Model design in R: Class 5

Alexandre Cremers

August 16th, 2019

Today's plan

- Introduce and discuss yet another problem with SemPrag data: participants bimodal distributions.
- Propose a technical solution with Stan, and other possible options within plain R.
- Recap of the week / open questions / ...

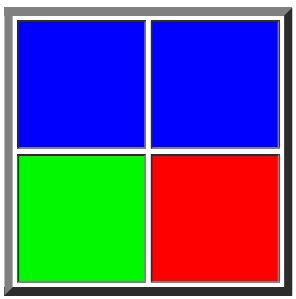
Usual situation

- Usual truth-value judgment task: Ambiguous targets forcing participants to choose between a true and false interpretation.
- Participants tend to stick to their initial decision.
- ☞ Participants split into two groups: *true responders* and *false responders*.
- Participants within each group are usually consistent, but there may be differences in consistency between the two groups.

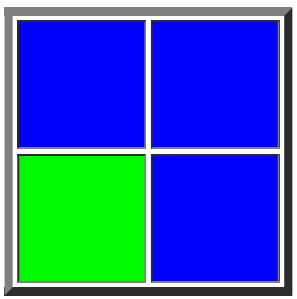
Tavano & Kaiser 2010

Example

(1) John knows which squares are blue



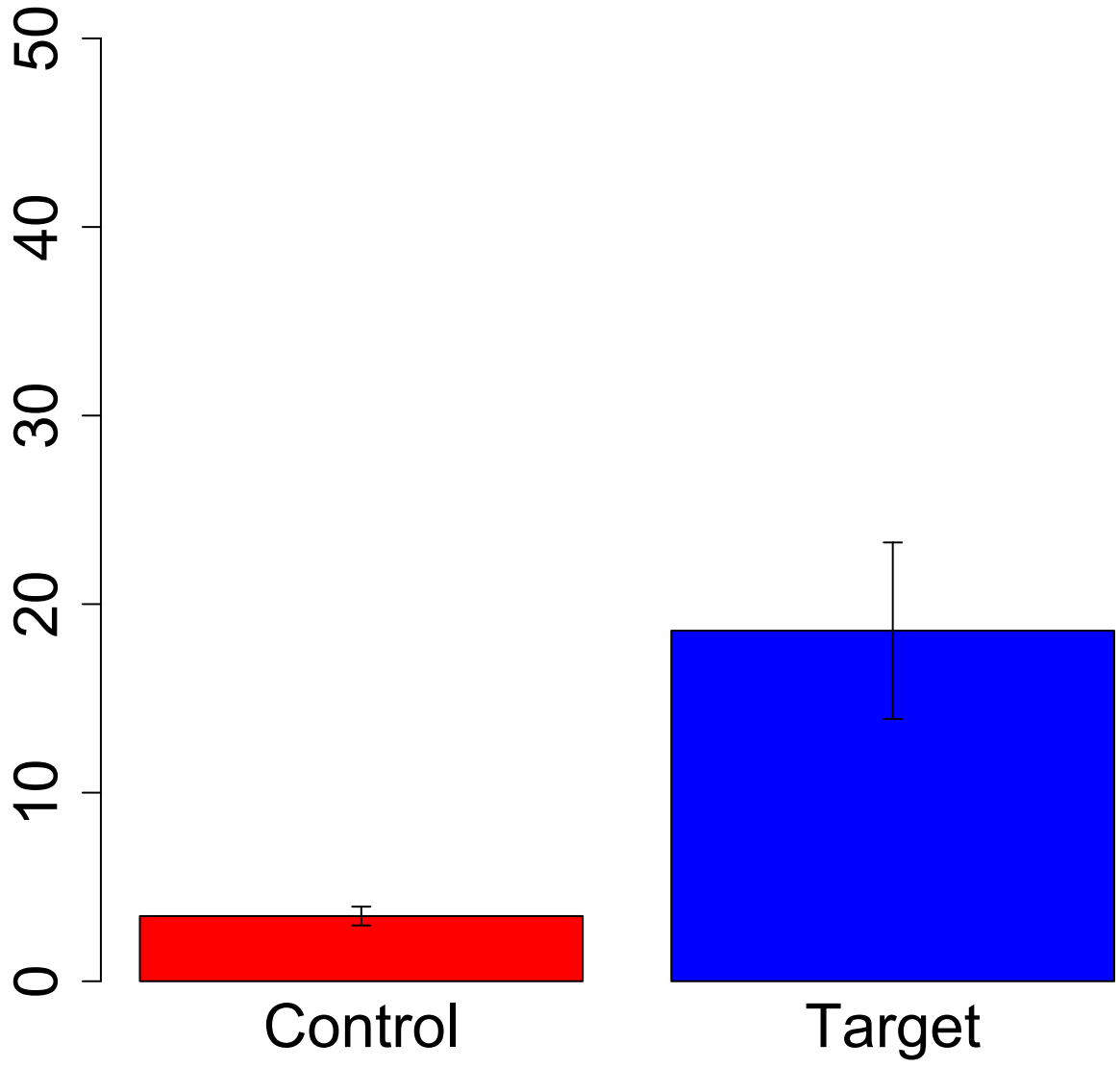
The actual card



John's beliefs

- Uncontroversial false reading (IE/SE)
- Controversial true reading (Weak Exhaustive)
- Research question: is this sentence judged true more often than a false control?

Results (Cremers&Chemla 2016)



First stats

- Mixed-effects logistic regression with:

```
Cond=if_else(Type=="false",-1,+1)
glmer(Resp ~ Cond + (1+Cond|Subj),
      family=binomial(link="logit"))
```

- Comparison to a model without `Cond`: $\chi^2(1) = 4.6, p = .032$
- Looking at the model summary:

```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.3425     0.8067  -6.623 3.53e-11 ***
Cond         -1.6831     0.8041  -2.093   0.0363 *
---
```

- `glmer` fitted a *negative* effect of `Cond`!

The problem in detail

Predicted probability of True responses:

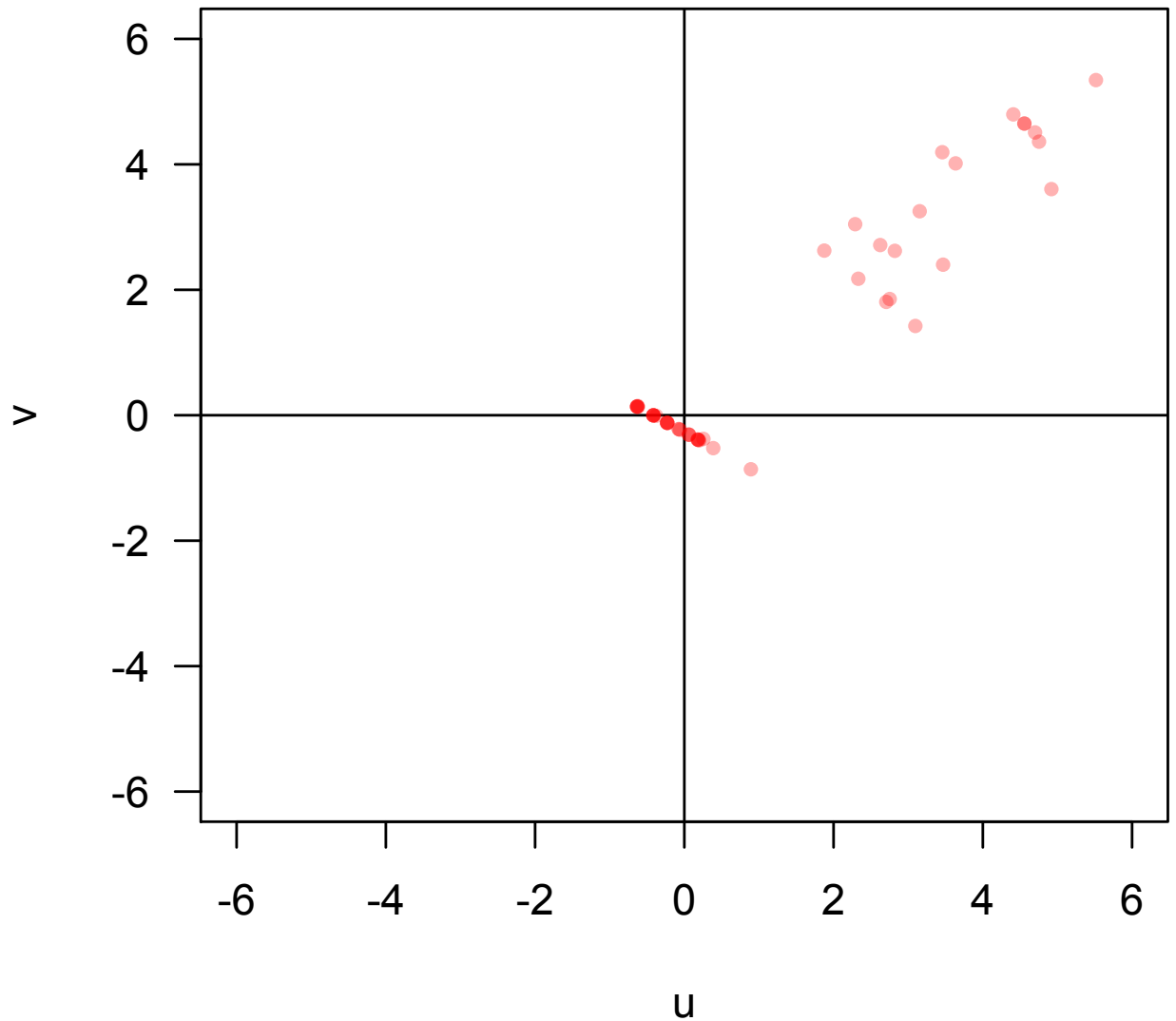
On false controls:	`plogis(-5.3425+1.6831) = 0.02510164`
On targets:	`plogis(-5.3425-1.6831) = 0.0008880447`

Reminder: actual rate of True responses on targets: 24%!

Where did the model hide the True responses?

Random effects

In the random effects, which are clearly not centered on 0.



Why did this happen?

- `lme4` assumes a normal distribution of the random effects (on the logit scale in case of logistic regression)
- We made it fit a GLMM on our data in order to show that there were more True responses on targets than on controls.
- This higher rate of True responses was meant to prove that the targets are ambiguous, i.e. that the participants are split into two groups!

☞ In short: we asked `lme4` to prove that the hypotheses for a `lme4` model are violated!

A (rather) technical solution

Stan: a probabilistic programming language for Bayesian statistics.

- Will allow us to specify the statistical model manually, including "weird" things like bimodal distributions.
- Can be integrated in `R` code.

Problematic `lme4` assumption:

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right)$$

Hand-coding RE structure

New assumption:

- A random intercept $\boldsymbol{u}$ with normal distribution
- A random slope $\boldsymbol{v}$ following a mixture of two gaussians (bimodal distribution).

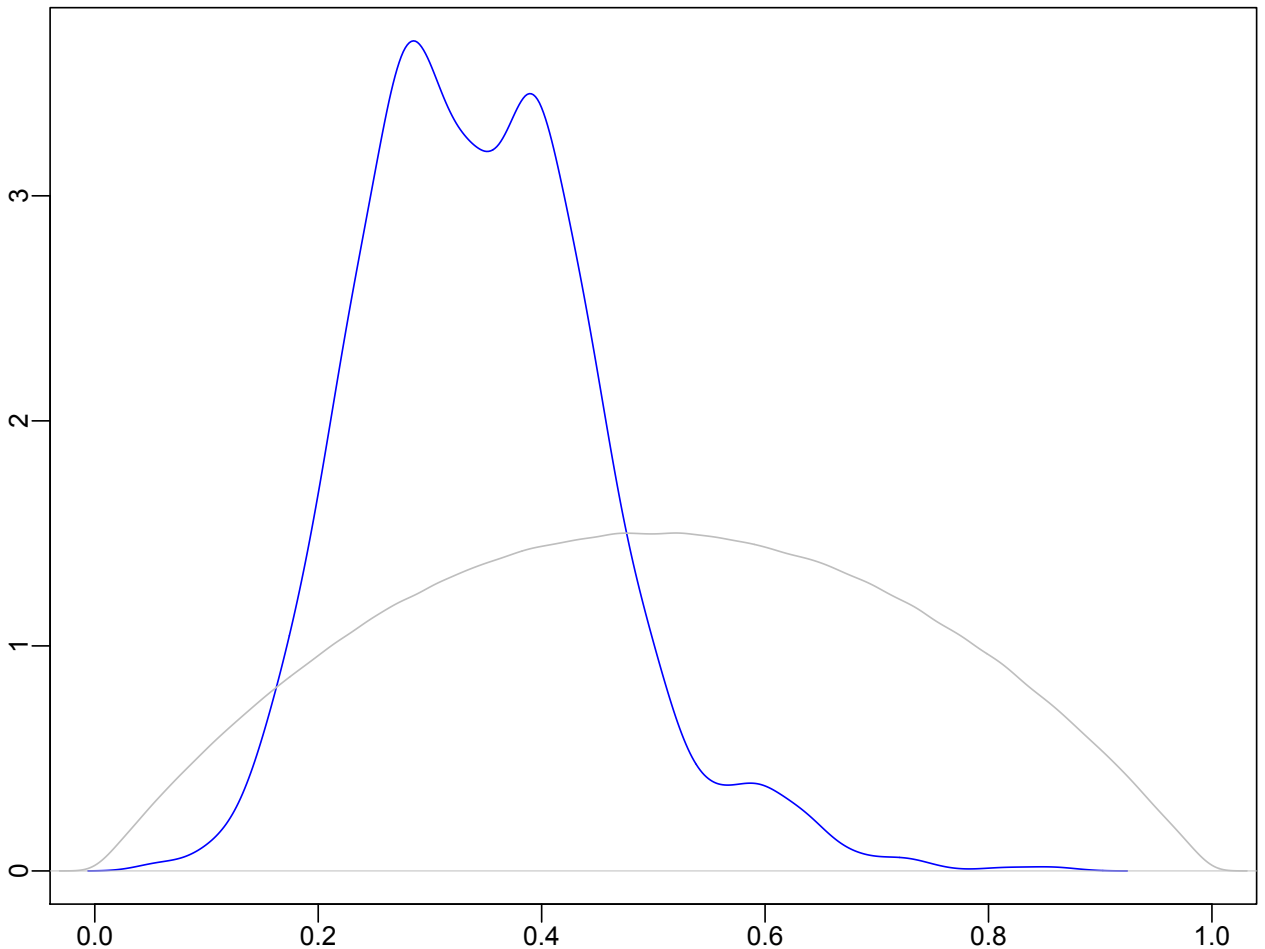
$$P(V = v) = \tau P(V = v|C = c_1) + (1 - \tau)P(V = v|C = c_2)$$

Relevant piece of Stan code (possibly outdated):

```
for (s in 1:Nsubj) {
                  S[s][1] ~ normal(0,S0_var);
                  increment_log_prob
                          (log_sum_exp(log(S1_tau) +
                          normal_log(S[s][2],S1_mu[1],S1_var_a),
                          log(1-S1_tau) +
                          normal_log(S[s][2],S1_mu[2],S1_var_b)));
}
```

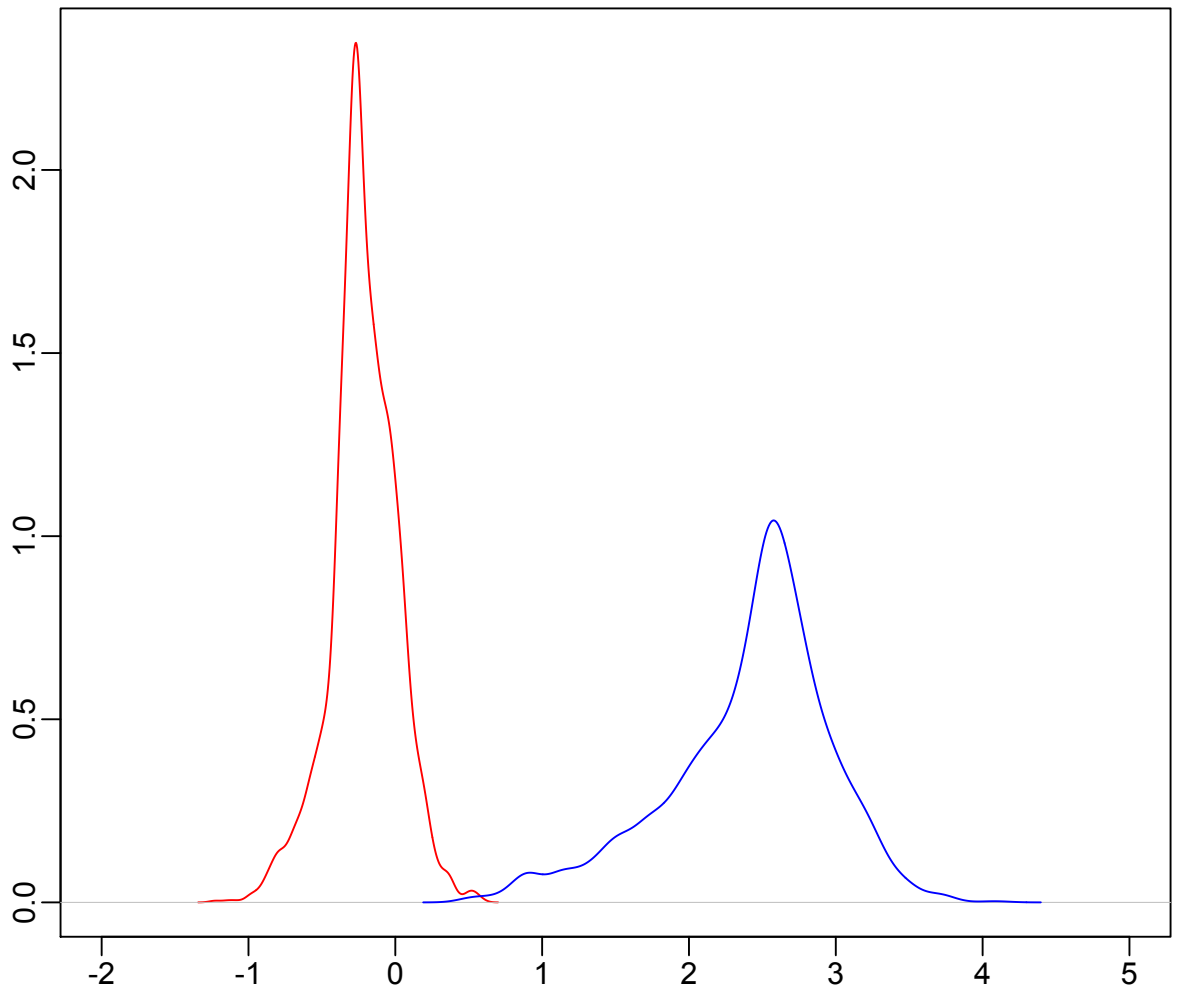
Stan results

Posterior distribution of τ (proportion of WE responders):



Stan results

Posterior distribution of slopes:
(difference between false control and target;
red: IE/SE responders, blue: WE responders)

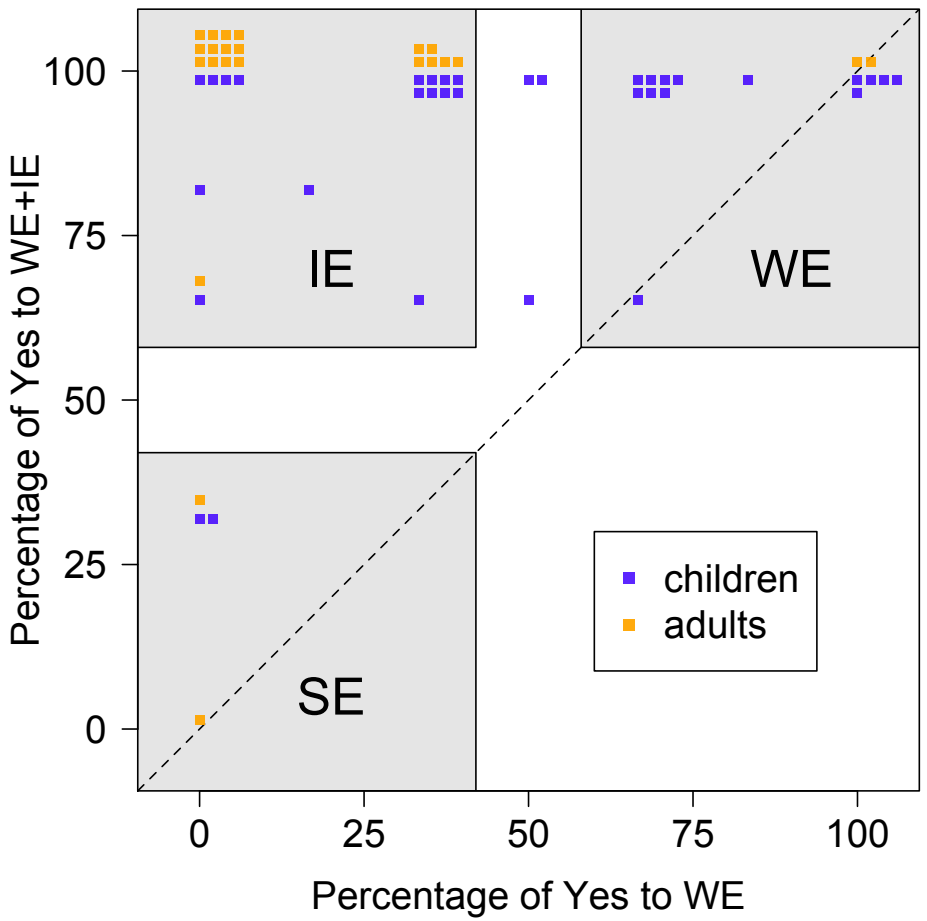


Conclusion

- With more sophisticated tools, it's possible to fit exactly the model we have in mind.
- However, the solution is pretty technical, even with prior knowledge of Stan (fitting mixture models is tricky, see discussion here)
- Alternatives?

How to deal with bimodal distributions?

Easier solution: classify your participants.



Classifying participants

The idea:

- If your participants are very consistent, you can reduce all data from one participant to a single label (e.g., WE/IE/SE).
- You can then run simple analyses on the count of participants in each category (χ^2 tests, Poisson models, see example here).
- This is very useful for acquisition studies, where you typically have little data per participant anyway.
- Downside: you lose data on participants who are inconsistent, which may be problematic if there are many of them, or if you are interested in these particular participants.

Conclusion

- The data we collect in SemPrag has a number of peculiarities which make it tricky to analyse.
- More sophisticated modelling methods allow us to specify precisely what our questions are, which...
  1. makes the statistical analyses more informative,
  2. avoids incorrect use of "default" statistical tools outside their range of application.
- It's always better to think before blindly applying a model (and never too late after either). When things go wrong:
  1. there's usually a reason why they do!
  2. better options often exist (more technical or actually simpler).