NUI Galway
OÉ Gaillimh

Semester I Examinations 2019/2020

Exam Codes	1BDS1, 1BY1, 1EM1, 1EV1, 2BA1, 2BCM1, 2BCT1, 2BME1 2BMU1, 2BPT1, 2BS1, 2EH1, 2EHS1, 2MR1, 3BS9
Exams	First Science, Second Arts, Second Science and Third Science
Module	**Statistics for Data Science 1**
Module Codes	**ST2001**
External Examiner	Prof. S. Wilson
Internal Examiner(s)	Dr. D. Roshan Prof. John Newell
Instructions:	**Answer all FIVE questions.**
Duration	2 hours
No. of Pages	14 pages, including this one
School	School of Mathematics, Statistics and Applied Mathematics
Release to Library:	Yes ☑ No □
Release in Exam Venue	Yes ☑ No □

Requirements:

Statistical Tables/ Log Tables	Log Tables are available if requested in the exam
Other Materials Allowed	A calculator is allowed (non-programmable and not capable of storing text).

Question 1

(a). A company manufactures cardboard containers for packaging frozen fruit juices. The manufacturing process forms the containers by joining cardboard stock to a metal base. The container can be inspected to determine if there might be leaks along the cardboard seam or where the metal base is attached. It is known from experience that **18%** of containers produced by this process are defective. A random sample of 50 containers was selected from the production process, and it was found that **12%** of the selected containers were defective. Because this seemed to indicate an improvement, another sample was selected of the same size, and of those **21%** of the containers were defective. Identify which of the numbers in bold are parameters and which are statistics?

[3 marks]

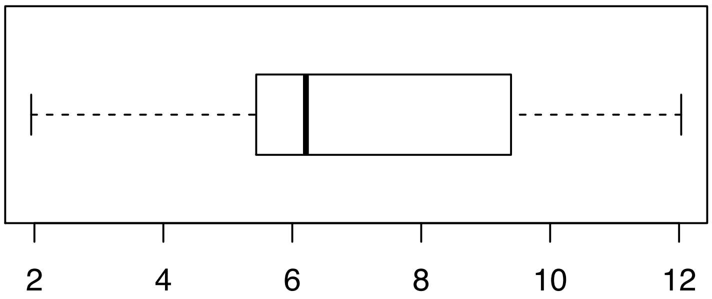
(b). A doctor is involved in a trial comparing a new medication to the current standard medication. Before deciding which medication a patient is allocated, the doctor examines the patient. Do you have any concerns with the treatment allocation for this trial? If so, explain these concerns.

[3 marks]

(c). An opinion poll is to be given to a sample of 90 members of a local gym. The members are first divided into men and women, and then a simple random sample of 45 men and a separate simple random sample of 45 women are taken. What sampling scheme was used here?

[3 marks]

(d). The following boxplot shows distribution of a data which has a median of 6.21.



What do you expect to be the relationship between the mean and median for this data.

[2 marks]

(e). If $P(A) = 0.3$, $P(B) = 0.4$ and $P(A|B) = 0$. Justify whether or not

(i) A and B are mutually exclusive?

(ii) A and B are independent?

[4 marks]

(f). A student has to sell 2 textbooks from a collection of 6 maths, 7 stats, and 4 physics books. How many choices are possible if both textbooks are to be on the same subject?

[4 marks]

(g). Mark visits campus every evening. However, some evenings the parking garage is full, often due to college events. There are academic events on 35% of evenings, sporting events on 20% of evenings, and no events on 45% of evenings. When there is an academic event, the garage fills up 25% of the time, and it fills up 70% of evenings when there are sporting events. On evenings when there are no events, it only fills up 5% of the time. For a randomly selected evening, if Mark comes to campus,

(i) What is the probability that the garage is full?

(ii) Given the garage is full, what is the probability that there is a sporting event?

[6 marks]

Total: 25 marks

Question 2

Note: relevant formulas and probabilities of the Binomial and Poisson distributions are given at the back of this paper.

(a). The return in € of a gambling game is given by a random variable X, where X has the following probability distribution

x	-4	-2	0	2	4
$P(X=x)$	0.2	0.1	0.3	0.3	??

(i) Calculate the missing probability, $P(X=4)$.

(ii) Find the probability that the return on the game is at least €2.

(iii) Find the expected return, $E(X)$.

(iv) Find the variance of the return, $V(X)$.

[7 marks]

(b). Sarah sells a magazine which is produced in order to raise money for homeless people. The probability of making a sale is 0.09 and is assumed to be independent for each person she approaches. Given that she approaches 40 people, find the probability that she will make:

(i) exactly 4 sales.

(ii) more than 5 sales.

[4 marks]

(c). The number of accidents occurring on a motorway each day is Poisson distributed with a mean of 3 accidents per day. Find the probability that:

(i) there are at least 2 accidents tomorrow.

(ii) there are exactly 10 accidents in the next week.

[4 marks]

Total: 15 marks

Question 3

(a). What is the minimum number of Galwegians that should be sampled at random so that with probability (at least) 0.95 the sample proportion of smokers will not differ from the unknown population proportion of smokers by more than ± 0.04?

[3 marks]

Note: `qnorm(0.95)=1.645` and `qnorm(0.975)=1.96`

(b). It has been reported that the average hotel check-in time from arrival to the hotel is 12 minutes. Jack has just arrived at the hotel for check-in. Assuming check-in times follow a Normal distribution with mean of 12 and a standard deviation of 2.0 minutes:

(i) What is the probability that Jack's check-in time will be greater than 14 minutes?

(ii) What is the probability that Jack's check-in time will be between 10 and 15 minutes?

(iii) Determine, to two decimal places, the check-in time exceeded by 67% of guests in the hotel.

Note: relevant formulas and probabilities of the Normal distributions are given at the back of this paper.

[8 marks]

(c). Assume $X_1, X_2, \ldots, X_{100}$ to be 100 independent and identically distributed random variables with a mean 10 (i.e. $\mu = 10$) and variance of 49 (i.e. $\sigma^2 = 49$). According to the Central Limit Theorem, write down the sampling distribution of the sample means, i.e. $\bar{X}$.

[2 marks]

(d). If a random variable X, follows a Binomial distribution with $n = 500$ and $p = 0.1$, i.e. $X \sim B(500, 0.1)$. Write down the Normal distribution which would be a good approximation of the distribution of X?

[2 marks]

Total: 15 marks

Question 4

(a). A researcher conducted a large sample two-sided test of the null hypothesis that $\mu = 50$. She reports a p-value of 0.04. Which one of the following is correct?

- **(i)** The null hypothesis is rejected at $\alpha = 0.01$.
- **(ii)** The null hypothesis is not rejected at $\alpha = 0.05$.
- **(iii)** The 95% confidence interval for μ would contain 50.
- **(iv)** The 99% confidence interval for μ would contain 50.

[2 marks]

(b). Suppose you have obtained a confidence interval for a population mean, but wish to obtain a greater degree of precision. Which of the following would result in a narrower confidence interval:

- **(i)** increasing the confidence level while keeping the sample size fixed.
- **(ii)** decreasing the confidence level while keeping the sample size fixed.
- **(iii)** decreasing the sample size while keeping the confidence level fixed.

[2 marks]

(c). In a particular cola canning factory, the mean fill of cans is set at 300 ml but there is concern that the population mean fill may not in fact be 300ml. A random sample of 100 cans showed a sample mean of 299.75 and a sample standard deviation of 1.09.

The key question of interest is whether the population mean fill is 300 ml.
Using the output below answer the following questions.

- **(i)** State the appropriate null and alternative hypotheses.

[2 marks]

- **(ii)** Define Type I and Type II errors and discuss the implication of making these errors in this application.

[3 marks]

- **(iii)** Interpret carefully the relevant 95% Confidence Interval for the mean fill in the population of interest.

[4 marks]

- **(iv)** Use the relevant interval estimate and p-value to decide whether you think there is sufficient evidence in the samples provided to claim that the population mean fill is 300 ml?

[5 marks]

(**v**) What are the assumptions underlying the one sample t-test presented?

[2 marks]

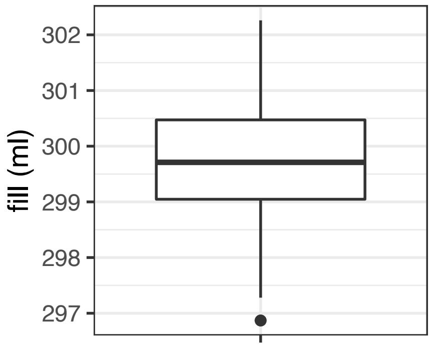
(**vi**) Explain why or why not the assumptions look justified based on the output provided.

[2 marks]

(**vii**) Explain how the bootstrap technique can be used to obtain a 95% confidence interval for the mean fill in the population of interest?

[3 marks]

```
# boxplot of the fill of cans
ggplot(data=cans_dat,aes(y=fill,x=""))+
  geom_boxplot()+
  theme_bw()+
  ylab("fill (ml)")+
  xlab("")
```



```
# One sample t-test output.
t.test(cans_dat$fill,mu=300,alternative = "two.sided",conf.level = 0.95)
```

```

    One Sample t-test

data:  cans_dat$fill
t = -2.2952, df = 99, p-value = 0.02383
alternative hypothesis: true mean is not equal to 300
95 percent confidence interval:
 299.5314 299.9660
sample estimates:
mean of x 
 299.7487 
```

Total: 25 marks

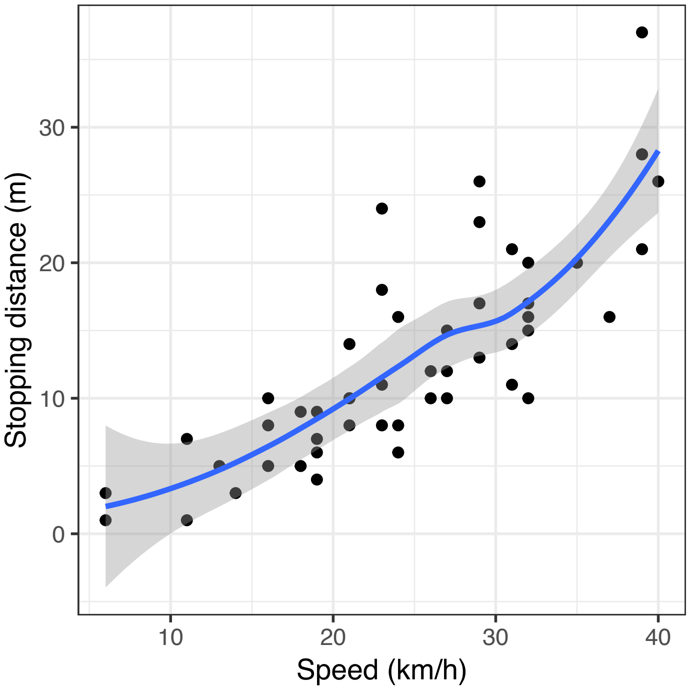
Question 5

Data were collected on the speed of cars (km/h) and the distances taken to stop (m) for a sample of 50 cars. Data from the first six cars are listed below. The aim of this study is to determine whether the speed of a car (X) can be used to predict their stopping distance (Y).

```
Cars %>% head()
```

```
  speed distance
1     6        1
2     6        3
3    11        1
4    11        7
5    13        5
6    14        3
```

(a). Using the scatterplot (with a smoother) and the correlation coefficient provided below answer the following three questions.



```
cor(Cars$speed,Cars$distance) %>% round(2)
```

```
[1] 0.81
```

(i) In one word ("yes" or "no") would you consider the above data to contain outliers?

[1 mark]

(ii) Interpret the correlation coefficient between the speed of a car and its stopping distance.

[3 marks]

(iii) What does the smoother suggest regarding the suitability of a simple linear regression model in this context?

[2 marks]

(b). Assume that a model of the form $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ (for $i = 1, ..., 50$), together with the usual assumptions is appropriate in this context. The output below provides the fit of a regression analysis carried out on these data. Answer the following questions.

```
# Simple Linear Regression model
lm(distance ~ speed, data=Cars)
```

```
Call:
lm(formula = distance ~ speed, data = Cars)

Coefficients:
(Intercept)        speed
    -5.3168       0.7424
```

(i) Explain briefly what the term ϵ_i represents in a simple linear regression model.

[2 marks]

(ii) Write down the equation of the line of best fit and also provide an interpretation of the slope and intercept of the corresponding line of best fit. Justify if the intercept is or is not a useful estimate in this analysis.

[6 marks]

(iii) Predict the stopping distance for a car with a speed of 13 km/h.

[2 marks]

(iv) Why is the result in part (iii) different from 5, the value in the table of data presented above i.e. the stopping distance when the speed is 13 km/h?

[2 marks]

(v) The predicted stopping distance for a car driving with a speed of 100 km/h is found to be 68.92m. Explain briefly if you have a concern regarding this prediction.

[2 marks]

Total: 20 marks

Useful formulae and Probabilities

Binomial distribution

The Binomial distribution has a probability parameter p and an associated sample size n

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $X = 0, 1, 2, \ldots, n$ with

$$E(X) = np$$

$$Var(X) = np(1-p)$$

and

```r
dbinom(0,size=40,prob=0.09) %>% round(4)
```

```
[1] 0.023
```

```r
dbinom(1,size=40,prob=0.09) %>% round(4)
```

```
[1] 0.091
```

```r
dbinom(2,size=40,prob=0.09) %>% round(4)
```

```
[1] 0.1754
```

```r
dbinom(3,size=40,prob=0.09) %>% round(4)
```

```
[1] 0.2198
```

```r
dbinom(4,size=40,prob=0.09) %>% round(4)
```

```
[1] 0.2011
```

```r
dbinom(5,size=40,prob=0.09) %>% round(4)
```

```
[1] 0.1432
```

p.t.o.

Poisson distribution

The Poisson(λ) distribution has a single parameter λ which is the mean rate over a given time interval

$$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where $X = 0, 1, 2, \ldots$ with

$$E(X) = \lambda$$

$$Var(X) = \lambda$$

and

```r
dpois(0,lambda = 3) %>% round(4)
```

```
[1] 0.0498
```

```r
dpois(1,lambda = 3) %>% round(4)
```

```
[1] 0.1494
```

```r
dpois(2,lambda = 3) %>% round(4)
```

```
[1] 0.224
```

```r
dpois(10,lambda = 3) %>% round(4)
```

```
[1] 0.0008
```

```r
dpois(5,lambda = 21) %>% round(4)
```

```
[1] 0
```

```r
dpois(10,lambda = 21) %>% round(4)
```

```
[1] 0.0035
```

```r
dpois(15,lambda = 21) %>% round(4)
```

```
[1] 0.0395
```

```r
dpois(20,lambda = 21) %>% round(4)
```

```
[1] 0.0867
```

Normal distribution

The Normal distribution has two parameters, the mean μ is a measure of location and the standard deviation σ is a measure of spread

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

where $X \in \mathbb{R}$

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

and

```
pnorm(1, mean = 0, sd = 1, lower.tail = T) %>% round(4)
```

```
[1] 0.8413
```

```
pnorm(-1, mean = 0, sd = 1, lower.tail = T) %>% round(4)
```

```
[1] 0.1587
```

```
pnorm(1.5, mean = 0, sd = 1, lower.tail = T) %>% round(4)
```

```
[1] 0.9332
```

```
pnorm(-1.5, mean = 0, sd = 1, lower.tail = T) %>% round(4)
```

```
[1] 0.0668
```

```
qnorm(0.33, mean = 0, sd = 1, lower.tail = T) %>% round(4)
```

```
[1] -0.4399
```

```
qnorm(0.95, mean = 0, sd = 1, lower.tail = T) %>% round(4)
```

```
[1] 1.6449
```

```
qnorm(0.975, mean = 0, sd = 1, lower.tail = T) %>% round(4)
```

```
[1] 1.96
```

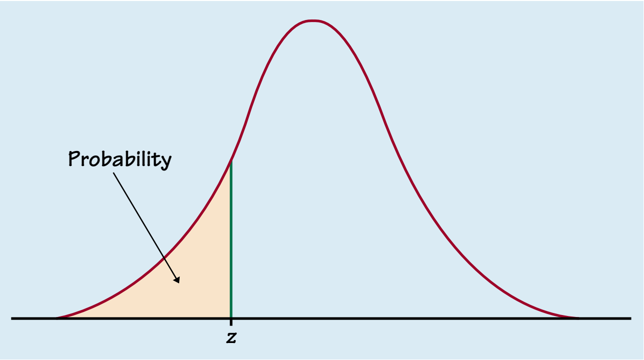


Table entry for *z* is the probability lying below *z*.

TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Normal Table 2

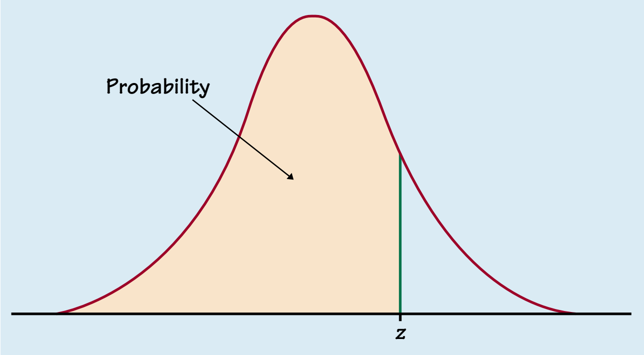


Table entry for z is the probability lying below z.

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