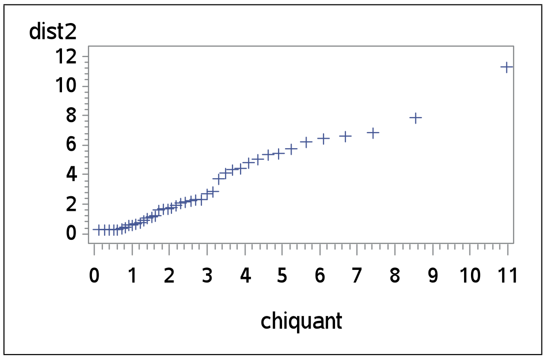
STAT505 Assessment #3

1. (*data from J&W Exercise 1.6*) The data in "air.dat" are 42 measurements on air-pollution variables in Los Angeles. The columns are x_1 = wind, x_2 = solar radiation, x_3 = CO, x_4 = NO, x_5 = NO_2, x_6 = O_3, and x_7 = HC.

   (a) Using notation consistent with our lessons, explain briefly what X_{ij} and μ_j represent in terms of the variables here. Similarly, explain briefly what σ_{jk} represents. What does it mean if $\sigma_{jk} = 0$?

   X_{ij} is ith pollution measurement for the jth variable. μ_j is the population mean of the jth variable, and σ_{jk} is the population covariance between variables j and k. If $\sigma_{jk} = 0$, then variables j and k are uncorrelated.

   (b) Considering only solar radiation, CO (carbon monoxide), and NO_2 (nitrogen dioxide), can these variables reasonably be assumed to be multivariate normal? Justify your answer with appropriate plots.

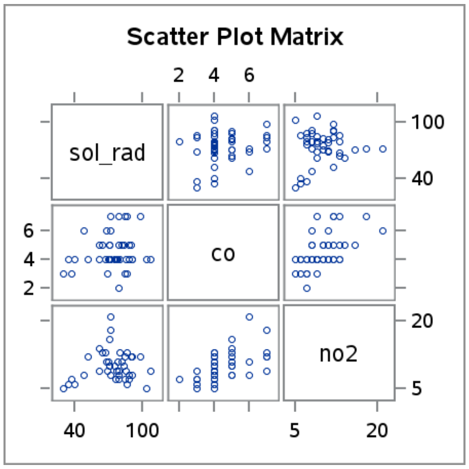
   The "dist2" quantities are the statistical distances between a point $\mathbf{x}_i$ and the sample mean vector $\overline{\mathbf{x}}$, given by

   $$(\mathbf{x}_i - \overline{\mathbf{x}})' S^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}) \tag{1}$$

   If $\mathbf{X}$ is multivariate normal, then the quantities in (1) are approximately chi-square distributed with 3 degrees of freedom and should strongly correlate with quantiles from that distribution. These χ^2_3 quantiles are the "chiquant" values in the above plot. Although there is slight curvature, the overall relationship is reasonably straight line and does not strongly contradict the normality assumption.

   (c) Provide a matrix of scatterplots for the three variables above. Also, report the numeric correlation for each pair of variables. Is any pair significantly correlated? Answer this with separate confidence intervals of correlation. Use a Bonferroni adjustment for multiplicity so that your confidence for all intervals simultaneously is 95%.

   The scatterplot matrix shows strong correlation between CO and NO_2. The other pairs are apparently strong.



The correlations are tabulated below. The confidence levels are Bonferroni adjusted to $1 - .05/3 = 98.3\%$ to account for the multiplicity. Only the interval for CO NO_2 does not include 0, so it alone is considered significant.

Variable	With	Corr	Lower	Upper
sol_rad	co	0.18279	-0.194787	0.513190
sol_rad	no2	0.11573	-0.259830	0.460883
co	no2	0.55658	0.240864	0.765782

2. Measurements of biochemical oxygen demand (Y_1) and suspended solids (Y_2) were obtained from the discharge of $n = 11$ municipal wastewater treatment plants into the rivers of Wisconsin. The values are recorded in the data set “water.dat” and can be read into SAS with the following code. The “firstobs=2” tells SAS to skip the first row, which includes labels.

```
data water;
infile ’v:\water.dat’ firstobs=2;
input trt y1 y2;
run;
```

Assume that these data are sampled from a bivariate normal population with mean vector μ and covariance matrix $\mathbf{\Sigma}$.

(a) Find the sample mean vector and the sample covariance matrix.

$$\overline{\mathbf{x}} = \begin{bmatrix} 34.64 \\ 33.18 \end{bmatrix} \quad \text{and} \quad \mathbf{S} = \begin{bmatrix} 109.25 & 120.37 \\ 120.37 & 363.76 \end{bmatrix}$$

(b) Find 95% confidence intervals for the population means using

i. One at a time multiplier

ii. Bonferroni multiplier

iii. Simultaneous confidence region multiplier

Provide an interpretation of each of these intervals.

i. With 95% confidence, the mean oxygen demand is within 27.61 to 41.66. With 95% confidence, the mean suspended solids is within 20.37 to 45.99.

ii. With 95% confidence, the mean oxygen demand is within 26.34 to 42.94, and simultaneously the mean suspended solids is within 18.04 to 48.33.

iii. With 95% confidence, the mean oxygen demand is within 24.94 to 44.33, and simultaneously the mean suspended solids is within 15.5 to 50.87. These are more conservative than those in part ii because the multiplier may be re-used for any number of intervals without sacrificing confidence.

	loone	upone	lobon	upbon	losim	upsim
Y_1	27.61	41.66	26.34	42.94	24.94	44.33
Y_2	20.37	45.99	18.04	48.33	15.5	50.87

(c) Compute the sample correlation r between oxygen demand and suspended solids.

$$r = 0.60381$$

(d) Test $H_0 : \rho = 0$ against $H_a : \rho \neq 0$ at the $\alpha = 0.01$ level. What are your conclusions?

The test statistic for this is $z = 0.69912$ with p-value .0480. At the .01 level of significance, this is not enough evidence to reject H_0 and conclude that $\rho \neq 0$.

(e) Give a 95% confidence interval for ρ.

With 95% confidence, ρ falls within 0.006166 to 0.883626. Note that this does not include 0 since our p-value was less than .05. However, if we were to compute a 99% confidence interval, it would include 0 in agreement with our conclusion in part (d).

```
proc corr data=water cov nocorr out=corr_out;
  var y1 y2;
  run;
proc iml;
  use corr_out;
  read all var _NUM_ where(_TYPE_="MEAN") into xbar[colname=varnames];
  read all var _NUM_ where(_TYPE_="COV") into S;
  read all var _NUM_ where(_TYPE_="N") into n;
  varnames=t(varNames);
  s2=vecdiag(S);
  n=n[1];
  p=nrow(S);
  xbar=t(xbar);
  t1=tinv(1-.025,n-1);
  tb=tinv(1-.025/p,n-1);
  loone=xbar-t1*sqrt(s2/n);
  upone=xbar+t1*sqrt(s2/n);
  lobon=xbar-tb*sqrt(s2/n);
  upbon=xbar+tb*sqrt(s2/n);
  print varNames loone upone lobon upbon;
  f=finv(0.95,p,n-p);
  losim=xbar-sqrt(p*(n-1)*f*s2/(n-p)/n);
  upsim=xbar+sqrt(p*(n-1)*f*s2/(n-p)/n);
  print varNames losim upsim;
  quit;
proc corr data=water nocorr nosimple fisher(biasadj=no);
  var y1 y2;
  run;
```