Meeting 14 and 15:
The decisive approach to statistical inference.

The decisive approach to statistical inference, part I

Point estimation of an unknown parameter θ:

The <u>decision rule</u> is a point estimator (the functional form): $\delta(\tilde{\boldsymbol{x}}) = \hat{\theta}(\tilde{\boldsymbol{x}})$

The <u>action</u> is a particular point estimate. $a = \hat{\theta}_{obs} = \hat{\theta}(\boldsymbol{x})$

<u>State of nature</u> is the true value of θ.

The loss function is a measure of how far away the estimator is from θ:

$$L(\delta(\tilde{\boldsymbol{x}}), \theta) = L\left(\hat{\theta}, \theta\right)$$

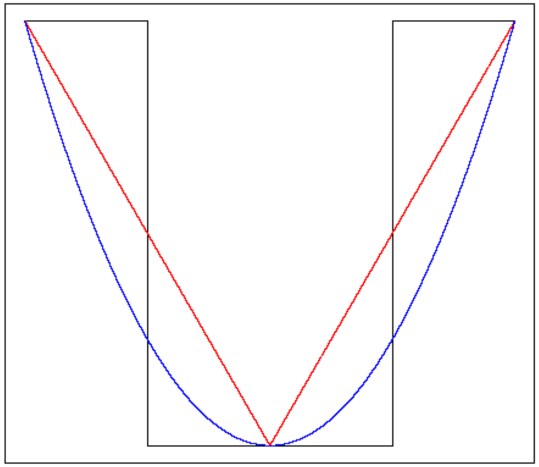
<u>Prior information</u> is quantified by the prior distribution (pdf/pmf) $f'(\theta)$.

<u>Data</u> is the random sample $\boldsymbol{x}$ from a distribution with (pdf/pmf) $f(\boldsymbol{x}|\theta)$.

Three simple loss functions (univariate case)

Zero-one loss:

$$L\left(\hat{\theta},\theta\right)=\begin{cases}0 & \left|\hat{\theta}-\theta\right|<m\\ k & \left|\hat{\theta}-\theta\right|\geq m\end{cases}\quad k,m>0$$



Absolute error loss:

$$L\left(\hat{\theta},\theta\right)=k\cdot\left|\hat{\theta}-\theta\right|\quad k>0$$

Quadratic (error) loss (or squared loss):

$$L\left(\hat{\theta},\theta\right)=k\cdot\left(\hat{\theta}-\theta\right)^{2}\quad k>0$$

Bayes estimators:

A Bayes estimator is the estimator that minimizes the expected posterior loss:

$$E\left(L\left(\hat{\theta}(\boldsymbol{x})\middle|\boldsymbol{x}\right)\right) = \int_{\theta} L\left(\hat{\theta}(\boldsymbol{x}), \theta\right) \cdot f''(\theta|\boldsymbol{x})d\theta$$

$$\Rightarrow \hat{\theta}_B(\boldsymbol{x}) = \min_{\delta}\left(\int_{\theta} L(\delta(\boldsymbol{x}), \theta) \cdot f''(\theta|\boldsymbol{x})d\theta\right)$$

Minimization with respect to different loss functions will result in measures of location in the *posterior* distribution of θ.

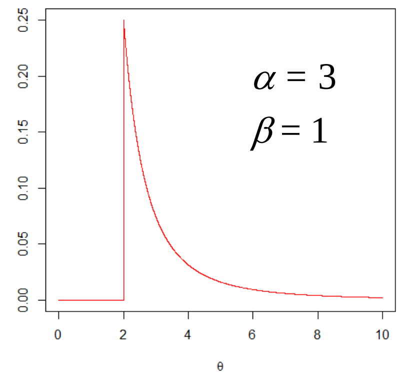
Zero-one loss: $\hat{\theta}_B(\boldsymbol{x})$ is the posterior mode of $\tilde{\theta}$ given $\boldsymbol{x}$

Absolute error loss: $\hat{\theta}_B(\boldsymbol{x})$ is the posterior median of $\tilde{\theta}$ given $\boldsymbol{x}$

Quadratic loss: $\hat{\theta}_B(\boldsymbol{x})$ is the posterior mean of $\tilde{\theta}$ given $\boldsymbol{x}$: $E\left(\tilde{\theta}\middle|\boldsymbol{x}\right)$

Example

Assume we have a sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ from $U(0, \theta)$ and that a prior density for θ is the Pareto density



$$f'(\theta|\alpha,\beta) = (\alpha-1)\cdot\beta^{\alpha-1}\cdot\theta^{-\alpha} \quad ,\theta \geq 2; \alpha > 1; \beta > 0$$

What is the Bayes estimator of θ under quadratic loss?

The posterior distribution is also Pareto with

$$x_{(n)} = max\{x_1, \ldots, x_n\}$$

$$f''(\theta|n,\boldsymbol{x},\alpha,\beta) = (\alpha+n-1)\cdot\left(\max\{\beta,x_{(n)}\}\right)^{\alpha+n-1}\cdot\theta^{-(\alpha+n)} \quad ,\theta \geq \max\{\beta,x_{(n)}\}$$

$$\Rightarrow \hat{\theta}_B = E(\tilde{\theta}|\boldsymbol{x}) = \int_{\theta=\max\{\beta,x_{(n)}\}}^{\infty} \theta\cdot(\alpha+n-1)\cdot\left(\max\{\beta,x_{(n)}\}\right)^{\alpha+n-1}\cdot\theta^{-(\alpha+n)}d\theta =$$

$$= (\alpha+n-1)\cdot\left(\max\{\beta,x_{(n)}\}\right)^{\alpha+n-1}\cdot\int_{\theta=\max\{\beta,x_{(n)}\}}^{\infty} \theta\cdot\theta^{-(\alpha+n-1)}d\theta =$$

$$= \frac{\alpha+n-1}{\alpha+n-2}\max\{\beta,x_{(n)}\}$$

Compare with $\hat{\theta}_{MLE} = x_{(n)}$

Minimax estimators:

Find the value of θ that maximizes the expected loss with respect to the sample values, i.e. that maximizes the risk over the set of estimators.

Then, the particular estimator that minimizes the risk for that value of θ is the minimax estimator.

$$\hat{\theta}_{\text{minimax}} = \operatorname*{argmin}_{\delta}\left(\max_{\theta} D(\delta, \theta)\right)$$

Usually difficult to find minimax estimators, but there is one method to find it <u>via a Bayes' estimator.</u>

<u>Theorem</u> (actually a corollary of a theorem both presented in "Lehmann E.L. *Theory of point estimation*. Wiley, 1983")

If a Bayes' estimator has constant risk. [*i.e. not dependent on* θ] it is also a minimax estimator.

Example We wish to estimate the parameter p under quadratic loss in a binomial sampling model for sample size n. Hence we (will) have observed a random variable $\tilde{r}$ that is $Bi(n, p)$.

We (should) know that (with n fixed) the maximum-likelihood estimator of p is

$$\hat{p}_{MLE} = \frac{\tilde{r}}{n}$$

but can we find a minimax estimator under quadratic loss?

A Bayes' estimator under quadratic loss is the posterior mean of p, hence we need to specify a prior distribution – Natural to use the conjugate beta distribution with density function

$$f'(p|a,b) = \frac{p^{a-1} \cdot (1-p)^{b-1}}{B(a,b)} \quad 0 \leq p \leq 1 \qquad B(a,b) = \frac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a+b)}$$

The posterior density function is then

$$f''(p|n,r,a,b) = \frac{p^{a+r-1} \cdot (1-p)^{b+n-r-1}}{B(a+r, b+n-r)}$$

and the Bayes' estimator is its mean:

$$\hat{p}_B = \hat{p}_B(\tilde{r}) = \frac{a+\tilde{r}}{a+r+b+n-r} = \frac{a+\tilde{r}}{a+b+n}$$

$$\hat{p}_B = \frac{a+\tilde{r}}{a+b+n}$$

The risk function for this estimator is

$$D(\hat{p}_B, p) = E_{\tilde{r}}\left((\hat{p}_B - p)^2\right) = \sum_{r=0}^{n} (\hat{p}_B - p)^2 \cdot \binom{n}{r} \cdot p^r \cdot (1-p)^r =$$

$$= \sum_{r=0}^{n} \left(\frac{a+r}{a+b+n} - p\right)^2 \cdot \binom{n}{r} \cdot p^r \cdot (1-p)^r =$$

$$= \sum_{r=0}^{n} \left[\left(\frac{a^2}{(a+b+n)^2} - \frac{2\cdot a\cdot p}{a+b+n} + p^2\right) + \left(\frac{2\cdot a}{(a+b+n)^2} - \frac{2\cdot p}{a+b+n}\right)\cdot r +\right.$$

$$\left. + \frac{1}{(a+b+n)^2}\cdot r^2\right] \cdot \binom{n}{r} \cdot p^r \cdot (1-p)^r = \frac{a^2}{(a+b+n)^2} - \frac{2\cdot a\cdot p}{a+b+n} + p^2 +$$

$$+ \left(\frac{2\cdot a}{(a+b+n)^2} - \frac{2\cdot p}{a+b+n}\right) \cdot \underbrace{\sum_{r=0}^{n} r \cdot \binom{n}{r} \cdot p^r \cdot (1-p)^r}_{E(\tilde{r}|p)} + \frac{1}{(a+b+n)^2} \cdot \underbrace{\sum_{r=0}^{n} r^2 \cdot \binom{n}{r} \cdot p^r \cdot (1-p)^r}_{\substack{E(\tilde{r}^2|p) = \\ Var(\tilde{r}|p) + \left(E(\tilde{r}|p)\right)^2}}$$

$$E(\tilde{r}|p) = n \cdot p \qquad\qquad \hat{p}_B = \frac{a + \tilde{r}}{a + b + n}$$

$$Var(\tilde{r}|p) = n \cdot p \cdot (1 - p)$$

$$\Longrightarrow D(\hat{p}_B, p) = \frac{a^2}{(a+b+n)^2} - \frac{2 \cdot a \cdot p}{a+b+n} + p^2 + \left(\frac{2 \cdot a}{(a+b+n)^2} - \frac{2 \cdot p}{a+b+n}\right) \cdot n \cdot p +$$

$$+ \frac{1}{(a+b+n)^2} \cdot (n \cdot p \cdot (1-p) + (n \cdot p)^2) = \cdots =$$

$$= \frac{1}{(a+b+n)^2} \cdot \left(a^2 + \big(n - 2 \cdot a \cdot (a+b)\big) \cdot p + \big((a+b)^2 - n\big) \cdot p^2\right)$$

This risk function will be constant (for fixed *n*) if

$$\big(n - 2 \cdot a \cdot (a+b)\big) = 0 \qquad \text{and} \qquad \big((a+b)^2 - n\big) = 0$$

$$\Leftrightarrow a = b = \frac{\sqrt{n}}{2}$$

Hence, the estimator

$$\hat{p}_B = \frac{\frac{\sqrt{n}}{2} + \tilde{r}}{\frac{\sqrt{n}}{2} + \frac{\sqrt{n}}{2} + n} = \frac{\sqrt{n} + 2 \cdot \tilde{r}}{\sqrt{n} + n}$$

is a Bayes' estimator with constant risk, and according to the theorem above it is also a minimax estimator.

The value of the constant (but actually n-dependent) risk is

$$\frac{1}{\left(\frac{\sqrt{n}}{2} + \frac{\sqrt{n}}{2} + n\right)^2} \cdot \left(\frac{n}{4} + 0 \cdot p + 0 \cdot p^2\right) = \frac{n}{4 \cdot (\sqrt{n} + n)^2} = \frac{1}{4 \cdot (1 + \sqrt{n})^2}$$

Exercise:

Is $\hat{p}_B$ unbiased? What is the risk of the unbiased $\hat{p}_{MLE} = \tilde{r}/n$? For which range of n is the risk of $\hat{p}_B$ lower than the risk of $\hat{p}_{MLE}$?

In an inferential setup we may work with *propositions* or *hypotheses*.

A hypothesis is a central component in all building of science.

The "standard situation" would be that we have two hypotheses at a time:

H_0 The forwarded hypothesis

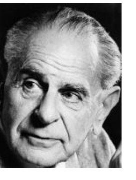
H_1 The alternative hypothesis

These must be mutually exclusive.

Successive falsification of hypotheses (cf. Popper[1]) until only one is left is one strategy for science building.

From a perspective of statistical inference "falsification" is never a decision with 100% certainty, and there are different ways of handling this uncertainty.

[1]Popper K., Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge, London, 1963



Classical statistical hypothesis testing

(Neyman J. and Pearson E.S. , 1933)

The two hypotheses are different explanations to the *Data*.
$\Rightarrow$ Each hypothesis provides *model(s)* for *Data*

The purpose is to use *Data* to try to falsify H_0.

Decision is in one direction only.

Type-I-error: Falsifying a true H_0
Type-II-error: Not falsifying a false H_0

Size or Significance level: $\alpha = P(\text{Type-I-error})$

If each hypothesis provides one and only one model for *Data*:

Power: $1 - P(\text{Type-II-error}) = 1 - \beta$

Both hypotheses are then referred to as *simple hypotheses*

Most powerful test for ***simple*** hypotheses (Neyman-Pearson lemma):

$$\text{Reject (falsify) } H_0 \text{ when } \frac{\mathcal{L}(H_1|Data)}{\mathcal{L}(H_0|Data)} \geq A$$

where $\mathcal{L}(H_0|Data)$ and $\mathcal{L}(H_1|Data)$ are the likelihoods of H_0 and H_1 respectively (notation with calligraphic $\mathcal{L}$ to not confuse with loss function).

… and where $A > 0$ is chosen so that

$$P\left(\frac{\mathcal{L}(H_1|\text{Data})}{\mathcal{L}(H_0|\text{Data})} \geq A \,\middle|\, H_0\right) = \alpha$$

This minimises β for fixed α.

Note that the probability is taken with respect to ***Data*** , i.e. with respect to the probability model for *Data* given H_0.

Extension to *composite* hypotheses: Uniformly most powerful test (UMP)

Example: A seizure of pills, suspected to be Ecstasy, is sampled for the purpose of investigating whether the proportion of Ecstasy pills is "around" 80% or "around" 50%.

In a sample of 50 pills, 39 proved to be Ecstasy pills.

As the forwarded hypothesis we can formulate

H_0: Around 80% of the pills in the seizure are Ecstasy

and as the alternative hypothesis

H_1: Around 50% of the pills in the seizure are Ecstasy

The likelihood of the two hypotheses are

$\mathcal{L}$ (H_0 | *Data*) = Probability of obtaining 39 Ecstasy pills out of 50 sampled when the seizure proportion of Ecstasy pills is 80%.

$\mathcal{L}$ (H_1 | *Data*) = Probability of obtaining 39 Ecstasy pills out of 50 sampled when the seizure proportion of Ecstasy pills is 50%.

Assuming a large seizure these probabilities can be calculated using a binomial sampling model *Bin*(50, *p*), where H_0 states that $p = p_0 = 0.8$ and H_1 states that $p = p_1 = 0.5$.

In generic form, if we have obtained *x* Ecstasy pills out of *n* sampled:

$$\mathcal{L}(H_0|Data) = \mathcal{L}(H_0|x,(n)) = \binom{n}{x} \cdot p_0^x \cdot (1 - p_0)^{n-x}$$

$$\mathcal{L}(H_1|Data) = \mathcal{L}(H_1|x,(n)) = \binom{n}{x} \cdot p_1^x \cdot (1 - p_1)^{n-x}$$

The Neyman-Pearson lemma now states that the most powerful test is of the form

$$\frac{\mathcal{L}(H_1|Data)}{\mathcal{L}(H_0|Data)} \geq A \Rightarrow \frac{p_1^x \cdot (1-p_1)^{n-x}}{p_0^x \cdot (1-p_0)^{n-x}} = \left(\frac{p_1}{p_0}\right)^x \cdot \left(\frac{1-p_1}{1-p_0}\right)^{n-x} \geq A$$

$$\Leftrightarrow$$

$$x \cdot \ln\left(\frac{p_1}{p_0}\right) + (n-x) \cdot \ln\left(\frac{1-p_1}{1-p_0}\right) \geq \ln A$$

$$\Leftrightarrow$$

$$x \leq \frac{\ln A - n \cdot \ln\left(\frac{1-p_1}{1-p_0}\right)}{\ln\left(\frac{p_1}{p_0}\right) - \ln\left(\frac{1-p_1}{1-p_0}\right)}$$

$$= C(n) \quad \left\langle \text{since } \; p_1 < p_0 \Rightarrow \ln\left(\frac{p_1}{p_0}\right) - \ln\left(\frac{1-p_1}{1-p_0}\right) < 0 \right\rangle$$

Hence, H_0 should be rejected in favour of H_1 as soon as $x \leq C$

How to choose C?

Normally, we would set the significance level α and the find C so that

$$P\left(X \leq C \middle| H_0\right) = \alpha$$

If α is chosen to 0.05 we can search the binomial distribution valid under H_0 for a value C such that

$$\sum_{k=0}^{C} P\left(X = k \middle| H_0\right) \leq 0.05 \Rightarrow \sum_{k=0}^{C} \binom{50}{k} \cdot 0.8^k \cdot 0.2^{50-k} \leq 0.05$$

MSExcel:

`BINOM.INV(50;0.8;0.05)` returns the lowest value of B for which the sum is <u>at least</u> 0.05 $\Rightarrow$ *35*

`BINOM.DIST(35;50;0.8;TRUE)` $\Rightarrow$ *0.06072208*
`BINOM.DIST(34;50;0.8;TRUE)` $\Rightarrow$ *0.030803423*

$\Rightarrow$ Choose $C = 34$. $\Rightarrow$ Since $x = 39$ we cannot reject H_0

Drawbacks with the classical approach

- *Data* alone "decides". Small amounts of data $\Rightarrow$ Low power
- Difficulties in interpretation:

When H_0 is rejected, it means

"If we repeat the collection of data under (in principal) identical circumstances

then in (at most) 100α % of all cases when H_0 is true $\dfrac{\mathcal{L}(H_1|Data)}{\mathcal{L}(H_0|Data)} \geq A$ "

Can we (always) repeat the collection of data?

- "Falling off the cliff" – What is the difference between "just rejecting" and "almost rejecting" ?
- "Isolated" falsification (or no falsification) – Tests using other data but with the same hypotheses cannot be easily combined

The Bayesian Approach

There is always a process that leads to the formulation of the hypotheses.
$\Rightarrow$ A *prior probability* exists for each of them:

$$p_0 = P(H_0|I) = P(H_0)$$
$$p_1 = P(H_1|I) = P(H_1)$$
$$p_0 + p_1 = 1$$

Simpler expressed as *prior odds* for the hypothesis H_0:

$$Odds(H_0|I) = \frac{p_0}{p_1} = \frac{P(H_0|I)}{P(H_1|I)}$$

Non-informative priors: $p_0 = p_1 = 0.5$ gives prior odds = 1

Data should help us calculating *posterior odds*

$$Odds(H_0|Data,I) = \frac{P(H_0|Data,I)}{P(H_1|Data,I)} = \frac{q_0}{q_1}$$

$$\Rightarrow$$

$$q_0 = P(H_0|Data,I) = \frac{Odds(H_0|Data,I)}{Odds(H_0|Data,I) + 1}$$

The "hypothesis testing" is replaced by a judgement upon whether q_0 is

- small enough to make us believe in H_1 (*falsifying* H_0)
- large enough to make us believe in H_0 (*falsifying* H_1)

Confirming/Undermining support of H_0.

i.e. no pre-setting of the decision direction is made.

How can we obtain the posterior odds?

The odds ratio (posterior odds/prior odds) is know as the *Bayes factor*:

$$B = \frac{Odds(H_0|Data, I)}{Odds(H_0|I)} = \frac{P(H_0|Data, I)/P(H_1|Data, I)}{P(H_0|I)/P(H_1|I)}$$

$\Rightarrow$

$$Odds(H_0|Data, I) = B \cdot Odds(H_0|I)$$

Hence, if we know the Bayes factor, we can calculate the posterior odds (since we can always set the prior odds).

There are different situations depending on the complexities of the hypotheses and the probability measure applicable to *Data*.

1. Both hypotheses are simple, i.e. they each give one and only one model for *Data*

   a) Distinct probabilities can be assigned to *Data*

   Bayes' theorem on odds-form then gives

$$\frac{P(H_0|Data,I)}{P(H_1|Data,I)} = \frac{P(Data|H_0,I)}{P(Data|H_1,I)} \cdot \frac{P(H_0|I)}{P(H_1|I)}$$

   Hence, the Bayes factor is

$$B = \frac{P(Data|H_0,I)}{P(Data|H_1,I)}$$

   The probabilities of the numerator and denominator respectively can be calculated (estimated) using the model provided by respective hypothesis.

b) *Data* is the observed value $\boldsymbol{x}$ of a continuous (possibly multidimensional) random variable

It can be shown that

$$\frac{P(H_0|Data,I)}{P(H_1|Data,I)} = \frac{f(\boldsymbol{x}|H_0,I)}{f(\boldsymbol{x}|H_1,I)} \cdot \frac{P(H_0|I)}{P(H_1|I)}$$

where $f(\boldsymbol{x} \mid H_0, I)$ and $f(\boldsymbol{x} \mid H_1, I)$ are the probability density functions given by the models specified by H_0 and H_1 respectively.

Hence, the Bayes factor is

$$B = \frac{f(\boldsymbol{x}|H_0,I)}{f(\boldsymbol{x}|H_1,I)}$$

Known (or estimated) density functions under each model can then be used to calculate the Bayes factor.

In both cases we can see that the Bayes factor is a *likelihood ratio* since the numerator and denominator are likelihoods for respective hypothesis.

$$\Rightarrow \quad B = \frac{\mathcal{L}(H_0|Data, I)}{\mathcal{L}(H_1|Data, I)}$$

Example Ecstasy pills revisited

H_0: Around 80% of the pills in the seizure are Ecstasy
H_1: Around 50% of the pills in the seizure are Ecstasy

The likelihoods for the hypotheses are

$$\mathcal{L}(H_0|Data) = \binom{50}{39} \cdot 0.8^{39} \cdot 0.2^{11} \approx 0.1271082$$

$$\mathcal{L}(H_1|Data) = \binom{50}{39} \cdot 0.5^{39} \cdot 0.5^{11} \approx 3.317678e - 05$$

$$\Rightarrow B \approx \frac{0.1271082}{3.317678e - 05} \approx 3831$$

Hence, *Data* are 3831 times more probable if H_0 is true compared to if H_1 is true.

Assume we have no particular belief in any of the two hypothesis *prior* to obtaining the data.

$$\Rightarrow Odds(H_0) = 1$$

$$\Rightarrow Odds(H_0|Data) \approx 3831 \cdot 1$$

$$\Rightarrow P(H_0|Data) = \frac{3831}{3831 + 1} \approx 0.9997$$

Hence, upon the analysis of data we can be 99.97% certain that H_0 is true.

Note however that it may be unrealistic to assume only two possible proportions of Ecstasy pills in the seizure!

2. The hypothesis H_0 is simple but the hypothesis H_1 is *composite*, i.e. it provides several models for *Data* (several explanations)

The various models of H_1 would (in general) provide different likelihoods for the different explanations.
$\Rightarrow$ We cannot come up with one unique likelihood for H_1.

If in addition, the different explanations have <u>different prior probabilities</u> we have to <u>weigh</u> the different likelihoods with these.

If the composition in H_1 is in form of a set of <u>discrete</u> alternatives, the Bayes factor can be written

$$B = \frac{\mathcal{L}(H_0|Data)}{\sum_i \mathcal{L}(H_{1i}|Data) \cdot P\left(H_{1i}\middle|H_1\right)}$$

where $P(H_{1i} \mid H_1)$ is the conditional prior probability that H_{1i} is true given that H_1 is true (*relative prior*) , and the sum is over all alternatives H_{11} , H_{12} , …

$$B = \frac{\mathcal{L}(H_0|Data)}{\sum_i \mathcal{L}(H_{1i}|Data) \cdot P\left(H_{1i}\middle|H_1\right)}$$

If the relative priors are (fairly) equal the denominator reduces to the *average* likelihood of the alternatives.

If the likelihoods of the alternatives are equal the denominator reduces to that likelihood since the relative priors sum to one.

If the composition is defined by a <u>continuously</u> valued parameter, θ we must use the <u>conditional prior density</u> of θ given that H_1 is true: $f'\left(\theta\middle|H_1\right)$ and integrate the likelihood with respect to that density.

$\Rightarrow$ The Bayes factor can be written

$$B = \frac{\mathcal{L}(H_0|Data)}{\int_{"\theta \in H_1"} \mathcal{L}(\theta|Data) \cdot f'\left(\theta\middle|H_1\right) d\theta}$$

3. Both hypothesis *are composite*, i.e. each provides several models for *Data* (several explanations)

This gives different sub-cases, depending on whether the compositions in the hypotheses are discrete or according to a continuously valued parameter.

The "discrete-discrete" case gives the Bayes factor

$$B = \frac{\sum_j \mathcal{L}\left(H_{0_j} | Data\right) \cdot P\left(H_{0j} \middle| H_0\right)}{\sum_i \mathcal{L}(H_{1i} | Data) \cdot P\left(H_{1i} \middle| H_1\right)}$$

and the "continuous-continuous" case gives the Bayes factor

$$B = \frac{\int_{"\theta \in H_0"} \mathcal{L}(\theta | Data) \cdot f'\left(\theta \middle| H_0\right) d\theta}{\int_{"\theta \in H_1"} \mathcal{L}(\theta | Data) \cdot f'\left(\theta \middle| H_1\right) d\theta}$$

where $f'\left(\theta \middle| H_0\right)$ is the conditional prior density of θ given that H_0 is true.

Example Ecstasy pills revisited again

Assume a more realistic case where we from a sample of the seizure shall investigate whether the proportion of Ecstasy pills is <u>higher than</u> 80%.

$\Rightarrow$

H_0: Proportion $\theta > 0.8$

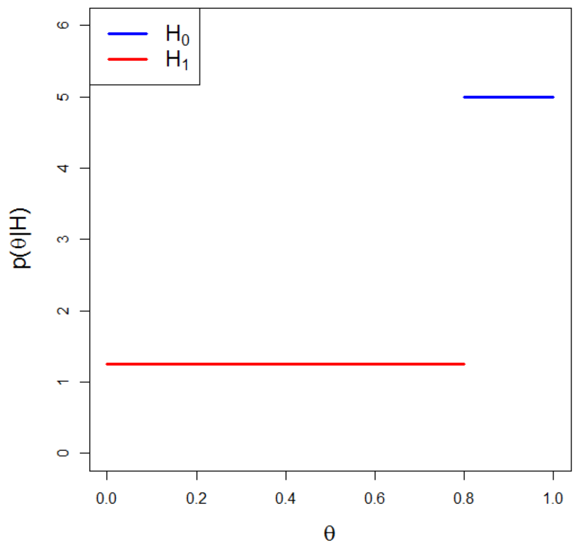
H_1: Proportion $\theta \leq 0.8$

i.e. both are composite

We further assume that all θ within the region of each hypothesis are equally likely, hence having uniform distributions. The <u>conditional prior densities</u> for θ under each hypothesis can thus be defined as

$$f'(\theta|H_0) = \begin{cases} \dfrac{1}{1-0.8} = 5 & 0.8 < \theta \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f'(\theta|H_1) = \begin{cases} \dfrac{1}{0.8-0} = 1.25 & 0 \leq \theta \leq 0.8 \\ 0 & \text{otherwise} \end{cases}$$



The likelihood *function* is (irrespective of the hypotheses)

$$\mathcal{L}(\theta|Data) = \binom{50}{39} \cdot \theta^{39} \cdot (1-\theta)^{11}$$

Then, the Bayes factor is

$$B = \frac{\int_\theta \mathcal{L}(\theta|Data) \cdot f'(\theta|H_0) d\theta}{\int_\theta \mathcal{L}(\theta|Data) \cdot f'(\theta|H_1) d\theta} = \frac{\int_{0.8}^{1} \binom{50}{39} \cdot \theta^{39} \cdot (1-\theta)^{11} \cdot 5 d\theta}{\int_{0}^{0.8} \binom{50}{39} \cdot \theta^{39} \cdot (1-\theta)^{11} \cdot 1.25 d\theta}$$

$$=$$

$$= 4 \cdot \frac{\int_{0.8}^{1} \theta^{39} \cdot (1-\theta)^{11} \cdot 1 d\theta}{\int_{0}^{0.8} \theta^{39} \cdot (1-\theta)^{11} \cdot 1 d\theta}$$

How do we solve these integrals?

The Beta distribution:

(We should know that) a random variable is said to have a Beta distribution with parameters *a* and *b* if its probability density function is

$$\boxed{\begin{array}{l} f(x) = C \cdot x^{a-1} \cdot (1-x)^{b-1} \;;\; 0 \le x \le 1 \\ \text{with } C = \displaystyle\int_0^1 x^{a-1} \cdot (1-x)^{b-1} dx = \mathrm{B}(a,b) \end{array}}$$

Hence, we can identify the integrals of the Bayes factor as proportional to different probabilities of <u>the same</u> beta distribution

$$\frac{\int_{0.8}^{1} \theta^{39} \cdot (1-\theta)^{11} d\theta}{\int_{0}^{0.8} \theta^{39} \cdot (1-\theta)^{11} d\theta} = \frac{\int_{0.8}^{1} C \cdot \theta^{39} \cdot (1-\theta)^{11} d\theta}{\int_{0}^{0.8} C \cdot \theta^{39} \cdot (1-\theta)^{11} d\theta}$$

$$= \frac{\int_{0.8}^{1} C \cdot \theta^{40-1} \cdot (1-\theta)^{12-1} d\theta}{\int_{0}^{0.8} C \cdot \theta^{40-1} \cdot (1-\theta)^{12-1} d\theta}$$

namely a beta distribution with parameters a = 40 and b =12.

```
> num <- 1-pbeta(q=0.8,shape1=40,shape2=12)
> den <- 1 - num
> num
[1] 0.314754
> den
[1] 0.685246
> ratio <- num/den
> B <- 4*ratio
> B
[1] 1.83732
```

Hence, the Bayes factor is 1.83732.

With even prior odds ($Odds(H_0) = 1$) we get the posterior odds equal to the Bayes factor and the posterior probability of H_0 is

$$P(H_0|Data) = \frac{1.83732}{1.83732 + 1} \approx 0.65$$

$\Rightarrow$ *Data* does not provide us with evidence clearly against any of the hypotheses.

Finite action problems revisited

So far the confirming/undermining of a hypothesis has been made by the calculation of the *posterior odds*:

$$\frac{P(H_0|\boldsymbol{x})}{P(H_1|\boldsymbol{x})} = B \cdot \frac{P(H_0)}{P(H_1)}$$

Concluding which of H_0 and H_1 should be the hypothesis to be retained has thus been a question about whether the posterior probability of one of the hypothesis is "high enough".

Coupling the posterior probabilities with losses (or utilities) will define a decision problem.

The loss function is

Action	State of nature	
	H_0 true	H_1 true
Accept H_0	0	c_0
Accept H_1	c_1	0

c_0 : Cost of accepting H_0 when H_1 is true

c_1 : Cost of accepting H_1 when H_0 is true

The Bayes action is the action that minimises the expected posterior loss:

Action	Expected posterior loss
Accept H_0	$0 \cdot \Pr(H_0 \mid \boldsymbol{x}) + c_0 \cdot \Pr(H_1 \mid \boldsymbol{x}) = c_0 \cdot \Pr(H_1 \mid \boldsymbol{x})$
Accept H_1	$c_1 \cdot \Pr(H_0 \mid \boldsymbol{x}) + 0 \cdot \Pr(H_1 \mid \boldsymbol{x}) = c_1 \cdot \Pr(H_0 \mid \boldsymbol{x})$

Example: Return again to the example with dye on banknotes

The posterior probabilities were obtained before (Meeting 1):

P(“Dye is present”|”Positive detection) = 0.047

P(“Dye is not present”|”Positive detection”) = 0.953

The loss function proposed in Meeting 10 was (transformed from disutilities):

Action	State of the world	
	Dye is present (H_0)	Dye is not present (H_1)
Destroy banknote	0	100
Use banknote	5000	0

Hence,

Action	Expected posterior loss
Destroy banknote	$0 \cdot 0.047 + 100 \cdot 0.953 = 95.3$
Use banknote	$5000 \cdot 0.047 + 0 \cdot 0.953 = 235.0$

Minimising the expected posterior loss gives the action “Destroy banknote”.

How low must the fine be for the action to be changed?

General decision-theoretic approach

A loss function of “0 – k“ type is used (there may be two different values of k):

Action	States of the world	
	H_0 is true	H_1 is true
Accept H_0	0	L(Type-II-error)= L(II)
Accept H_1	L(Type-I-error)= L(I)	0

Expected posterior losses (assuming availability of data ***x***):

$$\text{Action is "Accept } H_0\text{":} \quad 0 \cdot \Pr(H_0|\mathrm{x}) + L(\mathrm{II}) \cdot \Pr(H_1|\mathrm{x}) = L(\mathrm{II}) \cdot \Pr(H_1|\mathrm{x})$$
$$\text{Action is "Accept } H_1\text{":} \quad L(\mathrm{I}) \cdot \Pr(H_0|\mathrm{x}) + 0 \cdot \Pr(H_1|\mathrm{x}) = L(\mathrm{I}) \cdot \Pr(H_0|\mathrm{x})$$

Hence the optimal action would be "Accept H_0" when

$$L(\mathrm{II}) \cdot P(H_1|\mathrm{x}) < L(\mathrm{I}) \cdot P(H_0|\mathrm{x}) \Leftrightarrow \frac{P(H_0|\mathrm{x})}{P(H_1|\mathrm{x})} > \frac{L(\mathrm{II})}{L(\mathrm{I})}$$

$$\Leftrightarrow$$

$$B \cdot \frac{P(H_0)}{P(H_1)} > \frac{L(\mathrm{II})}{L(\mathrm{I})} \Leftrightarrow B > \frac{P(H_1)}{P(H_0)} \cdot \frac{L(\mathrm{II})}{L(\mathrm{I})}$$

… and the optimal action would be "Accept H_1" when

$$L(\mathrm{II}) \cdot P(H_1|\mathrm{x}) > L(\mathrm{I}) \cdot P(H_0|\mathrm{x}) \Leftrightarrow \frac{P(H_0|\mathrm{x})}{P(H_1|\mathrm{x})} < \frac{L(\mathrm{II})}{L(\mathrm{I})}$$

$$\Leftrightarrow$$

$$B \cdot \frac{P(H_0)}{P(H_1)} < \frac{L(\mathrm{II})}{L(\mathrm{I})} \Leftrightarrow B < \frac{P(H_1)}{P(H_0)} \cdot \frac{L(\mathrm{II})}{L(\mathrm{I})}$$

Return to example with banknotes:

P(“Dye is present”|”Positive detection) = $P(H_0 \mid \boldsymbol{x}) = 0.047$

P(“Dye is not present”|”Positive detection”) = $P(H_0 \mid \boldsymbol{x}) = 0.953$

$L(\mathrm{I}) = 5000$

$L(\mathrm{II}) = 100$

$$\Rightarrow \frac{P(H_0|\mathrm{x})}{P(H_1|\mathrm{x})} = \frac{0.047}{0.953} \approx 0.049 \quad ; \quad \frac{L(\mathrm{II})}{L(\mathrm{I})} = \frac{100}{5000} = 0.02$$

Since $0.049 > 0.02$ we should accept H_0, i.e. believe that dye is present, and hence destroy the banknote.

For accepting H_1 (and use the banknote), then fine ($L(\mathrm{I})$) must satisfy

$$\frac{0.047}{0.953} < \frac{100}{L(\mathrm{I})} \quad \Rightarrow \; L(\mathrm{I}) < \frac{100 \cdot 0.953}{0.047} \approx 2028$$