Part 4: Models

Lecture: Introduction to Statistics

Winter term 2019/2020

Part 4: Models

Lecture: Introduction to Statistics

Winter term 2019/2020

Contents

1 Overview

In the first section we introduce the *model chapter* based on the debate between *Frequentists* and *Bayesians*. After a short overview and background with the main ideas of each approach we formalize the conceptual ideas and the components of the presented model. We finish by discussing three models: *binomial*, *factorial-design* and *simple linear regression* model.

2 Two notions of probability

What are probabilities? We look at two viewpoints: probabilities exist "outside in the world" or they are "subjective beliefs" [@kruschke2015]. Although both notions imply different approaches how to deal with probabilities, the mathematical properties are quite similar [@kruschke2015].

2.1 Frequentism — Probabilities as properties of the world

We now assume that probabilities are existent and properties of the world. When observing a particular object, then in the long run, we should be able to deduce from our observations the "inherent" probability of an outcome of this object. Frequentism is an approach that searches for relative frequencies in a large number of trials [@vallverdu2015]. Tossing a coin results in head or tail (for simplicity we do not consider the possibility that it stands on its edge). A Frequentist would toss a coin many times, take a note on each observation and, finally, calculate the relative frequency of each outcome. That is, the observed outcome divided by the number of tosses. In the long run, the outcome "head" or "tail" will be observed around 50% (given the coin is fair, the toss is not manipulated, no external factors like wind etc. influences the outcome, ...).

Therefore it is mandatory

> to obtain *a large number of trials* from which emerged the relative frequency from an event. [@vallverdu2015, p.50] (emphasis is taken from the original)

In other words: For Frequentists probabilities (relative frequencies) describe *objective properties of the world* rather than subjective beliefs.

2.2 Bayesianism — Probabilities as subjective beliefs

Another notion of probabilities is to think of them as "beliefs" inside our head. For example considering again the coin flip example, we could have a relatively strong belief that a coin is fair or we actually believe that a coin is fair but we are not so certain about that or we "just do not know" (see Fig. **??** from left to right).

2.2.0.1 Bayes' Theorem

The core of Bayesian methods is Bayes' theorem which tell us how prior belief is combined with observed data:

$$P(H|Data) = \frac{P(Data|H) * P(H)}{P(Data)}$$

or in plain language,

$$Posterior = \frac{Likelihood * Prior}{Marginal\ Likelihood},$$

where

- $P(H|Data)$ is the conditional probability (*posterior probability*) of the hypothesis H given a particular result (Data);

- $P(Data|H)$ is the conditional probability (*likelihood*) of a result(data) given the hypothesis H;
- $P(H)$ is the *prior probability* of the hypothesis H;
- $P(Data)$ is the *marginal (or average) likelihood*[1] of the data irrespective of the truth of any hypothesis.

The job of the "marginal Likelihood" in the denominator is to standardize the posterior, to ensure it sums up to one (integrates to one). Therefore, the key lesson of Bayes' theorem is [@mcelreath2015]:

$$Posterior \propto Likelihood * Prior$$

What does that mean? Consider again the coin flip experiment. We have observed a certain amount of coin flip trials, thus we can describe a probability distribution: a list of possible outcomes and their corresponding probabilities (Likelihood) [@kruschke2015]. Furthermore, as we have stated above, we express a prior belief about our hypothesis, for example we might belief, that the coin is a trick coin and therefore extremely biased. Again, the mathematical way of expressing our belief is by assigning numbers to a set of mutually exclusive events (the prior). From that, finally, the rules of probability theory define a uniquely logical posterior for the prior, likelihood and data (Posterior).

At this point we see an important difference between Frequentist and Bayesian methods: A parameter in Bayesian methods is conceptualized as a random variable with its own distribution (the posterior) that summarizes the current state of knowledge. The expected value of the posterior is the best guess about the true value of the parameter and its variability reflects the amount of uncertainty [@kline2013]. In Frequentist statistics, a parameter is seen as a constant that should be estimated with sample statistics [@kline2013].

3 Likelihood, Prior, & Posterior

So far we wanted to clearly state the most important conceptual differences between Frequentism and Bayesianism and to introduce the core of Bayesian methods: Bayes' theorem. Before we go further into the topic of modelling, some conceptual notions are neccessary.

3.1 Probability density function vs. Likelihood function

As we already know from the "probability"-lecture, for a coin flip the probability of each outcome can be described with the *Bernoulli distribution*[2], as there exist two discrete outcomes (head or tail) and a constant probability θ:

$$p([X=x]|\theta) = \theta^{[x]}(1-\theta)^{(1-[x])}$$

where

- θ is the probability of "head" for the coin flip;
- the bracket -[ ] indicates that we assume this parameter as unknown.

With the formula above we assume that we know the probability θ and derive from this the distribution of possible outcomes. But we might be interested in a different perspective, that is the assumption that we have the data fixed but θ is unknown: the *likelihood function* — a mathematical formula that specifies the plausibility of the data. It states the probability of any possible observation:

$$p(X=x|[\theta]) = [\theta]^{x}(1-[\theta])^{(1-x)}$$

In the first case, we assume the outcome to be an unknown parameter (variable), whereas in the latter case the unknown parameter is θ. Please be aware that through exchanging the roles of x and θ in the second equation (likelihood function) this function is no longer a probability distribution and thus does not integrate to 1.

[1] Sometimes also referred to as "Evidence"

[2] we need here a distinction between Bernoulli and Binomial...

3.2 Priors

General introduction to priors... A prior is a initial probability assignment for each possible value of the parameter [@mcelreath2015].

3.3 Posterior

4 Modeling

4.1 Introductory example

As introductory example a coin flip experiment is considered. The question is if a particular coin is *biased*. In order to investigate this question a coin is flipped x times (=trials) and the number of success (i.e. number of "head") k is recorded. This is repeated n times (=observations).

```
#load packages
library(tidyverse)
library(brms)
library(rethinking)
library(ggthemes)
```

```
#simulate coin flip data set
sample.space <- c(0,1)
theta <- 0.5              # probability of a success (here: head)
X <- 30                   # number of trials in the experiment
n <- 10                   # number of observations
k <- 0                    # number of heads [initialization]

## repeat experiment N-times
for (i in 1: n) {
  k[i] <- sum(sample(sample.space, size = X, replace = TRUE,
                     prob = c(theta, 1 - theta)))
}

## show results in a tibble
coin.flip <- tibble("n" = seq(from=1, to=n, by=1),
                    "k" = k,
                    "x" = X
) %>%
  print()
```

```
## # A tibble: 10 x 3
##        n     k     x
##    <dbl> <dbl> <dbl>
##  1     1    14    30
##  2     2    13    30
##  3     3    15    30
##  4     4    20    30
##  5     5    15    30
##  6     6    13    30
##  7     7    13    30
##  8     8    13    30
##  9     9    18    30
## 10    10    15    30
```

The above table shows the observed outcome, but how the underlying probability of coming up *heads* can be derived from that data set?

4.2 Steps of Data Analysis

The approach described here is based on [@mcelreath2015, @kruschke2015]. Although the approach is introduced in a Bayesian context, it can be used as a general guideline (with some caveats):

- Identify the relevant variables according to the hypothesis. (Measurement scales, predicted vs. predictor variables).
- Define the descriptive model for the relevant variables.
- likelihood distribution (distribution of each outcome variable that defines the plausibility of individual observations)
- parameters (define and name all parameters of the model in order to relate the likelihood to the predictor variable(s))
- Bayesian context: Specify a prior distribution.

Further steps that will be subject of later chapters:

- Inference and interpretion of the results.
- Model checking (Is the defined model adequate?)

The hypothesis for the introductory example: *The coin is fair, thus, the underlying probability of heads coming up is 0.50.*

First step is to *identify the relevant variables.* For the coin flip experiment a coin is flipped n times, wherby each observation consists of x trials. The variable *coin flip* Y is dichotomous with the possible outcomes "head" and "tail". For each observation the outcome is recorded: "0" for coming up tail and "1" for coming up head. The data are summarized for each observation. The variable k indicates the number of heads coming up in x trials.

In **the second step** a *descriptive model for the described variables* has to be defined. An underlying probability θ is assumed, indicating the probability of heads coming up $p(y = 1)$. The probability that the outcome is head, given a value of parameter θ, is the value of θ [@kruschke2015, p.109]. Formally, this can be written as

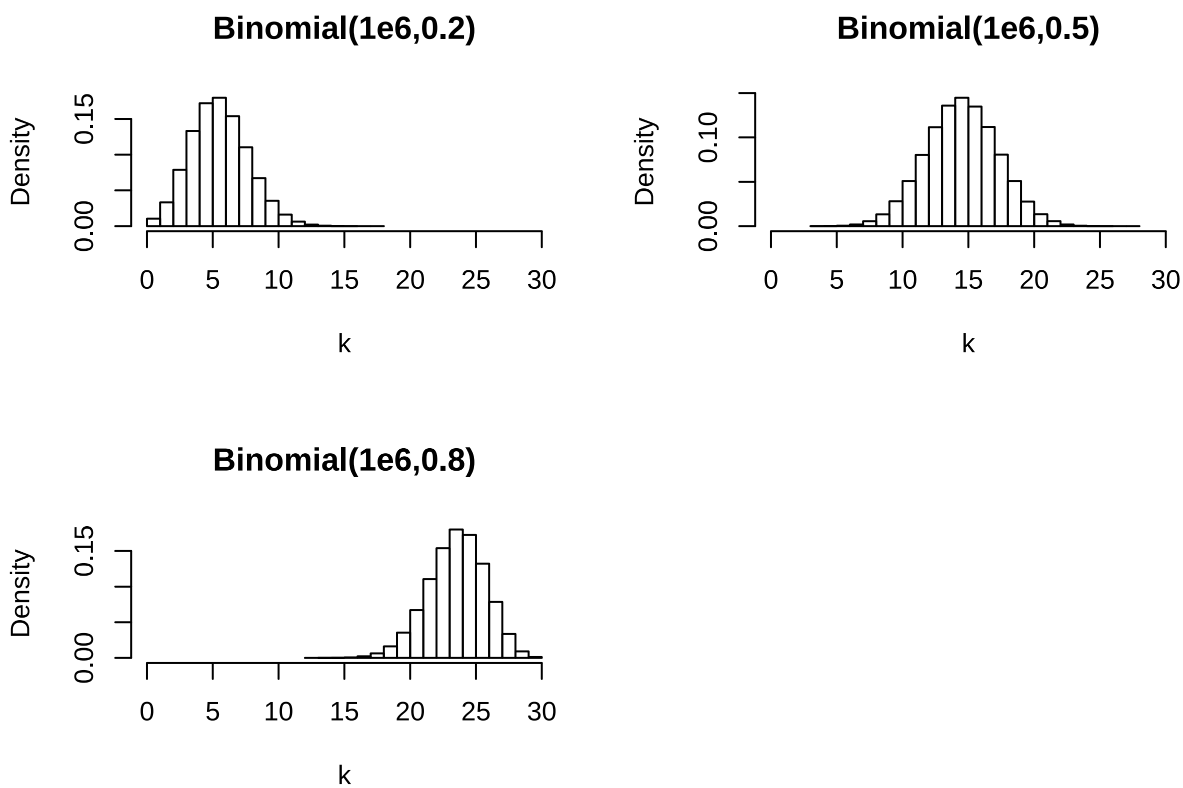
$$p(y = 1|\theta) = \theta$$

As only two outcomes of Y exists, the probability that the outcome is tail is the complementary probability $1 - \theta$. Both probabilities can be combined in one probability expression:

$$Pr(Y|n, \theta) = \frac{n!}{y!(n-y)!}\theta^y(1-\theta)^{n-y}.$$

This probability distribution is called the **Binomial distribution**. The fracture at the beginning indicates how many ordered sequences of n outcomes a count y have, therefore the important conceptional part is the latter one.

```
#Plot probability distribution: What would be the expected observed number
#of "head" given the underlying prob. theta?

par(mfrow=c(2,2))
## theta=0.2
hist(rbinom(n=1e6,size=30,prob=0.2), xlab="k", main="Binomial(1e6,0.2)",
     xlim = c(0,30), freq=FALSE)
## theta=0.5
hist(rbinom(n=1e6,size=30,prob=0.5), xlab="k", main="Binomial(1e6,0.5)",
     xlim = c(0,30), freq=FALSE)
## theta=0.8
hist(rbinom(n=1e6,size=30,prob=0.8), xlab="k", main="Binomial(1e6,0.8)",
     xlim = c(0,30), freq=FALSE)
```



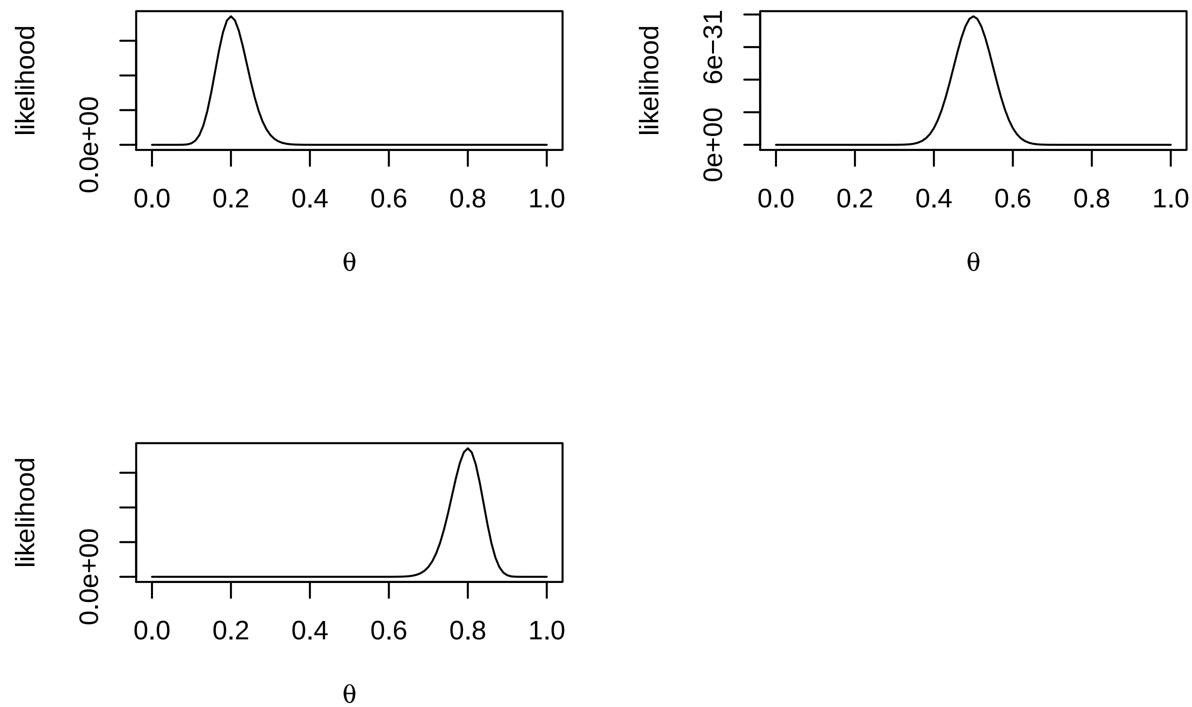
When the coin is flipped only once, then the probabilty can be written as:

$$Pr(Y|\theta) = \theta^y (1 - \theta)^{1-y}.$$

This special variant of the Binomial distribution is the so-called **Bernoulli distribution**. To see the connection to the first considerations: When the outcome "head" is observed the equation reduces to $Pr(y = 1|\theta) = \theta$ and when the outcome "tail" is observed the equation results in $Pr(y = 0|\theta) = (1 - \theta)$.

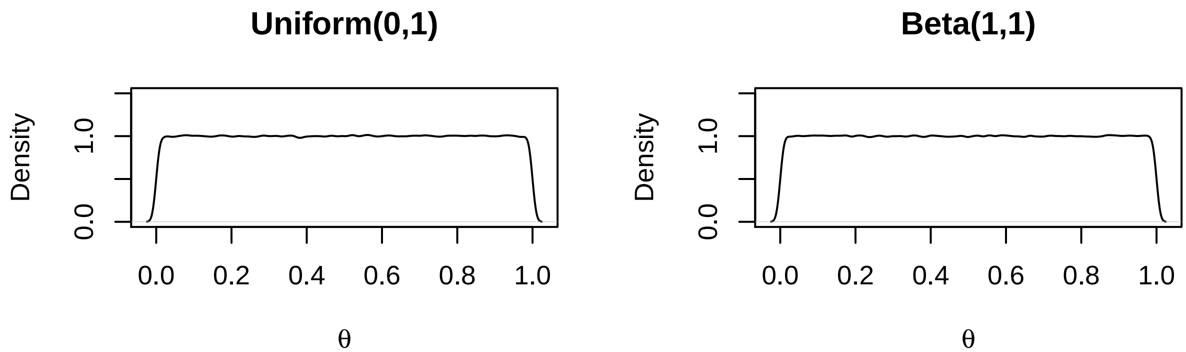
Accordingly, for the introductory example it can be noted that the coin flip variable Y *is distributed as* Binomial distribution. (Note: For Bayes' rule the *likelihood function* is needed. Remember, the likelihood function treats θ as unknow and the data as known, while this role of parameter is exchanged in a probability distribution.)

```
# calculate the Liklihood function
binomial.likelihood <- function(n, k, theta){  theta^k*(1-theta)^(n-k)}
theta <- seq(from=0, to=1, by=0.01)

# Plot likelihood: What would be the expected underlying prob. theta given
# observed number of "head" in 100 observations?
par(mfrow=c(2,2))
plot(theta,binomial.likelihood(100,20,theta), xlab=expression(theta),
     ylab="likelihood", type="l")
plot(theta,binomial.likelihood(100,50,theta), xlab=expression(theta),
     ylab="likelihood", type="l")
plot(theta,binomial.likelihood(100,80,theta), xlab=expression(theta),
     ylab="likelihood", type="l")
```



The third step is solely a *Bayesian idea*, that is the *incorporation of prior knowledge*. What do we believe about the coin bias θ before seeing the data? Assuming that no expectation about θ exists a priori, indicating that all values of θ between 0 and 1 are equally probable. This can be modeld by a uniform distribution or a Beta distribution with parameters a=1 and b=1 (see following figure).

```
# Modelling prior knowledge "ignorance"

par(mfrow = c(2, 2))
## simulated a uniform(0,1) distribution
dens(runif(n=1e6,min=0, max=1), ylim = c(0,1.5),
     xlab=expression(theta), main="Uniform(0,1)")
## simulates a beta(1,1) distribution
dens(rbeta(n=1e6,shape1=1,shape2=1), ylim = c(0,1.5),
     xlab=expression(theta), main="Beta(1,1)")
```



So far, the coin flip model is define conceptionally. In the following some notational considerations have to made.

5 Notation

5.1 Textual notation

In the textual notation, first the prior assumptions (if the Bayesian perspective is taken) are indicated. For the coin flip example this is:

$$\theta \sim Beta(1,1).$$

The symbol "$\sim$" means "is distributed as", thus, the above equation says before seeing the data all possible values of θ between 0 and 1 are assumed to be equally likely.

Subsequently, the descriptive model for the data has to be defined. As already described in the section above, it is assumed that the observed data (upcoming of heads k) are distributed as Binomial distribution with given n (number of observations) and unknown θ. This relation is denoted symbollically as

$$k \sim Binomial(\theta|n).$$

To summarize the current model (whereby the prior knowledge is only considered from a Bayesian perspective):

$$\theta \sim Beta(1,1),$$

$$k \sim Binomial(\theta|n).$$

5.2 Graphical notation

When models get very complex and incorporate many parameters it can be difficult to tease out all relations between the model components. In such a situation a graphical notation of a model might be helpful. In the following the convention described in Wagenmakers and Lee's *Bayesian Cognitive Modelling* (2014) is used: The graph structure is used to indicate dependencies between the variables, with children depending on their parents [@leeWagen2014]. General conventions:

- Nodes - problem relevant variables,
- shaded nodes - observed variables,
- unshaded nodes - unobserved variables,
- circular nodes - continuous variables,
- square nodes - discrete variables,
- single line - stochastic dependency, and
- double line - deterministic dependency.

For the introductory example this indicates:

- relevant variables: number of trials (n), number of success (k) and probability for a success (θ),
- observed variables: n and k,
- unobserved variables: θ,
- continuous variable: θ,
- discrete variables: n and k.

In the next step the dependencies have to be determined:

The number of success k depends on the probability of a success θ as well as on the number of trials n.

Finally, the graphical structure together with the textual notation can be represented:

6 Further examples

6.1 Difference between two groups

In the introductory example we asked for the underlying probility θ of a single coin that was flipped repeatedley. Consider now, that a second coin y_2 is introduced. One question that arises might be for example: *How different are the biases of the two coins?*

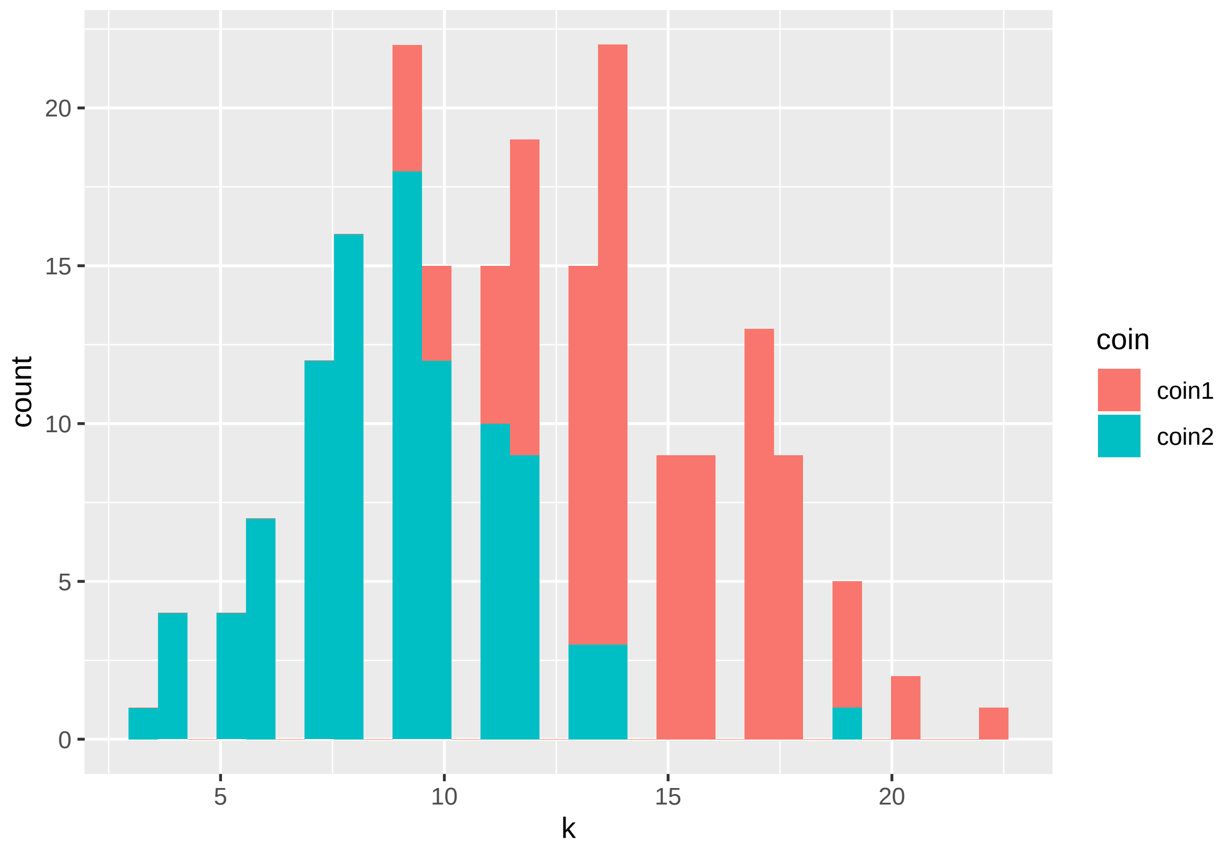
```
#simulate flipps of two coins
sample.space <- c(0,1)
##First coin:
theta1 <- 0.5              # probability of a success (here: head)
X1 <- 30                   # number of trials in the experiment
n1 <- 100                  # number of observations
k1 <- 0                    # number of heads [initialization]

for (i in 1: n1) {
  k1[i] <- sum(sample(sample.space, size = X1, replace = TRUE,
                     prob = c(theta1, 1 - theta1)))
}
##Second coin:
theta2 <- 0.7              # probability of a success (here: head)
X2 <- 30                   # number of trials in the experiment
n2 <- 100                  # number of observations
k2 <- 0                    # number of heads [initialization]

## repeat experiment N-times
for (i in 1: n2) {
```

```r
  k2[i] <- sum(sample(sample.space, size = X2, replace = TRUE,
                     prob = c(theta2, 1 - theta2)))
}

## show results in a tibble
coin.flip2 <- tibble("coin" = c(replicate(n1,"coin1"),replicate(n2,"coin2")),
                    "n" = c(seq(from=1, to=n1, by=1),seq(from=1, to=n2, by=1)),
                    "k" = c(k1,k2),
                    "x" = c(replicate(n1,X1),replicate(n2,X2))
) %>%
  print()
```

```
## # A tibble: 200 x 4
##    coin      n     k     x
##    <chr> <dbl> <dbl> <dbl>
##  1 coin1     1    17    30
##  2 coin1     2    13    30
##  3 coin1     3     9    30
##  4 coin1     4     9    30
##  5 coin1     5    15    30
##  6 coin1     6    15    30
##  7 coin1     7    10    30
##  8 coin1     8    12    30
##  9 coin1     9    14    30
## 10 coin1    10    16    30
## # ... with 190 more rows
```

```r
#Plotting the observed results
ggplot(data=coin.flip2,mapping = aes(x=k, fill=coin ))+
          geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



6.1.1 Conceptual steps for modeling

We suppose that the underlying probabilities of the two coins correspond to *different* latent variables θ_1 and θ_2.

First step is again the *identification of the relevant variables* according to the research question. As already indicated for the "one coin" example we have:

- the observed number of heads k_1 and k_2 (for each coin, respectively), which is influenced by
- the number of observations n_1 and n_2 and by
- the underlying probabilities θ_1 and θ_2.

Furthermore, from a conceptional perspective, we are interested in the *difference between the coin biases*. Therefore a further variable will be introduced δ, defined by:

$$\delta = \theta_1 - \theta_2.$$

The *distributional assumptions*, according to the **second and third step**, can be adopted from the "one coin" example, such that the graphical notation (including the textual notation) can be denoted as follows:

6.1.2 Notation Beta-Binomial Model - Two Groups

6.2 Simple linear regression with one metric predictor

The following example originates from a data set in which speed of cars and the distance taken to stop was recorded. It is a simple data set good for introducing the basic ideas for simple linear regression.

```r
#The "cars" data set
data(cars)
#take a look at the variables included in the data set
str(cars)
```

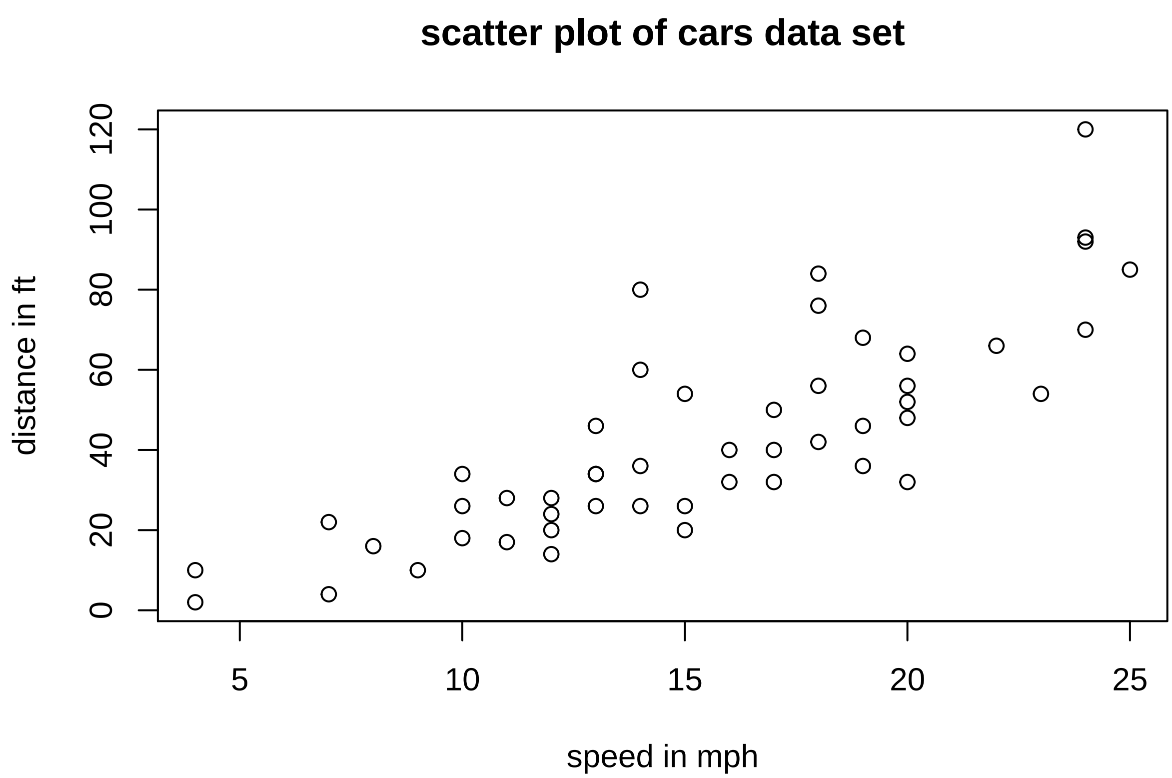
```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

One possible question could be how much the stopping distance increases when the speed of a car increases.

6.2.1 Conceptual steps for modeling

First step is to **identify the relevant variables**. In this case these are "speed" measured in mph and "distance" measured in ft, thus, both variables are metric variables. As distance will be predicted from speed. The *predicted variable* is "distance" and the *predictor variable* is "speed". A scatter plot can visualize a possible relationship between both variables.

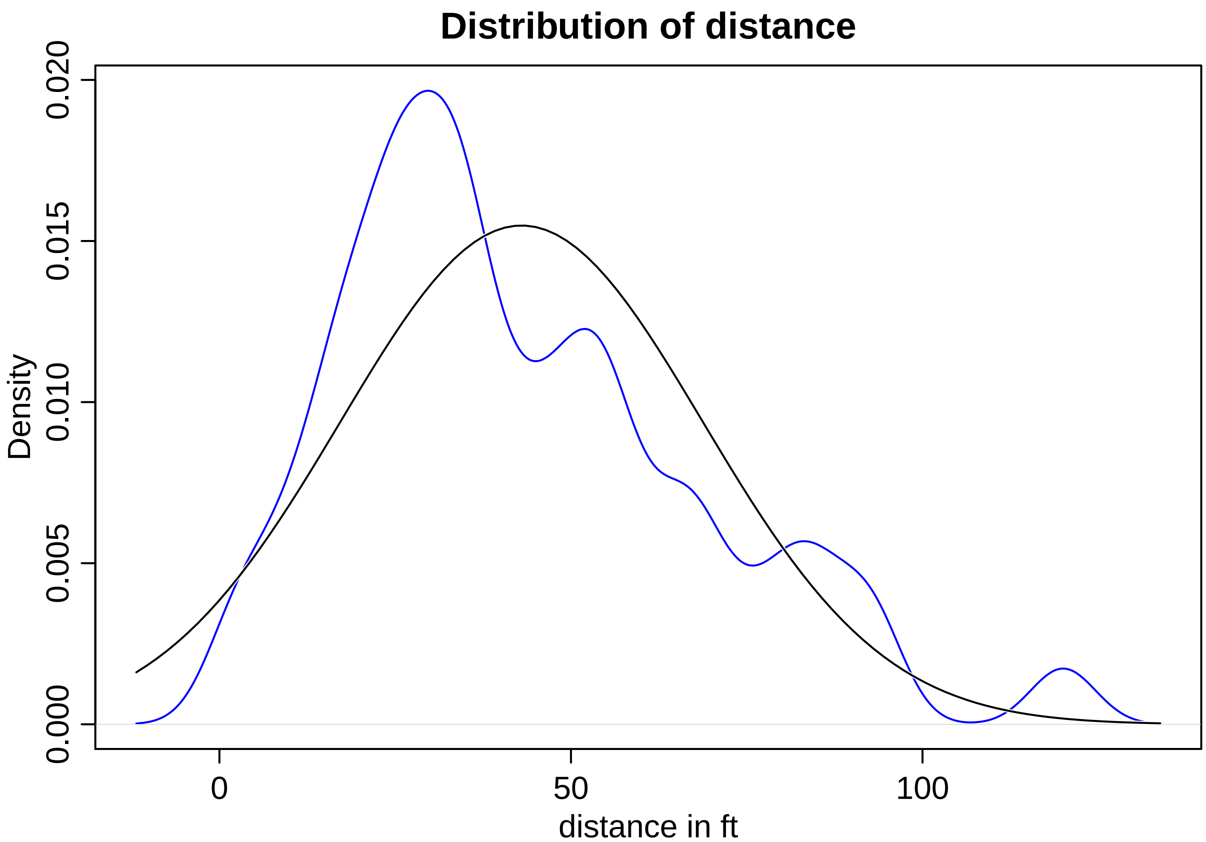
```r
plot(x=cars$speed,y=cars$dist, type="p", main="scatter plot of cars data set",
     ylab="distance in ft", xlab="speed in mph")
```



Next step is to define a **descriptive model of the data**. According to the scatter plot it is not too absurd to think that distance might be proportional to speed. Therefore, a linear relationship between both variables can be assumed, where speed is used i order to predict distance. But how can the distribution of the predicted variable "distance" be described? The following plot shows in blue the density of the actual distance values.

```r
#density of distance values in blue
#(in black simulation of a normal distribution)
dens(cars$dist, col="blue", norm.comp = TRUE, main="Distribution of distance",
```

```
    xlab="distance in ft")
```



Although the distribution of "distance" values is not identical to the corresponding normal distribution, it can be assumed that the values follow a *normal distribution*. The underlying consideration is that the distance values y_i are distributed randomly according to a normal distribution around the predicted value $\hat{y}$ and with a standard deviation denoted with σ. This can be denoted as:

$$y_i \sim Normal(\mu, \sigma).$$

The index i indicates each element (i.e. car) of the list y, which in turn is the list of distances.

6.2.1.1 Small excursos: "iid"

The short model description above incorporates often already an assumption about the distribution of distance-values: *identically and independently ditributed*. Often the abbreviation *iid* can be found for this assumption:

$$y_i \overset{\text{iid}}{\sim} Normal(\mu, \sigma).$$

The abbreviation *iid* indicates that each value y_i has the same probability function, independent of the other y values and using the same parameters [@mcelreath2015].

In the third step, a Bayesian perspective is taken the **prior knowlege** (before seeing the data) has to be defined. The parameters of the current model are the predicted value μ and the standard deviation σ. For the parameter μ a normal distribution can be assumend with parameters that reflect the estimated values from the sample.

```
#descriptive statistics from the sample
tibble(variables=c("speed", "distance"),
```

```
       mean=c(mean(cars$speed),mean(cars$dist)),
       sigma = c(sd(cars$speed), sd(cars$dist)))
```

```
## # A tibble: 2 x 3
##   variables  mean sigma
##   <chr>     <dbl> <dbl>
## 1 speed      15.4  5.29
## 2 distance   43.0 25.8
```

$$\mu \sim Normal(43, 26)$$

For the standard deviation σ a uniform distribution is assumed:

$$\sigma \sim Uniform(0, 40)$$

6.2.2 Notation Simple Regression model

7 Further elaboration on modeling (in anticipation of the topic "estimation")

7.1 Beta-Binomial model - one group (revisited)

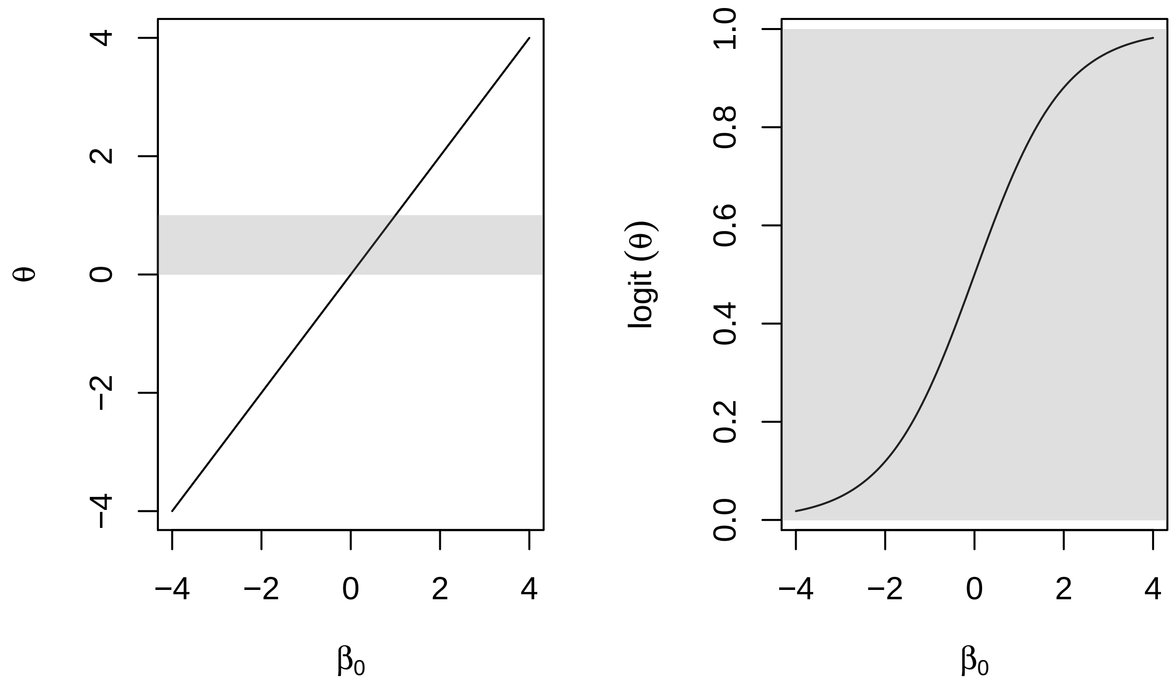
Sofar the existence of the underlying probability θ for observing head as outcome of a coin flip has been discussed. But the estimation of θ has been ignored until yet. Although "estimation" will be topic of next chapter, it is helpful at this point to discuss the introduced models further. In order to estimate θ *parameter(s)* are needed. When it comes to estimation exactly this/these parameter(s) will be the result(s), therefore is is important to see already the connection to the models that were developed in this chapter.

For the coin flip example the value of interest is the underlying probability, thus, only one parameter is needed:β_0. (Note: Latin letters are used when we refer to the sample, greek letters are used when we refer to the population.)

How is β_0 linked to the latent variable θ?

Considering for example the simplest case: a *linear relationship* (see next plot left side). The problem which arises at this point is that θ represents a probability, and is therefore bounded to the range 0-1 (grey shaded area).

```
#Different relationships between the parameter and expected value
x <- seq(from=-4, to=4, length.out = 100)
y <- x                    #linear relationship
y.log <- logistic(x)   #logistic relationship

par(mfrow = c(1, 2))   #set both plot beside each other
plot(x,y,type="l", ylab=expression(theta), xlab=expression(beta[0]))
rect(-5,0,5,1,col = rgb(0.5,0.5,0.5,1/4), border = NA)
plot(x,y.log,type="l", ylab=expression(logit~(theta)), xlab=expression(beta[0]))
rect(-5,0,5,1,col = rgb(0.5,0.5,0.5,1/4), border = NA)
```



A mathematical transformation is needed such that the parameter β_0 can take any value while θ is bounded to the range 0-1. One transformation that offers exactly this possibility is the *logit link function* (see aboth plot right side)

$$logit(\theta) = \beta_0.$$

As the underlying assumption maps the parameter to the latent variable θ (and not the other way around) from a conceptional point of view the *inverse link function* is more appropriate, which is the *logistic link* in this case:

$$\theta = logistic(\beta_0).$$

It is defined as

$$\theta = \frac{exp(\beta_0)}{1 + exp(\beta_0)}.$$

Both expression, *logit* and *logistic* link achieve mathematically the same result but it is conceptionally just a different matter of emphasis [@kruschke2015].

7.1.1 Notation of beta-binomial model - one group (revisited)

The current descriptive model incorporates the idea that parameter β_0 is estimated from the given sample. It defines the latent variable θ. The parameter is maped to θ by a logistic link function. The underlying probability θ designates the observed number of upcoming heads. The number of upcoming heads in turn, is assumed to be distributed as Binomial distribution.

7.2 Beta-Binomial model - two groups (revisited)

In the above model for two coins the latent variable δ was already introduced. It is defined by the difference between the underlying probabilities $\theta_1 - \theta_2$. Which parameters should be used in order to estimate the difference between both groups? As we will see, it turns out that the same mathematical form can be used, as one would use for simple linear regression:

$$\theta_j = \beta_0 + \beta_1 * X_{Group_j},$$

$$\text{with } X_{Group_j} = \begin{cases} 0, \text{if coin 2,} \\ 1, \text{if coin 1.} \end{cases}$$

Considering *coin 2*, the above equation would result in

$$\theta_2 = \beta_0,$$

which is the *intercept* and indicates the proportion of head coming up for coin 2.

Considering by contrast coin 1, then the equation would result in:

$$\theta_1 = \beta_0 + \beta_1.$$

The proportion of coming up head for coin 1 has to be calcuated by summing up the *intercept* β_0 and the *slope* β_1.

Taken togehter: *What is the interpretation of the slope* β_1*?* The difference $\delta = \theta_1 - \theta 2$ is

$$\theta_1 - \theta_2 = (\beta_0 + \beta_1) - \beta_0 = \beta_1 = \delta,$$

the slope β_1, thus, we can see that this parameterization enables us to estimate the difference between two groups. When in comes to estimation and interpretation the results will be the intercept b_0 and the slope b_1.

7.3 Simple linear regression model (revisited)

–

8 References