**An Introduction to Using Routinely Collected Data With R**

**Introduction to Statistics**

**Solutions to Practical Sessions**

**Copyright Notice**

Copyright © 2016 by Robertson Centre for Biostatistics, University of Glasgow, Glasgow G12 8QQ. Telephone 0141 330 4744. This publication or any part thereof may not be reproduced or transmitted in any form or by any means electronic or mechanical, including photocopying, recording, storage in an information retrieval system or otherwise, without the prior written permission of the author.

**Disclaimer**

While the author takes great care to ensure the accuracy and quality of these materials, all material is provided without any warranty whatsoever, included but not limited to, the implied warranties of merchantability or fitness for a particular purpose. Further, the author reserves the right to revise this publication and make changes to its content at any time, without obligation to notify any person or entity of such revisions or changes.

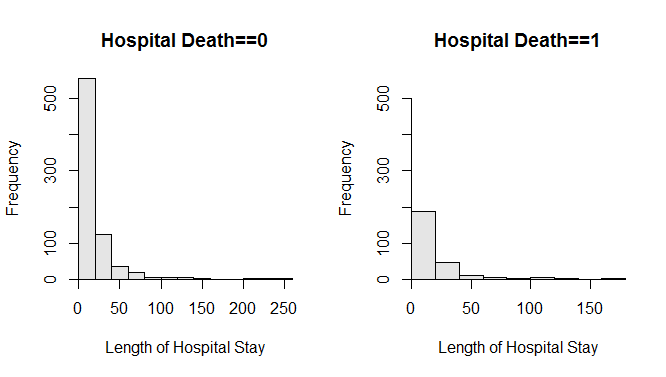
1. Practical Session 1: Basic Data Description

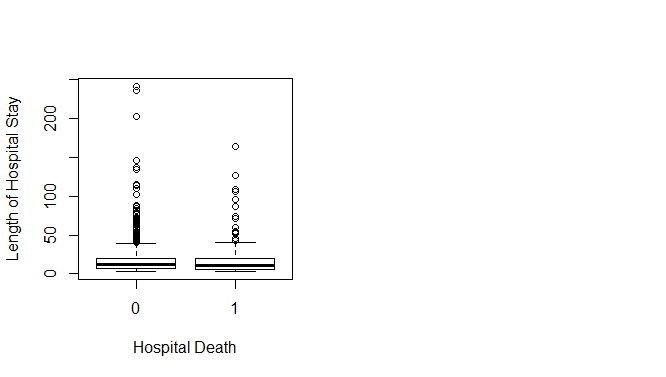
1.1

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistics** | **Overall** | **Males** | **Females** |
| N | 1000 | 562 | 438 |
| Mean | 62.47 | 61.94 | 63.16 |
| Standard deviation | 16.11 | 15.47 | 16.90 |
| Median | 64.90 | 64.90 | 64.87 |
| 95th percentile | 86.00 | 83.90 | 87.97 |
| Maximum | 101.85 | 96.71 | 101.85 |

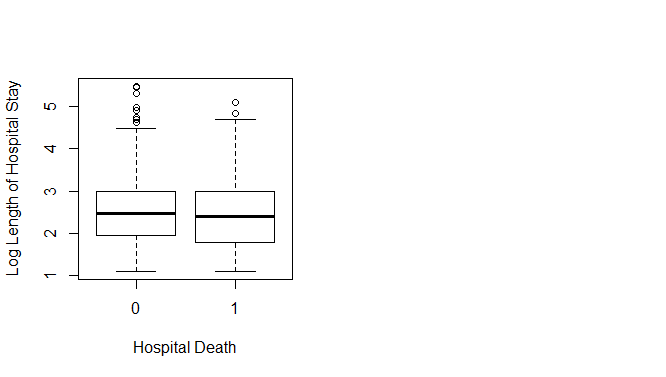
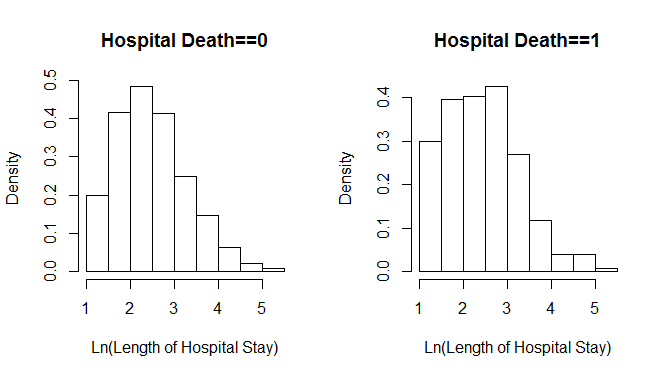
1.2

1. The length of hospital stay is not normally distributed





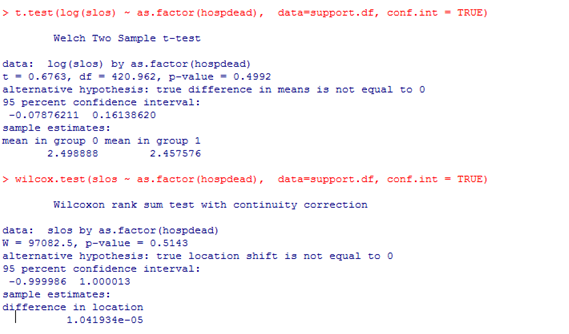
1. Log length of hospital stay looks more Normal than the raw length,.



1. There is very little difference, the difference in the medians is around 0.087, and the difference in mean is around 0.041.
2. Practical Session 2: Hypothesis Testing
   1. The sample mean and median between the groups are fairly similar, and the plots of the transformed variables are approximately normal, therefore a 2-sample t-test is sufficient. The null hypothesis for the test is that the population mean difference of log(slos) between the two hospital death groups is zero (equivalently, the population group means are equal).

The confidence interval shows that the population differences between log length of hospital stay in those that died in hospital and those that didn’t die could be 0.07 in one direction or 0.16 in the other direction. Since this interval contains 0, and the p-value is greater than 0.05, we can conclude that there is no evidence of a difference between log lengths of hospital stay (and therefore, length of hospital stay itself) for the two hospital death groups.

The Mann-Whitney test gives very similar results to the t-test.



* 1. Yes, there is a highly significant difference (p<0.001 for both tests). Eastern Europe and USSR and Africa are more likely to have death rates above 10 per 1000 population, while the other four country groups are more likely to have low death rates.

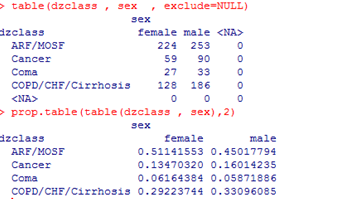


Generally one should use Fisher’s exact test because some of the expected counts are less than 5. This can be tested using ExpFreq() from the package DescTools (see lecture for example). It’s probably safer to use this test; however, on this occasion it makes no difference to the results.

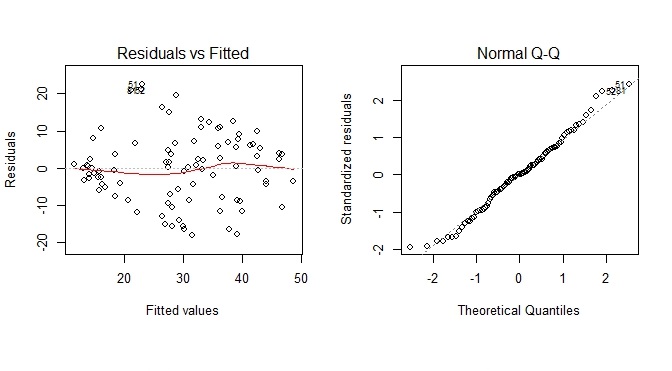
* 1. The null hypothesis: In the population, admission type is independent of gender. Equivalently, in the population, the proportion of patients with each admission type is equal for each gender.

The proportion for cancer and coma are reasonably similar but females tend to be slightly more likely to be admitted with ARF/MOSF and males with COPD/CHF/Cirrhosis.

However, the difference in the distributions of patients across the admission types between gender is not significant (p=0.2392).



1. Practical Session 3: Correlation and Regression
   1. Pearson’s correlation co-efficient is most appropriate here. The sample correlation coefficient is -0.89 and there is strong evidence to reject the null hypothesis that the population correlation coefficient is zero (p<0.05). The confidence interval is -0.93 to -0.85, suggesting that in the wider population, the correlation is likely to lie between -0.85 and -0.93. There is evidence of a strong negative association between female life expectancy and birth rate.
2. There is a moderately strong negative association between the birth rates and log10GNP (ie. as wealth increases, birth rate decreases). The relationship appears reasonably linear, though there is a reasonable amount of variability.



The normal probability plot indicates that the residuals are normally distributed.

There may be slightly less variability at the extremes of the residuals plot than in the middle, suggesting that the assumption of constant variability may be questionable. However, this phenomenon is not particularly unusual, nor is it particularly extreme in this case, and so we will accept this assumption.

There is no curvature in the residuals so the assumption of linearity holds.

1. We could transform the birth rate variable, eg. take a log transformation.
2. H0: There is no relationship between birth rate and log10GNP in the population.
3. The intercept is 75.5, so this is the average birth rate for a log10GNP of 0, or a corresponding GNP of 1.

The slope is -14.1, with a conﬁdence interval of -16.8 to -11.4. This is signiﬁcantly diﬀerent from zero, so there is evidence of a relationship between birth rate and GNP. In the population, we would expect the birth rate to decrease by between 11.4 and 16.8 per 1000 population for each 1 unit increase in log10GNP (this equates to multiplying GNP by 10).

Adjusted R squared is 53.7%, so this is a moderately well-ﬁtting model.

A standardized residual outside the interval (-2,2) is considered an outlier, and there are four countries that have residuals of greater than 2.

* 1. The average birth rate in the population for countries with GNP of 10000 is likely to be between 16.2 and 21.8 per 1000 population, while an individual country with GNP of 10000 is likely to have a birth rate of between 0.3 and 37.8 per 1000 population.