Advanced Statistical Topics B

Bayesian statistics

Claus Thorn Ekstrøm
UCPH Biostatistics
ekstrom@sund.ku.dk

Slides @
biostatistics.dk/teaching/advtopicsB/

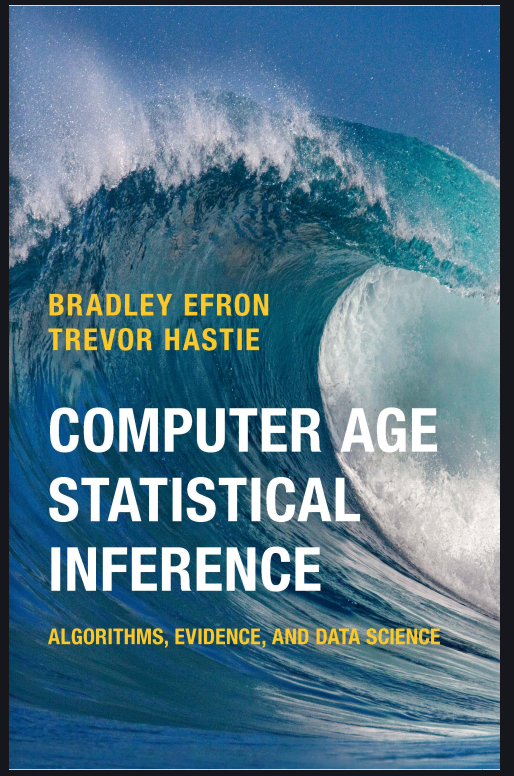
Practicalities

- Every day from **8.15 to 15**
- This year's cock-up
- R
- Additional software extensions
- Breaks/exercises as we go along
- Lunch roughly at 11.
- Please read before classes

Sister course: Advanced Statistical Topics A

Course overview

1. Bayesian Statistics (Monday)
2. Network analysis (Wednesday)
3. Neural networks and deep learning (Tuesday)
4. Principal component analysis (PCA) and partial least squares (PLS) (Thursday)



web.stanford.edu/~hastie/CASI/

Scientific questions

- Is there life on Mars?
- Will it rain tomorrow?
- What is the average height of females?
- What is the effect of this treatment?

What is a probability?

Long run interpretation: $\frac{\#\text{pin down}}{\#\text{throws}}$.

Concepts from frequentist statistics

> *What is the average height of Danish adult females?*

- Parameter estimation
- p values
- Confidence intervals
- *Precise answer to the wrong question*

What is the average female height?

> *I don't know what the mean female height is. However, I know that its value is **fixed** (not a random one). Therefore, I cannot assign probabilities to the mean being equal to a certain value, or being less than/greater than some other value. The most I can do is collect data from a sample of the population and estimate its mean as the value which is most consistent with the data.*

Example: female height

Estimation - maximum likelihood:

$$\hat{\theta} = \arg\max_{\theta} L(\theta; Y)$$

Given the model and the observed data. Find the parameter(s) that would make the likelihood largest.

Note: $L(\theta; Y) \approx P(Y; \theta)$

Sample of $N = 10$ individuals with $\bar{y} = 168.2$ cm and $\text{sd}(y) = 8$ cm.

Typical analysis summary

If we assume that female height can be approximated by a normal distribution then our best estimate for the **population** mean is identical to the **sample** mean, $\hat{\mu} = 168.2$.

We are 95% confident that the interval

$$[\bar{y} \pm 1.96 \times \text{sd}(y)/\sqrt{N}] = [166.63; 169.77]\text{cm}$$

includes the true **population** mean.

What is a p value?

Given a statistical model and an assumption about the world H_0.

Let T_{obs} be a statistic, that summarises the relevant parameter(s) of the data. The p value is defined as

$$P(|T| \geq |T_{\text{obs}}||H_0)$$

IF the null hypothesis is correct *and* **IF** the statistical model is correct then what is the probability to observe a test statistic that is *further* away from H_0 than the observed test statistic.

The p value

Recall the research hypothesis:

> *What is the average height of Danish adult females?*

We are not really answering the research question.

We know the estimate result in the sample. But that will vary with each sample ...

What we are doing: **IF** we know the full DGP what might the sample data look like?

Confidence intervals

A 95% confidence interval is *exactly* the range of values that - if they were tested as a null hypothesis - would *not* be rejected.

A combination on sample variation.

Note: this statement is about CIs *in general*. We do not know anything about our particular CI.

Another note: A CI provides the same evidence for/against the null hypothesis for all values in the CI range.

Bayesian analysis

Probability

- A number between 0 and 1 which encompasses my (our?) statement about uncertainty / certainty
- 1 is complete certainty something is the case
- 0 is complete certainty something is *not* the case

It is a **subjective** measure.

Can probabilities be subjective?

If the payoff is $1.00 I would be willing to bet $0.30 that it will rain

The concepts of Bayesian statistics

> *What is the average height of Danish adult females?*

- Start with prior distribution of my certainty of the population value. Sample data and update beliefs to obtain posterior distrib. of belief
- Maximum a posteriori (MAP) estimation (or some others)
- Credibility intervals
- Bayes factors
- *Subjective answer to right question*

What is the average female height?

> *I agree that the mean is a fixed but since it is unknown, I see no problem in representing **my** uncertainty probabilistically. I will do so by defining a probability distribution over the possible values of the mean and use sample data to update this distribution.*

Can probabilities be subjective?

Priors

The prior distribution expresses **my** initial belief about a parameter.

- Prior knowledge. Classical/hierarchical Bayes. Specify prior knowledge.
- Flat. Essentially a uniform or near-uniform distribution. Results in the MLE estimator.
- Empirical bayes. Use data to guess the prior.

The choice of prior will have an impact on the posterior distribution. This impact will diminish with increasing sample size.

Bayes' formula

$$
\begin{aligned}
P(\theta|D) &= \frac{P(D|\theta) \times P(\theta)}{P(D)} \\
&\propto \underbrace{P(D|\theta)}_{\text{Model / DGP}} \times \underbrace{P(\theta)}_{\text{Prior}}
\end{aligned}
$$

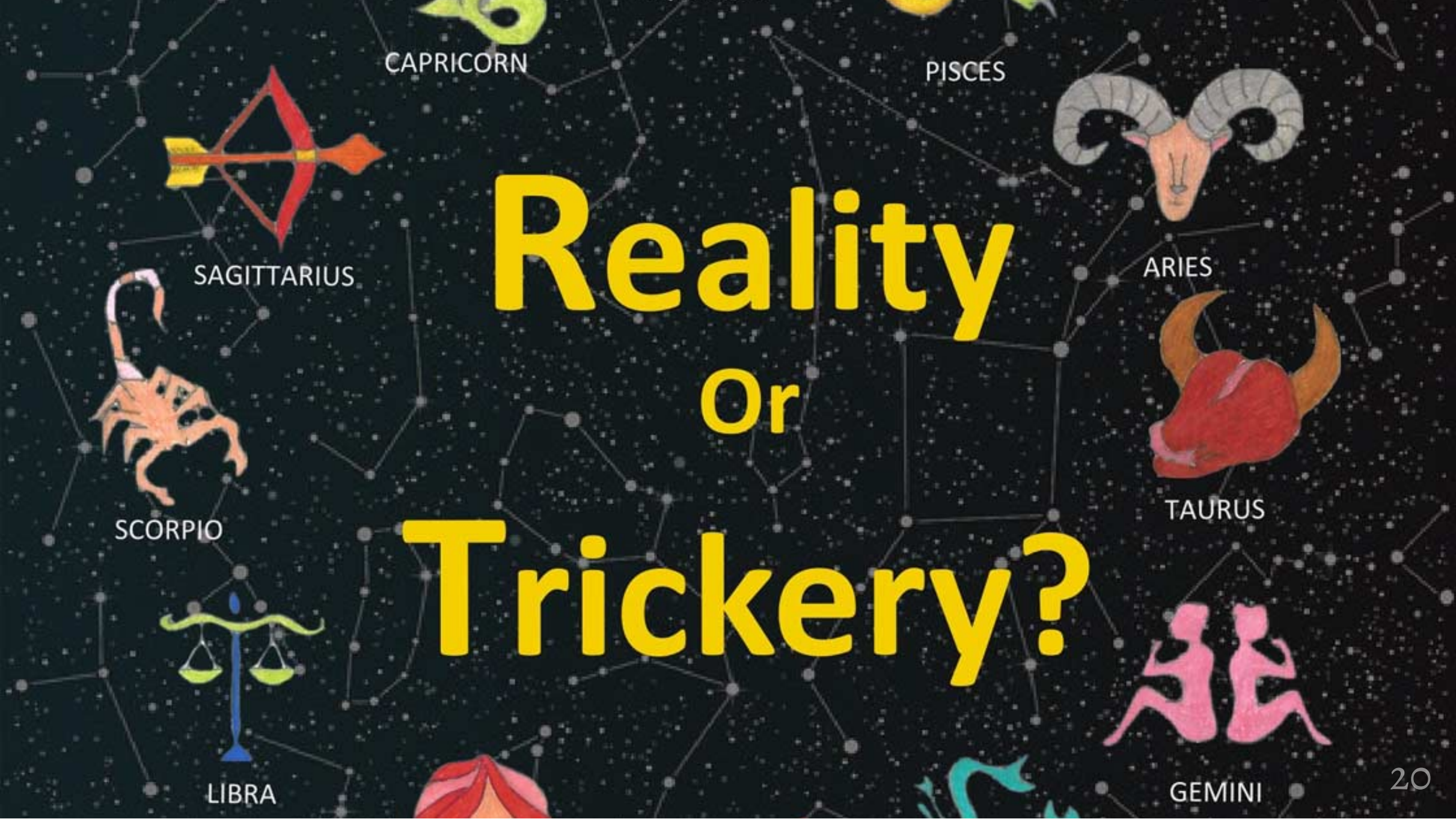
where the marginal likelihood, $P(D)$, in the denominator is the probability of obtaining the data D but without assuming anything about the actual value of θ.

Differences in approaches

In the Bayesian framework we make probability statements about model parameters.

In the frequentist framework, parameters are fixed non-random quantities and the probability statements concern the data.

> Note that credible intervals can be interpreted in the more natural way that there is a probability of 0.95 that the interval contains μ rather than the frequentist conclusion that 95% of such intervals contain μ.


CAPRICORN
PISCES
SAGITTARIUS
ARIES
Reality
Or
Trickery?
SCORPIO
TAURUS
LIBRA
GEMINI

The binomial model

Assumptions

- N independent trials
- Two possible outcomes: succcess and failure
- Same probability of success, θ, in every trial

Estimate:

$$\hat{\theta} = \frac{\#\ \text{Relevant}}{\#\ \text{Possible}}$$

Frequentist analysis

```
prop.test(27, 84, correct=FALSE)
```

```
    1-sample proportions test without continuity
    correction

data:  27 out of 84, null probability 0.5
X-squared = 10.714, df = 1, p-value = 0.001063
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.2312612 0.4272144
sample estimates:
        p
0.3214286
```

Bayesian analysis

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

Naive (but correct) approximation.

1. Assume prior distribution
2. Draw θ from the prior distribution.
3. Draw a sample outcome given θ.
4. If the sample outcome matches the actual outcome then save θ
5. Repeat from 2 a large number of times.

Bayesian analysis

Bayes factors (alternative testing hypotheses)

Bayes factor is defined as the relative likelihood of the data under two different hypotheses. It is defined as:

$$BF = \frac{P(D|H_1)}{P(D|H_2)} = \frac{\frac{P(H_1|D)}{P(H_2|D)}}{\frac{P(H_2)}{P(H_1)}}$$

Independent of sample size, and shows support for the two hypotheses.

Classical hypothesis testing gives the null hypothesis preferred status, and only considers evidence against it.

Wordings

K	Strength of evidence
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

Note: H_1 vs H_2. Otherwise reverse.

Alternatives exist!

Software

- `rstan`. Check out the Stan website. The supreme workhorse
- `brms`. Bayesian Regression Models using Stan
- `rstanarm`. Bayesian Applied Regression Modeling via Stan

Volume of cherry trees

```
data(trees)
```

rstanarm

```
library("rstanarm")
f1 <- stan_glm(Volume ~ Height, data=trees,
               family = gaussian(),
               chains = 4, cores = 2,
               seed = 12, iter = 4000)
```

Default weak priors: $N(0, 10)$ for intercept, $N(0, 5)$ otherwise. Check out `prior`, `prior_intercept` and `prior_aux`.

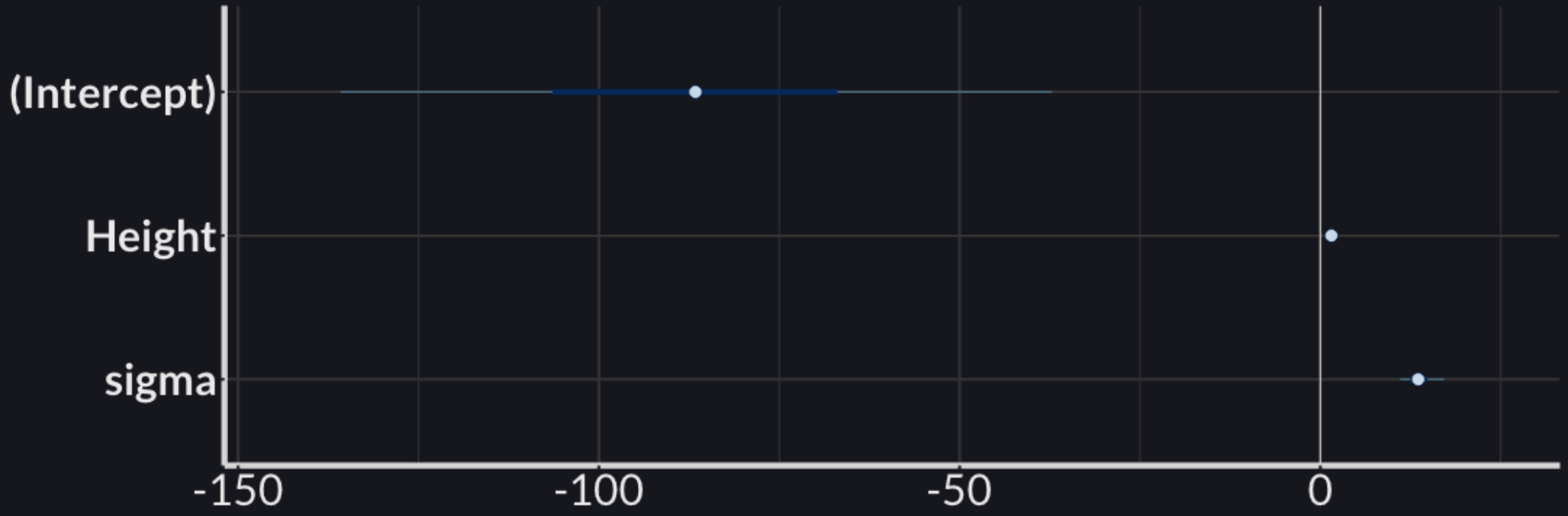
```
f1
```

```
stan_glm
 family:       gaussian [identity]
 formula:      Volume ~ Height
 observations: 31
 predictors:   2
------
            Median MAD_SD
(Intercept) -86.6   29.2
Height        1.5    0.4

Auxiliary parameter(s):
      Median MAD_SD
sigma 13.6    1.8

------
  For help interpreting the printed output see ?print.stanreg
  For info on the priors used see ?prior_summary.stanreg
```

```
plot(f1)
```



50% and 90% Credibility intervals

How does estimation work in practice?

In practice we should do as we did previously. However, it is inefficient as we have seen.

Need to sample from (complex) posterior distribution:

$$P(\alpha, \beta, \sigma^2 | y)$$

Involves multidimensional integral. Sometimes easy. Generally use alternative: sample from conditional posterior distributions.

MCMC methods

Conditional posterior distributions

$$P(\alpha|y,\beta,\sigma^2),\ P(\beta|y,\alpha,\sigma^2),\ P(\sigma^2|y,\alpha,\beta),$$

Markov chain Monte Carlo (MCMC) methods comprise a class of algorithms for sampling from a probability distribution.

Sample from these independently. Consequence: sampling approach generates dependent samples from the joint posterior distribution.

MCMC concepts

- **Chains.** A positive integer specifying the number of Markov chains. The default is 4.
- **Iterations.** A positive integer specifying the number of iterations for each chain (including warmup). The default is 2000.
- **Warm-up.** A positive integer specifying the number of warmup iterations per chain.
- **Thinning.** A positive integer specifying the period for saving samples. The default is 1.

Posterior distribution of parameters

```
posterior <- as.matrix(f1)
head(posterior)
```

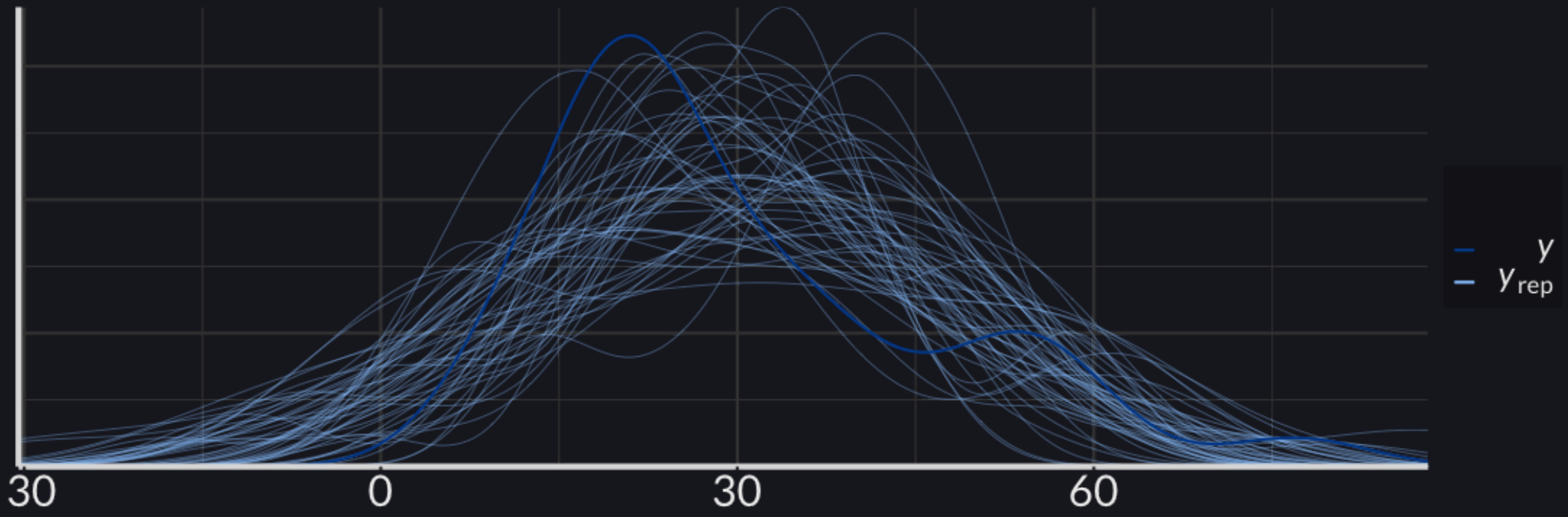
```
           parameters
iterations (Intercept)    Height    sigma
      [1,]   -21.26092 0.7002313 18.88568
      [2,]   -71.00485 1.3821234 15.55591
      [3,]   -62.52229 1.2556886 14.31037
      [4,]   -97.14594 1.6386454 10.82033
      [5,]   -23.87646 0.7047129 12.58199
      [6,]  -102.69781 1.7756832 13.89995
```

Plot distributions

```
library("bayesplot")
mcmc_areas(posterior, pars=c("Height"), prob = 0.8)
```

Sample outcomes from posterior

```
postsamp <- posterior_predict(f1, draws = 500)
color_scheme_set("brightblue")
ppc_dens_overlay(trees$Volume, postsamp[1:50, ])
```



Testing hypotheses about parameters

Define criteria for hypothesis.

Consider the posterior distribution.

```
# Probability that the height regression is
# larger than 2
posterior <- as.matrix(f1)
mean(posterior[,2]>1)
```

```
[1] 0.915375
```

shinystan

```
launch_shinystan(f1)
```

BRMS

```
library("brms")
f2 <- brm(Volume ~ Height + Girth, data=trees, refresh = 0)
```

Model checking

```
f2
```

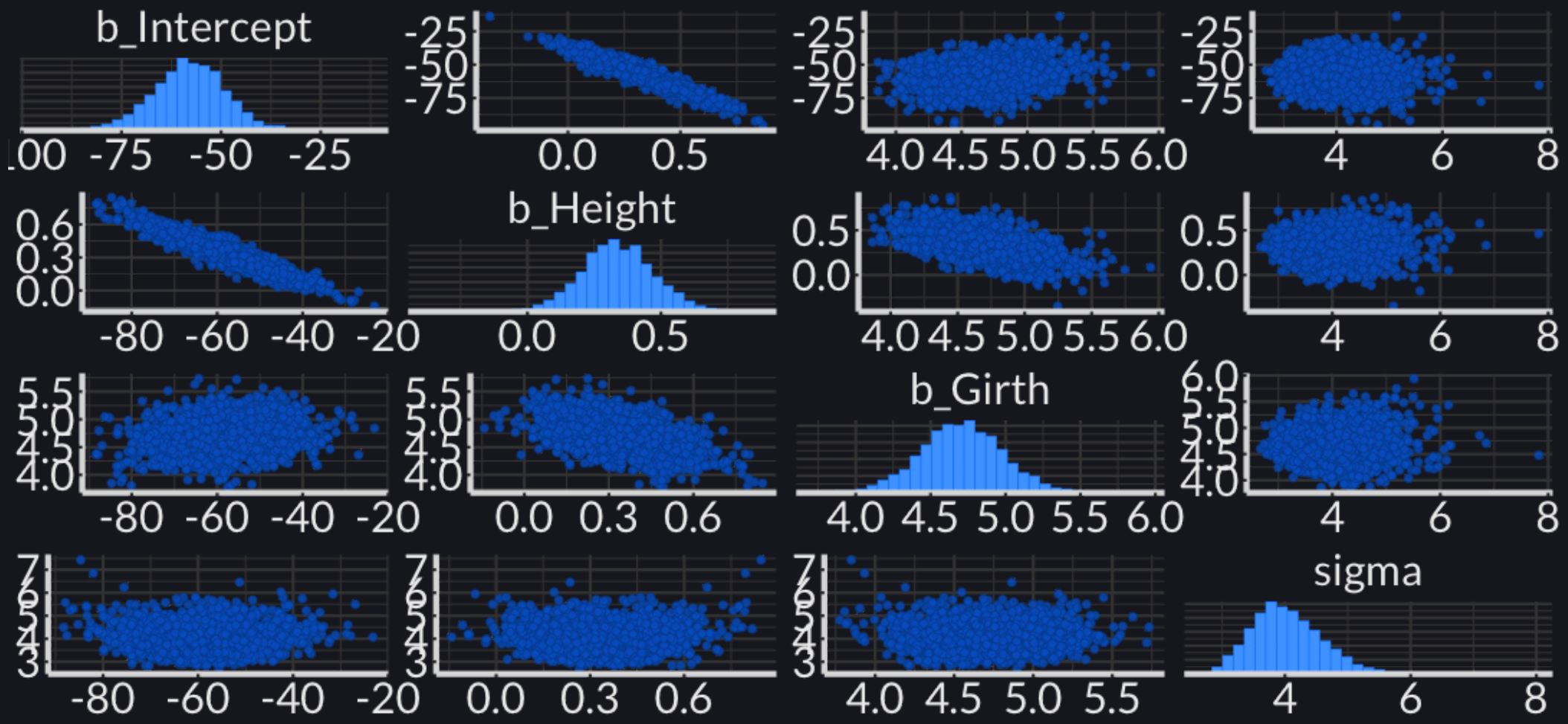
```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: Volume ~ Height + Girth
   Data: trees (Number of observations: 31)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat
Intercept   -58.01      9.24   -76.83   -40.71 1.00
Height        0.34      0.14     0.07     0.62 1.00
Girth         4.71      0.28     4.16     5.26 1.00
          Bulk_ESS Tail_ESS
Intercept     3645     2994
```

```
plot(f2, pars = c("Height", "Girth"))
```

```
Warning: Argument 'pars' is deprecated. Please use
'variable' instead.
```

```
pairs(f2)
```



Marginal effects

```
plot(conditional_effects(f2, effects = "Girth"))
```

stan

```
data {
  int<lower=1> N;
  vector[N] y;
  vector[N] x;
}
parameters {
  real alpha;
  real beta;
  real<lower=0> sigma;      // Note lower limit
}
model {
  alpha ~ normal(0, 10);    // Prior
  beta ~ normal(0, 10);     // distributions
  sigma ~ cauchy(0, 2.5);   // defined here
  y ~ normal(alpha + beta * x, sigma);
}
```

Run model

```
library("rstan")
DF <- list(x=trees$Height, y=trees$Volume, N=nrow(trees))
lm1 <- stan_model("linreg.stan") # Compile the stan program
f1 <- sampling(lm1, iter = 300, data = DF, show_messages=FALSE) # Samp
```

```
SAMPLING FOR MODEL '3692bcd2cb7b25b92b6f645d4a5fcc85' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 9e-06 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:   1 / 300 [  0%]  (Warmup)
Chain 1: Iteration:  30 / 300 [ 10%]  (Warmup)
Chain 1: Iteration:  60 / 300 [ 20%]  (Warmup)
```

Empirical Bayes

A practical challenges is that it requires a statistician to hold some prior belief of the parameter that they are trying to estimate.

What if we only have data and no prior information?

Empricial Bayes approximates hierarchical Bayes by using the data to form our prior *and then* data to form posterior beliefs.

Fast, approximate inference for hierarchical models.

Empirical Bayes

Approximate the marginal of θ using the maximum likelihood estimate (MLE).

Set parameters of prior distribution to obtain this estimate.

The point estimates for the prior (i.e. mean/MAP) will look like a weighted average of the sample estimate and the prior estimate (likewise for estimates of the variance).

(Other versions exist)