Exercise 6.1

a). Mean and variance transformation formulae

Obtain mean and variance of fibrinogen

```
. tabstat fbg , s(n mean var)

    variable |         N      mean   variance
-------------+------------------------------
         fbg |       235  2.469362   .2789718
--------------------------------------------

. summ fbg , detail

                       Fibrinogen, g/l
-------------------------------------------------------------
      Percentiles      Smallest
 1%         1.45           1.25
 5%          1.7            1.4
10%         1.85           1.45       Obs                 235
25%          2.1           1.55       Sum of wgt.         235

50%         2.35                      Mean           2.469362
                        Largest       Std. dev.      .5281778
75%          2.8            3.9
90%          3.1              4       Variance       .2789718
95%         3.25            4.2       Skewness       .8484106
99%            4            4.8       Kurtosis          4.591
```

Natural logarithm transformation

$$E[log_e(X)] \approx log_e(\mu) - \frac{\sigma^2}{2\mu^2} \qquad Var[log_e(X)] \approx \sigma^2\left(\frac{1}{\mu}\right)^2 = \frac{\sigma^2}{\mu^2}$$

```
Approx mean of log FBG

. disp log(2.469362)-0.5*0.2789718/2.469362^2
.88108471

Approx variance of log FBG
. disp 0.2789718/2.469362^2
.04574998
```

Square root transformation

$$\mathrm{E}\left[\mathrm{X}^{\frac{1}{2}}\right] \approx \mu^{\frac{1}{2}} - \frac{\sigma^2}{8\mu^{\frac{3}{2}}} \qquad \mathrm{Var}\left[\mathrm{X}^{\frac{1}{2}}\right] \approx \sigma^2\left(\frac{1}{2\mu^{\frac{1}{2}}}\right)^2 = \frac{\sigma^2}{4\mu}$$

```
Approx mean of square root FBG
. disp sqrt(2.469362)-(1/8)*0.2789718*2.469362^(-3/2)
1.5624337

Approx variance of square root FBG
. disp 0.2789718/(4*2.469362)
.02824331
```

Create new transformed variables

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

. tabstat log_fbg sqrt_fbg , s(mean var) col(stat)

    variable |      mean  variance
-------------+--------------------
     log_fbg |  .8820977  .0436069
    sqrt_fbg |  1.562823  .0270617
----------------------------------
```

Transformation	Transformation formulae		Sample	
	Mean	**Variance**	**Mean**	**Variance**
Log_e	0.881	0.0457	0.882	0.0436
Square-root	1.562	0.0282	1.563	0.0271

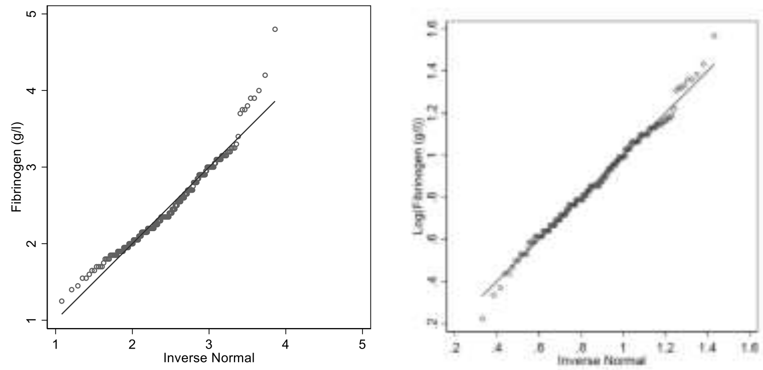
Sample mean and SD for transformed variables are quite close to those from the transformation formulae above.

b). Checking normality of fibrinogen

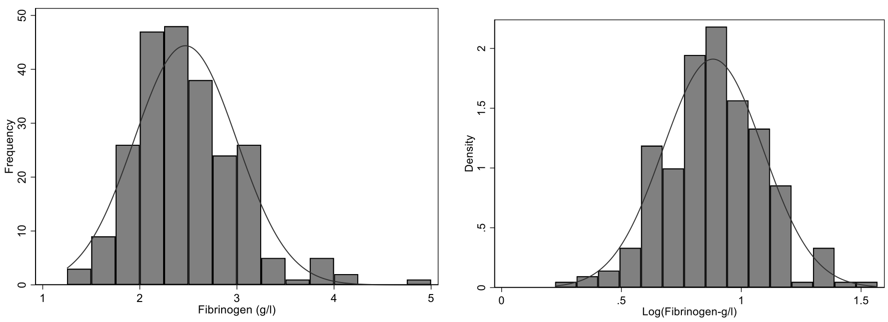
```
. summarize  fbg, detail

                      Fibrinogen (g/l)
-------------------------------------------------------------
      Percentiles      Smallest
 1%         1.45           1.25
 5%          1.7            1.4
10%         1.85           1.45       Obs                 235
25%          2.1           1.55       Sum of Wgt.         235

50%         2.35                      Mean           2.469362
                        Largest       Std. Dev.      .5281778
75%          2.8            3.9
90%          3.1              4       Variance       .2789718
95%         3.25            4.2       Skewness       .8484107
99%            4            4.8       Kurtosis          4.591
```

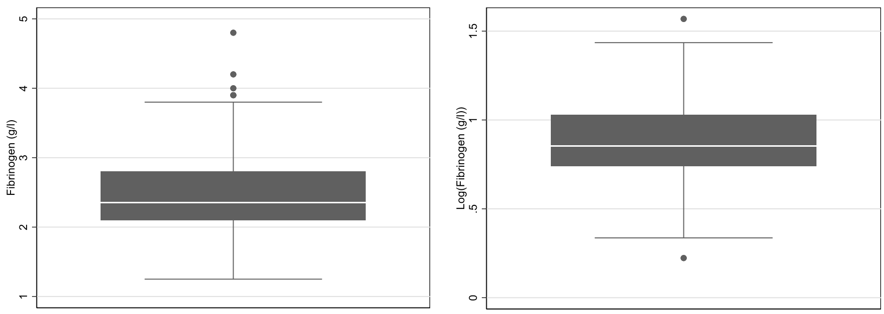
The skew is 0.85 – so slightly positively skewed. The kurtosis is 4.59, slightly greater than 3, indicating heavy tails – this is likely to be mainly to the right as the distribution is positively skewed. Notice that the distance between the 50th and 75th percentiles (0.45) is substantially greater than that between the 50th and 25th percentiles (0.25). This is another indication of skewness.

Normal plots for fibrinogen and log(fibrinogen).



Looking at the normal plot for fibrinogen (left) the largest values are more extreme than they would be under normality and the smallest values are less extreme than they would be under normality. This indicates positive skew. The markers in the normal plot for log fibrinogen lie much closer to the line of equality, i.e. the log transformation reduces skewness and hence improves approximation of the distribution to normal.

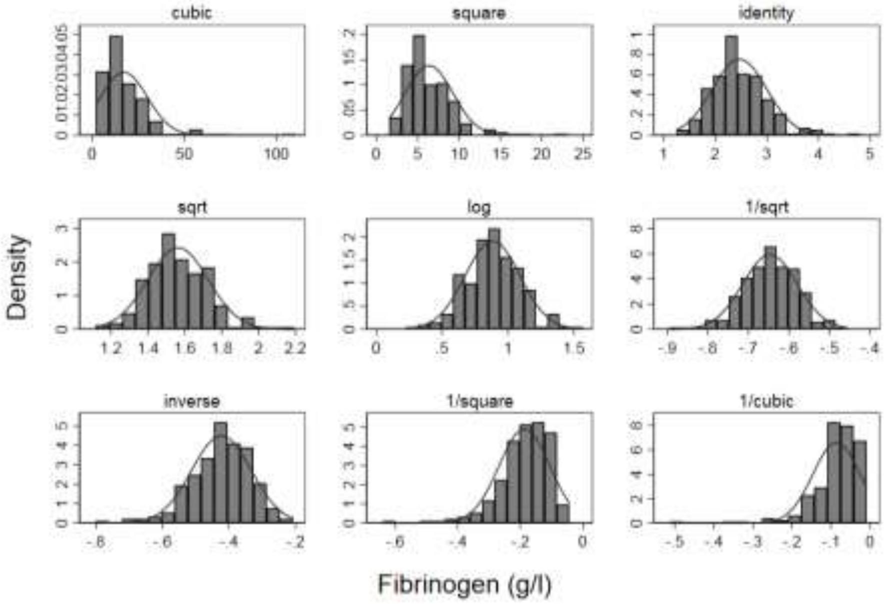


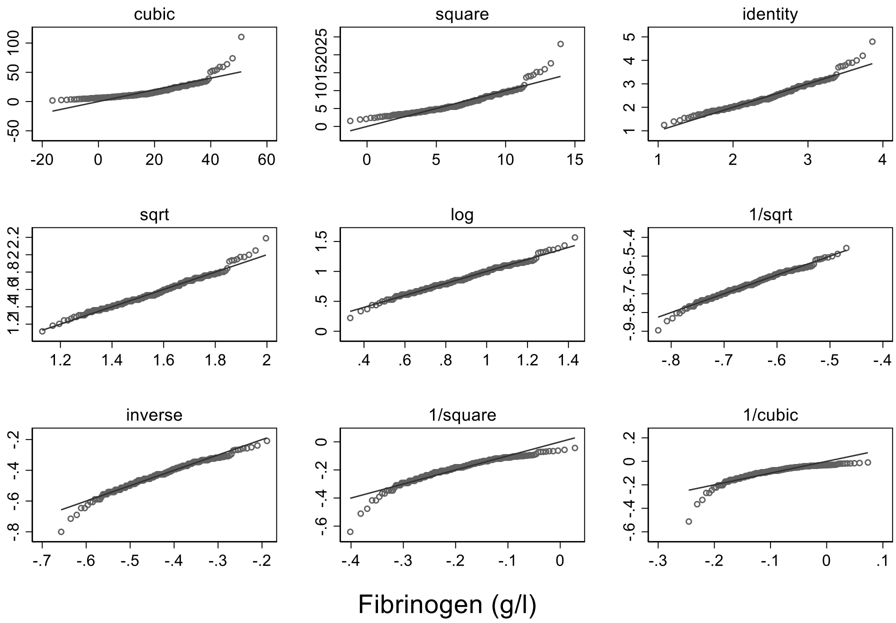


It is not very easy to choose between the untransformed and log-transformed variables from the histogram or box-plots.

Output from the `ladder`, `gladder` and `qladder` commands suggests that the log transformation of fibrinogen provides the best approximation to normality.

```
. ladder fbg

Transformation          formula                 chi2(2)       P(chi2)
---------------------------------------------------------------------
cubic                   fbg^3                         .         0.000
square                  fbg^2                         .         0.000
identity                fbg                       27.18         0.000
square root             sqrt(fbg)                 10.41         0.005
log                     log(fbg)                   1.83         0.401
1/(square root)         1/sqrt(fbg)                5.84         0.054
inverse                 1/fbg                     19.53         0.000
1/square                1/(fbg^2)                 65.46         0.000
1/cubic                 1/(fbg^3)                     .         0.000
```





Quantile-Normal plots by transformation

Geometric mean and 95% CI for fibrinogen

```
. tabstat log_fbg, stat(n mean sd)

    Variable |         N      Mean        SD
-------------+------------------------------
     log_fbg |       235  .8820977  .2088227
--------------------------------------------

. disp exp(0.8820977)
2.4159625

dis invt(234,0.975)
1.9701536

. disp exp(0.8820977-1.9702*0.2088227/sqrt(235))2.3519861
2.3519845

. disp exp(0.8820977+1.9702*0.2088227/sqrt(235))
2.481679
```

Geometric mean [95% CI] = 2.416 [2.352, 2.482]

We can use the ameans command in Stata to check answer.

```
. ameans  fbg

    Variable |    Type        Obs        Mean       [95% Conf. Interval]
-------------+----------------------------------------------------------
         fbg | Arithmetic     235    2.469362         2.401481   2.537242
             |  Geometric     235    2.415962         2.351986   2.481679
             |   Harmonic     235    2.364299         2.302088   2.429965
------------------------------------------------------------------------
```

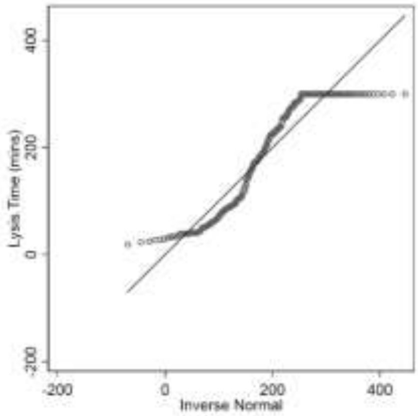
c). Checking normality of lysis times

First need to recode the 999 as a missing value.

```
. mvdecode lys, mv(999)

. summ lys , d

                      Lysis Time (mins)
-------------------------------------------------------------
      Percentiles      Smallest
 1%           25             19
 5%           39             23
10%           43             25       Obs                 234
25%           92             27       Sum of Wgt.         234

50%          204                      Mean           188.5598
                        Largest       Std. Dev.      98.06671
75%          300            300
90%          300            300       Variance        9617.08
95%          300            300       Skewness      -.2646756
99%          300            300       Kurtosis        1.54092
```

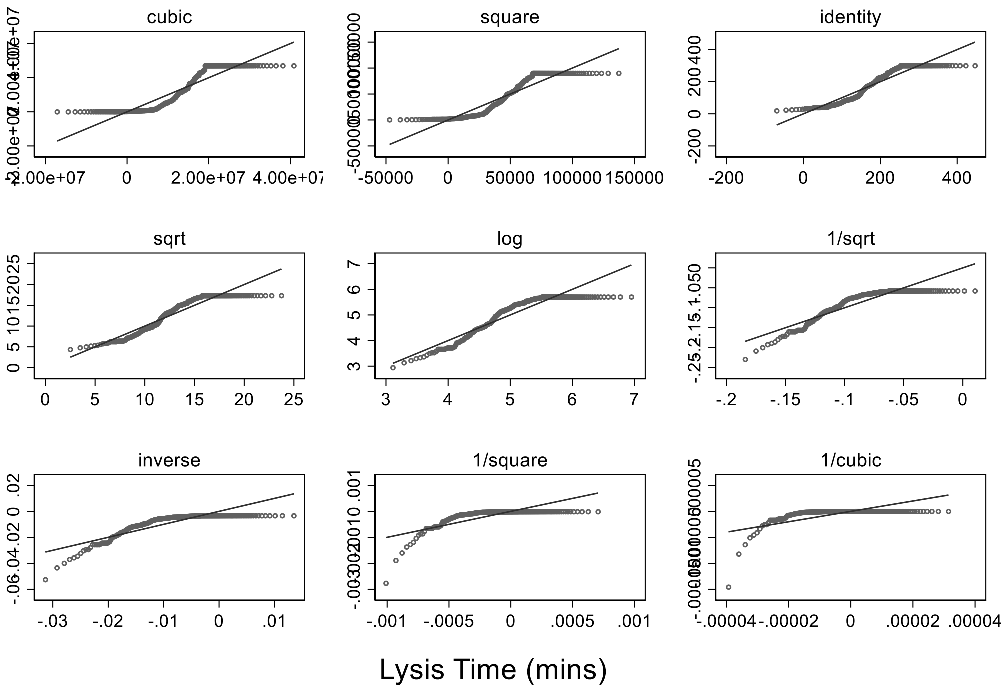
Normal plots for lysis times

```
. qnorm lys  , ms(oh) name(norm_lys, replace) aspect(1)
```



For lysis times (left) extreme observations at both ends of the distribution are not as extreme as they would be were the data normally distributed, i.e. tails are light = kurtosis<3.

No POWER transformation can deal with the cluster of points at 300 minutes.



Quantile-Normal plots by transformation

Exercise 6.2: Simulated data and distributional plots

a). Generate data

Generate data from four different distributions.

```
. clear
. set obs 500
. set seed 20171204
. gen n=rnormal(50,3)
. gen u=runiform(0,1)
. gen c=rchi2(8)
. gen b=rbeta(8,2)
```

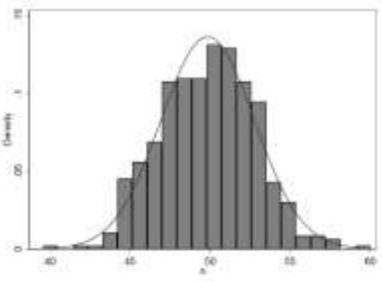
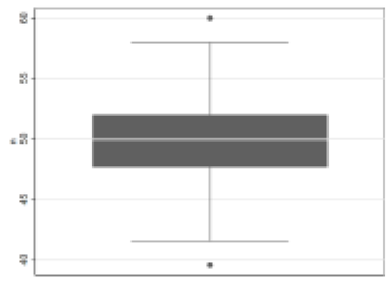
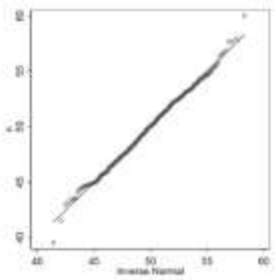
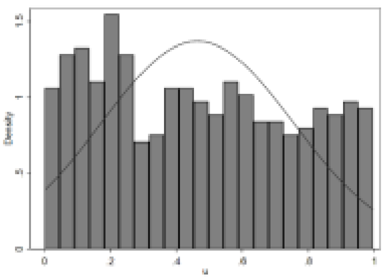
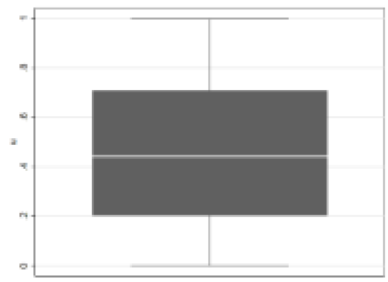
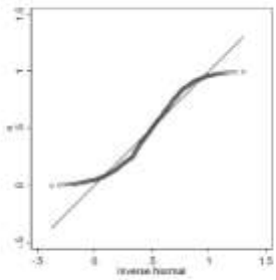
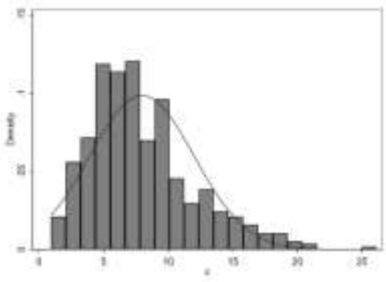
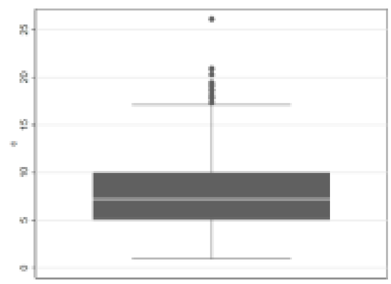
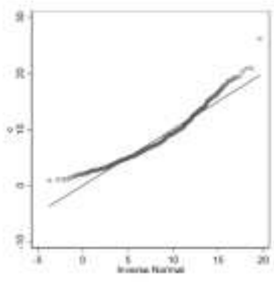
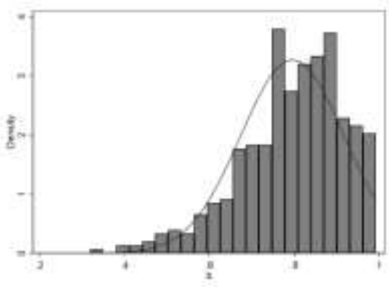
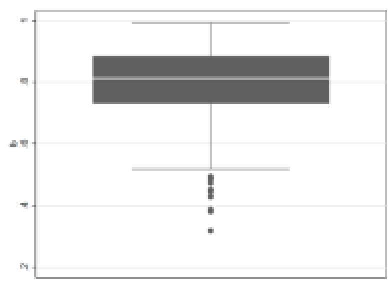
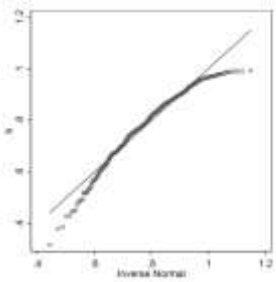
b). Create plots using a loop

```
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) ms(oh) aspect(1)
}
```

```
. tabstat n u c b , s(mean median skew kurtosis) col(stat) format(%4.3f)
```

variable	mean	p50	skewness	kurtosis
n	49.839	49.920	0.013	2.985
u	0.464	0.442	0.170	1.792
c	7.963	7.192	0.967	3.979
b	0.796	0.812	-0.814	3.631

See next page for distributional plots

<table>
<tr><th>Histograms</th><th>Box plots</th><th>Normal plots</th></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td colspan="3">Normal – distribution is symmetric and tails as expected under normal. Note that even with data sampled from a truly normal distribution there will be some departure from line of equality, particularly toward extremes.</td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td colspan="3">Uniform (0,1) – all values lie between 0 and 1. Distribution is symmetric but tails are lighter than expected under normality, i.e. observations at either end of distribution are not as extreme as expected.</td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td colspan="3">Right skewed – asymmetry can be seen in all three plots. Lowest values are less extreme, and largest values are more extreme than expected under normality.</td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td colspan="3">Left skewed – asymmetry can be seen in all three plots. Lowest values are more extreme, and largest values are less extreme than expected under normality.</td></tr>
</table>