May 2010

Ph.D. COMPREHENSIVE EXAMINATIONS
DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO

APPLIED STATISTICS COMPREHENSIVE EXAMINATION

May 13, 2010, 12:30 p.m. – 4:30 p.m.

Sidney Smith Hall

(#questions = 5); (#pages = 12 including cover page)

1. ATTEMPT ALL QUESTIONS.

   It is not necessary to completely solve every problem to achieve a good performance. *Emphasize what you do know.*

2. Please work neatly and legibly.

3. **Start each question in a new book, with your name and the number of the question on the front cover.** If there is more than one book for a question, then also indicate which is the first book and which second, e.g., Jane Smith, Question 5, Book 1 of 2.

4. The questions are not in any special order, nor are they all of equal difficulty.

5. The problems may be improperly phrased or may contain a misprint. Should this happen, reflect it in your discussion. Faculty members are not available to answer questions during the exam.

6. You are NOT permitted any aids (e.g., books, notes, etc.) **aside from a single non programmable calculator.**

7. Probability tables that may be useful are appended at the end of the exam paper, after all questions.

8. Good luck!

1. [1 page] Assume that in a clinical trial for an anti-arrhythmic drug, patients are randomized into a Treatment and a Control group. Patients in the treatment group receive the anti-arrhythmic drug, while those in the control group receive an inactive placebo. Cardiologists obtain an ECG for each patient for a certain time period, then count the number of extrasystolic (abnormal) heart beats that have been recorded during the specified time period. The goal is to determine whether the drug under consideration is effective in the sense of lowering the numbers of abnormal events compared to those in the control group. Besides the treatment information (yes or no), age, gender and length of the time period are available as covariates for 24 patients. It has been justified that the average occurrences of extrasystolic heart beats is roughly proportional to the length of the time period, given other covariates the same. In a first approximation, assume the number of extrasystolic heart beats is Poisson distributed. However, it is found that the Poisson parameter is in fact also random due to individual variation.

   (a) Denote the individual Poisson parameter by Z_i. To have a constant over-dispersion, what assumption is needed for Z_i? Express the constant dispersion parameter for this case.

   (b) Suppose that only main effects are significant, write down the GLM model with the canonical link and the categorical predictors expressed by dummy variables.

   (c) Suppose that you only have three quantities available: the estimate of the treatment effect, the deviance D_1 of the model in (b) and the deviance D_0 of the model without the treatment term. Write the appropriate null and *one-sided* alternative hypotheses in terms of the model parameters. Specify the rejection region of this test with level 0.05 in terms of the three available quantities.

2. [1 page] A fast food chain is considering a change in the blend of coffee beans they use to make their coffee. To determine whether their customers prefer the new blend, the company plans to select a random sample of n coffee-drinking customers and ask them to taste coffee made with the new blend and with the old blend, in cups marked "A" and "B." Half the time the new blend will be in cup A, and half the time it will be in cup B.

   Letting θ denote the true probability that a customer will prefer the new blend, and $\widehat{\theta}$ the proportion of customers in the sample who chose the new blend, company market researchers will compute

$$Z = \sqrt{n}\,(2\widehat{\theta} - 1).$$

   They will reject the null hypothesis and proceed to test market the new blend if $Z > z_0$. Please answer the following questions. A table of the standard normal distribution is provided, and you may use a calculator. Don't bother to correct for continuity.

   (a) Give a brief justification of the test statistic based on Central Limit Theorem.

   (b) Derive a general formula for the power function of the test that has an approximate significance level of $\alpha = 0.05$. Your answer is an expression involving n, z_0, θ and Φ, the cumulative distribution function of the standard normal distribution. *Show all your work and circle your final answer.*

   (c) Show that for fixed values of z_0 and $\theta > 0.50$, the power function is a strictly increasing function of n. *Show all your derivation.*

   (d) Suppose that $\theta = 0.60$ and $\alpha = 0.05$. What is the minimum sample size required so that the power of the test is at least 0.80? Your answer is a positive integer. *Show your work and circle your answer.*

3. [3 pages] The data in Table 1 below show the clotting time for two different samples of blood products, called "lot1" and "lot2", at different doses of a clotting agent, u. The researcher ran a regression of `log(time)` on `x = log(u)` and a variable `lot` that was created as a two-level factor, with levels "1" and "2". The R code and output for this is given in the Figure 1 on the next page.

Table 1: Data on clotting time of blood.

u	5,10,15,20,30,40,60,80,100
lot1	118,58,42,35,27,25,21,19,18
lot2	69,35,26,21,18,16,13,12,12

(a) The researcher was puzzled about the residual plots given by the software, and was worried about whether or not the needed assumptions for the analysis were satisfied. The researcher's supervisor said everything was fine because the R^2 for the regression was very large (0.9539). Write two or three sentences for the researcher and her supervisor discussing the validity of the linear model.

(b) The statistician they consulted suggested using instead a generalized linear model with gamma distributed errors. Results from this fit are given in the code in Figure 2. Write the equations for the two fitted lines for $\hat{\mu}^{-1}$, one for each lot, as functions of x, showing the standard errors for the estimated coefficients in parentheses. For example:

$$\text{for lot 1}: \qquad \hat{\mu}^{-1} = 3(\pm 1) + 7(\pm 2)x,$$

although of course $3, 1, 7$ and 2 are not correct.

(c) Explain how to carry out a test of constant slope for x in the two lots. Does the data provide evidence against the assumption of a constant slope? Give an approximate 95% confidence interval for the difference between the two slopes.

(d) Write a brief summary of the analysis for the researchers, explaining how the generalized linear model fit in (b) differs from the linear model fit in (a), and which model is to be preferred for this data.

Figure 1: R output for a linear model fit to the data in Table 1

```
> clottime = c(118,58,42,35,27,25,21,19,18,69,35,26,21,18,16,13,12,12)
> u = c(5,10,15,20,30,40,60,80,100,5,10,15,20,30,40,60,80,100)
> x = log(u)
> lot = factor(c(rep("1",9),rep("2",9)))

> lm.clot <- lm(log(clottime) ~ lot + x + log:x)
> summary(lm.clot)

Call:
lm(formula = log(clottime) ~ x + lot + x:lot)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16452 -0.10730 -0.04348  0.07614  0.25353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.47806    0.18073  30.312 3.62e-14 ***
x           -0.59704    0.05252 -11.367 1.87e-08 ***
lot2        -0.58179    0.25558  -2.276   0.0391 *
x:lot2       0.03383    0.07428   0.455   0.6558
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1482 on 14 degrees of freedom
Multiple R-squared: 0.9539,Adjusted R-squared: 0.944
F-statistic: 96.47 on 3 and 14 DF,  p-value: 1.371e-09

> par(mfrow=c(2,2),oma = c(0, 0, 1.1, 0))

> plot(lm.clot, las = 1)
```

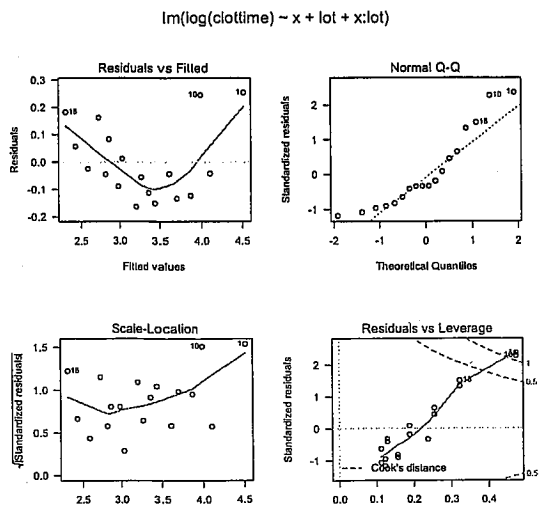
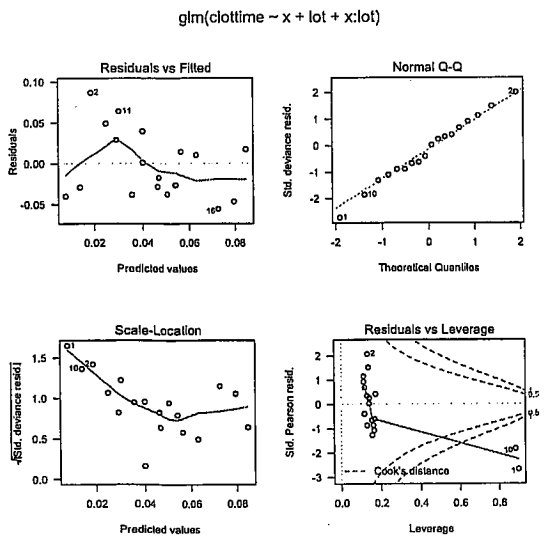


Figure 2: R code and output for fitting a gamma model to the clotting time data.

```
> summary(glm.clot)

Call:
glm(formula = clottime ~ x + lot + x:lot, family = Gamma(link = "inverse"))

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.055738  -0.035480  -0.008216   0.026073   0.086411

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0165544  0.0008655 -19.127 1.97e-11 ***
x            0.0153431  0.0003872  39.626 8.85e-16 ***
lot2        -0.0073541  0.0016780  -4.383 0.000625 ***
x:lot2       0.0082561  0.0007353  11.228 2.18e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for Gamma family taken to be 0.002129707)

    Null deviance: 7.708667  on 17  degrees of freedom
Residual deviance: 0.029401  on 14  degrees of freedom
AIC: 63.195

Number of Fisher Scoring iterations: 3
```



4. [1 page] Given a sample consisting of 15 hospitals in Toronto area and 15 in Montreal area, each with 10 patients that potentially would have heart attacks in 2009, three treatments (A, B, C) were randomly assigned to this sample, i.e., each treatment group consists of 10 hospitals, then the numbers of patients encountering heart attacks were observed. Also included as predictor is the the average age of the patients in each hospital. It is of interest to test whether the treatment effect varies (across group A, B, C) in the same pattern for hospitals in Toronto and Montreal. Consider a Binomial model with the logistic link function.

   (a) Write down the full model and the reduced model involved in the above test, explicitly specifying the model components, predictors and parameters. Then express the null hypothesis in terms of regression parameters. (*Hint: you need to use suitable dummy variables to describe the models.*)

   (b) The residual deviances of the full model and the reduced model are estimated as 75 and 57, respectively. The average numbers of heart attack patients in the hospitals within eachcombination of treatment and area are also provided below. Describe how you will perform this test, i.e., specifying the value of the test statistic and its distribution under the null hypothesis.

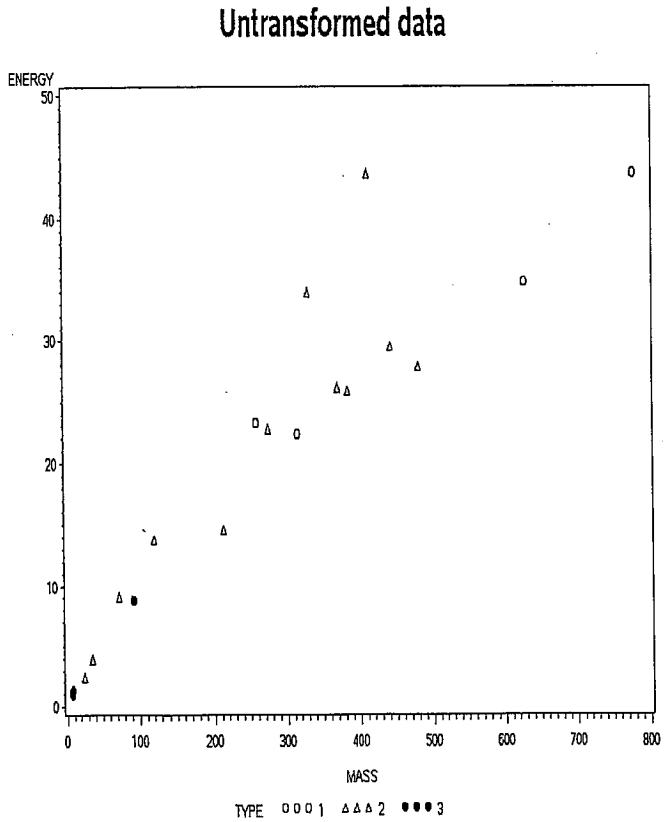
	Trt. A	Trt. B	Trt. C
Toronto	3.2	2.8	1.9
Montreal	4.5	3.3	0.0

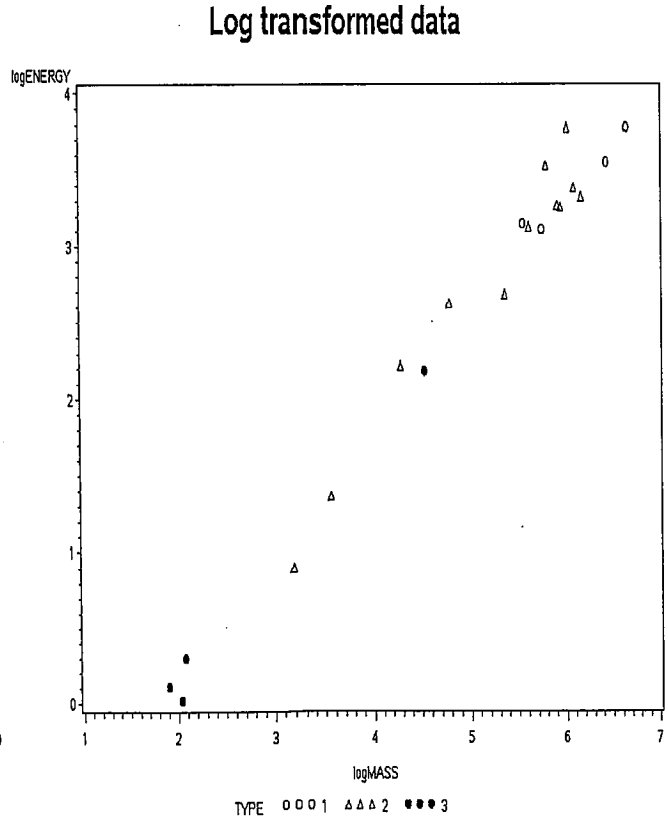
5. [3 pages] Echolocation is a method used by some bats for navigation and hunting. Echolocating animals send out calls and use the information received from echoes to locate and identify objects. Zoologists are interested in whether the combined energy needed for echolocation and flight in bats is the sum of the energy needed for flight plus the energy needed for echolocation, or whether the combined energy is less than the sum. Data were collected on in-flight energy expenditure and body mass for three types of animals: echolocating bats (4 observations), non-echolocating bats (4 observations), and non-echolocating birds (12 observations).

   The dataset includes the following variables for the 20 observations:
   - MASS - the body mass of the animal in grams
   - ENERGY - flight energy expenditure in Watts
   - TYPE - 1=non-echolocating bats, 2=non-echolocating birds, 3=echolocating bats

   The following output was obtained from SAS and is given on the next 3 pages:
   - scatterplots of ENERGY versus MASS and logENERGY versus logMASS, where logENERGY and logMASS are the natural log transformations of ENERGY and MASS, respectively
   - output from two linear models (full and reduced) using the log-transformed variables
   - a series of model diagnostic plots for the reduced model

   Questions begin after the SAS output.





FULL MODEL

Number of Observations Used 20

Dependent Variable: logENERGY

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	29.46993210	5.89398642	163.44	<.0001
Error	14	0.50486797	0.03606200		
Corrected Total	19	29.97480007			

R-Square	Coeff Var	Root MSE	logENERGY Mean
0.983157	7.650468	0.189900	2.482201

Source	DF	Type III SS	Mean Square	F Value	Pr > F
logMASS	1	3.37875395	3.37875395	93.69	<.0001
TYPE	2	0.04122471	0.02061236	0.57	0.5773
logMASS*TYPE	2	0.04844952	0.02422476	0.67	0.5265

REDUCED MODEL

Number of Observations Used 20

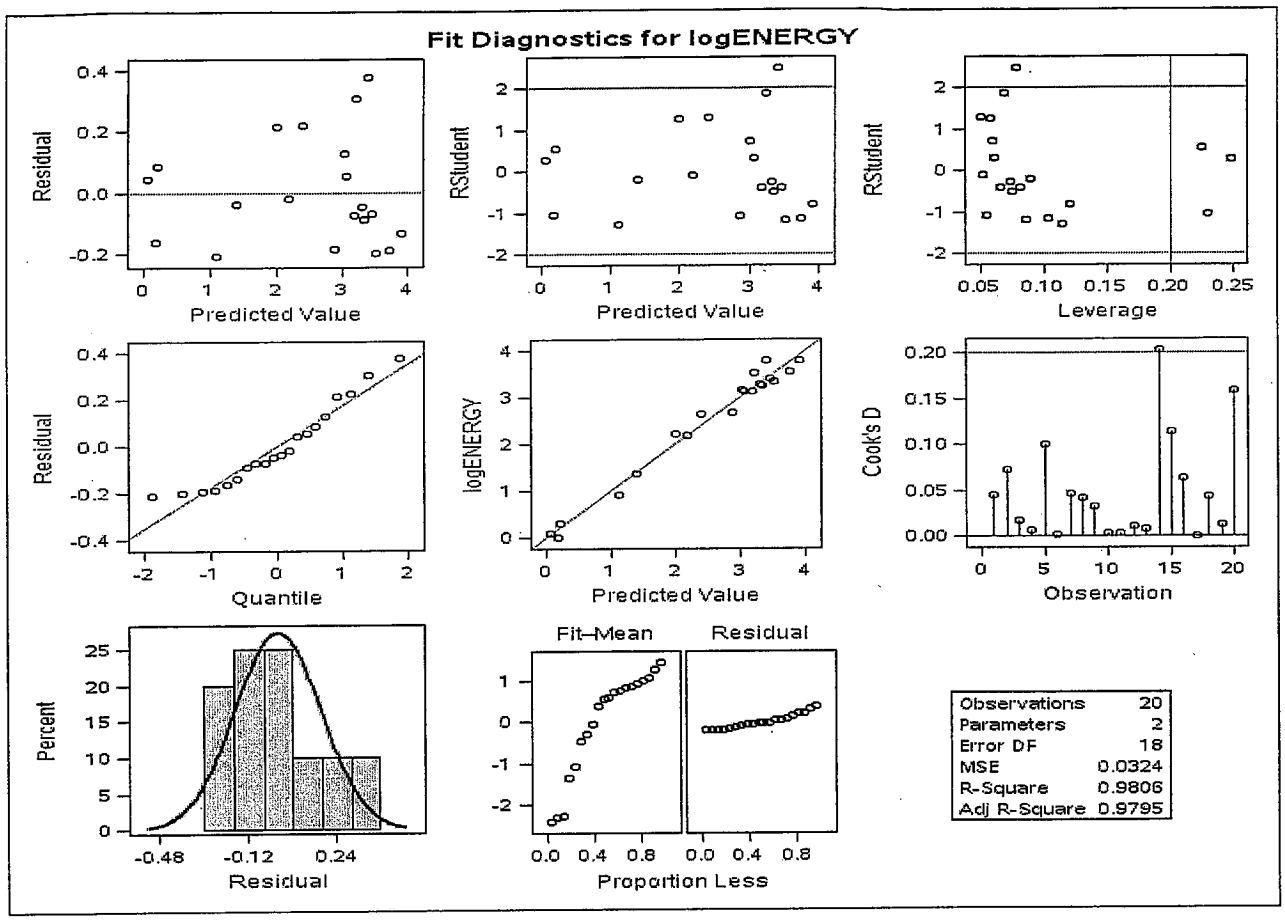
Dependent Variable: logENERGY

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	29.39190898	29.39190898	907.64	<.0001
Error	18	0.58289110	0.03238284		
Corrected Total	19	29.97480007			

R-Square	Coeff Var	Root MSE	logENERGY Mean
0.980554	7.249709	0.179952	2.482201

Source	DF	Type III SS	Mean Square	F Value	Pr > F
logMASS	1	29.39190898	29.39190898	907.64	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-1.468258423	0.13716178	-10.70	<.0001
logMASS	0.808609747	0.02684000	30.13	<.0001



(a) The purpose of the analysis was to examine the differences in energy expended among the 3 types of animals. However body mass of the animals was also included in the linear models. Why was it included?

(b) Linear models were fit using the log-transformed variables. Explain why it is appropriate to use the transformed variables.

(c) For the full model, what is being tested by the test with p-value 0.5625? What do you conclude about the logENERGY-logMASS relationship?

(d) Carry out an hypothesis test with null hypothesis that the coefficients are 0 for all of the terms in the full model that are not in the reduced model. What do you conclude from this test about the logENERGY-logMASS relationship and the energy required for echolocation?

(e) The fitted reduced model is $\widehat{\textbf{logENERGY}} = -1.47 + 0.81\,\textbf{logMASS}$. Give a practical interpretation of the coefficient of logMASS in terms of the relationship between ENERGY and MASS.

(f) For the reduced model, what conditions must hold for the inferences to be correct? Do the conditions appear to hold or not? If you need more information to answer this, indicate what information you would like to have.

TABLE A Table entry for z is under the standard normal curve to the left of z.

TABLE A Standard normal probabilities (*continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

VassarStats: Table of Critical F Values (p. 1)

[top entry for .05 level; bottom entry for .01 level]

df denominator	df numerator													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	161	199	216	225	230	234	237	239	241	242	243	244	245	245
	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6083	6107	6126	6143
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42
	99	99	99	99	99	99	99	99	99	99	99	99	99	99
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71
	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.98	26.92
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87
	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.31	14.25
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64
	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.82	9.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96
	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53
	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24
	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03
	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86
	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74
	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64
	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55
	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48
	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42
	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56