STATISTICS DEPARTMENT
M.S. EXAMINATION

PART I
CLOSED BOOK

Friday, May 25, 2001

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

School of Education Conference Room, AE 123

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.

1. (a) Suppose that we have 2 different groups of patients with a certain disease, 101 patients in each group. In one group, they take diagnostic test #1 and 95 patients test positive for the disease; in the 2^{nd} group, they take diagnostic test #2 and 88 test positive for the disease. Assume that the total of 202 patients are randomly chosen from some population and are randomly assigned to the two groups. Can one conclude that the probability of a testing positive differs between the two diagnostic tests? Show your work.

   (b) Suppose in another scenario that 101 patients randomly chosen from some population took both diagnostic tests. 82 tested positive for both tests, 13 tested positive for diagnostic test #1 but not for #2, a (+, –) result, and 6 tested positive for #2 but not for #1, a (–, +) result.

      (i) Suppose we focus our attention on only those 19 patients where there was a lack of agreement between the two tests. Can one conclude that the probability of a (+, –) result (versus a (–, +) result) is not equal to .5? Show your work.

      (ii) Explain why a "YES" answer to (i) would allow one to establish that it is not the case that the probability of a positive result is the same for both diagnostic tests.

---

2. A professional gambler expects to have a 'winning day' 4 out of 5 days when he plays a certain gambling game. He wins $5000 on his winning days; on his losing days, he loses $10,000. Assume independence between different plays of the game.

(a) What is the probability that his 5^{th} loss occurs on the 15^{th} play of the game?

(b) Assuming that his 5^{th} loss occurs on the 15^{th} play, how much will his gain be after 15 plays?

(c) Given that the gambler has just suffered his 5^{th} loss, how many total games would we have expected him to play?

(d) Given that the gambler has just suffered his 5^{th} loss, what is his expected gain?

(e) Suppose X = number of plays that it takes to achieve the gambler's w^{th} loss. Explain why for w, a large integer, X is approximately normally distributed.

3. Sam installs TV cables. He is about to install a cable in an apartment and a cable in a house. Let X be the time that he requires to install a cable in the apartment and let Y be the time that he requires to install a cable in the house. Suppose that $0 < a < b < \infty$ and that X has a uniform distribution on the interval (a, b). Suppose that $0 < \tau < \infty$ and that Y has the probability density function f_Y defined by

$$f_Y(y) = \begin{cases} 2\tau^2 y^{-3} & \text{for } \tau < y < \infty \\ 0 & \text{elsewhere.} \end{cases}$$

5 (a) Find $P(Y > b)$.

5 (b) Find $E(Y - X)$.

(c) Suppose that X and Y are independent and suppose that $0 < \tau < a$.

5 (i) Find $E(\frac{1}{X})$.

5 (ii) Let $U = \frac{Y}{X}$. Find the probability density function of U.

4. Let $X_1, X_2, .., X_n$ be independent and identically distributed two-parameter exponential random variables with p.d.f.

$$f(x|\theta,\eta) = \frac{1}{\theta}e^{-(x-\eta)/\theta} \tag{1}$$

where $x > \eta$ and $\theta > 0$.

(a) Calculate the method of moments estimators of θ and η. (Hint: Recall the exponential and gamma densities when computing $E[X]$. Also recall that a location parameter does not change the variance of a random variable. Therefore, the variance of the two-parameter exponential is the same as the one-parameter exponential.)

(b) Calculate the maximum likelihood estimator, $\hat{\theta}$, of θ.

(c) Calculate the maximum likelihood estimator, $\hat{\eta}$, of η.

(d) Find the sufficient statistics for θ and η.

(e) Explain why the maximum likelihood estimators are functions of the sufficient statistics.

For the rest of the problem, assume that $\eta = 0$.

(f) Compute the Cramer-Rao Lower Bound for θ. Show that the MLE of θ, $\hat{\theta}$, achieves the CRLB.

(g) Give an asymptotic $100(1-\alpha)\%$ confidence interval for θ.

(h) Suppose the sample size is $n = 5$. Explain why the confidence interval given above may not be accurate in terms of its coverage probability. As an alternative method of calculating a confidence interval for θ, describe how you would implement the parametric bootstrap to produce the bootstrap estimate of θ, $\bar{\hat{\theta}}^*$, a bootstrap estimate of the standard error, $s_{\hat{\theta}}$, and a 95% Bootstrap confidence interval for θ.

5. Suppose that n is a positive integer and that $0 < p < 1$ Let N be a binomial random variable with parameters n and p. For each positive integer i, let X_i be a random variable whose probability density function is

$$f_X(x) = \begin{cases} \frac{1}{2}x^2 \exp(-x) & \text{if } x > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

Suppose that N, X_1, X_2, $\cdots$ are independent. Let

$$Y_k = \begin{cases} 1 & \text{if } k = 0 \\ \prod_{i=1}^{k} X_i & \text{if } k = 1, 2, \cdots \end{cases}$$

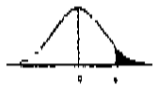
7 (a) Find $P(Y_1 > 2)$.

7 (b) Find $E(Y_N)$ in terms of $\mu = E(X_1)$, n, and p. You need not determine the value of μ.

6 (c) For $k = 1, 2, \cdots$ let $Z_k = \dfrac{Y_k}{E(Y_k)}$. Show that Z_k, $k = 1, 2, \cdots$ is a Martingale.

CUMULATIVE DISTRIBUTION OF CHI-SQUARE*

Degrees of Freedom	Probability of a Greater Value												
	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005
1	....	....	...	...	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.80	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.64	107.56	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17



Areas for the Normal Curve

02	.03	.04	.05	.06	.07	.08	.09
920	.4880	4840	.4801	.4761	.4721	4681	.4641
522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
129	.4090	.4052	4013	.3974	.3936	.3897	.3859
745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
676	.2643	2611	.2578	.2546	.2514	2483	.2451
358	.2327	2296	2266	.2236	.2206	.2177	.2148
061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
539	.1515	.1492	1469	.1446	.1423	.1401	.1379
314	.1292	1271	.1251	.1230	.1210	.1190	.1170
112	.1093	.1075	1056	.1038	.1020	.1003	.0985
)934	0918	.0901	0885	.0869	.0853	.0838	.0823
)778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
)643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
)526	.0516	.0505	0495	.0485	.0475	.0465	.0455
)427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
)	.0336	.0329	0322	.0314	.0307	.0301	.0294
)2	.0268	.0262	.0256	.0250	.0244	.0239	.0233
)217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
)170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
)132	0129	.0125	.0122	.0119	.0116	.0113	.0110
0102	.0099	.0096	.0094	0091	.0089	.0087	.0084
0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
0044	.0043	.0041	.0040	.0039	.0038	0037	.0036
0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

–
3
7
0
9

Percentage Points of the *t* Distribution

df	a = .1	a = .05	a = .025	a = .01	a =
1	3.078	6.314	12 706	31.821	63
2	1 886	2.920	4.303	6.965	9
3	1 638	2.353	3.182	4.541	5
4	1 533	2.132	2.776	3.747	4.
5	1.476	2.015	2.571	3.365	4.
6	1.440	1.943	2.447	3.143	3.
7	1.415	1.895	2.365	2.998	3.
8	1.397	1.860	2.306	2.896	3.
9	1.383	1.833	2.262	2.821	3.
10	1.372	1.812	2 228	2.764	3.
11	1.363	1.796	2.201	2.718	3
12	1.356	1.782	2.179	2.681	3.
13	1 350	1.771	2.160	2.650	3.
14	1.345	1.761	2.145	2.624	2.
15	1.341	1.753	2.131	2.602	2.
16	1.337	1.746	2.120	2.583	2.
17	1.333	1.740	2.110	2.567	2.
18	1.330	1.734	2 101	2.552	2.
19	1.328	1.729	2.093	2.539	2.
20	1.325	1.725	2.086	2.528	2.
21	1.323	1.721	2 080	2.518	2.
22	1.321	1.717	2 074	2.508	2.
23	1.319	1.714	2.069	2.500	2.
24	1.318	1.711	2.064	2.492	2.
25	1.316	1.708	2.060	2.485	2.
26	1.315	1.706	2.056	2.479	2.
27	1.314	1.703	2.052	2.473	2.
28	1.313	1.701	2.048	2.467	2.
29	1.311	1.699	2.045	2.462	2.
30	1.310	1.697	2.042	2.457	2.
40	1.303	1.684	2.021	2.423	2.7
60	1.296	1.671	2.000	2.390	2.6
120	1.289	1.658	1.980	2.358	2.6
240	1.285	1.651	1.970	2.342	2.5
[illegible]	1.282	1.645	1.960	2.326	2.5

closed book #1 Answer

(a) Suppose that we have 2 different groups of patients with a certain disease, 101 patients in each group. In one group, they take diagnostic test #1 and 95 patients test positive for the disease and in the 2nd group, they take diagnostic test #2 and 88 test positive for the disease. Assume that the total of 202 patients are randomly chosen from some population and are randomly assigned to the two groups. Can one conclude that it is not the case that the probability of a positive result is the same for both diagnostic tests? Show your work.

(b) Suppose in another scenario that 101 patients randomly chosen from some population took both diagnostic tests. 82 of them tested positive for both tests, 13 of them tested positive for diagnostic test #1 but not for #2, a (+, –), result, and 6 of them tested positive for #2 but not for #1, a (–, +) result.

(i) Supposing we focused our attention on only those 19 patients where there was a lack of agreement between the two tests. Can one conclude that the probability of a (+, –) result (versus a (–, +) result) is not equal to .5? Show your work.
(ii) Explain why if your answer to (i) is "YES", then you have established that it is not the case that the probability of a positive result is the same for both diagnostic tests.

```
>>> DA

!  Datafile: C:\backup\E Drive\MSSOLUT\msp2s012nd.cy3

!  Table   1 of   1
```

Table1	+	-	TOTAL
test #1	95 (91.5)	6 (9.5)	101
test #2	88 (91.5)	13 (9.5)	101
TOTAL	183	19	202

Pearson's

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 2\left\{\frac{(3.5)^2}{91.5} + \frac{(3.5)^2}{9.5}\right\} = 2.847$$

```
>>> RC CH/EX

Datafile: C:\backup\E Drive\MSSOLUT\msp2s012nd.cy3

PEARSON CHI-SQUARE TEST

Statistic based on the observed 2 by 2 table(x) :
 CH(X): Pearson Chi-Square Statistic =      2.847(      2.091 with Yates CC)

Asymptotic p-value: (based on Chi-Square distribution with 1 df )
   Two-sided : Pr { CH(X) .GE.        2.847 } =      0.0916
   One-sided : 0.5 * Two-sided                =      0.0458

Exact p-value and point probabilities :
   Two-sided : Pr { CH(X) .GE.        2.847 } =      0.1466
               Pr { CH(X) .EQ.        2.847 } =      0.0952
   One-sided : Let y be the value in Row 1 and Column 1
        y =95 min(Y) =82 max(Y) =101 mean(Y) =       91.50 std(Y) =       2.080

          Pr { Y .GE. 95 } =       0.0733
          Pr { Y .EQ. 95 } =       0.0476
```

p-value = .0916
Thus at $\alpha = .05$
one cannot conclude
that $P(+)$ is different
for the 2 tests.

```
>>> DA

!  Datafile: C:\backup\E Drive\MSSOLUT\msp2s01.cy3

!  Table   1 of   1
```

diagnostic test 2

Table1	+	-	TOTAL
diagnostic test 1 +	82	13	95
-	6	0	6
TOTAL	88	13	101

```
>>> PS MC/EX

Datafile: C:\backup\E Drive\MSSOLUT\msp2s01.cy3

MCNEMAR'S TEST
```

Statistic based on the observed 2 by 2 table(x) :

Min	Max	Mean	Std-dev	Observed	Standardized
-19.00	19.00	0.0000	4.359	7.000	1.606

Asymptotic Inference:
One-sided p-value: Pr { Test Statistic .GE. Observed } = 0.0541
Two-sided p-value: 2 * One-sided = 0.1083

Exact Inference:
One-sided p-value: Pr { Test Statistic .GE. Observed } = 0.0835
Pr { Test Statistic .EQ. Observed } = 0.0518
Two-sided p-value: 2*One-Sided = 0.1671

asymptotic

Newan $\chi^2 = \frac{(13-6)^2}{13+6} = \frac{49}{19} = (1.606)^2$; We can also use the chi-square goodness of fit test

p-value $= .1083$ (2-sided)

Cannot conclude at $\alpha = .05$

exact

$X = \#$ of $(+,-) = 13$

$n = 19$

p-value $= 2P(X \geq 13) = 2(.0835) = .1671$; Cannot conclude at $\alpha = .05$
(from binomial table, $n=19$, $X=13$)

ii) Suppose you conclude that $P((+,-)) > P((-,+))$

then $P(+, \text{test } 1) = P((+,+)) + P((+,-))$

$> P((+,+)) + P((-,+)) = P(+, \text{test } 2)$.

CSX I
Openbook #2 Answer

A professional gambler figures to have 'winning day' 4 out of 5 days that he plays a certain gambling game. He wins \$5000 on his winning days; otherwise, on his losing days, he loses \$10,000. Assume independence between different plays of the game.

(a) What is the probability that his 5th loss occurs on the 15th play of the game?
(b) Assuming that his 5th loss occurs on the 15th play, how much will be his gain after 15 plays?
(c) After the gambler suffers his 5th loss, how many total games would we expect to have been played by then?
(d) What is the gambler's expected gain after suffering his 5th loss?
(e) Suppose X = number of plays that it takes to achieve the gambler's w th loss. Explain why for w, a large integer, X is approximately normally distributed.

(a) X = # of games until (and including) the 5th loss is hypergeometric.

$$P(X=15) = \binom{14}{4}(.8)^{10}(.2)^5$$

(b) $10(5000) - 5(10000) = 0$ (even, steven)

~~(c)~~ $\cancel{\dfrac{5}{\left(\frac{1}{5}\right)} = 25 = E(X)}$

(d) $G = (X-5)\,5000 - 5(10{,}000)$

$E(G) = (20)\,5000 - 5(10{,}000) = 50{,}000$

(e) Let Y_1 = # of plays thru 1st loss

$i \geq 1$; Y_{i+1} = # of plays after i^{th} loss thru $(i+1)^{st}$ loss.

$Y_1, Y_2, \ldots$ are iid geometrics

$X = Y_1 + Y_2 + \cdots + Y_w$ is Normal by the central limit theorem.

Open book #3

3. Sam installs TV cables. He is about to install a cable in an apartment and a cable in a house. Let X be the time that he requires to install a cable in the apartment and let Y be the time that he requires to install a cable in the house. Suppose that $0 < a < b < \infty$ and that X has a uniform distribution on the interval (a,b). Suppose that $0 < \tau < \infty$ and that Y has the probability density function f_Y defined by

$$f_Y(y) = \begin{cases} 2\tau^2 y^{-3} & \text{for } \tau < y < \infty \\ 0 & \text{elsewhere.} \end{cases}$$

5 (a) Find $P(Y > b)$. $-\ P[Y>b] = \begin{cases} \int_b^\infty 2\tau^2 y^{-3}\,dy = \frac{2\tau^2 y^{-2}}{-2}\Big|_{y=b}^{y=\infty} = \frac{\tau^2}{b^2} & \text{if } b \geq \tau \\ 1 & \text{if } b < \tau \end{cases}$

5 (b) Find $E(Y-X)$. $E(Y-X) = E(Y) - E(X)$, $E(X) = \frac{a+b}{2}$, $E(Y) = \int_\tau^\infty y\, 2\tau^2 y^{-3}\,dy = \frac{2\tau^2 y^{-1}}{-1}\Big|_{y=\tau}^{\infty} = 2\tau$. Hence $E(Y)-E(X) = 2\tau - \frac{a+b}{2}$

(c) Suppose that X and Y are independent and suppose that $0 < \tau < a$.

~~4 (i) Find $P(Y>X)$.~~ $P[Y>X] = \int_a^b \frac{P[Y>x]\,dx}{b-a} = \frac{\int_a^b \frac{\tau^2}{x^2}dx}{b-a} = \frac{\tau^2\left(\frac{1}{-x}\right)\Big|_{x=a}^{x=b}}{b-a} = \frac{\tau^2\left(\frac{1}{a}-\frac{1}{b}\right)}{b-a} = \frac{\tau^2}{ab}$

5 (i) ~~(ii)~~ Find $E(\frac{Y}{X})$. $E\left(\frac{Y}{X}\right) = E(Y)\,E\left(\frac{1}{X}\right) = 2\tau \int_a^b \frac{1}{x}dx/(b-a) = \frac{2\tau}{b-a}\left(\ln(b)-\ln(a)\right)$

5 (ii) ~~(iii)~~ Let $U = \frac{Y}{X}$. ~~Find the probability density function of U~~ $P[U>1]$

~~$P[U>u] = P[Y>uX]$~~

$$= \frac{\int_a^b P[Y>ux]\,dx}{b-a}$$

note U always exceeds $\frac{\tau}{b}$:
$U = \frac{Y}{X} > \frac{Y}{b} > \frac{\tau}{b}$.

$$= \frac{\int_a^b \frac{\tau^2}{(ux)^2}dx}{b-a} = \frac{1}{u^2}\frac{\tau^2}{ab} \quad \text{if } ua \geq \tau$$

for $\frac{\tau}{b} \leq u < \frac{\tau}{a}$,

$$P[U>u] = \frac{\int_a^b P[Y>ux]\,dx}{b-a}$$

$$= \frac{\int_a^b I_{(-\infty,\tau]}(ux)\,dx}{b-a} + \frac{\int_a^b \frac{\tau^2}{(ux)^2} I_{(\tau,\infty)}(ux)\,dx}{b-a}$$

$$\frac{\int_a^{\frac{\tau}{u}} dx}{b-a} + \frac{\int_{\frac{\tau}{u}}^b \left(\frac{\tau}{u}\right)^2 \frac{1}{x^2}dx}{b-a} = \frac{\frac{\tau}{u} - a + \left(\frac{\tau}{u}\right)^2 \left(\frac{-1}{x}\Big|_{x=\frac{\tau}{u}}^{x=b}\right)}{b-a}$$

$\frac{\tau}{u} - a + \left(\frac{\tau}{u}\right)^2\left(\frac{u}{\tau} - \frac{1}{b}\right)$ $\quad \frac{2\tau}{u} - a - \left(\frac{\tau}{u}\right)^2\frac{1}{b}$ $\quad \therefore f_U(u) = \frac{2}{u^3}\frac{\tau^2}{ab} I_{[\frac{\tau}{a},\infty)}(u)$

$2\tau \quad 2\tau^2$

$u = \frac{x}{y}$, $v = y$

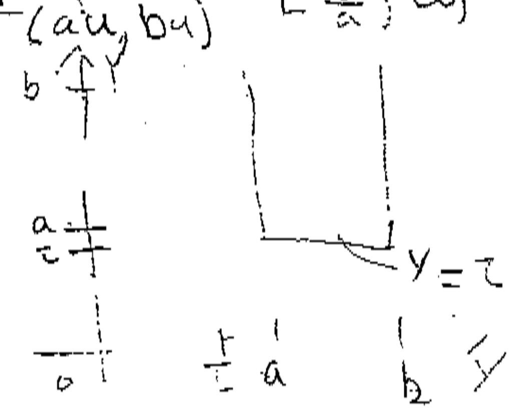
$x = \frac{v}{u}$, $y = v$

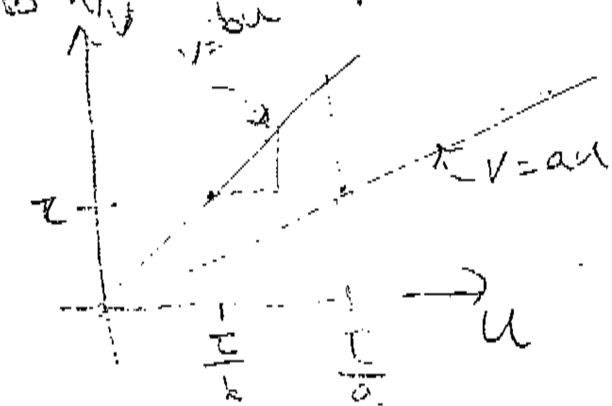
$$J = \begin{vmatrix} -\frac{v}{u^2} & \frac{1}{u} \\ 0 & 1 \end{vmatrix} = -\frac{v}{u^2}$$

$$f_{U,V}(u,v) = f_{X,Y}(x(u,v), y(u,v))\,|J|$$

$$= \frac{1}{b-a} I_{(a,b)}\left(\frac{v}{u}\right) 2\tau^2 v^{-3} I_{(\tau,\infty)}(v) \frac{v}{u^2}$$

$$\cancel{\frac{1}{u^2(b-a)} 2\tau^2 v^{-2} I_{(au,bu)}(v)\, I_{[\frac{\tau}{a},\infty)}(u) + \frac{1}{u^2(b-a)} 2\tau^2 v^{-2} I_{(\tau,bu)}(v)\, I_{(\frac{\tau}{b},\frac{\tau}{a})}(u)}$$





$x = a \Rightarrow \frac{v}{u} = a$, $v = au$

$$f_U(u) = \frac{2\tau^2}{u^2(b-a)}\left[\int_{au}^{bu} v^{-2}\,dv\, I_{[\frac{\tau}{a},\infty)} + \int_{\tau}^{bu} v^{-2}\,dv\, I_{(\frac{\tau}{b},\frac{\tau}{a})}(u)\right]$$

$$= \frac{2\tau^2}{u^2(b-a)}\left[-\frac{1}{v}\Big|_{v=au}^{v=bu} I_{[\frac{\tau}{a},\infty)}(u) + -\frac{1}{v}\Big|_{v=\tau}^{v=bu} I_{(\frac{\tau}{b},\frac{\tau}{a})}(u)\right] =$$

$$= \frac{2\tau^2}{u^2(b-a)}\left[\left(\frac{1}{au} - \frac{1}{bu}\right) I_{[\frac{\tau}{a},\infty)} + \left(\frac{1}{\tau} - \frac{1}{bu}\right) I_{(\frac{\tau}{b},\frac{\tau}{a})}(u)\right]$$

$$E(U) = \frac{2\tau^2}{ab}\int_{\frac{\tau}{a}}^{\infty} u^{-2}\,du + \frac{2\tau}{b-a}\int_{\frac{\tau}{b}}^{\frac{\tau}{a}} u^{-1} - \frac{\tau}{b}u^{-2}\,du$$

$$= \frac{2\tau^2}{ab}\left[-u^{-1}\Big|_{u=\frac{\tau}{a}}^{u=\infty}\right] + \frac{2\tau}{b-a}\left[\ln(u) + \frac{\tau}{b}u^{-1}\right]\Bigg|_{u=\frac{\tau}{b}}^{u=\frac{\tau}{a}}$$

$$= \frac{2\tau^2}{ab}\left[\frac{a}{\tau}\right] + \frac{2\tau}{b-a}\left[\ln\left(\frac{\tau}{a}\right) - \ln\left(\frac{\tau}{b}\right) + \frac{\tau}{b}\left(\frac{a}{\tau} - \frac{b}{\tau}\right)\right]$$

$$= \frac{2\tau}{b} + \frac{2\tau}{b-a}\left[\ln(b) - \ln(a)\right] - \frac{2\tau}{b}$$

— closed book #4

a) $E[X] = \int_{\eta}^{\infty} x \frac{1}{\theta} e^{-\frac{x-\eta}{\theta}} dx = \frac{1}{\theta}\int_{\eta}^{\infty} x\, e^{-\frac{x-\eta}{\theta}} dx$

$$= \frac{1}{\theta}\int_{\eta}^{\infty} (x-\eta+\eta)\, e^{-\frac{x-\eta}{\theta}} dx$$

$$= \frac{1}{\theta}\int_{\eta}^{\infty} (x-\eta)\, e^{-\frac{x-\eta}{\theta}} dx + \frac{\eta}{\theta}\int_{\eta}^{\infty} e^{-\frac{x-\eta}{\theta}} dx$$

let $y = x-\eta$, $dy = dx$ let $z = x-\eta$, $dz = dx$

$$= \frac{1}{\theta}\int_{0}^{\infty} y\, e^{-\frac{y}{\theta}} dy + \frac{\eta}{\theta}\int_{0}^{\infty} e^{-\frac{z}{\theta}} dz$$

$$= \theta\Gamma(2) \underbrace{\int_{0}^{\infty} \frac{1}{\theta^2\Gamma(2)} y^{2-1} e^{-\frac{y}{\theta}} dy}_{Gamma(\theta, 2)} + \eta \underbrace{\int_{0}^{\infty} \frac{1}{\theta} e^{-\frac{z}{\theta}} dz}_{Exp(\theta)}$$

$\Gamma(2) = 1$

$$= \theta + \eta$$

Using the hint.

$$E[X^2] = Var(X) + (E[X])^2 = \theta^2 + (\theta+\eta)^2$$

To find the MME's solve

$$E[X] = \theta + \eta = \frac{1}{n}\sum x_i$$

$$E[X^2] = \theta^2 + (\theta+\eta)^2 = \frac{1}{n}\sum x_i^2$$

① $\tilde{\eta} = \frac{1}{n}\sum x_i - \tilde{\theta} = \bar{x} - \tilde{\theta}.$

② $\theta^2 + (\theta+\eta)^2 = \frac{1}{n}\sum x_i^2$

$$2\theta^2 + 2\theta\eta + \eta^2 = \frac{1}{n}\sum x_i^2$$

$$2\theta^2 + 2\theta(\bar{x} - \theta) + (\bar{x}-\theta)^2 = \frac{1}{n}\sum x_i^2$$

$$\cancel{2\theta^2} + \cancel{2\theta\bar{x}} - \cancel{2\theta^2} + \bar{x}^2 - \cancel{2\bar{x}\theta} + \theta^2 = \frac{1}{n}\sum x_i^2$$

$$\theta^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

$$\theta^2 = \frac{\sum x_i^2 - \frac{(\sum x)^2}{n}}{n}$$

$$\tilde{\theta} = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n}}.$$

b) $L(\theta,\eta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{x_i-\eta}{\theta}}$ $\quad x_{(1)} \cdots x_n > \eta$

$= \theta^{-n} e^{-\frac{1}{\theta}\sum(x_i-\eta)}$ $\quad x_{(1)} > \eta.$

$\ell(\theta,\eta) = -n \log\theta - \frac{1}{\theta}\sum(x_i-\eta)$

$\frac{\partial}{\partial\theta}\ell(\theta,\eta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum(x_i-\eta) = 0$

$\frac{n}{\theta} = \frac{1}{\theta^2}\sum(x_i-\eta)$

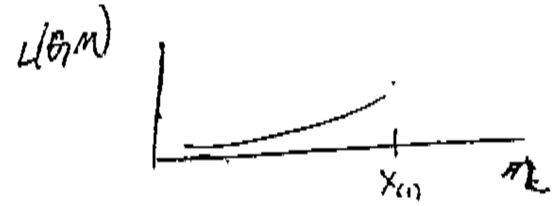
$\hat\theta = \frac{\sum(x_i-\hat\eta)}{n.}$ //

---

c) irregular estimation

$L(\theta,\eta) = \theta^{-n} e^{-\frac{1}{\theta}\sum x_i} e^{+\frac{\eta}{\theta}}$

$L(\theta,\eta) \propto e^{+\eta} \quad \eta < x_{(1)}$



$\hat\eta = x_{(1)}$

d) $f(x_1,\dots,x_n|\theta,\eta) = \theta^{-n} e^{-\frac{1}{\theta}\sum(x_i-\eta)} I_{(\eta,\infty)}(x_{(1)})$

$= \theta^{-n} e^{+\frac{\eta}{\theta}} e^{-\frac{\sum x_i}{\theta}} I_{(\eta,\infty)}(x_{(1)})$

$\therefore (\sum X_i, X_{(1)})$ are suff. for (θ, η)

e) From the factorization theorem, the likelihood is $g(T,\theta)\,h(\underset{\sim}{x})$, which depends on θ only through T. Therefore, to maximize this quantity we need only maximize $g(T,\theta)$.

f) $f(x|\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$, $\log f(x|\theta) = -\log\theta - \frac{x}{\theta}$

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

$\frac{d^2}{d\theta^2} \log f(x|\theta) = \frac{1}{\theta^2} - \frac{2x}{\theta^3}$

$I(\theta) = -E\left[\frac{d^2}{d\theta^2}\log f(x|\theta)\right] = -E\left[\frac{1}{\theta^2} - \frac{2x}{\theta^3}\right]$

$= -\frac{1}{\theta^2} + \frac{2}{\theta^3} E[X] = -\frac{1}{\theta^2} + \frac{2}{\theta^3}\theta = \frac{1}{\theta^2}$

$CRLB = \frac{1}{nI(\theta)} = \frac{\theta^2}{n}$.

For the $Exp(\theta)$, the MLE is $\hat{\theta} = \bar{X}$.

Since $Var(\hat{\theta}) = Var(\bar{X}) = \frac{\theta^2}{n}$, $\hat{\theta}$ achieves

the CRLB

g) $\hat{\theta} \pm z_{\alpha/2} \frac{1}{\sqrt{nI(\hat{\theta})}} \Rightarrow \hat{\theta} \pm z_{\alpha/2} \frac{\hat{\theta}}{\sqrt{n}}$ where $\hat{\theta} = \bar{X}$

e) The asymptotic CI computed in g may not be accurate since it requires a large sample. The coverage probability may be low since the CI would be too narrow as a result of the small sample size.

To implement the parametric bootstrap:

① Using the MLE $\hat{\theta} = \bar{X}$ generate B samples from $Exp(\hat{\theta})$. ② Compute ~~$\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ as the sample mean of~~ each bootstrap sample. ③ The bootstrap estimate of θ is $\bar{\hat{\theta}}^* = \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i^*$ ④ The bootstrap estimate of $S_{\hat{\theta}}$ is $S_{\hat{\theta}} = \sqrt{\frac{\Sigma(\hat{\theta}_i^* - \bar{\hat{\theta}}^*)^2}{B-1}}$

⑤ An approximate 95% bootstrap CI can be obtained from the 0.025 and 0.975 percentiles of the $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

closed book #3

5 Suppose that n is a positive integer and that $0 < p < 1$. Let N be a binomial random variable with parameters n and p. For each positive integer i, let X_i be a random variable whose probability density function is

$$f_X(x) = \begin{cases} \frac{1}{2}x^2 \exp(-x) & \text{if } x > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

Suppose that N, X_1, X_2, $\cdots$ are independent. Let

$$Y_k = \begin{cases} 1 & \text{if } k = 0 \\ \prod_{i=1}^{k} X_i & \text{if } k = 1, 2, \cdots \end{cases}$$

(a) Find $P(Y_1 > 2)$. $P[Y_1 \geq 2] = P[X_1 \geq 2] = \int_2^\infty \frac{1}{2}x^2 e^{-x}\,dx = -\frac{1}{2}x^2 e^{-x}\Big|_2^\infty + \int_2^\infty x e^{-x}\,dx$
$= 2e^{-2} - xe^{-x}\Big|_2^\infty + \int_2^\infty e^{-x}\,dx = 2e^{-2} + 2e^{-2} - e^{-x}\Big|_2^\infty = 5e^{-2}$ or $P[Y > 2] = P[N(2) < 3] = e^{-2} + 2e^{-2} + \frac{2^2 e^{-2}}{2!}$

~~(b) Find $E(Y_N \mid N = 10)$ in terms of $\mu = E(X_1)$. You need not determine the value of μ.~~
~~$E(Y_N \mid N = 10) = E\left(\prod_{i=1}^{10} X_i\right) = \prod_{i=1}^{10} E(X_i) = \prod_{i=1}^{10} \mu = \mu^{10}$~~

(b) Find $E(Y_N)$ in terms of $\mu = E(X_1)$, n, and p. You need not determine the value of μ. $E(Y_N) = \sum_{k=0}^{n} E(Y_k \mid N = k)\binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^{n} \mu^k \binom{n}{k} p^k (1-p)^{n-k} = (1 - p + \mu p)^n$

and $\mu = 3$

(c) For $k = 1, 2, \cdots$ let $Z_k = \frac{Y_k}{E(Y_k)}$. Show that Z_k, $k = 1, 2, \cdots$ is a Martingale.

$$E(Z_{k+1} \mid Z_1 \cdots Z_k) = E\left(Z_k \frac{X_{k+1}}{E(X_{k+1})} \,\Big|\, Z_1 \cdots Z_k\right)$$
$$= Z_k\, E\left(\frac{X_{k+1}}{E(X_{k+1})} \,\Big|\, Z_1, \ldots, Z_k\right)$$
$$= Z_k\, E\left(\frac{X_{k+1}}{E(X_{k+1})}\right)$$
$$= Z_k$$

STATISTICS DEPARTMENT
M.S. EXAMINATION

PART II
OPEN BOOK

Tuesday, May 29, 2001

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

Statistics Department Computer Lab, SC S152

Instructions: Complete problems *one, two,* and *three* and choose between problems *four* and *five*. Each problem counts 25 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 company wants to improve the lifetimes of a kind of delicate electronic device it manufactures. Engineers have proposed two new variations on the manufacturing process Method B and Method C, hoping that one of them will be better than the present Method A. In one investigation of the methods, 15 devices made by each method were subjected to stress tests, and times to failure (in days) were recorded for each of the 45 devices tested. The data are shown below, and are available in the file LIFETEST.TXT.

	A	B	C
	0.72	0.41	5.60
	2.63	1.54	1.87
	3.95	0.39	2.99
	0.02	1.25	0.54
	0.04	1.40	0.53
	0.14	0.74	1.42
	1.86	1.21	1.64
	0.07	0.16	2.77
	0.25	1.36	0.36
	0.99	1.33	1.95
	1.38	0.44	1.54
	0.44	0.04	0.95
	0.67	0.97	0.62
	2.12	0.37	1.39
	1.80	1.22	2.42
Mean	1.139	0.855	1.773

(a) For now, suppose that the usual assumptions for an ANOVA are satisfied. What is the name of the appropriate ANOVA design for this experiment? State the assumptions explicitly.

(b) Perform the ANOVA of part (a). Show the ANOVA table. Do the methods differ significantly at the 5% level? 10% level?

(c) It would be considered a major improvement if time to failure in the stress test could be increased by 1 day. How many observations on each of 3 groups would be required (instead of $n = 15$ here) to achieve a power of 80% for the ANOVA F-test. Use the 5% level of significance and assume that the population standard deviations are $\sigma = 1.1$ days.

(d) Are the assumptions of the ANOVA model met? Give the name(s) and P-value(s) of the test(s) you use.

(e) Past experience with similar test data has shown that it is sometimes appropriate to do a log or rank transform of the data. (That is either to replace each observed time with its logarithm, say base 10; or to replace each of the 45 observations with its rank, a number from 1 through 45.) Try both transformations and comment whether either is beneficial.

(f) Recall that the purpose of this experiment is to determine whether either of the two new methods (B or C) gives significantly longer times to failure than current Method A. Based on your results in (a) - (e), or the results of any other statistical procedure(s) you deem appropriate, what is your answer to that question? State clearly what data and procedures lead to your conclusion.

(g) Suppose that the company has four sites where an experiment such as the one described here can be performed (3 methods, 15 observations for each method). Suppose that the same three methods are tested at each site (giving a total of $4 \times 3 \times 15 = 180$ observations). Write the ANOVA model for this larger experiment.

2. Professor Leo Kahane, Department of Economics, collected the wage data partially shown below. Use the complete data set to answer questions a-g.

Wage	**Education**	**Female**	**Married**	**Experience**	**Black/Hispanic**
11	8	male	married	42	other
20.4	17	male	unmarried	3	other
9.1	13	male	unmarried	16	black/hispanic
13.75	14	male	married	21	other
24.98	16	male	married	18	other
7.7	13	female	unmarried	8	other
10	10	female	married	25	other
...	...	...	...	...	...
6.25	9	male	unmarried	30	other

Wage is 1991 adjusted thousands of dollars.
Education is years of schooling completed.
Experience is years of labor market experience.

a. Write a SAS program to successfully read in the data from the text file named wages. Even if your program fails, paste the data into a data step and CONTINUE!
b. Fit a model predicting wages based solely on years of education.
c. Test whether education is a significant predictor of wages at the 5% level.
d. Find a 95% confidence interval for the rate at which wages increase due to an increase of 1 year of education.
e. Include years experience in the prediction model. Are education and years experience both significant predictors of wages?
f. Create dummy variables for the other variables and find the "best" prediction model for these data that predicts wages. Cite your criterion for "best".
g. Based on your model, what can you conclude about the importance of each of the variables in determining wages.

3. Consider a Markov chain $X_0, X_1, X_2, \ldots$ with state space $S = \{1, 2, 3, 4, 5, 6\}$, initial probability vector $\pi_0 = (1, 0, 0, 0, 0, 0)$, and transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 2/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 3/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 4/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 0 & 5/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{bmatrix}.$$

(a) Evaluate the probabilities $\mathrm{P}\{X_2 = 1\}$ and $\mathrm{P}\{X_{n+2} = 3 \mid X_n = 3\}$.
[*Save time:* Do only the *necessary* arithmetic.]

(b) What does it mean for a chain to be irreducible? Is this chain irreducible?
What does it mean for a state to be persistent (recurrent)? Transient?
For each state in this chain, say whether it is persistent (recurrent) or transient.
What does it mean for a chain to be ergodic? Is this chain ergodic?

(c) Carefully explain how to verify, using hand computations and a theorem about Markov chains, that the vector $\lambda = (1/30, 1/20, 1/12, 1/6, 1/2, 1/6)$ gives the *limiting* distribution of this chain.
Explain briefly how you could approximate λ by finding higher and higher powers of $\mathbf{P}$ on a computer.

(d) Use the information in part (c) to evaluate the following limits (as $n \to \infty$):
$\lim \mathrm{P}\{X_n = 5 \mid X_0 = 3\}$ and $\lim \mathrm{P}\{X_n = 3\}$.
Also, evaluate μ_1, the mean number of stages after the chain leaves state 1 until it next returns to state 1.

4. An experiment was done to measure the effects of ozone, a component of smog. A group of twenty-two 70-day-old rats was kept in an environmental containing ozone for 7 days, and their weight gains (in grams) were recorded. Another group of twenty-three rats of similar age was kept in an ozone-free environment for a similar time, and their weight gains were recorded. Using the Splus program available at

   http://www.telecom.csuhayward.edu/~esuess/MEsp01ob.ssc

   analyze the data to determine the effect of ozone.

   (a) Describe the shapes of the histograms for the ozone and control groups.

   (b) What are the sample means for each group? What are the sample standard deviations?

   (c) Perform an independent two-sample t-test to see if there is a difference in the weight gains between the ozone group and control group. Before performing the test, check the assumptions of equal variance and normality.

   (d) What is the 95% independent two-sample t-confidence interval for $\mu_{oz} - \mu_{con}$?

   (e) Use the nonparametric bootstrap to estimate the difference in the means, to calculate a standard error estimate of the difference in the two means, and to calculate a 95% empirical bootstrap confidence interval.

   (f) Comment on the difference you observe between the 95% confidence intervals produced in parts (d) and (e).

   (g) Does ozone influence the rats' weight gains? If so, give an estimate of how much.

5. Table A: gives the birth weights (in pounds) of Poland China pigs in eight litters. The data come from *Statistical Methods* by George Snedecor.

a. Construct the analysis of variance table. Test at the 0.10 level the hypothesis of no difference between mean weights in the eight litters. (See file pig.txt.)
b. Suppose litters 1,3,4 were sired by one boar and the other five litters were sired by another boar. Is there significant difference between the mean weights in those two groups? Assume this is a planned comparison.
c. Is there a difference between the mean weight for large litters (numbers 1,2,3,4) and small litters (numbers 5,6,7,8)? Assume this is a comparison suggested by the data.

Table A

1	2	3	4	5	6	7	8
2.0	3.5	3.3	3.2	2.6	3.1	2.6	2.5
2.8	2.8	3.6	3.3	2.6	2.9	2.2	2.4
3.3	3.2	2.6	3.2	2.9	3.1	2.2	3.0
3.2	3.5	3.1	2.9	2.0	2.5	2.5	1.5
4.4	2.3	3.2	3.3	2.0		1.2	
3.6	2.4	3.3	2.5	2.1		1.2	
1.9	2.0	2.9	2.6				
3.3	1.6	3.4	2.8				
2.8		3.2					
1.1		3.2					

Overall Grading: Total points will probably not be divisible by the number of parts, which is 7. To avoid fractional points, 1 more point each may be given to some earlier parts than to some later parts.

(a) This is a one-way ANOVA (completely randomized design) with three levels of the fixed factor. Assumptions are that the data are normal and that the variances of the three groups are equal. Although it is not requested here, the formal model could be written as

$$Y_{ij} = \mu + \alpha_i + e_{ij};\ i = 1, 2, 3 \text{ (methods)}, ; j = 1, 2, \ldots, 15 \text{ (replications)};$$

where $\Sigma\, \alpha_i = 0$ and the e_{ij} are independent, identically distributed as $N(0, \sigma^2)$.

Grading: No points for anything but the one-way model. Substantial penalty for not mentioning independence, normality, and equal variances. Some penalty for saying it's a random effects model, but none for omitting to say the one effect is fixed. (If wrong model, try to give partial credit below for consistent adherence to whatever model is incorrectly stated.)

(b) In Minitab, we stacked the data into a column 'Life' with subscripts in 'Method'. The ANOVA tables follows, along with the Fisher LSD comparisons, which are useful in later parts. We also stored residuals for later use. Results are significant at the 10% level, but not the 5% level.

Grading: Heavy penalties for incorrect DF, MS, or F. Heavy penalty if interpretation is not consistent with P-value

```
Analysis of Variance for Life
Source     DF        SS        MS        F        P
Method      2      6.62      3.31     2.95    0.063
Error      42     47.05      1.12
Total      44     53.67
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level       N      Mean     StDev  -----+---------+---------+---------+-
1          15     1.139     1.146       (---------*--------)
2          15     0.855     0.511  (---------*--------)
3          15     1.773     1.336                 (---------*---------)
                                   -----+---------+---------+---------+-
Pooled StDev =    1.058                0.60      1.20      1.80      2.40

Fisher's pairwise comparisons
    Family error rate = 0.120
Individual error rate = 0.0500

Critical value = 2.018

Intervals for (column level mean) - (row level mean)

                 1           2

     2      -0.497
             1.063

     3      -1.414      -1.697
             0.146      -0.137
```

(c) Minitab's power/ sample size procedure shows that we would need 25 observations per method, to have 80% power.

Grading: Heavy penalties for using incorrect α, σ or power. Because comparison of A with B and of A with C is the crux, it would be OK to use the sample size formula, with correct parameters, for two-sample t-tests.

```
ie-way ANOVA

Sigma = 1.1   Alpha = 0.05   Number of Levels = 3

           Sample  Target  Actual  Maximum
SS Means     Size   Power   Power  Difference
     0.5       25  0.8000  0.8116  1
```

(d) A normal probability plot of the residuals shows a distinct curvature characteristic of nonnormal data. The P-value of the Anderson-Darling test is 1.1%. Bartlett's test for homoscedasticity performed on the original data (by Method) gives P-value 0.4%. Neither assumption seems to be met.

Grading: Roughly half credit each for some valid test of normality and some valid test of equal variances. Heavy penalty for doing Anderson-Darling, or other normality test, on the original data instead of the residuals. Partial credit on normality for rough description "curvature" in normal probability plot. (Bartlett would usually be done on the data, but for this model results should be the same if done on the residuals; no penalty.) There is really no way to check whether groups are independent. Control charts, runs tests, etc. might be used to check for independence within each group. But this is not expected; bonus points against errors in this part if discussed or correctly performed.

(e) With log-transformed data, the P-value is 5.6% for the F-test in the ANOVA table; 2% for Anderson-Darling on residuals, and 1.7% for Bartlett on the log-transformed data (by Method). With rank-transformed data, the P-value is 4.7% for the F-test, 11.3% for Anderson-Darling on residuals, and 19.5% for Bartlett on ranks (by Method). The rank-transformed data meet the ANOVA assumptions better and give a slightly smaller P-value for the F-test.

Grading: Somewhat more than 1/2 credit for doing *either one* of the transformations correctly all the way through.

(f) Fisher's LSD comparisons seem appropriate for comparing B with A and C with A. (Other comparison methods OK.) Only in the case of the log-transformed data does LSD find a significant difference between A and one of the new methods (specifically, between A and C). *There is evidence to suggest, but by no means to confirm, that Method C may be superior to Method A.* As seen in part (c) we do not really have enough data here reliably to detect differences of a size deemed important. The words "delicate" and "stressed" in the description of the problem were intended to give a clue that these devices may be failing because of "accidents" rather than by "wearing out," and hence may have exponentially distributed life times. If so, the log transformation may be defensible on theoretical grounds. (Bear in mind, however, that differences in means of logged lifetimes correspond to ratios different from 1 of geometric means of actual lifetimes.)

Grading: On purpose, sort of an open-ended and vague question. Credit strongly based on reasonable conclusions based on specific work done above. Penalty for making strong claims based on P-values around 5%, or making a strong distinction between 5% and 6% levels.

(g) This is a standard two-factor ANOVA (both factors fixed, crossed) with three levels of Method and four of Site. Interaction is supported. Model is

$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ij}$; $i = 1, 2, 3$ (methods); $j = 1, 2, 3, 4$ (sites); $k = 1, \ldots, 15$ (replications).
Restrictions: $\Sigma_i\, \alpha_i = \Sigma_j\, \beta_j = \Sigma_i\, (\alpha\beta)_{ij} = \Sigma_j\, (\alpha\beta)_{ij} = 0$. Distribution: e_{ij} i.i.d. $N(0, \sigma^2)$

Grading: No points if two factors not mentioned. Heavy penalty for not giving the model, since it was asked for. Heavy penalty for turning replications into a third effect, or for inventing bogus nesting. Some penalty if ranges of subscripts not given, interaction omitted, any factors declared random, restrictions omitted, or distributional information missing.

Source: Data were randomly generated as follows: All three methods exponential, A and B with mean 1 and C with mean 2. Modeled after a real situation with proprietary data.

Open book # 2

Solution to the regression problem:

a. SAS program. Many versions will work.

```
data new;
 infile 'i:\courswrk\stat\bcrumbo\exam questions spring 2001\reg.txt'
delimiter='09'x MISSOVER;
input wage educ gender $ marital $ exper blackhis $;
proc print;run;
```

b. The SAS System 21:55 Tuesday, April 24, 2001 4

The REG Procedure
Model: MODEL1
Dependent Variable: wage

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	368.23327	368.23327	15.19	0.0003
Error	43	1042.28064	24.23908		
Corrected Total	44	1410.51391			

Root MSE	4.92332	R-Square	0.2611
Dependent Mean	8.65444	Adj R-Sq	0.2439
Coeff Var	56.88777		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-5.12727	3.61126	-1.42	0.1629
educ	1	1.08423	0.27817	3.90	0.0003

Notice that the F ratio is highly significant. The equation is
wage = -5.127 + 1.08423 * education.

c. Since the above model is significant, education must be a significant predictor of wage. P=.0003

d. Using the estimate of the slope and its standard error from the print out, at 95% ci for the coefficient of education in the equation is 1.08±1.96*.278 or 1.08 ± .54.

The SAS System 21:55 Tuesday, April 24, 2001 5

The REG Procedure
Model: MODEL1
Dependent Variable: wage

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	432.04007	216.02004	9.27	0.0005

Error	42	978.47384	23.29700
Corrected Total	44	1410.51391	

Root MSE	4.82670	R-Square	0.3063
Dependent Mean	8.65444	Adj R-Sq	0.2733
Coeff Var	55.77130		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-9.40101	4.38214	-2.15	0.0378
educ	1	1.26301	0.29333	4.31	<.0001
exper	1	0.11009	0.06652	1.65	0.1054

Although education remains a significant predictor, experience is not significant since p=.1054 is >.05.

f.

The REG Procedure
Model: MODEL1
Dependent Variable: wage

C(p) Selection Method

Number in Model	C(p)	R-Square	Variables in Model
4	4.2049	0.4025	educ female blkhis exper
3	4.4751	0.3679	educ female blkhis
2	4.9743	0.3298	educ blkhis
3	5.1474	0.3577	educ female exper
3	5.3201	0.3550	educ blkhis exper
5	6.0000	0.4056	educ female married blkhis exper
4	6.1234	0.3733	educ female married blkhis
2	6.5170	0.3063	educ exper
4	6.6818	0.3647	educ female married exper
3	6.8570	0.3316	educ married blkhis
2	6.9094	0.3003	educ female
4	7.2806	0.3556	educ married blkhis exper
1	7.4852	0.2611	educ
3	8.0581	0.3133	educ female married
3	8.3332	0.3091	educ married exper
2	9.0350	0.2679	educ married
2	16.2444	0.1580	female blkhis
3	17.7226	0.1660	female married blkhis
3	18.2442	0.1581	female blkhis exper
1	18.3738	0.0951	blkhis
4	19.7179	0.1661	female married blkhis exper
2	20.2079	0.0976	married blkhis
2	20.2496	0.0970	blkhis exper
1	20.2583	0.0664	female
2	21.0168	0.0853	female married
3	22.0528	0.1000	married blkhis exper
2	22.0849	0.0690	female exper
3	22.9234	0.0867	female married exper
1	23.9768	0.0097	married

1	24.6088	0.0001	exper
2	25.9764	0.0097	married exper

Using the CP criterion several models could be useful, but the first in this list is a model containing education, a black Hispanic indicator, a female indicator, and work experience.

The SAS System 21:55 Tuesday, April 24, 2001 9

The REG Procedure
Model: MODEL1
Dependent Variable: wage

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	567.72968	141.93242	6.74	0.0003
Error	40	842.78423	21.06961		
Corrected Total	44	1410.51391			

Root MSE	4.59016	R-Square	0.4025
Dependent Mean	8.65444	Adj R-Sq	0.3427
Coeff Var	53.03823		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-5.79887	4.41281	-1.31	0.1963
educ	1	1.14697	0.28353	4.05	0.0002
female	1	-2.48446	1.39358	-1.78	0.0822
blkhis	1	-3.05441	1.76283	-1.73	0.0909
exper	1	0.09957	0.06543	1.52	0.1359

While education is still the only significant factor in this model, the CP statistics suggests that this is the best model to use.

g. If we rely on the model above rather than traditional significance, we note that being female reduces ones wages by 2.48 thousand dollars, and being black or Hispanic reduces you wages by just over $3000, while work experience increases salary about $90 for each year worked. But education increases salary over $1000 for each year of school completed up to grade 12 which is all that is included in this data.

We have NOT looked at assessment factors and we have not tried to see how well the model does on a hold out sample.

openbook # 3

Answers:

(a) $P\{X_2 = 1\} = P\{X_2 = 1 \mid X_0 = 1\} = (1/6)^2 + (1/6)^2$. More formally, element 1 of $\pi_0 P^2$.
$P\{X_{n+2} = 3 \mid X_n = 3\} = 10/36 = 5/6$. This is element (3, 3) of P^2.

(b) A chain is irreducible if it has only one class of intercommunicating states; that is, in some number of stages each state can lead to any other states. This chain is irreducible because the path $1 \to 2 \to 3 \to 4 \to 5 \to 6 \to 1$ is possible. Another way to look at it is that all elements in P^2 are positive, so any state can be reached from any other state in 2 stages.

A state is persistent if it is visited infinitely often over the long run. It is transient if it will surely be visited for a last time. In a finite irreducible chain all states are persistent.

A chain is ergodic if it has a limiting distribution. A finite irreducible chain is ergodic if it is aperodic of period d. (A state is periodic if the X-process can return to that state only at multiples of $d > 1$ stages. Here the process can return to any state in either 2 or 3 stages; there can be no periodicity because the greatest common divisor of 2 and 3 is 1.

(c) It is easy to verify that $\lambda P = \lambda$. Thus λ gives the steady-state (stationary) distribution of the chain. For an ergodic chain, the steady-state distribution is also the limiting distribution.

Each *row* in higher powers of P approaches λ. The approach is geometrically fast and errors decrease monotonically. Stop when you have the required number of decimal places of accuracy. In this case, entries in P^{32} are accurate to 5 or more decimal places.

(d) Hence, $\lim P\{X_n = 5 \mid X_0 = 3\} = \lambda_5 = 1/2$ and $\lim P\{X_n = 3\} = \lambda_3 = 1/12$.
Also, the average return time for a state is the reciprocal of the limiting probability that the process will be in that state: $\mu_1 = 1/\lambda_1 = 30$.

Note: Here is an algorithm for simulating this chain with a fair die:

For stage 0; because of the initial vector, we begin in state 1.

For stage 1 roll the die; go to the state indicated on the die.

For subsequent stages roll the die:

> If now in any state other than 6, go to the state corresponding to the maximum number seen on the die since the last visit to 6.
>
> If now in state 6, go to the state indicated on the die (which becomes the new "maximum").

[illegible] #4

(4) a) Ozone non normal
Control: skewed right, outlier left.

b) Ozone: $\bar{x} = 11.01$, $s = 19.02$
Control: $\bar{x} = 22.43$, $s = 10.78$

c) equal variance? No
normality? No
$H_0: \mu_{oz} = \mu_{con}$ $t = -2.492$ p-value = .0166
$H_A: \neq$
Reject H_0, There is a difference

d) 95% CI $(-20.657, -2.177)$

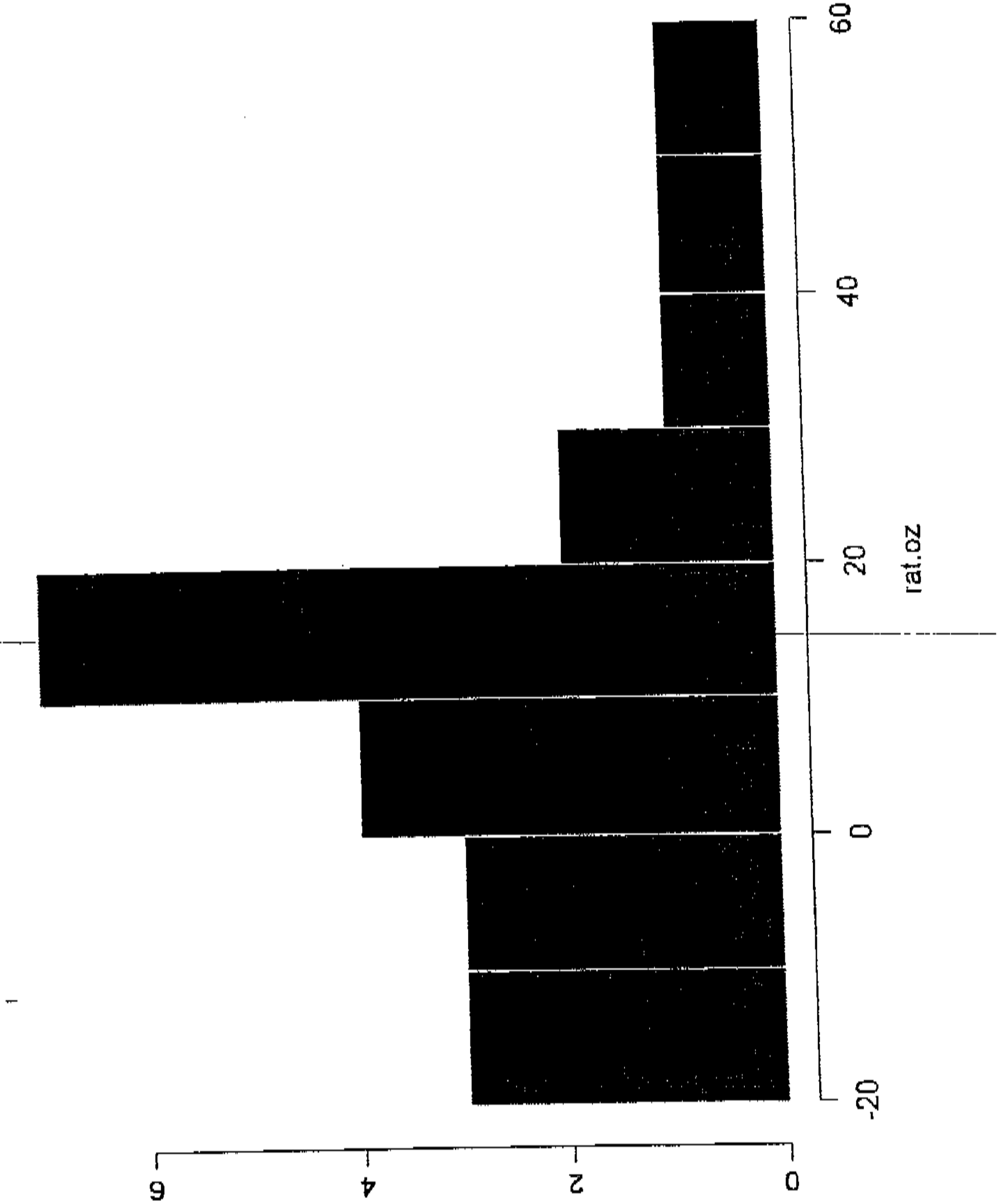
e) $\bar{\hat{\theta}}^* = -11.4$, $s_{\hat{\theta}} = 4.563$, 95% Bootstrap CI $(-20.16, -2.228)$

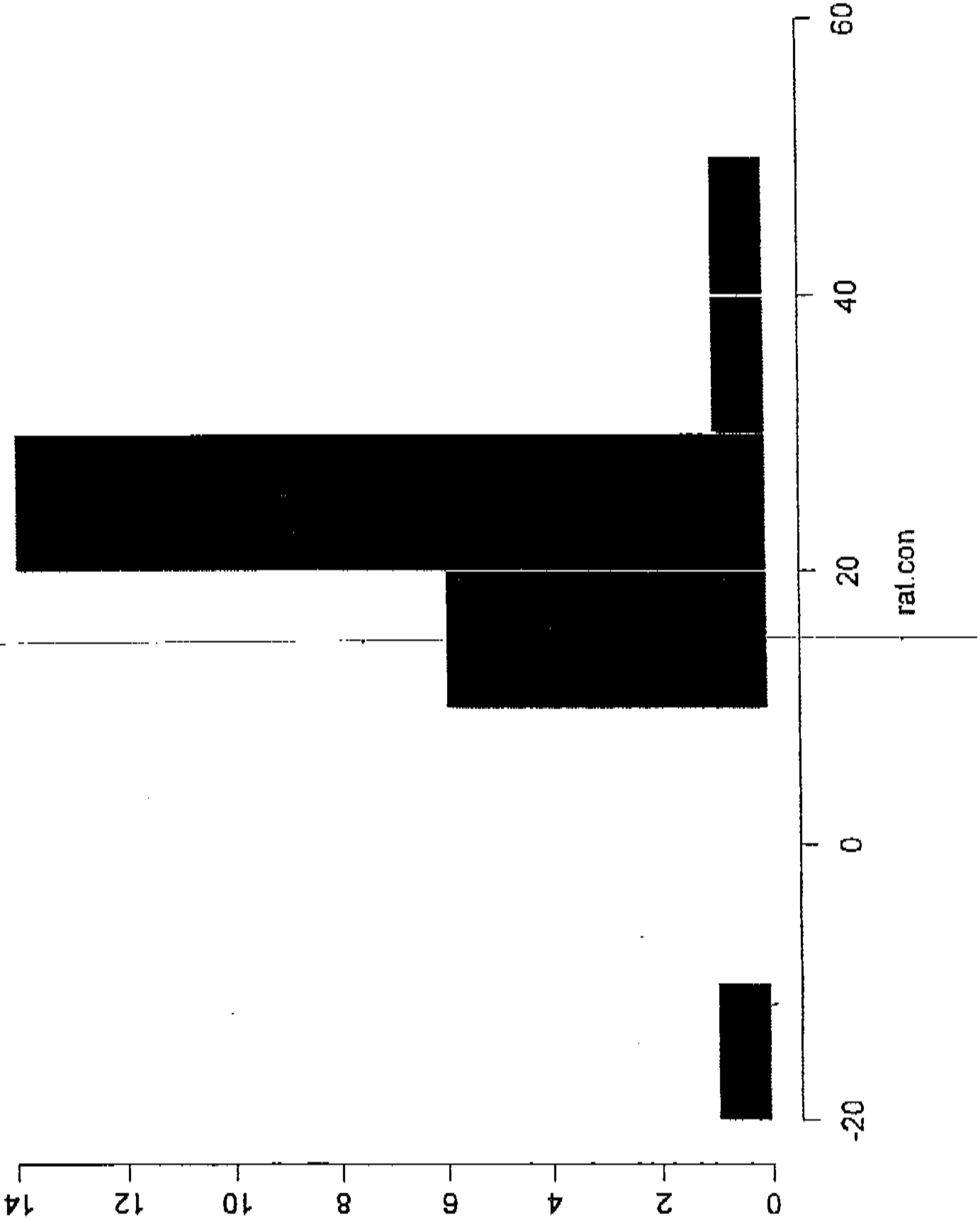
f) The Bootstrap CI is slightly more narrow.

g) Yes, There is an approximate decrease of -11.4 grams.

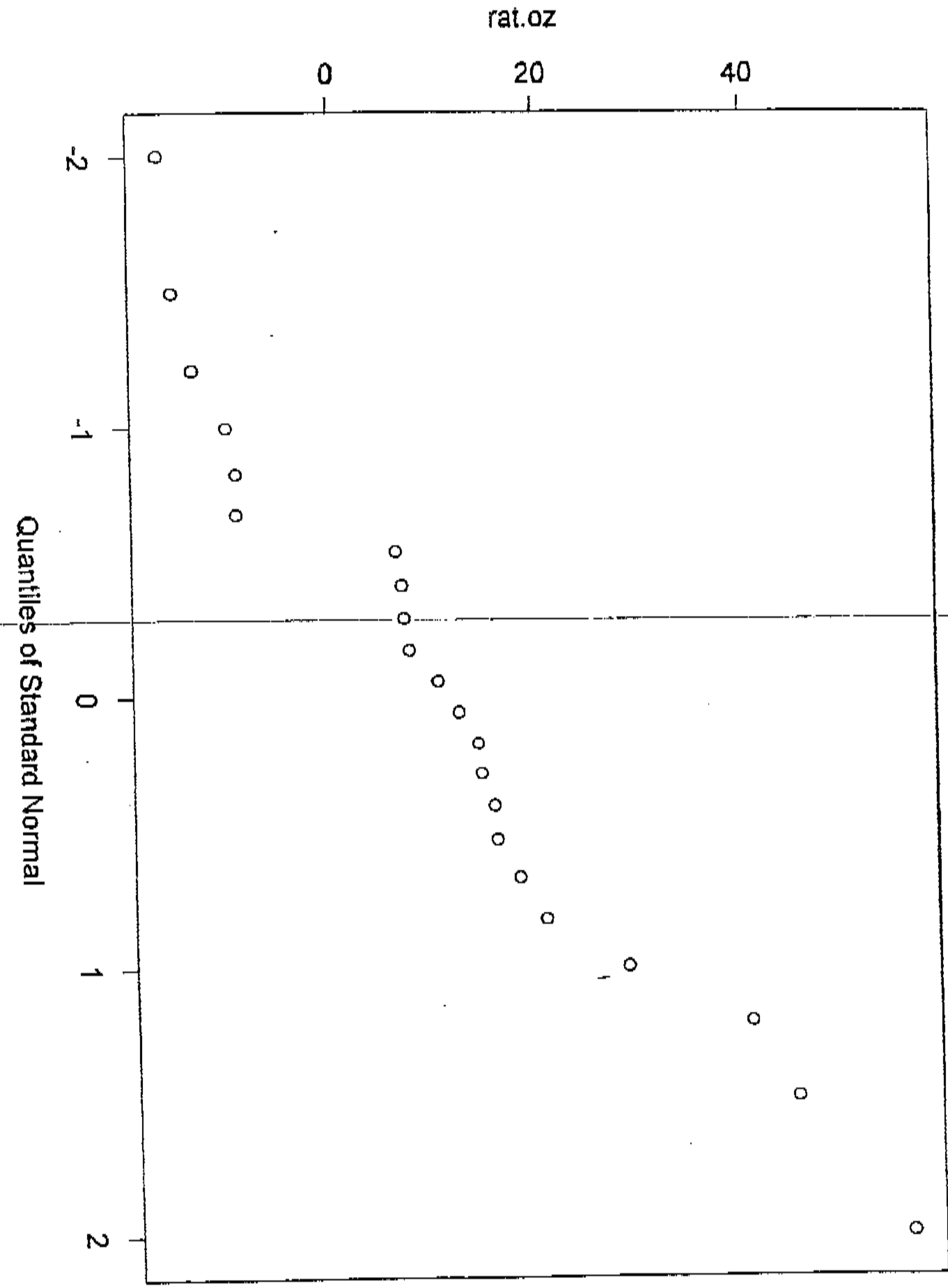
```
# ozone data from rats

rat.oz <- c(10.1, 6.1, 20.4, 7.3, 14.3, 15.5, -9.9, 6.8, 28.2, 17.9, -12.9, 14.0, 6.6, 12.1,
    15.7,
    39.9, -15.9, 54.6, -14.7, 44.1, -9.0, -9.0)

rat.con <- c(41.0, 38.4, 24.9, 25.9, 21.9, 18.3, 13.1, 27.3, 28.5, -16.9, 17.4, 21.8, 15.4,
    27.4,
    19.2, 22.4, 17.7, 26.0, 29.4, 21.4, 22.7, 26.0, 26.6)

# histograms

hist(rat.oz,xlim=c(-20,60))
hist(rat.con,xlim=c(-20,60))

# means

mean(rat.oz)
mean(rat.con)

mean(rat.oz) - mean(rat.con)

# standard deviations

sqrt(var(rat.oz))
sqrt(var(rat.con))

# check normality

qqnorm(rat.oz)
qqnorm(rat.con)

# independent two sample test and confidence interval

t.test(rat.oz,rat.con)

# nonparmetric bootstrap confidence interval

B <- 10000        # number of bootstrap samples

rat.oz.n <- length(rat.oz)    # rat.oz sample size
rat.con.n <- length(rat.con)  # rat.com sample size

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

for(i in 1:B){
    oz.samp <- sample(rat.oz,rep=T,size=rat.oz.n)     # bootstrap sample from rat.oz
    con.samp <- sample(rat.con,rep=T,size=rat.con.n) # bootstrap sample from rat.con
    mean.diff.boot[i] <- mean(oz.samp)-mean(con.samp)    # calculation of the bootstrap
    differences in the means
}

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

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

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

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

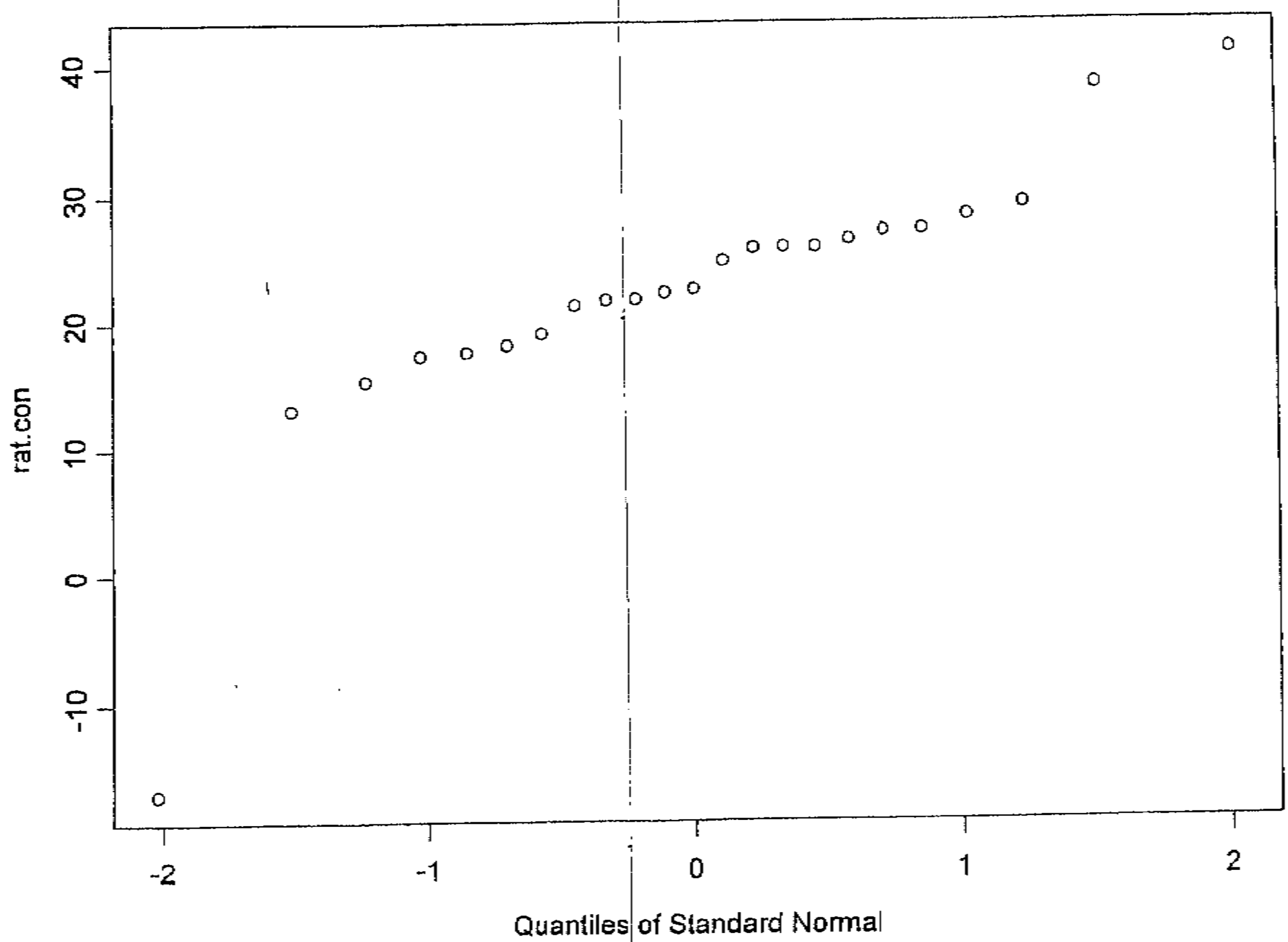
```
> # ozone data from rats
rat.oz <- c(10.1, 6.1, 20.4, 7.3, 14.3, 15.5, -9.9, 6.8, 28.2, 17.9, -12.9, 14, 6.6, 12.1,
    15.7, 39.9, -15.9, 54.6, -14.7, 44.1, -9, -9)
> rat.con <- c(41, 38.4, 24.9, 25.9, 21.9, 18.3, 13.1, 27.3, 28.5, -16.9, 17.4, 21.8, 15.4,
    27.4, 19.2, 22.4, 17.7, 26, 29.4, 21.4, 22.7, 26, 26.6) # histograms
> hist(rat.oz, xlim = c(-20, 60))
> hist(rat.con, xlim = c(-20, 60))  # means
> mean(rat.oz)
[1] 11.01
> mean(rat.con)
[1] 22.43
> mean(rat.oz) - mean(rat.con)   # standard deviations
[1] -11.42
> sqrt(var(rat.oz))
[1] 19.02
> sqrt(var(rat.con))   # check normality
[1] 10.78
> qqnorm(rat.oz)
> qqnorm(rat.con)   # independent two sample test and confidence interval
> t.test(rat.oz, rat.con) # nonparametric bootstrap confidence interval

    Standard Two-Sample t-Test

data:  rat.oz and rat.con
t = -2.492, df = 43, p-value = 0.0166
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -20.657  -2.177
sample estimates:
 mean of x mean of y
     11.01     22.43

> B <- 10000 # number of bootstrap samples
> rat.oz.n <- length(rat.oz) # rat.oz sample size
> rat.con.n <- length(rat.con)   # rat.com sample size
> mean.diff.boot <- numeric(B)   # vector for bootstrap differences in the means
> for(i in 1:B) {
    oz.samp <- sample(rat.oz, rep = T, size = rat.oz.n) # bootstrap sample from rat.oz
    con.samp <- sample(rat.con, rep = T, size = rat.con.n)
    # bootstrap sample from rat.con
    mean.diff.boot[i] <- mean(oz.samp) - mean(con.samp)
    # calculation of the bootstrap differences in the means
}
> mean(mean.diff.boot) # bootstrap estimate of the difference in the means
[1] -11.4
> sqrt(var(mean.diff.boot))   # bootstrap standard error of the difference in the means
[1] 4.563
> hist(mean.diff.boot) # bootstrap distribution of the difference in the means
> quantile(mean.diff.boot, c(0.025, 0.975))
    # calculation of the empirical bootstrap confidence interval
    2.5%   97.5%
 -20.16 -2.228
```

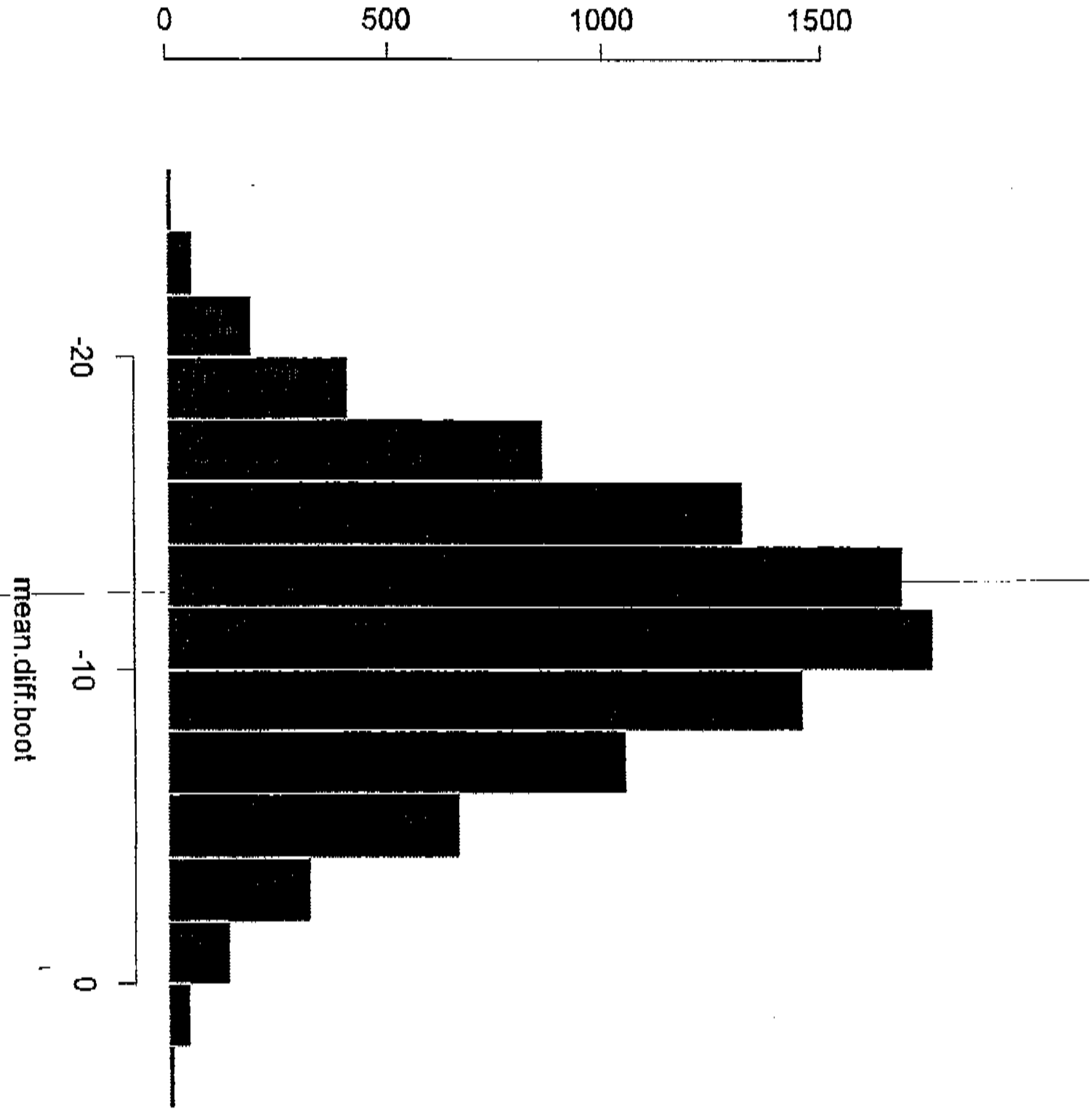

rat.oz
-20
0
20
40
60
0
2
4
6



rat.con
-20
0
20
40
60
0
2
4
6
8
10
12
14


rat.oz
0
20
40
Quantiles of Standard Normal
-2
-1
0
1
2


rat.con
40
30
20
10
0
-10
-2
-1
0
1
2
Quantiles of Standard Normal


0
500
1000
1500
-20
-10
0
mean.diff.boot

Openbook #5

Solution to open book version:

a.

```
Test of mu = 0 vs mu not = 0

Variable          N      Mean     StDev   SE Mean
C1               10     0.660     1.858     0.587

Variable              95.0% CI            T        P
C1             ( -0.669,   1.989)      1.12    0.290
```

One-way ANOVA: C1, C2, C3, C4, C5, C6, C7, C8

```
Analysis of Variance
Source     DF        SS        MS        F        P
Factor      7     7.489     1.070     2.99    0.011
Error      48    17.165     0.358
Total      55    24.654
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level       N      Mean     StDev  -----+---------+---------+---------+
C1         10    2.8400    0.9536                    (-----*------)
C2          8    2.6625    0.7050                (------*------)
C3         10    3.1800    0.2741                          (-----*-----)
C4          8    2.9750    0.3196                     (-------*------)
C5          6    2.3667    0.3830          (-------*--------)
C6          4    2.9000    0.2828                 (---------*---------)
C7          6    1.9833    0.6274    (-------*-------)
C8          4    2.3500    0.6245       (---------*---------)
                                   -----+---------+---------+---------+
Pooled StDev =   0.5980                1.80      2.40      3.00      3.60
```

According to this MINITAB output, There is a significant difference at 5% and at 10% between the means of the 8 litters. That is, we reject the hypothesis that all eight litter means are equal.

b. This is a planned comparison so we can use one of our degrees of freedom to test whether the means of the two sires are equal. If the sample sizes were the same in each group, we would form the comparison as $(\mu_1+\mu_3+\mu_4)/3-(\mu_2+\mu_5+\mu_7+\mu_8)/4$ and test this comparison as to whether it is zero. However, the sample sizes are different and we have to compute a weighted mean comparison. The standard error of the comparison $\Sigma\lambda*$ is $\sqrt{(s^2\Sigma(\lambda_i^2/n_i))}$. So using MSE as s^2, we have $\sqrt{(0.358(1/9/10+1/16/8+1/9/10+1/9/8+1/16/6+1/16/4+1/16/6+1/16/4))}=.186$ as the standard error for the comparison $1/3(8.995)-1/4(13.56)=-.3924$. so the t-statistics with error df is $-.3924/.186$ or $t=-2.109$ with 48 degrees of freedom. This is significant at 5%.

c. Since this comparison is a post hoc test, we need to use Scheffe's test. We compare the comparison divided by its standard error with $\sqrt{(8-1)}*F_{7,48}$ or 3.93. The linear comparison is: $1/4(2.8400+2.6625+3.1800+2.9750)-1/4(2.3667+2.9000+1.9833+2.3500)$ and the se is $\sqrt{(0.358(1/16/10+1/16/8+1/16/10+1/16/8+1/16/6+1/16/4+1/16/6+1/16/4))}=.47$ Hence the ratio of comparison to se is 1.09 which is smaller than 3.93 and hence the null hypothesis of no difference is not rejected.