A_3	A_4	A_5
B	B_1	B_2	B_3	B_4	

suppose that

$$A_4 < A_1 < A_2 < B_3 < A_5 < B_1 < B_4 < A_3 < B_2$$

then

rank	1	2	3	4	5	6	7	8	9
value	A_4	A_1	A_2	B_3	A_5	B_1	B_4	A_3	B_2

$$W = 4 + 6 + 7 + 9 = 26$$

e.g. $W = 10$

arrangement 1:	B_1	B_2	B_3	B_4	A_1	A_2	A_3	A_4	A_5
arrangement 2:	B_2	B_1	B_3	B_4	A_1	A_2	A_3	A_4	A_5
arrangement 2:	B_1	B_2	B_3	B_4	A_2	A_1	A_3	A_4	A_5
etc.									

$$P(W = 10 | n_A = 5, n_B = 4) = \frac{4!5!}{9!} = 0.00794$$

coin #	1	2	3	4	5
density (1 st collection)	19.07	19.09	19.17	19.18	19.31
density (2nd collection)	19.17	19.30	19.31	19.32	

1. H_0 (null hypothesis) $\text{median}(\text{sample 1}) = \text{median}(\text{sample 2})$

H_a (alternative hypothesis): $\text{median}(\text{sample 1}) \neq \text{median}(\text{sample 2})$

rank	1	2	3.5	3.5	5	6	7.5	7.5	9
density	19.07	19.09	19.17	19.17	19.18	19.30	19.31	19.31	19.32

$$W = 3.5 + 6 + 7.5 + 9 = 26$$

3.

W	10	11	12	14	16	19	22	24	26	27	28
$P(w \leq W)$	0.01	0.025	0.05	0.1	0.25	0.5	0.75	0.9	<i>0.95</i>	0.975	0.99
$P(w \geq W)$	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	<i>0.05</i>	0.025	0.010

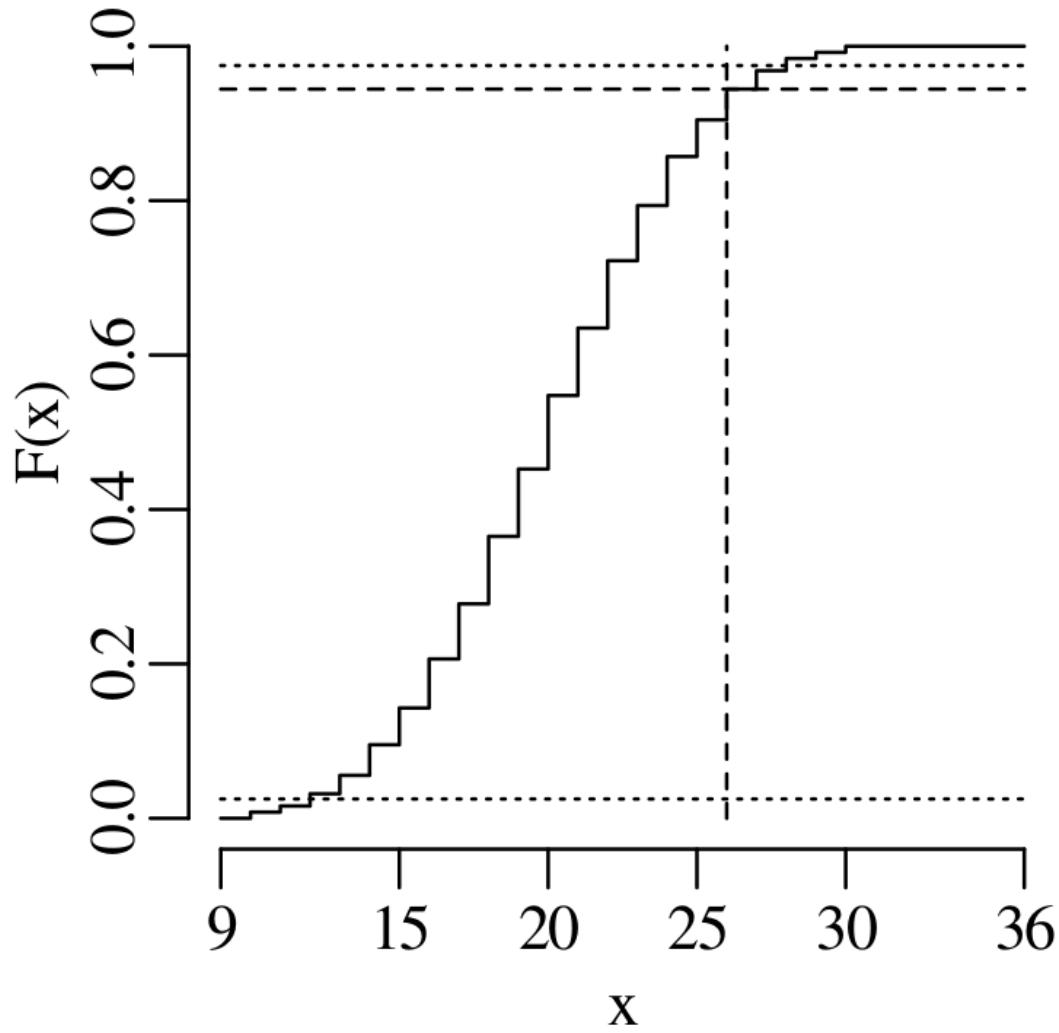
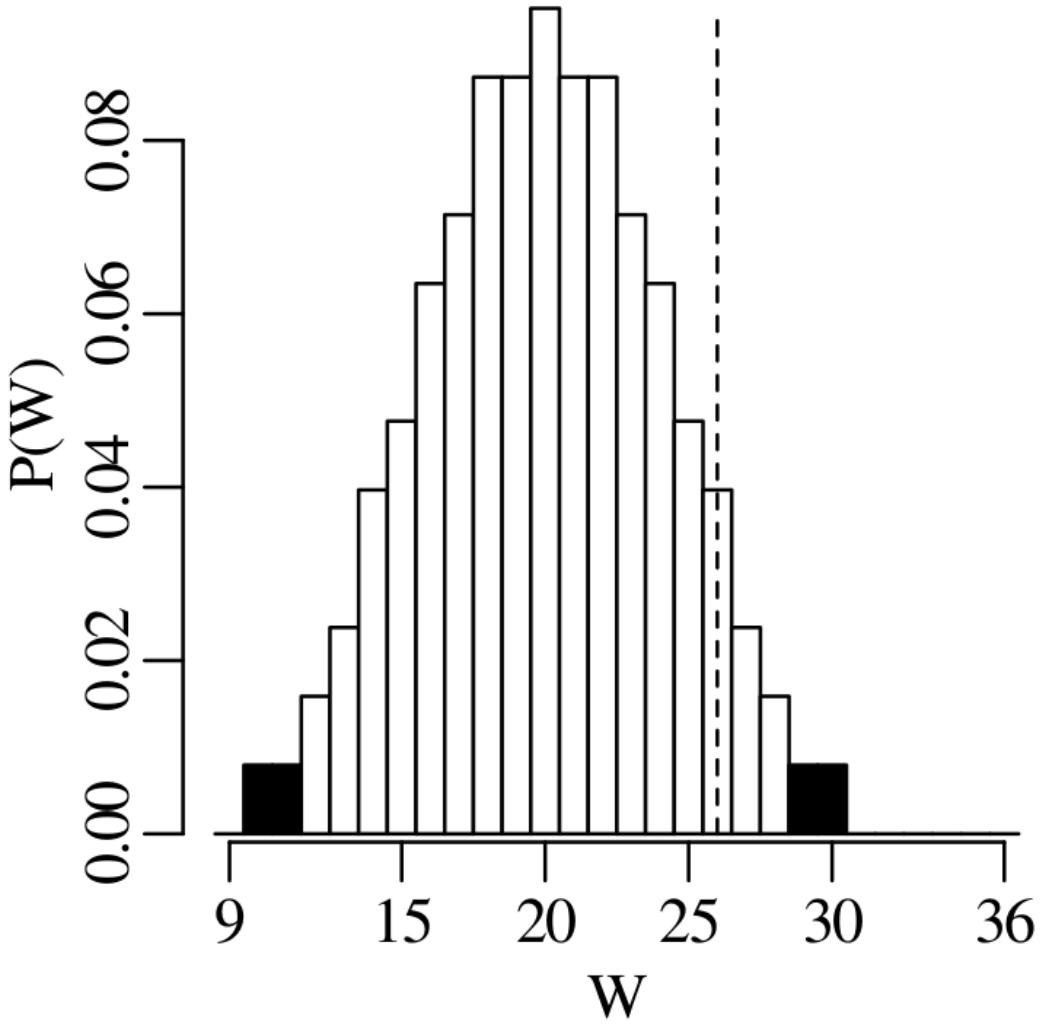
4. $\alpha = 0.05$

5.	W	10	11	12	14	16	19	22	24	26	27	28
	$P(w \leq W)$	0.01	0.025	0.05	0.1	0.25	0.5	0.75	0.9	0.95	0.975	0.99
	$P(w \geq W)$	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01

$$R = \{ W = 10 \text{ and } W > 27 \}$$

6. $W = 26 \notin R$

7. p-value = 0.10 > α



2. Kolmogorov-Smirnov test

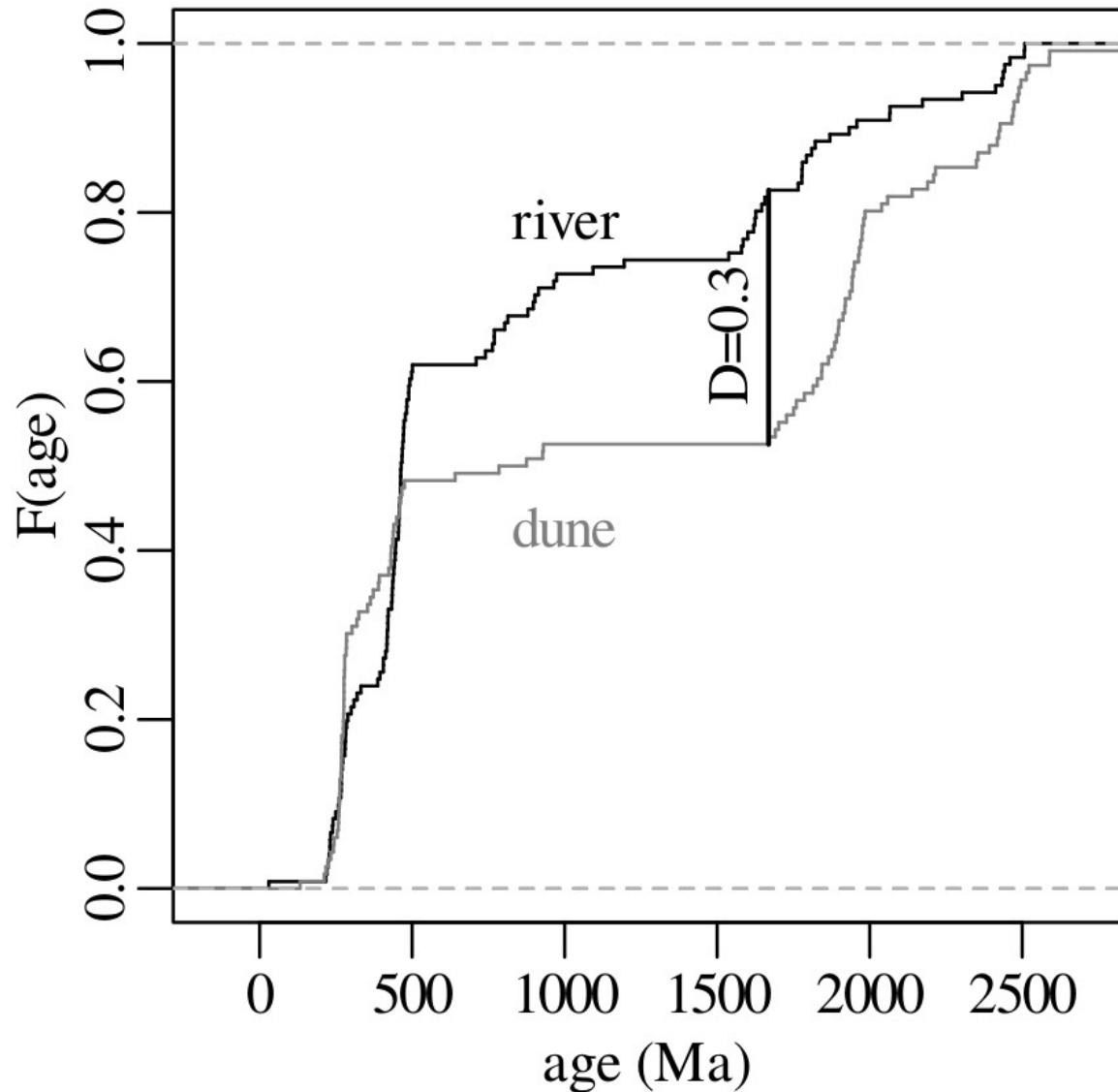
Given

$$x = \{x_1, x_2, \dots, x_n\}$$

and

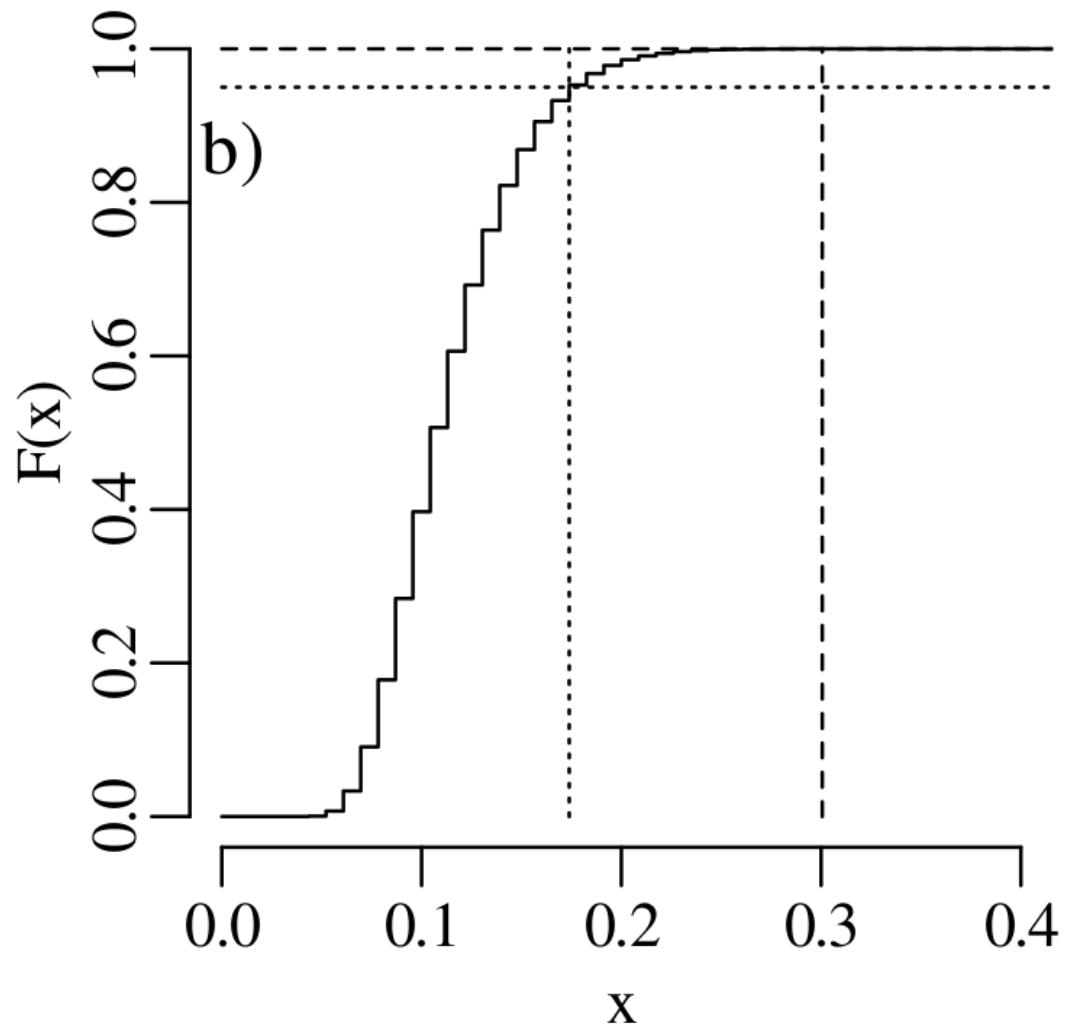
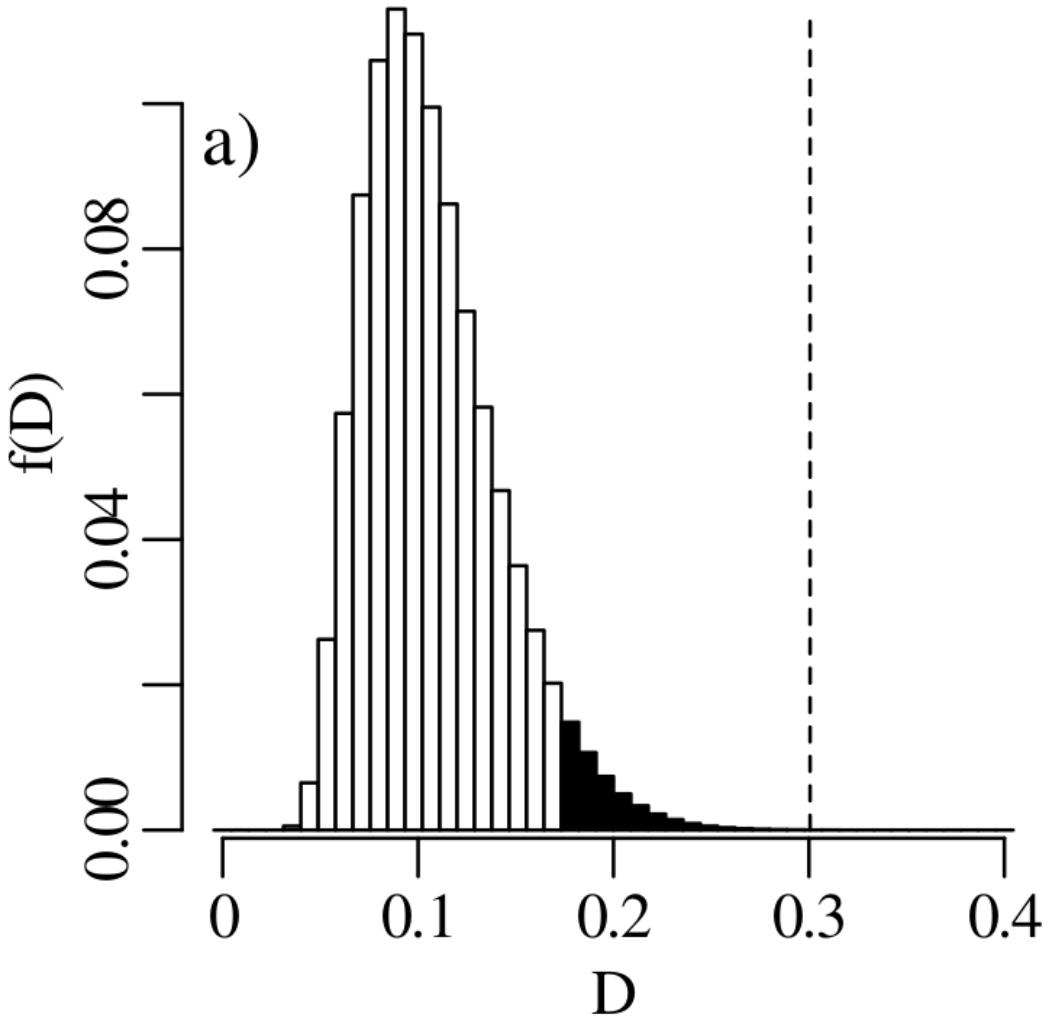
$$y = \{y_1, y_2, \dots, y_m\}$$

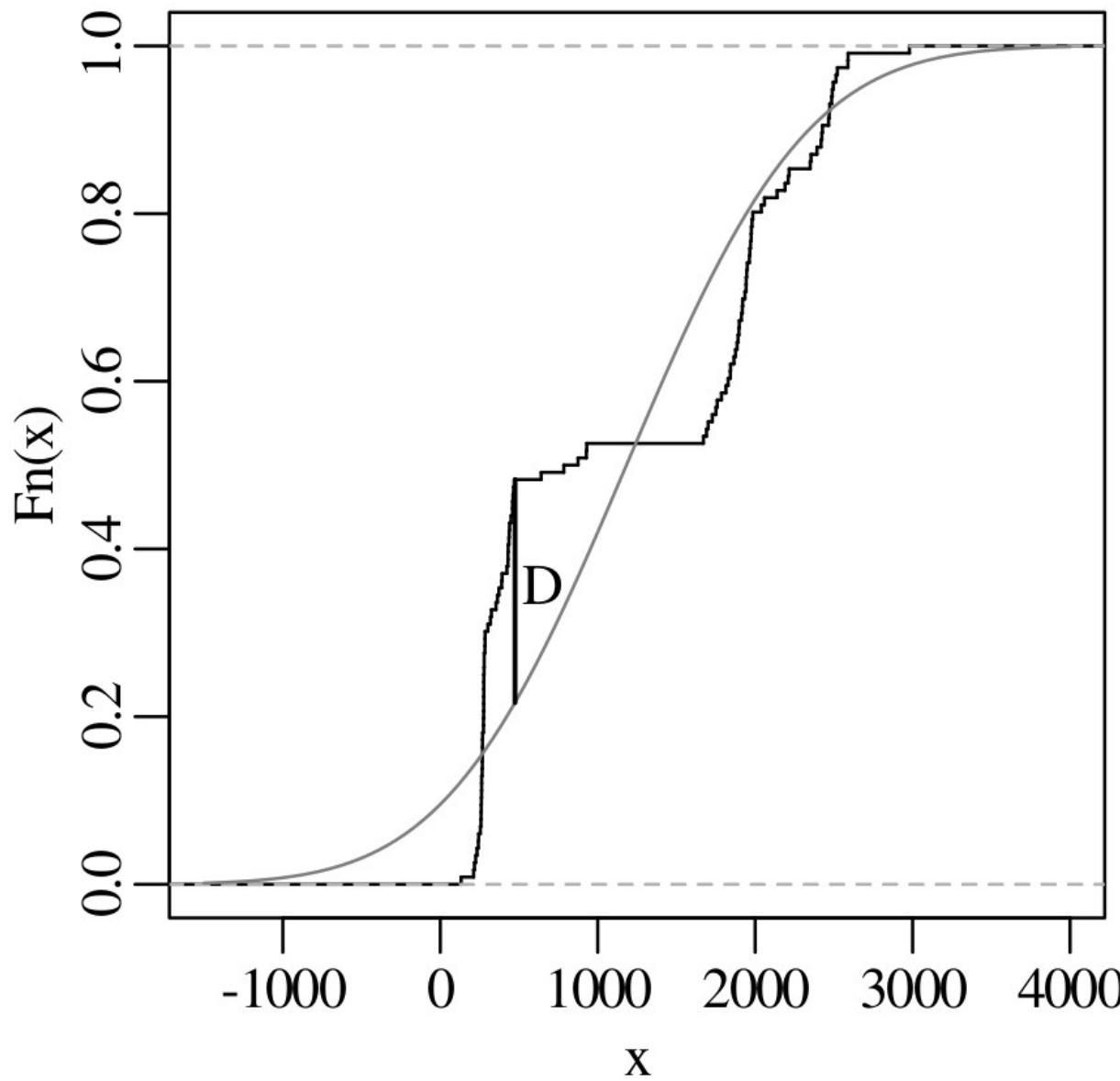
$$D = \max_z |F_x(z) - F_y(z)|$$



- H_0 (**null hypothesis**): samples 1 and 2 were drawn from the same distribution
 H_a (**alternative hypothesis**): samples 1 and 2 were drawn from different distributions
- $D = 0.30$
- | D | 0.061 | 0.061 | 0.070 | 0.078 | 0.087 | 0.104 | 0.130 | 0.157 | 0.174 | 0.191 | 0.209 | 0.301 |
|---------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------------------|
| $P(d \leq D)$ | 0.01 | 0.025 | 0.05 | 0.1 | 0.25 | 0.5 | 0.75 | 0.9 | 0.95 | 0.975 | 0.99 | 0.99996 |
| $P(d \geq D)$ | 0.99 | 0.975 | 0.95 | 0.9 | 0.75 | 0.5 | 0.25 | 0.1 | 0.05 | 0.025 | 0.010 | 4.5×10^{-5} |
- $\alpha = 0.05$
- | D | 0.061 | 0.061 | 0.070 | 0.078 | 0.087 | 0.104 | 0.130 | 0.157 | 0.174 | 0.191 | 0.209 | 0.301 |
|---------------|-------|-------|-------|-------|-------|-------|-------|-------|--------------|--------------|--------------|----------------------|
| $P(d \geq D)$ | 0.99 | 0.975 | 0.95 | 0.9 | 0.75 | 0.5 | 0.25 | 0.1 | 0.05 | 0.025 | 0.010 | 4.5×10^{-5} |

 $R = \{D > 0.174\}$
- $D = 0.301 \in R$
- p-value = $4.5 \times 10^{-5} < \alpha$





Statistics for geoscientists

An introduction to R