STATISTICS DEPARTMENT
M.S. EXAMINATION

PART I
CLOSED BOOK

Friday, November 9, 2001

9:00 a.m. - 1:00 p.m.

School of Science Conference Room, ScN137

Instructions: Complete *all five* problems. Each problem counts 20 points. Unless otherwise noted, points are allocated approximately equally to lettered parts of a problem. Spend your time accordingly.

Begin each problem on a new page. Write the problem number and the page number in the specified locations at the top of each page. Also write your chosen ID code number on every page. Please write only within the black borderlines, leaving at least 1" margins on both sides, top and bottom of each page. Write on one side of the page only.

At the end of this part of the exam you will turn in your answers sheets, but you will keep the question sheets and your scratch paper.

Tables of some distributions are provided. Use them as appropriate.

1. An individual claims that she can identify the Coca-Cola drink when given an unmarked pair of drinks, one being a Coca-Cola and one being a Pepsi-Cola. The individual states that she is not perfect but can make a correct identification of the Cola drink with probability greater than 0.75. We present this individual with 60 pairs of drinks and note the number of correct identifications (out of 60).

(a) Using the binomial table output given below, at least how many correct identifications would this individual need to have, for us to conclude with at least 95% confidence that her claim is true?

(b) Obtain an approximate answer to (a) using only standard normal tables.

(c) Obtain an approximate 90% confidence interval for the P (correct identification) for this individual based on 53 correct identifications out of the 60 pairs.

Binomial Probabilities: n = 60, P(Success) = .75; missing entries have 'exact probability' = '.000000'.

k	P(Y = k)	P(Y<=k)
27	0.000001	0.00000
28	0.000002	0.00000
29	0.000006	0.00001
30	0.000018	0.00003
31	0.000053	0.00008
32	0.000145	0.00022
33	0.000368	0.00059
34	0.000877	0.00147
35	0.001954	0.00342
36	0.004071	0.00749
37	0.007922	0.01542
38	0.014385	0.02980
39	0.024343	0.05414
40	0.038340	0.09248
41	0.056108	0.14859
42	0.076146	0.22474
43	0.095626	0.32036
44	0.110839	0.43120
45	0.118228	0.54943
46	0.115658	0.66509
47	0.103354	0.76844
48	0.083975	0.85242
49	0.061696	0.91411
50	0.040719	0.95483
51	0.023953	0.97879
52	0.012437	0.99122
53	0.005632	0.99685
54	0.002190	0.99904
55	0.000717	0.99976
56	0.000192	0.99995
57	0.000040	0.99999
58	0.000006	1.00000
59	0.000001	1.00000

2. Suppose that X has the probability density function

$$f_X(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Let Y have the probability density function

$$f_Y(y) = \begin{cases} \frac{1}{y^2} & \text{if } y > 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Assume that X and Y are independent.

(a) Find the probability that $2X$ exceeds Y.

(b) Find $E\left[(XY)^{\frac{1}{2}}\right]$.

(c) Let $Z = -\log(X)$. Find the probability density function of Z.

(d) Let $W = \frac{X}{Y}$. Find the probability density of W.

4. Suppose that X has the probability density function $p(x) = P[X = x] = \theta^{x-1}(1-\theta)$ for $0 < \theta < 1$ and $x = 1, 2, \cdots$. Assume that $Y_1, Y_2, \cdots$ are random variables with the probability density function

$$f(y) = \begin{cases} \exp(-y) & \text{for } y > 0 \\ 0 & \text{elsewhere.} \end{cases} \sim \exp(1)$$

Suppose that the random variables X, $Y_1, Y_2, \cdots$ are independent.

(a) Find the probability density function of $Y_1 + Y_2$.

(b) Find $E\left[\sum_{j=1}^{X} Y_j\right]$.

(c) Find $P\left[\sum_{j=1}^{X} Y_j > 10 \mid X = 4\right]$.

(d) Find the moment generating function of $\sum_{j=1}^{X} Y_j$.

(e) Is Y_1 independent of $Y_1 + Y_2$? Prove your answer.

5. It can be shown that

$$\int_0^1 x^{a-1}(1-x)^{b-1}\,dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},\ \textit{where } \Gamma \textit{ is the gamma function.}$$

Using the above result,

(a) write down the density function of a beta random variable X with parameters a and b.

(b) Obtain $\mu = E(X)$, the expected value or population mean of X.

(c) Obtain the population variance of X and show that it is equal to $[\mu(1-\mu)]/[K+1]$, where $K = a + b$.

(d) Suppose that X refers to the proportion of a store's inventory unsold during a particular week. Over a fifteen week period, the results are

0.1066
0.0000
0.5828
0.4562
0.4301
0.3912
0.5557
0.4625
0.5904
0.3975
0.3097
0.5893
0.5332
0.2652
0.6582 .

Note: Sum of observations = 6.3286, Sum of squares of observations = 3.1575.

Using the naive approach (referred to as 'method of moments') of setting the population mean and variance equal to the sample mean and variance, fit the above X to a beta model and determine values for a and b.

#1 CB Solution

(a) Let X = # of correct identification

51, since $P(X \geq 51) = 1 - P(X \leq 50) = 1 - .95483 \leq .05$.

(b) need $$Z = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} = \frac{X - 60(.75)}{\sqrt{60(.75)(.25)}} = \frac{X - 45}{3.354} \geq 1.65$$

or $X \geq 50.53 \simeq 51$. (same answer as above)

(c) $$\underbrace{\hat{\pi} - 1.65\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{60}}}_{\simeq .81} \leq \pi \leq \underbrace{\underset{\overset{\uparrow}{\hat{\pi}}}{\boxed{\frac{53}{60}}} + 1.65\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{60}}}_{.95}$$

$\hat{\pi} \simeq .883$

CB Solution

4. Suppose that X has the probability density function

$$f_X(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Let Y have the probability density function

$$f_Y(y) = \begin{cases} \frac{1}{y^2} & \text{if } y > 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Assume that X and Y are independent.

(a) Find the probability that $2X$ exceeds Y. $P(2X > Y)$

(b) Find $E\left[(XY)^{\frac{1}{2}}\right]$.

(c) Let $Z = -\log(X)$. Find the probability density function of Z.

(d) Let $W = \frac{X}{Y}$. Find the probability density of W.

DF

1. Suppose that X has the probability density function

$$f_X(x) = \begin{cases} 1 & \text{if } 0<x<1 \\ 0 & \text{elsewhere.} \end{cases}$$

Let Y have the probability density function.

$$f_Y(y) = \begin{cases} \frac{1}{y^2} & \text{if } y>1 \\ 0 & \text{elsewhere.} \end{cases}$$

Assume that X & Y are independent

(a) Find the probability that $2X$ exceeds Y.

(b) Find $E[(XY)^{\frac{1}{2}}]$.

(c) Let $Z = -\log(X)$. Find the probability density function of Z.

(d) Let $W = X/Y$. Find the probability density function of W.

#3
c B
F.S

3 a) $L(\theta) = f(x_1, \ldots, x_n | \theta) = \prod^n f(x_i | \theta)$

$$= \prod^n (\theta+1) x_i^\theta = (\theta+1)^n \left[\prod^n x_i\right]^\theta$$

$$l(\theta) = n \log(\theta+1) + \theta \sum_i^n \log(x_i)$$

$$\frac{d}{d\theta} l(\theta) = \frac{n}{\theta+1} + \sum^n \log(x_i) = 0$$

$$\frac{n}{\theta+1} = -\sum^n \log(x_i)$$

$$\hat{\theta} = -1 - \frac{n}{\sum \log(x_i)}$$

b) $f(x|\theta) = (\theta+1) x^\theta$

$$\log f(x|\theta) = \log(\theta+1) + \theta \log(x)$$

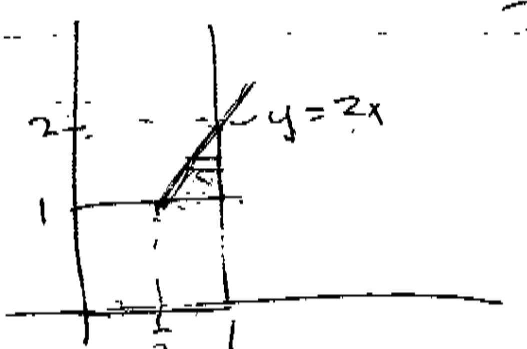
$$\frac{d}{d\theta} \log f(x|\theta) = \frac{1}{\theta+1} + \log(x)$$

$$\frac{d^2}{d\theta^2} \log f(x|\theta) = -(\theta+1)^{-2} = -\frac{1}{(\theta+1)^2}$$

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\right] = \frac{1}{(\theta+1)^2}$$

$$AV = \frac{(\theta+1)^2}{n}$$

(a)



$$f_{X,Y}(x,y) = \frac{1}{y^2} I_{(0,1)}(x)\, I_{(1,\infty)}(y)$$

$$P[Y \le 2X]$$

$$= \int_1^2 \int_{\frac{y}{2}}^1 \frac{1}{y^2}\, dx\, dy$$

$$= \int_1^2 \frac{1}{y^2}\left(1-\frac{y}{2}\right) dy$$

$$= -\frac{1}{y} - \frac{1}{2}\ln(y)\Big|_1^2$$

$$= (1-\frac{1}{2}) = \frac{1}{2}(\ln(2) - \ln(1$$

(b) $$E\left[(XY)^{\frac{1}{2}}\right] = E(X^{\frac{1}{2}})\, E(Y^{\frac{1}{2}})$$

$$= \int_0^1 x^{\frac{1}{2}}\, dy \int_1^\infty y^{\frac{1}{2}}\, y^{-2}\, dy$$

$$= \left[\frac{1}{\frac{1}{2}+1} x^{\frac{1}{2}+1}\Big|_0^1\right]\left[\frac{y^{-2+\frac{1}{2}+1}}{-2+\frac{1}{2}+1}\Big|_1^\infty\right]$$

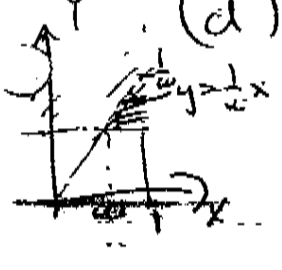
$$= \left[\frac{2}{3}\right]\left[-\frac{1}{-\frac{1}{2}}\right] = \frac{4}{3}$$

(c) $$F_Z(z) = P[Z \le z] = P[-\log(X) \le z]$$

$$= P[X \ge e^{-z}] = 1 - e^{-z} \text{ so } f_Z(z) = e^{-z}$$

$0 < w < 1 \Rightarrow$

(d) $$1 - F_W(w) = P[X/Y > w] = P[Y < \frac{1}{w} X]$$



$$= \int_1^{\frac{1}{w}} \int_{wy}^1 \frac{1}{y^2}\, dx\, dy = \int_1^{\frac{1}{w}} (1-wy)\frac{1}{y^2}\, dy$$

$$= \left(-\frac{1}{y}\right)\Big|_1^{\frac{1}{w}} - w\ln(y)\Big|_1^{\frac{1}{w}} = 1 - w - w\ln(\tfrac{1}{w})$$

$$f_W(w) = 1 + \ln(w$$

$$= -\ln(w$$

c) $\hat{\theta} \pm z_{\alpha/2} \frac{1}{\sqrt{n I(\hat{\theta})}}$

$\hat{\theta} \pm z_{\alpha/2} \frac{\hat{\theta}+1}{\sqrt{n}}$

d) $n=12$ is a small sample.

Parametric Bootstrap:

Get $\hat{\theta}$ MLE

① $x_i^* \sim f(x|\hat{\theta}) = (\hat{\theta}+1)x^{\hat{\theta}}$

$i=1,\dots,B$

② List of $\hat{\theta}_i^* = -\frac{n}{\sum \log(x_i^*)} - 1$

$i=1,\dots,B$

③ $s_{\hat{\theta}} = \sqrt{\frac{\sum(\hat{\theta}_i^* - \bar{\theta}^*)^2}{B-1}}$

④ $[\hat{\theta}^*_{(.025)}, \hat{\theta}^*_{(.975)}]$

DF Solution (15 VP

4 2. Suppose that X has the probability density function $p(x) = P[X = x] = \theta^{x-1}(1-\theta)$ for $0 < \theta < 1$ and $x = 1,2,\cdots$. Assume that $Y_1, Y_2, \cdots$ are random variables with the probability density function

$$f(y) = \begin{cases} \exp(-y) & \text{for } y > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

Suppose that the random variables X, $Y_1, Y_2, \cdots$ are independent.

(a) Find the probability density function of $Y_1 + Y_2$.

(b) Find $E\left[\sum_{j=1}^{X} Y_j\right]$.

(c) Find $P\left[\sum_{j=1}^{X} Y_j > 10 \;\middle|\; X = 4\right]$.

9 (d) Find the moment generating function of $\sum_{j=1}^{X} Y_j$.

(e) Is Y_1 independent of $Y_1 + Y_2$? Prove your answer.

DF

4 Suppose that X has the probability function $p(x) = P[X=x] = \theta^{x-1}(1-\theta)$ for $x = 1, 2, \ldots$. Assume that $Y_1, Y_2, \ldots$ are random variables with the probability density function

$$f(y) = \begin{cases} \exp(-y) & y > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

Suppose that the random variables $X, Y_1, Y_2, \ldots$ are independent.

(a) Find the probability density function of $Y_1 + Y_2$.

(b) Find $E\left(\sum_{j=1}^{X} Y_j\right)$

(c) Find $P\left[\sum_{j=1}^{X} Y_j > 10 \,\middle|\, X = 4\right]$

(d) Find the moment generating function of $\sum_{j=1}^{X} Y_j$.

(e) Is Y_1 independent of $Y_1 + Y_2$? Prove your answer.

4. (a) $f_{Y_1+Y_2}(z) = z\exp(-z)\, I_{(0,\infty)}(z)$

pf. m.g.f of Y_1+Y_2 is $\frac{1}{(1-t)^2}$

& this is m.g.f of $z\exp(-z)\, I_{(0,\infty)}(z)$

(b) $E(\sum_{j=1}^{X} Y_j) = E(X)\, E(Y_1)$

$= \frac{1}{1-\theta}\, 1 = \frac{1}{1-\theta}$

(c) $P[\sum_{j=1}^{X} Y_j > 10 \mid X=4]$

$= P[\sum_{j=1}^{4} Y_j > 10]$

m.g.f of $\sum_{j=1}^{4} Y_j = \frac{1}{(1-t)^4}$

$E(e^{tY_1})$
$= \int_0^\infty e^{ty}\, e^{-y} dy$
$= \frac{1}{1-t}\int_0^\infty (1-t)\, y e^{-(1-t)y} dy$
$= \frac{1}{1-t}$

$\therefore$ density of $\sum_{j=1}^{4} Y_j = U$ is $\frac{u^3 e^{-u}}{3!} I_{(0,\infty)}(u)$

$P[U > 10] = \sum_{j=0}^{3} \frac{10^j e^{-10}}{j!}$
integ by Parts

(d) $E(e^{t(\sum_{i}^{X} Y_i)}) = \sum_{x=1}^{\infty} E[e^{tY_1}]^x\, \theta^{x-1}(1-\theta)$

$= (1-\theta)E(e^{tY_1}) \sum_{x=1}^{\infty} (\theta E(e^{tY}))^{x-1} = \frac{(1-\theta)E(e^{tY_1})}{1-\theta E(e^{tY_1})} = \frac{(1-\theta)\frac{1}{(1-t)}}{1-\theta\frac{1}{(1-t)}}$.

(e) No. $\mathrm{Cov}(Y_1, Y_1+Y_2) = \mathrm{Var}(Y_1) > 0$

$\therefore$ not indep.

#5 Solutions

(a)

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0<x<1$$
$$= 0, \text{ elsewhere}$$

(b)

$$\mu = E(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^{a}(1-x)^{b-1}\,dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}$$
$$= \frac{a}{a+b}$$

(c) $E(x^2) = \dfrac{(a+1)a}{(a+b+1)(a+b)}$

$$Var(x) = \frac{(a+1)a(a+b) - a^2(a+b+1)}{(a+b+1)(a+b)^2} = \frac{a[(a+1)(a+b) - a(a+b+1)]}{(a+b+1)(a+b)^2} = \frac{ab}{(a+b+1)(a+b)^2}$$
$$= \frac{\mu(1-\mu)}{K+1}$$

(d) 1^{ST} $\bar{x} = .4219$, $s^2 = (.18[illegible])^2$

Set $\mu = \dfrac{a}{a+b} = \dfrac{a}{K} = .4219$ or $a = .4219K$

Set $Var(x) = \dfrac{\mu(1-\mu)}{K+1} = (.18[illegible])^2$

or $K = \left[\dfrac{\mu(1-\mu)}{(.18[illegible])^2} - 1\right] = \dfrac{(.4219)(.5781)}{(.18[illegible])^2} - 1$

$= \underline{\underline{6}}$

Thus $a \cong .42(6) = 2.52$
$b = 6 - a \cong 3.48$

STATISTICS DEPARTMENT
M.S. EXAMINATION

PART II
OPEN BOOK

Tuesday, November 13, 2001

9:00 a.m. - 1:00 p.m.

Statistics Department Computer Lab, SC S152

Instructions: Complete *all five* problems. Each problem counts 20 points. Unless otherwise noted, points are allocated approximately equally to lettered parts of a problem. Spend your time accordingly.

The web site address for data and program files for this exam is:

http://www.telecom.csuhayward.edu/~esuess/MSexam/MSexam.htm

Begin each problem on a new page. Write the problem number and the page number in the specified locations at the top of each page. Also write your chosen ID code number on every page. Please write only within the black borderlines, leaving at least 1" margins on both sides, top and bottom of each page. Write on one side of the page only.

At the end of this part of the exam you will turn in your answer sheets, but you will keep the question sheets and your scratch paper.

You may use a computer to work any of the problems, but your answers must be handwritten on standard paper provided for the examination. Printers may *not* be used during the exam, and pages printed out by computer may *not* be submitted. As indicated, some problems have data files available on disk.

1. A professional wine tasting is conducted for "brands" A, B, C, and D of 1998 zinfandel wine. The brands are all produced in the same region and are similar in price. Each of four judges, 1, 2, 3, and 4, rate the wines independently of the other judges and without being told which wineries make the four brands. The results are as follows:

Judge	A	B	C	D
1	77	93	85	92
2	86	93	76	86
3	81	89	82	81
4	85	97	75	78

(No computer file available for these data. Type carefully.)

Each score shown above is a sum of five component ratings for different properties of the wine. The component scores can range from 0 to 20, so the total scores can range from 0 to 100. Higher ratings correspond to better wines. Because judging is not an exact process and because of variations from bottle to bottle of the same brand, it is supposed that slightly different results would be obtained if this wine tasting were repeated the next day. The main purpose of the tasting is to determine whether some of the wines are clearly judged superior to others in spite of such random variations.

(a) What is the name usually given to the statistical procedure appropriate for analyzing these data in the circumstances described. State an appropriate statistical model and null hypothesis that match the "main purpose" stated above.

(b) In terms of your model, what does it mean to say that the scores are normally distributed? What feature of the description above leads you to believe that the scores may be approximately normally distributed.

(c) Assuming that the data are normal as stated in part (b), carry out the appropriate analysis and say whether there are any statistically significant differences (5% level) among the four brands.

(d) If there are significant differences among the wines, perform an appropriate procedure to identify the best one(s). If you found no significant differences, then what is the probability that a true difference of 5 points or greater went undetected?

(e) The organization that conducted this wine tasting is interested to know whether some judges give systematically higher or lower scores than others. Do the data give an opportunity to test for this? If so, give the result. If not, explain why not.

(f) In spite of your answer to part (b), suppose you doubt that the scores are really normal and wish to perform a procedure to answer part (c) that does not assume the data to be normal. Give the name of an appropriate alternate procedure and state the result. Does this procedure allow you to answer part (e)? If so, give the result. If not, explain why no such result can be obtained.

\
58

Answers:

(a) This is a randomized block design. Model: $Y_{ij} = \mu + \alpha_i + B_j + e_{ij}$; where $i = 1, 2, 3, 4$; $j = 1, 2, 3, 4$, and e_{ij} are i.i.d. NORM(0, σ^2). Here α indicates the brand effect, which we consider to be fixed, because the particular brands of wine are of interest. Also, B indicates the judge effect, which we consider to be random, assuming that judges were selected at random. (No points off for failure to discuss whether the effects are fixed or random because in this simple model it doesn't matter; if correctly addressed this can be used to offset a small flaw elsewhere in this problem.) Interaction is confounded with error because there is one observation per cell. (No credit for *this part* for saying it's a one-way ANOVA; the rest graded on the basis on consistency. Only token credit for *the whole problem* for saying it's a chi-squared contingency table; even if consistently computed; the rest of the parts clearly show that's a nonsense interpretation.)

(b) Because of the Central Limit Theorem, one would expect the sum of five components to be nearly normal.

(c) The Minitab printout for the appropriate ANOVA is as follows:

```
Analysis of Variance for Score

Source        DF         SS          MS        F       P
Brand      4-1 3     408.50      136.17     5.23   0.023
Judge      4-1 3      30.00       10.00     0.38   0.767 > 0.05
Error  [illegible] 9     234.50       26.06
Total         15     673.00
          16-1
```

The Brand effect is significant at the 5% level, but not at the 1% level.

(d) Row and column means are as follows:

```
Brand         A         B         C         D              Mean
-----------------------------------------------------------------
 1       77.000    93.000    85.000    92.000            86.750
 2       86.000    93.000    76.000    86.000            85.250
 3       81.000    89.000    82.000    81.000            83.250
 4       85.000    97.000    75.000    78.000            83.750
-----------------------------------------------------------------
Mean     82.250    93.000    79.500    84.250            84.750
```

Multiple comparisons needed. *One* possible multiple comparison procedure would be Fisher's LSD: LSD = $t_{0.025,9}$ (2 MS(Err)/b)$^{1/2}$ = 2.262 [2(26.06)/4]$^{1/2}$ = 8.17 < |93.00 – 84.25| = 8.75; also, LSD > other absolute differences. Thus, the summary diagram is:

B(93.00) $\underline{\text{D(84.25) A(82.25) C(79.50)}}$

That is, Brand B is rated significantly better than any of the others, which do not differ significantly among themselves.[It is *wrong* to use a one-way procedure, Minitab's or any other, to find LSD; these use the inflated MS(Err) of a one-way ANOVA instead of the correct MS(Err) of the ANOVA for the block design. Anyhow, LSD$_{\text{one-way}}$ would not find even the one difference detected above. The penalty for this unthinking use of the wrong procedure, leading to a nonsensical answer, is half the points for this part.]

(e) The ANOVA table in (c) shows that the Judge effect is not significant.

(f) The Friedman test can be used to test whether population median scores are different for the four Brands. Adjusted for ties, the result is (just barely) significant at the 5% level. Because this test is based on ranks for each row separately, it is not possible to use it to test for differences among Judges.

2. The dataset *cities.txt* contains $n = 49$ data points, each corresponding to a city in the United States, the pairs being 1920 and 1930 populations in thousands of each city, which we will denote by u and x respectively. The dataset *cities.txt* and an Splus program *cities.ssc* are available at

   http://www.telecom.csuhayward.edu/~esuess/MEexam/MSexam.htm

   (a) Using the statistical software package of your choice compute the relevant summary statistics and plot the relevant graphical presentations for estimating the mean difference in city populations between 1930 and 1920.

   (b) Compute a 95% confidence interval for $\mu_{1930} - \mu_{1920}$ using the appropriate formula. Verify the normality assumption for your CI calculation.

   (c) Using the Splus code to produce a bootstrap CI for $\mu_{1930} - \mu_{1920}$.

      i. Explain how to use the nonparametric bootstrap to estimate $\mu_{1930} - \mu_{1920}$.

      ii. Describe what the code is doing.

      iii. Run the code and plot a histogram of the bootstrapped mean difference.

      iv. Report the bootstrap estimate of $\mu_{1930} - \mu_{1920}$ and report the bootstrap CI for $\mu_{1930} - \mu_{1920}$.

      v. Compare your results with the CI you computed above.

      vi. Does there appear to be a statistically significant increase in the mean population in 1930 from 1920? Explain.

   (d) Suppose that we are now interested in the ratio of means because this would enable us to estimate the total population in 1930 from the 1920 figure. If the cities form a random sample with (U, X) denoting the pair of population values for a randomly selected city, then the total 1930 population is the product of the total 1920 population and the ratio of expectations, $\theta = E[X]/E[U]$. By examining the scatterplot of the 1930 versus 1920 populations, there is no obvious parametric model for the joint distribution of (U,X), so it is natural to estimate θ by the estimator $T = \bar{X}/\bar{U}$. We might also be interested in the uncertainty in T. Use the Splus code to produce a bootstrap confidence interval for θ.

      i. Explain how to use the nonparametric bootstrap to estimate θ.

      ii. Run the code and plot a histogram of the bootstrap ratios.

      iii Get the bootstrap estimate of θ and a 95% bootstrap CI for θ.

      iv. Comment on what would be needed to determine the comparable large sample CI for θ.

```
### Bootstrap the cities data.

cities <- matrix(c(138, 143, 93, 104, 61,  69, 179, 260, 48,  75,
37,  63, 29,  50, 23,  48,
 30, 111,  2,  50, 38,  52, 46,  53, 71,  79, 25,  57, 298, 317,
 74,  93, 50,  58, 76,  80,
381, 464, 387, 459, 78, 106, 60,  57, 507, 634, 50, 64, 77, 89,
64,  77, 40,  60, 136, 139, 243, 291, 256, 288, 94,  85, 36,  46,
45,  53, 67,  67, 120, 115, 172, 183, 66,  86, 46,  65, 121, 113,
44,  58, 64,  63, 56, 142, 40,  64, 116, 130, 87, 105, 43,  61,
43,  50, 161, 232, 36,  54), byrow = T, ncol=2)

n <- 49

u <- cities[,1] x <- cities[,2]

# scatter plot of

plot(u,x)

# mean difference

d <- mean(u) - mean(x)

# ratio of means

t <- mean(u)/mean(x)

# Bootstrap difference and ratio

B <- 1000 # number of bootstrap samples

cities.boot <- matrix(NA,ncol=2,nrow=n)
    # storage matrix for bootstrap sample pairs

diff.boot <- numeric(B)
    # vector for bootstrap differences in the means
ratio.boot <- numeric(B)
    # vector for bootstrap ratio in the means

for(i in 1:B){
    i.boot <- sample(seq(1:49), replace=T)
            # sample the index vector [1,2,...,49] with replacement
    for(j in 1:n){
        cities.boot[j,] <- cities[i.boot[j],]
```

```
            # create the matrix of sampled pairs
    }
    diff.boot[i] <- mean(cities.boot[,2]) - mean(cities.boot[,1])
            # calculation of the bootstrap differences in the means
    ratio.boot[i] <- mean(cities.boot[,2])/mean(cities.boot[,1])
            # calculation of the bootstrap ratios in the means
}

# Bootstrap analysis of differences

mean(diff.boot)
            # bootstrap estimate of the difference in the means

sqrt(var(diff.boot))
            # bootstrap standard error of the difference in the means

hist(diff.boot)
            # bootstrap distribution of the difference in the means

quantile(diff.boot, c(0.025,0.975))
            # calculation of the empirical bootstrap confidence interval

# Bootstrap analysis of ratios

mean(ratio.boot)
            # bootstrap estimate of the ratio in the means

sqrt(var(ratio.boot))
            # bootstrap standard error of the ratio in the means

hist(ratio.boot)
            # bootstrap distribution of the ratio in the means

quantile(ratio.boot,c(0.025,0.975))
            # calculation of the empirical bootstrap confidence interval
```

#2
DB
ES

2a

b. (16.43, 32.88)

c i) resample pairs, differences $\bar{x} - \bar{u}$

(i)

(ii)

(v) bootstrap estimate $\hat{\mu}_{1930} - \hat{\mu}_{1920} = 24.57$

bootstrap CI (17.22, 32.94)

v) same

vi) yes

d. i) resample pairs, ratios $\bar{x}/\bar{u}$

ii)

iii) bootstrap estimate $\theta = 1.24$

bootstrap CI (1.176, 1.31)

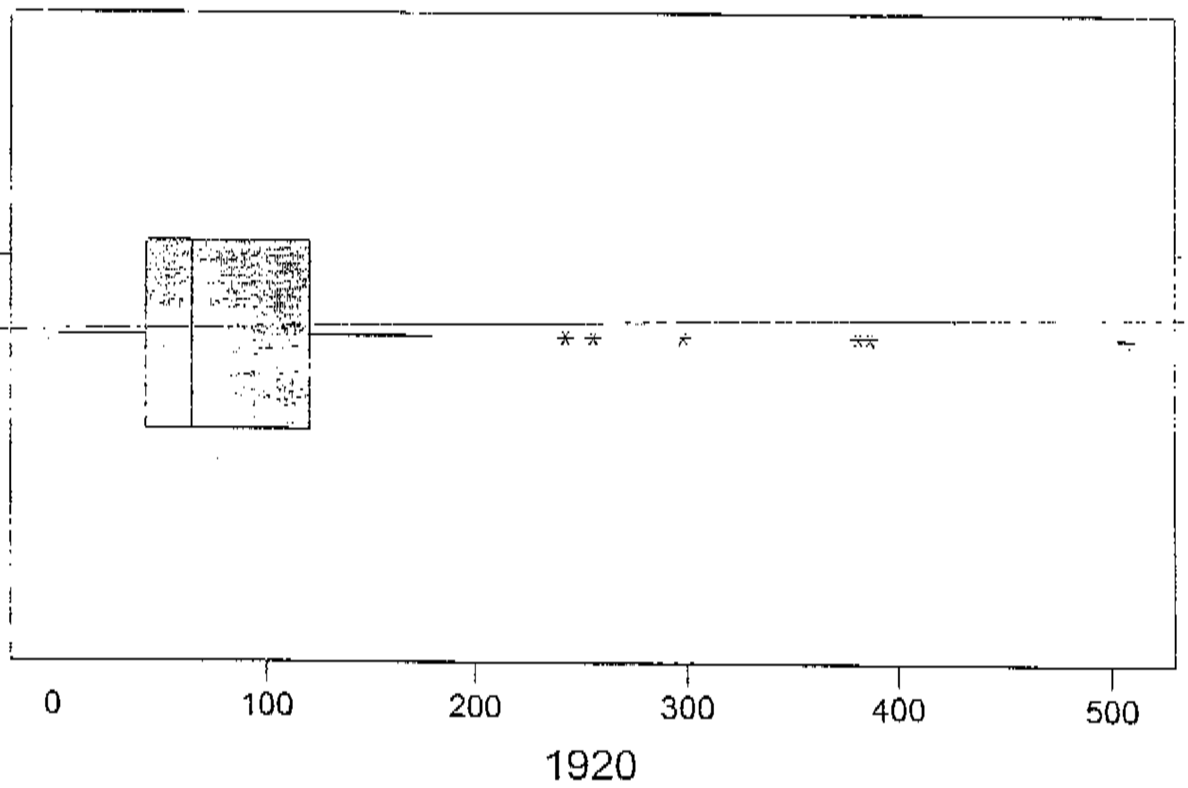
iv) distribution of $\bar{x}/\bar{u}$

[11/5/2001 13:49] E:\MastersExam\2001\MastersExamF2001\citesoutput.txt

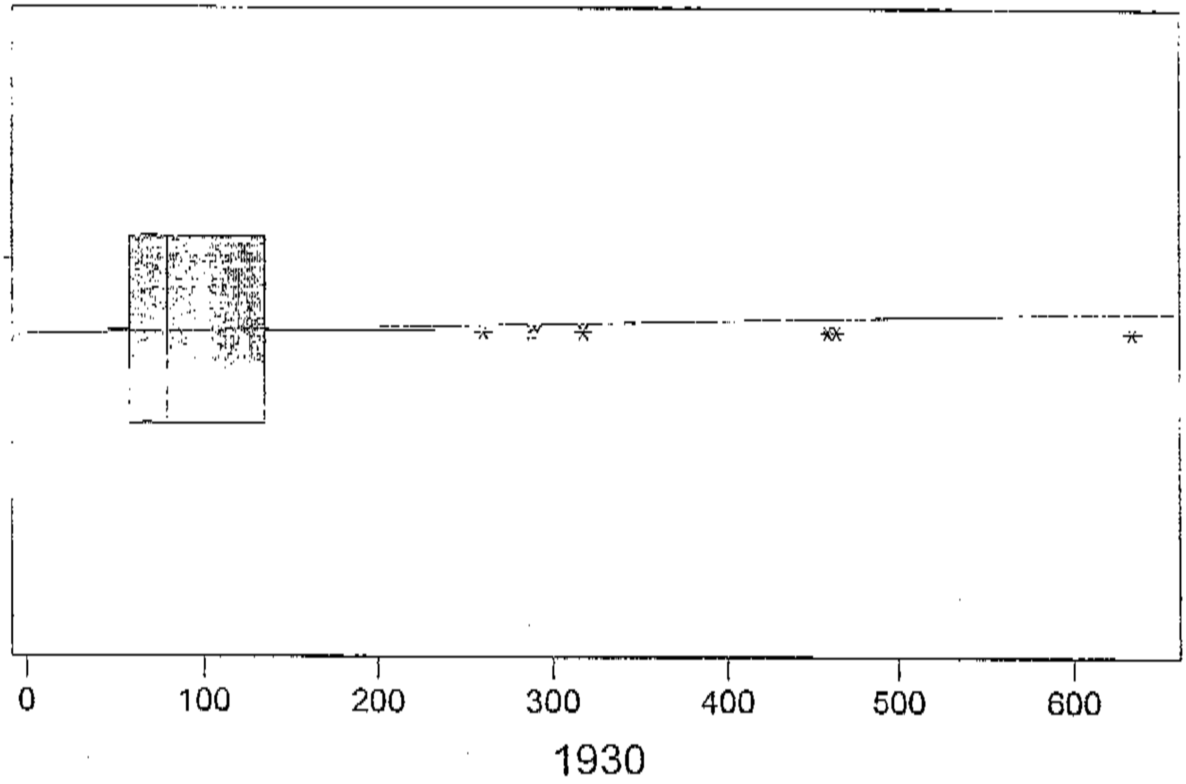
#2

```
a)

Descriptive Statistics: 1920, 1930, diff


Variable             N       Mean     Median     TrMean      StDev    SE Mean
1920                49      103.1       64.0       91.9      104.4       14.9
1930                49      127.8       79.0      112.7      123.1       17.6
diff                49      24.65      18.00      22.49      28.63       4.09

Variable       Minimum    Maximum         Q1         Q3
1920               2.0      507.0       43.0      120.5
1930              46.0      634.0       58.0      134.5
diff             -9.00     127.00       8.00      27.50

b)

Paired T-Test and CI: 1930, 1920

Paired T for 1930 - 1920

                  N      Mean     StDev   SE Mean
1930             49     127.8     123.1      17.6
1920             49     103.1     104.4      14.9
Difference       49     24.65     28.63      4.09

95% CI for mean difference: (16.43, 32.88)
T-Test of mean difference = 0 (vs not = 0): T-Value = 6.03  P-Value = 0.000
```

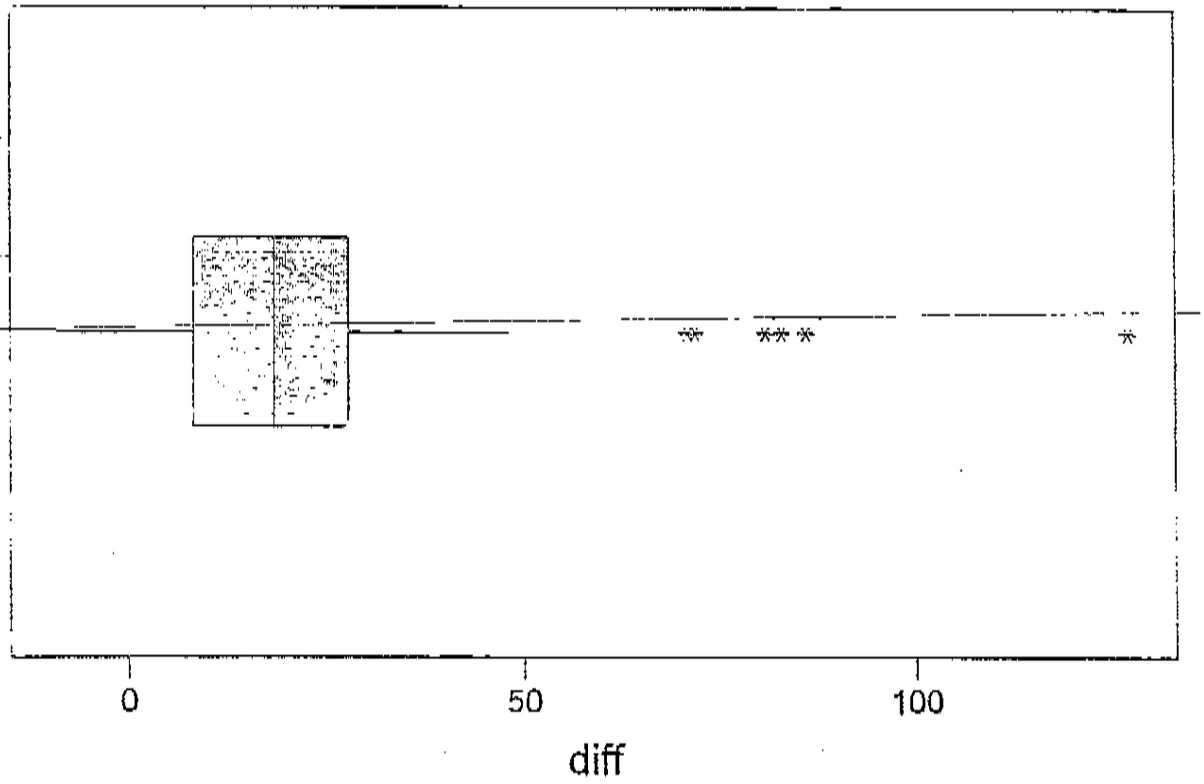
Boxplot of 1920



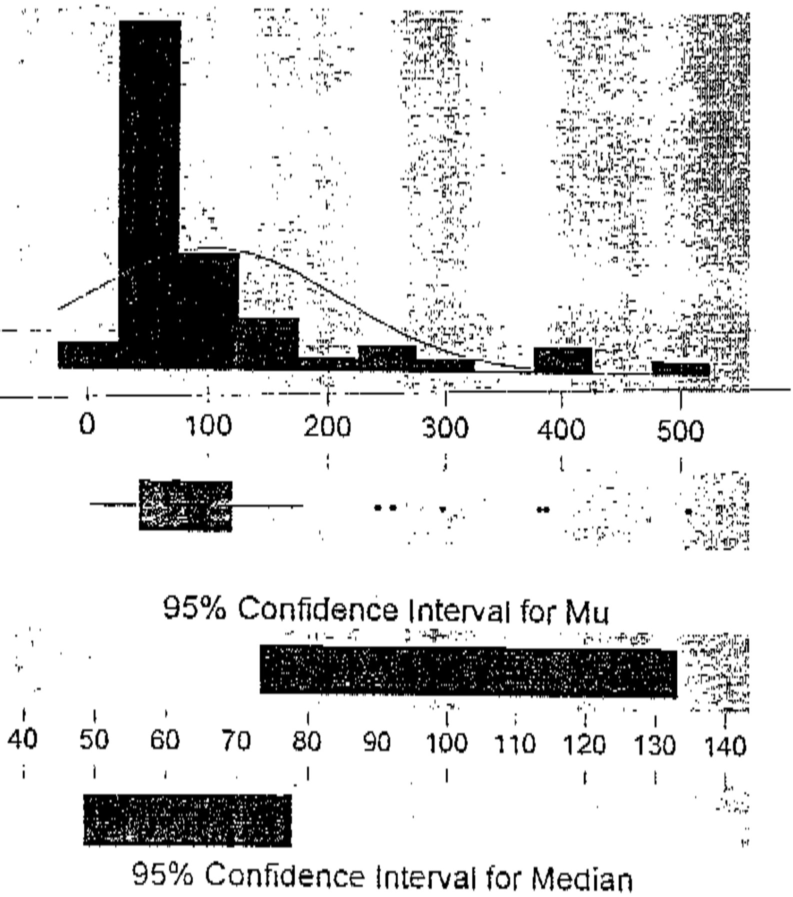
Boxplot of 1930



Boxplot of diff



Descriptive Statistics



Variable: 1920

Anderson-Darling Normality Test

A-Squared:	5.046
P-Value:	0.000
Mean	103.143
StDev	104.405
Variance	10900.4
Skewness	2.25528
Kurtosis	5.14124
N	49
Minimum	2.000
1st Quartile	43.000
Median	64.000
3rd Quartile	120.500
Maximum	507.000

95% Confidence Interval for Mu

73.154 133.131

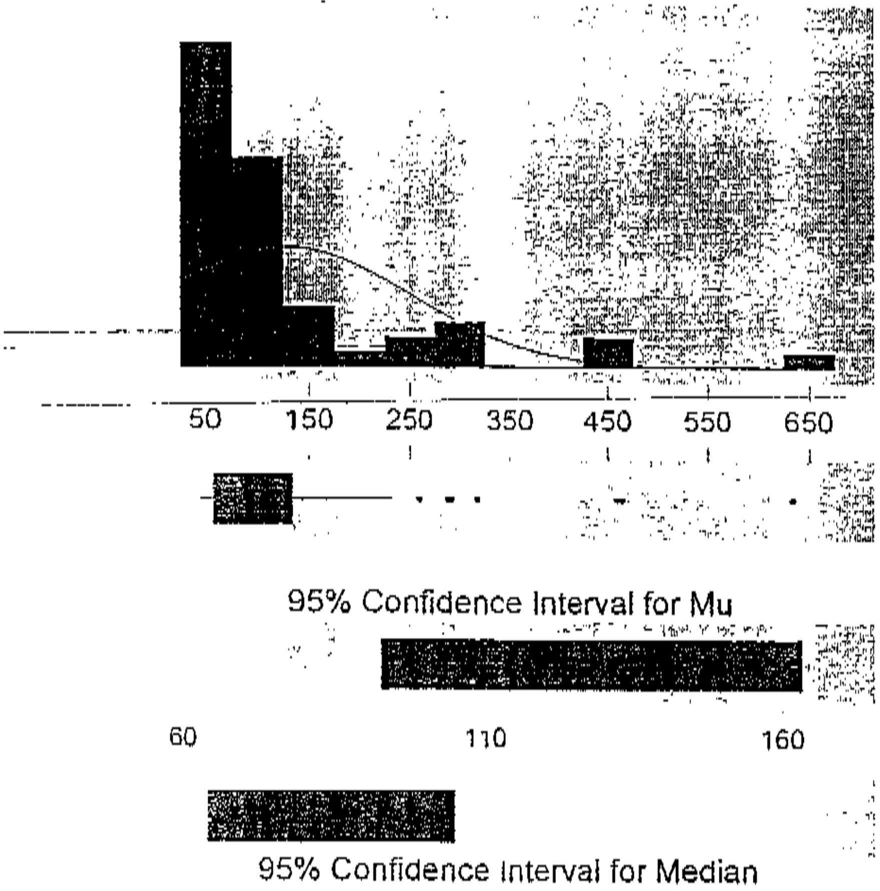
95% Confidence Interval for Sigma

87.066 130.433

95% Confidence Interval for Median

48.429 77.786

Descriptive Statistics



Variable: 1930

Anderson-Darling Normality Test

A-Squared:	6.211
P-Value:	0.000
Mean	127.796
StDev	123.121
Variance	15158.8
Skewness	2.49290
Kurtosis	6.47667
N	49
Minimum	46.000
1st Quartile	58.000
Median	79.000
3rd Quartile	134.500
Maximum	634.000

95% Confidence Interval for Mu

92.431 163.160

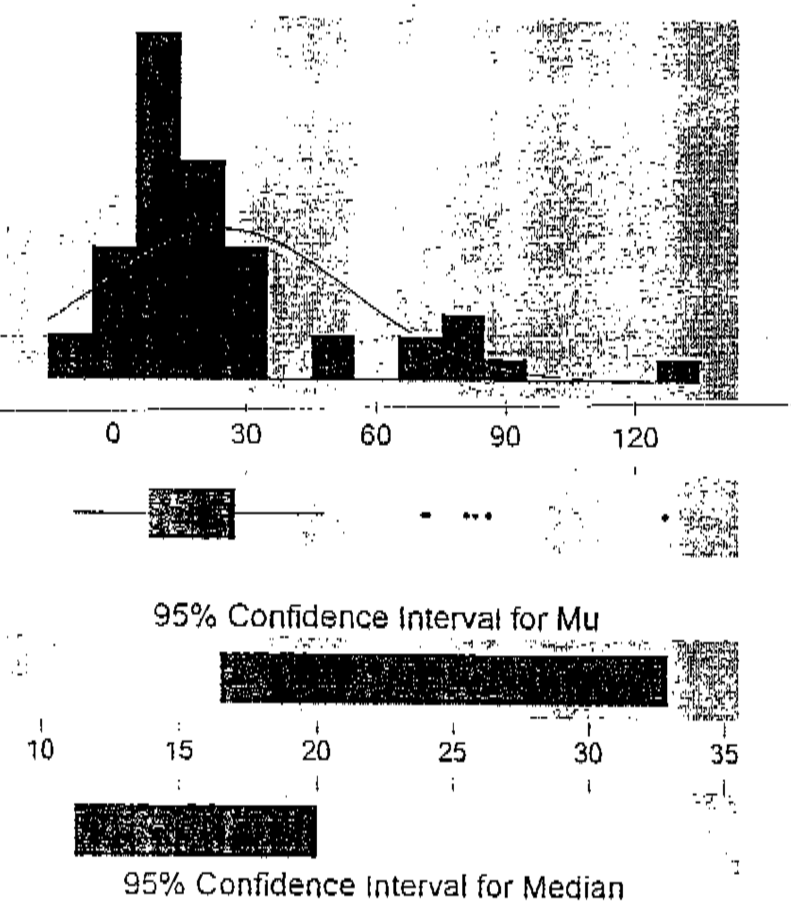
95% Confidence Interval for Sigma

102.673 153.815

95% Confidence Interval for Median

64.000 104.786

Descriptive Statistics

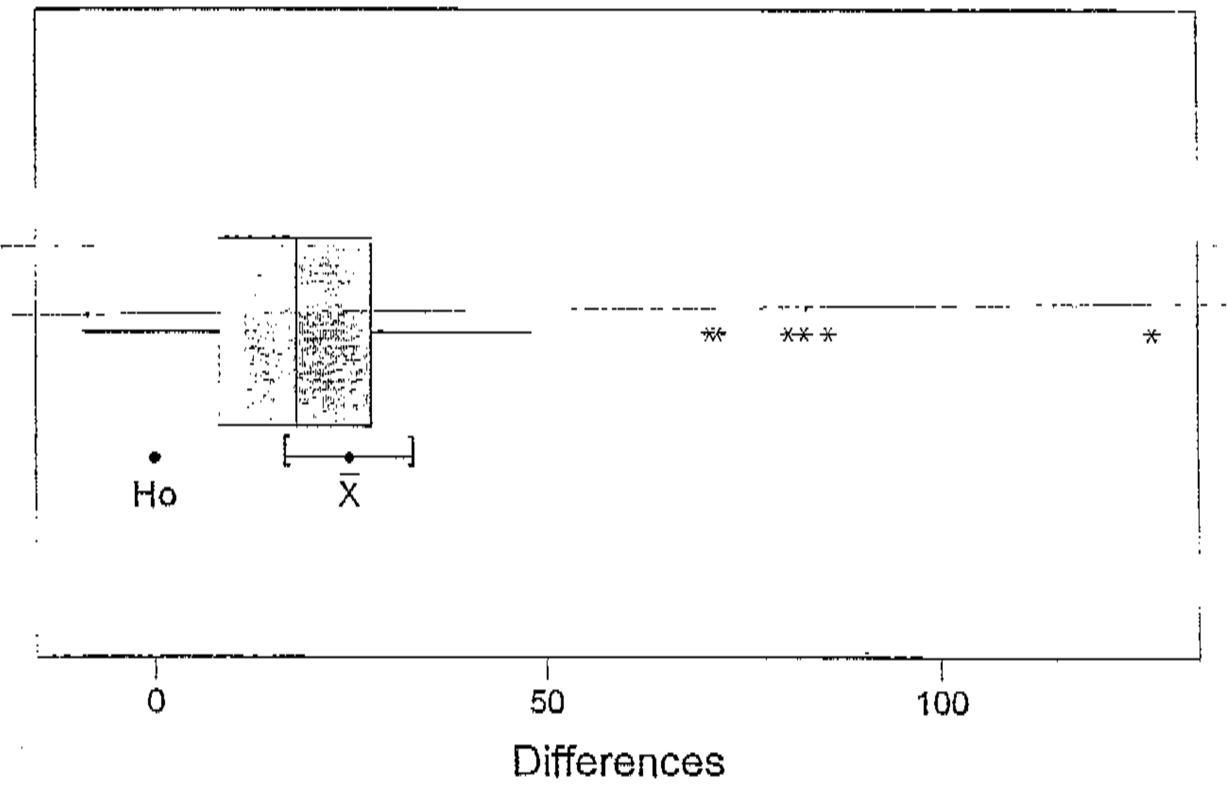


Variable: diff

```
Anderson-Darling Normality Test
    A-Squared:        3.665
    P-Value:          0.000

    Mean            24.6531
    StDev           28.6302
    Variance        819.690
    Skewness        1.72606
    Kurtosis        2.83716
    N                    49

    Minimum          -9.000
    1st Quartile      8.000
    Median           18.000
    3rd Quartile     27.500
    Maximum         127.000
95% Confidence Interval for Mu
    16.430           32.877
95% Confidence Interval for Sigma
    23.875           35.768
95% Confidence Interval for Median
    11.214           20.000
```

Boxplot of Differences

(with Ho and 95% t-confidence interval for the mean)



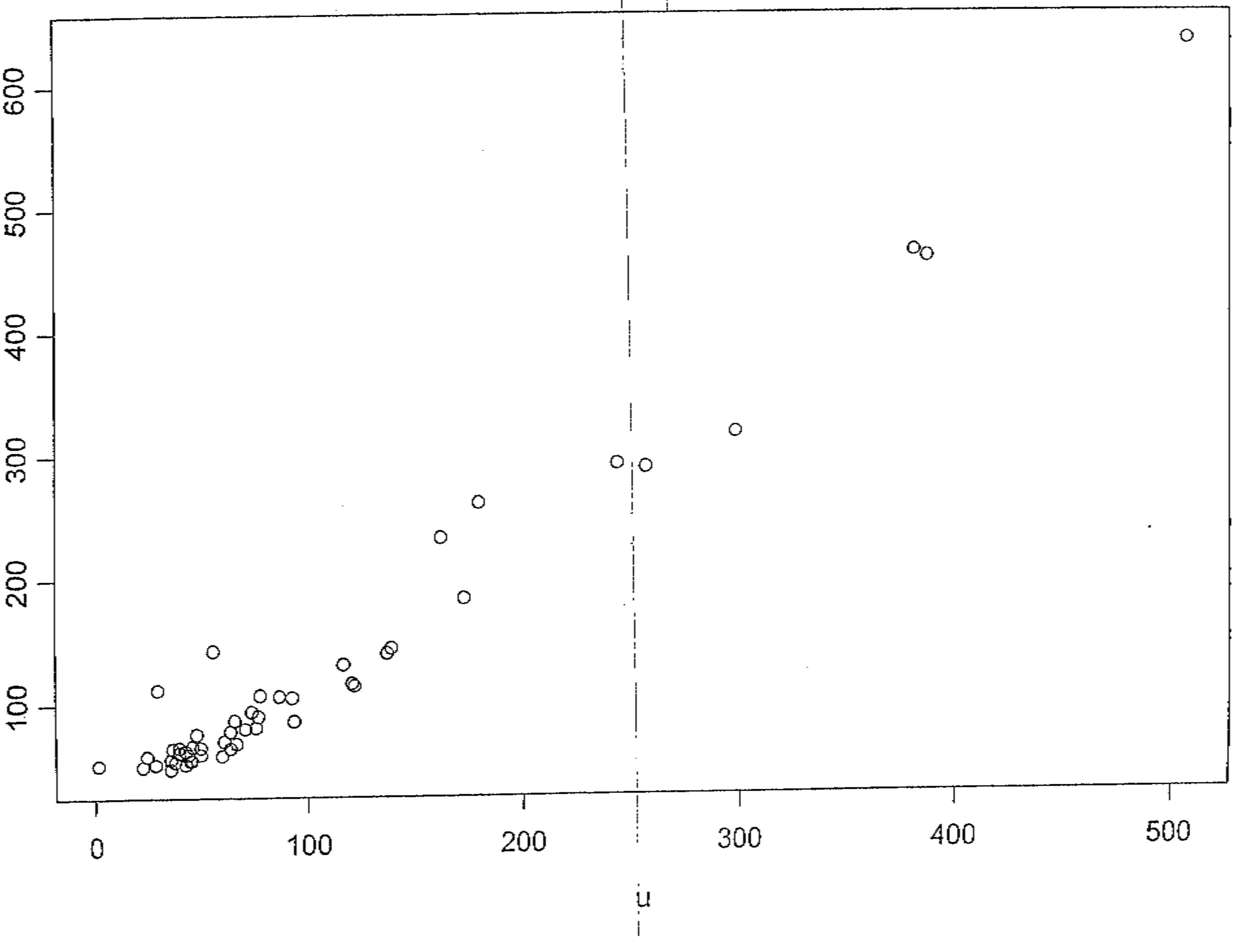
c)

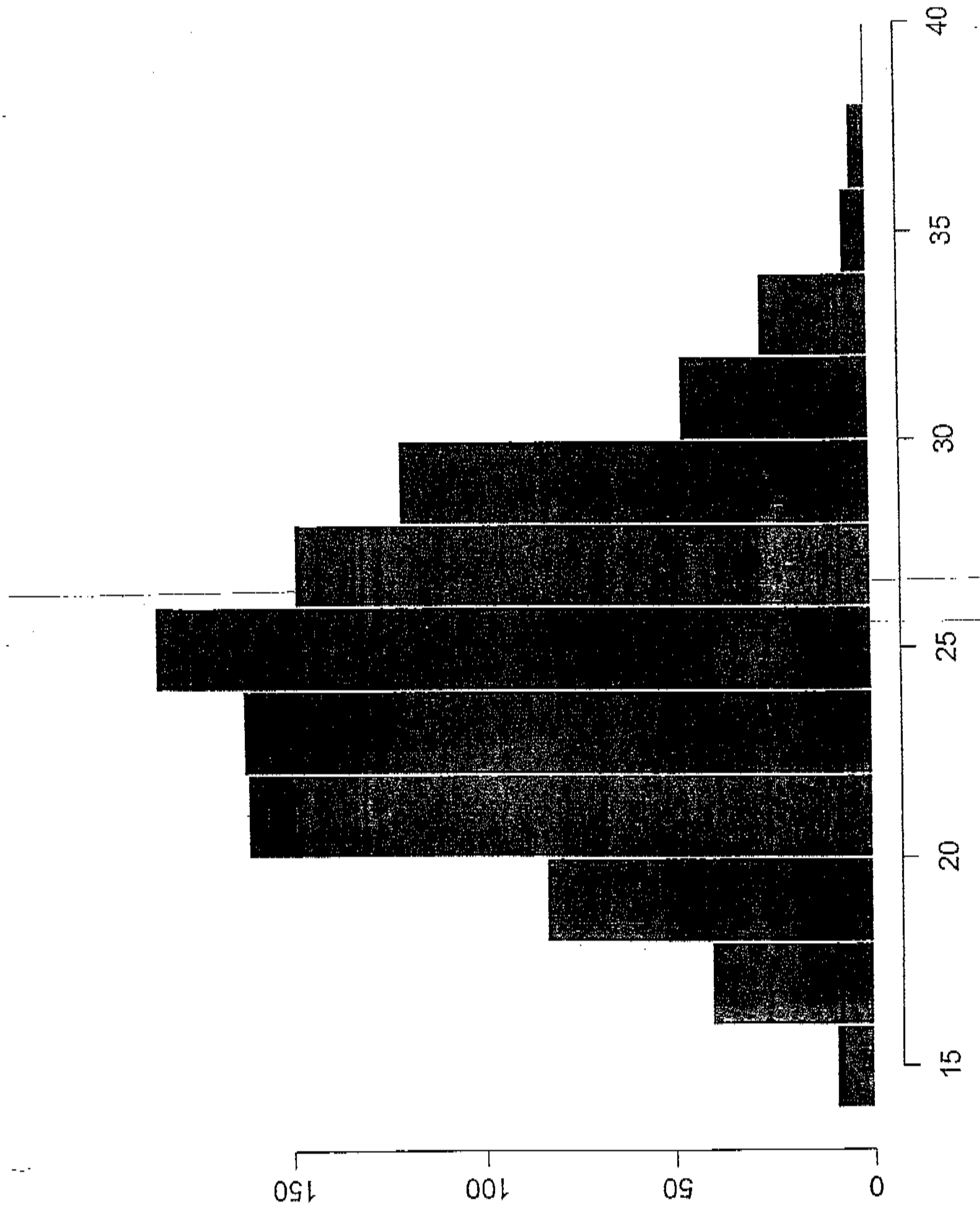
```
  ### Bootstrap the cities data.
cities <- matrix(c(138, 143, 93, 104, 61, 69, 179, 260, 48, 75, 37, 63, 29, 50, 23, 48, 30,
    111, 2, 50, 38, 52, 46, 53, 71, 79, 25, 57, 298, 317, 74, 93, 50, 58, 76, 80, 381,
    464, 387, 459, 78, 106, 60, 57, 507, 634, 50, 64, 77, 89, 64, 77, 40, 60, 136, 139,
    243, 291, 256, 288, 94, 85, 36, 46, 45, 53, 67, 67, 120, 115, 172, 183, 66, 86, 46,
    65, 121, 113, 44, 58, 64, 63, 56, 142, 40, 64, 116, 130, 87, 105, 43, 61, 43, 50,
    161, 232, 36, 54), byrow = T, ncol = 2)
> n <- 49
> u <- cities[, 1]
> x <- cities[, 2]  # scatter plot of
> plot(u, x) # mean difference
> d <- mean(u) - mean(x)   # ratio of means
> t <- mean(u)/mean(x) # Bootstrap difference and ratio
> B <- 1000  # number of bootstrap samples
> cities.boot <- matrix(NA, ncol = 2, nrow = n)   # storage matrix for bootstrap sample pairs
> diff.boot <- numeric(B)  # vector for bootstrap differences in the means
> ratio.boot <- numeric(B) # vector for bootstrap ratio in the means
> for(i in 1:B) {
    i.boot <- sample(seq(1:49), replace = T)
    # sample the index vector [1,2,...,49] with replacement
    for(j in 1:n) {
        cities.boot[j,  ] <- cities[i.boot[j],  ]
    # create the matrix of sampled pairs
    }
    diff.boot[i] <- mean(cities.boot[, 2]) - mean(cities.boot[, 1])
    # calculation of the bootstrap differences in the means
    ratio.boot[i] <- mean(cities.boot[, 2])/mean(cities.boot[, 1])
    # calculation of the bootstrap ratios in the means
}
```

```
# Bootstrap analysis of differences
  mean(diff.boot)   # bootstrap estimate of the difference in the means
[1] 24.57
> sqrt(var(diff.boot)) # bootstrap standard error of the difference in the means
[1] 4.087
> hist(diff.boot)   # bootstrap distribution of the difference in the means
> quantile(diff.boot, c(0.025, 0.975))
   # calculation of the empirical bootstrap confidence interval
  2.5% 97.5%
 17.22 32.94
```

d)

```
# Bootstrap analysis of ratios
> mean(ratio.boot)  # bootstrap estimate of the ratio in the means
[1] 1.24
> sqrt(var(ratio.boot))# bootstrap standard error of the ratio in the means
[1] 0.03514
> hist(ratio.boot)  # bootstrap distribution of the ratio in the means
> quantile(ratio.boot, c(0.025, 0.975))
   # calculation of the empirical bootstrap confidence interval
  2.5% 97.5%
 1.176  1.31
```


600
500
400
300
200
100
0
100
200
300
400
500
u


15
20
25
30
35
40
diff pot
0
50
100
150

3. The table below indicates the numbers of patients from a study of five hospitals for a surgical procedure designed to improve the functioning of certain joints impaired by disease. In this study the meanings to be attached to "no improvement." "partial functional restoration," and "complete functional restoration" were carefully defined in terms of measurable phenomena.

TABLE: Results of surgical procedure for patients at five area hospitals.

	HOSPITAL				
Results of surgical procedure	A	B	C	D	E
No improvement	13	5	8	21	43
Partial functional restoration	18	10	36	56	29
Complete functional restoration	16	16	35	51	10

a. What does the number 13 represent?

b. What model is appropriate for each cell?

c. Test the hypothesis that the distribution of results of surgical procedure is the same across hospitals.

d. Display estimates for all 15 cells of this table under the model given in part b. You may provide the information from your computer analysis without showing the computations.

e. Write a summary paragraph indicating the target population, the sample used in this study, a summary of the information obtained in your analysis, and a conclusion about surgical procedure. You may refer to the estimates from part d if useful.

OB
#3
JN

Solution 3.

TABLE 1: Results of surgical procedure for patients at five area hospitals.

			HOSPITAL		
Results of surgical procedure	**A**	**B**	**C**	**D**	**E**
No improvement	13	5	8	21	43
Partial functional restoration	18	10	36	56	29
Complete functional restoration	16	16	35	51	10

a. 13 is the number of patients undergoing a specific surgical procedure in hospital A who showed no improvement after the surgery. This number 13 can be converted to a point estimate for the joint probability that a randomly chosen patient will show no improvement and have had the surgery at area hospital A by dividing by the total of patients 367.

b. Every cell in column i and row j in the table above can be written as a frequency or converted to a joint probability. The model for the joint probability is phat(hospital i, results of surgery j)=phat (hospital i) * phat (surgery result j). The values for phat (hospital i) and phat (surgery result j) are the marginal probabilities for the two variables in the table. The model for the frequency in each cell is n*phat(for the cell). Sometimes these models are written as expected value of the cell = row total * column total/n. (A similar model for the cell probability would require that same result divided by n^2 rather than n.)

$$f_{ij} = f_{i.} * f_{j}/n \qquad \text{expected cell frequency}$$

c. We want to test whether hospitals and outcomes are independent or that proportions are equally distributed across hospitals. The chi-squared test is the best to use here. The following SAS program will generate the necessary test and the estimates (expected cell frequencies) requested in part d.

Code goes here and output.

d. Display estimates for all 15 cells of this table under the model given in part b. You may provide the information from your computer analysis without showing the computations.

TABLE 2: Expected result of surgical procedure for patients at five area hospitals.

			HOSPITAL		
Results of surgical procedure	**A**	**B**	**C**	**D**	**E**
No improvement	11.53	7.60	19.37	31.39	20.11

Partial functional restoration	19.08	12.59	32.07	51.97	33.29
Complete functional restoration	16.39	10.81	27.55	44.64	28.60

e. An study compared outcomes across five hospitals for a particular surgical procedure to restore patient functionality. The study classified patient outcomes by results of the procedure and hospital as shown in the data table 1 above. The target population was all patients with severe functionality loss who might benefit from this surgery. The sample is the 367 patients who took part in the study. Table 2 gives the expected or fitted frequency under the hypothesis of independence between hospitals and results or stated another way,.under the hypothesis that results are equally distributed across hospitals. When a chi-squared statistics is used to evaluate the hypothesis we find that the test statistic is large enough to reject the null hypothesis of independence. The chance that such a large test statistic (or a higher one) could have occurred by chance is less than 1 chance in 10,000 so we conclude that an equal distribution over hospitals is unlikely.

It is difficult to determine the cause of this uneven distribution without further study. It could be that the hospitals are of different quality or that the patients that go to the hospitals are in different states of health. It is even possible that the assignments of final improvement could have been handled unevenly. At a glance one might try to have this surgery at hospital C or D and avoid hospital E. But this response may be unfair as hospital E may be an excellent experimental hospital that takes many patients who are otherwise at too high a risk for this surgery. It may be the BEST choice for someone whose state of health is poor.

4. In all parts of this problem each "server" requires an exponentially distributed length of time, with mean 5 minutes, to complete his/her work (and thus to be ready to serve another customer if one is waiting to be served). Servers work independently.

[NOTE: The lettered parts below are independent and can be worked in any order. Part (a) is ***required;*** it counts 8 points. Work ***any two*** of parts (b), (c), and (d); 6 points each. When a numerical answer is required, you may use an analytical method to obtain the exact answer, a computer software function to give an answer correct to several places, or a computer simulation of sufficient scope to obtain a close approximation. If you use an analytical method, show the formula and computation briefly. If you use a computer procedure or simulation, say what software you use and describe the process briefly.]

(a) A bank has one waiting line that leads to any one of four tellers (servers). At the moment all four tellers are busy. John is at the head of the line and will be served as soon as any one of the four tellers finishes with his/her current customer.

> Find the distribution of the length of time John has to wait to begin service (to start being served by a teller). Show your derivation.
>
> State the mean and variance of this distribution. (No derivation required.)
>
> What is the probability that John will have to wait more than 2 minutes to begin service?
>
> What is the expected length of time before he *finishes* service?

(b) It is closing time at a bank. The bank has three tellers, and all of them are currently busy. There are no more customers waiting to be served. When each teller finishes with the customer he/she is currently serving, that teller will immediately go home.

> Find the mean and variance of the length of time until the last teller finishes and goes home.

(c) Mary arrives at a bank to make a complex transaction that requires the attention of three different tellers in sequence (each with different skills, but all with the same exponential service time distribution given at the top). None of the three tellers are busy and there are no customers other than Mary.

> Find the distribution of the length of time before Mary finishes with all three tellers. Show your derivation.
>
> State the mean, and variance of this distribution. (No derivation required.)
>
> What is the probability that it will take Mary more than 6 minutes to finish.

(d) A small bank has only one teller. If necessary, customers form a line to wait for the teller to serve them. Customers arrive according to a Poisson process at an average rate of one every 6 minutes.

> This is a well-known kind of queuing system. What is it called? M/M/1
>
> At steady state, what proportion of the time will the teller be idle? [Hint: Use the standard formula $P_n = (\lambda/\mu)^n\{1 - (\lambda/\mu)\}$, with appropriate values of n, λ, and μ.]
>
> At steady state, what is the expected number of people "in the system" (in line and being served)? [There is a standard formula for this; you may state and use it without derivation.] What would the answer be if customers arrived at the rate of one every 4 minutes?

Answers: #4 oB

(a) Let X be the service time of one teller; the cdf of X is $F(x) = 1 - e^{-x/5}$. John's wait W is the minimum of four independent random variables with this cdf.

$$P(W > w) = P(X_1 > w, X_2 > w, X_3 > w, X_4 > w) = [P(X > w)]^4 = (e^{-w/5})^4 = e^{-4w/5}.$$

Therefore, the cdf of W is $G(w) = 1 - e^{-4w/5}$, and W has an exponential distribution with mean 5/4 and variance $(5/4)^2 = 25/16$.

The probability that John waits more than 2 minutes for service is $P(W > 2) = e^{-8/5} = 0.2019$.

The expected time for John to finish service is $E(W + X) = 5/4 + 5 = 6.25$ minutes.

(b) Here the waiting time W is the maximum of three independent random variables X_i each with the cdf F of part (a). By an argument similar to that of (a) the waiting time T_1 for the first of these is exponentially distributed with mean 5/3, the additional waiting time T_2 for the second is exponentially distributed with mean 5/2, and the remaining waiting time T_3 for the third is exponentially distributed with mean 5.

Thus $E(W) = E(T_1 + T_2 + T_3) = 5/3 + 5/2 + 5 = (10 + 15 + 30)/6 = 55/6 = 9.167$ minutes. By the no-memory property of exponentials, the T_i are independent. This gives $V(W) = (5/3)^2 + (5/2)^2 + 5^2 = 34.028$.

Alternatively, $P(W \le w) = P(X_1 \le w, X_2 \le w, X_3 \le w) = [P(X \le w)]^3 = [F(w)]^3 = [1 - e^{-w/5}]^3$. Differentiating, we can obtain the pdf of W and hence find its mean and variance.

Here is a computer simulation that will find the mean and standard deviation with acceptable accuracy. (One run gave: $E(W) = \mu \approx 9.199$ and $SD(W) = \sigma \approx 5.8317$, so $V(W) = \sigma^2 \approx (5.8317)^2 = 34.009$. Runs longer than $m = 100{,}000$ will tend to give better accuracy.)

```
MTB > random 100000 c1-c3;
SUBC> expo 5.
MTB > rmax c1-c3 c4
MTB > desc c4

Descriptive Statistics: C4

Variable            N       Mean     Median     TrMean      StDev    SE Mean
C4             100000     9.1991     7.9249     8.7181     5.8317     0.0184

Variable      Minimum    Maximum         Q1         Q3
C4             0.1073    60.1512     5.0024    11.9837
```

The minimum (a) and sum (c) could be done using the subcommands `rmin` and `rsum`, respectively.

(c) Here the waiting time W is the sum of three independent X_i with the cdf F of part (a). The mgf of W is $m(t) = [(1-5t)^{-1}]^3 = (1-5t)^{-3}$, which is the mgf of a gamma random variable with shape parameter 3 and scale parameter 5. It should be clear that $E(W) = 3(5) = 15$ and $V(W) = 3(5)^2 = 75$. Minitab gives $P(W > 6) = 1 - 0.1205 = .8795$.

```
MTB > cdf 6;
SUBC> gamma 3 5.

Cumulative Distribution Function

Gamma with a = 3.00000 and b = 5.00000

         x      P( X <= x )
    6.0000         0.1205
```

(d) This is an M/M/1 queue.

The derivation was not requested, but here is a sketch. Letting $P_n = P(n \text{ in system})$, the balance equations are $\lambda P_0 = \mu P_1$, and $(\lambda + \mu)P_n = \lambda P_{n-1} + \mu P_{n+1}$, for $n = 1, 2, 3, \ldots$. Expressing P_n in terms of P_0, and then solving for P_n, we have $P_n = (\lambda/\mu)^n[1 - (\lambda/\mu)]$, for $n = 0, 1, 2, \ldots$, from which the formula for L follows. Because $\lambda = 1/6$ and $\mu = 1/5$, $P_0 = 1 - 5/6 = 1/6$

The formula for the average number in the system (being served and in the queue) at steady state is $L = \lambda/(\mu - \lambda)$, where $\lambda = 1/6$ and $\mu = 1/5$. Thus $L = (1/6)/(1/5 - 1/6) = 5$. Almost any book that even introduces queues talks about the M/M/1 queue and discusses these formulas.

5. A pharmaceutical company ran a preliminary study on the effects of a hormone supplement on growth rates. The first phase tests the hormone on experimental rats. From similar investigations it was believed that over the range of dosage the response would be approximately linear, and a straight line was routinely fitted to the data, yielding **fitted growth rate=86.44-0.20*supplement dosage**. Two rows of the data are shown below.

```
OBS    SUPPLMNT     GROWTHRT     LACKFIT
1      10           73           1
2      10           78           1
```

a. Write a SAS program to read in the data from the flat file growth rate. If you are unable to write the SAS program, copy the data and proceed with the rest of the problem using any computer package of your choice. 3 points

b. Verify the linear model given in the problem above and discuss parameter estimates and their significance. 3 points

c. Include the additional variable lackfit as a main effect to check the adequacy of the linear model (lack of fit). 3 points

d. What is the purpose of the variable lackfit? How does it work? 2 points

e. Use other error checking methods to determine what model might be better suited to the data without using the lack of fit variable. 3 points

f. Fit the model that you propose in part e above and assess the fit. 3 points

g. The table below gives an ANOVA table for some model. Write the model and discuss where each SS comes from and relate it to your model(s). 3 points

Analysis of Variance for growth rate data

Source	SS	DF	MS
Model	67428.6	2	33714.3
Mean	67404.1	1	67404.1
Linear	24.5	1	24.5
Residual	686.4	8	85.8
Lackfit	659.4	4	164.8
Error	27.0	4	6.8
Total	68115.0	10	

Solution 5. Open Book

a. Here is a version of the program that will work.

```
data growth;
infile 'I:/courswrk/stat/(folder path)/growth rate';
input obs supplmnt growthrt lackfit;
run;
```

b. To verify the linear model run a linear regression like the one below:

```
proc glm;
model growthrt=supplmnt;
```

The parameter estimates from this program match those given in the problem up to round-off error. Oddly the linear component is not significantly different from zero with a p-value of nearly 70% with 8 degrees of freedom.

c. Use the program below or a similar one in SPSS or MINITAB.

```
proc glm;
class lackfit;
model growthrt=supplmnt lackfit;
random lackfit;
```

```
Source                  DF        Type I SS     Mean Square    F Value    Pr > F
SUPPLMNT                 1       24.5016598      24.5016598       3.63    0.1295
LACKFIT                  4      659.3983402     164.8495851      24.42    0.0045
```

While supplement is still not significant, the lack of fit term is significant. This fact leads us to question the original model.

d. The term lack of fit uses the replicated supplement values to partition the error term farther to see if the error term can be attributed to a lack of fit in the original model. The values for lackfit reflect the 6 unique values of supplement in the data.

e. The best tool is a graph like the one below. I used the following SAS code to obtain this graph.

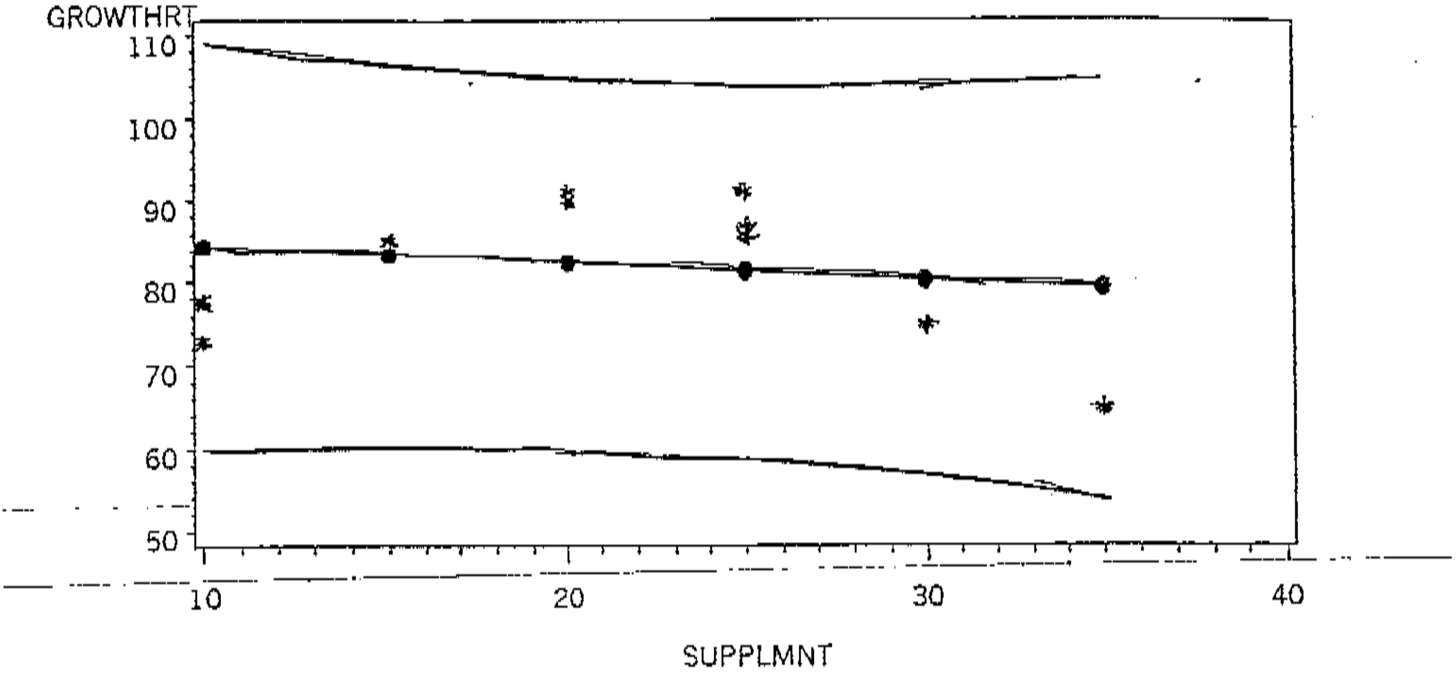
```
proc sort;
by supplmnt;
symbol1 v=star cv=blue;
symbol2 v=dot cv=black;
symbol3 i=join ci=blue;
symbol4 i=join ci=red;
proc gplot;
plot growthrt*supplmnt=1 growhat*supplmnt=2 growhat*supplmnt=3
     lcl*supplmnt=4 ucl*supplmnt=4/overlay;
title  'GRAPH: Growth Rate versus Supplement dosage';
title2 'Fits are black dots connected by blue lines.';
title3 'Upper and lower confidence limits are shown in red.';
footnote 'Observed growth rate appears as blue star.';
run;
```

```
proc glm;
class lackfit;
model growthrt=supplmnt supsq lackfit;
random lackfit;
run;
```

Note the curvature in the observed values. Replicated supplement dosages do not vary exceedingly in the graph. The curvature is causing the lack of fit here.

GRAPH: Growth Rate versus Supplement dosage

Fits are black dots connected by blue lines.
Upper and lower confidence limits are shown in red.



Observed growth rate appears as blue star.

A graph of residuals also shows the lack of fit for a linear model. You can use a program like the following:

```
data resids;
set resids;
zero=0;
proc gplot;
plot rstudent*growhat=1 zero*growhat=4/overlay;
title 'GRAPH: Studentized deleted residuals';
title2 'Studentized deleted residuals versus fitted values are shown as
stars.';
title3 'Reference line for residuals of zero is shown in red.';
run;
```

Notice that in this plot also the curvature is easily identified. We should try a supplement^2 term.

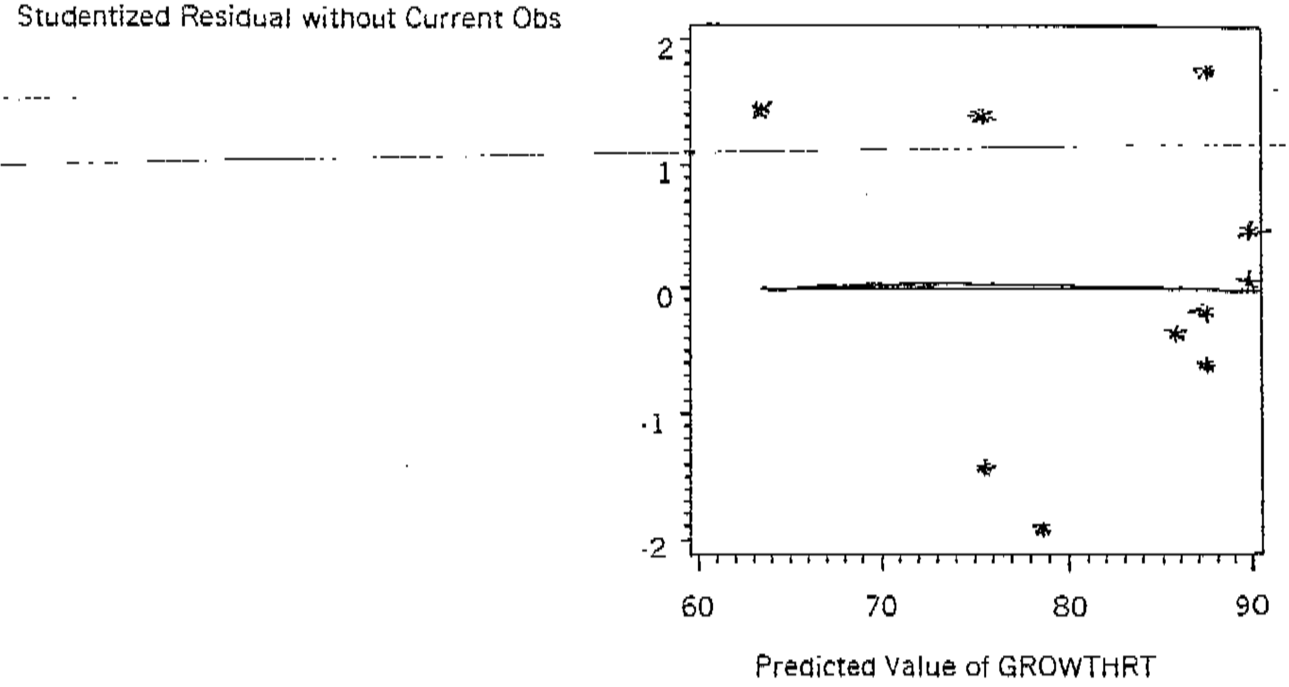
```
title2 'Studentized deleted residuals versus fitted values are shown as
stars.';
title3 'Reference line for residuals of zero is shown in red.';
run;
```

OR Lack of fit could be entered using a general linear model with supplement and supplement squared as covariates in the model. Now supplement and lack of fit under-cut each other since they both represent the linear trend. Neither is significant. Only supplement-squared is significant in this model.

```
proc glm;
class lackfit;
model growthrt=supplmnt supsq lackfit;
output out=residsl p=growhat r=resids lcl=lcl ucl=ucl rstudent=rstudent;
```

GRAPH: Studentized deleted residuals

Studentized deleted residuals versus fitted values when model is quadratic are shown as stars.
Reference line for residuals of zero is shown in red.



Observed growth rate appears as blue star.

g. The ANOVA table is the same as the one obtained in part c, however, it is not corrected for the mean and so the total SS and the model SS appears larger than that shown in the computer. SPSS does give this total SS. All of the remaining values for linear (supplement), residual, lack of fit and error are just like the ones in the computer output in part c.

```
The REG Procedure

                                           Model: MODEL1
                                   Dependent Variable: GROWTHRT

                                       Analysis of Variance

                                                       Sum of           Mean
                  Source                   DF         Squares         Square    F Value
Pr > F

                  Model                     2       665.70617      332.85309      51.56
<.0001
                  Error                     7        45.19383        6.45626
                  Corrected Total           9       710.90000


                           Root MSE              2.54092    R-Square     0.9364
                           Dependent Mean       82.10000    Adj R-Sq     0.9183
                           Coeff Var             3.09491


                                       Parameter Estimates

                                        Parameter       Standard
                     Variable     DF     Estimate          Error    t Value    Pr > |t|

                     Intercept     1     35.65744        5.61793       6.35      0.0004
                     SUPPLMNT      1      5.26290        0.55802       9.43      <.0001
                     supsq         1     -0.12767        0.01281      -9.97      <.0001
```

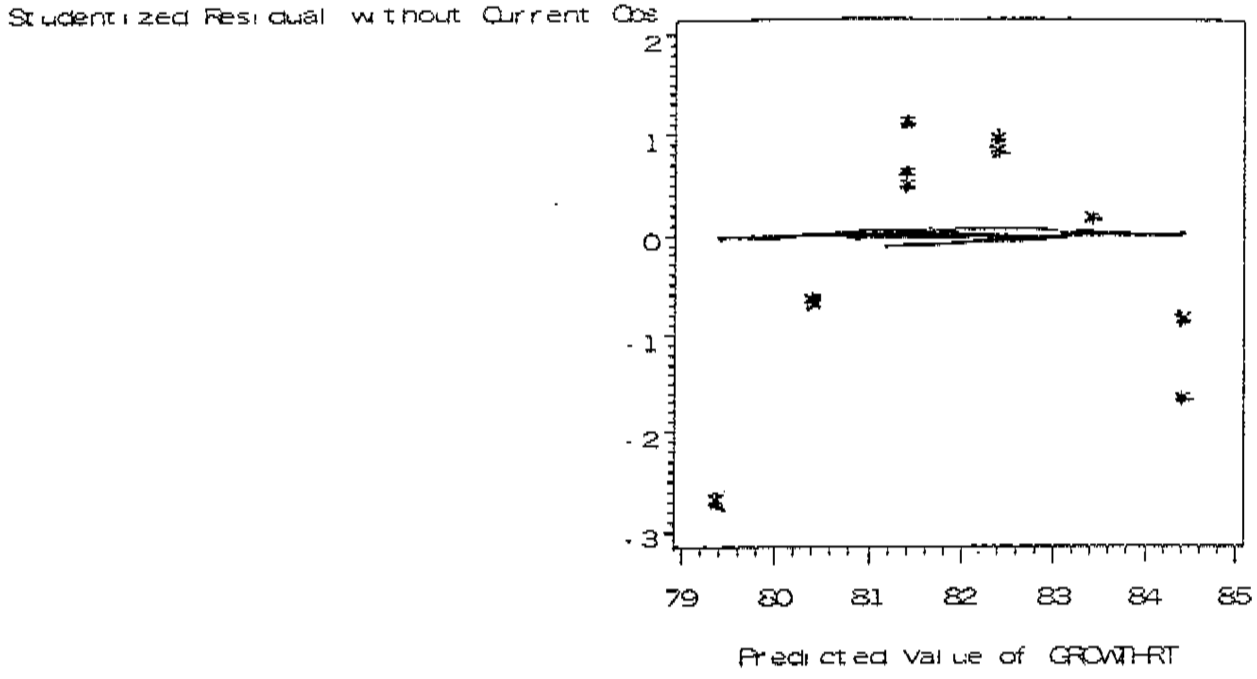
Note that both supplement and supplement-squared are not zero in this model. Also note in the graph of residuals versus the fit shown at the very end of the solution that the residuals show a random pattern.

Lack of fit could be entered in the regression as well, but dummy variables for the six values of lack of fit would have to be created and included as in the following SAS program. None of the terms in this model are significant.

```
data growth;
set growth;
supsq=supplmnt*supplmnt;
zero=0;
if lackfit=1 then lack1=1; else lack1=0;
if lackfit=2 then lack2=1; else lack2=0;
if lackfit=3 then lack3=1; else lack3=0;
if lackfit=4 then lack4=1; else lack4=0;
if lackfit=5 then lack5=1; else lack5=0;
*Note that there is no lack6 because when
       all lack1-lack5 are 0 then lackfit must be 6;
proc reg;
model growthrt=supplmnt supsq lack1-lack5;
output out=resids2 p=growhat r=resids lcl=lcl ucl=ucl rstudent=rstudent;
~~proc sort;~~
by supplmnt;
proc gplot;
plot rstudent*growhat=1 zero*growhat=4/overlay;
title 'GRAPH: Studentized deleted residuals';
```

GRAPH: Studentized deleted residuals

Studentized deleted residuals versus fitted values are shown as stars.
Reference line for residuals of zero is shown in red.



Observed growth rate appears as blue star.

f. Introducing the term supplement-squared into the model can be done with or without supplement in the model. Here is a SAS program that will generate the quadratic model and a final regression that plots the residuals from that model.

```
data growth;
set growth;
supsq=supplmnt*supplmnt;
zero=0;
proc reg;
model growthrt=supplmnt supsq;
output out=resids2 p=growhat r=resids lcl=lcl ucl=ucl rstudent=rstudent;
proc sort;
by supplmnt;
proc gplot;
plot rstudent*growhat=1 zero*growhat=4/overlay;
title 'GRAPH: Studentized deleted residuals';
title2 'Studentized deleted residuals versus fitted values are shown as
stars.';
title3 'Reference line for residuals of zero is shown in red.';
run;
```