Open Access Library Journal
2016, Volume 3, e3049
ISSN Online: 2333-9721
ISSN Print: 2333-9705

An Application of Bootstrapping in Logistic Regression Model

Isaac Akpor Adjei[1]*, Rezaul Karim[2]

[1]Department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[2]Department of Statistics, Jahangirnagar University, Savar, Bangladesh
Email: *isaac.adjei@gmail.com

How to cite this paper: Adjei, I.A. and Karim, R. (2016) An Application of Bootstrapping in Logistic Regression Model. *Open Access Library Journal*, **3**: e3049.
http://dx.doi.org/10.4236/oalib.1103049

Received: September 8, 2016
Accepted: September 25, 2016
Published: September 28, 2016

Abstract

Computer intensive methods have recently been intensively studied in the field of mathematics, statistics, physics, engineering, behavioral and life sciences. Bootstrap is a computer intensive method that can be used to estimate variability of estimators, estimate probabilities and quantile related to test statistics or to construct confidence intervals, explore the shape of distribution of estimators or test statistics and to construct predictive distributions to show their asymptotic behaviors. In this paper, we fitted the classical logistic regression model, and performed both parametric and non-parametric bootstrap for estimating confidence interval of parameters for logistic model and odds ratio. We also conducted test of hypothesis that the prevalence does not depend on age. Conclusions from both bootstrap methods were similar to those of classical logistic regression.

Subject Areas

Applied Statistical Mathematics, Statistics

Keywords

Nonparametric Bootstrap, Parametric Bootstrap, Logistic Regression, Confidence Interval, Test of Hypothesis

1. Introduction

Knowing the distribution of test statistic of random sample drawn from population of interest provides clues as to the methods to be employed in analyzing such data. Statisticians normally have to make decision on the nature of the distribution for the population of which the sample was obtained. A good guess of the nature of the population distribution leads to powerful test. However, high price is paid if the assumption of the distribution is wrong. Under normal circumstances it is not possible to validate the distribution of sample by re-sampling from given population due to high cost of implementation. It is very important to consider other methods of analyzing data, which are flexible with the choice of the distribution and based on this, bootstrap methods was introduced.

Bootstrap methods are computer-based methods for assessing measures of accuracy to statistical estimates like sample mean and standard errors (Efron and Tibshirani, 1994 [1]). The general idea is based on re-sampling from a given sample. There are three types of bootstrap: non-parametric bootstrap which does not assume any distribution of the population; semi-parametric bootstrap, which partly has an assumption on the distribution on parameter and whose residuals have no distributional assumption; and finally parametric bootstrap which assumes a particular distribution for the sample at hand. In this paper parametric and non-parametric bootstrap are considered for the given dataset.

Objectives

The aims of this paper are to formulate a logistic regression model and estimate the probability of infection as function of age using a Generalized Linear Model for binary data, construct 95% confidence intervals for the unknown parameters of the model and test the hypothesis that the prevalence does not depend on age using both classical and bootstrap (non-parametric and parametric) methods.

2. Methodology

2.1. Data

The dataset *Keil* (see Appendix), is a serological data of Hepatitis A from Bulgaria. It contains information about the age of the subject (in age group of one year), the number of seropositive (number of infected by hepatitis A), and sample size at each are group.

2.2. Logistic Regression Model

Bootstrapping is rapidly becoming a popular alternative tool to estimate parameters and standard errors for logistic regression model (Ariffin and Midi, 2012 [2]). Fitrianto and Cing (2014) [3] asserts that logistic regression is a popular and useful statistical method in modeling categorical dependent variable. Logistic regression is a statistical modeling approach used to investigate the relationship between the independent variable(s) and dichotomous dependent variable (Kleinbaum and Klein, 2010 [4]). In this section, the response variable of interest is the number of infected (Y_i) with Hepatitis A where $Y_i \sim Bin(n_i, \pi_i)$. $Age(x)$ is considered as covariate. The logistic regression model for binary response variable takes the form

$$\text{logit}[\pi_i] = \beta_0 + \beta_1 Age_i, \quad i = 1, 2, \cdots, n \tag{1}$$

where π_i is the proportion of Hepatitis A infection and β_1 is the effect of Age on the log odds of prevalence of infection. The β_0 and β_1 are unknown parameters to be estimated with 95% confidence intervals and a test of hypothesis $H_0:\beta_1=0$ by classical (Agresti and Kateri, 2011 [5]) and bootstrap methods are performed.

2.3. Parametric Bootstrap

In applications where the standard asymptotic theory does not hold, the null reference distribution can be obtained through parametric bootstrapping (Reynolds and Templin, 2004 [6]). Here a parametric model is fitted to the data, often by maximum likelihood, and random samples are drawn from this fitted model. Then the estimates of interest are computed from these data. This sampling process is repeated many times. The use of a parametric model at the sampling stage of the bootstrap methodology leads to procedures which are different from those obtained by applying basic statistical theory to inference for the same model.

Parametric bootstrap confidence interval

Using an algorithm by (Zoubir and Iskander, 2004 [7]; Carpenter and Bithell, 2000 [8]), a parametric bootstrap confidence interval is obtained as follows:

1) Estimate parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$) of logistic model (1) using the observed data and estimate π:

$$\hat{\pi}_i=\exp\left(\hat{\beta}_0+\hat{\beta}_1 Age_i\right)\Big/\left(1+\exp\left(\hat{\beta}_0+\hat{\beta}_1 Age_i\right)\right),\quad i=1,2,\cdots,n. \tag{2}$$

2) Draw bootstrap sample $\left(x,y^*\right)_b=\left(\left(x_1,y_1^*\right),\cdots,\left(x_n,y_n^*\right)\right)$ where $y_i^*\sim B\left(n_i,\hat{\pi}_i\right)$ for ($b=1,\cdots,B$).

3) For each $b=1,\cdots,B$ estimate the bootstrap sample statistics $\hat{\theta}_1^*,\cdots,\hat{\theta}_B^*$ where $\hat{\theta}_b^*=\left(\hat{\beta}_0^*,\hat{\beta}_1^*,\mathrm{e}^{\hat{\beta}_0^*},\mathrm{e}^{\hat{\beta}_1^*}\right)$ by refitting model (1).

4) Estimate the bootstrap mean and standard error of $\hat{\theta}=\left(\hat{\beta}_0,\hat{\beta}_1,\mathrm{e}^{\hat{\beta}_0},\mathrm{e}^{\hat{\beta}_1}\right)$.

$$\bar{\theta}^*=\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^* \quad\text{and}\quad \hat{s}e\left(\hat{\theta}\right)=\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^*-\bar{\theta}^*\right)} \tag{3}$$

5) Estimate $(1-\alpha)100\%$ bootstrap confidence interval by finding quantile of bootsrap replicates

$$\left(\hat{\theta}_L,\hat{\theta}_U\right)=\left(\hat{\theta}_{(b)}^{*\left(\frac{\alpha}{2}\right)},\hat{\theta}_{(b)}^{*\left(1-\frac{\alpha}{2}\right)}\right) \tag{4}$$

2.4. Non-Parametric Bootstrap

The non-parametric bootstrap belongs to the general sub-field non-parametric statistics that is defined by (Dudewicz, 1976 [9]) as the sub-field of statistics that provides statistical inference procedures, which rely on weaker assumptions (or no assumptions at all) about the underlying distribution of the population. Statistical practitioners should use non-parametric procedures only in so far as the assumptions about the underlying distribution are seriously doubtful in their validity. Efron (1979) [10] states that the bootstrap is a way to pull oneself up (from an unfavorable situation) by ones bootstrap, to provide trustworthy answers despite of unfavorable circumstances. However, when assumptions are not violated, non-parametric procedures will usually have greater variance (in point estimation), less power (in hypothesis testing), wider intervals (in confidence interval estimation), lower probability of correct selection (in ranking and selection) and higher risk (in decision theory) when compared to a corresponding parametric procedure (Efron and Tibshirani, 1994 [1]).

The idea called substitution principle or the plug-in rule gives explicit recognition of the fact that frequentist inference involves replacement of an unknown probability distribution by an estimate. In the simplest setting a random sample is available and the nonparametric estimate is the empirical distribution function, while a parametric model with a parameter of fixed dimension is replaced by its maximum likelihood estimate (Davison *et al.*, 2003 [11]).

Non-Parametric bootstrap confidence interval

Using a procedure proposed by (Zoubir and Iskander, 2004 [7]; Carpenter and Bithell, 2000 [8]), an algorithm for non-parametric confidence interval can be written as follows:

1) Make a new dataset for binary response with covariate(s) (x, y) from group data.

2) Draw bootstrap sample by sampling the pairs with replacements from new the dataset $(x, y)_b^* = \left((x_1, y_1)^*, \cdots, (x_n, y_n)^*\right)$ for ($b = 1, \cdots, B$).

3) For each $b = 1, 2, \cdots, B$ estimate the bootstrap sample statistics $\hat{\theta}_1^*, \cdots, \hat{\theta}_B^*$ where $\hat{\theta}_b^* = \left(\hat{\beta}_0^*, \hat{\beta}_1^*, e^{\hat{\beta}_0^*}, e^{\hat{\beta}_1^*}\right)$ by refitting model (1).

4) Estimate the bootstrap mean and standard error of $\hat{\theta} = \left(\hat{\beta}_0, \hat{\beta}_1, e^{\hat{\beta}_0}, e^{\hat{\beta}_1}\right)$.

$$\bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^* \quad \text{and} \quad \hat{se}\left(\hat{\theta}\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \bar{\theta}^*\right)} \tag{5}$$

5) Estimate $(1-\alpha)100\%$ bootstrap confidence interval by finding quantile of bootstrap replicates

$$\left(\hat{\theta}_L, \hat{\theta}_U\right) = \left(\hat{\theta}_{(b)}^{*\left(\frac{\alpha}{2}\right)}, \hat{\theta}_{(b)}^{*\left(1-\frac{\alpha}{2}\right)}\right) \tag{6}$$

2.5. Bootstrap Test

In many applications, significance testing can be used to assess the plausibility of certain hypothesis. The likelihood ratio test, the score test and the Wald test are three asymptotically equivalent test procedures. For regular cases, their null distribution is a χ^2 distribution with the appropriate degrees of freedom. This χ^2 distribution is an approximate null distribution. The true null distribution converges to the χ^2 distribution as the sample size tends to infinity. It is not always clear whether this approximation is accurate enough or even valid in all cases. The bootstrap can offer an alternative way to determine an approximate null distribution. The bootstrap based null distribution also improves as the sample size increases, but there are theoretical and simulation results showing that it is often at least as accurate as its χ^2 counterpart (Davison and Hinkley, 1997 [12]). The bootstrap mechanism should reflect the original data generation mechanism and the bootstrap simulation should satisfy the null hypothesis. The idea is to generate (repeatedly) new bootstrap data, reflecting the null hypothesis, recalculate the test statistic and in this way to simulate the null distribution of the test statistic. These bootstrap test values can then be used to compute a p-value.

2.5.1. Parametric Bootstrap for Test of Hypothesis

The algorithm for parametric test of hypothesis given by (Fox, 2015 [13]) is as follows:

1) Estimate parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$) of logistic model (1) using the observed data and calculate observed (obs) test statistic $t_{\beta_1}^{obs} = \hat{\beta}_1 / se(\hat{\beta}_1)$. Let $\hat{\theta} = (\hat{\beta}_1, t_{\beta_1}^{obs})$.

2) Estimate π under the $H_0 : \beta_1 = 0$

$$\hat{\pi}_i = \exp(\hat{\beta}_0) / (1 + \exp(\hat{\beta}_0)), \quad i = 1, 2, \cdots, n \quad \text{under } H_0 : \beta_1 = 0 \tag{7}$$

3) Draw bootstrap sample $(x, y^*)_b = ((x_1, y_1^*), \cdots, (x_n, y_n^*))$ where $y_i^* \sim B(n_i, \hat{\pi}_i \mid H_0)$ for ($b = 1, \cdots, B$).

4) For each $b = 1, \cdots, B$ estimate the bootstrap sample statistics $\hat{\theta}_1^*, \cdots, \hat{\theta}_B^*$ where $\hat{\theta}_b^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, t_{\beta_1}^*)$ by refitting model (1).

5) Calculate bootstrap P-value by

$$P\text{-value} = \frac{1 + \#(|\hat{\theta}_b^*| > |\hat{\theta}|)}{B + 1} \quad \text{for } \hat{\theta}_b^* = (\hat{\beta}_1^*, t_{\beta_1}^*) \tag{8}$$

where # represent the number of times.

2.5.2. Non-Parametric Bootstrap for Test of Hypothesis

An algorithm for non-parametric test of hypothesis given by (Fox, 2015 [13]) is as follows:

1) Make a new dataset for binary response with covariate(s) (x, y) from group data.

2) Estimate parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$) of logistic model (1) using the observed data and calculate observed (obs) test statistic $t_{\beta_1}^{obs} = \hat{\beta}_1 / se(\hat{\beta}_1)$. Let $\hat{\theta} = (\hat{\beta}_1, t_{\beta_1}^{obs})$.

3) By fixing x, draw bootstrap sample by sampling from only y with replacements form new dataset $(x, y^*)_b = ((x_1, y_1^*), \cdots, (x_n, y_n^*))$ for ($b = 1, \cdots, B$). This breaks the correlation between x and y.

4) For each $b = 1, \cdots, B$ estimate the bootstrap sample statistics $\hat{\theta}_1^*, \cdots, \hat{\theta}_B^*$ where $\hat{\theta}_b^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, t_{\beta_1}^*)$ by refitting model (1).

5) Calculate bootstrap P-value by

$$P\text{-value} = \frac{1 + \#(|\hat{\theta}_b^*| > |\hat{\theta}|)}{B + 1} \quad \text{for } \hat{\theta}_b^* = (\hat{\beta}_1^*, t_{\beta_1}^*) \tag{9}$$

where # represent the number of times.

3. Results

3.1. Logistic Regression Model

The parameter estimates together with standard errors (s.e) and confidence intervals (C.I) by logistic model (1) using classical approach are shown in **Table 1**. It is observed that per year increase in age increases the log odds of infection by 0.0838. This result is statistically highly significant with $p\text{-value} < 0.0001$ at 5% level of significance. The 95% confidence interval for the odds ratio of the effect of age is (1.0735, 1.1015). This means the odds of infection increased about 7% to 10%.

3.2. Parametric Bootstrap

Results obtained from the logistic model (1) by parametric bootstrap are shown in Table 2. This results lead to similar conclusion from classical method. The estimated odds ratio obtained is 1.088. The confidence interval depicts the odds of infection increases about 7% to 10%.

3.3. Non-Parametric Bootstrap

The parameter estimates together with standard errors (s.e) and confidence intervals (C.I) of the logistic model (1) by using non-parametric bootstrap approach are presented in Table 3. These results lead to similar conclusion as in the case of classical and parametric bootstrap methods. The estimated odds ratio obtained by this approach is also is 1.088. This also means with regards to the confidence interval, the odds of infection increased about 7% to 10%.

3.4. Comparisons: Classical and Bootstrap Parameter Estimates

From Tables 1-3, it can be observed that the parameter estimates are very close. The standard errors of estimates for parametric bootstrap were slightly smaller compared to that of non-parametric bootstrap but very close to that obtained from classical approach. This is due to the fact that in both Classical and Parametric bootstrap methods,

Table 1. Parameter estimates of logistic model using GLM.

Parameter	estimate	s.e	95% C.I	z value	Pr(>|z|)
Intercept (β_0)	−1.4301	0.1736	(−1.7704, −1.0899)	−8.24	<0.0001
AGE (β_1)	0.0838	0.0066	(0.0709, 0.0967)	12.72	<0.0001
$\exp(\beta_1)$	1.0874	0.0072	(1.0735, 1.1015)		

Table 2. Confidence intervals and p-value by parametric bootstrap.

Parameter	estimate	s.e	95% C.I	P-value
β_0	−1.4412	0.1756	(−1.7988, −1.1119)	
β_1	0.0844	0.0066	(0.0722, 0.0981)	<0.0001
$\exp(\beta_1)$	1.0880	0.0072	(1.0748, 1.1031)	

Table 3. Confidence intervals and p-value by non-parametric bootstrap.

Parameter	estimate	s.e	95% C.I	P-value
β_0	−1.4400	0.1813	(−1.8107, −1.1007)	
β_1	0.0843	0.0070	(0.0716, 0.9878)	<0.0001
$\exp(\beta_1)$	1.0880	0.0076	(1.0743, 1.1038)	

the design matrix for the covariate Age was fixed.

The *P*-values obtained for testing hypothesis $H_0 : \beta_1 = 0$ by both Non-Parametric and Parametric methods are shown in Table 2 and Table 3 respectively. It is observed that in both situations, the effect of age is highly significant.

Comparing the length of confidence intervals for the three methods, it was observed that the interval length for non-parametric method is wider compared to that of classical and parametric methods. The classical and that of parametric methods have similar interval length.

The 95% confidence intervals for predicted prevalence by using both parametric and non-parametric methods are presented in Figure 1 and Figure 2 respectively. It can generally be concluded that the probability for infection increases with age. The length of the interval reduces with increase in Age. This means prediction for higher age is more precise.

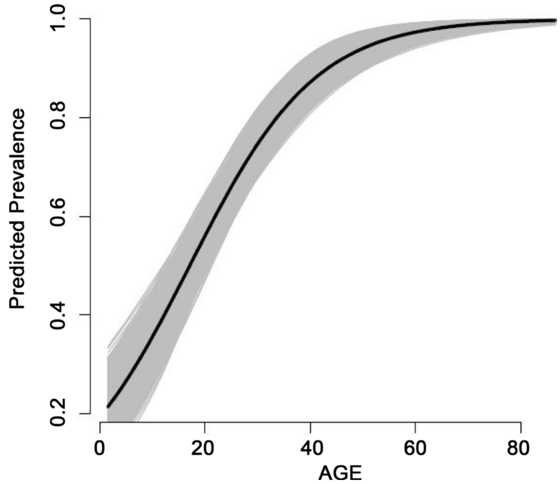

Figure 1. The 95% point wise confidence intervals for the predicted values of the prevalence by using parametric bootstrap.

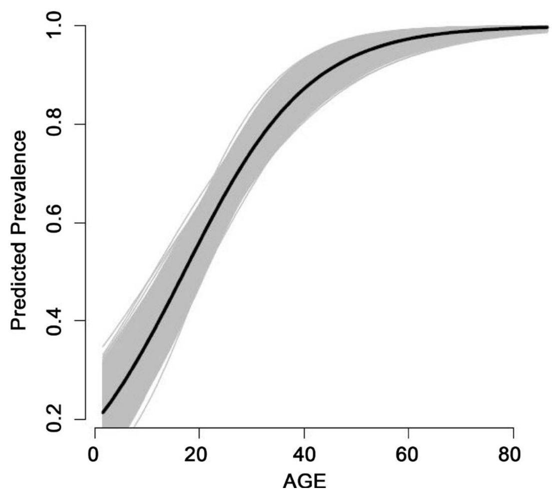

Figure 2. The 95% point wise confidence intervals for the predicted values of the prevalence by using non-parametric bootstrap.

4. Conclusions

The bootstrap technique used for estimation and testing produced flexible results. Most of the results were similar to the classical results established under probability theory.

From the classical logistic regression model estimates, it was observed that the prevalence of Hepathitis A infection increased with age. Parametric and non-parametric methods used to investigate the effect of age gave similar results.

We conclude that this computer intensive method gives us an idea about the asymptotic behavior of estimators and also it is easy in implementation based on simulations.

References

[1] Efron, B. and Tibshirani, R.J. (1994) An Introduction to the Bootstrap. Chapman and Hall/CRC, UK.

[2] Ariffin, S.B. and Midi, H. (2012) Robust Bootstrap Methods in Logistic Regression Model. 2012 *International Conference on Statistics in Science*, *Business*, *and Engineering* (*ICSSBE*), Langkawi, 10-12 September 2012, 1-6. http://dx.doi.org/10.1109/ICSSBE.2012.6396613

[3] Fitrianto, A. and Cing, N.M. (2014) Empirical Distributions of Parameter Estimates in Binary Logistic Regression Using Bootstrap. *International Journal of Mathematical Analysis*, **8**, 721-726. http://dx.doi.org/10.12988/ijma.2014.4394

[4] Kleinbaum, D.G. and Klein, M. (2010) Modeling Strategy Guidelines. In: *Logistic Regression*, *Part of the Series Statistics for Biology and Health*, Springer, Berlin, 165-202. http://dx.doi.org/10.1007/978-1-4419-1742-3_6

[5] Agresti, A. and Kateri, M. (2011) Categorical Data Analysis. Springer, Berlin. http://dx.doi.org/10.1007/978-3-642-04898-2_161

[6] Reynolds, J.H. and Templin, W.D. (2004) Comparing Mixture Estimates by Parametric Bootstrapping Likelihood Ratios. *Journal of Agricultural, Biological, and Environmental Statistics*, **9**, 57-74. http://dx.doi.org/10.1198/1085711043145

[7] Zoubir, A.M. and Iskander, D.R. (2004) Bootstrap Techniques for Signal Processing. Cambridge University Press, Cambridge. http://dx.doi.org/10.1017/CBO9780511536717

[8] Carpenter, J. and Bithell, J. (2000) Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians. *Statistics in Medicine*, **19**, 1141-1164. http://dx.doi.org/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F

[9] Dudewicz, E.J. (1976) Introduction to Statistics and Probability. Holt, Rinehart and Winston, New York.

[10] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7**, 1-26. http://dx.doi.org/10.1214/aos/1176344552

[11] Davison, A.C., Hinkley, D.V. and Young, G.A. (2003) Recent Developments in Bootstrap Methodology. *Statistical Science*, **18**, 141-157. http://dx.doi.org/10.1214/ss/1063994969

[12] Davison, A.C. and Hinkley, D.V. (1997) Bootstrap Methods and Their Application, Volume 1. Cambridge University Press, Cambridge. http://dx.doi.org/10.1017/CBO9780511802843

[13] Fox, J. (2015) Applied Regression Analysis and Generalized Linear Models. Sage Publications, Thousand Oaks, California.

Appendix

Dataset

```
kei1<-list(
AGE=c(
   1, 2, 3, 4, 5, 6, 7, 8,  9, 10, 11, 12, 13, 14, 15, 16, 17, 18,
   19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34,
   35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50,
   51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66,
   67, 68, 69, 70, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83,
   86)+0.5,
POSITIVE=c(
   3, 3, 3, 4, 7, 4, 3, 4, 7, 8, 2, 3, 2, 0, 5, 13, 1, 3, 15, 22,
   15, 7,  8, 7, 12, 5, 10, 15, 9, 9, 9, 8, 9, 8, 9, 13, 6, 15, 11,
   6, 8, 13, 7, 5, 7, 9, 9, 22, 6, 10, 6, 13, 8, 7, 13, 11, 8, 8,
   9, 13, 5, 5, 5, 5, 10, 8, 4, 5, 4, 8, 9, 1, 4, 7, 6, 2, 3, 2, 4,
   1, 1, 2, 1),
NEGATIVE=c(
   16, 15, 16, 13, 12, 15, 12, 11, 10, 15, 7, 7, 11, 1, 16, 41, 2,
   6, 32, 37, 24, 10, 10, 11, 15, 10, 13, 19, 12, 9, 14, 10, 11, 9,
   14, 14, 7, 16, 13, 8, 8, 14, 10, 5, 7, 9, 9, 22, 7, 10, 6, 14, 8,
   7, 13, 11, 8, 8, 10, 16, 5, 6, 5, 5, 10, 8, 4, 5, 5, 8, 9, 1, 4,
   7, 6, 2, 3, 2, 4, 1, 1, 2, 1))
```