Meeting 14
Classification and decision

LINKÖPING UNIVERSITY

What is classification?

Common is that courses in statistics comprise hypothesis testing of

- the value of a specific parameter (one population)
- the equality in value of two parameters (two populations)

This is part of the classic approach to statistical inference (point estimation, interval estimation and hypothesis testing of one or two or several parameters).

What if the general question is whether an object belongs to a specific population or if it belongs to another population?

At first we assume that we have measured the object with the measurement in form of a real number *x*, say.

If the potential populations are all characterised by a probability density (or mass) function depending on the value of a parameter, θ (one- or multidimensional), but with the same functional form we could use the approach of testing a forwarded value of this parameter:

H_0: $\theta = \theta_0$
H_1: $\theta \neq \theta_0$

The Bayes factor will then be

$$B = \frac{f\left(x \middle| \theta_0\right)}{\int_{\theta \neq \theta_0} f\left(x \middle| \theta\right) p(\theta) d\theta}$$

and we can multiply the Bayes factor with the prior odds for H_0 to obtain the posterior odds and subsequently the posterior probability of H_0 .

But if that’s not the case?

The measurement *x* may be distributed completely differently in the potential populations.

Let c_i denote the *i*th population (class) to which the object may belong and let i_0 be the subscript of the class addressed by H_0.

The hypotheses may be formulated as

$$H_0 : \text{object} \in c_{i_0}$$

$$H_1 : \text{object} \notin c_{i_0} \Leftrightarrow \text{object} \in \bigcup_{i \neq i_0} c_i$$

The expression $\bigcup_{i \neq i_0} c_i$ says here that the object may belong to *any* of the classes

that are different from c_{i_0} but it is assumed to belong to <u>*one*</u> class only.

Now, if the object belongs to class c_i the measurement value x is assumed to have a probability distribution that depends on (is specific for) that class.

$$F(x) = F_i\left(x \middle| c_i\right)$$

The probability density function (or probability mass function) is of course also specific for each class

$$f(x) = f_i\left(x \middle| c_i\right)$$

The Bayes factor can then be written

$$B = \frac{f_{i_0}\left(x \middle| c_{i_0}\right)}{\sum_{i\neq i_0} f_i\left(x \middle| c_i\right)\cdot \underbrace{\Pr\left(\text{object}\in c_i \,\middle|\, \text{object}\in \bigcup_{i\neq i_0} c_i\right)}_{\text{…but these values may be difficult to assign/obtain}}}$$

The simple case: Two classes

$$H_0 : \text{object} \in c_0$$

$$H_1 : \text{object} \in c_1$$

The Bayes factor (with measurement value *x*) is then

$$B = \frac{f_0\left(x \middle| c_0\right)}{f_1\left(x \middle| c_1\right)}$$

and note that the functional forms in the numerator and the denominator may be completely different

It may be the case that f_0 or f_1 (or both) depends on the value of an unknown parameter θ, but it is not the value of θ that would separate the classes.

A more general form of the Bayes factor is then

$$B = \frac{\int_{\theta \in \Theta_0} f_0\left(x \middle| c_0, \theta\right) p_0(\theta) d\theta}{\int_{\theta \in \Theta_1} f_1\left(x \middle| c_1, \theta\right) p_1(\theta) d\theta} = \frac{f_{0*}\left(x \middle| c_0\right)}{f_{1*}\left(x \middle| c_1\right)}$$

where Θ_0 is the parameter space for θ under H_0, Θ_1 is the parameter space for θ under H_1, and p_0 and p_1 are the prior densities for θ under respective hypothesis.

The functions f_{0*} and f_{1*} are the densities of the *predictive* distributions of x.
(*Note!* Not posterior predictive)

Instead of integrating with the prior densities, we could plug in “ordinary” point estimates of θ (one per class)

$$B = \frac{f_0\left(x \middle| c_0, \hat{\theta}^{(c_0)}\right)}{f_1\left(x \middle| c_1, \hat{\theta}^{(c_1)}\right)}$$

When would this be reasonable?

An example

Is a (seized) coin forged?

$\boldsymbol{H_0}$: The ten-krona piece is forged – belongs to class c_0
$\boldsymbol{H_1}$: The ten-krona piece is genuine – belongs to class c_1

What could be measurements/observations?

- Observed coinages, stripes on the edge, etc.
- Weight
- Measured dimensions: thickness, diameter
- Physico-chemical measurements: conductivity, elemental composition, magnetic resonance, …

Specifications from Riksbanken (*The Swedish Central Bank*):

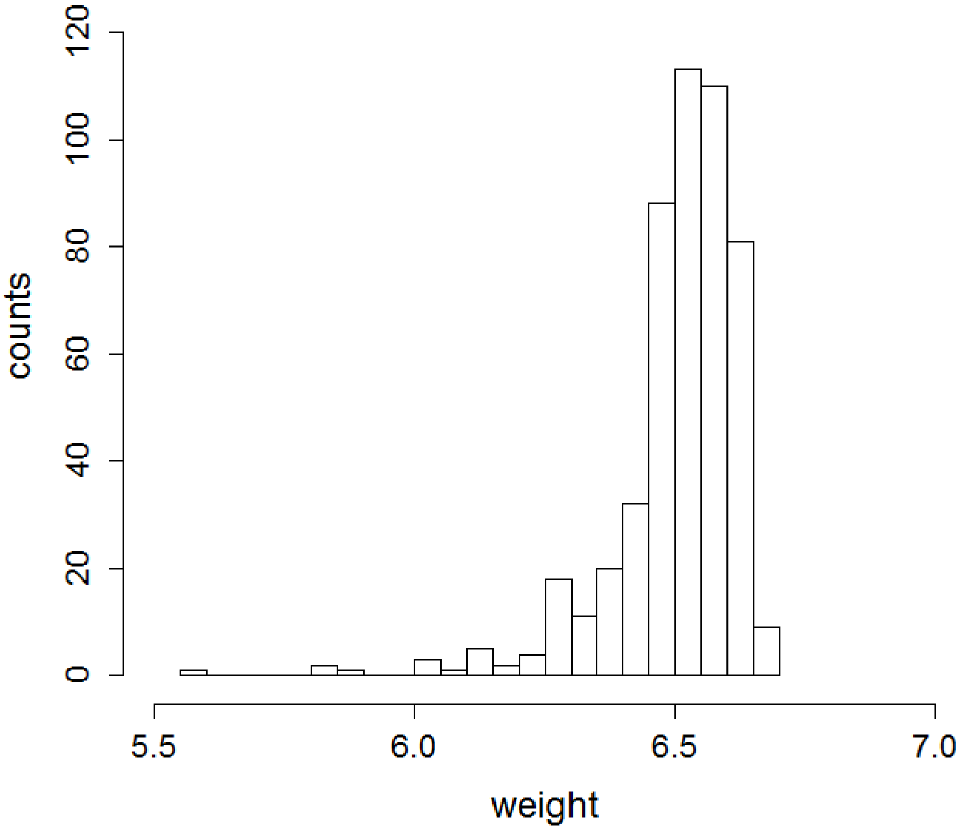
Colour: Gold.
Metal content: Alloy of copper, aluminium, zinc and tin.
Weight: 6.60 grammes.
Diameter: 20.5 millimetres.
Thickness: 2.90 millimetres.
Edge: Partially smooth, partially milled.

$\boldsymbol{H_0}$: The ten-krona piece is forged

$\boldsymbol{H_1}$: The ten-krona piece is genuine

Assume we have weighed the coin with the result
x = 6.59 gram

From a set of 501 ten-krona pieces, part of the general circulation, but considered to be genuine, we get the following empirical distribution of the variable weight (i.e. 501 weights $z_1, z_2, \ldots , z_{501}$)



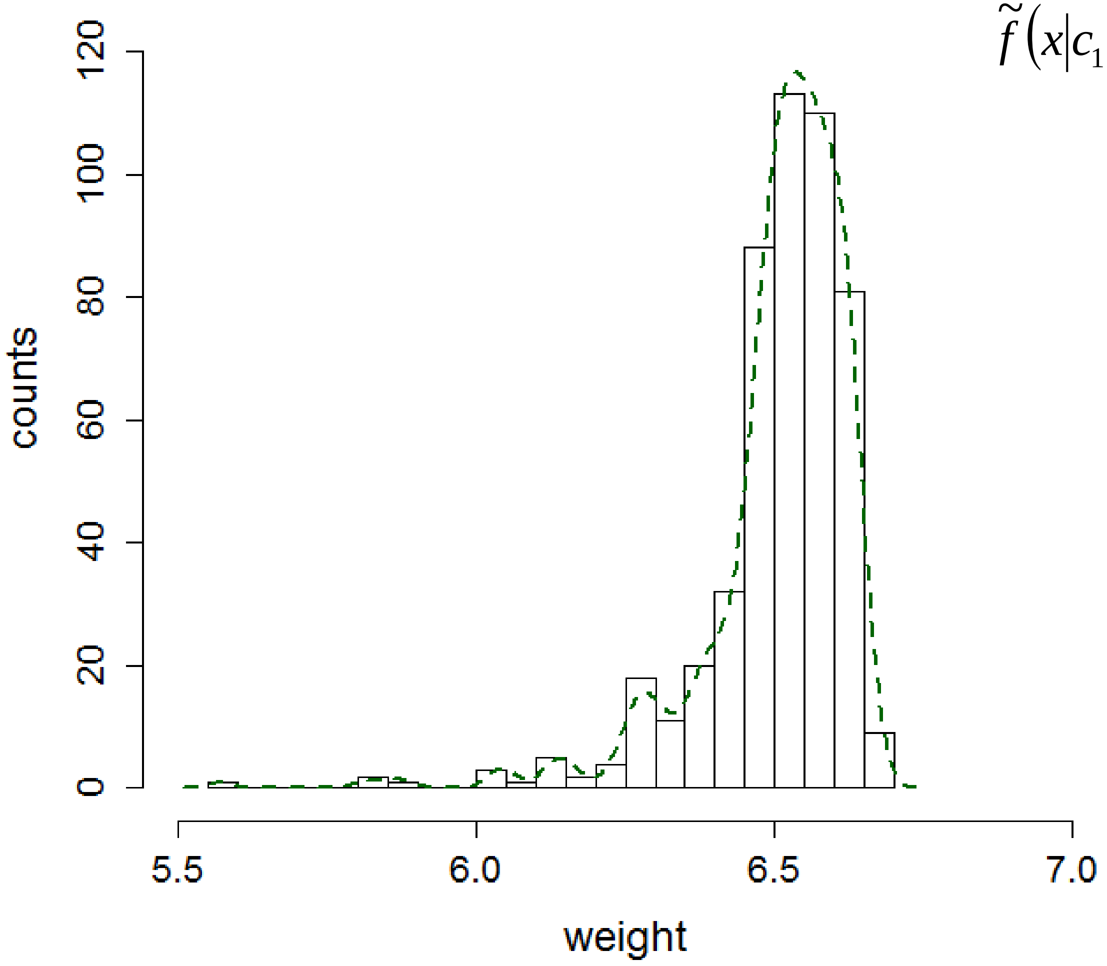
Reflects $f_1(x|\ c_1)$

H_0 : The ten-krona piece is forged

H_1 : The ten-krona piece is genuine

A kernel density estimate (using default optimal bandwidth, *h*):

$$\tilde{f}\left(x \mid c_1\right) = \frac{1}{501}\sum_{i=1}^{501}\frac{1}{h}K\left(\frac{x-z_i}{h}\right)$$

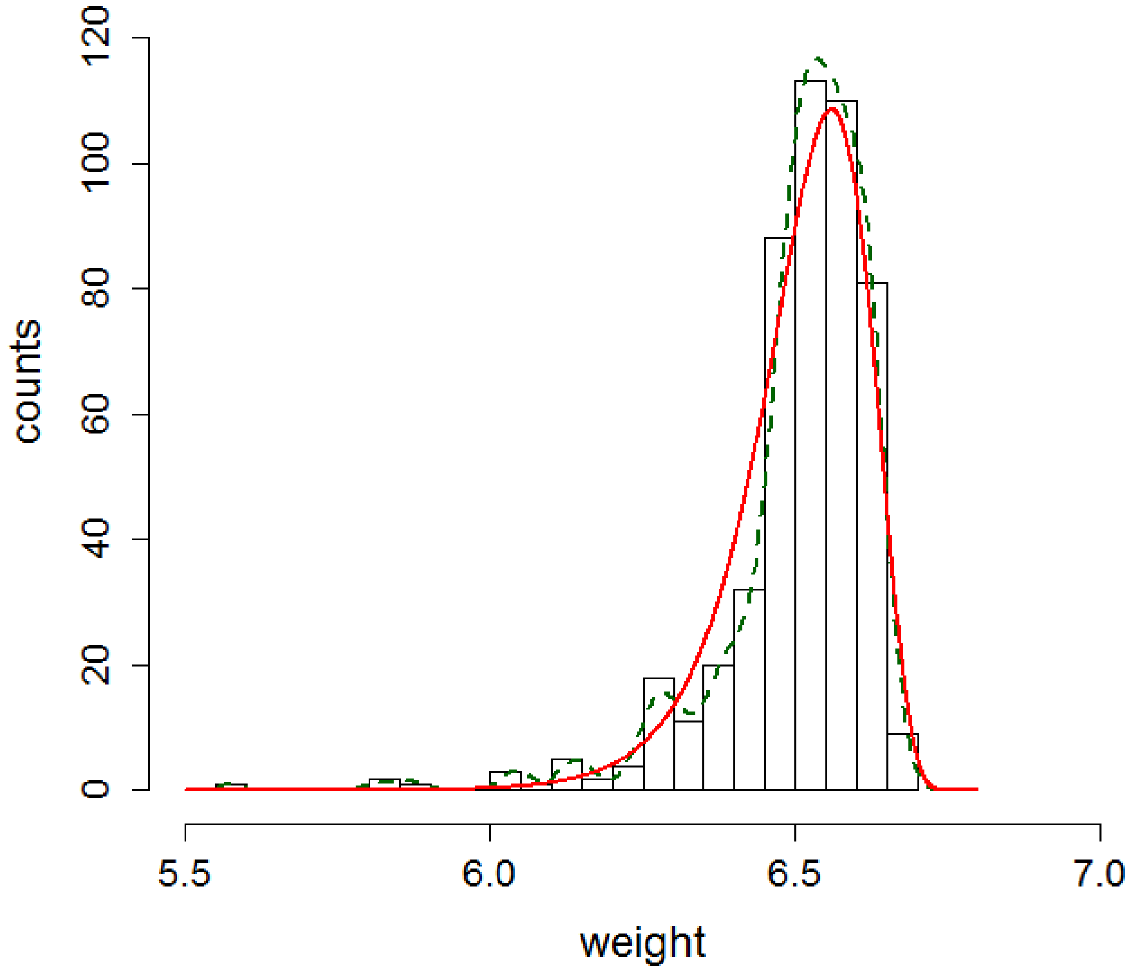


Reasonable?

$\boldsymbol{H_0}$: The ten-krona piece is forged

$\boldsymbol{H_1}$: The ten-krona piece is genuine

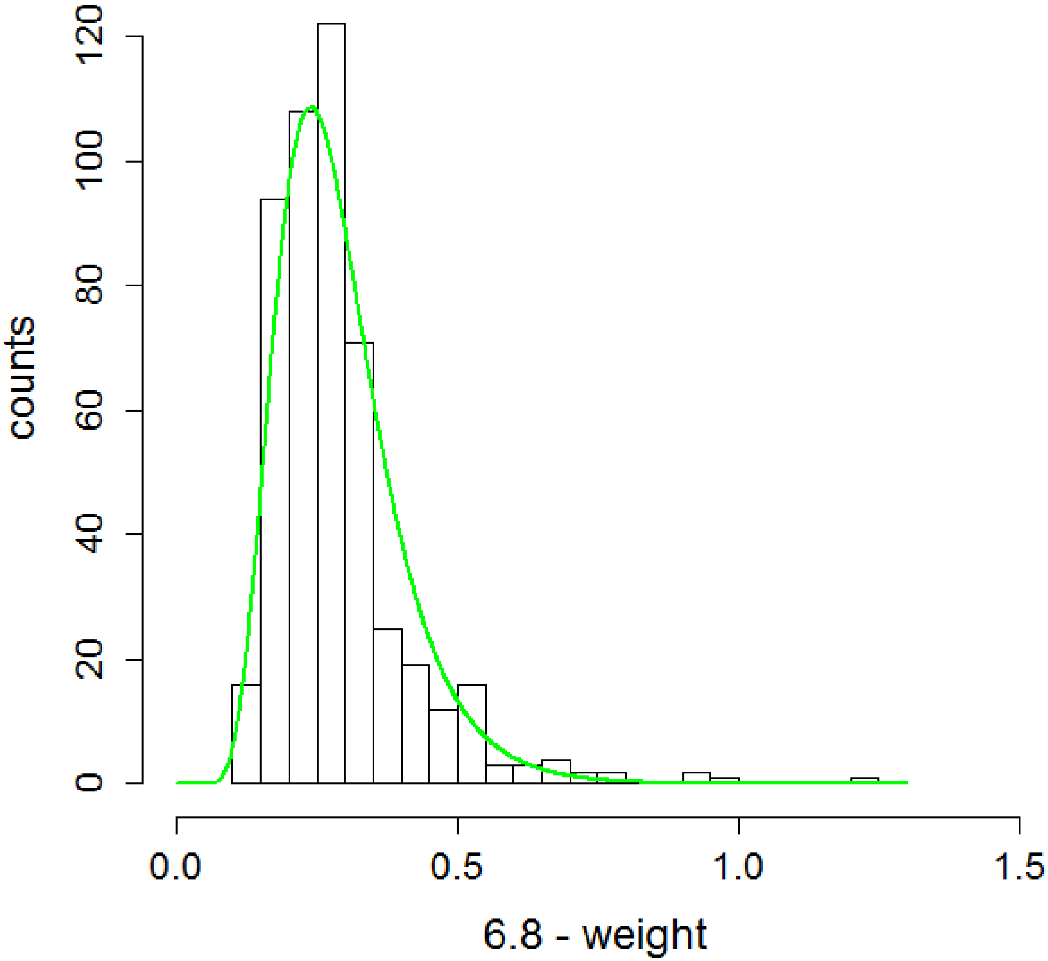
Compare with a parametric distribution fit:



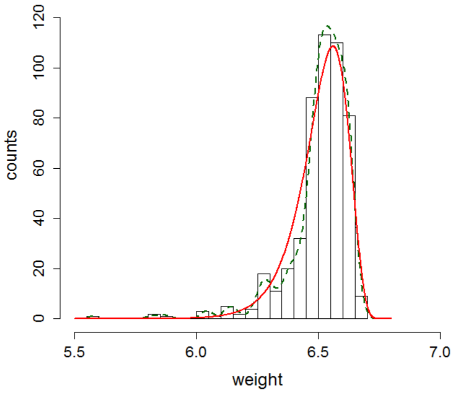
H_0 : The ten-krona piece is forged

H_1 : The ten-krona piece is genuine

The parametric fit is a fit of a lognormal distribution to the transformed variable $y = 6.8 – x$



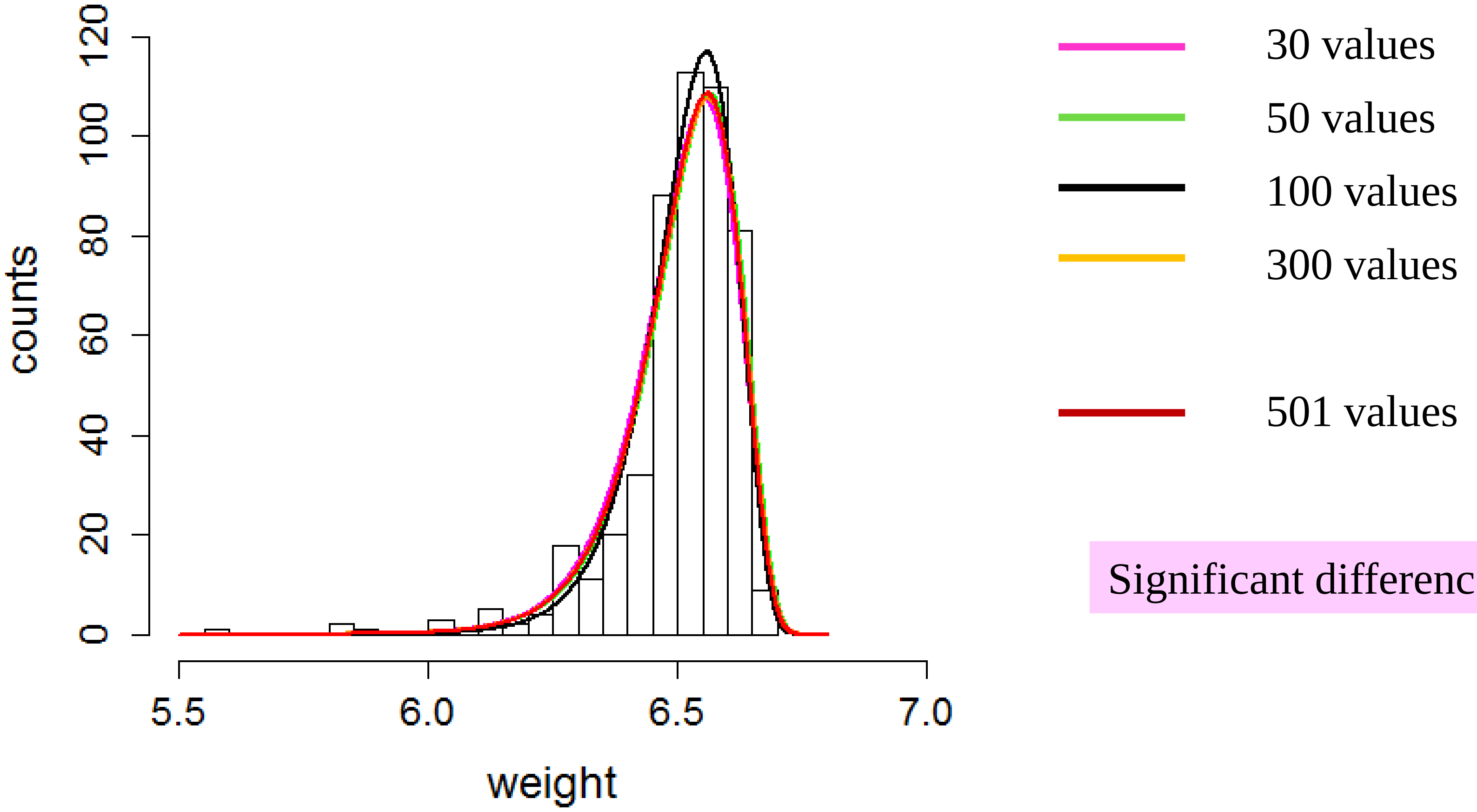
…followed by a back-transformation.



H_0 : The ten-krona piece is forged
H_1 : The ten-krona piece is genuine

How sensible is the model to the amount of background data?

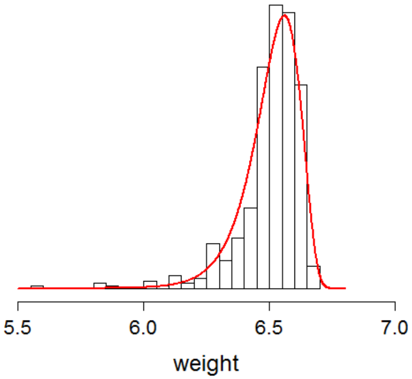
Trying to fit the lognormal distribution to subsets of 30, 50, 100 and 300 values



Significant differences?

H_0 : The ten-krona piece is forged
H_1 : The ten-krona piece is genuine

Now, assuming the fitted transformed lognormal density describes the variation in weight among genuine ten-krona pieces…



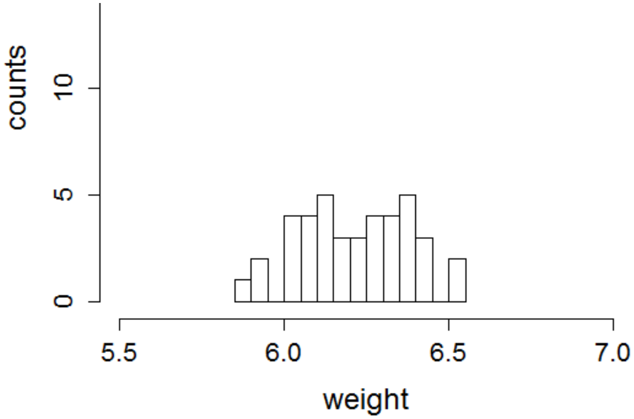
…we will need a corresponding density function describing the variation in weight among forged ten-krona pieces.

Assume we have also collected some forged ten-krona pieces from different seizures (which are undoubtedly classified as fakes). They need not to be that many, but should represent the variation.

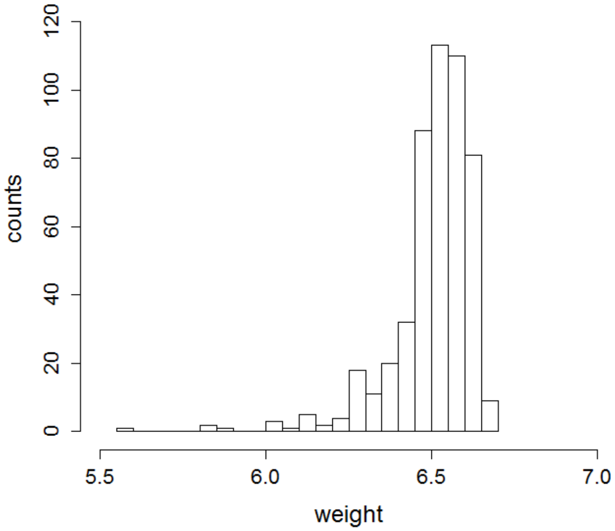
H_0 : The ten-krona piece is forged

H_1 : The ten-krona piece is genuine

Assume the variation among 40 forged pieces is like the following:



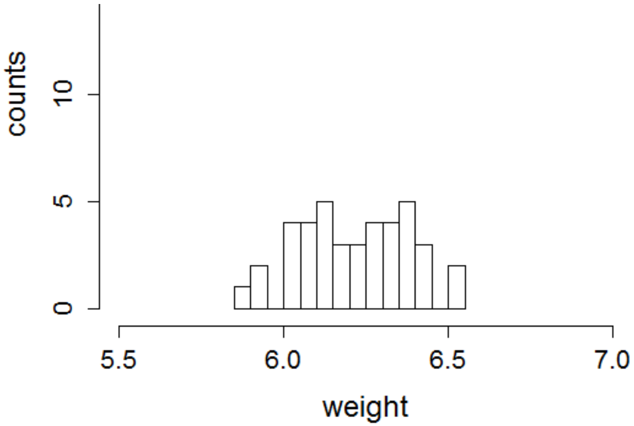
Compare with the variation among genuine pieces:



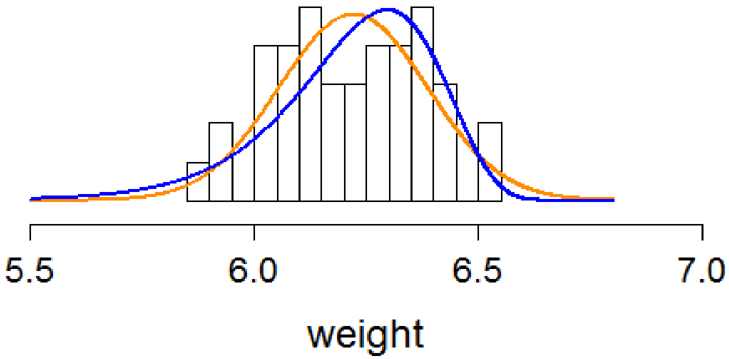
Does the difference in symmetry reflect a true difference or is it just due to different amounts of background data?

H_0 : The ten-krona piece is forged
H_1 : The ten-krona piece is genuine

Is the normal distribution a reasonable model for the variation in weight among forged pieces?



Compare the fit with a normal density to the fit with a transformed lognormal density (like for the genuine pieces)



🟧 Normal density

🟦 Transformed lognormal density

Significant differences?

H_0 : The ten-krona piece is forged
H_1 : The ten-krona piece is genuine

To be consequent, let us choose the same basic model for both data sets, i.e. model $y = 6.8 - x$ as being lognormally distributed

$$\Rightarrow \quad f_0(x|c_0) = \frac{1}{(6.8-x)\sigma_0\sqrt{2\pi}} e^{-\frac{(\ln(6.8-x)-\mu_0)^2}{2\sigma_0^2}} = f(x|\mu_0,\sigma_0)$$

$$f_1(x|c_1) = \frac{1}{(6.8-x)\sigma_1\sqrt{2\pi}} e^{-\frac{(\ln(6.8-x)-\mu_1)^2}{2\sigma_1^2}} = f(x|\mu_1,\sigma_1)$$

$$P(\tilde{x} \le x) \Leftrightarrow P(6.8 - \tilde{y} \le x)$$
$$\Leftrightarrow 1 - P(\tilde{y} \le 6.8 - x)$$
$$\Rightarrow f_{\tilde{x}}(x) = -f_{\tilde{y}}(6.8-x)\cdot(-1) =$$
$$f_{\tilde{y}}(6.8-x)$$

Hence, we assume the same functional form of the two densities (but note that this is not necessary).

The (approximate) Bayes factor is then obtained by plugging in estimates of μ_0, σ_0, μ_1 and σ_1 from the background data.

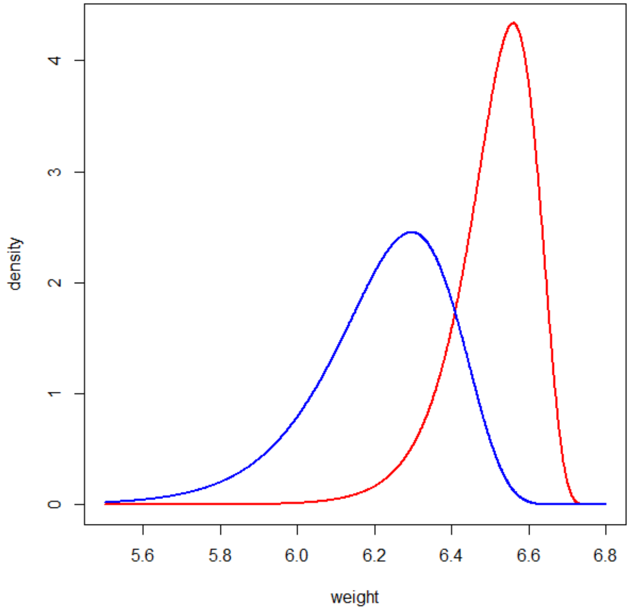
$$B = \frac{\hat{f}_0(x|c_0)}{\hat{f}_1(x|c_1)} = \frac{f(x|\hat{\mu}_0,\hat{\sigma}_0)}{f(x|\hat{\mu}_1,\hat{\sigma}_1)} = \frac{\hat{\sigma}_1}{\hat{\sigma}_0} e^{-\left[\frac{(\ln(6.8-x)-\hat{\mu}_0)^2}{2\hat{\sigma}_0^2} - \frac{(\ln(6.8-x)-\hat{\mu}_1)^2}{2\hat{\sigma}_1^2}\right]}$$

Why “approximate”?

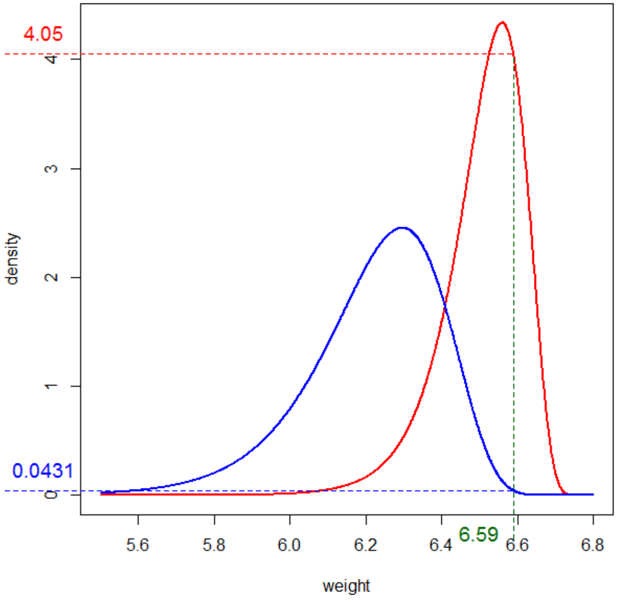
$\boldsymbol{H_0}$: The ten-krona piece is forged

$\boldsymbol{H_1}$: The ten-krona piece is genuine

To illustrate, we plot the two estimated density functions in the same diagram:



…and read of the density values at $x = 6.59$

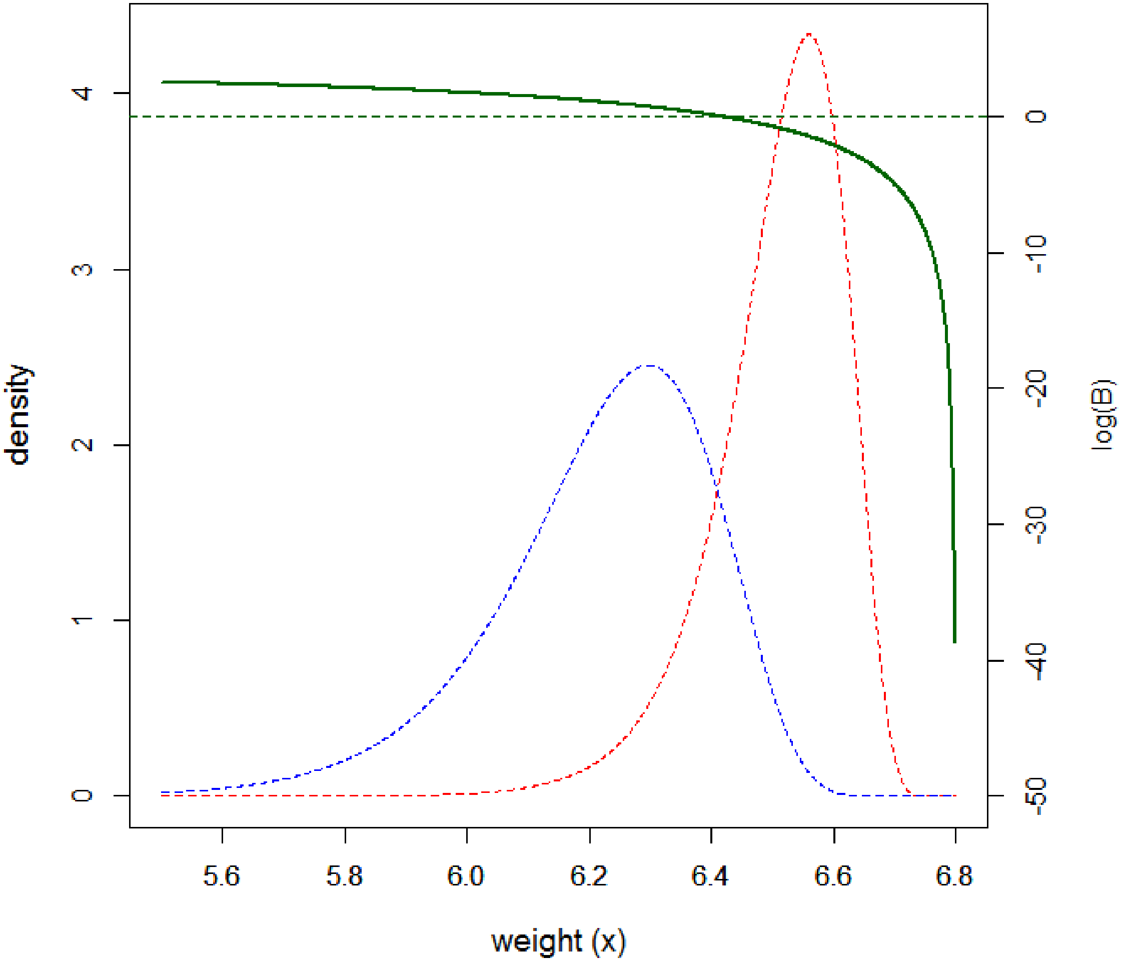


$$B = \frac{f_0\left(x \middle| c_0, \hat{\mu}^{(c_0)}, \hat{\sigma}^{(c_0)}\right)}{f_1\left(x \middle| c_0, \hat{\mu}^{(c_1)}, \hat{\sigma}^{(c_1)}\right)} \approx \frac{0.0431}{4.05} \approx 0.011$$

H_0 : The ten-krona piece is forged

H_1 : The ten-krona piece is genuine

The Bayes factor as a function of *x* (i.e. measured weight)

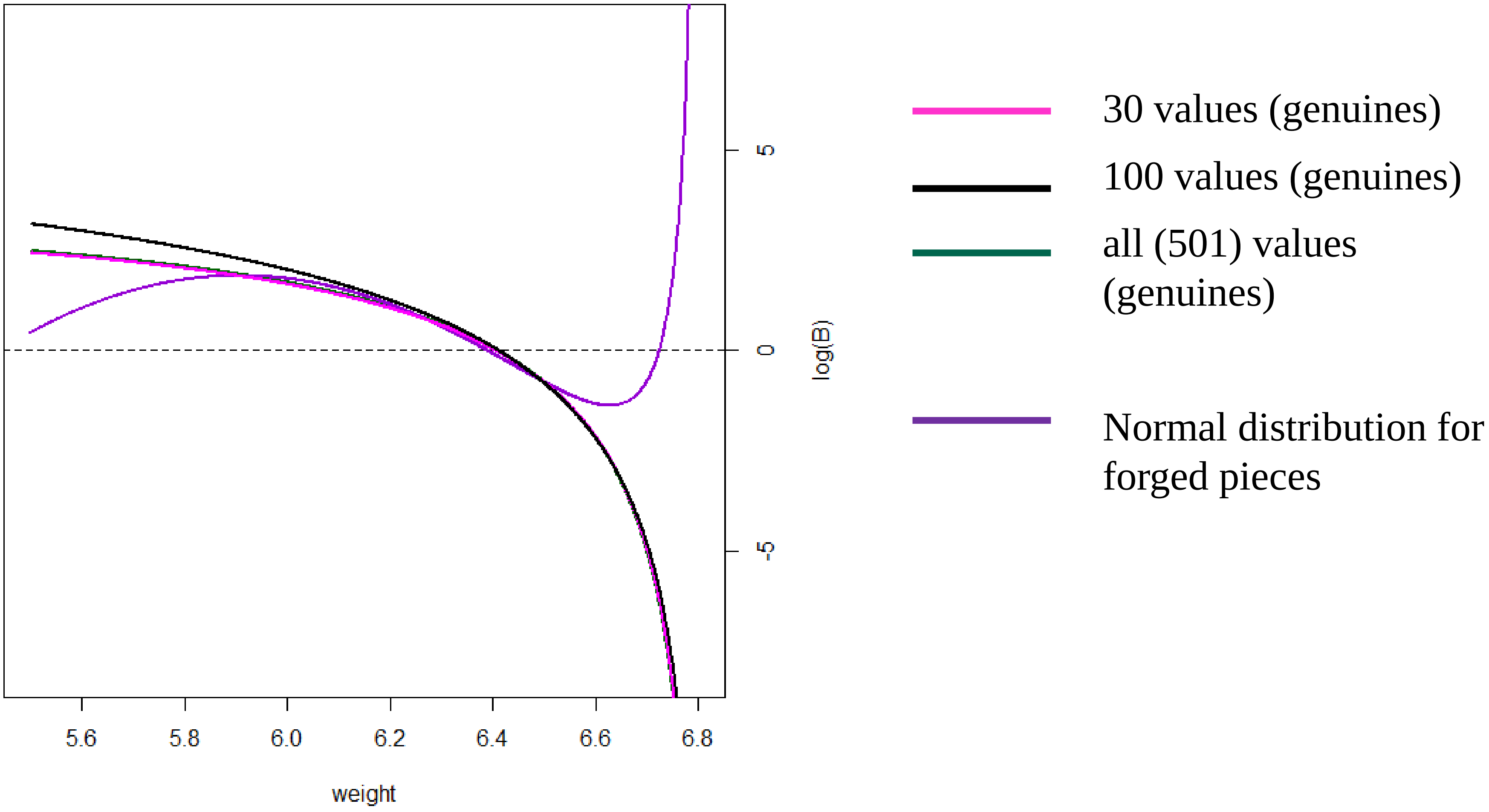


H_0 : The ten-krona piece is forged

H_1 : The ten-krona piece is genuine

Sensitivity to

- the amount of background data?
- the choice of probability distribution of weights of forged pieces?



H_0 : The ten-krona piece is forged
H_1 : The ten-krona piece is genuine

Decision-theoretic solution

As is common for all finite action problems we consider a loss function of "0 – k" type:

	Class is c_0 (forged piece)	**Class is c_1 (genuine piece)**
Decision is c_0 (forged piece)	0	$L(\text{II})$
Decision is c_1 (genuine piece)	$L(\text{I})$	0

Expected posterior losses: Decision is c_0 : $0 \cdot \Pr(H_0|x) + L(\text{II}) \cdot \Pr(H_1|x) = L(\text{II}) \cdot \Pr(H_1|x)$

Decision is c_1 : $L(\text{I}) \cdot \Pr(H_0|x) + 0 \cdot \Pr(H_1|x) = L(\text{I}) \cdot \Pr(H_0|x)$

Hence, the decision would be "forged piece" if $L(\text{II}) \cdot \Pr(H_1|x) < L(\text{I}) \cdot \Pr(H_0|x)$ and it would be "genuine piece" if $L(\text{II}) \cdot \Pr(H_1|x) > L(\text{I}) \cdot \Pr(H_0|x)$

H_0 : The ten-krona piece is forged
H_1 : The ten-krona piece is genuine

$$L(\text{II})\cdot \Pr(H_1|x) < L(\text{I})\cdot \Pr(H_0|x)$$

$$\Leftrightarrow$$

$$\frac{\Pr(H_0|x)}{\Pr(H_1|x)} > \frac{L(\text{II})}{L(\text{I})} \Leftrightarrow B\cdot\frac{\Pr(H_0)}{\Pr(H_1)} > \frac{L(\text{II})}{L(\text{I})} \Leftrightarrow B > \frac{L(\text{II})}{L(\text{I})}\cdot\frac{\Pr(H_1)}{\Pr(H_0)}$$

and

$$L(\text{II})\cdot \Pr(H_1|x) > L(\text{I})\cdot \Pr(H_0|x) \Leftrightarrow B < \frac{L(\text{II})}{L(\text{I})}\cdot\frac{\Pr(H_1)}{\Pr(H_0)}$$

Now, what would be reasonable values of L(II) and L(I) ?

$\boldsymbol{H_0}$: The ten-krona piece is forged
$\boldsymbol{H_1}$: The ten-krona piece is genuine

If the piece is forged and we make the decision that it is genuine we may get into big trouble if we use the coin.

If the piece is genuine and we make the decision that it is forged we simply lose its monetary value, i.e. SEK 10.

$\Rightarrow$ It may be reasonable to assume that $L(\mathrm{I}) >> L(\mathrm{II}) \Rightarrow L(\mathrm{II}) / L(\mathrm{I}) << 1$

Assume we have no reason beforehand to believe that one of the classes is more probable than the other $\Rightarrow \Pr(H_0) = \Pr(H_1) = 0.5$

Hence, we should make the decision "forged piece" if $B > L(\mathrm{II}) / L(\mathrm{I})$ and the decision "genuine piece" if $B < L(\mathrm{II}) / L(\mathrm{I})$.

Now $B \approx 0.011$ and $L(\mathrm{II}) / L(\mathrm{I})$ is assumed to be much lower than 1.
What shall we decide?

Another example: Quality of rice

A country grows and sells rice and the rice is grown in two areas of the country, A and B. It is well-known that rice from region A has much higher quality than rice from region B.

An importer of this rice cannot deem on whether the rice comes from region A or from region B just by observing or smelling the product.

Hence, fraud in the exporting of rice from this country would occur now and then (rice originating from region B is labelled as originating from region A).

Advanced analysis of a batch of rice for export is needed for a decision on its origin.

Using a decision-theoretical approach

$\boldsymbol{H_0}$: The rice in question originates from region A
$\boldsymbol{H_1}$: The rice in question originates from region B

Data to be used for the decision making:

Measurements of the contents (in mg/kg or %) of 11elements:

As (x_1), Br (x_2), K (x_3), Mn (x_4), Rb (x_5), Zn (x_6), Ca (x_7), Cu (x_8), Fe (x_9), Mg (x_{10}), P (x_{11})

Measurement of the so-called *stable isotope ratio* δ^{18}O (x_{12})

Definition: $$\delta^{18}\mathrm{O} = \frac{^{18}\mathrm{O}}{^{16}\mathrm{O}} - 1$$

The ratio of the amount of oxygene with mass nuclear number 18 and the amount of oxygene with mass number 16

Usually measured in terms of per mille (positive or negative deviation from 1)

H_0 : The rice in question originates from region A

H_1 : The rice in question originates from region B

Some background data:

Region A:

	As (mg/kg)	**Br (mg/kg)**	**K (%)**	**Mn (mg/kg)**	**Rb (mg/kg)**	**Zn (mg/kg)**	**Ca (mg/kg)**	**Cu (mg/kg)**	**Fe (mg/kg)**	**Mg (mg/kg)**	**P (mg/kg)**	**δ18O (o/oo)**
1	0,237	0,782	0,182	24,68	8,85	24,31	80,00	1,77	3,56	434	1366	26,10
2	0,211	0,318	0,172	16,36	5,73	23,03	70,00	1,90	2,92	412	1382	26,45
3	0,121	0,231	0,156	14,01	9,11	21,88	71,00	1,39	7,01	370	1268	26,25
4	0,150	0,364	0,126	11,37	2,05	22,11	88,00	0,60	3,11	285	813	27,57
5	0,250	0,262	0,121	12,47	3,71	29,96	91,00	2,27	12,64	485	1498	26,05
|	|	|	|	|	|	|	|	|	|	|	|	|
67	0,169	0,236	0,131	16,60	6,33	21,48	70,35	2,708	5,321	413	1337	26,69

Region B:

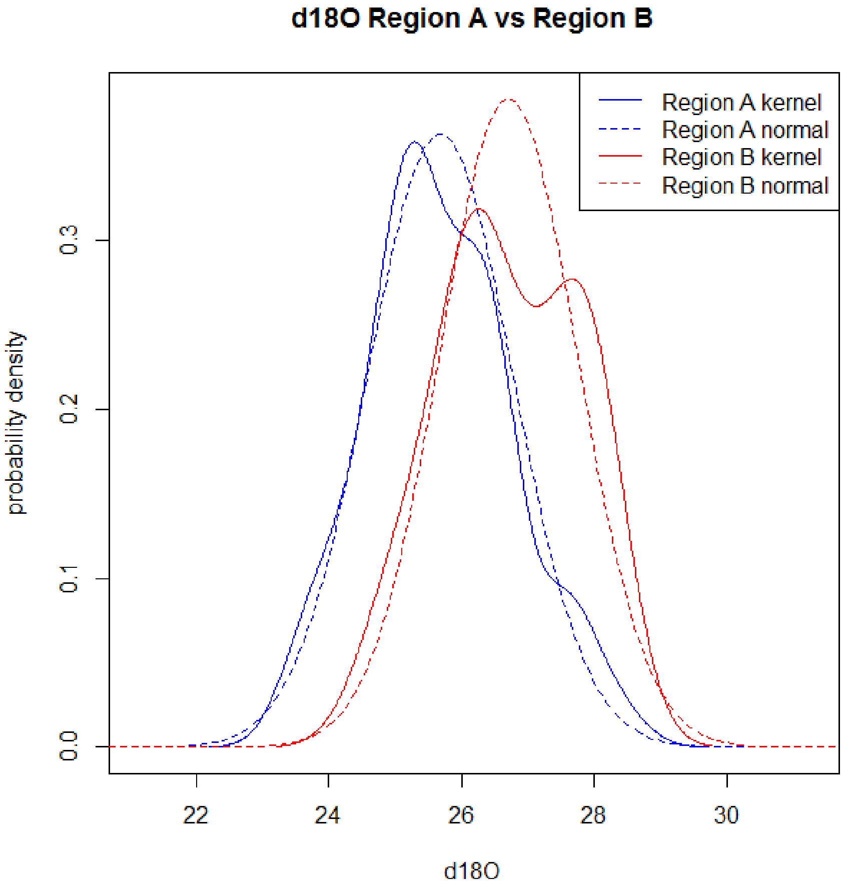
	As (mg/kg)	**Br (mg/kg)**	**K (%)**	**Mn (mg/kg)**	**Rb (mg/kg)**	**Zn (mg/kg)**	**Ca (mg/kg)**	**Cu (mg/kg)**	**Fe (mg/kg)**	**Mg (mg/kg)**	**P (mg/kg)**	**δ18O (o/oo)**
1	0,227	0,296	0,136	13,39	14,32	21,42	41,00	1,090	5,310	320	1230	26,25
2	0,160	0,287	0,073	9,23	18,65	23,35	50,00	1,340	8,130	347	1286	26,44
3	0,157	0,174	0,107	10,27	18,06	21,58	67,00	0,690	5,680	406	1396	27,10
4	0,262	0,239	0,079	9,75	21,68	21,67	63,00	0,780	21,310	307	1190	26,22
5	0,246	0,171	0,103	11,27	15,39	19,74	93,00	0,670	11,070	442	1529	25,90
|	|	|	|	|	|	|	|	|	|	|	|	|
17	0,164	0,219	0,112	12,74	15,86	16,12	58,03	1,996	2,545	407	1471	28,02

H_0 : The rice in question originates from region A

H_1 : The rice in question originates from region B

Which probability models may be applied?

At first, consider variable x_{12}, i.e. δ^{18}O



Normal distribution reasonable?

Sufficient with only one variable?

H_0 : The rice in question originates from region A
H_1 : The rice in question originates from region B

Using more variables

If we can anticipate normally distributed data (also in several dimensions)…

…not an unusual assumption within Chemometrics – on the contrary even exaggerated …

Use multivariate normal densities for subsets of variables that are too strongly correlated – multiply densities for such sets.

$$f_i(\mathbf{x}|c_i) = f_i(x_1, x_2, \ldots, x_{12}|c_i) = \prod_j f_{i,j}(\mathbf{x}_j|c_i), \quad i = 0,1 \text{ (Region A, Region B)}$$

where the product is over the sets of strongly correlated variables and where $f_{i,j}(\mathbf{x}_j\,|c_i\,)$ is the joint multivariate density over the set $\mathbf{x}_j$ of strongly correlated variables.

Too strongly correlated??

Partial correlation

We can obtain the unique correlation between x_i and x_j by calculating the *partial correlation coefficient* .

Assume a set of random variables $\boldsymbol{x} = \{x_1, \ldots , x_m\}$

$$\rho_{x_i, x_j | \boldsymbol{x} \backslash \{x_i, x_j\}} = \frac{E\left[x_i \cdot x_j \middle| \boldsymbol{x} \backslash \{x_i, x_j\}\right] - E\left[x_i \middle| \boldsymbol{x} \backslash \{x_i, x_j\}\right] \cdot E\left[x_j \middle| \boldsymbol{x} \backslash \{x_i, x_j\}\right]}{\sqrt{Var\left[x_i \middle| \boldsymbol{x} \backslash \{x_i, x_j\}\right] \cdot Var\left[x_j \middle| \boldsymbol{x} \backslash \{x_i, x_j\}\right]}}$$

The pairwise partial correlation can be obtained from the inverse of the covariance matrix, $\mathbf{\Omega} = \mathbf{C}(\boldsymbol{x})$ provided it is positive definite**.**

If so, let $\mathbf{P} = \mathbf{\Omega}^{-1}$

$$\rho_{x_i, x_j | \boldsymbol{x} \backslash \{x_i, x_j\}} = -\frac{\mathbf{P}_{ij}}{\sqrt{\mathbf{P}_{ii} \cdot \mathbf{P}_{jj}}}$$

Graphical modelling

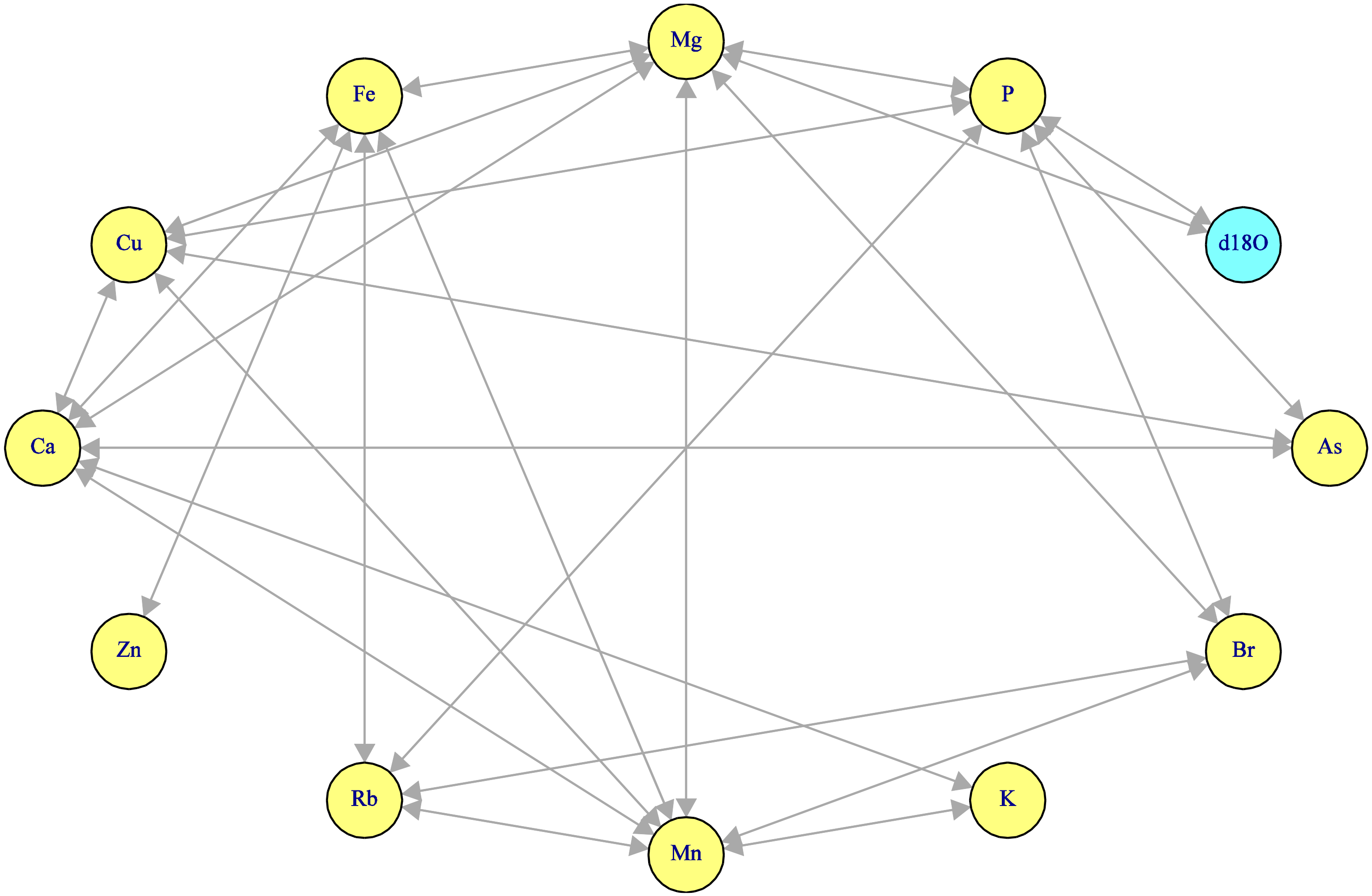
Upon decision about how high a calculated partial correlation can be, still anticipating approximate independence between the variables we may join all pairs of variables considered to be dependent.

This is one way of defining a graphical model for the variables.

H$_0$: The rice in question originates from region A

H$_1$: The rice in question originates from region B

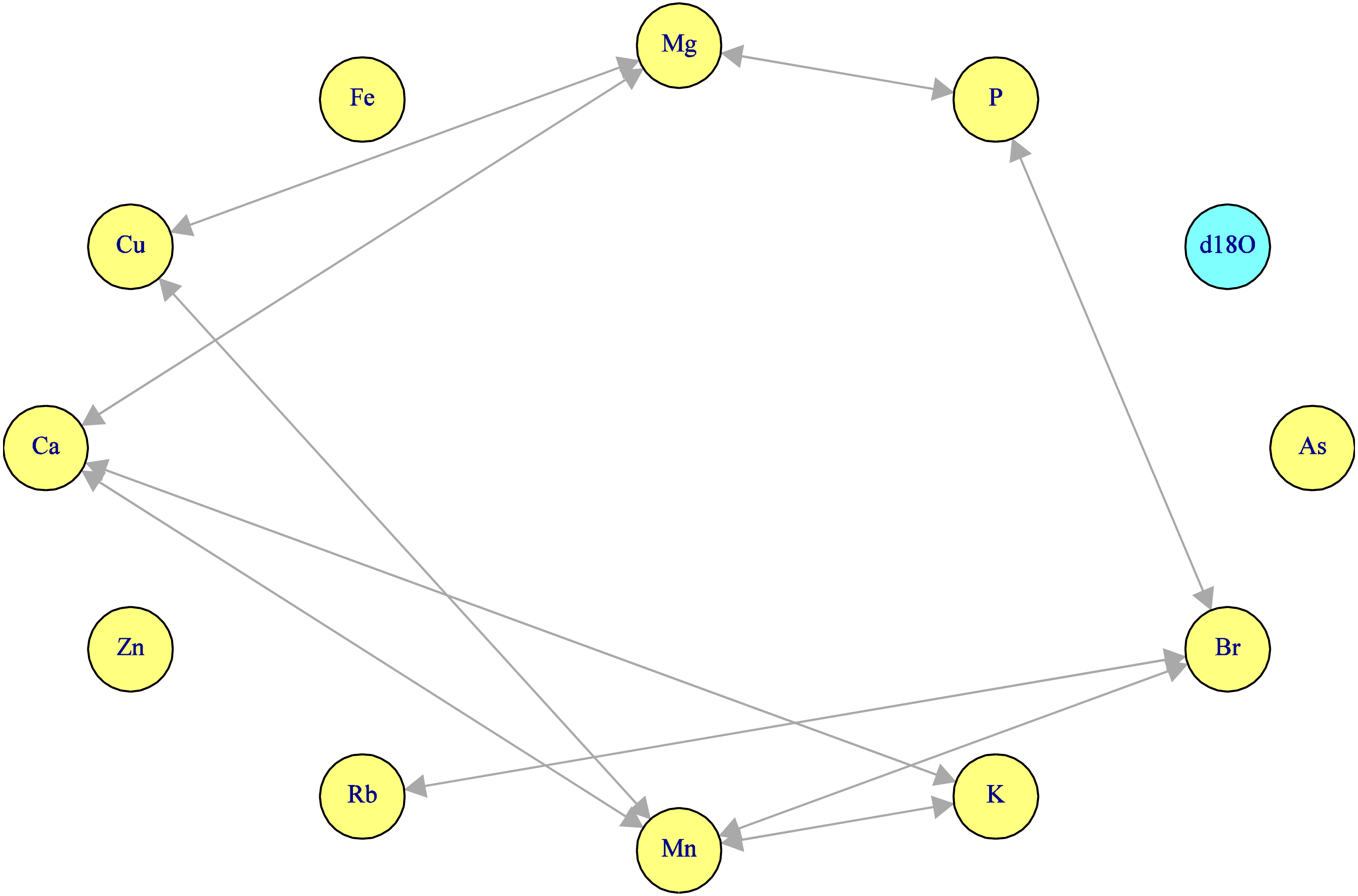
Ignoring partial correlations less than 0.2:



H_0 : The rice in question originates from region A

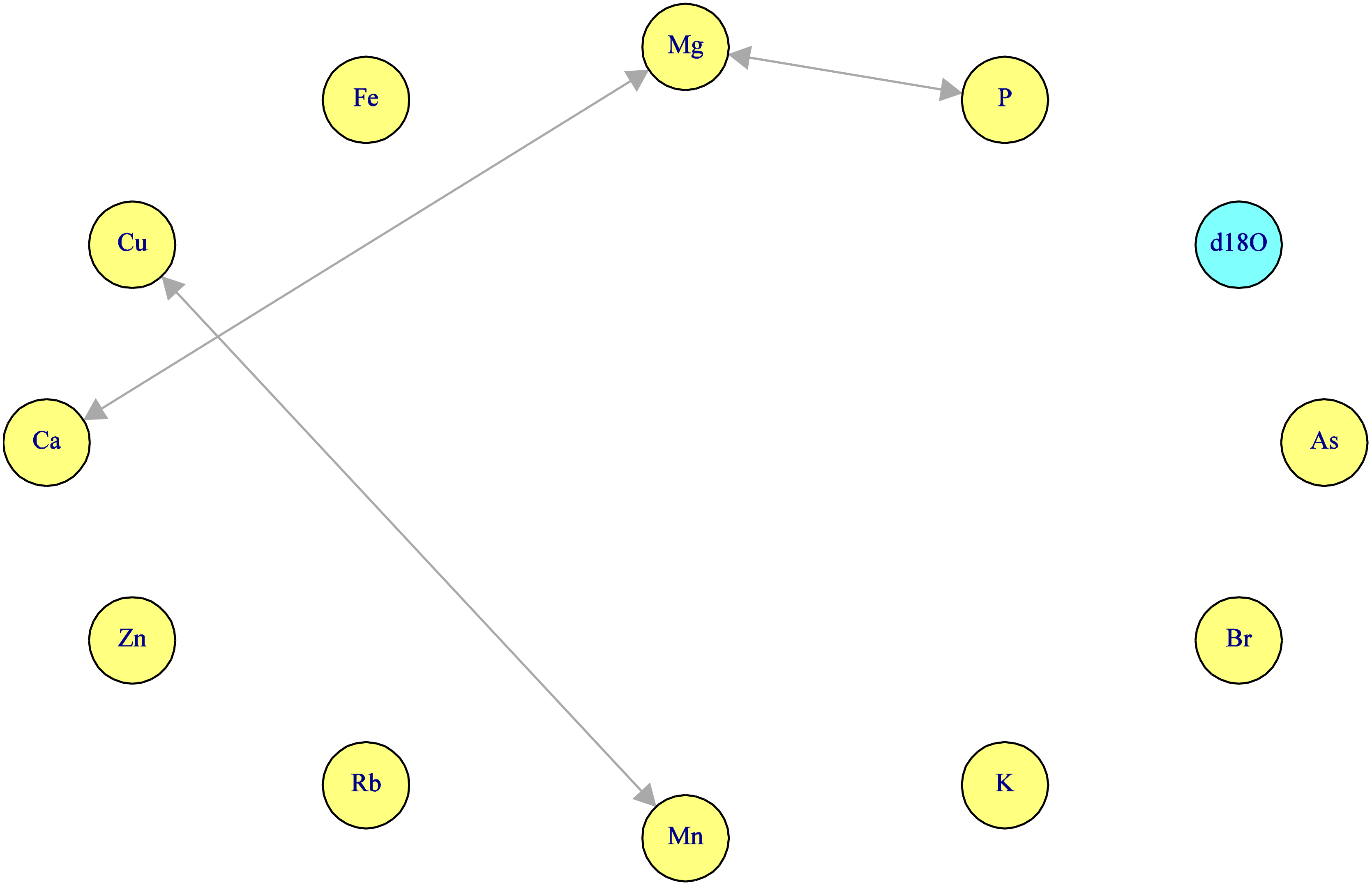
H_1 : The rice in question originates from region B

Ignoring partial correlations less than 0.3:



H_0 : The rice in question originates from region A
H_1 : The rice in question originates from region B

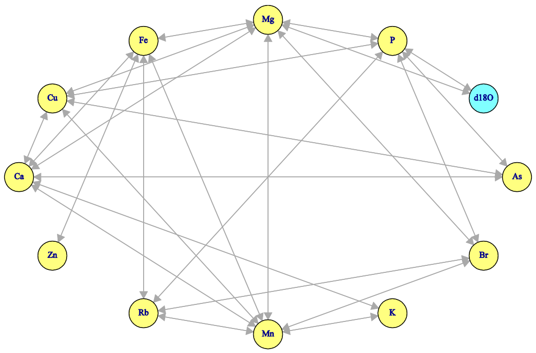
Ignoring partial correlations less than 0.4:



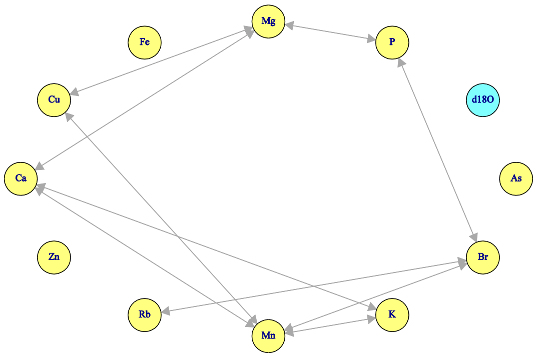
H_0 : The rice in question originates from region A

H_1 : The rice in question originates from region B

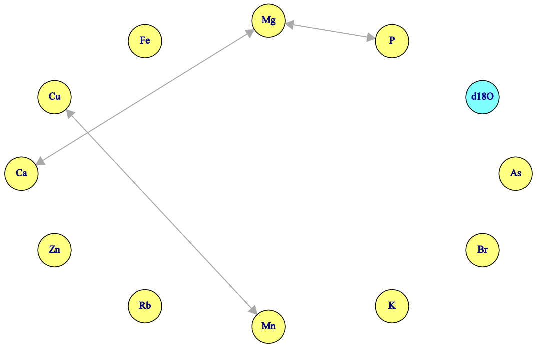
Ignoring partial correlations less than 0.2:



Ignoring partial correlations less than 0.3:



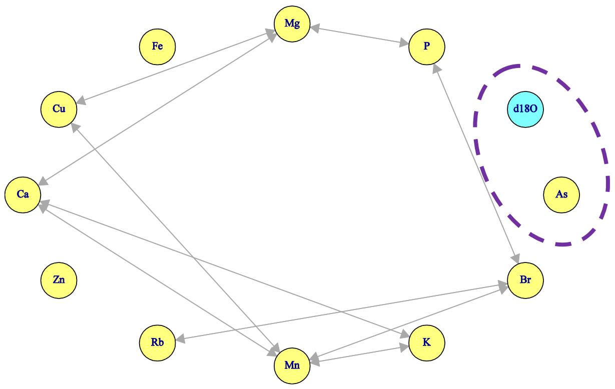
Ignoring partial correlations less than 0.4:



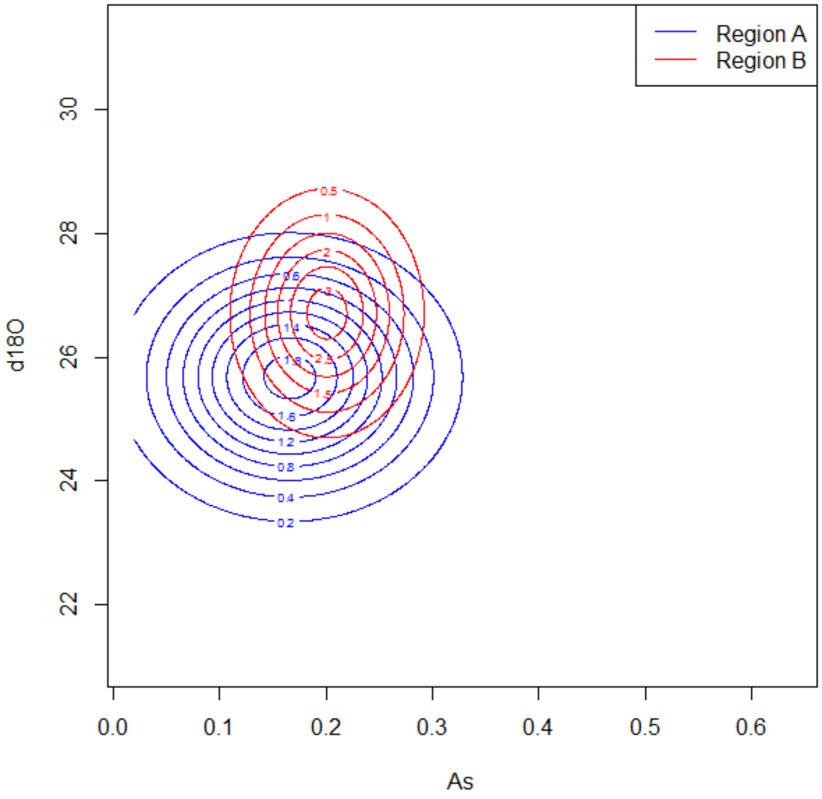
$\boldsymbol{H_0}$: The rice in question originates from region A

$\boldsymbol{H_1}$: The rice in question originates from region B

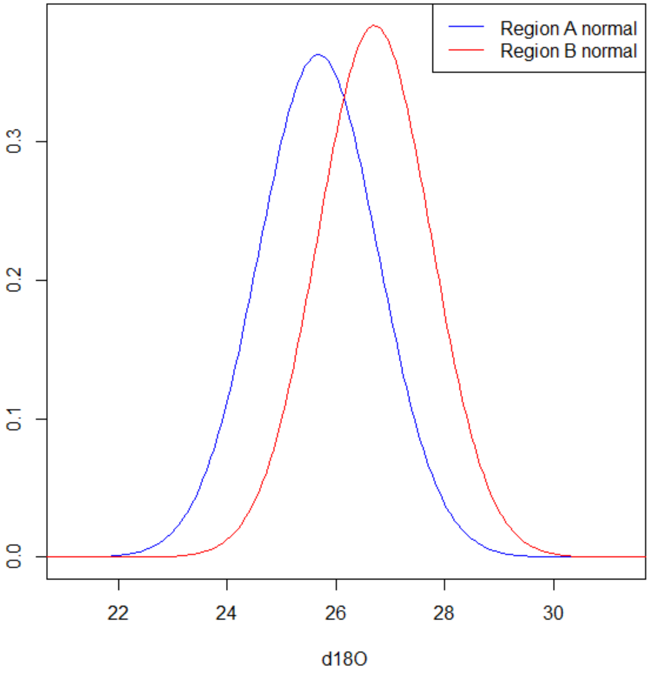
If we ignore partial correlations less than 0.3 we may (for instance) consider δ18O and As to be approximately independent.



Bivariate densities:



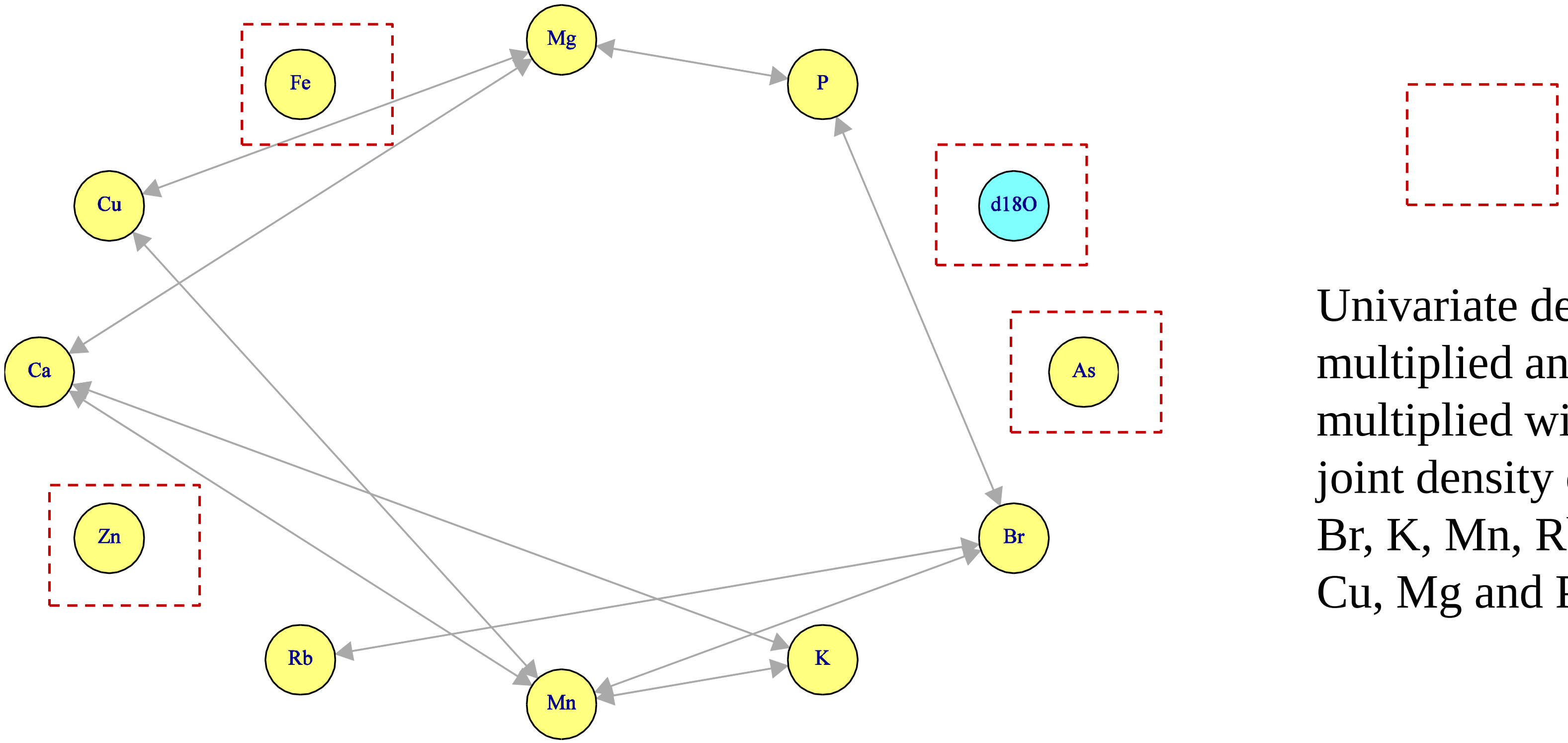
Compare with univariate densities:



H_0 : The rice in question originates from region A

H_1 : The rice in question originates from region B

Using all variables and the graphical model ignoring partial correlations less than 0.3



Univariate densities multiplied and multiplied with the joint density of Br, K, Mn, Rb, Ca, Cu, Mg and P

$$f_i(\mathbf{x}|c_i) = f_{i,1}(x_{12}|c_i) \times f_{i,2}(x_1|c_i) \times f_{i,3}(x_6|c_i) \times f_{i,4}(x_9|c_i) \times f_{i,5}(x_2, x_3, x_4, x_5, x_7, x_8, x_{10}, x_{11}|c_j)$$

H_0 : The rice in question originates from region A

H_1 : The rice in question originates from region B

Assessing the method by leave-one-out cross validation:

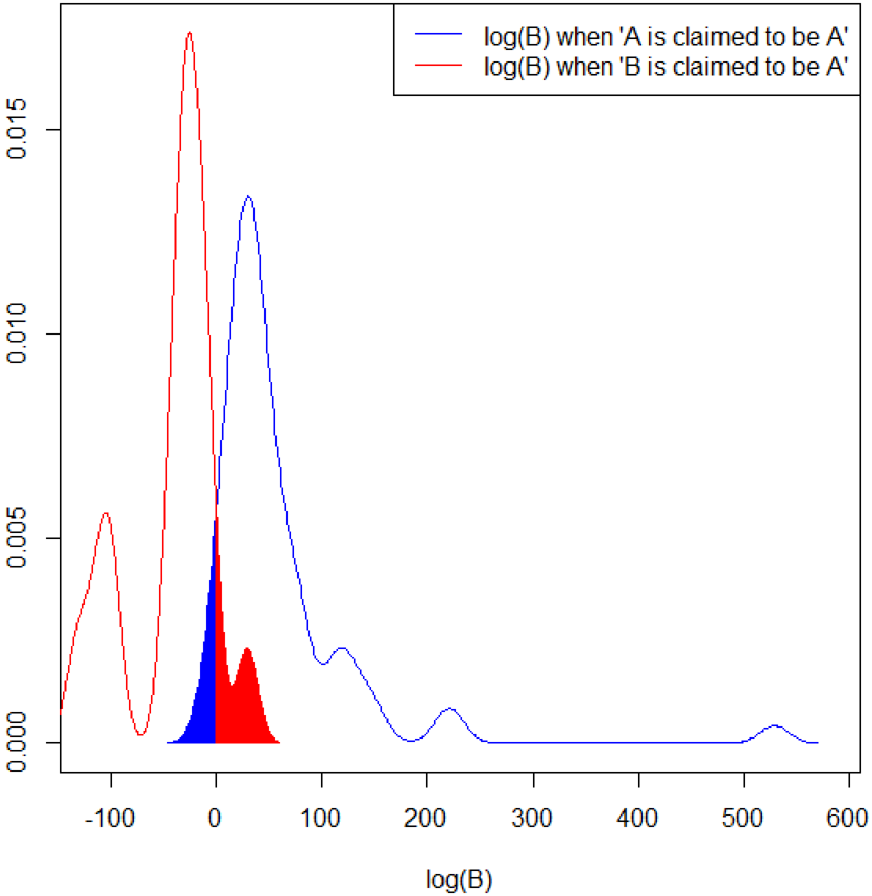
- Take out sample 1, fit the probability density functions $f(\boldsymbol{x}\,|c_0)$ and $f(\boldsymbol{x}\,|c_1)$
- Calculate the Bayes factor for the excluded sample
- Repeat with next sample etc.
- Group the Bayes factors according to the known ground truth (known region)

As (mg/kg)	Br (mg/kg)	K (%)	Mn (mg/kg)	Rb (mg/kg)	Zn (mg/kg)	Ca (mg/kg)	Cu (mg/kg)	Fe (mg/kg)	Mg (mg/kg)	P (mg/kg)	δ18O (o%)
0.227	0.296	0.136	13.39	14.32	21.42	41.00	1.090	5.310	320	1230	26.25
0.160	0.287	0.073	9.23	18.65	23.35	50.00	1.340	8.130	347	1286	26.44
0.157	0.174	0.107	10.27	18.06	21.58	67.00	0.690	5.680	406	1396	27.10
0.262	0.239	0.079	9.75	21.68	21.67	63.00	0.780	21.310	307	1190	26.22
0.246	0.171	0.103	11.27	15.39	19.74	93.00	0.670	11.070	442	1529	25.90
0.199	0.212	0.098	12.14	32.88	21.20	76.00	1.170	4.440	391	1394	26.20
0.157	0.190	0.118	12.94	13.13	19.30	66.00	1.010	12.000	399	1441	26.25
0.253	0.286	0.110	9.76	17.23	21.66	54.00	0.640	5.460	319	1186	25.60
0.177	0.451	0.115	16.21	26.83	19.99	96.00	2.410	9.700	546	1651	25.30

H_0 : The rice in question originates from region A

H_1 : The rice in question originates from region B

Results:



False positive rate: 6 %

False negative rate: 3 %

H_0 : The rice in question originates from region A
H_1 : The rice in question originates from region B

Assume the importer considers buying a shipment of rice labelled as coming from Region A and hence at a corresponding price, 10 000 $ say.

If the importer decides to buy the rice he may count on a profit of 3000 $ from selling all of it.
Assume this will happen if the rice is from Region A, but if the rice is from Region B consumers are expected to learn about the lower quality after a while, and we can assume that only half the shipment will be sold (the rest will eventually have to be destroyed).

Hence, the loss function is

	Rice is from Region A	**Rice is from Region B**
Decision is Buy	0	$L(\text{II}) = 5000 + 1500 = 6500$
Decision is Do not buy	$L(\text{I}) = 3000$	0

H_0 : The rice in question originates from region A
H_1 : The rice in question originates from region B

Like before the optimal decisions with respect to minimum posterior loss are

Buy if $B > \dfrac{L(\mathrm{II})}{L(\mathrm{I})} \cdot \dfrac{\Pr(H_1)}{\Pr(H_0)} = \dfrac{6500}{3000} \cdot \dfrac{\Pr(H_1)}{\Pr(H_0)}$

Do not buy if $B < \dfrac{6500}{3000} \cdot \dfrac{\Pr(H_1)}{\Pr(H_0)}$

Assume the importer has no background information on whether rice shipments from the supplier tend to be falsely labelled or not.

$$\Rightarrow \frac{\Pr(H_1)}{\Pr(H_0)} = 1$$

$\Rightarrow$

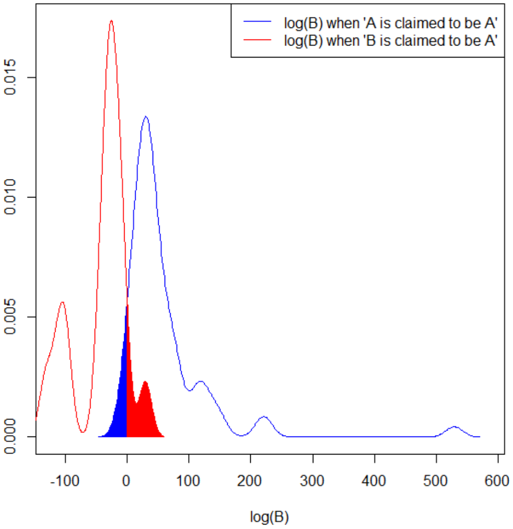
If $B > 6500/3000 \approx 2.17$ then Buy

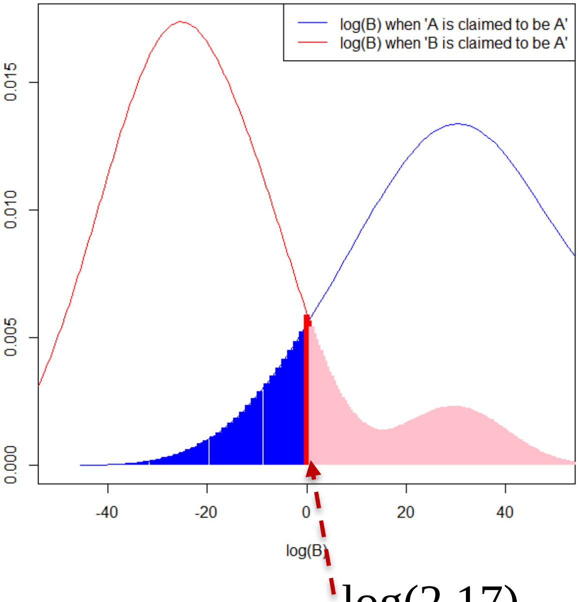
If $B < 2.17$ then Do not buy

How reliable is this decision rule?

H_0 : The rice in question originates from region A

H_1 : The rice in question originates from region B





log(2.17)

The rate of false positives were 6%, and the rate of taking the decision to Buy while the rice is falsely labelled is about the same (pink shaded area almost coincides the red-shaded area in the left picture [when the same scale is used]).

$\boldsymbol{H_0}$: The rice in question originates from region A
$\boldsymbol{H_1}$: The rice in question originates from region B

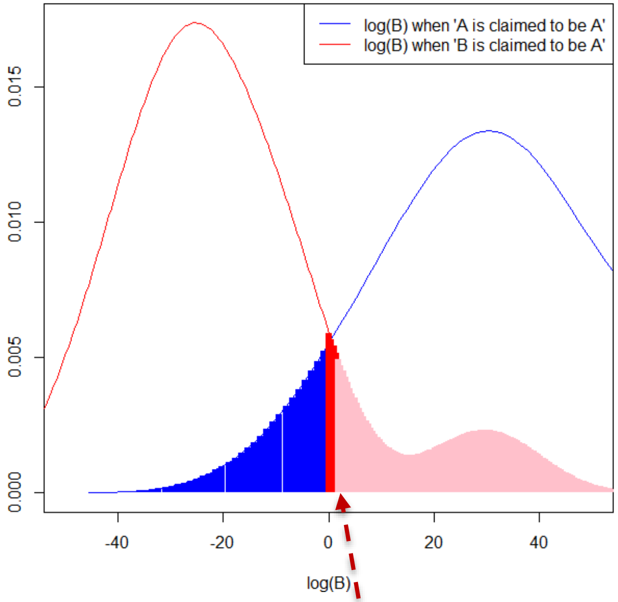
What if the importer suspects the supplier to that extent that his prior odds of H_0 is only 1 to 2 against (i.e. the ratio equals 0.5)?

$$\Rightarrow \frac{\Pr(H_1)}{\Pr(H_0)} = 2$$

$$\Rightarrow$$

If $B > (6500/3000) \cdot 2 \approx 4.33$ then Buy

If $B < 4.33$ then Do not buy



log(4.33)

Slightly better, but maybe not satisfactory.