

Selection of samples

Anders Nordgaard

The Swedish National Laboratory of
Forensic Science (SKL)



Where/When do we sample?

- When the population (seizure, consignment) is too large to be analyzed in its entirety:
 - because of limitations in time and/or resources (personnel, money)



- When the analysis of a single unit means destruction (cf. Mavridis/Aitken)
- When seizing is equivalent to sampling (Isn't it always?)



- When the population is “infinite”

What do we sample?

- Drugs (pills, plastic bags, capsules, phials)
- Bank-notes and coins
- CD-ROMs
- Crops (suspected cannabis)

- Individuals
- Glass
- Fibres



Objectives of sampling

- To do forensic analysis in particular cases
- To establish data bases (reference material) for use with evidence evaluation
- For quality assurance reasons



How do we sample?

That's the question!



Two “general” cases

1. The population is expected to be (large and) heterogeneous
 - Difficult to make prior assumptions about population parameters
 - Sample size must usually be large (...to reflect the heterogeneity)
 - ➔ Normal approximations are valid and sample size determination can be done with the “frequentist” approach
2. The population is expected to be homogeneous
 - More easier to make prior assumption about population parameters
 - Sample size needs not to be large (e.g. if we are sure about that all elements in the population are of the same kind, we only need to sample one unit)
 - Bayesian approach to sample size determination is more attractable.



The heterogeneous case

- undesirable for forensic analysis in particular cases
- expected when data bases are to be established

Sampling of individuals, glass, fibres etc.

Should be carried out with careful use of knowledge from survey theory:

- Comparison of frame population with true population
- Choice of sampling design (simple random sampling, stratified sampling, cluster sampling,...)
- Efficient prevention and post-handling of non-response



The homogeneous case

- Main “Objective” of the current presentation
- Required for efficient sampling in daily case-work

Sampling of drug pills, bank-notes, CD-ROMs etc. for further analysis

General desire: To keep the sample size very small (5-10 units)

Sampling under experimental conditions for concluding upon proportions

General desire: To keep the sample size as small as possible



Some examples from drug sampling

1. Homogeneity expected from visual inspection and experience

Consider a case with a seizure of 5000 pills, all of the same colour (blue), form (circular) and printing (e.g. the Mitsubishi trade mark)



The forensic scientist would say “this is a seizure of Ecstasy pills”

So, what do we know about blue pills (supposed to be Ecstasy)?

Consider historical cases with blue pills

Group the cases into M clusters with respect to another parameter, e.g. the print on the pill.

Find an estimate of the *prior distribution* for the proportion θ of Ecstasy pills among blue pills.

Nordgaard A. (2006) Quantifying experience in sample size determination for drug analysis of seized drugs. *Law, Probability and Risk* **4**: 217-225



Cluster	Accumulated size of seizure	Accumulated size of sample	Number of Ecstasy pills	Number of Non-Ecstasy pills
1	N_1	n_1	x_1	$n_1 - x_1$
2	N_2	n_2	x_2	$n_2 - x_2$
...
M	N_M	n_M	x_M	$n_M - x_M$

Use a generic *beta prior* for the proportion θ of Ecstasy pills in the current seizure:

$$f(\theta | \nu_1, \nu_2) = \frac{\theta^{\nu_1-1} \cdot (1-\theta)^{\nu_2-1}}{B(\nu_1, \nu_2)} ; 0 \leq \theta \leq 1$$



Use the grouped data to estimate the parameters ν_1 and ν_2 of this beta prior.

This can be done by the *maximum likelihood method* using the fact that the probability of obtaining x_i Ecstasy pills in cluster i is

$$\Pr(x_i) \approx \frac{\binom{\lfloor N_i \cdot \theta \rfloor}{x_i} \cdot \binom{\lfloor N_i \cdot (1-\theta) \rfloor}{n_i - x_i}}{\binom{N_i}{n_i}}$$

Hypergeometric distribution

where “ $\lfloor \cdot \rfloor$ ” stands for rounding downwards to nearest integer and

$$\binom{a}{b} = \frac{a!}{b! \cdot (a-b)!} \quad \text{for integers } a \geq b$$

The obtained point estimates of ν_1 and ν_2 can be assessed with respect to *bias* and *variance* using *bootstrap resampling*.

In Nordgaard (2006) original point estimates of ν_1 and ν_2 for historical cases of blue pills at SKL are

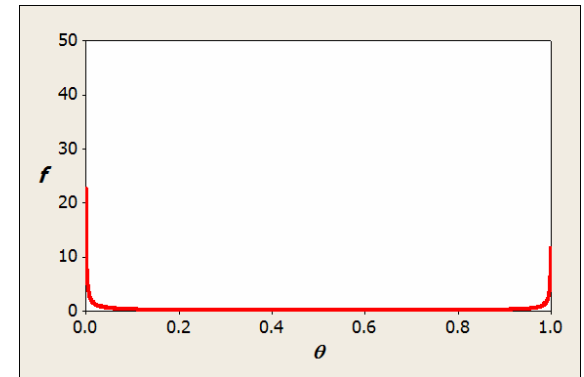
$$\hat{\nu}_1 = 0.075 \quad \text{and} \quad \hat{\nu}_2 = 0.224$$

Bias adjusted estimates are

$$\hat{\nu}_1^* = 0.038 \quad \text{and} \quad \hat{\nu}_2^* = 0.133$$

and upper 90% confidence limits for the true values of ν_1 and ν_2 are

$$\nu_1 \leq 0.062 \quad \text{and} \quad \nu_2 \leq 0.262$$



Now, assume the forthcoming sample of n units will consist entirely of Ecstasy pills. (*Otherwise the case will be considered “non-standard”*)

The sample size is determined so that the *posterior* probability of θ being higher than a certain proportion, say 50% is at least say 99% (referred to as 99% *credibility*)

For large seizures the posterior distribution of θ given all n sample units consist of Ecstasy is also *beta*:

$$f(\theta | n, v_1, v_2) = \frac{\theta^{v_1+n-1} \cdot (1-\theta)^{v_2-1}}{B(v_1+n, v_2)} ; 0 \leq \theta \leq 1$$

Thus we solve for n

$$\int_{0.50}^1 f(\theta | n, \nu_1, \nu_2) d\theta \geq 0.99$$

\Leftrightarrow

$$\frac{\int_{0.50}^1 \theta^{\nu_1+n-1} \cdot (1-\theta)^{\nu_2-1} d\theta}{B(\nu_1+n, \nu_2)} \geq 0.99$$

where ν_1 and ν_2 are replaced by their (adjusted) point estimates or upper confidence limits.

For the above case we find that with the bias-adjusted point estimates ($\hat{v}_1^* = 0.038$ and $\hat{v}_2^* = 0.133$) the required sample size is at least **3** and with the upper confidence limits used instead (i.e with 0.062 and 0.262) the required sample size is at least **4**

There are at SKL usually no large differences between different choices of estimated parameters, nor between different colours of Ecstasy pills.

A general sampling rule of $n = 5$ can therefore be used to state with 99% credibility that at least 50% of the seizure consists of Ecstasy pills. For a higher proportion, a sample size around 12 appears to be satisfactory.



For smaller seizures it is more wise to rephrase the requirement in terms of the number of Ecstasy units in the non-sampled part of the seizure.

The posterior beta distribution is then replaced with a *beta-binomial* distribution.

More details and ready-to-use Excel macros for calculating the required sample size can be found in the ENFSI booklet “Guidelines on representative drug sampling”

2. Homogeneity stated upon inspection

Consider now a case with a (large) seizure of drug pills of which the forensic scientist cannot directly suspect the contents.

Visual inspection → All pills seem to be identical

Can we substitute the “experience” from the Ecstasy case?

UV-lightning

Pills can be inspected under UV light.

The fluorescence differs between pills with different chemical composition and looking at a number of pills under UV light would thus reveal (to greatest extent) heterogeneity.

Uncertainty of this procedure lies mainly with the person who does the inspection

➔ Experiment required!



Assume a prior $g(\theta)$ for the proportion of pills in the seizure that contains a certain (but possibly unknown) illicit drug.

For sake of simplicity, assume that pills may be of two kinds (the illicit drug or another substance).

Let Y be a random variable associated with the inspection such that

$$Y = \begin{cases} 0 & \text{if inspection gives "all pills are identical"} \\ 1 & \text{if inspection gives "differences among pills"} \end{cases}$$

Relevant case is $Y = 0$

(Otherwise the result of the UV-inspection has rejected the assumption of homogeneity.)

Now, $\Pr(Y = 0 | \theta)$ for $0 < \theta < 1$

is the *false positive probability* as a function of θ (if a positive result means that no heterogeneity is detected)

while $\Pr(Y = 0 | \theta = 0) + \Pr(Y = 0 | \theta = 1)$ is the true positive probability



The prior g can be updated using this information (when available)

$$h(\theta | Y = 0) = \frac{\Pr(Y = 0 | \theta) \cdot g(\theta)}{\int_0^1 \Pr(Y = 0 | \lambda) \cdot g(\lambda) d\lambda}$$

Note that an *non-informative prior* (i.e. $g(\theta) \equiv 1$; $0 \leq \theta \leq 1$) can be used.

The updated prior (i.e. the posterior after UV-inspection) can then be used analogously to the previous case (Ecstasy)

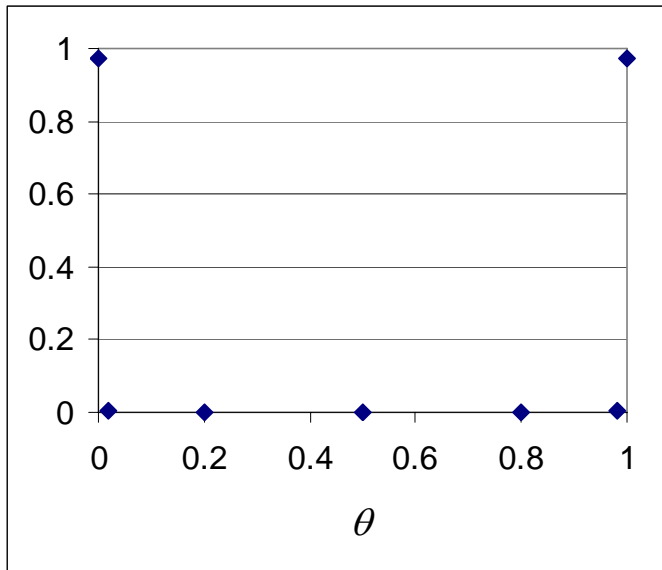
To get estimates of the false negative probabilities, an experiment was conducted at SKL.

- 9 different pills were used to form 9 different mixes of 2 pills.
- Each mix was prepared by randomly shuffling 100 pills of the current proportions on a tray that was put under UV-light
- 10 case-workers made inspections in random order such that a total of 114-117 inspections were made for each mix

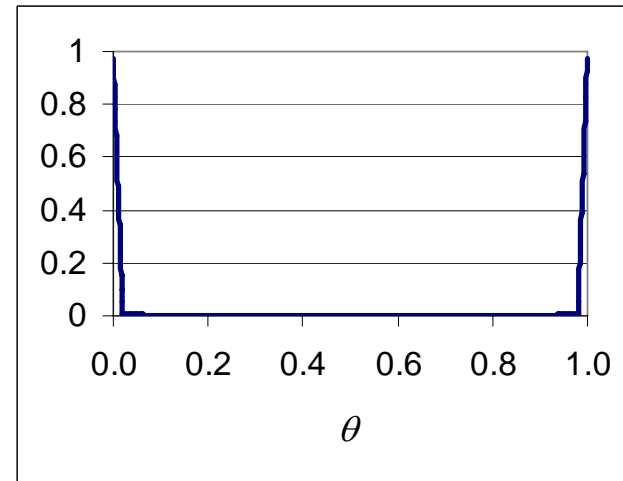
Results from experiment:

<i>Mix</i>	θ	<i>Counts of “all equal”</i> ($Y = 0$)	<i>Counts of “differences noted”</i> ($Y = 1$)
2% Noskapiin / 98% Oxascand 25 mg	0.02/0.98	0	114
2% Depolan / 98% Trimetoprim	0.02/0.98	1	116
5% Enalapril / 95% Lehydan	0.05/0.95	0	116
5% Pargitan / 95% Oxascand 15 mg	0.05/0.95	0	115
20% Oxascand 25 mg / 80% Noskapiin	0.20/0.80	0	118
20% Trimetoprin / 80% Depolan	0.20/0.80	0	114
50% Enalapril / 50% Lehydan	0.50	0	116
50% Pargitan / 50% Oxascand 15 mg	0.50	0	117
100% Egazil	0/1	114	3

Data can be illustrated by plotting estimated probabilities for $Y = 0$ vs. θ



Linear interpolation gives

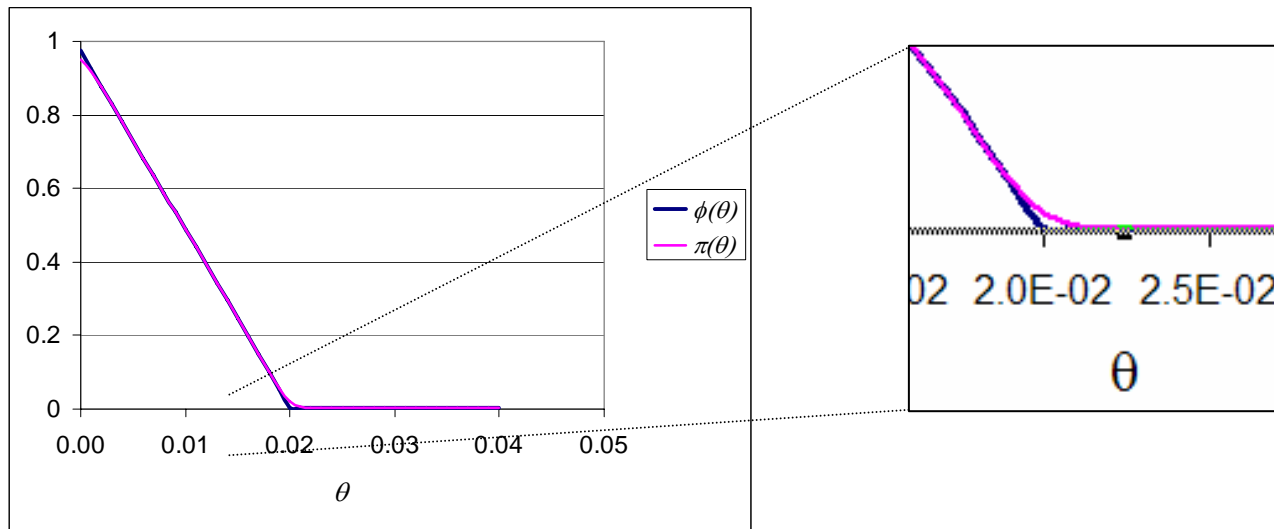


$$\hat{P}(Y = 0 | \theta) = \underline{\underline{\phi(\theta)}} = \begin{cases} 0.97 - 48.5 \cdot \theta & 0 \leq \theta \leq 0.02 \\ 0.005 - 0.024 \cdot \theta & 0.02 \leq \theta \leq 0.20 \\ 0 & 0.20 < \theta < 0.80 \\ -0.019 + 0.024 \cdot \theta & 0.80 \leq \theta < 0.98 \\ -47.5 + 48.5 \cdot \theta & 0.98 \leq \theta \leq 1 \end{cases}$$

To avoid the vertices at $\theta = 0.02, 0.20, 0.80$ and 0.98 , the linearly interpolated values are smoothed using a Kernel function:

$$\pi(\theta) = \int_0^1 K(\theta - \lambda) \cdot \phi(\lambda) d\lambda$$

where $K(x)$ is a symmetric function integrating to one over its support.



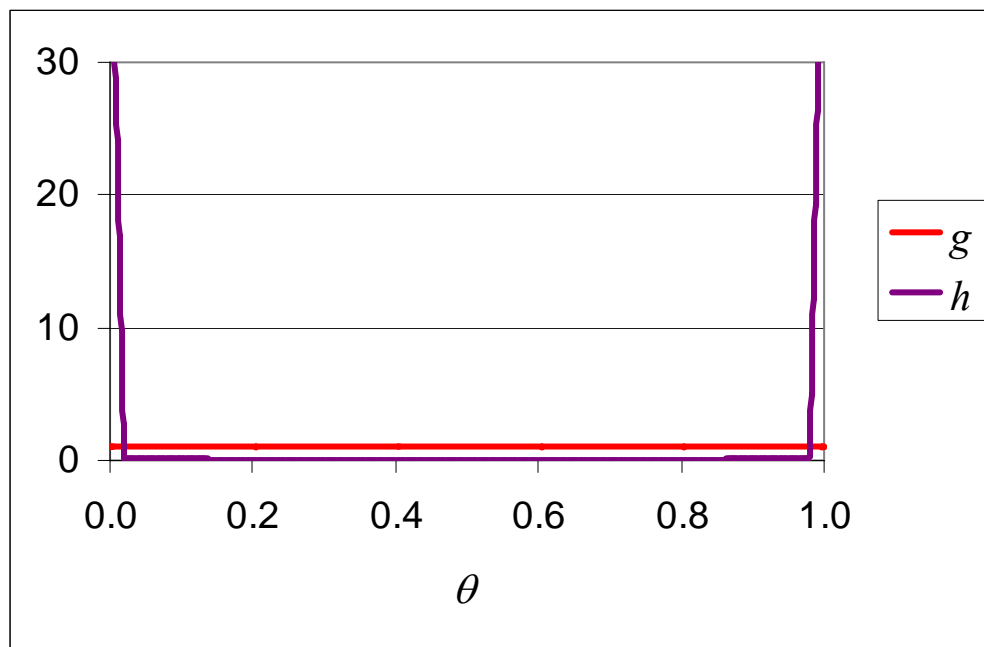
Now, the prior can be updated using this smoothed function as an estimate of $\Pr(Y = 0 | \theta)$, i.e.

$$h(\theta | Y = 0) = \frac{\pi(\theta) \cdot g(\theta)}{\int_0^1 \pi(\lambda) \cdot g(\lambda) d\lambda}$$

(With a non-informative prior g , this simplifies into

$$h(\theta | Y = 0) = \frac{\pi(\theta)}{\int_0^1 \pi(\lambda) d\lambda} \quad)$$

Comparison of the non-informative prior and the updated prior



Now, let x be the number of illicit drug pills found in a sample of n pills.

Analogously with the Ecstasy case n should be determined so that if $x = n$ a 99% credible lower limit for θ is 50% (or even higher).

With the updated prior derived the following table of posterior probabilities is obtained

n	$\Pr(\theta > 0.5 x = n, Y = 0)$
3	0.99996032237
4	0.99999475894
5	0.99999924614
6	0.99999988597
7	0.99999998211
8	0.99999999711
9	0.99999999952
10	0.99999999992

Thus, a sample size of $n = 3$ units is satisfactory.

Slightly higher values may be recommended due to the limits of the experiment

A more complicated situation (*more maths*)

Suppose we have a large seizure consisting of a number of sacs with pills.

Each sac seems to contain the same type of pills but the colour and printing of pills differ between sacs.

The whole seizure however is suspected to be Ecstasy.

Should we apply the previously introduced sampling rules to each sac one at a time, or..

Could we create a “combined” sampling rule?

For sake of simplicity, assume there are only two sacs

Define each sacs “proportion” of the whole seizure:

Alt. 1 Count the number of pills in each sac $\rightarrow N_1$ and N_2

Alt. 2 (if all pills have equal weights) Weigh the sacs $\rightarrow W_1$ and W_2



Such measures are used to calculate the proportions:

$$\alpha_1 \text{ and } \alpha_2 = (1 - \alpha_1)$$

where $\alpha_i = N_i / (N_1 + N_2)$ or $W_i / (W_1 + W_2)$ depending on the choice of measure.

As there are only two sacs we write the proportions α and $(1 - \alpha)$ respectively.

Strategy 1: Separate sampling from the sacs (stratified sampling)

Let θ_i be the proportion of Ecstasy pills in sac i ; $i = 1, 2$

Use separate independent beta priors for each θ_i :

$$f(\theta_i | \nu_1^{(i)}, \nu_2^{(i)}) = \theta_i^{\nu_1^{(i)}-1} \cdot (1 - \theta_i)^{\nu_2^{(i)}-1} / B(\nu_1^{(i)}, \nu_2^{(i)})$$

Let x_i be the number of obtained Ecstasy pills in a sample of n_i units from sac i

→
$$\Pr(x_i = k | \theta_i) = \binom{n_i}{k} \cdot \theta_i^k \cdot (1 - \theta_i)^{n_i - k}$$

Separate and independent binomial distributions

The proportion of Ecstasy pills in the whole seizure is now

$$\theta = \alpha \cdot \theta_1 + (1 - \alpha) \cdot \theta_2$$

Assume as before that we expect to get samples consisting entirely of Ecstasy pills

The inequality to be solved for n_1 and n_2 then becomes

$$\Pr(\theta > k \mid x_1 = n_1, x_2 = n_2) > p$$

With more mathematical details:

$$\Pr(\alpha \cdot \theta_1 + (1 - \alpha) \cdot \theta_2 > k \mid x_1 = n_1, x_2 = n_2) > p$$

\Leftrightarrow (as the priors are independent)

$$\iint_{\alpha \cdot \theta_1 + (1 - \alpha) \cdot \theta_2 > k} f(\theta_1 \mid x_1 = n_1, x_2 = n_2) \cdot f(\theta_2 \mid x_1 = n_1, x_2 = n_2) d\theta_1 d\theta_2$$

\Leftrightarrow (as x_i has no impact on θ_j when $i \neq j$)

$$\iint_{\alpha \cdot \theta_1 + (1 - \alpha) \cdot \theta_2 > k} f(\theta_1 \mid x_1 = n_1) \cdot f(\theta_2 \mid x_2 = n_2) d\theta_1 d\theta_2$$

where

$$f(\theta_i \mid x_i = n_i) = \frac{\theta_i^{n_i + \nu_1^{(i)} - 1} \cdot (1 - \theta_i)^{\nu_2^{(i)} - 1}}{B(n_i + \nu_1^{(i)}, \nu_2^{(i)})}; i = 1, 2$$



Strategy 2: Combined sampling

Merge the sacs into one seizure (mix carefully)

Let as before θ be the proportion of Ecstasy pills in the whole seizure.

Assuming beta priors as above for the separate sacs the *combined prior* is

$$f(\theta \mid v_1^{(1)}, v_2^{(1)}, v_1^{(2)}, v_2^{(2)}) = \alpha \cdot f(\theta \mid v_1^{(1)}, v_2^{(1)}) + (1 - \alpha) \cdot f(\theta \mid v_1^{(2)}, v_2^{(2)})$$

We would now take a sample of n units from the merged seizure.

Let x be the number of Ecstasy pills in this sample.

$$\rightarrow \Pr(x = k | \theta) = \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k}$$

Assume as before that we expect $x = n$ and that the inequality to be solved is

$$\Pr(\theta > k | x = n) > p$$

The posterior distribution of θ given $x = n$ becomes

$$f(\theta | x = n) = \frac{\theta^n \cdot (\alpha \cdot f(\theta | \nu_1^{(1)}, \nu_2^{(1)}) + (1 - \alpha) \cdot f(\theta | \nu_2^{(2)}, \nu_2^{(2)}))}{\int_0^1 \lambda^n \cdot (\alpha \cdot f(\lambda | \nu_1^{(1)}, \nu_2^{(1)}) + (1 - \alpha) \cdot f(\lambda | \nu_2^{(2)}, \nu_2^{(2)})) d\lambda} =$$

$$= \beta \cdot \frac{\theta^{n+\nu_1^{(1)}-1} \cdot (1-\theta)^{\nu_2^{(1)}-1}}{B(n+\nu_1^{(1)}, \nu_2^{(1)})} + (1-\beta) \cdot \frac{\theta^{n+\nu_1^{(2)}-1} \cdot (1-\theta)^{\nu_2^{(2)}-1}}{B(n+\nu_1^{(2)}, \nu_2^{(2)})}$$

where

$$\beta = \frac{\alpha \cdot \frac{\Gamma(\nu_1^{(1)} + \nu_2^{(1)}) \cdot \Gamma(n + \nu_1^{(1)})}{\Gamma(\nu_1^{(1)}) \cdot \Gamma(n + \nu_1^{(1)} + \nu_2^{(1)})}}{\alpha \cdot \frac{\Gamma(\nu_1^{(1)} + \nu_2^{(1)}) \cdot \Gamma(n + \nu_1^{(1)})}{\Gamma(\nu_1^{(1)}) \cdot \Gamma(n + \nu_1^{(1)} + \nu_2^{(1)})} + (1 - \alpha) \cdot \frac{\Gamma(\nu_1^{(2)} + \nu_2^{(2)}) \cdot \Gamma(n + \nu_1^{(2)})}{\Gamma(\nu_1^{(2)}) \cdot \Gamma(n + \nu_1^{(2)} + \nu_2^{(2)})}}$$



Thus the posterior is in the same class (weighted beta distributions) as the prior.

Solving this inequality becomes more complex but which strategy is best?

...at an experimental stage right now

An example from quality assurance

Project on automatic identification of substances from chromatograms (SKL: SYREN project)

Idea is to avoid manual inspection of chromatograms (a time-consuming process)

Requirements:

Moderate false negative probabilities can be accepted
(will send the chromatogram back to manual inspection)

False positive probabilities must be extremely low

A quality experiment is to be carried out. How many trials do we need to run to state whether the requirements are fulfilled?



Assume our prior assumption about the *false positive probability*, FPP is $1/50000$.

Can inference about such a small probability be done with a number of trials lower than 50000?

We need statements of the kind

$$\Pr(FPP < FPP_{\max} \mid \text{Experimental data}) > p$$

i.e. a $100p$ % upper credible limit for FPP satisfies that FPP is lower than a defined limit FPP_{\max}



Let θ_S depict the true *FPP* of the automatic identification method for a certain substance S .

Let n be the number of trials sought.

Let x_S be the number of *false* identifications of substance S .

$$\rightarrow \Pr(x_S = k \mid n, \theta_S) = \binom{n}{k} \cdot \theta_S^k \cdot (1 - \theta_S)^{n-k} ; k = 0, 1, \dots, n$$

*Binomial
distribution*

Let the prior *assumption* about θ_S be that it is around 1/50000.

How can this assumption be transformed into a suitable *prior distribution*?

First, the most proper distribution class is the beta distributions

$$f(\theta_S | \nu_1, \nu_2) = \frac{\theta_S^{\nu_1-1} \cdot (1-\theta_S)^{\nu_2-1}}{B(\nu_1, \nu_2)} ; 0 \leq \theta_S \leq 1$$

The question (as before) is how to choose values of ν_1 and ν_2 .

A first suggestion:

The mean of the beta distribution is $E(\theta_s) = \frac{v_1}{v_1 + v_2}$

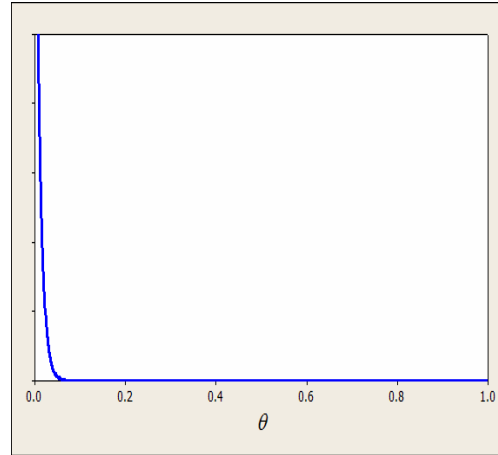
We may substitute the mean by the assumption 1/50000 (i.e. method of moments estimation):

$$\frac{\hat{v}_1}{\hat{v}_1 + \hat{v}_2} = \frac{1}{50000}$$

→ One equation, two unknowns ?



The shape of the beta distribution we are looking for should be of the following kind



Such shapes are achieved by setting ν_1 equal to 1 and $\nu_2 > 1$

→ Solve

$$\frac{1}{1 + \hat{\nu}_2} = \frac{1}{50000}$$

$$\Rightarrow \hat{\nu}_2 = 49999 \quad (\approx 50000)$$

In general, for a prior assumption of a very low *FPP* (or any proportion) to be θ_{prior} , the prior distribution can be chosen to be

$$f(\theta) = \frac{(1 - \theta)^{\theta_{prior}^{-1} - 1}}{B(1, \theta_{prior}^{-1})}; 0 \leq \theta \leq 1$$

Analogously, for a prior assumption of a very high *FPP* the prior distribution can be chosen to be

$$f(\theta) = \frac{\theta^{\theta_{prior}^{-1} - 1}}{B(\theta_{prior}^{-1}, 1)}; 0 \leq \theta \leq 1$$

Now, the prior distribution in our case is $f(\theta_S) = \frac{(1 - \theta_S)^{49998}}{B(1, 49999)}$

To decide upon the number of trials we assume (analogously with the drug sampling cases) that the number of false identifications will be zero, i.e. that $x_S = 0$.

Then choose n such that

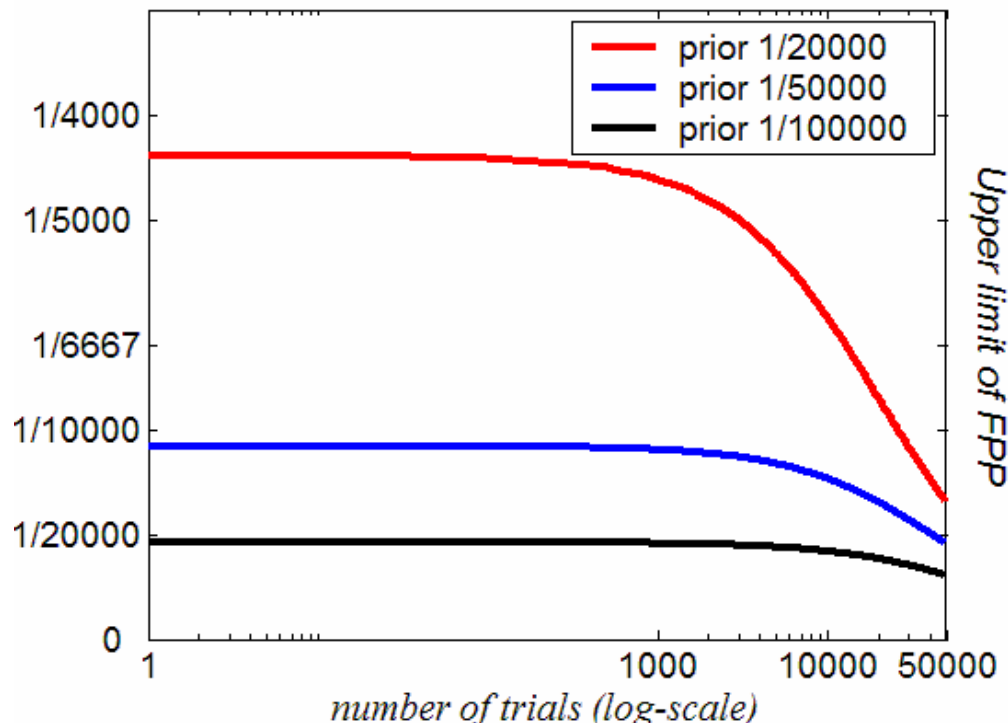
$$\Pr(\theta_S < \theta_{\max} \mid x_S = 0, n) > p$$

where θ_{\max} is a defined upper limit for the *FPP*

In practice, there will be a balancing between the number of trials and the upper limit of the *FPP*.

Fix p , i.e. the degree of credibility and use the posterior probability to illustrate the relationship between n and θ_{\max} .

Below is graphed such relationships for $p = 0.99$ and for 3 different choices of θ_{prior} , i.e. the prior assumption on θ_S (1/20000, 1/50000 and 1/100000).



Things start to “happen” for quite a large number of trials



Other suggestions for the choice of prior parameters

An assumption about the *FPF* being approximately 1/50000 can be interpreted as an observation that the first false identification has come after 50000 runs.

Assume a non-informative prior for θ_S ($f(\theta_S) \equiv 1$) and *update* this prior using the above reasoning.

Let y be the number of runs until the first false identification is made.

→ y is *geometrically* distributed, i.e.

$$\Pr(y = m | \theta_S) = (1 - \theta_S)^{m-1} \cdot \theta_S$$

The update is then the posterior distribution of θ_S given $y = m$

$$\begin{aligned} f(\theta_S | y = m) &= \frac{(1 - \theta_S)^{m-1} \cdot \theta_S \cdot 1}{\int_0^1 (1 - \lambda)^{m-1} \cdot \lambda \cdot 1 d\lambda} = \\ &= \frac{\theta_S^{2-1} \cdot (1 - \theta_S)^{m-1}}{B(2, m)} \end{aligned}$$

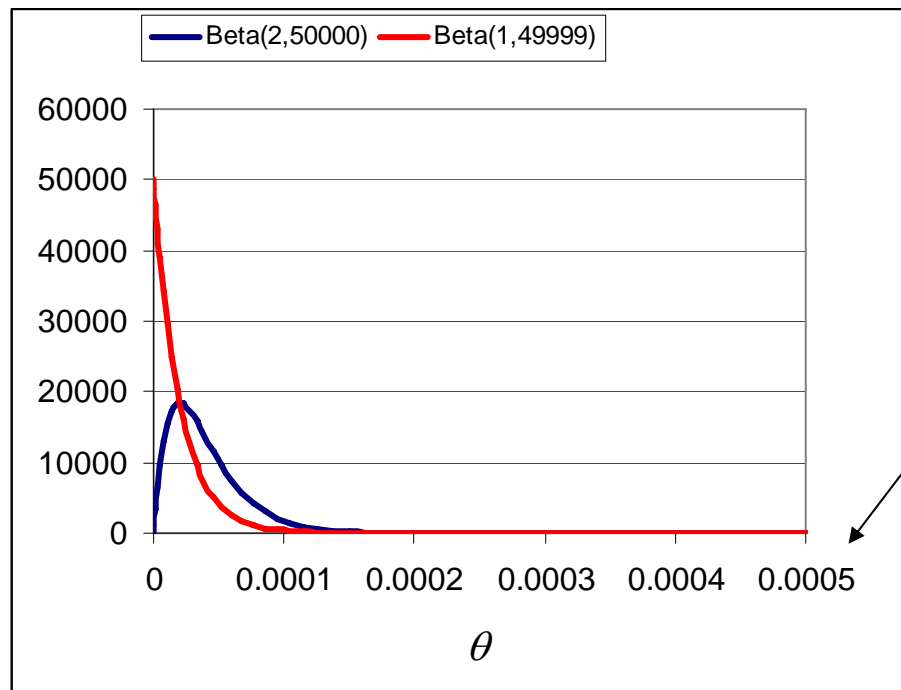
i.e. a beta distribution with $\nu_1 = 2$ and $\nu_2 = m$

More specifically, with $m = 50000$ we come up with the prior

$$Beta(2, 50000)$$

to compare with the previous $Beta(1, 49999)$

The differences are small but may still have impact



Note that the scale does not extend to 1

An application on increasing the evidence value

Suppose we have a bank-note with stains of dye suspected to have come from a safe deposit box for bank-notes. (*Bank-notes should be impregnated with such dye if somebody tries to break into the box*)

The hypothesis put forward by the prosecution is

H_P : The stains on the bank-note consists of dye from a deposit box

The hypothesis put forward by the defence is

H_D : The stains on the bank-note consists of dye from some other type of source



Some assumptions need to be made:

- All potential deposit boxes from which the bank-note could have got its stains has the same type of impregnation dye
- The impregnation dye used in deposit boxes cannot be found anywhere else.
- There is a negligible risk that the stains is a result of contamination, i.e. that the dye (or some part of it) has been transferred from another impregnated bank-note and persisted.



Now, let

E_1 = “The composition of colours in the stains matches visually the types of colour used in deposit box impregnating dye”

E_2 = “The impregnation pattern on the bank-note is of the same kind that can be seen on bank-notes known to have been tinted with deposit box impregnating dye”

The likelihood ratio for the combination of these two pieces of evidence is assumed to exceed 10000, i.e.

$$\frac{\Pr(H_P | E_1, E_2)}{\Pr(H_D | E_1, E_2)} > 10000 \times \frac{\Pr(H_P)}{\Pr(H_D)}$$

On the scale of conclusions used at SKL this is expressed as strong support for the prosecutor's hypothesis.

Another piece of evidence is to be introduced with purpose to increase the posterior odds further:

$E_3 =$ “There is a high concentration of triethyl phosphate in the bank-note”

Triethyl phosphate (TEP) is one of the components in deposit box impregnation dye, but can also be found elsewhere, e.g. in flame retardants.

Under the assumptions of H_P as well as H_D , E_3 is conditionally independent of E_1 and E_2 .



$$\frac{\Pr(H_P | E_1, E_2, E_3)}{\Pr(H_D | E_1, E_2, E_3)} = \frac{P(E_3 | H_P)}{P(E_3 | H_D)} \times \frac{\Pr(H_P | E_1, E_2)}{\Pr(H_D | E_1, E_2)}$$

Now, $\Pr(E_3 | H_P)$ can be approximately be set to 1

We would expect $\Pr(E_3 | H_D)$ to be very small as the presence of TEP on bank-notes without dyeing can only be explained with contamination, and the concentration would then be fairly low.

We need a survey of non-dyed banknotes from which we can estimate $\Pr(E_3 | H_D)$ or actually give an upper limit of this probability with a certain credibility.

As other evidence are present and fairly strong we would be satisfied if we could state that

$$\Pr\{\Pr(E_3 | H_D) < 0.01 \mid \text{Survey results}\} > 0.99$$

This will then increase the evidence value with a factor 100.

On the scale of conclusions used at SKL this would be expressed as support with certainty for the prosecutor's hypothesis.



Again, we may assume that a sample of non-dyed bank-notes would not contain any units with high concentrations of TEP.

Depict $\Pr(E_3 | H_D)$ by θ and let x be the number of bank-notes in a sample of n units.

As before

$$f(\theta | \nu_1, \nu_2) = \theta^{\nu_1-1} \cdot (1-\theta)^{\nu_2-1} / B(\nu_1, \nu_2)$$

Beta prior

$$\Pr(x = k | n, \theta) = \binom{n}{k} \cdot \theta^k \cdot (1-\theta)^{n-k}$$

Binomial distribution

Then we should choose n so that

$$\Pr(\theta < 0.01 | x = 0, n) > 0.99$$

Upper 99% credible limit

Different priors and corresponding sample sizes

v_1	v_2	Prior information	Prior mean	Sample size
1	1	Non-informative	0.5	458
1	10	High conc. TEP more uncommon	< 0.1	449
1	100	High conc. TEP rare	< 0.01	359
1	250		<0.004	209
1	500	High conc. TEP extremely rare	< 0.002	1