Statistical Rethinking Week 2 Notes

Kanishka Misra

5/7/2019

Gaussian Model for Height

The data in this example are collected from a survey of Kalahari Foragers, collected by Nancy Howell.

```
data("Howell1")
d <- Howell1

precis(d) %>%
  rownames_to_column(var = "variable") %>%
  as_tibble() %>%
  select(-histogram) %>%
  kable(caption = "Descriptives about the Data")
```

Table 1: Descriptives about the Data

variable	mean	sd	5.5%	94.5%
height	138.2635963	27.6024476	81.108550	165.73500
weight	35.6106176	14.7191782	9.360721	54.50289
age	29.3443934	20.7468882	1.000000	66.13500
male	0.4724265	0.4996986	0.000000	1.00000

In this example, we look at the height of all adults present in the data. The prior for Height we use is:

$$h_i \sim Normal(\mu, \sigma) \tag{1}$$

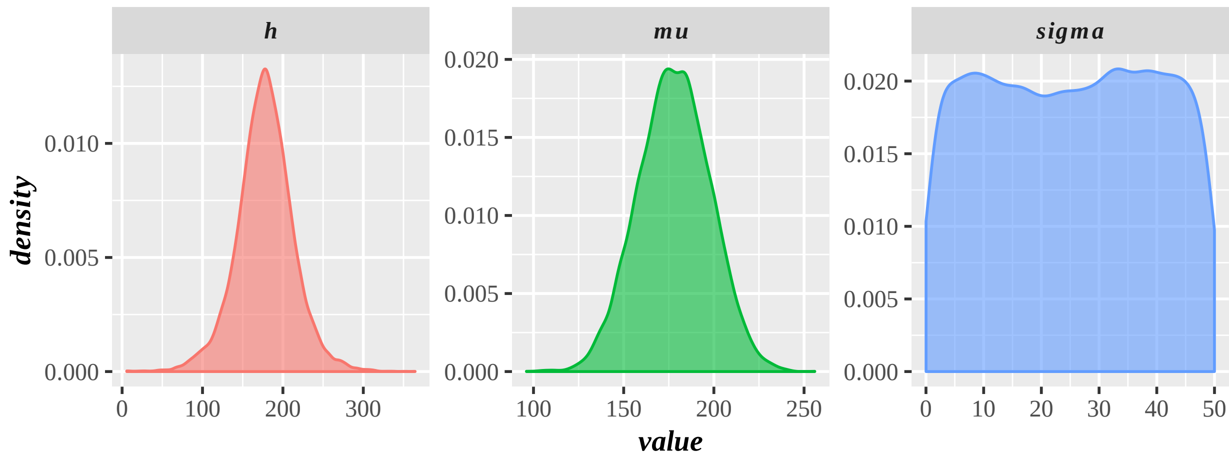
Here, height is observed, while μ and σ are not, and thus have to be inferred from h_i. Since it is a bayesian model, we assert their priors as:

$$\mu \sim Normal(178, 20) \tag{2}$$
$$\sigma \sim Uniform(0, 50) \tag{3}$$

We can now simulate from these distributions to check what the model believes prior to seeing the data, this is known as *Prior Predictive Distribution*. This is the best way to check and see what the prior means. This is not *p-hacking* since data is not used. We typically use these models from scientifically available resources.

```
ppd_height <- tibble(
  sample_mu = rnorm(1e4, 178, 20),
  sample_sigma = runif(1e4, 0, 50),
  prior_h = rnorm(1e4, sample_mu, sample_sigma)
)

ppd_height %>%
  gather("variable", "value") %>%
```

```
  mutate(
    variable = case_when(
      variable == "prior_h" ~ "h",
      variable == "sample_mu" ~ as.character(bquote(mu)),
      variable == "sample_sigma" ~ as.character(bquote(sigma))
    )
  ) %>%
  ggplot(aes(value, fill = variable, color = variable)) +
  geom_density(alpha = 0.6, show.legend = F) +
  facet_wrap(~variable, scales = "free") +
  theme_latex()
```



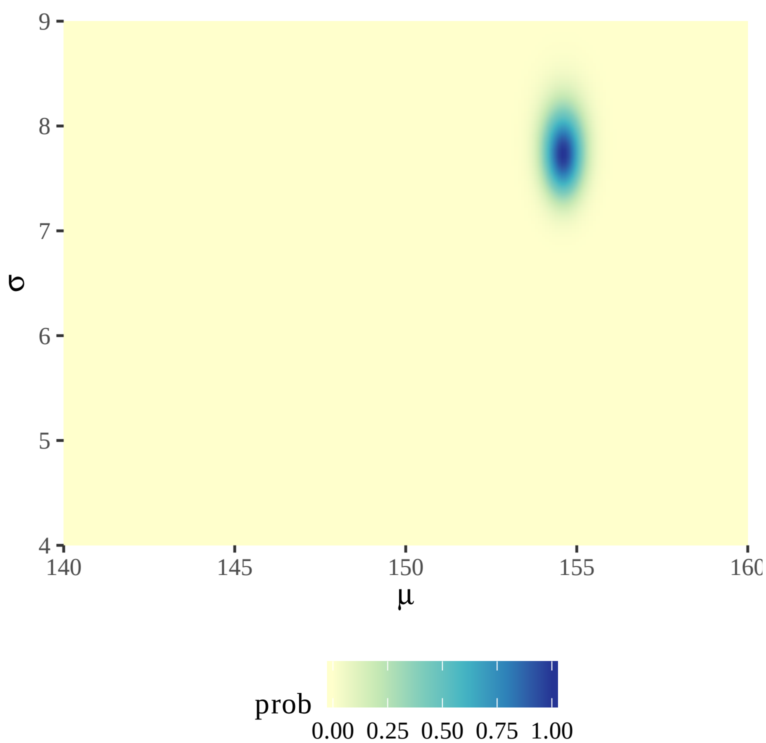
While this prior is not very informed, it still lies in the realm of possibility ($h_i > 0$ cm), however, this prior seems to have a lot of unnaturally tall people.

Computing the Posterior Distribution using Grid Approximation

Usually, we never use grid approximation in practical scenarios, but since we only have 2 variables, we do it as follows, for 200×200 possibilities.

```
# Subset d for adults
d2 <- d %>%
  filter(age >= 18)

n <- 200

d_grid <- tibble(
  mu = seq(140, 160, length.out = n),
  sigma = seq(4, 9, length.out = n)
) %>%
  expand(mu, sigma)

compute_grid <- function(mu, sigma, variable) {
  dnorm(variable, mean = mu, sd = sigma, log = T) %>%
    sum()
}

d_grid <- d_grid %>%
  mutate(
```

```r
    log_likelihood = map2_dbl(mu, sigma, compute_grid, variable = d2$height)
  ) %>%
  mutate(
    prior_mu = dnorm(mu, mean = 178, sd = 20, log = T),
    prior_sigma = dunif(sigma, min = 0, max = 50, log = T),
    product = log_likelihood + prior_mu + prior_sigma,
    prob = exp(product - max(product))
  )

d_grid %>%
  ggplot(aes(x = mu, y = sigma)) +
  geom_raster(aes(fill = prob),
              interpolate = T) +
  # scale_fill_viridis_c() +
  scale_x_continuous(expand = c(0,0), limits = c(130, 165)) +
  scale_y_continuous(expand = c(0,0)) +
  scale_fill_distiller(palette = "YlGnBu", direction = 1) +
  labs(x = expression(mu),
       y = expression(sigma)) +
  coord_cartesian(xlim = range(d_grid$mu),
                  ylim = range(d_grid$sigma)) +
  theme_latex() +
  theme(
    legend.position = "bottom"
  )
```



We now sample from μ and σ

Quadratic Approximation of the Posterior

Adding a Predictor: Weight