

古谷数学教室第 20 回

統計的な推測

2024 年 9 月 8 日

1 基礎事項

1.1 二項分布

同じ条件のもとでの試行の繰り返しを反復試行という。反復試行において、ある事象がおける回数 X の確率分布について考える。

1 回の試行で事象 A が起こる確率を p とする。この試行を n 回行う反復試行において、事象 A の起こる回数を X とすると、 X は確率変数で、その確率分布は表 1 のようになる。

表 1 n 回の反復試行における確率分布表

X	0	1	...	n	計
P	${}_nC_0(1-p)^n$	${}_nC_1p(1-p)^{n-1}$...	${}_nC_np^n$	1

表 1 で与えられる確率分布を**二項分布**といい、 $B(n, p)$ で表す¹⁾。

確率変数 X が二項分布 $B(n, p)$ に従うとき、期待値 $E(X)$ 、分散 $V(X)$ 、標準偏差 $\sigma(X)$ は、一般に、次のことが成り立つ：

二項分布に従う確率変数の期待値と分散

$$\begin{aligned}E(X) &= np, \\V(X) &= np(1-p), \\ \sigma(X) &= \sqrt{np(1-p)}.\end{aligned}$$

1) $B(n, p)$ の B は、「二項分布」を意味する英語 binomial distribution の頭文字である。

1.2 正規分布

表 2 は、あるクラス 40 人の身長を調べた結果で、度数分布表において、各階級の相対度数をまとめたものである。この 40 人の中から 1 人を選ぶとき、その生徒の身長を X とする。たとえば、 X が 174cm 以上 178cm 未満の階級に属する確率は相対度数 0.175 であると考えてよい。このように考えると、 X が各階級に属する確率は、各階級の相対度数に一致する。したがって、 X は各階級に対して属する確率の定まる確率変数と考えられる。

表 2 あるクラス 40 人の身長の度数分布表

身長の階級 (cm)	度数	相対度数
158 以上 162 未満	2	0.050
162 以上 166 未満	4	0.100
166 以上 170 未満	10	0.250
170 以上 174 未満	16	0.400
174 以上 178 未満	7	0.175
178 以上 182 未満	1	0.025
計	40	1.000

図 1 は、表 2 で示した相対度数を柱状のグラフで示したもの（ヒストグラム）である。たとえば、図 1 の灰色部分が $162 \leq X < 170$ となる確率を表す。

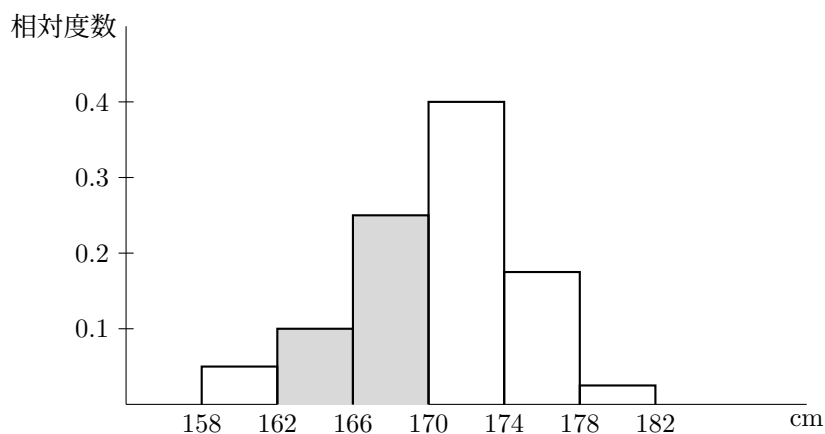


図 1 あるクラス 40 人の身長のヒストグラム

データの総数を増やし、階級の幅を細かく分けて同様なヒストグラムをつくっていくと、ヒストグラムの形はある曲線に近づくと考えられる。

一般に、連続した値をとる確率変数 X を**連続型確率変数**という。連続型確率変数 X の確率分布を考える場合は、 X に 1 つの曲線 $y = f(x)$ を対応させ、 $a \leq X \leq b$ となる確率

$$P(a \leq X \leq b)$$

が、図の灰色部分の面積で表されるようにする。この曲線 $y = f(x)$ を、 X の**分布曲線**といい、関数 $f(x)$ を**確率密度関数**という。

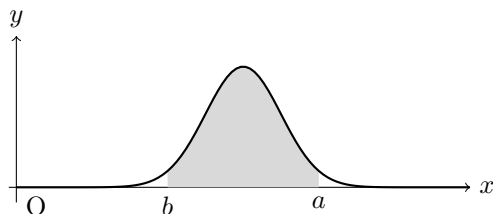


図2 分布関数

確率密度関数 $f(x)$ は、次のような性質をもつ：

$$\text{常に、} f(x) \geq 0 \text{ で } P(a \leq X \leq b) = \int_a^b f(x)dx,$$

$$X \text{ のとる値の範囲が } \alpha \leq X \leq \beta \text{ のとき } \int_{\alpha}^{\beta} f(x)dx = 1.$$

m を実数、 σ を正の実数とする。このとき、関数

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

は、連続型確率変数 X の確率密度関数となることが知られている。このとき、 X は**正規分布**に従うといい、 $N(m, \sigma^2)$ で表す。ここで、 e は無理数の定数で、次の式で定義される：

$$e = \lim_{k \rightarrow 0} (1 + k)^{\frac{1}{k}}.$$

曲線 $y = f(x)$ を**正規分布曲線**という。

確率変数 X が正規分布 $N(m, \sigma^2)$ に従うとき、確率について

$$P(m - \sigma \leq X \leq m + \sigma) = 0.6827,$$

$$P(m - 2\sigma \leq X \leq m + 2\sigma) = 0.9545,$$

$$P(m - 3\sigma \leq X \leq m + 3\sigma) = 0.9973.$$

また、確率変数 X が正規分布 $N(m, \sigma^2)$ に従うとき、期待値 $E(X)$ 、標準偏差 $\sigma(X)$ について、次のことが知られている：

正規分布の期待値、標準偏差

$$E(X) = m,$$

$$\sigma(X) = \sigma.$$

確率変数 X が正規分布 $N(m, \sigma^2)$ に従うとき、 X の分布曲線 $y = f(x)$ は、次のような性質をもつ：

1. 直線 $x = m$ に関して対称であり、 y は $x = m$ で最大値をとる。
2. x 軸を漸近線とし、 x 軸と分布曲線との面積は 1 である。
3. 標準偏差 σ が大きくなると曲線の山は低くなって、横に広がる。 σ が小さくなると曲線の山は高くなって、直線 $x = m$ の周りに集まる。

確率変数 X が正規分布 $N(m, \sigma^2)$ に従うとき、 X の 1 次式 $aX + b$ は正規分布 $N(am + b, a^2\sigma^2)$ に従う確率変数であることが知られている。とくに、

正規分布と標準正規分布

$$Z = \frac{X - m}{\sigma}$$

とおくと、確率変数 Z は正規分布 $N(0, 1)$ に従い、 Z の確率密度関数は、

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

となる。

正規分布 $N(0, 1)$ を、**標準正規分布**という。

標準正規分布 $N(0, 1)$ に従う確率変数 Z に対し、確率 $P(0 \leq Z \leq u)$ をこのプリントでは $p(u)$ で表す。3 節の表 5 は、いろいろな u の値に対する $p(u)$ の値 (図 4 の青い部分の面積が $p(u)$) である。

表 5 によると、たとえば

$$P(0 \leq Z \leq 1.23) = p(1.23) = 0.3907$$

である²⁾。

u03	
⋮		↓	
1.2	→	0.3907	
⋮			

図 3 $p(1.23)$ のときの標準正規分布表の見方

2) $p(1.23)$ のときの見方に関しては、図 3 を参照。

また、次の等式が成り立つ：

$$P(-u \leq Z \leq 0) = P(0 \leq Z \leq u) = p(u),$$

$$P(Z \leq 0) = P(Z \geq 0) = 0.5.$$

二項分布と正規分布の関連について、ここで詳しく調べる。1回の試行で事象 A の起こる確率を p とするとき、この試行を n 回行う反復試行において、 A の起こる回数を X とすれば、 X は二項分布 $B(n, p)$ に従う確率変数である。 $X = r$ である確率 P_r は

$$r = 0, 1, \dots, n, \quad P_r = {}_n C_r p^r (1-p)^{n-r}$$

であり、 X の期待値 $E(X)$ 、標準偏差 $\sigma(X)$ は、次のようになる。

$$E(X) = np,$$

$$\sigma(X) = \sqrt{np(1-p)}.$$

横軸 r 、縦軸 P_r の二項分布のグラフの形は、 n が大きくなるにつれて正規分布曲線に近づくことが知られている。

一般に、次のことが成り立つ：

二項分布の正規分布による近似

1. 二項分布 $B(n, p)$ に従う確率変数 X は、 n が大きいとき、近似的に正規分布 $N(np, np(1-p))$ に従う。
2. 二項分布 $B(n, p)$ に従う確率変数 X に対し、 $Z = \frac{X - np}{\sqrt{np(1-p)}}$ は、 n が大きいとき、近似的に標準正規分布 $n(0, 1)$ に従う。

1.3 母集団と標本

統計的な調査において、対象全体からデータを集めて調べる方法を**全数調査**という。これに対して、調査の対象全体からその一部を抜き出して調べる**標本調査**という方法がある。

標本調査の場合、調査の対象全体を**母集団**という。母集団に属する個々の対象を**個体**といい、個体の総数を**母集団の大きさ**という。また、調査のため母集団から抜き出された個体の集合を**標本**といい、母集団から標本を抜き出すことを**抽出**という。標本に属する個体の総数を**標本の大きさ**という。

母集団の各個体を等しい確率で抽出する方法を**無作為抽出**といい³⁾、無作為抽出によって選ばれた標本を**無作為標本**という。

3) 無作為抽出では、乱数さいや乱数表などが使われる。

母集団から標本を抽出するのに、毎回もとにもどしながら個体を1個ずつ抽出する⁴⁾ことを復元抽出という。これに対して、個体をもとにもどさないで標本を抽出することを非復元抽出という。

統計的な調査の対象には、身長、時間、不良品などのように、特定の性質がある。これを特性といい、ある特性を表す数量を変量という。大きさ N の母集団において、変量 x のとりうる異なる値を

$$x_1, x_2, \dots, x_r$$

とし、それぞれの値をとる個体の個数を

$$f_1, f_2, \dots, f_r$$

とすると、この母集団における変量 x の度数分布表は、表3のようになる。

表3 母集団の度数分布表

x の値	度数
x_1	f_1
x_2	f_2
\vdots	\vdots
x_r	f_r
計	N

表3から1個の個体を無作為に抽出して、変量 x の値を X とするとき、 X は確率変数である。 X の確率分布表は表4のようになり、 P は母集団における変量 x の相対度数と一致する。この X の確率分布を母集団分布という。また、確率変数 X の期待値、標準偏差を、それぞれ母平均、母標準偏差といい、それぞれ m 、 σ で表す。この母平均 m 、母標準偏差 σ は、母集団における変量 x の平均値、標準偏差にそれぞれ一致する。

表4 母集団分布

X	x_1	x_2	\dots	x_r	計
P	$\frac{f_1}{N}$	$\frac{f_2}{N}$	\dots	$\frac{f_r}{N}$	1

1.4 標本平均の分布

母集団から大きさ n の無作為標本を抽出し、それらの変量 x の値を X_1, X_2, \dots, X_n とするとき、

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

を標本平均という。 n を固定すると、標本平均 \bar{X} は1つの確率変数になる。

4) たとえば、1から100までの番号札100枚の中から、大きさ5の標本を復元抽出するときの標本の総数は、 100^5 通りである。

標本平均 \bar{X} の確率分布と母集団分布の関係を調べる。母平均 m 、母標準偏差 σ の母集団から、大きさ n の無作為標本を復元抽出し、それらの変数 x の値を X_1, X_2, \dots, X_n とする。各 X_k は、どれも大きさ 1 の標本で、母集団分布に従う確率変数である。よって、

$$E(X_1) = E(X_2) = \dots = E(X_n) = m,$$

$$\sigma(X_1) = \sigma(X_2) = \dots = \sigma(X_n) = \sigma.$$

したがって、

$$E(\bar{X}) = m.$$

また、復元抽出の場合、 X_1, X_2, \dots, X_n は互いに独立な確率変数であるから、

$$V(\bar{X}) = \frac{\sigma^2}{n},$$

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

これまでのことをまとめると、次のことがいえる：

標本平均の期待値と標準偏差

母平均 m 、母標準偏差 σ の母集団から大きさ n の無作為標本を抽出するとき、その標本平均 \bar{X} の期待値 $E(\bar{X})$ と標準偏差 $\sigma(\bar{X})$ は

$$E(\bar{X}) = m,$$

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

非復元抽出の場合も、標本の大きさ n に比べて母集団の大きさが十分大きいときは、復元抽出と同様に扱ってよい。

標本平均 \bar{X} の分布について、次の性質があることが知られている：

標本平均の分布

母平均 m 、母標準偏差 σ の母集団から抽出された大きさ n の無作為標本について、標本平均 \bar{X} は、 n が大きいとき、近似的に正規分布 $N(m, \frac{\sigma^2}{n})$ に従うとみなすことができる。

正規分布 $N(m, \frac{\sigma^2}{n})$ の標準偏差は $\frac{\sigma}{\sqrt{n}}$ であるから、標本平均 \bar{X} に対して、 $Z = \frac{\bar{X}-m}{\sigma/\sqrt{n}}$ は、 n が大きいとき、近似的に標準正規分布 $N(0, 1)$ に従う。

一般に、母集団の中である特性 A をもつものの割合を、その特性 A の**母比率**という。また、抽出された標本の中で特性 A をもつものの割合を**標本比率**という。

特性 A の母比率が p である十分大きな母集団から、大きさ n の無作為標本を抽出し、それらに

対して、 X_1, X_2, \dots, X_n の値を次のように定める。

特性 A をもつとき $X_k = 1$ ($k = 1, 2, \dots, n$),

特性 A をもたないとき $X_k = 0$ ($k = 1, 2, \dots, n$).

このとき、 $T = X_1 + X_2 + \dots + X_n$ を考えると、 T は大きさ n の標本の中で特性 A をもつものの個数を表す確率変数である。また、標本平均 $\bar{X} = \frac{T}{n}$ は、特性 A の標本比率 R を表す。 T は二項分布 $B(n, p)$ に従うから、

$$E(R) = p,$$
$$V(R) = \frac{p(1-p)}{n},$$

である。 n が大きいとき、 T は近似的に正規分布 $N(np, np(1-p))$ に従うとみなすことができる。

一般に、標本の大きさ n を限りなく大きくしていくと、標本平均 \bar{X} の分布を近似する正規分布の標準偏差 $\frac{\sigma}{\sqrt{n}}$ は、限りなく 0 に近づき、 \bar{X} の分布は母平均 m の近くに限りなく集中する。すなわち、 \bar{X} が m に近い値をとる確率が 1 に近づく。

次の事実を、**大数の法則**という：

大数の法則

母平均 m の母集団から大きさ n の無作為標本を抽出するとき、 n が大きくなるに従って、その標本平均 \bar{X} はほとんど確実に母平均 m に近づく。

1.5 推定

標本平均を用い母平均を推測する方法を考える。

母平均 m 、母標準偏差 σ をもつ母集団から抽出された大きさ n の無作為標本の標本平均 \bar{X} に対して、確率変数

$$Z = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$$

は n が大きいとき、近似的に標準正規分布 $N(0, 1)$ に従う。

表 5 によると

$$P(|Z| \leq 1.96) = 0.95$$

であるから

$$P\left(|\bar{X} - m| \leq 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

すなわち、不等式

$$\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \quad (1)$$

の成り立つ確率は 0.95 である。

母平均の推定

式 (1) で示される範囲を、母平均 m に対する信頼度 95% の信頼区間といい、次のように表す：

$$\left[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \right].$$

母平均 m に対して信頼度 95% の信頼区間を求めることを、「母平均 m を信頼度 95% で推定する」ということがある。また、必要に応じて、信頼度 99% や 90% の信頼区間を求めることがある。

\bar{X} は確率変数であるから、標本から実際に得られる平均値 \bar{x} は、抽出される標本によって異なる。しかし、大きさ n の無作為標本を繰り返して抽出して、得られた平均値から信頼区間を多数作ると、その中には m を含むものが 95% あることが期待される。これが、信頼度 95% の信頼区間の意味である。

「母平均の推定」では、母標準偏差 σ を用いて、母平均を推定しているが、実際には σ の値もわからない場合が多い。標本の大きさ n が大きいときは、母標準偏差 σ の代わりに標本偏差 S を用いても差し支えないことが知られている。

次に、標本比率から母比率を推定する方法を調べる。ある特性 A の母比率が p である母集団から無作為に抽出された大きさ n の無作為標本について、特性 A の標本比率 R は、 n が大きいとき、近似的に正規分布 $N(p), p(1-p)/n$ に従う。したがって、母平均の推定と同様にして、

$$P\left(R - 1.96\sqrt{\frac{p(1-p)}{n}} \leq p \leq R + 1.96\sqrt{\frac{p(1-p)}{n}}\right) \quad (2)$$

式 (2) において、 n が大きいときは、大数の法則により R は p に近いとみなしてよい。よって、式 (2) の根号の中の p を R でおき換えることにより、次の結果が得られる：

母比率の推定

標本の大きさ n が大きいとき、標本比率を R とすると、母比率 p に対する信頼度 95% の信頼区間は

$$\left[R - 1.96\sqrt{\frac{R(1-R)}{n}}, R + 1.96\sqrt{\frac{R(1-R)}{n}} \right].$$

1.6 仮説検定

母集団分布に関する仮定を**仮説**という。**仮説検定**とは、標本から得られた結果によって、仮説が正しいかどうか判断する手法である。仮説が正しくないと判断することを、仮説を**棄却**するという。

仮説検定において、仮説を棄却する基準となる確率を**有意水準**という。それより確率が小さい事象が起こると仮説を棄却する (**危険率**ともいう)。

有意水準 α に対し、その範囲の確率が α になるような確率変数の値の範囲を、**棄却域**という。

実現した確率変数の値が棄却域に入れば仮説を棄却する。

棄却域を分布の両側にとる検定を**両側検定**といい、棄却域を分布の片側にとる検定を**片側検定**という。

2 演習問題

1. 二項分布 $B(8, \frac{1}{2})$ の平均と分散を求めよ。
2. 2枚の硬貨を同時に8回投げるとき、2枚とも表が出た回数を X とするとき、確率変数 X の従う二項分布を $B(n, p)$ の形に表せ。
3. 確率変数 X の確率密度関数 $f(x)$ が $f(x) = 2x$ ($0 \leq x \leq 1$) であるとき、確率 $P(0.3 \leq X \leq 0.5)$ と $P(0.2 \leq X \leq 0.8)$ を求めよ。この確率変数 X の期待値 $E(X)$ 、分散 $V(X)$ 、標準偏差 $\sigma(X)$ を、それぞれ求めよ。
4. 確率変数 Z が標準正規分布 $N(0, 1)$ に従うとき、確率 $P(-1 \leq Z \leq 2)$ を求めよ。
5. 1、2、3の数字を記入した玉が、それぞれ2個、3個、5個の計10個の袋の中に入っている。これを母集団として、次の問いに答えよ。
 - (1) 玉に書かれている数字の母集団分布を求めよ。
 - (2) 母平均 m 、母分散 σ^2 、母標準偏差 σ を求めよ。
6. 16歳の男子生徒の体重 X [kg] は、平均値 59.8kg、標準偏差 6.9kg の正規分布に従うという。この母集団から無作為に25人からなる標本を取り出すとき、その標本平均 \bar{X} の期待値と標準偏差を求めよ。
7. ある県の高校生を母集団とするとき、その身長は平均 165cm、標準偏差 4cm の正規分布をなしていた。この母集団から無作為に64人の標本を抽出したとき、その標本平均が 164cm 以上 166cm 以下である確率を求めよ。
8. ある試験を受けた高校生の中から、100人を任意に選んだところ、平均点は 58.3点であった。母標準偏差を 13.0点として、母平均の信頼区間を信頼度 95%で推定せよ。

3 標準正規分布表

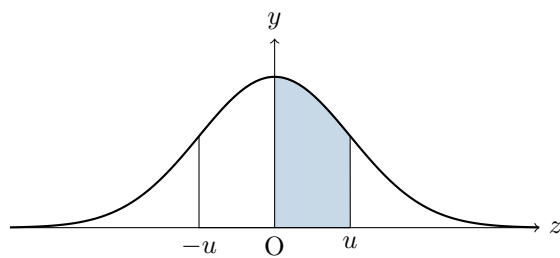


图 4 標準正規分布

表 5 正規分布表

u	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4874	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4943
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4972	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4980	0.4980	0.4981
2.9	0.4981	0.4982	0.4983	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990