Exercises 1

1. Let $y_1, \ldots y_n$ be a random sample from $N(\mu, \sigma^2)$ where mean μ and variance σ^2 are unknown. Show that $\hat{\mu} = \bar{y} = \sum_{i=1}^{n} y_i/n$ and $\hat{\sigma}^2 = \sum_{i=1}^{n}(\bar{y} - y_i)/n$ maximize the likelihood function

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2} \right) \right]$$

2. Suppose that X_1 and X_2 are independent unit exponential random variables. Show that the distribution of $Y = \log(X_1/X_2)$ is

$$f_y = \frac{\exp(y)}{(1 + \exp(y))^2}$$

for $-\infty < y < \infty$.

3. Show that the logistic density

$$f(x) = \frac{\exp(x)}{(1 + \exp(x))^2}$$

is symmetrical about zero. Find the cumulative distribution function and show that the $100p$ percentile occurs at

$$x_p = \log(p/(1 - p))$$

4. Generate a data matrix of 2010 rows where

   - covariate x obtains values $-1, -0.99, -0.98, \ldots, 0.98, 0.99, 1$ and each value of x is present 10 times,
   - response Y follows the model $\mathrm{E}(Y_i \mid x_i) = 0.2 + 3x_i$ and
   - the error term is normally distributed with mean 0 and variance 0.8.

   Fit a linear model to the data and compare the estimated model parameters to the true model parameters. (Useful R commands: `lm`, `rnorm`, `seq`, `rep`, `plot`, `summary`.)

5. Find a recent article where a GLM is used in **data analysis**. Identify the application area, the type of the GLM, the response variable, the explanatory variables and the sample size. Example:
   Article: J. Karvanen, K. Silander, F. Kee, L. Tiret, V. Salomaa, K. Kuulasmaa, P.-G. Wiklund, J. Virtamo, O. Saarela, C. Perret, M. Perola, L. Peltonen, F. Cambien, J. Erdmann, N. J. Samani, H. Schunkert, A. Evans, for the MORGAM Project, The impact of newly-identified loci on coronary heart disease, stroke and total mortality in the MORGAM prospective cohorts. Genetic Epidemiology, Volume 33, pages 237-246, 2009
   Application area: epidemiology
   Type of GLM: Logit model for binary response (one of the models used in the article)
   Response variable: disease history at baseline (=at the beginning of the study)
   Explanatory variables: single nucleotide polymorphism, age at baseline, sex and cohort
   Sample size: 5613 (case-cohort design)