Meeting 18:
Sampling issues (without and with a decision-analytic approach), Robustness

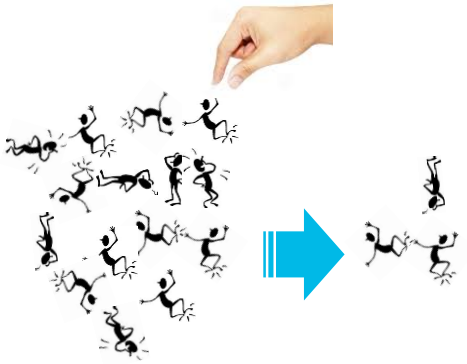
In front of the oral exam

- Try to form groups of two persons – if not possible then a single person is the alternative
- Suggest a few time-points comprising 2 hours for two persons and one hour for one person, and send a mail to me with those suggestions
- There is in principal no deadline – the exam can be taken at any time (as long as the course exists (i.e. having course code 732A66))!
- An example of excerpts from a dialogue at the oral exam has been uploaded to the course web

Sampling issues in classical inference

Sampling in a general sense can be of two kinds:

I. Sampling without replacement from a finite population



II. Sampling from an infinite population/from a process/(with replacement from a finite population)

Case II also covers making several measurements on a specific object.

For simple random sampling …

How many units should be sampled?

Depends to the objective of sampling:

- to estimate the value of a parameter?
  - the number of units should be chosen from a desired bound on the point estimate (desired length of a confidence interval)
  - requires prior knowledge of the population variance

$$n_0 \geq \frac{4 \cdot \left(z_{\alpha/2} \cdot \sigma\right)^2}{D^2} \qquad n = \frac{n_0}{1 + (n_0 - 1)/N}$$

D is the desired length of a confidence interval, N is the size of the population

- to be able to reject a (range of) value(s) of a parameter with a high probability when the true parameter value is at a certain distance from that (range of) value(s)?
  - the number of units should be chosen so that the power function of a hypothesis test at such a distance is at least as high at that probability
  - requires prior knowledge of the population variance

For stratified sampling and cluster sampling …

How many units should be sampled, and if the population is stratified, how should they be allocated over strata?

- to estimate the value of a parameter?
  - the number of units should be chosen from a desired bound on the point estimate (desired length of a confidence interval)
  - requires prior knowledge of the population variance, stratum variances and sizes, numbers and sized of clusters

For sampling from several populations (and planning of experiments) …

How many units should be sampled from each population? How many measurements should be made for each experimental setup?

- to be able to detect with a certain probability a difference in a parameter between two populations? to be able to reject the hypothesis of no differences in effect between the different set-ups of an experiment?
  - the number of units should be chosen so that the power function of the hypothesis test used at such a distance is at least as high at that probability
  - requires prior knowledge of the population variances

The Bayesian approach to sampling

How many units should be sampled to be able to state with a certain probability p_0 that

- a parameter is at least/most a certain value or within a specific range?
- the difference in a parameter between two populations is at least/ at most a certain value or within a specific range?
- …

$$P\left(\theta \geq \theta_0 \middle| n, y_1, \ldots, y_n\right) = \int_{\theta \geq \theta_0} f''(\theta | n, y_1, \ldots, y_n; \psi) d\theta$$

$$f''(\theta | n, y_1, \ldots, y_n; \psi) = \frac{f(y_1, \ldots, y_n | \theta; \psi) \cdot f'(\theta | \psi)}{\int_{u \in \Theta} f(y_1, \ldots, y_n | u; \psi) \cdot f'(u | \psi) du}$$

Solve for n $\quad P\left(\theta \geq \theta_0 \middle| n, y_1, \ldots, y_n\right) \geq p_0$

Examples from drugs sampling

Consignments of drugs in forms of pills, capsules, ampoules or plastic bags can be very extensive (e.g. thousands of pills in big sacs)

Analysis must by legal reasons be made “unit-wise” and is time-consuming $\Rightarrow$ as small sample sizes as possible are desired.

However, drug seizures are usually homogeneous with respect to the active substance in each unit.

For so-called *identifying analysis* what is of interest is whether a unit contains the active (illicit) substance or not – percentages of that substance or presence of other substances is of minor importance.

Hence, the sampling scheme (hypergeometric or approx. binomial) is expected to render homogeneous samples, i.e. $x = n$ (all units are illicit ones) or $x = 0$ (no unit is illicit).

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”.

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:

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

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

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 that the probability of obtaining x_i Ecstasy pills in cluster *i* is

$$P(\tilde{x}_i=x_i|\theta,n_i)\approx\frac{\binom{\lfloor N_i\cdot\theta\rfloor}{x_i}\cdot\binom{\lfloor N_i\cdot(1-\theta)\rfloor}{n_i-x_i}}{\binom{N_i}{n_i}}$$

Hypergeometric distribution

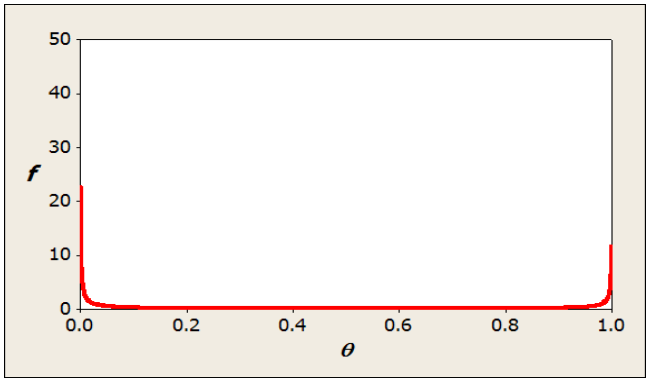
where "$\lfloor\cdot\rfloor$" stands for rounding downwards to nearest integer

The likelihood function of ν_1 and ν_2 in light of observed numbers in all clusters ($\boldsymbol{x}=(x_1,\ldots,x_M)$) then becomes

$$\mathcal{L}(\nu_1,\nu_2|\mathrm{x})=\prod_{i=1}^{M}P(\tilde{x}_i=x_i|\nu_1,\nu_2,n_i)=\prod_{i=1}^{M}\int_0^1 P(\tilde{x}_i=x_i|\theta,n_i)\cdot p(\theta|\nu_1,\nu_2)d\theta$$

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 (*now* NFC*)* are



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

Bias adjusted estimates are

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

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

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

Should confidence limits be used in empirical Bayes?

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*:

$$f''\left(\theta\middle|n,\nu_1,\nu_2\right)=\frac{\theta^{\nu_1+n-1}\cdot(1-\theta)^{\nu_2-1}}{B(\nu_1+n,\nu_2)}\ ;\ 0\le\theta\le 1$$

Thus we solve for *n*

$$\int_{0.50}^{1} f\left(\theta\middle|n,\nu_1,\nu_2\right)d\theta\ge 0.99 \quad\Leftrightarrow\quad \frac{\int_{0.50}^{1}\theta^{\nu_1+n-1}\cdot(1-\theta)^{\nu_2-1}d\theta}{B(\nu_1+n,\nu_2)}\ge 0.99$$

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

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

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 in general 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.

2. Homogeneity stated upon inspection only

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

Visual inspection $\Rightarrow$ 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.

does not work for capsules and ampoules

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

$\Rightarrow$ 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, $P(Y = 0|\theta)$ 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 $P(Y = 0|\theta = 0) + P(Y = 0|\theta = 1)$

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 upon UV-inspection) can then be used analogously to the previous case (Ecstasy).

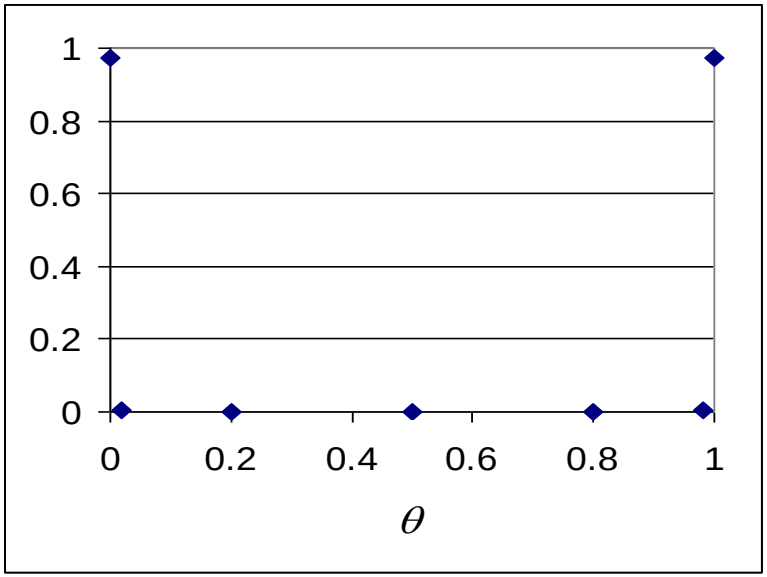
Example Experiment (conducted at SKL (*now* NFC))

- 8 types of pills with different substances were used to form 9 different mixtures (i.e. of two proportions) of 2 types of pills
- Each mixture was prepared by randomly shuffling 100 pills with 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 mixture

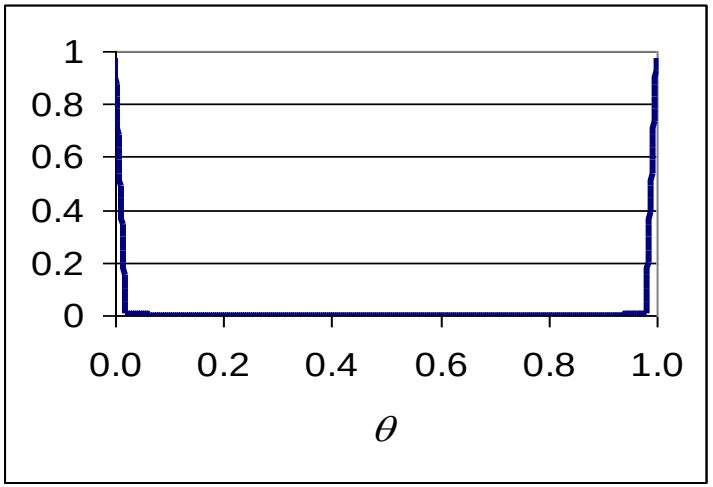
Data:

Mix	Mix code (within experiment)	Counts of "all equal" (Y = 0)	Counts of "differences noted" (Y = 1)
2% Noskapin / 98% Oxascand 25 mg	2A	0	114
2% Depolan / 98% Trimetoprim	4A	1	116
5% Enalapril / 95% Lehydan	1A	0	116
5% Pargitan / 95% Oxascand 15 mg	3A	0	115
20% Oxascand 25 mg / 80% Noskapin	2B	0	118
20%Trimetoprin / 80% Depolan	4B	0	114
50% Enalapril / 50% Lehydan	1B	0	116
50% Pargitan / 50% Oxascand 15 mg	3B	0	117
100% Egazil	5	114	3

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



Linear interpolation gives

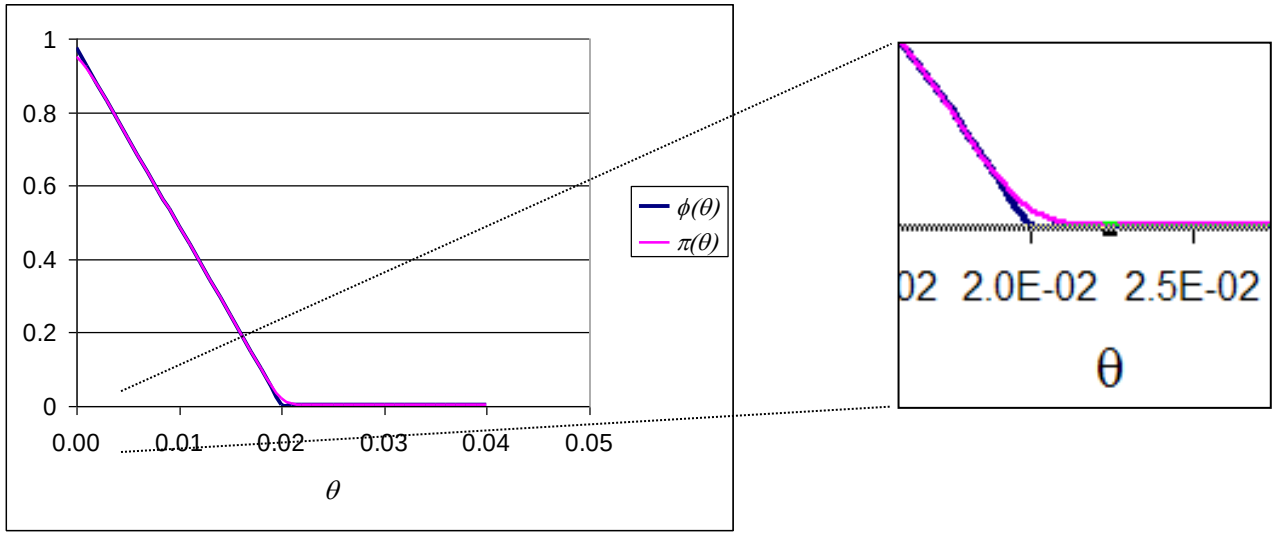


$$\hat{P}(Y=0\mid\theta)=\underline{\underline{\phi(\theta)}}=\begin{cases}0.97-48.5\cdot\theta & 0\le\theta\le 0.02\\ 0.005-0.024\cdot\theta & 0.02\le\theta\le 0.20\\ 0 & 0.20<\theta<0.80\\ -0.019+0.024\cdot\theta & 0.80\le\theta<0.98\\ -47.5+48.5\cdot\theta & 0.98\le\theta\le 1\end{cases}$$

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

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

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

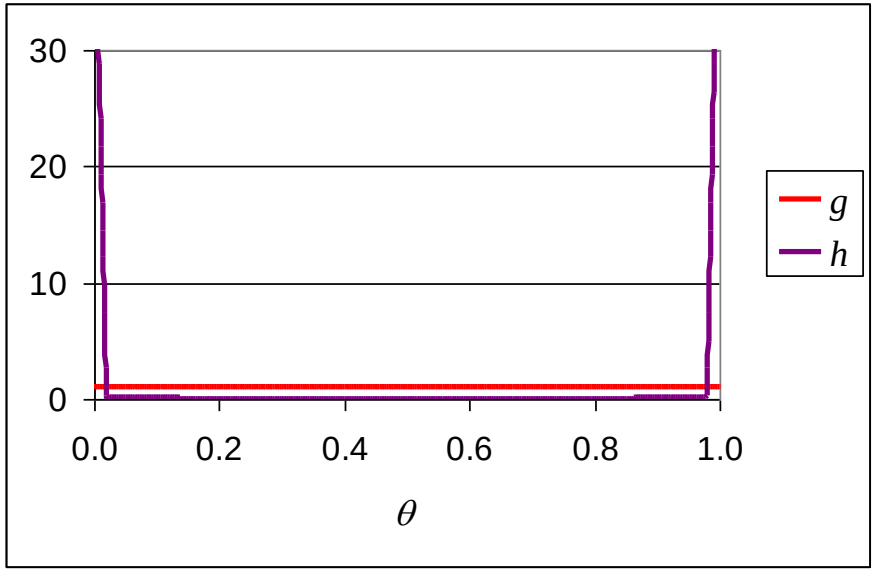


Now, the prior can be updated using this smoothed function as an estimate of $\Pr(Y = 0|\theta)$, i.e.

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

(With a non-informative prior g, this simplifies into $h(\theta|Y = 0) = \frac{\pi(\theta)}{\int_0^1 \pi(\lambda) d\lambda}$)

Comparison of the non-informative prior g and the updated prior h



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	$\Pr(\theta > 0.5 \mid x = n, Y = 0)$
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

The decision-theoretic approach

As was previously taken up, the decision about sampling (and how much to sample) builds on the expected value of sample information, EVSI(n), and the optimal sample size is the value of n for which the *expected net gain of sampling*

$$\text{ENGS}(n) = \text{EVSI}(n) - \text{CS}(n)$$

is maximised.

Particularly, from Meeting 11:

$$\text{VSI}(y) = E''R(a''|y) - E''R(a'|y) = E''L(a'|y) - E''L(a''|y)$$

$$\text{EVSI} = \int \text{VSI}(y) \cdot f(y)dy = \left\langle \begin{matrix}\text{One}\\ \text{variant}\end{matrix} \right\rangle = \cdots = E''L(a') - E''L(a'')$$

Example: Return to the examples with illicit pills

Assume we should make a decision on whether the proportion, θ, of Ecstasy pills in a seizure of 1000 pills is less than or at least 50 %.

The possible actions are a_1 = "$\theta < 50$ %" and a_2 = "$\theta \geq 50$ %"

Assume a "0–k_i" loss function as

	$\theta < 50$ %	$\theta \geq 50$ %
a_1	0	1
a_2	20	0

Assume a prior distribution of θ as *Beta*(ν_1, ν_2) with $\nu_1 =$ 0.038 and $\nu_2 = 0.133$ (the point estimates from the empirical Bayes procedure)

$$\Rightarrow P(\theta < 0.50 | \nu_1 = 0.038, \nu_2 = 0.133) = \int_0^{0.5} \frac{\theta^{0.038-1} \cdot (1-\theta)^{0.133-1}}{B(0.038,0.133)} d\theta$$

$$\approx 0.780$$

$$\Rightarrow$$

$$EL(a_1) = 0 \cdot 0.780 + 1 \cdot 0.220 = 0.22$$

$$EL(a_2) = 20 \cdot 0.780 + 0 \cdot 0.220 = 15.6$$

$$\Rightarrow a'_{opt} = a_1$$

The number of Ecstasy pills in a sample of *n* pills is *Bin*(*n*, θ).

Pre-assuming the sample to be completely homogeneous, i.e. either all are Ecstasy pills or all are non-Ecstasy pills gives the posterior distribution to be any of

Beta($\nu_1 + n$, ν_2) [all are Ecstasy] and *Beta*(ν_1, $\nu_2 + n$) [all are non-Ecstasy]

With *Beta*($\nu_1 + n,\ \nu_2$) as posterior the expected posterior losses are

$$EL^{(i)}(a_1|n) = 0 \cdot \Pr(\theta < 0.5|n,n) + 1 \cdot \Pr(\theta \geq 0.5|n,n)$$

$$= \int_{0.5}^{1} \frac{\theta^{0.038+n-1}(1-\theta)^{0.133-1}}{B(0.038+n,0.133)}\, d\theta$$

$$EL^{(i)}(a_1|n) = 20 \cdot \Pr(\theta < 0.5|n,n) + 0 \cdot \Pr(\theta \geq 0.5|n,n)$$

$$= 20 \int_{0}^{0.5} \frac{\theta^{0.038+n-1}(1-\theta)^{0.133-1}}{B(0.038+n,0.133)}\, d\theta$$

With *Beta*($\nu_1,\ \nu_2 + n$) as posterior the expected posterior losses are

$$EL^{(ii)}(a_1|n) = 0 \cdot \Pr(\theta < 0.5|0,n) + 1 \cdot \Pr(\theta \geq 0.5|0,n)$$

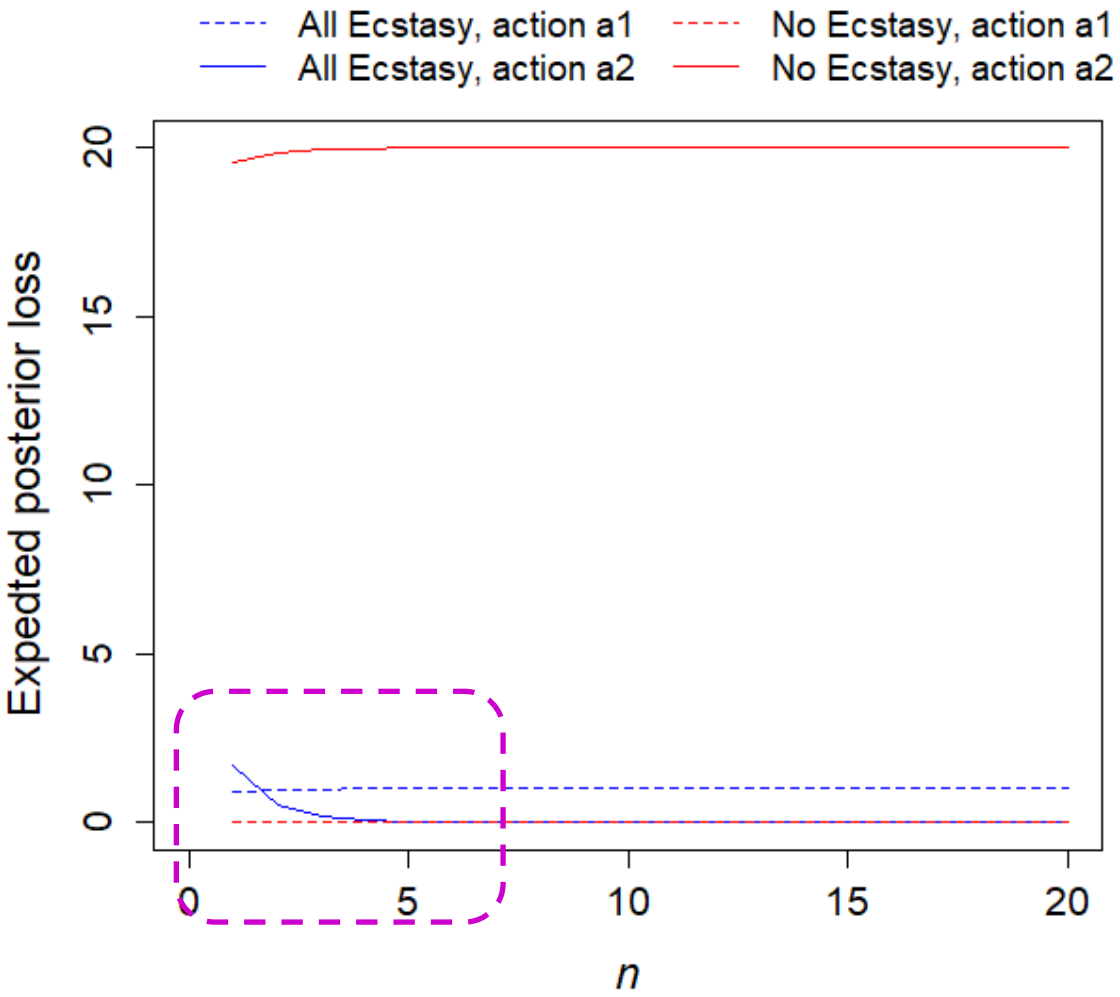
$$= \int_{0.5}^{1} \frac{\theta^{0.038-1}(1-\theta)^{0.133+n-1}}{B(0.038,0.133+n)}\, d\theta$$

$$EL^{(ii)}(a_2|n) = 20 \cdot \Pr(\theta < 0.5|0,n) + 0 \cdot \Pr(\theta \geq 0.5|0,n)$$

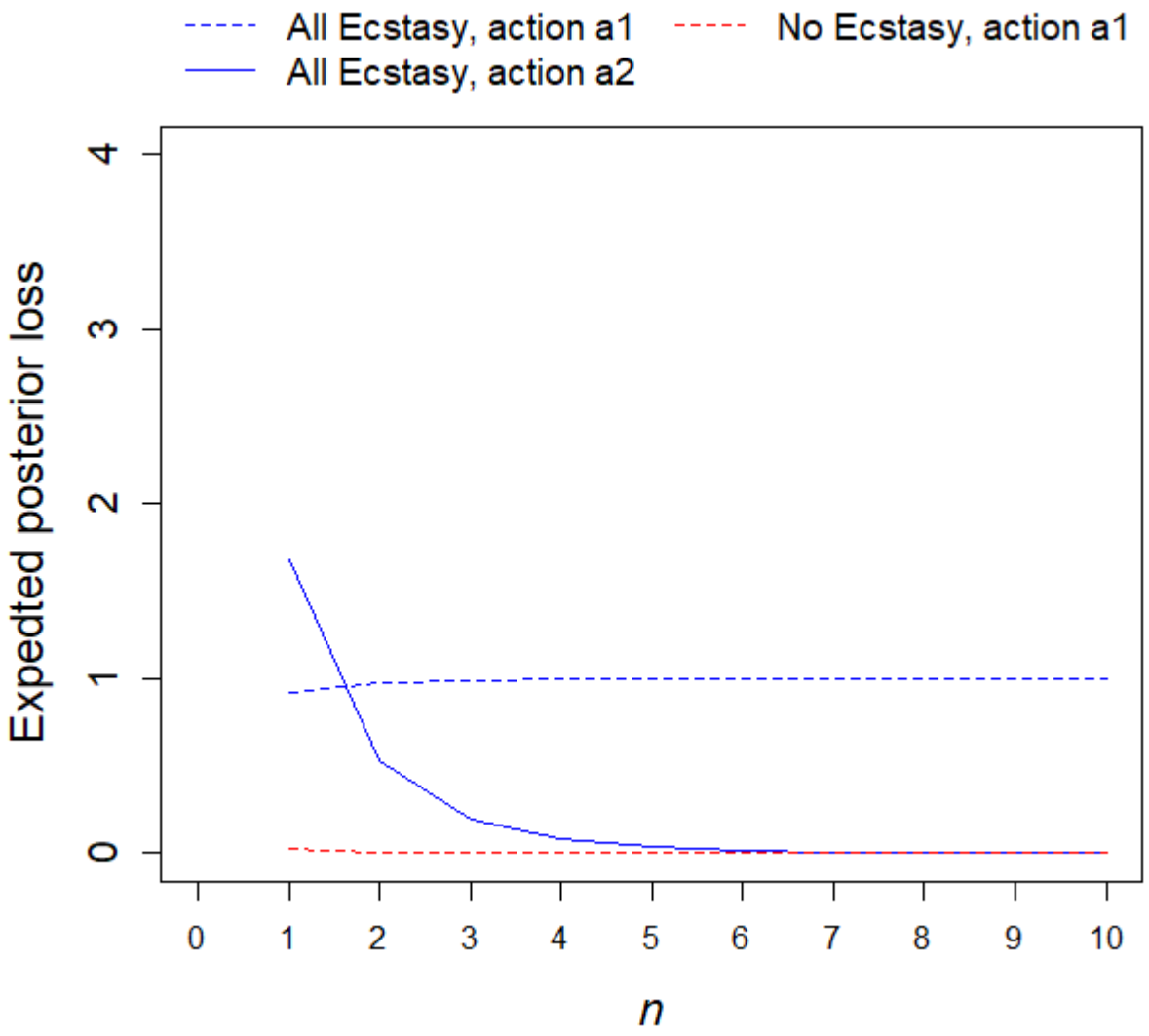
$$= 20 \int_{0}^{0.5} \frac{\theta^{0.038-1}(1-\theta)^{0.133+n-1}}{B(0.038,0.133+n)}\, d\theta$$

How would we obtain the optimal sample size?

Graph the posterior expected losses against *n*



Not an issue when all sampled pills are non-Ecstasy



With a sample size of 2 all should be clear. However with no costs involved minimising ENGS is not the question here.

Some future perspectives on decision making and data acquisition

The Bayesian view of decision making is that the rational decision is the decision (action) that maximises the expected utility (or minimises the expected loss) with respect to the available (subjective) probability distributions over the possible states of the world.

However, this applies to a situation where the decision maker is *forced* to make a decision

In mathematical terms the optimal action conditional on the available information is

$$\arg\max_{a}\left\{\int U(a,\theta)g(\theta|I)d\theta\right\}$$

where $g(\theta|I)$ is the probability distribution over the possible states of the world (θ) given the available information I.

An "easy" case may be if one of the possible actions is "leave things as they are for now", which signals a possibility to postpone the decision.

But what if EU ("leave things as they are for now"$|I$) < $EU(a|I)$ for some action a ?

Decision in court is an important example, medical decisions may be another.

In court the possible actions are two: *convict* or *acquit*

However, the evidence that has been forwarded by the prosecutor and the pleas made by the prosecutor and the defense may lead to that (a member of) the jury [or the court] finds

$$EL(\text{acquit}|I) > EL(\text{convict}|I)$$

but despite this chooses to acquit the defendant.

The point is that "acquit" may stand for two actions:

acquit when $EL(\text{acquit}|I) < EL(\text{convict}|I)$

acquit when $EL(\text{acquit}|I) > EL(\text{convict}|I)$ but I is deemed non-sufficient

acquit when $EL(\text{acquit}|I) < EL(\text{convict}|I)$

acquit when $EL(\text{acquit}|I) > EL(\text{convict}|I)$ but I is deemed non-sufficient

Recall the presentation at Meeting 2:

We wrote $I = I(m) = \bigcup_{k=1}^{m} I_k$, where I_1, I_2, … are (mutually exclusive) pieces of background information

Say that in a case where the decision is acquit of the second kind $m = m_0$
$\Rightarrow I = I(m_0)$

Hence, deeming I to be non-sufficient means deeming m_0 to be too small.

The question is: For which $m > m_0$ would $I(m)$ be sufficient

This can also be projected to subordinal decision problems: Have the prosecutor proven that the defendant's shoe made the shoe mark recovered from the crime scene?

Assume there is a case where a door has been forced with some kind of breaking tool. On the door there are green stains of paint in the toolmarks. A green-painted crowbar has been seized from a suspect.

The hypotheses are

$\boldsymbol{H_m}$: The green stains of paint (S) originate from the seized crowbar (G)
$\boldsymbol{H_a}$: The green stains of paint (S) originate from another breaking tool

The forensic scientist that analysed the stains and the crowbar gives the following report:

"The observed characteristics [E] of the stains [features of the paint] are between 200 and 300 times more probable if the stains originate from the seized crowbar compared to if they originate from another breaking tool"

This means that the likelihood ratio of $\boldsymbol{H_m}$ versus $\boldsymbol{H_a}$ is *somewhere* between 200 and 300 – only one value but not precisely which one.

$$200 \leq \frac{P(E|\boldsymbol{H_m}, I)}{P(E|\boldsymbol{H_a}, I)} \leq 300 \quad \text{or} \quad 200 \leq \frac{f(E|\boldsymbol{H_m}, I)}{f(E|\boldsymbol{H_a}, I)} \leq 300$$

depending on the scale of the measurements

$\boldsymbol{H_m}$: The green stains of paint (S) originate from the seized crowbar (G)
$\boldsymbol{H_a}$: The green stains of paint (S) originate from another breaking tool

$$200 \leq \frac{P(E|\boldsymbol{H_m}, I)}{P(E|\boldsymbol{H_a}, I)} \leq 300$$

Now, if we put ourselves into the factfinder's (jury, judge) situation…

The report provide quite a large likelihood ratio (Bayes' factor).

With even prior odds, i.e. $\frac{P(\boldsymbol{H_m}|I)}{P(\boldsymbol{H_a}|I)} = 1$ and if we assume it can be nothing else than a breaking tool used to force the door, then $P(\boldsymbol{H_m}|E, I) + P(\boldsymbol{H_a}|E, I) = 1$ and the posterior probability of $\boldsymbol{H_m}$ will be

$$P(\boldsymbol{H_m}|E, I) \geq \frac{200}{200 + 1} \approx 0.995$$

which among many judicial decision makers would be considered very high.

$\boldsymbol{H_m}$: The green stains of paint (S) originate from the seized crowbar (G)
$\boldsymbol{H_a}$: The green stains of paint (S) originate from another breaking tool

$$200 \leq \frac{P(E|\boldsymbol{H_m}, I)}{P(E|\boldsymbol{H_a}, I)} \leq 300$$

The alternative view is to go via the loss function

Action	States of the world	
	$\boldsymbol{H_m}$ is true	$\boldsymbol{H_a}$ is true
Accept $\boldsymbol{H_m}$	0	L(II)
Accept $\boldsymbol{H_a}$	L(I)	0

$$200 \leq B \leq 300$$

where the decision would be "Accept $\boldsymbol{H_m}$" if $B > \frac{P(\boldsymbol{H_a}|I)}{P(\boldsymbol{H_m}|I)} \cdot \frac{L(\text{II})}{L(\text{I})}$

With $\frac{P(\boldsymbol{H_m}|I)}{P(\boldsymbol{H_a}|I)} = 1$ this decision rule reduces to $B > \frac{L(\text{II})}{L(\text{I})}$

Hence, as long as $L(\text{II}) < B \times L(\text{I}) \geq 200 \times L(\text{I})$ the decision will be "Accept $\boldsymbol{H_m}$"

When L(II) is almost equal to $200 \times L(\text{I})$ an equivalent decision rule is to

accept $\boldsymbol{H_m}$ if $P(\boldsymbol{H_m}|E, I) \geq 0.995$

H_m: The green stains of paint (S) originate from the seized crowbar (G)
H_a: The green stains of paint (S) originate from another breaking tool

$$200 \leq \frac{P(E|H_m, I)}{P(E|H_a, I)} \leq 300$$

So, e.g. the loss function

Action	States of the world	
	H_m is true	H_a is true
Accept H_m	0	1
Accept H_a	< 200	0

with a Bayes factor of 200 and prior odds equal to 1 ("1 to 1 on") would lead to the decision "Accept H_m".

…provided the decision maker can trust on that the Bayes factor is robust.

One argument supporting such a view can be that the forensic scientists reports a range in which the Bayes factor lies (and 200 would then be interpreted as the absolute lower limit).

$\boldsymbol{H_m}$: The green stains of paint (S) originate from the seized crowbar (G)
$\boldsymbol{H_a}$: The green stains of paint (S) originate from another breaking tool

$$200 \leq \frac{P(E|\boldsymbol{H_m}, I)}{P(E|\boldsymbol{H_a}, I)} \leq 300$$

But reporting an interval instead of a single value may give the impression that the Bayes factor has random variation (which is not true!).

If the forensic scientist is convinced that the Bayes factor is at least 200, then giving the benefit of the doubt to the defendant the fact finder should use 200 and the interesting question is – how robust is that value?

$$B = \frac{P(E|\boldsymbol{H_m}, I)}{P(E|\boldsymbol{H_a}, I)}$$

The numerator of B is naturally of less interest here since (most often) deviations from the value 1 is due to sources that can be controlled for, e.g.

- laboratory issues
- quality of the evidentiary material
- choices of items if there is more than one ("The green stains originate from a crowbar belonging to the suspect" and the suspect has 3 crowbars)

$\boldsymbol{H_m}$: The green stains of paint (S) originate from the seized crowbar (G)
$\boldsymbol{H_a}$: The green stains of paint (S) originate from another breaking tool

$$B = \frac{P(E|\boldsymbol{H_m}, I)}{P(E|\boldsymbol{H_a}, I)}$$

The denominator of *B* is the crucial component. The assignment of it depends on the available information and knowledge about the variation of the features of the paint analysed among breaking tools.

One possibility of obtaining *B* = 200 is that the forensic scientist has previously investigated in total 200 breaking tools of which one had paint with these features. The relative frequency of the features becomes 1/200 which (with $P(E|\boldsymbol{H_m}, I) = 1$) gives a Bayes factor of exactly 200.

Would that value of *B* be robust?

What would happen with the Bayes factor if another breaking tool with these paint features is found?

$\boldsymbol{H_m}$: The green stains of paint (S) originate from the seized crowbar (G)
$\boldsymbol{H_a}$: The green stains of paint (S) originate from another breaking tool

$$B = \frac{P(E|\boldsymbol{H_m}, I)}{P(E|\boldsymbol{H_a}, I)}$$

Another possibility is that the forensic scientist has made a market research on crowbars and found out that on the average 137 out of 20 000 sold crowbars had green paint.
Moreover, the forensic scientist had bought 5 randomly chosen green crowbars and found out that 3 of these had paint with the same features as the seized crowbar.

With these data the numerator of B can be assigned as

$$P(E|\boldsymbol{H_a}, I) = P(\text{"green and with observed features"}|\boldsymbol{H_a}, I)$$

$$= P(\text{"observed features"}|\text{"green"}, \boldsymbol{H_a}, I) \times P(\text{"green"}|\boldsymbol{H_a}, I) = \frac{3}{5} \times \frac{137}{20000} \approx$$

$$\approx 0.00411$$

which (with $P(E|\boldsymbol{H_m}, I) = 1$) gives a Bayes factor of approx. 243 rounded down to 200.

Would that value of B be robust?

Is it sufficient to do the market research on crowbars only?

$\boldsymbol{H_m}$: The green stains of paint (S) originate from the seized crowbar (G)
$\boldsymbol{H_a}$: The green stains of paint (S) originate from another breaking tool

$$B = \frac{P(E|\boldsymbol{H_m}, I)}{P(E|\boldsymbol{H_a}, I)}$$

Going back to the situation where the forensic scientist has investigated 200 breaking tools of which one had paint with the features of interest.

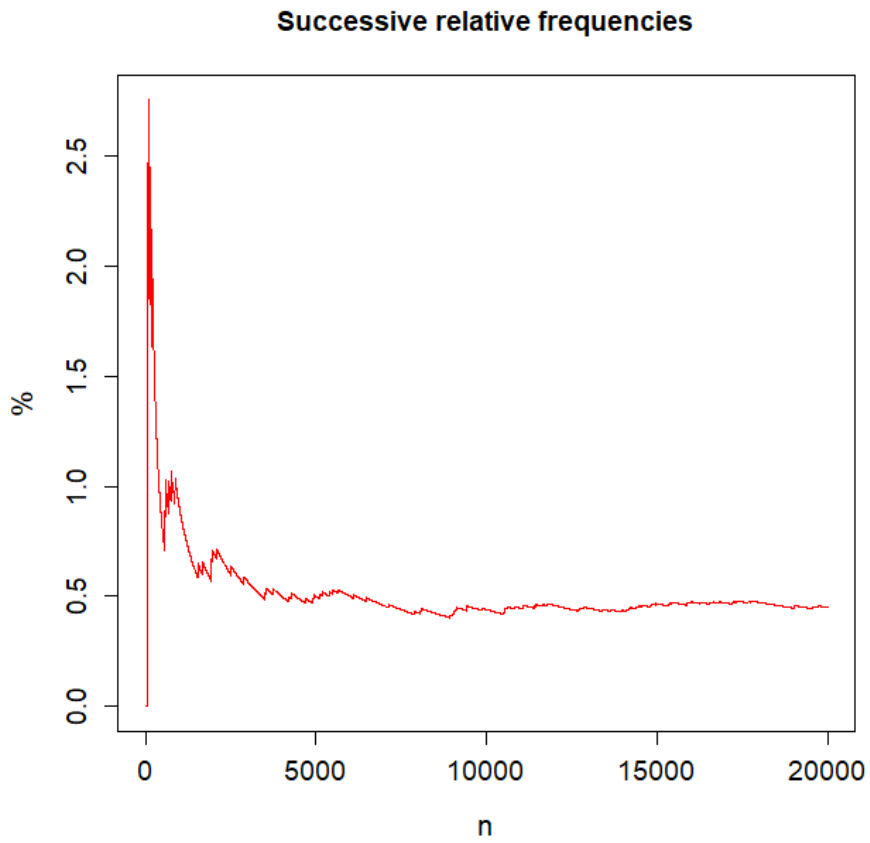
Assume that the true proportion of breaking tools with green paint and the features of interest (E) is 1/200 = 0.5 %.

Simulate, say 20 000 instances of observing the paint features of (independent) crowbars with that proportion (0.5%) of E.

Plot the successively obtained relative frequencies.

$\boldsymbol{H_m}$: The green stains of paint (S) originate from the seized crowbar (G)
$\boldsymbol{H_a}$: The green stains of paint (S) originate from another breaking tool

$$B = \frac{P(E|\boldsymbol{H_m}, I)}{P(E|\boldsymbol{H_a}, I)}$$



How big must *n* be so that $P(E|\boldsymbol{H_a}, I(n))$ can be considered as sufficiently robust for the fact finder to make the decision "Accept $\boldsymbol{H_m}$"?

That value of *n* would be the size of the sample in a complementary study of breaking tools needed.

Note 1: Have we by this suddenly become frequentists?

Note 2: Wouldn't it be better to find the expression for, say a 99 % credible interval for the proportion of interest and what value of *n* would make this interval as narrow as requested?