4.1 Data Exploration

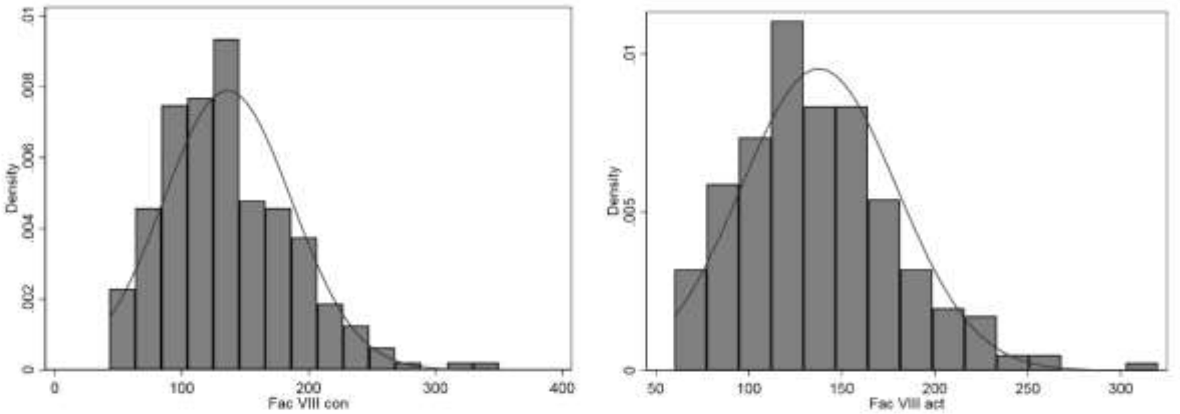
The commands `describe` and codebook give a helpful overview of the dataset in memory. There are no missing values.

```
. describe

Contains data from factor8.dta
 Observations:           235
    Variables:             4                  1 Oct 2021 14:14
--------------------------------------------------------------------------------Variable
Storage   Display    Value
    name         type    format    label      Variable label
--------------------------------------------------------------------------------
id              int     %10.0g                Patient No.
sex             byte    %8.0g      sexlab     Sex
con             float   %9.0g                 Factor 8 Concentration
act             float   %9.0g                 Factor 8 Activity
--------------------------------------------------------------------------------
```

```
. codebook

-------------------------------------------------------------------------------
id                                                                  Patient No.
-------------------------------------------------------------------------------

                  Type: Numeric (int)

                 Range: [1,258]                     Units: 1
         Unique values: 235                     Missing .: 0/235

                  Mean: 132.791
             Std. dev.: 73.9813

           Percentiles:     10%       25%       50%       75%       90%
                             29        69       135       197       234

-------------------------------------------------------------------------------
sex                                                                         Sex
-------------------------------------------------------------------------------

                  Type: Numeric (byte)
                 Label: sexlab

                 Range: [1,2]                       Units: 1
         Unique values: 2                       Missing .: 0/235

            Tabulation: Freq.   Numeric  Label
                          199         1  Male
                           36         2  Female

-------------------------------------------------------------------------------
con                                                      Factor 8 Concentration
-------------------------------------------------------------------------------

                  Type: Numeric (float)

                 Range: [43,350]                    Units: 1
         Unique values: 133                     Missing .: 0/235

                  Mean: 136.072
             Std. dev.: 50.5719

           Percentiles:     10%       25%       50%       75%       90%
                             76        99       130       168       202
```

```
------------------------------------------------------------------------
act                                                    Factor 8 Activity
------------------------------------------------------------------------

                  Type: Numeric (float)

                 Range: [60,320]                     Units: 1
         Unique values: 122                      Missing .: 0/235

                  Mean: 137.711
             Std. dev.: 41.8438

           Percentiles:     10%       25%       50%       75%       90%
                             89       107       131       161       196
```

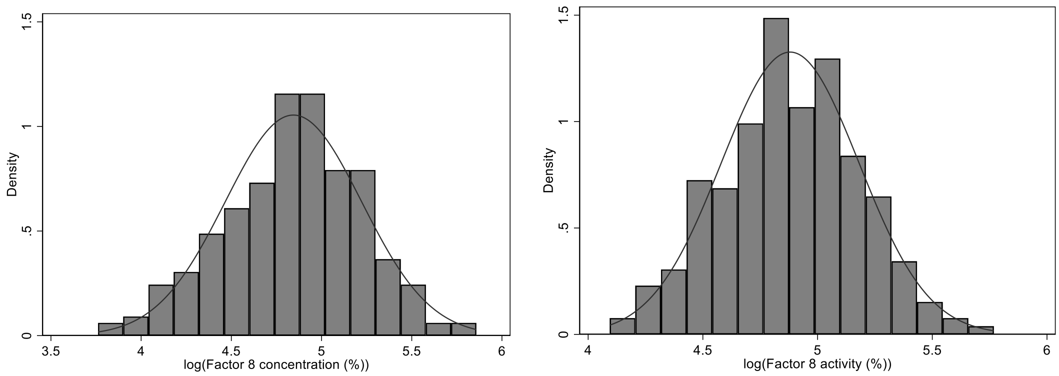
```
. histogram con
. histogram act
```



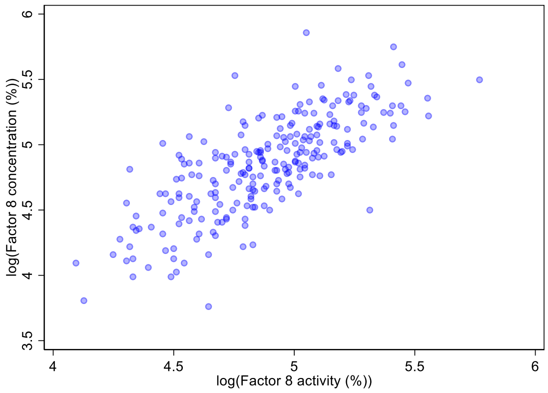
Clear evidence of skewness in the untransformed variables.

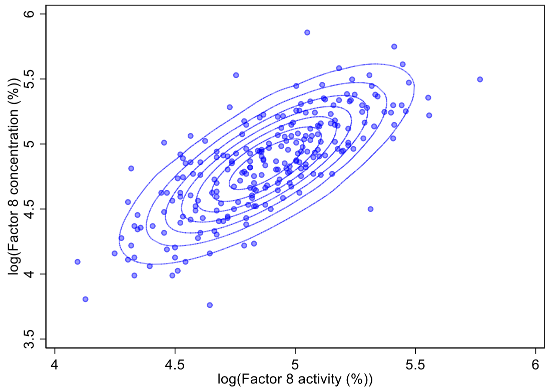
A logarithmic transformation will reduce skewness and it is a ‘natural’ transformation to use with data on a % scale (the log makes 50% and 200% equally far from 100%, which is often intuitively desirable).

```
. gen log_con=log(con)
. gen log_act=log(act)
. label variable  log_con "log(Factor 8 concentration (%))"
. label variable  log_act "log(Factor 8 activity (%))"
. hist  log_con
. hist  log_act
```



Improved normality in log-transformed variables





Visually no evidence against an assumption of bivariate normality. See practical dofile for commands to produce scatter plot and contour plot.

4.2 Association between Factor VIII concentration and activity

```
. corr log_con log_act
(obs=235)

             |  log_con  log_act
-------------+------------------
     log_con |   1.0000
     log_act |   0.7583   1.0000
```

Test of H_0: $\rho = 0$

$$T = r\sqrt{\frac{n-2}{1-r^2}} = 0.7583 \times \sqrt{\frac{233}{1-0.7583^2}} = 17.76 \sim t_{n-2}$$

Comparison with t_{233} gives p<0.0001 .

Calculating a 95% confidence interval for ρ

$$Z_r = tanh^{-1}(r) = \frac{1}{2} log_e\left(\frac{1+r}{1-r}\right) = \frac{1}{2} log_e\left(\frac{1+0.7583}{1-0.7583}\right) = 0.9922$$

$$SE(Z_r) = \sqrt{\frac{1}{n-3}} = 0.0657$$

Thus a 95% CI for Z_ρ is 0.9922 ± 1.96×0.0657 = (0.863, 1.121)

Back transformation gives 95% CI for ρ:

$$\frac{\exp(2\times 0.863)-1}{\exp(2\times 0.863)+1}, \frac{\exp(2\times 1.121)-1}{\exp(2\times 1.121)+1} = \tanh(0.863), \tanh(1.121) = (0.698, 0.808)$$

There is a user written Stata command (`corrci`) that calculates confidence intervals for a correlation coefficient. You will need to install the command first. Type `net search corrci.ado` and click on the latest update to install.

```
. corrci l_con l_act
(obs=235)

              correlation and 95% limits
l_con l_act      0.758    0.698    0.808
```

As calculated by hand!

We conclude from this that the observed data is consistent (at the 5% level of statistical significance) with postulated correlation coefficients in the range 0.70 to 0.81. There is clear evidence of a substantial level of association between these two variables.

```
. bysort sex:corr log_con log_act

-> sex = Male
(obs=199)

             |  log_con  log_act
-------------+------------------
     log_con |   1.0000
     log_act |   0.7620   1.0000


-> sex = Female
(obs=36)

             |  log_con  log_act
-------------+------------------
     log_con |   1.0000
     log_act |   0.7246   1.0000
```

Correlation coefficients very similar in males and females. Given overall CI calculated above it is unlikely that there will be any evidence against the null hypothesis of the true correlations being the same in males and females. Best to test this in a regression model including appropriate interaction terms i.e. to test for different slopes.

4.3 Association between gender and high concentration

```
. gen high=con>150 if con<.
. tab high sex, chi2 exact

           |          Sex
      high |      male     female |     Total
-----------+----------------------+----------
         0 |       135         22 |       157
         1 |        64         14 |        78
-----------+----------------------+----------
     Total |       199         36 |       235

          Pearson chi2(1) =   0.6223   Pr = 0.430
           Fisher's exact =                 0.446
   1-sided Fisher's exact =                 0.272
```

Since the *p*-value (from both tests) is > 0.05 there is no evidence of an association between a 'high' Factor VIII concentration and gender.

The estimated odds ratio ($\hat{\psi}$) relating gender (females vs. males) to a 'high' Factor VIII concentration is $(135 \times 14)/(64 \times 22) = 1.34$.

Calculating a 95% CI for the odds ratio

Working on a logarithmic scale:

$$\log(\hat{\Psi}) = \log\left(\frac{135 \times 14}{64 \times 22}\right) = 0.2944 \qquad \mathrm{SE}(\log(\hat{\Psi})) = \sqrt{\frac{1}{135} + \frac{1}{14} + \frac{1}{64} + \frac{1}{22}} = 0.3741$$

Therefore a 95% CI for log(Ψ) is 0.2944 ± 1.96×0.3741 = (-0.4387, 1.0276) and a 95% confidence interval for Ψ is (exp(-0.4387), exp(1.0276)) = (0.64, 2.79).

Note that this confidence interval could have been computed in Stata using the tabodds command (designed for use in epidemiological case-control studies). The relevant part of the output is as follows.

```
. tabodds high sex, or woolf
--------------------------------------------------------------------------
            |                                 odds          Woolf
        sex |       cases     controls       ratio      [95% Conf. Interval]
------------+-------------------------------------------------------------
       male |          64          135     1.00000              .          .
     female |          14           22     1.34233        0.64486    2.79417
--------------------------------------------------------------------------
```

We conclude from this analysis that the observed data is consistent (at the 5% level of statistical significance) with population odds ratios in the range 0.64 to 2.79. We estimate that the odds of a female having a Factor VIII concentration above 150% are 34% higher than those of a male, but this increase is not statistically significant and is consistent both with a much greater increased risk in females (odds up to 2.79 times that of males) and with somewhat reduced risk (odds reduced by as much as 36% compared with males).