Statistical Methods
(summer term 2024)

Testing hypotheses I

(based on original lectures by Prof. Dr. N. Christlieb and Dr. Hans-G. Ludwig)

Dr Yiannis Tsapras

ZAH – Heidelberg

Overview

- Testing hypotheses I: concepts of Bayesian and orthodox (classical) testing
- Testing hypotheses II: array of orthodox tests ...

Testing hypotheses: Basic ideas by example

Consider tossing two similar coins:
coin 0: p(heads)=0.5 (fair coin)
coin 1: p(heads)=0.7 (unfair coin)

Problem: somebody picks up one of the coins, tosses it 10 times in a row and tells you the number of heads they got. On the basis of the number of heads alone, your task is to decide whether it was the fair (0) or the unfair coin (1)

How do you decide?

(FYI: For those wondering what coin this is, it's a gold Double Eagle $20 coin from 1849 and it's worth a pretty penny)

Testing hypotheses: Basic ideas by example

More formal:
Hypothesis H_0: coin 0 was tossed
Hypothesis H_1: coin 1 was tossed

- *Simple hypothesis*: we know which probability distribution to use
  - both alternatives are simple to express
  - Binomial distribution: $P(X = x) = \binom{n}{k} p^k (1-p)^{n-k}$, X number of heads, likelihoods $P(x|H_0)$?, $P(x|H_1)$?

x	0	1	2	3	4	5	6	7	8	9	10
coin 0	.0010	.0098	.0439	.1172	.2051	.2461	.2051	.1172	.0439	.0098	.0010
coin 1	.0000	.0001	.0014	.0090	.0368	.1029	.2001	.2668	.2335	.1211	.0282

- Does that really help? In the end we would rather just like to know $P(H_0|x)$ and $P(H_1|x)$ after observing the number of heads ("posterior probabilities")

Testing hypotheses: Basic ideas by example

- Use Bayes' theorem for "inverting" probability (for generative model H_0)

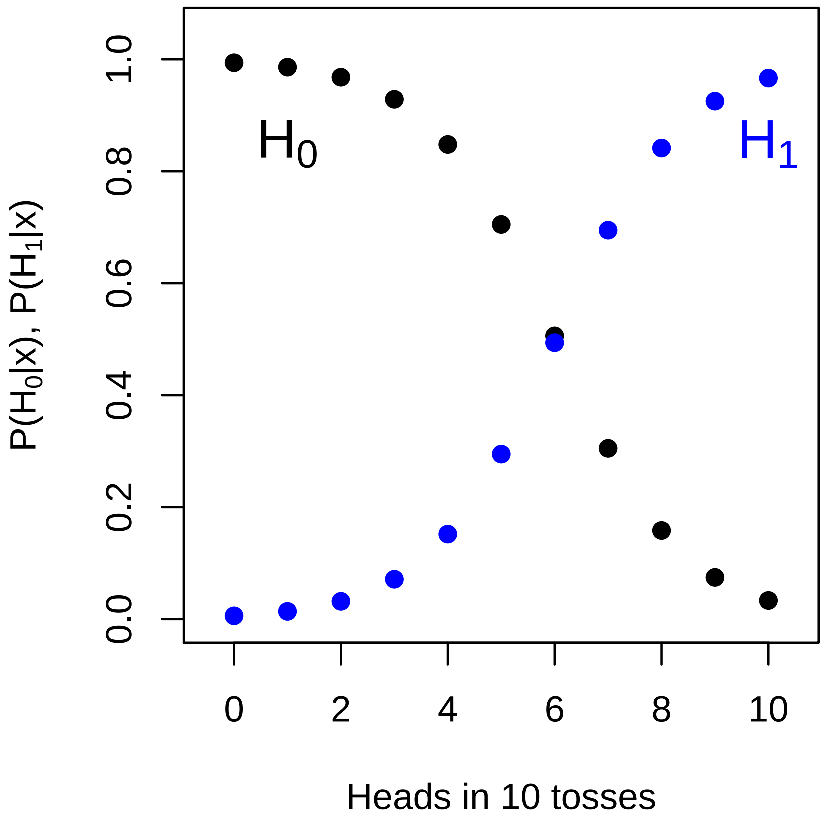
$$P(H_0|x) = \frac{P(x|H_0)\,P(H_0)}{P(x)}$$

  … and similarly for $P(H_1|x)$

- Ok, but what is $P(H_0)$ ("a priori probability")?
  - we may come up with a value assuming a certain behaviour for the person choosing which coin to toss, but let's assume they are fair and unbiased ...
    - ⋆ so let's just go for the simplest reasonable *assumption* $P(H_0) = P(H_1) = 0.5$ and call it our *'null hypothesis'*

- And what is $P(x)$? Probability of the data x?
  - Law of total probability: $P(x) = \sum_k P(x|H_k)P(H_k)$, so the evidence is the probability of observing the data, considering all possible hypotheses
    $\{H_k\}$ exhaustive and mutually exclusive
    $P(x) = P(x|H_0)P(H_0) + P(x|H_1)P(H_1)$

Testing hypotheses: Basic ideas by example



- Note: here $H_0(x) + H_1(x) = 1$
- This is a super-idealized scenario. We know all possible hypotheses, which is hardly ever the case!
  - Typically, H_0 ("null hypothesis") tested against alternative hypothesis H_1 but $H_1 \neq \neg H_0$, as it does not cover all possible alternative hypotheses to H_0
  - Since H_1 is not a complete negation of H_0, we cannot really employ the law of total probability for $P(x)$:(
- Next best thing: look at the (posterior) **odds ratio** and then the probability of the data $P(x)$ cancels out: $\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(x|H_0)}{P(x|H_1)} \frac{P(H_0)}{P(H_1)}$ = ratio of likelihoods × ratio of prior probabilities :)
  - The posterior **odds ratio** provides a direct comparison between the hypotheses without needing to compute the evidence $P(x)$
  - the ratio of likelihoods above is called the **Bayes factor**

Testing hypotheses: Basic ideas by example

The ratio of a priori probabilities cancels out (we set both to 0.5), so evaluating the likelihood ratio $\frac{P(x|H_0)}{P(x|H_1)}$ gives us:

x	0	1	2	3	4	5	6	7	8	9	10
$P(x\|H_0)$	.0010	.0098	.0439	.1172	.2051	.2461	.2051	.1172	.0439	.0098	.0010
$P(x\|H_1)$	.0000	.0001	.0014	.0090	.0368	.1029	.2001	.2668	.2335	.1211	.0282
$P(x)$	.0005	.0050	.0227	.0631	.1209	.1745	.2026	.1920	.1387	.0654	.0146
$P(H_0\|x)$	.9940	.9861	.9681	.9287	.8480	.7051	.5061	.3052	.1584	.0746	.0334
$P(H_1\|x)$	.0060	.0139	.0319	.0713	.1520	.2949	.4939	.6948	.8416	.9254	.9666
$\frac{P(x\|H_0)}{P(x\|H_1)}$	165.4	70.88	30.38	13.02	5.579	2.391	1.025	.4392	.1882	.0807	.0346

- To decide between H_0 and H_1, choose the hypothesis with larger posterior probability!

- Choose H_0 if . . .

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(x|H_0)}{P(x|H_1)}\frac{P(H_0)}{P(H_1)} > 1$$

  or equivalently for the (marginalized) likelihood ratio $\frac{P(x|H_0)}{P(x|H_1)} > c$, where the constant c depends on the a priori model probabilities

Testing hypotheses: Basic ideas by example

- In our toy problem, what are the consequences of using $c = 1$ as (a somewhat arbitrary) threshold?
  - H_0 (fair coin) is accepted as long as $X \leq 6$, and is rejected in favor of H_1 (unfair coin) for $X > 6$, where X is the number of heads tossed

- There are two possible ways to be wrong . . . Can you confirm the numbers?
  - reject H_0 when it is true (type I error, "false positive")

$$P(\text{reject}\, H_0|H_0) = P(X > 6|H_0) = \sum_{x=7}^{10} P(x|H_0) = 0.17$$

  - accept H_0 when it is false (type II error, "false negative")

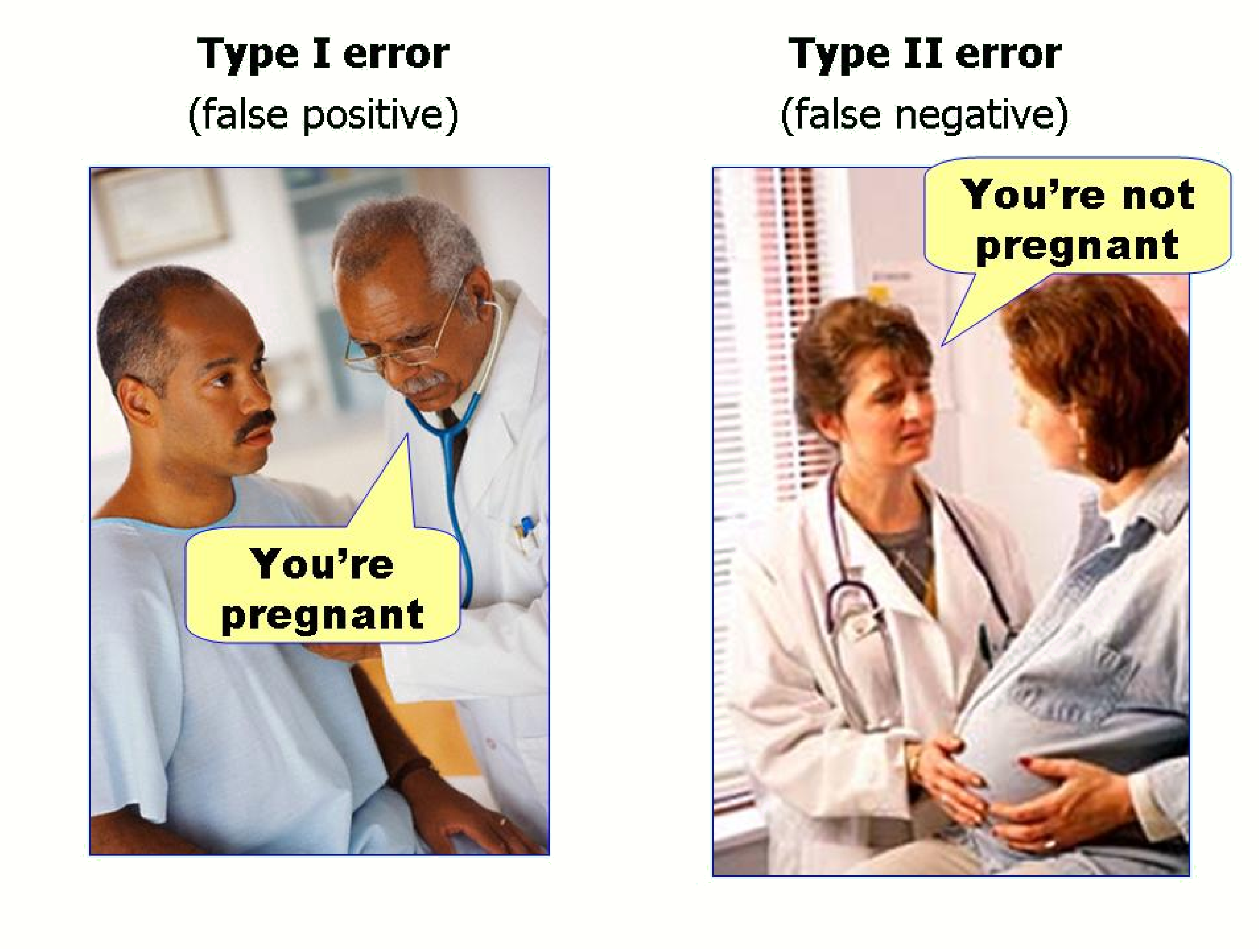
$$P(\text{accept}\, H_0|H_1) = P(X \leq 6|H_1) = \sum_{x=0}^{6} P(x|H_1) = 0.35$$

- Note that one error is not just the complement of the other

Interlude: Wikipedia on the naming of things . . .

- A type I error (false positive, error of the first kind) is the incorrect rejection of a true null hypothesis. Usually a type I error leads one to conclude that a supposed effect or relationship exists when in fact it doesn't. Examples of type I errors include a test that shows a patient to have a disease when in fact the patient does not have the disease, a fire alarm going on indicating a fire when in fact there is no fire, or an experiment indicating that a medical treatment should cure a disease when in fact it does not.

- A type II error (false negative, error of the second kind) is the failure to reject a false null hypothesis. Examples of type II errors would be a blood test failing to detect the disease it was designed to detect, for a patient who really has the disease; a fire breaking out and the fire alarm does not ring; or a clinical trial of a medical treatment failing to show that the treatment works when really it does.



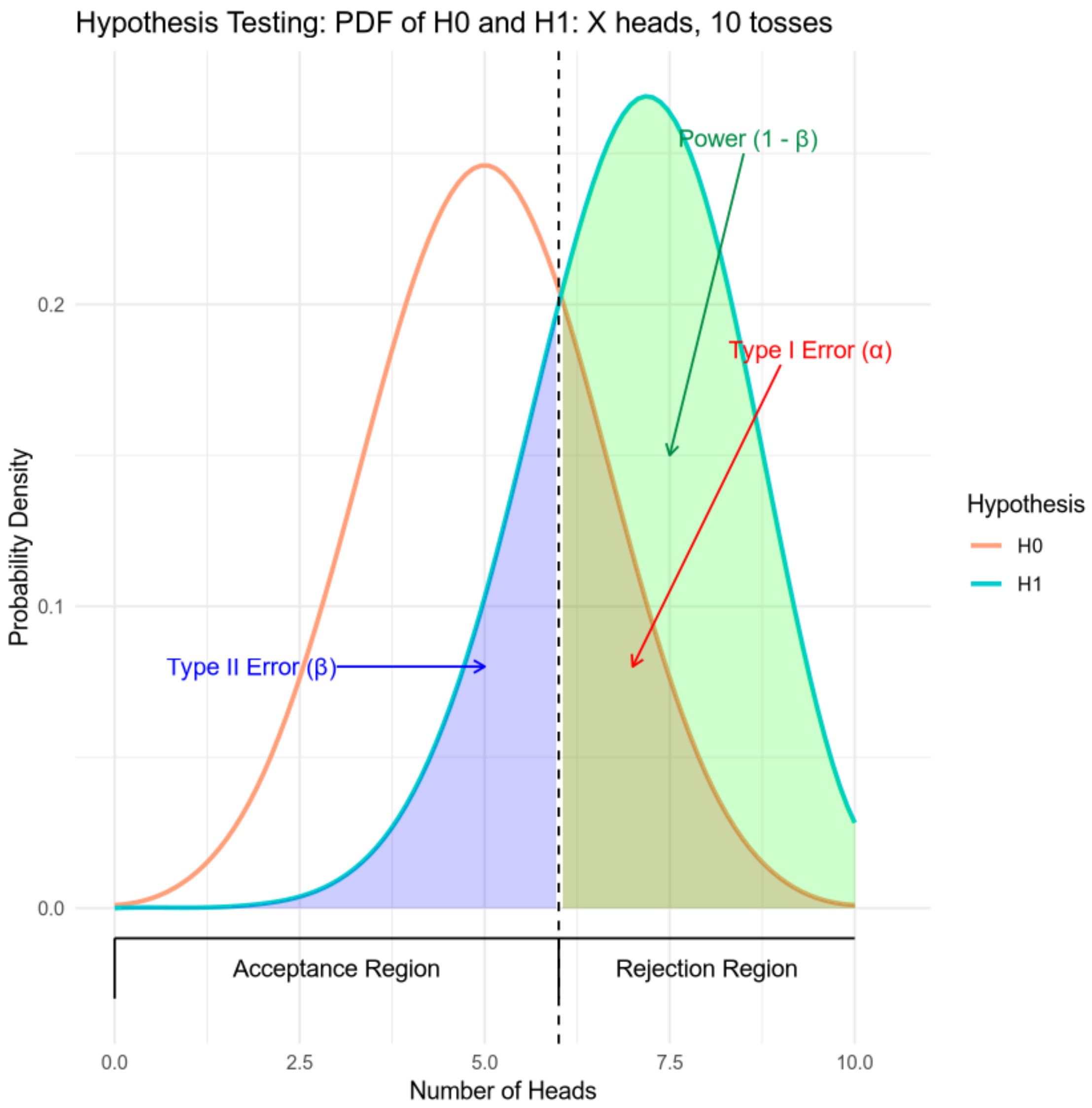
Type I error
(false positive)

Type II error
(false negative)

Terminology

- The significance level, α, is the probability of making a type I error
  - this is like a threshold you set for how willing you are to risk making a false positive error
- The power of a test is $1 - \beta$ where β is the probability of making a type II error
  - e.g. in a medical test, high power means the test is good in detecting disease
- A test statistic is the statistic that is used for deciding whether to accept (or reject) the null hypothesis H_0
  - in our previous example with the coins, this was the *odds ratio* (of the posteriors)
- rejection region is the set of parameter values of the test statistic for which the null hypothesis, H_0 is rejected
  - its complement is the acceptance region
- null distribution: the probability distribution of the test statistic assuming the null hypothesis is true

Terminology (for the coin example)



test statistic used:

$$\frac{P(H_0|x)}{PH_1|x} > 1$$

Different approaches to Hypothesis testing

- **Bayesian Hypothesis Testing:**
  - Compares hypotheses by calculating their posterior probabilities given the data
  - Requires knowledge of the full probability distribution for both the null (H_0) and alternative (H_1) hypotheses
  - Often used when prior information is available or when integrating evidence across different models

- **Neyman-Pearson (NP) Paradigm:**
  - Focuses on rejecting or not rejecting the null hypothesis (H_0) in favor of an alternative (H_1)
  - Works with $simple$ hypotheses, where the probability distribution is fully specified, but can be extended to handle $composite$ hypotheses
  - The NP approach is best suited to scenarios where we want to control the probability of making a type I error

- **Fisher's Approach to Hypothesis Testing (Frequentist):**
  - Emphasizes testing H_0 against an unspecified alternative hypothesis
  - Focuses on p-values as a measure of evidence against H_0
  - Only requires the distribution of the test statistic under H_0
  - Makes the problem asymmetric by focusing on the null hypothesis alone

Neyman-Pearson lemma

- In the Neyman-Pearson paradigm, the specification of priors is not necessary
  - Neyman-Pearson lemma:
    Suppose H_0 and H_1 are simple hypotheses. A likelihood ratio test rejects H_0 whenever the likelihood ratio is less than c at significance level α. Then, for any other test of H_0 with a significance level $\leq \alpha$, its power against H_1 is at most the power of this likelihood ratio test

- Many test statistics are possible...

- The likelihood ratio test is optimal, meaning it has the largest discriminating power
  - unfortunately, problems are rarely formulated as alternatives between simple hypotheses

- How does one select an appropriate test statistic for more complicated problems?
  - it depends on the specific context and the hypotheses being tested
  - different disciplines often rely on different types of tests

Choosing the null hypothesis

- Usually, H_0 is chosen to indicate that “nothing special is going on”
  - This means that H_0 represents a state of no effect or no difference, serving as a baseline for comparison

- The test statistic under the null hypothesis needs to be known
  - a test statistic is a standardized value (calculated from the sample data during the hypothesis test)
  - knowing its distribution under H_0 allows us to determine the significance of the observed data

- Additional considerations:
  - In science, simpler hypotheses are preferred because they are more easily testable (potentially falsifiable)
  - Occam’s razor: other things being equal, the simplest explanations should be preferred over more complex ones
  - Popper’s falsification principle: a hypothesis/theory must be falsifiable for it to hold any (scientific) value
  - Gravity of errors! Consider the consequences of Type I and Type II errors

The p-value

- The p-value is the probability, given H_0 is true, of obtaining data (or a value of the statistic) at least as extreme as the one observed, or more extreme
  - In other words, it quantifies how surprising or extreme the observed data is under the null hypothesis

- p-value often used to characterize the significance level of a test
  - The smaller the p-value, the stronger the evidence against the null hypothesis

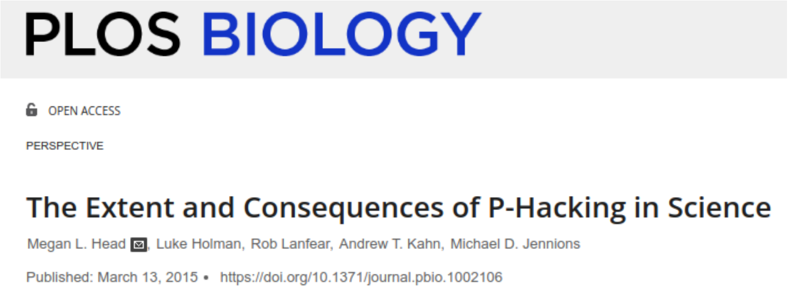
- It is important to note that ...:
  - the p-value is not the probability that H_0 is true
  - the p-value is not the probability of observing the data that we have
    - ⋆ both of these are common sources of misinterpretation in hypothesis testing!

Consider the coin tossing example: if 8 heads are observed, what is the p-value?

The trials and tribulations of the p-value

- p-values often used for selection:
  - common thresholds for p-values are 0.05 (5%), 0.01 (1%), and 0.001 (0.1%)
  - a p-value less than 0.05 typically indicates strong evidence against the null hypothesis, leading to its rejection *

- * Beware of p-hacking!:
  - (also known as data dredging, data fishing, data snooping or data butchery)
  - p-hacking refers to manipulating data or analyses until non-significant results appear significant
  - it involves employing loosely defined or multiple hypotheses or data manipulation until a p-value less than 0.05 is obtained
  - it undermines the integrity of scientific research and remains a problem in today's 'publish or perish' academic culture



PLOS BIOLOGY

OPEN ACCESS

PERSPECTIVE

The Extent and Consequences of P-Hacking in Science

Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, Michael D. Jennions

Published: March 13, 2015 • https://doi.org/10.1371/journal.pbio.1002106

Bayesian hypothesis testing – model selection

- In the Bayesian context, hypotheses/models are not rejected or accepted but rather alternative models are compared

- For both alternatives, one must know the generative models (and their PDFs)

- We have already encountered the **odds ratio**

$$\frac{P(M_1|x)}{P(M_2|x)} = \frac{P(x|M_1)}{P(x|M_2)} \frac{P(M_1)}{P(M_2)}$$

- Typically, models depend on parameters θ. Using the law of total probability, we can write

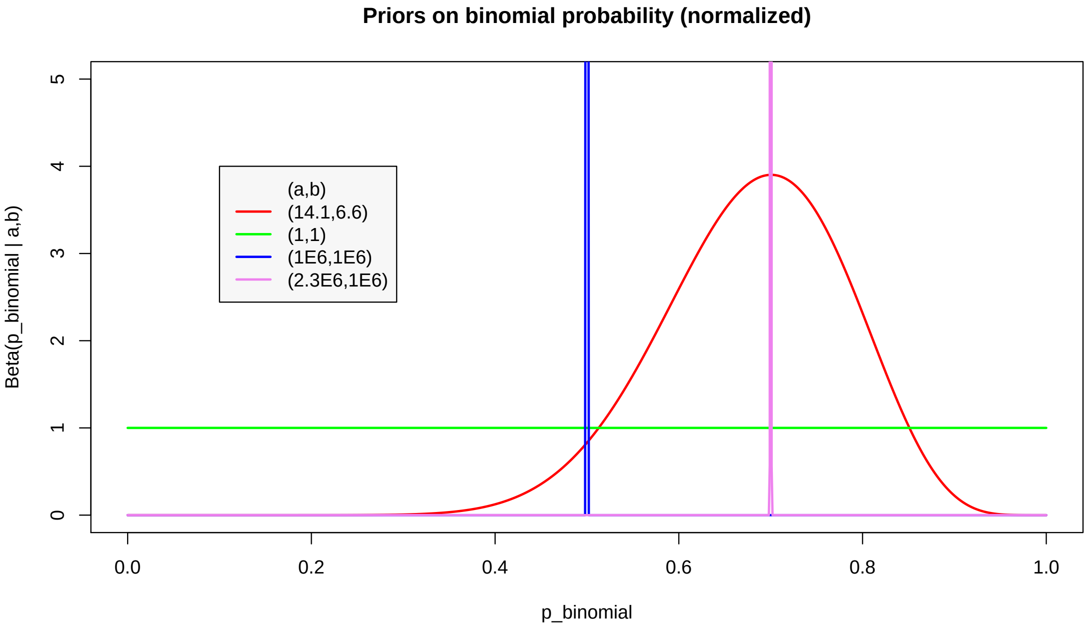
$$\frac{P(M_1|x)}{P(M_2|x)} = \frac{\int P(x|\theta_1, M_1)\, P(\theta_1|M_1)\, d\theta_1}{\int P(x|\theta_2, M_2)\, P(\theta_2|M_2)\, d\theta_2} \frac{P(M_1)}{P(M_2)}$$

Bayesian model selection – what is what?

$$\frac{P(M_1|x)}{P(M_2|x)} = \frac{\int P(x|\theta_1, M_1)\, P(\theta_1|M_1)\, d\theta_1}{\int P(x|\theta_2, M_2)\, P(\theta_2|M_2)\, d\theta_2}\, \frac{P(M_1)}{P(M_2)}$$

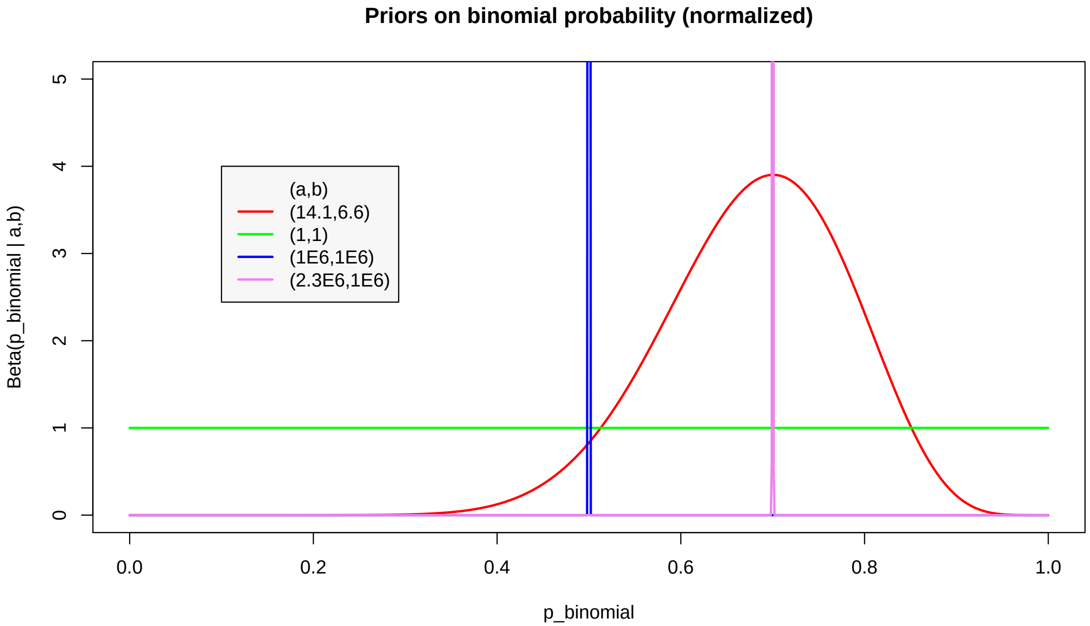
- $P(x|\theta, M)$ is the likelihood: the probability of the data given the parameters and model
- $P(\theta|M)$ is the prior of the parameters: the probability distribution of the parameters for model M, before observing the data
- $P(M)$ is the prior of the model: our initial belief in the model, before observing the data
- If there is no reason to prefer one model over the other, then assign equal weight to their priors, $P(M_1) = P(M_2)$, and the last ratio above (in blue) cancels out. In that case, the **odds ratio** $=$ **Bayes factor** (ratio of marginal likelihoods)
- $P(x|M) = \int P(x|\theta, M) P(\theta|M) d\theta$ is the so-called **marginal likelihood** or evidence of the observed data, given the model
- Ratios > 10 are considered strong evidence for preferring one model/proposition over the other (but the exact threshold depends on the circumstances)

Comparing models when tossing coins – again



- Now let's toss *four* coins of unknown fairness level, assuming different priors
- Again, 10 tosses are performed, and the total number of heads x is recorded
- coin 0 is a fair coin with $p_{\text{bin}} = 0.5$. Here, p_{bin} is the probability of the underlying binomial distribution describing the tossing
  - corresponds to prior probability density $\mathrm{P}(p_{\text{bin}}) = \delta(p_{\text{bin}} - 0.5)$, where δ stands for a δ function
  - can be approximated by a Beta-distribution → see plot above

Comparing models when tossing coins – again



- Coins are assigned priors with the following normalized continuous probability densities $\mathrm{P}(p_{\mathrm{bin}})$
  - coin 0: $\mathrm{P}(p_{\mathrm{bin}}) = \delta(p_{\mathrm{bin}} - 0.5)$ $\rightarrow$ precisely known parameter
  - coin 1: $\mathrm{P}(p_{\mathrm{bin}}) = \delta(p_{\mathrm{bin}} - 0.7)$
  - coin 2: $\mathrm{P}(p_{\mathrm{bin}}) = \mathrm{Beta}(p_{\mathrm{bin}} | \alpha = 14.1, \beta = 6.6)$ $\rightarrow$ falls around $p_{\mathrm{bin}} \approx 0.7$
  - coin 3: $\mathrm{P}(p_{\mathrm{bin}}) = 1$ $\rightarrow$ no prior knowledge at all
- All coins are tossed equally likely
- Always comparing pairs of coins: which one was tossed? What is the evidence?

Comparing models when tossing coins – again

number of heads x	0	1	2	3	4	5	6	7	8	9	10
evidence coin 0	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001
evidence coin 1	.000	.000	.001	.009	.037	.103	.200	.267	.233	.121	.028
log(Bayes 1/0)	-2.218	-1.851	-1.483	-1.115	-.747	-.379	-.011	.357	.725	1.093	1.461
evidence coin 2	.000	.002	.009	.027	.065	.121	.182	.218	.201	.130	.045
log(Bayes 2/0)	-.677	-.721	-.706	-.633	-.500	-.307	-.051	.270	.659	1.123	1.667
evidence coin 3	.091	.091	.091	.091	.091	.091	.091	.091	.091	.091	.091
log(Bayes 3/0)	1.971	.969	.316	-.110	-.353	-.432	-.353	-.110	.316	.969	1.971

- Taking the assumed fair coin 0 as baseline one obtains
  - the unfair coin 1 is the preferred model when observing a large number of heads
  - similarly, the more “flexible” coin 2
  - the “ultra flexible” coin 3 is preferred only in extreme cases

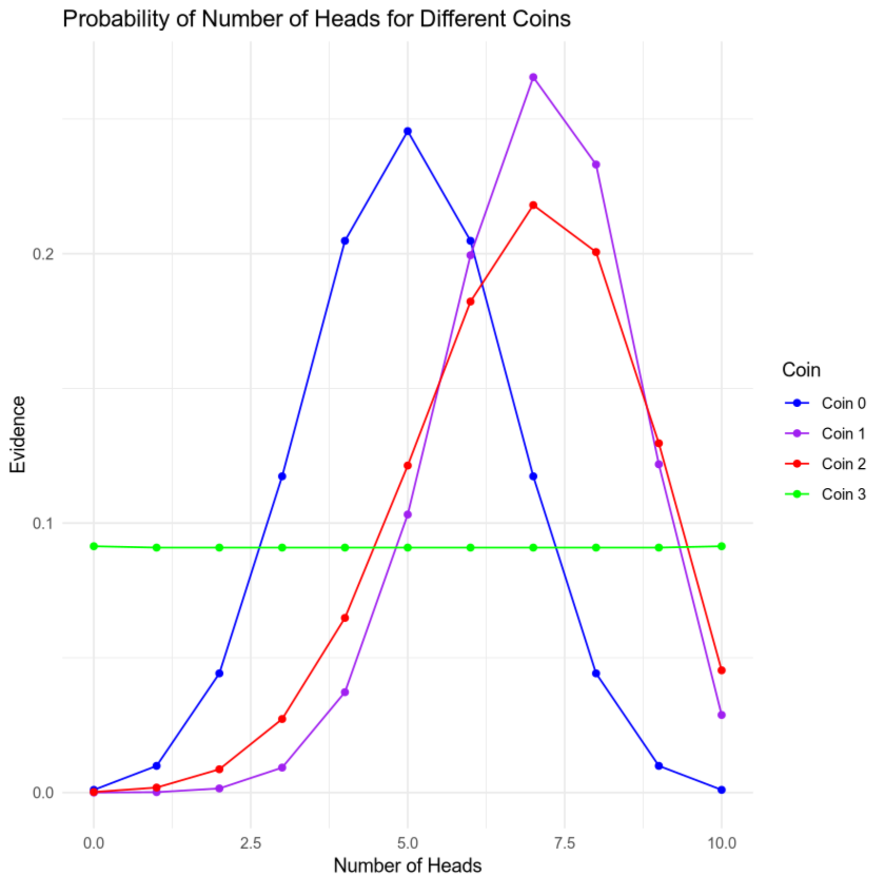
- Remarkably, coin 3 is able to represent the data best as shown by the likelihoods below:

number of heads x	0	1	2	3	4	5	6	7	8	9	10
$P(x\|H_0)$	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001
$P(x\|H_3, p = x/10)$	1.000	.387	.302	.267	.251	.246	.251	.267	.302	.387	1.000

  - this is a consequence of the fact that p can be adjusted to match the observations
  - the presence of the free parameter p is penalized when calculating the evidence
  - preference to simpler model reasonably representing the data → Occam’s razor

Comparing models when tossing coins – again

number of heads x	0	1	2	3	4	5	6	7	8	9	10
evidence coin 0	.001	.010	.044	.117	.205	.246	.205	.117	.044	.010	.001
evidence coin 1	.000	.000	.001	.009	.037	.103	.200	.267	.233	.121	.028
log(Bayes 1/0)	-2.218	-1.851	-1.483	-1.115	-.747	-.379	-.011	.357	.725	1.093	1.461
evidence coin 2	.000	.002	.009	.027	.065	.121	.182	.218	.201	.130	.045
log(Bayes 2/0)	-.677	-.721	-.706	-.633	-.500	-.307	-.051	.270	.659	1.123	1.667
evidence coin 3	.091	.091	.091	.091	.091	.091	.091	.091	.091	.091	.091
log(Bayes 3/0)	1.971	.969	.316	-.110	-.353	-.432	-.353	-.110	.316	.969	1.971


Probability of Number of Heads for Different Coins
Evidence
Number of Heads
Coin
Coin 0
Coin 1
Coin 2
Coin 3