Exercise 6.1

The file **fibrinogen.dta** contains data on some haemostatic factors (factors related to blood clotting) on 235 patients with angina (chest pain). There are three variables:

id: Patient number

fbg: Fibrinogen (g/l): A protein which promotes the clotting of blood; high levels can increase the risk of heart disease.

lys: Lysis time (minutes): The time a blood clot takes to dissolve; long times may also be related to an increased risk of heart disease.

For practical reasons, the lysis time measurements were censored at 5 hours (300 minutes). Thus a value of 300 should be interpreted as '≥ 300'. The value 999 indicates a missing value.

Launch Stata, open a do-file and add commands to change the working directory to the folder containing the data and to load the dataset **fibrinogen.dta**.

a) Mean and variance transformation formulae

Here we will investigate the accuracy of the mean and variance transformation formulae for the *fibrinogen* data. Use `tabstat` or `summarize` to obtain the sample mean $(\hat{\mu})$ and variance $(\hat{\sigma}^2)$ of fibrinogen (fbg). Use the mean and variance transformation formulae (see below and see slides/notes) to calculate the approximate mean and variance for $\log_e$ and square root transformations. Write your answers in the table below.

$$\mathrm{E}[\mathrm{g}(\mathrm{X})] \approx \mathrm{g}(\mu) + \frac{\sigma^2}{2}(\mathrm{g}''(\mathrm{X})|_{\mathrm{x}=\mu}) \qquad \mathrm{Var}[\mathrm{g}(\mathrm{X})] \approx \sigma^2 \times \left[\mathrm{g}'(\mathrm{X})|_{\mathrm{x}=\mu}\right]^2$$

Transformation	Transformation formulae		Sample	
	Mean	Variance	Mean	Variance
$\log_e$				
Square-root				

Create two new variables that are the $\log_e$ and square-root transformations of fibrinogen;

```
gen log_fbg=log(fbg)
gen sqrt_fbg=sqrt(fbg)
```

Use the `summarize` command to obtain the sample mean and variance of the transformed variables and complete the table above. How close are the mean and variance calculated using the transformation formulae to the sample means and sample variances for the transformed variables?

b) Checking normality of fibrinogen

Use `summarize` with the option `detail` to obtain a statistical summary of fbg. What can you deduce about the distribution of fbg from this numerical summary?

Carry out an informal graphical inspection using a normal plot to assess whether fibrinogen is approximately normally distributed. What do you conclude?

```
qnorm fbg, xlab(1(1)5) ylab(1(1)5) aspect(1) ms(oh)
```

Produce a histogram and box plot of fbg. Do you find these plots more or less informative than the normal plot?

```
histogram fbg , width(0.25) normal freq
graph box fbg
```

Using the same commands, judge whether log fibrinogen is more nearly normally distributed than fibrinogen itself. What do you conclude?

Use the `ladder`, `gladder` and `qladder` commands in Stata to assess what would be the best transformation of fbg in order to improve normality.

Calculate the mean and 95% confidence interval for log fibrinogen and back transform to calculate the geometric mean fibrinogen and its 95% confidence interval. Use the command `ameans` to check your results.

c) Checking normality of lysis times

We now turn to the second haemostatic variable lysis times (lys). Use `summarize` and `qnorm` to obtain numerical and graphical descriptions of the distribution of lysis times. What issues are there?

Recode the 999 as a missing value and then construct a normal plot of lysis times. Interpret the plot.

```
mvdecode lys, mv(999)
qnorm lys
```

Does the normal plot exhibit skewness or kurtosis? Try to sketch a histogram of the distribution of lysis times using the information from the normal plot. Use the `histogram` command to check your sketch.

Create a new variable (*log_lys*) that is the natural logarithm of lysis times and construct a normal plot of log lysis times. Try using `gladder` or `qladder`. What conclusions can you draw from this?

Exercise 6.2: Simulating data and distribution plots (optional)

In this exercise we will simulate data from different distributions. We will then produce graphical summaries to inspect the distributions. The goal is to understand the appearance of each plot given data from differently shaped distributions.

a) Generate simulated data

Here we simulate 500 observations from (i) a normal distribution with mean 50 and SD 5, (ii) a uniform(0,1) distribution, (iii) a chi-squared distribution with 8 degrees of freedom, (iv) a negatively skewed beta distribution.

```
clear
set obs 500                    determines the number of observations to be generated
set seed 20201203              enables simulated results to be reproduced

gen n=rnormal(50,5)            random sample from normal distribution with mean=50, SD=5
gen u=runiform(0,1)            random sample from uniform(0,1) distribution
gen c=rchi2(8)                 random sample from chi-squared distribution with 8df
gen b=rbeta(8,2)               random sample from beta distribution with shape parameters 8,2
```

b) Distributional plots

For each of these 4 variables produce a histogram, box plot and normal plot. We will save some time by using a loop.

```
foreach V of varlist n u c b {
histogram `V' , normal name(hist_`V', replace)
graph box `V' , name(box_`V' , replace)
qnorm `V', name(norm_`V' , replace)
}
```

Note that the local macro V evaluates as n on the first loop, u on the second loop etc. The local macro must be punctuated correctly with a left single quote before, and right single quote after.

Which of these plots do you find most helpful in assessing normality?