STAT505 Assessment #2

1. Water and mosquito fish samples were collected at a sample of 96 sites in the Florida Everglades in the spring of 1996. The resulting data are stored in the file "marsh.txt", available on ANGEL. The data are provided in the following columns: 1) station code (alpha-numeric), 2) longitude, 3) latitude, 4) total mercury, 5) methyl mercury, 6) turbidity, 7) total phosphorus, 8) mercury in fish.

   These data may be input into SAS using the following code:

```
data marsh;
infile "v:\marsh.txt";
input station $ lon lat thg mehg turbid tpw fish;
run;
```

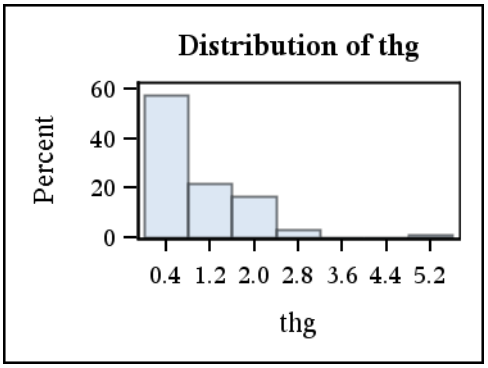
   The "$" sign after station is used to indicate that this is an alpha-numeric variable.

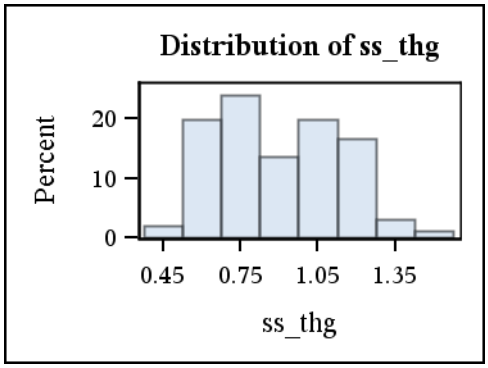
   (a) Produce histograms for total mercury, methyl mercury, turbidity, phosphorus, and fish.

      i. Should a normalizing transformation be applied to any of the variables? If so, what transformation (if any) is most appropriate (e.g., square root, quarter root, log) for each variable.

      ii. If transformations are appropriate, produce histograms for each of the transformed variables.

   The normality of total mercury (thg) is improved with a quarter-root transformation as seen below.



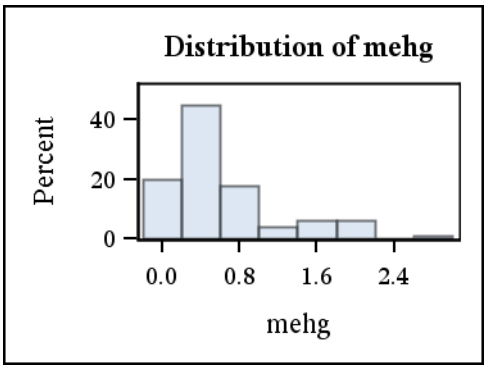


total mercury quarter-root transformation

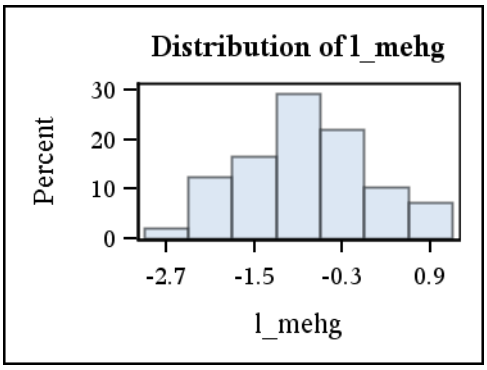
   The relevant SAS code for such a plot is

```
proc univariate data=marsh noprint;
histogram thg;
run;
```

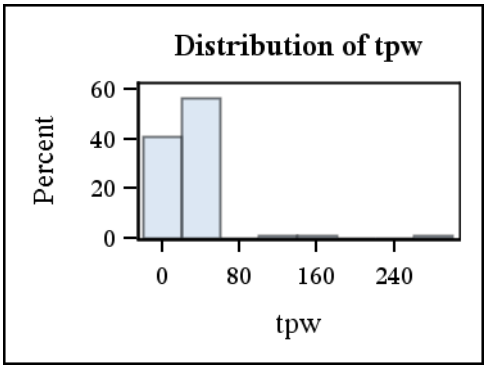
   Other variables are treated similarly.



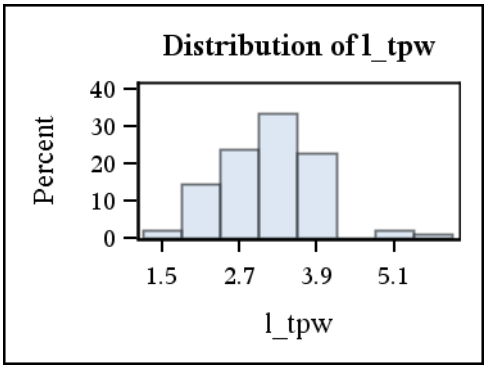
methyl mercury



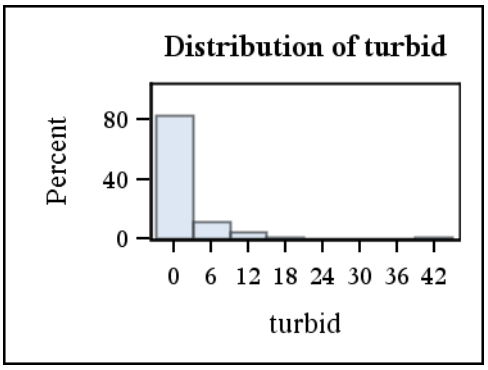
log transformation



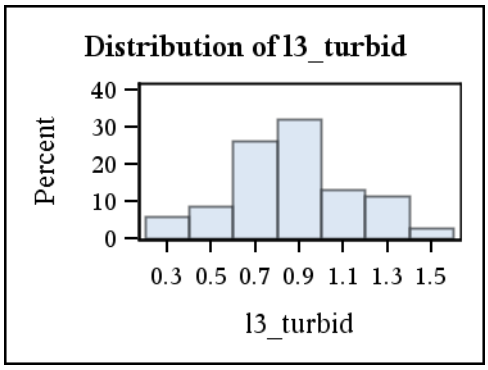
total phosphorus



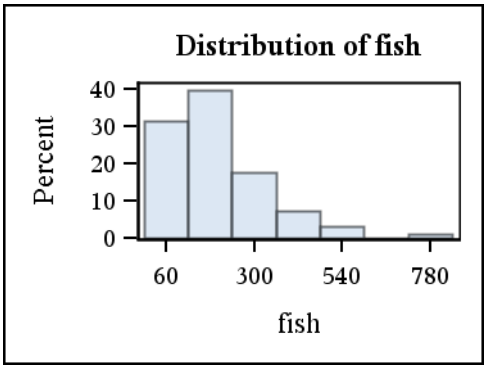
log transformation



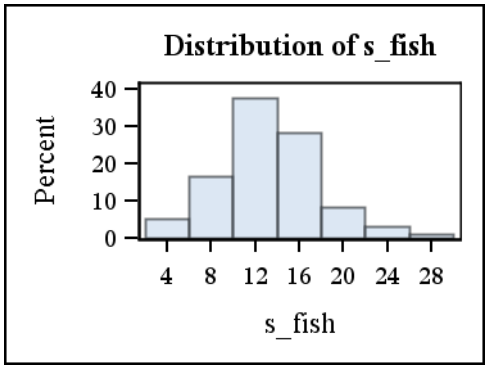
turbidity



log + third-root transformation



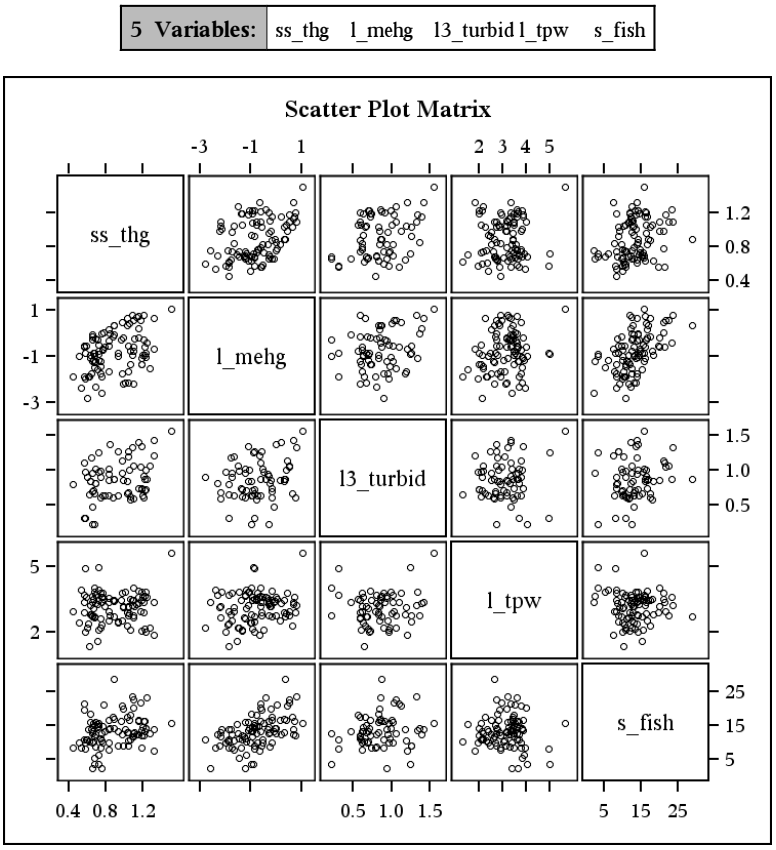
fish



square-root transformation

(b) Using the transformations from part (a), produce a matrix of scatter plots for total mercury, methyl mercury, turbidity, phosphorus, and fish. Are there any outliers? In terms of the strengths of linear relationship, how would you interpret these plots?

The scatterplot matrix is below. There appears to be an outlier in plot *l_tpw* by *l_mehg* and in plot *l_tpw* by *ss_thg*. The station code for both is "M001".

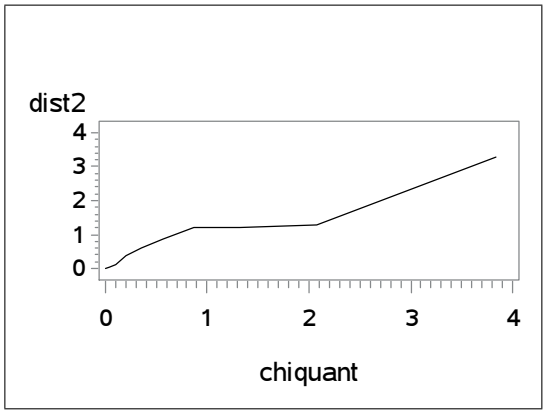


Judging by the clouds of points, the variables *l_mehg*, *s_fish* and *ss_thg* seem to be slightly positively correlated. The others do not appear to be related at all. The SAS code for this plot is

```
proc corr nosimple nocorr data=marsh plots=matrix;
var ss_thg l_mehg l3_turbid l_tpw s_fish;
run;
```

2. (*adapted from J&W Exercise 4.24a*) (the data for this problem is in the file "companies.dat") Exercise 1.4 contains data on three variables for the world's 10 largest companies as of April 2005. For the sales (x_1) and profits (x_2) data, construct a $Q-Q$ plot. Do these data appear to be normally distributed? Explain.

   Plotting squared Mahalanobis distances d_i^2 versus chi-square quantiles $q_{(i)}$ should display a reasonably straight line if the data were sampled from a bivariate normal distribution.



The curvature suggests evidence that these data did not come from such distribution. This plot could also be obtained with the following SAS commands:

```
proc princomp data=e424a std out=pcresult;
  var sales profits;
  run;
data mahal;
  set pcresult;
  dist2=uss(of prin1);
  run;
proc sort;
  by dist2;
  run;
data plotdata;
  set mahal;
  prb=(_n_ -.5)/10;
  chiquant=cinv(prb,1);
  run;
proc gplot;
  plot dist2*chiquant;
  run;
  quit;
```

3. Obtain the 95% prediction ellipse for each of the following covariance matrices:

$$\Sigma_1 = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{bmatrix},$$

$$\Sigma_3 = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 2.0 \end{bmatrix}, \quad \Sigma_4 = \begin{bmatrix} 1.0 & -0.5 \\ -0.5 & 1.0 \end{bmatrix}.$$

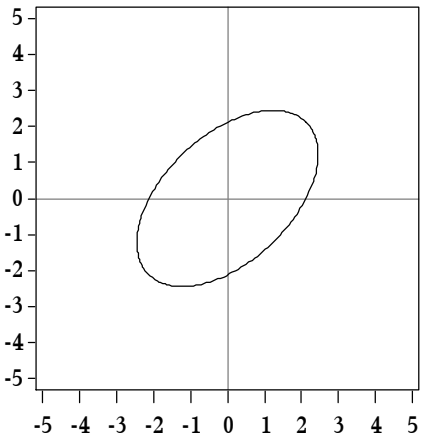
You may assume $\mu_1 = \mu_2 = 0$ in each case. With Σ_1 as a reference, comment on the comparison with

(a) Σ_2?
The variances are equal in both cases, but the covariance is stronger for Σ_2, which is represented by its narrower shape along the 45° line.

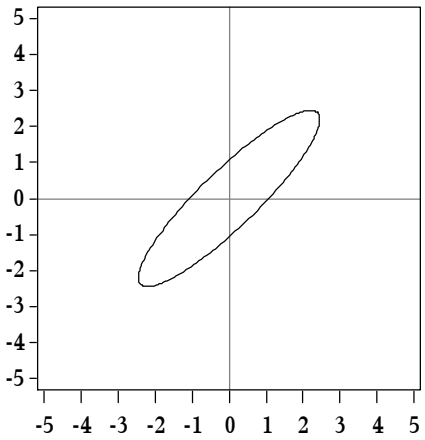
(b) Σ_3?
The covariances are equal in both cases, but the variance for the second variable is larger for Σ_3, which is represented as a wider shape along the horizontal axis.

(c) Σ_4?

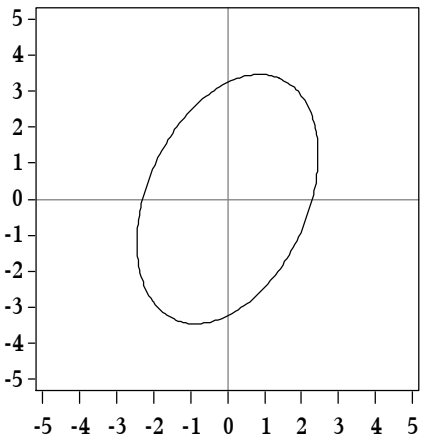
The covariance is negative here, so the ellipse is directed in the perpendicular direction.



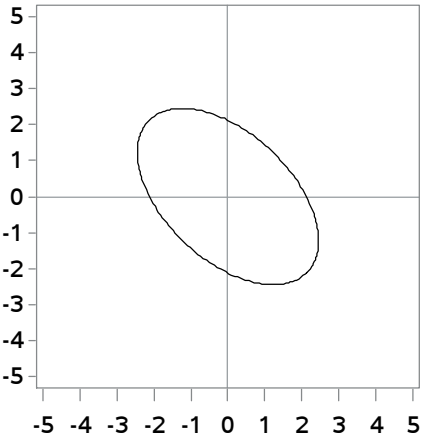
Ellipse for Σ_1



Ellipse for Σ_2



Ellipse for Σ_3



Ellipse for Σ_4

The SAS code below is for plotting Σ_1. The others are similar.

```
data b;
  pi=constant('PI');
  do i=0 to 200;
    theta=pi*i/100; *theta ranges from 0 to 2pi;
    u=cos(theta);
    v=sin(theta);
    output; *theta, u, and v created as variables in data set b;
  end;
  run;
proc iml;
  create c var{x y};
  start ellipse;
    mu={0,0};
    sigma={1.0000 0.5000,
       0.5000 1.0000};
    lambda=eigval(sigma);
    e=eigvec(sigma);
    d=diag(sqrt(lambda));
    z=z*d*e`*sqrt(5.99); *5.99=chisq(2,.95) for 95% probability;
    do i=1 to nrow(z);
      x=z[i,1];
      y=z[i,2];
      append; *adds new values for x and y to data set c;
    end;
  finish;
  use b;
  read all var{u v} into z;
  run ellipse; *assigns (x,y) point for each theta;
  quit;
proc gplot data=c;
  axis1 order=-5 to 5 length=3 in;
  axis2 order=-5 to 5 length=3 in;
  plot y*x / vaxis=axis1 haxis=axis2 vref=0 href=0;
  symbol v=none l=1 i=join color=black;
  run;
```