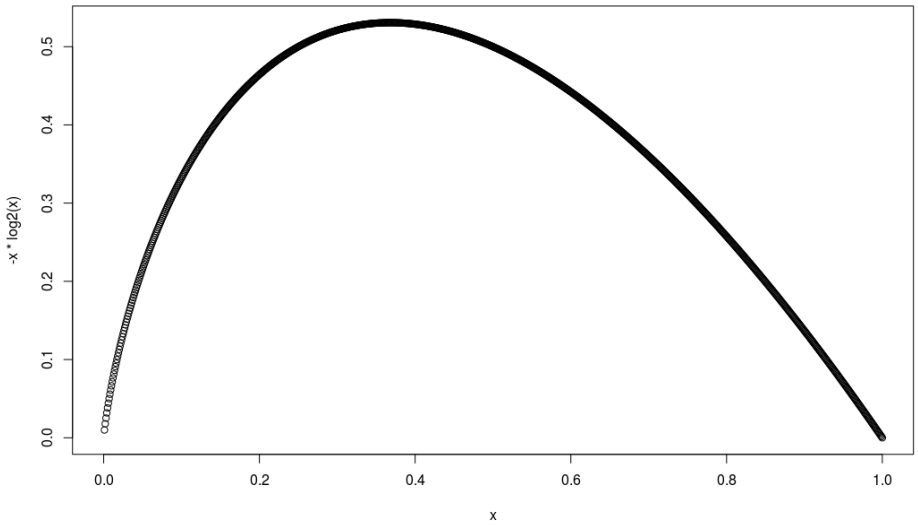
Exercises in R on probability distributions – a gentle tutorial

NPFL 054 lab session (Hladká and Holub, 2016)

Part I – ENTROPY

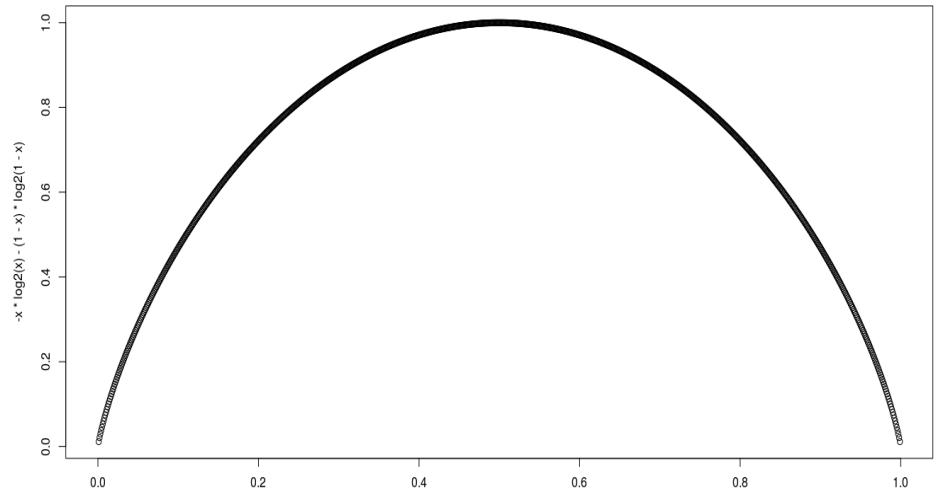
Example 1: Draw the graph of function $p \cdot \log_2(p)$ for $0 \leq p \leq 1$.

```
> p = (0:1000)/1000
> plot(p, -p*log2(p))
```



Example 2: Draw the graph of function H(p, 1-p) for $0 \leq p \leq 1$.

```
> p = (0:1000)/1000
> plot(p, - p*log2(p) - (1-p)*log2(1-p))
```



Part II – Working with probability distributions in R

Hint: Get familiar with R functions related to probability distributions, both discrete and continuous.

- `dbinom(), pbinom(), qbinom(), rbinom()`
- `dnorm(), pnorm(), qnorm(), rnorm()`

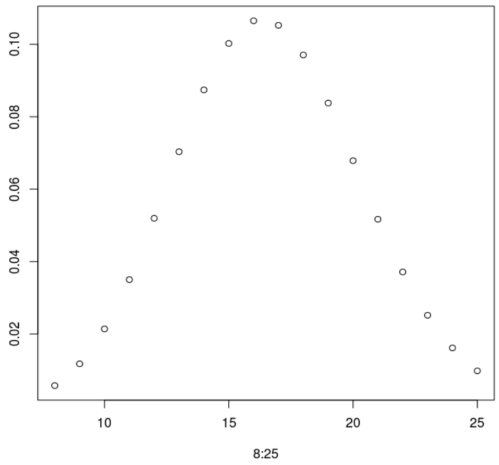
Exercise 1 (empirical, discrete distribution)
Read and analyze data file xy.100.csv

- Read file `xy.100.csv` – there are 100 observations of two random variables X and Y.
- Compute the estimate of marginal distributions p(x), p(y).
  – i.e. estimate the probabilities p(x), p(y) for all values of X and Y
- Draw histograms for both variables.
  – hints: `plot(<factor>), barplot(table(...)), hist(...)`
- Compute the estimate of joint and conditional distributions p(x,y), p(x | y), p(y | x).
  – i.e. estimate those probabilities for all pairs (x,y)
- Compute entropy and mutual information
  - H(X), H(Y)
  - H(X | Y), H(Y | X)
  - I(X;Y)
- Question: Are distributions X and Y statistically independent?

Exercise 2 (standard, discrete, binomial distribution)
Is the die true?

You roll a die 100 times and get just 10 sixes… – Is the die true?

- What is the probability of getting just 10 sixes?
- What is the probability of getting 10 or fewer sixes?
- Draw the probability distribution.
- Simulate the described experiment 1000 times and compute the empirical distribution. Compare it to the theoretical one. Then do the same with 1,000,000 simulations.



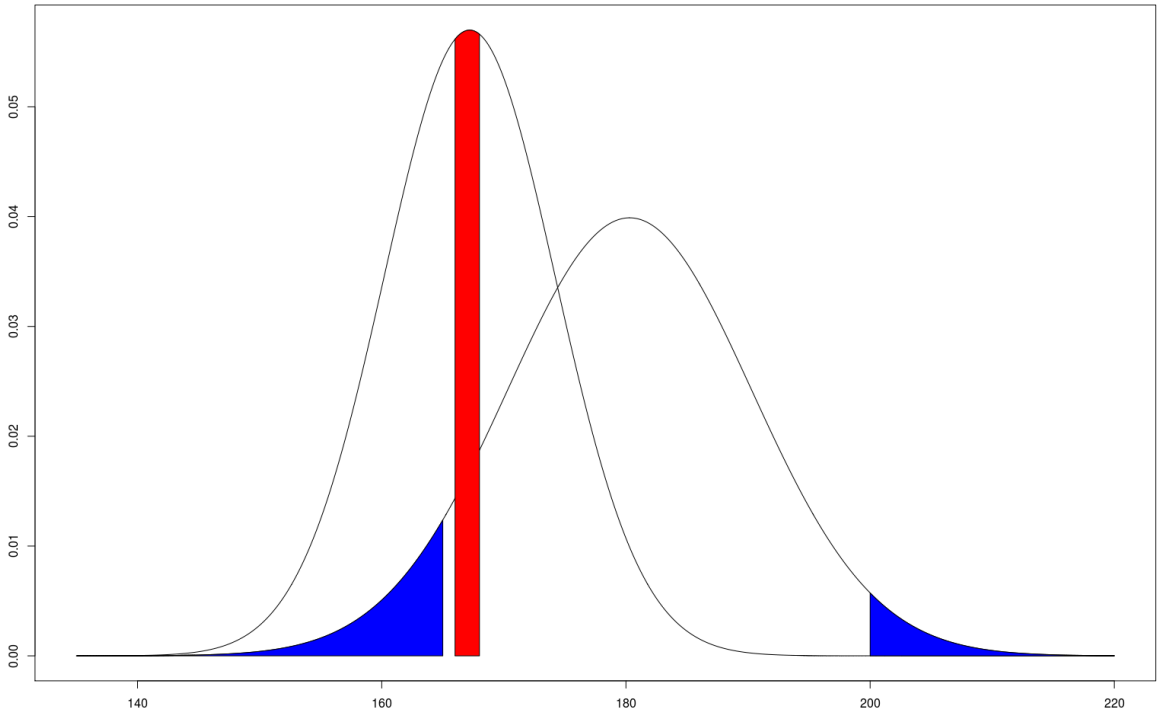
Exercise 3 (standard, continuous, normal distribution)
Which observation is more likely?

In 2001 it was found that the average male height in the Czech Republic is 180.3 cm, while the average female height is 167.2 cm. We assume that both distributions are normal with mean = 180.3 and variance = 100, and with mean = 167.2 and variance = 50, respectively.

Question: When you randomly meet a man, and then a woman – independently(!) –, what is more likely?

- that the man will be bigger than 200 or smaller than 165?, or
- that the woman will be between 166 and 168?

First, compute the probabilities, and then draw a picture with the density functions and display the probabilities as areas under the density curves. Hint: to fill the area under curve use polygon().



Exercise 4 (standard, continuous, normal distribution)
Who is more typical?

Assume the same height distributions as in the previous exercise. Then, who is more typical?

- a 178 cm man?, or
- a 169 cm woman?