

Introduction

Factor VIII is necessary to make blood clot; haemophiliacs cannot produce their own factor VIII and thus need to be given factor VIII to prevent bleeding. However, high levels of factor VIII may increase the risk of having a heart attack. Factor VIII can be measured in two ways: as a protein concentration and as a clotting activity (*i.e.* by measuring the time taken for blood to clot). Both concentration and activity results are expressed as a percentage (%) of that of a reference standard. In this practical we look at some data on factor VIII in patients with angina (chest pain).

Data from 235 angina patients who were admitted to hospital for investigation of their chest pain are in the Stata file **factor8.dta**. This can be downloaded from the AT Moodle page. Open Stata and load the data. You may wish to open a dofile and save all the commands from this practical along with any comments.

There are four variables:

- id:** Patient number
- sex:** Gender (1=Male, 2=Female)
- con:** Factor VIII concentration (%)
- act:** Factor VIII activity (%).

4.1 Data exploration

1. Familiarise yourself with the data e.g. how many variables/observations are there, are there any missing values? How many men/women are there in this sample?
2. Produce histograms for factor VIII concentration and activity. Do these two variables appear to be normally distributed?
3. Log-transform each of the factor VIII variables. Do the transformed variables appear to be normally distributed?
4. Using the log-transformed values, plot factor VIII concentration against factor VIII activity. Would you accept an assumption of bivariate normality?

4.2 Association between factor VIII concentration and factor VIII activity

1. Use the command **corr** to obtain the Pearson correlation between factor VIII concentration and activity (use the log-transformed variables). Interpret your result.
2. Use the **summarize** command to calculate the standard deviation of the two factor VIII variables and hence calculate their covariance. You can check your answer by using the **corr** command with the option **cov**.
3. By hand, carry out a hypothesis test of whether the true correlation is zero. What assumptions are needed for this test to be valid?

4. By hand, calculate a 95% CI for the true correlation coefficient. What do you conclude? What assumptions are needed for this CI to be valid? [See last sheet of practical for how to download and use a user written command for calculating CIs for a correlation coefficient.]
5. Do you think the association between the two Factor VIII variables differs between men and women? How would you test whether there is a difference?

4.3 Association between gender and 'high' factor VIII concentration

In 4.2 we investigated the association between two continuous variables. Here we look at the association between two binary variables – gender and 'high' factor VIII concentration. We will define 'high' factor VIII concentrations to be those above 150%.

1. Generate a new binary variable that takes the value 1 if Factor VIII concentration is greater than 150% and 0 otherwise. What proportion of patients have a high concentration?
2. Using the **tabulate** command obtain a two-way frequency tabulation of high concentration and sex (add the options **col**, **chi2** and **exact**). What do you conclude from the table of results and the hypothesis tests reported here? Which of the two tests (chi square and Fisher's exact) is the most appropriate here?
3. Calculate, by hand, the odds ratio and its 95% CI relating high factor VIII concentration to gender. What do you conclude from this?
4. Use the **tabodds** command with the options **or** and **woolf** to check your results. Type **help tabodds** for more information.

Stata commands needed for practical

Summarising data:

```
describe
codebook
summarize          (for continuous variables)
tabulate           (for categorical variables)
```

Creating a histogram

```
histogram con          (add option normal for overlaid normal curve)
```

Creating a log-transformed variable

```
generate logcon=log(con)
```

Correlation between two variables

```
corr var1 var2
corr var1 var2 , cov    (asks for covariance to be displayed)
```

Functions for hyperbolic tangent (tanh) and inverse hyperbolic tangent (\tanh^{-1})

```
display tanh(0.6)      (tanh)
display atanh(0.7)     ( $\tanh^{-1}$ )
```

Confidence intervals for a correlation coefficient

```
net search corrci.ado
corrci log_con log_act
```

Chi-square and Fisher's exact test

```
tab var1 var2 , chi exact
```

Creating a binary (yes/no) variable from a continuous variable

```
generate highcon=(con>150) if con<.
```

Odds ratios for two-by-two table

```
tabodds highcon sex , or
```