STAT 237
Prior and Posterior Predictive Checks

April 18-22, 2022

Colin Reimer Dawson

Outline

Outline

Review: Hierarchical Model and Posterior Inference

Selecting a Prior and Sanity Checks

Recall Peter Venkman's experiment to test for extra-sensory perception: Each of 1000 people attempted to call the result of 10 consecutive coin flips. How can we model this data?

- Data: $y_{n|s}$: **outcome of trial** n for person s

$$y_{1s} \dots y_{Ns} \mid \theta_s \overset{i.i.d.}{\sim} \text{Bernoulli}(\theta_s)$$

- θ_s: **long-run success chance** for person s

$$\theta_s \mid \omega, \kappa \overset{i.i.d.}{\sim} \text{Beta}(\kappa\omega, \kappa(1-\omega))$$

- ω: **mean success rate** across the population

$$\omega \sim \text{Beta}(\gamma_\omega \mu_\omega, \gamma_\omega(1-\mu_\omega))$$

- κ: **homogeneity of the population**

$$\kappa \sim \text{Gamma}(\gamma_\kappa \mu_\kappa^2, \gamma_\kappa \mu_\kappa)$$

- μ_ω, μ_κ: Prior means of ω and κ
- $\gamma_\omega, \gamma_\kappa$: (Roughly) prior "precision" of ω and κ

Posterior Distribution

- The **joint posterior density** over the θs, ω and κ is obtained by the product rule and then normalization (though we don't actually need the normalized value)

$$p(\boldsymbol{\theta}, \omega, \kappa \mid \mathbf{Y}) = \frac{p(\omega, \kappa, \boldsymbol{\theta}) p(\mathbf{Y} \mid \omega, \kappa, \boldsymbol{\theta})}{p(\mathbf{Y})}$$

$$= \frac{p(\omega) p(\kappa) \prod_{s=1}^{S} p(\theta_s \mid \omega, \kappa) \prod_{s=1}^{S} \prod_{n=1}^{N} p(y_{n|s} \mid \theta_s)}{p(\mathbf{Y})}$$

- We obtain T **samples** from this distribution using MCMC (as implemented by something like Stan):

$$(\boldsymbol{\theta}^{(t)}, \omega^{(t)}, \kappa^{(t)}), \qquad t = 1, \ldots T$$

- We can then use these samples to estimate Expected Values of various functions of the parameters

$$\mathbb{E}\left[g(\boldsymbol{\theta}, \omega, \kappa) \mid \mathbf{y}\right] \approx \frac{1}{T} \sum_{t=1}^{T} g(\boldsymbol{\theta}^{(t)}, \omega^{(t)}, \kappa^{(t)})$$

Outline

The Prior Predictive Distribution

- As models get more complicated and there are more parameters (and hyperparameters) to think about, it gets harder to have good intuitions about what our "top-level" choices imply for our model

The Prior Predictive Distribution

- As models get more complicated and there are more parameters (and hyperparameters) to think about, it gets harder to have good intuitions about what our "top-level" choices imply for our model
- For example: It might be easy to choose μ_ω, but what about μ_κ, γ_ω and γ_κ?

The Prior Predictive Distribution

- As models get more complicated and there are more parameters (and hyperparameters) to think about, it gets harder to have good intuitions about what our "top-level" choices imply for our model
- For example: It might be easy to choose μ_ω, but what about μ_κ, γ_ω and γ_κ?
- Rather than try to figure this out by intuition, it's generally easier to **see what kinds of data** our choice of prior generates

The Prior Predictive Distribution

- As models get more complicated and there are more parameters (and hyperparameters) to think about, it gets harder to have good intuitions about what our "top-level" choices imply for our model
- For example: It might be easy to choose μ_ω, but what about μ_κ, γ_ω and γ_κ?
- Rather than try to figure this out by intuition, it's generally easier to **see what kinds of data** our choice of prior generates
- The **prior predictive distribution** of our model is the **marginal distribution** over data variables. In our example:

$$\begin{aligned} p(\mathbf{y}) &= \int \int \int p(\mathbf{y}, \boldsymbol{\theta}, \omega, \kappa)\, d\boldsymbol{\theta}\, d\omega\, d\kappa \\ &= \int p(\kappa) \int p(\omega) \int p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \omega, \kappa)\, d\boldsymbol{\theta}\, d\omega\, d\kappa \end{aligned}$$

The Prior Predictive Distribution

- The **prior predictive distribution** of our model is the **marginal distribution** over data variables. In our example:

$$p(\mathbf{y}) = \int p(\kappa) \int p(\omega) \int p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \omega, \kappa) \, d\boldsymbol{\theta} \, d\omega \, d\kappa$$

The Prior Predictive Distribution

- The **prior predictive distribution** of our model is the **marginal distribution** over data variables. In our example:

$$p(\mathbf{y}) = \int p(\kappa) \int p(\omega) \int p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \omega, \kappa) \, d\boldsymbol{\theta} \, d\omega \, d\kappa$$

- This looks... scary

The Prior Predictive Distribution

- The **prior predictive distribution** of our model is the **marginal distribution** over data variables. In our example:

$$p(\mathbf{y}) = \int p(\kappa) \int p(\omega) \int p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \omega, \kappa)\, d\boldsymbol{\theta}\, d\omega\, d\kappa$$

- This looks... scary
- Fortunately, we don't need to work with this integral directly: we can again **sample** from the distribution

The Prior Predictive Distribution

- The **prior predictive distribution** of our model is the **marginal distribution** over data variables. In our example:

$$p(\mathbf{y}) = \int p(\kappa) \int p(\omega) \int p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \omega, \kappa) \, d\boldsymbol{\theta} \, d\omega \, d\kappa$$

- This looks... scary
- Fortunately, we don't need to work with this integral directly: we can again **sample** from the distribution
- We don't even need MCMC for this, because the sampling is purely **top-down**

Simulating Data From the Prior Predictive

At each iteration:

1. Starting at the "roots", sample each "parentless" parameter from its prior distribution
2. Then sample each parameter that has only those variables as "parents" from its prior, conditioning on the parent values
3. Continue down the tree until we have the direct parents of the data variables, then sample the data variables from their conditional distribution

Simulating Data From the Prior Predictive

At each iteration:

1. Starting at the "roots", sample each "parentless" parameter from its prior distribution
2. Then sample each parameter that has only those variables as "parents" from its prior, conditioning on the parent values
3. Continue down the tree until we have the direct parents of the data variables, then sample the data variables from their conditional distribution

In our example, for $t = 1, \ldots, T$:

1. Sample $\omega^{(t)}$ and $\kappa^{(t)}$ from their priors, $p(\omega)$ and $p(\kappa)$
2. Sample $\theta_1^{(t)}, \ldots, \theta_S^{(t)}$ from $p(\theta_s \mid \omega, \kappa)$
3. Sample $y_{1s}^{(t)}, \ldots, y_{Ns}^{(t)}$ from $p(y_{ns} \mid \theta_s)$

What do we do with the results?

- Useful to **visualize** the resulting datasets

What do we do with the results?

- Useful to **visualize** the resulting datasets
- Our main consideration is that the results cover **all of the kinds** of datasets that we think we're likely to see

What do we do with the results?

- Useful to **visualize** the resulting datasets
- Our main consideration is that the results cover **all of the kinds** of datasets that we think we're likely to see
- It's ok if some of the results include some datasets that we don't think we'd see: Better to be overly inclusive than too restrictive

What do we do with the results?

- Useful to **visualize** the resulting datasets
- Our main consideration is that the results cover **all of the kinds** of datasets that we think we're likely to see
- It's ok if some of the results include some datasets that we don't think we'd see: Better to be overly inclusive than too restrictive
- That said, if **most** of the results are implausible, our prior is probably **too broad**