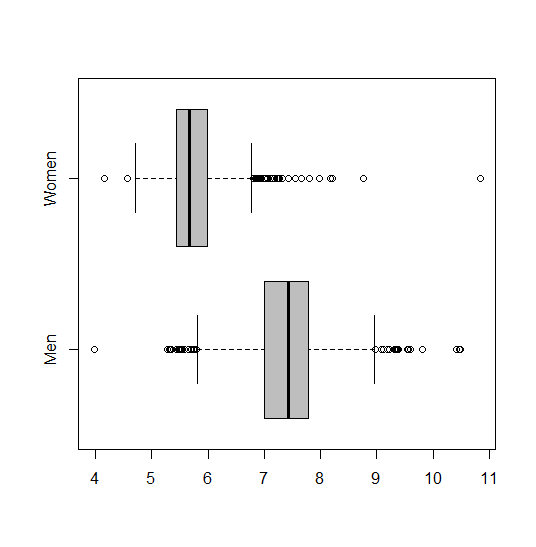
**Statistical Report Writing Sample No.3.**

**Introduction.** The food intake by Danish people was studied in a comprehensive survey by Haraldsdottir et al. (1985), including data from 2224 persons. Many variables were registered in this study. The summary statistics of a variable called BMR, related to the basal metabolic rate, are shown below. Note that the mean BMR are different, around 7.4 and 5.7, respectively.

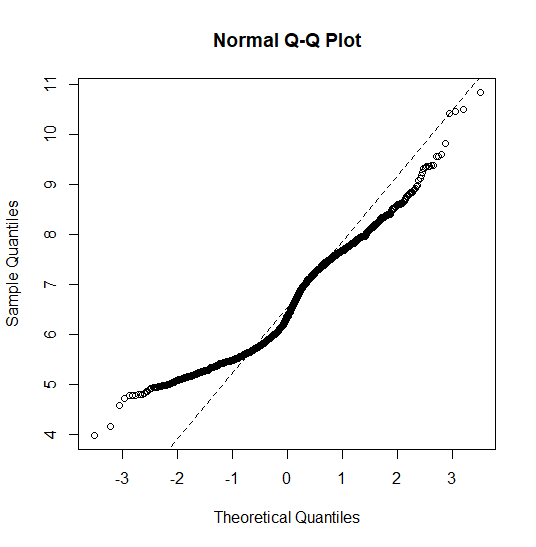
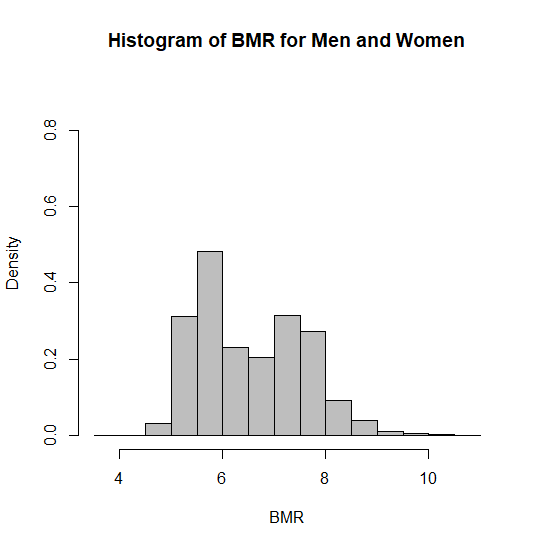
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Size | Mean | SD | 1st Quartile | Median | 3rd Quartile |
| Men | 1079 | 7.386 | 0.723 | 6.997 | 7.424 | 7.786 |
| Women | 1145 | 5.747 | 0.498 | 5.444 | 5.677 | 5.991 |

Parallel boxplots for men and women are shown below. The summary statistics and the boxplots tell the same story: BMR values are generally higher for men compared to women. The result might be different, though, if we repeat the experiment and use new sample. The purpose of the statistical analysis is to clarify whether the observed difference between men and women is caused by an actual

difference.



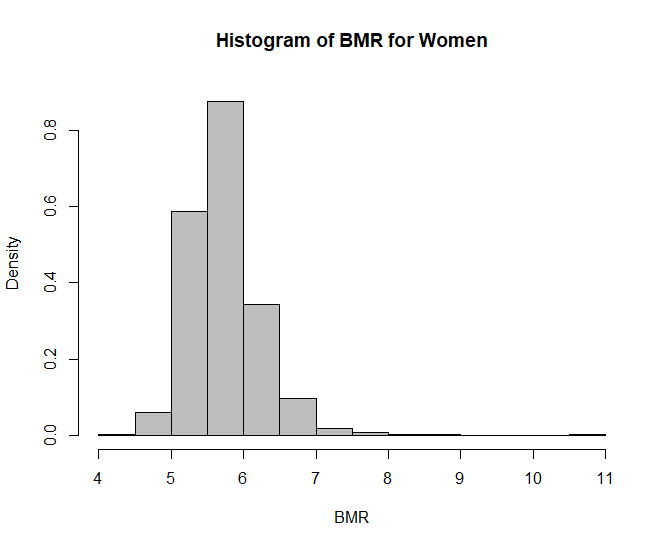
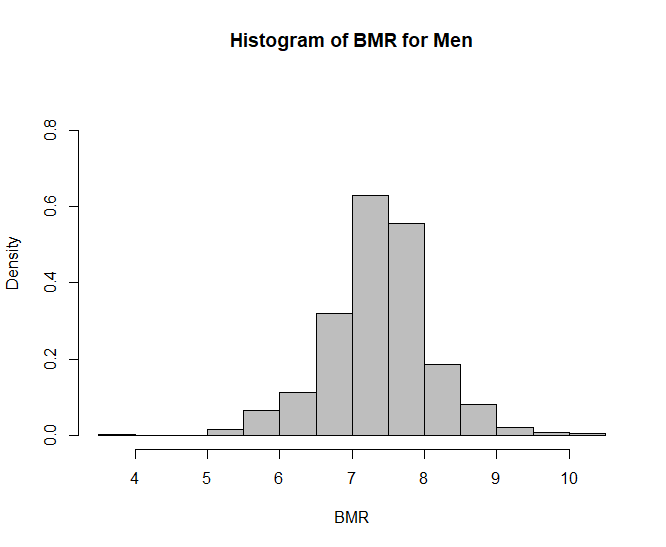
The histogram (below left) includes all data, and a QQ-plot is shown in the lower right figure. The distribution is clearly bimodal and the normal approximation is thus not appropriate.



**Statistical model and data analysis.**  The study is the comparison of two groups, men and women, in which we want to identify the parameters αMen and αWomen of mean BMR for men and women. The model is given by

*y*i = αg(i) + *e*i

where *y*i ’s are the observed BMR, *e*i‘s are the reminder terms, and *g(i)* is either “Men” or “Women.” Note that the reliability of statistical inference is dependent on the ability to assume that the reminder terms *e*i‘s come from a normal distribution. Thus, we need to assess the normality for the variable BMR since it is used to construct a confidence interval.



The two separate histograms (left and right above) for men and women, respectively show reasonably bell-shaped (although not perfect) corresponding a normal density curve, and therefore, the normal distribution was adequate to describe the data for each of the samples. The distribution for the sample of men seems to be slightly wider than for the sample of women. Thus, we may assume that their variances (or equivalently, the standard deviations) are different. The means of the BMR variable are 7.386 for men and 5.747 for women, respectively, and the difference (αMen − αWomen) of the expected BMR is estimated by

7.386 − 5.747 = 1.639

The general procedure yields the standard error SE(αMen − αWomen) of 0.0265. The degrees of freedom is obtained as 1899.6, and the 95% confidence interval for the difference becomes (1.588, 1.692). The alternative hypothesis is HA: αMen ≠ αWomen, and test statistic is 61.941, which is highly unlikely. We firmly reject the null hypothesis.

We instead may assume that the variances are the same for men and women. Then the corresponding standard error SE(αMen − αWomen) is 0.0262. The degrees of freedom is obtained as 2222, and the 95% confidence interval becomes (1.589, 1.691). Due to the additional assumption, the expected difference is estimated with slightly higher precision.

**Conclusion.** Regardless of the additional assumption the p-value ( < 2.2×10-16 ) is extremely small, and therefore, the result is highly significant.If we do not assume that the variances are equal for two groups, the corresponding method (general procedure) produces a slightly wider confidence interval. In general, the least accurate result is viewed as a conservative choice. Thus, the confidence interval (1.588, 1.692) is the conservative result. In either case, the confidence interval indicates that there is significant difference of mean BMR values between men and women. In particular, we see that the confidence interval lies in positive values. In other words, men have a higher BMR than women.