MAT120：数量的推論　第14講ハンドアウト
正規分布

前回の講義では，データをどのように記述するかを学ぶことで統計学を始め，中心傾向と変動という二つの中核的な考え方を導入しました。今回の講義では，最も有名な連続型確率分布である正規分布を学びます。正規曲線とは何か，標準正規分布とは何か，そしてz-score を使ってどのように異なる正規分布のあいだを移ることができるのかを学びます。また，なぜ連続変数に対する確率は棒の高さではなく曲線の下の面積 によって測られるのかも学びます。次回の講義では，これらの道具を使って仮説検定を始めます。

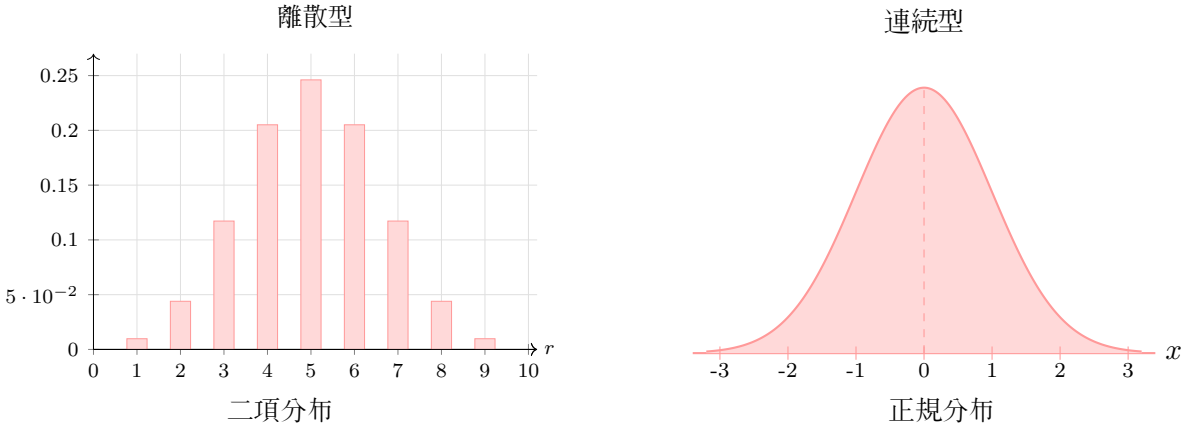
1　正規分布とは何か

1.1　離散型確率変数と連続型確率変数

確率変数 とは，その値が偶然によって決まる変数のことです。この段階では，次の二つの重要な型を覚えておけば十分です。

- **離散型確率変数：** $0, 1, 2, 3, \ldots$ のような飛び飛びの値を取る。
- **連続型確率変数：** ある区間の中の任意の実数値を取りうる。

第12講では，離散型確率変数から作られる確率分布，すなわち二項分布を学びました。今日は，その代わりに連続型確率変数から作られる確率分布へ進みます。



連続型確率分布は，これまで見てきた離散型分布よりも多くの構造を持っています。それはつまり，話すべきことがより多いということであり，したがってより混乱しやすくもあります。そこで，私たちの方針は，極限 を使って連続の場合を，すでに知っていることと結びつけながら学ぶことです。

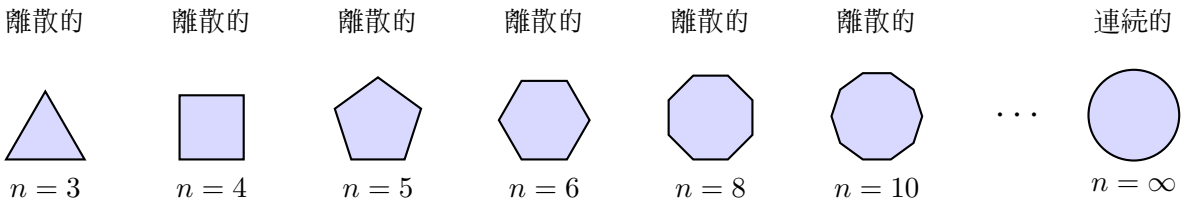
1.2　極限のたとえ

断面の形がそれぞれ異なる丸太がたくさんあるとしましょう。

最初の丸太は断面が三角形なので，転がすのがとても不便です。次の丸太は断面が四角形ですが，少し転がしやすくなっています。その後，五角形の断面をもつ丸太はさらに転がしやす

く，六角形の断面をもつ丸太はそれよりもっと転がしやすくなります。断面の辺の数が増えるにつれて，丸太はどんどん転がしやすくなることが分かります。もちろん，完全な円形の丸太が最も転がしやすいでしょう。

もし丸太が非常にたくさんあって，辺の数をどんどん増やしていくと，次のような列が得られるかもしれません。

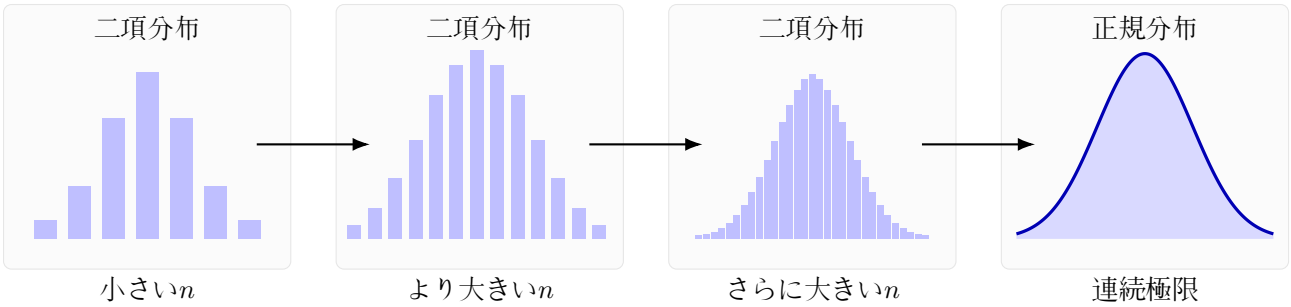


この操作の極限は円になります。円は，無限に多くの辺をもつ図形と考えることができます。辺の数を離散量として解釈すると，図形に辺をどんどん加えていく過程の中で，多角形は極限において連続的な図形へ近づいていくことが分かります。

1.3 極限過程

丸太を転がすたとえには，確率変数にも応用できる深い考え方があります。すなわち，試行回数を増やし，したがって棒の本数も増えていくような二項分布の列を考えることで，正規分布に「近づく」ことができるのです。言い換えれば，連続的な曲線は，多数の離散的な棒の極限として記述できるのです。

二項分布では，グラフの棒の総数は$n + 1$ 本であることを思い出してください。可能な値が0 からn まで並ぶからです。n が大きくなると，二項分布の棒はより細く，より多くなります。分布全体はますます滑らかに見えるようになり，やがて曲線のように見え始めます。

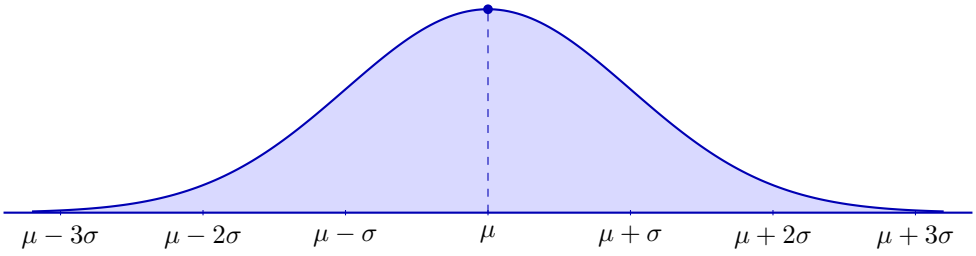


この極限で現れる鐘型の曲線を正規分布 と呼びます。

1.4 正規分布と正規曲線

正規密度関数の グラフは正規曲線 と呼ばれます。その形から，しばしばベル曲線 とも呼ばれま

す。



正規分布は統計学において最も重要な対象の一つです。実際には，人間の特徴，物理的測定値，工業データ，測定誤差など，多くの文脈で正規分布が現れます。

正規分布にはいくつかの重要な特徴があります。

正規分布の重要な性質

1. 曲線は鐘型であり，その最も高い点は平均μ の上にある。
2. 曲線はμ を通る鉛直線に関して対称である。
3. 両端の裾は水平軸に近づくが，決して接したり交わったりしない。
4. 曲線全体の下の面積は1 である。

1.5 数学的記述

正規分布はグラフで表されるので，対応する関数が何であるかを問うことができます。この関数には特別な名前があり，正規密度関数 と呼ばれます。私たちはこの関数を直接扱う必要はほとんどありませんが，次のように書かれます。

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

この関数は見た目にはかなり複雑ですが，いくつか重要な観察をしておくべきです。まず，この関数には変数x が一つだけあります。これはグラフの横軸になる連続型確率変数です。次に，x 以外に，関数$f(x)$ が必要とする重要な数が二つあることが分かります。

- μ は分布の平均である。
- σ は分布の標準偏差である。

上の関数を見ると分かるように，μ とσ は（もちろんπ のような定数は別として）この関数の中に現れる唯一のパラメータ情報です。したがって，すべての正規分布は，数μ とσ によって完全に特徴づけられる のです。つまり，μ とσ の選び方が異なれば，異なる正規曲線が得られます。このため，正規分布はしばしば次のように書かれます。

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

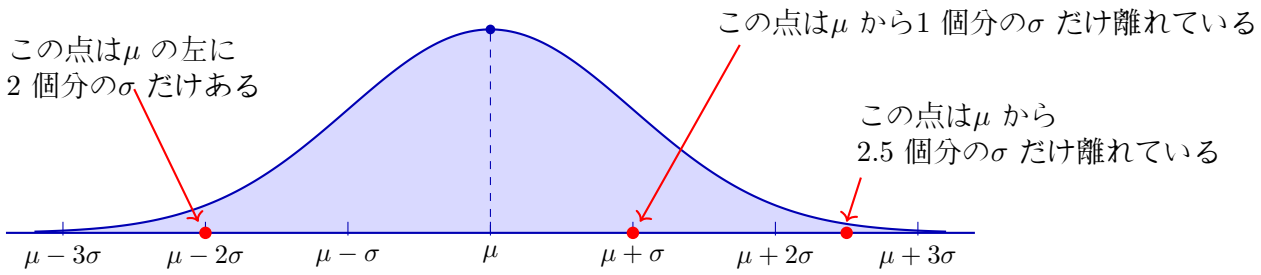
言葉で言えば，これは「連続型確率変数X は，平均μ，分散σ^2 をもつ正規分布に従う」と読みます。

これは正規分布の非常に特別な特徴です。母平均μ は分布の中心がどこにあるかを表し，母標準偏差σ はデータの広がり具合を測る量です。ふつう，これらはデータセットを記述するための二つの数にすぎません。しかし，正規分布の場合には，μ とσ という二つの数が分布を完全に記述する のです。したがって，曲線の上を読み取るために必要な情報は，記号μ とσ のみなのです。

2 標準正規分布

2.1 z-score

ここまでの議論から，任意の正規分布$\mathcal{N}(\mu, \sigma^2)$ のx 軸は，μ とσ を用いて記述できることが分かります。もし確率変数X の値をランダムに一つ選ぶとすると，その値はx 軸上のどこかに落ちます。x 軸はμ を中心とし，σ を単位として測られているので，その値が中心からどれだけ離れているかという形で位置を表すことができます。たとえば次のようになります。



z-score は，標準得点 とも呼ばれ，この情報を純粋な数に変換します。すなわち，ある測定値が平均から何個分の標準偏差だけ離れているかを表します。

z-score の定義

もしx が平均μ，標準偏差σ をもつ正規分布からの値ならば，そのz-score は

$$z = \frac{x - \mu}{\sigma}.$$

この公式は，部分ごとに次のように解釈できます。

- $x - \mu$ は，x が平均からどれだけ離れているかを測る。
- σ で割ることで，その距離を「標準偏差何個分か」に変換する。
- 結果は正にも負にもなり，それはそれぞれ平均の右側か左側かに対応する。

たとえば，上の図では，

- $\mu - 2\sigma$ にある点のz-score は-2 になる。
- $\mu + \sigma$ にある点のz-score は1 になる。
- $\mu + 2.5\sigma$ にある点のz-score は2.5 になる。

逆に，z-score が分かっていて元の値x を求めたいこともあります。公式

$$z = \frac{x - \mu}{\sigma}$$

から出発して整理すると，

$$x = \mu + z\sigma.$$

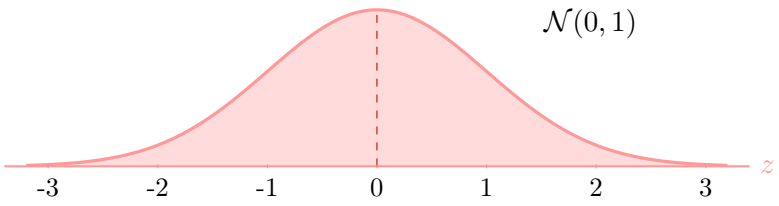
が得られます。z-score と元の得点は，同じ情報をただ二つの異なる書き方で表しているにすぎません。

2.2 標準正規分布

標準正規分布 とは，平均が0，標準偏差が1 の正規分布です。ふつう，これは

$$Z \sim \mathcal{N}(0, 1).$$

と書かれます。ここでは，標準正規分布に対応する確率変数には，先ほどまで使っていたX ではなく，特別にZ という文字を使うことに注意してください。慣例として，標準正規分布$\mathcal{N}(0, 1)$ は赤 で描き，より一般の正規分布は青 で描くことにします。こうすることで，視覚的に区別しやすくなります。標準正規分布は次のように描かれます。



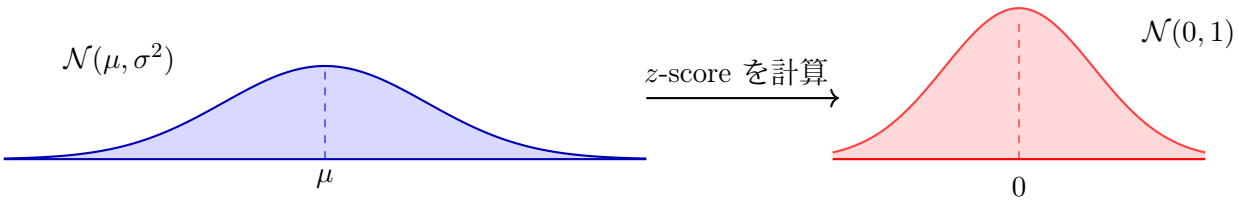
ここで横軸を見ると，中心が0 であり，そこから1 刻みで左右に数えていることが分かります。

2.3 なぜ標準正規分布が役に立つのか

どの正規分布も，測定値をz-score に変換することで標準正規分布に変換することができます。これこそが標準正規分布が重要である理由です。正規曲線ごとに異なる面積表を覚える代わりに，一つの普遍的な基準曲線を使えばよいのです。

中心となる考え方

すべての正規分布は，標準正規の世界へ移し替えることができる。そうしてしまえば，確率は一つの標準正規分布の面積表から読み取ることができる。



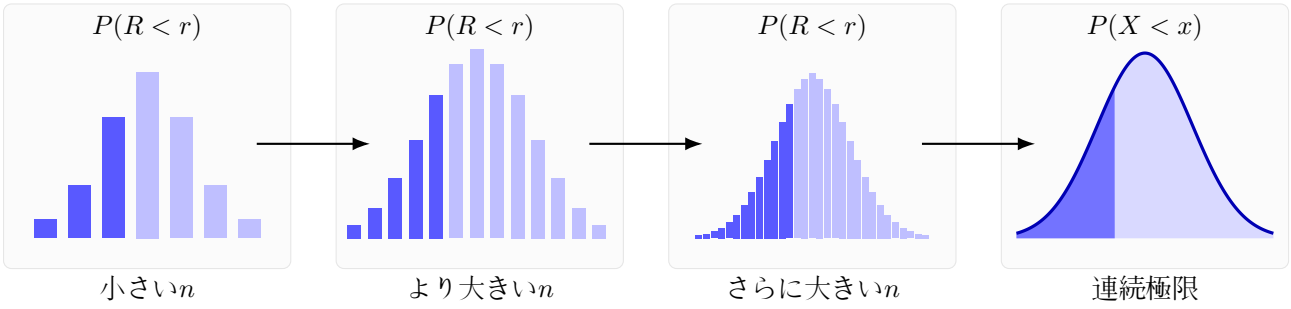
3　面積と確率

3.1　棒の高さから面積へ

二項分布を扱うとき，各棒の高さは，離散型確率変数がその値を取る正確な確率を表していました。たとえば，$n = 10$，$p = 0.5$ の二項分布では，

- ちょうど$r = 5$ となる確率は，$r = 5$ の棒の高さであり，約0.24 である。
- $r < 5$ となる確率は，$r = 0, 1, 2, 3, 4$ の棒の高さを足し合わせることで得られ，約0.38 になる。

n がどんどん大きくなると，この多数の細い棒を足し合わせる過程は，しだいに連続的な面積のように見えてきます。



上の図が示しているように，連続型確率変数では，確率は棒の高さ（あるいはその和）ではなく，曲線の下の面積 として考えます。

離散型確率と連続型確率

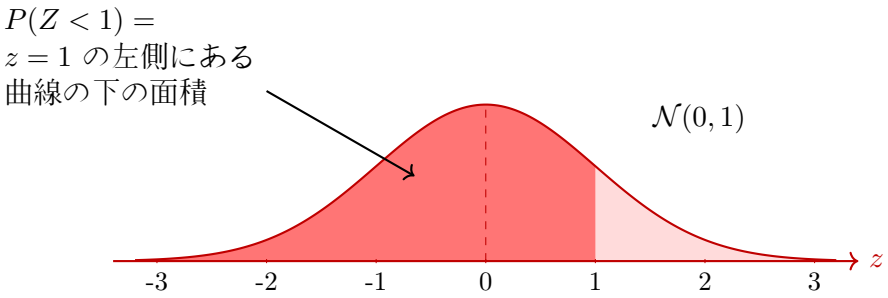
- **離散型** の場合，確率は個々の結果の確率を足し合わせることで求められる。
- **連続型** の場合，確率は曲線の下の面積を測ることで求められる。

3.2　面積はどのように求めるか

再び標準正規分布$\mathcal{N}(0, 1)$ を考えましょう。上の議論によれば，

$$P(Z < z),$$

と書くとき，それは標準正規曲線のうち，点z の左側にある面積を意味します。たとえば，確率$P(Z < 1)$ は次の塗られた面積で表されます。



もし微積分を知っていれば，これらの面積は積分を用いて求めることができます。しかし，これはしばしば難しいので，実際にはz 表 を使うことが多いです。

z 表とは，標準正規分布のよく使われる領域に対する答えを一覧にしたものだと考えることができます。この講義では，主に次の表を使います。

z	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
$P(Z<z)$	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.98

ここで，二行目は特定のz の値に対して，その左側の面積を与えています。たとえば，上の図では，この表は塗られた面積が0.84 であることを教えてくれます。つまり，曲線全体の下の面積のうち84% が$z=1$ の左側にあるということです。確率の言葉で言えば，確率変数Z が1 より小さい値を取る確率がおよそ84% であるということです。

3.3 面積の理解

z 表は，標準正規確率変数Z に対する標準正規曲線の下の面積を一覧にしたものです。この表は，

$$P(Z<z),$$

すなわちz の左側 の面積についての近似値を与えるものと解釈します。正規分布の基本的性質を使えば，この表から他の面積も記述できます。

他の面積の求め方

- **z の左側の面積：** $P(Z<z)$ をそのまま読む。
- **z の右側の面積：** $P(Z>z)=1-P(Z<z)$ を使う。
- **二つの値のあいだの面積：** $P(a<Z<b)=P(Z<b)-P(Z<a)$ を使う。
- **負の値：** 対称性より，$a>0$ に対して，

$$P(Z<-a)=1-P(Z<a).$$

連続型確率変数では，ただ一つの正確な点 を取る確率は0 です。したがって，たとえば

$$P(Z=0.5)=0.$$

これは，その点が不可能だという意味ではありません。ただ，一つの点は幅を持たないので，面積も0 になるという意味です。

3.4　他の正規分布に対する確率の計算

第2節で，z-score の公式

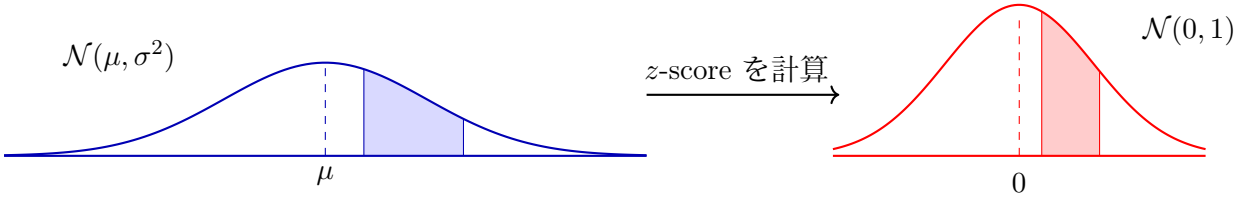
$$z = \frac{x - \mu}{\sigma}$$

を用いれば，任意の正規分布$\mathcal{N}(\mu, \sigma^2)$ を標準正規分布$\mathcal{N}(0, 1)$ に変換できると述べました。ここで有用なのは，この変換が面積の比率も保つということです。つまり，$\mathcal{N}(\mu, \sigma^2)$ における確率を，z-score と上のz 表を使って求められるということです。確率変数$X \sim \mathcal{N}(\mu, \sigma^2)$ に対しては，

$$P(X < x) = P\left(Z < \frac{x - \mu}{\sigma}\right).$$

が成り立ちます。

言い換えれば，$P(X < x)$ の値を求めるには，点x のz-score を計算し，その対応する面積を上のz 表で調べればよいのです。この考え方は他の面積にも同様に使えます。たとえば，$P(x_1 < X < x_2)$ のような確率も，x_1 とx_2 に対するz-score を用いて同じように求めることができます。



これこそがz-score の本当の強みです。標準正規表を使って，どんな 正規分布でも扱えるようにしてくれるのです。

例題

$X \sim \mathcal{N}(10, 2^2)$ とする。ランダムに選ばれた値が11 と14 のあいだに入る確率を求めなさい。記号では$P(11 \leq X \leq 14)$ を求めよ。

解答

Step 1: x の値をz の値に変換する。

$$z_1 = \frac{11-10}{2} = 0.5, \qquad z_2 = \frac{14-10}{2} = 2.0.$$

Step 2: 表から左側確率を読む。

$$P(Z < 0.5) = 0.69, \qquad P(Z < 2.0) = 0.98.$$

Step 3: 引き算して区間の確率を求める。

$$P(11 \le X \le 14) = P(Z < 2.0) - P(Z < 0.5) = 0.98 - 0.69 = 0.29.$$

したがって，求める確率は

$$0.29$$

である。言葉で言えば，この正規モデルからランダムに選ばれた値が11 と14 のあいだに入る確率は約29% です。

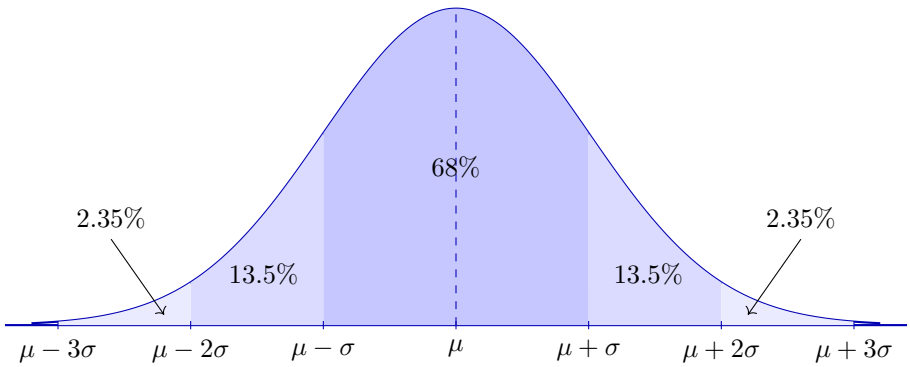
3.5 68-95-99.7 ルール

正規分布では，標準偏差がデータについて非常に有用な幾何学的目安を与えます。

経験則

正規分布では，

- データのおよそ68% は平均から1 標準偏差以内にある。
- データのおよそ95% は平均から2 標準偏差以内にある。
- データのおよそ99.7% は平均から3 標準偏差以内にある。



残りの0.3% は両端の極端な裾にあり，それぞれの裾に約0.15% ずつ入っています。

3.6 例題：標準正規曲線の下の面積

上の簡易z 表と対称性を使って，次の面積を求めなさい。

問題

(a) $z = -1.00$ の左側の面積を求めよ。
(b) $z = 2$ の左側の面積を求めよ。
(c) z 表を使わずに，$z = 0$ の左側の面積を求めよ。
(d) z 表を使わずに，曲線全体の下の面積を求めよ。
(e) $z = 1$ の右側の面積を求めよ。
(f) 区間$-1 < Z < 2$ の面積を求めよ。

解答

(a) 表より，

$$P(Z < -1.00) = 0.16.$$

(b) 表より，

$$P(Z < 2) = 0.98.$$

(c) 対称性より，面積全体の半分が0 の左側にあるので，

$$P(Z < 0) = 0.50.$$

(d) 曲線全体の下の面積は常に1 である。

(e)

$$P(Z > 1) = 1 - P(Z < 1) = 1 - 0.84 = 0.16.$$

(f)

$$P(-1 < Z < 2) = P(Z < 2) - P(Z < -1).$$

表を使うと，

$$P(-1 < Z < 2) = 0.98 - 0.16 = 0.82.$$