# Meeting 15:
# Bayesian inference:
## Comparisons with forensic examples

LINKÖPING UNIVERSITY

# Comparisons

Consider the issue to test

$H_0 : \theta_1 \geq \theta_2$ (or $\theta_1 \leq \theta_2$ ) against
$H_1$: $\theta_1 < \theta_2$ (or $\theta_1 < \theta_2$)

where $\theta_1$ and $\theta_2$ are of the same parameter type, but from two different populations.

By rewriting $H_0$ as $\theta_1 - \theta_2 \geq 0$ (and correspondingly $H_1$ as $\theta_1 - \theta_2 < 0$ we have transferred the problem back to a test of the value of the compound parameter $\theta = \theta_1 - \theta_2$

Statistical calculations for the “merging” of two populations will then be needed to sort out the problem – independent populations and independent samples gives product priors and product likelihoods.

# Two-sided testing

So far the hypotheses specified through parameters have been of the kind

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1 \; (= \Theta \setminus \Theta_0)$$

where $\Theta$ is the parameter space.

This may be viewed as a classification problem – the issue is to choose between two hypotheses (two models) the one that best describes the data given prior probabilities for both hypotheses.

The specification through a parameter induces the need for integrating the likelihood function as soon as any of the hypotheses is composite.

Very often, though, the issue is to analyse two sets of data material for making inference about whether they originate from the same model.

$$H_0 : \text{Model for data set 1} = \text{Model for data set 2}$$
$$H_1 : \text{Model for data set 1} \neq \text{Model for data set 2}$$

Not so much about the models themselves, but essentially about whether they are common or not – *n.b.* completely specified models, not just up to the value of a parameter.

*Examples*

- Are the effects of the two treatments equal or not?

- Do the two seizures of amphetamine come from the same manufacturing batch?

- Was the recovered shoeprint made by the suspects’ shoe ?

In many situations a parametric model is common for the two data sets.

The difference is then expressed in terms of one or several parameters.

$$H_0 : \theta_1 = \theta_2$$
$$H_1 : \theta_1 \neq \theta_2$$

Assuming independent data sets the general likelihood function is

$$L\left(\theta_1, \theta_2 \middle| Data\right) = L\left(\theta_1 \middle| Data\ set\ 1\right) \cdot L\left(\theta_2 \middle| Data\ set\ 2\right)$$

However, the likelihood function for $H_0$ is a function of one parameter value only:

$$L\left(H_0 \middle| Data\right) = L\left(\theta \middle| Data\ set\ 1\right) \cdot L\left(\theta \middle| Data\ set\ 2\right)$$

since $H_0$ states $\theta_1 = \theta_2 = \theta$.

With prior density $p(\theta)$ for $\theta$ (common to both hypothesis) the Bayes factor becomes

$$B = \frac{\int_\Theta L\left(\theta \middle| Data\ set\ 1\right)\cdot L\left(\theta \middle| Data\ set\ 2\right)\cdot p(\theta)d\theta}{\int_\Theta L\left(\theta_1 \middle| Data\ set\ 1\right)\cdot L\left(\theta_2 \middle| Data\ set\ 2\right)\cdot p(\theta_1)\cdot p(\theta_2)d\theta_1\, d\theta_2} =$$

$$= \frac{\int_\Theta L\left(\theta \middle| Data\ set\ 1\right)\cdot L\left(\theta \middle| Data\ set\ 2\right)\cdot p(\theta)d\theta}{\int_\Theta L\left(\theta \middle| Data\ set\ 1\right)\cdot p(\theta)d\theta \cdot \int_\Theta L\left(\theta \middle| Data\ set\ 2\right)\cdot p(\theta)d\theta}$$

Lindley,
*Biometrika*, 1977

*Example*

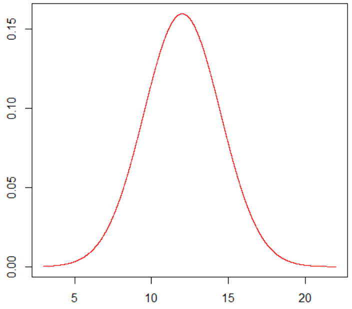
Assume we are comparing two seizures of cannabis with respect to their mean concentration of THC (*Tetrahydrocannabinol, the active narcotic substance*).

Denote the two means $\mu_1$ and $\mu_2$ respectively.

The concentration for the types of material in question (typically buds of cannabis plants) varies normally between 5 and 20 % with a peak around 12 %. We therefore assign a prior density for the mean concentration of *any* material of this type as

$$\mu \sim N\left(\text{mean} = 12\,, \text{standard deviation} = \frac{20-5}{6}\right) = N(12{,}2.5)\,[\%]$$

For symmetric close-to-normal distributions a reasonable estimate of the standard deviation is Range/6.
The range is almost covered by the mean ± 3 standard deviations for a normal distribution.



To simply formulas denote the prior $N(\phi\,,\ \tau\,)$ where $\phi = 12$ and $\tau$, the standard deviation, is 2.5.

$$\Rightarrow p(\mu) = \frac{1}{\tau\sqrt{2\pi}} \cdot \exp\left\{-\frac{(\mu-\phi)^2}{2\tau^2}\right\}$$

Now, let's say we have a method of measurement that gives a value $x$, which is normally distributed $N(\mu,\ \sigma\,)$ , where $\mu$ is the true mean concentration and $\sigma$ is the standard deviation.

Let's further assume that the method has been validated to provide a standard deviation of about 0.1 percentage points ➔ $x \sim N(\mu,\ 0.1\ )$

For $n_1$ measurements $x_{11}, \ldots , x_{1n_1}$ on the first seizure we obtain the likelihood function

$$L\left(\mu_1 \middle| x_{11},\ldots,x_{1n_1}\right) = \prod_{i=1}^{n_1} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{-\frac{(x_{1i}-\mu_1)^2}{2\sigma^2}\right\} =$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^2 \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(x_{1i}-\mu_1)^2\right\}$$

This can be shown to be

$$L\left(\mu_1 \middle| x_{11},\dots,x_{1n_1}\right) = g_1\left(x_{11},\dots,x_{1n_1},\sigma\right)\cdot \exp\left\{-\frac{n_1}{2\sigma^2}\left(\overline{x_1}-\mu_1\right)^2\right\} =$$

$$= g_1 \cdot \exp\left\{-\frac{n_1}{2\sigma^2}\left(\overline{x_1}-\mu_1\right)^2\right\}$$

where the function $g_1$ does not depend on $\mu_1$.

Analogously, for $n_2$ measurements $x_{21}, \dots , x_{2n_2}$ on the second seizure we obtain the likelihood function

$$L\left(\mu_2 \middle| x_{21},\dots,x_{2n_2}\right) = g_2\left(x_{21},\dots,x_{2n_2},\sigma\right)\cdot \exp\left\{-\frac{n_2}{2\sigma^2}\left(\overline{x_2}-\mu_2\right)^2\right\} =$$

$$= g_2 \cdot \exp\left\{-\frac{n_2}{2\sigma^2}\left(\overline{x_2}-\mu_1\right)^2\right\}$$

Hence, the Bayes factor is

$$B=\frac{\int_{-\infty}^{\infty} g_1\cdot\exp\left\{-\frac{n_1}{2\sigma^2}\left(\bar{x}_1-\mu\right)^2\right\}\cdot g_2\cdot\exp\left\{-\frac{n_2}{2\sigma^2}\left(\bar{x}_2-\mu\right)^2\right\}\cdot\frac{1}{\tau\sqrt{2\pi}}\cdot\exp\left\{-\frac{(\mu-\phi)^2}{2\tau^2}\right\}d\mu}{\int_{-\infty}^{\infty} g_1\cdot\exp\left\{-\frac{n_1}{2\sigma^2}\left(\bar{x}_1-\mu\right)^2\right\}\cdot\frac{1}{\tau\sqrt{2\pi}}\cdot\exp\left\{-\frac{(\mu-\phi)^2}{2\tau^2}\right\}d\mu\cdot\int_{-\infty}^{\infty} g_2\cdot\exp\left\{-\frac{n_2}{2\sigma^2}\left(\bar{x}_2-\mu\right)^2\right\}\cdot\frac{1}{\tau\sqrt{2\pi}}\cdot\exp\left\{-\frac{(\mu-\phi)^2}{2\tau^2}\right\}d\mu}=$$

$$=\frac{\int_{-\infty}^{\infty}\exp\left\{-\frac{n_1}{2\sigma^2}\left(\bar{x}_1-\mu\right)^2\right\}\cdot\exp\left\{-\frac{n_2}{2\sigma^2}\left(\bar{x}_2-\mu\right)^2\right\}\cdot\exp\left\{-\frac{(\mu-\phi)^2}{2\tau^2}\right\}d\mu}{\frac{1}{\tau\sqrt{2\pi}}\cdot\int_{-\infty}^{\infty}\exp\left\{-\frac{n_1}{2\sigma^2}\left(\bar{x}_1-\mu\right)^2\right\}\cdot\exp\left\{-\frac{(\mu-\phi)^2}{2\tau^2}\right\}d\mu\cdot\int_{-\infty}^{\infty}\exp\left\{-\frac{n_2}{2\sigma^2}\left(\bar{x}_2-\mu\right)^2\right\}\cdot\exp\left\{-\frac{(\mu-\phi)^2}{2\tau^2}\right\}d\mu}$$

*A little bit tedious to sort out*

# Simplification when comparing two normal means

Express the hypotheses as

$$H_0 : \mu_1 - \mu_2 = \delta = 0$$
$$H_1 : \mu_1 - \mu_2 = \delta \neq 0$$

Above, it was shown that $L\left(\mu_1 \middle| x_{11}, \ldots, x_{1n_1}\right) = g_1 \cdot \exp\left\{ -\frac{n_1}{2\sigma^2}\left(\overline{x_1} - \mu_1\right)^2 \right\}$

If we “reduce” the sample $x_{11}, \ldots, x_{1n_1}$ to its sample mean, i.e. $\overline{x_1}$
the likelihood function for $\mu_1$ becomes

$$L\left(\mu_1 \middle| \overline{x_1}\right) = \frac{1}{\left(\sigma / \sqrt{n_1}\right)\sqrt{2\pi}} \cdot \exp\left\{ -\frac{\left(\overline{x_1} - \mu_1\right)^2}{2 \cdot \left(\sigma^2 / n_1\right)} \right\}$$

since $\overline{x_1} \sim N\left( \text{mean} = \mu_1, \text{standard deviation} = \frac{\sigma}{\sqrt{n_1}} \right)$

Note that

$$L_1\left(\mu_1 \middle| x_{11},\ldots,x_{1n_1}\right) \propto \exp\left\{-\frac{n_1}{2\sigma^2}\left(\overline{x_1}-\mu_1\right)^2\right\}$$

$$L\left(\mu_1 \middle| \overline{x_1}\right) = \propto \exp\left\{-\frac{\left(\overline{x_1}-\mu_1\right)^2}{2\cdot\left(\sigma^2/n_1\right)}\right\} = \exp\left\{-\frac{n_1}{2\sigma^2}\left(\overline{x_1}-\mu_1\right)^2\right\}$$

Both likelihood functions are proportional to a common essential part, i.e. the part that contains $\mu_1$.

This is due to that the sample mean is a *sufficient statistic* for the population mean.

Analogously

$$L\left(\mu_2 \middle| \overline{x_2}\right) = \propto \exp\left\{-\frac{n_2}{2\sigma^2}\left(\overline{x_2}-\mu_1\right)^2\right\}$$

Now, for independent samples $\overline{x_1} - \overline{x_2} \sim N\left( \mu_1 - \mu_2 , \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$

With $\mu_1 - \mu_2 = \delta$ and $\sigma_1 = \sigma_2 = \sigma$ we get

$$\overline{x_1} - \overline{x_2} \sim N\left( \delta, \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

$\overline{x_1} - \overline{x_2}$ is a sufficient statistic for $\mu_1 - \mu_2 = \delta$ and “reducing” the two data sets to $\overline{x_1} - \overline{x_2}$ gives the likelihood function for $\delta$ as

$$L\left(\delta \mid \overline{x_1} - \overline{x_2}\right) = \frac{1}{\sigma \sqrt{1/n_1 + 1/n_2}\,\sqrt{2\pi}} \cdot \exp\left\{ -\frac{\left(\overline{x_1} - \overline{x_2} - \delta\right)^2}{2 \cdot \sigma^2 \cdot \left(1/n_1 + 1/n_2\right)} \right\}$$

The Bayes factor becomes

$$B=\frac{L\left(\delta=0\middle|\overline{x_1}-\overline{x_2}\right)}{\int_{-\infty}^{\infty}L\left(\delta\middle|\overline{x_1}-\overline{x_2}\right)\cdot p(\delta)d\delta}=$$

$$=\frac{\frac{1}{\sigma\sqrt{1/n_1+1/n_2}\sqrt{2\pi}}\cdot\exp\left\{\frac{\left(\overline{x_1}-\overline{x_2}\right)^2}{2\cdot\sigma^2\cdot\left(1/n_1+1/n_2\right)}\right\}}{\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{1/n_1+1/n_2}\sqrt{2\pi}}\cdot\exp\left\{\frac{\left(\overline{x_1}-\overline{x_2}-\delta\right)^2}{2\cdot\sigma^2\cdot\left(1/n_1+1/n_2\right)}\right\}\cdot\frac{1}{\tau\sqrt{2}\sqrt{2\pi}}\cdot\exp\left(-\frac{\delta^2}{4\tau^2}\right)d\delta}=$$

$$=\frac{\exp\left\{\frac{\left(\overline{x_1}-\overline{x_2}\right)^2}{2\cdot\sigma^2\cdot\left(1/n_1+1/n_2\right)}\right\}}{\frac{1}{\tau\sqrt{2}\sqrt{2\pi}}\cdot\int_{-\infty}^{\infty}\exp\left\{\frac{\left(\overline{x_1}-\overline{x_2}-\delta\right)^2}{2\cdot\sigma^2\cdot\left(1/n_1+1/n_2\right)}\right\}\cdot\exp\left(-\frac{\delta^2}{4\tau^2}\right)d\delta}$$

A little bit easier to sort out. Can be shown to be

$$B=\sqrt{1+\frac{2\cdot n_1\cdot n_2\cdot\tau^2}{\sigma^2\cdot\left(n_1+n_2\right)}}\cdot\exp\left\{-\frac{1}{2}\cdot\left(\frac{1}{\sigma^2\cdot\left(1/n_1+1/n_2\right)}-\frac{1}{\sigma^2\cdot\left(1/n_1+1/n_2\right)+2\tau^2}\right)\cdot\left(\overline{x_1}-\overline{x_2}\right)^2\right\}$$

Inserting the values of $\sigma$ (= 0.1) and $\tau$ (= n2.5) we obtain

$$B=\sqrt{1+\frac{12.5\cdot n_1\cdot n_2}{0.01\cdot\left(n_1+n_2\right)}}\cdot\exp\left\{-\frac{1}{2}\cdot\left(\frac{1}{0.01\cdot\left(1/n_1+1/n_2\right)}-\frac{1}{0.01\cdot\left(1/n_1+1/n_2\right)+12.5}\right)\cdot\left(\overline{x}_1-\overline{x}_2\right)^2\right\}$$

Let’s say we have obtained the sample means 18.1 % and 18.2 % and that our prior odds for $H_0$ are 1 (non-informative prior).

| $n_1$ | $n_2$ | $B$ | Posterior odds | Classical *P*-value |
|---|---|---|---|---|
| 2 | 2 | 21.46 | 21.46 | 0.3173 |
| 2 | 3 | 21.27 | 21.27 | 0.2733 |
| 5 | 5 | 16.03 | 16.03 | 0.1138 |
| 5 | 7 | 14.05 | 14.05 | 0.0877 |
| 10 | 10 | 6.49 | 6.49 | 0.0253 |
| 30 | 30 | 0.08 | 0.08 | 0.0001 |
| 100 | 100 | 0.00 | 0.00 | 0.0000 |

# Comparison of multivariate means

Assume there are two multivariate populations with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ respectively. The statistical model is that each population can be described in terms of a *p*-dimensional random vector

$$\boldsymbol{X} = (X_1, \dots , X_p)^{\mathrm{T}}$$

with a probability density function (or probability mass function) depending on the mean $\boldsymbol{\mu}$ : $f\,(\mathbf{x} \mid \boldsymbol{\mu}\,) = f(x_1, \dots , x_p \mid \boldsymbol{\mu}\,)$

The pair of hypotheses for comparing the two (vector) means is:

$$H_0\text{: } \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$
$$H_1\text{: } \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

The random variation can be decomposed into (at least) two *levels*:

1. Within-variation: Accounts for the variability in $\boldsymbol{X}$ given a specific value of $\boldsymbol{\mu}$. This is modelled by $f\,(\boldsymbol{x} \mid \boldsymbol{\mu})$.
2. Between-variation: Accounts for the uncertainty about the true value of $\boldsymbol{\mu}$. This can be modelled in terms of a prior density for $\boldsymbol{\mu}$, i.e. $p\,(\boldsymbol{\mu})$, $\boldsymbol{\mu} \in \mathbf{M}$

This is referred to as a *two-level model* for the variation in which the variation at different levels is stochastic (cf. analysis of variance with random effects).

The Bayes factor for assessing the pair of hypothesis ($H_0$, $H_1$) from two sets of data, one from each population (*Data set* 1, *Data set* 2) is

$$B = \frac{\int_{\mathbf{M}} L\left(\boldsymbol{\mu} \middle| Data\ set\ 1\right) \cdot L\left(\boldsymbol{\mu} \middle| Data\ set\ 2\right) \cdot p\left(\boldsymbol{\mu}\right) d\boldsymbol{\mu}}{\int_{\mathbf{M}} L\left(\boldsymbol{\mu} \middle| Data\ set\ 1\right) \cdot p\left(\boldsymbol{\mu}\right) d\boldsymbol{\mu} \cdot \int_{\mathbf{M}} L\left(\boldsymbol{\mu} \middle| Data\ set\ 2\right) \cdot p\left(\boldsymbol{\mu}\right) d\boldsymbol{\mu}}$$

i.e. the same structure as for univariate data

When (vector) means are compared, sufficient statistics from the data sets are the two sample (vector) means:

$$\textit{Data set } 1:$$

$$\boldsymbol{x}_1 = \left(\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}\right) \quad ; \quad \boldsymbol{X}_{1i} \sim f\left(\boldsymbol{x}_{1j} \middle| \boldsymbol{\mu}_1\right)$$

$$\text{Sufficient statistic}: \quad \overline{\boldsymbol{X}}_{1,n_1} \sim f\left(\overline{\boldsymbol{x}}_{1\cdot} \middle| \boldsymbol{\mu}_1\right)$$

$$\textit{Data set } 2:$$

$$\boldsymbol{x}_2 = \left(\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n_2}\right) \quad ; \quad \boldsymbol{X}_{2i} \sim f\left(\boldsymbol{x}_{2j} \middle| \boldsymbol{\mu}_2\right)$$

$$\text{Sufficient statistic}: \quad \overline{\boldsymbol{X}}_{2,n_2} \sim f\left(\overline{\boldsymbol{x}}_{2\cdot} \middle| \boldsymbol{\mu}_2\right)$$

$$\Rightarrow L\left(\boldsymbol{\mu}_i \middle| \textit{Data set } i\right) = k\left(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}\right) \cdot f\left(\overline{\boldsymbol{x}}_{i\cdot} \middle| \boldsymbol{\mu}_i\right)$$

This in turn gives

$$B = \frac{\int_{\mathbf{M}} f\left(\overline{\boldsymbol{x}}_1 \middle| \boldsymbol{\mu}\right) \cdot f\left(\overline{\boldsymbol{x}}_2 \middle| \boldsymbol{\mu}\right) \cdot p\left(\boldsymbol{\mu}\right) d\boldsymbol{\mu}}{\int_{\mathbf{M}} f\left(\overline{\boldsymbol{x}}_1 \middle| \boldsymbol{\mu}\right) \cdot p\left(\boldsymbol{\mu}\right) d\boldsymbol{\mu} \cdot \int_{\mathbf{M}} f\left(\overline{\boldsymbol{x}}_2 \middle| \boldsymbol{\mu}\right) \cdot p\left(\boldsymbol{\mu}\right) d\boldsymbol{\mu}}$$

Now, assume that both within- and between-variation is normal:

$$\boldsymbol{X}_{ij}\,\big|\,\boldsymbol{\mu_i},\boldsymbol{\Sigma}_i \sim N_p\left(\boldsymbol{\mu_i};\boldsymbol{\Sigma}_i\right)$$

where $\boldsymbol{\Sigma}_i$ is the (variance-)covariance matrix.

This gives

$$\overline{\boldsymbol{X}}_i \sim N_p\left(\boldsymbol{\mu_i};n_i^{-1}\boldsymbol{\Sigma}_i\right) \quad \Rightarrow$$

$$f\left(\overline{\boldsymbol{x}}_i\,\big|\,\boldsymbol{\mu_i}\left[,\boldsymbol{\Sigma}_i\right]\right)=\left(2\pi\cdot\det\left(n_i^{-1}\boldsymbol{\Sigma}_i\right)\right)^{-p/2}\cdot\exp\left\{-\left(\overline{\boldsymbol{x}}_i-\boldsymbol{\mu_i}\right)^{\mathrm{T}}\cdot n_i\boldsymbol{\Sigma}_i^{-1}\cdot\left(\overline{\boldsymbol{x}}_i-\boldsymbol{\mu_i}\right)\right\}$$

## *Full Bayesian approach:*

$\boldsymbol{\mu}_i \sim N_p\left(\boldsymbol{\theta}_i;\boldsymbol{\Omega}_i\right)$ Prior distribution is multivariate normal with hyper mean $\boldsymbol{\theta}_i$ and hyper covariance matrix $\boldsymbol{\Omega}_i$

$\boldsymbol{\Sigma}_i \sim W_p^{-1}\left(d_i;\mathbf{T}^{-1}\right)$ Prior distribution is p-dimensional Inverse Wishart with $d_i$ degrees of freedom and precision matrix $\mathbf{T}^{-1}$

# *Simplified approaches*

Assume that the covariance matrix $\mathbf{\Sigma}$ is common to both populations. This is a most reasonable assumption when $\mathbf{\Sigma}$ represents measurement uncertainty.

It is also a reasonable assumption when the variation in a population does not depend substantially on the mean (*read normal distributions*) .

The comparability of the populations is self-understood – why would we otherwise compare the means?

Assume further that the prior distribution of $\boldsymbol{\mu}$ is common to both populations – why could that be reasonable?

Instead of applying the full Bayesian approach with these simplifications we may use estimates from sufficient background data of

- $\boldsymbol{\Sigma}$ (to avoid the use of an Inverse Wishart distribution)
- $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ (*empirical Bayes* approach)

Assume we have a background data set of *N* observed vectors from each of *M* different populations

$\{\boldsymbol{y}_{l,r}\}_{l=1,\ldots,M;r=1,\ldots,N}$ where

$\boldsymbol{y}_{l,r} = \left(y_{l,r,1},\ldots,y_{l,r,p}\right)^{\mathrm{T}}$ is assumed to be $N_p\left(\boldsymbol{\mu};\boldsymbol{\Sigma}\right)$ and

$\boldsymbol{\mu}$ is assumed to be $N_p\left(\boldsymbol{\theta};\boldsymbol{\Omega}\right)$

An estimate of $\boldsymbol{\Sigma}$ is then obtained as

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{(N-1)\cdot M}\sum_{l=1}^{M}\sum_{r=1}^{N}\left(\boldsymbol{y}_{l,r}-\bar{\boldsymbol{y}}_{l,\cdot}\right)\cdot\left(\boldsymbol{y}_{l,r}-\bar{\boldsymbol{y}}_{l,\cdot}\right)^{\mathrm{T}} = \frac{\mathbf{S}_w}{(N-1)\cdot M}$$

where $\bar{\boldsymbol{y}}_{l,\cdot} = N^{-1}\sum_{r=1}^{N}\boldsymbol{y}_{l,r}$

The analogue of *SSE* (residual sum of squares) for multivariate data.

An estimate of $\boldsymbol{\Omega}$ is

$$\hat{\boldsymbol{\Omega}} = \frac{1}{M-1}\sum_{l=1}^{M}\left(\bar{\boldsymbol{y}}_{l,\cdot}-\bar{\boldsymbol{y}}_{\cdot,\cdot}\right)\cdot\left(\bar{\boldsymbol{y}}_{l,\cdot}-\bar{\boldsymbol{y}}_{\cdot,\cdot}\right)^{\mathrm{T}} - \frac{\hat{\boldsymbol{\Sigma}}}{N} = \frac{\mathbf{S}_{*}}{M-1} - \frac{\mathbf{S}_w}{N^2(M-1)}$$

where $\bar{\boldsymbol{y}}_{\cdot,\cdot} = (N\cdot M)^{-1}\sum_{l=1}^{M}\sum_{r=1}^{N}\boldsymbol{y}_{l,r}$

…and an estimate of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \bar{\boldsymbol{y}}_{\cdot,\cdot} = (N\cdot M)^{-1}\sum_{l=1}^{M}\sum_{r=1}^{N}\boldsymbol{y}_{l,r}$$

For obtaining the Bayes factor

$$B=\frac{\int_{\mathbf{M}} f\left(\bar{\boldsymbol{x}}_1\middle|\boldsymbol{\mu}\right)\cdot f\left(\bar{\boldsymbol{x}}_2\middle|\boldsymbol{\mu}\right)\cdot p(\boldsymbol{\mu})d\boldsymbol{\mu}}{\int_{\mathbf{M}} f\left(\bar{\boldsymbol{x}}_1\middle|\boldsymbol{\mu}\right)\cdot p(\boldsymbol{\mu})d\boldsymbol{\mu}\cdot\int_{\mathbf{M}} f\left(\bar{\boldsymbol{x}}_2\middle|\boldsymbol{\mu}\right)\cdot p(\boldsymbol{\mu})d\boldsymbol{\mu}}$$

this implies

$$\bar{\boldsymbol{X}}_i \text{ is assumed to be } N_p\left(\boldsymbol{\mu};n_i^{-1}\hat{\boldsymbol{\Sigma}}\right) \Rightarrow$$

$$f\left(\bar{\boldsymbol{x}}_i\middle|\boldsymbol{\mu}\right)=\left(2\pi\cdot\det\left(n_i^{-1}\hat{\boldsymbol{\Sigma}}\right)\right)^{-p/2}\cdot\exp\left\{-\left(\bar{\boldsymbol{x}}_i-\boldsymbol{\mu}\right)^{\mathrm{T}}\cdot n_i\hat{\boldsymbol{\Sigma}}^{\mathbf{-1}}\cdot\left(\bar{\boldsymbol{x}}_i-\boldsymbol{\mu}\right)\right\}$$

and

$$\boldsymbol{\mu} \text{ is assumed to be } N_p\left(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\Omega}}\right)\Rightarrow$$

$$p(\boldsymbol{\mu})=\left(2\pi\cdot\det\left(\hat{\boldsymbol{\Omega}}\right)\right)^{-p/2}\cdot\exp\left\{-\left(\boldsymbol{\mu}-\hat{\boldsymbol{\theta}}\right)^{\mathrm{T}}\cdot\hat{\boldsymbol{\Omega}}^{\mathbf{-1}}\cdot\left(\boldsymbol{\mu}-\hat{\boldsymbol{\theta}}\right)\right\}$$

Aitken CGG & Lucy D (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics* **53**(1): 109-122

– use the full likelihoods (i.e. not the densities of the sufficient means) in

$$B = \frac{\int_{\mathbf{M}} L\left(\boldsymbol{\mu}\middle| Data\ set\ 1\right)\cdot L\left(\boldsymbol{\mu}\middle| Data\ set\ 2\right)\cdot p\left(\boldsymbol{\mu}\right)d\boldsymbol{\mu}}{\int_{\mathbf{M}} L\left(\boldsymbol{\mu}\middle| Data\ set\ 1\right)\cdot p\left(\boldsymbol{\mu}\right)d\boldsymbol{\mu}\cdot\int_{\mathbf{M}} L\left(\boldsymbol{\mu}\middle| Data\ set\ 2\right)\cdot p\left(\boldsymbol{\mu}\right)d\boldsymbol{\mu}}$$

i.e.

$$L\left(\boldsymbol{\mu}\middle| Data\ set\ i\right)=\prod_{j=1}^{n_i} f\left(\boldsymbol{x}_{ij}\middle|\boldsymbol{\mu}\right)=\prod_{j=1}^{n_i}\left(2\pi\cdot\det\left(\hat{\boldsymbol{\Sigma}}\right)\right)^{-p/2}\cdot\exp\left\{-\left(\boldsymbol{x}_{ij}-\boldsymbol{\mu}\right)^{\mathrm{T}}\cdot\hat{\boldsymbol{\Sigma}}^{\mathbf{-1}}\cdot\left(\boldsymbol{x}_{ij}-\boldsymbol{\mu}\right)\right\}$$

– derive explicit formulas for the Bayes factor when

- prior distribution of $\boldsymbol{\mu}$ is normal
- prior distribution of $\boldsymbol{\mu}$ is estimated by a kernel density

*From the “real” world*

# Comparison of amphetamine seizures

- Issue:
  - To evaluate findings from two seizures of amphetamine against the propositions

$\boldsymbol{H_m}$ : The two seizures originate from the same precipitation batch
$\boldsymbol{H_a}$ : The two seizures originate from different precipitation batches

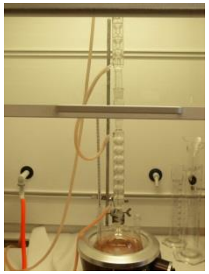
- Framework constraint:
  - The findings should be based on impurity profiling for a set of 30 impurities agreed on in a European cooperation program

- Problem:
  - Today’s “standard of evaluation” is manual inspection of profiles – time consuming, especially since in a case there are usually pairwise comparisons of several seizures

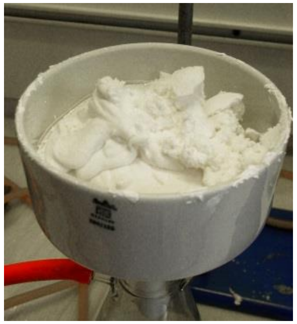
# The production of amphetamine

- Choose a recipe

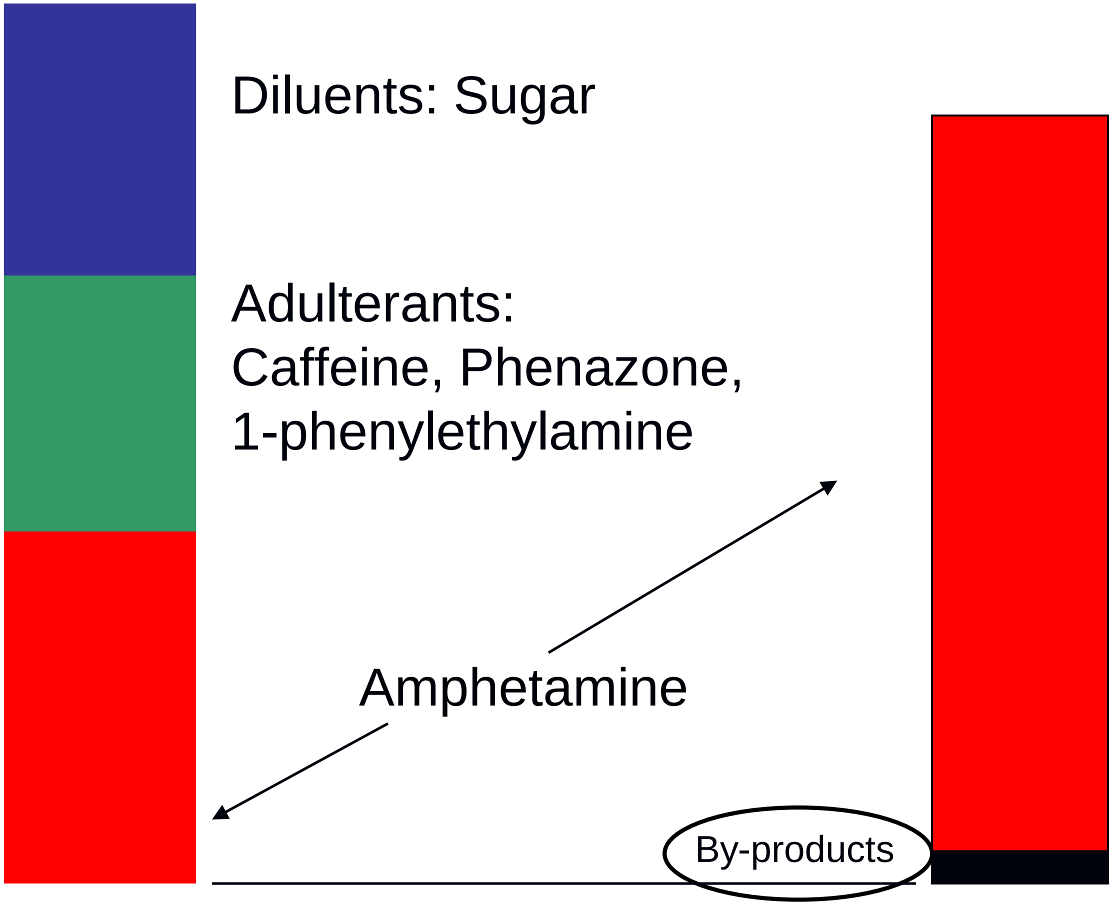
- Produce amphetamine oil



- Precipitate the amphetamine powder ⇒ ***<u>precipitation batch</u>***



# *Impurities (by-products) come with production:*



Diluents: Sugar

Adulterants:
Caffeine, Phenazone,
1-phenylethylamine

Amphetamine

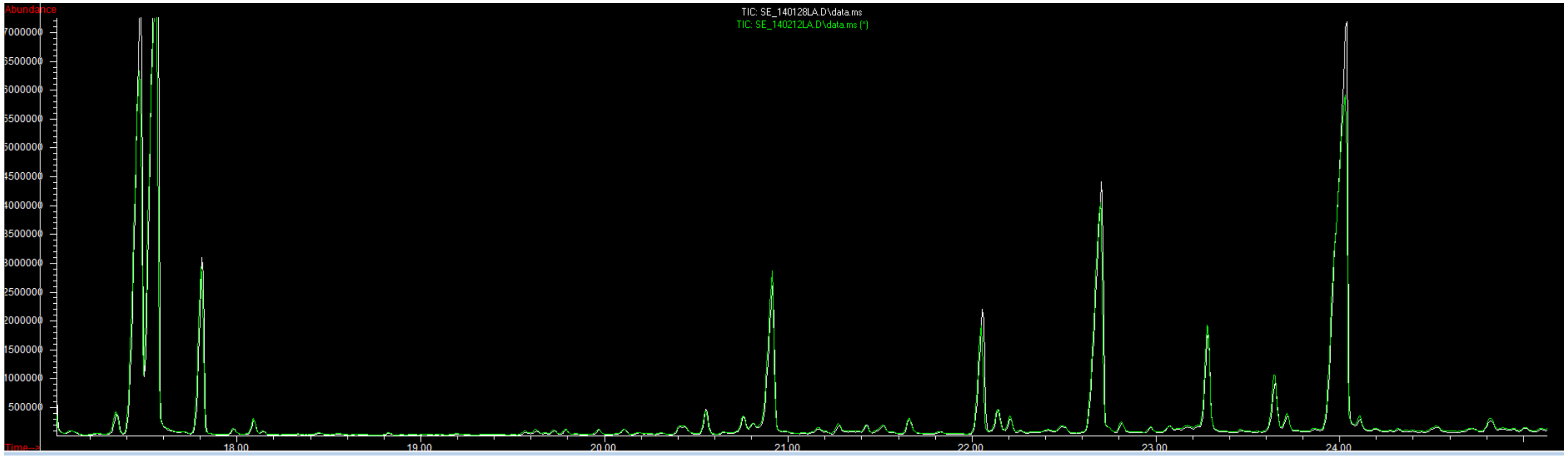
By-products

# Manual inspection/comparison of impurity profiles

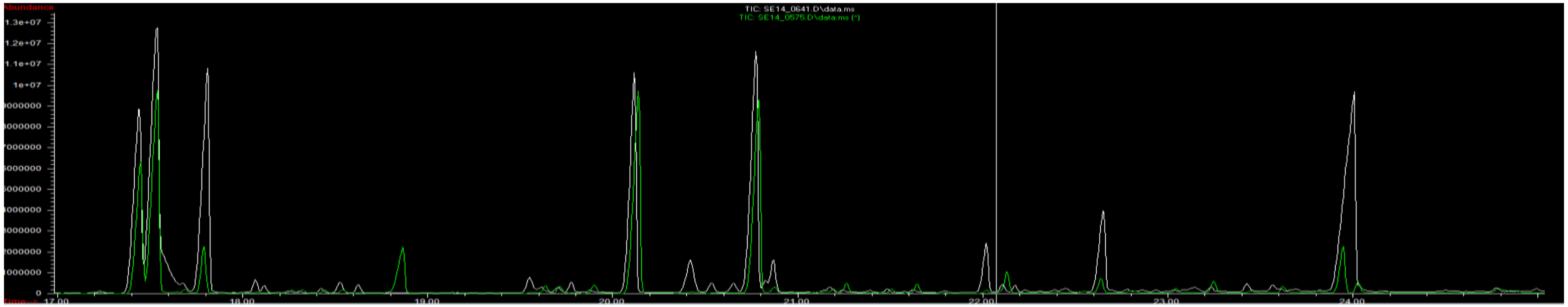
Inspecting overlaid *chromatograms*

**Green** and **violet** peaks representing equal or different profiles?

I.



II.



# Evaluative statement

The National Forensic Centre (NFC) uses a common (verbal) scale of conclusions based on intervals of likelihood ratios (LR) [or Bayes factors].

“The findings from the comparison of the two seizures...

| | |
|---|---|
| …*support extremely strongly*… | $(10^6 \le \mathrm{LR})$ |
| …*support strongly*… | $(6000 \le \mathrm{LR} < 10^6)$ |
| …*support*… | $(100 \le \mathrm{LR} < 6000)$ |
| …*support to some extent*… | $(6 \le \mathrm{LR} < 100)$ |
| … that the two seizures originate from the same precipitation batch | |
| …*support neither* that the two seizures originate from the same precipitation batch *nor* that they originate from different precipitation batches | $(1/6 < \mathrm{LR} < 6)$ |
| …*support extremely strongly/strongly//to some extent* that the two seizures originate from different precipitation batches | $\mathrm{LR} \le 1/6$, $1/100$, $1/6000$, $1/10^6$ |

Nordgaard et al, *Law, probability and Risk* , 2012

# How do we know that the correct magnitude of the likelihood ratio has been assigned?

Always address the likelihood ratio when evaluating:

$$LR = \frac{\Pr\left(E \middle| H_m\right)}{\Pr\left(E \middle| H_a\right)}$$

with *E* representing the findings from the inspection/comparison

The probabilities Pr(*E* | $H_m$ ) and Pr(*E* | $H_a$) are assigned from experience and knowledge…

…but are by necessity objected to a substantial amount of subjectivity.

Manual comparisons should always be made, but it would be good to have the evaluation assisted by an objective probabilistic model.

# Applying a probabilistic model to the findings

Monitoring 30 peaks in a chromatogram: *Multivariate inference*

*In theory:* Model the multivariate distributions of the ***peak areas*** under each of the propositions and compute the Bayes factor for the comparison:

With $\boldsymbol{x}$ = observed vector of peak areas for seizure 1

$\boldsymbol{y}$ = observed vector of peak areas for seizure 2

$\boldsymbol{\theta}$ **=** true mean vector of peak areas

$p(\boldsymbol{\theta})$ = prior density for $\boldsymbol{\theta}$

$$\boxed{B = \frac{\int f\left(\mathbf{x}\middle|\boldsymbol{\theta}\right)\cdot f\left(\mathbf{y}\middle|\boldsymbol{\theta}\right)\cdot p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int f\left(\mathbf{x}\middle|\boldsymbol{\theta}\right)p(\boldsymbol{\theta})d\boldsymbol{\theta}\times\int f\left(\mathbf{y}\middle|\boldsymbol{\theta}\right)p(\boldsymbol{\theta})d\boldsymbol{\theta}}}$$

# Simplifying…

*Assume…*

- the peak area variation between samples from the same precipitation batch – the *within-variation* – is normal

- the peak area variation between samples from different precipitation batches – the *between-variation* – is not (necessarily) normal

*Graphical description of experimental data (see below) do not contradict these assumptions.*

With these assumptions Aitken & Lucy (*Applied Statistics*, 2004) give explicit functions for the numerator and denominator of the Bayes factor.

$$\boxed{B = \frac{f_m\left(\boldsymbol{x}, \boldsymbol{y} \middle| \boldsymbol{\mu}, U, C\right)}{f_a\left(\boldsymbol{x}, \boldsymbol{y} \middle| \boldsymbol{\mu}, U, C\right)}}$$

where $f_m$ and $f_a$ are density functions obtained by integrating multivariate normal distributions of $\boldsymbol{x}$ and $\boldsymbol{y}$ with a kernel estimate of the prior density of $\boldsymbol{\theta}$.

## *The Multivariate Kernel Likelihood Ratio (MVK)*

*B* will then depend on

- the mean vector of the prior distribution of peak areas: $\boldsymbol{\mu}$
- the variance-covariance matrix for the within-variation: $U$
- the variance-covariance matrix for the between-variation: $C$

*Taking these as hyperparameters we need stable estimates!*

*Could also be assigned prior distributions, but that is not done here.*

# Experimental study

- Five different recipes were chosen
- Five different oils were produced with each recipe
- Three precipitation batches per oil were taken
- *One precipitation was not successful*

⇒ In total 74 different precipitation batches

From each precipitation batch 4-9 replicate samples were taken for profiling.

4-5 replicates were taken when the batch was fresh. For a number of batches another 4 replicates were taken upon storage (in freezer) and airing.

Ignoring the full hierarchical structure of the data…

Two levels of variation:

- within precipitation (between replicates from the same batch)
- between precipitations (sum of variance components from recipe, oil and precipitation)

# Harmonising and variance stabilizing transformations:

- When profiling using GC/MS an *internal standard* is run in each sequence

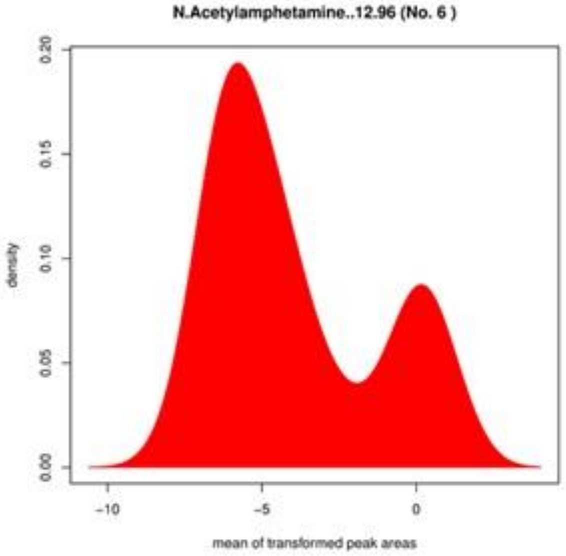
- Each sample is diluted before analysis so that the linear range of the instrument is not exceeded. How much depends on the content of by-products in each sample. The degree of dilution is represented as a *sample multiplier* (factor)

- Moreover, the dry concentrations of the precipitation batches will vary substantially

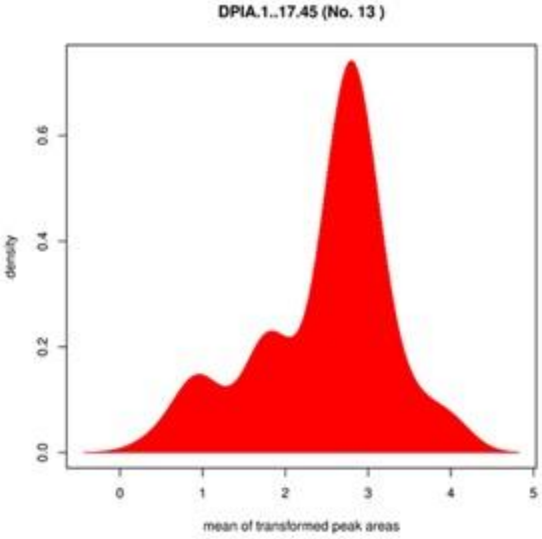
To account for these observable sources of variation and to allay the effect of outliers (w.r.t. assumed normal within-variation) the original peak area (*PA*) is transformed as

$$\boxed{pa = \log\left(\frac{PA}{PA_{\text{internal standard}}} \times \frac{\text{sample multiplier}}{\text{dry concentration}}\right)}$$

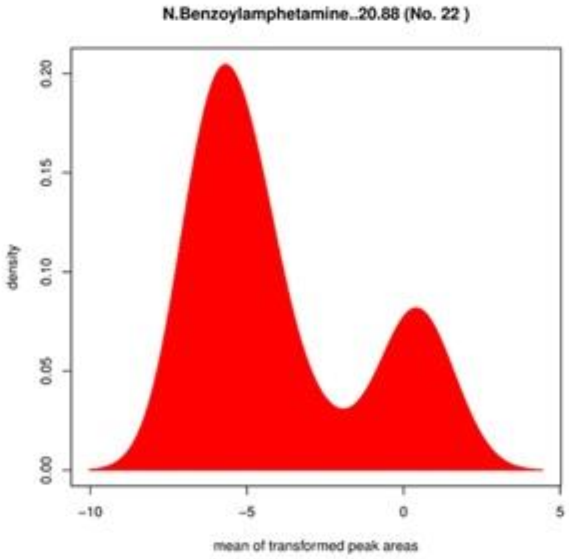
# Between-variation for some of the impurities:



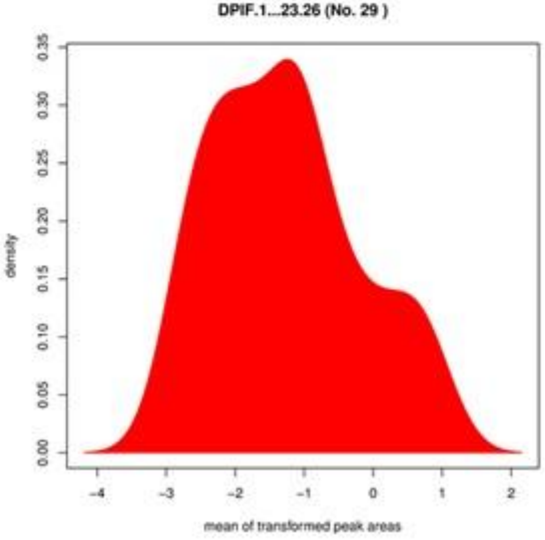
N.Acetylamphetamine..12.96 (No. 6 )



DPIA.1..17.45 (No. 13 )



N.Benzoylamphetamine..20.88 (No. 22 )



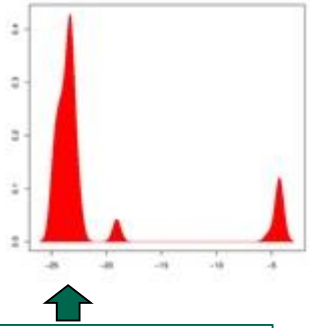
DPIF.1...23.26 (No. 29 )

# Data reduction

With 74 different batches and sometimes only 4 replicate measurements on each it is not numerically feasible to estimate variance-covariance matrices for a 30-dimensional vector.

Several of the impurities could however be discarded by “expert reasons”:

- Some impurities are known to be volatile
- In experimental (and casework) data some impurities show very low entropy (are seldom detected)



zero peak area

Upon such considerations 18 impurities remain – still too many!

# Solution used here:

Expert choice of 12 impurities, from experience known to vary between seizures indisputably from different sources.

| |
|---|
| N.Acetylamphetamine |
| N.Formylamphetamine |
| Benzylamphetamine |
| DPIA.1 |
| alfa.Methyldiphenetyletylamine |
| N.Benzoylamphetamine |
| Unknown.B2 |
| X2.6.Dimethyl.3.5.diphenylpyridine |
| X2.4.Dimetyl.3.5.diphenylpyridine |
| Pyridine.7.and.14 |
| X2.6.Diphenyl.3.4.dimethylpyridine |
| DPIF.1 |

*Of interest primarily for chemists and people involved in the profiling cooperation projects*

Works numerically, but stable estimates may still be hard to obtain.

# Investigating the stability by resampling

**Resample** the experimental data by drawing batch samples with replacement within batches.

Compute estimates (MLE) of *U* and *C* (from original data and resampled data).

$$\Rightarrow \hat{U}, \hat{C} \text{ and } \left\{U^*\right\}, \left\{C^*\right\}$$

Assess the traces and determinants of $\{U^*\}$ and $\{C^*\}$ with respect to relative standard deviation over the resamples for three size of set of impurities:

(i) 3 impurities (ii) 6 impurities (iii) 12 impurities (as was used)

## Results for 500 resamples:

| | **3 impurities** | **6 impurities** | **12 impurities** |
|---|---|---|---|
| **tr(U), %σ** | 4.9 | 4.8 | 6.0 |
| **det(U), %σ** | 18.4 | 19.4 | 21.9 |
| **tr(C), %σ** | 1.3 | 1.7 | 1.5 |
| **det(C), %σ** | 5.1 | 5.5 | 9.2 |

# Does it work for the evaluation?

How can we assess whether the suggested method works or not?

*In general:*

- The false positive rate should be very low (*Do we accept non-zero rates?)*
- The false negative rate should be very low (*Needs to be of the same magnitude as the false positive rate?*)

> “It is better that ten guilty escape than one innocent suffer.”
>
> *Lord William Blackstone*

- Evidence strongly supporting a proposition should generate a large Bayes factor and vice versa

Recall the NFC scale of conclusions:

| ***Level*** | *Interval of B* | ***Level*** | *Interval of B* |
|---|---|---|---|
| **+4** | $10^6 \leq B$ | **–1** | $1/100 < B \leq 1/6$ |
| **+3** | $6000 \leq B < 10^6$ | **–2** | $1/6000 < B \leq 1/100$ |
| **+2** | $100 \leq B < 6000$ | **–3** | $1/10^6 < B \leq 1/6000$ |
| **+1** | $6 \leq B < 100$ | **–4** | $B \leq 1/10^6$ |
| 0 | $1/6 < B < 6$ | | |

Hence, what matters here is

- whether a level at the "correct" side of the scale is attained **and**
- whether findings on seizures with *a common source* generally lead to high levels (+3, +4); and whether findings on seizures with *different sources* generally lead to low levels ( –3, –4)

# Cross-validatory study:

$$B = \frac{f_m\left(\mathbf{x},\mathbf{y}\middle|\boldsymbol{\mu},U,C\right)}{f_a\left(\mathbf{x},\mathbf{y}\middle|\boldsymbol{\mu},U,C\right)}$$

1. (a) Leave out one precipitation batch at a time, estimate $\boldsymbol{\mu}$, $U$ and $C$ from the rest
   (b) Split the replicates of the left-out batch into two “seizures with common origin” (use different combinations for the split)
   (c) Apply the estimated “model” to each combination and compute the median Bayes factor

⇒ 74 values of *B* for assessing the false negative rate and the “strength consistency” for a common origin.

2. (a) Leave out two precipitation batches at a time (use all combinations) to be two “seizures” with different origins, estimate $\boldsymbol{\mu}$, $U$ and $C$ from the rest
   (b) Apply the “estimated model” to the two seizures, compute the Bayes factor

⇒ 2701 values of *B* for assessing the false positive rate and the “strength consistency” for different origins.

# Results:

Analysis made with R-package *comparison* (D. Lucy)

***Assessment of the false negative rate***

For the 74 (median) values of the Bayes factor, *B* for the proposition of a common origin :

| *Scale level* | **–4** | **–3** | **–2** | **–1** | **0** | **+1** | **+2** | **+3** | **+4** |
|---|---|---|---|---|---|---|---|---|---|
| *Frequency* | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 28 | 44 |

Hence, the false negative rate is 1/74 ≈ 1.4 %

Using the average Bayes factor instead of the median renders level +4 only

## *Assessment of the false positive rate*

For the 2701 values of the Bayes factor, *B* for the proposition of a common origin :

| *Scale level* | **–4** | **–3** | **–2** | **–1** | **0** | **+1** | **+2** | **+3** | **+4** |
|---|---|---|---|---|---|---|---|---|---|
| *Frequency* | 2633 | 4 | 3 | 0 | 2 | 4 | 4 | 12 | 39 |

For the two occurrences of level 0 the Bayes factor is 0.39 and 0.34 respectively.

Hence, the false positive rate is 59/2701 ≈ 2.2 %

*Disturbing:* The false positive results are dominated by high scale levels (+3 and +4)

Further inspection of the comparisons rendering levels +3 and +4 shows that

- all but 3 are comparisons of batches from the same oil (the 3 concerns two oils from the same recipe)

- the Bayes factor for the comparison is mostly of a lower magnitude when comparing samples that were stored and aired compared to the corresponding comparison of fresh samples – in a few cases the level shifted from +3 to –2

- manual comparisons of the chromatograms
    - also resulted mostly in high scale levels (+3, +4) for the fresh samples
    - resulted in lower levels (from –2 to +3) for stored and aired samples

## *Assessment of the strength consistency*

## Empirical cross-entropy, *ECE*

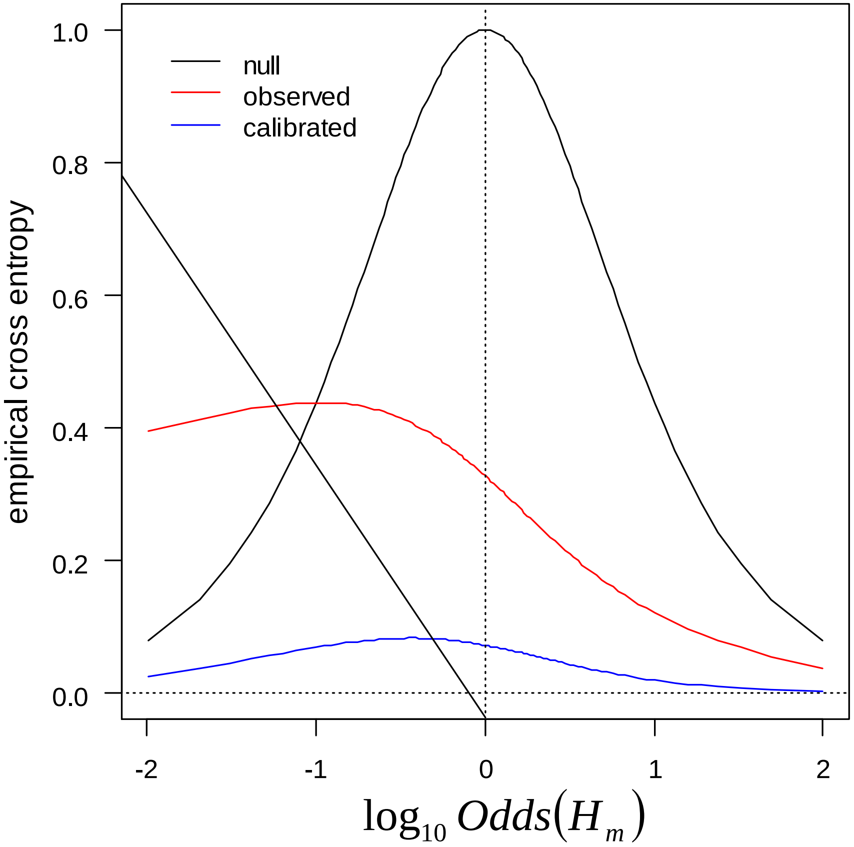
In the cross-validatory study, let

$S_m$ = The set of comparisons of samples with common origin
$N_m$ = Number of comparisons of samples with common origin (i.e. 74)
$S_a$ = The set of comparisons of samples with different origins
$N_a$ = Number of comparisons of samples with different origins (i.e. 2701)

$$
\begin{aligned}
ECE = & \frac{\Pr(H_m)}{N_m} \sum_{i \in S_m} \log_2 \left( 1 + \frac{1}{B_i \cdot Odds(H_m)} \right) + \\
& + \frac{\Pr(H_a)}{N_a} \sum_{j \in S_a} \log_2 \left( 1 + B_j \cdot Odds(H_a) \right)
\end{aligned}
$$

Ramos et al, *J.Forensic Sci*, 2013

# ***ECE* plot** (provided by R-package *comparison*)



The closer the **red** curve (observed *ECE*) is to the **blue** curve (calibrated *ECE*) the more robust are the Bayes factors that are calculated this way.

For prior odds where the **red** curve exceeds the **black** curve (representing *ECE* for non-informative evidence – *B* is always = 1), the Bayes factors can be misleading.

Hence, the Bayes factors produced with the 12 selected impurities can be misleading if the prior odds are below 0.1.

# References

Aitken C.G.G, Lucy D. (2004). Evaluation of trace evidence in the form of multivariate data. *Appl.Statist.* 53(1): 109-122.

Lindley D.V. (1977). A problem in forensic science. *Biometrika* 64(2): 207-213.

Nordgaard A, Ansell R., Drotz W., Jaeger L. (2012). Scale of conclusions for the value of evidence. *Law, probability and Risk* 11(1): 1-24.

Ramos D., Gonzales-Rodriguez J., Zadora G., Aitken C. (2013). Information-Theoretical Assessment of Likelihood Ratio Computation Methods. *J.Forensic Sci* 58(6): 1503-1518.