HIGH DIMENSIONAL STATISTICS

BAYESIAN INFERENCE

Msc programme in Mathematics and Statistics

University of Helsinki

Matti Pirinen

LARGE SCALE INFERENCE – WHAT IS THE QUESTION?

- *P*-value: Probability of at least as extreme data sets under the null, $\Pr(D_{\geq} \mid H_0)$
  - We can control false positive rate, the probability that a null variable is labelled as significant, but we have no direct control over what happens to non-null variables
- *Q*-value: Proportion of false discoveries among at least as extreme data sets
  - We can control false discovery rate, the proportion of null variables among the significant ones, but we do not have a direct control on the probability that a particular variable is null
- How could we compute $\Pr(H_0 \mid \text{Data})$, the probability that the null hypothesis holds for one particular variable based on the observed data?
  - This is done with Bayesian statistics
  - It requires specifying
    - Alternative possible hypotheses ("models")
    - Prior probabilities of these models

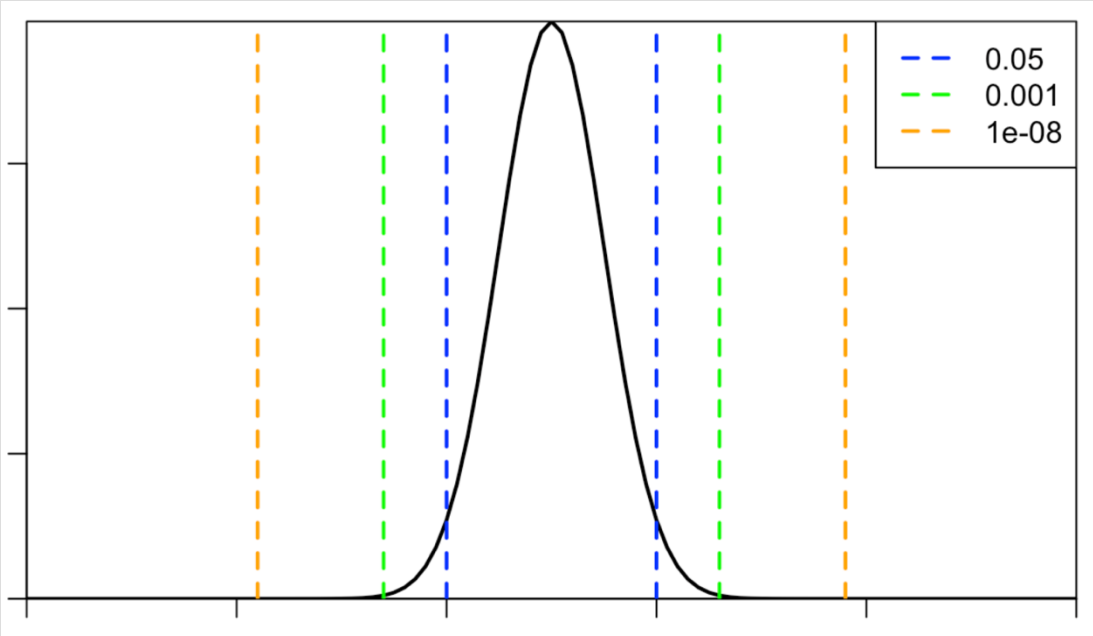
MULTIPLE TESTING EXAMPLE

- Plan: get data from UK Biobank on n = 500,000 individuals and their $p = 10^6$ genetic variants to test the association with diabetes status.
- Suppose only $q = 10^3$ variables are downloaded before the connection drops
- Do tests for q variables, the smallest P-value is 10^{-5} for variable v.
- Should we call v "statistically significant"?

MULTIPLE TESTING EXAMPLE

- Plan: get data from UK Biobank on n = 500,000 individuals and their $p = 10^6$ genetic variants to test the association with diabetes status.
- Only $q=10^3$ variables are downloaded before the connection drops
- Do tests for q variables, smallest P-value is 10^{-5} for variable v.
- Should we call v "statistically significant"?
- If we use FWER, then the significance threshold would have been $0.05/p$ had we got all the data, but since we only did q tests, could it now be $0.05/q$, which is 10^3 times larger than $0.05/p$.
- Still, in both cases the data that we had about the variable v is exactly same so is it reasonable that the "significance" label would depend on whether we managed download the rest of the data?

Null distribution of test statistic



Which threshold should we use?
The one defined by the original plan of testing p variables or the one defined by the actual number of tests (q) we have done?

BAYES RULE

$$\Pr(B)\Pr(A \mid B) = \Pr(A, B) = \Pr(A)\Pr(B \mid A) \implies \Pr(A \mid B) = \frac{\Pr(A)\Pr(B \mid A)}{\Pr(B)}.$$

- *A* and *B* are binary "events"
- Before knowing whether *B* happened, we consider that probability of *A* is $\Pr(A)$, **the prior probability** of *A*
- After knowing that *B* has happened, we consider that probability of *A* is now $\Pr(A \mid B)$, **the posterior probability** of *A* given *B*
- Bayes rule tells how we learn from observation:
  Prior probability is multiplied by factor $\frac{\Pr(B \mid A)}{\Pr(B)}$ to turn it to the posterior probability.

EXAMPLE: DIAGNOSTIC TEST

Example 4.2. Suppose we have a medical test for a disease D which has sensitivity of $\alpha = 0.99$, that is, gives (correctly) a positive result for 99% of true cases, and specificity of $\beta = 0.99$, that is, gives (correctly) a negative result for 99% of healthy individuals. What is the probability that an individual who in a population screening tests positive $(+)$ truly has the disease, when the prevalence K of the disease is (a) 10% or (b) 0.1%?

Answer 4.2.

- $\Pr(+ \mid D) = \alpha$ (sensitivity is α)
- $\Pr(+ \mid \text{no}D) = 1 - \beta$ (specificity is β)
- $\Pr(D) = K$ (prevalence is K, this is our prior probability of having D *before* we see the test result)
- $\Pr(+) = \Pr(+ \mid D)\Pr(D) + \Pr(+ \mid \text{no}D)\Pr(\text{no}D) = \alpha K + (1 - \beta)(1 - K)$

Bayes rule says that the posterior probability of D, also called **positive predictive value** (PPV), is

$$\Pr(D \mid +) = \Pr(D)\frac{\Pr(+ \mid D)}{\Pr(+)} = \frac{K\,\alpha}{K\,\alpha + (1 - K)\,(1 - \beta)}.$$

```
a = 0.99
b = 0.99
K = c(0.1, 0.001)
data.frame(preval = K, PPV = a * K / (a * K + (1 - b) * (1 - K)))
```

```
##   preval        PPV
## 1  0.100 0.91666667
## 2  0.001 0.09016393
```

TESTING SETUP

- $S = \{P \leq \alpha\}$, is the event that test gives a significant result (*P*-value $\leq \alpha$)
- *T* is the event that the variable is truly non-null
- *N* is the event that the variable is truly null
- $\Pr(T) = 1 - \Pr(N)$
- Bayes rule: $\Pr(T \mid S) = \frac{\Pr(T)\Pr(S|T)}{\Pr(S)}$ and $\Pr(N \mid S) = \frac{\Pr(N)\Pr(S|N)}{\Pr(S)}$
- $\frac{\Pr(T \mid S)}{\Pr(N \mid S)} = \frac{\Pr(T)\Pr(S|T)}{\Pr(N)\Pr(S|N)} = \text{prior}-\text{odds} \times \frac{\text{power}}{\alpha}$
- Interpretation: When we get a significant result, odds that it is a true discovery rather than a false discovery is proportional to prior-odds of a true discovery and statistical power of the test, and inversely proportional to the significance threshold α
- If we want that the inference procedure produces largely only true discoveries we can
  - Increase prior-odds by pre-selecting variables that are more likely non-null
  - Increase statistical power by increasing sample size or effect size (if possible)
  - Decrease significance level

MULTIPLE TESTING EXAMPLE CONTINUED

- We assume that there are only a few real diabetes variants among 10M genetic variants and set the prior probability of non-zero effect to $\Pr(T) \approx 10^{-6}$
- For a well-powered test (power $\approx$ 1), if we want that the significant variables are true positives with 95% probability (posterior-odds = 0.95/0.05 = 19), we should choose as our significance threshold

$$\alpha = \frac{\text{prior}-\text{odds}}{\text{posterior}-\text{odds}} = \frac{10^{-6}}{19} \approx 5 \cdot 10^{-8}$$

- Importantly, this threshold is independent of how many test we are doing in any one experiment and hence solves the conceptual problem of thresholds that depend on the number of tests
- If we studied only variables that have higher prior probability than 10^{-6}, then we could increase the significance threshold, but that would not be because we do fewer tests but rather because we now have different prior-odds for our tests.

BAYESIAN INFERENCE FOR A PARAMETER

- θ is probability that coin lands heads up, considered as an unknown parameter
- Before tossing coin, we assume ***prior*** distribution $\theta \sim \text{Uniform}(0,1)$
- We toss the coin *i* times and get *H(i)* heads
- The sampling model of the experiment is $H(i) | \theta \sim \text{Bin}(i, \theta)$
- Bayes rule gives that posterior distribution is $\theta \mid H(i) \sim \text{Beta}(H(i) + 1, i - H(i) + 1)$

This is computed by plugging into

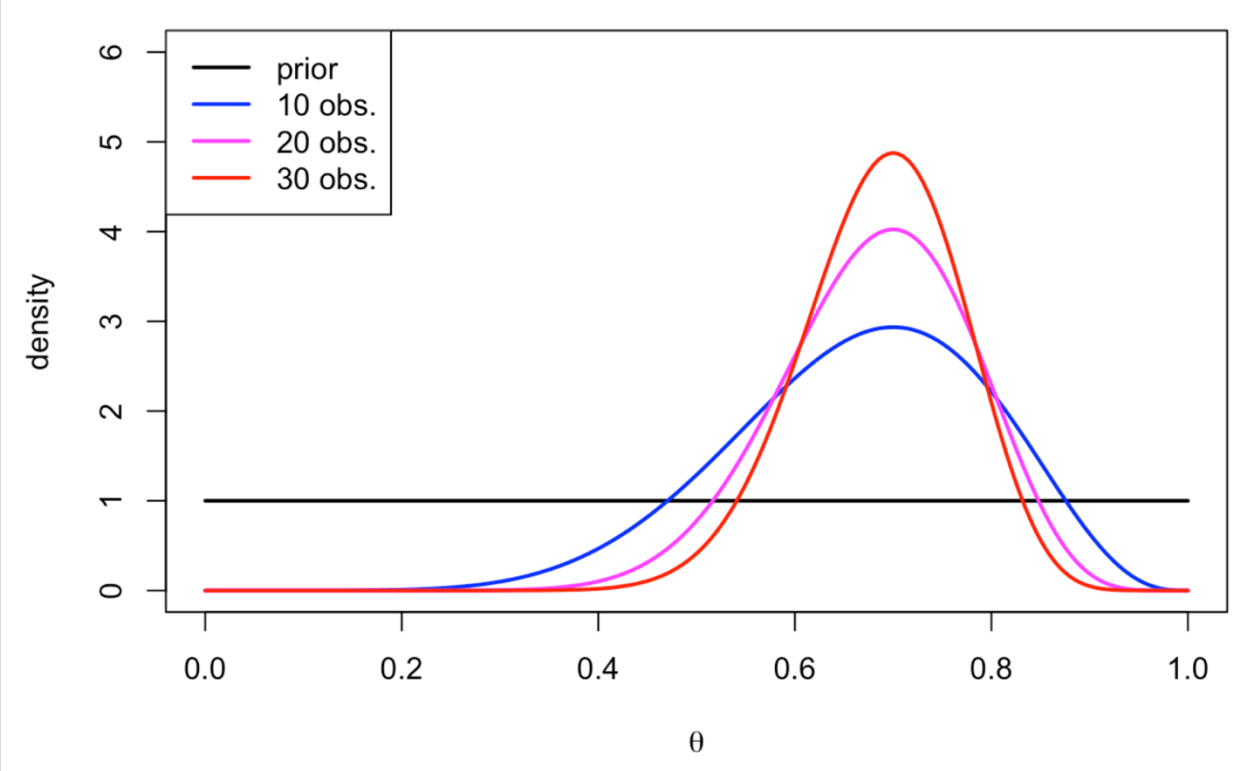
Bayes formula $p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$

the prior distribution $p(\theta) = 1$ and

the likelihood function $p(D|\theta) = \binom{i}{H(i)} \theta^{H(i)}(1-\theta)^{i-H(i)}$

BAYESIAN INFERENCE FOR A PARAMETER

- Before tossing coin, we assume ***prior*** distribution $\theta \sim \text{Uniform}(0,1)$
- We toss the coin *i* times and get *H(i)* heads
- Posterior distribution is $\theta \mid H(i) \sim \text{Beta}(H(i)+1, i-H(i)+1)$



Prior is flat, not favoring any region over others.

As we get more observations, the posterior gets more concentrated reflecting that we are learning from data.

Here
H(10) = 7
H(20) = 14
H(30) = 21
all leading to point estimate of $\hat{\theta} = 0.7$ but with different precision.

FROM SIGNIFICANCE TO OBSERVED DATA

- *D* is the observed data (and not anuymore just event whether *P*-value is below a threshold)
- H_0 is null hypothesis ("null model")
- H_1 is alternative hypothesis ("alternative model")
- Finally, we can talk about probability that the null hypothesis is true for this data set:

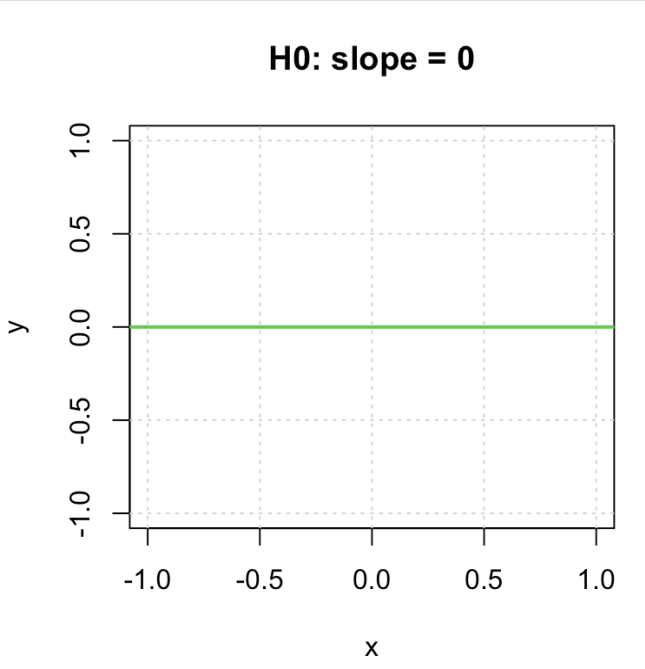
$$\Pr(H_0 \mid \mathcal{D}) = \frac{\Pr(H_0)\Pr(\mathcal{D} \mid H_0)}{\Pr(\mathcal{D})} = \frac{\Pr(H_0)\Pr(\mathcal{D} \mid H_0)}{\Pr(\mathcal{D} \mid H_0)\Pr(H_0) + \Pr(\mathcal{D} \mid H_1)\Pr(H_1)}$$

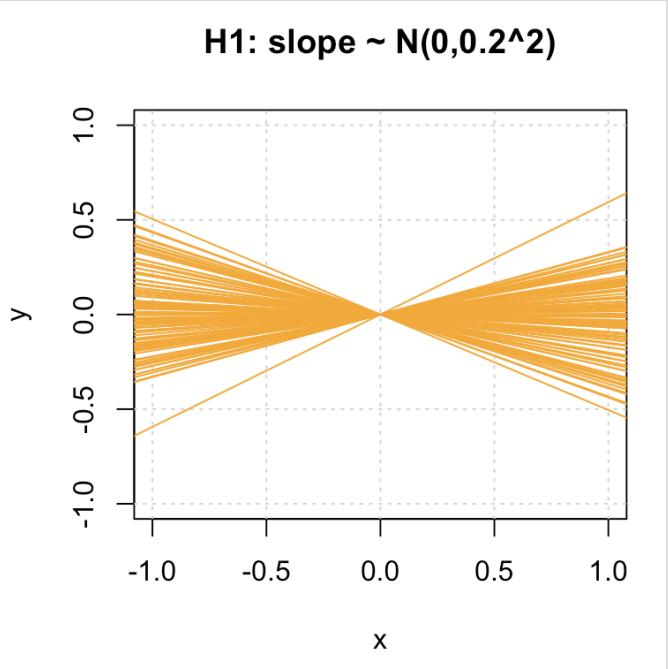
- Term $\Pr(D \mid H)$ is **(marginal) likelihood** of the hypothesis *H* and is the probability mass / density value of the observed data under the hypothesis *H*
- When hypothesis *H* sets a prior distribution $\Pr(\theta \mid H)$ for parameter θ then

$$\Pr(D \mid H) = \int \Pr(D \mid \theta) \cdot \Pr(\theta \mid H)\, \mathrm{d}\theta$$

EXAMPLE: MODEL COMPARISON FOR SLOPE

- Null model H_0 says that the slope in linear regression is 0
- Alternative model H_1 says that the slope has prior distribution $\beta \sim N(0, \tau_1^2)$





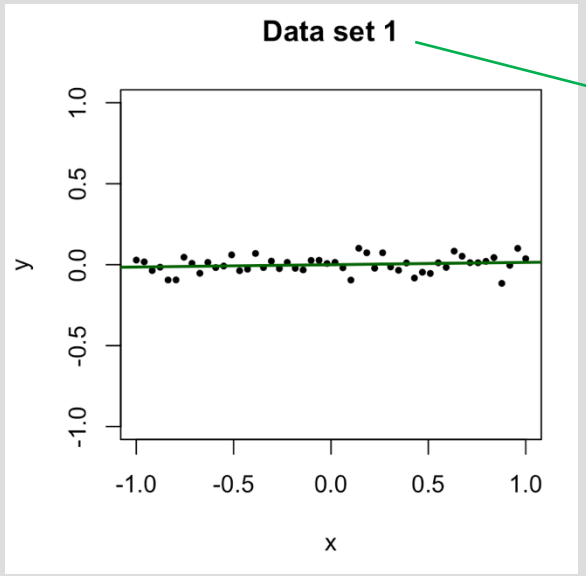
100 simulations from hypotheses H_0 and H_1 ($\tau_1^2 = 0.04$)

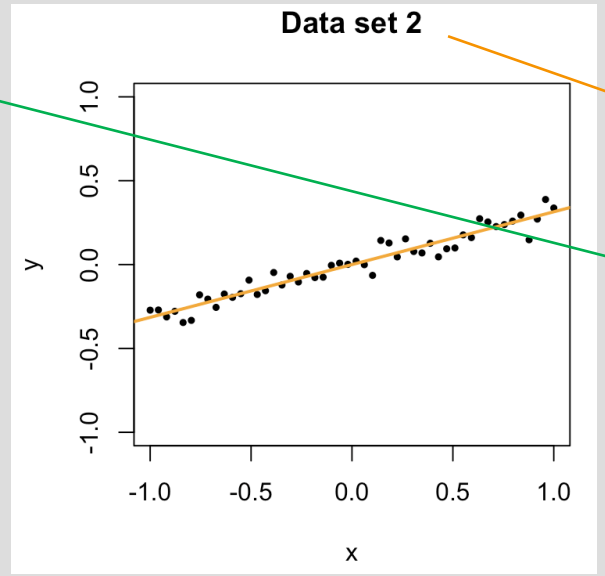
EXAMPLE: MODEL COMPARISON FOR SLOPE

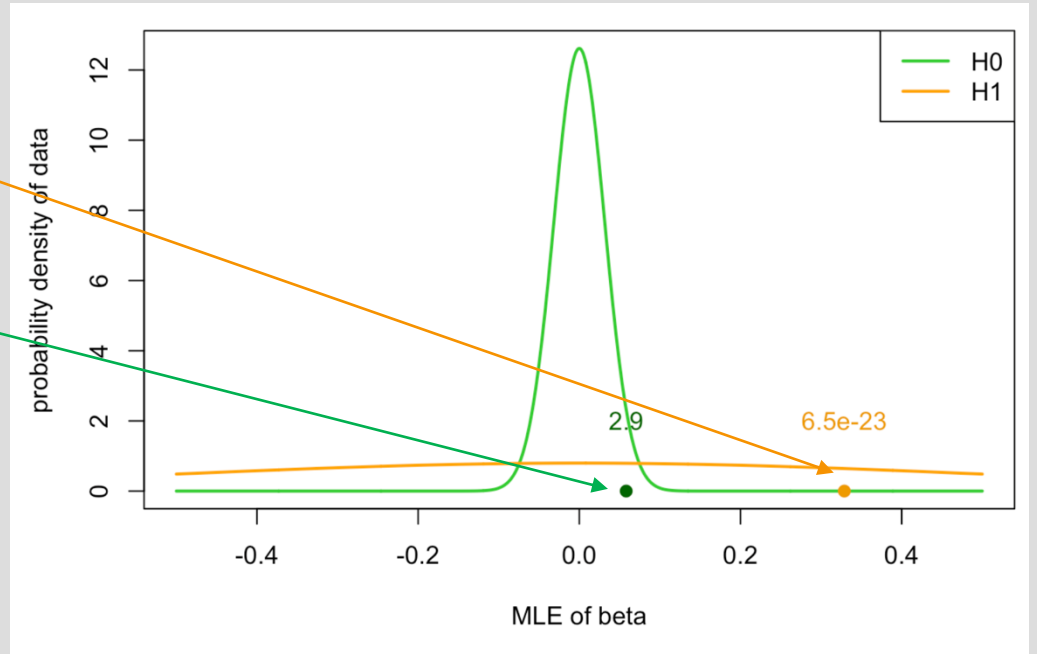
- Null model H_0 says that the slope in linear regression is 0
- Alternative model H_1 says that the slope has prior distribution $\beta \sim N(0, \tau_1^2)$
- We can compute the marginal likelihoods
  - SE is standard error of slope estimate $\hat{\beta}$
  - c is a constant

$$\Pr(\mathcal{D} \mid H_0) = c \cdot \mathcal{N}\left(\widehat{\beta};\, 0,\, \mathrm{SE}^2\right)$$

$$\Pr(\mathcal{D} \mid H_1) = c \cdot \mathcal{N}\left(\widehat{\beta};\, 0,\, \tau_1^2 + \mathrm{SE}^2\right)$$





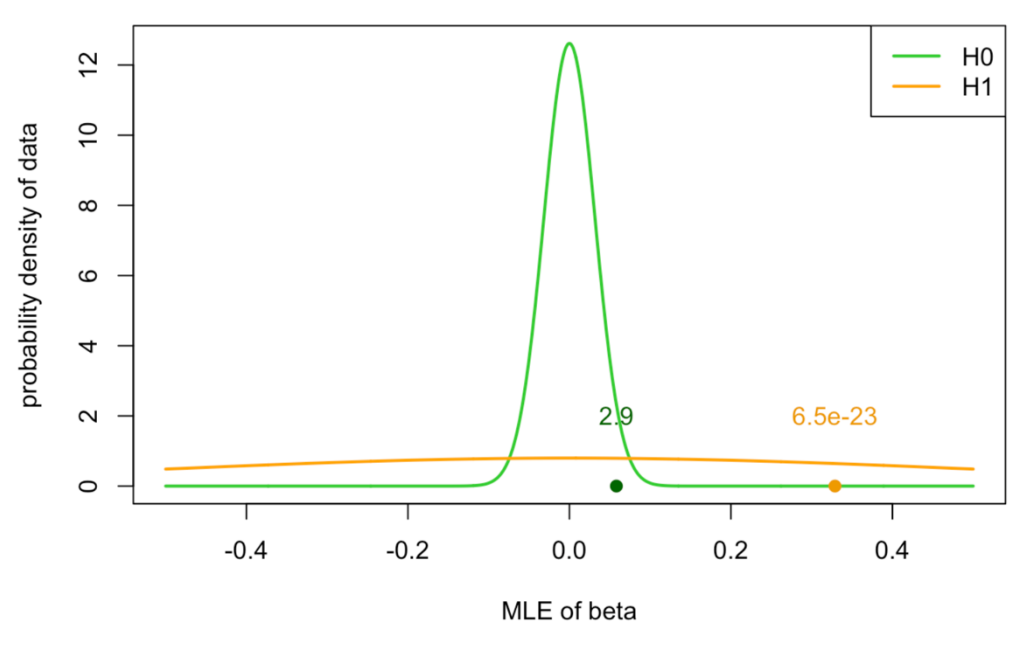


EXAMPLE: MODEL COMPARISON FOR SLOPE

- Null model H_0 says that the slope in linear regression is 0
- Alternative model H_1 says that slope has prior distribution $\beta \sim N(0, \tau_1^2)$
- We can compute the marginal likelihoods
  - SE is standard error of slope estimate $\hat{\beta}$
  - c is a constant

$$\Pr(\mathcal{D} \mid H_0) = c \cdot \mathcal{N}\left(\widehat{\beta};\, 0,\, \mathrm{SE}^2\right)$$

$$\Pr(\mathcal{D} \mid H_1) = c \cdot \mathcal{N}\left(\widehat{\beta};\, 0,\, \tau_1^2 + \mathrm{SE}^2\right)$$

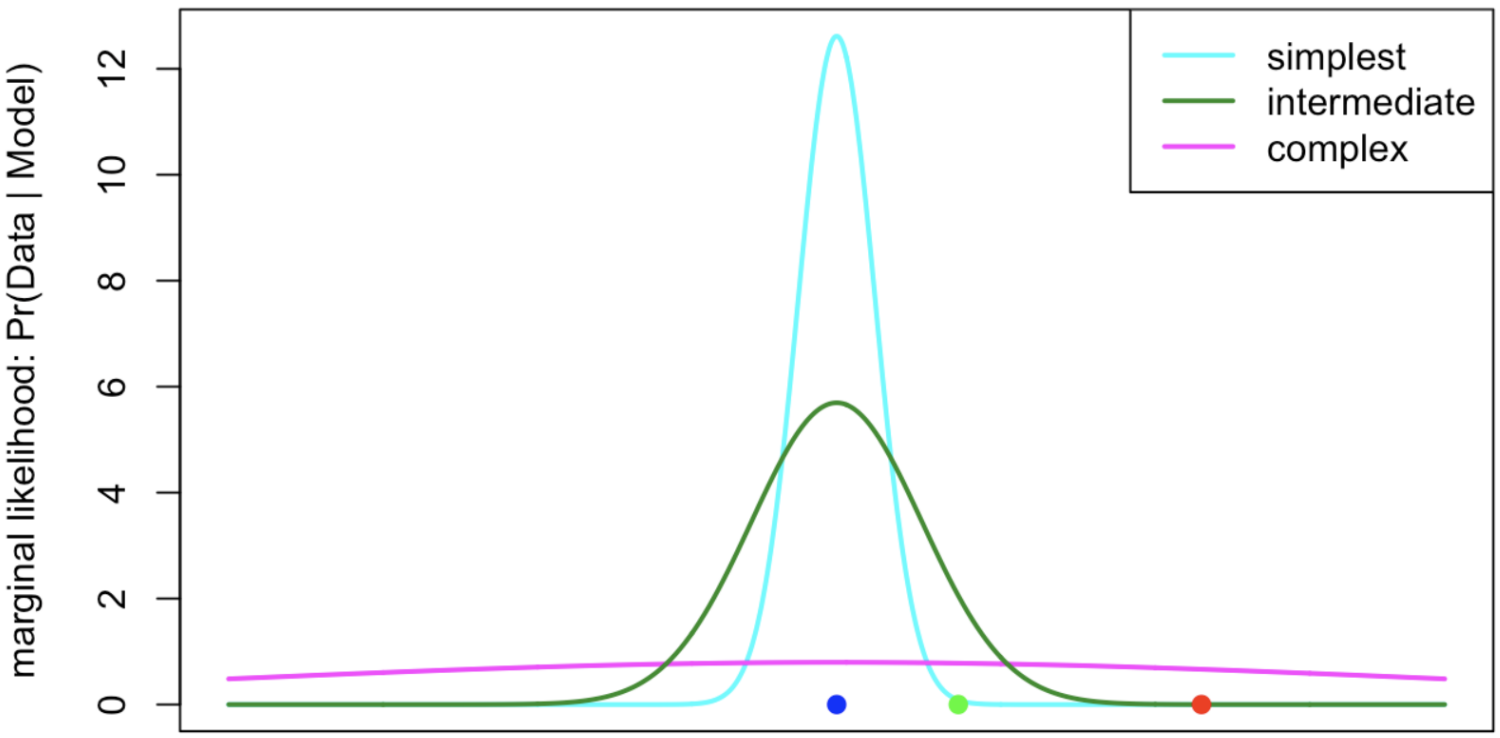


To compare the models, we multiply prior-odds with likelihood ratio ("Bayes factor", BF) to get posterior-odds

$$\frac{\Pr(H_0 \mid \mathcal{D})}{\Pr(H_1 \mid \mathcal{D})} = \frac{\Pr(\mathcal{D} \mid H_0)}{\Pr(\mathcal{D} \mid H_1)} \frac{\Pr(H_0)}{\Pr(H_1)}$$

BF for two data sets are shown on the Figure.

MARGINAL LIKELIHOOD



Marginal likelihood for 3 models are shown.

Since likelihood as function of data is prob. density, it integrates to 1.

Simple model spreads the mass to a small collection of possible data sets and for those sets the density is relatively high (cyan).

Complex model can model many types of data sets and hence the relative density for any particular data sets tends to be small (magenta).

$$\frac{\Pr(\mathcal{D} \mid H_0)}{\Pr(\mathcal{D} \mid H_1)}$$

If simple model can explain observed data well, Bayes factor will favor simple model over a more complex model. Thus, Bayesian model comparison protects from overfitting.

EXAMPLE: COMPARING COINS

- Let's consider two series of coin tosses of length n, where the number of observed heads are h_1 and h_2. We want to infer whether the two series are conducted with a similar coin.
- H_0: $\theta_1 = \theta_2 \sim \text{Uniform}(0,1)$ (i.e., probability of heads is the same)
- H_1: $\theta_1 \sim \text{Uniform}(0,1)$ and $\theta_2 \sim \text{Uniform}(0,1)$ and $\theta_1 \perp \theta_2$
- Sampling model is

$$\Pr(h_1, h_2 \mid \theta_1, \theta_2) = \prod_{i=1}^{2} \binom{n}{h_i} \theta_i^{h_i} (1-\theta_i)^{n-h_i}$$

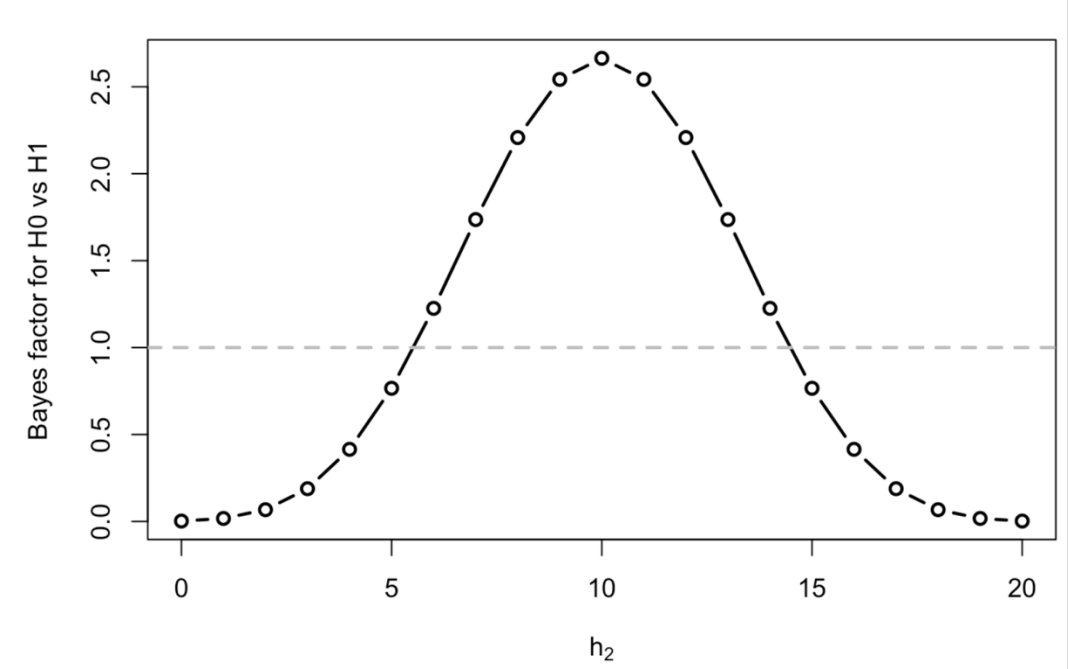
- Marginal likelihoods can be presented using Beta-functions (B) and their ratios is the Bayes factor comparing H_0 against H_1

$$\text{BF}_{01} = \frac{\Pr(h_1, h_2 \mid H_0)}{\Pr(h_1, h_2 \mid H_1)} = \frac{B(h_1 + h_2 + 1, 2n - h_1 - h_2 + 1)}{B(h_1 + 1, n - h_1 + 1) B(h_2 + 1, n - h_2 + 1)}$$

EXAMPLE: COMPARING COINS

- Let's consider two series of coin tosses of length n, where the number of observed heads are h_1 and h_2. We want to infer whether the two series are conducted with a similar coin.
- H_0: $\theta_1 = \theta_2 \sim \text{Uniform}(0,1)$ (i.e., probability of heads is the same)
- H_1: $\theta_1 \sim \text{Uniform}(0,1)$ and $\theta_2 \sim \text{Uniform}(0,1)$ and $\theta_1 \perp \theta_2$

NOTE:
Bayesian model comparison
has a built-in control against overfitting as it favors a simpler model, when the simpler model is "good enough".



Suppose $h_1=10$.
BF favors H_0 when $5 < h_1 < 15$

In addition to BF, the prior probabilities of the hypotheses will also affect the posteriors.

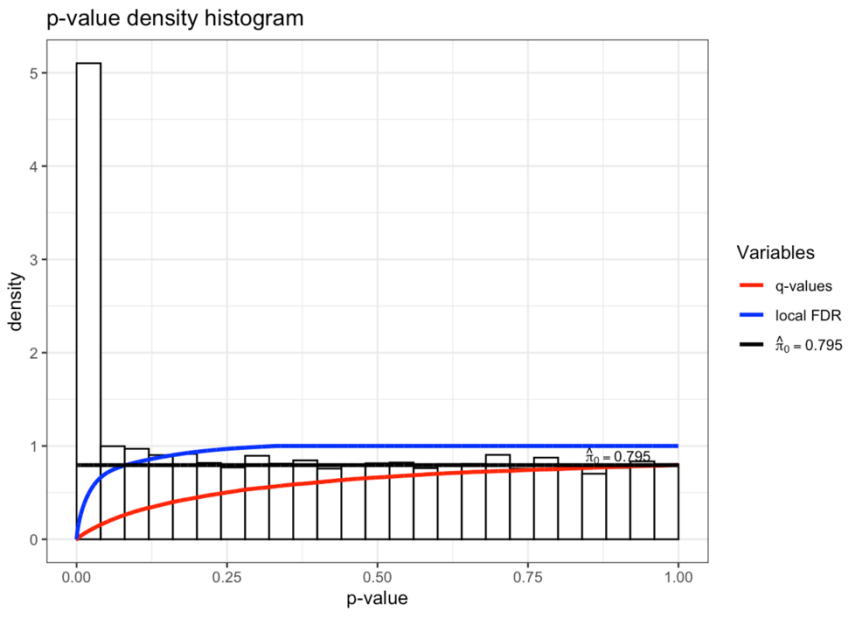
$$\frac{\Pr(H_0 \mid \mathcal{D})}{\Pr(H_1 \mid \mathcal{D})} = \frac{\Pr(\mathcal{D} \mid H_0)}{\Pr(\mathcal{D} \mid H_1)} \frac{\Pr(H_0)}{\Pr(H_1)}$$

LOCAL FALSE DISCOVERY RATE

- LFDR applies Bayes rule locally to the observed *P*-value distribution
  - Assume we have estimated $\hat{\pi}_0$ (as in HDS3) and also empirical density function of *P*-values $\hat{f}(P)$ on the interval (0,1)

$$\text{lfdr}_j = \Pr(H_j \mid P_j, \widehat{f}, \widehat{\pi}_0) = \frac{\Pr(H_j \mid \widehat{f}, \widehat{\pi}_0) \cdot \Pr(P_j \mid H_j, \widehat{f}, \widehat{\pi}_0)}{\Pr(P_j \mid \widehat{f}, \widehat{\pi}_0)} = \frac{\widehat{\pi}_0 \cdot 1}{\widehat{f}(P_j)} = \frac{\widehat{\pi}_0}{\widehat{f}(P_j)}.$$

H_j is the null hypot. for variable *j*



At every point on x-axis, LFDR compares the estimated $\hat{\pi}_0$ to the estimated P-value density at that point.

If there are similar density of *P*-values as expected based on $\hat{\pi}_0$, then LFDR = 1, whereas if there are more *P*-values than expected by $\hat{\pi}_0$, then LFDR < 1.