UNIVERSITÄT LEIPZIG

European Research Council

Introduction to Bayesian Statistics

Natalia Levshina©

Leipzig University

December 4 2018, Ghent

Course outline

1. Basic concepts of Bayesian statistics

2. A simple illustration: Binomial proportions and online dating

3. Bayesian regression with brms

4. A linguistic illustration: help + (to) Infinitive

Why Bayesian?

- You can the research hypothesis directly, instead of trying to reject the null hypothesis.
  - e.g. How confident can I be that eating too many sweets will lead to diabetes?
  - Compare with the frequentist approach: can I reject the null hypothesis that eating too many sweets will not lead to diabetes?
- No *p*-values and hence no *p*-value hacking!
- Every result is interpretable and useful, not only the 'significant' ones. Good for scientific progress.

Types of probabilities

- p(x) is the probability of event x
  - The probability of getting the bullet when playing Russian roulette
  - The probability of a random person in this room being a linguist, liking beer, crime series, etc.
- p(x, y) is the joint probability that events x and y will happen together
  - E.g. the probability that a random person in this room is a linguist and loves ice-cream; the probability that a random person in this room likes beer and crime series.
- p(x|y) is the conditional probability of event x given event y, i.e. that event x will happen if y happens
  - E.g. The probability of finding an ice-cream fan if one picks a linguist; the probability that a person who likes beer also likes crime series.
  - Can be computed as $p(x|y) = p(x,y)/p(y)$

Bayes' rule

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

The mysterious
Thomas Bayes (1702 – 1761)

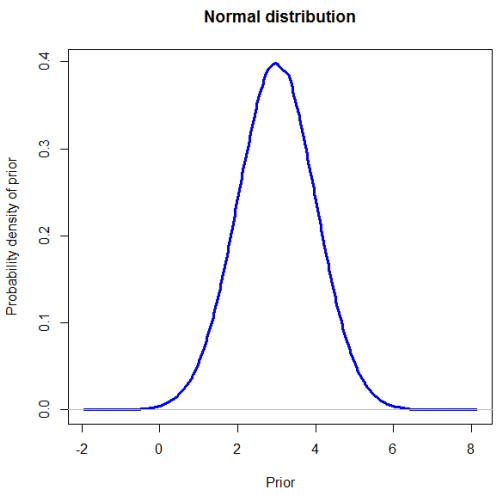
Why would you care?

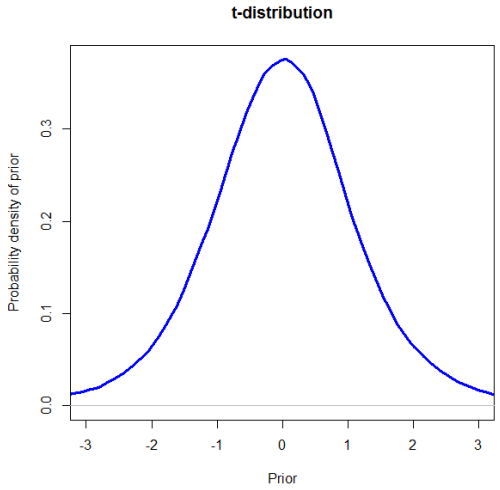
$$\boldsymbol{p\,(beliefs \mid data)} = \frac{\boldsymbol{p\,(data \mid beliefs)\,p\,(beliefs)}}{\boldsymbol{p\,(data)}}$$

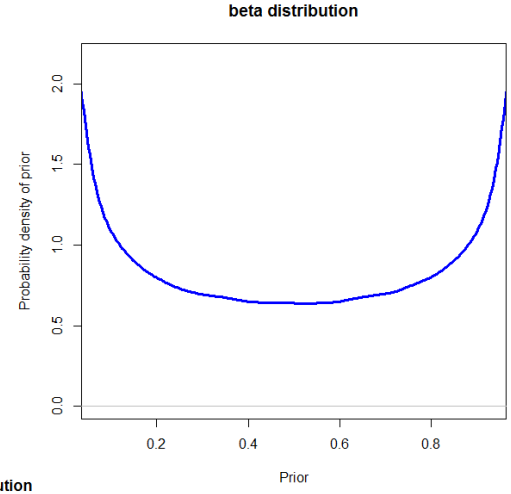
- $p\,(beliefs)$ – prior probability, or just prior
- $p\,(beliefs \mid data)$ – posterior probability, or posterior
- $p\,(data \mid beliefs)$ – likelihood
- $p\,(data)$ – evidence (or prior predictive, or marginal likelihood)

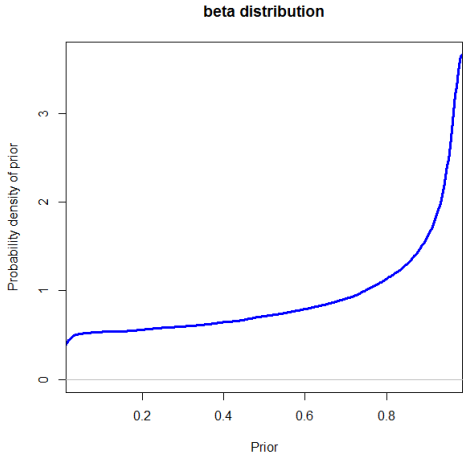
The posteriors express directly the probability of our beliefs given the data!

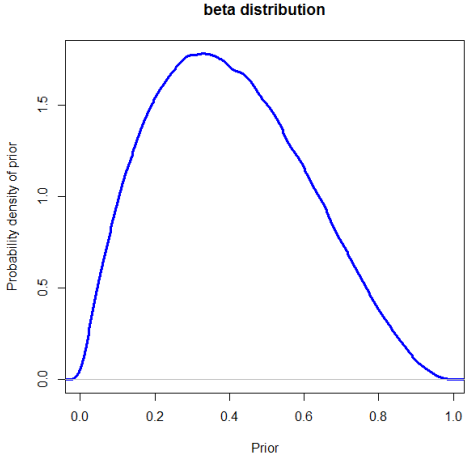
Beliefs as probabilities





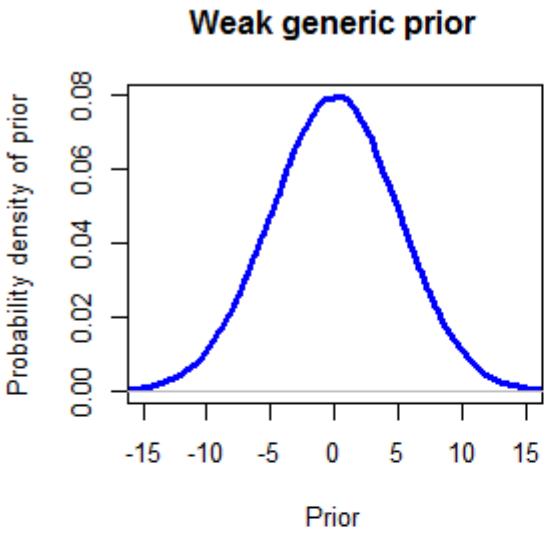


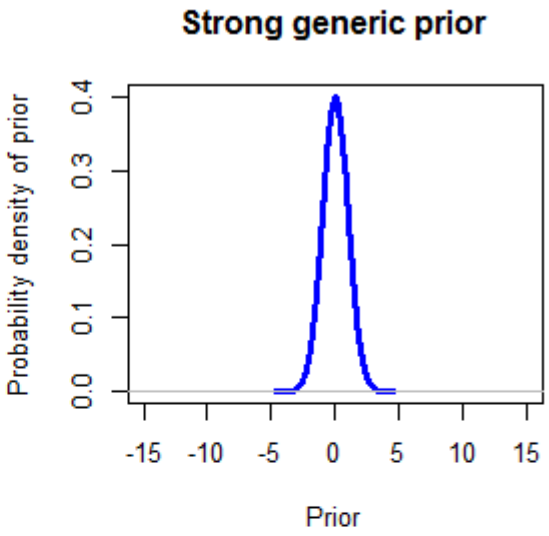


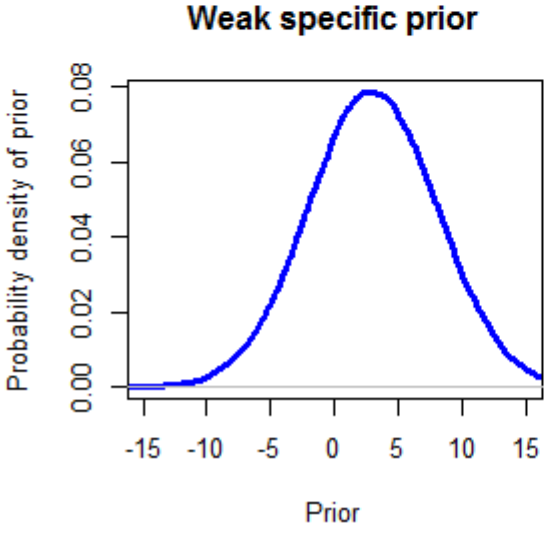


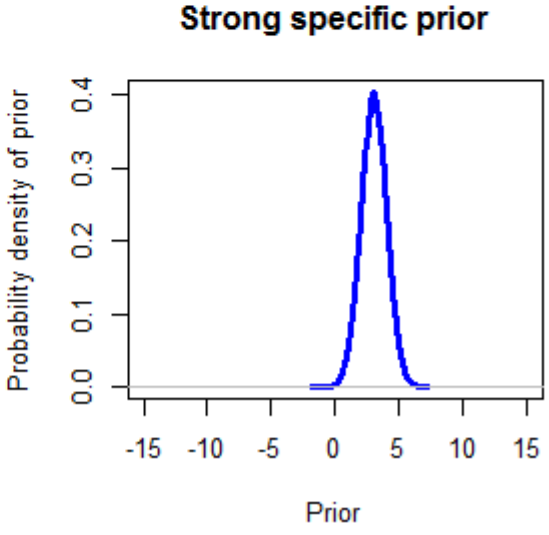
Types of prior distributions

- Flat, or uniform non-informative priors have no impact on the posteriors. All values have the same likelihood. The results are nearly identical to the ones obtained with the help of frequentist methods.
- Informative priors have impact on the posteriors:
  - weak vs. strong (e.g. normal distribution with sd = 10 vs. normal distribution with sd = 1)
  - generic vs. specific (e.g. normal distribution with sd = 5 and mean = 0 vs. normal distribution with sd = 5 and mean = 3).


Weak generic prior
Probability density of prior
0.00
0.02
0.04
0.06
0.08
-15
-10
-5
0
5
10
15
Prior


Strong generic prior
Probability density of prior
0.0
0.1
0.2
0.3
0.4
-15
-10
-5
0
5
10
15
Prior


Weak specific prior
Probability density of prior
0.00
0.02
0.04
0.06
0.08
-15
-10
-5
0
5
10
15
Prior


Strong specific prior
Probability density of prior
0.0
0.1
0.2
0.3
0.4
-15
-10
-5
0
5
10
15
Prior

Probability distributions

- What are your prior beliefs about a 10 cent coin? Is it fair?
- What are your prior beliefs about the heights of Belgian men?
- About the IQ of the world's population?
- About word frequencies in a corpus?
- About the probabilities (from 0 to 1) generated by a random probabilities generator for internet poker?

Posterior probabilities

- Posteriors are the conditional probabilities of a parameter (or model) after the relevant evidence (data) has been taken into account.

- Posteriors are probability distributions, too.

- One can use 95% Highest Density Intervals / Credible Intervals for inference, where 95% of the posterior distribution lies. Do not confuse them with confidence intervals in frequentist statistics!

Course outline

1. Basic concepts of Bayesian statistics

2. A simple illustration: Binomial proportions and online dating

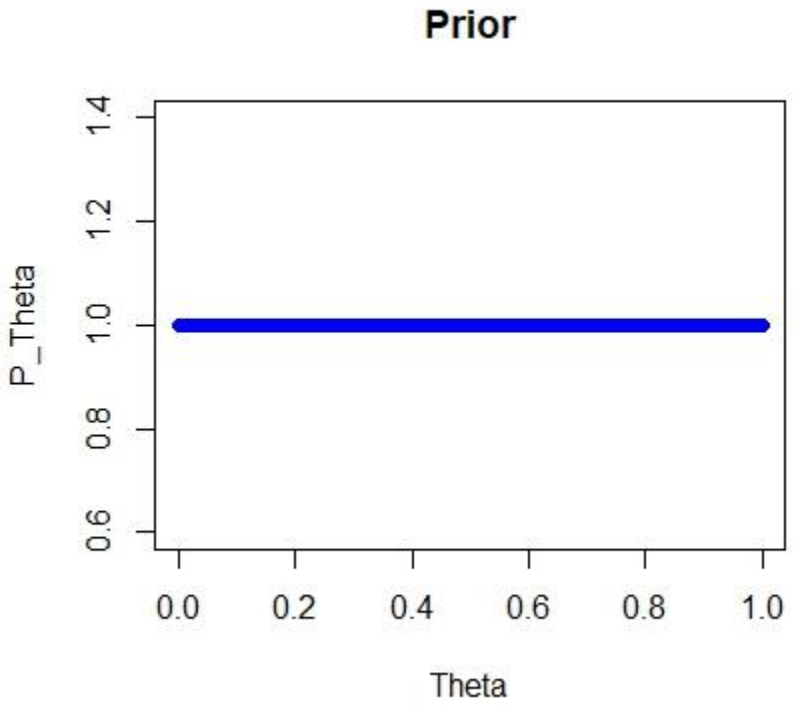
3. Bayesian regression with brms

4. A linguistic illustration: help + (to) Infinitive

Success or failure?

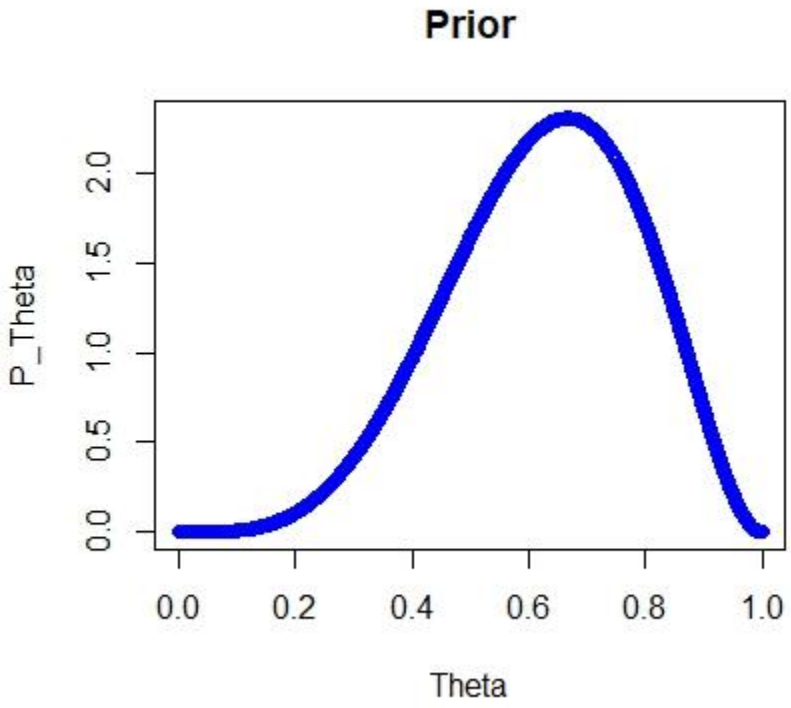
- Imagine a user of an online dating app. The user sends invitations to the persons he or she is interested in. They may reply (‘success’) or not reply (‘failure’).
- This represents a binomial distribution with a certain number of trials, a number of successes and a number of failures (similar to a toss of a coin). For example, there are 10 trials, 6 successes and 4 failures.
- Everybody has some prior expectations about his or her attraction powers, based on their previous romantic experience.
- But the real world experience may change them -> posterior beliefs.

Prior beliefs: uniform



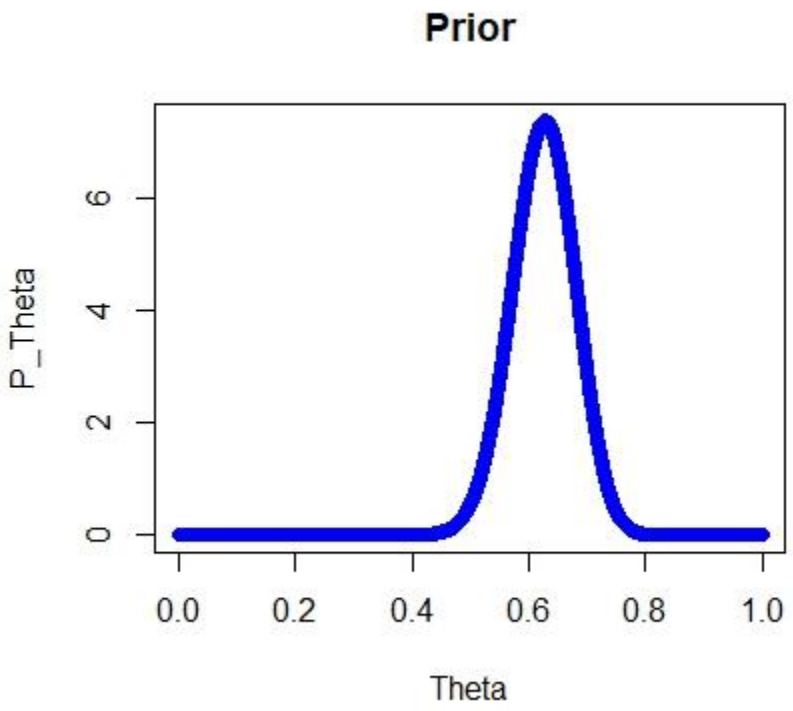
I have no clue how attractive I am.

Prior beliefs: cautiously optimistic



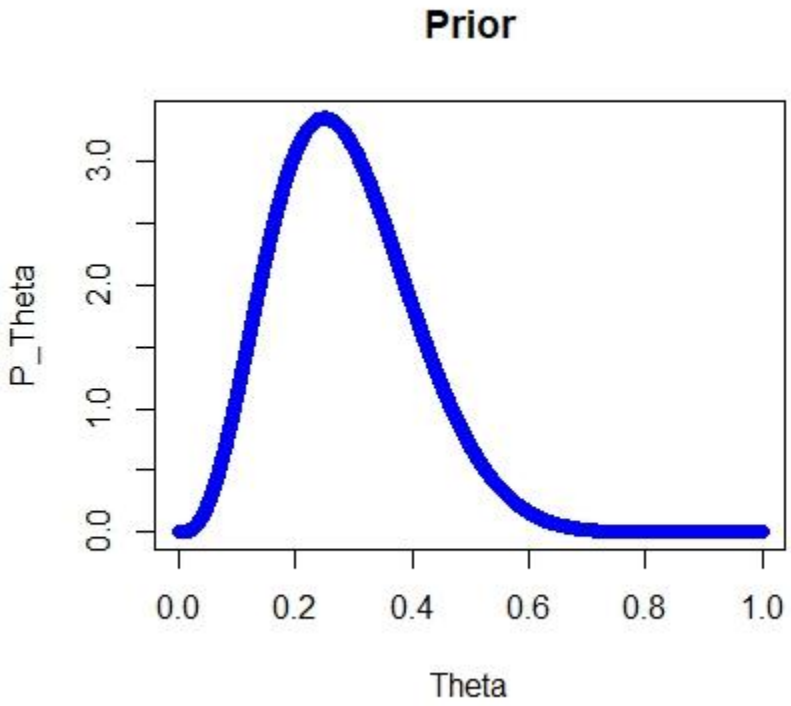
I think I'm rather attractive…

Prior beliefs: experienced optimistic user



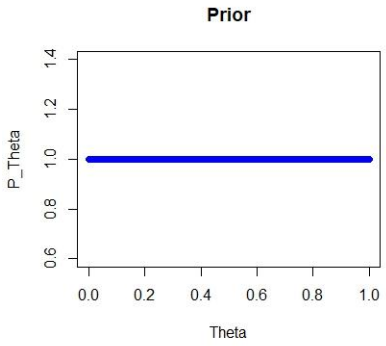
I know my chances very well.

Prior beliefs: pessimistic user

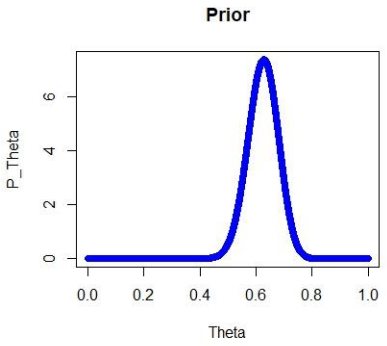


I don’t expect much…

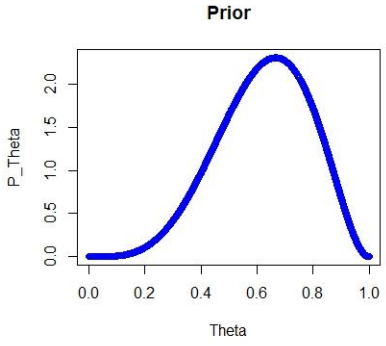
Beta distributions



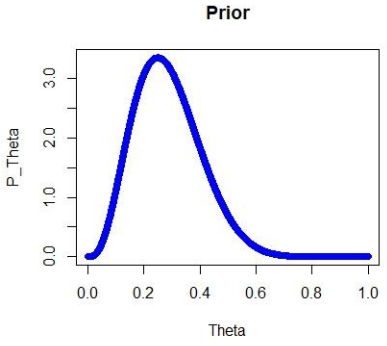
a = 1, b = 1



a = 50, b = 30



a = 5, b = 3



a = 4, b = 10

Beta distributions in R

- E.g. for the uniform prior:

```
a <- 1
b <- 1

Theta <- seq(from = 0, to = 1, by = 0.0001)

P_Theta <- dbeta(Theta, a, b)

plot(Theta, P_Theta, main = "Prior", col =
"blue")
```

Exercise

- Create a beta distribution that would reflect your prior beliefs whether or not you would understand Bayesian statistics.

OR

- Create a beta distribution that would reflect your prior beliefs whether or not it will rain on Christmas eve.

Data: first experience with the dating app

- The user sent 5 invitations and received 4 replies. One person did not reply.

```
Success <- 4
Failure <- 1
Total_N <- Success + Failure
```

How to compute the posteriors

- The posteriors are influences both by the priors and the data.

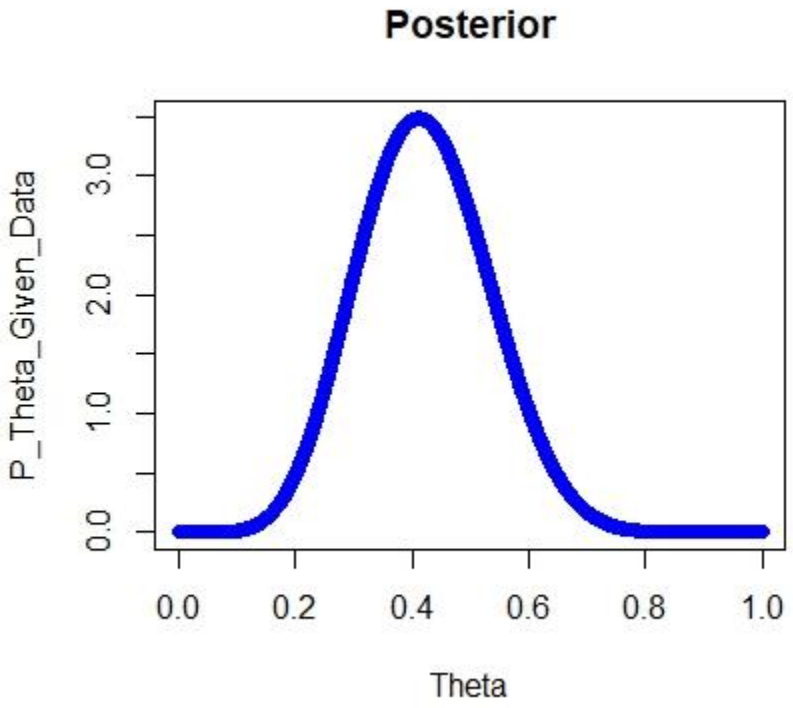
- If the priors were of the type “pessimistic user”, i.e.

```
a <- 4
b <- 10
```

- … then the probabilities of theta given the data are computed as follows:

```
P_Theta_Given_Data <- dbeta(Theta, a + Success, b
+ Failure)
plot(Theta, P_Theta_Given_Data, main =
"Posterior", col = "blue")
```

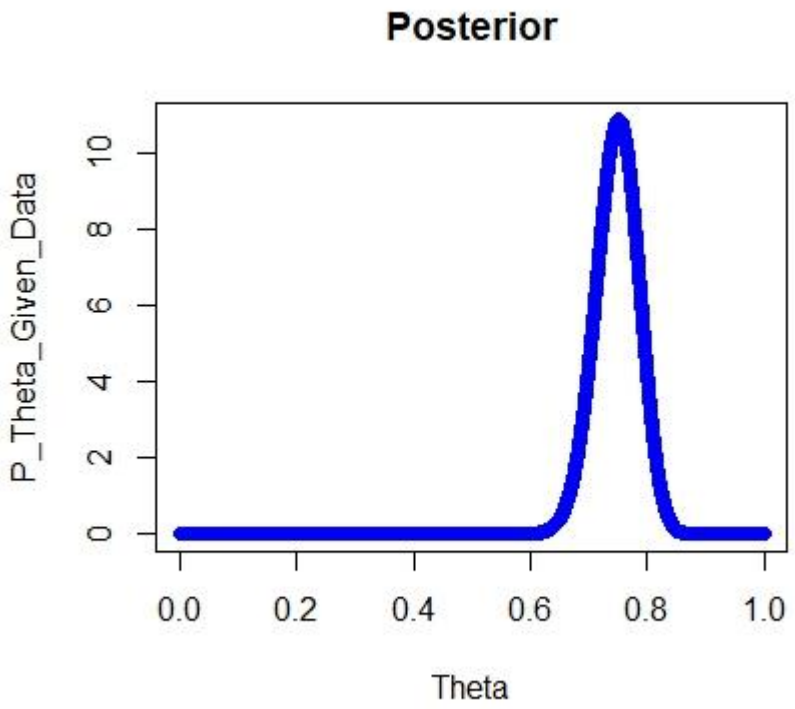
Posteriors: first encouraging experience…



Data: more experience, the same ratio

```
Success <- 100
Failure <- 25
```

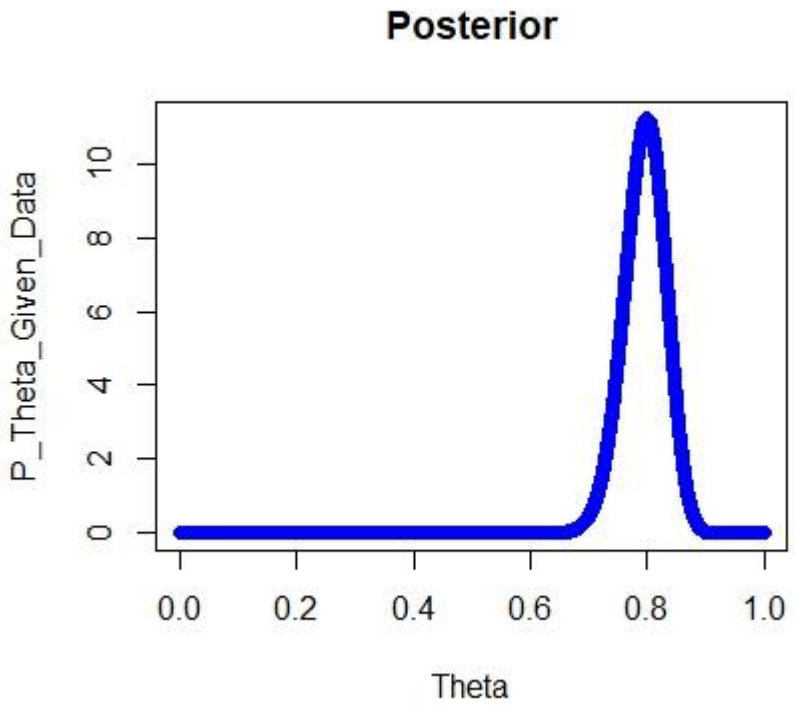
What a change!



Compare with the uniform prior

```
a <- 1
b <- 1
```

Important: with uniform priors, only the data influence the posteriors!

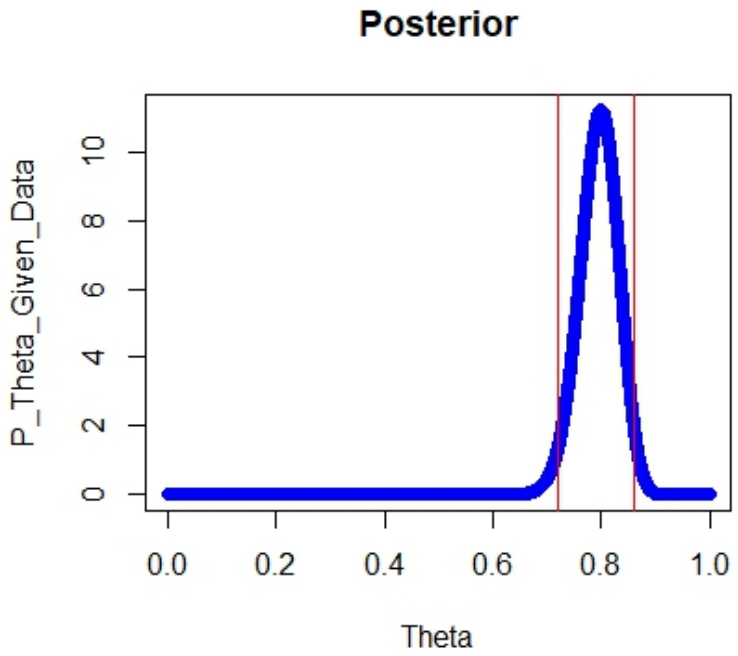


Exercise

- Try some more combinations of the priors and the data and see how the posteriors change. What are your conclusions…
  - about the influence of data size?
  - about the influence of different types of priors?

95% credible intervals

```
abline(v = qbeta(0.975, a + Success, b +
Failure), col = "red")
abline(v = qbeta(0.025, a + Success, b +
Failure), col = "red")
```



Course outline

1. Basic concepts of Bayesian statistics

2. A simple illustration: Binomial proportions and online dating

3. Bayesian regression with brms

4. A linguistic illustration: help + (to) Infinitive

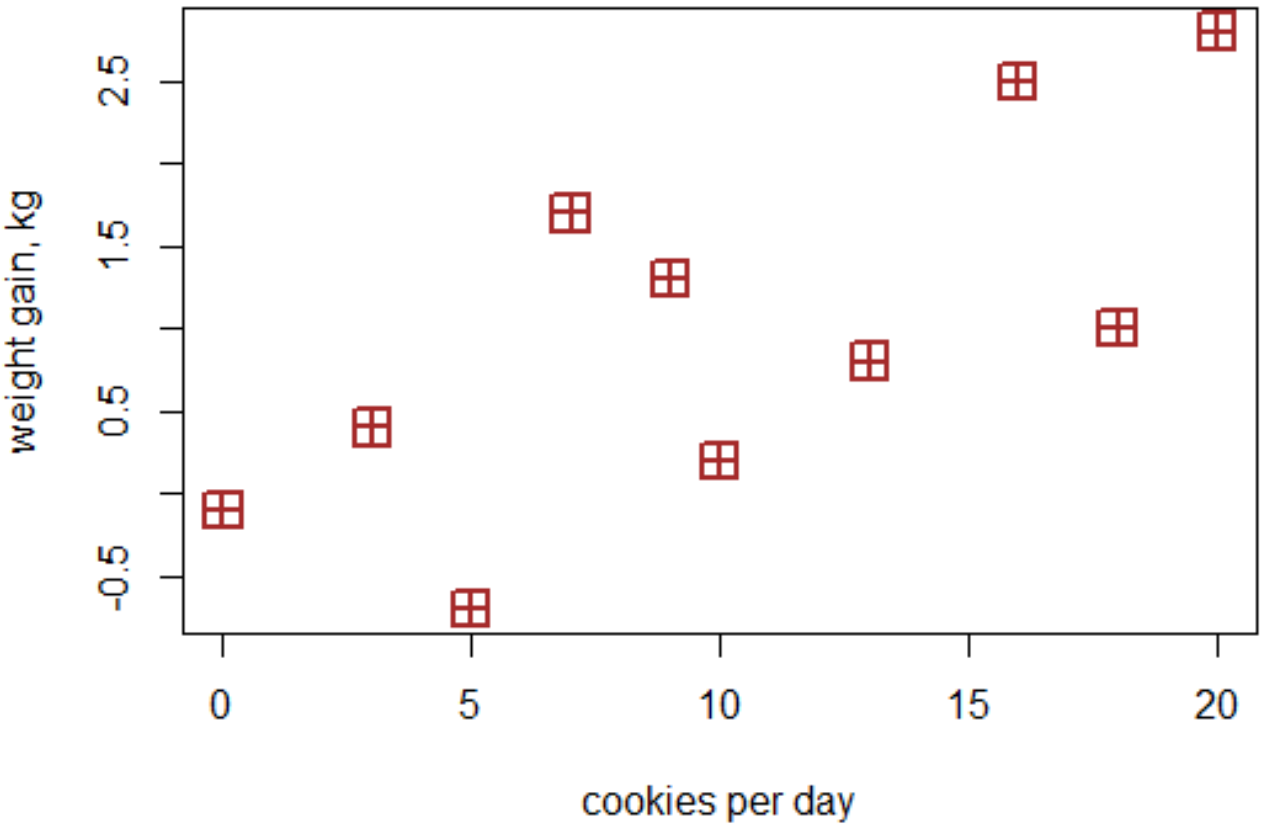
Before we start…

- Think about the following question: What is the effect of eating cookies on one's weight? In other words, what will happen to your weight if you eat more cookies per day? Fewer cookies?

- Write down your hypothesis: The more cookies one eats, …

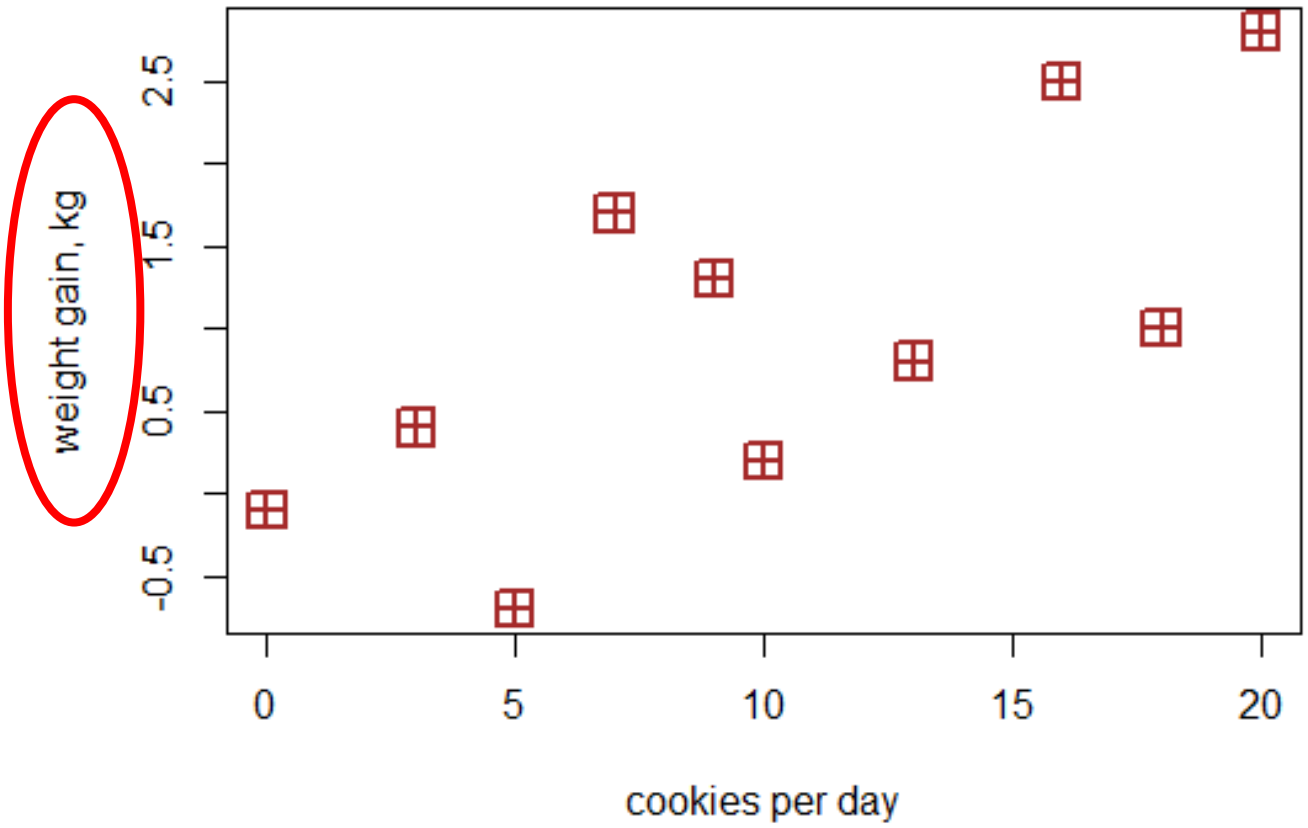
A cookie diet

Name	Cookies eaten per day	Kilos gained
John	0	-0.1
Mary	3	0.4
Bill	5	-0.7
Jane	7	1.7
Laura	9	1.3
Ann	10	0.2
Chris	13	0.8
Eve	16	2.5
Peter	18	1.0
Steve	20	2.8


weight gain, kg
-0.5
0.5
1.5
2.5
0
5
10
15
20
cookies per day

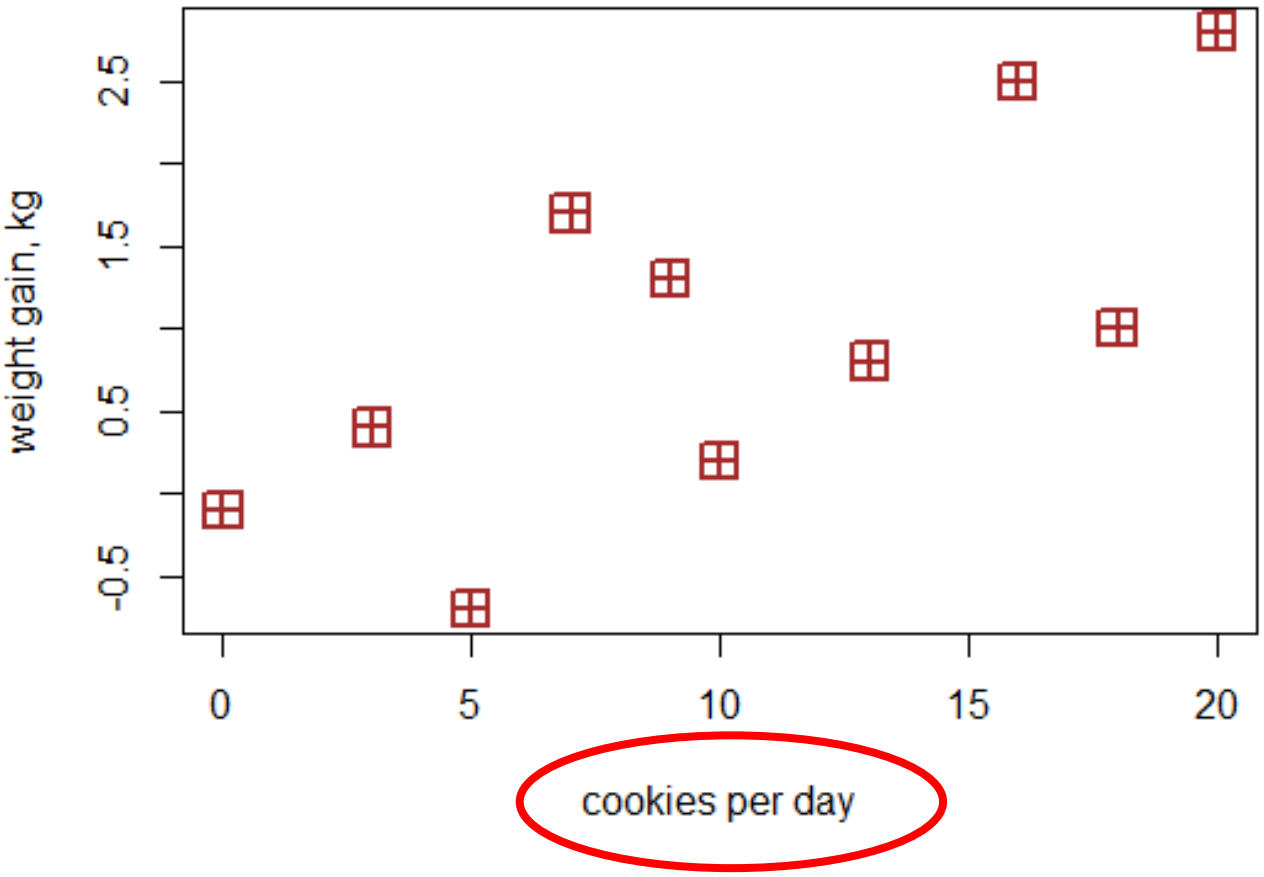
Fundamental concepts of regression

- Dependent variable (response): weight gain


weight gain, kg
-0.5
0.5
1.5
2.5
0
5
10
15
20
cookies per day

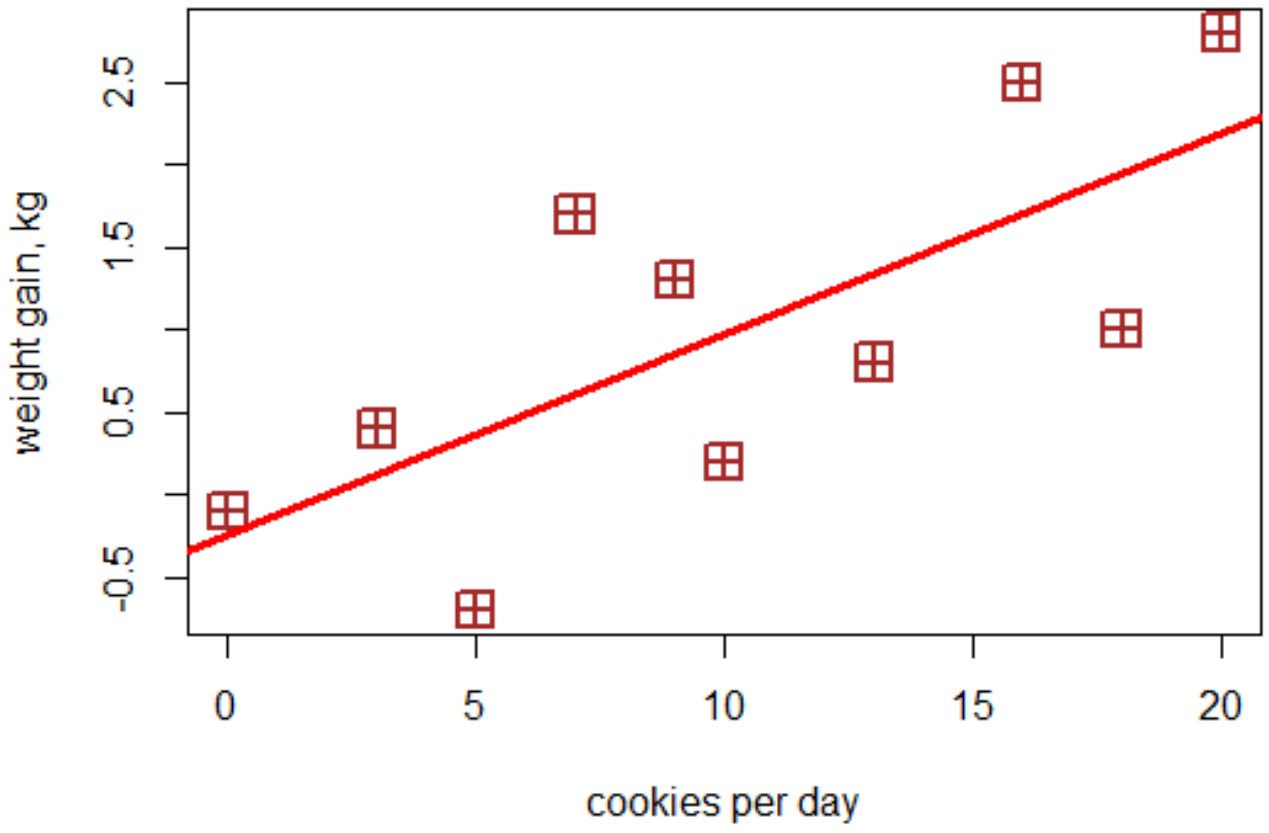
Fundamental concepts of regression

- Dependent variable (response): weight gain
- Independent variable (predictor): cookies


weight gain, kg
-0.5
0.5
1.5
2.5
0
5
10
15
20
cookies per day

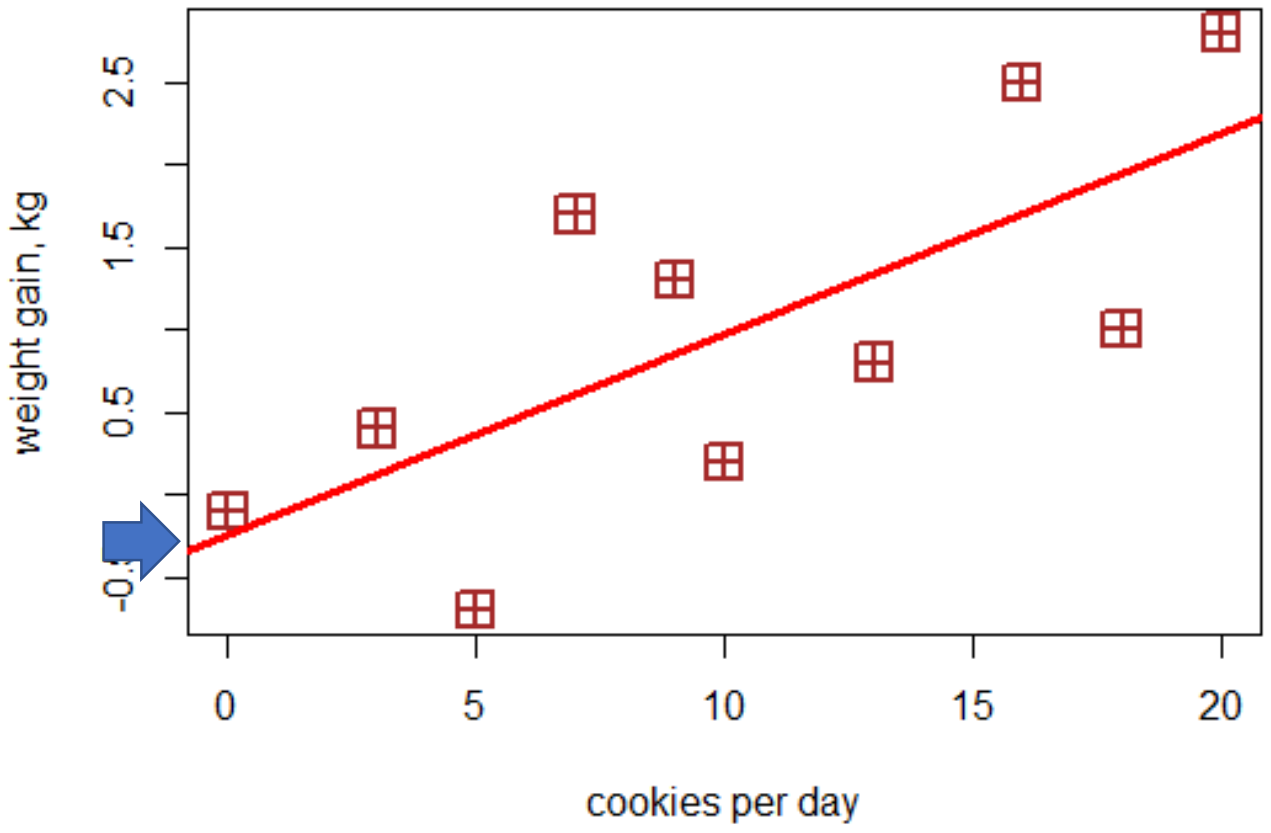
Fundamental concepts of regression

- Dependent variable (response): weight gain
- Independent variable (predictor): cookies
- Regression line: line that can be drawn through the cloud of points such as the distances between the line and all points are as small as possible


weight gain, kg
-0.5
0.5
1.5
2.5
0
5
10
15
20
cookies per day

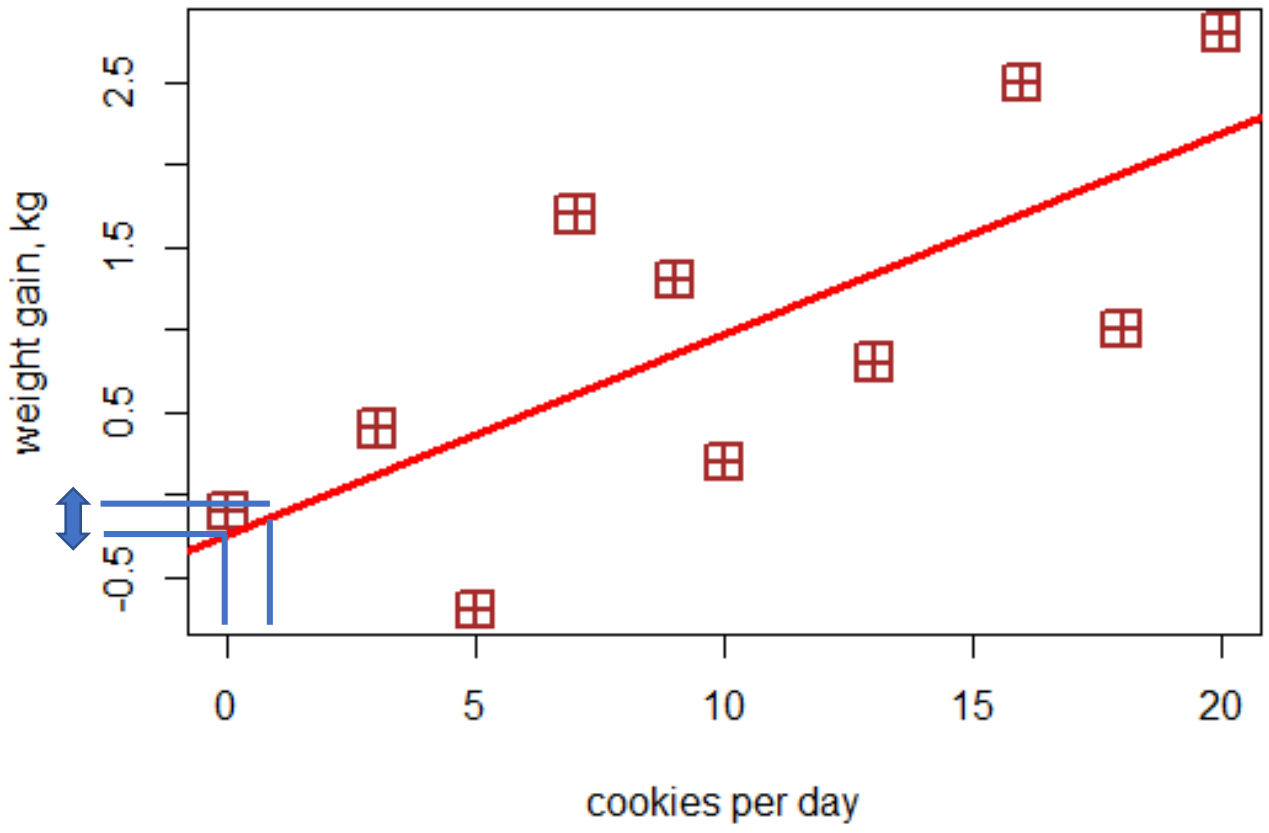
Fundamental concepts of regression

- Dependent variable (response): weight gain
- Independent variable (predictor): cookies
- Regression line: line that can be drawn through the cloud of points such as the distances between the line and all points are as small as possible
- Intercept: value of y where the line crosses the y-axis


weight gain, kg
2.5
1.5
0.5
0
5
10
15
20
cookies per day

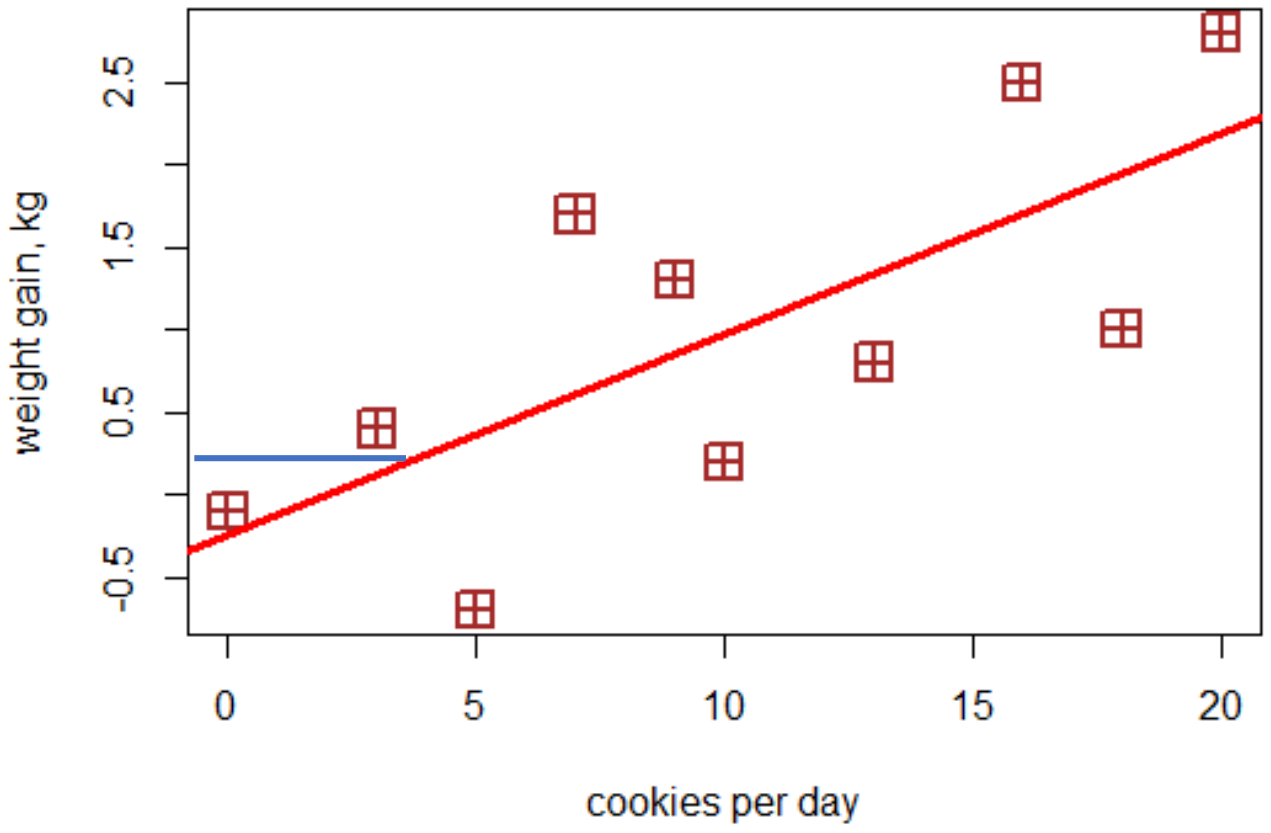
Fundamental concepts of regression

- Dependent variable (response): weight gain
- Independent variable (predictor): cookies
- Regression line: line that can be drawn through the cloud of points such as the distances between the line and all points are as small as possible
- Intercept: value of y where the line crosses the y-axis
- Slope: increase of y per unit of x on the line


weight gain, kg
-0.5
0.5
1.5
2.5
0
5
10
15
20
cookies per day

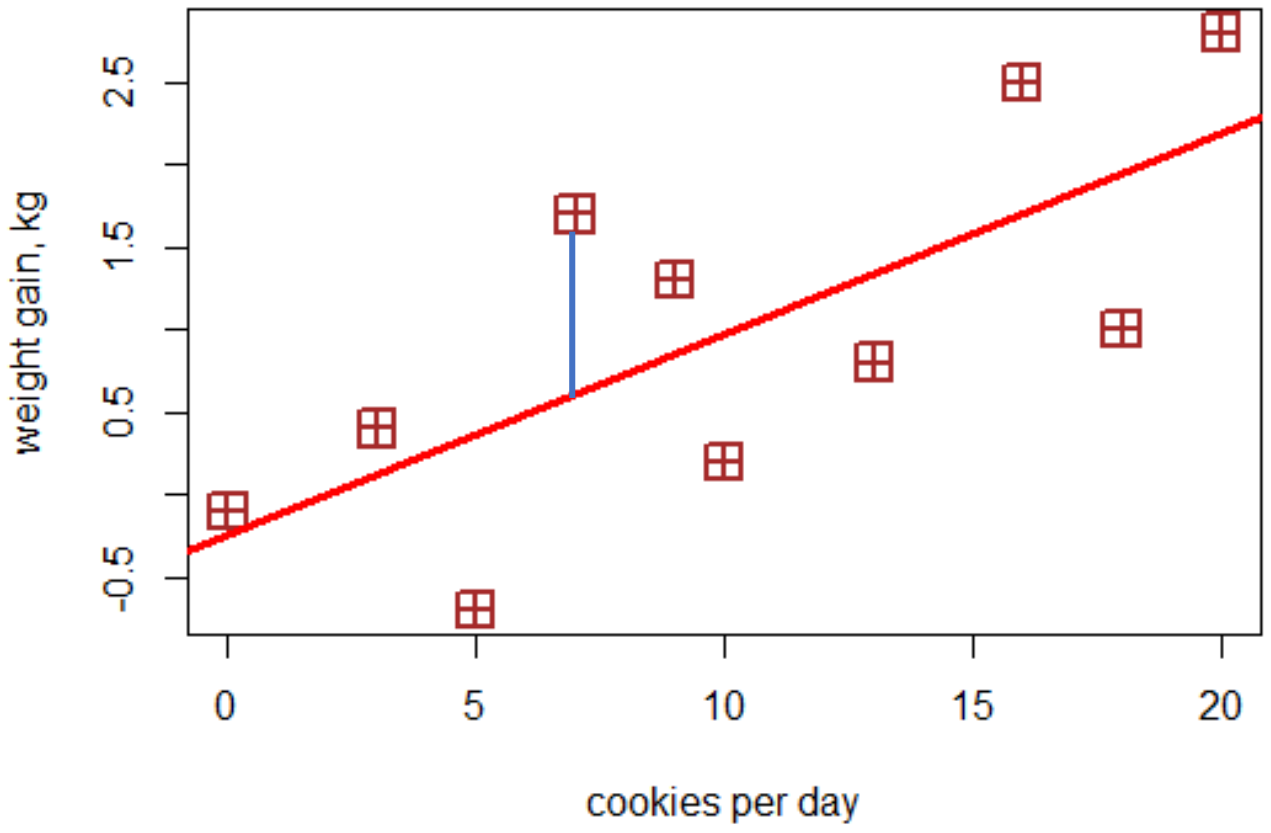
Fundamental concepts of regression

- Dependent variable (response): weight gain
- Independent variable (predictor): cookies
- Regression line: line that can be drawn through the cloud of points such as the distances between the line and all points are as small as possible
- Intercept: value of y where the line crosses the y-axis
- Slope: increase of y per unit of x
- Fitted values: the y-coordinates of the projections of the points on the line


weight gain, kg
-0.5
0.5
1.5
2.5
0
5
10
15
20
cookies per day

Fundamental concepts of regression

- Dependent variable (response): weight gain
- Independent variable (predictor): cookies
- Regression line: line that can be drawn through the cloud of points such as the distances between the line and all points are as small as possible
- Intercept: value of y where the line crosses the y-axis
- Slope: increase of y per unit of x
- Fitted values: the y-coordinates of the projections of the points on the line
- Residuals: the differences between the observed and fitted response values (the distances between the points and their projections on the line)


weight gain, kg
-0.5
0.5
1.5
2.5
0
5
10
15
20
cookies per day

The magic of linear regression

Observed value of **y** =

Intercept **α**

+

Slope **β***value of **x**

+

Residual **ε**

} Fitted value **$\bar{y}$**

$$y = \alpha + \beta x + \varepsilon = \bar{y} + \varepsilon$$

Exercise

- Your colleague fitted a regression model which shows that happiness measured on a scale from 0 to 100 depends on Belgian chocolate (in grams).
- The formula looks as follows:

$$\text{Happiness} = 38 + 0.5 * \text{Chocolate} + \text{Error}$$

- How can you interpret these numbers?
- How happy will you be if you eat 50 grams, as predicted by the model? If you eat 100 grams?

Enter the data

```
subjects <- c("John", "Mary", "Bill", "Jane",
"Laura", "Ann", "Chris", "Eve", "Peter",
"Steve")
cookies <- c(0, 3, 5, 7, 9, 10, 13, 16, 18, 20)
kilos <- c(-0.1, 0.4, -0.7, 1.7, 1.3, 0.2, 0.8,
2.5, 1, 2.8)
```

```
mydata <- data.frame(kilos, cookies)
```

Simple linear regression in R

```
lm(kilos ~ cookies)

Call:
lm(formula = kilos ~ cookies)

Coefficients:
(Intercept)      cookies
    -0.2364       0.1214
```

Simple linear regression in R: Summary

```
summary(lm(kilos ~ cookies))
Call:
lm(formula = kilos ~ cookies)
Residuals:
    Min      1Q  Median      3Q     Max
-1.0707 -0.7189  0.2043  0.5668  1.0864
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.23645    0.49341  -0.479   0.6446
cookies      0.12143    0.04151   2.925   0.0191 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.8228 on 8 degrees of freedom
Multiple R-squared:  0.5169,  Adjusted R-squared:  0.4565
F-statistic: 8.558 on 1 and 8 DF,  p-value: 0.01913
```

Bayesian regression in R: some examples

- brms (a wrapper for Stan)
- rstan (for advanced, requires programming in Stan)
- MCMCglmm (logistic regression was a bit tricky, in my experience)
- arm
- blme

Bayesian linear regression

```
library(brms)
mybrm <- brm(kilos ~ cookies, data = mydata)
```

```
Compiling the C++ model
SAMPLING FOR MODEL 'bb9e16ac0c58e0aeba05bd726f2e6628'
NOW (CHAIN 1).
…
Chain 1: Adjust your expectations accordingly!
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
…
```

What is going on?

Markov Chain Monte Carlo

- We cannot always derive the posterior distribution mathematically, as was the case with beta distributions (the dating example).
- Instead, one can get the posterior distribution by sampling a large number of representative points from the posterior distribution with the help of a Markov Chain Monte Carlo algorithm
  - Monte Carlo simulation: any simulation that draws random values from a distribution, e.g. `rnorm()`, `rt()` in R.
  - Markov Chain process: a random walk when the next step does not depend on the steps before the current position.

Metropolis algorithm

- President X of country Y (no names!) wants to sell weapons in order to promote peace in the world.
- Goes to a rich country Z to negotiate the deals with several sheikhs.
- Some of them have more money, some have less.
- Obviously, the more money, the more X should be interested!
- But X doesn’t know how much money each has (he doesn’t know much, in general).

The Sheikhs’palaces

This consumption is very conspicuous: the more money, the more splendid the palace.

10bn

20bn

30bn

40bn

50bn

60bn

The random walk

- Imagine the president arrives first in Sheikh 3's palace. Let's call him the Current Sheikh (CS). Flips a coin if he should go to the sheikh on the left or to the one on the right. The Sheikh selected in this process is called the Proposed Sheikh (PS).
- X doesn't know how much money each sheikh has, but he is not totally stupid. He can look from the window of the CS's palace and compare the sizes of the palaces, computing the ratio PS/CS with the help of a top secret supercomputer.
- If the PS's palace is larger and more lavish, PS/CS > 1 and the president goes to the PS.

The random walk (continued)

- If the PS's palace is smaller and more modest, $0 < PS/CS < 1$, then X generates randomly the probability from 0 to 1 using a top secret random probability generator.
- If the randomly generated probability is less than the PS/CS ratio, then he moves to the PS. If the probability is greater, then he stays, and the steps are repeated again.
- An amazing fact: In the long run, the frequency of stays at each palace will approximate the sheikhs' relative wealth!

Some implications for Bayesians

- If some value (e.g. the wealth of the current sheikh) is very large, the algorithm may get stuck at it and not traverse the space quickly enough. One should use diagnostic plots for that purpose.
- The results of the first walks are usually excluded (aka burn-in, or warm-up period) because the initial position can introduce substantial bias.
- It’s recommended to have several Markov chains and check if they behave similarly.

How to control Markov chains and warm-up

```
chains = 2 #to speed up the things a little bit, 4 Markov chains by default
iter = 500 #number of iterations in the random walk, 2000 by default. Again, we just want to speed the things up.
warmup = 200 #default first half, i.e. 250 with 500 iterations
```

Experiment with different settings

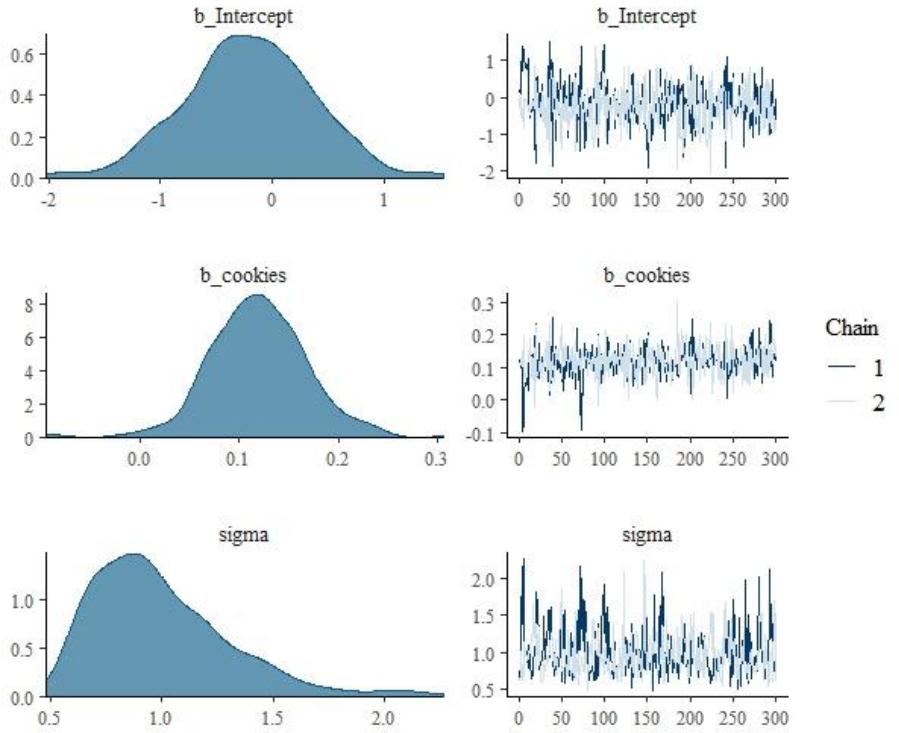
```
mybrm1 <- brm(kilos ~ cookies, data = mydata,
chains = 2, iter = 500, warmup = 200)
```

Density and trace plots in brms

```
plot(mybrm1)
```

The density plots should not be bimodal (with two humps).

The trace plots should look like fat hairy caterpillars, not bending anywhere.



Effective sample size and r-hat

- Effective sample size: the number of posteriors in Markov chains discounted for autocorrelation between them (when the chain gets stuck). The greater effective sample size, the more reliable the results.
- R-hat metric: the ratio of the between-chain and the within-chain variability of posteriors. If the chains have converged, these measures will be similar; otherwise, the between-chain variability will be larger. R-hat should be 1.

Effective sample size and rhat in brms

summary(mybrm1)

```
Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI
Intercept    -0.21      0.58    -1.34     0.86
cookies       0.12      0.05     0.03     0.22
          Eff.Sample Rhat
Intercept        348 1.01
cookies          416 1.01

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample
sigma     1.00      0.31     0.58     1.79        311
      Rhat
sigma 1.00
```

…

Back to the default settings

```
mybrm <- brm(kilos ~ cookies, data = mydata)
plot(mybrm)
```

The rhat indices are also better!

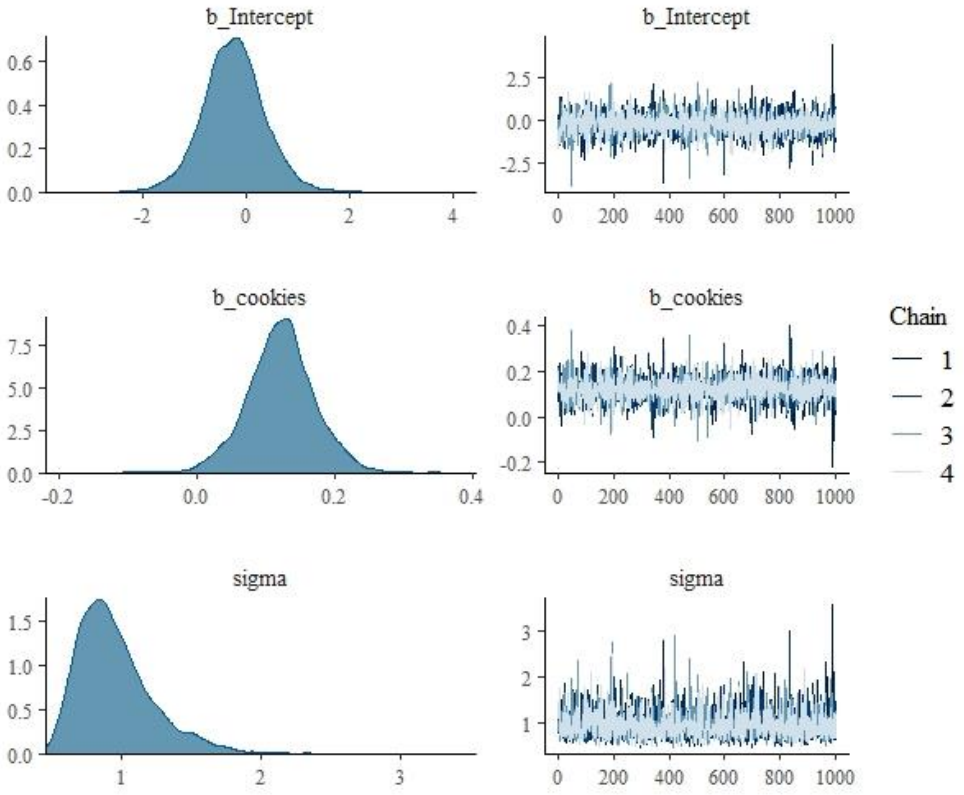


Table of coefficients in the summary

`summary(mybrm)`

…

```
Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept    -0.24      0.62    -1.44     0.98        2318 1.00
cookies       0.12      0.05     0.02     0.22        2479 1.00
```

…

Interpreting the summary output

- Estimate: the mean of the posterior distribution
- Est.error: standard error of the posterior distribution
- l-95% CI: the lower boundary of the 95% credible interval
- u-95% CI: the upper boundary of the 95% credible interval

Posterior distribution

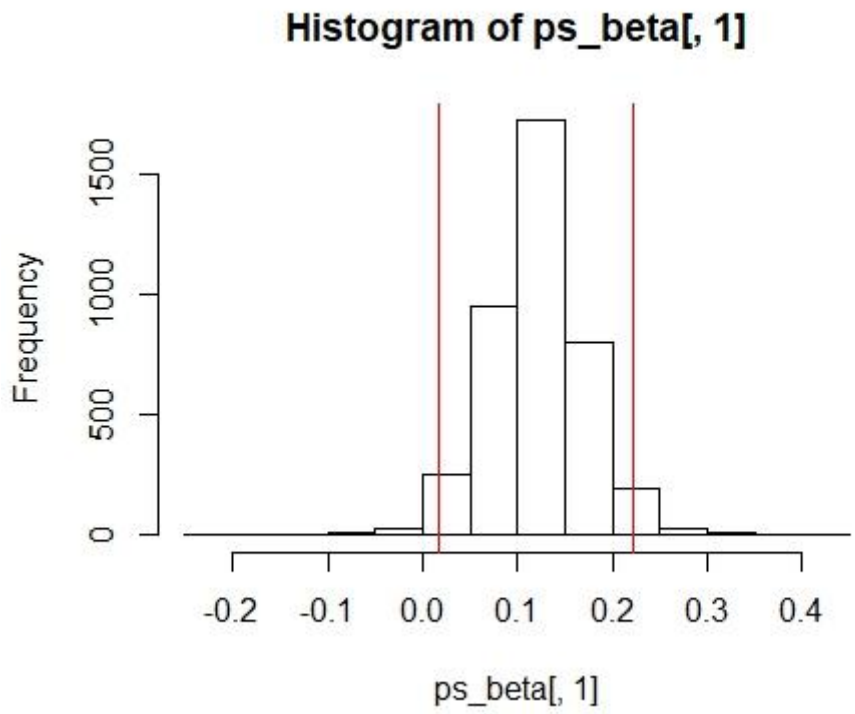
```
ps_beta <- posterior_samples(mybrm, pars =
"cookies")

dim(ps_beta)
[1] 4000    1
summary(ps_beta)
 b_cookies
 Min.   :-0.21978
 1st Qu.: 0.09086
 Median : 0.12186
 Mean   : 0.12142
 3rd Qu.: 0.15130
 Max.   : 0.40287
```

compare with the table!

Posterior distribution and 95% CI

```
hist(ps_beta[, 1])
abline(v = quantile(ps_beta[, 1], 0.025), col
= "red")
abline(v = quantile(ps_beta[, 1], 0.975), col
= "red")
```

Posterior distribution and 95% CI



Testing the research hypothesis

Our expectation was that the effect of cookies on weight is positive: the more cookies, the more weight.

Now we can easily compute the proportion of all posterior estimates that are positive:

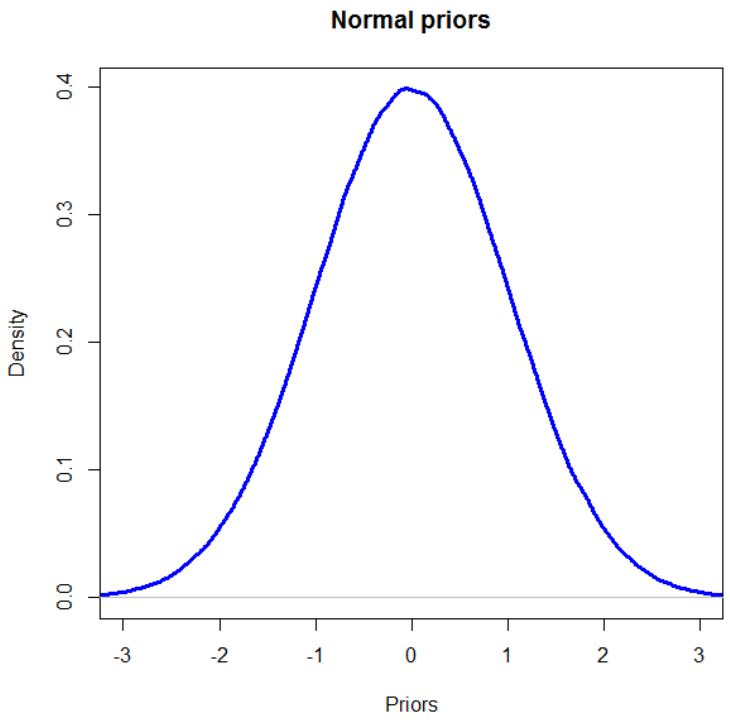
```
mean(ps_beta[, 1] > 0)
[1] 0.98925
```

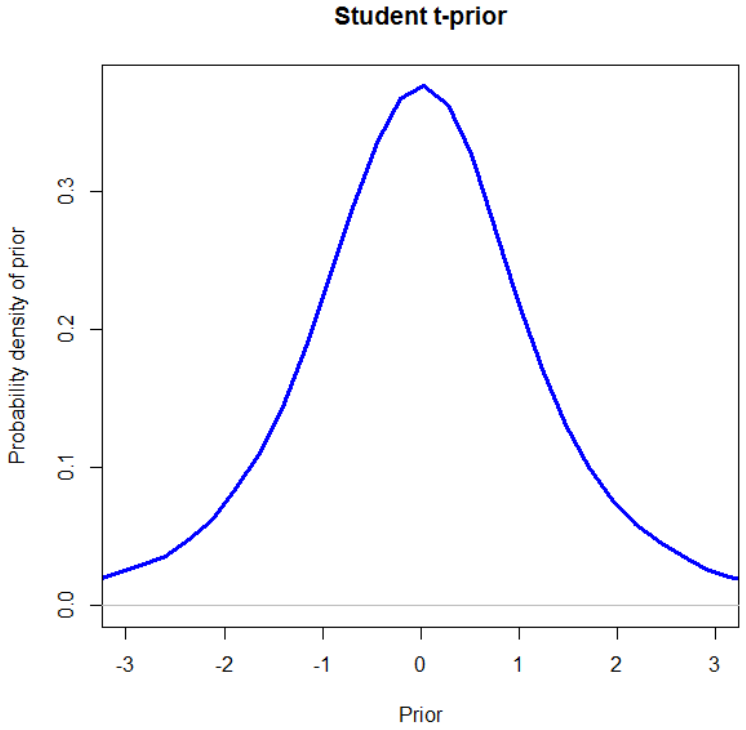
This means that there is almost a 99% probability that the effect of cookies on weight is positive. This is the probability of our research hypothesis.

Popular informative priors

- `normal(mean, scale)`
- `student_t(df, mean, scale)`
- `cauchy(mean, scale)`
- See examples in `?set_prior` and `?brm`

Watch the tails





Student t-priors allow for outliers.

Specific informative priors

- Purely hypothetically, imagine that there is some previous research that suggests that the estimate of ‘cookies’ is -2. That is, we expect that eating cookies leads to weight loss. You are highly confident about this result and expect very little deviation from this effect.

- You want to use this information as your priors. How to do it?

Using specific informative priors

```
mybrm2 <- brm(kilos ~ cookies, data = mydata,
prior = prior(normal(-2, 0.3), class = b))

ps_beta2 <- posterior_samples(mybrm2,
"cookies")
summary(ps_beta2[, 1])
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-2.62506 -1.67767 -1.44587 -1.42823 -1.19162  0.02805
```

How to visualize your priors

- Normal distribution:

```
plot(density(rnorm(n = 10000, mean = -2, sd = 0.3)))
```

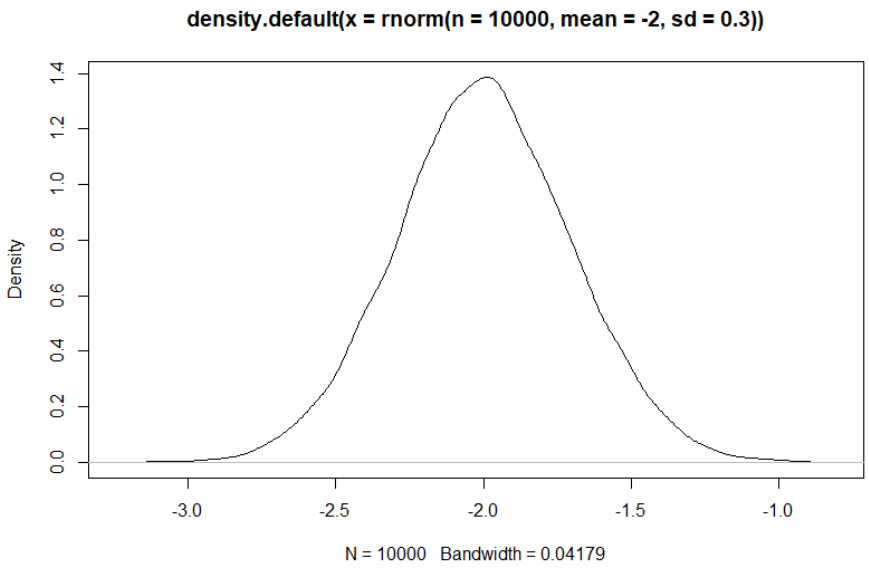
Mean is the expected estimate (the central value), sd is standard deviation, which shows how much certainty you have.

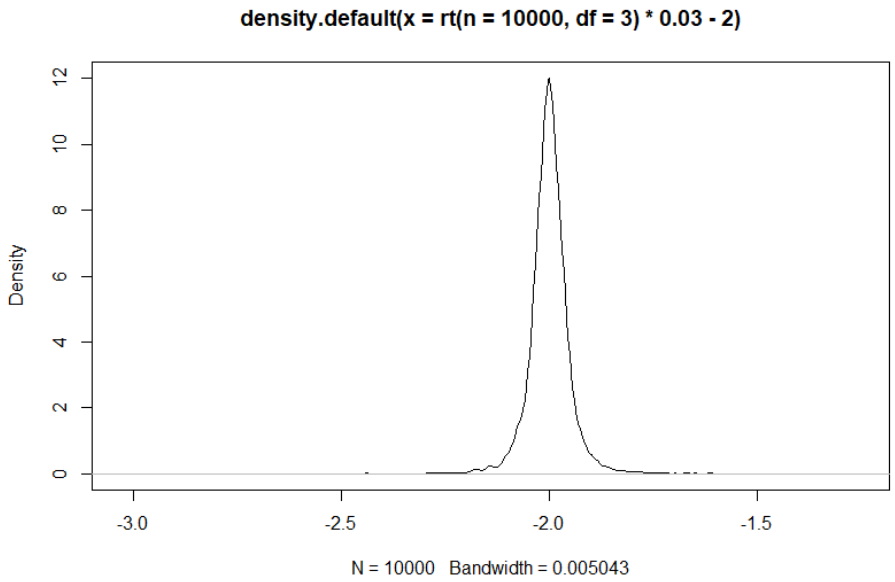
- Student t-distribution:

```
plot(density(rt(n = 10000, df = 3)*0.03 - 2))
```

where -2 is the central value, df = 3 shows the thickness of tails (the lower, the thicker), and 0.03 is scaling parameter (certainty).

Resulting plots





Goodness of fit

- R^2 is the best known goodness-of-fit measure, which ranges from 0 (useless model) to 1 (perfect fit). Its evaluation depends on the discipline. The brms package provides a Bayesian version of this statistic.

```
bayes_R2(mybrm)
     Estimate Est.Error        Q2.5     Q97.5
R2 0.4646035 0.1813101 0.02738844 0.6722349
```

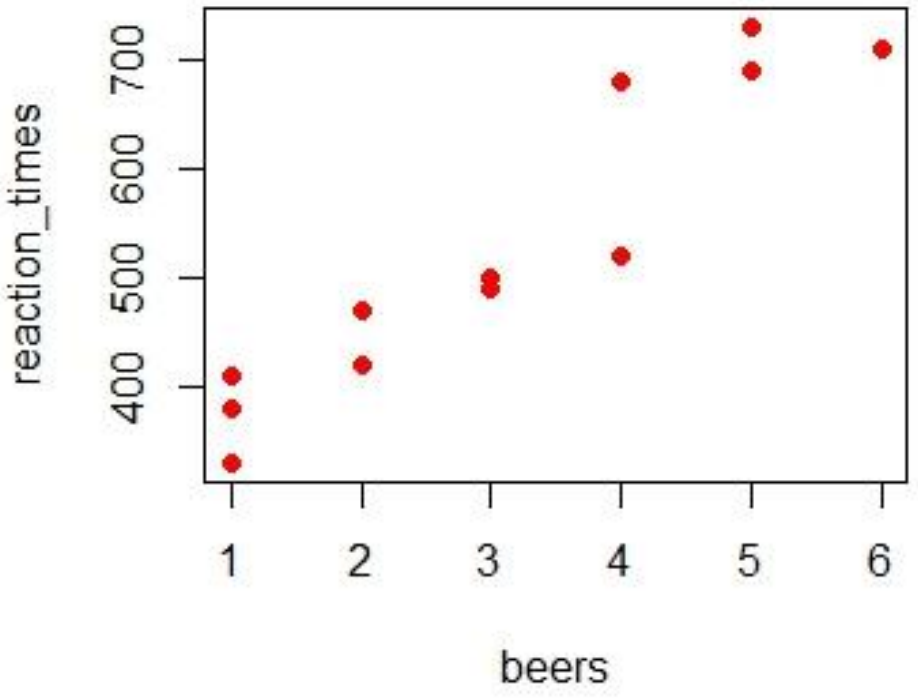
Exercise

- Examine the following model. What do you think it shows? Are there any problems? If yes, how would you fix them?

```
beers <- c(1, 1, 1, 2, 2, 3, 3, 4, 4, 5,
5, 6)
reaction_times <- c(330, 410, 380, 420,
470, 500, 490, 680, 520, 730, 690, 710)

exper_data <- data.frame(beers,
reaction_times)
```

Exercise (cont.)

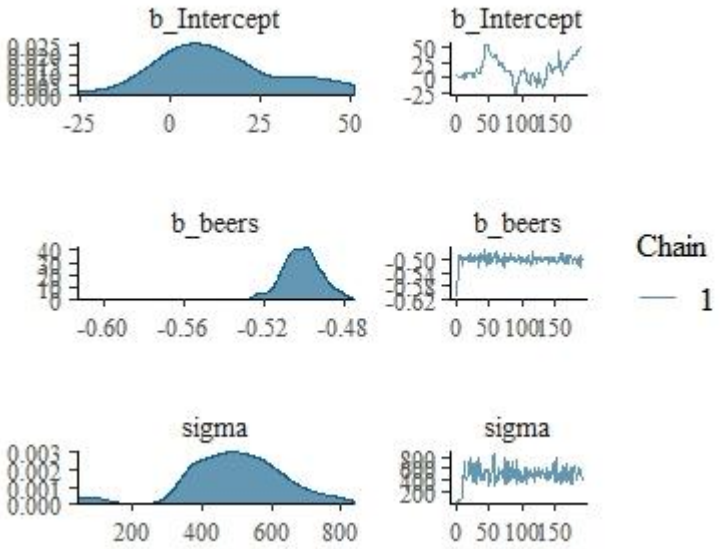
```
plot(beers, reaction_times, pch = 16, col
= "red")
```



Exercise (cont.)

```
exper_brm <- brm(reaction_times ~ beers, data
= exper_data, warmup = 10, chains = 1, iter =
200, prior = prior(normal(-0.5, 0.01), class
= b))
```

```
plot(exper_brm)
```



Exercise (cont.)

summary(exper_brm)

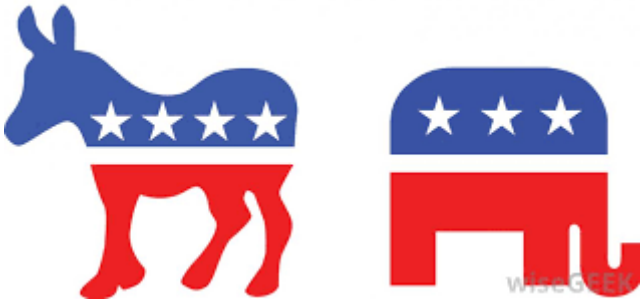
```
…
Samples: 1 chains, each with iter = 200; warmup = 10; thin = 1;
         total post-warmup samples = 190

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample
Intercept    13.28     17.47   -16.04    50.31          9
beers        -0.50      0.01    -0.52    -0.48        178
          Rhat
Intercept 0.99
beers     1.00
```

Additional useful features in brms

- Generalized Linear Models (e.g. logistic)
- Random effects, like in mixed models
- Smooth terms, like in Generalized Additive Models
- Many useful visualization tools

Logistic regression

- Situations with two or more categorical outcomes:

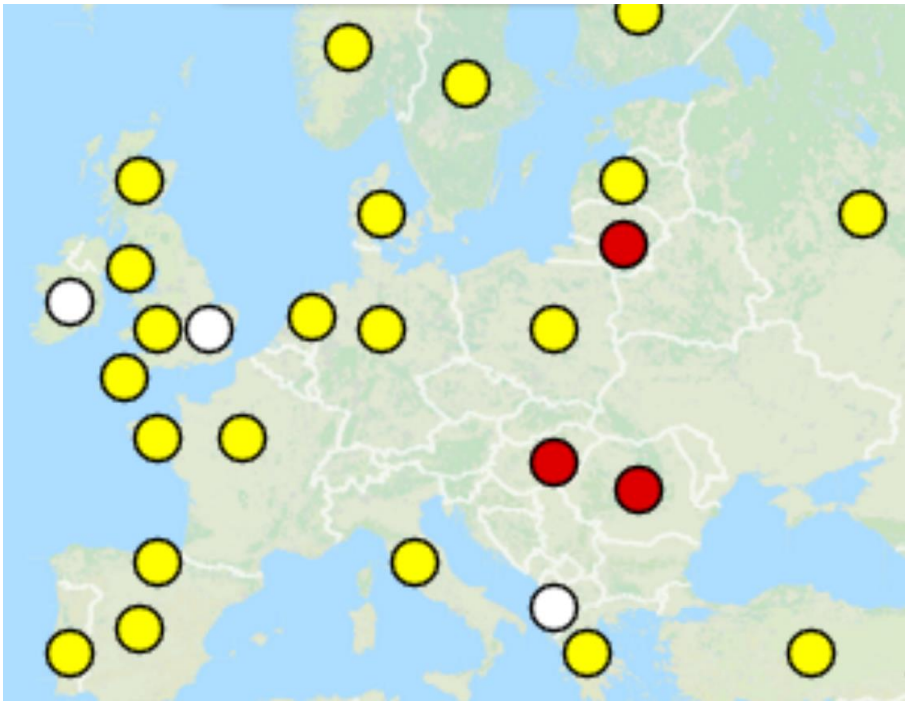


T or V?

- The distinction is present in most European languages
  - T forms: informal, familiar, e.g. French *tu*, German *du*, Russian *ty* + Verb 2nd SG
  - V forms: formal, polite, e.g. French *vous*, German *Sie*, Russian *vy* + Verb 2nd PL or 3rd SG/PL

Cross-linguistic research

- WALS Chapter 45, Helmbrecht 2013



Values

○	No politeness distinction	136
●	Binary politeness distinction	49
●	Multiple politeness distinctions	15
●	Pronouns avoided for politeness	7

Power and solidarity (Brown and Gilman 1960)

- Power dimension:
  - Based on “older than”, “richer than”, “parent of”, etc.
  - T = lower status of the Hearer, V = higher status of the Hearer
  - Systematic distinction already in the late Middle Ages. Everyone had his/her fixed place in the society.

- Solidarity dimension:
  - Based on “the same age/family/class as”.
  - T = closeness, V = distance
  - Emerged with social mobility and egalitarian ideology. Starting from the French revolution (*Citoyen*, *tu*).
  - Currently dominates in major European languages, but there are subtle cross-linguistic differences.

Data for the case study

- T/V forms
- 50 subjects, 18 questions
- Communicative situations:
  - Q_ID (ID of the question in the questionnaire)
  - Familiarity (Close, Middle and Far)
  - H_Age (Younger, Older and Same)
- Subject's characteristics
  - S_ID (Subject's ID)
  - S_Extrav (index of extraversion)
  - S_Age (age)

Data in R

str(tvdata)

```
'data.frame':  900 obs. of  9 variables:
$ S_ID        : Factor w/ 50 levels "1","2","3","4",..:
$ S_Year      : int  1934 1934 1934 1934 1934 1934 1934
$ S_Extrav    : int  74 74 74 74 74 74 74 74 74 74 ...
$ S_Age       : num  84 84 84 84 84 84 84 84 84 84 ...
$ Q_ID        : int  1 2 3 4 5 6 7 8 9 10 ...
$ Familiarity: Factor w/ 3 levels "Close","Far",..: 1 1
$ H_Age       : Factor w/ 3 levels "Older","Same",..: 3 2
$ Form        : Factor w/ 2 levels "T","V": 2 1 2 2 2 2
$ Name        : Factor w/ 3 levels "Short","First"
```

A mega-model with all kinds of complications

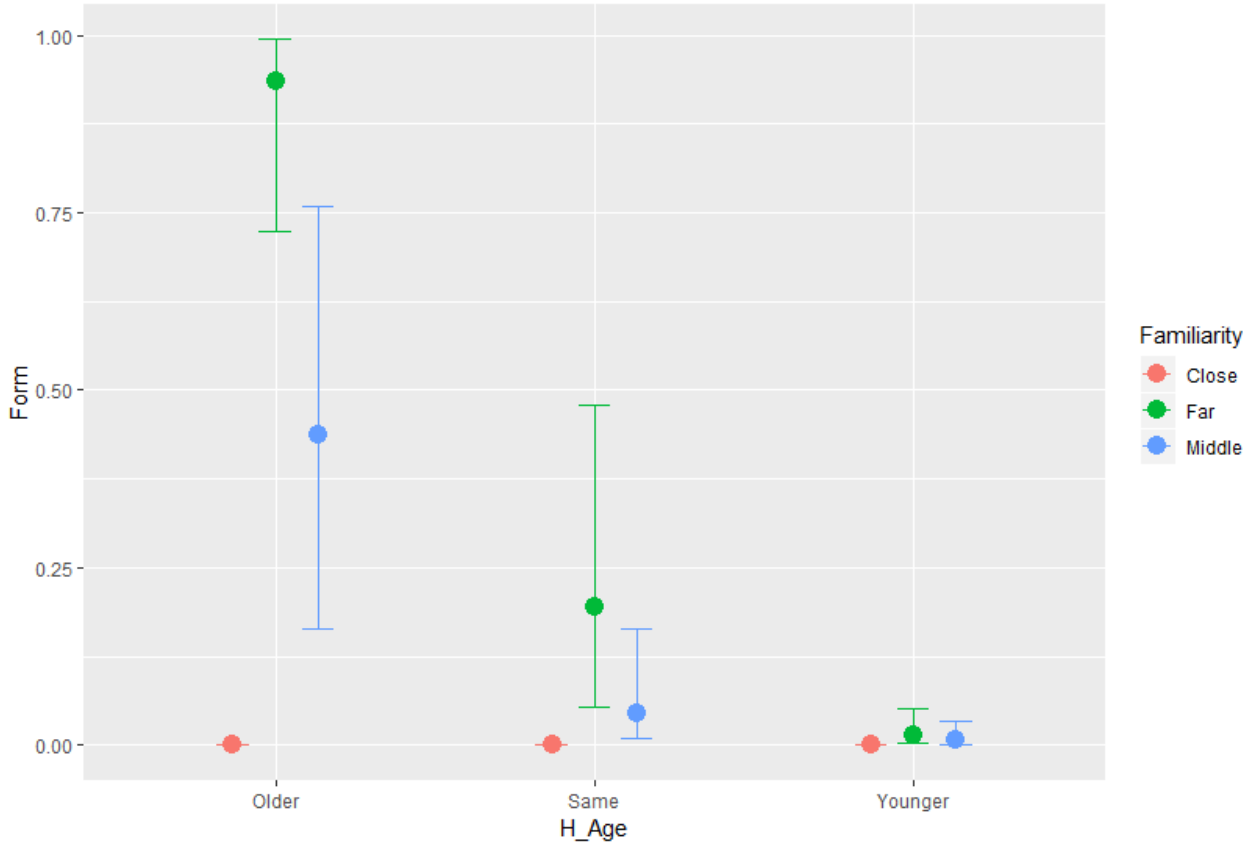
- For logistic models, add `family = Bernoulli`.
- Random effects are specified as in lme4.
- For smooths, one can currently use either `s()` or `t2()`:

```
mega_brm <- brm(Form ~ t2(S_Age, S_Extrav)
+ H_Age*Familiarity + (1|S_ID), data =
tvdata, family = bernoulli, chains = 2, iter
= 1000, warmup =  200)
```

Visualization of interactions

```
marginal_effects(mega_brm)
```

A selected plot:



Adjustments per subject

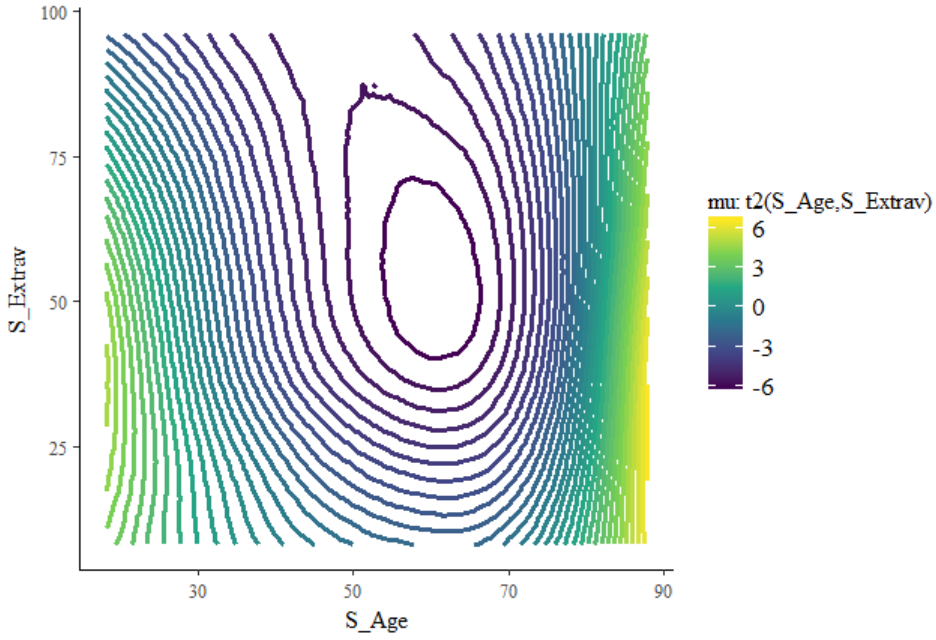
ranef(mega_brm)

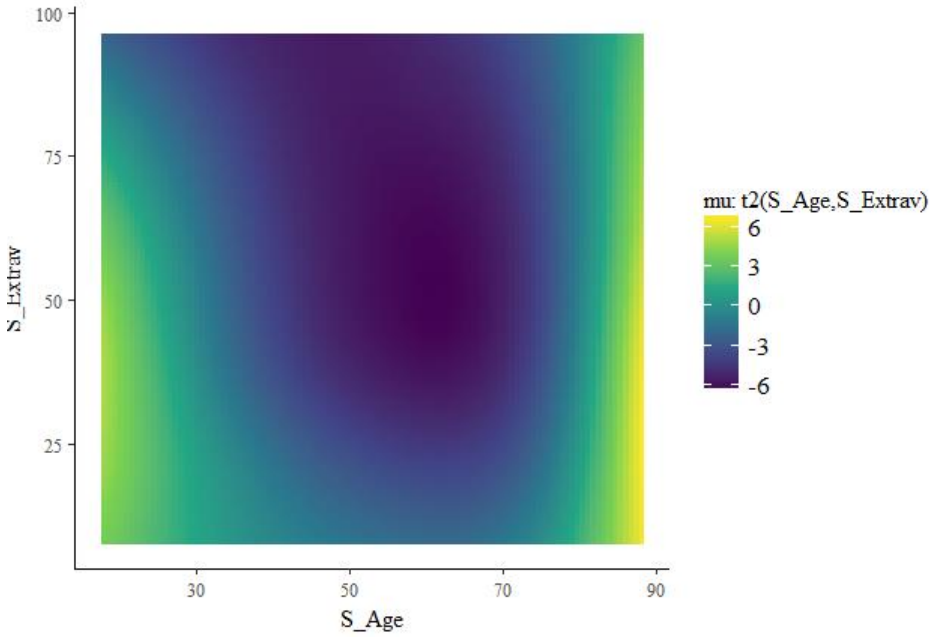
```
$S_ID
, , Intercept

      Estimate Est.Error        Q2.5     Q97.5
1   0.67845283 1.0264882 -1.25134855 2.8637459
2   0.12872995 0.7484718 -1.39487999 1.6670851
3  -0.05758030 0.8080955 -1.73774628 1.5519841
4   0.10880918 0.7484347 -1.36317190 1.6178879…
```

Plotting the smooths

```
ms <- marginal_smooths(mega_brm) #to save
time, because it's computationally intensive
plot(ms, stype = "contour")
plot(ms, stype = "raster")
```

Marginal smooth plots





Course outline

1. Basic concepts of Bayesian statistics

2. A simple illustration: Binomial proportions and online dating

3. Bayesian regression with brms

4. A linguistic illustration: help + (to) Infinitive

Help + (to) Infinitive

a. *If this book does not **help** you **to survive** the Zombie Apocalypse, a full refund may be obtained from the author.*

b. *Just to be on the safe side you might want to start doing these 8 exercises that will **help** you **survive** the zombie apocalypse.*

https://www.amazon.co.uk/Z-Day-UK-surviving-Apocalypse-Britain/dp/1490389873

http://steadystrength.com/8-exercises-that-will-help-you-survive-the-zombie-apocalypse/

Data and method

- GloWBE (7 geographic varieties of English)
- Samples of 5 to 7 thousands observations with *help* (to) Infinitive in each country
- Bayesian models with mixed effects (random intercepts for individual verbs and websites). The default priors (very weak generic ones).

Relevant factors

- Linguistic distance (the principle of cognitive complexity, Rohdenburg 1996)
  - …it's a way for me to make a contribution, to **help** the country in a small way **to get back** on its feet.
- Avoidance of identity, or horror aequi (Rohdenburg 2003)
  - Sorry, but how is this supposed **to help** answer the question?
- Register (formal -> to-infinitive)
- Inflectional form (helping +to-infinitive)
- Presence or absence of the Helpee (presence -> bare infinitive)
- And a few other factors.

Linguistic distance

Country	Posterior mean	2.5%	97.5%	P(β > 0)
Australia	0.46	0.3	0.63	100%
Ghana	0.7	0.57	0.84	100%
Great Britain	0.56	0.41	0.71	100%
Hong Kong	0.69	0.56	0.83	100%
India	0.82	0.66	0.99	100%
Jamaica	0.48	0.3	0.65	100%
USA	0.38	0.22	0.54	100%

Note: positive estimates mean the higher likelihood for the to-infinitive, negative estimates mean the higher likelihood of the bare infinitive.

Horror aequi

Country	Posterior mean	2.5%	97.5%	P(β > 0)
Australia	−1.33	−1.51	−1.15	0%
Ghana	−1.29	−1.46	−1.14	0%
Great Britain	−1.24	−1.4	−1.09	0%
Hong Kong	−1.13	−1.27	−0.98	0%
India	−1.35	−1.54	−1.17	0%
Jamaica	−1.62	−1.82	−1.44	0%
USA	−1.3	−1.52	−1.09	0%

Note: positive estimates mean the higher likelihood for the to-infinitive, negative estimates mean the higher likelihood of the bare infinitive.

Formality (average word length in a text)

Country	Posterior mean	2.5%	97.5%	P(β > 0)
Australia	0.3	0.04	0.57	98.9%
Ghana	0.1	−0.17	0.36	76.1%
Great Britain	0.59	0.33	0.86	100%
Hong Kong	0.27	0.03	0.52	98.7%
India	−0.35	−0.63	−0.07	0.8%
Jamaica	0.3	−0.02	0.64	96.8%
USA	0.12	−0.2	0.44	76.6%

Note: positive estimates mean the higher likelihood for the to-infinitive, negative estimates mean the higher likelihood of the bare infinitive.

Informativity effects in language

- There is ample evidence that more predictable elements are reduced in language production.

- E.g. Jaeger (2010): if a complement clause is predictable given the verb, the complementizer *that* is more likely to be omitted.

  - I think you're right.
  - I show that alternatives exist.

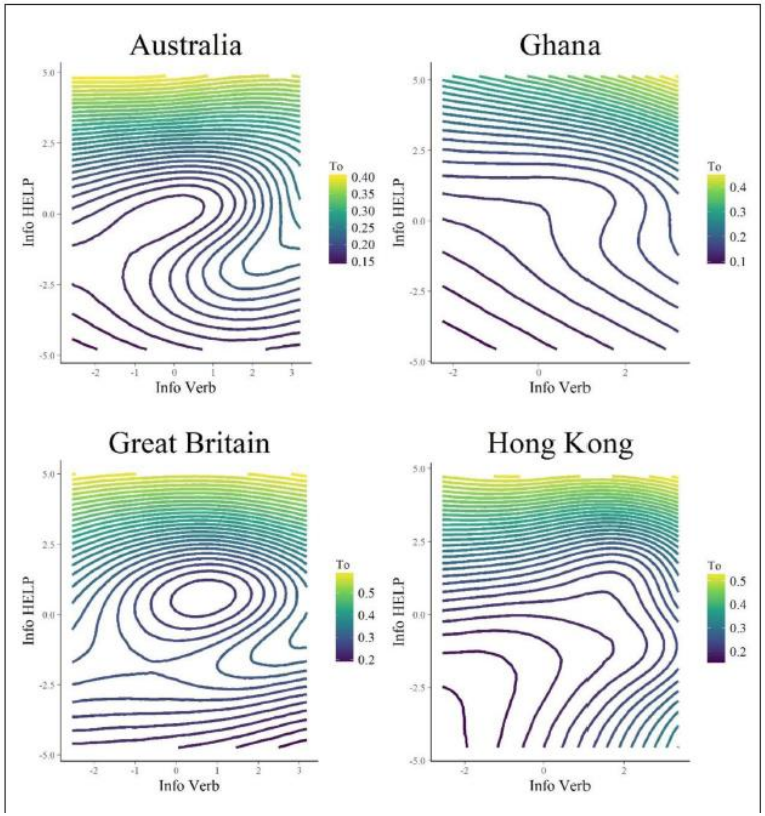
https://doi.org/10.5281/zenodo.1542857

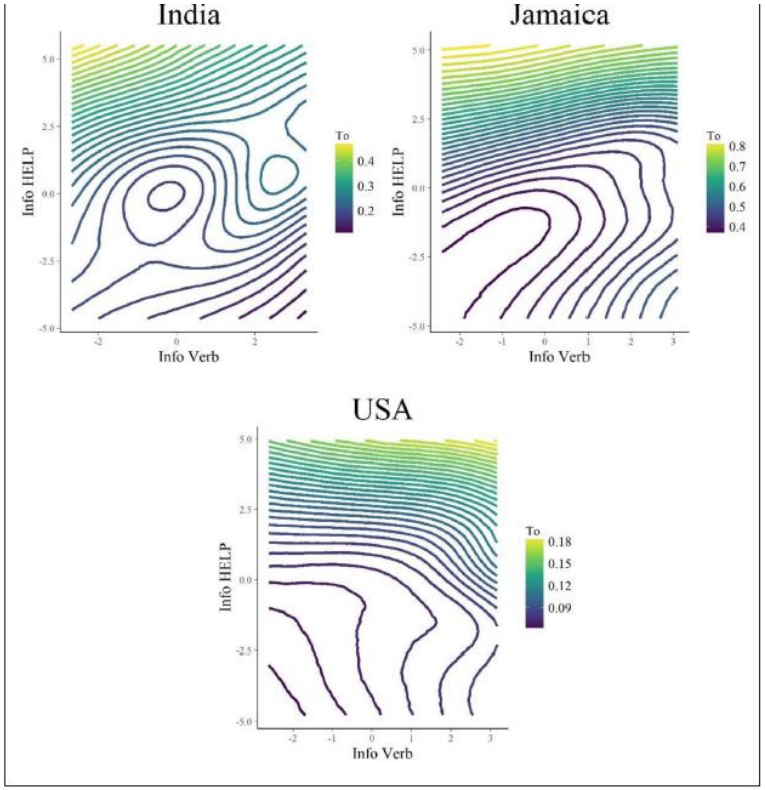
Slot-filler predictability

- InfoVerb: -log predictability of the verb given HELP

- InfoHelp: -log predictability of HELP given the verb

Slot-filler predictability





Interpretation

- Remarkably, all varieties show some kind of effects of using to when the predictability of HELP given the verb is low.
- This is the case of high-frequency verbs, e.g.

  - *Oh Lord, help me to be pure, but not yet.* (St. Augustine)

More details

Levshina, Natalia. 2018. Probabilistic grammar and constructional predictability: Bayesian generalized additive models of *help* + (to) Infinitive in varieties of web-based English. *Glossa* 3(1). 55.1-22. DOI: 10.5334/gjgl.294/