統計学 2020 Lecture 6: 正規分布とその性質

北門　利英（東京海洋大学海洋生物資源学科）

2020 年 6 月 17 日

Attention:

- 授業 HP に授業に関する情報をアップデートしていきますので参照ください．(URL: https://toshihidekitakado.github.io/STAT2020/index.html)
- 今回およびこれまでの授業に関してわからないことがあれば，メールかリアルタイム接続時に遠慮なく質問してください（6 月 17 日も 13:00-14:00 とします）．
- 分からないことがあったら，そのままにしないで，毎回しっかり確認してください．

Point: 次の用語をしっかり理解すること．

- 離散型確率変数と連続型確率変数の違い
- 正規分布の定義とその性質，そして確率の計算ができること

1　導入

1.1　例題 1: 魚の体長の母集団の推測

例題 1: 魚の体長の母集団の推測

養殖場で商用に飼育している魚種があり，体長 20cm 以上が商品サイズとして考えられているとする．飼育場の母集団の体長分布を調べるとともに，20cm 以上の個体数の割合も知りたいとする．どうすればそれらを知ることができるであろうか？ただし，魚はすべて同じ年齢とする．

東京海洋大学の大泉ステーションではニジマスを飼育している．決して例題のような商用ではないが，ここでは大泉のニジマスを例にとって上記の問題について考えてみる．

大泉のニジマスは，冬に孵化し 1 年半経過した夏には，体長が 20 センチ近くまで成長する．成長の様子には個体差があり，体長の大きな個体もいれば，小さい個体もいる．もちろん平均値に近い体長をもつ個体の頻度が高い．このような様子を数学的に表現する 1 つの手段が正規分布の利用である．

ところで，いま知りたいのは大泉のニジマス全体（これを母集団とよぶ）の体長の分布であるが，すべてのニジマスの体長を測定するのは効率が良くない．そこで，通常は母集団から無作為に（ランダムに）個体を抽出し（これをサンプリングという），サンプル個体の体長を測定する．すなわち「一部を調べて全体の様子を知

る」というアイデアである．この考えは Lecture3 の視聴率調査でも述べたが，今回は知りたい対象が率だけではなく，母集団の体長分布なので少し設定が異なるが，「一部を調べて全体の様子を知る」という考えでは同様である．このようなサンプリングというアイデアを模式図で表したのが図 1 である．

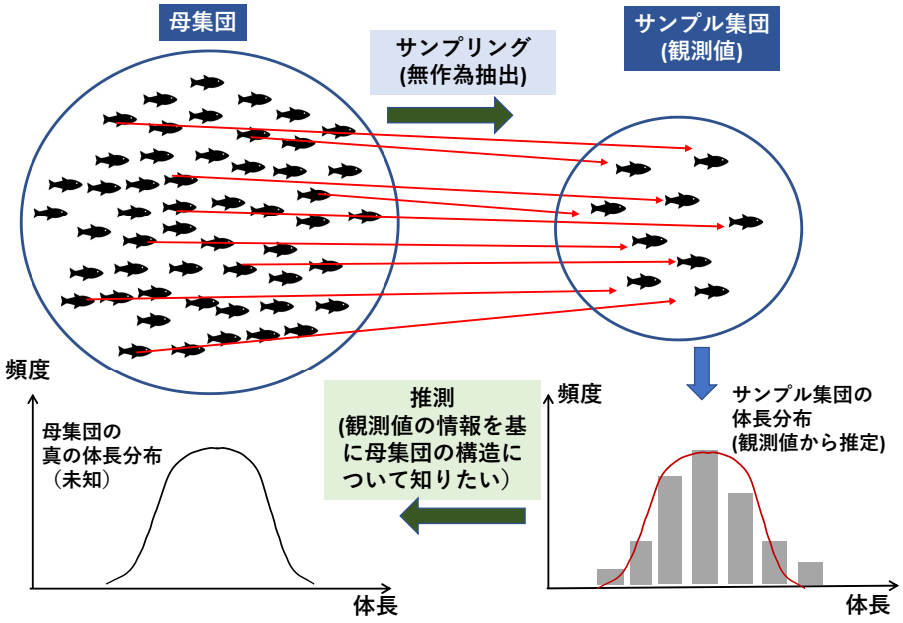


図1　母集団からのサンプリングの概念図

以下は令和元年の実データの度数分布表とヒストグラムである．サンプルの数は約 1000 である．そのヒストグラムに，データから推定した正規分布を重ね合わせたのが図 2 である．非常に綺麗に当てはまっていることが分かる．

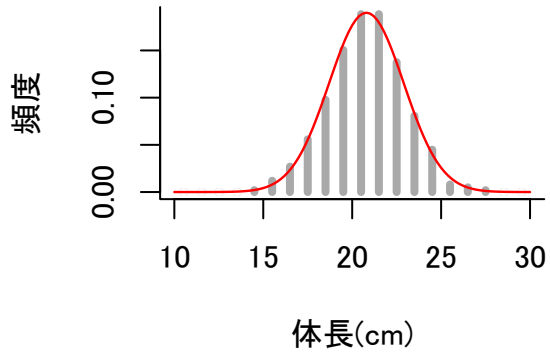


図2　大泉ニジマス体長測定結果 (令和 1 年度) と正規分布の当てはめ結果

実際，このデータから平均値 $\mu = 20.8$, 標準偏差 $\sigma = 2.1$ と推定される．また，20cm 以上の魚の割合は

$$P(Y \geq 20) \fallingdotseq 0.648$$

となることが分かる．さてどのように計算したのでしょうか．次の節では，その正規分布について説明する．

2 正規分布とは

2.1 連続型確率分布における確率

正規分布は連続型確率分布の中で最も重要な役割を果たす．Lecture 2 で学んだことであるが，連続型の確率分布の場合，$Y = y$ のように各値に対して確率を定義することができず，したがって

$$P(a \leq Y \leq b) = \int_a^b f(y)dy$$

のように区間に対して確率密度関数 $f(y)$ を積分して確率を求める．

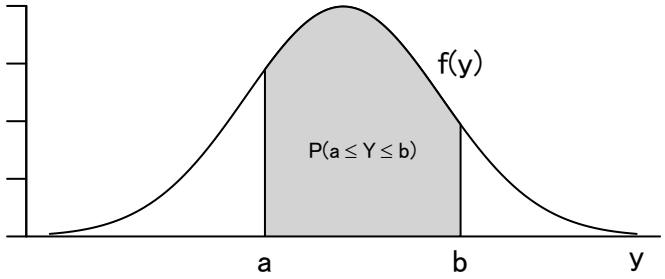


図3 連続型確率分布における確率の計算イメージ

2.2 正規分布の定義

確率密度関数 $f(y)$ が確率分布を規定しており，特に以下の形の確率密度関数を持つとき，確率変数は正規分布にしたがうという．

連続型分布 1 確率変数 Y の確率密度関数が

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < \infty \quad \left(\begin{array}{c} -\infty < \mu < \infty \\ 0 < \sigma < \infty \end{array} \right) \tag{1}$$

となるとき，Y は正規分布 (normal distribution) $N(\mu, \sigma^2)$ にしたがうという．

この $f(y)$ はパラメータ $y = \mu$ に関して対称な釣鐘型の関数である．また，μ と σ^2 はそれぞれ，分布の位置および広がりを表現するパラメータ，すなわち期待値および分散を表すパラメータである．

連続型の場合の期待値は離散型のシグマ記号が積分記号になっただけなので，一見難しく見えるが決してそうではない．

定義 1　[確率変数の期待値]　確率変数 Y の期待値は次のように定義される．

$$E[Y] = \begin{cases} \displaystyle\sum_{y=0}^{\infty} y f(y) & \text{(離散型確率分布のとき)} \\ \displaystyle\int_{-\infty}^{\infty} y f(y) dy & \text{(連続型確率分布のとき)} \end{cases} \tag{2}$$

分散の定義も同様で

定義 2　[確率変数の期待値]　確率変数 Y の期待値は次のように定義される．

$$V[Y] = \begin{cases} \displaystyle\sum_{y=0}^{\infty} (y - E[Y])^2 f(y) & \text{(離散型確率分布のとき)} \\ \displaystyle\int_{-\infty}^{\infty} (y - E[Y])^2 f(y) dy & \text{(連続型確率分布のとき)} \end{cases} \tag{3}$$

である．

計算の詳細は割愛するが，確率変数 Y が正規分布 $N(\mu, \sigma^2)$ にしたがうとき，$E[Y] = \mu$ と $V[Y] = \sigma^2$ である．必ず覚えて下さい．

2.3　正規分布の確率密度関数の概形

正規分布 $N(\mu, \sigma^2)$ の確率密度関数の概形は以下の通り．平均 μ に対して左右対称．また，平均値が違っても平行移動するだけで形は変わらない．また分散が大きくなると密度関数が平たくなるが，これは散らばり具合が大きくなることと，全体を積分すると 1 になるからである．

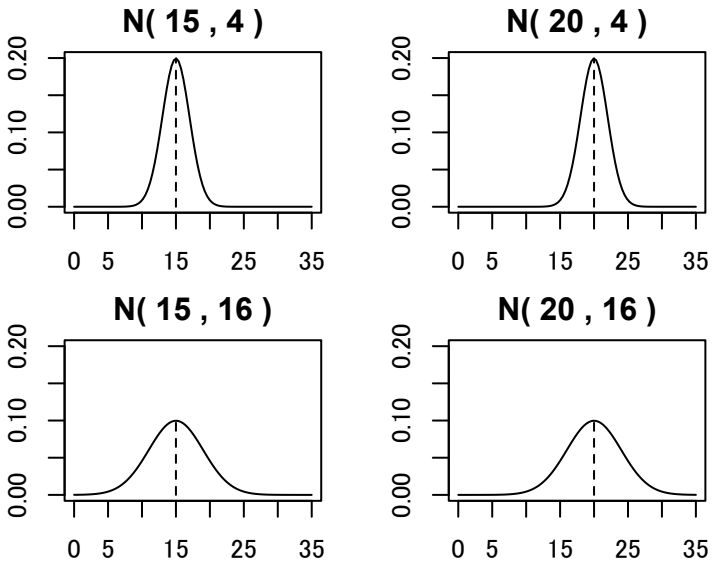


図4　正規分布の確率密度関数の概形

☞ 正規分布の確率密度関数の式，概形，そしてパラメータの意味は必ず覚えること．

正規分布の密度関数に対して

$$\int_{-\infty}^{\infty} f(y) = 1 \tag{4}$$

が成り立つが，数学的な証明はこの授業のスコープから外れるので割愛する (例えば極座標変換を利用して確かめることができる．詳しくは「数理科学入門」（恒星社厚生閣）の第 1 章など参照のこと).

2.4 正規分布にしたがう「確率変数の観測値」と「確率分布」の関係

確率分布と観測値との関係をもう少し理解するために，以下のようなアニメーション資料を用意した．ここでは，平均値 $\mu = 20$, 分散 $\sigma^2 = 2^2$ （標準偏差 $\sigma = 2$) を仮定し，この正規分布 $N(\mu, \sigma^2)$ にしたがう観測値（乱数）を生成してみる．

```
[1] 17.58587 20.55486 22.16888 15.30860 20.85825 21.01211 18.85052 18.90674
[9] 18.87110 18.21992
```

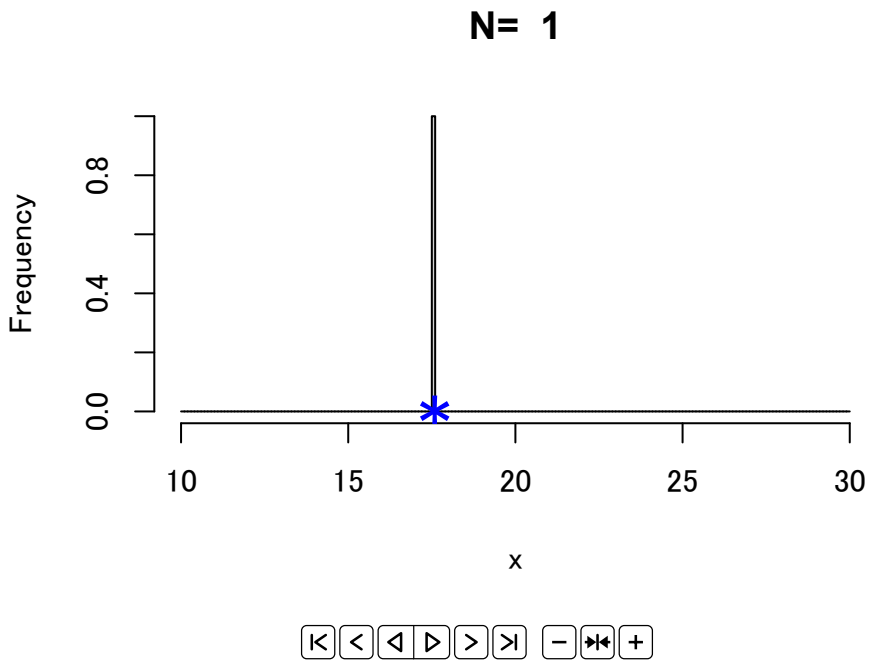


図5 正規分布 N(20,4) にしたがう乱数を 1 個ずつ生成し順番にヒストグラムとして積み上げたアニメーション．全部で 200 個の乱数を生成.

☞ アニメーションの再生には右向き三角のボタンを押してください.

次に，乱数の数を 10000 個まで増やし，元の正規分布の確率密度関数と重ね合わせた．乱数の数が多くなるとヒストグラムが確率密度関数とほぼ一致することが分かる．すなわち，データの数を大きくすると，母集団の近似としてどんどん正確になることが分かる.

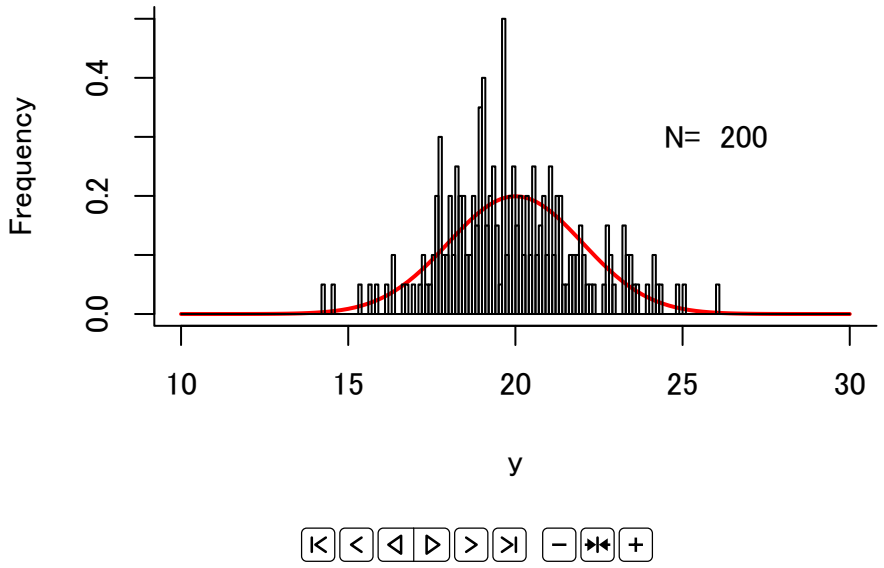


図6　正規分布 N(20,4) にしたがう乱数を 200 個ずつ生成し順番にヒストグラムとして積み上げたアニメーション．全部で 10000 個の乱数を生成．赤線は母集団分布 N(20,4) の確率密度関数.

☞ 次のサイトにも正規分布の補助教材を保存しています.

https://kitakado.shinyapps.io/Lecture06_S2/

3　正規分布の性質

Lecture 6 では次の標準正規分布を覚えて下さい．正規分布の特別な形ですが，非常に重要です.

3.1　標準正規分布

性質 1　確率変数 Y が $N(\mu, \sigma^2)$ にしたがうとき,

$$Z = \frac{Y - \mu}{\sigma} \tag{5}$$

は $N(0,1)$ にしたがう．この変換を標準化 (standardization)，また $N(0,1)$ を標準正規分布 (standard normal distribution)という.

上記の性質は，積分の変数変換 (1 変数の場合はいわゆる置換積分) を利用して示すことができる．すなわち，$z = (y - \mu)/\sigma$ とおくと $dz/dy = 1/\sigma$ より

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2}} \frac{dy}{dz} dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

となり，したがって Z が標準正規分布の $N(0,1)$ の密度関数をもつことがわかる．

定理 1 確率変数 Y が $N(\mu,\sigma^2)$ にしたがうとき，$a+bY$ は $N(a+b\mu,b^2\sigma^2)$ にしたがう．

さて，最初の例題に戻ると，知りたい確率は $Y\sim N(20.8,2.1^2)$ に対して $P(Y\geq 20)$ の確率であった．この確率は標準正規分布の性質を用いて次のように計算することができる．

$$
\begin{aligned}
& P(Y\geq 20)\\
= \ & P\left(\frac{Y-20.8}{2.1}\geq\frac{20-20.8}{2.1}\right) \quad \text{(括弧の中は事象を表すので両辺四則演算可，標準化する)}\\
= \ & P(Z\geq -0.381) \quad (Z=\tfrac{Y-20.8}{2.1}\text{とおくと}Z\sim N(0,1))\\
= \ & P(-0.381\leq Z\leq 0)+P(Z\geq 0) \quad (Z\sim N(0,1)\text{の数表（付録）を用いるために確率を分割})\\
= \ & P(0\leq Z\leq 0.381)+0.5 \quad (Z\sim N(0,1)\text{の確率密度関数は 0 に対して左右対称})\\
\fallingdotseq \ & P(0\leq Z\leq 0.38)+0.5 \quad (P(0\leq Z\leq 0.38)\text{は付録の数表から 0.148 ともとまる})\\
= \ & 0.648
\end{aligned}
$$

(注)1 点の確率は 0 なので不等号の「等号」は気にしないでください．

3.2 正規分布の再生性

正規分布にしたがう独立な確率変数の和もまた正規分布にしたがう．独立同一な n 個の正規分布にしたがう確率変数の和もまた正規分布にしたがう．これらの性質は区間推定と仮説検定のところで再度説明するので，今回は定理の存在だけ覚えておいてください．

定理 2 確率変数 Y_1,Y_2 が独立でそれぞれ $N(\mu_1,\sigma_1^2)$, $N(\mu_2,\sigma_2^2)$ にしたがうとき，aY_1+bY_2 は $N(a\mu_1+b\mu_2,a^2\sigma_1^2+b^2\sigma_2^2)$ にしたがう．

定理 3 確率変数 $Y_1,Y_2,\ldots,Y_n$ が独立同一に $N(\mu,\sigma^2)$ にしたがうとき，$\bar{Y}=\sum_{i=1}^{n}Y_i/n$ は $N(\mu,\sigma^2/n)$ にしたがう．

4 番外編：正規分布の混合

これは定期試験には出ませんが，水産のデータを扱う際にはよく利用するものです．

例 1 ヒトの身長や魚の体長などは正規分布にしたがうと考えられる典型的な例である．年齢が経過するにつれ母集団の平均身長は大きくなる．したがって，この場合，正規分布の期待値 μ は年齢に依存する．特に水産生物の場合，生まれてからの経過年数 t の関数として von Bertalanffy 式

$$\mu(t)=L_\infty\left\{1-e^{-K(t-t_0)}\right\}$$

を利用することが多い．また，年齢が大きいほど個体間の身長や体長のばらつきも大きくなるであろう．したがって分散 σ も年齢と関係したパラメータと考えられる．下図は $L_\infty=50,K=0.5,t_0=0$ の下でのシミュレーションデータである．

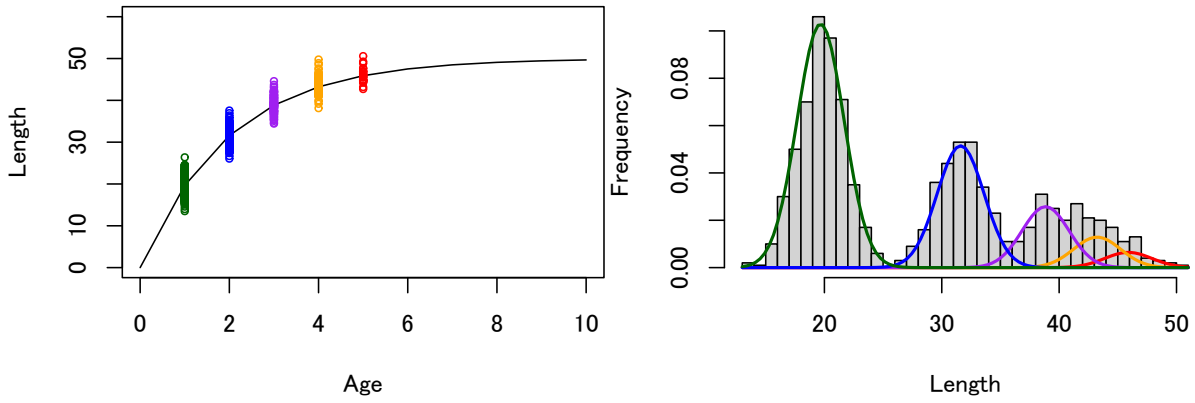


図7　異なる年齢の体長データと混合正規分布の様子

5　演習問題

標準正規分布表を利用すること．

練習問題 1　ある魚の体長 Y(cm) が正規分布 $N(20, 5^2)$ にしたがうとき, $P(10 \leq Y \leq 30)$ および $P(Y > 25)$ を求めよ．(答え:0.9544 と 0.1587)

練習問題 2　統計学の定期試験の点数が正規分布 $N(65, 4^2)$ に従うとする．60 点以上で合格とするとき，合格者の割合はいくらか？(フィクションです) (答え:0.8944)

練習問題 3　あるアザラシ種は，資源量が 1000 頭以下である確率が 30 パーセント以上あるとき，絶滅危惧種とみなされるとする．本種に対する最新の調査報告書によると，資源量推定値の自然対数が確率分布 $N(7.2, 0.6^2)$ に従うと記載されていた．この調査の結果，この種は絶滅危惧種とみなされるか？(これもフィクションです)

クジラの資源量を N とおくとき，調査の結果，$Y = \log N \sim N(7.2, 0.6^2)$ が成り立つ．いま知りたい確率は $P(N \leq 1000)$ であるから

$$\begin{aligned} P(N \leq 1000) &= P(\log N \leq \log 1000) = P(Y \leq \log 1000) = P\left(\tfrac{Y-7.2}{0.6} \leq \tfrac{\log 1000 - 7.2}{0.6}\right) \\ &= P(Z \leq -0.487) \fallingdotseq P(Z \leq -0.49) = 0.3121 \end{aligned}$$

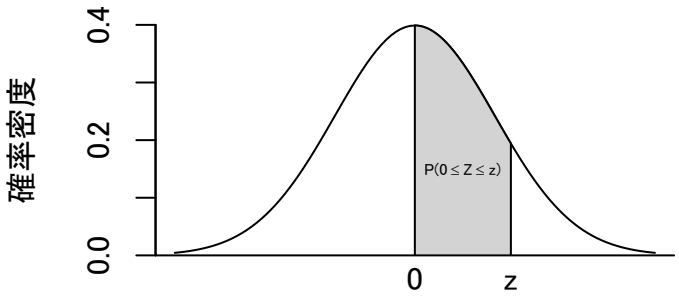
と計算され，したがって本種は絶滅危惧種とみなされることになる．

次の問題は提出課題です．次回以降の提出となりますので今回は提出不要です．

練習問題 4　あるチョコレート工場で 100g の板チョコを生産しているが，製品によってばらつきが生じ，正規分布 $N(102, 2^2)$ に従うとされている．100g 未満のチョコレートは出荷できないとき，生産したチョコレートの何%が不良品となるか？また，不良品率を 1 パーセント以下にしたいとき，板チョコの重さの平均値をいくらにするように生産工程を変えればよいか？(平均値を変えてもばらつき，すなわち分散は変わらないとする)

付録: 標準正規分布表 (このページは試験の際に印刷して持ち込むこと)

以下の表は，$Z \sim N(0,1)$ の標準正規分布について，$I(z) = P(0 \leq Z \leq z)$ の値を各 z の値に対して与えている．たとえば $z = 1.96$ における $I(z)$ の値を知りたい場合には，表の行 1.9，列 0.06 の交わったところの数値を見れば $I(1.96) = 0.4750$ であることが分かる．



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
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	0.4875	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	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	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