

## Statistics and Data Analysis

**Professor William Greene**

Phone: 212.998.0876

Office: KMC 7-78

Home page: [www.stern.nyu.edu/~wgreene](http://www.stern.nyu.edu/~wgreene)

Email: [wgreene@stern.nyu.edu](mailto:wgreene@stern.nyu.edu)

Course web page: [www.stern.nyu.edu/~wgreene/Statistics/Outline.htm](http://www.stern.nyu.edu/~wgreene/Statistics/Outline.htm)

### Final Examination

This is an open book, open notes test. You may use any notes or printed material you wish. You may use a calculator. No use of cell phones, PDAs or computers is permitted, however.

There are 120 points in this test.

A table of probabilities for the normal distribution appears at the end of the test.

### Part I. Multiple Regression Model

[10 points]

In a carefully worded paragraph, explain how the multiple linear regression model explains the relationships between a dependent variable and several independent variables. As part of your explanation, describe how the coefficients in the model are interpreted. Also as part of your explanation, describe how to interpret the  $R^2$  in a multiple regression equation.

Note: As we discussed in class, I assume that you have crafted and printed an answer to this question before coming to this exam. You may simply hand in your printed answer; you need not rewrite the answer in your exam booklet. Be sure that your name appears on this additional sheet.

The multiple regression model is a device for analyzing the covariation of a variable to be explained and a set of variables that are believed to explain its variation. The important feature of the approach is that it allows the researcher to study the “effect” of an explanatory variable on the dependent variable while holding all other variables constant. This is an experiment that might not actually exist in the data, for example, variation education while holding constant. The coefficients in the model are interpreted as the “partial effects.” Each of them is interpreted to measure the effect of a unit change in the variable in question on the dependent variable, again, while holding all other variables constant. The  $R^2$  in the regression measures the strength of the association between the dependent variable and the independent variables. Formally, it is the proportion of the variation in the dependent variable that is accounted for by variation of the independent variables. A useful way to interpret  $R^2$  is as the square of the correlation between the predictions of the model and the actual values of the dependent variable.

## Part II.

[15 points]

1. When the filling process is in control, the fill amounts that go into Pure Pacific beer are supposed to be normally distributed with a mean of 12.0 ounces and a standard deviation of 0.1 ounces. Every 30 minutes, a bottle is selected randomly from the filling line and its contents are noted precisely. If the fill amount is found to be above 12.15 or below 11.85 ounces, then the process is declared out of control, the line is stopped (and the staff has to drink all the beer that has already been filled).

a. If the process really is in control, what is the probability that it will be declared out of control when the next bottle is inspected.

$$P(x > 12.15) + P(x < 11.85) = P(z > 1.5) + P(z < -1.5) = .0668 + .0668 = .1336$$

b. If 16 bottles are pulled off the line in an 8 hour shift, what is the probability that exactly one bottle will be outside the control range (11.85 to 12.15 ounces).

$$\text{Binomial probability, } P = .1336, n = 16, \text{Prob}(x = 1) = {}_{16}C_1 \times .1336 \times (1-.1336)^{15} = .2487$$

c. Suppose that the process has become out of control – the mean has shifted to 11.80 though the standard deviation is the still the same. If the management pulls a bottle randomly off the line and uses the test described above (in part a), what is the probability that they will fail to detect that the process is out of control.

$$P(x > 12.15|\mu=11.8) + P(x < 11.85|\mu=11.8) = P(z > 3.5) + P(z < 1) = .0002 + .6915 = .6917$$

This is the probability that they will conclude the process is out of control. The probability that they will not conclude it is out of control is  $1 - .6917 = .3083$ .

## Part III. Answer exactly one of these two questions.

[5 points]

1. Speaking of problems at the big three. In 1971, General Motors introduce the Chevrolet Vega, a small car designed to compete with Japanese imports. The Vega was manufactured (from 1971 to 1977) at a new plant at Lordstown, Ohio. Though the rumors were untrue and unfair, it was rumored that a significant number of defective cars were made at Lordstown, especially on Mondays and Fridays. Consider the following proportions (completely fictional just for the purpose of this example) for the manufacture of Chevrolet cars

	Chevrolet Cars	
	Lordstown	Elsewhere
Defects	.04	.03
No Defects	.16	.77

a. If a car is found to be defective, what is the probability it was built at Lordstown?

$$P(L|D) = P(L,D)/P(D) = .04 / .07 = 4/7$$

b. What is the probability that a Chevrolet car was not made at Lordstown?

$$P(\text{not } L) = P(E) = .03 + .77 = .8$$

[5 points]

2. Twenty years ago, a group of Harvard researchers tallied the marriage rates for women of varying ages and found that white, college-educated women who failed to marry in their 20s faced abysmal odds of ever tying the knot. According to the research, a woman who remained single at 30 had only a 20 percent chance of ever marrying. By 35, the probability dropped to 5 percent. Not surprisingly, the story made the cover of People Magazine. In the story's most infamous line, NEWSWEEK reported that a 40-year-old single woman was "more likely to be killed by a terrorist" than to ever marry. That comparison wasn't in the study, and even in those pre-9/11 days, it struck many people as an offensive analogy. Nonetheless, it quickly became entrenched in pop culture and is still routinely cited in TV shows and news stories. (<http://www.houselustthebook.com/articles/marriage-by-the-numbers/>)

In one of the more astute analyses of the research, (singer and aspiring statistician) Cher stated that the study couldn't possibly be right because she was over 30 and she was about to marry. Is Cher right in her assessment of the study? Comment.

Cher should not quit her day job. Just because an event is unlikely does not mean it is impossible. Cher appears not to understand the notion of probability. She is wrong.

#### Part IV. Answer exactly 2 of these 4 questions.

[10 points]

1. A (hypothetical) Senatorial vote on whether to bail out the maraschino cherry industry was split as follows:

	Republican	Democrat	Total
Voted Yes	12	46	58
Voted No	28	14	42
Total	40	60	100

Did the senators vote (roughly) along party lines? Carry out a test of the hypothesis that vote and party were independent.

Chi squared =  $n \times \text{sum for all 4 cells, } (\text{Observed} - \text{Expected})^2 / \text{Expected}$

Observed = the proportion in the table.

Expected = row proportion  $\times$  column proportion.

These are  $(.40)(.58) = .232$   $(.60)(.58) = .348$

$(.40)(.42) = .168$   $(.60)(.42) = .252$

Chi squared =  $100[(.12 - .232)^2 / .232 + (.46 - .348)^2 / .348 + (.28 - .168)^2 / .168 + (.14 - .252)^2 / .252]$   
= 21.46

The critical value is 3.84, so the hypothesis that party and vote are independent is rejected.

Yes, they voted by party.

[10 points]

2. 32 high school students took a test of economic literacy. 14 of them had previously taken a special economics course, 18 had not. The test scores produced the following information

	Mean	Standard Deviation	Sample Size
Did not take the course	21.5	4.003	18
Did take the course	24.3	3.857	14

- a. Form a 95% confidence interval for the population mean score of students who took the course.

$$24.3 \pm t[13] \times 3.857 / \sqrt{14} = 24.3 \pm 2.160 \times 1.031 = 24.3 \pm 2.227 \\ (22.073 \text{ to } 26.527)$$

- b. Test the hypothesis that the population mean for those who did not take the course equals 20. Explain your assumptions and show your calculations in detail.

$$H_0: \mu = 20$$

$$H_1: \mu \text{ not equal to } 20.$$

Assume normally distributed, but small sample.

Rejection region is sample means far from 20

$$21.5 \text{ is } (21.5 - 20) / [4.003 / \sqrt{18}] = 1.415 \text{ standard errors away from } 20$$

The critical t value with 17 degrees of freedom is 2.110, so the hypothesis

That the mean is 20 cannot be rejected.

- c. How would you test the hypothesis that the economic literacy course raises test scores? Show how you would do the computations. What is the rejection region for your test?

The hypothesis is equivalent to the hypothesis that the mean of the second group

Is larger than that of the first, or that the difference is greater than zero. The

Difference in the means is  $24.3 - 21.5 = 2.8$ .

Estimate the standard error of the difference with

$$\sqrt{4.003^2/18 + 3.857^2/14} = 1.397.$$

2.8 is  $(2.8 - 0)/1.397 = 2.003$  standard errors above zero.

We are doing a one tailed test, so we want all 5% in the upper side of the

Distribution. So, the critical value from the normal distribution is 1.645.

So, we will reject the hypothesis that the difference in the means is less

Than or equal to 0.0.

**[10 points]**

3. Computation of a linear regression produces the following Analysis of Variance Table:

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Statistic
Regression	3	1.7	C	D
Residual	A	15.3	E	
Total	123	B	F	

a. Fill in the missing values A, B, ..., F in the table.

$$A=120, B = 17, C = 1.7/3 = .5666, E = 15.3/120 = .1275, F = 17/123 = .1382, \\ D = C/E = .5666/.1275 = 4.444$$

b. What is the  $R^2$  for the regression?

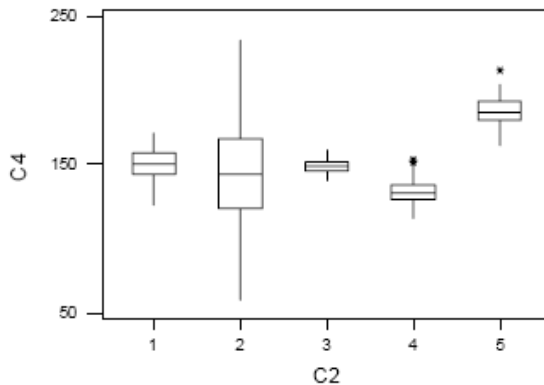
$$R^2 = 1.7/B = 1.7/17 = .1000$$

c. Do you think we would reject the hypothesis that all the coefficients in the model are zero? Explain your answer.

The F statistic is 4.0, there are 3 predictors, and the sample is pretty large. The critical F is going to be close to 2, so yes, we will be rejecting the hypothesis. To be precise, we need a table, but we can be confident based just on this.

**[10 points]**

4. Here is a display showing five boxplots:



The five variables pictured here have the following selected numerical summaries:

	N	Mean	Median	StDev	PLOT
ARTICHOKE	123	149.03	148.95	3.83	3
BERRY	117	150.42	150.32	10.02	1
COLLAR	124	144.06	143.39	33.30	2
DIANTHUS	119	185.91	185.44	8.78	5
EGGPLANT	117	131.93	130.83	8.31	4

Match the boxplots to the variables.

**Part V.**  
**[20 points]**

In a study of airline costs, a regression (not using Minitab) of the log of costs on the log of output and the log of the price of fuel produces the following results:

```

+-----+
| LogCost      Mean          = 13.31689 |
|              Standard deviation = 1.218336 |
| Number of observations = 246 |
| Residuals    Sum of squares = 5.328249 |
|              Standard error of e = .1480775 |
| Fit          R-squared      = .9853484 |
|              Adjusted R-squared = .9852278 |
| Model test   F[ 2, 243]     = 8171.12 |
|              P value for F    = .0000 |
+-----+
+-----+-----+-----+-----+
|Variable| Coefficient | Standard Error |t-ratio |P value |
+-----+-----+-----+-----+
|Constant| 9.53520*** | .16169637     | 58.970 |.0000 |
|LQ      | .83908***  | .00775014     | 108.266|.0000 |
|LPF     | .37578***  | .01248221     | 30.105 |.0000 |
+-----+-----+-----+-----+

```

The following results are obtained when three variables, load factor, stage length and points served, are added to the model.

```

+-----+
| LogCost      Mean          = 13.31689 |
|              Standard deviation = 1.218336 |
| Number of observations = 246 |
| Residuals    Sum of squares = 3.387771 |
|              Standard error of e = .1188096 |
| Fit          R-squared      = .9906843 |
|              Adjusted R-squared = .9904903 |
| Model test   F[ 5, 240]     = 5104.61 |
|              P value for F    = .0000 |
+-----+
+-----+-----+-----+-----+
|Variable| Coefficient | Standard Error |t-ratio |P value |
+-----+-----+-----+-----+
|Constant| 9.82216*** | .14186370     | 69.237 |.0000 |
|LQ      | .88541***  | .01182029     | 74.906 |.0000 |
|LPF     | .38403***  | .01184508     | 32.421 |.0000 |
|LOADFCTR| -.70199*** | .19236111     | -3.649 |.0003 |
|STAGE   | -.00029*** | .0000456915   | -6.281 |.0000 |
|POINTS  | .00264***  | .00032388     | 8.165  |.0000 |
+-----+-----+-----+-----+

```

a. Is output a significant predictor of costs? How do you know?

Yes. The t ratio on logOutput is 108.266 in the first model and 74.9 in the second. This is highly significant.

b. Would you say that the first regression is a good predictor of the dependent variable?

The two coefficients are highly significant and  $R^2$  is .9853. And, the regression model makes Sense. So, I would say yes.

c. Do the three additional variables in the second regression provide a significant additional fit in the model? How can you determine this?

This needs the F test done in class.  $F = [(.9906843 - .9853484)/3] / [(1 - .9906843)/(240)] = 45.822$ . This is a large F. Much larger than 2. I would say that the additional variables Do add significantly to the fit, even though it does not seem like  $R^2$  goes up by much.

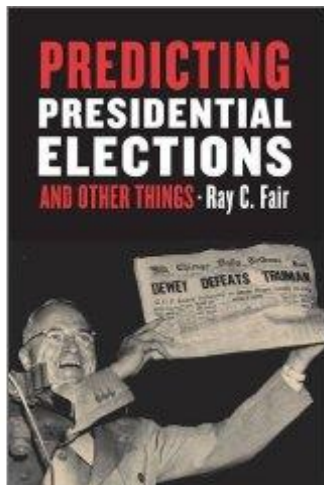
d. How would you interpret the coefficient on POINTS in the second regression model?

The coefficient on points gives the change in the log of costs given one more point served. A change in a log is a percent change. So, this says that costs go up by 0.2% when one more Point (city) is served.

e. Is it possible to determine the regression sum of squares in the second regression? How?

It is. Since  $R^2 = \text{Regression Sum of Squares} / \text{Total Sum of Squares}$ ,  
Regression Sum of Squares =  $R^2 \times \text{Total sum of squares}$ ,  
 $R^2 = .9906843$ .  $N = 246$ . The standard deviation of LogCost is 1.218336, so the Variance is the square, 1.4843. The total variation is 245 times this, or 363.664. So, the Regression sum of Squares =  $363.664 \times .9906843 = 360.276$ .

## Part IV. Model Building [50 points]



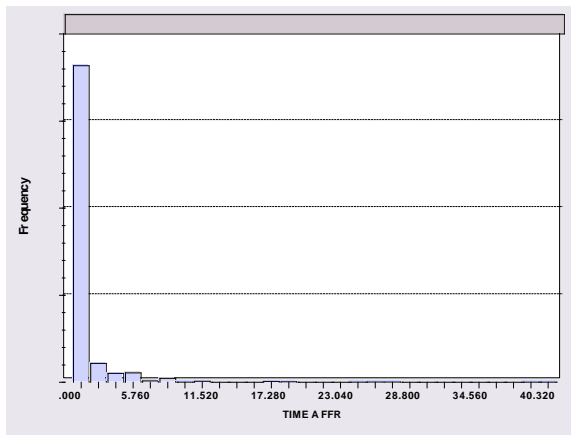
### Predicting Presidential Elections and Other Things Ray Fair, Stanford U. Press, 2002

In addition to “big” topics—presidential elections, Federal Reserve behavior, and inflation—and “not quite so big” topics—wine quality—the book takes on questions of more direct, personal interest. **Who of your friends is most likely to have an extramarital affair?** How important is class attendance for academic performance in college?

This book, published in 2002, describes many of Ray Fair’s [see <http://fairmodel.econ.yale.edu/>] research projects over a 35 year career as an economist. Among Ray’s most famous papers is “A Theory of Extramarital Affairs,” in which he examined from the point of view of an economist the time that married people spent in extramarital activity. As part of that research, he examined survey data sets gathered by two magazines, *Psychology Today* and *Redbook*. We are going to examine the second of these. The variables in this data set, which included both men and women are:

RATEMARR = rating of marriage, coded 1 to 5  
AGE = age in years  
YRSMARR = number of years in present marriage  
RELIG = religiosity coded 1 to 4 with 1 least religious and 4 most  
NUMKIDS = number of children recorded as 5 if more than 5  
EDUCYRS = years of education  
WIFE OCC = wife’s occupation on Hollingshead scale

HUSBOCC = husband's occupation on Hollingshead scale  
 The Hollingshead scale rates occupations 1-6 essentially by prestige  
 TIMEAFFR = time spent in extramarital affairs



This is a histogram of the dependent variable, TIMEAFFR that he (we) will study.

- Are the data skewed left or right?
- The sample mean of the variable is 0.71. What would be a good estimate of the standard deviation, 0.2, 2.0, or 20.0. Explain.
- Estimate the median of TIMEAFFR.

1.

- The data are skewed in the direction of the long tail, to the right.
- The range of the variable looks to be about 0 to 10 or so, so only 2.0 makes sense. The range should be about 4 standard deviations.
- Almost everyone has a zero, so the median is zero.

2. The first set of regression results (not computed with Minitab) is

```

+-----+
| LHS=TIMEAFFR Mean           =   .7053739 |
|                               Standard deviation =   2.203374 |
| Model size Degrees of freedom =       5392 |
| Fit R-squared                =   .0540677 |
|                               Adjusted R-squared =   .0533240 |
| Model test F[ 7, 5392] (prob) = 72.71 (.0000) |
+-----+
+-----+-----+-----+-----+-----+
|Variable| Coefficient | Standard Error | b/St.Er. | P Value |
+-----+-----+-----+-----+-----+
|Constant| 3.60945*** | .23280645      | 15.504   | .0000   |
|RATEMARR| -.42004*** | .02836033      | -14.811  | .0000   |
|AGE      | -.01477*   | .00878538      | -1.682   | .0926   |
|YRSMARR | -.01457    | .00961659      | -1.515   | .1297   |
|RELIG    | -.24364*** | .03111409      | -7.830   | .0000   |
|NUMKIDS | -.01849    | .02969534      | -.622    | .5336   |
|WIFEOCC | .06577**   | .03131989      | 2.100    | .0357   |
|HUSBOCC | .00405     | .02076489      | .195     | .8454   |
+-----+-----+-----+-----+-----+

```

- How many observations were used to compute the regression  
 Degrees of freedom =  $N - \text{\# predictors} - 1$ .  $5392 = N - 7 - 1$ , so  $N = 6000$
- Do you think this regression provides a “significant” explanation of time spent in extramarital affairs?  
 The R squared is pretty low, only about 0.05. But, a couple of the variables are very significant, And the F statistic is extremely large with a P value of 0.000. So, yes, it seems to.
- What are the significant predictors of time spent in extramarital affairs in these results?  
 Those with t statistics larger than 2 are RATEMARR, RELIG and WIFEOCC.
- What does the reported value  $p = 0.0926$  reported with AGE mean?

This is the probability that I would observe a coefficient as large as -.01477 (in absolute value) if the real coefficient on AGE was zero.

e. Can you provide a specific interpretation to the reported value of -0.24364 reported with RELIG in the results?

RELIG is coded 1,2,3,4. The coefficient says that for each increase in this variable of 1 unit, the Amount of time spent in extramarital affairs falls by .24364. (Time was measured in hours per month.)

f. Explain the meaning of the *F* statistic and the associated *p* value reported.

The F statistic is used to test the hypothesis that all of the coefficients are zero (that there is no Model. The F value shown is 72.71 and the P value is 0.0000. This says that there is zero Probability that I would observe the regression coefficients as large as these if they were Actually all zero.

g. Based on these results, test the hypothesis that the coefficient on Age in the model is zero.

We did this in part d., The t statistic is -1.682. The critical value for 6,000 degrees of freedom Is 1.96. Therefore, I would not reject the hypothesis that the coefficient on AGE is zero. The Fact that the P value is .0926 > 0.0500 says the same thing.

h. The book description states something about “who ... is most likely to have an ... affair.” Of course, the linear regression model does not do that at all. I created another variable, AFFAIR = 1 if the time spent in affairs is greater than 0 and 0 if it is 0. In this data set, using 6000 observations, the proportion of people who reported having an affair is .322. What would you report to the reporter from People Magazine who is interested in a plausible range of values for the proportion of people in the population who are having extramarital affairs.

This is the same as the computation we did for election returns. The estimated mean is .322 The estimated standard error will be  $\sqrt{.322(1-.322)/6000} = .006$ . So, our confidence interval Would be  $.322 \pm 1.96(.006) = .322 \pm .0118 = .3102$  to  $.3338$ .

i. These are the results of a logistic regression. What does the model suggest is the best variable for predicting whether someone will have an affair? Explain.

I would bet on RATEMARR. Not only is it the most significant (largest t ratio) it is also The largest coefficient.

```

+-----+
| Binary Logistic Model for AFFAIR |
+-----+
+-----+-----+-----+-----+
|Variable| Coefficient | Standard Error |b/St.Er.|P value |
+-----+-----+-----+-----+
|Constant| 3.72572*** | .29876337 | 12.470 | .0000
|YRSMARR | .11002*** | .01094293 | 10.054 | .0000
|RATEMARR| -.71611*** | .03143062 | -22.784 | .0000
|AGE | -.06049*** | .01027798 | -5.885 | .0000
|NUMKIDS | -.00423 | .03161398 | -.134 | .8935
|RELIG | -.37516*** | .03476335 | -10.792 | .0000
|EDUCYRS | -.03922** | .01548038 | -2.533 | .0113
|WIFE0CC | .16023*** | .03397089 | 4.717 | .0000
|HUSBOCC | .01240 | .02292554 | .541 | .5886
+-----+

```

## Cumulative Normal Probabilities

The table entry is  $\text{Prob}[Z \leq z]$ . Note that this is not the same as the table in your text. In this table, the 0.5000 has been added to the table values.

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.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	.9988	.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