

Statistics and Data Analysis

Professor William Greene

Phone: 212.998.0876

Office: KMC 7-78

Home page: www.stern.nyu.edu/~wgreene

Email: wgreene@stern.nyu.edu

Course web page: www.stern.nyu.edu/~wgreene/Statistics/Outline.htm

Fall, 2008 Midterm Examination Solutions

Instructions.

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 100 points in this test. The 8 questions provide for 95 points in total.

For the remaining 5 points, give a correct answer to the following question on your exam bluebook: **What is your name?**

90 minutes (1:30 – 3:00) is allotted for this test (on Thursday, October 30, 2008).

For question 1 in part I, since you had this question in advance, if you composed and printed out an answer to this question at home, you may submit that answer on the sheet that you brought with you. Be sure to put your name on that separate sheet.

NOTE: In all cases where a numerical answer is obtained, you can just show how to obtain the answer, without actually carrying it out. For example, if the answer is $12345/0.43984$, you may just leave your answer in that form.

Part I. Statistical Methodology

[10] 1. The following quote was taken on October 16 from

<http://www.cnn.com/2008/POLITICS/10/15/debate.poll/index.html>

Fifty-eight percent of debate watchers questioned in a CNN/Opinion Research Corp. poll said Democratic candidate Obama did the best job in the debate, with 31 percent saying Republican Sen. John McCain performed best.

The post-debate polls do not reflect the views of all Americans. They only represent the views of people who watched the debates. The CNN/Opinion Research Corp. poll was conducted by telephone Wednesday night, with 620 adult Americans who watched the debate questioned. The survey's sampling error is plus or minus 4 percentage points.

Answer each of these with a short description. (No more than a couple sentences.)

(a) How can CNN claim to provide useful information about millions of Americans based on a sample of only 620 observations?

As long as the sample is drawn randomly, a sample of 620 observations will provide a reasonably good guess of the true population proportion. The law of large numbers – 620 is pretty large – allows us to use random samples to learn about populations. There remains sampling variability, so we also compute a margin of error. In a sample of 620, the margin of error is not very large.

(b) How is the margin of error computed and what does it mean?

For an estimate of a population proportion, the margin of error is computed as ± 1.96 standard errors where the standard error is computed as $\sqrt{p(1-p)/n}$ where p is the sample estimate (the 0.58 above, for example). The margin of error is used to state that the proportion of those polled (such as 58%) stated something, and we believe that the true proportion is between this proportion plus and minus the margin of error. The value 1.96 is chosen so that we are 95% confident in this statement. With a wider interval, we would be more confident. For example, if we used 2.54 standard errors instead, we could be 99% certain. If we used the entire range from 0 to 1, we could be 100% certain, but this would not be very informative.

(c) The quote and the statistical procedure behind it rely on two fundamental results, the *law of large numbers* and the *central limit theorem*. How?

The *law of large numbers* is used to justify using a random sample to learn about the population. The law states that a sample statistic computed from a random sample will resemble the population counterpart, and the resemblance will be closer the larger is the sample. The sample of 620 is deemed to be large enough to provide a usable estimate.

The *central limit theorem* states that the sampling distribution of a sample mean will be approximately normal. We used this result to decide how many standard errors to use to construct our margin of error. That is, we chose 1.96 standard errors to form the margin of error because the mean plus and 1.96 standard deviations encompasses 95% of the probability in the normal distribution.

[5] 2. The following was taken from

<http://www.msnbc.msn.com/id/27339545/>
An msnbc.com guide to presidential polls
Why results, samples and methodology vary from survey to survey

WASHINGTON - A poll is a small sample of some larger number, an estimate of something about that larger number. For instance, what percentage of people reports that they will cast their ballots for a particular candidate in an election? A sample reflects the larger number from which it is drawn. Let's say you had a perfectly mixed barrel of 1,000 tennis balls, of which 700 are white and 300 orange. You do your sample by scooping up just 50 of those tennis balls. If your barrel was perfectly mixed, you wouldn't need to count all 1,000 tennis balls — your sample would tell you that 30 percent of the balls were orange.

Is the underlined sentence true or false? Explain your answer.

The sentence is false. My sample of 50 will contain anywhere from 0 to 50 orange balls no matter how well mixed the barrel is. This is sampling variability. Because of the points discussed in question 1, we would expect that a random sample of 50 balls would contain approximately 30% (fifteen) orange balls, but there is no assurance that it would contain exactly 15 orange balls. The end result is that my sample will suggest some value near 30% but it will not be exact.

Part II. Regression Models

[20] 3. This Minitab output shows the result of regressing *Profits/Sales* on *R&D/Sales* for a sample of industries.

Regression Analysis: Profit_S versus R&D_S

The regression equation is
Profit_S = 68.0 + 0.676 R&D_S

Predictor	Coef	SE Coef	T	P
Constant	68.04	16.46	4.13	0.001
R&D_S	0.6763	0.4480	1.51	0.151

S = 42.6753 R-Sq = 12.5% R-Sq(adj) = 7.0%

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	4149	4149	2.28	0.151
Residual Error	16	29139	1821		
Total	17	33288			

You may use the response “no way to tell” if you believe that the information cannot be determined from the results.

- (a) What is the value of s_e , the standard error of the regression?
42.6753
- (b) What is the standard deviation of *Profits/Sales*? (Just show the computation)
 $\text{sqr}(33288/17) = 44.2506$
- (c) How many data points were used in this regression?
18. The Total DF of 17 is n-1.
- (d) What is the value of R^2 for this regression? Show how it is computed.
 $0.125 = \text{Regression SS}/\text{Total SS} = 4149/33288 = 0.1246395$
- (e) What is the average *R&D/Sales* in this data set?
That information is not in the table. It can't be determined.
- (f) What is the correlation coefficient r between *Profits/Sales* and *R&D/Sales*. Is it positive or negative? Explain.
 $r = \sqrt{R^2} = +0.35304$. It is positive, because the slope of the regression (the coefficient on R&D_S) is positive.
- (g) The slope of the regression line is +0.676. How would you interpret this coefficient?
**It estimates the impact of an additional dollar of R&D/Sales on Profit/Sales.
An additional \$1 of R&D is estimated to be associated with an additional \$0.676 profit.**
- (h) Does this estimated regression provide convincing evidence of a statistically significant relationship between *Profits/Sales* and *R&D/Sales*? Explain.
Not really. The R^2 is not very high, but what is telling is the F statistic. 2.28 is well under the benchmark value of 4.0 that we are using for a regression of one variable on another. The “t statistic” is only 1.51, well under 2. (Of course, $F = t^2$ so this is the same information.)

[20] 4. The figure below plots $P = \text{price}$ (the y variable) vs. size (the x variable) in square feet for 50 houses. I will compute a linear least squares regression of P on size :

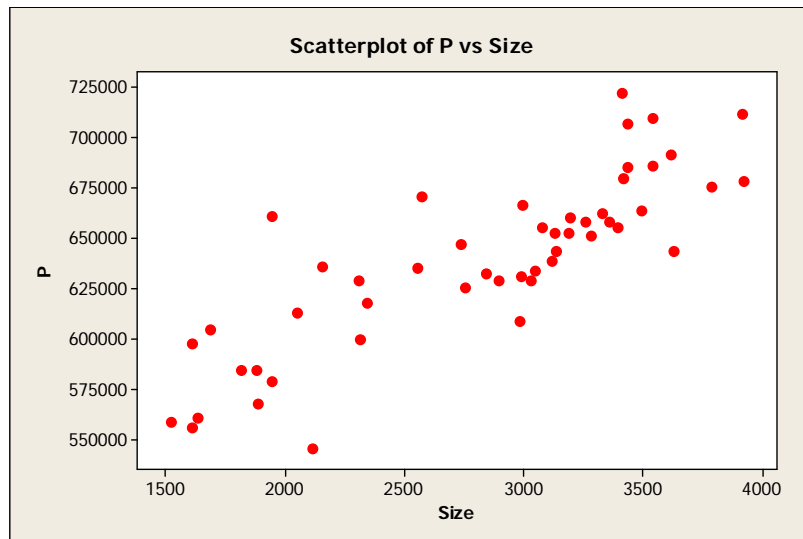
- (a) Do you think that the slope in the regression will be negative or positive? What would this mean in terms of your interpretation of the data?

Based on the way the points fall in the figure, ranging from lower left to upper right, it is clear that a regression line will have a positive slope.

- (b) Which of these values would be a good guess of the R^2 in this regression?

(i) 0.010, (ii) 0.200, (iii) 0.700, (iv) 0.999, (v) there is no way to know.

There is a decently good arrangement of the points along a line. .01 is much too weak. 0.999 is obviously too good. Maybe it could be 0.200 or 0.700. As noted, the fit to a line looks like it will be pretty good. I'll choose 0.700. (There is a way to know, based on the examples we looked at in class.)



(c) The actual regression results and some descriptive statistics are shown below. Fill in the missing values (A), (B), (C), (D) and (E). (Note (A) and (B) appear in two places.)

(B) Regression slope $b = \text{Covariance}(\text{Size}, P) / \text{Variance}(\text{Size})$
 $= 24850515 / 2815.2^2$
 $= 3.13557$

(A) Constant term $a = \text{Mean of } P - b \times \text{Mean of Size}$
 $= 638742 - 3.13557(2812.5)$
 $= 629923.2$

(C) $S = \sqrt{\text{Residual SS} / \text{DF}}$
 $= \sqrt{26082710113 / 48}$
 $= 7371.5$

(D) $R^2 = \text{Regression SS} / \text{Total SS}$
 $= 62579058669 / 88661768782$
 $= 0.7058$

(E) $F = (n-2)R^2 / (1-R^2)$
 $= 48 (.7058) / (1 - .7058)$
 $= 115.1543$

F is also equal to T^2 where $T = 10.73$.

Regression Analysis: P versus Size

The regression equation is

$$P = \text{(A)} + \text{(B)} \text{ Size}$$

Predictor	Coef	SE Coef	T	P
Constant	<u>(A)</u>	13879	35.60	0.000
Size	<u>(B)</u>	4.789	10.73	0.000

$$S = \text{(C)} \quad R\text{-Sq} = \text{(D)} \quad R\text{-Sq}(\text{adj}) = 70.0\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	62579058669	62579058669	<u>(E)</u>	0.000
Residual Error	48	26082710113	543389794		
Total	49	88661768782			

Descriptive Statistics: Size, P

Variable	N	Mean	StDev
Size	50	2815.2	695.4
P	50	638742	42537

$\text{Cov}(\text{Size}, P) = 24,850,515$

Part III. Probability

[10] 5. A study described at <http://www.medscape.com/viewarticle/518415> 3 states that in a (large) sample of individuals, the mean body mass index (BMI) is 26.3 and the standard deviation is 3.9. (BMI is measured as weight in kilograms divided by the square of height in meters.) Assume that BMI is normally distributed among people. The World Health Organization defines “Overweight” as BMI greater than 25 and “Obese” as BMI greater than 30.

(a) If the sample mean and standard deviation are actually the true population mean and standard deviation (or, really precise estimates of them), what proportion of the population is overweight?

$$\begin{aligned}\text{Assuming normality with mean 26.3 and standard deviation 3.9,} \\ \text{Prob[BMI} > 25] &= \text{Prob}[(\text{BMI} - 26.3)/3.9 > (25 - 26.3)/3.9] \\ &= \text{Prob}[z > -.33] \\ &= \text{Prob}[z < .33] \\ &= 0.6293 \text{ (plus a bit of rounding error)}\end{aligned}$$

(b) What proportion of the population is overweight but not obese?

$$\begin{aligned}\text{Prob}[25 < \text{BMI} < 30] &= \text{Prob}[-.33 < z < (30 - 26.3)/3.9] \\ &= \text{Prob}[-.33 < z < .95] \\ &= .8023 - \text{Prob}[z < -.33] \\ &= .8023 - (1 - .6293) \\ &= .4316\end{aligned}$$

(c) If a sample of 100 individuals is drawn at random from the population, what is the probability that the average of their BMI values will be in the “normal” (or underweight) range, less than or equal to 25?

$$\begin{aligned}\text{The mean of a sample of 100 would have a standard error of } 3.9/\text{sqr}(n) \text{ or } .39. \\ \text{Prob[mean} \leq 25] &= \text{Prob}[(\text{mean} - 26.3)/.39 < (25 - 26.3)/.39] \\ &= \text{Prob}[z < -3.33] \\ &= 1 - \text{Prob}[z < 3.33] \\ &= 1 - .9996 \\ &= .0004\end{aligned}$$

(d) There are 2.2 pounds per kilogram and 3.28 feet per meter. If BMI were measured in English units, that is pounds per square foot, what would be the mean and standard deviation of the sample? Explain.

The random variable BMI has mean 26.3 and standard deviation 4.9. If we measured in Pounds and square feet, the new random variable would be $(2.2 / 3.28^2) = .204$ times BMI. The mean and standard deviation would both be multiplied by 0.204.

If you need it, a table for the normal distribution appears at the end of this exam booklet.

[10] 6. An insurance company has two kinds of customers, high risk and low risk. (Though it knows it has two types of customers, it does not know which type any particular customer is.) It does know the following: 10% of its customers are high risk: $P(H)=.10$, 90% are low risk: $P(L)=.90$. The proportion of high risk customers that will make a claim is .05: $P(C|H)=.05$; the proportion of low risk customers that will make a claim is .01: $P(C|L)=.01$.

a. what proportion of customers who make claims come from the high risk group; that is, what is $P(H|C)$?

Given:

$$P(H) = 0.1, P(L) = 0.9$$

$$P(C|H) = 0.05, P(C|L) = 0.01$$

To find $P(H|C)$ use Bayes theorem:

$$P(H|C) = P(H \text{ and } C)/P(C) = [P(C|H)P(H)] / P(C) \text{ The numerator is } 0.05(0.10) = .005$$

The denominator is

$$P(C) = P(C|H)P(H) + P(C|L)P(L) = .05(.1) + .01(.9) = .005+.009 = .014.$$

$$\text{So, } P(H|C) = .005/.014 = 0.357$$

b. What proportion of customers will file claims; that is, what is $P(C)$?

$$\text{We found this above. } P(C) = .014$$

c. Suppose the firm has 10,000 customers. How many claims should it expect to be made in each year (assuming that each customer who makes a claim makes exactly one claim).?

$$\text{This would be } 10,000 \text{ times } .014, \text{ or } 140.$$

Part IV. Descriptive Statistics

[10] 7. The figure below shows a histogram of the 59 midterm grades in a different course that I once taught.. The sample average grade for the full sample was 84.6.

(a) Approximately what was the median grade for the class? Explain.

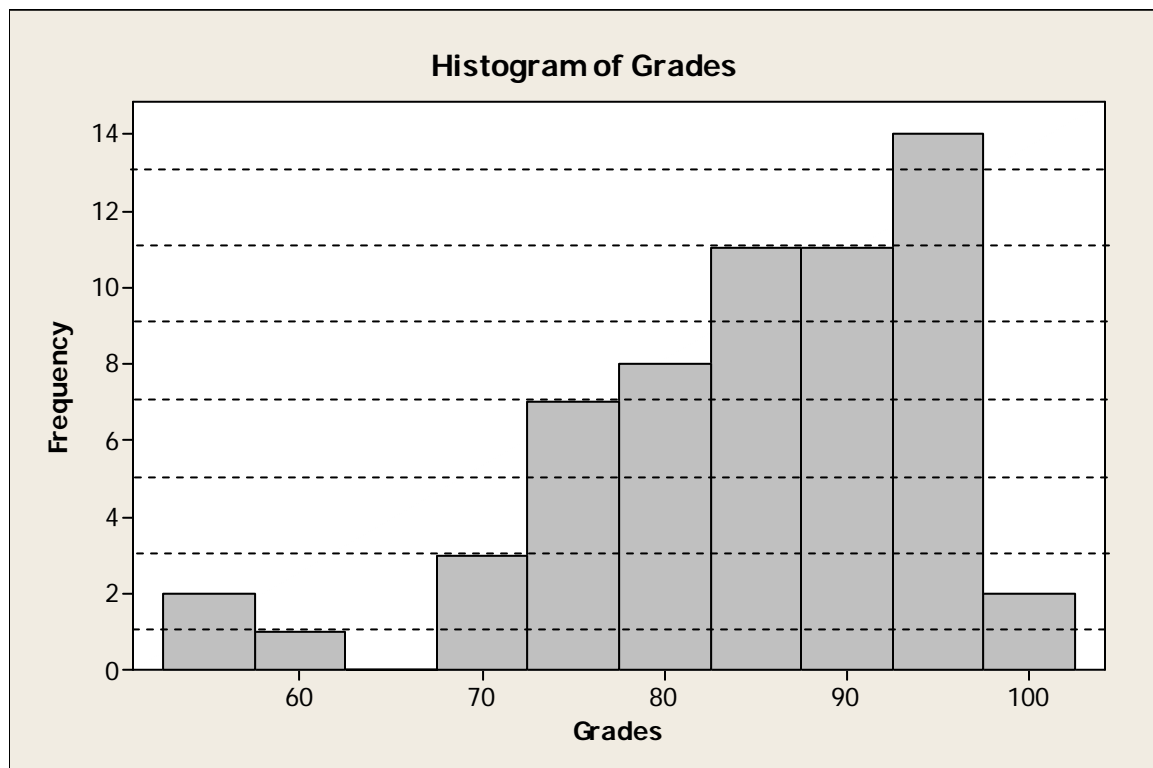
There are 59 grades. Counting the number of students in the blocks, we are looking for the 29th. This is about 88 or so. We know the median is greater than 84.6. (Next part.)

(b) Is the mean (sample average) grade less than, equal to, or greater than the median grade? Explain how you can answer this question without actually computing the mean, just by looking at the figure.

The data are skewed to the left, so the mean is less than the median.

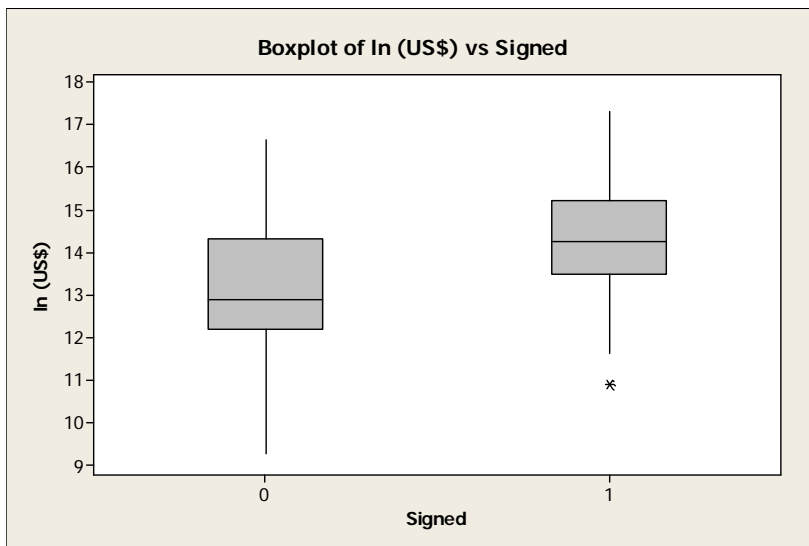
(c) The standard deviation of the midterm grades was approximately which of these four? (i) About 100.0, (ii) About 10.0, (iii) About 1.0 (iv) It is impossible to tell. Explain how you arrived at the answer.

The range of the data will be about 6 standard deviations around the mean (using our empirical rule) Both 100 and 1.0 are obviously wrong, leaving 10.0. (It is possible to tell.)



[10] 8. The box plots in the figure below describe the sale prices of the Monet paintings that we discussed in class on Tuesday, October 28th. The sample of paintings included some that were signed (identified as “Signed = 1” in the figure) and some that were not signed (marked “Signed = 0”). Based on this figure, mark the following statements as True, False, or Neither (meaning it is not possible to tell). (No explanation needed for the answers.)

- (a) The average price of a signed painting is higher than that of an unsigned one.
Neither
- (b) The median price of a signed painting is higher than that of an unsigned one.
True
- (c) The interquartile range is larger for signed paintings.
False
- (d) Minitab thinks one of the signed paintings is underpriced.
False. Minitab has no idea. It does not think.
- (e) The highest priced signed painting is priced higher than the highest priced unsigned one.
True (apparently, since no outliers are shown in the upper range)
- (f) The number of signed paintings is larger than the number of unsigned ones.
Neither – sample sizes are not given
- (g) For the unsigned paintings, about half the prices are above the median for unsigned paintings while for the signed paintings, about half the prices are below the median for signed paintings.
True – this is the definition of the median.



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>	<i>.00</i>	<i>.01</i>	<i>.02</i>	<i>.03</i>	<i>.04</i>	<i>.05</i>	<i>.06</i>	<i>.07</i>	<i>.08</i>	<i>.09</i>
.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	.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