```r
str(res)
```

```
'data.frame':	1950 obs. of  2 variables:
 $ ID        : num  1 1 1 1 1 1 1 1 1 1 ...
 $ feb_1_2015: num  0.884 -0.421 -0.302 -0.306 -0.351 ...
```

Exercise 3: Wokring with raster data

Calculate the mean and standard deviation of elevation for each German state.

4 Random Variables

4.1 Preparation

- Read: Altman, D. G., and J. M. Bland. 1999. "Statistics Notes: Variables and Parameters." BMJ (Clinical Research Ed.) 318 (7199): 1667.

4.2 The basics

Experiment The process of observation or measurement. This could be asking a person for their political opinion, measuring the DBH (diameter at breast height) of a tree, or counting the number of plants of a given species in a predefined plot.

Outcome The result of an experiment is called an outcome.

Sample space S This is the set of all possible outcomes of an experiment. And each outcome in the sample space S is called an **element of the sample space**.

For example, if an experiment consists of one role of a die, then the sample space would be:

$$S_1 = \{1, 2, 3, 4, 5, 6\}$$

A sample space is said to be **discrete** if it contains a countable number of elements. A sample space is said to **continuous** if the number of elements of the sample space are infinite and uncountable.

Events Often we are not interested in a single outcome but in a set of outcomes (= **event**) from a sample space.

For example, an event an event B occurs when the number of a die is divisible by 3. For S_1, event $A = \{3, 6\}$

- $P(A) \geq 0$
- $P(S) = 1$

4.3 Basic rules for probabilities

1. $0 \leq P(A) \leq 1$ for any event A.
2. $P(S) = 1$
3. Addition rule: If events A and B are disjoint[1], then $P(A \text{ or } B) = P(A) + P(B)$.
4. Complement rule: For any event A, $P(A^c) = 1 - P(A)$.
5. Multiplication rule: If events A and B are independent[2], then $P(A \text{ and } B) = P(A)P(B)$.

Exercise 1: Probabilities

A probability distribution is, simply speaking, *a list of all mutually exclusive outcomes of a random trial.* A very simple example is a dice roll, or a coin toss. Note, the probability of all outcomes **must** sum to 1.
A useful function to get all possible outcomes of an experiment is the function `expand.grid()`. If the experiments is to flip a coin three times, then we can easily get all possible combinations with:

```r
expand.grid(toss1 = c("H", "T"), toss2 = c("H", "T"),
            toss3 = c("H", "T"))
```

```
  toss1 toss2 toss3
1     H     H     H
2     T     H     H
3     H     T     H
4     T     T     H
5     H     H     T
6     T     H     T
7     H     T     T
8     T     T     T
```

Consider the very simple random trial of rolling four dices. Use R to answer the following questions:

1. What is the probability that the sum of the four dices is more than 10?
2. What is the probability that the sum of the four dices is less or equal to 5?

[1]Events are said to be **disjoint** if the have no outcomes in common.
[2]Events are **independent**, if knowing that one event occurs does not change the probability of other events.

4.4 Random variables

Random variable (RV) We can think of a random variable as a variable that probabilistically takes a value.[3]

- RV take on values, have types and domains.
- RV are usually denoted with capital letters (e.g., X, Y), while lower case letters (e.g., x, y) are used to denote a specific value of a RV.
- For example, we write $P(X = x)$ to denote the probability that the RV X is equal to x.

Consider the following example, when tossing two fair coins $S = \{TT, HH, TH, HT\}$.

The probabilitiy for each element in the sample space is given by:

Element of S	Probability	x
TT	0.25	2
HH	0.25	0
TH	0.25	1
HT	0.25	1

We define X as a random variable that counts the number of tails.

x	P(X = x)
2	0.25
0	0.25
1	0.50

4.5 Probability distributions

Probability mass function (PMF) If X is a discrete random variable, the PMF gives for each x within the range of X the probability, i.e. $f(x) = P(X = x)$.

Any function can be used as a PMF as long as it satisfies:

1. $f(x) \geq 0$ for each x within the its domain;
2. $\sum_x f(x) = 1$ for all x's in the domain of X.

[3] More formal, a random variable is a **real-valued** function that assigns each element of the sample space S to a number.

4.5.1 Example

Let's repeat a coin tossing experiment, but this time a coin is tossed four times and we count the number of heads. The sample space S consists of 16 elements (2^4).

```
o1 <- c("H", "T")
S <- expand.grid(o1, o1, o1, o1)

# or more generically
S <- expand.grid(replicate(4, o1, simplify = FALSE))
```

```
head(S)
```

```
  Var1 Var2 Var3 Var4
1    H    H    H    H
2    T    H    H    H
3    H    T    H    H
4    T    T    H    H
5    H    H    T    H
6    T    H    T    H
```

```
x <- rowSums(S == "T")
table(x) / 16
```

```
x
     0      1      2      3      4
0.0625 0.2500 0.3750 0.2500 0.0625
```

```
# Or better
table(x) / nrow(S)
```

```
x
     0      1      2      3      4
0.0625 0.2500 0.3750 0.2500 0.0625
```

Instead of "manually" calculating these probabilities, we can think of a function (=PMF) that gives us $P(X = x)$.

For the previous example we could also use:

$$f_1(x) = \frac{\binom{4}{x}}{16}$$

for $x = 0, 1, 2, 3, 4$.

Binomial coefficient and factorial

The binomial coefficient $\binom{n}{k}$ is an abbreviated way of writing

$$\frac{n!}{k!(n-k)!}$$

for $n \geq k \geq 0$ where $n!$ is the factorial of n. $\binom{n}{k}$ is read as "n choose k" and gives the number of ways k elements can be chosen from a set of n elements.
In R there is the function `choose()` to calculate the binomial coefficient.
The factorial of a natural number n is defined as

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$$

By definition $0! = 1$.
Example: the factorial is for $n = 4$ is defined as $n! = \prod_{i=1}^{n} i = 4 \cdot 3 \cdot 2 \cdot 1 = 24$.
In R there is a function `factorial()` to do this.

```
factorial(4)
```

```
[1] 24
```

Let's verify that this is a true PMF

```
f1 <- function(x) {
  choose(4, x) / 16
}
```

```
(res <- sapply(0:10, f1))
```

```
 [1] 0.0625 0.2500 0.3750 0.2500 0.0625 0.0000 0.0000 0.0000 0.0000 0.0000
[11] 0.0000
```

1. $f(x) \geq 0$ for each x within the its domain;

```
all(res >= 0)
```

```
[1] TRUE
```

2. $\sum_x f(x) = 1$ for all x's in the domain of X.

```
sum(res)
```

```
[1] 1
```

Exercise 2: PMF

Use R to verify that $f(x) = \frac{2x}{k(k+1)}$ can serve as a PMF for a random variable X, with $x = 1, 2, 3, \dots, k$.

4.6 Cumulative Distribution Function (CDF) of RVs

While the PMF gives $P(X = x)$, the Cumulative Distribution Function (CDF; F) gives $P(X \leq x)$.

For discrete RV this is given by

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$$

What is the difference between PMF and CDFs? With the PMF you can calculate the probability of $X = x$. For example, what is the probability of obtaining a 4 from a dice roll. If you are interested in the probability of rolling a number that is smaller or equal to 4 you will need the CDF. Note, that for discrete RV the CDF is the sum of all PMFs of $X \leq x$ (see the formula above).

4.7 Continuous RV

We only worked with discrete RV up to here (i.e., S was countable). Continuous RV are very similar, with two distinct differences:

1. The PMF becomes the Probability Density Function (PDF) and is now defined for a an interval. Instead of $P(X = x) = f(x)$ we have for any PDF $P(a \leq x \leq b) = \int_a^b f(x)dx$.

2. In the distribution function becomes $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$

4.8 Statistical distributions

There are many statistical distributions for discrete and continuous RV that we can use and usually we do not come up with our own distributions.

- PMFs for discrete random variables: Bernoulli, Binomial, Poisson or negative Binomial distribution.
- PDFs for continuous random variables. E.g., exponential distribution, normal, t, gamma distribution.
- Many distributions are related to each other.
- Distributions usually have parameters to make them more flexible (as we have seen before)

4.8.1 Example: Coin toss

We can generalize our coin toss experiment even more by using a binomial distribution. The binomial distribution models a series of Bernoulli trials (each Bernoulli trial can have exactly two outcomes with probability p. p is the probability of success).

4.8.2 Example Poisson distribution

- The PMF of a Poisson distribution is given as

  $$f(k;\lambda) = \frac{\lambda^k \exp(-\lambda)}{k!}$$

  with rate $\lambda \in (0, \infty)$ and $k \in \mathbb{N}_0$.
- Using this function, we can calculate the probability to observe any k given λ.
- In R, we can use the function `dpois()` to do this (instead of plugin into the formula from above).

```
lambda <- 2
dpois(4, lambda)
```

```
[1] 0.09022352
```

```
(lambda^4 * exp(-lambda)) / factorial(4)
```

```
[1] 0.09022352
```

4.8.3 Normal distribution

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

and the standard normal distribution

$$Z = \frac{X - \mu}{\sigma}$$

Z is distributed as standard normal distribution with $\mu = 0$ and $\sigma = 1$, this is often written as $Z \sim N(0, 1)$.

Exercise 3: Distributions

Many bird species migrate between winter and summer ranges. Assume we have a sample of 20 independent birds and we know that the probability of arriving at the winter range is 0.86.

- What is the probability that at *exactly* 10 birds arrive at their winter range?
- What is the probability that *at least* 10 birds arrive at their winter range?
- What is the probability that more than 5 and less than 16 birds arrive at their winter range?

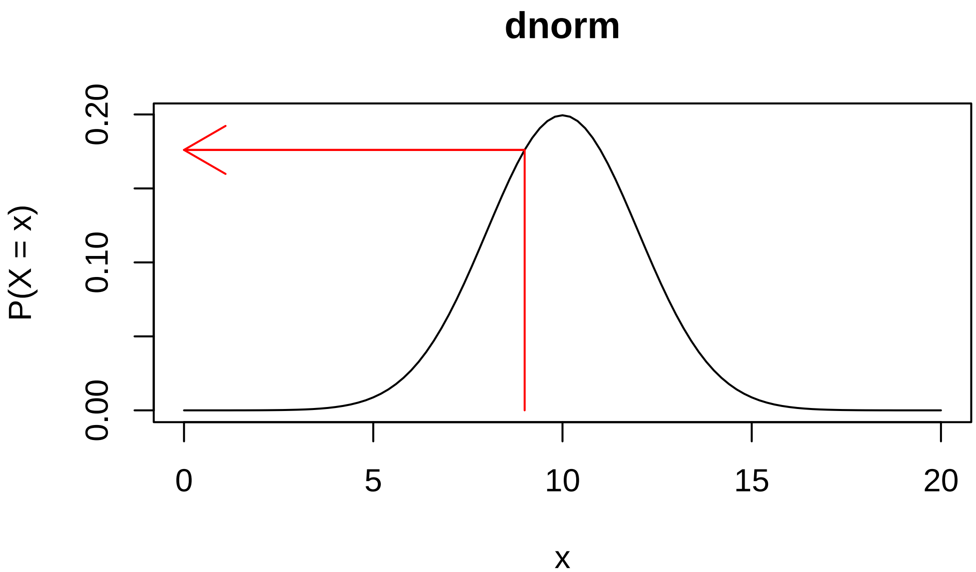
4.9 Working with distributions in R

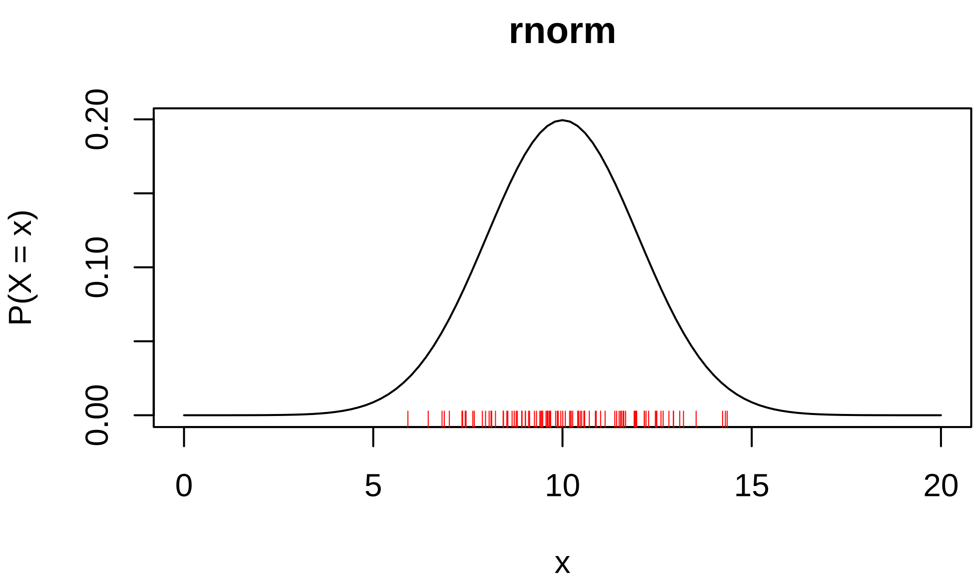
Most statistical distributions have four functions in R. For example for the normal distribution:

- `dnorm()`: Density function (the PMF or PDF)
- `pnorm()`: Cumulative distribution function (CDF)
- `qnorm()`: Quantiles of a distribution (the inverse of the CDF)
- `rnorm()`: Random numbers from a distribution

For a Poisson distribution the same would apply, but the functions are named (`dpois()`, `ppois()`, `qpois()`, `rpois()`).

Consider the following distribution: $X \sim N(10, 2)$ (i.e., a normal distribution with mean 10 and sd 2).


dnorm
P(X = x)
0.00
0.10
0.20
0
5
10
15
20
x


rnorm
P(X = x)
0.00
0.10
0.20
0
5
10
15
20
x