Quantitive Methods 2, ZHAW

Jürgen Degenfellner

2025-01-15

Contents

Chapter 1

Introduction

This script is a continuation of the first one for Quantitative Methods 1 at ZHAW.

In the first part, we learned about the basics of probability theory, descriptive statistics, Bayesian statistics, and hypothesis testing.

In this script, we will dive into the basics of statistical modeling - a world of aesthetic wonder and surprises.

This script is a first draft as you are the first group to be working with it.

Please feel free to send me suggestions for improvements or corrections.

This **should be a collaborative effort** and will (hopefully) never be finished as our insight grows over time.

The script can also be seen as a pointer to great sources which are fit to deepen your understanding of the topics. Knowledge is decentralized, and there are many great ressources out there.

For the working setup with R, please see this and the following sections in the first script.

The complete code for this script can be found here.

1.1 Books we will heavily borrow from are:

- (Free) Statistical Rethinking, YouTube-Playlist: Statistical Rethinking 2023
- (Free) Understanding Regression Analysis: A Conditional Distribution Approach
- Data Analysis Using Regression and Multilevel/Hierarchical Models

- (Free) Doing Bayesian Data Analysis

Chapter 2

Introduction

2.1 What is statistical modeling and what do we need this for?

Typically, one simplifies the complex reality (and loses information) in order to make it better understandable, mathematically treatable and to make predictions.

Underlying our models, there are theories which should be falsifiable and testable. For instance, I would be really surprised if I pull up my multimeter and measure the voltage (V) and electric current (I) at a resistence (R) in a circuit and find that Ohm's law $V = IR$ is not true. This **law** can be tested over and over again and if one would find a single valid counterexample, the law would be falsified. It is also true that the law is probably not 100% accularate, but an extremely good approximation of reality. Real-world measurements carry measurement errors and when plotting the data, one would see that the data points might not lie exactly on a straight line. This is not a problem.

A statistical model is a mathematical framework that represents the relationships between variables, helping us understand, infer, and predict patterns in data. It acts as a bridge between observed data and the real-world processes that generated them. In health research, where variability and uncertainty are inherent, statistical models are valuable tools for making sense of complex phenomena. You can watch this as short intro.

Depending on the task at hand, we would use different models. In any case, logical reasoning and critical thinking comes first, then comes the model. **It makes no sense to estimate statistical models just for the sake of it**.

All models are wrong, but some are useful. Or to quote George Box:

> "Since all models are wrong the scientist cannot obtain a 'correct' one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity."

In my opinion, statistical modeling is an art form: difficult and beautiful.

One goal of this course is to improve interpretation and limitations of statistical models. They are not magical turning data into truth. Firstly, the rule gargabe in, garbage out (GABA) applies. Secondly, statistical models are based on data and their variability and have inherent limitations one cannot overcome even with the most sophisticated models. This is expressed for instance in the so-called bias-variance trade-off. You can't have it all.

2.1.1 Explanatory vs. Predictive Models

I can recommend reading this article by Shmueli et al. (2010) on this topic.

Statistical models serve different purposes depending on the research question. Two primary goals are **explanation** and **prediction**, and each requires a different approach:

Explanatory Models focus on understanding causal relationships. These models aim to uncover mechanisms and answer **"why"** questions. For example:

- Does smoking increase the risk of lung cancer? **Yes**. (If you want to see what a large effect-size looks like, check out this study.)
- How large is the "effect" of smoking on lung cancer? **Large**.
- Does pain education and graded sensorimotor relearning improve disability (a question we ask in our Resolve Swiss project)?

Explanatory models are **theory-driven**, designed to test hypotheses. Here, one wants to understand the underlying mechanisms and the relationships between variables and hence often uses (parsimonious) models that are more interpretable, like linear regression.

Predictive Models prioritize forecasting future outcomes based on patterns in the data. These models aim to answer **"what will happen?"** For instance:

- Gait analysis using Machine Learning (ML)?
- Skin cancer detection using neural networks?

Predictive models are **data-driven**, often using complex algorithms to achieve high accuracy. Their success is measured using metrics like Root Means Square

Error (RMSE), Area Unter the Curve (AUC), or **prediction error on new, unseen data**. Any amount of model complexity is allowed. One could for instance estimate a neural network ("just" another statistical model) with many hidden layers and neurons in order to improve prediction quality. Interpretability of the model weights is not a priority here.

While explanatory and predictive goals often complement each other, their differences highlight the importance of clearly defining the purpose of your analysis. In applied health research, explanatory models help identify causal mechanisms, while predictive models can guide real-world decisions by providing actionable forecasts. Together, they enhance both our understanding of phenomena and our ability to make informed decisions in complex environments.

2.1.2 Individual vs. Population Prediction

Another important distinction is between **individual vs. population** prediction. In the smoking example above, we can be very sure about the mean effects that smoking has on lung cancer. On an individual level, it is harder to predict the outcome. Nevertheless, individual predictions will be (notably) better than random guessing. We will discuss this in greater detail.

2.1.3 Practical Use of Statistical Models

In my optinion, we should never be afraid to test our statistical models (as honestly as possible) against reality. We could for instance ask ourselves:

- "How much better does this model classify than the arithmetic mean? (i.e., the linear model with just an intercept)"
- "How much better does this model classify than random guessing?"
- Is it worth the effort to collect data and estimate this model by using hundreds of hours of our time?

In some cases, these questions can be answered straightforwardly.

- In advertising (Google, Facebook, …), a couple of percentage points in prediction quality might make a difference of millions of dollars in revenue offsetting the statistitians salary.
- Improved forecasts of a few percentage points in the stock market or just being slightly better than the average, will make you faboulously rich.
- Improved cancer forecasting might save lives, money and pain and is not only measured in money.

2.1.4 Start at the beginning

What do we actually want to do in general? Very broadly speaking we want to: **describe** the association of variables to each other that carry variability. Hence, the relationship is not deterministic like

$$y = 2x + 3$$

but rather we need to "loosen up" the relationship to account for variability (in x and y). So, the values 2 and 3 are not fixed but aflicted with uncertainty. Depending on your philosophical view, you might say you want to find the "true" but unknown relationship between variables. This is what we do in simulation studies all the time: We know the true relationship, simulate data by adding variability and then try to estimate the true relationship we assumed in the first place. For some practical applications, we can get a really nice and complete answer to our question (for instance sample size for proportions).

So we are looking for a function f such that

$$Y = f(X)$$

where

- Y is the "outcome", "dependent variable" or "response".
- X are the "predictors". X can be a single Variable x or many variables $x_1, x_2, ..., x_p$.

It is important to be aware of the notation here: "Predict" does **not necessarily** mean that we can predict the value in the future. It merely means we estimate the value (or mean) of Y given X.

- This can be done at the same time points, known as **cross-sectional** analysis ("What is the maximum jumping height of a person given their age at a certain point in time, whereas both variables are measured at the same time?");
- or at different time points, known as **longitudinal analysis** ("What is the maximum jumping height of a person 10 years later (t_2) given their baseline health status at time t_1?").

The **simplest statistical model** would be the mean model where Y is "predicted" by a constant: $Y = c$ which (at least in the classical linear regression) turns out to be $c = \bar{x}$. This simple model is often surprisingly good, or, to put it in other words, models with more complexity are often not that much better with regards to multiple metrics.

2.2 A (simple) model for adult body heights in the Bayesian framework

As repetition, read the parts about Bayes statistics from QM1 again to refresh your memory about the Bayesian framework.

It's recommendable to read the beginning of the book Statistical rethinking up until page 39 as well. We are not completely new to the topic of Bayes due to QM1.

We want to **start building our first model** right away.

Let's begin with the example in Statistical rethinking using data from the !Kung San people starting on page 79.

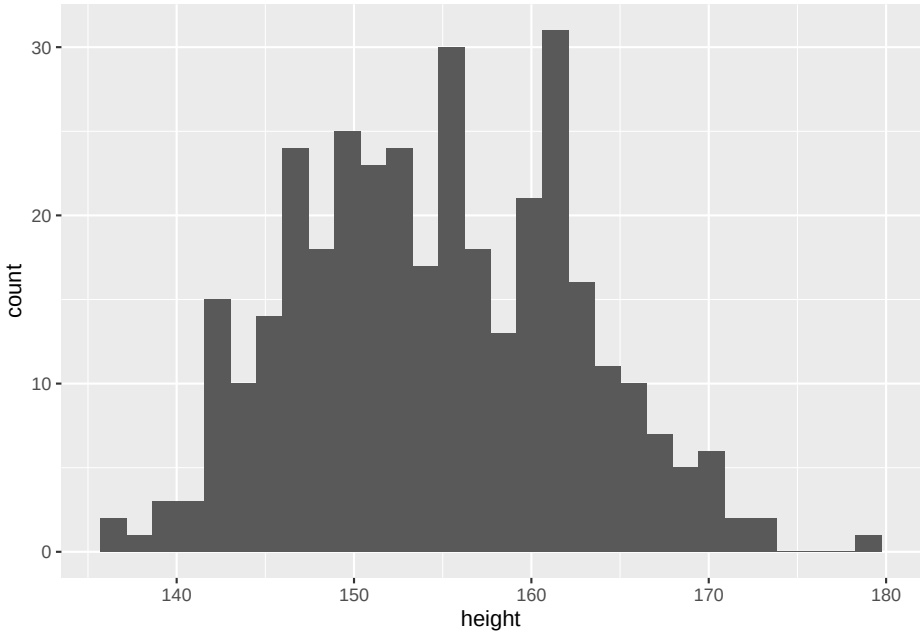
```
library(rethinking)
data("Howell1")
d <- Howell1
str(d)
```

```
## 'data.frame':    544 obs. of  4 variables:
##  $ height: num  152 140 137 157 145 ...
##  $ weight: num  47.8 36.5 31.9 53 41.3 ...
##  $ age   : num  63 63 65 41 51 35 32 27 19 54 ...
##  $ male  : int  1 0 0 1 0 1 0 1 0 1 ...
```

```
d2 <- d[d$age >= 18, ] # only adults
```

We want to model the adult height of the !Kun San people using prior knowledge (about the Swiss population) and data.

```
library(tidyverse)
d2 %>% ggplot(aes(x = height)) + geom_histogram()
```



Since we already have domain knowledge in this area, we can say that heights are usually normally distributed, or at least a mixture of normal distrubutions (female/male). We assume the following model:

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

As in QM1, we want to start with a Bayesian model and hence, we need some priors.

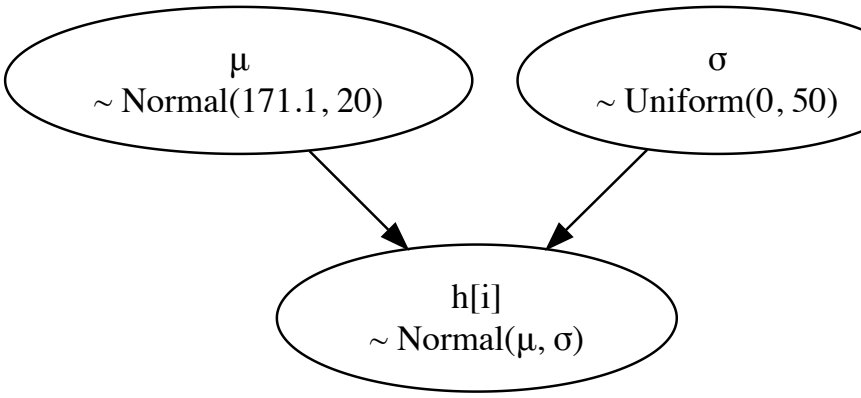
Since we are in Switzerland and just for fun, we use the mean of Swiss body heights as expected value for the **prior for the mean**. According to the link (Bundesamt für Statistik), the mean height of $n = 21,873$ people in the Swiss sample is 171.1 cm. We choose the same σ for the prior of the normal as in the book not to deviate too much from the example at hand.

Next comes our **model definition in the Bayesian framework**, which I often find more intuitive than the frequentist approach:

$$\begin{aligned} h_i &\sim \text{Normal}(\mu, \sigma) \\ \mu &\sim \text{Normal}(171.1, 20) \\ \sigma &\sim \text{Uniform}(0, 50) \end{aligned}$$

Description of the model definition: The heights are normally distributed with unknown mean and standard deviation. As our current knowledge about the mean height, we use a prior distribution for the mean (we do not know but

want to estimate) by assuming the mean of a population we know and a standard deviation of 20 cm which allows are rather large range of possible values for μ. σ is also unknown and a priori we restrict ourselves to values between 0 and 50 cm, whereas we assign equal plausibility to all values in this range (which can and should be critically discussed).



Vizualisation of the model structure:

Mind that there is a conceptual difference between the normal distribution of the heights and the normal prior distribution of the mean. The latter expresses our prior knowledge/insecurity about the unobserved mean. The normal distribution says we expect the heights to be normally distributed but we do not know the parameters (μ and σ) yet.

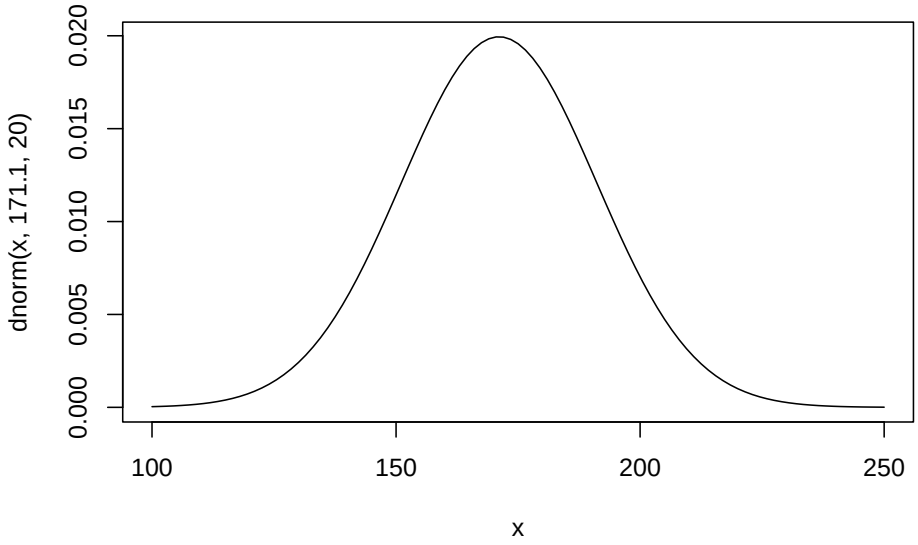
Of course we would not need the prior here due to the large sample size, but let's do it anyways for demonstration purposes. We are not completely uninformed about body heights and express our knowledge with the prior for μ. The 20 in the prior for the mean expresses our range of possible true mean values and aknowledge that there are a variety of different subpopulations with different means.

Using the Swiss data in the link one could estimate that the standard deviation of the heights from $21,873$ Swiss people is around is 25.6553 cm (Exercise 1).

Remember, in the Baysian world, there is no **fixed but unknown** parameter, but instead we define a distribution over the unobserved parameter.

We **visualize the prior for** μ.

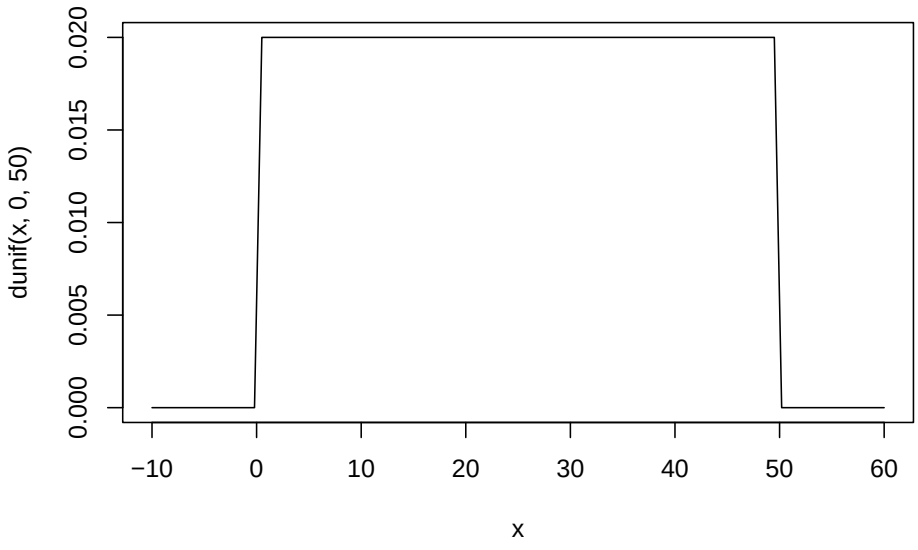
```
curve(dnorm(x, 171.1, 20), from = 100, to = 250)
```



The **prior for** σ is uniform between 0 and 50 cm. This is a very wide prior and just constrains the values to be positive and below 50 cm. This could be stronger of course.

Visualization of the prior for σ:

```
curve(dunif(x, 0, 50), from = -10, to = 60)
```



Note, we didn't specify a prior probability distribution of heights directly, but once we've chosen priors for μ and σ, these imply a prior distribution of individual heights.

Without even having seen the **new data**, we can check what our prior (model) for heights would predict. This is important. If the prior already predicts impossible values, we should reconsider our priors and/or model.

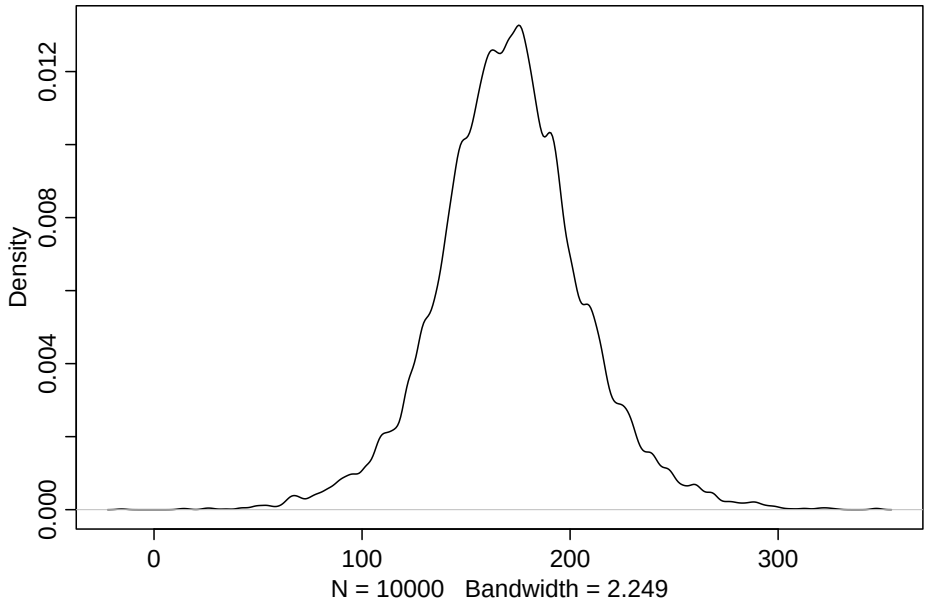
So, we simply draw μ and σ from the priors and then draw heights from the normal distribution using the drawn parameters.

Vizualisation of the prior for heights:

```
sample_mu <- rnorm(10^4, 171.1, 20)
sample_sigma <- runif(10^4, 0, 50)
prior_h <- rnorm(10^4, sample_mu, sample_sigma)
length(prior_h)
```

```
## [1] 10000
```

```
dens(prior_h)
```



The prior is not itself a Gaussian distribution, but a distribution of relative plausibilities of different heights, before seeing the data.

Now, there are a couple of different ways to estimtate the model incorporating the new data. For didactic reasons, grid approximation is often used (as in the book). For many parameters, grid approximation becomes more and more infeasible (due to combinatorial explosion).

We will skip that for now and use quadratic approximation instead which works well for many common procedures in applied statistics (like linear regression).

Later, you'll probably use (or the software in the background) mostly Markov chain Monte Carlo (MCMC) sampling to get the posterior. Pages 39 and the following explain the 3 concepts grid approximation, quadratic approximation and MCMC.

In short, **quadratic approximation** assumes that our posterior distribution of body heights can be approximated well by a normal distribution, at least near the peak.

Please read the addendum to get a clearer picture of what a bivariate normal distribution is.

Using the library `rethinking` we can estimate the model using quadratic approximation. First, we define the model in the `rethinking` syntax (see R code 4.25 in the book).

```
library(rethinking)
flist <- alist(
  height ~ dnorm(mu, sigma),
  mu ~ dnorm(171.1, 20),
  sigma ~ dunif(0, 50)
)
```

Then we estimate/fit the model using quadratic approximation.

```
m_heights <- rethinking::map(flist, data = d2)
```

Now let's take a look at the fit *maximum a posteriori* model:

```
precis(m_heights)
```

```
##             mean        sd       5.5%      94.5%
## mu    154.604101 0.4119941 153.945655 155.262548
## sigma   7.731329 0.2913856   7.265638   8.197019
```

Above, we see the mean of the posterior for μ **and** σ; and a 89% credible interval for those parameters.

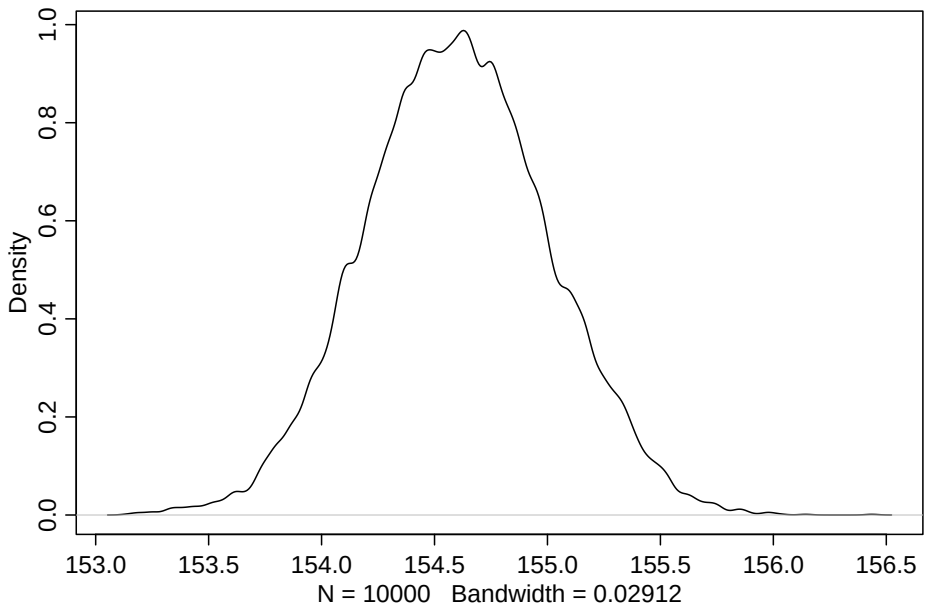
We can now plot the posterior distribution of the mean and the standard deviation.

```
post <- extract.samples(m_heights, n = 10^4)
head(post)
```

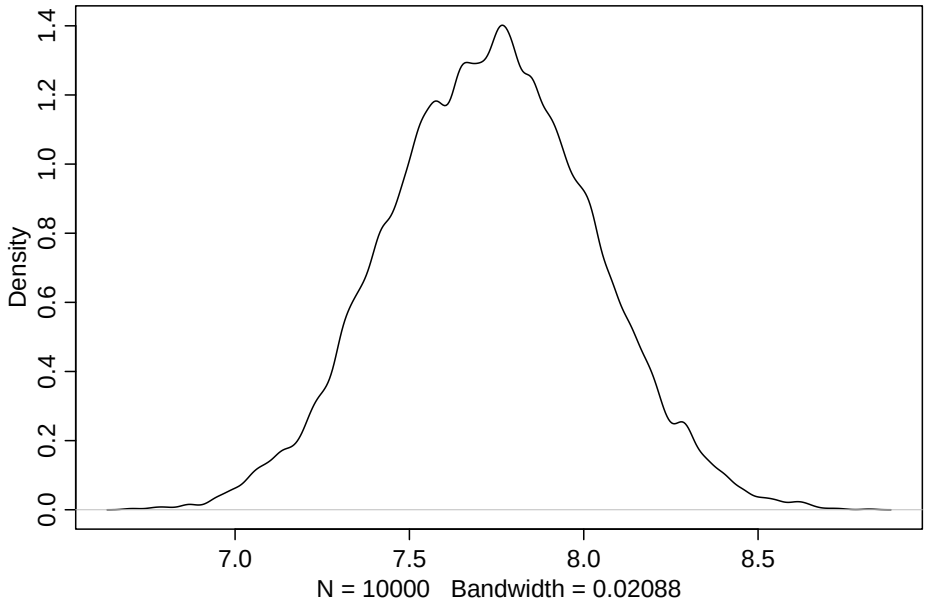
```
##         mu    sigma
```

```
## 1 153.8285 7.662522
## 2 154.6729 7.813708
## 3 154.2961 7.237938
## 4 154.2593 6.768533
## 5 155.0438 7.594867
## 6 154.4195 7.442121
```

```
dens(post$mu)
```




```
dens(post$sigma)
```



Note, that **these samples come from a multi-dimensional posterior distribution**. In our case, we approximated the posterior with a bivariate normal distribution. They are not necessarily independent from each other, but in this case they are. We know this from the model definition above. μ and σ are both defined as normal respectively uniform distributions and by definition do not influence each other. This is also visible in the vizualisation of the model structure: There is no confounding variable or connection between those priors. One could think of a common variable Z that influences both μ and σ. This could be genetic similarity which could influence both μ and σ.

Let's verify that μ and σ are uncorrelated:

```
vcov(m_heights)
```

```
##                mu        sigma
## mu    0.1697391350 0.0001538681
## sigma 0.0001538681 0.0849055527
```

gives you the variance-covariance matrix of the parameters of the posterior distribution. In the diagonal you see the variance of the parameters.

```
diag(vcov(m_heights))
```

```
##         mu      sigma
## 0.16973913 0.08490555
```

And we can compute the correlation matrix easily:

```
cov2cor(vcov(m_heights))
```

```
##                mu       sigma
## mu    1.000000000 0.001281709
## sigma 0.001281709 1.000000000
```

Let's plot the posterior in 3D, because we **can**:

WebGL is not supported by your browser - visit https://get.webgl.org for more info

How beautiful ist that?

We see in the 3D plot, that the "mountain" is not rotated, indicating graphically that the parameters are independent from each other.

We also see in the correlation matrix, the correlation of the parameters is ~ 0. In the context of a joint normal distribution, this means that the parameters are independent.

And, it is not an accident that the posterior looks like this. Using quadratic approximation, we used the bivariate normal distribution to **approximate** the posterior.

2.3 Classical approach for the simplest model

We have seen, how we could use domain and prior knowledge to fit a very simple model for body heights of a population (!Kung San) in the Bayesian framework.

Now, let's start at the same point in the classical framework. Here, we do not use any prior knowledge, at least not that explicitely.

The classical approach to fit a regression line is the so-calle least squares method.

The model is:

$$Y_i = height_i = c$$

for some $c \in \mathbb{R}$. But what is c? We determine the shape of the model ourselves (constant model, or mean model) and then estimate the parameter c.

We choose the c which minimizes the sum of squared errors from the actual heights. This has the advantage that deviations upper and lower from the actual height are equally weighted. The larger the deviation the (quadratically) larger the penalty.

SSE (Sum of Squared Errors) $= (height_1 - c)^2 + (height_2 - c)^2 + ... + (height_n - c)^2$

2.4 Exercises

2.4.1 [E] Exercise 1

Use the Swiss body heights data to determine - the 95% "Vertrauensintervall" for μ and - calculate the standard deviation of the heights from $21,873$ Swiss people.

2.5 Addendum

2.5.1 The bivariate normal distribution

As a refresher, you can look into the old QM1 script and read the chapter "4.7 Gemeinsame Verteilungen". Maybe this video also helps.

The bivariate normal distribution is a generalization of the normal distribution to two dimensions. Now, we look at the distribution of two random variables X and Y **at the same time**.

Instead of one Gaussian bell curve, we have a 3D bell curve. This curve defines how plausible different combinations of X and Y are.

Single points (like (3,6)) still have probability zero, because now the **volume** over a single point (x, y) is zero. The probability of a certain area is now the **volume** under the curve compared to the **area** under the density curve in the one-dimensional case.

Example: The following plot shows the density of a bivariate normal distribution of two variables X and Y with $\mu_X = 0$, $\mu_Y = 0$, $\sigma_X = 1$, $\sigma_Y = 1$ and $\rho = \frac{2}{3}$.

Below is the correlation matrix of the bivariate normal distribution.

```
##           [,1]      [,2]
## [1,] 1.0000000 0.6666667
## [2,] 0.6666667 1.0000000
```

If you move the plot around with your mouse, you see that there is a positive correlation between X and Y ($\rho = \frac{2}{3}$). This means that if X is above its mean, Y is also more likely to be above its mean. The variances of X and Y are both 1. That means, that if you cut through the plot in $X = 0$ or $Y = 0$, you see the same form of normal distribution. If you look at if from above, we have hihglighted the section on the surface over the area $X \in [0.5, 2]$ and $Y \in [0.5, 2]$. The volume over this area under the density curve is the probability of this area: $P(X \in [0.5, 2] \text{ and } Y \in [0.5, 2])$

Chapter 3

Literature

Here is a review of existing methods.

Chapter 4

Methods

We describe our methods in this chapter.

Math can be added in body using usual syntax like this

4.1 math example

p is unknown but expected to be around 1/3. Standard error will be approximated

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{1/3(1-1/3)}{300}} = 0.027$$

You can also use math in footnotes like this[1].

We will approximate standard error to 0.027[2]

[1] where we mention $p = \frac{a}{b}$

[2] p is unknown but expected to be around 1/3. Standard error will be approximated

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{1/3(1-1/3)}{300}} = 0.027$$

Chapter 5

Applications

Some *significant* applications are demonstrated in this chapter.

5.1 Example one

5.2 Example two

Chapter 6

Final Words

We have finished a nice book.