Introduction to probabilistic inference
Coryn Bailer-Jones, MPIA, 14 & 21 April 2023

Session 1

- concepts of inference
- zero- and one-parameter models
- assigning priors

Session 2

- multiple parameter estimation
- line fitting with MCMC
- model comparison

Interpreting test results

Interpreting test results

A test for a disease gives either a positive or a negative result, and is 90% reliable (true positive rate; sensitivity).

You test positive. What is the probability that you have the disease?

- 90%?
- >90%?
- <90%?

Interpreting test results

A test for a disease gives either a positive or a negative result, and is 90% reliable (true positive rate; sensitivity).

The test has a false positive rate of 7%.

You test positive. What is the probability that you have the disease?

- 90%?
- 93%?
- 83%?
- other?

Interpreting test results

A test for a disease gives either a positive or a negative result, and is 90% reliable (true positive rate; sensitivity).

The test has a false positive rate of 7%.

The base rate is 0.8%.

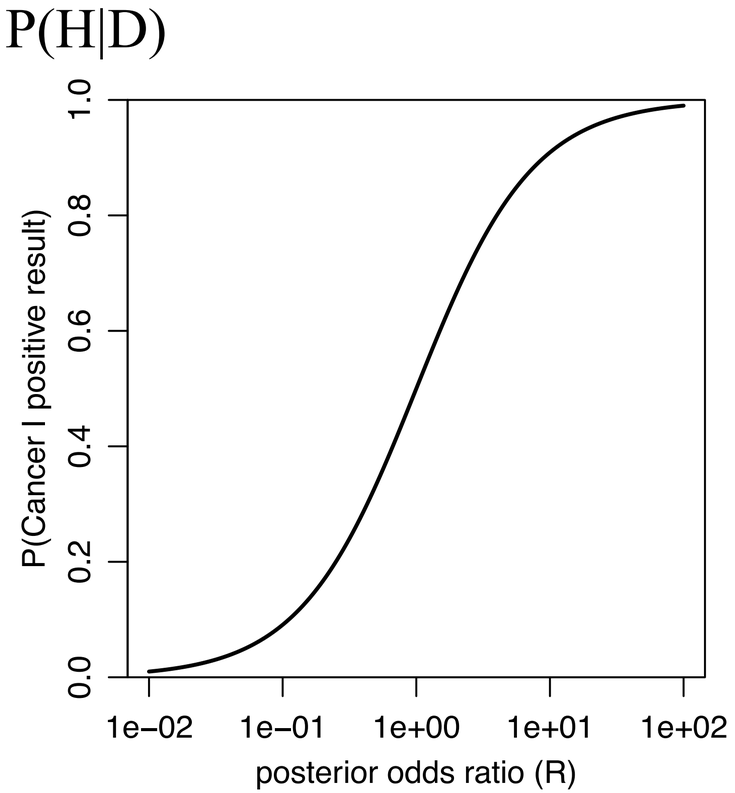
You test positive. What is the probability that you have the disease?

- 90%?
- 93%?
- >90%?
- <10%?
- other?

Interpreting test results

Result D	Hypothesis H true?	
	yes	**no**
positive	true positive $P(D\|H)$	false positive $P(D\|\overline{H})$
negative	false negative $P(\overline{D}\|H)$	true negative $P(\overline{D}\|\overline{H})$

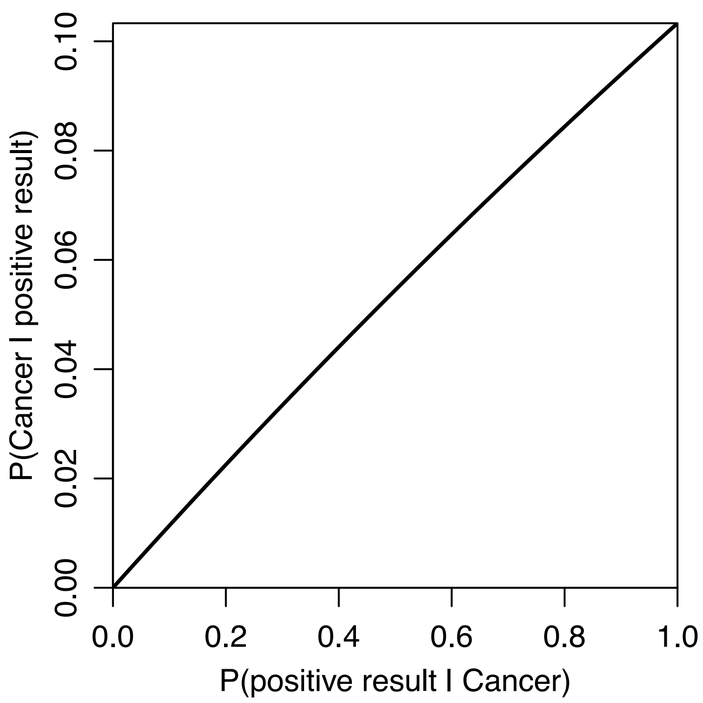
$$P(H|D) = \frac{1}{1 + \frac{1}{R}} \qquad R = \frac{P(D|H)P(H)}{P(D|\overline{H})P(\overline{H})}$$



Interpreting test results

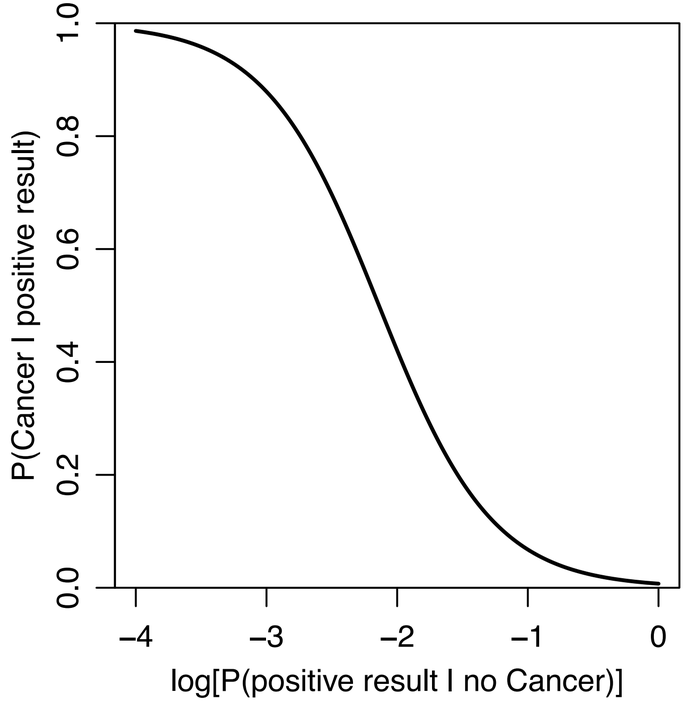
Nominal values
P(D|H) = 0.900
P(D|!H) = 0.070
P(H) = 0.008

P(H|D)



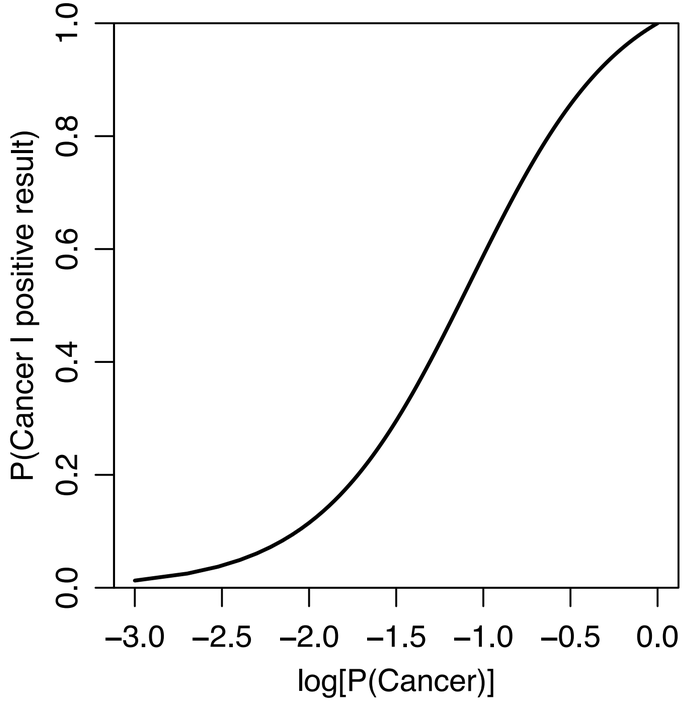
P(D|H)

P(H|D)



P(D|!H)

P(H|D)



P(H)

Thinking in terms of frequencies

Single parameter estimation

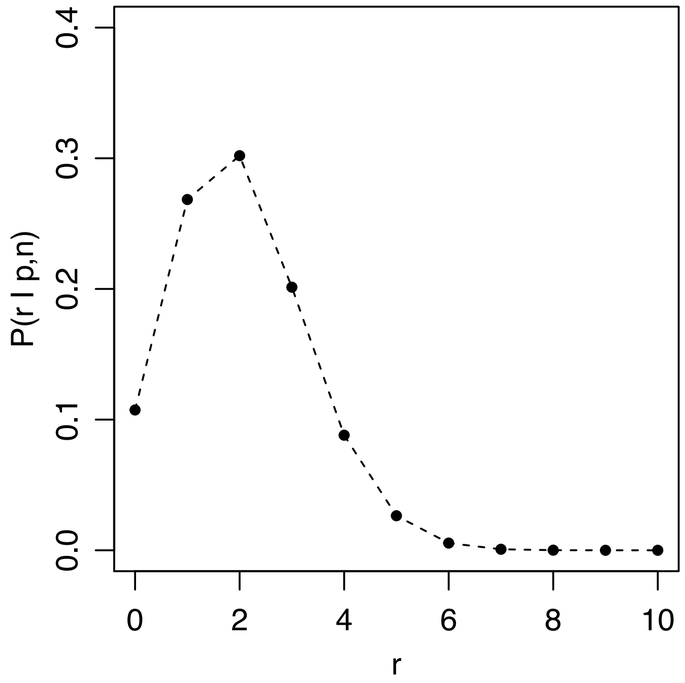
- Occurrence of objects in a large population, e.g. brown dwarfs among all stars
- Survey just n objects, of which r are brown dwarfs
- What is the true population fraction, p, of brown dwarfs?

- Cast this as an inference problem
- First question: how were the data generated?
- P(D|H)
- P(H)
- P(H|D)

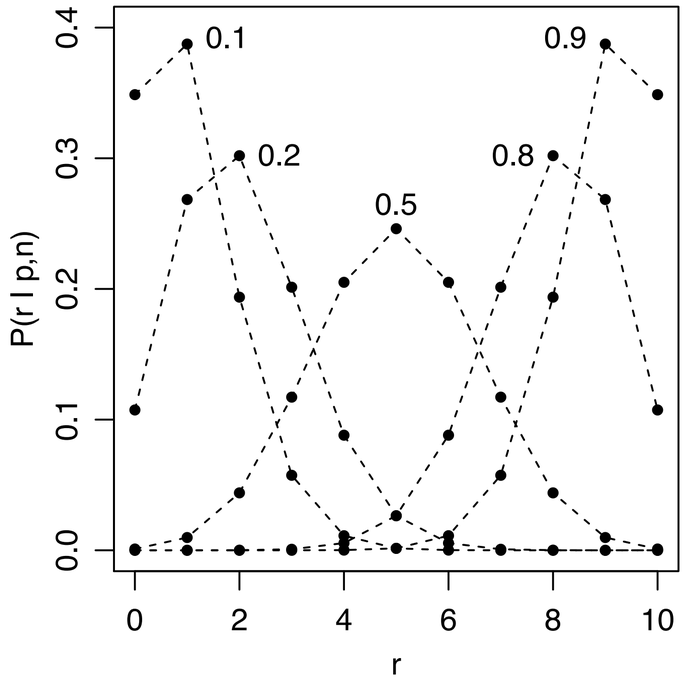
Binomial distribution

$$P(r|p,n) = \frac{n!}{r!(n-r)!}\, p^r (1-p)^{n-r} \quad \text{where} \quad n \geq 0,\ r \geq 0,\ 0 \leq p \leq 1.$$

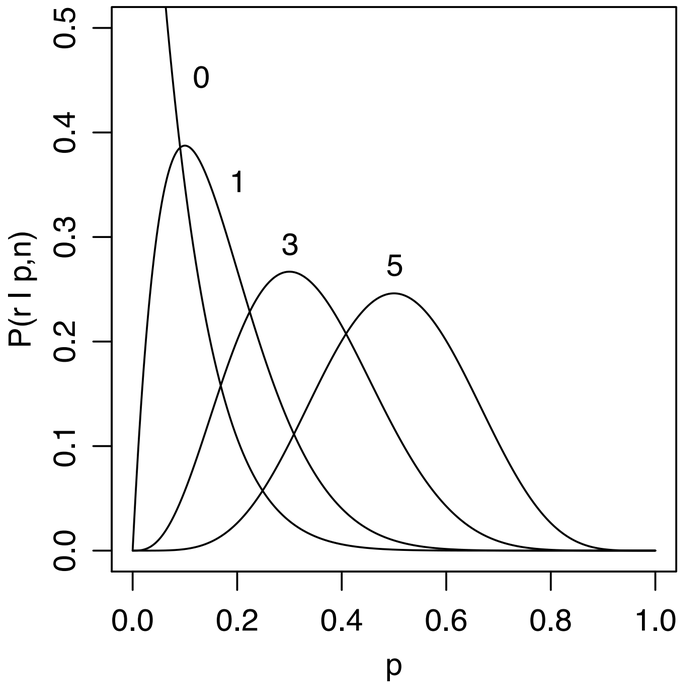
n=10, p=0.2



n=10, p=(0.1,0.2,0.5,0.8,0.9)



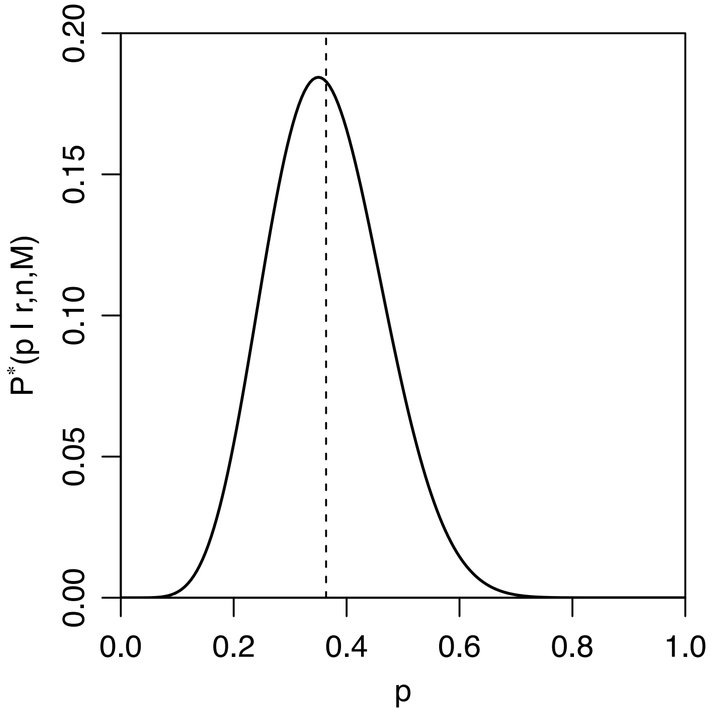
not a PDF in p
n=10, r=(0,1,3,5)



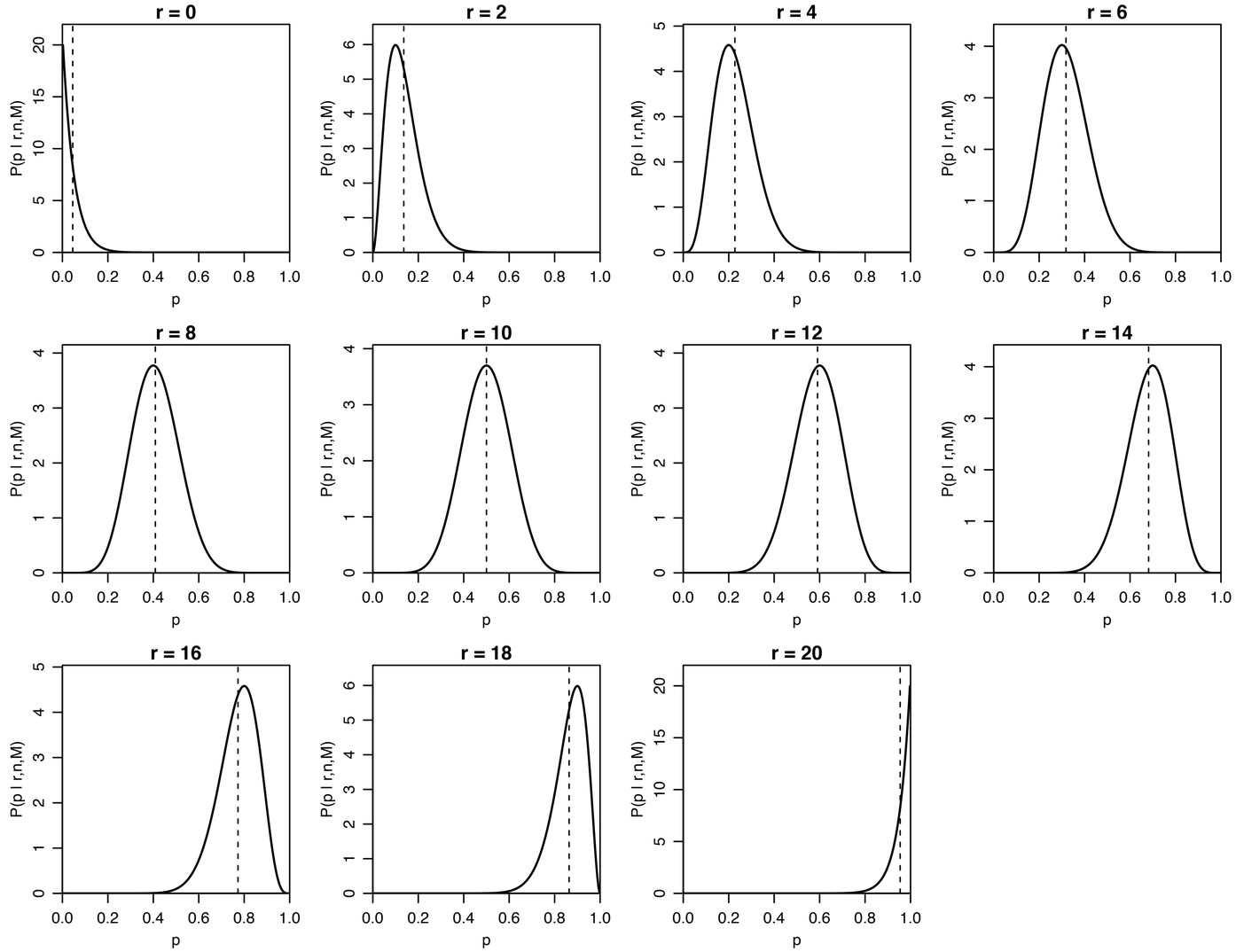
Posterior with uniform prior

$$P^*(p|r,n) \;=\; \frac{n!}{r!(n-r)!}\, p^r (1-p)^{n-r}$$

n=20, r=7, uniform prior
unnormalized posterior



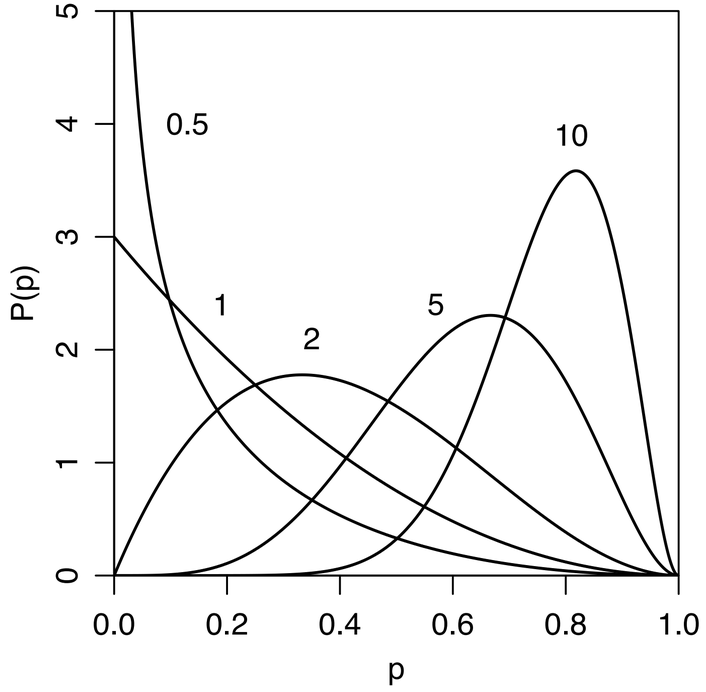
n=20, various r, uniform prior
normalized posterior



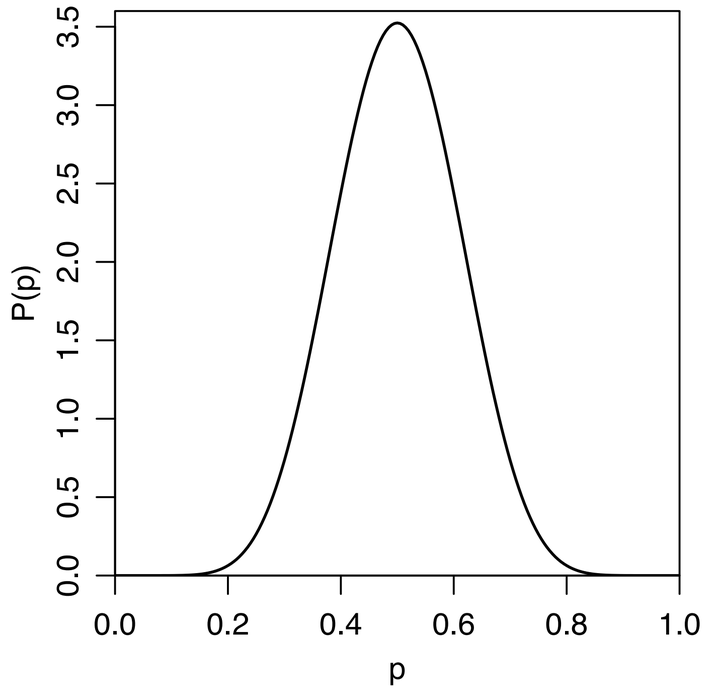
Beta distribution prior

$$P(p) = \frac{1}{B(\alpha,\beta)}\, p^{\alpha-1}(1-p)^{\beta-1} \quad \text{where} \quad \alpha > 0,\ \beta > 0,\ 0 \le p \le 1$$

α=(0.5,1,2,5,10), β=7



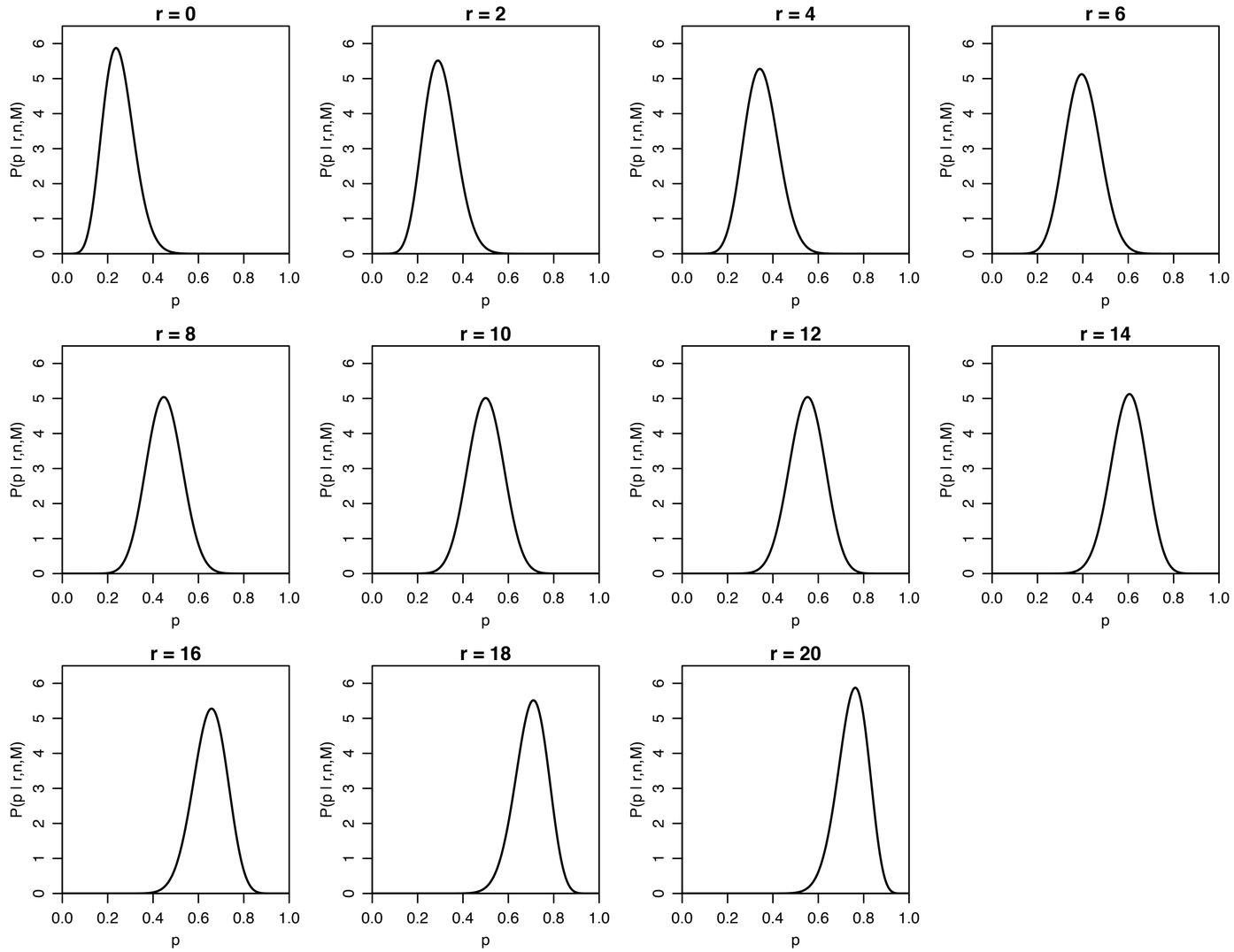
α=10, β=10



Posterior with beta prior

$$P^*(p|r,n) = p^{r+\alpha-1}(1-p)^{n-r+\beta-1}$$

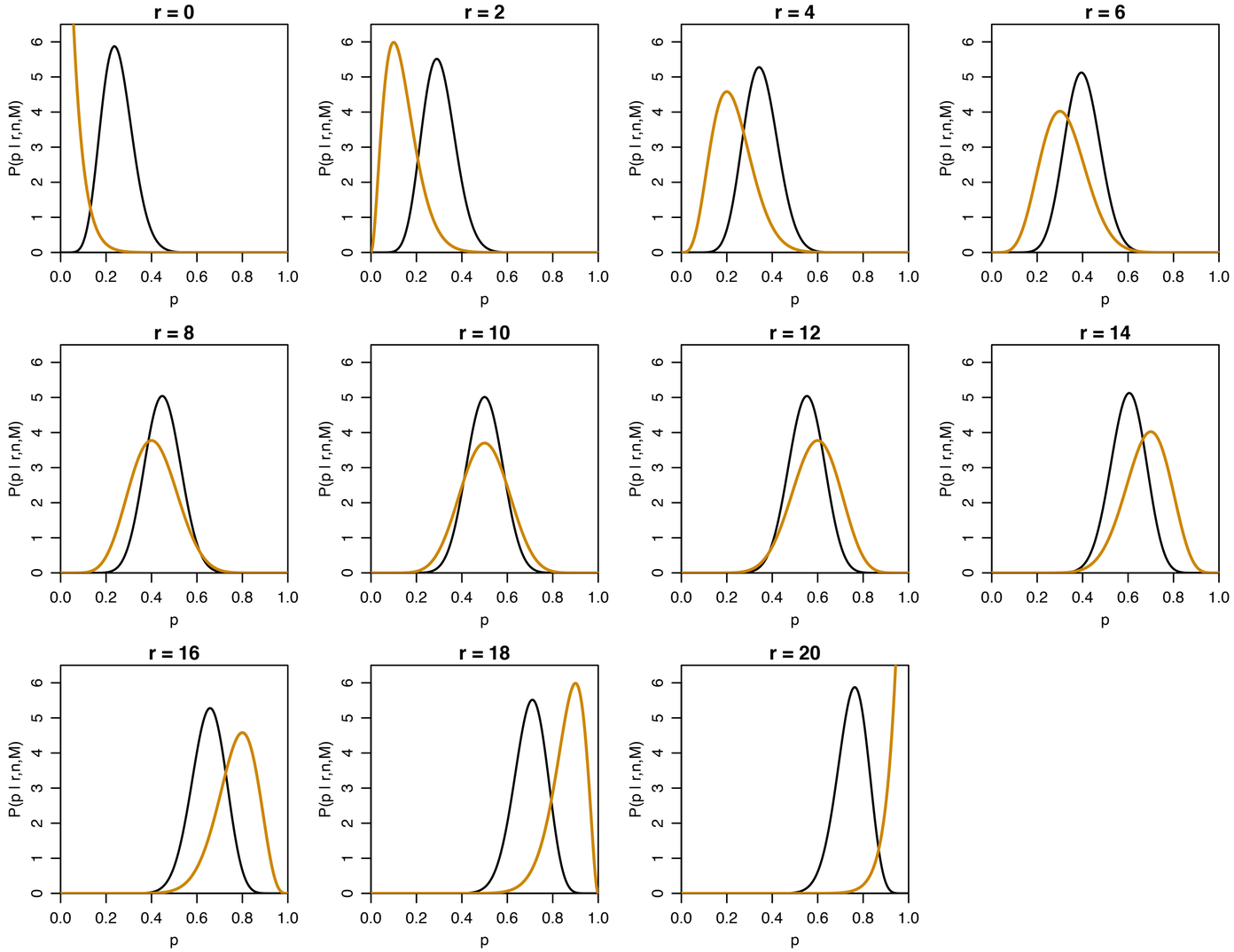
n=20, various r, normalized posterior with α=10, β=10;



Posterior with beta prior

$$P^*(p|r,n) = p^{r+\alpha-1}(1-p)^{n-r+\beta-1}$$

n=20, various r, normalized posterior
black=with α=10, β=10; orange=with α=1, β=1 (uniform)



Posterior with beta prior

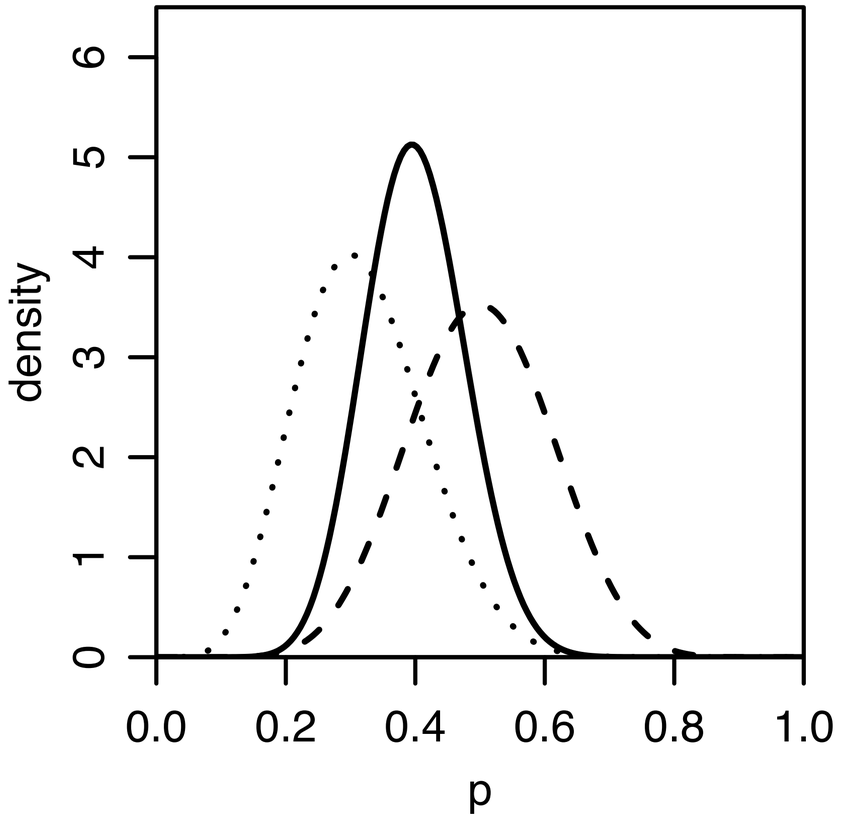
$$P^*(p|r,n) = p^{r+\alpha-1}(1-p)^{n-r+\beta-1}$$

$$P^*(p) = p^{\alpha-1}(1-p)^{\beta-1}$$

$$P^*(r|p,n) = p^r(1-p)^{n-r}$$

(beta) prior is *conjugate* for this (binomial) likelihood

n=20,r=6, with α=10, β=10
solid=posterior, prior=dashed, likelihood*=dotted



*likelihood is not a PDF over p
so cannot be shown
properly normalized.
Shown here with unit area.

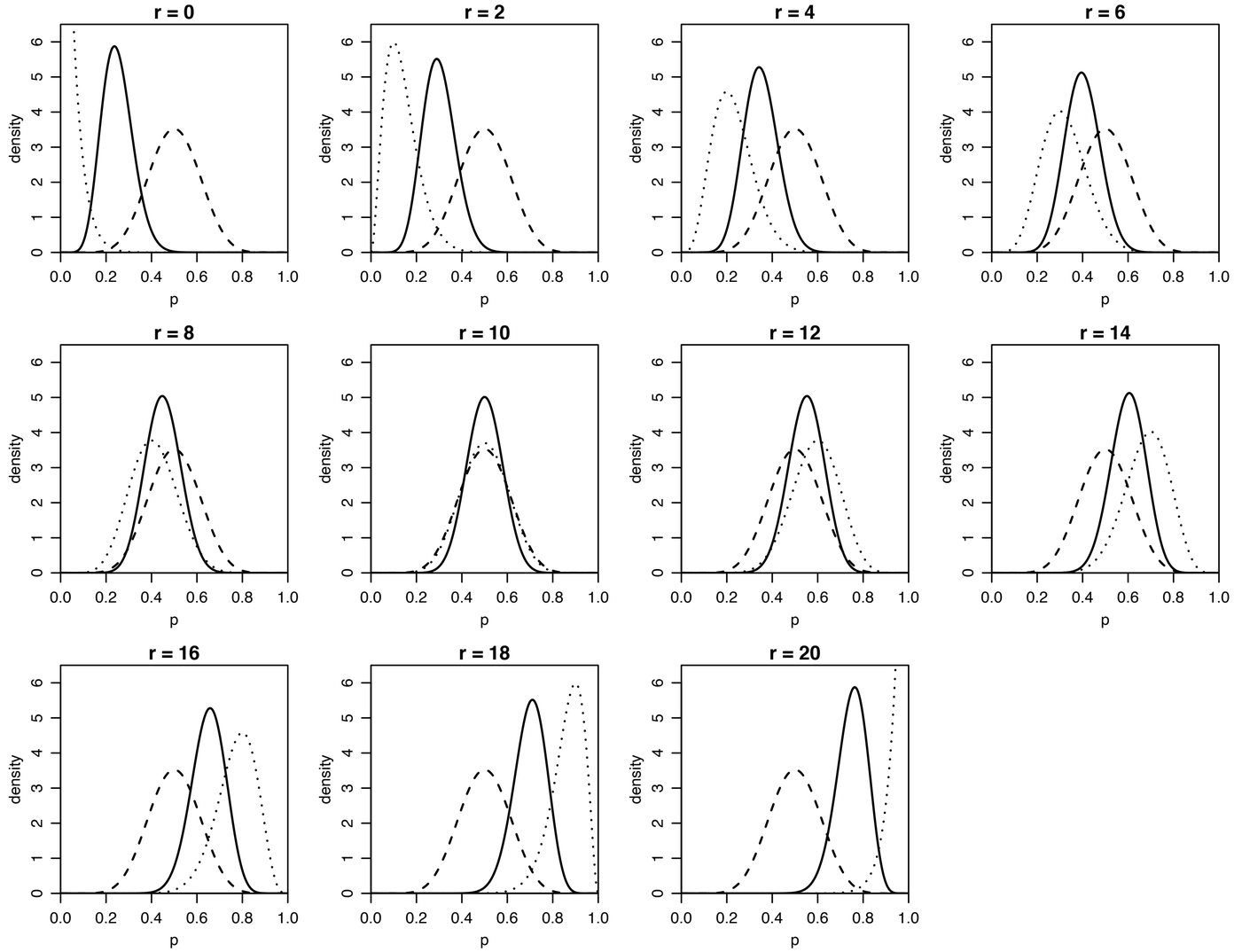
Posterior with beta prior

$$P^*(p|r,n) = p^{r+\alpha-1}(1-p)^{n-r+\beta-1}$$

$$P^*(p) = p^{\alpha-1}(1-p)^{\beta-1}$$

$$P^*(r|p,n) = p^r(1-p)^{n-r}$$

n=20, various r, with α=10, β=10
solid=posterior, prior=dashed, likelihood*=dotted



*likelihood is not a PDF over p
so cannot be shown
properly normalized.
Shown here with unit area.

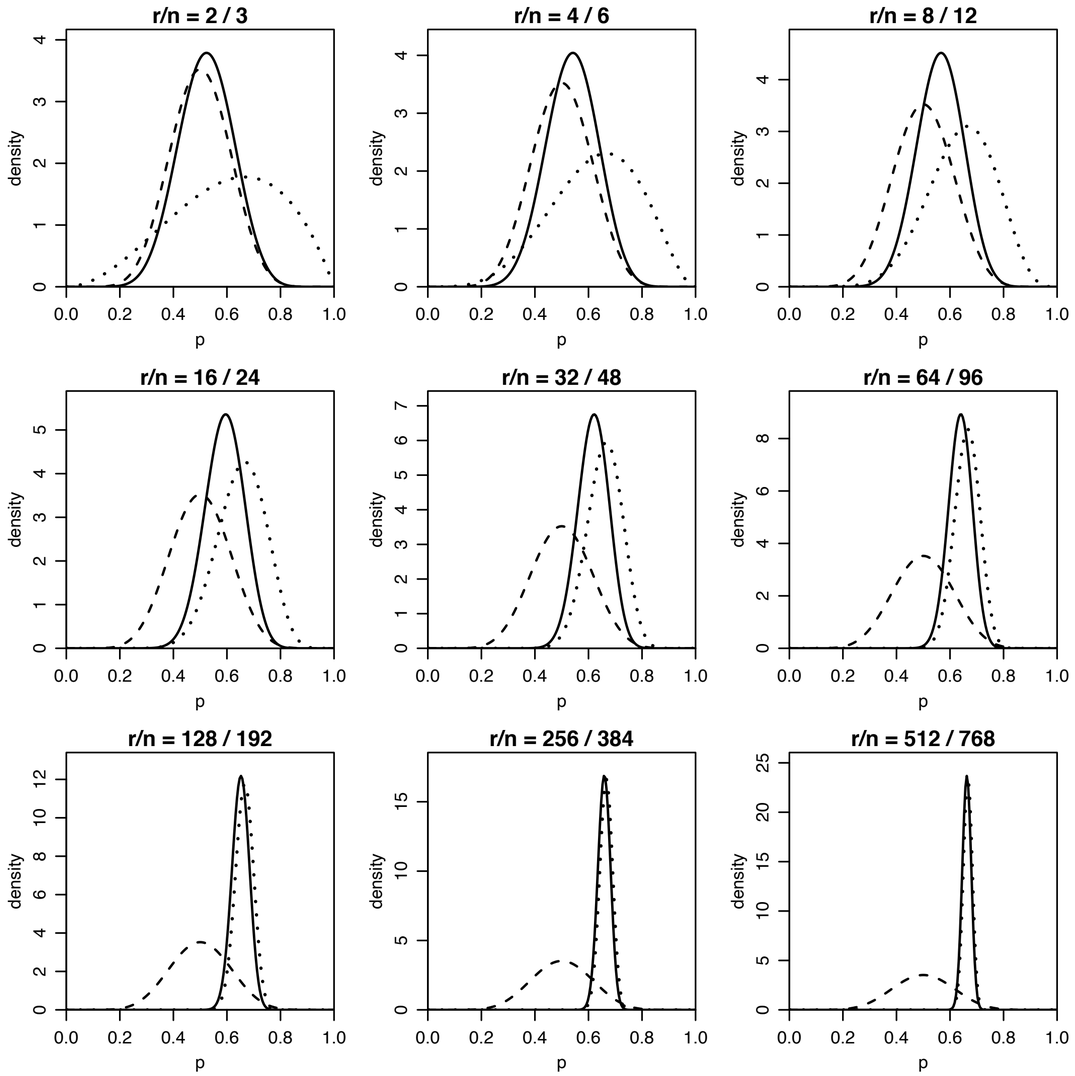
Varying the amount of data

$$P^*(p|r,n) = p^{r+\alpha-1}(1-p)^{n-r+\beta-1}$$

$$P^*(p) = p^{\alpha-1}(1-p)^{\beta-1}$$

$$P^*(r|p,n) = p^r(1-p)^{n-r}$$

r/n=2/3, various n, with α=10, β=10
solid=posterior, prior=dashed, likelihood*=dotted



*likelihood is not a PDF over p
so cannot be shown
properly normalized.
Shown here with unit area.

Summarizing the posterior

$$P^*(p|r,n) = p^{r+\alpha_p-1}(1-p)^{n-r+\beta_p-1}$$

$$\text{mean} = \frac{\alpha_p + r}{\alpha_p + \beta_p + n}$$

$$\text{mode} = \frac{\alpha_p + r - 1}{\alpha_p + \beta_p + n - 2}$$

Assigning priors

Multiple parameter estimation

- Set of measurements, D, assumed to be generated from model M with parameters θ
- We wish to estimate θ using D, i.e. determine the posterior $P(\theta \mid D,M)$
- Procedure
  - Decide on generative ("forward") model for D, $f(\theta)$
  - Decide on likelihood, $P(D \mid \theta,M)$
  - Decide on prior, $P(\theta \mid M)$
  - Product is **un**normalized posterior $P^*(\theta \mid D,M) = P(D \mid \theta,M)\, P(\theta \mid M)$
  - Normalization constant, $P(D \mid M)$, doesn't depend on θ, so often not required to estimate θ
  - Evaluate (compute/sample/visualize) the posterior
  - Summarize the posterior, e.g. median, mode, confidence intervals, (mean, st.dev.)

Estimating the parameters of a Gaussian

- M is a Gaussian
- θ is the mean and standard deviation
- D are draws from the Gaussian
- Example: 1D velocity dispersion of a population of stars

Estimating the parameters of a Gaussian

$D = \{x_i\}$ assumed drawn from a Gaussian with unknown mean μ and standard deviation σ

$$P(D|\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right]$$

$$= \frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left[-\frac{1}{2\sigma^2}(N(\overline{x}-\mu)^2 + NV_x)\right] \quad \text{where} \quad \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and}$$

$$V_x = \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2$$

Estimating the parameters of a Gaussian

$$P(D|\mu,\sigma) = \frac{1}{(2\pi)^{N/2}\sigma^N}\exp\left[-\frac{1}{2\sigma^2}(N(\overline{x}-\mu)^2+NV_x)\right]$$

likelihood

$$P(\mu) \propto \text{const}$$

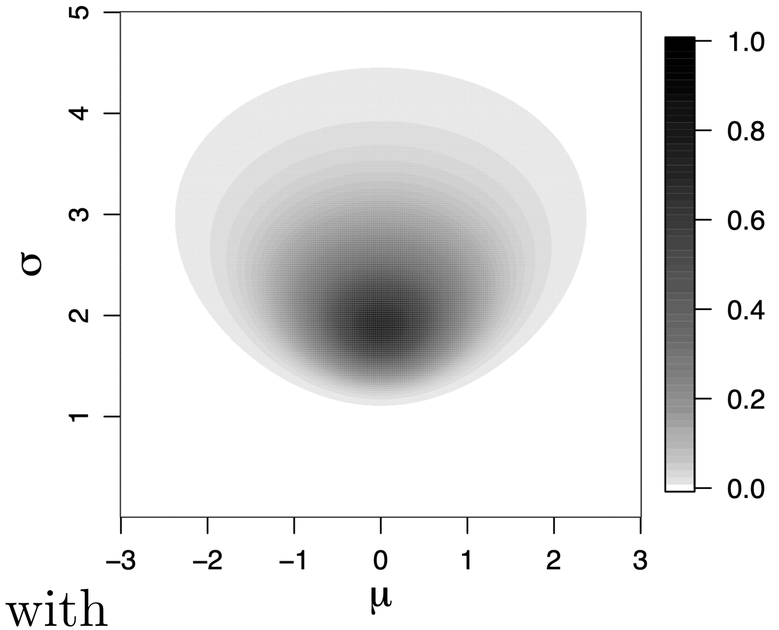
$$P(\sigma) \propto \frac{1}{\sigma}$$

$$P(\mu,\sigma) \propto \frac{1}{\sigma}$$

prior

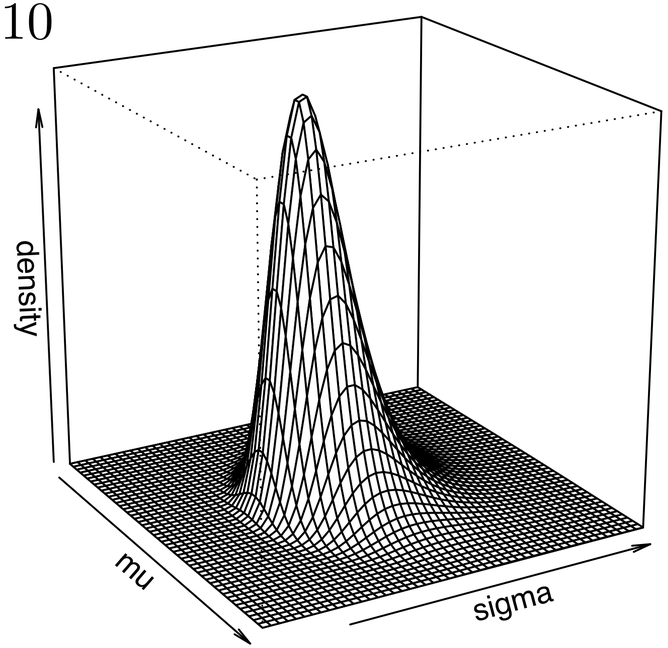
$$P(\mu,\sigma|D) \propto \frac{1}{\sigma^{N+1}}\exp\left[-\frac{1}{2\sigma^2}\left(N(\overline{x}-\mu)^2+NV_x\right)\right]$$

posterior

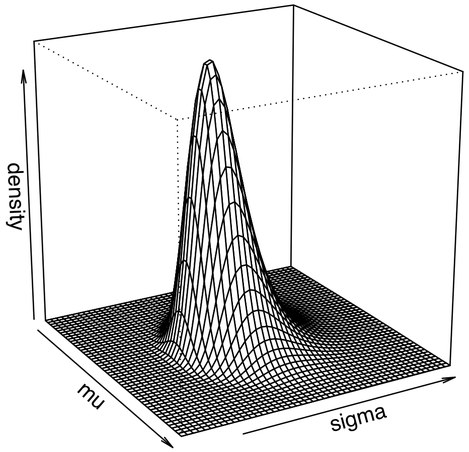


posterior with

$$\overline{x}=0, V_x=2^2, N=10$$



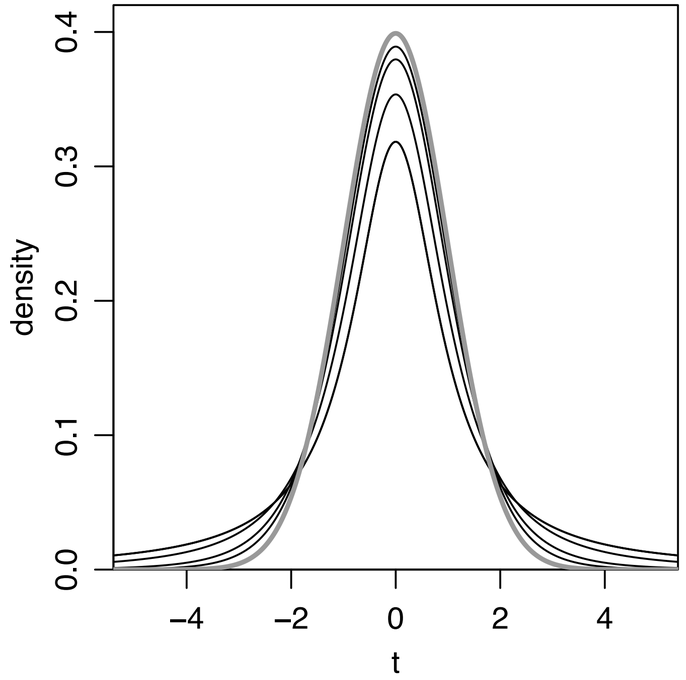
Marginal posterior distribution: mean



$$P(\mu|D) = \int_0^\infty P(\mu, \sigma|D)\, d\sigma$$

$$\propto \left[1 + \frac{(\overline{x} - \mu)^2}{V_x}\right]^{-N/2}$$

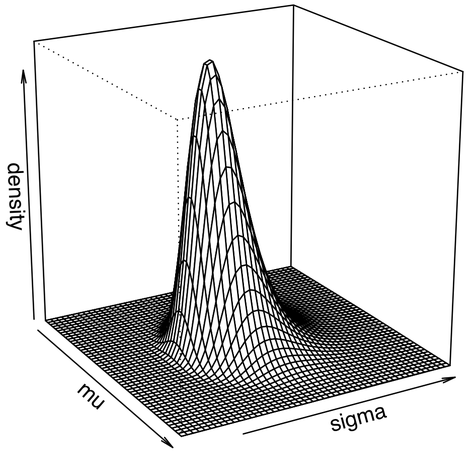
Student t distribution with 1,2,5,10 d.o.f
(thick grey line is unit Gaussian)



This is a Student t distribution with N-1 degrees of freedom (d.o.f) and

$$t = \frac{(\overline{x} - \mu)}{\sqrt{V_x/(N-1)}}$$

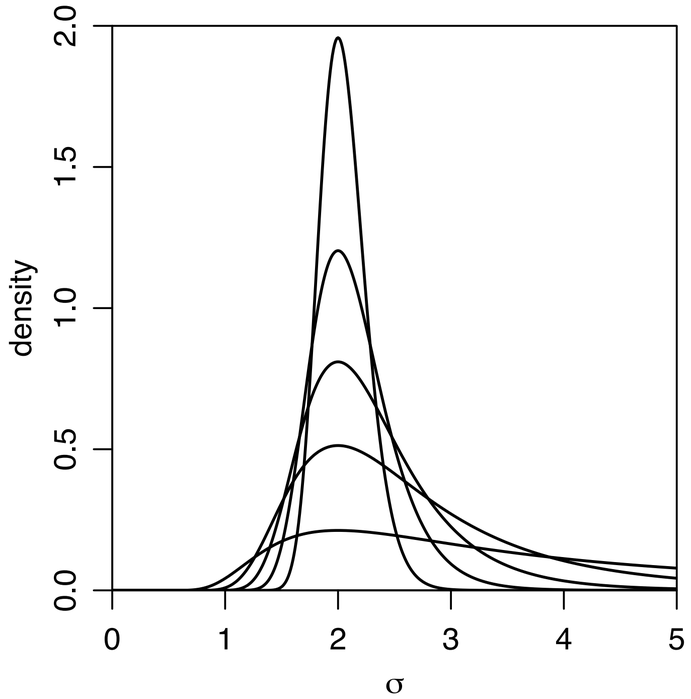
Marginal posterior distribution: standard deviation



$$P(\sigma|D) = \int_{-\infty}^{\infty} P(\mu, \sigma|D)\, d\mu$$

$$\propto \frac{1}{\sigma^N} \exp\left[-\frac{NV_x}{2\sigma^2}\right]$$

$V_x = 2^2$, N=2,5,10,20,50



$$P(\sigma^2|D) = \frac{1}{2\sigma} P(\sigma|D)$$

is an inverse gamma distribution

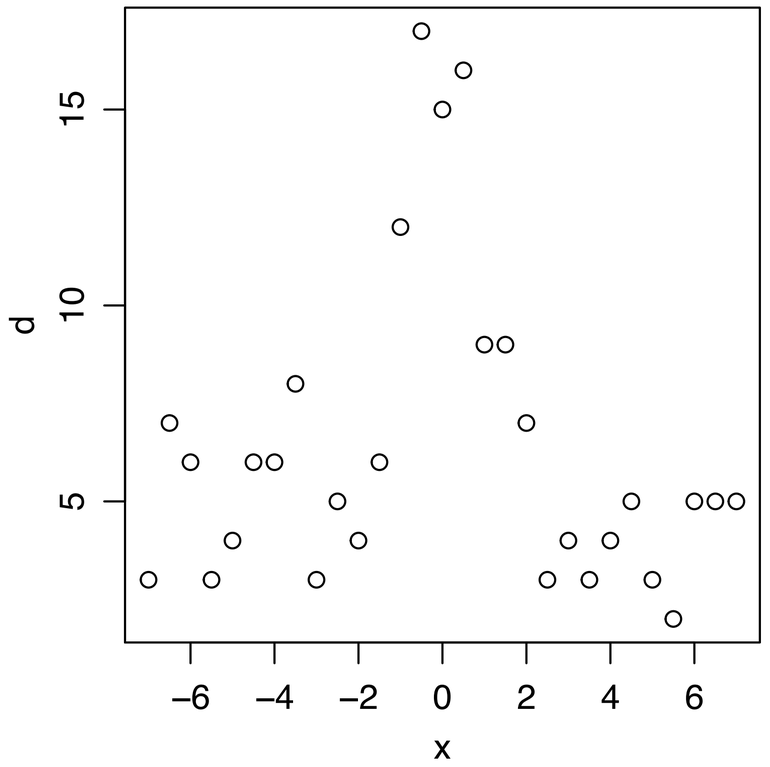
Exercise: estimating signal and background

Measure number of photons d as a function of wavelength x.
Expected signal is a combination of a Gaussian-shaped line and a background.
For exposure time t:

$$s = t\left[a\exp\left(-\frac{(x-x_0)^2}{2w^2}\right)+b\right]$$

You are given data D = $\{d_i(x_i)\}$ for known exposure time t.
The line profile (x_0 and w) is known.
Find the amplitude of the line.

x_0=0, w=1, t=5



Exercise: estimating signal and background

- What process generated the data?
  - What is the likelihood?
- What are sensible priors?
- How do you sample the posterior?
- How do you visualize the posterior?
- How do you compute a useful summary of this posterior?

https://github.com/bailer-jones/PBI-MPIA2023

Exercise: estimating signal and background

$$P(d|s) = \frac{s^d e^{-s}}{d!}$$

$$P(D|x_0, w, t, a, b) = \prod_i \frac{s_i^{d_i} e^{-s_i}}{d_i!}$$

Prior: uniform for a,b >=0, zero otherwise

$$P(a, b|D, M) = \begin{cases} \frac{1}{Z} \prod_i \frac{s_i^{d_i} e^{-s_i}}{d_i!} & \text{if } a \geq 0 \text{ and } b \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Exercise: parameter estimation via grid sampling

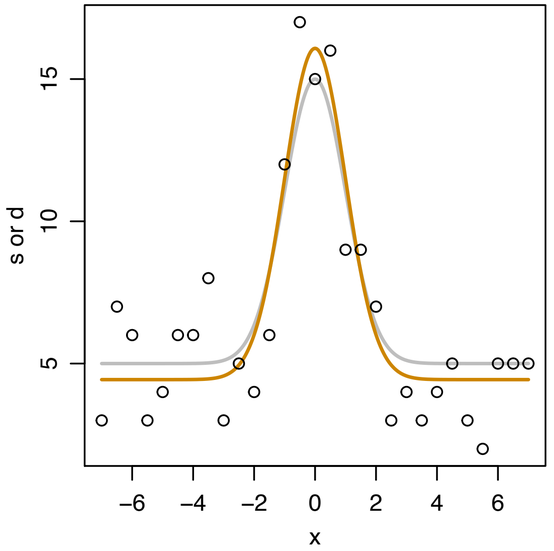
$$
\begin{aligned}
P(a_j|D) &\simeq \delta b \sum_{k=1}^{K} P(a_j, b_k|D) \\
\mu_a &= \int a P(a|D) da \simeq \delta a \sum_{j=1}^{K} a_j P(a_j|D) \\
\sigma_a^2 &= \int (a-\mu_a)^2 P(a|D) da \simeq \delta a \sum_{j=1}^{K} (a_j-\mu_a)^2 P(a_j|D) \\
\mathrm{cov}(a,b) &= \iint (a-\mu_a)(b-\mu_b) P(a,b|D)\, da\, db \\
&\simeq \delta a \delta b \sum_{j=1}^{K}\sum_{k=1}^{K} (a_j-\mu_a)(b_k-\mu_b) P(a_j, b_k|D) \\
\rho &= \frac{\mathrm{cov}(a,b)}{\sigma_a \sigma_b}
\end{aligned}
$$

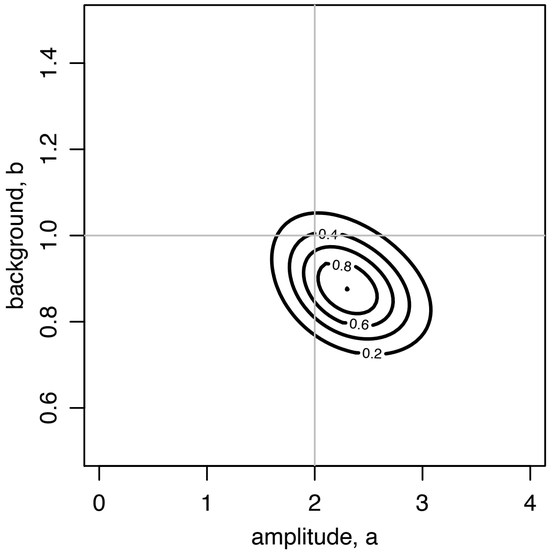
Exercise: estimating signal and background

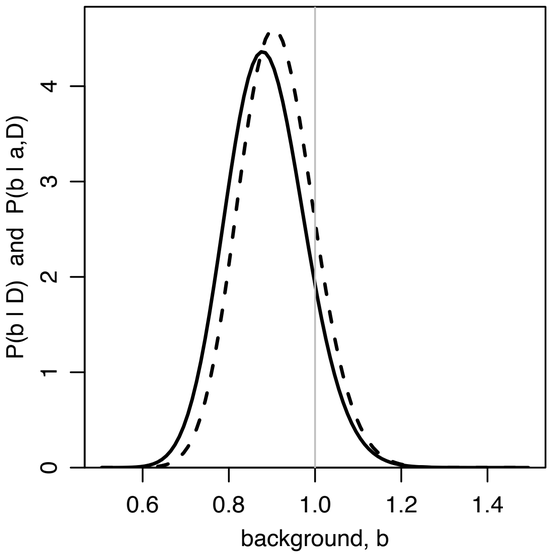
grey=true
orange=mean

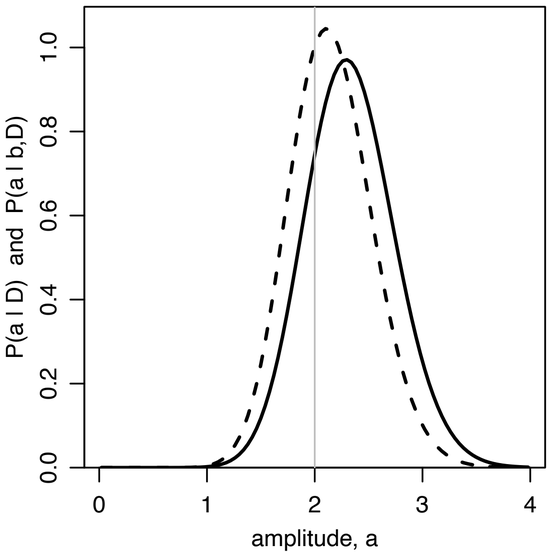
Inference (mean ± s.d.):
a = 2.33 ± 0.41 [true=2]
b = 0.887 ± 0.092 [true=1]
ϱ = -0.35

solid=marginalized
dashed=conditional
on true









Example adapted from Sivia & Skilling (2006)

Why we can’t always sample on a grid
or
Why we need efficient sampling

Fitting a line

$$y = f(x) + \epsilon \qquad \text{where}$$
$$f(x) = b_0 + b_1 x$$

forward model

$$P(y_i|x_i, \theta, M) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{[y_i - f(x_i; b_0, b_1)]^2}{2\sigma^2}\right]$$

likelihood for one data point

$$\ln P(\{y_i\}|\{x_i\}, \theta, M) = \sum_{i=1}^{N} \ln P(y_i|x_i, \theta, M)$$

log likelihood for many data points assuming independence

Fitting a line: priors

$$y = f(x) + \epsilon \quad \text{where}$$
$$f(x) = b_0 + b_1 x$$
$$P(y_i|x_i, \theta, M) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{[y_i - f(x_i; b_0, b_1)]^2}{2\sigma^2}\right]$$

forward model

likelihood for one data point

- b_0, intercept
- b_1, gradient
- σ, standard deviation

what priors shall we use?

$$P^*(\theta = (b_0, \alpha, \sigma)) = \exp\left[-\frac{(b_0 - m)^2}{2s^2}\right] \times 1 \times \log\sigma$$

unnormalized prior

Fitting a line

- See code (linear_model_posterior.R)

Fitting a line: posterior predictive distribution

$$P(y_p \,|\, x_p, \hat{\theta})$$

$$P(y_p \,|\, x_p, D) = \int \underbrace{P(y_p \,|\, x_p, \theta)}_{\text{likelihood } y_p} \underbrace{P(\theta \,|\, D)}_{\text{posterior}} \, d\theta$$

$$= \frac{1}{P(D)} \int \underbrace{P(y_p \,|\, x_p, \theta)}_{\text{likelihood } y_p} \underbrace{P(D \,|\, \theta)}_{\text{likelihood } D} \underbrace{P(\theta)}_{\text{prior}} \, d\theta$$

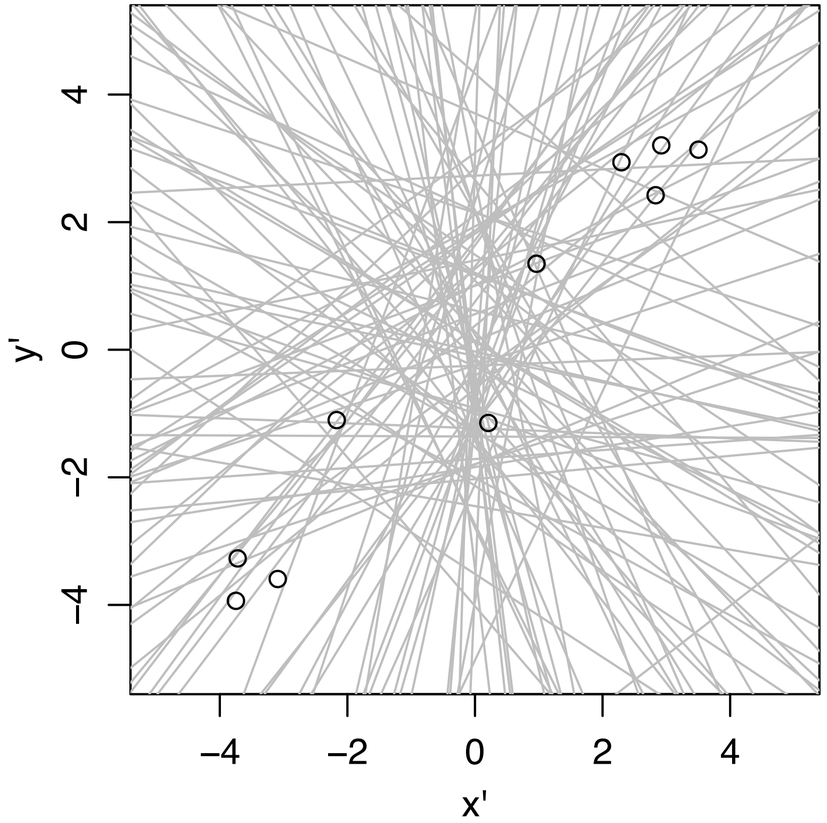
Exercise: line fitting

- Experiment with changing
  - number of data points
  - priors
- Questions
  - what is the impact on the inference as we increase/decrease the amount of data?
  - what about changing the priors?
  - how (conceptually) can we infer the “error bars” (sigma in the likelihood)?
  - what happens if we only have 2, 1, or 0 data points - less (fewer?) data than parameters?

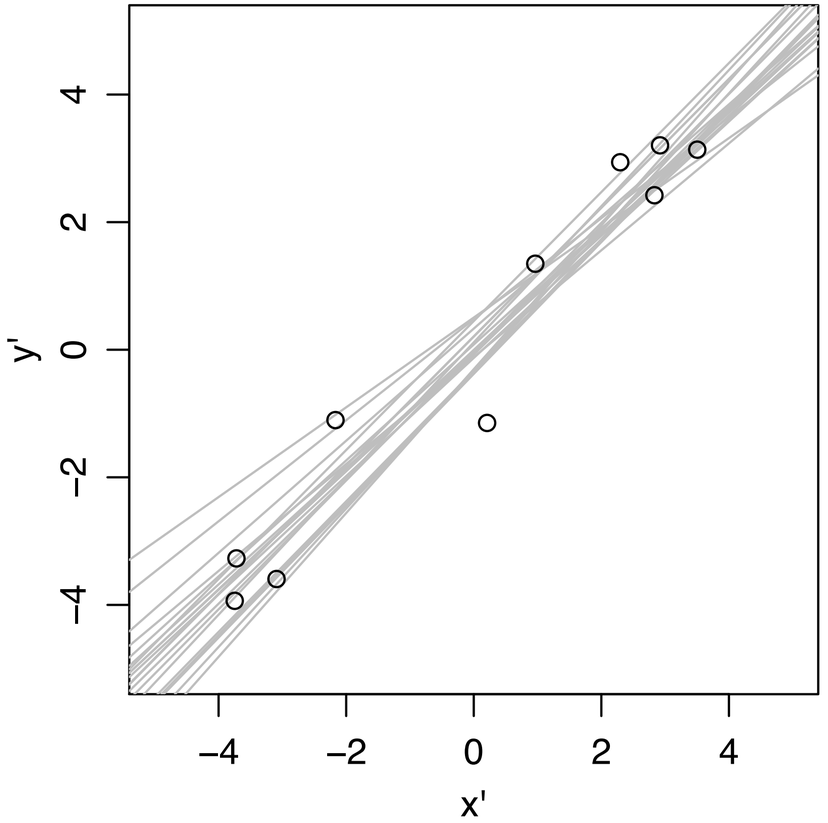
https://github.com/bailer-jones/PBI-MPIA2023

Exercise: line fitting

samples from the prior



samples from the posterior



Model comparison

$$P(\theta|D,M) = \frac{P(D|\theta,M)P(\theta|M)}{P(D|M)}$$

$$P(D|M) = \int \underbrace{P(D|\theta,M)}_{\text{likelihood}} \underbrace{P(\theta|M)}_{\text{prior}}\, d\theta$$

“marginal likelihood” or “evidence” of a model

$$BF_{12} = \frac{P(D|M_1)}{P(D|M_2)}$$

“Bayes factor” between two models

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

“posterior odds ratio” between two models

Model comparison: line fitting

Model 2
Straight line with any gradient

$$
\begin{aligned}
P(b_0) &= \mathcal{N}(0, 1) \\
P(\alpha) &= \mathcal{U}(0, 2\pi) \\
P(\log\sigma) &= \mathcal{U}(\log 0.5, \log 2)
\end{aligned}
$$

Model 1
Horizontal line

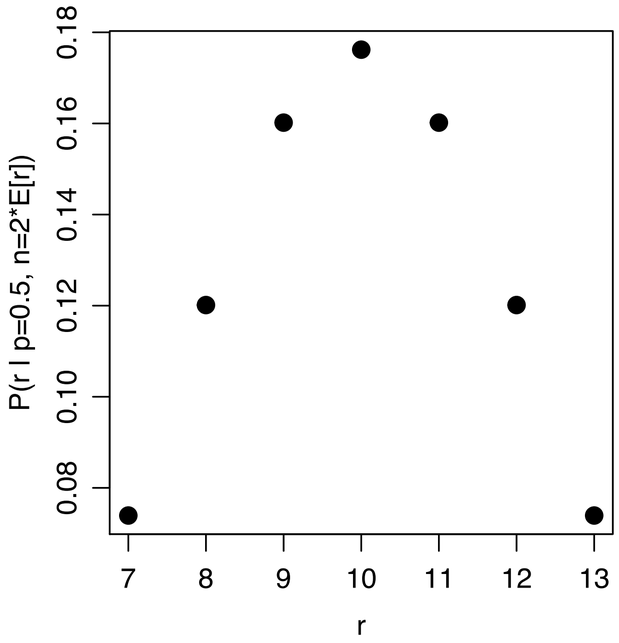
$$
\begin{aligned}
P(b_0) &= \mathcal{N}(0, 1) \\
P(\alpha) &= \delta(\alpha) \\
P(\log\sigma) &= \mathcal{U}(\log 0.5, \log 2)
\end{aligned}
$$

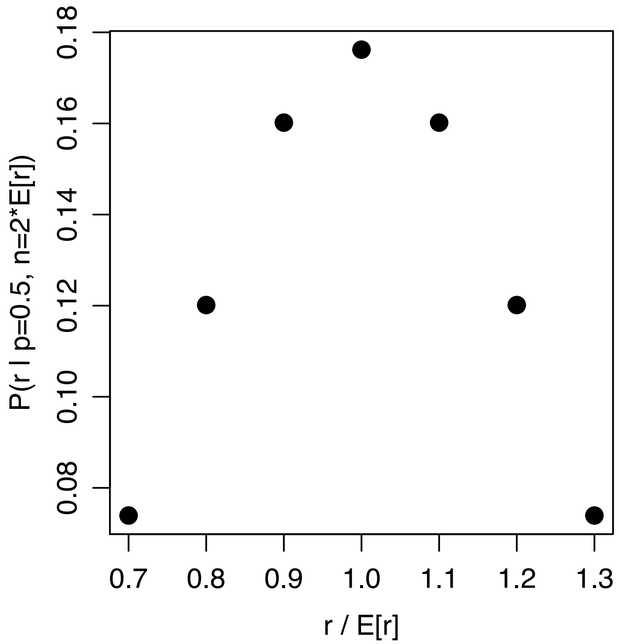
- See code (linear_model_evidence.R)

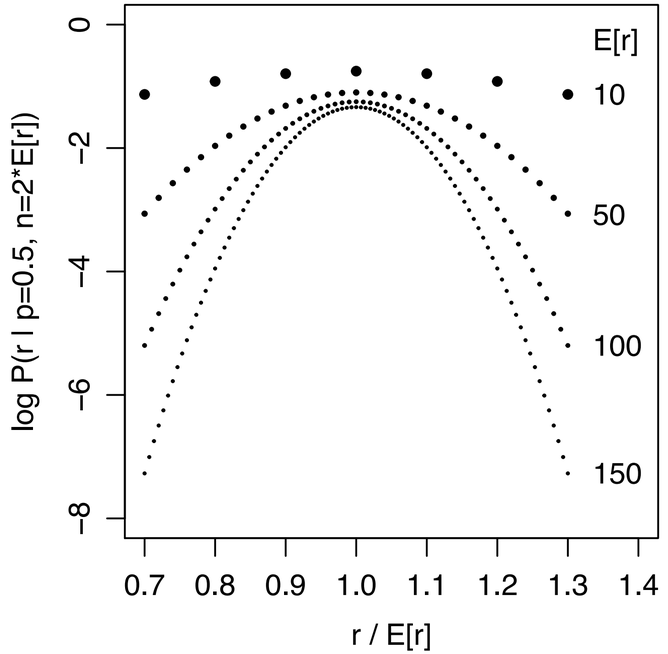
Model comparison: coin tossing

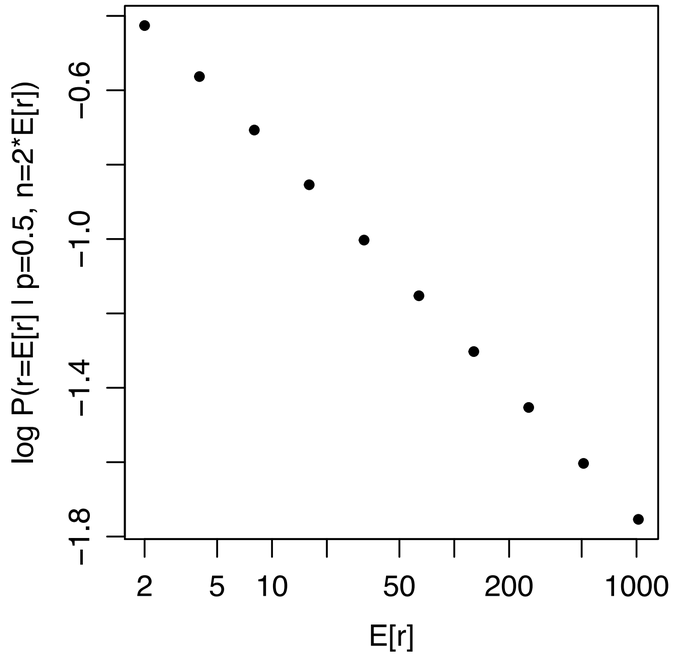
$$P(r|p,n) = \frac{n!}{r!(n-r)!}\, p^r (1-p)^{n-r}$$ likelihood

n=20, p=0.5 => E[r]=10









Model comparison: coin tossing

Model 1: fair coin (p=0.5)

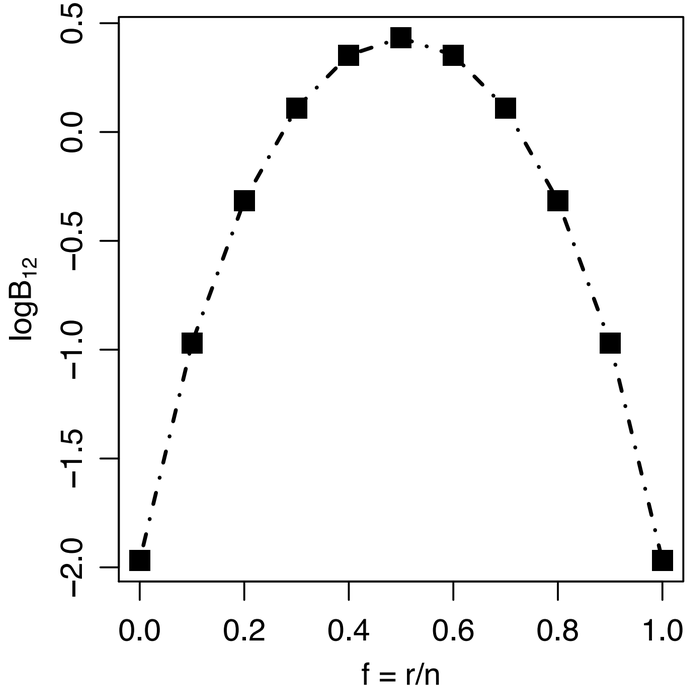
$P^*(p \mid M_1) = \delta(p-0.5)$

Model 2: p unknown

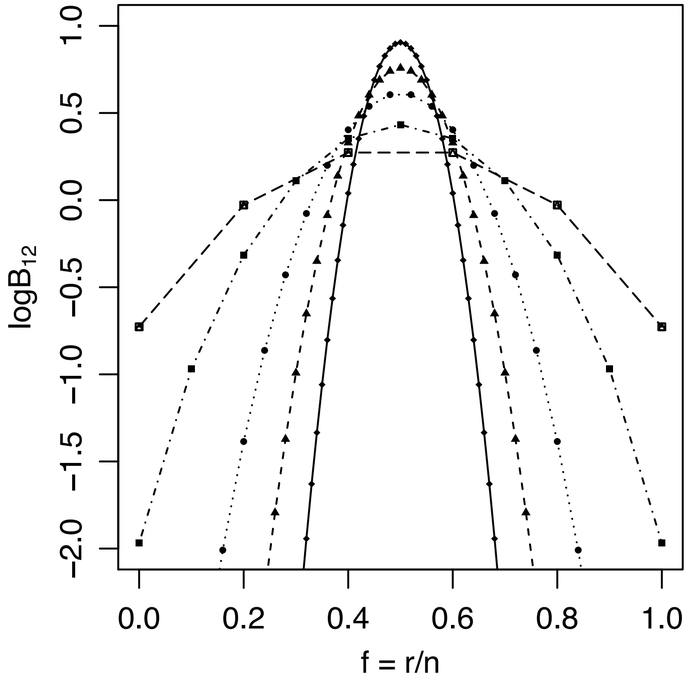
$P^*(p \mid M_2) = 1$ (i.e. uniform)

Toss coin n times and get r heads

n = 10



n = (5,10,25,50,100)
higher is narrower

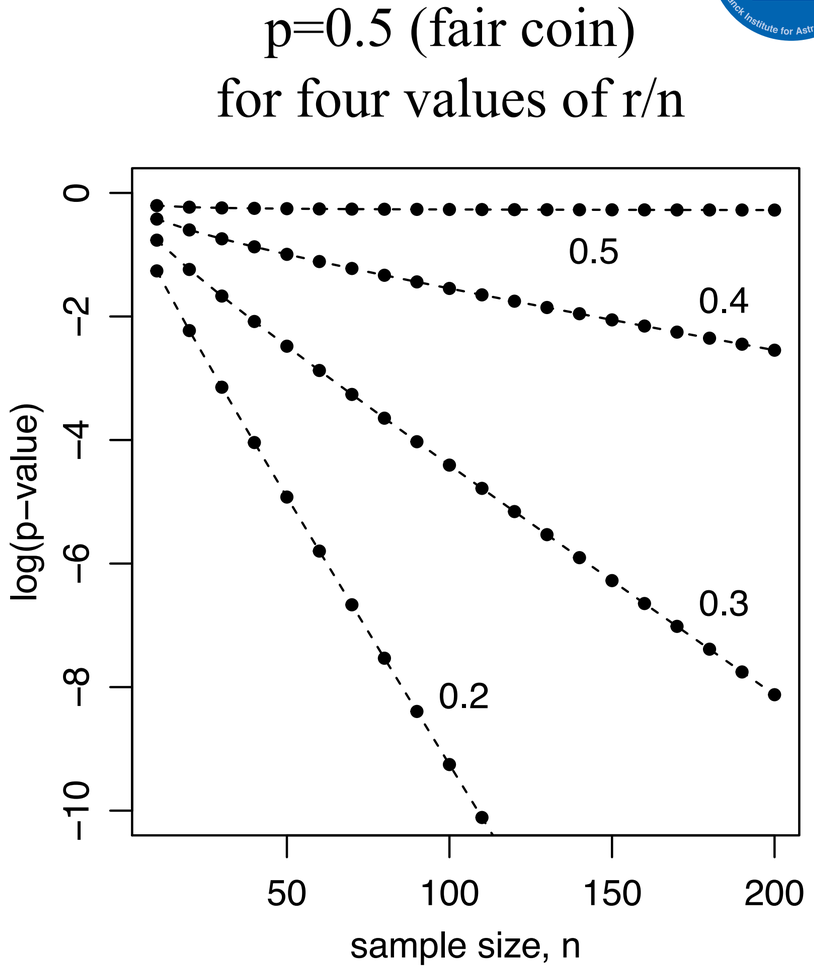


log B_{12} = 0 means model equally-favoured

The danger of p-values: coin tossing

$$P(r|p,n) = \frac{n!}{r!(n-r)!}\, p^r (1-p)^{n-r} \qquad \text{likelihood}$$

$$\text{p value} = \sum_{r'=0}^{r} P(r\,|p,n)$$



Take-home messages

- likelihoods get smaller for larger data sets
- absolute value of the likelihood is irrelevant…
- …yet p-value depend on these
- data often very unlikely under true model
- what's relevant is the *ratio* of likelihoods: Bayes factors

Source of most material in this course